E-Commerce and Retail B2B Case Study

Shamkumar Yedelli

Problem Statement

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.

To understand how to approach this problem using data science, let's first understand the payment process at Schuster now. 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.

Business Goals

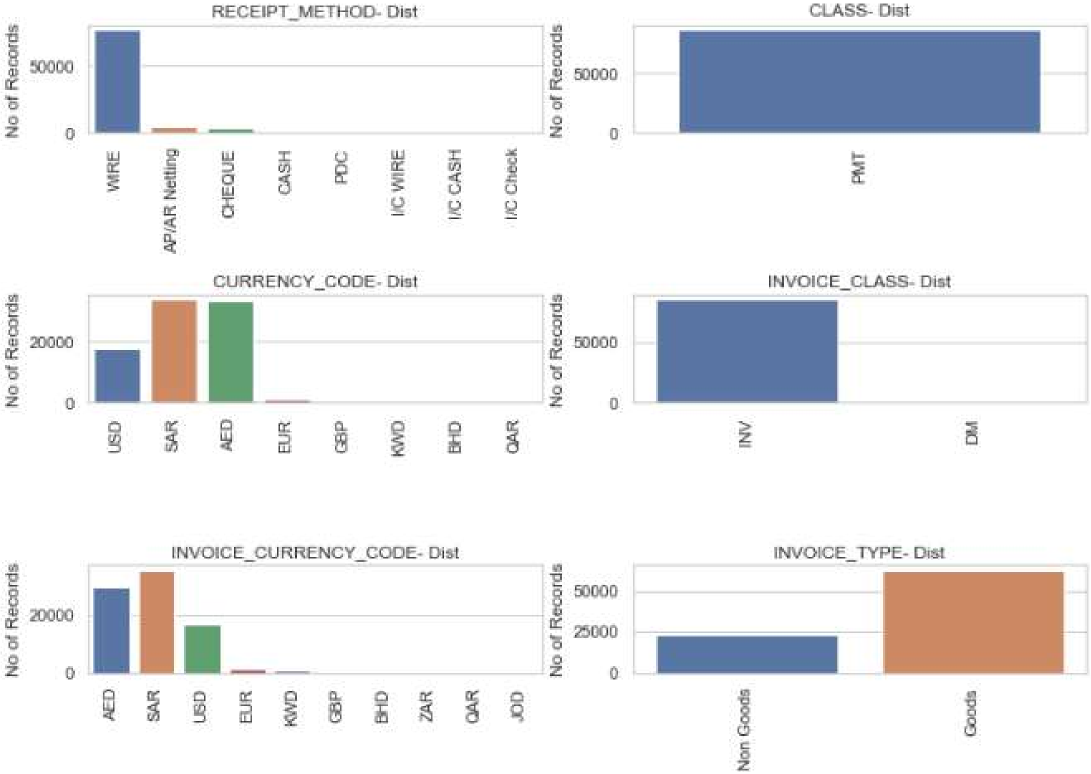
A classification model is to be built for the company which can be used by the company to target the potential customers.

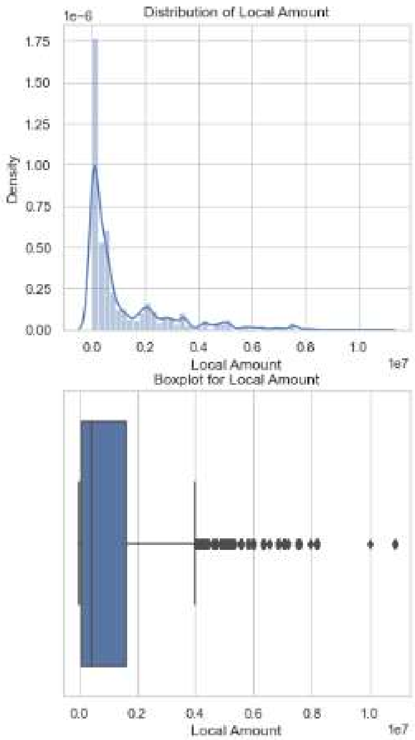
Goals:

