E-COMMERCE AND RETAIL B2B CASE STUDY

VIKHYAT NEGI

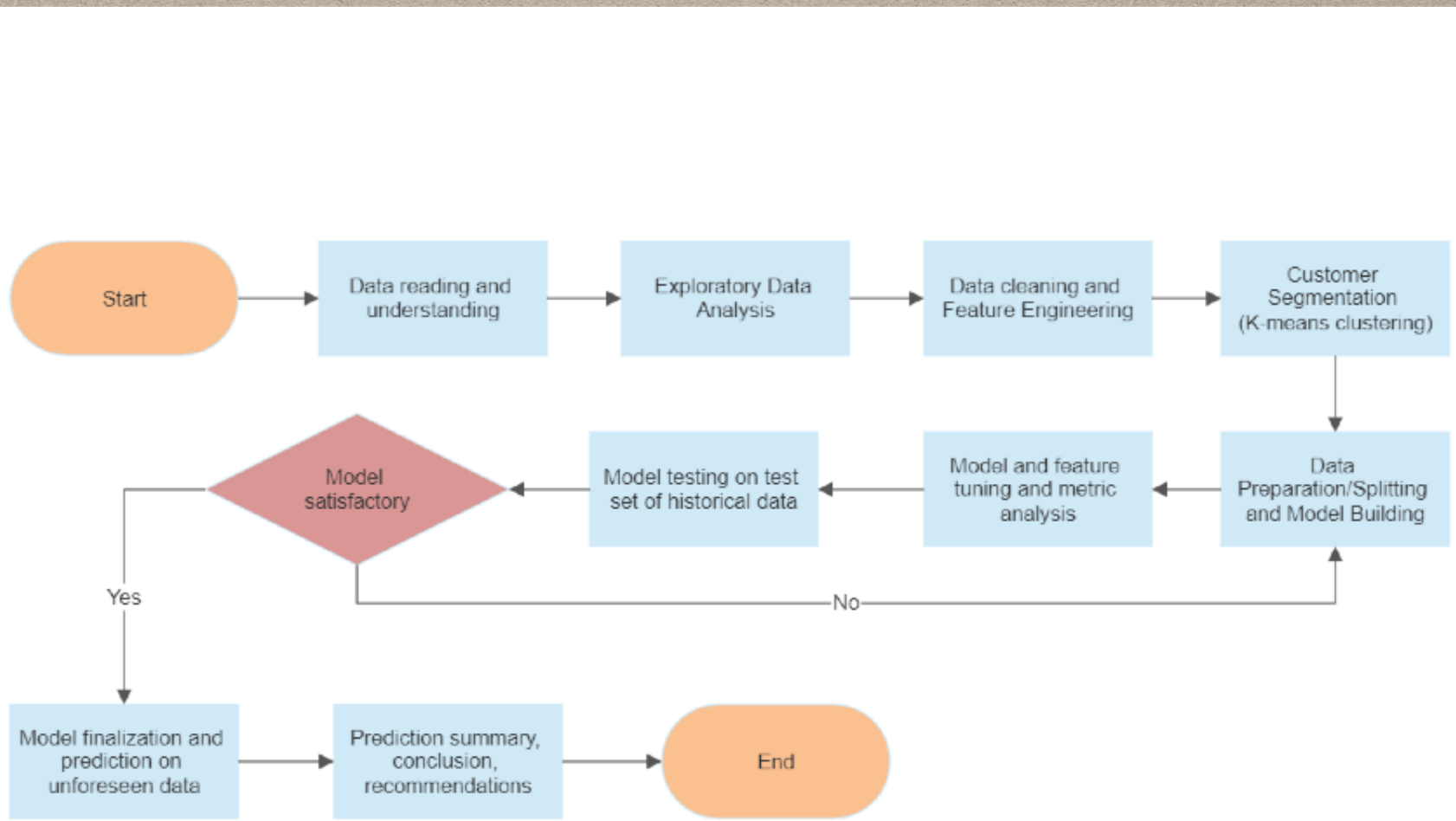
PROBLEM STATEMENT

Schuster is a multinational retail company dealing in sports goods and accessories. Schuster conducts significant business with hundreds of its vendors, with whom it has credit arrangements. Unfortunately, not all vendors respect credit terms and some of them tend to make payments late. Schuster levies heavy late payment fees, although this procedure is not beneficial to either party in a long-term business relationship. The company has some employees who keep chasing vendors to get the payment on time; this procedure nevertheless also results