Retail and Ecommerce case – study

Problem statement

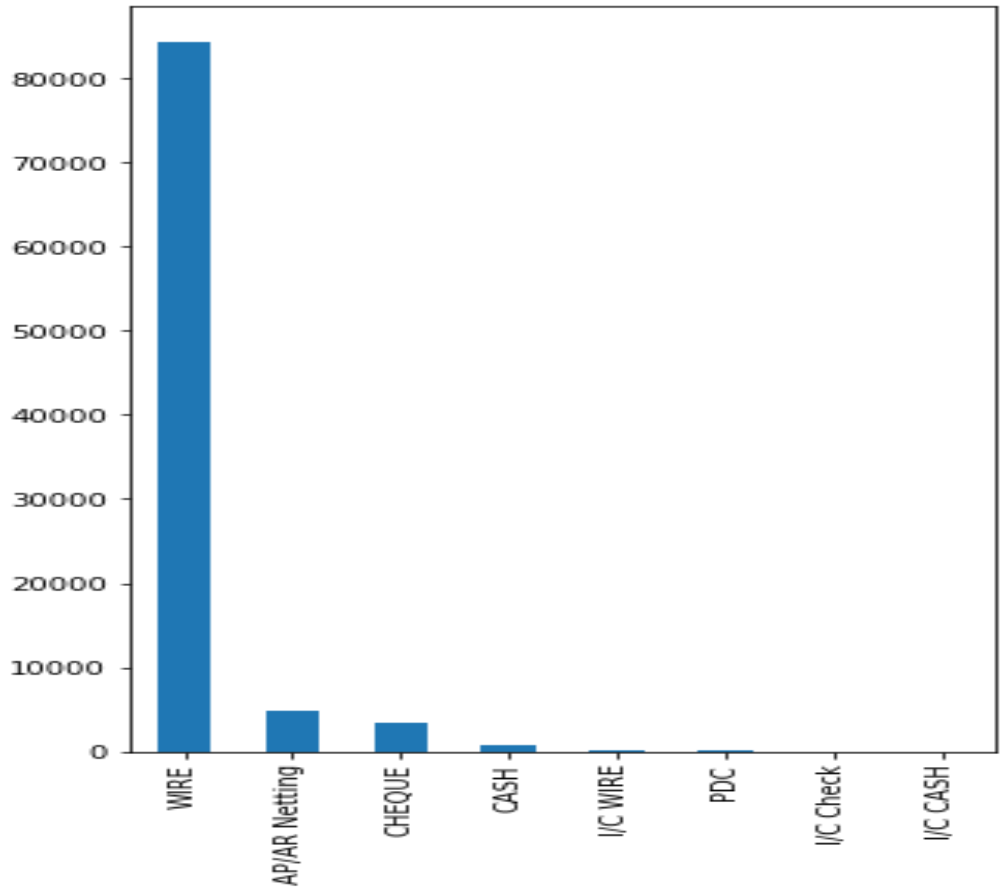
Schuster is a multinational retail company dealing in sports goods and accessories. Schuster conducts significant business with hundreds of its vendors, with whom it has credit arrangements. Unfortunately, not all vendors respect credit terms and some of them tend to make payments late. Schuster levies 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.

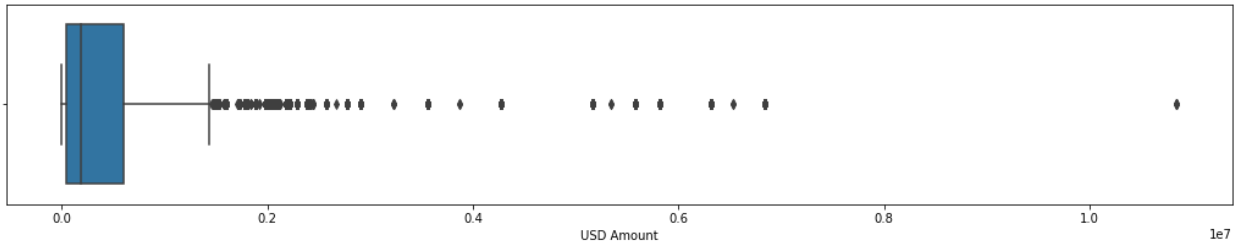
This case – study was solely and completely done by "saptarshi Das". Other 2 group member did not showed any involvement in the assignment. Since the entire case study was done by me, due to time constraint I cannot able to do it up to my perfection, but I have tried my level best.

