E-Commerce & Retail B2B Case Study

Muqeet Sultan Khan

Mohammed Ayaz Khan

Kevin Mathews

Understanding the issue & defining the goals

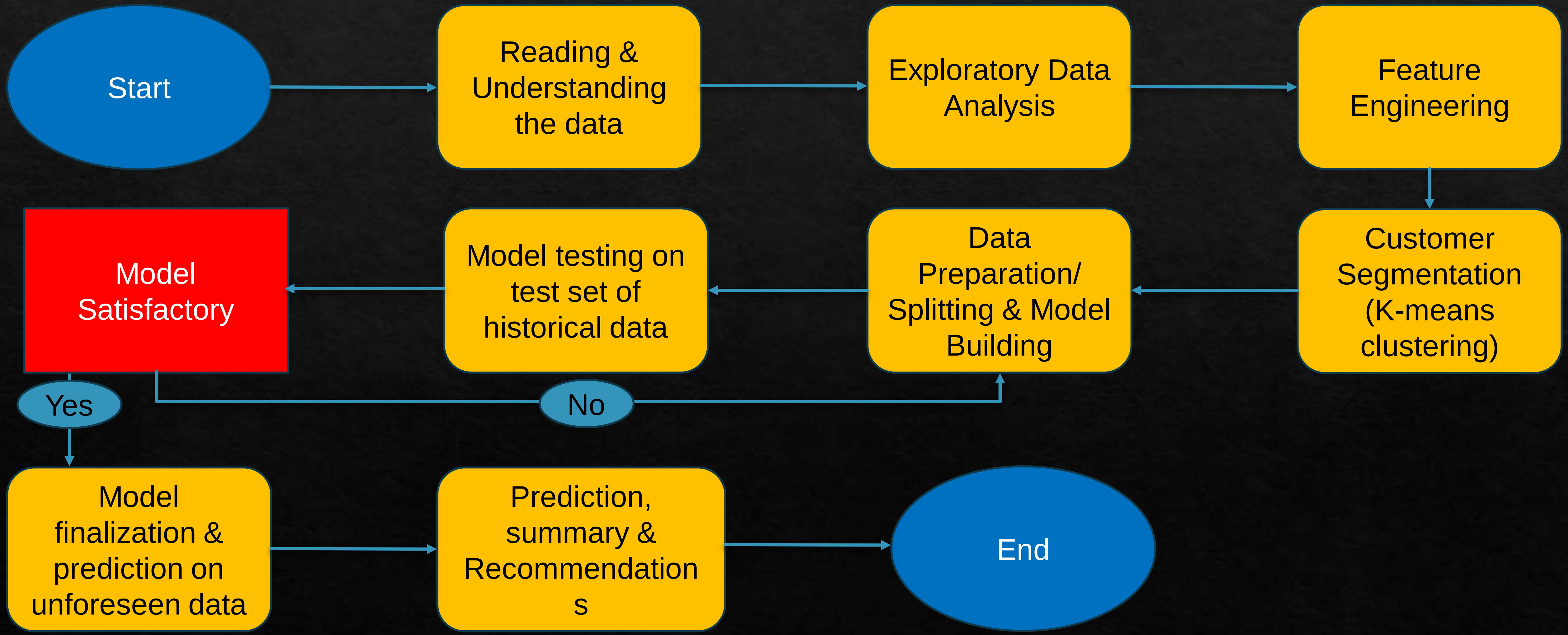
Problem Statement

- Schuster is a multinational retailer of sports goods & garments. It conducts significant business with hundreds of vendors with whom it has arranged a credit arrangement.
- However, not all its vendors respect the credit terms and tend to make late payments. While Schuster imposes a fine for every late payment made by the vendor, but such an approach is not seemingly beneficial for the long-term relationships of either of the parties.
- The collectors must keep chasing the vendors to ensure the payment is received on time. In case of late payments, they spend considerable time coordinating the payments. This resulted in many non-value-added activities, loss of time and effort and financial impact on Schuster.

