Schuster

E-COMMERCE & RETAIL B2B

Palak Paliwal
Prabin Pal
Muskan Mohanty

Domain
Oriented Case
Study

Identifying the Problem and Establishing Targets

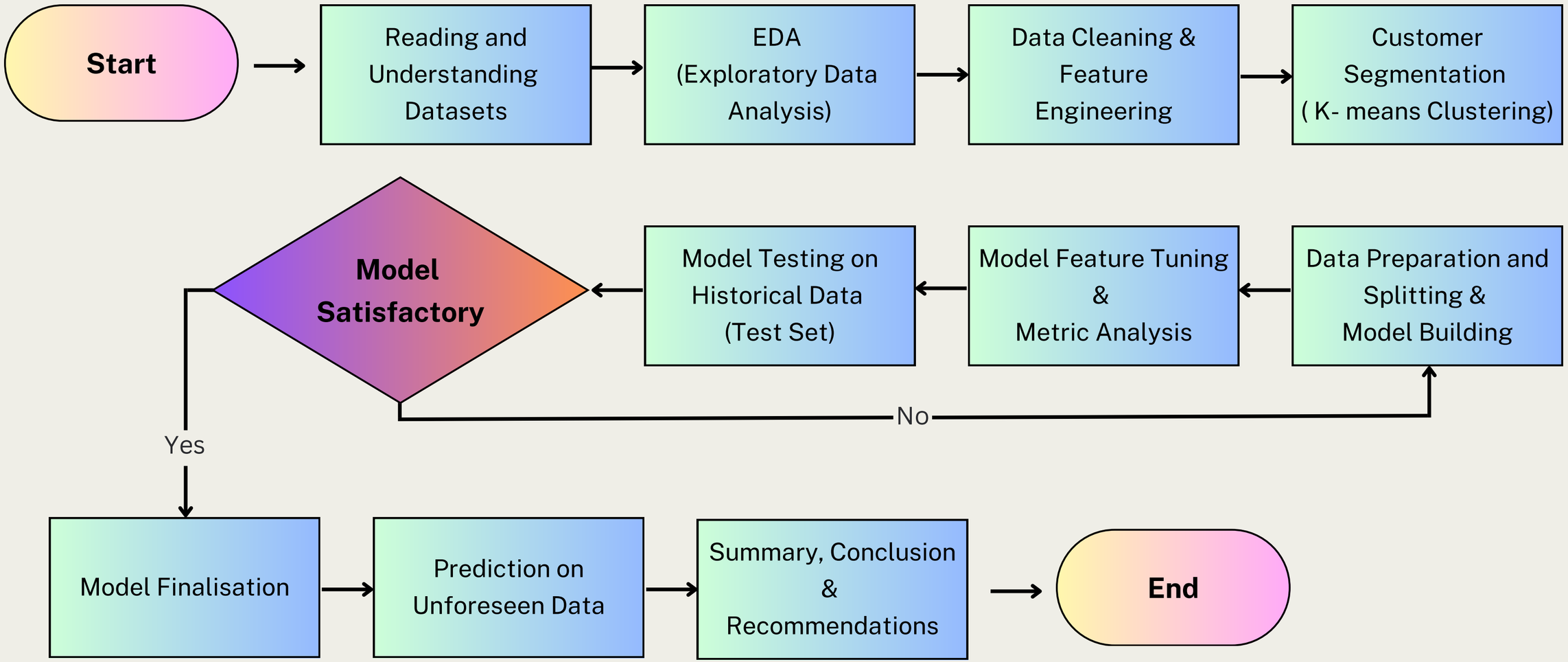
Problem Identification

- Schuster, a sports retail company engaged in B2B transactions, often extends credit to its vendors, who may not always adhere to the agreed-upon payment deadlines.
- When vendors delay their payments, it causes financial disruptions, leading to cash flow issues that hinder efficient business operations.
- Moreover, company staff members are burdened with the task of pursuing outstanding payments for extended periods, diverting resources from more productive activities and resulting in unnecessary expenditure.

Targets

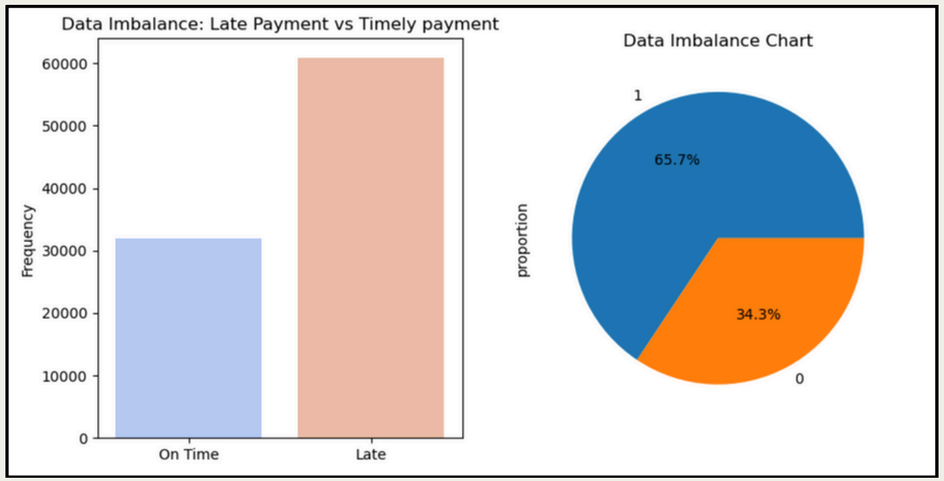
- Conducting customer segmentation to analyze payment behaviors.
- Leveraging historical transaction data, the company aims to forecast delayed payments for transactions that have not yet surpassed their due dates.
- The goal is to use these predictions to optimize resource allocation, speed up credit recovery, and minimize time spent on low-value tasks.

