# E- Commerce and retail B2B case study

Submission by

Jayaram K R

Jayant Kashid

Jinal Gohil

# Problem identification and Business objectives

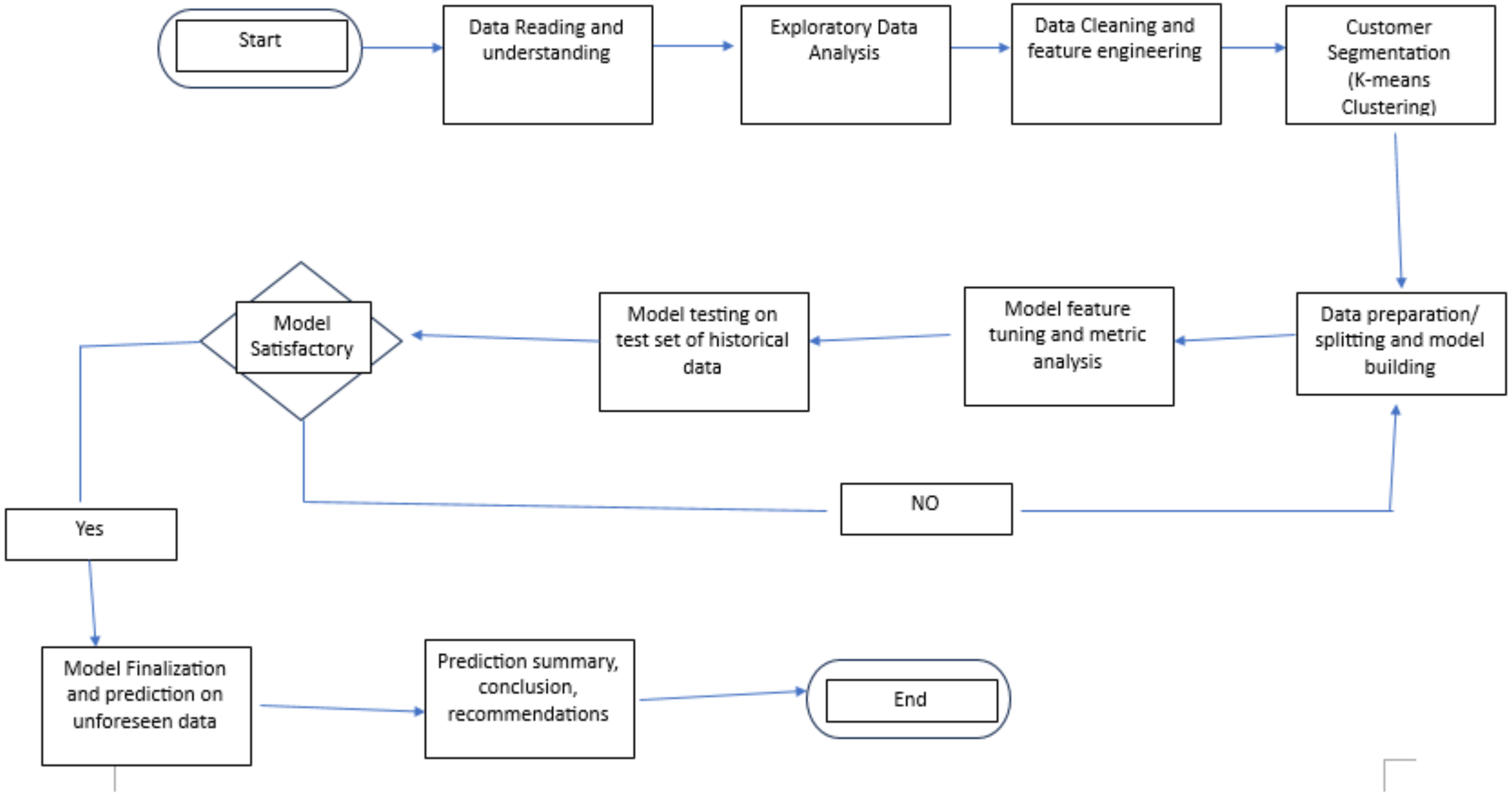
**Problem identification**

- 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 behaviour
- 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 activates

# Approach followed


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 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

