E-Commerce & Retail B2B Case Study

Understanding the customers’ payment behaviour

Group Members - Aakriti Singh, Aamir Farhan Sheikh, Kashif Khalid Ahmed

The Problem Statement

Company

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.

Problem statement

Unfortunately, not all vendors respect credit terms and some of them tend to make payments late. The company has some employees who keep chasing vendors to get the payment on time; this procedure nevertheless also results in non-value-added activities, loss of time and financial impact.

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.

Data Available

Received Payment Data

This data contains the information of all the transactions that have been performed with various vendors in the past.

Open Invoice Data

This data essentially contains the information of all the invoices that are open, i.e. that haven't been paid yet.

Data Dictionary

The data dictionary Excel workbook for this assignment contains two worksheets, which have the data dictionaries for the two datasets.

Case Study Goal

To summarise, as a business analyst, you want to find the answer to these questions:

- How can we analyse the customer transactions data to find different payment behaviours?
- In which way can you segregate the customers based on their previous payment patterns/behaviours?
- Based on the historical data, can you predict the likelihood of delayed payment against open invoices from the customers?
- Can you draw any business insights based on your developed model?

Execution

Steps Followed

Importing all libraries and datasets for understanding data

customer segmentation based on their past payment patterns

Model Evaluation, Finalizing the model, Model Evaluation on Unseen Data

Reading Data

EDA

Customer Segmentation

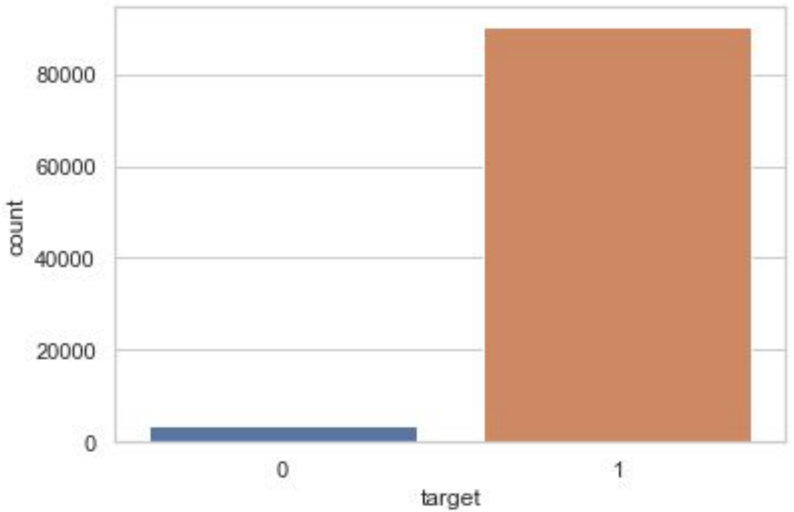
Data Preparation

Building Model

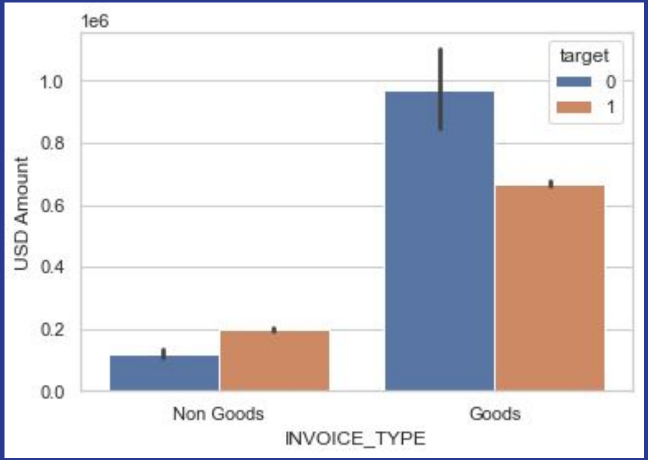
Performing Univariate and Bi-Variate analysis on Train data