Goal

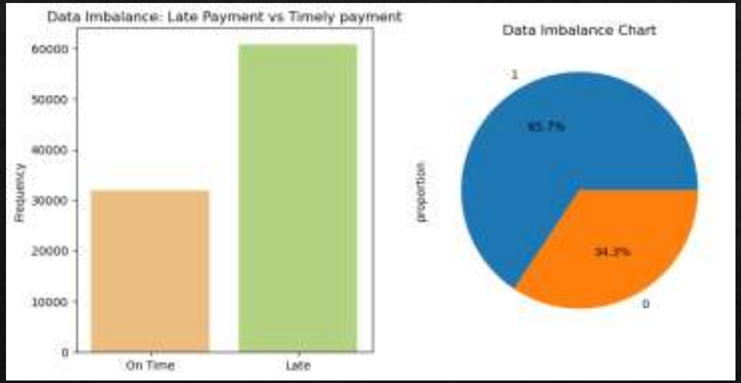
- Schuster would like to better understand the customers' payment behaviour based on their past payment patterns (customer segmentation).
- Using historical information, it wants to be able to predict the likelihood of delayed payment against open invoices from its customers.
- It wants to use this information so that collectors can prioritise their work in following up with customers beforehand to get the payments on time.

Mapping the strategy to the problem

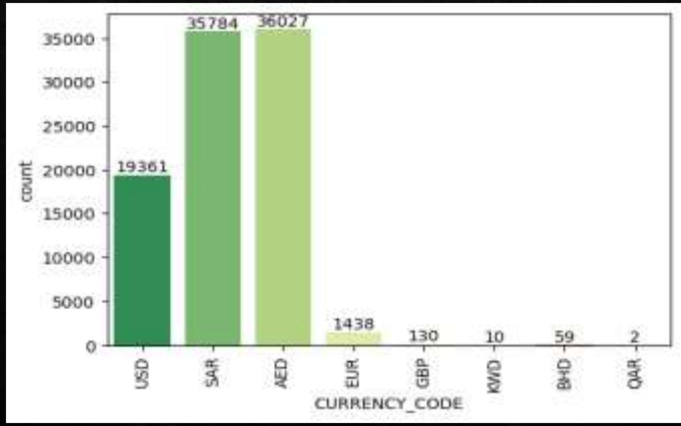


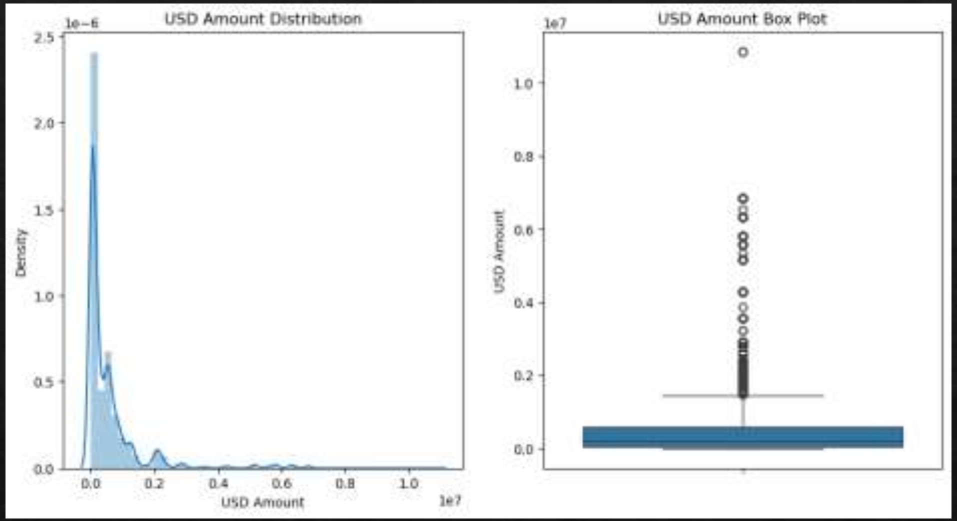
Univariate Analysis

- The class imbalance from the figure shown is 65.7% towards delayers which is acceptable and does not need imbalance treatment.



- The top 3 currencies with which the company makes deals are: AED,SAR & USD, suggesting more deals made in the middle-east.

