olist OLIST E-COMMERCE

OLIST E-COMMERCE DATA CHURN ANALYSIS

Purwadhika Digital Technology School

Prepared by: **Phi**
1. Muhammad Agisni
2. Sheila Riani Putri
3. Adib Fardan Auli

olist

TABLE OF CONTENT

PHI TEAM

MUHAMMAD AGISNI

SHEILA RIANI PUTRI

ADIB FARDAN AULI

01

OLIST COMPANY

Founded: 2015, Brazil

Business Model: Marketplace integrator helping small businesses sell on major e-commerce platforms.

Key Offerings: Logistics, customer service, and data insights for sellers.

Growth: Partnered with platforms like MercadoLibre, B2W, and Magazine Luiza.

Mission: Empower small businesses with access to larger audiences and competitive tools.

BUSINESS OVERVIEW

02

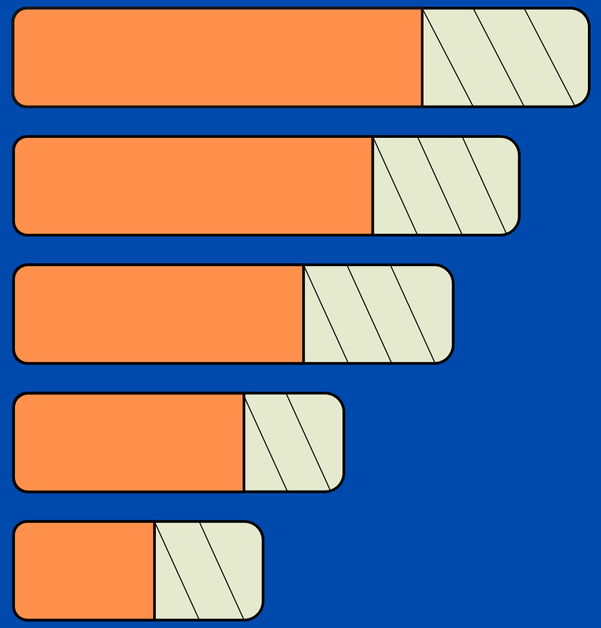
About Olist
A Brazilian e-commerce platform connecting small businesses with larger marketplaces.

Business Model
Acts as an intermediary, helping sellers list their products across multiple marketplaces.

Key Services
Order fulfillment, customer service, and logistics support.

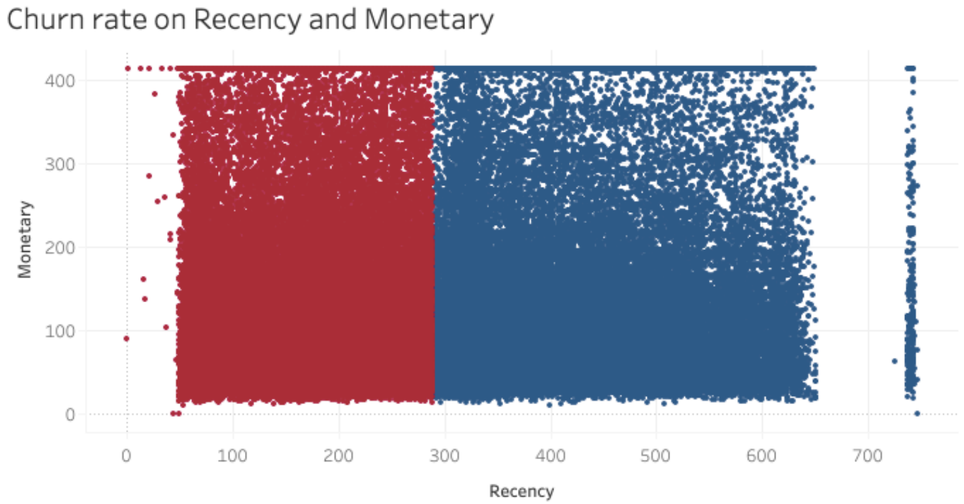
Challenges
High competition, logistics inefficiencies, and customer satisfaction management.

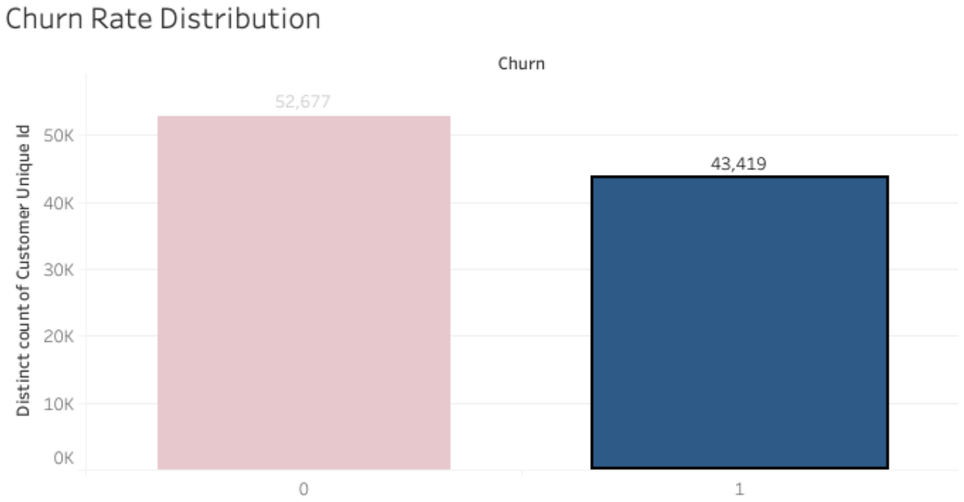
03



EXPLORATORY DATA ANALYSIS (EDA)

