E-COMMERCE & RETAIL B2B CASE STUDY

Priyanka Raut, Sandhya K & Radha J

OVERVIEW

COMPANY PROFILE

- Schuster is a multinational retail company dealing in sports goods and accessories.
- It 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.

BUSINESS CASE & OBJECTIVE

- Schuster 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.
- It would thus try to understand its customers' payment behavior and predict the likelihood of late payments against open invoices (customer segmentation)
- It wants to use this information so that collectors can prioritize their work in following up with customers beforehand to get the payments on time.

upGrad

APPROACH

DATA READING & UNDERSTANDING

1. Importing necessary Packages and Libraries
2. Reading dataset
3. Basic Information - Using methods like .describe(), .info() to get basic info of the data

DATA CLEANING & PREPARATION

1. Missing Values: Identifying missing data through imputation or removal
2. Outliers: Detecting outliers, if appropriate
3. Inconsistencies: Correcting any inconsistencies in data
4. Data Types: Ensuring each column is of the correct data type

EXPLORATORY DATA ANALYSIS

1. Univariate analysis
2. Bivariate analysis
3. Co-relation matrix using heatmaps to get insights on variables

FEATURE ENGINEERING

1. Dummy Variable Creation
2. One Hot Encoding
3. Standard Scaling to transform the data to make it more suitable for modeling

APPROACH

TRAIN TEST SPLIT

1. Dividing Dataset in Train data and Test Data in the ratio of 70:30
2. Cross Validation dataset to ensure that the model does not overfit

MODEL BUILDING

1. Building a simple Model first and check the evaluation metrics
2. Check p-value and VIF levels, in case of Logistics Regression
3. Drop variable one by one and rebuild the model
4. Hyperparameter tuning using GridSearch CV to arrive at best hyperparameters

MODEL EVALUATION

1. Review Accuracy, Recall, Precision, Specificity and Sensitivity on Train data, in case of Logistics Regression model
2. Optimal Cut-off / Threshold using Precision Recall Trade-off
3. Look at Classification report in case of Random Forests

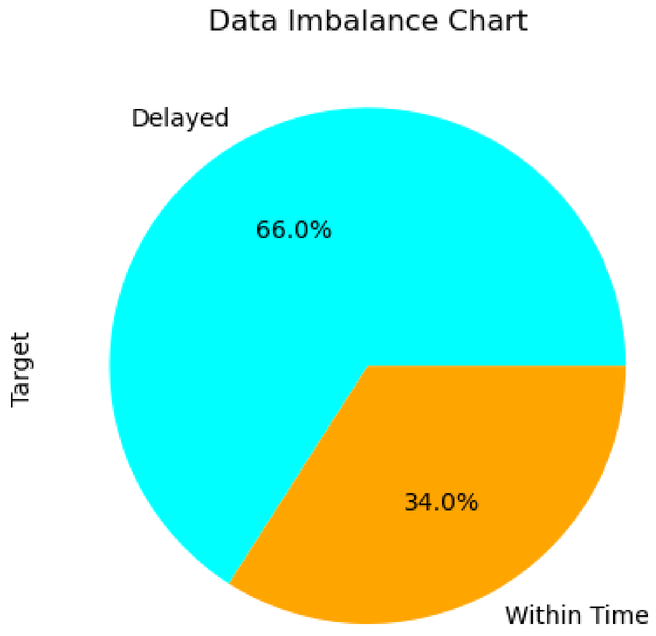
PREDICTIONS ON TEST DATA

1. Review Accuracy, Recall, Precision, Specificity and Sensitivity on Test data
2. Accuracy score of Test should be within +/-5% range of Train data to ensure that model has not overfit
3. Look at Feature importance of best model, in case of RandomForests

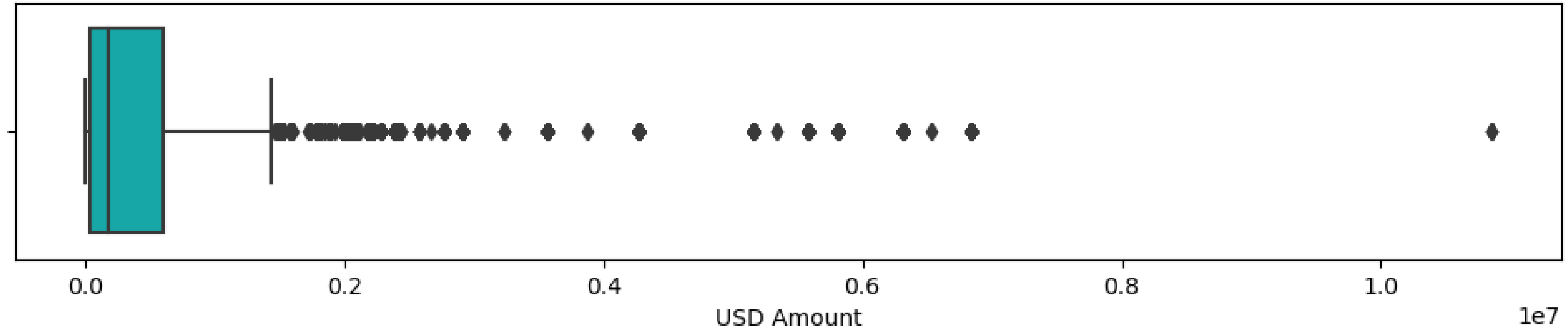
upGrad

UNDERSTANDING DATASET

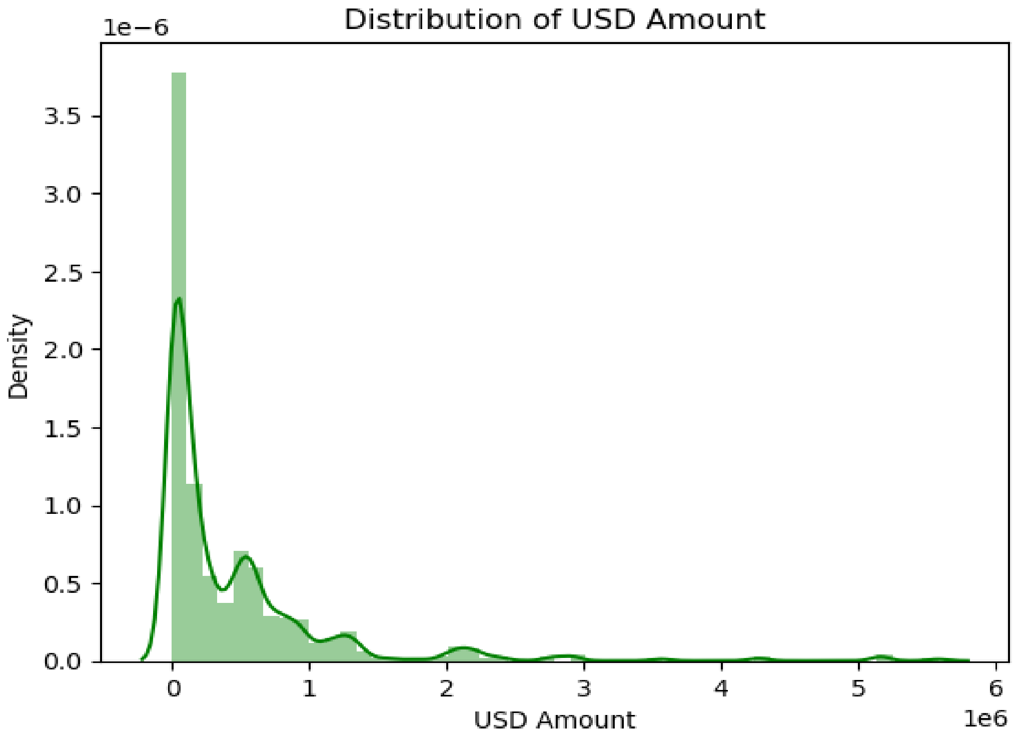
- Dataset has total 16 columns and 93937 rows.
- There are total 12 object values, 1 integer value int64 & 3 float values.
- Outliers are treated by Capping Method
- Null values : The "RECEIPT_DOC_NO" col has 0.03% null values was dropped as it is not important for model building.
- Target variable : Derived by checking whether payment receipt date falls within, or post due date.
- Data Imbalance : 66% of total was late payment.

