`

**Mod 7: Capstone Project - Research Paper**

Scott Foran Hall ID#326360

CSU Global

MIS 581 – Capstone: Business Intelligence and Data Analytics

Professor – Dr. Justin Bateh

**Contents**

**Abstract 3**

**Introduction 4**

**Project Overview 5**

**Objectives 5**

**Research Questions and Hypotheses 8**

**Literature Review 9**

**Research Design 13**

**Ethical Considerations 13**

**Limitations 14**

**Analytics, Results and Findings 15**

**Conclusion 25**

**Recommendations 26**

**Reference List 28**

**Abstract**

Businesses everywhere have a need to know their customers better, providing a compelling product or service is clearly a top priority as the business would not exist without it. From there, it is particularly critical to understand when and where your customers need you. Missing an opportunity to connect with a current or prospective customer can mean the difference between gaining new or existing business together or losing potential clients. Many businesses have used trial and error or “gut feel” to determine when and where to staff, introducing the potential for waste and error in the process. But now, with cutting edge technology and state of the art analytics tools, the gap can be closed by enabling predictive models that can provide better planning and insights for new tactics to meet these customers when and where they are. This research project aims to explain why understanding these patterns is important, what methods may work best to achieve it, when and how it should be implemented, and some additional ideas around enhancements, risks, and limitations. Ultimately, the goal is to take dependable, reliable business insights, and develop operational tactics that will ultimately improve the customer experience and improve efficiencies for the business.

**Predicting Consumer Demand in an Insurance Environment**

**Introduction**

Many organizations and industries can benefit from the power that advancements in data analytics can now provide. Areas of the business where leaders used to take their best guess or rely on intuition can now tap into actionable data that is already at their fingertips and convert it into actionable insights. And while one can spend time and effort researching a multitude of new things, the best outcomes from research efforts are often a result of combining analysis with experience and drive.

For those reasons, and more, this research is focused on the property/casualty insurance industry, specifically focused on AmFam Group’s consumer insurance business. Having experience within the industry, including areas where day-to-day pain points provides an opportunity to bring new insights. Many problems exist within this business that can be solved or improved upon such as the focus on increasing broad coverage with multiple products for long-term customers.

Other examples of increased challenges to the business include coverage risks which are becoming more extreme due to changes in weather and climate, social unrest, price inflation hitting parts, supply chain reliability and labor costs. And perhaps most important, understanding the customers’ needs, and meeting them where they are to solve for those needs is definitely a top priority.

**Overview**

With great options, the top priorities of the business should determine an area of focus. For that reason, the case study that being for this exercise is a use case related to business planning needs and associated projections for customer demand traffic, and the associated impacts to sales, servicing, and claims team’s ability to meet that demand. The process described throughout required gaining an understanding of overall volume, and the things that influence customers like holidays, pay days, day of the week, etc. The idea is that the more enabled the operations teams are with predictive models that can help them plan for when customers will be looking for their assistance, the better they will be enabled to meet various client’s needs. So, this business problem has great potential both for learning, and applying insights for an existing business.

**Objectives of the Research**

The objectives of the research will therefore be to l*everage reliable, accessible data and use it to create a reliable demand model to plan around*, with the following supporting objective components in mind:

* Improve understanding of reliable customer behavior, associated patterns
* Plan for sales and service staffing needs
* Provide insights around seasonality, holiday demand changes
* Understand intraweek, monthly trends and variability

To add little more information regarding goals, this stated objective increases the likelihood that the aim of gaining better understanding of customer behavior, and the associated aims of understanding and creating plans around those trends can be achieved. Additionally, the insurance industry is becoming more data rich every day, offering understanding more about prospects, clients, traffic, and claims. And while there is no shortage of business problems including how to maximize potential prospects, ensure that client needs are handled well, and navigating new challenges related to supply shortages, inflation, and expenses, many of those questions are already being explored, while the aim of better understanding customer demand traffic remains largely un-tapped.

In terms of academics, the aim is always to learn and expand our abilities, there can be powerful effects from bringing new techniques, processes and analytics to familiar industries and businesses. Business needs for leveraging predictive analytics in the insurance industry spread far and wide, but purposeful use of such analytics remains somewhat scattered. And as a bonus, when new methods of analysis can be tested, discovered, even proven out, sometimes research can be put into practice in significant ways.