04 EDA-KEY FOUND





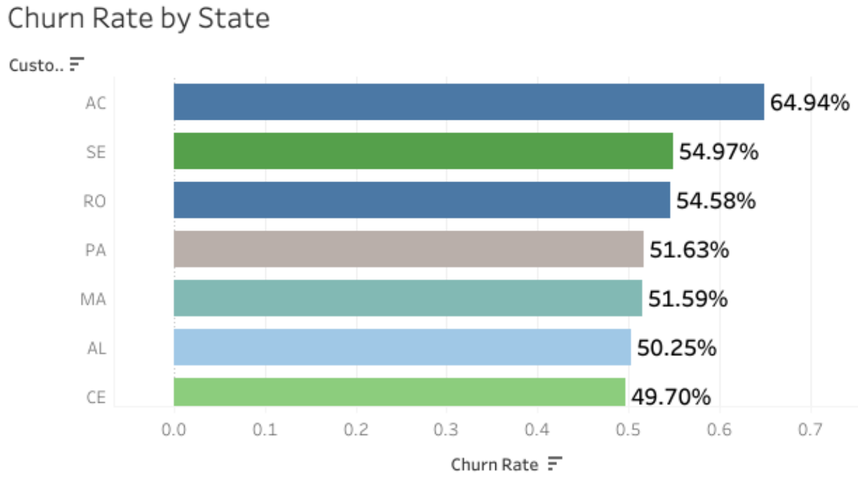
Customer Churn Insights

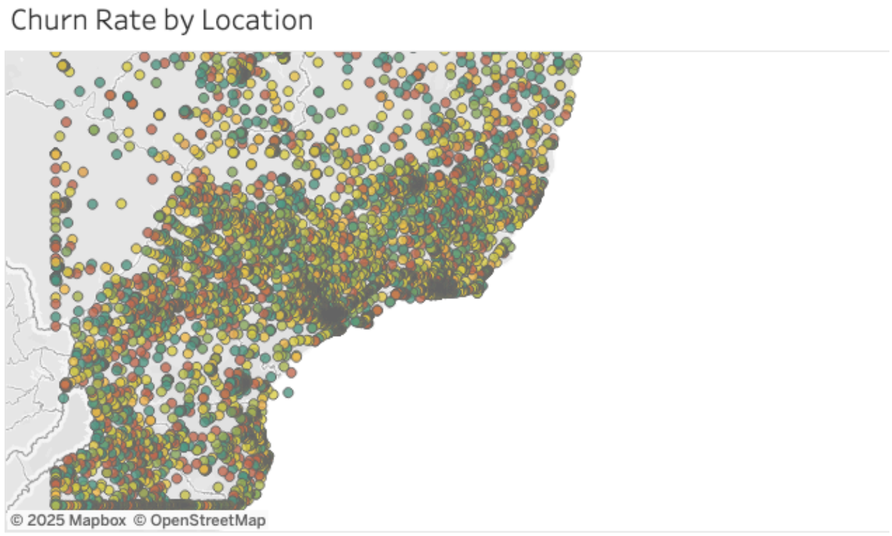
- Churned customers (red) tend to have lower recency values → Customers who haven't made purchases recently are more likely to churn.
- Active customers (blue) have higher recency values → Recent purchases indicate stronger engagement and retention.
- Churned customers show a wide range of spending → Even high-spending customers are leaving, signaling potential issues with customer satisfaction or engagement.
- Overall churn rate is ~45% → A moderately high churn rate that may require retention strategies to improve customer loyalty.

05

EDA-KEY FOUND

CHURN RATE BY STATE





Top States with Highest Churn:

- Acre (AC): Highest churn rate at 64.94%
- Sergipe (SE) & Rondônia (RO): 54.97% & 54.58%, respectively
- Other states range between 48.64% - 51.63%, with Rio Grande do Sul (RS) having the lowest churn at 48.64%

Key Observations:

- Churn rates vary significantly across states, indicating regional factors (competition, customer service, economy) might impact retention.
- States with higher churn (AC, SE, RO) need targeted interventions like loyalty programs or improved support.

Geographic Churn Distribution:

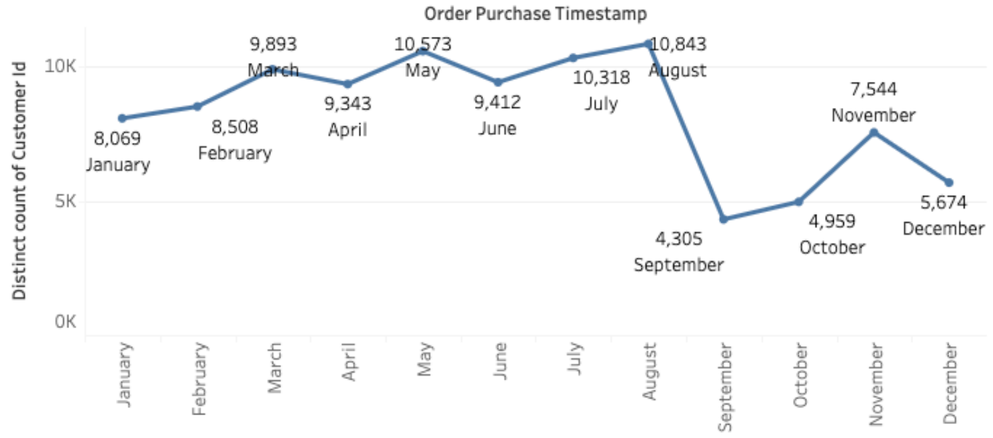
- Yellow = High Churn, Green = Low Churn
- High churn is not uniformly distributed, even in densely populated areas.
- Certain local clusters with high churn may signal service quality issues or strong market competition.

06

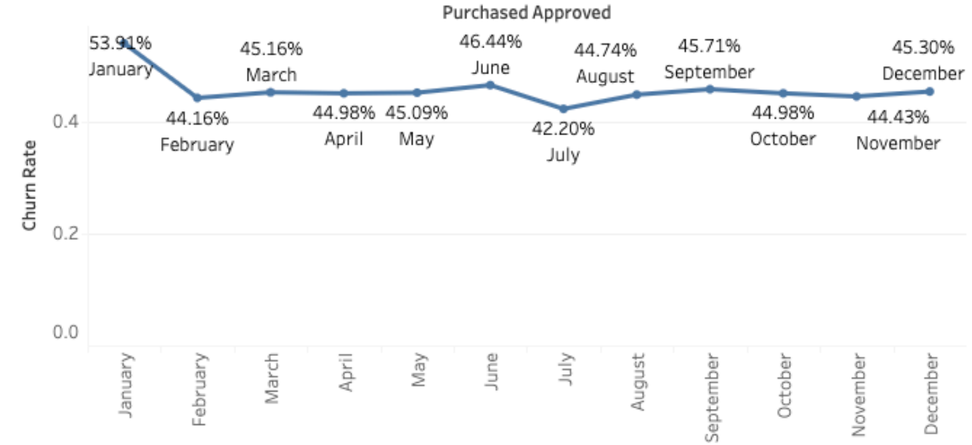
EDA-KEY FOUND

CHURN TRENDS

Customer Trend



Churn Trend



📈 Growth & Decline Patterns:

- Strong growth from January (5,069) to August (10,843).
- Sharp decline in September (4,305), partial recovery in November (7,544), but another dip in December (5,674).
- Possible causes: seasonality, retention challenges, or external factors.

📉 Churn Rate Trends:

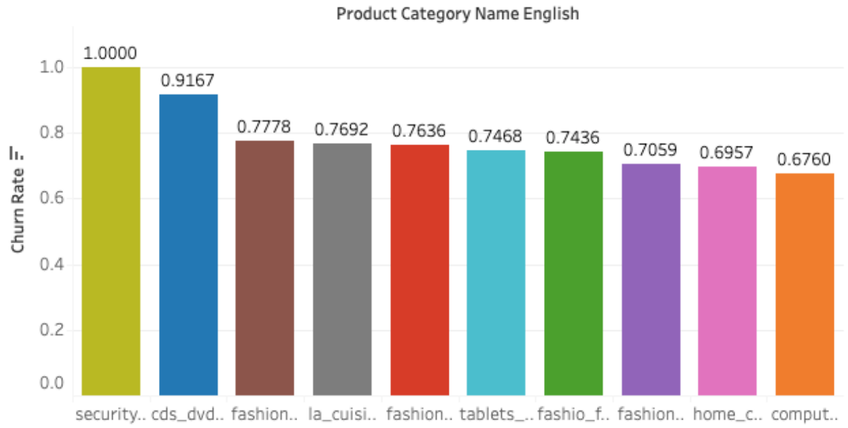
- January: Highest churn (53.91%), then drops to 44.16% in February.
- Fluctuates between 44-46%, hitting its lowest in July (42.2%).

07

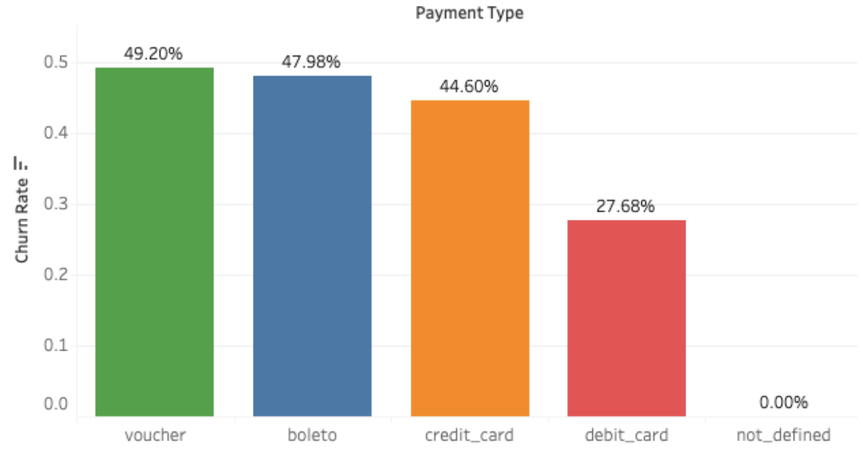
EDA-KEY FOUND

CHURN INSIGHTS BY CATEGORY & PAYMENT METHOD

Product category churn rate



Churn Rate by Payment Type



High-Churn Categories:

- Security & automation, media, and fashion have higher churn.
- May require better engagement, post-purchase support, or loyalty programs.
- Physical media (CDs, DVDs) is declining—consider digital transformation or product bundling.