Strategic Approach

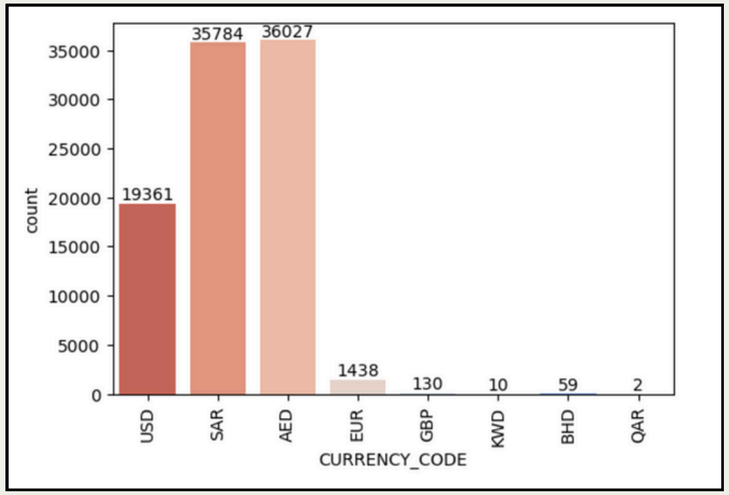


Univariate Analysis

Class Imbalance and Transaction Insights



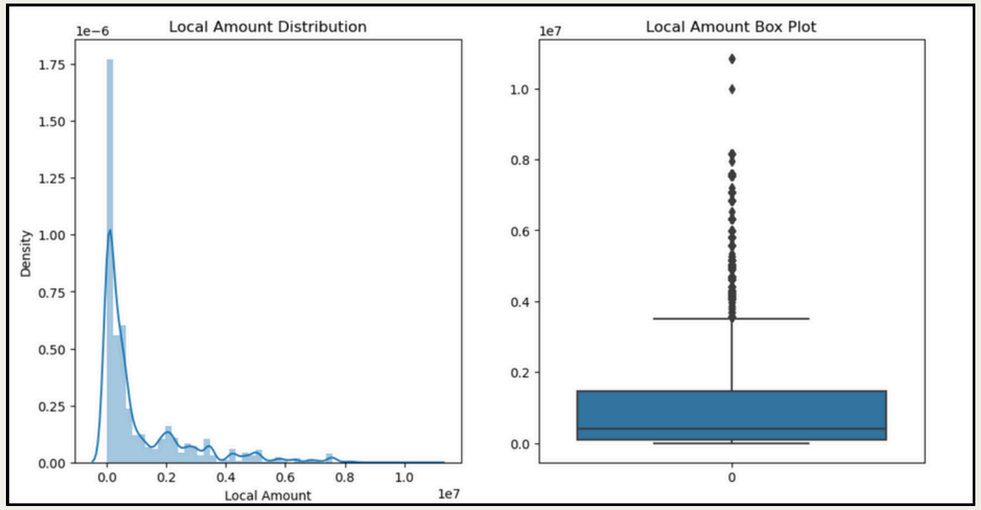
- The class imbalance is approximately 65.7% in favor of payment delayers, which is considered a reasonable imbalance and does not require any specific treatment for class imbalance.



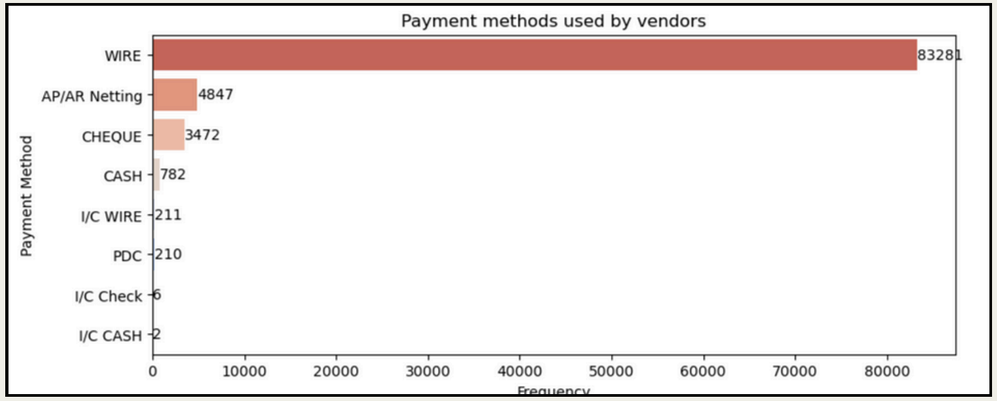
- The company primarily conducts transactions in three currencies: AED, SAR, and USD, with AED being the most frequently used. This suggests that the company has a significant volume of business with the Middle East region.

