# Table of Content


I. Introduction & Business Context 2
II. Tools & Libraries 2
III. Data Quality & Preprocessing 3
IV. Exploratory Data Analysis Techniques 3
V. Analytical Methods & Validation 4
    1. Statistical Analysis 4
    2. Machine Learning Models 4
    3. Customer Segmentation (Unsupervised Learning) 4
VI. Key Findings from Business Questions 5
    1. Which membership tiers churn more? 5
    2. Do churn rates differ by gender, and how is each churn group composed by gender? 5
    3. Do customers with higher average transaction value churn more frequently? 6
    4. Do customers with higher wallet balances churn less? 7
    5. Is there a churn difference between customers with zero vs. non-zero wallet points? 8
    6. Does higher login frequency correlate with lower churn risk? 9
    7. Which preferred offer type is best for customer satisfaction? 10
    8. Do customers with past complaints churn more? 11
    9. Which features (customer behaviors) have the most impact on churn vs not-churn? (RandomForest + SHAP value) 13
    10. How many separate customer groups based on their behaviors? (Kmean Clustering) 14
VII. Business Recommendations 16
    1. Retention & Loyalty Programs 16
    2. Engagement Initiatives 16
    3. Wallet Incentives 16
    4. Targeted Marketing by Segment 17
    5. Complaint Management 17
VIII. Conclusion 17
IX. References 18

# I. Introduction & Business Context

Our client is a B2B SaaS company operating on a subscription-based revenue model. In such businesses, recurring revenue is the main growth engine, and customer **churn** (customers discontinuing their subscription) poses a direct threat to profitability.

- **Business Problem:** High churn reduces customer lifetime value and increases customer acquisition costs.
- **Objective:** To identify churn drivers, detect high-risk customer groups, and recommend targeted retention strategies.
- **Dataset:** The company provided a dataset (saas_customer_churn.csv) with ~37,000 customer records, including demographic information (age, gender, region), behavioral data (login frequency, time spent), financial activity (transaction value, wallet balance), service interactions (complaints, resolutions), and churn risk scores.

This project applies an **explanatory data analysis pipeline** to uncover insights and guide data-driven decision-making.

# II. Tools & Libraries

The analysis was conducted in **Python (Jupyter Notebook)**, using the following key libraries:

- Data Handling & Preprocessing
  - pandas, numpy: data manipulation, missing values, transformations.
  - datetime, re: date/time parsing and cleaning irregular strings.
- Exploratory Data Analysis (EDA)
  - matplotlib, seaborn: visualizations such as bar charts, box plots, histograms, KDE density plots.
  - PercentFormatter: formatting proportions on charts.
- Machine Learning Models
  - scikit-learn: train/test splitting, RandomForest classifier, pipelines, clustering with KMeans, PCA dimensionality reduction.
  - shap: interpretability and feature importance analysis.
- Clustering & Segmentation
  - KMeans: customer segmentation.

    - silhouette_score: validation of clustering quality.

These libraries allow a **full pipeline**: from cleaning raw data → exploratory analysis → predictive modeling → interpretability → customer segmentation.

# III. Data Quality & Preprocessing

Initial Audit Findings

- Missing values in key features: region (~15%), wallet points (~9%), preferred offers (~1%).
- Inconsistent categorical values (e.g., “Yes/No/True/Unknown” variations).
- Numeric errors: negative wallet balances, invalid login days, strings like “Error” in numeric fields.
- Outliers: very high transaction values (up to ~100,000).
- Mixed date formats (dd-mm-yyyy, yyyy-mm-dd, dd/Mon/yyyy).

Preprocessing Steps

- **Drop high-cardinality IDs:** security_no and referral_id removed as not useful.
- **Standardize categorical variables:** Normalized Yes/No values, replaced ?, XXXXXXX, Q with “Unknown.”
- **Dates:** Converted joining_date to ISO format (YYYY-MM-DD).
- Numeric corrections:
    - Removed negative balances.
    - Converted “Error” values to zero in login frequency.
    - Removed $ signs from transaction values.
- **Missing value imputation:** Median used for numeric fields; mode/“Unknown” for categorical.
- **Outlier handling:** Values capped at the 99th percentile.
- **Final dataset:** 36,992 records × 22 cleaned columns. No missing values remained.

# IV. Exploratory Data Analysis Techniques

A combination of **statistical summaries, visualization techniques, and cross-tabulation** was applied.

- Univariate Analysis

    - Distribution of age, transaction values, wallet balances.
    - Outlier detection using describe() and histograms.
- Bivariate Analysis
    - Churn vs membership category (stacked bar plots).
    - Churn vs gender (nested donut chart).
    - Churn vs spending (KDE density plot).
    - Churn vs wallet balance (boxplots).
- Multivariate Analysis
    - Correlation analysis using Spearman's rank correlation.
    - Feature importance using RandomForest + SHAP.
    - KMeans clustering with PCA projection for visualization.

# V. Analytical Methods & Validation

1. Statistical Analysis

    - Descriptive statistics: mean, median, variance, percentiles for numeric fields.
    - Outlier analysis: values capped at 99th percentile.
    - Frequency tables for categorical features.

2. Machine Learning Models

- **RandomForest Classifier:** Built on numerical + encoded categorical features.
- Validation:
    - Accuracy and interpretability prioritized over complex models.
    - **SHAP values** used to interpret feature importance.

    Key Insights from SHAP:

    - Wallet balance → strongest predictor (higher = lower churn).
    - Membership tier → Premium/Platinum reduce churn, No/Basic increase churn.
    - Transaction value and login frequency → moderate importance.
    - Demographics (age, gender) → minimal effect.

3. Customer Segmentation (Unsupervised Learning)

    - **KMeans clustering** with standardized features.
    - Validation methods:

- **Elbow method** → optimal clusters between k=4 and k=5.
- **PCA visualization** → clusters displayed clearly in 2D space.

