Predicting Delays in the Supply Chain with the Use of Machine Learning

Raghad Al-Saghir
rsa1390@rit.edu

Follow this and additional works at: https://scholarworks.rit.edu/theses

Recommended Citation

This Master's Project is brought to you for free and open access by RIT Scholar Works. It has been accepted for inclusion in Theses by an authorized administrator of RIT Scholar Works. For more information, please contact rit scholarworks@rit.edu.
Predicting Delays in the Supply Chain with the Use of Machine Learning

by

Raghad Al-Saghir

A Capstone Submitted in Partial Fulfillment of the Requirements for the
Degree of Master of Science in Professional Studies: Data Analytics

Department of Graduate Programs & Research

Rochester Institute of Technology
RIT Dubai
Fall 2022
RIT

Master of Science in Professional Studies:
Data Analytics

Graduate Capstone Approval

Student Name: Raghad Al-saghir
Graduate Capstone Title: Predicting Delays in the Supply Chain with the Use of Machine Learning

Graduate Capstone Committee:

Name: Dr. Sanjay Modak      Date: 18/01/2023
Chair of committee

Name: Dr. Khalil Al-Husseini Date: 18/01/2023
Member of committee
Acknowledgments

First, I would like to express my thanks to my parents and my siblings for motivating me throughout this journey. Then, I would like to show my gratitude to everyone who helped me succeed in my educational endeavors. Sincere thanks to Dr. Khalil for his direction and assistance during the process. I also want to express my gratitude to my programs for their support and collaboration during our research. Last but not least, I want to express my gratitude to Rochester Institute of Technology for providing this program and fostering a fantastic learning atmosphere.
Abstract

The supply chain includes all the steps in an item's life cycle starting from raw material and all the way to the customers. Delays in these steps have always been a problem faced by businesses due to the huge financial loss these delays account for. These delays frequently cause significant financial losses, irritate customers, and loss of customer confidence. The aim of this study is to construct a prediction model that predicts late deliveries before its occurrence with the use of big data and machine learning. The study used Dataco Supply Chain dataset and cleaned, visualized, and trained it using several classification machine learning algorithms. Finally, potential models were compared and the best one was chosen based on accuracy and recall values.

**Key words:** Supply chain, Late Delivery Prediction, Machine Learning, Big Data
Contents

ACKNOWLEDGMENTS .................................................................................................................. II
ABSTRACT .................................................................................................................................... III
LIST OF FIGURES .................................................................................................................. V
LIST OF TABLES .................................................................................................................... V
LIST OF EQUATIONS ........................................................................................................... V

CHAPTER 1 .................................................................................................................................. 1
  1.1 BACKGROUND INFORMATION .................................................................................. 1
  1.2 PROJECT GOALS ......................................................................................................... 2
  1.3 AIMS AND OBJECTIVES ........................................................................................... 2
  1.4 RESEARCH METHODOLOGY .................................................................................... 3
  1.5 LIMITATIONS OF THE STUDY ................................................................................ 4

CHAPTER 2 – LITERATURE REVIEW ................................................................................. 5

CHAPTER 3- PROJECT DESCRIPTION ............................................................................. 9
  3.1. DATA COLLECTION .................................................................................................. 9
  3.2. DATA DESCRIPTION ............................................................................................... 9

CHAPTER 4- DATA ANALYSIS ........................................................................................ 13
  4.1. DATA EXPLORATORY ............................................................................................. 13
  4.2. DATA CLEANSING .................................................................................................. 14
  4.3. DATA VISUALIZATION ........................................................................................... 16
  4.4. MODEL BUILDING .................................................................................................... 24
  4.4.1. ALGORITHMS USED ......................................................................................... 24
  4.4.2 MODELLING ......................................................................................................... 28
  4.4.3 MODELS COMPARISON ...................................................................................... 31

CHAPTER 5 ................................................................................................................................ 34
  5.1 CONCLUSION ............................................................................................................. 34
  5.2 RECOMMENDATIONS ............................................................................................... 34
  5.3 FUTURE WORK .......................................................................................................... 35

BIBLIOGRAPHY ..................................................................................................................... 36

APPENDIX ............................................................................................................................. 38
EXTRA WORK .......................................................................................................................... 38
R STUDIO CODE ..................................................................................................................... 43
List of Figures
Figure 1: supply chain steps (Connected Supply Chain, n.d.) ................................................................. 1
Figure 2: counts of attributes missing values generated by R studio ...................................................... 15
Figure 3: Counts of the delivery status classes ...................................................................................... 16
Figure 4: Method of payment bar plot .................................................................................................... 16
Figure 5: Shipping Mode Bar plot ......................................................................................................... 17
Figure 6: Department name bar plot ...................................................................................................... 17
Figure 7: Chi square test result of Department name ............................................................................. 18
Figure 8: Chi square test result of shipping mode ................................................................................ 18
Figure 9: Chi square test result of order region ................................................................................... 19
Figure 10: Chi square test result of Type of Payment ........................................................................... 20
Figure 11: Histograms of the numeric attributes ..................................................................................... 20
Figure 12: Location map of the orders in the United States ................................................................... 22
Figure 13: Correlation matrix of some numeric features is shown below ............................................. 23
Figure 14: ANOVA test results .............................................................................................................. 24
Figure 15: difference between LM and LG, (Krantiwadmare, 2021) ..................................................... 26
Figure 16: SVM illustration, (Support Vector Machine Algorithm, 2021) ................................................. 26
Figure 17: Linear Regression Output ...................................................................................................... 28
Figure 18: Linear Regression Confusion Matrix ...................................................................................... 29
Figure 19: Linear Regression MAE ......................................................................................................... 29
Figure 20: Support Vector Machine Confusion Matrix ............................................................................ 30
Figure 21: Logistic Regression Confusion Matrix .................................................................................. 30
Figure 22: Naive Bayes Confusion Matrix .............................................................................................. 31

List of Tables
Table 1: Attributes Description ............................................................................................................. 12
Table 2: Frequency and the percentage of appearance ......................................................................... 13
Table 3 counts of delivery statuses on the sampled data ...................................................................... 14
Table 5: Delivery status classes after encoding .................................................................................. 16

List of Equations
Equation 1: Multiple Linear Regression, (Deepanshi, 2021) ................................................................ 25
Equation 3: Accuracy Equation, (google, 2022) ......................................................................................... 31
Equation 4: Accuracy for binary classification Equation, (google, 2022) .................................................. 32
Equation 5: Recall equation, (google, 2022) ............................................................................................... 32
Chapter 1

1.1 Background Information

The transfer of raw materials from the supplier passing through the manufacturer and all the way to the customer is included in the supply chain (Lutkevich, 2021). A supply chain includes all the steps of a product's life cycle starting from issuing the raw material until the receipt of the products by the customer as shown in figure 1 (E-commerce supply chain). Any malfunction in any of the supply chain steps will result in huge financial losses that may result from product damages, inventory and holding costs, retailers and distributors shortages or surpluses, and of course customer dissatisfaction at the end of the day. In this study, delays in shipping to customers step are tackled. As online shopping or what is called E-commerce has risen significantly these days reaching around 80% of the USA population, not all shipments were shipped or delivered in time efficiently leading to customer disappointment and losing the customer in most cases (Aviso, 2022). One to three supply chain delays or interruptions, according to 58% of consumers, would cause them to completely stop purchasing from a company (Stanley, 2022). Consequently, attracting a new customer in the place of the customer lost is more expensive than maintaining the previous one. Hence, solving this delay issue has become a must to protect a business from such losses. Predicting the delay before it occurs is a great potential solution. Machine learning can be implemented in such cases to predict delays based on some predetermined information. The goal of machine learning, a subfield of artificial intelligence (AI) and computer science, is to simulate human learning processes using data and algorithms, steadily increasing the accuracy of the results (Education, 2020). In machine learning tasks, algorithms are trained using statistical techniques to create classifications or predictions and to discover important insights.

