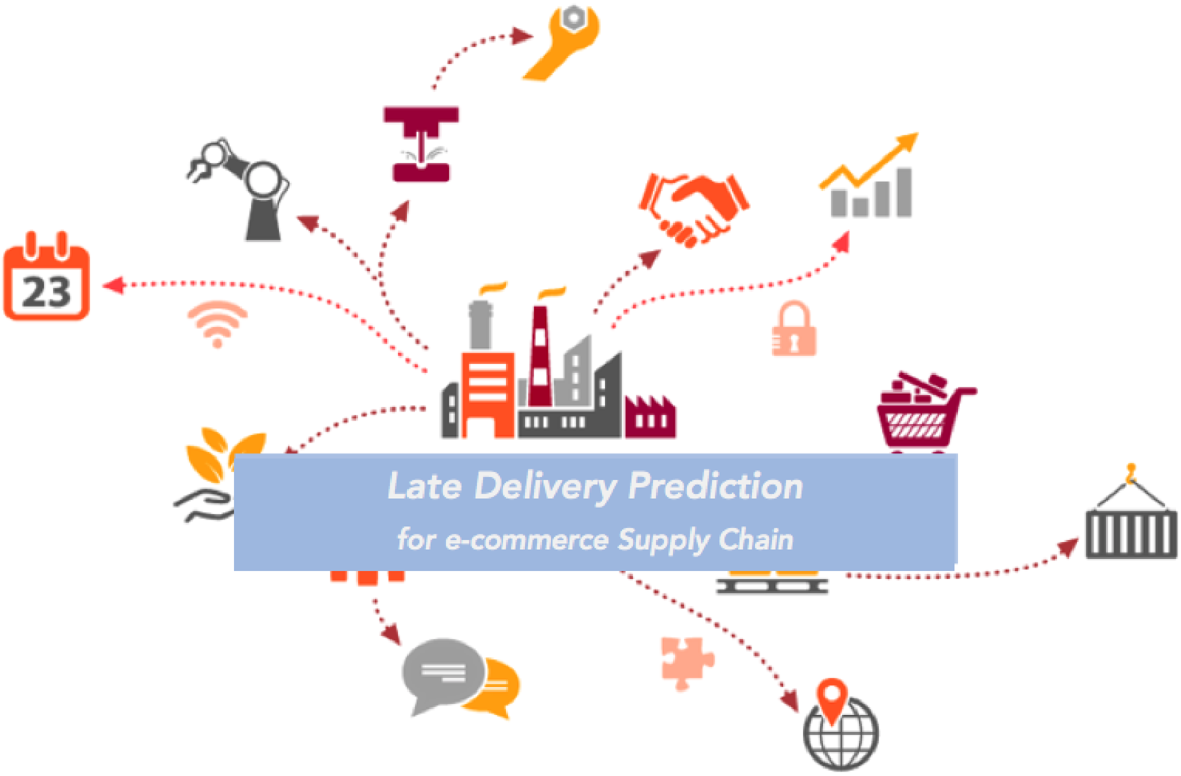


Capstone 2

Late Delivery Prediction

For e-commerce supply chain

Final Report

1. The Problem

1.1. The Problem Statement

Our target business is an e-commerce supply chain business. The business has two major concerns in its business operations:

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

The business has decided to consult a data analytics firm (us) and wants to figure out:

- Any insights they can find on their business in terms of their product catalog and customer base.
- Which customers can they 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 using the order/sale data before they occur.

1.2. The Opportunity (Benefit to the customer)

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.

Detecting late deliveries before they occur can greatly benefit the supply chain business by rerouting and taking proactive steps to ensure on-time delivery and can also set appropriate customer expectations in-terms of delivery time so that customers feel dissatisfied.

Fraudulent transactions directly impact the company's profitability. This is simply money lost for the organization and it is in the best interest of the organization to curtail fraudulent transactions before they occur.

1.3. The Solution

In order to solve the problems that our customer is facing, we will cover three major steps in our analysis:

1. Exploratory data analysis to detect trends in sales, product pricing and segments, markets and regions, and insights related to late delivery and fraudulent transactions
2. Customer segmentation analysis using RFM technique providing a systematic approach to customer loyalty programs.

3. A reliable machine Learning model that the company can deploy to detect late deliveries and improve customer satisfaction.

We will explore various industry standard classification algorithms, perform hyper-parameter tuning and compare their performance before choosing the best one for our model. We will also look at the feature importance to gain insight into our model.

2. The Dataset

The data used is the transactions of all sales by product from January 2015 to January 2018 for a reputed ecommerce business. The data is published by DataCo and can be downloaded from :

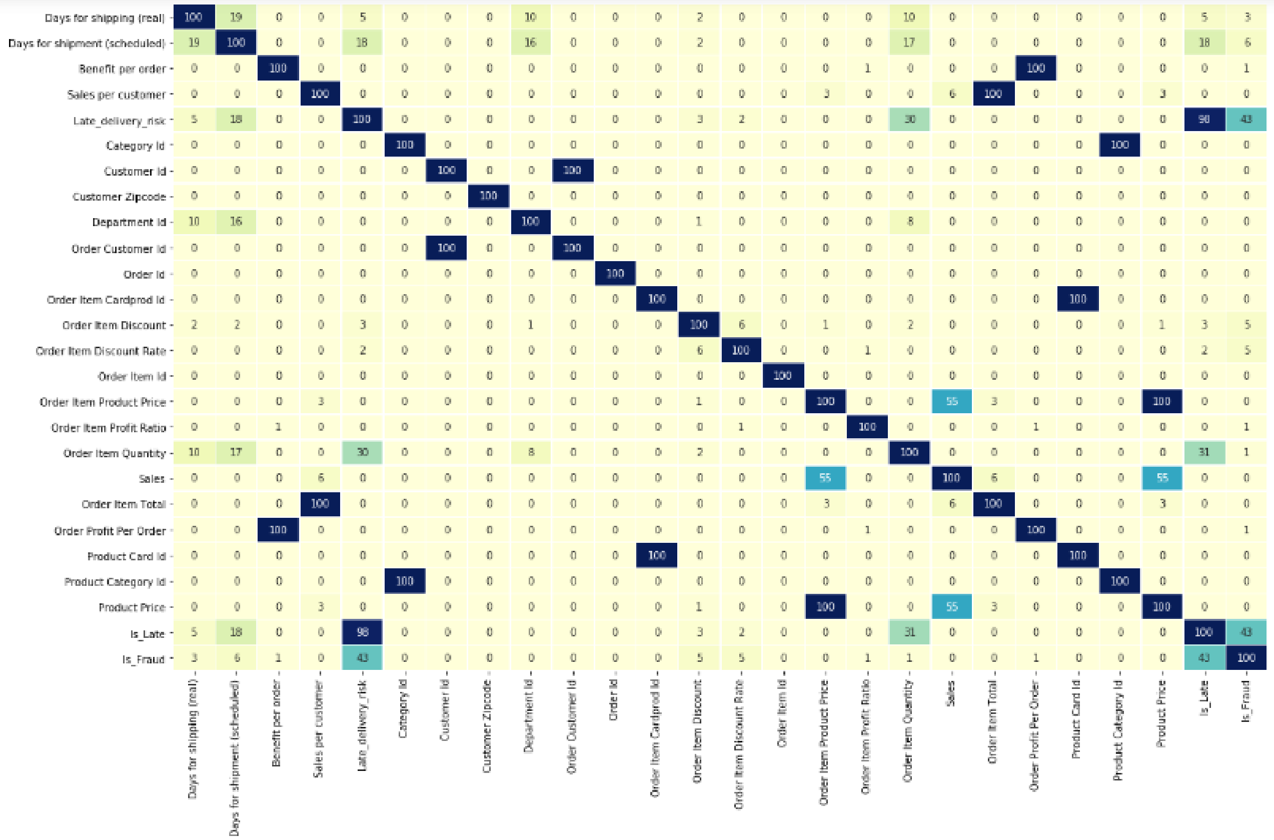
https://data.mendeley.com/datasets/8gx2fvg2k6/5

3. Data Wrangling

The data set was relatively clean with only 4 out of 45 columns having any null values.

