E-commerce & Retail Case Study

By
Abhishek Singh Chauhan,
Tara Thapa Magar,
Yadeesh KR,
Kamran Masood.

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 behavior and predict the likelihood of late payments against open invoices.

Objectiv
e

- 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.

Solution Methodology

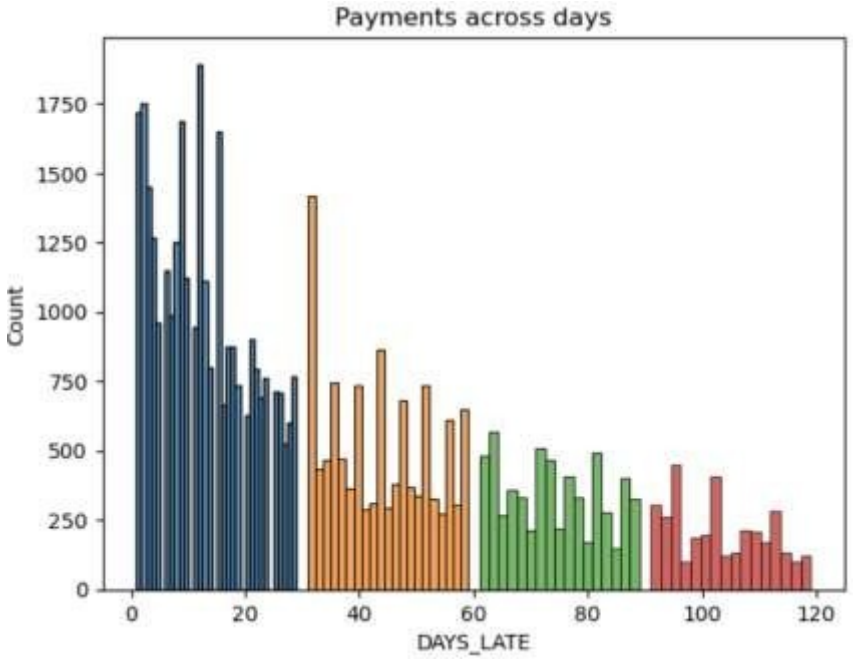
1. Reading and Understanding the data
2. Data Cleaning
   - Delete null values
   - Dropping columns which contains only one value
   - Dropping duplicated columns
   - Dropping columns which are not important for the analysis
3. Exploratory Data Analysis
   - Data imbalance check
   - Creating derived metrics (Ex: overdue_days, credit_period)
4. Clustering
5. Data Preparation
   - Outlier treatment
   - Creating dummy variables Feature scaling
   - Train Test split
6. Model Building
7. Model Evaluation
8. Conclusion

Grouped data

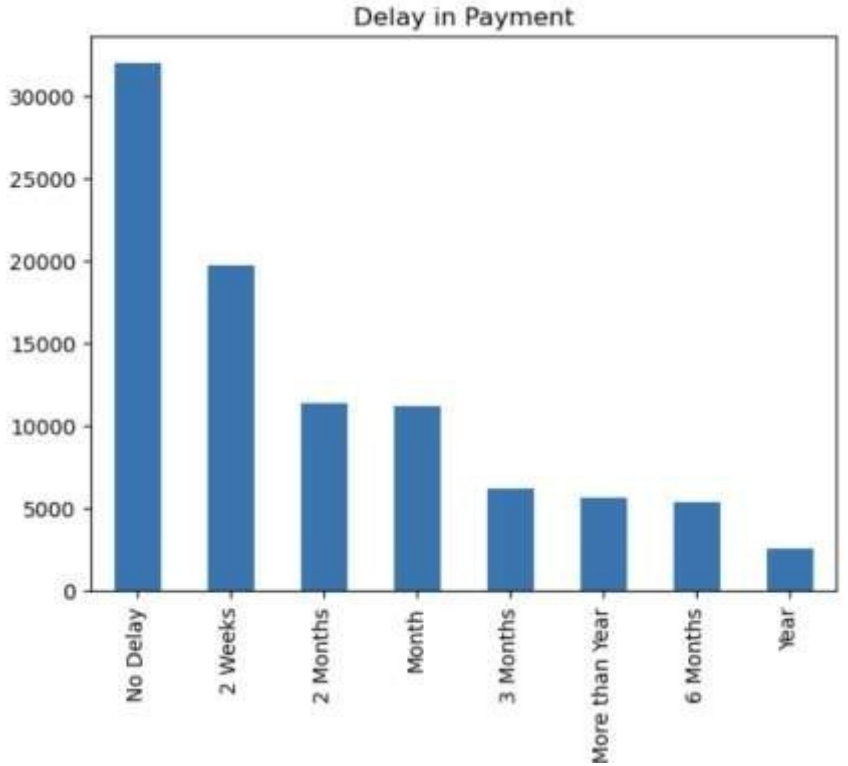
- Grouped customers based on mean credit period and standard deviation of credit
  period
- Mean credit period is following a normal distribution
- Standard deviation of credit period is skewed towards left
- The average number of days from invoice date to due date is 38

