E-Commerce & Retail B2B 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 would like to better understand the customers' payment behaviour based on their past payment patterns (customer segmentation).
- Using historical information, it wants to be able 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.

Business Goal

- Schuster would like to better understand the customers' payment behaviour based on their past payment patterns (customer segmentation).
- Using historical information, it wants to be able 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.

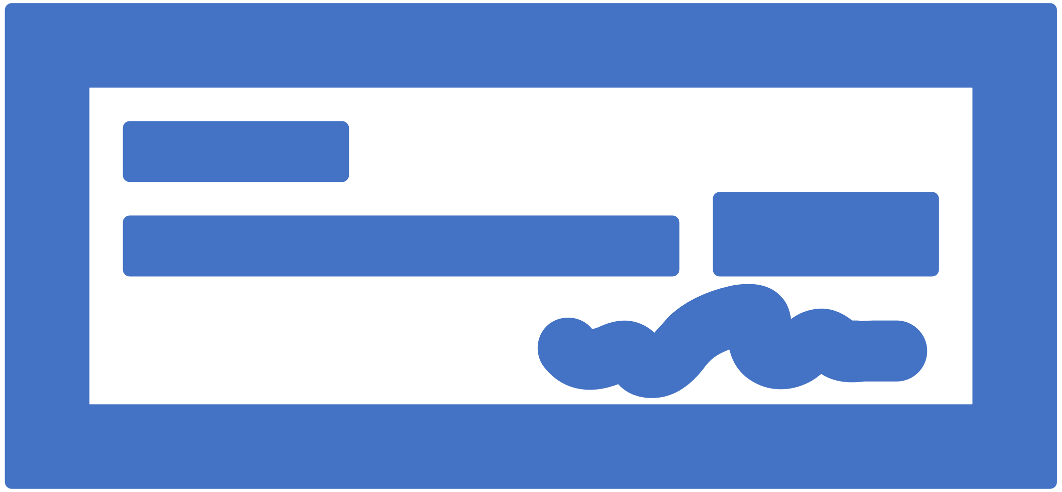
Data Understanding

- There are two specific tables, Past transactions record and open invoices.
- Using Past transactions record we need to do segmentation of the customers by deriving more variables if needed.
- Train the model based on past transactions data and predict if a particular customer will be prone to late payment or will pay in time.
- Aggregate the probability of the customers late payment on customer level

Data Preprocessing

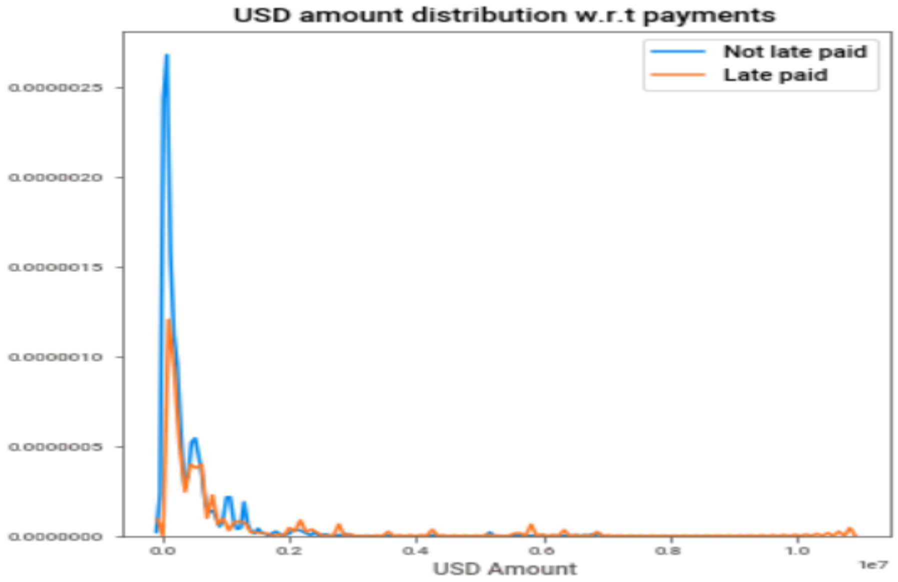
- Observed many columns like RECEIPT_DOC_NO, RECEIPT_METHOD, CLASS have mostly only one data point which is not useful for analysis. So dropped these columns.
- Derived the target variable using DUE_DATE and RECEIPT_DATE. Target variable – 1 (Late Payments)