First, started with received payment dataset , import some library and did descriptive analytics into it,

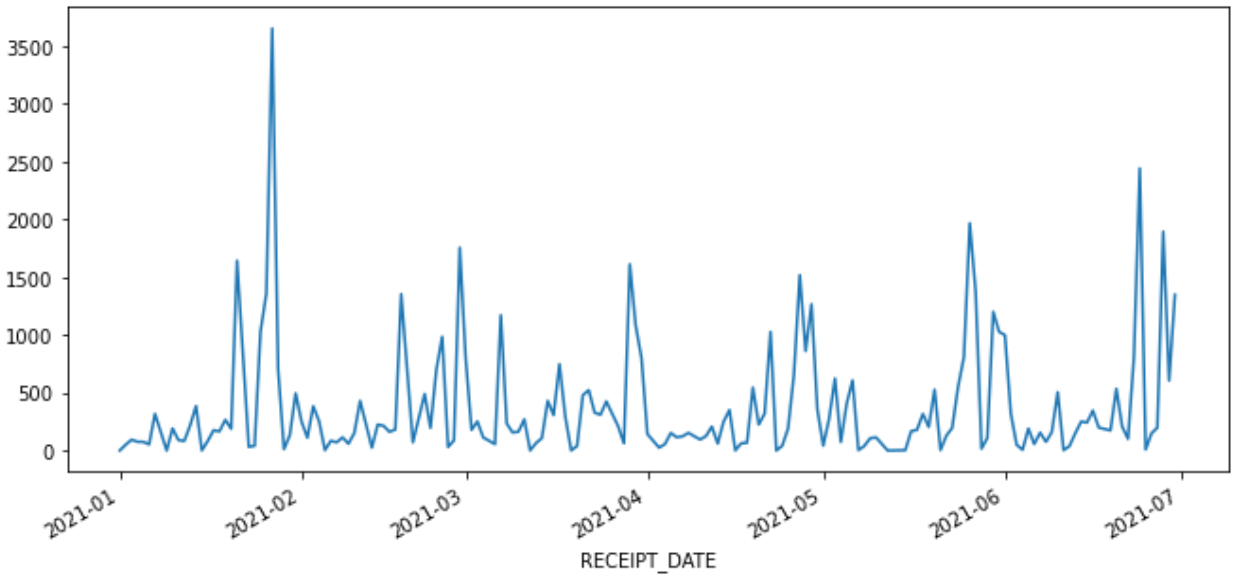
The data- columns data types were not proper so fixed that , fixed missing values , and did some visualization

