



REVA
UNIVERSITY

Bengaluru, India

A Project Report on
Minimizing Losses on Trials of “Strategy Builder”
Tool using Business Analytics

Submitted in Partial Fulfilment for Award of Degree of
Master of Business Administration
In Business Analytics

Submitted By
Tushar Nigam
R18MBA57

Under the Guidance of
Dr. Rashmi Agarwal
Assistant Professor & Mentor
RACE, REVA University

REVA Academy for Corporate Excellence - RACE
REVA University
Rukmini Knowledge Park, Kattigenahalli, Yelahanka, Bengaluru - 560 064
race.reva.edu.in

August, 2022



Candidate's Declaration

I, **Tushar Nigam** hereby declare that I have completed the project work towards the second year of Master of Business Administration in Business Analytics at, REVA University on the topic entitled **Minimizing Losses on Trials of “Strategy Builder” Tool using Business Analytics** under the supervision of **Dr. Rashmi Agarwal, Assistant Professor**. This report embodies the original work done by me in partial fulfilment of the requirements for the award of degree for the academic year 2022.

Place: Bengaluru

Date: 27th Aug 2022

Name of the Student: Tushar Nigam

Signature of Student



Certificate

This is to Certify that the Project work entitled **Minimizing Losses on Trials of “Strategy Builder” Tool using Business Analytics** carried out by **Tushar Nigam** with **R18MBA57**, is a bonafide student of REVA University, is submitting the second year project report in fulfilment for the award of **MBA** in Business Analytics during the academic year 2022. The Project report has been tested for plagiarism, and has passed the plagiarism test with the similarity score of less than 15%. The project report has been approved as it satisfies the academic requirements in respect of the project work prescribed for the said Degree.

Signature of the Guide

Name of the Guide

Dr. Rashmi Agarwal

Signature of the Director

Name of the Director

Dr. Shinu Abhi

External Viva

Names of the Examiners

1. Vaibhav Sahu, Strategic Cloud Engineer, Google
2. Abhishek Sinha, Data Science Manager, Capgemini

Place: Bengaluru

Date: 27th August 2022



Acknowledgement

I would like to thank our Chancellor Dr. P Shyama Raju, Vice Chancellor Dr. M. Dhanamjaya, Registrar Dr. N. Ramesh, Controller of Examinations Dr. Beena, and Director of Corporate Trainings Dr. Shinu Abhi from REVA University without whom I would neither have been able to complete this project nor would I have made it through my MBA degree.

I would want to thank my mentor Dr. Rashmi Agarwal, Assistant Professor for guiding me throughout the project and for showing me the path I could walk for this industrial project. I would also thank all the professors at REVA Academy of Corporate Excellence for teaching us awesome methodologies throughout the academic year. Finally, I would like to thank my family for patiently supporting me while I was busy working on this project.

Place: Bengaluru

Date: 27th Aug 2022



Similarity Index Report

This is to certify that this project report titled “Minimizing losses on trials of ‘Strategy Builder’ Tool using Business Analytics” was scanned for similarity detection. Process and outcome are given below:

Software Used: **Turnitin**

Date of Report Generation: 25th Aug 2022

Similarity Index in %: 8%

Total word count: 6646

Name of the Guide: Dr. Rashmi Agarwal

Place: Bengaluru

Date: 27th Aug 2022

Name of the Student: Tushar Nigam

Signature of Student

Verified by: M N Dincy Dechamma

Signature

Dr. Shinu Abhi,

Director, Corporate Training

List of Abbreviations

Sl. No	Abbreviation	Long Form
1	B2B	Business To Business
2	Q1	Quarter 1
3	Pvt. Ltd.	Private Limited
4	SQL	Structured Query Language
5	API	Application Programming Interface
6	EDA	Exploratory Data Analysis
7	USP	Unique Selling Point
8	MBA	Market Basket Analysis

List of Figures

No.	Name	Page No.
Figure No. 5.1	CRISP-DM Flow	17
Figure No. 7.1	ER Diagram of Sales System in RLP	22
Figure No. 7.2	Sample of raw data used for the study	22
Figure No. 7.3	Sales count of new product	23
Figure No. 7.4	Sales data by Domain	24
Figure No. 7.5	Sales data by Region	24
Figure No. 7.6	Sales data by No_of_Products_already_purchased	25
Figure No. 7.7	Sales data by No_of_Years_Of_Association_Binned	25
Figure No. 7.8	Sales data by No_of_Bugs_Reported	26
Figure No. 7.9	Heatmap of Correlations	26
Figure No. 8.1	Unique values of Nominal Variable - Region	27
Figure No. 8.2	Encoded Region data	27
Figure No. 8.3	Unique values of Nominal Variable	28
Figure No. 8.4	Encoded Domain data	28
Figure No. 8.5	Unique values of Variable - No_Of_Years_Of_Association_Binned	28

Figure No. 8.6	Unique values of Variable - No_Of_Bugs_Reported_Binned	28
Figure No. 8.7	Encoded No_Of_Years_Of_Association data	29
Figure No. 8.8	Encoded No_Of_Bugs_Reported_Binned data	29
Figure No. 8.9	Correlation Heat Map after Data Preparation	29
Figure No. 9.1	Decision Tree Outcome	31
Figure No. 9.2	Logistic Regression Coefficients	32
Figure No. 9.3	Combination of multiple Decision Tress in Random Forest	33
Figure No. 9.4	Subset of data in XGBoost	34
Figure No. 9.5	Understanding Metrics of Market Basket Analysis	35
Figure No. 9.6	Market Basket Analysis Output	35
Figure No. 9.7	Optimal no. of Clusters	36
Figure No. 10.1	Decision Tree Model Evaluation Metrics	37
Figure No. 10.2	Logistic Regression Model Evaluation Metrics	37
Figure No. 10.3	Bernoulli Naïve Bayes Model Evaluation Metrics	38
Figure No. 10.4	Random Forest Model Evaluation Metrics	38
Figure No. 10.5	XGBoost Model Evaluation Metrics	39
Figure No. 11.1	Model Deployment Design	40

List of Tables

No.	Name	Page No.
Table No. 10.1	Model Accuracy Results	40

Abstract

RLP Software Pvt. Ltd. has developed a new product “Strategy Builder” and targets to sell it to its existing customers. The company has a customer base of around 11000+ customers. The new product was launched in Q1,22 and focuses on Customer Incentive Management. It aims to give indefinite ease and flexibility to create rules/schemes for a business to incentivize their customers for their orders/purchases. To demonstrate the significance and usability of the product, the company provides a trial version of the tool to potential customers for a specific time.

In past ~1.5 years, the trial version was allocated to 6560 customers, out of which only 2886 have onboarded and the rest of them did not sign the deal. The conversion rate of the trial is only 44%. Moreover, for each trial, there is a cost associated, and every time a customer does not sign the deal after the trial, this amount gets wasted. The loss that the company is facing in these trials had been huge and the company is looking forward to ways to reduce this.

Therefore, this project focuses on applying Business Analytics techniques for solving the problem. With the existing sales data, useful insights could be found to give recommendations to the company on where they are doing good and what they must change to improvise their sales strategies. The ideal scenario for reducing the loss is to only give trials to the customer who will buy the product for sure. While the future cannot be exactly predicted but the probability can be found with the help of predictive analytics. Hence, **Binary Classification** techniques were applied to the historic data to identify the most probable customers.

Another way to reduce this loss is by cross-selling this new product. The idea is to find the right set of products with which the sale of new product is more so that the customers who own this set of products or purchase the same set-in future can be considered as a probable customers of the new product, as well as the trial can be given to them with fewer chances of losses. To achieve that, **Market Basket Analysis** was done for understanding the purchasing patterns of existing customers.

Additionally, **Clustering** was done to segment customers based on similar characteristics and understand which cluster of customers are buying the product most and the one buying the least

so that measures could be taken to move the customers from low-performing clusters to high-performing ones.

With the application of the above-mentioned techniques, there were multiple ways found to reduce these losses and improve the sales of Strategy Builder. Random Forest came out to be the best performing predictive model with an accuracy of 82.56% which can evaluate the probability of a customer's purchase intention. Clustering provided the set of customers having similar traits and are least responsive in purchasing the product. This data will lay the foundation for targeting these customers and improving sales. From the purchase pattern of the customers, it was found that customers who bought products "CLM" and "CPQ" both are most likely to buy "Strategy Builder" and hence become the best candidate for cross-selling the new product.

Keywords: Data Mining, Binary Classification, Clustering, Market Basket Analysis, Ensemble Models, One-Hot Encoding, Cross Selling

Contents

Candidate's Declaration.....	2
Certificate.....	3
List of Abbreviations	6
List of Figures	6
List of Tables	7
Abstract	8
Chapter 1: Introduction	11
Chapter 2: Literature Review.....	12
Chapter 3: Problem Statement	16
Chapter 4: Objectives of the Study	17
Chapter 5: Project Methodology	18
Chapter 6: Business Understanding	21
Chapter 7: Data Understanding.....	23
Chapter 8: Data Preparation.....	28
Chapter 9: Data Modeling.....	31
Chapter 9: Data Evaluation.....	38
Chapter 10: Deployment	41
Chapter 11: Analysis and Results	42
Chapter 12: Conclusions and Recommendations for future work	43
Bibliography	44
Appendix.....	46
Plagiarism Report.....	46
Publications in a Journal/Conference Presented/White Paper	48

Chapter 1: Introduction

Product trial is a unique form of advertising as it provides a direct experience to the customer. It is much better than documentation, PPT and theories as it's the working model of the solution the product provides. With the advancement of technology, new products are emerging constantly and rapidly. But not all of them become popular and profitable. Studies mention that 95% of new products fail (Emmer, 2018). Not enough exposure to the customers could be one of the major reasons for this as the customers are not familiar with the capability and potential of the product (Sun, 2017). Therefore, a trial becomes very important for customers to evaluate the product and make purchase decisions.

Although the trial is an effective way of promotion, it is also a very expensive way of marketing. There are a lot of resources involved in the trials and there is a cost associated with every resource. These resources could be user training, operations, technical support, cloud computation and storage, licenses, etc. The company loses revenue if the conversion rate of customers after trials is low.

Therefore, there should be a strategy based on which the trials are given to the customer so that losses are minimal. Purchase intent is one of the main factors that need to be looked at before giving any trial to the customer (Contributor, 2017). It is the probability that a customer will buy a product. Once the purchase intent is evaluated, the decision of giving a trial or not can be taken. Purchase Intent can be evaluated with the help of predictive analytics.

RLP Software Pvt. Ltd., a software development company that had launched a new product named Strategy Builder. The conversion ratio after the trial is just 44%. Based on the data provided, the business has lost \$8,450,200 on unsuccessful trials. The objective of the study is to minimize these losses. Predictive Analytics would be used to identify the customers with higher purchase intentions. It would also be interesting to know if the new product is purchased more by the customers who own a specific set of existing products. Affinity analysis is a powerful tool to get this information. Clustering can also provide useful insights into the purchase trends of customers with similar traits.

Chapter 2: Literature Review

Existing published works were reviewed to identify the right set of solutions for this problem. The Literature review done for this project is divided into 5 parts. The first part focuses on the importance of trials after the launch of a product, the second one explains the concept of Purchase Intentions and the last 3 parts are dedicated to various machine learning algorithms to improve the trial conversion.

Importance of Product Trial

Whenever a new product is launched, there is always an expectation of great sales by the company. Most of the time the results are not as expected and from there, the retrospect and improvisation start. For customers who have not purchased the product, the company can sell products through the improvement of product satisfaction. However, the more the product satisfaction is improved, the more services need to be paid for, which increases the cost of the company. The company wants to sell its products successfully with minimal cost. To maximize the benefits for the company, a data mining system has to be developed to help companies focus on the unpurchased customers with the strongest purchase intentions (Mai, 2021).

Purchase Intention

It is the willingness of the customer to buy a certain product. It's a dependent variable based on several independent factors and is a measure of a potential customer's attitude towards purchasing a product. It is an important metric in designing marketing strategies (Team, 2021). Price had been an important variable in influencing purchase intentions but other variables such as knowledge, experience and quality are important in the process of customer's purchase decisions too (Mirabi, 2015). For that matter, the duration of association of customers with the company also plays an important role. Another factor is the perceived value as described in Figure No. 2.1, which is considered directly proportional to the purchase intent (Fathy, 2015). With these vast possibilities, the first step is to identify the factors affecting customers' purchase intention.

