E-Commerce and Retail B2B Case Study

Insha Durwesh

Introduction

Problem Statement

Schuster, a global retailer specializing in sports goods and accessories, works with hundreds of vendors under established credit terms. However, some vendors frequently miss payment deadlines, resulting in substantial late fees. While these fees impose a financial penalty, they're ultimately detrimental to sustaining strong, long-term vendor relationships. Currently, Schuster dedicates resources to follow up on overdue payments, a process that adds limited value, consumes time, and incurs costs. The company now aims to analyze vendor payment behaviors and forecast the likelihood of delayed payments on outstanding invoices.

Objective

- Schuster aims to gain deeper insights into vendor payment behaviors by analyzing historical payment patterns and segmenting its customers accordingly.
- By leveraging this historical data, Schuster wants to predict the probability of delayed payments on outstanding invoices.
- With these predictions, collectors can better prioritize follow-ups, proactively reaching out to vendors to ensure timely payments.

Solution Methodology

1. **Data Familiarization:** We started by analyzing the dataset to understand its structure, variables, and context, ensuring a strong foundation for the analysis.

2. **Data Cleaning:** We enhanced data quality by removing null values, dropping columns with single values, eliminating duplicates, and excluding irrelevant fields.

3. **Exploratory Data Analysis (EDA):** During EDA, we checked for data imbalances and created derived metrics like **overdue_days** and **credit_period** for deeper insights.

4. **Clustering:** Clustering techniques helped us identify natural groupings within the data, revealing key patterns.

5. **Data Preparation:** We treated outliers, created dummy variables, scaled features, and split the dataset into training and test sets.

6. **Model Building:** Using the prepared data, we developed models to meet project goals and maximize predictive accuracy.

7. **Model Evaluation:** We assessed model performance using key metrics to ensure effectiveness and reliability.

8. **Conclusion:** Finally, we synthesized our findings and identified actionable insights to guide recommendations.

Insights from Exploratory Data Analysis

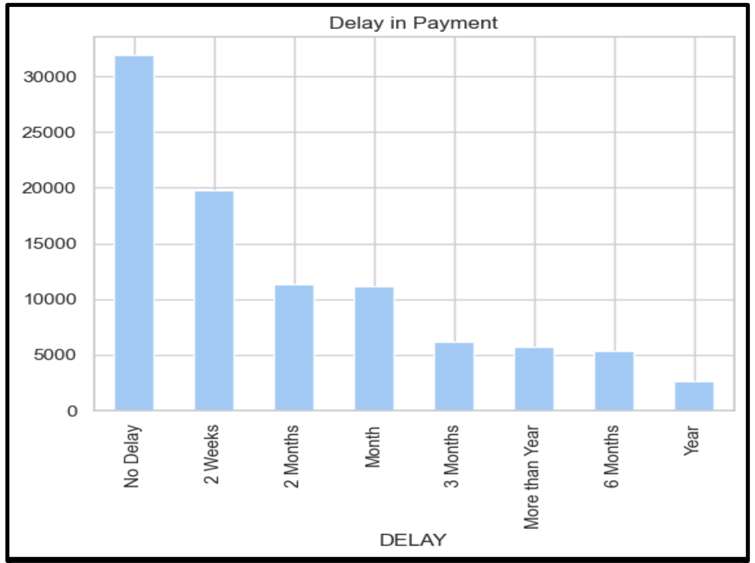
Customers were segmented based on the mean credit period and the standard deviation of their credit periods.

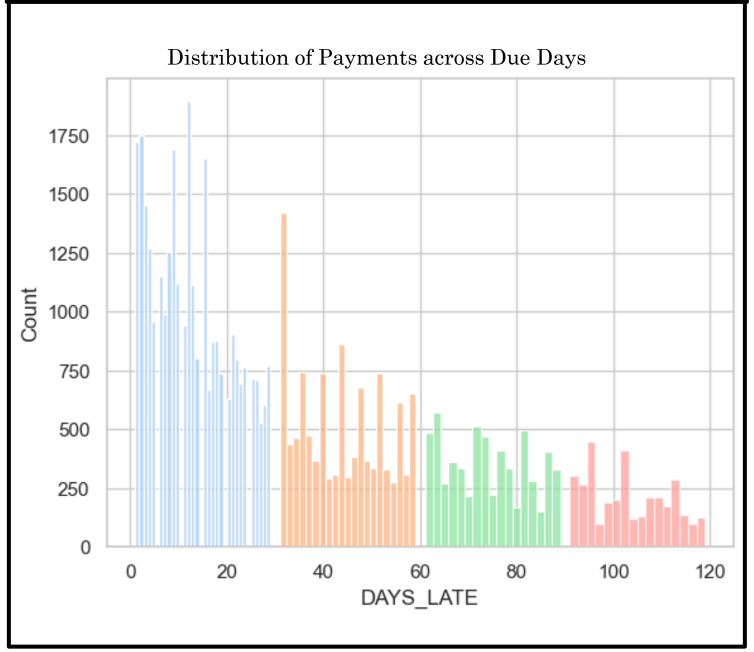
The mean credit period follows a normal distribution. The standard deviation of the credit period is left-skewed.

