E-Commerce and Retail B2B Case Study

- Niharika Trivedi
- Abhishek Sharma

Problem Statement

Schuster, a multinational retailer of sports goods and accessories, conducts extensive business with numerous vendors under credit arrangements. However, some vendors delay payments, leading Schuster to impose late fees, which are not conducive to long-term partnerships. Currently, employees spend significant time following up on payments, resulting in non-value-added activities, time loss, and financial impact. Schuster aims to analyze customer payment behavior to predict late payments on open invoices.

Objective

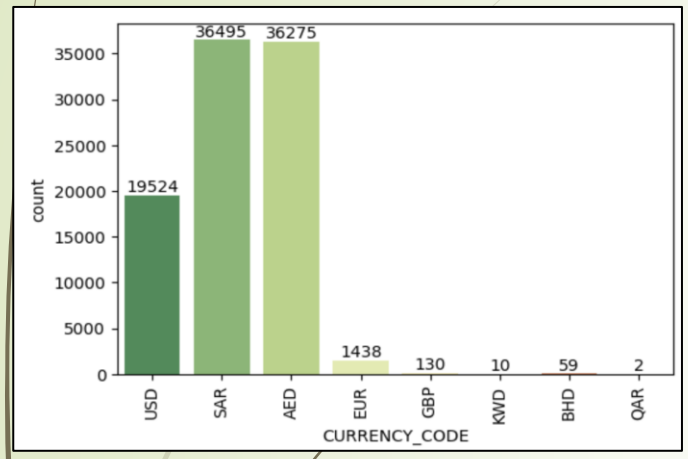
- Schuster aims to better understand customer payment behavior through past payment patterns (customer segmentation).
- Using historical data, it seeks to predict the likelihood of delayed payments on open invoices.
- This information will help collectors prioritize follow-ups to ensure timely payments.

Approach

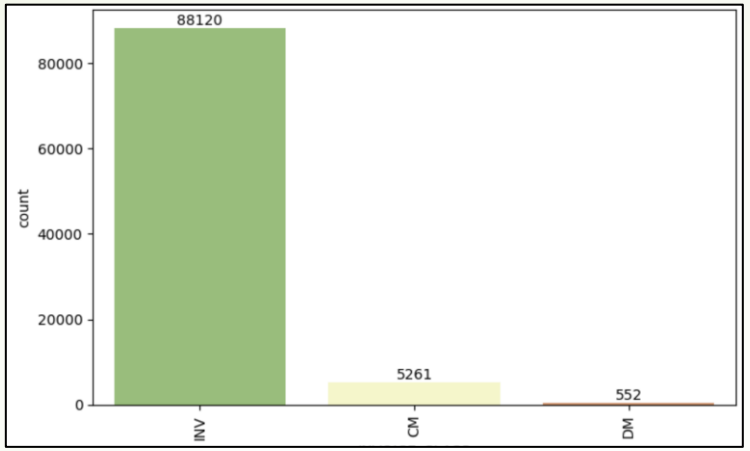
1. Reading and Understanding the data
2. Exploratory Data Analysis
   - Data imbalance check
   - Creating derived metrics
3. Data Cleaning and Feature Engineering
   - Delete null values
   - Dropping columns which contains only one value
   - Dropping duplicated columns
   - Dropping columns which are not important for the analysis
4. Customer Segmentation (K-means Clustering)
5. Data Preparation/Splitting and Model Building
6. Model and Feature Tuning
7. Model Testing on test set of historical Data
8. Model Satisfaction
9. Model Evaluation
10. Prediction Summary, conclusion and Recommendation.

Exploratory Data Analysis and Data Imbalance

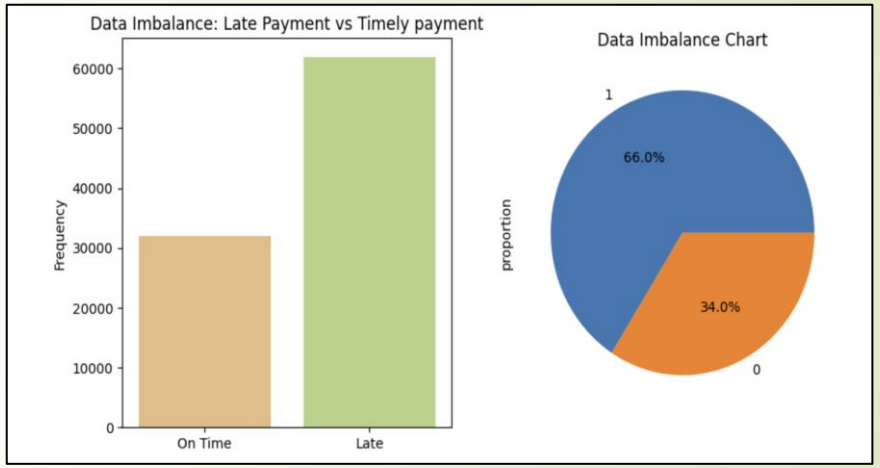
(univariate)



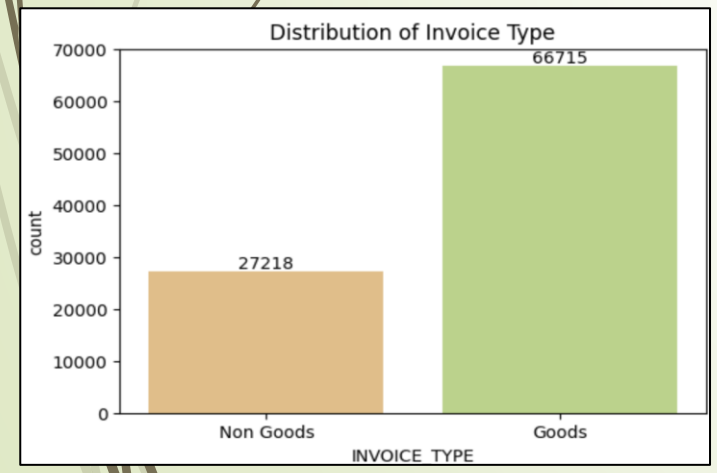
Currency code: Most Bill Payments are in USD, SAR and AED