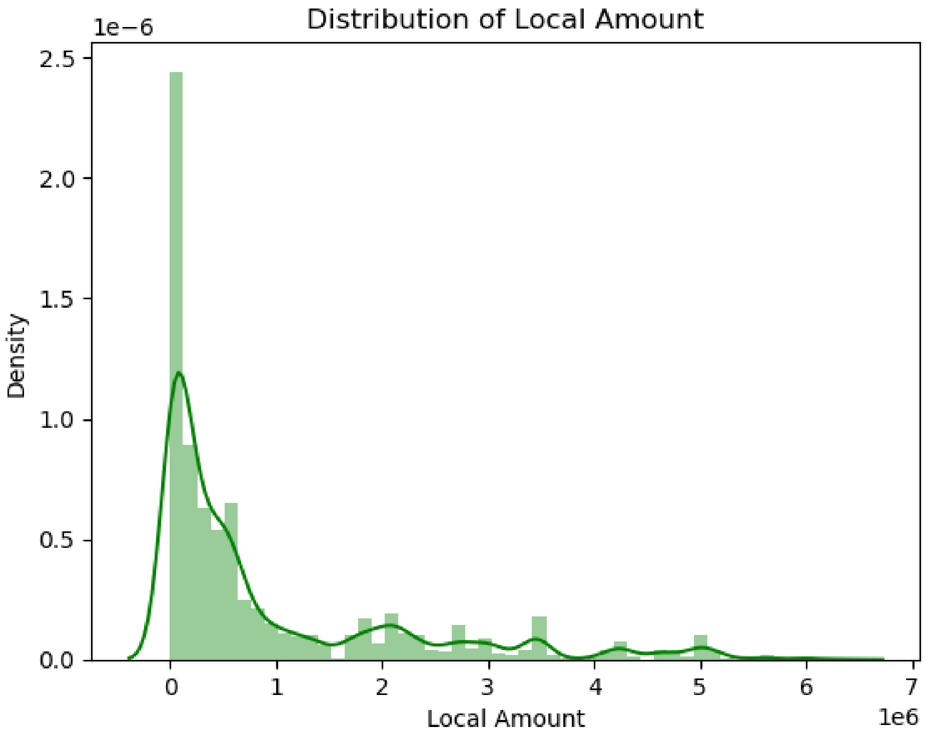


Outlier Detection using Boxplot



ANALYSIS - UNIVARIATE

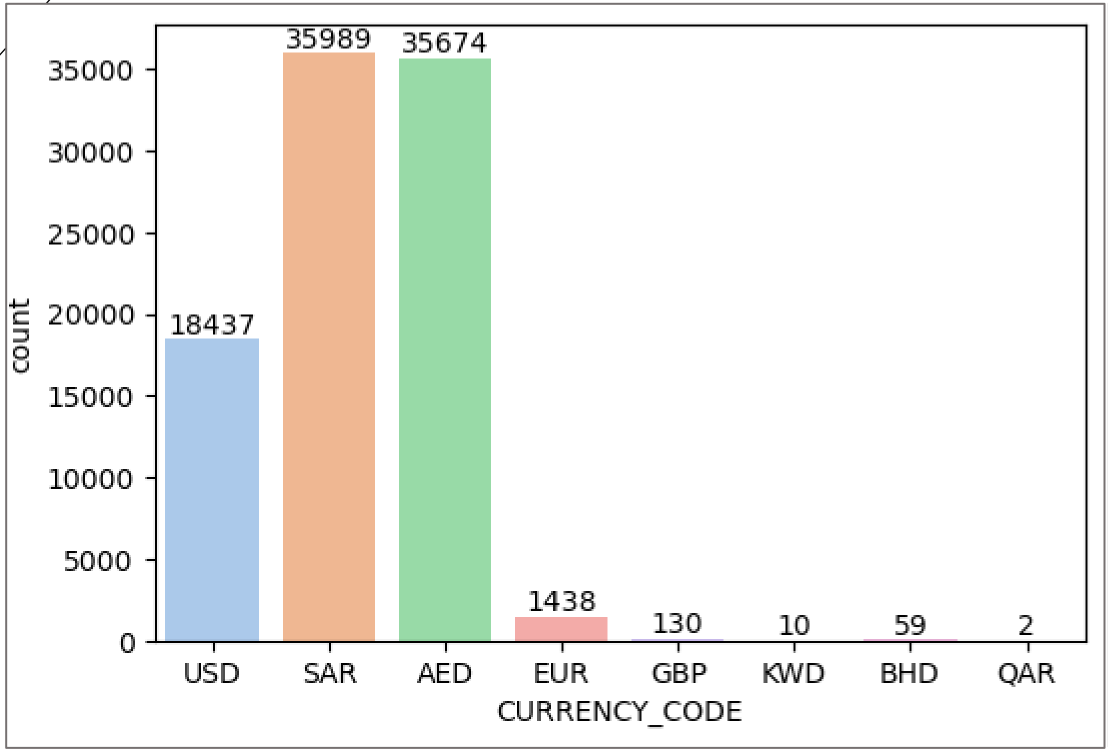

Distribution of USD Amount

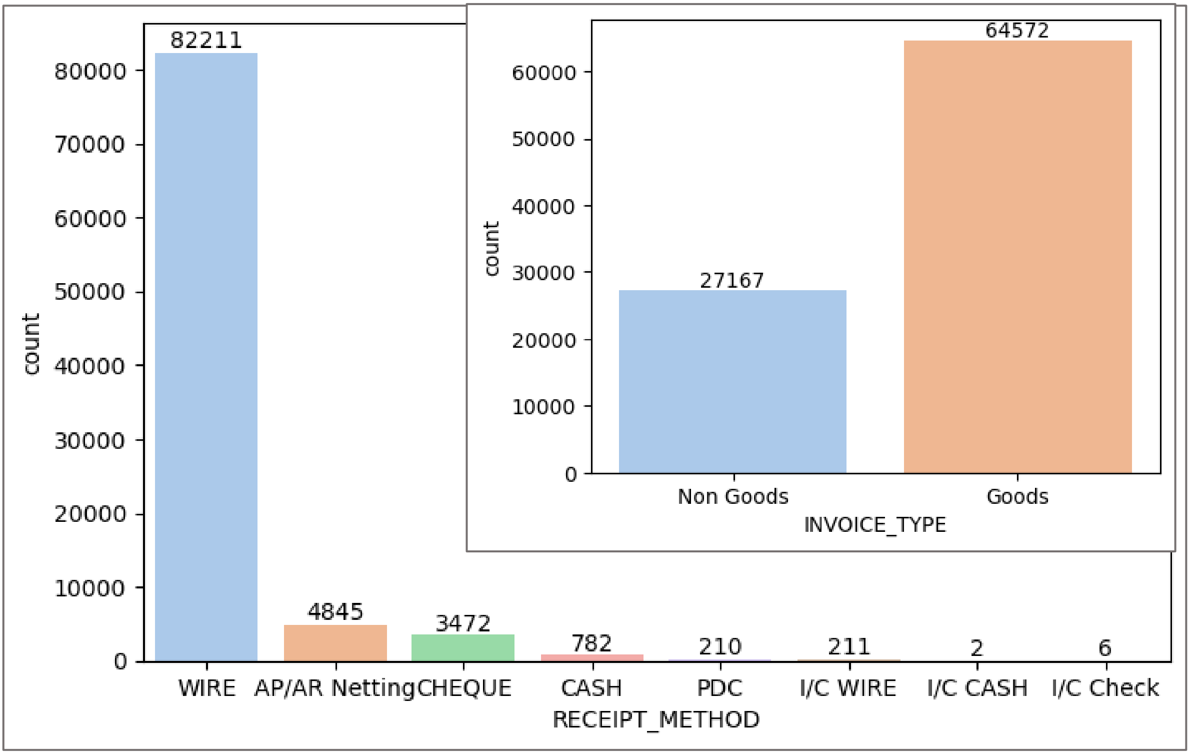

Distribution of Local Amount

Inferences:

- Since Local Amount column contains amount for different currency, we have utilized USD amount column and safely dropped Local Amount