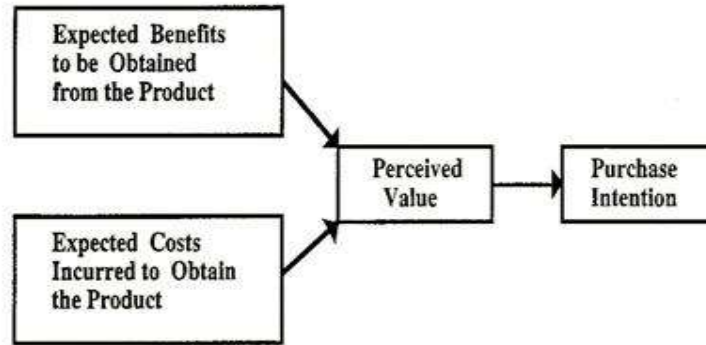


Figure No. 2.1 Perceived value strongly influences Purchase Intention (Team, 2021)

Binary Classification

It is a powerful technique that could be used to predict purchase intention. From the business point of view, two prediction models are needed (Jaiswal, 2011). The prediction model which uses some or all behavioral data could only be applied to the company's existing client base, for whom all data is available - the so-called cross sale (Bole & Papa, 2011). On the other hand, it makes sense to build a separate prediction model using only the socio-demographic data if the company were interested in acquiring customers that have not yet established a business relationship with the company. Since the modeling objective became clear, the next step was to identify the type of classification.

Binary classifiers are a widely used option, however, there is an alternate method available called one-class classification, which works only with a single class of data (Bellinger, 2012). This technique is used when class distribution is imbalanced i.e. one of the classes is much higher in data than the other. In such cases, the former may not perform well. Considering the data in this study the classes are divided in a ratio of 44:56, which makes it a balanced dataset. Therefore, it was decided to go with the binary classifiers only for this study.

Market Basket Analysis

Once the purchase intent is evaluated, several other important patterns can be discovered from data collected from the business. Affinity analysis is one such technique for achieving that. For identifying the affinity between the products there have been multiple pieces of research done. Most of them involved using Market Basket Analysis. Prior study (Kaur & Kang, 2016) shows that this way is useful in finding out interesting patterns from a large amount of data, predicting

future association rules as well as the right methodology to find out outliers. Among various methods for affinity analysis, the apriori algorithm is found to be better for association rule mining. There are various complexities involved in the apriori algorithm like the exhaustive scan of the database multiple times and as the input becomes larger, the computation time increases significantly (Gupta & Mamtara, 2014). Still, this is a popular choice among data scientists. The data collected from the business is moderate in size, hence these limitations will have negligible effects on our study.

Some studies tried a completely new way of improvising the MBA. One such study used the Map/Reduce of Cloud Computing (Woo & Xu, 2011). The algorithm has been executed on EC2 small instances of AWS with nodes 2, 5, 10, 15, and 20. The execution times of the experiments show that the proposed algorithm gets better performance while running on a large number of nodes to a certain point. However, from a certain point, Map/Reduce does not guarantee to increase the performance even though we add more nodes because there is a bottleneck for distributing, aggregating, and reducing the data set among nodes against the computing powers of additional nodes. Hence, the scope of the study is limited to the basic algorithm for finding associations.

Clustering

The solution for the problem that the business is facing could also be achieved by clustering which could help in understanding similarities between the customers who are or are not purchasing the new product. This technique divides the input into a certain number of clusters based on similar traits. The idea is to take advantage of these common traits to identify the solution. Clustering is an unsupervised technique where the solution is not unique and it strongly depends upon the analyst's choices. For instance, in k-means clustering, the results depend on the no of clusters passed as input to the algorithm. Its analyst's efficiency that an optimal number of the clusters has to be passed. Clustering always provides groups, even if there is no group structure (Omran, 2007).

K-Means arguably is the most popular clustering method. This is why studying its properties is of interest not only to the classification, data mining and machine learning communities but also to the increasing numbers of practitioners in marketing research, bioinformatics, customer management, engineering and other application areas (Kodinariya & Makwana, 2013). The

algorithm first starts with randomly selected centroids which act as a starting point of the cluster, then iterative calculations are performed to optimize the positions of centroids (Bu, 2018). These iterations stop when the defined number is achieved or if there is no movement of the centroid, which means the clustering is successful.

The information gained from the above studies laid the foundation for understanding the actual problem and identifying the right objectives for solving the problem that the business is facing.

Chapter 3: Problem Statement

Considering the Sales of the new product as per historic data, the overall loss so far is as below:

Total Number of Customers who got the trials: 6560

Customers who purchased the product: 2886

Cost of each trial: \$2300

Loss = (Total Number of Customers who got the trials) – (Customers who purchased the product) * (Cost of each trial)

Based on the above formula, the loss comes around to \$8,450,200. This is a big amount and the company loses out on Revenue because of this. Hence, the company is looking out for ways to “**Minimize the Losses on Trials of Strategy Builder Tool**”.

Chapter 4: Objectives of the Study

Understanding the problem is one of the major steps in solving it. Sometimes, the problem mentioned is not the actual problem but is just a symptom of it. In this case, the company mentioned they wanted to minimize the losses on trials, which can only be done by increasing the sales, but whether that can be done directly, or it requires the problem to be broken into multiple verticals was the first challenge. After analysis, it was found that this problem cannot be solved directly as the sales are dependent on multiple factors like Marketing, Competition, Pricing, Purchase Patterns, Usability, Interface, USP, etc. and each of them needs to be understood and improved. Since these factors are high in number, the most important ones are picked for applying Analytics Techniques. Below are the objectives that the study deals with:

1. **Identify the most probable customers to whom this product can be sold.** Losses can be minimized when most of the customers buy the product after trial. Technically, if the trial is given only to the most probable customers there are higher chances that the product gets sold and there is no loss on trial. Predictive analytics is a powerful tool for such scenarios.
2. **Identify the combination of products with which the new product gets sold frequently.** This can enable the business to find the right customers based on their purchase history. Also, this information can lay the foundation for cross-selling. Market Basket Analysis brings out such patterns from historic data.
3. **Identify the set of customers who are least responsive in buying the new product.** Clustering the customers can give significant information as the groups are formed based on similar characteristics. Identifying the cluster in which the least number of the customers have purchased the new product can be targeted similarly for Marketing, Promotion, Discounts, etc. to move them from lower performing cluster to the higher performing one.

After identifying the above objectives, the next step was to choose the right framework for proceeding. CRISP-DM methodology was selected for this study and is explained in the next chapter.

Chapter 5: Project Methodology

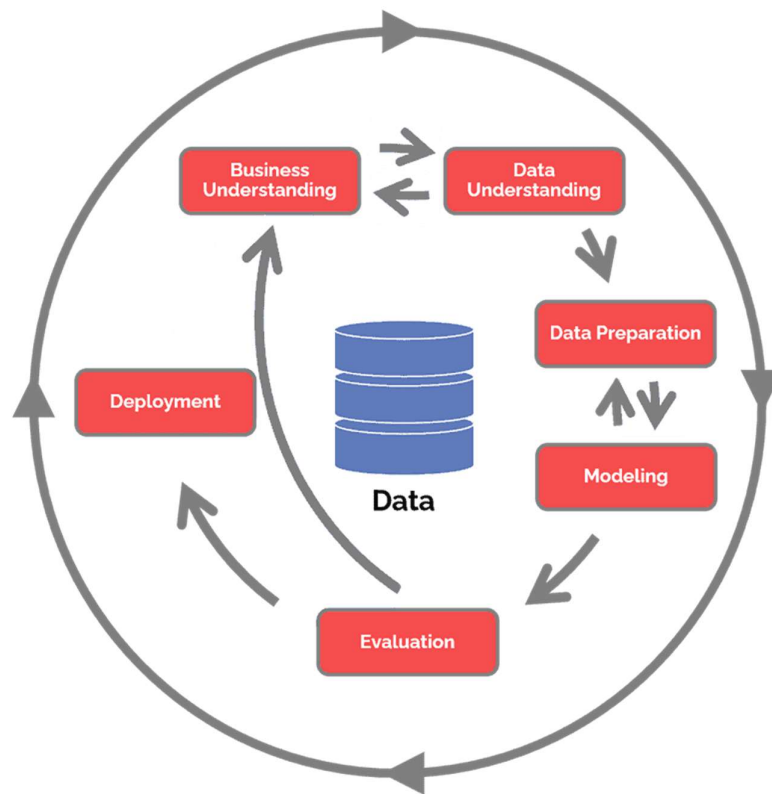


Figure No. 5.1 CRISP-DM Flow (Hotz, 2022)

CRISP-DM stands for **C**ross **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining (CRISP-DM). It is a powerful framework for solving any business problem (Hotz, 2022). Figure No. 5.1 shows the 6 stages of this framework which are explained below:

1. Business Understanding

Business Understanding focuses on the initial study of the problem. None of the problems could be solved without having proper knowledge around them. This also is a stage where functional or domain knowledge can be gathered that would be useful in solving the problem. This stage has the below set of actions:

- A. **Understanding Business Objectives:** It is a combination of understanding what the business is all about and the problem being faced.

- B. **Assessing the situation:** Gaining knowledge around any problem. Finding the answers to relevant questions like What is the problem? Who is impacted? When did the problem arise? etc.
- C. **Identifying Problem Statements:** Once the required knowledge is gained, the actual problem can be broken into multiple and relevant subproblems. These problems should be independent.
- D. **Identifying solutions:** After problem statements are found, a crisp solution needs to be identified for them.

2. Data Understanding

This stage is all about understanding data that the business has provided for solving the problem and has the below set of actions:

- A. **Collection of data:** The right solution to a problem can only be identified if the data collected has all the required information. It should contain all the relevant metrics that could help in solving the problem.
- B. **Data description:** Data collected should be understood properly for its significance. Functional understanding of variables is important.
- C. **Data Exploration:** This step is all about finding insights from the data collected. Charts, Graphs, etc. are widely used tools for this.
- D. **Data Quality Inspection:** This step verifies if the quality of data is good enough for business analytics techniques. The scope should be understood for cleaning, imputation, and other data preparation techniques.

3. Data Preparation

This stage also known as Data Munging is the most important part of this framework. It is the most time taking activity in the whole data mining process as it forms the foundation for Modeling. It has five tasks:

- A. **Feature Extraction:** Identify which set of data is important and will be used for data mining activities. The features which do not seem to contribute much to solving the problem can be excluded.
- B. **Data Cleaning:** Imputing missing values, finding outliers are some of the activities done in this step.

C. **Deriving new features:** Creating the new features from existing features. Binning, Ratios, and Encoding are some of the activities involved in this step.

D. **Data Formatting:** Not every algorithm accepts all kinds of data. Some of them are restricted to using only numeric values. So, categorical variables are converted to numeric variables in such cases.

4. Modeling

Modeling is the task of running algorithms on the data to get patterns and insights. The steps involved in the stage are:

A. **Identifying the right modeling techniques:** There are multiple algorithms developed. But not all can be applied to any data. Data Understanding and Problem statements are important factors in deciding the appropriate technique.

B. **Data Split:** Data can be split into the test, train and validation datasets for training, testing and understanding the accuracy of models.

C. **Model building:** It is all about running the right algorithm on the prepared data.

D. **Accuracy Assessment:** Test data is used for evaluating accuracy. The more the model results match with testing data, the higher the accuracy.

5. Evaluation

As there can be multiple models or techniques applied to the same data, the next task is to identify which is working for the problem the most. This stage is all about achieving this. Accuracy, AUC, ROC, Lift, Support and Confidence are some of the metrics that are used for evaluation. Based on the results, the selected model can be used for deployment in the customer environment.

6. Deployment

All the above activities are useless if the model is not deployed in the customer environment. Keeping the model as API is one of the techniques of deployment. One important point to note is that even after the model is deployed there is ample scope for improvisation based on fresh data. Evaluation, Improvisation and Maintenance go hand in hand.

Chapter 6: Business Understanding

RLP Software Pvt Ltd. headquartered in San Mateo, California is a **B2B** company developing business process automation solutions. The organization was founded in 2006 and since then it has been a popular name in the **Revenue Lifecycle** catering products industry. It offers the Commercial Operations Suite, which enables businesses to optimize quotes and digital commerce, manage contracts and documents, as well as automate revenue management. The company caters to the energy, financial services, healthcare, media, retail, and other sectors.

