Domain Oriented Cast Study

E-Commerce & Retail B2B

upGrad & IIITB | Data Science Program

By:
Ranjith Madhavan
Gopi Jagini
Satya Ranjan Padhiary

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, some of the vendors tend to make payments late.
- Schuster would thus try to understand its customers' payment behavior and predict the likelihood of late payments against open invoices.
- To help sales team chase the vendors to complete the payment before due date.

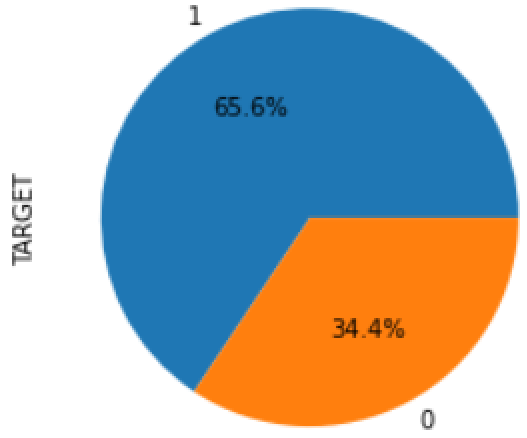
Goal of the Case Study

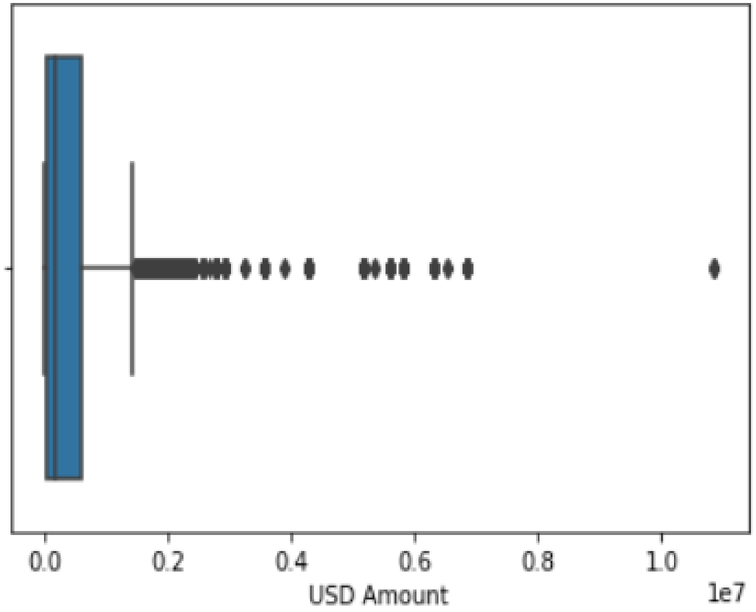
- To segment the customers based on their past payments data to better understand the customers' payment behavior.
- Design a predictive analytical model to predict the likelihood of delayed payment against open invoices.
- Apply new age ML techniques to classify and predict the default customers.
- To draw business insights that can help Schuster redefine the business approach accordingly.