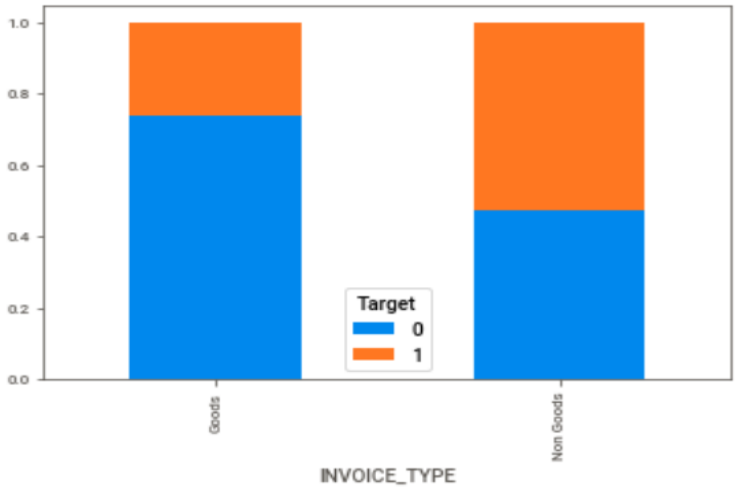
  Target variable – 0 (Not Late Payments)
- Derived Payment term variable from due_date and invoice creation date