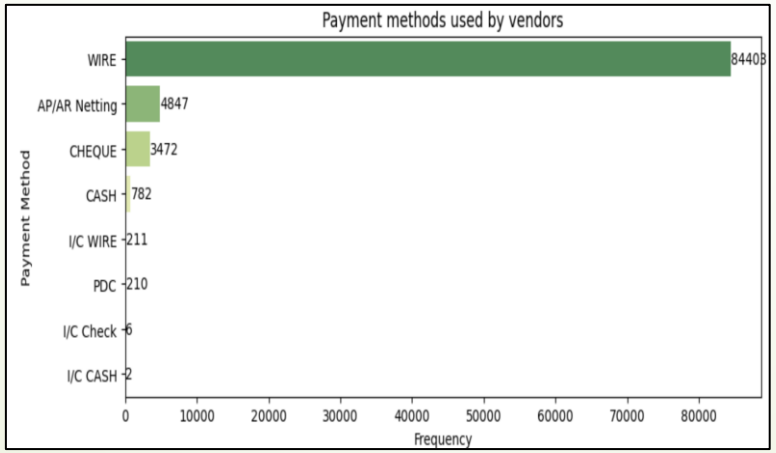
INVOICE_CLASS: INV has the maximum number of bills



Class Imbalance is 66% towards Late Payers which is acceptable and does not need imbalance treatment.



INVOICE_TYPE: Good have more invoices generated as compared to Non-Goods



RECEIPT_METHOD: Most payments are done in Wired method and least in cash

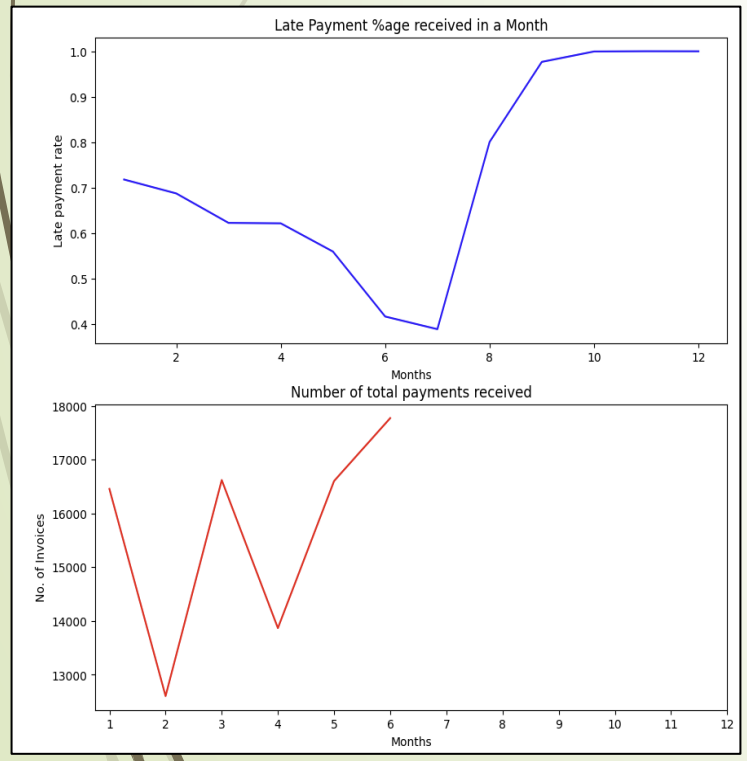
Characteristics of Late payment types (Bivariate)

- Monthly Late Payment rate impact-

From the first graph, 7th Month(July)7 has the very lowest late payment rate

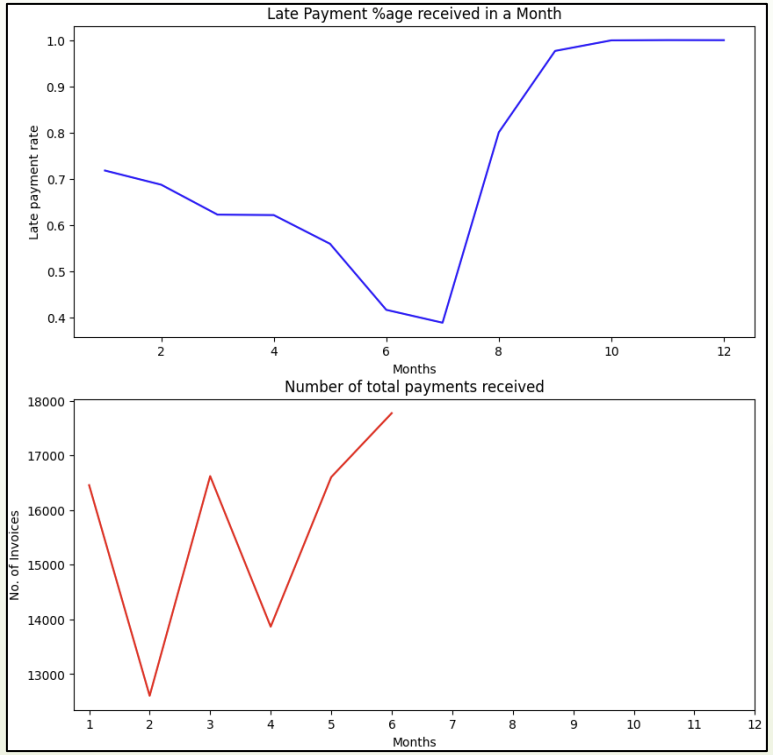
Starting 7th month, the late payment rate increases steeply. Also, the number of invoices are comparatively lower than the first half of the year.