The following steps were performed to prepare the dataset for use:

1. Columns with large missing values were dropped - Order Zipcode & Product Description
2. Unnecessary columns with useless information were dropped such as Customer Password, Customer Street, Product Image etc.
3. First Name and Last Name were combined to remove repetitiveness.
4. It was suspected that there are multiple features with duplicate values. To confirm and address this, a heat map was made of comparison of all columns. 100% matching columns were removed to deal with duplicates .



Heat map for column similarity

5. Two new columns - Is_Late & Is_Fraud were created as target variables for the machine learning algorithms
6. Standardization of the data set was also performed before implementing machine learning algorithms.

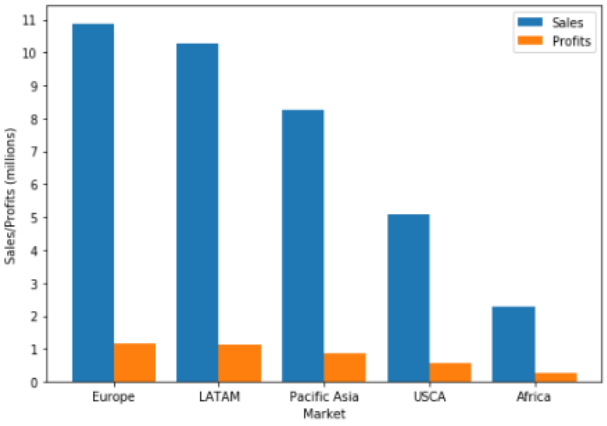
4. Exploratory Data Analysis

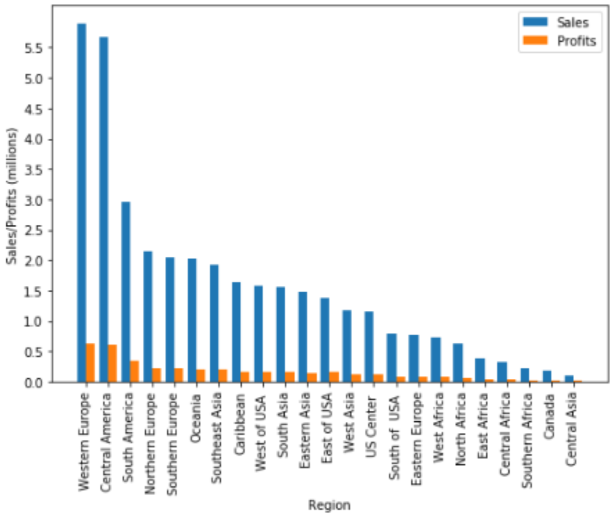
Since the number of feature variables is large and since we can approach the company's problems from multiple angles, an in-depth EDA was done on the data set. The goal was to detect trends in sales, product pricing and segments, markets and regions, and insights related to late delivery and fraudulent transactions.

A customer segmentation analysis was also performed using RFM technique to get a glimpse of the company's problem of customer churn and to provide a systematic approach to customer retention using discounts and loyalty programs.

Sales by Continent and Sales by Region:

Let's look at how different geographical locations compare in terms of sales and profits.

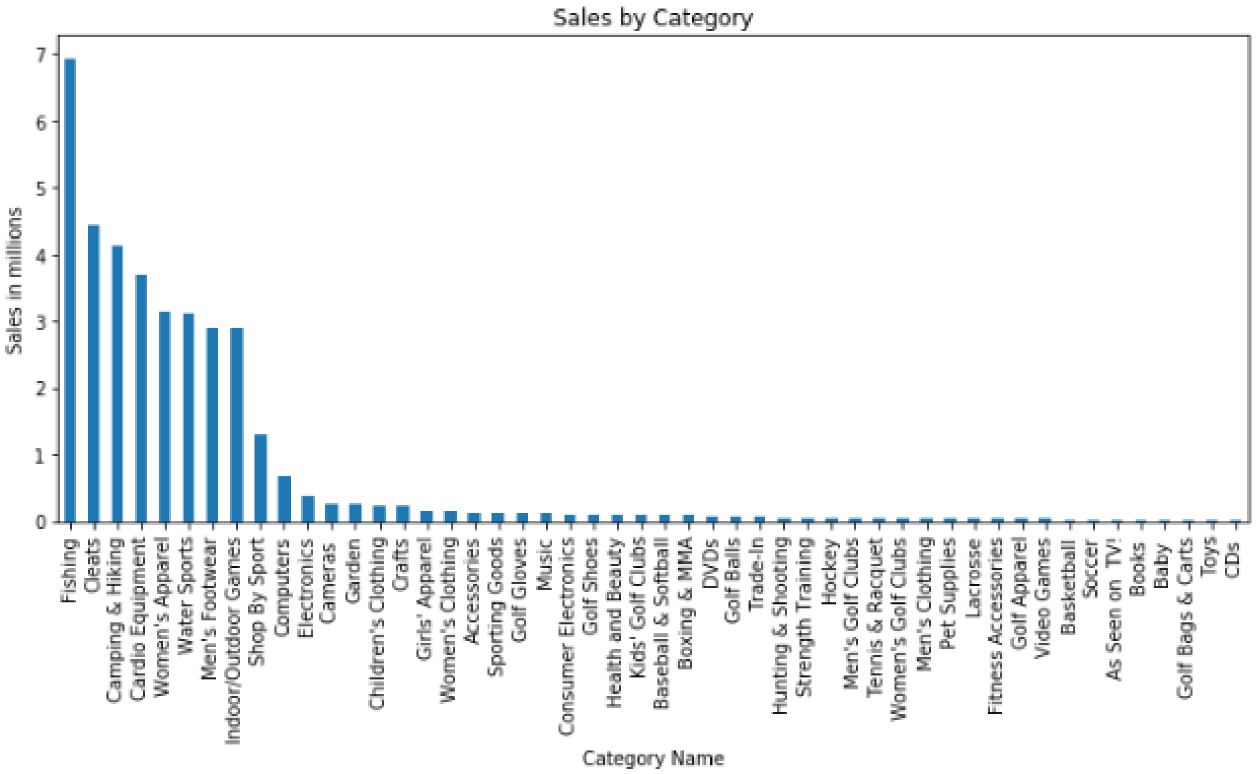




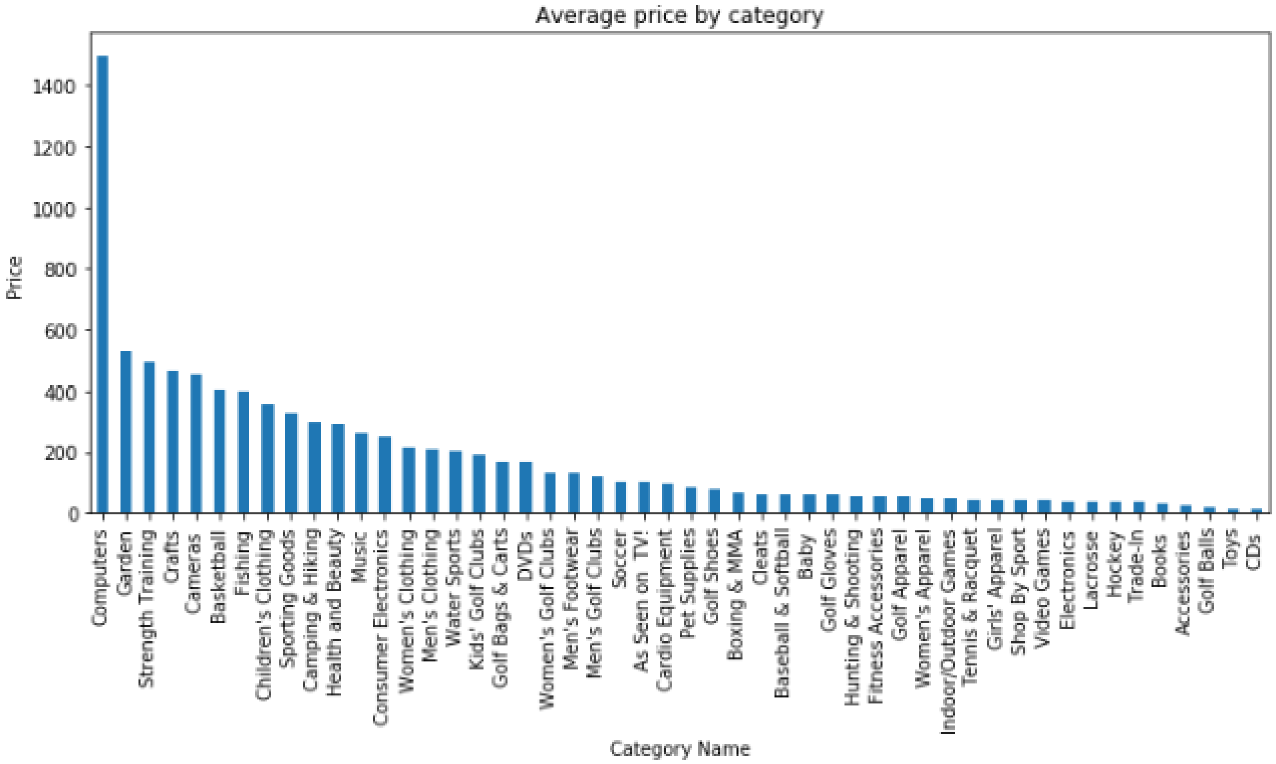
European market has the most sales followed by Latin America whereas Africa has the least. In these markets Western Europe and Central America recorded the highest sales.

