E-Commerce & Retail B2B Case Study

Mini Saxena
Prudhvi Mounika Mythri
Neha Manjrekar

Problem Statement & Business Goals

Schuster is a multinational retail company dealing in sports goods and accessories. Schuster conducts significant business with hundreds of its vendors, with whom it has credit arrangements. Unfortunately, not all vendors respect credit terms and some of them tend to make payments late. Schuster levies heavy late payment fees, although this procedure is not beneficial to either party in a long-term business relationship. The company has some employees who keep chasing vendors to get the payment on time; this procedure nevertheless also results in non-value-added activities, loss of time and financial impact. Schuster would thus try to understand its customers' payment behaviour and predict the likelihood of late payments against open invoices.

Goals:

- Schuster would like to better understand the customers' payment behaviour based on their past payment patterns (customer segmentation).
- Using historical information, it wants to be able to predict the likelihood of delayed payment against open invoices from its customers.
- It wants to use this information so that collectors can prioritise their work in following up with customers beforehand to get the payments on time.

Payment Process at Schuster

- Every time a transaction of goods takes place with a vendor □ the accounting team raises an invoice and shares it with the vendor.
- This invoice contains the details of the goods, the invoice value, the creation date and the payment due date based on the credit terms as per the contract.
- Business with these vendors occurs quite frequently. Hence, there are always multiple invoices associated with each vendor at any given time.

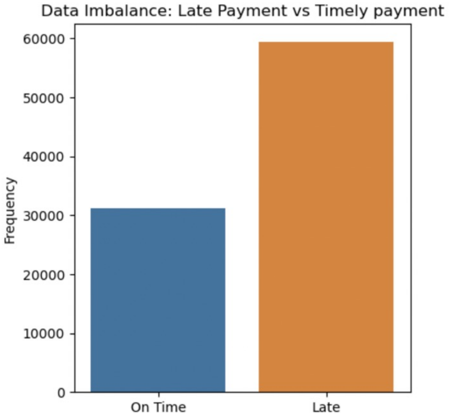
Methodology

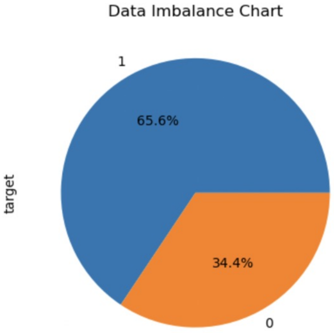
- Reading and Understanding the data
- Data Cleaning and Preparation
- Handling Missing Values
- Handling Outliers
- Data Cleaning
- Exploratory Data Analysis
- Univariate Analysis
- Data Imbalance
- Bivariate Analysis
- Feature Engineering
- Dummy Variable creation
- Clustering

- Model building
- Train- Test Split
- Feature Scaling
- Logistic Regression
- Plotting the ROC curve
- Random Forest
- Hyperparameter Tuning
- model evaluation
- Checking feature importance
- Final Prediction based on Random Forest model

Dataset Properties

- Dataset has total 16 columns and 93937rows.
- There are total 12 object values, 1 integer value int64 & 3 float values.
- Outliers
- Null values : The "RECEIPT_DOC_NO" col has 0.03% null values. Dropping the columns as it is not important for model building.
- Target variable : Derived by checking whether the payment receipt date falls within, or after the due date.
- Data Imbalance : 65% of the total payment was late payment.

