Schuster Retail B2B Case Study

DSC 55 Batch

Submission on 05-12-2023

by

Bhawana Singh

Dhiraj Kumar Nayak

Priya G Rao

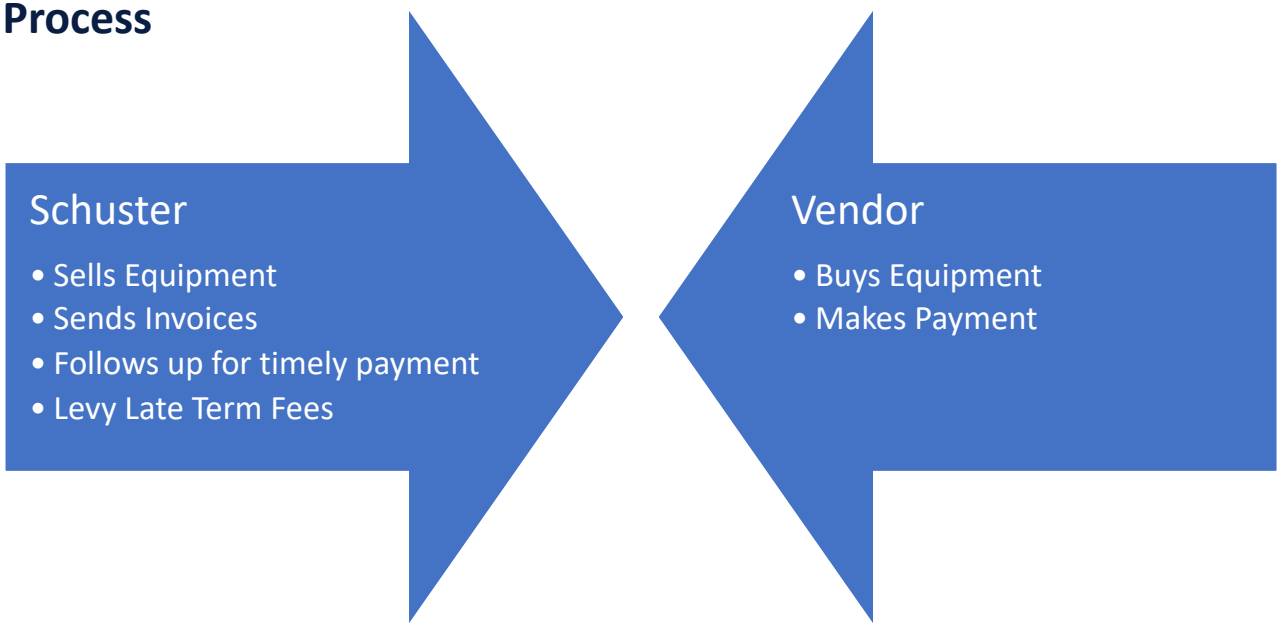
Schema

- Business Understanding and Objective
- Approach
- EDA Insights
- Customer Segmentation
- Logistic Regression Model
- Random Forest Model
- Model Comparison and Selection
- Predictions on Deployment
- Customers to Focus/Recommendations

Business Understanding and Objective

Schuster is a multinational retail company selling sports goods and accessories to its vendors on credit. Unfortunately, not all vendors respect credit terms and some of them tend to make payments late.

Process

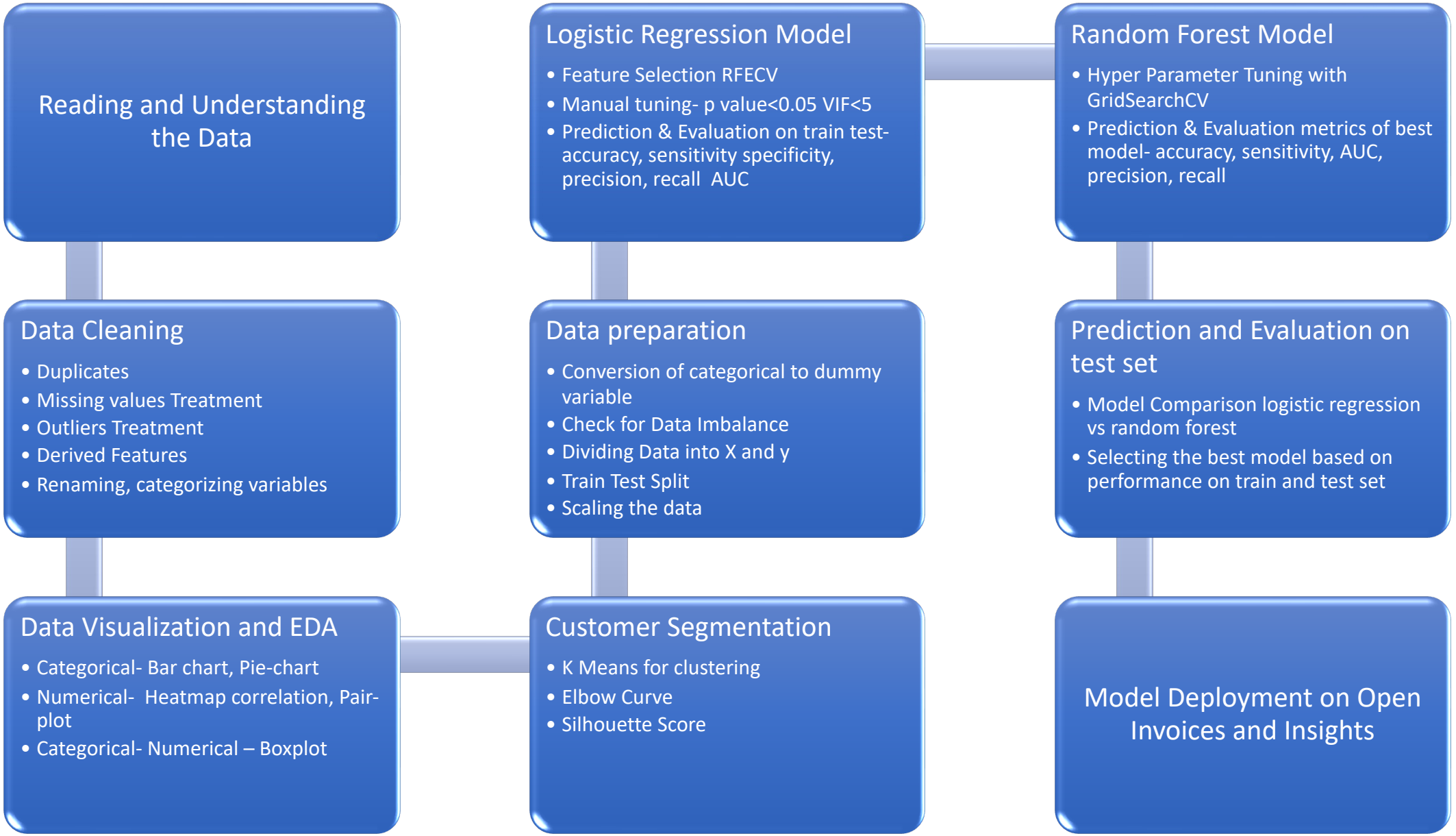


Objective

- Understand the customers' payment behaviour based on their past payment patterns (customer segmentation).
- Using historical information, Schuster wants 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

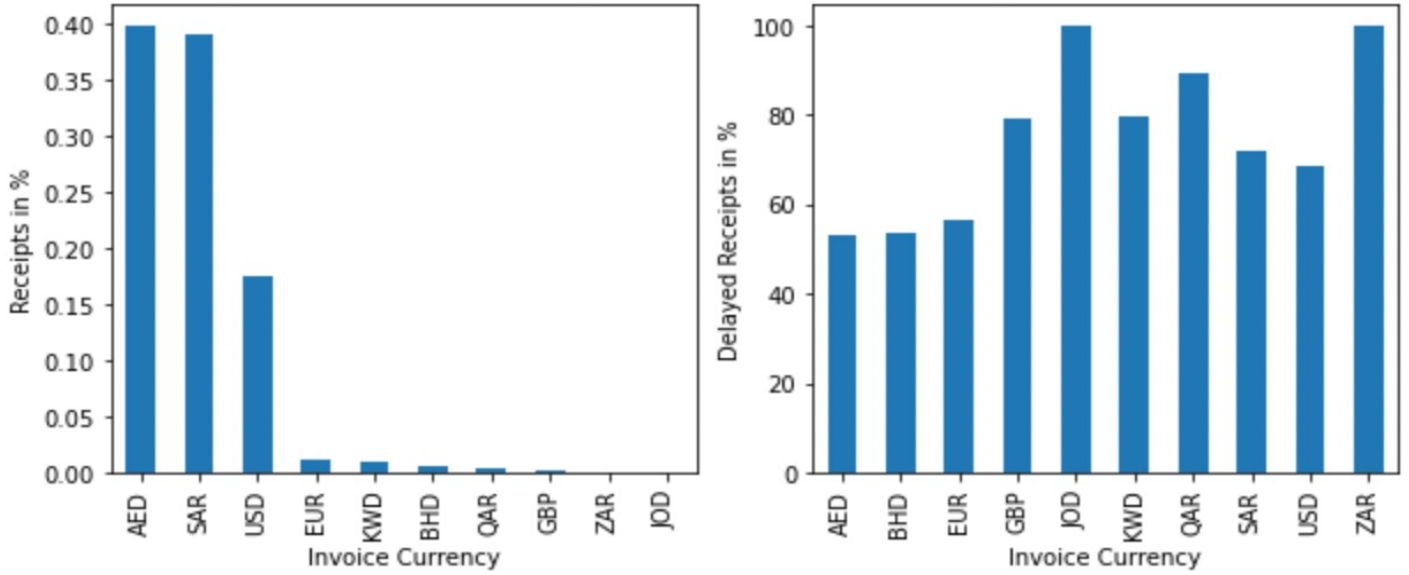
Requirement: To build a classification model with the primary objective of identifying important predictor attributes that will help the business understand indicators of late payment

Approach


Reading and Understanding the Data
Data Cleaning
• Duplicates
• Missing values Treatment
• Outliers Treatment
• Derived Features
• Renaming, categorizing variables
Data Visualization and EDA
• Categorical- Bar chart, Pie-chart
• Numerical- Heatmap correlation, Pair-plot
• Categorical- Numerical – Boxplot
Customer Segmentation
• K Means for clustering
• Elbow Curve
• Silhouette Score
Data preparation
• Conversion of categorical to dummy variable
• Check for Data Imbalance
• Dividing Data into X and y
• Train Test Split
• Scaling the data
Logistic Regression Model
• Feature Selection RFECV
• Manual tuning- p value<0.05 VIF<5
• Prediction & Evaluation on train test- accuracy, sensitivity specificity, precision, recall AUC
Random Forest Model
• Hyper Parameter Tuning with GridSearchCV
• Prediction & Evaluation metrics of best model- accuracy, sensitivity, AUC, precision, recall
Prediction and Evaluation on test set
• Model Comparison logistic regression vs random forest
• Selecting the best model based on performance on train and test set
Model Deployment on Open Invoices and Insights

