E-Commerce and Retail B2B Case Study

Tiana Viola Pinto
Ujwal Kant
Tushar Parwani

Objective

Problem Identification

- A sports retail company Schuster dealing in B2B transactions often deals with vendors on a credit basis, who might or might not respect the stipulated deadline for payment.
- Vendors delaying their payments result in financial lag and loss which becomes detrimental to smooth business operations
- Additionally, company employees are set up chasing around for collecting payments for a long period resulting in no value-added activities and wasteful resource expenditure

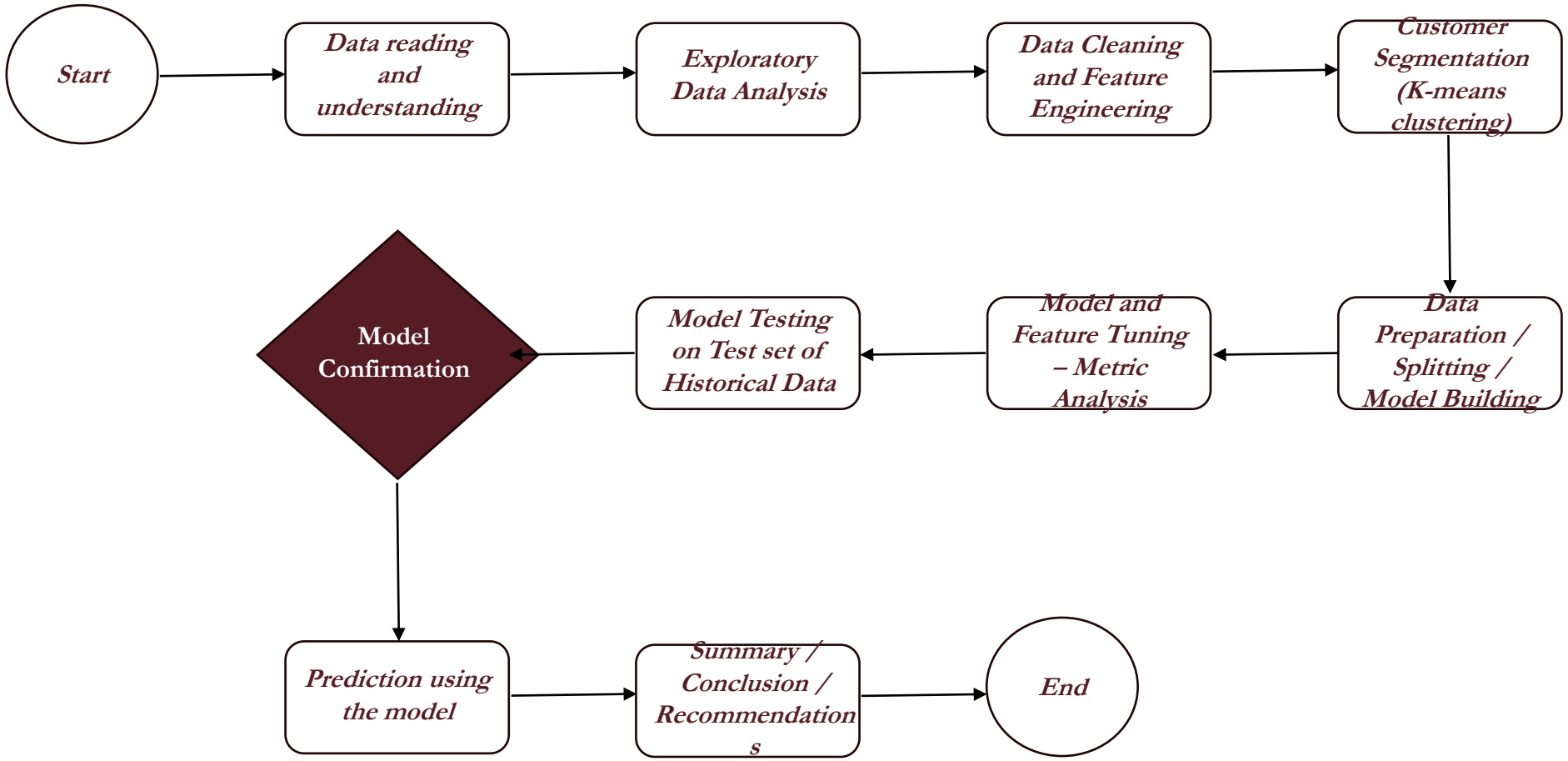
Business Objectives

- Customer segmentation to understand the customer’s payment behaviour
- Using historical information, the company requires prediction of delayed payment against an unforeseen dataset of transactions with due dates yet to be crossed
- The company requires the prediction for better resource delegation, quicker credit recovery and reduction of low value-adding activities

