E-CommerceandRetail B2B Case Study

Varshini N
Upasana Mishra

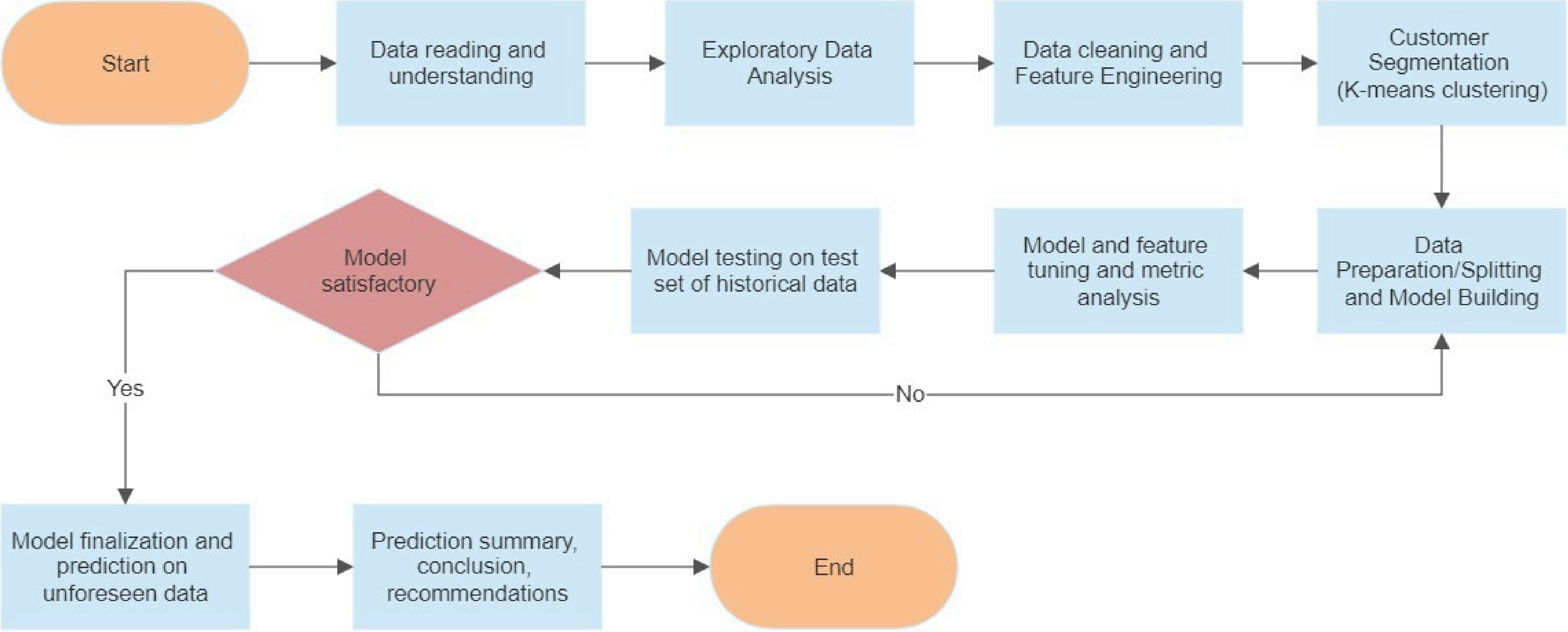
Problem Statement

- Schuster, a sports retail company specializing in B2B transactions, often extends credit to vendors, who may or may not adhere to the agreed payment deadlines.
- Payment delays from vendors lead to financial setbacks and losses, disrupting smooth business operations.
- Additionally, company employees spend extended periods chasing overdue payments, leading to unproductive tasks and inefficient use of resources.

Business Objective

- Segmenting customers to analyze their payment behavior.
- The company needs to predict delayed payments for an upcoming dataset of transactions with due dates that have not yet passed, based on historical data.
- The company seeks predictions to enable more efficient resource allocation, faster credit recovery, and streamlined operations.

Problem 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 and transaction insights (univariate)

Fig. 1

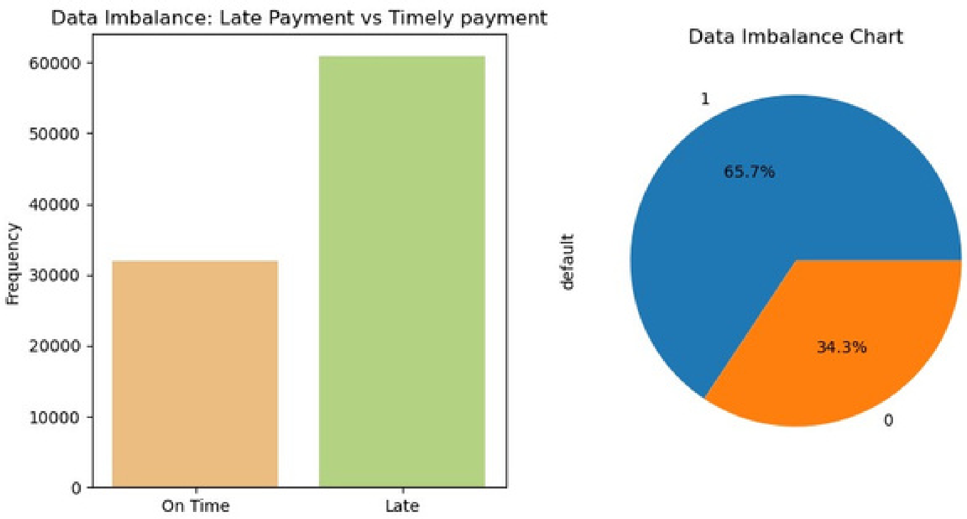
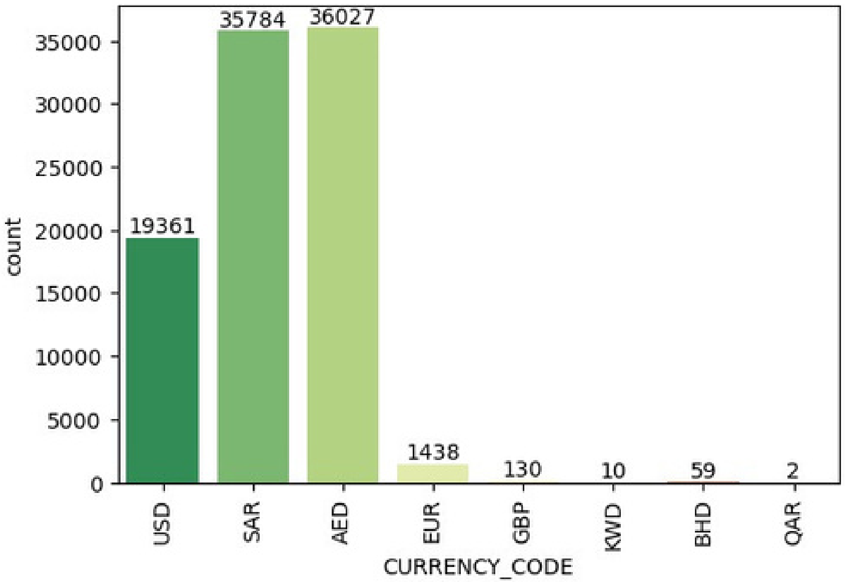


Fig. 2



From Figures 1 and 2:

- There is a class imbalance of 65.7% toward payment delayers, which is acceptable and does not require adjustment.
- The top three currencies in which the company conducts transactions are AED, SAR, and USD, with AED being the most frequently used, indicating a higher volume of transactions in the Middle East.