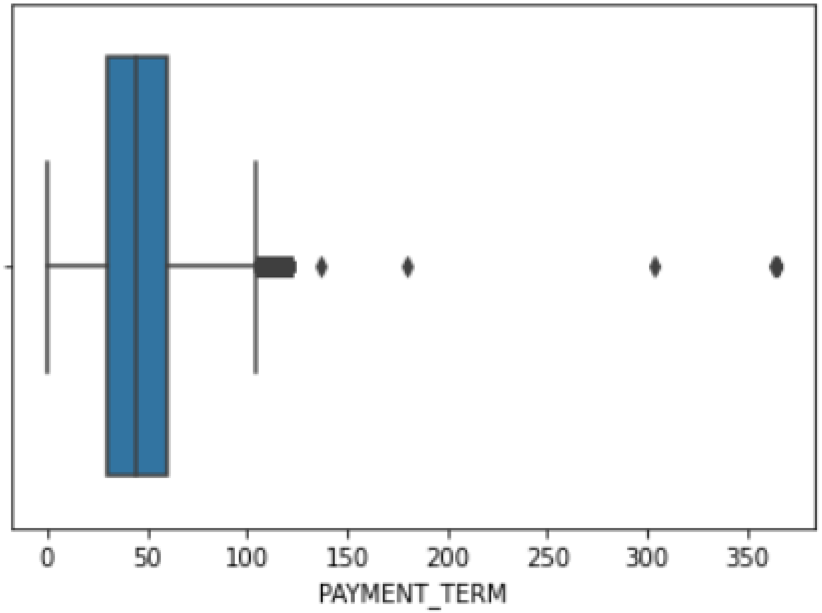
Data Preparation

- Started with the Received_Payments_Data file as input, which contains past payments data of Schuster's vendors
- A few values of the RECEIPT_DOC_NO were missing (0.03%). So, removed respective rows
- Converted CUSTOMER_NUMBER from 'int' to 'string' type
- Converted all date columns to 'datetime' type
- Calculated the days of PAYMENT_TERM as the difference of DUE_DATE and INVOICE_CREATION_DATE
- Dropped the rows with invoice amount zero or negative
- Imputed the rows with negative values in the PAYMENT_TERM column. Having lesser data than the invoice data means no sense in this analysis
- Created a new column 'AGE' as the days of difference RECEIPT_DATE and INVOICE_CREATION_DATE
- Removed unwanted columns and columns which are not present in both Received_Payments_Data and Open_Payments files
- Created the TARGET variable to map defaulters with 1 and on time payers with 0
- Modified the values of Open_Payments columns to match with the Received_Payments_Data columns

Exploratory Data Analysis (EDA): Univariate







TARGET variable shows 66% customers with late payment and 34% with on time payments

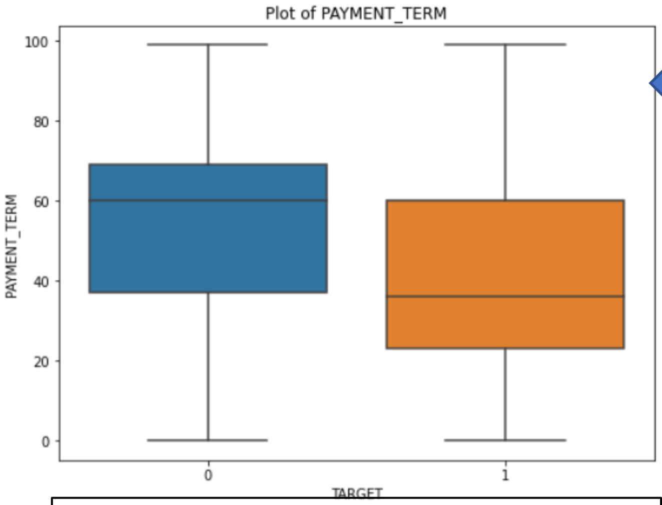
Median of the TRANSACTION AMOUNT is 20000USD, but there are a few outliers with large amounts

75% of the PAYMENT_TERM days are less than 60 days

Note: Most common method of payment is WIRE, followed by AP/AR NETTING and CHEQUE. All other payment methods are negligible

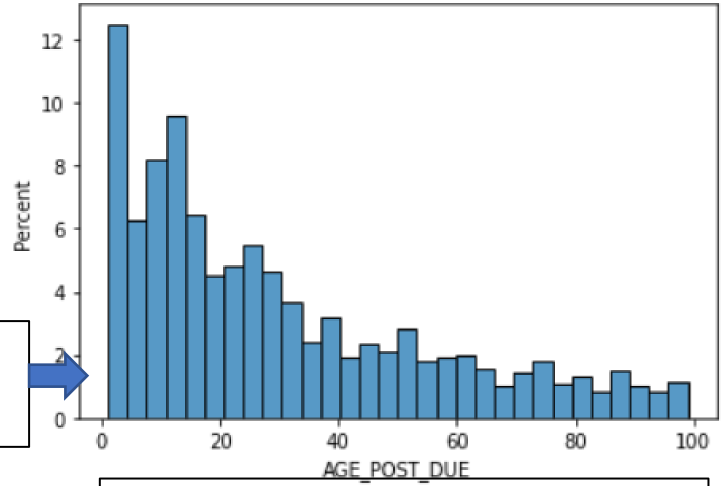
* Refer Python file for detailed analysis

