E-Commerce and Retail B2B Case Study

DEEPALI PANDIT

HARSHINI KOTAPATI

KASHYAP RAGHVENDRA

Problem Statement & Objective

Problem statement:

- A sports retail firm, Schuster, engaged in B2B transactions, frequently conducts business with vendors on credit, who may or may not adhere to the agreed payment deadlines.
- Vendors delaying their payments result in financial lag and loss which becomes detrimental to smooth business operations
- Additionally, company employees are set up chasing around for collecting payments for a long period of time resulting in no value-added activities and wasteful resource expenditure

Objective:

- Customer segmentation to understand the customer's payment behavior
- The business needs to predict late payments using past data against an unexpected dataset of transactions with unmet due dates.
- The company requires the prediction for better resource delegation, quicker credit recovery and reduction of low value-adding activities

Univariate Analysis

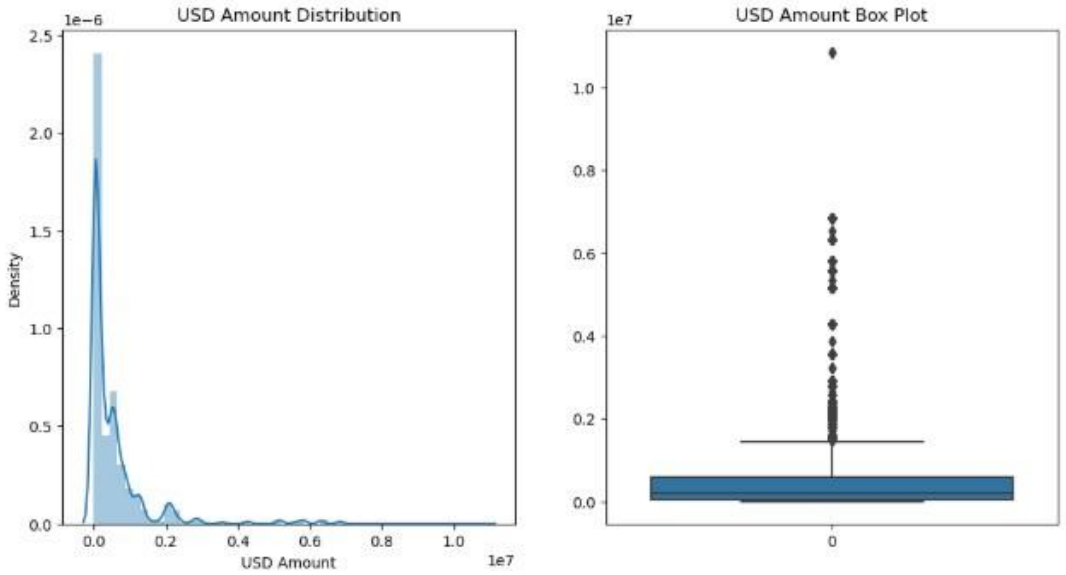


Fig 1: USD amount distribution