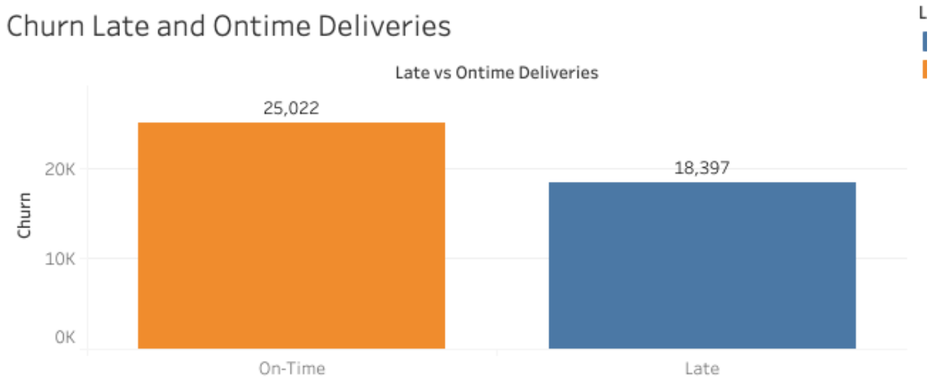
Payment Method Impact on Churn:

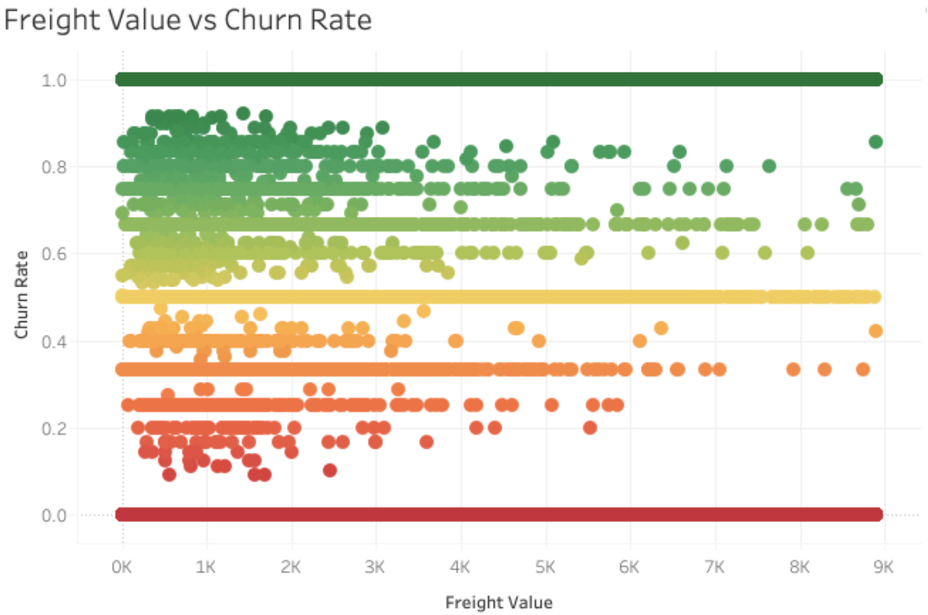
- Voucher & boleto users churn more, suggesting a need for incentives to encourage repeat purchases.
- Credit card users show stable churn, but better engagement strategies can help.
- Debit card users are more loyal, possibly preferring direct payments over deferred options.

08

EDA-KEY FOUND

CHURN INSIGHTS: DELIVERY TIMING & FREIGHT VALUE





Churn is higher for on-time deliveries (25,022) than late deliveries (18,397).

- This challenges the assumption that late deliveries cause higher churn.
- Possible reasons:
- On-time deliveries may be to low-engagement customers who churn due to poor product experience, pricing, or lack of incentives.
- Late deliveries might occur for loyal customers who continue purchasing despite delays.
- Customers expecting delays may not be as likely to churn.

Optimizing Customer Retention:

- Pricing Strategies: Mid-range freight value customers may need better incentives or service improvements.
- Customer Lifetime Value (CLV): High freight value customers who churn—are they one-time buyers or potential long-term customers?
- Retention Efforts: Mid-range freight value customers churn significantly, so better follow-up, loyalty programs, and personalized offers could improve retention.

UNDERSTANDING CUSTOMER CHURN IN OLIST E-COMMERCE TO IMPROVE RETENTION

Objective:
This research aims to analyze customer churn patterns in the Olist e-commerce platform using machine learning. By identifying key factors that influence churn, we can develop targeted retention strategies to reduce customer attrition and improve long-term business growth.
Why is this important?

- High churn rates lead to revenue loss and increased acquisition costs.
- Predicting churn allows proactive customer engagement and intervention.
- Optimizing strategies such as personalized offers, better delivery experiences, and improved customer service can enhance retention.
-

Approach:

1. Exploratory Data Analysis (EDA): Understanding trends in customer behavior.
2. Feature Engineering & Selection: Identifying key churn indicators.
3. Machine Learning Modeling: Predicting churn with high accuracy.
4. Actionable Insights: Developing business strategies based on findings.

This research will help Olist and similar e-commerce platforms enhance customer retention, optimize marketing efforts, and boost overall profitability.

10

DATA PREPARATION

Datasets Loaded:

ORDERS
CUSTOMERS

PRODUCTS
PAYMENTS

SELLERS
GEOLOCATION

REVIEWS

Data Cleaning Steps:

- Handled missing values
- Removed duplicate records
- Adjusted data types for efficiency
- Outlier handling

DATA PREPARATION

MISSING VALUE INSIGHT

11

- **ORDERS:** CONVERTS DATE COLUMNS, FILLS MISSING ORDER_DELIVERED_CARRIER_DATE USING MEDIAN PROCESSING TIME, AND IMPUTES CATEGORICAL VALUES WITH MODE.
- **REVIEWS:** FILLS MISSING SCORES WITH MEDIAN, REPLACES EMPTY COMMENTS WITH "NAO_REVEJA", AND DROPS SPARSE COLUMNS.
- **PRODUCTS:** USES MEDIAN FOR MISSING PRODUCT DETAILS AND ASSIGNS "UNKNOWN" TO MISSING CATEGORIES.
- **PAYMENTS:** FILLS MISSING PAYMENT_VALUE WITH MEDIAN.

