E-COMMERCE & RETAIL B2B CASE STUDY

Submitted BY: Pavithra S
YEAR : 2024

Objectives

GAIN DEEPER INSIGHTS INTO CUSTOMERS' PAYMENT BEHAVIORS BY ANALYZING PAST PAYMENT PATTERNS AND SEGMENTING CUSTOMERS.

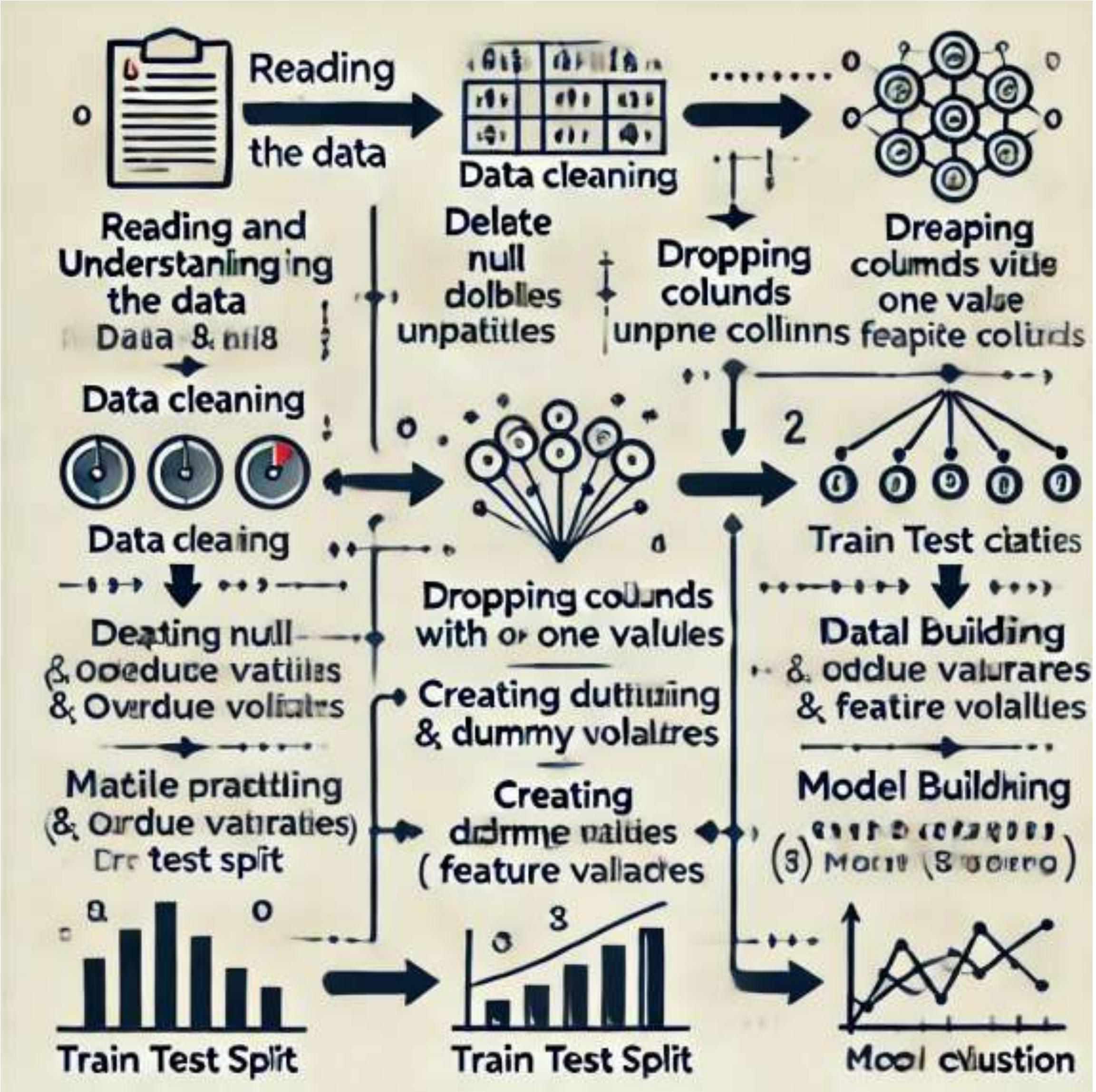
LEVERAGE HISTORICAL DATA TO PREDICT THE PROBABILITY OF DELAYED PAYMENTS FOR OUTSTANDING INVOICES.

USE PREDICTIVE INSIGHTS TO HELP COLLECTORS PRIORITIZE THEIR EFFORTS AND ENABLE PROACTIVE FOLLOW-UPS TO ENSURE TIMELY PAYMENTS.

