University of New Haven
TAGLIATELA COLLEGE OF ENGINEERING
Electrical & Computer Engineering and Computer Science

Course: Introduction to Data Science (DSCI-6002-01)
Major: Master of Science in Data Science
Department: Electrical & Computer Engineering & Computer Science (ECECS)

FINAL PROJECT
TECHNICAL REPORT

Fall 22

CONTENTS

Executive Summary

The project aims at analyzing our target business from e-commerce supply chain industry, facing high customer churn due to late deliveries and comparing and deploying multiple classification machine learning models for late deliveries predictions

Team Members:
Sahil Sehgal
Yegneshwar Rao Ginjupalli
Bello Salisu
Sai Kumar Gude

Questions?
Contact: ssehg1@unh.newhaven.edu

Abstract

Customer satisfaction is the bread and butter of ecommerce businesses. It is the repeat business that is most beneficial to e-commerce businesses and leads to maximum profit since there are no new acquisition costs involved in repeat business. Thus, prevention of customer churn, which is directly linked to customer satisfaction, is a key driver of a supply chain company's profitability.

Our target business, the DataCo company, is an ecommerce supply chain business, where customer segmentation analysis was performed in this project which helps the company to better understand its customers and target them to increase customer responsiveness and the company's revenue.

The business has two major concerns in its business operations:

1. It is facing high customer churn of unsatisfied customers due to late deliveries
2. It is facing losses due to fraudulent transactions

Some Questions to ask from the Data is:

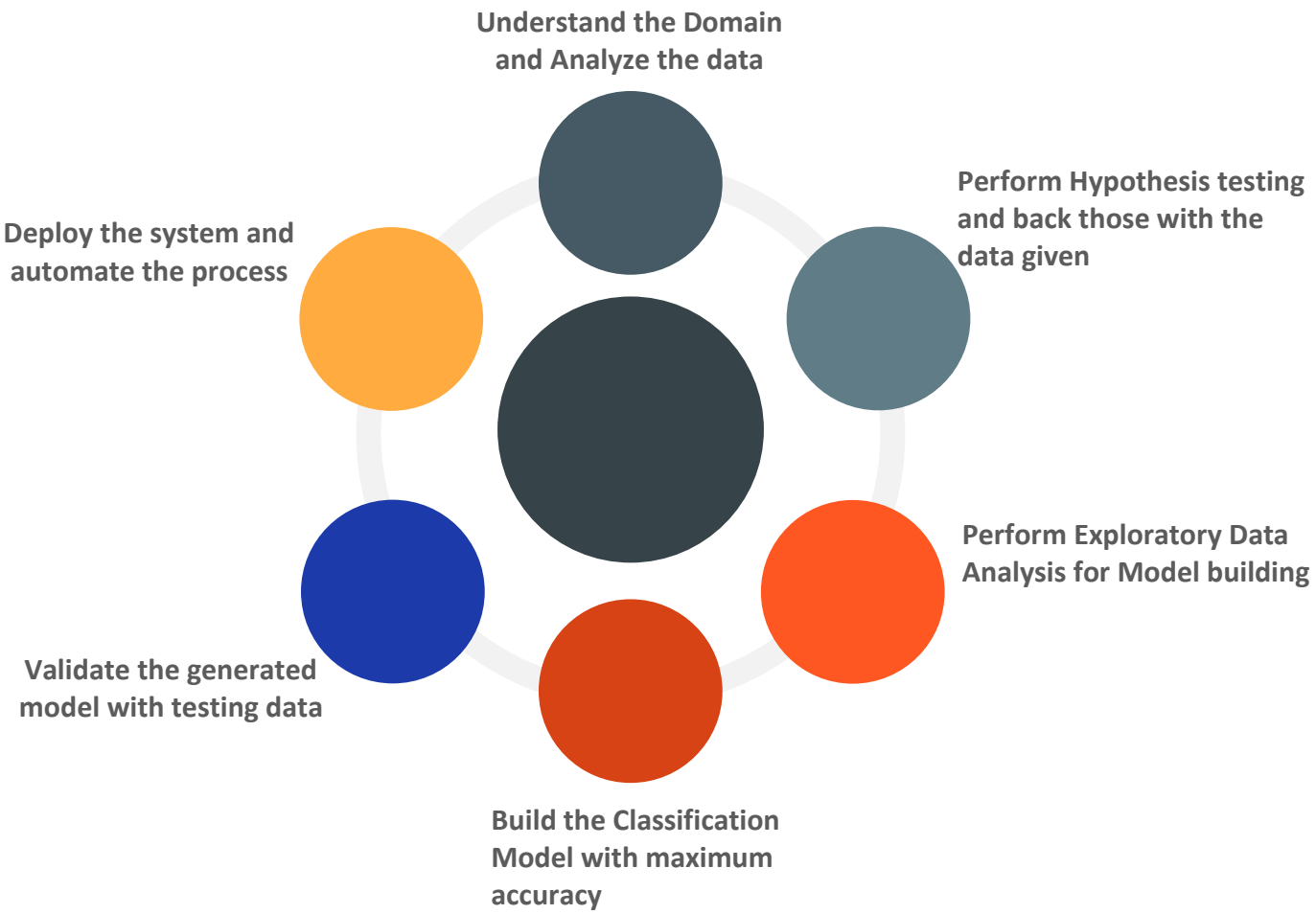
- Any insights on their business in terms of their product catalog and customer base
- Which customers to target immediately to reduce churn and boost sales
- Do any patterns emerge from their customer base and transactional data that can point to patterns in late deliveries and fraudulent transactions.
- A prediction model that can predict late deliveries sing the order/sale data before they occur.