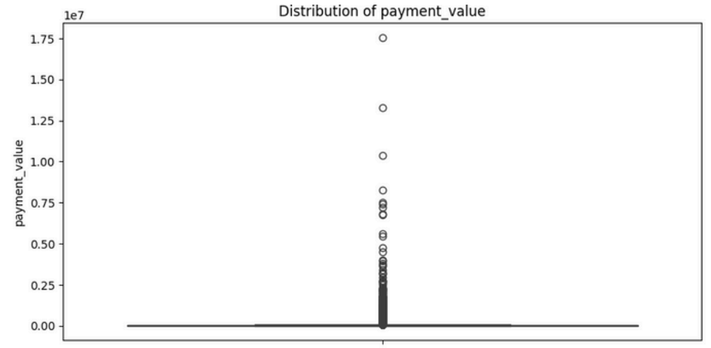
Detailed Column Summary for All Olist Datasets

dataset_name	feature	qtd_null	percent_null	dtype	qtd_cat
orders	order_id	0	0.000000	object	99441
orders	customer_id	0	0.000000	object	99441
orders	order_status	0	0.000000	object	8
orders	order_purchase_timestamp	0	0.000000	datetime64[ns]	0
orders	order_approved_at	0	0.000000	datetime64[ns]	0
orders	order_delivered_carrier_date	146	0.146821	datetime64[ns]	0
orders	order_delivered_customer_date	0	0.000000	object	95665
orders	order_estimated_delivery_date	0	0.000000	datetime64[ns]	0
customer	customer_id	0	0.000000	object	99441
customer	customer_unique_id	0	0.000000	object	96096
customer	customer_zip_code_prefix	0	0.000000	int64	0
customer	customer_city	0	0.000000	object	4119
customer	customer_state	0	0.000000	object	27
geolocation	geolocation_zip_code_prefix	0	0.000000	int64	0
geolocation	geolocation_lat	0	0.000000	float64	0
geolocation	geolocation_lng	0	0.000000	float64	0
geolocation	geolocation_city	0	0.000000	object	8011
geolocation	geolocation_state	0	0.000000	object	27
order_items	order_id	0	0.000000	object	98666
order_items	order_item_id	0	0.000000	int64	0
order_items	product_id	0	0.000000	object	32951
order_items	seller_id	0	0.000000	object	3095
order_items	shipping_limit_date	0	0.000000	object	93318
order_items	price	0	0.000000	float64	0
order_items	freight_value	0	0.000000	float64	0
payments	order_id	0	0.000000	object	99440
payments	payment_sequential	0	0.000000	int64	0
payments	payment_type	0	0.000000	object	5
payments	payment_installments	0	0.000000	int64	0
payments	payment_value	0	0.000000	float64	0
reviews	review_id	0	0.000000	object	98410
reviews	order_id	0	0.000000	object	98673
reviews	review_score	0	0.000000	int64	0
reviews	review_creation_date	0	0.000000	object	636
reviews	review_answer_timestamp	0	0.000000	object	98248
products	product_id	0	0.000000	object	32951
products	product_category_name	0	0.000000	object	74
products	product_name_lenght	0	0.000000	float64	0
products	product_description_lenght	0	0.000000	float64	0
products	product_photos_qty	0	0.000000	float64	0
products	product_weight_g	0	0.000000	float64	0
products	product_length_cm	0	0.000000	float64	0
products	product_height_cm	0	0.000000	float64	0
products	product_width_cm	0	0.000000	float64	0
sellers	seller_id	0	0.000000	object	3095
sellers	seller_zip_code_prefix	0	0.000000	int64	0
sellers	seller_city	0	0.000000	object	611
sellers	seller_state	0	0.000000	object	23
product_category_translation	product_category_name	0	0.000000	object	71
product_category_translation	product_category_name_english	0	0.000000	object	71

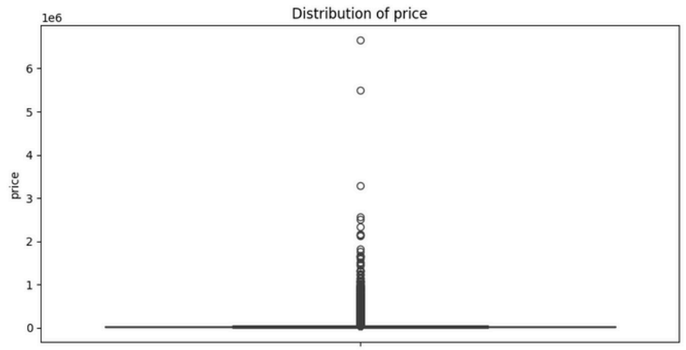
DATA PREPARATION

OUTLIER INSIGHT

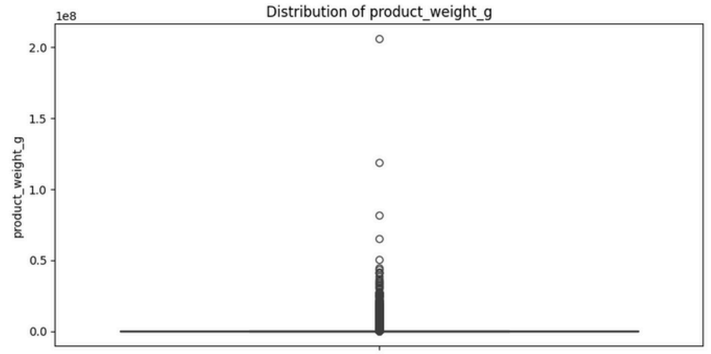
12



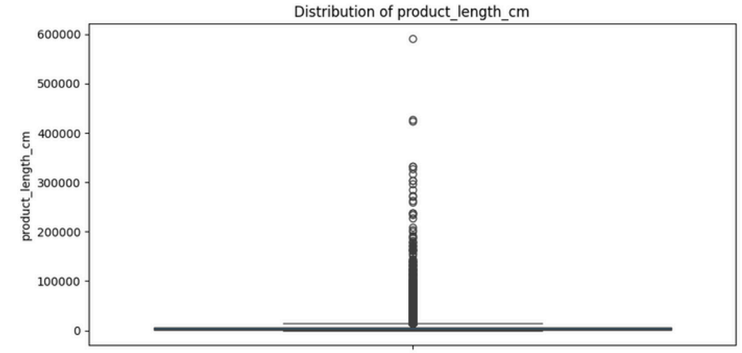
Skewness of review_score: -1.3613795801860509
Standard deviation of review_score: 1.3419908820998652



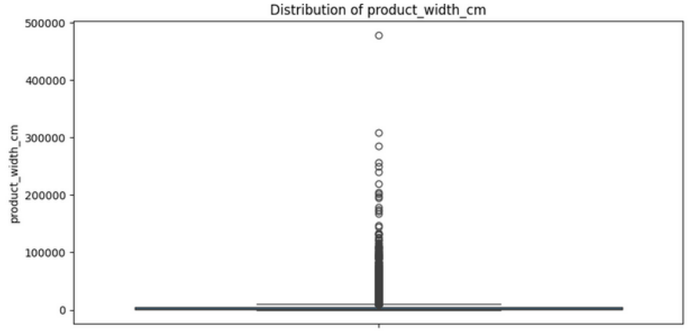
Skewness of freight_value: 23.26153211145121
Standard deviation of freight_value: 7026.622946840956



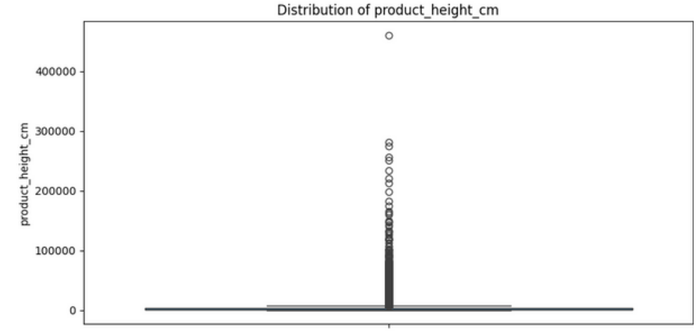
Skewness of product_length_cm: 11.965820487175028
Standard deviation of product_length_cm: 9949.452819046066



Skewness of product_height_cm: 14.677177426249566
Standard deviation of product_height_cm: 6332.266454117043



Skewness of geolocation_lat: 1.631595270481175
Standard deviation of geolocation_lat: 5.621435729526562