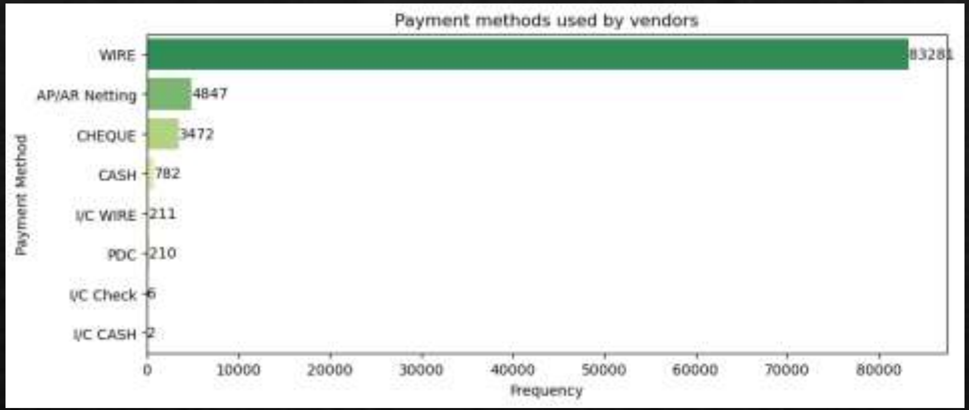




- The figure shows that the transactions tend to lie between a range of $1 and $2.5m
- The frequency of transactions is higher below $1.75m

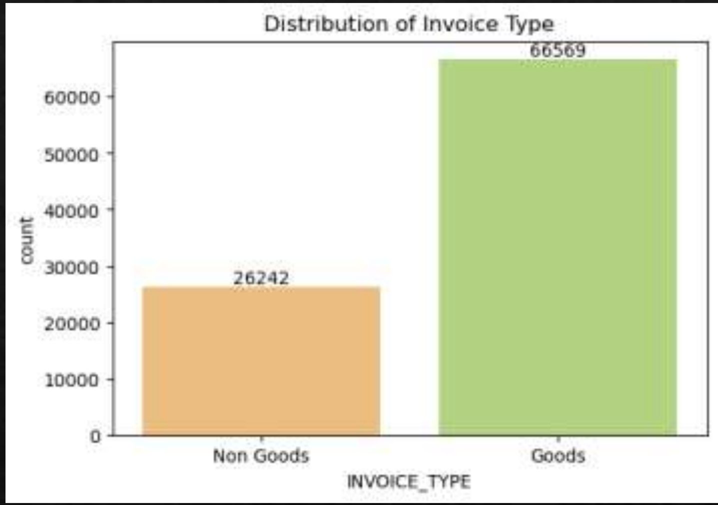
Univariate Analysis

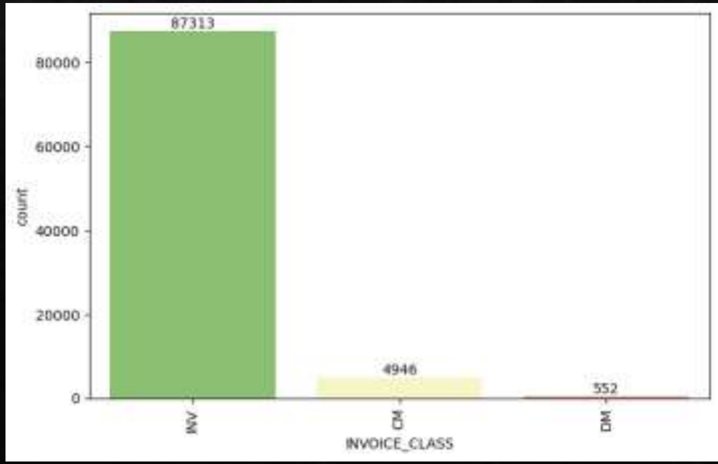
- Wire payment method is the most common payment method received by the company, followed by netting, cheque and cash.