The analysis in this notebook is aimed at providing solutions to the above problems and presents the following analysis:

- Exploratory data analysis to detect trends in sales, product pricing and segments, markets and regions, and insights related to late delivery and fraudulent transactions
- Customer segmentation analysis using RFM technique providing a systematic approach to customer loyalty programs.
- A reliable machine Learning model that the company can deploy to detect late deliveries and facilitate customer satisfaction.

Methodology

Following CRISP-DM methodology for Data Science Process



- From the Operations perspective if we can analyze the factors affecting the deliveries of the orders and the transactions, we can predict the future delivery and order status. We need to work on the below mention steps
- Research around Supply Chain sector and dig deep into logistics domain to understand the pipeline of an order (every chain involved from seller to buyer)
- Analyze the dataset to understand the relationship between different features affecting the delivery of the order and also causing suspicion to the transaction

Methodology

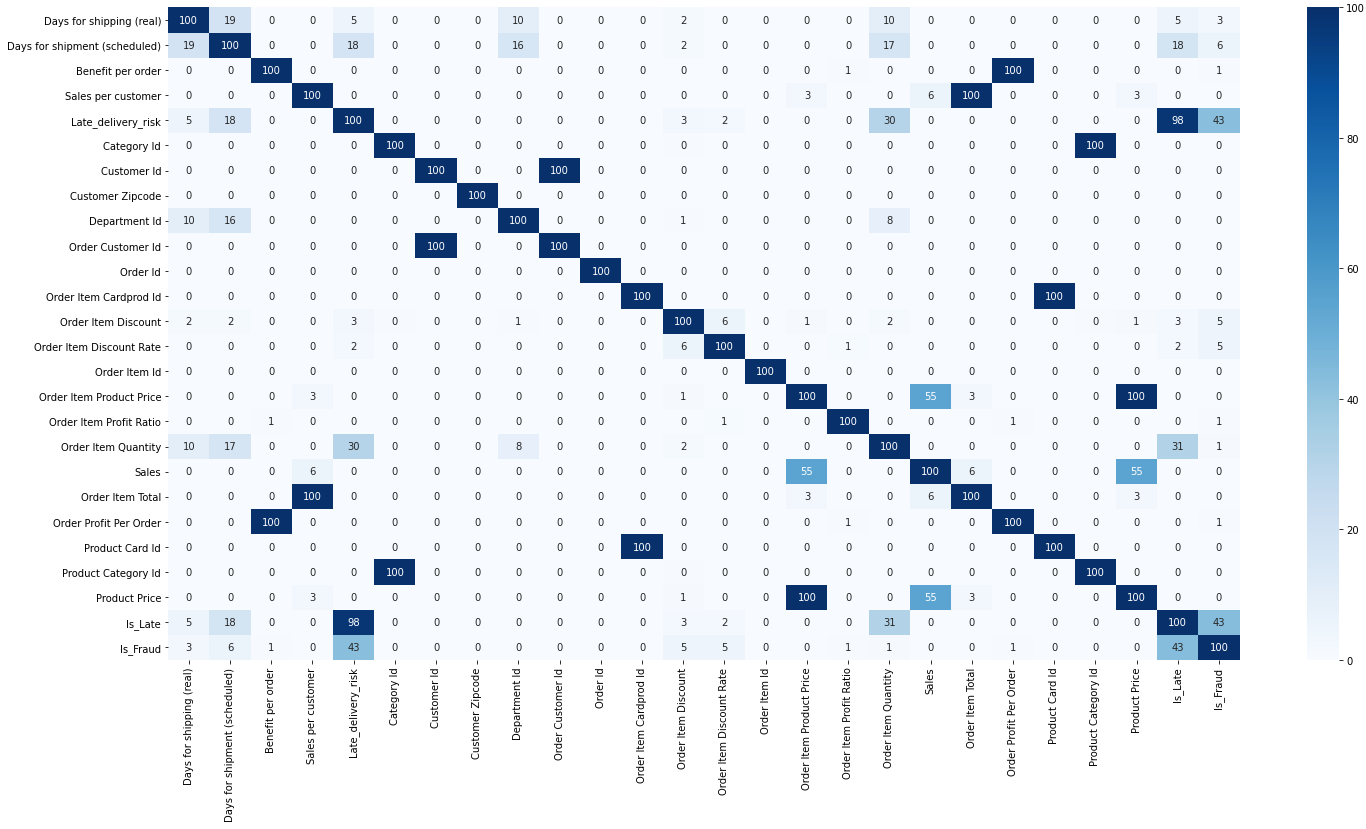
- Identify factors like payment modes, shipping methods, order and customer regions, sales, order quantity, days for shipment etc. to back any hypothesis
- To find any improvement/implementations in the overall logistics pipeline which can prevent the above challenges
- At last, build three classification models and compare them based on different hyper parameters (user selection) to classify any order (given order details) to fall under on-time delivery or late delivery and any order to fall under fraudulent transaction or valid transaction

