E- Commerce and Retail B2B Case Study

Submission By:

Namrata Sidharth Khobragade
Deepa M
Royal Singh

Addressing the problem statement & defining objectives:

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 levies heavy late payment fees, although this procedure is not beneficial to either party in a long-term business relationship. 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. Schuster would thus try to understand its customers' payment behaviour and predict the likelihood of late payments against open invoices.

Goal

- Schuster would like to better understand the customers' payment behavior 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 prioritize their work in following up with customers beforehand to get the payments on time.

Approach :

❑ Data Reading & understanding

❑ EDA

❑ Data Cleaning & Feature Engineering

❑ Customer Segmentation (K- Means Clustering)

❑ Data Preparation/Splitting & Model Building

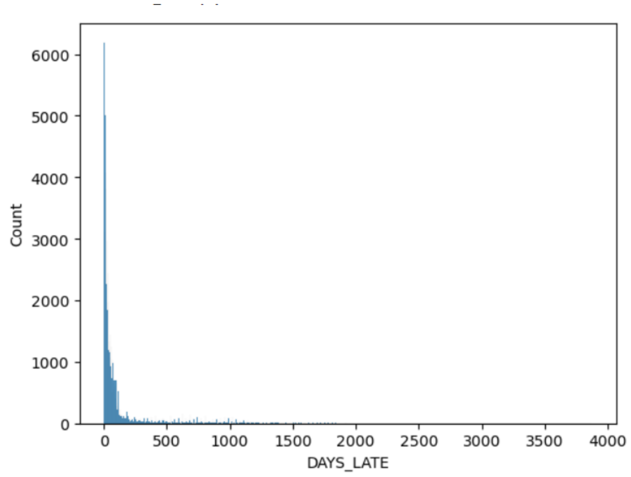
❑ Model & feature tuning and Metric analysis

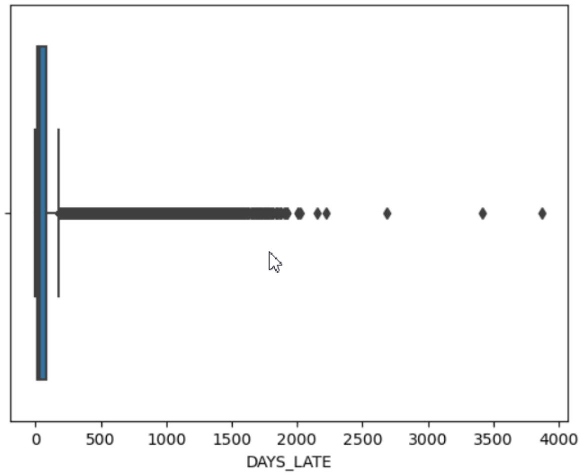
❑ Model Testing on test set of historical data

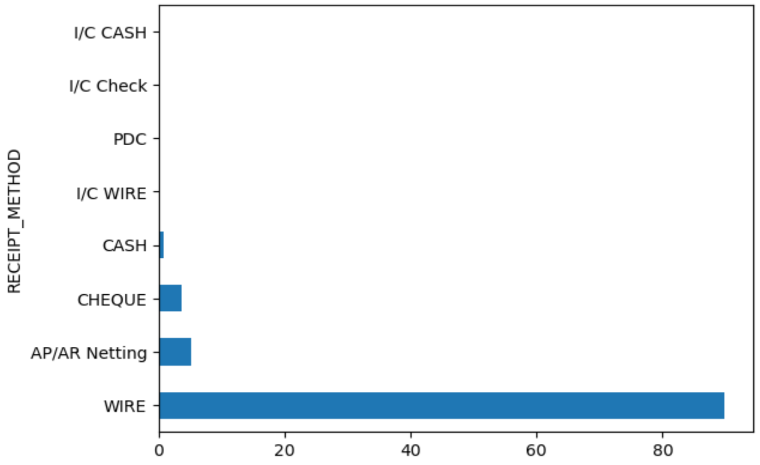
❑ Model Finalization and prediction on unforeseen data