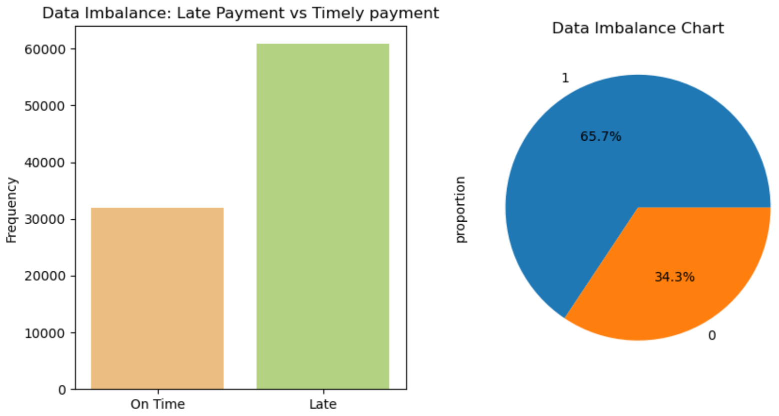


Fig 2

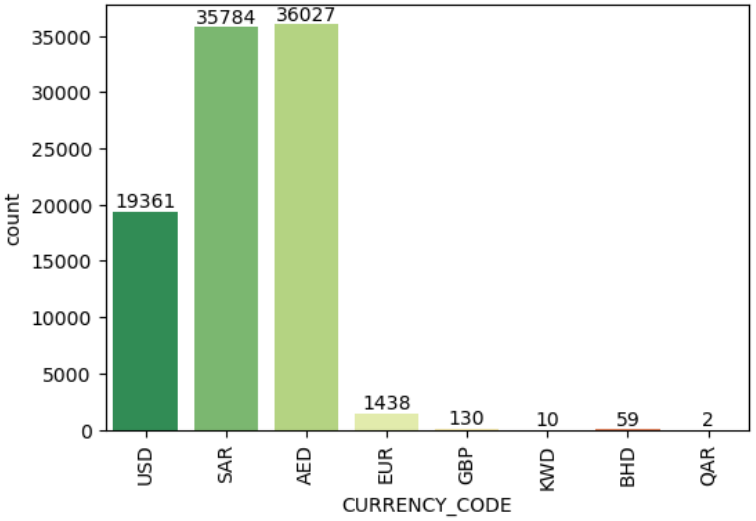


**Fig 1 and 2 Suggests**

- 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 3

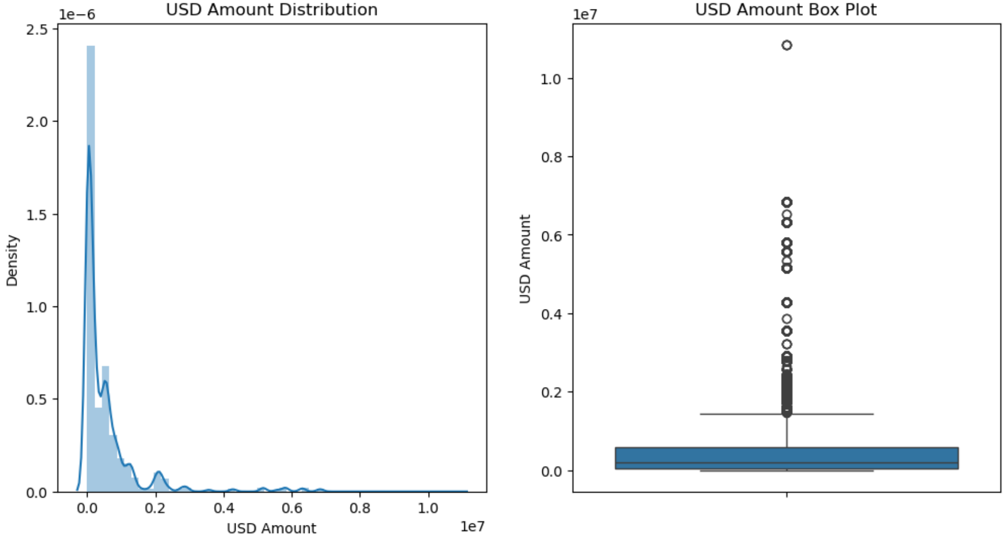


**Fig -3 suggests**

- 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

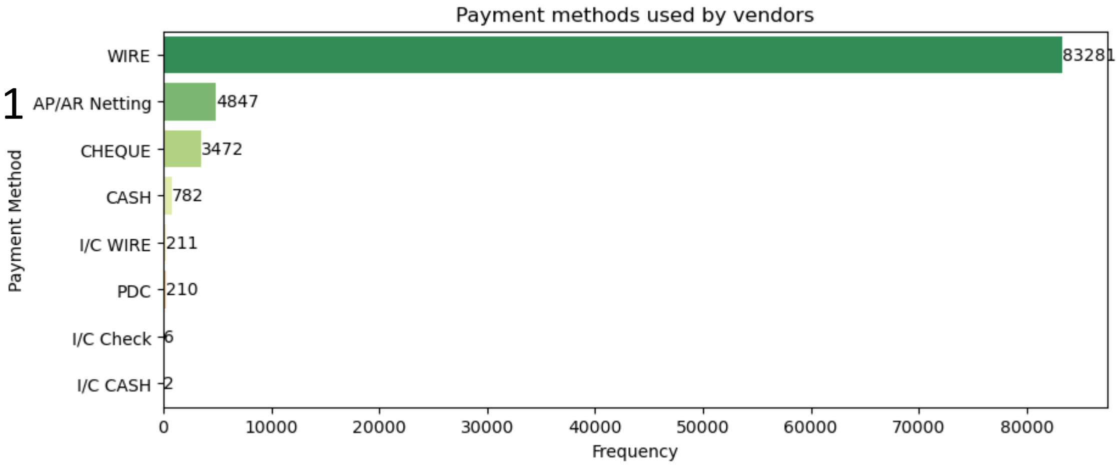
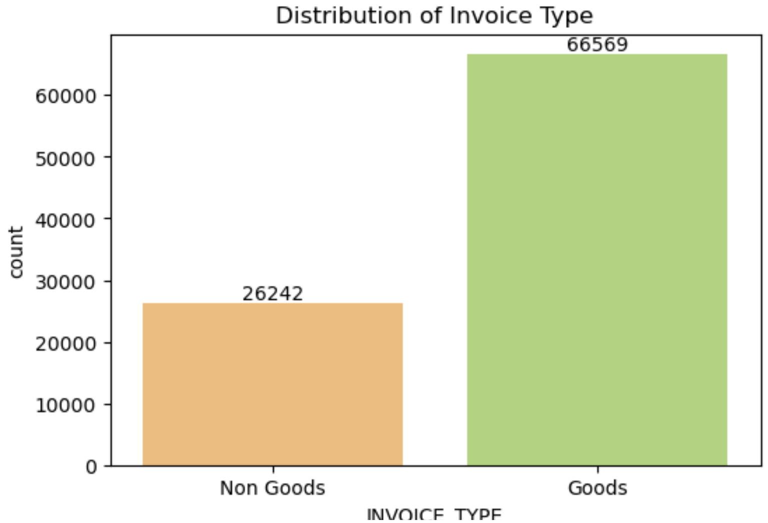


Fig 2



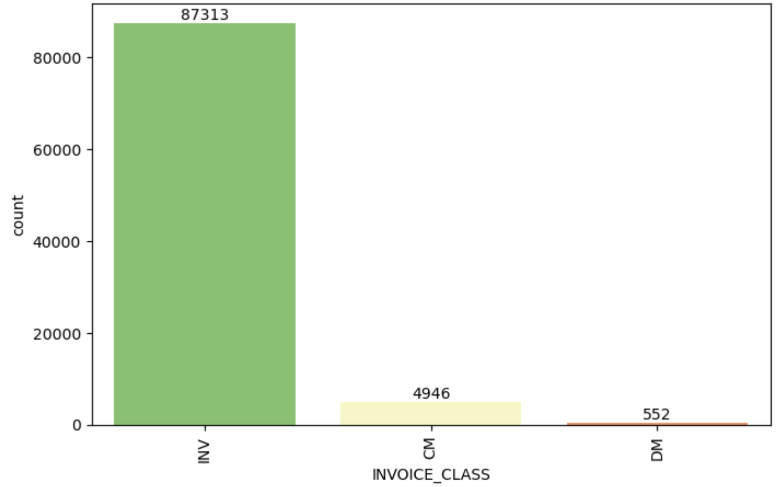
