



ASHIRVADA

Final Project



APP

Dashboard

Notebook

Zarah Sabrina Larasati

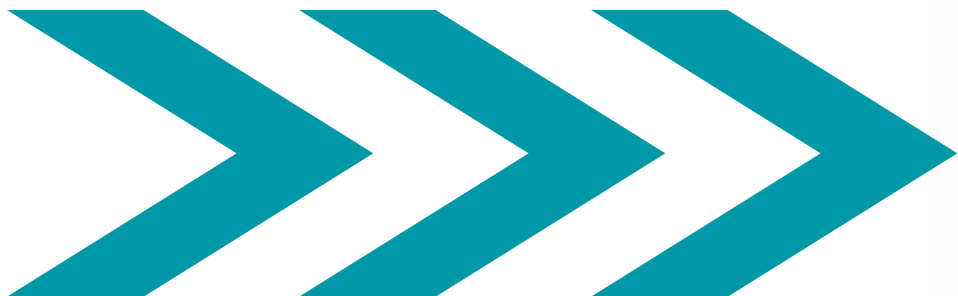


zarahsabrina@gmail.com

Presented By

Mentor : Muhammad **Hanif** Fajari

OUTLINE



Introduction

Problem Statement, Goals, and Objective

Understanding Dataset

Data Pre-Processing

Exploratory Data Analysis (EDA)

Data Integration

Feature Selection

Modeling

Explainability and Error Analysis

Model Evaluation

Impact

Business Recommendation

Conclusion



INTRODUCTION

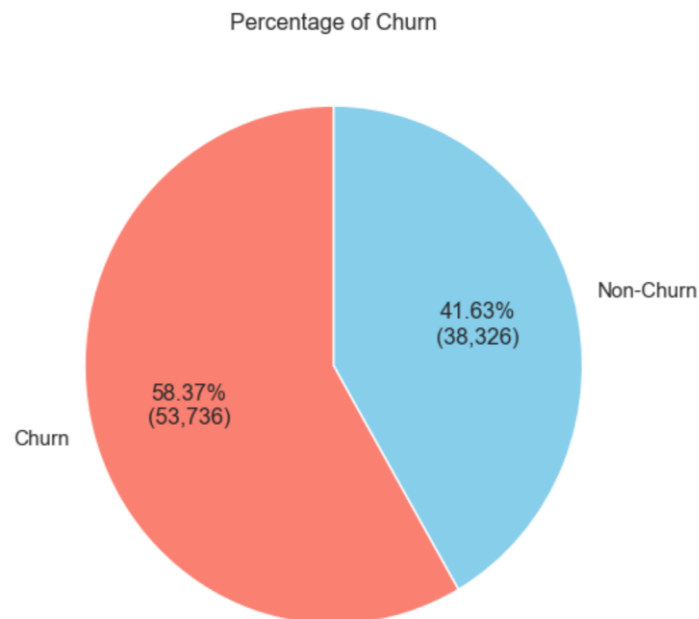
INTRODUCTION

In today's competitive e-commerce market, especially in India, businesses are facing a serious challenge : customers are not coming back. Many users make only one purchase and never return, making it difficult to build long-term customer relationships. The high churn rate leads to lost revenue opportunities. As a result, improving customer retention has become a critical business priority.

PROBLEM STATEMENT

Customer retention is a major challenge for ASHIRVADA e-commerce. This challenge is reflected in the dataset used for this project, which shows a **high churn rate of 58.37%** from 92,062 customers indicating low loyalty and weak retention. This project defines churn as no purchases in the past six months to better identify truly inactive users. Early detection is key to sustaining profitability. Machine learning techniques offer a scalable way to anticipate churn and enable proactive engagement through data-driven strategies.

It is also important to **consider the high Customer Acquisition Cost (CAC)**, while customers only stay for a short time due to churn. **Retaining customers longer becomes more cost-effective**, and retention strategies that reduce churn and increase Customer Lifetime Value (CLV) can enhance profitability. Therefore, understanding and reducing churn through efficient retention strategies is a key step in maximizing long-term profits.





GOALS

To reduce customer churn and increase customer loyalty through prediction model and personalized marketing strategies by identifying high-risk customers and enabling timely retention actions.

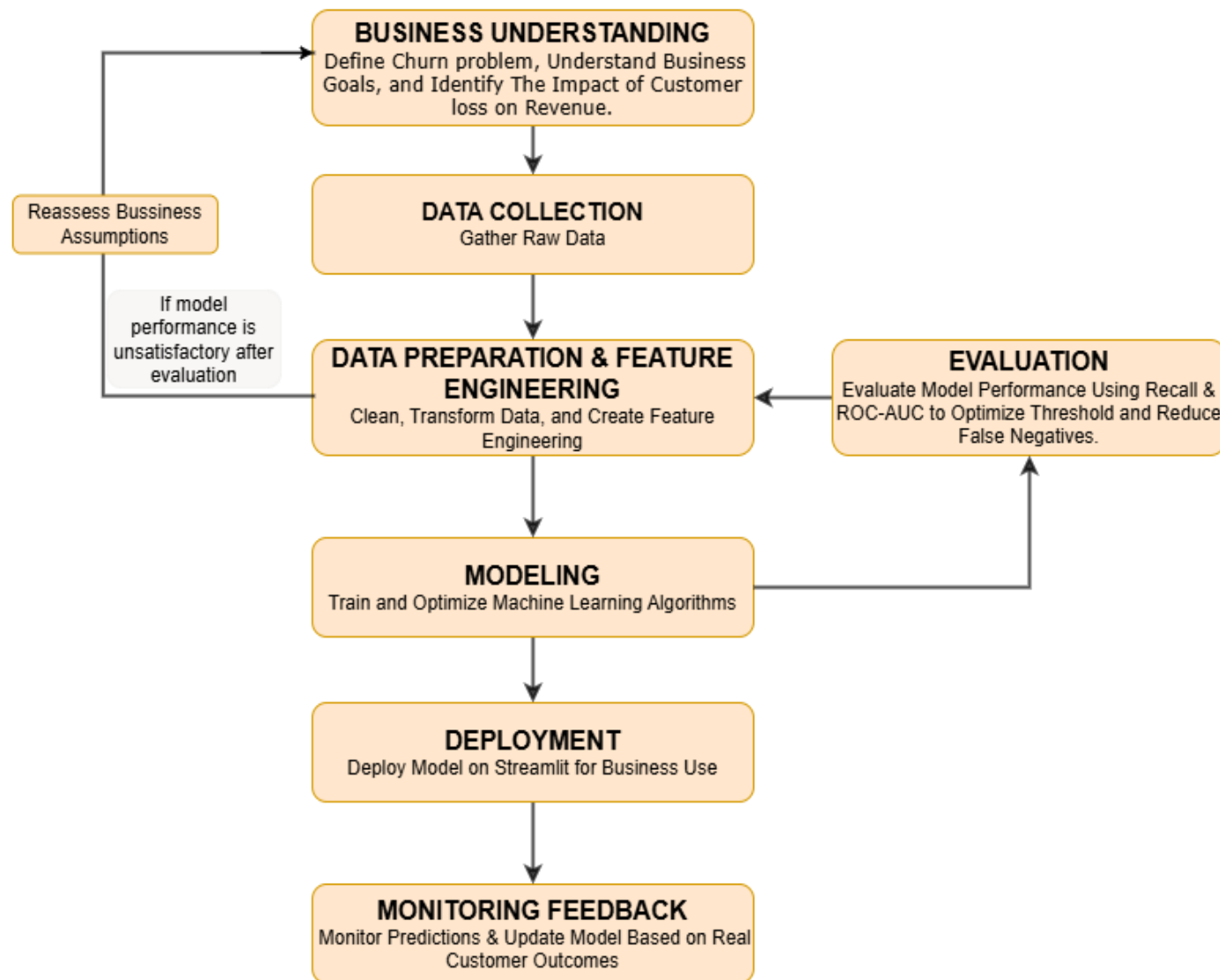
OBJECTIVES

- ✓ Analyze Historical Customer Behavior
- ✓ Identify Key Features for Churn Prediction
- ✓ Build and Train a Churn Prediction Model
- ✓ Evaluate model effectiveness using key performance metrics, including Recall and ROC-AUC, with a baseline target of 70%.
- ✓ Segment customers based on their churn risk scores for targeted retention strategies.
- ✓ Support the marketing team with monthly updates and strategy tuning based on model predictions and real churn behavior.



MACHINE LEARNING CYCLE

This flowchart illustrates the **end-to-end process of predicting customer churn** using machine learning.

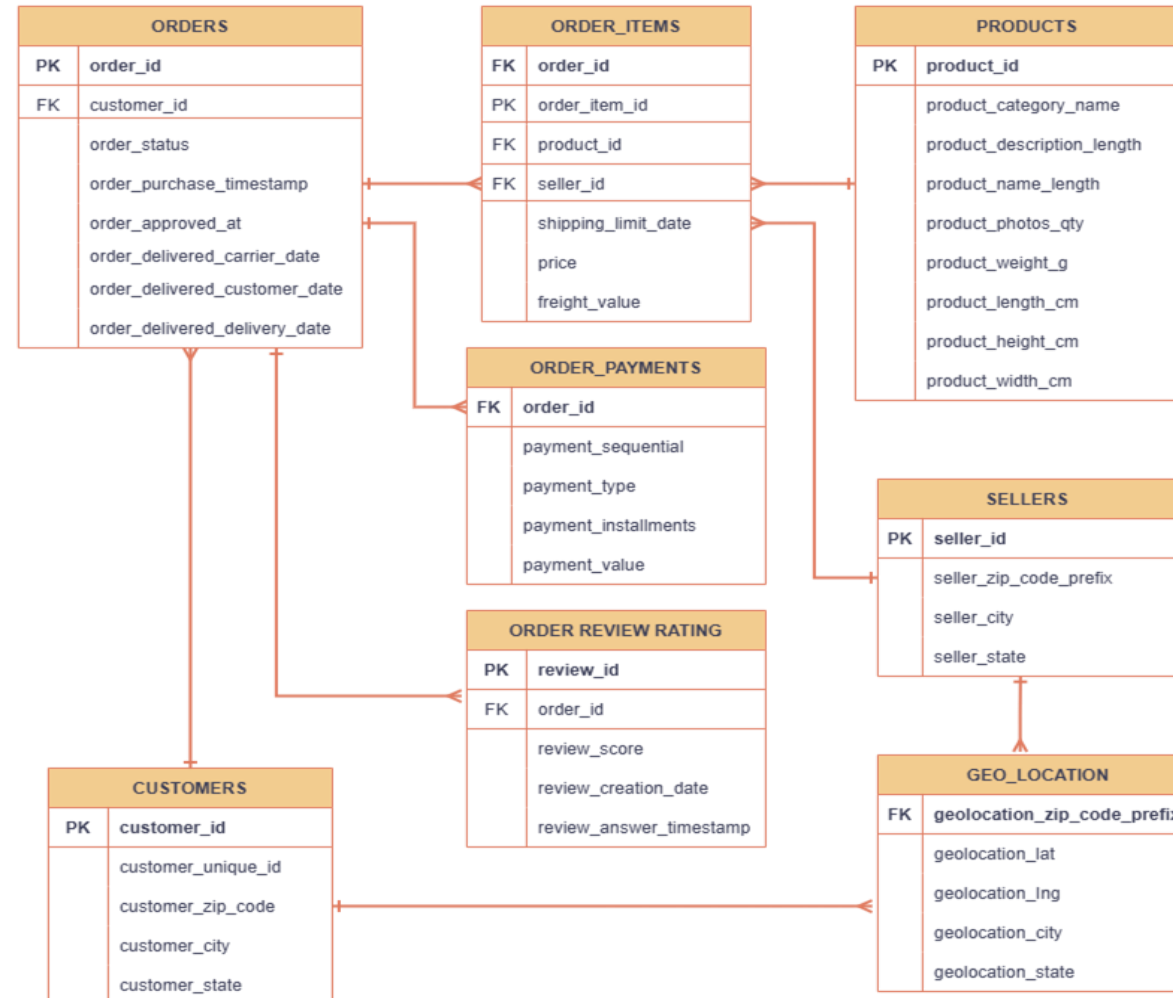




UNDERSTANDING DATASET

ENTITY RELATIONSHIP DIAGRAM

The diagram illustrates the **relationships between e-commerce tables** such as orders, order items, customers, products, and sellers, connected through primary and foreign keys. Each table holds specific data, including order details, product information, payments, locations, and customer reviews.