For the 3rd month, the number of invoices is the highest and late payment rate is comparatively lower than other months with large number of invoices.



- Monthly Late Payment rate impact-

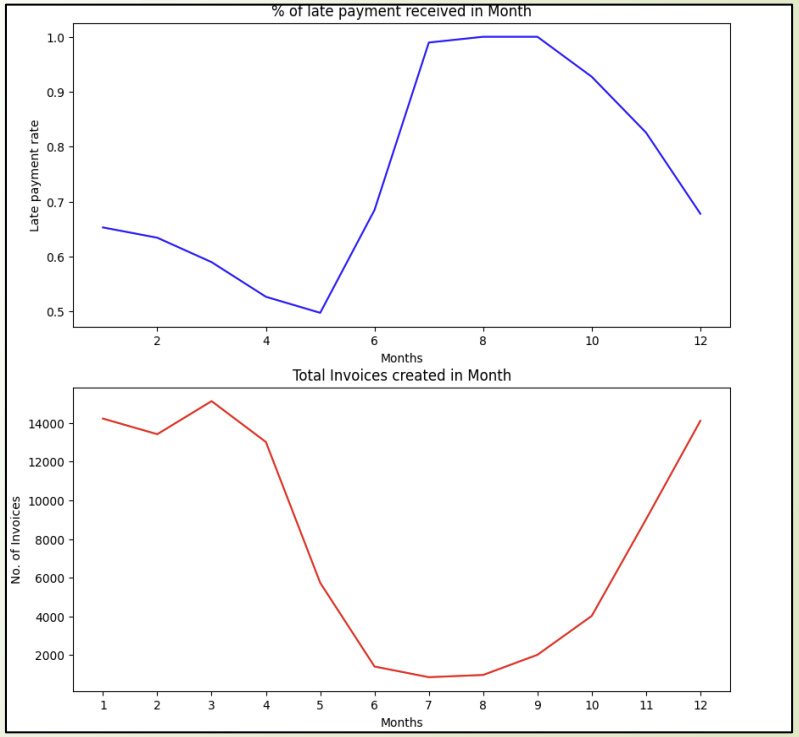
No payment received against any invoices from 7th month onwards.

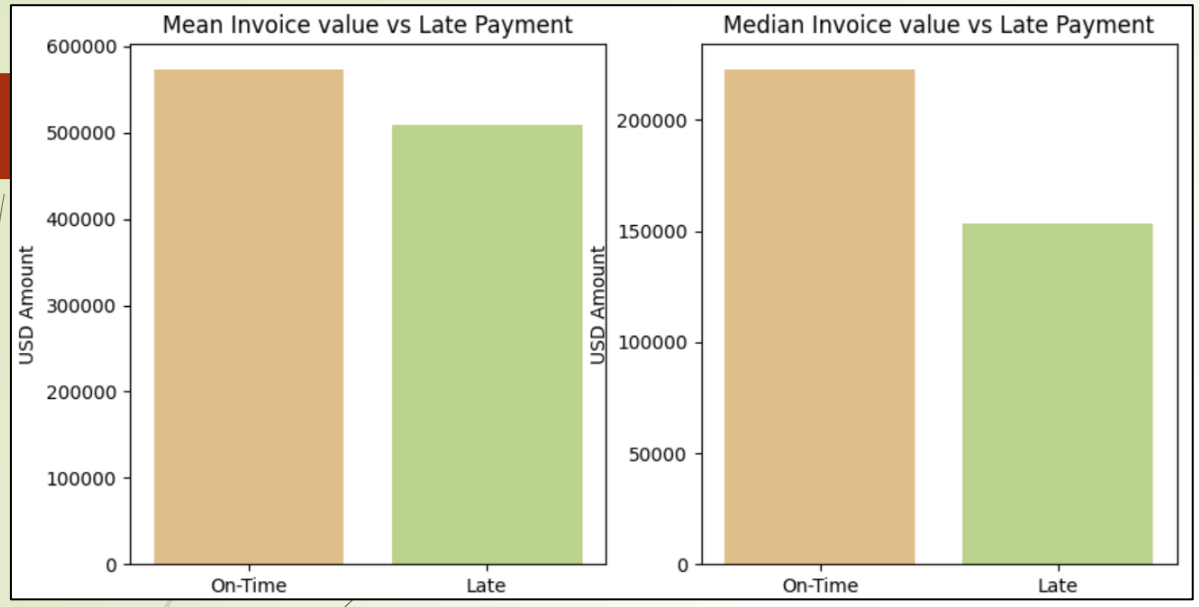


- Monthly Late Payment rate impact-

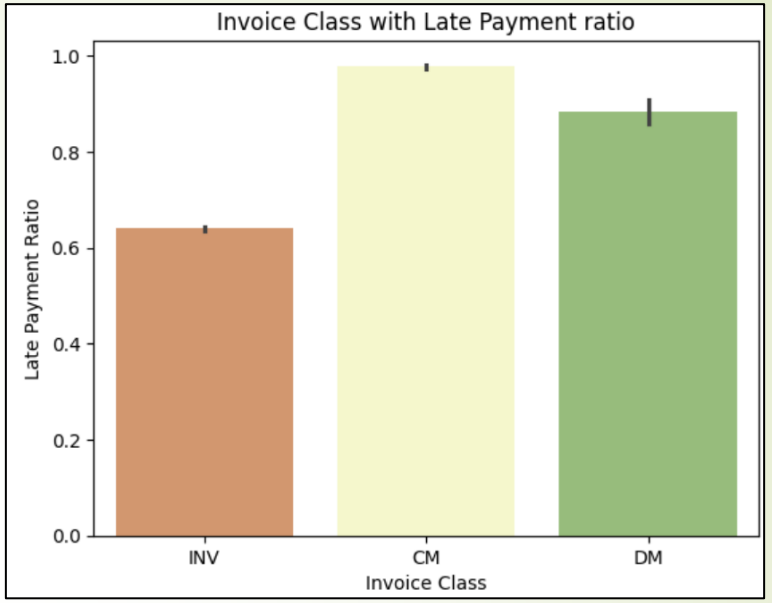
Late payment rate is decreases from 1st to 5th month and then shows a very high increase from 7th month

For the months 7, 8 and 9, the late payment rate is very hig and then decreases for 11th and 12th months.

