رواد مصر الرقمية

SALES FORECASTING AND DEMAND PREDICTION

Members

01 Dina Magdy Tolba

02 Dina Mohsen Amin

03 Merna Ashraf Gaied

04 Fadwa Mohamed Mahmoud

05 Noha Mostafa Sayed

06 Heba Abdelsalam Abdelmotalb

Agenda

Project Overview

In this project, we're trying to predict whether a customer will churn after his/her subscription expires. A typical subscription lasts 30 days. The goal of this project is to predict whether a customer will renew his/her subscription within 30 days of the previous one expiring.

Dataset Description

- Source: Kaggle – Customer Retention Datathon (Riyadh Edition)
- Size: 2.02 GB
- Service: Music streaming platform

Goal:

Predict renewal within 30 days after membership expiration

User Behavior:

- Auto-renew or manual re-subscription
- Cancellation doesn’t always mean churn
- Churn: No valid subscription within 30 days

Dataset Overview

File	Description
Train_data.csv	Main training set, containing user IDs (msno) and target churn label (is_churn).
Members.csv	User demographic details: age (bd), city, gender, registration method, and registration date.
Transactions.csv	Main training set, containing user IDs (msno) and target churn label (is_churn).
User_logs.csv	Daily listening behavior: number of songs played by completion percentage and total seconds listened.
kaggle_test_data.csv	Test set for prediction, providing only user IDs (msno) to generate churn probabilities.

Pre-processing

1- Handling Missing Values:

- We found missing values in the gender column and replaced them with the value "other" to maintain consistency.
- The bd (birthdate) column had unrealistic or missing values, which were also handled appropriately.

2-Handling Age Outliers:

- Outliers were detected in the age column where ages were less than 13 or greater than 99. These unrealistic ages were considered outliers and replaced with the median of valid age values (between 13 and 99).

Pre-processing

3- Aggregating Log Data:

- User activity logs were aggregated by user (msno), including total counts of songs played and average daily activity.

4-Feature Engineering:

- New feature(active_days) was calculated as the number of unique days a user was active.
- An (age_bucket) feature was created using age ranges to support visualization and analysis.

Pre-processing

5- Merging Data Sources:

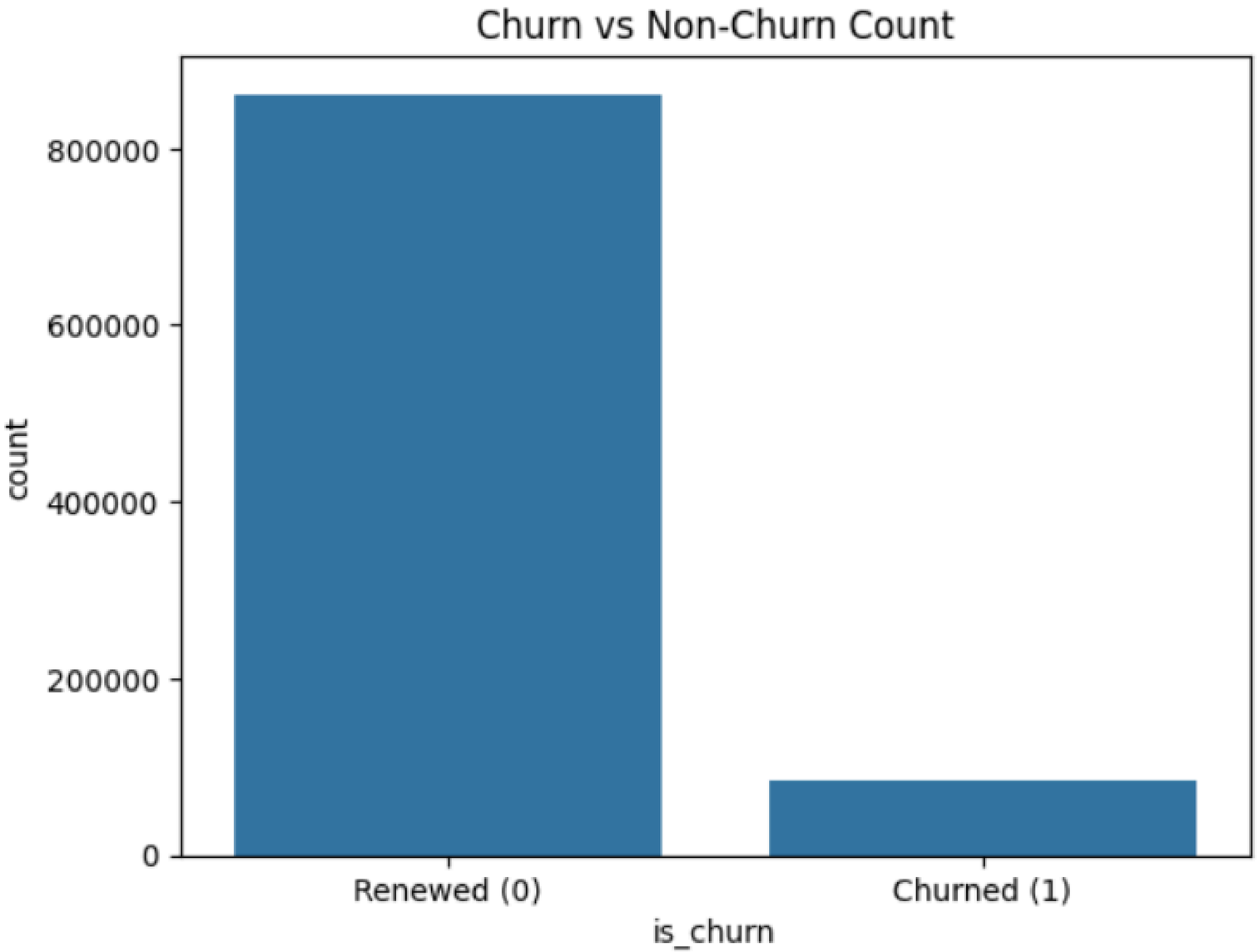
- The main (train_data) was merged with cleaned (members), aggregated (user_logs), and latest (transactions) data using msno as a key.