**Organizational Benefits**

In terms of specifics not only around long-and medium-term planning, but also for day-to-day decision making, there have been a few key innovations that have shifted management from making pure judgment calls to leveraging data-based insights that help guide those decisions. Examples of this include customer demand forecasting, where different types of forecasts are utilizing customer traffic patterns to determine whether demand will be up or down and allow management to re-allocate resources accordingly. Another is around customer quote follow up, where specific queries focused on what stage in the quoting process a customer may have walked away. Depending how far they got, systems are set up to reach out with an email or text message and quickly connect them with a live agent if there is further interest. This is where better demand planning can provide new benefits.

As it related to functional value, these are just a small sample of some organization benefits from this research:

* Understanding demand related data points needed to drive key results
* Building measurable metrics that are understandable and within the span of control
* Creation of demand related reporting and dashboards that are accessible and actionable
* Initiate review results regularly and responding to changes and trends

As mentioned, the organization stands to benefit from the proposed research in multiple ways, and more will be explored later on with final recommendations.

**Research Questions and Hypotheses**

**Problem Statement**

Continuing with the background and facts mentioned, the specific problem being researched is well focused. As mentioned, operations teams have a need to better understand the customer and prospect demand needs. This will rely on data being collected in a few ways, most of which are very reliable. Beginning with physical locations, agencies do track customer calls and “pop in’s”, and have good, automated capture of every quote being generated. Additionally, all channels involving technology, such as mobile and web traffic, and importantly, contacts to live agents such as chats, emails and phone calls are captured in detail.

With that in mind, here are the resulting problem statement and associated business question(s):

* Problem statement: *“The business has a need to understand consumer demand, therefore understanding reliable patterns related to that demand should be pursued.”*
  + Research Questions:
    - *Does the day of the week impact consumer behavior?*
    - *Do seasonal changes cause volatility?*
    - *Can the resulting information be actionable?*

This problem statement helps define a clear focus, and associated business questions listed above should help determine whether that problem can be solved. And the aim will be to determine whether those questions can lead to clear answers.

**Hypothesis**

Having identified an area of focus, it’s important to package these various theories, questions, and ideas into a concise set of hypotheses for the research. With that in mind, the hypothesis is as follows:

* Hypothesis (H1): *Statistical modeling, leveraging 15-20+ customer behavior attributes will allow creation of a propensity model that provides trends around when and where customers need assistance.*
* Null Hypothesis: *Statistical modeling, leveraging 15-20+ customer behavior attributes will NOT allow creation of a propensity model that provides trends around when and where customers need assistance.*

It is worth pointing out this hypothesis is very specific, as it relates to what is being modeled (customer traffic), why – to confirm theorized patterns within that behavior, and what is driving that traffic.

The outcome is that specific elements previously mentioned will be utilized within models, with a couple of examples of model types under consideration that will be discussed momentarily.

**Literature Review**

In support of the project, associated problem statements and hypotheses in mind, the following pieces of literature will contribute to the research project, providing strong ideas, technical insights and overall support.

**Scholarly Literature Piece #1**

The first piece of literature selected to assist with this research project was chosen to assist with general concepts in taking a data driven approach in operations planning, and it is *Profit Driven Business Analytics: A Practitioner's Guide to Transforming Big Data Into Added Value* by Verbeke/Bassens/Bravo. This text was chosen due to the focus on helping senior leadership in management with critical decisions. Some of the topics include *Value Centric Perspective in Analytics, Business Applications* and more.

A topic that is likely to be reviewed closely is *Analytical Techniques*, which contains a topic on “denormalizing data for analysis”. This topic focuses heavily on the need to get various data elements into a clean, “structured table”, and touches on associated elements of ETL and other interesting analytics techniques. The book also touches on important elements such as selecting the right mix of analysts and data scientists, and finding the right skillsets for your business problems, overall solid contributions to the research project.