DATA PRE PROCESSING

HANDLING MISSING VALUE

ORDERS

	Missing Values	Percentage (%)
order_delivered_customer_date	2965	2.981668
order_delivered_carrier_date	1783	1.793023
order_approved_at	160	0.160899
order_id	0	0.000000
customer_id	0	0.000000
order_status	0	0.000000
order_purchase_timestamp	0	0.000000
order_estimated_delivery_date	0	0.000000

PRODUCTS

	Missing Values	Percentage (%)
product_category_name	623	1.890686
product_name_length	610	1.851234
product_description_length	610	1.851234
product_photos_qty	610	1.851234
product_weight_g	2	0.006070
product_length_cm	2	0.006070
product_height_cm	2	0.006070
product_width_cm	2	0.006070
product_id	0	0.000000

SELLERS

	Missing Values	Percentage (%)
seller_city	57	1.84168
seller_state	57	1.84168
seller_id	0	0.00000
seller_zip_code_prefix	0	0.00000

Remove missing values because missing values with a **very small percentage (< 5%)** will not affect trends or patterns in the data. This aligns with Little, R.J.A. & Rubin, D.B. (2002) in Statistical Analysis with Missing Data, which states that if missing data is minimal and random, its removal will not significantly impact the analysis.



HANDLING DATA TYPES

Converted to String

- Customer zip code prefix
- Geolocation zip code prefix
- Sellers zip code prefix

Converted to Datetimes

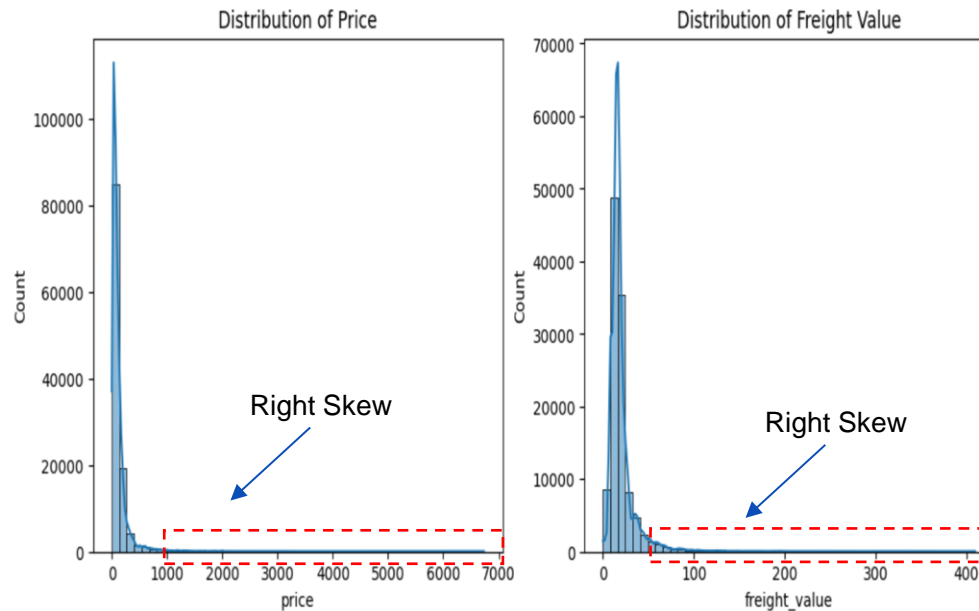
- Shipping limit date
- Review creation date
- Review answer timestamp
- Order purchase timestamp
- Order approved at
- Order delivered carrier date
- Order delivered customer date
- Order estimated delivery date



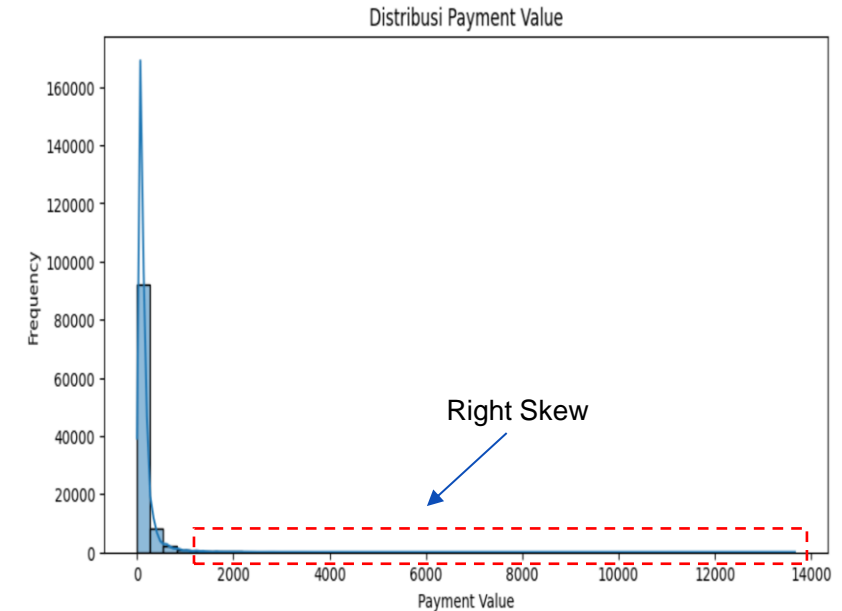
EXPLORATORY DATA ANALYSIS (EDA)

UNIVARIATE ANALYSIS

ORDER_ITEMS



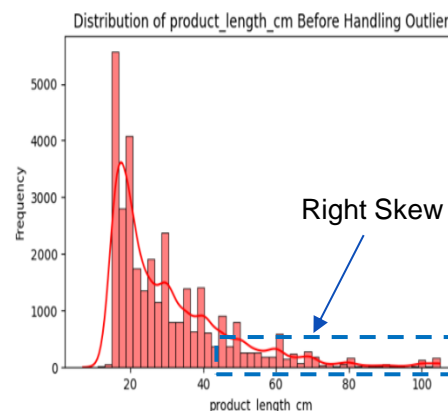
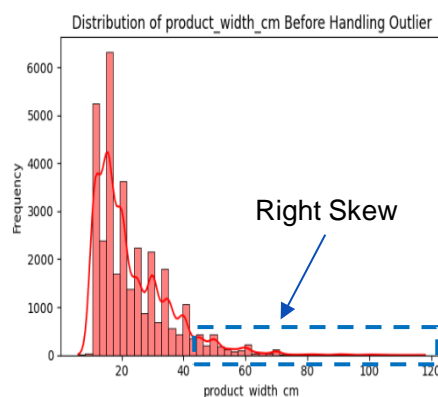
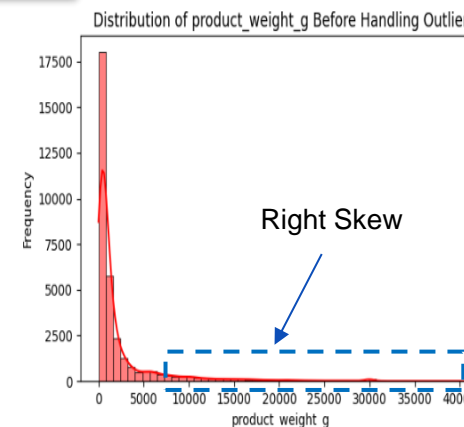
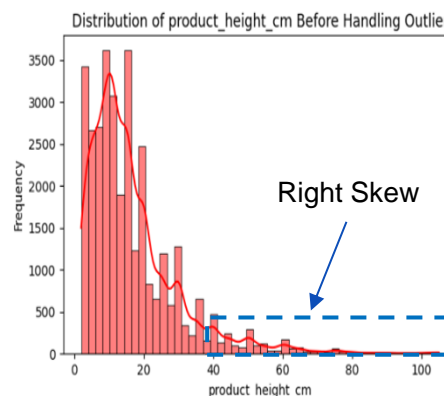
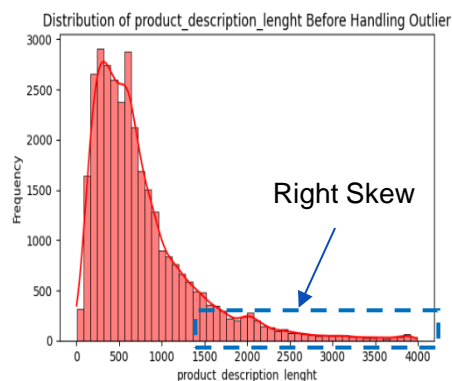
ORDER_PAYMENTS



The distribution of price, freight value, and payment value is highly **skewed to the right**, indicating the presence of extreme values (outliers). Skewness to the right means that **a large number of transactions have low values, while a small proportion of transactions have very high values**, causing the long tail to extend to the right. These outliers may require special attention in pricing, shipping, or payment strategies to address the potential impact on overall analysis and prediction.

UNIVARIATE ANALYSIS

PRODUCTS



Various product features like product description length, product height, product weight, product length, and product width all exhibit **right-skewed** distributions. This indicates the presence of extreme values or outliers in the data. Specifically, right-skewed distributions imply that **while most products have lower values (in terms of dimensions or weight), there are a few products with significantly higher values**, which stretch the distribution tail to the right.



DATA INTEGRATION

DATA INTEGRATION

customer_unique_id	total_payment_value	mean_price	avg_review_score	customer_state	payment_type	product_category_name
0000366f3b9a7992bf8c76cfd3221e2	141.90	129.90	5.0	Gujarat	credit_card	Bed_Bath_Table
0000b849f77a49e4a4ce2b2a4ca5be3f	27.19	18.90	4.0	Andhra Pradesh	credit_card	Health_Beauty
0000f46a3911fa3c0805444483337064	86.22	69.00	3.0	Andhra Pradesh	credit_card	Stationery
0000f6ccb0745a6a4b88665a16c9f078	43.62	25.99	4.0	Andhra Pradesh	credit_card	Telephony
0004aac84e0df4da2b147fca70cf8255	196.89	180.00	5.0	Chhattisgarh	credit_card	Telephony
...
ffcf5a5ff07b0908bd4e2dbc735a684	4134.84	785.00	5.0	Maharashtra	credit_card	Health_Beauty
fffea47cd6d3cc0a88bd621562a9d061	84.58	64.89	4.0	Andhra Pradesh	credit_card	Baby
ffff371b4d645b6ecea244b27531430a	112.46	89.90	5.0	Andhra Pradesh	credit_card	Auto
ffff5962728ec6157033ef9805bacc48	133.69	115.00	5.0	Maharashtra	credit_card	Watches_Gifts
ffffd2657e2aad2907e67c3e9daecbeb	71.56	56.99	5.0	Gujarat	credit_card	Perfumery

The selected features were chosen because they are **business-relevant and directly aligned with churn prediction**. Each feature provides actionable insights into customer behavior, enabling targeted retention strategies and measurable impact on customer value.

FEATURE ENGINEERING

customer_region	product_category_group	spending_category	churn
South	Home & Furniture	Low Spend	1
South	Home & Furniture	Low Spend	1
South	Home & Furniture	Low Spend	1
North	Fashion, Beauty & Food	Medium Spend	0
North	Automotive, Tools & Industrial	Medium Spend	0

Customer Region

- Grouping customers by region in India (North, South, East, West) using a mapping from state names to regions.