Well-known products used by **11000+ customer** organizations are:

1. CPQ (Configure, Price, Quote)
2. CLM (Contract Lifecycle Management)
3. IWA (Intelligent Workflow Approvals), etc.

Strategy Builder: A new product of RLP launched in Q1 2021 which focuses on Customer Incentive Management. This tool aims to give indefinite ease and flexibility to create rules/schemes for a business to incentivize their customers for their orders/purchases. Billing is an important part of the Revenue Lifecycle, and this is where this tool comes into the picture.

Significant Use Cases of tool (but not limited to):

Buy X Get X – Discounted prices on the product if a certain quantity is purchased.

Buy X Get Y – Discounted price on a different product if another product is bought in a certain quantity.

Package Savings – Discounted prices if a set of products are bought together as a package.

Subscription – Discounted price if the product is purchased regularly.

Cross-selling “Strategy Builder”: RLP software is interested in selling its new product to existing customers for improving its existing Revenue Lifecycle. This in turn would increase their revenue and deepen the customer relationship.

Trial Version of Product: To demonstrate the significance and usability of the product, the company provides a trial version of the tool to potential customers for a specific time.

Onboarding Numbers: In the past ~1.5 years, the trial version was allocated to 6560 customers and 2886 have onboarded, the rest of them did not sign the deal.

Expenses on a trial: Whenever a trial is given it requires resources and every resource has a cost associated with it. Below are some of the expense areas:

1. **DevOps:** It is the combination of cultural philosophies, practices, and tools that increases an organization's ability to deliver applications and services at high velocity. It involves software developers and operations both.
2. **User Training:** It helps the user in efficiently operating the system. It involves both documentation as well as physical sessions.
3. **Third-party Licencing:** The product makes use of third-party software which requires a license to be used. License is mapped to each user and each one has a cost associated with it.
4. **Cloud Computing and Storage:** Cloud computing is the on-demand availability of computer system resources, especially data storage and computing power, without direct active management by the user. Large clouds often have functions distributed over multiple locations, each location being a data centre.
5. **Technical Support:** For the whole duration of the trial, there is some time dedicated to each customer. Each hour spent with the customer is a cost.

The above-mentioned information was provided by the Sales Team of RLP and was crucial in understanding the real problem. Not to mention, the right questions were asked during multiple discussions to gain this precise knowledge and move forward in the right direction.

After gaining a Business understanding, the next phase is to investigate the data provided by the business. The data needs to be understood for the significance of variables, find insights from historic data and identify if any cleaning is required.

Chapter 7: Data Understanding

The data was scattered into multiple entities in the SQL database. While most of the sales data were available as read-only to internal employees but there were a couple of entities that had restricted access as they contained confidential information regarding the Licenses, Revenue and In Progress deals. So, this study is done on data that does not violate the confidentiality clauses. Figure No. 7.1 shows a glimpse of the Entities and their relationships that were within the scope of this study.

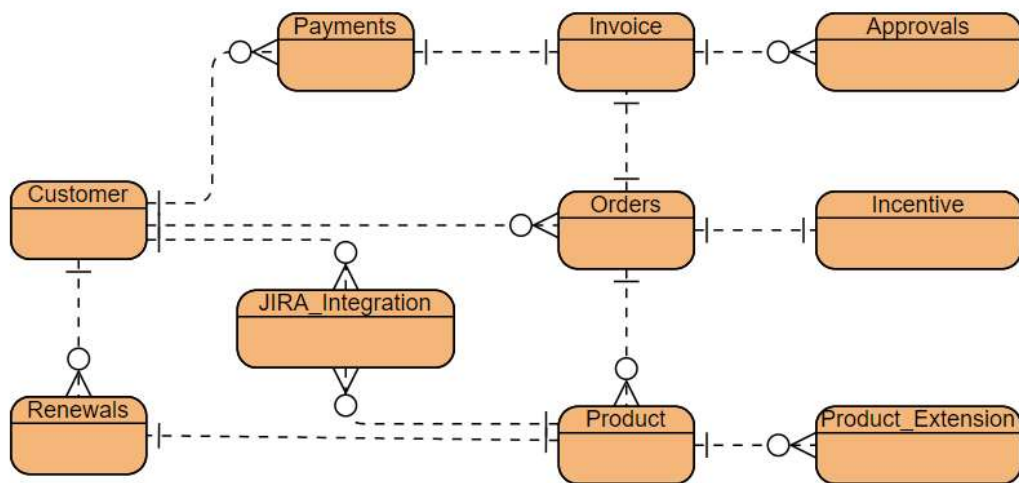


Figure No. 7.1 ER Diagram of Sales System in RLP

Trial Data Extraction: Sales team provided the list of customers who were granted the trial versions of Strategy Builder. The data had two columns 'Customer Id' and 'Purchased' indicating whether the customer bought the product or not after trial. The rest of the data was extracted from the Sales database based on the id of the customer. The entity names in the actual database were too technical, so the labels are changed to make them more understandable and meaningful.

The data extracted had 6560 observations of the customers on which trials were done. There are 12 variables out of which 10 are independent, 1 is dependent and 1 is a customer id column which would not be considered in the study further.

Significance of the variables in the data:

1. **Domain:** The functional area of the customer company.
2. **Region:** The region in which the company is headquartered. This data contains continent names.
3. **No_of_Products_already_purchased:** This is the count of all the products that the customer has already purchased from RLP.
4. **Using_Similar_Product:** Whether the company is using a similar product from any other company. This information is based on the proposal feedback.
5. **CPQ_Customer:** Whether the customer purchased the product CPQ from RLP.
6. **CLM_Customer:** Whether the customer purchased the product CLM from RLP.
7. **IWA_Customer:** Whether the customer purchased the product IWA from RLP.
8. **No_of_times_of_Renewal_:** No of times the customer has renewed the product in past.
9. **No_Of_Years_Of_Association:** No. of years the customer is associated with RLP.
10. **No_Of_Bugs_Reported:** No of Bugs reported from the customer while using the existing products.
11. **Purchased:** Whether the customer has purchased the new product or not.

Data Exploration

1. Count of customers that purchased the new product.

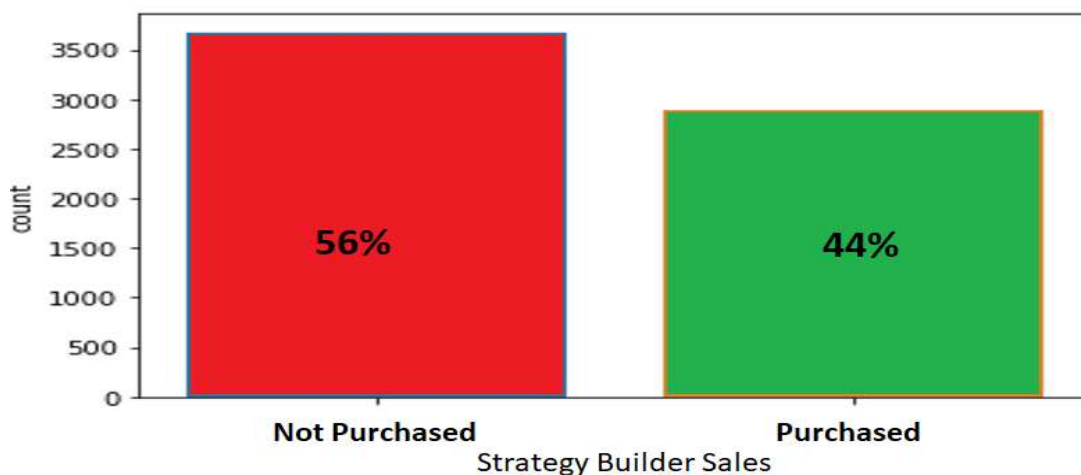


Figure No. 7.3 Sales data for Strategy Builder

Figure No. 7.3 shows the count of customers that purchased the product vs those who did not. Out of 6560 customers on whom the trial was done, 3674 customers did not purchase the product. The conversion rate of trials is just 44%.

2. % Of Customers that purchased the new products by domain.

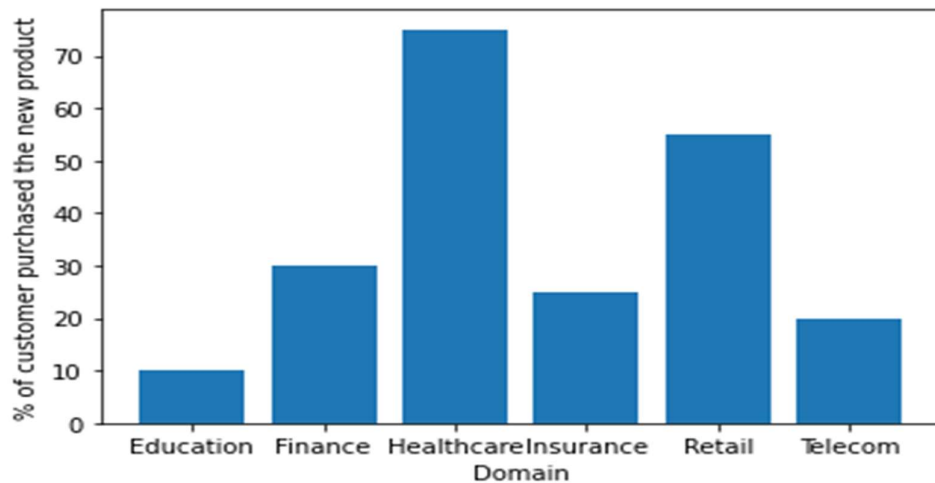


Figure No. 7.4 Sales data by Domain

Figure No. 7.4 shows that existing customers from Healthcare and Retail are purchasing the product most while Education and Telecom customers have shown the least response.

3. % Of Customers that purchased the new products by region.

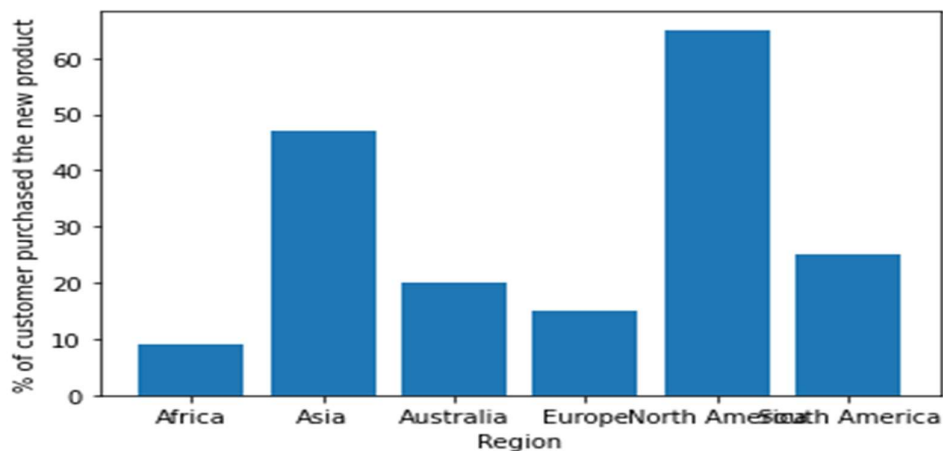


Figure No. 7.5 Sales data by Region

Figure No. 7.5 shows that existing customers from Asia and North America are purchasing this new product more while the sales in the African and European regions are least.

4. % Of customers that purchased the product based on No. of products already purchased.

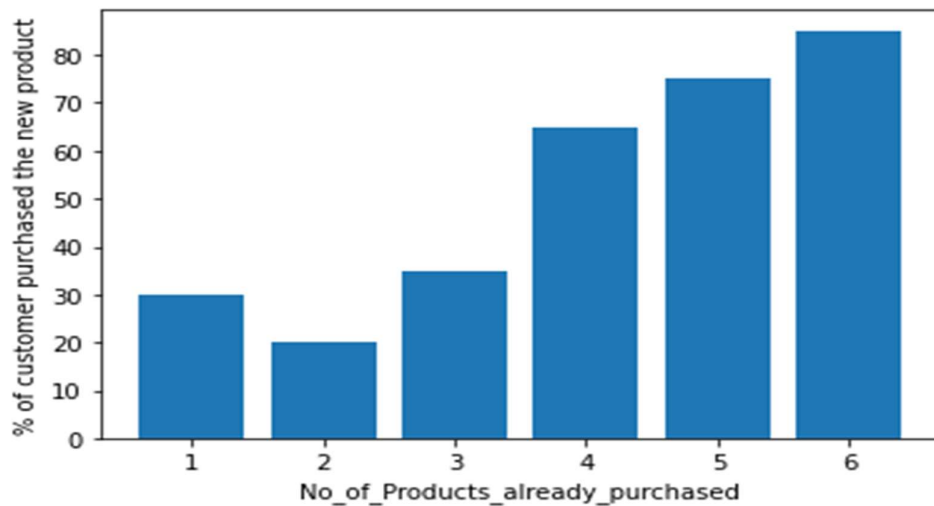


Figure No. 7.6 Sales data by No_of_Products_already_purchased

Figure No. 7.6 shows that the more the customer uses the existing products, the more are they purchasing the new product.

