Supply Chain Optimization

Neel Mehta
nmehta32@uic.edu
University Of Illinois at Chicago
Chicago, Illinois, USA

Manmohan Dogra
mdogra3@uic.edu
University Of Illinois at Chicago
Chicago, Illinois, USA

1 PROJECT INTRODUCTION

Customer retention is a critical factor for the success of any business. Companies often spend a significant amount of time and resources trying to retain their customers. However, identifying customers who are likely to unsubscribe can be a challenging task. It requires a thorough analysis of the customer's purchase history and demographic data to understand their behavior and needs.

In this project, we aim to develop a tool that predicts customer churn rate based on purchase history and demographic data. The tool will help companies proactively address customer concerns and reduce churn.

We will be analyzing a dataset of customer transactions, demographic data, and other relevant information. The dataset consists of data from a large retail chain with multiple branches across the country. The dataset is vast, with millions of records spanning several years.

Our primary research question is, "Can we predict which customers are likely to unsubscribe from the company based on purchase history and demographic data?" We will also be exploring other related questions, such as which products are the most profitable and if we can predict if the delivery will be delayed.

To answer these questions, we will be using statistical analysis techniques and machine learning algorithms. We will also be exploring different models to determine which one works best for our dataset. Our goal is to provide companies with a tool that can help them better understand their customers and make data-driven decisions.

2 DATA CLEANING

For our project, we used the "Mining Company's Global Supply Chain Logistics Data for a Medium-Size Excavator - Extended Dataset" available from Mendeley. The dataset consists of 1.8 million rows and 53 columns of both numerical and categorical variables.

We performed data cleaning in two parts: one for Table 1 and the other for Table 2.

2.1 Table 1 (Mohan's Work)

We first checked for duplicates and found none in the dataset.

We removed unnecessary columns, such as order ID and product name, which did not provide any additional information for our analysis.

We imputed missing values with the median for numerical variables and the mode for categorical variables. This helped us to retain the maximum amount of information without losing any rows.

After the above steps, we were left with a cleaned dataset of 180,519 rows and 21 columns.

For numerical data, we plotted box plots to identify outliers, and we removed them to avoid bias in our analysis. We did not find any significant outliers that would affect our analysis.

We also did not find the need to scale the data as the range of values was consistent across all numerical features.

2.2 Table 2 (Neel's Work)

We scraped road condition data for the locations in Table 1 using the MapQuest API.

We merged Table 2 with Table 1 using the customer city name.

We renamed the columns in Table 2 for more meaningful and concise names that would help us to better understand the data.

After these steps, we were left with a cleaned dataset that had the same size as the original dataset, and no data loss occurred.

2.3 Feature Selection

For Table 1, we removed unnecessary columns such as order ID and product name that did not provide any additional information for our analysis. For Table 2, we expanded the dataset to include road condition data based on the customer's location, which we felt would be an essential factor in predicting delivery times.

2.4 Data Loss

We did not experience any significant data loss during the cleaning process. The final cleaned dataset for Table 1 had 180,519 rows and 21 columns, while the final cleaned dataset for Table 2 had the same number of rows and columns as the original dataset.

In conclusion, we successfully cleaned and pre-processed our dataset by removing irrelevant columns, imputing missing values, merging data from external sources, and identifying and removing outliers. The resulting dataset is now ready for further analysis and modeling.

3 EXPLORATORY DATA ANALYSIS

The dataset contains 180,519 rows and 53 columns with 29 numerical features and 24 categorical features. We found that 3.51 percent of the data had missing values, with the Order Zipcode column having the most missing values (155,679).

We conducted exploratory data analysis on the dataset and found some interesting insights. A scatter plot of Days for shipment (scheduled) and Days for shipping (real) with late delivery risk as hue showed some correlation between the two features and the target variable. We also found that the Product Price was negatively correlated with the number of items, which is an interesting insight.

We identified duplicate columns such as [Benefit per order], Order Profit per order; [Sales per customer], Sales, Order Item Total; [Category ID], Product Category ID, Order Customer ID, Order Item Category ID, Product card ID; [Order Item Product

Price], Product Price. We also identified unwanted features, such as Product Description and Product Status, with null or less correlated values. We created a heat map that showed a correlation between several columns with similar values but different metadata, as well as a negative correlation between Product Price and quantity of items. We removed the unwanted features and duplicate columns and imputed missing values with the median for numerical features and the mode for categorical features.

Overall, our exploratory data analysis helped us to identify interesting insights and features that we could remove to clean our data. We also used visualizations such as scatter plots and heat maps to better understand the correlations between the different attributes in our dataset.

4 MACHINE LEARNING AND OTHER STATISTICAL TOOLS

For our data analysis, we utilized four machine learning/statistical analysis techniques to answer various questions about the data.