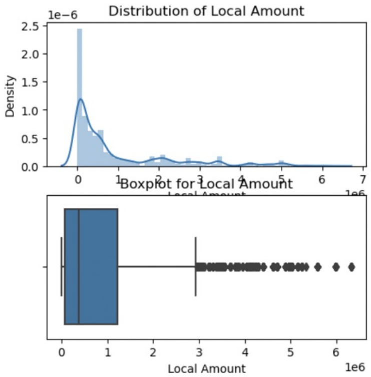


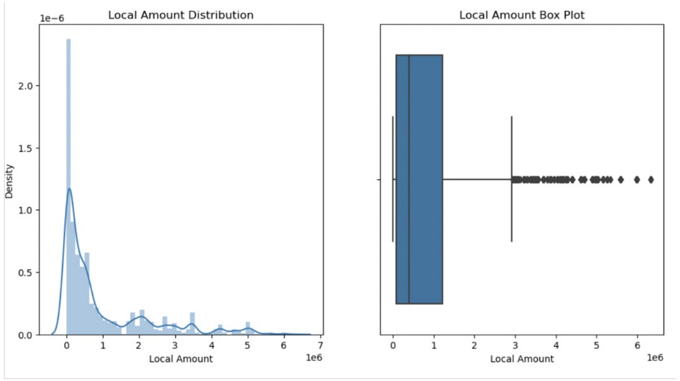


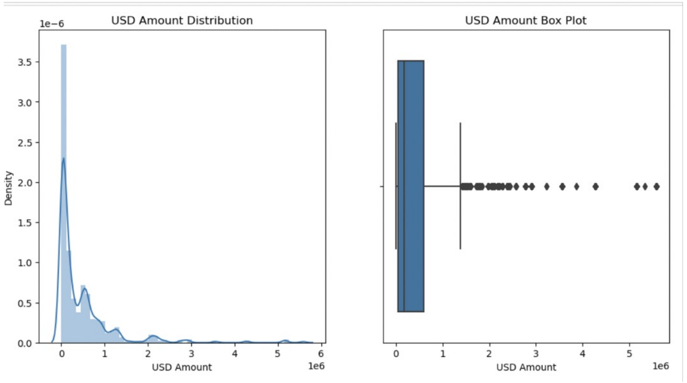
EDA

Received_Payments_Data.csv
Open_Invoice_data.csv

Visualizing the distribution of local & USD amount

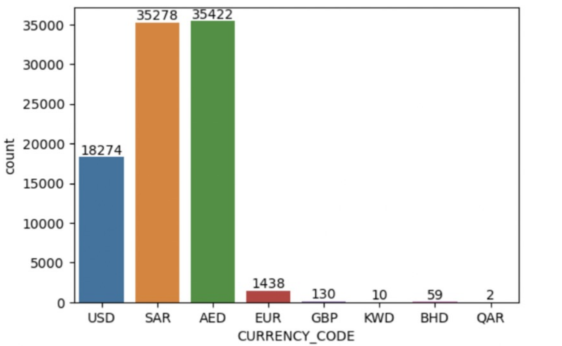


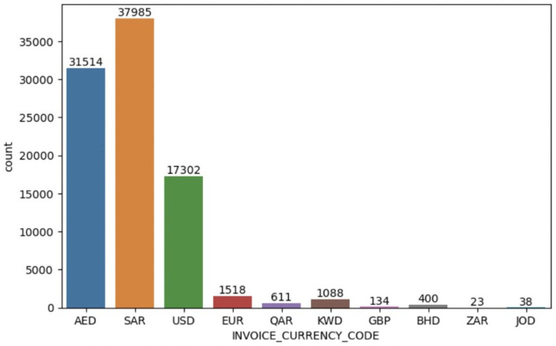


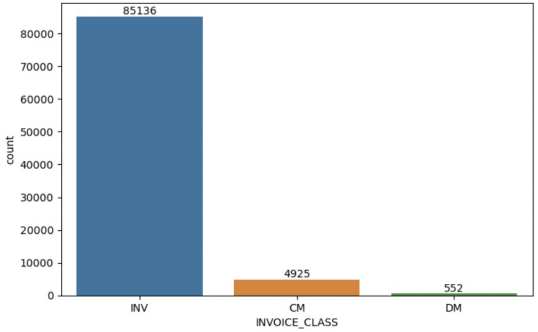


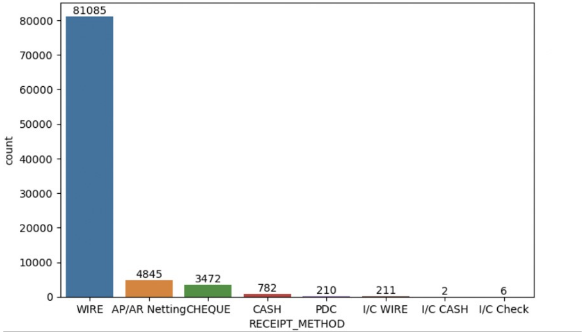
- The transaction values seem to lie between a range of $1 and $2m and are most frequent below ~$1.75m