The average interval from invoice date to due date is 38 days.

The distribution of overdue days shows a right-skewed pattern, with most customers having shorter delays. However, there is a significant tail towards longer delays, indicating a subset of customers with severe payment issues.

It's evident that a significant number of customers make timely payments (No Delay). However, there's a notable proportion of customers who experience delays ranging from 2 weeks to over a year.





Insights from Distribution Analysis

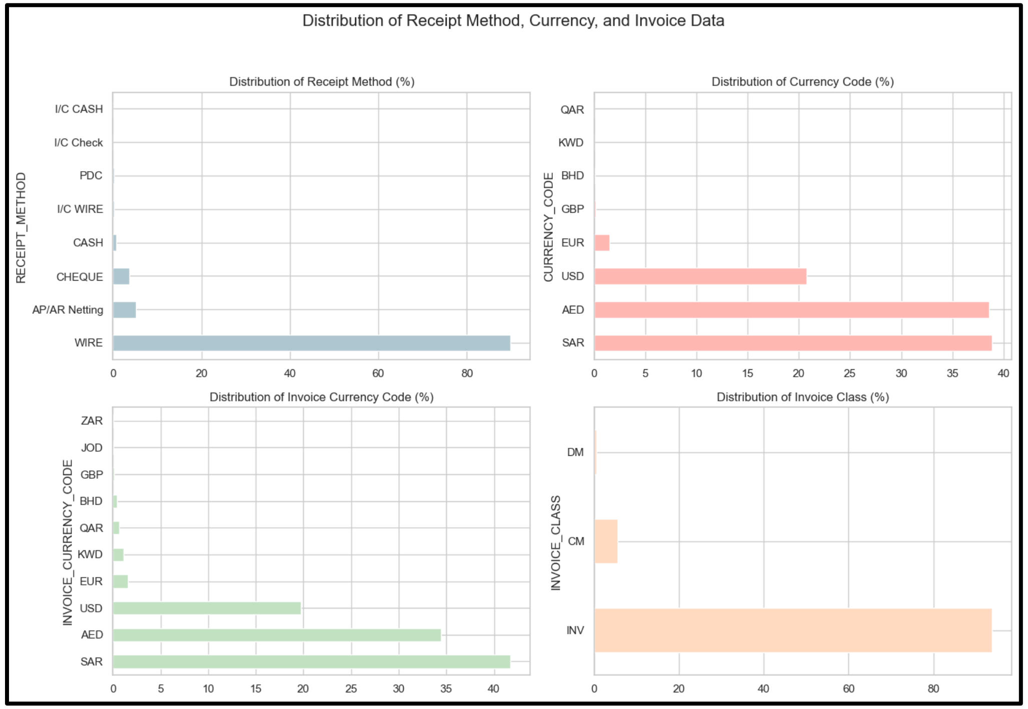
1. **Receipt Method Distribution:** The majority of payments are received via **WIRE** transfer, accounting for over 80% of total receipts. Other methods, such as **Cheque**, **AP/AR Netting**, and **Cash**, are minimally used.

2. **Currency Code Usage**: **AED** and **SAR** are the most commonly used currencies, each representing around 35-40% of the transactions. **USD** is also prevalent, while other currencies like **GBP** and **EUR** are used to a lesser extent.

3. **Invoice Currency Code Preference:** Consistent with the currency code distribution, **AED**, **SAR**, and **USD** dominate the invoice currency codes. Limited use is observed for currencies such as **JOD**, **ZAR**, and **KWD**.

4. **Invoice Class Distribution**: **Invoice (INV)** is the primary invoice class, making up around 90% of total invoices, with **Credit Memos (CM)** and **Debit Memos (DM)** being used sparingly.

These insights suggest a high reliance on wire transfers and specific currency codes (AED, SAR, USD) for transactions, which may influence payment processing times and international payment dynamics.