Problem Statement

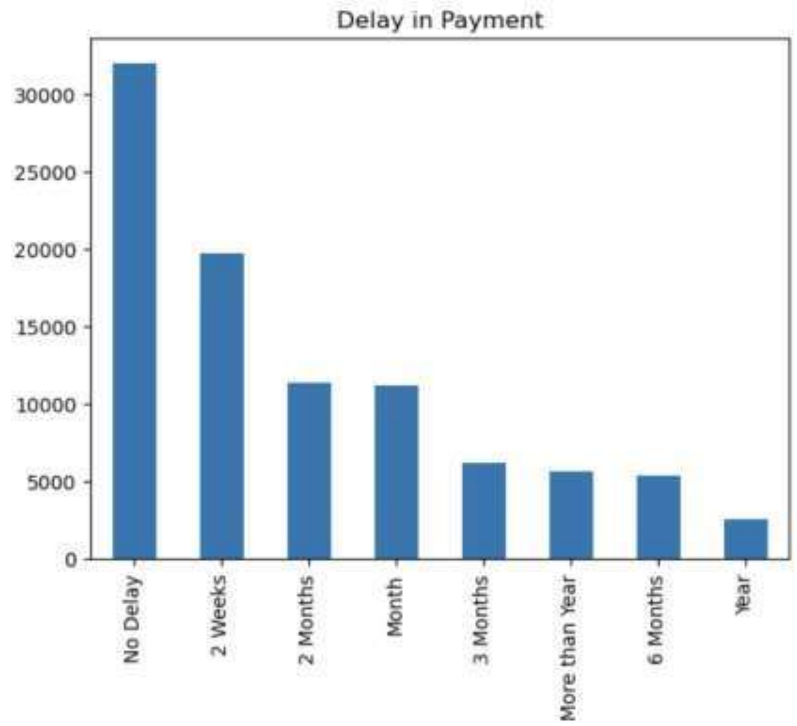
Schuster is a global retail company specializing in sports goods and accessories. It has extensive business relationships with hundreds of vendors, many of whom have credit agreements with the company. However, not all vendors adhere to the agreed-upon credit terms, and some consistently make late payments. While Schuster imposes substantial late payment fees, this approach does not foster long-term, mutually beneficial relationships. The company also employs staff who regularly follow up with vendors to ensure timely payments, but this results in time-consuming, non-value-added activities and financial costs. To address this, Schuster aims to better understand its vendors' payment behavior and predict the likelihood of late payments on open invoices.

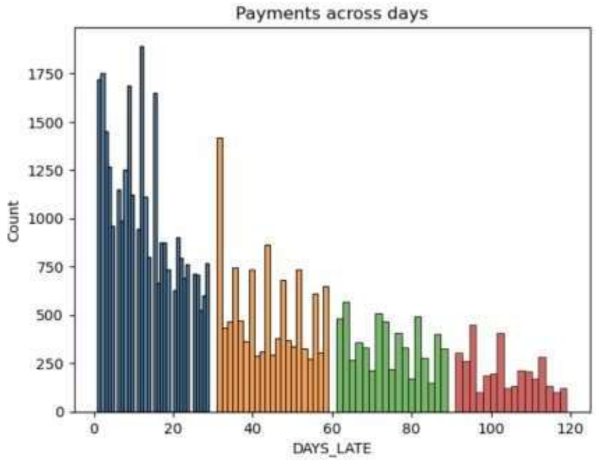
Approach to the Problem



EDA

- **Distribution of Payments Across Due Days:**Most payments are cleared on time.
- Payments delayed by two weeks are above the average delay rate.

