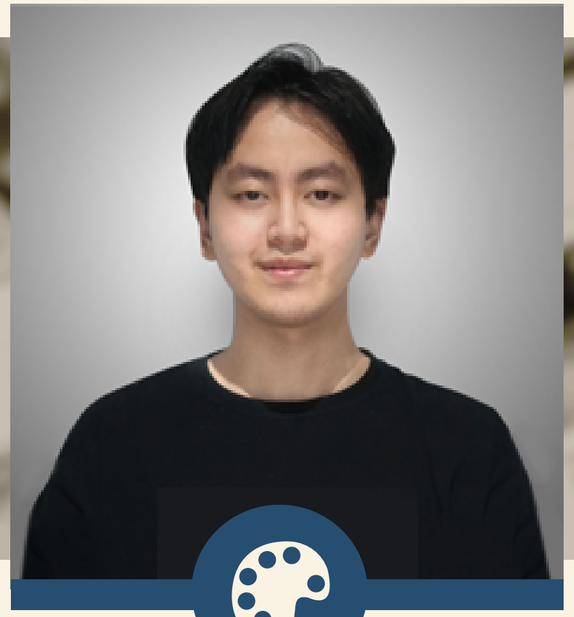
E-COMMERCE CUSTOMER CHURN ANALYSIS AND PREDICTION

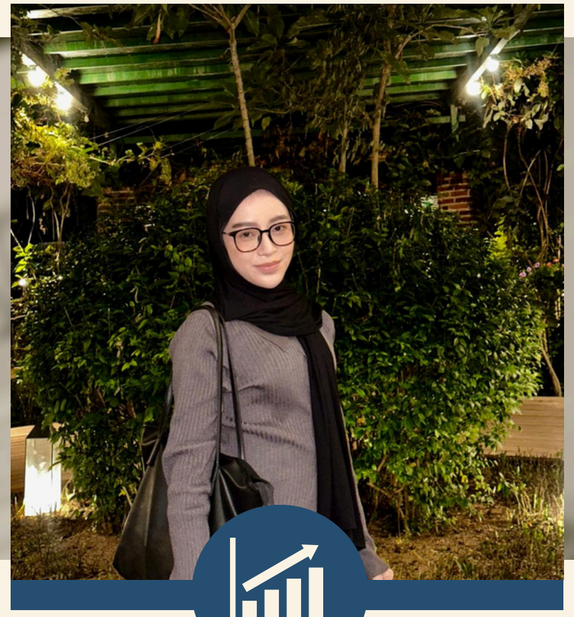
Purwadhika Data Science Final Project
Group Alpha – JCDSOL018

OUR TEAM



Aldino Dian Mandala Putra

JCDSOL-018-010



Liswatun Naimah

JCDSOL-018-016

Zulfi Nadhia Cahyani

JCDSOL-018-006

OUTLINES

01 Business Understanding

02 Data Understanding and Preparation

03 Exploratory Data Analysis (EDA)

04 Modeling and Evaluation

05 Business Impact

06 Conclusion and Recommendations

BUSINESS UNDERSTANDING

Presents the context, business problem, goals, and the analytical approach used to analyze the data and develop a machine learning model for predicting customer churn in an e-commerce platform.

E-Commerce Industry & Churn Challenge

E-Commerce Landscape

- Fast-growing but highly competitive
- Customer loyalty is fragile due to low switching costs

Churn in E-Commerce

- Churn = customer becomes inactive or stops buying
- Often silent and hard to detect
- Common causes:
  - Better offers from competitors
  - Poor Service or delivery experience
  - Low customer engagement

Why Churn Prediction Matters

16.8% of customers have churned, signaling revenue loss risk

- Churn increases Customer Acquisition Cost (CAC)
- Retaining customers is 5–25× cheaper than acquiring new ones (Harvard Business Review, 2014)
- No predictive system to detect churn patterns
- A data-driven solution is needed for early intervention

Project Goal

Build a predictive model to reduce churn

- Detect churners early
- Support targeted retention
- Improve CRM decisions

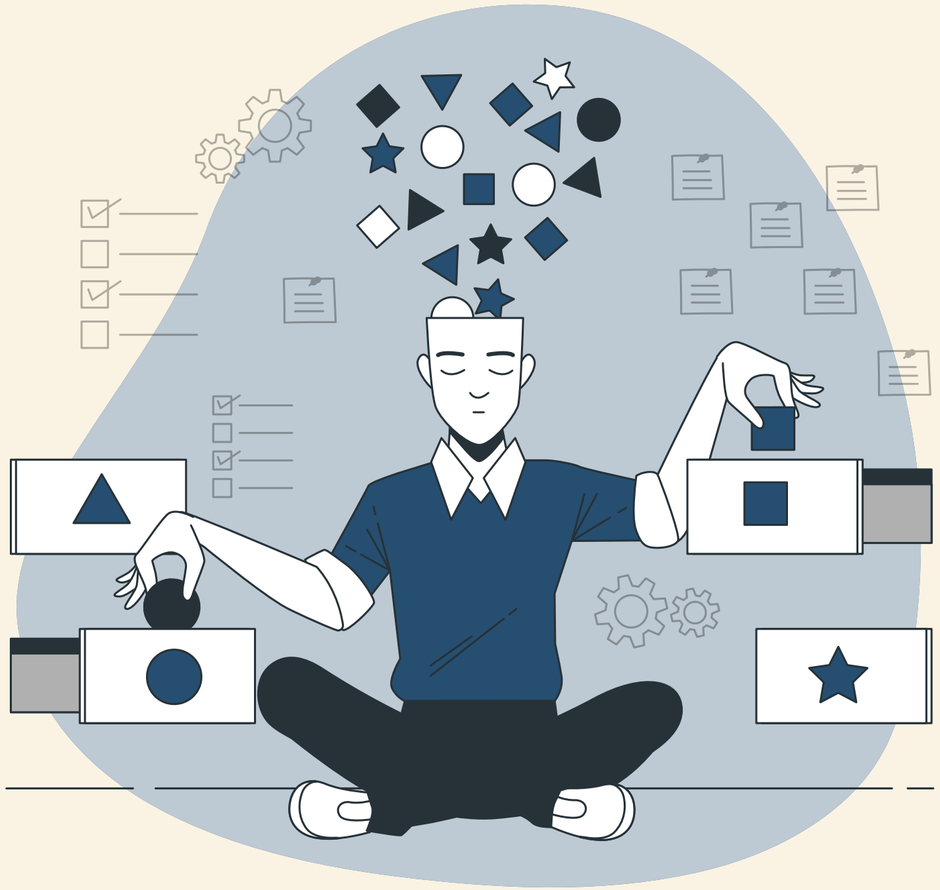
Analytical Focus

Business-driven machine learning approach

- This project builds a binary classification model to predict:
  - Will a customer churn? → Yes (1) or No (0)
- However, no predictive model is 100% accurate.
  - Even the best model will occasionally make mistakes in its predictions.

Error Type	Business Risk
False Negative *(missed churn)*	Silent loss of revenue, no chance to intervene in time.
False Positive *(misclassified loyal)*	Unnecessary incentive spend on customers who would have stayed.

- In churn prediction, missing actual churners (FN) poses the highest risk, which is why:
  - Minimizing false negatives is the main objective.
- Priority: F2-Score as the main metric (recall-oriented), followed by Recall and PR-AUC to minimize false negatives.



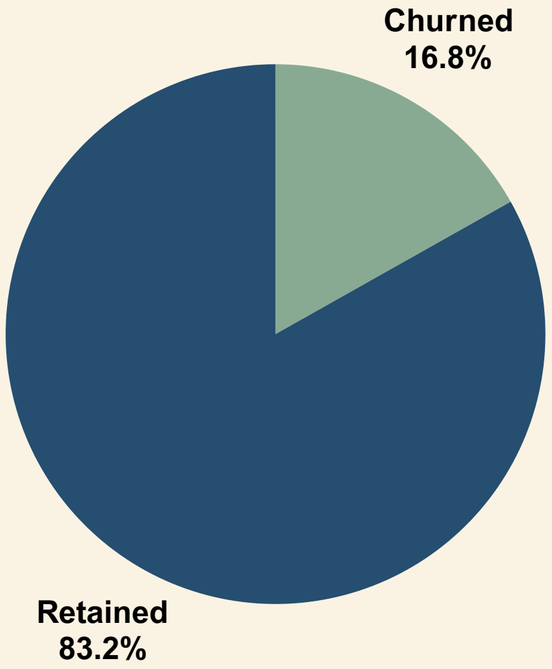
DATA UNDERSTANDING AND PREPARATION

Details the process of understanding the dataset, followed by data cleaning and preprocessing to ensure the data is accurate, consistent, and ready for modeling.