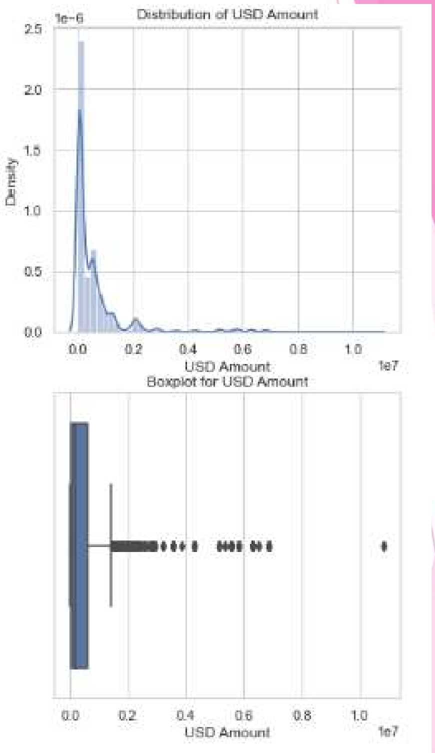
1) Schuster would like to better understand the customers' payment behaviour based on their past payment patterns (customer segmentation).
2) Using historical information, it wants to be able to predict the likelihood of delayed payment against open invoices from its customers.
3) It wants to use this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

Exploratory Data Analysis

Distribution of Categorical Variables





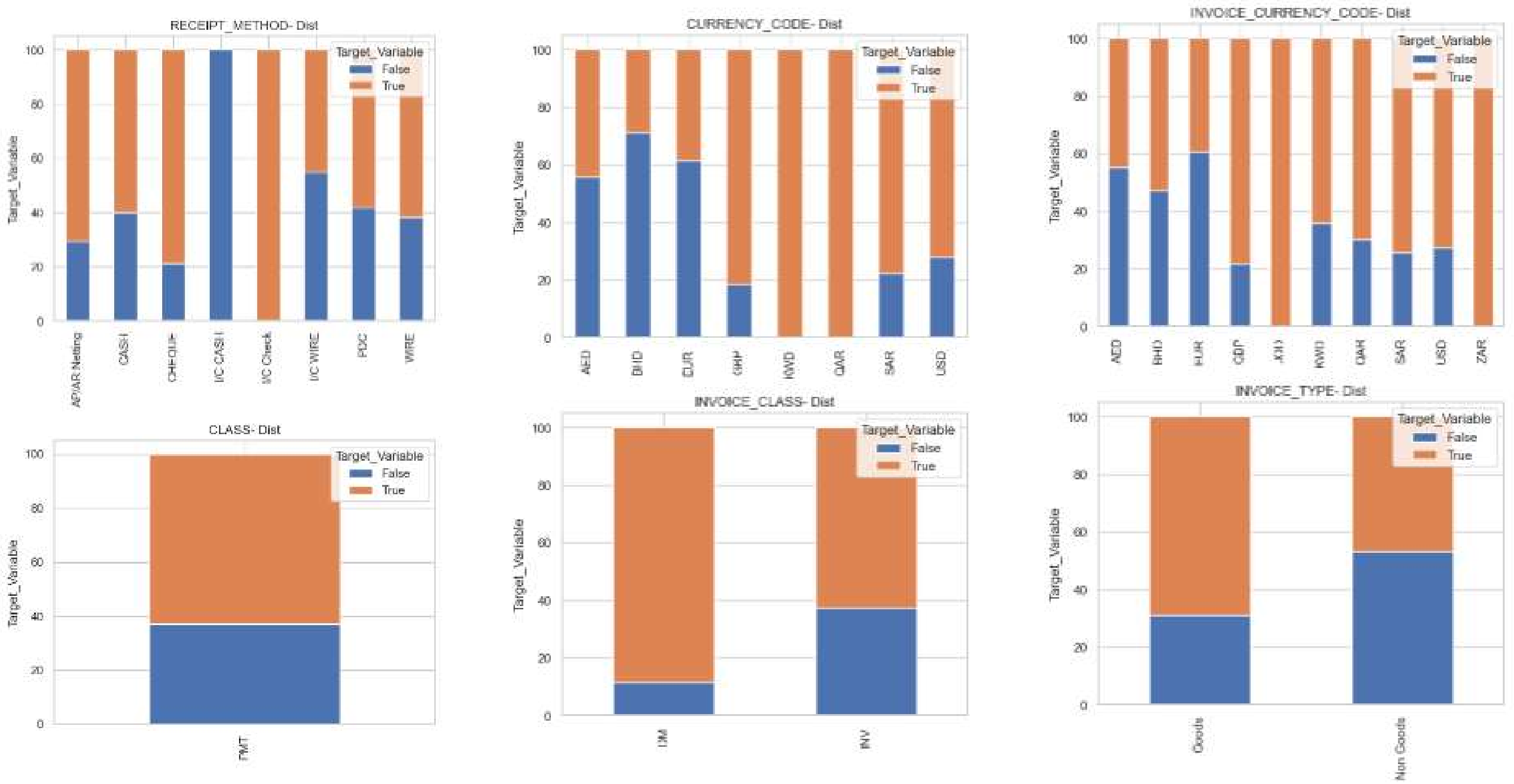


Distribution of Local Amount

As seen from above, we can see potential outliers in the dataset which are then further removed