# VI. Key Findings from Business Questions

1. Which membership tiers churn more?

- **Visualization:**

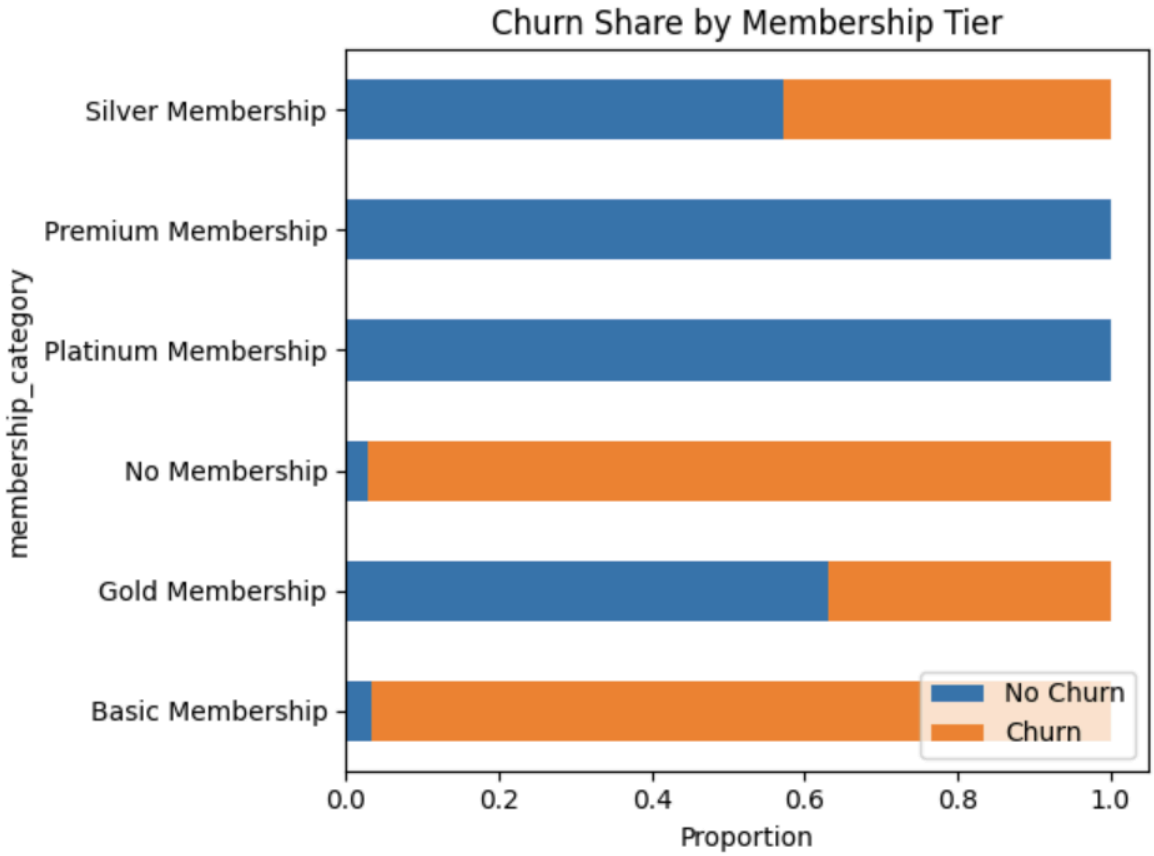


*Figure 1: Membership and Churn chart*

- **Result:**
  - Entry-level (No/Basic/Silver) customers have the highest churn. Premium/Platinum customers show strong loyalty.