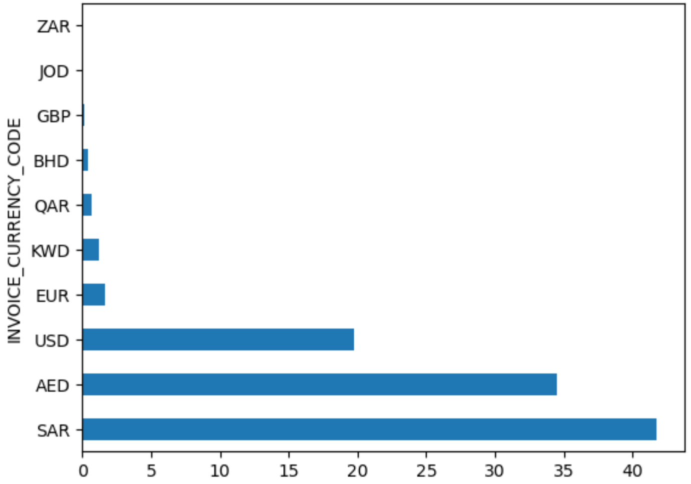
❑ Prediction summary, conclusion and recommendations

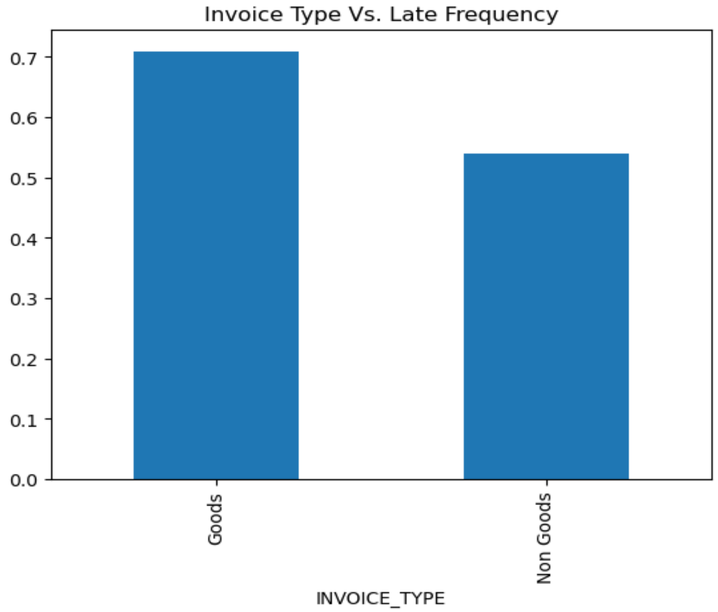
Transaction Insights (Univariate & Bivariate Analysis) :