Sales by Product Category:

Next we look at the total sales in different product categories.

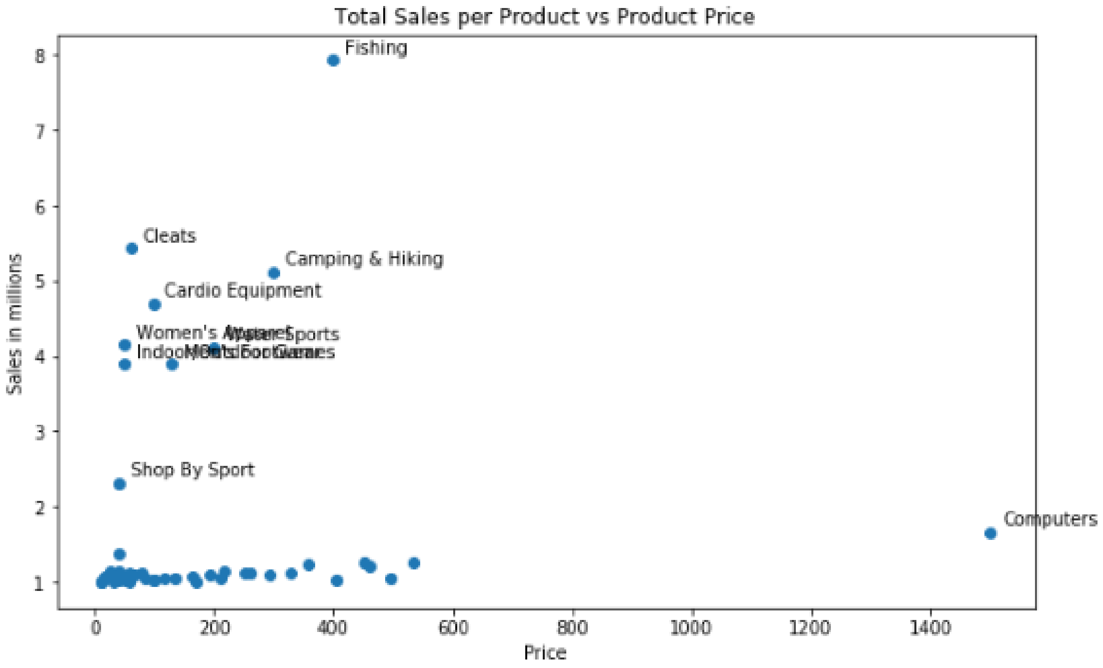


We also take a look at the average product prices.

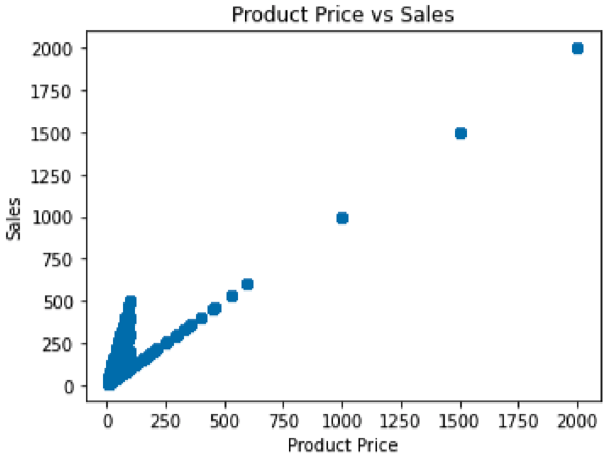


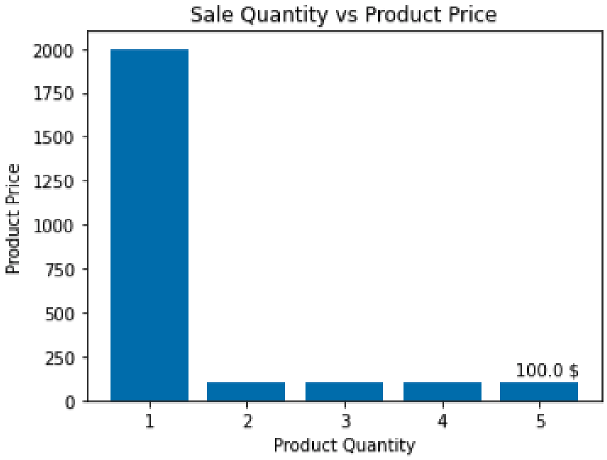
Looks like Fishing and Cleats top the sales by product category while computers have the maximum mean price among all categories.

Is there any correlation between the average price and sales of a product category?



There doesn't seem to be any direct correlation between product prices and their total sales. Frequency of purchase is driven by demand. Lets see if we can spot any relationships between purchase quantity and product price.

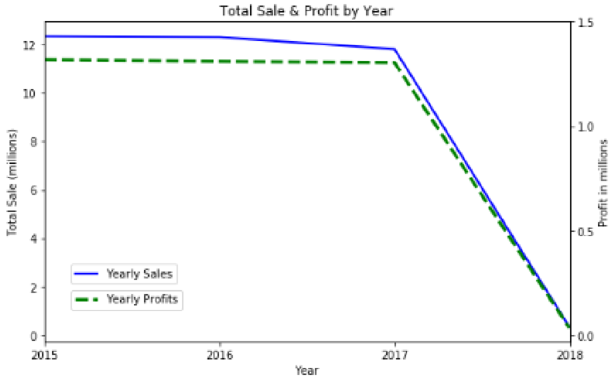


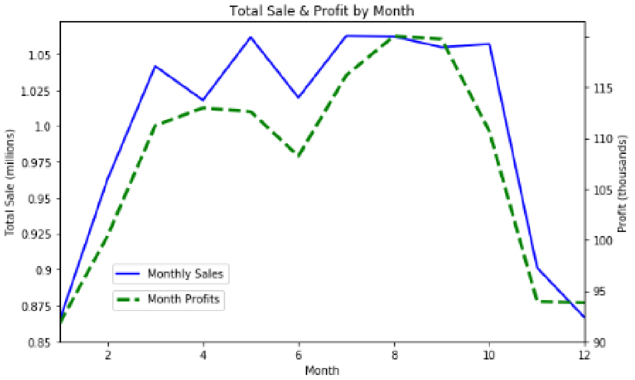


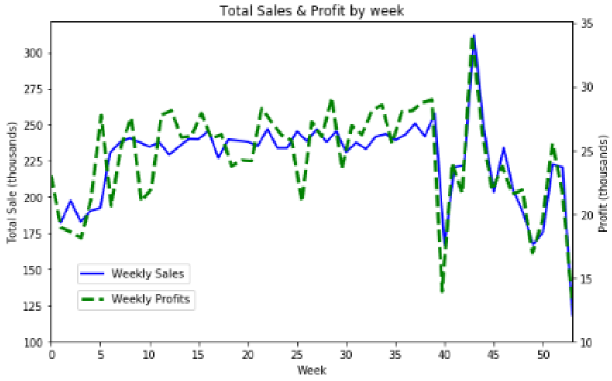
We can observe something interesting about the buying pattern of the customers looking at the maximum product price vs product quantity purchased. From the figure on the right we can see that customers bought more than 1 nos. (2, 3, 4 & 5) of only those items which are within the 100$ price brand. Hence we see that the slope of the scatter plot is steeper within 100$ for the plot on the left. This is an interesting insight in terms of consumer behaviour and can be used to pivot or boost sales on the platform.