Class imbalance and transaction insights (univariate)

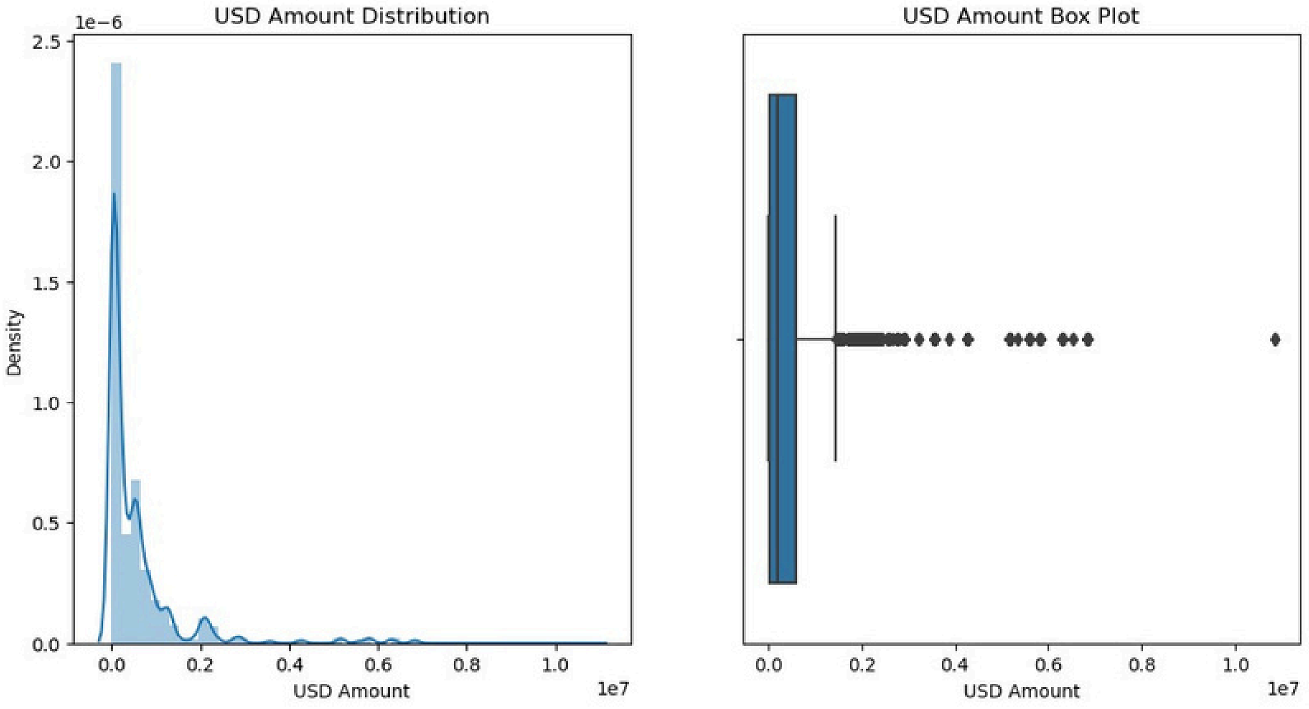


Fig. 3

From Fig. 3, we observe,

- Transaction values range between $1 and $3 million, with the majority of transactions occurring below approximately $1.75 million.

Class imbalance and transaction insights (univariate)

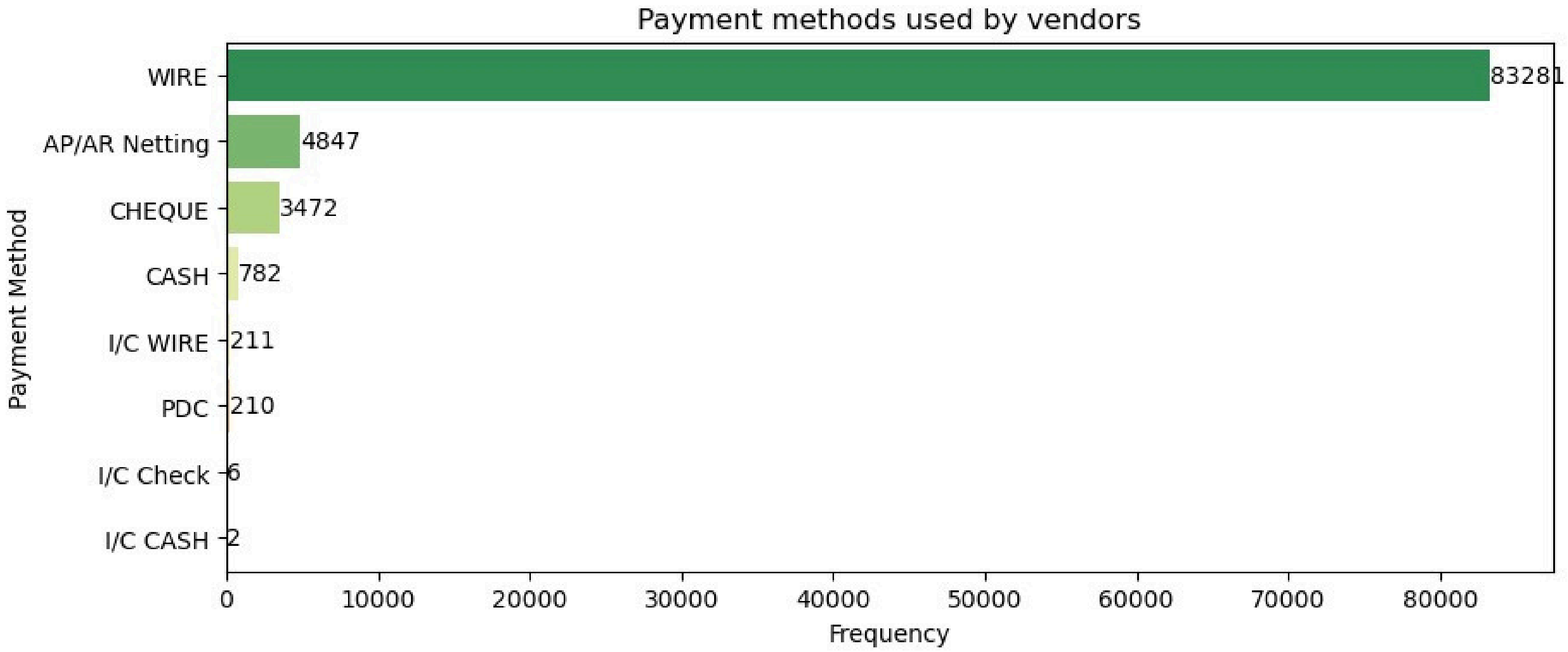


Fig. 1

From Figure 1, we observe that wire transfers are the most common payment method received by the company, followed by netting, cheques, and cash.