### Fig 1 Suggest that

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

### Fig 2 & 3 Suggests

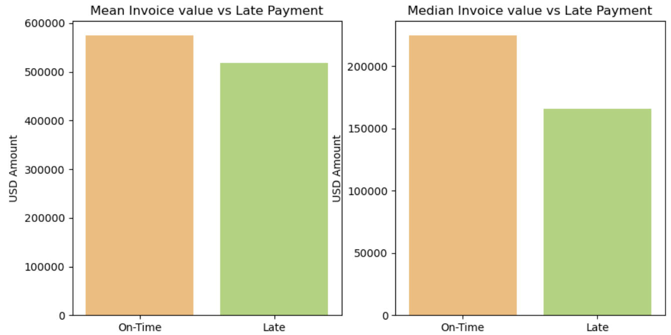
- Goods type invoices comprise of the major share of invoices generated
- The major invoice class is ‘Invoice’ with the rest having very low percentages of the share

Fig 3



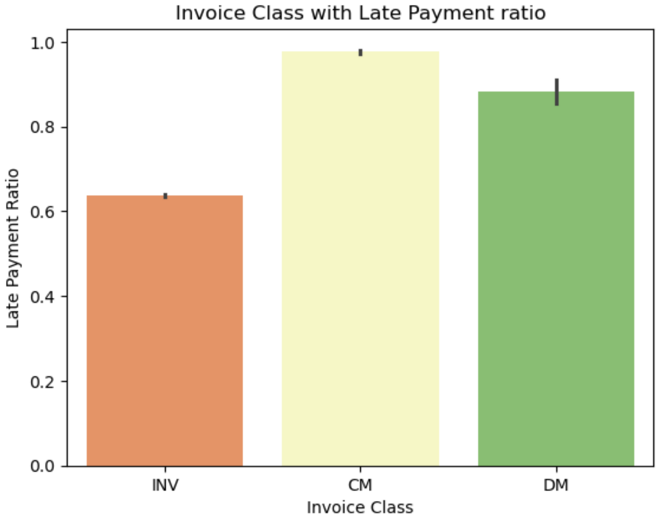
# Default payment type characteristics (Bivariate)