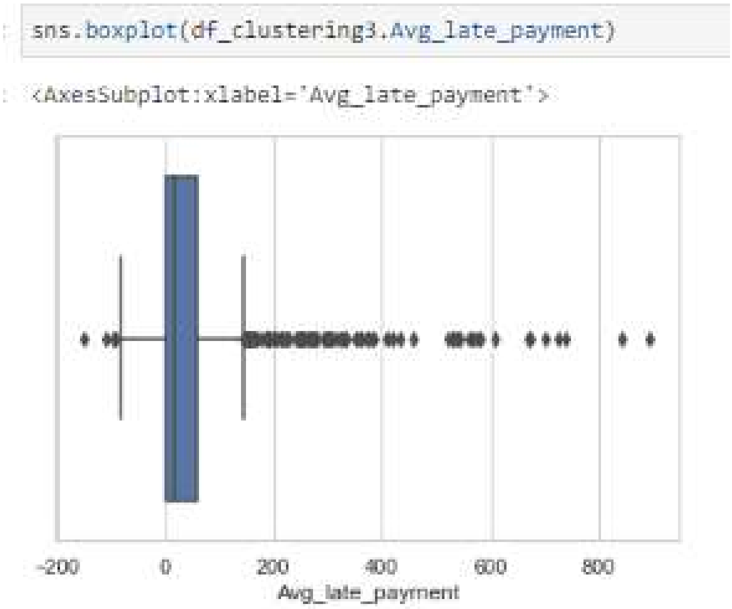
Distribution of USD Amount

As seen from above, we can see potential outliers in the dataset which are then further removed



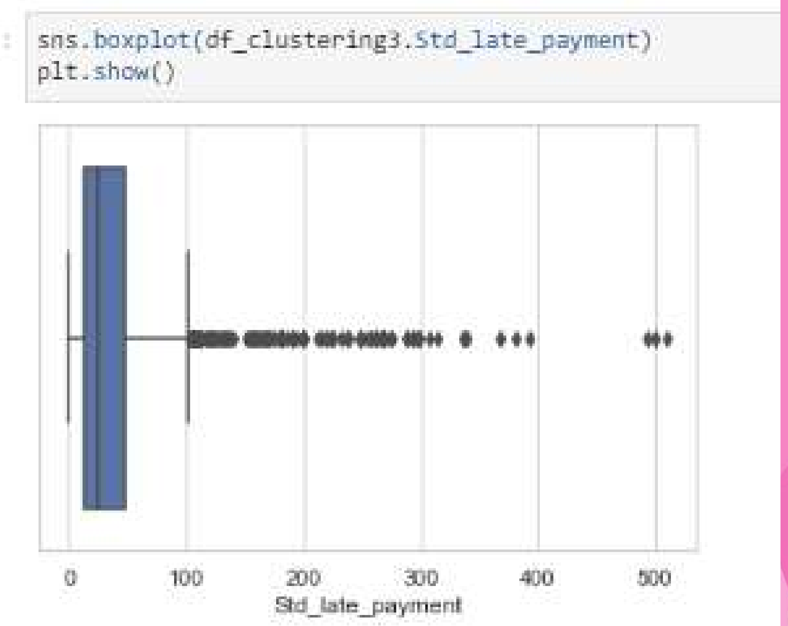
Target Variable Distribution basis different categorical columns. As seen above, it gives us clear indication about the target variables among the different data variables.

