E-Commerce and Retail B2B Case Study

Presented By :

Prathamesh Salunke
Sarthak kapoor
Sandeep Chintala

DS C67 April 2024

Addressing the issue and defining objectives

Problem Identification

- Schuster, a multinational sports retail company, deals with hundreds of vendors in the B2B space.
- Payments are often provided on credit terms, but many vendors fail to pay on time.

- Issues caused by late payments:

- Financial lag, leading to cash flow disruptions.
- Employees waste time chasing overdue payments.
- Long-term business relationships are strained due to heavy late payment penalties.

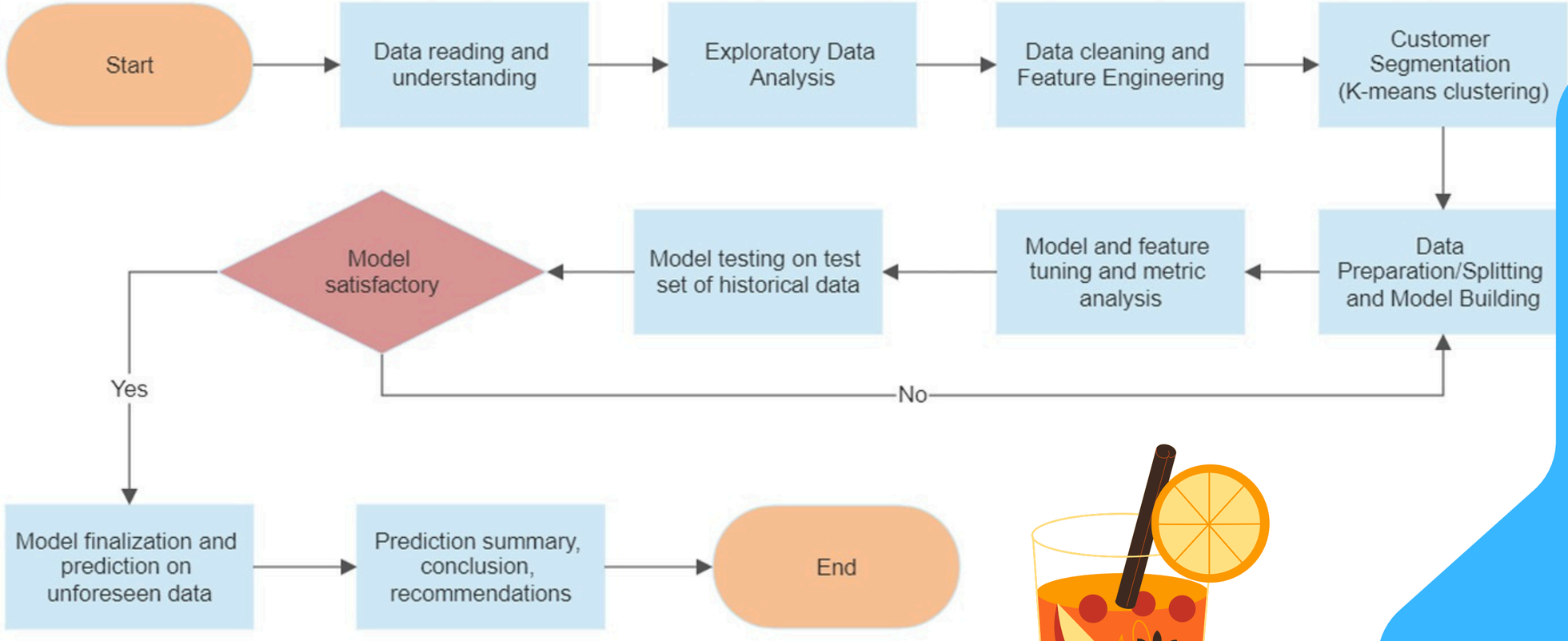
Business Objectives

- Segment customers based on historical payment behavior to understand patterns.
- Use past transaction data to predict the likelihood of late payments for open invoices.

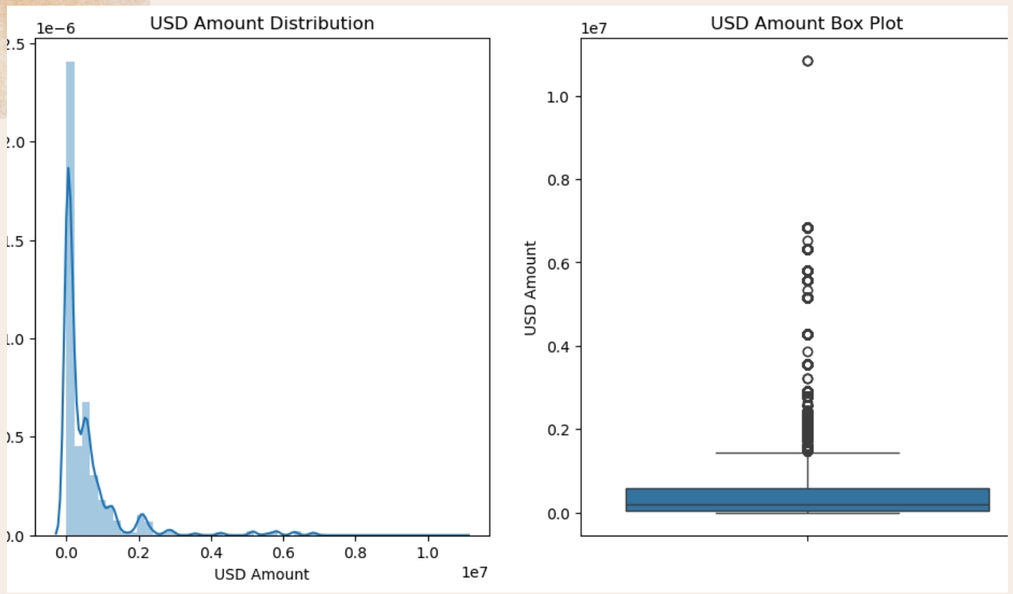
- Enable resource prioritization for:

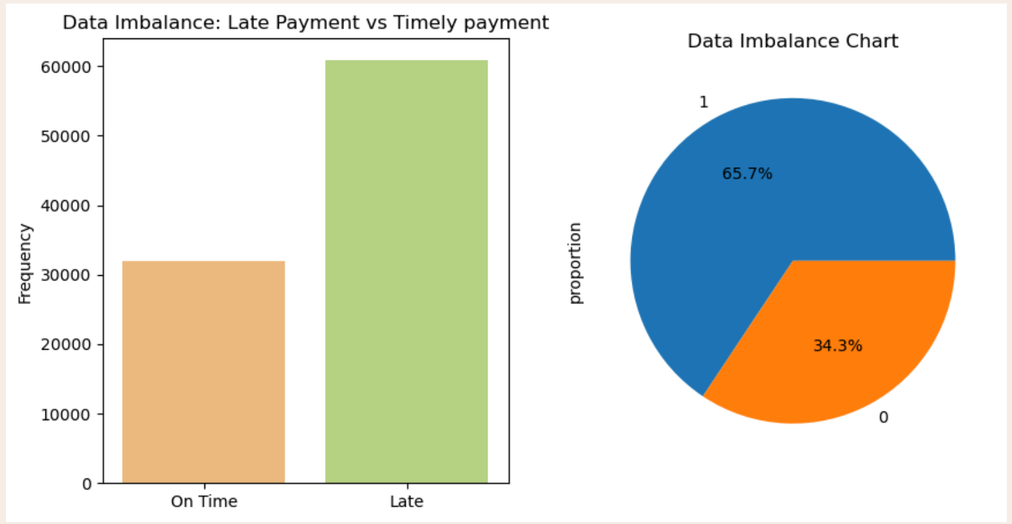
- Proactive follow-ups with high-risk customers.
- Reduction of non-value-added activities.
- Improve overall cash flow management and operational efficiency.

Approach Strategy to the Problem



Class imbalance and transaction insights (univariate)





Most transactions are below $1.75M, with a few high-value outliers above $2M.

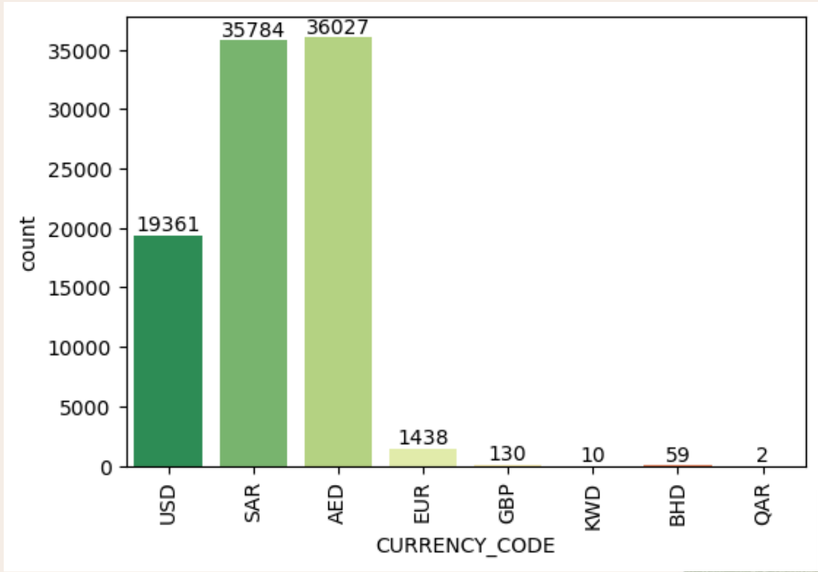
Low-value transactions dominate and are more prone to delays.

High-value transactions show a lower risk of late payments.

Currencies: AED > SAR > USD.