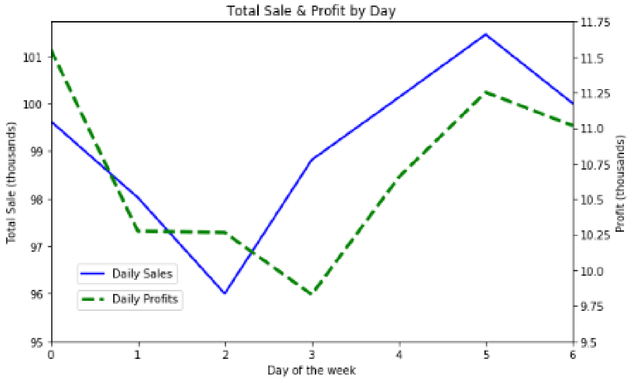
Sales and Profits vs Time Periods:

Let's look at the sales through different time periods. We look at the total sales summed over year (fig 1) and the yearly average of total sales summed over month, week and day (figures 2, 3, and 4).









From the yearly sales graph, we observe that sales were consistent from 2015 until 2017 but suddenly dipped in 2018. This could have resulted from a hindrance of operations due to reasons such as financial shortages. It is worth examining what the cause of this dip is.

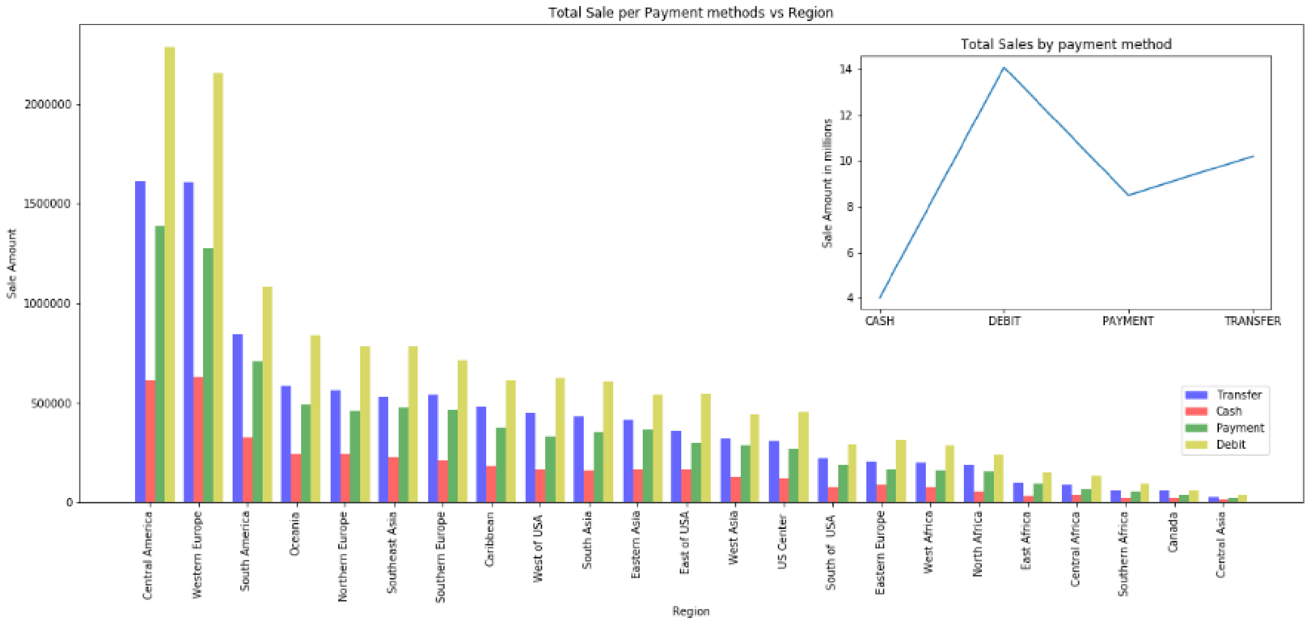
From the total monthly sales averaged over the years, we see that maximum sales happen from March till October. The months of January and December have the least sales. As far as the profits are concerned we see an increment in the 2nd half of the year, typically picking up after June, which is the worst month for profits in between March and October.

Looking at the weekly sales, it reflects what we have gathered from the monthly sales i.e. the sales dip during the start and end of the year. We see consistent sales till week 39 after which we see large fluctuations from week 40 to 51. This is most probably reflective of the consumer behaviour in and around the festive weeks during the last few months of the year.

The daily Sales show clearly that weekends have the highest sales and profits. Tuesday has the lowest sale and Wednesday has the lowest profits. This could be due to discounts offered.

Analysis of Payment methods vs Sale amount and Regions:

Next, let's look at the sales done via different payment methods over different regions.

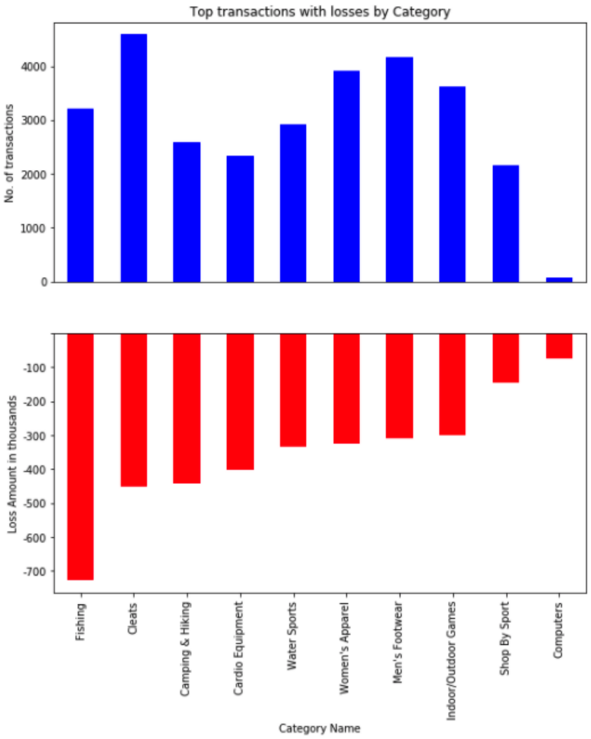


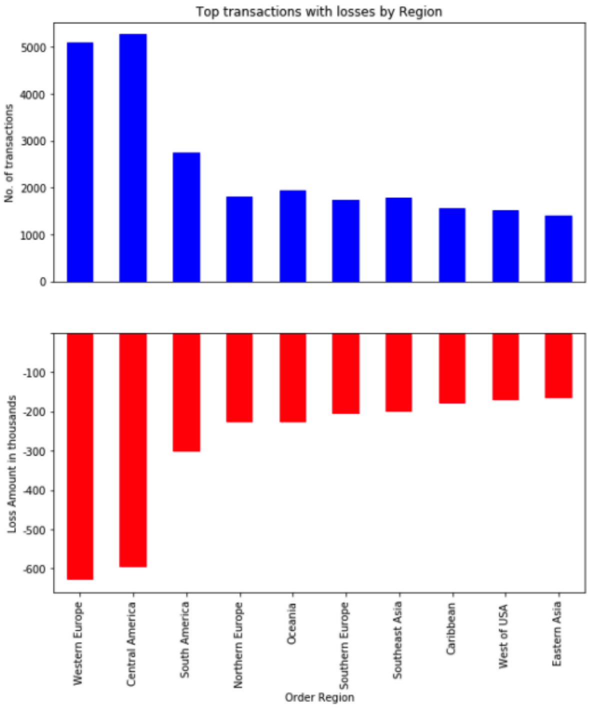
Debit type is the most preferred payment method by people throughout all regions, cash payment being the least preferred method. This is reflected in both the number of transactions and the transaction amount.

It is quite surprising to see all regions showing the same proportion of sales in terms of transaction type.

Exploration of Loss Generating Transactions:

Let's look at the product category wise distribution of loss making transactions. Below we look at the top 5 categories that are responsible for the maximum amount and the frequency of loss making transactions.





The total loss sales are approximately 3.9 Millions which is a huge amount.

