E-Commerce and Retail B2B Case Study

Sanchari Saha
Sanket Shrivas
Sania Kapoor

Handling the issue and explaining intention

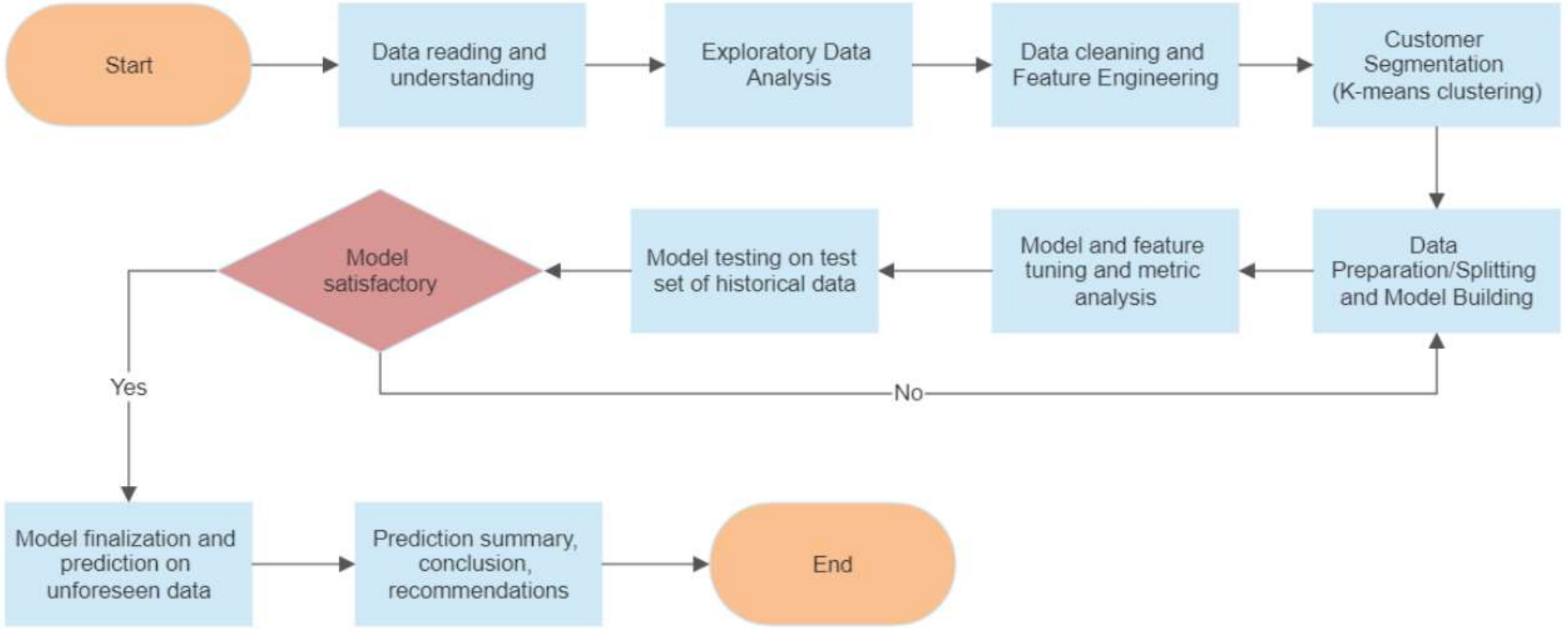
Identifying the issue:

- Schuster is a sports retail company operating through business-to-business transactions.
- They frequently deal with vendors who may not adhere to agreed payment deadlines.
- Delayed payments from vendors result in financial setbacks and operational disruptions.
- Employees spend considerable time chasing overdue payments, diverting resources from value-added activities.
- This inefficiency leads to wasted time and resources, complicating business operations further.

Business intention:

- Implement customer segmentation to analyze and comprehend customer payment patterns effectively.
- Utilize historical data to forecast delayed payments for transactions with pending due dates.
- Enhance resource allocation and expedite credit recovery through accurate prediction models.
- Reduce low-value activities by optimizing processes based on predicted payment behaviors.