in non-value-added activities, loss of time and financial impact. Schuster would thus try to understand its customers' payment behaviour and predict the likelihood of late payments against open invoices.

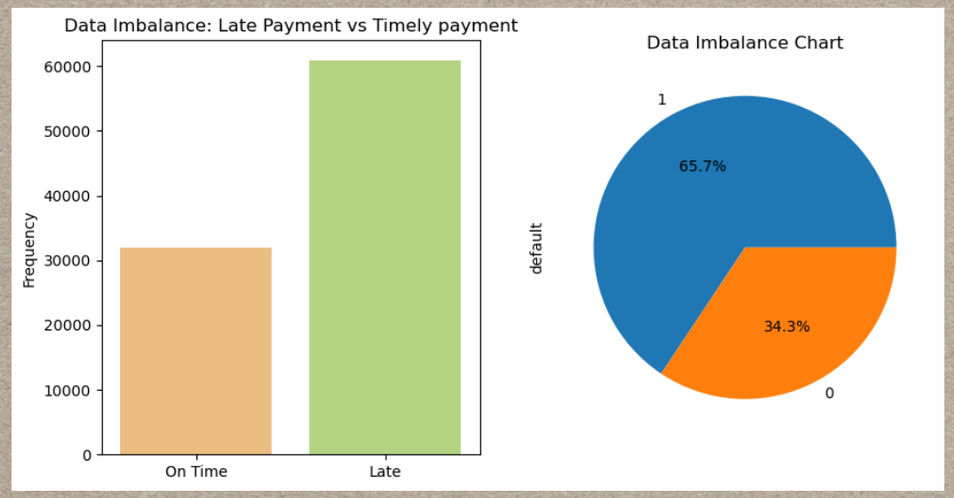
GOAL

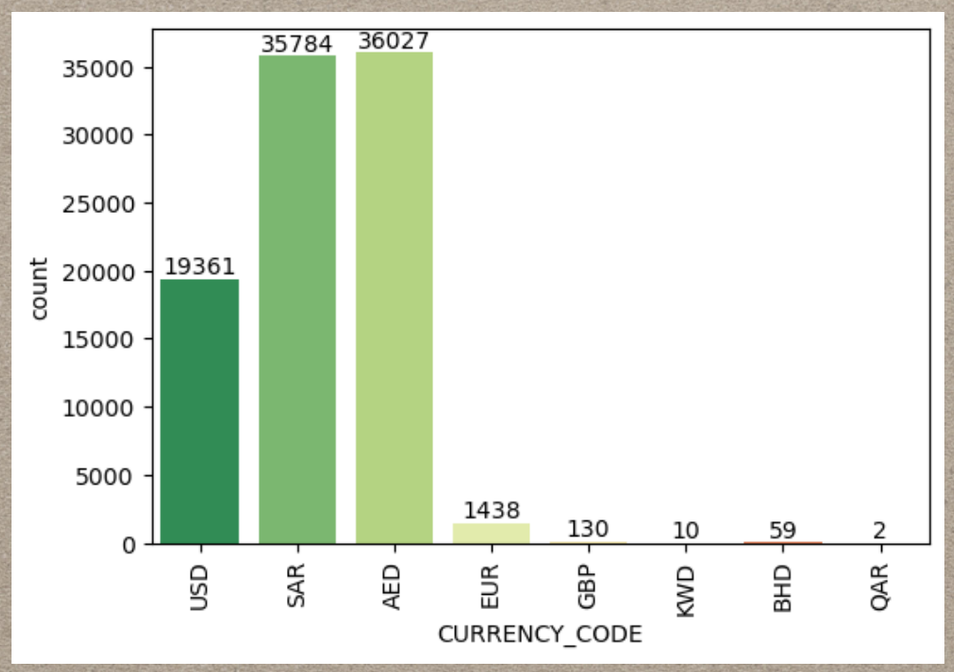
- Schuster would like to better understand the customers' payment behaviour based on their past payment patterns (customer segmentation).
- Using historical information, it wants to be able to predict the likelihood of delayed payment against open invoices from its customers.
- It wants to use this information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

