E-Commerce & Retail B2B

Thach Thao Ho | DSC65

Problem Statement

- Schuster is a multinational retail company specializing in sports goods and accessories.
- The company has credit arrangements with hundreds of vendors.
- Some vendors fail to respect credit terms and make late payments.
- Schuster imposes heavy late payment fees, which are not beneficial for long-term business relationships.
- Employees spend time chasing vendors for on-time payments, leading to non-value-added activities, time loss, and financial impact.
- Schuster aims to understand customers' payment behavior and predict the likelihood of late payments on open invoices.

Business **Goal**

- Understand customers' payment behavior through past patterns (customer segmentation).
- Predict likelihood of delayed payments using historical data.
- Enable collectors to prioritize follow-ups for timely payments.

Solution Methodology

1. **Read and understand the data.**
2. **Clean the data:**
   - Delete null values.
   - Drop columns with only one value.
   - Remove duplicated columns.
   - Eliminate columns not important for analysis.
3. **Conduct Exploratory Data Analysis:**
   - Check data imbalance.
   - Create derived metrics (e.g., overdue_days, credit_period).
4. Cluster the data.
5. Prepare the data:
   - Treat outliers.
   - Generate dummy variables.
   - Scale features.
   - Split into train and test sets.
6. Build the model.
7. Evaluate the model.
8. Conclude findings.

Grouped data

- **Grouped customers** by the average credit period and the variation (standard deviation) in their credit periods.
- The **mean credit period** follows a normal distribution.
- The **standard deviation** of the credit period shows a left-skewed distribution.
- On average, the time between the invoice date and due date is **38 days**.