Resolution Approach



Univariate Analysis of Class Disparity and transaction Trends

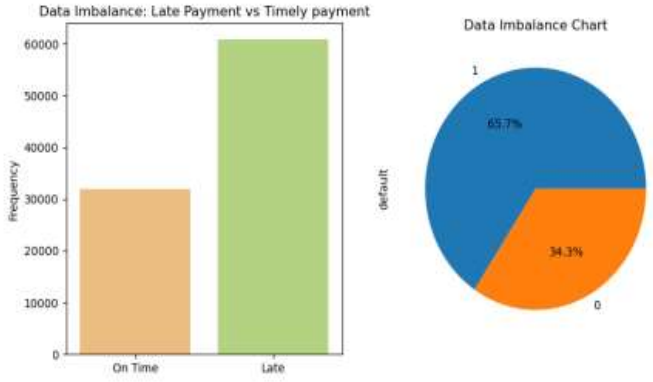


Fig. 1

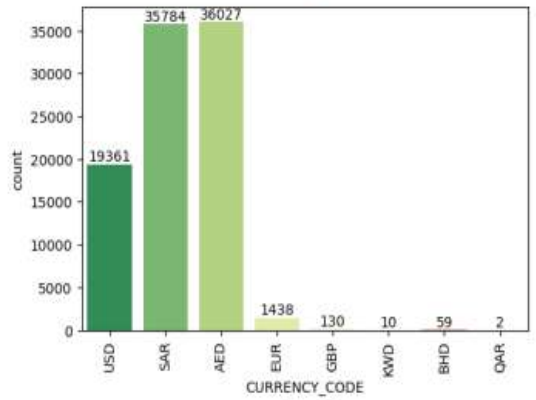


Fig. 2

From Figure 1 and 2:

- Class imbalance stands at 65.7%, predominantly favoring payment delayers, indicating an imbalance level considered acceptable without requiring specific corrective measures.

- The primary currencies transacted by the company are AED, SAR, and USD, with AED being the most prevalent. This observation implies a higher volume of transactions occurring with the Middle-Eastern region.

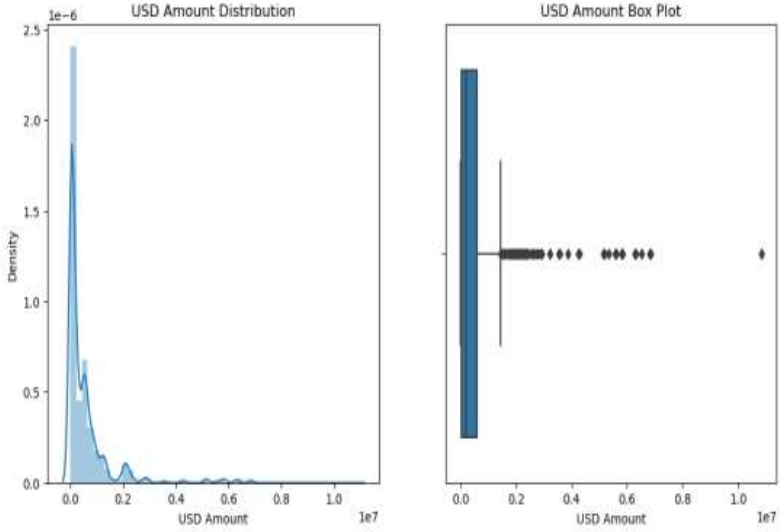
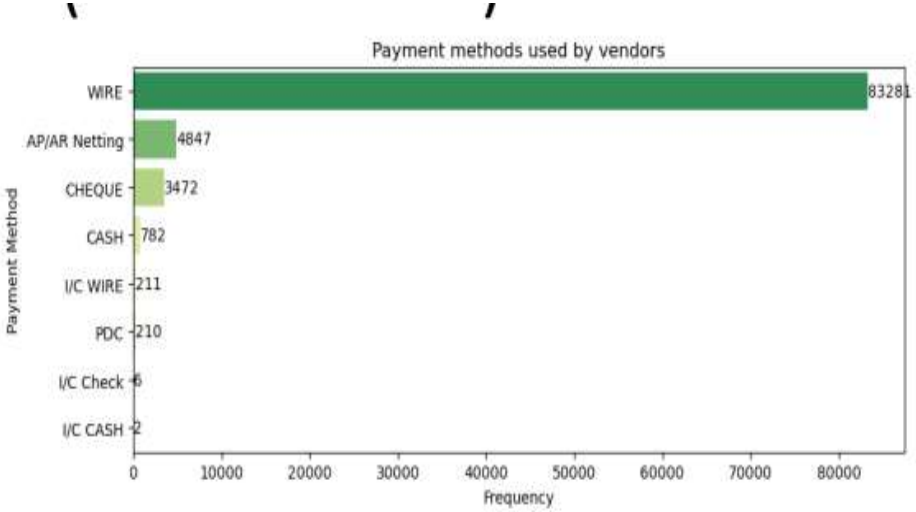


Fig. 1

From Figure 3, it's evident that:

- Transaction values typically fall within the range of $1 to $3 million.
- The highest frequency of transaction values occurs below approximately $1.75 million.