Fig 1



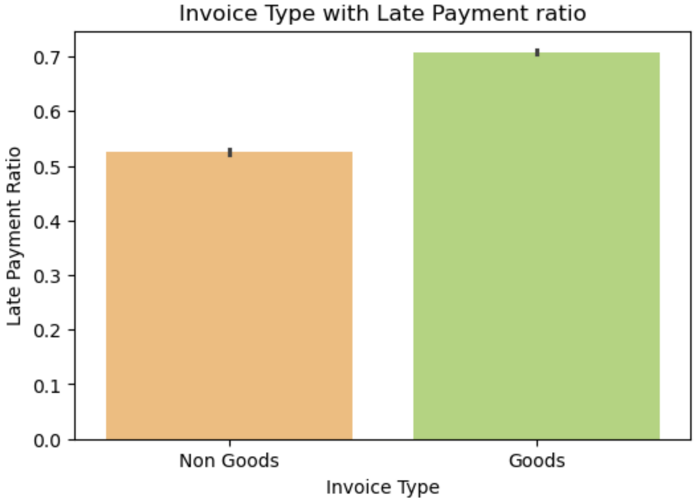
- From fig. 1, the mean and median of
the payment amount is higher for payers who pay on time than late, suggesting that higher value transactions show lesser delay risk than lower value transactions

Fig 2



- 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

Fig 3



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

# K- Means clustering –Customer segmentation

```
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 0.40252553894618825
For n_clusters=7, the silhouette score is 0.4069490441271981
For n_clusters=8, the silhouette score is 0.4151884768372497
```

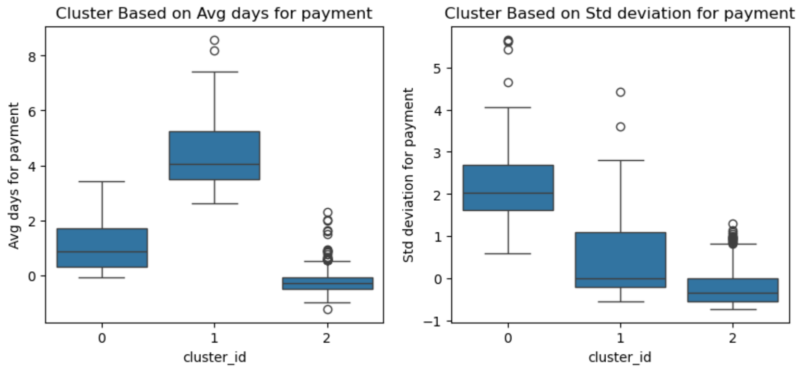


Fig 1

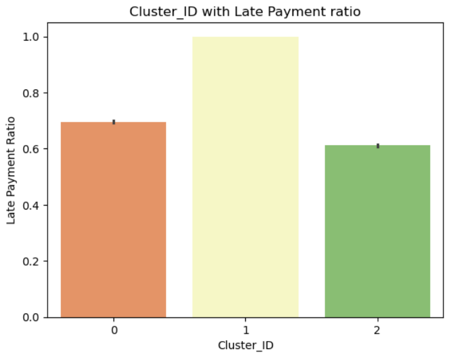
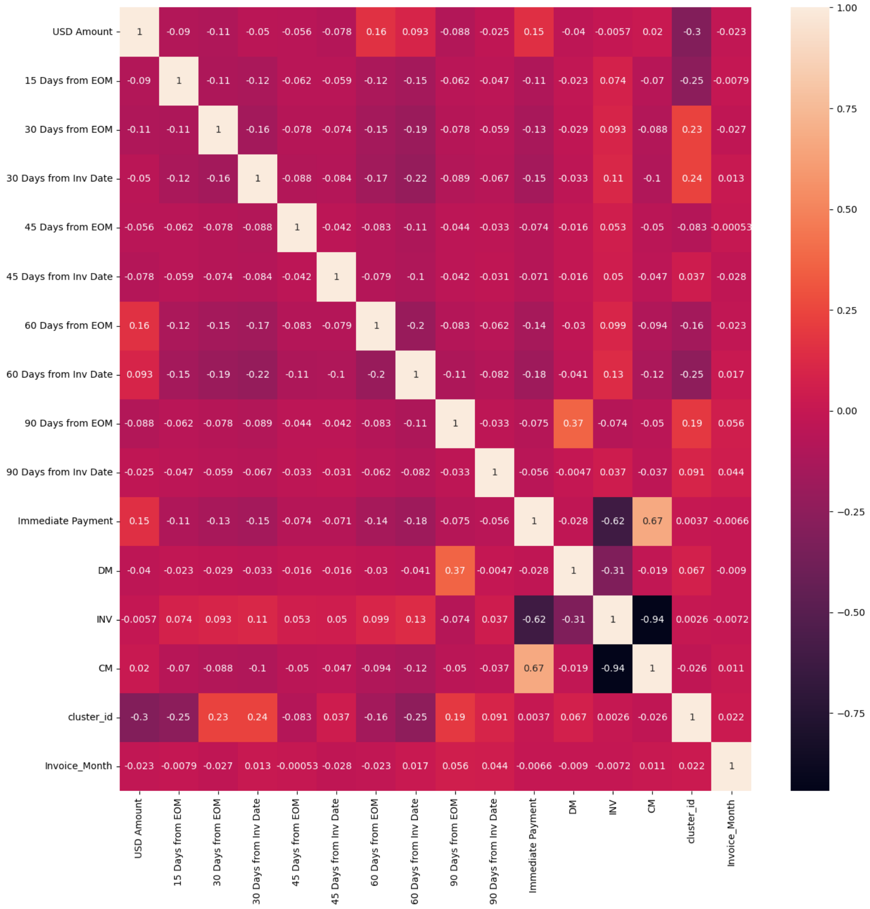