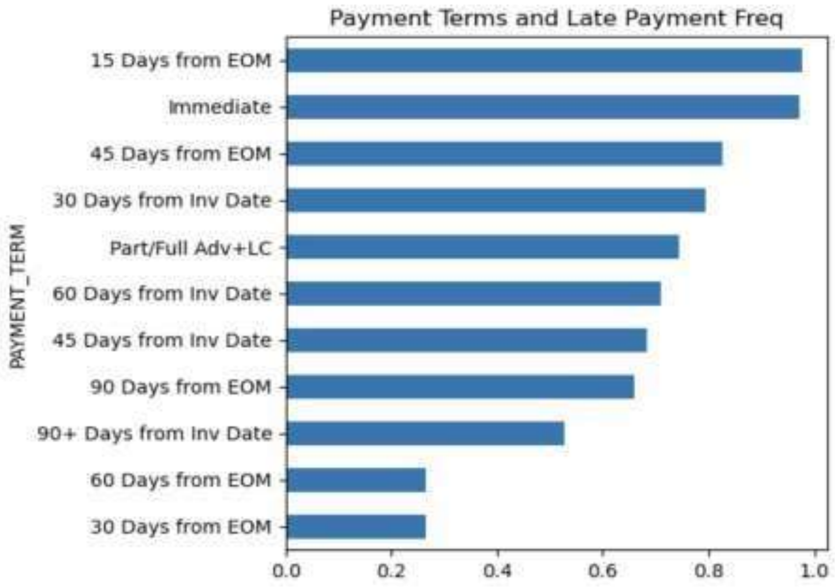


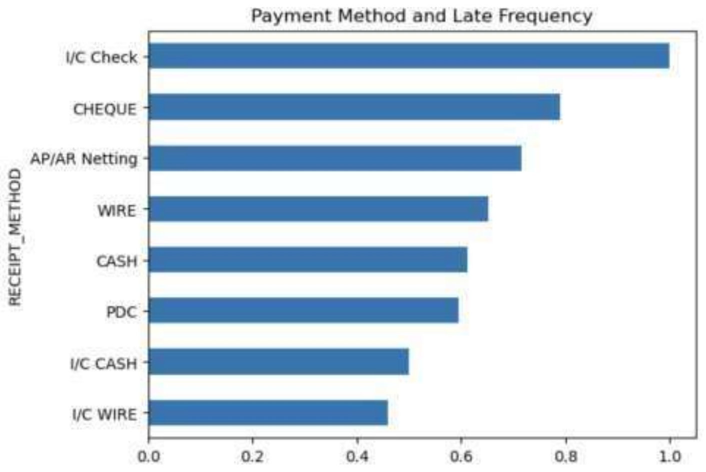


EDA

- Late payment frequency by Payment
- Method

- Late payment frequency by Payment
- term

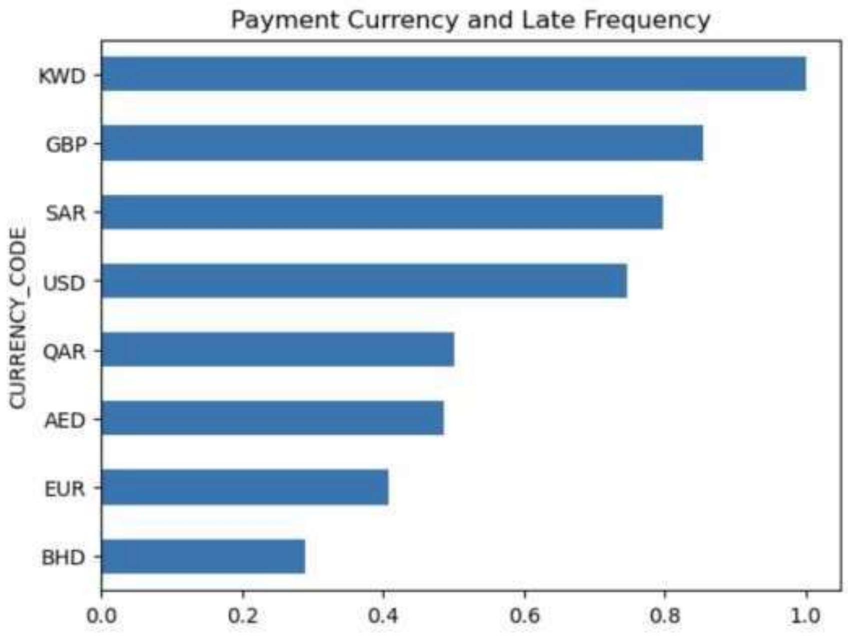


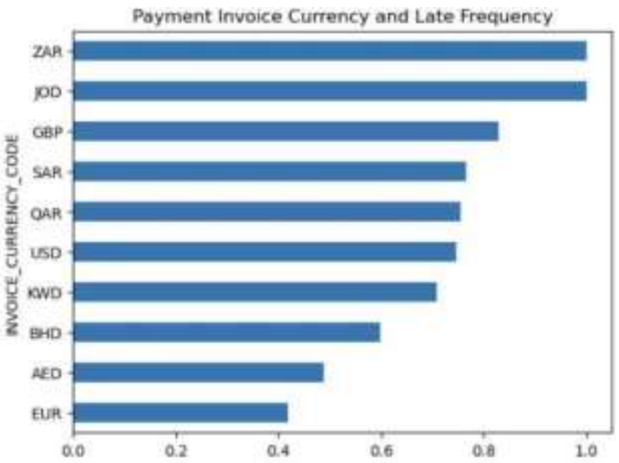


EDA

Late payment frequency by Payment Currency

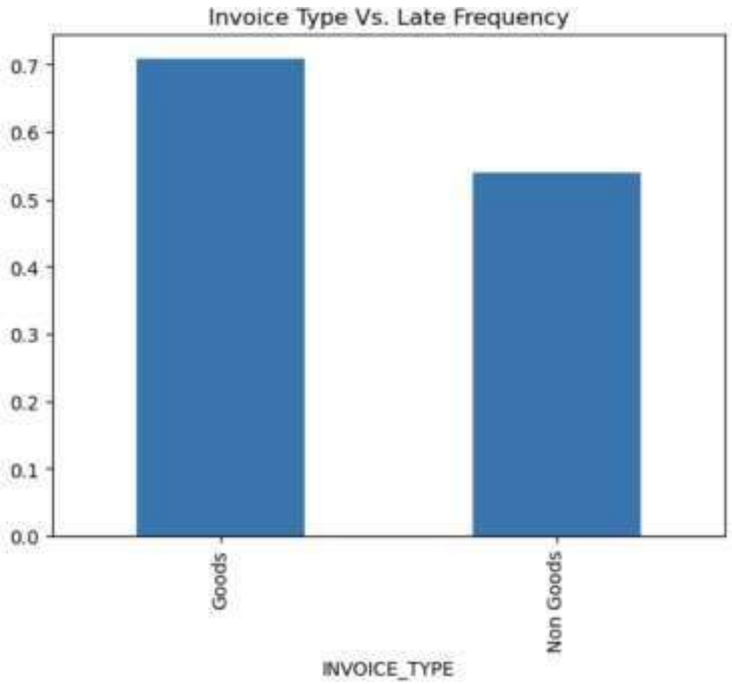
Late payment frequency by Payment invoice currency