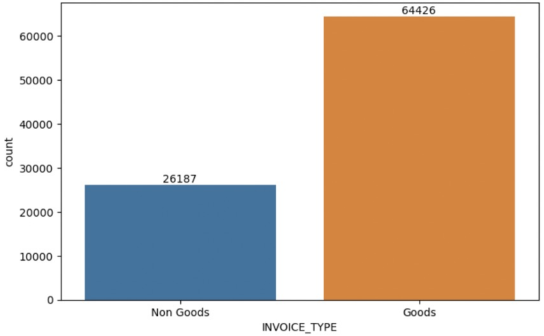
Analyzing Numeric Variables





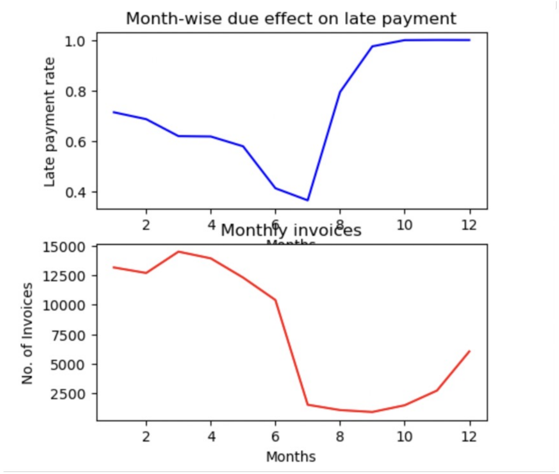


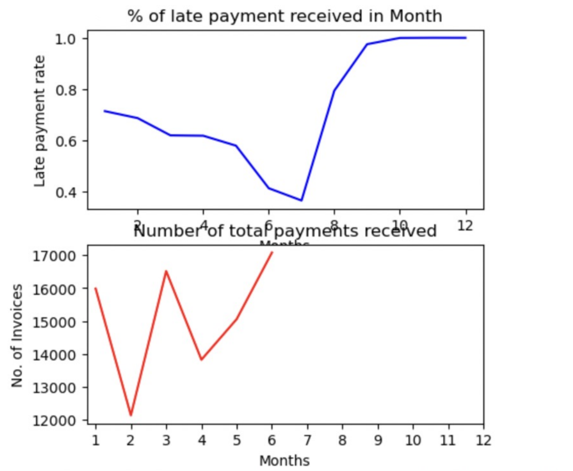


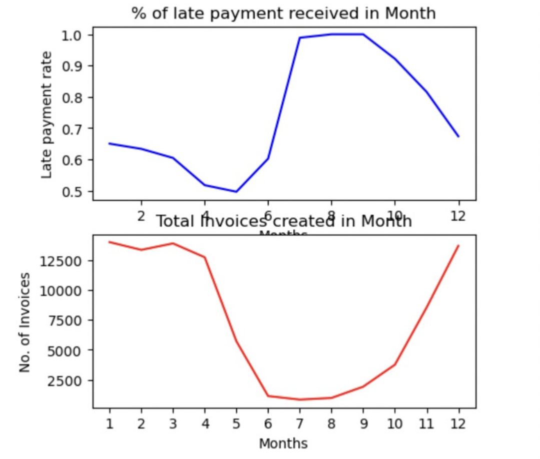


- Currency used for bill payments are mostly USD, SAR or AED.
- INV has the maximum number of bills in INVOICE_CLASS column.
- Majority of the invoices are raised for Goods.
- The most preferred payment method for bill payment is WIRE.