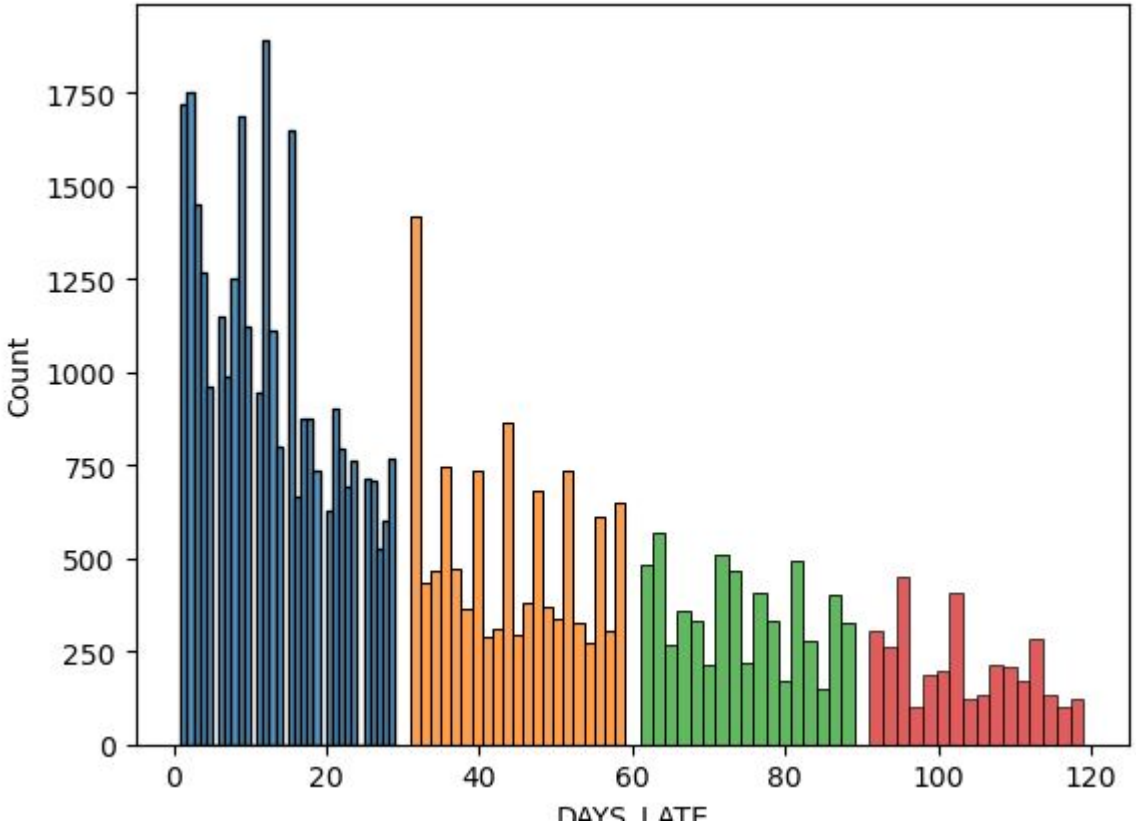
Solution Methodology

1. **Read and understand the data.**
2. **Clean the data:**
   - Delete null values.
   - Drop columns with only one value.
   - Remove duplicated columns.
   - Eliminate columns not important for analysis.
3. **Conduct Exploratory Data Analysis:**
   - Check data imbalance.
   - Create derived metrics (e.g., overdue_days, credit_period).
4. Cluster the data.
5. Prepare the data:
   - Treat outliers.
   - Generate dummy variables.
   - Scale features.
   - Split into train and test sets.
6. Build the model.
7. Evaluate the model.
8. Conclude findings.

Grouped data

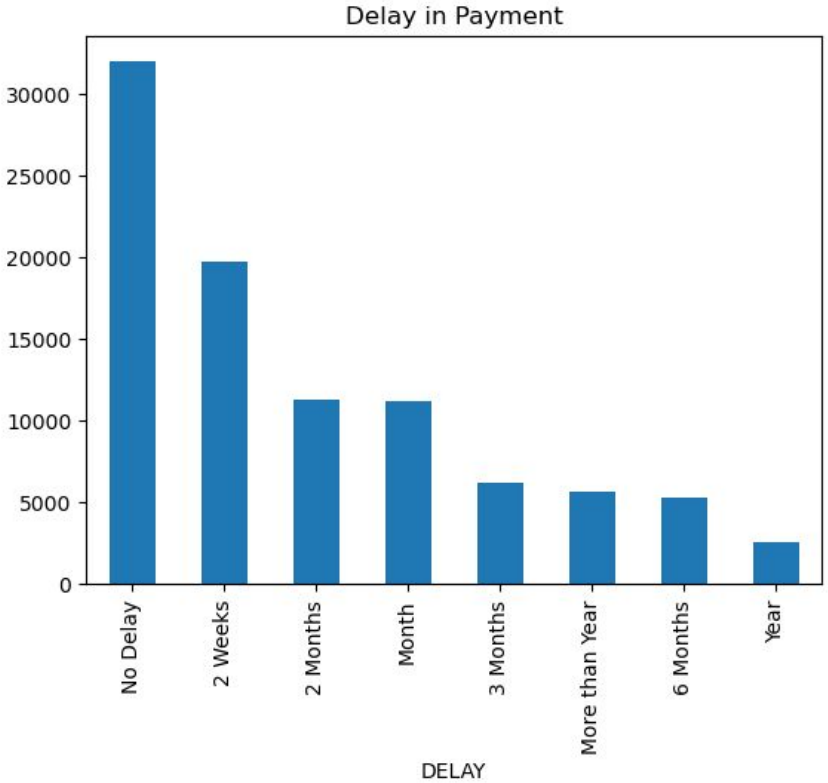
- **Grouped customers** by the average credit period and the variation (standard deviation) in their credit periods.
- The **mean credit period** follows a normal distribution.
- The **standard deviation** of the credit period shows a left-skewed distribution.
- On average, the time between the invoice date and due date is **38 days**.