ANALYSIS - BIVARIATE

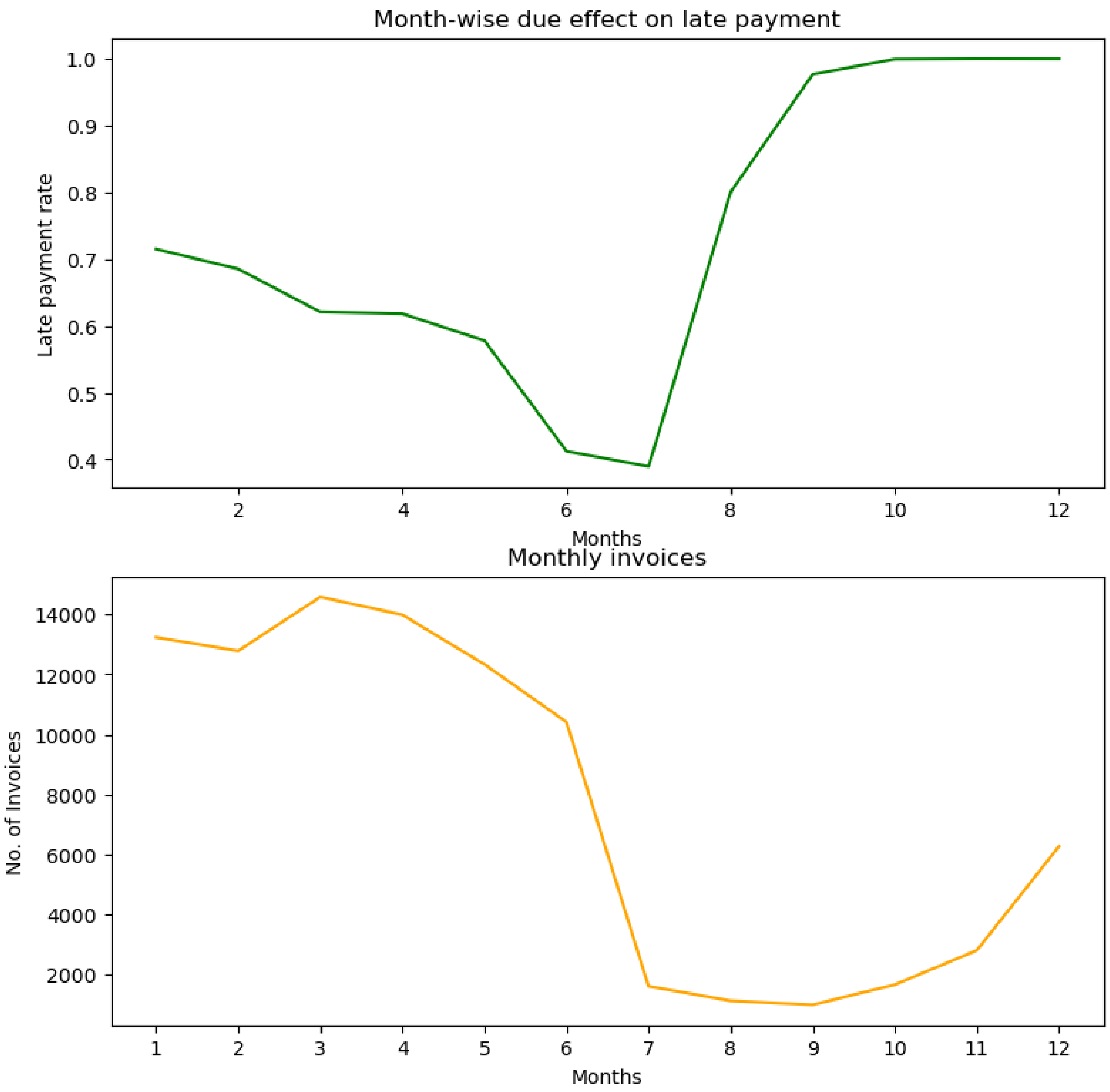




Inferences:

- 1st plot depicts that Currency used for bill payments are mostly USD, SAR or AED.
- 2nd plot depicts Majority of the invoices are raised for Goods.
- 3rd plot depicts that the most preferred payment method for bill payment is WIRE

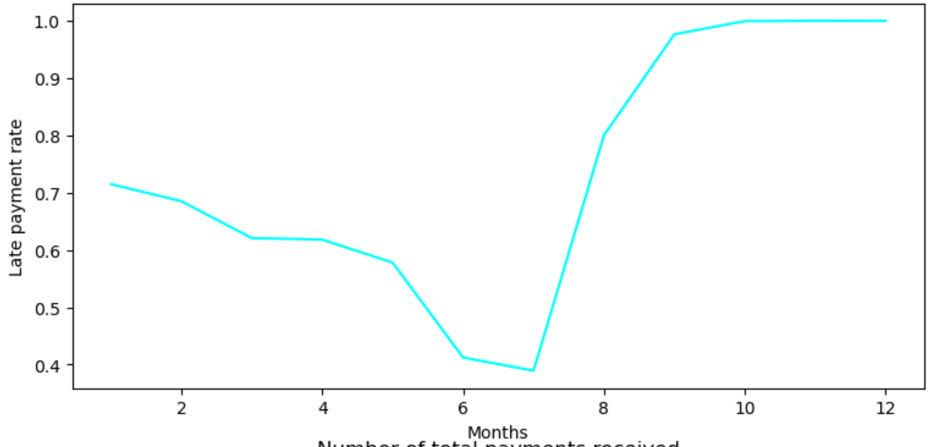
ANALYSIS - BIVARIATE

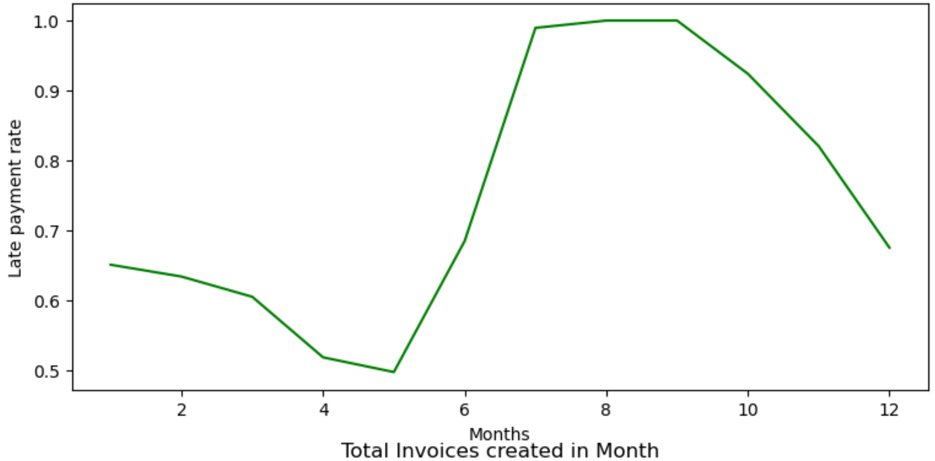


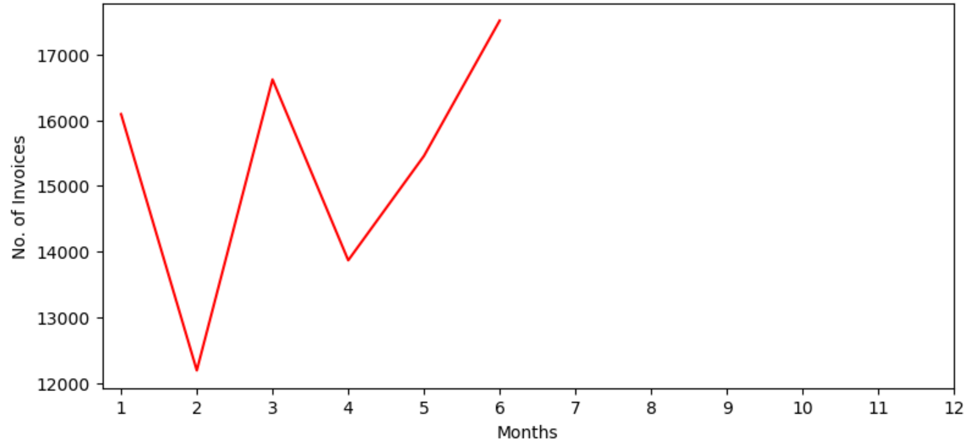
Inference: Graphs depicts that

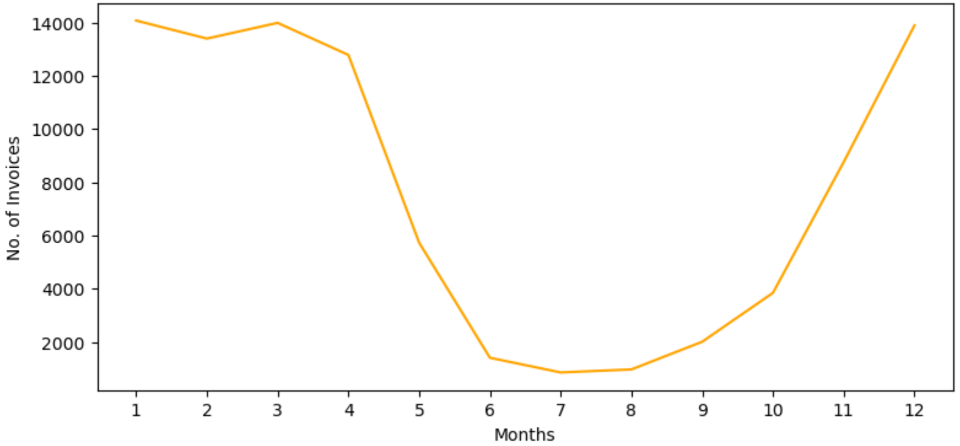
- For the 3rd month, the number of invoices is the highest and late payment rate is comparatively lower than other months with large number of invoices.
- Month 7 has the very low late payment rate, this can be because of the fact that the number of invoices is also low.
- In the 2nd half of the year,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