**Scholarly Literature Piece #2**

The next article is *The Use of Deseasonalization Techniques in Demand Forecasting* and is focused on understanding time series patterns, seasonality and how they impact customer behavior. A description of deseasonalization was provided in *Deseasonalization of a Time Series*, which states; “Deseasonalization is carried out in a similar manner. However, here we divide the original series, Y, by the seasonal index for corresponding months. For example, the first cell in column Y/S, cell G3, contains the formula =D3/Jan. Jan is a name I defined for the January seasonal index computed in the previous recipe” (Bourg, para 3). Below is an excerpt from the text that goes into detail about time series forecasting and the associated techniques:

* “*In a manufacturing context, demand forecasting can be seen as a proactive process of determining production needs … Within the wide range of available forecasting techniques, the one based on time series (or historical series) are particularly relevant. A time series or historical series {Yi} is a sequence of values Y 1,..,Yi,..,Yn (representing, for instance, demand for a given product) sampled at regular intervals T1,..,Ti,..,Tn for a given time period. Several components can be identified in time series.*
* *The trend component describes the tendency of the demand in the considered time interval. As an example, Figure 1 shows a time series with a strong and easily identifiable growing trend component. The cyclic component is due to typical variations over long-time intervals (generally years), mostly related to business cycles. Figure 2 shows a time series with a strong cyclic component, in which it is possible to highlight a recurrent demand pattern, with a typical cycle length of two years…*”

A picture containing text, line, screenshot, plot

Description automatically generated

A picture containing text, screenshot, line, plot

Description automatically generated

**Scholarly Literature Piece #3**

The next scholarly resource is an eBook called *Marketing Analytics: A Practical Guide to Improving Consumer Insights Using Data Techniques*. This book was chosen for its focus on understanding how to predict consumer behavior with data analytics. From the description, the techniques are designed to help analysts “*leverage predictive techniques to measure and improve marketing performance. By exploring real-world marketing challenges, it provides clear, jargon-free explanations on how to apply different analytical models for each purpose. From targeted list creation and data segmentation, to testing campaign effectiveness, pricing structures and forecasting demand, this book offers a welcome handbook on how statistics, consumer analytics and modelling can be put to optimal use*” (Grigsby, 2018).

Specifically, the areas of segmentation and forecasting demand, along with focus on some particular analytics techniques, such as logistic and multivariate regression modeling will be utilized. The chapters focused on predictive analytics are likely to provide value as well, and there is good content around other techniques, and tools.

**Scholarly Literature Piece #4**

The other formally reviewed piece of scholarly literature that was reviewed is focused on overall operations planning, and will be purposed around how to interpret, utilize and implement the resulting consumer trend data and insights, and apply tactics to improve the customer experience. This literature is a case study called *Design of a Sales and Operations Planning (S&OP) process* (Avila/Lima/Moriera/Pires/Bastos, 2019). From the abstract, this study is focused on companies need to “develop and increase coordination between operational functions to respond rapidly and accurately to customer requests”. There is solid content related to planning tactics such as staffing models and integrating IT/new technology solutions to assist with these tactics. This resource will assist with the finishing touches, helping bring the customer demand model to life by way of new tactics and decisions used to take advantage of those insights.

**Other Research**

Other areas of research may include further examples of organizations that are utilizing consumer demand models to create and implement predictive trends in operations.

Additionally, it will be interesting to understand how technology and automation can also assist with these efforts. In summary, these scholarly resources should provide a lot of added value to all of the intended areas of research previously mentioned and should bring new beneficial insights into business planning methods and techniques.

**Research Design**

**Methodology**

In terms of tools and tactics, several were deployed. First, Alteryx, a tool used within the organization, was utilized to extract and connect data elements together, for a useable data set. And while Alteryx is a unique tool, the methods it uses to tie data elements together often utilize SQL language and functionality, with a visual interface for added benefit.

Additionally, Python is a tool that in this case can take advantage of a structured, pliable dataset (supplied by Alteryx), and perform a variety of different model types, to determine the best possible solutions and fit. Some of the models reviewed were multivariate regression test, which is designed to evaluate how multiple variables influence a particular outcome, and partial least squares, which helps “perfect” the amount of input from each variable by pursuing the best approach. For instance, with this research, does it matter if a holiday falls on a Monday in the Summer, and how do those three different variables (Holiday, Day of Week, Time of Year) interact or combine to affect customer behavior.

Logistic regressions are being utilized to determine whether certain variables have any influence at all on customer traffic. For example, is there any residual impact on behavior three days after a minor holiday. Finally, MS Excel is used as a way to sort, clean and check some of the data and the associated tests mentioned above. To complete some of those checks, the solver tool was utilized as well.