Product Category Group

- Grouping product categories into 6 major groups such as "Electronics & Gadgets", "Fashion, Beauty & Food", etc based on an existing category mapping

Spending Category

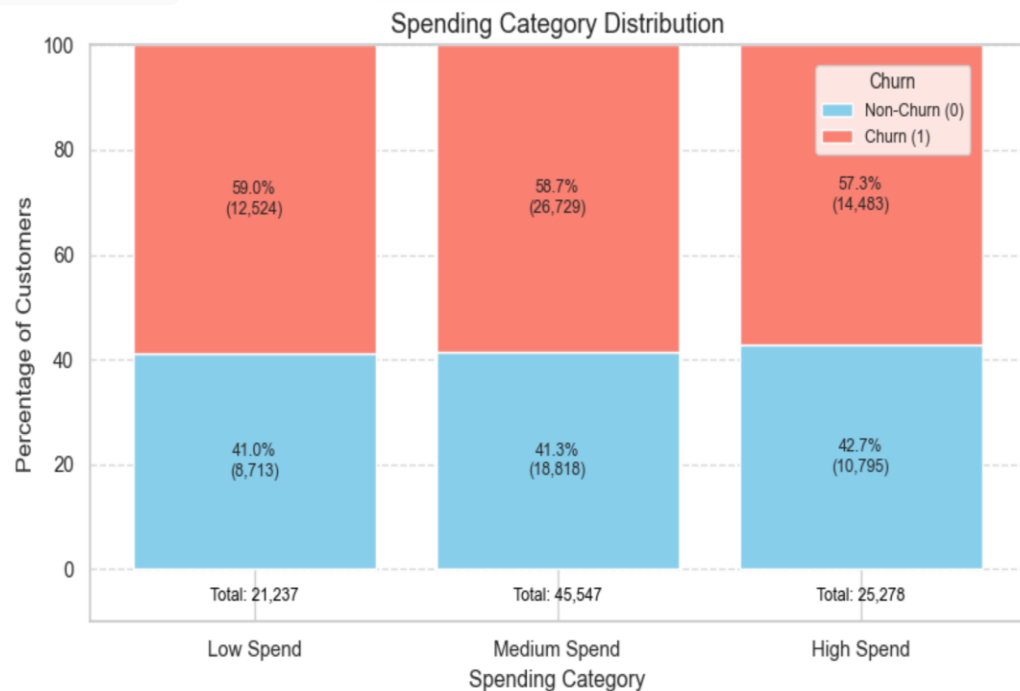
- Dividing payment values into 3 spending categories based on quartiles: "Low Spend", "Medium Spend", and "High Spend".

Churn Label

- No transactions in the last 6 months by assigning a value of 1 for churned customers and 0 for active customers.

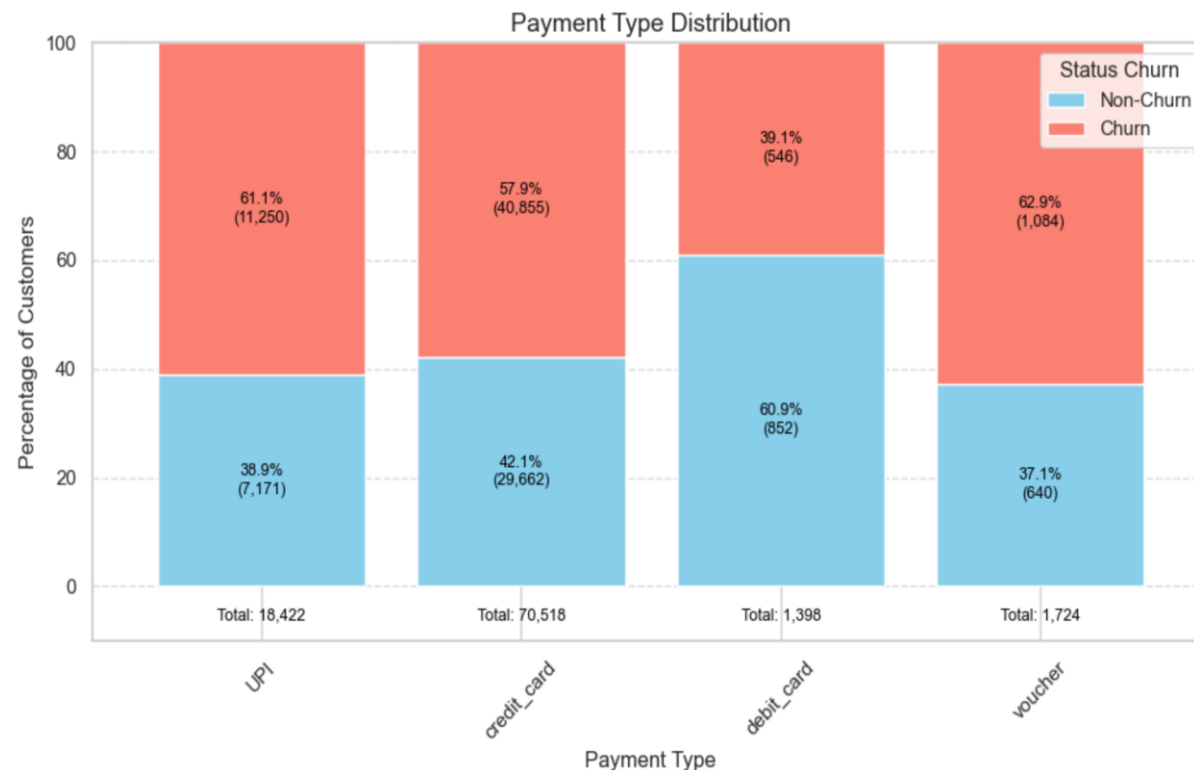
MULTIVARIATE ANALYSIS

Spending Category



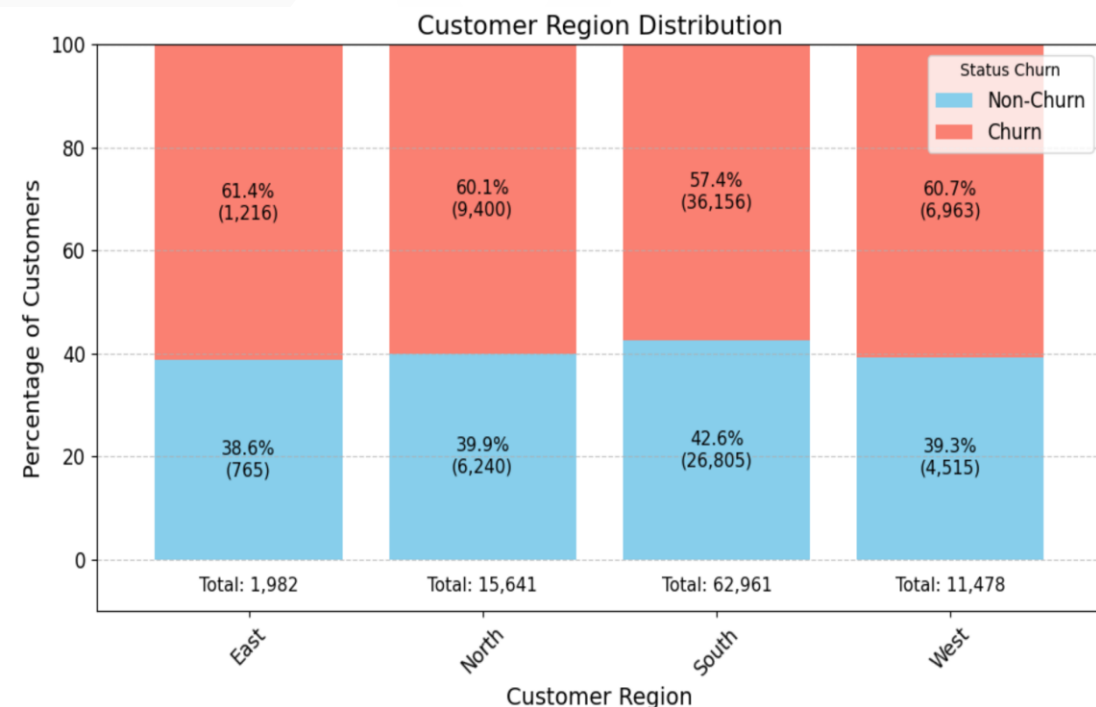
Although the low spending category has the highest percentage of churn rate within its own group, it contributes less to the overall churn because it has fewer customers. On the other hand, categories with more customers, like **Medium Spend and High Spend**, have a bigger impact on total churn. Therefore, churn reduction efforts should focus on these larger groups to keep more customers and reduce future losses

Payment Type



The chart shows that **credit card and UPI** payment types have the highest churn customers, indicating better customer retention in those categories. Therefore, focusing on retaining UPI and credit card users while maintaining the strategies that work for debit card and voucher users could help reduce overall churn.

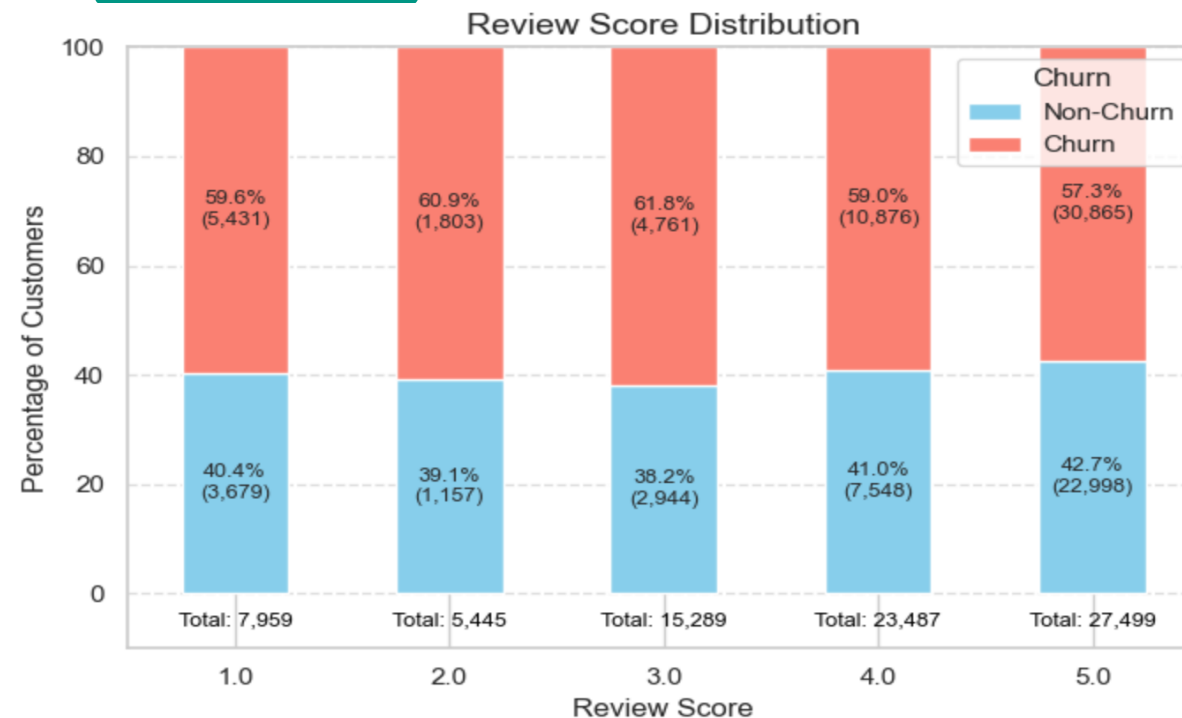
Customer Region



Although the East region has a higher percentage of churn rate, it contributes less to overall churn due to its smaller customer base, while the **South region**, with a larger customer base, has the highest number of churned customers (26,805). Therefore, churn reduction strategies should prioritize the South region while also addressing churn in the other regions.

