CAPSTONE PROJECT 3

E-COMMERCE CHURN PREDICTION

BY: Ajeng Aulia Salshabila

TABLE OF CONTENTS

BUSINESS PROBLEM

The Company currently lacks a systematic way to identify customers at high risk of churning, leading to retention promotions being deployed broadly and inefficiently. This results in inflated marketing costs and missed opportunities to prevent revenue loss.

FINAL GOALS

- *Build a classification model to identify customers likely to churn.*

- *Generate churn probability scores to help prioritize retention efforts.*

- *Minimize false negatives, ensuring customers who are at risk of churning are not overlooked.*

STAKEHOLDERS

- *Marketing & Retention Team*
- *Product & Service Development Team*
- *Executive Management (C-Levels)*

PRIMARY **METRIC**

Target / Label

1 : Churn
0 : Not churn

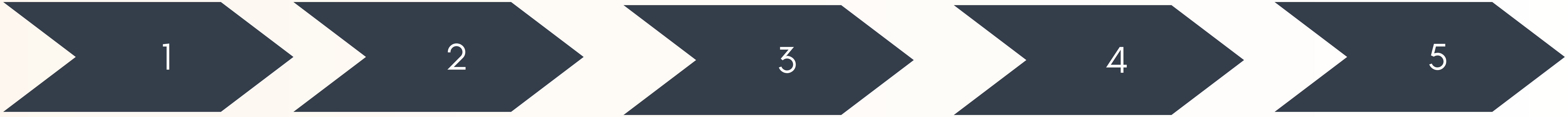
The model prioritizing **Recall** to minimize False Negatives (FN). This strategy aligns with the core business objective: reducing FN is crucial for maximizing customer retention and profitability.

DATA UNDERSTANDING

Columns	Description
Tenure	Duration (in months or years) the customer has been with the company
WarehouseToHome	Distance between the warehouse and the customer's home
NumberOfDeviceRegistered	Total number of devices registered under the customer's account
PreferedOrderCat	Customer's most preferred order category in the last month
SatisfactionScore	Customer satisfaction rating based on service experience
MaritalStatus	Customer's marital status
NumberOfAddress	Total number of addresses added by the customer
Complaint	Indicates whether the customer raised any complaint in the last month
DaySinceLastOrder	Number of days since the customer's most recent order
CashbackAmount	Average cashback received in the last month
Churn	Churn

DATA CLEANING

& Feature Engineering



- **Category Standardization**

Cleaned by lowercasing, removing extra characters, and merging equivalent categories into a single standardized label.

- **Missing Value Analysis**
  - Missing values found in Tenure, WarehouseToHome, and DaySinceLastOrder.
  - Used missingno visualizations

