E-Commerce and Retail B2B Case Study

By:
Vaibhav Vijay

Problem identification

Schuster, a sports retail company engaged in B2B transactions, frequently extends credit to vendors who may or may not adhere to the stipulated payment deadlines.

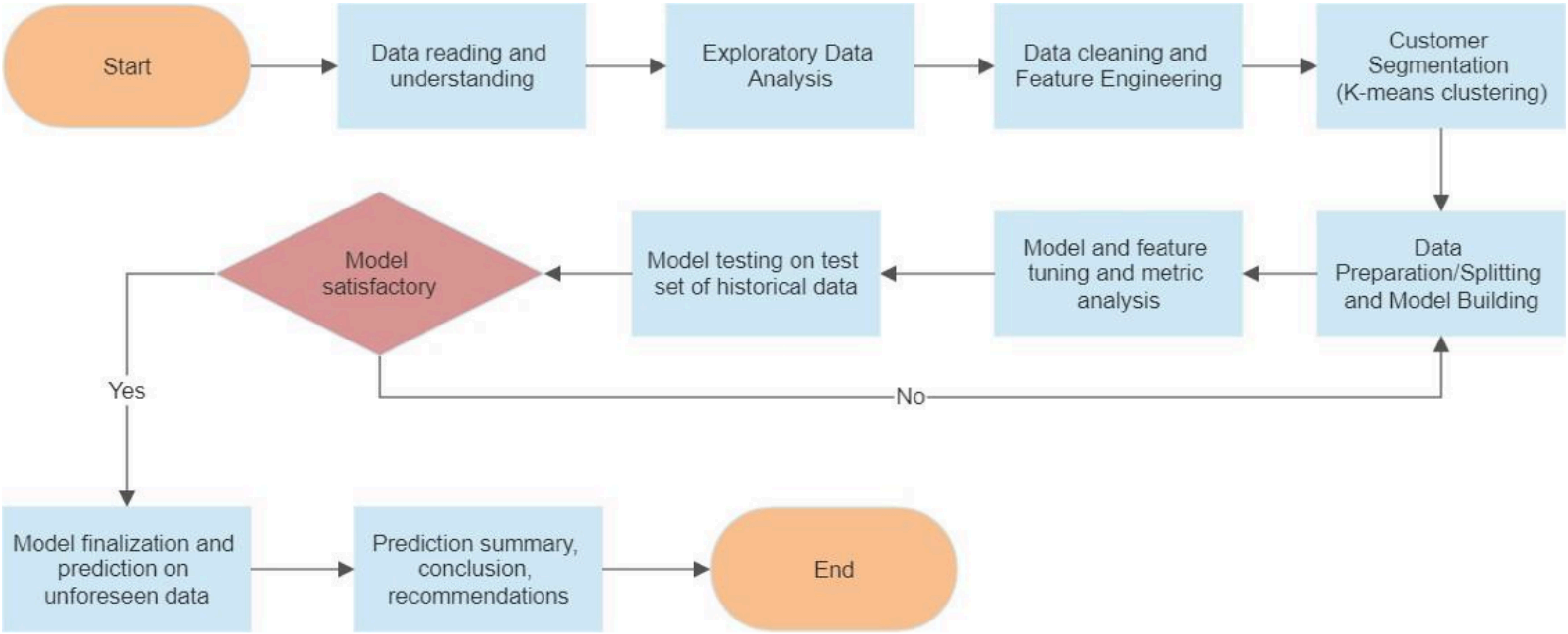
- Delayed payments from vendors cause financial setbacks and losses, disrupting smooth business operations.
- Furthermore, company employees spend significant time chasing overdue payments, leading to unproductive efforts and wasted resources.

Problem identification

Schuster, a sports retail company engaged in B2B transactions, frequently extends credit to vendors who may or may not adhere to the stipulated payment deadlines.

- Delayed payments from vendors cause financial setbacks and losses, disrupting smooth business operations.
- Furthermore, company employees spend significant time chasing overdue payments, leading to unproductive efforts and wasted resources.

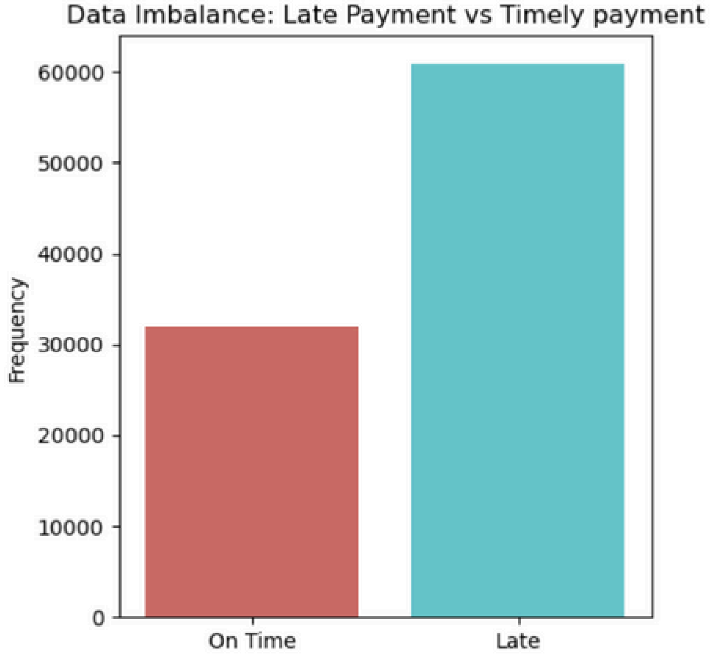
Approach

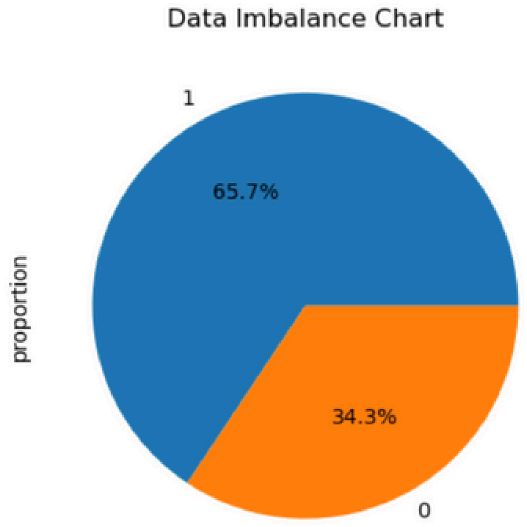


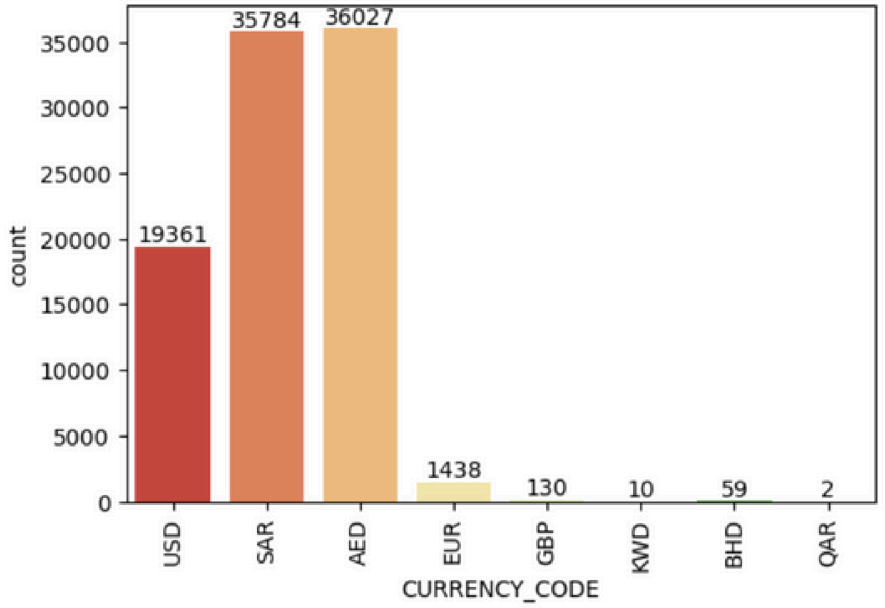
Start
Data reading and understanding
Exploratory Data Analysis
Data cleaning and Feature Engineering
Customer Segmentation (K-means clustering)
Data Preparation/Splitting and Model Building
Model and feature tuning and metric analysis
Model testing on test set of historical data
Model satisfactory
No
Yes
Model finalization and prediction on unforeseen data
Prediction summary, conclusion, recommendations
End