customer_region	customer_state
0 East	Arunachal Pradesh, Orissa, West Bengal
1 North	Chhattisgarh, Delhi, Haryana, Himachal Pradesh, Jammu & Kashmir, Punjab, Rajasthan, Uttar Pradesh, Uttaranchal
2 South	Andhra Pradesh, Karnataka, Kerala, Tamil Nadu
3 West	Goa, Gujarat, Madhya Pradesh, Maharashtra

Review Score

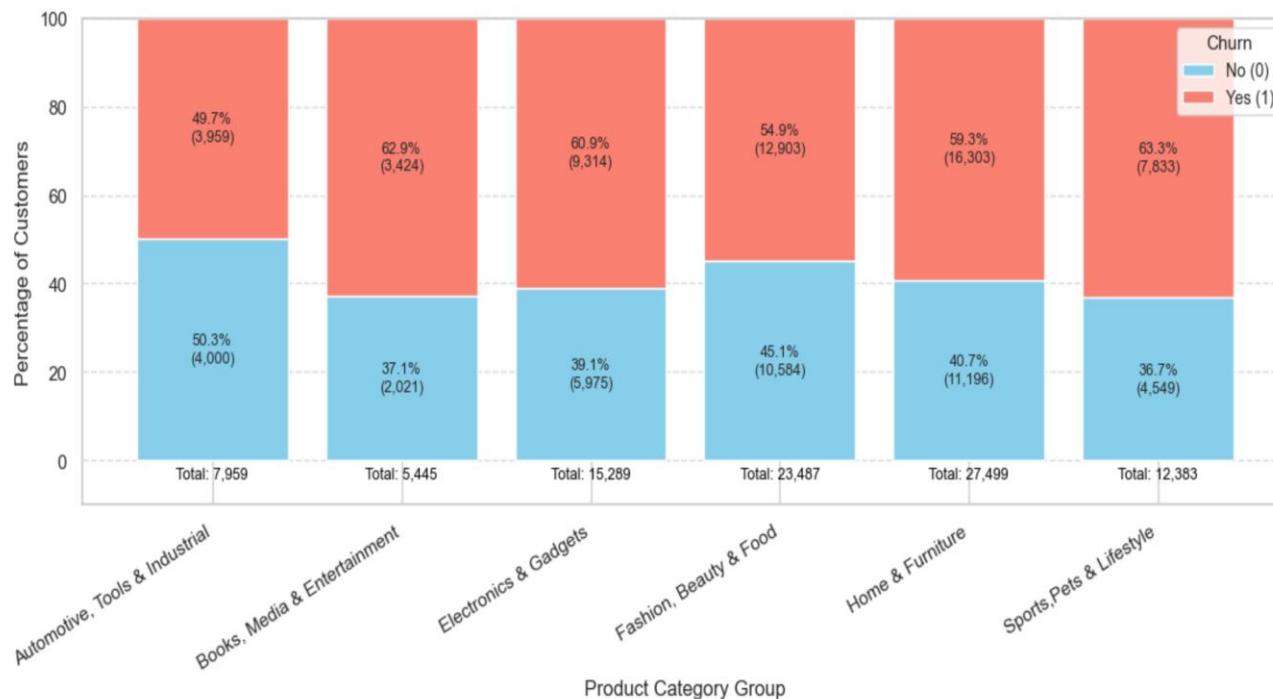


Although review scores (3.0) have a higher percentage of churn, they contribute less to overall churn due to their smaller customer base. The **larger customer base in the 5.0 review score category**, however, contributes significantly to total churn. Therefore, strategies to reduce churn should focus on improving retention across all review score categories, especially in the 5.0 review score group, where the absolute number of churned customers is the highest.

MULTIVARIATE ANALYSIS

Product Category Group

Product Category Group Distribution



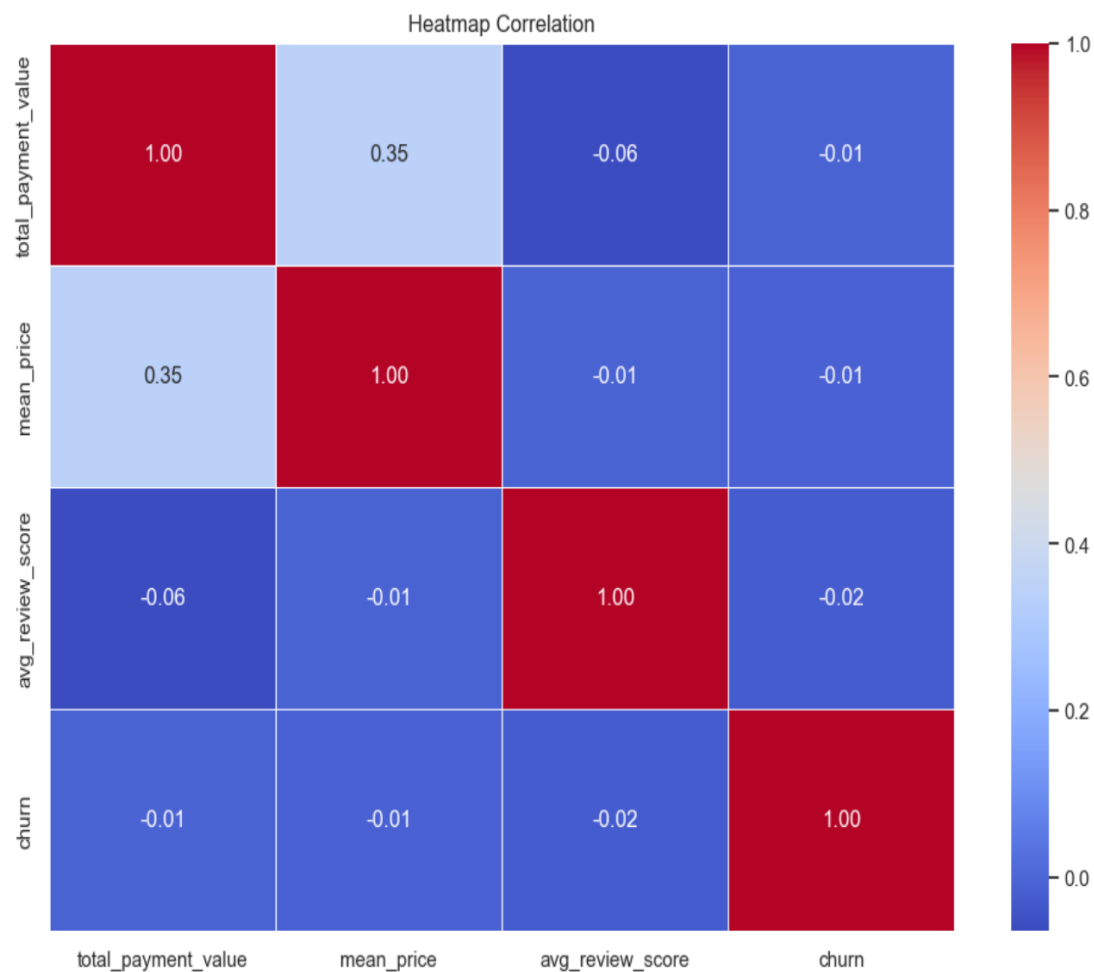
Even though categories like Sports, Pets & Lifestyle and Books have the highest percentage churn rates within their own groups, they contribute less to overall churn due to their smaller customer bases. In contrast, the **Home & Furniture** with larger customer bases, contribute the most to total churn. Therefore, efforts to reduce churn should focus on these larger categories, as retaining customers in these groups would have the greatest impact on overall churn reduction.

product_category_group	product_category
0 Automotive, Tools & Industrial	Agro_Industry_And_Commerce, Auto, Construction_Tools_Construction, Construction_Tools_Lights, Construction_Tools_Safety, Costruction_Tools_Garden, Costruction_Tools_Tools, Industry_Commerce_And_Business, Market_Place, Security_And_Services, Stationery
1 Books, Media & Entertainment	Art, Arts_And_Craftmanship, Books_General_Interest, Books_Imported, Books_Technical, Cds_Dvds_Musicals, Dvds_Blu_Ray, Music, Musical_Instruments, Signaling_And_Security, Toys
2 Electronics & Gadgets	Audio, Cine_Photo, Computers, Computers_Accessories, Consoles_Games, Electronics, Fixed_Telephony, Small_Appliances, Small_Appliances_Home_Oven_And_Coffee, Tablets_Printing_Image, Telephony
3 Fashion, Beauty & Food	Baby, Diapers_And_Hygiene, Drinks, Fashion_Bags_Accessories, Fashion_Childrens_Clothes, Fashion_Female_Clothing, Fashion_Male_Clothing, Fashion_Shoes, Fashion_Sport, Fashion_Underwear_Beach, Food, Food_Drink, Health_Beauty, Luggage_Accessories, Perfumery, Watches_Gifts
4 Home & Furniture	Air_Conditioning, Bed_Bath_Table, Furniture_Bedroom, Furniture_Decor, Furniture_Living_Room, Furniture_Mattress_And_Upholstery, Garden_Tools, Home_Appliances, Home_Appliances_2, Home_Comfort_2, Home_Confort, Home_Construction, Housewares, Kitchen_Dining_Laundry_Garden_Furniture, La_Cuisine, Office_Furniture
5 Sports, Pets & Lifestyle	Christmas_Supplies, Cool_Stuff, Flowers, Party_Supplies, Pet_Shop, Sports_Leisure



FEATURE SELECTION

HEATMAP CORRELATION



Feature such as total payment value, mean price, dan review score have weak correlation with churn, which may suggest a non-linear relationship.

STATISTIC ANALYSIS

ANOVA & CHI SQUARE TEST

ANOVA is used to determine whether a significant difference in means across multiple groups (region or spending category) related to a continuous variable like churn scores. Chi-Square tests the dependence between two categorical variables with churn.

===== ANOVA Test Results (Churn vs Categorical Variables) =====
 customer_region: F=24.8757, P=4.42e-16 → There is a significant difference
 spending_category: F=8.5266, P=1.98e-04 → There is a significant difference

===== Chi-Square Test Results (Churn vs Categorical Variables) =====
 customer_region: Chi2=74.5698, P=4.48e-16 → There is a significant dependence
 spending_category: Chi2=17.0505, P=1.98e-04 → There is a significant dependence

T-TEST

T-Test is applied to assess whether the average values of numerical features (mean price, total payment value, average review score) significantly differ between churned and non-churned customers.

===== T-Test Results (Numerical vs Churn) =====
 mean_price: T=-2.8357, P=4.57e-03 → There is a significant difference
 total_payment_value: T=-2.1687, P=3.01e-02 → There is a significant difference
 avg_review_score: T=-6.9272, P=4.32e-12 → There is a significant difference

WOE AND IV TEST

WOE and IV Test is used to measure the predictive power of categorical features in relation to churn, helping in feature selection for modeling.

===== Information Value (IV) Scores =====

customer_region: IV=0.0033 → Low
 spending_category: IV=0.0008 → Low
 payment_type: IV=0.0128 → Low
 product_category_group: IV=0.0256 → Medium

===== WOE Table for product_category_group =====

	Total	Churn	Non-Churn	% Churn \
product_category_group				
Automotive, Tools & Industrial	7959	3959	4000	0.073675
Books, Media & Entertainment	5445	3424	2021	0.063719
Electronics & Gadgets	15289	9314	5975	0.173329
Fashion, Beauty & Food	23487	12903	10584	0.240118
Home & Furniture	27499	16303	11196	0.303391
Sports,Pets & Lifestyle	12383	7833	4550	0.145768

	% Non-Churn	WOE	IV
product_category_group			
Automotive, Tools & Industrial	0.104368	-0.347859	0.010677
Books, Media & Entertainment	0.052732	0.188936	0.002076
Electronics & Gadgets	0.155899	0.105915	0.001846
Fashion, Beauty & Food	0.276157	-0.139784	0.005038
Home & Furniture	0.292125	0.037825	0.000426
Sports,Pets & Lifestyle	0.118718	0.205108	0.005548

- **Significance of Feature:** Features like **customer region, spending category, mean price, total payment value, average review score, payment type, product category group** are significant in predicting churn, indicated by low P-values in tests.
- **Predictive Power in IV Test :** **Product category group has moderate predictive power.** Other Features like customer region, payment type, and spending category have lower predictive power but still significant influence.