5. % Of customers that purchased the product based on No. of products already purchased.

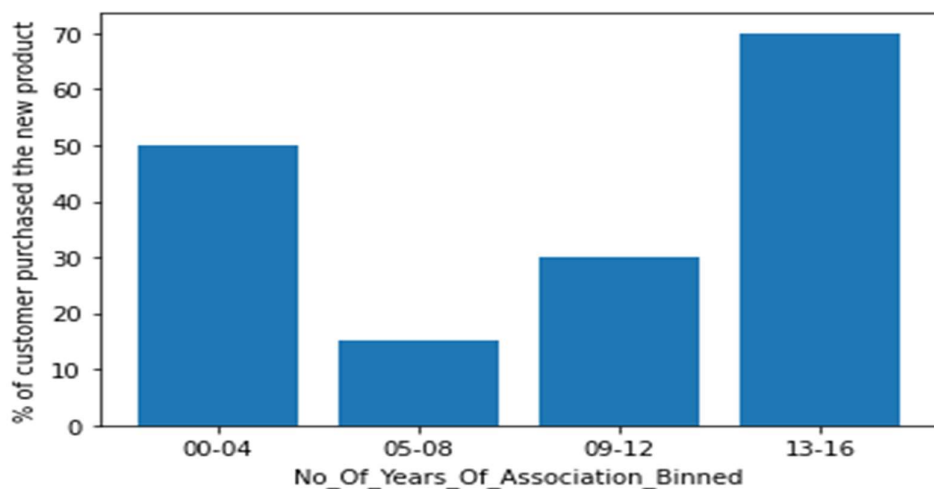


Figure No. 7.7 Sales data by No_of_Years_Of_Association_Binned

Figure No. 7.7 shows that either the customer who is recently associated or has been a very old customer is purchasing the new product more.

6. % Of customers that purchased the product based on No of bugs reported so far from the existing products.

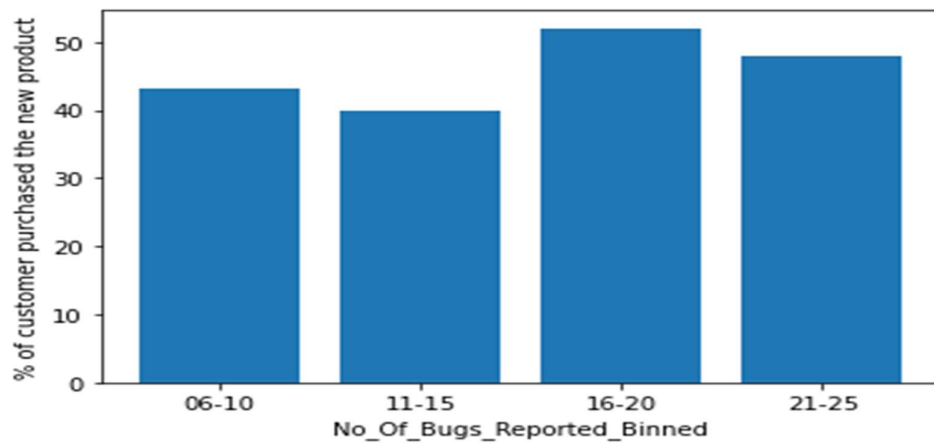


Figure No. 7.8 Sales data by No_Of_Bugs_Reported_Binned

Figure No. 7.8 shows no useful insight from this metric, almost all the ranges have similar purchase trends.

7. Heatmap as shown in Figure No. 7.9 indicates that there is a high correlation between the variable 'Purchased' and 'No of Products Already Purchased'.

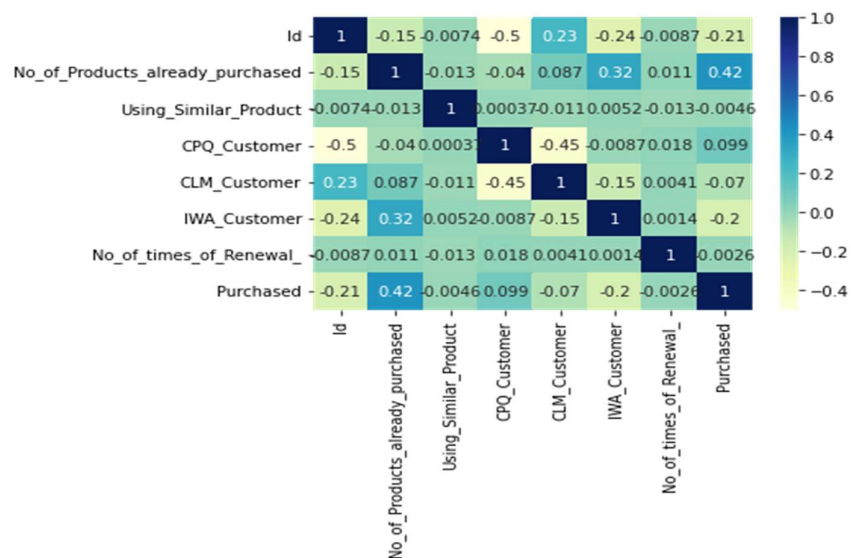


Figure No. 7.9 Heatmap of Correlations

Chapter 8: Data Preparation

Data Preparation is an extremely important stage because the model and its accuracy depend on this a lot. This involves cleaning data, deriving new features and formatting data. The model usually understands numeric data; hence the categorical variables must be converted to numeric ones.

Nominal Variable: is a categorical variable that has two or more categories, but there is no order followed by them. This must be converted to numeric data before this can be fed to the model. In the above dataset, there are 2 nominal variables.

One-Hot Encoding: It is the process of creating dummy variables. This technique is used for categorical variables where order does not matter. One-Hot encoding technique is used when the features are nominal (do not have any order). In one hot encoding, for every categorical feature, a new variable is created.

Encoding 'Region' variable:

The raw data contains 6 unique values for 'Region' as shown in Figure No. 8.1.

```
[13] data['Region'].unique()

array(['Africa', 'Asia', 'Australia', 'Europe', 'North America',
       'South America'], dtype=object)
```

Figure No. 8.1 Unique values of Nominal Variable - Region

After encoding, the new Region variables got created as shown in Figure No. 8.2.

Region_Africa	Region_Asia	Region_Australia	Region_Europe	Region_North America	Region_South America
1.0	0.0	0.0	0.0	0.0	0.0
1.0	0.0	0.0	0.0	0.0	0.0
1.0	0.0	0.0	0.0	0.0	0.0
1.0	0.0	0.0	0.0	0.0	0.0
1.0	0.0	0.0	0.0	0.0	0.0

Figure No. 8.2 Encoded Region data

Encoding 'Domain' variable:

The raw data contains 6 unique values for 'Domain' as shown in Figure No. 8.3

```
[12] data['Domain'].unique()

array(['Education', 'Finance', 'Healthcare', 'Insurance', 'Retail',
      'Telecom'], dtype=object)
```

Figure No. 8.3 Unique values of Nominal Variable - Domain

After encoding, the new Domain variables got created as shown in Figure No. 8.4

Domain_Education	Domain_Finance	Domain_Healthcare	Domain_Insurance	Domain_Retail	Domain_Telecom
1.0	0.0	0.0	0.0	0.0	0.0
1.0	0.0	0.0	0.0	0.0	0.0
1.0	0.0	0.0	0.0	0.0	0.0
1.0	0.0	0.0	0.0	0.0	0.0
1.0	0.0	0.0	0.0	0.0	0.0

Figure No. 8.4 Encoded Domain data

Binning: It is a technique for reducing the cardinality of continuous and discrete data. Binning groups related values together in bins to reduce the number of distinct values.

Binning 'No_Of_Years_Of_Association':

No_Of_Years_Of_Association had to be binned as it had a lot of discrete values. Figure No 8.5 shows the values of 'No_Of_Years_Of_Association' after binning.

```
[100] data['No_Of_Years_Of_Association_Binned'].unique()

array(['00-04', '13-16', '09-12', '05-08'], dtype=object)
```

Figure No. 8.5 Unique values of Variable - No_Of_Years_Of_Association_Binned

Binning 'No_Of_Bugs_Reported':

No_Of_Bugs_Reported also had to be binned as it had a lot of discrete values. Figure No 8.6 shows the values of 'No_Of_Bugs_Reported' after binning.

```
[102] data['No_Of_Bugs_Reported_Binned'].unique()

array(['06-10', '21-25', '16-20', '11-15'], dtype=object)
```

Figure No. 8.6 Unique values of Variable - No_Of_Bugs_Reported_Binned

The above-binned variables were used for EDA which was shown in Data Understanding Chapter.

Ordinal Variables: It is a categorical, statistical data type where the variables have natural, ordered categories and the distances between the categories are not known. This variable needs to be converted to numeric before it can be fed to Model. The above-binned variables must be converted to numeric.

Converting 'No_Of_Years_Of_Association_Binned' to numeric:

Figure No. 8.7 shows the original and encoded values for the variable.

```
[105] # Convert No_Of_Years_Of_Association_Binned to Numerical Values
      data['No_Of_Years_Of_Association_Binned'].replace(["00-04", "05-08", "09-12", "13-16"],
                                                         [1, 2, 3, 4], inplace=True)
      data.head()
```

Figure No. 8.7 Encoded No_Of_Years_Of_Association data

Converting 'No_Of_Bugs_Reported_Binned' to numeric:

Figure No. 8.8 shows the original and encoded values for the variable.

```
# Convert No_Of_Bugs_Reported_Binned to Numerical Values
data['No_Of_Bugs_Reported_Binned'].replace(["06-10", "11-15", "16-20", "21-25"],
                                             [1, 2, 3, 4], inplace=True)
data.head()
```

Figure No. 8.8 Encoded No_Of_Bugs_Reported_Binned data

Correlation Heat Map after Data Preparation:

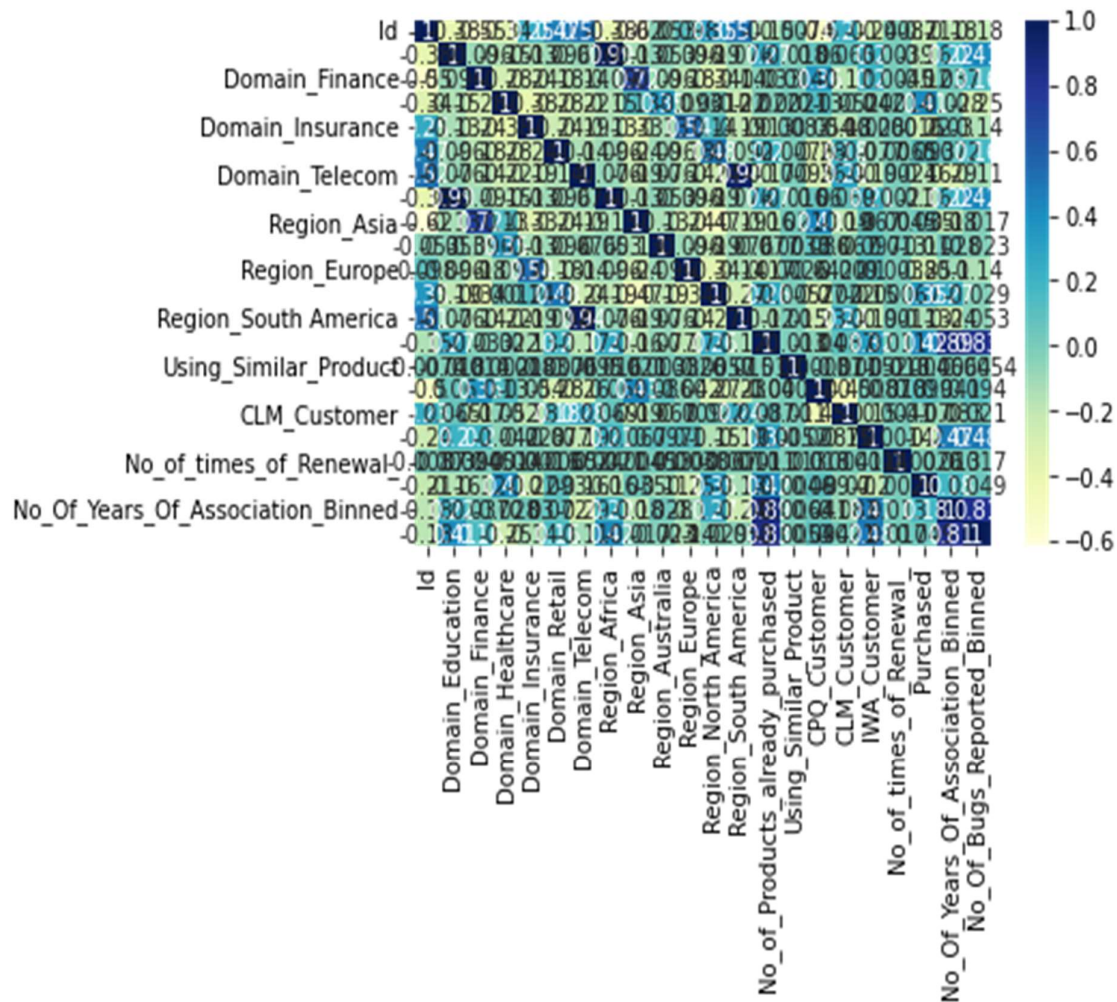


Figure No. 8.9 Correlation Heat Map after Data Preparation

Correlation heatmap from raw data is already shown in the data understanding chapter where the “No_of_Product_already_purchased” variable came out to be strongly correlated with the sale of the product. Since, after applying data preparation techniques, multiple new variables are introduced. It would be interesting to see if there are any strong correlations discovered from them. Figure No. 8.9 shows that in addition to the “No_of_Product_already_purchased” variable, Region_Australia, No_of_Years_of_Association_Binned, and Domain_Healthcare also show a high correlation with Purchase Variable.