Univariate Analysis of Class Disparity and transaction Trends



From Fig. 1, we observe,

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

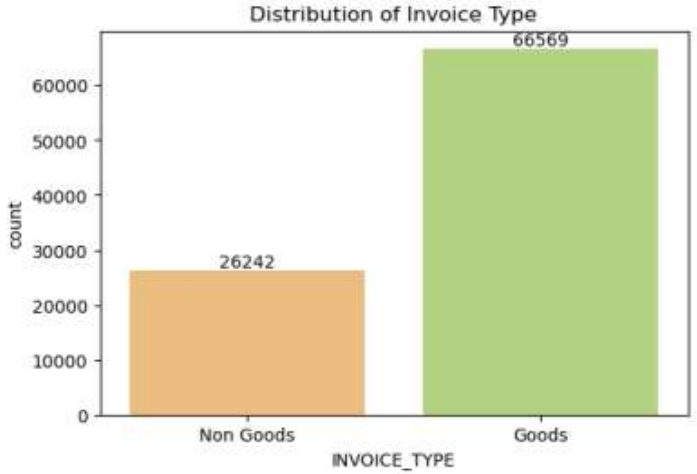


Fig. 2

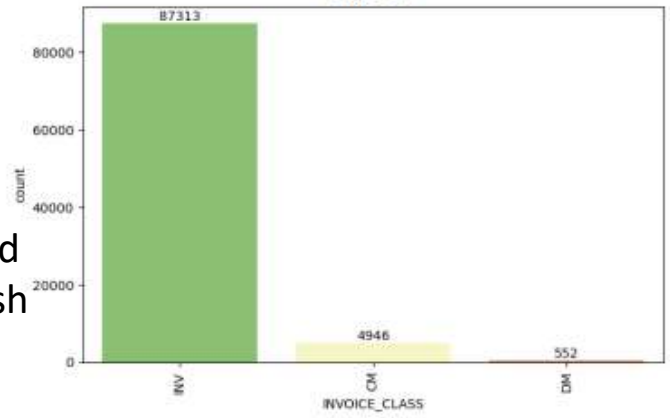


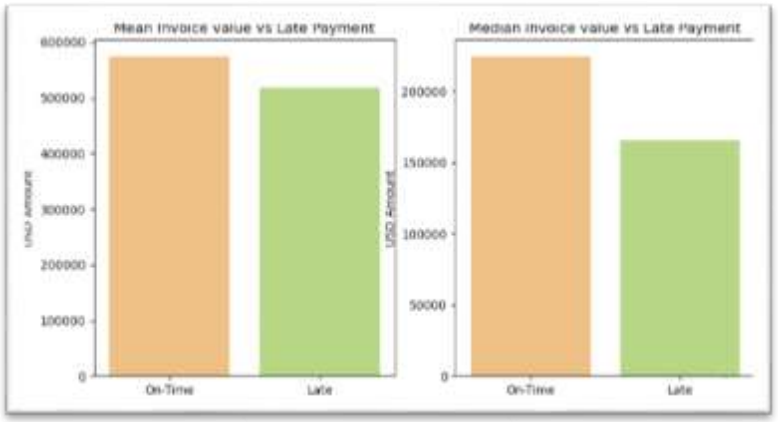
Fig. 3

From Figures 2 and 3, the following observations can be made:

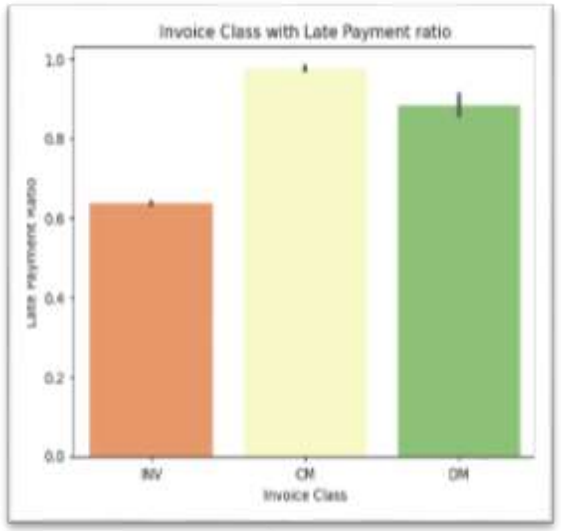
- Invoices related to goods type transactions represent the majority of invoices generated by the company.