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

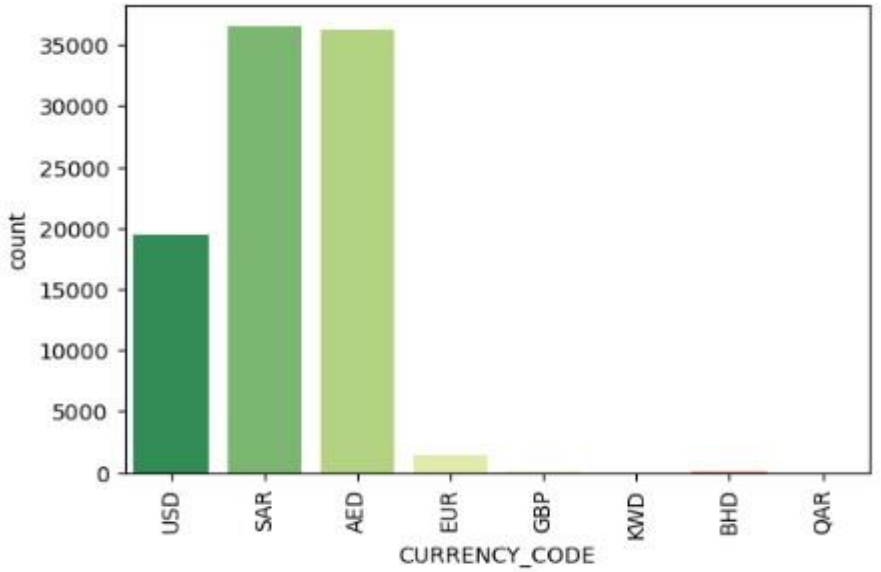


Fig 2

Fig 2: 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

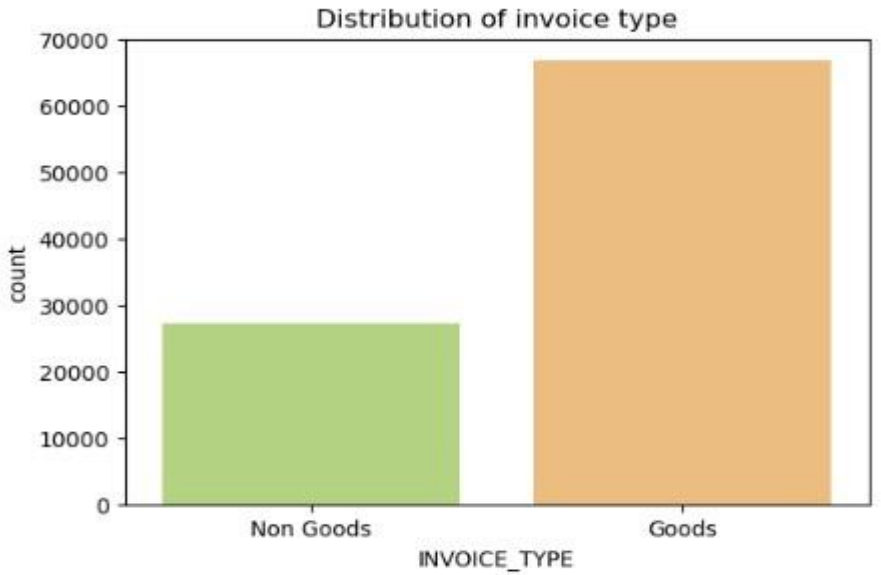


Fig 3

Max generated invoices were for goods product

Univariate Analysis

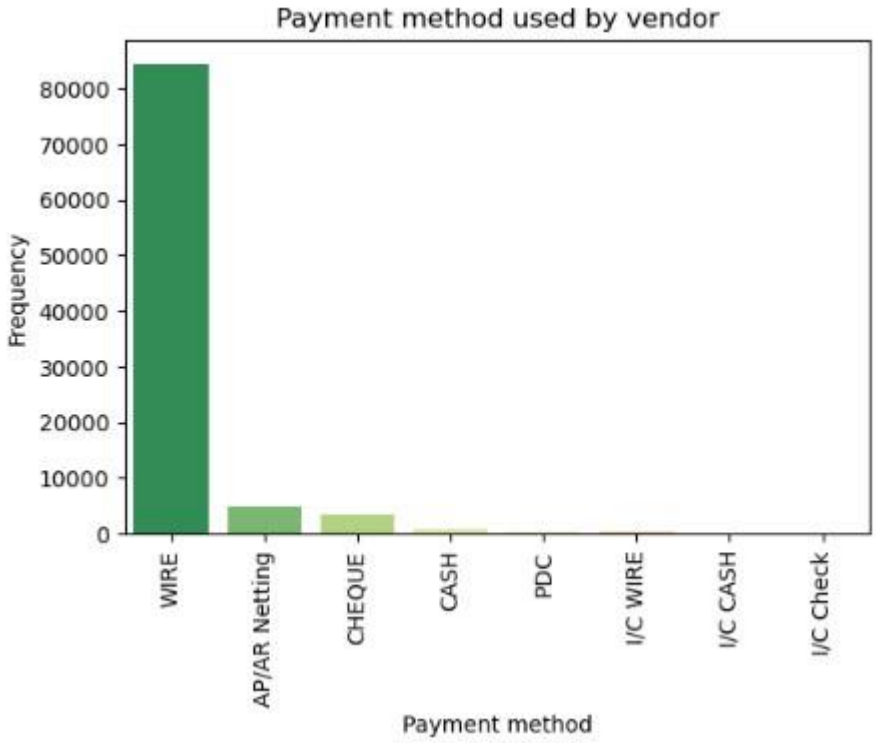


Fig 4

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

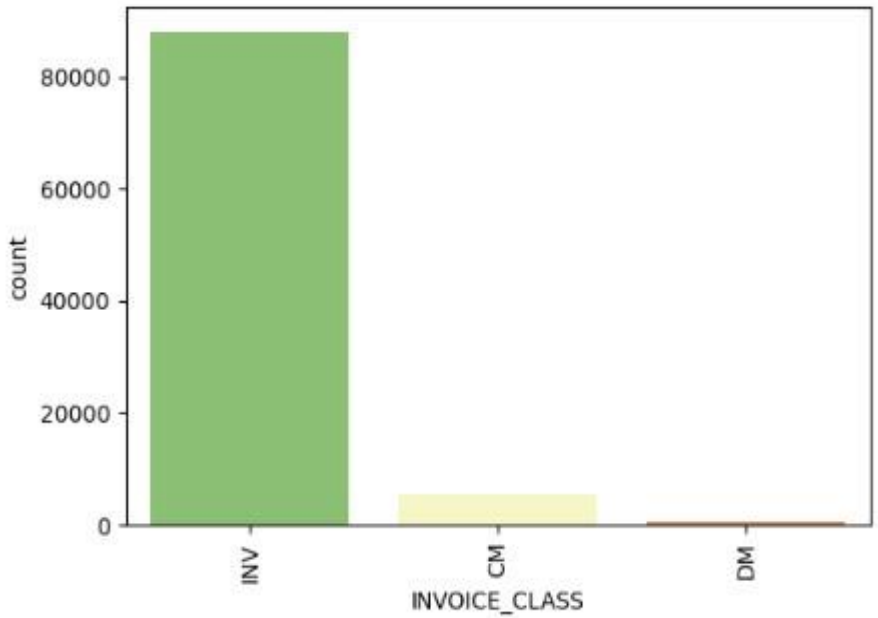


Fig 5

The major invoice class is ‘Invoice’ with the rest having very low percentages of the share