Exploratory Data Analysis (EDA): Bivariate



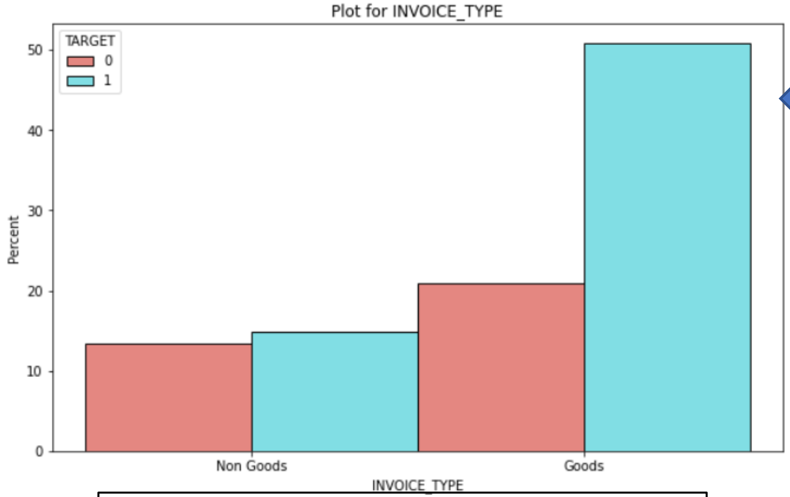
PAYMENT_TERM vs TARGET

Default Payments tend to happen when payment term is shorter



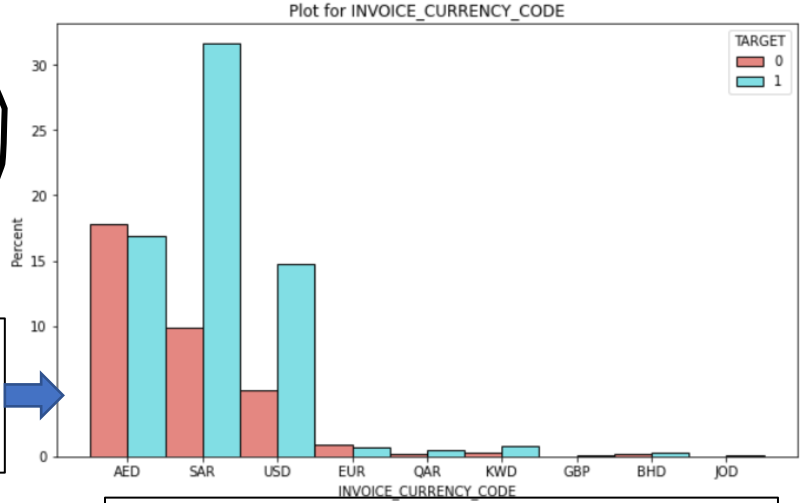
AGE_POST_DUE vs TARGET

75% of the people who default make payment within 78 days post Due date



INVOICE_TYPE vs TARGET

Goods category has generally higher chances of defaulting compared to Non-Goods



INVOICE_CURRENCY_CODE vs TARGET

SAR & USD have more percentage of defaulters compared to other currency

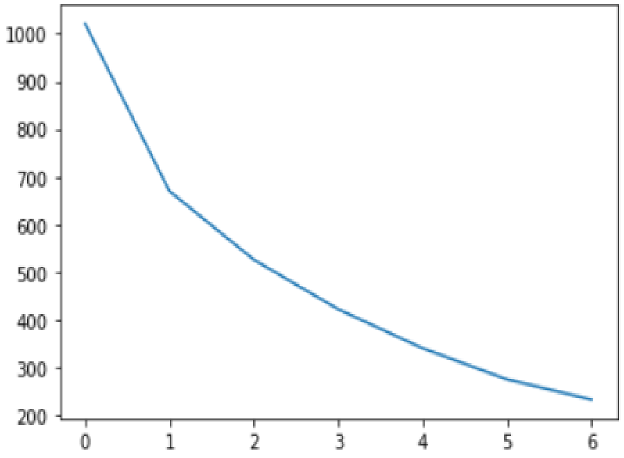
* Refer Python file for detailed analysis

