E-COMMERCE & RETAIL B2B CASE STUDY

To Identify late payment customers for precautionary measures

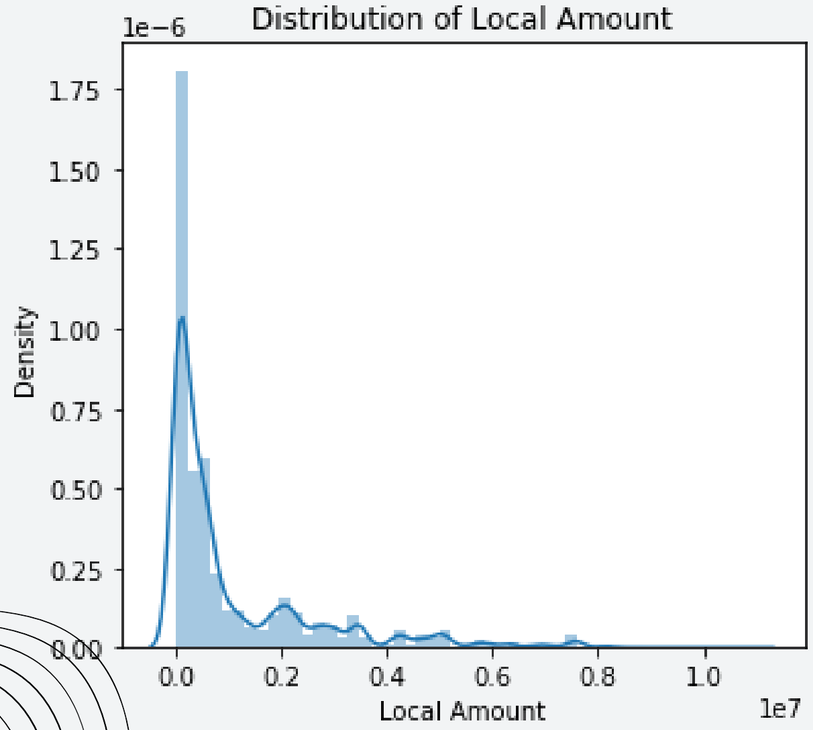
BY-SATYAM KHORGADE

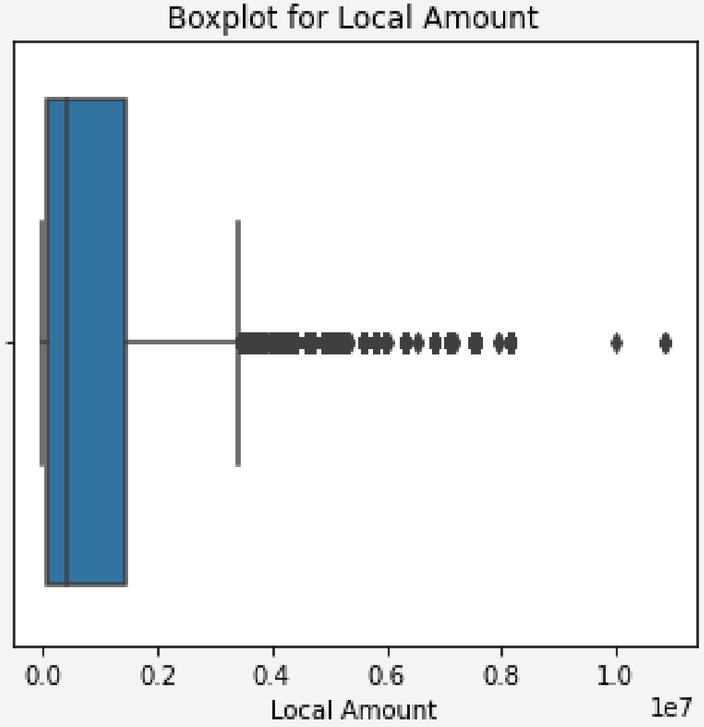
READING AND UNDERSTANDING DATA

- 1) Data Type Checks
- 2) Treating Missing Values
- 3) Dropping unnecessary columns
- 4) Outlier detection
- 5) Handling Outliers
- 6) Derived Columns

- NUMERIC COLUMNS - [USD_AMOUNT]
- CATEGORICAL COLUMN - PAYMENT TERM,
- INVOICE_CLASS.
- DATE COLUMNS - RECIEPT_DATE, DUE_DATE,
- INVOICE_DATE
- CREATED TARGET VAR - LATE_PAY
- DROPPED COLUMNS LIKE
- LOCAL_AMOUNT,RECEIPT_DOC_NO ,CUSTOMER_
- NAME,CLASS_CURRENCY_CODE,INV _CURR_CODE,
- RECIEPT_METHOD WHICH WERE NOT CONTRIBUTING
- TO OUR TARGET VARIABLE

EDA





Local amount

EDA

Categorical Columns

Top 10 customers based on frequency

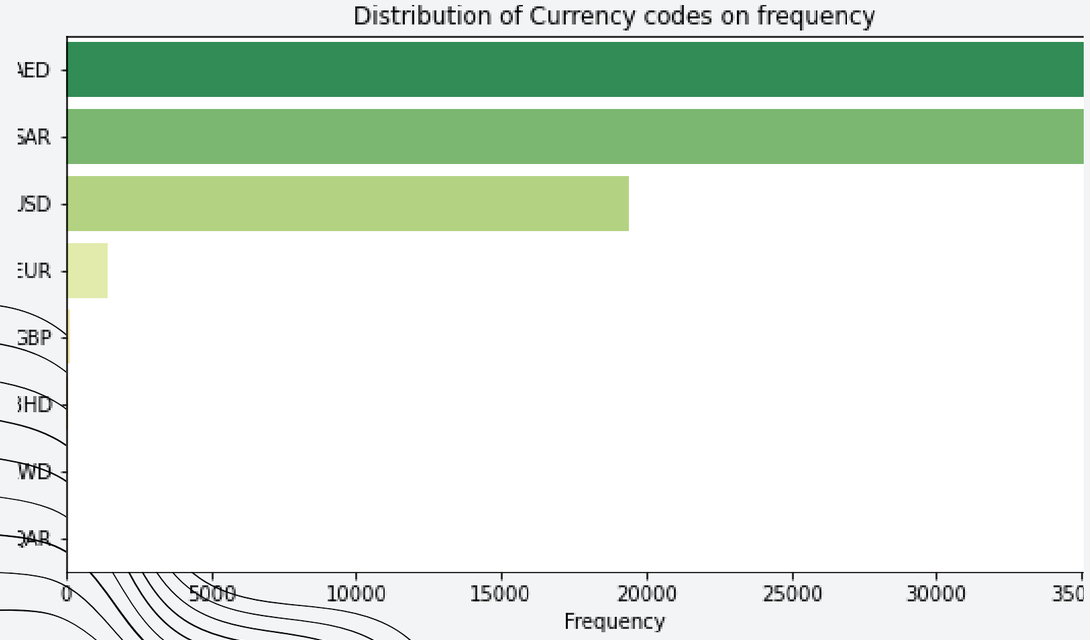
- SEPH Corp 23075
- FARO Corp 15004
- PARF Corp 6624
- ALLI Corp 5645
- AREE Corp 2224
- DEBE Corp 2133
- RADW Corp 1647
- YOUG Corp 1480
- HABC Corp 1402
- CARR Corp 952