FEATURE SELECTION

Mean Price

- ✓ Statistically **significant in T-Test** ($p = 0.004$) for differentiating churn behavior
- ✓ **Moderately correlated** with total payment value (0.35) but provides **non-redundant information**
- ✓ **Reflects customer preference** for higher- or lower-priced products

Average Review Score

- ✓ **Highly significant** (T-Test $p < 0.001$) in relation to churn
- ✓ **Commonly used and practical feature** for predicting loyalty or churn from previous purchases

Total Payment Value

- ✓ **Statistically significant based on T-Test** ($p = 0.03$)
- ✓ **Directly reflects** the customer's overall purchasing power

Customer Region

- ✓ Shows **significant impact based on ANOVA and Chi-Square tests** ($p < 0.001$)
- ✓ **Easy to encode (only 4 categories)** without the risk of sparsity like product categories
- ✓ **Captures regional behavioral differences** relevant for churn analysis
- ✓ Although its IV is low (0.003), it remains valuable for segmentation purposes





FEATURE ENCODING

customer_region_East	customer_region_North	customer_region_South	customer_region_West
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1



One Hot Encoding

One-Hot Encoding is used for categorical columns without an order, such as [customer region](#).



MODELING

SPLIT TRAIN-TEST

FEATURE IN
TRAINING
DATA



total_payment_value	mean_price	avg_review_score	customer_region_East	customer_region_North	customer_region_South	customer_region_West
186.82	170.00	4.0	0	0	1	0
103.11	35.95	5.0	0	0	0	1
175.87	160.00	5.0	0	0	1	0
143.84	25.00	4.0	0	1	0	0
41.72	33.00	4.0	0	0	1	0

- The training data uses selected numerical and encoded categorical features such as total payment value, mean price, average review score, and one-hot encoded customer regions.
- The process of splitting the dataset for churn prediction and the features (X) and target (y) are separated. It is then **split into 70% training and 30% testing** sets using stratified sampling to ensure balanced churn distribution.

CHECKING OUTLIER IN DATA TRAINING

```

📊 Outlier Percentage in X_train:

📌 Column: mean_price
Outlier count   : 4941
Outlier percentage: 7.67% out of 64443 records

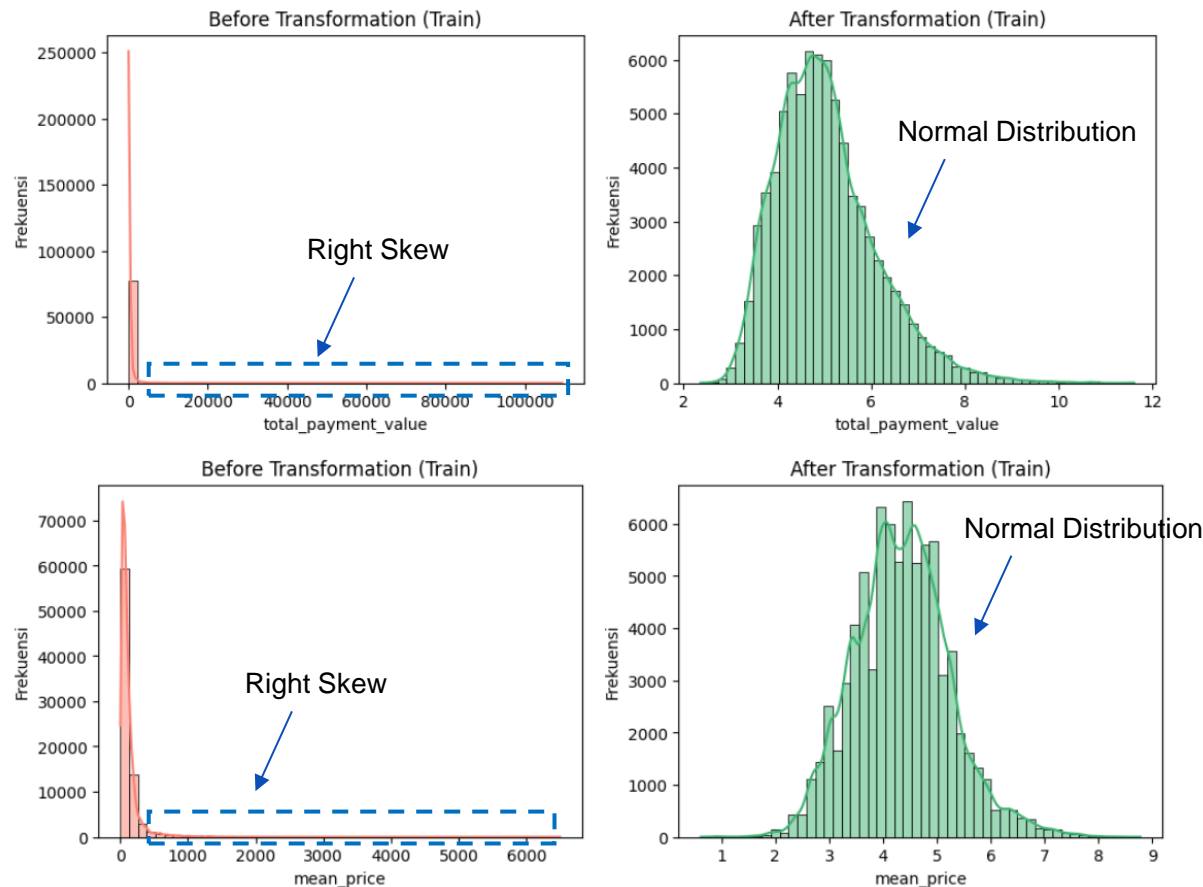
📌 Column: total_payment_value
Outlier count   : 6282
Outlier percentage: 9.75% out of 64443 records

📌 Column: avg_review_score
Outlier count   : 8345
Outlier percentage: 12.95% out of 64443 records
  
```

Outliers were not removed from the training data because they represent real variations in customer behavior, which are important for churn prediction. Removing them could result in loss of valuable signals, especially in a business context where extreme values (high spending or low review scores) may strongly influence churn decisions. Moreover, the outlier percentage is relatively moderate and can be managed through robust models.

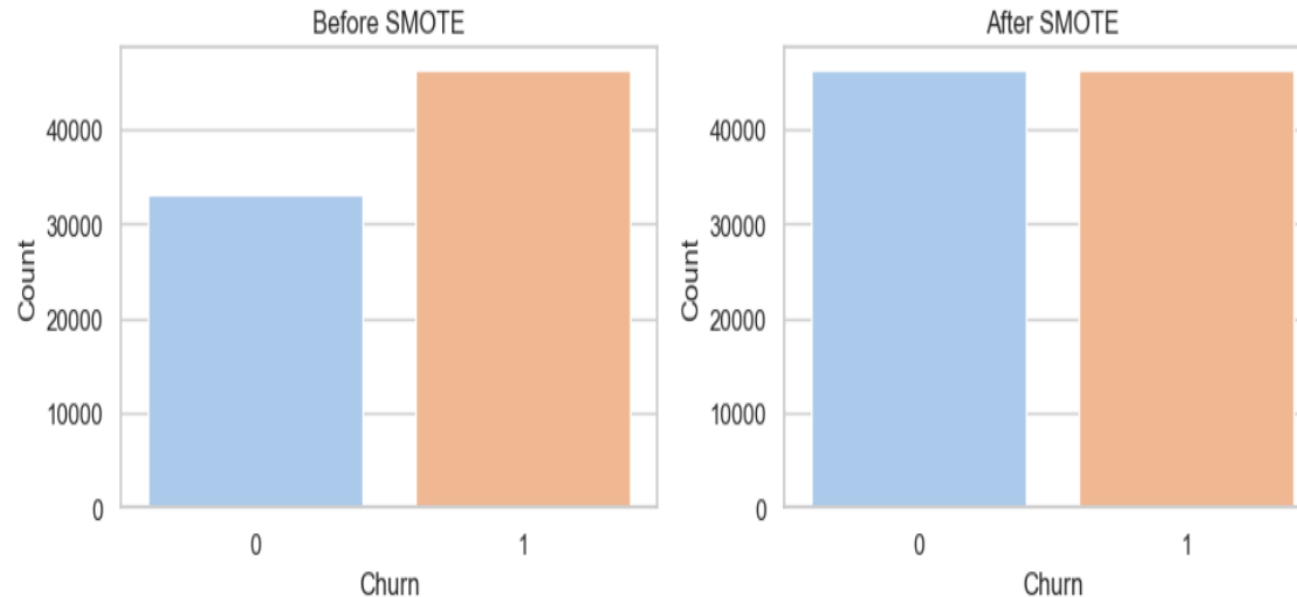


TRANSFORMATION



Log transformation was applied to **reduce skewness and normalize the distribution** of numerical features such as **total payment value and mean price**. This helps improve model performance by minimizing the impact of extreme values and making the data more suitable for machine learning algorithms.

HANDLING CLASS IMBALANCE



Class imbalance handling was performed because the **churn class was not evenly distributed**. Before SMOTE, the training set had **37,615 churn (58.37%)** and **26,828 non-churn (41.63%)** records. After applying SMOTE, both classes were balanced to 37,615 samples each, allowing the model to learn equally from both churn and non-churn customers, and reducing bias toward the majority class.



EVALUATION METRIC

In e-commerce, churn models should reach at **least 70% recall** to identify most at-risk customers and prevent revenue loss. A **ROC AUC of 70% or higher** is also preferred, as it shows the model can reliably separate churned from active users. Models below these thresholds may offer limited business impact and lead to ineffective retention efforts. (Saha, et.al, 2023)

Recall

Recall is a metric used to measure the model's ability to detect customers at risk of churn. A high Recall helps minimize False Negatives (FN), ensuring that customers who are likely to churn are identified and given the attention needed to retain them.

ROC-AUC

Evaluates the model's overall ability to distinguish between churn and non-churn across various threshold values, providing a more comprehensive measure of performance.

MODEL PERFORMANCE

NO	MACHINE LEARNING MODEL	BEFORE TUNING				AFTER TUNING			
		TRAINING		TESTING		TRAINING		TESTING	
		RECALL	ROC-AUC	RECALL	ROC-AUC	RECALL	ROC-AUC	RECALL	ROC-AUC
1.	Logistic Regression	0,4644	0,5272	0,4611	0,5214	0,4653	0,5272	0,4618	0,5214
2.	Random Forest	0,9984	0,9990	0,7392	0,7724	0,8786	0,9630	0,7037	0,7399
3.	LGBM	0,9835	0,6679	0,9812	0,6346	0,9492	0,7354	0,9377	0,6874
4.	XGBOOST	0,9602	0,6911	0,9524	0,6547	0,9132	0,7869	0,8800	0,7121
5.	Ensamble (Random Forest + XGBoost)	0,9844	0,7271	0,9797	0,6718	0,9868	0,7202	0,9823	0,6703
6.	CATBOOST	0,6152	0,6958	0,5987	0,6600	-	-	-	-
7.	ADABOOST	0,5105	0,5568	0,5005	0,5443	-	-	-	-



SELECTED MODEL

XGBOOST

High recall and improved ROC-AUC after tuning, helping detect churn accurately while maintaining balance

Performs more consistently in data training and testing than LGBM or Random Forest and avoids major overfitting

Robust to outlier data

Robust to encoded categorical data

TUNING HYPERPARAMETER

GridSearchCV



`n_estimators` (100, 200), `max_depth` (3, 5, 7), and `learning_rate` (0.01, 0.05, 0.1) for optimization.

Subsample Parameter



Evaluates `subsample` (0.8, 1.0) and `colsample_bytree` (0.8, 1.0) for regularization.

