Machine Learning Explained: A Practical Guide to Data-driven Decision Making

ABDELHAMID ZAIDI
RENATO RACELIS MAALIW III
MRS. K. P. MAHESWARI
DR. HAEWON BYEON

Xoffencer

MACHINE LEARNING EXPLAINED: A PRACTICAL GUIDE TO DATA-DRIVEN DECISION MAKING

Authors:

- Abdelhamid ZAIDI
- Renato Racelis Maaliw III
- Mrs. K. P. Maheswari
- Dr. Haewon Byeon

Xoffencer

www.xoffencerpublication.in

ISBN-13: 978-81-19534-45-6 (Paperback)

Publication Date: 30 October 2023

MRP: ₹450/-

Published by:

Xoffencer International Publication

Behind shyam vihar vatika, laxmi colony

Dabra, Gwalior, M.P. – 475110

Cover Page Designed by:

Satyam soni

Contact us:

Email: mr.xoffencer@gmail.com

Visit us: www.xofferncerpublication.in

Author Details

Abdelhamid ZAIDI

Abdelhamid ZAIDI is an Associate Professor in the College of Science at Qassim University in Saudi Arabia. He has a PhD in Statistics and Stochastic Modeling from University Grenoble-Alpes (France) and an Engineering degree in Computer Science and Applied Mathematics from ENSIMAG Grenoble (France). He works mainly on the development of computational methods applied to various subjects of signal and image processing and artificial intelligence. He also has many contributions in the field of artificial intelligence. His research work was published in many top ranked journals. He is also the author of three books covering numerical analysis, algorithmic, probability, and statistics.

Renato Racelis Maaliw III

Renato Racelis Maaliw III is an Associate Professor and Researcher at the College of Engineering in Southern Luzon State University, Lucban, Quezon, Philippines. He has a doctorate degree in Information Technology with specialization in Machine Learning, a Master's degree in Information Technology with specialization in Web Technologies, and a Bachelor’s degree in Computer Engineering. His area of interest is in artificial intelligence, computer engineering, web technologies, software engineering, data mining, machine learning, and analytics. He has published original research articles, a multiple time best paper awardee for various IEEE sanctioned conferences; served as technical program committee for world-class conferences, author, editor and peer reviewer for reputable high-impact research journals.

Mrs. K. P. Maheswari

Mrs. K. P. Maheswari MCA., M.Phil., NET (Computer Science) is Assistant Professor of Computer Applications at Fatima College, Madurai. She has rich experience in teaching. She has received "Women Transforming Nation Awards 2023 – Certificate of Appreciation for Dedication" from Women Lead. Her research interests include Machine Learning, Deep Learning, Artificial Intelligence and Network Security. She has authored book chapters and several publications in a reputed journals. She has presented papers in various Conferences and seminars at National and International level. She has also served as a subject matter expert for workshops and seminars.Her professional achievements have included obtaining Microsoft (MTA - Microsoft Technology Associate) International certifications in Python Programming, HTML 5, and Security Fundamentals.

Dr. Haewon Byeon

Dr. Haewon Byeon received the Dr Sc degree in Biomedical Science from Ajou University School of Medicine. Haewon Byeon currently works at the Department of Medical Big Data, Inje University. His recent interests focus on health promotion, AI-medicine, and biostatistics. He is currently a member of international committee for a Frontiers in Psychiatry, and an editorial board for World Journal of Psychiatry. Also, He were worked on 4 projects (Principal Investigator) from the Ministry of Education, the Korea Research Foundation, and the Ministry of Health and Welfare. Byeon has published more than 343 articles and 19 books.

Preface

The text has been written in simple language and style in well organized and systematic way and utmost care has been taken to cover the entire prescribed procedures for Science Students.

We express our sincere gratitude to the authors not only for their effort in preparing the procedures for the present volume, but also their patience in waiting to see their work in print. Finally, we are also thankful to our publishers **Xoffencer Publishers, Gwalior, Madhya Pradesh** for taking all the efforts in bringing out this volume in short span time.

Contents

CHAPTER 1

INTRODUCTION

During the course of the process of making a choice, we rely on a variety of presumptions, premises, and the circumstances; all of this is directed by the goal that is related with the decision itself. However, the premises and the knowledge of the corporation are dependent on our data since they are an essential component of our organization as a system. The context and the assumptions are both external factors that are beyond the control of any decision maker. Both the background and the assumptions represent outside forces that are not within the control of any decision maker. A prominent example of a conceptual error is the misunderstanding that exists between data and information, which in reality correspond to entirely distinct ideas. This misunderstanding is a common occurrence. In point of fact, information and data cannot in any way be substituted for one another in any context.

To put this another way, there is no guarantee that the data will be consistent, comparable, or traceable, despite the fact that we are able to collect data from a broad variety of distinct data sources. This is because there are so many diverse data sources. Because of this, in order for us to make a decision, we need to have a good comprehension of both the component that is presently being examined and the data that is linked with it at the present time. Only then will we be able to make an informed choice. The identification of the system itself is the first step that must be taken before any other aspects of the system, such as its boundaries, context, subsystems, feedback, inputs, and outputs, can be determined. Because of this, it is significant because, according to the point of view connected with general system theory, it is necessary to identify the system that is being discussed. In order to get a more in-depth understanding of the system, we must first begin by defining it. After that, we may proceed to quantifying each associated quality in order to achieve this goal. This would make it possible for us to have a better understanding of the system.

Because of this, in order for us to collect information on the topic of the research, we will initially need to measure it in order to quantify the characteristics that are associated with it. For this, we will need to perform certain measurements on the subject. After that, we will establish the indicators that will be applied for the purpose of determining the value of each measure, and we will do so by utilizing the results of

the stage that came before it. Within the context of this method, the Measurement and Evaluation (M&E) process can gain an advantage from making use of a conceptual framework that is built on top of an underlying ontology. The M&E framework makes it possible to describe the basic ideas, which prepares the way for a measurement process to be carried out in a manner that is consistent and repeatable. This is made possible by the fact that the framework makes it possible to specify the essential concepts.

The ability of a measuring process to be automated is of the utmost significance, even if it is required for a measuring process to give findings that are consistent, comparable, and traceable. The ability of a measuring process to be automated is of the utmost relevance. Because the activities that take place in today's economy take place in real time, we need to pay considerable attention to the use of online monitoring in order to notice and avoid a variety of different scenarios while they are happening. Because of this, we will be able to reduce risk while maximizing our efficiency. In this regard, the functionality of the measurement and evaluation frameworks is an extremely valuable asset, as they make it possible to organize and automate the process of measuring in a manner that is consistent. This makes the frameworks an exceptionally helpful asset. As a result of this, the frameworks are a very useful asset.

As soon as it is feasible to guarantee that the measurements are comparable, consistent, and traceable, the method of decision-making will naturally be based on their history, which will consist of the measurements collected throughout the years. This will be the case as soon as it is possible to guarantee that the measurements are comparable, consistent, and traceable. This will take place as soon as it is practical to assure that the measurements are comparable, consistent, and traceable. In this regard, the organizational memory is of special importance due to the fact that it makes it possible to store prior organizational experience and knowledge in order to get ready for future proposals (that is, as the foundation for a range of different assumptions and premises, among other things).

In this regard, the organizational memory is of particular use. Because of this, the organizational memory is a component that is of very high importance. Measurements and the experiences that are associated with them provide continuous nourishment for the organizational memory, and the organizational memory provides the foundation for the feedback that is utilized in the process of decision making. On the other hand, given that the Organizational Memory is only a model, it is feasible that there will be no

instructions (or previous instances) to follow in the event of a brand new circumstance (such a natural disaster). This is due to the fact that the Organizational Memory does not function as a true store of knowledge. The fact that the Organizational Memory is a model is the reason why things are the way they are. It is of the highest significance to have this in mind since, in the context of smart cities, there are scenarios linked with measurement and evaluation procedures on infrastructure, and one of those circumstances is the potential for gathering incomplete data.

Keeping this in mind is of the utmost importance. Keeping this in mind is absolutely necessary due to the fact that it is one of the potential outcomes. This is due to the fact that it is highly probable that there are no documents surviving from a period of time that came before this one. The last image depicts Santa Rosa, which is a city in Argentina's La Pampa province. Santa Rosa is located in Santa Rosa Province. When the city got the amount of water that it normally anticipates receiving over the course of an entire year in the span of only one week, it was at a loss for what actions to do next. Even if they had prior knowledge of the regularity and intensity of the rains, the city was nonetheless made worthless by them.

During this invited session, we are going to speak about the effect that the multiple pieces of data and information have along the road to decision-making, as well as how those pieces of data and information may be utilized in the future. In addition, we place an emphasis on the measurement and evaluation process as a significant asset connected with knowing the entities under investigation (such as a business process, a person, a system, and so on), their contexts, and the technique in which the process may be automated. This is because we believe that the measurement and evaluation process is one of the most important aspects of understanding the entities under examination. This is due to the fact that we are of the opinion that the process of measuring and evaluating is directly tied to the knowledge of the entities that are the subject of the research. We place a large amount of importance on the function of the organizational memory known as providing a knowledge basis for proposals, as this is the role that the Organizational Memory performs in the process.

The following outline is going to provide a description of the format of this article that you just read. An evaluation of the relevance of measurement and assessment as the major impulse behind making decisions based on data is offered in Section 2, which also serves as a summary of the subject matter. This section also serves as a summary of the subject matter. In the third section of this essay, the social implications of making

decisions based on data in the context of smart cities are investigated. In the subsequent section of this article, we are going to discuss a possible application scenario that makes use of the Autochthonous Institute of Housing, which is located in La Pampa, Argentina. Following the presentation of several works that are connected to this issue in the fifth part, the findings and recommendations for additional research are examined.

1.1 MEASUREMENT AND EVALUATION'S IMPORTANCE

An intriguing way to get the conversation going is to start it off by asking questions about the concepts and the applications of those conceptions. This is a great way to get the topic rolling. To put it another way, if we don't measure anything, what type of results can we possibly anticipate? What are the reasons for the significance of us measuring these things? What are the positive outcomes that result from doing such actions? When it comes to engineering, common sense tells us that we need to define an idea or an object in order to know the concrete and conceptual features of the thought or the thing that is in issue. This is true even though defining an idea or an object is not strictly necessary. Once we are familiar with all of the qualities, it will be helpful to quantify each one so that we may investigate the behavior in a number of scenarios. This can be done once we have completed the previous step.

As soon as we are comfortable with all of the qualities, this will be carried out as quickly as humanly feasible. Because of this, and on the basis of our evaluation of each occurrence, we are able to comprehend what makes a normal scenario and what defines an abnormal circumstance. This understanding is essential for detecting and avoiding effects that are not desired. This is due to the fact that humans have the ability to differentiate between what would be considered a typical condition and what would be considered an abnormal circumstance. As a consequence of this, the preventable occurrences and the increase of resource utilization provide us with an interesting social and economic point of view as a beneficial cause for the measurement. In each and every one of these scenarios, the issues of the precision of the data continue to be a significant focus of research.

It is now possible, as a result of the measurement, to carry out a quantitative analysis of the qualities of an entity that is the subject of the investigation. This entity could be a system, it might be a component, or it might be something totally else. However, in order to guarantee that the results are accurate, we will need to concentrate our attention

particularly on the principles underlying the measuring process. Only then can we be confident that the findings can be trusted. To put this another way, the usefulness of the measurement is contingent upon the measurements being consistent and comparable, as well as the process of measuring being repeatable. In addition, the utility of the measurement is dependent upon the technique of measuring being repeated. This is due to the fact that the usefulness of the measurement relies heavily on its consistency as well as its comparability.

As a consequence of this, it is absolutely necessary for each and every one of us to have the same meaning with regard to the ideas of measurements, metrics, and indicators, amongst other things. In this regard, the measurement and evaluation frameworks make particularly excellent sense because they enable us to come to an understanding on the concepts that we want to use throughout the process of measurement and to communicate in the same language, thereby reducing the likelihood that misunderstandings will take place. In other words, they make sense because they allow us to talk about the same things, which is made possible by the fact that they exist.

For instance, if we want to monitor an organization as a system, we could use the Balanced Scorecard perspective that was developed by Kaplan and Norton; the Goal-Question Metric method; the C-INCAMI framework (which is an acronym that stands for Context- Information Need, Concept Model, Attribute, Metric, and Indicator); or any one of a number of other frameworks. These are just some examples. These are only a few examples of the many various kinds of structures that may be utilized. There are many more. It is possible for every strategy to have both strengths and limitations, but the method in which these traits present themselves will alter depending on the circumstance in which we choose to use the approach. It is possible that any strategy might have both strengths and weaknesses. Even if we are free to choose from a variety of methods, it is essential that we maintain the same standards of measurement over the course of time. This is true even though we have the option to choose from a variety of methods. Because of this, our ideas and assessments will be consistent with one another and will be able to be compared to one another.

There are some characteristics that are tied to the data itself, and these are the only ones that matter in the context of the Data Quality standard that is based on ISO 27001. This illustrates that the characteristics are fully dependent on the data, as is the case with the correctness, completeness, consistency, credibility, and currentness of the data. Furthermore, this demonstrates that the qualities are entirely dependent on the data. On

the other hand, there are other qualities that are concurrently dependent on the system as well as the data. A few examples of these include the data's correctness and precision, as well as its accessibility, compliance, and secrecy. Other examples are its exactness and precision. Efficiency and accuracy are two further instances that illustrate the data's qualities. It is essential that this point be highlighted because the data are a component of the system, and the quality of the data is influenced not only by the data themselves but also by the system in which they are processed. It is vital that this point be emphasized because the quality of the data is impacted not only by the data themselves but also by the system in which they are processed. Because the data are an integral part of the system, it is critical that this aspect get adequate attention.

The process of basing judgments on the study of data rather than solely relying on one's intuition is referred to as "data-driven decision making," and it may also be defined as "the practice of basing decisions on the analysis of data." Data-driven decision making is a term that was coined by Harvard Business School professor Michael D. Norton. It is easy to comprehend how the quality of the data will have an immediate and direct impact on the process of decision-making if the process of decision-making is based on the data. If the process of decision-making is based on the data, it is simple to grasp how the quality of the data will have an influence. When seen from this perspective, it is simple to understand how the standard of the data will have an immediate and direct effect on the procedure.

Due to this fact, it is of the highest importance to conduct out monitoring at each and every stage of the data life cycle. That is to say, we need to build protocols inside the organization with the intention of monitoring the gathering of data, the processing of data, the analysis of data, the preservation of data, as well as the reusing, using, or deleting of data. These are the kinds of activities that require careful observation and oversight. In this regard, there are appealing ideas connected to a wide variety of viewpoints that are associated with the key methods for preserving the data quality. Such models include, for instance, the data maturity model produced by the CMMI Institute as well as the CALDEA model, which is based on maturity models. Other examples of such models include the models described in the previous sentence. Both of these examples serve as excellent illustrations of advice that should be taken into consideration.

In order to bring this discussion full circle and return to the beginning of the article, the question that has to be addressed is as follows: Why is it helpful to measure? It is

helpful to have a broad comprehension of the entity that will serve as the focus of the study. The process of measuring determines the process of collecting, and throughout the whole life cycle of the data, the measuring process has a direct relationship to the process of data collection. Because of this, if we concentrate our efforts on the phase of the life cycle during which the data is obtained, we have a significant opportunity to significantly enhance the performance of the other stages of the life cycle. This is due to the fact that we focus the majority of our attention on the data at the stage in which it is being obtained.

To frame this another way, if we were able to decrease the risk of mistakes appearing in the data while it was still being collected at the source, we would be able to limit the impact of errors spreading to later stages of the life cycle. This is because we would have lowered the chance of mistakes appearing in the data while it was still being obtained at the source. To put it another way, we would be in a position to forestall the occurrence of mistakes in the first place. In conclusion, as a result of this, it would be possible for us to reduce the risks associated with the quality of the data when we make decisions based on the data. Consistency issues and other concerns of a comparable type are included among these dangers.

In spite of this, the measurement merely refers to the procedure by which we collect the data; it says nothing about the method by which we evaluate the outcomes of the measures. For instance, in accordance with the parameters of C-INCAMI, the assessment necessitates the formalization of the organizational knowledge by way of the decision criteria that are embedded inside the indicators. In this approach, each indication is provided with a enough number of concepts for grasping each value of the related measure and arriving at a conclusion by making use of the organizational knowledge that is available to the individual conducting the analysis.

Consequently, the sequence in which we may establish the measurement and assessment technique has major bearing on the outcome. As a direct result of this, we are in a position to take advantage of techniques such as GOCAME (Goal-Oriented Context-Aware Measurement and Evaluation) and SiQinU (Strategy for Understanding and Improving Quality in Use). Both of these strategies are aimed at helping us understand and improve the quality of the services we provide.

When it comes to making judgments based on data, the organization makes use of its data in order to fuel a variety of different decision-making processes. In this particular

instance, the history that is connected to the data is a very valuable resource that has the potential to aid in the process of decision-making. As a direct result of this fact, the memories of the organization could be taken into consideration when modeling the experiences of the organization based on historical measurements and assessments. In addition, while the choice is being made, a case-based line of reasoning may be retrieved from the organizational memory and utilized in order to provide support for the proposal. This may occur at any point during the decision-making process.

1.2 THE IMPACT OF DATA-DRIVEN DECISION MAKING ON SOCIETY

A "smart city" is a city that integrates and monitors its critical infrastructure via the use of "smart computing" in order to provide required services to the general population. This is the characteristic that distinguishes a city as a "smart city." On the other hand, the concept of smart cities is not just related with a technology component; rather, it is connected with a number of other elements as well, including the economy and the government, amongst other things.

Every choice that is made along the numerous various services or infrastructure in a city needs to be established on the previous experiences of the city's people. Leaning on our past experiences as well as the information we've gathered here in the city, we are able to recognize various deviations from the norm that fall within the typical range. As a direct result of this, we are in a position to identify circumstances that do not correspond to the criteria that we have established as acceptable. The monitoring of a city's infrastructure and services is vital to attaining the ultimate goal, which is to either prevent potentially harmful conditions from arising or, if the worst should happen, to notice them while they are happening in real time. Either way, the goal is to achieve this goal as quickly as possible.

When it comes to acquiring a better understanding of the present condition of any service or component of the city's infrastructure that may be in issue, the process of measuring and assessing is extremely essential. To put this another way, if we don't measure, we won't be able to determine the state of each individual component, which is why it's so important that we do. This is of the ultimate necessity for correctly orienting the process of making decisions based on data, as each decision has to be based on the current state of the numerous components that are positioned along the city. This is because each decision needs to be based on the current status of the various elements that are positioned along the city.

What are the possible outcomes in the event that we are unable to access the data? Since we do not have any proof in the form of data or records to back up the several other alternatives, it is quite probable that the decisions are led by intuition. This is because we do not have any evidence to support the numerous other possibilities. The difficulty with relying on one's intuition when making judgements is that it is an extremely subjective process, and we do not have any records or experiences from the past that we may use in order to explain the chosen course of action. The people in charge of making decisions face a difficult task as a result. There is a possibility that persons in positions of authority may do things that will have a detrimental impact on society if they base their decisions on their feelings rather than on what they know to be true rationally. For instance, the city of Santa Rosa may be found in Argentina's La Pampa region. Santa Rosa is located in La Pampa.

The country of Argentina is home to La Pampa. When the city needs to determine how much water is falling during a certain storm, it can use a pluviometer as a point of reference to determine how much precipitation is falling at any given moment. This makes it possible to collect data, but it does not make it possible to monitor in real time the problem associated with the water level decreasing along the city or the quantity of water moving through the sewers. Nevertheless, it enables the collection of data to be carried out.

The city got an amount of water from nature in the month of March in 2017 that was comparable to one year's worth of precipitation, yet it only fell on one week of that month. This occurred despite the fact that the precipitation fell over an extended period of time. According to the age-old proverb that "a picture is worth more than a thousand words," Figure 1 demonstrates that the use of intuition in the decision-making process has obvious implications for the city of Santa Rosa. The north zone and the capital of the province of La Pampa continue to bear the brunt of the repercussions that the flooding has had to this day and age.

This is the case despite the fact that the water has subsided somewhat. This is as a result of the fact that these regions were among the very first to be impacted by the problem. Even while it may not always be possible to stop natural catastrophes from occurring, it is still possible to keep an eye on the city's infrastructure and services in order to plan any necessary repairs in plenty of time in advance. This is something that can be done. This will guarantee that the city is ready for anything may come its way by ensuring that it is prepared.

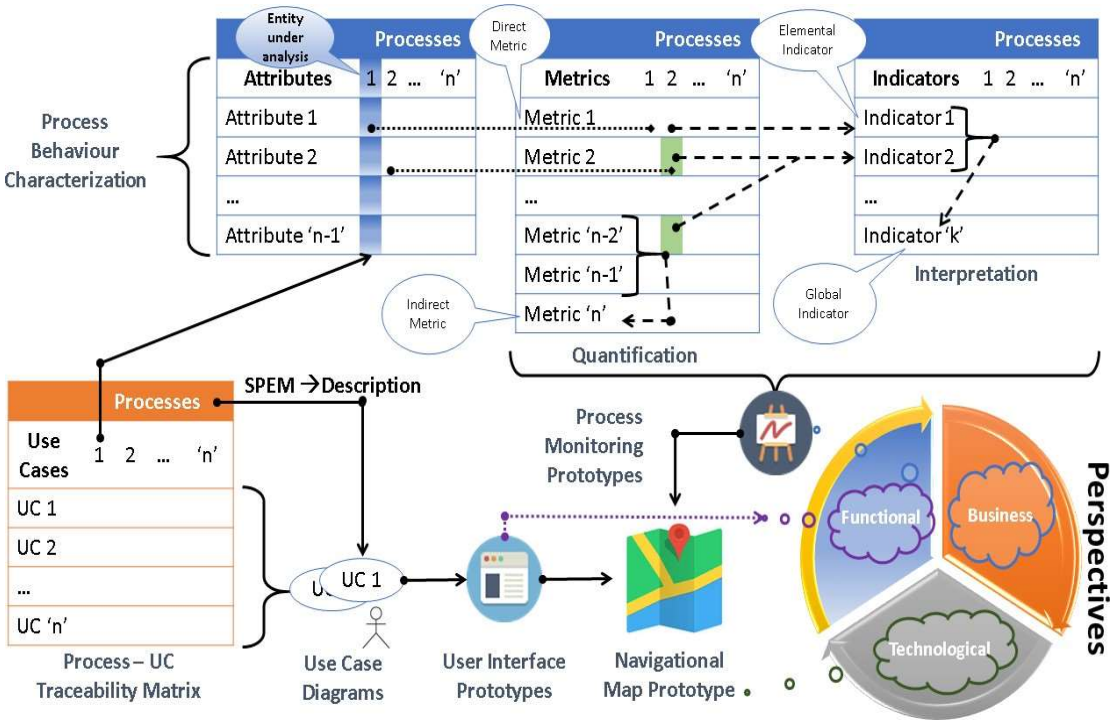


Fig. 1.1 The concept of "Conceptual Perspective" in relation to "Business Process Monitoring" A Method of Decision-Making That Is Based on Data

source: *guide data driven, data collection and processing through by Arman Qureshi (2019)*

The monitoring of the population's infrastructures and services, on the other hand, might at the very least improve the quality of life of the population, foresee disasters, or even save their lives.

1.3 THE AUTARCHIC INSTITUTE OF HOUSING OF LA PAMPAS IS AN APPLICATION CASE

The Autonomic Institute of Housing of La Pampa, which is also known as AIHLP and often abbreviated as such, is the public entity that is in charge of the development of housing over the entirety of the region that is regulated by the government of the province. This institution is also frequently abbreviated as AIHLP. AIHLP is the name that most people use for it, and you could also see it shortened as such. Even in the event that the organism receives financial assistance from the local, state, or national

governments, it is still exclusively responsible for its own financial management and is unaffected by the acts of any third parties. This is the case even in the event that the organism receives support from all three levels of government. The family that will eventually own each property are ones that do not have the financial resources to get access to the official system that banks employ to carry out their operations. This means that these families will be the ones who own each property after it is all said and done.

As a direct consequence of this, every house will be sold to a family that is not in a financial position to purchase it on the open market. These families will, in the end, be the ones who wind up being the owners of each property as a direct result of this situation. In this sense, there is a constant demand associated with it, and there is always a gap between the supply and the need for it. In other words, the supply is never sufficient to meet the demand. To put it another way, there is never enough supply to satisfy the level of demand.

Because there is such a large demand for the dwellings, the AIHLP has designed a one-of-a-kind system for assigning them that takes into account the particular circumstances of each family. This action is taken in order to satisfy the extremely high degree of consumer demand. This step is made in order to accommodate the very high level of customer demand that has been expressed. The purpose of this tactic on the part of the authorities was to arrive at judgments based on the basis of prior experience in order to gather information from previous scenarios, make the most out of the resources that were available, and prevent the repeating of errors repeatedly.

we want to get the ball rolling on a reengineering of our business processes, the major goal of which will be to enhance the visibility of the activities that are related with each process. We will be using SPEM, which is an abbreviation that refers for Software Process Engineering Meta-Model. SPEM will serve as the modeling language that we will be using. We use a process that is referred to as EIb PREME (Integrated approach that is based on Processes, Requirements, Measurements, and Evaluations), which stands for "Integrated approach that is based on Processes, Requirements, Measurements, and Evaluations." The focus of this strategy is on the processes themselves, as opposed to the requirements or assessments. The objective of this method is to monitor the activities of the organization that is the focus of the inquiry. In order for this method to be successful, the C-INCAMI framework, which serves as a measuring and evaluation framework, will need to be utilized.

After determining that each step could be effectively tested with our stakeholders in a manner that was both satisfactory and effective, we immediately began the process of designing the measurement and evaluation project employing the GOCAME technique. This was done as soon as we had concluded that each stage could be tested effectively. When seen from this perspective, there is no longer a mystery that the entity that was the major focus of our inquiry was each modeled process. After that, we detailed the method in which the authorities want each business process to be categorised, as well as the point of view that was associated with it via our cooperation with them. In addition, we included the point of view that was linked with it through our collaboration with them. This was done due to the fact that we were collaborating with them on this project. Figure 2 illustrates how the idea may be comprehended by breaking it down into its component parts. The initial phase is the defining of the process's characteristics, and the steps that follow are dedicated to the implementation of the three monitoring perspectives: functional, technical, and business.

As can be seen in figure 2, each characteristic or quality that may be used to characterize a process was assigned a metric that might be used to assess that particular characteristic or quality. This was done in order to ensure that the metrics would be useful. This was done in order to guarantee that the attributes and traits would be equivalent to one another. Because of this, it was feasible to conduct an analysis of the process while taking into consideration the many various aspects of it. After that, an indication for interpreting each connected measure in each operation was formed, and this was done while keeping close connection with the authorities that were vital to the case. Ultimately, the case was successful. At this point, it was very necessary for the authorities to have a solid grasp of business in order to effectively incorporate the choice criteria into the design of the indicator. This may be accomplished by having a great understanding of business.

We were able to successfully construct a prototype of the visual scorecard so that we could include it in the process of incorporating metrics and indicators into the monitoring of the operations. It was the final phase from a commercial point of view, and it was this step that brought about the outcomes that were necessary. The step that brought about these outcomes was the creation of a web-enabled and multi-device command board. It is possible to derive the model of use cases in parallel from each individual phase of the process. This includes both the activities and the tasks. It is possible to accomplish this goal. After getting consent from the end user, the user

interface prototypes that were designed based on the use cases were put into production. These prototypes were conceptualized with the use cases in mind as their primary source of motivation.

There is a natural link between the user interface and the command board due to the fact that the processes comprise the logical viewpoint connected with the data collection, processing, and reporting. This is due to the fact that the processes entail the collection of data, processing of such data, and reporting of said data. Because the user interface and the command board are both components of the same system, there is a link between the two of them. This link is responsible for the connection that exists between the two.

In addition to the perspectives of economics and logic, the authorities have contemplated the possibility of making the data and information interoperable across the myriad of entities that constitute the provincial government. This has been done in conjunction with the analysis of economics and logic. This was done in consideration of the fact that a variety of creatures come together to form the governance of the province. They have labeled this novel way of looking at things as the "technological perspective." In order to accomplish the task of effectively completing the data interchange, the concept of "software as a service" was utilized in this fashion. As a result of this, it became possible for any authorized organism to have access to the data in a manner that was both direct and automated, and it did not require the participation of any intermediaries in order to do this.

In the end, a visual coordination of the three viewpoints (technical, business, and functional) was achieved by employing navigational charts and working closely with the necessary authorities. This was successful in accomplishing the goal. This led to the achievement of a fruitful conclusion. An increased trend in the number of families that registered can be seen taking place over the course of time between November of 2014 and October of 2015, as seen in figure 3, which presents these statistics. This trend can be observed taking place over the length of time. As can be seen, the rate of increase is rather rapid, and there is still a considerable way to travel before it reaches a level that can be considered stable. Nowadays, the "simple" data that is presented on the command board offers them the opportunity to design a plan for the works that need to be done and a projection of the demand in the years to come. Both of these capabilities were previously impossible. It is just one of a few measures that might potentially perform this purpose.

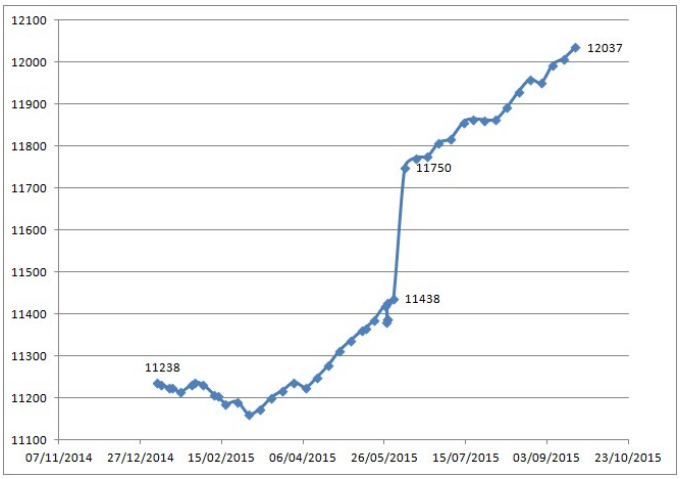


Fig. 1.1 Aio HLP - Evolution of the volume of Registered Families

source: *guide data driven, data collection and processing through by Arman Qureshi (2019)*

As a direct result of the benefits that were achieved, the AIHLP is now in a position to make decisions that are backed by data, monitor their company operations in real time, and be interoperable with any authorized entity that needs the information. All of these capabilities were made possible as a result of the benefits that were acquired.

During the course of this invited session, we answered the issues that were asked to us by the audience and offered our viewpoint on the part that measurement and evaluation play in the process of data-driven decision making. In terms of the quality of the data, and in accordance with ISO 250012, we may have aspects that depend solely on the data; however, we may also have aspects that depend solely on the system; alternatively, we may have both types of aspects. In any case, we may have both types of aspects. When seen from this vantage point, the decision-making process that is driven by data is dependent not just on governance and other connected challenges, but also on the quality of the data itself. In the case that the data are of low quality and the technique of measuring is flawed, this may be an indication that the process of making judgments is also incorrect due to the fact that both of these aspects are related. As a consequence of this, the Data Management Maturity Model that was developed by the CMMI Institute is an attractive option, at the very least for the purpose of doing research on it. This is because the CMMI Institute was the one responsible for developing it in the first place.

During the course of this conversation, we presented a real-life situation that took place in the city of Santa Rosa in La Pampa, Argentina, to demonstrate the catastrophic results that may arise from making decisions based only on one's gut sense. In this scenario, a man was shot and killed after he refused to listen to the advice of a doctor. The court case was based on a true occurrence that took place in Argentina at the time it was being portrayed in the drama. There is a strong probability that on a daily basis, we read articles that describe the benefits of basing decisions on data. In these articles, the advantages are discussed. It is intriguing from the point of view of measuring up and making evident the effects, both positive and bad, that are associated with the presence and absence of data-driven decision making. To put it another way, it is fascinating because it poses the question of whether or not decision-making should be driven by facts.

In this specific instance of an application, which is connected to the Autonomic Institute of Housing of La Pampa, we exhibit a circumstance in which the business processes are believed to be the entity that is being researched. This particular illustration was created to demonstrate a situation in which the business processes were thought to be the entity. In this setting, we are going to discuss a strategy to measuring and assessing based on C-INCAMI that was developed in order to make the process of decision-making more streamlined and straightforward.

We will continue to make progress on the implementation of different research scenarios linked with data-driven decision making, as well as constraints and ramifications, as part of our ongoing endeavor. This will be carried out as a component of the continuing effort that we are making.

1.4 DATA-DRIVEN, MACHINE LEARNING-POWERED DECISION-MAKING

Up until this point, decisions in the world of business have been made on the basis of the facts and numbers, with trends that have been identified in the data from the past being projected into the future. There will be a rising number of possibilities to make use of the power that is made accessible by modern information technology as the current era of information technology continues to advance. These opportunities will reveal themselves as the current era of information technology continues to progress. The most modern technology, which is based on machine learning and artificial intelligence, goes beyond just aiding in the process of decision-making and has reached

a stage where the majority of decision-making is both autonomous and carried out by itself. This is a significant advancement from previous generations of technology, which were limited to only assisting in the process. This is made feasible by the fact that the technology was built from the ground up to achieve this purpose.

1.4.1 Business Value of Machine Learning

Because it makes it possible to receive and evaluate information in a much more timely manner, information technology in general plays an essential part in the productivity and performance of enterprises. This helps to further support the argument that information technology is a significant component. Investing in digital transformation has the potential to boost a company's value in each and every area. This is true for companies of all different sizes. On the other hand, there are some of them that can be easily identified, such as the amount of time and money that is saved, there are some of the traits that cannot be accurately articulated, such as a person's speed and agility.

These are not readily quantifiable, but they are considered to be crucial due to the fact that bad decision-making is a factor that cripples the firm and enables competitors to capture market share. This is the reason why it is essential to pay attention to these aspects. However, increasing the speed at which choices are made should not come at the price of decreasing the quality of those decisions. The successful execution of a flawed idea does not result in the development of new value and, in the long run, will lead to failure. Using cutting-edge technology like machine learning is one way to ensure that the process you employ to create judgments is both effective in its use of time and well-structured in its approach. Another way is to utilize a decision-making tool that is designed specifically for this purpose.

Before they can perform their tasks effectively, machine learning systems need not only a substantial amount of data but also insights that are important to the tasks they are intended to perform. They utilize these specific pieces of data as building blocks for the models that they are developing. We are able to differentiate between four separate types of analytics, each of which, depending on the amount of data and the type of data that you have, may offer a different set of insights than the others. Depending on the amount of data and the type of data that you have, we are able to differentiate between four distinct types of analytics.

1. **Descriptors and categories used in analytical research:** The most fundamental kind of analytics, which relies entirely on information from the past in order to

provide an account of what occurred and when it took place. In order to complete this analytics assignment, the massive amount of data will be divided up into smaller, more manageable bits. This is accomplished by adding descriptive statistics to the data that is already available in order to render the material simpler to comprehend and more suitable for presentation to the relevant stakeholders.

2. **Methods pertaining to analysis and diagnosis:** In order to give an explanation for why something took happened, a more in-depth examination of the past and its events is performed. It explains not just why particular factors had an impact on the final result, but also how those factors really brought about that impact. The forms of analytics that make the most use of classification and regression training algorithms are the ones that have been described below. In spite of the fact that it does not provide any insights that can be put into practice, it does provide a greater understanding of the causal relationships.

3. **Analytical forecasting and prediction:** By determining the probability that particular occurrences will take place in the future using this method of analysis, one might arrive at predictions about the future. The combination of statistics and algorithms that were developed specifically for machine learning is what makes predictive analytics so effective at producing accurate projections. There are times when it can even help with more complex estimates in the areas of marketing and sales.

4. **The application of predictive analytics in practice:** The most advanced degree of analytical expertise is represented by this sort of analysis since it makes a recommendation for a course of action on the basis of predictive analysis. In addition to this, it is possible for it to work in both directions, proposing particular activities that should be undertaken in order to attain a favorable outcome and predicting an outcome based on the actions that are actually taken in the situation. It's likely that the optimization of recommendation engines is the usage of this kind that's the most well-known.

1.4.2 Business Process Automation

Businesses that place a significant emphasis on data are often very active in the processes of collecting information, storing information, and administering information. These firms also make great use of the data they collect in order to improve

the procedures that they utilize to manage their organizations. If a company increases the quantity of data that it gathers, there is a possibility that it will be able to improve the quality of the analytics that it offers to its customers. In addition, there is a potential that the company will be able to improve its capacity to make judgements internally if it automates some of the processes that it now employs. This possibility exists because there is a possibility that the company will be able to enhance its capacity to make judgments.

Take, for example, the efforts that are being made in the Innovation Department of your business to improve the organization's use of automation. When it comes to automating processes, the best place to begin is by making use of software that has been specifically built for the administration of innovation and that stores data relevant to the activities that you carry out in the field of innovation. Because it makes the most effective use of resources, this is the best place to start. In order to make effective use of this data and study the common patterns that emerge during the process of decision-making, you will want a machine learning framework. For instance, the automatic approval process that is made accessible by Innovation Cloud Enterprise helps to prevent the creation of bottlenecks in the workflow, which enables the work to continue without interruption.

When there are too many people involved in the decision-making process, the method for approving anything might often take longer than it should. On the other hand, in certain circumstances, individuals accountable are merely overburdened with an overwhelming amount of work to complete. The occurrence of difficulties is quite possible in either of these two scenarios. The process of making judgments is slowed down in one way or another, and the most effective technique for preventing this from occurring is to automate decisions that are well-structured. When the innovation project has to be pushed further into development, machine learning algorithms may educate decision-makers about what their regular reactions in the past have been for a specific circumstance. This may educate decision-makers about what their normal responses have been in the past. This helps to increase the likelihood that the project will be successful overall.

CHAPTER 2

MACHINE LEARNING FOR PUBLIC POLICY MAKING

2.1 MACHINE LEARNING

When one is in charge of making decisions on public policy, how does one approach the challenging task of offering accurate predictions? In this context, one approach that may be pursued would consist of, in addition to employing one's intuition, making use of the human capacity for experience. For instance, the local police department may make the assumption that there is a higher possibility of criminal activity in certain portions of the city due to prior experience or other variables. This assumption might be based on a number of different causes. If they decide to station police personnel in a number of key locations, there is a good chance that this strategy will prove to be highly effective for them.

After all, this is something that a variety of different departments within a variety of different organizations have kept up as a consistent practice over the course of the history of their individual companies. On the other hand, there is always the possibility that it will not function in the most efficient manner that is conceivable given the circumstances. Would it not be amazing if we were able to improve the police department's capacity to predict crimes by utilizing the data that is now available on crime rates in addition to other information that is already accessible? If we could do this, would that not be an incredible achievement?

In other words, it would be incredible if we were able to achieve it, don't you think so? The idea of machine learning starts to become more prominent in the gameplay at this time. Machine learning gives us the ability to get insight into data that we do not presently have access to by utilizing the knowledge and information that we already have at our disposal. We are able to achieve this goal by making good use of the information and expertise that is currently in our possession. Forecasts are produced by basing their assumptions on the information that has been gathered throughout time. An algorithm's ability to learn from the data it is given increases in proportion to the amount of information that is delivered to it. The strategy in question has been given the moniker "machine learning" as a direct result of this particular line of reasoning. Machine learning gives computers the ability to learn from their own experiences,

much to how individuals learn from the events and circumstances that occur in their own lives. In addition, computers now have the ability to learn from their own data thanks to machine learning. This research was carried out with the intention of providing a high-level overview of what machine learning is and what processes are involved in the process of training a machine learning model. More specifically, the emphasis of this research will be placed on the connection between data and the judgments made by algorithms. This study's objective is to give an explanation of what machine learning comprises in its most general form as a means of fulfilling its mission.

When humans attempt to model data, they typically turn to more traditional approaches. Machine learning, on the other hand, is much simpler to understand in comparison to these other methods. Linear regression models are shorthand for more complex statistical models that have the only purpose of describing linear correlations between independent variables and dependent variables. The models in question are examples of what are known as "simple statistical models." It is necessary for a large portion of the population that is responsible for making decisions on public policy to have a comprehensive understanding of these frameworks.

On the one hand, there are a lot of major differences that can be found between the models that are used in machine learning and linear regressions and other traditional data models. On the other hand, there are a lot of parallels that can be found between linear regressions and other conventional data models and the models that are used in machine learning. Because of this, it is of the highest necessity to carry out a more in-depth comparison and contrast of the similarities and differences that exist between these two cultures' approaches to data modeling.

Both human data modeling and machine learning, in their own distinct ways, make an effort to model relationships within the data that they are working with. However, the two approaches represent these relationships differently. Both of these methods make use of statistical models in order to determine the nature of the link that does, in fact, exist between the independent factors in the data and the dependent variables that are the focus of the current inquiry. Statistical models are used to determine the nature of the relationship between the independent factors in the data and the dependent variables. The key difference that can be established between human data modeling and machine learning is the approach that is used to pick a model that represents a link. This is the fundamental differentiation that can be established. This is due to the fact that people are responsible for modeling human data.

The most important difference that can be made between the two is based on this particular aspect. In the process of modeling human data, the human modeler will choose a stochastic model, such as a linear regression model, and will then apply this model to the data in order to fit the data to the model. This process is known as "fitting the data to the model." Because of this, the difficulty of the problem of fitting an essentially arbitrary function to the data may be reduced to the effort of finding a restricted set of parameters that optimizes the fit of a given functional form to the data. This would be an improvement over the previous situation, in which the complexity of the issue was not decreased. This would be an improvement over the prior condition, in which the complexity of the problem was equal to the difficulty of the work itself.

This would be an improvement over the previous situation. The method of doing an analysis based on linear regression makes use of these two components, which may also be referred to as the axis intercept and the slope, respectively. The slope and the axis intercept are both names that may be used to refer to these characteristics. It is not unduly difficult to identify the values of the model's parameters using methods such as least squares, which will generate the best possible fit between the model and the data. This will be accomplished by determining the values that will produce the best possible fit between the model and the data. This is something that may be performed in a very uncomplicated fashion. This is something that can be done in a way that does not need an excessive amount of difficult steps.

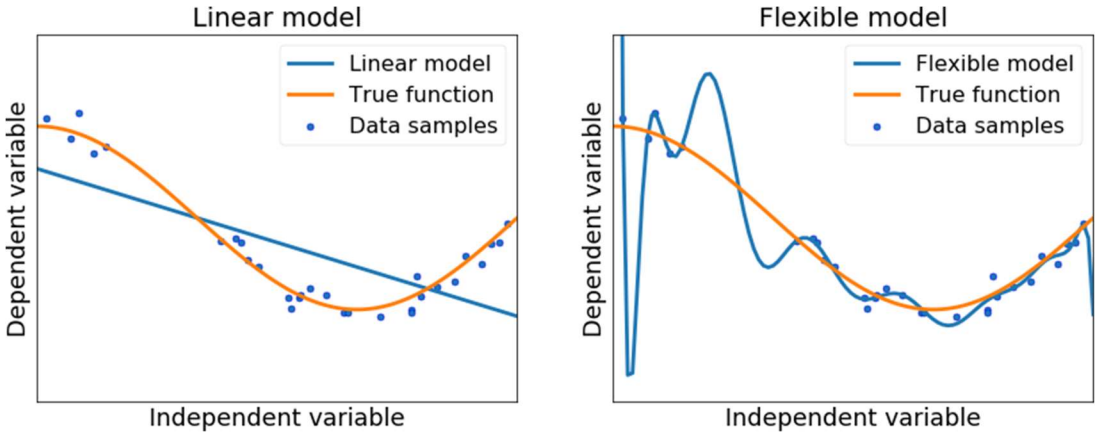


Figure 2.1: A linear model and a more flexible model fit to the same dataset

source: *machine learning for public policy making, data collection and processing through by Umesh Verma (2018)*

The linear model does not offer an appropriate representation of the data because it is not flexible enough to handle the curved connection that arises between the independent and the dependent variable. This curvilinear relationship develops because the independent variable affects the dependent variable. The model that has a larger degree of adaptability, on the other hand, is unnecessarily flexible, which leads to an overfit to the data.

If you focus an inordinate amount of your attention on the sampled data points, you will end up with a highly undulating curve that does not correspond very well to the function that is actually responsible for generating the data. This is something that will occur if you devote an abnormally high level of attention to the data points that have been sampled. Because of this, the ideal model would have more adaptability than the linear model shown on the left, but it would have less adaptability than the model shown on the right, which has an excessive degree of flexibility. This is because the linear model on the left has a fixed set of parameters that must be adhered to.

When it comes to machine learning, a person is necessary to select the algorithm that will be used to depict the relationship between the variables that are under their control and those that are not under their control. This duty is the obligation of the individual responsible for carrying out the machine learning. On the other hand, algorithms that make use of machine learning have the potential to acquire highly flexible functional forms from data without the need for people to explicitly explain such forms. This is a possibility due to the fact that these algorithms are able to learn. In contrast to this, human data modeling requires the participation of humans throughout the process of form creation. These forms cannot be designed without human input. Because of this, machine learning has a significant advantage over other methods: the algorithms that underpin machine learning can automatically adjust sophisticated models to fit data. This allows machine learning to be more accurate. Because of this, machine learning is far more effective than traditional approaches.

The selection of a machine learning approach places significantly less of a limitation on the variety of possible connections that may be taught in comparison to the selection of a human-chosen model, such as a linear model. This is because the selection of a human-chosen model places more restrictions on the potential connections that can be learned. There is a cost associated with the linear model's simplicity, as well as the simplicity of any other data model that people are able to construct using straightforward mathematics. If the relationship between the independent variables and

the dependent variables is not linear (or is produced according to whatever relationship the human modeler has selected), then the model will not fit the data very well and will not be very useful. If the relationship between the independent variables and the dependent variables is linear, then the model will fit the data extremely well. This compromise emerges as a result of the ease with which a linear model may be established. As a result, this compromise is costly.

Naturally, the incorporation of nonlinear transformations of the independent variables into an equation that explains linear regression is something that is achievable. This is a possibility that ought to be considered. However, this does not change the fact that the connections that may be learnt are still restricted to the small range of functional forms that a human modeler chooses to concentrate on. This is the case despite the fact that this new information has been presented. Machine learning, on the other hand, makes it possible for a computer to automatically fit models that are more flexible in order to represent the link between variables that are independent of each other and variables that are dependent on them. This is achieved by making it possible for a computer to automatically fit models that are more complex. Since machine learning models contain a far greater number of parameters than other, more fundamental models, such as linear regressions, the process of fitting models to data has become significantly more difficult as a result of this. Because of this, fitting models becomes a substantially more difficult task.

However, if machine learning algorithms are provided with sufficient training data, they are able to match more complex linkages between independent and dependent variables than any data model that could ever be developed by a person. This is because machine learning algorithms are able to learn from examples. This is due to the fact that the algorithms used in machine learning are able to match more intricate relationships between independent variables and dependent variables than traditional learning algorithms are.