Exploratory Data Analysis

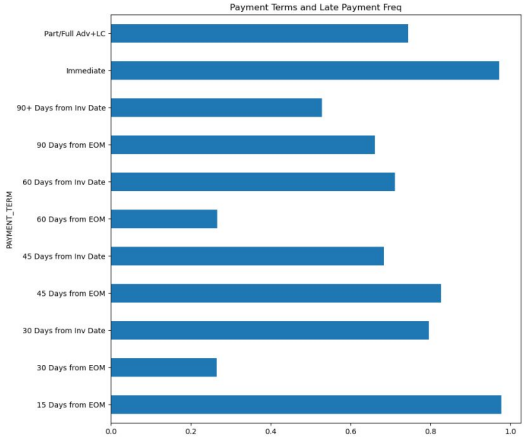


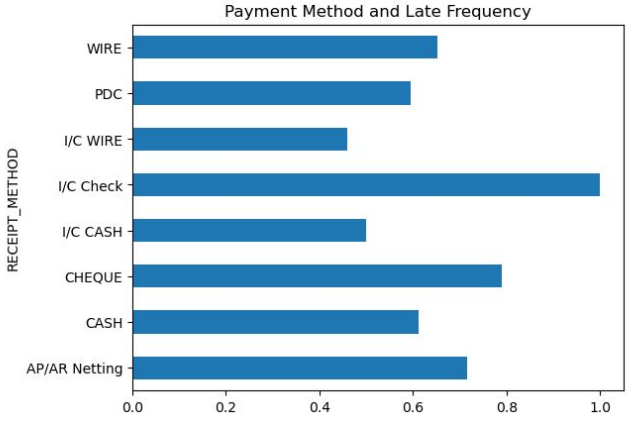
Distribution of payments across due days

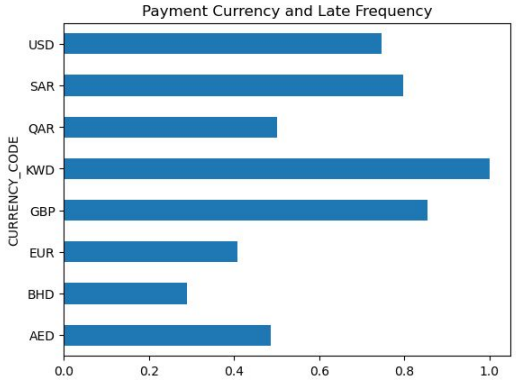
Exploratory Data Analysis

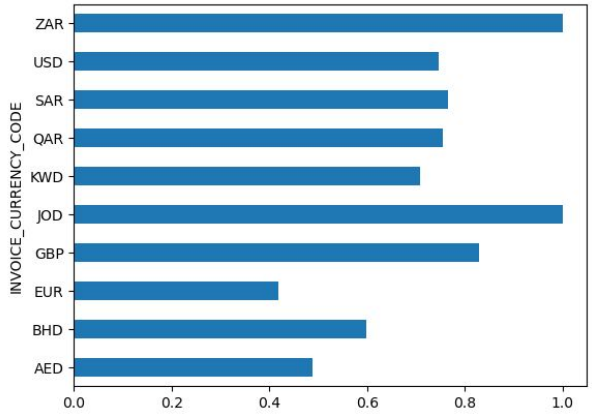


- The majority of payments are made on time.
- Payments delayed by two weeks are higher than the average.