Cross validation



`Cross-validation (cv=3)` ensures generalization by training on different data subsets.

ADJUSTMENT THRESHOLD

XGBOOST After Tuning

☒ **Data Training**

Recall : 0,9132

ROC-AUC : 0,7869

☒ **Data Testing**

Recall : 0,8800

ROC-AUC : 0,7121

Threshold 0.4

XGBOOST After Tuning

☒ **Data Training**

Recall : 0,9603

ROC-AUC : 0,7869

☒ **Data Testing**

Recall : 0,9383

ROC-AUC : 0,7121

Threshold 0.35

Threshold 0.35 is chosen to ensure that no at-risk customers are overlooked.

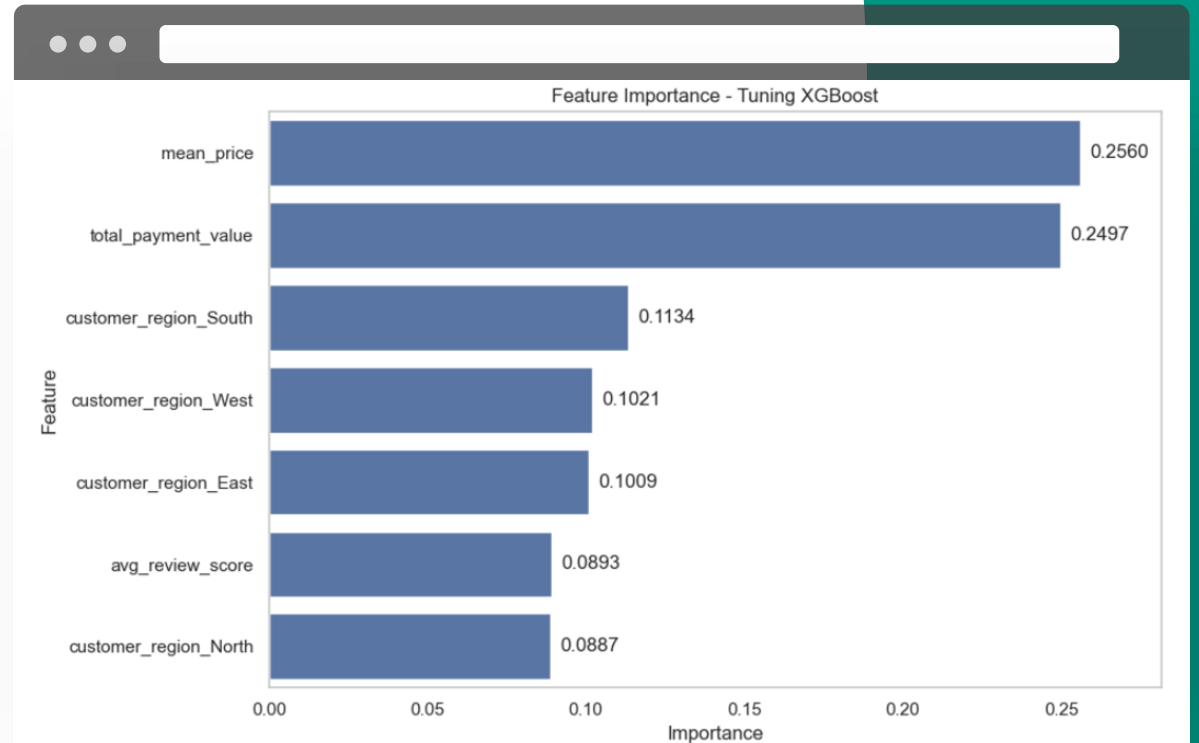


EXPLAINABILITY AND ERROR ANALYSIS



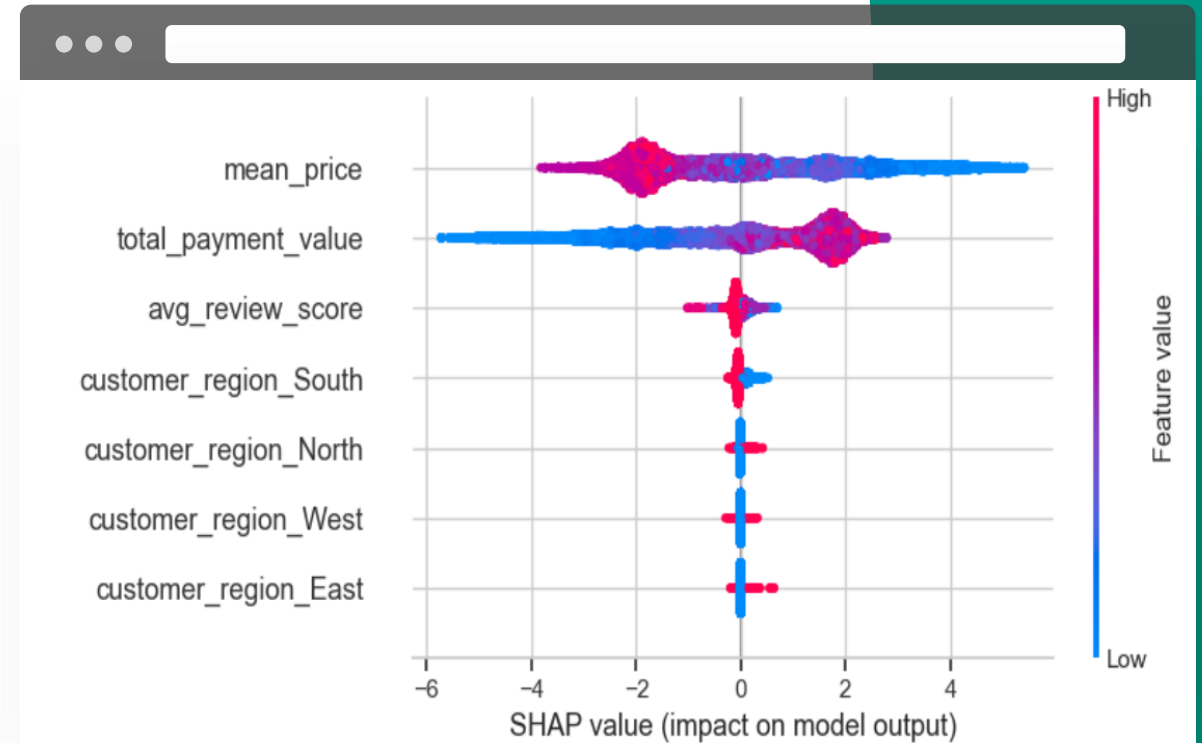
FEATURE IMPORTANCE

The most important factors in predicting churn are **mean price** and **total payment value**, showing that spending habits affect loyalty. Region and review score also play a role, meaning satisfaction and location matter. These insights help the business create more targeted and effective retention strategies.



SHAP ANALYSIS

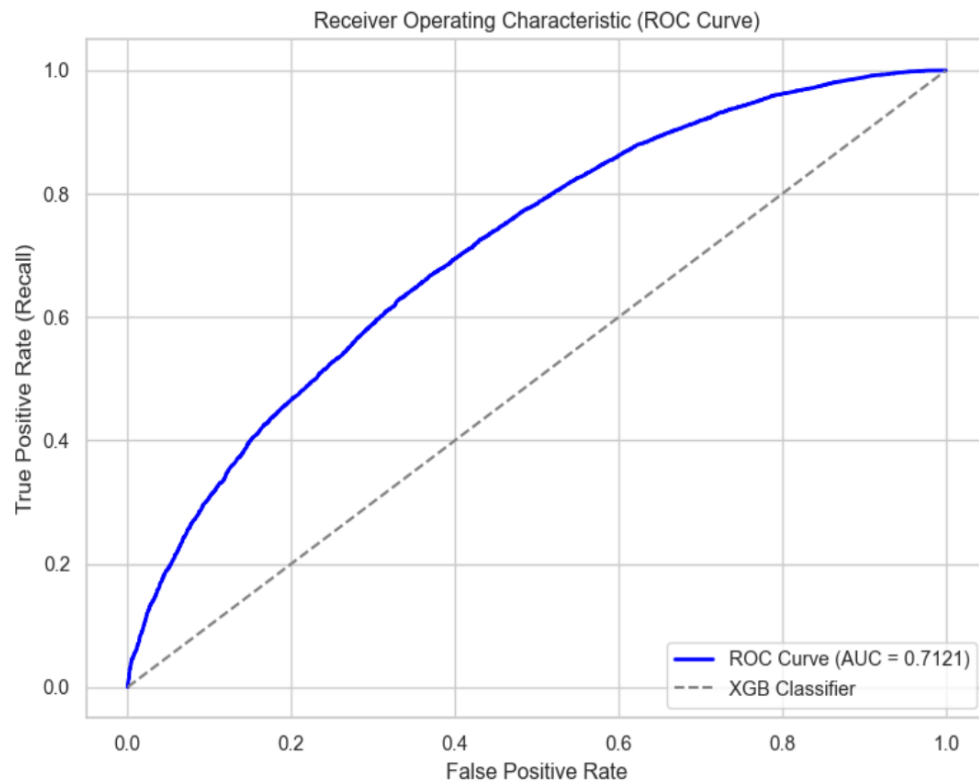
The **mean price** and **total payment value** are the **most influential features**, significantly affecting the model's predictions. Other features like average review score and customer region attributes have a smaller but still noticeable impact.





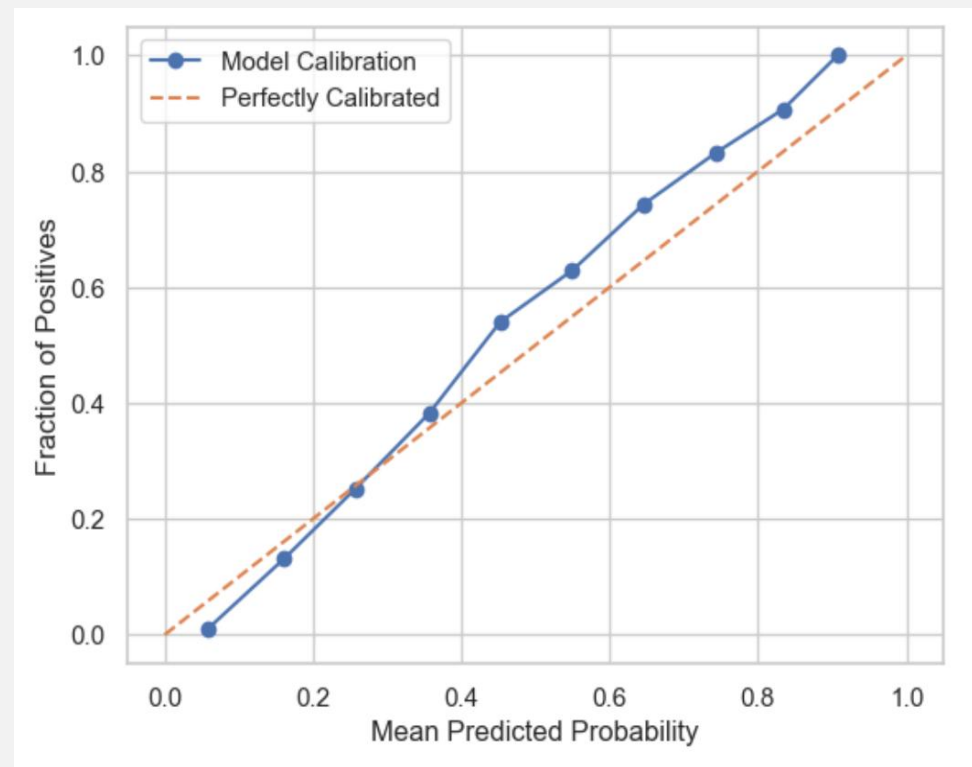
ROC AUC CURVE

The ROC curve shows the performance of the XGBoost after tuning model in distinguishing between churn and non-churn customers. With an **AUC of 0.7121**, the model demonstrates a **good level of class separation—significantly better than random guessing (AUC = 0.5)**. This indicates that the model can reliably prioritize customers based on their churn risk for more effective retention strategies.