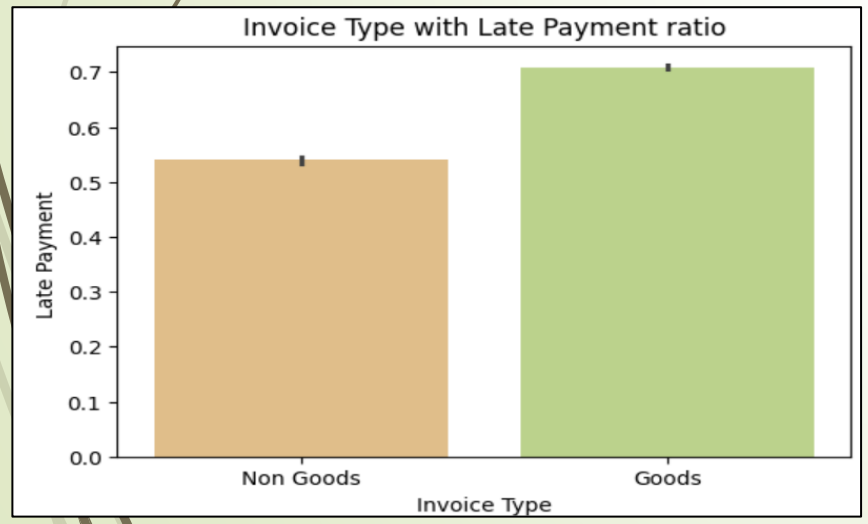




Clearly, the mean and median of invoice value of on-time bill payment is higher than late payment.



Late payment ratio for Credit Note transaction types are maximum, followed by Debit Note and Invoice. This indicates a high late_pay risk in Credit and Debit note invoice classes



Late payment ratio for Goods is higher than Non-Goods.

Customer Segmentation (K-means Clustering)

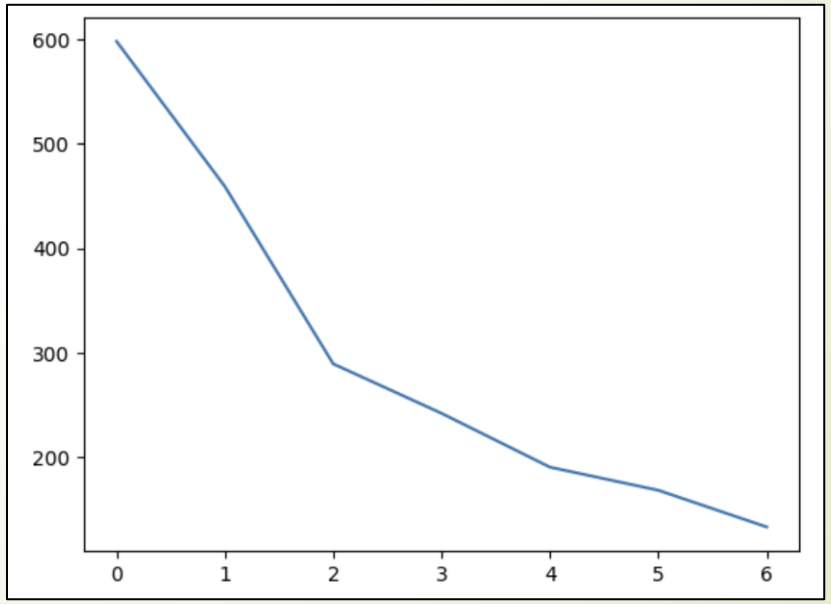
Recommendation: Customer-level attributes could also be important independent variables to be included in the model. A customer-level attribute can be determined via customer segmentation. You have to segment your customers based on two derived variables: the average payment time in days for a customer and the standard deviation for the payment time.

