SCHUSTER CASE STUDY

Submitted By :

Sakshi Priyadarshi

Sanat

Sajid Ahmed

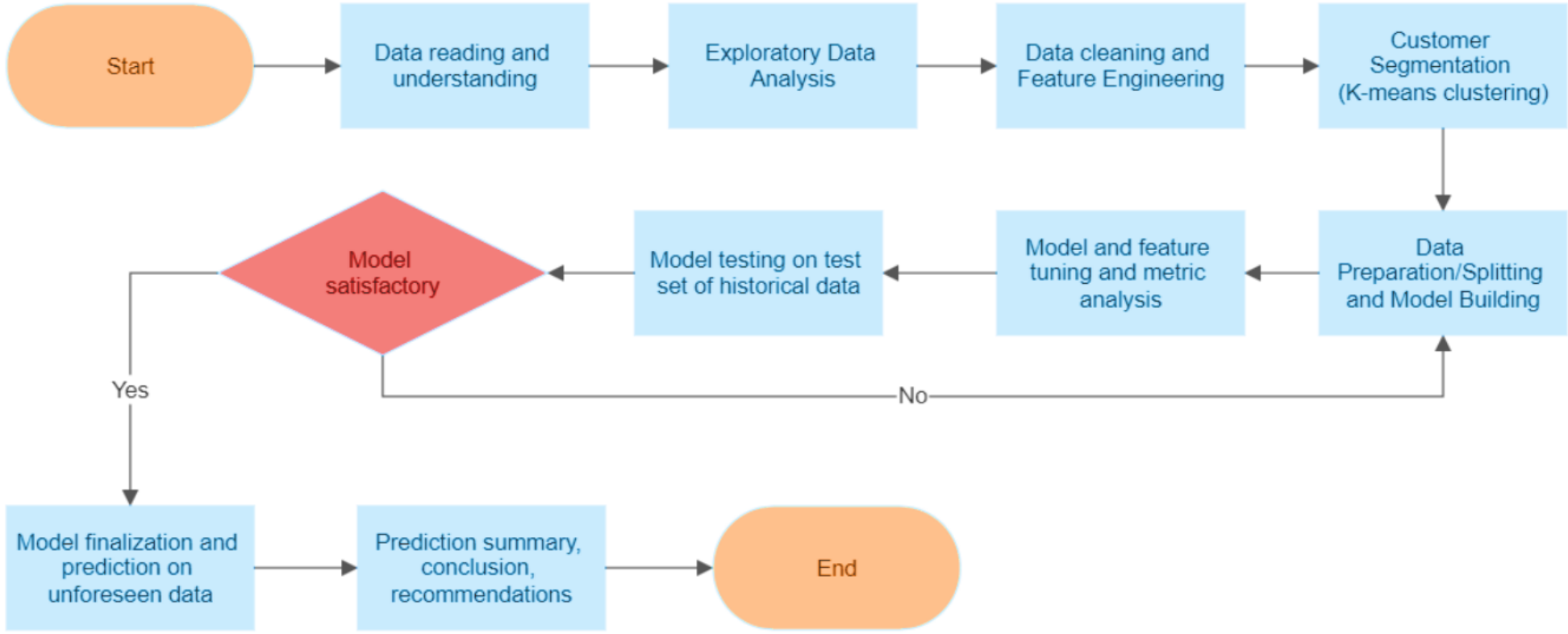
Problem Statement

Schuster, a multinational retail company, faces challenges with late payments from vendors, resulting in non-value-added activities and financial impact. The company seeks to understand vendor payment behavior, predict the likelihood of late payments against open invoices, and prioritize collections efforts efficiently.

