E-commerce *and B2B Case Study*

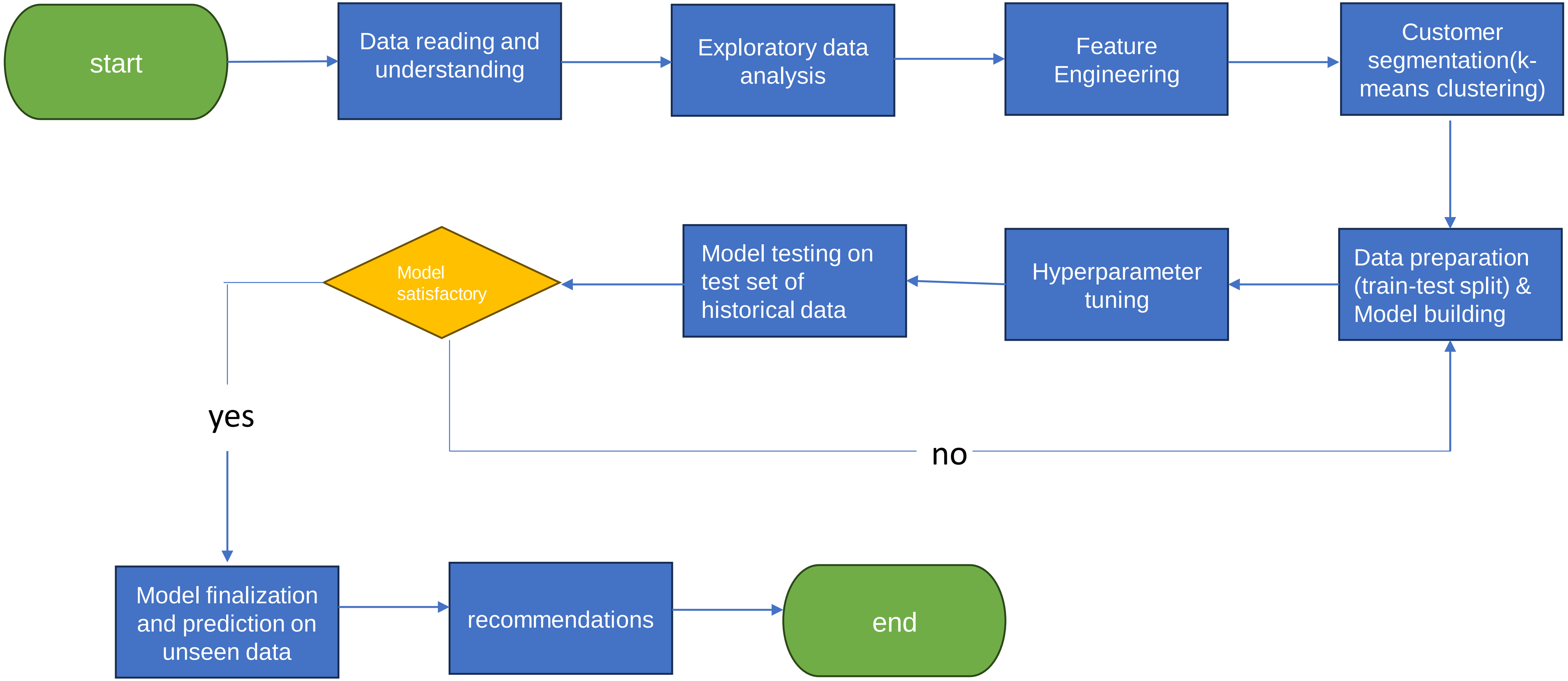
Problem statement

- A sports retail company Schuster dealing in B2B transactions often deals with vendors on a credit basis, who might or might not respect the stipulated deadline for payment
- 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

Business Objectives

- Customer segmentation to understand the customer’s payment behavior
- Using historical information, the company requires prediction of delayed payment against an unforeseen dataset of transactions with due date yet to be crossed
- The company requires the prediction for better resource delegation, quicker credit recovery and reduction of low value-adding activities

STEPS INVOLVED



start

Data reading and understanding

Exploratory data analysis

Feature Engineering

Customer segmentation(k-means clustering)