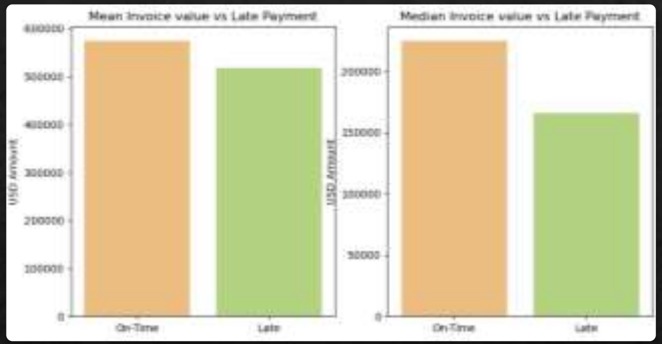
- From the figure on the right, “Invoice” is the major invoice class, while the rest have very low percentages.

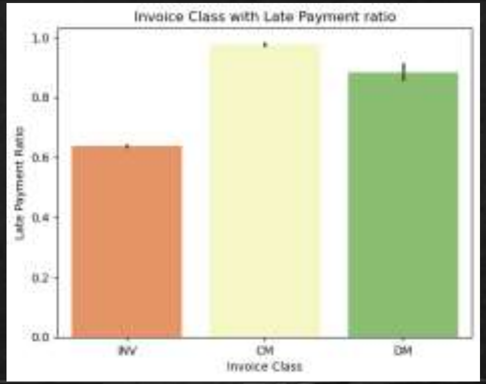
- Goods type of invoices comprise a major share of invoices generated.

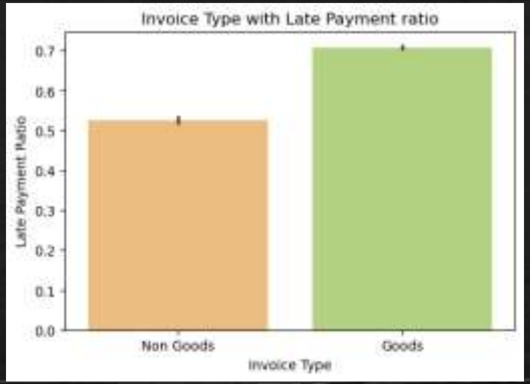




Bivariate Analysis







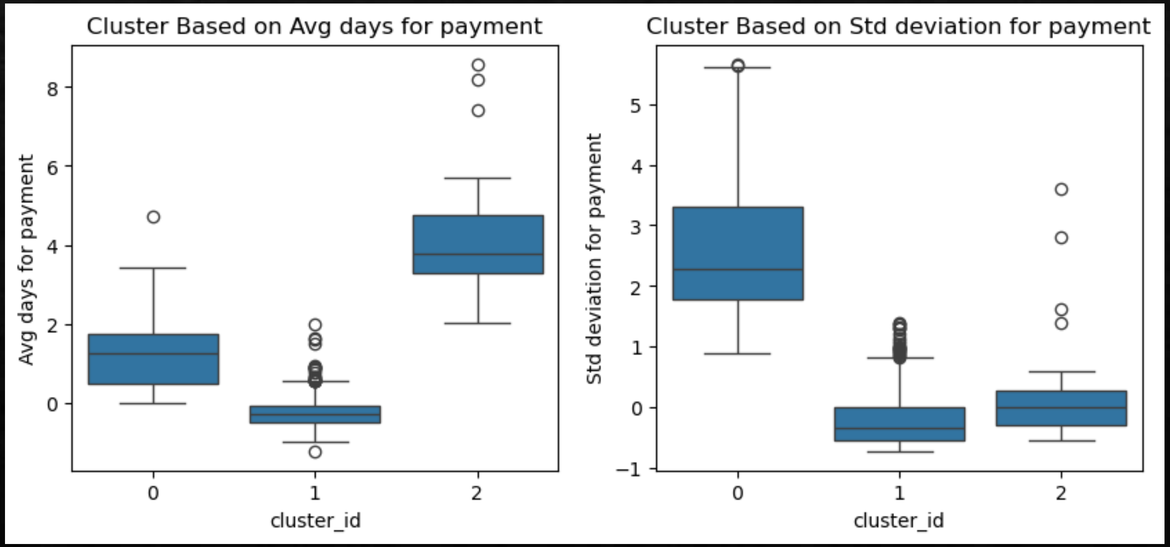
The mean and median of the payment amount is higher for payers who pay on time than late, suggesting that higher value transactions show lesser delay risk than lower value transactions.