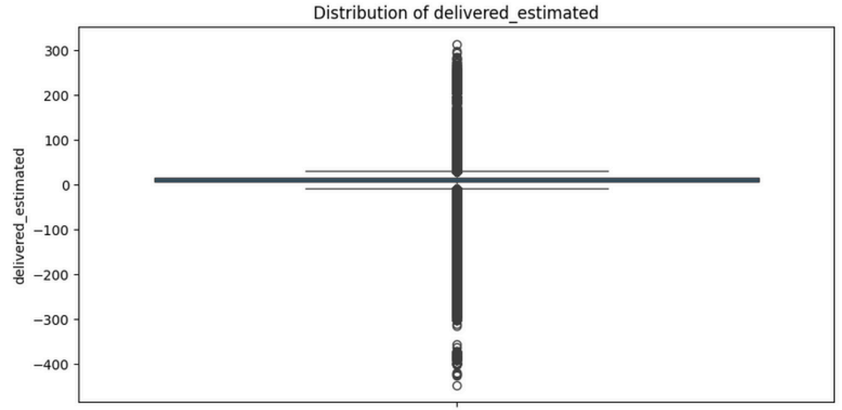
Skewness of product_width_cm: 11.59223186824621
Standard deviation of product_width_cm: 7200.956758063494

PRICE, PAYMENT VALUE, AND PRODUCT DIMENSIONS HAVE EXTREME OUTLIERS, WHICH NEED TRANSFORMATION OR REMOVAL.

DATA PREPARATION

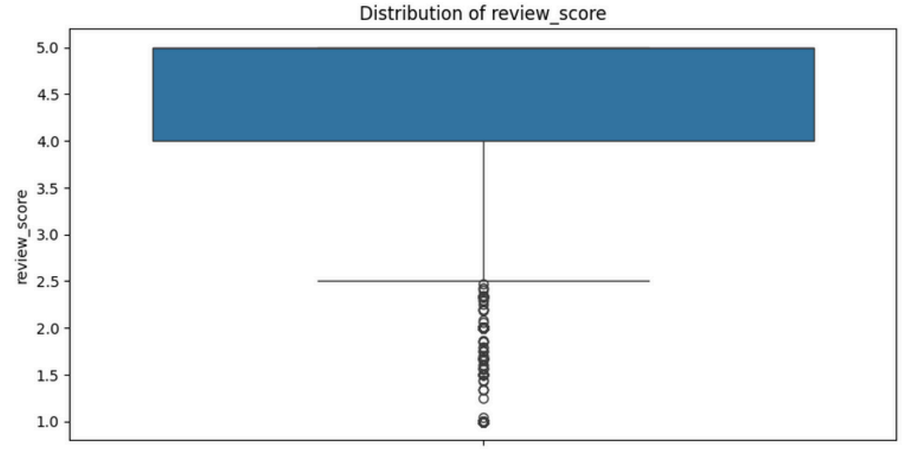
OUTLIER INSIGHT

13



Skewness of purchased_delivered: 3.6795837296810427
Standard deviation of purchased_delivered: 30.72348800550564

Skewness of product_id: 6.600163817912844
Standard deviation of product_id: 0.3280884722980742



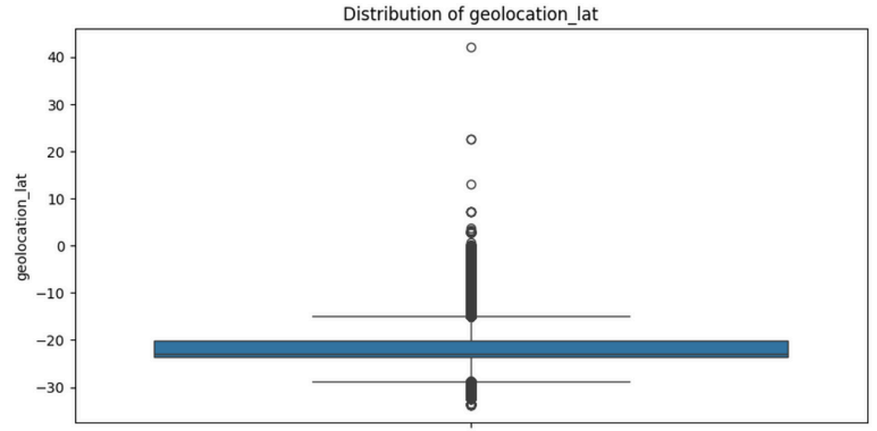
Skewness of Recency: 0.44545210864069507
Standard deviation of Recency: 153.41786878790546

DELIVERY TIME AND REVIEW SCORES ARE NEGATIVELY SKEWED, SHOWING EARLY DELIVERIES AND GENERALLY HIGH RATINGS.

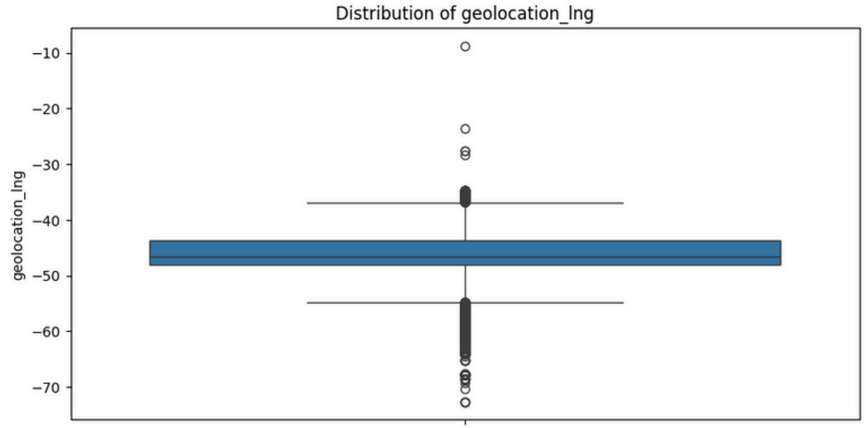
DATA PREPARATION

OUTLIER INSIGHT

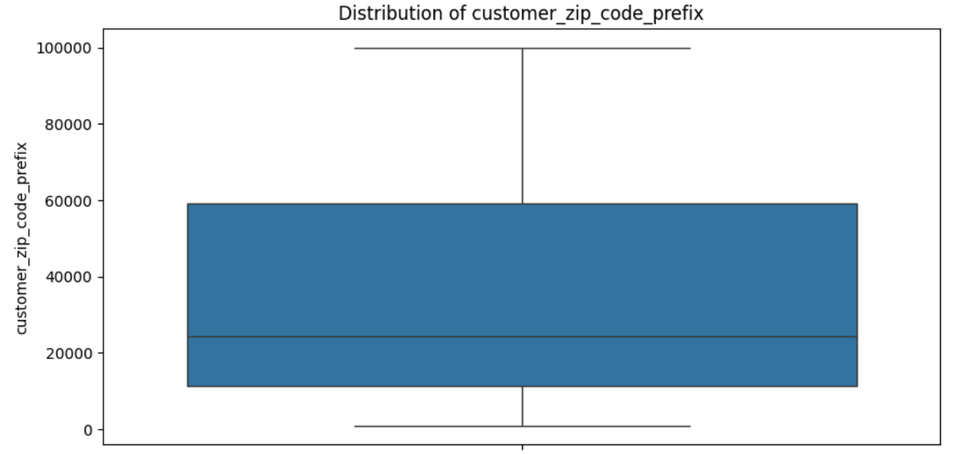
14



Skewness of geolocation_lng: 0.017939844544967415
Standard deviation of geolocation_lng: 4.0588322788673565



Skewness of payment_installments: 1.594304437661753
Standard deviation of payment_installments: 2.724316155626932



Skewness of order_id: 11.862593280003077
Standard deviation of order_id: 0.2143837547943183