- Among the invoice classes, 'Invoice' dominates significantly, with other classes contributing only a small percentage of the overall share.

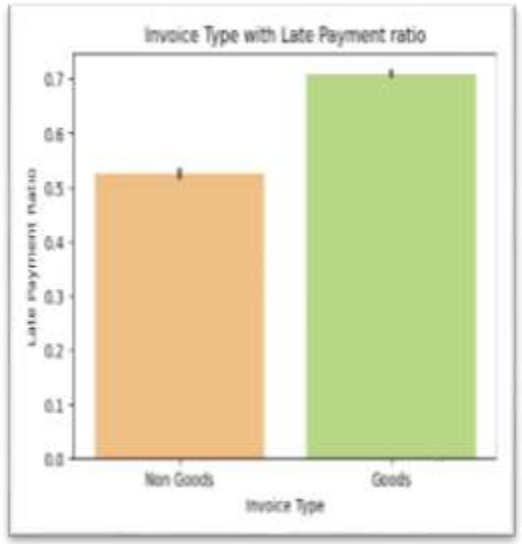
Identifying the characteristics of defaulter payment types



In this first figure we can see that the mean and the median amount is higher for the payers who are paying on time than who are paying late, suggesting that high value transactions shows less risk as compared to low value transactions.



This figure tells us that late payment ratio for Credit Note transaction types are highest, followed by Debit note and Invoice. Which implies that there is a higher delay risk Credit and Debit Note as compared to Invoice



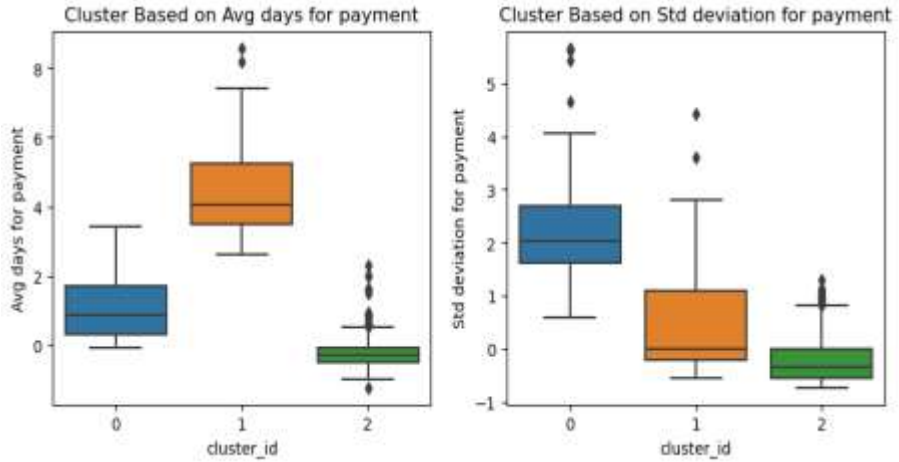
Invoice types for Goods show a greater late payment ratio as compared to Non Goods, which suggests that there is a high chance of late payment delay

Customer segmentation using K means method

- For n_clusters=2, the silhouette score is 0.7557759850933141
- For n_clusters=3, the silhouette score is 0.73503646233166
- For n_clusters=4, the silhouette score is 0.6182691953064194
- For n_clusters=5, the silhouette score is 0.6209288452882942
- For n_clusters=6, the silhouette score is 0.40252553894618837
- For n_clusters=7, the silhouette score is 0.4069490441271981
- For n_clusters=8, the silhouette score is 0.4151884768372497