Class imbalance and transaction insights (univariate)

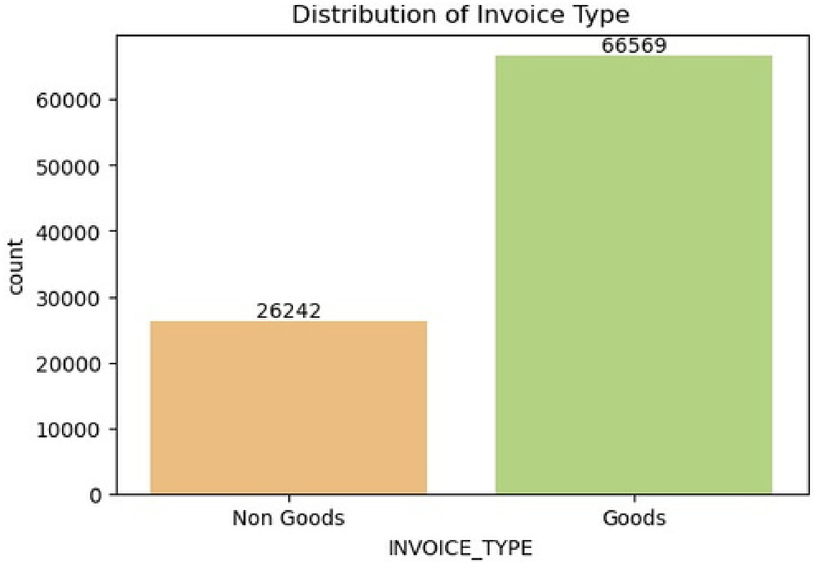


Fig. 2

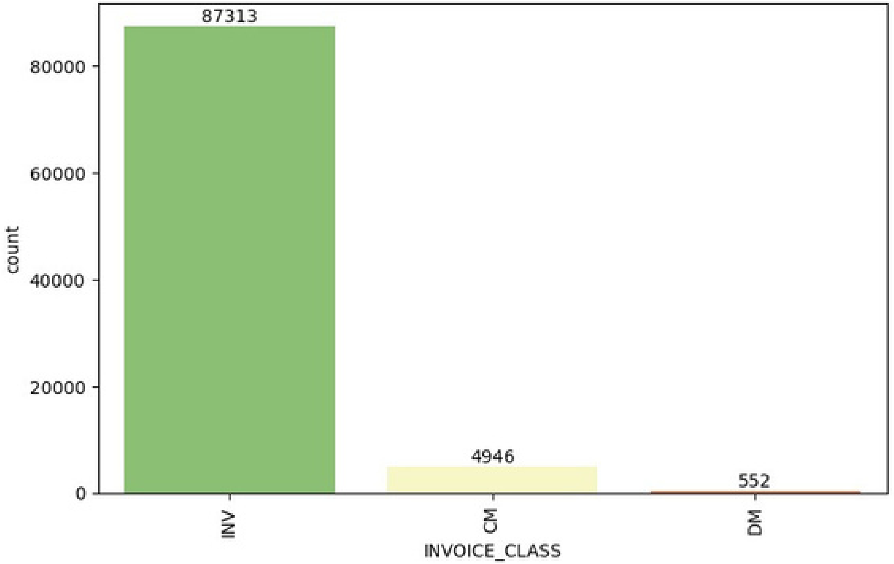


Fig. 3

From Figures 2 and 3:

- The majority of invoices generated are classified as "Invoice," with other types representing a much smaller share.
- Goods-type invoices make up the largest portion of invoice categories.

Identifying characteristics of defaulterpayment types (Bivariate)