Data Understanding

This dataset represents customer records from an **e-commerce platform**, including demographic details, app usage behavior, and transaction history.

- Dataset Name: Ecommerce Customer Churn Analysis and Prediction
- Author: Ankit Verma
- Published: Circa 2021
- Number of Records: 5,630 rows
- Number of Features: 20 columns
- Source: Kaggle - E-Commerce Churn Dataset
- General Description: This dataset captures behavioral and profile data from e-commerce platform users, including login behavior, order history, and churn status (whether a customer left the platform).



***Churn Rate** PieChart*

Data Limitation

Data limitation refers to the constraints or shortcomings in the data used for a research, analysis, or project. These limitations can affect the accuracy, completeness, and validity of the results.

Limitations:

- **No Timestamp or Date Field** → Prevents time-based.
- **Unspecified Units for Some Feature** → The dataset lacks unit descriptions for several numeric fields, so we need to make assumptions about the features.
- **Churn Definition is Not Clearly Stated** → It may be based on lack of app login, inactivity in transactions, or both—but no definitive rule is provided in the dataset.
- **Missing Values** → 4.5% to 5.5% missing values.
- **Class Imbalance** → Churn (16.8%) and Retained (83.2%).

Data Cleaning

Missing Value

7 numerical columns were found to have low missing values, each ranging between **4.46%** and **5.45%,** Solution : **Removed Rows** (OrderCount), **Filled”0"** (CouponUsed), Median Imputation.

Incorrect Data Type

Tenure, CouponUsed, OrderCount, and DaySinceLastOrder were converted **from floats to integers** to **ensure consistency** and **prevent misinterpretation** by the classification model.

Duplicate Data

Dropping total of **542 duplicate rows** were identified based on a subset of key columns (excluding CustomerID) that represent each customer’s unique interaction.

Inconsistent Data

Converting unique values with the same meaning, such as **'phone'** to **'mobile phone'**, **'mobile'** to **'mobile phone'**, **'cc'** to **'credit card'**, and **'cod'** to **'cash on delivery'.**

Handling Outlier

Outliers in **10 numerical** features were **kept**, as they reflect valid customer behavior, with **robust modeling techniques** recommended to handle their impact effectively.

Data Preprocessing

Data Transformation

Robust Scaler: This scaler was used on numerical features and is resistant to outliers.

One-Hot Encoding: applied to nominal categorical features that lack intrinsic order such as **PreferredLoginDevice, PreferredPaymentMode, Gender, MaritalStatus,** and **PreferedOrderCat.**

Ordinal Encoding: used for ordinal categorical features with natural rank order, such as **CityTier** and **SatisfactionScore.**

Feature Engineering

Recency-Based Engagement Metrics: features that measure how recently a customer interacted with a business (**RecencyRatio** and **IsActiveUser).**

UnhappyCustomer: is a feature that combines complaints and low satisfaction scores to detect customers who, despite continuing to purchase, show hidden signs of churn risk due to negative sentiment.

Loyalty and Tenure Binning: **TenureGroup,** segments continuous tenure into four meaningful stages based on evolving customer loyalty and churn risk.

EXPLORATORY DATA ANALYSIS (EDA)

Details the exploratory data analysis conducted to gain business insights related to customer churn and to support the development of machine learning models.

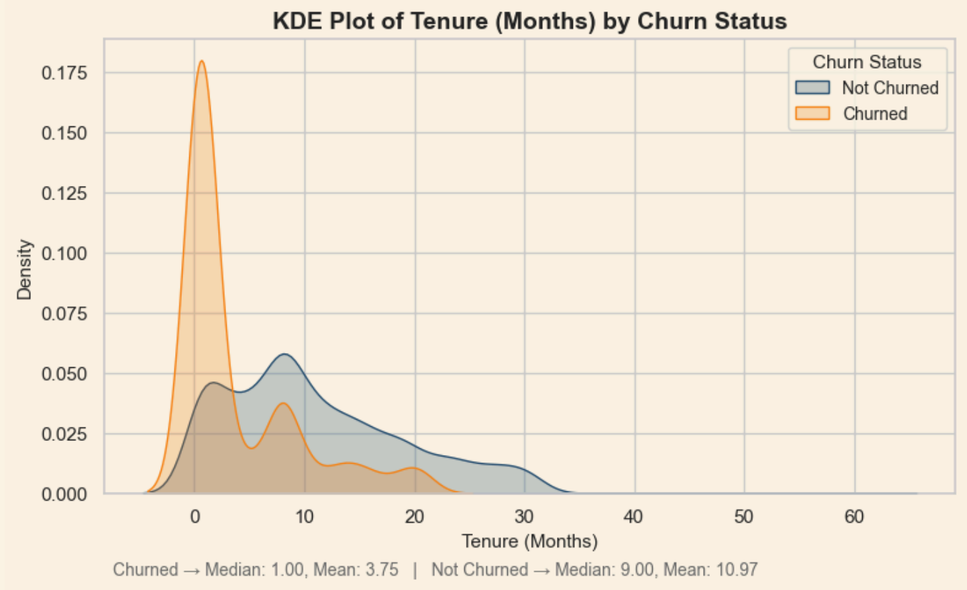
Distribution of Key Numerical Features

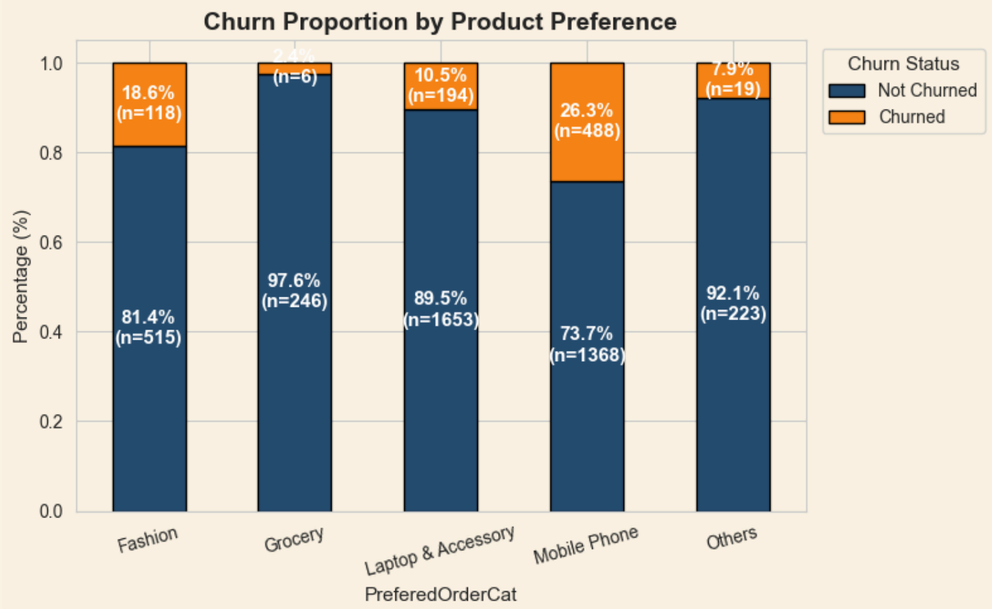
Most numerical features show non-normal, skewed distributions, especially in behavioral metrics like Order Count and Coupon Used. Due to this, we applied the Mann-Whitney U Test to assess whether these features differ meaningfully between churned and retained customers

Which Features Truly Drive Customer Churn?