Figure 1: supply chain steps (Connected Supply Chain, n.d.)
1.2 Project goals

The main objective of the project is to propose a model that predicts delays in shipments before they occur using machine learning techniques with the use of historical data. A secondary objective is determining what factors affect or can lead to delays in shipment receipts directly or indirectly by examining a group of attributes. By knowing these factors, we can perform analysis to know what leads to delaying deliveries and how much of an impact these factors have. Consequently, recommendations on how to avoid these delays or reduce them will be endorsed.

1.3 Aims and Objectives

The main goal of this project is to develop a model that predicts the delay in delivery before it occurs based on collected data and with the use of machine learning. The model will take predetermined data of specific product delivery and predicts its status whether the delivery will reach on time, early, or late. This model is expected to work with any type of delivery method and any product category since the required data is fed into the model
1.4 Research Methodology

The first four of the six phases in the CRISP-DM framework for data mining and analysis will be followed in this project's approach. This framework was chosen because of its structural qualities. It enables the analyst to guarantee that all necessary actions for achievement are being taken. The actions are as follows:

(1) Business understanding:
Starting with this phase, we will be focusing on determining and understanding the business objectives and needs of this study. A preliminary plan will be developed in this phase.

(2) Data understanding:
In this phase, data will be explored, and attributes will be understood. Hypothesis testing, ANOVA and correlation can be used in this phase to study the dependency of the attributes to reduce the variables.

(3) Data preparation:
In this phase the dataset will get sampled, noise will be removed, and qualitative attributes will get encoded. Data will be separated to testing and training data.

(4) Modelling:
Finally, in this phase the model will be created. Many techniques will be tested such as Support Vector Machine, logistic regression, naïve bayes and linear regression. The best technique will be selected based on its and accuracy and recall values. R studio will be used to code the model.
1.5 Limitations of the Study

The main limitation of this study is the dataset. The dataset used lacks some important factors that effects the shipping process. Moreover, not all factors are useful, and some factors cannot always be provided before the delivery get shipped. In fact, getting a supply chain dataset that includes more relatable factors would have helped and released higher accuracies.
Chapter 2 – Literature Review

Delays in the supply chain are a major challenge faced by businesses all over the world. They result in products reaching at their destination later than expected. ATRI (American Transportation Research Institute) estimates that in 2016, traffic congestion alone resulted in over 1.2 billion hours of delays for trucking operations in the United States. This many delays results in increased operational expenditures of $74.5 billion (Hooper, 2018). This delay problem has risen more recently because of the Covid-19 pandemic as it restricted and mandated the availability of workers (Mckay, 2022). Moreover, the lockdown period the world survived increased the dependency of people on online shopping and home delivery. On various online platform, COVID-19 has influenced consumer shopping patterns. The pandemic has had a huge impact on consumer behavior, sales volume, and the entire supply chain (Galhotra & Dewan, 2020). Hence, solving the delays in the supply chain is crucial to minimize customer dissatisfaction, financial losses and maximize the profit of all these huge online orders.

Many previous studies have tackled this issue by proposing models to predict delays in the supply chain before it occurs. A came across study (Dias, Rezende, & Salles, 2021) suggested a solution of an analytical model based on absorbing the semi-Markov process for expecting delivery delays. A Markov renewal process is a random process that generalizes the idea of Markov jump processes in probability and statistics. In many ways, a semi-Markov process is like a Markov renewal process, with the exception that it defines a state for each time, rather than simply at the leap times (Yu, 2016). Another paper study mentioned a way for avoiding delays by predicting journey speeds which is total distance travelled over the total time, by proposing an innovative method for modelling the interactions between automobiles and urban freight trucks based on the multi-modal macroscopic fundamental diagram (MFD) for journey speed prediction in urban environments (Allister, 2020).

Digging more into the field of predicting delays in supply chains revealed that an extremely huge portion of the studies all implemented big data and machine learning models in their predictions. It was proven that big data has a variety of applications in the field of the supply chain since large data sets are produced because of logistics providers' massive management of product flow. Daily, worldwide delivery networks track the origin, destination, size, weight, content, and
position of packages where all are recorded in databases (Mikavica, Kostić-Ljubisavljevića, & Dogatovića, 2015). For instance, a study in the field of port logistic supply chain employed machine learning tools of regression, clustering, and correlation to construct an integrated management model for prediction purposes (Ouyang, 2019). Using the concepts of the Linear Regression Algorithm, Data Mining Prediction analysis, and Rapid Miner application tools, research (Wahyudi & Arroufa, 2022) attempted to provide a plan to cope with delays in the delivery of goods by employing qualitative research in their study. Moreover, an interesting study (Zhang & Perry, 2016) provided the procedure for obtaining the Oracle EBS system's inventory shipping data, then used Primitive KNN algorithms (classification) and time series forecasting algorithms (i.e., ARMA) to estimate future shipments for a collection of inventory items. The study went for the Euclidean distance and a K=1 for the KNN, which showed significantly better results than the ARMA method.

More machine learning techniques including boosting algorithms were used in this field. For instance, a study (Khiari & Olaverri-Monreal, 2022) findings demonstrated that three boosting algorithms—gradient boosting, CatBoost, and histogram gradient boosting—are the ones that scale the best in terms of runtime in comparison to the other while predicting delivery time. They have chosen based on the MAE and RMSE values. Moreover, researchers (Viellechner & Spinler, 2020) did a comparison of machine learning techniques, that are used in predicting delays in vessels, including classification and regression. For the classification models, Support Vector Machine, Random Forest, Neural Network, and Logistic Regression reflected higher accuracies. Moreover, the study used basic statistics and lasR with glmnet for variable reduction and Random Forest for validity confirmation. However, most papers tackled the prediction of a delivery time itself not the status of delivery. Furthermore, a compiling paper in the supply chain field did a literature review of all machine learning methods studies used in the supply chain field and endorsed the fact that around 88% of the machine learning techniques used in previous supply chain studies went for Hyper Box Classifier, Support Vector Machine and Neural Networks (BOUSQAOUI, ACHCHAB, & TIKITO, 2017). Another study with support vector machine application predicted outward logistics transit durations that enables providers to schedule timely risk mitigation during shipment planning. The predictive model is made up of a classifier that has been specifically trained using historical shipment, weather, and social media data for each unique source-destination pair. The Support Vector Machine model calculates the transit times for
upcoming shipments (Ben Miled, King, & Kim, 2020). Going more specifically to transportation and distribution supply chain studies, Neural Networks, Support Vector Machine, Decision Trees and K Nearest Neighbour are the top in this area.

Narrowing our search process to customer delivery delay prediction, a study in the E-commerce field employed two machine learning models to predict the time delivery of products (Bristy, Durjoy, & Hasan, 2021). The two models are the Random Forest classifier and the Support Vector Machine. The paper searched the independence of the variables using chi-squared and then found that the Support Vector Machine is the best model as it reflected higher accuracy. Another study in the same field used Support Vector Machine with and without consecutive look-back and periodic look-back techniques to forecast the delivery time (Erkmen, Nigiz, Sarı, Arlı, & Akay, 2022). The addition of the look-back techniques increased the accuracy of the model proposed. Another study in the area of predicting delays in E-commerce deliveries suggested an improved hybrid voting-based classification model with Trees, Ensemble techniques such as bagging and boosting, and shipping mode and scheduled shipment time-enabled features (Wani, Singh, Vivekanad, & Manoj, 2022). The study revealed excellent results with high accuracies, however its models included some unrealistic attributes that cannot be provided in real life cases.