**Ethical Considerations**

From an ethical perspective, the biggest factor within this research is that customer behavior is being tracked and analyzed. That being said, there is no personal privacy concern for individuals, as this data would be utilized and aggregated by day, and by customer channel (emails, chats, phone calls, social media, etc). With that in mind, many companies are aggregating customer behavioral information for better planning and customer habits, and the data for this experiment is no different. Ultimately the planning and related benefits to the overall customer experience should be a good trade off, however, it may make sense to ensure consumers are notified and aware of the usage of this data, and its various purposes.

**Limitations**

While the proposed research, hypothesis, and reliability of the aforementioned predictive approach to customer demand planning is generally sound, there are some limitations to this approach.

First, the variables or “inputs” being utilized are mainly time-series driven. This means that while many reliable, repetitive events can be studied and planned around, new activities that are not set to a calendar can create challenges from a reliability perspective.

Additionally, a model like this will not be able to account for changes in outside situations or environment. For example:

* Economic challenges and recessions
* Unemployment volatility
* Civil unrest

As can be interpreted, these situations are all “disruptors” and could throw off a model that banks on reliable historical patterns.

Another challenge is rooted in the disconnect between why customers may be contacting the organization in the first place. Transactional data capture of customer behavior tied to when they contact the business tells you when they did what they did, but it does not tell you why. Potential downstream solutions for this could include combining transactional contact data with existing or new surveys with contact reason dispositioning, but that would be a later phase for this initiative. For these reasons, and more, it is unlikely that the statistical margin of error will come in better than +- 5-6% with the current iteration of the recommended model.

**Analytics Results and Findings**

The analysis for this program involves several steps, they are:

1. Data elements - collection methods, description
2. Gathering the needed data elements into one location, via ETL
3. Determining model type
4. Setting up and running the model
5. Reviewing the results and identifying key factors
6. Analysis/conclusions

**Data Elements - Collection**

This data is collected in a few ways, some more reliable than others. Beginning with physical locations, agencies do track customer calls and “pop in’s”, and have good, automated capture of every quote or customer contact being generated.

The data categories will fall into the following groups:

* Customer Sales Inquiries
  + Count per day
  + Internet vs retail
  + Other contact channels
* Customer service inquiries
  + Phone Support
    - General inquiries
    - Pay my bill
    - Request ID cards
  + Chat Support

The proposed dataset, which represents a broad scope within the business, should enable the aforementioned research, and ultimately provide the fuel needed to answer the associated business questions, and determine whether the hypothesis is true or false.

**Dataset Description**

The outline below describes the data elements that are currently being captured and expect to be utilized:

* Description of the type of the data elements
  + Interval – Time series, Daily aggregation
    - Many types of date/time variables
      * Month, Week (of the year)
      * Day of week, etc
  + Positive integer (customer contact volume)
    - Counts, aggregated, grouped
    - Work types
      * Email, chat, phone, agency
  + Meticulous capture of factors related to calendar events
    - Seasonality
    - Holidays
      * Business days before and after holidays
    - Other planned events ie. Superbowl, World Cup, etc

As can be interpreted, the data for this work is reliant on time series, calendar related events, which can be as basic as “day of week”, ie Friday, to as intricate as the third day after a minor holiday. This is because there is data that has been tracked going back many years, and those data elements, assuming they are clean, should be able to point to very reliable patterns in behavior which can then be planned around. For example, a qualitative theory could be that current or prospective consumers slow their consumptive behavior around the holidays between Christmas and New Year, and then increase their activity after New Year’s Day, but what does the quantitative data show.

For a sample visual model, an example data dictionary is shown in the chart below.