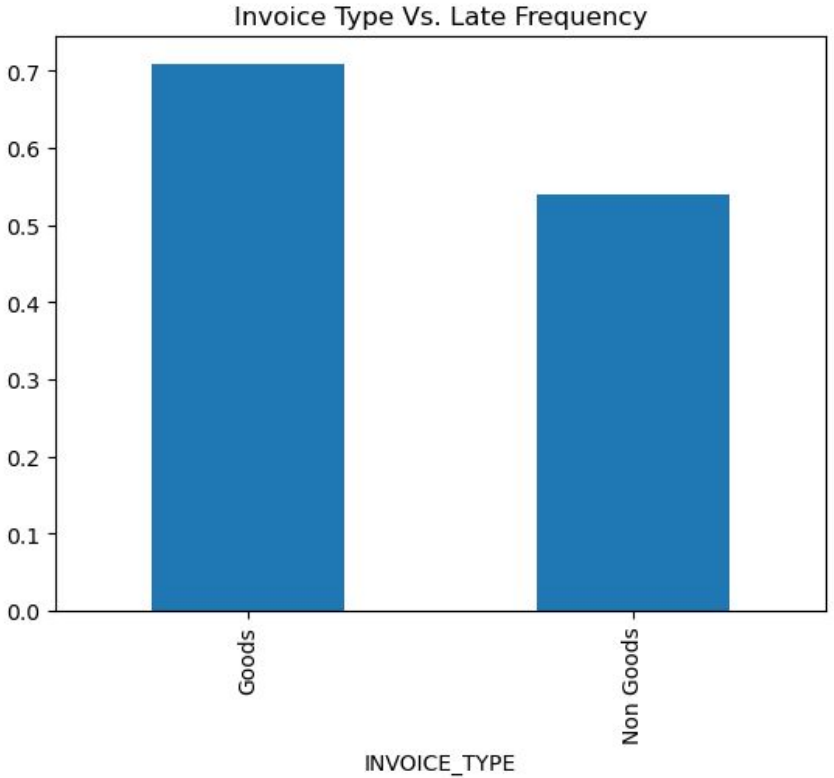

Payment Terms and Late Payment Freq
PAYMENT_TERM
Part/Full Adv+LC
Immediate
90+ Days from Inv Date
90 Days from EOM
60 Days from Inv Date
60 Days from EOM
45 Days from Inv Date
45 Days from EOM
30 Days from Inv Date
30 Days from EOM
15 Days from EOM
0.0
0.2
0.4
0.6
0.8
1.0

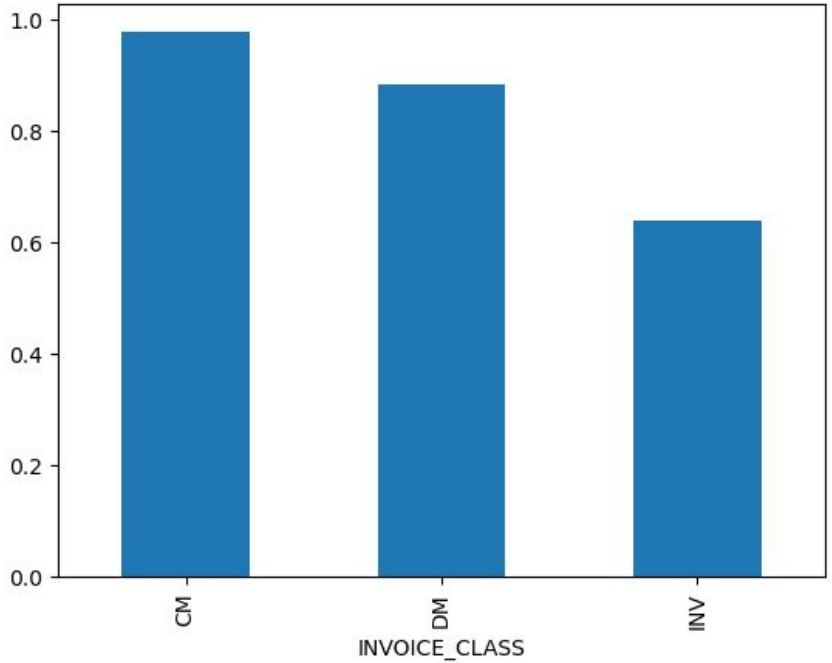

Payment Method and Late Frequency
RECEIPT_METHOD
WIRE
PDC
I/C WIRE
I/C Check
I/C CASH
CHEQUE
CASH
AP/AR Netting
0.0
0.2
0.4
0.6
0.8
1.0


Payment Currency and Late Frequency
CURRENCY_CODE
USD
SAR
QAR
KWD
GBP
EUR
BHD
AED
0.0
0.2
0.4
0.6
0.8
1.0