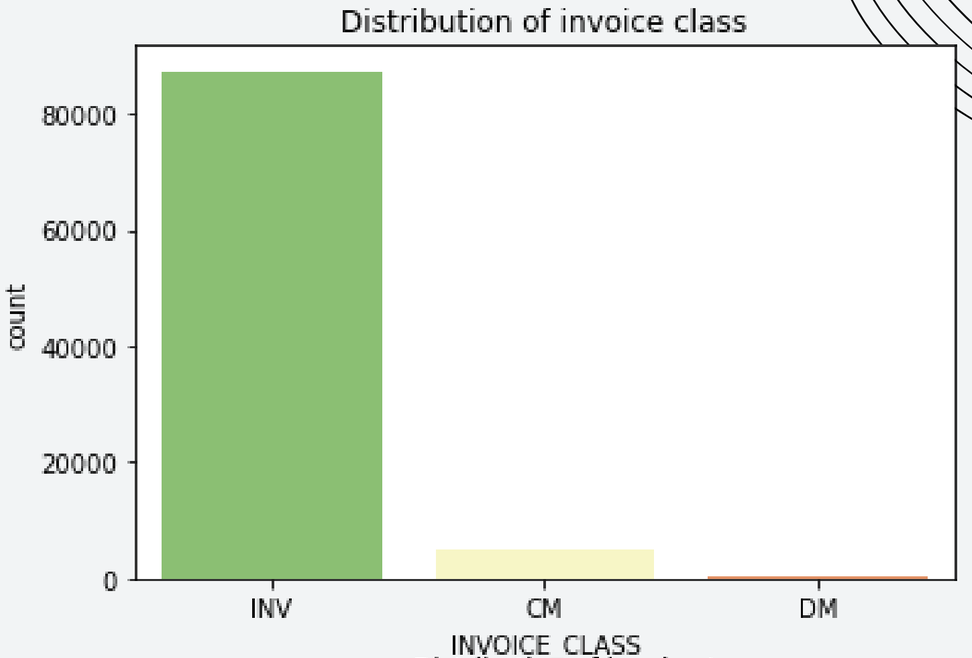
Top 10 customers based on amount

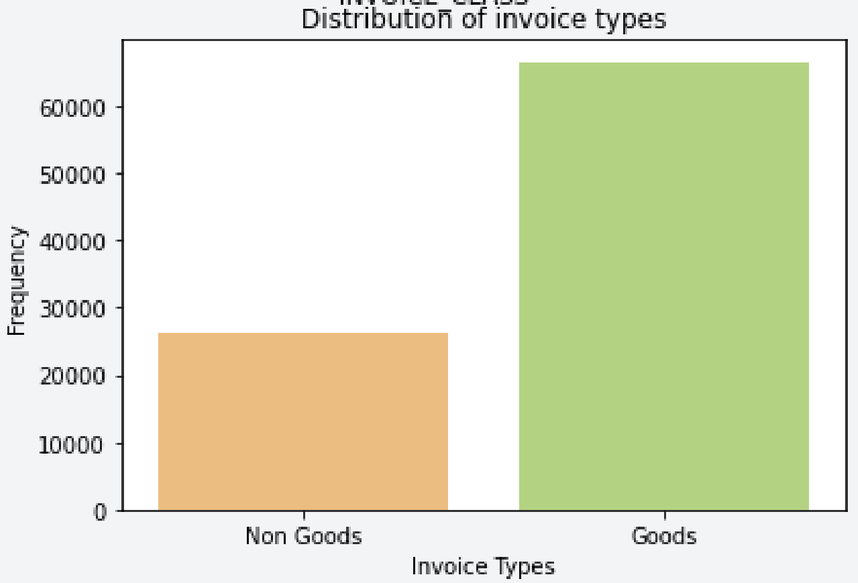
- SEPH Corp 32,533,709,059.000
- FARO Corp 5,790,071,209.000
- PARF Corp 3,200,510,261.000
- ALLI Corp 2,580,740,593.000
- AREE Corp 1,125,144,489.000
- HABC Corp 534,321,619.000
- RADW Corp 362,237,576.000
- L OR Corp 295,550,941.000
- CGR Corp 279,516,184.000
- PCD Corp 246,606,985.000