Figure 3 – Data dictionary, customer contacts

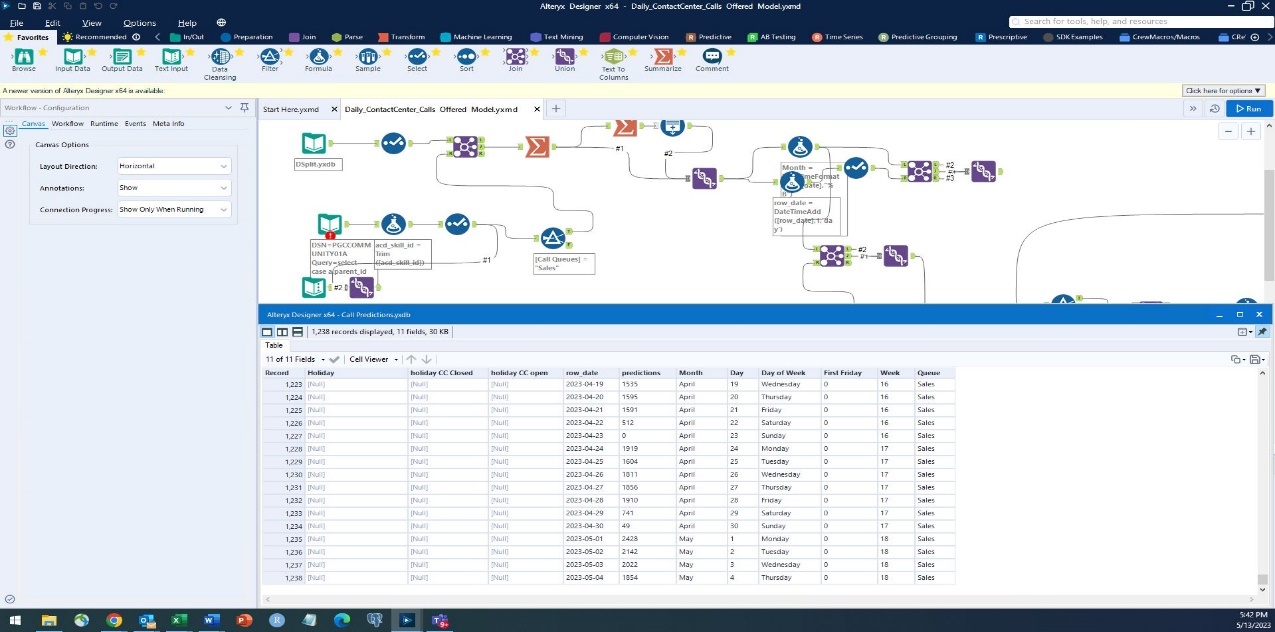


Note: Data dictionary chart supplied by Scott Hall

**Data Extraction (ETL)**

To gather the various data elements, and get them into one clean table view for further analysis, an ETL process was required, utilizing Alteryx. Without going into too many details, Alteryx enabled joining of disparate datasets using join points, in this case a phone number or other unique set of customer identifiers. As shown below, the several data sources being joined together were able to produce a clean dataset, which is shown briefly in figure 4.

Figure 4 – Alteryx workflow data results



Note – Data results provided by Scott Hall

As can be interpreted, the process for getting the dataset is somewhat lengthy, but provides a good base to work within. Regarding data cleansing, the data is generally clean, but it was reviewed to ensure there are no incomplete results, and that data falling out of normal range (+- 20% from the base mean) is accurate and explainable. As an example of some further analysis with this data, Figure 5 below shows a sample of a scatter plot comparing insurance sales with customer contacts.

This can be seen visually below:

Figure 5 – Scatter plot of data sample

Note – Graphic supplied by Scott Hall

As shown, the data points in the scatter plot are relatively well organized, with only a few outliers. While those outlier data points on the right side of the Y axis, may require further analysis, this particular comparison achieved an R^2 value of .377, showing a loose correlation, but not enough to focus on with this research so we will put that to the side for now.

**Determining Model Type(s)**

Regarding models reviewed and utilized, there were two that were the focus of this research.

The first is a linear regression done in Python that is embedded in the Alteryx workflow above, and a small snapshot from some of the code and process from that model which is shown below:

Figure 6 – Sample from Python linear regression

A screenshot of a computer program

Description automatically generated with medium confidence

Note – Screenshot supplied by Scott Hall

As a result of running in some cases many years of data through the model, the resulting outputs provide some significant relationships for some of the variable categories. For example, P values for Day of week, and time of year (Week) ran in the 0.6-0.8 range, which is fairly high. For holidays, and associated before/after days, the P values were lower, running in the 0.04-0.06.

Figure 7 – Prediction output from linear regressionA screenshot of a computer

Description automatically generated

Note – Data from linear regression providing “predictions output”

As shown, the linear regression supplies a forecasted volume, based on past demand, and considering several factors such as holidays, day of week, month, etc. Interestingly, this model did not require a lot of “training”, a phenomenon described as “benign overfitting” (Zhou, Ge, 2023, para 2-4), so overall this model is fairly solid, and has provided a prediction output (shown below) that provides a forecast that is consistently +- ~7% when compared to actuals afterwards. For those reasons, this model could be an option for the business to build around.