Approach Strategy

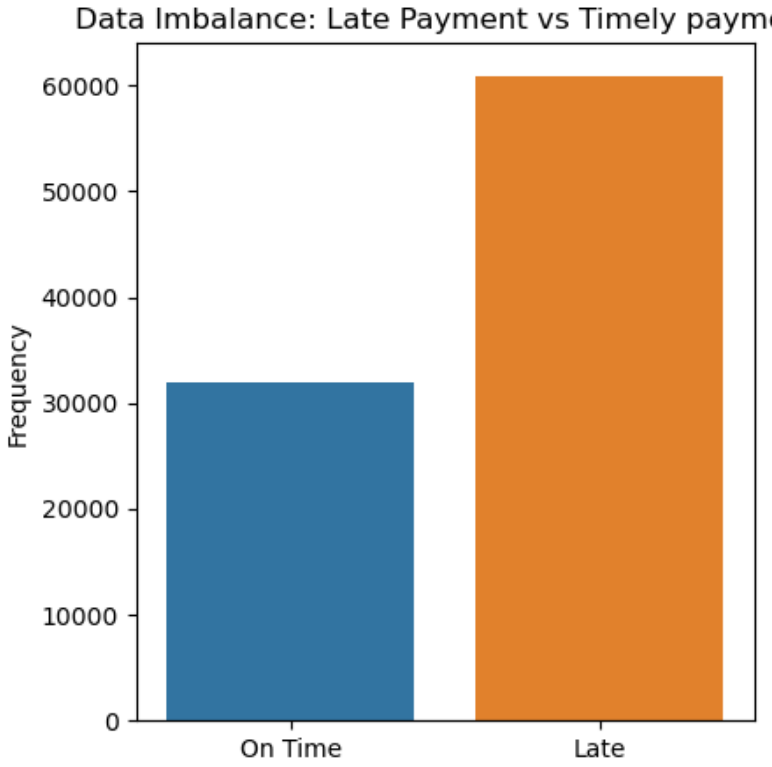
Process Summary

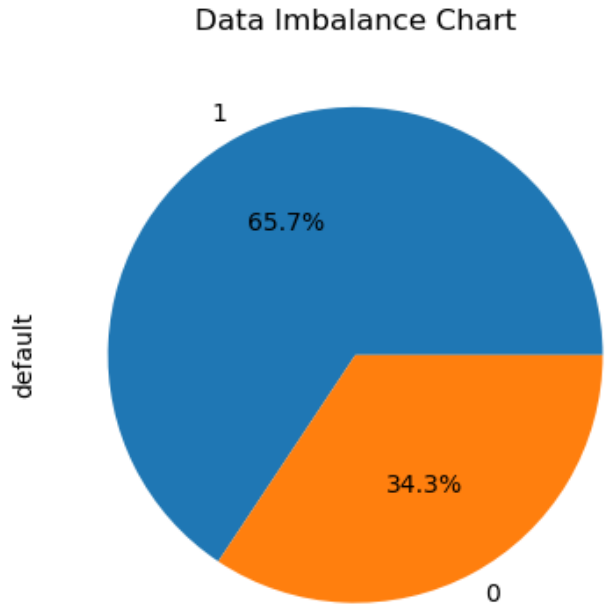
Steps Involved in Model Building



EDA & Data Analysis

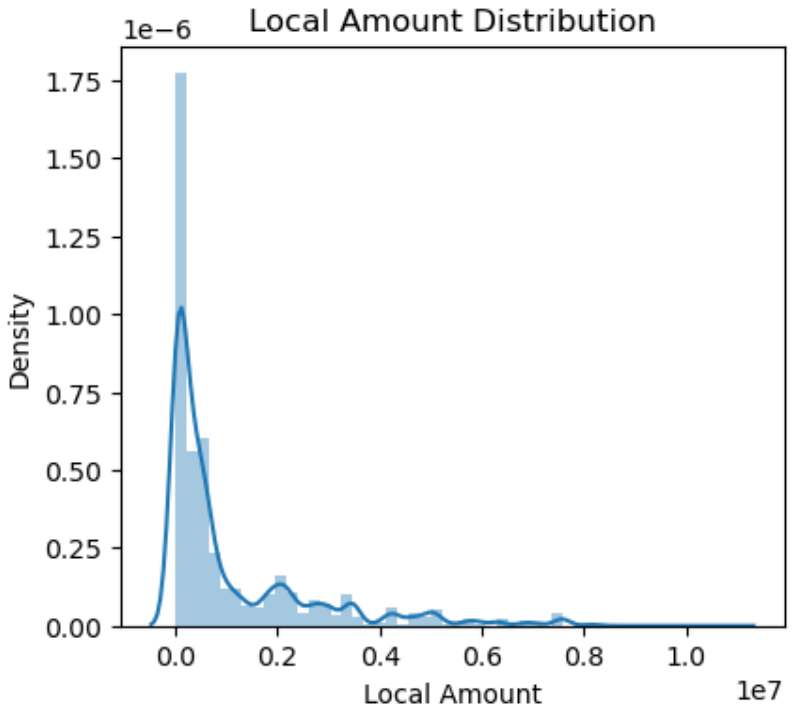
Univariate Analysis Observation

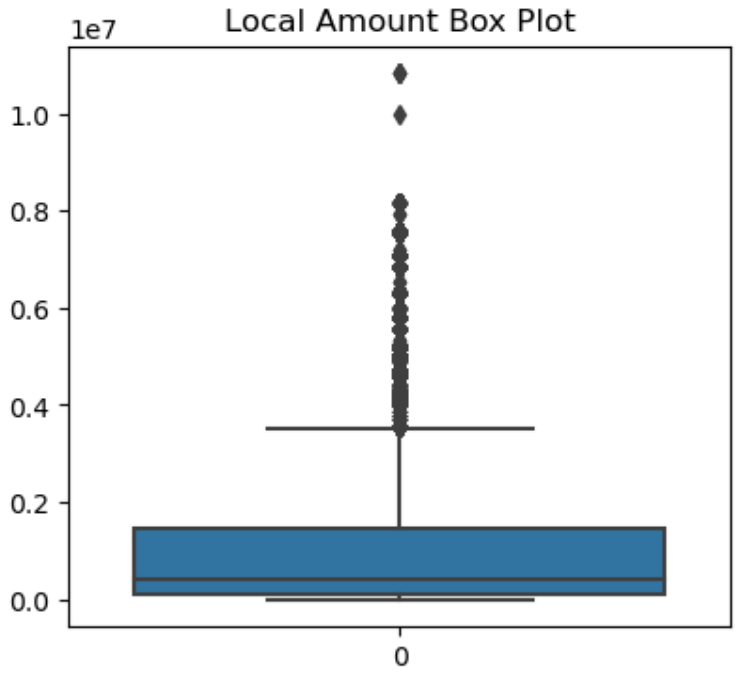




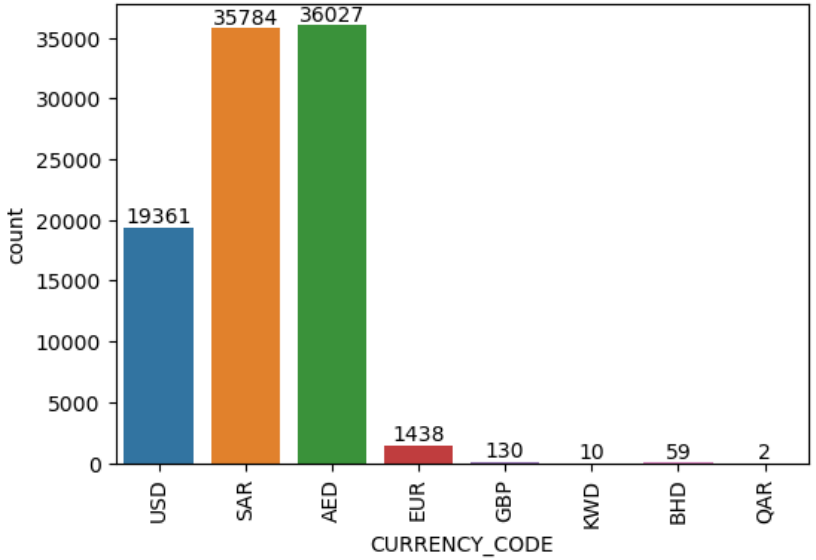
Payment delayers make up 65.7% of the class imbalance, which is acceptable and does not require imbalance management.