Due of the similarities between human data modeling and machine learning, it is essential to address the reality that there is some overlap between the two subjects. This is due of the nature of the two disciplines. Even using the approach of least squares to fit a straightforward linear model is an example of machine learning in its purest form. Although a person makes the decision to use the linear model, the computer employs the least squares method in order to fit the model to the data. The decision to use the linear model was made by the human. In spite of the fact that the individual chooses to

operate with a linear model, this is the result. On the other hand, machine learning provides users with access to a far larger selection of algorithms, each of which is able to model increasingly intricate aspects of data connections. Unfortunately, in order for you to make use of them, you will be required to pay a charge of some kind.

It's conceivable that the models employed in machine learning be very flexible, to the point where they fit particular data too flawlessly. This would be an issue if the goal of machine learning is to automate tasks. In the same way that it is conceivable for this to happen when attempting to represent a complicated link using a linear model, it is also possible that this will not be appropriate. Figure 2.1 illustrates this issue in a way that is not only clear but also condensed and straightforward.1. Underfitting is when a model is not flexible enough to account for the true link that exists between the variables that are independent and those that are dependent. This occurs when the model does not have enough flexibility. This is something that can occur if the model does not have sufficient flexibility.

On the other hand, the phenomenon that is known as "overfitting" occurs when a model is so adaptable that it can account for each and every data point in the dataset to which it was applied and sampled. This happens when a model is so adaptable that it can account for each and every data point in a dataset. This is a possibility whenever a model is so malleable that it can give an explanation for each and every data point contained in a dataset. The basic objective of machine learning is to arrive at a point of equilibrium that is achieved by striking a balance between the two. A model should neither be so constrained that it fails to accurately depict the real link that does exist between independent and dependent variables, nor should it be so flexible that it produces results that are too good for the data that it was designed to analyze. Instead, a model should strike a balance between these two extremes. Instead, a model ought to achieve a balance between these two extremes; that is, it ought to be just the appropriate amount of restricted while also being just the correct amount flexible.

How can we make predictions using machine learning that are appropriate to the world as it actually exists? An instance of the process of developing and employing a machine learning model may be seen in Figure 2.2. This graphic depicts high-level process phases as high-level workflow activities. Some examples of high-level workflow phases are data preparation, model training, and prediction making. So that we can have a better understanding of how each of these processes operates, let's take a look at the components that go into each of these processes. To provide public policy makers with

a foundational grasp of machine learning, the content that will be presented in the pages that follow has been prepared with this objective in mind. The collection of published work that relates to the method of machine learning offers access to a large amount of content that may be accessed. provide an excellent introduction, while Hastie gives a thorough discussion of the many different methods that are involved in machine learning.

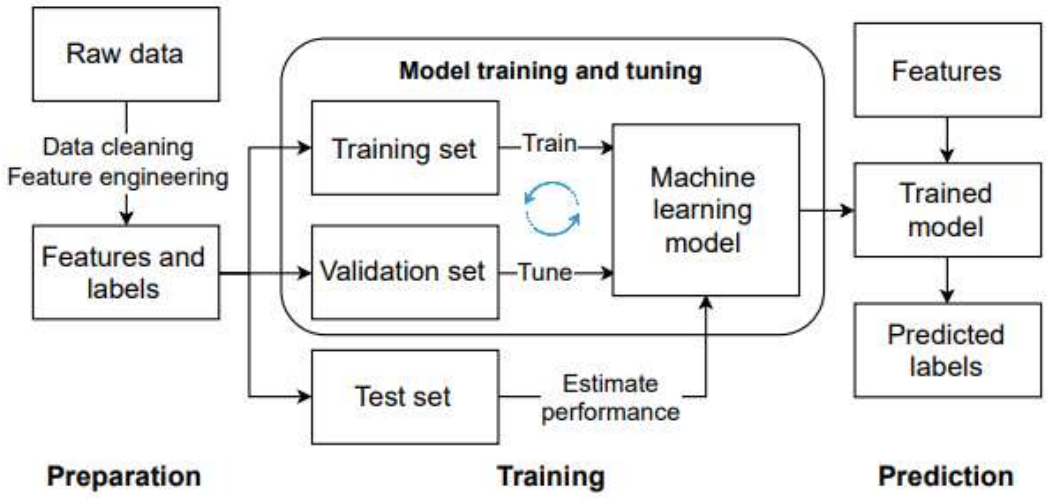


Figure 2.2: Overview of the machine learning workflow from data preparation and model training to prediction

source: *machine learning for public policy making, data collection and processing through by Umesh Verma (2018)*

2.1.1 Data preparation

The gathering of raw data that we have recognized as having the potential to be of some value in improving our ability to anticipate the outcome of the experiment that we are interested in is the first stage in the process of machine learning. We have determined that these data have the potential to be of some benefit to us. Rarely does machine learning get data in its raw form in a format that is immediately useful to the process. In the vast majority of cases, it is necessary to carry out operations such as cleaning the data, addressing any issues that may have been discovered regarding the data's quality, and combining information obtained from a range of sources into a single dataset.

When this stage is finished, a dataset will often take the shape of a table, with the variables organized into columns and the observations organized into rows. This process is called "descriptive statistics."

The following stage of the process is known as "feature engineering." This is due to the fact that variables are also referred to as features in the field of machine learning. This stage of the workflow is referred to as the Feature Engineering phase. This is due to the fact that the phase that came before this one was referred to as variable engineering due to the purpose for which it was carried out. In data analysis, the process of extracting higher-level features from lower-level features, which are also referred to as raw variables, is referred to as feature engineering. These characteristics are sometimes referred to as lower-level characteristics or characteristics. This is done in order to provide an algorithm for machine learning with information that is more appropriate to the topic that is currently being tackled. In order to continue the discussion on the prediction of bail, let us assume that we are interested in determining whether or not a defendant is likely to appear in court after being granted bail. This would imply that we have an interest in determining the possibility that the defendant would return to the scene of the crime.

We collect information on previous bail judgments as well as the actions that criminals took after being released from detention so that we may better understand their behavior. When we have finished organizing and cleaning the data into a single dataset, one of the characteristics that will be included in that dataset will be the allegation for which the defendant was arrested. We will add this feature once we have finished organizing and cleaning the data. Once we have finished sorting and cleansing the data, we will proceed with the addition of this functionality. According to the information provided in Table 2.1, a higher-level feature would be an extra characteristic that determines whether or not a defendant is charged with a violent offense. This determination is based on the arrest charge. This would be considered a more advanced feature. This would be considered a more significant part of the crime, as stated in the arrest charge that was brought against the suspect.

If we feel that the category of violent crime is beneficial for predicting whether or not a defendant will come before the court if released on bond, then this kind of feature engineering might contribute important information to our dataset. This information could be used to make decisions about how to proceed with our investigation. To put it another way, if we feel that the category of violent crime is helpful in forecasting

whether or not a defendant would show up in court after being released on bail, then we would say that the category of violent crime has an advantage. In addition to the features, our dataset has to have information on the result that we want to forecast in order for a machine learning algorithm to comprehend the patterns that lead to a defendant failing to appear before the court. This information can be found in the "outcome that we want to forecast" column.

When we talk about machine learning, the variable in which we are most interested, known as the target variable, is referred to by that term. This variable is a label for something else. At the conclusion of the process of producing the data, there ought to be a selection of essential characteristics and a one-of-a-kind label affixed to each observation in the dataset. This should take place. This step ought to be performed when the process of preparing the data has been finished in its entirety. This is the way things are supposed to operate.

Table 2.1: Made-up examples of features that could be used for predicting whether a defendant would appear before court if released on bail

	Features							**Label**
ID	**Age**	**Gender**	**Race**	**Arrest county**	**Arrest charge**	**Violent crime**	**Prior arrests**	**FTA**
1	28	Male	White	Bronx	Murder	Yes	Drugs	Yes
2	35	Female	Hispanic	Queens	Robbery	Yes	-	No
3	21	Male	Black	Brooklyn	Fraud	No	Guns	No
…	…	…	…	…	…	…	…	…

FTA is the label to be predicted and stands for “failure to appear”. Some of the features such race might actually not be included in a predictive model because it has become politically unacceptable to use them.

2.2 MODEL TRAINING

After the data has been cleaned up, relevant features have been assigned to each observation in the dataset, and labels have been assigned to each observation, we will be ready to begin the process of actually training the machine learning model. It is possible to extract a training set, a validation set, and a test set from the initial dataset once it has been divided into three smaller datasets that do not overlap with one another. These datasets are called the validation set, the test set, and the training set,

respectively. These sets are created using the primary dataset as a starting point. During the process of machine learning, a model is "trained" on the training set, "validated" on the validation set, and "tested" on the test set. This is done in the order that is suggested by the names of the various sets. Specifically, the titles.

According to the concept of a firewall ensures that none of the data that was used to train the model is subsequently used to assess it. This is something that may be accomplished by preventing access to the data. An algorithm for machine learning is responsible for carrying out the initial phase of the process of adjusting the parameters of a statistical model to match the data. The first step of this algorithm involves extracting the features and the labels that are part of the training set. Learning through machine analysis makes an effort to construct a link between the features and the label, which may be a relationship that is substantially more complex than a simple linear one. This may be the case since the analysis is performed in a sequential fashion.

Machine learning, which seeks to discover the link between the features and the label in the same way that linear regression finds the ideal linear relationship between the independent factors and the dependent variable, finds the ideal linear relationship between the independent variables and the dependent variable. In other words, machine learning identifies the ideal linear relationship between the independent variables and the dependent variable. It's possible, for instance, that those who are accused of committing serious crimes have a decreased possibility of showing up in court.

Or juvenile delinquents with a history of imprisonment for acts involving weapons who were imprisoned in Brooklyn and who have a record of having served time there. The degree of complexity that may be associated with the connection between the characteristics and the label does not have a theoretical upper restriction since there is no upper bound at all. This means that there is no limit to the amount of complexity that may be associated with this relationship. When machine learning algorithms are provided with a significant amount of data, they are able to automatically detect these patterns without the requirement for a human to define a restricted number of functional forms that are to be fitted to the data. This eliminates the need for a person to be involved in the process. The fact that these algorithms are able to educate themselves from their own blunders is what makes this achievement attainable.

However, there are a number of obstacles that need to be conquered before machine learning can be considered successful. If the model has sufficient flexibility, it will be

feasible to fit any dataset in any acceptable fashion that can be arbitrarily chosen. This is dependent on the model having an appropriate amount of adaptability. The graph that can be seen on the right-hand side of Figure 2.1 demonstrates that the level of flexibility that a model possesses is directly proportional to the degree to which it is able to accurately represent the data that it is given. The figure gives an illustration of this capability. Sadly, this is not at all what we have in any way, shape, or form planned for the future in any capacity.

In machine learning, the objective is not to provide results that are an exact match to the data that has been provided; rather, the objective is to produce predictions by making use of data that has never been examined before. In contrast to this, classical learning has the objective of achieving a level of data matching that is as close as is practically possible. a machine learning model has to have the ability to generalize its findings beyond the confines of the dataset it was trained on. At this point in the process, the validation set is just beginning to acquire some level of significance. The intercept and slope of a linear regression are two examples of the numerous parameters that are used in models for machine learning. These two examples are taken from linear regressions.

These are only two instances of the numerous factors. There are many more. In order to properly fit these parameters to the data, a dataset is also utilized. Machine learning models, on the other hand, contain something called hyperparameters in addition to the normal model parameters. Adjusting the behavior of the model is accomplished with the help of these hyperparameters. To phrase this another way, the flexibility of a model as well as how well it fits the data is dictated by the model's hyperparameters. After we have completed the process of teaching a machine learning model on the training set, we can use the validation set to evaluate how accurate the model is at predicting labels that have never been seen before. This may be determined by contrasting the model's forecasts with the outcomes that actually occurred. To begin, we start by allowing the machine learning model to make predictions for us based on the attributes that are part of the validation set.

The following step involves determining how accurate those predictions are by comparing them to the actual labels that are part of the validation set. This provides us with some insight into the generalizability of the model, which is of great use to us. If the model cannot generalize beyond the training set (that is, if the predictions on the training set are noticeably better than the predictions on the validation set), then it is

likely that the model has been overfit. This may be determined by comparing the forecasts on the training set to the predictions on the validation set. It is possible to find out the answer to this question by contrasting the forecasts on the training set with the predictions on the validation set. one approach to resolving the issue of overfitting is to modify the values of the model's hyperparameters in such a manner that the model becomes less flexible and is better able to generalize the results of its investigations. This stage is highly crucial when it comes to the process of training a machine learning model. Models that do not generalize effectively are not likely to give correct predictions when they are utilized in an environment that is representative of the actual world. We end training the model when we are pleased with how effectively it generalizes from the training set to the validation set.

At that point, we move on to the validation phase of the process. An objective and all-encompassing evaluation of the model's capacity to produce correct predictions may be carried out by exchanging the training set for the test set. By comparing the labels that the model predicts for the test set data to the actual labels that are contained within the test set, we are able to evaluate how well the model is able to make predictions for data that was not utilized during the training phase. This allows us to assess how well the model is able to generate predictions for data that was not utilized during the testing phase. Because of this, we are able to evaluate how effectively the model can create predictions for data that was not used during the training phase.

2.2.1 Constructing an Estimate

At this point in time, it is very necessary for us to have a machine learning model that has some degree of promise. We are now in a position to start the process of making forecasts based on the present state of the globe. This will allow us to better prepare for the future. To continue with the example of bail prediction, the goal at this step is to use the machine learning model that has been developed in order to estimate the likelihood that a fresh defendant would return to court after being released on bond. This estimate will be made using the data that was collected during the previous stage of the process. To do this, it is necessary to calculate the chance that a newly charged defendant will show up in court after being granted release on bail. It is far more efficient to use a machine learning model that has already been trained to create predictions than it is to develop a model from the ground up from scratch. Rather than beginning from scratch, it is preferable to use a model that has already been trained. The only attributes of a defendant that we utilize as features are those characteristics

themselves, and the model uses those characteristics to make a forecast about the possibility that the defendant will appear in court at some point. After that, the court may use the prediction to determine whether or not to release the defendant on bail, or at the very least, it may provide the judge some more information that will aid her in reaching her judgment. After that, the court may use the prognosis to decide whether or not to release the defendant on bail.

2.3 PREDICTING HYGIENE VIOLATIONS

Food safety is an issue that has to be addressed in virtually every single city throughout the whole world. This is true on a global scale. Your health should not be placed in peril as a result of eating the food that is offered at dining facilities such as restaurants and other places where meals are served. As a result of this, several municipalities have hygiene inspectors working for them. These inspectors go to a variety of restaurants, seek for any issues related to hygiene that they could discover, and then strive to fix the issues that they find. However, there is a problem: cities usually do not have enough inspectors to visit every restaurant regularly enough, which results in certain sanitary breaches being undetected and poses a risk to the health of the general public. This is an issue since cities sometimes do not have enough inspectors to visit every restaurant frequently enough. This condition presents a number of challenges. The environment that the case study is being conducted in is the same one that was discussed earlier.

Our objective is to improve the distribution of hygiene inspectors across restaurants in order to detect as many instances of hygiene breaches as is reasonably practicable. In order to do this, our goal is to improve the distribution of hygiene inspectors. In order to do this, we want to work toward more evenly dispersing hygiene inspectors around the country. In an ideal world, hygiene inspectors would concentrate the majority of their time, attention, and effort on the establishments that had the most potential for breaking standards governing cleanliness. Techniques from machine learning might be utilized to make a determination regarding this danger. Because food authorities are aware that hygiene inspectors are an effective causal method of avoiding hygiene violations, it is necessary for them to obtain correct information on the possibility of a hygiene violation occurring. This is in reference to the difference that may be made when one predicts something vs when one infers a causal link between two events.

In this specific case study, I make use of a machine learning model that has been trained on historical data from the City of Seattle in an effort to forecast future violations of

cleanliness ordinances. The data used to train the model came from the City of Seattle. The data that was utilized to train the model originated from the city of Seattle. The information was taken from a report on a research study with the snappy title "Where Not to Eat? The study titled "Improving Public Policy by Predicting Hygiene Inspections Using Online Reviews" was written by Kang, Kuznetsova, Luca, and Choi. It was published in 2013 by Kang and the other co-authors of the article. 3 This dataset contains information that was obtained from a variety of sources, including the results of cleanliness inspections carried out by the City of Seattle as well as reviews of restaurants that were published on Yelp.

The information was collected in order to create this dataset. Yelp is a website that gives members of the general public the opportunity to provide ratings and reviews based on their personal experiences with local businesses. They found a correlation between the two sets of data, which led them to the conclusion that there was a relationship between the two sets. This led to the compilation of a dataset that included 152,153 Yelp reviews and 13,299 inspections of 1,756 distinct dining establishments. This dataset is regarded to be quite moderate in size when compared to the standards of machine learning, despite the fact that personally conducting an investigation of this dataset would involve a substantial amount of time.

First and foremost, we need to have a deeper understanding of the context before we can go on to the next step of the machine learning model's training process, which is to begin. How exactly are the ratings for Seattle's sanitary inspections determined, and what specific factors are taken into account in the process? When a restaurant in Seattle is submitted to a health inspection, the inspector will give the institution a score out of one hundred to indicate the degree to which the establishment complies with various regulations on public health and the preparation of food. The score will be based on the inspector's overall assessment of the restaurant's ability to meet the requirements of the regulations. This can be seen in Figure 2.3, which illustrates the distribution of the scores across the entirety of the dataset. The conclusion is less pleasing the higher the score that was received.

On the other hand, having a low number of good inspection ratings does not always indicate the presence of hygienic issues that might be harmful to customers. Infractions such as erroneous labeling will be found by inspectors, which will result in a higher inspection penalty score but will not pose a significant risk to the health of the general population. As a result, in an effort to make the subject of prediction easier to

understand, I do not make an attempt to anticipate the score on the inspection. Instead, I concentrate on determining whether or not a score is high enough to be considered a major breach of hygiene standards. Because of this, I am able to decide whether or not a score is high enough to make the process of prediction easier. The difficulty of producing a forecast as a result of this flips from being one of regression to being one of classification with two distinct classes. This is because of the effect that this has.

I use a percentage that is somewhere around 10% of the most serious hygiene infractions to characterize what I consider to be major hygiene breaches since there is no predefined barrier beyond which violations of hygiene are no longer seen as important. This is because there is no fixed threshold beyond which violations of hygiene are no longer regarded as significant. As a direct consequence of this, the cutoff point for a penalty score of 33 has been decided upon as the point at which infractions are considered to have reached a level of severity that warrants a substantial punishment. This particular category is accountable for exactly 10.3% of the total cleanliness scores that were discovered within the dataset.

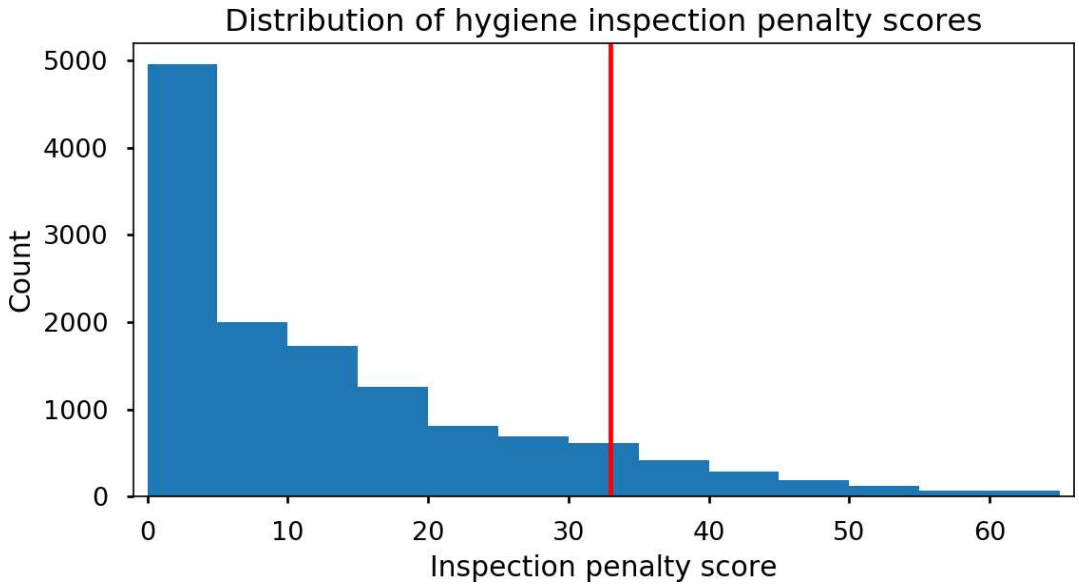


Figure 2.3: Distribution of hygiene inspections penalty scores resulting from 13,299 hygiene inspections of 1,756 restaurants in Seattle between 2006 and 2013

source: *machine learning for public policy making, data collection and processing through by Umesh Verma (2018)*

The gravity of the situation has a direct bearing on the degree of the punishment that is awarded. The cutoff point for the penalty score in the case study was set at 33, and the red line depicts where the cutoff point was located. If a violation of hygiene has a score that is either equal to or higher than 33, then the breach of hygiene is considered to be severe. This category is to blame for 10.3% of all of the violations that were found in the dataset, and it's liable for a total of 133 of them.

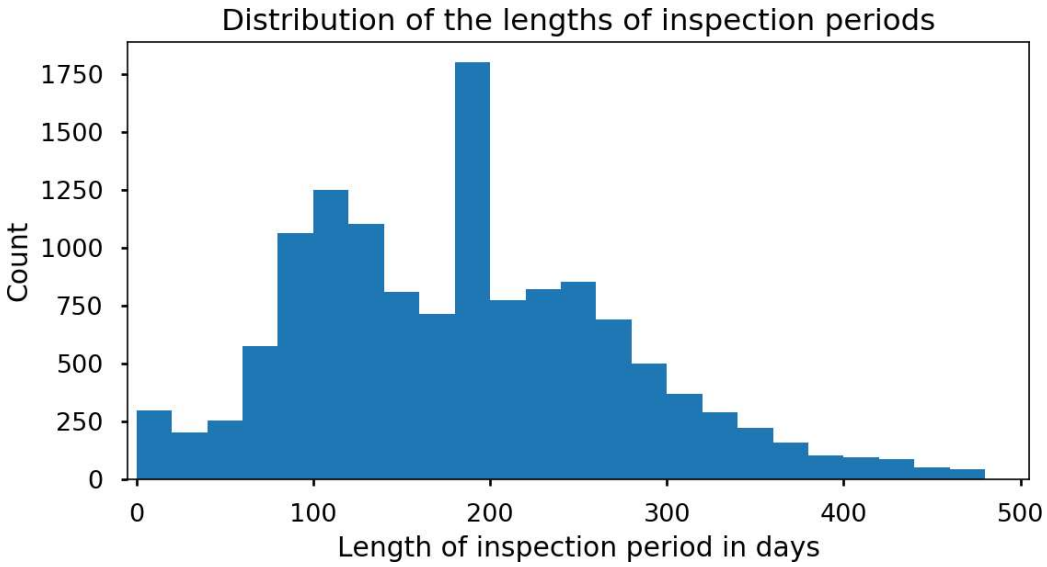


Figure 2.4: Distribution of the lengths of the inspection periods in the hygiene prediction dataset in days.

source: *machine learning for public policy making, data collection and processing through by Umesh Verma (2018)*

Due to the fact that there are far more inspection scores than there are restaurants in the dataset, a significant number of restaurants have been given multiple inspection scores at a variety of times. As a direct consequence of this fact, the total score for the dataset now possesses a far greater degree of variation. As a consequence of this, I stick to and define the inspection period of an inspection score as the period of time that begins one day after the conclusion of the inspection that came before the inspection that is being discussed here and ends one day before the inspection that is being discussed here. This is the period of time that constitutes the inspection period of an inspection score. This is the span of time that determines how well an inspection was performed. The "first inspection period" is the period of time that begins immediately before the inspection

takes place. This term is used in the context of eating establishments that have never been subjected to an audit in the past. The distribution of the different lengths of time that make up the inspection periods is shown in Figure 2.4, which was created in line with this description.

Inspection up to the day of the inspection at issue and includes that day. inspection up until the day of the inspection at issue. In the case of eating places that have never been subjected to an inspection in the past, the first inspection period comprises the time period that falls immediately before the first inspection (which explains why the histogram indicates an extremely high peak around the period of 180 days).

Estimating how the hygiene inspection will go at the conclusion of each inspection period is proving to be a challenge for us in terms of our ability to produce accurate forecasts. This presents us with a fresh challenge that must be surmounted. Due to the fact that there are only two classes involved, this prediction issue is really straightforward. How serious are the infractions of the hygiene requirements that were discovered during the inspection? Which of these pieces of information will be most useful to us in formulating an answer to the question that has been posed? The attributes that were used in the process of training the machine learning model that was deployed for this case study are presented in Table 4.1. The simplest way to examine all of them at once is to have them organized in the form of a large table, with each inspection period acting as a row and each attribute acting as a column. This will allow you to see all of them at once.

Table 2.2: Features and label contained in the dataset used in the hygiene violation case study

Data	Explanation
ZIP Code	The ZIP Code of the restaurant.
Cuisines	The cuisines offered in the restaurant according to Yelp, such as Japanese,Mexican, Pizza, Sandwiches etc.
Length of the inspection period indays	The inspection period of a hygiene inspection ranges from the day after theprevious inspection until the day of the inspection in question. The first inspection period of a restaurant spans the six months before the first hygiene inspection of that restaurant.
Number of reviews	The number of reviews of a restaurant that users posted on Yelp during theinspection period.

Average review rating	The average rating of the reviews of a restaurant posted during the inspection period (ranging from one to five stars).
Number of negative reviews	The number of reviews of a restaurant with a rating below or equal to threestars that users posted on Yelp during the inspection period.
Average previous inspection penaltyscore	The average of the hygiene inspection penalty scores assigned to a restaurant before the inspection in question (zero if there has been no previous inspection).
Previous inspection penalty score	The hygiene inspection penalty score assigned to a restaurant in the last inspection before the inspection in question (zero if there has been no previous inspection).
Review text	The concatenated texts of all reviews posted during the inspection period.
Inspection penaltyscore	The inspection penalty score assigned to a restaurant in the inspection period in question. The goal is to predict if this score is equal to or greater than 33, in which case a hygiene violation is labeled as severe.

The authors Kang et al. (2013) have completed a significant portion of the necessary data preparation for this case study. They compared the information they found about restaurants, including the review texts found on Yelp, to the cleanliness inspection ratings found on the website maintained by the city of Seattle. This is a mandatory step that must be carried out as it is an important component of the process. In spite of this, there is still further work that has to be done on the engineering of the features.

Quantitative representations of the data are required in order to train machine learning algorithms effectively. The majority of the data that is displayed in Table 4.1 is numerical in nature; hence, the table is organized in a manner that is appropriate for the presentation of numerical information. This comprises the total number of reviews, the number of days that the inspection period lasted, and the postal code. If the culinary feature and the hygiene violation label are encoded in the system as dummy variables, then the conversion process from dummy variables to numeric characteristics may be completed quickly and efficiently. This will guarantee that the conversion goes off without a hitch and in the most efficient manner possible. The review texts, on the other hand, will need to be converted into numerical data. How exactly should this be carried out?

Figure 2.5: A Yelp review that might indicate hygiene problems in a restaurant

source: *machine learning for public policy making, data collection and processing through by Umesh Verma (2018)*

As can be seen in Figure which can be seen here, a user of Yelp has the ability to write practically whatever they want in a review. This information is available here. This necessitates deviating from standard rules of syntax and spelling, in addition to conceiving up wholly new expressions in their place. Before feeding the text into a machine learning algorithm, it is strongly recommended that it be cleaned up, as doing so is likely to produce superior results. Even if machine learning is successful in finding patterns in the jumbled data, this fact remains unchanged. As a direct consequence of this, I conceived of the review texts in an organized and sequential manner, making use of the appropriate algorithms throughout the process.5. I went through each review and made sure that all of the punctuation marks had been deleted and that none of the words had their beginning letters capitalized. I also made sure that all of the terms were spelled correctly.

In order to make the assessments look more organized, I removed any and all superfluous words that served no use. Stopwords are words like "the" and "but" and "for" that are used so frequently in the English language that they do not communicate very much information that is useful for prediction. Examples of stopwords are "the" and "but" and "for." Stopwords include words like "the," "but," and "for," among other common instances.

Following the process of placing the remaining words into their dictionary form, I lemmatized them by first eliminating any inflectional ends and then proceeding with the process. As an illustration, the word "go" has been substituted for the words "goes," "went," and "going" in every occurrence in which it was found.

After completing up with this section of the preprocessing, the numerical elements of the review texts may subsequently be removed from them. The term frequency-inverse document frequency (TFIDF) statistic was applied to the concatenated texts of the reviews that were carried out during the course of an inspection period. I employed this statistic. This was done so that we could determine the frequency with which particular terms were mentioned in each review. The calculation of how frequently a word appears in the review texts during a specific inspection period, followed by the division of that frequency by a figure that represents the percentage of all review texts in which the term is used, is the fundamental principle that supports TFIDF6. This calculation is performed by first determining how frequently a word appears in the review texts, and then doing the division. Calculation of the TFIDF is the name given to this process.

If we simply computed the frequency of a word in a review text without normalizing it using the frequency of that word across all review texts, then terms that are more prevalent in the English language would naturally be overrepresented in comparison to words that are less frequent in the review texts. This would result in an inaccurate representation of the data. Even if we were to merely compute the frequency of a term in a single review text, we would still find that this is the case. TFIDF is able to circumvent this issue since it provides a realistic approximation of the frequency with which a phrase appears in a review text as opposed to the frequency with which it appears in all review texts. This allows it to more accurately compare the two frequencies.

I computed the TFIDF by not only using single words, which are referred to as "unigrams," but also by applying it to word pairs and word triples, which are respectively referred to as "bigrams" and "trigrams." This allowed me to get a more accurate picture of the frequency of each word combination. Because of this, I was able to offer the machine learning algorithm with information that was more appropriate to the purpose for which it was being employed. In the following table, which can be seen below, there is an explanation of the processes that must be taken in order to extract these n-grams from a text that has been preprocessed.

Table 2.3: The text preprocessing steps used in the hygiene violation case study applied to an example sentence

Original sentence	**We enjoyed our food, although parts of it were burned.**
Preprocessed sentence	enjoy food part burn
Unigrams	enjoy, food, part, burn
Bigrams	enjoy food, food part, part burn
Trigrams	enjoy food part, food part burn

The steps of eliminating all punctuation marks, changing all capital letters to lowercase, deleting stopwords, and lemmatizing the remaining words in the text are included in preprocessing. Preprocessing also lemmatizes the document. These procedures are carried out before any further processing of the document takes place.

The only thing that I added to a dataset that was already in existence were the TFIDFs of the 5,000 unigrams, bigrams, and trigrams that occur the most frequently. The tabular dataset that was produced as a direct consequence of this still consists of 13,299 rows (one for each inspection period), but it now has 5,135 columns: one for each characteristic, including those in Table (with the ZIP code and the cuisines encoded as dummy variables), plus the TFIDFs of the 5,000 most common uni-, bi-, and trigrams. As a direct consequence of this, a substantial dataset was produced, and it had 5,135 characteristics for each inspection period. However, standard modeling techniques could struggle to deal with datasets in which the number of variables is on the same order of magnitude as the number of observations. This is because the number of variables tends to grow exponentially with the number of observations. On the other hand, machine learning is able to manage datasets in which the number of variables is on the same order of magnitude as the number of observations. In other words, it can handle datasets with a high degree of complexity.

After the dataset was constructed, I utilized 80% of it for the training and validation sets, which included a total of 10,639 inspection periods, and I utilized the remaining 2,660 inspection periods for the test set. The dataset was composed of these sets in their combined form. Following that, I moved on to the next step of the machine learning process, which was to train a few different models by making use of both the training set and the validation set. It was discovered that the method of classifying random forest classifier was the one that was the one that was the most successful overall. On the other hand, the existing body of research on machine learning has a wealth of data that

provides adequate explanations in a variety of contexts. It is not within the purview of this specific study to investigate the inner-workings of machine learning algorithms such as this one or any others. In any event, the selection of a specific method for machine learning is often of less relevance than the quality of the feature engineering and the availability of pertinent data to begin with. In other words, the choice of a method is not as important as the other two factors together. In the majority of cases, having access to trustworthy data is more useful than having an innovative approach at one's disposal since it allows for better informed decision making.

To what degree of accuracy does the trained random forest model's forecasting lead us to expect results? In the field of machine learning, it is standard practice to employ a statistic that is known as the area under the receiver operating characteristic curve (often abbreviated as AUC) when measuring how well a model can predict the results of an experiment. The outcomes of this statistic served as the basis for my assessment of how successfully this strategy had been implemented. Our random forest model has an area under the curve (AUC) of 0.7, as shown in Figure2.6, which is a level of performance that may be considered to be rather excellent. In spite of the fact that it has a name that may seem puzzling to some, the AUC methodology is actually rather simple in its method of issue resolution. The area under the curve (AUC) is the percentage of randomly selected pairings for which the model predicts a larger hygiene violation risk for the inspection period that resulted in a major hygiene violation.

This percentage is expressed as a percentage of the total number of pairings. Taking the absolute value of the area under the curve is the way to determine this percentage's value. To compute this, first pick at random one inspection period from among all inspection periods that end with a serious hygiene violation. Next, select at random another inspection period from among all inspection periods that do not end with a severe hygiene violation. Finally, multiply the two numbers together to get the total number of inspection periods. The disparity in length between these two inspection periods serves as an indicator of the frequency with which a serious hygiene violation was found within the relevant inspection periods.

To put it another way, if we are provided with a random pair of inspection periods, one of which has a hygienic violation and the other of which does not, there is a 70% chance that our model will assign a higher risk rating to the inspection period that has the severe violation than it will to the inspection period that does not have a severe violation. This is because our model takes into account the severity of the violation

when determining which inspection period has the higher risk. This is because the severity of the infraction is taken into account by our model when calculating which inspection period has the larger risk. These discoveries have provided evidence that demonstrates the validity of our concept. A perfect model would properly rank each and every one of these pairings and have an AUC of 1, but having a likelihood of 70% of ranking correctly is much better than having a chance of just 50% of correctly ranking the pairs in question if you guessed at random. An ideal model would have an area under the curve (AUC) of 1.

In a perfect world, the area under the curve (AUC) would equal 1. It would appear that the information needed by our model to acquire knowledge through machine learning has been gathered.

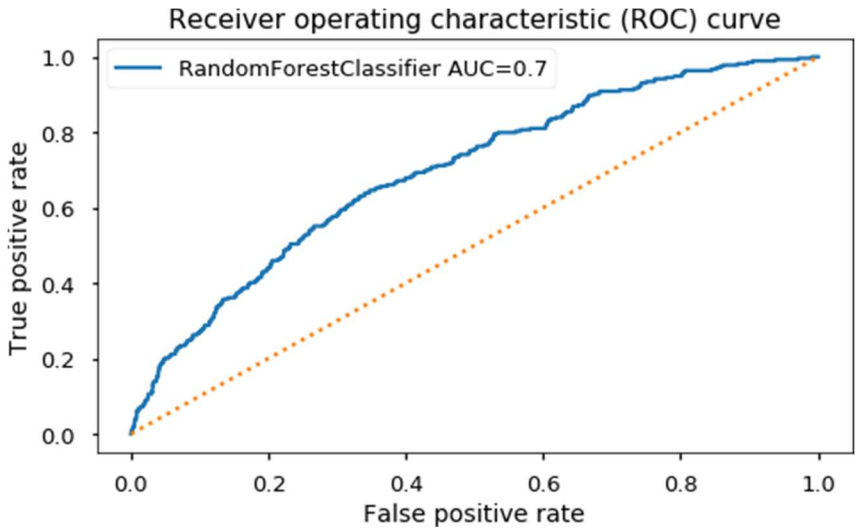


Figure 2.6: Receiver operating characteristic curve (ROC) for the random forest classifier used in the hygiene prediction case study

source: *machine learning for public policy making, data collection and processing through by Umesh Verma (2018)*

This graphic offers exact definitions of the true positive rate and the false positive rate, which may be found in the statistical literature. The true positive rate is the percentage of tests that result in a positive diagnosis. For the purposes of this case study, it will be enough to ensure that the classifier achieves an area under the receiver operating characteristic curve that corresponds to 0.7 of the goal value.

This will fulfill all of the requirements for success. This indicates that, given a random pair of inspection periods, one of which has a hygiene violation and the other of which does not, there is a 70% chance that the model will rank the inspection period that has the severe violation as having a greater risk than the inspection period that does not have the violation. This indicates that there is a 70% chance that the model will rank the inspection period that has the severe violation as having a greater risk than the inspection period that does not have the violation. This is due to the fact that when making its conclusions, the model takes into account the level of seriousness associated with the infraction. This is because the model takes into consideration the extent to which the rule was breached, which explains why we see this result. When an AUC of 0.5 is utilized, the line that is illustrated by the dots represents the ROC that is anticipated to be the absolute worst case scenario. If the ROC had a value of one in each and every site, then the AUC would also be one, and this would show that the ROC was perfect. If the ROC did not have a value of one in any of the sites, then the AUC will always be zero.

How accurate is our model when it is really put into practice and used to analyze something? The discipline of machine learning frequently makes use of a statistic that is referred to as AUC, which stands for "area under the curve." In spite of the fact that it is a common statistic, it does not tell us very much about the ability of our model to forecast events in the world that actually exists. It makes perfect sense to take a look at the myriad of various sorts of prediction mistakes that our model generates in order to obtain a more accurate picture of the accuracy of the forecasts that it gives. This will allow us to better understand how accurate its predictions are. The predictions that our model produces are highly inaccurate and come in a wide variety of flavors. The confusion matrix, sometimes known as "the confusion matrix," may be found presented in table 4.3. This matrix makes a comparison between the predictions made by the model and the actual labels that are contained inside the test set.

It is possible to make a prediction mistake of either the false-positive or false-negative form. Both types of errors are possible. Both of these types of mistakes are entirely conceivable. When a model incorrectly predicts a significant hygiene violation when, in reality, there was no such serious violation, this type of error is referred to as a false positive error. When a model mistakenly forecasts a major breach in food safety, for instance, this is an example of the sort of error that can occur. When a model predicts that there will not be any serious hygiene breaches during an inspection period, but

there really are such violations, this is an example of a false negative error. False positive errors are the opposite of false negative errors. This can take place if a model projects that there would be no major hygiene breaches throughout the period that is being inspected. According to the confusion matrix, our model achieves an accuracy level of 75.5 percent across the board for all 2,660 of the different inspection intervals that are included in the test set.

This would imply that 126 more of the model's forecasts were true than the previous total of 1,882, bringing the overall number of accurate predictions generated by the model to a total of 2,008. There are still 504 erroneous positive predictions and 148 incorrect negative predictions, and these two categories combined account for 24.5% of all of the predictions that were made using the test set. Despite the fact that this is not a terrible outcome, it does not change the fact that there are still incorrect predictions. As a consequence of this, it is probably not a good idea to place a complete reliance on this technique in the process of assigning hygiene inspectors to restaurants in the Seattle area. It is feasible that the forecasts may be used to offer more information to hygiene inspectors on which restaurantss are likely to commit hygiene violations; however, given that every fourth prediction is incorrect, it is probably not reliable enough to be utilized on its own.

There is a possibility that the predictions may be used to offer further information to hygiene inspectors on which restaurantss are likely to commit hygiene violations. One use for the projections would be to offer more information to hygiene inspectors about which restaurants are most likely to violate hygiene regulations.

Table 2.4 : Confusion matrix for the random forest classifier trained using the training and validation sets of in the hygiene violation case study

Accuracy: 75.5% (n= 2,660)		**Predicted label**	
		Severe violation	**Not severe**
Actual label	**Severe violation**	**126** (true positives)	**148** (false negatives)
	Not severe	**504** (false positives)	**1,882** (true negatives)

The accuracy of these forecasts may be determined by applying them to the findings of the 2,660 inspection periods that comprise the test set. This will allow for a more

comprehensive analysis of the data. This collection is responsible for twenty percent of the entirety of the dataset that was examined.

Despite this, the case study demonstrates that making use of data-driven predictive modeling is one method that has the potential to be lucrative when it comes to increasing the distribution of hygiene inspectors across various eateries. This is due to the fact that it enables a more precise calculation of the locations at which inspectors should be deployed. Inspections of hygiene are a great illustration of a prediction issue that could be amenable to being solved by machine learning in the future. When it comes to inspections, there are recurring choices that need to be made; there are new data sources that can be accessed in the form of Yelp reviews that include projected trends; and it is simple to evaluate hygiene violation predictions by sending an inspector to a restaurant. Because Yelp provides a standard online platform not just for the city of Seattle but also for a great number of other cities, it would not be difficult to expand the machine learning model that was created for this case study so that it could be used in other parts of the country. This was done for the purpose of demonstrating the applicability of the model.

This was done with the goal of determining whether or not machine learning might be utilized to accurately forecast the actions of consumers. There is a good chance that this would make the model better able to make accurate predictions, and it would also make it possible for the model to have an influence in the world that actually exists. A recent study, for example, found that the City of Boston may be between 30 and 50 percent more successful in allocating hygiene inspectors to restaurants if it adopted the winning algorithm from an online machine learning tournament (Glaeser et al. 2016). The research was conducted by Glaeser and colleagues. Glaeser and his colleagues were the ones who carried out the research. The findings of the research that Glaeser and his colleagues conducted can be credited with leading to this conclusion. This exemplifies how machine learning may make the process of public policy making more successful, which is not only a significant advantage for financially struggling local governments but also an example of how machine learning may make the process of public policy making more successful.

2.3 MACHINE PREDICTIONS

When applied to problems that call for prediction as a solution, machine learning may be used to find solutions, which can then be used to generate machine predictions as

the end result of this process. We have demonstrated that there are difficulties associated with prediction in the process of developing public policy, and we have discussed how machine learning may be a method that may be applied to confront these difficulties in a way that is applicable in a manner that is relevant in the real world. In this piece of study, we are going to take a more in-depth look at the precise situations under which machine predictions may be advantageous for the formulation of public policy, as well as the benefits that machine learning brings in such scenarios.

Additionally, we are going to examine the specific conditions under which machine predictions may be advantageous for the creation of private policy. This will be done in combination with an investigation of the circumstances under which the application of machine predictions to the formulation of public policy may prove to be beneficial. This will be done in conjunction with an assessment of the benefits that may be obtained via the implementation of machine learning in environments similar to these. In addition to this, we will examine a large number of additional instances of prediction challenges that have previously been solved using the use of machine learning in the process of formulating public policy. These examples will be examined below. After we have finished going over the prior content, we will move on to this next step.

2.3.1 Suitable Prediction Problems

We will be concentrating on those aspects of the process of formulating public policy that do not lend themselves very well to being resolved by machine learning, as these are the aspects on which we will be focusing here. Before the findings that machine predictions generate can be put to any kind of practical use, it is required to satisfy a number of preconditions. One of the requirements is that the occurrence of the event that is to be predicted must be frequent enough for statistical methods to be applicable. This is a criterion that must be met. The purpose of this criterion is to guarantee that one can make an accurate prediction of the results. The decision-makers who are in charge of public policy would like to know the outcome of a number of unknown occurrences; yet, some of these occurrences are so exceptional that the approach of data-based modeling is not sufficient for forecasting their outcomes. Who will come out on top in the upcoming battle, and why is it important?

Will a disagreement be significant enough to thwart the efforts of the party that controls the government to achieve its objectives? When faced with the following decision, what course of action do you think an unpredictably powerful leader of a state would take?

These are really important issues, but unfortunately, machine learning is unable to provide answers to them. This is due to the fact that the occurrences in question are extremely uncommon, and every single one is unique in its own manner. As a result, it is impossible for machine learning to offer satisfactory responses to these inquiries. The outcomes that are to be predicted ought to be sufficiently similar to leave comparable patterns in the data behind after they have occurred. After then, a program designed for the goal of machine learning, which is a branch of artificial intelligence, may recognize these patterns and use them accordingly. This leads one to believe that the objective that is to be expected should not be of a particularly complicated nature on its own.

Sending an inspector into a restaurant allows for the possibility of discovering sanitary infractions similar to those outlined in the case study. Because all dining establishments are obligated to operate in line with the same regulations and standards, this is an approach that is shown to be effective. Predictions about bail are also not overly difficult because they depend mostly on the judge's assessment of the likelihood that the detainee will comply with the terms of their release and appear in court again after having been granted bail. The reason for this is that the judge takes into consideration the amount of time that has transpired since the prisoner was granted release on bail. Predicting the outcomes of other types of judicial judgements, on the other hand, is a far more challenging task to do. When determining the length of a sentence, it is common practice to take into consideration a number of different criteria.

Deterrence, punishment, and regret are some of the elements that are taken into consideration; however, some of these characteristics are difficult to measure and are not included in this discussion. When determining a defendant's sentence, the court takes into account a range of criteria, one of which is the defendant's potential for future criminal behavior, which may be forecast. As one of the key reasons, this is one of the primary reasons why machine learning is not a sufficient answer for the sentencing problem, as stated by. One of these requirements is that a model must offer some method for determining how accurate the predictions it generates are going to be.

Even if an algorithm has an excellent predictive performance on the data that was used to train it, this will not be of much help if the predictions that it creates in the actual world are too inaccurate. If this is the case, the algorithm will not be of much use. It is essential to examine not just the accuracy of an algorithm's predictions but also the degree to which those predictions may be effectively extended. This is because

accuracy is only half the equation. If, for example, an algorithm projected that a defendant would not commit a crime after being released from jail, but the defendant ended up committing a crime, then it is evident that the algorithm's forecast was erroneous. As we will see in the next section on Study, which is entirely devoted to the issue of prediction evaluation, it is not always as straightforward as one may imagine it to be to evaluate predictions.

2.3.2 Advantages of Machine Learning

What distinct benefits does machine learning have over the myriad of other methods that may be employed to model data in particular? It shouldn't come as much of a surprise to anyone that one of the primary benefits of machine learning is that it is incredibly successful at prediction. This is one of the fundamental advantages of machine learning. This talent was polished over the course of time owing to the efforts that were combined from a number of distinct contributing sources. Machine learning can deal with high-dimensional data, which is a situation in which the number of variables is equivalent to the total number of observations. This type of data presents a challenge for traditional statistical methods. This is one of the contributing factors, and it applies to situations in which the number of variables is equal to the total number of observations. Another one of the contributing elements is that the total number of observations is equal to the number of variables. Learning machines are highly skilled when it comes to spotting important patterns in data and establishing the connection between those patterns and the goal of producing predictions.

Learning machines are also extremely proficient when it comes to providing predictions based on what they have learned. The fact that the data that is utilized in machine learning may arrive in a number of forms is an advantage that is closely tied to this one and is one of the reasons why it is advantageous. This is a perk that is of great assistance to the user. A straightforward one is comprised of nothing more than an enormous table, with the observations being listed in the table's rows and the variables being stated in the table's columns. But because big data is often the output of the innumerable computer-mediated transactions that take place in our modern world, it may also consist of items such as text, images, videos, sounds, sensor data, and a great many other things – in general, any information that is digital or can be digitized. This is due to the fact that large amounts of data are typically the result of the countless computer-mediated transactions that take place in our contemporary environment.