This field is rich with good papers that implement machine learning and big data in building predictive models. An interesting study by Berrones-Sanz believes that late deliveries is caused by the drivers themselves, hence in order to better understand how driving performance is related to work, he proposed a study that look at how working conditions may contribute to late deliveries and reduces noncompliance in deliveries attributable to driver-related factors in a manufacturing company (Berrones-Sanz, 2021). The binary logistic regression was used to model and forecast truckers' on-time deliveries in his study based on the availability of variables that affect the working circumstances of drivers. Finally, the last paper came across in this field by Bouhadi, Azmani, fiouh mentions that there are many internal and external factors that causes delays in deliveries, and the internal factors are more of in the effect of the external ones. Internal factors include decisions made by the transportation industry about the choice of resources and the applicability of planning. The external factors include traffic jams, the state of the weather, and the availability of delivery bays. Moreover, the study constructed of a fuzzy-bayesian strategy that foresees potential delays impacting the efficient operation of a delivery operation using predictive
analysis integrating Bayesian networks (BNs) and fuzzy logic. These two papers gave a good idea of other potential attributes that can be implemented to advance my prediction model and obtain more accurate predictions.

In conclusion, the research process revealed that a gap exists in implementing machine learning only in foreseeing the delays by predicting the status of delivery before the shipment is shipped using realistic attributes. The search procedure yielded a number of takeaways, including:

- Machine learning is a powerful tool that can be used in solving any problem a supply chain can face Based on big data such as delays which is a major challenging issue.
- Delays in any step-in supply chain will cause enormous delays in the following steps leading to huge financial losses.
- Delays can be predicted either qualitatively by predicting the delivery time or quantitatively by predicting the status of delivery.
- Other tools can be combined with machine learning to increase the accuracy of the prediction such as look-back approaches.
- For classification delay prediction problems, SVM, logistic regression, Neural Networks, and naïve bayes are the best techniques that showed high accuracies in all previous studies.
- Basic statistics and chi-square are good approaches to studying the dependence of variables, hence reducing them.
- Techniques in predicting delays can be compared using MAE, RMSE, accuracy and recall values.
- Factors such as street congestion, weather conditions, and driver circumstances do effect meeting the delivery deadlines.
Chapter 3- Project Description

This project will involve multiple phases to construct a model that predicts delays in the deliveries before they occur. First, the dataset will get explored for better understanding then cleaned to avoid any trivial errors while modelling and to make it more understandable. After that, the data visualization phase comes for better understanding of the attributes and to remove all doubts regarding the dependency of the variables. This step will lead to reducing my attributes and keeping the relevant ones. Finally, machine learning techniques will be employed to develop the delay prediction model.

In this chapter, all information regarding the dataset will be discussed. The source of data and how was it collected will be detailed in this phase. Moreover, the description of all the variables in the chosen dataset will be defined here.

3.1. Data Collection

The analysis utilized a Dataset of Supply Chains from the organization DataCo Global. The dataset was accessed through Kaggle. The dataset was collected in the USA in the period between January 2015 and September 2017. The dataset includes the status of delivery which is our aim of prediction including to 52 more variables.

3.2. Data Description

The dataset contains 53 different variables of customers information, and around 180,000 observations. The dataset is mainly about three types of products which are clothing, electronic supplies, and Sports. This dataset was chosen because of the diversity of its variables as it includes many good information about the customer and the delivery process, which makes the prediction process effective and realistic. The following table describes all the attributes with the data type in each attribute.
<table>
<thead>
<tr>
<th>FIELDS</th>
<th>DESCRIPTION</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>Type of transaction made</td>
<td>Character</td>
</tr>
<tr>
<td>Days for shipping (real)</td>
<td>Actual shipping days of the purchased product</td>
<td>Numeric</td>
</tr>
<tr>
<td>Days for shipment (scheduled)</td>
<td>Days of scheduled delivery of the purchased product</td>
<td>Numeric</td>
</tr>
<tr>
<td>Benefit per order</td>
<td>Earnings per order placed</td>
<td>Numeric</td>
</tr>
<tr>
<td>Sales per customer</td>
<td>Total sales per customer made per customer</td>
<td>Numeric</td>
</tr>
<tr>
<td>Delivery Status</td>
<td>Delivery status of orders: Advance shipping , Late delivery , Shipping cancelled , Shipping on time</td>
<td>Character</td>
</tr>
<tr>
<td>Late_delivery_risk</td>
<td>Categorical variable that indicates if sending is late (1), it is not late (0).</td>
<td>Numeric</td>
</tr>
<tr>
<td>Category Id</td>
<td>Product category code</td>
<td>Numeric</td>
</tr>
<tr>
<td>Category Name</td>
<td>Description of the product category</td>
<td>Character</td>
</tr>
<tr>
<td>Customer City</td>
<td>City where the customer made the purchase</td>
<td>Character</td>
</tr>
<tr>
<td>Customer Country</td>
<td>Country where the customer made the purchase</td>
<td>Character</td>
</tr>
<tr>
<td>Customer Email</td>
<td>Customer's email</td>
<td>Character</td>
</tr>
<tr>
<td>Customer Fname</td>
<td>Customer name</td>
<td>Character</td>
</tr>
<tr>
<td>Customer Id</td>
<td>Customer ID</td>
<td>Numeric</td>
</tr>
<tr>
<td>Customer Lname</td>
<td>Customer last name</td>
<td>Character</td>
</tr>
<tr>
<td>Customer Password</td>
<td>Masked customer key</td>
<td>Character</td>
</tr>
<tr>
<td>Customer Segment</td>
<td>Types of Customers: Consumer, Corporate , Home Office</td>
<td>Character</td>
</tr>
<tr>
<td>Customer State</td>
<td>State to which the store where the purchase is registered belongs</td>
<td>Character</td>
</tr>
<tr>
<td>Customer Street</td>
<td>Street to which the store where the purchase is registered belongs</td>
<td>Character</td>
</tr>
<tr>
<td>Customer Zipcode</td>
<td>Customer Zipcode</td>
<td>Numeric</td>
</tr>
<tr>
<td>Department Id</td>
<td>Department code of store</td>
<td>Numeric</td>
</tr>
<tr>
<td>Department Name</td>
<td>Department name of store</td>
<td>Character</td>
</tr>
<tr>
<td>Field</td>
<td>Description</td>
<td>Type</td>
</tr>
<tr>
<td>------------------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>-----------</td>
</tr>
<tr>
<td>Latitude</td>
<td>Latitude corresponding to location of store</td>
<td>Numeric</td>
</tr>
<tr>
<td>Longitude</td>
<td>Longitude corresponding to location of store</td>
<td>Numeric</td>
</tr>
<tr>
<td>Market</td>
<td>Market to where the order is delivered: Africa, Europe, LATAM, Pacific Asia, USCA</td>
<td>Character</td>
</tr>
<tr>
<td>Order City</td>
<td>Destination city of the order</td>
<td>Character</td>
</tr>
<tr>
<td>Order Country</td>
<td>Destination country of the order</td>
<td>Character</td>
</tr>
<tr>
<td>Order Customer Id</td>
<td>Customer order code</td>
<td>Numeric</td>
</tr>
<tr>
<td>order date (DateOrders)</td>
<td>Date on which the order is made</td>
<td>Numeric</td>
</tr>
<tr>
<td>Order Id</td>
<td>Order code</td>
<td>Numeric</td>
</tr>
<tr>
<td>Order Item Cardprod Id</td>
<td>Product code generated through the RFID reader</td>
<td>Numeric</td>
</tr>
<tr>
<td>Order Item Discount</td>
<td>Order item discount value</td>
<td>Numeric</td>
</tr>
<tr>
<td>Order Item Discount Rate</td>
<td>Order item discount percentage</td>
<td>Numeric</td>
</tr>
<tr>
<td>Order Item Id</td>
<td>Order item code</td>
<td>Numeric</td>
</tr>
<tr>
<td>Order Item Product Price</td>
<td>Price of products without discount</td>
<td>Numeric</td>
</tr>
<tr>
<td>Order Item Profit Ratio</td>
<td>Order Item Profit Ratio</td>
<td>Numeric</td>
</tr>
<tr>
<td>Order Item Quantity</td>
<td>Number of products per order</td>
<td>Numeric</td>
</tr>
<tr>
<td>Sales</td>
<td>Value in sales</td>
<td>Numeric</td>
</tr>
<tr>
<td>Order Item Total</td>
<td>Total amount per order</td>
<td>Numeric</td>
</tr>
<tr>
<td>Order Profit Per Order</td>
<td>Order Profit Per Order</td>
<td>Numeric</td>
</tr>
<tr>
<td>Order Region</td>
<td>Region of the world where the order is delivered: Southeast Asia, South Asia, Oceania, Eastern Asia, West Asia, West of USA, US Center, West Africa, Central Africa, North Africa, Western Europe, Northern, Caribbean, South America, East Africa, Southern Europe, East of USA, Canada, Southern Africa, Central Asia, Europe, Central America, Eastern Europe, South of USA</td>
<td>Character</td>
</tr>
<tr>
<td>Order State</td>
<td>State of the region where the order is delivered</td>
<td>Character</td>
</tr>
<tr>
<td>Order Status</td>
<td>Order Status: COMPLETE, PENDING, CLOSED, PENDING_PAYMENT, CANCELED</td>
<td>Character</td>
</tr>
<tr>
<td>Table 1: Attributes Description</td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Product Card Id</td>
<td>Product code</td>
<td>Numeric</td>
</tr>
<tr>
<td>Product Category Id</td>
<td>product category code</td>
<td>Numeric</td>
</tr>
<tr>
<td>Product Description</td>
<td>Product Description</td>
<td>Numeric</td>
</tr>
<tr>
<td>Product Image</td>
<td>Link of visit and purchase of the product</td>
<td>Character</td>
</tr>
<tr>
<td>Product Name</td>
<td>Product Name</td>
<td>Character</td>
</tr>
<tr>
<td>Product Price</td>
<td>Product Price</td>
<td>Numeric</td>
</tr>
<tr>
<td>Product Status</td>
<td>Status of the product stock: If it is 1 not available, 0 the product is available</td>
<td>Numeric</td>
</tr>
<tr>
<td>Shipping date (Date Orders)</td>
<td>Exact date and time of shipment</td>
<td>Numeric</td>
</tr>
<tr>
<td>Shipping Mode</td>
<td>The following shipping modes are presented: Standard Class, First Class, Second Class, Same Day</td>
<td>Character</td>
</tr>
</tbody>
</table>
Chapter 4- Data Analysis