APPROACH STRATEGY TO THE PROBLEM



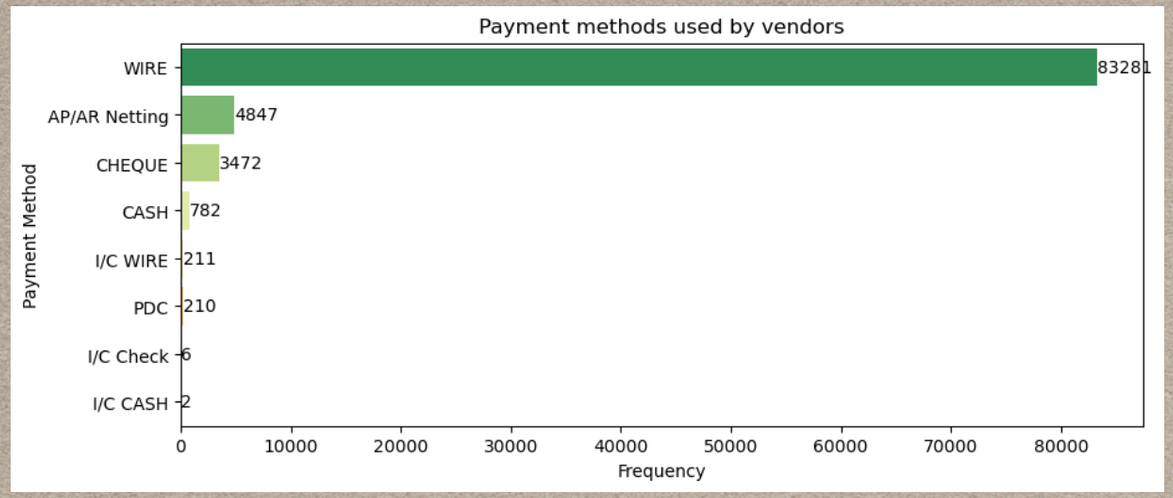
CLASS IMBALANCE AND TRANSACTION INSIGHTS (UNIVARIATE)