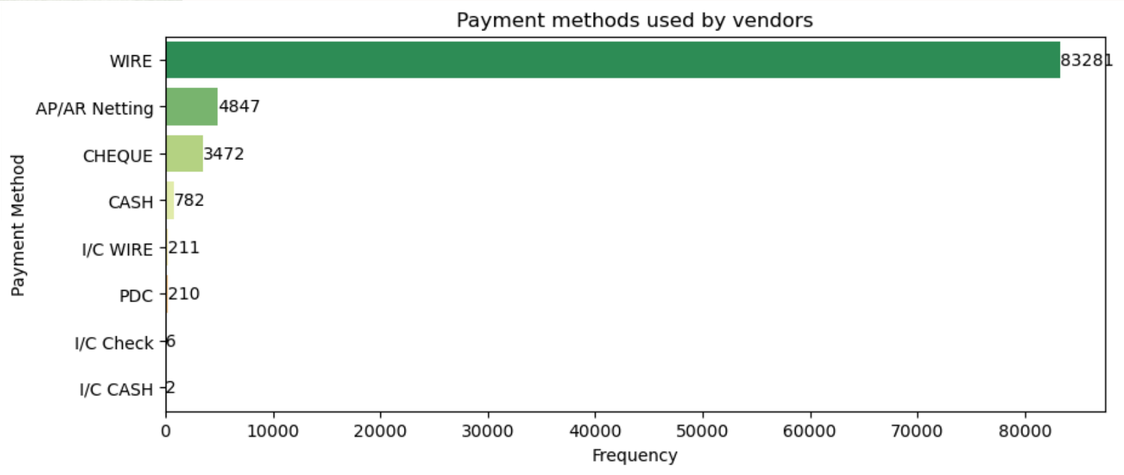
Invoice Types: Goods dominate and show higher late payment rates.

Payment Method: Wire transfers are the most common.

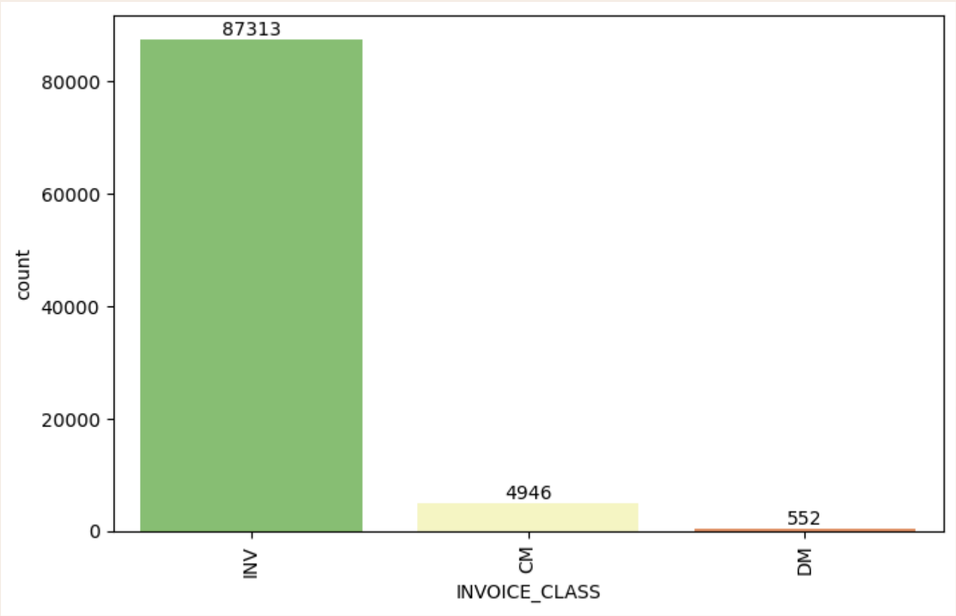


Late Payments: 65%
On-Time Payments: 35%
Imbalance is acceptable for this analysis.

univariate

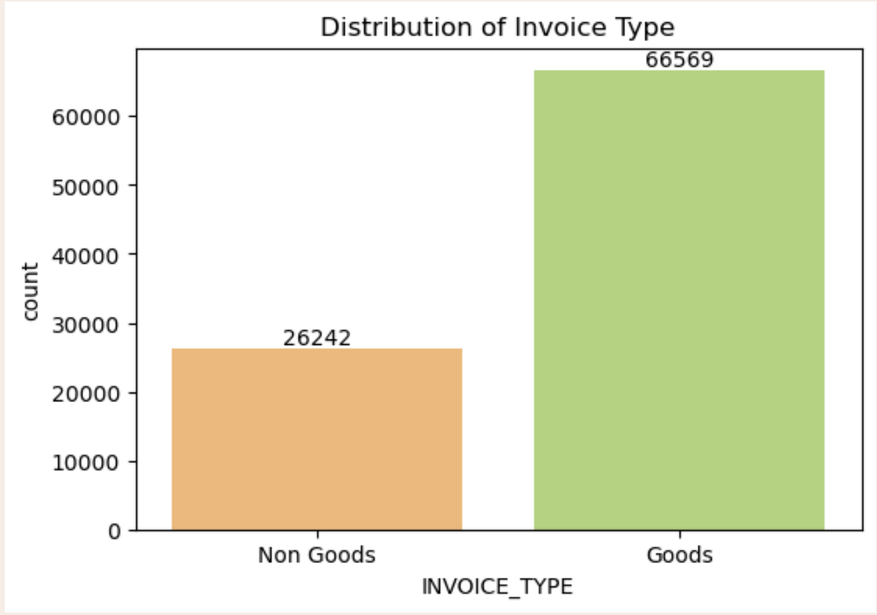


Wire transfer is the dominant payment method, accounting for the majority of transactions (83,281).
The prevalence of wire transfers indicates a streamlined payment process, but late payments may still occur due to credit terms.



