E-commerce & Retail B2B Case Study

Identify Late Payment Customers to ensure proper cash flow and avoid the levy of penalties

Sumneet Khanna

Tejaswi Avuthu

Tresa Shetey

Step involved in the Process of Model Building

01
Reading the Dataset
Importing necessary Libraries and reading the dataset to work upon

02
Understanding the data
Check for Shape, datatypes and informatics of the data.

03
EDA
Dropping columns with more than 45% missing values, replace NAN, Outliers univariate & Bivariate Analysis

04
Creation of Dummies Feature Engineering and Scaling Train Test Split
After all the adjustments to the data we proceed with Data Split Train Test and Scaling.

05
Evaluate the model Random Forest & Logistic Regression
Perform statistical , RFE VIF, Cross Validation for feature elimination and check the confusion Matrix and Accuracy score by building 2-3 Models

Reading and Understanding the Data

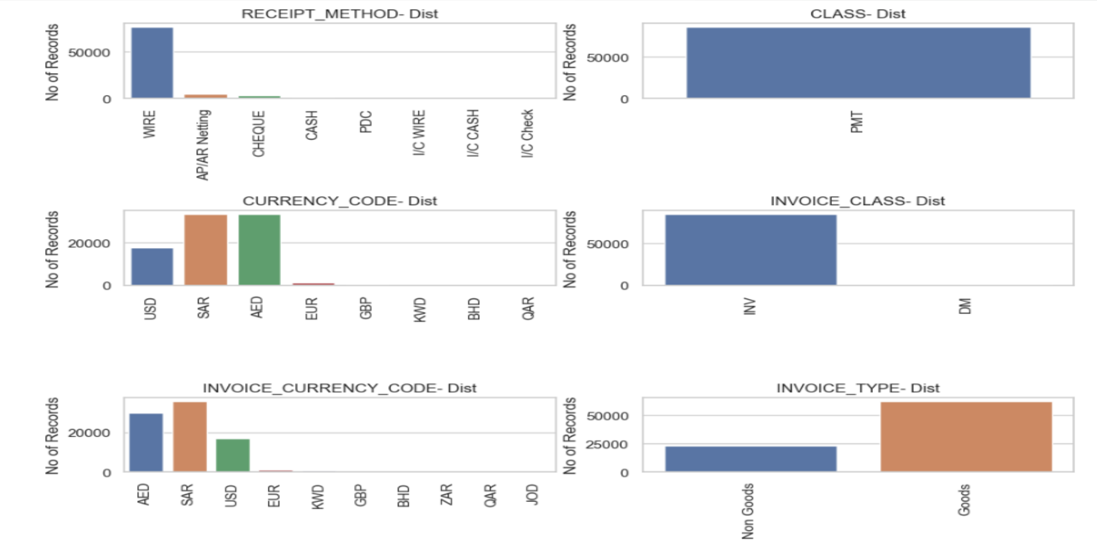
- The data has 93937 row and 16 columns
- Check for duplicate rows
- Receipt_Doc_No has missing values of 0.003% we can go ahead and drop this column
- Receipt_Date, Due_Date and Invoice_Creation_Date needs to be converted to Date type.
- From USD Amount we need to remove days extension.