This is owing to the fact that vast volumes of data are routinely created as a side consequence of the numerous computer-mediated transactions that take place in our modern environment. This is because of the proliferation of computer technology in our modern environment. For instance, each and every transaction that takes place over the internet leaves digital traces behind, which may be of great aid when it comes to the process of formulating new regulations for the general public. some individuals characterize "big data" in terms of the so-called "three Vs," which are larger data volumes, higher data velocities, and more data varieties.

In other words, greater data volumes, higher data velocities, and more data variations. These aspects include a greater volume of data, a faster velocity of data, and a greater variance of data. conventional data are frequently available only after a longer amount of time has passed, and they come in a form that is less thorough than the form that big data do. On the other hand, large amounts of data are frequently available with a much reduced temporal lag. When data are aggregated at the country or regional level, the potential of machine learning is wasted since the process of aggregation, which is still utilized frequently in the process of public policy making, for example, results in the loss of a sizeable quantity of information.

This is because aggregation happens on a higher level than on individual devices, which explains why this is the case. If public policy makers were to employ new sources of granular big data rather than aggregated information, this may possibly provide them with a significant increase in their own capacity to predict outcomes. The final advantage that may be acquired via the use of machine learning is disclosed once a model has been polished and improved to its fullest potential. expanding the scope of the model's application is a straightforward process in the context of this particular scenario.

After being trained, a model is able to make predictions on fresh data with a marginal cost that is extremely near to $0. This is because the cost of training the model was $0. This is due to the fact that the model is able to gain knowledge from its prior mistakes. It is not essential to have anything more advanced than a system that is capable of storing data and producing predictions based on a model. This is the minimum level of sophistication that is required. Because of this, scaling up a model that utilizes machine learning is straightforward, and the program gives more accurate forecasts with fewer variables than individuals do. However, it is necessary to ascertain the accuracy of the projections that are created from the additional data that has been provided.

2.3.4 Examples from Public Policy Making

In the context of utilizing public policy in an effort to make accurate forecasts of the future, what kinds of challenges may possibly materialize? Table provides an overview of the challenges associated with prediction that are present in a range of sectors that are related to the formulation of public policy, as well as an explanation of how these obstacles can be overcome by employing data-driven models. The solutions to these problems are discussed in further detail in Table 5.1. The purpose of the prediction, the data and techniques that were used, as well as the findings of the investigation that was carried out are all covered in this section of the report. While some of the study makes use of more sophisticated machine learning methodologies, other parts of it relies on more conventional regression models. In the process of data analysis, both kinds of models are considered.

In any event, machine learning stands out as a method that has the potential to be a viable alternative for the process of producing predictive models. This is true despite the fact that it is still a relatively new field of research. As the chart that follows shows, data-driven modeling and machine learning both have a variety of possible uses in the process of formulating public policy. The policies that are being called into question include, amongst other things, those that are related to the economy, taxation, and health care, as well as those that are related to agriculture, education, and public engineering. They originate not just from nations whose economies are emerging but also from those whose economies are already well-developed.

It is possible to use a wide variety of data sources, such as satellite images, electronic health records, and data collected in classrooms, in order to address problems involving prediction. Concerns regarding prediction are necessary at all different scales, from the local to the global. The table that follows provides an illustration of the multiple challenges connected with forecasting that come up all along the process of formulating public policy. It also provides an illustration of the possible contributions that machine learning may make toward the resolution of these challenges.

When looking at this table, there is still one question that has to be answered: what kinds of practical applications of machine learning are available for resolving prediction challenges that public policy makers are up against? Even the majority of quantitative public policy makers who are experienced with data modeling have probably never created a machine learning algorithm on their own since machine

learning has only recently garnered a lot of attention. This is because machine learning has only just recently gained a lot of attention. This is due to the fact that there has been a lot of focus placed on machine learning only very lately. This is because, in recent times, the topic of machine learning has garnered a lot of interest from researchers and industry leaders. Although it is feasible for policy makers with an interest in statistics to learn how to apply machine learning through the use of books and (online) courses, one strategy that has a better chance of being successful is to engage with consultants who are specialists in machine learning. It is probably the easiest method to tackle a prediction problem by making use of machine learning, and that is to discuss the prediction difficulties that are now at hand with them, while also being conscious of the hurdles and constraints that are addressed in Study.

In the event that the cost of this solution proves to be unfeasible, a workable and more cost-effective alternative would be to collaborate with members of the data science community who contribute their time. The vast majority of the time, individuals who specialize in data science and machine learning are more than happy to devote their time to causes that are geared toward making the world a better place as a whole. It is feasible that beginning the process of debugging machine learning difficulties on one of the several websites that conduct machine learning contests, such as Kaggle7 or DataKind8, might prove to be advantageous in the long run. These platforms post datasets alongside a particular prediction challenge, and they then ask the community of machine learning experts to collaborate on the problem in order to find a solution.

A reward will be given to the person who proposes the idea that proves to be the most practical response. One may get a decent understanding of the kind of prediction issues that have already been solved by utilizing this strategy by reading through the finished prediction competitions that are presented on these websites. These websites offer a variety of prediction contests. Competitions for making predictions are hosted on these websites.

Table 2.5 : Examples of machine learning being applied to prediction problems from different areas of public policy making

	Study	**Method, data and goal**	**Result**
	Combining satellite	Convolutional neural networks	The inexpensive and scalable

	imagery and machine	are trained on publicly available	model can explain up to 75% of
	learning to predict	high-resolution satellite imagery	the variation in local-level
	poverty	to estimate local consumption	economic outcomes. It could
	(Jean et al. 2016)	expenditure and asset wealth in	transform how poverty is
		five African countries.	targeted and tracked in
			developing countries.
	Predicting poverty	Automatic feature engineering	In regions where censuses and
	and wealth from	and elastic net regularization are	household surveys are rare, the
	mobile phone	used on an individual's past	method allows to gather
	metadata	history of mobile phone use to	inexpensive, localized and
	(Blumenstock,	infer her socioeconomic status.	timely information (at a finer
	Cadamuro, and On		level than satellite imagery).
	2015)		

Agricultural policy	Random Forests for Global and Regional Crop Yield Prediction (Jeong et al. 2016)	Random Forests are compared to multiple linear regressions for their ability to predict crop yields of wheat, maize, and potato using climate and biophysical variables at global and regional scales.	Random Forests outperformed multiple linear regression benchmarks in all performance statistics, making them an effective and versatile machine learning method for crop yield prediction.

Health policy	Predictive Modeling for Public Health: Preventing Childhood Lead Poisoning (Potash et al. 2015)	Logistic regressions, support vector machines and random forests are used to predict the risk of children being poisoned by lead in their homes in Chicago. Data comes from blood tests, home lead inspections, property value assessments and censuses.	The models allow the Department of Public Health toprioritize which households to target when trying to prevent lead poisoning before it occurs. This is a better method than waiting for blood tests to indicate poisonings after the fact.
	Prediction mode-lingusing EHR data: Challenges, strategies, and a comparison of machine learning approaches (Wu, Roy, and Stewart 2010)	Logistic regressions, support vector machines and boosting are used on data from electronichealth records (EHRs) to detect heart failure before the actual date of clinical diagnosis.	The models are able to predict heart failure more than 6 months before the actual clinical diagnosis reasonably well. This means that a patient's health history can be used to predict future illnesses and target interventions.
Engineering	Water pipe condition assessment: a hierarchical beta process approach for sparse incident data (Li et al. 2014)	Bayesian nonparametric learning and existing infrastructure data are used to predict the failure probability of water pipes in a city to establish a ranking for inspections.	Experimental results show that the model does better than current best practice methods, leading to substantial savings on reactive repairs and maintenance.
Public hiring	Productivity and Selection of Human Capital with Machine Learning (Chalfin et al. 2016)	Stochastic gradient boosting and regression with Lasso regularization are used to improve police hiring decisions and teacher tenure decisions. The data used includes surveys as well as socio-demographic and classroom data.	Using machine learning models for hiring decisions can potentially reduce the excessive use of force by police and improve police-community relations. Similarly, students would benefit from better teacher hiring decisions.

Tax policy	Collaborative information acquisition for data-driven decisions (Kong and Saar-Tsechansky 2014)	Combinations of multiple learners and a variety of data sources are used to improve the cost efficiency of tax audit decisions.	The approach could increase sales tax profits by an average of 4 percent, strengthening this revenue source for governments.

2.4 CHALLENGES AND LIMITATIONS

The findings of this study up to this point have provided data that demonstrates the viability of utilizing machine learning in the process of making public policy. There is a chance that predictive modeling will completely alter the approach that public policy makers use to addressing difficulties associated with prediction. However, machine learning is not a panacea by any stretch of the imagination. In this part of the article, we will talk about the challenges and constraints that come along with machine predictions. Those who are responsible for developing public policy really must have a thorough understanding of the challenges and constraints that now exist.

In point of fact, the majority of public policy makers will engage with professional sellers of prediction software or consultants rather than putting into operation a machine learning solution on their own rather than doing so. This is because of the expertise and experience these individuals bring to the table. According to, firms that market solutions that include machine learning have a tendency to oversell their wares by making assurances that are hard to maintain. This is because businesses that market solutions that include machine learning have a tendency to oversell their wares.

In a situation such as this one, it is of the utmost importance to make certain that the right questions are being asked and to refrain from accepting any claim at face value. This study was conducted with the intention of enlightening those individuals who are in charge of formulating public policy. The hurdles that machine predictions face in the process of developing public policy may be broken down into three categories: the limits of prediction, technological and human barriers, and ethical and legal issues. Each of these categories has its own unique set of obstacles and constraints. Every one of these spheres comes with its own unique set of difficulties and constraints.

2.4.1 The Boundaries of Our Predictions

Do we have the capacity to anticipate all that may occur? Not in the least, as expected. Even though we live in a highly connected society, in which technological systems engage with the social character of its users (Vespignani 2009), and in which ever-increasing amounts of data are produced as a result of these interactions, there are still some things that cannot be expected. Even though we live in a society in which technological systems interact with the social character of its users, there are still some things that cannot be predicted. Predictions regarding the behavior of complex techno-social systems, such as those that public policy makers are forced to deal with, can never be guaranteed to be accurate to a one hundred percent degree. the complex systems that exist in the modern world are made up of a large number of discrete units, each of which acts in a manner that is, to some extent, random and unexpected.

Additionally, Brunner states that these discrete units each have their own unique characteristics and behaviors. The world is continually changing for a number of different causes, including those that are associated with advancements in technology, changes in the environment, and societal changes. As a consequence of this change, there is no assurance that things that were true yesterday will continue to be true today or tomorrow. Even if we had access to the most powerful algorithms and the most huge quantities of data, we still wouldn't be able to predict the behavior of such complex systems with exact precision. This is because such systems are inherently unpredictable. This basic hurdle cannot be overcome by even the most sophisticated techniques of machine learning.

What does this involve, speaking in words that are more common? When applied to a system whose behavior we seek to anticipate, it is likely that a machine learning model that has been trained on data from the past would not be able to accurately forecast future data. This is because the system is too dynamic for the model to be able to consistently predict the behavior of the system. It is not feasible to exclude this danger in its entirety from the scenario. In point of fact, one of the reasons why it is so difficult to foresee the behavior of complex techno-social systems is due to the frequent meddling of policymakers. This interference can take many forms. Take, for example, the case study that involved generating predictions regarding the level of cleanliness. By gathering information on eateries and utilizing data from Yelp reviews, the machine learning program was able to generate accurate predictions regarding hygiene violations in the Seattle area. In the hypothetical situation in which this model was put

into effect, public health inspectors would be sent to those restaurants that the algorithm identified to be the most significant. These restaurants would be prioritized for inspection. Unfortunately, this intervention might render some of the model's future forecasts inaccurate.

This suggests that potentially hazardous eateries realize they can no longer escape hygiene inspections, which leads to an improvement in the current hygiene conditions at such places. This would be the best-case situation. The worst-case scenario is that restaurants find out that Yelp reviews might be enlightening, and as a result, they offer incentives to their customers to encourage them to leave favorable evaluations, or they even buy phony reviews, all in an effort to hide the opinions of their actual customers. there is evidence that the use of big data and innovative prediction approaches have not resulted in an increase in the accuracy of the forecasts provided for some reactive systems. This is the conclusion reached by the authors based on their examination of the available data. When compared side-by-side, predicting the behavior of systems that react to forecasts is a far more challenging endeavor than forecasting the weather. Those unfortunate souls who are tasked with the task of developing public policy nearly invariably have to cope with the second scenario.

In the event that the projected patterns in the data don't materialize, what countermeasures do we have at our disposal? The only solution is to do routine updates on the machine learning models that are currently being utilized. This is analogous to the behavior that humans would have in a similar situation. In the absence of a predictive algorithm, it is necessary for hygiene inspectors to choose restaurants for inspection on their own. Hygiene inspectors must choose which restaurants they will check. In some of the restaurants they investigate, they do not find any substantial violations of sanitation regulations, but in others they do. This affords the inspectors the chance to reflect on their prior experiences and draw conclusions about the areas that are most likely to violate the hygiene regulations. It is feasible to say that inspectors build their very own internal models in order to foresee breaches of hygiene requirements. This is something that is done in order to make predictions.

They should alter their internal beliefs in light of any new information that comes to light, whether it be the discovery of a violation in an area where they had not expected finding one or the lack of a violation in a location where they had anticipated finding one. Either way, they should take into account the new information when making these adjustments. It is necessary to make use of a strategy that is somewhat comparable to

what is being presented here in order to put machine learning into action. Models need to have their parameters updated on a regular basis so that they may incorporate any newly gathered information. This process is referred to as "retraining." it is possible to do so by keeping up with the changes in predictive patterns and ensuring that a model continues to make accurate predictions using this strategy.

2.4.2 Technical and Human Obstacles to Overcome

State that some of the technological challenges in the field of machine learning include data availability, data management, and processing. The creation of a predictive model requires the utilization of training data; however, public policymakers generally do not have access to the same wealth of data that their counterparts in the private sector have. It may be difficult for public policy makers to gain access to appropriate big data for the purpose of machine learning because of the common suspicion that the government is gathering an excessive quantity of personal information. Even if individuals in control of policy are given access to the necessary huge data, the subsequent procedures probably won't be much easier. A substantial quantity of storage space as well as processing power is required for the management of huge volumes of data and the performance of computations using that data.

As the quantity of datasets continues to increase, even conceptually straightforward jobs take a noticeably longer amount of time and need a noticeably greater amount of labor. Even processes that seem relatively basic, such as removing and summarizing variables from vast quantities of data and analyzing the links between them, can take a considerable amount of tim). even the normally well-paying private sector experiences problems when it comes to hiring qualified people for tasks like these. This is the case despite the fact that the private sector has a larger labor pool., one of the most significant challenges that the public sector may face is the lack of available people who possess the necessary skills.

One other technological obstacle that has to be conquered is determining how accurate machine forecasts are. Decision makers want to be presented with data that confirms the accuracy of computer forecasts before they will utilize them. There are many different facets of the process of making public policy that lend themselves well to the execution of experiments. The use of randomized controlled trials, which are particularly common in the education or public health sectors of poor nations, can be put to use to evaluate hypotheses. This can be done in a number of different fields.

They are the most efficient approach for assessing the accuracy of machine predictions, as stated by Kleinberg, Ludwig, and Mullainathan (2016). However, there are other disciplines, like the subject of criminal justice, in which it is not possible to conduct experiments. Validating the accuracy of the machine's predictions under these kinds of conditions is a challenging task.

Take, for example, the situation with the forecast of the bail. As was mentioned at the beginning of this inquiry, the authors of this paper assert that by making use of their machine learning algorithm, "crime can be reduced by up to 24.8% with no changes in jailing rates, or jail populations can be reduced by 42% with no increases in crime rates". This assertion was made in light of the fact that it was discussed at the beginning of this inquiry. This contradicts the authors' earlier statement that the use of their algorithm will lead to "no increases in crime rates."

How are they able to make that determination? They most definitely did not carry out an experiment in which they put people in jail or let them out based on their predictions, since that would be completely unethical. It is extremely difficult to know what the behavior of those who were permitted to remain in jail would have been if they had been permitted to post bail and be freed from detention. In order to find a solution to this problem, have to resort to more sophisticated econometric methods. These strategies take advantage of the fact that judges are assigned to defendants in a way that is basically random, and it is well-known that some judges are known to be more rigorous in their bail rulings than others. This examination will not get into the intricacies of their econometric technique; rather, the example that was shown earlier indicates how challenging it may be to evaluate machine predictions in the context of the social domain.

To be truthful, conducting one's research based on the projections of humans is not always the simplest alternative. Because of this, it is extremely challenging to evaluate the accuracy of the forecast made by a single judge on the possibility that a defendant would commit another crime in the future. Human forecasts, on the other hand, are not only far more prone to unexpected outcomes, but they are also frequently influenced by factors that have no business being involved in the decision-making process. In contrast to this, an algorithm accurately predicts the same outcome each time, provided that the same inputs are used in the calculation. There is evidence that judges are effected by factors that should be extraneous to court rulings, including literally "what the judge ate for breakfast". One such factor is literally "what the judge ate for

breakfast". One of these factors is figuratively speaking "what the judge ate for breakfast". In such circumstances, the use of huge volumes of data and the precise evaluation of projections is a problem, but it also presents an opportunity to bring greater rigor to essential areas of decision making in ways that were previously inconceivable.

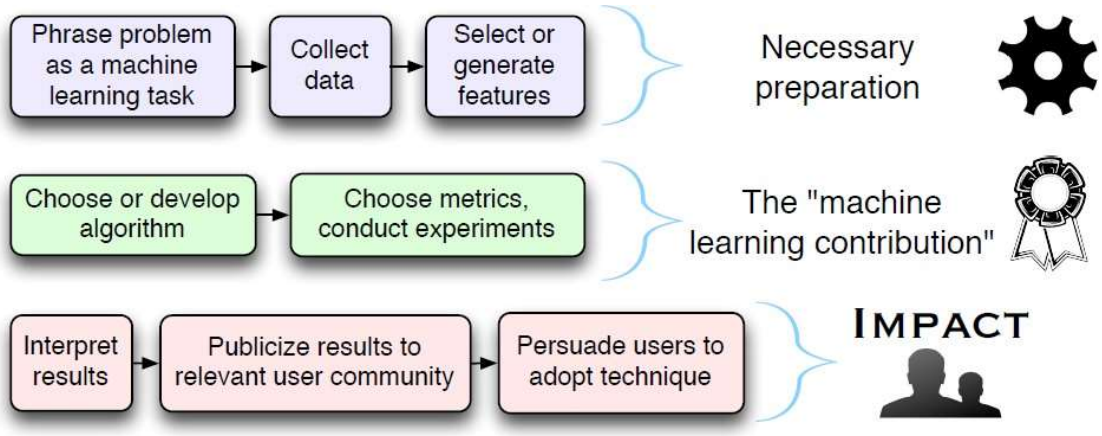


Figure 2.7 : The long way from identifying a prediction problem suitable for machine learning to real- world impact

source: *machine learning for public policy making, data collection and processing through by Umesh Verma (2018)*

The ability for machines to make accurate predictions is not something that is just tough to achieve from a technology viewpoint. In addition to it, there is a significant element of human involvement. Figure 2.7 makes it abundantly evident that there is a substantial distance to travel between identifying a prediction problem that is susceptible to machine learning and creating data-driven forecasts that have an influence in the actual world. This is made clear by the fact that there is a large gap between the two. In a broad sense, the activities and procedures that take place in the time period between the gathering of data and the carrying out of experiments to evaluate the accuracy of machine predictions are referred to as technical phases. They are in the greatest need of the expertise and experience of professionals who have been trained in machine learning and possess the requisite competencies. The problem, on the other hand, is not so much technology as it is human when it comes to the phases at the beginning and the end of the journey to genuine influence in the world.

These stages concern the beginning and the end of the path. Each and every execution of a machine prediction solution must necessarily start with the participation of human beings and must ultimately come to a finish with their participation. Everything begins with the identification of a prediction problem and the framing of that problem as a challenge for machine learning. From that point on, everything else is constructed upon that foundation. If public policy experts have never heard of how machine forecasts are possible, it is probable that they will struggle to perceive opportunities of this type. There have been various attempts made to equip individuals with a mindset that is more data-driven; but, due to the fact that data science is still a relatively young field of study, these efforts are still in their infancy and have not yet proven successful. If a public policy maker reads this article and takes away nothing else from it, the knowledge that data and machine learning can be leveraged to address prediction challenges is already a huge step in the right direction. This paper was written to educate public policy makers.

It is necessary for people to factor these projections into their decision-making processes so that the predictions provided by machines can have an impact in the world as we know it. It is unlikely that merely stating that the forecasts created by an algorithm are correct will be enough to convince people to utilize the predictions that were generated by the algorithm. Imagine a situation in which a judge who has devoted her whole career to deciding whether or not defendants should be allowed to post bail is notified all of a sudden that a computer can make more accurate predictions than she can. This would be a shocking development for the judge. In this particular situation, it will be necessary to implement a strategy that is successful in order to convince the court to accept the machine projections.

A strategy of this nature should not only teach students that computer forecasts are typically superior to those produced by people, but it should also show students that machine predictions may be inaccurate. This is because computer forecasts are frequently superior to those made by humans because computer forecasts are more accurate. To give you an example, hygiene inspectors had access to data that the machine learning algorithm that was being utilized in the case study did not know about. When a consumer phones the hygiene authorities to complain about an issue at a restaurant, they offer the human inspectors with essential information that the model that is based on online evaluations is unable to incorporate without further processing. The most efficient strategy for resolving this issue is to combine predictions produced

by people with those made by machines in such a way that the combined outcome is superior to the results obtained by each methodology when used on their own.

2.4.3 Concerns Relating to Ethical and Legal Matters

Even if we are successful in overcoming the challenges that have been addressed up to this point, there is still a fundamental question that has to be answered: do we really want to use machine predictions?, the mere fact that we are able to make a prediction about anything does not signify that we are confident in our capacity to depend on that forecast. This is because just because we are capable of making a prediction does not mean that we are certain in our ability to depend on that prediction. In this section, we will only be able to scratch the surface of the ethical and legal challenges that are created by machine predictions; but, public policy makers who wish to apply machine learning should clearly be aware that these concerns exist.

Using computer predictions might lead to bias, which is the most important ethical and legal concern that could occur as a result. There are several possible points of entry through which bias may be introduced into the predictions that machines make. The first strategy involves using machine learning techniques directly to the algorithm. The particular machine learning algorithm that is utilized makes a difference in the outcome of one's efforts to address an issue involving prediction. There is a wide range of capabilities across algorithms, with some being able to model more sophisticated interactions than others. If a predictive model is extremely simple in contrast to the predictive correlations that are present in the data, then the model will underfit the data, which will bring bias into the predictions that the model generates. Only by carefully analyzing the many different forms of machine learning algorithms will it be possible to steer clear of this problem and choose the one that is most ideally suited to the prediction task.

The second way that bias may sneak into the output of machine learning is by selecting an erroneous prediction objective as the goal of the learning process. In the training phase of a machine learning algorithm, the major focus is always on optimizing some statistic that is connected to the training data. This is done in order to achieve the best possible results. For instance, the machine learning approach that was utilized in the cleanliness inspection case study optimized for the receiver operating characteristics area under the curve (AUC). The model that produced the highest AUC was deemed to be the superior model since it was the one that achieved the best results.

However, the metrics that are used to evaluate a machine learning model in the real world are typically different from those that are used when the machine learning algorithm is being trained. This is because the actual world presents a variety of challenges that cannot be simulated in a laboratory setting. Instead of using a complex statistic such as AUC, which is difficult to understand for anyone other than professionals in the field of machine learning, the quality of hygiene forecasts should be evaluated based on how much they improve the distribution of hygiene inspectors to restaurants and how much of an improvement there is in food safety. According to The Economist (2016b), it is essential for those in charge of public policy to sit down with those who specialize in machine learning in order to formulate an optimization goal that accurately reflects the true measure of interest to the greatest extent that is practical. This is important in order to connect the prediction goal of a model with the real-world result that is of interest to the researcher.

The utilization of biased training data is the third potential entry point for bias into predictions. This is due to the fact that biased training data are used. This is the most difficult obstacle to go through since skewed data typically signals that there is also bias in the mechanisms that are already in place that generate the data, and this is the one that requires the greatest effort. Imagine for a second that we are interested in calculating the likelihood that a certain individual would commit a crime at some point in the near or distant future.

We first obtain information on arrests from the criminal justice system as well as other significant data, and then we utilize this information to construct a machine learning model utilizing the data that we gathered. What kinds of problems could be brought on as a result of this? The fact that the records from the criminal justice system only notify us when someone was caught in the process of committing a crime is the root cause of the problem that we have. On the other hand, being caught in the act of committing a crime is not always a reliable predictor of whether or not the crime was really carried out.

For instance, if black people in the past have been the target of discrimination at the hands of the police, such as by over-policing neighborhoods in which a significant portion of the black population resides, and if this discrimination is reflected in the data, then it is highly likely that machine learning models will pick up on this discrimination and bias their predictions against black people there are already accusations that some of the crime prediction software utilized in the United States is

prejudiced and gives projections that are less favorable for blacks than they are for whites. This is despite the fact that blacks are statistically more likely to commit crimes than whites are. It is quite challenging to gather data without introducing some sort of bias into the process. It's possible that, for reasons relating to justice, we don't want elements like race, religion, or gender to have any effect in the judgments that are made based on projections. Eliminating the immediate effects that these elements have on the system is not very difficult; nevertheless, doing it in a comprehensive manner may be extremely complex.

The problem is that these sorts of facts routinely find their way into forecasts, which is the source of the problem. For instance, research consistently demonstrates that one's racial origin and area of residence, in addition to one's gender and choice of work, all have significant connections with one another. If we excluded from a dataset every variable that correlates with another variable that ought not to play a role in the process of prediction, we would usually be left with very little data on which to make our forecasts. This is because removing variables that correlate with other variables that ought not to play a part in the process of prediction may be difficult.

On the other hand, the answer to the opposite of this dilemma can likewise result in bias. It is possible for bias to enter predictions via factors that are not included in the dataset. This might lead to inaccurate results. This is an alternate approach of introducing bias into a dataset by means of the variables that are contained within it. Despite the fact that it is difficult to get all of the necessary data for a forecast, it is possible for inaccurate conclusions to be drawn from predictions that are based on only a portion of the significant components. This can occur even if it is difficult to acquire all of the necessary data.

What steps may be taken to ensure that biased assumptions are not formed by computer programs? There could be some help available if there is more transparency, if additional data can be gathered, and if there is an ongoing investigation into how well machines can anticipate outcomes. If this weren't the case, nobody would have any indication that, for example, some prediction models are more accurate for whites than they are for blacks. They wouldn't be able to tell the difference between the two groups. But even if the accuracy of a model fluctuates depending on which segments of the population it is applied to, could this be sufficient grounds to throw away the model altogether? No, this is not usually the case; however, it does always depend on the options that are provided to the machine predictions. This is not the case in general.

If the predictions that are generated by machines are less likely to be affected by mistake or bias than those that are made by people, then we could decide to go with the machine predictions even if they aren't always accurate to the letter. It is important to recognize the limitations of predictions in order to take the first step toward improving their accuracy, but it is also possible that neither human forecasts nor machine forecasts can ever be completely free of bias. When creating projections based on data that was historically created, public policy makers should, in general, exercise great caution because some parts of the data might be biased against disadvantaged people. This is because certain components of the data could be skewed against marginalized populations. There is a possibility that there is no way to completely prevent prejudice; however, there are methods available to at least lessen the problem.

The fact that many machine learning algorithms are "black boxes" that do not explain the rationale behind their predictions is a significant second significant ethical and legal barrier that machine predictions must meet. This barrier makes the problem of prejudice, which was discussed before, far worse. To fully understand a predictive model, one must first be in the position to do so with specific conditions met. For instance, when there are only a few variables involved, it is straightforward to explain how the linear model arrived at its predictions. The prediction may be obtained by first totaling all of the model's coefficients, followed by multiplying those sums by the values of the variables that are independent of each other. By looking at the sign of a coefficient as well as its absolute value, it is straightforward to determine how much each of the independent variables contributes to a prediction. This may be done by looking at the sign of a coefficient. the difficulty lies in the fact that ever more sophisticated models of machine learning almost inevitably involve non-linearities and intricate interactions between variables. This is a significant hurdle.

In circumstances such as these, there is no one particular mathematical formula that can be used to describe how a prediction is generated. This is because predictions are typically based on a combination of factors. Because of this, it is hard for any individual to grasp such models in their entirety due to the complexity involved. If we want to use these models for decision making, it is abundantly evident that the current circumstance is not the best possible one. According to it may be difficult to create trust in a paradigm that we do not completely know since we do not have all of the information necessary to do so. What steps should be taken when a machine learning model is far too complicated for a person to comprehend? There are two different avenues open to you.

One of them is to get an approximation of how a model would behave in the neighborhood of a certain prediction that we are interested in. researchers came up with a method that explains the predictions of any machine learning model in the region of the data that is surrounding a prediction of interest by fitting a local interpretable model to this region. This methodology was created. This strategy does not explain the predictions that the model makes as a whole, but it does provide some insight into how the model creates the prediction that is being discussed here. The other approach that may be taken is to make use of machine learning models that provide an indication of the level of certainty associated with a forecast. Some models not only produce forecasts, but they also evaluate the chance that each prediction will turn out to be accurate. Even if this does not enable us to explain the predictions that the model generates, at least it gives us an indication of the degree of confidence that the model itself has in a particular forecast.

After then, humans might investigate those forecasts in which the model has just a moderate amount of confidence. If we want more accurate forecasts, perhaps this is the only path open to us. the complexity of machine learning models may simply be the price that we have to pay in order to have high predicted accuracy. According to Lipton (2016), there is a trade-off between the interpretability of a model and the accuracy of the predictions that it makes for a variety of situations that arise in the actual world. It's possible that using straightforward linear connections to depict the complicated realities of our world isn't going to cut it.

There is a good chance that many forecasts made by machines will be more accurate than those made by humans for the simple reason that they represent complicated relationships that are beyond the scope of human comprehension. Even if we don't have a complete understanding of why a model predicts particular outcomes, we may chose to utilize more complicated models as long as the benefits of this complexity outweigh the drawbacks of this complexity in the form of more accurate predictions.

The question of accountability is the final obstacle that has to be overcome in this part. Who is accountable for the forecasts made by machines? The answer to this issue is important from both an ethical and a legal point of view, so keep that in mind while you think about it. Some forecasts, like the one about the bail, can have a significant impact on the lives of the individuals who hear them. There is no cause for concern if a forecast turns out to be accurate; but, just like human predictions, computer predictions can occasionally be inaccurate. Who should be held responsible for an

inaccurate prediction? Who are the people who employ an algorithm? Who was it exactly that designed the model? Who exactly was the source of the data? There is not currently a legally acceptable response that satisfies these issues. The current legal structures are only gradually adapting to the new reality, but first steps toward change are now being taken.

People have the right to relevant information about the logic that is involved in automated decision-making that affects them, which is one of the most far-reaching legislation in this area. This makes the new General Data Protection Regulation from the European Union one of the most far-reaching regulations in this area. In a similar vein, the City of New York has only recently announced the formation of a task group that will investigate the algorithms that the City employs in order to ensure justice, fairness, and accountability. In order for our public systems to be completely prepared to cope with machine predictions, certain further actions need to be taken after these good first efforts that have been taken to address this crucial issue.

After the proper laws have been enacted, the accountability that is associated with machine forecasts does not necessarily have to be lower than that associated with human predictions. People have a tendency to compare technical solutions to perfection, yet it is acceptable for humans to have imperfections that are inherent to being human. But studies and predictions that are based on data may also be examined for biases and other forms of undesirable behavior. Mathematical models can even be more trustworthy and transparent than people, who frequently provide flimsy reasons for their actions rather than in-depth explanations of the causal relationships behind those decisions. It will not be simple for a legal system that is accustomed to dealing with human agency to figure out how to properly establish the legal basis for algorithmic responsibility, but it appears that there is no way to get around this difficulty.

Our conversation about the difficulties and restrictions posed by machine predictions for the formulation of public policy has come to an end. There are some other concerns that should be taken into account in this context, such as privacy, appropriate encryption, and data protection. These are certainly issues that play a part for machine learning because of all the data that is required to train a model. Taking these concerns into account will help ensure that machine learning is used appropriately. Other comparable issues that may benefit from some consideration are concerns regarding gameability and security. These problems, however, are not specific to machine

predictions; rather, they concern data management and policy making systems in general. As a result, we will not go into detail about them here. In spite of this, a public policy maker who wishes to employ machine learning to solve a prediction problem should be aware of both the ones addressed in this paper as well as the ones that are not discussed here.

This study has shown that employing machine learning to address prediction challenges that arise throughout the process of developing public policy is a realistic approach, as proven by the findings of this research. Despite the fact that there are still many roadblocks and constraints, data-based predictive modeling has the ability to change the way that public officials handle prediction challenges. This is the case despite the fact that there are still many impediments. For the very first time in the entirety of human history, large amounts of data on the social behavior of humans on a massive scale are currently accessible. It is no longer statistics agencies with their aggregated datasets that make it possible to get the majority of insights into human behavior; rather, it is the fine-grained data that is produced in today's techno-social systems.

Being part of a culture that places such a strong emphasis on the collection and analysis of data presents a number of opportunities for improvement in the process of developing public policy. In a perfect world, putting a focus on the facts may lead to more rigorous processes that are also more equitable. Since the dawn of time, humans have been obliged to confront the challenge of prediction in the process of formulating public policy; nevertheless, these predictions are virtually never put through a full examination and are nearly never tested to ensure that they are accurate. This could someday change in the era of big data, with machine learning playing a large part in the generation of superior forecasts for the benefit of society as a whole. This would be to everyone's advantage.

However, this does not negate the fact that machine learning is not an ideal answer. To begin, it shouldn't be a solution looking for a problem to solve in the first place. However, there is a possibility that machine learning cannot solve all of the problems associated with making predictions, and individuals in charge of formulating public policy face a far wider range of obstacles than just those associated with making forecasts. In an ideal world, we would combine data-based modeling with human intuition and experience in order to capitalize on the benefits of both techniques while avoiding the drawbacks of each one. This would be the case in order to capitalize on the benefits of both approaches while avoiding the drawbacks of each. It is quite

unlikely that we will be able to depend only on algorithms in the not-too-distant future to tackle challenging prediction challenges such as bail projections. These difficulties require a lot of data and analysis. But in a lot of other areas that aren't as controversial, including violations of hygienic regulations, this might very well be the case very soon.

Unfortunately, there is also a large risk that predictive modeling might be utilized in a manner that is not acceptable. This technology might be utilized by authoritarian governments in order to maintain control over their citizenry. For instance, China is planning to introduce a mandatory social credit system beginning in the year 2020, according to a recent announcement made by the country. Following the installation of the necessary infrastructure, algorithmic assessments of every citizen's behavior will be carried out. the social score that is produced can later be used to "incentivize" behavior that is seen to be "desirable" from the point of view of the government.

The coupling of this social credit system with machine learning has the potential to lead to an increase in the government's level of control over the individuals who make up that government to an undefined degree. Instead of reacting to unpleasant behavior after it has already taken place, the government may decide to adopt preventative steps in order to put a stop to undesirable conduct before it ever occurs. This would be preferable than reacting to undesirable behavior after it has already occurred. The only way to know for certain whether or not machine predictions will be utilized more frequently for use cases that are comparable to those detailed in this study is to wait till the future to find out the answer to this question. Nevertheless, machine learning and big data are not going away any time soon, and individuals who make choices about public policy should be aware of the possibility to utilize data-driven predictive modeling for the good of society.

CHAPTER 3

MACHINE LEARNING IN DATA- DRIVEN PRICING

3.1 INTRODUCTION

In recent years, there has been a notable increase in the total amount of data as a result of the continual development of new technologies and new methods to the gathering of data. This growth may be attributed to the fact that there are now more ways than ever before to collect data. As a direct result of the fact that this information can be gathered from nearly everything and everything that takes place in the globe, virtually anything and everything may be observed and researched as a direct result of this. 2017 research by Wamba and colleagues. there is absolutely no sign of a slowdown in the rate at which the quantity of data is growing. In point of fact, forecasts presented in Figure 3.1 indicate that the volume of data will nearly double over the course of the next two years and increase by a factor of five over the course of the subsequent five years.

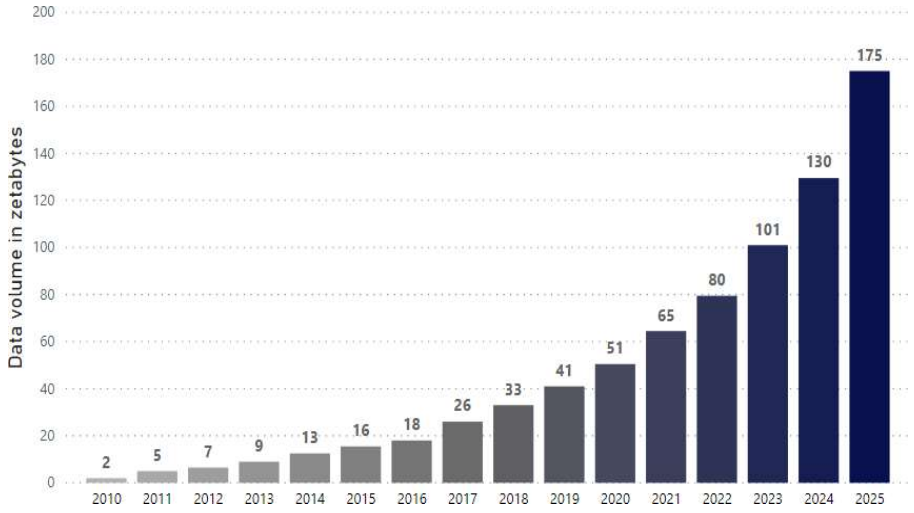


Figure 3.1 Volumes of data in zetabytes (Statista 2020)

source: *machine learning in data- driven pricing, data collection and processing through by Amar Singh (2018)*

Concurrently, the process of globalization is accelerating, which is making the world a place that is more linked than ever before. this is having the direct effect of heightening

the level of competitiveness in nearly every industry. Because of this, it is the obligation of businesses to encourage creative thinking and to make efficient use of all of their resources, including the data that they collect. By utilizing information obtained from a wide variety of sources, businesses may improve their capacity for making decisions and their ability to effectively compete in their respective markets. As a direct result of this change, it is even plausible to state that data is becoming one of the key markers of whether or not a company will continue to be relevant. This is a possibility since data is becoming increasingly important.

The pricing of goods and services is one of the most productive use for this data. Pricing is one of the most significant parts of competition since it controls both the consumption of a product by customers as well as the income of a business at the same time. Data-driven pricing is not a newly established idea; nonetheless, its application has been extremely common in recent years. As a consequence of the increasing quantity of data as well as the simplicity with which it can be accessed, businesses are now in a position to design pricing systems that take into consideration a range of critical parameters at the same time. These pricing systems can, in the best of all conceivable worlds, be completely dynamic, which indicates that the price is continually customized in accordance with the scenario and the consumer.

For the full potential of dynamic pricing to be realized, it is necessary to have systems that are able to automatically evaluate and integrate the many forms of information that have been discussed above. It's possible that one answer to this problem is to make advantage of machine learning. one of the most popular research topics at the moment is machine learning, and new applications for it are being invented all the time. Additionally, it is a technology that many companies are interested in utilizing, despite the fact that they do not have a complete understanding of what it includes or the potential applications that it may have. One example of the enormous potential that this technology holds is the prospect of completely dynamic pricing that is determined by data. This is only one example of the enormous potential that this technology possesses. When it comes to pricing, machine learning may be put to use for a variety of purposes, including the segmentation of clients, the staging of prices, the forecasting of demand, and the automation of a variety of functions. There are several authors in total.

3.2 PRICING CONSIDERATIONS AFFECTED BY FACTORS

Decisions about pricing have a direct bearing on a company's ability to earn a profit and can have a significant amount of that influence. Given this information, it is plainly

clear that decisions about price play a very significant role in the operation of a corporation. According to research, a gain of around nine percent in operational profits may be attributed to an increase in price of one percent. This is based on the assumption that there will be no reduction in volume. Despite this, the great majority of companies devote an insignificant amount of attention to their pricing strategies and make no attempt to identify the deal that represents the best possible value. According to the findings of a survey that was carried out in 2019, as many as 78 percent of firms are aware that their pricing structures have the potential to be significantly improved and be based on a greater number of dimensions. This is most likely because standard human pricing methods make it incredibly difficult to recognize new price patterns that can bring additional value. This is one of the reasons why automated pricing systems are becoming increasingly popular.

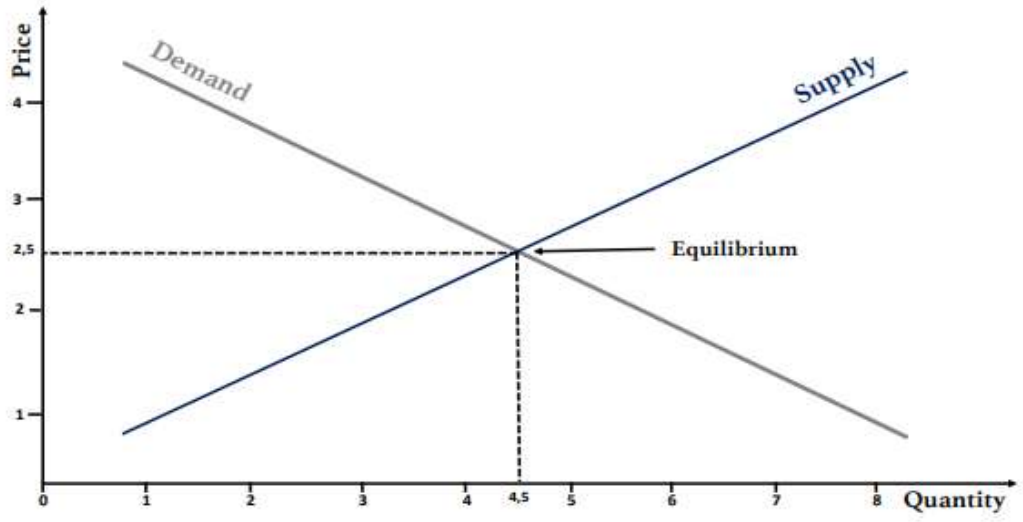


Figure 3.2 Law of supply and demand

source: *machine learning in data- driven pricing, data collection and processing through by Amar Singh (2018)*

Keeping up with the intricacy of various pricing variables that are always shifting for hundreds of different products is nearly impossible, especially for large firms. This is especially true for businesses that operate only online. a report from 2014 by Baker et al. Price is frequently established by taking a look at supply and demand (Figure 3.2), and for the vast majority of products and services, it is a crucial factor in determining

whether or not an order will be made. Figure 2: Price is frequently determined by taking a look at supply and demand. This suggests that when there is a high demand for a certain product but there is a limited supply, the price may be set relatively high, and when there is a low supply but a high demand, the price can be set relatively low. The converse is true when there is a high demand but a low supply. The term "equilibrium" refers to the point in time at which the supply of a good or service and the demand for it are equal, and it is at this point that prices are most usually determined.

The process of assigning a price to anything may be difficult due to the fact that supply and demand are influenced by a wide number of factors, all of which are in a state of continuous transition. When there are multiple varieties of the same product, each of which has its own supply and demand, it may be rather difficult to pinpoint the point at which supply and demand are in equilibrium.. a. The factor that determines how much anything is worth is the price of the object when it is broken down into its component parts. In addition to this, it is a factor that defines the degree to which the product is competitive, lucrative, and where it is positioned in the market. In other words, it is a factor that decides where the product stands in the market. businesses have to be very specific when determining prices and strategies for pricing products and services. This is owing to the fact that pricing may be seen as a measurement of a variety of various things, including a wide variety of other things. Customers, the company itself, competing companies, and collaborative partners are the four key categories that may be used to identify the components that impact pricing (Figure 3.3). Each of these four categories has the potential to have an effect on the price.

Figure 3.3 Factors affecting price

source: *machine learning in data- driven pricing, data collection and processing through by Amar Singh (2018)*

Naturally, one of the factors that affects price is the company. Costs, goals, and company strategy are three factors that have a significant impact on the ultimate pricing. Taking into account one's costs is of utmost significance, as profit is created

by subtracting one's expenditures from one's selling price. Because the profit margins on different items might vary quite a bit from one another, it is extremely vital to accurately evaluate the expenses of each product. There is also a significant relationship between the pricing and the company's goals and strategies. If a corporation, for instance, has the goal of building a high-quality brand, then it is quite probable that the pricing would reflect that goal as well. This type of pricing is known as value-based pricing. Pricing is determined by how useful consumers believe the good or service to be, which is the core tenet of the value-based pricing strategy. A pricing strategy that is solely focused on manufacturing costs is the polar opposite of this. This kind of pricing is known as cost-based pricing, and it is utilized frequently by businesses that have high production volumes but poor profit margins in an effort to achieve cost leadership.

Given that pricing is one of the most important variables in winning and qualifying orders, it should come as no surprise that competition has a significant impact on prices. Price becomes an even more essential factor when there are several rivals selling items that are identical to the one being sold. Because of this, it is essential to do ongoing price comparisons with other competitors. Additionally, businesses are required to evaluate the quality of their services in comparison to the quality of other comparable products or alternatives.

Customers have varying characteristics depending on factors such as their geographical region and socioeconomic standing. It is essential to determine the qualities of the consumer, such as the maximum amount they are ready to pay and whether or not they are seeking for a high-quality brand or just the alternative with the lowest price. (Brandt 2018, etc.) Because customers are the only stakeholders who can actually result in revenues, conducting research on them is of utmost significance. In addition to this, the consumption habits of customers are always shifting, which means that the study of their behavior needs to be highly dynamic. cited in:

Collaborators, especially in terms of their pricing, are a factor in the creation of prices. For instance, the margins of collaborators need to be made large enough to ensure good functionality, but at the same time, these costs need to be factored into the company's own prices. This is how prices are influenced when a company sells products or services through third parties. In a similar vein, the suppliers of raw materials and the subcontractors have an effect on pricing since any changes in their prices are immediately reflected in the costs incurred by the firm.

The product's inherent qualities also have a significant bearing on how changes in price affect demand for the product. The responsiveness of the market to changes in the price that are caused by nothing else but the price change is measured by price elasticity. It generally relies on the significance of the product or service, as well as the qualities associated with it, such as whether it is a need or a luxury. Products that have a high price elasticity are more sensitive to changes in price, and even a slight rise in price can drastically affect demand for those products. This also indicates that as prices go down, there will be a big rise in the amount of demand.