Univariate Analysis

Class Imbalance and Transaction Insights



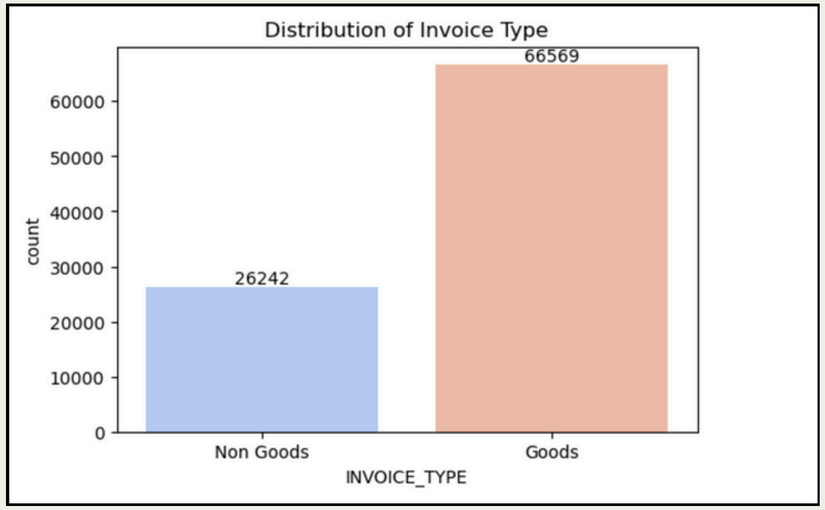
- Transaction values range from $1 to $3 million, with the most common values clustered below approximately $1.75 million.



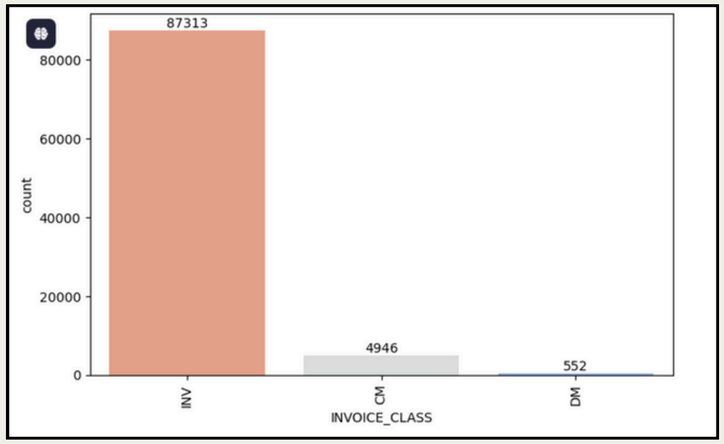
- The wire payment method is the most commonly used payment method for transactions received by the company, followed by netting, cheque, and cash payments.

Univariate Analysis

Class Imbalance and Transaction Insights



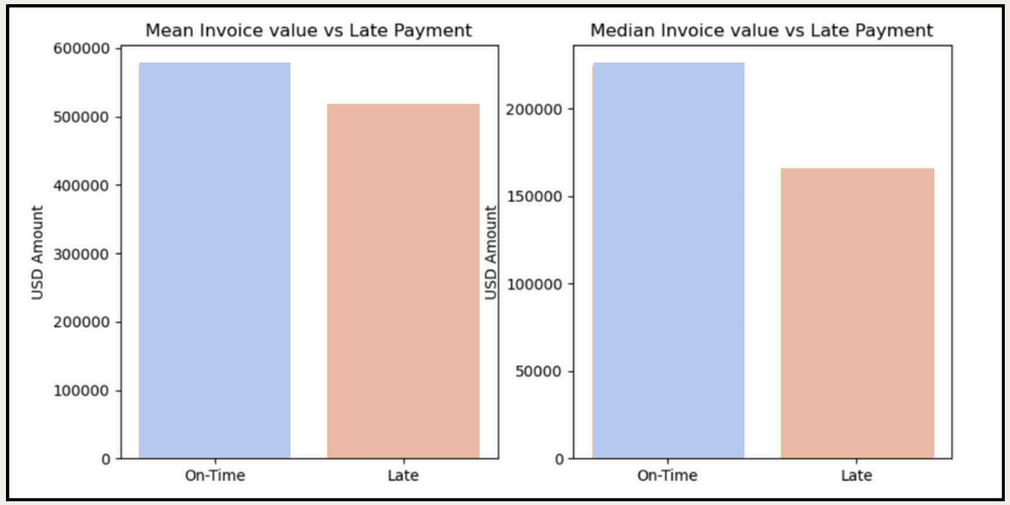
- Goods-type invoices account for the largest proportion of the invoices generated by the company.



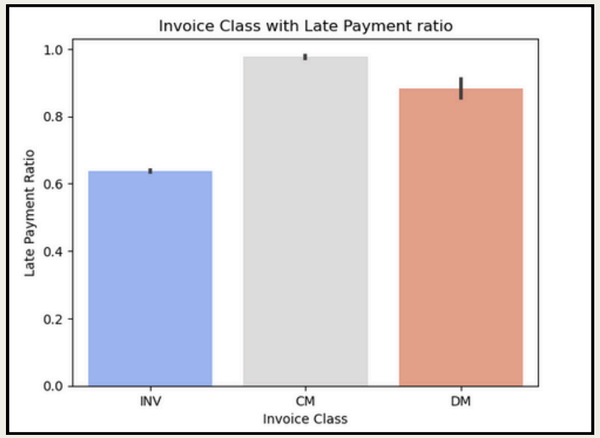
- The primary invoice class is ‘Invoice’, which holds the majority of the share, while other classes have minimal representation.

