E-Commerce & Retail B2B Case Study

Team Members :

Jatin Sethi

Janaki T R

Harsh Nevatia

Problem statement

- Schuster is a multinational retailer of sports goods and garments. It conducts significant business with hundreds of vendors with whom it has arranged a credit arrangement.
- However, not all its vendors respect the credit terms and tend to make late payments. While Schuster imposes a fine for every late payment made by the vendor, but such an approach is not seemingly beneficial for the long-term relationships of either of the parties.
- The collectors (Schuster employees who collect and reconcile the payments against the invoices raised) must keep chasing the vendors to ensure the payment is received on time. In case of late payments, they spend considerable time coordinating the payments. This resulted in many non-value-added activities, loss of time and effort and financial impact on Schuster.

Approach Strategy to the Problem

1. Data Exploration & Preprocessing

Data Ingestion and Understanding

Exploratory Data Analysis (EDA)

Data Cleaning and Feature Engineering

2. Segmentation & Model Development

Customer Segmentation (K-means Clustering)

Data Splitting and Model Training

Feature Tuning and Performance Optimization

3. Model Evaluation & Insights

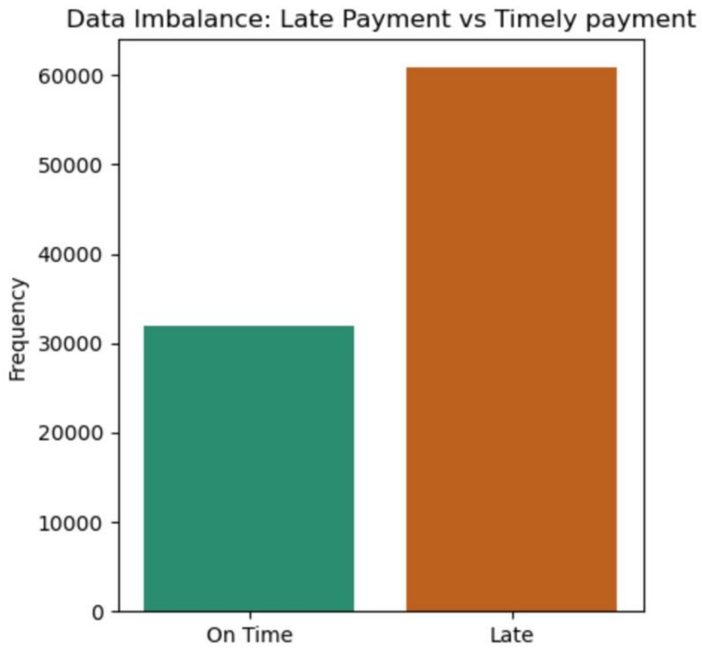
Model Validation on Historical Data

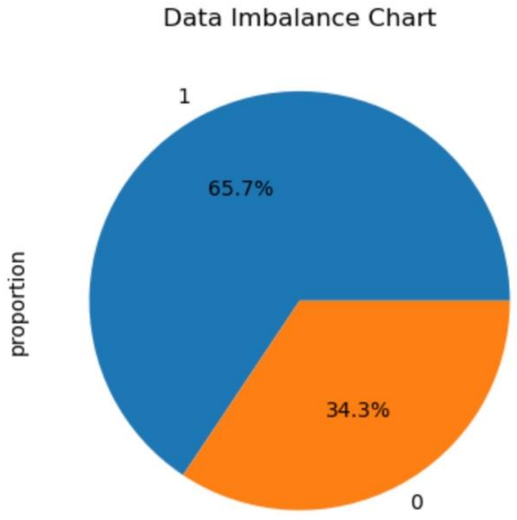
Final Model Deployment and Prediction

Summary of Insights and Recommendations

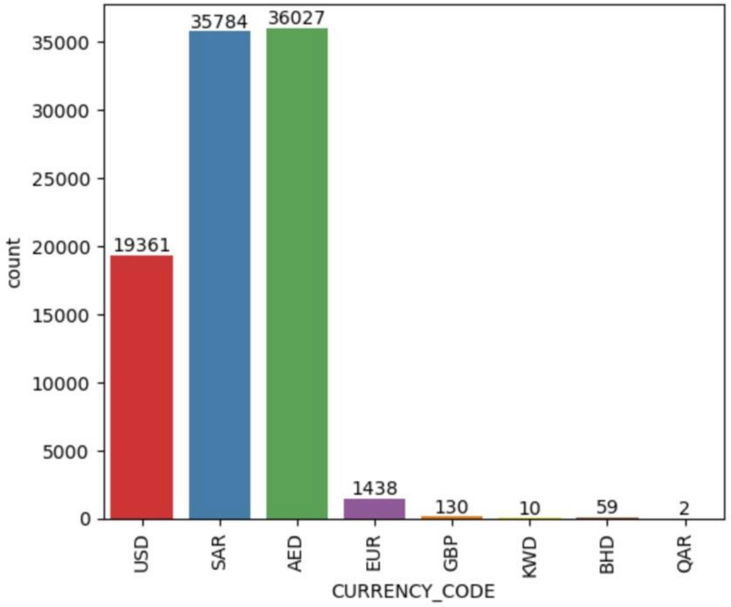
Class Distribution & Transaction Patterns (Univariate Analysis)

- There is a 65.7% imbalance towards payment delayers, which is within an acceptable range and does not require imbalance correction.