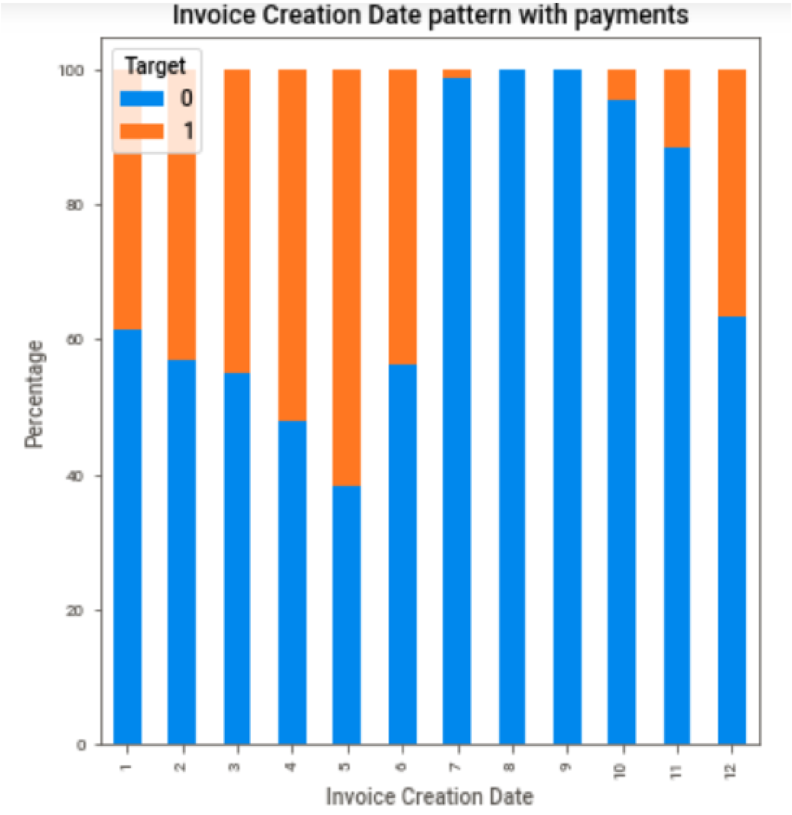
Data visualization

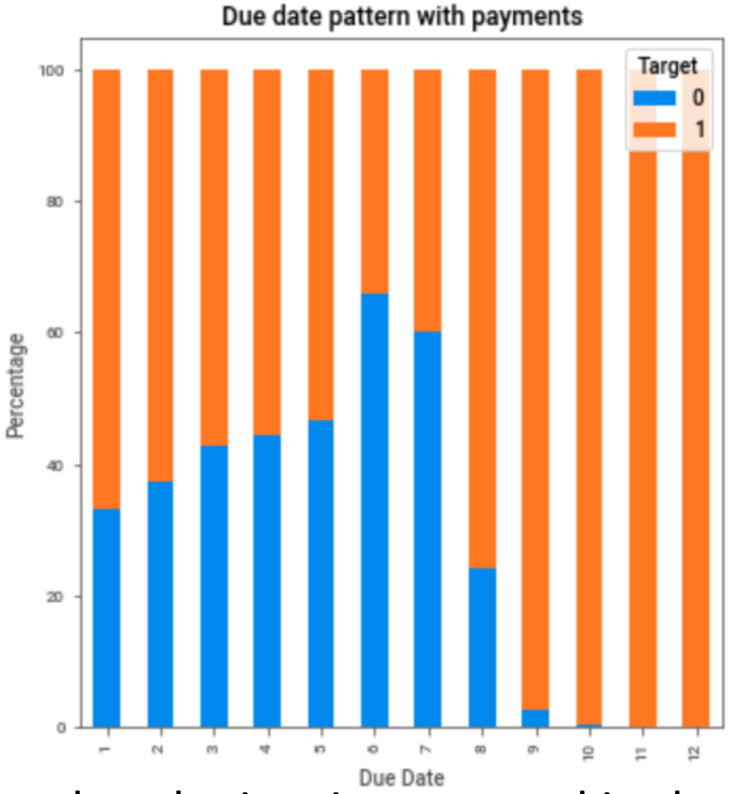
- In below plot, We can observe that higher value orders are being paid in time.





- Non Goods invoice type has more late payments when compared to Goods



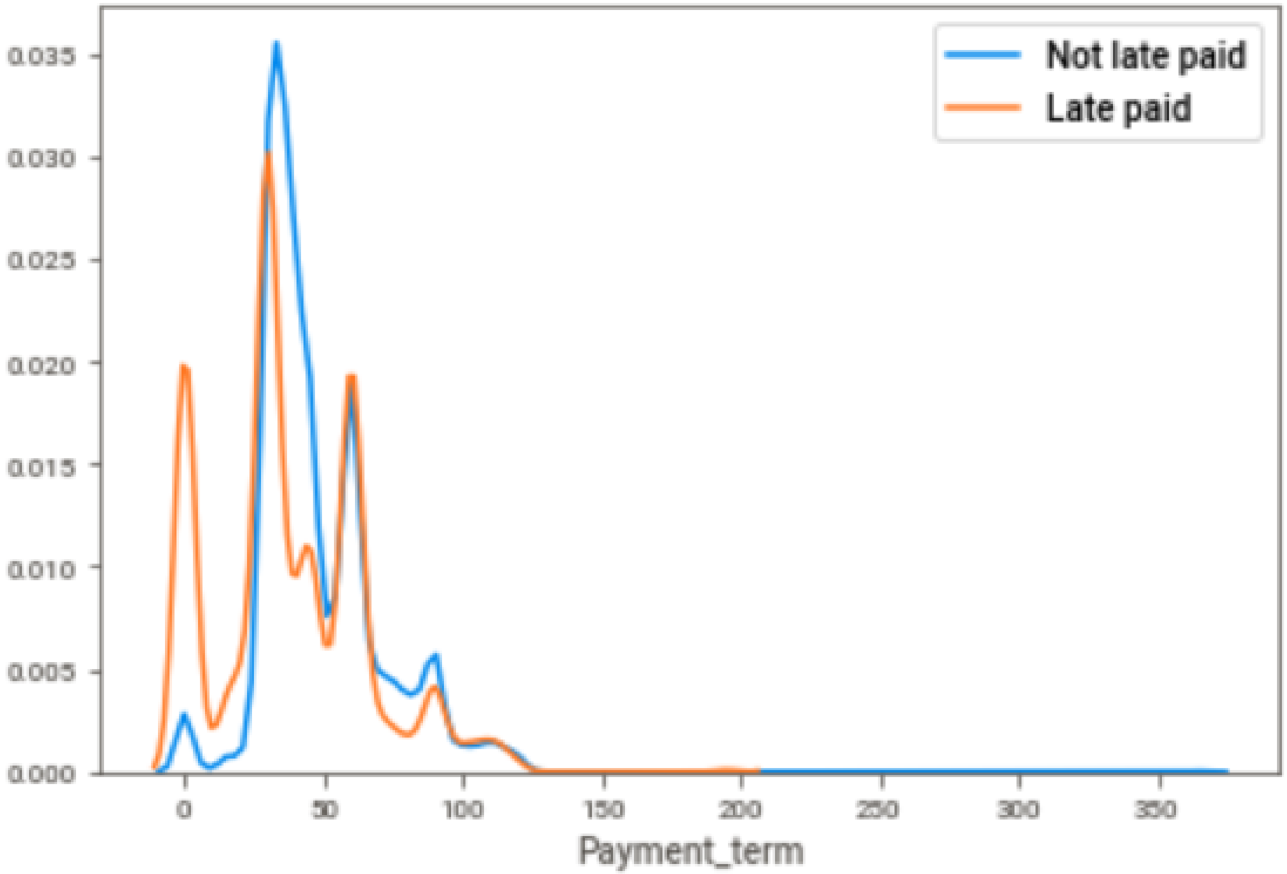


Data visualization

- We can observe that the invoices created in the months of July, Aug, Sept, Oct have high chances of not paying within the due date.
- Also when the due dates lie in the months of Sept, Oct, Nov and Dec. They are mostly not getting paid in time