These preprocessing steps ensured that the dataset was clean, consistent, and ready for exploratory data analysis and model training.

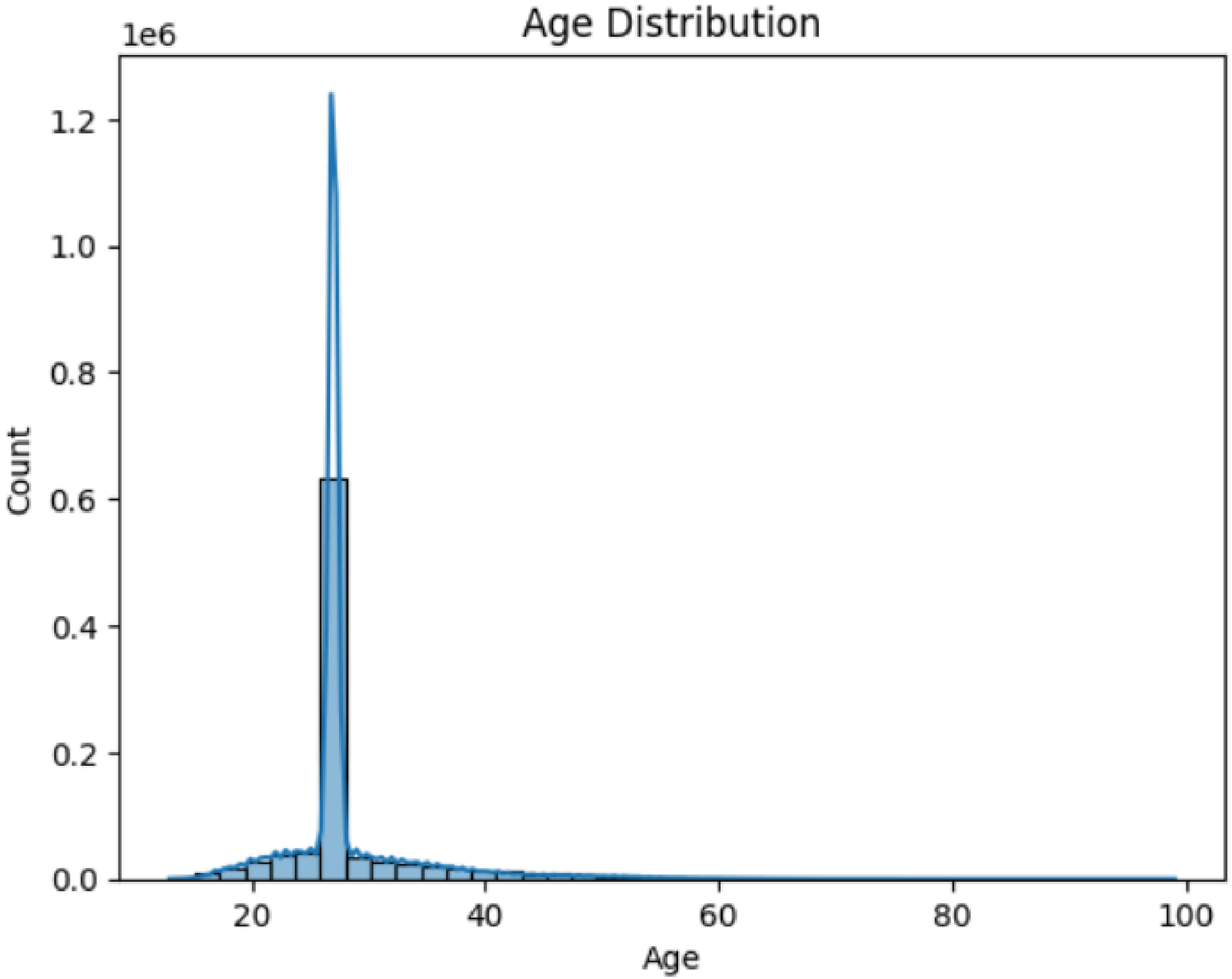
Exploratory Data Analytics

A sample of the analyses done on the dataset to understand the data in greater detail.