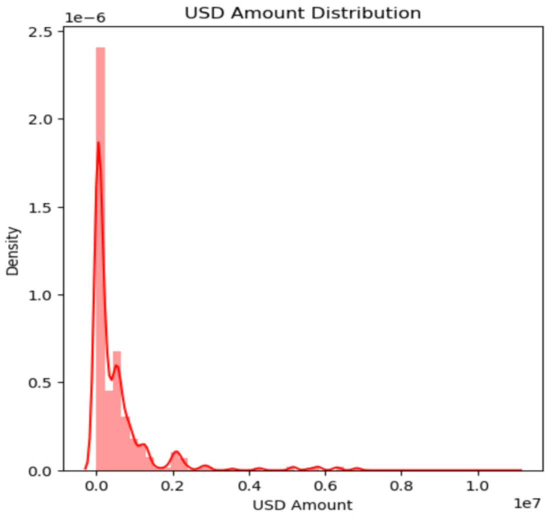


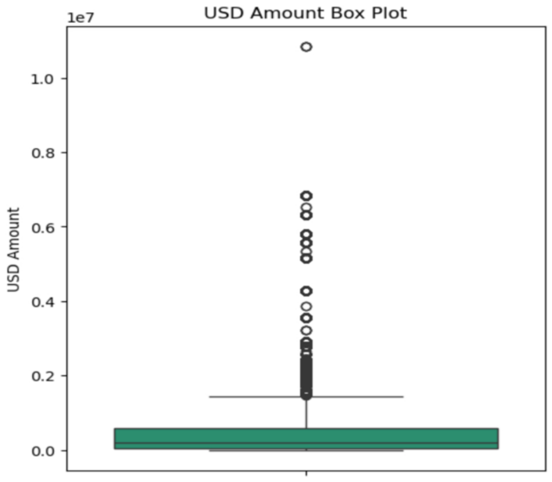


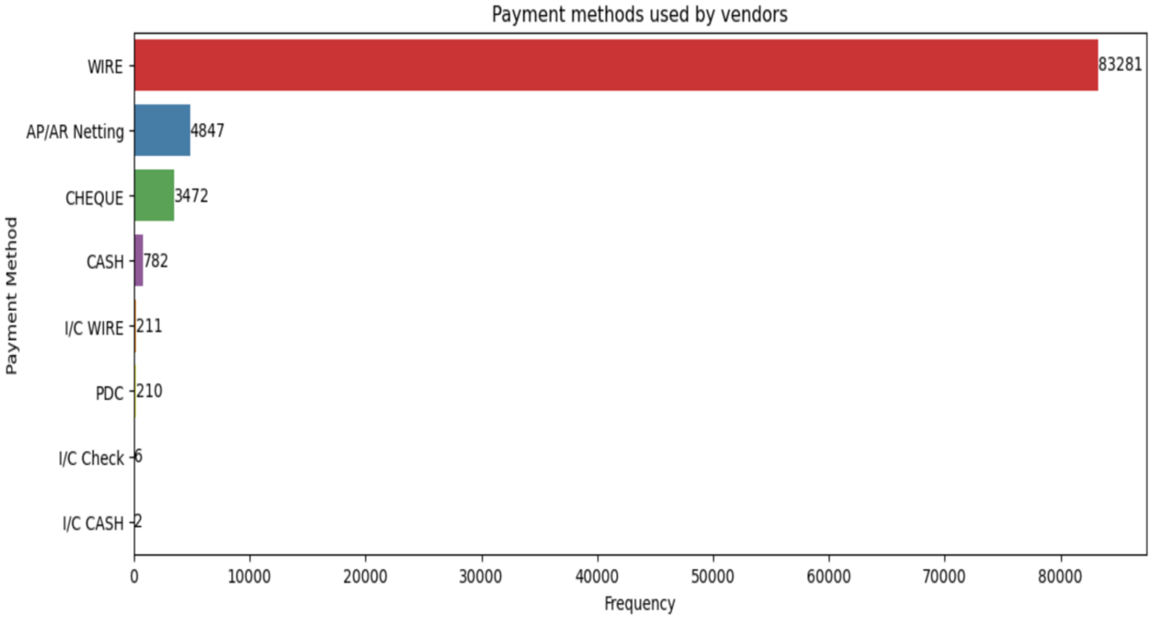
- The top three currencies used in transactions are AED, SAR, and USD, with AED being the most frequently used, indicating a higher volume of transactions with the Middle East.



- Transaction values typically range between $1 and $3 million. The majority of transactions occur at values below approximately $1.75 million.

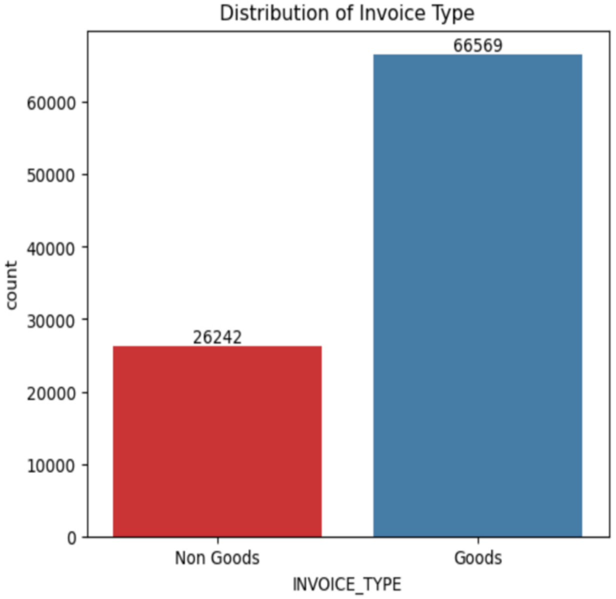


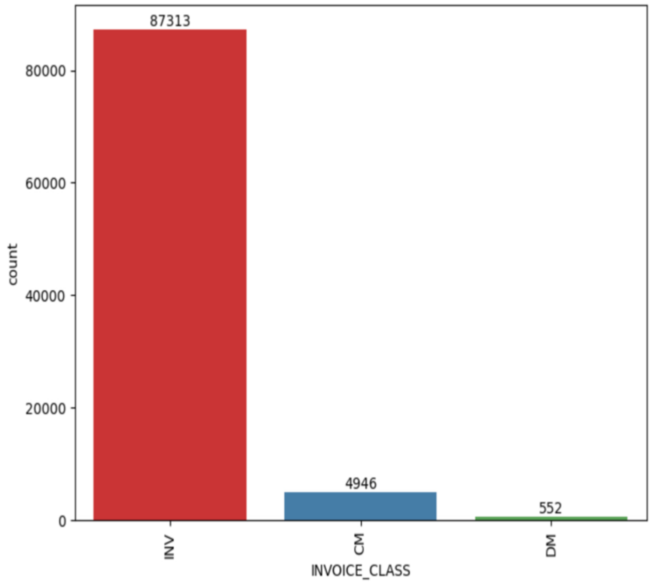




- The Wire payment method is the most frequently received by the company, followed by Netting, Cheque, and Cash.