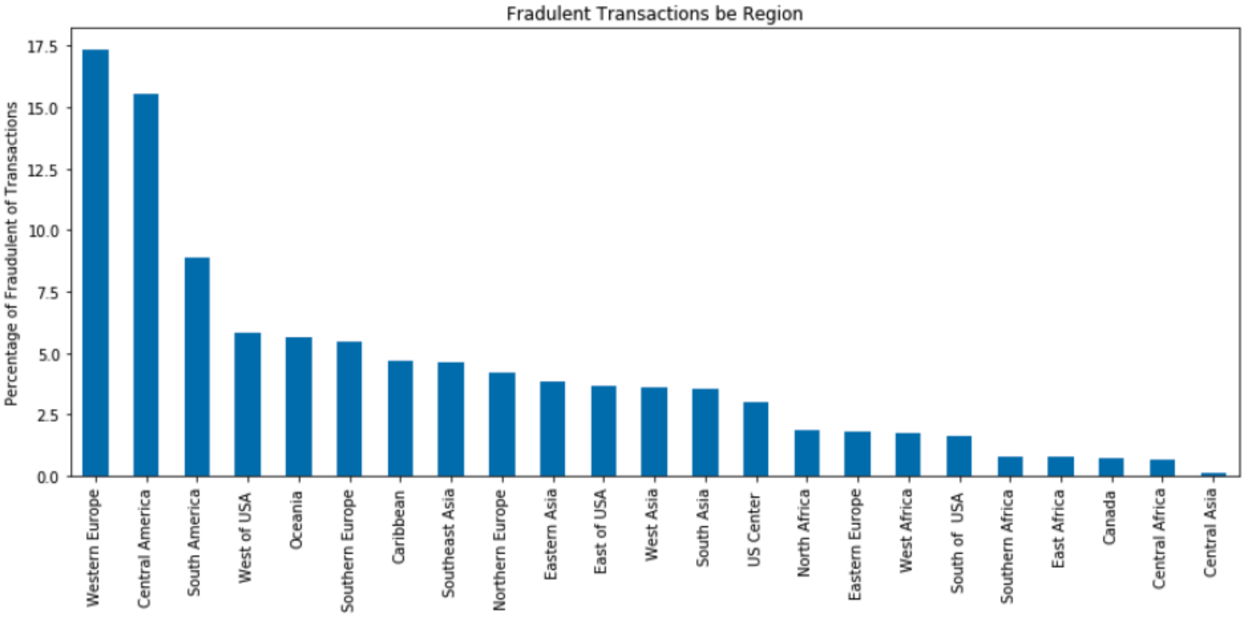
It can be seen that Cleats is the category with maximum frequency of loss generating transactions followed by Mens footwear. Fishing records the highest loss in loss generating transactions.

Most lost sales are happening in Western Europe & Central America region. This lost sales may have happened due to suspected frauds or late deliveries.

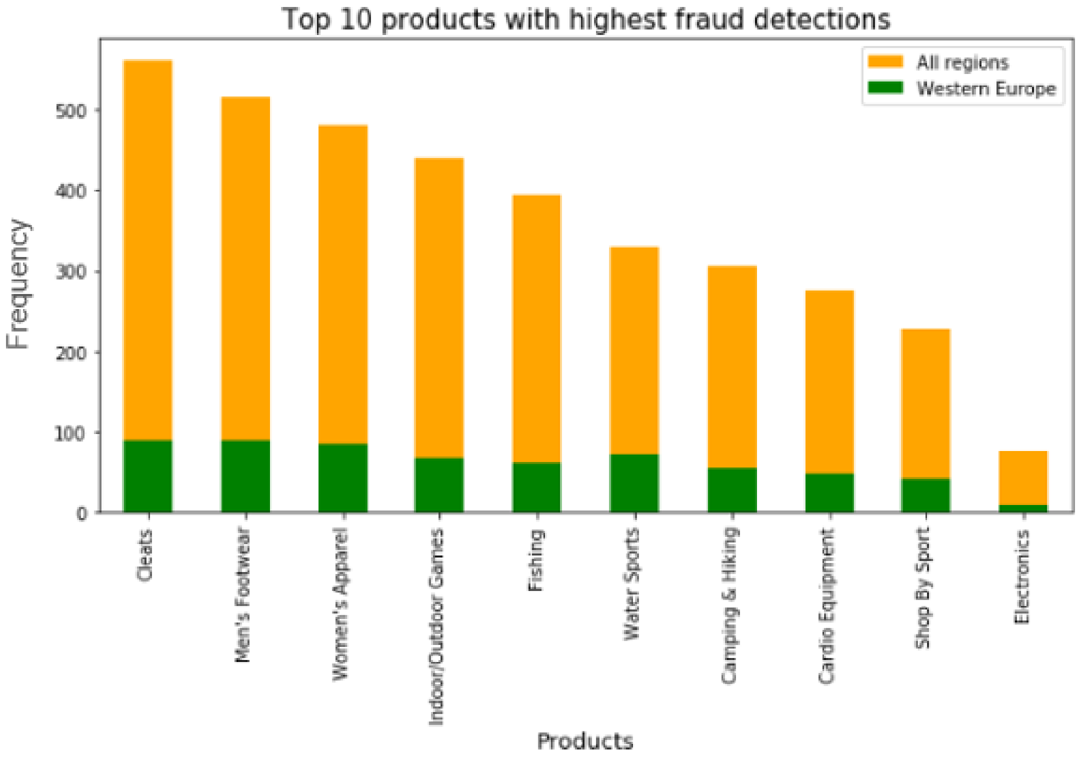
Exploration of Fraudulent Transactions:

Fraudulent transactions is a big challenge for any supply chain company. Let's see if we can find any patterns that can help us in this direction.

Below we look at the distribution of ‘Suspected Fraud’ transactions over different regions.

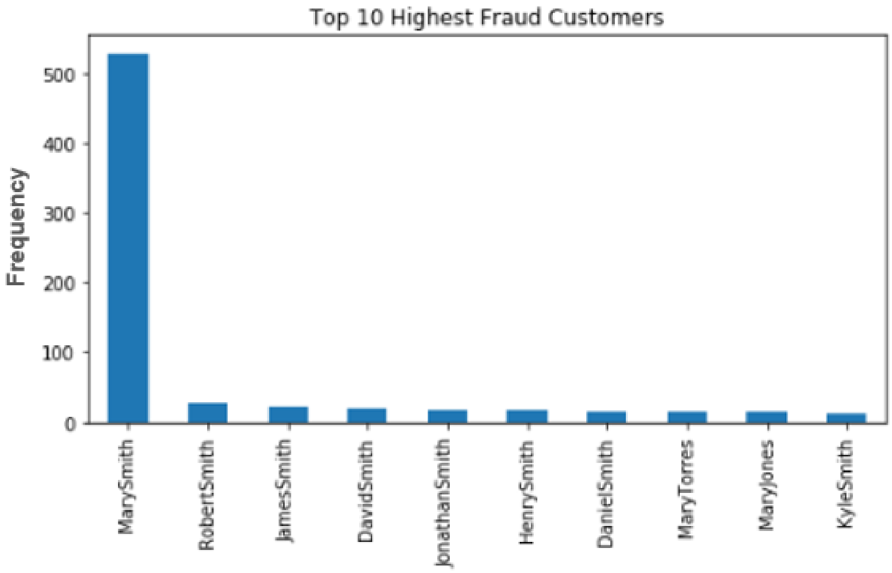


It can be observed that the highest number of suspected fraud orders are from Western Europe which is approximately 17.4% of total orders followed by Central America with 15.5%. Which product is being suspected of fraud the most?



Cleats department is being suspected of fraud the most followed by Men's footwear in all the regions and also in Western Europe.

