
Business to Business
B2B

E-Commerce and Retail B2B Case Study

SIDDHARTH SINGHAL

Addressing the issue and defining objectives

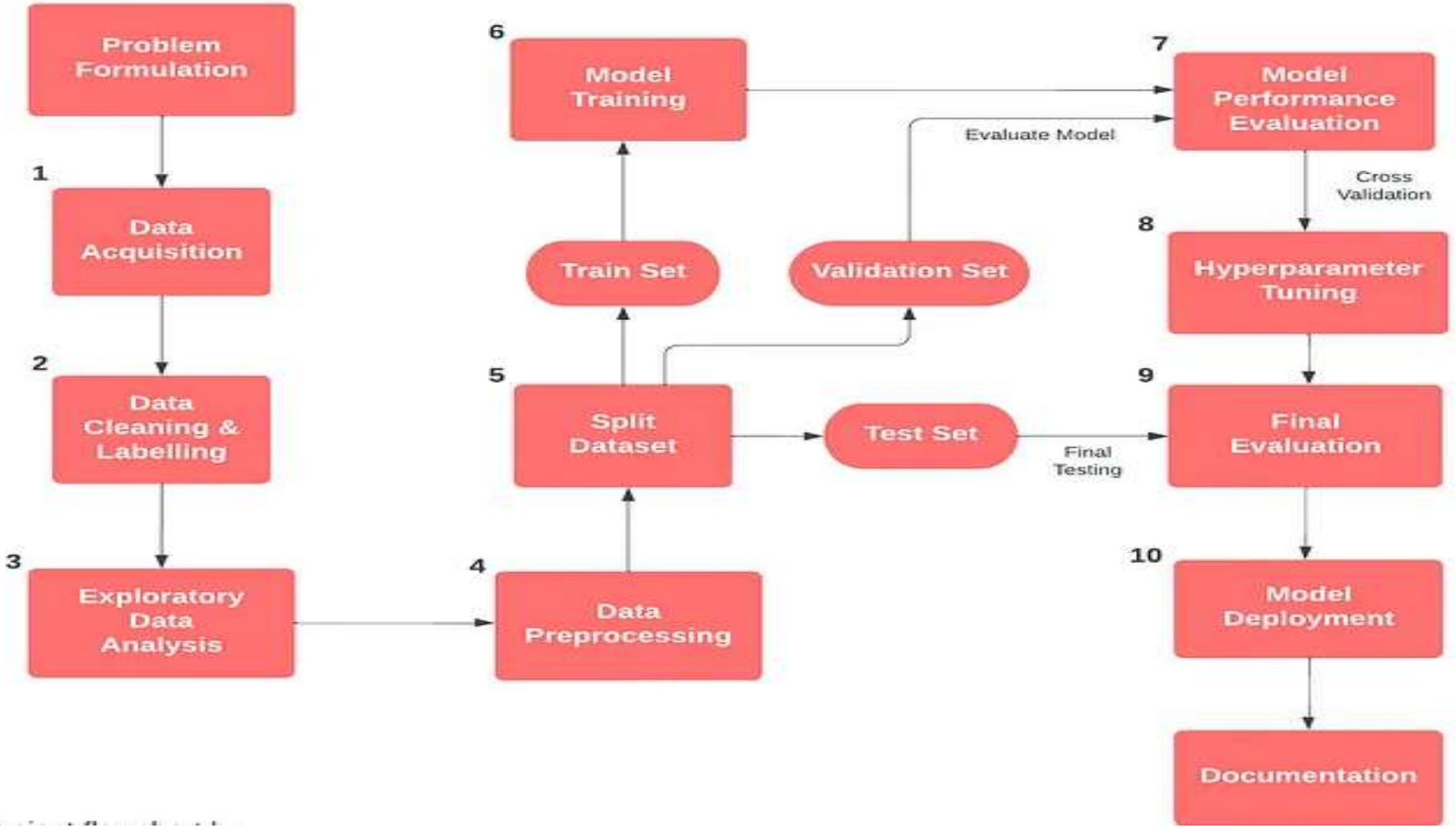
Problem Identification

•Schuster, a sports retail company engaged in B2B transactions, frequently extends credit to its vendors. However, some vendors fail to adhere to the agreed payment deadlines.
•Payment delays by vendors cause financial bottlenecks, adversely impacting smooth business operations.
•Employees are often occupied with lengthy payment collection efforts, leading to inefficient resource utilization and reduced focus on value-adding activities.