- **Distribution Check**

Identified skewed distributions

- **Imputation**

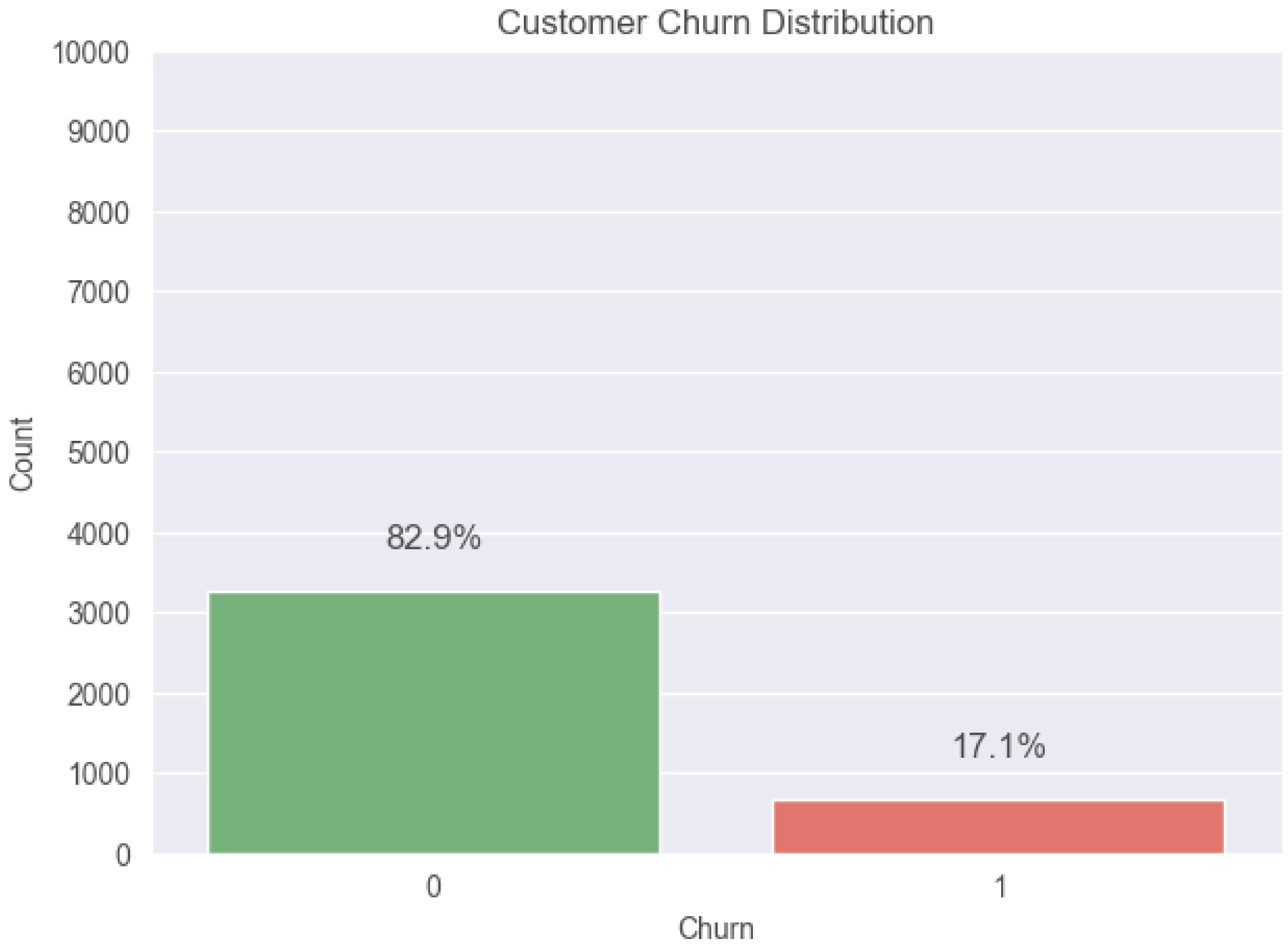
Median chosen to preserves the original distribution better than the mean.

- **Feature Engineering**

Performed train-test split, applied OneHotEncoder for categorical features, and used StandardScaler to normalize numerical features before modeling.