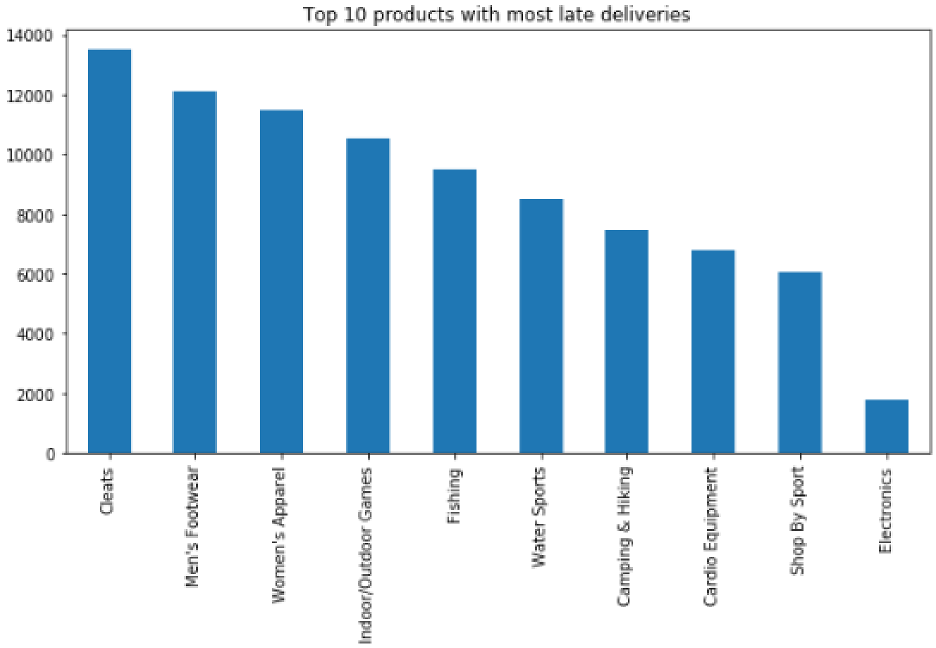
Let's look at the consumers who are suspected of fraud.



Looks like the customer named Mary Smith alone was responsible for trying to conduct fraud 528 times which is very shocking. The total amount in fraudulent transactions was almost 102k which is a huge amount. Looks like Mary was using a different address every time when placing orders, a new customer id was issued each time which makes it difficult to identify the customer and ban them. All these parameters should be taken into consideration to improve fraud detection algorithms so fraud can be identified more accurately.

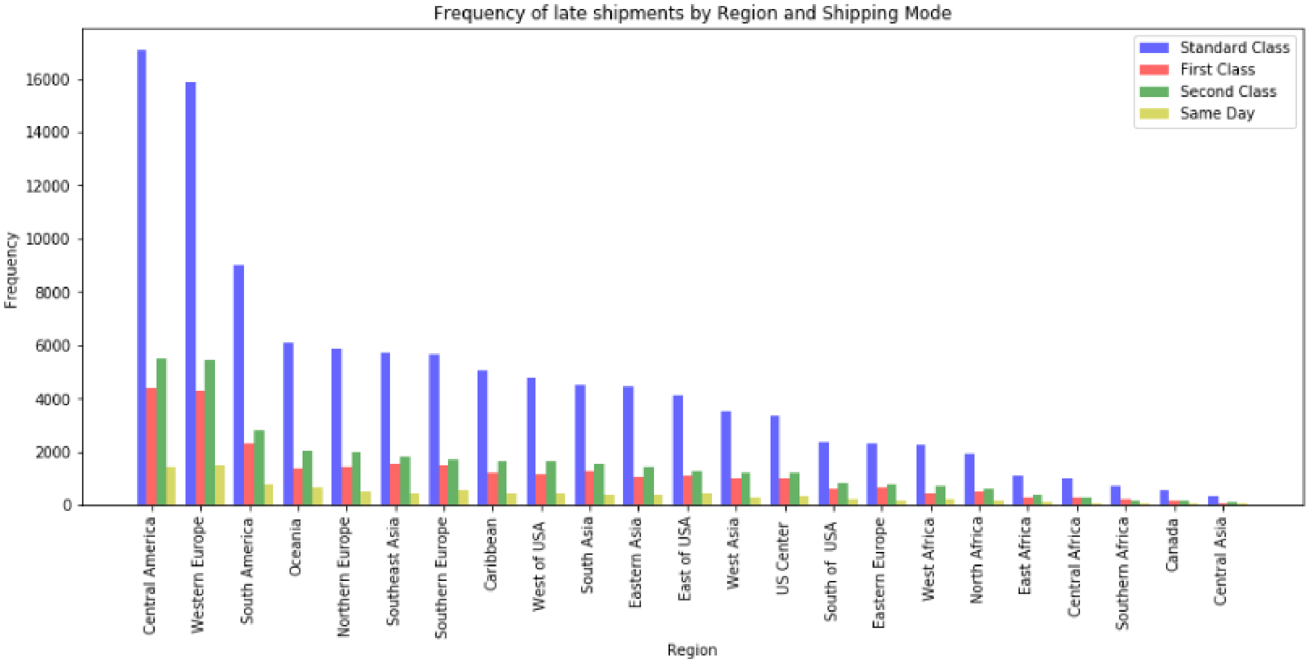
EDA of the Logistics:

Delivering products to customers on time without late delivery is another important aspect for a supply chain company and is a big driver for customer satisfaction. What category of products are being delivered late the most?



It can be seen that orders with the Cleats department getting delayed the most followed by Men's Footwear. There could be multiple aspects involved in driving a product's delivery period from demand, supply, product times and seasonality and of course the logistics ecosystem.

Next we look at how the different shipping methods compare with respect to late deliveries in different geographical regions:



It's not a surprise to see the Standard Shipping mode lead in terms of late deliveries. Standard shipping is followed by Second Class and First Class which are close in terms of late deliveries. The least late deliveries are for Same Day mode. We can see that the costlier the delivery method, the higher the chance of on-time delivery.

As for the regions, Central America has higher late deliveries than Western Europe even though the latter has higher over sales. This clearly points to the fact that Central America has worse logistics infrastructure compared to Western Europe.

Customer Segmentation:

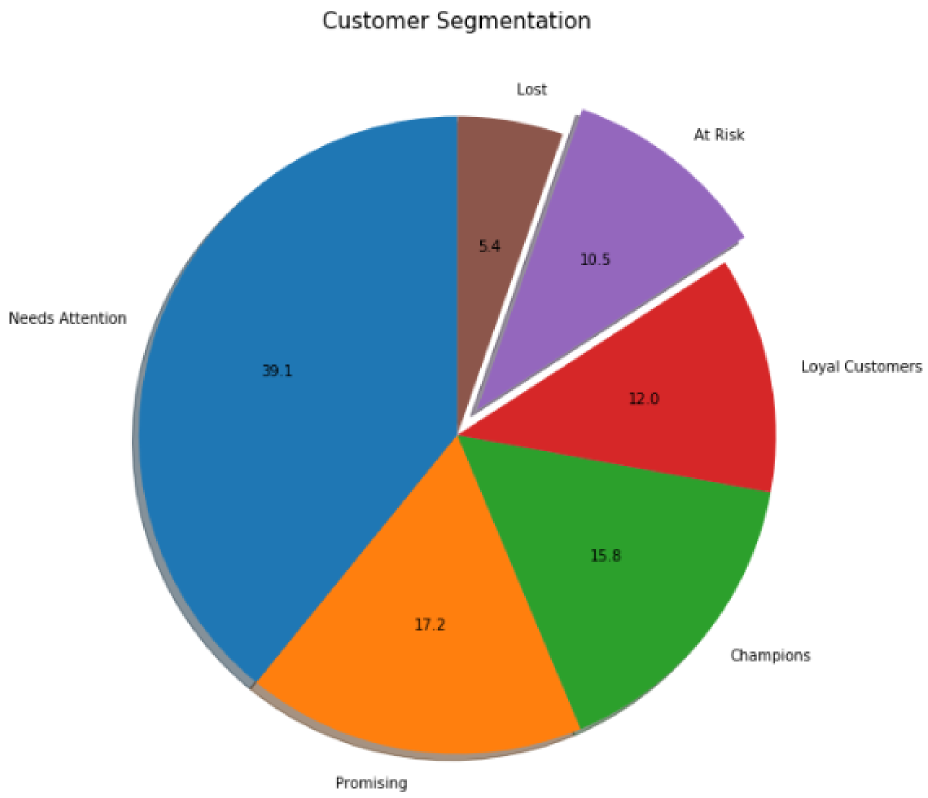
We use the RFM (Recency, Frequency, Monetary) technique to segment customers. We exploit the purchase data available to calculate:

- Time elapsed since last purchase (Recency)
- Frequency or purchase (Frequency)
- Total value of goods purchased (Monetary)

To obtain a single value for customer comparison & segmentation, we bin the R, F and M scores into ‘quartiles’ and sum these into a single score for each customer.

Categorizing these scores with labels, we observe that 10.5% of the customers are at risk of attrition while 39.2% need attention to be converted to regular customers. We see a churn of about 5.4%.