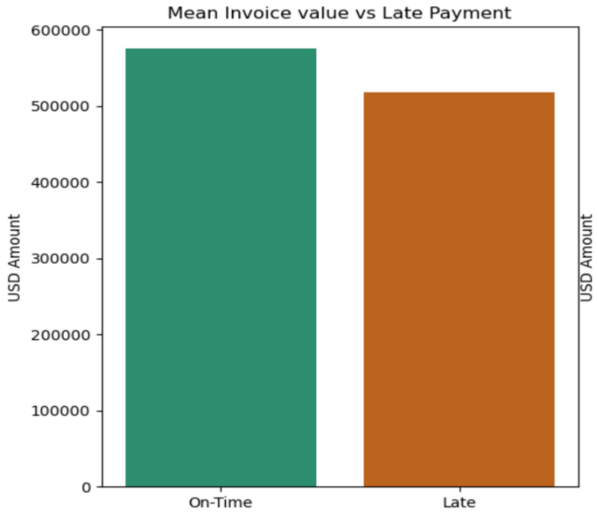
- Goods-type invoices make up the majority of the total invoices generated.
- The 'Invoice' category holds the largest share, while other invoice classes contribute only a small percentage.

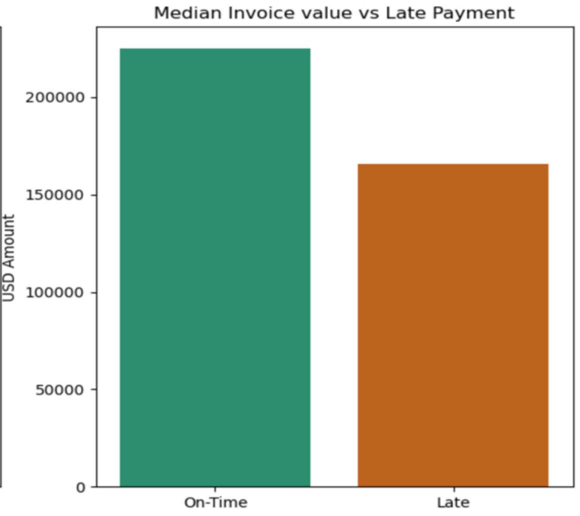


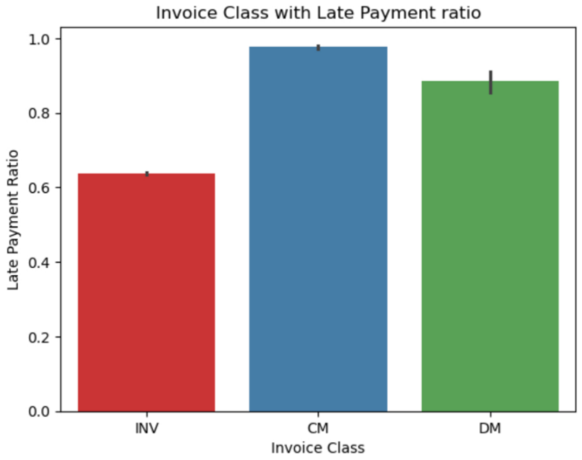


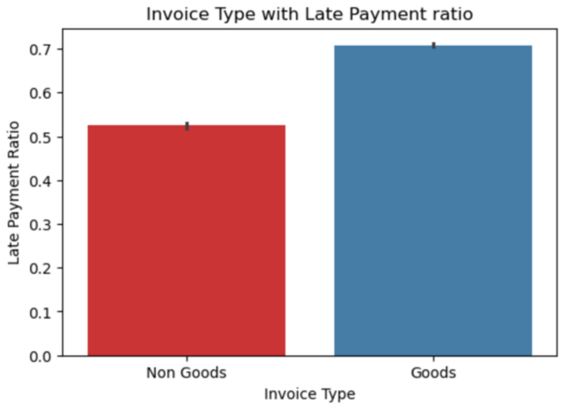
Defaulter Payment Patterns (Bivariate Analysis)

- The mean and median payment amounts are higher for on-time payers compared to late payers, indicating that higher-value transactions are less prone to delays than lower-value ones.
- The late payment ratio is highest for Credit Note transactions, followed by Debit Notes and Invoices, suggesting that Credit and Debit Note invoice classes carry a higher risk of payment delays.
- Goods-type invoices exhibit a higher late payment ratio compared to non-goods invoices, indicating an increased likelihood of payment delays in goods-related transactions.

