Problem Statements & Assignment Planing

NGUYEN HUU LIEM

Problem Statement

Stakeholders: Schuster is a multinational retail company dealing in sports goods and accessories.

Business Requirements: Identify Late Payment invoices

Data Analyst Requirements: build a model with the primary objective of identifying important predictor attributes that will help the business understand indicators of late payment.

Model Requirements: Classification Model, Regression Model, with Evaluation, and able to predict Open Invoices Data.

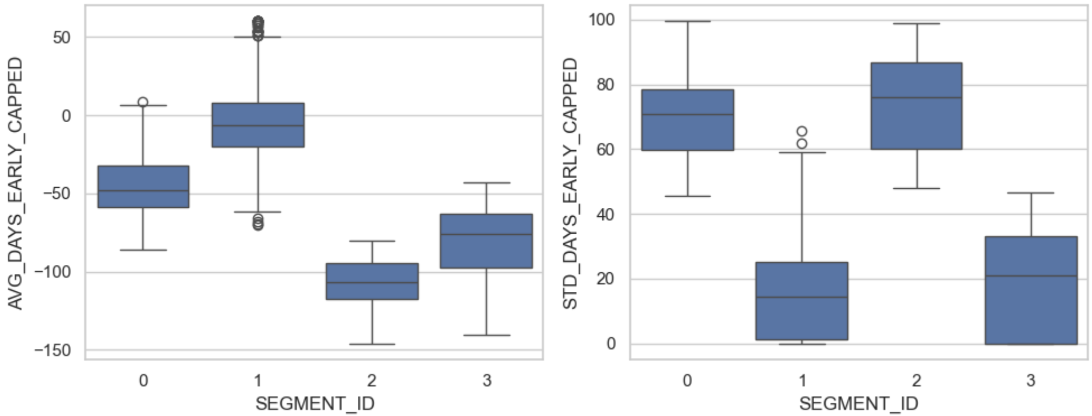
Data Preparation and EDA

NGUYEN HUU LIEM

Data Preparation

- Remove unecessary columns based on the Data Dictionary- Drop cols with just a single variable
- format date variables
- calculate TARGET variable
- Calculate Payment Term in days
- Identify and remove Outliers from “USD Amount” and “PAYMENT_TERM_DAYS” using percentiles. Removing outliers will provide a better Scaler later on.

Customers Segmentaion



We have ourselve 4 customers segments with different characteristics:

- `SEGMENT_ID` 0 : Customers who make very late payments, with inconsistent late payment durations.

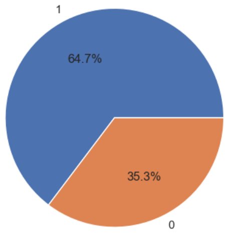
- `SEGMENT_ID` 1 : Customers who consistently make on-time payments.

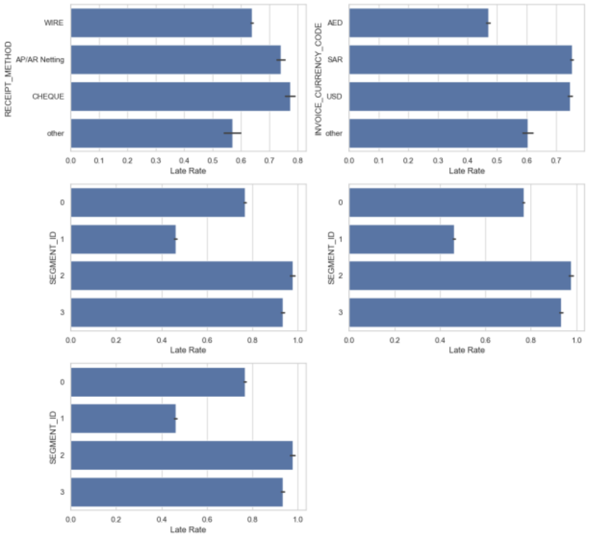
- `SEGMENT_ID` 2 : Customers who are kinda late, and have inconsistent late payment durations.

- `SEGMENT_ID` 3 : Customers who consistently make late payments.

EDA

- About 2/3 payments are Late!