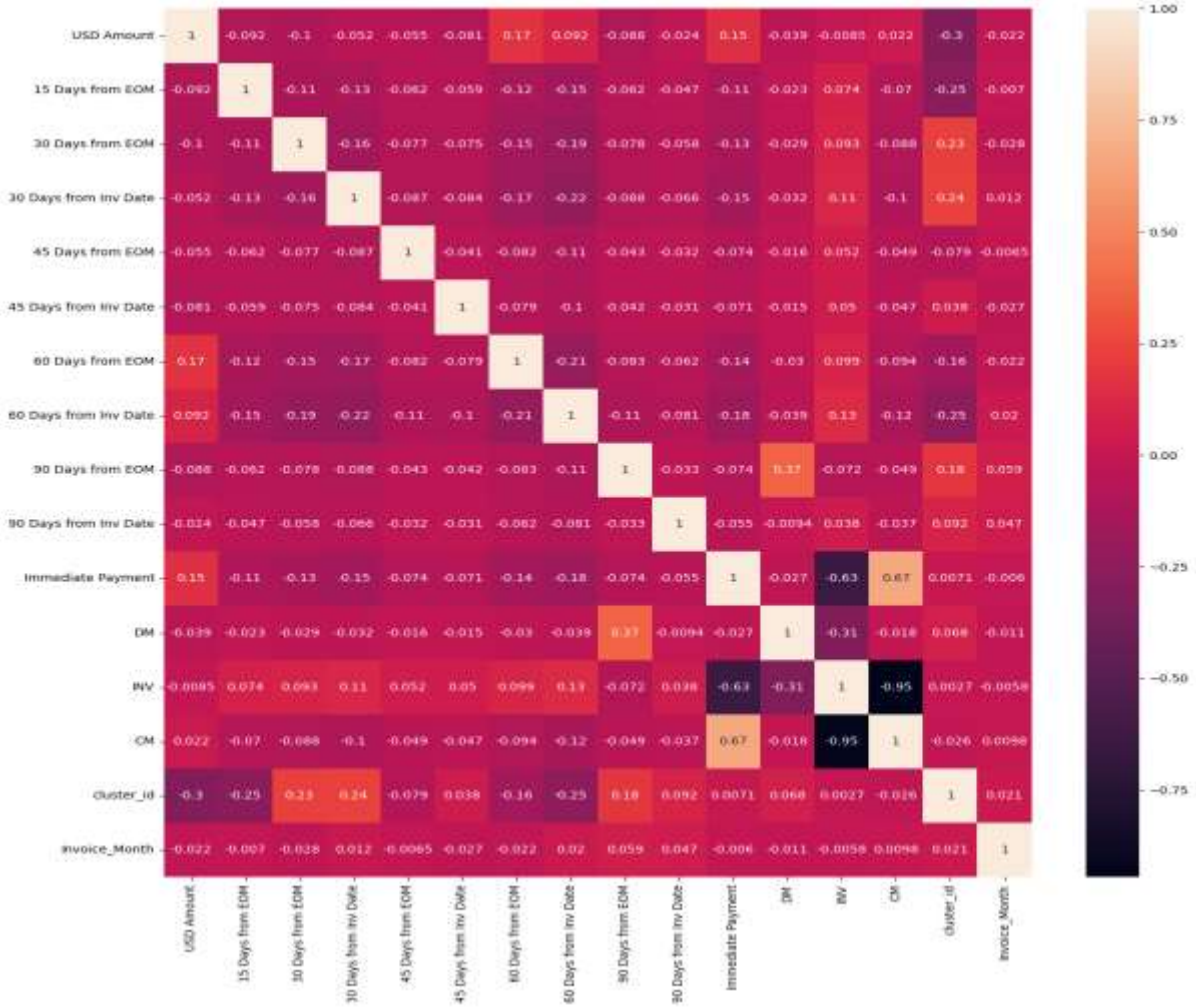
One of the primary reason was to create a category to understand the customer payment behavior, which was achieved using K-mean clustering method, using average and standard deviation based on the number of days it took to complete the payment.