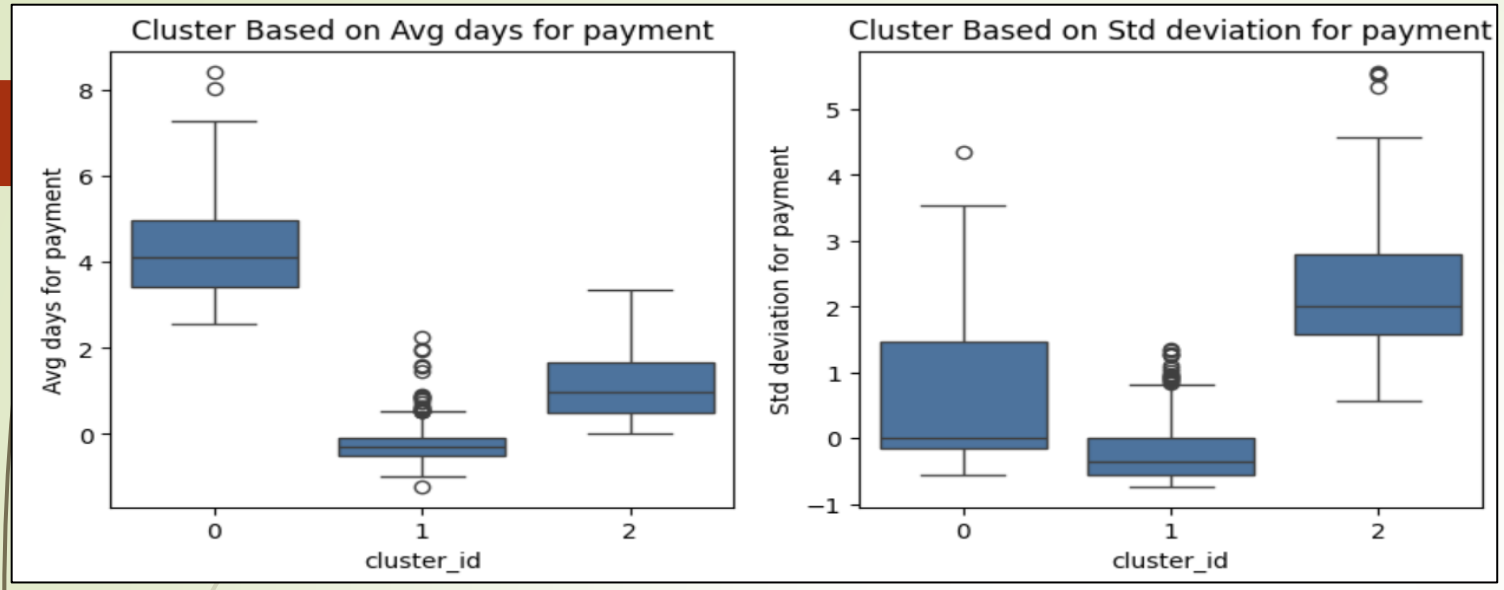
Using Silhouette analysis, we found the Optimal number of Cluster

Too many cluster will loose its importance so choosing k=3 by analyzing Silhouette score

```
For n_clusters=2, the silhouette score is 0.7512198959242521
For n_clusters=3, the silhouette score is 0.6083477287571657
For n_clusters=4, the silhouette score is 0.6114775297760283
For n_clusters=5, the silhouette score is 0.3987005610663089
For n_clusters=6, the silhouette score is 0.3980707870888025
For n_clusters=7, the silhouette score is 0.37108367800016495
For n_clusters=8, the silhouette score is 0.4154930610650793
```