Feature	Test Used	p-value	Significant?	Stakeholder Insight Summary
Gender	Chi-Square	0.6156	❌	Churn rate is statistically similar between genders.
Marital Status	Chi-Square	0	✅	Single customers are significantly more likely to churn.
City Tier	Chi-Square	0	✅	Customers in Tier 3 cities are at the highest risk of churn.
Tenure	Mann-Whitney U	0	✅	Short-tenure users churn more; long-tenure users are more retained.
Satisfaction Score	Mann-Whitney U	0	✅	Higher satisfaction scores do not consistently lower churn.
Complain	Chi-Square	0	✅	Customers who file complaints have higher churn risk.
Hour Spent on App	Mann-Whitney U	0.2907	❌	Daily time spent on the app does not statistically affect churn.
Preferred Login Device	Chi-Square	0.0008	✅	Desktop users are more prone to churn than mobile users.
Preferred Order Category	Chi-Square	0	✅	Preference for Fashion or Grocery is linked to higher churn.
Preferred Payment Mode	Chi-Square	0	✅	Cash on Delivery (COD) users show significantly higher churn rates.
Order Amount Hike (YoY)	Mann-Whitney U	0	✅	Users with increased order value tend to stay longer.
Order Count	Mann-Whitney U	0.5443	❌	Order frequency alone is not a reliable churn predictor.
Coupon Used	Mann-Whitney U	0.3649	❌	Using more coupons does not significantly impact churn status.
Cashback Amount	Mann-Whitney U	0	✅	Low cashback linked to higher churn; higher cashback improves loyalty.
Warehouse to Home	Mann-Whitney U	0	✅	Longer delivery distances increase the risk of customer churn.
Number of Devices Registered	Mann-Whitney U	0	✅	Users with more devices registered show significantly higher churn.
Number of Addresses	Mann-Whitney U	0	✅	Customers with more delivery addresses show higher churn rates.
Day Since Last Order	Mann-Whitney U		✅	Customers with longer inactivity tend to churn more frequently.

Bivariate Analysis





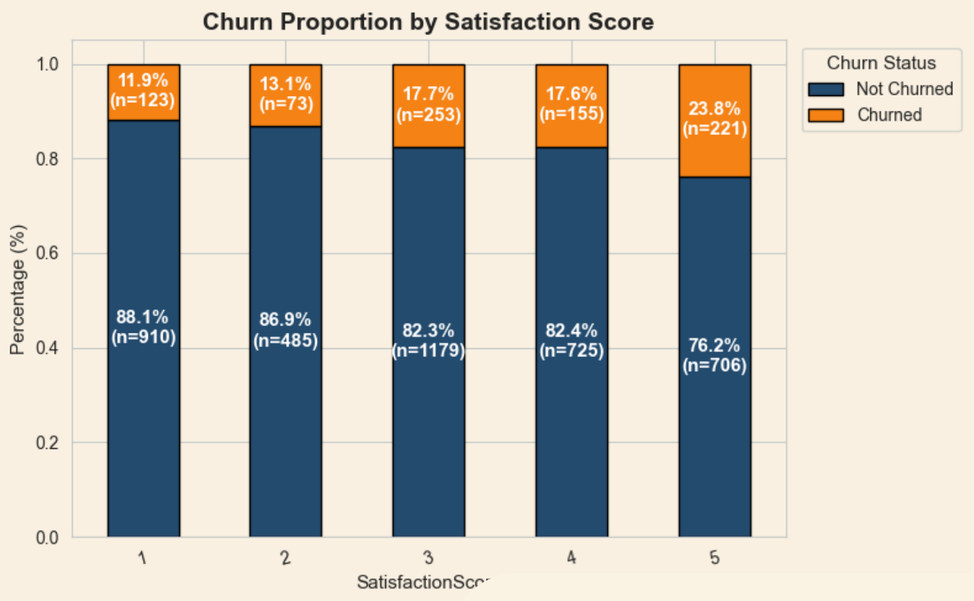
Insight

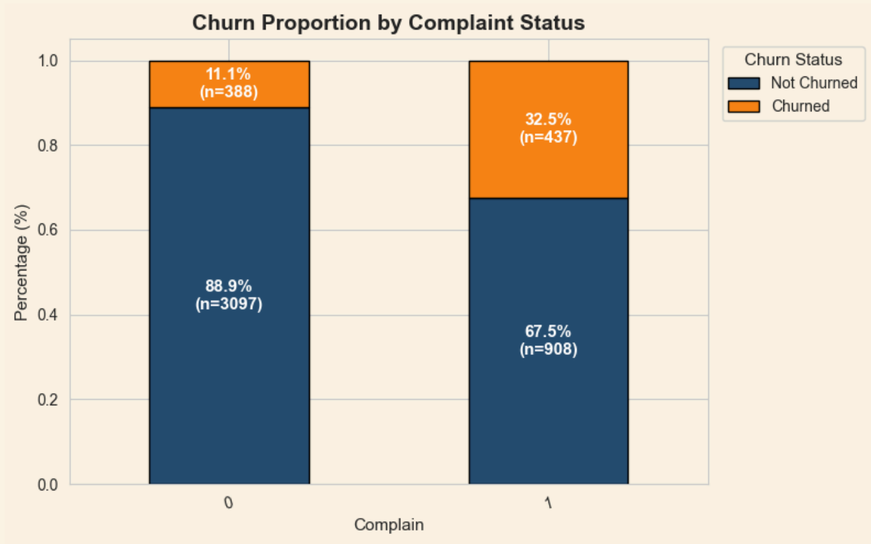
- New customers (low tenure) are much more likely to churn, confirming the need for early-stage engagement strategies.
- The Mobile Phone product category has the highest churn rate (26.3%), suggesting potential issues with customer satisfaction or loyalty in this segment.

Suggestion

- Personalized Offers: Provide early loyalty rewards, discounts, or feature unlocks to encourage continued usage.
- Exclusive Deals: Offer limited-time discounts or loyalty points for repeat purchases in the mobile category.