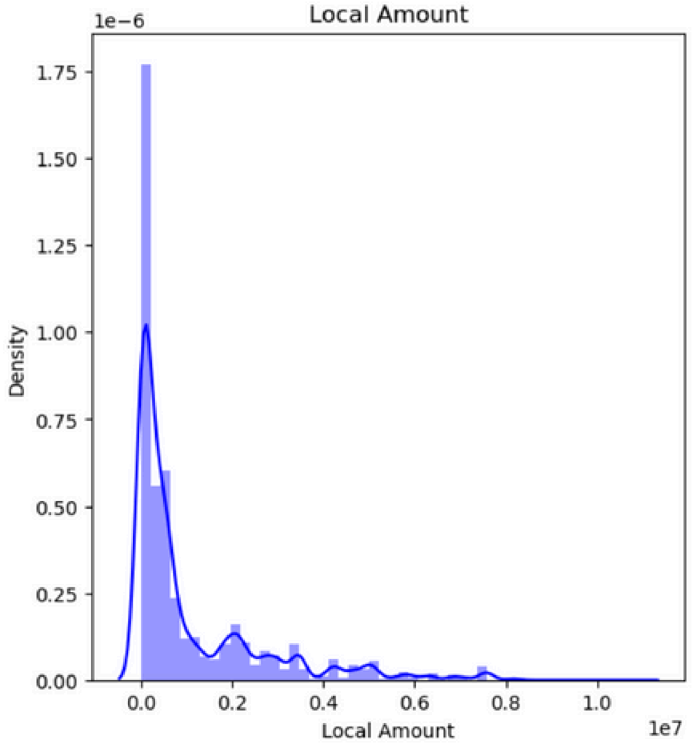
Class Imbalance & Transaction Insights

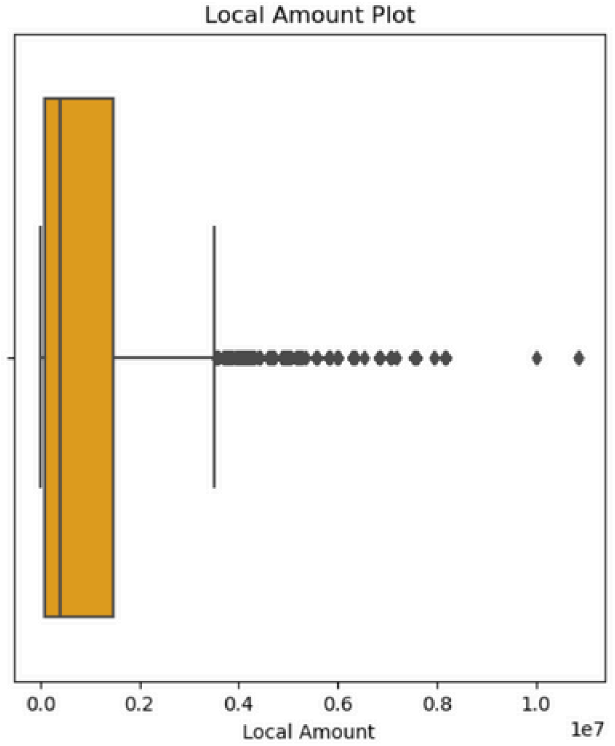
- The class is imbalanced, with 65.7% of transactions involving payment delayers. This level of imbalance is acceptable and does not require correction.
- The company's top three currencies are AED, SAR, and USD, with AED being the most commonly used, indicating a higher volume of transactions in the Middle East.
- Transaction values mostly range from $1 to $3 million, with the majority below $1.75 million.
- Transactions in this range are the most frequent.