Bivariate Analysis

Identifying the Characteristics of Payment Types Associated with Defaulters



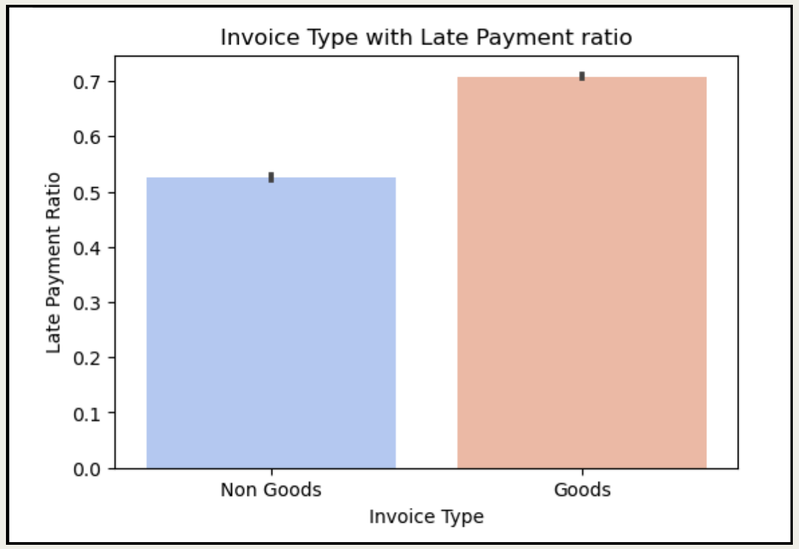
- The average and median payment amounts are higher for customers who make timely payments compared to those who delay payments. This indicates that transactions with higher values tend to be less likely to face delays, while lower-value transactions exhibit a higher risk of delayed payments.



- The late payment ratio is highest for Credit Note transactions, followed by Debit Note and Invoice types. This suggests that Credit Note and Debit Note invoices carry a higher risk of delayed payments compared to standard Invoice transactions.

Bivariate Analysis

Identifying the Characteristics of Payment Types Associated with Defaulters



- Goods-type invoices exhibit a higher late payment ratio compared to non-goods invoices, indicating an increased likelihood of payment delays for goods-related transactions.

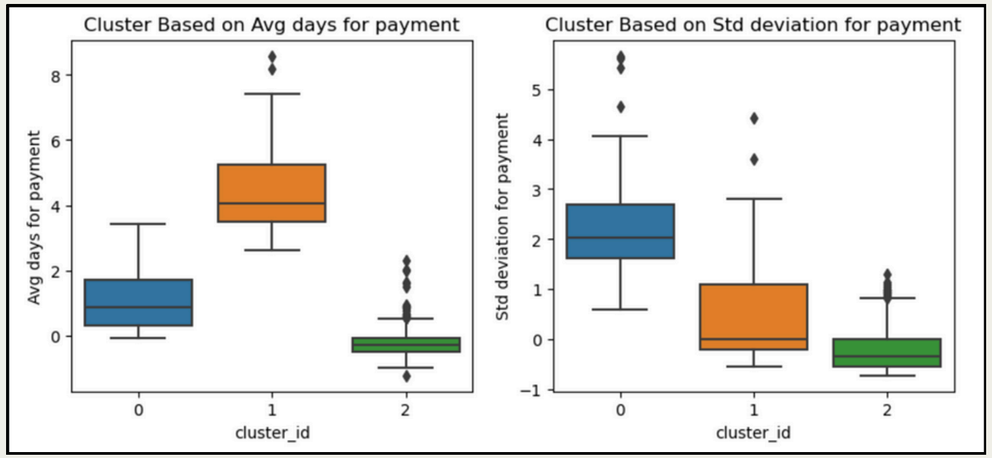
Customer Segmentation using K-means Clustering

One of the objectives was to categorize customers to understand payment behaviors which was achieved by K-means clustering using average and standard deviation of number of days it took for the vendor to make payment.