Insights from Payment Terms and Late Payment Analysis

1. **Payment Terms and Late Payment Frequency:** Terms like **45 Days from EOM** and **60 Days from EOM** have higher late payment frequencies, indicating that longer terms may contribute to delays. Shorter terms (e.g., **15 Days from EOM**) show fewer late payments.

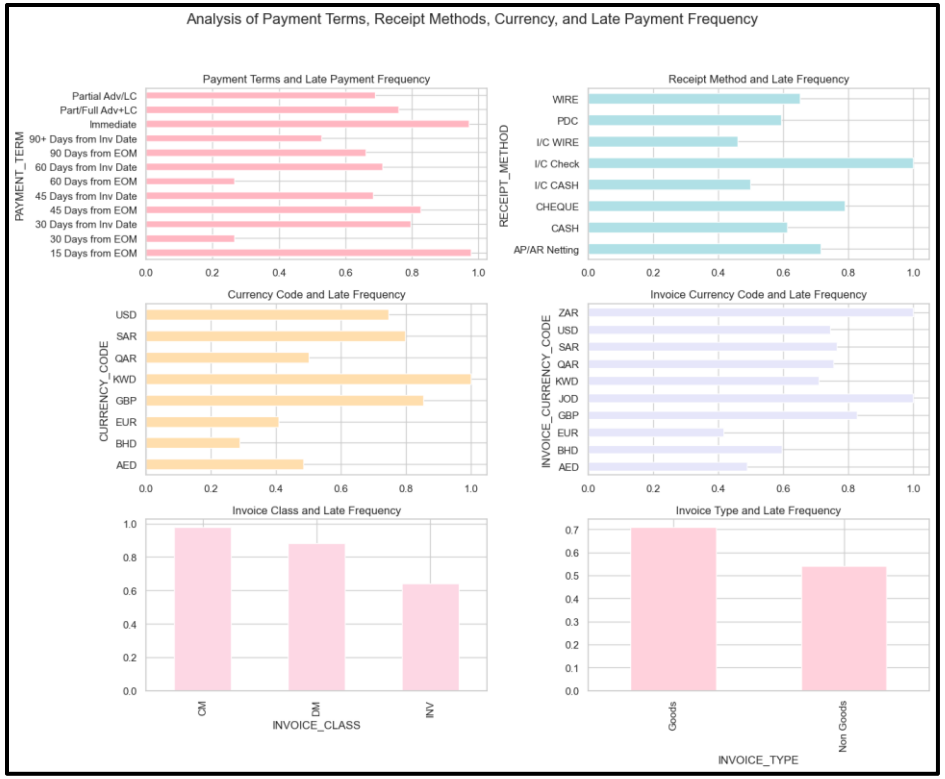
2. **Receipt Method Impact on Late Payments:** Late payments are common across various receipt methods, with **Wire** and **I/C Cash** methods showing slightly higher late frequencies. This suggests that the payment method alone may not be a decisive factor in timely payments.

3. **Currency Code and Late Payment Frequency**: **KWD** and **QAR** have higher frequencies of late payments compared to other currencies. **AED** and **USD** show relatively lower late payment rates, potentially indicating a more reliable payment performance in these currencies.

4. **Invoice Class and Late Payment Frequency**: **Credit Memos (CM)** and **Debit Memos (DM)** have higher late payment frequencies, while **Invoices (INV)** tend to be more punctual. This implies that standard invoices may be easier to process on time than adjustments or corrections.

5. **Invoice Type and Late Frequency:** Late payments are more frequent for **Goods** invoices compared to **Non-Goods** invoices, suggesting a need for special attention to goods-related transactions to ensure timely payments.

These insights highlight the need for evaluating specific payment terms, receipt methods, and currency codes to optimize payment timeliness and reduce late payment occurrences.



Preparing Data for Modelling

- **Outlier Removal and Data Scaling:** Outliers were removed using the Interquartile Range (IQR) method to ensure reliable clustering. The data was then scaled using the Standard Scaler to standardize features, ensuring equal contribution to the clustering algorithm.

- **Clustering Approach:** K-Means clustering was applied, with the optimal number of clusters determined using the Elbow Curve and Silhouette Score. The Elbow Curve helped identify the point of diminishing returns, while the Silhouette Score suggested that too many clusters reduce interpretability.

- **Optimal Cluster Selection:** Based on both metrics, k=4 was chosen for its balance between model complexity and clarity.

- **Data Imbalance Observation:** The dataset showed a moderate imbalance, with 66% delayed entries and 34% non-delayed. This imbalance may affect clustering results and should be considered in interpretation. Further adjustments may be needed for a balanced dataset.