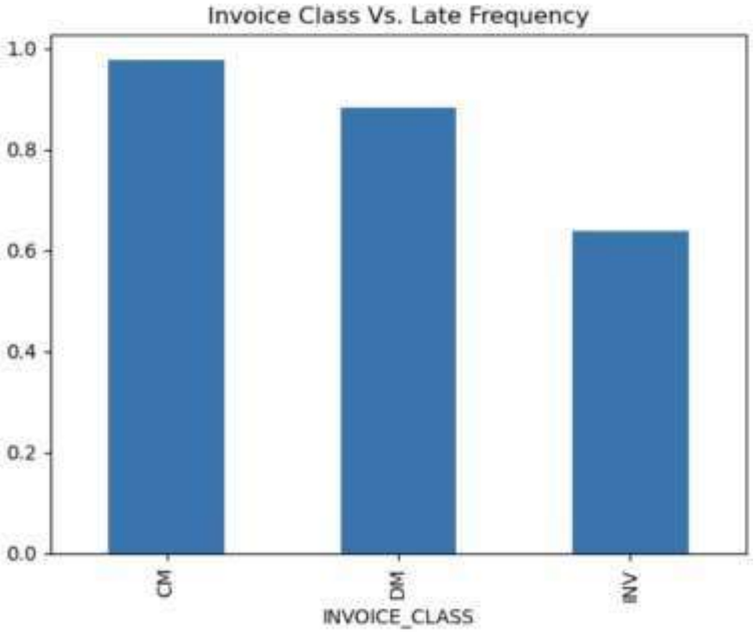


EDA

- Late payment frequency by Invoice type

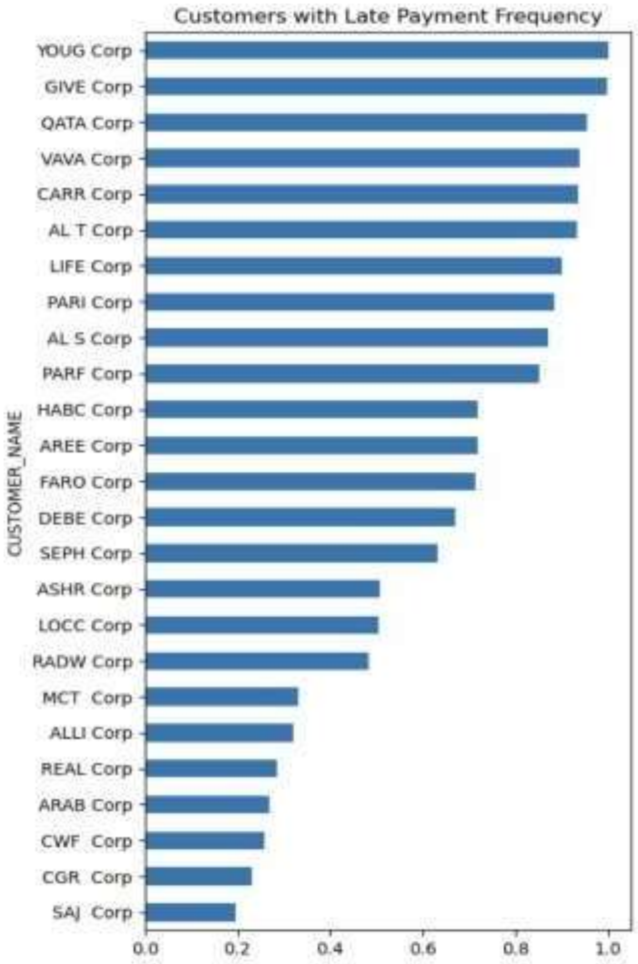


- Late payment frequency by Invoice class