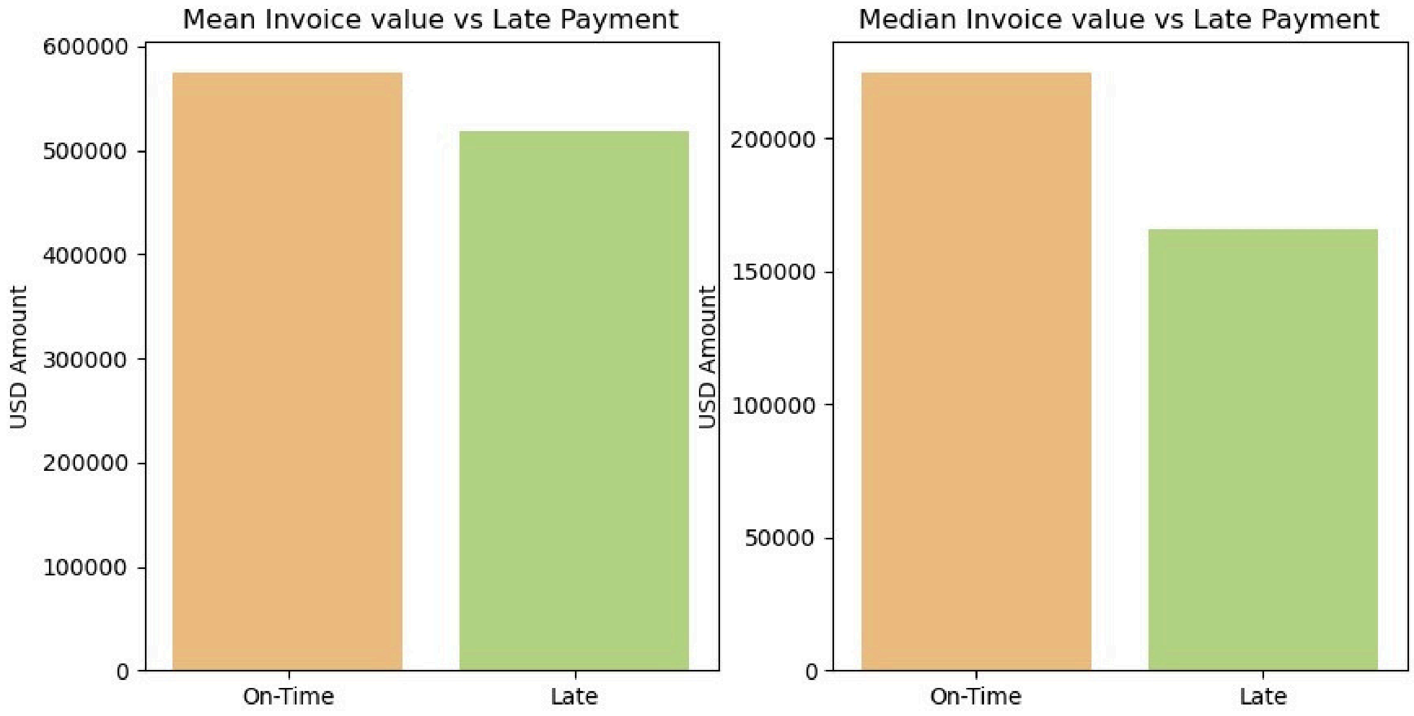


Fig. 1

From Figure 1, the mean and median payment amounts are higher for on-time payers compared to late payers, suggesting that higher-value transactions carry a lower risk of delay than lower-value ones.

Identifying characteristics of defaulter payment types (Bivariate)

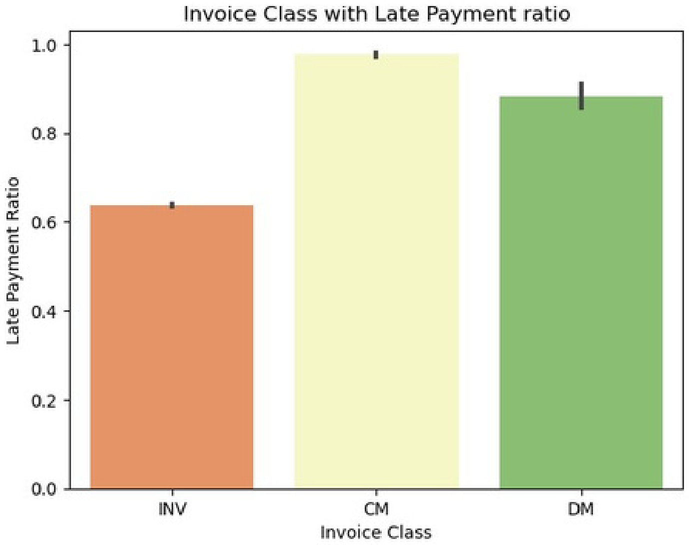


Fig. 2

From Figure 2, late payment...
The ratio of Credit Note transactions is the highest, followed by Debit Notes and Invoices, indicating a greater risk of payment delays in Credit and Debit Note invoice categories.

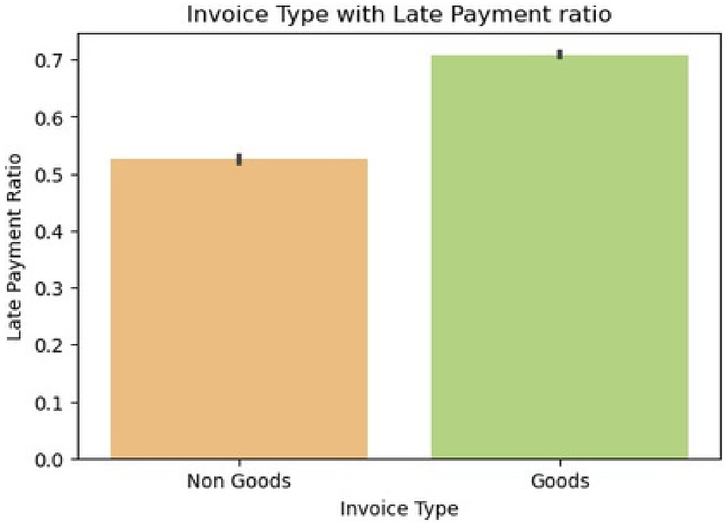
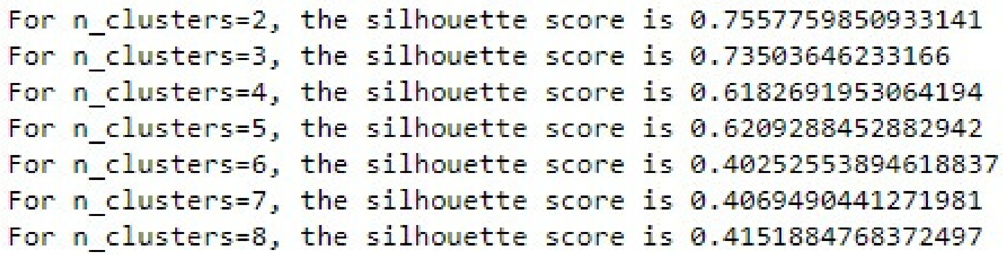


Fig. 3

From Figure 3,Goods-type invoices have a higher late payment ratio compared to non-Goods invoices, indicating an increased likelihood of payment delays.

Customer segmentation using K-means Clustering

- An objective was to categorize customers to analyze payment behaviors, which was accomplished through K-means clustering using the average and standard deviation of days taken by each vendor to make payments.
- The number of clusters was set to 3, as adding more clusters beyond this point led to a significant drop in the silhouette score.



```
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
```