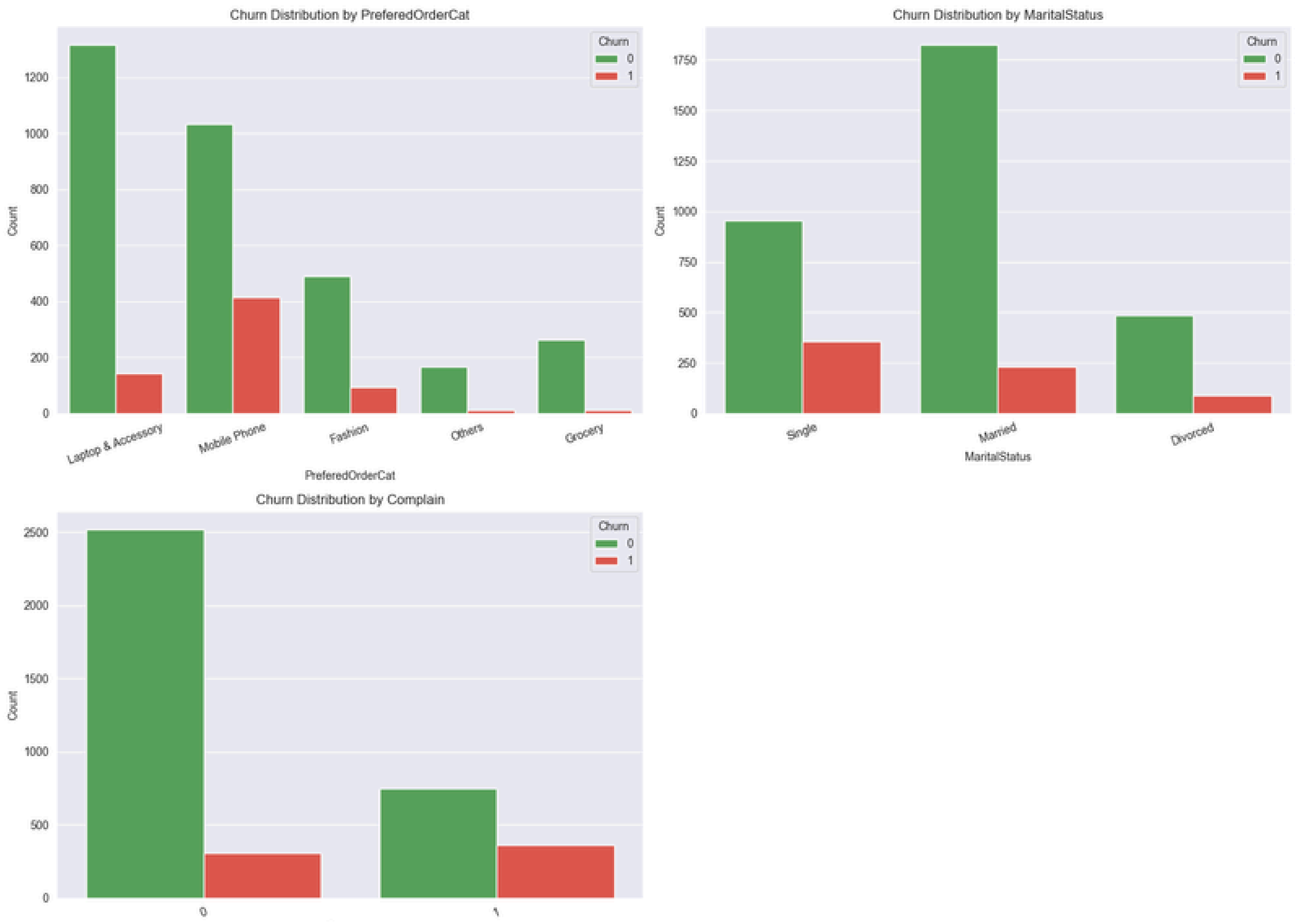
EXPLORATORY DATA ANALYSIS

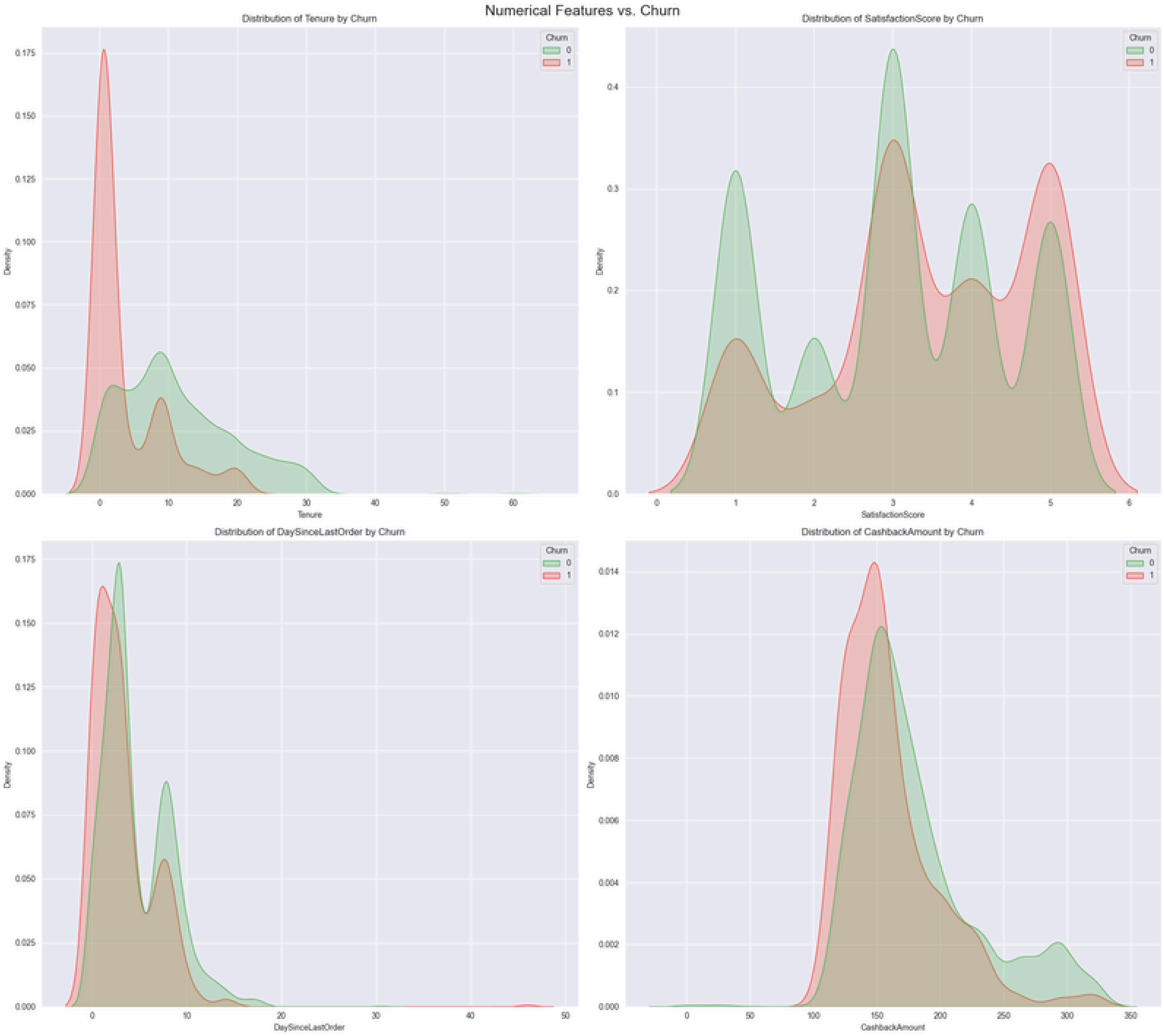
The target distribution is highly imbalanced, with only around ~17% churn cases.