2. Do churn rates differ by gender, and how is each churn group composed by gender?

- **Visualization:**

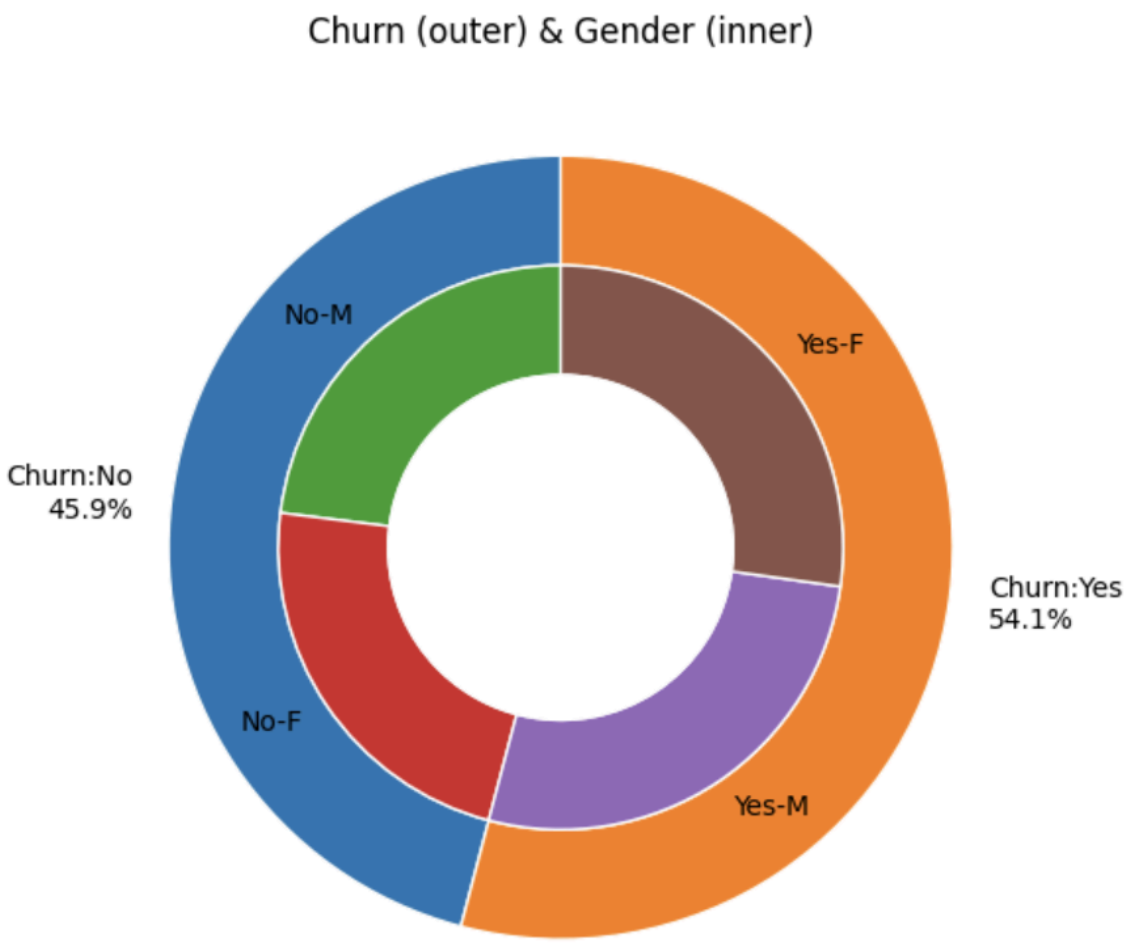


*Figure 2: Gender and Churn chart*

- **Result:**
    - EntryChurn is similar across male and female customers. Gender is not a strong predictor.

3. Do customers with higher average transaction value churn more frequently?

- **Visualization:**

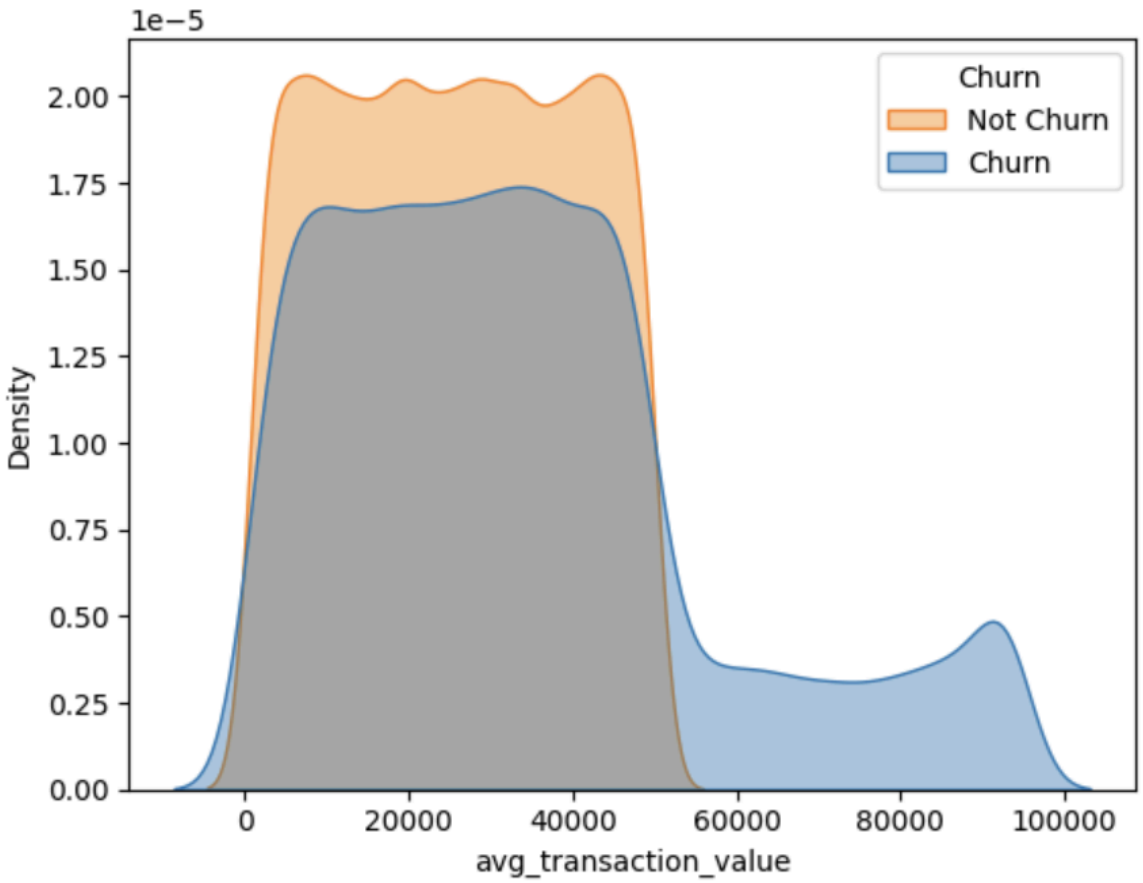


*Figure 3: Average Transaction Value and Churn chart*

- **Result:**
    - High spenders show higher churn sensitivity — potential dissatisfaction with pricing.