EDA

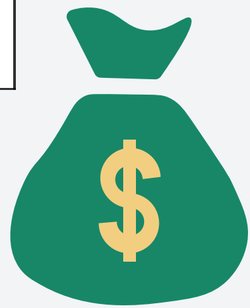
Categorical Columns





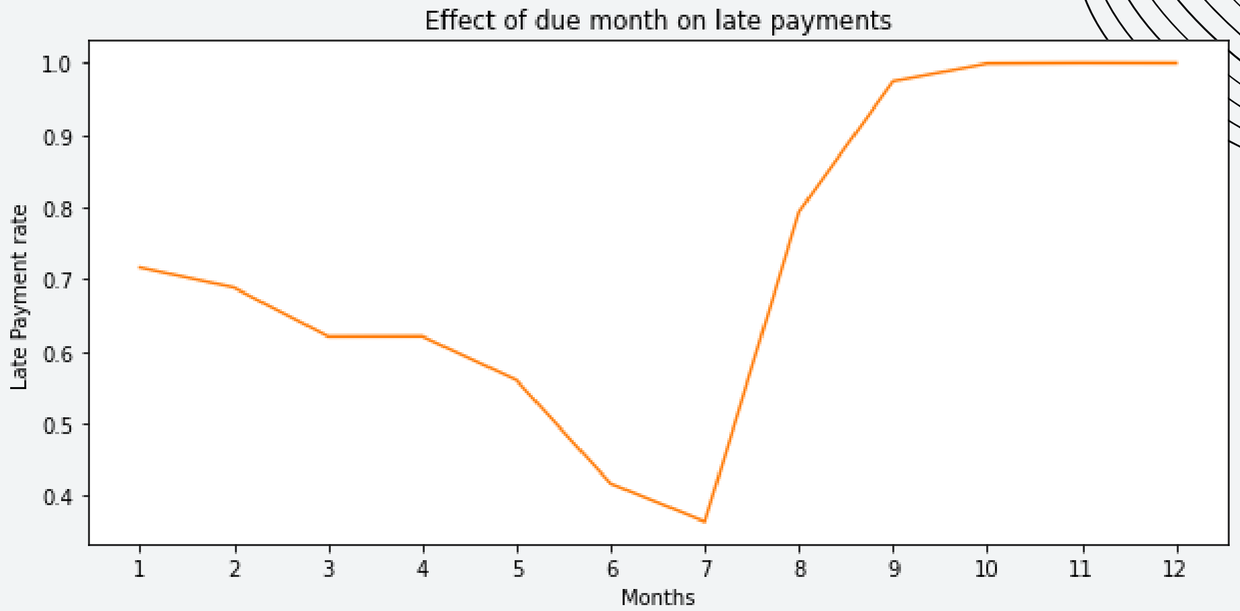


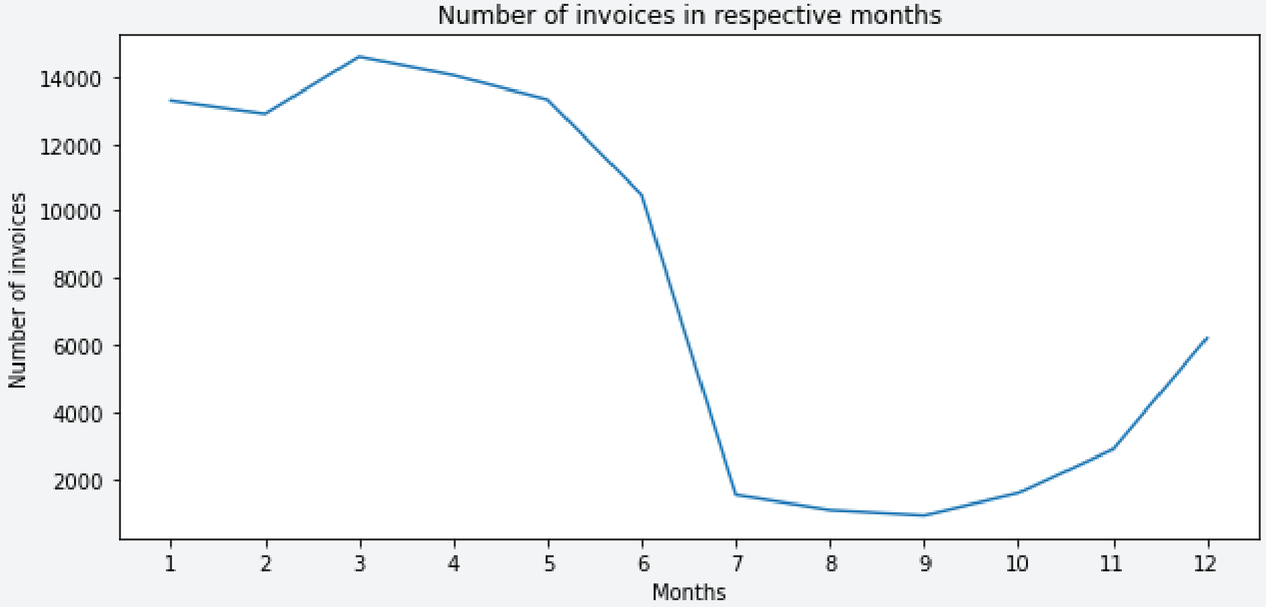
invoice



EDA

Bivariate Analysis

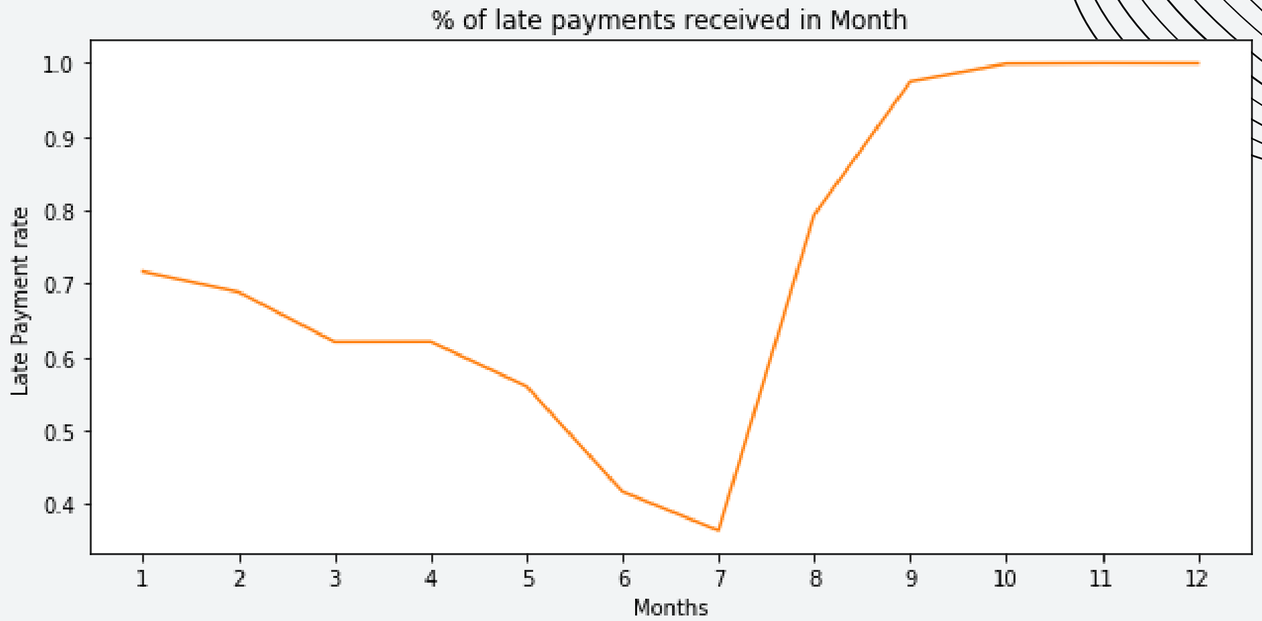


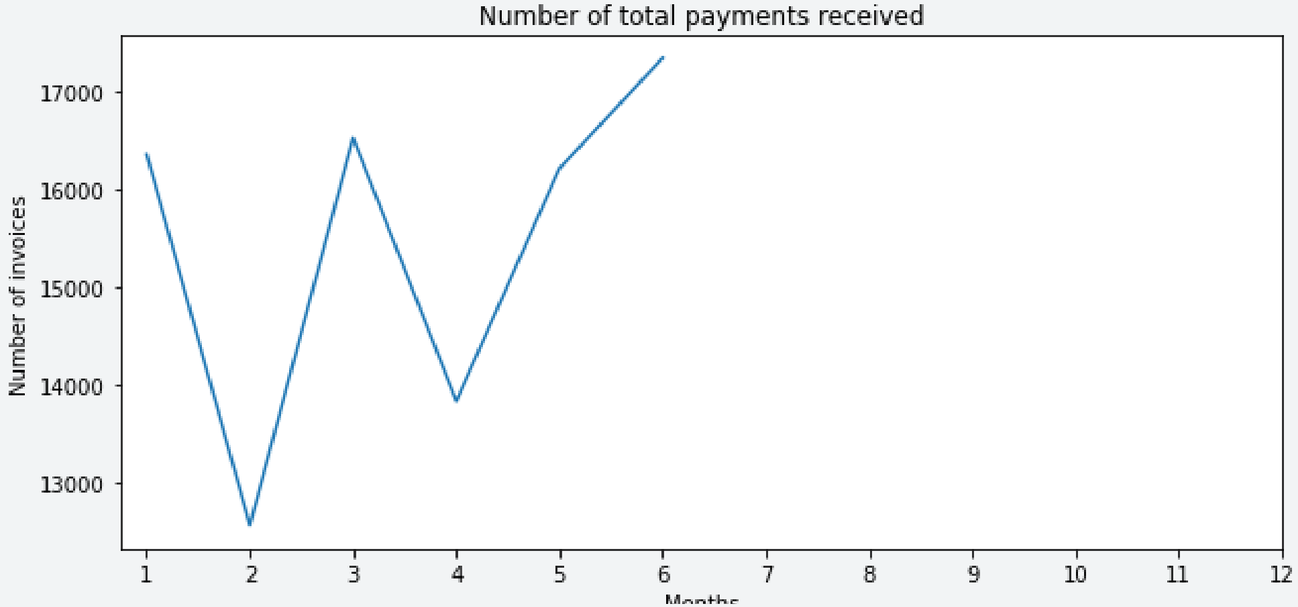


Monthly affects on payments and invoices.