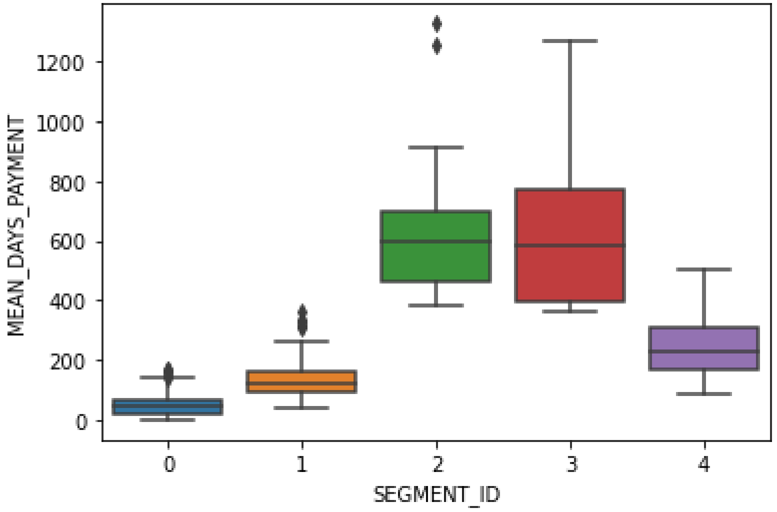
Customer Segmentation: K-Means Clustering

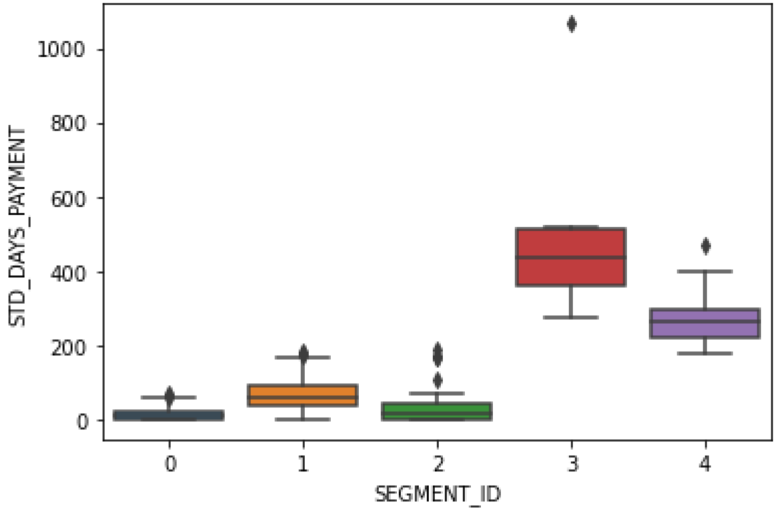
Note: Two new variables, MEAN_DAYS_PAYEMENT (mean) & STD_DAYS_PAYMENT (Standard Deviation) have been derived against each CUSTOMER_NUMBER to be given as input to the K-Means clustering for Customer Segmentation.

AGE variable the difference between Payment date and Invoice Date.

Rescaling through Standardization has been performed on the derived variables for clustering







SSD Curve

The curve is stabilizing after index 3 (k=5). So optimal no of clusters is 5.

Insight

From Above Graphs: SEGMENTS '0' and '1' have more good customers having less MEAN and STD of days of payment

*Note : Segment labels may vary every time K-Means clustering is re-run

* Refer Python file for detailed analysis

Model Building: Random Forest Classification

- Built Random Forest Classifier to predict the payment defaulters
- Performed feature importance to identify the critical features for the model
- Also, performed Hyperparameter tuning to improve the model accuracy

Before hyper-parameter tuning: Train Data	Before hyper-parameter tuning: Test Data		After hyper-parameter tuning: Train Data	After hyper-parameter tuning: Test Data
83.0%	82.6%	**ACCURACY**	87.6%	86.1%
0.63	0.64	**RECALL**	0.74	0.72