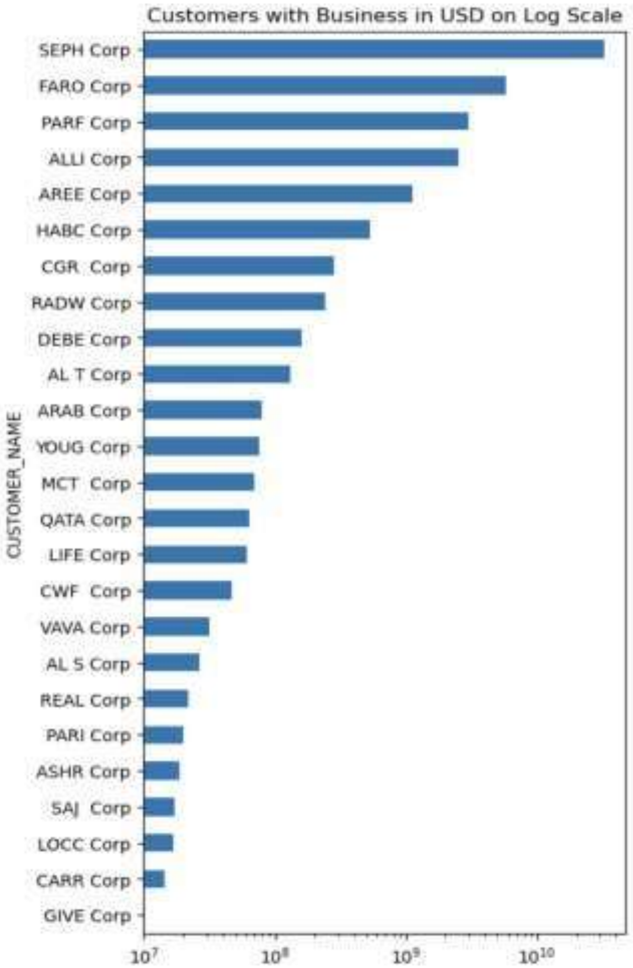


EDA

- Late payment frequency by Customers

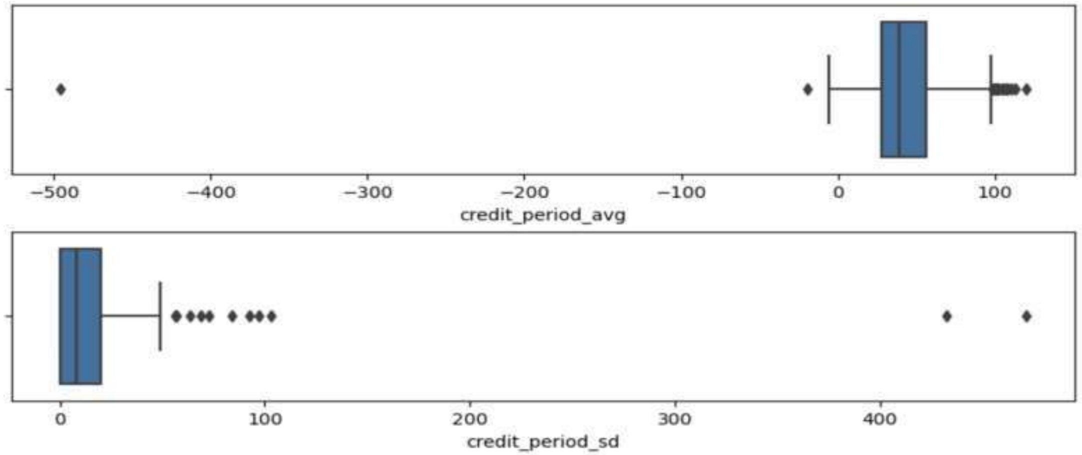


- Customer count with Business in USD (log scaled)