CATEGORY VS CHURN :

MAINLY INFLUENCED BY PRODUCT CATEGORY CHOICE, MARITAL STATUS, AND CUSTOMER COMPLAINTS.





NUMERICAL VS **CHURN:**

HIGHEST AMONG SHORT-TENURE USERS WITH LOW SATISFACTION, RECENT ORDERS, AND MINIMAL CASHBACK.

MODEL SELECTION

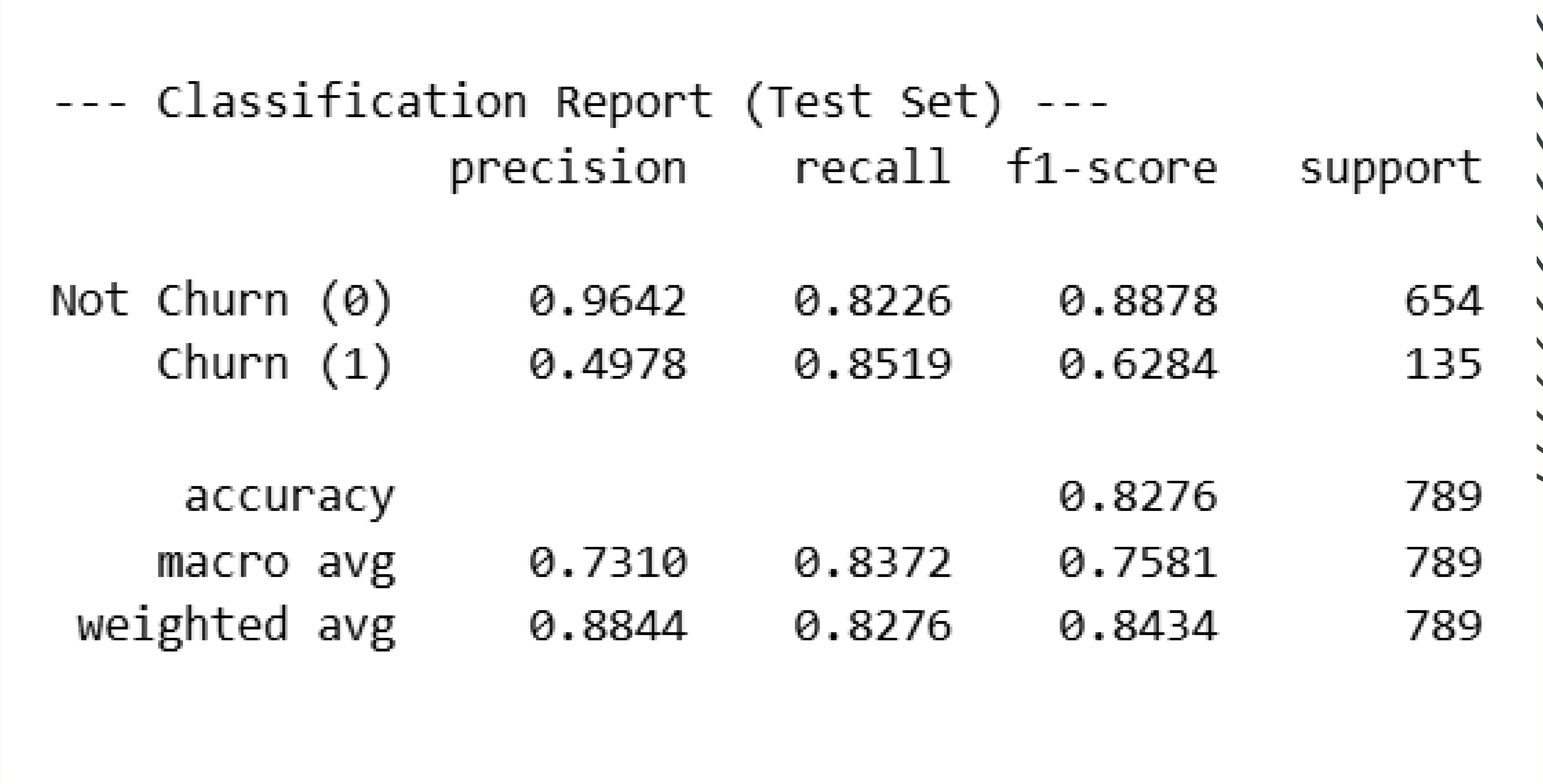
Model	Mean Accuracy	Mean Accuracy	Mean Recall	Mean F1
KNN	0,825824	0,495279	**0,833091**	0,620832
SVM	0,77569	0,420169	**0,816355**	0,554661
Logistic Regression	0,786796	0,433892	**0,81073**	0,565257
XGBoost	0,926711	0,80035	0,762513	0,780718
Decision Tree	0,899105	0,684739	0,760592	0,720615
Random Forest	0,925125	0,797398	0,755088	0,775053
LightGBM	0,923859	0,794961	0,749619	0,771
Gradient Boosting	0,900693	0,704924	0,721703	0,713081

HYPERPARAMETER TUNING

KNN

- model__metric: euclidean
- model__n_neighbors: 13
- model__weights: distance

Best recall: 0.8887504326756662



```
--- Classification Report (Test Set) ---
               precision    recall  f1-score   support

Not Churn (0)     0.9642    0.8226    0.8878       654
    Churn (1)     0.4978    0.8519    0.6284       135

     accuracy                         0.8276       789
    macro avg     0.7310    0.8372    0.7581       789
 weighted avg     0.8844    0.8276    0.8434       789
```