Customer segmentation using K-means Clustering

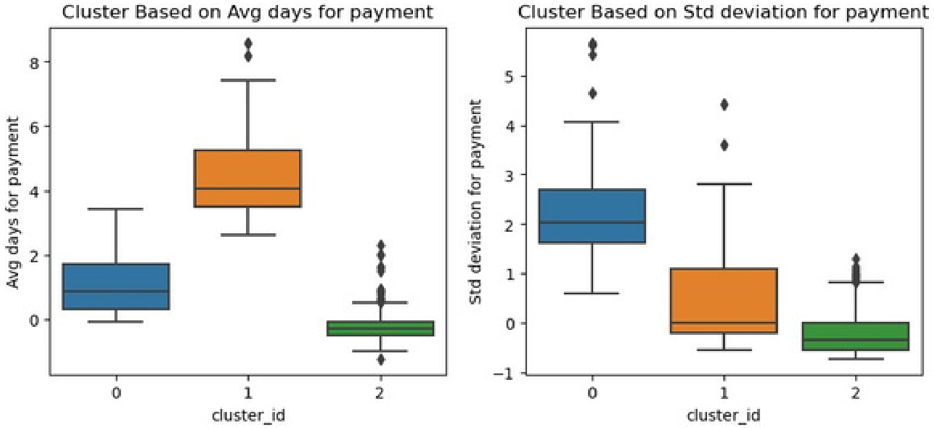


Fig. 1

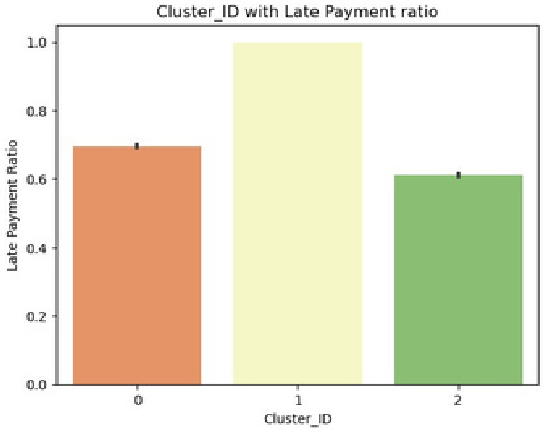
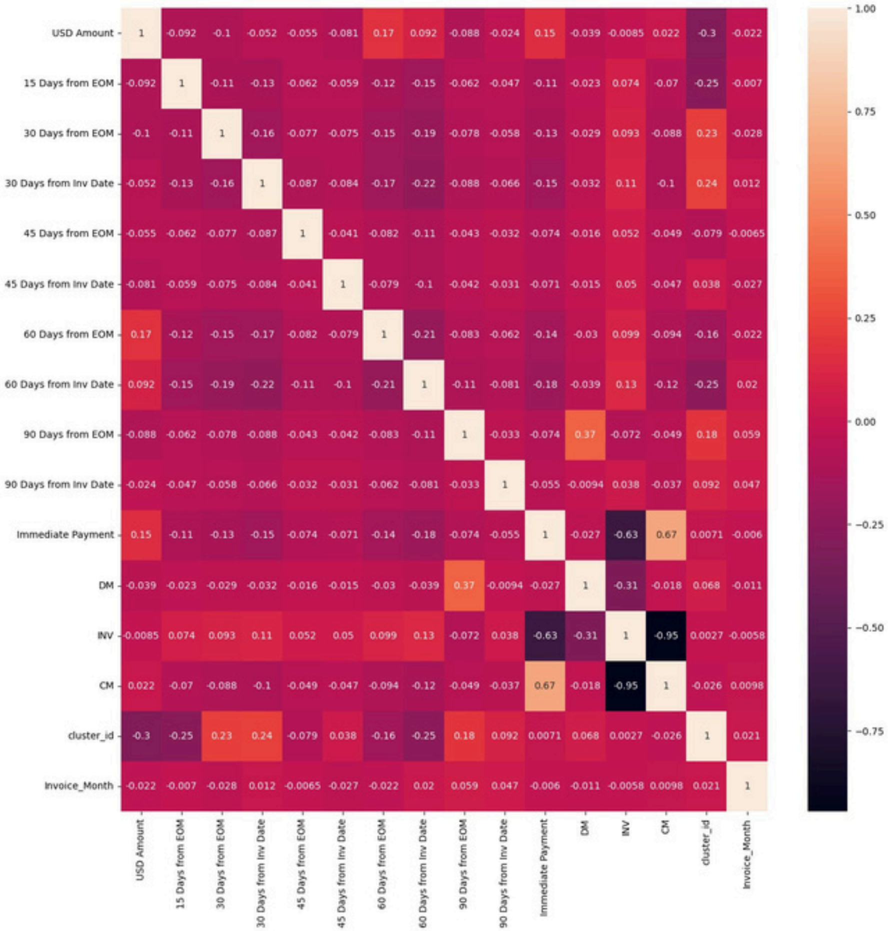


Fig. 2

- Category 2 consists of early payers with the shortest average payment time, while Category 1 includes delayed payers with the longest average payment time. Category 0 falls between the two and is labeled as medium-duration payers.
- It was also noted that prolonged payers historically exhibit significantly higher rates of payment delays compared to early or medium-duration payment transactions.(see Fig. 2).