This approach ensures effective clustering while preserving data integrity and clarity.

Customer Segmentation Insights

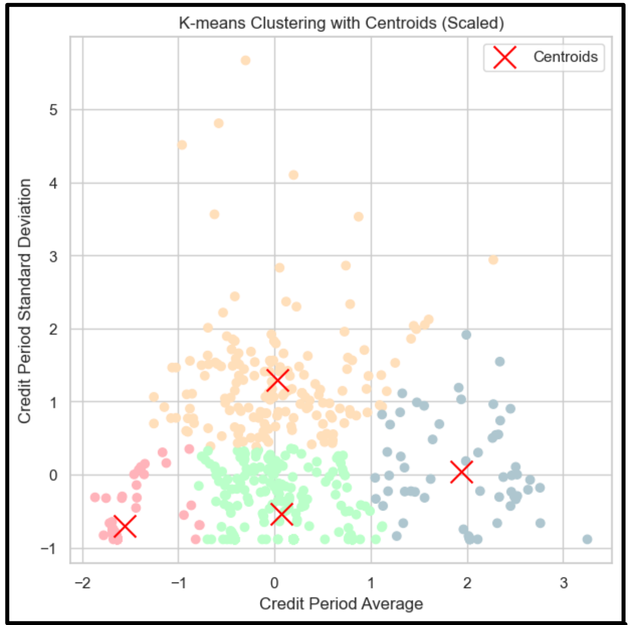
There are 4 distinct clusters of customers with varying average credit periods and standard deviation.

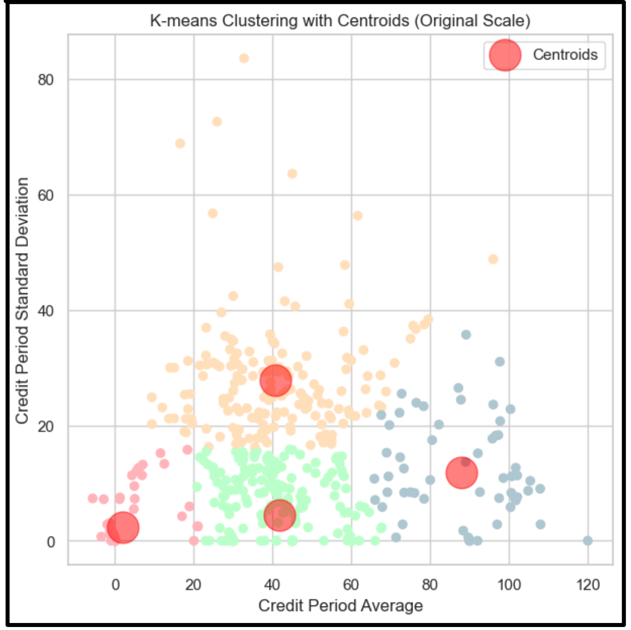
Most customers fall within an average credit period of 20 to 60 days, with moderate variability (Blue and Yellow clusters).

Customers with an average credit period under 20 days show low variability in their credit usage, indicating consistent short-term credit users (Green cluster).

Customers with an average credit period above 60 days have moderate variability, suggesting stable long-term borrowers (Yellow cluster).

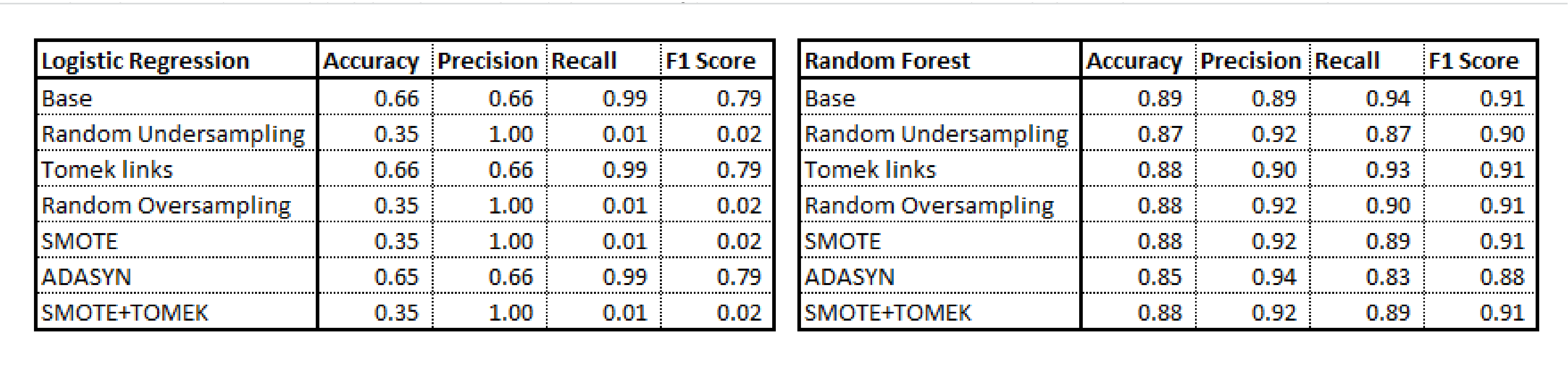
The highest variability is observed among customers with average credit periods of 20 to 40 days, indicating irregular or fluctuating credit patterns (Purple cluster)