Univariate Analysis Observations

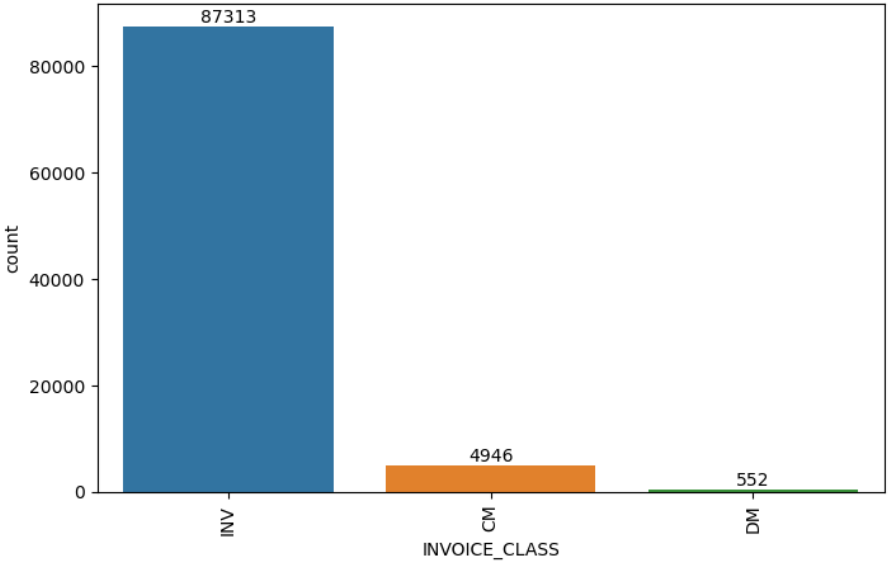




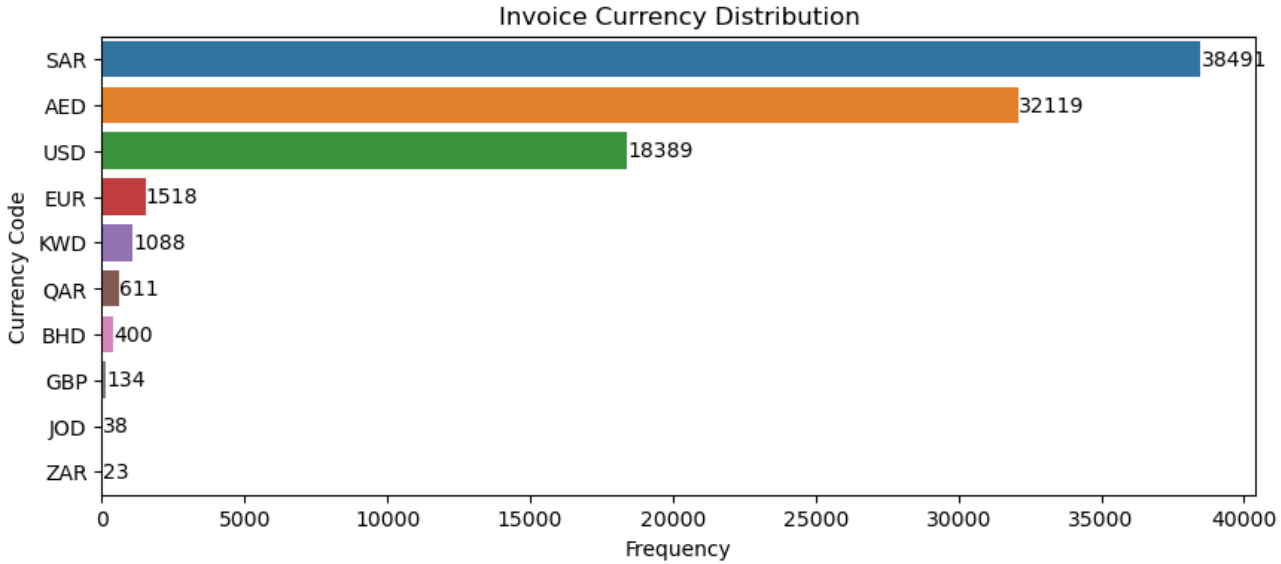
Since there are no currency values in the 'Local Amount' columns and an alternate column 'USD Amount' containing data conveys the same information regarding the bill amount, the column 'Local Amount' seems redundant and was hence dropped.



The currency used for bill payments is mostly USD, SAR or AED.

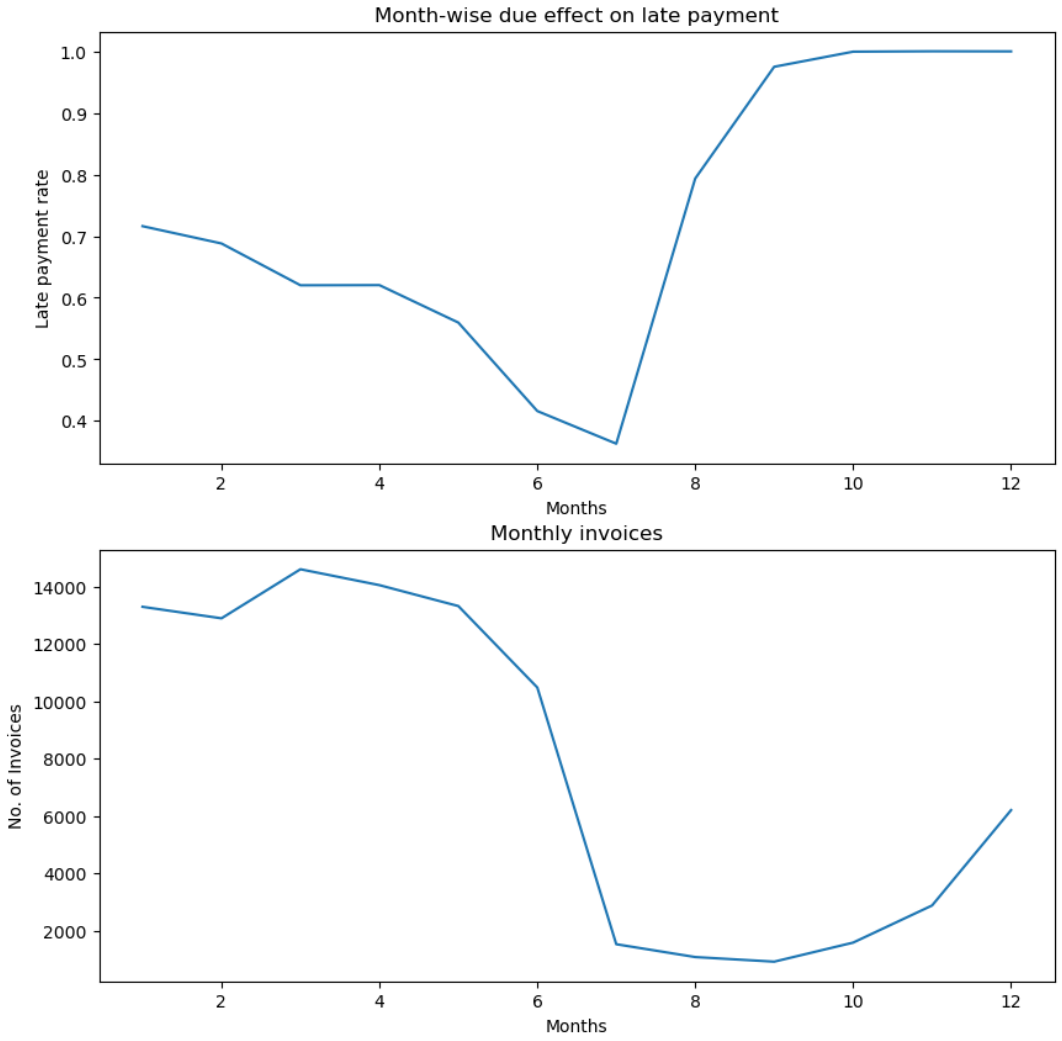


INV has the maximum number of bills in the INVOICE_CLASS column.