Effect of month on late payment rate

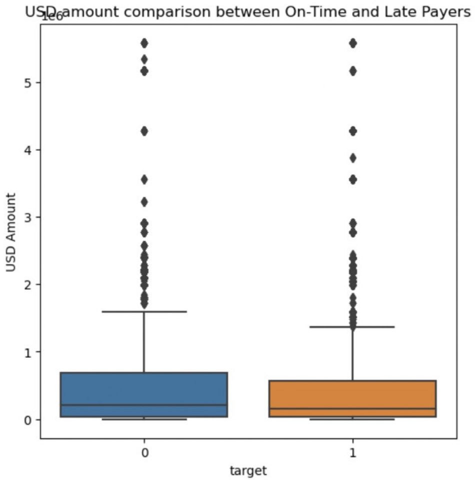


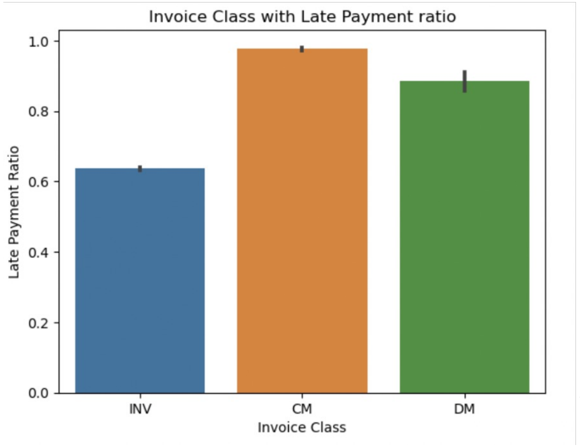


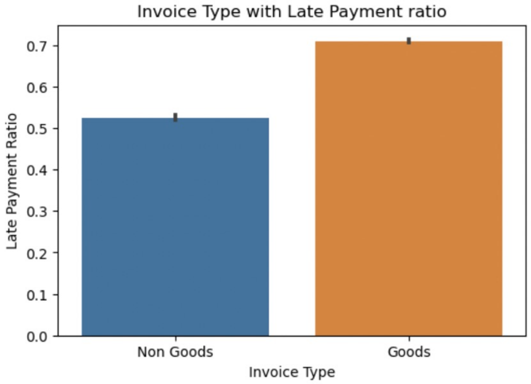


- For the 3rd month, the number of invoices is the highest and late payment rate is comparatively lower than other months with large number of invoices.
- Month 7 showed very low late payment rate this might be because of the number of invoices is also low.
- In the 2nd half of the year, the late payment increases steeply from 7th month onwards. The number of invoices are comparatively lower than the first half of the year.
- No payment received against any invoices from 7th month onwards.
- Late payment rate is decreases from 1st to 5th month.
- For the months 7, 8 and 9, the late payment rate is very high.

Analyzing Late Payers

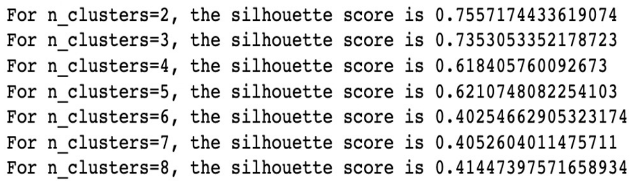






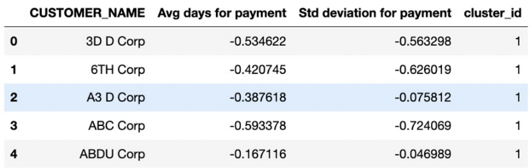
- The Median of on time and late payer is more or less the same.
- Credit memo has the higher late payment ratio.
- Majority of the invoice type was for Goods.

Clustering



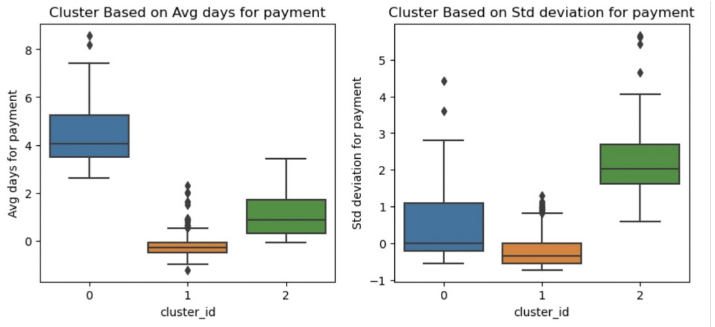
```
For n_clusters=2, the silhouette score is 0.7557174433619074
For n_clusters=3, the silhouette score is 0.7353053352178723
For n_clusters=4, the silhouette score is 0.618405760092673
For n_clusters=5, the silhouette score is 0.6210748082254103
For n_clusters=6, the silhouette score is 0.40254662905323174
For n_clusters=7, the silhouette score is 0.4052604011475711
For n_clusters=8, the silhouette score is 0.41447397571658934
```

- From the above results, we can see that for 3 clusters, the silhouette score is decent. Hence selecting n_clusters as 3.