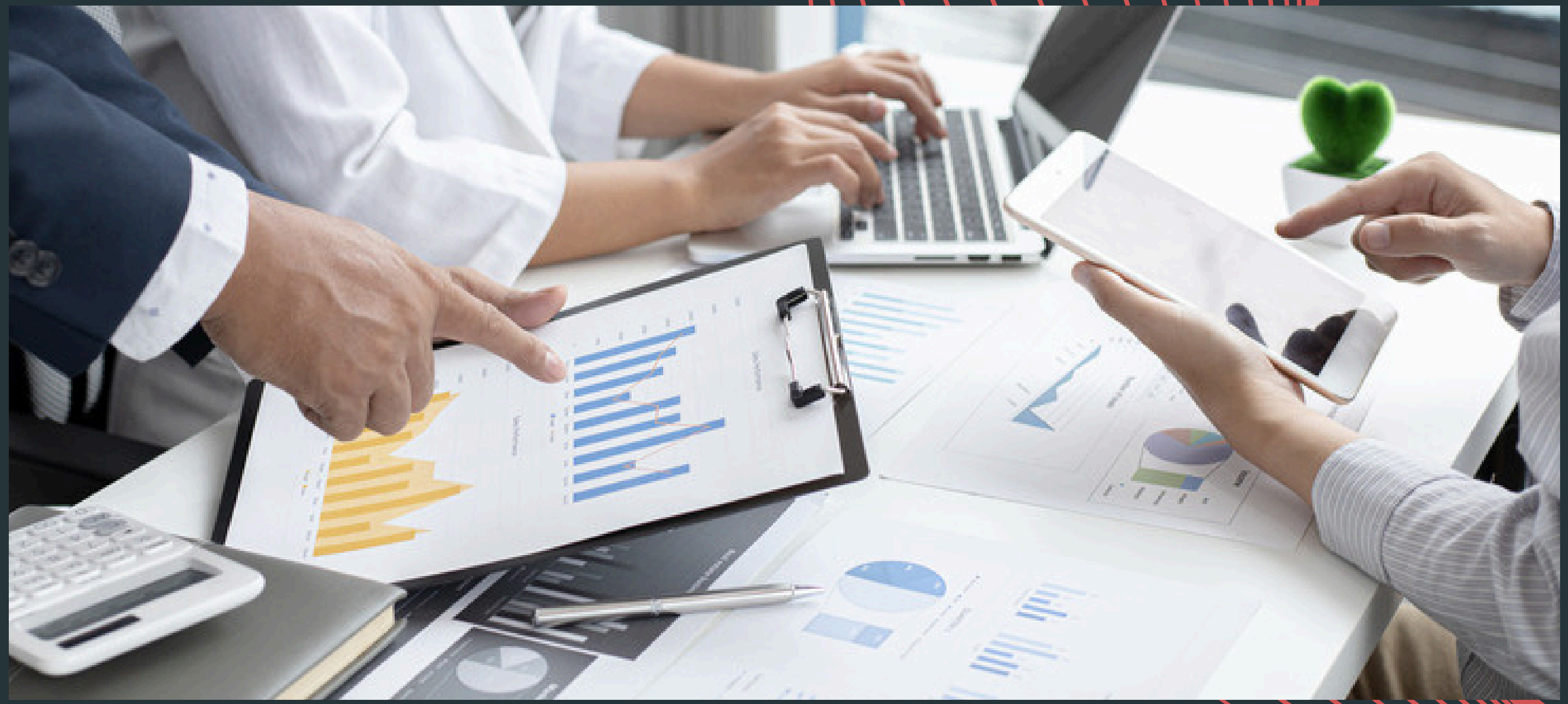
- **DATA TEST**

HOW KNN WORKS

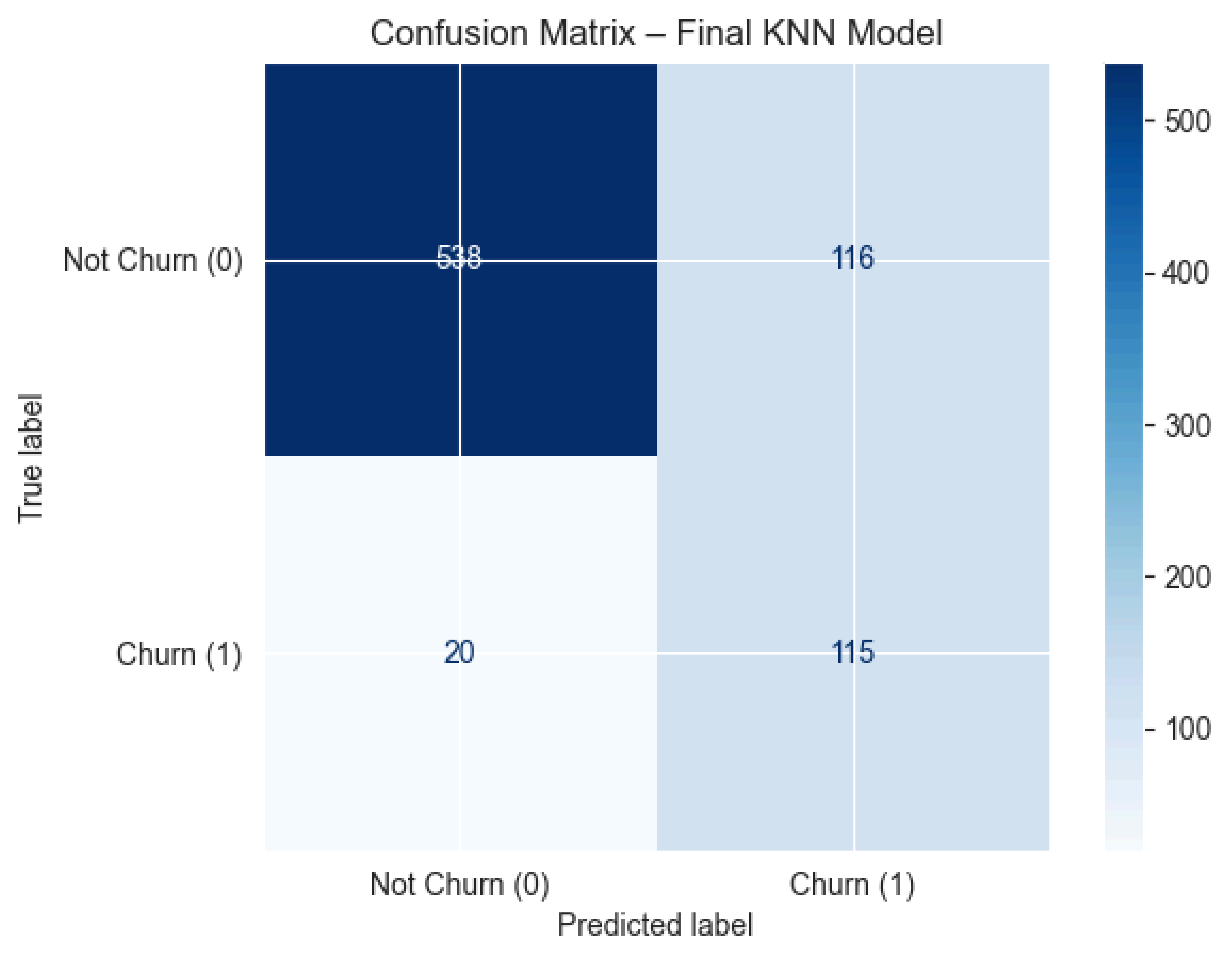
Feature Space (each dot = customer)

```
    o   o    ← Non-Churn
 o     o



      X ?     ← New customer
   ▲  ▲  ▲
   |    |    |
   o    o    o    ← Churn
```

KNN looks at the k closest customers and predicts based on the majority class.



BUSINESS EVALUATION



- *The Cost without Model:*

135 customer churn x Rp 700,000
TOTAL LOSS = Rp 94,500,000

- *Cost for Scenario 2 (With Model):*

Cost of 'Missed' (FN): 20 customers x Rp 700,000 = Rp 14,000,000
Cost of 'Wrong Intervention' (FP): 116 customers x Rp 200,000 = Rp 23,200,000
TOTAL COST = Rp 37,200,000

COST–BENEFIT ANALYSIS RESULT (per 789 customers)	
Total Loss (Without Model):	**Rp94,500,000**
Total Cost (With Model):	**Rp37,200,000**
POTENTIAL SAVINGS:	**Rp57,300,000**
Total Intervention Cost (Investment):	**Rp46,200,000**
Return on Investment (ROI):	**12.403%**

OVERALL **CONCLUSION**

This project successfully developed a high-recall churn prediction model and uncovered clear behavioral drivers of customer churn.

The analysis confirms a significant retention gap, with churn concentrated among short-tenure customers, recent one-time buyers, low-satisfaction users, and customers who submitted complaints, while higher cashback acts as a retention lever.

The tuned KNN model was selected for its superior Recall, enabling the business to proactively identify high-risk customers and improve retention strategies more efficiently.

BUSINESS RECOMMENDATION

1. Proactive, Targeted Retention

- Focus interventions on customers predicted to churn.
- Use personalized offers: targeted discounts, increased cashback, or priority support.

2. Optimize Retention Budget

- Reduce unnecessary spending by avoiding promotions for customers unlikely to churn.
- Direct incentives only to high-risk segments, where they create real impact.

3. Act on Key Churn Drivers

- Closely monitor customers with short tenure, low satisfaction, recent complaints, or declining activity.
- Use these signals as triggers for early retention outreach.

4. Strengthen Cashback Strategy

- Cashback is shown to support retention; use it intentionally.
- Provide higher or customized cashback for the most at-risk customers.

THANK YOU!