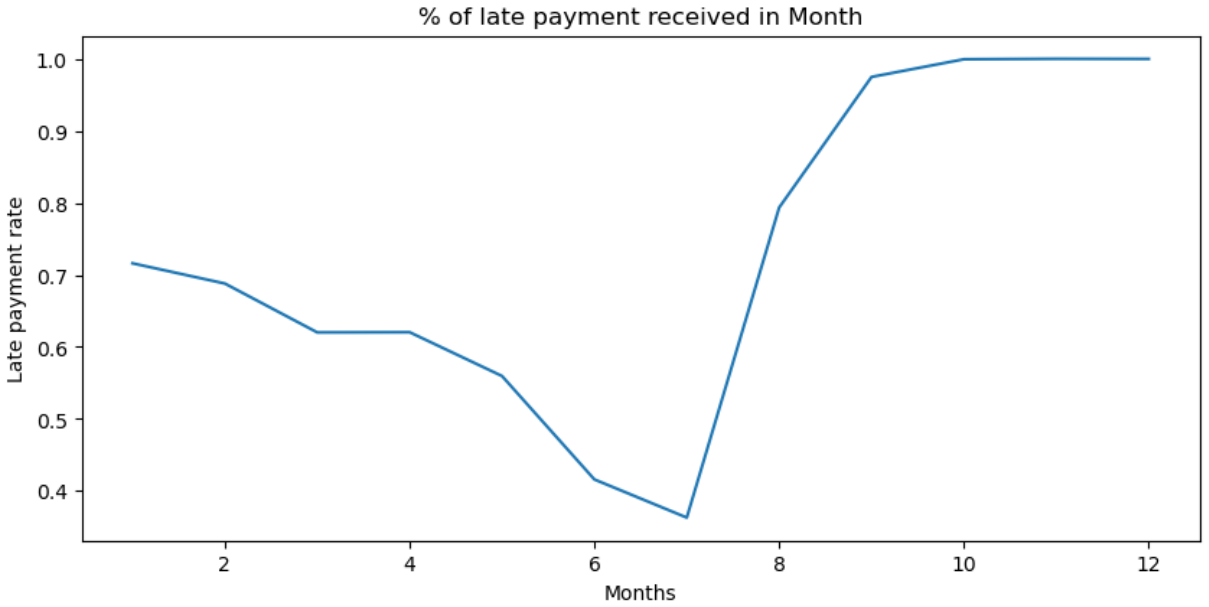


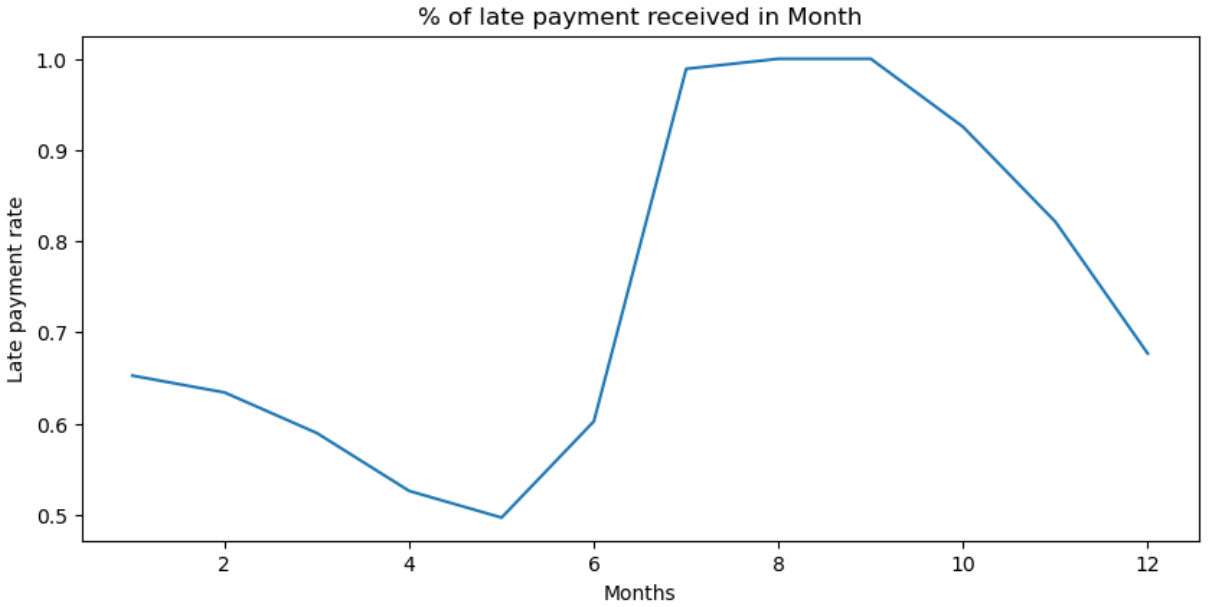
The most preferred payment method for bill payment is WIRE.

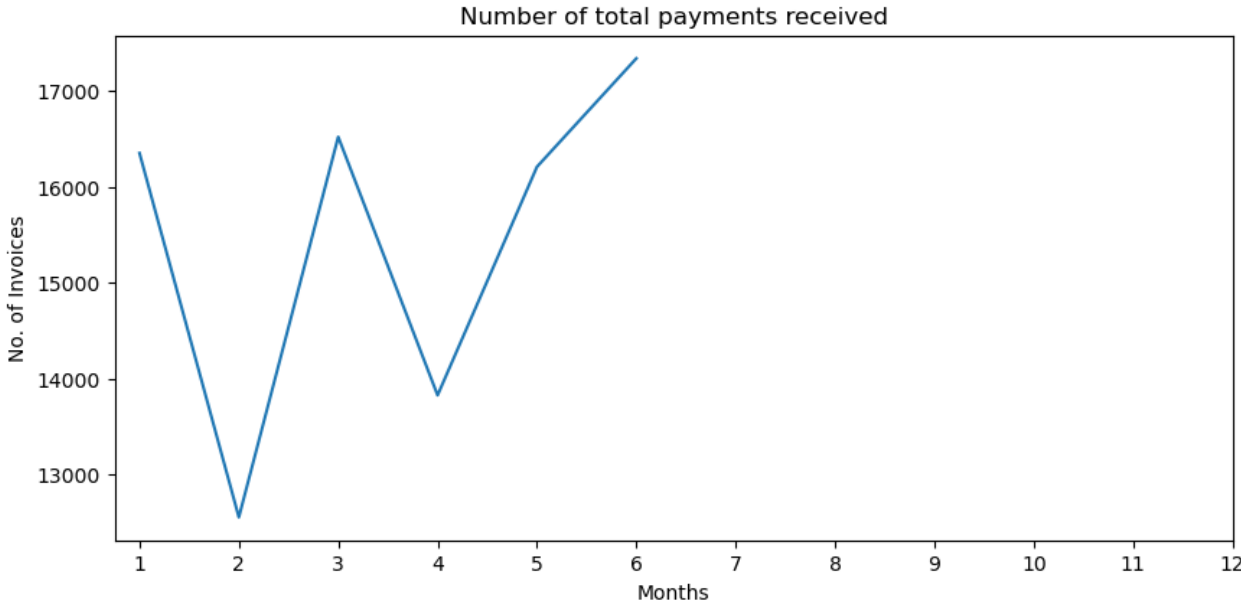
Bivariate Analysis Observation

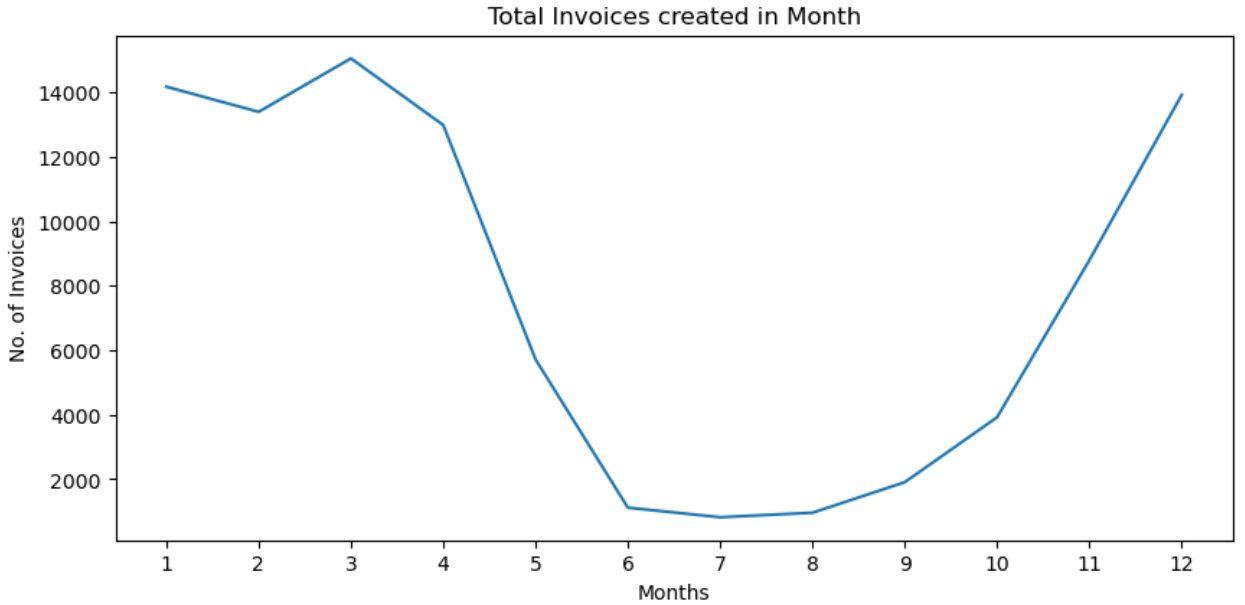


The third month appears to have the most invoices, however, compared to other months with high invoice counts, this month has a low late payment rate. Because there are fewer bills in the seventh month, there is a very low late payment rate. Even though there are comparably fewer invoices in the second half of the year than in the first, the late payment rate rises sharply beginning in the seventh month.