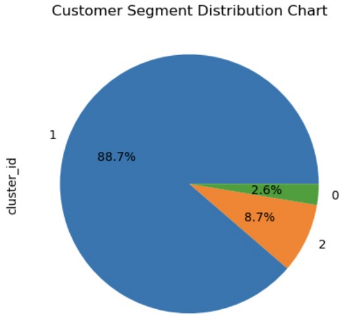
	CUSTOMER_NAME	Avg days for payment	Std deviation for payment	cluster_id
0	3D D Corp	-0.534622	-0.563298	1
1	6TH Corp	-0.420745	-0.626019	1
2	A3 D Corp	-0.387618	-0.075812	1
3	ABC Corp	-0.593378	-0.724069	1
4	ABDU Corp	-0.167116	-0.046989	1

- Assigning labels to the cust_seg dataframe



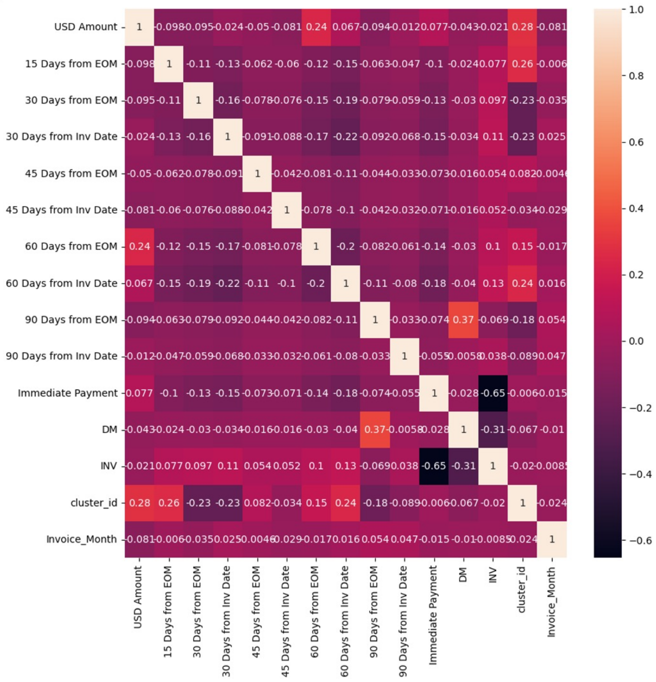
- '0' Cluster -- Prolonged Invoice Payment
- '1' Cluster -- Early Invoice Payment
- '2' Cluster -- Medium Invoice Payment

- Median for prolonged invoice payment is higher than Early invoice payment .

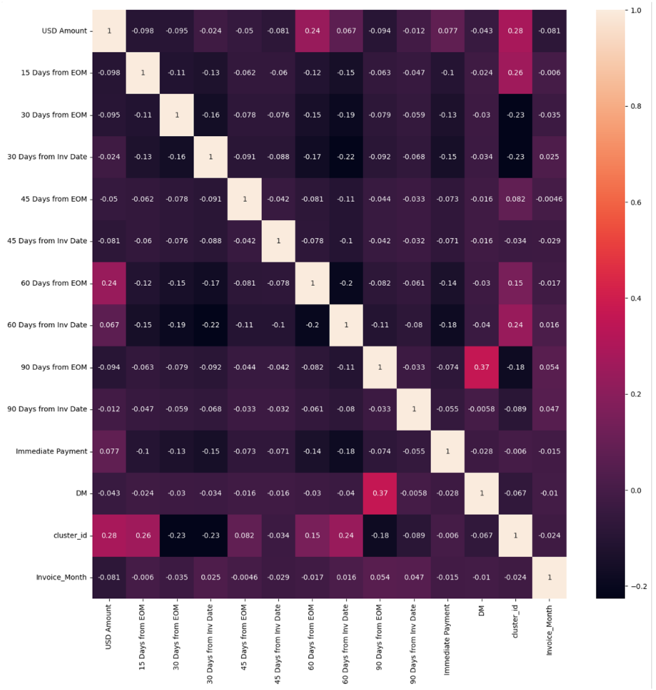


- From the above we can see that Early customers comprise of 88.7% of customers whereas medium and prolonged payers are 11.3% in total

Heat map of X_train dataset



Before dropping dropping INV - INV & Immediate Payment has high multicollinearity

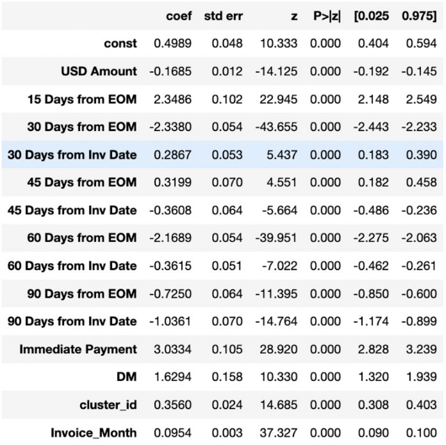


After dropping INV –
data has no high multicollinearity.