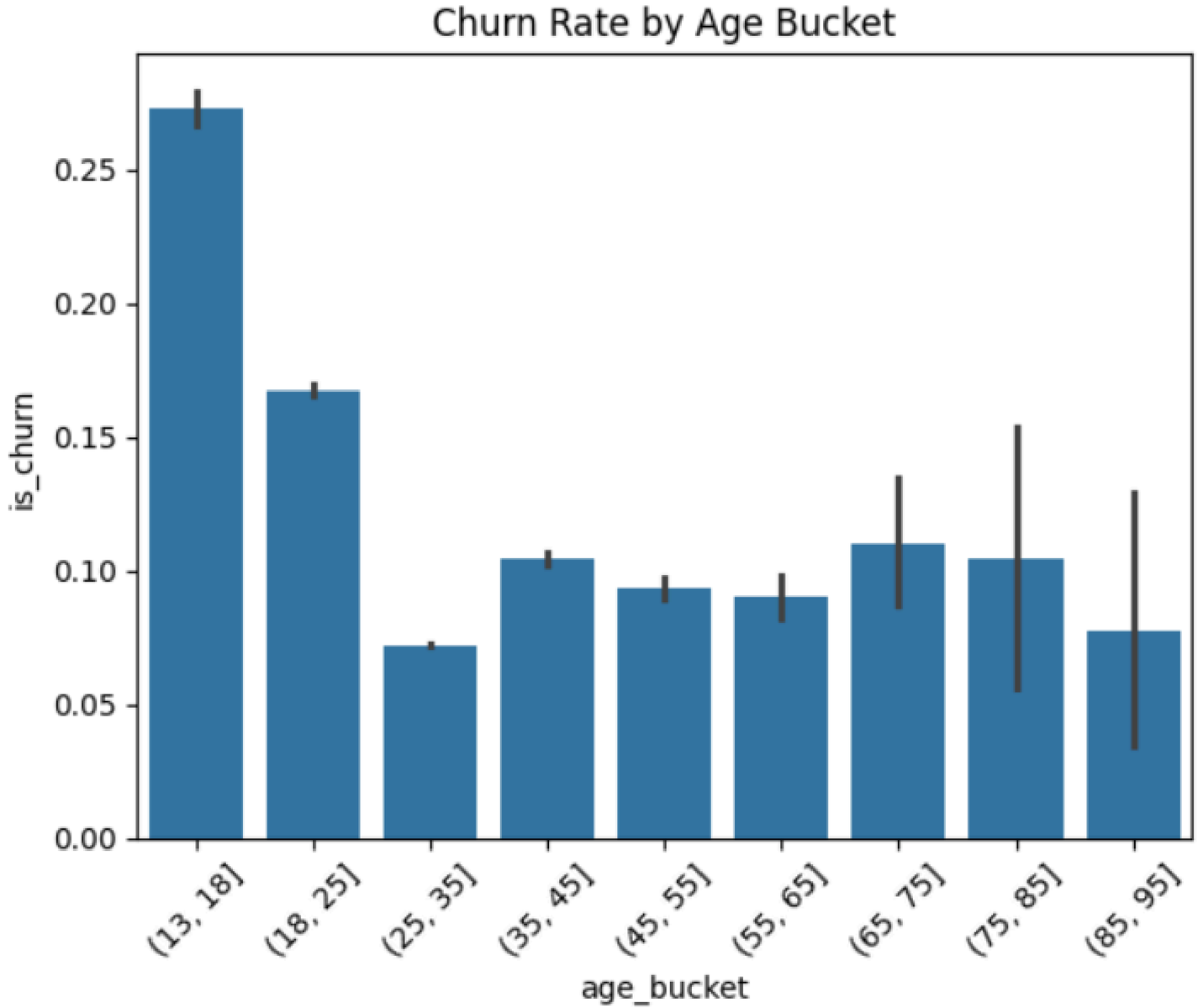
User Churn Count



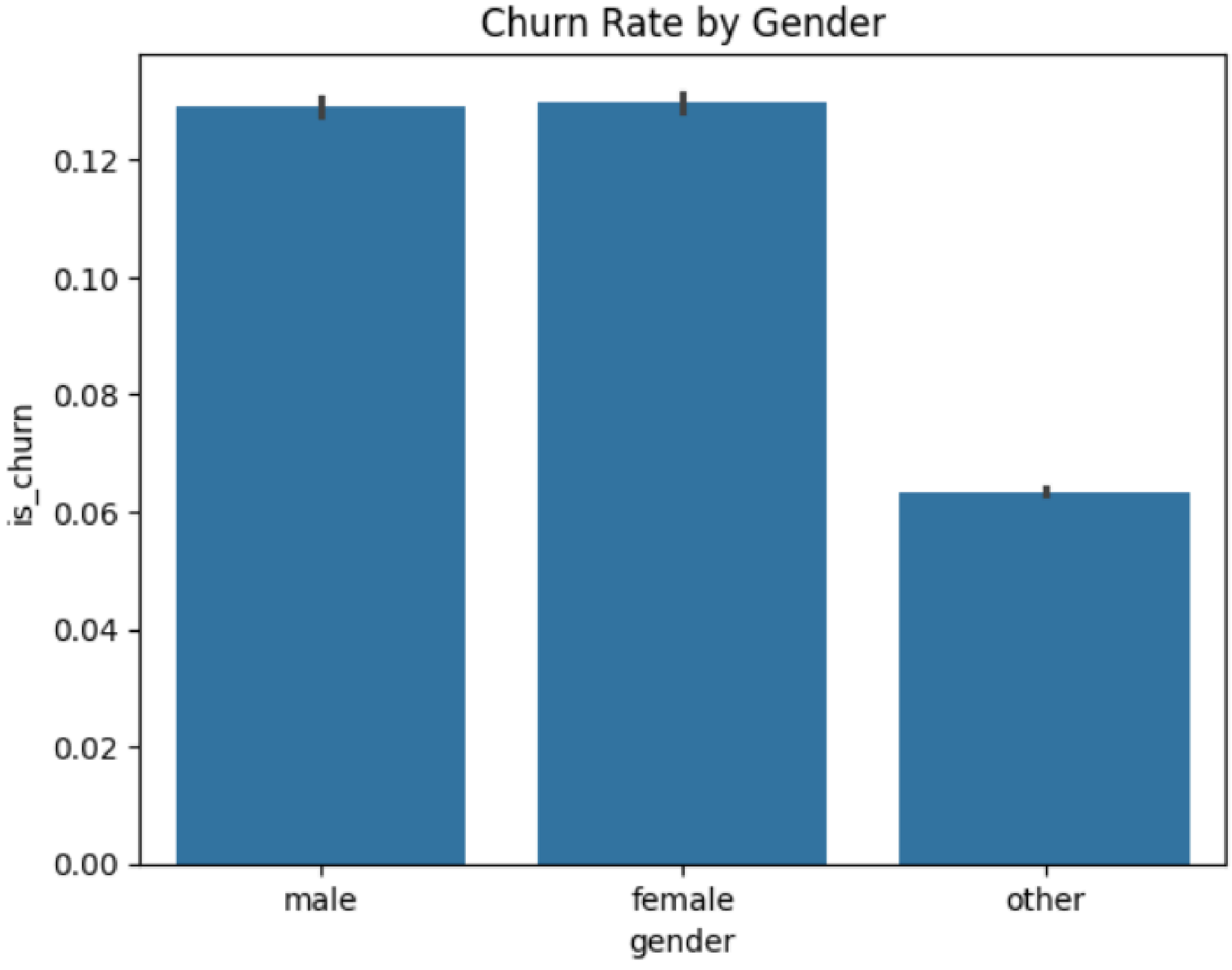
User Age Distribution



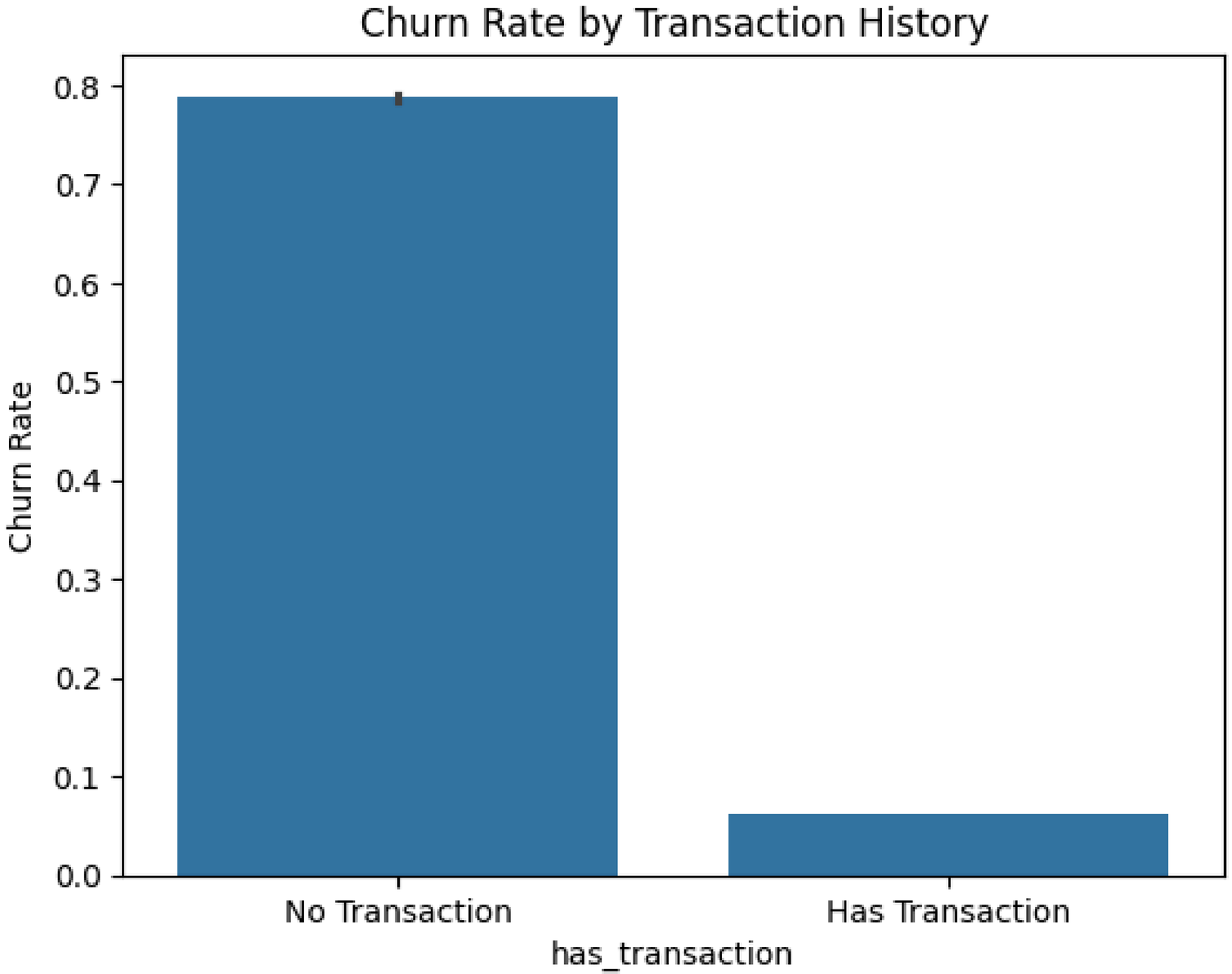