Data imbalance

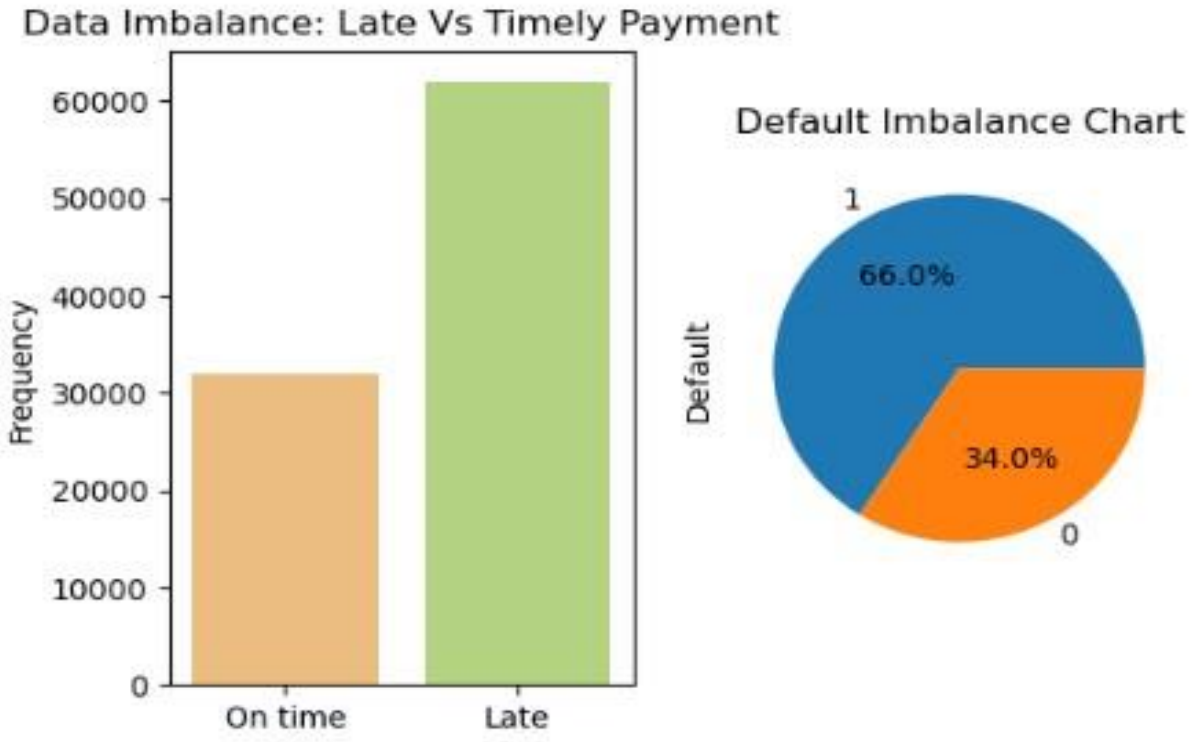


Fig 6

- The class imbalance is 65.7% towards payment delayers which is an acceptable imbalance and does not need imbalance treatment