- The number of clusters were decided to be 3 as there was a significant dip in the silhouette score post 3 clusters

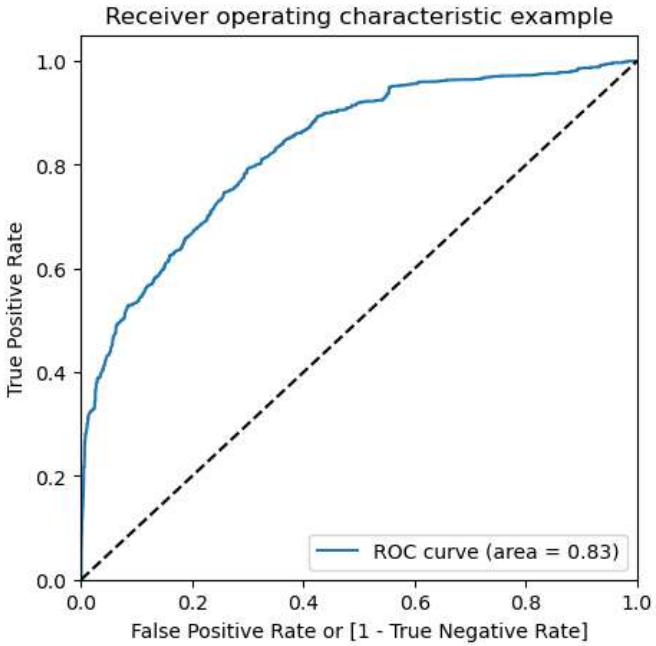


In category 2 there were early payers with least number avg days taken for payment, in category 1 there were prolonged payers with highest number of days taken to pay. Whereas category 0 lies somewhere between the other two category hence titled as medium duration payers.

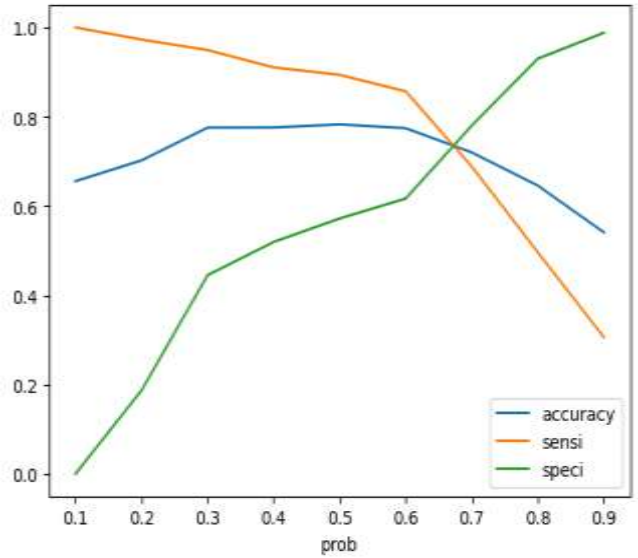
Model Building



CM and INV, INV and Immediate Payment, DM and 90 days from EOM have high multicollinearity, therefore we are dropping these columns to avoid multicollinearity effect



- We created a logistic regression model after dropping all the unnecessary variables and multicollinearity variables, which resulted in giving us the variables with acceptable p-values and VIF figures. Hence kept the remaining features with no further elimination and also a good ROC curve area of 0.83



- The trade of plot between accuracy, sensitivity, and specificity showed us an optimum probability cutoff of ~0.6, which helped us to predict which transaction would result in delay payment in the received payment dataset.

Comparison between the two models, logistic regression and random forests

- A random forest was build using the same parameters as the logistic regression with hyper- parameter tuning, which gave us the following parameters

```
Best hyperparameters: {'max_depth': 30, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 150}
Best f1 score: 0.9394084954678357
```

  - The above parameter helped us to create a random forest model, whose metrics were compared to the logistic regression model and the final model was finalized

Random Forest found better than Logistic Regression

```
# Let's check the overall accuracy.
accuracy_score(y_pred_final.default, y_pred_final.final_predicted)
```

```
0.7754632955035196
```

```
#precision score
precision_score(y_pred_final.default, y_pred_final.final_predicted)
```

```
0.8115658179569116
```

```
# Recall Score
recall_score(y_pred.default, y_pred.final_predicted)
```

```
0.8569416073818412
```

Fig. 1 (Logistic Regression Metrics - Test Set)

	precision	recall	f1-score	support
0	0.92	0.85	0.88	9502
1	0.93	0.96	0.94	18342
accuracy			0.92	27844
macro avg	0.92	0.91	0.91	27844
weighted avg	0.92	0.92	0.92	27844

Fig. 2 (Random Forest Metrics - Test Set)