Churn Rate by Age Group



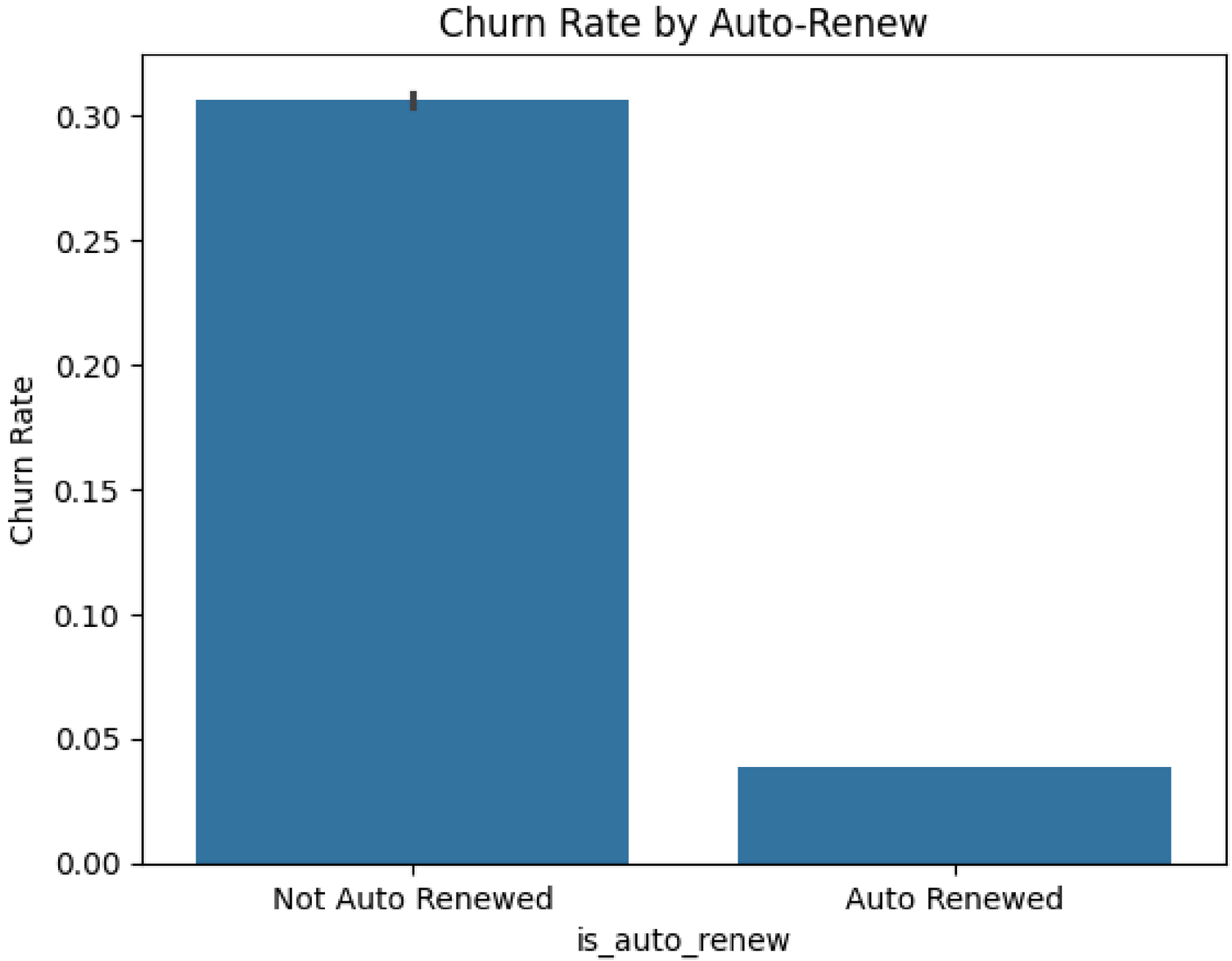
Churn Rate by Gender



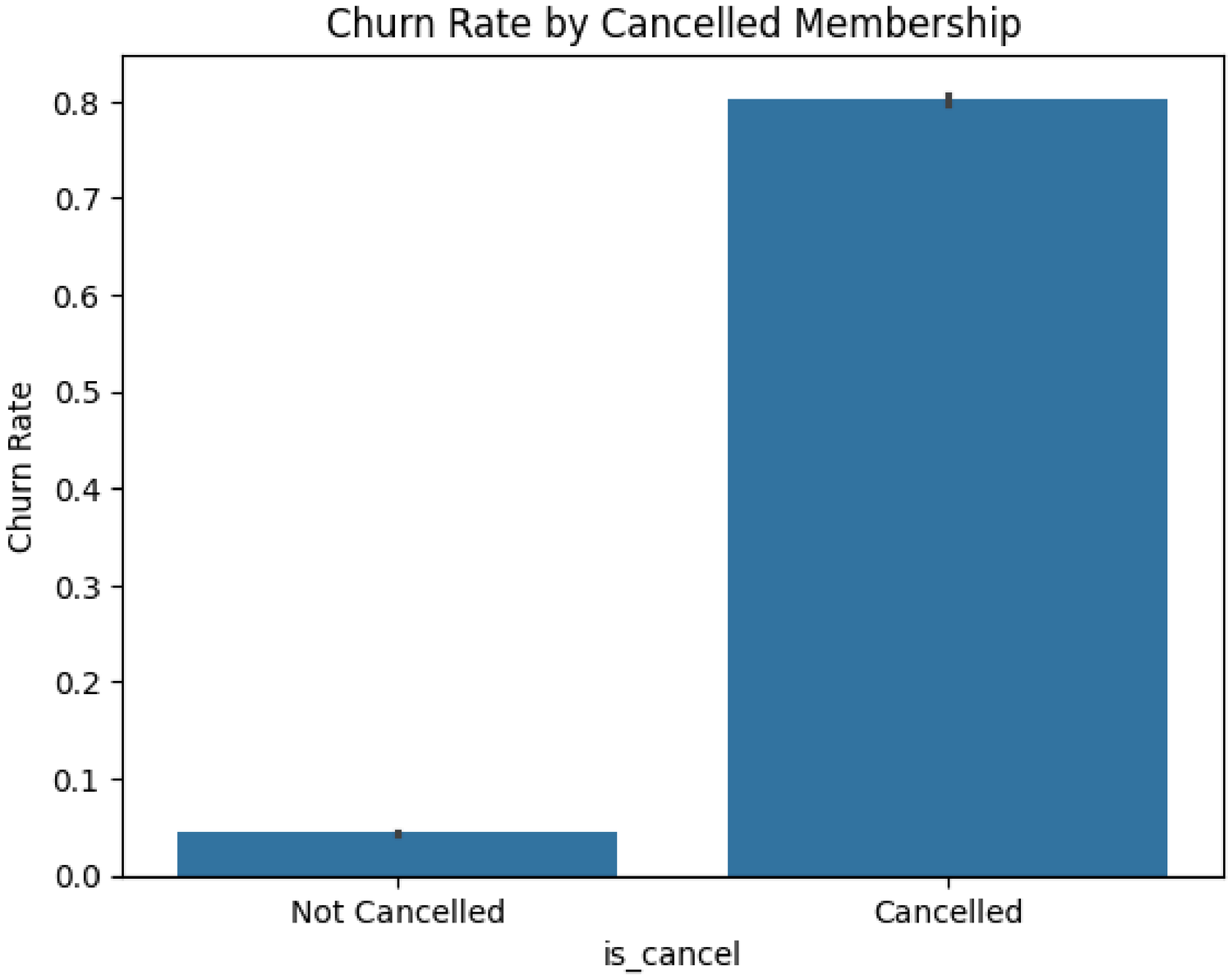