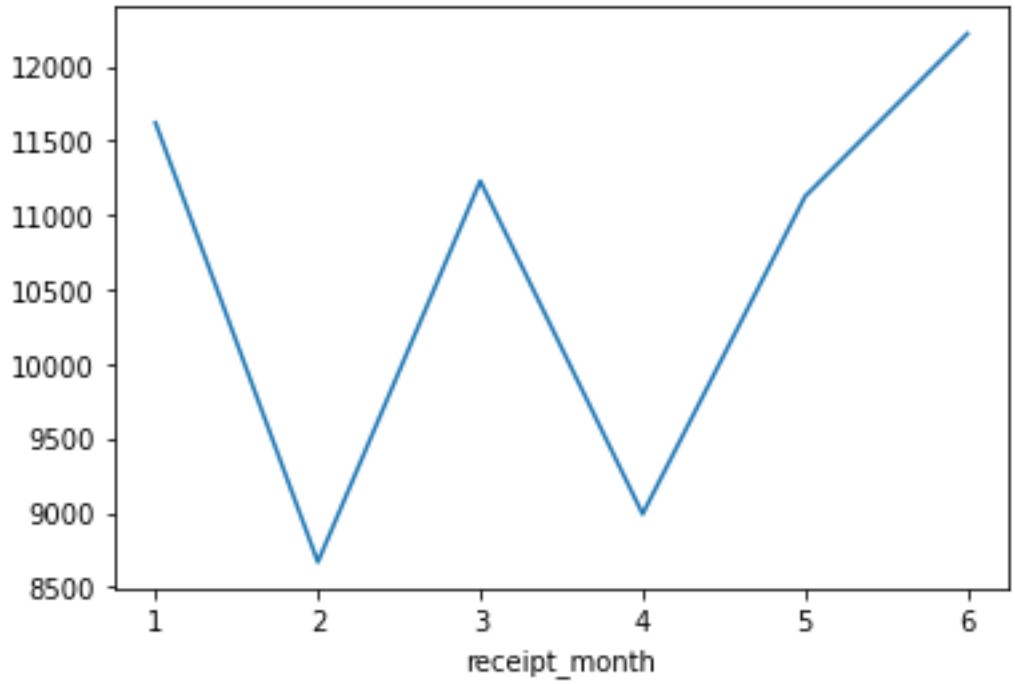


Visualized the receipt payment mode and found max payment was from WIRE mode.



The local amount has a skewed dataset with some extreme high values and same With the USD amount also.

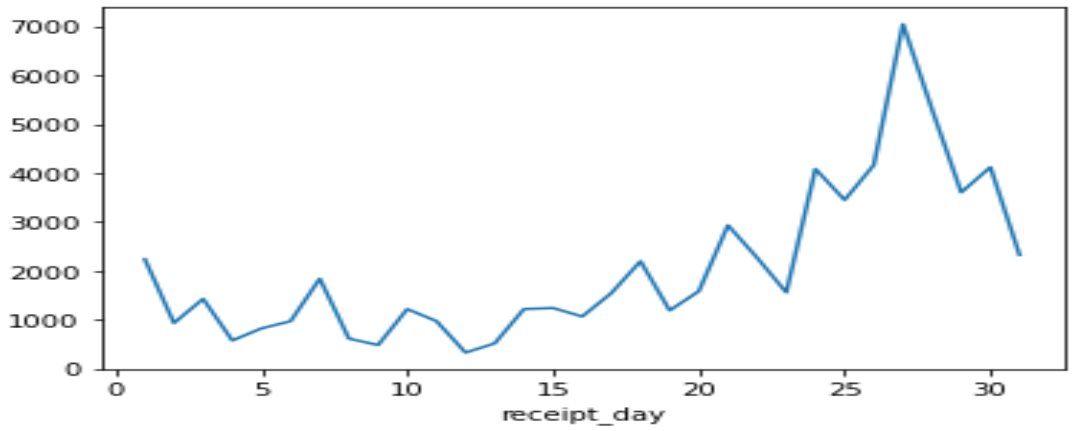




Then I did some day and monthly analysis on the due date and found that end of the months like from 25 to 30 has max customer to did not pay before the due_date you can see a peak in the end on each month

And in monthly chat , its swinging jan , march and may June has high default rate

See the down fig, has from 25 to 28 day has high default rates



After missing values , did some feature eng, extracted the day and month , year from due date, grouped the customer name on the frequencies and one- hot encoding was done to convert the categorical to numeric.

I applied logistic regression method for classification

Applied the RFE tech, and choose 15 best features,

For numeric feature applied log transformation as there were outliers for outliers treatment .

According to the RFE and manual p values and VIF comparison found that , receipt modes , currency , loan amt, due date , invoice class , payment term and types acts As an imp features for model building.