GEOLOCATION AND CUSTOMER ZIP CODES ARE MILDLY SKEWED, INDICATING REGIONAL CONCENTRATION OF CUSTOMERS.

olist

15

MACHINE LEARNING APPROACH

Objective
Predicting customer churn

Models Used

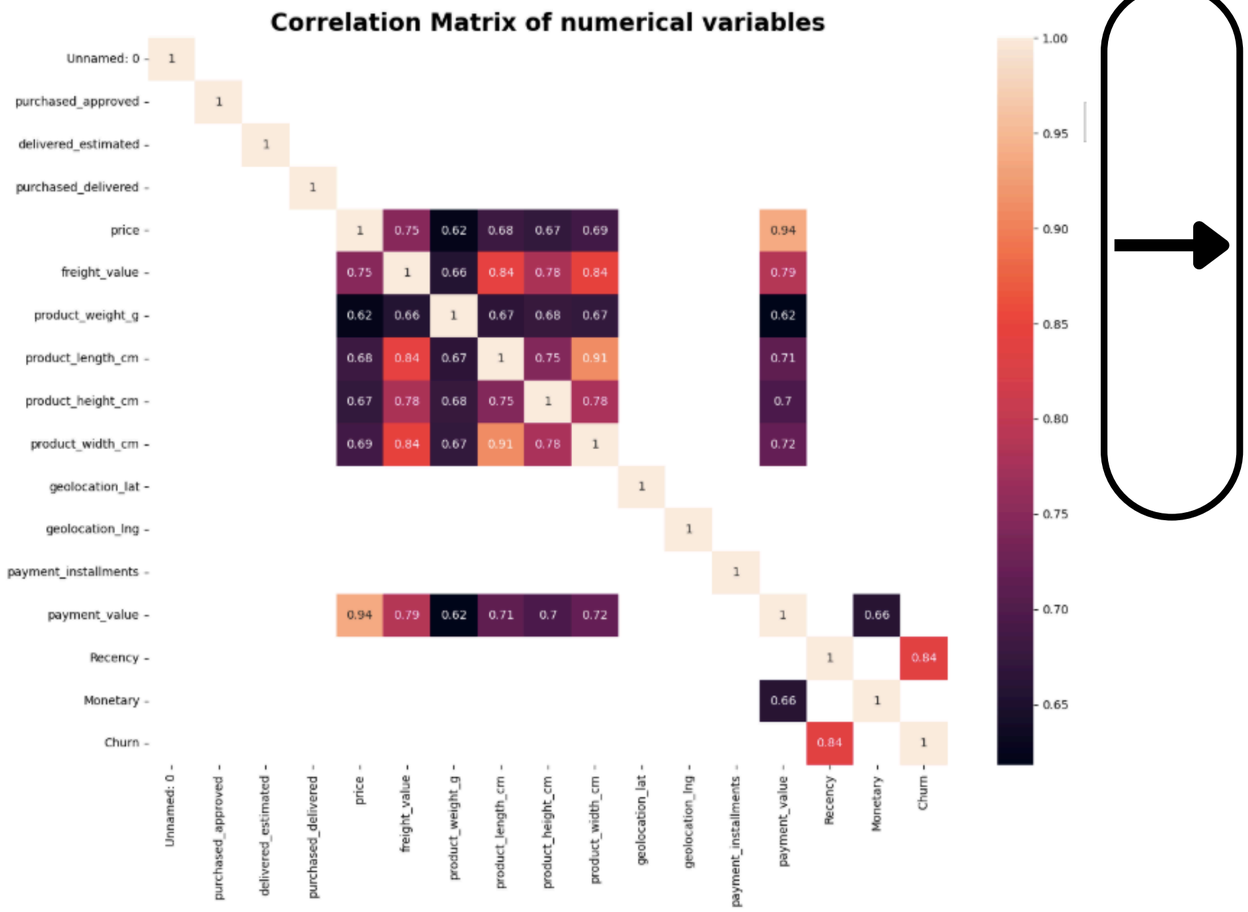
- Logistic Regression
- Random Forest
- Gradient Boosting
- XGBoost
- CatBoost
- LightGBM

16

MACHINE LEARNING
CORRELATION MATRIX

High Correlation Between Some Variables:

- **Price & Freight Value (0.75):** More expensive products generally have higher shipping costs.
- **Product Dimensions & Freight Value (0.84):** Larger products tend to have higher shipping fees.
- **Payment Value & Installments (0.94):** Customers paying in installments typically have higher total payments, indicating redundancy.
- **Monetary & Recency (0.84):** High-spending customers tend to make more recent purchases.



17

MACHINE LEARNING
STATISTICAL RESULT

Key Findings:

- **Customer State & Churn:** Significant dependency, indicating churn rates vary by region.
- **Payment Type & Churn:** Payment methods influence churn behavior.
- **Payment Installments & Churn:** Higher installment payments correlate with churn.
- **Customer City & Churn:** Regional differences impact churn rates.
- **Freight Value, Product Dimensions & Churn:** Shipping costs and product size influence churn.
- **Recency & Monetary Value & Churn:** Recent, high-spending customers show different churn patterns.
- **Price & Churn:** No significant correlation, suggesting pricing alone does not drive churn.

Variable	Test Statistic	p-value	Degrees of Freedom	Conclusion
Customer State & Churn	330.82	1.37e-54	26	✅ Dependent (Churn varies by region)
Payment Type & Churn	283.45	4.03e-60	4	✅ Dependent (Payment method influences churn)
Payment Installments & Churn	550.35	9.34e-113	9	✅ Dependent (Installment payments affect churn)
Customer City & Churn	4637.44	1.83e-08	4118	✅ Dependent (Regional differences impact churn)
Purchased Approved & Churn	-	-	-	✅ Correlated
Purchased Delivered & Churn	-	-	-	✅ Correlated
Freight Value & Churn	-	-	-	✅ Correlated (Shipping cost affects churn)
Product Weight & Churn	-	-	-	✅ Correlated
Product Dimensions (Length, Width, Height) & Churn	-	-	-	✅ Correlated (Larger products impact churn)
Geolocation (Latitude & Longitude) & Churn	-	-	-	✅ Correlated (Location affects churn)
Recency & Churn	-	-	-	✅ Correlated (Recent transactions impact churn)
Monetary Value & Churn	-	-	-	✅ Correlated (Spending behavior affects churn)
Delivered Estimated & Churn	-	-	-	❌ Not Correlated (Delivery estimate not a factor)
Price & Churn	-	-	-	❌ Not Correlated (Pricing alone doesn't affect churn)

18

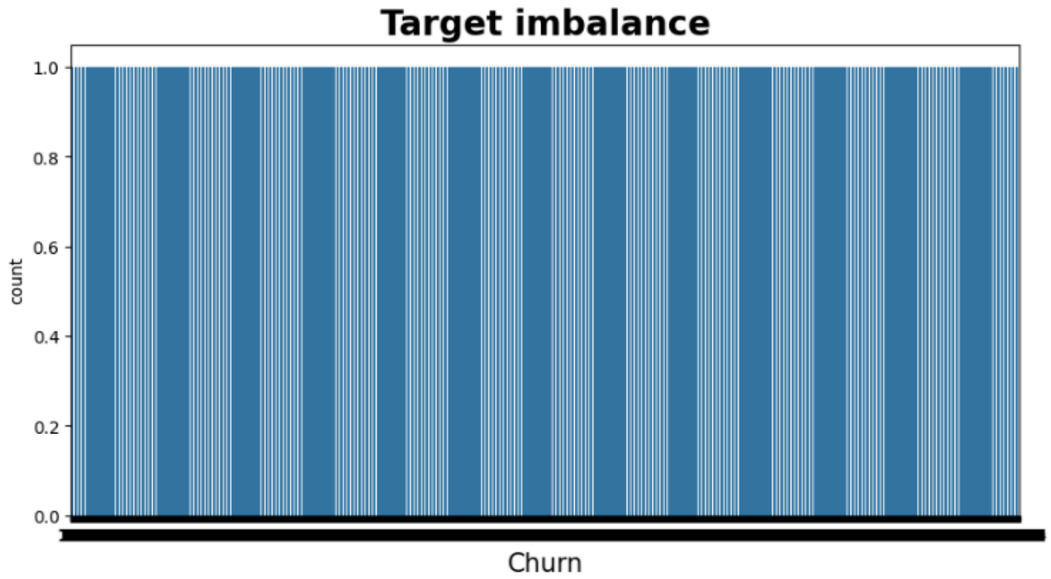
MACHINE LEARNING
CHECKING IMBALANCE

Checking Imbalance

```
final_outlierTreated.Churn.value_counts(normalize = True) * 100
```