Cluster Segmentation : K-means

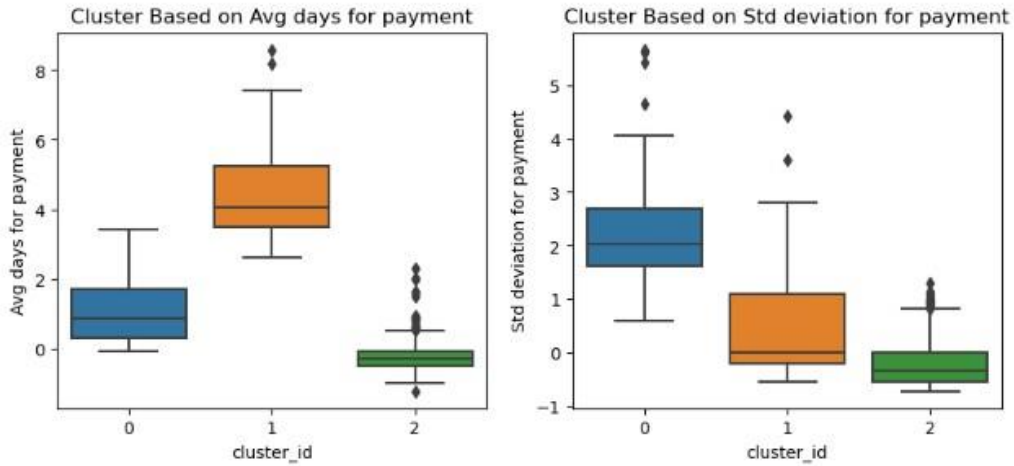


Fig 7

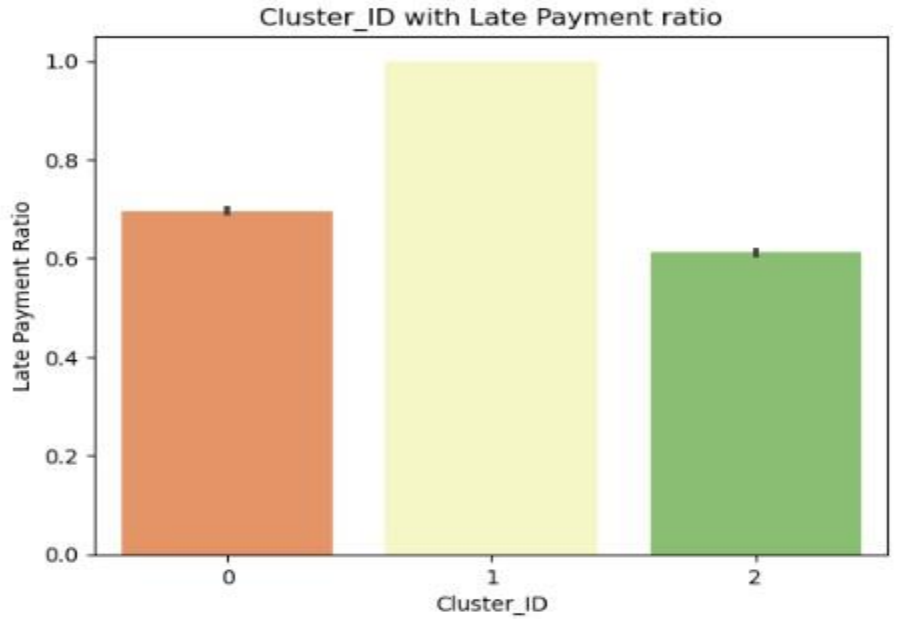


Fig 8

Three clusters were chosen because the silhouette score significantly decreased as the number of clusters increased after three.

It was also observed that prolonged players historically have significantly greater rates of delay in payment than early or medium duration payment transactions