Elbow curve

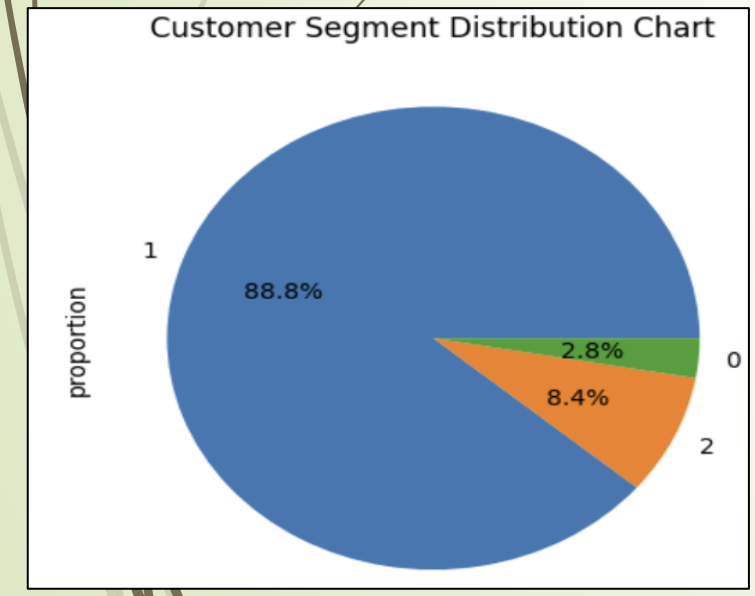


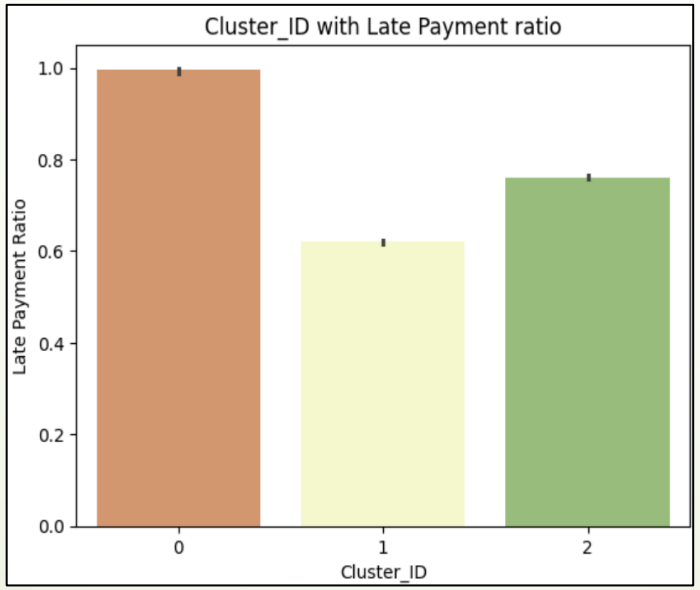


0 Cluster -- Medium Invoice Payment - These customers usually pay their invoices on time or within 20 to 60 days

1 Cluster -- Early Invoice Payment - These customers usually pay their invoices on time

2 Cluster -- Delay or Late Invoice Payment - # The variability is highest when credit period is between 20 to 60 days



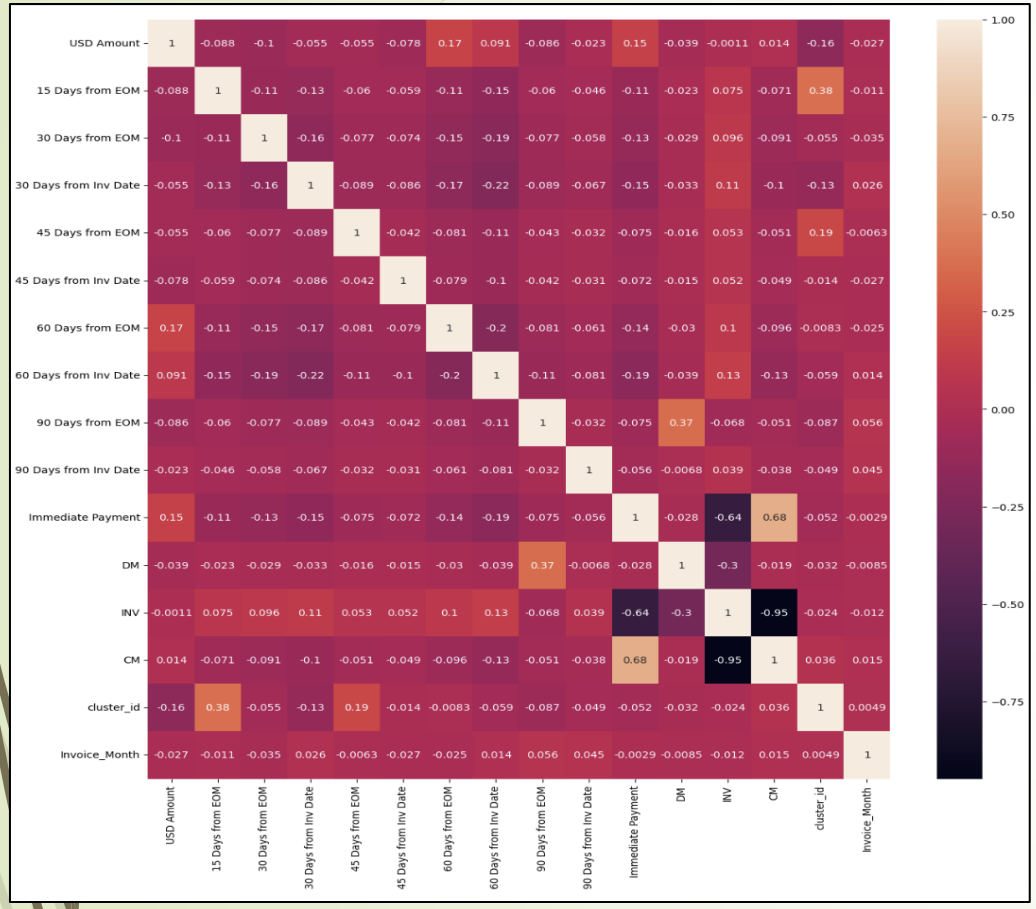


From PIE chart and Bar Chart here-

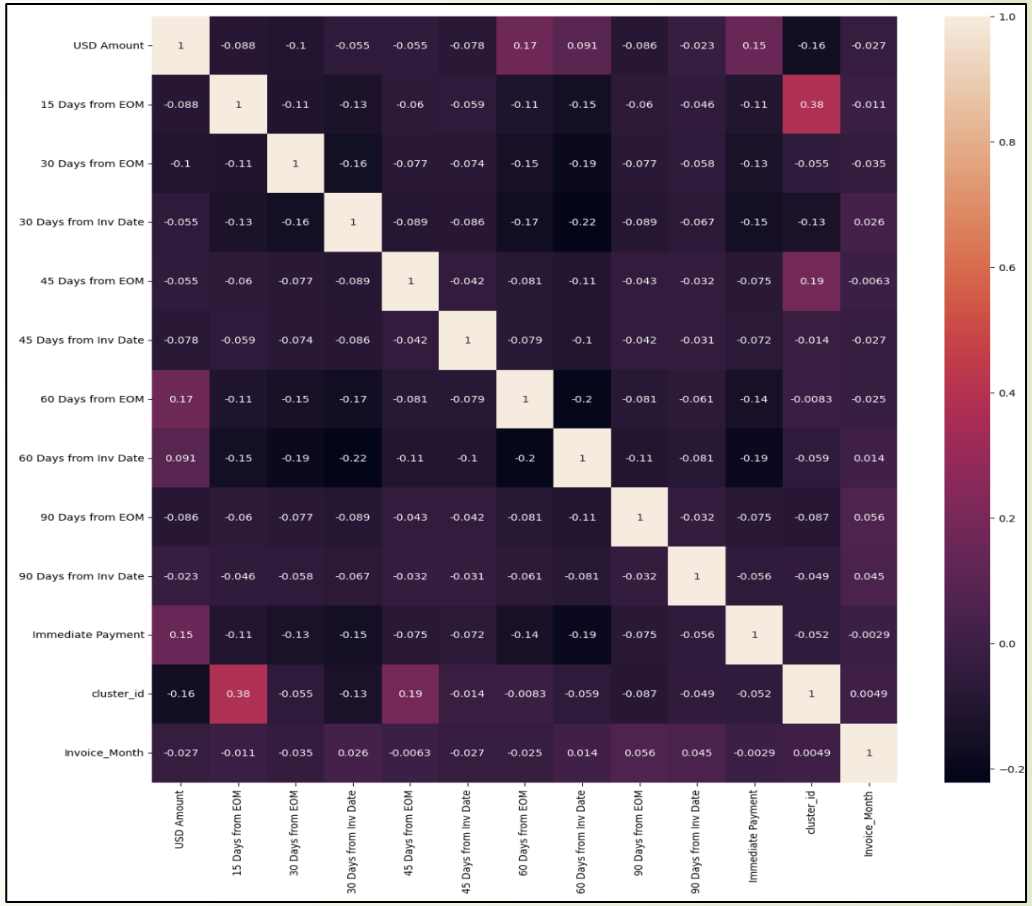
- 1 Cluster -- Early Invoice Payment
- 2 Cluster -- Delay or Late Invoice Payment
- 0 Cluster -- Medium Invoice Payment