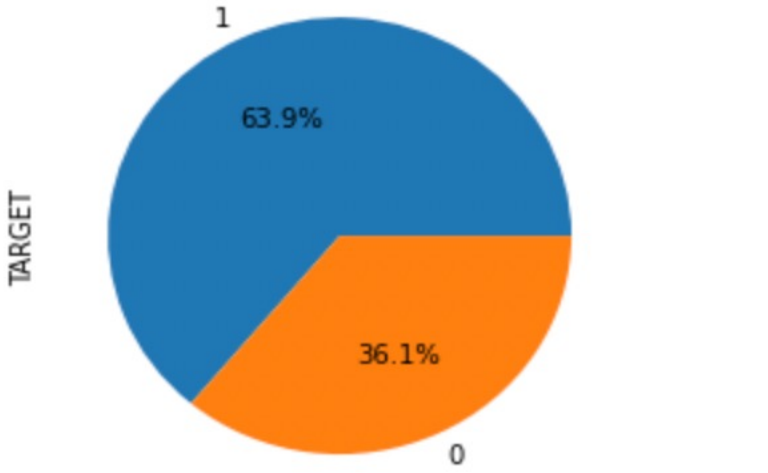
EDA Insights

Bar Chart showing the currency of invoice and % of delayed receipt by invoice currency

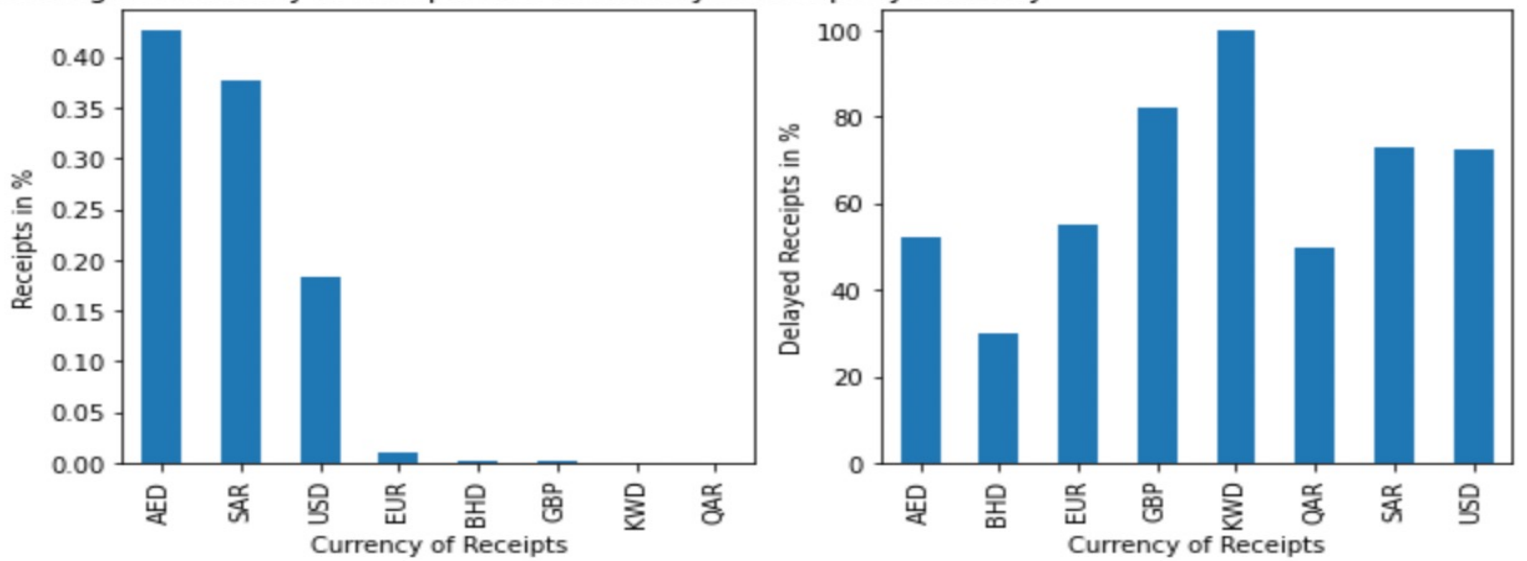


- No major data imbalance in the target variable(64:36)
- Invoices are mainly raised in AED,SAR and USD.
- Almost 100% delay where invoicing currency is JOD or ZAR
- Delays are high where payments are received in KWD followed by GBP

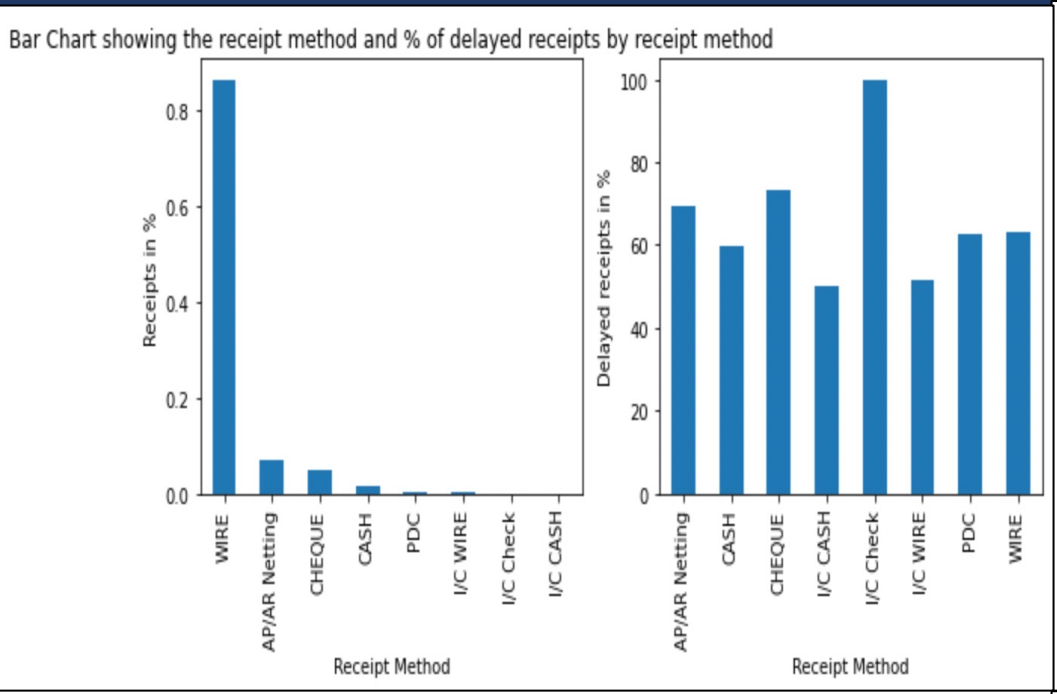
Pie Chart showing the split of the target variable