ANALYSIS - BIVARIATE


% of late payment received in Month


% of late payment received in Month

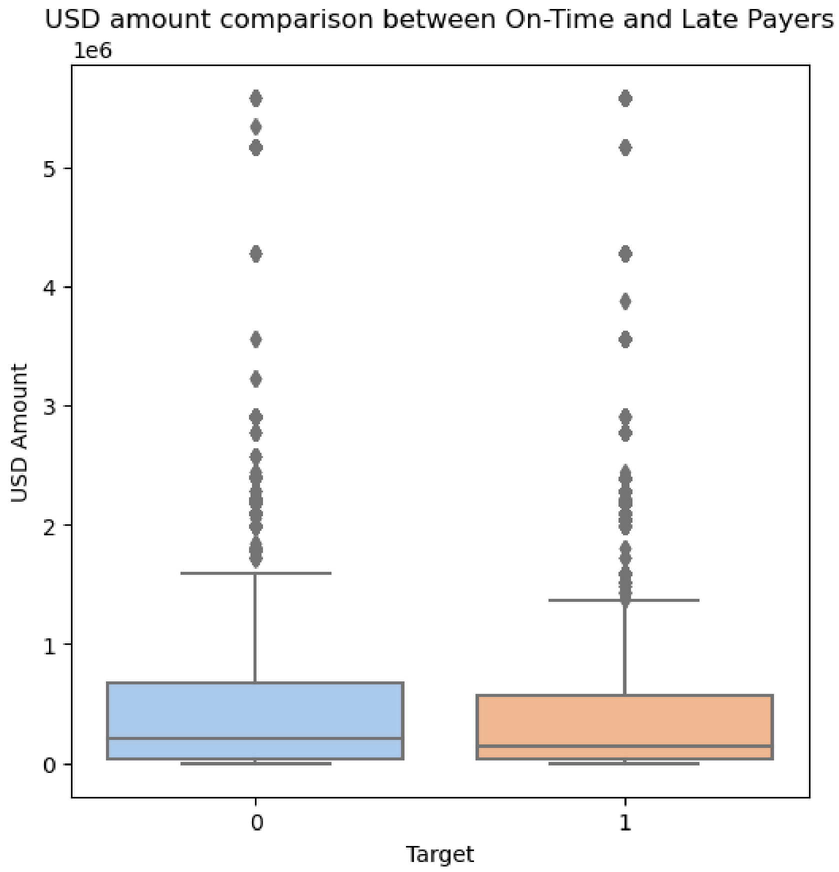

Number of total payments received

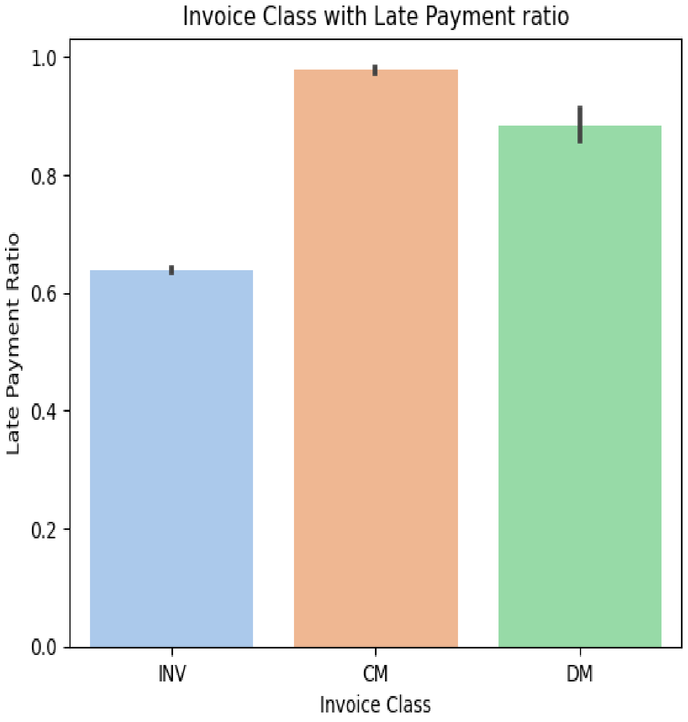

Total Invoices created in Month

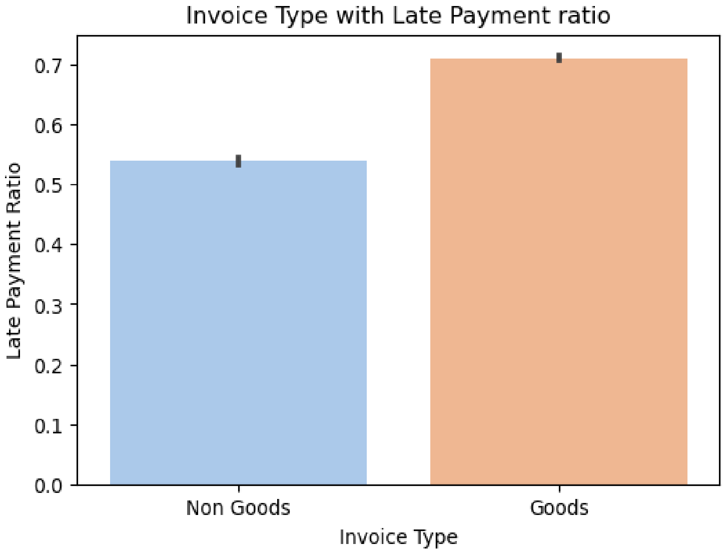
Inference:

- From 7th month onwards, none of the invoices are paid i.e. No payment received against any invoices after 7th month
- Late payment rate decreases from 1st to 5th month.
- For the months 7, 8 and 9, the late payment rate is very high.

ANALYSIS - BIVARIATE


USD amount comparison between On-Time and Late Payers

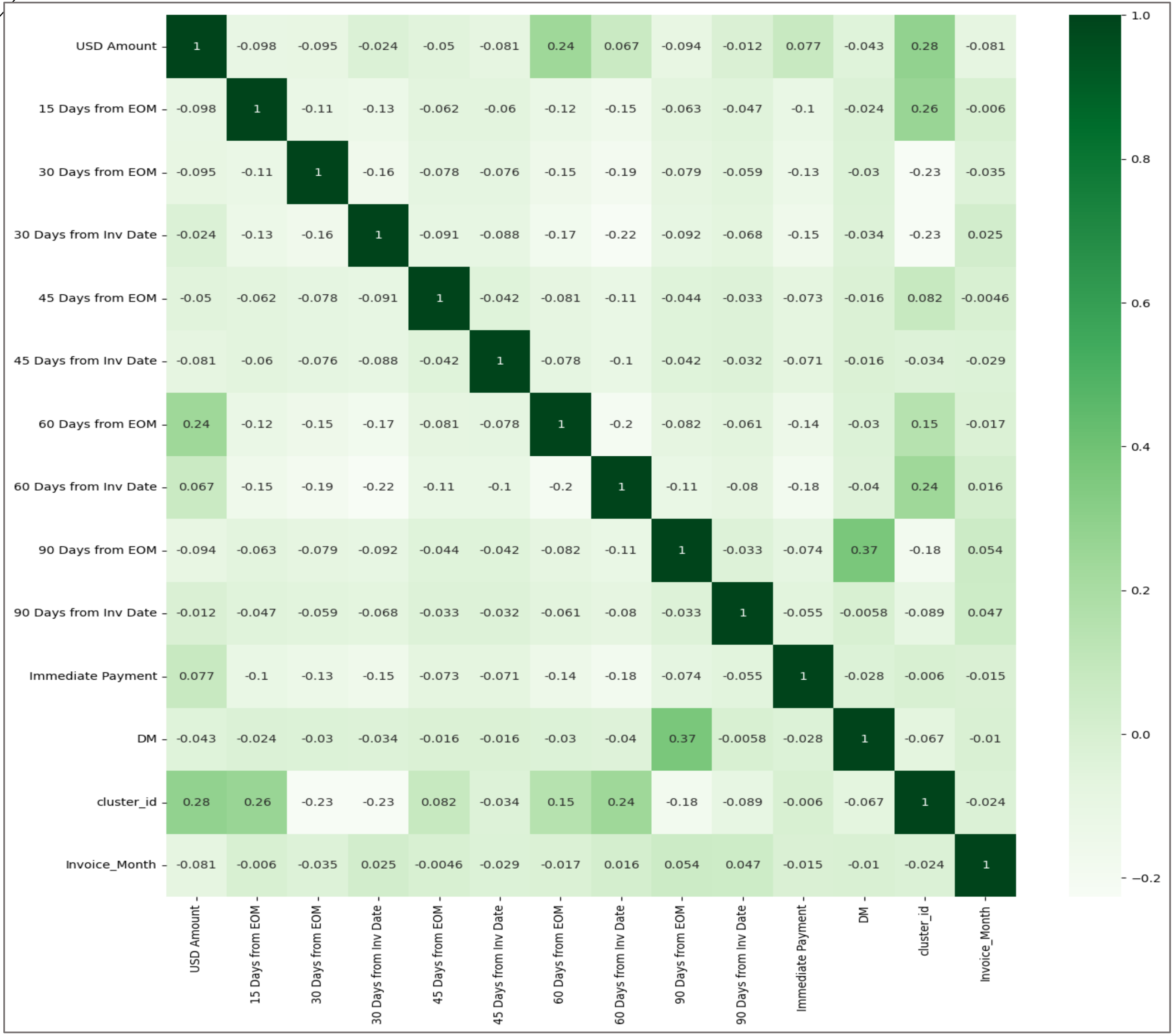

Invoice Class with Late Payment ratio


Invoice Type with Late Payment ratio

Inference:

- Median of on time and late payer is approximately the same.
- Late payment ratio for INVOICE_CLASS is very high for CM and lowest for INV.
- Late payment ratio for INVOICE TYPE is very high for GOODS than Non Goods

CO-RELATION BETWEEN VARIABLES



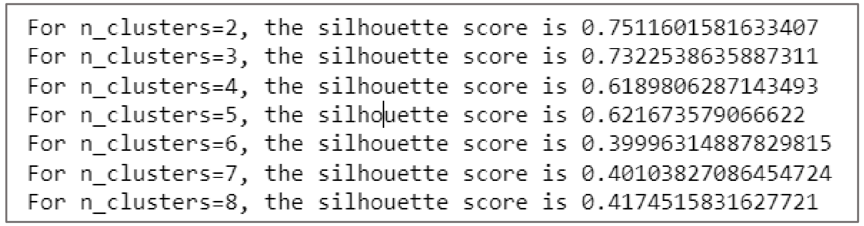
	USD Amount	15 Days from EOM	30 Days from EOM	30 Days from Inv Date	45 Days from EOM	45 Days from Inv Date	60 Days from EOM	60 Days from Inv Date	90 Days from EOM	90 Days from Inv Date	Immediate Payment	DM	cluster_id	Invoice_Month
USD Amount	1	-0.098	-0.095	-0.024	-0.05	-0.081	0.24	0.067	-0.094	-0.012	0.077	-0.043	0.28	-0.081
15 Days from EOM	-0.098	1	-0.11	-0.13	-0.062	-0.06	-0.12	-0.15	-0.063	-0.047	-0.1	-0.024	0.26	-0.006
30 Days from EOM	-0.095	-0.11	1	-0.16	-0.078	-0.076	-0.15	-0.19	-0.079	-0.059	-0.13	-0.03	-0.23	-0.035
30 Days from Inv Date	-0.024	-0.13	-0.16	1	-0.091	-0.088	-0.17	-0.22	-0.092	-0.068	-0.15	-0.034	-0.23	0.025
45 Days from EOM	-0.05	-0.062	-0.078	-0.091	1	-0.042	-0.081	-0.11	-0.044	-0.033	-0.073	-0.016	0.082	-0.0046
45 Days from Inv Date	-0.081	-0.06	-0.076	-0.088	-0.042	1	-0.078	-0.1	-0.042	-0.032	-0.071	-0.016	-0.034	-0.029
60 Days from EOM	0.24	-0.12	-0.15	-0.17	-0.081	-0.078	1	-0.2	-0.082	-0.061	-0.14	-0.03	0.15	-0.017
60 Days from Inv Date	0.067	-0.15	-0.19	-0.22	-0.11	-0.1	-0.2	1	-0.11	-0.08	-0.18	-0.04	0.24	0.016
90 Days from EOM	-0.094	-0.063	-0.079	-0.092	-0.044	-0.042	-0.082	-0.11	1	-0.033	-0.074	0.37	-0.18	0.054
90 Days from Inv Date	-0.012	-0.047	-0.059	-0.068	-0.033	-0.032	-0.061	-0.08	-0.033	1	-0.055	-0.0058	-0.089	0.047
Immediate Payment	0.077	-0.1	-0.13	-0.15	-0.073	-0.071	-0.14	-0.18	-0.074	-0.055	1	-0.028	-0.006	-0.015
DM	-0.043	-0.024	-0.03	-0.034	-0.016	-0.016	-0.03	-0.04	0.37	-0.0058	-0.028	1	-0.067	-0.01
cluster_id	0.28	0.26	-0.23	-0.23	0.082	-0.034	0.15	0.24	-0.18	-0.089	-0.006	-0.067	1	-0.024
Invoice_Month	-0.081	-0.006	-0.035	0.025	-0.0046	-0.029	-0.017	0.016	0.054	0.047	-0.015	-0.01	-0.024	1

Inference:

- We formed a heatmap 1st with all important variables and dropped INV variable since INV & Immediate Payment had high multicollinearity
- The 2nd heatmap which is shown here depicts that data has no high multicollinearity and further model building can be done.

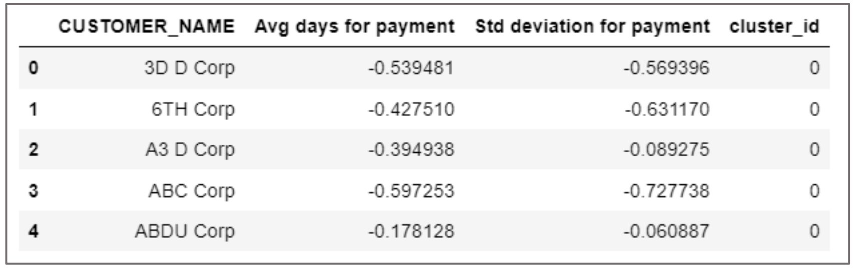
PART 1 OF CASE STUDY: CUSTOMER SEGMENTATION BY CLUSTERING

Silhouette analysis to decide n_clusters



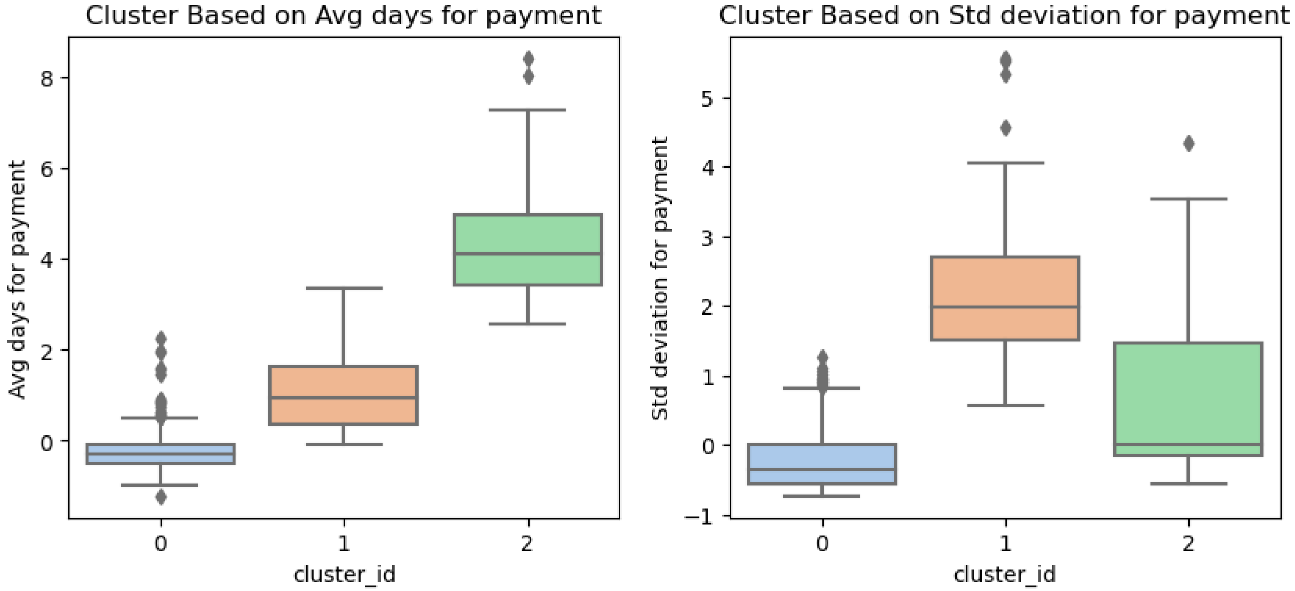
```
For n_clusters=2, the silhouette score is 0.7511601581633407
For n_clusters=3, the silhouette score is 0.7322538635887311
For n_clusters=4, the silhouette score is 0.6189806287143493
For n_clusters=5, the silhouette score is 0.621673579066622
For n_clusters=6, the silhouette score is 0.39996314887829815
For n_clusters=7, the silhouette score is 0.40103827086454724
For n_clusters=8, the silhouette score is 0.4174515831627721
```