EDA

Bivariate Analysis

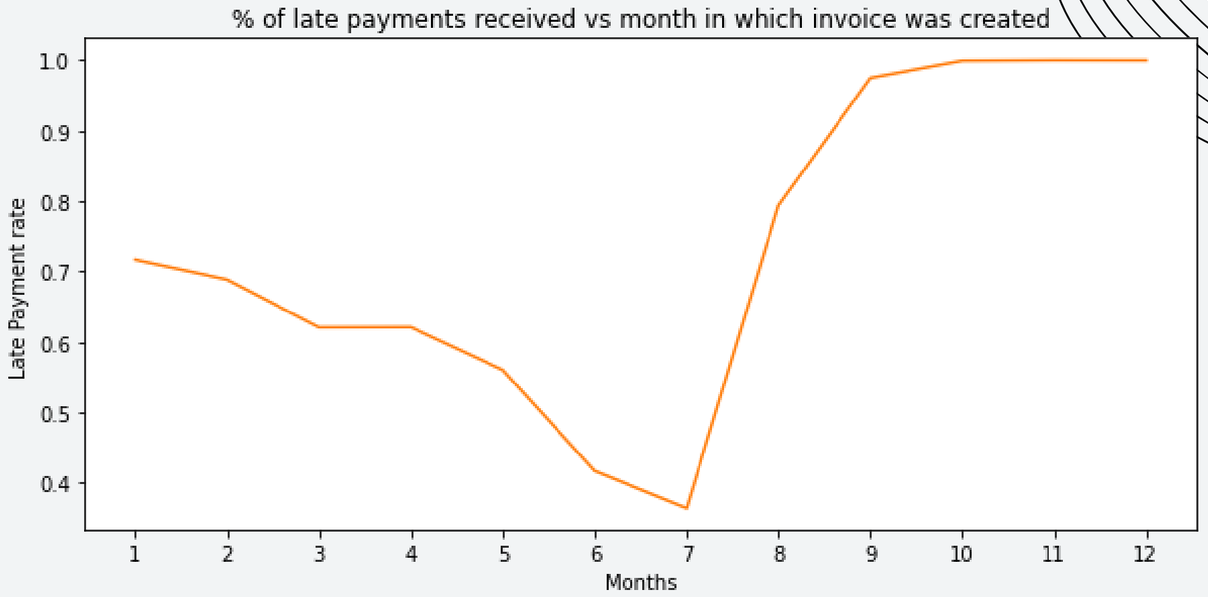


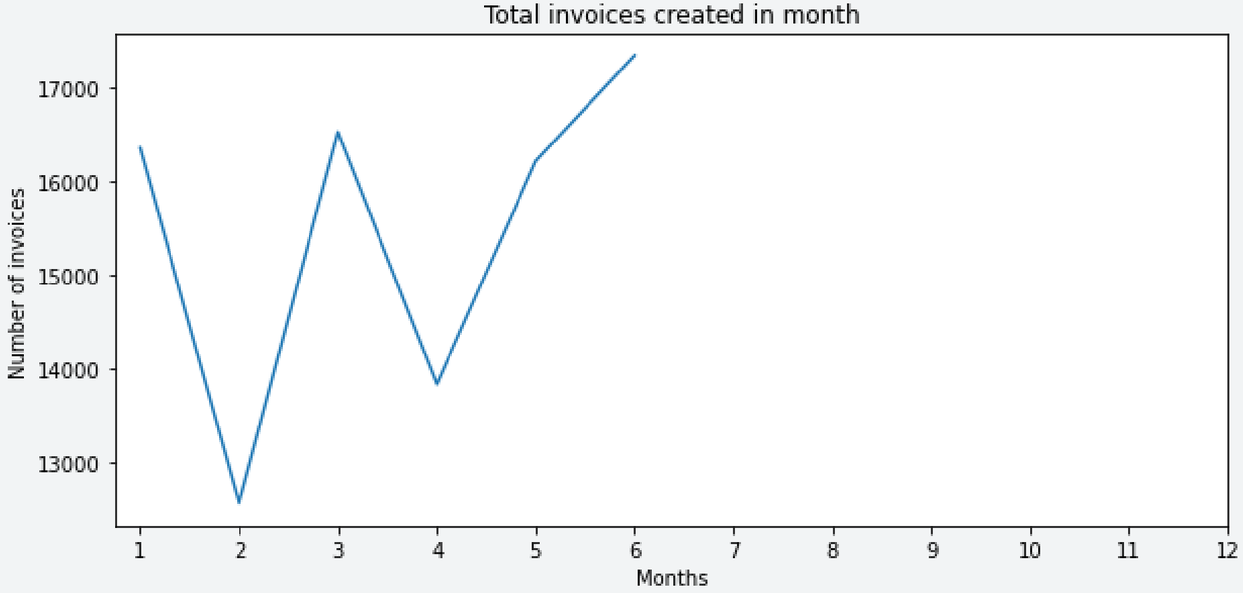


A stark effect is noted here, which provides that all payments are received in the first half of the year only