Objectives

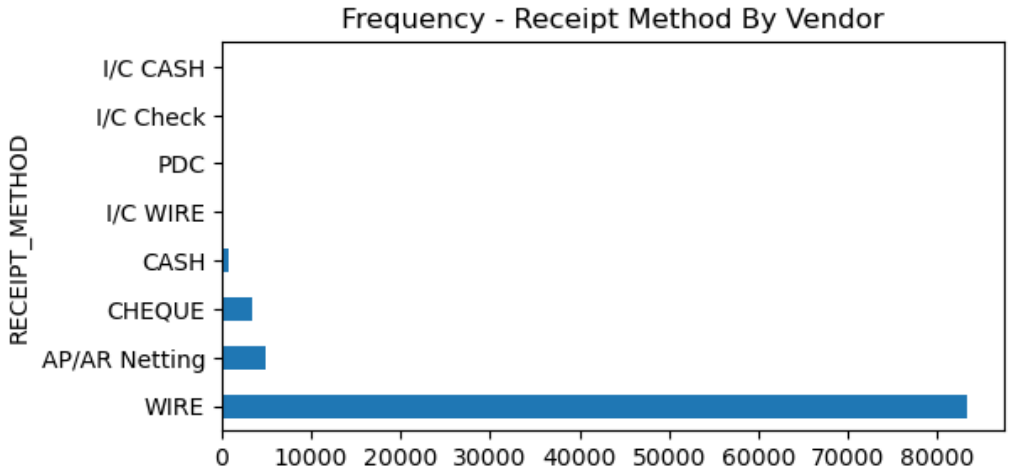
- Analyze customer transaction data to identify various payment patterns.
- Segment customers based on their historical payment behaviors.
- Develop a predictive model to forecast the probability of delayed payments for open invoices from customers.
- Extract actionable insights from the developed model.