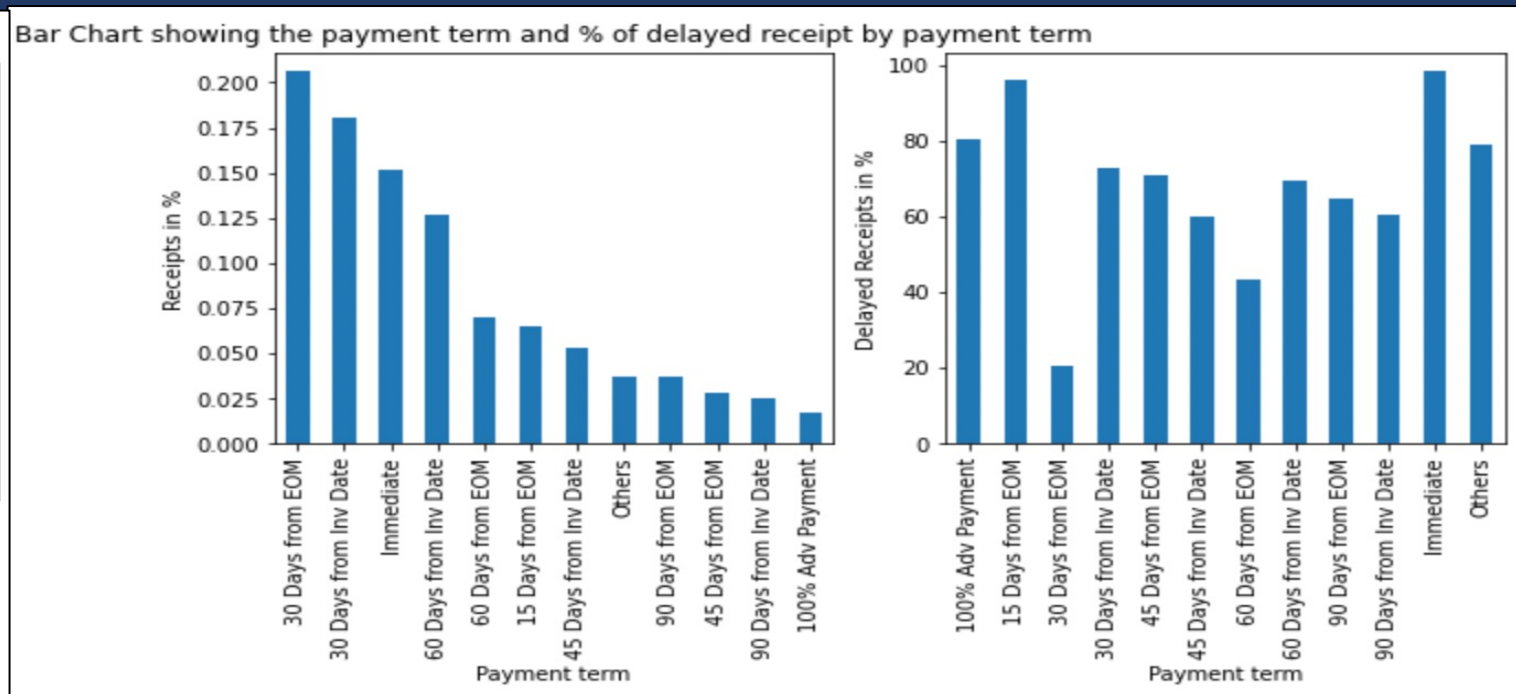


Bar Chart showing the currency of receipts and % of delayed receipt by currency

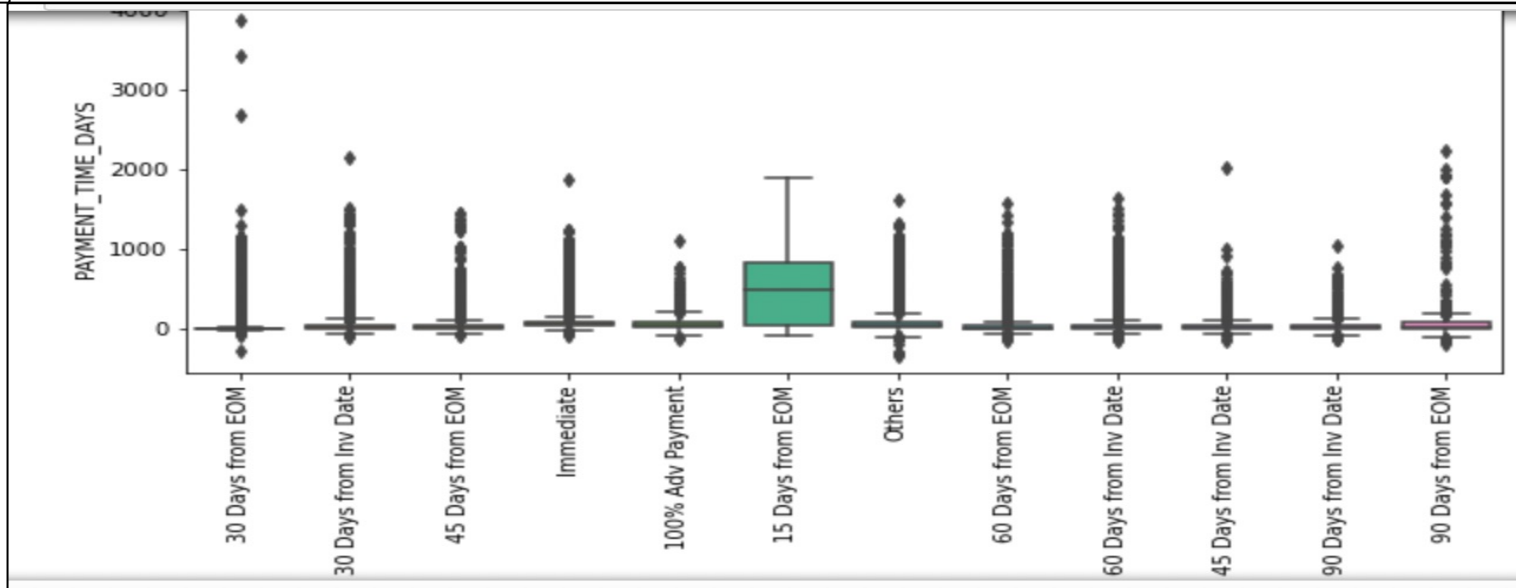


EDA Insights

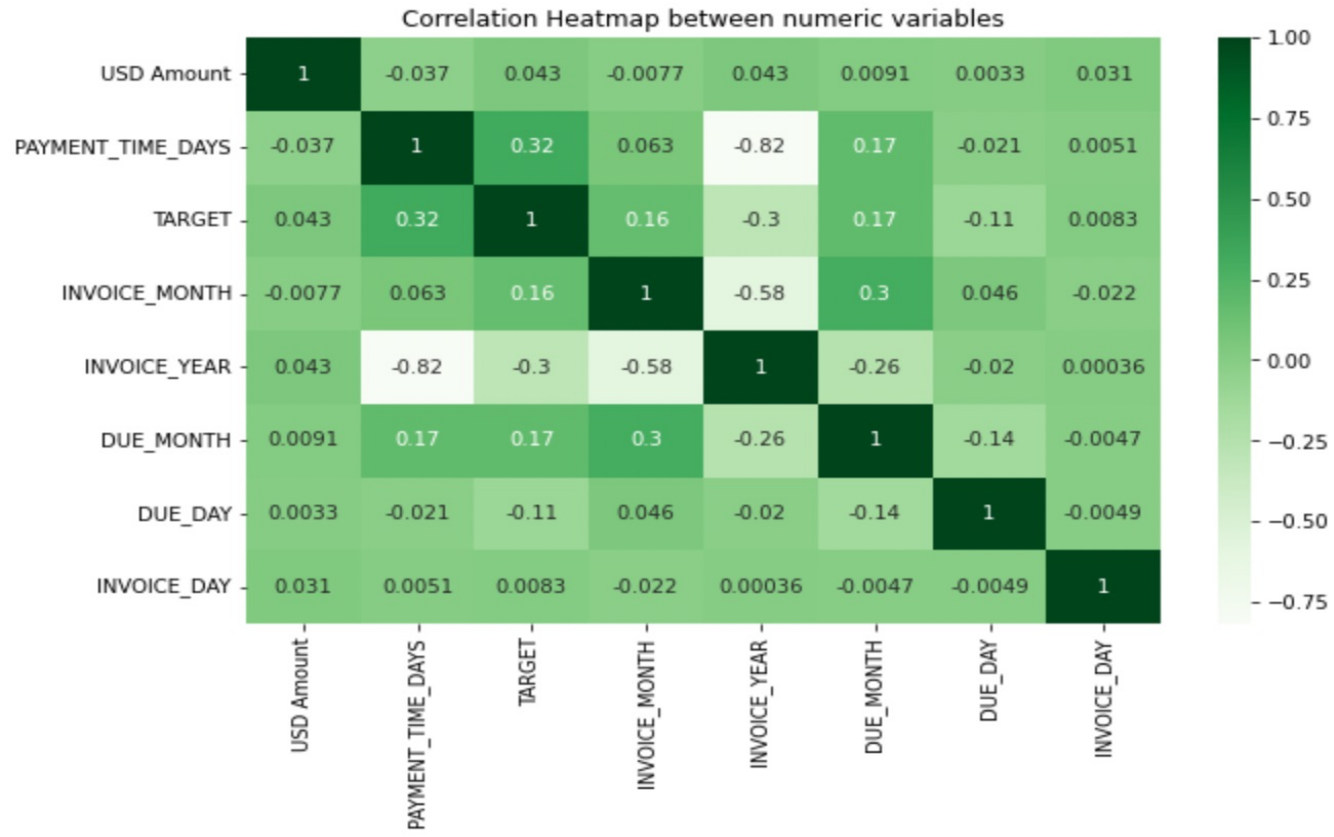


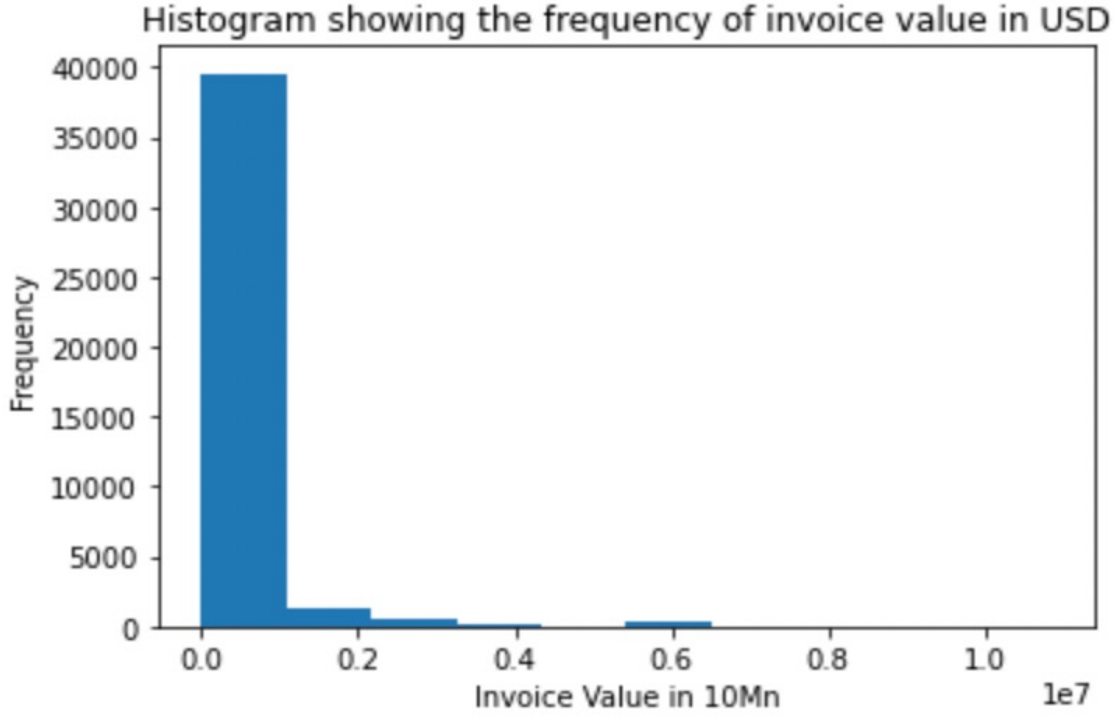


- WIRE is popular method of payment, but delays are high in case of L/C Check
- Payment delays are high in case payment terms are 15 days from EOM or Immediate
- The median value of time taken for payment is high in case of 15 days from EOM