Assigning cluster_id to customers in df



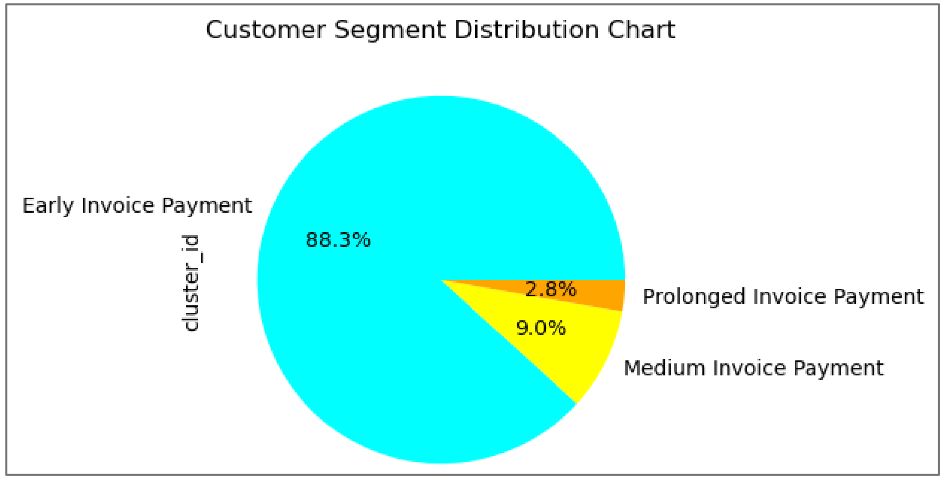
	CUSTOMER_NAME	Avg days for payment	Std deviation for payment	cluster_id
0	3D D Corp	-0.539481	-0.569396	0
1	6TH Corp	-0.427510	-0.631170	0
2	A3 D Corp	-0.394938	-0.089275	0
3	ABC Corp	-0.597253	-0.727738	0
4	ABDU Corp	-0.178128	-0.060887	0

Avg days for payment and Std deviation for payment



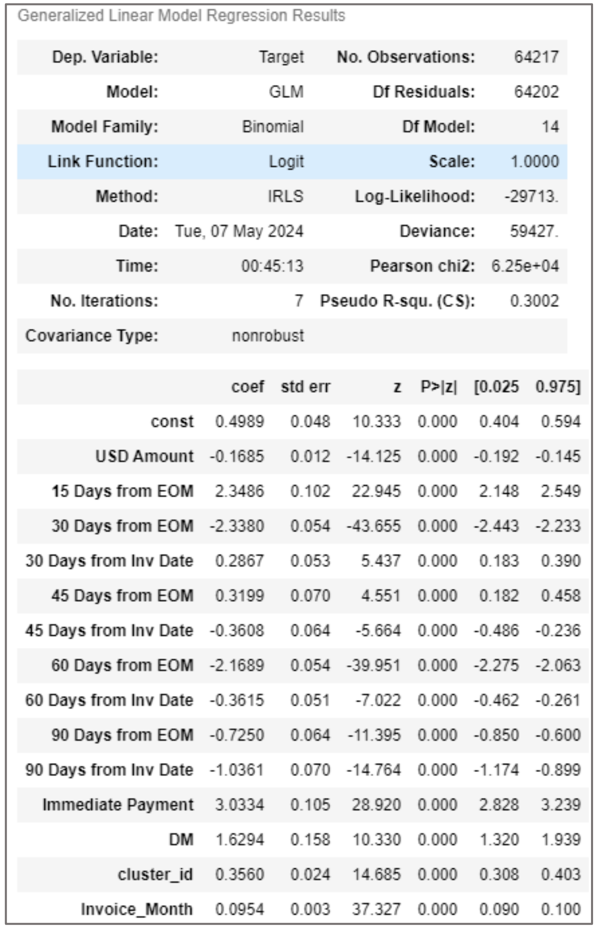
Inference:

- For 3 clusters, silhouette score is decent
- Median for prolonged invoice payment is higher than Early invoice payment
- Early customers comprise of 88.3% whereas medium & prolonged payers are 11.3% in total



PART 2 MODEL BUILDING – LOGISTIC REGRESSION

1st Model

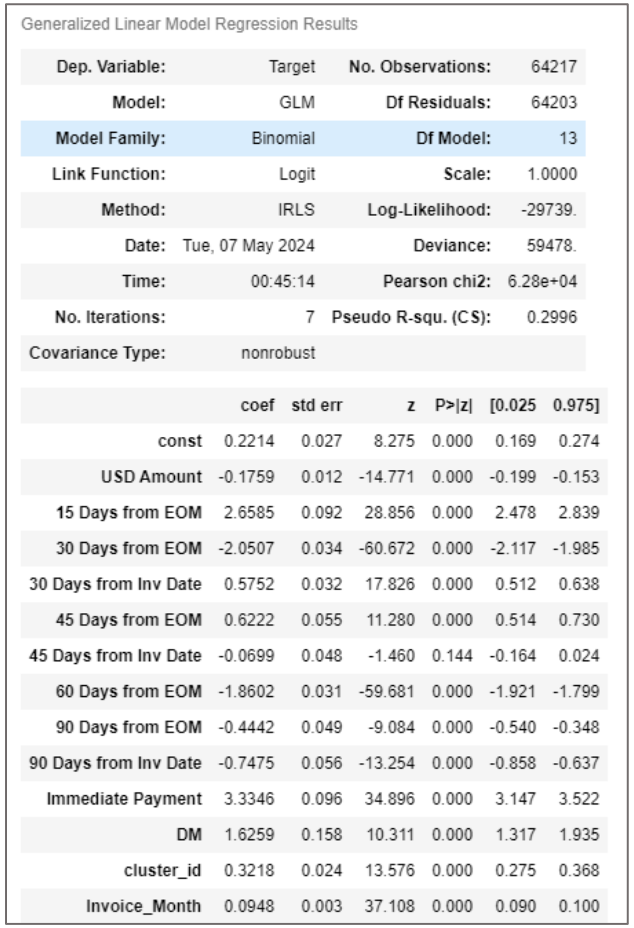


Generalized Linear Model Regression Results

Dep. Variable:	Target	No. Observations:	64217
Model:	GLM	Df Residuals:	64202
Model Family:	Binomial	Df Model:	14
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-29713.
Date:	Tue, 07 May 2024	Deviance:	59427.
Time:	00:45:13	Pearson chi2:	6.25e+04
No. Iterations:	7	Pseudo R-squ. (CS):	0.3002
Covariance Type:	nonrobust		

	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

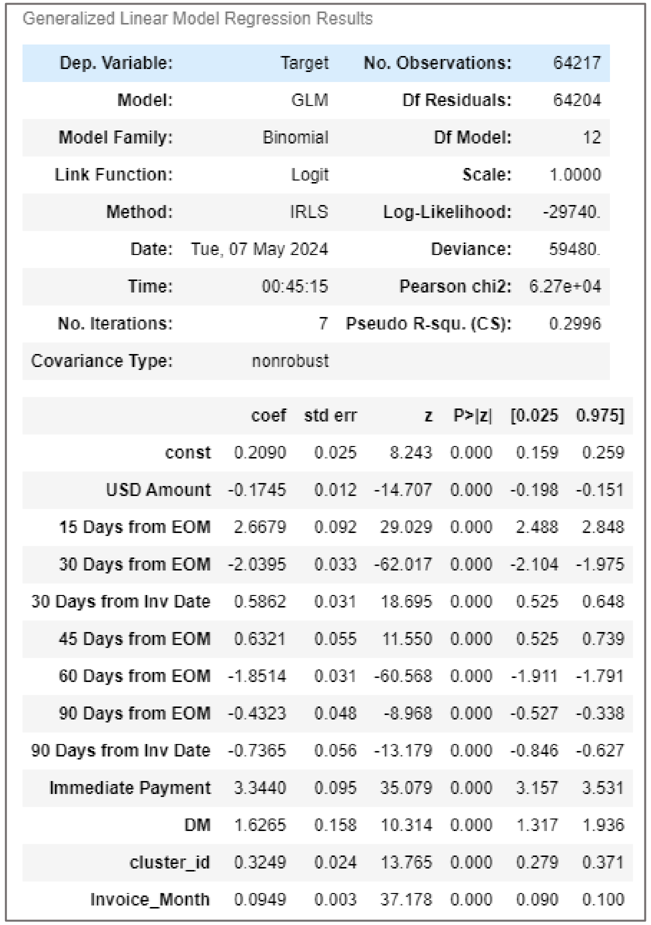


Generalized Linear Model Regression Results

Dep. Variable:	Target	No. Observations:	64217
Model:	GLM	Df Residuals:	64203
Model Family:	Binomial	Df Model:	13
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-29739.
Date:	Tue, 07 May 2024	Deviance:	59478.
Time:	00:45:14	Pearson chi2:	6.28e+04
No. Iterations:	7	Pseudo R-squ. (CS):	0.2996
Covariance Type:	nonrobust		

	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

3rd Model



Generalized Linear Model Regression Results

Dep. Variable:	Target	No. Observations:	64217
Model:	GLM	Df Residuals:	64204
Model Family:	Binomial	Df Model:	12
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-29740.
Date:	Tue, 07 May 2024	Deviance:	59480.
Time:	00:45:15	Pearson chi2:	6.27e+04
No. Iterations:	7	Pseudo R-squ. (CS):	0.2996
Covariance Type:	nonrobust		