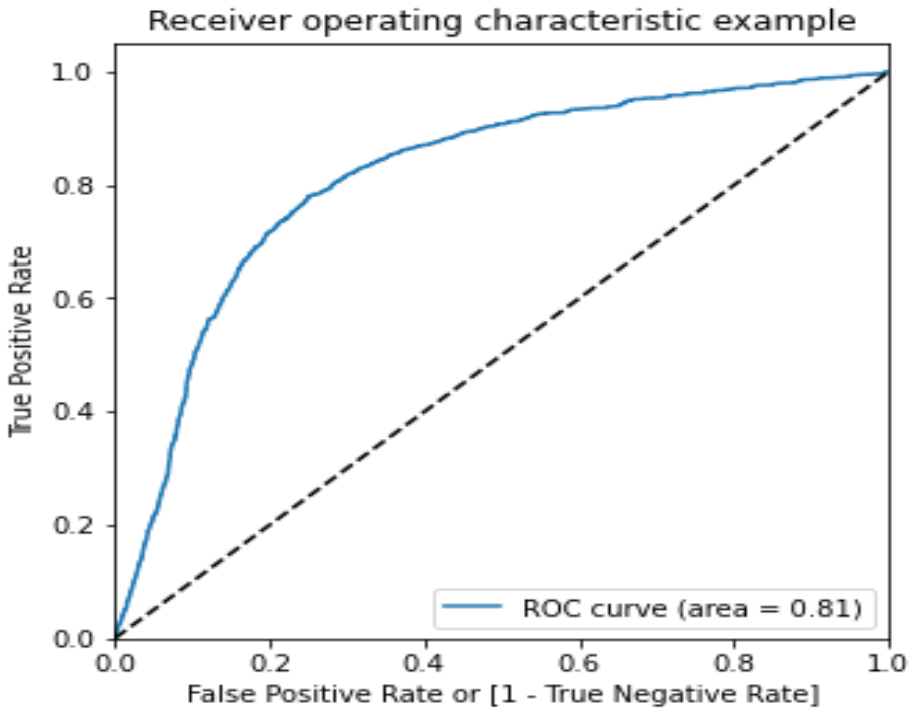
On the model building part on the train set , choosing the threshold as 0.5

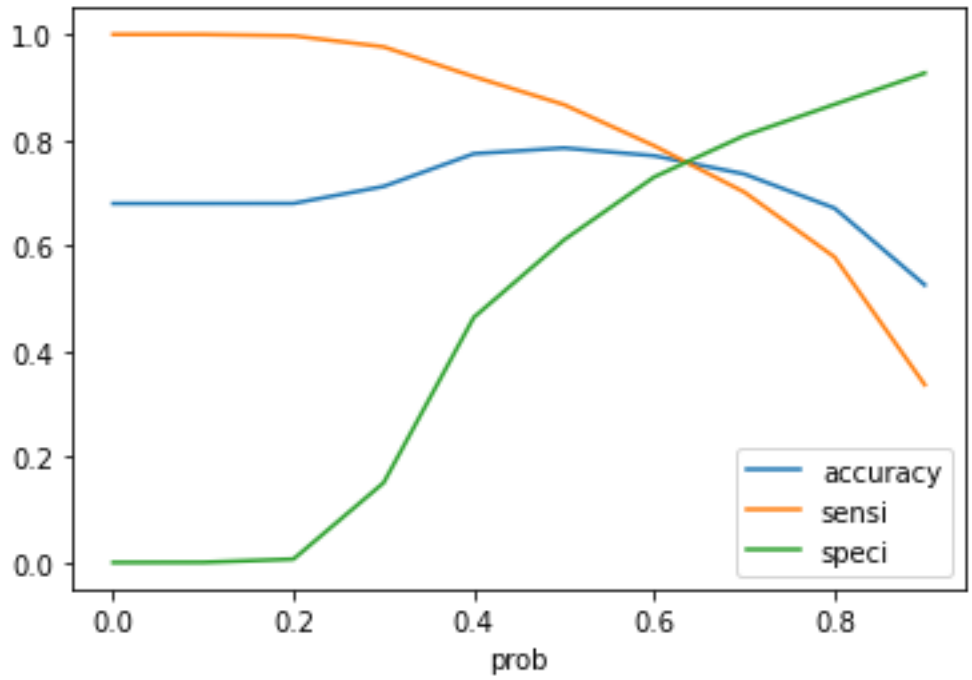
Model accuracy – 78%
Sensitivity – 80%

Then plot a ROC curve

The ROC curve below looks quite good.