4. Do customers with higher wallet balances churn less?

- **Visualization:**

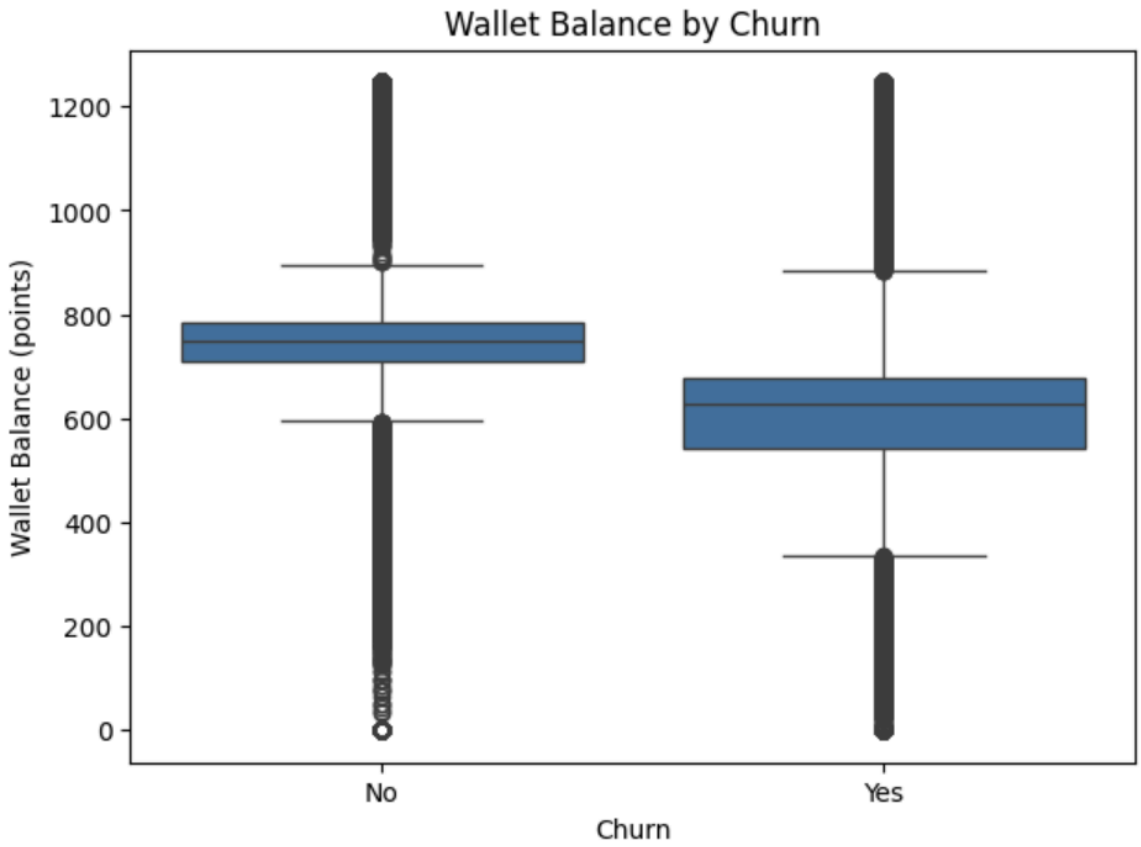


*Figure 4: Wallet Balance and Churn chart*

- **Result:**
    - Higher balances strongly correlate with reduced churn.

5. Is there a churn difference between customers with zero vs. non-zero wallet points?

- **Visualization:**

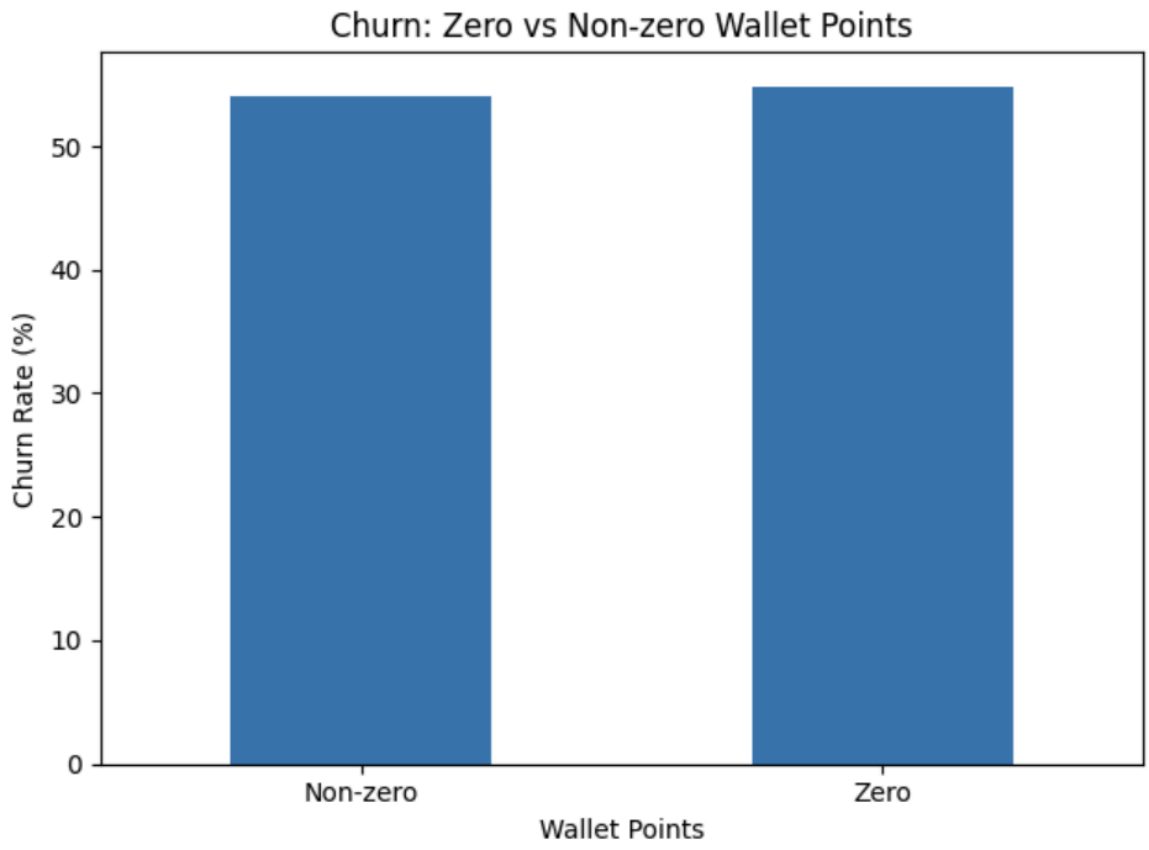


*Figure 5: Zero and Non-Zero Wallet Point chart*

- **Result:**
    - Little impact; balance size matters more.