Data preparation (train-test split) & Model building

Hyperparameter tuning

Model testing on test set of historical data

Model satisfactory

yes

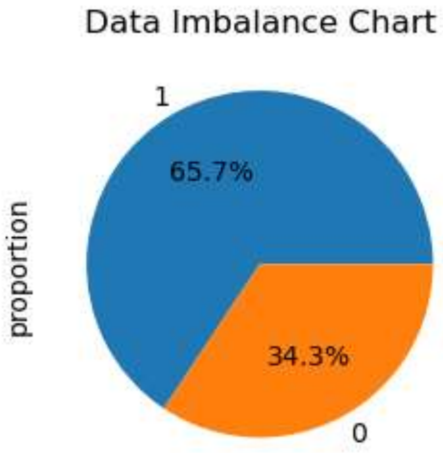
no

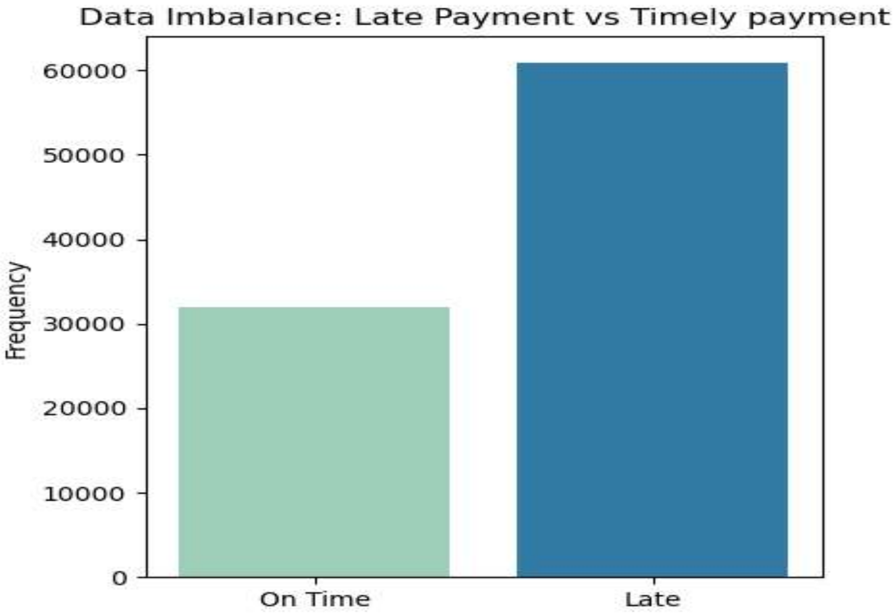
Model finalization and prediction on unseen data

recommendations

end

Data imbalance





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

Univariate analysis

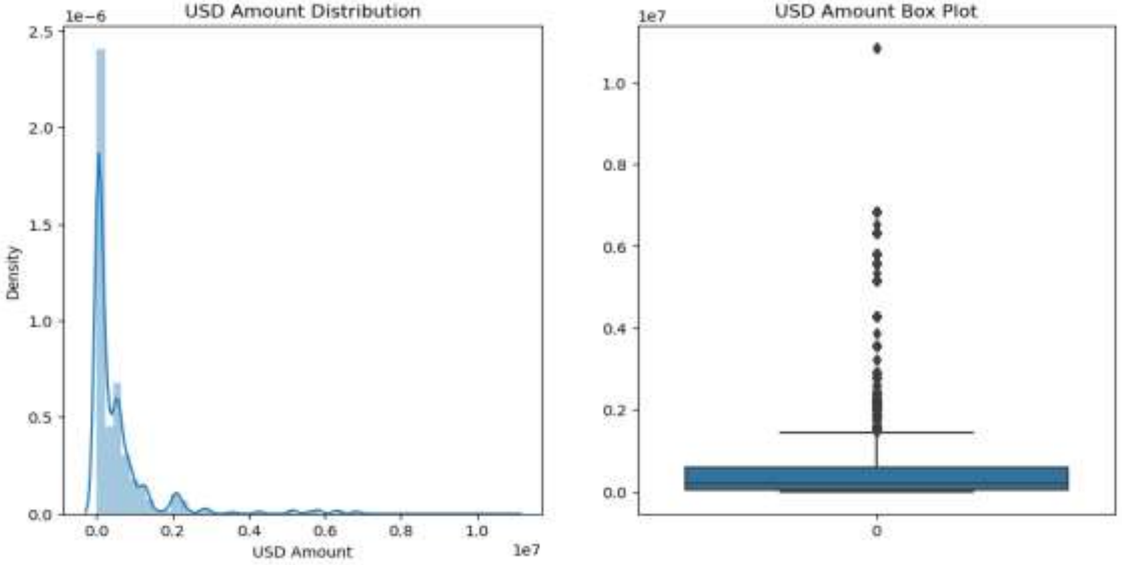


Fig 1.

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

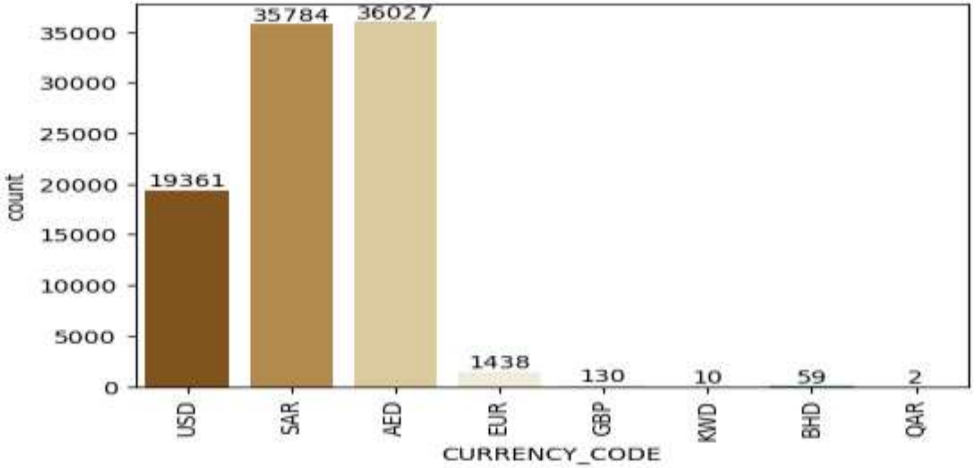


Fig 2.

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.

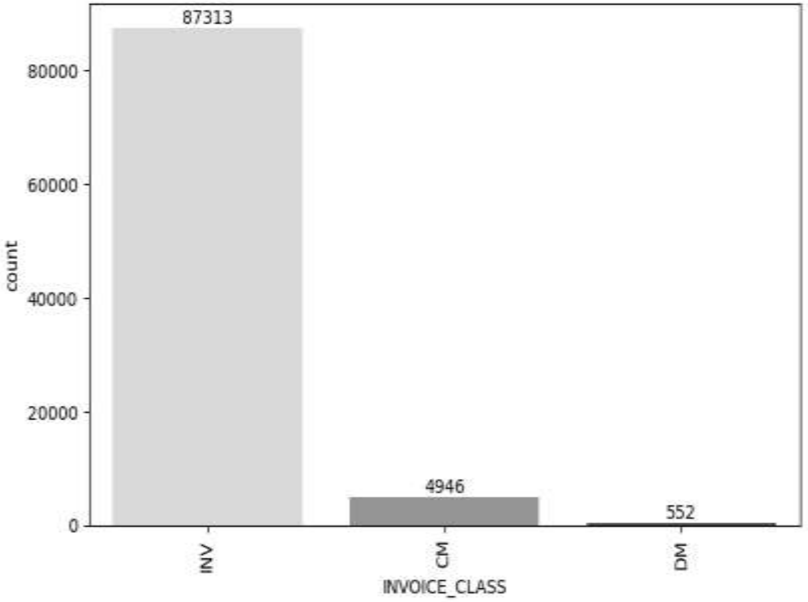


Fig 3.