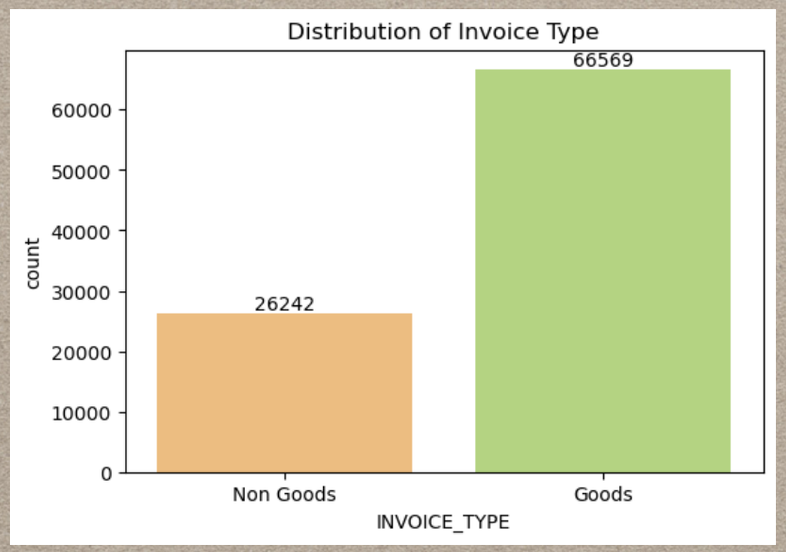


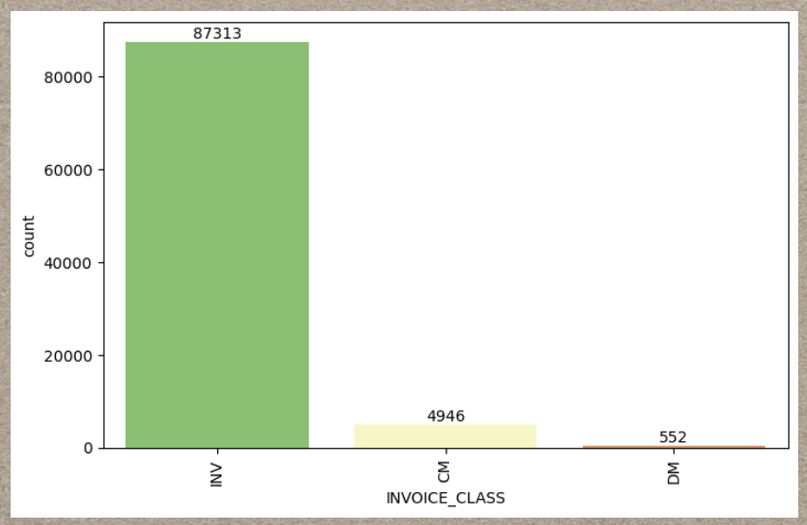


The class imbalance is 65.7% towards payment delayers which is an acceptable imbalance and does not need imbalance treatment

The top three currencies in which the company deals are AED, SAR and USD with AED as the most dealt currency suggesting greater transactions with the middle-east