"INV" (Invoices) form the majority class with 87,313 transactions, followed by Credit Memos (CM) and Debit Memos (DM).

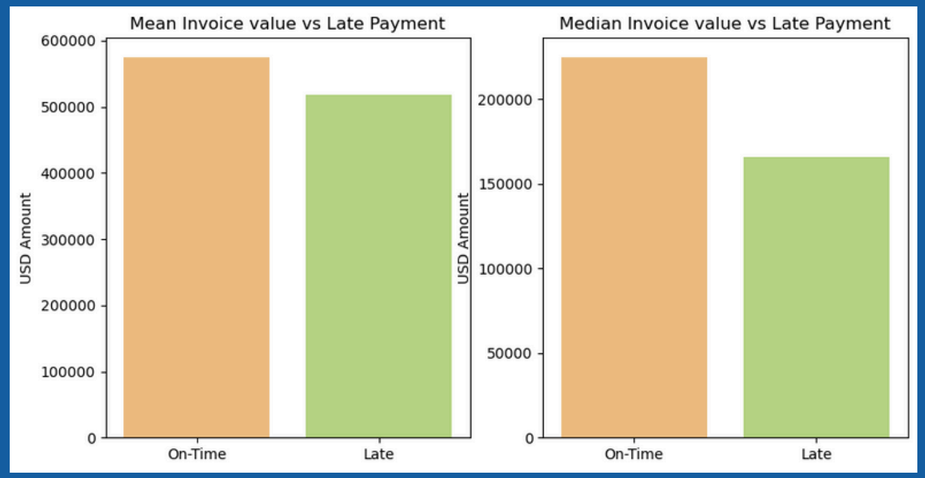
Late payments are more frequent for Credit Memos and Debit Memos, requiring tailored credit control policies.



Insight: Goods invoices dominate, making up 71.7% of the total transactions (66,569 out of 92,811).

Goods invoices show higher late payment risks compared to non-goods invoices, warranting stricter policies for goods-related transactions.

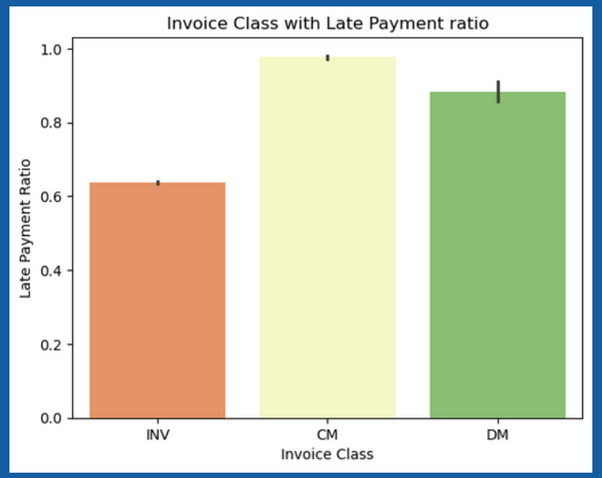
Identifying characteristics of defaulterpayment types (Bivariate)



Higher mean and median invoice values (~$600,000 and ~$250,000).

Lower values, indicating smaller transactions are more prone to delays.

Smaller invoice amounts require additional attention to mitigate payment delays.

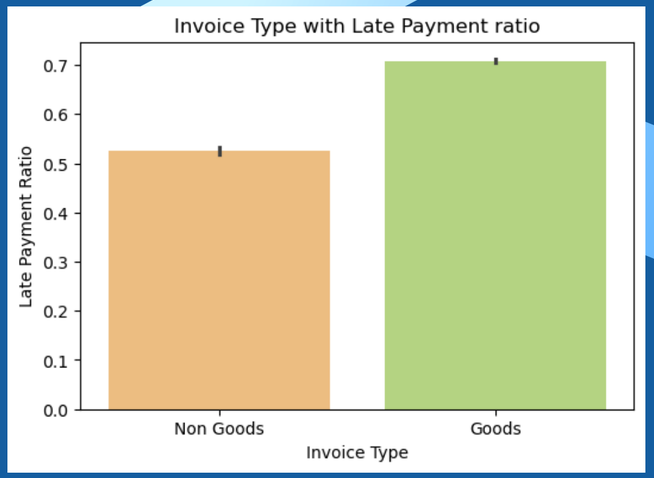


Credit Memos (CM): Highest late payment ratio (~90%).

Debit Memos (DM): Second-highest late payment ratio (~85%).