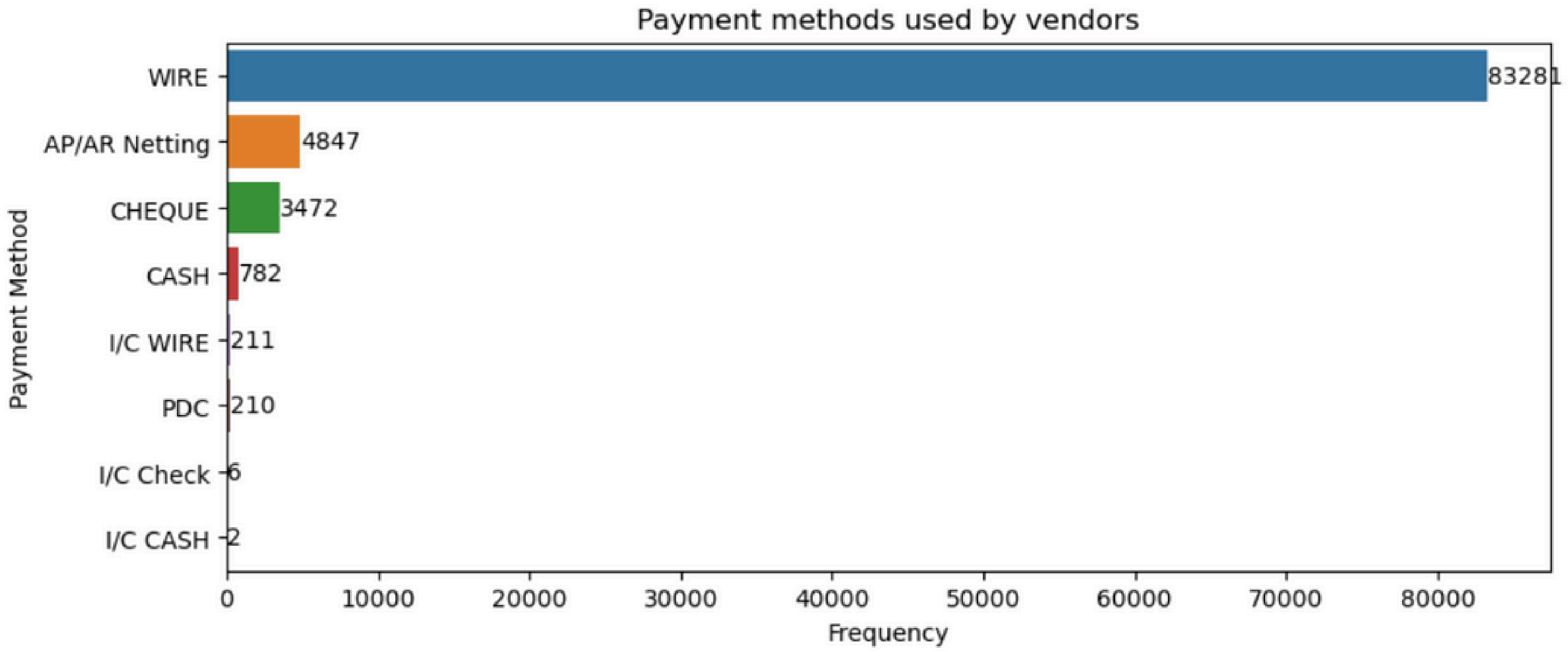




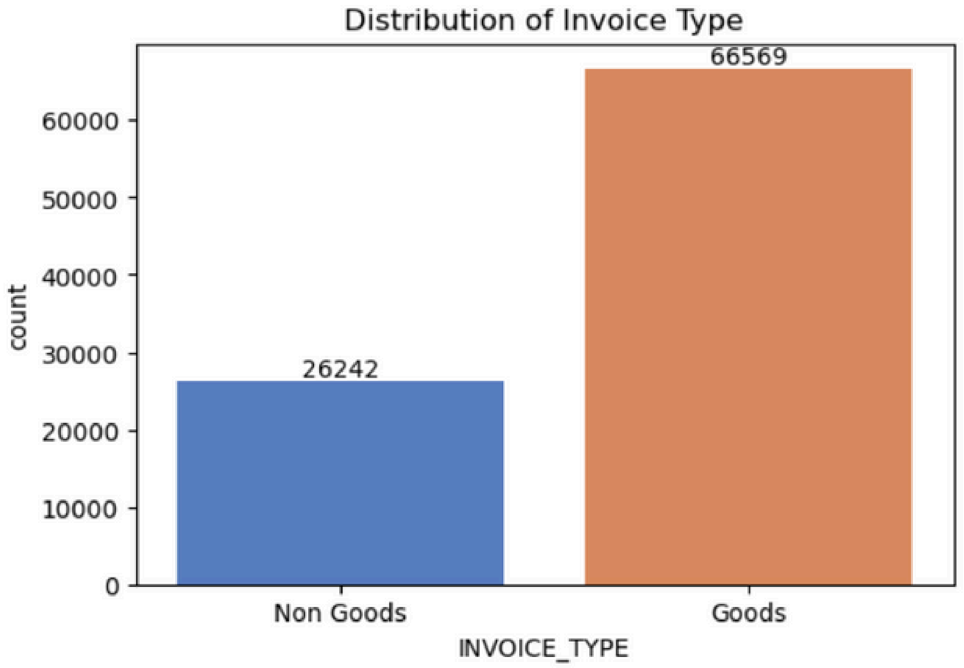


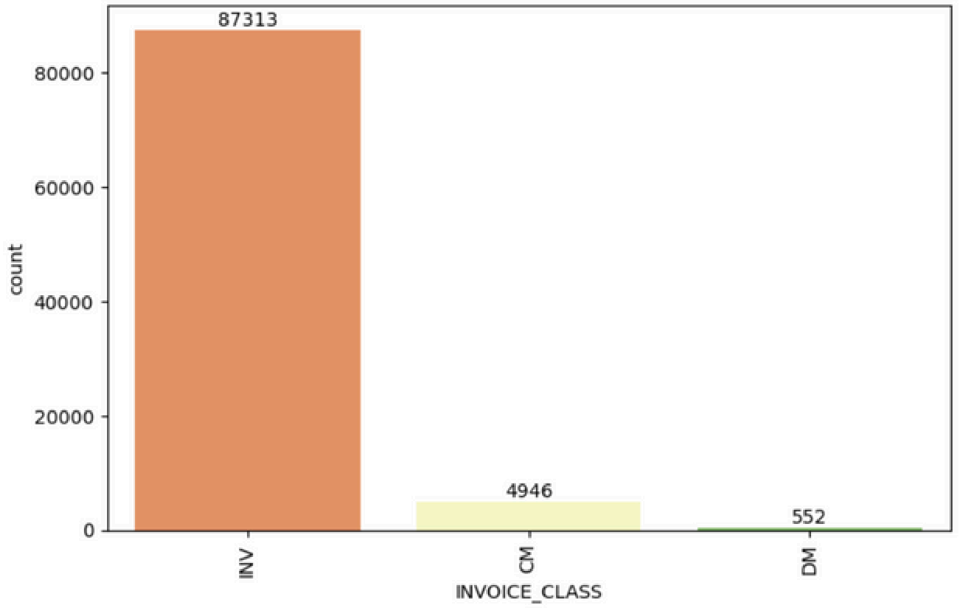


Class Imbalance & Transaction Insights

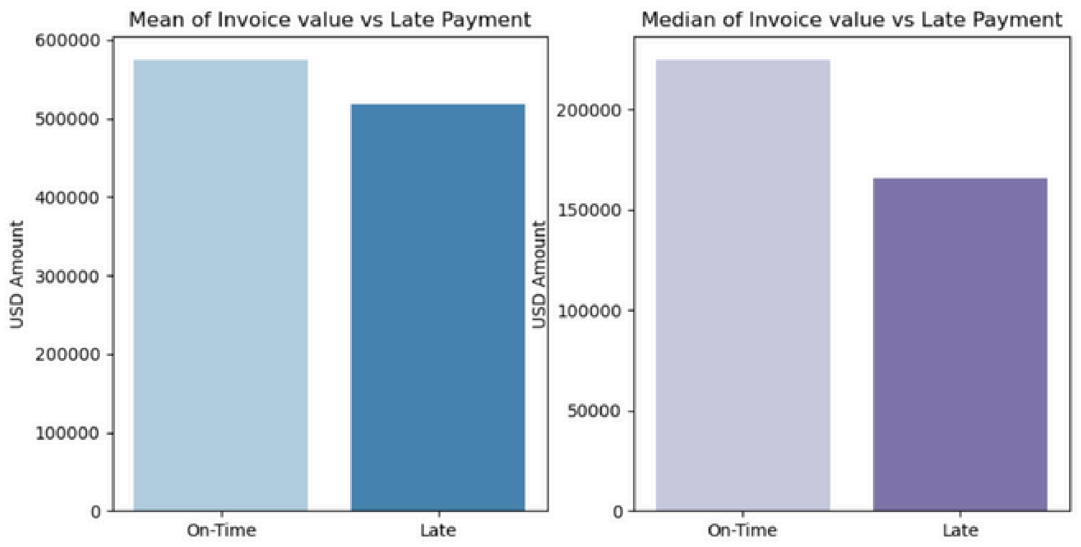


- The company receives most payments through wire transfers, followed by netting, cheques, and cash.
- Goods-type invoices make up the majority of all invoices generated.
- The most common invoice category is "Invoice," while the other categories account for a very small percentage.