Clearly, cluster 1, were early payers with least number of average days taken to pay and cluster 2 were delayed or late payers

Data Preparation/Splitting and Model Building



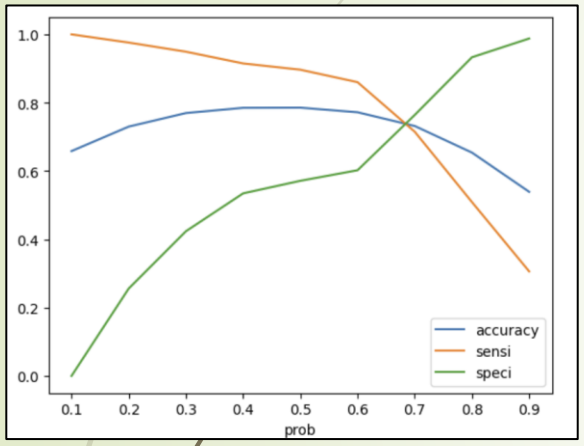
CM & INV, INV & Immediate Payment, DM & 90 days from EOM has high multicollinearity, hence dropping these columns.



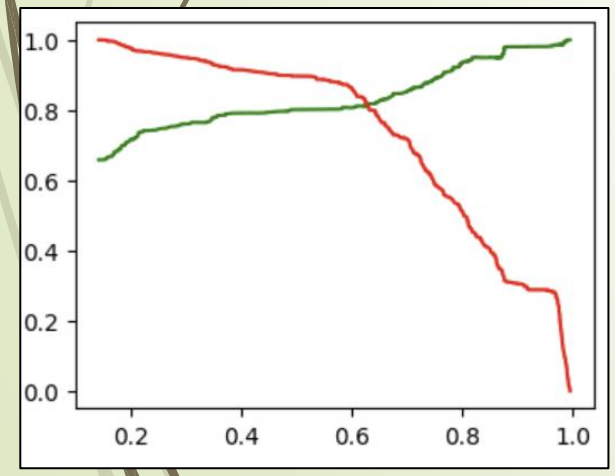
After removing the high multicollinearity in the model

Model Building: Logistic Regression and Random Forests

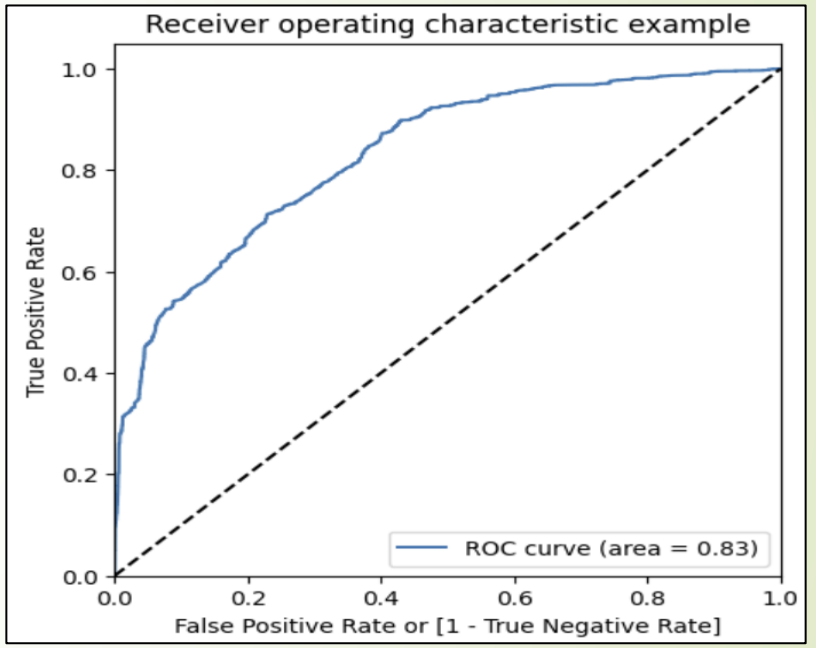
Logistic Regression



Graph showing accuracy, sensitivity and specificity for various probabilities
0.6 is the optimum point to take it as a cutoff probability.



On Precision & Recall trade off we found optimal cutoff of between 0.6 & 0.7 .
Hence keeping the optimal cutoff 0.6.



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

Random Forest

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: 0.9389846480412654
              precision    recall  f1-score   support

           0       0.97      0.91      0.94     22469
           1       0.95      0.98      0.97     43284

    accuracy                           0.96     65753
   macro avg       0.96      0.95      0.95     65753
weighted avg       0.96      0.96      0.96     65753
```

Comparison between Logistic Regression and Random Forest

Logistic Regression

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

```
0.7759403832505323
```

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

```
0.8119895129575476
```

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