Business Objectives

•Segment customers based on their payment behavior to gain deeper insights.
•Leverage historical data to predict delayed payments for transactions with pending due dates in a new dataset.
•Utilize these predictions to optimize resource allocation, accelerate credit recovery, and minimize non-value-adding activities.

Approach Strategy to the Problem



Class imbalance and transaction insights (univariate)

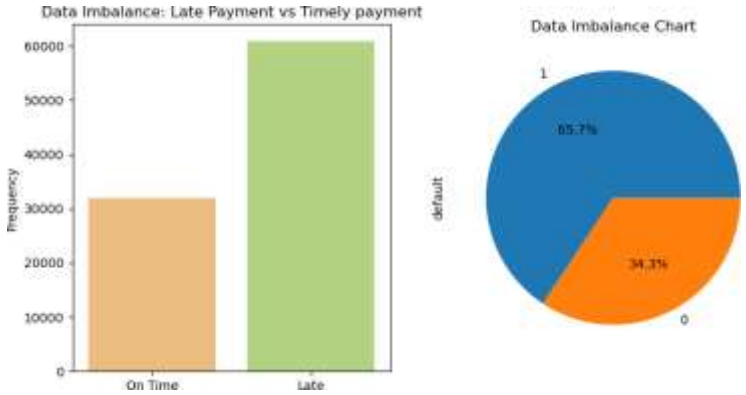


Fig. 1

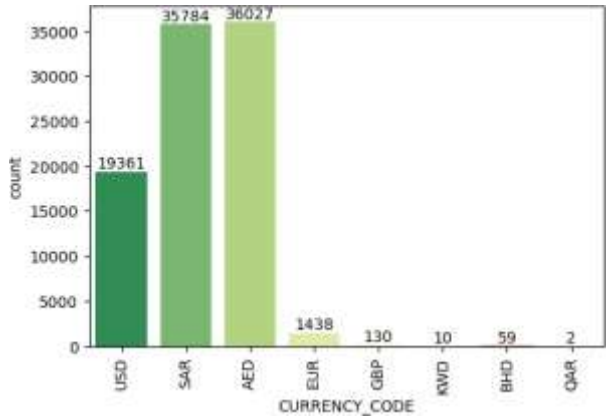


Fig. 2

From Fig. 1 and 2:

- The class imbalance is 65.7% towards payment delayers which is an acceptable imbalance and does not need imbalance treatment
- The top three currencies in which the company deals are AED, SAR and USD with AED as the most dealt currency suggesting greater transactions with the middle-east