EDA

- Distribution of payments across due days

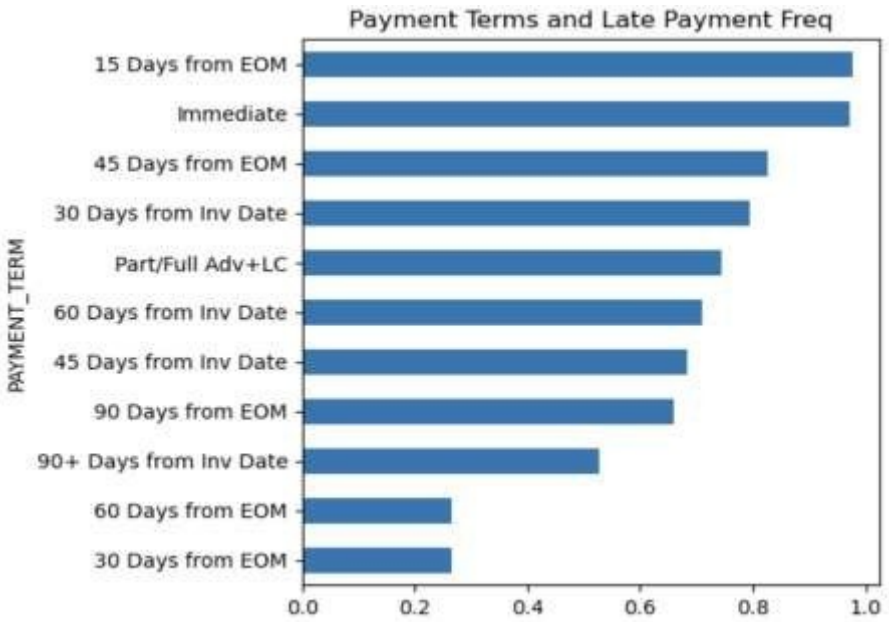


- Delay in payment
- Most of payments are cleared with no delay
- 2 weeks delayed payments are above the average