After the data cleaning and preparation techniques are applied, the modified dataset is now ready for the Modelling phase.

Chapter 9: Modeling

The data collected from the business was divided into training data for fitting the model and testing data for evaluation of the model.

Training and Test data in this study was divided in the ratio of **75:25**. Below Modelling techniques were applied to the data.

1. Binary Classification for Identifying the most probable customer for the new product:

It is a technique for identifying rules for the classification of data. The rules are formed from the patterns learned from the training dataset. There are multiple models which can be applied to this problem. For this study, 5 different models have been executed.

A. Decision Tree

Figure No. 9.1 shows the graphical representation of the decision tree that came as a result.

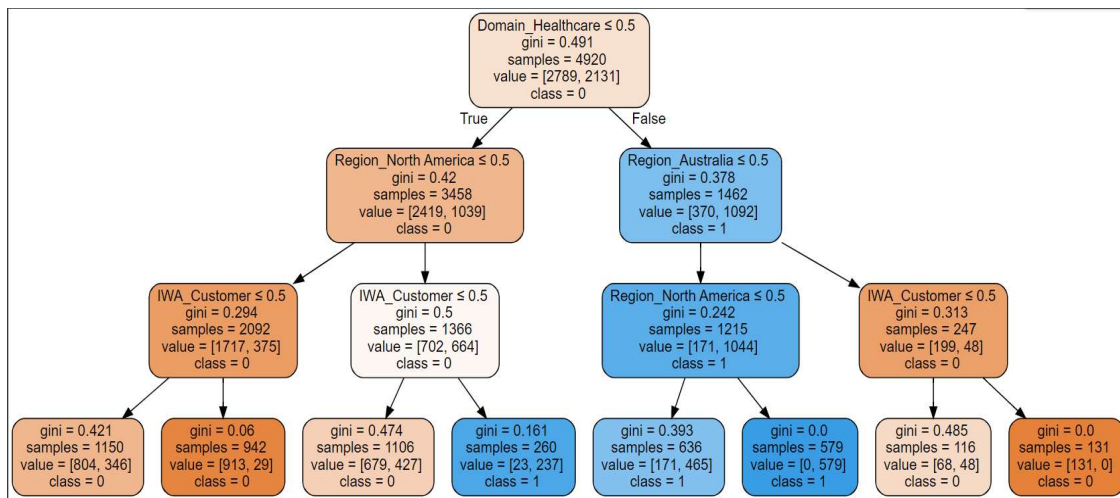


Figure No. 9.1 Decision Tree Outcome

B. Logistic Regression

Figure No. 9.2 shows the coefficient values that came as an output from Logistic Regression Model.

	Variable	Coeff
11	Region_South America	3.902359
2	Domain_Healthcare	3.540338
10	Region_North America	2.826785
1	Domain_Finance	2.012421
0	Domain_Education	1.221863
14	No_of_times_of_Renewal_	-0.015568
12	Using_Similar_Product	-0.033094
13	IWA_Customer	-0.957716
7	Region_Asia	-1.078810
9	Region_Europe	-1.107822
4	Domain_Retail	-1.250271
6	Region_Africa	-1.383339
3	Domain_Insurance	-1.531157
8	Region_Australia	-3.159712
5	Domain_Telecom	-3.993732

Figure No. 9.2 Logistic Regression Coefficients

Based on the coefficients received, the Region and Domain are very strong variables impacting the sales. If the customer is in South America or the Healthcare domain, the sale is likely to happen as these are favorable factors, whereas if the customer is in the Australian region or the Telecom domain, the probability of sale is less.

C. Bernoulli Naïve bayes

The final dataset that was used for Modeling is having all the input attributes as Boolean values, hence Bernoulli Naïve Bayes algorithm was used.

D. Random Forest Classifier

This modeling technique creates multiple decision trees as shown in Figure No. 9.3 and the decision tree which gives the most accurate result is used for prediction. This is a form of the Ensemble model.

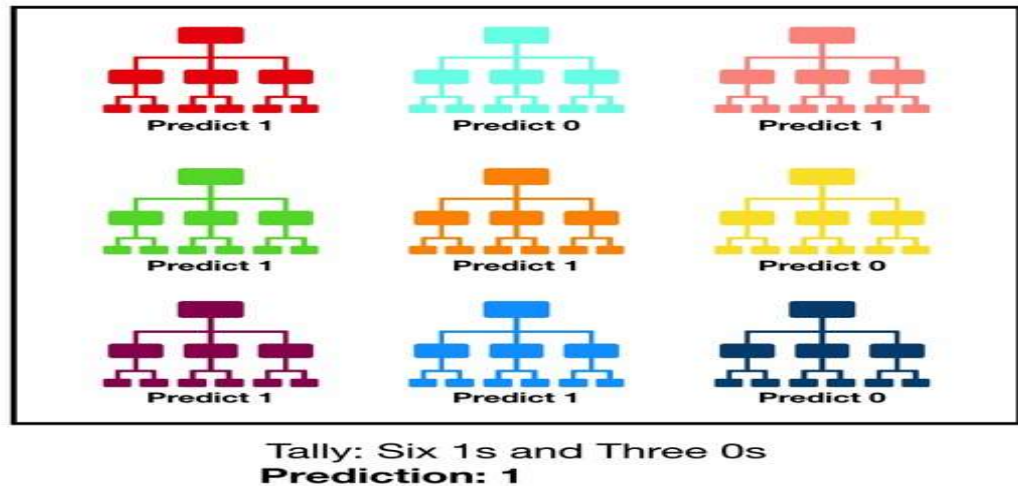


Figure 9.3 Combination of multiple Decision Tress in Random Forest (Yiu, 2019)

E. XGBoost

This is also an Ensemble model and works by creating multiple decision trees as shown in Figure No. 9.4. Here the importance is given to the weight of independent variables. Each variable is assigned a weight and based on the result it is fed to the second decision tree and so on.

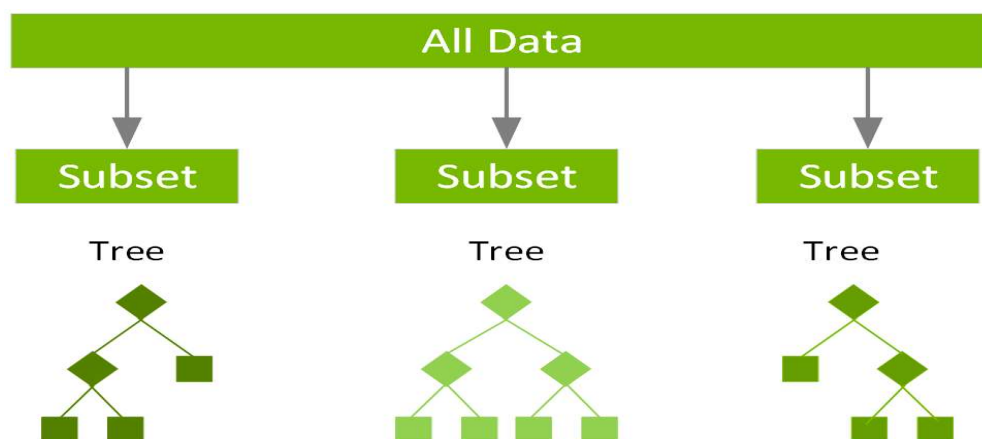


Figure No. 9.4 Subset of data in XGBoost (Nvidia, 2022)

The above-mentioned algorithms were applied to the data. The evaluation of these is done in the next chapter where the best performing model would be selected for identifying the most probable customers.

2. Market Basket Analysis

It is a technique to identify the purchase pattern of products. In other words, it analyses the affinity between the products, so that it can be discovered which product is frequently sold and more importantly the combination of products that get sold together. Support, Confidence and Lift as shown in Figure No. 9.5, are the metrics that help us measure the affinity.

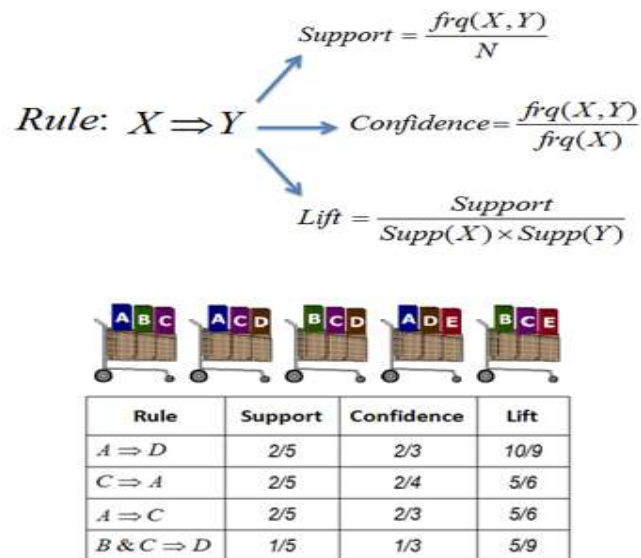


Figure No. 9.5 Metrics of Market Basket Analysis (Li, 2017)

The data in this study contains the purchasing history of the customers. The above-mentioned algorithm was applied to discover meaningful insights into purchasing patterns. Figure No. 9.6 shows the outcome of the analysis.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
4	(CLM_Customer, CPQ_Customer)	(Purchased)	0.213262	0.439939	0.117988	0.553252	1.257566	0.024165	1.253640
5	(Purchased)	(CLM_Customer, CPQ_Customer)	0.439939	0.213262	0.117988	0.268191	1.257566	0.024165	1.075059
2	(CLM_Customer, CPQ_Customer)	(IWA_Customer)	0.213262	0.334146	0.083537	0.391708	1.172266	0.012276	1.094629
3	(IWA_Customer)	(CLM_Customer, CPQ_Customer)	0.334146	0.213262	0.083537	0.250000	1.172266	0.012276	1.048984
0	(Purchased)	(CPQ_Customer)	0.439939	0.572713	0.276220	0.627859	1.096288	0.024261	1.148184
1	(CPQ_Customer)	(Purchased)	0.572713	0.439939	0.276220	0.482300	1.096288	0.024261	1.081825

Figure No. 9.6 Market Basket Analysis Output

Observations:

1. CPQ and Strategy Builder has the highest support, so they are highly popular products amongst the others.
2. Based on the lift and confidence value, the customers who have purchased CLM and CPQ both are more likely to purchase the new product.

3. Clustering of Customers

Clustering is the task of dividing the data into n number of clusters. The clusters formed are based on similar traits. The algorithm starts by randomly putting the centroid in the dataset as per the number configured. Then the iteration of comparison starts, and more similar elements are grouped.