Analysis

Data Wrangling

Analyzed and removed duplicate columns

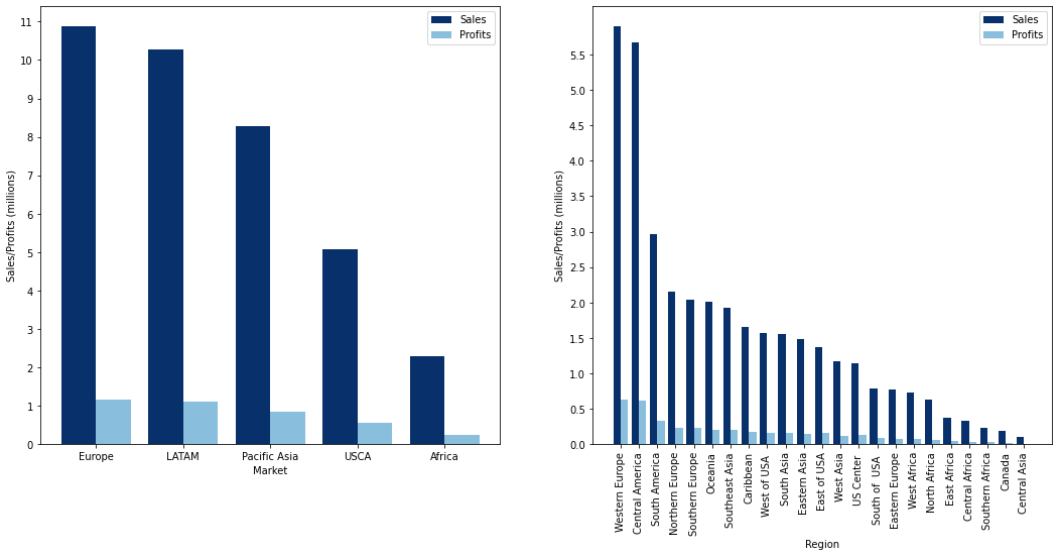
```
In [8]: # Comparison function for columns. Returns dataframe with percentage of values that differ
def remove_duplicate_cols(df):
    df_num = df.select_dtypes(include='number')
    dup_df = pd.DataFrame(columns=df_num.columns)
    for col_row in df_num.columns:
        for col_col in df_num.columns:
            dup_df.loc[col_row, col_col] = (df_num[col_row] == df_num[col_col]).sum() * 100 / len(df_num)
    dup_df = dup_df.astype('float')
    return dup_df

# Plot heatmap with duplicate percentages and remove columns with 100
cleanded_data = remove_duplicate_cols(data)
plt.figure(figsize=(24,12))
sns.heatmap(cleanded_data, annot=True, fmt='.0f', cmap='Blues', edgecolors='black')
plt.show()
```



Analysis

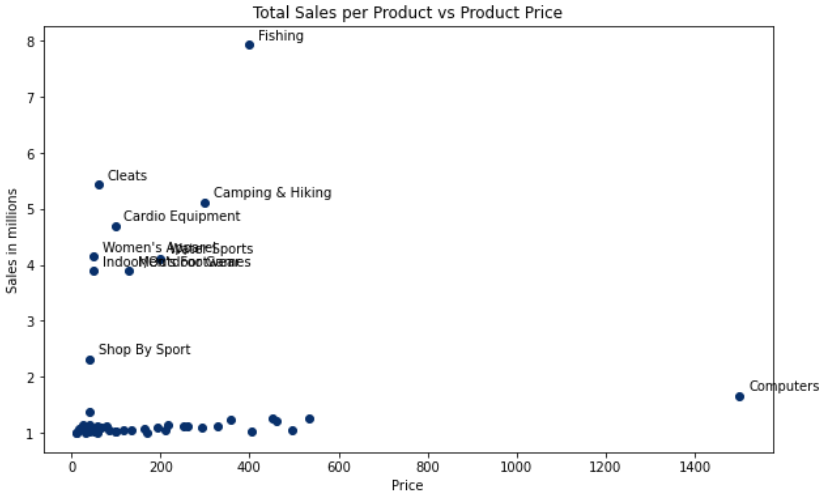
Data Visualization

Sales by Region



Note: It can be seen from the graph that European market has the greatest number of sales followed by Latin America whereas Africa has the least. In these markets western Europe regions and central America recorded highest sales.