Bivariate Analysis





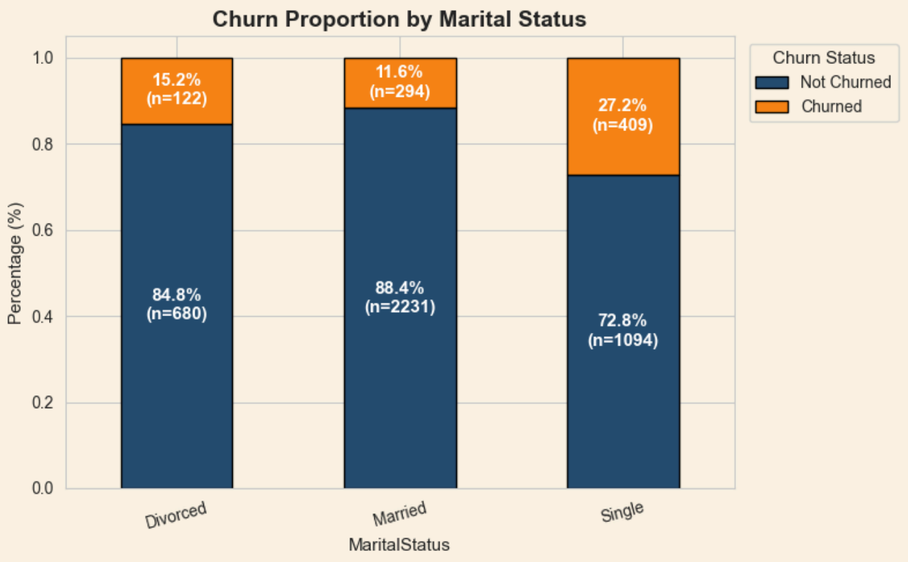
Insight

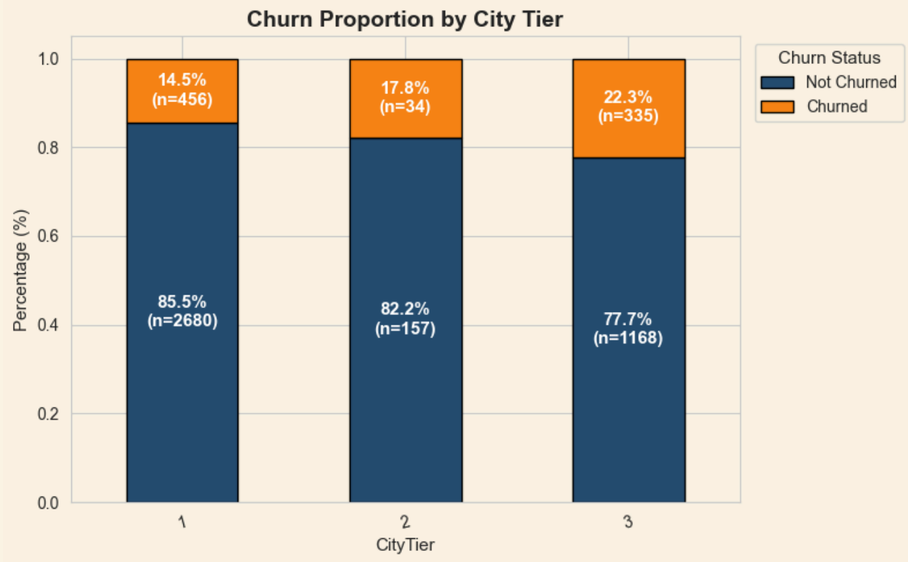
- Customers with low satisfaction scores (1–2) surprisingly show lower churn, while scores 4–5 see a sharp increase, suggesting the presence of silent churners who appear satisfied but silently disengage.
- Customers who have submitted complaints are significantly more likely to churn (32.5%) than those who haven't (11.1%), indicating that poor complaint resolution is a major churn driver.

Suggestion

- Faster Resolution SLAs: Set strict timelines for responding and resolving complaints (e.g., within 24–48 hours).

Bivariate Analysis





Insight

- Single customers have the highest churn rate (27.2%), possibly due to lower brand attachment or price sensitivity.
- Churn increases as city tier decreases; customers in Tier 3 cities show the highest churn (22.3%), likely due to logistics or service limitations in lower-tier regions.

Suggestion

- Logistics Improvements: Partner with reliable local couriers or create regional hubs to speed up deliveries and reduce service gaps.

Customer Insight

Complain Rate by Satisfaction Score

SatisfactionScore	TotalCustomers	TotalComplaints	ComplaintRate(%)
1	1033	322	31%
2	558	157	28%
3	1432	391	27%
4	880	213	24%
5	927	262	28%

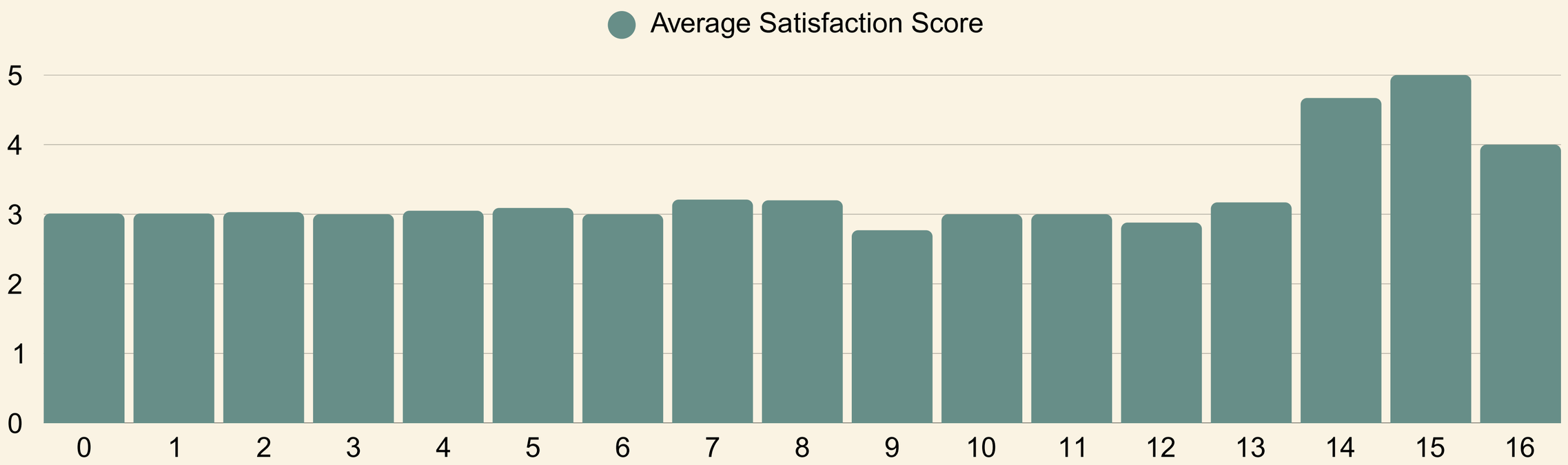
Low Satisfaction Scores Correlate with High Complaint Rates (~30%)

Score 5 unexpectedly high complaint rate may reflect rising expectations or isolated problems that need deeper investigation.

Suggestion:
Segment the Data: Break down Score 5 customers by product, region, and tenure to see where complaints cluster.

Customer Insight

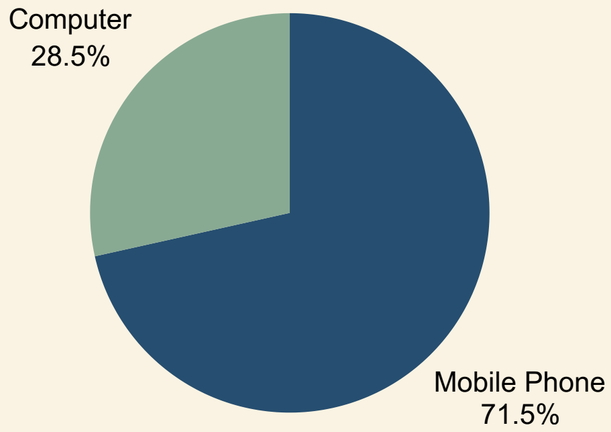
Coupon Usage Increases User Satisfaction

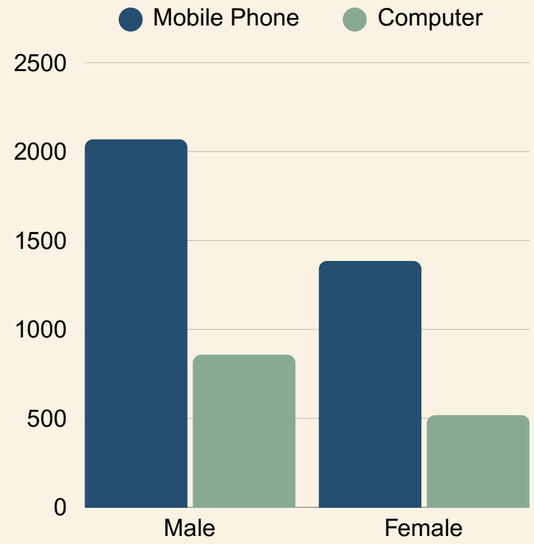


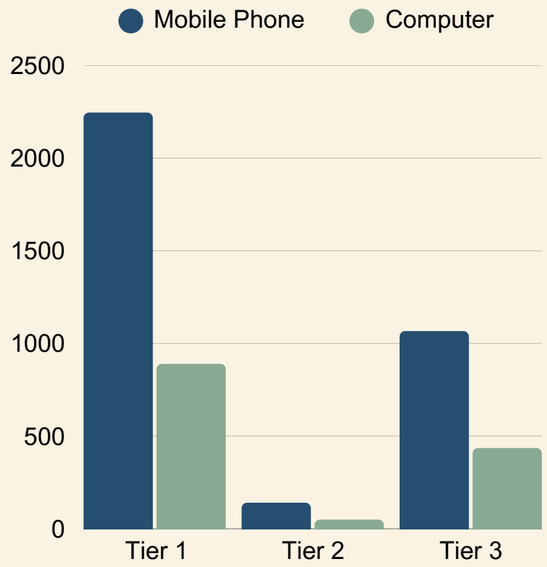
Frequent coupon users **(14–16 coupons)** show significantly higher satisfaction, indicating that promotional engagement can enhance customer experience and loyalty.

Suggestion: Give extra coupons to repeat customers to make them feel valued.