4.1. Data Exploratory

To get started, I have chosen Delivery Status attribute to be my response or what this study is aiming to predict. Status delivery is a qualitative variable with four main values which are: late delivery, advance shipping, shipping on time and shipping cancelled. The following table summarizes the frequency and the percentage of appearance.

<table>
<thead>
<tr>
<th>Delivery Status</th>
<th>Count of Delivery Status</th>
<th>Count of Delivery Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Late delivery</td>
<td>98977</td>
<td>55%</td>
</tr>
<tr>
<td>Advance shipping</td>
<td>41592</td>
<td>23%</td>
</tr>
<tr>
<td>Shipping on time</td>
<td>32196</td>
<td>18%</td>
</tr>
<tr>
<td>Shipping canceled</td>
<td>7754</td>
<td>4%</td>
</tr>
<tr>
<td>Grand Total</td>
<td>180519</td>
<td>100%</td>
</tr>
</tbody>
</table>

Table 2: Frequency and the percentage of appearance

We are only interested in predicting whether the delivery will be late or on time. Hence, we will sample our dataset to use the observations with late and on time delivery. Late delivery and on time delivery forms around 55% and 18%, which is a total of 73% of our original dataset and 75% and 25% of our sampled data set as shown in the table below.
Table 3 counts of delivery statuses on the sampled data

<table>
<thead>
<tr>
<th>Delivery Status</th>
<th>Count of Delivery Status</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Late delivery</td>
<td>98977</td>
<td>75%</td>
</tr>
<tr>
<td>Shipping on time</td>
<td>32196</td>
<td>25%</td>
</tr>
<tr>
<td>Grand Total</td>
<td>131173</td>
<td>100%</td>
</tr>
</tbody>
</table>

Having a 75% as a percentage of late delivery is insane. An actual solution must be found, and immediate actions must be taken to decrease this percentage or even eliminate it.

4.2. Data Cleansing

Data cleansing is one of the significant steps in data analytics as it improves our data quality and in doing so, the overall productivity is increased, and costly errors are avoided in some cases. First, we will start by taking a sample of 16383 observations from our dataset. Data cleansing can be done through checking six quality dimensions which are: accuracy, completeness, consistency, timeliness, validity, and uniqueness.

Luckily my dataset met all the data quality dimensions except the completeness where is showed some missing values for some attributes as shown in the below table generated by R studio.
For Figure 2: counts of attributes missing values generated by R studio

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Count</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Product Description</td>
<td>0</td>
<td>Days for shipping (real)</td>
</tr>
<tr>
<td>Sales per customer</td>
<td>0</td>
<td>Benefit per order</td>
</tr>
<tr>
<td>Late_delivery_risk</td>
<td>0</td>
<td>Delivery Status</td>
</tr>
<tr>
<td>Category Name</td>
<td>0</td>
<td>Category Id</td>
</tr>
<tr>
<td>Customer Country</td>
<td>0</td>
<td>Customer City</td>
</tr>
<tr>
<td>Customer Fname</td>
<td>0</td>
<td>Customer Email</td>
</tr>
<tr>
<td>Customer Lname</td>
<td>0</td>
<td>Customer Password</td>
</tr>
<tr>
<td>Customer Segment</td>
<td>0</td>
<td>Customer State</td>
</tr>
<tr>
<td>Customer Street</td>
<td>0</td>
<td>Customer Zipcode</td>
</tr>
<tr>
<td>Department Id</td>
<td>0</td>
<td>Department Name</td>
</tr>
<tr>
<td>Latitude</td>
<td>0</td>
<td>Longitude</td>
</tr>
<tr>
<td>Market</td>
<td>0</td>
<td>Order City</td>
</tr>
<tr>
<td>Order Country</td>
<td>0</td>
<td>Order Customer Id</td>
</tr>
<tr>
<td>order date (DateOrders)</td>
<td>0</td>
<td>Order Id</td>
</tr>
<tr>
<td>Order Item Cardprod Id</td>
<td>0</td>
<td>Order Item Discount</td>
</tr>
<tr>
<td>Order Item Discount Rate</td>
<td>0</td>
<td>Order Item Id</td>
</tr>
<tr>
<td>Order Item Product Price</td>
<td>0</td>
<td>Order Item Profit Ratio</td>
</tr>
<tr>
<td>Order Item Quantity</td>
<td>0</td>
<td>Sales</td>
</tr>
<tr>
<td>Order Item Total</td>
<td>0</td>
<td>Order Profit Per Order</td>
</tr>
<tr>
<td>Order Region</td>
<td>0</td>
<td>Order State</td>
</tr>
<tr>
<td>Order Status</td>
<td>0</td>
<td>Order Zipcode</td>
</tr>
<tr>
<td>Product Card Id</td>
<td>14438</td>
<td>Product Category Id</td>
</tr>
<tr>
<td>Product Name</td>
<td>65383</td>
<td>Product Image</td>
</tr>
<tr>
<td>Product Image</td>
<td>0</td>
<td>Product Price</td>
</tr>
<tr>
<td>Product Status</td>
<td>0</td>
<td>shipping date (DateOrders)</td>
</tr>
<tr>
<td>Shipping Mode</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

As shown above, Product Description and Order Zip code have some missing values. However, since these two attributes are unimportant, they will be omitted with some other irrelevant attributes including: Customer Email, Product Status, Customer Password, Customer Street, Customer Fname, Customer Lname, Product Description, Product Image, Order Zipcode, Shipping Date, Order Customer Id, order date (DateOrders), Order Id, Order Item Cardprod Id, Order Item Discount, Order Item Id, Product Card Id.
Next step was encoding our response to (1,0) as shown in the table below. The counts of each of the classes are shown in figure.