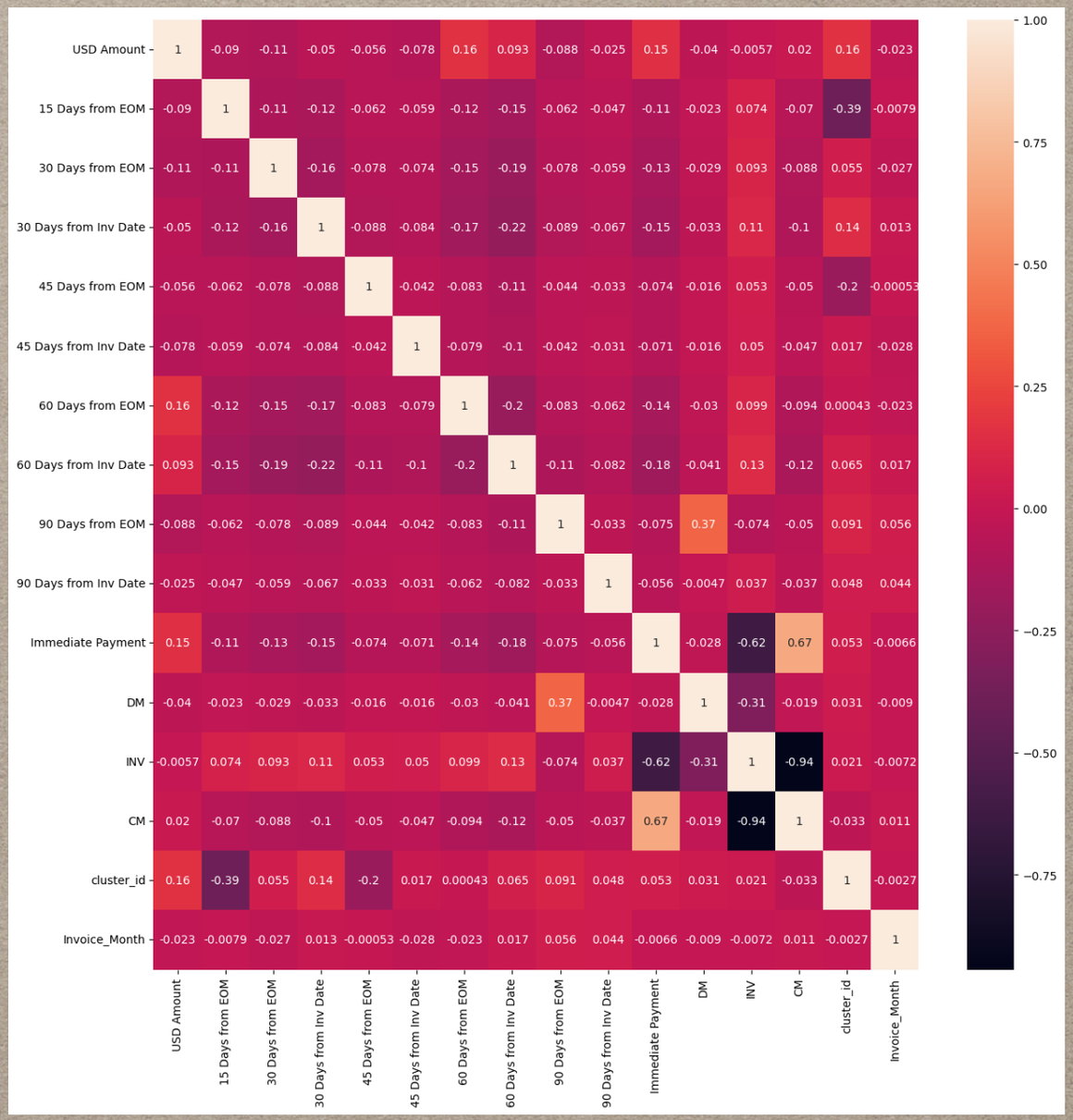
Wire payment method is the most common payment method received by the company, followed by netting , cheque and cash