EDA Insights

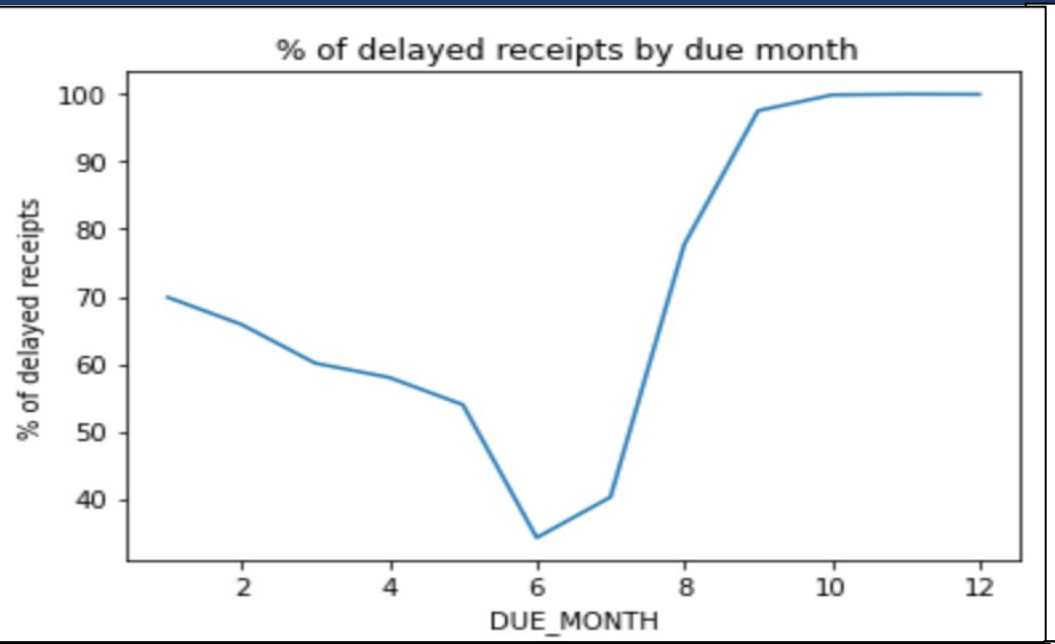


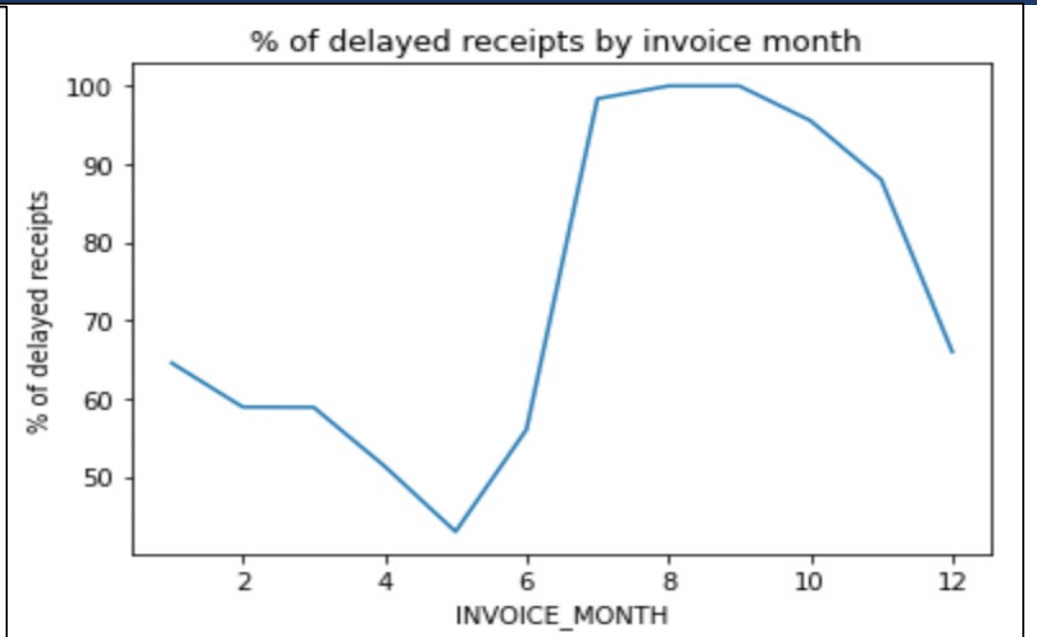


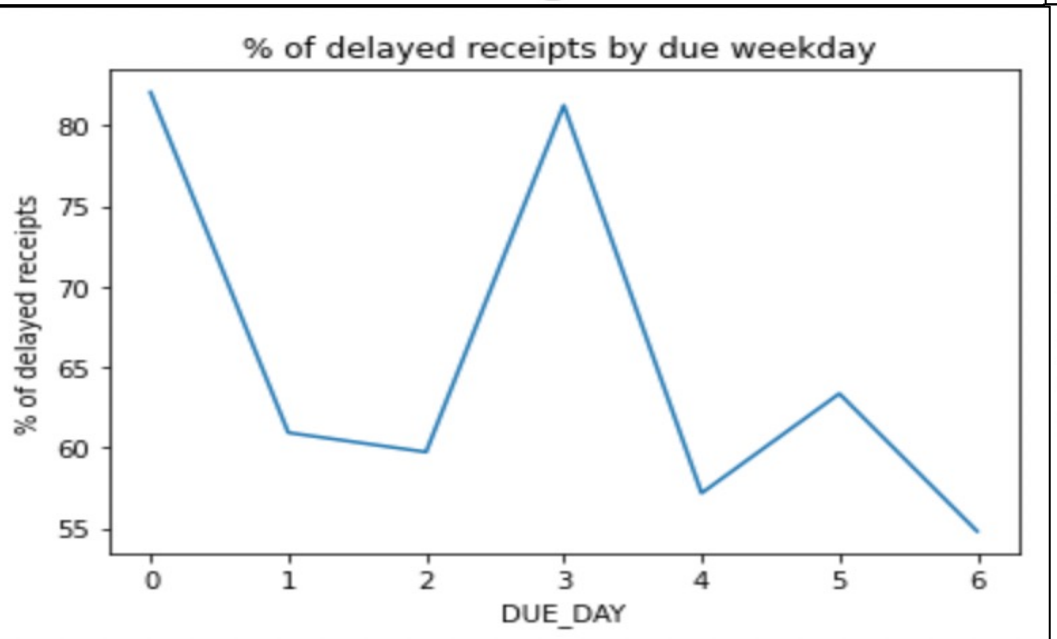
- Most of Schuster's Billings are for value less than 1Mn
- Schuster has few high value customers
- No strong correlation found among the numeric variables

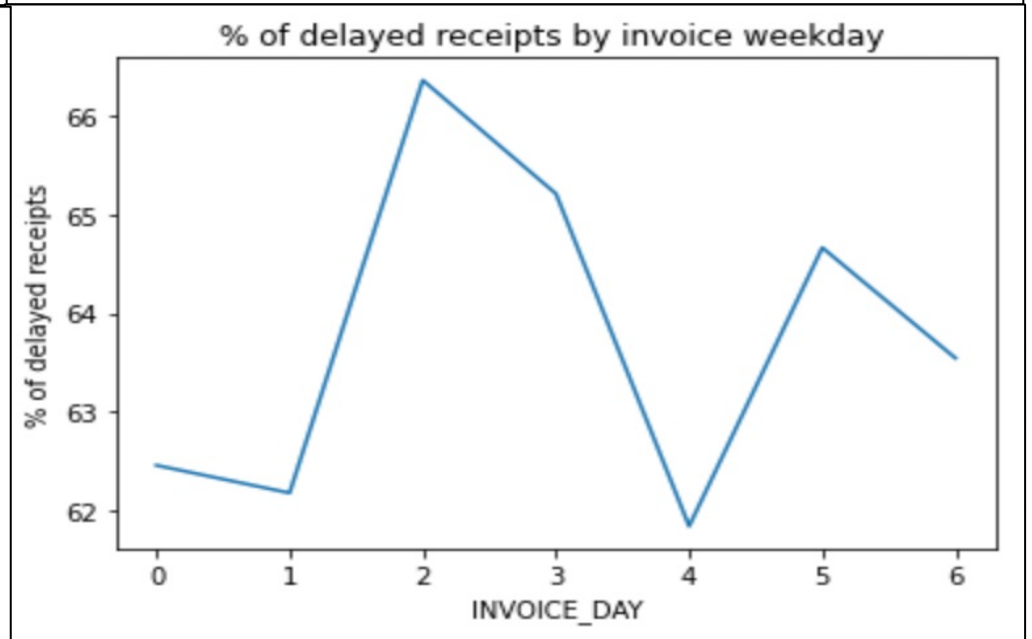
CUSTOMER_NAME	
SEPH Corp	6.153360e+09
ALLI Corp	1.652239e+09
FARO Corp	1.492450e+09
PARF Corp	6.955001e+08
CGR Corp	2.508334e+08
RADW Corp	2.430579e+08
HABC Corp	2.323890e+08
AREE Corp	2.012016e+08
PCD Corp	1.426100e+08
DUBA Corp	1.003589e+08