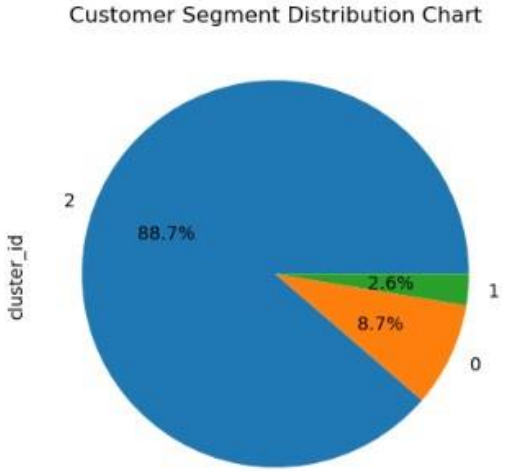


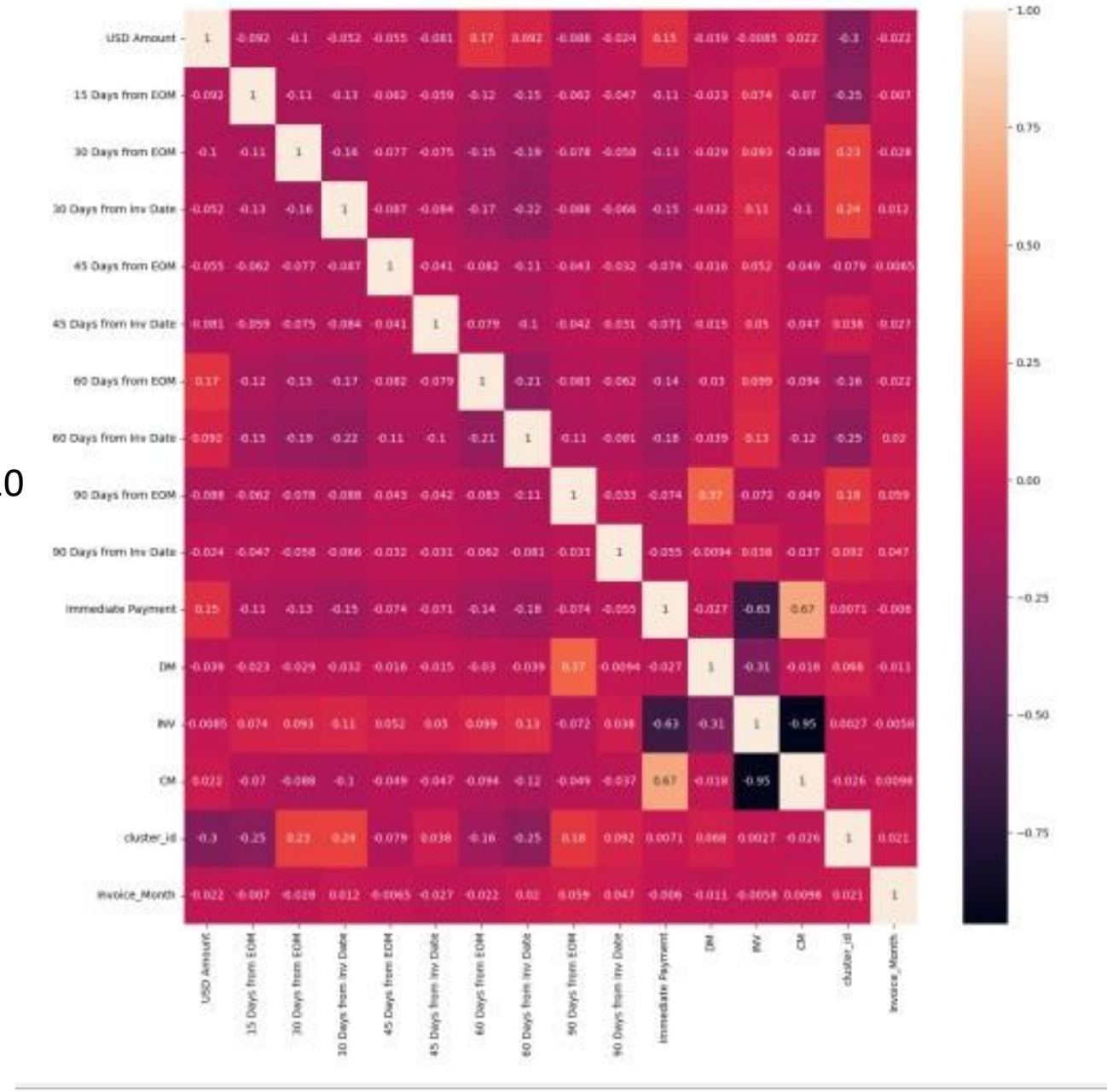
Fig 9

Early customers comprise of 88.7% of customers whereas medium and prolonged payers are 11.3% in total

Model Building

Due to their high levels of multicollinearity, CM & INV, INV & Immediate Payment, DM & 90 days from EOM have been dropped in order to avoid the multicollinearity effect.

Fig 10



Random Forest Vs Logistics Regression

```
In [188]: # Let's check the overall accuracy.
          accuracy_score(y_pred_final.Default, y_pred_final.final_predicted)

Out[188]: 0.7754632955035196

In [189]: #precision score
          precision_score(y_pred_final.Default, y_pred_final.final_predicted)

Out[189]: 0.8115658179569116

In [190]: # Recall Score
          recall_score(y_pred.Default, y_pred.final_predicted)

Out[190]: 0.8569416073818412
```

Fig 11

	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 12

- It is evident that the Random Forest model outperformed the logistic regression model in terms of overall precision and recall scores. Furthermore, as it was crucial to raise the percentage prediction of late payers to be targeted, recall score were particularly significant in this case.
- Since the data is heavy on categorical variables, random forest is better suited to the job than logistic regression
- Therefore, random forest model was finalized to be the model of choice and go forward with predictions

Random Forest Feature Rating

```
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)
```

Fig 13

The random forest was then used to find out the feature rankings which shows that the top 5 features to predict delay which included