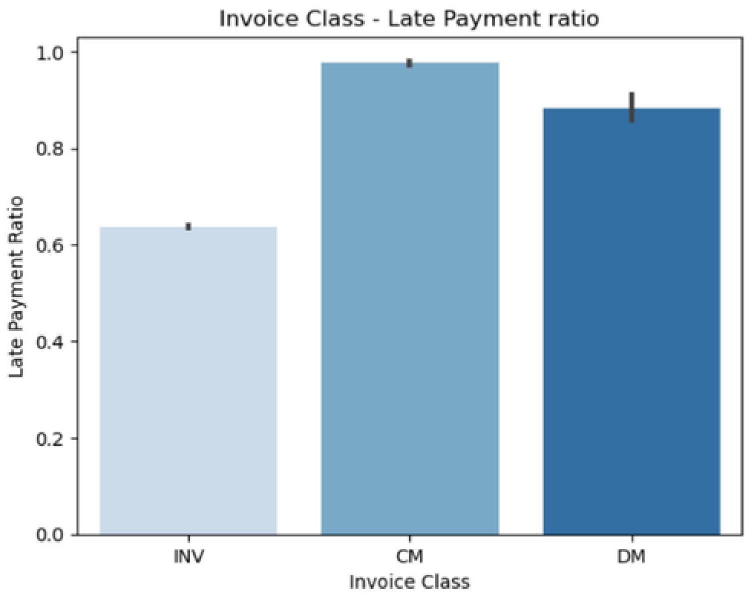




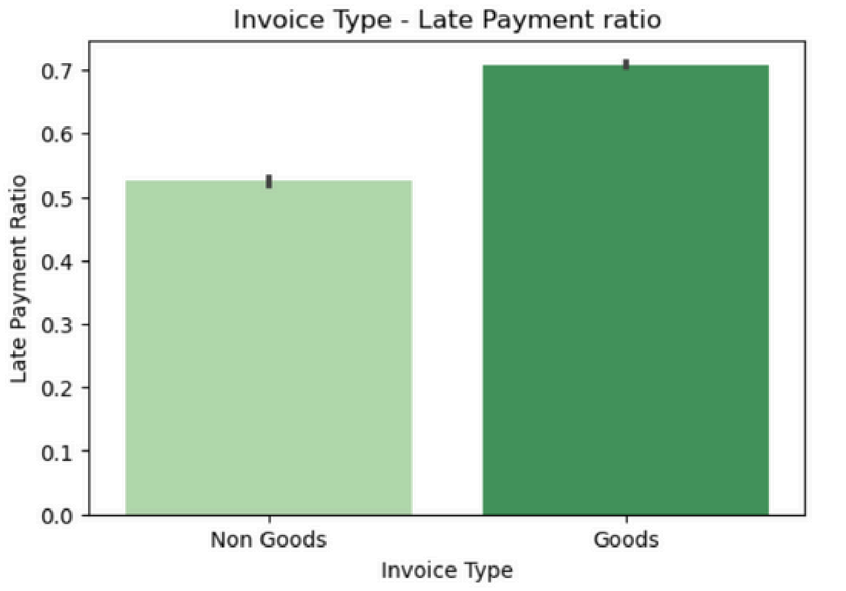
Characteristics of payment types of Defaulters



The average and median payment amounts are higher for on-time payers compared to late payers, indicating that higher-value transactions are less likely to be delayed than lower-value ones.



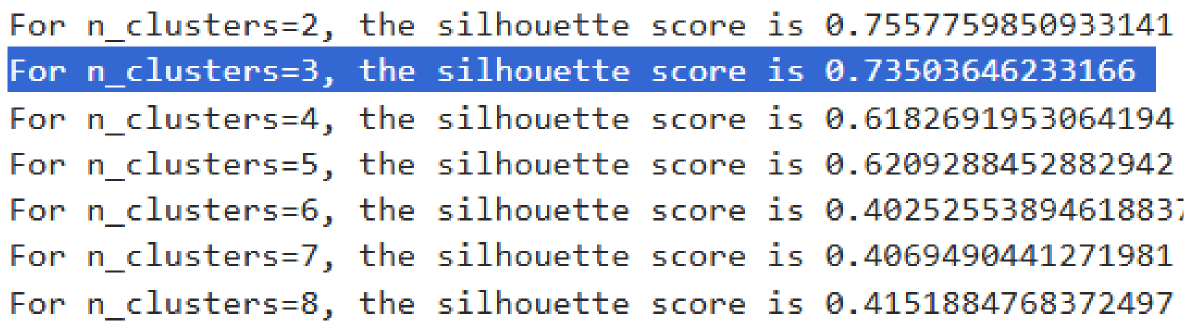
Credit Note transactions have the highest late payment ratio, followed by Debit Notes and Invoices, suggesting a higher risk of delays in Credit and Debit Note invoice categories.



Goods-type invoices have a higher late payment ratio than non-goods, indicating a greater likelihood of payment delays.

Customer segmentation using K-means clustering

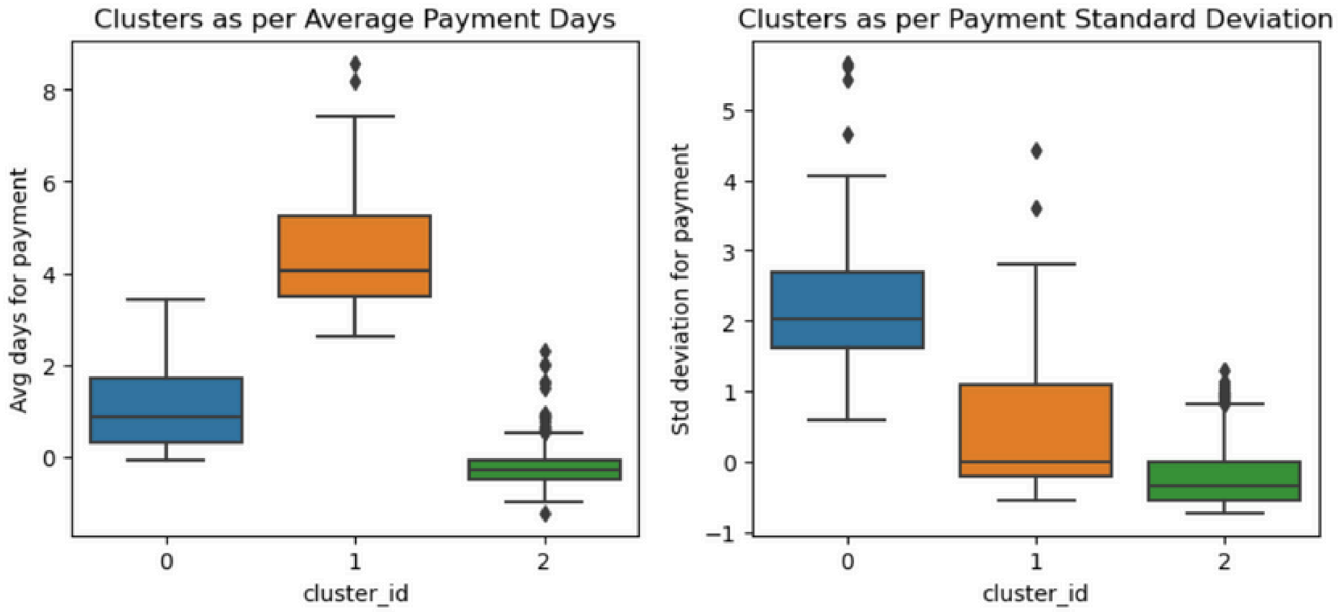
- One of the objectives was to classify customers based on their payment behaviors. This was accomplished using K-means clustering, considering the average and standard deviation of the number of days vendors took to make payments.