To get a product's price elasticity, just divide the percentage of change in the amount demanded by the percentage of change in the price. It is used to anticipate how different price adjustments would effect demand, as well as for the purpose of attempting to discover the optimum feasible price that will result in the most profits. it is possible to determine it with relative ease by using data on past sales and prices. As a result, it also plays an important part in the process of data-driven pricing.

3.3 THE PROCESS OF SETTING PRICES

it is essential to establish a distinct process and infrastructure for pricing that takes into consideration the aforementioned difficulties. This is because the elements that affect price can change quite a bit and are always shifting. the general pricing process (Figure 3.4) is comprised of three primary stages: fundamental analysis, strategy matching, and determining the final price. This model represents the pricing process in its most general context and does not, for instance, take into consideration the connections between the various stages of the process. The formulation of a pricing that takes into consideration a number of criteria, both internal and external, is the purpose of this procedure. Although the method of pricing a product or service may be different based on the pricing strategy, product, or service in question, the fundamental idea would not change.

The procedure for pricing starts out with a step known as the preliminary analysis phase. At this stage, a number of the components that were covered in chapter 2.1 are being investigated in more detail. An examination of the aspects of the operational environment that have an impact on demand, cost, and competition are to be carried out as part of this work. These aspects include everything that has an influence on demand and costs, as well as things like the price elasticity of demand and the quality of service in contrast to that of other comparable goods or substitutes. At this stage in

the process, it is also quite important to carry out an analysis of the factors that are associated with the product by itself. This entails looking at the product's life cycle from beginning to end, as well as estimating how much money will be made and how much it will cost during the course of the cycle. The study that Grimmer et al. published in 2015

Figure 3.4 Pricing process pyramid (adapted: Järvenpää & Partanen 2010, p. 199)

source: *machine learning in data- driven pricing, data collection and processing through by Amar Singh (2018)*

During the stage that is known as "strategy matching," the company's strategy and overall goals are taken into consideration when it comes to pricing. These should be in harmony with one another since, depending on the company's aims, pricing might take many different shapes and, as a result, should accurately represent those differences. when a company's primary objective is to achieve the greatest potential rate of growth, the pricing strategy it employs will appear significantly different than when the company's primary objective is, for example, to realize the most possible profits. The major objective of this stage is to identify certain criteria that will be utilized in further talks on data-driven pricing systems and machine learning algorithms. These criteria will be used in later debate. For instance, the price may be capped so that it does not go below a certain amount or rise over a particular threshold, depending on the overall strategy and objectives of the company. On the other hand, the selection of dynamic pricing that is driven by data is represented at this stage because it is also a price decision that is made strategically. This is because the selection of dynamic pricing is also a price decision.

The last stage of the process is referred to as "setting the price," and it is at this stage that the total price of the product is determined. In order to achieve effective pricing, it is vital to have a full understanding of customers and their viewpoints on the price, in addition to any price modifications. Gorodnichenko and Talavera published their findings in 2017 The price or value that a customer associates with a product is not always the same as the price that the customer actually pays for that goods. This is the case in many instances. If companies are aware of the factors that customers consider when making price comparisons, they will be able to establish pricing in a manner that is noticeably more competitive. Before agreeing on a price for a product or service, it is essential to do market research in order to identify what exactly consumers want and how much they are willing to pay for it. This information can then be used to inform pricing decisions. (2008) Supposedly, in accordance with Interhuber.

On a more operational level, the process of pricing may be split down into numerous phases, and the progression of these stages might go in either way. Pricing is a process that needs to be ongoing and iterative so that it can account for the ever-changing nature of the factors that impact demand, as well as the prices that competitors are offering for their products. Pricing is in point of fact an ongoing practice that is practically duplicated on a continuous basis. 2017 research carried out by Fisher et al. As a result of this, price should always be up to negotiation.

3.4 PREREQUISITES FOR DYNAMIC PRICING

Data is the most critical precondition for dynamic data-driven pricing since, without it, there is no empirical verification of the operational environment. If there is no empirical verification, dynamic data-driven pricing cannot occur. Given this, having access to the data is the most crucial need that must be met. In the absence of appropriate data, it is impossible to put into practice a dynamic pricing model that is driven by the data. As was discussed in the chapter that came before this one, coming to a determination on the cost of an item is often the result of carrying out a great deal of research. On the other hand, the vast majority of businesses do not include any of their data into the pricing process in any manner, shape, or form. a paper that was published in 2014 by Baker et al. This may sometimes lead to a scenario in which identical products of the same sort are offered for sale at prices that are drastically different from one another, even to persons who are a part of the same client base as one another. If the profit margins aren't high enough, it's possible that some of the items will have to be sold at a considerable loss in order to keep the firm alive. This would be the case in the event

that the profit margins aren't high enough. For this reason, it is of the utmost importance for businesses to initiate the utilization of data analytics for pricing at an early stage in order to optimize their revenues.

Data is defined as any collection of information, including words, numbers, measurements, observations, or even simple descriptions of objects. This definition encompasses all types of information, including but not limited to descriptions of things. As was said before, advancements in technology have made it feasible to collect data from virtually any location. This was formerly impossible. The data can be organized in a wide variety of distinct ways; nevertheless, it can often be broken down into the three categories shown in Figure 3.5: structured data, unstructured data, and semi-structured data. In all, there are a few different authors. Authors from a variety of backgrounds.

Figure 3.5 Different data types

source: *machine learning in data- driven pricing, data collection and processing through by Amar Singh (2018)*

Structured data is the easiest to analyze since it is given in a consistent format that consists of columns and rows. This makes structured data the most preferable type of data to work with. Structured Query Language, or SQL for short, is a computer language that is used for relational databases. SQL is also a more common name for Structured Query Language. Managing structured data often requires the use of SQL, which is a database query language. Due to the fact that information is in a normalized form, it is a great deal easier to study and assess. Spreadsheets created using Excel and database tables are two common places to look for this particular sort of information. There are many different sorts of structured data, some examples of which are sales data, address details, demographic information, and location data from smart devices. Authors from a variety of backgrounds.

It is not feasible to store unstructured data in rows and columns because unstructured data does not have an associated data model. Websites, audio files, pictures, the content of social media platforms, PDF files, and textual survey responses are all instances of what consider to be examples of unstructured data. Since the data does not have a consistent structure to begin with, it is rather tough to handle and analyze unstructured data. This may make the process feel quite overwhelming. Because of this, the vast majority of companies ignore the process of doing analysis on unstructured data and do not make effective use of it. The findings of Baars and Kemper are as follows:

The mix of structured and unstructured data is referred to as semi-structured data, and it is the third category of data. In layman's words, it may be characterized as a blend of structured and unstructured data. Even while it does have some structured qualities, semi-structured data does not correlate to a structure that might be kept in a relational database even if it does have some structured properties. This is because semi-structured data includes certain ordered characteristics that make it simpler to arrange. A very excellent example of this would be an email, the content of which is entirely unstructured, but which does include some structured data, such as the sender's email address and the current time stamp. This is a very good demonstration of what we are talking about.

If it is not processed first, raw data, in whatever form it may take, is generally nothing more than a collection of numbers or digits and does not disclose anything meaningful about the world. This is the case regardless of the quantity of data that is at our disposal. Because there is such a large quantity of data, not all of it can be employed, and, to tell you the truth, the vast majority of it is not valuable. Therefore, the underlying challenge is not the absence of data; rather, it is the way of identifying and evaluating significant data that can aid in making pricing decisions. This can be seen as a positive development, as it suggests that the true difficulty is not the absence of data.

The processing of the raw data and the organization of that data into a format that can be understood more easily are both necessary steps in the process of transforming raw data into information. The distribution format options that are appropriate for this kind of information include things like reports and dashboards, for example. the term "insight" refers to information that not only holds an added value but also possesses the capacity to aid in the process of decision-making. In a word, insights are reasonable inferences that are made from the information that is supplied by the data. The act of analyzing vast volumes of raw data can result.

In the discovery of insightful tidbits of information that are useful in solving problems and making decisions. According to the utilization of these insights offers firms with distinct and practical rules or principles that make decision making less complicated and more rational. For instance, the data may indicate that a certain group of customers has a habit of making a much higher volume of purchases at the beginning of each new calendar month. This specific group of customers often has a propensity to have more cash on hand at the beginning of the month, which is the reason why the majority of their shopping is finished around this time of the month. The conclusion that can be derived from this is that because of this tendency, the majority of their shopping is completed around this time of the month. These "aha" moments can also be used to produce more optimum pricing if they are exploited properly. In light of the example given above, the price may be changed to make the demand more evenly distributed, or the profit margins might be expanded if the demand is particularly high. Both of these options are possible.

The fact that not all data is created equal is the factor that makes gaining these insights a far more difficult task than it would otherwise be. Because the results of analytics are only as accurate as the data that was provided, companies need to exercise extreme caution with regard to the type of data that they are utilizing in their operations. If the data is incorrect in any manner, reviewing it will also result in erroneous reporting and bad decision-making. This is because the flawed data will be evaluated. Throughout the course of history, there have been several occasions in which the exploitation of wrong or defective data has resulted in huge challenges.

3.5 MACHINE LEARNING

Structured data, which is presented in a manner that is consistent and consists of columns and rows, is said to be the data that is the simplest to analyze, according to Baars and Kemper (2008). Because of this, structured data is the form of data that is recommended to deal with the most. Structured Query Language, sometimes known as SQL for short, is a type of computer language that is utilized for relational database management. Structured Query Language is also known by its more frequent abbreviation, SQL. SQL, which is short for structured query language, is typically utilized while managing structured data because it is a database query language. Studying and evaluating the information is made a great deal simpler as a result of the fact that it has been normalized and is in a standardized format. Excel spreadsheets and database tables are two frequent locations to check for this specific kind of information.

Both of these kinds of sites may be found online. There is a wide variety of structured data, some examples of which include sales data, address details, demographic information, and location data from smart devices. There are also several more types of structured data. Authors coming from a range of different backgrounds.

Because unstructured data does not have an associated data model, it is not possible to store unstructured data in rows and columns. This makes the storage of unstructured data in rows and columns impossible. some examples of unstructured data include websites, audio files, photographs, the content of social media platforms, PDF files, and textual survey replies. All of these types of data may be found on the internet. It is not easy to handle and evaluate unstructured data since the data does not have a consistent structure to begin with. This makes the task more difficult. Because of this, the procedure might feel quite daunting to you. Because of this, the great majority of businesses overlook the process of performing analysis on unstructured data and do not effectively utilize it as a result. The following sums up what Baars and Kemper found in their research:

The third kind of data consists of semi-structured data, which is a combination of structured and unstructured data. This type of data is referred to as semi-structured data. It is possible to describe it as a combination of organized and unstructured data using language that is more accessible to the general public. Even while it does have some structured characteristics, semi-structured data does not correspond to a structure that might be stored in a relational database even if it does have some structured qualities. This is because semi-structured data does have some structured features. This is due to the fact that semi-structured data consists of specific ordered properties that make it easier to organize. One of the best illustrations of this would be an email, the body of which is completely unstructured but which does contain some structured data, such as the sender's email address and the current time stamp. This is a fantastic illustration. This serves as an excellent illustration of the topic that we have been discussing.

Raw data, in whatever form it may exist, is typically nothing more than a collection of numbers or digits and does not reveal anything relevant about the world if it is not processed first. This is true regardless of the shape that raw data may take. It makes no difference how much information we have access to; this will always be the case. Because there is such a great quantity of data, not all of it can be utilized, and, to tell you the truth, the vast majority of it is not valuable. Therefore, the fundamental obstacle is not the lack of data; rather.

It is the method of discovering and analyzing key data that may assist in making pricing decisions. This is because the absence of data is not the root of the problem. This might be considered as a positive development since it implies that the real challenge is not the lack of data. This could be seen as a positive development. In order to turn raw data into information, there are a few phases that must first be completed, including the processing of the raw data and the organizing of that data into a format that is simpler to comprehend. Both of these steps are required for the process. Things like reports and dashboards, for instance, are examples of the kinds of distribution formats that are suitable for the sort of information being discussed here. The investigation that Baye and his coworkers carried out in 2007 the term "insight" refers to knowledge that not only has an additional benefit but also has the potential to be of assistance in the process of making decisions. In a nutshell, insights are logical conclusions that may be drawn from the information that is provided by the data.

The process of analyzing large amounts of raw data can lead to the discovery of nuggets of information that are insightful and can be used to help solve issues and make choices. the implementation of these insights provides businesses with clear and actionable rules or principles that simplify and rationalize the decision-making process. For instance, the data may suggest that a certain category of clients has a pattern of making a much greater number of purchases right at the start of each and every brand-new month of the calendar year. Due to the fact that this particular category of consumers has a tendency to have more cash on hand at the beginning of the month, the majority of their purchasing is completed around this time of the month.

Because of this propensity, the vast majority of their shopping is finished around this time of the month, which is the conclusion that can be drawn from the information shown here. If they are leveraged in the right way, these "aha" moments have the potential to potentially be used to provide more optimal pricing. In light of the previous illustration, it is possible to adjust the price in order to make the demand more evenly distributed, or the profit margins may be increased if the demand is very strong. Both of these options are worth considering. Both of these alternatives are open to consideration. in the most ideal scenario, the process of locating these significant ideas would be completely automated.

The fact that not all data is produced equal is the component that makes acquiring these insights a significantly more challenging endeavor than it otherwise would be if it weren't for the fact that not all data is made equal. Because the results of analytics are

only as accurate as the data that was supplied, businesses need to exercise extra caution with regard to the kind of data that they are employing in their operations. This is because the accuracy of the results of analytics is only as good as the data that was provided. Reviewing the data might lead to inaccurate reporting as well as poor decision-making if the data is flawed in any way. This is due to the fact that the inaccurate data will be examined. Throughout the course of history, there have been a number of instances in which the utilization of inaccurate or deficient data has resulted in significant difficulties. In each of these instances, the consequences were significant.

According to what Haug and his colleagues indicated in 2011, Machine Learning (ML) is a collection of separate algorithms and statistical models that are used in the area of data science to automate, predict, and solve a broad variety of problems and processes, ML is a collection of distinct algorithms and statistical models. In this chapter, the essential notions of machine learning are deconstructed and analyzed, particularly from the point of view of dynamic pricing. In general, machine learning algorithms are the ones to thank for the creation of mathematical models through the utilization of training data. After that, these models are utilized to produce automatic judgements and forecasts without being expressly trained to carry out any tasks whatsoever. The foundation of machine learning is mathematics that is not unduly difficult and can be comprehended in a rational manner, despite the fact that machine learning is a very demanding topic of study. there are three basic classifications that may be used to machine learning: supervised learning, unsupervised learning, and reinforcement learning. These categories can be shown in Figure 3.6.

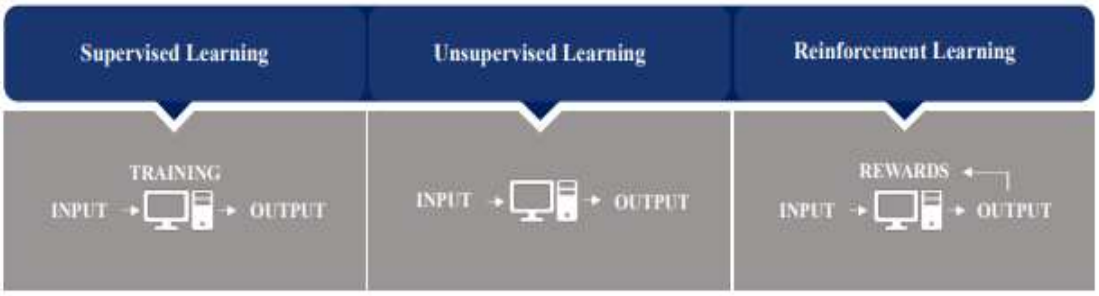


Figure 3.6 Machine learning types

source: *machine learning in data- driven pricing, data collection and processing through by Amar Singh (2018)*

When both the output (o) and the input (i) of the learning process are already known, this type of learning is referred to as "supervised learning." On the basis of these pairs of values (i and o), the algorithm for machine learning discovers which individual traits and combination of features produce a certain outcome. When these algorithms have been trained with a considerable quantity of data, they are able to generate incredibly accurate forecasting models, in which an output may be predicted based on any input that is supplied. This is possible because these models are able to develop extremely accurate prediction models. When it comes to pricing, this information may be used in a variety of ways, for example, to create predictions about the ways in which changes in price affect customer demand. It is possible to achieve this objective by supplying the model with output and input pairs that consist of price and the demand that corresponds to it. Because of this, the model is able to develop the capacity to gain the ability to grasp the link that exists between them.

Unsupervised learning is a kind of education in which the learner is only provided with information about the input. This approach makes no attempt to foresee or guarantee a certain result in any circumstance. Finding patterns and structures within the data is the goal of this approach, which aims to do this through providing an understanding of the data. One area of application that might profit from utilizing unsupervised learning is the process of customer segmentation. 2019 if we are to believe Vickery This is achieved by supplying the algorithm with data that encompasses a variety of dimensions of the consumer. Using the information that is presented in this section, it is possible to divide customers into a number of distinct subgroups. This is a really big step forward in terms of pricing, as it enables algorithms to zero in on certain customer groups and provide the most advantageous price for those segments.

Learning by reinforcement is a strategy in which the goal of the algorithm is not specified in advance. It is the one that has the most sophisticated technology, but at the same time, it is also the one that is the most difficult to understand and make sense of it is involved in the process of continually making decisions in an unknown environment and plays a role in that process. The data serves as the input for the algorithm, and in the beginning, the algorithm generates output and data at random for it. The merit of the outcomes that the algorithm generates will decide the amount of compensation that it receives. The Reinforcement algorithm is programmed to maximize the quantity of rewards it receives; as a consequence, it is continually changing to its surroundings in real time as it learns what sorts of actions result in

positive and negative reinforcement. This ensures that it obtains the greatest possible benefit from the algorithm. 2009, based on information provided by Norvig and Russell This is especially important in terms of pricing, as it has been mentioned several times before that the factors that drive price and demand are always moving. As a result, pricing may more accurately reflect actual market conditions.

One of the most important advantages of reinforcement learning is that it does not require a model of the environment to be set in advance. This is possibly the most important advantage. These interstate links are not taught, but rather learned via a process of dynamic interaction with the world that is around them. In situations in which both the external and the internal components are susceptible to rapid change, the application of reinforcement learning is particularly useful. This is because it helps the learner better adapt to their environment. These algorithms are able to dynamically optimize the pricing by taking into consideration a variety of criteria. Some examples of these elements include the length of time that is still available and the number of vacant seats on a flight.

Deep learning is a type of machine learning that is far more sophisticated than the models that we have been talking about up until this point. It is a subset of machine learning that can have as some of its components supervised learning, unsupervised learning, or reinforcement learning. Each of these types of learning is discussed more below. Due to the fact that it may be utilized in such a broad variety of different ways, it can be used to solve such a vast variety of different types of problems. Deep learning is a mathematical function that, in its most fundamental form, aims to copy the working of the human brain and to arrive at judgements in the same way. This imitating and arriving at conclusions in the same way is called deep learning.

The fundamental components of neural networks are referred to as layers. The input layer, the hidden layer, and the output layer are the three unique categories that may be used to classify these layers (Figure 3.7). These layers can be divided down into these categories. The initial layer of the neural network is responsible for processing the data that was input, and the data that was processed is then sent to the other layers of the network as output. The second layer, often known as the hidden layer, is responsible for processing the data that has been brought up from the lower layers. After then, this layer will produce a new output that is determined by the data from the layers below it. This procedure is repeated as many times as necessary until the desired end result, such as the final price, is reached. Due to the fact that only 20% of data is structured, the fact

that the use of deep learning and neural networks has made it feasible for some to use unstructured data is a good thing from the standpoint of pricing. This is a really encouraging new turn of events.

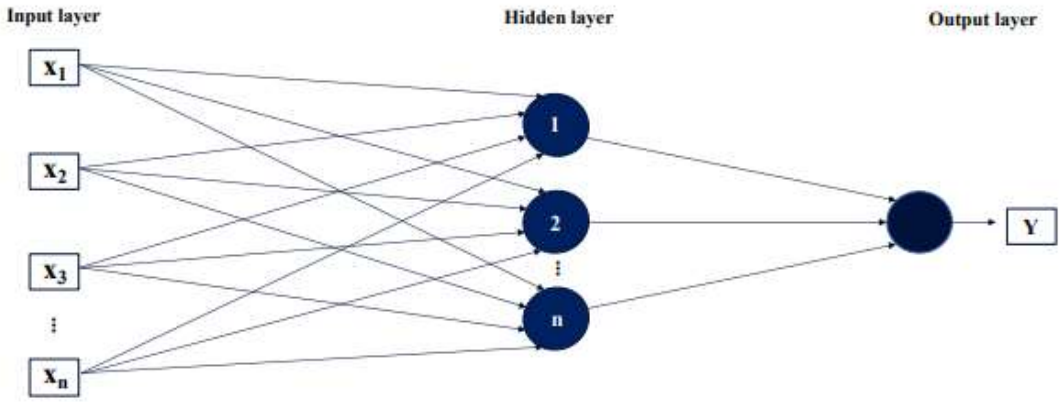


Figure 3.7 Neural network (adapted: Yiu 2019)

source: *machine learning in data- driven pricing, data collection and processing through by Amar Singh (2018)*

One component that is shared by all different types of machine learning is the concept that the results generated by these algorithms are only as trustworthy as the data with which they are presented. This is one of the aspects that is considered to be a defining characteristic of machine learning. Second, an increase in the total quantity of data that each of these models gets leads to an improvement in the accuracy of those models. Predictions that are based on insufficient data are prone to inaccuracy because they are obtained from only a restricted cross-section of the real world. In other words, the real world is not sufficiently represented in the data. Because of this, the development of any form of machine learning algorithm is also a process that lays a focus on the relevance of the gathering, storing, and processing of data.

3.5.1 Data storing and collection

Companies are able to more effectively forecast the future demand and interests of their customers when they have a bigger quantity of data not just about their customers, but also about their competitors. This is because businesses are able to compare the data they collect on both their customers and their competitors. Despite the massive volumes

of data that are always being transferred around, it may be a great challenge to process valuable data from such a big number of data. This is due to the fact that there is so much data. It is challenging to establish which data should be prioritized for examination due to the vast number of data that is currently accessible. Because purchasing and storing data costs money, companies have a responsibility to carefully decide what information to maintain and how much of it to keep. This creates a problem not just with the quality of the product but also with the amount of it.

Relational databases are excellent for storing structured data, but they are unable to process unstructured or semi-structured data. However, relational databases are amazing for storing structured data. Relational databases are an excellent option for storing data that has been organized. This presents a challenge since, according to AWS 2020, in order to get the most out of the data that is gathered, it should be consolidated in a location where almost all of the data can be analyzed as a whole rather than in its component parts. Since this is not the case, it is difficult to get the most out of the data that is collected. Lakes of data have emerged as the answer to this problem during the past few years. A data lake is a repository of data that stores the data in the format in which it was originally produced. This type of repository stores data in its native format. because they are able to manage many kinds of data, they are a good alternative for maintaining all of a company's data in a one spot because they are so versatile.

Companies that implement data lakes offer themselves with a centralized place from which all of their data can be retrieved and analyzed. This location is known as the "data lake." Because of this breakthrough, data scientists are now able to make use of this lake in order to conduct more comprehensive analytics and machine learning. Despite the fact that data lakes are effective for storing and analyzing vast amounts of data, they have been the target of a substantial amount of criticism. One of the primary reasons for this is that data lakes are not distributed. This criticism is a result of the fact that some enterprises simply toss everything into the lake in the expectation that they would benefit in some manner from it. As a result of this, companies lose knowledge of the data that exists, where it is stored, and the form that it takes. It is recommended that this data be saved in these lakes so that we can get the most use out of it; however, we should be careful to only save the data that is really necessary there.

Before price decisions can be made based on the data, it is required to first make the data accessible. In this context, "data collection" refers to the process of amassing and analyzing a variety of different types of information. There are a number of methods

that may be used to collect this data; however, the method that frequently results in the conclusions that are the most accurate is the one that combines a number of different kinds of information. This information may be obtained by the firm in a variety of ways, including directly from customers, from other companies, or even, for example, through the company's very own manufacturing lines. In addition to gathering data on their own, businesses have the option of obtaining data through the use of professional third-party data platforms that collect data from a range of sources. These platforms gather information from a wide range of different sources.

companies are able to alter their pricing to the proper level if they collect data, such as that which is offered by their nearest rivals. This data may be gathered by surveying customers and seeing how other firms price their products and services. However, there are cases in which companies need to be more creative in order to collect the essential information. The simplest technique for accomplishing this work is to browse through the websites of one's rivals. This data is, at best, easily accessible through APIs (Application User Interfaces), which indicates that gaining access to it does not call for a considerable amount of work (Mon 2018).

Internet of Things (IoT)-based solutions make it easier to evaluate internal challenges and dynamically forecast aspects such as production costs, for example. The Internet of Things relies on sensors, which are often relatively small devices that may be utilized to monitor a wide variety of phenomena in real time. (2016) They may, for example, be fitted with sensors to detect acceleration, pressure, temperature, or humidity. This is in accordance with Han et al. the data that is produced by sensors that are connected to the Internet of Things is often relatively easy to upload and collect into nearly any form of data warehouse. Using these sensors has a lot of advantages, and this is just one of them. They are able to gather data from the production lines, which can subsequently be used for a variety of purposes, including calculating the costs of any future maintenance that may be required.

state that there are three ways to collect data about customers: directly questioning consumers, indirectly monitoring customer behavior, or adding data from other sources of customer data to the data that the company already possesses regarding customers. Information gained through data marketplaces and open source information are two other kinds of data. Both of these types of information are freely available to anybody who has an interest in obtaining them. Customers' online activity and click data may be gleaned from websites and social media pages, which are two of the most prominent

and effective sources of such information. (Deshpande 2019). While the data received from social media may reveal extremely sensitive information about, for example, a person's interests or work, the click-data data that is obtained from websites is helpful in predicting consumer behavior. There is a broad selection of tools available on the market, such as Google Analytics and Hotjar, that can automatically collect this information for you. You may take use of this. One is also able to develop data specific to one's own place by making use of the same technology.

It is possible to keep track of and apply the information from earlier sales in order to generate forecasts on future purchases. It is possible to monitor this data simply at the level of an entire company or an item; but, in order to track the transactions that are performed by individual consumers, some kind of identification for the purchase is required. This identification might be achieved, for example, by demanding the use of a customer profile at an online store or the usage of a loyalty card at a brick-and-mortar store. Another option would be to ask for a driver's license number or other identifying information. Asking for the holder's driver's license number is an additional approach that may be taken to accomplish this identification. Be advised that the process of purchase may be halted for some customers if you insist that they provide authentication documents before making a purchase. Because of this, it is essential to give serious consideration to the question of whether the advantages of authentication exceed the possibility of a reduction in sales.

Figure 3.8 Customer profiles (adapted: Doppsen 2020; Deshpande 2019)

source: *machine learning in data- driven pricing, data collection and processing through by Amar Singh (2018)*

The construction of personalized customer profiles and the provision of tailored pricing structures to unique clientele or customer groups are both capabilities that may be made available to businesses as a result of the collecting of data from a wide range of sources

by those firms (Figure 3.8). The establishment of pricing is only one of many possible applications for these profiles; rather, they may also be used to accomplish a wide variety of other objectives, such as the formulation of targeted marketing and communication strategies. (Deshpande 2019). Utilizing machine learning, which will result in the profiles being more dynamic and accurate, is one method that may be used to improve them even more.

In addition to this, it is of the utmost importance to be aware that data that is no longer relevant should not be destroyed and that there is no time restriction on the preservation of this data. When looking at time series, and especially when looking at structured data, the structure of the data will typically remain pretty consistent. This is especially true when looking at structured data. This makes it possible to make comparisons between data from the past and data from more recent times. It is necessary that previous data be retained in order to guarantee the accuracy of these forecasts. The majority of forecasting and machine learning algorithms are produced via the study of historical data; thus, it is essential that historical data be preserved.

3.6 DYNAMIC PRICING

3.6.1 Data-driven pricing

The concept of a situation being "data-driven" refers to one in which the majority of decisions and actions are directed by the analysis of data. When there is a significant amount of data to process, this approach operates at its most productive level of performance. This is The Wall Street 2020. It is possible for a company's data to serve as the basis for almost all of its activities when the company has access to an adequate amount of information that has been obtained from the relevant sources. The availability of a big amount of data gives a fantastic insight into the world, allowing businesses to make improvements to their pricing decisions, and enables demand forecasting for the future. As a direct result of the recent explosion in the quantities of data and the development of technology for assessing it, many firms have adopted data-driven decision making as an important component of their day-to-day strategic and operational operations. This is a direct outcome of the recent explosion in the amounts of data.

As soon as the relevant data is made accessible, it is possible to include it into the process of pricing the product. The process of pricing is one that consists of a number

of different procedures, one of which is the examination of a range of different elements, as was discussed in chapter 2.2. Given that it is feasible for data to deliver answers that are devoid of ambiguity, it is plainly obvious that it should be employed at every point of the pricing process. A report from 2014 by Baker et al. The phrase "data-driven pricing" refers to a way of setting prices in which the bulk of the stages in the pricing process are determined by data rather than by person experience, "gut feel," or traditional business methods. This pricing method has been increasingly popular in recent years.

Data-driven pricing seeks to achieve the goal of presenting the relevant price to the suitable clients at the appropriate moment. To cite: Because of this, it is feasible to boost profitability and demand, in addition to delivering an improved experience to clients. Thanks to contemporary methods of data collection and analysis, retailers are now able to take into account a wide range of factors, including the competition, the season, ERP-data, currency rates, operational cost, demand, and even corporate objectives. Even the weather may play a role in these considerations. These components offer helpful insights on both the external world and the internal constraints.

In business-to-customer (B2C) marketplaces, such as retail, there are several competitors that offer exactly the same things, which further heightens the amount of rivalry that occurs in these markets. This is because retail is a business-to-customer (B2C) marketplace. In addition, customers may easily compare prices, and there are even websites that have been designed specifically to carry out this cost comparison job for customers in an automated fashion. As a direct result of this, numerous retail chains have even changed to basing their prices almost completely on the information supplied by their competitors. This transition has occurred as a direct consequence of this. However, basing pricing only on those of competitors is a very risky strategy since it does not take into consideration the company's own expenses or the strategic goals it has set for itself. This leaves the firm vulnerable to a number of potential outcomes.

The demand for a number of different goods is very variable according on the time of year. When pricing these kinds of things, you need to use demand data from previous years, and you should aim for the highest possible level that price elasticity will let you achieve. There are also specific local events that have the potential to cause large spikes in demand for the related goods and services. In addition to taking into account the possibility of such events occurring while making orders and increasing inventory

levels, product pricing need to be modified so that they are in line with the surge in demand for the product. Off-season pricing should be lower in accordance with the law of supply and demand because firms want to create annual demand equilibrium that is as near to ideal as is practically possible.

Because seasonal shifts aren't the only thing that might have an effect on demand, the weather should also be taken into consideration when setting prices. Demand is a vital factor that is influenced by the weather on a local level. As a clear example, sunny weather may significantly raise demand for some products (Figure 3.9), such as sunscreen or ice cream; on the other hand, the peak demand for umbrellas comes during periods of precipitation. it will be feasible to increase the price as a result of the rise in demand, particularly the urgent necessity, which will also make it possible to do so.

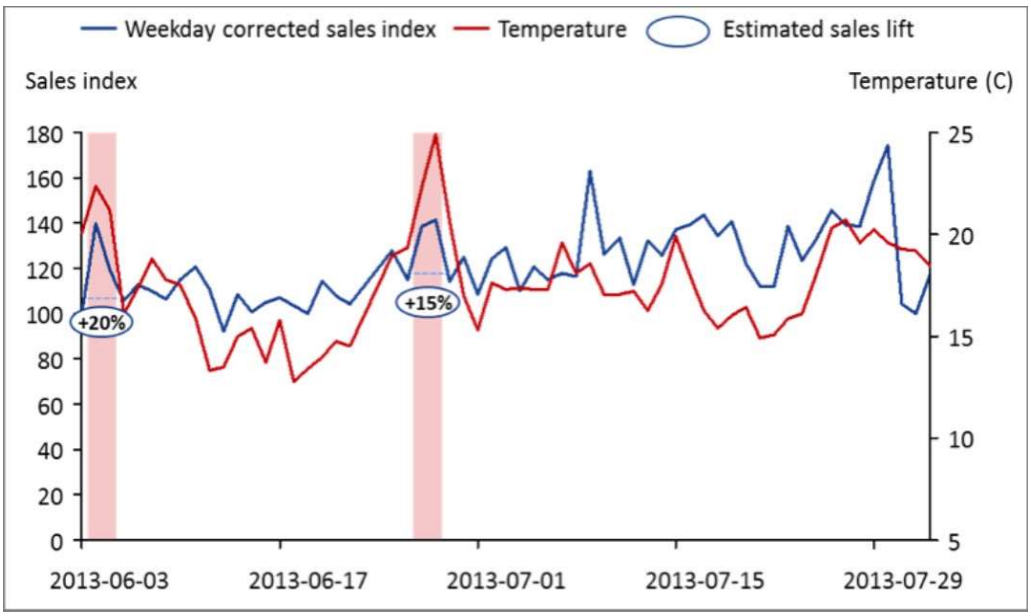


Figure 3.9 Example of weather impact in sales of all fresh products. (Ylinen 2014)

source: *machine learning in data- driven pricing, data collection and processing through by Amar Singh (2018)*

The data that is gathered from the quantities of items in stock may also be usefully applied to the process of price setting. Enterprise Resource Planning systems, often known as ERPs, are generally capable of providing a real-time image of inventory

levels in addition to providing information on inventory turnover rates. ERPs are owned and maintained by enterprises most of the time. It is conceivable to declare, on the whole, that it is intended for stocks to move as quickly as is humanly possible to do so. This is something that can be done.

Businesses are able to swiftly evaluate this data, and as an illustration, when there is an increase in the supply levels for a certain product, those businesses may choose to lower their prices for that product. This also works in the opposite direction; when inventory levels are too low and there is an excessive length of time until the next delivery, prices can be raised in order to limit demand and boost profits. This is done when there is an excessive amount of time between the next delivery. 2018 research by Tiwari and coworkers. The same strategy of optimizing becomes valuable for firms operating in the service sector as well, which frequently do not have physical inventory on hand but are instead faced with a fixed deadline by which all of their things must be sold. The service industry enterprises may benefit from optimizing their processes in the same way. Companies that are involved in a diverse set of economic activities fall under this category. Hotels and airlines, for example, are examples of such businesses

Because the profit made by the firm is determined by the difference between the selling price and the expenses, it is essential that these costs be handled extremely carefully in order to achieve the highest possible level of profitability. Utilizing the information provided by the organization regarding its variable and fixed expenditures is not a tough task. Nevertheless, a strong foundation in cost accounting is essential in order to effectively assign the relevant charges to the appropriate items. With the use of sensors connected to the Internet of Things, production costs may be monitored in a nearly real-time manner. Additionally, other aspects that contribute to production costs, such as broken equipment, can be foreseen.. The ERP systems owned by the firm provide extremely real-time data on the prices paid by the company's suppliers, which are, of course, reflected in the expenditures incurred by the company itself.

The value of one currency relative to another can have a large influence on a company's bottom line, an observation that is especially valid for companies that engage in significant levels of international trade. It is possible to acquire data on the exchange rate in a very short amount of time and does not require a significant amount of research. If businesses are willing to take this information into consideration, they will be able to design pricing systems that are capable of automatically adjusting prices in response to shifts in exchange rates. It is essential to keep in mind that when making

purchases, businesses have an additional responsibility to take into account variations in currency rates. This is because exchange rate fluctuations have a direct impact on the costs that are spent by the business.

one of the most important aspects that plays a role in establishing prices is the level of demand. In accordance with the principle of supply and demand, companies have the flexibility to charge higher prices for their goods and services when there is a larger demand for such goods and services. By continually monitoring the many factors that contribute to demand, data-driven pricing might lead to significantly higher profit margins or greater sales volumes as a result of increased competition. Demand is one of these components, but there are many other considerations to consider. Companies need to keep a close check on all of these factors collectively rather than monitoring them separately so they can account for the fact that excessive price increases might, of course, lead to a fall in demand in accordance with price elasticity. This is because of the relationship between price and demand.

3.7 UTILIZING MACHINE LEARNING IN DATA-DRIVEN PRICING

As was covered in chapter two, the pricing structure is established by a variety of factors, all of which are vulnerable to transformation at any time. When the world gets more globalized, companies need to be able to react to the changes that are taking place all over the world at all hours of the day and night. These changes can take place anywhere in the world at any time. Automated pricing systems are used by businesses since manually determining prices for products based on the vast number of available features is practically impossible to do. It is possible to arrive at such a data-driven dynamic pricing system through the use of machine learning, which makes it possible to do so. Dynamic data-driven pricing is a pricing model in which prices are continuously recalculated based on new data as it becomes available. Traditional data-driven pricing is distinguished from this sort of pricing by the fact that the systems that establish the prices are able to do so in a manner that is entirely under their own control and in real time.

The data obtained from customers may be used to develop multi-dimensional segmentation of the consumer base if the appropriate software is applied. Learning through machine interaction not only makes it feasible to dynamically update these segments, but it also makes it possible to use these segments in pricing. This is because learning makes it possible for machines to communicate with one another. This helps

the firm gain a deeper understanding of its clientele, which may, in turn, result in an increase in income for the organization. it is feasible to use predictive models in order to estimate the future consumption patterns and purchasing habits of these diverse groups of individuals. These groups of people may be broken down into many categories. This paints a picture that is, to some extent, accurate of the kind of demand that is anticipated as well as the kind of pricing that should be set up as a result of those expectations. This chapter covers a variety of alternative techniques to machine learning that might be utilized as viable strategies for the implementation of dynamic pricing.

An effective strategy for segmenting customers into distinct categories, the k-means clustering algorithm is shown here. The first thing that has to be done is settle on the overall number of clusters that will be used. The number of clusters that will be employed may be represented by the variable k, and each of these clusters will have a point in the centre that will be referred to as the mean. The mean will be the value that is most representative of the cluster as a whole. After then, the K-means algorithm will divide all of the inputs into clusters with the intention of making sure that each input is placed in the cluster that has the mean that is the most comparable to it. After that point, the algorithm will begin to optimize itself in such a manner that the positions of the means will move in such a way that the overall distance from each input to the mean will be as near to zero as is possible for a human to achieve. After this point, the algorithm will have reached its maximum potential.

The figure that can be seen lower down (graphic 10) provides an illustration of the outcomes that may be achieved via the utilization of k-means clustering. Because there are a total of three k:s present in this particular scenario, the result is the establishment of three separate clusters. In the context of pricing, these may be three separate customer groups, each of which exhibits a different pattern of purchasing behavior and varies in the degree to which they are able to pay the price. In this particular situation, the two axes may be broken down into the following categories: The y-axis depicts the level of activity on the online store, while the x-axis represents revenue. successful k-means algorithms contain a significant lot more of these facets than less successful ones do. This is done in order to ensure that the maximum number of consumers are able to comprehend the information. Customer segmentation, such as the one discussed in this article, is one of the applications of k-means clustering that gets the most usage and is one of the applications that sees the most use. 2019 for the Sagar

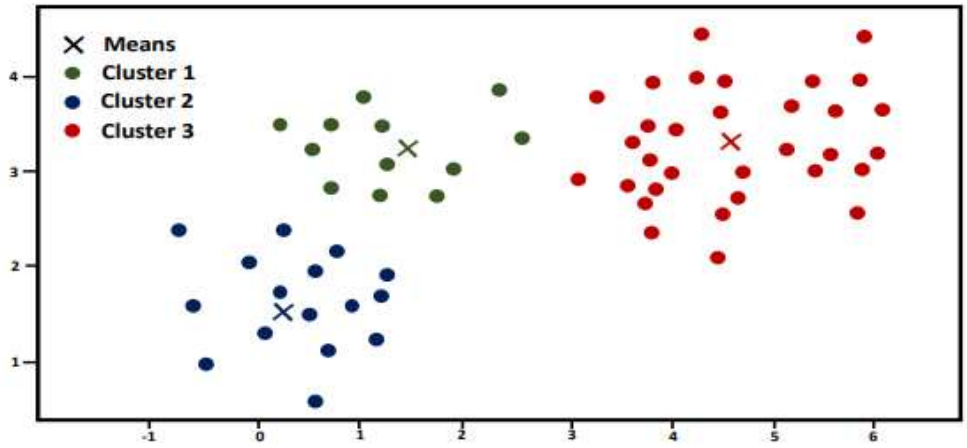


Figure 3.10 K-means clustering example (adapted: Sagar 2019)

source: *machine learning in data- driven pricing, data collection and processing through by Amar Singh (2018)*

In addition, machine learning makes it possible to create forecasts about the future purchases of particular customers by making use of the customers' previous transactional data. The decision tree learning algorithm is one of the predictive models that is widely used to estimate future outcomes of X based on prior input and output data. Other predictive models include neural networks and Markov models. Research findings that were given in Gupta suggest that the output of a decision tree is governed by a variable called X that travels from the tree's root node to one of the tree's leaves. At each intersection, the decision tree contains a node, and at each of those nodes, the successor outcome is determined based on the splitting. In the vast majority of cases, this splitting is defined by one of the qualities of X, however in other cases, it is determined by a specified set of splitting criteria. X can be divided in a number of different ways.

An example of how to determine whether or not a certain consumer will purchase a product at a particular price is presented in the diagram (Figure 3.11), which can be seen below. For this particular set of circumstances, the value of the outcome leaf will either be "SALE" or "NO SALE." Before deciding whether or not the consumer made a purchase, the decision tree will initially establish the client's gender. This is done in order to facilitate the taking of the relevant action. At this point, the tree divides into two separate branches, each of which asks the same set of questions but arrives at a

different collection of conclusions. The next piece of information that is collected by the tree is the age of the individual who is using the product. The assumption that the customer would not make a purchase is made by the decision tree when it comes to the female branch of the decision tree. This assumption is based on the fact that the client's age must be at least 20 years old. When the consumer's age is over 20, the decision tree analyzes whether or not the individual is currently enrolled in an educational program. In the event that the client is a student, the transaction will not take place; on the other hand, if the customer is not a student, the transaction will be carried out.

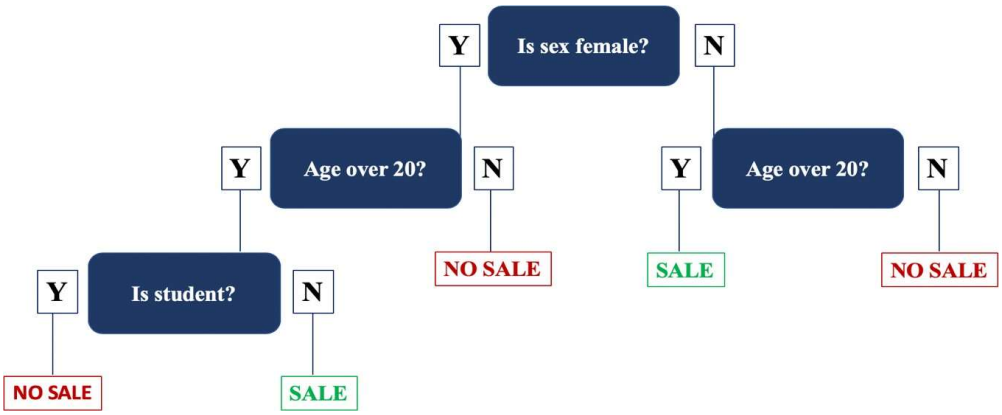


Figure 3.11 Desicion Tree Learning Example

source: *machine learning in data- driven pricing, data collection and processing through by Amar Singh (2018)*

The dynamic price fluctuation is the most significant item that can be performed with the assistance of machine learning. To achieve this objective, one approach that might be employed is machine learning algorithms based on reinforcement. reinforcement algorithms are appropriate for pricing since they were created for contexts in which a high number of both external and internal components are susceptible to frequent change. As a result, reinforcement algorithms are ideally suited for pricing. These sorts of algorithms are able to be "trained" using a significant amount of historical data, and then they may be optimized to identify the pricing that will result in the greatest potential profit. Before reaching a judgment, these algorithms are able to take into account a myriad of different aspects of the situation. Because it is a computer, it is able to make these decisions with a constant supply of information and take into consideration a greater number of factors than a human brain could ever hope to handle.

Which may be located at this location, features an analysis of an example of dynamic pricing being applied. When calculating the price that will be charged, a straightforward neural network like the one depicted in the accompanying figure takes into account the four inputs listed below. In order for this model to be able to generate pricing judgements that are as close to perfect as is humanly conceivable, it must first be trained on data pertaining to previous pricing decisions. For example, because of the resulting margin and the demand, it has been rewarded, and as a result, it has improved. This is because the consequent margin and the demand. The components that flow into this particular network include things like estimated demand, client segmentation, the weather, and the amount of inventory that is currently available. In order to create a model that is a more accurate representation of reality, there has to be a significant increase in the number of these components. One of this model's numerous features that makes it particularly appealing is its capacity to promptly generate fresh pricing that is best suited to the conditions at hand. This is only one of its many advantages.

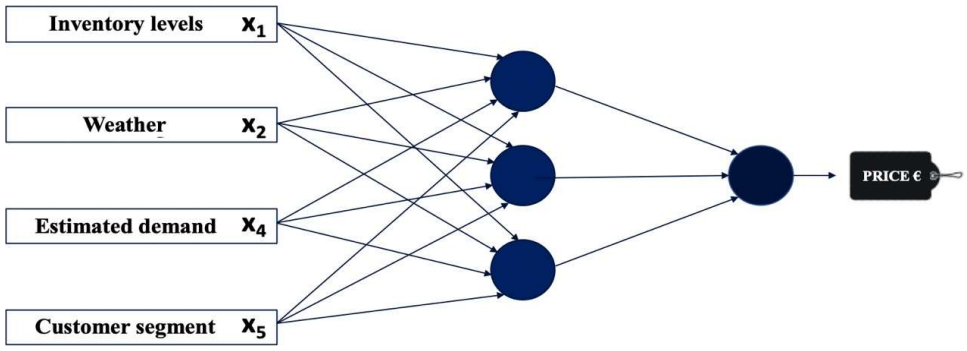


Figure 3.12 Pricing neural network

source: *machine learning in data- driven pricing, data collection and processing through by Amar Singh (2018)*

When these algorithms are included into the pricing structure of a corporation, the price begins to take on a dynamic quality in the truest sense. Some companies have begun using electronic price tags in traditional stores. These tags are located on the shelves. Electronic price tags like this are dynamically updated in the same way as their online equivalents are. The application of dynamic pricing is not limited to the sphere of online shopping alone. Despite this, dynamic pricing is being used a great lot more commonly in online merchants, notably in the transaction of products and services connected to transportation, such as purchasing plane tickets.