The pie chart below shows the percentage of customers falling in different categories. Such an analysis can be really useful in tailoring discount and loyalty programs in order to offer something extra to each category, driving the customer upwards in the sales cycle and boosting sales.



5. In-Depth Analysis and Machine Learning

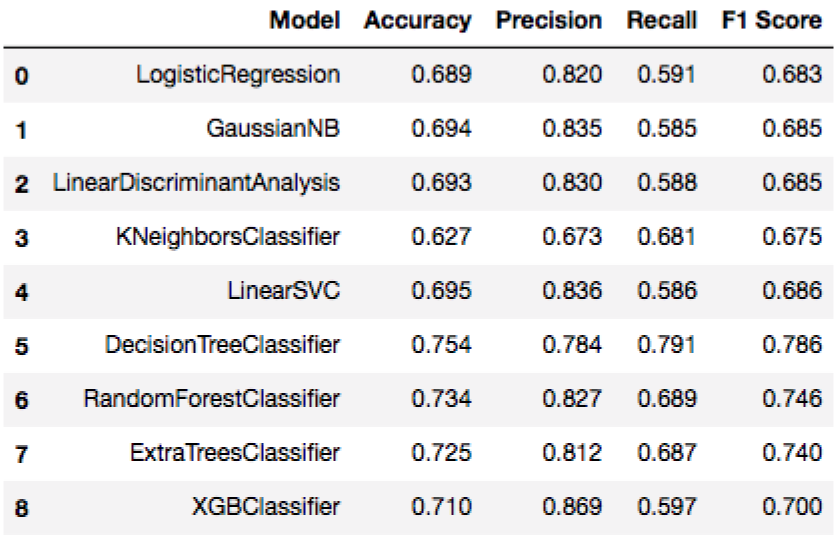
In this section we explore different machine learning classifiers to predict late deliveries given the transactional sale data. We select the most promising classifiers and perform hyperparameter tuning to tweak our models' performance.

5.1. Data Modelling, Evaluation and Validation

- We drop some redundant features and convert all categorical variables to numeric type using Scikit Learn’s LabelEncoder.
- The feature array is standardized using the StandardScaler.
- We choose the f1-measure as our evaluation metric for comparison of different models. We also look at Accuracy, Recall and Precision from these modelsl.
- We use a 5-fold Cross Validation technique to evaluate our model efficacy

5.2. Model Comparison

The following table shows the different models trained on the data and their classification scores:



	Model	Accuracy	Precision	Recall	F1 Score
0	LogisticRegression	0.689	0.820	0.591	0.683
1	GaussianNB	0.694	0.835	0.585	0.685
2	LinearDiscriminantAnalysis	0.693	0.830	0.588	0.685
3	KNeighborsClassifier	0.627	0.673	0.681	0.675
4	LinearSVC	0.695	0.836	0.586	0.686
5	DecisionTreeClassifier	0.754	0.784	0.791	0.786
6	RandomForestClassifier	0.734	0.827	0.689	0.746
7	ExtraTreesClassifier	0.725	0.812	0.687	0.740
8	XGBClassifier	0.710	0.869	0.597	0.700

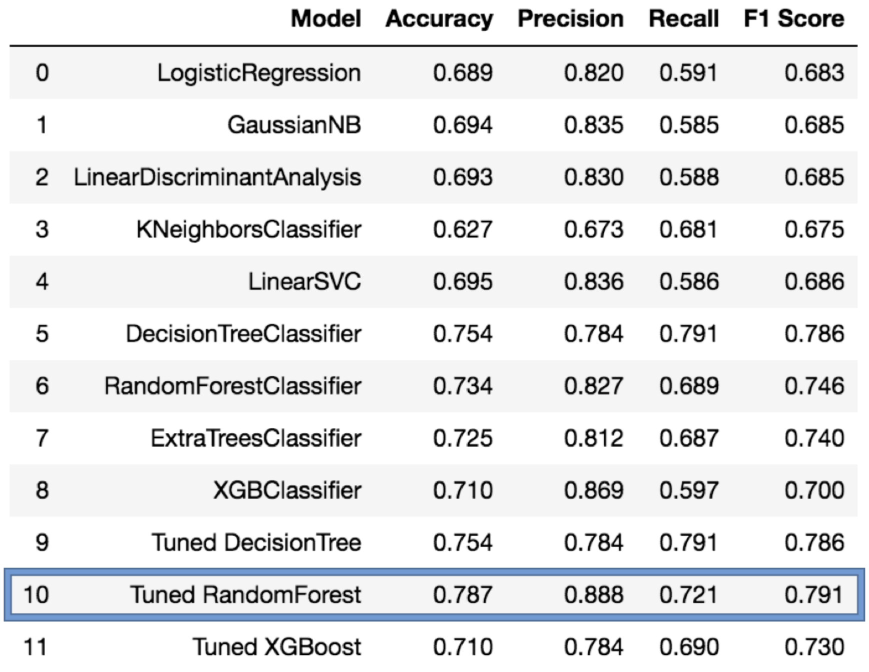
**Dimensionality reduction using PCA was done for KNN and LinearSVC models due to large training times.
We can clearly see that the tree based non-linear classifiers are performing better than the other models. We take these models further by tuning their hyperparams and

5.3. HyperParameter Tuning

Hyperparameter tuning is performed on the Decision Tree, Random Forest and XGBoost Classifiers. The table below summarizes the Parameter Grids explored for each model and the best parameters returned after tuning.

Model	Param Grid	Best Params
DecisionTreeClassifier	max_depth: [10, 20, 30, 40 , 50, 60, 70, None], min_samples_split: [2, 10, 25, 50]	max_depth: 70, min_samples_split: 2
RandomForestClassifier	n_estimators:[100, 500, 1000], max_features: [5, 10, 15, 20]	n_estimators: 1000, max_features: 20
XGBClassifier	n_estimators: [100, 500, 1000, 2000], learning_rate: [0.05, 0.1, 0.2, 0.3]	n_estimators: 1000, learning_rate: 0.3

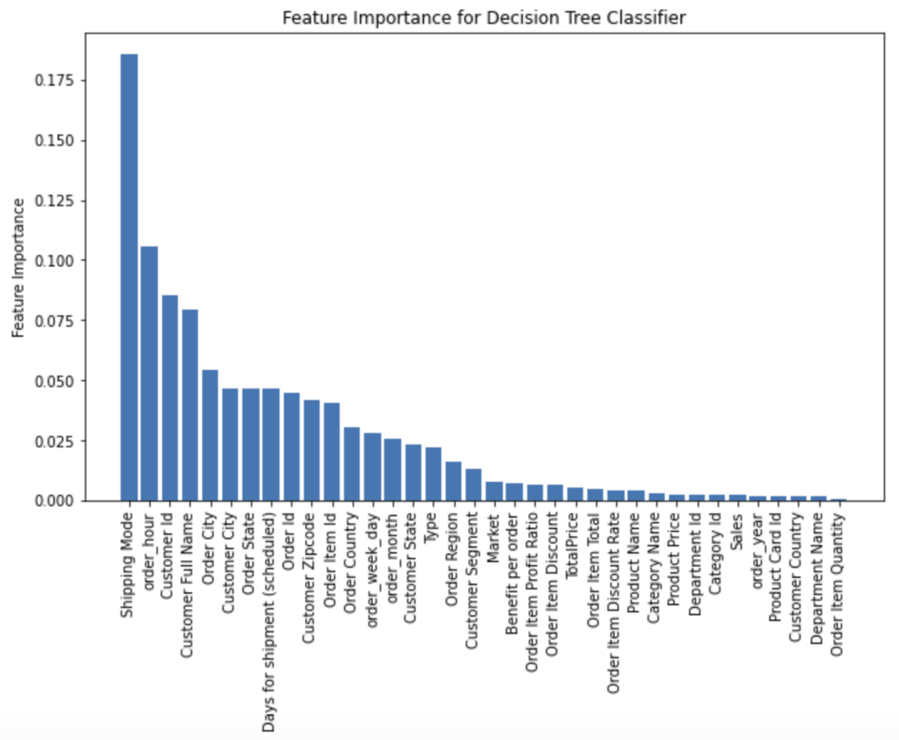
We can see that random forest after tuning the hyperparams emerges as the best model for predicting late deliveries with an F1 score of 0.791.