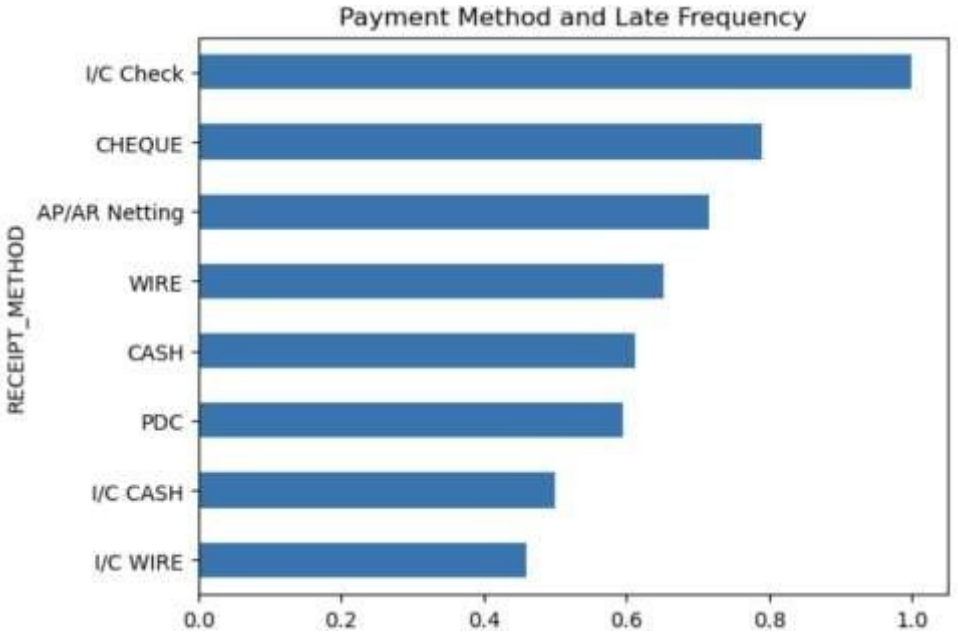


EDA

- Late payment frequency by Payment term

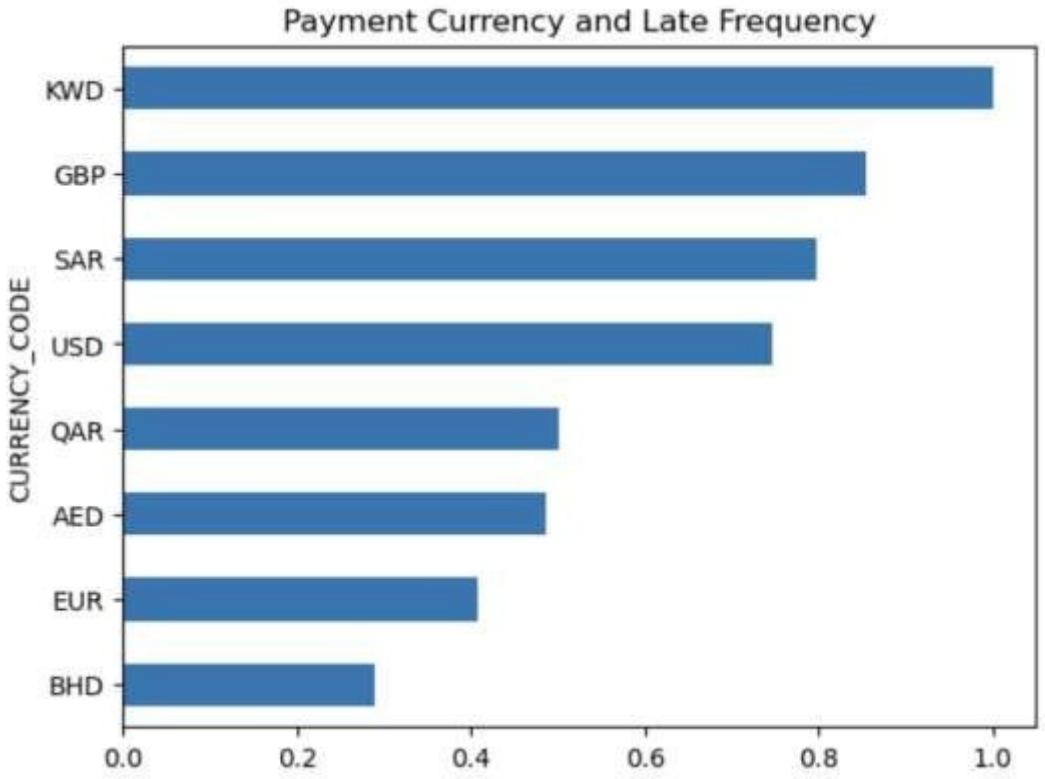


- Late payment frequency by Payment method

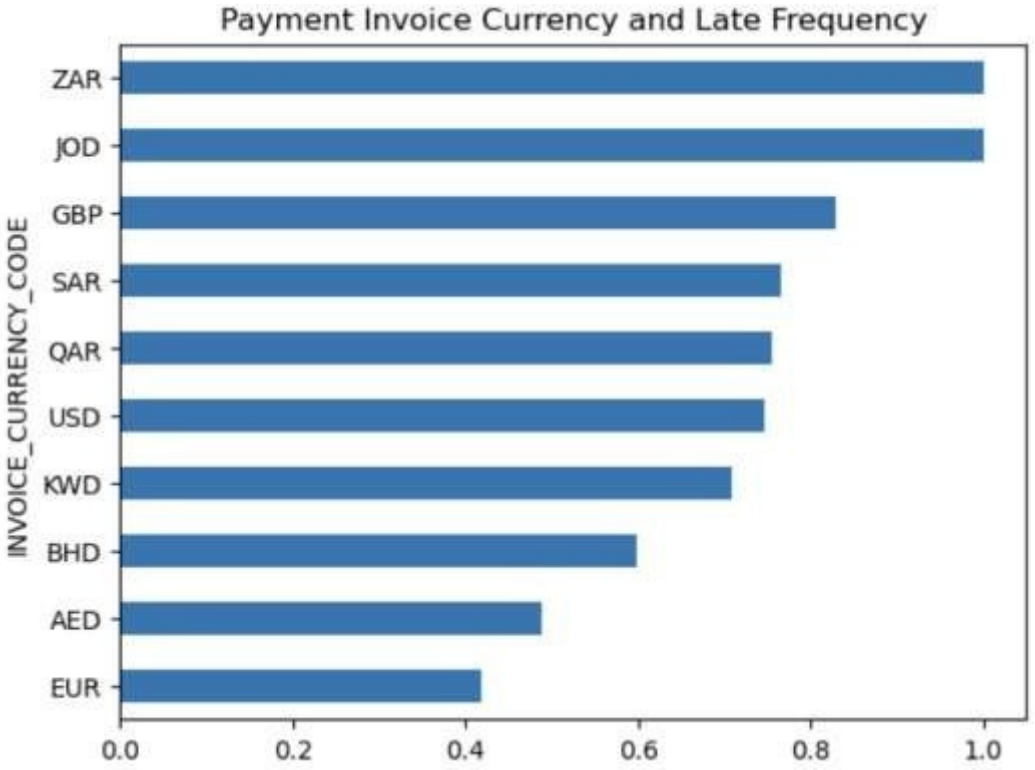


EDA