Invoices (INV): Relatively lower ratio (~65%).

Payment delays are more common in credit and debit memos, highlighting the need for stricter policies.



Goods Invoices: Higher late payment ratio (~70%).

Non-Goods Invoices: Lower ratio (~55%).

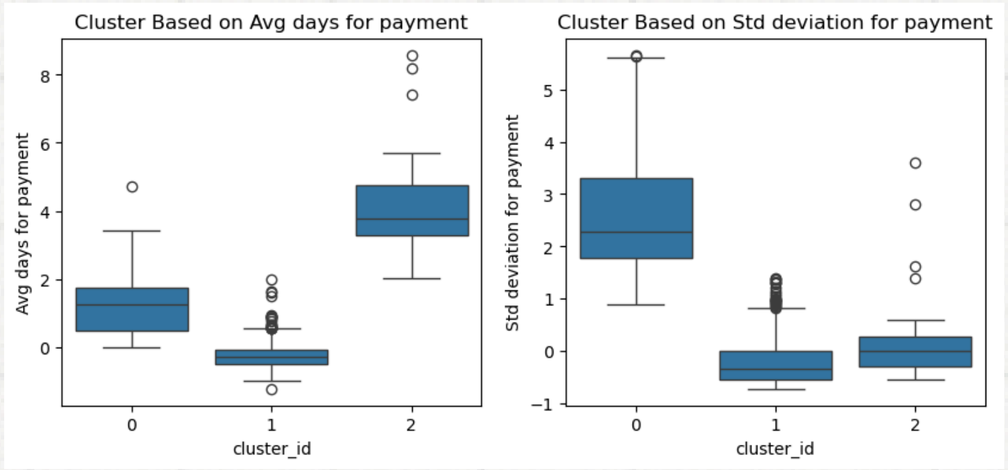
Goods-related transactions are riskier, warranting stricter payment terms or proactive follow-ups.

Customer Segmentation Using K-means Clustering

Cluster 0: Medium-duration payers with inconsistent behavior.

Cluster 1: Early payers with consistent habits.

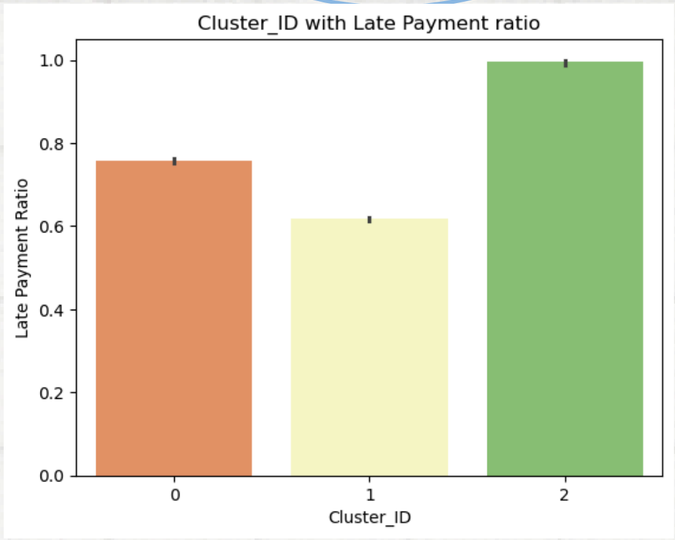
Cluster 2: Prolonged payers with high delays and consistent late payment behavior.



Left Graph: Cluster distribution based on average payment days.

Right Graph: Cluster distribution based on payment standard deviation.