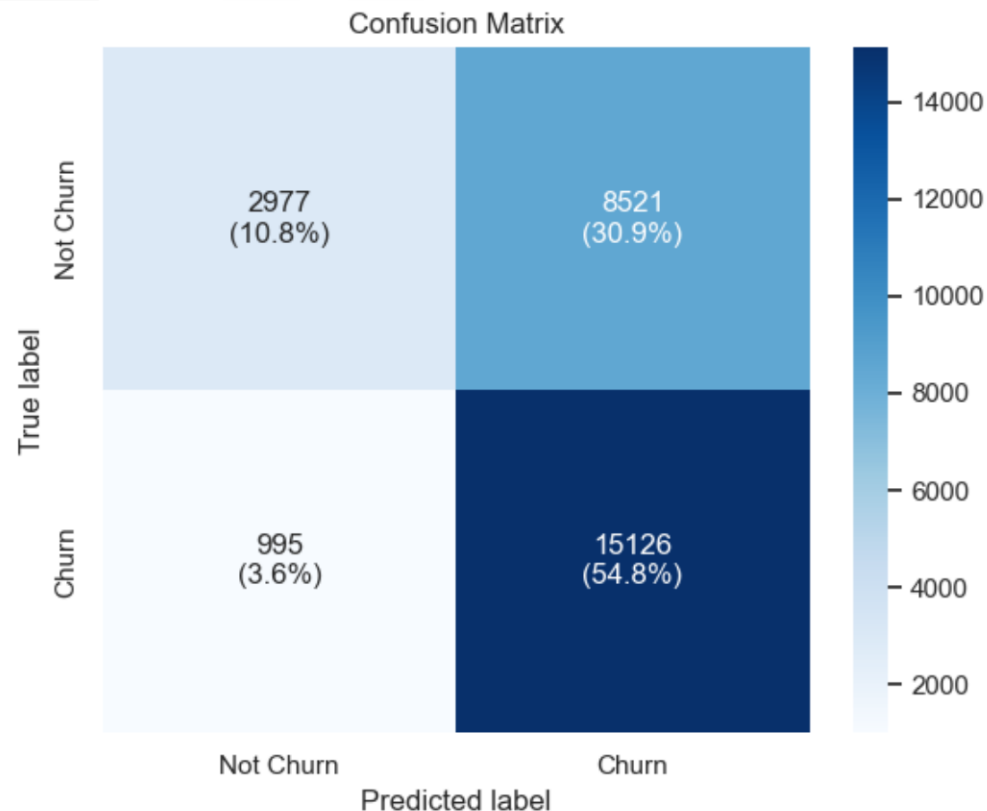


CALIBRATION CURVE

The calibration plot shows that the model is **fairly well-calibrated overall**. At higher probability levels, the model's predictions align closely with actual outcomes, indicating good confidence in identifying churn. However, at lower probabilities, the model slightly miscalibration for low-risk predictions.



CONFUSION MATRIX



This model has high recall, successfully identifying **15.126 at-risk customers**, although 8.521 customers who did not churn were predicted as churn. **Many false positives are acceptable as long as intervention costs remain low and the model generates a positive business impact.** (Saha, et.al, 2023)

Adjusted Confusion Matrix

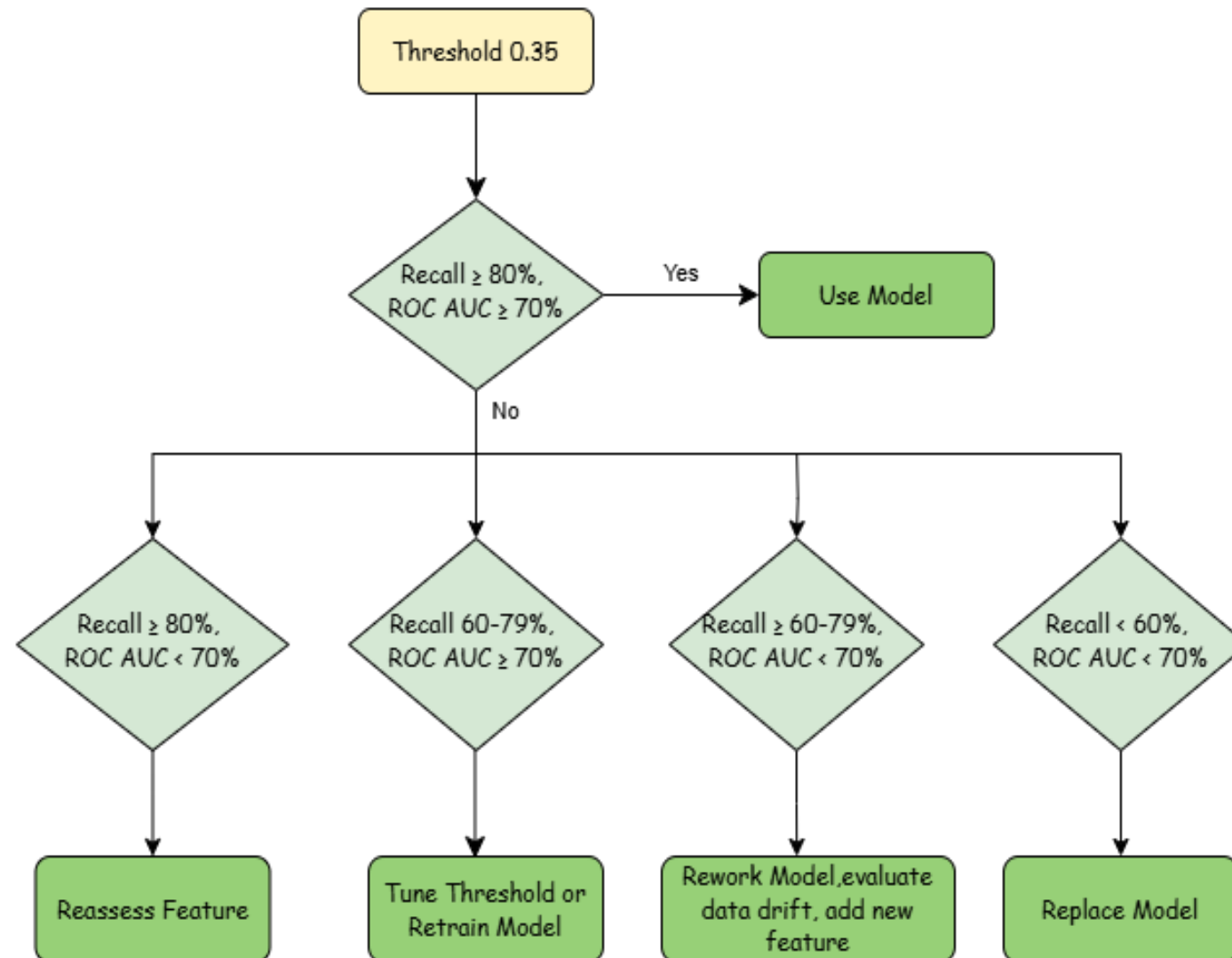
Confusion Metric	Value
Adjusted Scaled TP	50419
Adjusted Scaled FP	28403
Adjusted Scaled FN	3317
Adjusted Scaled TN	9923
Total After Adjustment	92062

The **Adjusted Confusion Matrix projects** the model's prediction results **into the full customer base of 92,062**. This scaling helps estimate real-world business impact by showing how many customers would be correctly or incorrectly classified.



MODEL EVALUATION

MODELING EVALUATION WORKFLOW



IMPACT

CUSTOMER CHURN IMPACT

Current Churn Rate

(Before Modeling)

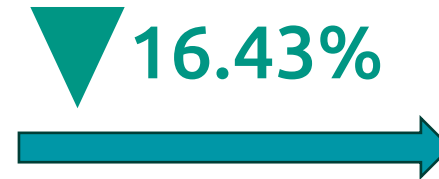
58.37 %



Projected Churn Rate

(After Modeling)

41.94 %



This model can be applied to future marketing strategies based on churn risk predictions, continuously reducing churn through customer insights and targeted retention efforts. With an estimated recovery rate of 30%, the churn prediction model is projected to reduce the churn rate from 58.37% (53,736 customers from 92,062 customers) to 41.94% (38,611 customers), successfully retaining **15,125 customers**, representing a **16.43% decrease** in churn.

Churn Evaluation	Value
Total Actual Churn Customers (TP + FN)	53736
Estimated Recovered Customers (30% of TP)	15125
Current Churn Rate	58.37%
Churn Rate Decrease	16.43%
Projected Churn Rate (Post-Intervention)	41.94%

Key Assumptions :

Recovery Rate : 30 % (of TP)

(Thomas, Blattberg, & Fox, 2004)

$$\text{Current Churn Rate} = \left(\frac{TP+FN}{\text{Total Customers}} \right) \times 100 \%$$

$$\text{Recovered Customer} = (TP \text{ adjusted} \times \text{recovery rate TP})$$

$$\text{Projected Churn Rate} = \left(\frac{TP \text{ adjusted} \times (1 - \text{recovery rate TP}) + FN}{\text{Total Customers}} \right) \times 100 \%$$

REVENUE IMPACT

Money Evaluation	Value
Total Churned Revenue	\$11,196,505.69
TP Net Gain (30% recovery rate)	\$2,511,337.83
FP Net Gain (40% loyalty rate)	\$2,409,001.89
Net Revenue Gain (TP + FP)	\$4,920,339.72
% Net Revenue Gain (post intervention)	43.95%
FN Lost Revenue	\$669,574.32
FN Loss % of Churned Revenue	5.98%
Total Discount Given (TP)	\$2,105,386.27
Total Marketing Cost (TP + FP)	\$78,822.00
Total Cost (Discount + Marketing)	\$2,184,208.27
Projected Retained Revenue	\$16,116,845.41

Net Gain

Cost

Total Churned Revenue

(Before Modeling)

\$ 11.19 M



Net Revenue Gain

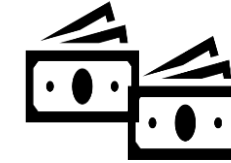
43.95 %
(\$ 4.92 M)

Total Cost
(\$ 2.18 M)

Projected Retained Revenue

(After Modeling)

\$ 16.11 M



Before modeling, the total churned revenue was \$11.19M with 53,736 churned customers. By targeting a 30% customer recovery rate and a 40% loyalty rate, the company is projected to retain 26,487 customers with a marketing cost of \$1 per customer and a discount rate of 20% (applied only to TP). This intervention is expected to **retain \$ 4.92 M in revenue**, resulting in a 43.95% net gain, which brings the total projected retained revenue to \$16.11 M. With cost of \$ 2.18 M, this strategy successfully reduces churn and significantly boosts revenue retention, demonstrating a strong return on investment.

$$\text{ROI} = (\text{Net Revenue Gain} / \text{Total Cost}) \times 100 \% = (4.92 / 2.18) \times 100 \% = 225.23 \%$$

Net Revenue Gain = (30% recovery rate x TP revenue) + (40% loyalty rate x FP Revenue)

Projected Retained Revenue = Total Churned Revenue + Total Net Gain

Cost = Total Discount FP + Total marketing cost

Key Assumptions :

Discount Rate : 20 %
Marketing Cost : 1 \$ per customer
Recovery Rate : 30 % (of True Positive(TP))
Loyalty Rate : 40 % (of False Positive (FP))

(Thomas, Blattberg, & Fox, 2004)