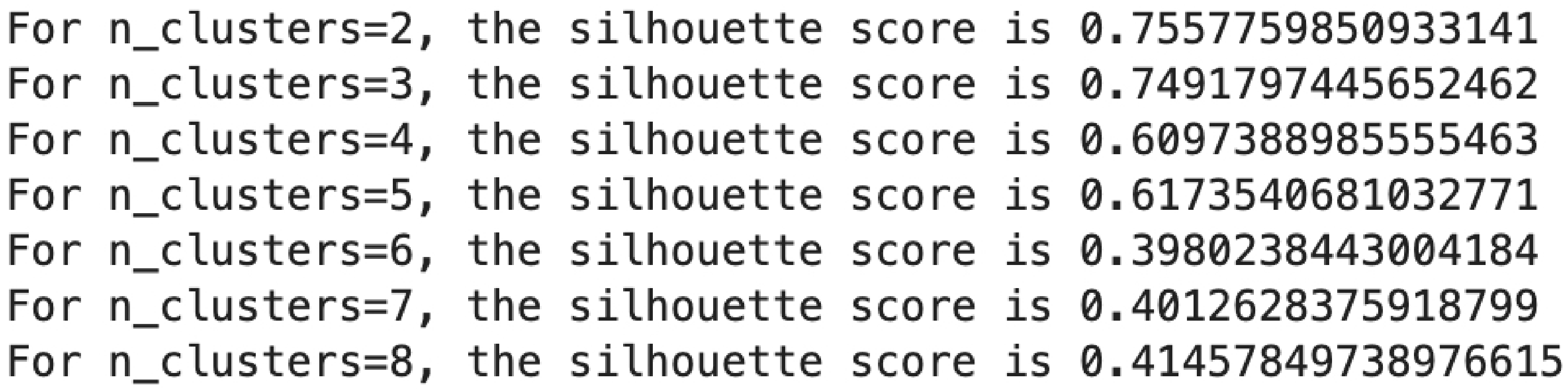








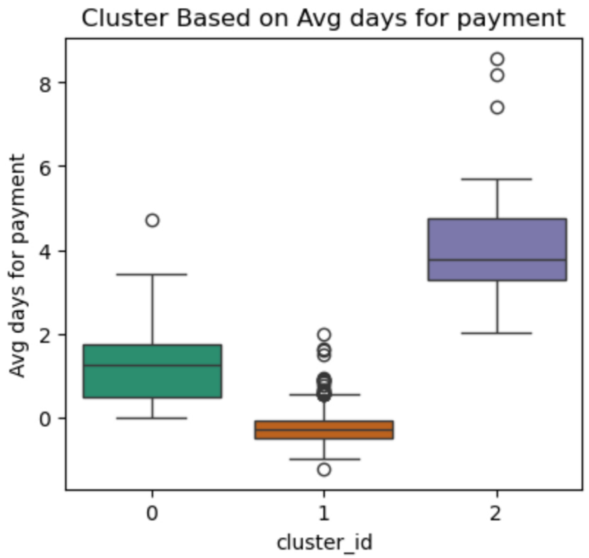
Customer Categorization via K-Means Clustering



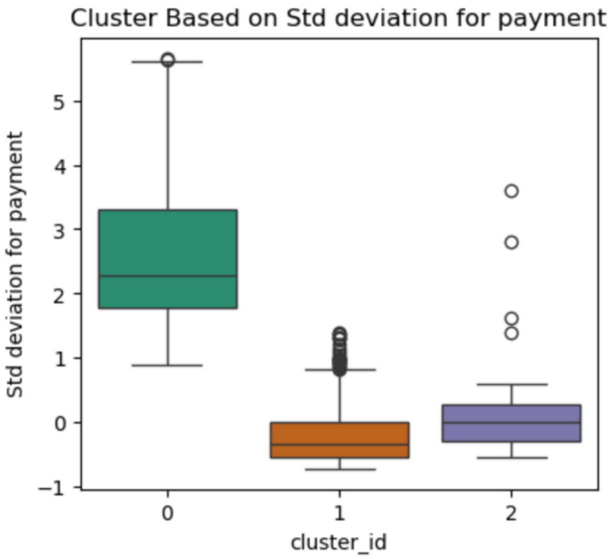
```
For n_clusters=2, the silhouette score is 0.7557759850933141
For n_clusters=3, the silhouette score is 0.7491797445652462
For n_clusters=4, the silhouette score is 0.6097388985555463
For n_clusters=5, the silhouette score is 0.6173540681032771
For n_clusters=6, the silhouette score is 0.3980238443004184
For n_clusters=7, the silhouette score is 0.4012628375918799
For n_clusters=8, the silhouette score is 0.41457849738976615
```

One of the key objectives was to classify customers based on their payment behaviour, which was achieved through K-Means clustering using the average and standard deviation of the number of days vendors took to make payments.

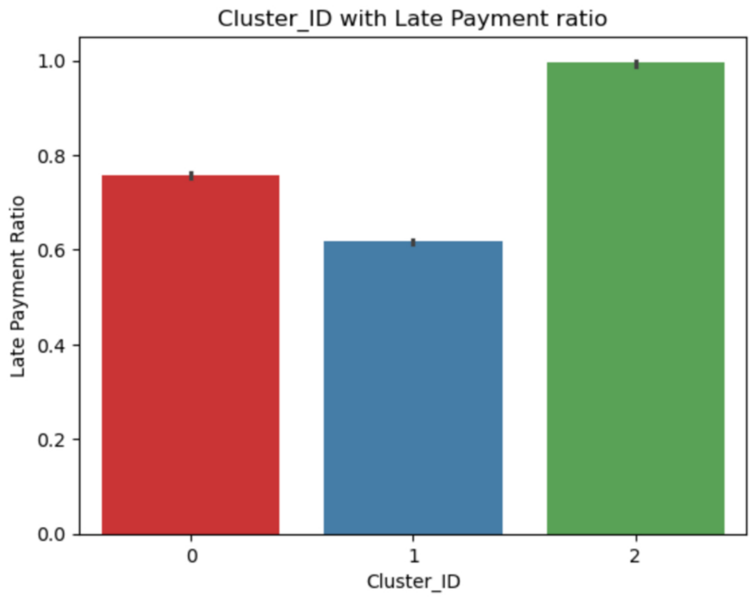
- Three clusters were chosen as the optimal number since increasing beyond three resulted in a significant drop in the silhouette score.



