RETAIL E-COMMERCE

Group Activity - December 2022

DSC43 - Srivatsa / kiran sai / MOHAMMED TAHER

Problem Statement

- 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.

Understand Historical Data of Received Payments

- Understand Historical Data of Received Payments
  - 15 Columns with 93937 Rows
  - Convert object Datatypes to Dates as required
  - Removed Unnecessary Columns

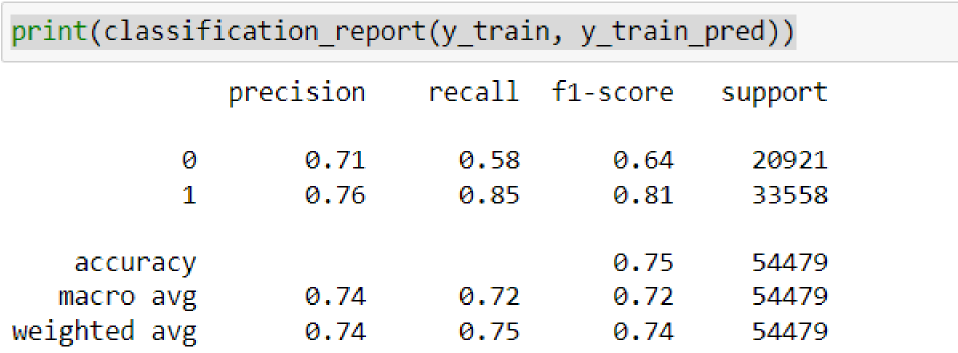
    'RECEIPT_DOC_NO','PAYMENT_TERM','CLASS','RECEIPT_METHOD','Local Amount','INVOICE_ALLOCATED’
  - Derived PAY_TERM based on Invoice creation and Due date
  - Derived Target (DELAYED) based on RECEIPT DATE after DUE DATE
  - Cleaned Data
    - PAY_TERM Cannot be negative
    - USD_Amount cannot be <= Zero

Customer Segmentation

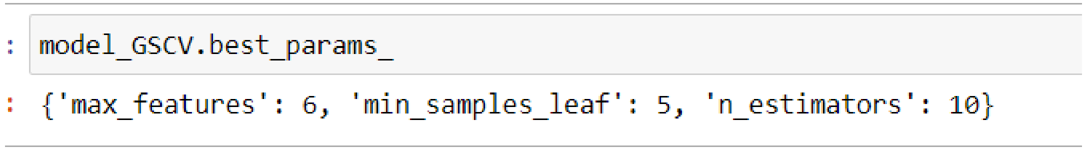
- Based on Mean / Std Deviation of PAY_TERM – Per Customer
- Prepared data by Scaling
- Using Kmeans
  - Determine optimum K value from Silhoutte Analysis
  - Identified Cluster_id for each customer (5 Clusters)
  - Renamed 0 – 4 as Customer Segment A to E

Model Preparation

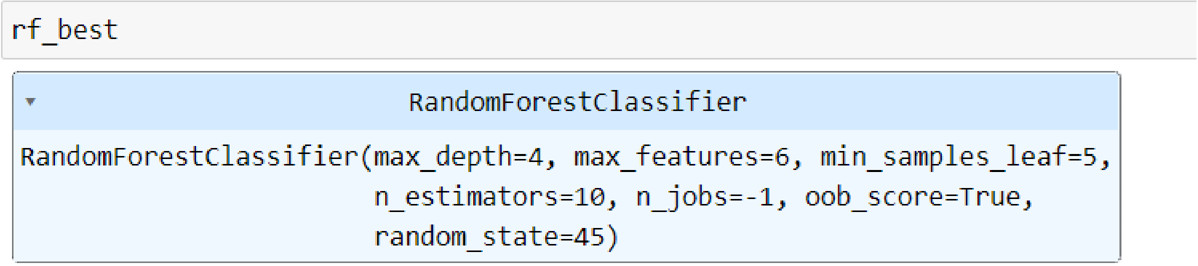
- Created Dummy Variables
- 70 : 30 Train Test Split performed
- Scaling of Numerical Variables – StandardScaler
- Used RandomForestClassifier
- Hyperparameter using GridSearchCV
  - Fitting 5 folds for each of 48 candidates, totalling 240 fits
- Identified Best Score of 0.74



```
print(classification_report(y_train, y_train_pred))

              precision    recall  f1-score   support

           0       0.71      0.58      0.64     20921
           1       0.76      0.85      0.81     33558

    accuracy                           0.75     54479
   macro avg       0.74      0.72      0.72     54479
weighted avg       0.74      0.75      0.74     54479
```



```
: model_GSCV.best_params_

: {'max_features': 6, 'min_samples_leaf': 5, 'n_estimators': 10}
```



```
rf_best

▾ RandomForestClassifier
RandomForestClassifier(max_depth=4, max_features=6, min_samples_leaf=5,
                       n_estimators=10, n_jobs=-1, oob_score=True,
                       random_state=45)
```

PREDICTING ON INVOICE DATA

- Removed Rows where Age is Positive – Already Delayed
- Derive the PAY_TERM based on AGE + (DUE_DATE - AS_ON_DATE)
- USD Amount – has Comma Seperator Addressed that – converted to int
- Recreated the Customer Segmentation on the Invoice Data
- Ensured that the Column names and Values of Categorical Variables are inline with Received Payment Dataset
- Created the Dummy Variables
- Predicted Using the Same Rf_Best Model that was selected earlier

- Out of **495** Unique Customers - Over **200** Customers are Predicted to Delay Payments