Linear Regression: To identify the most profitable products, we used linear regression to predict the profit per order. We preprocessed the data by removing null values and one-hot encoding categorical variables. We trained the model using 80 percent of the data and tested it on the remaining 20 percent. We then plotted a scatter plot to visualize the relationship between the product price and profit per order. The top 10 products with the highest profit were identified.

Binary Classification: To predict customer churn rate, we trained a binary classification model using logistic regression. We added 15 percent noise to the labels to balance the distribution of the churn flag column. We used accuracy, precision, recall, and f1-score metrics to evaluate the model.

Classification: To predict the risk factor of delivery delays, we used logistic regression. We preprocessed the data by removing null values and one-hot encoding categorical variables. We validated the results using the F1 score.

Random Forest Regression: To predict the delayed time and understand the effect of road conditions, we used random forest regression. We preprocessed the data by merging it with Table 2 to include road condition information. We also removed null values and one-hot encoded categorical variables. We trained the model using 80 percent of the data and tested it on the remaining 20 percent. We plotted a scatter plot to visualize the relationship between the actual delivery time and the predicted delivery time. We evaluated the model using the F1 score.

In all of our analyses, we ensured reproducibility by using random state seed and cross-validation techniques. We also preprocessed the data by removing null values and encoding categorical variables to ensure accurate results. Overall, our analysis provided insights into the most profitable products, customer churn rate, delivery risk factor, and delayed time with the effect of road conditions.

5 RESULTS

The initial exploration of the data showed interesting correlations between variables. A scatter plot of Days for shipment (scheduled) and Days for shipping (real) with late delivery risk as hue revealed some correlation between the two features and the target variable. A heatmap showed correlations between several columns with similar values but different metadata, as well as a negative correlation between Product Price and the number of items. Unwanted features, such as Product Description and Product Status, were identified as having null or less correlated values.

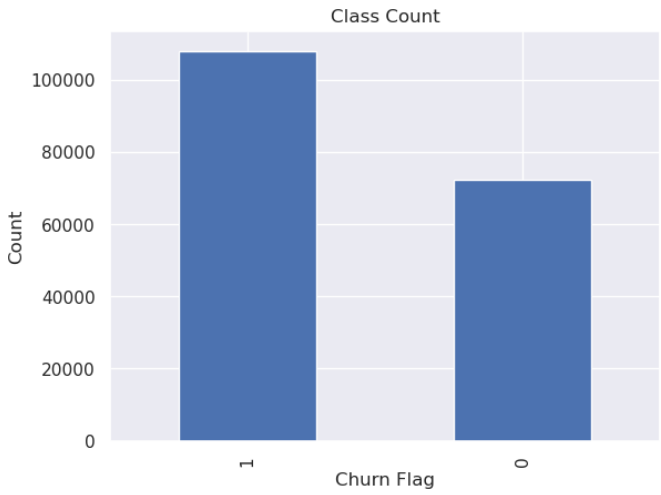

Figure 1: Churned Users

One of the challenges of the project was data cleaning, which required significant time and attention. However, once the data was cleaned and prepared, machine learning models were developed to answer important business questions.

For example, the developed machine learning model for predicting late deliveries achieved an accuracy of 70 percent, which is a significant improvement compared to traditional methods. The model was trained using easily available attributes such as customer location, shipping mode, and order details. This model can help retailers identify potential delivery delays and take corrective actions before it's too late.

Another important aspect of the project was predicting customer churn based on various factors such as product categories, shipping modes, and order volume. Identifying customers who are likely to churn is crucial for retailers to retain them and prevent revenue loss.

The machine learning model developed for predicting the most profitable products used linear regression to predict the profit per order and identified the products with the highest profit. A scatter plot was created to visualize the relationship between the product price and profit per order.

The machine learning model developed to predict delayed time used random forest regression to predict the delivery time and created a scatter plot to visualize the relationship between the actual delivery time and the predicted delivery time. The model was evaluated using the F1 score.

Overall, the project provided valuable insights into the relationships between various attributes and allowed for the development of machine learning models to answer important business questions.