K-means Clustering with Centroids (Scaled)


K-means Clustering with Centroids (Original Scale)

Model Building and Selection

- We fitted logistic regression and random forest and experimented with several methods such as undersampling and oversampling using SMOTE, ADASYN, and Tomek links.
- The Base model (without implementing any class imbalance technique) and Tomek links produced the best results compared to other class imbalance methods.
- The Random Forest model demonstrated significantly higher performance, achieving superior Accuracy, Precision, Recall, and F1 Score.
- Logistic Regression exhibited better Recall than Random Forest but had lower Accuracy and Precision.
- Based on these observations, we will proceed with the Random Forest model without applying any class imbalance techniques.



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

Insights from the Model

- Hyperparameter tuning is achieved using GridsearchCV method.
- We will finalize the model for future use because it demonstrates very high accuracy, precision, recall, and F1-score.
- The recall is especially high, indicating that the model effectively predicts a significant proportion of delayed payments.
- The most significant contributor to delayed payments is the USD Amount, followed by factors such as the credit period and payment terms, with specific payment terms like "30 Days from EOM" and "60 Days from EOM" also playing notable roles. These factors highlight key areas for improving payment timelines.

Conclusion on the Model

The developed classification model, based on the Random Forest algorithm, demonstrates strong performance in predicting customer payment behavior.

Key Performance Metrics:

- Accuracy: The model accurately predicts 85% of cases.
- Precision: The model correctly identifies 83% of customers as not delayed and 86% as delayed.
- Recall: The model correctly identifies 70% of customers who are not delayed and 93% of customers who are delayed.
- F1-Score: The model achieves an F1-score of 0.76 for not delayed customers and 0.89 for delayed customers, indicating a good balance between precision and recall.

Model Implications:

- Effective Customer Segmentation: The model can be used to identify customers at risk of payment delays, enabling targeted interventions and personalized strategies.
- Improved Decision-Making: By accurately predicting customer behavior, businesses can optimize their credit and collection processes, reducing losses and improving cash flow.
- Enhanced Customer Relationships: Proactive outreach to customers at risk of delay can strengthen customer relationships and prevent negative consequences.

Business Recommendations

Adopt Milestone or Staggered Invoicing

- Consider invoicing in stages based on project milestones, rather than waiting to bill for the entire order at once.

Recommended Payment Terms

- PAYMENT_TERM_ 180 Days from Invoice Date
- PAYMENT_TERM_ Advance with Discount
- PAYMENT_TERM_ 120 Days from End of Month (EOM)
- PAYMENT_TERM_ 7 Days from EOM
- PAYMENT_TERM_ Standby Letter of Credit (LC) at 30 Days

Caution with Certain Payment Terms

- Be cautious with terms like 30 Days from EOM and 60 Days from EOM, as these often lead to delayed payments.

Prioritize Reliable Currencies

- Prefer using **ZAR, QAR, and GBP** for invoicing.
- Exercise caution with **SAR** and **USD** as they are more prone to delayed payments.

Best Practices Recommendations for Schuster

- **Establish Clear Payment Schedules:** Define specific milestones and associated payment deadlines to set clear expectations from the start.
- **Maintain Open Communication:** Regularly update clients on invoice statuses and payment timelines to avoid misunderstandings and keep everyone informed.
- **Use Automated Payment Reminders:** Set up automated reminders for upcoming payments to minimize delays and support consistent cash flow.
- **Implement Late Payment Fees:** Clearly outline late fees in contracts to encourage timely payments and reinforce the importance of prompt invoicing.

By implementing transparent invoicing practices and structured payment terms, clients can build stronger financial relationships and contribute to smoother project completion.

Thank You!

Turning chaos into clarity, one datapoint at a time.