Sales by Product Price

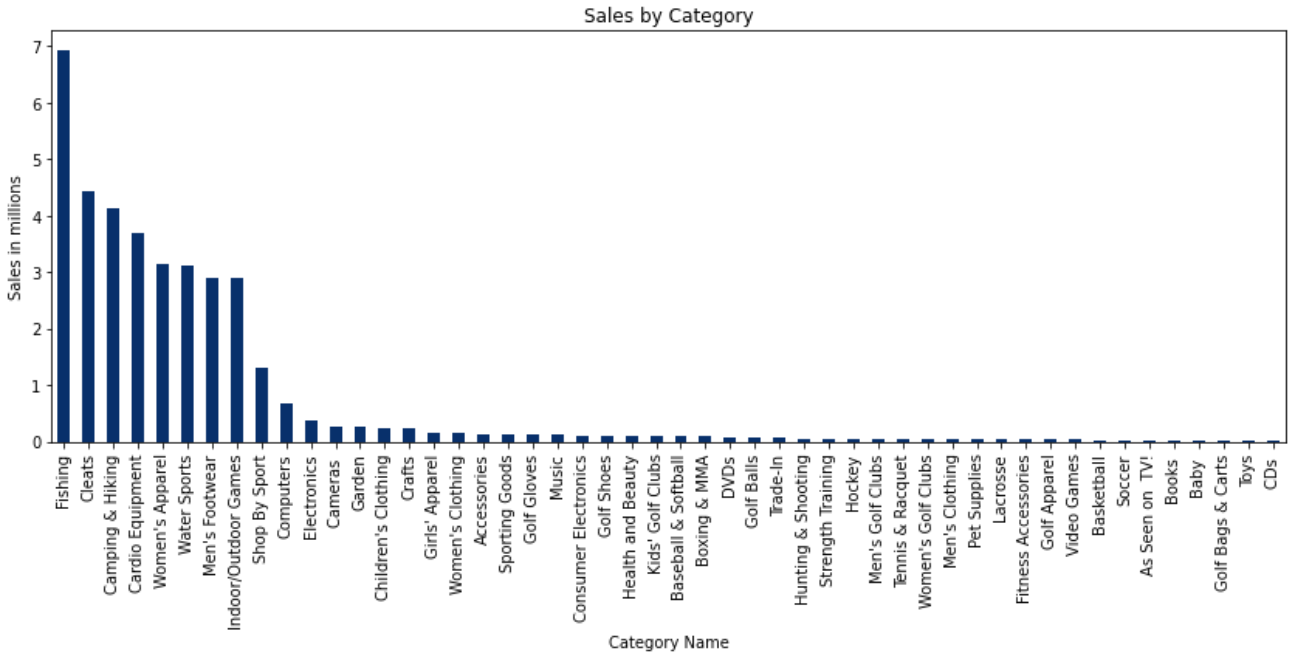


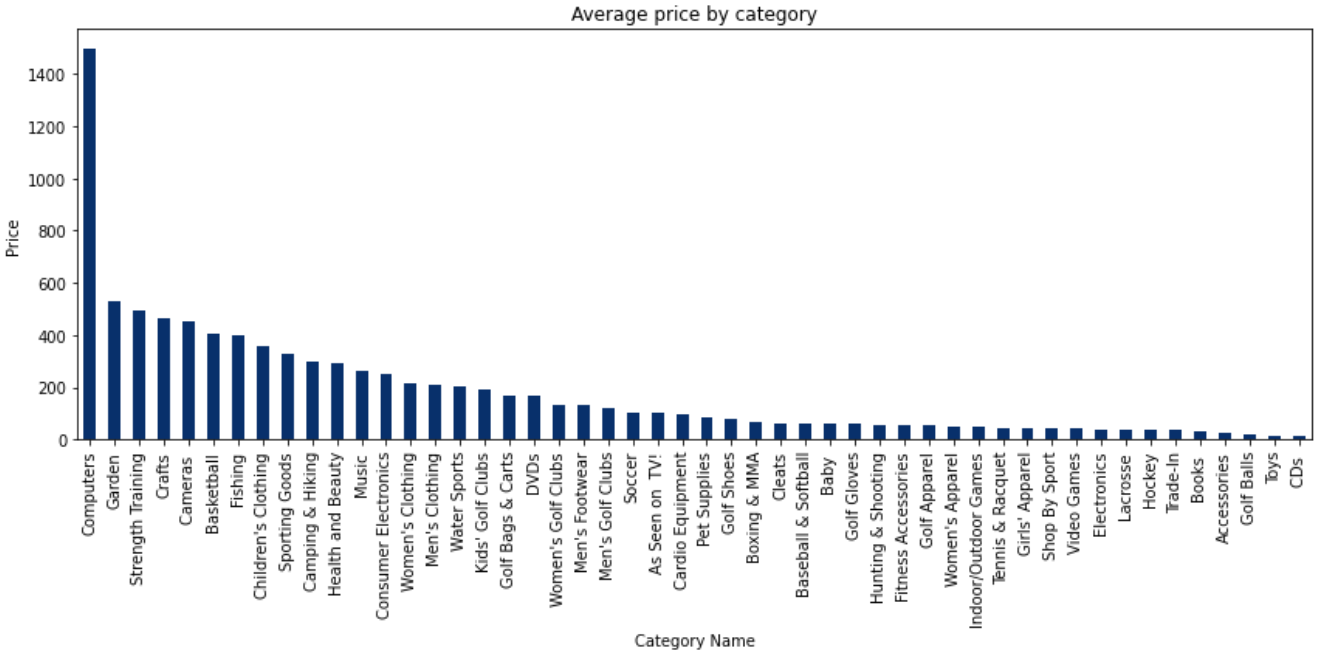
We can observe some overlap between the top products in terms of total sales and average price. Although the top selling products seem to be driven by consumer demand than anything else and no causation exists between the product price and product sales.

We can say that demand is the primary factor driving the sales of a product category.