- The major invoice class is 'INV' with the rest having very low percentages of the share

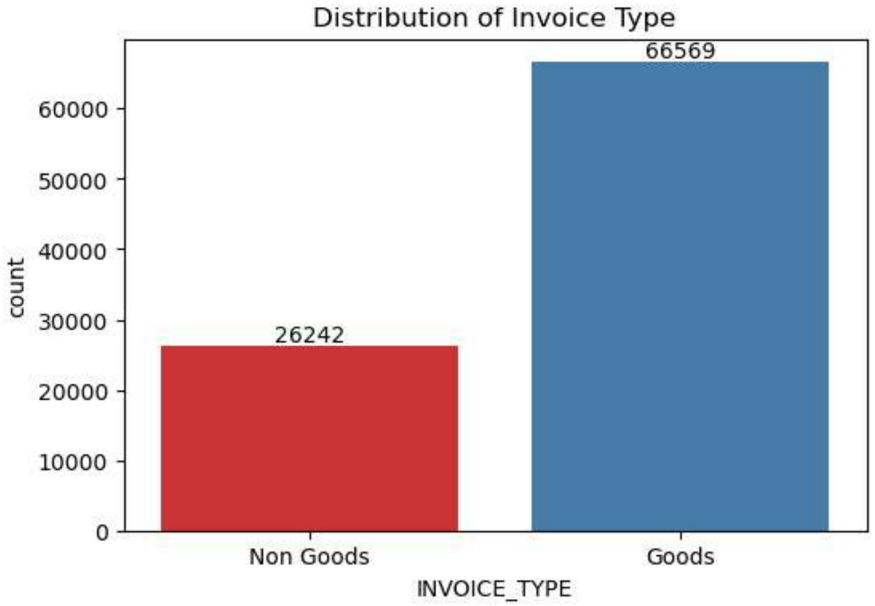


Fig 4.

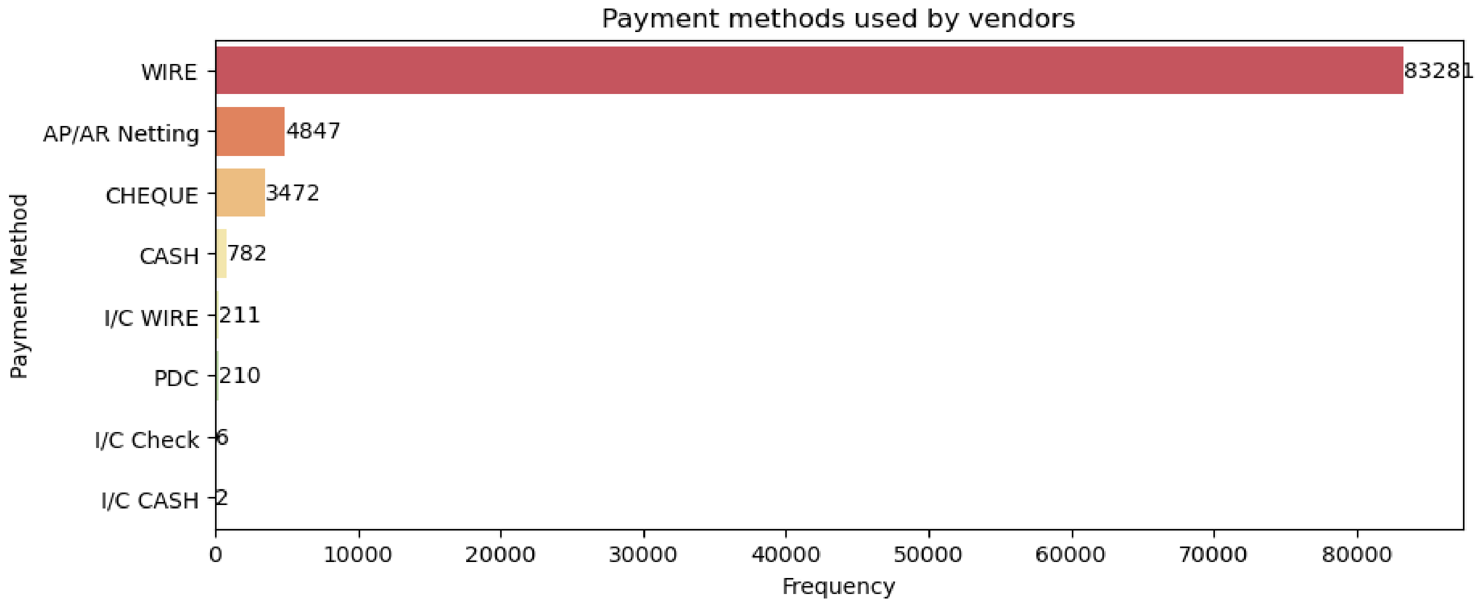


Fig 5.