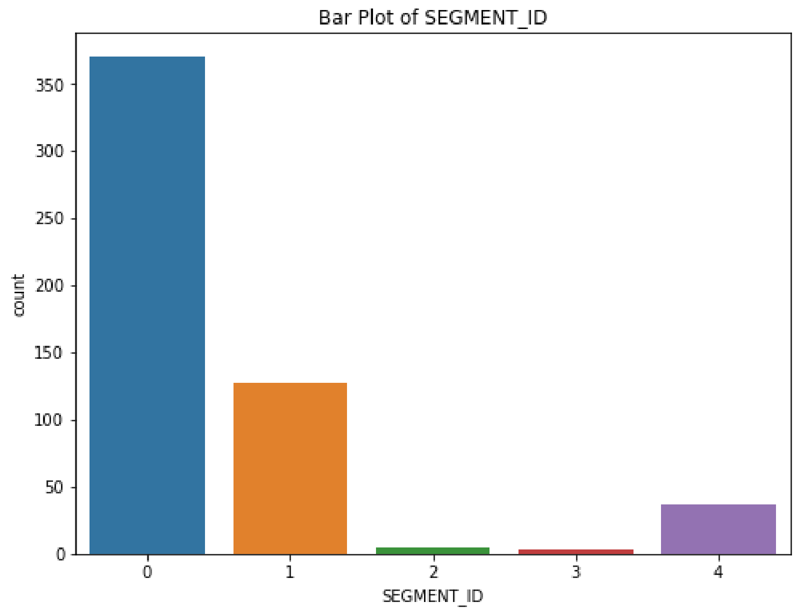
Before hyper-parameter tuning **After hyper-parameter tuning**

Train Data Test Data

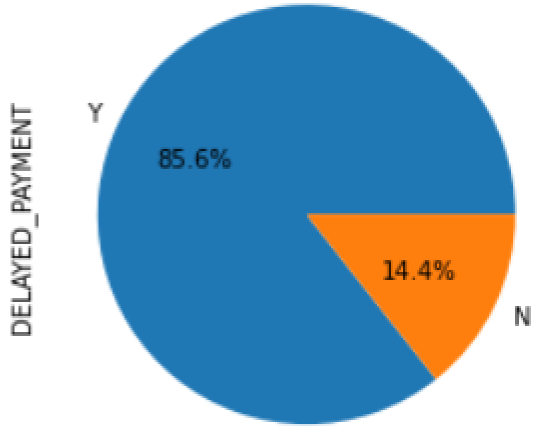
* Refer Python file for detailed analysis

Analyzing Open Invoice Data

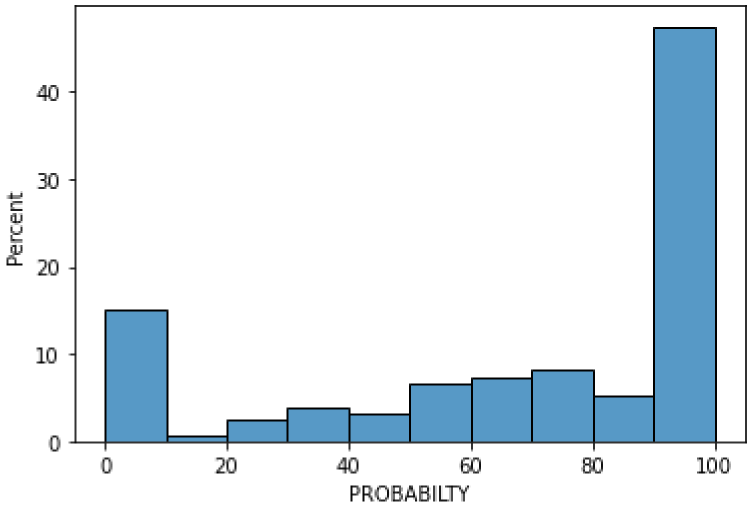
- Performed Data Pre-Processing before applying the classification model
- Applied the Optimized Random Forest Classifier model on Open Invoice Data
- Concatenated customer number with each prediction
- The prediction is performed at transaction level
- Prediction on late payments is aggregated at customer level