Problem Approach

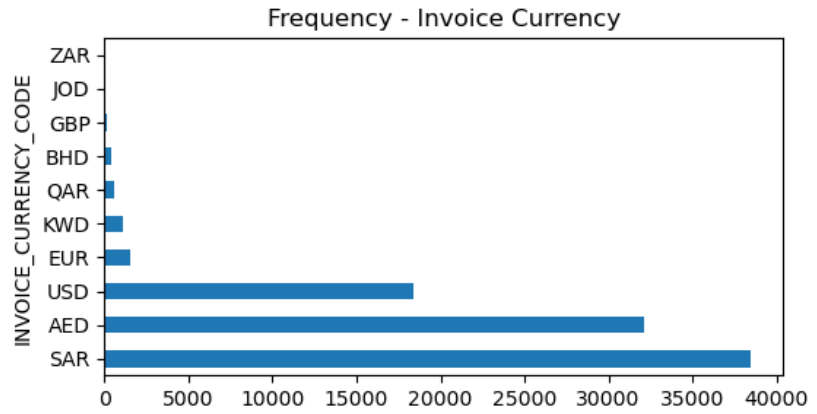


Insights from Data

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

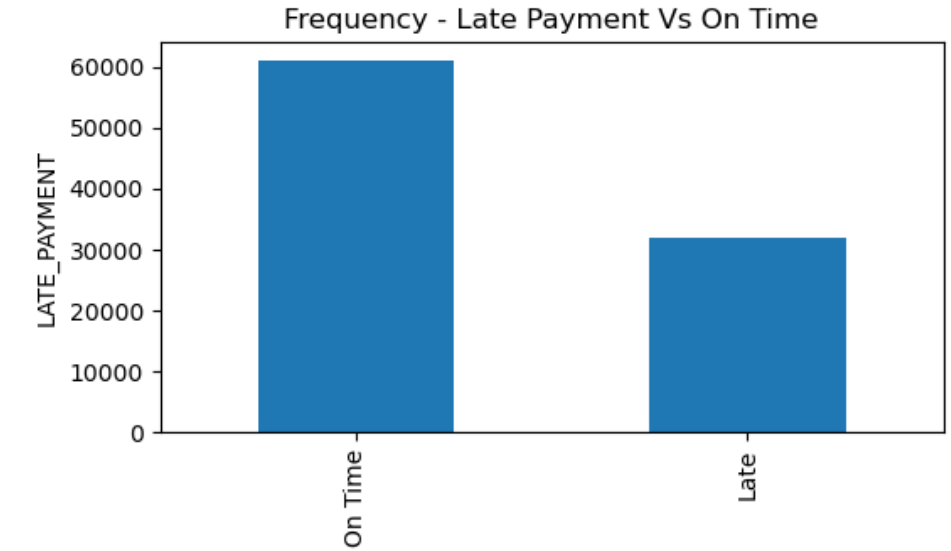


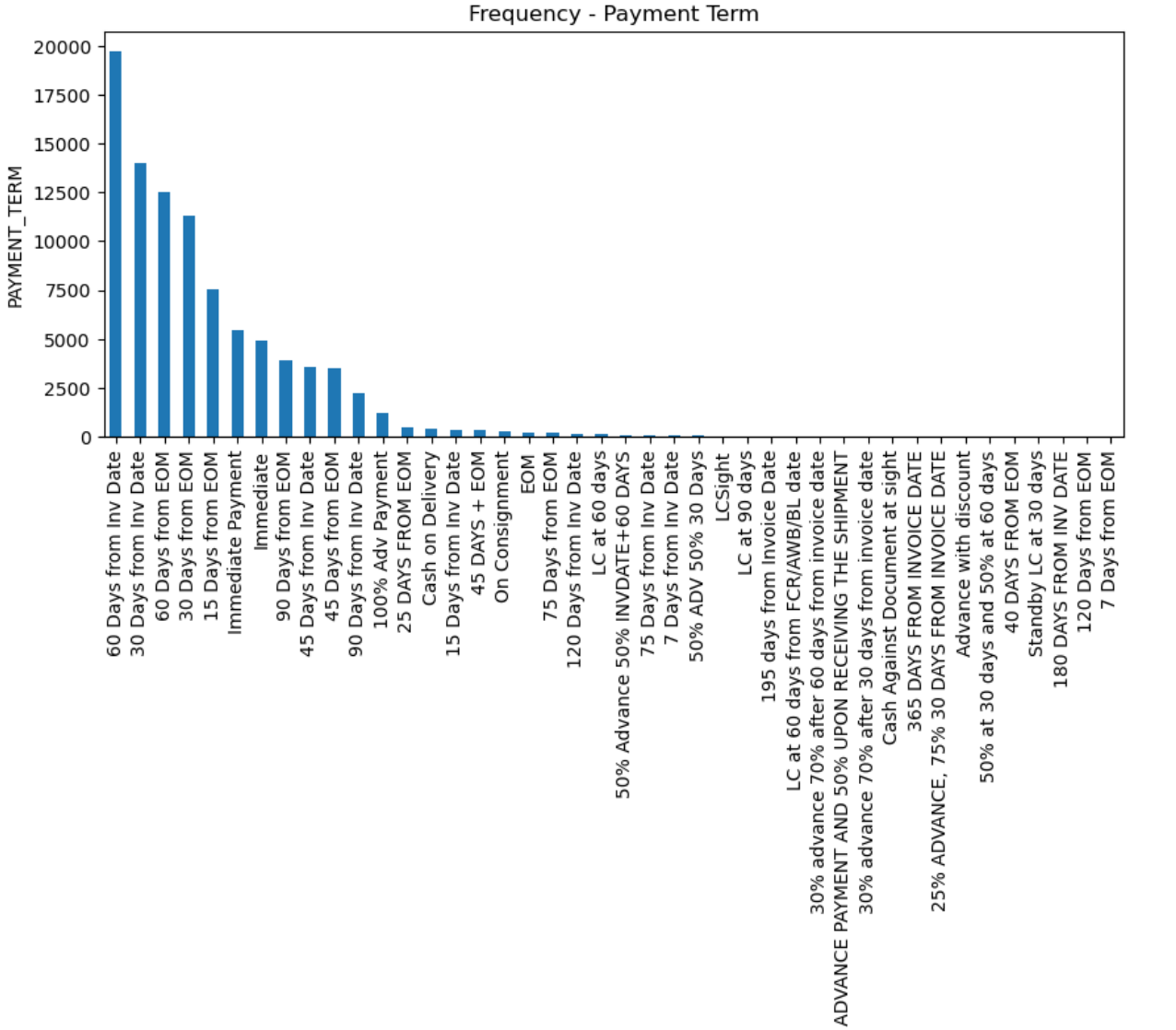
SAR, AED and USD are the top three currencies in which the company deal.