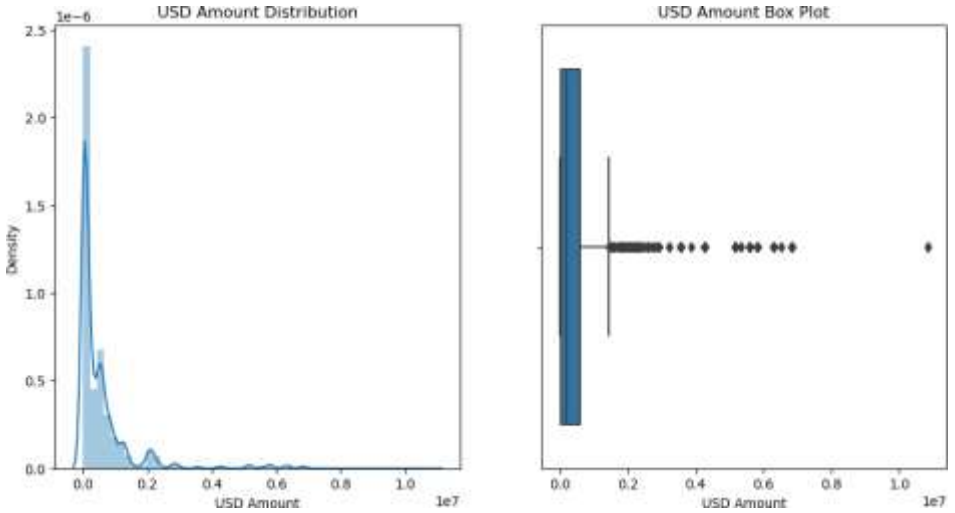


Fig. 1

From Fig. 3, we observe,

- The transaction values seem to lie between a range of $1 and $3m
- The transaction values are most frequent below ~$1.75m

Class imbalance and transaction insights (univariate)

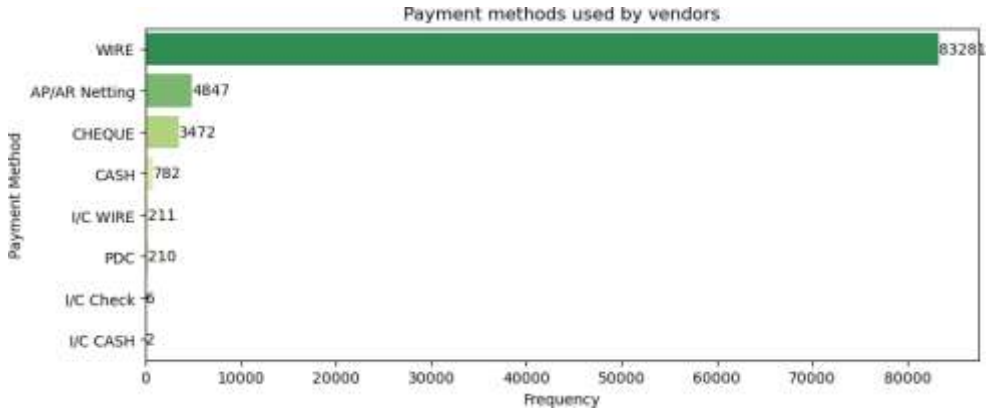


Fig. 1

From Fig. 1, we observe,

- Wire payment method is the most common payment method received by the company, followed by netting , cheque and cash

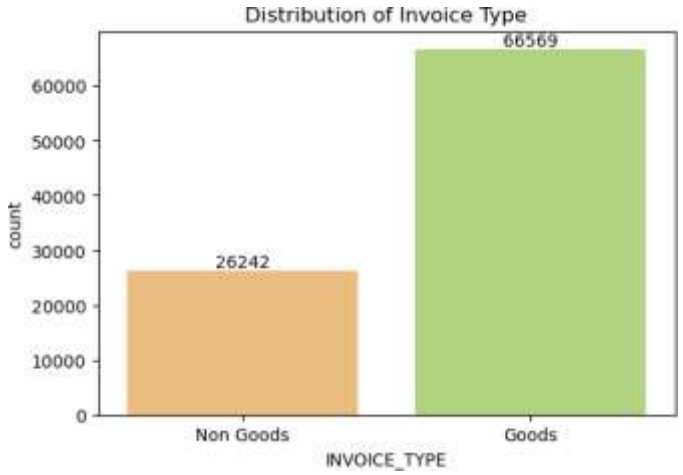


Fig. 2

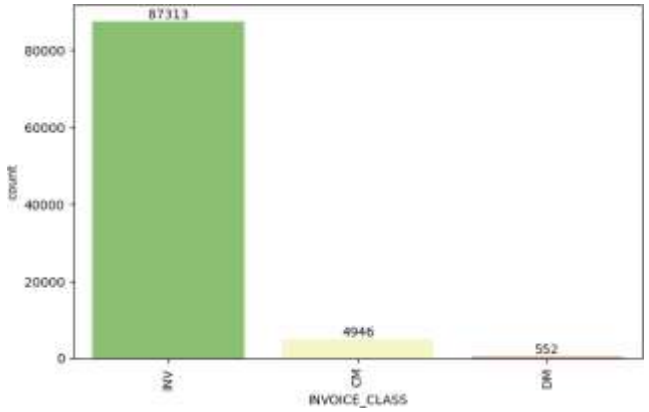


Fig. 3

From Fig. 2 and 3:

- Goods type invoices comprise of the major share of invoices generated
- The major invoice class is ‘Invoice’ with the rest having very low percentages of the share

Identifying characteristics of defaulter payment types (Bivariate)

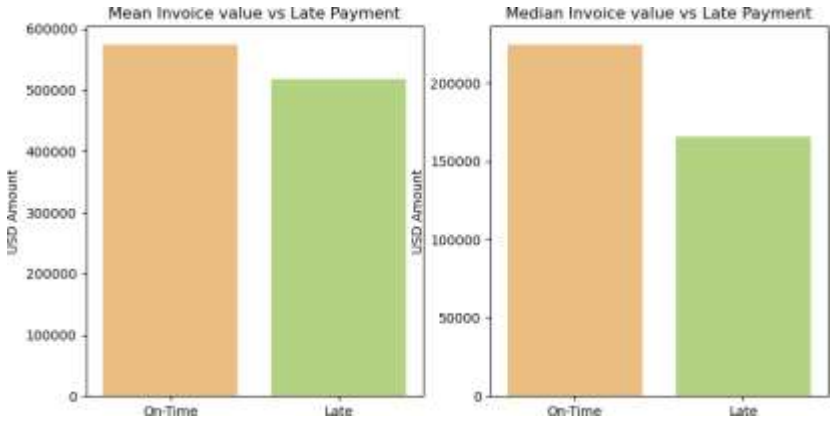


Fig. 1

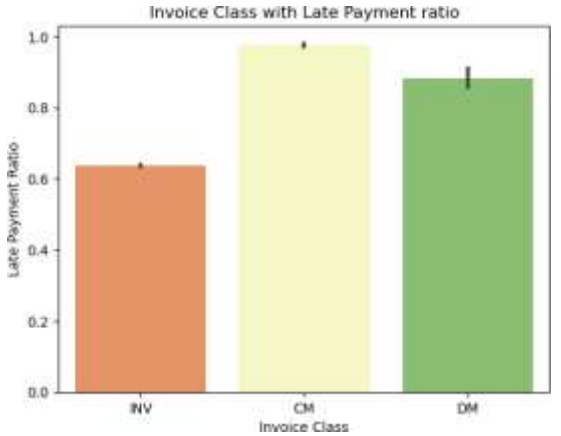


Fig. 2

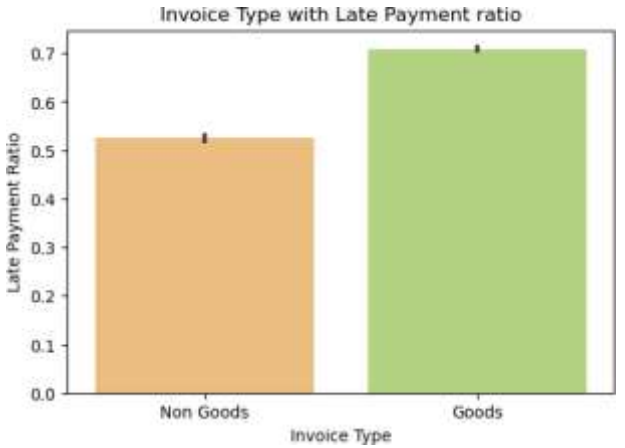


Fig. 3

- From fig. 1, the mean and median of the payment amount is higher for payers who pay on time than late, suggesting that higher value transactions show lesser delay risk than lower value transactions

- From fig. 2, late payment ratio for Credit Note transaction types are maximum, followed by Debit Note and Invoice suggesting higher delay risk in Credit and Debit note invoice classes

- From fig. 3, Goods type invoices show greater late payment ratio than non-goods hence showing increased chances of payment delay