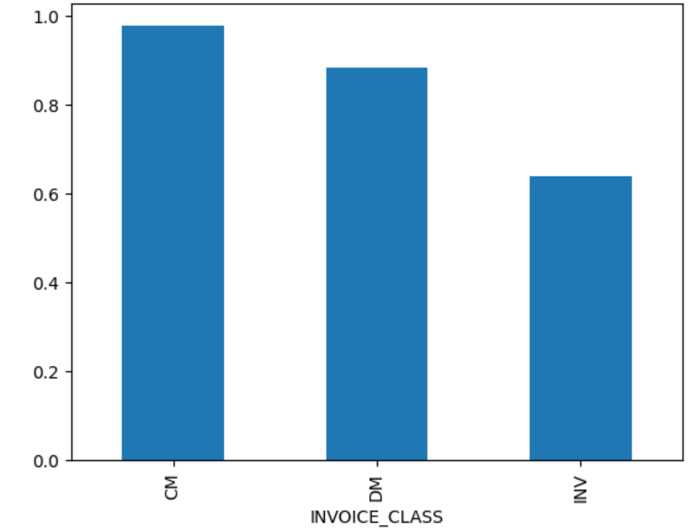






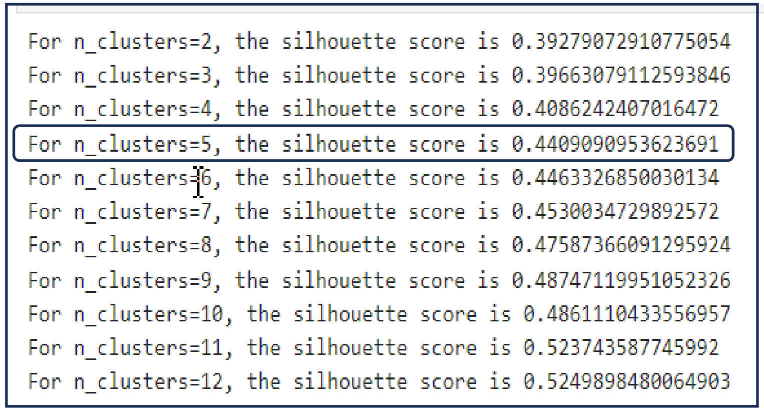






Customer segmentation using K-means clustering

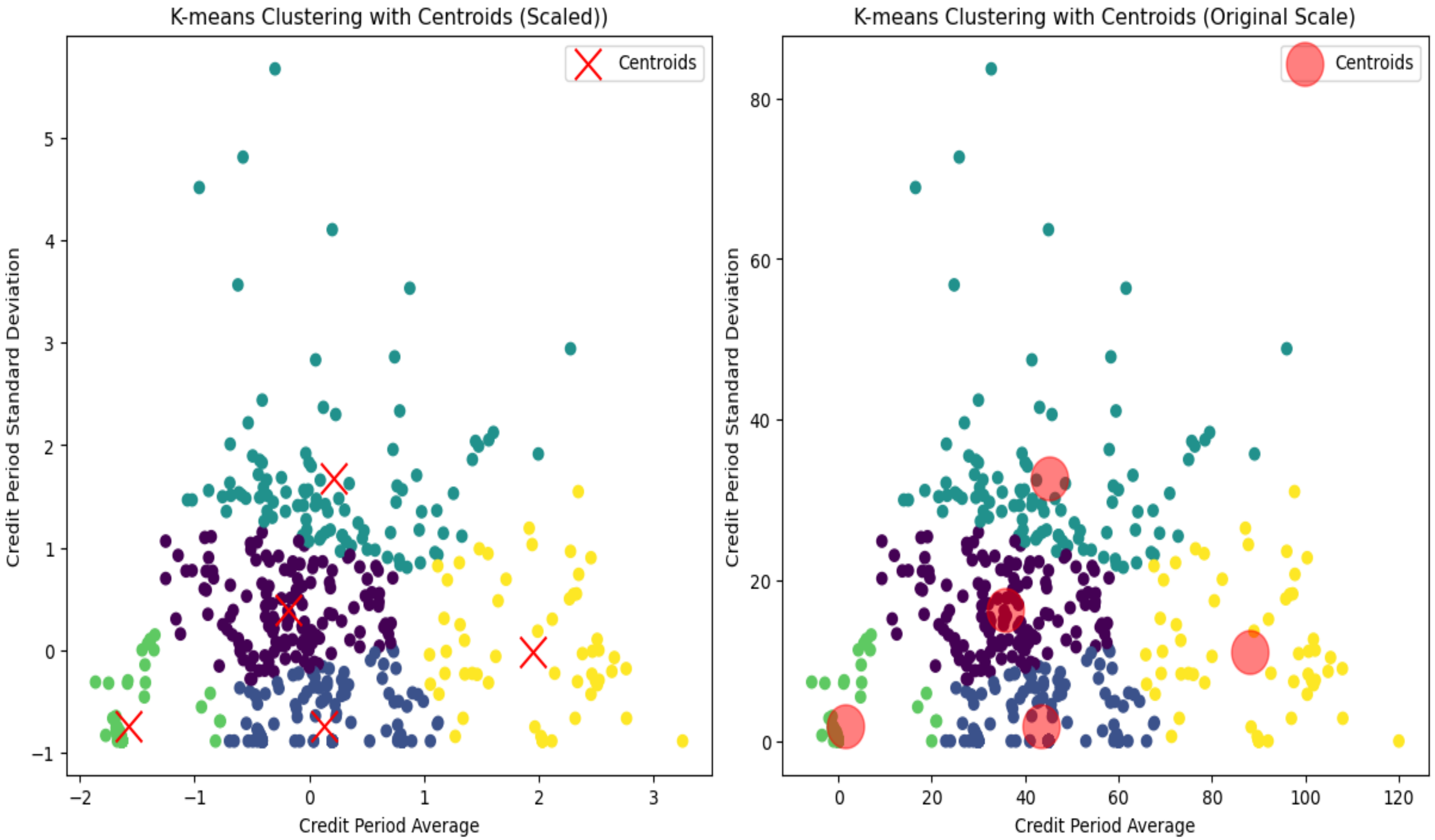
Clustering:



```
For n_clusters=2, the silhouette score is 0.39279072910775054
For n_clusters=3, the silhouette score is 0.39663079112593846
For n_clusters=4, the silhouette score is 0.4086242407016472
For n_clusters=5, the silhouette score is 0.4409090953623691
For n_clusters=6, the silhouette score is 0.4463326850030134
For n_clusters=7, the silhouette score is 0.4530034729892572
For n_clusters=8, the silhouette score is 0.47587366091295924
For n_clusters=9, the silhouette score is 0.48747119951052326
For n_clusters=10, the silhouette score is 0.4861110433556957
For n_clusters=11, the silhouette score is 0.523743587745992
For n_clusters=12, the silhouette score is 0.5249898480064903
```

We will go with k=5 as the score is close to 0.47. There are score greater than this as well for clusters 8 to 12, however, too many cluster will loose its importance.