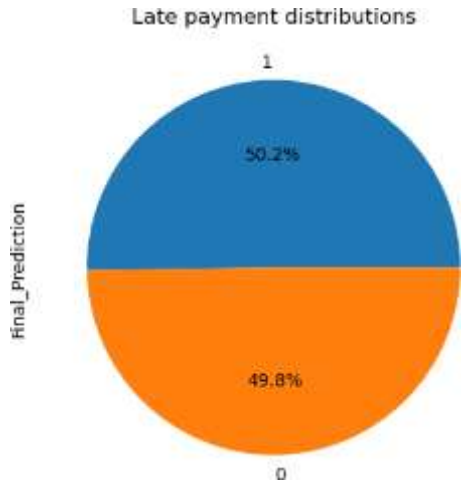
- It can be observed that the overall precision and recall scores of the Random forest model has far-exceeded the logistic regression model. Also, recall scores were more essential in this case since it was important to increase the percentage prediction of late payers to be targeted.
- Since the data is heavy on categorical variables, randomforest is better suited for this job than logistic Regression.
- Therefore, randomforest model was finalized to be the model of choice and go forward with predictions.

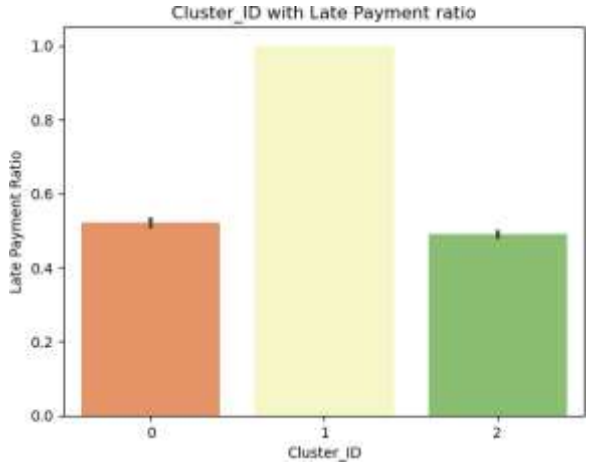
Random Forest Feature Ratings

```
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 used after that to find out the feature rankings which shows that the top 5 features to predict delay includes -
  - USD Amount
  - Invoice Month
  - 60 Days from EOM (Payment Term variable)
  - 30 Days from EOM (Payment Term variable)
  - Cluster-ID (which in turn is dependent on average and standard deviation of days required to make payment)
- The customers clubbed with cluster ID were then applied to the open-invoice data as per the customer name and then, predictions were made on that.

50% payments were predicted to be delayed as per Open- invoice data, prolonged payment days contributed to alarmingly high delay rates.





- Predictions made by the final model suggests that there are probable 50.2% transactions where payment delays can be expected, which can cause a shocking lag to business operations.

- Customer segment with historically prolonged payment days are anticipated to have the most delay rate (~100%) than historically early or medium days payment transactions, this is similar to the results found based on historical outcomes as well.

Customers with the highest delay probabilities

Customer_Name	Delayed_Payment	Total_Payments	Delay%
ALSU Corp	7	7	100.0
LVMH Corp	4	4	100.0
MILK Corp	3	3	100.0
MUOS Corp	3	3	100.0
MAYC Corp	3	3	100.0
ROVE Corp	3	3	100.0
AMAT Corp	3	3	100.0
TRAF Corp	3	3	100.0
CITY Corp	3	3	100.0
DAEM Corp	3	3	100.0

- Predictions here 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

Customer_Name	Delayed_Payment	Total_Payments	Delay%
ALSU Corp	7	7	100.0
LVMH Corp	4	4	100.0
MILK Corp	3	3	100.0
MUOS Corp	3	3	100.0
MAYC Corp	3	3	100.0
ROVE Corp	3	3	100.0
AMAT Corp	3	3	100.0
TRAF Corp	3	3	100.0
CITY Corp	3	3	100.0
DAEM Corp	3	3	100.0

Fig. 1

From our clustering analysis we can make the following inferences-

•Credit Note Payments observe the greatest delay rates compared to the Debit Note or Invoice type invoice classes, hence company policies on payment collection could be 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 to reduce the delays.

•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 this segment. The company can apply penalties depending on billing amount, the lesser the bill, the greater the percentage of penalty on late payments. This has to be the last resort, of course.

•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 extensively focussed.

- The companies in Fig 1. with the greatest probability and total & delayed payment counts should be the first priority and should be focused more due to such high probability instances.