The second model reviewed is slightly different and pulls in some different tools as well. For this approach, an adjusted least squares method was used. This approach applies associated coefficients for each of the identified components of the multiple variables that being considered, and after taking the absolute value of the error rate, uses the solver within excel to come up with the lowest “Sum of Squares” as shown below in Figure 8 which provides a snapshot of the model and related variables set up in excel, and shows how the solver was set up to find the solution. Figure 8 – Partial Least Squares ProcessA screenshot of a computer

Description automatically generated

Note - Screenshot of the model and associated variables set up in excel by Scott Hall

As shown, this method was achieved in excel, leveraging the solver, which utilizes powerful computing power to review thousands of different combinations of coefficients, finding the absolute value of positive and negative combinations by “squaring” the result, then adding the results to determine the “best fit” based on finding the lowest “error rate”. This is technically process falls into the multi variate non-linear regression category, but the least squares methods have become a popular tool for utilizing dozens of relevant variables in the same model to find the best solution. A sample of the results from this model are shown below.

Figure 9 – Partial Least Squares model

A screenshot of a computer

Description automatically generated with medium confidence

Note - Output of monthly summary data from regression model by Scott Hall

**Results**

The monthly forecast (highlighted cells) from the least squares forecast model is quite reliable with most months coming within +- 2-3% variance, aside from some outlier months such as November 2022, which was impacted by localized economic stimulus programs. The daily results from the model are similar to the linear regression, running +- ~6-8%, however there was some skewing related to seasonality that would likely improve the accuracy over time following some additional tuning and perhaps a new variable or two.

In short, the process of utilizing historical customer demand data, and applying time series factors such as day of week and holidays does appear to prove the theories and answer the associated questions. More on final analysis and recommendations is summarized below.

**Conclusion**

In summary, the tools and methods mentioned above take advantage of powerful new(er) technology and statistical modeling approaches that help drill into specific business questions with data driven answers, helping organizations come to firm conclusions that can drive key decisions.

As previously explained in detail, two different types of models, linear regression, and partial least squares both provided significant capabilities for predictive modeling consumer demand in the insurance environment, with the least squares model slightly outperforming the linear regression, but both achieving the desired result of providing directional guidance within a reasonable margin of error, allowing the business to plan for expected customer behavior.

And to revisit the Null Hypothesis, the statement is as follows:

* Null Hypothesis: *Statistical modeling, leveraging 15-20+ customer behavior attributes will NOT allow creation of a propensity model that provides trends around when and where customers need assistance.*

Therefore, the null hypothesis is deemed to be false, with a high level of confidence. This tells us that the data utilized in the study, and the associated tests, research and analysis are a sound foundation for this business practice and process adoption.

And with the related business questions aimed at providing better availability of resources when and where clients need them, there are opportunities to enable tactics that will improve planning, and ultimately offer a better customer experience.

**Recommendations**

Regarding next steps, the following recommendations should be considered:

* The ROI step should be taken first, this step should be done to evaluate the hard costs from current inefficiencies, including things such as opportunities to capture and service abandoned customers such as Customers who think the wait time is too long at certain times, as well as Customers who disconnect after waiting in line. Additionally, a determination should be made regarding the costs from lower occupancy during certain days/times. Lastly, a summarized total of the expense and lost revenue opportunities should be compared to determine the overall return on investment. From there, a determination of what a 3-4% improvement in those categories would provide in total expense/revenue benefit, then subtract the cost of two full time analysts to determine total cost benefit. Assuming the ROI is positive, the business should implement a full-scale predictive model for customer traffic.
* The predictive model should utilize organized data capture of customer contacts from all channels with the following attributes: date, day of week, customer ID, channel, IVR/other path. From there, the organization should deploy a multivariate model, using either the linear or least squares regression method using software tool(s) of choice (Python, Excel, Alteryx, etc.). P values and error rates utilizing r^2 values should be completed for model decision. The capture baseline (average daily volume) by channel is an additional step that should be taken, and the capture of co-efficient factors for every measured variable (day of week, time of month, day after major holiday, etc). The model should apply layered factors to baseline volume to enable future daily forecast
* Measurement – the process should be tested and measured after 4-6 weeks, then reviewed for accuracy. Once the model is averaging ~+-5% forecast accuracy on most days (85+%), it can be deployed. Model accuracy should improve slightly over time as more data is ingested
* Operational actions - once reliable, staffing adjustments can be made based on higher/lower forecasts, and using creative staffing techniques such as rotating weekends, flexible schedules, etc. Management can also explore supplementing human resources with tier 1 automation options, improved self-servicing tools, Mobile App improvements and Voice Ai and other augmentation.