Cluster Based on Avg days for payment



Cluster Based on Std deviation for payment



Cluster_ID with Late Payment ratio

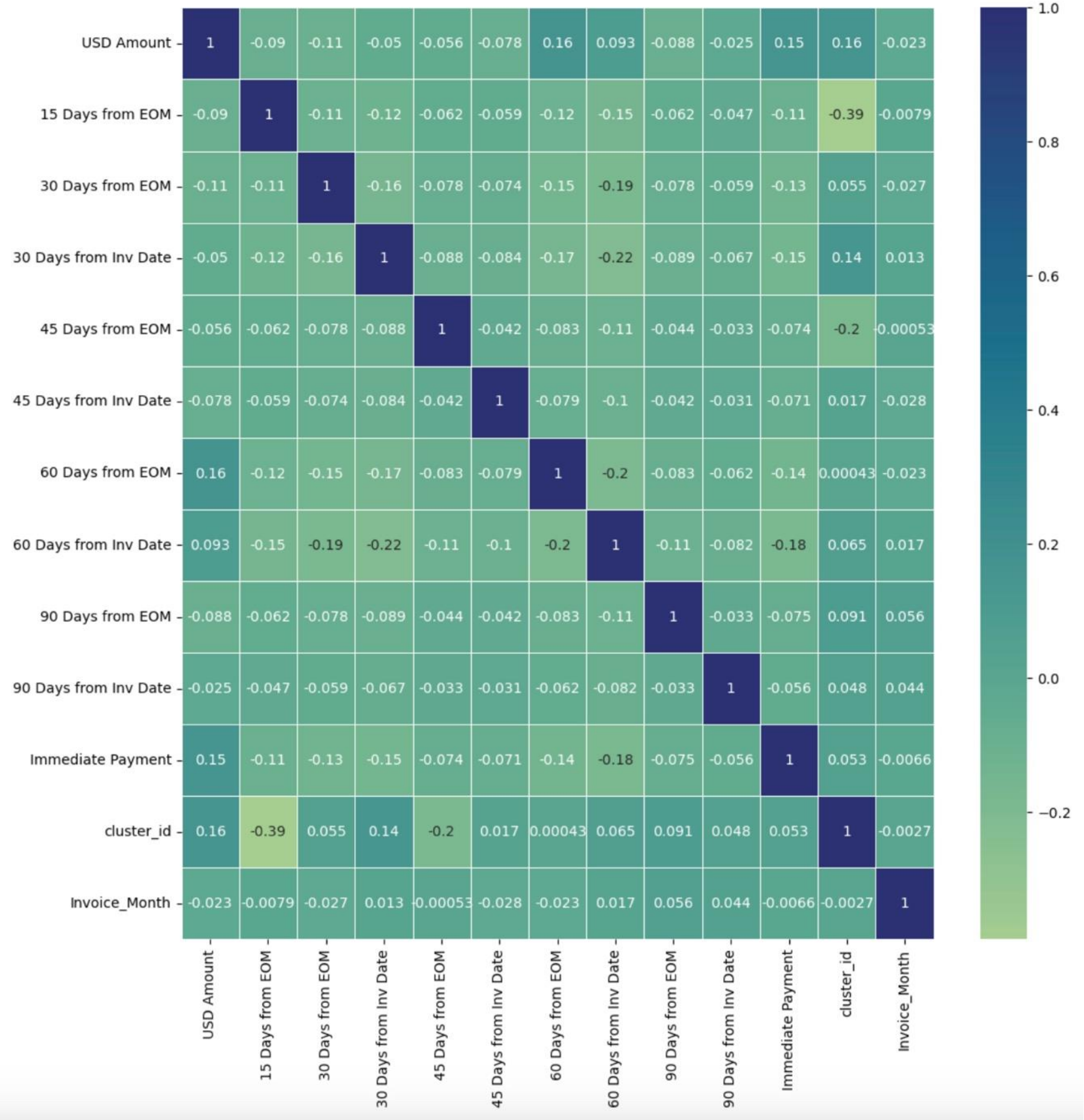
- Cluster 2 represents early payers, who took the least number of days to make payments.
- Cluster 1 consists of prolonged payers, who took the longest time to settle payments.
- Cluster 0 falls between the two and was categorized as medium-duration payers.
- Additionally, prolonged payers have historically shown a much higher rate of payment delays compared to early or medium-duration payers.

Model Building

High multicollinearity was observed in the following pairs:

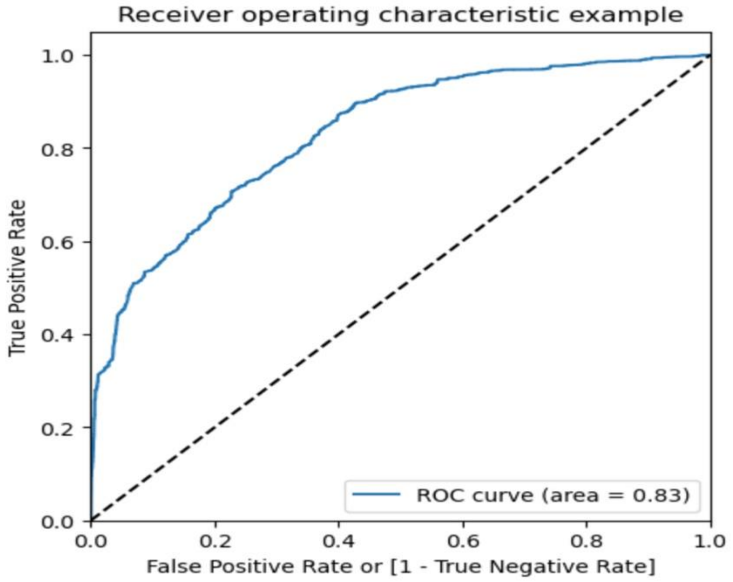
- CM & INV
- NV & Immediate Payment
- DM & 90 Days from EOM

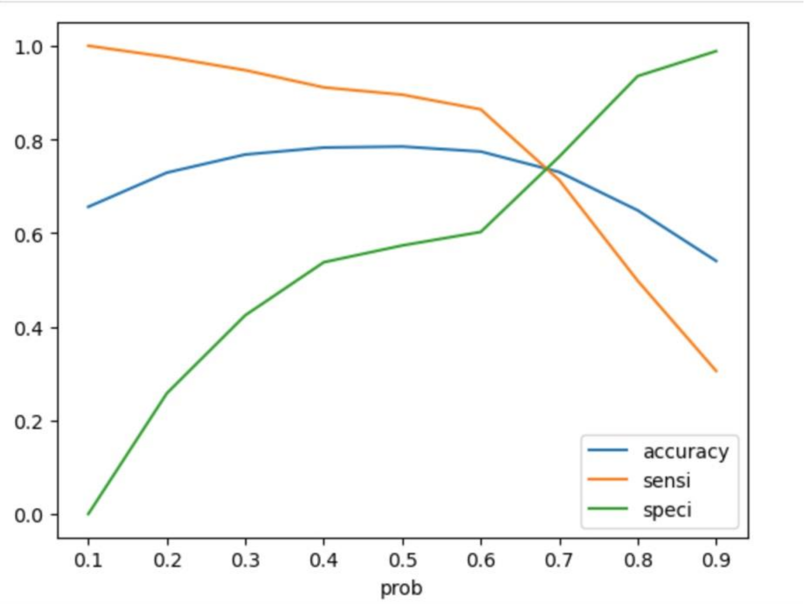
To prevent the adverse effects of multicollinearity, these columns were dropped from the analysis.