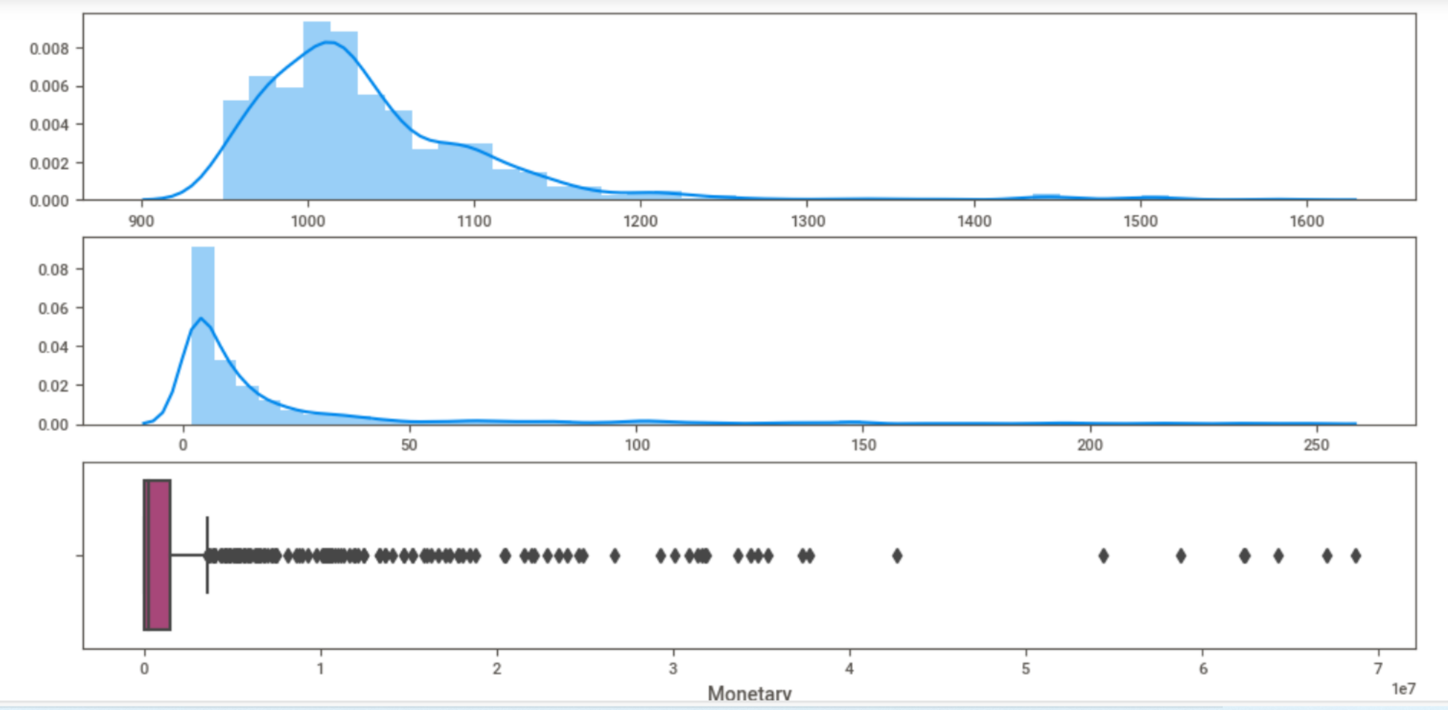
Data Visualization

- The invoices with payment term more than 200 days are being paid within due date