Churn	proportion
0	54.817058
1	45.182942

dtype: float64



The churn variable is nearly balanced, with 54.82% non-churned customers (0) and 45.18% churned customers (1). The distribution suggests that no severe class imbalance exists

19

MODEL EVALUATION

Evaluation Metrics:

- Accuracy
- Precision
- Recall
- F1-score

Feature Importance:

Delivery time significantly impacts review scores.

20

MACHINE LEARNING MODEL APPROACH

	test_accuracy	train_accuracy	test_precision	train_precision	test_recall	train_recall	test_kappa	train_kappa	f1_score	roc_auc_score
CatBoost	0.755671	0.757128	0.740205	0.740590	0.710828	0.710988	0.505429	0.507879	0.725219	0.839232
XGBoost	0.678668	0.767704	0.652033	0.751801	0.625258	0.724590	0.349482	0.529465	0.638365	0.674132
LightGBM	0.682882	0.731828	0.659530	0.712964	0.621932	0.679433	0.357140	0.456272	0.640179	0.757848
RandomForest Classifier	0.660822	1.000000	0.649612	1.000000	0.547603	1.000000	0.306499	1.000000	0.594262	0.651206
Logit_FullModel	0.580853	0.586867	0.557494	0.558840	0.400296	0.394567	0.136076	0.142913	0.465995	0.627612

Best Model for Churn Prediction

- CatBoost performs best with 75.57% accuracy and 83.92% ROC-AUC, balancing precision (74.02%) and recall (71.08%) to minimize false positives while capturing actual churners.
- XGBoost & LightGBM show good recall (>61%) but less consistency in predictions (lower kappa scores).
- Random Forest overfits – perfect training accuracy but test accuracy drops to 66.48%.
- Logistic Regression underperforms → low recall (40.03%) & kappa (13.60%), making it ineffective for churn prediction.

Conclusion

- CatBoost is the most reliable model due to its balance of accuracy, recall, and ROC-AUC.
- XGBoost & LightGBM can be alternatives but have lower consistency.
- Random Forest needs tuning to reduce overfitting.
- Logistic Regression is not suitable for churn prediction due to its low recall and accuracy.

21

HYPERPARAMETER TUNING

```python
# Hyperparameter tuning
cat = CatBoostClassifier()
parameters = {
    'iterations': [1000, 1200, 1500],
    'depth': [6, 8, 10],
    'learning_rate': [0.01, 0.1, 0.2]
}

gcv_cat = GridSearchCV(estimator=cat, param_grid=parameters, cv=5, n_jobs=-1)
gcv_cat.fit(xtrain_cat, ytrain_cat)
best_params_cat = gcv_cat.best_params_
print("Best Parameters:", best_params_cat)
```

```
Best Parameters: {'depth': 6, 'iterations': 1000, 'learning_rate': 0.1}
```

	test_accuracy	train_accuracy	test_precision	train_precision	test_recall	train_recall	test_kappa	train_kappa	f1_score	roc_auc_score
CatBoost	0.755671	0.757128	0.740205	0.740590	0.710828	0.710988	0.505429	0.507879	0.725219	0.839232

Best Model for Churn Classification

- CatBoost model performed the best with strong accuracy (0.7557) and AUC-ROC (0.8392).

Key metrics:

- Precision (0.7402) → Ensures correct churn predictions, reducing false alarms.
- Recall (0.7108) → Helps capture actual churners, allowing proactive retention.
- F1-score (0.7252) → Balances precision and recall for reliable predictions.

21

MODEL RESULTS

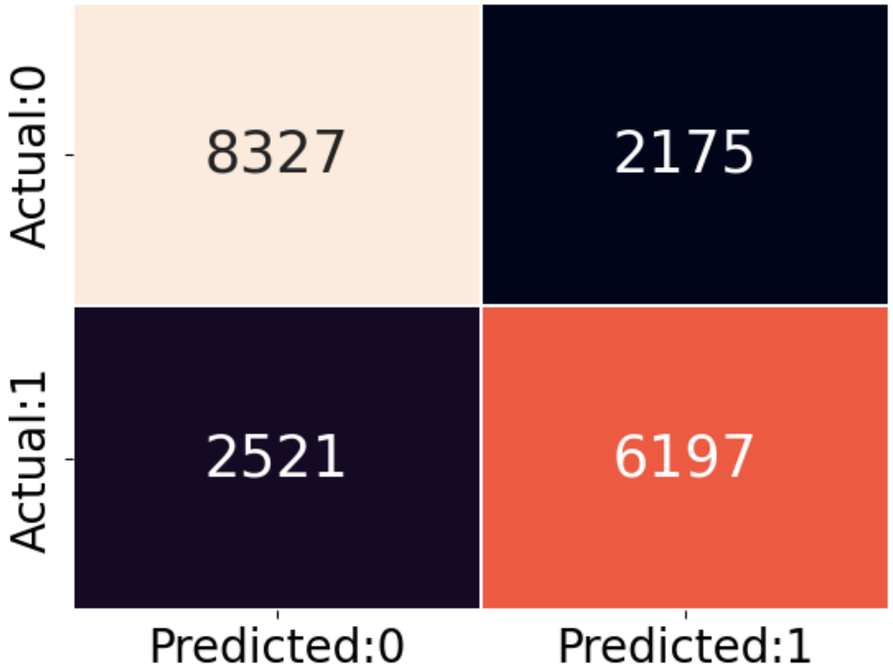
Evaluation Metrics: Recall

RFM Segmentation

Feature Importance:
Delivery time significantly impacts review scores.

22

MODEL RESULTS



If Focused on Recall → Capture as Many Churners as Possible

Suitable for businesses aiming to minimize customer loss as much as possible.

Strategy & Consequences:

✅ Detects almost all customers who are truly going to churn (low False Negatives, FN).

✅ Ideal if losing customers has a significant impact on revenue.

❌ More False Positives (FP) → Customers who wouldn't actually churn are predicted to churn → May lead to unnecessary retention costs (discounts, promotions), requiring a substantial budget.

23

RFM SEGMENTATION
FOR CHURN ANALYSIS

RFM segmentation ranks customers based on:

- Recency (R): How recently a customer made a purchase
- Frequency (F): How often they purchase
- Monetary (M): How much they spend

Key Insight:

- Recency is more critical than Monetary value in predicting churn.
- Churn is measured based on Recency.