Customer Insight







Mobile Device Dominates Across All Demographics and Locations

Across all city tiers, a clear majority of users prefer Mobile Phones (around 72%), while Computers account for the remaining ~28%. This trend is consistent in City Tier 1 (71.62% vs. 28.38%), City Tier 2 (73.82% vs. 26.18%), and City Tier 3 (70.99% vs. 29.01%).

Suggestion:

Simplify Navigation For Mobile: Make product search, filtering, and checkout accessible within minimal taps.

MODELING AND EVALUATION

Focuses on developing and optimizing benchmark models for predicting customer churn in an e-commerce platform, selecting the best-performing model, and conducting diagnostics and interpretation to ensure reliable and actionable insights.

Modeling Workflow

1. Data Preprocessing
2. Model Benchmarking
3. Class Imbalance Handling
4. Hyperparameter Tuning
5. Best Model Selection & Evaluation
6. Feature Importance Analysis

DATA SPLITTING

Full cleaned dataset: 4,830 records

Benchmark Model

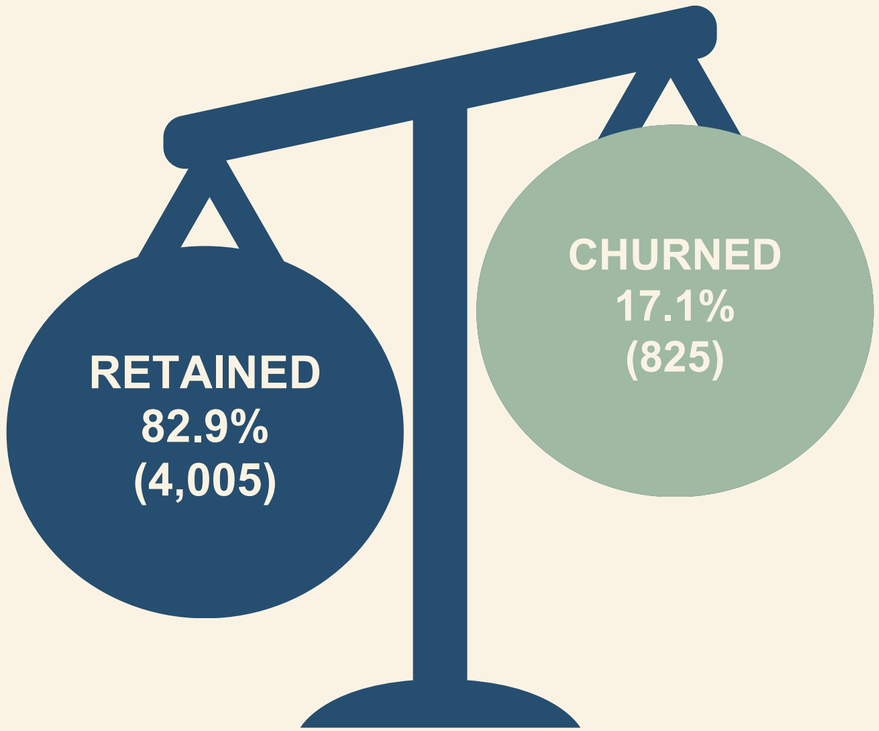
MODEL	APPROACH
Logistic Regression	Linear classification with probability outputs
K-Nearest Neighbors (KNN)	Predicts based on majority class among nearest neighbors
Decision Tree	Recursive rule-based splitting
Bagging Classifier	Ensemble of base classifiers trained on bootstrapped samples
Random Forest	Ensemble of decision trees with feature randomness
Gradient Boosting	Sequential learners correcting predecessor errors
XGBoost	Highly optimized gradient boosting
AdaBoost	Sequential model with emphasis on misclassified instances
CatBoost	Gradient boosting with native categorical support
LightGBM	Leaf-wise gradient boosting with histogram optimizations

Metrics Priority Order:

F2-Score → **Recall** → **PR-AUC** → **F1-Score** → **Precision** → **ROC-AUC** → **Accuracy**

Handling Class Imbalance

CLASS IMBALANCE



SMOTE-ENN

STEP	METHOD	BENEFIT
SMOTE	Generates synthetic churn samples using k-nearest-neighbor interpolation	Helps the model learn minority (churn) patterns
ENN	Removes samples that disagree with their nearest neighbors (noise cleanup)	Sharpens class boundaries and reduces confusion

SMOTE (Synthetic Minority Oversampling Technique)
ENN (Edited Nearest Neighbors)

Hyperparameter Tuning

Hyperparameter tuning was performed using randomized search with 5-fold cross-validation.

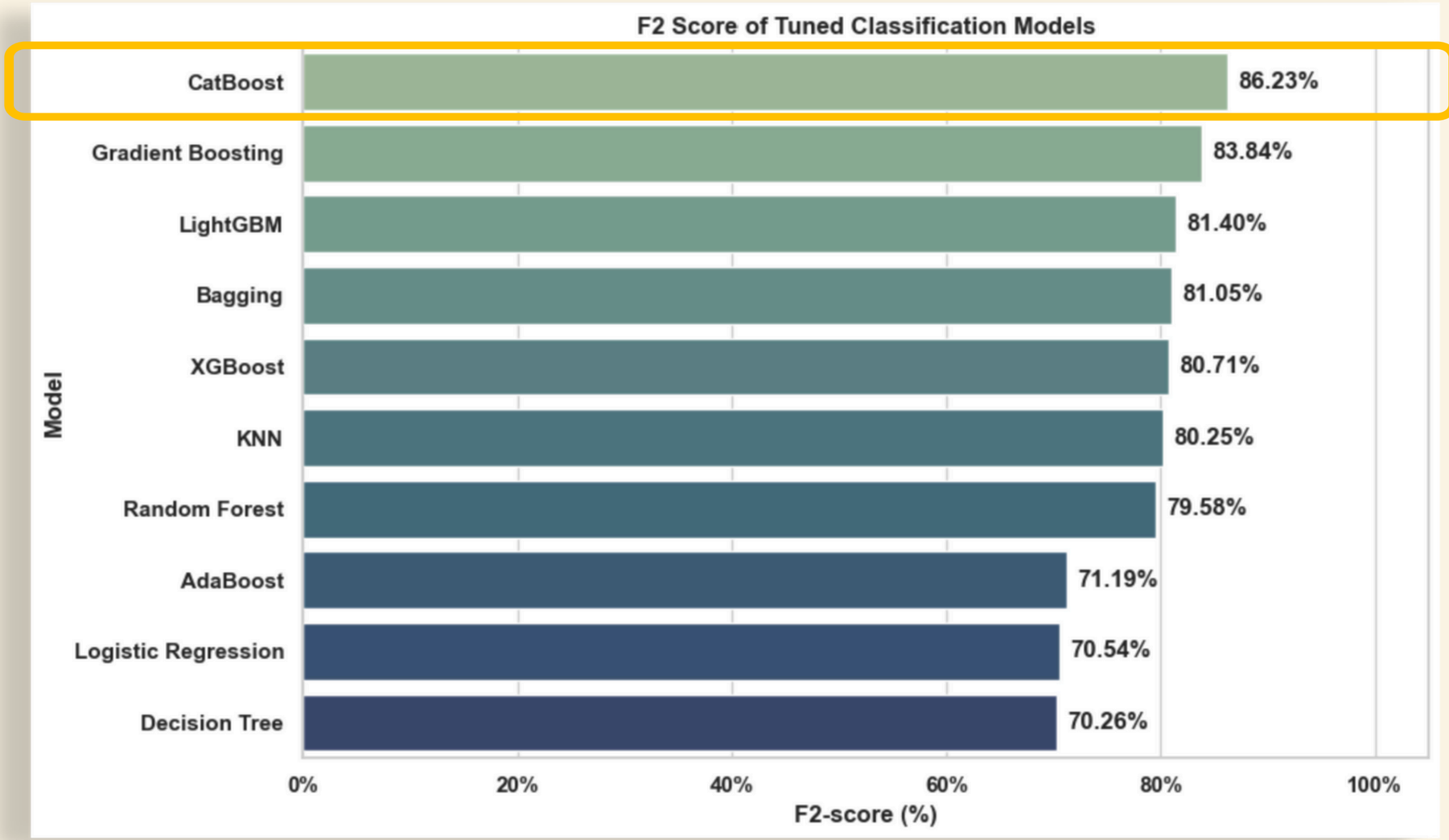
BEFORE

Model	F2-Score (%)	Recall (%)	Precision (%)
Bagging	**79.40%**	83.83%	65.70%
Gradient Boosting	**78.96%**	82.77%	66.71%
CatBoost	**78.22%**	81.58%	67.20%
LightGBM	**77.03%**	80.08%	66.87%
KNN	**76.77%**	92.96%	45.27%
XGBoost	**76.63%**	79.93%	65.84%
Logistic Regression	**70.66%**	85.18%	42.02%
Random Forest	**69.05%**	76.34%	50.00%
AdaBoost	**68.48%**	80.39%	43.00%
Decision Tree	**66.80%**	76.80%	44.31%