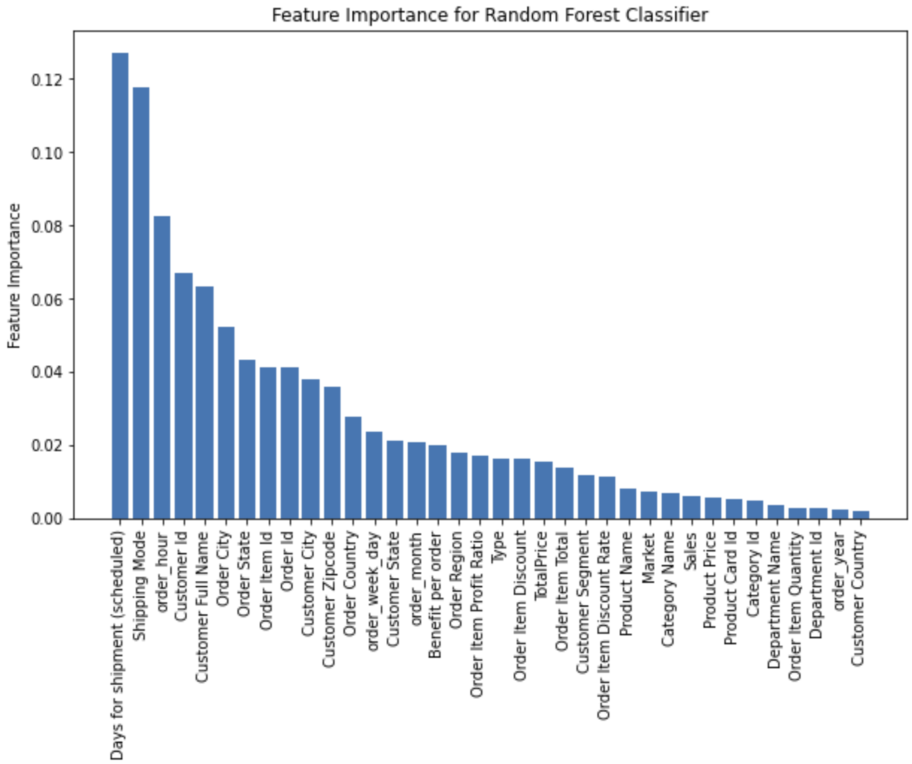


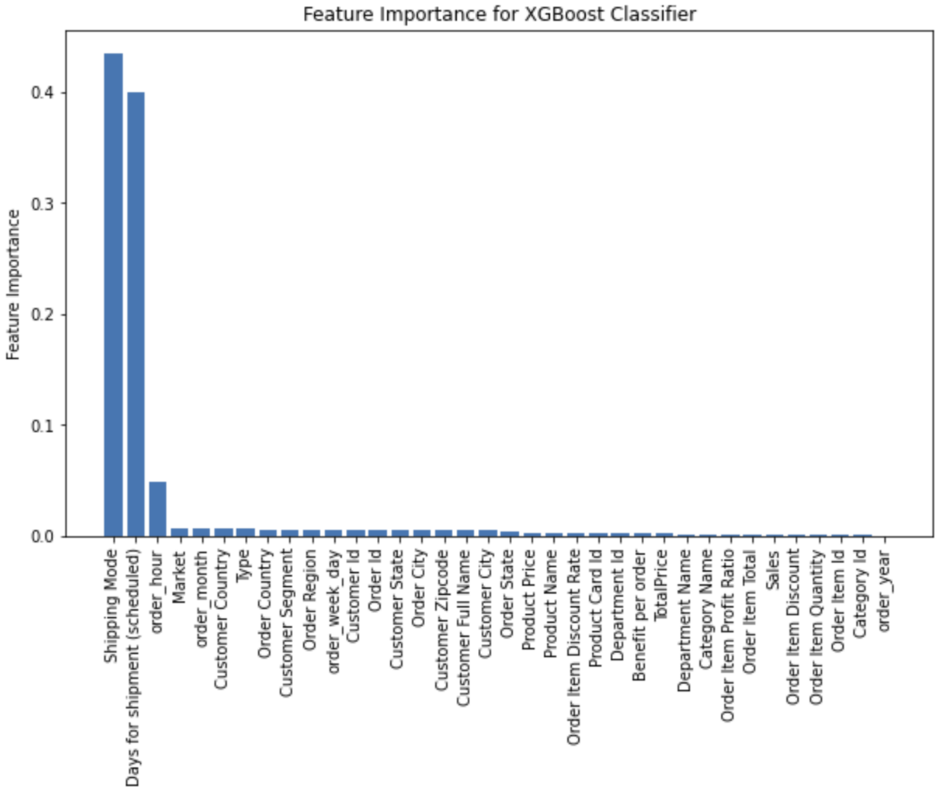
	Model	Accuracy	Precision	Recall	F1 Score
0	LogisticRegression	0.689	0.820	0.591	0.683
1	GaussianNB	0.694	0.835	0.585	0.685
2	LinearDiscriminantAnalysis	0.693	0.830	0.588	0.685
3	KNeighborsClassifier	0.627	0.673	0.681	0.675
4	LinearSVC	0.695	0.836	0.586	0.686
5	DecisionTreeClassifier	0.754	0.784	0.791	0.786
6	RandomForestClassifier	0.734	0.827	0.689	0.746
7	ExtraTreesClassifier	0.725	0.812	0.687	0.740
8	XGBClassifier	0.710	0.869	0.597	0.700
9	Tuned DecisionTree	0.754	0.784	0.791	0.786
10	Tuned RandomForest	0.787	0.888	0.721	0.791
11	Tuned XGBoost	0.710	0.784	0.690	0.730

5.4. Feature Importances

Finally we look at feature importances assigned by our three tuned models and take a look at the path taken by our decision tree model. We conclude with suggestions on what can be done to tackle the features impacting late deliveries the most.




Feature Importance for Random Forest Classifier
Feature Importance
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Days for shipment (scheduled)
Shipping Mode
order_hour
Customer Id
Customer Full Name
Order City
Order State
Order Item Id
Order Id
Customer City
Customer Zipcode
Order Country
order_week_day
Customer State
order_month
Benefit per order
Order Region
Order Item Profit Ratio
Type
Order Item Discount
TotalPrice
Order Item Total
Customer Segment
Order Item Discount Rate
Product Name
Market
Category Name
Sales
Product Price
Product Card Id
Category Id
Department Name
Order Item Quantity
Department Id
order_year
Customer Country


Feature Importance for XGBoost Classifier
Feature Importance
0.4
0.3
0.2
0.1
0.0
Shipping Mode
Days for shipment (scheduled)
order_hour
Market
order_month
Customer Country
Type
Order Country
Customer Segment
Order Region
order_week_day
Customer Id
Order Id
Customer State
Order City
Customer Zipcode
Customer Full Name
Customer City
Order State
Product Price
Product Name
Order Item Discount Rate
Product Card Id
Department Id
Benefit per order
TotalPrice
Department Name
Category Name
Order Item Profit Ratio
Order Item Total
Sales
Order Item Discount
Order Item Quantity
Order Item Id
Category Id
order_year

We see that Shipping mode, Schedule shipment days and the hour of placing the order have high significance in all these three models. We have also concurred from our EDA of different shipment modes vs. late frequencies that shipping mode is a significant factor in late deliveries.

Suggestions to tackle these three factors and improve logistics efficacy are presented in the section below.

6. 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 transactions 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 Quarter 3 following which the sales suddenly dipped by almost 65% in 2018 quarter 1. 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.

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. When compared with other classification machine learning models, the DecisionTree did a good job of identifying orders with later delivery with an f1 score of 78.6%. When we tuned the RandomForest, it showed a much better prediction accuracy for late deliveries with an F1 score of nearly 80%. 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 importances 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 (Will be late or not?). This was extremely evident in the XGBoost model where the model chose to assign negligible importances to all other features except the three listed above.

The company can take multiple steps to improve on these three features. Two critical suggestions in this regard would be:

- Improve or switch the logistics for increased reliable performance on lower priced shipping modes.
- Incorporate the order traffic and same day dispatch feasibility (these parameters could be indicative of 'order_hour') into the shipping estimates when orders are placed by customers. This will improve the feature response with both ‘days for shipment (scheduled)’ as well as ‘order_hour’.