Customer Segments Identified:

Customer Segmentation Based on RFM (Recency & Monetary)

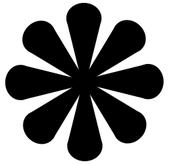
Rank (R, M)	Segment	Customer Percentage	Interpretation
(1,1)	Best Customers	6.27%	Very active customers who spend a lot of money.
(1,2)	High-Value & Recent Customers	6.40%	Recently purchased and have high spending.
(2,1)	High-Value & Recent Customers	6.14%	Still quite active and have high spending.
(2,2)	Engaged Customers	6.27%	Recently purchased with moderate spending.
(1,3)	Potential Loyalists	6.00%	Very recent customers but with lower spending.
(3,1)	At-Risk High-Value Customers	6.12%	Haven't purchased in a while, but previously spent a lot.
(2,3)	At-Risk Customers	5.95%	Fairly recent, but with low spending.
(3,3)	Lost Customers	6.25%	Not recent and low spending.
(4,1)	Lost High-Value Customers	6.05%	Haven't purchased in a long time but used to spend a lot.
(4,3) & (4,4)	Churned Customers	6.30% & 6.25%	Haven't purchased for a long time and have low spending, highly at risk of churn!

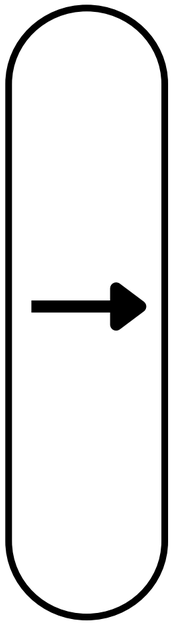
24

RECOMMENDATIONS

25

RECOMMENDATION





Understanding RFM Segmentation Data

From the ranking, we can categorize customers into two major groups:

1. High-value and potential customers (R1-M1, R1-M2, R2-M1).
2. Customers at risk of churn (R3-M4, R4-M4, R4-M3).

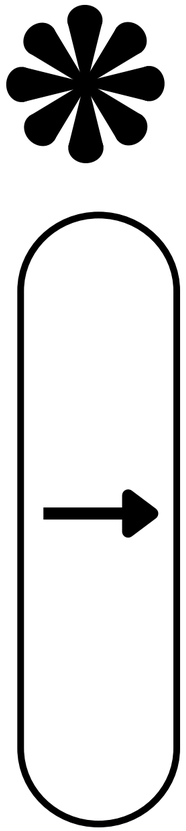
Connecting RFM with Feature Importance:

- Purchased Approved & Delivered: Measures customer satisfaction with transactions and delivery.
- Delivered Estimated & Freight Value: Shipping time and cost impact customer retention.
- Product Length CM: Potentially linked to product category and purchase experience.
- Payment Type & Installments: Payment method and installment options influence customer behavior.

26

HIGH-VALUE & LOYAL CUSTOMERS

(R1-M1, R1-M2, R2-M1, R2-M2)



Characteristics:

- Recently made purchases and spent a significant amount of money.
- Highly valuable customers who should be retained.

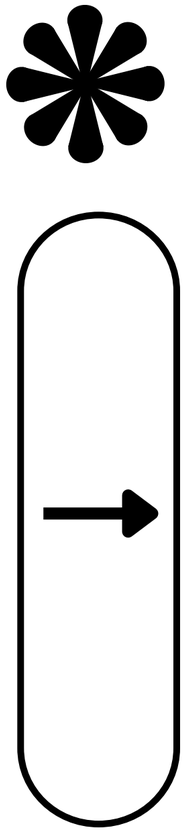
Recommended Retention Strategies:

- Loyalty Programs & Exclusive Offers
  - Provide exclusive discounts and early access to promotions.
  - Offer priority customer support to enhance the shopping experience.
- Personalized Product Recommendations
  - Use AI-based recommendations based on frequently purchased products.
  - Send targeted emails and push notifications with relevant product offers.
- Optimize Delivery and Checkout Experience
  - Ensure accurate delivery time estimates to prevent dissatisfaction.
  - Offer free shipping or fast delivery upgrades for loyal customers.

27

POTENTIAL CUSTOMERS WHO CAN MOVE UP

(R1-M3, R2-M3, R3-M1)



Characteristics:

- Recently made purchases but have not spent a significant amount yet.
- Have potential to become loyal customers with the right incentives.

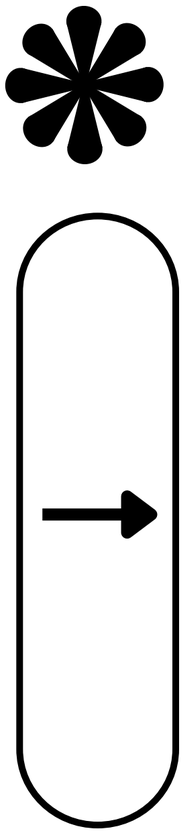
Recommended Retention Strategies:

- Upsell & Cross-Sell Strategies
  - Provide bundle deals or discounts on their next purchase to encourage spending.
- Flexible Payment Options
  - Promote installment plans or cashback offers to increase spending.
  - Align with feature importance results related to payment type & installments.

28

CUSTOMERS AT RISK OF CHURN

(R3-M3, R3-M4, R4-M2)



Characteristics:

- Have not made a purchase in a long time but have a moderate spending history.
- If no action is taken, they are likely to churn.

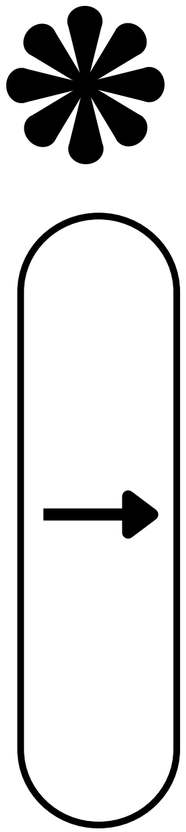
Recommended Retention Strategies:

- Re-Engagement Campaigns
  - Send "We Miss You" emails with exclusive discounts to encourage return purchases.
  - Use push notifications to remind them of past interest in products.
- Improve Purchase & Delivery Satisfaction
  - Analyze past product reviews and delivery experiences to identify potential dissatisfaction.
  - Provide free shipping vouchers or cashback if they had issues with previous orders.

29

HIGH-RISK CHURN CUSTOMERS

(R4-M4, R4-M3, R4-M2)



Characteristics:

- Have not made a purchase in a long time and have low spending levels.
- Less likely to return, but there is still a chance for reactivation.

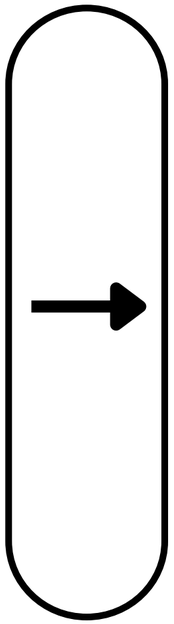
Recommended Retention Strategies:

- Segmented Targeted Discounts
  - If they have purchased a specific product category before, offer special discounts for similar products.
- Positive Shopping Experience Reminders
  - Send customer testimonials or highlight top-rated products they may be interested in.
- Last-Chance Campaigns
  - Inform them about major promotions that are ending soon to create urgency.

30

FUTURE WORK AND FURTHER RESEARCH

If the churn rate has significantly decreased or if the available budget is limited, the next focus should be on handling customers who only are highly likely to churn. Instead of targeting all churn segment, efforts should shift toward improving precision in churn prediction.



Future Research Recommendations:

- Precision-Focused Churn Prediction
  - Shift from a broad recall-based approach to a more targeted precision-based approach.
  - Prioritize interventions for customers who have the highest probability of churning.
- Sentiment Analysis for Personalized Retention
  - Conduct text analysis on customer reviews and feedback to better understand their needs and pain points.
  - Use natural language processing (NLP) to classify customer sentiments and integrate them into churn prediction models.
  - Provide tailored retention strategies based on sentiment-driven insights.

OLIST E-COMMERCE

Purwadhika
Digital Technology School

THANK YOU

DO YOU HAVE ANY QUESTION?