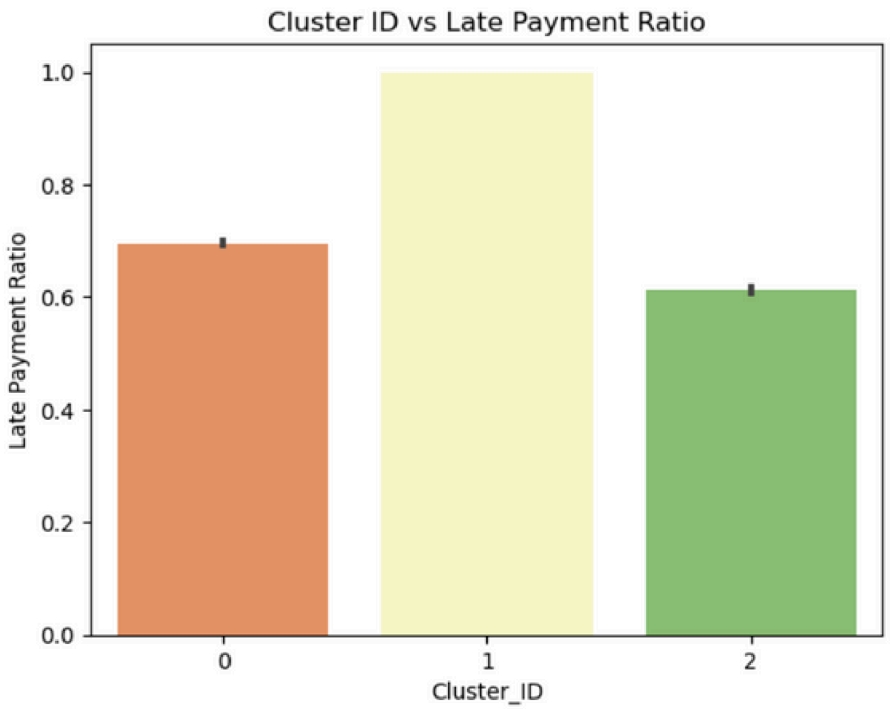


```
For n_clusters=2, the silhouette score is 0.7557759850933141
For n_clusters=3, the silhouette score is 0.73503646233166
For n_clusters=4, the silhouette score is 0.6182691953064194
For n_clusters=5, the silhouette score is 0.6209288452882942
For n_clusters=6, the silhouette score is 0.40252553894618837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

- The number of clusters was set to 3 because adding more than 3 clusters caused the silhouette score to drop significantly.

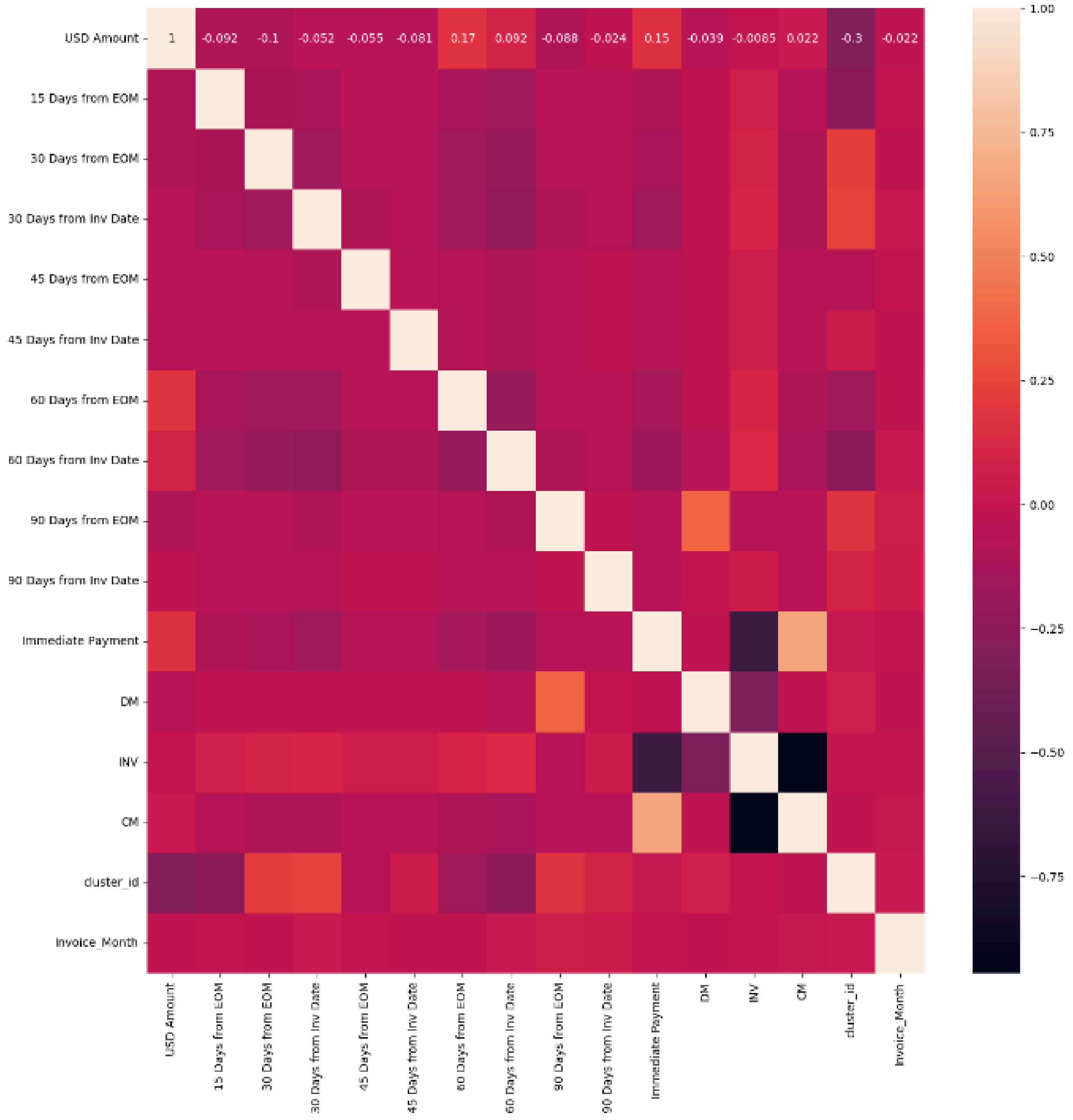
- Category 2 consists of early payers who take the least time to make payments. Category 1 includes prolonged payers who take the longest time to pay. Category 0 falls between the two and is labeled as medium-duration payers.





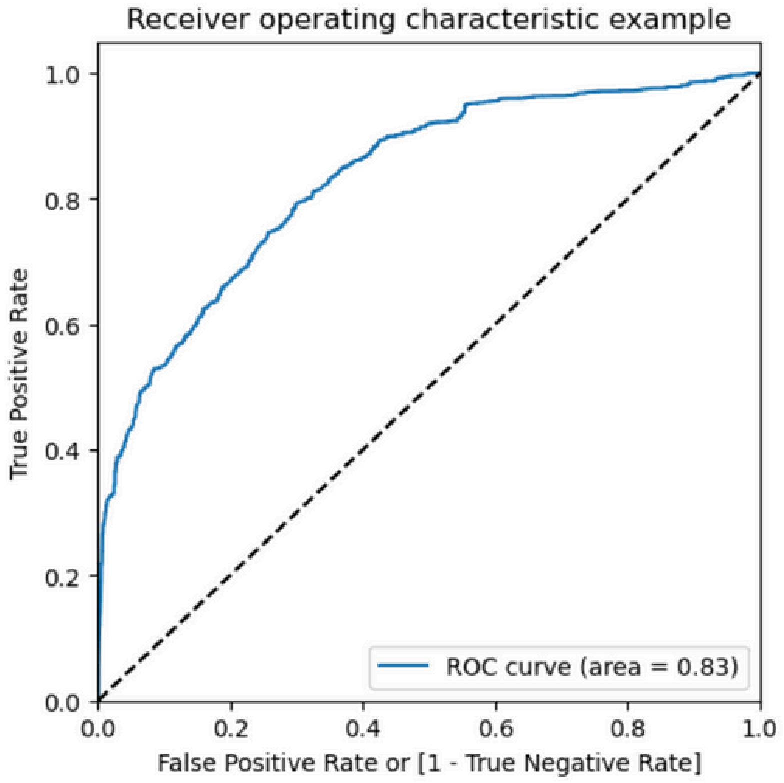
- It was observed that prolonged payers have historically shown much higher rates of payment delays compared to early or medium-duration payers

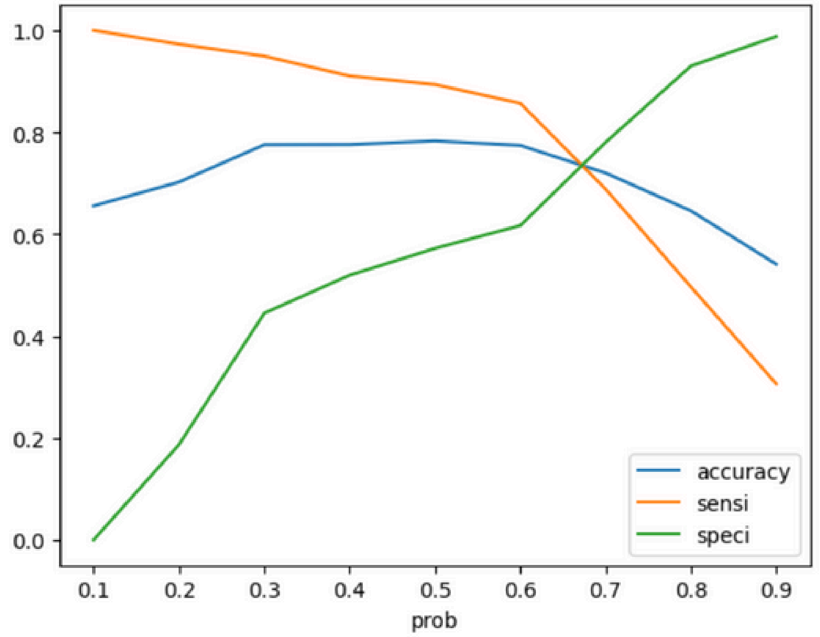
Model Building



- CM & INV, INV & Immediate Payment, and DM & 90 days from EOM show high multicollinearity. Therefore, these columns are being dropped to avoid the impact of multicollinearity.

Comparison between two models, logistic regression and random forests



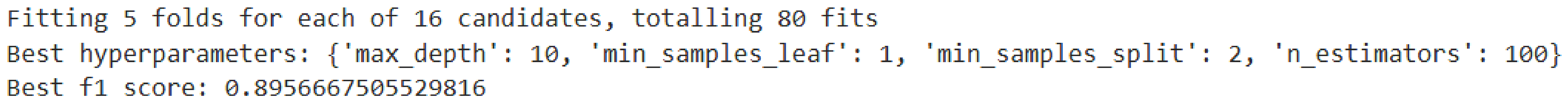


- After removing multicollinear and unnecessary variables, the logistic regression model retained variables with acceptable p-values and VIFs. No further feature elimination was needed, and the model achieved a good ROC curve area of 0.83.

- The trade-off plot between accuracy, sensitivity, and specificity identified an optimal probability cutoff of ˜0.6. This cutoff was used to predict delayed payments in the received payments dataset.

Comparison between two models, logistic regression and random forests

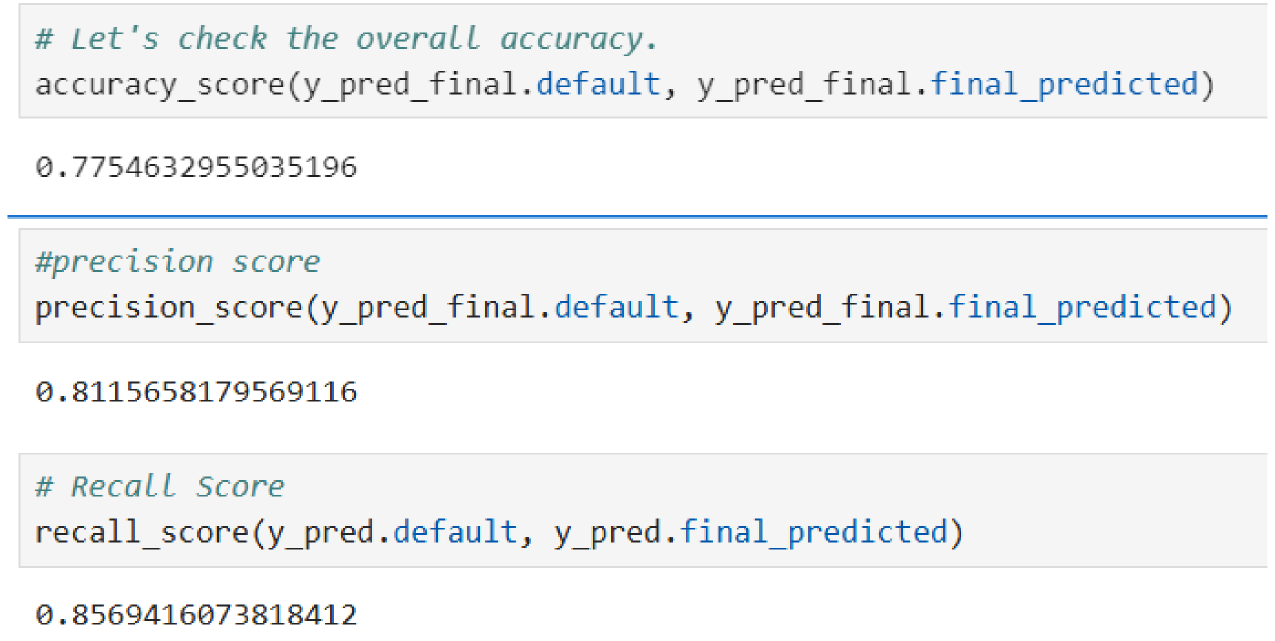
A random forest model was created using the same parameters as the logistic regression model, along with hyperparameter tuning, resulting in the following optimized settings.