Analysis

Sales by Product Category

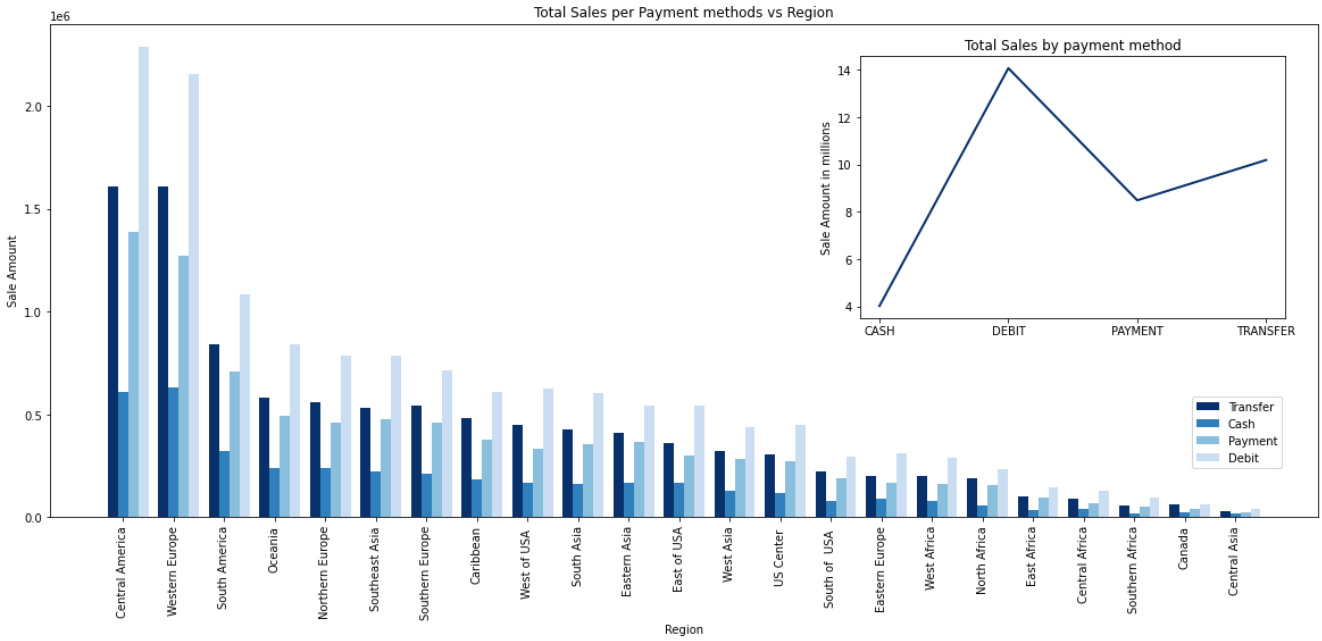




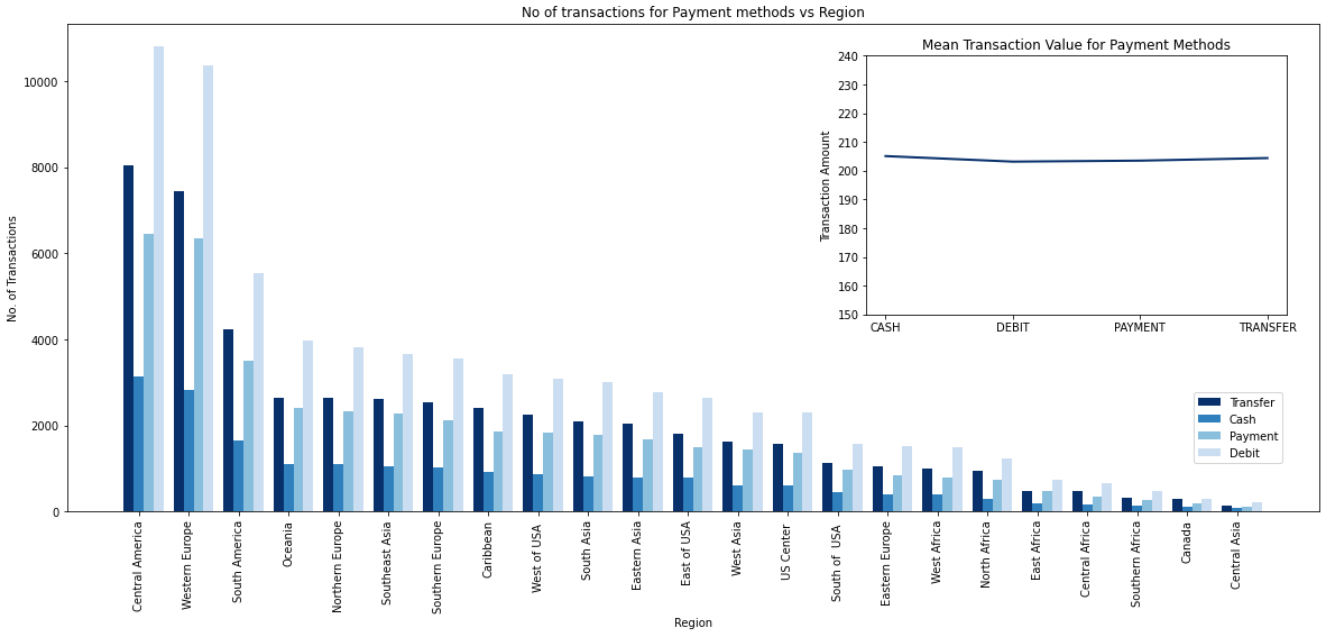
As we can see from the above figure, Fishing category had the greatest number of sales followed by the Cleats. However, it is surprising to see that top 7 products with highest price on average are the most sold products on average with computers having almost 1350 sales despite price being 1500$.

Analysis

Sales per Payment Method



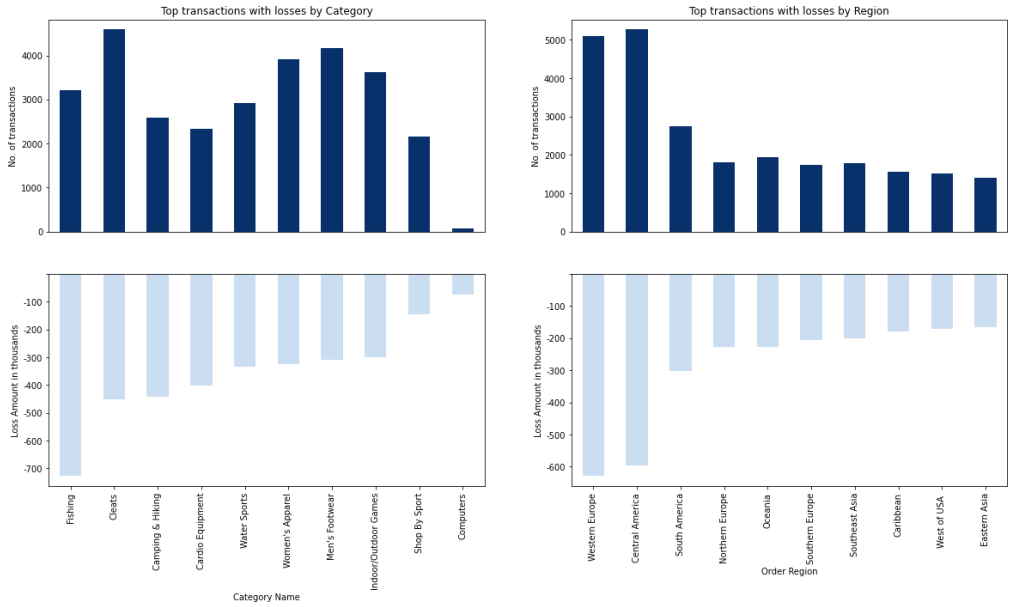
Number of Transactions per payment method



Debit type is most preferred payment method by people throughout all regions, Cash payment being the least preferred method. This is reflected in both the number of transactions and the transaction amount.
We also observe that the mean amount for each transaction type is nearly the same (~ 205$)