- Goods type invoices comprise of the major share of invoices generated

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

Bivariate analysis

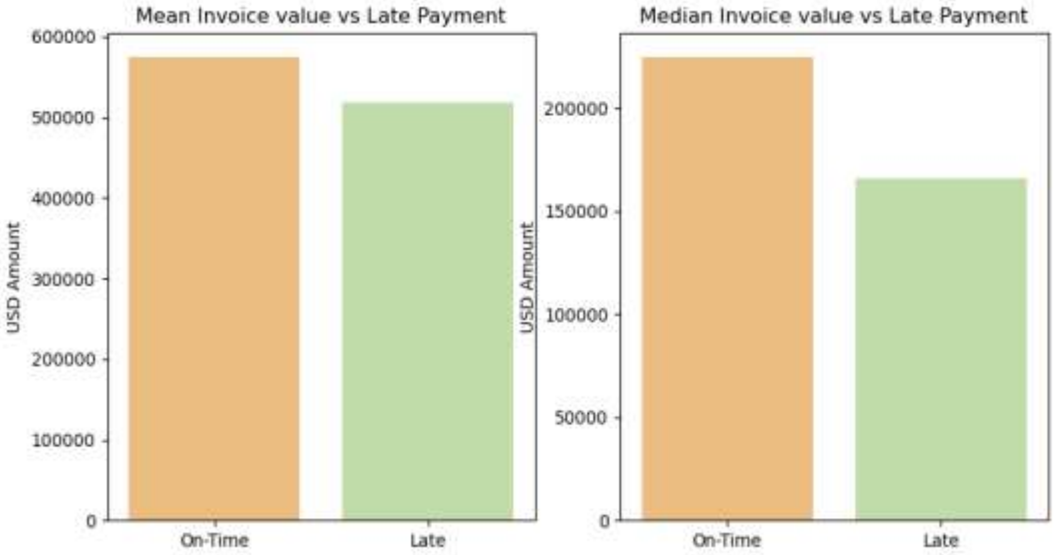


Fig 4.

- 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

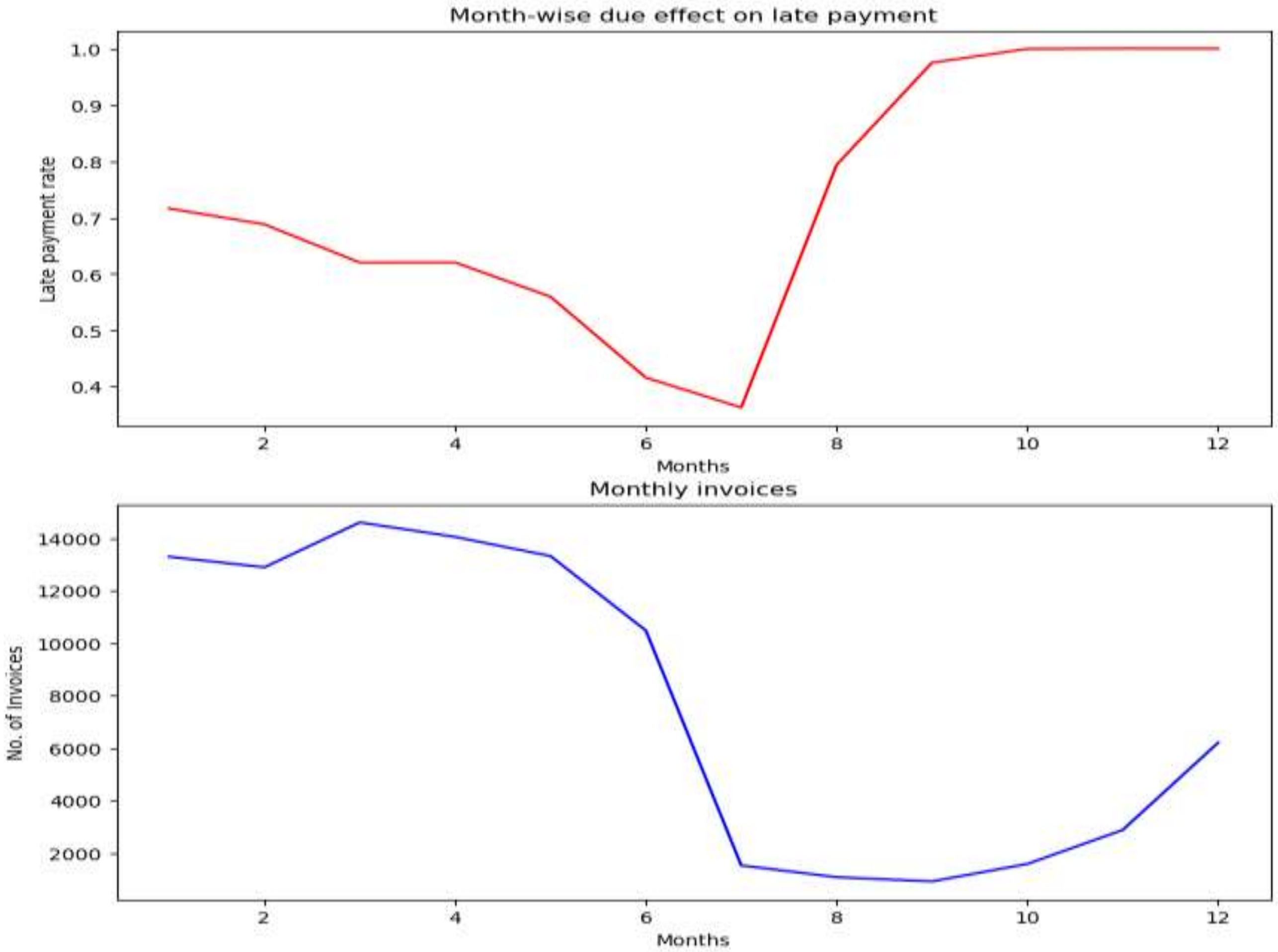


Fig 5.

- For the 3rd month, the number of invoices is the highest and late payment rate is comparatively lower than other months with large number of invoices.
- Month 7 has the very low late payment rate, this can be because of the fact that the number of invoices is also low.
- In the 2nd half of the year, the late payment increases steeply from 7th month onwards. The number of invoices are comparatively lower than the first half of the year.