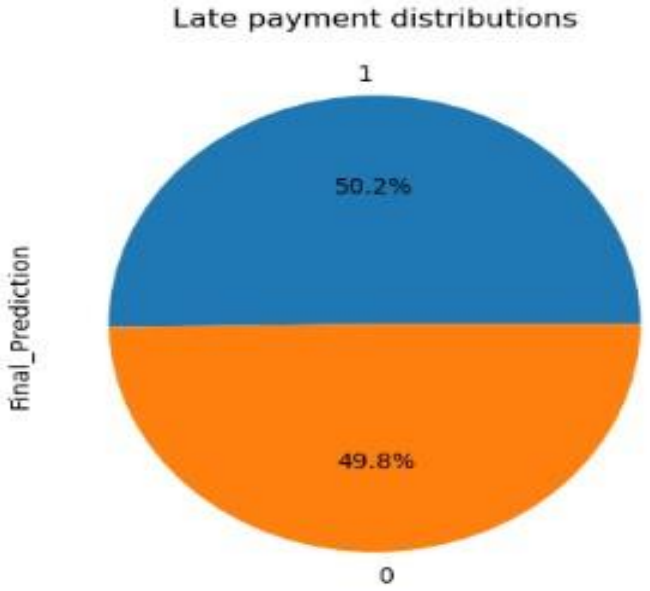
- USD Amount
- Invoice Month
- 60 Days from EOM (Payment Term variable)
- 30 Days from EOM (Payment Term variable)
- Cluster-ID (which in turn is dependent on average and standard deviation of days required to make payment)

The customers segmented with cluster ID was then applied to the open-invoice data as per the customer_name and predictions were made.

Final prediction with Open invoice dataset

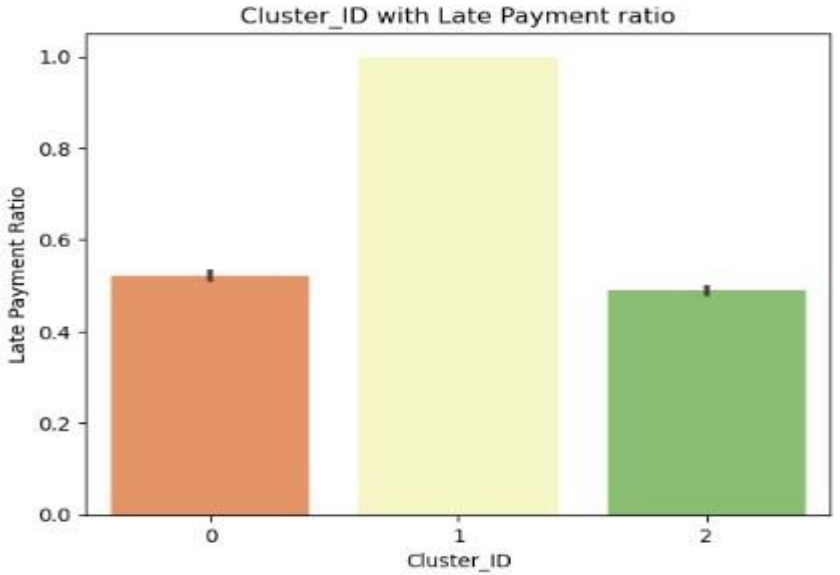
Predictions generated by the final model suggests that there is a possible 50.2% transactions where payment delay can be expected, which can cause a shocking lag to business operations

Fig 14



According to historical results, the customer segment with prolonged payment days is expected to have the highest delay rate (~100%) compared to historically early or medium-day payment transactions.

Fig 15



Customers with 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

According to predictions, the businesses listed in the table on the left have the highest chance of going into default, with the highest number of late and total payments.

Recommendations

- Compared to debit note or invoice type invoice classes, credit note payments have the highest delay rate; therefore, firm policies regarding payment collection should be rigid with regard to these invoice classes.
- Stricter payment regulations may be applied to goods-type invoices since they had noticeably higher payment delay rates than non-goods-type invoices.
- It is advised to concentrate more on lesser value payments because they make up the majority of transactions and are also more likely to have late payments. The company can apply penalties depending on billing amount.
- Three categories—0, 1, and 2—which stand for medium, prolonged, and early payment durations, respectively—were created by clustering customer segments. Cluster 1 consumers should receive special attention because their delay rates were noticeably higher than those of early and middle days of payment.
- Final 10 customers should be focused more as their probability is highest

Thank You