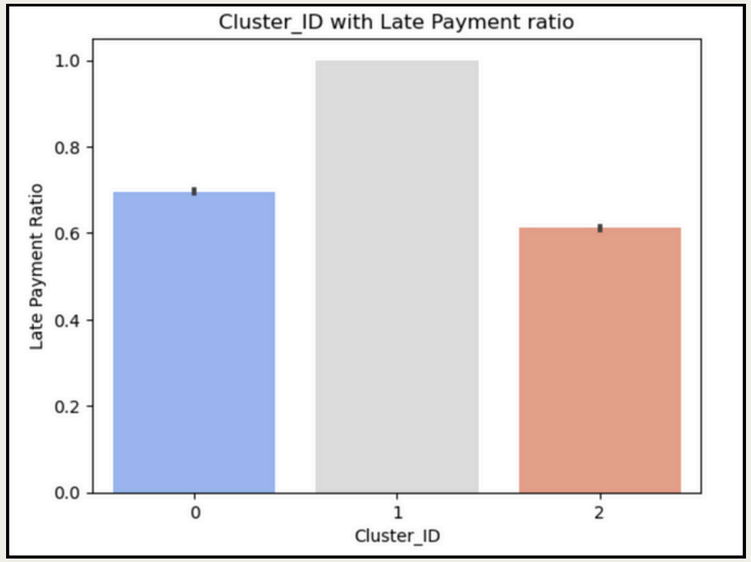
```
For n_clusters=2, the silhouette score is 0.7557759850933141
For n_clusters=3, the silhouette score is 0.735036462331666
For n_clusters=4, the silhouette score is 0.6182691953064194
For n_clusters=5, the silhouette score is 0.6209288452882942
For n_clusters=6, the silhouette score is 0.4025255389461882 5
For n_clusters=7, the silhouette score is 0.4069490441271981
For n_clusters=8, the silhouette score is 0.4151884768372497
```

- The decision to set the number of clusters to 3 was based on the significant decrease in the silhouette score observed when the number of clusters exceeded 3.
- This suggests that 3 clusters provided the best balance between compactness and separation in the dataset.
- Increasing the number of clusters beyond this point led to less distinct and more overlapping groupings, as indicated by the reduced silhouette score.

Customer Segmentation using K-means Clustering

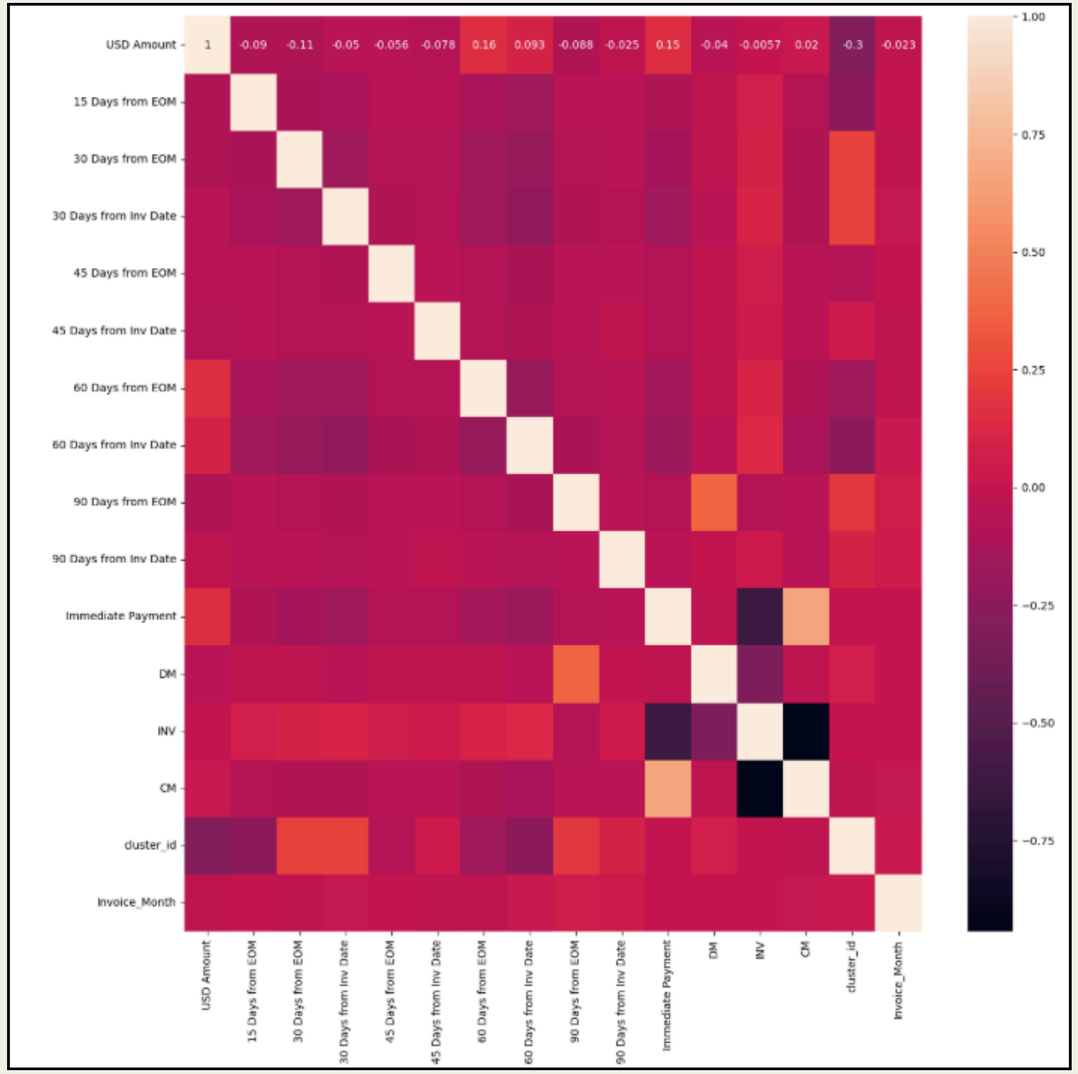


- Category 2 represents the early payers, characterized by the shortest average payment durations.
- Category 1 corresponds to the prolonged payers, who take the longest time to settle their payments.
- Category 0, which lies between the two extremes, is categorized as medium-duration payers, indicating an intermediate payment behavior.