AFTER

Model	F2-Score (%)	Recall (%)	Precision (%)
CatBoost	**86.23%**	89.82%	74.35%
Gradient Boosting	**83.84%**	87.13%	72.84%
LightGBM	**81.40%**	84.88%	69.91%
Bagging	**81.05%**	84.88%	68.64%
XGBoost	**80.71%**	84.58%	68.24%
KNN	**80.25%**	91.47%	53.83%
Random Forest	**79.58%**	83.53%	66.91%
AdaBoost	**71.19%**	76.95%	54.80%
Logistic Regression	**70.54%**	85.03%	41.95%
Decision Tree	**70.26%**	74.85%	56.43%

Best Model Selection



CatBoost is selected as the best model (highest F2-Score).

PARAMETERS BEFORE TUNING

- iterations=100
- learning_rate=0.1
- depth=6

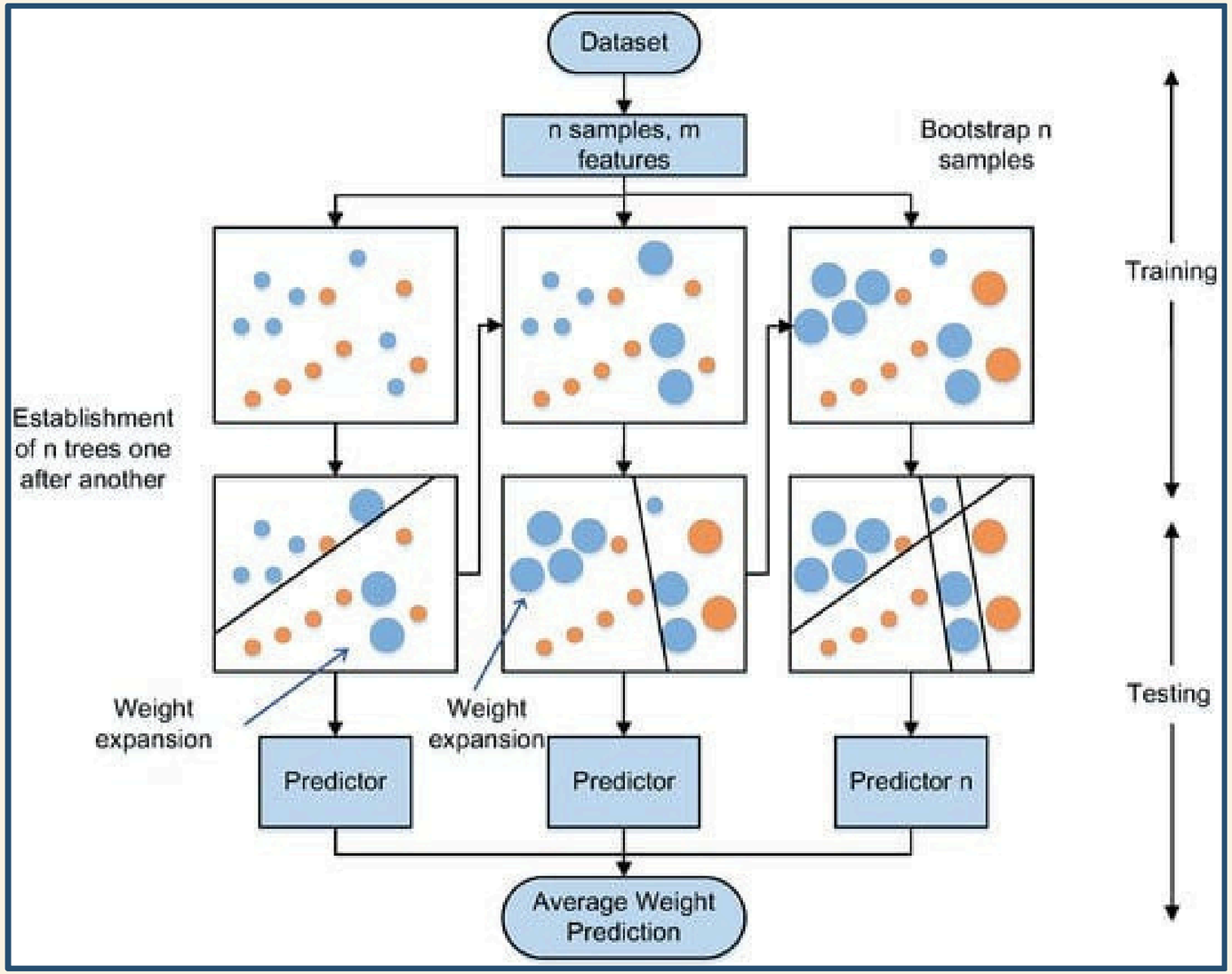
BEST PARAMETERS AFTER TUNING

- iterations=300
- learning_rate=0.1
- depth=8
- l2_leaf_reg=3
- random_strength=1
- bagging_temperature=0

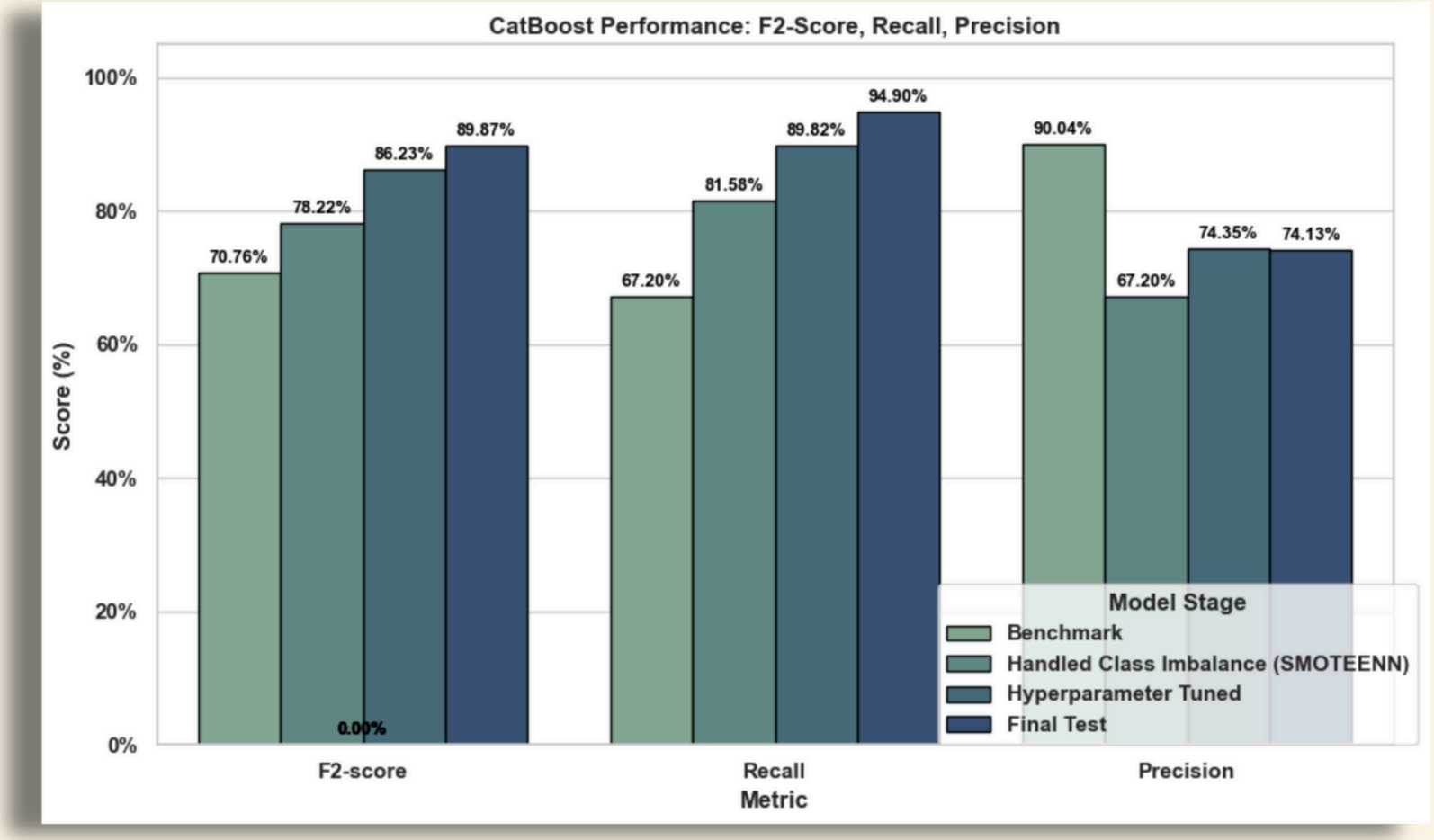
Best Model (CatBoost)