Analysis

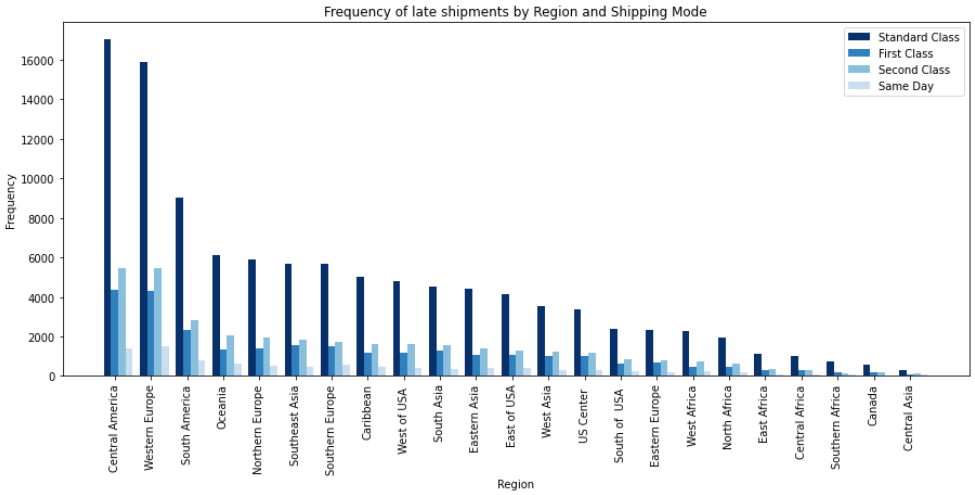
Fraudulent Transactions Analysis



Total losses recorded: -3883547

The total loss sales are approximately 3.9 Million which is a huge amount. It can be seen that Cleats is the category with maximum frequency of loss generating transactions followed by Men's footwear. Fishing records the highest amount lost in loss generating transactions. Most lost sales are happening in Western Europe & Central America region. This lost sale may have happened due to suspected frauds or late deliveries.

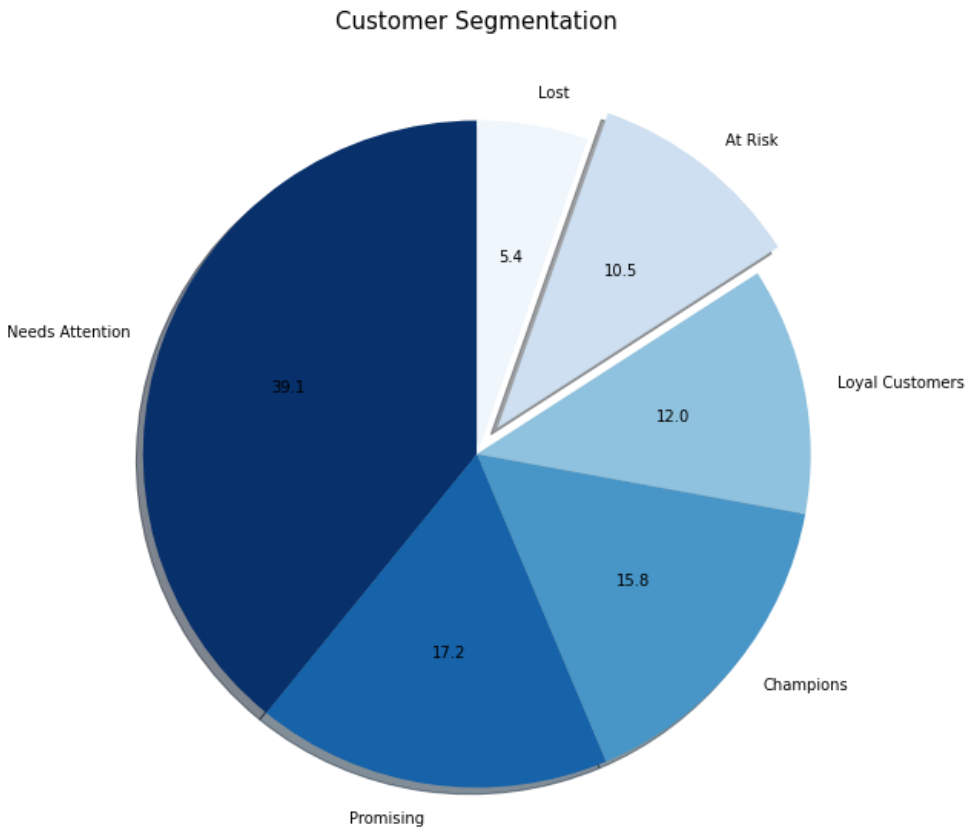
Late Delivery Analysis



It can be observed that highest number of suspected fraud orders are from Western Europe which is approximately 17.4% of total orders followed by Central America with 15.5%. Which product is being suspected fraud the most?

Analysis

Customer Segmentation (RFM Analysis)