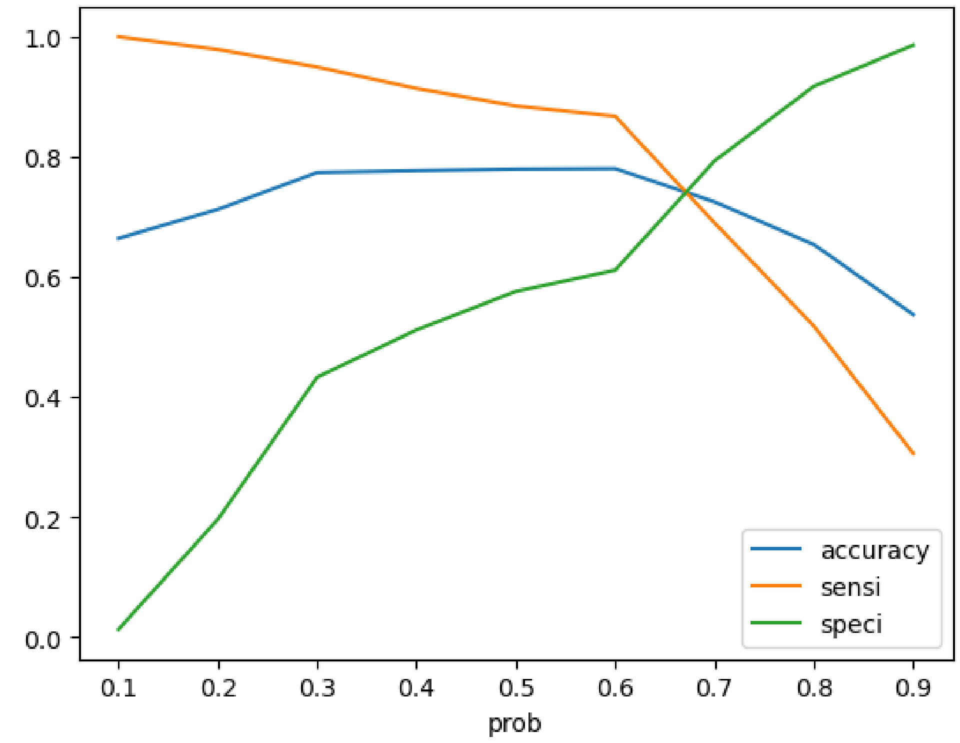
	coef	std err	z	P>|z|	[0.025	0.975]
const	0.2090	0.025	8.243	0.000	0.159	0.259
USD Amount	-0.1745	0.012	-14.707	0.000	-0.198	-0.151
15 Days from EOM	2.6679	0.092	29.029	0.000	2.488	2.848
30 Days from EOM	-2.0395	0.033	-62.017	0.000	-2.104	-1.975
30 Days from Inv Date	0.5862	0.031	18.695	0.000	0.525	0.648
45 Days from EOM	0.6321	0.055	11.550	0.000	0.525	0.739
60 Days from EOM	-1.8514	0.031	-60.568	0.000	-1.911	-1.791
90 Days from EOM	-0.4323	0.048	-8.968	0.000	-0.527	-0.338
90 Days from Inv Date	-0.7365	0.056	-13.179	0.000	-0.846	-0.627
Immediate Payment	3.3440	0.095	35.079	0.000	3.157	3.531
DM	1.6265	0.158	10.314	0.000	1.317	1.936
cluster_id	0.3249	0.024	13.765	0.000	0.279	0.371
Invoice_Month	0.0949	0.003	37.178	0.000	0.090	0.100

Inference:

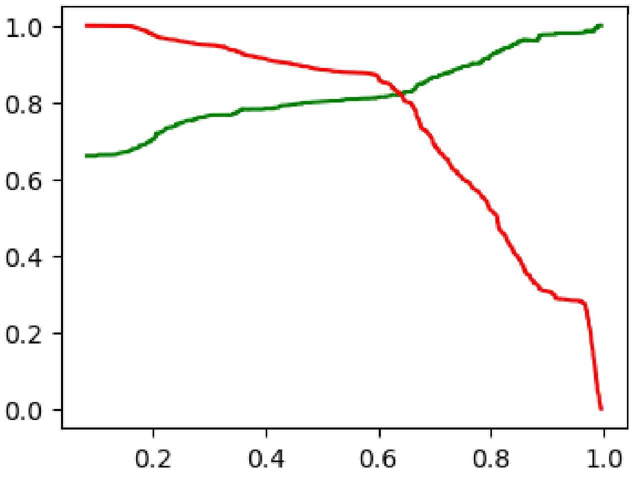
- In 1st Model we performed manual feature elimination of '60 Days from Inv Date’ (high VIF)
- In Model 2, we dropped '60 Days from Inv Date’ which had high p-value
- Our 3rd Model i.e. Final model has 11 features with p-value & VIF in acceptable range

MODEL EVALUATION

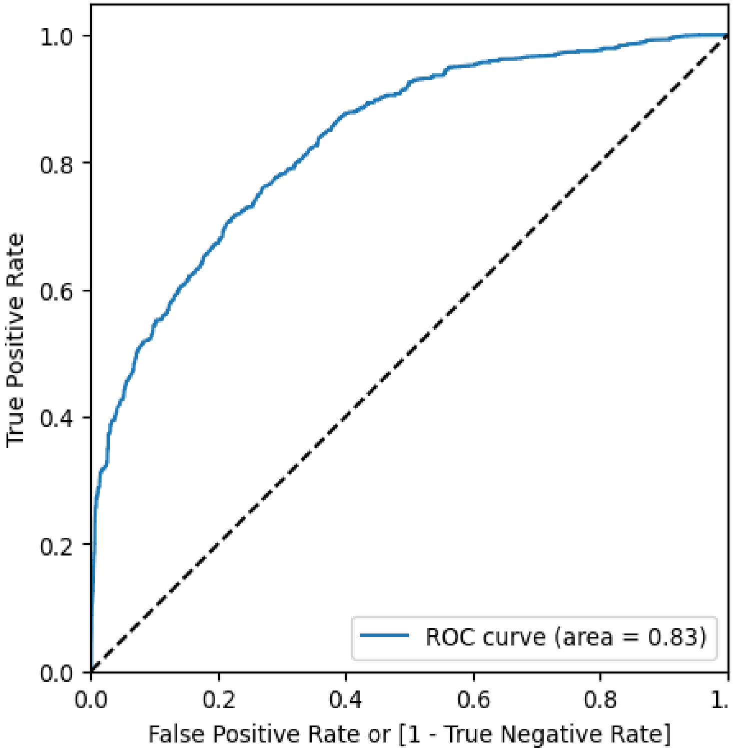
Accuracy, Sensitivity Specificity across probabilities



Precision Recall Trade-off



Receiver operating Characteristic

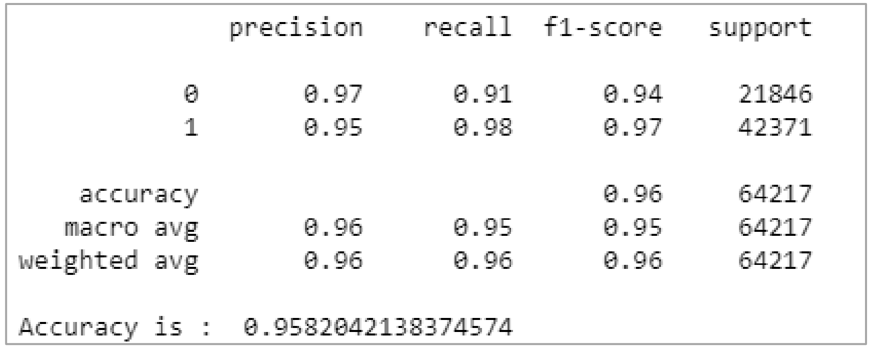


Inference:

- 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 taken it as 0.6.
- AUC = 0.83 which shows the model is good, Our train & test accuracy is almost same around 77-78 %

MODEL BUILDING – RANDOM FOREST

Classification Report on Train data prediction



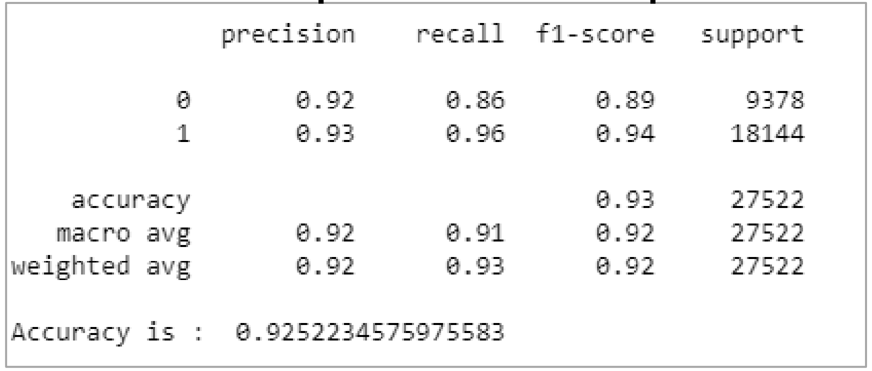
```
              precision    recall  f1-score   support

           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
```