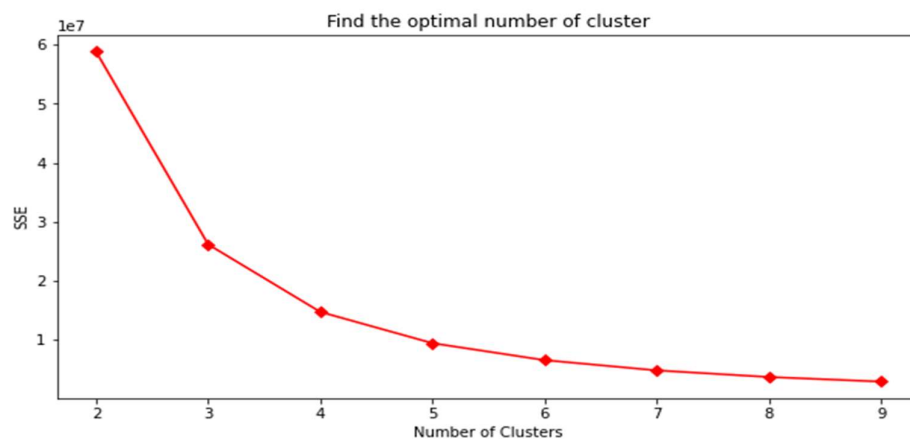


Figure No. 9.7 Optimal no. of Clusters

K-means clustering algorithm – This is a technique in which n number of records are divided into k number of clusters based on the nearest mean values.

Selecting the right value of k using the elbow method: Figure No. 9.7 shows that 3 is the optimal number of clusters that should be formed

After applying the algorithms to the dataset, the next step is to evaluate them. The next chapter focuses on evaluating these models based on accuracy scores.

Chapter 10: Model Evaluation

Models will be evaluated based on the accuracy achieved on the test data. Below are the metrics that came as an outcome of the respective modelling technique:

Decision Tree

Accuracy Score					
0.7792682926829269					
Precision/Recall Metrics					
	precision	recall	f1-score	support	
0	0.73	0.92	0.82	885	
1	0.87	0.61	0.72	755	
accuracy			0.78	1640	
macro avg	0.80	0.77	0.77	1640	
weighted avg	0.80	0.78	0.77	1640	
AUC					
0.7667826542447713					

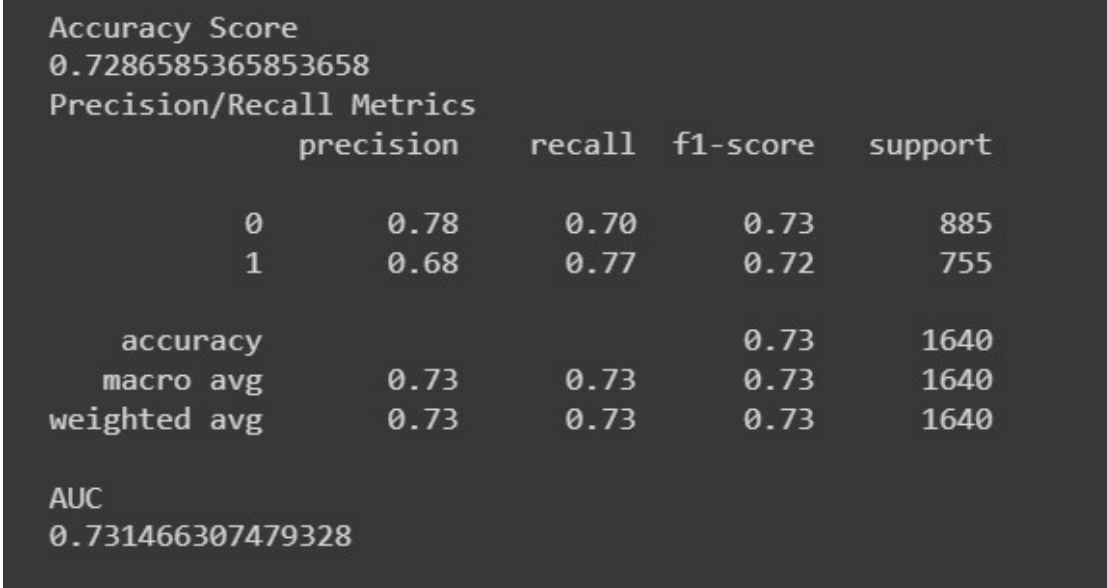
Figure No. 10.1 Decision Tree Model Evaluation Metrics

Logistic Regression

Accuracy Score					
0.7060975609756097					
Precision/Recall Metrics					
	precision	recall	f1-score	support	
0	0.72	0.74	0.73	885	
1	0.68	0.67	0.68	755	
accuracy			0.71	1640	
macro avg	0.70	0.70	0.70	1640	
weighted avg	0.71	0.71	0.71	1640	
AUC					
0.7034609196692484					

Figure No. 10.2 Logistic Regression Model Evaluation Metrics

Bernoulli Naïve Bayes



```
Accuracy Score
0.7286585365853658
Precision/Recall Metrics
      precision    recall  f1-score   support

     0       0.78      0.70      0.73       885
     1       0.68      0.77      0.72       755

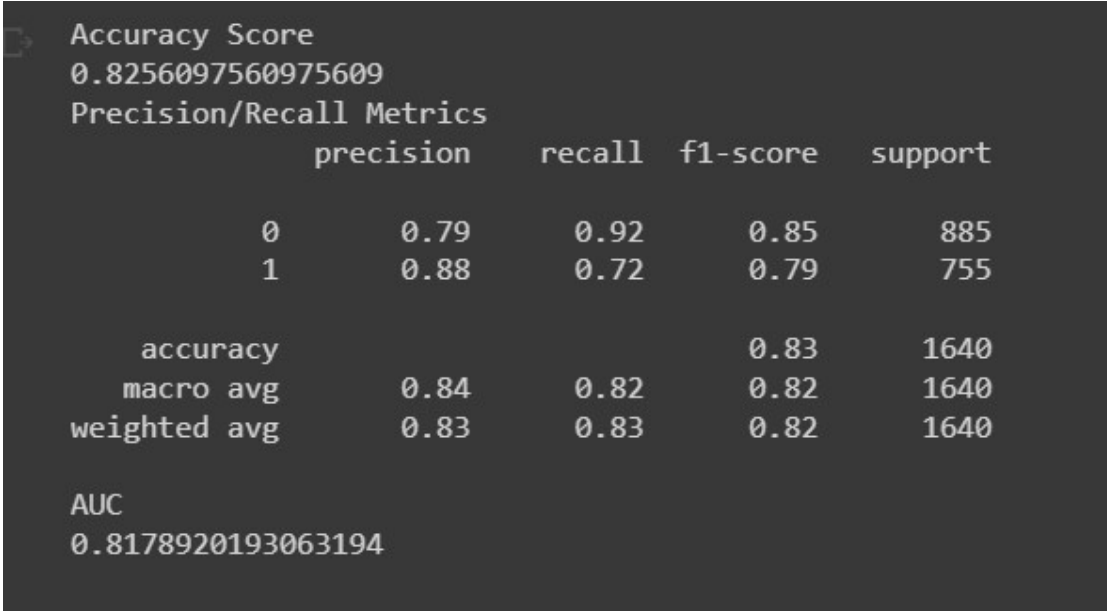
 accuracy          0.73      0.73      0.73      1640
  macro avg       0.73      0.73      0.73      1640
weighted avg       0.73      0.73      0.73      1640

AUC
0.731466307479328
```

The image shows a terminal window with a dark background and light gray text. It displays the evaluation metrics for a Bernoulli Naïve Bayes model. The metrics include Accuracy Score, Precision/Recall Metrics (a confusion matrix), and AUC. The confusion matrix shows that for class 0, the model correctly classified 885 instances with a precision of 0.78 and recall of 0.70. For class 1, it correctly classified 755 instances with a precision of 0.68 and recall of 0.77. The overall accuracy is 0.73, and the AUC is 0.731466307479328.

Figure No. 10.3 Bernoulli Naïve Bayes Model Evaluation Metrics

Random Forest



```
Accuracy Score
0.8256097560975609
Precision/Recall Metrics
      precision    recall  f1-score   support

     0       0.79      0.92      0.85       885
     1       0.88      0.72      0.79       755

 accuracy          0.83      0.83      0.83      1640
  macro avg       0.84      0.82      0.82      1640
weighted avg       0.83      0.83      0.82      1640

AUC
0.8178920193063194
```

The image shows a terminal window with a dark background and light gray text. It displays the evaluation metrics for a Random Forest model. The metrics include Accuracy Score, Precision/Recall Metrics (a confusion matrix), and AUC. The confusion matrix shows that for class 0, the model correctly classified 885 instances with a precision of 0.79 and recall of 0.92. For class 1, it correctly classified 755 instances with a precision of 0.88 and recall of 0.72. The overall accuracy is 0.83, and the AUC is 0.8178920193063194.

Figure No. 10.4 Random Forest Model Evaluation Metrics

XGBoost

Accuracy Score				
0.825				
Precision/Recall Metrics				
	precision	recall	f1-score	support
0	0.80	0.91	0.85	885
1	0.87	0.73	0.79	755
accuracy			0.82	1640
macro avg	0.83	0.82	0.82	1640
weighted avg	0.83	0.82	0.82	1640
AUC				
0.8179107269802073				

Figure No. 10.5 XGBoost Model Evaluation Metrics

Model Accuracy results:

<u>Model</u>	<u>Accuracy</u>
Decision Tree	77.92%
Logistic Regression	70.60%
Bernoulli naive bayes	72.86%
Random Forest	82.56%
XGBoost	82.5%

Table No. 10.1 Model Accuracy Results

Based on the Accuracy Scores as shown in Table No. 10.1, Random Forest and XGBoost Models are giving the highest and almost similar accuracy. Considering the decimal values of accuracy, the **Random Forest** Model's accuracy is high and should be considered for deployment.

Chapter 11: Deployment

Based on the results we have achieved from the model; Random Forest Model can be integrated with the existing system of sales. Whenever the sample of customers is picked for giving the trials, the result from this model must be checked and the trial should be extended accordingly.

Model as API: It is a business logic deployed on the server with a contract of input and output. It is platform-independent and can be called from any tech stack provided the contract is abode by. The selected model will be exposed as an API as shown in Figure No. 11.1, and it will return the response as true or false. True means the customer is a probable customer for the new product and false means vice versa.

Sample Python API using Flask:

```
from flask import Flask, jsonify, request, send_file
app = Flask ()
@app.route('/is-probable-customer', methods = ['GET'])
def isProbableCustomer():
    customer_data = request.args.get('customer_data')
    response = randomForestModel.predict(customer_data)
    return response
```

Endpoint for Caller: `http://127.0.0.1:8000/is-probable-customer`

Verb: Get

Body: Customer data object

Response: Boolean (True or False)

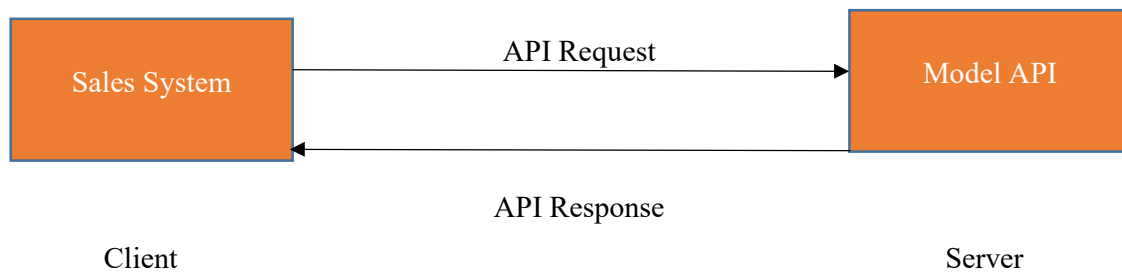


Figure No. 11.1 Model Deployment Design

Chapter 12: Analysis and Results

Predictive Analytics

After Applying multiple machine learning algorithms to the dataset, Random Forest came out to be the best model with an accuracy of 82.56%. This is a decent accuracy as it does not come under overfitting or poor accuracy. The objective of finding the most probable customer is achieved with the model finalized.

Market Basket Analysis

Another objective of the study was to find the association between the products. This information is now available for the customer to improvise its sales strategy. Based on the lift and confidence value of 1.25 and 0.62 respectively, the customers who have purchased CLM and CPQ both are more likely to purchase the new product. Hence, this combination of products can be considered for cross-selling Strategy Builder.