Model Building

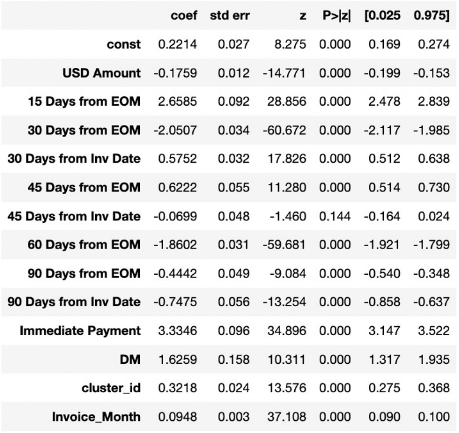
Model Building

1st Model



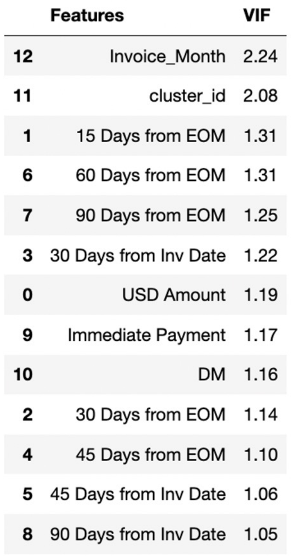
	coef	std err	z	P>|z|	[0.025	0.975]
const	0.4989	0.048	10.333	0.000	0.404	0.594
USD Amount	-0.1685	0.012	-14.125	0.000	-0.192	-0.145
15 Days from EOM	2.3486	0.102	22.945	0.000	2.148	2.549
30 Days from EOM	-2.3380	0.054	-43.655	0.000	-2.443	-2.233
30 Days from Inv Date	0.2867	0.053	5.437	0.000	0.183	0.390
45 Days from EOM	0.3199	0.070	4.551	0.000	0.182	0.458
45 Days from Inv Date	-0.3608	0.064	-5.664	0.000	-0.486	-0.236
60 Days from EOM	-2.1689	0.054	-39.951	0.000	-2.275	-2.063
60 Days from Inv Date	-0.3615	0.051	-7.022	0.000	-0.462	-0.261
90 Days from EOM	-0.7250	0.064	-11.395	0.000	-0.850	-0.600
90 Days from Inv Date	-1.0361	0.070	-14.764	0.000	-1.174	-0.899
Immediate Payment	3.0334	0.105	28.920	0.000	2.828	3.239
DM	1.6294	0.158	10.330	0.000	1.320	1.939
cluster_id	0.3560	0.024	14.685	0.000	0.308	0.403
Invoice_Month	0.0954	0.003	37.327	0.000	0.090	0.100

2nd Model



	coef	std err	z	P>|z|	[0.025	0.975]
const	0.2214	0.027	8.275	0.000	0.169	0.274
USD Amount	-0.1759	0.012	-14.771	0.000	-0.199	-0.153
15 Days from EOM	2.6585	0.092	28.856	0.000	2.478	2.839
30 Days from EOM	-2.0507	0.034	-60.672	0.000	-2.117	-1.985
30 Days from Inv Date	0.5752	0.032	17.826	0.000	0.512	0.638
45 Days from EOM	0.6222	0.055	11.280	0.000	0.514	0.730
45 Days from Inv Date	-0.0699	0.048	-1.460	0.144	-0.164	0.024
60 Days from EOM	-1.8602	0.031	-59.681	0.000	-1.921	-1.799
90 Days from EOM	-0.4442	0.049	-9.084	0.000	-0.540	-0.348
90 Days from Inv Date	-0.7475	0.056	-13.254	0.000	-0.858	-0.637
Immediate Payment	3.3346	0.096	34.896	0.000	3.147	3.522
DM	1.6259	0.158	10.311	0.000	1.317	1.935
cluster_id	0.3218	0.024	13.576	0.000	0.275	0.368
Invoice_Month	0.0948	0.003	37.108	0.000	0.090	0.100