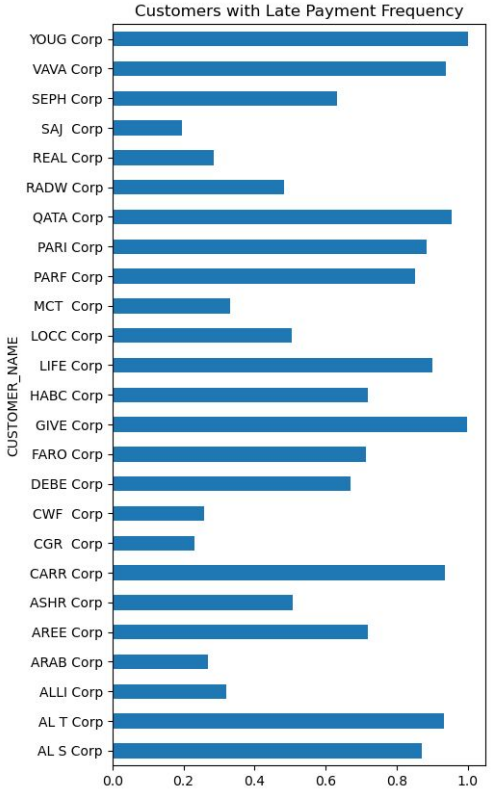

INVOICE_CURRENCY_CODE
ZAR
USD
SAR
QAR
KWD
JOD
GBP
EUR
BHD
AED
0.0
0.2
0.4
0.6
0.8
1.0

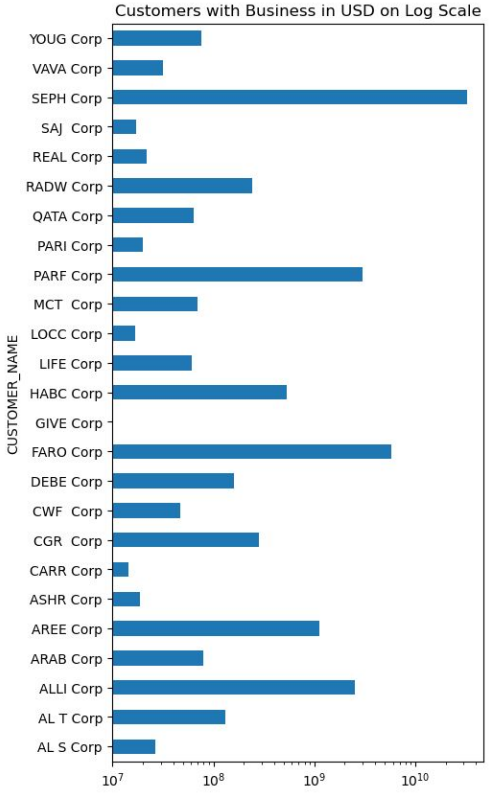
Exploratory Data Analysis



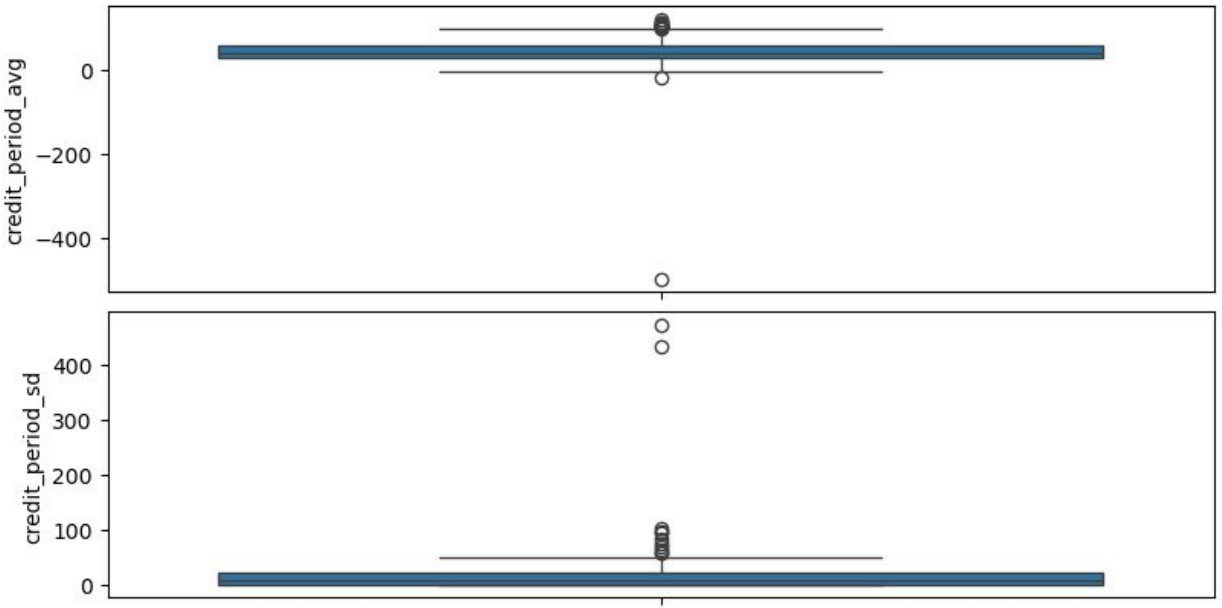


Exploratory Data Analysis

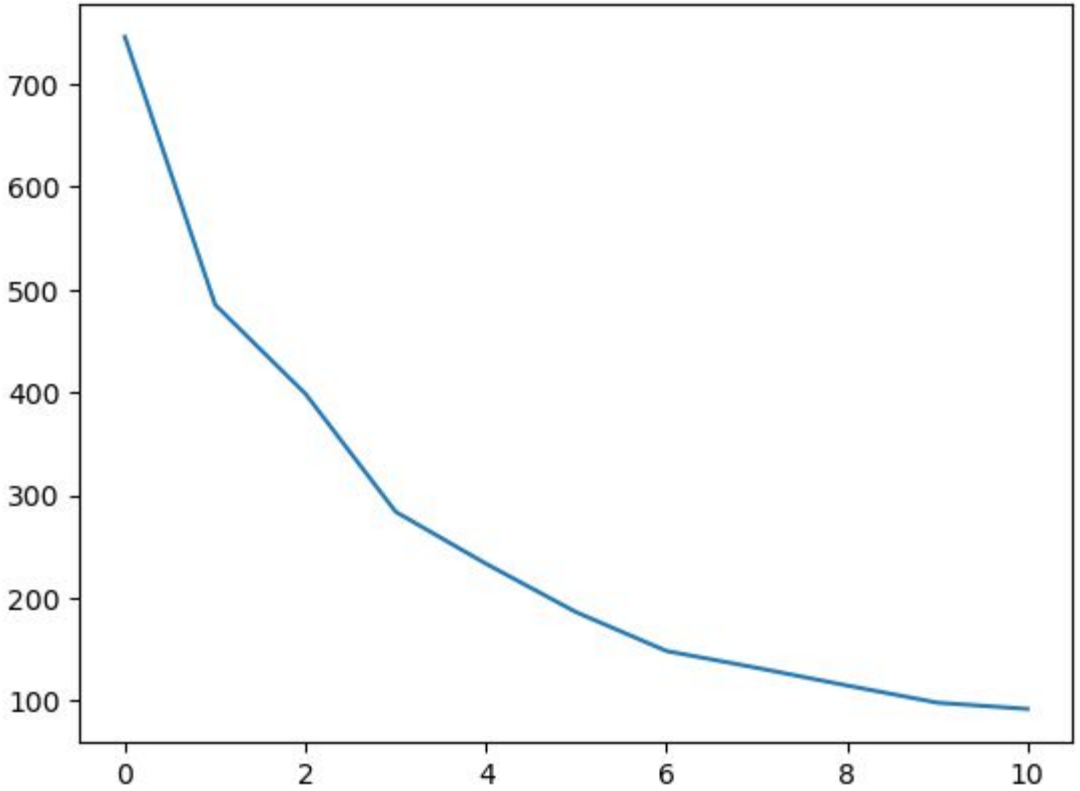




Outlier treatment and scaling

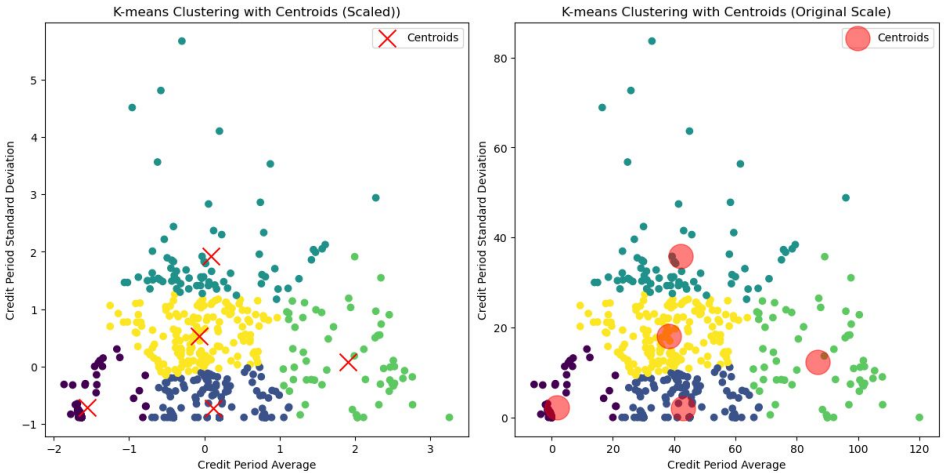


Customer Segmentation