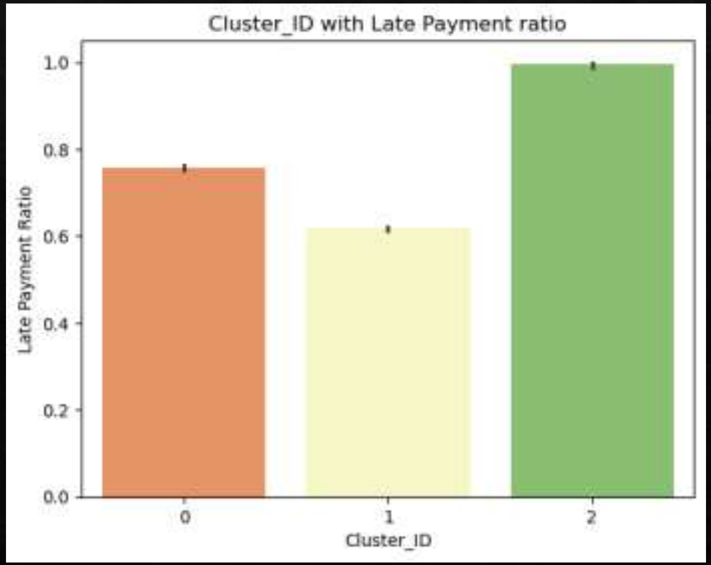
Late payment ratio for Credit Note transaction types are maximum, followed by Debit Note and Invoice suggesting higher delay risk in Credit and Debit note invoice classes.

Goods type invoices show greater late payment ratio than non-goods hence showing increased chances of payment delay.

Customer Segmentation(K-means Clustering)

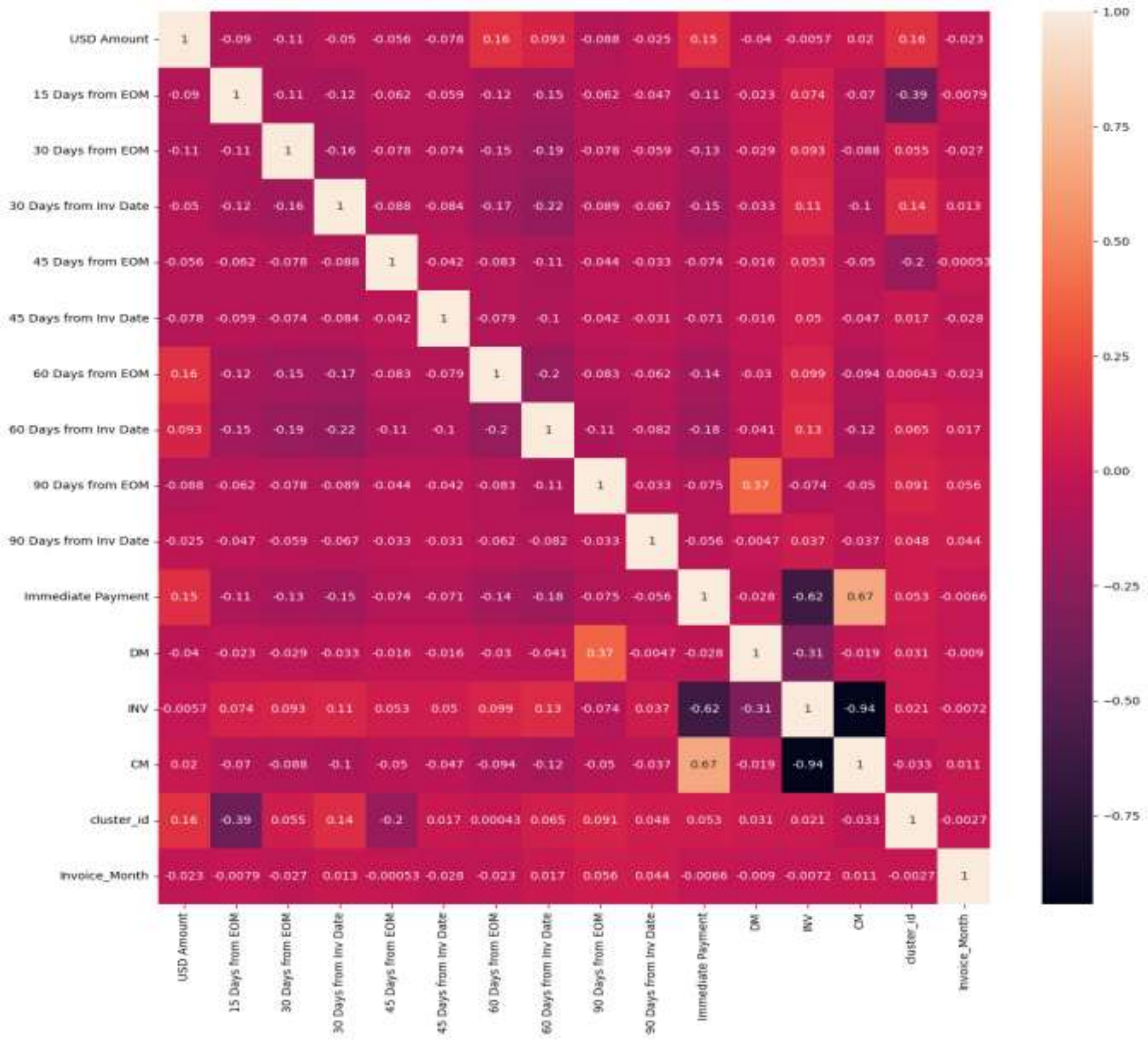
- One objective was to segment the customers to understand the payment behaviors which was done using K-means clustering method.
- The number of clusters were decided to be 3 since with increase in clusters post 3, there was a significant decrease in silhouette score.
- The category 1 were early payers with least number of average days taken to pay and category 2 were prolonged payers with greatest number of average days taken to pay. Category 0 lie in between the other two categories and hence labelled as medium duration payers.





