CUSTOMER RETENTION ANALYSIS USING LOGISTIC REGRESSION IN PYTHON

1. Background and Problem Statement

A mid-sized U.S.-based subscription box company offering monthly wellness and lifestyle products faced stagnant growth and rising customer churn. The management team lacked clarity on which customer behaviors or traits were driving cancellations. The company sought a data-driven solution to identify key predictors of customer churn and proactively improve retention strategies. Python was selected for this analysis due to its flexibility in handling logistic regression and visualizing binary outcomes.

2. Objectives

- To predict customer churn based on behavioral and demographic features
- To identify statistically significant variables that influence the likelihood of churn
- To interpret logistic regression coefficients to inform retention strategy
- To produce a report with clear visualizations and recommendations for stakeholder action

3. Methodology

3.1 Data Description

- Sample Size: 6,000 active and churned customers
- Features included:
    - Demographic: Age, gender, location
    - Behavioral: Number of months subscribed, average monthly spend, engagement score, complaints filed
    - Target Variable: Churned (1) or Active (0)

3.2 Data Preprocessing

- Null handling with mean imputation and mode filling
- Binary encoding of gender and location
- Standardization of continuous variables
- Outlier detection using IQR for spend and complaints

- Feature scaling with StandardScaler from sklearn

3.3 Logistic Regression Modeling

- Model built using LogisticRegression() from sklearn.linear_model
- Checked multicollinearity using Variance Inflation Factor (VIF)
- Evaluated model accuracy using metrics:
    - Accuracy
    - Precision
    - Recall
    - F1 Score
    - ROC-AUC

3.4 Model Interpretation

- Odds ratio interpretation of coefficients
- Features with the strongest predictive power:
    - Negative: Higher engagement scores and longer subscription tenure
    - Positive: High complaint count and low average spend

3.5 Validation

- 70/30 Train-Test Split
- k-Fold Cross-validation (k=5) to confirm generalizability
- ROC Curve and AUC visualized using matplotlib and seaborn

4. Key Findings

- Churn rate in sample: 32.6%
- Complaints filed were the most influential predictor (odds ratio: 3.9)
- Customers with fewer than 3 months tenure had 2.2x higher churn probability
- Engagement score negatively correlated with churn (r = -0.63)
- Model achieved 84% accuracy and AUC of 0.88

5. Actionable Recommendations

- Implement early engagement campaigns in the first 90 days
- Launch automated support workflows for complaint handling
- Create a predictive churn dashboard for real-time alerting
- Offer targeted discounts to low-engagement segments

6. Deliverables

- Python Jupyter notebook with all steps
- .pkl file for logistic regression model
- Visuals: Confusion matrix, feature importance plot, ROC curve
- PDF report for non-technical stakeholders summarizing analysis and business implications

7. Stakeholder Relevance

Academic:

- Illustrates the use of logistic regression in real-world customer analytics
- Demonstrates binary classification modeling and interpretation using Python

Corporate:

- Provides a replicable framework for churn modeling
- Suitable for SaaS, subscription, or recurring revenue-based businesses