Fig 2

- Categorising customers to understand the payment behaviour was achieved through K-means clustering using the average and standard deviation number of days it took for the vendor to make payments
- Number of clusters are decided to be 3 as any increase in the cluster number results in significant reduction of silhouette score
- Category 2 paid early where as category 1 prolonged the payments, category 0 lie in between the other two and hence are 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



| | 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 | -0.09 | -0.11 | -0.05 | -0.056 | -0.078 | 0.16 | 0.093 | -0.088 | -0.025 | 0.15 | -0.04 | -0.0057 | 0.02 | -0.3 | -0.023 |
| 15 Days from EOM | -0.09 | 1 | -0.11 | -0.12 | -0.062 | -0.059 | -0.12 | -0.15 | -0.062 | -0.047 | -0.11 | -0.023 | 0.074 | -0.07 | -0.25 | -0.0079 |
| 30 Days from EOM | -0.11 | -0.11 | 1 | -0.16 | -0.078 | -0.074 | -0.15 | -0.19 | -0.078 | -0.059 | -0.13 | -0.029 | 0.093 | -0.088 | 0.23 | -0.027 |
| 30 Days from Inv Date | -0.05 | -0.12 | -0.16 | 1 | -0.088 | -0.084 | -0.17 | -0.22 | -0.089 | -0.067 | -0.15 | -0.033 | 0.11 | -0.1 | 0.24 | 0.013 |
| 45 Days from EOM | -0.056 | -0.062 | -0.078 | -0.088 | 1 | -0.042 | -0.083 | -0.11 | -0.044 | -0.033 | -0.074 | -0.016 | 0.053 | -0.05 | -0.083 | -0.00053 |
| 45 Days from Inv Date | -0.078 | -0.059 | -0.074 | -0.084 | -0.042 | 1 | -0.079 | -0.1 | -0.042 | -0.031 | -0.071 | -0.016 | 0.05 | -0.047 | 0.037 | -0.028 |
| 60 Days from EOM | 0.16 | -0.12 | -0.15 | -0.17 | -0.083 | -0.079 | 1 | -0.2 | -0.083 | -0.062 | -0.14 | -0.03 | 0.099 | -0.094 | -0.16 | -0.023 |
| 60 Days from Inv Date | 0.093 | -0.15 | -0.19 | -0.22 | -0.11 | -0.1 | -0.2 | 1 | -0.11 | -0.082 | -0.18 | -0.041 | 0.13 | -0.12 | -0.25 | 0.017 |
| 90 Days from EOM | -0.088 | -0.062 | -0.078 | -0.089 | -0.044 | -0.042 | -0.083 | -0.11 | 1 | -0.033 | -0.075 | 0.37 | -0.074 | -0.05 | 0.19 | 0.056 |
| 90 Days from Inv Date | -0.025 | -0.047 | -0.059 | -0.067 | -0.033 | -0.031 | -0.062 | -0.082 | -0.033 | 1 | -0.056 | -0.0047 | 0.037 | -0.037 | 0.091 | 0.044 |
| Immediate Payment | 0.15 | -0.11 | -0.13 | -0.15 | -0.074 | -0.071 | -0.14 | -0.18 | -0.075 | -0.056 | 1 | -0.028 | -0.62 | 0.67 | 0.0037 | -0.0066 |
| DM | -0.04 | -0.023 | -0.029 | -0.033 | -0.016 | -0.016 | -0.03 | -0.041 | 0.37 | -0.0047 | -0.028 | 1 | -0.31 | -0.019 | 0.067 | -0.009 |
| INV | -0.0057 | 0.074 | 0.093 | 0.11 | 0.053 | 0.05 | 0.099 | 0.13 | -0.074 | 0.037 | -0.62 | -0.31 | 1 | -0.94 | 0.0026 | -0.0072 |
| CM | 0.02 | -0.07 | -0.088 | -0.1 | -0.05 | -0.047 | -0.094 | -0.12 | -0.05 | -0.037 | 0.67 | -0.019 | -0.94 | 1 | -0.026 | 0.011 |
| cluster_id | -0.3 | -0.25 | 0.23 | 0.24 | -0.083 | 0.037 | -0.16 | -0.25 | 0.19 | 0.091 | 0.0037 | 0.067 | 0.0026 | -0.026 | 1 | 0.022 |
| Invoice_Month | -0.023 | -0.0079 | -0.027 | 0.013 | -0.00053 | -0.028 | -0.023 | 0.017 | 0.056 | 0.044 | -0.0066 | -0.009 | -0.0072 | 0.011 | 0.022 | 1 |