Customer segmentation using K-means clustering

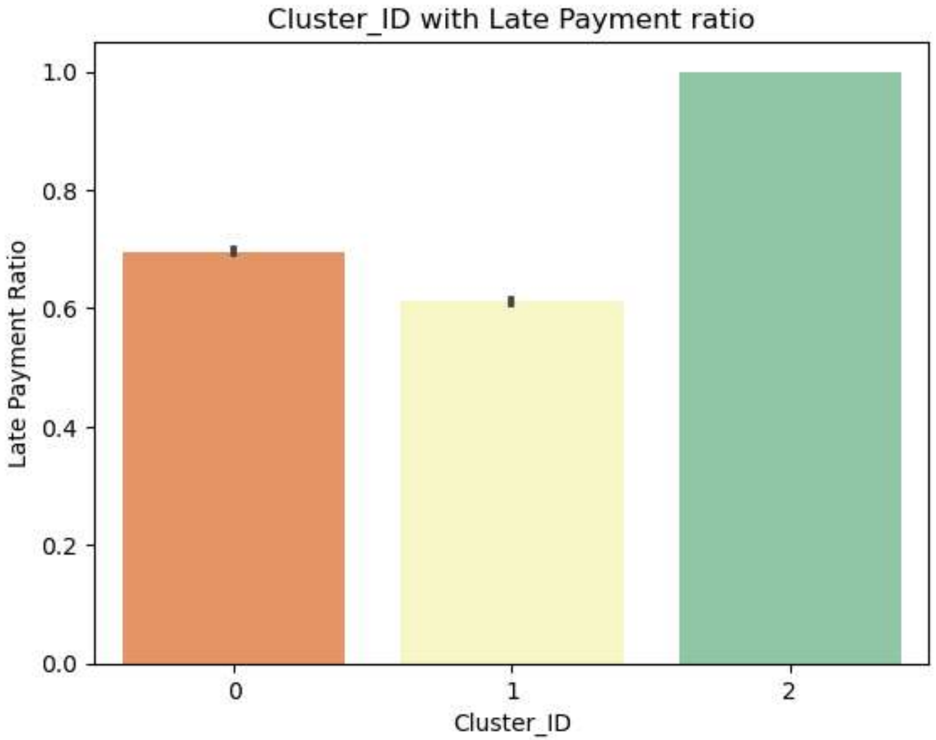


Fig 1.

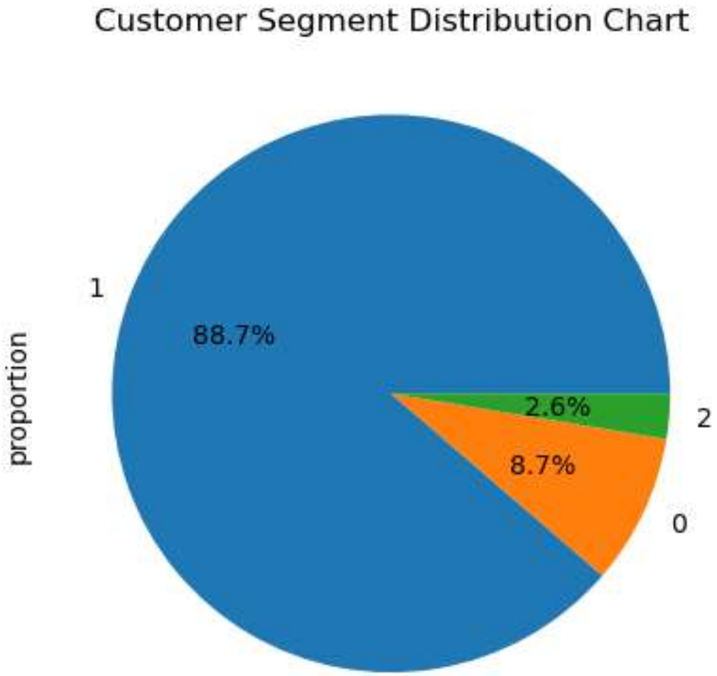
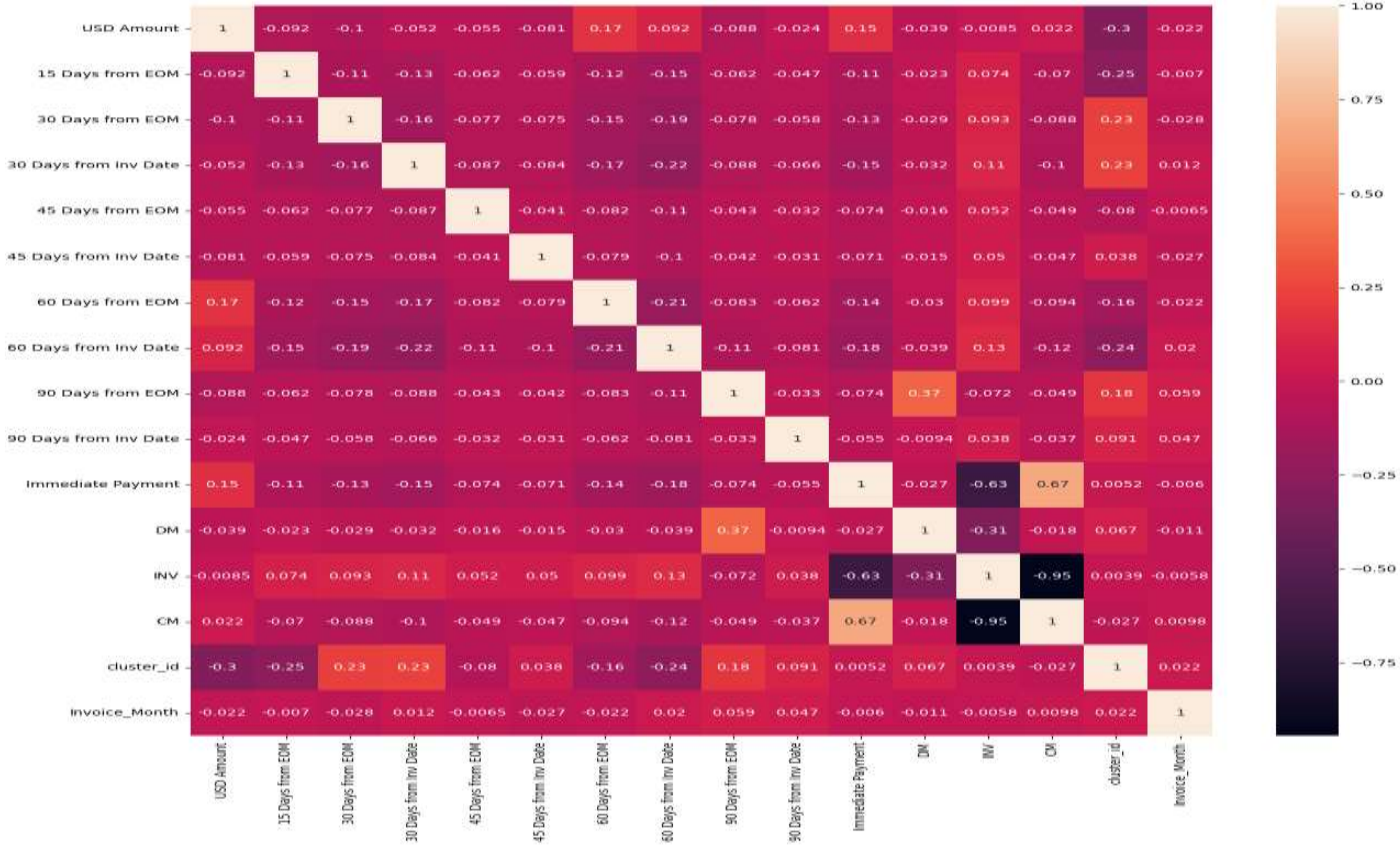