3.8 BENEFITS AND RISKS OF DYNAMIC DATA-DRIVEN PRICING

The benefits of machine learning are clearly evident when it comes to data-driven pricing, including the following: prices can be regulated in a dynamic and efficient manner; the suitable price can be identified for the relevant persons; and demand projections can become more exact. These adjustments in price and revisions to the projection can also be used to minimize the amount of stock that is kept on hand while simultaneously increasing the rate at which it is turned over. When taken into consideration as a whole, these aspects help to make the revenue management of a business more efficient and fruitful.

It is essential to determine the appropriate price for the appropriate people at the appropriate time since doing so enables the maximum amount of cash to be created from each individual customer as well as total demand. This is why it is necessary to locate the proper pricing for the appropriate people at the appropriate time. The price may be set exactly at the level at which the transaction is still completed in order to take use of the very accurate projections that may be generated by machine learning algorithms. Additionally, dynamic segmentation makes it feasible to categorize individuals into separate consumer groupings depending on the method in which they make purchases. This is accomplished via the use of a scoring system.

Even amongst these two groups, it is possible, with the right amount of control, to fix the price at the level that is most profitable. If all goes according to plan, these prices will be able to be adjusted to a fully personalized level based not just on client profiles, but also on the degree to which each particular consumer is price elastic. The most important advantage is that components that would not be factored into the decision-making process without the assistance of machine learning might be included in the price option. Another advantage is that the combination of the many different elements that, taken as a whole, determine how much is bought, when it is bought, and at what price it is bought at is also included. This is a benefit since it allows for more accurate decision making.

Despite this, there is not a guarantee that implementing dynamic pricing would result in a positive impact on the operations of the business. The results have the potential to be rather bad, particularly in situations in which it is applied inappropriately and not in the way that it was intended to be used. Because bad pricing decisions have the potential to result in significant revenue losses, firms simply cannot afford to make significant

mistakes. Forecasting errors, difficulties with pricing in general, and excessive price modifications are at the very least some of the potential risks that are linked with dynamic pricing. Demand estimates that are done poorly are simply wrong; this is why judgements that are made on the basis of them are likewise erroneous and not very good; they were based on projections that were inaccurate and were performed poorly. Pricing choices that are made on the basis of an erroneous assessment might subsequently result in a margin that is much too low or in prices that are much too high, both of which are reflected in demand that is even lower than before. one of the biggest risks posed by this circumstance is the possibility that the company might lose some of its prospective customers. It is possible that the firm may lose both its existing clients and any new consumers that it acquires in the future if people start to boycott the business because of the exorbitant prices it charges.

Businesses run the risk of making honest pricing errors when using dynamic pricing, which is a risk in and of itself. In spite of the fact that this risk is present in the same way when pricing is done using traditional techniques, it is increased in dynamic pricing systems that are driven by machine learning. this is owing to the fact that these systems are usually "black boxes," whose functions are difficult to fathom. It is of the highest importance how the many facets of the decision on pricing are weighed in order to arrive at a final decision. For example, the history purchasing patterns of a customer are likely to have a greater effect on the client's future purchases than the present exchange rates; yet, algorithms may improperly weight these elements, which results in projections that are not even close to being accurate. As a consequence of this, the price that is produced by the dynamic pricing model has the potential to be completely wrong, which may cause a lot of prospective customers to decide against making any purchases at all.

Price variations brought about by an excessive amount of market volatility might also be hazardous. In spite of the fact that in theory it yields the greatest results when the pricing is dynamically altered in response to the conditions, this technique has the potential to upset customers quite a little. If the price fluctuates too frequently—for example, more than once an hour—potential customers may view this as unfair and decide to postpone making a purchase decision in order to watch for lower prices. Alternatively, they may choose to wait until prices stabilize before making a purchase decision. (Dholakia) as of the year 2015 This position is obviously not ideal because businesses want demand to be as consistent as possible and sales prices to bring in the

largest potential profit margins. Because of this, the current situation is not ideal. The goal of businesses is to have demand that is as uniform as feasible. If consumers choose to do business with competitors whose prices are more stable over a protracted period of time, it is feasible that total demand will decline as a result of this shift in consumer behavior.

3.9 CREATING DYNAMIC PRICING SYSTEM

3.9.1 Dynamic pricing system suitability

If a corporation had so many different products available for purchase at the same time, it would be either completely impossible or, at the very least, extremely difficult to accurately price all of those products without utilizing a pricing system that was both dynamic and data driven. A dynamic pricing system is especially well suited for firms like these, and it may even be essential for them to have one. When you consider that certain businesses, like Amazon, sell more than 12 million distinct products at the same time, it is simple to see why automation and machine learning are essential in the retail industry (Dayton 2020). It is also conceivable to state that the benefits of dynamic pricing are further emphasized when the same company has several things; as a result, smaller organizations in particular need to think about whether or not dynamic pricing is profitable.

In order for a company to successfully implement dynamic pricing, they need to be willing to take a methodical approach and handle the process on their own. The system will not be effective on its own, and outstanding results will not be achieved without careful analysis of the business environment, planning of pricing strategy, and awareness of the factors that influence the relationship between demand and price. In order for a company to be prosperous, its employees need to have extensive expertise on a wide range of subjects, including data management, pricing, and machine learning. This is something that is very necessary for the organization. In addition to this, the company has the choice to collaborate with a consultant who will act on their behalf to handle the current predicament.

The relevance of data volume is also being pushed to the forefront as a direct result of the increasing dynamic nature of pricing. For big players, who often have thousands of commodities in their inventory, this data is more easily available. It is necessary for companies to collect huge amounts of data from the outside world in order for them to

have a proper grasp of both their customers and the operational environment in which they function. In order for price to be modified effectively in reaction to changes in inventory levels and costs, data from within the organization must also be made available. Because of this, it is possible that the construction of a model is not only lucrative but also perhaps practical only for organizations that have substantial quantities of sales to support the implementation of the model. Because there are a variety of machine learning algorithms that may create predictions that are mostly dependent on previous data, it is not even worth studying a pricing model if it does not have at least one component that is prediction-based.

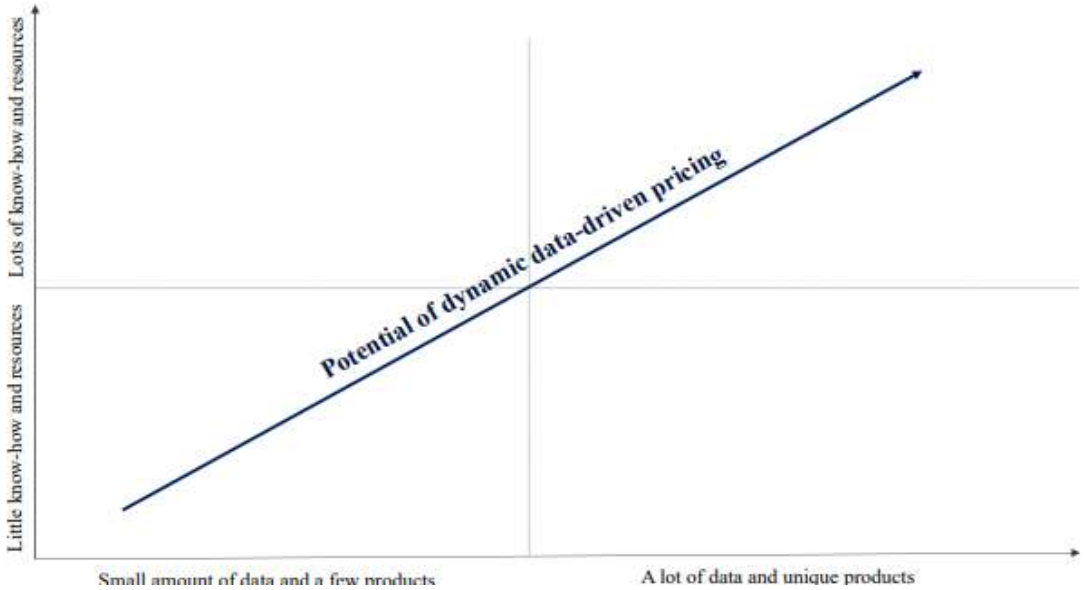


Figure 3.13 Potential of dynamic data-driven pricing

source: *machine learning in data- driven pricing, data collection and processing through by Amar Singh (2018)*

When considering the feasibility of adopting dynamic pricing that is based on such data, it is impossible to stress how important it is to make sure that one possesses data of a high quality. A dynamic pricing model that is driven by data cannot be developed in the absence of high-quality information as well as the orderly collecting and storage of such information. This is because the two are inextricably linked to one another. It is necessary to do quality assurance checks on the data not only at the time of the first import of the data into the system, but also on a continuous basis. As a result, one of

the requirements for the organization is the collection of diverse and well-organized data, as well as the construction of an operational environment for the storage of data, in which a range of data sources and types may be investigated. These are both requirements that must be met. In a scenario like this, a data lake is a fantastic alternative to consider.

On the basis of demands that were determined in advance, the accompanying figure (Figure 3.13) gives a condensed description of the applicability of dynamic data-based pricing in a wide range of various kinds of businesses. In conclusion, the possibility that dynamic pricing is an option that is suited for a firm improves in proportion to the quantity of data that the organization owns from a range of items, as well as the amount of expertise and resources that the company possesses.

3.10 DYNAMIC PRICING INFRASTRUCTURE

In order to reach the highest possible degree of efficiency, it is essential to approach the process of installing a data-driven dynamic pricing system in a logical manner. As a result of this, a detailed explanation of the approach is crafted with the purpose of acting as a sort of road map for the process. This technique, also known as dynamic pricing infrastructure, is comprised of a total of 11 separate procedures (see Figure 3.14). The steps in this procedure are normally carried out in the order that is stated, however some of them can be repeated in the reverse order if that is needed.

The approach begins with a thorough examination of the current state of the business climate as its first step. At this stage of the procedure, the objective is to carry out an in-depth study of both the external environment in which the organization works and its own internal affairs. This inquiry will focus on both the external environment and the internal affairs of the company. During this time period, the operating environment, the competition, and the items offered by rivals, in addition to the firm's own strengths and weaknesses, are examined. The purpose of the research is to collect data that can be useful for the firm's pricing strategy as well as for the many other stages of the process as a whole. The analysis is quite important since it plays a large part in deciding the kind of strategy that a company should begin developing and the kind of goals that it should follow. As a result, it is necessary.

The stage at which the pricing strategy and objectives are addressed is the stage that guides the price at which the entire company offers its products or services. It not only

outlines the objectives of the pricing strategy of the organization, but also the approaches that will be used in order to realize those objectives. At this stage in the procedure, a determination is made about the pricing strategy, and the utilization of dynamic pricing is one of the possibilities that are accessible. A conclusion must be drawn on the necessity of implementing dynamic pricing based on the findings of the business environment assessment. This conclusion must be reached before dynamic pricing may be implemented. The pricing strategy also determines certain boundaries, such as minimum and maximum margins, for the items that it sells in accordance with these parameters.

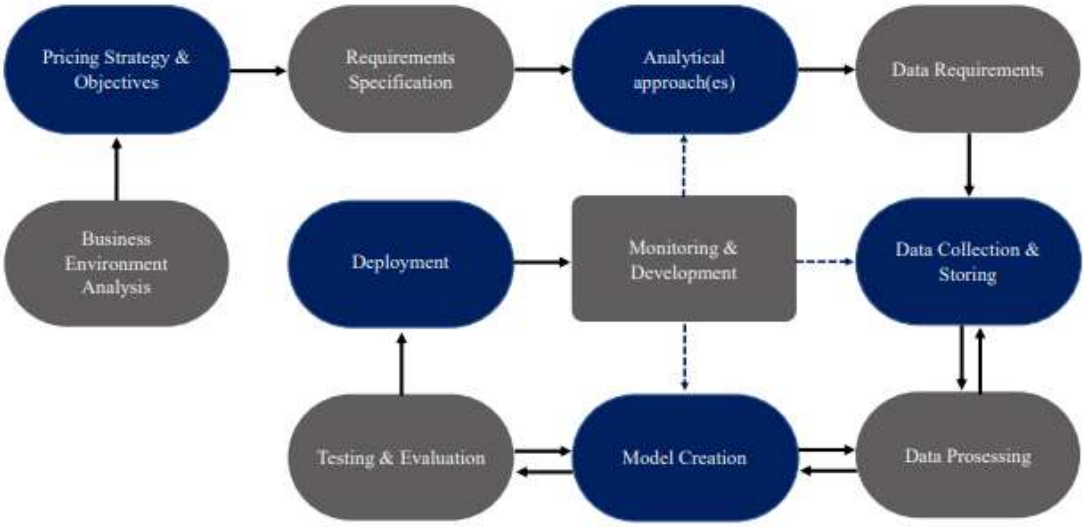


Figure 3.14 Dynamic pricing infrastructure

source: *machine learning in data- driven pricing, data collection and processing through by Amar Singh (2018)*

The third step consists of outlining the criteria for what has to be done. The purpose of this stage is to produce a list of all of the objectives and requirements that need to be fulfilled in order for the completed product to be regarded as successful. The functional and non-functional features are both taken into consideration in the evaluation criteria for this phase. The functional requirements lay forth the characteristics that the system need to possess so it can do its job well. This may entail, for example, offering estimates on an hourly basis or optimizing the price on at least two independent occasions each and every day. Alternatively, it could require both of these things. The standards for

evaluating non-functional characteristics might be qualitative or related to resources. The degree to which a dynamic pricing system should adhere to standards of reliability, security, and efficiency is going to be decided by the system's qualitative criteria. On the other hand, resource related requirements are what decide the amount of time and money the business is willing to invest in the deployment and maintenance of the pricing system.

During the stage of analytical approach, it is decided which statistical and machine learning techniques will be applied, as well as the specific sorts of those approaches. Because the problems govern the type of the patterns that need to be identified, the goal also entails assessing which issues are the most critical ones that demand solutions. This is necessary because the patterns that need to be found are dictated by the problems. Due to the fact that the methods used have such a substantial bearing on the outcomes, it is imperative that this stage be approached with extreme caution. If determining customer subgroups and basing prices on those categories is part of the plan, the analysis approach likely has to be one that is both descriptive and unsupervised. If, on the other hand, the goal is to anticipate future purchases and demand, a predictive model such as the decision tree could be the one that is the most suitable alternative. However, it is possible to acquire the most accurate results by integrating the findings of numerous separate models in order to, for example, categorize customers and simultaneously estimate future demand. This may be done in order to produce the most precise results possible.

Following the selection of the appropriate methods for conducting the analysis, the essential data should be given. Examining the many different kinds of analytical methods that have been applied in the past might help to provide light on the sorts of data that are necessary. At this point, not only the standards for the data's quality and the formats of the data, but also the locations from where it will be received are being decided upon. Because machine learning algorithms are completely ineffective if they do not have any data to work with, it is extremely necessary that all of the pertinent data be discovered and identified. After an analysis of the requirements for the data has been carried out, the following stage is to begin the process of data collection and storage. Having said that, it is essential to keep in mind that this particular phase is a continuing one given that new data is being produced as the procedure is being carried out. Given that dynamic pricing requires the collection of as much real-time information as is reasonably possible, it stands to reason that this data should be

continually obtained. This is because dynamic pricing needs the collection of as much real-time information as is reasonably possible. As a consequence, the term "step" refers, in its most fundamental sense, to the process of automating the gathering of data for a particular data warehouse or data lake.

The stage of processing data involves the processing and organization of the data that is going to be used by the machine learning algorithms in the next step. At this point, the most essential thing that needs to be done is to make sure that the data are of a high quality and are presented in the correct manner. Data exceptions are deleted since the information that they offer might be misleading; acceptable column names are provided; and data in general is cleaned up. This phase is a continuous process much like the one that came before it since the data must always be in the right format, not just when the model is being generated. This phase follows the same pattern as the phase that came before it. Because negative results achieved via machine learning are directly associated to poor data quality, it is impossible to overestimate the relevance of this phase due to the reason that it is impossible to overstate the significance of this phase.

After the data have been analyzed, the creation of the model might perhaps get underway. During this phase, both the machine learning algorithms and the price mechanism that will be derived from them will be developed. During this phase, the machine learning algorithms will be constructed. It is vital to create algorithms for machine learning that are based on specified analytical approaches in order to keep track of the plethora of factors and happenings that are taking place both outside and inside to the company. For example, when working with a predictive model such as a decision tree, the model is trained and improved with the use of historical data from the model's past iterations. This is done by using data from the model's previous iterations. These characteristics are collected and evaluated in real time by algorithms, which then utilize the data to produce price decisions that are dynamic in nature. The price is controlled at the intervals that are mentioned in the specification of the requirements on the basis of these judgements.

Following the completion of the model's construction, it will be put through a series of tests and examined. The findings that have been supplied by the pricing system as well as the results that have been created by the machine learning algorithms are both put to the test and assessed at this stage of the process. The outcomes that are created by the predictive models may be compared with actual data from the real world, and in

general, the pricing decisions that are made by the system, as well as the margins that are produced, can be approximately calculated. At this stage, it is straightforward to evaluate the additional value created by the model. For instance, one may investigate whether or not the sales revenues are greater when the pricing is set according to what is advised by the model.

However, it is essential to bear in mind that the model has not yet been implemented at this point in time; rather, it will be tested in combination with the systems that are already in place. This is something that should be kept in mind at all times. Because, as was said earlier, erroneous pricing choices can lead to really bad results, testing should take a significant amount of time in order to see the impact that various changes have. Due to this reason, it is extremely important that testing take place. In the case that the outputs are not sufficient and do not match to the goals that are outlined in the requirements specification, the process will need to be iterated in the other direction.

When all of the testing has been finished and it has been shown that the model generates successful results, it may then be deployed. This is a reference to the implementation process that takes place when the pricing model is included into the pricing structure that was previously in place. When presented with such a scenario, the machine learning model will begin to implement dynamic pricing control. However, in conventional stores, where it is impracticable to make manual modifications to the pricing of specific items, this demands the use of digital price tags since it is quick and easy to make price changes in online companies because of how easy it is to make such changes.

The second step, which will take place once the model has been put into action, will include continuous monitoring and enhancement. Putting such a dashboard, for instance, is a great way to simplify the process of keeping track of the results. Concerns that need to be monitored closely include, among other things, the margin for the product or product group, the number of price adjustments, the demand forecast in relation to the actual demand, and a general comparison of the outcomes to pricing that is not dynamic. The mechanism will need to be designed in the case that there are problems throughout the process of monitoring, such as the forecasts not being correct on any level. In general, the model needs to go through continuous evolution because there are always going to be new data sources and more powerful technologies being generated.

1. **What are the benefits of utilizing machine learning in data-driven pricing, in addition to the challenges and risks that it presents?**

The use of data-driven pricing in conjunction with machine learning is advantageous in a number of ways, as is readily apparent: When demand forecasting for particular customers becomes more exact, it will be possible to discover the optimal price for the suitable persons. Prices are able to be managed in a dynamic and effective manner. If these are included into an organization's pricing systems in order to enable dynamic price modifications, then the requirement for manually adjusting prices will no longer be necessary. Because of this, the company will become more efficient, and if the pricing plan is successful, it will also result in an increase in earnings and overall sales.

On the other hand, dynamic pricing comes with its own set of issues, such as the risk that some customers would feel dissatisfied by the extreme price fluctuations and perceive that it is unjust. This is just one example of the difficulties that might arise from using dynamic pricing. In addition, the findings of machine learning algorithms' projections and the conclusions they generate could be off-base at times. This is because machine learning algorithms are not entirely reliable. Another challenge that must be surmounted is determining the right weightings to be given to the various components that go into determining the pricing. In the event that these are inaccurate, pricing will unquestionably be a catastrophe that cannot be alleviated in any way.

2. **What different kinds of data are necessary for the establishment of a dynamic pricing system that is driven by data, and where can that data be collected?**

In order for machine learning algorithms to operate effectively, there must first be a substantial volume of data collected. In order to obtain reliable outcomes, it is necessary, first, that the data be of a high quality, and then, second, that they be presented in the suitable way. Because demand is affected by a large number of other variables, it is vital to assemble information on a wide variety of issues in order to guarantee that almost every facet is taken into consideration. This may be accomplished by gathering information about a wide range of topics. very price is based on the customer's knowledge, which is why data acquired from customers is of the biggest importance. [C] onsumer knowledge is very important in [(especially pricing]. These data can be collected in a variety of ways, such as through the websites of the firm itself, directly from open source sources, or by purchasing them through data marketplaces. Other potential sources include those that are freely available.

Additionally, the company's internal systems give a considerable number of data on manufacturing costs and inventory levels. Both of these aspects should be included into the process of pricing the product or service.

3. What requirements does a company need to fulfill in order to have pricing that is dynamic and driven by data?

In order for a company to adopt data-driven dynamic pricing, the company must first undertake meticulous planning of their pricing strategy and then collect precise and diverse data. Only then can the company apply data-driven dynamic pricing. Steps that must be taken in order to establish a successful strategy include doing a thorough analysis of the external conditions of the business and gaining an understanding of the factors that have an impact on both demand and price. Because of this, it is essential for the company to have a thorough understanding of the business environment, in addition to a solid awareness of both its own operations and those of its competitors. In addition to this, the organization has to have a significant level of competence and awareness regarding information technology-intensive subjects such as machine learning, data analytics, and system integrations. In general, creating data-driven dynamic pricing is not a process that is simple or quick. As a result, the implementation of such pricing requires a substantial amount of time, resources, and attention on the part of the business.

Because of the numerous advantages that come along with its implementation, dynamic pricing techniques are likely to grow increasingly popular in the near and far futures. One of these advantages is an improvement in the profit margins, and another is an increase in the demand from customers. The ever-increasing number of data as well as the development of technologies such as machine learning are contributing to the fact that it is becoming progressively less difficult to adopt dynamic pricing. As a consequence of this, there is a substantial probability that in the not-too-distant future, dynamic pricing won't be quite as exceptional, and its use will develop into something that is utilized more frequently.

As machine learning and artificial intelligence continue to make strides forward in their respective spheres of application, price optimization will become increasingly accurate and dynamic. The role that neural networks, and more especially deep learning, play in the process of pricing goods and services will receive increased attention in the not too distant future. It is of the utmost importance that extra efforts

be made in order to study both of them and their prospects from the perspective of the cost. As a result of the vast potential of the market, it is a certain that in the not-too-distant future, there will be more software that can automatically modify pricing according to the conditions of the market.

CHAPTER 4

DATA-DRIVEN SALES FORCE SCHEDULING

Businesses operating in a diverse variety of markets need to figure out how to make the best possible use of the limited number of sales force members they have available. This suggests that in order to maximize the expected revenues in the future, one must decide which of the various projects should be the major focus of one's efforts in order to maximize revenue potential. In the vast majority of instances, the profitability of these ventures is not the only element that differentiates them from one another; rather, they also have distinguishing traits, such as the specific sort of product or service that is supplied, the location, or past interactions with the target client. It is reasonable to speculate that the presence of these characteristics increases the likelihood of a company getting selected to work on a certain project. On the other hand, these characteristics may also assist to measure the degree to which increasing one's sales effort raises one's likelihood of securing a contract for a project ("the uplift") in the bid process. This may be accomplished by calculating the degree to which increased sales effort increases the likelihood of being awarded a contract or tender.

Using this method, one is able to assess the likely marginal advantage that will result from a salesperson making a visit to a potential customer. Starting with a massive data collection that contains both successful and unsuccessful efforts allows us to design a one-of-a-kind method that is data-driven and is based on scheduling the sales force. We combine machine learning algorithms for uplift prediction with routing and scheduling models by building on top of this data set. In particular, this method takes into account the fact that estimations of uplift are not foolproof and that the ensuing uncertainty must be taken into account when scheduling a sales force. In addition, this method takes into account the fact that there is a limited amount of time available to schedule a sales force. The planning process may be made more precise by taking these two considerations into account.

4.1 INTRODUCTION

In many different industries, including as the pharmaceutical industry, the construction industry, and the industrial services sector, businesses allocate a significant amount of their marketing budgets to activities that include their sales employees. In order for

businesses to effectively plan and organize the activities of their sales force, it is important for these businesses to identify and prioritize the projects in which increased sales efforts lead to greater predicted revenue increases. Because of this, the companies are able to organize and timetable the operations of their sales teams. In recent years, businesses have reduced the size of their sales teams while at the same time increasing their investments in digital technology. This is done in order to improve the efficiency of the sales agents who are still employed by the company and, more specifically, to improve their ability to target specific customers.

The conclusion is often a "priority list" of potential customers with the greatest likelihood of converting into paying customers. However, in order to organize sales operations, such information is insufficient if the required sales effort is not consistent across all of the available projects. A good illustration of this would be the amount of time invested in traveling to and from in-person sales meetings. The scope of travel activities will be determined, in part, by the geographical locations of customers and any other booked clients, in addition to any other potential customers' locations. As a consequence of this, scheduling the sales force becomes a process that involves numerous salesmen's routes, with the objective of improving income despite the fact that the input parameters are uncertain. However, the essential optimization framework for sales-force management has not yet been investigated in the research literature. This framework need to contain prediction and prescription, but it has not yet been researched.

In this work, we aim to close this research gap and present a data-driven approach for overcoming the constraints of integrated targeting and scheduling of sales forces. The driving force behind our work is a research project that we are collaborating on with DAW, a significant producer of paint and coating solutions in Germany. DAW increases its chances of being given contracts to sell paint, mortar, and other products linked to construction by conversing with a range of clients, such as painters, processors, or planners, via its direct sales channel. These conversations take place in order to ensure that DAW is able to meet the needs of its customers.

Not only are the projects unique in terms of the amount of revenue that they have the potential to create, but they are also diverse in terms of other elements, such as the specific sort of product or service, the location of the project, or the history of interactions with the related partners. Even if no further sales effort is made, it is permissible to conclude that these characteristics are, at the very least in part, predictive

of the businesses' chances of acquiring a particular project. This is the case even if there is no other sales effort done. This is due to the nature of the characteristics that are under consideration. However, it is difficult to estimate the likelihood of obtaining a project based on the particular qualities of the project, and it is even more difficult to anticipate how a particular sales activity would raise this probability (the "uplift"), which may also be impacted by the characteristics of the project. Nevertheless, it is possible to estimate the likelihood of obtaining a project by dividing the likelihood of obtaining the project by the particular qualities of the project. These two activities are equally challenging to do.

The process of scheduling is difficult not only due to the fact that the company does not have solid information about the uplift, but also because to the fact that the capacity required to visit different customers varies from project to project. This makes the procedure particularly challenging. Because of this, completing the assignment will be very difficult. This is mostly attributable to the fact that the locations of the potential clients are dispersed throughout a large geographical region. The distance between a client and a sales person's home base might affect the amount of time that is required from the sales professional during a customer visit. The larger the distance, the more time that will be required. Therefore, while scheduling its sales force, the business has to take into mind the fact that visiting a "promising" client far away from the home base may limit the number of other visits of customers who may look less promising but are situated closer to the home base. This is because of the fact that visiting a "promising" client far away from the home base may restrict the number of visits of customers who are placed closer to the home base.

An colleague of ours in the business world has spent the better part of the last several years accumulating a substantial quantity of data on prior project bids, including both those that were successful and those that were unsuccessful. On the basis of this data, we develop an end-to-end solution that leverages cutting-edge machine learning algorithms for the purpose of predicting uplifts and solving a routing issue in order to produce the most efficient schedule for the sales force. This solution also addresses a problem with the way that the sales force is scheduled.

An essential component of the answer to this problem is the projected rise in income for a project that is associated with an additional visit from a customer. In order to provide an accurate estimation of the rises, we provide a procedure that is divided into two stages: To begin, we put a predictive classification model through its paces by way

of training in order to ascertain the probability of being given a certain project. After that, our uplift approximation method makes use of this prediction model by including it as a core building element so that it may take use of the results. Given the current circumstances, it is not plausible for us to think that our estimations of the uplift would be accurate to the one hundred percent level. Because of this, the model that is used to schedule the work of the sales force has to take into account the residual uncertainty that is linked with the estimates of the uplift. In order to do this, we have devised an innovative way of weighing that is influenced by the issues that are associated with decision analysis. This technique has the benefit of being able to specifically adjust for the degree of confidence that is attributed to the prediction model. This is a distinct advantage over other available methods. In this part of the report, we will provide a detailed numerical analysis of the proposed strategy. A breakdown of the overarching approach may be seen in Figure 4.1.

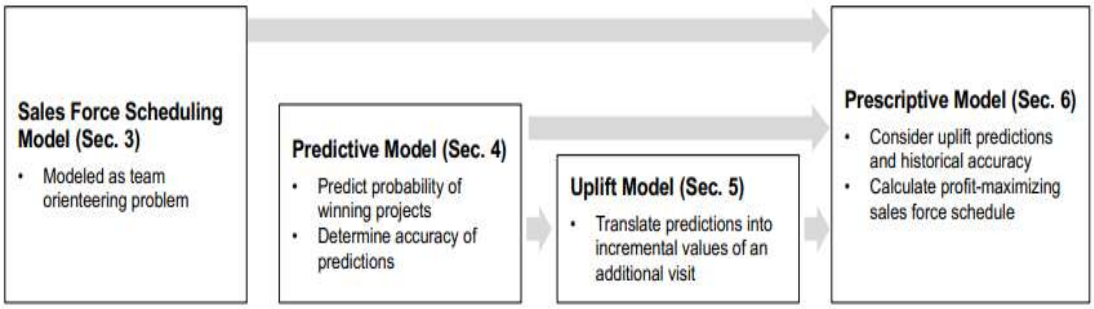


Figure 4.1: An overview of the suggested data-driven methodology

Source: *Data-driven Operations Management Data Collection and Processing through by M.Sc. Jan Maximilian Meller, In November 2019*

Our methodology is unique in that it takes into account the unpredictability of the uplift estimations, which is a feature that sets it apart from other approaches. Because of this, we are willing to acknowledge the inherent probabilistic character of the uplift forecasts, which may cause the sales force scheduling model to mistakenly favor customers with an excessively optimistic uplift prediction over customers with an accurate or too conservative estimate. We show the applicability of this method by using data that was collected from the actual world. This method should be applicable to a broad range of different firms who are experiencing difficulty in scheduling their sales staff since we used data that was collected from the real world. When seen in a

larger perspective, our technique establishes an important link between marketing and sales analytics and the more common challenges that emerge in the management of operations. This relationship is vital because it allows for more effective resolution of both sets of problems.

4.2 ISSUE DESCRIPTION

Imagine there is a company that is soliciting proposals for a number of different kinds of projects, each of which has a unique potential for financial gain. Every single one of the consumers, shown by the dollar sign c, is associated with a particular project. Customers may be associated with more than one project if they work with a major construction business that is often active in a variety of different building projects. Customers that work with such a company are more likely to be linked with many projects. Kc is the abbreviation that we use to refer to the group of projects that are associated with a certain customer c. Let's abbreviate the possibility that the company will be given project k as "Pk," which stands for "probability." In order for the company to increase its chances of being given a project, it may choose to implement a sales effort in the form of a visit to the potential customer in order to boost its odds of doing so. After an additional visit, we describe the rise (such as the shift in the chance of winning project k) with the sign pk, which stands for uplift. This is because pk is an abbreviation for uplift.

The company's sales force is comprised of sales agents that are based in a single headquarters location around the country. Each sales representative has a maximum capacity, which is denoted by the symbol and may be interpreted as the amount of hours of work they put in on a daily basis. This capacity is represented by the symbol. T^{travel}_{ij} is an abbreviation that stands for "total amount of time spent at a particular consumer." In addition, the amount of time necessary to travel between two locations (either clients or the home base) i and j may be determined by using the formula T^{travel}_{ij}. This time can be included into the total amount of time spent traveling.

The major goal of the company is to design a sales force strategy for the company that will enable it to generate the greatest amount of anticipated additional profits (expressed as vp'q) in connection with this timetable. Before the company can create a schedule, it must first choose a certain tour set denoted by T. In order to do this, one must take into consideration the uplifts pk, the profits k, as well as the capability of the sales force that is now accessible. The "tour" that a sales representative is on will dictate

the order in which they visit a group of customers over the course of a particular day. This "tour" will also decide which customers they visit first. The letter xijt is used to symbolize the binary variable that indicates whether or not the trip segment from location i to location j is included in tour t. The letter yckt is used to denote the binary variable that indicates whether or not customer c is visited on tour t in order to pitch project k. Both of these variables are considered to be optional. The letter t is used to represent both of these different variables.

The decision variables xijt and yckt thoroughly specify both the tours that are to be visited as well as the subset of customers that are to be visited. Because we are operating on the presumption that each sales rep is able to do exactly one tour, we have calculated that the overall capacity of the sales force for the next day will be more than the maximum allowed. As part of the second step of the procedure, a sales person is appointed to accompany each trip. In this analysis, we use the premise that sales professionals have preferences that are similar to one another and that the uplifts pk are not dependent on the particular sales person with whom the customer interacts. As a direct result of this, a sales representative may be appointed to any tour t P T, and a schedule is equivalent to a collection of tours T. This organization will work on optimizing the following process in order to provide the most effective schedule they possibly can:

$$\max v(\pi) = \sum_{c \in C} \sum_{t \in T} \sum_{k \in K_c} \Delta p_k \chi_k y_{ckt}$$

Taking into account the several restrictions that are listed below:

$$\sum_{j \in C} x_{0jt} = 1 \qquad \forall t \in T$$

$$\sum_{i \in C} x_{i0t} = 1 \qquad \forall t \in T$$

$$\sum_{j \in C, j \neq c} x_{ijt} \geqslant y_{ckt} \qquad \forall i \in C \;\&\; \forall k \in K_c \;\&\; t \in T$$

$$\sum x_{jit} \geqslant y_{ckt} \qquad \forall i \in C \;\&\; \forall k \in K_c \;\&\; t \in T$$

$$\sum_{j \in C, j \neq c} \sum_{s \in T} y_{ckt} \leqslant 1 \qquad \forall c \in C \;\&\; \forall k \in K_c$$

$$\sum_{j \in C} x_{0jt} \tau_{0j}^{travel} + \sum_{i \in C} x_{i0t} \tau_{i0}^{travel} + \sum_{c \in C, j \in C} x_{cjt} \tau_{cj}^{travel} + \sum_{c \in C, k \in K_c} y_{ckt} \tau_{c}^{visit} \leqslant \eta^{max} \qquad \forall t \in T$$

Constraints have been put in place, and we are making sure that each excursion starts and ends in the home base as much as possible. 0. Constraints, and ensure that each customer location has one in-going and one out-going connection if a sales person pitches at least one project linked to the customer on a given tour; alternatively, ensure that the customer location does not have any connections if the sales representative does not pitch any projects related to the client. The use of constraints helps to ensure that the full potential of the available trip durations is used. Additional limits on the deletion of subtours are required; but, for the purpose of brevity, we will not present them here. If the appropriate input parameters are supplied, optimal schedules for problems of a realistically sized may be constructed in a reasonable period of time by using commercial MIP solvers. This is granted that the parameters are accurate.

Table 4.1: Decision variables and input parameters

T	Set of tours.
C	Set of customer locations.
K	Set of projects.
K_c	Set of projects associated with customer c.
η^{max}	Maximum tour length (time).
τ_c^{visit}	Duration of a visit at customer c.
τ_{ij}^{travel}	Travel time between a pair of locations i and j.
Δp_k	Uplift generated by visiting customer c to discuss project k.
χ_k	Profit of winning craft k.

$y_{ckt} \in \{0, 1\}$	Indicates if customer c is visited on tour t regarding project k.
$x_{ijt} \in \{0, 1\}$	Indicates if travel segment between locations i and j is scheduled on tour t.

Despite the fact that the formulation presented above illustrates a very easy example of the problem, our model might be improved to take into account more intricate conditions. To begin, the daily activities of the sales team are now planned out using an online scheduling tool. This indicates that a strategy for the sales force is created each day in preparation for the next day. Nevertheless, our approach may naturally be extended to a planning horizon that covers more than one period at a time (days). The second presumption we make is that all sales reps are the same, both in terms of the preferences they have and the abilities they possess at the time.

On the other hand, in applications that take place in the real world, the uplifts can be depending on the person who is carrying out the visit. If we had access to historical data on the level of the sales person, we would be able to alter our model so that it takes into consideration the various kinds of promotions that are available. Third, we feel that it is appropriate to engage in conversation with customers at any point over the course of the day. In point of fact, despite appearances, it's possible that there are restrictions on the hours during which a customer may be reached. In order to take into account these limitations, we may formulate the issue as one requiring team orienteering and then find a solution to it by making use of time slots.

4.3 ANALYTICAL MODELING

In order to construct a schedule that allows the sales representatives to make the most efficient use of their time, the planner has to be aware of the uplift values associated with each project. We are unable to build a machine learning model to predict these uplift values since they cannot be observed. This prevents us from doing so. We are unable to make advantage of machine learning as a direct consequence of this. On the other hand, we may determine how likely it is that a certain project would be successful by instructing a predictive model to gain knowledge from the past and then feeding it the relevant information as its input. Following that, this prediction model will be used as a component in the uplift approximation method that will be presented in the following steps. In the following, we will begin by presenting the data that is available,

as well as our strategy for the engineering of features. After that, we will provide a variety of the machine learning approaches that we make use of, as well as analyze their individual performances, in order to choose the machine learning method that will offer the most precise estimate of the subsequent uplift, which will be discussed in Section 4.5.

4.3.1 Data set

Our affiliated company, DAW, has access to a database that details a variety of construction projects that were carried out in Germany between the dates of January 2015 and May 2017; the database may be seen here. Figure 4.2 provides a more in-depth breakdown of the relational data structure that lies under the surface.

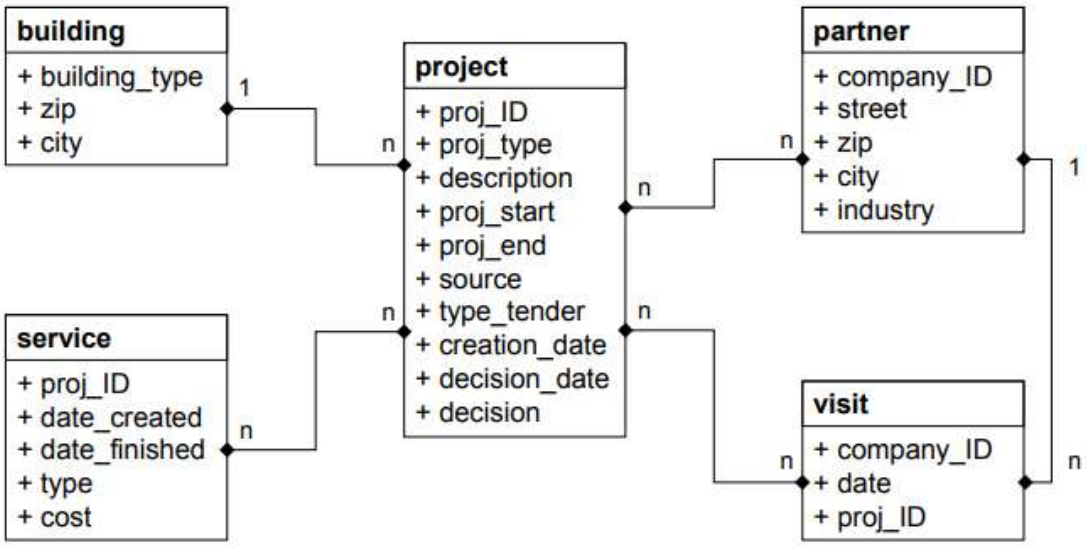


Figure 4.2: Overview of the existing data structure

Source: Data-driven Operations Management Data Collection and Processing through by M.Sc. Jan Maximilian Meller, In November 2019

- **Building Table:** It should come as no surprise that a number of minor development initiatives, such as painting a building's interior as well as outside, may be carried out concurrently with a more extensive development project. As a consequence of this, one of the tables provides information about the construction project in a more general sense, such as the category it falls under or the location it may be found in.

- **Project Table:** In this main table of the database, information is stored on several work packages that are connected to extensive development activities. As a result of the fact that these work packages will be the foundation of our future research endeavors, we shall refer to them as "projects" throughout the remainder of this section. Each entry in the database contains a description and classification of the kind of project being worked on (for example, painting works, plastering works, and repairs), information on timing such as the start and finish dates of construction, and characteristics of the project assignment (for example, public tender, direct assignment). Additionally, the eventual decision about the success or failure of a project is documented for future reference.
- **Partners Table:** It is common practice for large projects to need the involvement of a number of distinct partners, some of whom may also collaborate with one another on a number of other projects. The information that has been preserved on these firms includes specifics such as their locations, as well as the business sector in which they operate and the particular function that they play in the successful execution of a certain project.
- **Company Visits Table:** The exchanges that have taken place in the past between the company and its partners are documented in a fourth table. Every entry in the table of data represents a different appointment that will take place on a certain day. Some of these visits have access to more, more specific information on the project that was discussed, but not all of them do. This information is accessible for some of these visits.
- **Services Table:** Last but not least, our partner company, DAW, offers a wide range of services to potential partner companies. These services include, among other things, the provision of color samples and a tailored color advisory. Because these services are often adapted to the requirements of a certain project, it is more suitable to delegate responsibility for them to the specific project rather than to the partner business. The information regarding the completed services includes not only the kind of service that was done but also the date when it was performed as well as the value, in monetary terms, of the services that were supplied.
- **Data Cleaning and Processing:** On this collection of unprocessed data, the following essential data cleaning techniques were carried out. The majority of the 46,913 projects that were taken into consideration did not get a definitive verdict once the evaluation process was completed. We were able to deduce the choice for some of them by supposing that projects that were not granted funding within the stipulated time limit of 365 days had been scrapped. This gave us the ability to infer

the conclusion for some of them. After that point, the unassigned projects that were still being worked on were scrapped altogether.

After that, we got started on the process of eliminating records that were already present in the database. As can be seen in Figure 4.3, the outcomes of 2,929 of the 8,444 projects that are still active are decided in a time period that is much shorter than one week after the record is first submitted into the system. This is the case for the majority of the projects that are still active. In addition, out of these 2,929 efforts that were selected on the spot, only 73 are judged to be unsuccessful attempts. According to the findings of this study, a significant share of projects are settled upon – and often won – during the first encounter with the customer. These things are not included in our data set since the projects in question are never feasible possibilities for possible site visits and, as a consequence, would cause an unnecessary bias to be introduced into our model. As a result, our model does not take these things into consideration. The final data collection for the subsequent investigations included information on a total of 5, 515 projects, and DAW had been given a supply contract in 3, 828 of those instances.

In order to extract from the data the values of the project-specific uplift parameter pk, we must first generate some estimates for the probability p(k) of successfully finishing project k, where k is specified by some feature vector k. These estimates will then be used to calculate the values of pk. This possibility is clearly contingent on a broad range of other conditions, some of which are within our control (for example, the number of sales representatives' visits), and some of which are not (for example, the nature of the project, the partners participating, or the kind of project assignment). Figure 4.4 provides a timeline that illustrates an illustration of such a decision-making process as an example. The organization decides how much face-to-face sales effort (also known as the number of times a sales representative would visit a partner company in person) to put into a potential customer, also known as a lead. Additional services, on the other hand, are often requested by the partners and are thus considered exogenous.

We make use of supervised machine learning in order to develop a model that is able to provide predictions p(k) given a feature vector k. This is accomplished by using machine learning. In supervised machine learning, we train a model by providing it with a large quantity of historical data. This data consists of pairings of a feature vector with the information about whether the project was successful or unsuccessful. Our goal is to discover a relational structure that exists between the many different pieces of information that are included in the data. Next, we will present an explanation of

how we extract the features given the raw data, and then continue to train and evaluate various different machine learning algorithms. This will follow the discussion from the previous section.

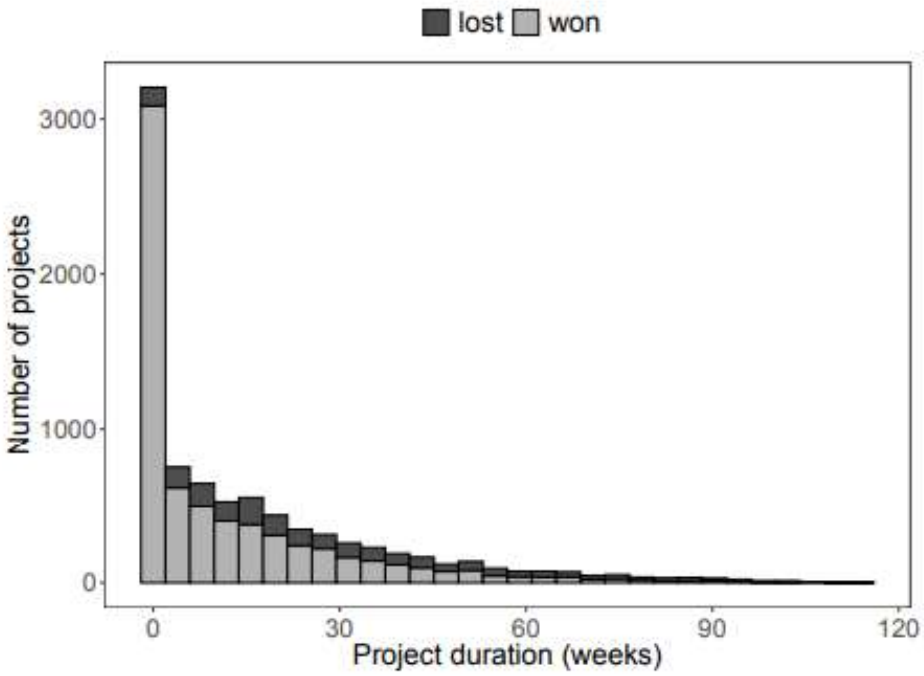


Figure 4.3: Waiting period before a decision is made in the system

Source: *Data-driven Operations Management Data Collection and Processing through by M.Sc. Jan Maximilian Meller, In November 2019*

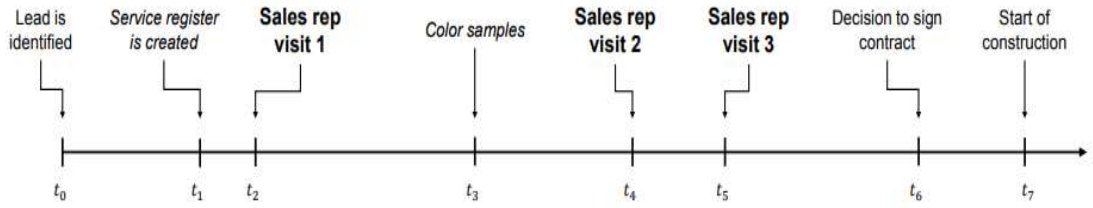


Figure 4.4: Timeline of a project illustrated

Source: *Data-driven Operations Management Data Collection and Processing through by M.Sc. Jan Maximilian Meller, In November 2019*

4.3.2 Enhancement engineering

In reference to the data structure that was described in section 4.3.1, it is evident that some of the data characteristics may be directly used as features (for instance, the information that is unique to the project and the static building). However, in order to make use of the information that is included in other entities (such as partners, services, or visitors), the data must first undergo the right transformations. The following features, which are direct ones, were taken into account:

- The kind of structure, which may range from an apartment complex to an office building, a church to a hotel, and so on.
- The nature of the endeavor that is being carried out, whether it one of reconstruction, renovation,
- Painting the inside and exterior of a building, installing thermal insulation, and other chores of a similar kind are examples of work packages.
- Leads that might be generated from several sources, such as architects, property owners, and general contractors, amongst others.
- The selection process may be done in a variety of ways, such as by going straight to the source, conducting a public or regional tender, and so on.