<table>
<thead>
<tr>
<th>Late Delivery</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shipping on time</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 4: Delivery status classes after encoding

4.3. Data Visualization

For better understanding of the dataset a visualization of some attributes will be done. First, to understand our customers' tends, the method of payment is visualized in a bar plot as shown below. It can be concluded that a huge majority of the customers go for the debit payment method which agrees with fact that the most popular payment methods are credit and debit cards (Shopify, 2022). Companies that offer credit to buyers include Visa, Mastercard, American Express, and Discover; they pay for the purchase price, and clients settle their card balances monthly.
Another categorical attribute is visualized for customer better understanding which is the shipping mode. In our dataset we mainly have 4 shipping modes which are the standard, first, second and same day. The bar plot of the shipping mode is shown below. It can be told that around 50 percent of our customers go for the default standard class type shipping mode.

![Shipping Mode Bar Plot](image)

Figure 5: Shipping Mode Bar plot

The following bar plot summarizes the department name of the product purchased by customers. It shows that most of the products delivered are from the apparel and golf departments while small proportion are from the health and beauty.

![Department Name Bar Plot](image)

Figure 6: Department name bar plot
Proceeding forward, we need to study the correlation of our variables and response. Our variables vary between categorical and numerical variables so the testing method will differ. Starting with our categorical variables, we tested the relation between Delivery Status (Response) and the department name, shipping mode, order region and type. chi-square test is used as shown below. R studio was used for this phase.

1) Department name: The P value was found to be greater than our 5 percent threshold as shown below from the R output. It can be concluded that delays in deliveries have nothing to do with the type of products.

![Figure 7: Chi square test result of Department name](image_url)

<table>
<thead>
<tr>
<th>Apparel</th>
<th>Late delivery</th>
<th>Shipping on time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Book Shop</td>
<td>4387</td>
<td>1425</td>
</tr>
<tr>
<td>Discs Shop</td>
<td>178</td>
<td>55</td>
</tr>
<tr>
<td>Fan Shop</td>
<td>1769</td>
<td>551</td>
</tr>
<tr>
<td>Fitness</td>
<td>682</td>
<td>208</td>
</tr>
<tr>
<td>Footwear</td>
<td>1411</td>
<td>473</td>
</tr>
<tr>
<td>Golf</td>
<td>3135</td>
<td>992</td>
</tr>
<tr>
<td>Health and Beauty</td>
<td>24</td>
<td>7</td>
</tr>
<tr>
<td>Outdoors</td>
<td>604</td>
<td>180</td>
</tr>
<tr>
<td>Pet Shop</td>
<td>43</td>
<td>15</td>
</tr>
<tr>
<td>Technology</td>
<td>157</td>
<td>44</td>
</tr>
</tbody>
</table>

Pearson’s Chi-squared test
X-squared = 3.4762, df = 10, p-value = 0.9679

2) Shipping mode: the p value was found to be way less than 5 percent as shown below. it can be concluded that delays in delivery is related with what type of shipping mode our customers chose. Hence, this variable will get one hot encoded to be used in our prediction model.

![Figure 8 Chi square test result of shipping mode](image_url)

<table>
<thead>
<tr>
<th>First Class</th>
<th>Late delivery</th>
<th>Shipping on time</th>
</tr>
</thead>
<tbody>
<tr>
<td>3563</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Same Day</td>
<td>508</td>
<td>412</td>
</tr>
<tr>
<td>Second Class</td>
<td>3549</td>
<td>873</td>
</tr>
<tr>
<td>Standard Class</td>
<td>5204</td>
<td>2674</td>
</tr>
</tbody>
</table>

Pearson’s Chi-squared test
X-squared = 17.24, df = 3, p-value < 2.2e-16
3) Order Region: The p value was found to be way less than 5 percent as shown below. It can be concluded that the region where the order is delivered to is related to the delivery status. Hence, this variable will get one hot encoded to be used in our prediction model. However, for the scope of this study, we will only consider orders in South America, Central America, and Western Europe.

![Chi-square test result of order region](image)

**Figure 9: Chi square test result of order region**

4) Type: The p value was found to be way less than 5 percent as shown below. It can be concluded that delays in delivery is related to the payment method our customers chose. However, this piece of information is not always accessed since it has to do with the finance department where some classify it under customer privacy. Moreover, payment method is constant for the same customer in most cases so it will not change much causing changes in the delivery status.
Moving on to our numerical variables, we will start by plotting a histogram of all of them to get some preliminary information about their values and to catch any unusual trends or behaviors. The diagram below included all the histograms of the numeric attributes and after encoding the shipment mode.

Figure 11: Histograms of the numeric attributes
From the histograms we can tell the following:

- The histogram of the product prices and sales variables are right skewed which means that most of the count density is concentrated in less-costed items.
- The maximum number of scheduled (promised) days is 4 which as well has the most counts, while the maximum number of real shipping days is 6 which shows lack of credibility and dishonesty with customers and emphasizes how serious this issue is.
- Most customers go for ordering only 1 item in each purchasing process which indicates that either they are new clients and are so specific in their orders, or the lack of confidence and comfort while purchasing online. However, such issues can be solved by investing more in marketing and applying attracting customers strategy.
- Some attributes are redundant such as Category ID which is like Department ID, and Late Delivery Status and Late Delivery Risk. Hence, Category ID and Late Delivery Risk will be dropped.
- Some attributes are meaningless such as Order Item ID, Customer Zip code and Customer ID. These attributes will be dropped as well.
To get an overview of the location of the orders in the data set, a location map of the orders in the United States was generated using the ggmap function in R studio and using the Key: "AIzaSyBHgGu3oGbS4sC4_m3wWuFs8R4kGQ8X6A8"

Figure 12: Location map of the orders in the United States
Moving forward, a correlation plot between the selected numerical variables was done to study the relation between the variables and reduce them if needed. A correlation matrix of some numeric features is shown below.

![Correlation Matrix](image)

**Figure 13: Correlation matrix of some numeric features is shown below**

From the correlation matrix we can conclude the following:

- There is strong relation between the scheduled days for shipment and the real days for shipment.
- Good relation between the sales per customer and the item per customer which make sense since the less the cost the more the demand.
- Moderate relation negative and positive relation between the delivery status and the scheduled and real shipments days.
- Zero correlation between the delivery status and any cost related attributes which means that prices and cost have nothing to do with the delivery shipped late.
To know more about the strength of the relationship between my response and the scheduled and real shipping days ANOVA tests were used and the results was as shown below. The p-value is less than 5 percent, hence we reject our null hypothesis concluding that they are strongly correlated.

![ANOVA test results](image)

To conclude this phase we can close it by saying that the shipping mode, Order Region, real and scheduled shipping days have affects on our response which is the Delivery status.

### 4.4. Model Building

Proceeding forward, and after getting our dataset cleaned, understanding our attributes and pointing out the significant attributes, we can start modeling. However, and to be precise in our solution, days of shipment (real) attribute is not always determined prior to the delivery process and if so it will be provided before the item gets shipped in a relatively small period of time. Hence, and to keep this study on the highest level of reality, we will drop this attribute. However, for experimenting purposes, we did include the days of shipment (real) in our modeling just to see the results and to test our models. Results of this modeling is found under section 6.1.

### 4.4.1. Algorithms Used

In this study, four machine learning algorithms are used to build the predictive classification model that predicts our delivery status based on some predetermined attributes. The four algorithms are: Linear Regression, Logistic Regression, Support Vector Machine, Naïve Bayes. A brief explanation of how these algorithms work is as follows.
**Linear Regression**

The supervised machine learning approach known as "linear regression" identifies the linear relationship between the dependent and independent variables by finding the best-fitting linear line between them (Deepanshi, 2021). There are 2 types of linear regression: simple and multiple. When only one independent variable is present, simple linear regression must be used to determine its linear connection to the dependent variable. In contrast, several independent variables are used in multiple linear regression to identify the relationships. In this study and since we have multiple variables, multiple linear regression will be used. The formula of the multiple linear regression is as follows:

\[
y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 \ldots + b_n x_n
\]

*Equation 1: Multiple Linear Regression, (Deepanshi, 2021)*

where \( y \) is the dependent variable, \( b_0 \) is the intercept, and \( b_1, b_2, b_3, b_4, \ldots, b_n \) are the coefficients or slopes of the independent variables \( x_1, x_2, x_3, \ldots, x_n \).