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

Late payment rate decreases from the 1st to the 5th month. For the months 7, 8 and 9, however, the rate is very high.

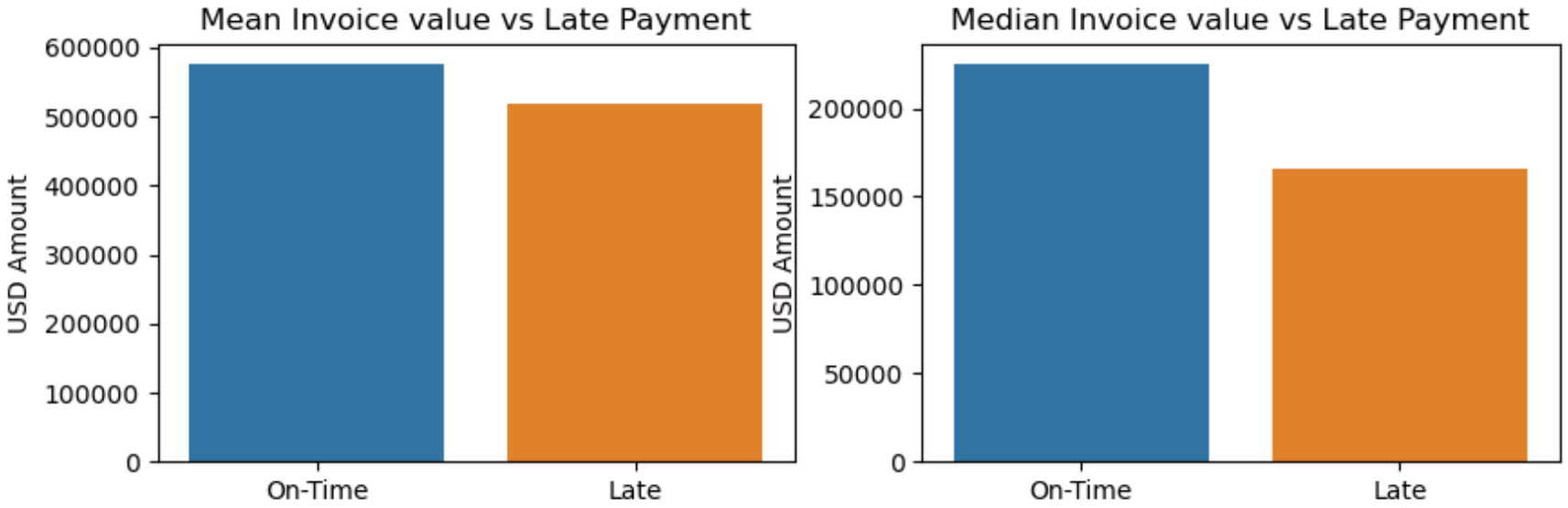


Fig1

The mean and median of invoice value of On-time bill payment is higher than late payment.

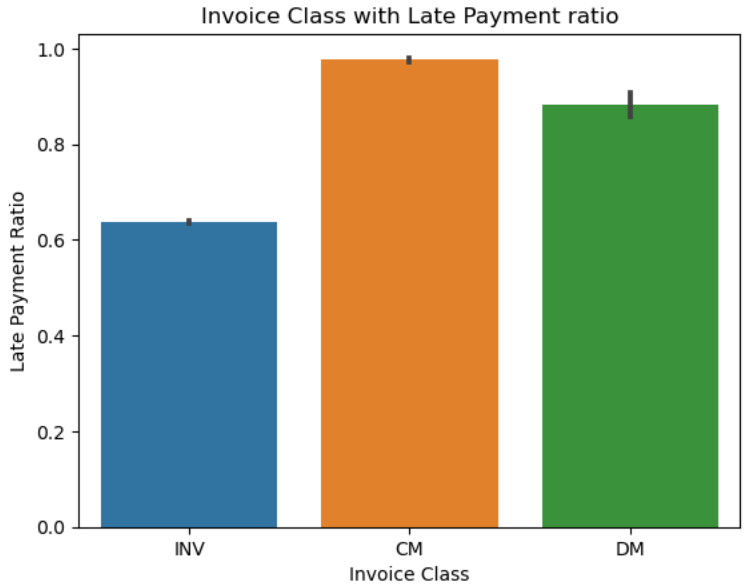
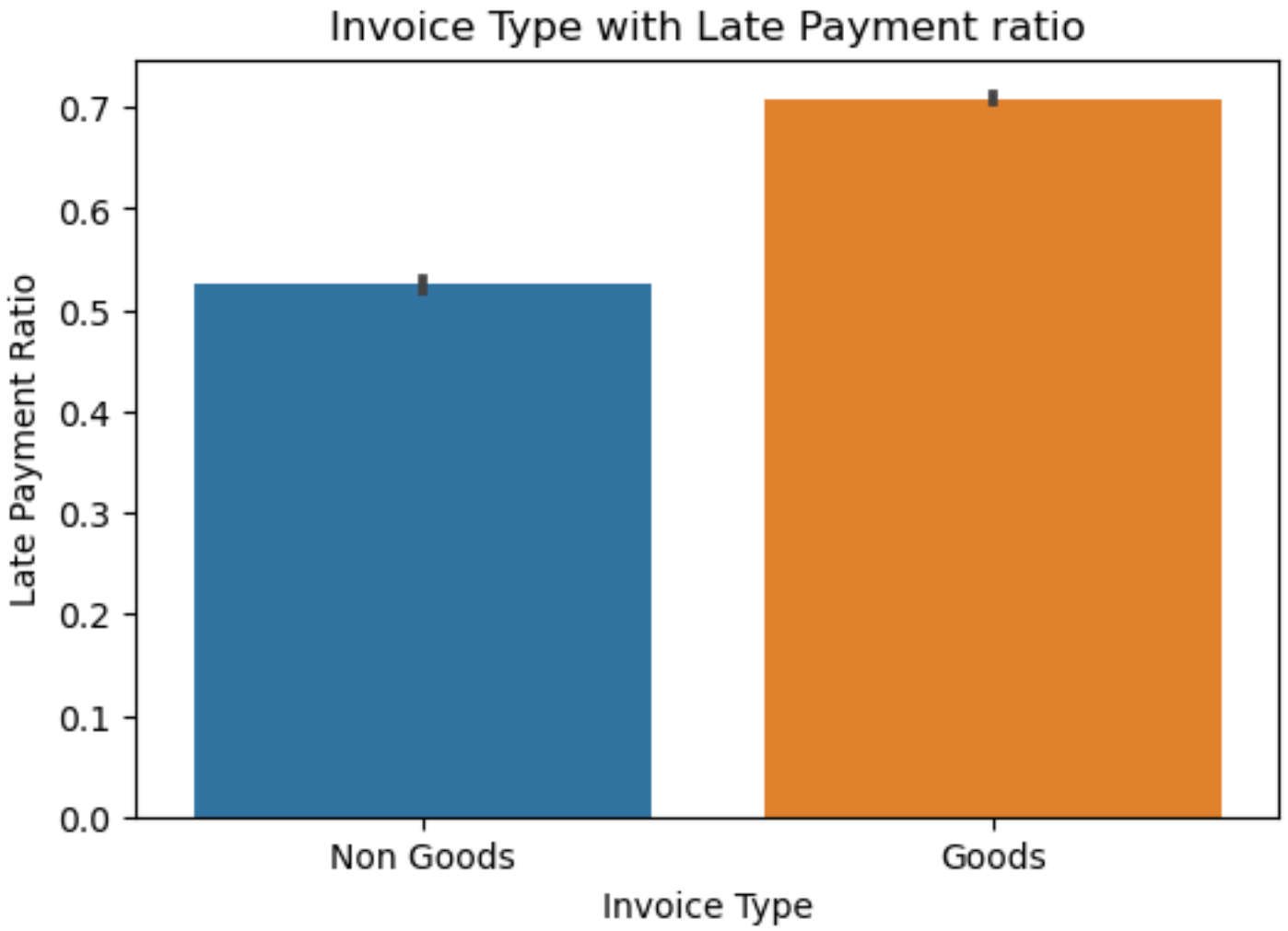