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

# Model comparison – Logistic Regression and Random Forests

Fig 1

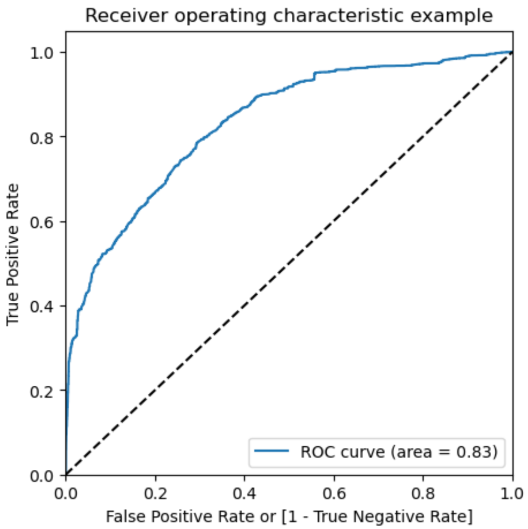


Fig 1

Logistic regression model subsequent to dropping multicollinearity resulted in acceptable p-value and VIF, hence retained the rest of the features with no further feature elimination and obtained a good ROC curve with AOC 0f 0.83

Fig 2

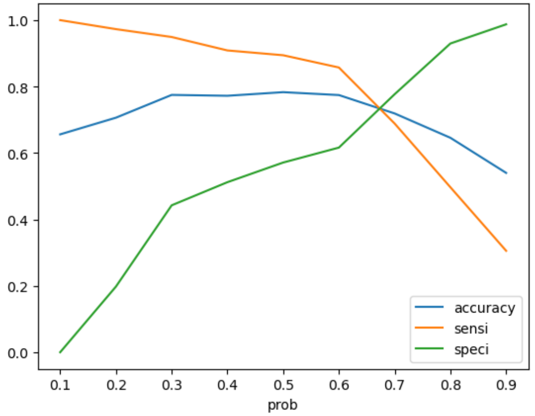
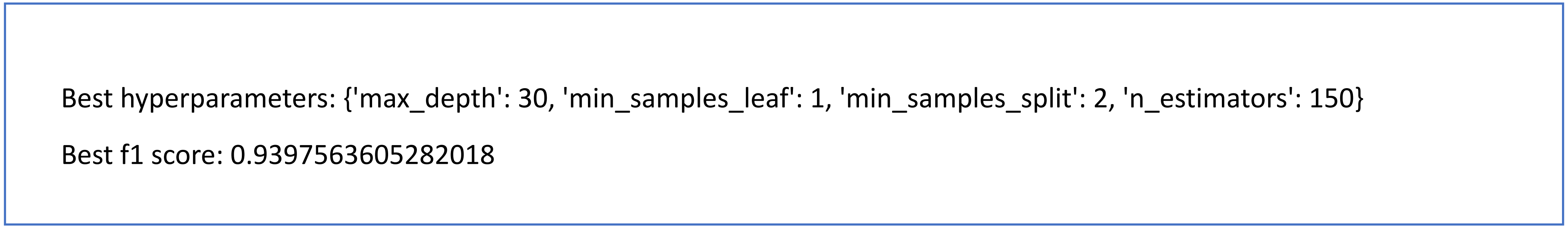


Fig 2

Accuracy, sensitivity and specificity trade off revealed an optimum probability cut off of ~0.6. This was used to further predict which transaction would result in delayed payments in the received payments dataset