**Logistic Regression**

logistic regression is a statistical technique used when the dependent variable is dichotomous, or binary (Banoula, 2022). Data and the relationship between one dependent variable and one or more independent variables are described using logistic regression. Nominal, ordinal, or interval types are all acceptable for the independent variables. Its name comes from the idea behind the logistic function it employs. The sigmoid function is another name for the logistic function. This logistic function has a value between 0 and 1. The graph below shows the difference between logistic and linear regression.
Support Vector Machine

The SVM algorithm's objective is to establish the best line or decision boundary that can divide n-dimensional space into classes, allowing us to quickly classify fresh data points in the future. A hyperplane is the name given to this optimal decision boundary. SVM selects the extreme vectors and points that aid in the creation of the hyperplane (Support Vector Machine Algorithm, 2021). Support vectors, which are used to represent these extreme instances, form the basis for the SVM method. Consider the diagram below, where a decision boundary or hyperplane is used to categorize two distinct categories:
**Naïve Bayes**

The Naive Bayes algorithm is a supervised learning method for classification problems that is based on the Bayes theorem. It is mostly employed in text categorization with a large training set. The Naive Bayes Classifier is one of the most straightforward and efficient classification algorithms available today. It aids in the development of quick machine learning models capable of making accurate predictions. Being a probabilistic classifier, it makes predictions based on the likelihood that an object will occur (Naïve Bayes Classifier Algorithm, 2021). The Bayes theorem is used to calculate the likelihood of a hypothesis given the available data. The conditional probability determines this. The Bayes theorem's formula is as follows:

\[
P(A|B) = \frac{P(B|A)P(A)}{P(B)}
\]


Where, \(P(A|B)\) is Posterior probability, \(P(B|A)\) is Likelihood probability, \(P(A)\) is Prior Probability and \(P(B)\) is Marginal Probability.
4.4.2 Modelling

In this section, we will start modelling after removing the day of shipment (Real attribute).

**Linear Regression**: Starting with the linear Regression, we will start by building our linear model. The following output was obtained.

![Linear Regression Output](image)

We obtained a small p value and a small residual error, as seen above, proving that our model can be fitted by a linear model. As a result, we can now use linear regression to make predictions. The confusion matrix and mean absolute error when using linear regression to forecast the delivery status are displayed in the result shown below.

![Figure 17: Linear Regression Output](image)
An accuracy of 74.84, a recall score of 50% and 0.375 MAE value were obtained from the Linear Regression.
Support Vector Machine: The second machine learning algorithm implemented is support vector machine. The below confusion matrix was obtained.

Confusion Matrix and Statistics

```
svm1.pred
0 1
0 9 507
1 13 1538
```

Accuracy : 0.7484
95% CI : (0.7291, 0.767)
No Information Rate : 0.5894
P-Value [Acc > NIR] : 1
Kappa : 0.0133
Mcnemar’s Test P-Value : <2e-16

Sensitivity : 0.409091
Specificity : 0.752078
Pos Pred Value : 0.017442
Neg Pred Value : 0.991618
Precision : 0.017442
Recall : 0.409091
F1 : 0.033457
Prevalence : 0.010643
Detection Rate : 0.004354
Detection Prevalence : 0.249637
Balanced Accuracy : 0.580585

'Positive' Class : 0

Figure 20: Support Vector Machine Confusion Matrix

An accuracy of 74.5%, a recall value of 41% were obtained from the support vector machine.

Logistic Regression: Next, the logistic regression algorithm was implemented on our dataset leaving us with the following confusion matrix

```
log1.predict
0 1
0 11 509
1 5 1542
```

Accuracy : 0.7513
95% CI : (0.7321, 0.7698)
No Information Rate : 0.5923
P-Value [Acc > NIR] : 1
Kappa : 0.0264
Mcnemar’s Test P-Value : <2e-16

Sensitivity : 0.687500
Specificity : 0.751828
Pos Pred Value : 0.021154
Neg Pred Value : 0.996768
Precision : 0.021154
Recall : 0.687500
F1 : 0.041045
Prevalence : 0.00741
Detection Rate : 0.005322
Detection Prevalence : 0.251572
Balanced Accuracy : 0.719664

'Positive' Class : 0

Figure 21: Logistic Regression Confusion Matrix
An accuracy of 75.13%, a recall value of 68% were obtained from the Logistic Regression.

**Naïve Bayes:** The second algorithm used is naïve bayes. The below confusion matrix was obtained.

![Naïve Bayes Confusion Matrix](image)

An accuracy of 46.35%, a recall value of 32% were obtained from the naïve bayes.

4.4.3 Models Comparison

The next step is to chose the best model between the models built in the previous section. A comparison table between the models will be done. The Accuracy and the Recall scores will be taken into consideration when choosing the best model to chose the model with the highest accuracy and recall.

**Accuracy:** One statistic for measuring the effectiveness of classification models is accuracy. The percentage of predictions our model correctly predicted is known as accuracy (google, 2022). Properly, accuracy is defined as follows:

\[
\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}
\]

*Equation 3: Accuracy Equation,* (google, 2022).
In our case of binary classification, accuracy is calculated in terms of negatives and positives as following:

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}
\]

*Equation 4: Accuracy for binary classification Equation, (google, 2022).*

Where \( TP = \) True Positives, \( TN = \) True Negatives, \( FP = \) False Positives, and \( FN = \) False Negatives.

**Recall:** Recall is a metric that measures the proportion of accurate positive predictions among all possible positive predictions (Brownlee, 2020). Recall gives an idea of the positive class's coverage in this way. The recall equation is as follows:

\[
\text{Recall} = \frac{\text{TruePositives}}{\text{TruePositives} + \text{FalseNegatives}}
\]

*Equation 5: Recall equation, (google, 2022).*

The below table summarizes the accuracy and recall values of our models:

<table>
<thead>
<tr>
<th>MODEL</th>
<th>ACCURACY</th>
<th>RECALL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear Regression</td>
<td>74.84%</td>
<td>50%</td>
</tr>
<tr>
<td>Support Vector Machine</td>
<td>74.84%</td>
<td>41%</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>75.13%</td>
<td>68%</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>46.355</td>
<td>32%</td>
</tr>
</tbody>
</table>

As shown above, logistic regression has the highest accuracy rate that’s around 75% in comparison with the rest of the models. Naïve Bayes showed an unacceptable rate that’s around
46.35% which makes it risky to implement. However, this low accuracy is not surprising since naïve Bayes assumes zero dependency between the attributes, which is not the case in our dataset as discussed in the correlation matrix of the numeric values in section 4.2 of this study. Going back to the logistic regression, it also shows the highest recall value among all the models. In conclusion we can infer that logistic regression is the best algorithm that can be used to build our prediction model.
5.1 Conclusion

A supply chain is a network of businesses that purchases raw materials, transforms them into intermediate products, and then deliver those end products to clients. Many delays occur in the supply chain steps and these delays are so unwanted by the organizations and the clients and leads to huge financial loses beside customer dissatisfaction and loss of customer trust. This study tracked the delays in the last step of the supply chain which is the customer delivery step. Predicting these delays is so challenging since it has the customer himself involved and is critical to avoid the mentioned losses. Machine learning and big data can be employed to develop a prediction model that predicts the late deliveries before they get shipped. A huge dataset was found through Kaggle that can be used for modelling. The dataset got cleaned by removing duplicates, visualized and unimportant attributes were dropped. Four machine learning classification algorithms were used to build four potential models which are: Linear Regression, Logistic Regression, Support Vector Machine and Naïve Bayes. Logistic Regression reflected the highest accuracy and a good recall score which makes it the best model.