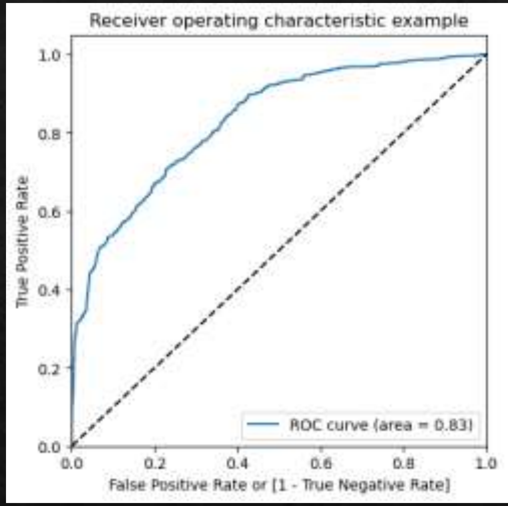
Model Building

CM & INV, INV & Immediate Payment, DM & 90 days from EOM has high multicollinearity, hence dropping these columns to prevent multicollinearity effect.

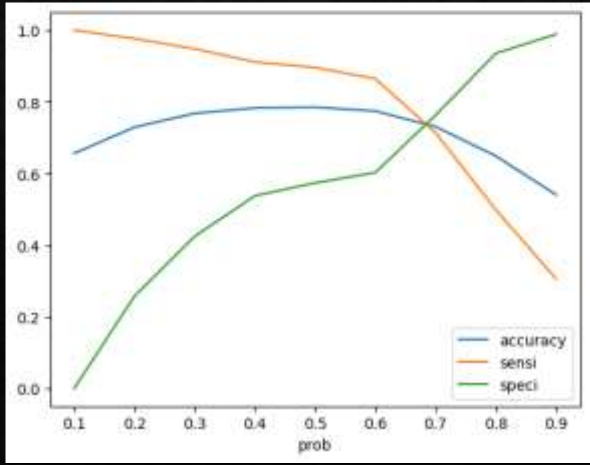


Comparison between Logistic Regression and Random Forests

- Logistic regression model formed after dropping multicollinearity and unnecessary variables resulted in remaining variables with acceptable p-value and VIF figures, hence retained the remaining features with no further feature elimination and a good ROC curve area of 0.83.