Most customers with open invoice belong to segment 0 and 1



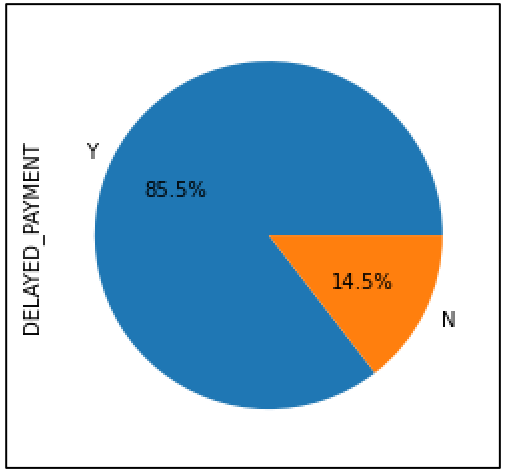
- 14% of the customers are predicted to pay on time on all transactions.
- 86% of the customers are likely to delay on at least one of their payments.



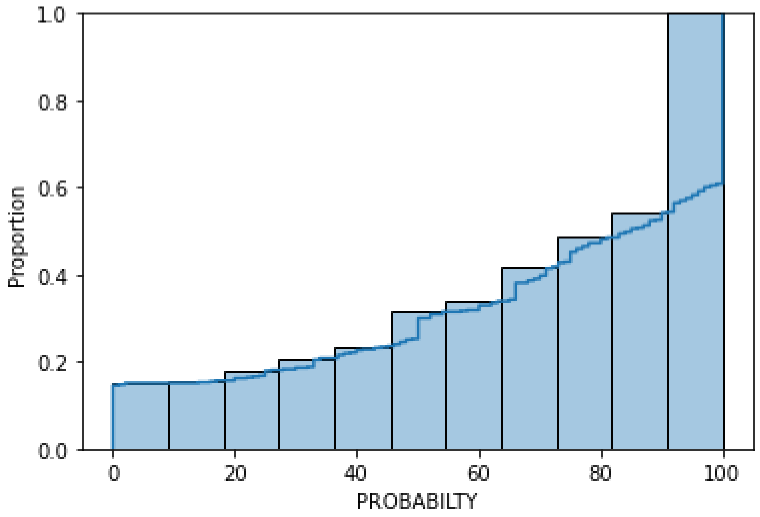
50% of the customers are likely to default on ~100% of their invoices

Business Insights

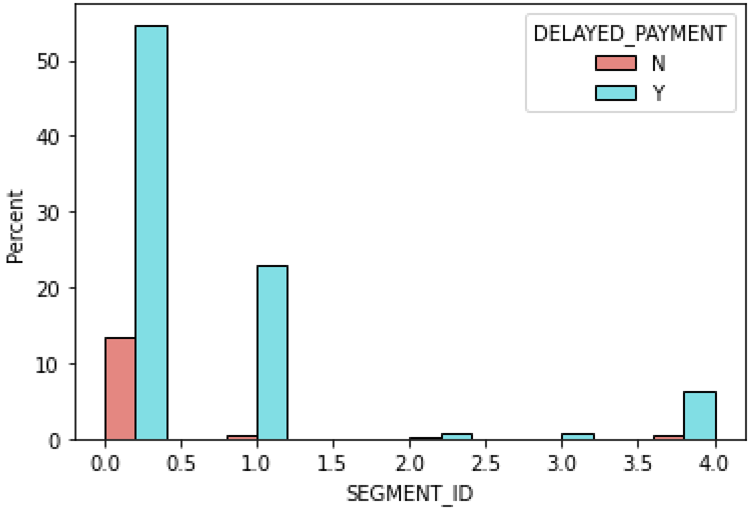
- 75% of Payment_Term is less than 60 days. And the late payments are close to the due date. So, relaxing the due date slightly will help improve the loyalty in customers
- Customers falling in GOODS category should be focused and targeted more to reduce the defaulter rate
- SAR has more no of defaulters and AED has less (in terms of percentage). The policies of AED customers should be replicated to other currency customers to reduce the payment issues
- Customers falling into segment '0' are more loyal than other segments with high percentage of on time payments. Both data sets have most no of customers falling in segments '0' and '1'
- Keeping a threshold of probability (E.g.: 50% or 60%) will help us identify critical defaulters and can be targeted more



86% the customers are likely to delay on at least one of their payments



70% of the customers default on more than 50% of the invoices



Segment 0 has most of the good customers