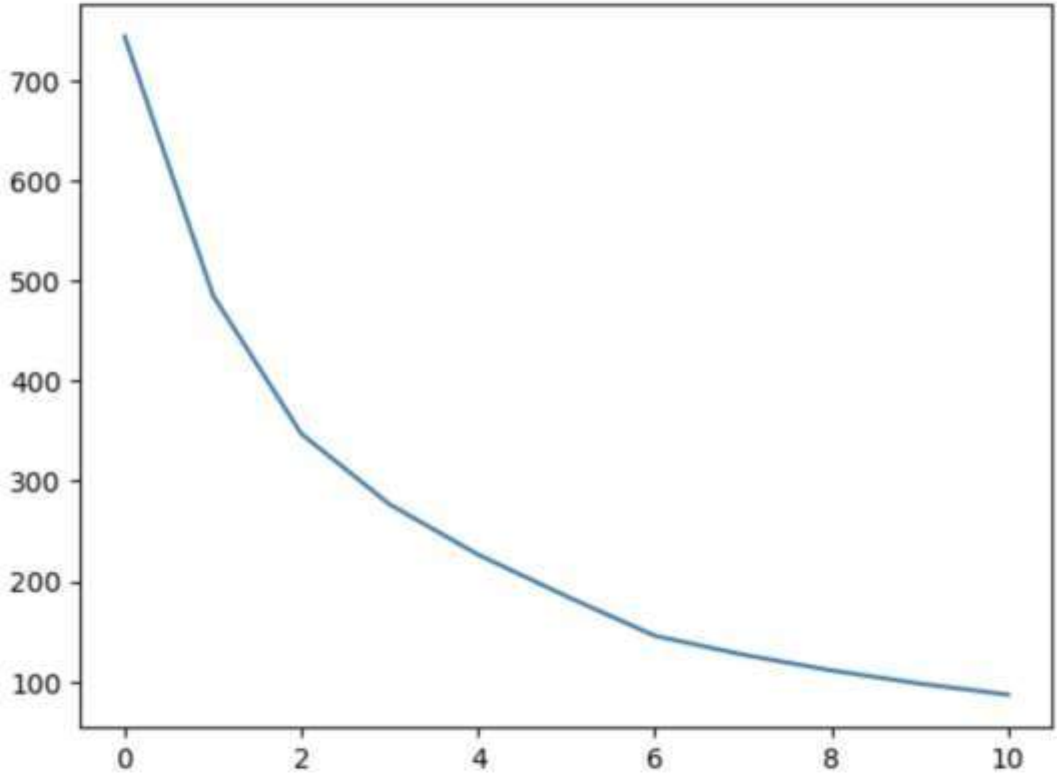
Outlier treatment and scaling

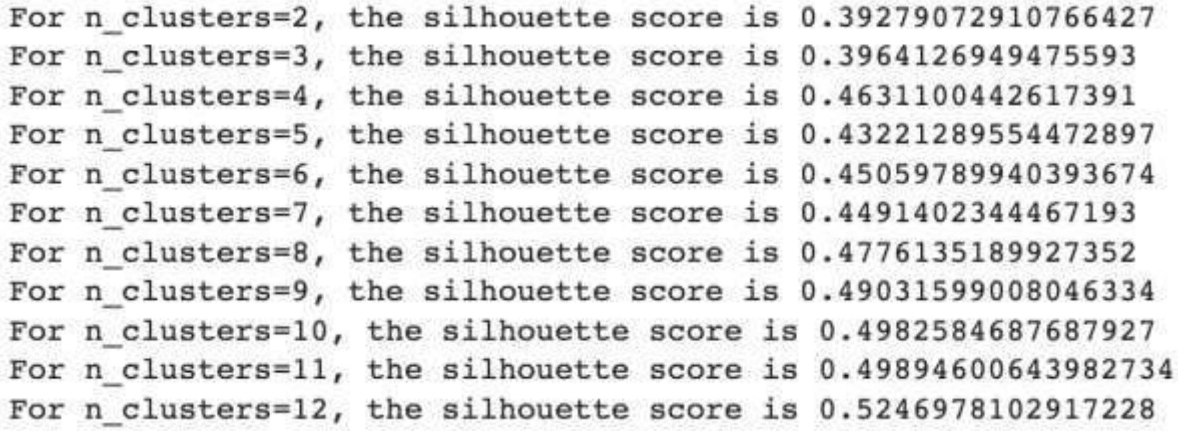
- Used IQR(Inter Quartile Range) method to remove outliers and scaled the data using
- Standard scalar method to do clustering



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

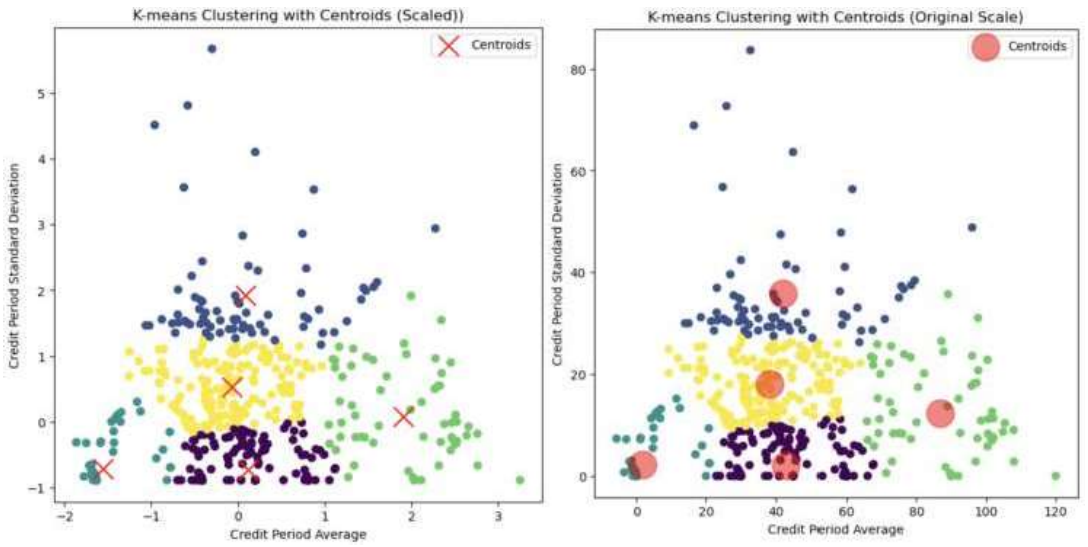
- There are clear 5 clusters of customers having different average payment days

- Most of the customers are offered between
- 20 to 60 days of payment terms on an
- average.(Blue, purple and yellow cluster)