Fig 2

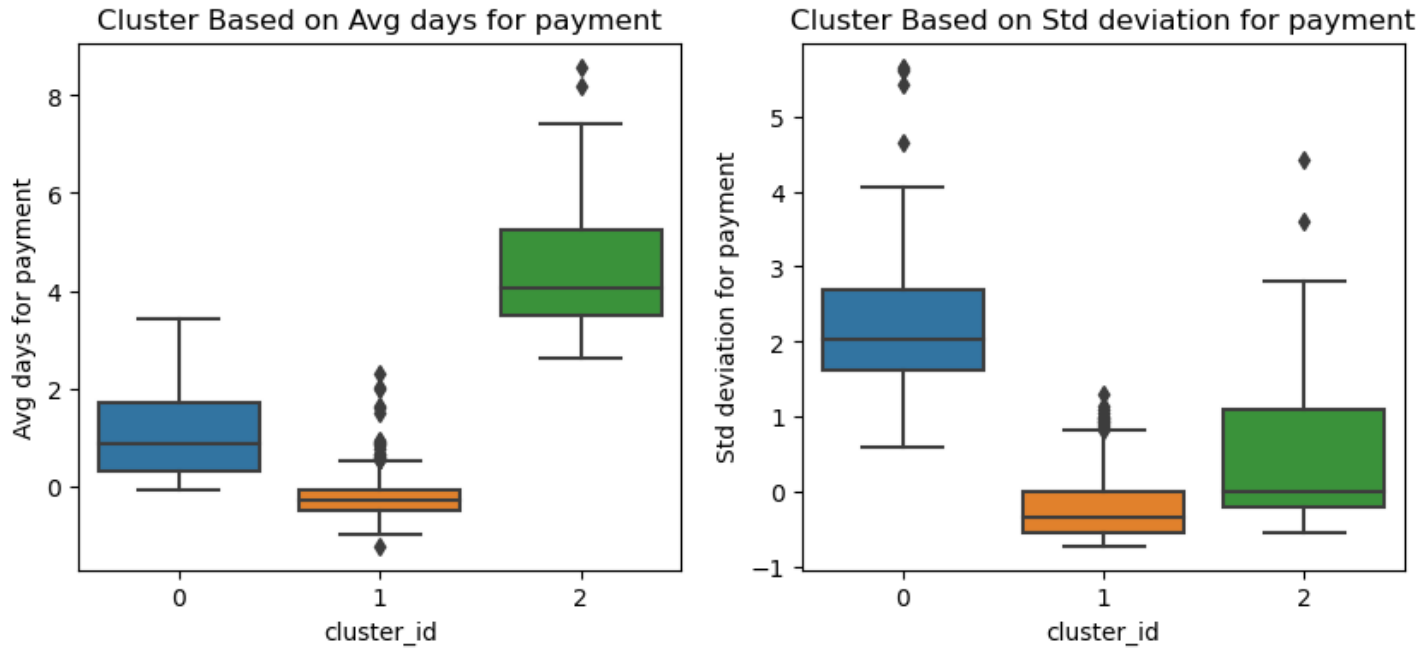
Late payment ratio is very high for CM and lowest for INV INVOICE_CLASS.



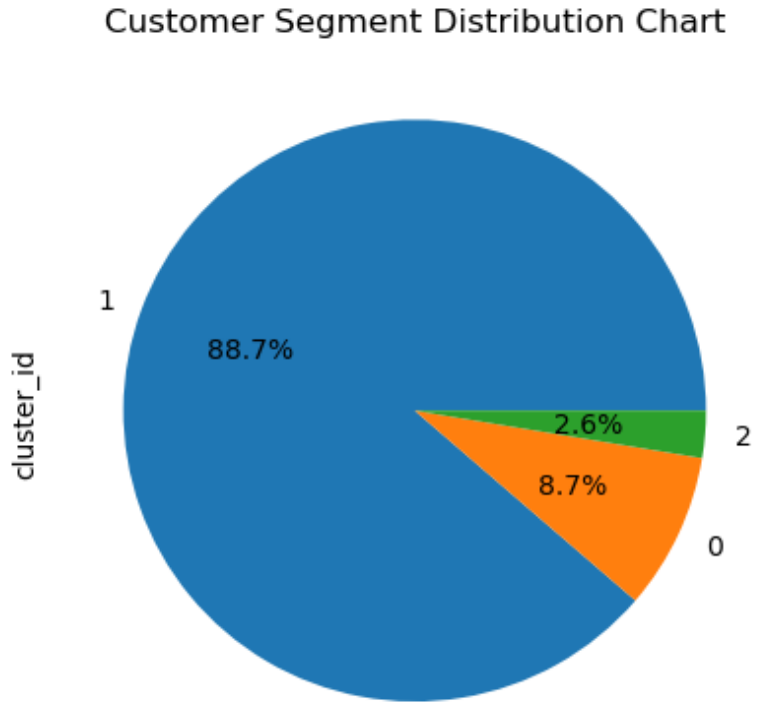
The l ate payment ratio for Goods is higher than for Non-Goods.

Customer Segmentation
K means Analysis

Enhancing model performance can be significantly achieved by incorporating customer-level attributes. By segmenting customers based on the average and standard deviation of their payment times, distinct customer groups can be identified. These segments can then be utilized as valuable features within the machine learning model.