# Model comparison – Logistic Regression and Random Forests

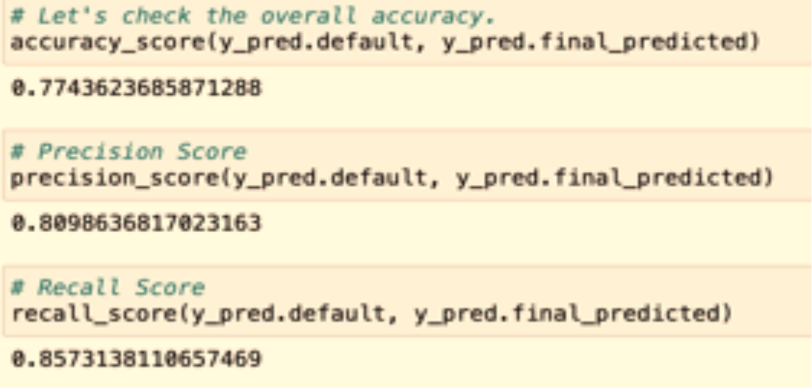
- Random forest was built with the same parameters as the Logistic regression with hyper parameter tuning resulting in the below mentioned parameters



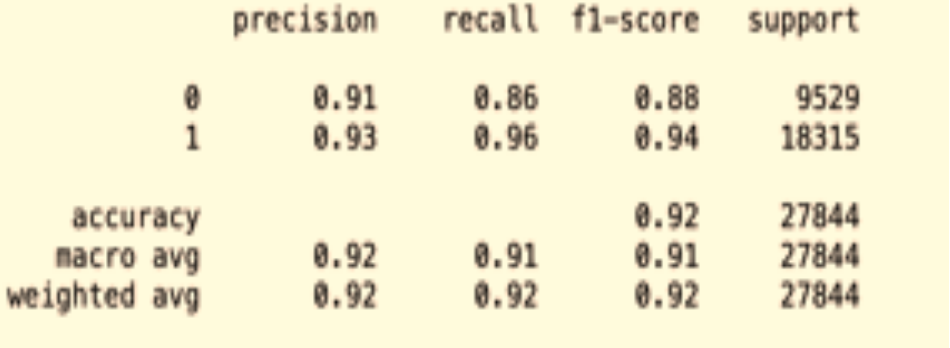
```
Best hyperparameters: {'max_depth': 30, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 150}

Best f1 score: 0.9397563605282018
```

- Random forest model was built using the above parameters, whose metrics were compared to the logistic regression model and the final model was arrived

# Comparison result – Random forest better than Logistic Regression



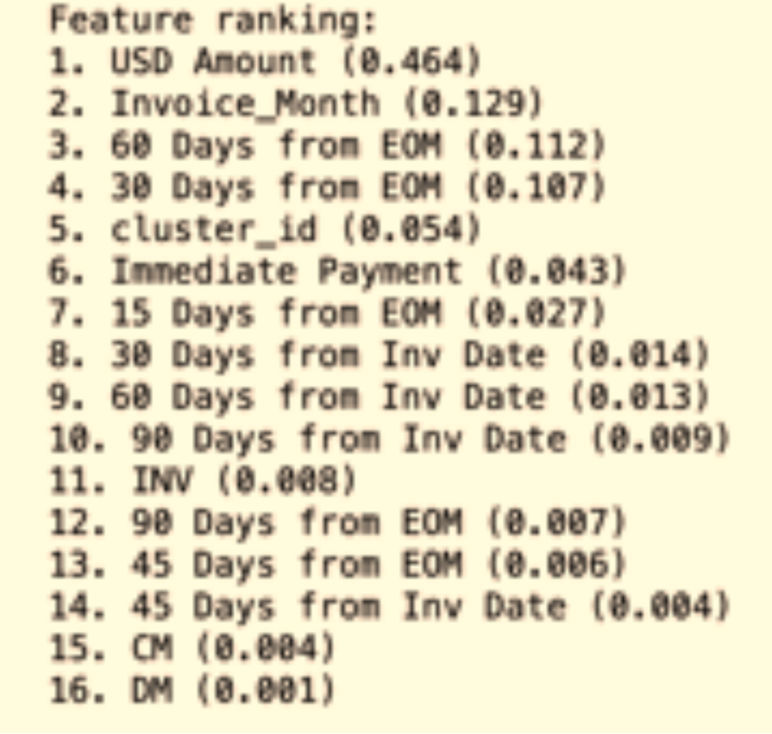
```
# Let's check the overall accuracy.
accuracy_score(y_pred.default, y_pred.final_predicted)

0.7743623685871288

# Precision Score
precision_score(y_pred.default, y_pred.final_predicted)

0.8098636817023163

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

0.8573138110657469
```



```
              precision    recall  f1-score   support

           0       0.91      0.86      0.88      9529
           1       0.93      0.96      0.94     18315

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

- 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 finalised to be the model of choice and go forward with predictions

# Feature Ratings – Random Forests



```
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 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

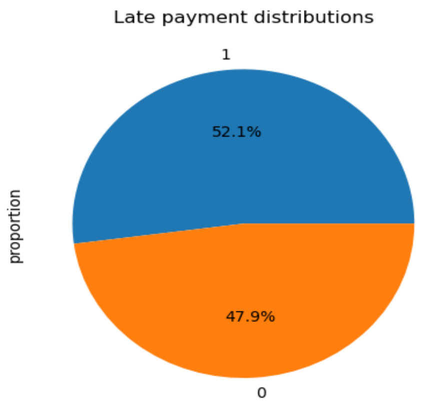
  - 60 Days from EOM (Payment Term variable)