EDA Insights



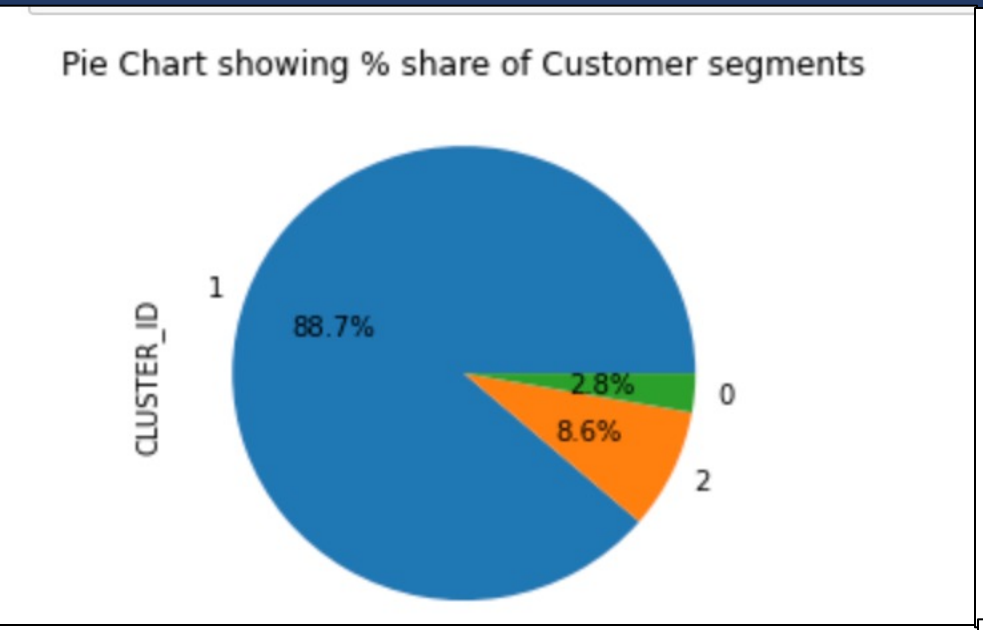






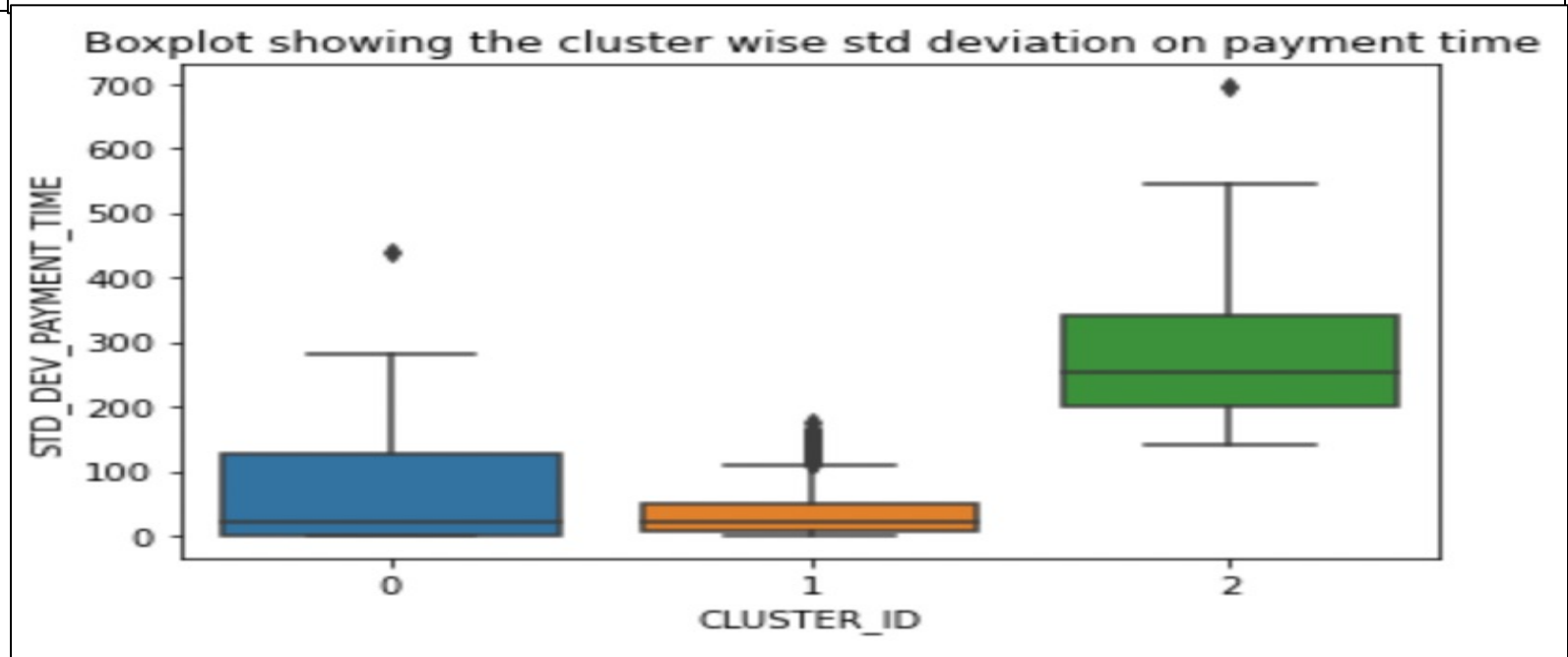
- Significant delays in receipts where invoices pertain to Jul-Sep or invoices due in the period Oct-Dec
- Invoices due on Monday and Thursday of the week have been delayed by the customer
- % of delayed receipts is high where invoices are raised on Wednesday
- *0 refers to Monday and 6 refers to Sunday*
- *Months Jan-Dec have the numbers 1-12*

Customer Segmentation



- Clusters
  - Cluster 0: Prolonged delays with moderate stand deviation
  - Cluster 1: Early Payments with low std deviation
  - Cluster 2: Moderate Delays with high standard deviation

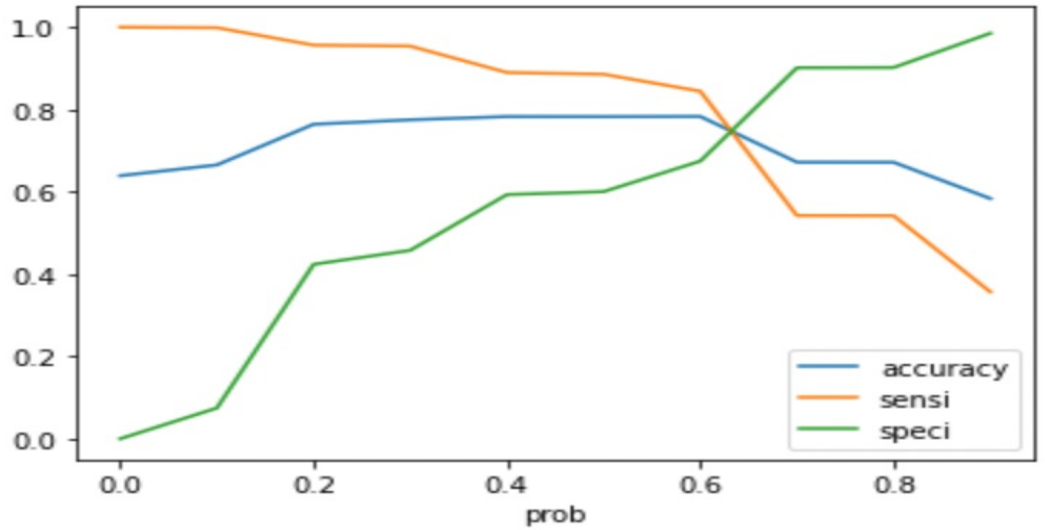
- Majority of the customers are in Cluster 1, focus should be on Cluster 0 and 2.



Logistic Regression Model

Variables considered in logistic regression model:

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.8321	0.022	37.401	0.000	0.788	0.876
PAYMENT_TERM_15 Days from EOM	2.1928	0.126	17.371	0.000	1.945	2.440
PAYMENT_TERM_30 Days from EOM	-2.2189	0.038	-57.902	0.000	-2.294	-2.144
PAYMENT_TERM_60 Days from EOM	-1.3068	0.051	-25.675	0.000	-1.407	-1.207
PAYMENT_TERM_Immediate	2.9537	0.112	26.344	0.000	2.734	3.173
DUE_MONTH_6	-1.3653	0.048	-28.488	0.000	-1.459	-1.271
DUE_MONTH_7	-1.7054	0.097	-17.627	0.000	-1.895	-1.516
DUE_MONTH_9	2.8743	0.348	8.261	0.000	2.192	3.556
DUE_MONTH_10	5.5862	1.004	5.563	0.000	3.618	7.554
CLUSTER_ID_2	0.6292	0.034	18.645	0.000	0.563	0.695



VIF <5 :

Features	VIF
const	2.787946
PAYMENT_TERM_15 Days from EOM	1.154744
PAYMENT_TERM_30 Days from EOM	1.153192
PAYMENT_TERM_60 Days from EOM	1.073112
PAYMENT_TERM_Immediate	1.121179
DUE_MONTH_6	1.024676
DUE_MONTH_7	1.010495
DUE_MONTH_9	1.011407
DUE_MONTH_10	1.011082
CLUSTER_ID_2	1.154136

prob	accuracy	sensi	speci
0.0	0.639096	1.000000	0.000000
0.1	0.665251	0.998724	0.074729
0.2	0.764164	0.956365	0.423812
0.3	0.774932	0.954026	0.457788
0.4	0.782779	0.889822	0.593224
0.5	0.782643	0.885570	0.600376
0.6	0.783084	0.844379	0.674541
0.7	0.671909	0.542280	0.901459
0.8	0.671807	0.541855	0.901929
0.9	0.583628	0.356843	0.985224

- Payment Terms 15 days from EOM and Immediate , due month 9 and 10 increases the probability of delay whereas Payment terms 30,60 days from EOM and due months 6 and 7 decrease the probability of delay.
- Optimal cut off for probability is 0.62
- Probability threshold considered as 0.5 as sensitivity is 0.89 at that level

Random Forest Model

Feature wise importance in random forest:

	VarName	Imp
2	PAYMENT_TERM_30 Days from EOM	0.411225
0	PAYMENT_TERM_Immediate	0.155492
3	DUE_MONTH_12	0.067502
1	DUE_MONTH_6	0.059655
5	PAYMENT_TERM_60 Days from EOM	0.049250
0	CLUSTER_ID_1	0.046122
4	PAYMENT_TERM_15 Days from EOM	0.031879
	CLUSTER_ID_2	0.031225
0	USD Amount	0.028208
3	PAYMENT_TERM_30 Days from Inv Date	0.017724

Parameters given for hyper parameter tuning :

```
Fitting 5 folds for each of 256 candidates, totalling 1280 fits

GridSearchCV(cv=5,
             estimator=RandomForestClassifier(n_jobs=-1, oob_score=True,
                                              random_state=42),
             n_jobs=-1,
             param_grid={'max_depth': [5, 10, 15, 20],
                         'max_features': [5, 10, 15, 20],
                         'min_samples_leaf': [200, 500, 1000, 1500],
                         'n_estimators': [25, 50, 80, 100]},
             verbose=1)
```

Parameters of the best random forest:

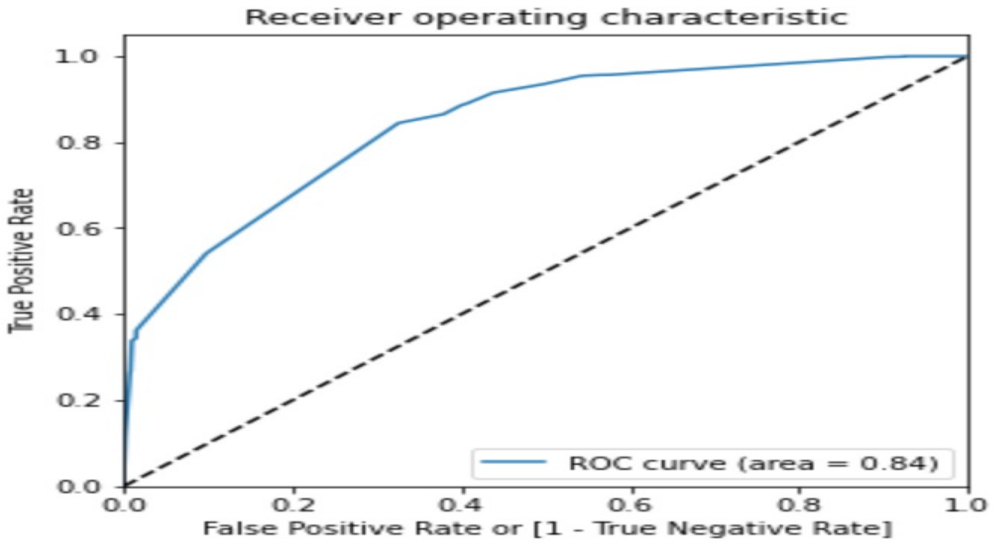
```
# finding the best random forest
rf_best=rf_cv.best_estimator_
rf_best
```

```
RandomForestClassifier(max_depth=10, max_features=10, min_samples_leaf=200,
                       n_jobs=-1, oob_score=True, random_state=42)
```