Insights from Data

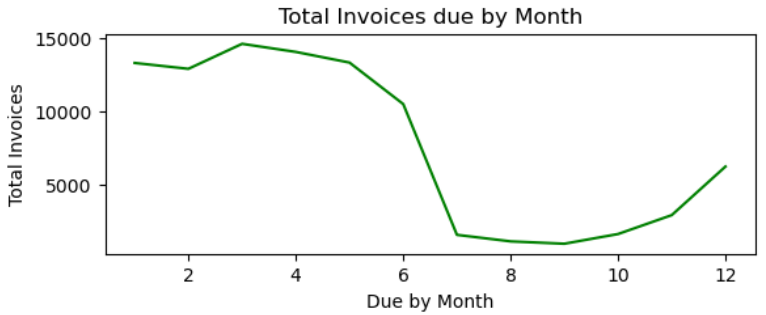
- Most of the invoice has “60 days from the Inv Date “ has the payment term.
- 34.34% of payments are delayed.

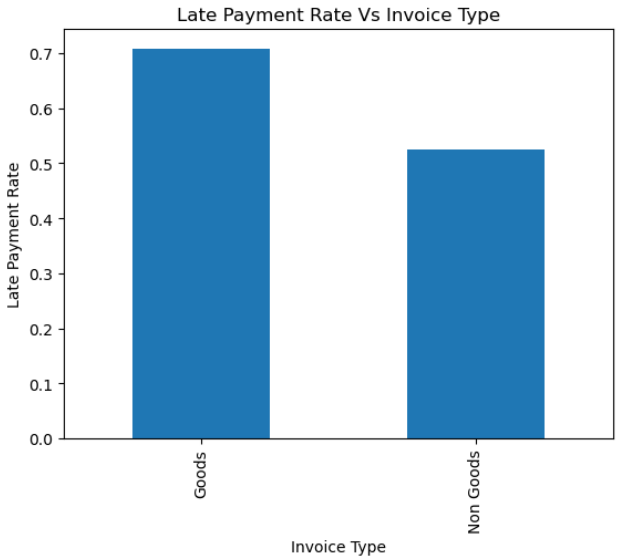


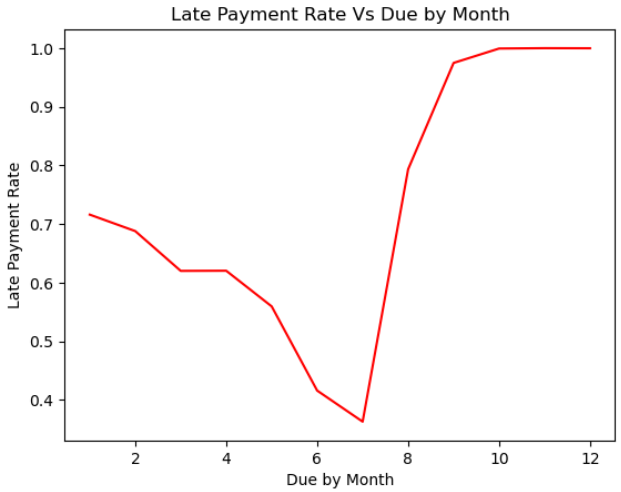


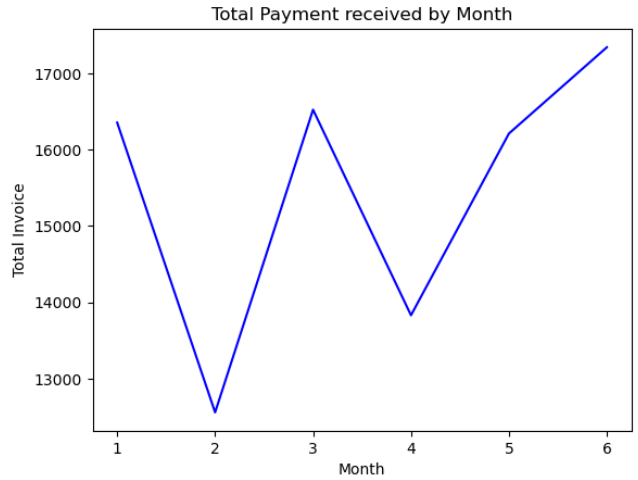
Insights from Data

- 70.8% of “INVOICE_TYPE” as “Goods” are delayed.
- Mostly invoices are due in first quarter
- Payment rate was delayed after second quarter significantly with a sudden spike.