- Late payment frequency by Payment currency

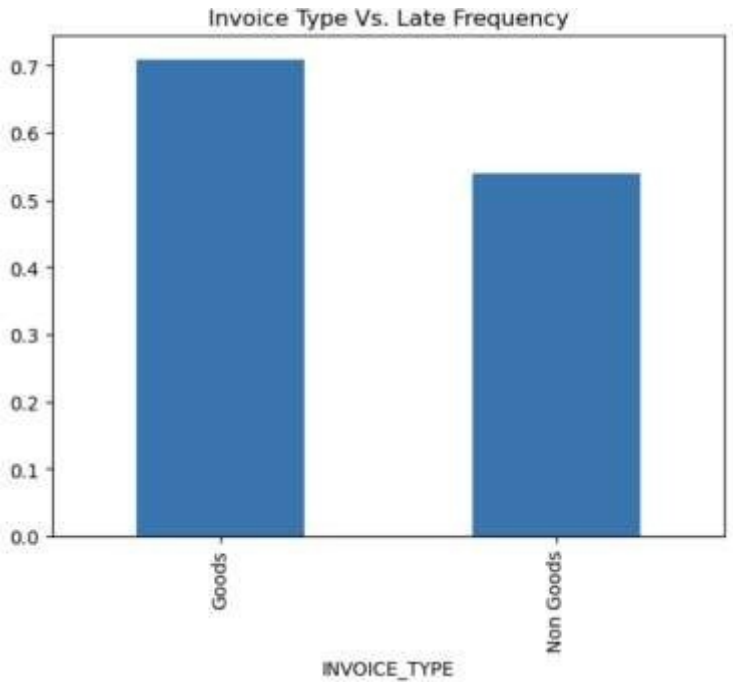


- Late payment frequency by Payment invoice currency

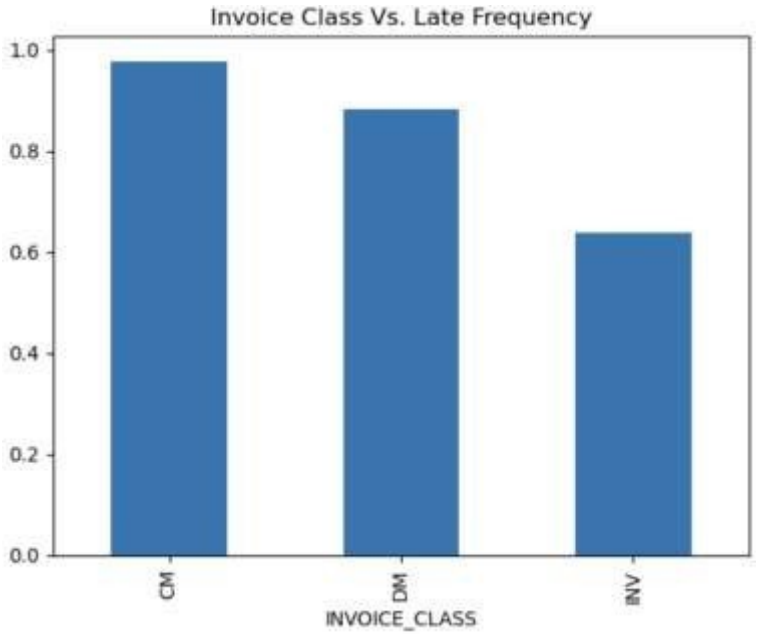


EDA

- Late payment frequency by Invoice type

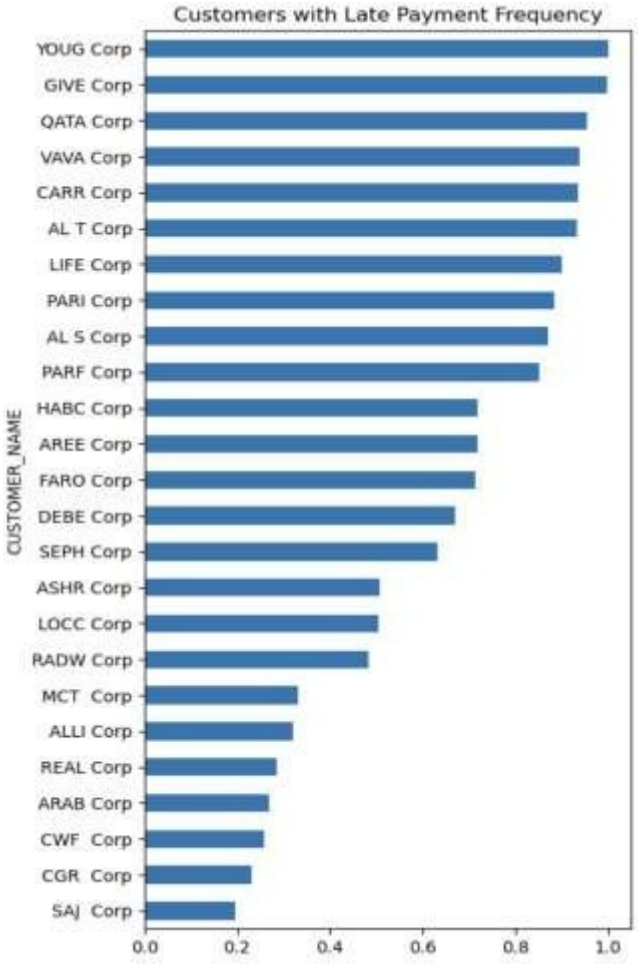