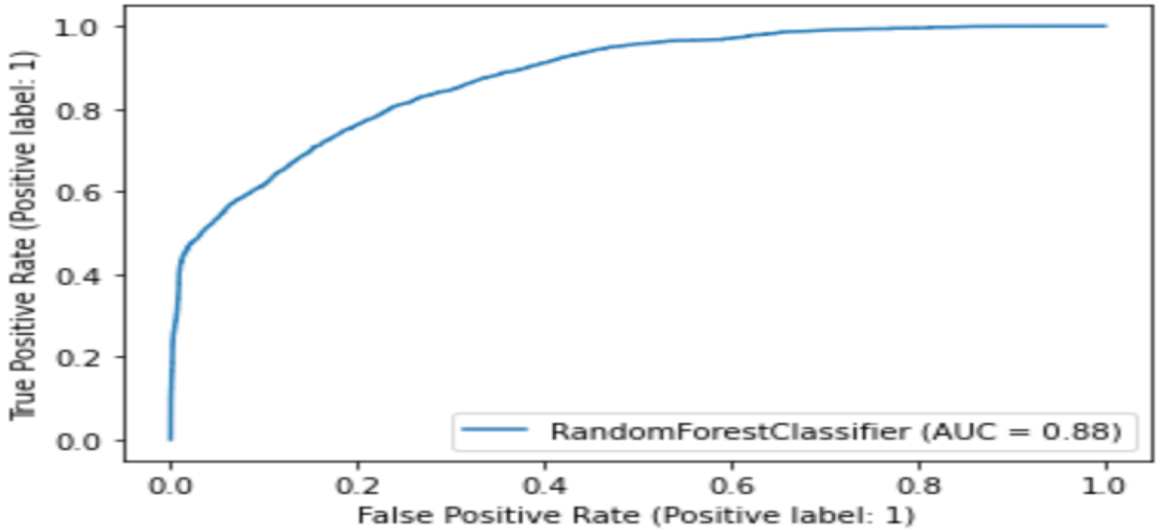
- OOB Score of the Random forest is 79.87% and cross val score 79.83% . There is no overfitting in this model.