These essential components serve as a record of the project's consistent and independent qualities. A significant amount of effort was put into the designing of the features so that dynamic relationships, in addition to interdependencies, may be included. These activities will be explored in further detail later down.

4.3.2.1 Entity-Relation Summaries' characteristics

In order to generate features that are of more significance, we made some straightforward adjustments to the data in accordance with the comprehensive technique that Kanter and Veeramachaneni detail in their article. In order to get started, we carried out a count operation on the partner table in order to get an accurate count of the total number of partner companies that took part in the event. During the second stage of the procedure, a tally was taken of the number of times that each partner company has collaborated with one another on past projects. Following that, we computed the mean value of this number across all of the partner companies that were participating to a certain project. Figure 4.5 depicts the process that must be followed in order to detect this attribute correctly.

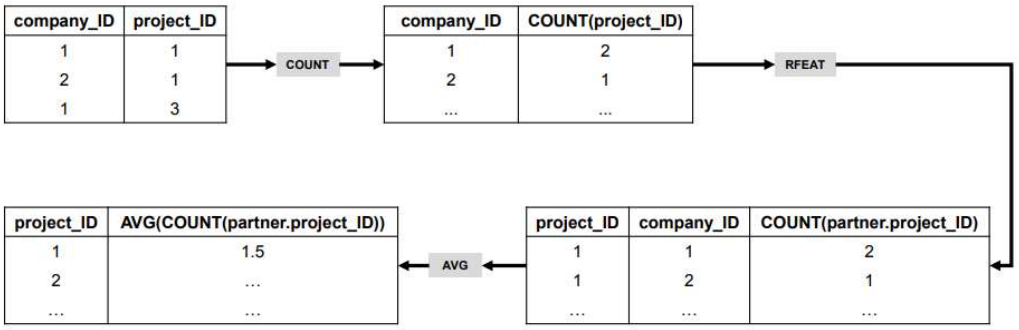


Figure 4.5: Exemplary calculation of the average number of historical projects with involved partner companies for a particular project

Source: *Data-driven Operations Management Data Collection and Processing through by M.Sc. Jan Maximilian Meller, In November 2019*

In addition, we determined the overall value of the additional services that were contributed to each project, and we recognized this as a quality that is exclusive to the process of working collaboratively. In the end, we decided to add binary features that specified whether or not a service register had been established as well as whether or not new services had been carried out.

4.3.2.2 Regularity of client visits connected to projects

The number of visits that have been made to the different participating partner companies over the course of the preceding t weeks prior to the decision being made on the project is one set of characteristics that stands out as being particularly notable. The first thing that has to be done in order to determine this number is to allocate a particular visit to a certain project. The data. Visits table provides a one-of-a-kind project identifier for visits that are exclusive to a project, despite the fact that its inclusion is voluntary. This is the case even though the inclusion of this identification is not required. In the event that such a project identification is stored, it will be able to allocate the whole of the visit to the project in question (weight w = 1); this will be done in the event that such a project identification is kept. For visits on the other hand that do not contain a particular project identification, we make the implicit assumption that any and all prospective partnerships with partner i were discussed, and we assign a weight to the visit that is equal to one more than the Ki value for each individual project. This gives visits that do not include a specific project identification a weight

that is one point higher than the Ki value. The actions that need to be taken are shown in Figure 4.6.

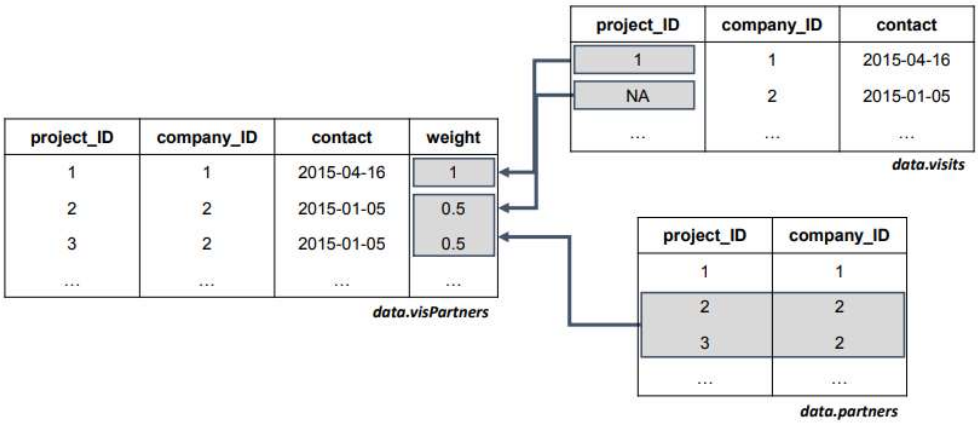


Figure 4.6: The weighted number of visits is calculated via feature engineering

Source: *Data-driven Operations Management Data Collection and Processing through by M.Sc. Jan Maximilian Meller, In November 2019*

We are operating on the assumption that the timing of these visits, in addition to the total number of visits, is an important factor in determining the chance of success. We have taken the number of visitors from the past and averaged them out across a wide range of various time periods, ranging from one week to two years, so that we can take into consideration the influence that time has on the data. Following that, each of these aggregations is then included into the design as its own unique component. Because of these qualities, there will unavoidably be some degree of multicollinearity present in the data. Although this presents a difficulty in explanatory modeling, where the objective is to comprehend the values of the coefficients, the major concentration of our efforts is directed at enhancing our capacity to produce accurate forecasts. In particular, algorithms for machine learning are considered robust due to their capacity to take multicollinearity into account. This is a characteristic that has contributed to their widespread adoption.

4.3.3 Models and Instruction

We put a large number of different models of prediction through their paces and evaluate how well they do their jobs in order to get a sense of how likely it is that we

will be given a project. In particular, we make use of both white-box approaches and black-box models. White-box techniques are methods in which the structure of the model and its fitted parameters can be interpreted by a human decision-maker. Black-box models often provide greater prediction accuracy at the price of limited interpretability. To put it another way, we make use of both black-box models and white-box approaches.

Our white-box baseline classifier is going to be based on logistic regression, which is what we rely on. The previously stated multicollinearity of covariates and separation, which takes place when a linear combination of characteristics is highly predictive of the result, are two typical challenges that arise in applied logistic regression. Separation takes place when a linear combination of characteristics is strongly predictive of the result. The factor of separation is also a contributor to the issue. As a result of this, we make use of a Bayesian variation of the logistic regression methodology. In spite of the fact that it is simple to comprehend, logistic regression is not as effective as more involved black-box models in the majority of practice-relevant contexts. This is due to the fact that logistic regression works on the premise that there is a steady and linear link between the variables that are being studied and the findings that they provide.

As a result of this, we evaluate its effectiveness by comparing it to the results of three different black-box models. The theory that was developed at Vapnik is the foundation for the first of these models, which is referred to as a support vector machine (SVM). The concept of statistical learning serves as the foundation for support vector machines. These machines accomplish their task by inserting hyperplanes into the feature space in order to effectively segregate the output classes of interest. One of the most important benefits of support vector machines is the large degree to which they are resistant to overfitting. This is one of the most important advantages.

In addition to this, we apply a model that is what is known as an artificial neural network (ANN). ANNs make use of non-linear functions, which are then applied to linear information combinations in order to provide accurate predictions. They are an efficient method of instruction that can be adapted to a number of settings and are thus very versatile in their use. In the end, we train a classification model called random forest, which is based on an ensemble version of decision trees. Because it has been shown to operate well across a wide range of use scenarios, this is the paradigm that we have decided to use in our work.

When we are training the models, we use a subsample that is equivalent to eighty percent of the overall data. The remaining twenty percent of the data have been split into two samples: the first sample is a test sample, and it will be used to evaluate the performance of the respective classification (15 percent). The second sample is an evaluation sample, and it will be used to evaluate our prescriptive scheduling model, which will be covered in Section 4.7. In order to fine-tune the models, we use 10-fold cross-validation inside the training data. This allows us to get the most accurate results possible.

4.3.4 Appraisal and Selection of Models

In order to evaluate the efficacy of the four different models that were researched, we offer the receiver operating characteristic (ROC), the associated area under the curve (AUC), the F1 score, and the phi coefficient (). A graphical depiction of the predictive performance of a binary prediction model is known as the area under the receiver operating characteristic curve (ROC). When comparing various classifiers, it is usual practice to calculate the area under the curve, which is also known as the AUC. This is because calculating the AUC enables one to synthesize the information contained in ROC curves in a single numerical metric. The F1 score is a statistic for determining how accurate a test is, and it takes into account both the precision and recall of the exam. Precision is the ratio of genuine positive predictions to the total number of positive samples, and recall is the ratio of genuine positive predictions to the total number of positive samples. Both of these ratios are related to the accuracy of a test.

In addition to the area under the curve (AUC) and the F1 statistic, the phi coefficient, which is also known as the Matthews correlation coefficient, is often acknowledged as being among the most accurate single-number measures of classification skill (Powers, 2011). This metric also goes by the names rho correlation coefficient and Matthews correlation coefficient, amongst others. This is particularly attributable to the fact that it is resistant to inequities that span different class levels. The coefficient is equal to the Pearson correlation coefficient when dealing with issues involving binary classification. The Pearson correlation coefficient evaluates the degree of correlation that exists between the actual and predicted results of binary classification and may offer values ranging from -1 to +1. When used as a metric for determining the efficacy of machine learning models, only the positive values of are relevant to take into consideration: A prediction with a value of +1 is seen to be absolutely accurate, in contrast to a forecast with a value of 0, which is thought to be completely arbitrary.

Figure 4.7 displays the receiver operating characteristic (ROC) curves for the four models that were tested. The Random Forest classifier works substantially better than its competitors throughout a significant percentage of the spectrum that was taken into considered, while the neural network operates most effectively within a specific section of the configurations that were taken into consideration. This is something that we are able to verify. The numerical performance metrics, which are shown in Table 4.2, provide credence to this assessment and confirm its conclusions. Because the AUC, F1, and scores that are produced by the Random Forest method are higher than those that are produced by the other techniques, we have chosen to utilize this model as the basis for our subsequent process for estimating uplift based on this model. This process will be described in more detail below.

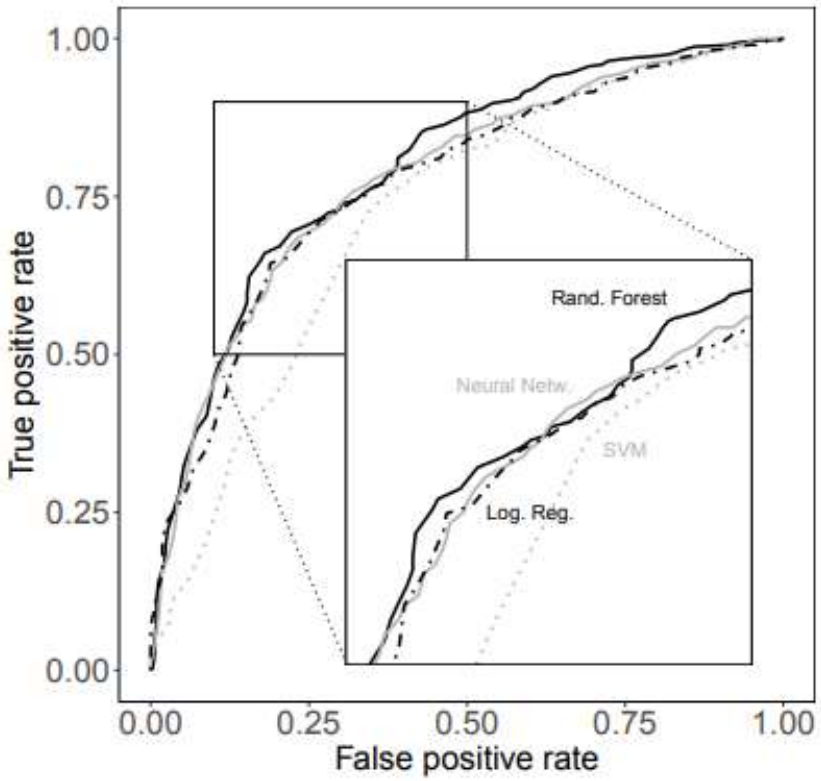


Figure 4.7: Comparison of the receiving-operational properties of the models

Source: *Data-driven Operations Management Data Collection and Processing through by M.Sc. Jan Maximilian Meller, In November 2019*

Table 4.2: Performance of classifiers is compared

	AUC	F1	ϕ
Support Vector Machine	0.72	0.82	0.30
Logistic Regression	0.77	0.82	0.34
Neural Network	0.78	0.82	0.37
Random Forest	**0.80**	**0.83**	**0.42**

4.4 APPROXIMATION OF UPLIFT

The specific values of a visit at a particular customer's location are a crucial input that need to be taken into consideration in order to find a solution to the scheduling problem that is mentioned in Section 4.3. Figure 4.8 depicts several exemplary trajectories for the probability p that was computed from the data. The Random Forest model discussed in Section 4.4 was used to generate these trajectories, which were given for your convenience. Evidently, the planner is not going to be able to optimize the schedules of her sales representatives based just on the data about the probability of success: Without knowing the actual "uplift" of an additional visit, one cannot make a trade-off between driving further distances for a large increase in the success probability of a single project and increasing the probabilities of multiple projects in the surroundings by a smaller margin (for example, compare the value, or the uplift, of the first visit to the one of the second and third visits in figure 4.8). This is because the uplift of the first visit is greater than the uplift of the second and third visits. Take, for instance, a comparison of the value, or the boost, that the first visit provided.

As a result, we are not concerned with the predicted probability p, but rather we are concentrating on the uplift pk that is connected with each probable visit. Calculating such values as the uplift is a particularly tough task due to the fact that the actual uplift cannot be observed. We will now explain our methods for estimating such uplifts in the following sections, taking into consideration the prediction model that was covered in Section 4.4.

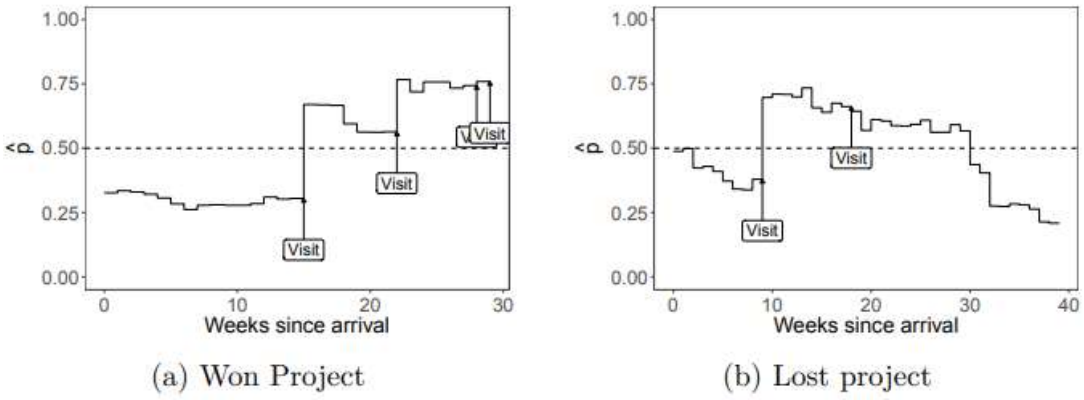


Figure 4.8: Examples of success probability trajectories

Source: *Data-driven Operations Management Data Collection and Processing through by M.Sc. Jan Maximilian Meller, In November 2019*

For the most part, accurate modeling of uplift value requires data from not one but two distinct groups: the treatment group and the control group. Due to the fact that the action variable, which in this instance is the number of customer visits, is numerical, we are unable to split our data into separate groups. This is due of the attribute of the action variable that allows for its numerical representation. Instead, we construct a "synthetic" treatment data set by adding fictitious client visits, and then we recalculate the success probability using the prediction model.

After that, we will be able to compute the uplift by comparing the probabilities of the synthetic data set to those of the original data set and then determining the difference between the two sets. This line of thinking is supported by prior research, including that which was conducted by Foster et al. and van de Geer et al. Our feature vector z may be enhanced if we understood it to be the union of a vector of endogenous "action features" a, which relates to the number of customer visits, and a vector of external characteristics z. This would allow us to better grasp what our feature vector z really represents. The accuracy of our feature vector will improve as a result of this interpretation. As a direct consequence of this, the vector that contains our forecasts is now represented as p() = pp=a, z). Take into consideration the fact that we use the very same prediction models that were covered in the part that came before this one. First, using the feature data set and the trained prediction models discussed in Section 4.4.4, we generate the fake "treatment" data set in which we raise ak by one. This step is

described in more detail in the previous section. The next paragraph will go into more information about this procedure that must be taken.

Figure 4.9 is a graphical representation of the distribution of the calculated uplift values, and it demonstrates how these values change depending on the number of previous visits ak. Finding that the number of visits, on average, has a positive impact on the chance of getting granted a project (p) is something that shouldn't come as much of a surprise to anybody. In addition to this, we see that the median value of the first visit is much higher when compared to the values of subsequent visits. This effect may be somewhat explained by the fact that the prior probabilities p(ak, zk) are rising with the number of visits, which eventually results in a decreased overall potential for uplift in the population. On the other hand, the number of visits that occurred before a visit is not a sufficient explanation for the boost that the visit offers. The boost that the visit delivers is what really matters. This result illustrates that other qualities capture a significant percentage of the information and offer helpful information that can be used to identify customers who are likely to be convinced. Additionally, this conclusion demonstrates that this information can be used to identify customers who are likely to be persuaded.

In order to acquire a deeper comprehension of the dependability of our technique, we make artificial modifications to the primary data set and then investigate the consequences these changes have on the distributional characteristics of the data. More specifically, we are investigating how the presence of this factor impacts things. In particular, we look at the discrepancies that exist between the distributions of projected probability for the original data and the distributions for the synthetically adjusted number of visits. This allows us to better understand the disparities between the two. In the circumstances when there was no visit to the area, the estimated probabilities in the original data set and the synthetic data set diverge, as shown in Figure 4.10.

We propose an explanation for this observation by making the assumption that the projects in the first data set that did not contain any visits are fundamentally unique from those that did include visits. This allows us to say that the projects that did involve visits are more likely to have been successful. The right panel demonstrates that regardless of whether there was one visit or several visits, the empirical CDFs exhibited behavior that was highly comparable to one another. After taking into consideration the many structural aspects of projects that were carried out without visits, we have come to the conclusion that the synthetic treatment approach does not systematically modify

the distributional qualities of success likelihood in a local region. This is the result that we have reached. As a consequence of this, we have arrived to the conclusion that our approach is able to provide estimations of uplift that are correct.

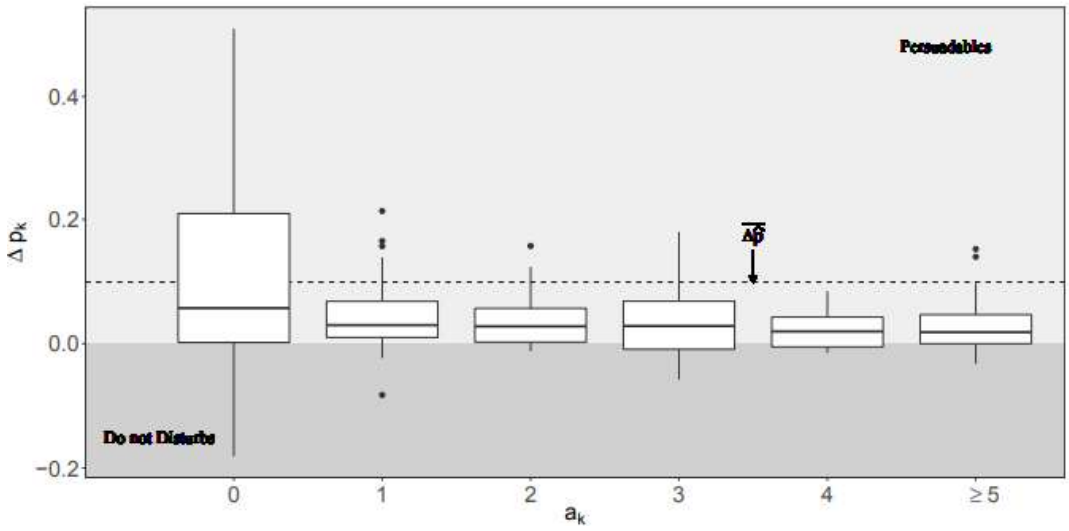


Figure 4.9: Depending on the quantity of earlier visits, uplift values p

Source: *Data-driven Operations Management Data Collection and Processing through by M.Sc. Jan Maximilian Meller, In November 2019*

This uplift information, in and of itself, already provides critical insights in a variety of different situations, and it can be used to do things like prioritize projects where it seems that additional effort would be advantageous (cf. van de Geer et al., 2018). This is especially true for the situation of sales operations in the insurance and financial sectors, which often include making outbound calls to potential customers. Sales activities like these are a common element of the sales process. However, in our environment, the effort, or the total time required to visit a particular client, is significantly influenced by the travel time of the sales person, and as a consequence, a great deal of reliance is placed on the location of the customer in addition to the locations of other customers. To put it another way, the total amount of time required to visit a particular client is directly related to the amount of time it takes the sales person to go there. As a consequence of this, we make use of a model for the scheduling of our sales force that takes into account the trade-off between improved uplifts and the amount of time that is required to visit each customer.

4.5 PRESCRIPTIVE SALES FORCE SCHEDULING WITH FORECAST UNCERTAINTY

We employ the uplift projections that are marked by pk in our prescriptive sales force scheduling system, while at the same time taking into account the accuracy of these predictions. When we talk about the "quality" of these forecasts, we are referring to the fact that they are susceptible to a certain level of uncertainty and might be linked to a larger or lesser margin of error. If we take it for granted that we have a flawless prediction model that consistently generates accurate estimates of uplift, then we should have no trouble resolving the optimization problem outlined in Section 4.3 by using the predictions pk as inputs to the optimization model. This is because we will have a model that is capable of making accurate predictions. If one did not have any information regarding the increase, the most prudent thing to do would be to maximize the overall amount of probable profits. Using a consistent uplift of pk for each project would be akin to finding a solution to the optimization challenge described in Section 4.3.

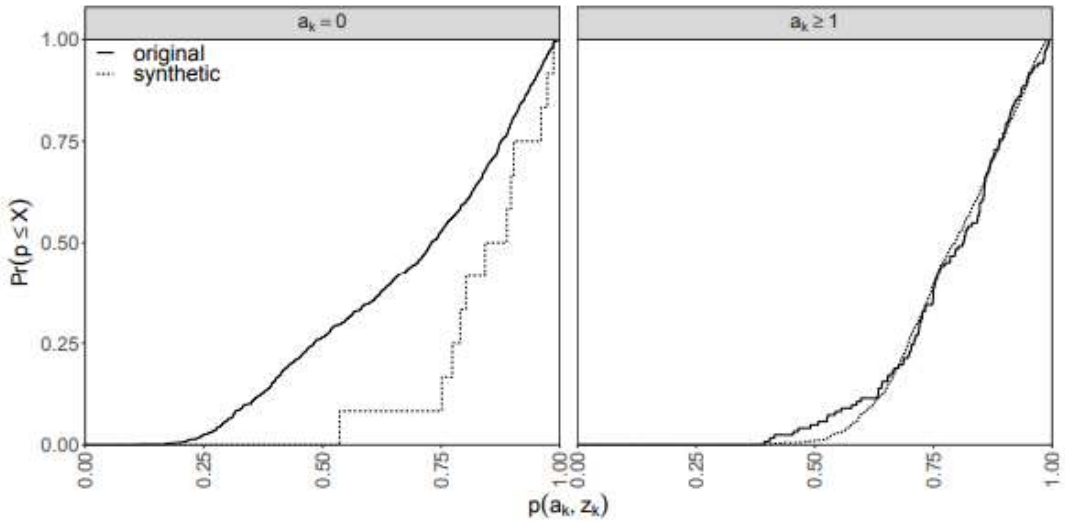


Figure 4.10: Comparison of the distribution of predicted probabilities. Individual panels correspond to number of prior visits a

Source: *Data-driven Operations Management Data Collection and Processing through by M.Sc. Jan Maximilian Meller, In November 2019*

The objective function of the optimization problem is simplified to when this happens:

$$\max v(\pi) = \sum_{c \in C} \sum_{t \in T} \sum_{k \in K_c} \alpha \chi_k y_{ckt}$$

In actual life, we are going to run across situations that are anywhere in the middle of these two extremes, such as the following: There is always going to be some room for mistake, regardless matter how sophisticated the machine learning prediction models are or how much data is used. If we choose to ignore this uncertainty, it might lead us to arrange diversions to see clients whose uplifts have been exaggerated, while at the same time preventing us from booking visits to customers whose uplift projections are too low.

It should come as no surprise that the quality of the forecast should be a factor in how much weight we place on the uplift estimates, and how much weight we should give to the profitability of the projects when deciding how to allocate our resources and how to rank our priorities. The Hodges-Lehmann criteria, which originate in the discipline of decision theory, provide a straightforward method for formally expressing the choice that must be made between the two distinct corner circumstances. It appears to indicate that a decision-maker will need to make a trade-off between a cautious worst-case approach and an optimistic alternative that maximizes expected value when making a choice in the face of inaccurate probability estimates. This is because the decision-maker will be making a choice in the face of flawed probability estimates.

For our purposes, the worst-case scenario is analogous to discarding the findings of the uplift prediction model. On the other hand, the most effective tactic for maximizing the value that is expected to be realized is to place one hundred percent of one's confidence in the results of the uplift prediction model. A linear weighting is assigned to these two objectives, and the value of that weighting is decided by a parameter marked by, which the person making the choice is responsible for determining. Following this line of reasoning, we are able to formulate an expression for the objective function as:

$$\max v(\pi_\lambda) = (1-\lambda) \sum_{c \in C} \sum_{t \in T} \sum_{k \in K_c} \overline{\Delta \hat{p}} \chi_k y_{ckt} + \lambda \sum_{c \in C} \sum_{t \in T} \sum_{k \in K_c} \Delta \hat{p}_k \chi_k y_{ckt},$$

where

$$\overline{\Delta \hat{p}} = \frac{1}{|K|} \sum_{k \in K} \Delta \hat{p}_k.$$

When the value of is equal to zero, we come up with the prudent worst-case policy, and when the value of is equal to one, we come up with the policy that maximizes the expected value by having entire trust in the uplift predictions. When is equal to one, we come up with the policy that maximizes the anticipated value. In order to standardize the scale, we decided to set p to be equal to the projected uplift on average for all of the projects that were included in the test.

Whoever is in charge of making the decision will unavoidably be faced with the difficulty of deciding what to do. According to our line of reasoning, selecting either the model quality or as your value for this parameter appears to be the most reasonable choice. Justifying this concept is made possible with the use of a simple linear regression model as a basis for comparison: The uplift predictions, represented by p, are considered as the independent variables, whereas the unknown true uplifts are interpreted as the dependent factors in this study. It is feasible to estimate the parameters of the regression, and the following are the results:

$$\hat{\alpha} = \overline{\mathbf{\Delta p}} - \hat{\beta}\overline{\mathbf{\Delta}\hat{p}} \qquad \text{and} \qquad \hat{\beta} = \rho_{\mathbf{\Delta}\hat{\mathbf{p}}\mathbf{\Delta p}} \frac{s_{\mathbf{\Delta p}}}{s_{\mathbf{\Delta}\hat{\mathbf{p}}}}.$$

Rich machine learning models, such as gradient boosting, in general have very small bias, which indicates that p is comparable to p. One example of such a model is the gradient boosting algorithm. The fact that our metric for measuring the quality of the prediction model,, is similar to the Pearson correlation coefficient that exists between the predicted values and the actual values enables us to make use of the fact that it for the purpose of estimating, we can make use of the fact that we can. Since there are no inconsistencies in the real units that serve as the foundation for the comparison, there is no need for us to adjust the coefficient of correlation using the ratio of standard deviations as a scaling factor. In addition, there is no need for us to do so.

$$\hat{\alpha} = (1 - \hat{\beta})\overline{\Delta\hat{p}} \qquad \text{and} \qquad \hat{\beta} = \phi.$$

When these estimations are taken into consideration, the target function of the model for scheduling the sales force may be stated as follows:

$$\max v(\pi_\lambda) = \sum_{c \in C} \sum_{t \in T} \sum_{k \in K_c} \chi_k \left[(1 - \phi)\, \overline{\Delta\hat{p}} + \phi \Delta\hat{p}_k \right] y_{ckt}.$$

The following is one interpretation that may be made on the thinking that went into developing the quality-adjusted forecasts: It is estimated that giving a presentation on project k at location i will increase the chance of being awarded the project by an amount equal to the average projected uplift of all projects included in the test data. This will result in a greater likelihood of being awarded the project. After that, we modify this baseline such that it leans more toward the forecasts dependent on how accurate the predictive model is proving to be.

4.6 COMPUTATIONAL EVALUATION

This section presents the findings obtained from performing exhaustive numerical calculations for the objectives of determining the robustness of our approach, evaluating the effectiveness of our prescriptive scheduling strategy in comparison to relevant benchmark policies, and developing additional structural and management insights. The calculations were carried out in order to achieve these goals. In the course of these investigations, we make use of the data that DAW has kindly supplied, which are described in Section 4.4. In section 4.6.1, we go through the steps that make up the first half of our review method.

In the following section, Section 4.6.2, we will evaluate the value of our prescriptive policy in comparison to the value of a predictive policy (which ignores the quality of the prediction model and assumes perfect uplift information) and the value of a zero-information policy (which does not take into consideration uplift information) for a base case scenario. This evaluation will be done for the purpose of determining whether or not our prescriptive policy is superior to the other two policies. This comparison is being conducted so that a decision may be made about which plan offers the most value. In the sections that follow, we will explore the ways in which various aspects, such as the heterogeneity of the profitability of the projects (Section 4.6.3), the sales force capacity (Section 4.6.4), and the quality of the prediction model (Section 4.6.5), influence the value of our prescriptive approach.

4.6.1 Evaluation Method

As part of our quantitative study, we build and evaluate sales force schedules for specific sales forces, each of which is situated in one of Germany's ten different home bases. These home bases are located throughout the country. Each sales force is comprised of a certain quantity of sales representatives, or Ts, who are tasked with

managing a particular sales zone. The evaluation process is outlined in Figure 4.11, which may be seen below.

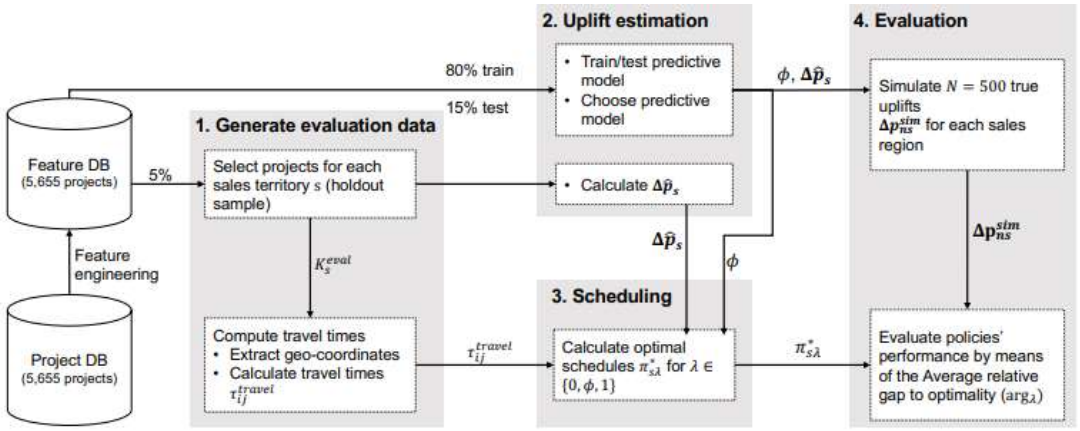


Figure 4.11: Evaluation process

Source: *Data-driven Operations Management Data Collection and Processing through by M.Sc. Jan Maximilian Meller, In November 2019*

We will begin the procedure by gathering the data that will be required for the analysis in the very first step of the process. The data that will be used for our evaluation will contain the characteristics of a holdout sample chosen at random from DAW's project data base. This sample will consist of five hundred projects. In this paper, Section 4.4 provides an explanation of these characteristics. These characteristics were not used in either the training or the testing of the prediction models. Before we began to train and test the prediction models, we divided the holdout sample into projects and assigned each one a sales region to assess as part of an evaluation set.

By using Google Maps API9, we were able to get the geocoordinates of the home bases 0s as well as all of the customer locations for the 500 projects. After that, we got the total journey times Travel between any two sites i, j in sales region s = 1,..., 10 by using the HERE API 10.

In the second step, we generated an estimate for the uplifts by looking at the characteristics of the individual projects that make up each sales area. This allowed us to have a better understanding of how much each territory would be affected. We were successful in accomplishing this goal by adhering to the procedures described in

Section 4.5, which are based on the random forest model. When applied to the validation data, this model produced the highest quality prediction performance that was achievable.

In the third stage, we solved the scheduling problem (see Section 4.6) for each sales region s and P t0,, 1u by using the expected uplifts ps (s 1,...10) obtained in the second phase. These uplifts were gained in the previous step. In order to do this, we took into consideration the quality of the model as well as the travel durations for ij. The phrasing of the reflects the prescriptive nature of our policy, which was covered in the part that came before this one. When = 1, a "predictive policy" is assumed, which implicitly presume precise predictions of uplift and maximizes the revenues that are predicted from any higher sales. This implies that accurate forecasts of uplift are necessary. In contrast, when = 0 is considered, uplift information is disregarded; the schedules that are produced as a consequence of this "zero-information policy" are intended to maximize the total profit that may be made from all of the projects combined. The study that will take place during this stage will result in the formulation of the optimal schedules *s (P t0,, 1u) for each sales area.

During the third phase, we will evaluate the degree of success that each of the three policies has had in being put into action. We are unable to conduct an exact comparison of the performance of the three policies since we do not have reliable information on the uplifts for the 500 projects. This prevents us from being able to make an appropriate comparison. We are going to continue in the following manner in order to circumvent this impediment and make certain that we are doing an objective comparison of performance: The use of simulation is what enables the development of N = 500 vectors of "true" uplift realizations psim sn for each sales region s. These vectors are generated for each sales region. We employ the Cholesky decomposition of the covariance matrix to ensure that the correlation between the estimated uplifts ps and the simulated actual uplifts psim sn is equal to. This allows us to be positive that the simulated real uplifts are accurate. Because of this, we are able to validate our hypothesis that the simulated real uplifts have the same impact as the calculated uplifts.

Then, we calculate vsnp sq for n = 1, 500, which is the additional expected profit made when the optimal schedule s of policy is applied, which was determined based on the uplift predictions ps, but the actual uplifts are psim sn. This is the case where the uplift forecasts were derived based on the uplift forecasts ps. The value that we use in the calculation of the extra anticipated profit is denoted by vsnp sq. As a performance

metric for the policy, we arrive at the average relative optimality gap (arg) by using vsnp and sq as our foundation.

$$\arg_{\lambda} = \frac{1}{S}\frac{1}{N}\sum_{s}\sum_{n}\left(1 - \frac{v_{sn}(\pi^{*}_{s\lambda})}{\hat{v}_{sn}}\right),$$

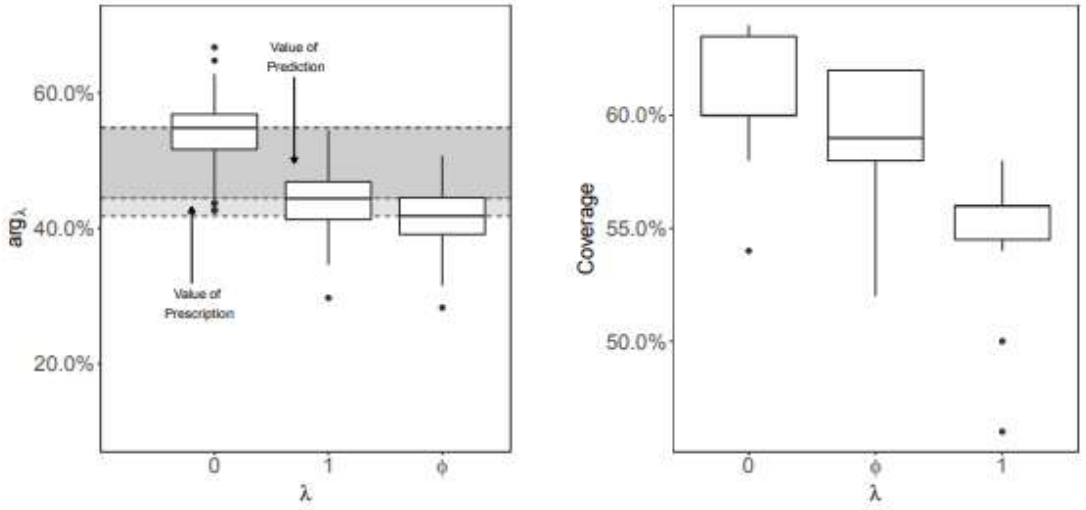
4.6.2 Base case value of the prescriptive policy

In the first stage of our inquiry, we are going to compare the predictive policy, the prescriptive policy, and the zero information policy to a base case scenario and see how well they perform. We take the results of our best prediction model, which had an accuracy of 0.42 (see Section 4.4), and we make the assumption that the size of the sales force in each sales region, which is represented by the notation |T|, is equal to 5. Unfortunately for us, our business partner DAW was unable to give us with specifics on the profitability of any one project. In the sake of this preliminary investigation, we shall proceed on the basis that the profits made by the different projects are equivalent to one another. This assumption will be represented by the notation k = 1 for all k € KEval s. In the next section, we will investigate the effect that the different types of profits have on the way the policies are carried out.

Image 4.12 (a) displays the arg for each approach, together with their distribution throughout all of the simulated runs presented in the image. This figure presents the results of all of the simulations. Both the predictive and prescriptive techniques provide substantially lower values of the arg when compared to the standard of having no prior information, which is denoted by the symbol = 0 in this context. This is quite evident and easy to observe. In addition, the prescriptive policy that takes into account the quality of the model (=) results in a performance that is slightly better than that of the predictive strategy (= 1); this is the case when comparing the two policies side by side. Within the context of this baseline situation, it is possible to assert that the presence of a potent predictive model (the "value of prediction") is accountable for the vast majority of the performance gain that was seen. When compared, the prescriptive method only delivers a negligible gain in total value, which we refer to as the "value of prescription" increase.

Figure 4.12 illustrates the relative coverage of the plans, which is often referred to as the proportional percentage of customers visited by each insurer. Since it optimizes the number of visits to be maximized in order to achieve maximum coverage even when

the profitability of all projects is the same, the fact that the zero-information benchmark achieves this outcome should not come as a surprise. The predictive policy, which takes into account just uplift estimates, ends up producing the lowest number of patient visits and, as a consequence, the smallest quantity of coverage. As a result of the fact that the prescriptive policy (=) makes a trade-off between the number of visits and the projects' uplifts based on the quality of the predictive model (= 0.42) – that is, how reliable the estimate of the uplifts are – it results in a coverage that is in between that of the zero-information policy and the predictive policy. This is due to the fact that the prescriptive policy makes a trade-off between the number of visits and the uplifts granted to the projects depending on the quality of the prediction model.



(a) Average relative gap to optimality (arg)

(b) Coverage

Figure 4.12: Profits that are homogeneous (|T| = 5, = 0.42)

Source: *Data-driven Operations Management Data Collection and Processing through by M.Sc. Jan Maximilian Meller, In November 2019*

When compared to the zero-information policy, the prescriptive strategy skips certain visits and "invests" more travel time in projects with greater projections of uplift. This is done in contrast to the zero-information approach, which skips all visits. However, it does so in a way that is more cautious than the predictive policy does, and this is because it takes into account the fact that the accuracy of the uplift projections cannot

be anticipated with full confidence. The fact that each initiative brings in the same amount of money makes these results less startling than they would otherwise be. However, we have seen that the results are accurate, and that the different rules result in different sales force schedules, which in turn result in varying degrees of success in relation to the goal. In the next section, we will analyze how these outcomes change when there is a difference in the profitability of the projects that are in issue.

4.6.3 Profit Heterogeneity's Impact

In this section, we will analyze how the varying potential profit margins of specific projects influence the relative success of the three techniques. Specifically, we will look at how these margins impact the effectiveness of the strategies. We draw a profit k from a (symmetric) triangular distribution with mode c = 1, support ra, bs, and width w = b a for each project k P KEval s in each sales region s. This distribution has mode c = 1, support ra, bs, and width w = b a. Since the scenario of consistent earnings was discussed in the prior part of this article, w = 0 is the appropriate value to use in this setting. In order to establish two distinct degrees of heterogeneity, we conduct experiments with varying quantities of support for the distribution, which are denoted by the notations w = 0.5 and w = 1. For the goal of guaranteeing that our research is reliable, we produce twenty distinct probable profit realizations—one for each degree of heterogeneity and sales area. This allows us to check whether or not our findings are accurate.

Figure 4.13 illustrates the outcomes of the policies applied to a number of different degrees of heterogeneity in the population. We discover that both the predictive and the prescriptive policies consistently outperform the zero-information benchmark (= 0q), and that the arg for both of these policies decreases as the profit heterogeneity grows. Additionally, we find that both of these policies frequently outperform the zero-information benchmark (= 0q) in the same way. It is abundantly clear that they are in a better position to select activities that will lead to high revenues and high uplifts.

Because of this modification, the effectiveness of the zero-information policy is enhanced for heterogeneity levels of w = 0.5 and w = 1, respectively. This is the case for both of these levels. This is because the policy has recently been changed to give more priority to projects that are anticipated to bring in a bigger amount of income. On the other hand, it does not take into consideration the uplift predictions, and as a consequence, it is conceivable that it will schedule visits for projects that have a high

profitability but a low uplift. As a direct consequence of this fact, the value of its arg is more than that of the predictive policy as well as the prescriptive policy combined. When there is a greater degree of fluctuation in profits, it is easier to see how adopting a prescriptive approach might be beneficial. Because of this criterion, there is a decreased likelihood that visits will be booked to high-profit projects that have forecasts of uplift that are implausibly high (and incorrect). When w is equal to zero in the first scenario, it is possible that a significant portion of the outperformance may be attributable to the capability of both the predictive and the prescriptive policies to make use of uplift forecasts. On the other hand, when there is a greater degree of variation in terms of profit, the value of the prescription increases, and it becomes more important to take into account the quality of the predictive model.

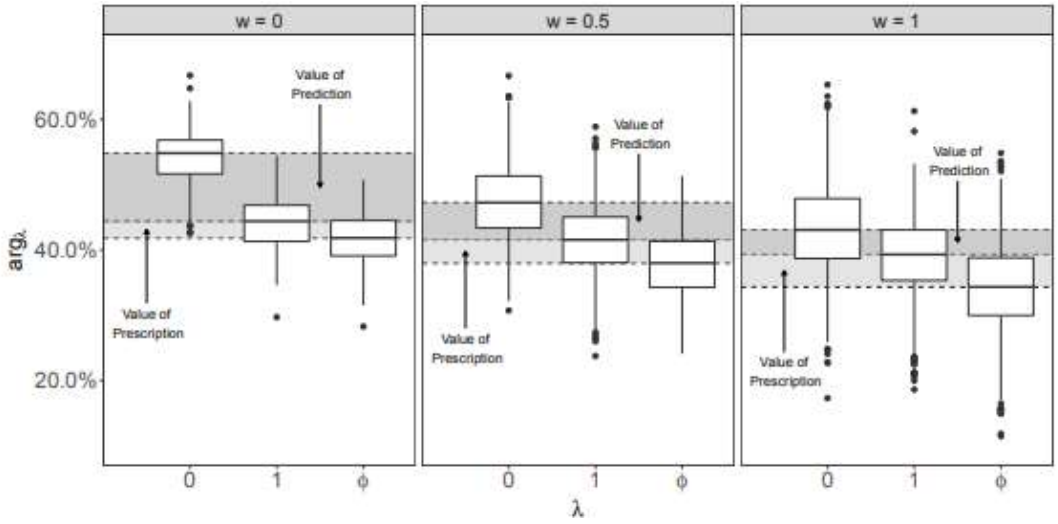


Figure 4.13: Average relative gap to optimality for varying levels of profit heterogeneity (|T| = 5, φ = 0.42)

Source: *Data-driven Operations Management Data Collection and Processing through by M.Sc. Jan Maximilian Meller, In November 2019*

4.6.4 The Impact of Sales Force Capacity

In this section, we will investigate the ways in which the three policies interact with one another in order to determine the extent to which the size of the sales force affects the effectiveness of the policies. We are going to do this by conducting an investigation on the influence that the size of the sales force has on the efficiency of the policies.

Because the size of a sales force is directly proportional to the amount of territory that it is able to cover, it is possible to determine the number of client visits that will be required by using the size of the sales force as a measuring stick. This is because the size of the sales force is directly related to the amount of territory that it is able to cover.

On the basis of a model quality of = 0.42 and varying degrees of profit heterogeneity, we discuss the effectiveness of the three methods. We do this by adjusting the amount of profit heterogeneity, as well as the size of the sales team in each area, which may range anywhere from one person all the way up to seven people depending on the particular region. Figure 4.14 demonstrates how the average revenue produced by each strategy differs depending on the size of the sales force and the degree to which profits originate from a range of sources. This variation in revenue may be attributed to the fact that different strategies target different customer demographics. This variation in arg might be because different strategies are focused at diverse types of customers, which would explain why there is so much diversity.

When there is a scarcity of capacity, the predictive and prescriptive policies are able to make greater use of the existing sales force capacity than the zero-information policy does. This is in contrast to the zero-information policy, which uses the information that is currently available. In contrast to this is the zero-information policy, which makes use of just the information that is now accessible. The zero-information policy, on the other hand, simply takes use of the information that is already available and is thus in contradiction to this method. As a result of the fact that the zero-information policy permits uninformed deployment of sales employees, the increase in performance that is associated with adding another sales rep is almost consistent. This is due to the zero-information policy that has been implemented. This is because the zero-information approach allows for the deployment of sales staff that are unaware of the goods they are selling. This is one of the reasons why this is the case.

On the other hand, both the predictive rules and the prescriptive rules demonstrate that the marginal values of adding more salespeople are decreasing. This is the case regardless of which rule is used. This is the case regardless of how correct the predictive policies may or may not be. When there is a significant amount of accessible sales force capacity, individual salespeople's performances begin to converge with those of the benchmark. This phenomenon is known as "sales force capacity convergence." This is because salespeople are meeting with a bigger number of customers, each of whom has a lower expected extra profit, which in turn generates a fall in the performance of

salespeople. The reason for this is related to the fact that salespeople are seeing a greater number of clients. This is due to the fact that expanding capacity eventually results in a rise in the total number of consumers who make purchases.