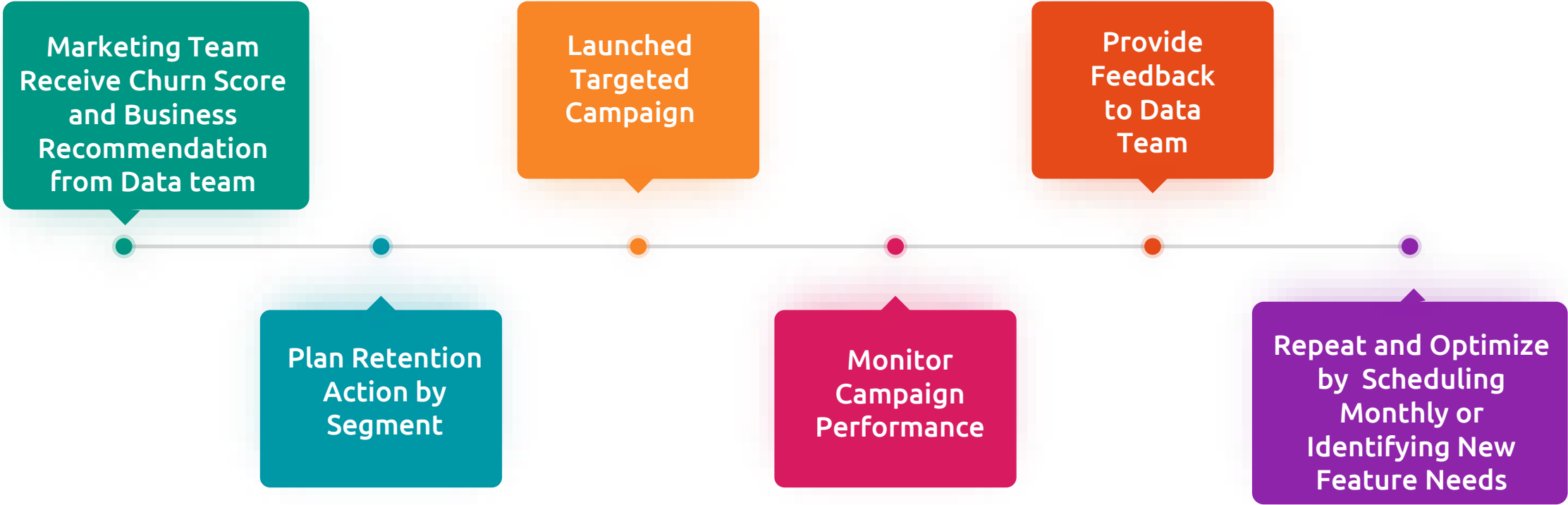
BUSINESS RECOMENDATION




BUSINESS RECOMMENDATION

Churn Risk	Support Insight (based on EDA)	Customer Region	Product Category Group	Spending Category	Review Score	Payment Type
High-Risk (Score ≥ 0.7)	<p>Region: South region has the highest churn due to its large customer base.</p> <p>Product Category: Home & Furniture needs attention due to its large customer base and churn.</p> <p>Spending Category: Medium Spend customers should be targeted with emotional engagement strategies.</p> <p>Review Score: 5.0 and 4.0 ratings show significant churn.</p> <p>Payment Type: Credit card and UPI users are at higher risk.</p>	<ul style="list-style-type: none"> - Focus on regions predicted to have high churn, specifically South Region. - Offer exclusive discounts to retain customers. 	<ul style="list-style-type: none"> - Offer exclusive discount in Home & Furniture category. - Address churn drivers in this category (e.g., limited variety, service issues). 	<ul style="list-style-type: none"> - Introduce exclusive discount to Medium Spend customers. - Focus on increasing emotional engagement with these customers. 	<ul style="list-style-type: none"> - Identify churn drivers among 5.0- and 4.0 rated customers, such as service gaps, and address them proactively. - Improving their satisfaction, and give exclusive discount, and free shipping. 	<ul style="list-style-type: none"> - Provide exclusive promotions for UPI and credit card users through cashback or loyalty rewards.
Medium-Risk (Score 0.35 – 0.7)	<p>Region: West and north region has a medium churn impact.</p> <p>Product Category: Fashion, Beauty & Food categories are quite significant for churn impact</p> <p>Spending Category: High Spend customers can be retained with value-based offers.</p> <p>Review Score: Focus on resolving issues for 1.0 ratings.</p> <p>Payment Type: Ensure smooth voucher redemption for customers.</p>	<ul style="list-style-type: none"> - Give personalized offer, such as discount coupons that can be used on specific products popular in that region. - Give discount for the next purchase. 	<ul style="list-style-type: none"> - Offer bundling offers or exclusive deals to engage customers in Fashion, Beauty & Food category. 	<ul style="list-style-type: none"> - Provide value-based offerings (e.g., discounts for bulk purchases or "buy one, get one free" offers) in High Spend customers to increase purchase frequency and raise the average order value. 	<ul style="list-style-type: none"> - Give apology email in 1.0 rated customers. - Personalized Product Recommendations with Discount 	<ul style="list-style-type: none"> - Review the voucher redemption process to ensure there are no barriers for customers to use their vouchers
Low-Risk (Score < 0.35)	<p>Region: The East region shows higher churn percentage, but has fewer customers.</p> <p>Product Category: Low churn in other categories like Books, Media & Entertainment, and so on.</p> <p>Spending Category: Low Spend category has low churn but should continue engagement to retain customers.</p> <p>Review Score: 2.0 and 3.0 ratings has low contribution to churn.</p> <p>Payment Type: Focus on debit card users with a seamless experience.</p>	<ul style="list-style-type: none"> - Give a regular newsletter with information on the latest products and limited-time offers in the Western and Eastern regions. - Give birthday discount reminders. 	<ul style="list-style-type: none"> - Offer points based, tiered memberships, referral bonuses, and birthday discounts to encourage repeat purchases and customer retention. 	<ul style="list-style-type: none"> - Continue engagement with newsletters, birthday reminders, and loyalty rewards to retain Low Spend customers and increase purchase frequency. 	<ul style="list-style-type: none"> - Offering Improvements Based on Feedback - Addressing Minor Service or Product Issues 	<ul style="list-style-type: none"> - Maintaining a smooth experience in debit card users to ensure that there are no new issues and reinforcing loyalty.

MARKETING WORKFLOW



CHURN RISK PREDICTION APP



ASHIRVADA

Customer Churn Risk Predictor

☒ Manual Input
 ☐ Input by Customer ID
 ☐ Batch CSV

Manual Input Prediction

🍎 Average Product Price

5.00

🔥 Total Payment Value

10.00

★ Average Review Score

1.00

3.00

5.00

📍 Customer State

maharashtra

🔍 Predict from Manual Input

Prediction Result

40.39%

🌟 This customer falls under the category: **Medium Risk Churn**

Customer Region: West

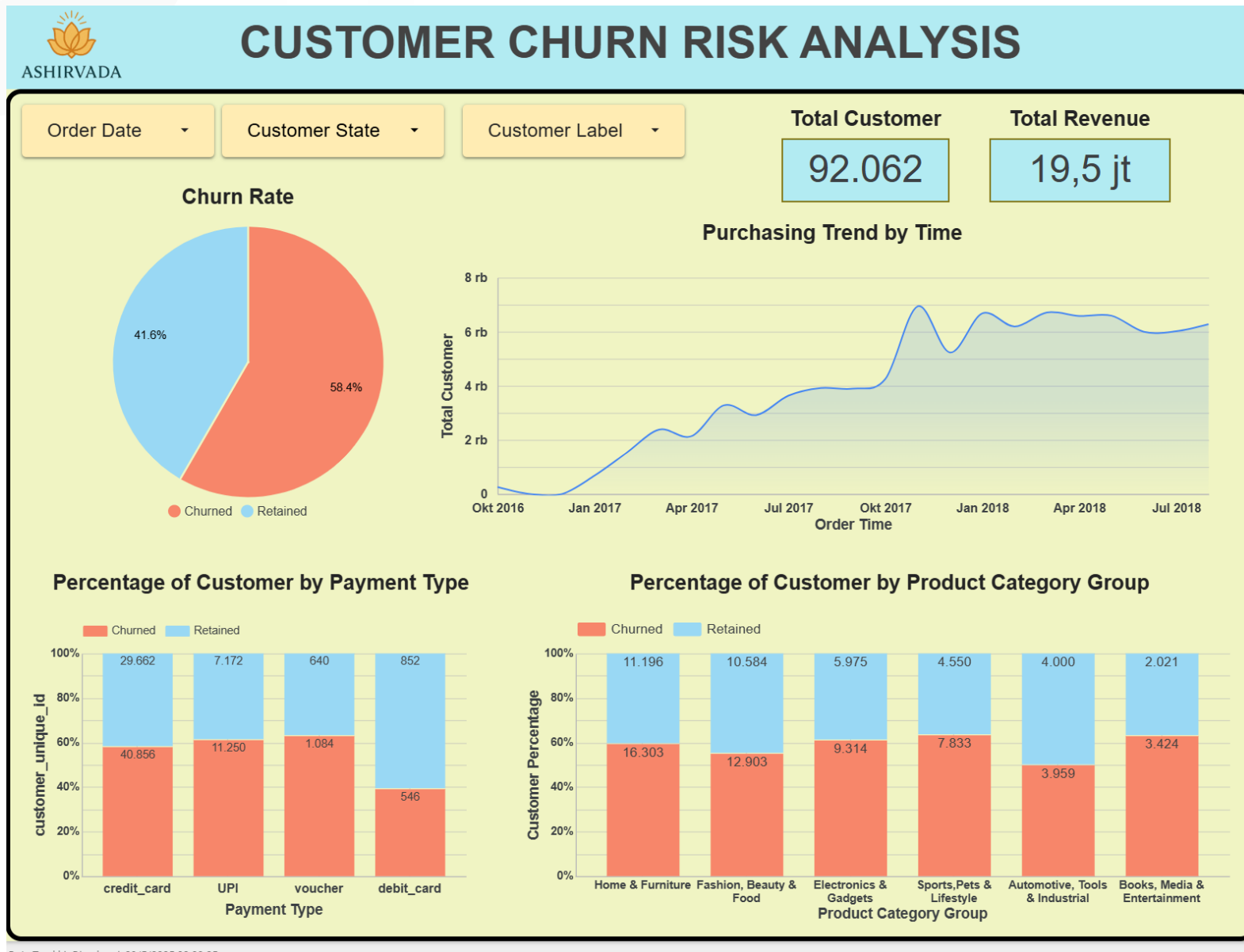
Spending Category: Low

Recommended Actions:

- 💡 Suggest products based on browsing and purchase history.
- 📋 Encourage joining loyalty programs with tiered rewards to increase stickiness.

Churn Risk Prediction
APP

CHURN RISK ANALYSIS DASHBOARD



Churn Risk Analysis
Dashboard



CONCLUSION



CONCLUSION

- The machine learning-based churn prediction model in this project, which uses **XGBoost**, achieved a **recall of 88%** and a **ROC-AUC of 71%** on the test data, demonstrating good performance in classifying churned customers.
- Features such as **mean price, total payment value, review score, and customer region** were proven to be **significant factors** influencing customer churn decisions.
- Implementing this model in business strategy can reduce the **churn rate in 16.43%** and preserve up to **\$4.92 million in revenue**, with only a low-cost intervention.
- Churn-risk-based segmentation enables companies to conduct **measured interventions**, ranging from aggressive promotions for high-risk segments to long-term engagement strategies for more stable customers.



REFERENCES

Saha, L., Tripathy, H. K., Gaber, T., El-Gohary, H., & El-kenawy, E.-S. M. (2023). Deep churn prediction method for telecommunication industry. *Sustainability*, 15(5), 4543. <https://doi.org/10.3390/su15054543>

Thomas, J., Blattberg, R. C., & Fox, E. (2004). Recapturing Lost Customers. *Journal of Marketing Research*, 41(1), 31–45. <https://doi.org/10.1509/jmkr.41.1.31.25086>.

Little, R.J.A. and Rubin, D.B. (2002) *Statistical Analysis with Missing Data*. 2nd Ed., Wiley Interscience, New York.



Thank You 