	USD Amount	15 Days from EOM	30 Days from EOM	30 Days from Inv Date	45 Days from EOM	45 Days from Inv Date	60 Days from EOM	60 Days from Inv Date	90 Days from EOM	90 Days from Inv Date	Immediate Payment	cluster_id	Invoice_Month
USD Amount	1	-0.09	-0.11	-0.05	-0.056	-0.078	0.16	0.093	-0.088	-0.025	0.15	0.16	-0.023
15 Days from EOM	-0.09	1	-0.11	-0.12	-0.062	-0.059	-0.12	-0.15	-0.062	-0.047	-0.11	-0.39	-0.0079
30 Days from EOM	-0.11	-0.11	1	-0.16	-0.078	-0.074	-0.15	-0.19	-0.078	-0.059	-0.13	0.055	-0.027
30 Days from Inv Date	-0.05	-0.12	-0.16	1	-0.088	-0.084	-0.17	-0.22	-0.089	-0.067	-0.15	0.14	0.013
45 Days from EOM	-0.056	-0.062	-0.078	-0.088	1	-0.042	-0.083	-0.11	-0.044	-0.033	-0.074	-0.2	-0.00053
45 Days from Inv Date	-0.078	-0.059	-0.074	-0.084	-0.042	1	-0.079	-0.1	-0.042	-0.031	-0.071	0.017	-0.028
60 Days from EOM	0.16	-0.12	-0.15	-0.17	-0.083	-0.079	1	-0.2	-0.083	-0.062	-0.14	0.00043	-0.023
60 Days from Inv Date	0.093	-0.15	-0.19	-0.22	-0.11	-0.1	-0.2	1	-0.11	-0.082	-0.18	0.065	0.017
90 Days from EOM	-0.088	-0.062	-0.078	-0.089	-0.044	-0.042	-0.083	-0.11	1	-0.033	-0.075	0.091	0.056
90 Days from Inv Date	-0.025	-0.047	-0.059	-0.067	-0.033	-0.031	-0.062	-0.082	-0.033	1	-0.056	0.048	0.044
Immediate Payment	0.15	-0.11	-0.13	-0.15	-0.074	-0.071	-0.14	-0.18	-0.075	-0.056	1	0.053	-0.0066
cluster_id	0.16	-0.39	0.055	0.14	-0.2	0.017	0.00043	0.065	0.091	0.048	0.053	1	-0.0027
Invoice_Month	-0.023	-0.0079	-0.027	0.013	-0.00053	-0.028	-0.023	0.017	0.056	0.044	-0.0066	-0.0027	1

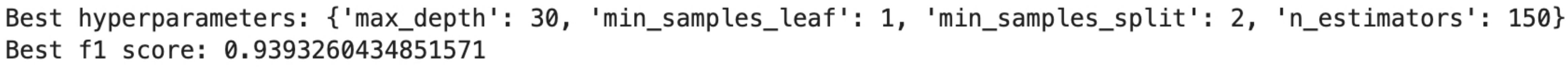
Performance Comparison: Logistic Regression vs. Random Forest





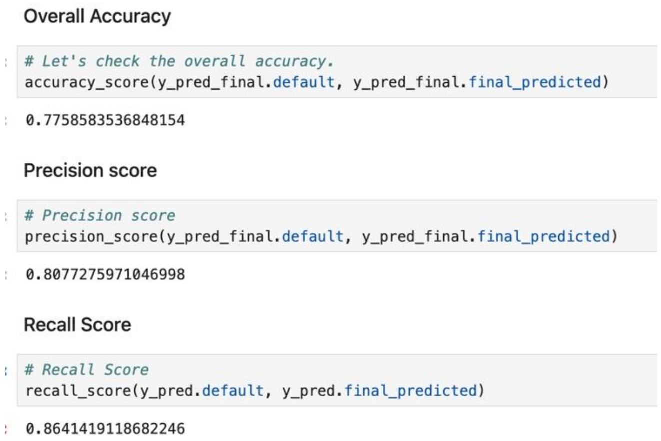
- After removing multicollinear and irrelevant variables, the Logistic Regression model retained only the features with acceptable p-values and VIF scores, ensuring a well-fitted model.
- The model achieved a strong ROC curve area of 0.83, indicating good predictive performance.
- The trade-off analysis between accuracy, sensitivity, and specificity identified an optimal probability cutoff of ~0.6, which was used to predict delayed payments in the received payments dataset.

- A Random Forest model was developed using the same parameters as the Logistic Regression model, along with hyperparameter tuning to optimize performance.
- The optimized Random Forest parameters were used to build the final model.
- The performance metrics of both models were compared, leading to the selection of the final model based on overall effectiveness.
- Below are the parameters we have used to built random forest model.



```
Best hyperparameters: {'max_depth': 30, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 150}
Best f1 score: 0.9393260434851571
```

Random Forest Outperforms Logistic Regression



Overall Accuracy

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

```
0.7758583536848154
```