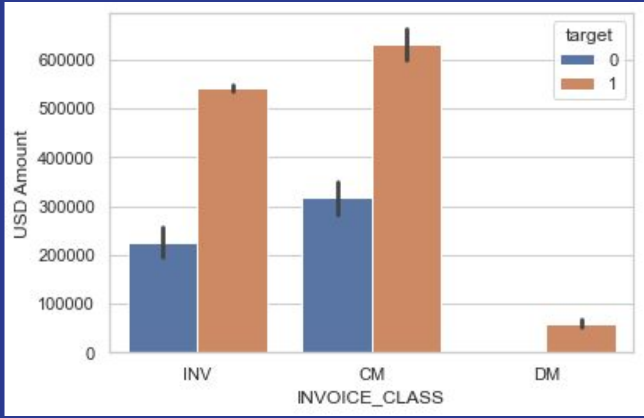
standardizing Data, Creating Dummy, Splitting data

Observations



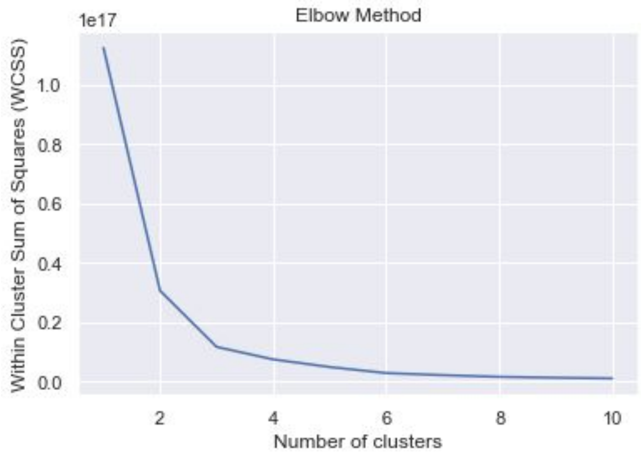
- Here we can see, approximately 4% of the customers are marked as 'Delayed'
- Clearly class imbalance is the issue and we will deal with it in the model building process



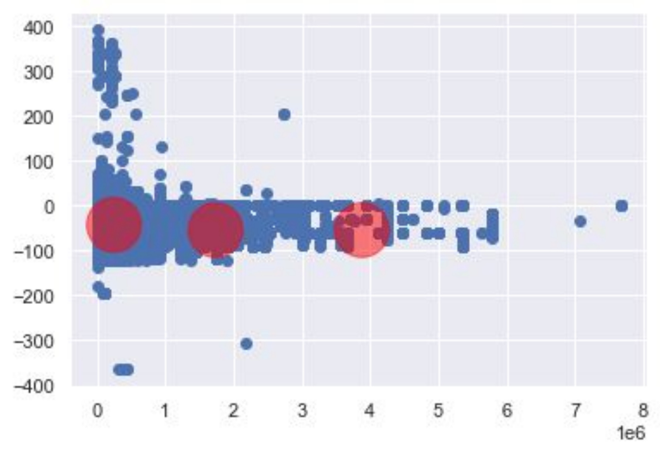


- Invoice amount showing pretty high for delayed payment customers in Goods invoice type
- Credit card payment mode accounts highest invoice amount across all the invoice classes for on-time customers

Observations



Based on the Elbow method, we could conclude that the number of clusters should be 2 or 3



We can see that average days of the payment time are segmented in three main zones:

a. 0-1 standard deviation of payment time
b. 1-2 standard deviation of payment time
c. 4 standard deviation of payment time

Observations

The first model we built is a Logistic regression model which acts as a baseline model for us. The scores obtained from this model are:

train score 0.9888981826477073
test score 0.9885032999787098

confusion Matrix is :

[[817   275]
[49 27041]]

ROC-AUC score test dataset: 0.9936543092689271
precision score test dataset: 0.9899326402108655
Recall score test dataset: 0.9981912144702842
f1 score test dataset :   0.9940447744734037

Challenges related to imbalanced dataset

1. Biased predictions
2. Misleading accuracy

We will check with two efficient techniques: ADASYN and SMOTE+TOMEK

Results after applying SMOTE+TOMEK Combining Oversampling and Undersampling

Accuracy: 0.962884110425094
F1 score: 0.980322070885695
Recall: 0.9617940199335548
Precision: 0.9995779943221055

clasification report:

	precision	recall	f1-score	support
0	0.51	0.99	0.67	1092
1	1.00	0.96	0.98	27090
accuracy			0.96	28182
macro avg	0.76	0.98	0.83	28182
weighted avg	0.98	0.96	0.97	28182

confussion matrix:
[[1081   11]
[1035 26055]]