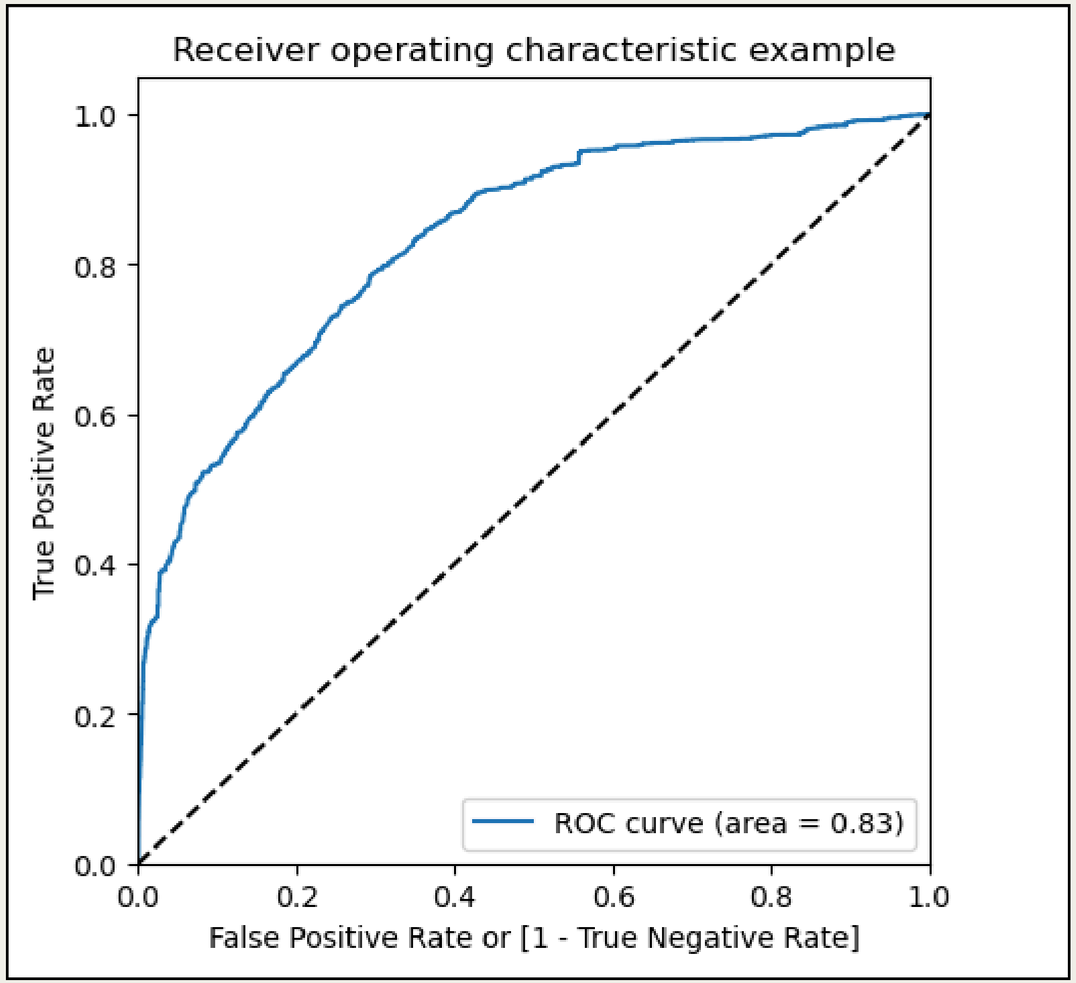
- It was also noted that customers in the prolonged payment category tend to exhibit significantly higher rates of delayed payments compared to both early and medium-duration payers.
- This suggests that the longer the payment period, the greater the likelihood of payment delays.