6. Does higher login frequency correlate with lower churn risk?

- **Visualization:**

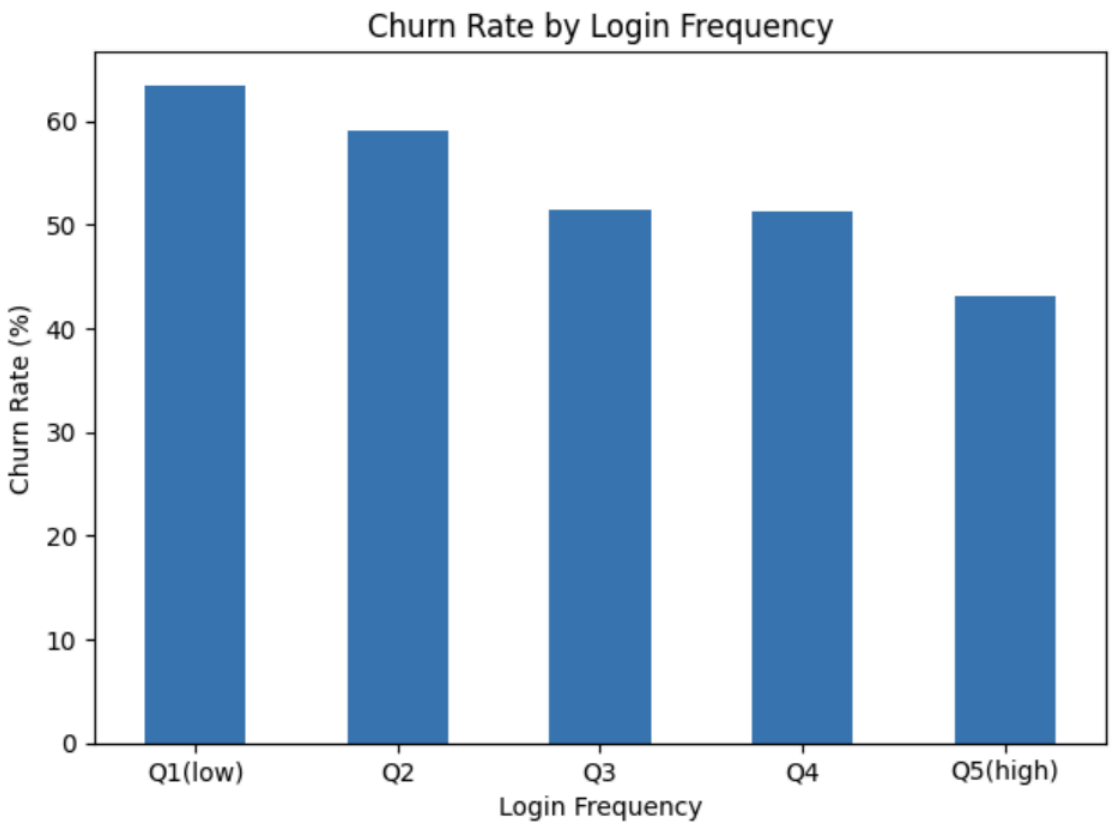


*Figure 6: Login Frequency and Churn chart*

- **Result:**
    - Strong negative correlation with churn. Frequent users churn less.

7. Which preferred offer type is best for customer satisfaction?

- **Visualization:**

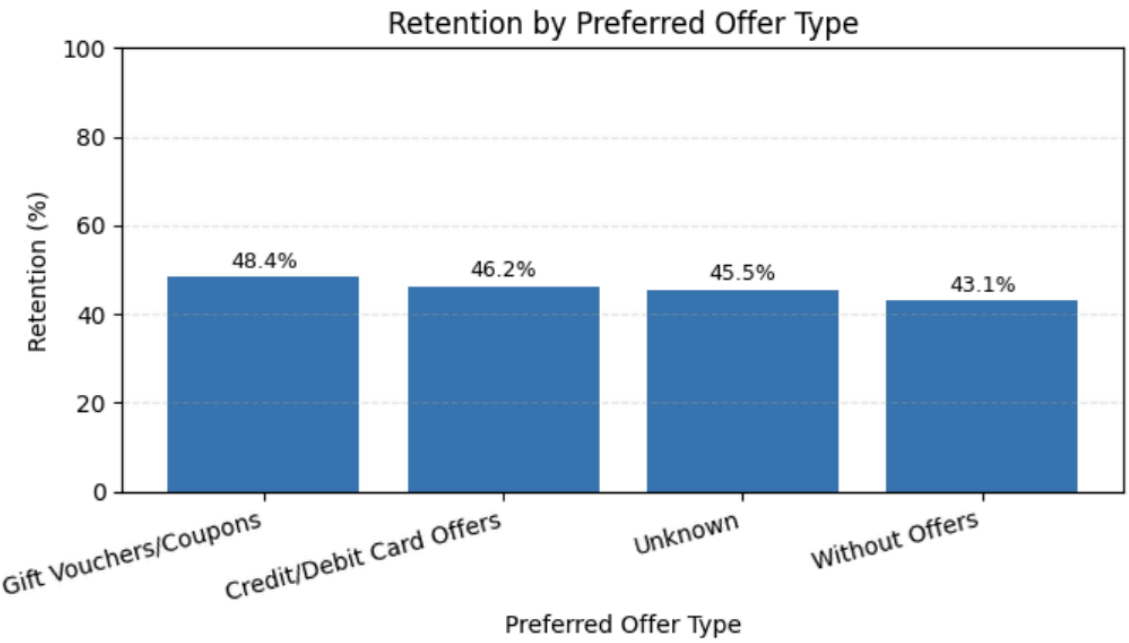


*Figure 7: Preferred Offer Type and Churn chart*

- **Result:**
  - Gift vouchers perform slightly better, though the effect is modest.