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 0.73503646233166
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 were decided to be 3 since with increase in clusters post 3, there was a significant decrease in silhouette score

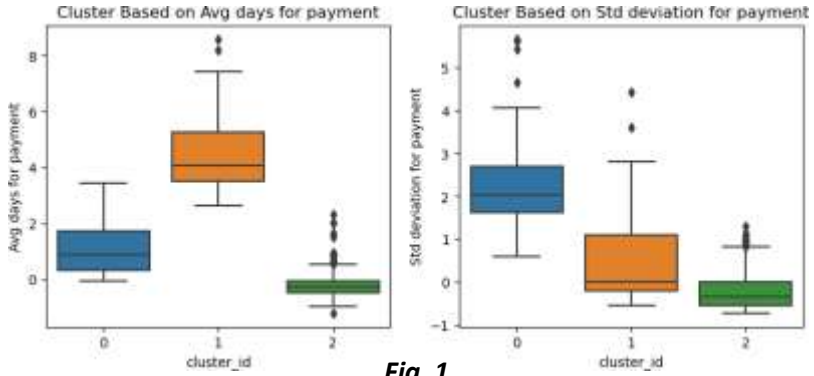


Fig. 1

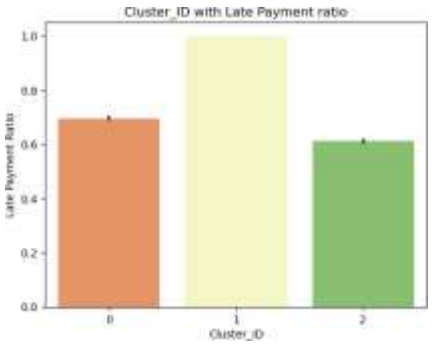
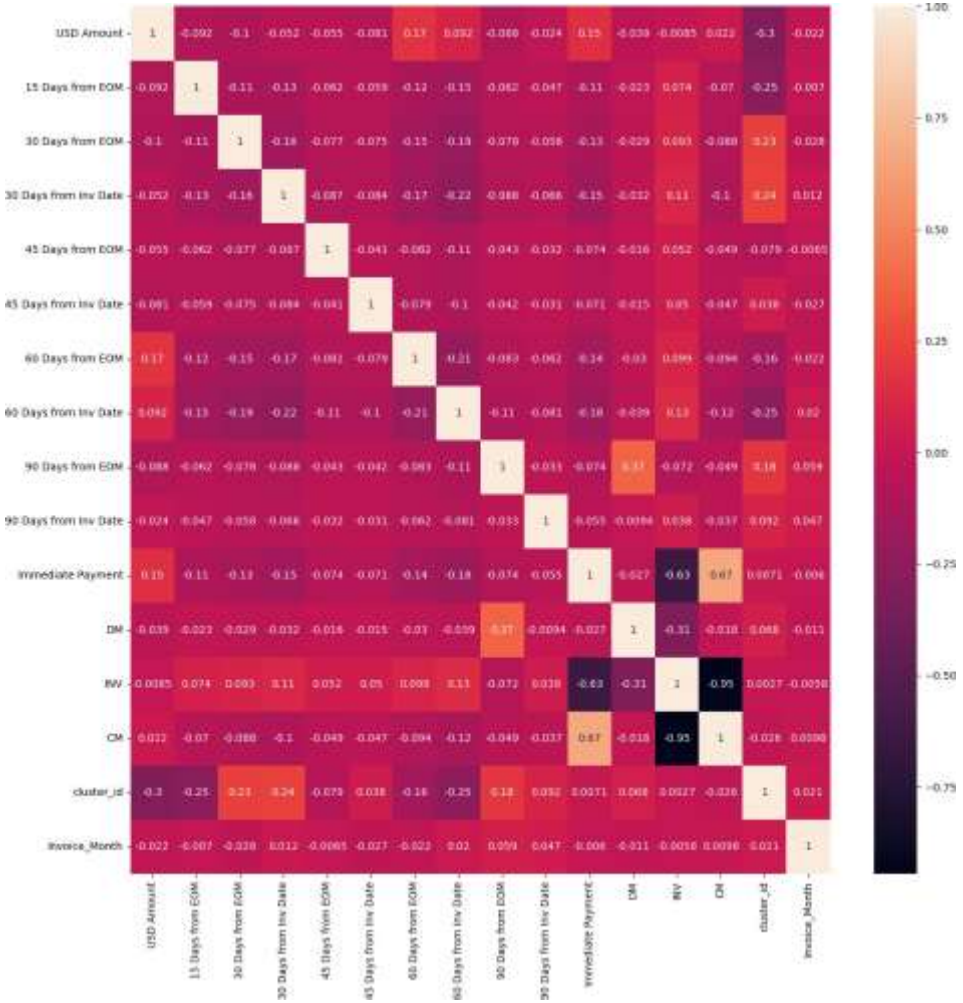