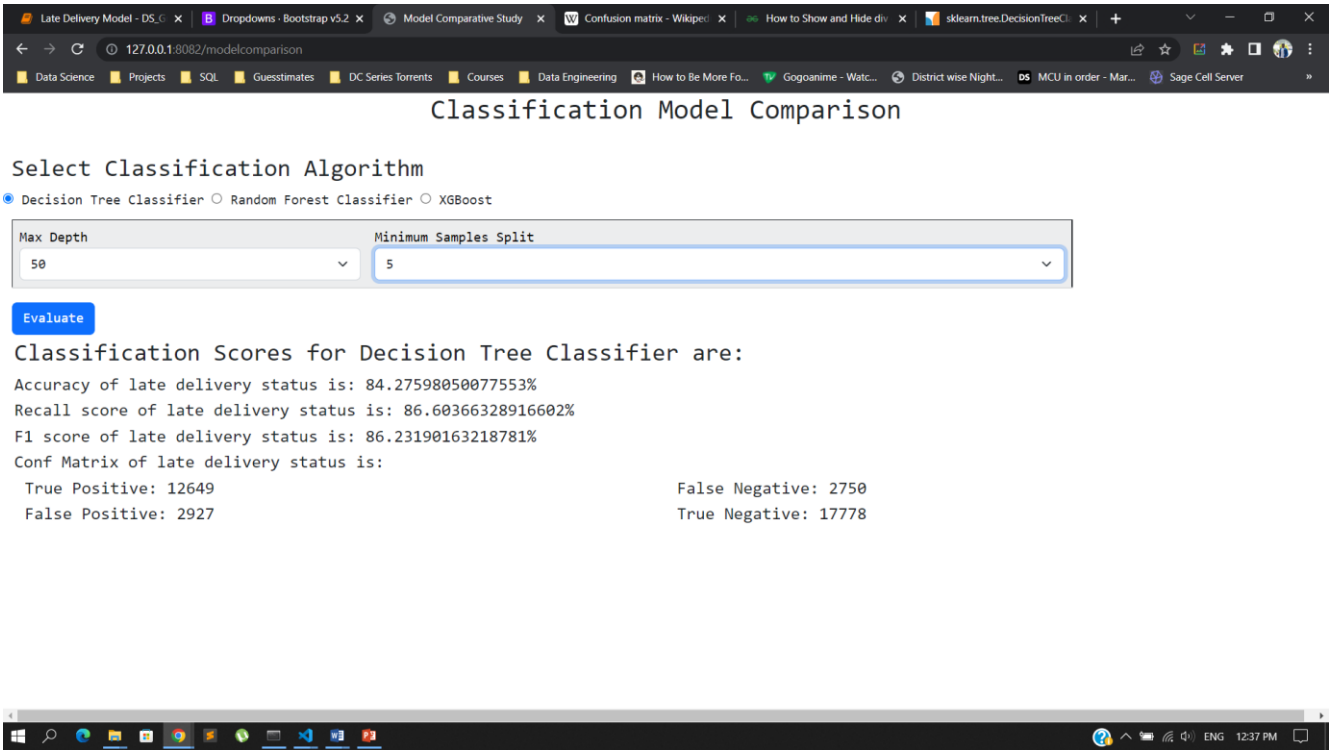
Based on the RFM segmentation technique, we see that nearly 10.5% of the customers are at risk of churn and nearly 39% need attention to be converted into promising customers. This can be done via promotional offers and discounts.

We see that 5.4% of the customers are already lost.

Model Deployment

Decision Tree Classifier

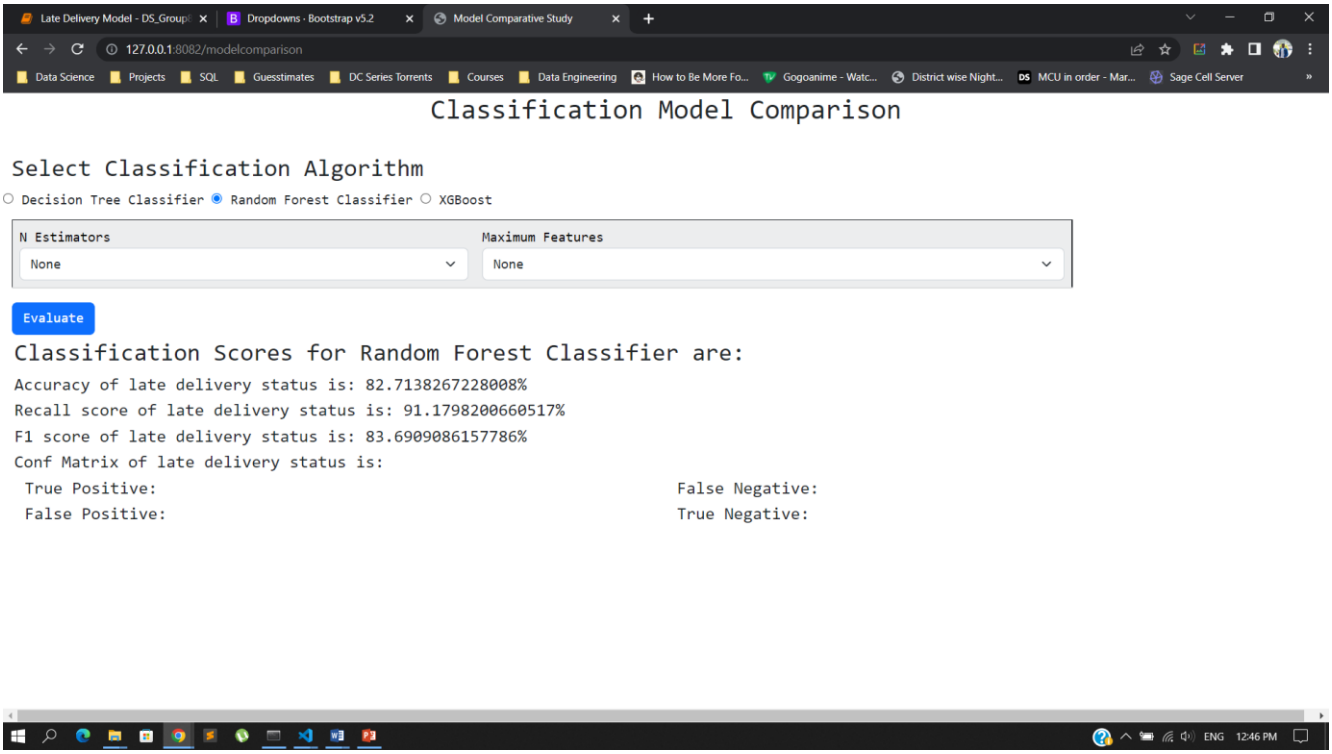
Deployed Decision Tree Classifier with hyper parameters Max Depth of 50 and Sample Split of 5 (flexible for user to choose) to get the Classification Report below