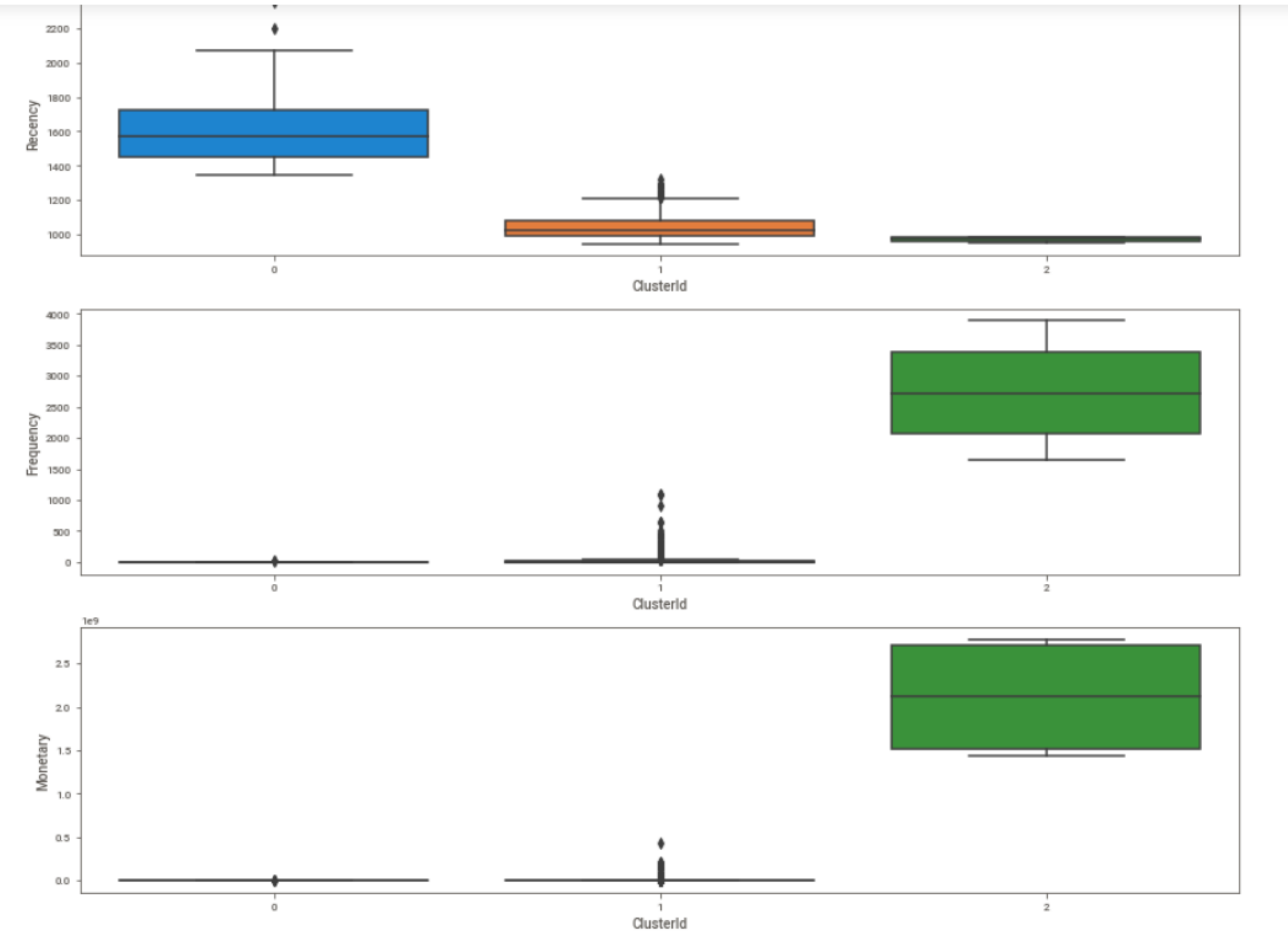
Customer Segmentation

- Derived new features i.e. Recency, Frequency and Monetary variables
- Applied boxcox transformation, outlier treatment and scaler transformation on the features before doing segmentation using k means clustering.
- After checking the elbow curve, decided on the cluster value as 3.
- Segmented the customer data into 3 clusters.



Customer segmentation

- The first cluster being high recency, low activity and low purchase values customers. They can be labelled as Occasional Engagers
- The second cluster being medium recency, medium activity and medium/low purchase valued customers. They can be labelled as Regular Customers
- The third cluster being low recency, high frequency and high monetary valued customers. They can be labelled as High-Value, High-Frequency Customers
- For the current problem statement we can concentrate more on high value customers payment as they bring more business and the loss if they dont pay their bills on time is huge.