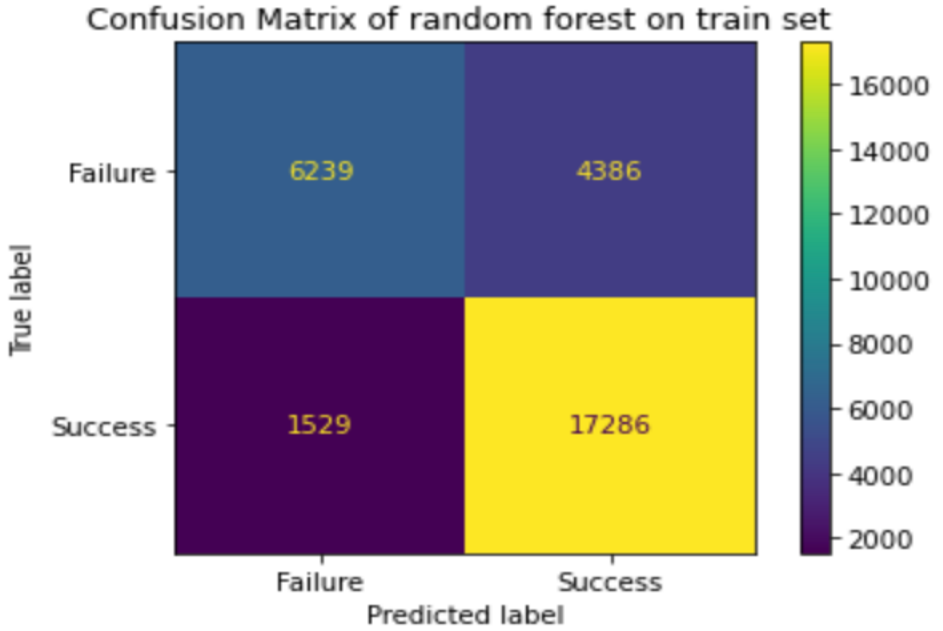
Random Forest vs Logistic Regression

Logistic Regression

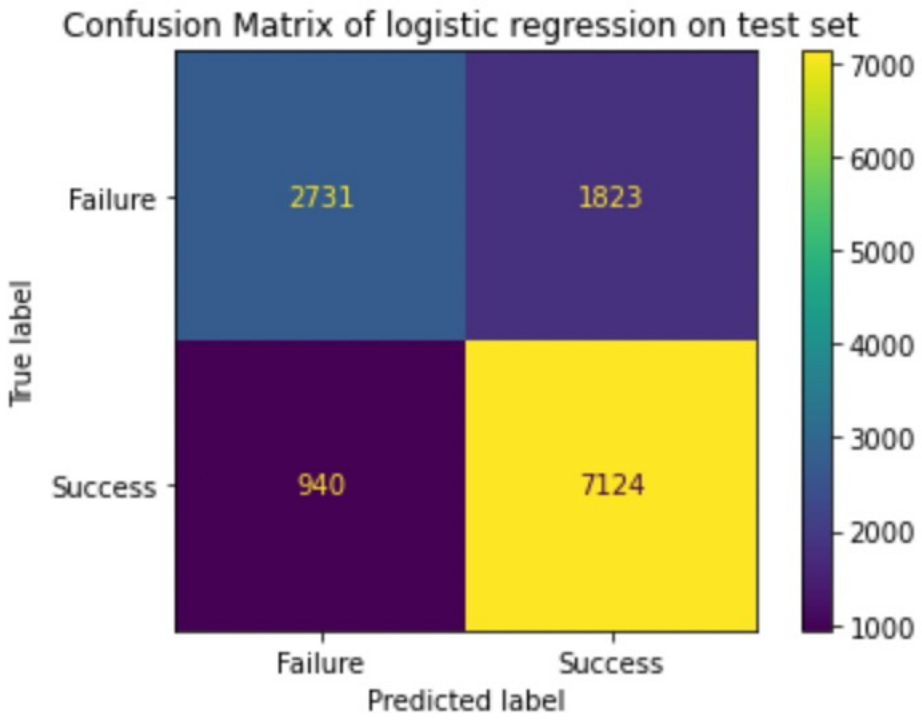


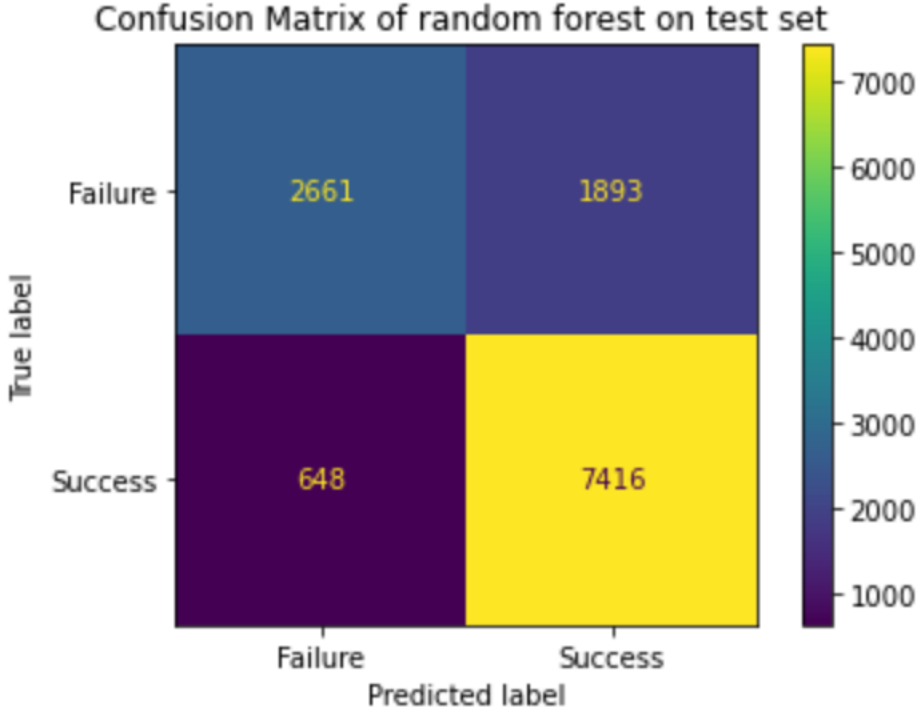
Random Forest





Random Forest vs Logistic Regression

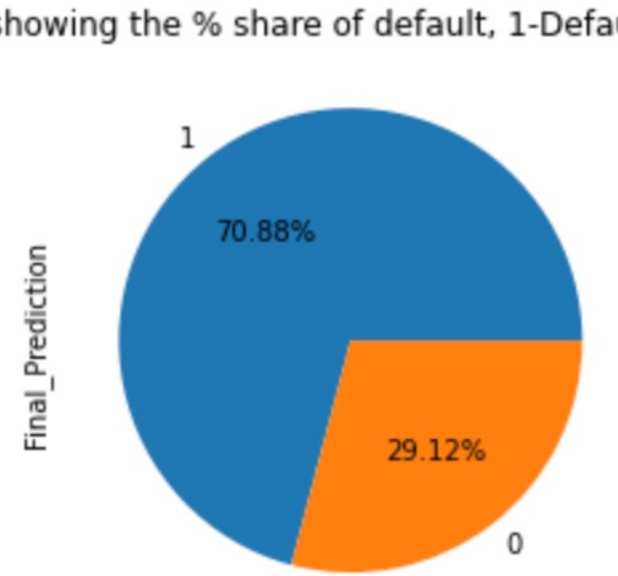


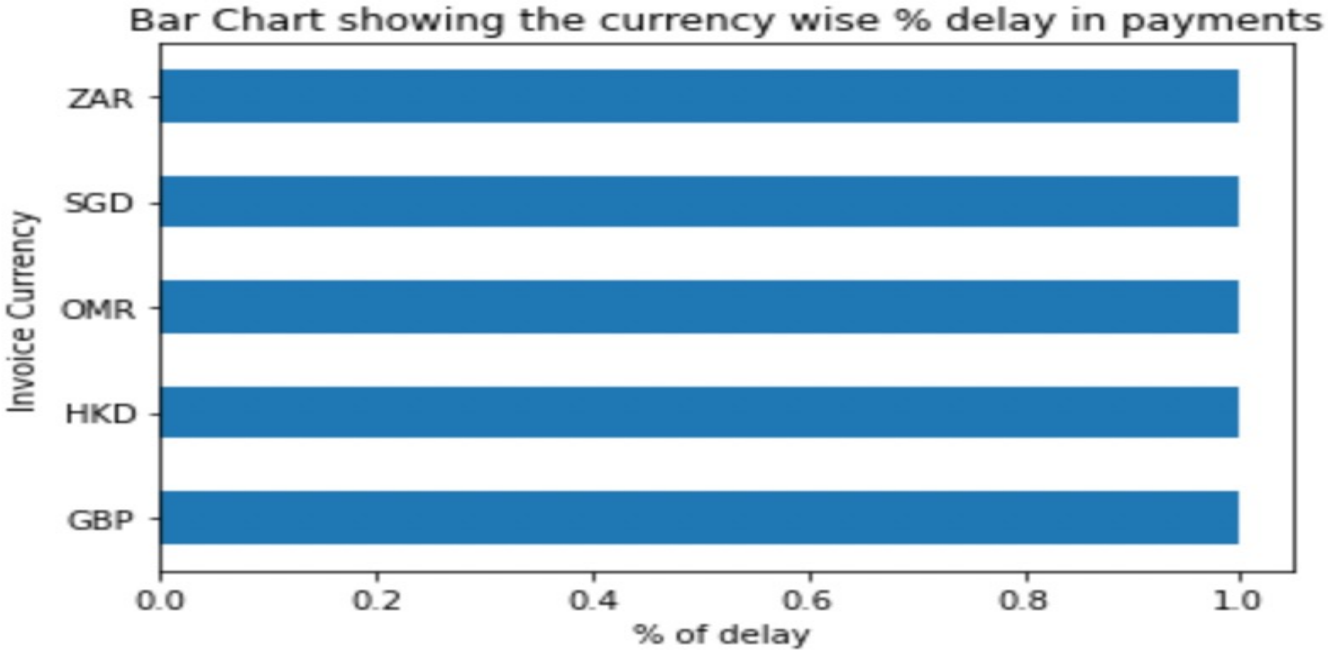


For this case study random forest is the preferred model for deployment since

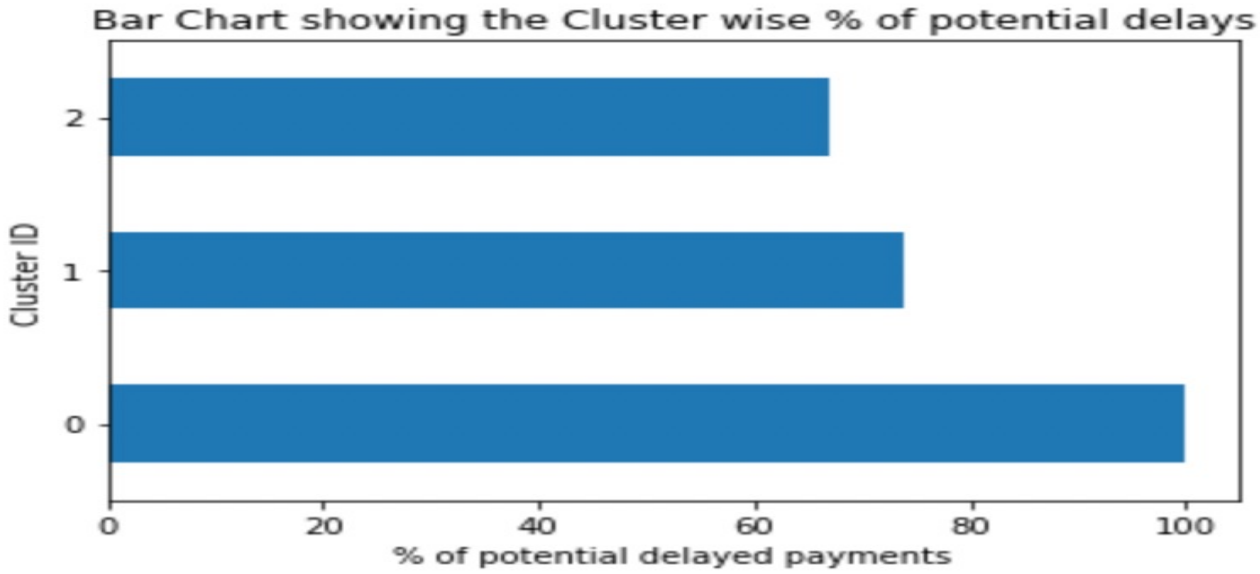
- AUC of random forest 0.88 is better than logistic regression(0.84)
- Sensitivity(Recall) of random forest is 0.92 on both train and test sets
- Sensitivity(Recall) of logistic regression is 0.89 on train and 0.88 on the test set.
- Precision 0.8 on train and test sets of Logistic Regression and Random forest
- Accuracy of random forest is also better than logistic regression.
- High sensitivity will ensure that the model will identify the cases of delayed payments much better thereby giving heads up for early followup.

Predictions on deployment

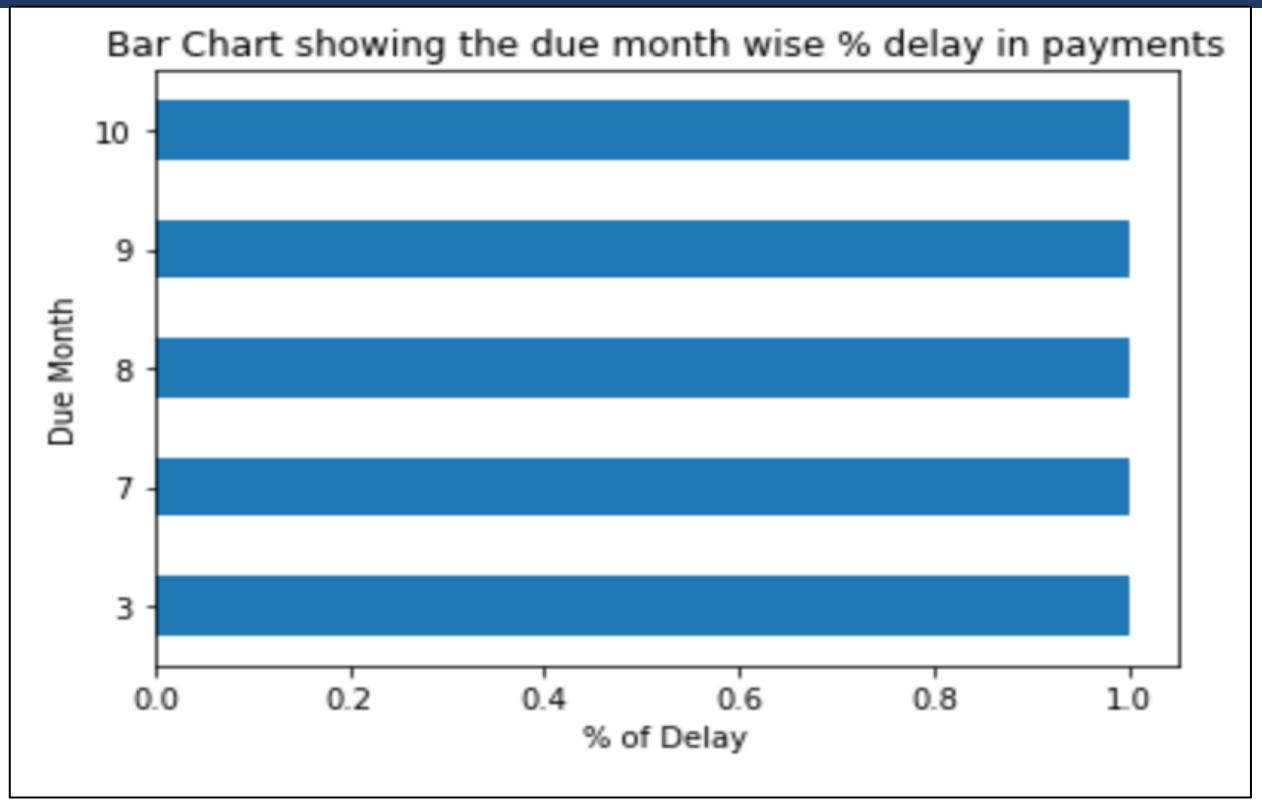


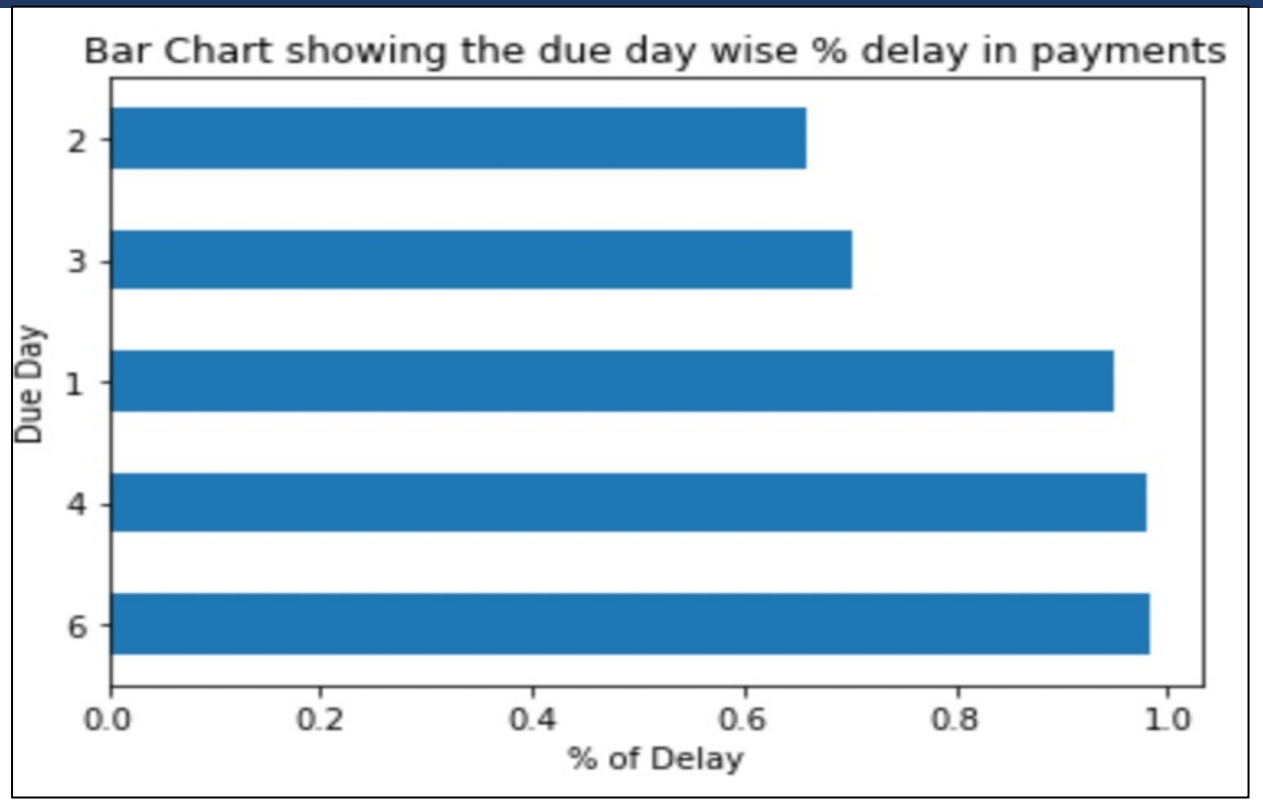


- On deployment, there's delay expected in 70.88% of the open invoices.
- 100% delays can be expected if
  - Invoice currency is GBP,HKD,OMR,SGD and ZAR
  - Invoice due on friday or sunday
  - Customers belonging to Cluster ID 0