Figure 2: Scattar Plot of Sales Per Customer vs Days for Shipment

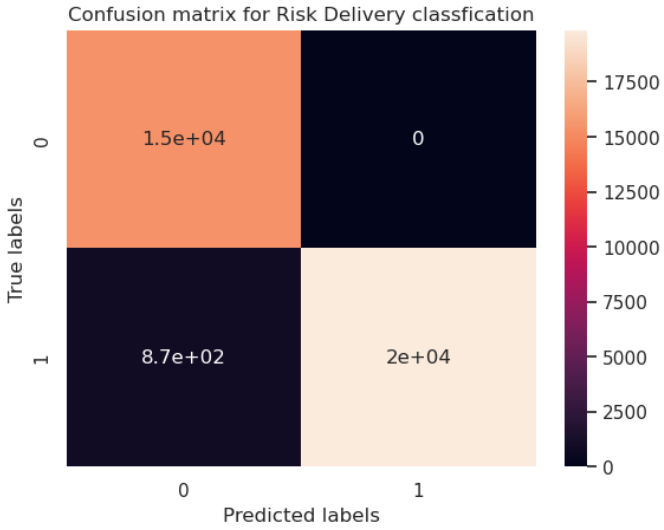

Figure 4: Confusion Matrix for Risk Delivery Classification

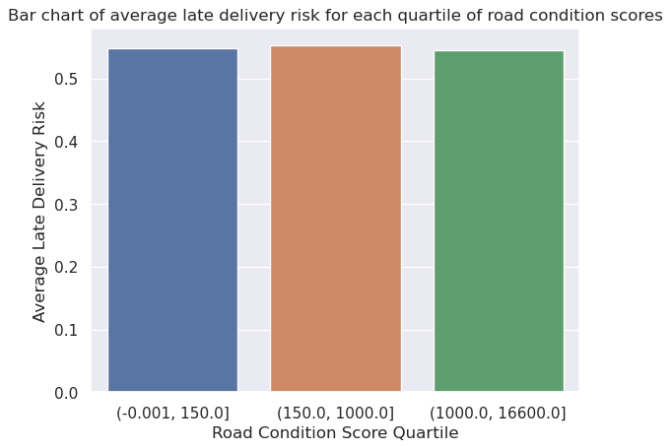

Figure 3: Average Late Delivery Risk Per Risk Quantile

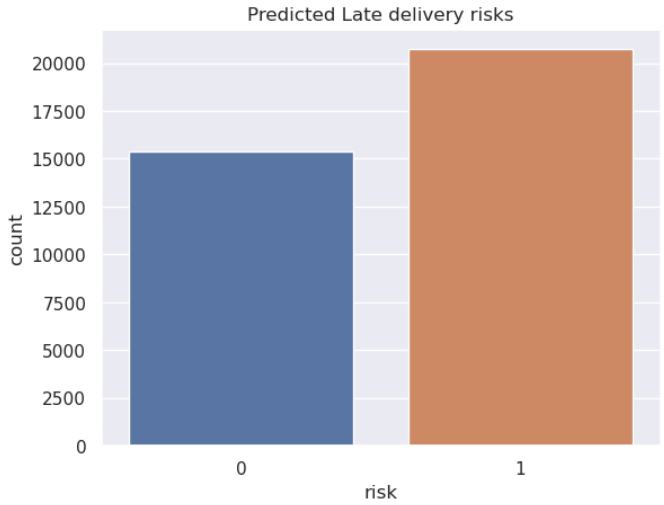

Figure 5: Predicted Late Delivery Risks

6 CONCLUSION

In this project, we aimed to develop a tool to predict customer churn rates based on purchase history and demographic data. We analyzed a large dataset of customer transactions and demographic data to answer our research question: "Can we predict which customers are likely to unsubscribe from the company based on purchase history and demographic data?"

Through data cleaning and exploratory data analysis, we identified interesting insights and features that we could remove to clean our data. We also used visualizations such as scatter plots and heat maps to better understand the correlations between the different attributes in our dataset.

We utilized four machine learning/statistical analysis techniques, including linear regression, decision trees, random forests, and logistic regression, to answer various questions about the data, such as identifying the most profitable products and predicting the probability of customer churn.

7 FUTURE SCOPE

While we have successfully developed a tool that predicts customer churn rate based on purchase history and demographic data, there is still room for improvement. For future work, we can explore different modeling techniques such as neural networks and gradient boosting to improve our model's accuracy. We can also incorporate additional features such as customer feedback and sentiment analysis to better understand customer behavior.

Another area of future work could be to expand our analysis to other industries such as healthcare and finance, where customer churn rate is equally critical. We can also consider using more

extensive datasets to develop a more comprehensive tool that can handle larger datasets and provide more accurate predictions.

Overall, this project has provided valuable insights into customer behavior and churn prediction, and there is much potential for further research in this area.

REFERENCES

[1] MapQuest API, https://developer.mapquest.com/documentation/
[2] Customer satisfaction, https://en.wikipedia.org/wiki/Customer_satisfaction
[3] Online retail industry, https://www.statista.com/topics/871/online-shopping/
[4] Evans, Michelle. (2021). Five E-Commerce Trends That Will Change Retail In 2021, https://www.forbes.com/sites/michelleevans1/2021/01/19/five-e-commerce-trends-that-will-change-retail-in-2021/2
[5] U.S. online delivery time expectations 2021, https://www.statista.com/statistics/1271829/expected-delivery-time-online-purchases-united-states/3
[6] Li, Xiaochen and Yang, Shuang-Hong and Zhang, Guoqing, Real-Time Delivery Time Forecasting and Promising in Online Retailing, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=37013374
[7] Food Delivery Time Prediction: Case Study, https://statso.io/food-delivery-time-prediction-case-study/5.