8. Do customers with past complaints churn more?

- **Visualization:**

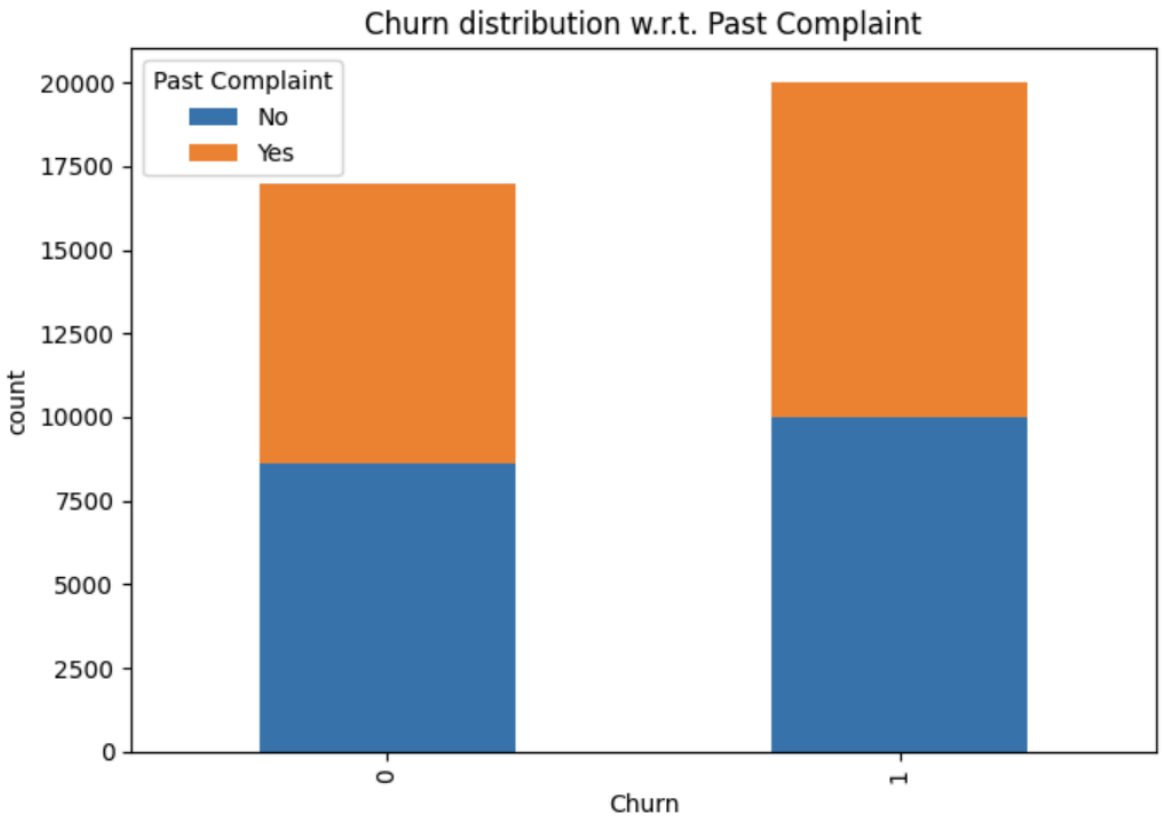


*Figure 8.1: Past Complaints Numbers and Churn chart*

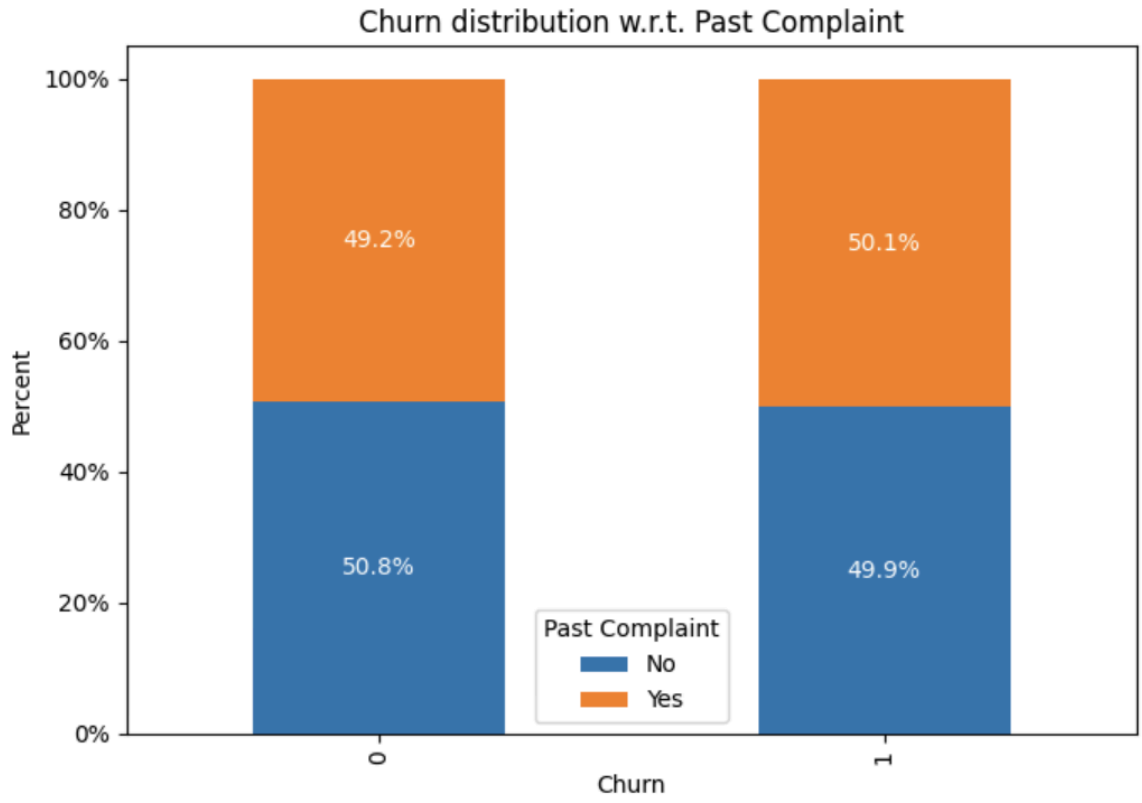


*Figure 8.2: Past Complaints Percentages and Churn chart*

- **Result:**
    - Do not drive churn; quick service recovery neutralizes risk.