- Performed K-Means clustering on the scaled data.
- Used the Elbow curve and Silhouette score to identify the optimal number of clusters.
- To maintain significance, selected k=5 clusters based on the analysis of the Silhouette score.

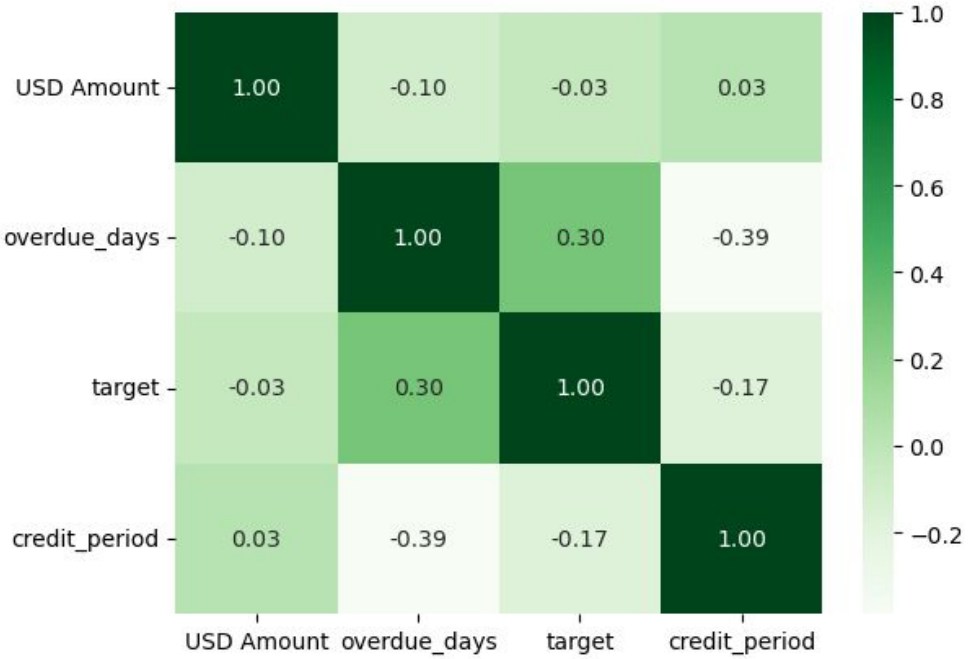
Clustering



5 distinct customer clusters have emerged, each with different average payment days.

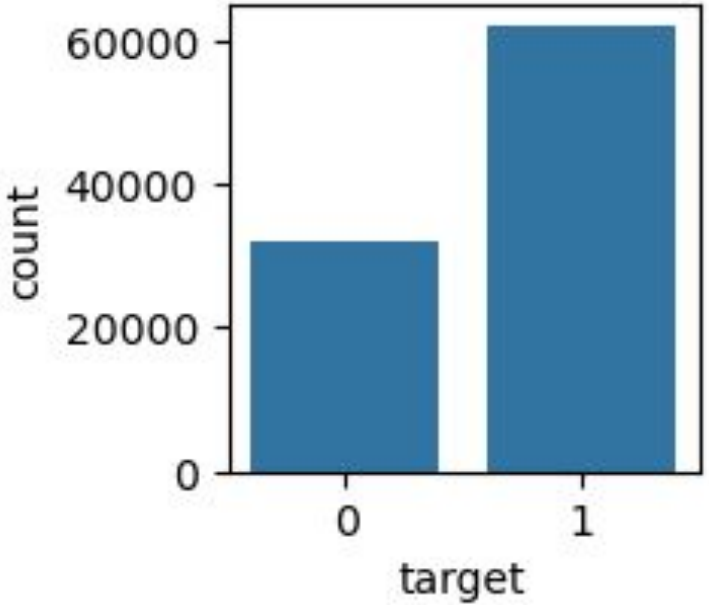
- The majority of customers receive payment terms between 20 to 60 days on average (represented by the blue, purple, and yellow clusters).
- When the average credit period is less than 20 days, there is minimal variation in the credit period (peacock blue cluster).
- For credit periods exceeding 60 days, there is moderate variability (green cluster).
- The greatest variability is observed in the 20 to 60-day credit period range.