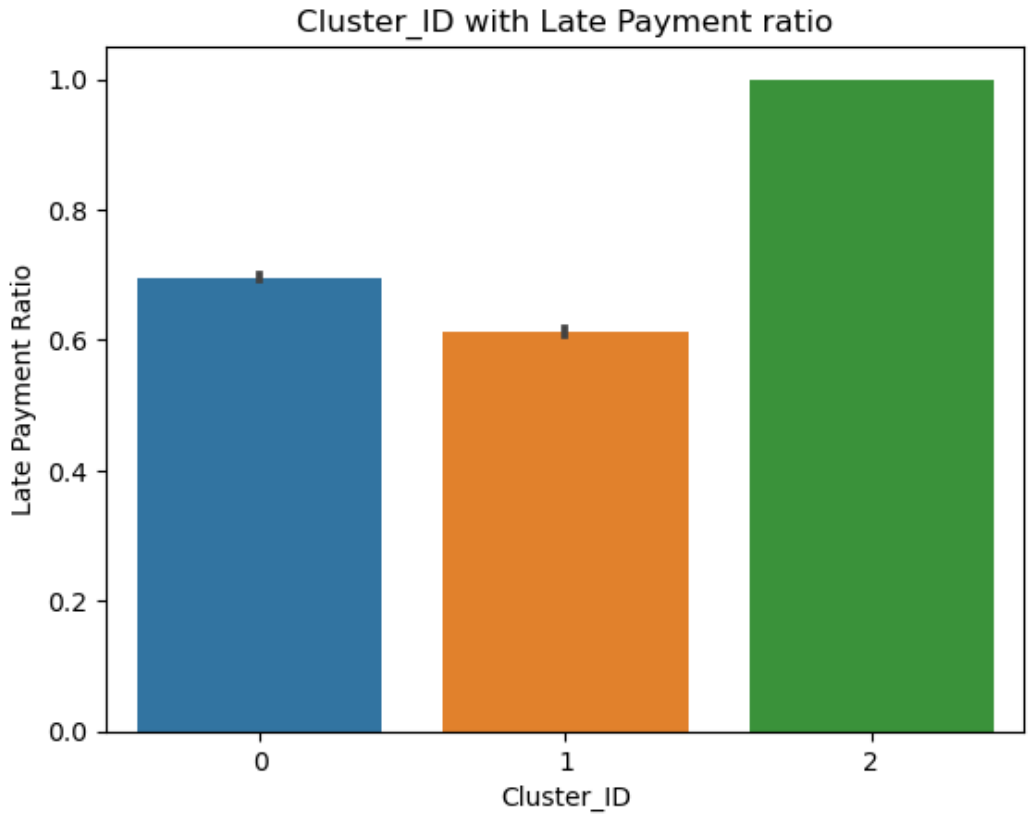


Cluster 0 - Medium Invoice Payment
Cluster 1 - Early Invoice Payment
Cluster 2 - Prolonged Invoice Payment



Cluster 0 - Medium Invoice Paymen
Cluster 1 - Early Invoice Payment
Cluster 2 - Prolonged Invoice Payment

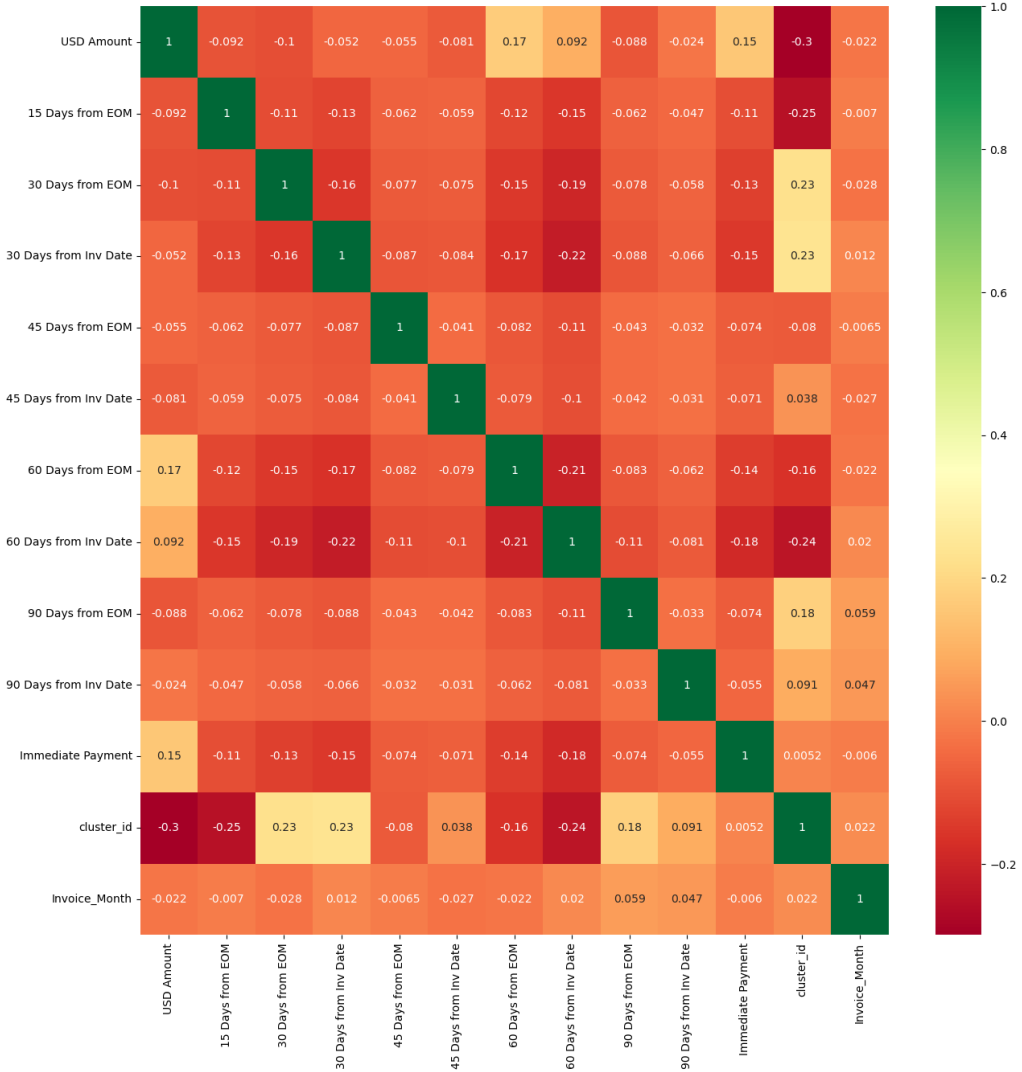
From the graphical distribution above, we can see that Early customers comprise of 88.7% of customers whereas medium payment duration has about 8.7%. Prolonged payers comprise the smallest distribution of 2.6%



Category 0 – Early Payers had the shortest average payment duration, while Category 2 – Extended Payers had the longest average payment duration.
Category 1 – Medium-Duration Payers.

Model Building

Model Building



Since the columns
- CM & INV
- INV & Immediate Payment
- DM & 90 days from EOM

These columns were dropped and the heatmap was plotted again with the remaining columns. The results presented hereby indicate that after dropping columns with high multicollinearity, the dataset is now ready for model building

Logistic Regression VS Random Forest

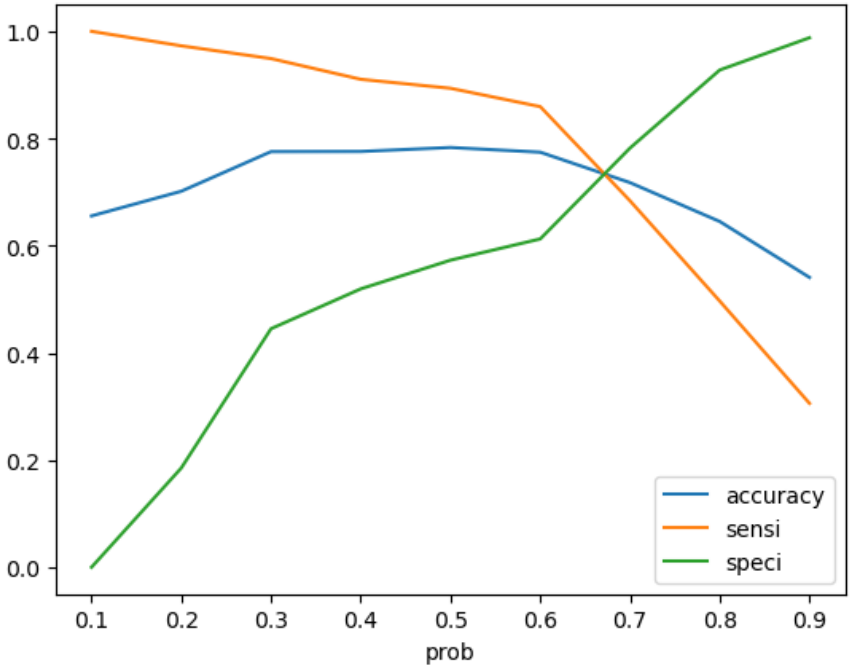


Fig 1

Fig 1: From the logistic regression plot above, 0.65 was taken to be the optimum point as a cutoff probability.