9. Which features (customer behaviors) have the most impact on churn vs not-churn? (RandomForest + SHAP value)

- **Visualization:**

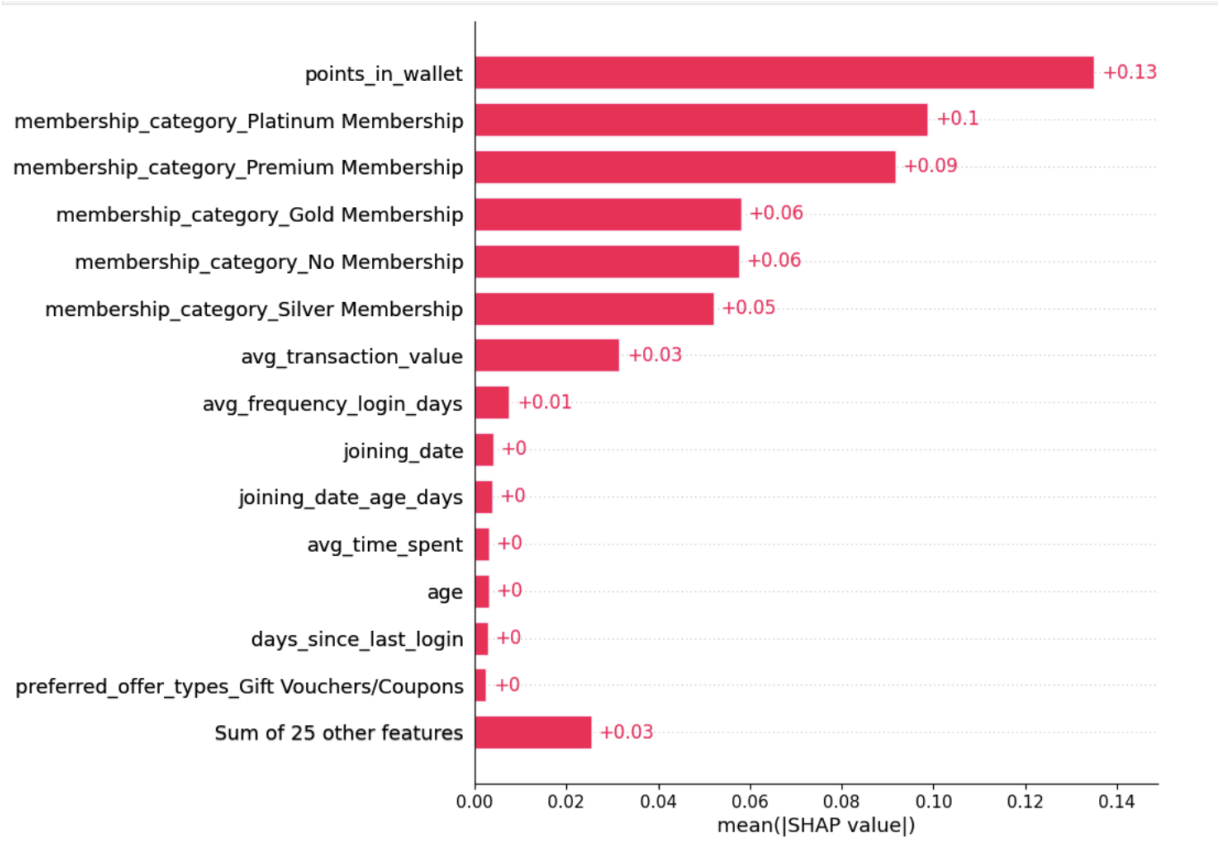


*Figure 9: Features (Customer Behaviors) have the most impact on Churn vs Not-Churn*

- **Result:**
  - Wallet balance, membership tier, and engagement metrics dominate churn prediction.