Check correlations between Numerical columns



Class imbalance and Data preparation

- Examined class imbalance by analyzing the percentage of delayed and non-delayed entries in the target column.
- The dataset shows a moderate imbalance, with around 66% of entries being delayed and 34% not delayed.



Feature selection

Chosen the top features with importance values greater than 0.02, while dropping other columns that contribute minimally.

USD Amount	0.501687
credit_period	0.190153
PAYMENT_TERM_30 Days from EOM	0.076405
PAYMENT_TERM_60 Days from EOM	0.067681
INVOICE_CURRENCY_CODE_SAR	0.028327
PAYMENT_TERM_15 Days from EOM	0.021997

Class imbalance and Model selection

Logistic Regression	Accuracy	Precision	Recall	F1 Score
Base	0.7	0.73	0.86	0.79
Random Understanding	0.75	0.83	0.79	0.81
Tomek links	0.7	0.74	0.84	0.79
Random Oversampling	0.74	0.84	0.74	0.79
SMOTE	0.72	0.81	0.75	0.78
ADASYN	0.72	0.83	0.72	0.77
SMOTE + TOMEX	0.77	0.81	0.86	0.83

Random Forest	Accuracy	Precision	Recall	F1 Score
Base	0.88	0.89	0.94	0.91
Random Understanding	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 + TOMEX	0.88	0.92	0.89	0.91

- The base model and Tomek links yielded the best results among the imbalance techniques.
- Random Forest achieved the highest accuracy, precision, recall, and F1 score, while Logistic Regression had a better recall but lower accuracy and precision.
- I'll proceed with Random Forest without applying any class imbalance adjustments.
- The dataset shows a slight imbalance with 64% delayed payments and 36% non-delayed payments.

Random Forest model and Hyperparameter tuning

Hyperparameter tuning was done using the GridsearchCV method. The model's performance on the training data shows an accuracy of 84.7%, and the validation accuracy is 85%, indicating no overfitting.

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

confussion matrix:

```
[[ 6665  2923]
[ 1319 17275]]
```

The model delivers strong accuracy, precision, recall, and F1-score, with particularly high recall, demonstrating its ability to predict a large proportion of delayed payments.

Conclusion & Recommendation

The analysis highlights the top 10 factors contributing to delayed payments, with the most influential being USD Amount, credit_period, and specific terms like PAYMENT_TERM_30 Days from EOM and PAYMENT_TERM_60 Days from EOM. Focusing on these key areas could help minimize payment delays effectively.

USD Amount	0.499642
credit_period	0.175609
PAYMENT_TERM_30 Days from EOM	0.088045
PAYMENT_TERM_60 Days from EOM	0.071716
INVOICE_CURRENCY_CODE_SAR	0.028733
PAYMENT_TERM_15 Days from EOM	0.022831
INVOICE_CURRENCY_CODE_USD	0.017957
PAYMENT_TERM_Immediate Payment	0.013858
PAYMENT_TERM_60 Days from Inv Date	0.013759
INVOICE_CLASS_INV	0.012659

The client should consider using milestone or staggered invoicing instead of waiting to invoice the entire order at once. Recommended payment terms include:

- 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

Exercise caution with PAYMENT_TERM_30 Days from EOM and PAYMENT_TERM_60 Days from EOM. For INVOICE_CURRENCY_CODE, the best options are ZAR, QAR, and GBP, while SAR and USD require careful consideration.

Thank You!