Clustering

Similar customers need similar solutions. With this intention, the third objective was achieved by finding the set of customers who have shown the least response in sales of Strategy Builder and have similar characteristics identified by k-means clustering. Out of the 3 clusters formed, the first one had the greatest and the second one had the lowest number of customers who purchased the new product.

Total Number of customers in 1st Cluster: 2192

Total customers who purchased the product in 1st Cluster: 1208

Total Number of customers in 2nd Cluster: 2181

Total customers who purchased the product in the 2nd Cluster: 672

All this work is incomplete if the right recommendations are not given to the business. The business should improvise its strategy based on these insights. With the help of the study done so far, recommendations are given in the next chapter.

Chapter 13: Conclusions and Recommendations for future work

Based on the above study, below are the conclusions and recommendations for RLP software:

1. Measures should be taken to attract customers from Education and Insurance domains as these are the lowest performing domains. It is strongly recommended to add some domain-specific features to the tool to attract these customers.
2. Africa and Europe are the lowest performing regions for this product, marketing and promotion should be strategized effectively to acquire the market share in these regions. Competition should be analyzed and measures to be taken accordingly.
3. Random Forest Model API to be integrated with RLP's Sales system and before giving a trial version to the customer, the probability must be checked.
4. For the customers who are less likely to buy the new product as per the model's response, they should be given a demo of the product (in premise) instead of giving a separate trial version, this is an attempt to introduce our product at a negligible cost. (After all, they are customers too!!)
5. The new product should be considered for cross-selling with CPQ and CLM products. This recommendation is for existing as well as future customers.
6. Customers in cluster 2 should be targeted for onboarding by applying similar strategies in terms of marketing, promotions, discounts, etc. as they have shown similar characteristics while clustering.

Future Work

The cluster of customers has already been created, also the lowest and highest performing clusters are identified. In the next round of study, these clusters should be explored for patterns so that strategies can be improved for these similar customers.

Objectives for subsequent research:

1. Identifying ways to move customers from the lowest performing cluster to the higher performing one.
2. Improve model based on fresh/future data.

Bibliography

- Bellinger, C. (2012). One-class versus binary classification: Which and when? *Proceedings - 2012 11th International Conference on Machine Learning and Applications, ICMLA 2012*, 2, 102–106. <https://doi.org/10.1109/ICMLA.2012.212>
- Bole, U., & Papa, G. (2011). Who are the Likeliest Customers: Direct Mail Optimization with Data Mining. *Contemporary Engineering Sciences*, 4(6), 259–268.
- Bu, A. (2018). *Education Ecosystem*. <https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1>
- Contributor, T. (2017). *What is purchase intent? - Definition from WhatIs.com*. <https://www.techtarget.com/whatis/definition/purchase-intent>
- Emmer, M. (2018). *Inc.com*. <https://www.inc.com/marc-emmer/95-percent-of-new-products-fail-here-are-6-steps-to-make-sure-yours-dont.html>
- Fathy, A. (2015). Identifying the Factors Affecting Customer Purchase Intention. *Global Journal of Management and Business Research: Administration and Management*, 15(2), 1–6.
- Gupta, S., & Mamtara, R. (2014). A Survey on Association Rule Mining in Market Basket Analysis. *International Journal of Information and Computation Technology*, 4(4), 409–414. <http://www.irphouse.com/ijict.htm>
- Hotz, N. (2022). *What is CRISP DM? - Data Science Process Alliance*. <https://www.datascience-pm.com/crisp-dm-2/>
- Jaiswal, S. (2011). *Javatpoint*. <https://www.javatpoint.com/machine-learning-models>
- Kaur, M., & Kang, S. (2016). Market Basket Analysis: Identify the Changing Trends of Market Data Using Association Rule Mining. *Procedia Computer Science*, 85(Cms), 78–85. <https://doi.org/10.1016/j.procs.2016.05.180>
- Kodinariya, T. M., & Makwana, P. R. (2013). Review on determining of cluster in K-means. *International Journal of Advance Research in Computer Science and Management Studies*, 1(6), 90–95. <https://www.researchgate.net/publication/313554124>
- Li, S. (2017). *A Gentle Introduction on Market Basket Analysis — Association Rules | by Susan Li | Towards Data Science*. <https://towardsdatascience.com/a-gentle-introduction-on-market-basket-analysis-association-rules-fa4b986a40ce>
- Mai, W. (2021). A data mining system for potential customers based on one-class support vector machine. *Journal of Physics: Conference Series*, 2031(1).

<https://doi.org/10.1088/1742-6596/2031/1/012066>

- Mirabi, V. (2015). A Study of Factors Affecting on Customers Purchase Intention Case Study : the Agencies of Bono Brand Tile in Tehran. *Journal of Multidisciplinary Engineering Science and Technology (JMEST)*, 2(1), 267–273.
- Nvidia. (2022). *What is XGBoost? | Data Science | NVIDIA Glossary*.
<https://www.nvidia.com/en-us/glossary/data-science/xgboost/>
- Omran, M. G. H. (2007). An overview of clustering methods. *Intelligent Data Analysis*, 11(6), 583–605. <https://doi.org/10.3233/ida-2007-11602>
- Sun, K. (2017). What Can Product Trial Offer?: Th[1] K. Sun, M. Zuo, and D. Kong, “What Can Product Trial Offer?: The Influence of Product Trial on Chinese Consumers’ Attitude towards IT Product,” <https://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/IJABIM>. *Https://Services.Igi-Global.Com/Resolvedoi/Resolve.aspx?Doi=10.4018/IJABIM.2017010102*, 8(1), 24–37.
<https://doi.org/10.4018/IJABIM.2017010102>
- Team, B. C. (2021). *MBA Skool*. <https://www.mbaskool.com/business-concepts/marketing-and-strategy-terms/10976-purchase-intention.html>
- Woo, J., & Xu, Y. (2011). Market Basket Analysis Algorithm with Map/Reduce of Cloud Computing. *The 2011 International Conference on Parallel ...*, April 2012.
<http://www.lidi.info.unlp.edu.ar/WorldComp2011-Mirror/PDP4494.pdf>
- Yiu, T. (2019). *Understanding Random Forest. How the Algorithm Works and Why it Is...* | by Tony Yiu | *Towards Data Science*. <https://towardsdatascience.com/understanding-random-forest-58381e0602d2>

Appendix

Plagiarism Report¹

Minimizing Losses on Trials of Strategy Builder Tool			
ORIGINALITY REPORT			
8%	7%	4%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS
PRIMARY SOURCES			
1	ukcatalogue.oup.com Internet Source	2%	
2	docplayer.net Internet Source	1%	
3	www.m-hikari.com Internet Source	1%	
4	Weijian Mai, Fengjie Wu, Fang Li, Wenjun Luo, Xiaoting Mai. "A Data Mining System for Potential Customers Based on One-Class Support Vector Machine", Journal of Physics: Conference Series, 2021 Publication	1%	
5	link.springer.com Internet Source	1%	
6	Submitted to Solihull College, West Midlands Student Paper	1%	
7	docs.oracle.com Internet Source	<1%	

¹ Turnitin report to be attached from the University.

8	Submitted to Rufus King International Baccalaureate High School Student Paper	<1 %
9	www.expatriates.com Internet Source	<1 %
10	Submitted to American University of the Middle East Student Paper	<1 %
11	Submitted to RDI Distance Learning Student Paper	<1 %
12	Submitted to University of Greenwich Student Paper	<1 %
13	Submitted to Georgia Military College Student Paper	<1 %
14	academyit.net Internet Source	<1 %
<div> <div>Exclude quotes</div> <div>On</div> <div>Exclude matches</div> <div>< 10 words</div> <div>Exclude bibliography</div> <div>On</div> </div>		

Publications in a Journal/Conference Presented/White Paper²

Paper Submitted:

Tushar Nigam and Rashmi Agarwal, “Predicting the Likeliest Customers; Minimizing Losses on Product Trials using Business Analytics”, EAI ICISML 2022 : 323048, 30th September 2022

² URL of the white paper/Paper published in a Journal/Paper presented in a Conference/Certificates to be provided.

Predicting the Likeliest Customers; Minimizing Losses on Product Trials using Business Analytics

Tushar Nigam¹[0000-0002-0095-6147] and Rashmi Agarwal²[0000-0003-1778-7519]

¹ RACE, REVA University, Bangalore, 560064, India
tushar.ba07@race.reva.edu.in

² RACE, REVA University, Bangalore, 560064, India
rashmi.agarwal@reva.edu.in

Abstract. A product trial is a great way of marketing. It offers a first-hand experience that a customer can use to evaluate the product and decide whether to buy it. However, there are expenses involved in every trial. If the buyer decides not to purchase the product after the trial, this expense gets wasted. The lower the conversion rate, the higher the loss that the company must face. To reduce these losses, a strategy should be developed on which studies should be conducted. One such business case is the subject of this paper. A software development company has launched a new product and given its current customers access to the product's trial version. A loss of \$8.5M resulted from 56% of trial customers not buying the product. The business is therefore looking for strategies to reduce this loss. This paper aims to present the work which is done to solve this business problem. The goal is accomplished by employing analytical methods to determine the most likely clients. The data provided by the company is used to build binary classification models. Another way to find the likeliest customer is based on the purchase pattern of the customers. Affinity analysis is done on the data to identify the set of products with which the new product is frequently sold to target the customers who have purchased those products for the sale of the new product. Among the binary classification models that were built, the Random Forest model outperformed other models with an accuracy of 83%. This enables the business to take calculated decisions while extending the trial to the customers. Alternatively, Market Basket Analysis, with an accuracy of 88%, discovered a set of two products, the existing buyers of which are more likely to buy the newly launched product. This information not only helped find the right customers but also paved the path for cross-selling of the new product.

Keywords: Likeliest Customers, Binary Classification, Market Basket Analysis, Cross Selling

1 Introduction

Product trial is a unique form of advertising as it provides a direct experience to the customer. It is much better than documentation, PPT, and theories as it's the working model of the solution the product provides. With the advancement of technology, new products are emerging constantly and rapidly. But not all of them become popular and profitable. Studies mention that 95% of new products fail [1]. Not enough exposure to the customers could be one of the major reasons for this as the customers are not familiar with the capability and potential of the product [2]. Therefore, a trial becomes very important for customers to evaluate the product and make purchase decisions.

Although the trial is an effective way of promotion, it is also a very expensive way of marketing. There are a lot of resources involved in the trials and there is a cost associated with every resource. These resources could be user training, operations, technical support, cloud computation and storage, licenses, etc. The company loses revenue if the conversion rate of customers after trials is low.

Therefore, there should be a strategy based on which the trials are given to the customer so that losses are minimal. Purchase intent is one of the main factors that need to be looked at before giving any trial to the customer [3]. It is the probability that a customer will buy a product. Once the purchase intent is evaluated, the decision of giving a trial or not can be taken. Purchase Intent can be evaluated with the help of predictive analytics.

A software development company, RLP Software Private Limited launched a new product 'Strategy Builder' and has extended the trial version of the tool to its existing customers. The conversion ratio after the trial is just 44%. Based on the data provided, the business has lost \$8,450,200 on unsuccessful trials. The objective of the study is

to minimize these losses. Predictive Analytics would be used to identify the customers with higher purchase intentions. It would also be interesting to know if the new product is purchased more by the customers who own a specific set of existing products. Affinity analysis is a powerful tool to get this information.

2 Literature Review

Existing published works were reviewed to identify the right set of solutions for this problem. The Literature review done for this case is divided into 4 parts. The first part focuses on the importance of trials after the launch of a product, the second one explains the concept of Purchase Intentions, and the last 2 parts are dedicated to various machine learning algorithms to improve the trial conversion.

2.1 Importance of Product Trial

Whenever a new product is launched, there is always an expectation of great sales by the company. Most of the time the results are not as expected and from there, the retrospect and improvisation start. For customers who have not purchased the product, the company can sell products through the improvement of product satisfaction. However, the more product satisfaction is improved, the more services need to be paid for, which increases the cost of the company. The company wants to sell its products successfully with minimal cost. To maximize the benefits for the company, a data mining system must be developed to help companies focus on unpurchased customers with the strongest purchase intentions [4].

2.2 Purchase Intention

It is the willingness of the customer to buy a certain product. It's a dependent variable based on several independent factors and is a measure of a potential customer's attitude towards purchasing a product. It is an important metric in designing marketing strategies [5]. Price had been an important variable in influencing purchase intentions but other variables such as knowledge, experience, and quality are important in the process of customer's purchase decisions too [6]. For that matter, the duration of association of customers with the company also plays an important role. Another factor is the perceived value, which is considered directly proportional to the purchase intent [7]. With these vast possibilities, the first step is to identify the factors affecting customers' purchase intention.