EDA

Bivariate Analysis

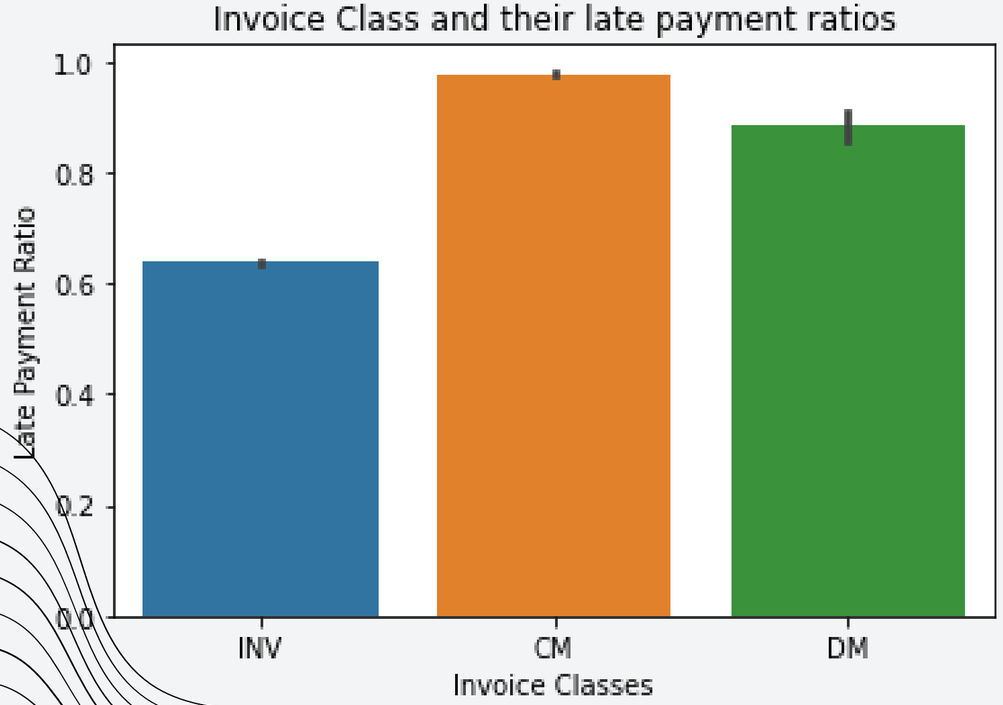




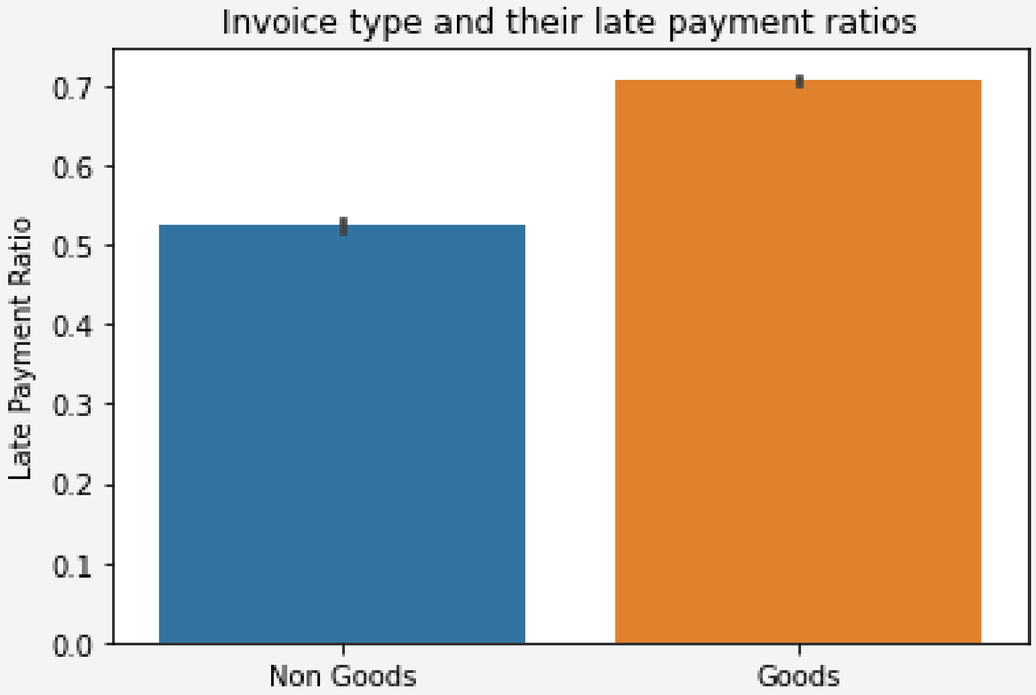
It is observed that all invoices were created in the first half of the year.

EDA

Bivariate Analysis



It is observed that both credit and debit memo have high late payment ratios, however it is also to be noted that there are only a few invoices with CM and DM Class



It is observed that late payment ratio is much higher for goods.

FEATURE ENGINEERING

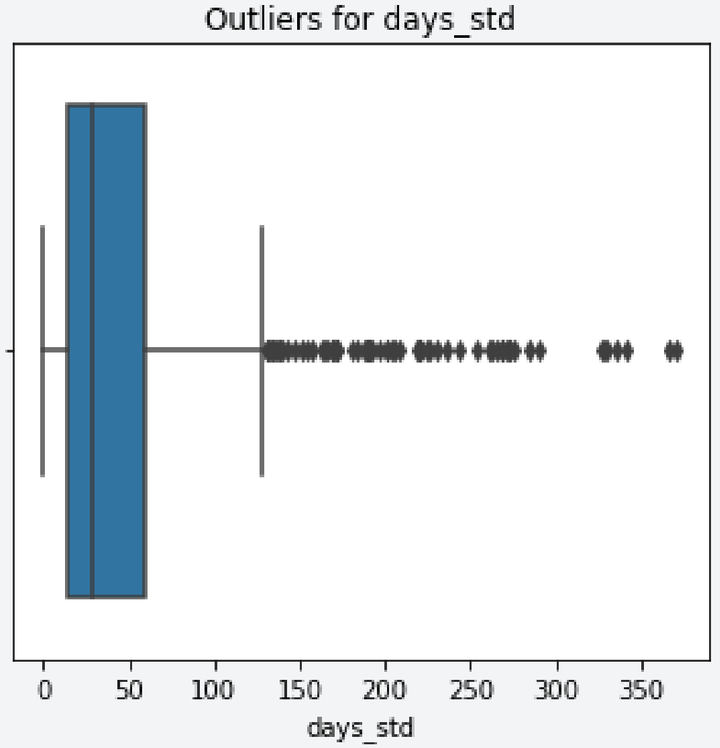
- Created Dummy variables for ‘Payment _Term’ and ‘Invoice_class’

- First combined similar payment terms and then clubbed every other payment term except top 10.

- Open_Invoice_Data table - removed unnecessary columns and created dummy variables
- Outlier treatment
- Removed about quantile 0.99.