WHAT IS CATBOOST?

- **CatBoost is a modern gradient boosting algorithm**
  Builds decision trees sequentially for strong predictive performance.
- **Handles categorical features natively**
  Eliminates the need for one-hot encoding with built-in support for categorical data – CatBoost (Categorical Boosting).
- **Uses ordered boosting to prevent overfitting**
  Avoids data leakage during training, making it highly effective for churn prediction.
- **Delivers fast, accurate, and stable performance**
  Efficient training with built-in regularization and minimal preprocessing.



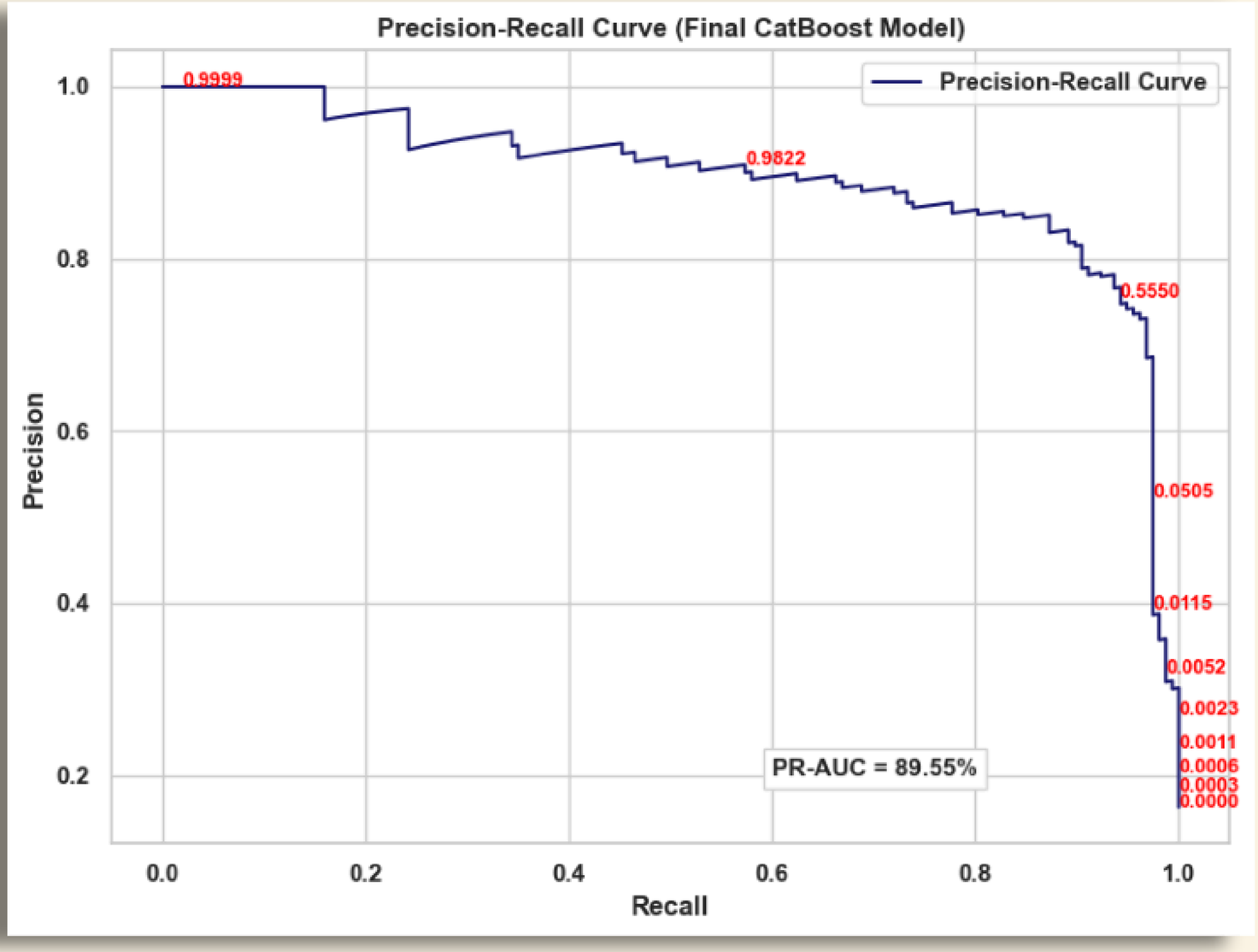
Model Evaluation



CATBOOST PERFORMANCE METRICS BY MODEL STAGE

- **F2-Score** rose from **70.76% (Benchmark) to 89.87% (Final Test),** highlighting significant gains in recall-focused optimization.
- **Recall** improved markedly from **67.20% to 94.90%,** confirming better detection of true churners.
- **Precision** slightly declined from **90.04% (Benchmark) to 74.13%,** a controlled trade-off for higher recall.
- **Key Takeaway:** Targeted class imbalance handling (with SMOTEENN) and hyperparameter tuning delivered a high-recall, deployment-ready model, crucial for minimizing undetected churn.