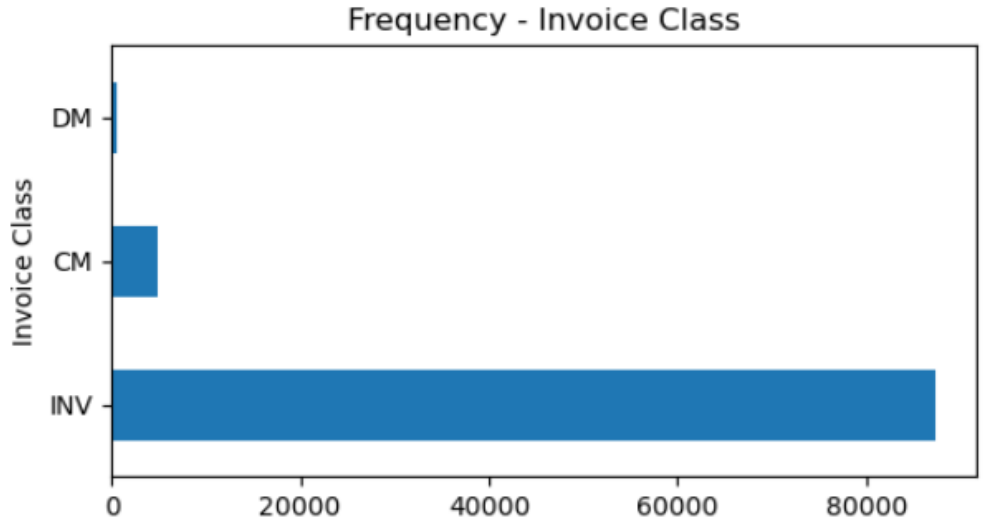




Feature Engineering

- Post data imputation, top 95% of Invoices falling under following Payment Terms
- Created Dummy variables for ‘Payment_Term’ and ‘Invoice_class’
- First combined similar payment terms and then clubbed every other payment term except top 10

30 Days from Inv Date	0.208324
60 Days from Inv Date	0.408791
Immediate Payment	0.593715
60 Days from EOM	0.686998
Immediate	0.770861
30 Days from EOM	0.831402
90 Days from EOM	0.861526
90 Days from Inv Date	0.889234
15 Days from EOM	0.909959
75 Days from EOM	0.928654
45 Days from Inv Date	0.946352

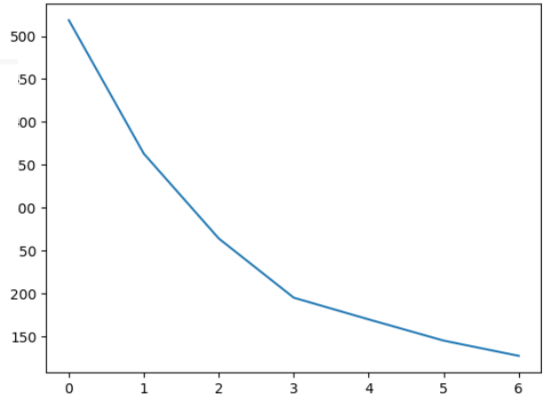


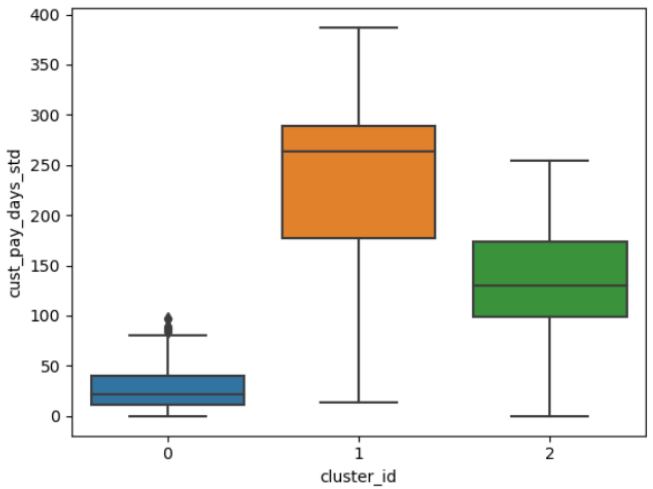
K-MEANS Clustering

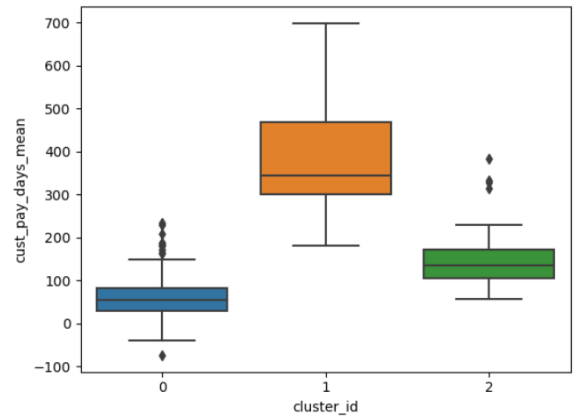
One of the objectives was to categorize customers to understand payment behaviors which was achieved by K-means clustering using average and standard deviation of number of days it took for the vendor to make payment.