Classification Report on Test data prediction



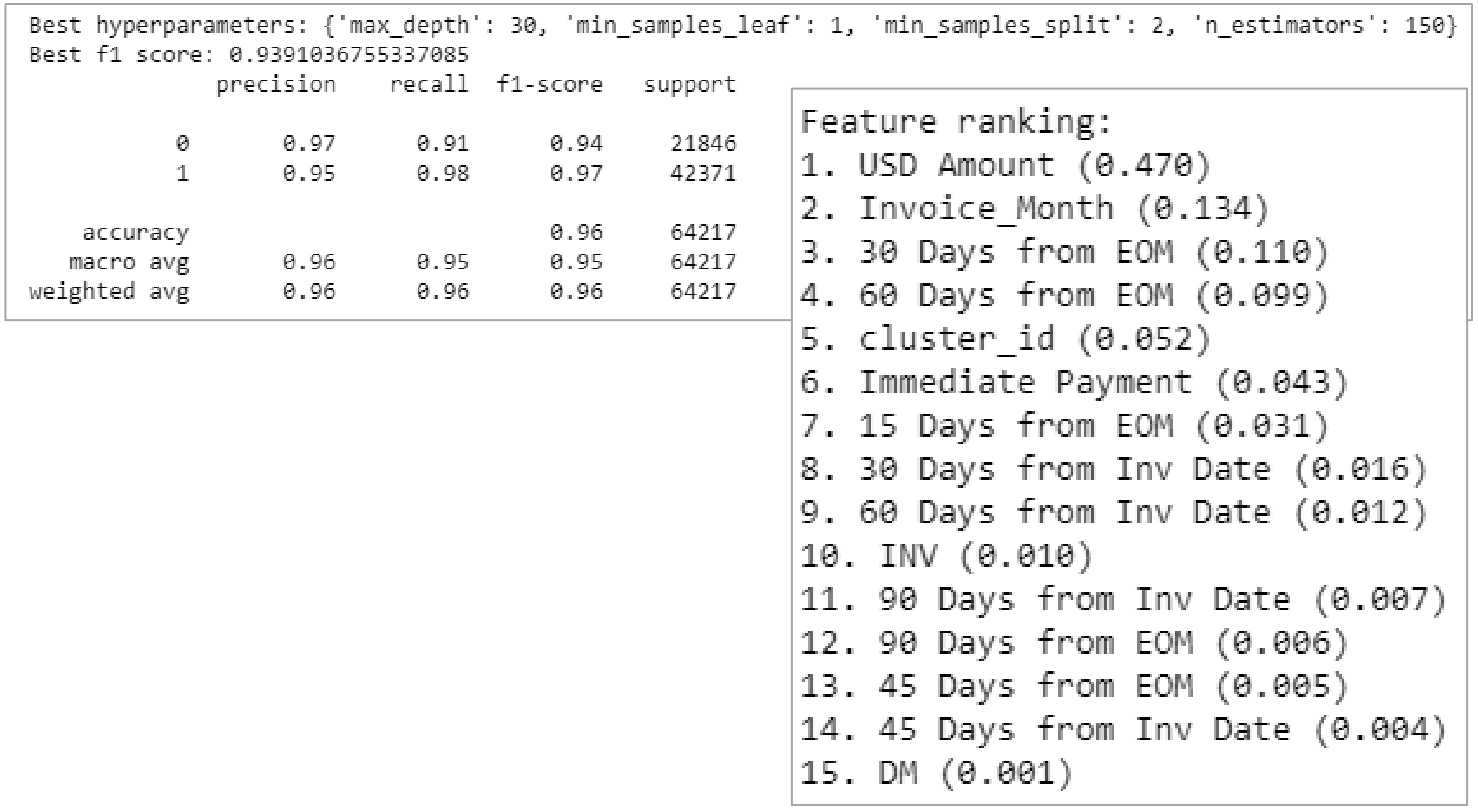
```
              precision    recall  f1-score   support

           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
```

Hyperparameter Tuning



```
Best hyperparameters: {'max_depth': 30, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 150}
Best f1 score: 0.9391036755337085
              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
```

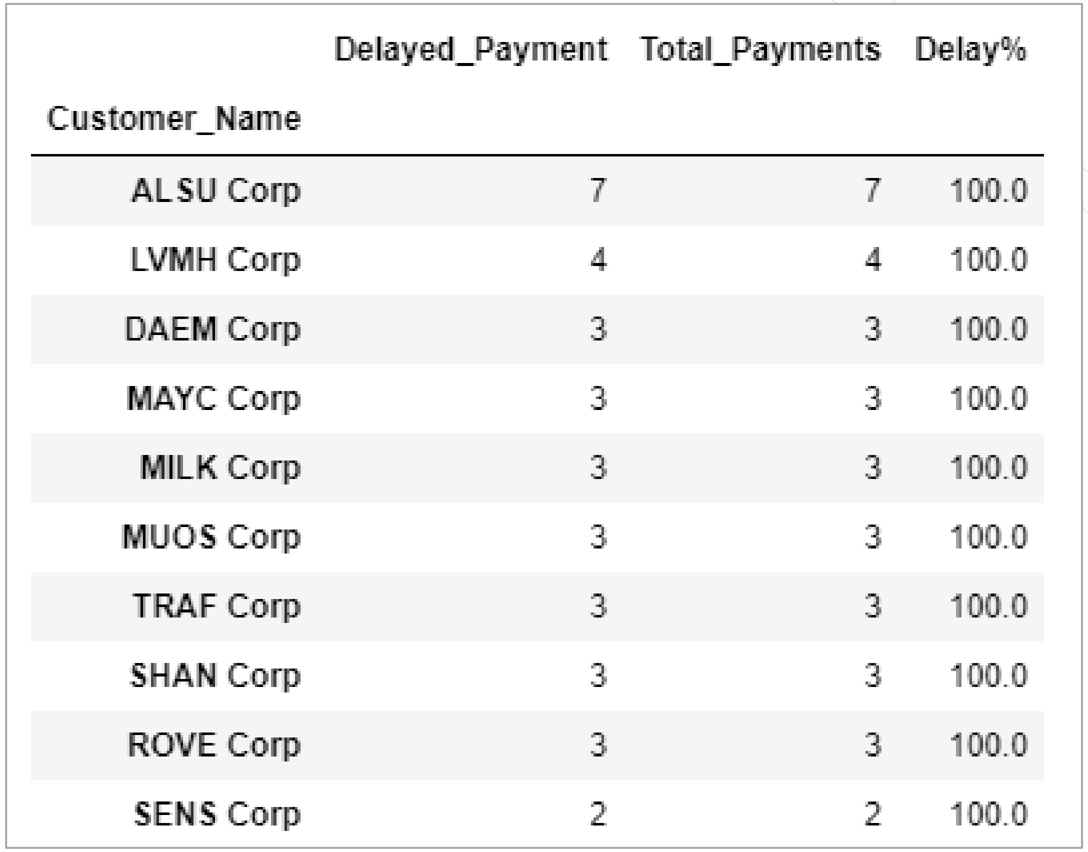
```
Feature ranking:
1. USD Amount (0.470)
2. Invoice_Month (0.134)
3. 30 Days from EOM (0.110)
4. 60 Days from EOM (0.099)
5. cluster_id (0.052)
6. Immediate Payment (0.043)
7. 15 Days from EOM (0.031)
8. 30 Days from Inv Date (0.016)
9. 60 Days from Inv Date (0.012)
10. INV (0.010)
11. 90 Days from Inv Date (0.007)
12. 90 Days from EOM (0.006)
13. 45 Days from EOM (0.005)
14. 45 Days from Inv Date (0.004)
15. DM (0.001)
```

Inference:

- Accuracy of Training data set is 95% and Accuracy of test data set is : 92%
- F1-score for train & test set is 0.96 and 0.93 respectively, implying it is a good model. Hence moving forward with this as final model for prediction.

CONCLUSIONS

- In 1st part, customer segmentation is done by Clustering Algorithm which tells us that Early late payers comprise of 88.3% hence should be proactively connected with
- Later in 2nd part, 2 models viz Logistic Regression and Random Forest are built, which predicted 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 RF model to make proactive calls to the customers to have them pay their invoice amount on time.
- According to the prediction, customers have maximum probability to default with maximum number of delayed payments. These companies should be focused more on payment collections



	Delayed_Payment	Total_Payments	Delay%
Customer_Name			
ALSU Corp	7	7	100.0
LVMH Corp	4	4	100.0
DAEM Corp	3	3	100.0
MAYC Corp	3	3	100.0
MILK Corp	3	3	100.0
MUOS Corp	3	3	100.0
TRAF Corp	3	3	100.0
SHAN Corp	3	3	100.0
ROVE Corp	3	3	100.0
SENS Corp	2	2	100.0

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 Early, medium and Prolonged payment duration respectively. It is observed that customers in cluster 2 (prolonged days) had significantly higher delay rates than early and medium days of payment, hence cluster 2 customers should be paid extensive focus

THANK YOU

upGrad

iiit-b
ज्ञानमुत्तमम्