```
Fitting 5 folds for each of 16 candidates, totalling 80 fits
Best hyperparameters: {'max_depth': 10, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 100}
Best f1 score: 0.8956667505529816
```

A random forest model was built using the parameters above, and its performance metrics were compared to those of the logistic regression model. The final model was then selected based on this comparison.

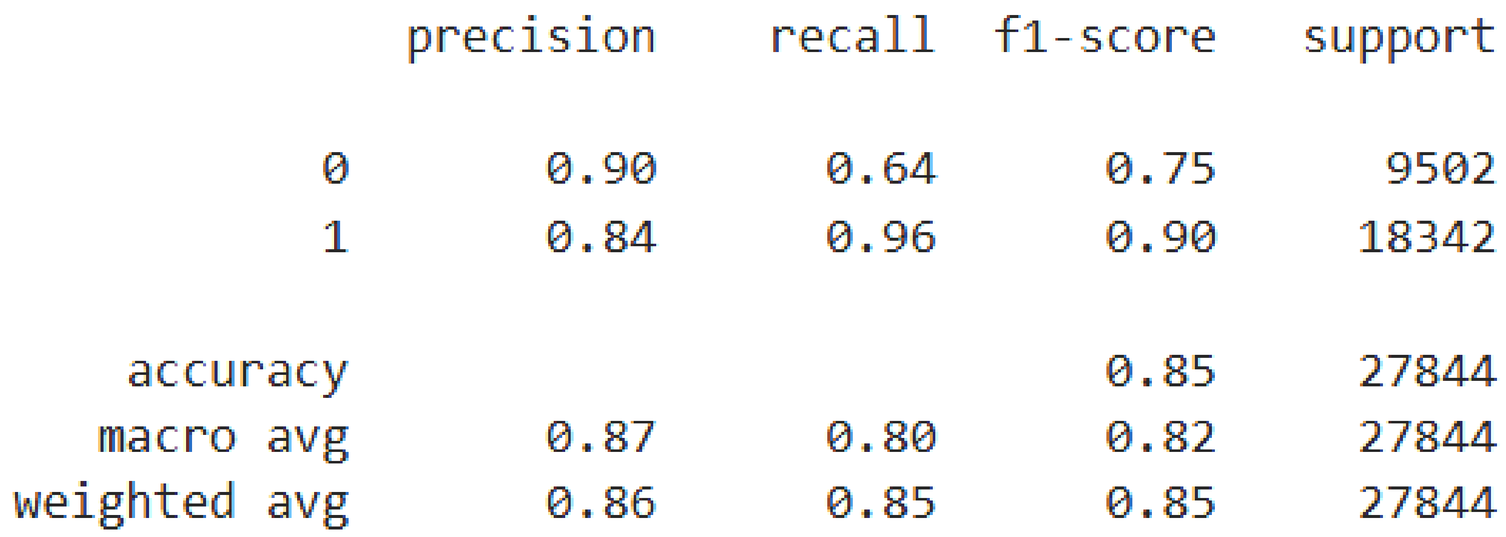
Random Forest found better than Logistic Regression