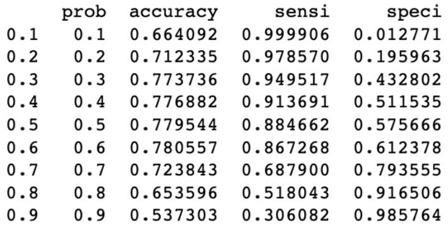
Final features



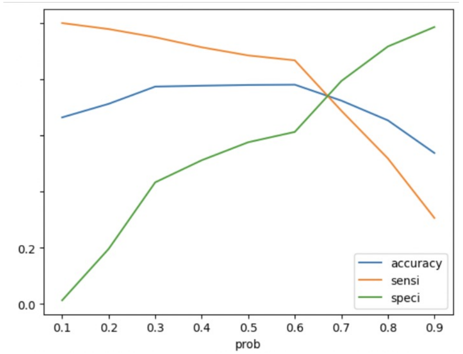
	Features	VIF
12	Invoice_Month	2.24
11	cluster_id	2.08
1	15 Days from EOM	1.31
6	60 Days from EOM	1.31
7	90 Days from EOM	1.25
3	30 Days from Inv Date	1.22
0	USD Amount	1.19
9	Immediate Payment	1.17
10	DM	1.16
2	30 Days from EOM	1.14
4	45 Days from EOM	1.10
5	45 Days from Inv Date	1.06
8	90 Days from Inv Date	1.05

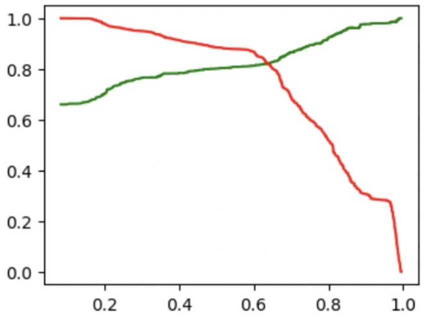
- After checking 1st Model we performed manual feature elimination of '60 Days from Inv Date’
- Final model has 13 features .
- All the P value VIF are in accepted range.
- We used this model to make predictions on traning and testing data set.

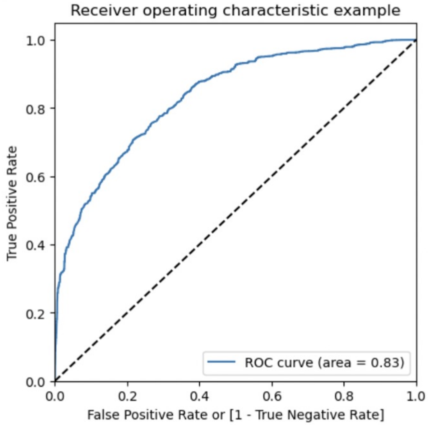
Model Evaluation and Observation



```
     prob  accuracy     sensi     speci
0.1   0.1  0.664092  0.999906  0.012771
0.2   0.2  0.712335  0.978570  0.195963
0.3   0.3  0.773736  0.949517  0.432802
0.4   0.4  0.776882  0.913691  0.511535
0.5   0.5  0.779544  0.884662  0.575666
0.6   0.6  0.780557  0.867268  0.612378
0.7   0.7  0.723843  0.687900  0.793555
0.8   0.8  0.653596  0.518043  0.916506
0.9   0.9  0.537303  0.306082  0.985764
```



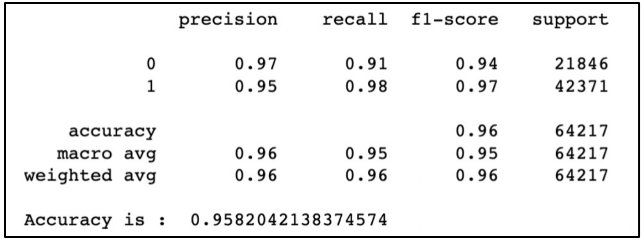




- From the curve, 0.6 is the optimum point to take it as a cutoff probability.
- On Precision & Recall trade off we found optimal cutoff of between 0.6 & 0.7 . Hence keeping the optimal cutoff 0.6.
- AUC = 0.83 which shows the model is good.
- Our train and test accuracy is almost same around 77-78 %

Random Forest

- Accuracy of Training data set is 95%
- Accuracy of test data set is : 92%
- f1-score for train and test set is 0.96 and 0.93 respectively , which implies that this is a good model. Hence moving forward with this as final model for prediction.