- Late payment frequency by Invoice class

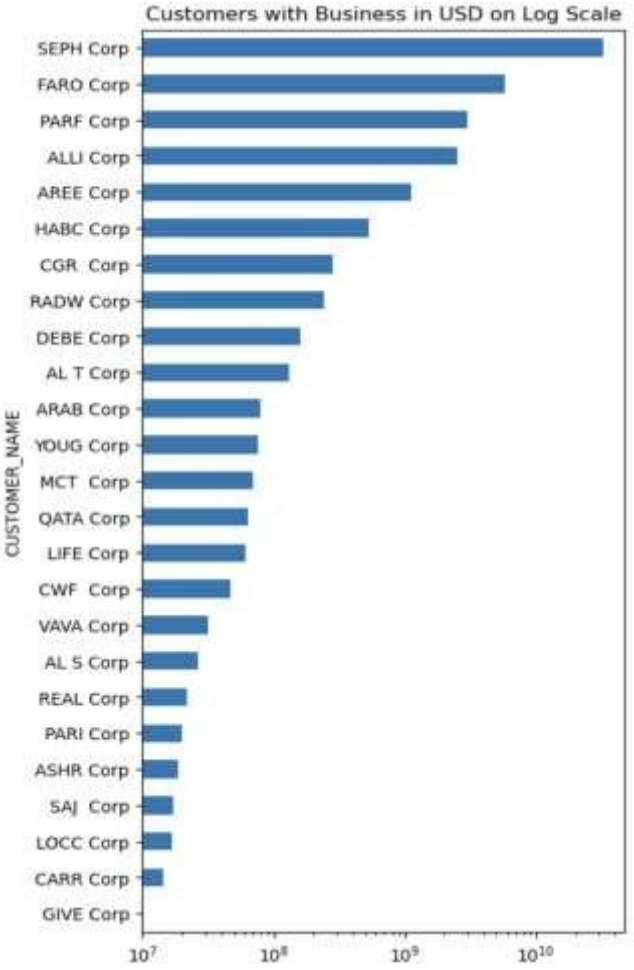


EDA

- Late payment frequency by Customers



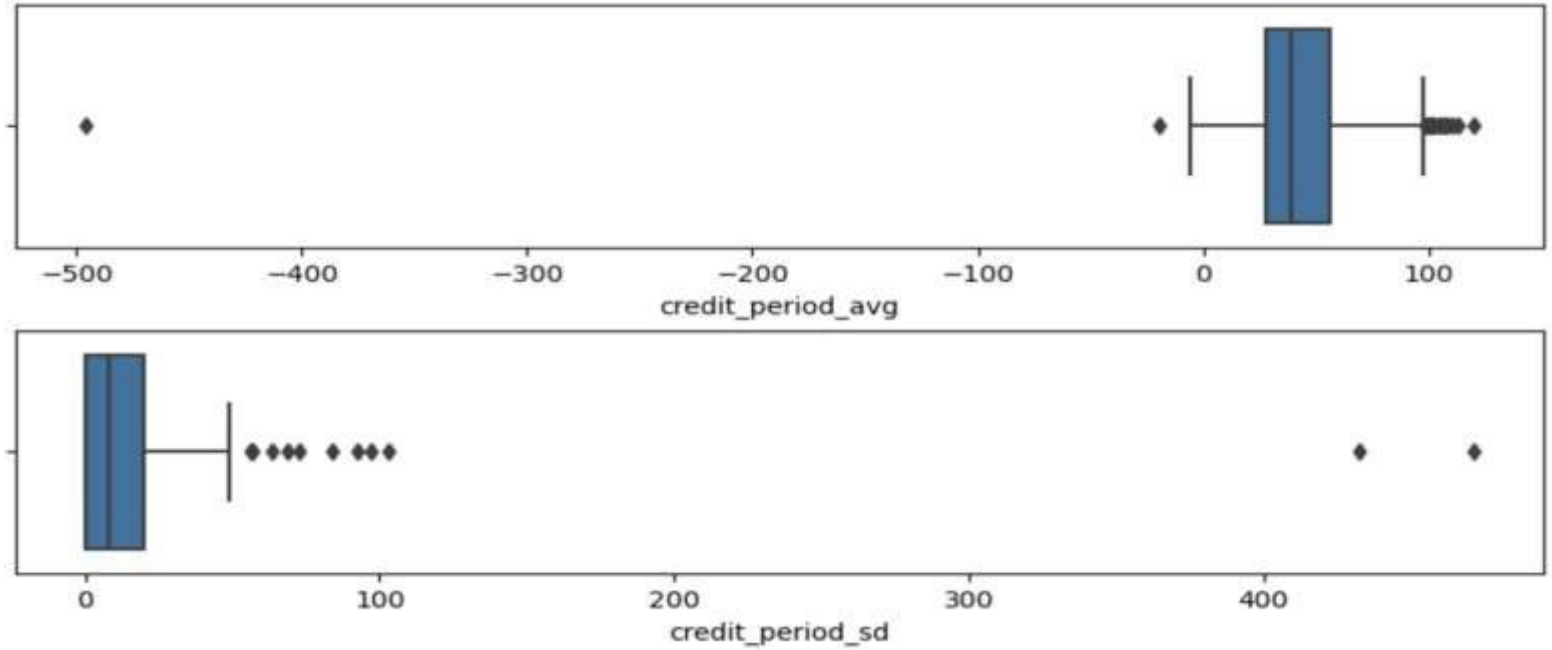
- Customer count with Business in USD (log scaled)