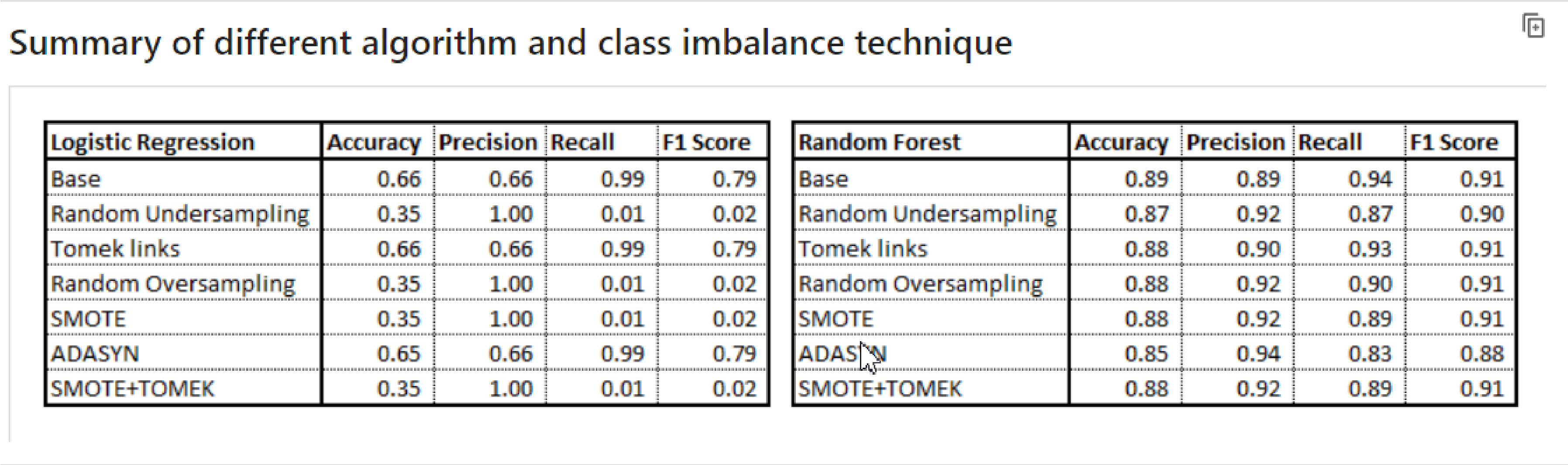
Plotting centroids:



Customer Segmentation:

- There are distinctly five clusters of customers, each with varying average payment days.
- On average, most customers are given payment terms ranging from 20 to 60 days, as observed in the blue, purple, and yellow clusters
- When the average credit period is under 20 days, there is very little variability in the credit terms, as seen in the peacock blue cluster
- highest variability is seen when credit period is between 20 to 60 days

Random Forest vs Logistics Regression :



Summary of different algorithm and class imbalance technique

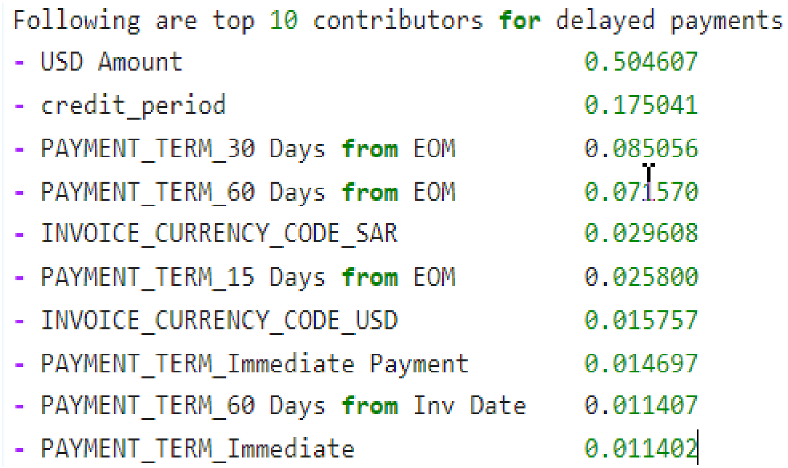
Logistic Regression	Accuracy	Precision	Recall	F1 Score
Base	0.66	0.66	0.99	0.79
Random Undersampling	0.35	1.00	0.01	0.02
Tomek links	0.66	0.66	0.99	0.79
Random Oversampling	0.35	1.00	0.01	0.02
SMOTE	0.35	1.00	0.01	0.02
ADASYN	0.65	0.66	0.99	0.79
SMOTE+TOMEK	0.35	1.00	0.01	0.02