Customer Segmentation basis Customer Behavior and Analysis



```
sns.boxplot(df_clustering3.Avg_late_payment)
```

```
<AxesSubplot:xlabel='Avg_late_payment'>
```



```
sns.boxplot(df_clustering3.Std_late_payment)
plt.show()
```

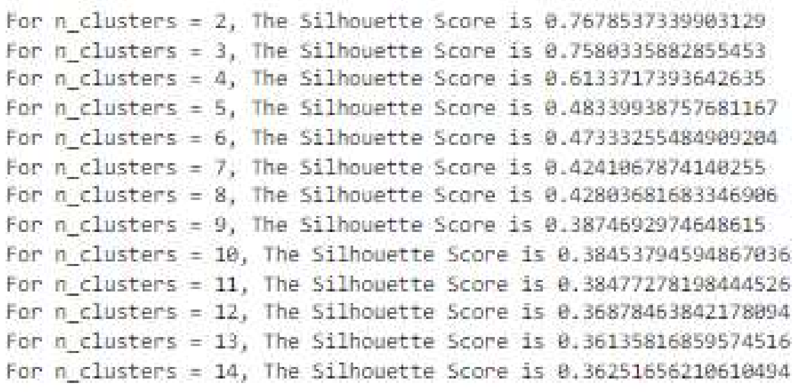
K Means Clustering

First, we scale the variables of the customer segmented columns (Avg_late_payment, Std_late_payment) using Standard Scaler and then fit transform the data.

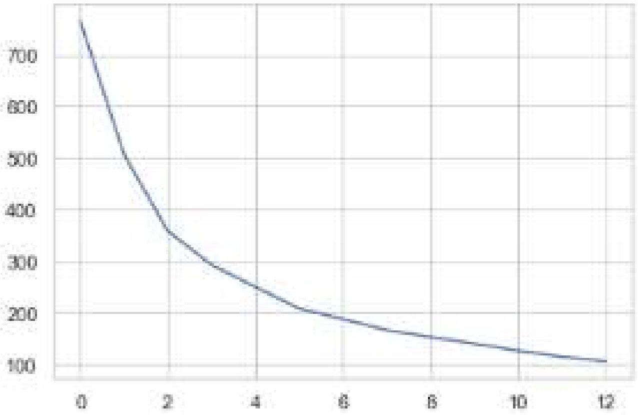
We then used K-Means Clustering to cluster the dataset into segments

Here is the elbow curve:

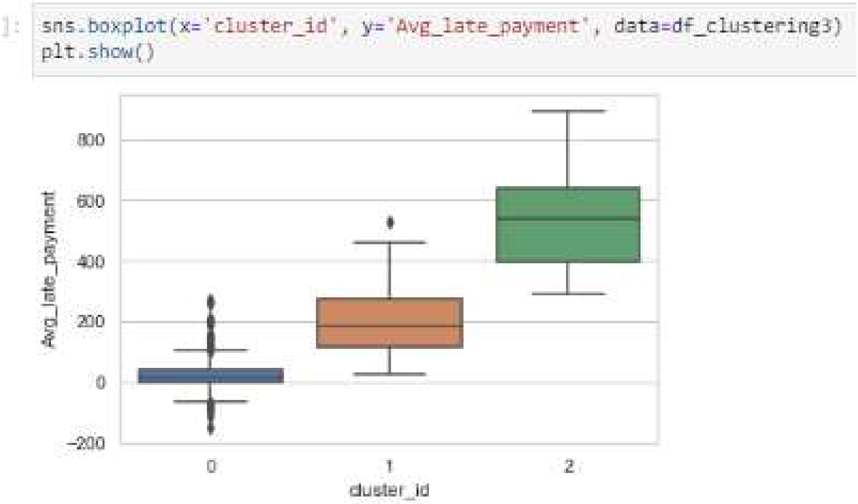
Then we did the Silhouette Analysis and finally to go ahead with building the final model we finalized k = 3 (no of clusters = 3) as per the Silhouette Score as seen below:



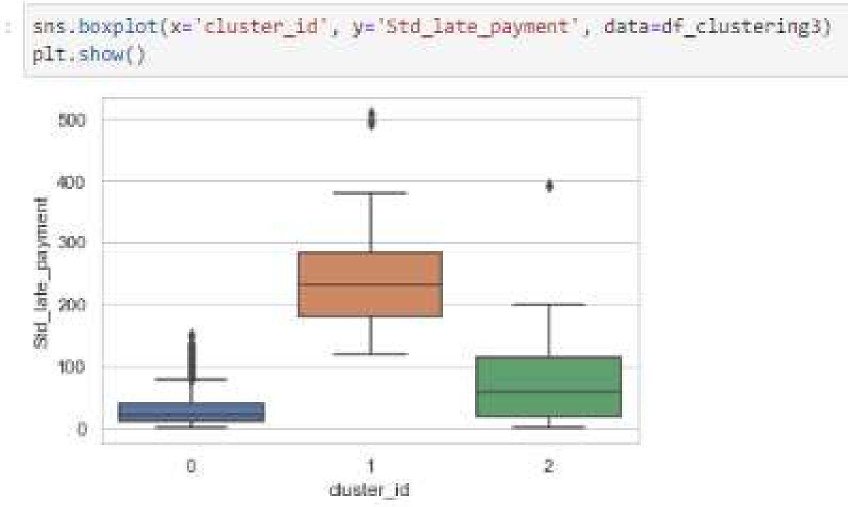
```
For n_clusters = 2, The Silhouette Score is 0.7678537339903129
For n_clusters = 3, The Silhouette Score is 0.7580335882855453
For n_clusters = 4, The Silhouette Score is 0.6133717393642635
For n_clusters = 5, The Silhouette Score is 0.48339938757681167
For n_clusters = 6, The Silhouette Score is 0.47333255484909204
For n_clusters = 7, The Silhouette Score is 0.4241067874140255
For n_clusters = 8, The Silhouette Score is 0.42803681683346906
For n_clusters = 9, The Silhouette Score is 0.3874692974648615
For n_clusters = 10, The Silhouette Score is 0.38453794594867036
For n_clusters = 11, The Silhouette Score is 0.38477278198444526
For n_clusters = 12, The Silhouette Score is 0.36878463842178094
For n_clusters = 13, The Silhouette Score is 0.36135816859574516
For n_clusters = 14, The Silhouette Score is 0.36251656210610494
```



Visualizing Late Payments across Clusters



```
sns.boxplot(x='cluster_id', y='Avg_late_payment', data=df_clustering3)
plt.show()
```



```
sns.boxplot(x='cluster_id', y='Std_late_payment', data=df_clustering3)
plt.show()
```

Customers belonging to Cluster Id 2 exhibit a higher likelihood of making late payments based on our analysis. It is crucial to prioritize collecting payments from these customers to mitigate the risk of delayed or outstanding payments.

Model Building - I

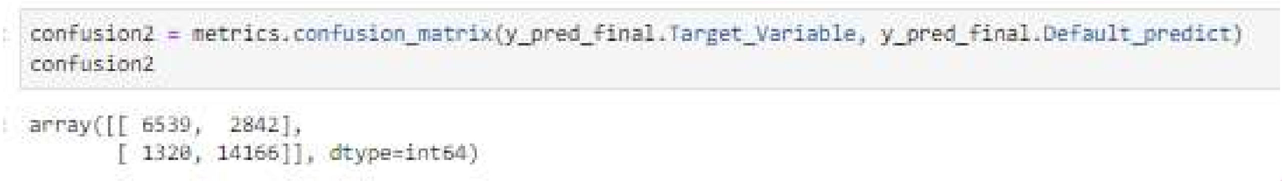
First, we will split the data randomly into train and test set where train_size = 0.7 and test _size = 0.3.

Then we scale the variables in the train set and build the first model which is Random Forest Classifier Model

We then used GridSearchCV() function and finally calculated the accuracy of the model to build it and test it on the testing set.

Accuracy: 83%

In the end, we built the confusion metrics. Refer below:



```
confusion2 = metrics.confusion_matrix(y_pred_final.Target_Variable, y_pred_final.Default_predict)
confusion2
```

```
array([[ 6539,  2842],
       [ 1320, 14166]], dtype=int64)
```

Finally we made the predictions on Open Invoice dataset using this model

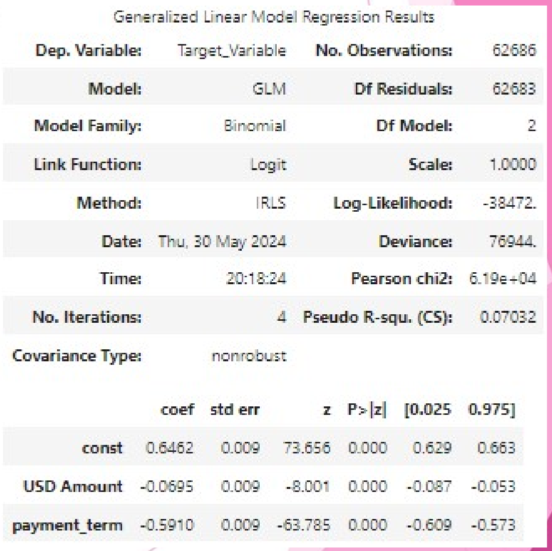
Model Building - II

First, we will split the data randomly into train and test set where train_size = 0.7 and test _size = 0.3.