CLUSTERING

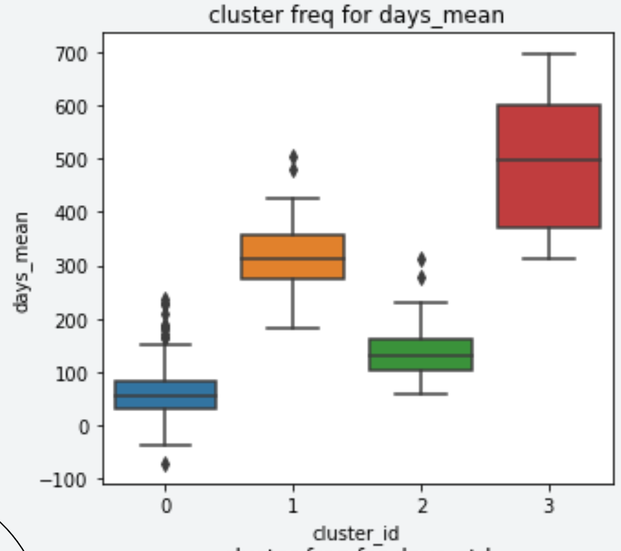
CUSTOMER SEGMENTATION

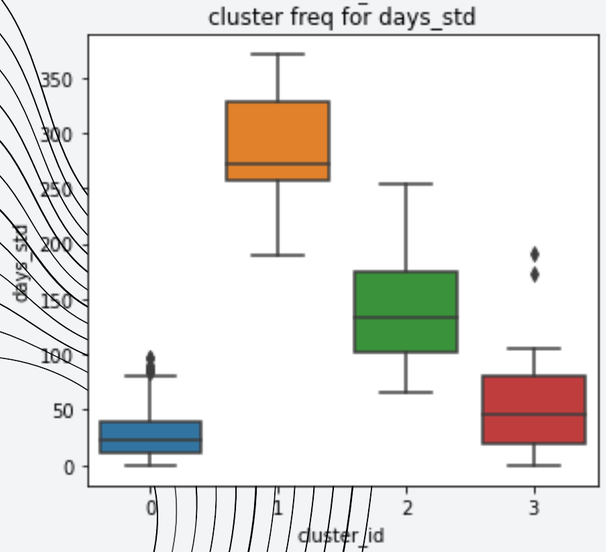


SCALING AND HOPKINS TEST

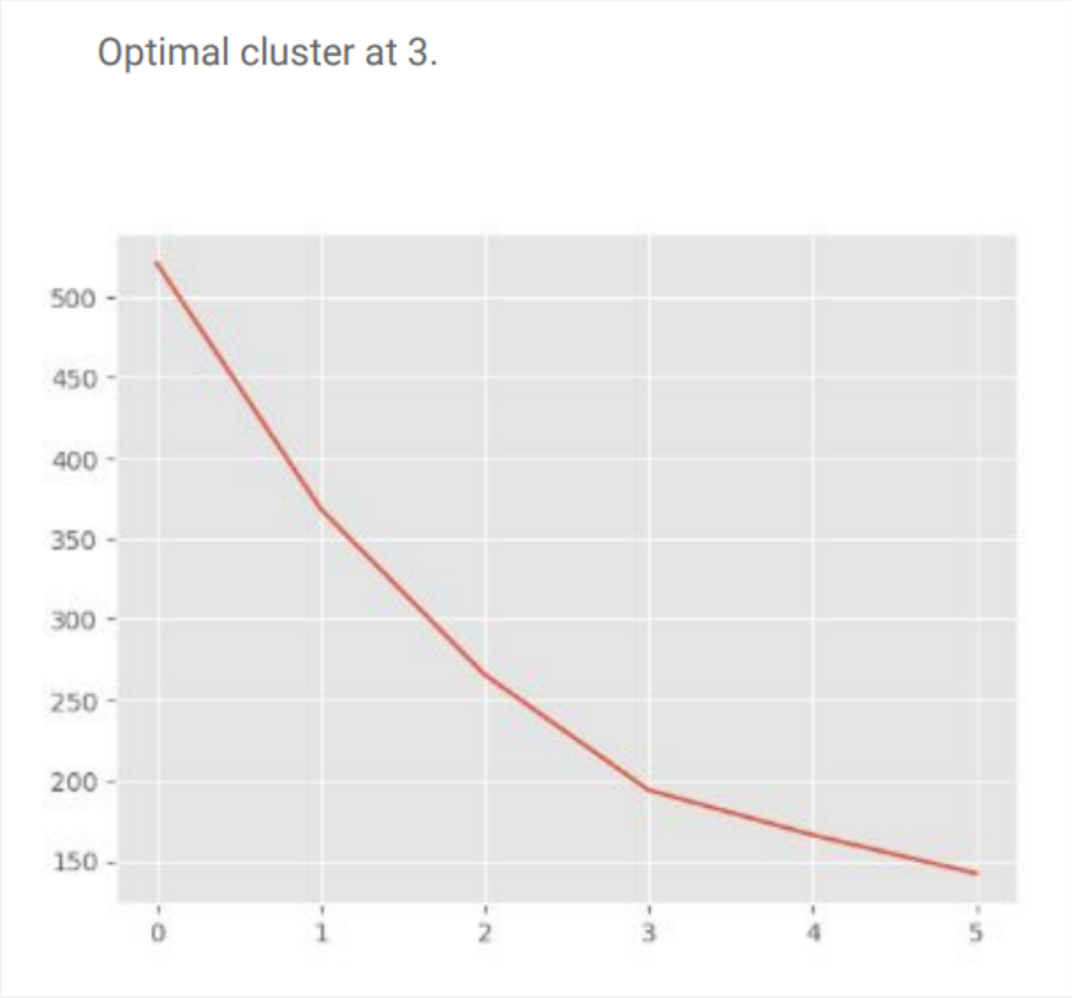
- We can see negative values for days mean which means that the customer has done immediate payment, while the invoice was created later.
- To maintain this data integrity, it would be advisible to use standardization over normalization for scaling.
- On running hopkins Test, we got a value of 0.91337.
- A value close to 1 tends to indicate the data is highly clustered, random data will tend to result in values around 0.5, and uniformly distributed data will tend to result in values close to 0.