Fig 2.

- The category 2 were prolonged payers with least number of average days taken to pay and category 1 were early 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

- From the above we can see that Early customers comprise of 88.7% of customers whereas medium and prolonged payers are 11.3% in total.

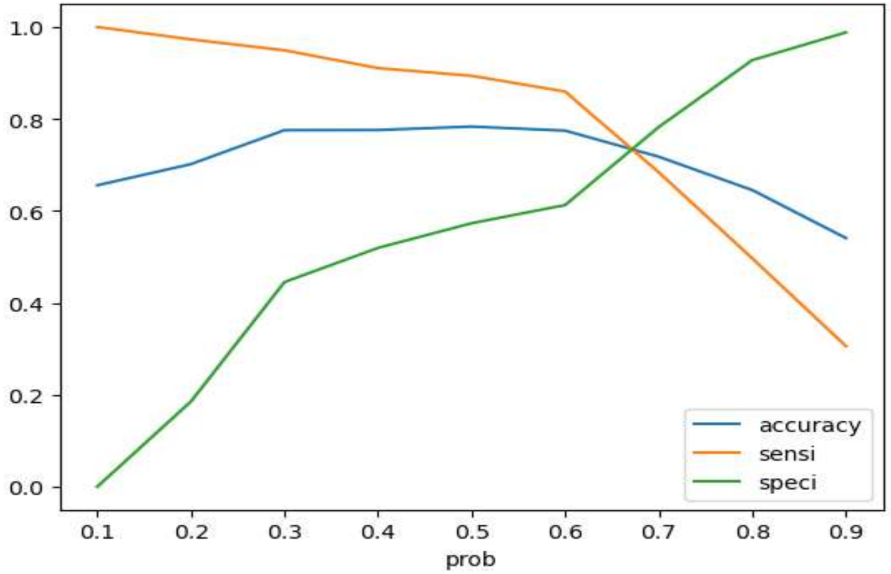
Model Building

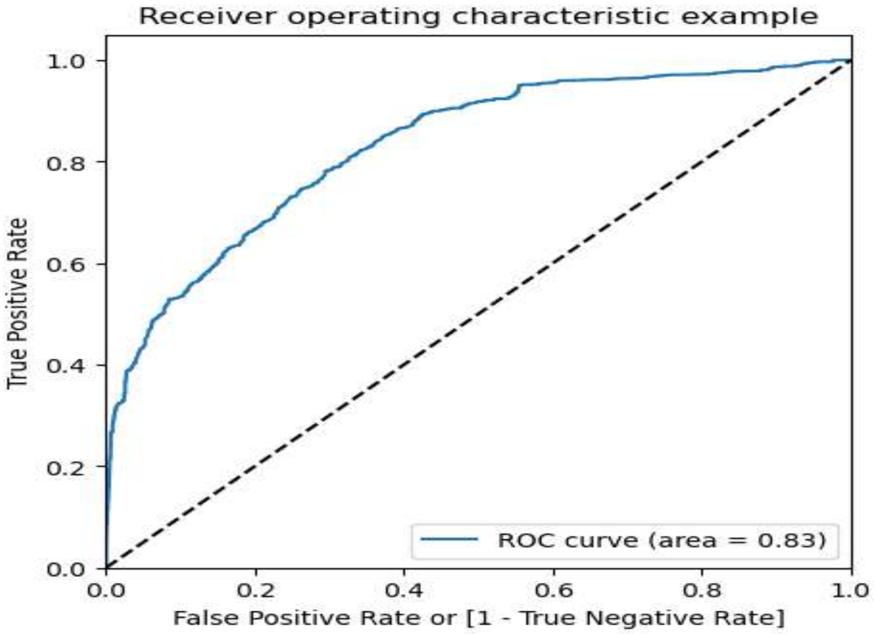


	USD Amount	15 Days from EOM	30 Days from EOM	30 Days from Inv Date	45 Days from EOM	45 Days from Inv Date	60 Days from EOM	60 Days from Inv Date	90 Days from EOM	90 Days from Inv Date	Immediate Payment	DM	INV	CM	cluster_id	Invoice_Month
USD Amount	1	-0.092	-0.1	-0.052	-0.055	-0.081	0.17	0.092	-0.088	-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	0.23	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	-0.08	-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	-0.24	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	0.091	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	0.0052	-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	0.067	-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	0.0039	-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	-0.027	0.0098
cluster_id	-0.3	-0.25	0.23	0.23	-0.08	0.038	-0.16	-0.24	0.18	0.091	0.0052	0.067	0.0039	-0.027	1	0.022
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	0.022	1