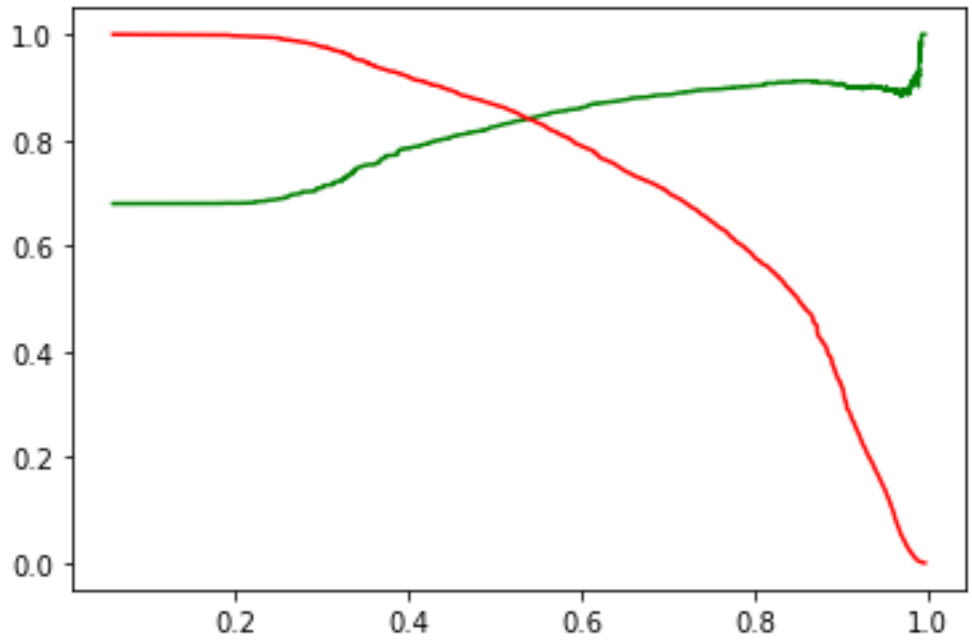
The ACCURACY , SENSI, SPECIFICITY intersection graph show a imp threshold value , which I considered to be 0.65 as per the graph





After applying the new threshold, the accuracy decreased a little but the main parameter sensitivity increased to 82% on the train set.

Also performed the precision and recall test to check the threshold it seems 0.6 is the best



After applying for the test data

Accuracy – 75%

Sensitivity – 76%

This is quite good I think most 80% defaulter we are capturing ,

Now the real invoice data is uploaded and all the same data preparation , feature engineering and all the thing are done

Sensitivity – 68% after the threshold of 0.5 ,

Couldn't get the time , not much happy with the model , could have improved a a lot after again checking the threshold or by applying random forest

IMP

But company should be notify the **wire** receipt payment mode customer more.

Certain currency like ,

CURRENCY_CODE_BHD	-2.5652	0.358	-7.166	0.000	-3.267	-1.864

CURRENCY_CODE_EUR	-0.7265	0.075	-9.714	0.000	-0.873	-0.580
CURRENCY_CODE_GBP	1.6908	0.333	5.077	0.000	1.038	2.343
CURRENCY_CODE_SAR	0.9615	0.024	40.866	0.000	0.915	1.008
CURRENCY_CODE_USD	0.6427	0.027	24.128	0.000	0.591	0.695

These currency payment are risky.

The due_dates by the end of the month are also risky

Invoice type and class DM (debit) also be careful about

These are my finding and business suggestions.