OPTIMAL CLUSTERS ELBOW METHOD



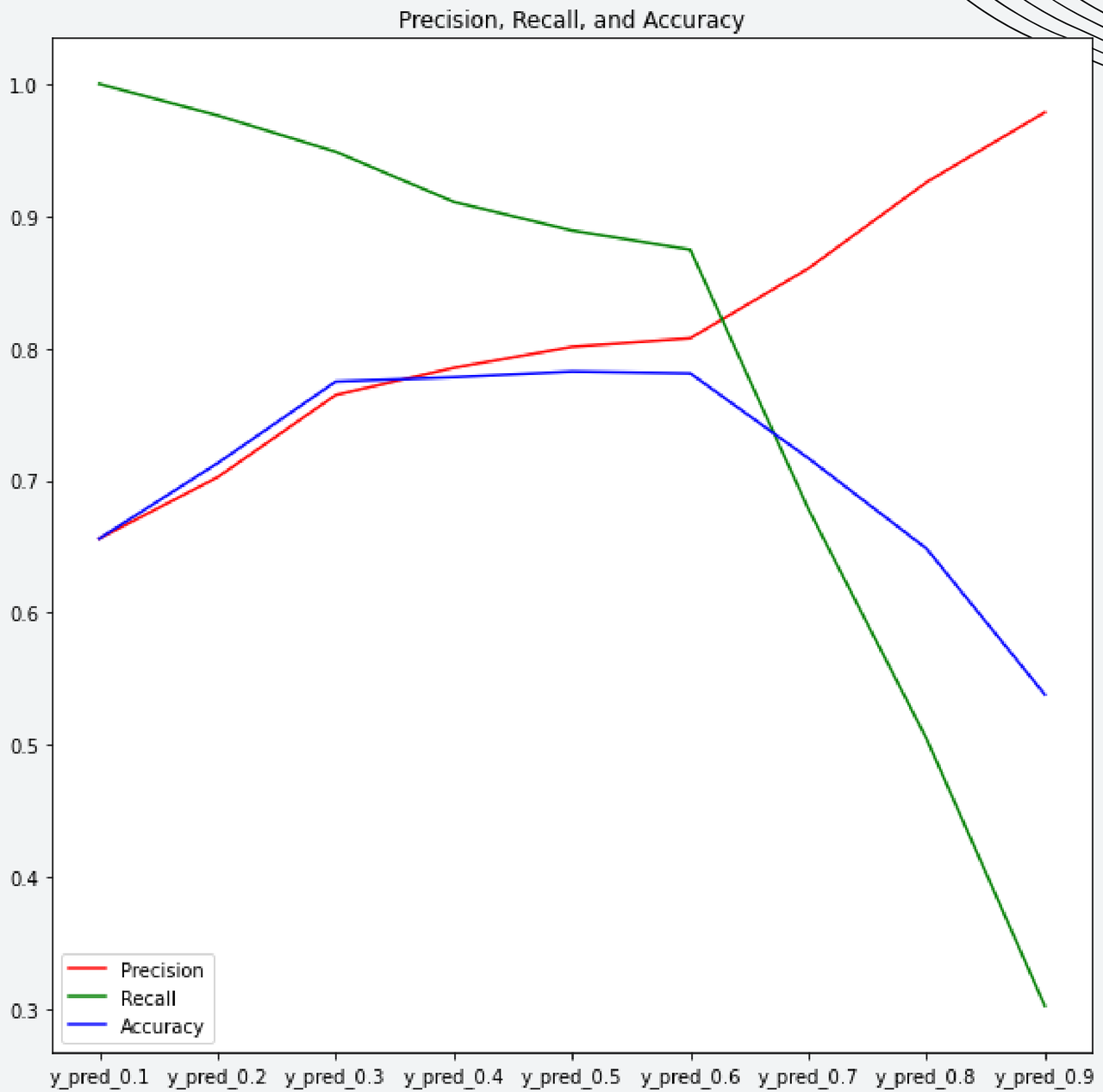


Summary Clustering



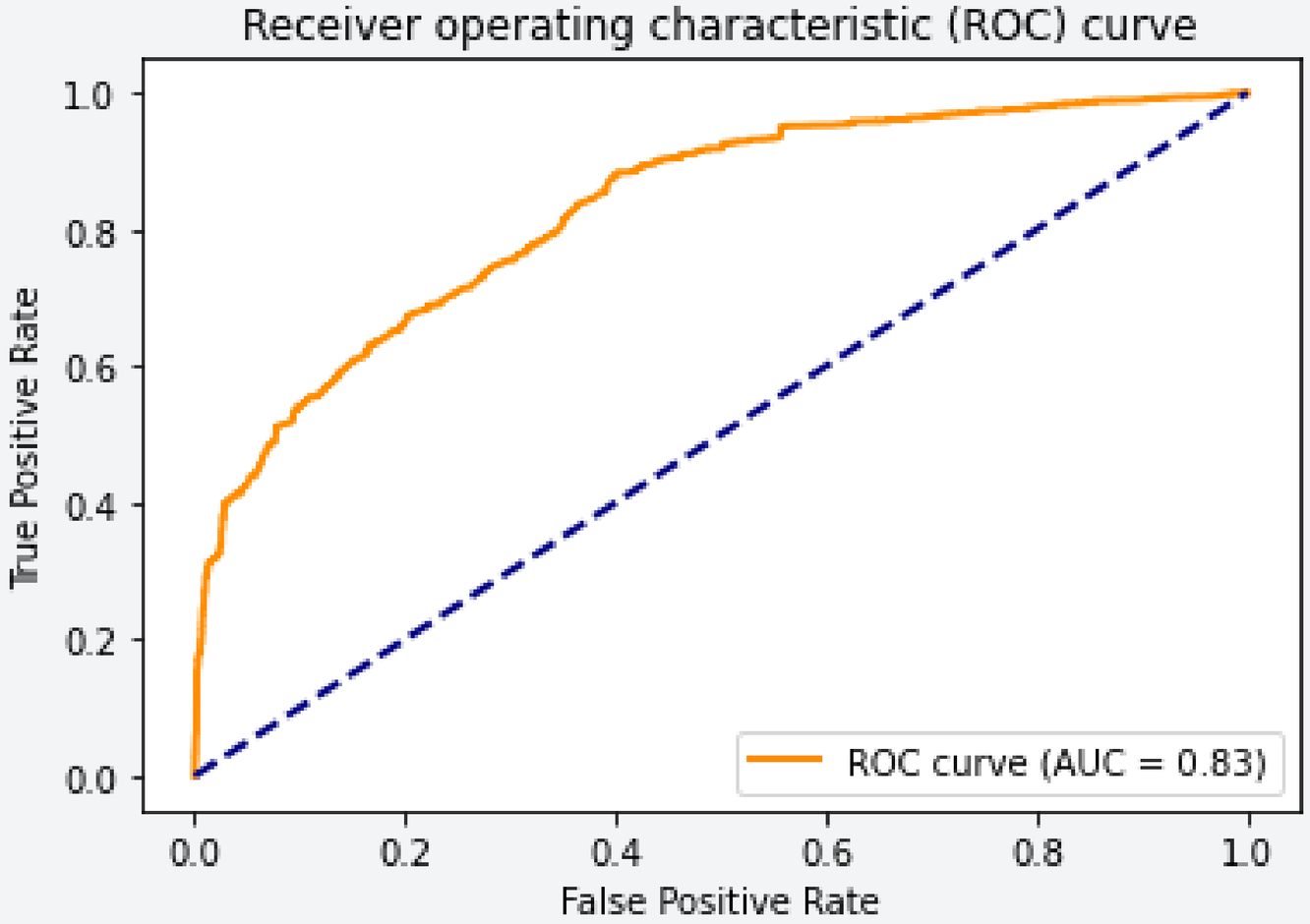
LOGISTIC REGRESSION

Precision , recall and accuracy curve give us the optimal cutoff at 0.5



LOGISTIC REGRESSION

The logistic regression algorithm is working, and has an AUC score of 0.83



LOGISTIC REGRESSION

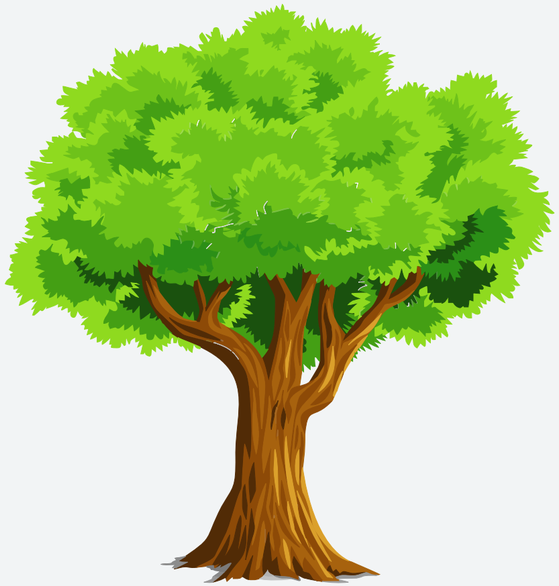
Evaluation metrics on training dataset

	precision	recall	f1-score	support
0	0.73	0.58	0.65	22349
1	0.80	0.89	0.84	42618
accuracy			0.78	64967
macro avg	0.77	0.73	0.74	64967
weighted avg	0.78	0.78	0.78	64967

	precision	recall	f1-score	support
0	0.73	0.58	0.65	9529
1	0.80	0.89	0.84	18315
accuracy			0.78	27844
macro avg	0.77	0.74	0.75	27844
weighted avg	0.78	0.78	0.78	27844

Evaluation metrics on the test set

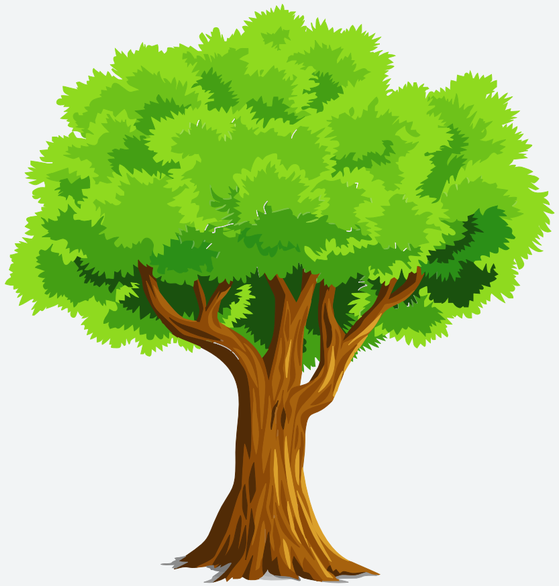
The data is almost similar and hence we can say that our algorithm is working as expected.



RANDOM FOREST

	PRECISION	RECALL	F1-SCORE	SUPPORT
0	0.91	0.85	0.88	9529
1	0.93	0.96	0.94	18315
ACCURACY			0.92	27844
MACRO AVG	0.92	0.90	0.91	27844
WEIGHTED AVG	0.92	0.92	0.92	27844

We see that the basic model itself has a high accuracy, recall, and precision. But here we focus on recall of the positive class in both the train and test set. The basic model itself is able to identify 93-94% of all positive instances.