Outlier treatment and scaling

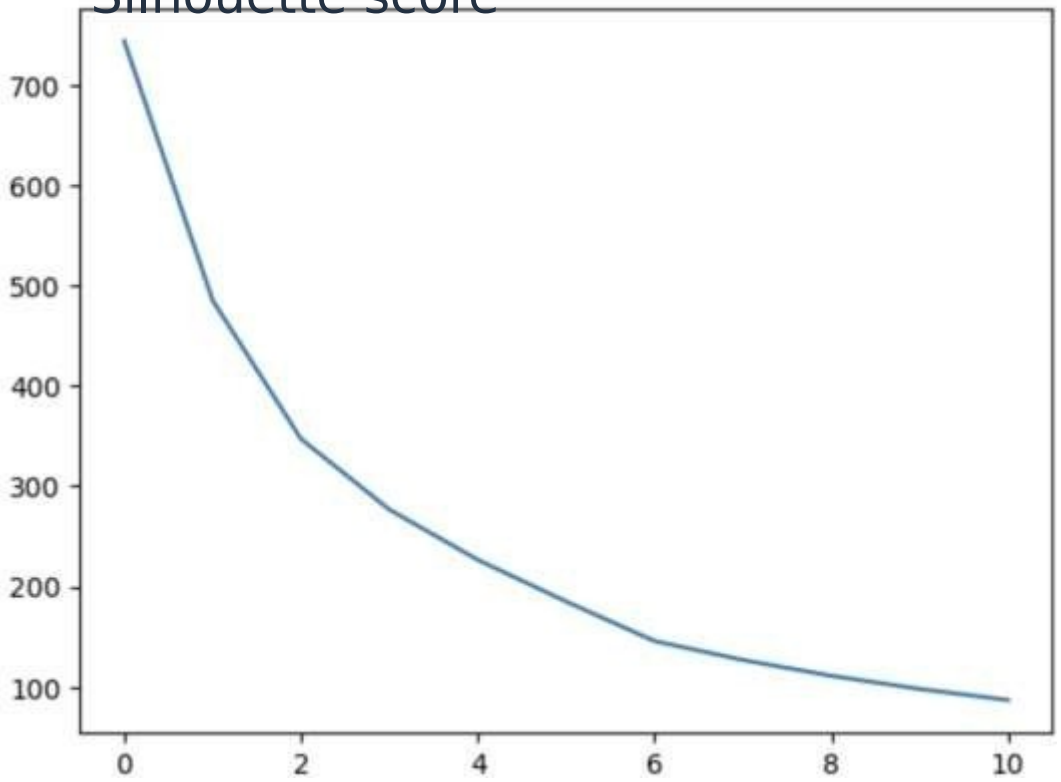
Used IQR(Inter Quartile Range) method to remove outliers and scaled the data using