Model Deployment

Random Forest Classifier

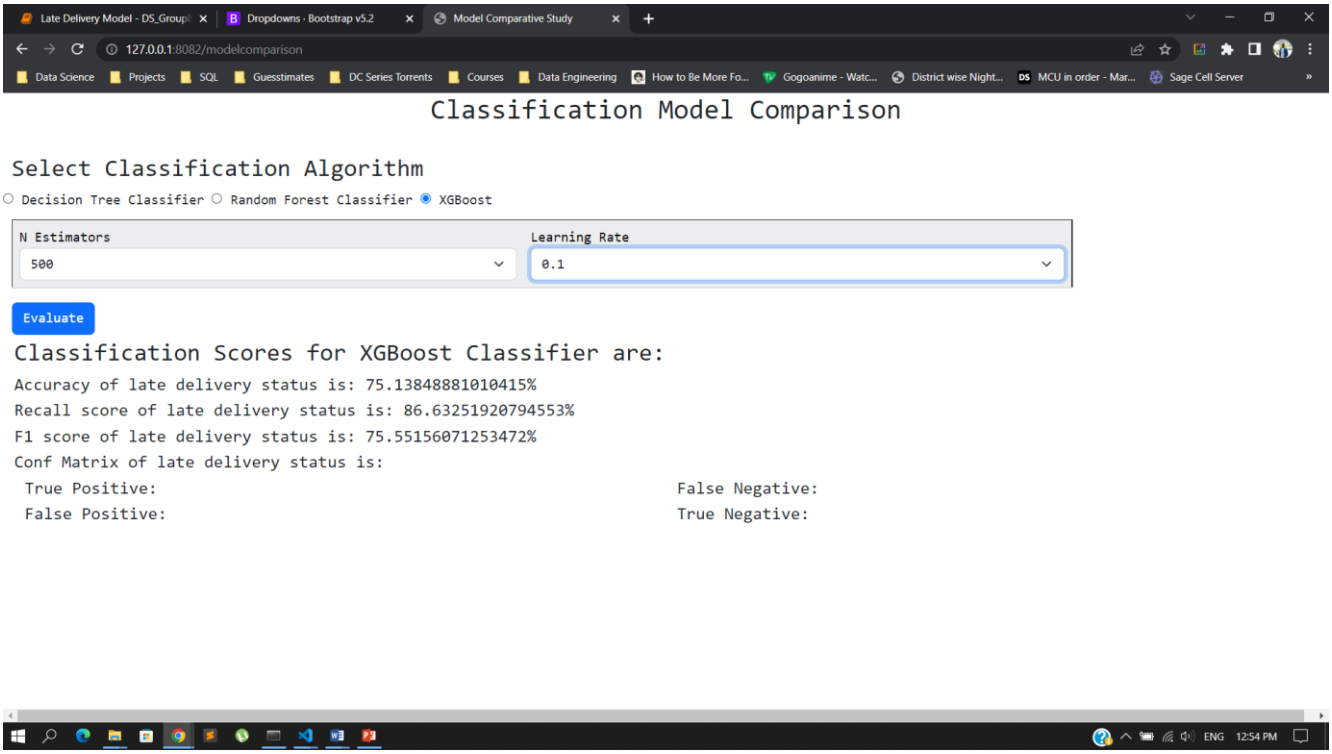
Deployed Random Forest Classifier using No hyper parameters (flexible for user to choose) to get the Classification Report below



Model Deployment

XGBoost Classifier

Deployed XGBoost Classifier with hyper parameters Estimators (500) and a Learning rate of 0.1 (flexible for user to choose) to get the Classification Report below



Conclusion

Conclusion

After analyzing the DataCo Company dataset we discovered that the highest sales were derived from the Western Europe and Central America regions. The frequency of late deliveries and fraudulent transaction were also proportionate with the frequency of sales by region, making Western Europe and Central America leaders in these categories too.

Sales: The total sales for the company were consistent and on the uptick until the 2017 Q3 following which the sales suddenly dipped by almost 65% in 2018 Q1. On average, July had the most sales in terms of monetary value while the profits peaked in the month of September.

Payments: Most customers preferred payments through debit cards and all fraud transactions were reported with wire transfer mode of payment. The company needs to set up checks and balanced to avoid these fraudulent transactions as we could see that the company was scammed with more than 100k by a single customer.

Customer Segmentation: The RFM technique deployed clearly shows a lack of customer retention techniques with almost 50% of the consumers falling under the 'At Risk' and 'Needs Attention' categories. This segmentation can help guide a comprehensive customer loyalty program which can help reduce the consumer churn significantly.

Logistics: Product categories: Cleats, Men's Footwear, and Women's Apparel lead in late deliveries. The supply chain of these products needs to be better optimized to tackle this.

Models: When compared with other classification machine learning models, the **Decision Tree** model did a good job of identifying orders with later delivery with an f1 score of **86.3%**. When we tuned the **Random Forest**, it showed a poor prediction accuracy for late deliveries with an F1 score of nearly **83%**. We had to limit the extent of hyper parameter tuning due to the computation power requirements. Although these models can be tuned further. Looking at the feature importance's from the tree-based models, it is evident that the 'Shipping Mode', 'Days for shipment (scheduled)' and 'order_hour' have the maximum weight in predicting the binary response of our target variable (late or not). This was extremely evident in the **XGBoost** model where the model chose to assign negligible importance's to all other features except the three listed above. The company can take multiple steps to improve on these three features. (Feature importance is shown in jupyter notebook)

Contributions/References

Dataset: https://data.mendeley.com/datasets/8gx2fvg2k6/5

Kaggle Notebook Reference:

https://www.kaggle.com/code/jaswanthhbadvelu/comparison-of-ml-models-with-rnn#Data-Modelling