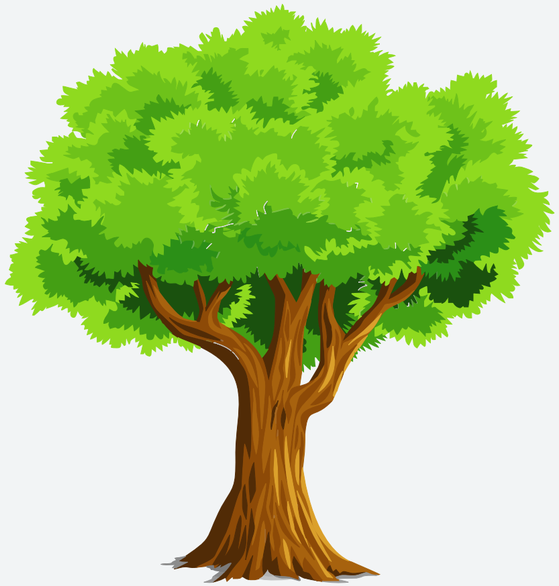


RANDOM FOREST

HYPERPARAMETER TUNING

	PRECISION	RECALL	F1-SCORE	SUPPORT
0	0.96	0.91	0.94	22349
1	0.95	0.98	0.97	42618
ACCURACY			0.96	64967
MACRO AVG	0.96	0.95	0.95	64967
WEIGHTED AVG	0.96	0.96	0.96	64967

GRID SEARCH CV



RANDOM FOREST

	PRECISION	RECALL	F1-SCORE	SUPPORT
0	0.91	0.85	0.88	9529
1	0.93	0.96	0.94	18315
ACCURACY			0.92	27844
MACRO AVG	0.92	0.90	0.91	27844
WEIGHTED AVG	0.92	0.92	0.92	27844

We have found the best model, which is giving great performance for us. All the metrics including accuracy, recall, and precision are great. We will use this model to make predictions on our invoices

RANDOM FOREST

ALTHOUGH FEATURE IMPORTANCE DOES NOT SHOW THE DIRECTION IN WHICH THE POSSIBILITY OF LATE PAYMENT IS AFFECTED (WHETHER IT WILL AFFECT IT POSITIVELY OR NEGATIVELY) IT DOES SHOW THAT FEATURES SUCH AS AMOUNT,
INVOICE MONTH, AND RECEIPT MONTH AFFECT
THE POSSIBILITY OF LATE PAYMENT

FEATURE RANKING:

1. USD AMOUNT (0.231)
2. INVOICE_MONTH (0.202)
3. RECIEPT_MONTH (0.139)
4. 60 DAYS FROM EOM (0.112)
5. 30 DAYS FROM EOM (0.107)
6. CLUSTER_ID (0.053)
7. IMMEDIATE PAYMENT (0.045)
8. 15 DAYS FROM EOM (0.031)
9. 60 DAYS FROM INV DATE (0.017)
10. 30 DAYS FROM INV DATE (0.015)
11. 90 DAYS FROM EOM (0.012)
12. 90 DAYS FROM INV DATE (0.010)
13. 45 DAYS FROM EOM (0.007)
14. 45 DAYS FROM INV DATE (0.006)
15. INV (0.006)
16. CM (0.006)
17. DM (0.001

RANDOM FOREST

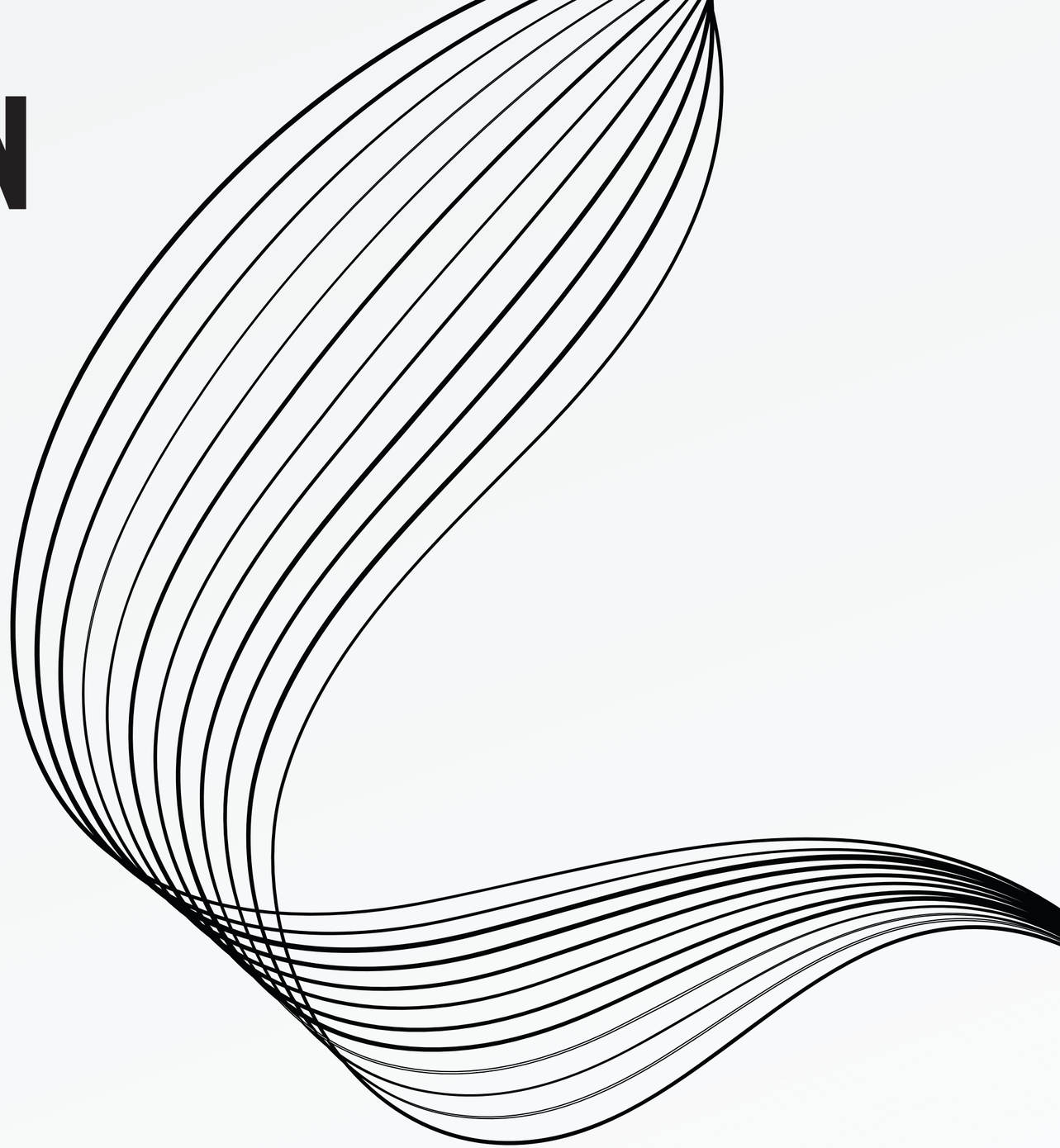
PREDICTIONS

- WE USE BOTH CLASSIFICATION MODELS TO
- CHECK PERFORMANCE.
- AS BOTH THE MODELS WERE PERFORMING WELL,
- WE NEED TO SELECT THE ONE WHICH IS MORE
- INTERPRETABLE.
- ALGORITHM WHICH HELPS US DEFINE THE LINEAR
- RELATIONSHIP OF FEATURES WITH TARGET
- VARIABLE.

Customer_Name	prob_logreg	Prob_rf
2H F Corp	0.0802	0.626667
3D D Corp	0.0000	0.296988
6TH Corp	0.0465	0.213571
ABDU Corp	0.0000	0.413614
ABEE Corp	0.4145	0.460000
...	...	...
ZAIN Corp	0.2711	0.740000
ZALL Corp	0.1761	0.287423
ZALZ Corp	0.0006	0.576915
ZINA Corp	0.1955	0.120000
ZUHA Corp	0.1716	0.305927

CONCLUSION

We have got two models predicting different probability values for a customer to have a late payment. It is to be noted that Random forest is performing much better than logistic regression.. we can take those into account and make pre-emptive calls to the customers to have them pay their invoice amounts on time. Anyone with a high value in Column in Prob_rf- shows that that customer has high probability of making a late payment. Since logistic regression shows linear relationship between probability and the features we can use it to find the relation



THANK'S FOR WATCHING

BY- Satyam Khorgade