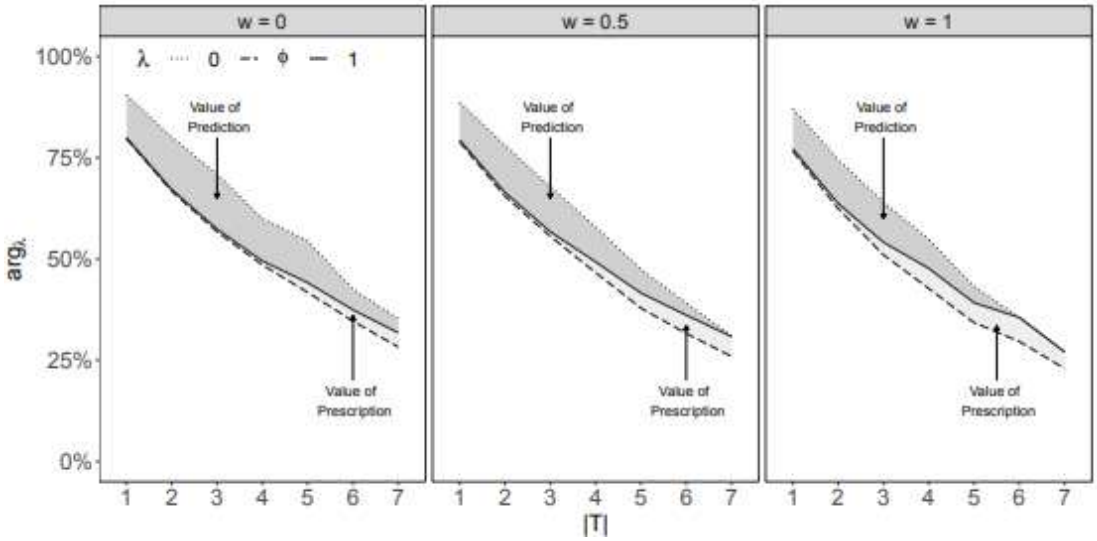


Figure 4.14: Average relative gap to optimality for varying levels of profit heterogeneity and sales force capacity (φ = 0.42)

Source: *Data-driven Operations Management Data Collection and Processing through by M.Sc. Jan Maximilian Meller, In November 2019*

When applied to low capacity levels, the prescriptive policy and the predictive policy both provide outcomes that are (almost) identical to one another. When capacity is at medium levels, the prescriptive policy leads to more performance than the predictive strategy, and the difference between the two policies grows as more capacity becomes available. a prescriptive policy leads to greater performance than a predictive strategy when capacity is at medium levels. It's possible that the various customer groups that each insurance policy chooses to survey may provide some light on the factors that contribute to these varying levels of performance. Figure 4.15 presents a graphical representation of the Jaccard coefficient of similarity for each and every conceivable combination of policies, with varying degrees of capacity and profit heterogeneity, and a tabular representation of the findings. To put it another way, the Jaccard coefficient is a measurement of the percentage of the total number of visits made by both policies to the total number of customers who are identical to those visited by both plans.

In other words, it compares the total number of visits made by both plans to the total number of customers who are the same as those visited by both plans. When capacity is at very low levels, both the predictive policy and the prescriptive policy will plan visits to a group of customers that is highly similar to the group being planned for. When w = 0.5 and w = 1, respectively, these clients are expected to provide high projected uplifts and high profit margins. The zero-information strategy, on the other hand, chooses a specific category of customers because it ignores the uplifts and focuses instead on the delicate balancing act that must be performed between profit margins and trip durations. This results in the selection of a different group of customers.

There is a difference in the decisions that are made between predictive and prescriptive policies when extra capacity becomes available: While the predictive policy continues to prioritize customers with (slightly) higher predicted uplifts or profits, the prescriptive policy protects against errors in prediction by striking a balance between the expected additional profits and the sales effort, which is reflected by the amount of time needed to visit a customer. In other words, the predictive policy continues to prioritize customers with (slightly) higher predicted uplifts or profits, while the prescriptive policy protects against errors in prediction. To put it another way, the prescriptive policy determines the length of time necessary to visit a client based on the anticipated higher earnings. This effect is particularly noticeable when there is a significant variance in the profits generated by the initiatives. This phenomenon takes place more often when capacity levels are high, and it takes place at medium and high levels of capacity as well.

When capacity levels are high, the prescriptive policy begins to resemble the zero-information policy more than it does the predictive policy. This behavior of the prescriptive approach explains why the "value of prescription" as indicated in Figure 4.14 develops both in the sales force capacity and the profit heterogeneity. Figure 4.14. The diagram for figure 4.14. As a consequence of this, we have arrived at the conclusion that the prescriptive policy should always be favored over the predictive strategy, and that the benefits of this policy are especially apparent in circumstances in which there is a significant degree of variance in the profits and the capacity of the sales force is not severely constrained. This conclusion was reached as a result of the fact that we have come to the realization that the prescriptive policy should always be favored over the predictive strategy.

4.6.5 Influence of Model Quality

There will be a wide range of quality in the forecasts that businesses get, which is represented here by the parameter. These traits are reliant on the availability of data, the predictiveness of that data, as well as the selection of a predictive model and the design of that model. In addition, these qualities are dependent on the design of the predictive model. In this part, we investigate how the performance of predictive and prescriptive policies may be influenced by the quality of the predictive model that is used for the purpose of projecting the uplifts in the future. In the first place, one of our key objectives is to figure out whether or not a low degree of model quality renders our prescriptive policy ineffective when applied in practice. In order to do this, we first examine the different policies based on the criteria described in Section 4.7.1, and then we duplicate "true" uplifts by employing values of P t0, 0.05, and 0.6u. Figure 4.15 depicts the effects of the policies when applied to homogeneous customers, with a sales force capacity of |T| 5, and varying model quality. This is the case when w = 0 is applied to the equation.

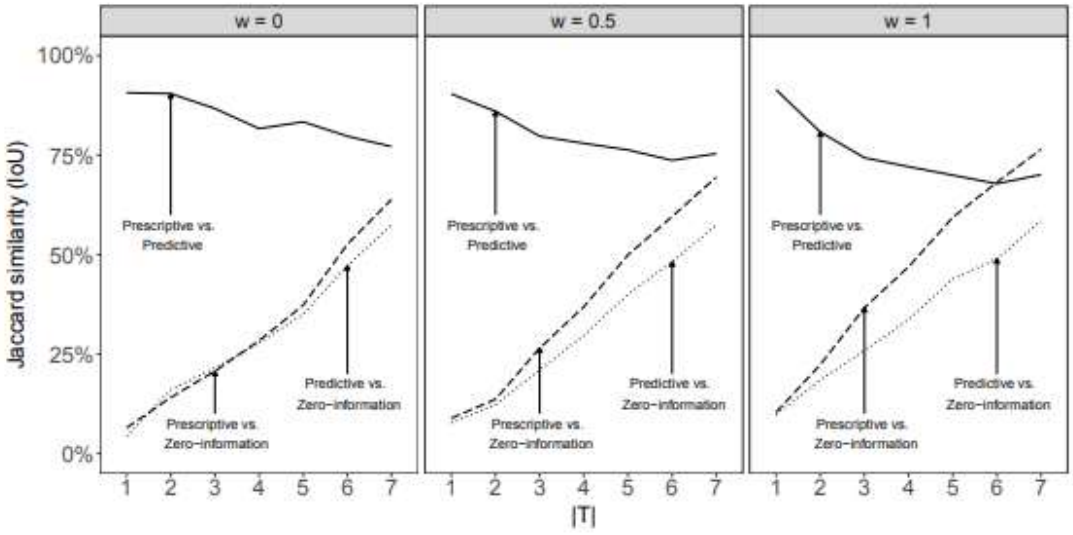


Figure 4.15: Similarity between the policies for varying levels of profit heterogeneity and sales force capacity (φ 0.42)

Source: *Data-driven Operations Management Data Collection and Processing through by M.Sc. Jan Maximilian Meller, In November 2019*

Its performance improves (almost linearly) together with the quality of the model and regularly beats the predictive policy, which may, at extremely low quality levels, lead to a performance that is worse than that of the zero-information policy. The prescriptive policy has a lower performance constraint at = 0 when its arg corresponds to that of the zero-information policy; its performance grows (almost linearly) in the model quality. This is because the prescriptive policy is defined as having this lower performance bound. One of the most compelling qualities of the prescriptive policy is its robust behavior, which makes it possible for a company to deploy predictive models despite the fact that such models may display only a very low degree of predicted accuracy. This is one of the reasons why the prescriptive policy is so popular.

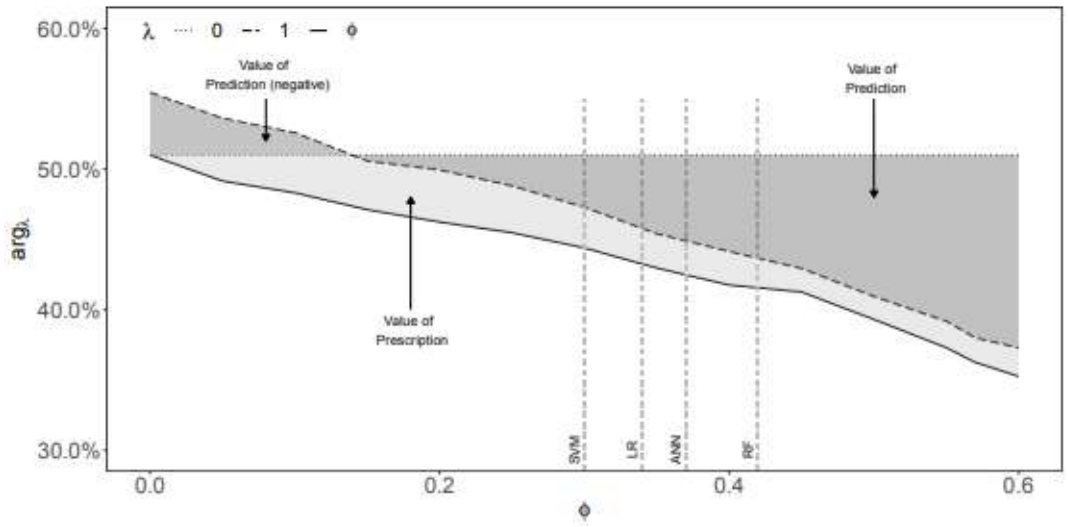


Figure 4.16: Average relative gap to optimality for varying model qualities (|T| = 5, w = 0)

Source: *Data-driven Operations Management Data Collection and Processing through by M.Sc. Jan Maximilian Meller, In November 2019*

In order to prove that this assumption is correct, we have emphasized, in Figure 4.16, the performance that would have been achieved if the various predictive models discussed in Section 4.4 had been used (please refer to the vertical dashed lines in Figure 4.16). In doing so, we have shown the correctness of this argument. We notice, for instance, that the prescriptive policy would result in considerable performance gains in contrast to the zero-information policy even if we employed models based on support

vector machines or logistic regressions, which result in the lowest quality — in order to make uplift predictions. This is because the zero-information policy is predicated on the assumption that no information is available. This is due to the fact that the zero-information policy is based on the premise that there is no information that is relevant to the situation.

The findings of this inquiry lend credibility to our past observations and assumptions, particularly in the following areas: Both the predictive scheduling approach and the zero-information policy tend to provide worse results when compared to the prescriptive scheduling strategy that we use, which consistently produces superior outcomes. Its performance advantage rises to a greater degree in proportion to the degree to which profits are diverse; it is at its most obvious when capacity is not severely constrained; and it predominates its rivals regardless of the quality of the prediction model.

We propose an integrated framework for prescriptive sales force scheduling as a solution to the problem of route planning that is encountered by sales representatives working in the construction industry. Our plan is based on a predictive model that makes use of an important machine learning categorization model in order to estimate the gains in revenue that can be ascribed to a single visit to a customer. These increases in revenue can be linked to the fact that our strategy is predicated on a predictive model that makes use of an essential machine learning categorization model. These uplift projections are used as an input for our prescriptive routing strategy, which takes the form of a task to be completed by a group engaging in orienteering. One of the defining qualities that sets our technique different from other methods is its capacity to tolerate the degree of inherent uncertainty in our uplift forecasts. This is one of the distinguishing features that set it apart from other methodologies. To make the optimization model more accurate and consistent with the accuracy of the underlying prediction model, we make use of a regularization parameter. This allows us to adjust the optimization model. This is carried out in accordance with the reliability of the model used for prediction.

In order to evaluate the efficacy of our technique, we undertake an investigation using a real-world data set that was acquired from a prominent European manufacturer of building paint and coating materials. We were able to show, via exhaustive numerical analysis, that the prescriptive scheduling technique outperforms two benchmarks in the base case scenario. This is the situation in which it is assumed that the revenues from

the many alternative projects are equivalent to one another. According to the findings of our subsequent sensitivity studies, the performance of the prescriptive policy is superior to that of the predictive policy under the following conditions: when profits from projects are divided more unevenly; when sales force capacity is increased; and when the quality of the uplift forecast is reduced. In addition, the prescriptive policy performs better when sales force capacity is increased than the predictive policy does. On the other hand, the standard of having no knowledge is easily surpassed in each and every one of the cases that were looked at.

According to the findings of our study, the framework for prescriptive sales force scheduling has the capability of extracting uplift information from historical sales data and analyzing it in conjunction with information regarding the uncertainty of various projections for the purpose of enhancing the schedules of salespeople who are currently engaged in sales activities. The determination of the factor, or the degree to which the uplift forecasts may be depended upon, is a crucial component of our technique. This assessment of the factor indicates the degree to which the projections can be relied upon. We have proposed a data-driven strategy for estimating this component, and we have done so on the basis that the individual who will be making the decision is risk neutral. In further study, we may adapt our approach to take into consideration the level of risk aversion shown by decision makers and the part that utility functions play in determining the value of this parameter. We would also take into account the role that utility functions play in determining this value.

CHAPTER 5

MACHINE LEARNING FOR INVENTORY MANAGEMENT

We examine and contrast two fundamentally different ways of taking data into consideration when making planning choices by using the scenario of a newsvendor dilemma as an example. In this scenario, observable qualities effect changes in demand, thus we use this scenario to compare and contrast the two approaches. Within the context of this situation, changes in demand are caused by the presence of visible attributes. The findings of our research contribute in two quite different ways to the overall expansion of the existing body of knowledge. To get started, we modified an existing machine learning method known as random forest in order to combine the two stages that are necessary for traditional segregated estimation and optimization (SEO) techniques. This allowed us to begin the process of combining the two phases.

Because of this, we are able to create a whole new approach that we refer to as joint estimation-optimization, or JEO. The first step is to produce an estimate of a model that is capable of anticipating demand, and the second step is to compute a safety buffer so that one can account for the uncertainty that is present in the forecasting model. In the traditional approach to search engine optimization (SEO), these steps are divided into two completely separate processes. In this second part of our research, we give an analysis of the elements that produce variations in the effectiveness of connected SEO and JEO implementations.

These variations can be attributed to a number of different causes. To be more specific, we are concentrating on which variables contribute the most to these differences. We offer the analytical and empirical findings from two separate investigations for the purpose of our performance assessments. One of these investigations was carried out in a regulated simulation setting, while the other was carried out on a data set collected from the actual world. Both of these investigations were carried out on the same data set. Both sets of study were conducted without collaboration between any of the researchers. When there is feature-dependent uncertainty and the cost structure of overage and underage costs is asymmetric, we demonstrate that JEO methods can lead to much better results than their SEO counterparts can. This is because JEO strategies are able to take advantage of the asymmetry in the cost structure of overage and underage costs.

This is due to the fact that JEO approaches take into consideration the uncertainty that is reliant on the features as well as the asymmetric cost structure. This is the circumstance that arises when the cost structure of overage and underage costs are asymmetrical. However, taking into account the real-world settings that were investigated, the size of these performance differences is limited because of the superimposition of effects that are in direct opposition to one another. These consequences include, respectively, the features of the remaining uncertainty and the cost structure. To put it another way, the magnitude of the performance differences is limited by a number of factors, one of which is the continued uncertainty, in addition to the cost structure.

5.1 INTRODUCTION

When looking at data for inventory management scenarios in which observable features impact changes in demand, we look at two concepts that couldn't be more fundamentally different from one another. The focus of research in the field of operations management has shifted away from approaches that rely on demand time-series from the past toward methods that may consider auxiliary data that may create variations in demand. This shift in focus has occurred as a result of an evolution in the direction of research. This development is in lockstep with the ever-increasing availability of data, and it has occurred as a direct result of the availability of more data. In other words, the availability of more data is directly responsible for this transformation. Studies that examine data on online clickstream to forecast the demand for door manufacturers; studies that derive daily demand from an examination of social media data; these are just a few examples of the use of such auxiliary data for making decisions regarding planning.

In the classic inventory-control literature, demand estimations of this sort are typically regarded to be the first stage in the process of making inventory choices. After then, the individual who makes the decision takes into account the uncertainty of the prediction (for instance, the empirical distribution of forecast inaccuracies) in addition to the costs that are connected with underage drinking and overage drinking. For clarity's sake, the person making the choice establishes an inventory level with the goal of reducing the potential expenses caused by mismatches between the two sets of inventory. This may be achieved by establishing a balance between the forecasted underage expenses connected with stock-out scenarios and the projected overage costs associated with extra inventory. This will allow for optimal financial management. In

the research that has been conducted on this topic, this concept is referred to as separated estimation and optimization, or SEO for short. One school of thought in the realm of academia advocates for the integration of these two stages, in contrast to the process of successively estimating a demand prediction model and optimizing inventory decisions based on the inputs of the former, which can be found in most textbooks and other resources. The estimated model has already taken into account the expected mismatch costs of the final inventory selection. This results in a single optimization problem that learns cost-optimal solutions based on past data. These are characteristics that are shared by all of their models.

In the context of settings containing newsvendors with parametric demand distributions, a number of research have shown that the most successful SEO strategies belong to a group of integrated approaches known as operational statistics. This conclusion was reached as a result of the findings of these studies. However, despite the fact that JEO methods are intuitively appealing because information is not lost between the prediction and optimization stages, there is not yet sufficient evidence to demonstrate that JEO approaches are superior to SEO approaches in a data-rich environment with non-parametric, feature-driven demand.

This is the case despite the fact that JEO methods are intuitively appealing due to the fact that information is not lost between the prediction and optimization stages. However, to the best of our knowledge, there has not been a full assessment of SEO and JEO techniques employing the same underlying machine learning technology and the same raw data. This is something that we believe needs to be done. The vast majority of the research that are currently available show that one JEO technique is superior than relatively straightforward standards such as sample average approximation.

We only acquire data that provide a credible foundation for comparing the SEO strategy with the JEO strategy in two of the experiments. In the first study, which was carried out by Ban and Rudin, the linear SEO strategy without regularization performed somewhat better than the JEO counterpart. In the second study, which was carried out by Huber and his colleagues, they discovered that there was no significant performance difference between a JEO technique based on artificial neural networks and its SEO counterpart. Both of these strategies have been devised with the end purpose of achieving the same objectives. Because of this, we believe that there is a gap in the study that necessitates a comprehensive investigation into the performance disparities

that exist between the implementations of the JEO concept and the SEO concept, as well as a quantification of the performance gap that exists in actual-world application settings. This is because of the fact that there is a difference between the JEO concept and the SEO concept.

The corpus of previously published information benefits from two distinct sorts of contributions made by our study, namely: In the first stage of this project, we are going to develop an entirely new JEO technique that will be based on an algorithm for machine learning called random forest. Second, we provide a critical and in-depth assessment of the structural differences and the factors that produce performance discrepancies between comparable SEO and JEO strategies for a number of different types of underlying machine learning algorithms. This part of the research focuses on the aspects that generate these performance disparities and how comparable SEO and JEO techniques generate them. We offer the analytical insights as well as the empirical results of two different investigations for the purpose of our performance evaluations. The first study was carried out in a controlled simulated environment, while the second study was carried out on a data set taken from the real world. Both of these studies were completed independently.

The discussion of implementations in part 5.3 follows the exposition of the theoretical underpinnings of the SEO and JEO ideas in section 5.2. The focus of this section is on two different underlying machine learning algorithms. Random forests and kernel optimization are two examples from the research literature that are included in these methods. In conclusion, the results of our investigation are presented and addressed in section 5.4 of the aforementioned document.

5.2 TWO CONCEPTS TO GET FROM DATA TO INVENTORY DECISIONS

Researchers in the field of operations management have been attempting to discover a solution to the problem of how to set inventory targets despite the unpredictability in consumer demand for a number of decades now. Uncertain demand is modeled in the conventional school of economic theory by making use of parameterized probability distributions, which, for the most part, are dealt with as though they are already known. However, given that the underlying demand distribution is often unknown in most actual scenarios, it is not possible to make such a strong assumption. This is because of the word "not feasible." Because of this, the assumption cannot be sustained. Not only is the manner of distribution unknown in a vast percentage of the occurrences that take

place in the actual world, but it is also patently evident that the level of demand does not remain the same throughout time. It is possible, for instance, that it is cyclical or seasonal, that it adheres to a pattern, that it is influenced by aspects such as weather, national holidays, and sales promotions, or that it may conform to a trend.

When confronted with a circumstance such as this one, one typical technique for dealing with it is to translate any information that could have predictive potential into features, which are essentially summary representations of the auxiliary data. This is one way to handle the situation. Take, for example, a model of demand known as an additive demand model. This kind of model has two parts: the demand level and an additional random component. We are able to show the concept of feature-driven demand with the help of this model. In this fundamental model, we make the assumption that the quantity of demand is both predictable and connected to the qualities of the data, which we will refer to as the vector x. This assumption is based on the fact that we believe the data can be represented as a vector. Internalizing any and all external uncertainty, which may also rely on the features, is the responsibility of the additional component that was added. As a result, demand D may be understood in the following manner:

$$D = \mu(\mathbf{x}) + \varepsilon$$
$$\text{with } \mathbb{E}[\varepsilon] = 0; \sigma_\varepsilon \sim \mathbf{x}$$
$$\mathbf{x} \in \mathbb{R}^k$$

Where f(x) is the function that defines the connection between the values of the characteristics x and the demand level, and where f(x) = E [D|X = x]. Where f(x) is the function that defines the link. The weekday, the month, and the temperature are some examples of data components that might be included in the vector x. Additionally, representations of additional factors that can have an influence on the amount of demand that is predicted could also be included.

Even if the function f(x) cannot be found in practice, we typically have access to a data collection of past observations. This is despite the fact that the function cannot be determined. These observations most frequently take the form of paired representations of demand and feature values. The notation T = (di,xi) is used to refer to the training data set, and the value of i can fall anywhere between 1 and n. Assuming there is an underlying demand model, such as in, we differentiate between two broad notions that

may be used to think about the learning data T in order to make decisions about inventory. These ideas may be utilized in order to make decisions about how much of a particular item should be kept in stock. In the next paragraphs, 5.2.1 and 5.2.2, we will delve even deeper into detail on these two distinct concepts.

5.2.1 Separate estimation and optimization (SEO) with auxiliary data

The search engine optimization method may be broken down into two stages: To begin, we will estimate a demand-forecasting model in order to see how well it can represent the connection that exists between the vector of data qualities x and the demand level x(x). To put it another way, by making use of a function that estimates x, we are able to arrive at an approximation of the function x. In view of the fact that we are unable to be absolutely confident that our model is error-free, in order to arrive at stocking decisions that are optimal for the circumstances, we regularly update our forecasts to take into consideration the unpredictability that is brought about by errors in our forecasting. For this reason, in order to offer a representation of the remaining uncertainty, we evaluate the prediction performance of the demand-forecasting model.

In order to attain this goal, one must first determine the distribution of the forecast errors. After then, the second distribution is utilized as an input for the inventory-optimization algorithm, which generates an additional safety stock by trading off forecast overage costs with expected underage costs. Ultimately, this helps reduce overall inventory costs. The algorithm that optimizes inventory levels will then decide how much more safety stock to keep on hand. Both the prediction that was generated by the forecasting model and the safety stock will be taken into consideration in order to arrive at the conclusion that will ultimately be made about the inventory.

To state it in a manner that is more official, the problem that has to be solved is

$$q^{*}_{SEO}(\mathbf{x}) \in \mathcal{M}\times\mathbb{R} = \underset{\hat{\mu}(\cdot)\in\mathcal{M}}{\arg\min}\ \mathbb{E}\left[L(\hat{\mu}(\mathbf{x}), D)\middle| X=\mathbf{x}\right] + \underset{z\in\mathbb{R}}{\arg\min}\ \mathbb{E}\left[C(z, D-\hat{\mu}(\mathbf{x}))\right]$$

The function μˆ, which predicts values, is selected from a function space M that translates from the set of all possible feature vectors X to real value outputs. Furthermore, L[μˆ(x), D) and C(z, D -μˆ(x)] are two separate loss functions that are independent of each other.

5.2.2 Joint estimation-optimization (JEO) with auxiliary data

In spite of the fact that it is frequently utilized in actual practice, the two-step SEO approach has a critical drawback, which is as follows: We have established two independent optimization issues by first optimizing a prediction model for the demand and then optimizing the inventory choice. These problems are not necessarily congruent with one another, which can lead to judgments that are not as excellent as they might be. JEO models, which immediately link the features with the final decision and hence avoid the intermediate phase of building a demand prediction model, have recently attracted attention as a consequence of this reason. JEO models directly link the characteristics with the final choice. The JEO models directly correlate the characteristics with the conclusion that is reached. Instead, the training of the demand prediction model and the decision-making process about inventory are combined into a single optimization problem that has to be addressed. The core concept of merging statistical estimating with optimization is often ascribed to Hayes, who estimated policies based on data by minimizing the expected overall operating cost. Hayes is also credited with having developed the idea.

Bertsimas and Kallus are the ones who come up with the concept for a framework for JEO models, and they formulate the problem as follows:

$$q^*_{JEO}(\mathbf{x}) \in \mathcal{Q} = \underset{q(\cdot)\in\mathcal{Q}}{\arg\min}\ \mathbb{E}\left[C(q(\mathbf{x}), D)|\mathbf{x}\right],$$

where q: X R is a decision function from the function space Q that maps from the set of all possible feature vectors X to real valued options; and C(q(x),D) is a loss function that returns costs given a decision q and a realization of demand D. The decision function q: X R translates from the set of all possible feature vectors X to real valued options. q: The translation from the set of all potential feature vectors X to real valued options is denoted by the letter X R. where q: X R is a decision function from the function space Q that maps from the set of all possible feature vectors X to real valued decisions. This decision function translates from the set of all possible feature vectors to the set of real valued decisions. q is a translation from the set of all potential feature vectors, X, into actual valued decisions. In contrast to SEO, JEO consists of a single optimization problem, the answer to which may be easily determined with reference to the real cost function C (q(x), D). This is the difference between the two that stands out as the more significant of the two and may be considered the deciding factor.

The functional relationship q (x) that exists between the choice and the characteristics is the essential factor that differentiates a few distinct examples of JEO processes that can be identified in the literature from one another. This relationship occurs between the decision and the characteristics. Both Beutel and Minner's and Ban and Rudin's respective contributions make use of the linear function q: X R: q (x) = T x in order to create a relationship between a feature vector x that has a length of k and the newsvendor quantity qpxq. This is done in order to determine whether or not a relationship exists between the two. In order to ascertain the relationship between a feature vector x and the newsvendor quantity, this relationship is formed so that it can be determined how the two are related to one another. They do this by making use of the learning data that they have available to them in order to optimize the weights j for each feature.

Ban and Rudin also present a second JEO strategy, which makes use of kernel functions to establish weights for each observation. This JEO method is referred to as the kernel function method. Both of the researchers put out this approach as a possible solution. This tactic has been laid out for the reader by the authors of the piece. After then, the decision is determined by taking a locally weighted average of the previous data in order to arrive at the best possible choice. We utilize the kernel approach in our study because it can be used to both SEO and JEO, which is a comparison that has not been published previously, and because it enables us to compare and contrast the data that we gather with our recently formed tree-based methodology. The kernel technique can be applied to both SEO and JEO. This comparison has not been published before. In contrast to Ban and Rudin, we place our primary emphasis on the differences between SEO and JEO and make an effort to isolate the primary factors that contribute to variation in performance. This stands in stark contrast to the fact that Ban and Rudin do not differentiate between these two options.

Deep learning is a type of artificial neural networks; to make it more accurate, combine it with a loss function that is comparable to that of a news vendor. Deep learning is a sort of machine learning. They apply their novel approach to a scenario that involves a newsvendor who sells a range of things and use that scenario as a test subject in order to evaluate how well their approach performs in contrast to other tried-and-true ways. They demonstrate that their method is effective in scenarios both with an appropriate number of training data and conditions with unknown underlying demand distributions. Additionally, they demonstrate that their method is effective in both cases

simultaneously. In addition to this, they provide evidence that demonstrates how successful their method is in both of the scenarios concurrently. They do not, however, evaluate their JEO strategy in comparison to a form of SEO, in which deep learning would be utilized to assess the demand from customers. This is not something that they participate in at all.

It is unclear what proportion of the cost gain they accomplish (in comparison to the benchmark approaches from the literature) is connected to the integration of estimate and optimization and what proportion is due to the superior and more challenging prediction strategy. As a consequence of this, it is uncertain what proportion of the cost gain they achieve (in comparison to the benchmark approaches from the literature). Deep learning algorithms, although being strong and usually providing good results, are unintelligible black boxes. This is despite the fact that deep learning algorithms regularly yield good outcomes. In spite of the fact that they frequently bring about favorable results, this continues to be the case. Because of this, they are not as appropriate for use in an analysis of the structural differences that exist between SEO and JEO as, for instance, tree-based approaches are. As a result of this, they are not as appropriate for use in a study of the structural differences that exist between SEO and JEO.

Bertsimas and Kallus have devised a concept for a tree-based strategy that is a hybrid of SEO and JEO. This strategy aims to improve website rankings. The shape of the decision tree that is generated by their model is determined by the way in which they make use of the conventional mean-squared error loss function. This capability has been available for a considerable amount of time. The authors take the sample of learning data that is placed in each leaf, and in the second phase of the process, they use it to solve an issue-specific version of the optimization problem presented in (5.3). They are therefore able to compute the answer for each individual leaf on the tree as a result of this. They apply this theory to the idea of random forests, which are groups of decision trees that, in general, give better outcomes than single trees would.

Although Bertsimas and Kallus' approach is the one that most closely resembles ours in terms of the underlying machine learning method, their tree-based method still contains the primary flaw of SEO models, which is the application of two independent optimization steps (first the structural learning, and then the actual cost "optimization"). Because of this error, the models are not as accurate as they otherwise could be. This error is referred to as "the application of two independent optimization steps," and we

call it "the application of two independent optimization steps." When they determine the structure of the decision tree, in contrast to us, they do not take into consideration the costs that are relevant to the scenario. On the other hand, this is not the case with us. There is no way that we will be able to produce a genuinely JEO solution that is built on random forests if we do not initially incorporate in these charges. There is no way. In the following paragraphs, we will offer a full description of the model that we have built to handle the problem of inventory that is faced by newsvendors. This problem has been identified as a barrier to growth for the industry.

5.3 APPLICATION TO THE NEWSVENDOR PROBLEM

We use a newsvendor scenario as a means of drawing attention to the fundamental performance differences that exist between SEO and JEO approaches. A problem that emerged in a real-world situation that took place at a restaurant chain served as the impetus for this discussion. In this particular scenario, the management of the restaurant is interested in determining the quantity q of a specific product that will have to be manufactured for the following working day. We take into consideration the non-stationary character of demand, which is impacted by external influences in the form of a k-dimensional feature vector x. These vectors may be thought of as having x as their dimension. Unsold quantities have to be disposed of at a cost of co per unit, and the projected cost of unmet demand is cu per unit. This means that the total cost of unmet demand is cu per unit. In accordance with what is said in (5.3), the goal here is to cut the entire estimated cost down by as much as is humanly possible:

$$\min_{q(\mathbf{x})\in\mathcal{Q}} \mathbb{E}[C(q(\mathbf{x}), D)]$$

With the specific newsvendor cost function

$$C(q(\mathbf{x}), D) = c_u(D - q(\mathbf{x}))^+ + c_o(q(\mathbf{x}) - D)^+,$$

Where D stands for the random demand and (.) + is a function that yields the value of the argument itself unless it is told that the argument it was given is negative, in which case it returns 0.

It is required for us to provide a more in-depth description of the function q(x) in order to be successful in finding a solution to this optimization problem. Following this, we

will present implementations utilizing two underlying functions (i.e., techniques for machine learning): the first implementation is based on random forests, while the second implementation is based on kernel regression. Both of these implementations will be discussed in more detail below. Both of these implementations will each have their own section below that describes them in further depth.

5.3.1 Implementation based on random forest

The machine learning technique of random forests, which was first introduced by Reiman, has been shown to have a high degree of prediction accuracy in a range of different scenarios. This was proved through a series of experiments. Tree-based approaches such as random forests are particularly useful for our research because we can use the final tree structures that they create to quantify heteroscedasticity, as was discussed in chapter 5.4.3. This is one of the reasons why tree-based methods are so beneficial. One of the subjects that is discussed in that subsection is this one.

A random forest is a specific kind of classification technique that, in general, consists of a number of trees referred to as T that partition the feature space into regions referred to as R that group instances whose characteristics have values that are comparable to one another. The forecast of a new instance that has not been seen before can be derived by first placing the instance into one of the regions based on the values of its characteristics and then ascribing a demand estimate, such as the mean demand of the other examples in this area, based on the other instances that are located in this region. This process can be repeated until the forecast is complete. The prediction of a new instance that has not been observed before is the result of this approach. This technique is based on the core premise that it is acceptable to infer that examples that are similar in known qualities of the data (the features) are also similar in unknown properties (such as the realized demand).

The concept that it is reasonable to suppose that instances that are similar in known qualities of the data are also similar in unknown properties (such as the realized demand) is what supports the methodology. The finding of the regions is accomplished by recursively conducting axis parallel splits on the training data set T in order to minimize a training loss function L ((x), D). In the following, we shall refer to as the parameter vector that sets the method in which a tree develops and R(x 1,) as the region of a single tree into which a new instance that is represented by x 1 would be sorted. Both of these terms will be used interchangeably throughout the rest of this section. We

are able to comprehend such a region as a forest-based adaptive neighborhood of x 1 that is outlined by the data-driven weights w i (x 1) of each historical observation. This allows us to comprehend how such a region works. This view is viable due to the fact that a forest delineates the boundaries of such a neighborhood.

Because it provides a data-driven method to re-weight previous observations in order to produce projections, the use of random forests is an essential component in the decision-making process pertaining to inventory. This is because random forests give a data-driven strategy. In the following, we will present a more in-depth description of how the core random forest mechanism may be employed via the lens of both the SEO technique and the JEO approach in order to create such judgements. This will be done in order to demonstrate how the SEO method and the JEO approach complement one another. We make note of two key characteristics that distinguish the two approaches: the manner in which regions are created by the training algorithm and the manner in which final judgements are made based on the particular neighborhoods. Both of these aspects are discussed in the following paragraphs.

5.3.1.1 SEO based on random forests

According to what is mentioned in Section 5.2, the generic SEO approach forecasts a different level of predicted demand on its own and then accounts for the uncertainty that is still there by computing an additional amount of safety stock. The quantity of this additional safety stock is something that will be determined by the manner in which forecast mistakes are distributed. The random forest algorithm is applied in line with this method in order to generate a forecast of the average demand, with the accuracy of the forecast being reliant on the actualization of feature vector x 1. When training tree topologies, the feature space is partitioned in such a way as to minimize the standard MSE loss function in order to achieve the best results. Because of this, it is possible to acquire the regions RSEO(x 1) that are necessary in order to forecast the conditional mean.

$$L(\hat{\mu}(\mathbf{x}), D) = L_{\mathrm{MSE}}(\hat{\mu}(\mathbf{x}), D) = \frac{1}{n}\sum^{n}(d_i - \hat{\mu}(\mathbf{x}_i))^2.$$

The subsequent phase of optimization will consist of locating an additional safety stock. This will enable the decision-maker to hedge against the possibility of making inaccurate projections by giving them the opportunity to exchange the costs of an

anticipated overage and underage, respectively. The solution to this issue is the same as the resolution to the problem of the plain data-driven newsvendor, which does not involve any features. In order for us to figure out how to approach this issue, we need an empirical distribution of the out-of-sample prediction errors. Therefore, after training the random forest on a subset of the training data, we evaluate the predictions using the entire set of data, which was not used for training the random forest. This ensures that the predictions are as accurate as possible. After that, the out-of-sample prediction errors, which are indicated by the symbol i, are computed, and the subsequent formula is utilized to derive the ultimate inventory choice using SEO-RF:

$$\hat{q}_{SEO-RF}(\mathbf{x}') = \sum_{i=1}^{n} w_i(\mathbf{x}')d_i + \inf\{\varepsilon : \hat{F}_n(\varepsilon) \geqslant \frac{c_u}{c_u + c_o}\},$$

5.3.1.2 JEO based on random forests

The approach that is taken by the JEO method, which is based on random forests (JEO-RF), is notably different from the one that is taken by the SEO random forest method (SEO-RF) in two key ways: To begin, the training algorithm's loss function already takes into consideration the cost structure of overage vs underage amounts. This is the case even though the system is still learning. Because of this, tree topologies are produced, which immediately reflect the second-stage optimization difficulty posed by the SEO approach. Second, in order to obtain the final inventory judgements when confronted with tree designs of this type, a novel method of taking into consideration the observations of the nodes that are next to the node being analyzed is implemented. Take into account the asymmetric loss function that is shown in the following:

$$L(q(\mathbf{x}), D) = C(q(\mathbf{x}), D) = \sum_{i=1}^{N} c_o(q(\mathbf{x}) - d)^+ + c_u(d - q(\mathbf{x}))^+$$

After becoming familiar with cost-aware tree structures, we implement the random forest kernel approach that was created in Scornet to define weight functions for the training instance as follows:

$$w_i(\mathbf{x}') = \sum_{t=1}^{T} \frac{\mathbb{1}_{(\mathbf{x}_i \in R_{JEO}(\mathbf{x}', \theta_t))}}{\sum_{t=1}^{T} N(\mathbf{x}', \theta_t)}.$$

Scornet claims that by employing this process, it is possible to avoid the development of rough estimations in regions of the feature space for which there is an absence of appropriate data. We are able to make use of these weights in order to generate data-driven neighborhoods for a fresh instance x 1 in a way that is comparable to that of the SEO method based on random forests. Now, we can make the final inventory decisions with JEO-RF by solving the following by applying Bertsimas and Kallus's architecture and inserting in our own loss function. This may be done by following these steps:

$$\hat{q}_{JEO-RF}(\mathbf{x}') = \arg\min_{q(\cdot)\in\mathcal{Q}} \sum_{i=1}^{N} C(q(\mathbf{x}'), d_i)$$
$$= \inf\{d : \sum_{i=1}^{N} w_i(\mathbf{x}')\mathbb{1}_{(d_i \leqslant d)} \geqslant \frac{c_u}{c_u + c_o}\}.$$

The realization that the difficulty that was brought about by this is connected to a quantile regression issue is what ultimately leads to the achievement of equality.

5.3.2 Implementation based on kernel optimization

In order to confirm the results that we acquire with our newly created approach that is based on random forests, we also use and evaluate the SEO and JEO ideas that are based on a kernel optimization (KO) method. The JEO-KO approach, which was established by, delivers the greatest findings in comparative research that makes use of a data set that originates from the actual world. This is because the method was developed by.

Kernel regression may trace its roots back to Nadaraya and Watson, who developed the essential notion that underpins the methodology. They introduced the concept of estimating a dependent variable, such as demand, by taking a locally weighted average of prior demands. This method is used to determine an estimate for the dependent variable. The degree to which the values of the characteristics of the historic observation are comparable to those of the instance that is being investigated is the factor that is used to determine the weights.

5.3.2.1 SEO with kernel regression

In the SEO-KO approach, also referred to as the kernel-based SEO technique, we stick to the SEO concept described in Section 5.2 and utilize kernel regression to estimate a demand-predicting function represented by f SEO-KO. This function takes a feature

vector x 1 as an input. The SEO-KO method is also known as the kernel-based SEO methodology. This function, which is known as the Nadaraya-Watson estimator, is characterized by the equation that is presented below:

$$\hat{q}_{SEO-KO}(\mathbf{x}') = \hat{\mu}_{SEO-KO}(\mathbf{x}') + \inf\{\varepsilon : \hat{F}_n(\varepsilon) \geqslant \frac{c_u}{c_u + c_o}\},$$

By utilizing the function μˆ SEO – KO, we assess the accuracy of the predictions made on the training data and then derive the out-of-sample prediction errors. Similar to the SEO-RF method, the ultimate determination of the inventory selection is made.

5.3.2.2 JEO with kernel optimization

The fundamental differentiation between the kernel-based Joint Estimation and Optimization approach (JEO-KO) and the Stochastic Estimation and Optimization approach (SEO-KO) is in their differing techniques. Ban and Rudin (Year) developed the JEO-KO method, which use the Nadaraya-Watson estimator (as outlined in equation 5.13) to estimate the cost linked to the newsvendor, instead of predicting the demand. The JEO-KO approach is afterwards introduced as:

$$\min_{q \geqslant 0} \frac{\sum_{i=1}^{N} K_w(\mathbf{x}' - \mathbf{x}_i) C(q, d_i)}{\sum_{i=1}^{N} K_w(\mathbf{x}' - \mathbf{x}_i)}.$$

According to Ban and Rudin (year), the problem at hand may be classified as a one-dimensional piecewise linear optimization problem. The authors provide a solution to this problem, which is provided as follows:

$$\hat{q}_{JEO-KO}(\mathbf{x}') = \inf\{q : \frac{\sum_{i=1}^{N} \kappa_i \mathbb{I}(d_i \leqslant q)}{\sum_{i=1}^{N} \kappa_i} \geqslant \frac{c_u}{c_u + c_o}\},$$

5.4 COMPARISON OF SEO AND JEO

In this part, we will examine the factors that contribute to the disparities in performance between the SEO and JEO methodologies. In the first subsection, we compare SEO and JEO by modeling the link between features and demand (SEO) and that between features and decision (JEO) as linear functions. This allows us to examine the similarities and differences between the two approaches. In this linear context, we are

able to demonstrate analytically that SEO results in outcomes that are less than optimum if the remaining prediction uncertainty follows a pattern that is not random. In accordance with the body of research that exists in the field of econometrics, we refer to this type of uncertainty as heteroscedasticity.

Our findings from the analytical investigation with linear models culminate in our hypothesis that heteroscedasticity is also the primary driver of performance disparities in the more complicated JEO and SEO techniques. This hypothesis was reached as a result of our examination of the data using linear models. Due to the fact that tree-based and kernel-based models do not permit analytical procedures that are comparable to those that are possible with linear models, the following analyses are based on two studies: A simulation experiment in which we evaluate the impact of various specifications of the data structures on the models' performance while controlling for exogenous, confounding effects; and a test of our findings on a real-world data set, in which we apply the two approaches to an inventory planning problem posed by a restaurant chain. Both of these experiments are controlled for exogenous, confounding effects.

5.4.1 Analytical examination

When modeling a connection between a dependent variable and a group of independent variables, it is common practice to make the assumption that the error term is homoscedastic. This is because it is easier to analyze the data if the error term is consistent across all of the variables. This is because we want to describe the relationship between the variable that is being dependent on and the variables that are being independent. This assumption implies that we are able to express the variation of the dependent variable as the sum of a term that is explained by the model, x, and a component of stochastic error that has a constant variance across all of the occurrences.

Specifically, this assumption says that we are able to describe the variation of the dependent variable as the sum of a term that is explained by the model, x. On the other hand, there is a high probability that this assumption of homoscedasticity will not be supported by the data. Breiman and Friedman explore the difficulty of predicting ozone levels for the following day, and they show that it is possible to anticipate these values with a far better degree of certainty on some days than on other days. When it comes to projecting demand, the same idea applies. For instance, the demand for a restaurant on a typical weekday may change far less than it does on the weekend.

In the next section, we will examine and contrast the impacts of heteroscedasticity on the cost performance of SEO and JEO. When the link between features and demand (SEO) and that between features and choice (JEO) are represented as linear functions, this section will focus on the effects of heteroscedasticity on SEO. More specifically, the cost performance of SEO and JEO is the topic that will be covered in this part. The linear SEO approach is comprised of the least squares estimate of the conditional mean function (x) and a sample quantile of all residuals. Both of these elements make up the total.

Beutel and Minner were the ones who initially presented the Linear JEO approach, but Ban and Rudin were the ones who ultimately perfected it. The conditional quantile is what distinguishes the technique.

$$\hat{q}_{JEO-Lin}(\mathbf{x}) = \mathbf{x}\hat{\beta}_{SL},$$

Koenker demonstrates that the quintile function, as shown in Equation, is comparable to the linear SEO strategy in the sense that the only difference between the two is a vertical displacement by the sample quintile of the error distribution q (SL). This is the case for a straightforward linear demand model with errors that are independent and identically distributed (iid) and do not depend on x. The conclusion that can be drawn from this is that in a situation that is homoscedastic and linear, the outcomes that are achieved using either approach are comparable to one another.

If there is any kind of feature-dependent uncertainty, the assumption of (iid) mistakes, which is necessary for the linear SEO approach, will not hold true. This is because the linear SEO strategy depends on it. On the other hand, if there is uncertainty that is not based on any features, then there is indeed uncertainty that is not dependent on any features. We will study the impact of heteroscedasticity on the uncomplicated univariate linear location-scale model by looking at both of the following approaches:

$$D|(X = x) = \beta x + (\gamma x)u$$

In this setting, the optimal newsvendor decision is given by

$$q^*(x) = x(\beta + \gamma F_u^{-1}(SL))$$

Proposition 5.1 For a linear location scale model with heteroscedasticity as in, the following hold:

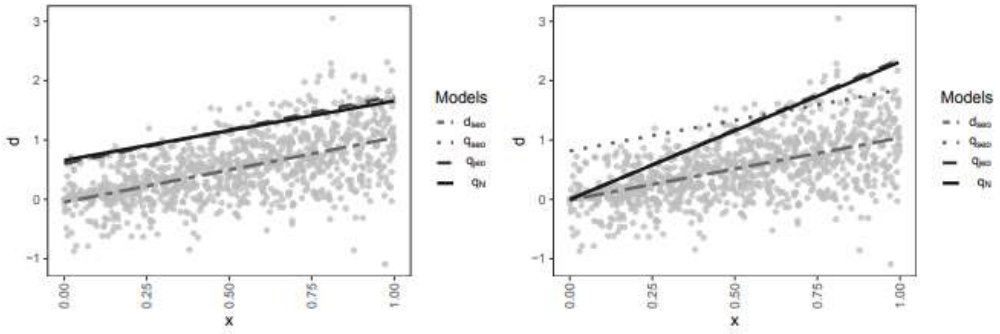
$$\mathbb{E}_{X\times D}\left[C(q_{JEO-Lin}(x), D)\right] \leqslant \mathbb{E}_{X\times D}\left[C(q_{SEO-Lin}(x), D)\right]$$

You may find the evidence for this thesis, as well as the proofs for all of the other propositions, in the appendix A. Figure 5.1 is an illustration of the scenario that has been given for X L ((x), D), 1q. It demonstrates that for the homoscedastic situation, both the SEO technique and the JEO approach perform well in the vicinity of the optimal choice quantile. In the case of heteroscedastic data, however, only JEO successfully captures the structure of the noise in an appropriate manner by altering the slope of the regression line. On the other hand, SEO leads to inefficiently high or low ordering judgments since there is only a parallel shift of the regression line.