So, we will finalize the SMOTE+TOMEK model as it's giving the better result across all the metrics

Observations

Top 20 features as per the feature-importance of Random Forest model

age
payment_term_immediate payment
cbrt_usd_amount
invoice_type_non goods
invoice_currency_code_sar
invoice_currency_code_usd
payment_term_immediate
invoice_class_inv
customer type_related party
payment_term_30 days from eom
payment_term_30 days from inv date
payment_term_60 days from inv date
payment_term_cash on delivery
invoice_currency_code_eur
payment_term_60 days from eom
invoice_class_dm
invoice_currency_code_bhd
payment_term_15 days from eom
invoice_currency_code_kwd
payment_term_90 days from eom

Results from Final Model

Accuracy: 0.9778936910084451
F1 score: 0.9883692709791841
Recall: 0.9771502399409376
Precision: 0.9998489140698772

clasification report:

	precision	recall	f1-score	support
0	0.64	1.00	0.78	1092
1	1.00	0.98	0.99	27090
accuracy			0.98	28182
macro avg	0.82	0.99	0.88	28182
weighted avg	0.99	0.98	0.98	28182

confussion matrix:
[[1088 4]
[619 26471]]

- So, we can observe that all score of the metrics got improved in this finalized model

Observations

Model Prediction on Unseen Data (Open Invoice Data)

	Cust id	actual	predicted	is_delayed
67288	34647	1	1	yes
60971	7530	1	1	yes
53170	7588	1	0	no
39162	45720	0	0	no
15138	2624	1	0	no
20187	20844	1	1	yes
59331	3997	0	1	yes
30267	34876	1	0	no
37858	45720	1	0	no
7244	3927	1	0	no

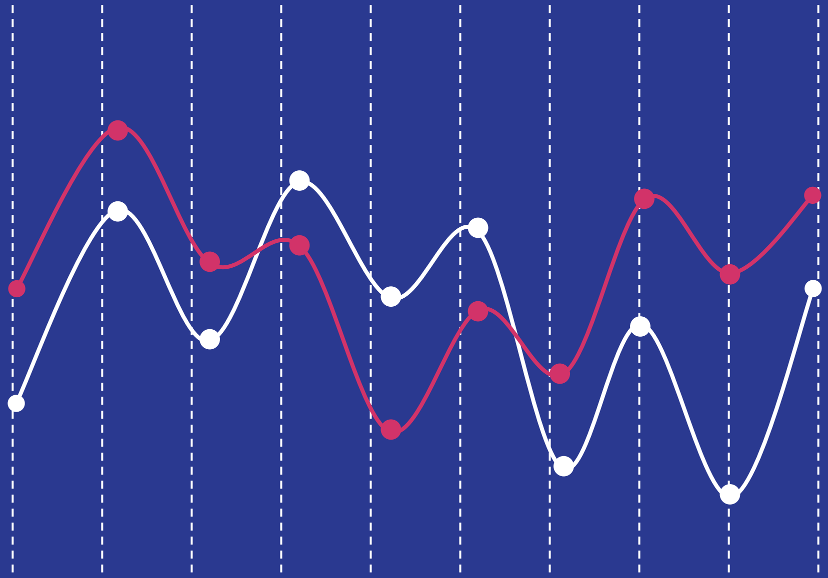
Finally, we observed that there are 28287 customers out of 88201 unseen records are predicted as delayed customers.

Top 10 factors / important predictors

- age
- payment_term_50% advance payment and 50% upon receiving the shipment
- payment_term_eom
- payment_term_lcsight
- payment_term_on consignment
- invoice_currency_code_eur
- invoice_currency_code_gbp
- invoice_currency_code_kwd
- invoice_currency_code_qar
- invoice_type_non goods

Business Recommendations

- We should focus more on the time difference between Due Date and Invoice Payment Date
- Payment terms: 50% advance payment and 50% upon receiving the shipment, eom, lcsight and on consignment variables need to be considered with greater attention.
- Where the invoice currency codes are eur, gbp, kwd and qar, the risk is higher of delay payment.
- Invoice type non-goods has lower impact than Goods invoice type in delayed payment.



Thank You!