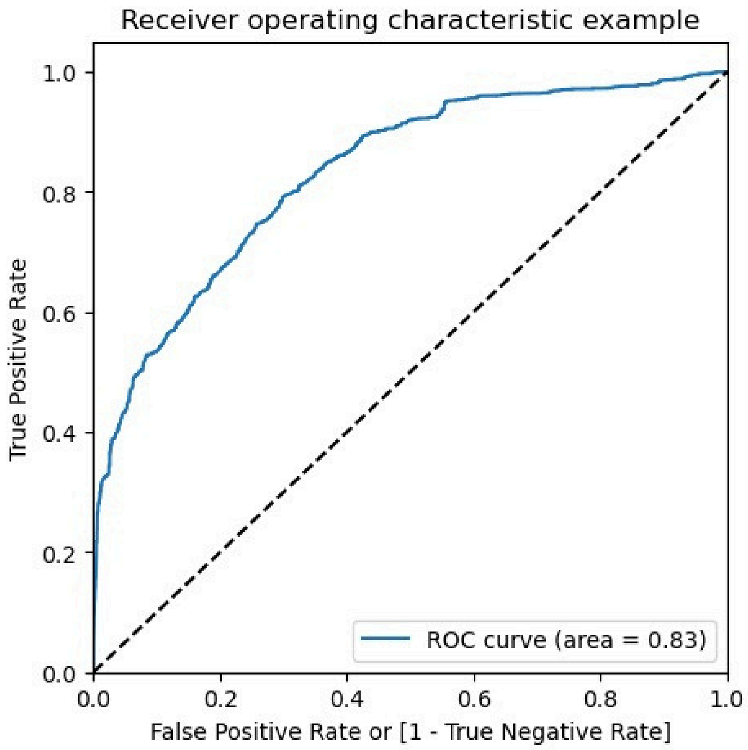
Model Building



	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	DM	INV	CM	cluster_id	Invoice_Month
USD Amount	1	-0.092	-0.1	-0.052	-0.055	-0.081	0.17	0.092	-0.088	-0.024	0.15	-0.039	-0.0085	0.022	-0.3	-0.022
15 Days from EOM	-0.092	1	-0.11	-0.13	-0.062	-0.059	-0.12	-0.15	-0.062	-0.047	-0.11	-0.023	0.074	-0.07	-0.25	-0.007
30 Days from EOM	-0.1	-0.11	1	-0.16	-0.077	-0.075	-0.15	-0.19	-0.078	-0.058	-0.13	-0.029	0.093	-0.088	0.23	-0.028
30 Days from Inv Date	-0.052	-0.13	-0.16	1	-0.087	-0.084	-0.17	-0.22	-0.088	-0.066	-0.15	-0.032	0.11	-0.1	0.24	0.012
45 Days from EOM	-0.055	-0.062	-0.077	-0.087	1	-0.041	-0.082	-0.11	-0.043	-0.032	-0.074	-0.016	0.052	-0.049	-0.079	-0.0065
45 Days from Inv Date	-0.081	-0.059	-0.075	-0.084	-0.041	1	-0.079	-0.1	-0.042	-0.031	-0.071	-0.015	0.05	-0.047	0.038	-0.027
60 Days from EOM	0.17	-0.12	-0.15	-0.17	-0.082	-0.079	1	-0.21	-0.083	-0.062	-0.14	-0.03	0.099	-0.094	-0.16	-0.022
60 Days from Inv Date	0.092	-0.15	-0.19	-0.22	-0.11	-0.1	-0.21	1	-0.11	-0.081	-0.18	-0.039	0.13	-0.12	-0.25	0.02
90 Days from EOM	-0.088	-0.062	-0.078	-0.088	-0.043	-0.042	-0.083	-0.11	1	-0.033	-0.074	0.37	-0.072	-0.049	0.18	0.059
90 Days from Inv Date	-0.024	-0.047	-0.058	-0.066	-0.032	-0.031	-0.062	-0.081	-0.033	1	-0.055	-0.0094	0.038	-0.037	0.092	0.047
Immediate Payment	0.15	-0.11	-0.13	-0.15	-0.074	-0.071	-0.14	-0.18	-0.074	-0.055	1	-0.027	-0.63	0.67	0.0071	-0.006
DM	-0.039	-0.023	-0.029	-0.032	-0.016	-0.015	-0.03	-0.039	0.37	-0.0094	-0.027	1	-0.31	-0.018	0.068	-0.011
INV	-0.0085	0.074	0.093	0.11	0.052	0.05	0.099	0.13	-0.072	0.038	-0.63	-0.31	1	-0.95	0.0027	-0.0058
CM	0.022	-0.07	-0.088	-0.1	-0.049	-0.047	-0.094	-0.12	-0.049	-0.037	0.67	-0.018	-0.95	1	-0.026	0.0098
cluster_id	-0.3	-0.25	0.23	0.24	-0.079	0.038	-0.16	-0.25	0.18	0.092	0.0071	0.068	0.0027	-0.026	1	0.021
Invoice_Month	-0.022	-0.007	-0.028	0.012	-0.0065	-0.027	-0.022	0.02	0.059	0.047	-0.006	-0.011	-0.0058	0.0098	0.021	1

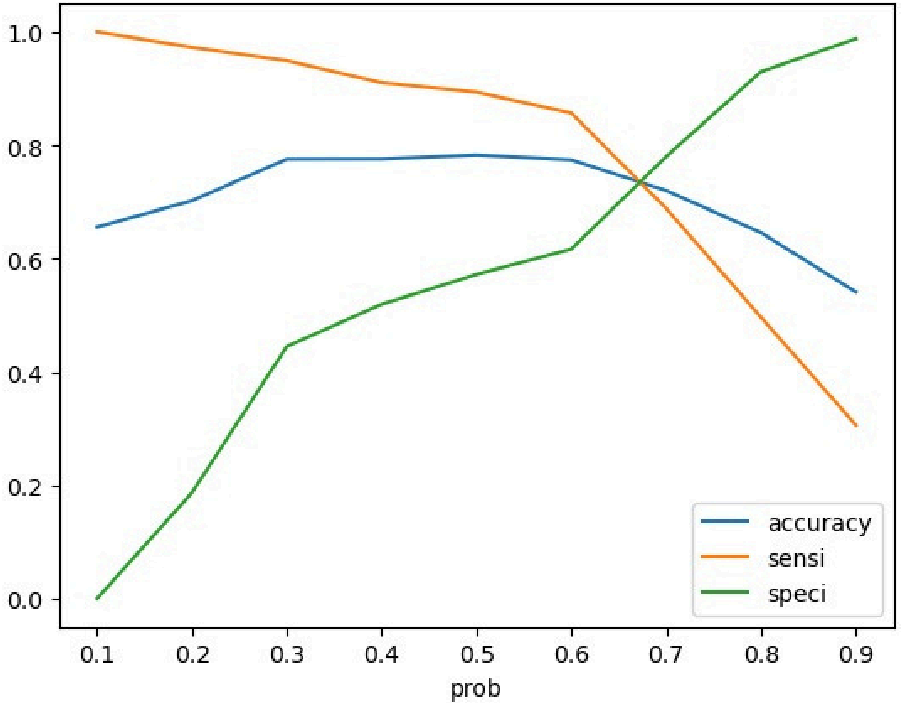
- CM & INV, INV & Immediate Payment, and DM & 90 days from EOM exhibit high multicollinearity; therefore, these columns will be dropped to mitigate the effects of multicollinearity.

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 VIF values. No further feature elimination was needed, and the model achieved a strong ROC curve area of 0.83.

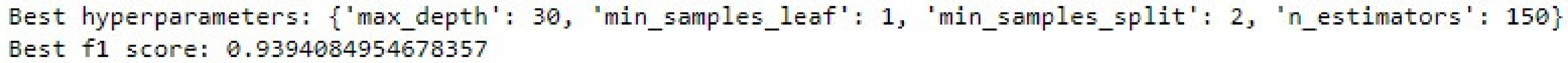
Comparison between two models, logistic regression and random forests



A cutoff of approximately 0.6 was applied to predict which transactions would lead to delayed payments in the payments dataset. The trade-off plot between accuracy, sensitivity, and specificity identified an optimal probability threshold.

Comparison between two models, logistic regression and random forests

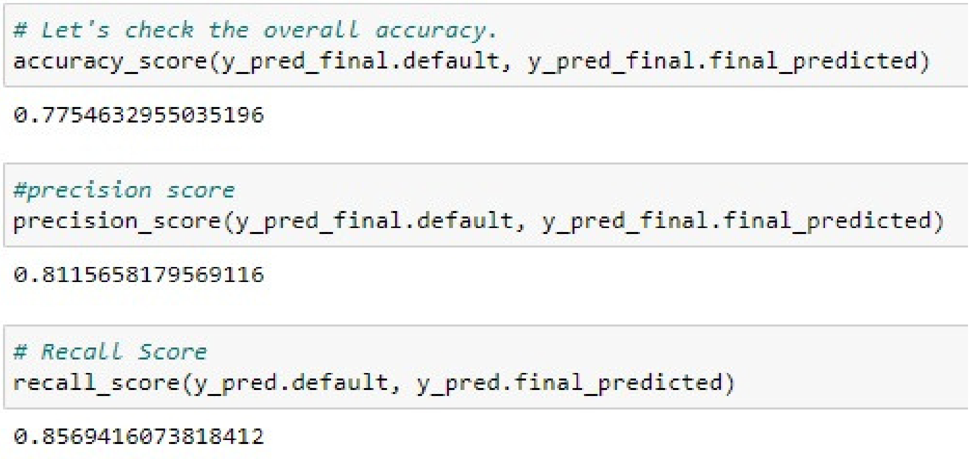
- A random forest model was developed using the same parameters as the logistic regression model, along with hyperparameter tuning, which yielded the following optimized settings.



```
Best hyperparameters: {'max_depth': 30, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 150}
Best f1 score: 0.9394084954678357
```