- The trade-off plot between accuracy, sensitivity and specificity revealed an optimum probability cutoff of ~0.7, which was used to further predict which transactions would result in delayed payments in the received payments dataset.



Comparison between Logistic Regression and Random Forests

- A random forest model was built using the same parameters as the logistic regression with hyper-parameter tuning, which resulted in the following parameters.

```
Best hyperparameters: {'max_depth': 30, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 150}
Best f1 score: 0.9392906020311104
```

- Using the above parameters, a random forest model was built, whose metrics were compared to the logistic regression model and the final model was finalized therefore

Random Forest Feature Rankings

```
Feature ranking:
1. USD Amount (0.491)
2. Invoice_Month (0.129)
3. 30 Days from EOM (0.114)
4. 60 Days from EOM (0.110)
5. Immediate Payment (0.041)
6. 15 Days from EOM (0.028)
7. cluster_id (0.027)
8. 60 Days from Inv Date (0.013)
9. 30 Days from Inv Date (0.011)
10. 90 Days from Inv Date (0.008)
11. INV (0.007)
12. 90 Days from EOM (0.007)
13. 45 Days from EOM (0.005)
14. CM (0.004)
15. 45 Days from Inv Date (0.004)
16. DM (0.001)
```

- The random forest was used to further find the feature ranking which shows the top 5 features to predict the delay, which are:
  1. USD Amount
  2. Invoice Month
  3. 30 Days from EOM
  4. 60 Days from EOM
  5. Immediate Payment

Customers with the highest delay in payment probabilities

- Predictions suggest that the companies mentioned in the table have the maximum probability to default with maximum number of delayed and total payments.

Customer_Name	Delayed_Payment	Total_Payments	Delay%
IL G Corp	13	13	100.0
RNA Corp	9	9	100.0
SHIS Corp	8	8	100.0
ALSU Corp	7	7	100.0
LVMH Corp	4	4	100.0
FINA Corp	4	4	100.0
V PE Corp	4	4	100.0
VIRT Corp	3	3	100.0
MANA Corp	3	3	100.0
TRAF Corp	3	3	100.0

Recommendations

Customer_Name	Delayed_Payment	Total_Payments	Delay%
IL G Corp	13	13	100.0
RNA Corp	9	9	100.0
SHIS Corp	8	8	100.0
ALSU Corp	7	7	100.0
LVMH Corp	4	4	100.0
FINA Corp	4	4	100.0
V PE Corp	4	4	100.0
VIRT Corp	3	3	100.0
MANA Corp	3	3	100.0
TRAF Corp	3	3	100.0

Based on our analysis, we can draw the following conclusions:

Payment Delays by Invoice Type:

Credit Note payments experience the highest delay rates compared to Debit Notes and other invoice types. It is advisable to enforce stricter payment collection policies for Credit Note invoices to mitigate these delays.

Goods vs. Non-Goods Invoices:

Goods-related invoices have shown significantly higher payment delay rates compared to non-goods invoices. Implementing more stringent payment policies for goods-related invoices could help in managing these delays effectively.

Focus on Lower Value Payments:

Lower value payments constitute the majority of transactions and also experience higher rates of delay. It is recommended to concentrate efforts on these transactions. Consider applying penalties based on the billing amount, where smaller bills incur higher penalty percentages for late payments. However, such measures should be a last resort.

Customer Segmentation and Payment Behavior:

Customers were segmented into three categories based on payment duration: Cluster 0 (medium), Cluster 1 (prolonged), and Cluster 2 (early). Customers in Cluster 1 (prolonged payment) exhibit significantly higher delay rates compared to those in early and medium payment clusters. Therefore, customers in Cluster 1 should be given special attention to address their payment issues.

Prioritization of High-Risk Companies:

Companies with the highest probability of delays and significant total and delayed payment counts should be prioritized. Focusing on these high-risk companies will help in managing and reducing overall payment delays effectively.

THANK YOU