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

logistic regression

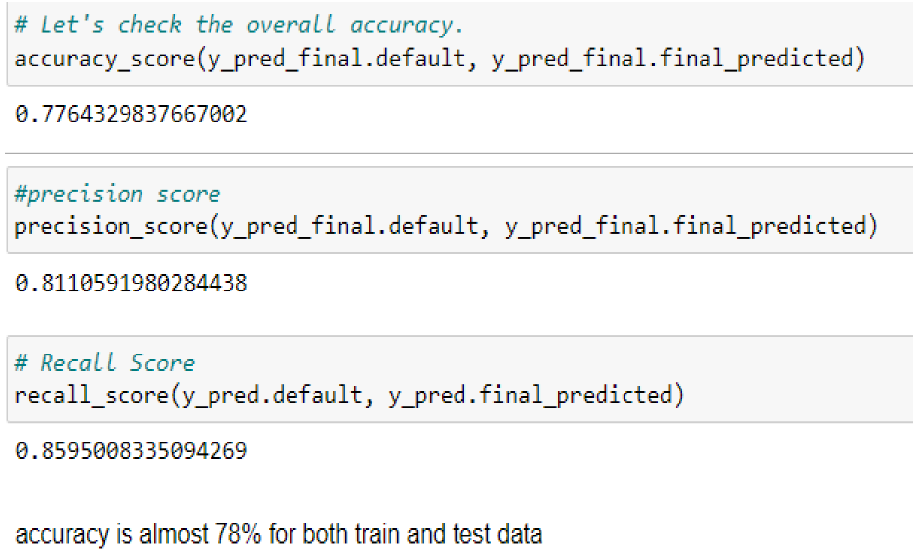




- The trade-off plot between accuracy, sensitivity and specificity revealed an optimum probability cutoff of ~0.6, which was used to further predict which transactions would result in delayed payments in the received payments dataset

- Logistic regression model formed after dropping multicollinearity and unnecessary variables resulted in remaining variables with acceptable p-value and VIF figures, hence retained the remaining features with no further feature elimination and a good ROC curve area of 0.83

Comparison between two models, logistic regression and random forests



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

```
0.7764329837667002
```

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

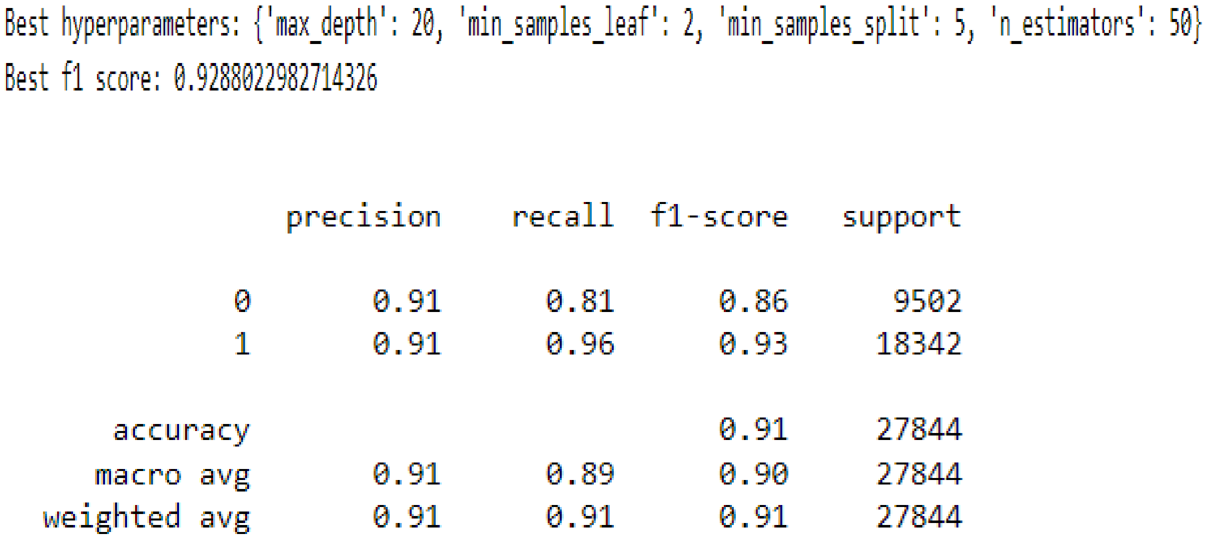
```
0.8110591980284438
```

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

```
0.8595008335094269
```

accuracy is almost 78% for both train and test data

Fig. 1 (Logistic Regression Metrics - Test Set)