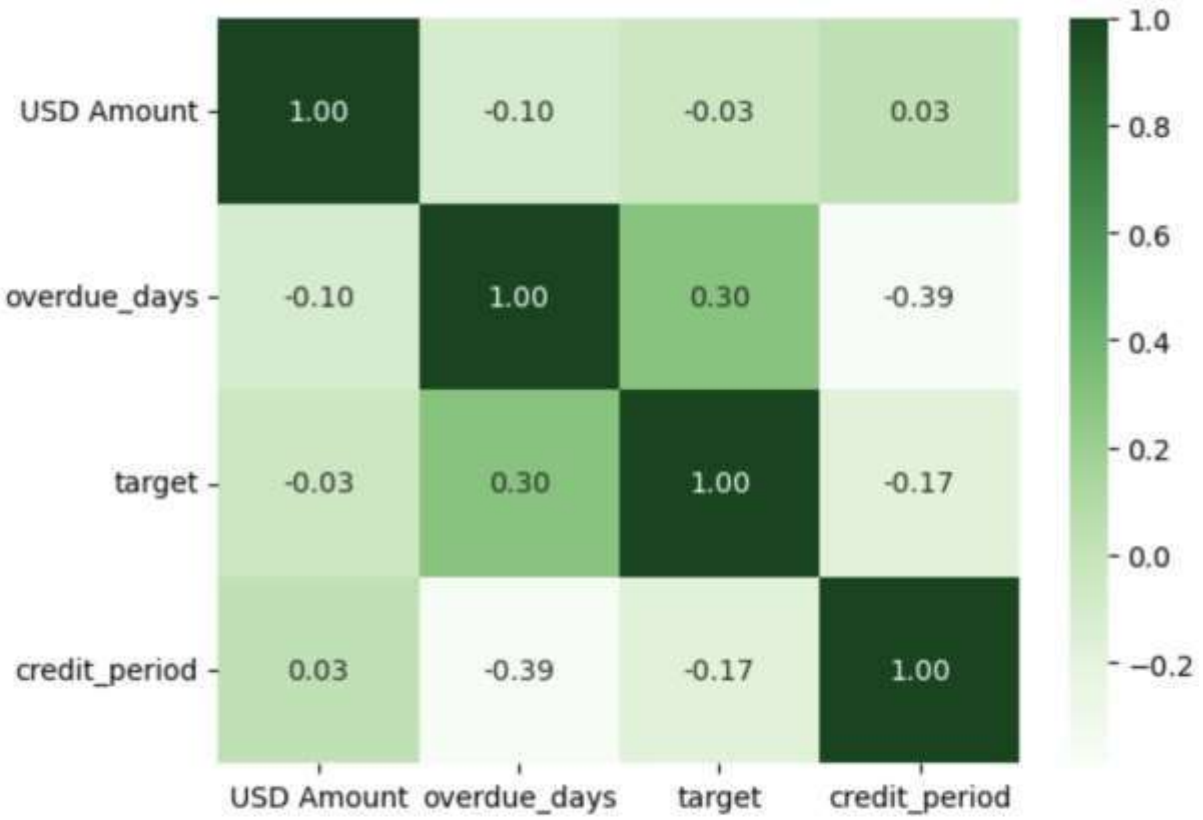
- When credit period is under 20 days on average, the variability in credit period is very less. (peacock blue cluster)