Goods type invoices comprise of the major share of invoices generated

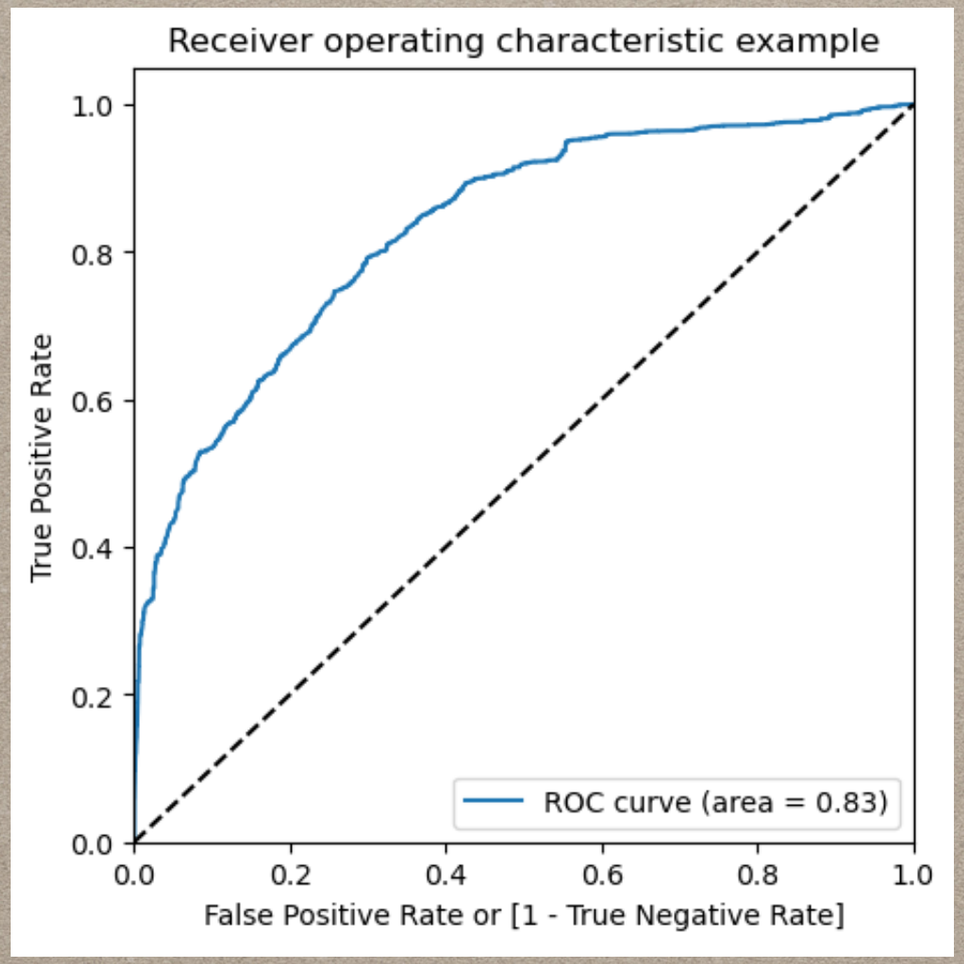
The major invoice class is ‘Invoice’ with the rest having very low percentages of the share