2.3 Binary Classification

It is a powerful technique that could be used to predict purchase intention. From the business point of view, two prediction models are needed [8]. The prediction model which uses some or all behavioral data could only be applied to the company's existing client base, for whom all data is available - the so-called cross sale [9]. On the other hand, it makes sense to build a separate prediction model using only the socio-demographic data if the company were interested in acquiring customers that have not yet established a business relationship with the company. Since the modeling objective became clear, the next step was to identify the type of classification.

Binary classifiers are a widely used option, however, there is an alternate method available called one-class classification, which works only with a single class of data [10]. This technique is used when class distribution is imbalanced i.e., one of the classes is much higher in data than the other. In such cases, the former may not perform well. Considering the data in this study the classes are divided in a ratio of 44:56, which makes it a balanced dataset. Therefore, it was decided to go with the binary classifiers only for this study.

2.4 Market Basket Analysis

Once the purchase intent is evaluated, several other important patterns can be discovered from data collected from the business. Affinity analysis is one such technique for achieving that. For identifying the affinity between the products there have been multiple pieces of research done. Most of them involved using Market Basket Analysis. A prior study [11] shows that this way is useful in finding out interesting patterns from a large amount of data, predicting future association rules as well as the right methodology to find out outliers. Among various methods

for affinity analysis, the apriori algorithm is found to be better for association rule mining. There are various complexities involved in the apriori algorithm like the exhaustive scan of the database multiple times and as the input becomes larger, the computation time increases significantly [12]. Still, this is a popular choice among data scientists. The data collected from the business is moderate in size, hence these limitations will have negligible effects on our study.

Some studies tried a completely new way of improvising the MBA. One such study used the Map/Reduce of Cloud Computing [13]. The algorithm has been executed on EC2 small instances of AWS with nodes 2, 5, 10, 15, and 20. The execution times of the experiments show that the proposed algorithm gets better performance while running on a large number of nodes to a certain point. However, from a certain point, Map/Reduce does not guarantee to increase the performance even though we add more nodes because there is a bottleneck for distributing, aggregating, and reducing the data set among nodes against the computing powers of additional nodes. Hence, the scope of the study is limited to the basic algorithm for finding associations.

3 Data Description

The data was scattered into multiple entities in the SQL database. While most of the sales data were available as read-only to internal employees but there were a couple of entities that had restricted access as they contained Confidential information. So, this study is done on data that does not violate the confidentiality clauses.

The sales team provided the list of customers who were granted trial versions of the new product. The data had two columns ‘Customer Id’ and ‘Purchased’ indicating whether the customer bought the product or not after the trial. The rest of the data was extracted from the Sales database based on the id of the customer.

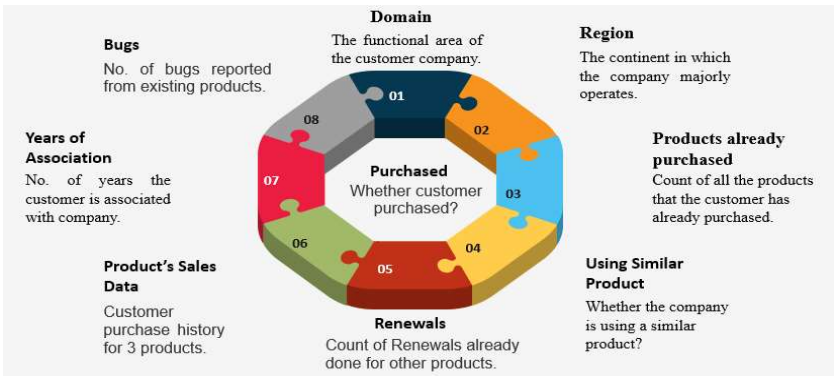


Fig. 1. Description of Variables

Fig. 1 shows the variables and their significance that were involved in the study. The data extracted had 6560 observations of the customers on which trials were done. There are 12 variables out of which 10 are independent, 1 is dependent and 1 is a customer id column which would not be considered in the study further. There were no missing values found in the dataset. The dependent Variable ‘Purchased’ had a balanced no of observations for each class with ratio of 44:56.

4 Data Preparation

Data Preparation is an extremely important stage because the model and its accuracy depend on this a lot. This involves cleaning data, deriving new features, and formatting data. The model usually understands numeric data; hence the categorical variables must be converted to numeric ones.

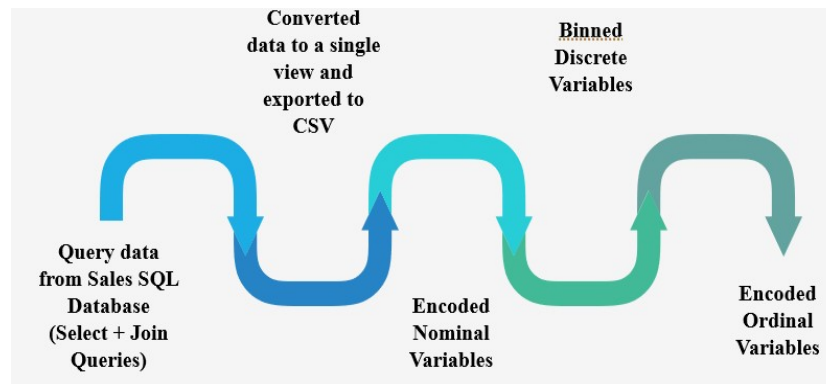


Fig. 2. Data Preparation Flow

Fig. 2 shows different techniques that were applied to preparing data for modeling. Encoding and Binning are the major ones among them followed by replacing the binned columns with numeric ones.

One-Hot Encoding: It is the process of creating dummy variables for categorical variables where order does not matter. For every categorical feature, a new variable is created.

Binning: It is a technique for reducing the cardinality of continuous and discrete data. Binning groups related values together in bins to reduce the number of distinct values.

With the application of these preprocessing techniques, the data was ready for the modeling stage.

5 Model Building

For achieving the objective of the study, there were 2 techniques identified. One was to build a predictive model, and another was to find the affinity of the new product with the set of existing products based on purchase history. The solution for each objective is explained below:

5.1 Identify the most probable customer for the new product.

Method: Binary Classification.

Dependent Variable: Purchased.

Independent Variables: 10 (All columns - Purchased)

Train and Test data ratio: 75:25

Metrics: Accuracy

Techniques: Decision Trees,
Logistic Regression,
Naïve Bayes,
Random Forest,
XGBoost.

5.2 Identify the set of products with which the new product gets sold frequently

Method: Market Basket Analysis.

Variables: Individual existing product purchase data for each customer.

Technique: Apriori algorithm for finding Association Rules

Metrics: Lift, Confidence and Support.

Train and Test data ratio: 75:25

6 Evaluation

After application of different machine learning algorithm, the next step is to evaluate them so that the most appropriate technique can be used for solving the business problem. Below are the evaluation results for the techniques applied:

6.1 Predictive Analytics

In the last step, multiple binary classification models were built, but not all can be used by the customer, hence, the most accurate one should be picked for predicting the likelihood. Table 1 shows the comparison of accuracy.

<u>Modeling Technique</u>	<u>Accuracy</u>
Decision Trees	77.9%
Logistic Regression	70.6%
Naïve Bayes	72.8%
Random Forest	83%
XGBoost	82%

Table 1. Accuracy Comparison

6.2 Market Basket Analysis

The combination of two products showed a high association with the new product in the training data. This result gave 88.28% accuracy when checked on the testing data.

7 Results

After evaluating the models and algorithms based on their respective metrics, like Lift, Confidence, and Support for Market Basket Analysis and Accuracy score for Binary Classification, below results and observations were found:

7.1 Predictive Analytics

After Applying multiple machine learning algorithms to the dataset, Random Forest came out to be the best model with the highest accuracy. The objective of finding the most probable customer is achieved with the model finalized.

7.2 Market Basket Analysis

Another objective of the study was to find the association between the products. This information is now available for the customer to improvise its sales strategy. Based on the lift and confidence values found from association rules, the customers who have purchased the products 'MLC' and 'QPC' both are more likely to purchase the new product. Therefore, the customers who have already purchased these products should be targeted for the trials of the new product. Additionally, this combination of products can be considered for cross selling the new product.

8 Conclusion

This paper gives an overview of data mining and modeling techniques applied to minimize the losses in product trials. Predictive analytics with the help of the Random Forest model gave an accuracy of 83% which is quite decent. There are no traces of underfitting and overfitting in the model. Therefore, the trials should be given to customers only based on the response of the model. For the customers, whose likelihood is less, should be given a demo/video of the product for marketing instead of a separate trial version, as they are the company's customers too and can be potential customers for the new product. Affinity analysis with the help of the Apriori algorithm discovered the fact that the new product gets sold more often by the customers who already own the existing 'MLC' and 'QPC' products of the company. Hence, the new product should be considered for cross-selling with these products.

9 Scope for future work

The solution for the problem that the business is facing could also be achieved by clustering, which could help in understanding similarities between the customers who are not purchasing the new product. This technique would divide the customers into a certain number of clusters based on similar traits. The idea is to take advantage of these common traits to identify the measures to onboard these customers.

References

- [1] M. Emmer, "Inc.com," 2018. <https://www.inc.com/marc-emmer/95-percent-of-new-products-fail-here-are-6-steps-to-make-sure-yours-dont.html> (accessed Aug. 10, 2022).
- [2] K. Sun, "What Can Product Trial Offer?: Th[1] K. Sun, M. Zuo, and D. Kong, "What Can Product Trial Offer?: The Influence of Product Trial on Chinese Consumers' Attitude towards IT Product," <https://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/IJABIM>," <https://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/IJABIM.2017010102>, vol. 8, no. 1, pp. 24–37, Jan. 2017, doi: 10.4018/IJABIM.2017010102.
- [3] T. Contributor, "What is purchase intent? - Definition from WhatIs.com," 2017. <https://www.techtarget.com/whatis/definition/purchase-intent> (accessed Aug. 10, 2022).
- [4] W. Mai, "A data mining system for potential customers based on one-class support vector machine," *J. Phys. Conf. Ser.*, vol. 2031, no. 1, 2021, doi: 10.1088/1742-6596/2031/1/012066.
- [5] B. C. Team, "MBA Skool," 2021. <https://www.mbaskool.com/business-concepts/marketing-and-strategy-terms/10976-purchase-intention.html> (accessed Aug. 10, 2022).
- [6] V. Mirabi, "A Study of Factors Affecting on Customers Purchase Intention Case Study : the Agencies of Bono Brand Tile in Tehran," *J. Multidiscip. Eng. Sci. Technol.*, vol. 2, no. 1, pp. 267–273, 2015.
- [7] A. Fathy, "Identifying the Factors Affecting Customer Purchase Intention," *Glob. J. Manag. Bus. Res. Adm. Manag.*, vol. 15, no. 2, pp. 1–6, 2015.
- [8] S. Jaiswal, "Javatpoint," 2011. <https://www.javatpoint.com/machine-learning-models> (accessed Aug. 11, 2022).
- [9] U. Bole and G. Papa, "Who are the Likeliest Customers: Direct Mail Optimization with Data Mining," *Contemp. Eng. Sci.*, vol. 4, no. 6, pp. 259–268, 2011.
- [10] C. Bellinger, S. Sharma, and N. Japkowicz, "One-class versus binary classification: Which and when?," in *Proceedings - 2012 11th International Conference on Machine Learning and Applications, ICMLA 2012*, 2012, vol. 2, pp. 102–106. doi: 10.1109/ICMLA.2012.212.
- [11] M. Kaur and S. Kang, "Market Basket Analysis: Identify the Changing Trends of Market Data Using Association Rule Mining," *Procedia Comput. Sci.*, vol. 85, no. Cms, pp. 78–85, 2016, doi: 10.1016/j.procs.2016.05.180.

- [12] S. Gupta and R. Mamtara, "A Survey on Association Rule Mining in Market Basket Analysis," *Int. J. Inf. Comput. Technol.*, vol. 4, no. 4, pp. 409–414, 2014, [Online]. Available: <http://www.irphouse.com/ijict.htm>
- [13] J. Woo and Y. Xu, "Market Basket Analysis Algorithm with Map/Reduce of Cloud Computing," *2011 Int. Conf. Parallel ...*, no. April 2012, 2011, [Online]. Available: <http://www.lidi.info.unlp.edu.ar/WorldComp2011-Mirror/PDP4494.pdf>