```
# 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
```

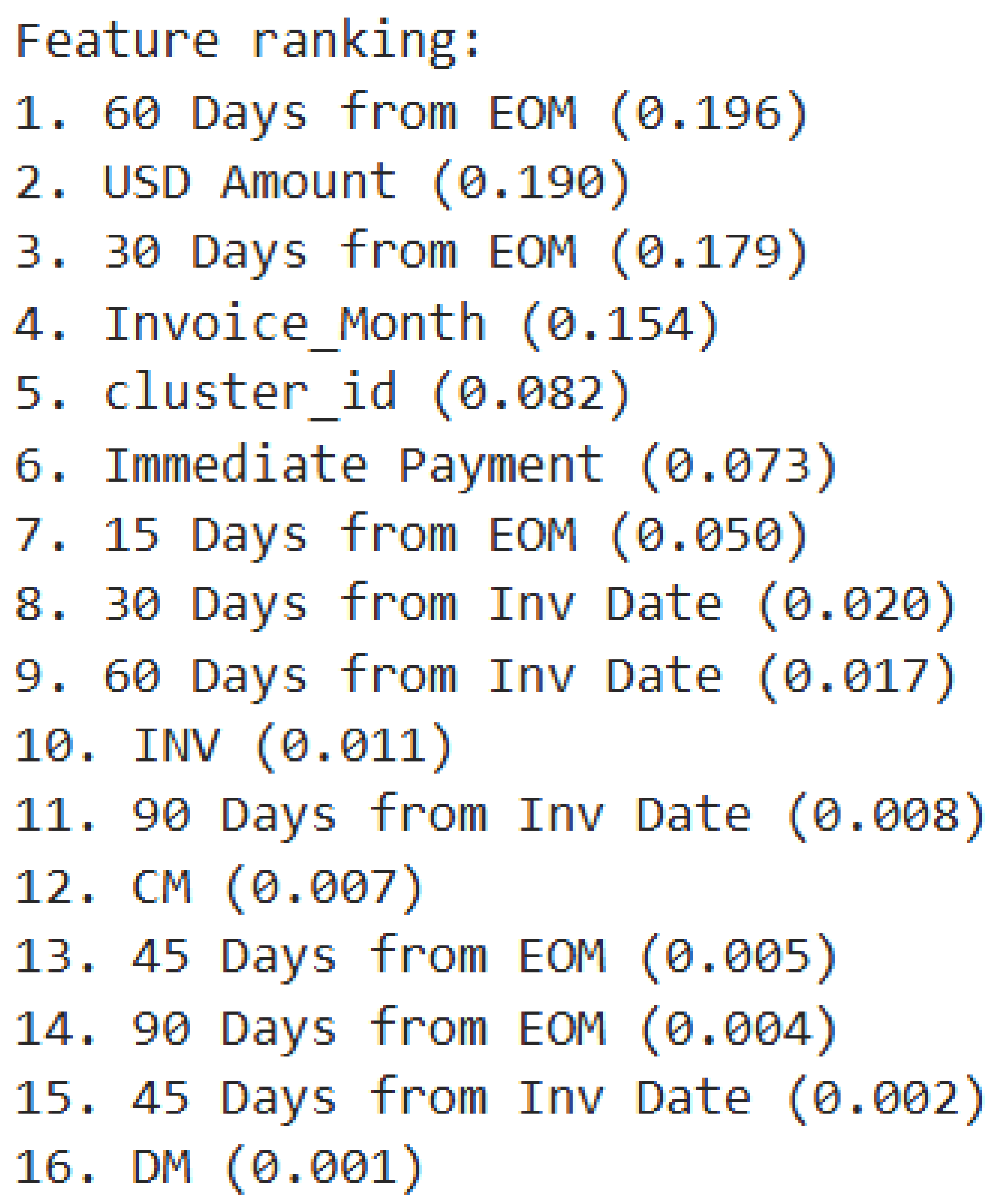
```
# Recall Score
recall_score(y_pred.default, y_pred.final_predicted)

0.8569416073818412
```



```
              precision    recall  f1-score   support

           0       0.90      0.64      0.75      9502
           1       0.84      0.96      0.90     18342