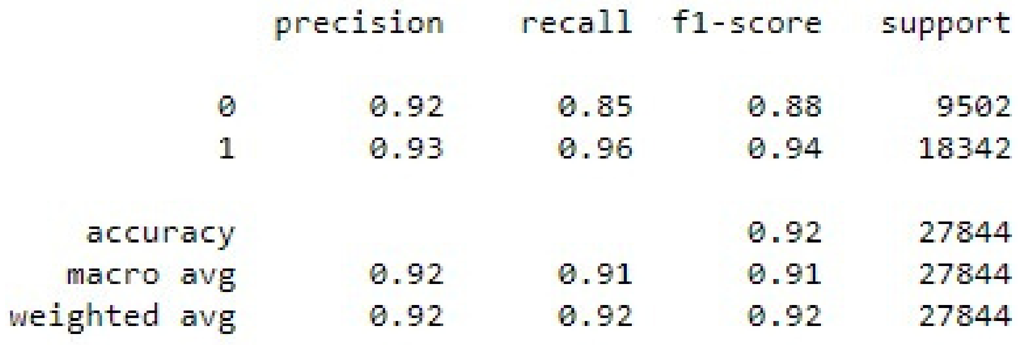
- With the optimized parameters, a random forest model was built and its metrics were compared to those of the logistic regression model, leading to the selection of the final model.

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
```

Fig. 1 (Logistic Regression Metrics -Test Set)



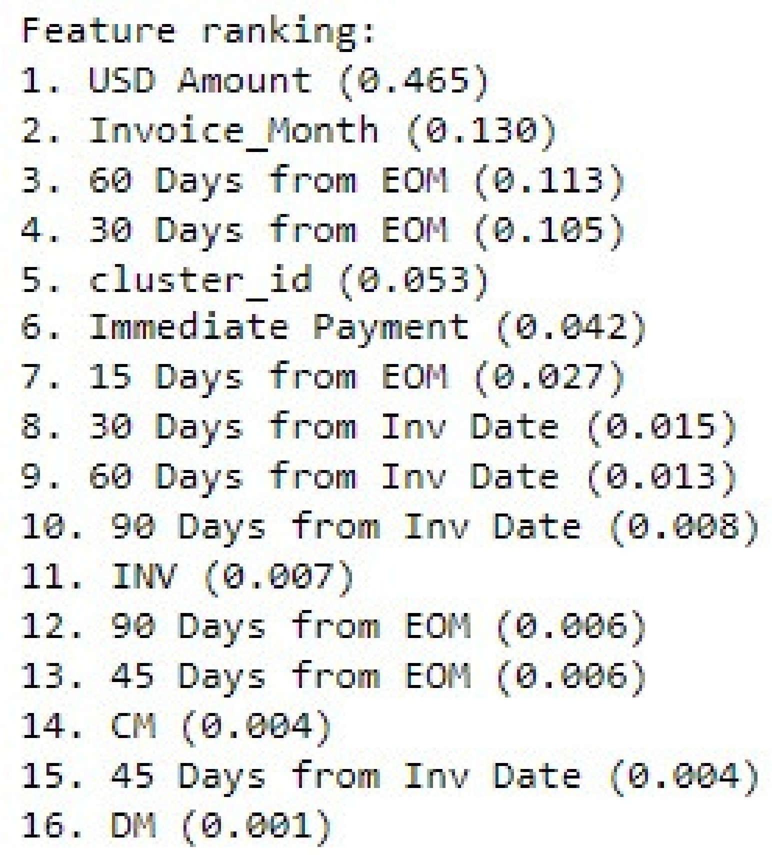
```
              precision    recall  f1-score   support

           0       0.92      0.85      0.88      9502
           1       0.93      0.96      0.94     18342

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

Fig. 2(Random Forest Metrics -Test Set)

Random Forest found better than Logistic Regression

- The random forest model showed significantly higher precision and recall scores compared to the logistic regression model.
- Recall was prioritised to improve the prediction accuracy for late payers, enabling better targeting.
- Due to the dataset’s high proportion of categorical variables, random forest was more suitable than logistic regression.
- As a result, the random forest model was selected as the final model for predictions.

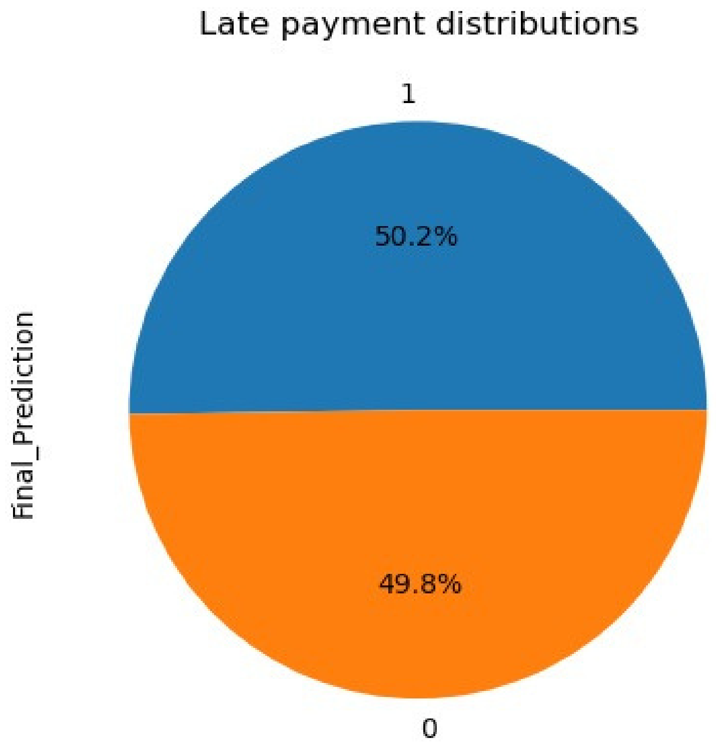
Random Forest FeatureRatings



```
Feature ranking:
1. USD Amount (0.465)
2. Invoice_Month (0.130)
3. 60 Days from EOM (0.113)
4. 30 Days from EOM (0.105)
5. cluster_id (0.053)
6. Immediate Payment (0.042)
7. 15 Days from EOM (0.027)
8. 30 Days from Inv Date (0.015)
9. 60 Days from Inv Date (0.013)
10. 90 Days from Inv Date (0.008)
11. INV (0.007)
12. 90 Days from EOM (0.006)
13. 45 Days from EOM (0.006)
14. CM (0.004)
15. 45 Days from Inv Date (0.004)
16. DM (0.001)
```