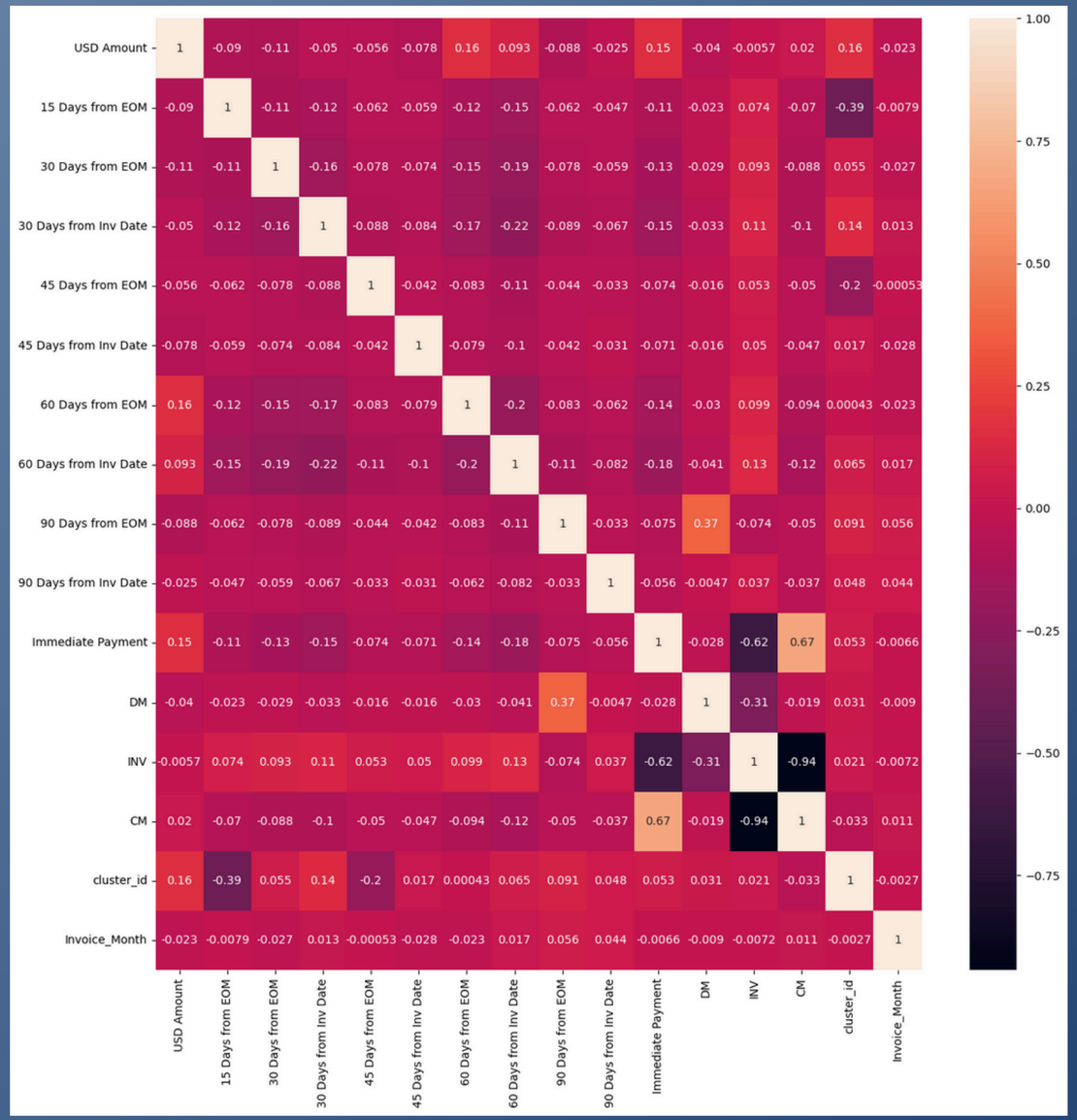
Prolonged payers show the highest delay rates.




```
For n_clusters=2, the silhouette score is 0.7557759850933141
For n_clusters=3, the silhouette score is 0.7491797445652462
For n_clusters=4, the silhouette score is 0.6097388985555463
For n_clusters=5, the silhouette score is 0.6173540681032771
For n_clusters=6, the silhouette score is 0.3980238443004184
For n_clusters=7, the silhouette score is 0.4012628375918799
For n_clusters=8, the silhouette score is 0.41457849738976615
```

3 clusters chosen based on silhouette score (optimal at 3 clusters with a score of 0.749).

Model Building



High Positive Correlations:

CM (Credit Memo) and Immediate Payment are strongly correlated (~0.67), indicating that immediate payments are likely for credit-related invoices.

Cluster ID shows a significant correlation with Avg Payment Days and Std Deviation, confirming its importance in predicting late payments.

High Negative Correlations:

INV (Invoices) and CM show a strong inverse correlation (~-0.94), highlighting their distinct payment patterns.

Minimal correlation between USD Amount and other predictors, suggesting its independent contribution to late payment prediction.

Business Context:

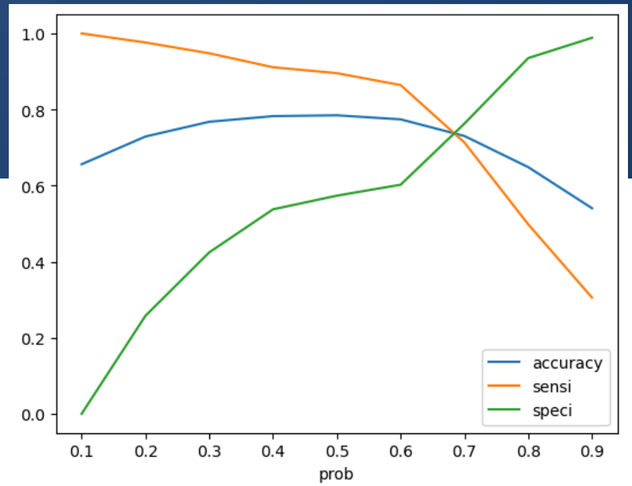
The heatmap validates the inclusion of key features like Cluster ID, Invoice Type, and Immediate Payment in the model.

Redundant or highly correlated variables, such as CM and INV, may require careful handling to avoid multicollinearity.