- The number of clusters were decided to be 3 since with increase in clusters post 3, there was a not significant change in silhouette score.
- The category 0 were early payers with least number of average days taken to pay and category 1 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
- It was also observed that prolonged players historically have significantly greater rates of delay in payment than early or medium duration payment transactions

```
For n_clusters=2, the silhouette score is 0.7019665026916853
For n_clusters=3, the silhouette score is 0.5952324617364421
For n_clusters=4, the silhouette score is 0.6141526990951806
For n_clusters=5, the silhouette score is 0.4274877610244625
For n_clusters=6, the silhouette score is 0.42220954720479476
For n_clusters=7, the silhouette score is 0.3991359850694904
For n_clusters=8, the silhouette score is 0.4028881474256595
```

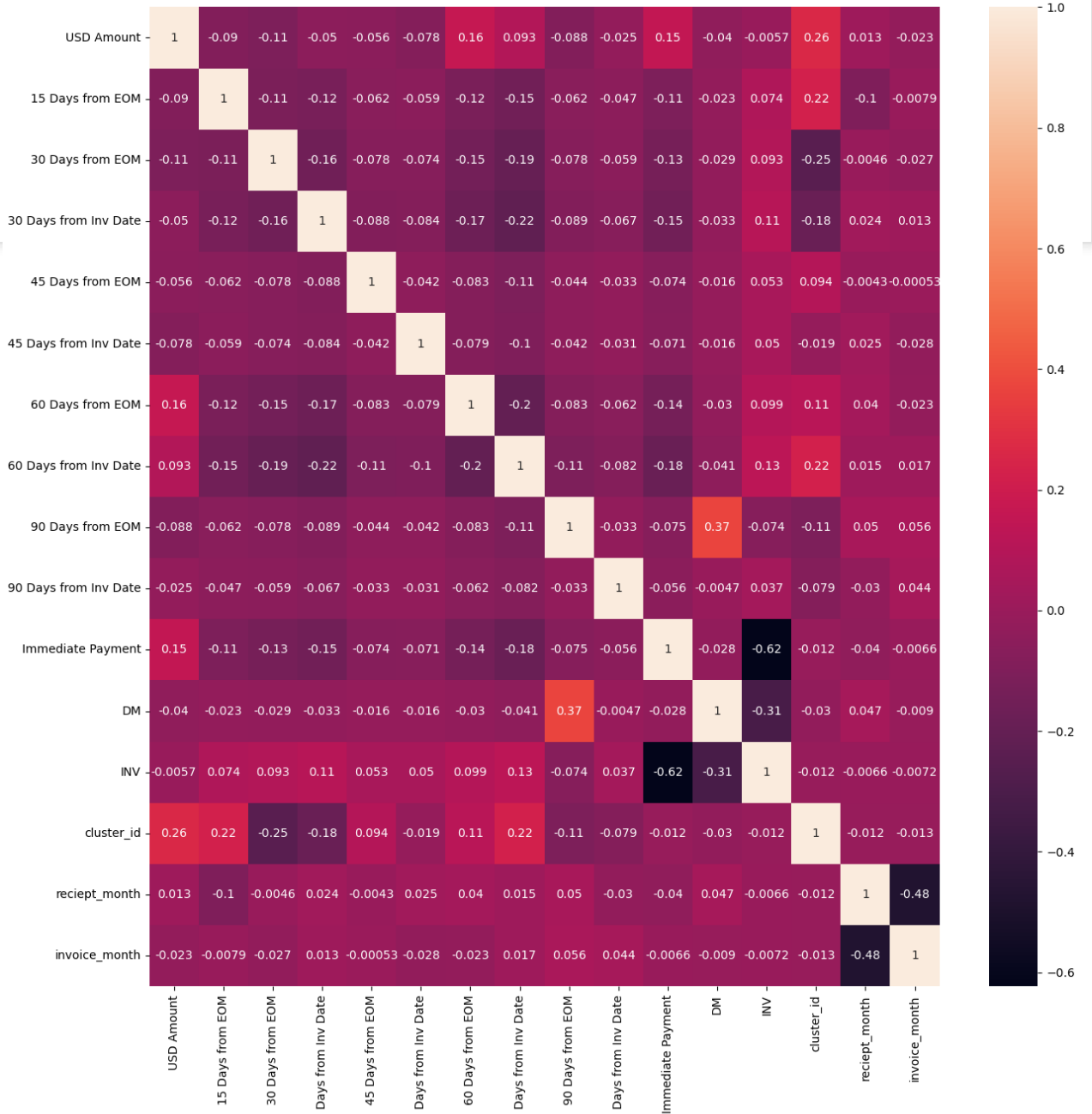






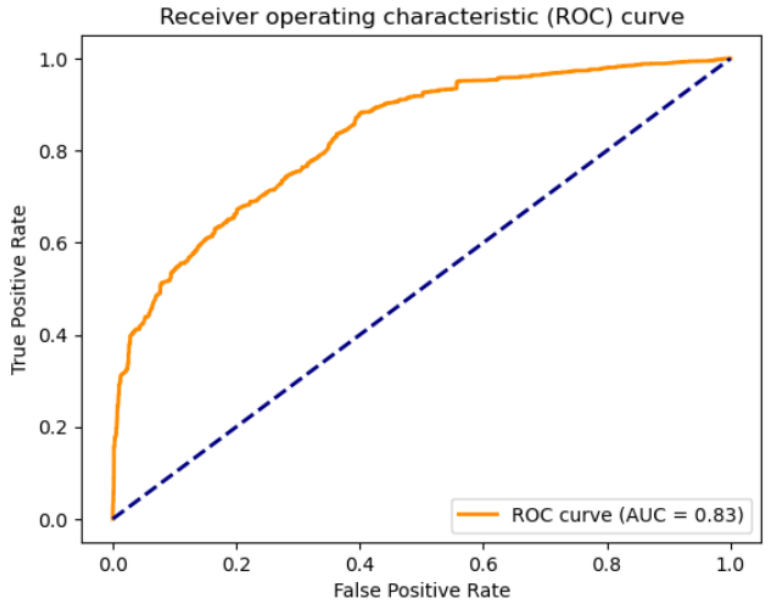
Model Building

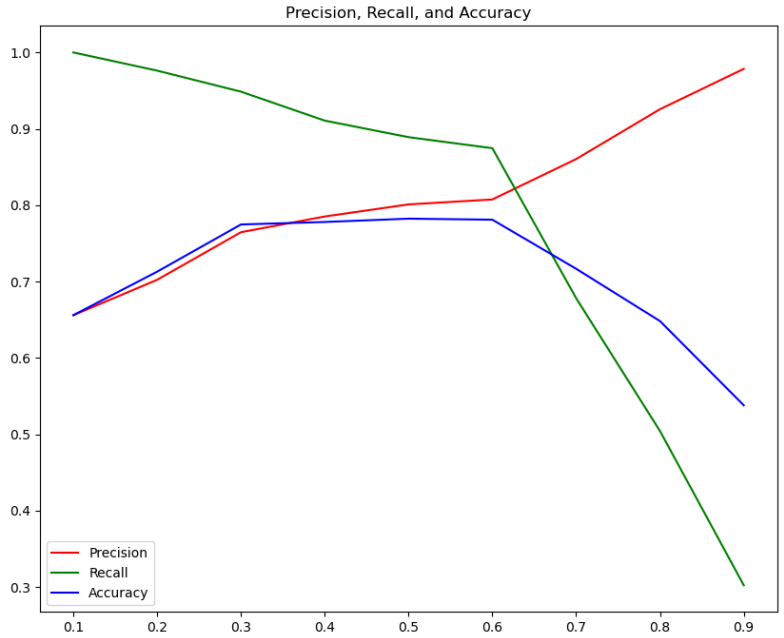
CM & INV, INV & Immediate Payment, DM & 90 days from EOM has high multicollinearity, hence dropping these columns to prevent multicollinearity effect