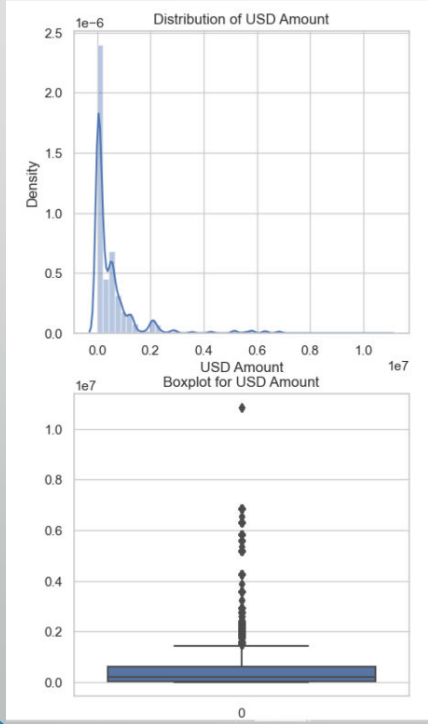
```
df_payment.info()
```

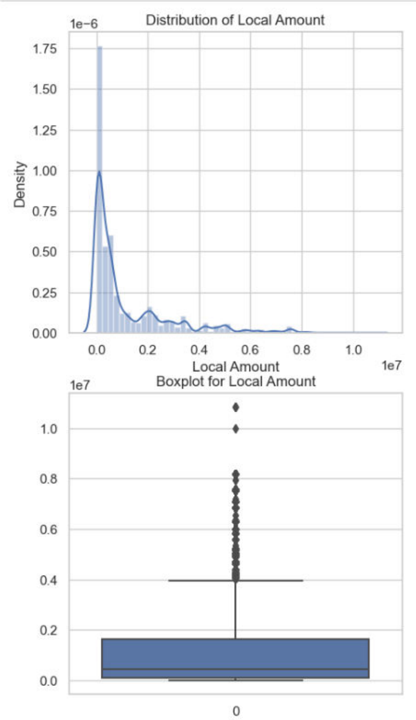
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 93937 entries, 0 to 93936
Data columns (total 16 columns):
 #   Column                 Non-Null Count  Dtype  
---  ------                 --------------  -----  
 0   RECEIPT_METHOD         93937 non-null  object 
 1   CUSTOMER_NAME          93937 non-null  object 
 2   CUSTOMER_NUMBER        93937 non-null  int64  
 3   RECEIPT_DOC_NO         93908 non-null  float64
 4   RECEIPT_DATE           93937 non-null  object 
 5   CLASS                  93937 non-null  object 
 6   CURRENCY_CODE          93937 non-null  object 
 7   Local Amount           93937 non-null  float64
 8   USD Amount             93937 non-null  float64
 9   INVOICE_ALLOCATED      93937 non-null  object 
 10  INVOICE_CREATION_DATE  93937 non-null  object 
 11  DUE_DATE               93937 non-null  object 
 12  PAYMENT_TERM           93937 non-null  object 
 13  INVOICE_CLASS          93937 non-null  object 
 14  INVOICE_CURRENCY_CODE  93937 non-null  object 
 15  INVOICE_TYPE           93937 non-null  object 
dtypes: float64(3), int64(1), object(12)
memory usage: 11.5+ MB
```

Univariate Analysis of Categorical

Shows the distribution of categories within each column and the count of each categories that influence the Customer Payment



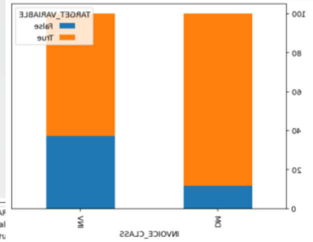


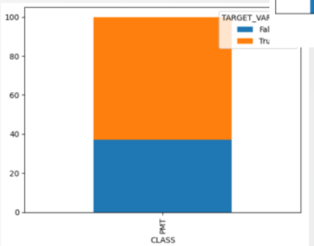


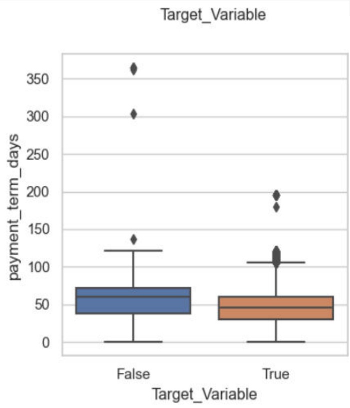
The distribution of
Numerical values within the dataset

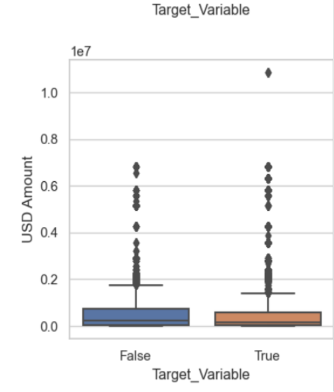
Calculation & Analysis of the Target Variable