In addition, the magnitude of the effect of heteroscedasticity is proportional to the level of service, or the degree of asymmetry present in the cost structure:

Proposition 5.2 With C(.) the e newsvendor cost function from Equation, 0 < y < 1 and X – unif (0,1) the following holds. For symmetric cost.

$$\mathbb{E}_{X\times D}\left[C(q_{JEO-Lin}(x), D)\right] = \mathbb{E}_{X\times D}\left[C(q_{SEO-Lin}(x), D)\right].$$



(a) Homoscedasticity (b) Heteroscedasticity

Figure 5.1: Comparison of the linear SEO and JEO approaches under homoscedastic versus heteroscedastic settings

Source: *Data-driven Operations Management Data Collection And Processing Through By M.Sc. Jan Maximilian Meller, In November 2019*

On the basis of these findings for linear models, we are able to draw two crucial conjectures, both of which will be studied further in the forthcoming study which will make use of more intricate underlying machine learning models:

Both the concept of homoscedasticity and the concept of heteroscedasticity are taken into consideration by Hypothesis 5.1. When both sets of results are compared in contexts with homoscedastic distributions, the performance of JEO does not significantly surpass that of SEO. When there are more shifts in a demand model that are sensitive to x, JEO's performance will improve in comparison to SEO's performance. To clarify this point further, this will take place whenever there are increasing degrees of heteroscedasticity.

The fifth and final hypothesis explores how much of a role the quality of the service plays, specifically looking at the degree to which it matters. When costs are symmetric, which is to say when the service level is 0.5, the presence of heteroscedasticity does not have a significant influence on the relative performance gaps that exist between SEO and JEO. It's feasible that the effect of heteroscedasticity will be shown to be more severe in circumstances in which there's a greater degree of asymmetry. This is something that may happen.

After this, we will investigate the structural distinctions that exist between SEO and JEO, with a particular emphasis on the more complicated machine learning models, such as random forests and kernel optimization, that are the driving force behind both of these approaches. These ideas serve as the foundation for both SEO and JEO, and both are developed upon them. This analysis will be carried out by comparing the models by means of a controlled simulation experiment with a real-world dataset taken from a restaurant chain. This will be done so as to determine whether or not the models accurately represent the actual world. These two collections of data are going to be utilized in combination with one another. This is because we are unable to generate proofs of propositions using these models in the same way that we did when we were working with the linear model. The reason for this is due to the fact that we are unable to generate proofs of statements using these models. Instead, we are going to use the linear model to figure out how to fix this issue.

5.4.2 Study 1: Simulation analysis

When we have a homoscedastic or heteroscedastic uncertainty structure, our initial numerical analysis is a controlled simulation experiment that enables us to quantify the

effect of feature-dependent demand uncertainty. This helps us to determine whether or not we should use a homoscedastic or heteroscedastic distribution. We are in the fortunate position of being able to identify and examine individual causes and consequences because to the tightly controlled environment. Our research on simulations is complemented by an investigation that makes use of a data collection derived from the actual world. Because it is susceptible to a number of extra factors, such as nonlinearity, heteroscedasticity, and misleading correlations between predictors and prescriptions, this kind of data does not give the same degree of insight as simulation research. Rather, it provides a lower level of understanding. We have a working hypothesis that the simulation approach that we used provides the potential for us to statistically validate our findings and permits the extraction of valuable insights on the factors that drive performance discrepancies.

In this section, we will start by discussing the equipment that was utilized in the experiment that we conducted. Next, we show the findings for the random forests technique first, and then we present the results for the kernel-based strategy. First, we discuss how we handle the feature-related uncertainty through our choice of a demand model and its parameterization.

5.4.2.1 Experimental setup

An additive demand model is employed in order to independently manage the link between features and demand, as well as the uncertainty associated with feature dependence. In a more formal manner, the determination of demand is expressed as follows: The parameter γ in the simulation defines the nature of the demand, with $\gamma = 0$ resulting in homoscedastic demand and γ values ranging from 0.1 to 1 resulting in discrete heteroscedastic demand with varying levels of heteroscedasticity. The coefficient of variation (CV) of noise is the parameter that governs the magnitude of noise. In the context of our simulation, we exercise control over the coefficient of variation (CV) noise due to its independence from the mean.

We consider heteroscedasticity with a two-population model for the uncertainty component $\varepsilon\gamma$ and a feature x0 that influences only the structure of the uncertainty and has no effect on the demand level. In reality, x0 could represent, for example, whether we consider a typical weekday or a weekend day, assuming that the mean is similar but the uncertainty around our predictions is higher on weekends. Via this modeling approach, γ controls the level of heteroscedasticity by affecting the difference of the standard deviations of ε^{0} and $\varepsilon^{1.}$

$$
\begin{aligned}
D &= \mu(\mathbf{x}) + \varepsilon_\gamma(\mathbf{x}) \\
\text{with } \mu(\mathbf{x}) &= x_1 + \ldots + x_k \\
\text{and } \varepsilon_\gamma(\mathbf{x}) &= \varepsilon_\gamma^0(1 - x_0) + \varepsilon_\gamma^1 x_0, \\
\text{where } \varepsilon_\gamma^0 &\sim \mathcal{N}\left(0, (1-\gamma)\sigma_{base}\right) \\
\text{and } \varepsilon_\gamma^1 &\sim \mathcal{N}\left(0, \sqrt{2-(1-\gamma)^2}\sigma_{base}\right) \\
\text{with } x_0 &\in \{0, 1\}, \\
\sigma_{base} &= \mathbb{E}\left[\mu(\mathbf{x})\right] cv_{noise},
\end{aligned}
$$

Term εγpxq as described in. Following this approach, we obtain a training dataset T_{Nsim} = {(d $_i$, x_i), i = 1, ..., N $_{sim}$}. To measure the performance of each model, we use the first Nsim 1 instances to train the model and then evaluate them for period Nsim. This procedure is repeated S times to achieve stable results. Mismatch costs incurred by model m P X, SEO-X with X either RF or KO are calculated for each simulation run s = 1, S via the cost function:

$$C(\hat{q}_m(x_s), d_s) = c_u(d_s - \hat{q}_m(x_s))^+ + c_o(\hat{q}_m(x_s) - d_s)^+,$$

per model m, and present the relative cost improvement associated with the JEO technique in comparison to the SEO strategy in the following manner:

$$\delta_{JEO} = \frac{\bar{c}_{\text{JEO-X}} - \bar{c}_{\text{SEO-X}}}{\bar{c}_{\text{SEO-X}}};$$

Table 5.1: Parameter settings for our experiments

Experiment	**Simulation**	**Real-world application**
	Section 2.4.2	Section 2.4.3
Parameters		
γ	$\{0, 0.25, \ldots, 1\}$	–
SL	$\{0.5, 0.8, 0.95, 0.99\}$	$\{0.5, 0.8, 0.95\}$
Controls		
cv_{noise}	$\{0.25, 0.5, 0.75, 1\}$	–
Model configs		
n_{trees}	$\{100, 500\}$	$\{100, 500\}$
min_{node}	$\{5, 15, 30\}$	$\{5, 15, 30\}$

In order to determine whether or not our hypotheses are accurate, we have conducted a variety of simulation experiments utilizing a wide range of parameter combinations, as given in Table 5.1. These experiments were carried out in order to assess whether or not our hypotheses are accurate. We investigate the implications that feature-dependent uncertainty has on the comparative performance of the JEO and SEO strategies, while concurrently controlling for the overall uncertainty level and the asymmetries that exist between the overage costs and the underage costs. Feature-dependent uncertainty can have a significant impact on the relative performance of these two techniques. The relative performance of the JEO and SEO techniques can be impacted by uncertainty that is depending on the features being considered.

In order to provide a clearer explanation, we tweak the value of the parameter to account for a wide range of possible combinations of service level SL and cvnoise. In every single one of our experiments, the sample size that we use is N sim = 501 observations, and the number of features that are responsible for calculating the demand levels is set to k = 3. Even though the number of features that are considered in real-world scenarios is typically much higher (for example, in our Yaz case study, we have k = 168), other studies have shown that tree-based methods such as random forests are especially likely to perform robustly even with noisy features. This is the case even though the number of features that are considered in real-world scenarios is typically much higher (for example, in our Yaz case study, we have k = 168). This is due to the fact that characteristics that are considered to be noisy are features that either do not have any predictive power at all or have just a little degree of predictive power. S = 100 will always be the value used for the number of iterations of the simulation that are performed on each parameter configuration and model.

R was the statistical programming language that our team utilized for developing the models that are detailed in Section 5.3. In order to make room for the random forest models, we had to make a few adjustments to the ranger package.

5.4.2.2 Results for random forest-based approaches

Figure 5.2 is an illustration of the relative performance increase that the JEO-RF technique possesses in comparison to the SEO-RF approach. This improvement is shown for increasing degrees of heteroscedasticity and for a range of parameterizations of noise parameters and service level parameters. There is no effect of increasing heteroscedasticity in situations in which the level of uncertainty is moderate (cv noise

= 0.25), and both approaches are equally successful in terms of extracting the underlying linear connections. If the degree of uncertainty is low, then it does not seem to make a difference whether or not there is any structure in the remaining uncertainty that may be useful for JEO-RF. This is because if the degree of uncertainty is low, then there will be less room for error.

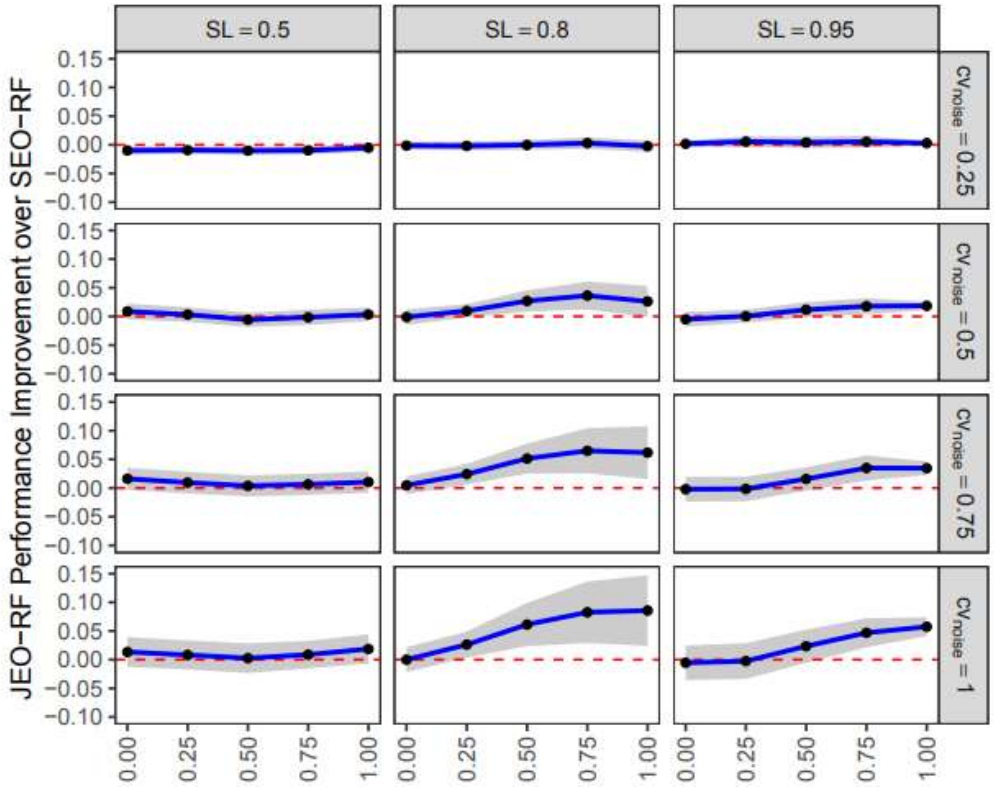


Figure 5.2: JEO-RF cost improvement over SEO-RF depending on γ (level of heteroscedasticity) in a linear demand setting for various service levels pSL = 0.5, 0.8 and 0.95 with different levels of base noise (cv $_{noise}$). The shaded area represents a 95% confidence interval around the mean improvement

Source: *Data-driven Operations Management Data Collection And Processing Through By M.Sc. Jan Maximilian Meller, In November 2019*

When a larger degree of uncertainty is being modeled, the effect of heteroscedasticity on the performance of JEO-RF is favorable in compared to SEO-RF. This is because JEO-RF is more robust to extreme values of the variables being modeled. This is due

to the fact that JEO-RF is more reliable. This is as a result of the positive effect that the influence of heteroscedasticity has on JEO-RF's performance. If the amount of uncertainty is sufficiently high and JEO-RF delivers significantly better results than SEO-RF under particular parameters (for instance, cv noise = 1 and SL = 0.95), then we are able to assert that Conjecture 5.1 is correct. JEO-RF is capable of achieving results that are noticeably superior than those achieved by SEO-RF.

In spite of this, we have showed that JEO-RF, even when operating in homoscedastic settings, has the potential to occasionally perform marginally worse than SEO-RF. This is especially true when working in conditions that are characterized by low service levels. This is especially true in situations in which the levels of service that are provided are insufficient. When compared to the JEO-RF, the SEO-RF performs better in terms of homoscedasticity due to the fact that it makes use of all of the residuals throughout the optimization step. This is due to the fact that it is able to base its conclusion on a larger sample, which is made possible by the fact that it incorporates all of the residuals in the step. As a result of this, it is able to provide more accurate results.

This gives it an edge over the JEO-RF that the latter does not have. Our investigation revealed that when the costs are symmetric—that is, when SL = 0.5—the existence of heteroscedasticity does not have a significant bearing on the performance of the strategies. This was one of the key takeaways from our investigation. This discovery is in agreement with the assumption that was stated in Conjecture 5.2. That assertion states that the performance of the techniques ought not to be significantly damaged by heteroscedasticity. This finding demonstrates that this conclusion is compatible with that assertion. This is mostly the result of the fact that JEO-RF does not make use of the feature that is responsible for the noise in order to maintain a balance in the way that its expenditures are distributed.

As a result of the fact that the error distributions are both symmetric about zero and differ from one another exclusively in terms of variance, splitting along this feature would not make a difference in terms of the costs that are involved in the process. An estimate of the sample mean can be obtained using SEO-RF with the MSE loss, whereas an estimate of the sample median may be obtained using JEO-RF. This is the explanation for the seemingly little variation in outcomes that one could observe when contrasting the two processes. Both of these approaches will receive a more in-depth analysis in the next portions of this article.

Conjecture 5.2 is supported by a number of outcomes, one of which is the fact that greater service levels make the impact that heteroscedasticity has on the relative performances of the approaches more evident. This is an additional consequence. The discovery provides evidence that the hypothesis is more plausible when there is a rise in the quality-of-service levels.

5.4.2.3 Results for kernel-based approaches

The relative performance increases JEO that the JEO-KO methodology delivers in contrast to the SEO-KO method are shown in Figure 5.3. These gains are presented for diverse degrees of heteroscedasticity as well as various parameterizations of the noise and the service level parameters. In addition, these gains are presented for a number of different noise levels. The setting in which the simulation takes place is identical to the one that was utilized when the random forest approach was put into practice. We make the discovery that the results obtained via the use of kernel optimization are, for the most part, identical to the results obtained through the use of random forests, but with implications that are not as significant.

Increasing the quantity of heteroscedasticity has minimal effect when there is already a low level of uncertainty in the data. The same holds true for random forest configurations. When dealing with higher levels of uncertainty, the existence of heteroscedasticity has a positive influence on the performance of JEO-KO when compared to that of SEO-KO, even though the impact is not nearly as clear as it is for random forests. This is because JEO-KO benefits more from the presence of heteroscedasticity than SEO-KO does. Despite this, JEO-KO performs significantly better than SEO-KO in a variety of settings. As a consequence of this, we are going to make the assumption that Conjecture 5.1 is correct given that the degree of uncertainty is high enough. On the other hand, if the homoscedasticity is perfect in a given setting, JEO-KO might not perform as well (cv noise = 0.75 and SL = 0.8).

with addition, with reference to Conjecture 5.2, the results obtained through the use of KO methods are equivalent to the results obtained via the use of random forests. When we take a look at the costs symmetrically (that is, with SL equal to 0.5), we do not observe any major differences. When higher service levels are taken into account, the performance of KO-JEO is strongly impacted by heteroscedasticity, in contrast to KO-SEO. The impact is more readily apparent when there is a higher quantity of ambient noise.

We have arrived at the conclusion that the most significant findings are related to the fundamental differences between the JEO and SEO principles, and that these discoveries are independent of the underlying machine learning methodology.

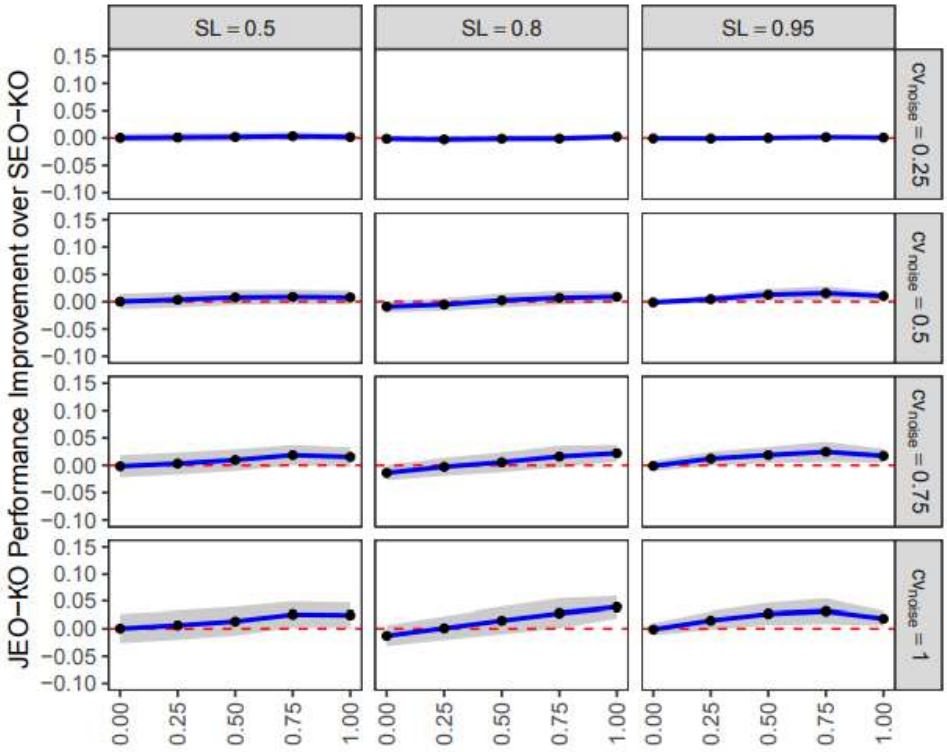


Figure 5.3: JEO-KO's cost improvement over SEO-KO depending on γ (level of heteroscedasticity) in a linear demand setting for various service levels pSL = 0.5, 0.8 and 0.95 with various levels of base noise (cv $_{noise}$). The shaded area represents a 95% confidence interval around the mean improvement

Source: *Data-driven Operations Management Data Collection And Processing Through By M.Sc. Jan Maximilian Meller, In November 2019*

5.4.3 Study 2: Prescriptive analytics at Yaz restaurant

In the following section 5.4.2, we present the results of a controlled experiment that we carried out in order to analyze the differences in performance that were seen between SEO and JEO. Even though we were able to evaluate the isolated impact of heteroscedasticity by using this approach while controlling for the amount of

uncertainty and cost asymmetries, the overall setting was far easier than the bulk of the circumstances that may emerge in the actual world. Even though we were able to study the isolated effect of heteroscedasticity by utilizing this technique while controlling for the amount of uncertainty and cost asymmetries. Specifically, the fact that we have distinguished the connection between features and demand from the connection between features and uncertainty is a major assumption that we have made. This is due to the fact that it is reasonable to assume that, given the conditions under which features are responsible for the overall unpredictability of demand, features will also have an effect on the total quantity of demand. As a consequence of this, it is not possible to monitor the effects of heteroscedasticity in the same manner that one could during a simulation experiment.

In this section, we evaluate how well JEO and SEO deal with a real-world inventory management challenge and compare their results. This issue encompasses a multitude of facets, each of which might have potentially intricate nonlinear yet unknown ties to demand. These types of linkages are frequently found in actual settings. In addition to that, there are a lot of unknown connections between this issue and demand. By contrasting the results of the two models' relative degrees of performance, we intend to validate the inferences that we drew from our simulation experiment and ensure that they are accurate.

Yaz, which is a network of quick-service eateries created in Germany, was the source of the data set. Meals provided by Yaz are composed of a limited number of fundamental components, each of which may be cooked in a huge range of different ways. Because these key components degrade in a short amount of time, Yaz needs to calculate how much of each she will require on a daily basis in order to produce them. Therefore, the structure of the problem (perishable commodities, per-unit overage, and underage fines), as was discussed before, culminates in the well-known predicament of the newsvendor.

In the following parts, you will first find a summary of the data sources that we employed, as well as the features that we derived from the data that was freely available. This will be followed by the results of our analysis, which can be found in the following sections. Following that, we will talk about our evaluation process, which is a fancy term for the thinking behind comparing and evaluating the two different approaches. In conclusion, we provide our findings on the performance of the two techniques that we designed for the application that would be used in the real world.

5.4.3.1 Data

Yaz was gracious enough to provide us with details on the income that was made at their flagship restaurant in Stuttgart, Germany, from September 27, 2013, all the way up to November 9, 2015. The demand structure for the products displays a significant level of variance with regard to both the mean demand and the coefficient of variation. This is the case regardless of whatever statistic is being considered. Because of this, we will demonstrate the performance of the model using three illustrative goods whose demand patterns are different from one another. Calamari, steak, and lamb are the three items that make up this list. As can be seen in Figure 5.4, the smoothed demand does not remain steady over time, which means that a normal newsvendor solution cannot be employed to tackle the inventory-management challenge at hand. This is due to the fact that implementing such a solution would presuppose that the distribution of demand is constant, which is not the case.

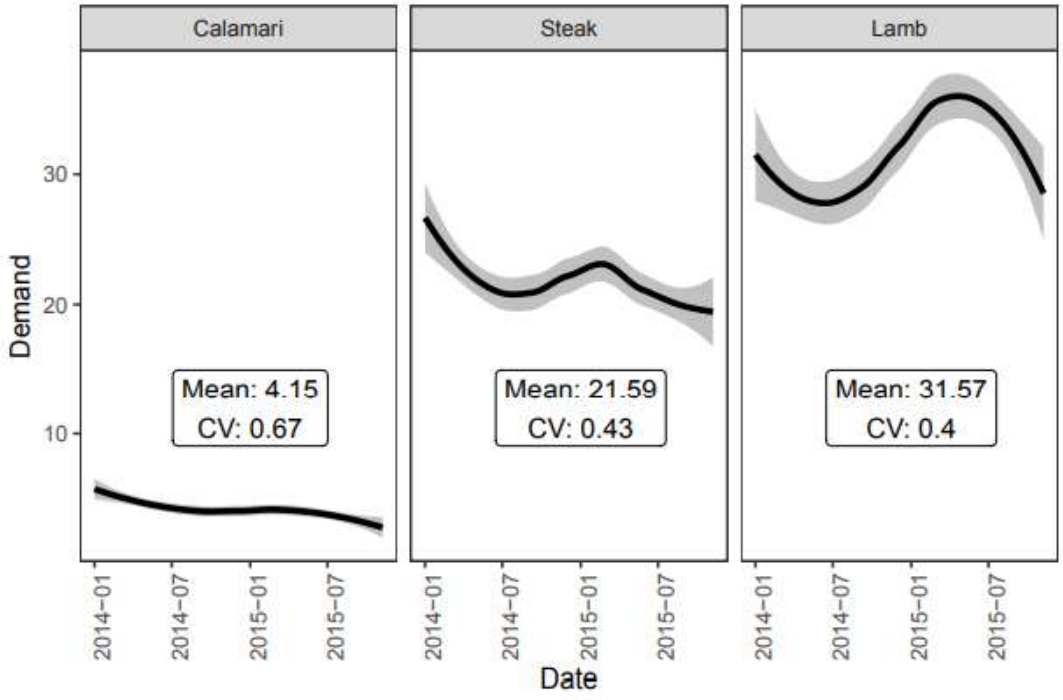


Figure 5.4: Evolution of the smoothed demand over time for different products

Source: *Data-driven Operations Management Data Collection And Processing Through By M.Sc. Jan Maximilian Meller, In November 2019*

In the past, the management of the restaurant requested that every single one of the establishment's services and/or products be available at all times. Because of this, the restaurant almost never ran out of any of its supplies, and Yaz only had to deal with stock-outs on a very seldom basis. Because there were only stock-out instances on 1.6% of the total number of days over the time range under consideration, this indicates that all three components were easily available on 98.4% of the occasions. As a result, in contrast to Bertsimas and Kallus, we do not make an attempt to adjust for data that has been repressed. We will be include a little amount of censored demand data in our comparison of JEO and SEO; however, we do not believe that this will have a significant bearing on the outcomes.

We gathered meteorological information from the databases maintained by the German Meteorological Service and then aggregated that information on a daily basis to reflect the same degree of detail as might be found in a hypothetical weather prediction for the next day. This was done because the restaurant management had a theory that the weather has a significant impact on customer demand, and he wanted to test his theory. Because the true weather predictions that were intended to be available for the next day weren't ready in time, we had no choice but to depend on the actual weather information from the day before instead. Even though this information would not be available at the time a choice is made, the features that we identified from this data are likely to be equivalent to a weather prediction for the next day.

This is the case despite the fact that we did not have access to this information. We were able to collect 168 characteristics for each product by extracting structural information about the underlying time series (for example, the rolling mean demand for the same daily). These 168 characteristics were obtained from the raw data. In Table 5.2, the most important qualities are dissected into their respective groups after being compiled there.

5.4.3.2 Evaluation procedure

We acquire a data set with the names TNY az = 1,..., N Y azu after cleaning and preprocessing the raw data. This data set contains NY az = 672 demand observations. We employ a five-fold cross-validation to test the performance of our model on this data set. This involves randomly dividing T NY az into five subgroups that are nearly the same size.

Table 5.2: Examples of relevant features for the product Steak

Source	Feature
Time Series	Average aggregate demand (for all products) on same weekday for the last two weeks
	Average aggregate demand (for individual products) on the same weekday over the last three weeks
	Aggregate demand (for all products) the day before
Calendar	Is December
	Is Saturday
	Is special day (Event, Holiday, etc.)
Weather	Air temperature two days ago
	Average Air temperature over last four days
	Average duration of sunshine over last five days

Let's use the notation to denote the indexing function that converts each unique observation to one of the five possible divisions in the data. After that, the prescription function is indicated by the notation L ((x), D), and it is calibrated with the k-th component of the data deleted from consideration. As a consequence of this, in order to assess how well our prescription model performs on the kth section of the data, we first calibrate it five times by making use of the various components that make up the training data set. Following this, we compute the estimated costs of the mismatch using the following formula:

$$\bar{c}_m = \frac{1}{N_{Yaz}} \sum_{i=1}^{N_{Yaz}} C(d_i, \hat{q}_m^{-\phi(i)}(x_i))$$

Again, in order to provide a more precise assessment of the capabilities of the models, we supply the number m, which is the percentage cost improvement compared to the sample average approximation (SAA) benchmark for each model. This value is provided so that the capabilities of the models may be evaluated more precisely.

After adjusting for cost asymmetry in terms of the service level, uncertainty within the data, and heteroscedasticity in our simulation experiments, which are detailed in section 5.4.2, we quantified these elements in our experiment based on real-world data. Our experiment was conducted in the actual world. In order to do this, we calculate the out-of-sample mean squared error (MSE) of the predictions produced by the SEO approach as a measure of the residual uncertainty (in other words, as a comparable metric to the cvnoise parameter in our simulations):

$$\varepsilon_{MSE} = \frac{1}{N_{Yaz}} \sum_{i=1}^{N_{Yaz}} (d_i - \hat{\mu}_{RF}(x_i))^2$$

We also quantify the heteroscedasticity in the residuals of the random forest predictions. In order to do this, we calculate the state-dependent coefficient of variation across all of the historical data d1,d nlt that have been sorted into a given leaf l in a tree t of our SEO random forest:

$$cv_{lt} = \frac{\sqrt{(\sum_i (d_{ilt} - \frac{1}{n_{lt}} \sum_i d_{ilt})^2}}{\frac{1}{n_{lt}} \sum_l d_{ilt}}$$

Then we determine the standard deviation for each tree t separately:

$$sd_t = \sqrt{\sum_l \left(cv_{lt} - \frac{1}{L_t} \sum_l cv_{lt} \right)^2}$$

Because it determines the degree to which the coefficient of variation varies based on the actual state (that is, the leaf into which an observation is sorted), this standard deviation is used to identify whether or not the residuals exhibit heteroscedasticity. To put it another way, this standard deviation is used to determine the degree to which the coefficient of variation fluctuates. After that, an indicator of heteroscedasticity is obtained by adding the standard deviations of all of the individual variables. RF:

$$\gamma_{RF} = \frac{1}{T} \sum_t sd_t$$

By applying this methodology, we are in the position to provide an estimation for the state-dependent uncertainty that is associated with the SEO-RF technique. This uncertainty may be thought of as a close approximation for the heteroscedasticity that exists in the residuals.

5.4.3.3 Results for random forest-based approaches

Following the implementation of JEO-RF and SEORF on the problem of inventory management at Yaz's firm, this section provides a summary of the most significant discoveries that were made. The percentage reductions in expenditures that have been experienced are displayed in Figure 5.5.

$$\delta_{m,SAA} = \frac{\bar{c}_m - \bar{c}_{SAA}}{\bar{c}_{SAA}} = \frac{\Delta_{m,SAA}}{\bar{c}_{SAA}},$$

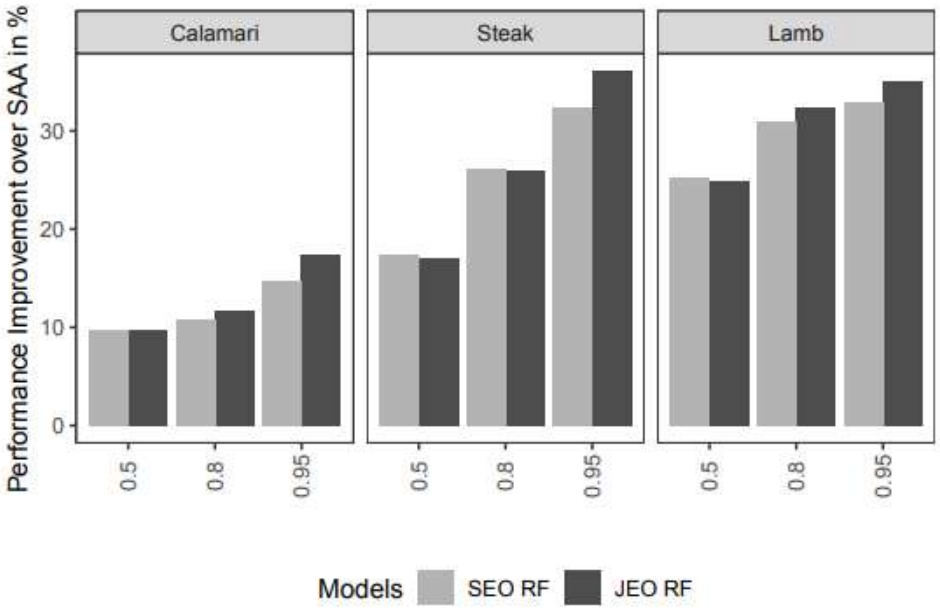


Figure 5.5: Percentage cost improvement $\delta_{m,SAA}$ over SAA for the SEO-RF and the JEO-RF models

Source: *Data-driven Operations Management Data Collection And Processing Through By M.Sc. Jan Maximilian Meller, In November 2019*

In connection with SAA for the many different service levels that are offered by JEO-RF and SEO-RF. As can be seen in Figure 5.5, both solutions offer a considerable improvement in terms of minimizing the costs associated with mismatches in compared to the SAA benchmark. This gain may be attributed to the fact that the number of mismatches is reduced. In addition, we discover that the performance of the two strategies is equivalent at the 0.5 service level, with the exception of the fact that SEO-RF has considerably lower expenses. These findings are in line with the findings of our simulation, which demonstrated that neither technique was preferable than the other for the 0.5 service level. These findings are compatible with the findings of our simulation. This occurred as a direct result of the fact that the symmetric mismatch cost structure that was developed as a result yielded prescriptions that were equivalent to one another.

When we examine the outcomes of our simulation experiments for more service levels, we find that they are equivalent to the following: As the gap between overage and underage fees continues to expand, the competitive advantage that JEO has in terms of pricing has the potential to grow even bigger. For instance, when comparing JEO-RF and SEO-RF for the 0.95 service level, we find that there are considerable differences between the two (for example, for steak, a cost improvement over SAA of 36% for JEO-RF and 32% for SEO-RF, and for calamari, a cost improvement over SAA of 17% vs 15% for SEO-RF).

Table 5.3: Measures for the forecast accuracy and heteroscedasticity of residuals of the SEO approach (upper part) and cost improvements of JEO over SEO for a 0.95 service level, and with the p-value results of a t-test (lower part)

	Calamari	Steak	Lamb
MAE	2.00	5.53	7.17
MSE	6.72	54.35	89.00
γ_{RF}	0.35	0.18	0.16
Δ_{JEO}	0.02	0.07	0.02
$\delta_{JEO}(\%)$	6.31	7.32	1.58
p-value	0.008	0.086	0.599

The outcomes of our simulations, which were provided in Section 5.4.2, revealed, in keeping with Hypothesis 5.1, that the degree of heteroscedasticity present in the residuals was the key factor responsible for the discrepancies in performance shown between the SEO-RF and JEO-RF techniques. This was shown to be the case by demonstrating that the residuals included heteroscedasticity. We were able to determine the source of the remaining uncertainty by applying descriptive statistics to the residuals of the SEO, which are shown in Table 5.3. The Yaz data set was used to confirm these results. This is a presentation of Table 5.3. We make the discovery that calamari, followed by steak and lamb, has the highest heteroscedasticity (as measured by YRF) in the leaf nodes. In third place is lamb. However, as can be seen in Table 5.3, we were able to achieve the greatest relative improvement with the steak (7.32%), followed by the calamari (6.31%), and finally the lamb (1.58%). This went against our assumptions, which were that JEO would have the biggest cost improvement for calamari goods. However, this turned out to be the case.

This finding may be explained by taking into account the combination of two effects that are in direct opposition to one another: While we find that calamari has the highest degree of heteroscedasticity, we also find that calamari has the best level of overall forecast accuracy. This is the case despite the fact that calamari has the highest amount of heteroscedasticity. This demonstrates that the amount of residual uncertainty for this product, which also affects the relative cost advantage of JEO in comparison to SEO, has reached its lowest point: The forecasting error of lamb is more than 3.5 times larger than the mean absolute error (MAE) of calamari, which significantly reduces the amount of variance in performance that can be accomplished utilizing the two different methodologies.

We have come to the conclusion that the performance advantage that JEO possesses over SEO in relation to calamari is statistically very significant, with a p-value of 0.008. As a direct result of this, we have arrived at the realization that the results of the simulation research and the conclusions that we obtained from the case study of the actual world are compatible with one another. This lends further credence to our theories, which indicate that heteroscedasticity is a crucial component leading to JEO's cost advantage over SEO. This finding gives further evidence for our hypothesis.

We tested the consistency of our findings over a large number of different model parameter combinations as the very last step in this process. Table 5.4 presents the mean absolute cost improvements (JEO) and the scaled absolute cost improvements

(JEO c SEO) for steak for a variety of combinations of service levels and the model-specific tuning parameters ntrees, which represents the number of trees, and minnode, which is the minimum number of observations in a node as an additional split. Additionally, the table also presents the scaled absolute cost improvements (JEO c SEO). In addition to this, the scaled absolute cost improvements are shown in Table 5.4 (JEO c SEO). We have demonstrated that the JEO method, when compared to the SEO technique, results in much reduced mean costs for all feasible parameter values, with the exception of the 0.5 service level. On the other hand, our cost decrease is only highly statistically significant in four of the different configurations that we analyzed.

Table 5.4: Mean absolute performance differences between SEO and JEO for steak, depending on model configurations. The last column divides the absolute performance difference by the mean mismatch cost of the SEO model for the specific configuration to illustrate the magnitude of the improvements

SL	n_{trees}	min_{node}	Δ_{JEO}	LB 95% CI	UB 95% CI	$\frac{\Delta_{JEO}}{\bar{c}_{SEO}}$
0.5	100	5	−0.01	−0.08	0.05	−0.00
0.5	100	15	0.00	−0.06	0.05	0.00
0.5	100	30	−0.03	−0.08	0.02	−0.01
0.5	500	5	−0.01	−0.06	0.05	−0.01
0.5	500	15	−0.02	−0.07	0.03	−0.02
0.5	500	30	−0.05	−0.09	0.00	−0.04
0.8	100	5	0.01	−0.05	0.07	0.00
0.8	100	15	0.03	−0.03	0.09	0.01
0.8	100	30	0.10	0.03	0.17	0.05
0.8	500	5	0.00	−0.05	0.05	0.00
0.8	500	15	0.02	−0.03	0.07	0.01
0.95	500	5	0.05	−0.02	0.12	0.06
0.95	500	15	0.07	−0.01	0.15	0.08
0.95	500	30	0.09	0.01	0.18	0.10

5.4.3.4 Results for kernel-based approaches

Following this, we will talk about the most important things that we learned from implementing JEOKO and SEO-KO to the issue of inventory management at Yaz's company. Both JEOKO and SEO-KO are programs that run in the web browser. The assessment method which we made use in our random forest strategy is identical to the one which we make use of in this particular setting. In compared to SAA, JEO-KO and SEO-KO are able to achieve a greater percentage of cost savings than SAA does (as shown in Figure 5.6). This cost savings is displayed for each of the available service levels. This comparison is carried out for each of the several service levels that are accessible.

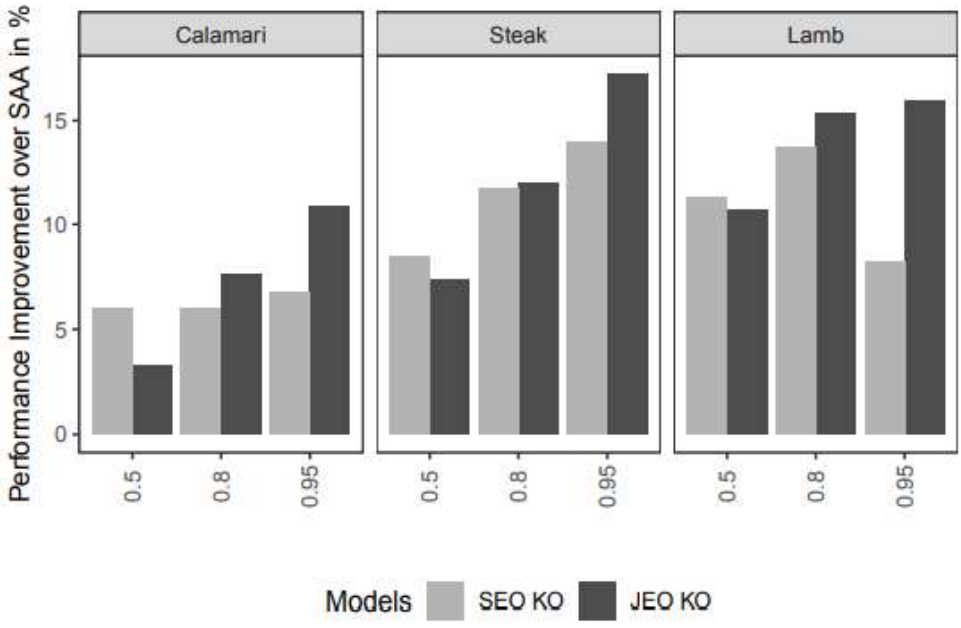


Figure 5.6: Percentage cost improvement $\delta_{m,SAA}$ over SAA for the SEO and the JEO kernel optimization models

Source: *Data-driven Operations Management Data Collection And Processing Through By M.Sc. Jan Maximilian Meller, In November 2019*

The fact that the mismatch costs for SEO-KO and JEO-KO are lower than they are for SAA leads us to the conclusion that the findings for KO are mostly compatible with those obtained using random forests. This is because the mismatch costs for SEO-KO

and JEO-KO are lower than they are for SAA. It has come to our attention that SEO-KO generates a superior cost improvement for SL = 0.5 in every scenario with the exception of random forest. In comparison, this is not the case with the random forest.

JEO-KO, which is an alternative to SEO, has demonstrated to deliver higher results when high service standards are involved. This is because JEO-KO takes into account user feedback. On the other hand, the kernel technique cannot evaluate heteroscedasticity in the same manner as the random forest approach can. This is because the random forest approach uses randomization. As a consequence of this, we are unable to draw any conclusions about the subject of whether or not any disparities in the performance of JEO-KO and SEO-KO when applying the kernel-based technique may be the result of heteroscedasticity in this specific practical setting.

When looking at the data for an inventory management problem similar to that of a newsvendor, in which shifts in demand are driven by observable features, we assessed the performance of two fundamentally different methodologies and contrasted the outcomes they produced. Our in-depth research gives the first comprehensive analysis of the performance differences between the two concepts. To do this, we compared the various implementations for SEO and JEO while employing two different underlying machine learning algorithms. As part of our inquiry into the connection between search engine optimization and job entry optimization, we drew these parallels and made these comparisons. In addition, the JEO methodology, which is a novel approach that was just recently introduced and is predicated on random forests, is an option that may be considered for the purpose of locating the optimal inventory quantities.

In an initial analysis, we established that the SEO and JEO approaches both achieve the same expected mismatch costs given that the residuals preserve homoscedasticity. On the other hand, we found that JEO outperforms SEO as the degrees of heteroscedasticity grow. We observed that the effect of heteroscedasticity was the same for the two more advanced nonlinear techniques in both a study that was based on a simulation and one that was based on a data set that was obtained from the actual world. Both of these studies were conducted by us. In conditions characterized by high heteroscedasticity and high remaining uncertainty (that is, low forecast accuracy), in conjunction with a highly asymmetric cost structure, the analysis of performance differences on our real-world data set suggests that both the random forest-based and the kernel optimization-based JEO approaches outperform their respective SEO counterparts. This is the case despite the fact that the random forest-based and kernel

optimization-based JEO approaches are both based on SEO. This is the case regardless of the fact that the JEO strategy that uses random forests is based on kernel optimization. In addition, we have demonstrated that the JEO strategy that is based on random forests works noticeably better than the JEO method that is based on kernels when the data is taken from the real world and applied to them. In addition to this, we developed a measure to assess the level of heteroscedasticity by making use of its tree-based structure in our analysis. This made it possible for us to get fresh insights into the nature of the unanswered questions that remained.

As a consequence of this, the use of JEO models is permissible within the parameters of settings that include high service levels, low forecasting accuracy, and the anticipation of heteroscedasticity. This is due to the fact that the internal structure of the JEO models has been adapted to circumstances that have these features. On the other hand, solutions for search engine optimization perform very well when implemented in conditions in which prediction accuracy is high and mismatch costs are symmetrical. In addition to the competitive performance of the existing SEO strategies in these kinds of environments, they are also adaptive in terms of the prediction model that lies behind them: SEO tactics, on the other hand, may directly benefit from advancements that lead to improved prediction models, but JEO methods need to be modified to the specific context in which they will be implemented. This is due to the fact that SEO tactics just serve as building blocks, while the optimization logic described in the next sentence is still applied.

Authors Details

ISBN: 978-81-19534-45-6

Abdelhamid ZAIDI, is an Associate Professor in the College of Science at Qassim University in Saudi Arabia. He has a PhD in Statistics and Stochastic Modeling from University Grenoble-Alpes (France) and an Engineering degree in Computer Science and Applied Mathematics from ENSIMAG Grenoble (France). He works mainly on the development of computational methods applied to various subjects of signal and image processing and artificial intelligence. He also has many contributions in the field of artificial intelligence. His research work was published in many top ranked journals. He is also the author of three books covering numerical analysis, algorithmic, probability, and statistics.

Renato Racelis Maaliw III, is an Associate Professor and Researcher at the College of Engineering in Southern Luzon State University, Lucban, Quezon, Philippines. He has a doctorate degree in Information Technology with specialization in Machine Learning, a Master's degree in Information Technology with specialization in Web Technologies, and a Bachelor's degree in Computer Engineering. His area of interest is in artificial intelligence, computer engineering, web technologies, software engineering, data mining, machine learning, and analytics. He has published original research articles, a multiple time best paper awardee for various IEEE sanctioned conferences; served as technical program committee for world-class conferences, author, editor and peer reviewer for reputable high-impact research journals.

Mrs. K. P. Maheswari, MCA., M.Phil., NET (Computer Science) is Assistant Professor of Computer Applications at Fatima College, Madurai. She has rich experience in teaching. She has received "Women Transforming Nation Awards 2023 – Certificate of Appreciation for Dedication" from Women Lead. Her research interests include Machine Learning, Deep Learning, Artificial Intelligence and Network Security. She has authored book chapters and several publications in a reputed journals. She has presented papers in various Conferences and seminars at National and International level. She has also served as a subject matter expert for workshops and seminars.Her professional achievements have included obtaining Microsoft (MTA - Microsoft Technology Associate) International certifications in Python Programming, HTML 5, and Security Fundamentals.

Dr. Haewon Byeon, received the Dr Sc degree in Biomedical Science from Ajou University School of Medicine. Haewon Byeon currently works at the Department of Medical Big Data, Inje University. His recent interests focus on health promotion, AI-medicine, and biostatistics. He is currently a member of international committee for a Frontiers in Psychiatry, and an editorial board for World Journal of Psychiatry. Also, He were worked on a 4 projects (Principal Investigator) from the Ministry of Education, the Korea Research Foundation, and the Ministry of Health and Welfare. Byeon has published more than 343 articles and 19 books.

Xoffencer International Publication
838- Laxmi Colony. Dabra,
Gwalior, Madhya Pradesh, 475110
www.xoffencerpublication.in

MRP: ₹ 450/-
ISBN
9 788119 534456