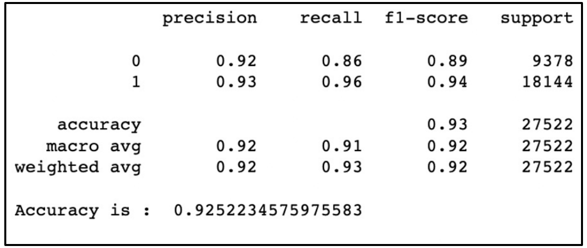


```
              precision    recall  f1-score   support

           0       0.97      0.91      0.94     21846
           1       0.95      0.98      0.97     42371

    accuracy                           0.96     64217
   macro avg       0.96      0.95      0.95     64217
weighted avg       0.96      0.96      0.96     64217

Accuracy is :  0.9582042138374574
```

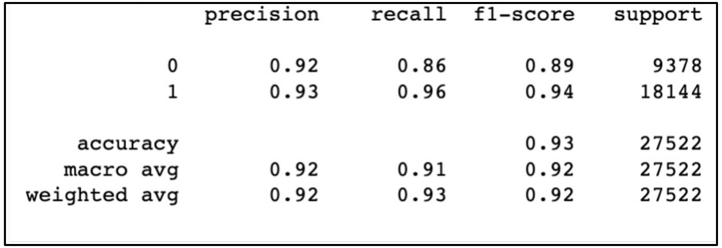


```
              precision    recall  f1-score   support

           0       0.92      0.86      0.89      9378
           1       0.93      0.96      0.94     18144

    accuracy                           0.93     27522
   macro avg       0.92      0.91      0.92     27522
weighted avg       0.92      0.93      0.92     27522

Accuracy is :  0.9252234575975583
```



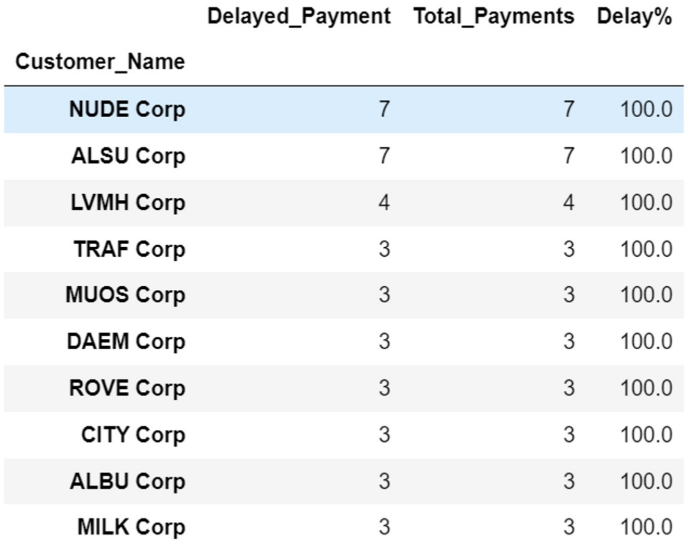
```
              precision    recall  f1-score   support

           0       0.92      0.86      0.89      9378
           1       0.93      0.96      0.94     18144

    accuracy                           0.93     27522
   macro avg       0.92      0.91      0.92     27522
weighted avg       0.92      0.93      0.92     27522
```

Conclusion

- We have got two models Logistic Regression and Random Forest predicting different probability values for a customer to have a late payment.
- It is observed that Random forest is performing much better than logistic regression. We can utilize this model and make pre-emptive calls to the customers to have them pay their invoice amounts on time.
- According to the prediction, companies in the right table has maximum probability to default with maximum number of delayed payments. These companies should be focused more.



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

Business Insights/Recommendations

- Credit Memo invoice class observed the high delay rate compared to Debit Memo or Invoice type invoice classes, hence company policies on payment collection could be made stricter around CM class.
- Goods type invoices had significantly higher payment delay rates than non-goods types and hence can be subjected to stricter payment policies.
- In the 2nd half of the year, late payments are observed to increase steeply from 7th months onwards even if the number of invoices are low. Company employees can pay more attention to actively chase customers to make payments.
- Customer segments were clustered into three categories, viz., 0,1 and 2 which prolonged, early and medium payment duration respectively. It is observed that customers in cluster 0 (prolonged days) had significantly higher delay rates than early and medium days of payment, hence cluster 0 customers should be paid extensive focus

Thank you!