```
0.86004066167637
```

Our train and test accuracy is almost same around 77.7 %

Random Forest

	precision	recall	f1-score	support
0	0.91	0.86	0.88	9490
1	0.93	0.96	0.94	18690
accuracy			0.92	28180
macro avg	0.92	0.91	0.91	28180
weighted avg	0.92	0.92	0.92	28180

- **The Random Forest model significantly outperformed the Logistic Regression model in both precision and recall scores. In this case, recall scores were prioritized to improve the identification rate of late payers for targeted action.**
- Since the data is heavy on categorical variables, random forest is better suited to the job than logistic regression

Feature Importance

With the help of Random Forest, we are able to find out features that have major impact on Late Payment. Listing them below based on ranking-

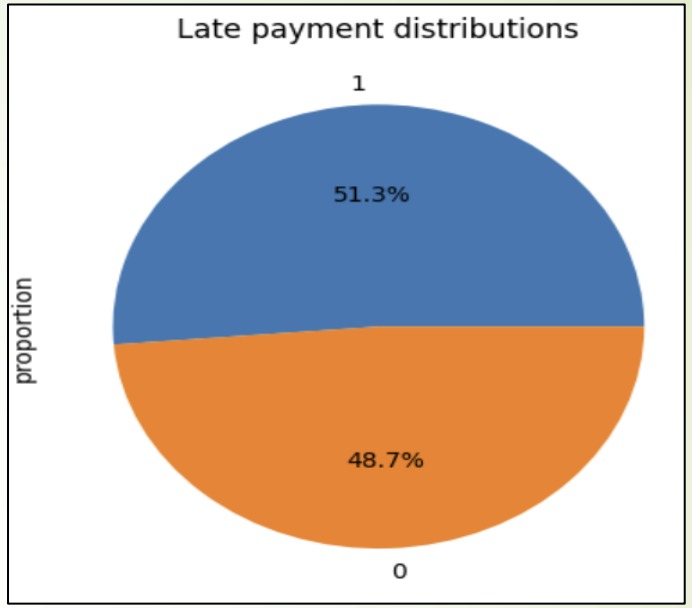
- USD Amount
- Invoice Month
- 30 Days from EOM
- 60 Days from EOM
- 15 Days from EOM
- Cluster-ID (which in turn is dependent on average and standard deviation of days required to make payment)

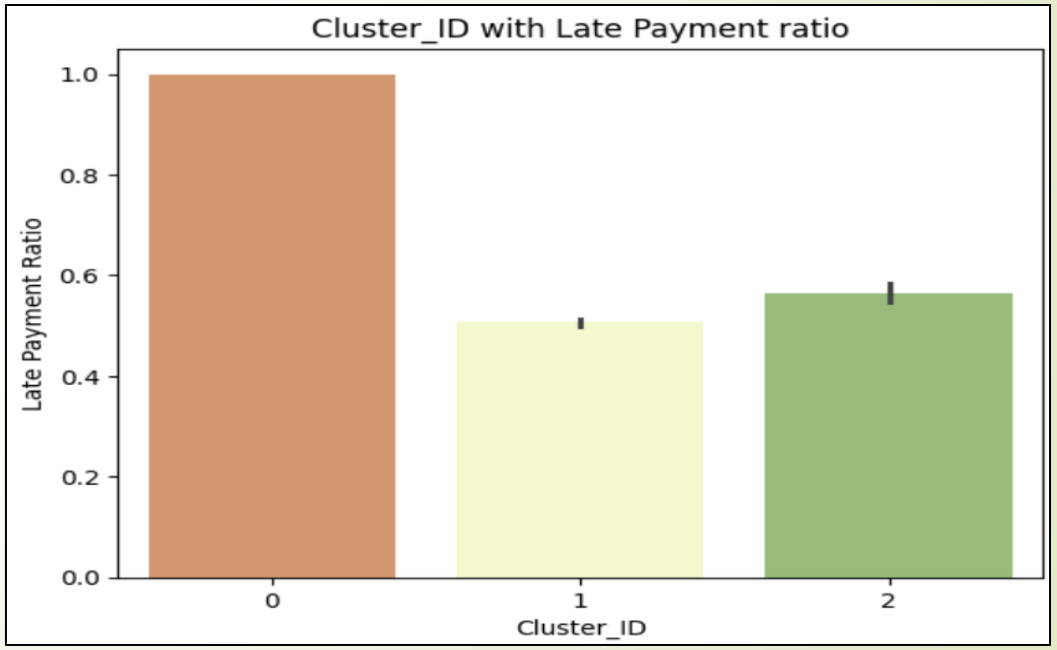
Important Feature Ranking

```
Feature ranking:
1. USD Amount (0.489)
2. Invoice_Month (0.131)
3. 30 Days from EOM (0.112)
4. 60 Days from EOM (0.111)
5. Immediate Payment (0.043)
6. 15 Days from EOM (0.028)
7. cluster_id (0.027)
8. 60 Days from Inv Date (0.012)
9. 30 Days from Inv Date (0.012)
10. INV (0.008)
11. 90 Days from EOM (0.007)
12. 90 Days from Inv Date (0.007)
13. 45 Days from EOM (0.005)
14. 45 Days from Inv Date (0.004)
15. CM (0.004)
16. DM (0.001)
```

- Based on final data set, we see that 51.3% payments predicted to be delayed

- Customer segment with historically delayed/late payment days most likely to have the most late payment rate than historically early or medium days payment transactions, this is similar to the result found based on historical outcomes





Top 10 customers with highest delay in payment

Customer_Name	Delayed_Payment	Total_Payments	Delay%
FORE Corp	8	8	100.0
ALSU Corp	7	7	100.0
LVMH Corp	4	4	100.0
SUND Corp	4	4	100.0
TRAF Corp	3	3	100.0
ROVE Corp	3	3	100.0
MAYC Corp	3	3	100.0
MUOS Corp	3	3	100.0
CITY Corp	3	3	100.0
WELL Corp	2	2	100.0

Recommendations

- Credit note payments show the highest delay rate compared to debit notes or invoices, suggesting stricter collection policies for these invoice types.
- Goods-related invoices also have higher delay rates than non-goods, warranting tighter payment policies.
- As lower-value transactions make up the majority and are more frequently delayed, the company should prioritize these, applying a scaled penalty based on billing amount (higher penalty percentage for lower bills) as a last resort.
- Customers were clustered into three segments: 0 (medium payment duration), 1 (early), and 2 (Late Payment). Cluster 2 customers, with the longest delay rates, should receive heightened attention.
- Top 10 companies with the highest probability and high delayed payment counts should be prioritized for focused collection efforts.

Thank you