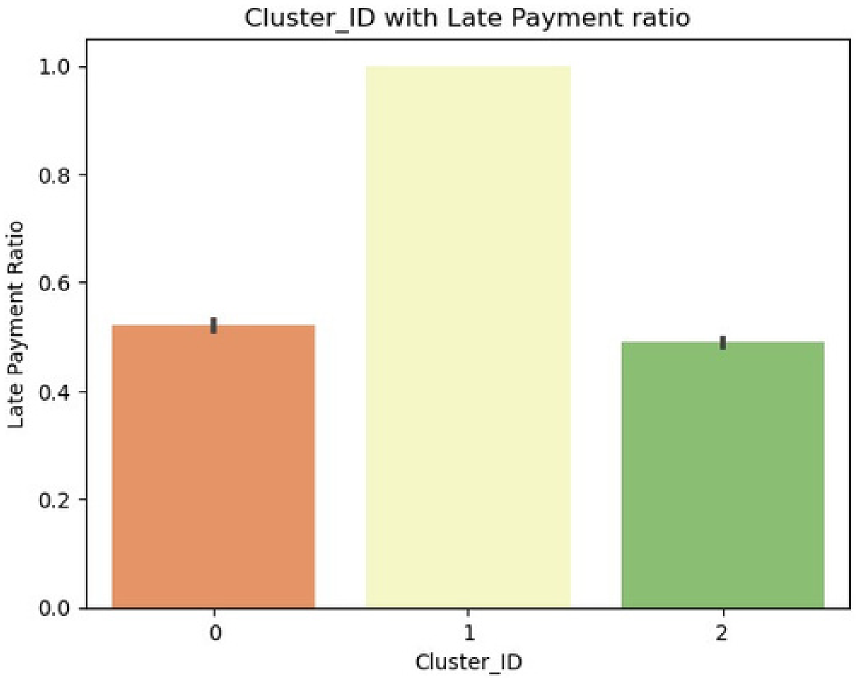
- The random forest model was used to determine feature rankings, identifying the top five features for predicting delays, which included:
- USD Amount
- Invoice Month
- 60 Days from EOM (a Payment Term variable)
- 30 Days from EOM (a Payment Term variable)
- Cluster ID (based on the average and standard deviation of days taken to make payment)
- The customer segmentation with Cluster ID was applied to the open-invoice data by customer name, and predictions were made accordingly.

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



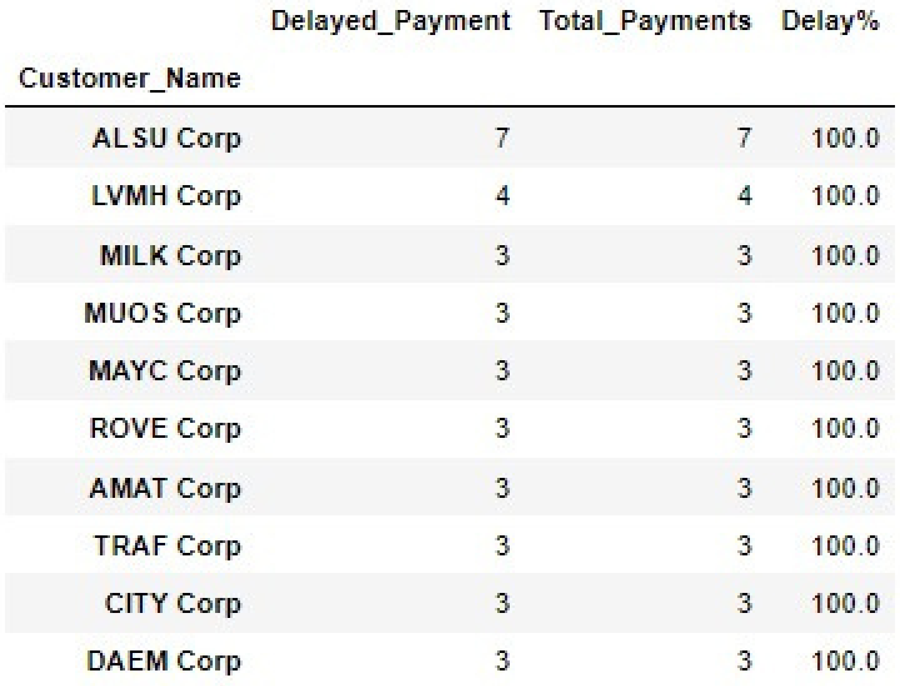
- Predictions from the final model indicate a probable 50.2% of transactions with expected payment delays, which could result in a significant disruption to business operations.

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



- Customers who have historically taken longer to make payments are predicted to have the highest delay rate (approximately 100%), consistent with findings from previous historical outcomes for early or medium payment transactions.

Customers with the highest delay probabilities



	Delayed_Payment	Total_Payments	Delay%
Customer_Name			
ALSU Corp	7	7	100.0
LVMH Corp	4	4	100.0
MILK Corp	3	3	100.0
MUOS Corp	3	3	100.0
MAYC Corp	3	3	100.0
ROVE Corp	3	3	100.0
AMAT Corp	3	3	100.0
TRAF Corp	3	3	100.0
CITY Corp	3	3	100.0
DAEM Corp	3	3	100.0

- Predictions indicate that the companies listed in the table on the left have the highest probability of default, accompanied by the greatest number of delayed and total payments.

Recommendations

- From our clustering analysis, we can draw the following insights:

- Credit Note payments have the highest delay rate compared to Debit Notes and Invoices. Therefore, stricter payment collection policies could be applied specifically to Credit Note transactions.

- Goods-type invoices show significantly higher delay rates than non-goods types, suggesting that stricter payment policies could also be applied to these invoices.

Recommendations

- Since most transactions are lower-value payments, and late payments are more common among them, focusing on these transactions is recommended. The company could consider applying penalties based on billing amounts, with smaller bills incurring higher penalty percentages for late payments—although this should remain a last-resort measure.

Recommendations

- Customers were segmented into three clusters: 0, 1, and 2, representing medium, prolonged, and early payment durations, respectively. Cluster 1 (prolonged payment duration) customers showed significantly higher delay rates than those in the early and medium categories, so they should receive additional attention.

- Companies with the highest probability and the largest counts of total and delayed payments, as shown above, should be prioritized for focused action due to their elevated probability of payment delays.

Thank You