5.2 Recommendations

As illustrated in section (4.4) our chosen model obtained an accuracy rate of 75%. This rate in considered acceptable and good as Hendricks points out that Industry benchmarks range from 70% to 90% where everything that is greater than 70% is considered to be realistic and useful model data output (Hendricks, 2022). However, higher rates remain always our aim and our proposed model has a room of improvement. As a first recommendation to boost our model, new attributes need to be introduced. These attributes can include the delivery method of transportation which is the type of vehicle that carries the shipment whether it’s a normal car, motor bike or a minivan, and how often does these vehicles breakdown (Victor, 2020). Another attribute is the weather conditions, where the weather impacts the street traffic and congestion. The Federal Highway Administration (FHWA) estimates that weather delays account for 23% of
all traffic delays (Weatheroptics, 2022). Furthermore, the driver condition and mental health impacts the delivery status. For instance, new drivers or dissatisfied drivers would likely perform less efficiently than a senior or satisfied driver. Consequently, collecting such information would enhance our model and generate more accurate predictions.

5.3 Future Work

Moving on to the next step, the mentioned recommendations will be taken into consideration to improve the model. Moreover, the model itself can get upgraded to anticipate other delivery statuses beside late delivery and on time delivery. The upgraded model will predict early deliveries as well. As late deliveries are unwanted by customers, early ones are as unwanted to. It was revealed that a significant portion of customers of around 31% see early deliveries inconvenient, and 60% said they would be hesitant to make another purchase from the same company when a delivery did not arrive within the specified timeframe regardless of it being early or late (Insights, 2022).
Bibliography


Why Early Isn’t the Same as On-Time for Last Mile Deliveries. Retrieved from DispatchTrack: https://www.dispatchtrack.com/blog/early-last-mile-deliveries


Appendix

Extra Work

In this section, the project modeling was done while including the days of shipment (Real) attribute. The following output were obtained.

**Linear regression:** as a start we will go with the linear regression model. The following output was generated by R studio.

![Linear regression output](image)

---

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.101e-15 on 4110 degrees of freedom
Multiple R-squared: 1, Adjusted R-squared: 1
F-statistic: 3.037e+31 on 21 and 4110 DF, p-value: < 2.2e-16
As shown above, we got a small p value, and a small residual error that indicates that our model can be fit in a linear model. Hence, we can now start predicting using linear regression. The below output shows the confusion matrix and the mean absolute error when implementing linear regression to predict the delivery status.

An accuracy of 99.82, an F1 value of 1 and 0.373 MAE value were obtained from the linear regression.

**Logistic Regression:** next, the logistic regression algorithm was implemented on path1 dataset leaving us with the following results.
After doing our predictions on our test dataset the following confusion matrix and MAE value were obtained.

Confusion Matrix and Statistics

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>508</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1558</td>
</tr>
</tbody>
</table>

Accuracy : 1
95% CI : (0.9982, 1)
No Information Rate : 0.7541
P-Value [Acc > NIR] : < 2.2e-16
Kappa : 1

McNemar's Test P-Value : NA

Sensitivity : 1.0000
Specificity : 1.0000
Pos Pred Value : 1.0000
Neg Pred Value : 1.0000
Precision : 1.0000
Recall : 1.0000
F1 : 1.0000
Prevalence : 0.2459
Detection Rate : 0.2459
Detection Prevalence : 0.2459
Balanced Accuracy : 1.0000

'Positive' Class : 0

[1] 26.18804

An accuracy of 99.82, an F1 value of 1.0 and 26.18 MAE value were obtained from the logistic regression.
Support Vector Machine: after that, a support vector machine algorithm is applied to generate our prediction model. The following confusion matrix and absolute mean error value were obtained from our support vector machine model.

<table>
<thead>
<tr>
<th>svm1.pred</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>505</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1558</td>
</tr>
</tbody>
</table>

Accuracy: 0.9985
95% CI: (0.9958, 0.9997)
No Information Rate: 0.7556
P-Value [Acc > NIR]: <2e-16
Kappa: 0.9961
Mcnemar's Test P-Value: 0.2482

Sensitivity: 1.0000
Specificity: 0.9981
Pos Pred Value: 0.9941
Neg Pred Value: 1.0000
Precision: 0.9941
Recall: 1.0000
F1: 0.9970
Prevalence: 0.2444
Detection Rate: 0.2444
Detection Prevalence: 0.2459
Balanced Accuracy: 0.9990

'Positive' Class: 0

[1] 0.3833377

An accuracy of 99.85, an fl value of 0.9970 and 0.3833 MAE value were obtained from the support vector machine.

Naïve Bayes: The last algorithm implemented in the Naïve Bayes Classifier. The below confusion matrix was obtained.
An accuracy of 99.81, a F1 score of 0.9961 were obtained from the naïve bayes.
R studio Code

---
title: "capstone_3"
author: "Raghad"
date: "2022-11-20"
output: pdf_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```  

## R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```{r cars}
library(ggplot2)
library(tidyverse)
library(lubridate)
library(dplyr)
library(scales)
library(plyr)
library(caTools)
library(class)
library(rpart)
library(dplyr)
library(tinytex)
```
library(latexpdf)
library(lubridate)
library(e1071)
library(caret)
library(rpart)
library(tinytex)
library(scales)

---

---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```
library(caTools)
library(class)
library(rpart)
library(dplyr)
library(tinytex)
library(latexpdf)
library(lubridate)
library(e1071)
library(caret)
library(rpart)
library(tinytex)
library(scales)

```r
str(DataCoSupplyChainDataset_version_2_)
```

```r
delaydata<-DataCoSupplyChainDataset_version_2_
delaydata$`Delivery Status`<- ifelse(delaydata$`Delivery Status`=="Late delivery",1,0)
summary(delaydata)
table(delaydata$`Delivery Status`)
```

```r
```
num_cols <- unlist(lapply(delaydata, is.numeric))  # Identify numeric columns
num_cols

...{
  pressure, echo=FALSE}
data_num <- delaydata[, num_cols]  # Subset numeric columns of data
data_num
summary(data_num)
...

...{
  pressure, echo=FALSE}
data(delaydata)
corr <- round(cor(data_num), 1)
head(corr[, 1:26])
...

...{
  pressure, echo=FALSE}
dat2<- data_num %>% select('Days for shipment (scheduled)', 'Days for shipping (real)', 'Benefit per order', 'Sales per customer', 'Delivery Status', 'Order Item Discount Rate', 'Order Item Product Price', 'Order Item Profit Ratio', 'Order Item Quantity', 'Order Item Total', 'Order Profit Per Order', 'Product Price', Shipping_standard, Shipping_first,
```r
data(data_num)
corr <- round(cor(dat2), 1)
head(corr[, 1:12])
```

```r
ggcorrplot(corr,
  hc.order = TRUE,
  type = "lower",
  lab = TRUE)
```

```r
supply_type <- table(DataCoSupplyChainDataset_version_2$_Type$, DataCoSupplyChainDataset_version_2$_`Delivery Status$")
supply_type
chisq.test(supply_type)
```

```r
supply_mode <- table(DataCoSupplyChainDataset_version_2$_`Shipping Mode`, DataCoSupplyChainDataset_version_2$_`Delivery Status")
supply_mode
```
chisq.test(supply_mode)

```
```{r pressure, echo=FALSE}
supply_region <- table(DataCoSupplyChainDataset_version_2_$`Order Region',DataCoSupplyChainDataset_version_2_$`Delivery Status')
supply_region
chisq.test(supply_region)
```

```
```{r pressure, echo=FALSE}
ggplot(DataCoSupplyChainDataset_version_2_, aes(x=Type, y=DataCoSupplyChainDataset_version_2_$`Sales per customer')) + geom_boxplot(fill='green')
```