Fig. 2

- The category 2 were early payers with least number of average days taken to pay and category 1 were prolonged payers with greatest number of average days taken to pay. Category 0 lie in between the other two categories and hence labelled as medium duration payers

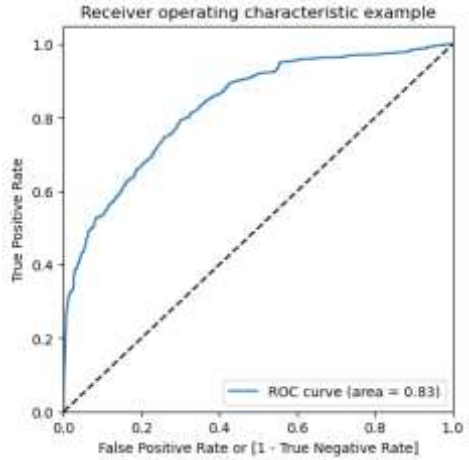
- It was also observed that prolonged players historically have significantly greater rates of delay in payment than early or medium duration payment transactions (Fig 2.)

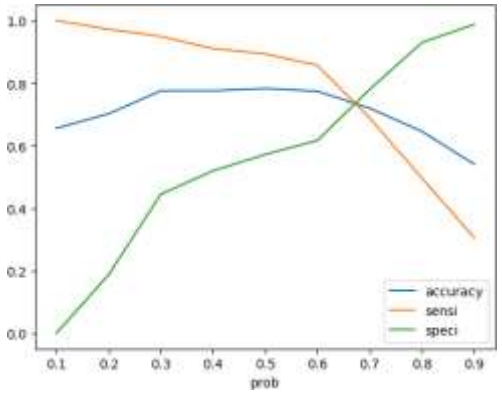
Model Building



- CM & INV, INV & Immediate Payment, DM & 90 days from EOM has high multicollinearity, hence dropping these columns to prevent multicollinearity effect

Comparison between two models, logistic regression and random forests





- The logistic regression model was refined by removing multicollinear and irrelevant variables. The remaining features exhibited acceptable p-values and VIF values, justifying their retention. This final model achieved a strong ROC curve area of 0.83, negating the need for further feature elimination.

- The trade-off plot between accuracy, sensitivity and specificity revealed an optimum probability cutoff of ~0.6, which was used to further predict which transactions would result in delayed payments in the received payments dataset

Comparison between two models, logistic regression and random forests

- A random forest model was built using the same parameters as the logistic regression with hyper-parameter tuning, which resulted in the following parameters

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

- 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

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)

Model Selection and Rationale

•Comparison of Precision and Recall Scores:

The Random Forest model significantly outperformed the Logistic Regression model in terms of both precision and recall scores.

•Importance of Recall:

Recall was prioritized in this case to maximize the identification of late payers, ensuring more accurate targeting for payment recovery efforts.

•Suitability of Random Forest:

Given the dataset's heavy reliance on categorical variables, the Random Forest model proved to be better suited for handling such data compared to Logistic Regression.