```
df_payment['late_days']=(df_payment['RECEIPT_DATE']-df_payment['DUE_DATE']).apply(lambda x:x.days)
df_payment['payment_term_days']=df_payment['DUE_DATE']-df_payment['INVOICE_CREATION_DATE']
df_payment['Target_Variable']=df_payment['late_days'] > 0
df_payment['payment_term_days']=df_payment['payment_term_days'].apply(lambda x:x.days)
```

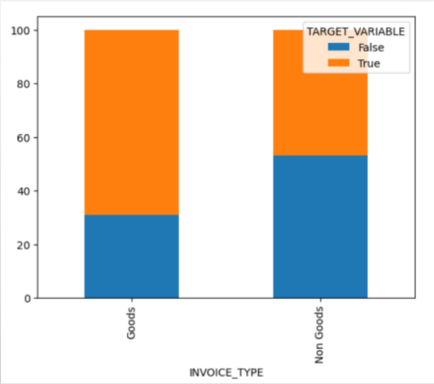


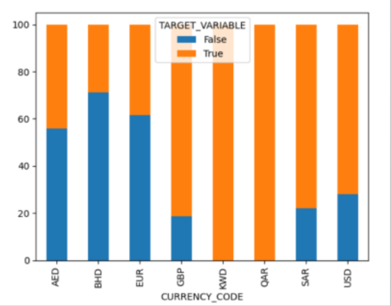


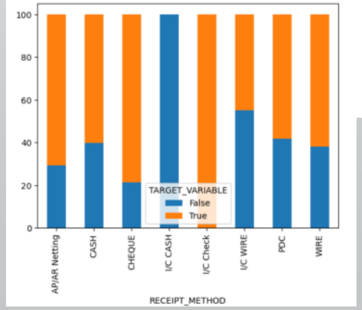




Analysis of Target variable with different variable within the dataset to understand the customer behavior.





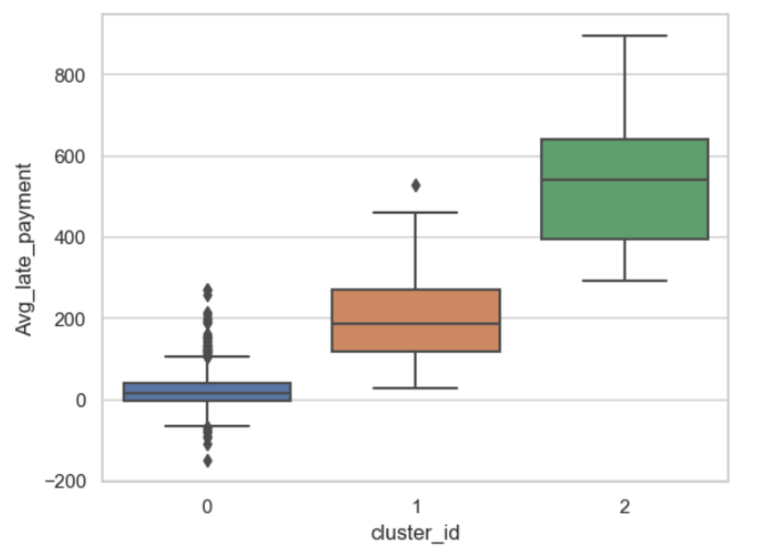


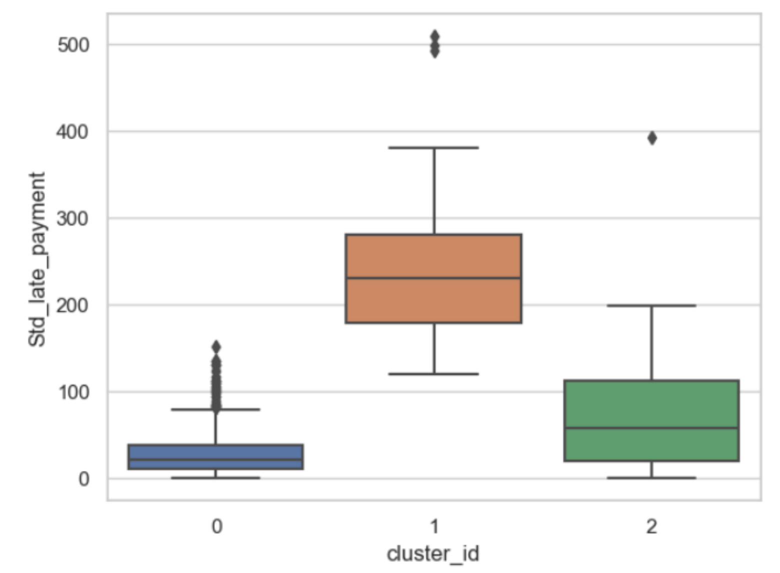