MODEL BUILDING

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

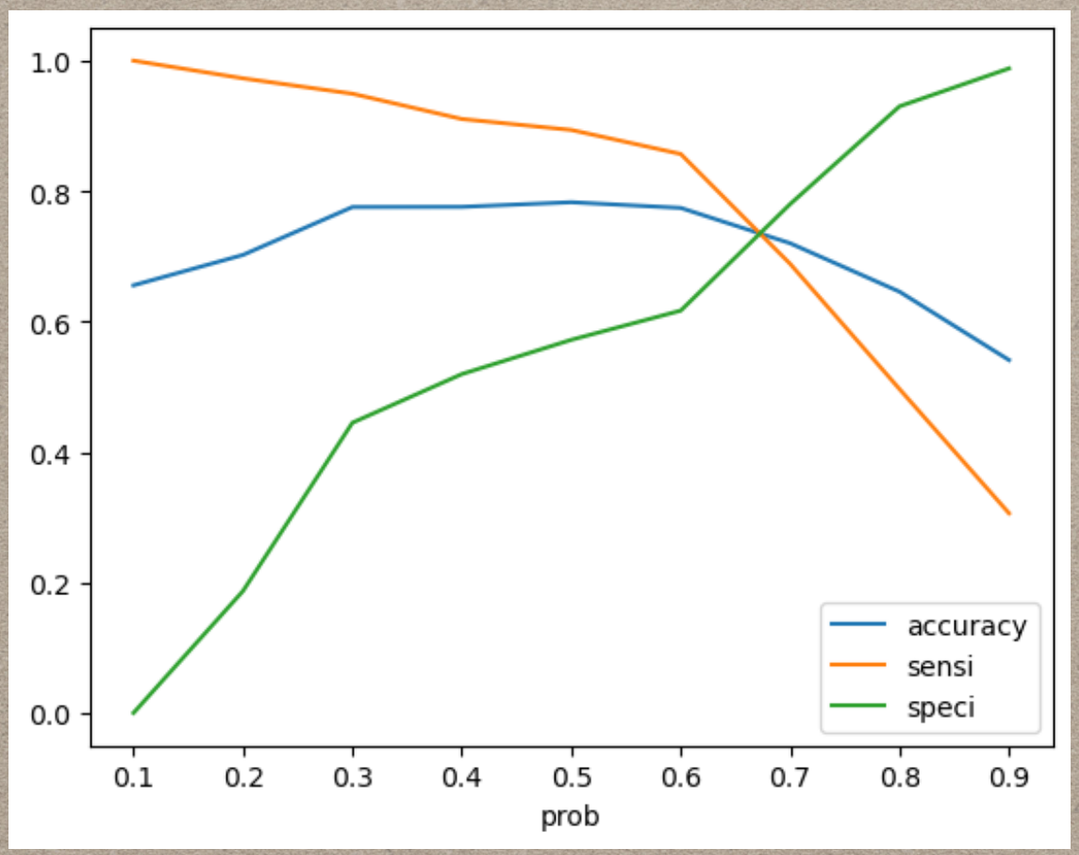


COMPARISON BETWEEN TWO MODELS, LOGISTIC REGRESSION AND RANDOM FORESTS



- 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

COMPARISON BETWEEN TWO MODELS, LOGISTIC REGRESSION AND RANDOM FORESTS



The trade-off plot between accuracy, sensitivity and specificity revealed an optimum probability cutoff of ~0.6, which was used to further predict which transactions would result in delayed payments in the received payments dataset

COMPARISON BETWEEN TWO MODELS, 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: 0.9394084954678357
```

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

Random Forest found better than Logistic Regression

It can be observed that the overall precision and recall scores of the Random forest model far- exceeded the logistic regression model. Also, recall scores were more important in this case since it was important to increase the percentage prediction of late payers to be targeted

Since the data is heavy on categorical variables, random forest is better suited to the job than logistic regression

Therefore, random forest model was finalised to be the model of choice and go forward with predictions

Random Forest Feature Ratings

```
Feature ranking:
1. USD Amount (0.465)
2. Invoice_Month (0.130)
3. 60 Days from EOM (0.113)
4. 30 Days from EOM (0.105)
5. cluster_id (0.053)
6. Immediate Payment (0.042)
7. 15 Days from EOM (0.027)
8. 30 Days from Inv Date (0.015)
9. 60 Days from Inv Date (0.013)
10. 90 Days from Inv Date (0.008)
11. INV (0.007)
12. 90 Days from EOM (0.006)
13. 45 Days from EOM (0.006)
14. CM (0.004)
15. 45 Days from Inv Date (0.004)
16. DM (0.001)
```

The random forest was then used to find out the feature rankings which shows that the top features to predict delay which included

USD Amount
Invoice Month
60 Days from EOM(Payment Term variable)
30 Days from EOM (Payment Term variable)

The customers segmented with cluster ID was then applied to the open-invoice data as per the customer name and predictions were made

CUSTOMERS WITH THE HIGHEST DELAY PROBABILITIES

Customer_Name	Delayed_Payment	Total_Payments	Delay%
LUXU Corp	37	37	100.0
FEND Corp	20	20	100.0
PARI Corp	14	14	100.0
IL G Corp	6	6	100.0
VALE Corp	6	6	100.0
SOFI Corp	6	6	100.0
CAIR Corp	4	4	100.0
MICH Corp	3	3	100.0
DOLC Corp	2	2	100.0
WELL Corp	2	2	100.0

Predictions suggest that the companies presented in the table to the left has the maximum probability of default with maximum number of delayed and total payments

####RECOMMENDATIONS

From our clustering analysis we can make the following inference

Credit Note Payments observe the greatest delay rate compared to Debit Note or Invoice type invoice classes, hence company policies on payment collection could be made stricter around such invoice classes

Goods type invoices had significantly greater payment delay rates than non-goods types and hence can be subjected to stricter payment policies

Since lower value payments comprise of the majority of the transactions, also late payments are seen more on lower value payments, it is recommended to focus more on those. The company can apply penalties depending on billing amount, the lesser the bill, the greater the percentage of penalty on late payments. Of course this has to be last resort

Customer segments were clustered into three categories, viz., 0,1 which mean medium, prolonged and early payment duration respectively. It was found that customers in cluster 1 (prolonged days) had significantly greater delay rates than early and medium days of payment, hence cluster 1 customers should be paid extensive focus

The companies with the greatest probability and total & delayed payment counts should be first priority and should be focused on more due to such high probability rates