Data Preparation and Modelling

Encode the categorical variables using dummy encoding

Many features like Recency, Frequency, monetary, and USD amount are right skewed. Hence need boxcox transformation on these.

Apply Minmax scalar on the transformed data.

Logistic regression

Use logistic regression model to get the initial model performance.

Train score : 0.65

Test score : 0.647

Confusion matrix and Classification report

is given here

```
[[1569 2974]
 [1232 6141]]
              precision    recall  f1-score   support

           0       0.56      0.35      0.43      4543
           1       0.67      0.83      0.74      7373

    accuracy                           0.65     11916
   macro avg       0.62      0.59      0.59     11916
weighted avg       0.63      0.65      0.62     11916
```

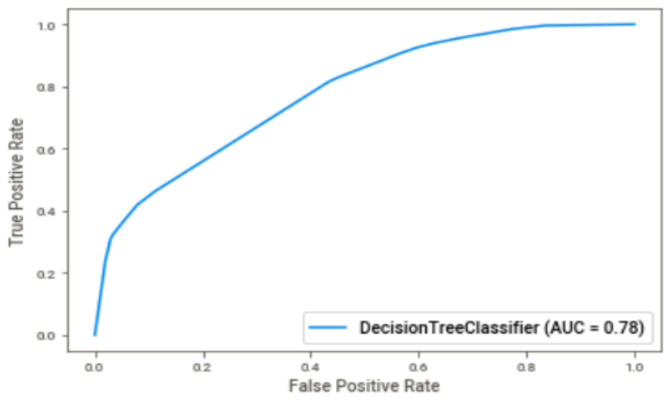
Data Modelling

- Using GLM technique, Checked the p values and dropped the columns with p value greater than 0.05 parallelly checking the VIF score(<5)
- Final important features are given in the right

Features	VIF
Frequency	3.21
Payment_term	3.07
INVOICE_CURRENCY_CODE_SAR	1.77
INVOICE_CURRENCY_CODE_USD	1.29
INVOICE_CURRENCY_CODE_KWD	1.02
INVOICE_CURRENCY_CODE_QAR	1.01
INVOICE_CURRENCY_CODE_GBP	1.00

Decision Tree Classifier

- Given is the ROC curve obtained after applying Decision tree classifier



Data Modelling

Using the grid search on decision tree classifier, got the best score as 0.80 and AUC as 0.92

Best model : DecisionTreeClassifier(max_depth=20, min_samples_leaf=5, random_state=42)

Random Forest

- Using RF got an oob score as 0.73 and AUC as 0.8
- Using grid search using RF got the best score as 0.8 and AUC as 0.91
- Best model : RandomForestClassifier(max_depth=20, min_samples_leaf=5, n_jobs=-1,
  random_state=42)

Prediction on Test data

- The Open invoice data column names were all adjusted according the train data.
- Filtered the data using AGE column having only positive AGE values
- Derived the features like RFM, clustered, payment_term for the test data.
- Applied boxcox and scaler transformation.
- Used the grid search on RF best model to predict the target variable.
- Also got the prediction probability against each predicted value

Summary

- Many of payments(62%) are being made after the due date, indicating a potential challenge in prompt payment adherence.
- A minority of payments are made on time, highlighting an opportunity for improvement in promoting prompt payment behavior.
- Notable trend indicates that invoices with due dates in September, October, November, and December are frequently paid after the due date.
- Late payments during these months could be influenced by seasonal factors, financial planning, or other external circumstances.

Business Insights for Customer segmentation

For Occasional engagers, Cross-selling or upselling initiatives to increase the overall transaction value.

For regular customers, Provide personalized offers and incentives to encourage repeat business for regular customers. Targeted promotions to maintain consistent transaction frequency.

For high value customers, Targeted promotions for high-value products or services.

Business Recommendations

- We could observe the higher value orders were being paid in time. To encourage this, Consider implementing a customer loyalty program or offering exclusive benefits to reward consistent and timely payments.
- Encourage even quicker payments by offering discounts or incentives for early payments. This can create a win-win situation for both the customer and your business.
- When non-goods invoices experience more late payments compared to goods invoices, it's essential to address the specific challenges associated with this invoice type.
- Provide flexible payment options for non-goods services. Offer online payment methods, electronic funds transfer (EFT), or automated clearing house (ACH) options to simplify the payment process.
- When invoices created in the months of July, August, September, and October have a higher likelihood of not being paid within the due date, it's important to address this pattern and implement strategies to improve timely payments

Thank you