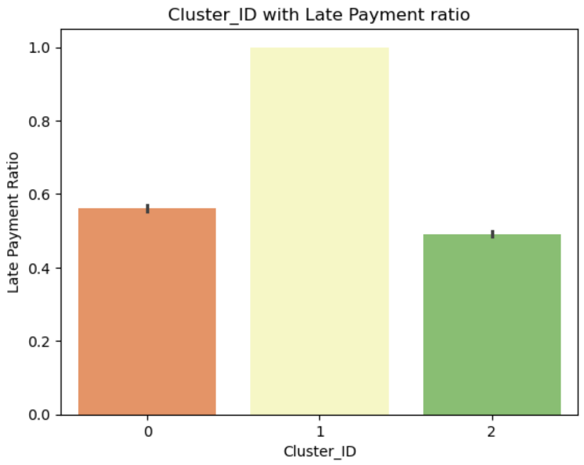
  - 30 Days from EOM (Payment Term variable)

  -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

# 52.1% payments predicted to be delayed as per open invoice data

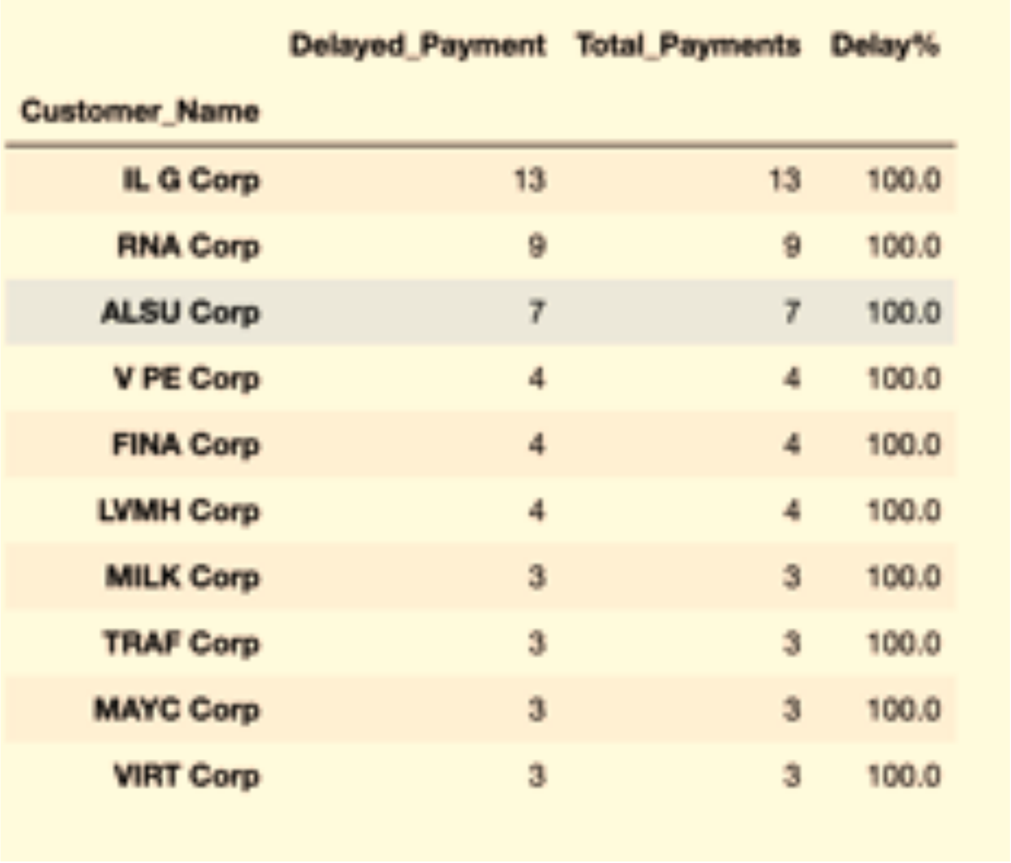




- Predictions made by the final model suggests that there is a probable 52.1% transactions where payment delay can be expected, which can cause a shocking lag to business operations

- 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 highest delay probabilities



| | Delayed_Payment | Total_Payments | Delay% |
|---|---|---|---|
| Customer_Name | | | |
| 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 |

- 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

**From our clustering analysis we can make the following inference-**

- 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. Of course this has to be last resort

- Customer segments were clustered into three categories, viz., 0,1 and 2 which mean medium, prolonged and early payment duration respectively. It was found that customers in cluster 1 (prolonged days) had significantly greater delay rates than early and medium days of payment, hence cluster 1 customers should be paid extensive focus

- The above 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