•Final Decision:

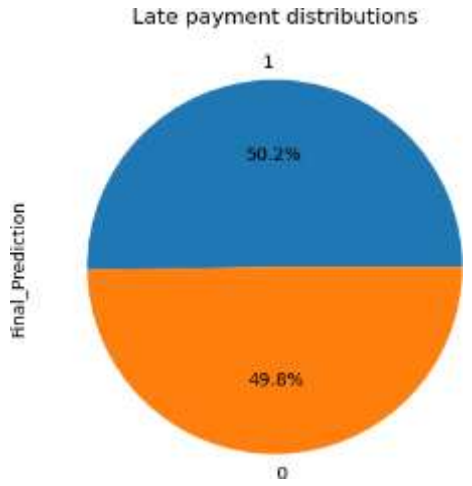
The Random Forest model was selected as the preferred model for making predictions, ensuring robust and actionable insights for targeting late payers effectively

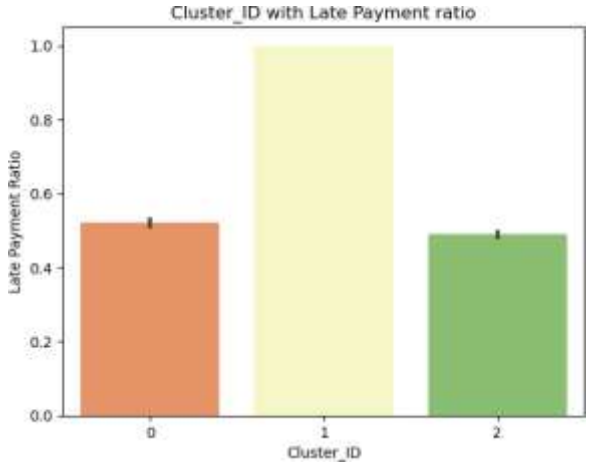
Random Forest Feature Ratings

```
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 was then used to find out the feature rankings which shows that the top 5 features to predict delay which included
  - USD Amount
  - Invoice Month
  - 60 Days from EOM (Payment Term variable)
  - 30 Days from EOM (Payment Term variable)
  - Cluster-ID (which in turn is dependent on average and standard deviation of days required to make payment)
- The customers segmented with cluster ID was then applied to the open-invoice data as per the customer name and predictions were made

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





- Predictions made by the final model suggests that there is a probable 50.2% transactions where payment delay can be expected, which can cause a shocking lag to business operations

- Customer segment with historically prolonged payment days are anticipated to have the most delay rate (~100%) than historically early or medium days payment transactions, this is similar to the result found based on historical outcomes

Customers with the highest delay probabilities

Customer_Name	Delayed_Payment	Total_Payments	Delay%
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 suggest that the companies presented in the table to the left has the maximum probability of default with maximum number of delayed and total payments

Recommendations

Customer_Name	Delayed_Payment	Total_Payments	Delay%
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

Fig. 1

Inferences from Clustering Analysis

•Higher Delay in Credit Note Payments:

Credit Note Payments exhibit the highest delay rates compared to Debit Notes or Invoice-type classes. The company can consider implementing stricter payment collection policies specifically for these invoice categories.

•Greater Delays in Goods-Type Invoices:

Goods-type invoices show significantly higher payment delays than non-goods types. Stricter payment policies can be enforced for these invoices to mitigate delays.

•Focus on Low-Value Payments:

Low-value payments form the majority of transactions and are more prone to delays. It is advisable to concentrate collection efforts on these transactions. As a last resort, penalties could be applied on late payments, with higher penalty percentages for lower bill amounts.

•Customer Clusters and Payment Behavior:

Customers were categorized into three clusters:

- Cluster 0: Medium payment duration
- Cluster 1: Prolonged payment duration
- Cluster 2: Early payment duration

  Cluster 1 customers, associated with the longest payment delays, should receive increased attention to address their high delay rates effectively.

•Priority Companies:

Companies identified in Fig. 1, which demonstrate high probabilities of delay and significant total and delayed payment counts, should be prioritized for focused collection efforts due to their high risk of delayed payments.