Clustering

Cluster 0 : Paying on Time
Prime Customers

Cluster 1 : Intermittently paying late
General Customers

Cluster 2 : Late payers
Problematic Customers





Random Forest Model

Random Forest Model is used for Model Building using Variables

- USD Amount
- Payment_Term_Days

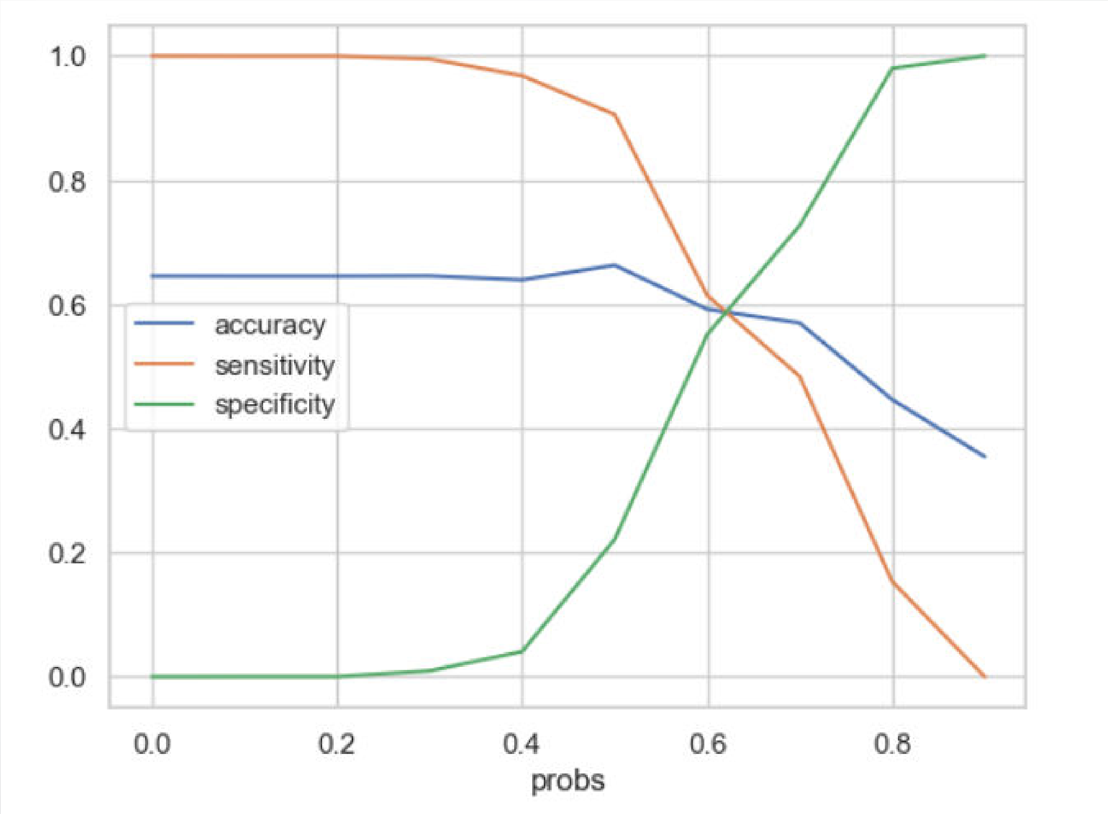
- Cross Validated Accuracy : 82%
- Accuracy of the Model : 83.64%

Logistic Regression Model

Logistic Regression Model is used for Model Building using Variables

- USD Amount
- Payment_Term_Days

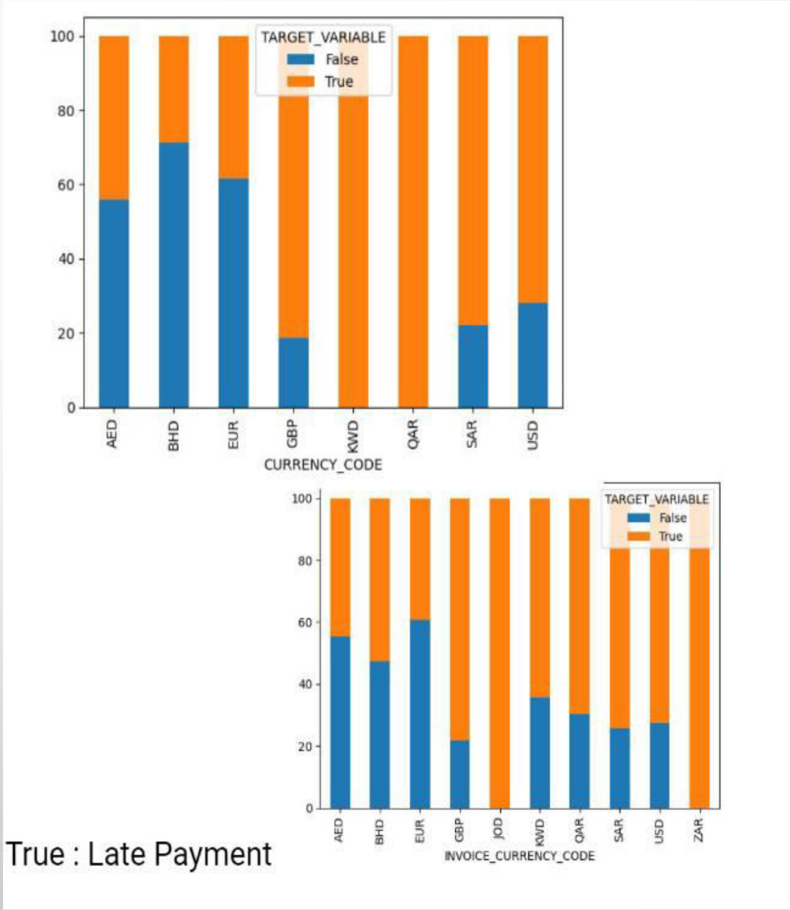
- Training Accuracy : 66%
- Test Accuracy : 65%



Observations & Suggestions