Standard scalar method to do clustering



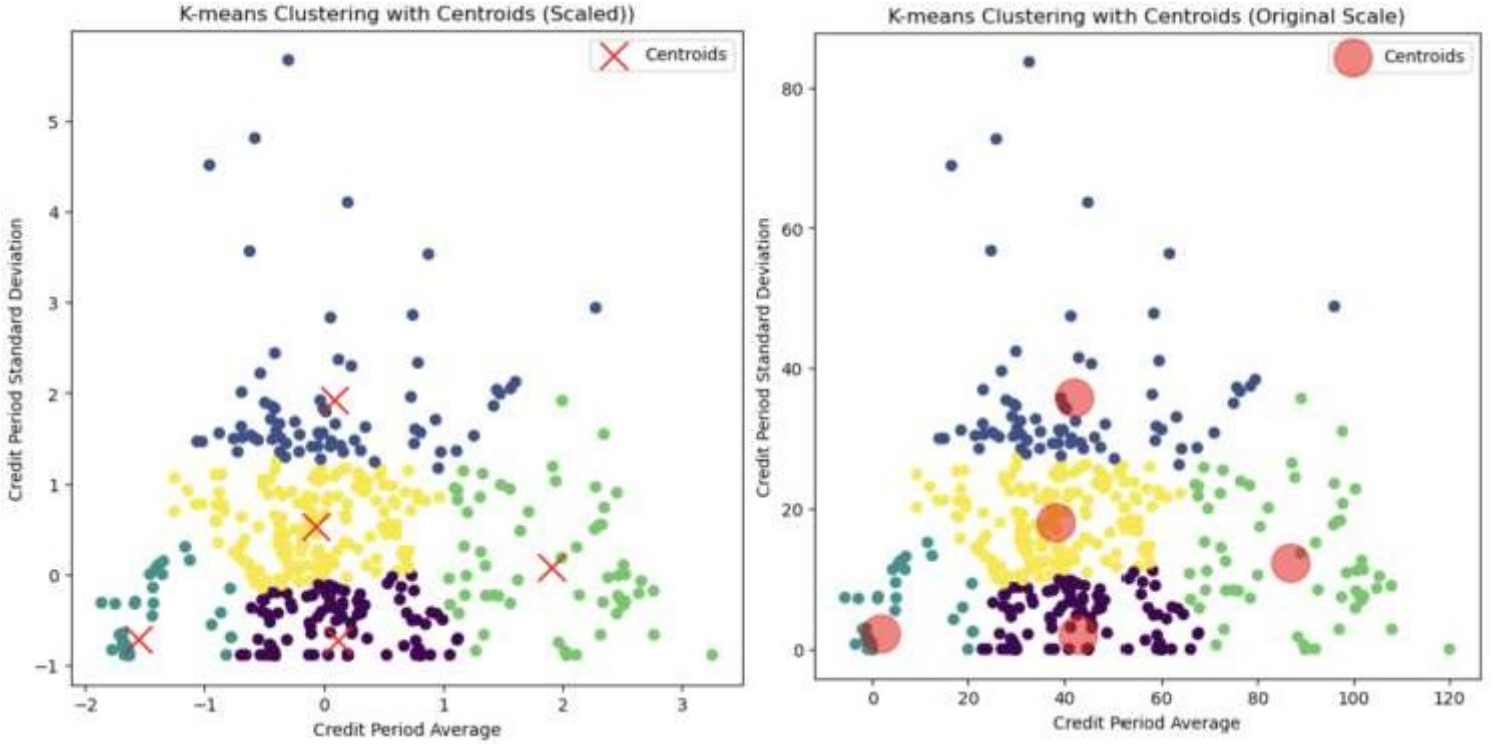
Customer segmentation

- K-Means clustering on scaled data.
- Elbow curve and Silhouette score to determine optimal cluster
- Too many cluster will loose its importance so choosing k=5 by analyzing Silhouette score



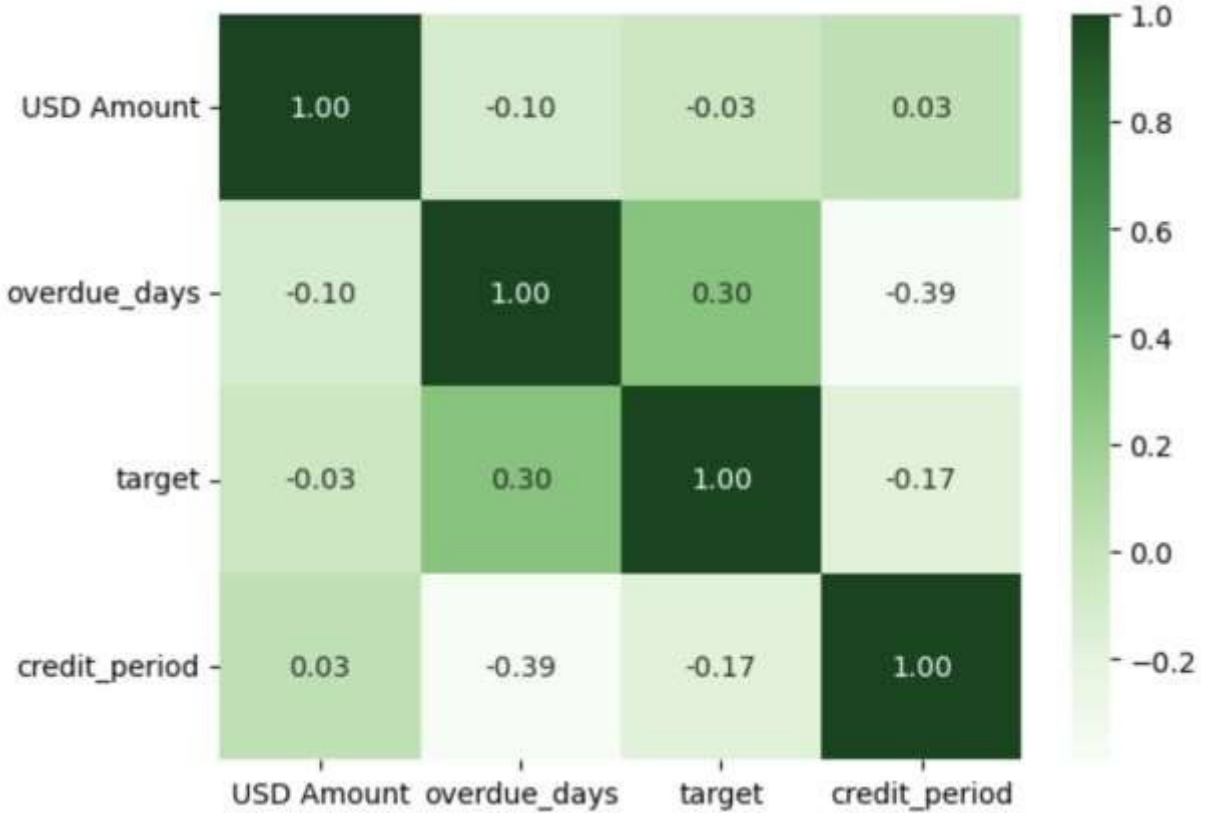

```
For n_clusters=2, the silhouette score is 0.39279072910766427
For n_clusters=3, the silhouette score is 0.3964126949475593
For n_clusters=4, the silhouette score is 0.4631100442617391
For n_clusters=5, the silhouette score is 0.43221289554472897
For n_clusters=6, the silhouette score is 0.45059789940393674
For n_clusters=7, the silhouette score is 0.4491402344467193
For n_clusters=8, the silhouette score is 0.4776135189927352
For n_clusters=9, the silhouette score is 0.49031599008046334
For n_clusters=10, the silhouette score is 0.4982584687687927
For n_clusters=11, the silhouette score is 0.49894600643982734
For n_clusters=12, the silhouette score is 0.5246978102917228
```