Precision score

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

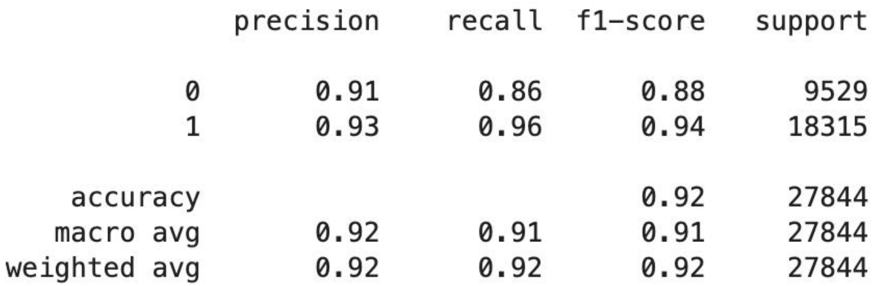
```
0.8077275971046998
```

Recall Score

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

```
0.8641419118682246
```

Logistic Regression Metrics - Test Set



```
              precision    recall  f1-score   support

           0       0.91      0.86      0.88      9529
           1       0.93      0.96      0.94     18315

    accuracy                           0.92     27844
   macro avg       0.92      0.91      0.91     27844
weighted avg       0.92      0.92      0.92     27844
```

Random Forest Metrics - Test Set

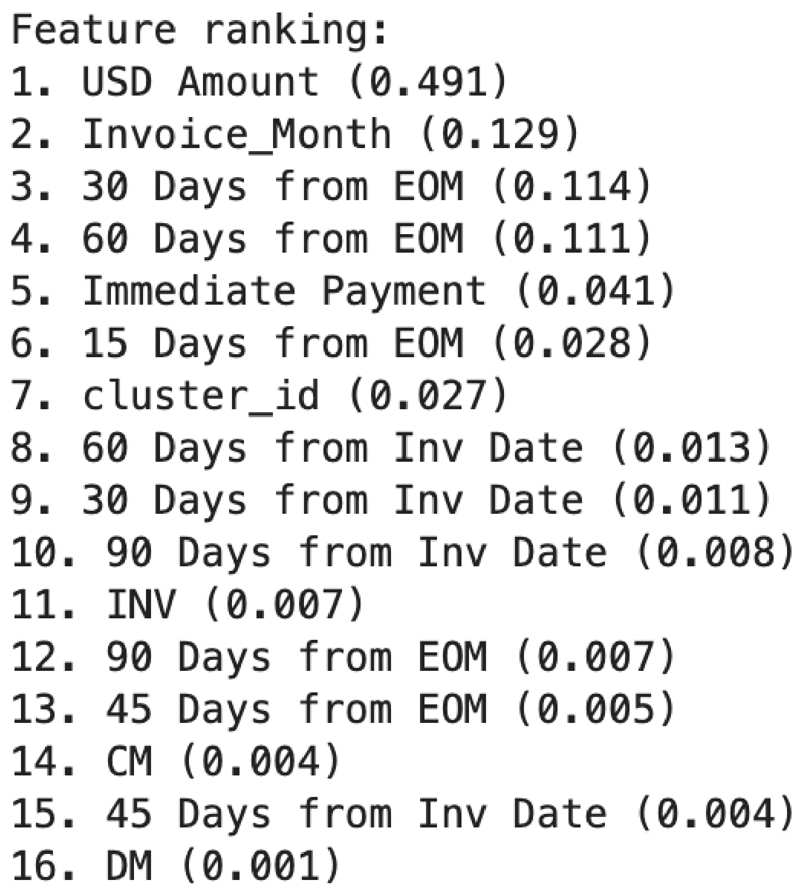
- Higher Precision & Recall: The Random Forest model significantly outperformed Logistic Regression in both precision and recall.
- Importance of Recall: Since predicting late payers accurately was a priority, recall played a crucial role in model selection.
- Better for Categorical Data: Given the dataset's high categorical variable count, Random Forest proved to be a more suitable choice.
- Final Model Selection: Based on superior performance, Random Forest was chosen for final predictions.

Random Forest Feature Rankings

Random Forest Feature Ranking identified the top 5 predictors of payment delays:

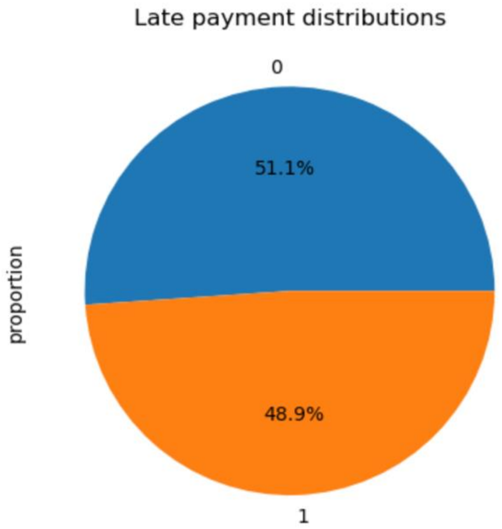
- USD Amount
- Invoice Month
- 60 Days from EOM (Payment Term)
- 30 Days from EOM (Payment Term)
- Cluster ID (derived from average and standard deviation of payment days)

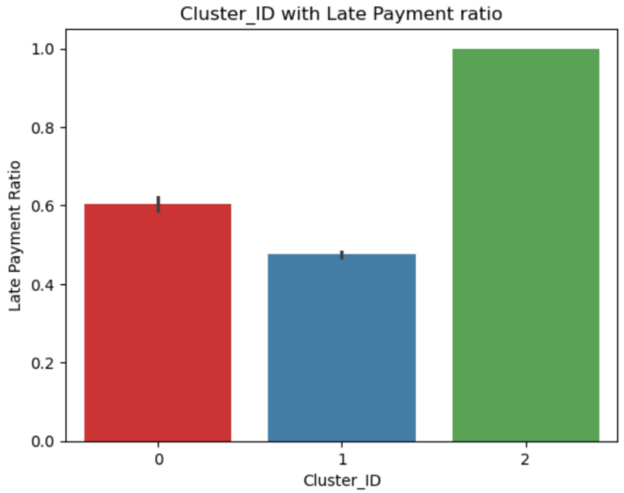
Cluster-based Segmentation was applied to open-invoice data based on customer name, enabling targeted payment delay predictions.



```
Feature ranking:
1. USD Amount (0.491)
2. Invoice_Month (0.129)
3. 30 Days from EOM (0.114)
4. 60 Days from EOM (0.111)
5. Immediate Payment (0.041)
6. 15 Days from EOM (0.028)
7. cluster_id (0.027)
8. 60 Days from Inv Date (0.013)
9. 30 Days from Inv Date (0.011)
10. 90 Days from Inv Date (0.008)
11. INV (0.007)
12. 90 Days from EOM (0.007)
13. 45 Days from EOM (0.005)
14. CM (0.004)
15. 45 Days from Inv Date (0.004)
16. DM (0.001)
```