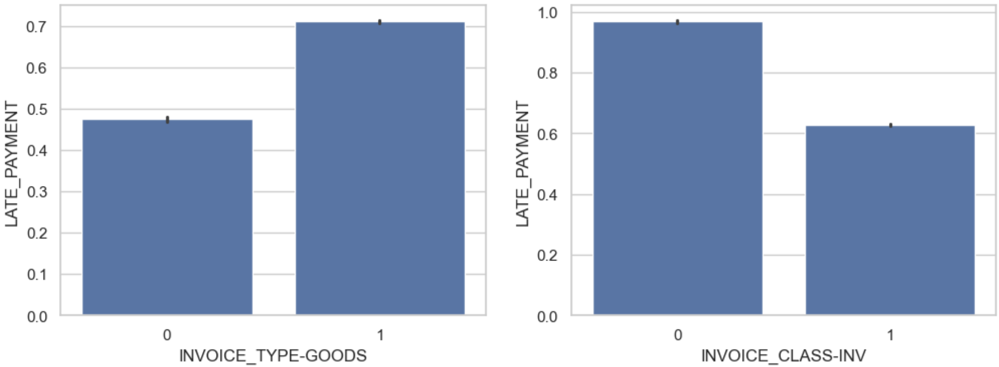




Insights on the Categoricals Variables Analysis:

- Payments with "Cheque" payment method have the highest chance of being late.
- Invoice's currency with "USD" and "SAR" currency have much higher change of being late, compared to "AED" and "other".
- Customers in SEGMENT_ID "3" have very high chance of paying late, while customers in SEGMENT_ID "1" are less likely to.

EDA



Insights on the binary variables:

- With Invoice type created for "physical goods", the change of getting a late payment is higher than Invoice created for "services"
- Invoice classes "Invoice" have lower change of a late payment, compared to "other" classes

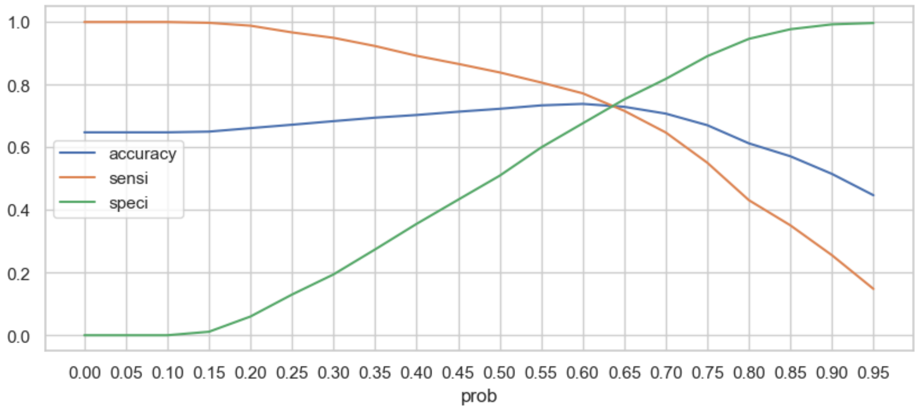
Model building & Evaluation

NGUYEN HUU LIEM

Test-train split and scaling

Logistic Regression

- The logistic regression model reach 0.73 accuracy, sensitivity, specificity
- This will be the benchmark for our next models



Random Forest

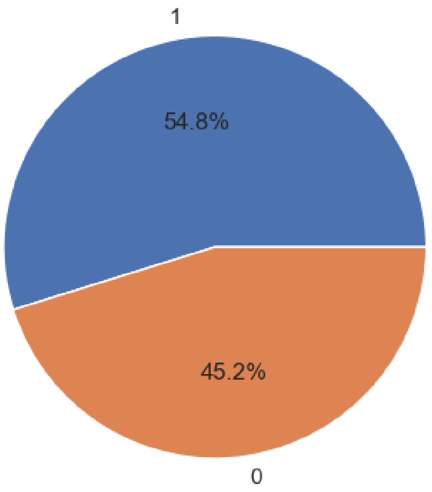
The Random Forest Classifier is built with the score of 0.92, and the params as follow:

- max_depth: 23
- max_features: 5
- min_samples_leaf: 1
- min_samples_split: 2
- n_estimators: 150

	precision	recall	f1-score	support
0	0.91	0.86	0.88	9166
1	0.92	0.95	0.94	16621
accuracy			0.92	25787
macro avg	0.92	0.90	0.91	25787
weighted avg	0.92	0.92	0.92	25787

- Good performance on class 1 (Late Payment) with high recall and F1-score. This means that our model can predict most of the late payments.
- High accuracy score on both train and test set, indicates a stable model have been built.

Prediction on Open Invoices



- Out of ~28k Invoices predicted, there are 55% of Invoices were predicted as will be late.

Thank You!