    accuracy                           0.85     27844
   macro avg       0.87      0.80      0.82     27844
weighted avg       0.86      0.85      0.85     27844
```

- The Random Forest model outperformed the Logistic Regression model in both precision and recall scores. Recall was particularly important in this case to better identify late payers for targeting.
- Given that the data heavily relies on categorical variables, Random Forest was more suitable for the task compared to Logistic Regression.
- Therefore, the Random Forest model was chosen as the final model for making predictions.

Random Forest Feature Ratings



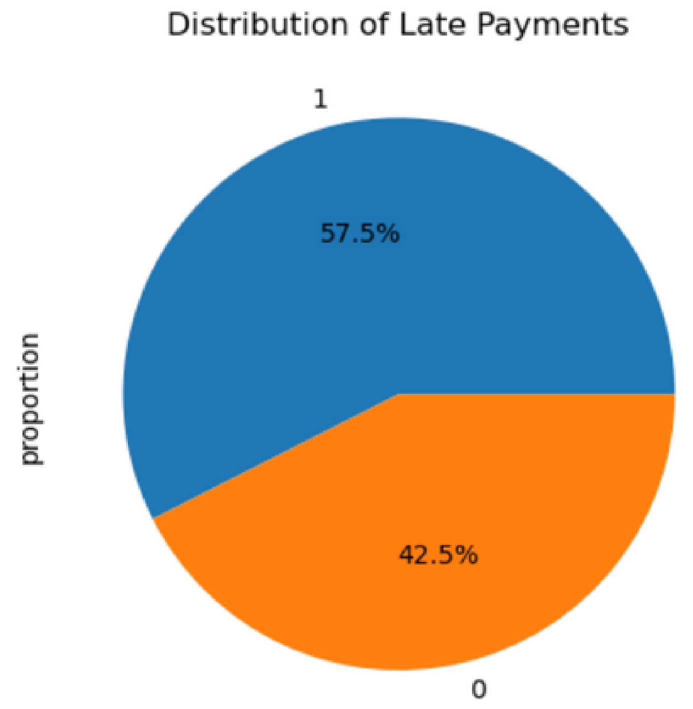
```
Feature ranking:
1. 60 Days from EOM (0.196)
2. USD Amount (0.190)
3. 30 Days from EOM (0.179)
4. Invoice_Month (0.154)
5. cluster_id (0.082)
6. Immediate Payment (0.073)
7. 15 Days from EOM (0.050)
8. 30 Days from Inv Date (0.020)
9. 60 Days from Inv Date (0.017)
10. INV (0.011)
11. 90 Days from Inv Date (0.008)
12. CM (0.007)
13. 45 Days from EOM (0.005)
14. 90 Days from EOM (0.004)
15. 45 Days from Inv Date (0.002)
16. DM (0.001)
```

The Random Forest model was used to rank features, identifying the top 5 factors for predicting payment delays:

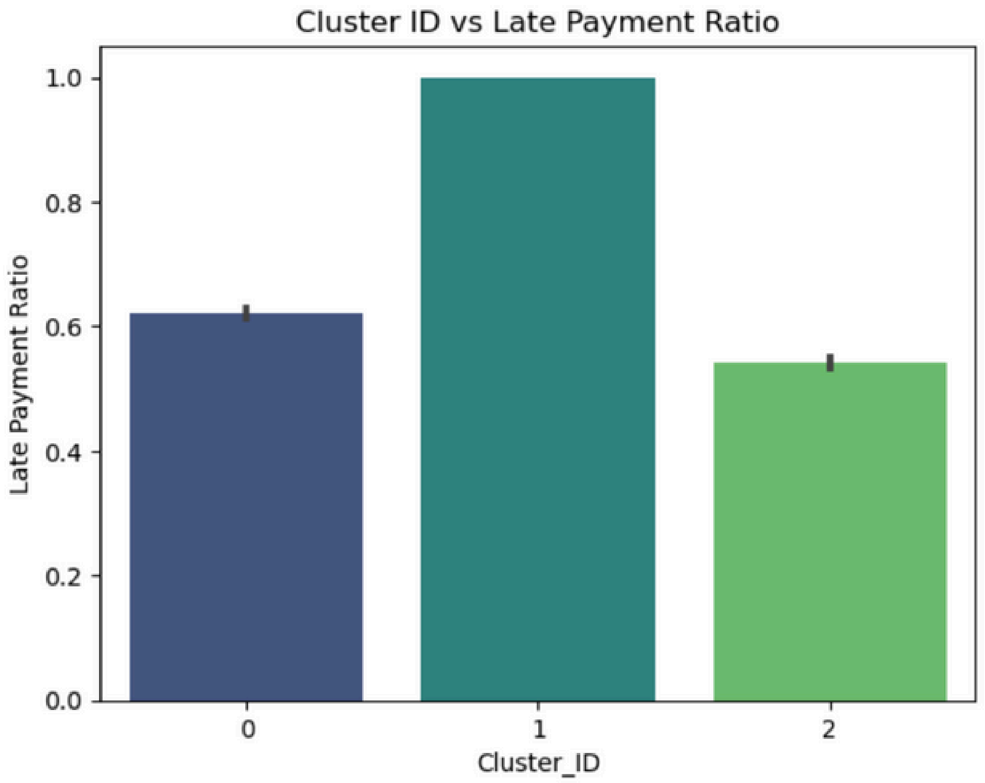
- 60 Days from EOM
- USD Amount
- 30 Days from EOM
- Invoice Month
- Cluster-ID (based on the average and standard deviation of days taken to make a payment)

The customer segments determined by Cluster-ID were then matched with the open-invoice data using customer names, and predictions were made accordingly.

50% payments predicted to be delayed as per Openinvoice data, prolonged payment days to observe alarmingly high delay rates

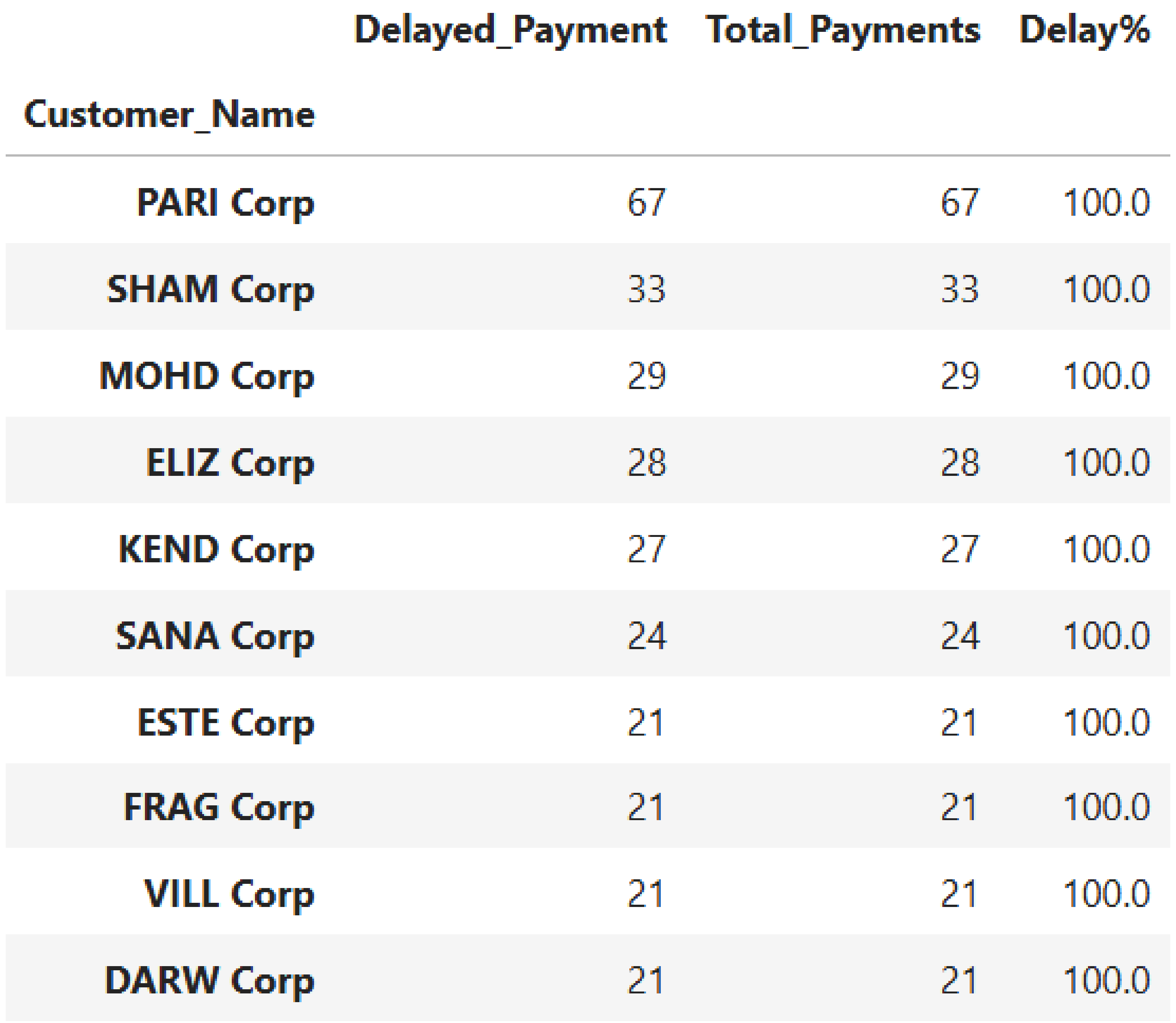


The final model predicts that approximately 57.5% of transactions are likely to experience payment delays, potentially causing significant disruptions to business operations.



Customers with a history of prolonged payment days are expected to have the highest delay rate (around 100%) compared to those with early or medium payment histories, consistent with past trends.

Customers with the highest delay probabilities



	Delayed_Payment	Total_Payments	Delay%
Customer_Name			
PARI Corp	67	67	100.0
SHAM Corp	33	33	100.0
MOHD Corp	29	29	100.0
ELIZ Corp	28	28	100.0
KEND Corp	27	27	100.0
SANA Corp	24	24	100.0
ESTE Corp	21	21	100.0
FRAG Corp	21	21	100.0
VILL Corp	21	21	100.0
DARW Corp	21	21	100.0

The predictions indicate that the companies listed in the table have the highest probability of default, with the most delayed and total payments.

Recommendations

Inferences from Clustering Analysis:

1. Credit Note Payments:
   - These observe the highest delay rate compared to Debit Note or Invoice type classes.
   - Recommendation: Implement stricter company policies for payment collection related to credit note invoices.
2. Goods vs Non-Goods:
   - Goods-type invoices have significantly higher payment delay rates than non-goods.
   - Recommendation: Stricter payment policies should be applied to goods-type invoices.
3. Focus on Lower-Value Payments:
   - Lower-value payments constitute the majority of transactions and show higher rates of late payments.
   - Recommendation: Introduce penalties based on billing amounts, with a higher penalty percentage for smaller bills, as a last resort.
4. Customer Segments:
   - Customers were clustered into three categories:
     - Cluster 0: Medium payment duration.
     - Cluster 1: Prolonged payment duration.
     - Cluster 2: Early payment duration.
   - Customers in Cluster 1 exhibit significantly higher delay rates.
   - Recommendation: Focus extensively on Cluster 1 customers to address delays.
5. High-Probability Companies:
   - Companies with the highest probability of delay and total delayed payment counts should be prioritized.
   - Recommendation: Pay close attention to these companies to mitigate risks and improve collection efforts.