- 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

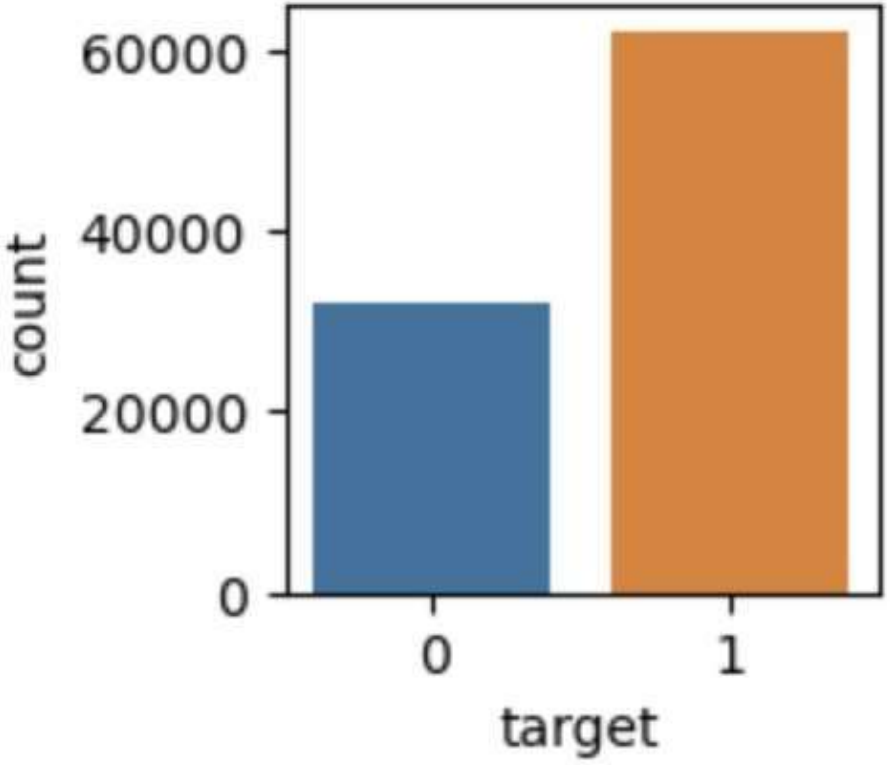


Correlation matrix



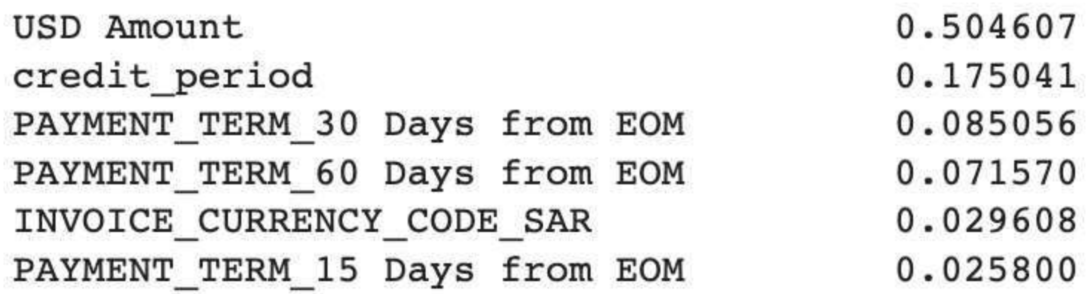
Class Imbalance and Data Preparation

- 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

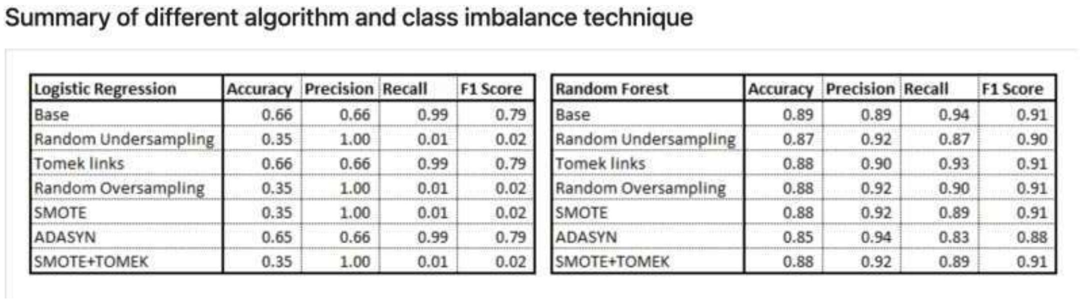
- Selected top features which are >0.02, other columns are dropped as it does not
- contribute much
- Features and its importance



```
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

- **Base (without class imbalance techniques)** and **Tomek Links** give the best results among all class imbalance techniques.
- **Random Forest** performs better with higher Accuracy, Precision, Recall, and F1 Score.
- **Logistic Regression** has better Recall than Random Forest but significantly lower Accuracy and Precision.
- We will proceed with **Random Forest** without implementing any class imbalance technique.
- The dataset is not highly skewed, with **64% delayed payments** and **36% non-delayed payments**.



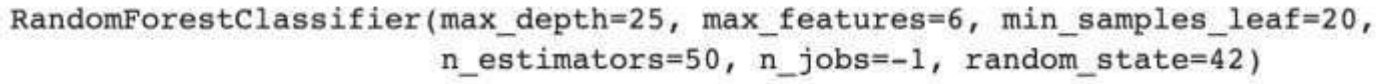
Summary of different algorithm and class imbalance technique

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

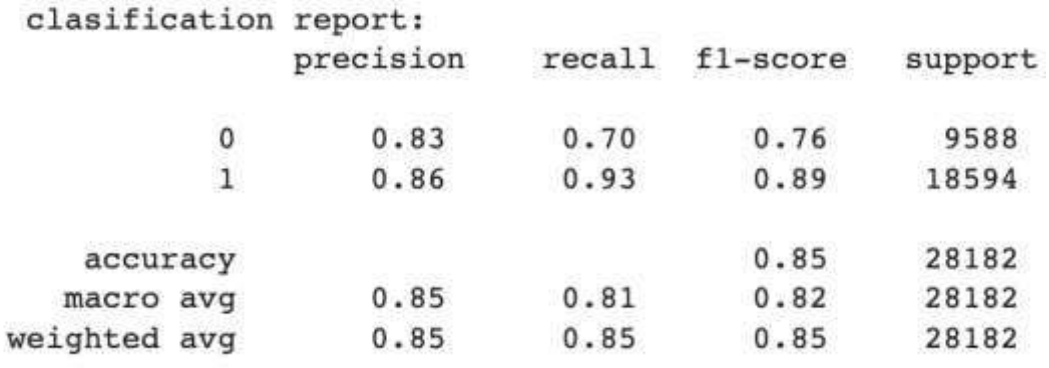
Random Forest	Accuracy	Precision	Recall	F1 Score
Base	0.89	0.89	0.94	0.91
Random Undersampling	0.87	0.92	0.87	0.90
Tomek links	0.88	0.90	0.93	0.91
Random Oversampling	0.88	0.92	0.90	0.91
SMOTE	0.88	0.92	0.89	0.91
ADASYN	0.85	0.94	0.83	0.88
SMOTE+TOMEK	0.88	0.92	0.89	0.91

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 highlights the top 10 contributors to delayed payments, with the most influential 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.** Addressing these factors could lead to a significant reduction in 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

THANK YOU