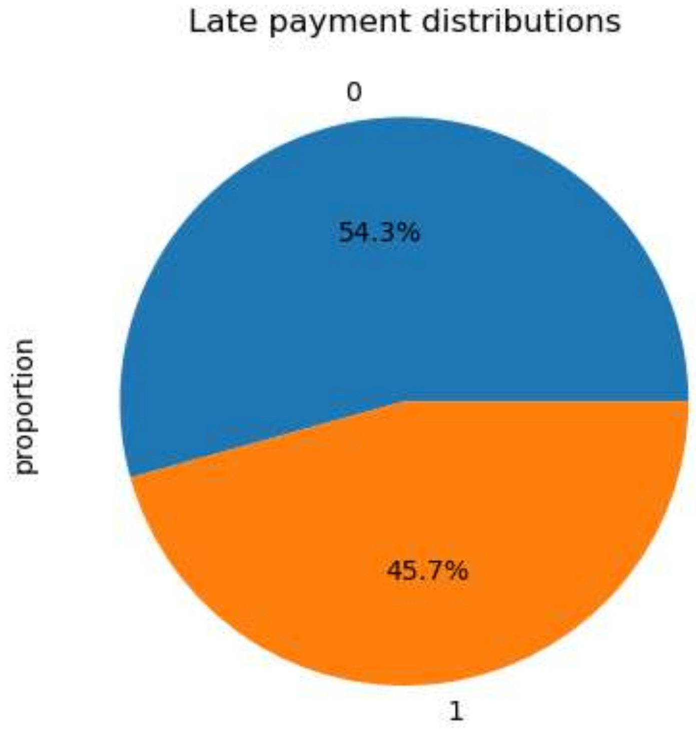
```
Best hyperparameters: {'max_depth': 20, 'min_samples_leaf': 2, 'min_samples_split': 5, 'n_estimators': 50}
Best f1 score: 0.9288022982714326
```

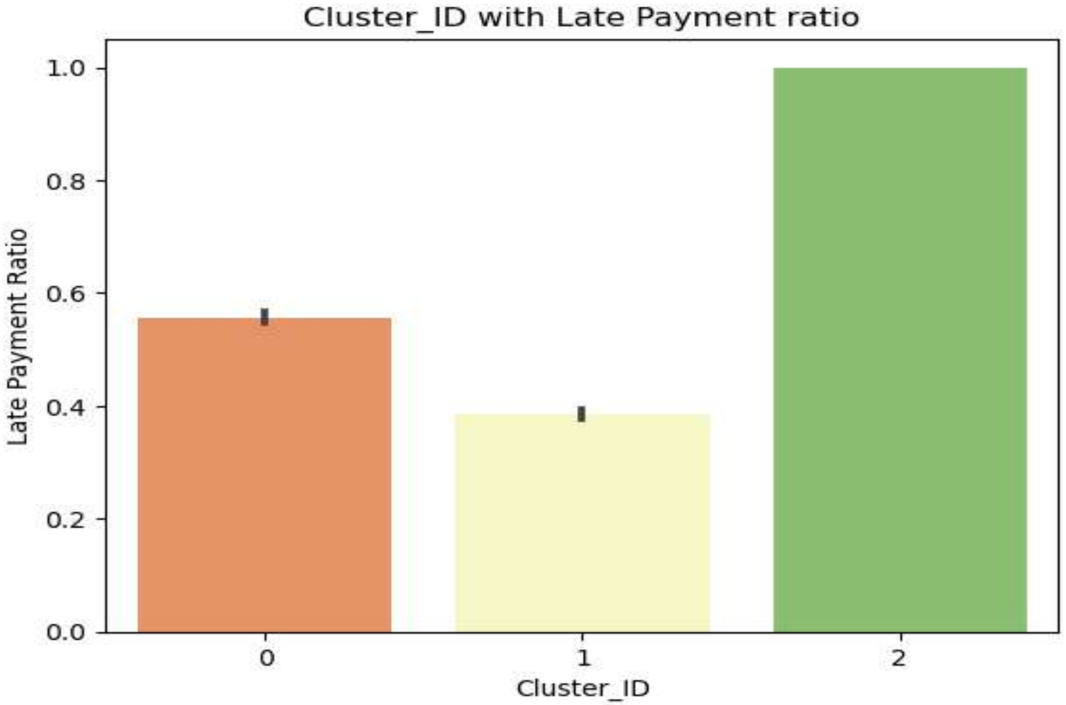
```
              precision    recall  f1-score   support

           0       0.91      0.81      0.86      9502
           1       0.91      0.96      0.93     18342

    accuracy                           0.91     27844
   macro avg       0.91      0.89      0.90     27844
weighted avg       0.91      0.91      0.91     27844
```

Fig. 2 (Random Forest Metrics - Test Set)

• It can be observed that the overall precision and recall scores of the Random forest model far exceeded the logistic regression model. Also, recall scores were more important in this case since it was important to increase the percentage prediction of late payers to be targeted • 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.

Final Prediction based on Random Forest model





observation: From the above pie chart, we can observe that 54.3% payments in the open invoice data with AGE value negative(indicating due date not crossed)

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

Recommendations

- Credit Note Payments observe the greatest delay rate compared to Debit Note or Invoice type invoice classes, hence company policies on payment collection could be made stricter around such invoice classes
- Goods type invoices had significantly greater payment delay rates than non-goods types and hence can be subjected to stricter payment policies
- Since lower value payments comprise of the majority of the transactions, also late payments are seen more on lower value payments, it is recommended to focus more on those. The company can apply penalties depending on billing amount, the lesser the bill, the greater the percentage of penalty on late payments. Ofcourse this has to be last resort
- Customer segments were clustered into three categories, viz., 0,2 and 1 which mean medium, prolonged and early payment duration respectively. It was found that customers in cluster 2 (prolonged days) had significantly greater delay rates than early and medium days of payment, hence cluster 2 customers should be paid extensive focus
- The following companies with the greatest probability and total & delayed payment counts should be first priority and should be focused on more due to such high probability rates ALSU Corp, SUND Corp, LVMH Corp, TRAF Corp, AMAT Corp, VENI Corp, MILK Corp, MAYC Corp, VIRT Corp, ROVE Corp