Comparison between two models, logistic regression and random forests

Accuracy-Sensitivity-Specificity Tradeoff

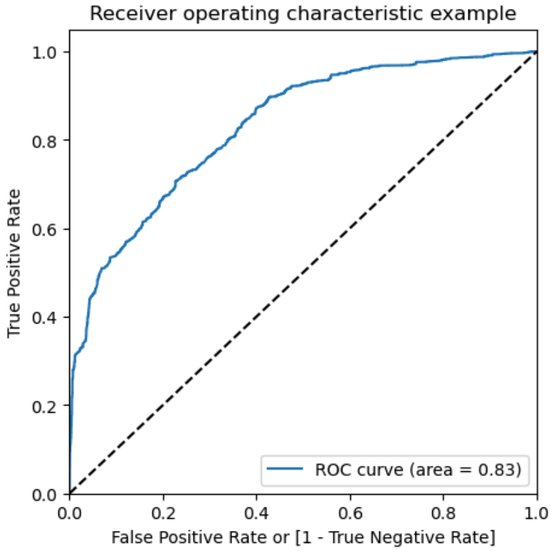
Accuracy (Blue Line): Peaks around 0.7 but drops at extreme probability thresholds.



Sensitivity (Orange Line): Decreases as the threshold increases, indicating a tradeoff in identifying true positives (late payments).

Specificity (Green Line): Increases with higher thresholds, showing better identification of non-defaulters.

Setting the threshold requires balancing sensitivity and specificity to align with business priorities (e.g., avoiding missed late payments).



Receiver Operating Characteristic (ROC) Curve

AUC (Area Under Curve): 0.83, reflecting strong discriminatory power of the Logistic Regression model.

The curve demonstrates consistent tradeoffs between false positive and true positive rates.

Logistic Regression performs well as a baseline model, offering reliable predictions with a clear sensitivity-specificity balance.

Comparison between two models logistic regression and random forests

Random Forest:

```
Best hyperparameters: {'max_depth': 30, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 150}
Best f1 score: 0.9393260434851571
```

Handles non-linearity and variable interactions effectively.
Higher recall and precision, making it better for identifying high-risk defaulters.
Computationally heavier and less interpretable.

"Using the above parameters, a random forest model was built, whose metrics were compared to the logistic regression model and the final model was finalized therefore"

Random Forest found better than Logistic Regression

(Random Forest Metrics - Test Set)

	precision	recall	f1-score	support
0	0.91	0.86	0.88	9529
1	0.93	0.96	0.94	18315
accuracy			0.92	27844
macro avg	0.92	0.91	0.91	27844
weighted avg	0.92	0.92	0.92	27844

(Logistic Regression Metrics - Test Set)

```
# Let's check the overall accuracy.
accuracy_score(y_pred_final.payment_status, y_pred_final.final_predicted)
```

```
0.7758583536848154
```

```
#precision score
precision_score(y_pred_final.payment_status, y_pred_final.final_predicted)
```

```
0.8077275971046998
```

```
# Recall Score
recall_score(y_pred.default, y_pred.final_predicted)
```

```
0.8641419118682246
```

Logistic Regression is a strong baseline model with good accuracy and precision but misses non-linear interactions.

Random Forest outperforms Logistic Regression in all metrics, especially recall, making it the preferred choice for identifying high-risk defaulters.

Random Forest Feature Ratings

```
Feature ranking:
1. USD Amount (0.491)
2. Invoice_Month (0.129)
3. 30 Days from EOM (0.114)
4. 60 Days from EOM (0.111)
5. Immediate Payment (0.041)
6. 15 Days from EOM (0.028)
7. cluster_id (0.027)
8. 60 Days from Inv Date (0.013)
9. 30 Days from Inv Date (0.011)
10. 90 Days from Inv Date (0.008)
11. INV (0.007)
12. 90 Days from EOM (0.007)
13. 45 Days from EOM (0.005)
14. CM (0.004)
15. 45 Days from Inv Date (0.004)
16. DM (0.001)
```

Cluster_ID (0.027): Customer segmentation effectively contributes to identifying defaulters, emphasizing the importance of clustering.

Other Payment Terms: Features like "90 Days from Inv Date" have minor influence but still contribute to model accuracy.

USD Amount (0.491): Most influential feature, indicating that higher transaction amounts significantly impact payment behavior.

Invoice_Month (0.129): Seasonal trends in payment behavior suggest specific months have higher late payment risks.

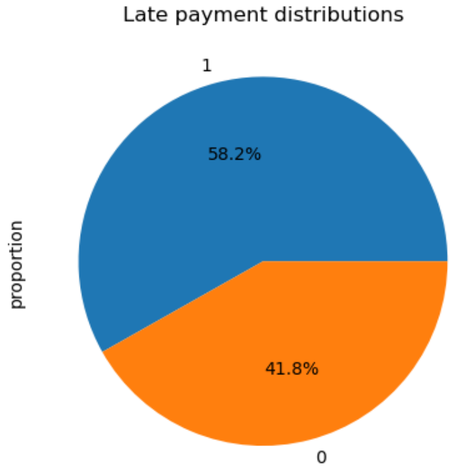
30 Days from EOM (0.114) and 60 Days from EOM (0.111): Payment terms related to the end of the month show a clear impact on timely payments.