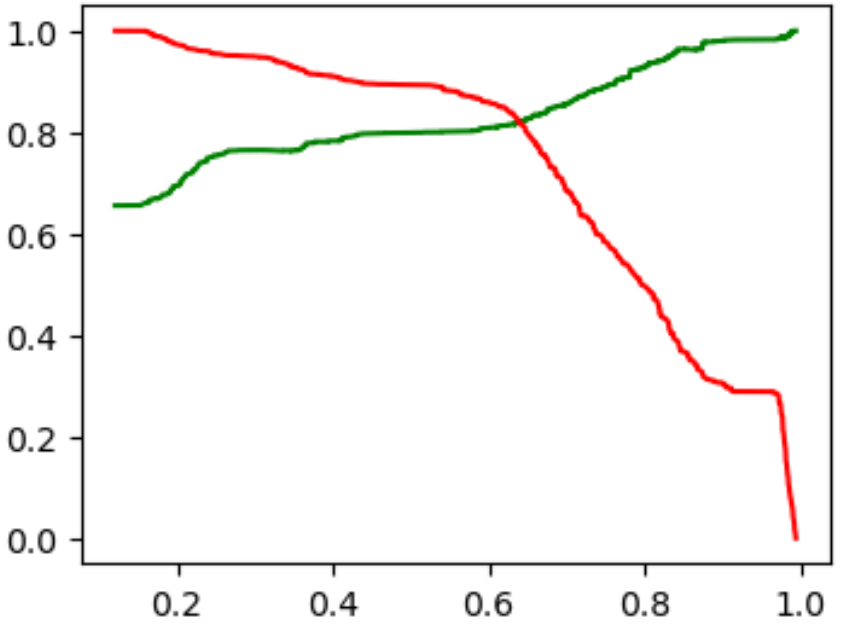
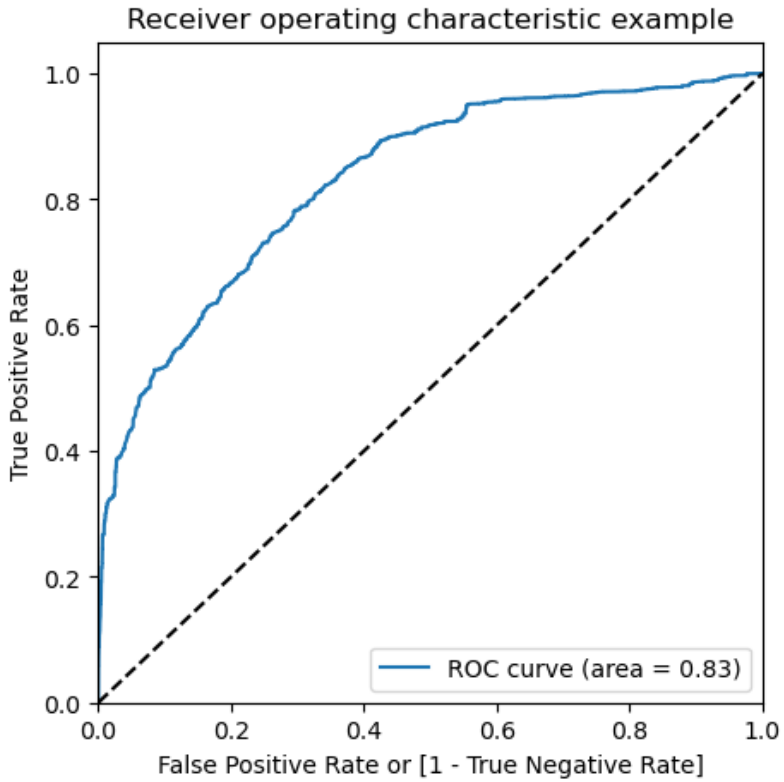


Fig 2

Fig 2: The plot from the random forest confirmed the feasibility of using 0.65 as the optimal point for the cutoff probability

ROC Curve



AUC = 0.83

This indicates that the model is acceptable

The following parameters were obtained by building a random forest model using the same parameters as the logistic regression and hyper-parameter tuning.

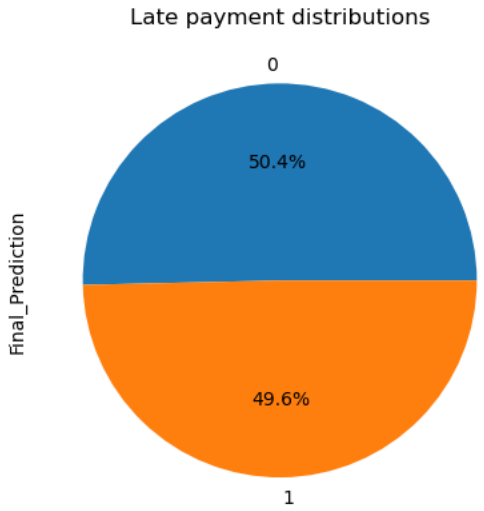
Random Forest outperformed Logistic Regression.

- It is evident that the Random Forest model outperformed the logistic regression model in terms of total precision and recall scores. Furthermore, because it was crucial to raise the percentage forecast of late payers to be targeted, memory ratings were particularly significant in this instance.
- Random forest is more appropriate for this task than logistic regression because the data primarily consists of categorical variables.
- As a result, the random forest model was decided upon as the preferred model and forecasts were made.

Comparison

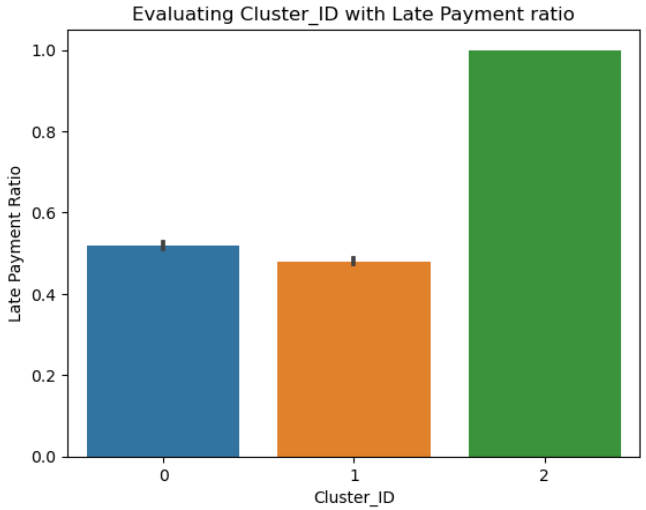
Using the above parameters, a random forest model was built, whose metrics were compared to the logistic regression model and the final model was finalized therefore

Predictions made by the Final Model



After applying the final model's predictions to the open-invoice data based on customer names, it is anticipated that 50.2% of transactions will experience a payment delay, potentially causing an unexpected lag in business operations.

According to Open invoice data, 50% of payments are expected to be delayed; extended payment days show startlingly high delay rates.



Customers with a history of long payment delays are expected to have the highest delay rate, approximately 100%, mirroring past observed patterns. In contrast, those with a history of early or medium payments are anticipated to experience fewer delays..

Recommendations

Observations & Recommendations

1. Credit Note Payments exhibit the highest delay rate compared to Debit Note or Invoice types, necessitating stricter company policies on payment collection for these invoice classes.

2. Lower-value payments constitute the majority of transactions and are more frequently delayed. It is recommended to focus on these, applying penalties based on billing amounts—smaller bills should incur higher penalty percentages for late payments, used as a last resort.

3. Goods-related invoices have significantly higher payment delays than non-goods types, warranting stricter payment policies.

4. Customer segments were categorized into three clusters: 0, 1, and 2, representing medium, and early payment durations, respectively. Customers in cluster 1 (prolonged payments) show significantly higher delay rates and should receive extensive attention.

5. Companies with the highest probability and total delayed payment counts should be prioritized due to their significant delay rates.

Thank you