Model Building



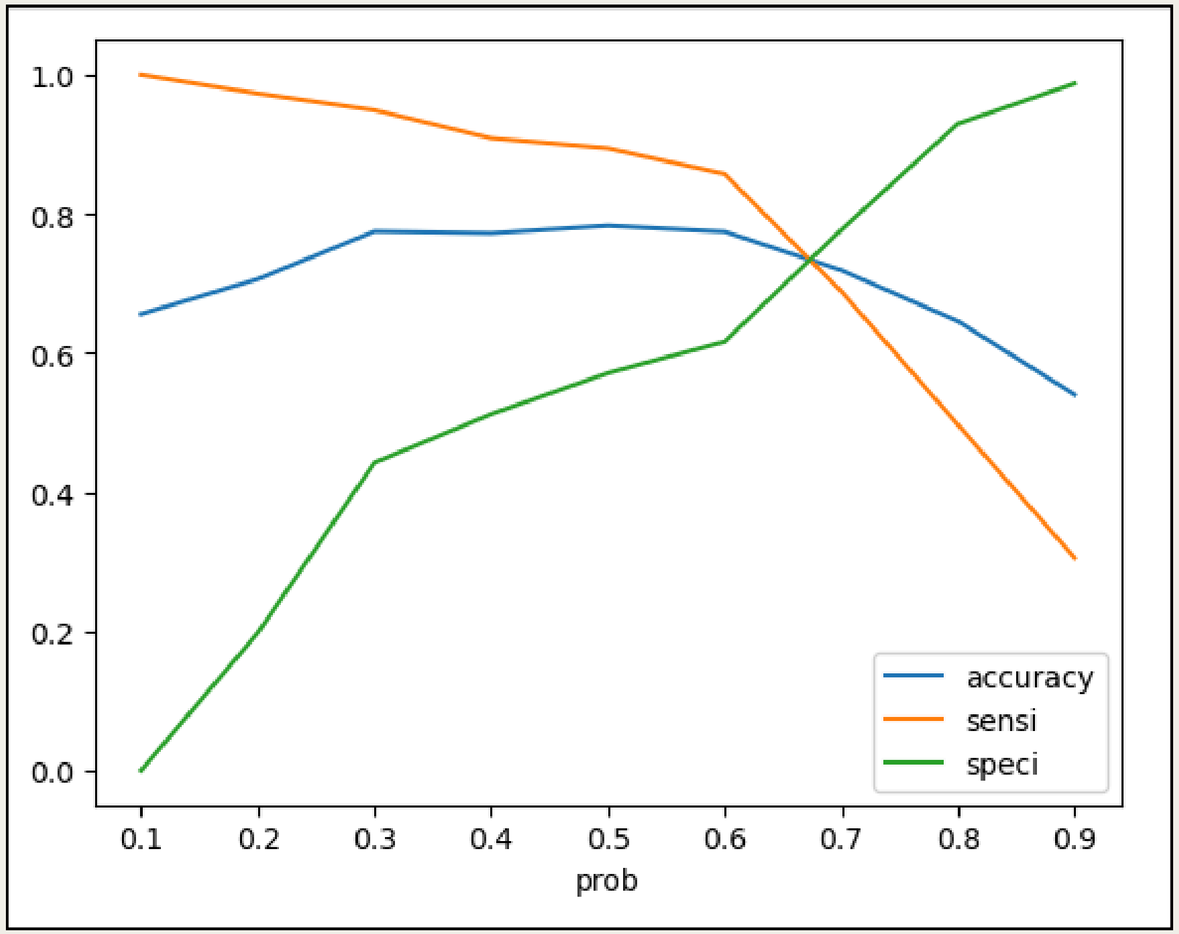
- Columns like CM & INV, INV & Immediate Payment, and DM & 90 Days from EOM were found to have high multicollinearity.
- To mitigate the effects of this multicollinearity, these columns have been removed from the dataset.
- This helps in improving the model's stability and ensures that no redundant or highly correlated features distort the predictive power of the machine learning model.

Model Building



- After addressing multicollinearity and removing unnecessary variables, the logistic regression model retained only the significant features, which showed acceptable p-values and Variance Inflation Factor (VIF) values.
- As a result, no further feature elimination was necessary. The final model demonstrated a solid performance with a good ROC curve area of 0.83, indicating its effectiveness in predicting the likelihood of late payments based on the selected features.

Model Building



- The trade-off plot between accuracy, sensitivity, and specificity helped determine the optimal probability cutoff of approximately 0.6.
- This threshold was used to classify transactions, enabling the prediction of which transactions in the received payments dataset were likely to result in delayed payments.
- This cutoff point balances the model's ability to correctly identify delayed payments (sensitivity) and the accuracy of predictions, ensuring the model is effective for practical use in payment predictions.

"Model Comparison: Logistic Regression vs. Random Forests"

A random forest model was developed using similar parameters to the logistic regression model, along with hyperparameter tuning, leading to the following optimized parameters.

```
Best hyperparameters: {'max_depth': 30, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 150}
Best f1 score: 0.9397563605282018
```

Using the parameters mentioned above, a random forest model was constructed and its performance metrics were compared to those of the logistic regression model. Based on this comparison, the final model was selected.

"Random Forest Outperforms Logistic Regression"

Logistic Regression - Test Set

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

[39]: 0.775211894842695

[40]: #precision score
      precision_score(y_pred_final.default, y_pred_final.final_predicted)

[40]: 0.811170761924427

[41]: # Recall Score
      recall_score(y_pred.default, y_pred.final_predicted)