Predicted Payment Delays & Business Impact





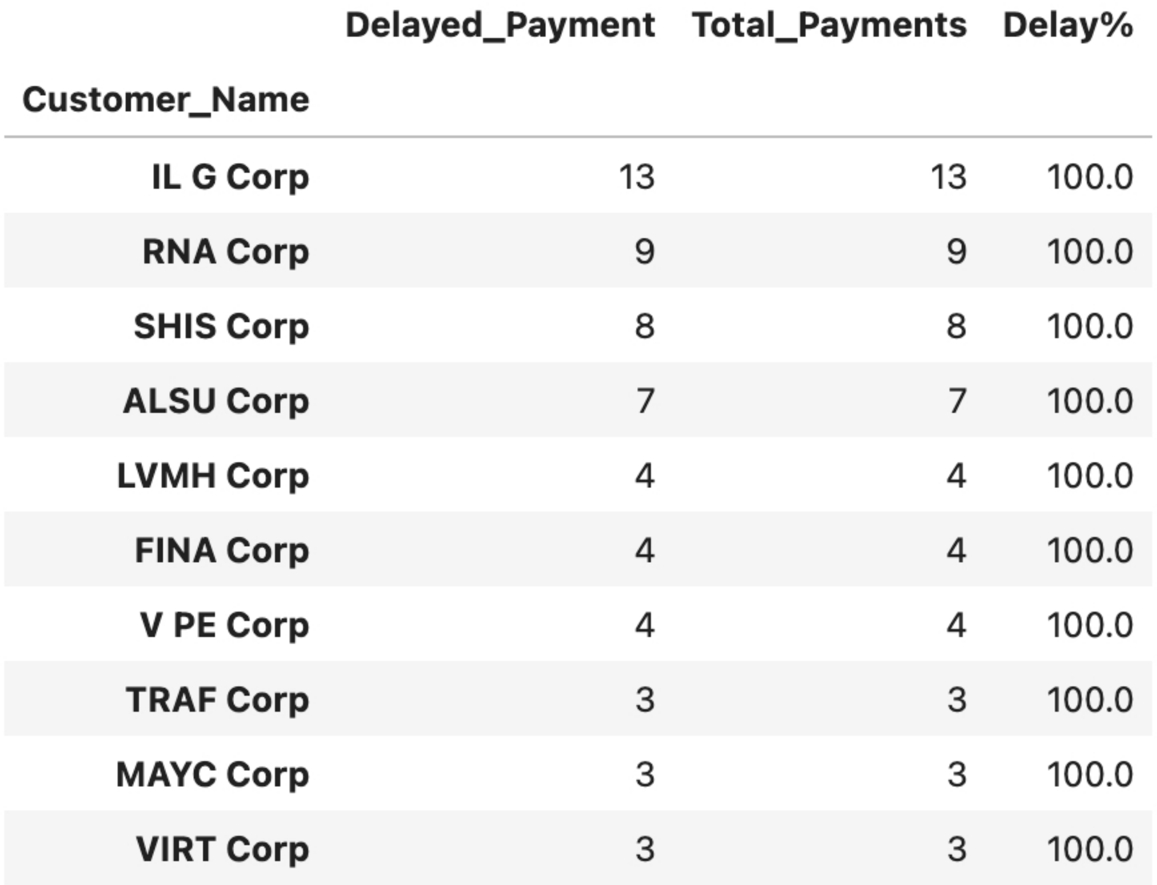
- The final model predicts that 50.2% of transactions are likely to experience payment delays, posing a significant risk to business operations.

- Customers with a history of prolonged payment delays are expected to have an alarmingly high delay rate (~100%), aligning with historical payment trends.

- Early and medium-duration payers show lower delay probabilities, reinforcing past observations.

High-Risk Customers with Maximum Delay Probability

- *Predictions identify companies with the highest likelihood of payment delays.*
- *These companies have the maximum number of overdue and total payments.*
- *They represent a high-risk segment for default, requiring closer monitoring.*



	Delayed_Payment	Total_Payments	Delay%
Customer_Name			
IL G Corp	13	13	100.0
RNA Corp	9	9	100.0
SHIS Corp	8	8	100.0
ALSU Corp	7	7	100.0
LVMH Corp	4	4	100.0
FINA Corp	4	4	100.0
V PE Corp	4	4	100.0
TRAF Corp	3	3	100.0
MAYC Corp	3	3	100.0
VIRT Corp	3	3	100.0

Key Recommendations from Clustering Analysis

❖ **Stricter Payment Collection for Credit Notes**

- Credit Note payments have the **highest delay rates** *compared to* Debit Note or Invoice categories. .
- Implement **stricter collection policies** to reduce risks.

❖ **Enhance Policies for Goods-Related Invoices**

- Goods-related invoices show **higher delay rates** than non-goods.
- Introduce **tighter payment terms & proactive follow-ups**.

❖ **Manage Lower-Value Transactions More Closely**

- **Majority of transactions** are lower in value and **prone to delays**.
- Consider a **tiered penalty system** (higher penalty for lower billing amounts, to be applied cautiously).

❖ **Focus on High-Risk Customer Segments (Cluster 1)**

- Customers in **Cluster 1 (Prolonged Payers)** show **significantly higher delay rates**.
- Strengthen **payment tracking & engagement** to minimize risks.

❖ **Prioritize High-Risk Companies for Action**

- Companies with the **highest probability of delayed payments** need targeted efforts.
- Implement **focused intervention strategies** to improve payment timeliness.