Then we scale the variables in the train set and build the second model which is Logistic Regression Model

We then did a RFE by fitting the model, looking at p-value of features and eliminating them with consideration of their VIF to avoid overfitting/multicollinearity.

After n number of iterations, we finally selected 2 best features for the model: USD_Amount and Payment_term.



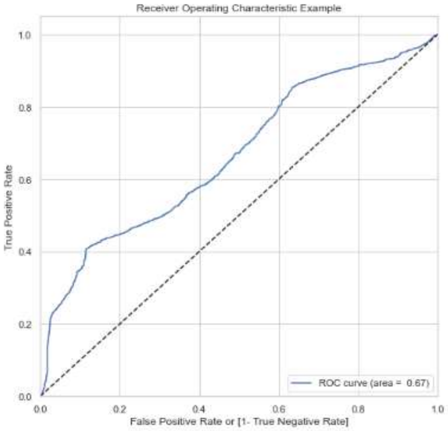
Generalized Linear Model Regression Results

Dep. Variable:	Target_Variable	No. Observations:	62686
Model:	GLM	Df Residuals:	62683
Model Family:	Binomial	Df Model:	2
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-38472.
Date:	Thu, 30 May 2024	Deviance:	76944.
Time:	20:18:24	Pearson chi2:	6.19e+04
No. Iterations:	4	Pseudo R-squ. (CS):	0.07032
Covariance Type:	nonrobust		

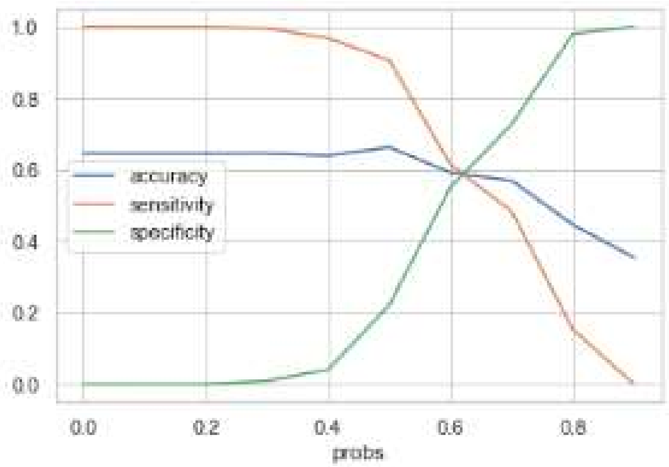
	coef	std err	z	P>|z|	[0.025	0.975]
const	0.6462	0.009	73.656	0.000	0.629	0.663
USD Amount	-0.0695	0.009	-8.001	0.000	-0.087	-0.053
payment_term	-0.5910	0.009	-63.785	0.000	-0.609	-0.573

Accuracy: 66%

ROC Curve:



Model Evaluation – Logistic Regression



ACCURACY, SENSITIVITY (TRAIN)

- Accuracy – 66.3%
- Sensitivity – 90.6%

We can see the optimal cutoff point is coming around 0.5, lets see values for other metrics with cutoff point as 0.5.

ACCURACY, SENSITIVITY (TEST)

- Accuracy – 65.98%
- Sensitivity – 90.52%

For the above observations the metric values are very close for Train and Test data so we **can accept this model.**

Conclusion

With the help of different models, we tried to find the potential customers for late payments on the basis of their late payment behaviors etc. against Open Invoices.

Considering the overall scenario, it would be more advantageous to choose the Random Forest Model. This model can effectively handle large datasets with higher dimensionality and is robust against overfitting.

Additionally, Random Forest provides important insights through feature importance metrics, aiding in better decision-making.

Hence, keeping in mind the advantages, predictions and accuracy scores of different models, we have finalized the Random Forest Classifier Model to be the accurate one to be used for predictions by Schuster.

Within the code we have selected the list of potential customers, some of them are listed here:
AL J Corp, ALLI Corp, MIDD Corp, SAUD CORP etc, along with their account number and transaction number.

The built Random Forest Classifier model shows ~**83% accuracy**.

With this, the model finds correct potential customers

Hence, we can say that the overall model built is accurate and hence one can achieve desired results with the same.