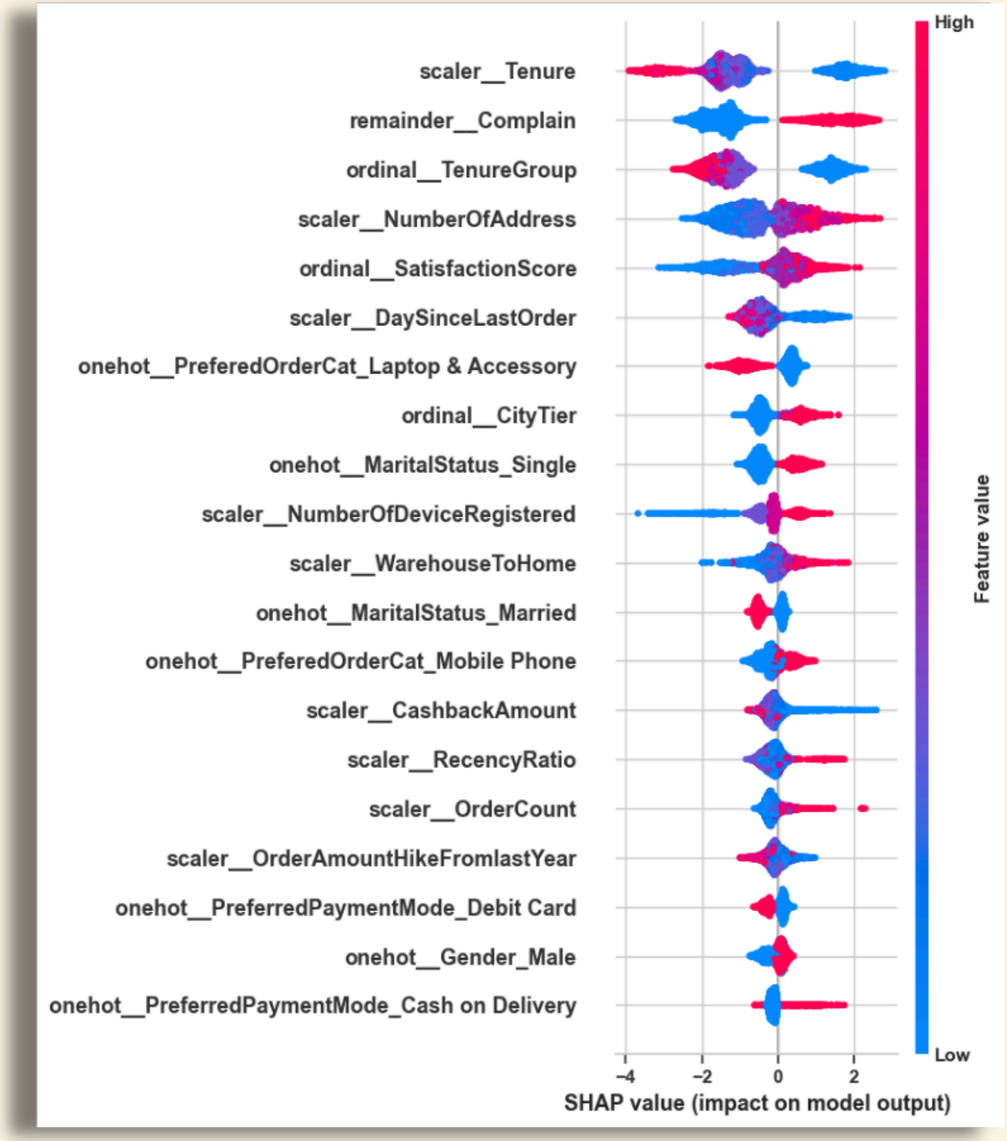
Model Evaluation



PRECISION-RECALL CURVE

- **Precision vs. Recall Curve** is used to assess performance on **imbalanced churn data.**
- Achieved a **PR-AUC of 89.55%,** reflecting strong ability to identify churners despite class imbalance.
- The curve shows high precision sustained across wide recall levels, with **minimal drop-off until ~90% recall.**
- **PR-AUC is more informative than ROC-AUC** in churn scenarios, emphasizing quality of positive predictions.
- **Insight:** High PR-AUC and curve stability confirm the model's effectiveness in capturing churners while minimizing false alarms, ideal for deployment.

Feature Important Analysis



SHAP ANALYSIS

Top Churn Drivers & Actionable Insights

1. **Short Tenure**
   - New customers are at high risk.
   - Prioritize onboarding and early engagement.
2. **Complaint History**
   - Complaints strongly signal churn risk.
   - Deploy fast recovery actions and SLA-based resolutions.
3. **Multiple Address Changes**
   - May reflect power users dropping off.
   - Improve delivery experience and address management.
4. **High Satisfaction Score**
   - Positive ratings don’t always mean loyalty.
   - Monitor behavior beyond satisfaction scores.
5. **Recent Purchase Activity**
   - Recent transactions don't prevent churn.
   - Implement post-purchase follow-ups to encourage repeat orders.

BUSINESS IMPACT

Estimates the financial implications of customer churn by conducting a cost-based retention analysis.

Retention Cost-Benefit Simulation

Objective

Compare the cost-efficiency and impact of retention campaigns with and without churn prediction using the same number of targets (256 customers).

Assumptions

- Retention cost per customer = Rp 75,000
- Customer Lifetime Value (CLV) = Rp 500,000
- Number of customers targeted in campaign = 256

Strategic Takeaway

- Same cost, but 12× higher value
- High recall (94.9%) ensures almost all churners are captured
- Significant ROI uplift supports smart, data-driven marketing decisions
- Validates real-world business impact of predictive churn modeling

Scenario A: Without Model (Manual / Random Targeting)

- Only ~20% are actual churners → 51 out of 256
- 205 loyal customers receive unnecessary promotions

💰 Retention cost: 256 × Rp 75K = Rp 19.2M
💡 Value saved: 51 × Rp 500K = Rp 25.5M
📊 Net benefit: Rp 6.3M
📈 ROI: +32.8%

Scenario B: With Model (CatBoost)

- Model predicts top 256 customers
- Captures 190 churners (recall 94.9%), with 66 false positives

💰 Retention cost: 256 × Rp 75K = Rp 19.2M
💡 Value saved: 190 × Rp 500K = Rp 95M
📊 Net benefit: Rp 75.8M
🚀 ROI: +395%

CONCLUSION AND RECOMMENDATIONS

The conclusion summarizes key findings and project limitations, while the recommendations focus on customer retention strategies and potential model enhancements.

Conclusion

- **Churn Risk:** 17.1% of customers churn, risking revenue and long-term growth.
- **Best Model:** CatBoost with F2-Score = 89.87%, Recall = 94.90%, Precision = 74.13%.
- **Impact:** Predictive targeting delivers 12× higher ROI (395%) using the same budget.
- **Key Drivers:**
  - Short tenure, complaints, complex delivery
  - Multi-device users, recent but low-frequency buyers
  - Higher risk in Tier-3 cities, COD, single male users

Project Limitation

- **No real-time features:** Behavior data (e.g., app hours) not updated live
- **No live validation:** Not tested in production or A/B campaigns
- **Platform-specific:** Model trained on one dataset, may not generalize
- **No churn feedback loop:** Can't improve based on false predictions yet

Model Recomendations

Ensure the churn model stays accurate, scalable, and business-aligned.

CRM Integration

- Deploy model in CRM for automated churn scoring
- Enable real-time targeting via scoring pipeline

Continuous Monitoring

- Retrain every 3–6 months to prevent model drift
- Update using latest transactional & behavioral signals

Outcome

Maintains high recall, minimizes false negatives, and ensures business alignment over time.

Threshold Optimization

- Keep low threshold (e.g., 0.1) to maximize recall
- Raise threshold if resources are limited to improve precision

Feature Enhancement

- Add real-time data (e.g., HourSpendOnApp, OrderCount)
- Run surveys to detect hidden churn reasons

Feedback & Testing

- Simulate campaign impact to align with business goals
- Use false positive/negative analysis to refine outreach

Business Recommendations

Use churn insights to improve retention efficiency and protect high-value users.

Segment by Churn Risk

- Prioritize high-risk users for retention actions
- Avoid over-targeting low-risk loyal customers

Target Early Risk Signals

- Focus on users with:
  - Short tenure, complaint history
  - Multiple addresses, long delivery distance
  - Multi-device usage, recent activity drop
- Offer preventive incentives before disengagement happens

Enhance Experience & Loyalty

- Personalize rewards for tech buyers & repeat users
- Offer follow-up nudges post-purchase (e.g., after first order)
- Improve UX for mobile apps and low-engagement users

Improve Fulfillment & Promotion Strategy

- Reassess coupon/cashback targeting efficiency
- Provide delivery subsidies for long-distance customers (e.g., Tier 3 cities)
- Reduce churn from Cash-on-Delivery (COD) by encouraging digital payments

Continuous Monitoring & Strategy Adjustment

- Track churn across segments over time
- Integrate churn prediction scores into monthly campaign planning
- Continuously refine marketing tactics using prediction feedback loops

MODEL DEPLOYMENT

ecommercechurnapp-groupalpha.streamlit.app

CUSTOMER DETAILS

Customer Engagement

Tenure (months)

12

Tenure Group: MidTerm

Hours on App/Week

1.00

0.00 5.00

Devices Registered

2

Days Since Last Order

10

Recency Ratio: 0.77

Active User: Yes

Customer Profile

City Tier

1

Gender

Female

Marital Status

Single

Behavioral Preferences

Login Device

Mobile Phone

Payment Mode

Debit Card

E-Commerce Churn Prediction App

This application estimates the likelihood that a customer will **churn**, meaning they may stop using the platform or making purchases.

Instructions: Enter customer details in the sidebar and click **Predict** to assess churn risk.

Predict

About This App

Feature Explanations

Contributors

THANK YOU

PROJECT LINKS

Github: E-Commerce Customer Churn Analysis and Prediction
Streamlit: E-Commerce Churn Prediction App
Tableau Dashboard: E-Commerce Churn Analysis Dashboard