In supposition, the implementation of a predictive model providing forecast data of customer demand traffic will provide the organization with great tools for improved planning and overall business optimization, for it’s clients, employees, and it’s bottom line. Thank you for your time and consideration.

**Portfolio Project References**

**Scholarly Articles and Resources**

Wouter Verbeke; Bart Baesens; Cristian Bravo. *Profit Driven Business Analytics : A*

*Practitioner's Guide to Transforming Big Data Into Added Value* Series: Wiley & SAS

Business Series. Hoboken, New Jersey : Wiley. 2017. eBook., Database: eBook

Collection (EBSCOhost)

Genovese, Andrea; Simpson, Mike. Series: Kogan Page Case Study: KTP Tyres : *The Use of*

*Deseasonalization Techniques in Demand Forecasting* Case Study Library. [London?] :

Kogan Page. 2016. eBook., Database: eBook Collection (EBSCOhost)

Grigsby, Mike. (2018). Marketing Analytics : *A Practical Guide to Improving Consumer*

*Insights Using Data Techniques*. Second edition. London : Kogan Page. 2018

Ávila, Paulo; Lima, Daniela; Moreira, Dália; Pires, António; Bastos, João. (2019). *Design of a*

*Sales and Operations Planning (S&OP) process* In Procedia CIRP. Case Study. Database: ScienceDirect

Bourg, David M. (2022). Deseasonalization of a Time Series. Flylib retrieved from

[Deseasonalization of a Time Series | Time Series Analysis (flylib.com)](https://flylib.com/books/en/2.22.1/deseasonalization_of_a_time_series.html)

Ben Ali, M., D’Amours, S., Gaudreault, J., Carle, M-A. (2018). *Configuration and evaluation of*

*an integrated demand management process using a space-filling design and Kriging*

*metamodeling*. Science Direct. Article Vol: 5:45-58

Zhou, Mo; Ge, Rong (2023). *Implicit Regularization Leads to Benign Overfitting for Sparse*

*Linear Regression*. Working Paper, Access URL: <http://arxiv.org/abs/2302.00257>

Bousebata, Meryem; Enjolras, Geoffroy; Girard, Stéphane. *Extreme partial least-squares*. In

Journal of Multivariate Analysis. March 2023 194 Language: English. DOI: 10.1016/j.jmva.2022.105101, Database: ScienceDirect

Siah Hwee Ang (2013). *In:Research Design for Business & Management.* SAGE, 2013

ISBN 144629336X, 9781446293362

Machi, L. MacEvoy, B. T. (2016). *The literature review – Six steps to success.* (3rd ed.). Corwin.

ISBN: 9781506336244

O’Leary, Z. (2017). *The essential guide to doing your research project* (3rd ed.). Sage

Publishing. ISBN: 9781473952089

**Additional Resources**

*How to create a data driven culture in your organization*, by Phocas Software, retrieved from

[https://www.phocassoftware.com/business-intelligence-blog/how-to-become-a-data driven-organization](https://www.phocassoftware.com/business-intelligence-blog/how-to-become-a-data%20driven-organization)

Waller, David (2020) *10 Steps to Creating a Data-Driven Culture, Harvard Business Review*

Retrieved from https://hbr.org/2020/02/10-steps-to-creating-a-data-driven-culture

Duan Yanqing, Cao Guangming, John S.Edwards. (2020) *Understanding the impact of business*

*analytics on innovation*. European Journal of Operational Research. Volume 281, Issue 3,

16 Pages 673-686.

Partial least squares regression. YouTube. https://www.youtube.com/watch?v=WKEGhyFx0Dg

Korstanje, J. (2021, July 22). Partial least squares. Medium. Retrieved April 21, 2023,

from https://towardsdatascience.com/partial-least-squares-f4e6714452a

Data Science and Big Data Analytics, Wiley; 1st edition (January 27, 2015)

Logistic and Multinomial logistic regression on SAS Enterprise Miner, 11/19/2016, Retrieved on

4/19/2023 <https://www.youtube.com/watch?v=9TQLIU3M0YE>