[41]: 0.8573138110657469
```

Random Forest Metrics - Test Set

	precision	recall	f1-score	support
0	0.96	0.91	0.94	22349
1	0.96	0.98	0.97	42618
accuracy			0.96	64967
macro avg	0.96	0.95	0.95	64967
weighted avg	0.96	0.96	0.96	64967

- It was observed that the Random Forest model outperformed the Logistic Regression model in terms of overall precision and recall scores.
- Specifically, recall scores were prioritized since accurately identifying late payers was crucial for targeting follow-up actions. Given that the data contained many categorical variables, Random Forest proved more suitable for the task compared to Logistic Regression.
- Therefore, the Random Forest model was selected as the final model to proceed with predictions.

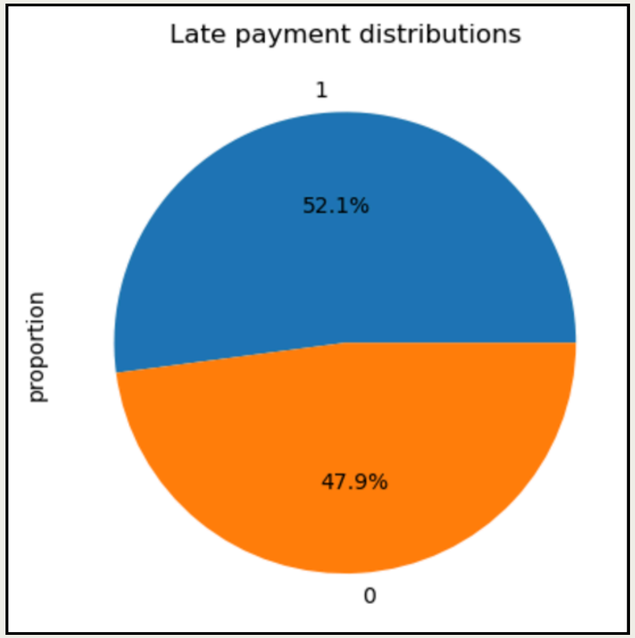
Ratings of Random Forest Features

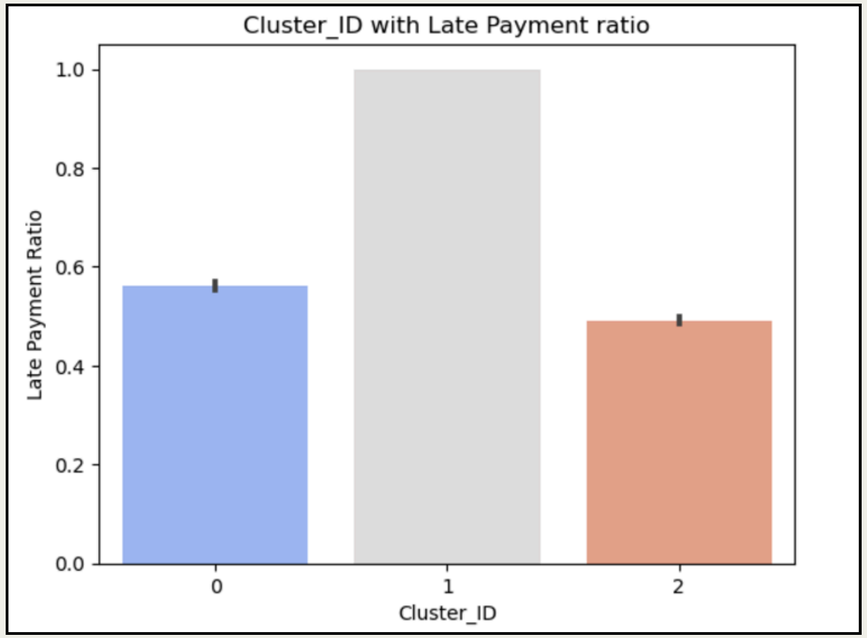
```
Feature ranking:
1. USD Amount (0.464)
2. Invoice_Month (0.129)
3. 60 Days from EOM (0.112)
4. 30 Days from EOM (0.107)
5. cluster_id (0.054)
6. Immediate Payment (0.043)
7. 15 Days from EOM (0.027)
8. 30 Days from Inv Date (0.014)
9. 60 Days from Inv Date (0.013)
10. 90 Days from Inv Date (0.009)
11. INV (0.008)
12. 90 Days from EOM (0.007)
13. 45 Days from EOM (0.006)
14. 45 Days from Inv Date (0.004)
15. CM (0.004)
16. DM (0.001)
```

- The Random Forest model was used to determine feature rankings, with the top five predictors of payment delay identified as:
  1. USD Amount
  2. Invoice Month
  3. 60 Days from EOM
  4. 30 Days from EOM
  5. Cluster-ID

- The customer segments based on Cluster-ID were applied to the open-invoice data, allowing predictions for delayed payments.

Predictions





- 52% of payments in the Openinvoice data are predicted to be delayed, with prolonged payment periods exhibiting significantly higher delay rates.
- The final model predicts that approximately 52.1% of transactions may experience payment delays, which could significantly disrupt business operations.

- The customer segment with a history of prolonged payment days is projected to have the highest delay rates (~100%), consistent with findings from historical data regarding payment behaviors. This trend aligns with previous observations of late payment tendencies among this segment.

Predictions

Customers with the highest likelihood of payment delays.

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

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

Recommendations

- Credit Note Payments exhibit the highest delay rates compared to Debit Note or Invoice types. This suggests a need for stricter policies for payment collection on credit notes to minimize delays.
- Goods-related invoices demonstrate significantly higher payment delays compared to non-goods types. Stricter payment collection measures should be implemented for these invoice categories.
- Since lower-value payments form a substantial portion of transactions and are more prone to delays, it is advisable to prioritize them.
- The company could implement a tiered penalty system based on invoice value, where smaller bills incur a higher percentage penalty for late payments. This should, however, be used as a last resort.
- Customer segmentation revealed three clusters categorized as early, medium, and prolonged payment durations. Cluster 1 (prolonged payments) showed the highest delay rates, requiring focused attention to improve compliance.
- Vendors with the highest probabilities of delay and significant delayed payment counts should be given top priority.
- These high-risk accounts need focused efforts to ensure timely payments.
- These insights can guide resource allocation and policy adjustments to enhance payment compliance and reduce delays.

Thank You!