```
```{r pressure, echo=FALSE}
ggplot(DataCoSupplyChainDataset_version_2_, aes(x=DataCoSupplyChainDataset_version_2_$`Delivery Status', y=DataCoSupplyChainDataset_version_2_$`Sales per customer')) + geom_boxplot(fill='green')
```

```
```{r pressure, echo=FALSE}
ggplot(DataCoSupplyChainDataset_version_2_, aes(x=reorder(Type, Type, function(x)-length(x)))) + geom_bar(fill='blue') + labs(x='payment method')
```

```
```{r pressure, echo=FALSE}
ggplot(DataCoSupplyChainDataset_version_2_, aes(x=reorder(DataCoSupplyChainDataset_version_2_$`Shipping Mode',Type, function(x)-length(x)))) + geom_bar(fill='red') + labs(x='shipping mode')
```
```{r pressure, echo=FALSE}

```
```
```
```
```
```
library(purrr)
library(tidyr)
library(ggplot2)

sample = sample.split(as.data.frame(delaydata)[,1], SplitRatio = 1/5)
sample1 = subset(delaydata, sample == TRUE)
sample2 = subset(delaydata, sample == FALSE)
sample1  %>%
  keep(is.numeric) %>%
gather() %>%
ggplot(aes(value)) +
  facet_wrap(~ key, scales = "free") +
  geom_histogram()

```
```{r pressure, echo=FALSE}
library(ggmap)
key <- "AIzaSyBHgGu3oGbS4sC4_m3wWuFs8R4kGQX6A8"
ggmap::register_google(key=key, write = TRUE)
nyc_map <- get_map(
  location = c(lon = -75, lat = 40),
  maptype = "terrain", zoom = 5
)

map <- ggmmap(nyc_map) +
  geom_point(
    data = delaydata, aes(x = delaydata$Longitude, y = delaydata$Latitude),
    size = 0.4, alpha = 0.2, color = "red"

```
ggtitle("order distribution in the USA") +
theme(plot.title = element_text(hjust = 0.5)) +
xlab("Longitude") + ylab("Latitude")
map

Part2:
```{r pressure, echo=FALSE}
delaydata2<-book5
delaydata2$`Delivery Status`<- ifelse(delaydata2$`Delivery Status` == "Late delivery", 1, 0)
summary(delaydata)
table(delaydata$`Delivery Status`)
```
```{r pressure, echo=FALSE}
um_cols2 <- unlist(lapply(delaydata2, is.numeric))  # Identify numeric columns
num_cols2
```
```{r pressure, echo=FALSE}
data_num2 <- delaydata2[, num_cols2]  # Subset numeric columns of data
data_num2
summary(data_num2)
```
'dat55<- data_num2

select('Days for shipment (scheduled)',
     'Benefit per order',
     'Sales per customer',
     'Delivery Status',
     'Order Item Discount Rate',
     'Order Item Product Price',
     'Order Item Profit Ratio',
     'Order Item Quantity',
     'Order Item Total',
     'Order Profit Per Order',
     'Product Price', Shipping_standard, Shipping_first,
     Shipping_second, Shipping_same, region_WE, region_CA, region_SA)

...

set.seed(330)
sample = sample.split(as.data.frame(dat55)[,1],SplitRatio =2/3)

train = subset(dat55, sample == TRUE)

...
lm1 <- lm(`Delivery Status` ~ ., data=train)

summary(lm1)

```
```{r}

lm1.probs <- predict(lm1, newdata = test)
lm1.pred <- ifelse(lm1.probs>0.5, 1, 0)
lm1.pred

lm1.res <- table(test$`Delivery Status`, lm1.pred)
lm1.res

```
```{r}

confusionMatrix(lm1.res, mode="everything")

```
```{r}

MAE(test$`Delivery Status`, predict(lm1))

```
```{r}

svm1 <- svm(`Delivery Status`~ ., data=train)

svm1.probs <- predict(svm1, test)

```
```{r}
svm1.pred = ifelse(svm1.probs>0.5, 1, 0)

svm1.res <- table(test$`Delivery Status`, svm1.pred)
svm1.res
...

```{r}
confusionMatrix(svm1.res, mode="everything")
```
...

```{r}
MAE(test$`Delivery Status`, predict(svm1))
```
...

```{r}
log1 <- glm(`Delivery Status` ~., data=train, family=binomial)
summary(log1)
...

```{r}
log1.probs = predict(log1, newdata = test, type="response")
log1.pred = ifelse(log1.probs>0.5, 1, 0)

log1.res <- table(test$`Delivery Status`, log1.pred)
log1.res
summary(log1.res)
...
```{r}
confusionMatrix(log1.res, mode="everything")
```

```{r}
MAE(test$`Delivery Status`, predict(log1))
```

```{r}
naivebayes1 <- naiveBayes(`Delivery Status` ~ ., data = train, usekernel = T)
naive_pred <- predict(naivebayes1, newdata=test)

nb1 <- table(test$`Delivery Status`, naive_pred)

confusionMatrix(nb1, mode="everything")
```

```{r}
MAE(test$`Delivery Status`, predict(naivebayes1))
```

```{r}
# Decision Tree

decision1 <- rpart(`Delivery Status` ~ ., data=train)
decision1.probs <- predict(decision1, test)
decision1.pred = ifelse(decision1.probs>0.5, 1, 0)

decision1.res <- table(test$`Delivery Status`, decision1.pred)
decision1.res
```
confusionMatrix(decision1.res, mode="everything")

```{r pressure, echo=FALSE}
dat56 <- data_num2 %>%
  select(`Days for shipment (scheduled)`,
          `Benefit per order`,
          `Sales per customer`,
          `Delivery Status`,
          `Order Item Discount Rate`,
          `Order Item Product Price`,
          `Order Item Profit Ratio`,
          `Order Item Quantity`,
          `Order Item Total`,
          `Order Profit Per Order`,
          `Product Price`, Shipping_standard, Shipping_first,
          Shipping_second, Shipping_same, region_WE, region_CA, region_SA)
```
train56 = subset(dat56, sample == TRUE)

test56 = subset(dat56, sample == FALSE)

```
```{r}
lm1 <- lm(`Delivery Status` ~., data=train56)

summary(lm1)

```{r}
lm1.probs <- predict(lm1, newdata = test56)

lm1.pred <- ifelse(lm1.probs>0.5, 1, 0)

lm1.pred

lm1.res <- table(test56$`Delivery Status`, lm1.pred)

lm1.res

```{r}
confusionMatrix(lm1.res, mode="everything")

```{r}
MAE(test56$`Delivery Status`, predict(lm1))

```
log1 <- glm('Delivery Status' ~., data=train56, family=binomial)

summary(log1)

log1.probs = predict(log1, newdata = test56, type="response")
log1.pred = ifelse(log1.probs>0.5, 1, 0)

log1.res <- table(test56$'Delivery Status', log1.pred)
log1.res
summary(log1.res)

confusionMatrix(log1.res, mode="everything")

MAE(test56$'Delivery Status', predict(log1))

svm1 <- svm('Delivery Status'~ ., data=train56)
svm1.probs <- predict(svm1, test56)
```{r}
svm1.pred = ifelse(svm1.probs>0.5, 1, 0)

svm1.res <- table(test56$'Delivery Status', svm1.pred)
svm1.res

```{r}
confusionMatrix(svm1.res, mode="everything")

```{r}
MAE(test56$'Delivery Status', predict(svm1))

```{r}
naivebayes1 <- naiveBayes('Delivery Status' ~ ., data = train56, usekernel = T)
naive_pred <- predict(naivebayes1, newdata=test56)

nb1 <- table(test56$'Delivery Status', naive_pred)

confusionMatrix(nb1, mode="everything")

```{r}
MAE(test56$'Delivery Status', predict(naivebayes1))

```