Predictions on deployment





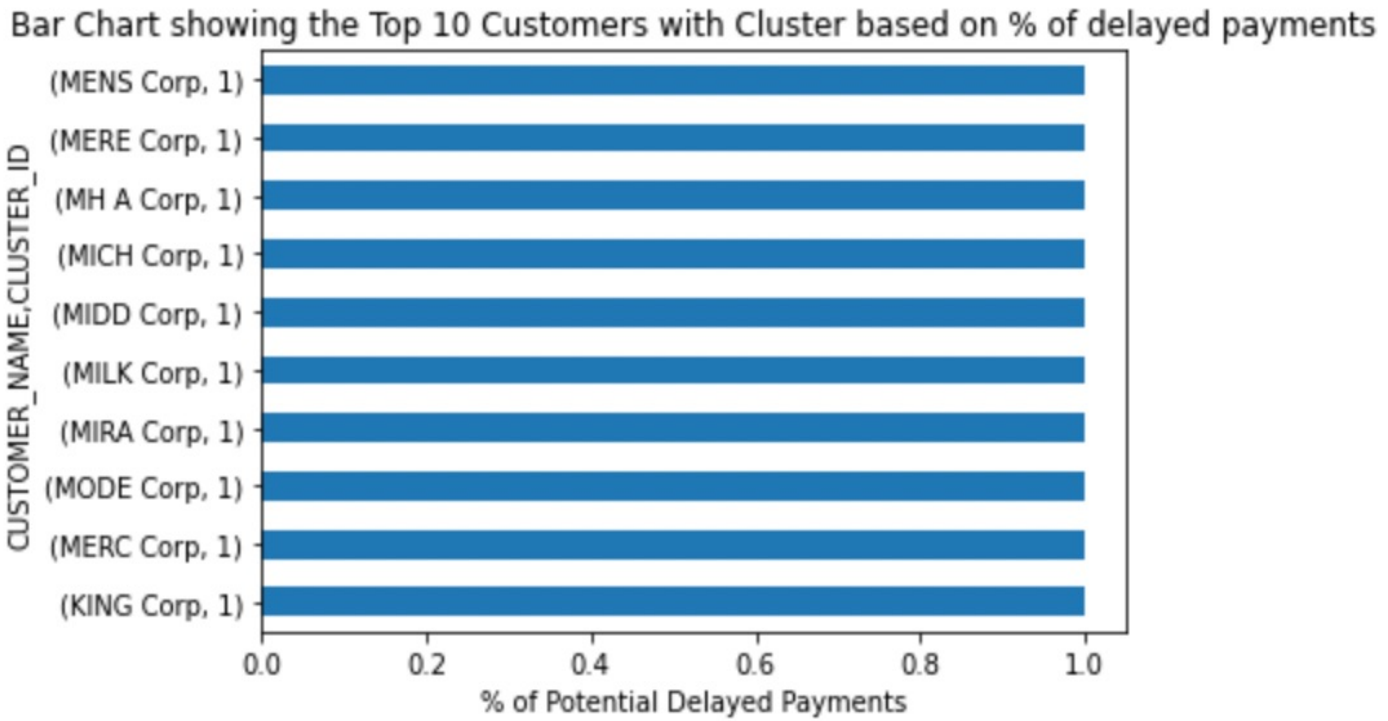
- Invoices due in March, Jul-Oct period will need advance follow up
- Delays can be expected on Invoices due on Sunday and Friday, proactive follow up required.

 # In Due Day 0 refers to Monday and 6 refers to Sunday
 # Months Jan-Dec have the numbers 1-12

Customers to Focus

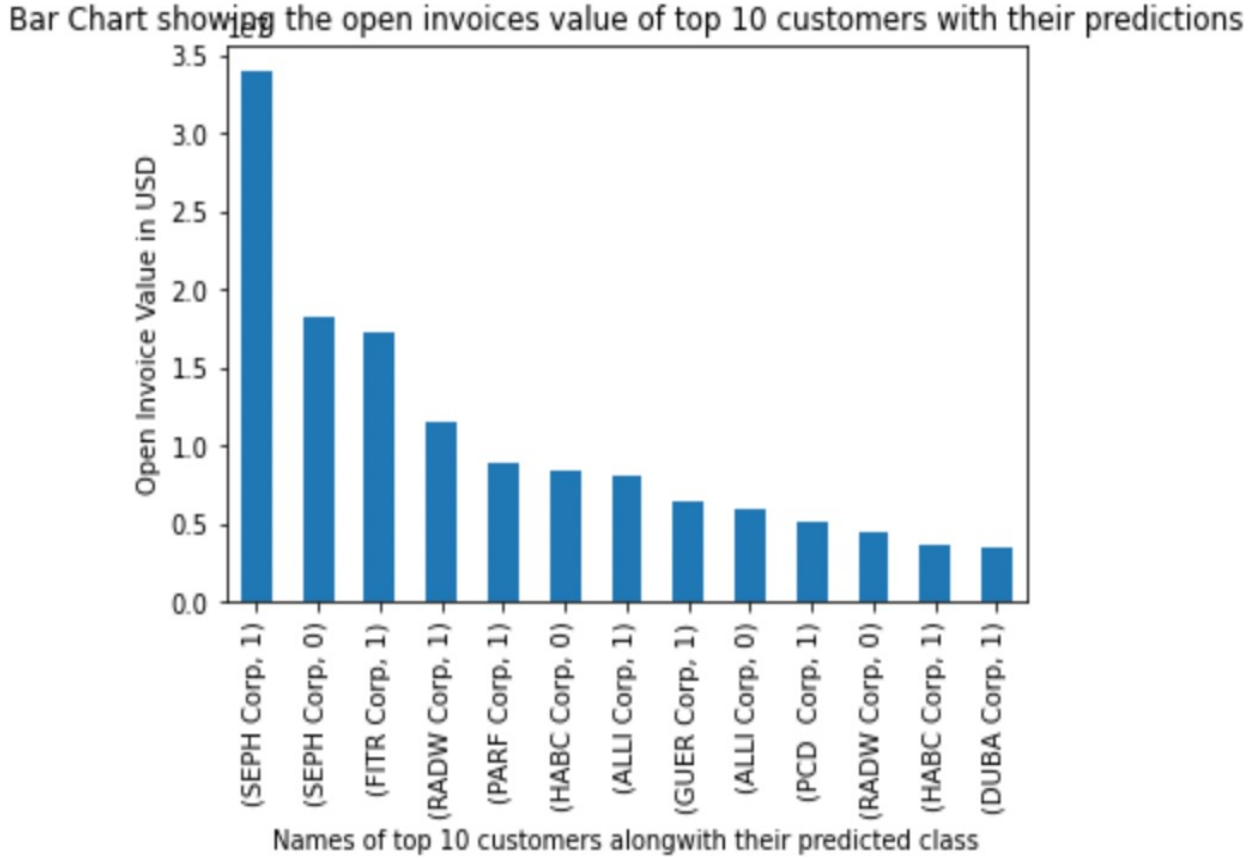
CUSTOMER_NAME	AVERAGE_PAYMENT_DAYS	STD_DEV_PAYMENT_TIME	CLUSTER_ID
ADMI Corp	575.384615	73.868733	0
ALAM Corp	459.000000	0.000000	0
ALSU Corp	565.500000	198.545405	0
ANTH Corp	553.625000	168.085809	0
BASI Corp	1176.666667	440.380801	0
EYEW Corp	917.000000	0.000000	0
HANI Corp	1142.000000	281.428499	0
I BE Corp	672.000000	21.213203	0
JUBA Corp	680.000000	0.000000	0
LINE Corp	517.000000	0.000000	0
MAY Corp	1296.500000	21.920310	0
NOUS Corp	609.000000	0.000000	0
QAWA Corp	834.600000	36.329052	0
QURA Corp	702.000000	22.627417	0
RABL Corp	504.222222	81.628392	0
REEM Corp	577.000000	0.000000	0
SAEE Corp	523.500000	13.435029	0
SAIF Corp	413.666667	26.102363	0
UAE Corp	840.727273	281.541503	0

- Focus on
  - customers in Cluster ID 0 as the delays in payment are prolonged historically
  - Customers who will default on all invoices as given below



Customers to focus

CUSTOMER_NAME	Number of Open Invoices	% of Predicted Delayed Invoices	Delayed Invoices Count
SEPH Corp	8260	0.648305	5355.0
FITR Corp	3454	0.866532	2993.0
PARF Corp	1717	0.845079	1451.0
AREE Corp	1117	0.482543	539.0
ALLI Corp	1042	0.435701	454.0
HABC Corp	517	0.696325	360.0
AL T Corp	584	0.578767	338.0
DEBE Corp	654	0.481651	315.0
RADW Corp	490	0.589796	289.0
CARR Corp	363	0.796143	289.0



- Customers with large volume of invoices will need better tracking
- Follow up with Customers having high value billing will help reduce the receivables

Recommendations

Focus Areas for invoice collection

- Dedicated person to manage the follow up of receivables of high value customers to maintain relationship

- Close monitoring of receivables of customers who belong to Cluster 0 who have historically delayed payments for long periods

- Focus on top customers who have high volumes in billing

- Receivables that fall due on Sunday or Friday or due in the period Jul-Oct or March will need advance follow up

- Schuster will need to understand the weekly payment cycle(day of the week when customer releases payment) of the customers to effectively follow up payments in advance.

Thank You