Random Forest	Accuracy	Precision	Recall	F1 Score
Base	0.89	0.89	0.94	0.91
Random Undersampling	0.87	0.92	0.87	0.90
Tomek links	0.88	0.90	0.93	0.91
Random Oversampling	0.88	0.92	0.90	0.91
SMOTE	0.88	0.92	0.89	0.91
ADASYN	0.85	0.94	0.83	0.88
SMOTE+TOMEK	0.88	0.92	0.89	0.91

- Here, Base (without implementing class imbalance technique) and Tomek links gives best result among all other class imbalance techniques.
- However, Random Forest performs much better with high Accuracy, Precision, Recall and F1 Score. Logistic Regression has better Recall than Random Forest but much lower Accuracy and Precision.
- So, we will go with Random Forest without implementing any class imbalance technique.
- We have also seen earlier that the dataset in not highly skewed with 64% delayed payments and 36% not delayed.

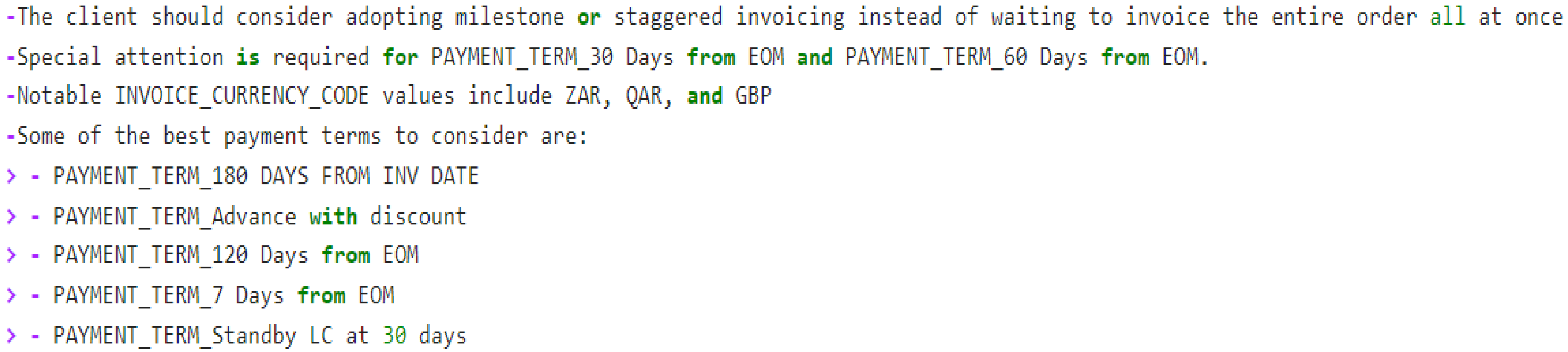
Conclusion & Recommendations:

Contributors for Delayed Payments:



```
Following are top 10 contributors for delayed payments
- USD Amount                          0.504607
- credit_period                       0.175041
- PAYMENT_TERM_30 Days from EOM       0.085056
- PAYMENT_TERM_60 Days from EOM       0.071570
- INVOICE_CURRENCY_CODE_SAR           0.029608
- PAYMENT_TERM_15 Days from EOM       0.025800
- INVOICE_CURRENCY_CODE_USD           0.015757
- PAYMENT_TERM_Immediate Payment      0.014697
- PAYMENT_TERM_60 Days from Inv Date  0.011407
- PAYMENT_TERM_Immediate              0.011402
```

Recommendation:



```
-The client should consider adopting milestone or staggered invoicing instead of waiting to invoice the entire order all at once
-Special attention is required for PAYMENT_TERM_30 Days from EOM and PAYMENT_TERM_60 Days from EOM.
-Notable INVOICE_CURRENCY_CODE values include ZAR, QAR, and GBP
-Some of the best payment terms to consider are:
> - PAYMENT_TERM_180 DAYS FROM INV DATE
> - PAYMENT_TERM_Advance with discount
> - PAYMENT_TERM_120 Days from EOM
> - PAYMENT_TERM_7 Days from EOM
> - PAYMENT_TERM_Standby LC at 30 days
```