Less common currencies are associated with higher late payment rates

Investigate whether it is related to currency transaction and process associated with it.



Observations & Suggestions

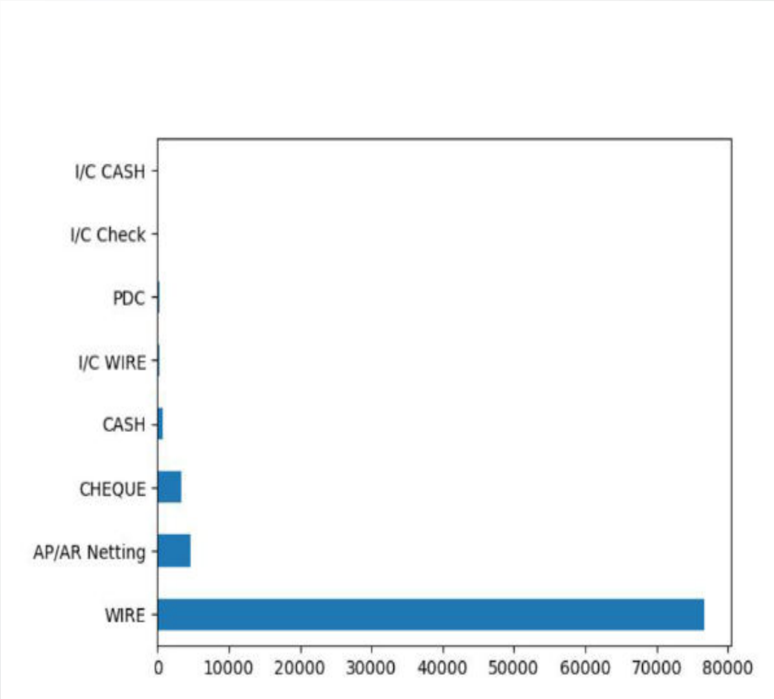
Most of the customers choose Wire as the mode of payment.

Can make wire transfer mode more faster & feasible.

Out of 85915 around 2761 payments were made before generation of invoice.

As we observe that the accuracy of Random Forest is better than Logistic Model we will use Random Forest Model to predict those customers who are likely to make late Payment

Timely follow –up of these customers can be ensured using the Model to facilitate the flow of cash.



Thank You