Logistic Regression Model

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 ~0.6, which was used to further predict which transactions would result in delayed payments in the received payments dataset

	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Precision	0.655995	0.702319	0.764634	0.785216	0.801044	0.807535	0.860422	0.925671	0.978510
Recall	1.000000	0.976207	0.948660	0.910742	0.889084	0.874630	0.678235	0.504364	0.302361
Accuracy	0.655995	0.712962	0.774763	0.778026	0.782382	0.781012	0.716749	0.648298	0.537996

Evaluation Metrics

Logistics Regression

Train Data

	precision	recall	f1-score	support
0	0.73	0.58	0.65	22349
1	0.80	0.89	0.84	42618
accuracy			0.78	64967
macro avg	0.77	0.73	0.74	64967
weighted avg	0.78	0.78	0.78	64967

Test Data

	precision	recall	f1-score	support
0	0.73	0.58	0.65	9529
1	0.80	0.89	0.84	18315
accuracy			0.78	27844
macro avg	0.77	0.74	0.75	27844
weighted avg	0.78	0.78	0.78	27844

- The Model produces almost similar result for both train & test data, which is indicative of the expectation of correct model
- Model shows linear relation between Probability & Features, which can be used to get the relation

Random Forest

Train Data

	precision	recall	f1-score	support
0	0.97	0.91	0.94	22349
1	0.95	0.98	0.97	42618
accuracy			0.96	64967
macro avg	0.96	0.95	0.95	64967
weighted avg	0.96	0.96	0.96	64967

Test Data

	precision	recall	f1-score	support
0	0.91	0.85	0.88	9529
1	0.93	0.96	0.94	18315
accuracy			0.92	27844
macro avg	0.92	0.90	0.91	27844
weighted avg	0.92	0.92	0.92	27844

- RF Model has high precision, recall & accuracy.
- Focus is on the recall of positive class in both train & test data
- The RF model is able to identify 92% of all positive classes with below mentioned hyper parameters

```
Best hyperparameters: {'max_depth': 30, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 150}
Best f1 score: 0.9546570200986905
==================================
```

Logistic Regression or Random Forest- Which one is better?

- 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 than logistic regression, however we can use both the classification model to check for the performance and interpretability.

```
Sensitivity =  0.8881458538645642
Specificity =  0.588661685086581
False Positive Rate =  0.41133831491341893
Positive Predictive Value =  0.8045871949664144
Negative Predictive Value =  0.734028901411594
```

Feature Ranking

The random forest was then used to find out the feature rankings which shows that the top 5 features to predict delay which included

- USD Amount
- Invoice Month
- Reciept Month
- 60 Days from EOM (Payment Term variable)
- 30 Days from EOM (Payment Term variable)
- Cluster-ID

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

```
Feature ranking:
1. USD Amount (0.235)
2. invoice_month (0.202)
3. reciept_month (0.140)
4. 60 Days from EOM (0.114)
5. 30 Days from EOM (0.106)
6. cluster_id (0.053)
7. Immediate Payment (0.045)
8. 15 Days from EOM (0.028)
9. 60 Days from Inv Date (0.016)
10. 30 Days from Inv Date (0.015)
11. 90 Days from EOM (0.013)
12. 90 Days from Inv Date (0.011)
13. INV (0.008)
14. 45 Days from EOM (0.007)
15. 45 Days from Inv Date (0.006)
16. DM (0.001)
```

Recommendations

From our analysis we can make the following inference :

- High value on prob_RF implies that customer has high probability of late payment
- 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 and 2 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 in Fig 1. with the greatest ability and total & delayed payment counts should be first priority and should be focused on more due to such high probability rates

Customer_Name	prob_LR	prob_RF
2H F Corp	0.0802	0.626667
3D D Corp	0.0000	0.255999
6TH Corp	0.0465	0.185782
ABDU Corp	0.0000	0.403053
ABEE Corp	0.4145	0.520000
...	...	...
ZAIN Corp	0.2711	0.730900
ZALL Corp	0.1761	0.283148
ZALZ Corp	0.0006	0.570705
ZINA Corp	0.1955	0.095556
ZUHA Corp	0.1716	0.252995

495 rows × 2 columns

Thank You!