Immediate Payment (0.041): Immediate payments play a role in mitigating delays.

Business Action: Focus on high-importance features like USD Amount, Invoice_Month, and payment terms to prioritize interventions and refine credit policies.

Model Efficiency: Leveraging top features reduces complexity while maintaining high accuracy.

Cluster 2 (Prolonged Payers) and the 58.2% delayed invoices represent areas of financial risk.
The predictions can guide targeted collection strategies, reducing overall late payment rates and improving cash flow.

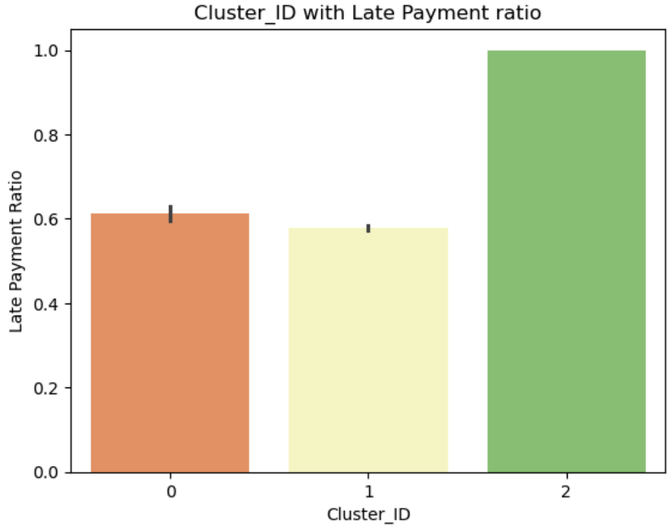


Late Payments: 58.2% of invoices are delayed, highlighting a widespread issue.
On-Time Payments: 41.8% of invoices are paid within terms.

Cluster 2 (Prolonged Payers): Nearly 95% of transactions result in late payments.

Cluster 1 (Early Payers): Lowest late payment ratio (~55%), indicating reliable customers.

Cluster 0 (Medium Payers): Moderate late payment ratio (~65%).



The Random Forest model effectively identifies Cluster 2 as the highest-risk group for late payments. Focused interventions, such as proactive reminders and stricter credit policies, are required for these customers.

CUSTOMERS WITH THE HIGHEST DELAY PROBABILITIES

Customer_Name	Delayed_Payment	Total_Payments	Delay%
AL K Corp	8	8	100.0
CUTE Corp	7	7	100.0
ALSU Corp	7	7	100.0
AMUR Corp	4	4	100.0
FINA Corp	4	4	100.0
LEAV Corp	4	4	100.0
AFRI Corp	4	4	100.0
ISMA Corp	4	4	100.0
MAYC Corp	3	3	100.0
MAJE Corp	3	3	100.0

Predictions suggest that the companies presented in the table to the left has the maximum probability of default with maximum number of delayed and total payments

Recommendations

Customer_Name	Delayed_Payment	Total_Payments	Delay%
AL K Corp	8	8	100.0
CUTE Corp	7	7	100.0
ALSU Corp	7	7	100.0
AMUR Corp	4	4	100.0
FINA Corp	4	4	100.0
LEAV Corp	4	4	100.0
AFRI Corp	4	4	100.0
ISMA Corp	4	4	100.0
MAYC Corp	3	3	100.0
MAJE Corp	3	3	100.0

1

1. Customer Payment Patterns Late Payment Ratio: Late payments comprise 65.7% of all transactions. Customers making delayed payments are more prevalent. Top Late Payers (High Priority Focus): Identified customers like YOUG Corp (100% delay), VAVA Corp (93.8%), and AL T Corp (91.7%) should be monitored closely. Actionable Recommendation: Initiate early follow-ups and stricter credit terms for customers with high late payment ratios.

2

1. Invoice Types and Classes Invoice Class Impact: Credit Notes (CM) observe the highest delay rate, while Invoices (INV) have the lowest delay rate. Action: Tighten policies on credit notes to reduce risks. Invoice Type Impact: Goods-based invoices (70% of total) show a higher delay ratio compared to non-goods types. Action: Introduce stricter credit terms for goods-based invoices.

3

1. Payment Terms Payment terms like "30 Days from Invoice Date" and "60 Days from Invoice Date" are most frequent. Late payments are prominent for these terms. Late Payment Clusters: Customers with payment terms exceeding 60 days often delay payments. Recommendation: Gradual penalties for delays in high-risk terms. Explore dynamic terms for repeat offenders.

4

1. Monthly Trends Seasonal Patterns: Late payments spike towards the end of the year (October to December) with a 99% delay rate. Actionable Insights: Introduce proactive measures for invoices issued in Q4. Increase collector focus during high-risk months.

5

1. Customer Segmentation (Clustering Analysis) Cluster Findings: Cluster 0: Medium payment duration. Cluster 1: Prolonged payment duration with high delays. Cluster 2: Early payments, low delays. Action: Focus on Cluster 1 customers for corrective actions like stricter terms or negotiations.

THANK YOU