Clustering



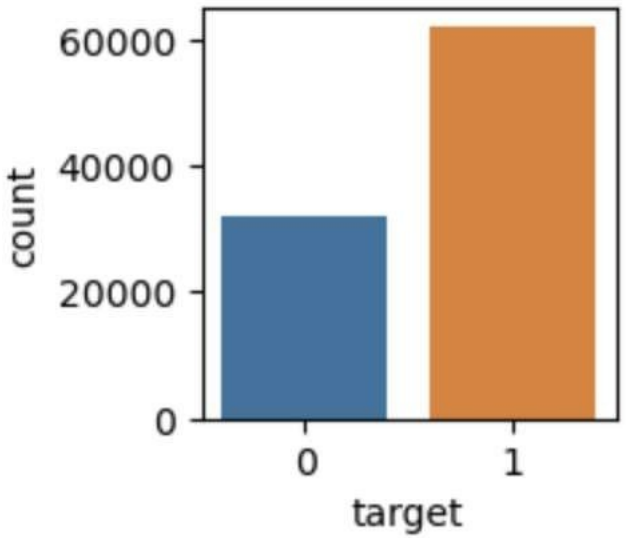
- There are clear 5 clusters of customers having different average payment days
  - Most of the customers are offered between 20 to 60 days of payment terms on an average.(Blue, purple and yellow cluster)
- When credit period is under 20 days on average, the variability in credit period is very less. (peacock blue cluster)
- When credit period is more than 60 days on average, there is relatively moderage variability in the offered credit period (green cluster)
- The variability is highest when credit period is between 20 to 60 days

Correlation matrix



Class Imbalance and Data Preparation

- Class imbalance by checking % of delayed and not delayed from target column
- The dataset is moderately imbalanced with approx. 66% delayed and 34% not delayed



Feature selection

- Selected top features which are >0.02, other columns are dropped as it does not
  contribute much
- Features and its importance

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

Class imbalance and Model selection

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.

Random Forest model and Hyperparameter tuning

- Hyperparameter tuning is achieved using GridsearchCV method
- Model performance on the training data
- Training accuracy- 84.7% Validation accuracy- 85% This clearly indicates the model is not overfitting

```
RandomForestClassifier(max_depth=25, max_features=6, min_samples_leaf=20,
                       n_estimators=50, n_jobs=-1, random_state=42)
```

clasification report:

	precision	recall	f1-score	support
0	0.83	0.70	0.76	9588
1	0.86	0.93	0.89	18594
accuracy			0.85	28182
macro avg	0.85	0.81	0.82	28182
weighted avg	0.85	0.85	0.85	28182

- **Confusion matrix**

	Positive	Negative
Positive	6665	2923
Negative	1319	17275

- Model is giving very high accuracy, precision, recall and f1-score.
- Out of these, recall is very high, which shows that model predicts very high proportion of delayed payments

Conclusion

- The analysis identifies the top 10 contributors to delayed payments, with the highest impact factors being USD Amount, credit_period, and specific payment terms such as PAYMENT_TERM_30 Days from EOM and PAYMENT_TERM_60 Days from EOM. This suggests that addressing these factors could significantly reduce payment delays.

FEATURES	IMPORTANCE
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

- To improve the payment process, it is recommended that the client consider adopting milestone or
  staggered invoicing strategies rather than waiting to invoice for the entire order at once.
- 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
- Additionally, it is important to exercise caution with payment terms such as PAYMENT_TERM_30 Days from EOM and PAYMENT_TERM_60 Days from EOM, as these have been identified as contributors to delayed payments.
- Regarding currency codes, it is advisable to prioritize INVOICE_CURRENCY_CODE like ZAR, QAR, and GBP, while being cautious when dealing with INVOICE_CURRENCY_CODE SAR and USD, as they can impact payment timeliness

Thank You