10. How many separate customer groups based on their behaviors? (Kmean Clustering)

- **Visualization:**

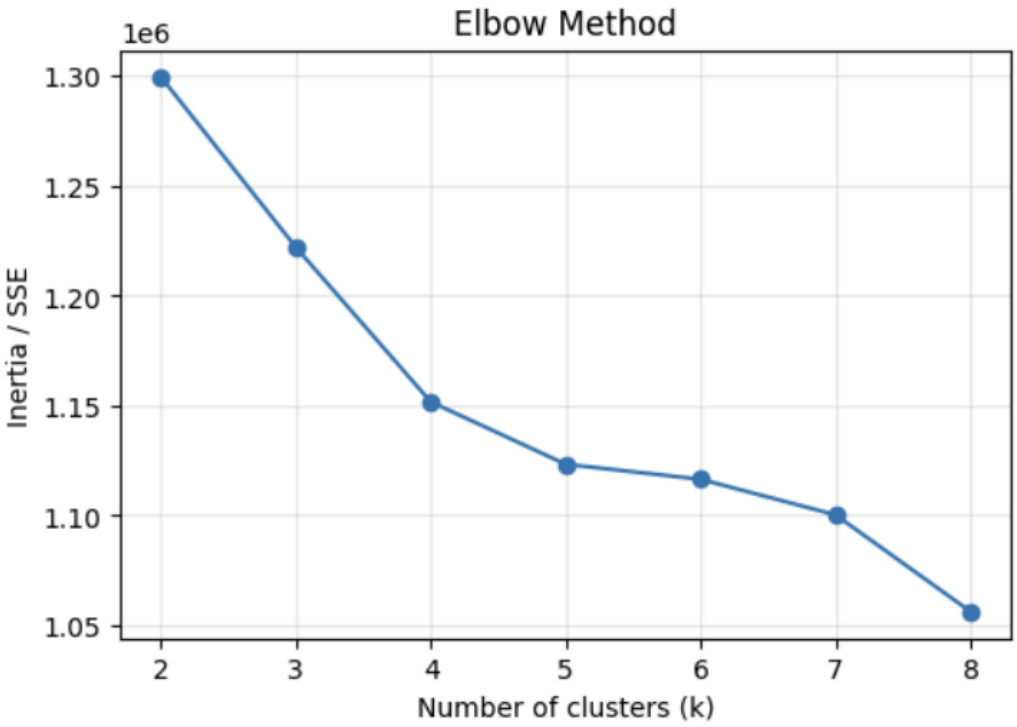


*Figure 10.1: Elbow Chart for optimum-k*

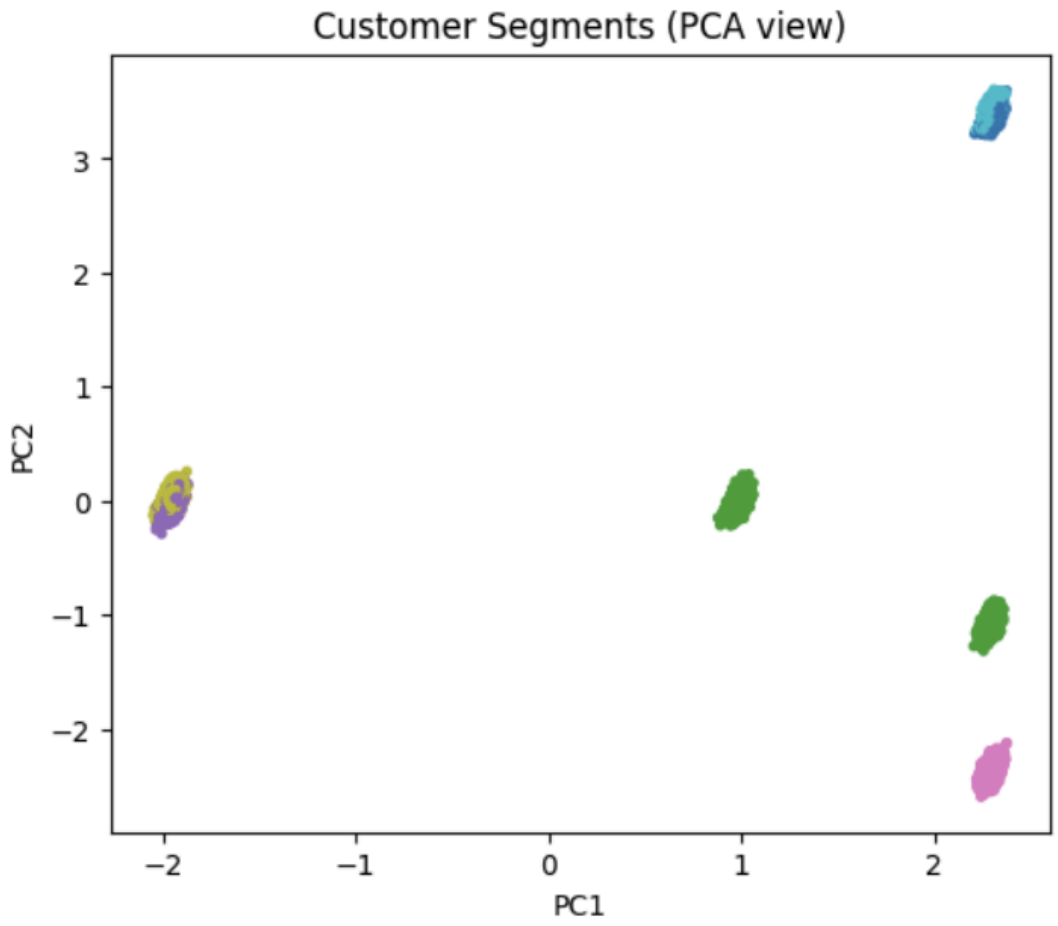


*Figure 10.2: Customer Clustering with k=5*

- **Result:**
  - Five meaningful customer groups with different behaviors and retention strategies.

# VII. Business Recommendations

1. Retention & Loyalty Programs
   - Encourage upgrades from No/Basic to higher tiers.
   - Protect Premium/Platinum members with exclusive rewards.
2. Engagement Initiatives
   - Gamification, reminders, and app campaigns to increase logins.
   - Special reactivation offers for low-frequency users.
3. Wallet Incentives
   - Introduce cashback, top-up bonuses, and loyalty points to raise balances.

- Focus on wallet-linked rewards for churn-prone groups.

4. Targeted Marketing by Segment
    - Cluster 0: Mobile, frequent users → push app-based offers.
    - Cluster 1: Follow-up resolution seekers → highlight customer service quality.
    - Cluster 2: Older loyal customers → build long-term retention programs.
    - Cluster 3: Offer-active users → upsell through time-limited promotions.
    - Cluster 4: Younger, discount-light customers → emphasize innovative bundles.
5. Complaint Management
    - Maintain fast SLA and follow-up procedures.
    - Use complaints as a loyalty touchpoint.

# VIII. Conclusion

This project demonstrates the value of combining **data quality checks, exploratory analysis, machine learning, interpretability, and clustering** to understand customer churn.

- Strengths of analysis:
    - Rigorous cleaning and preprocessing.
    - Multiple techniques (visual EDA, statistical analysis, ML, clustering).
    - Interpretability through SHAP values.
    - Actionable segmentation into 5 customer groups.
- Limitations:
    - Dataset provides churn risk scores rather than historical churn events.
    - External factors (market competition, pricing strategy) are not included.
    - Advanced model tuning was not the main focus.

**Final Takeaway:**
Customer churn can be reduced by focusing on:

- Upgrading entry-level members.
- Boosting wallet balances with incentives.

- Increasing engagement via logins.
- Maintaining strong service recovery.
- Deploying cluster-based personalized marketing.

These insights translate directly into **data-driven strategies** for marketing, product, and customer success teams, enabling the SaaS company to improve retention and long-term revenue.

# IX. References


1. Drive Link:

   https://drive.google.com/drive/folders/1uCo6YthRRWhnZnL-gSexny5WuyBnvwUP?usp=sharing

2. Github:

   https://github.com/tuanTaAnh/SaaS-Customer-Churn-Consulting

3. Video Link:

   https://drive.google.com/file/d/1yNLPKje8U1GLvebs7fPb5hQJMwDPIC5D/view?usp=sharing


*(*) You only need to access one of 2 links above for code file (Drive Link or Github)*
*(**) Because of my OS (MacOS), some installation instructions in Readme could cause problems for you if you are using Windows or Ubuntu.*