Churn Rate by Transaction History



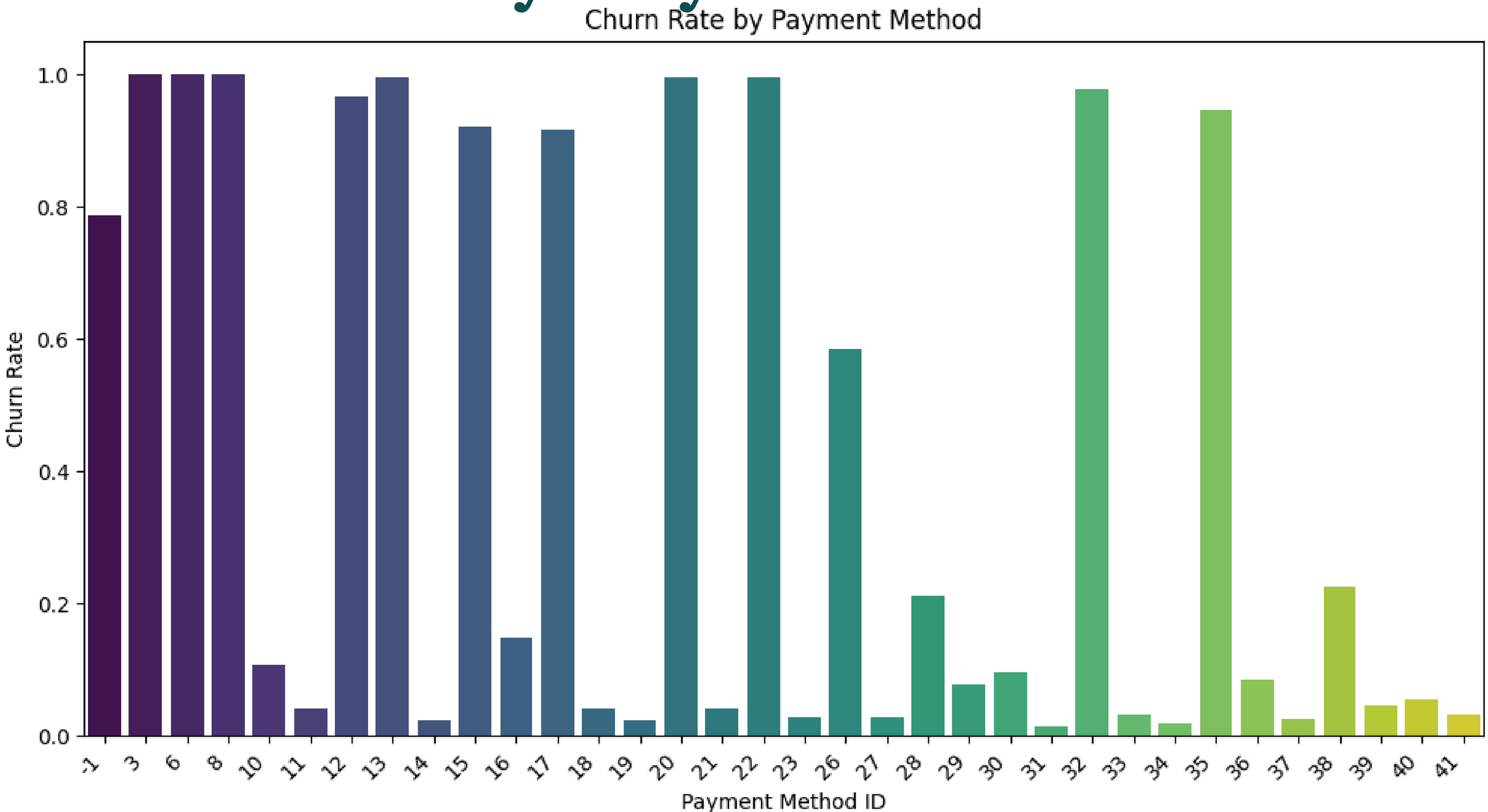
Churn Rate by Auo-Renew



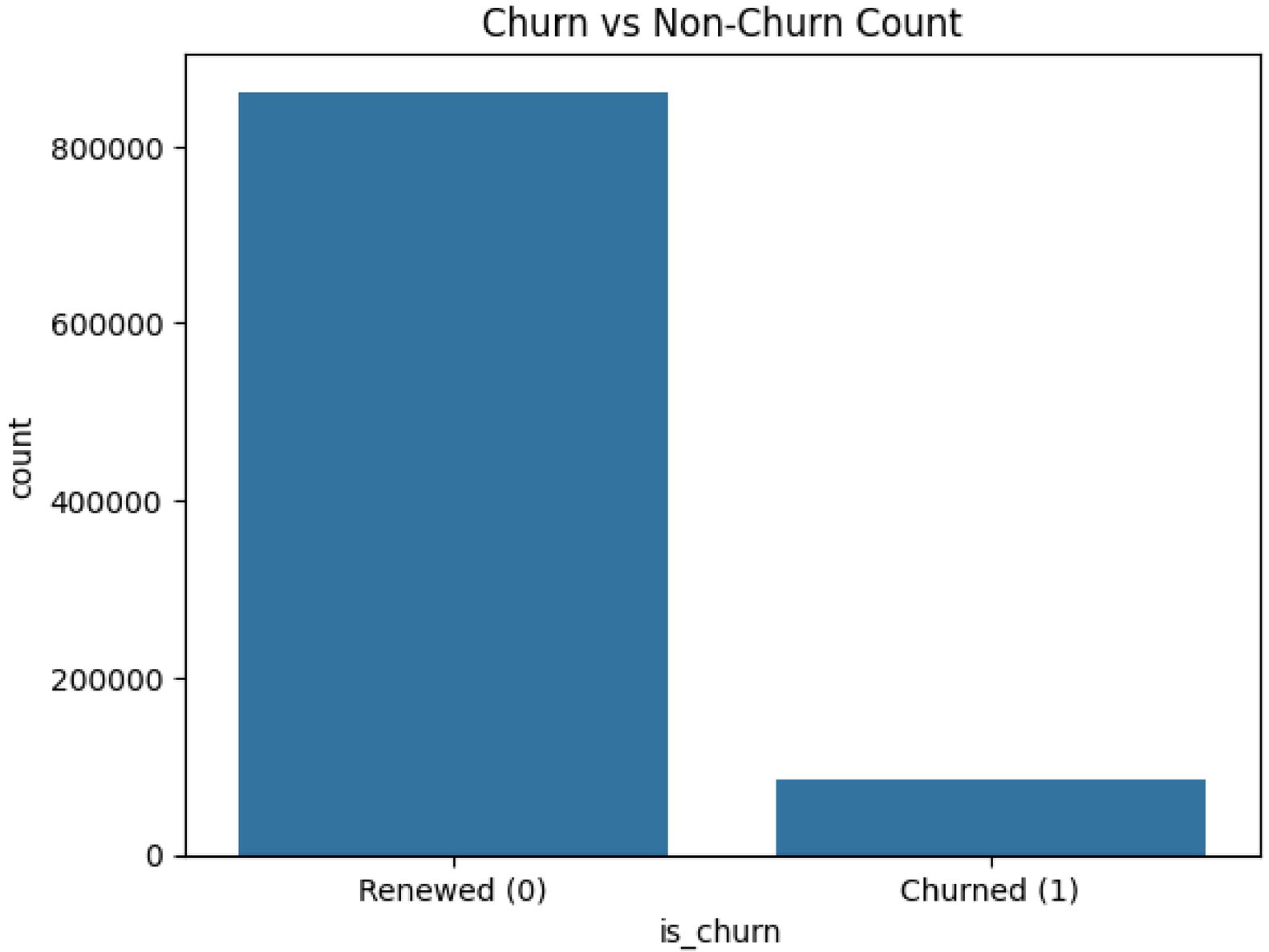
Churn Rate by Cancelled Membership



Churn Rate by Payment Method



Churn vs Non-churn Count



churn rate:
8.99%

Modeling

In this phase, we built and trained several machine learning models to predict whether a user will churn or not.

The models were trained using the cleaned and engineered dataset from the previous preprocessing steps. The target variable was is_churn, a binary label indicating if the user has churned (1) or not (0).

Modeling

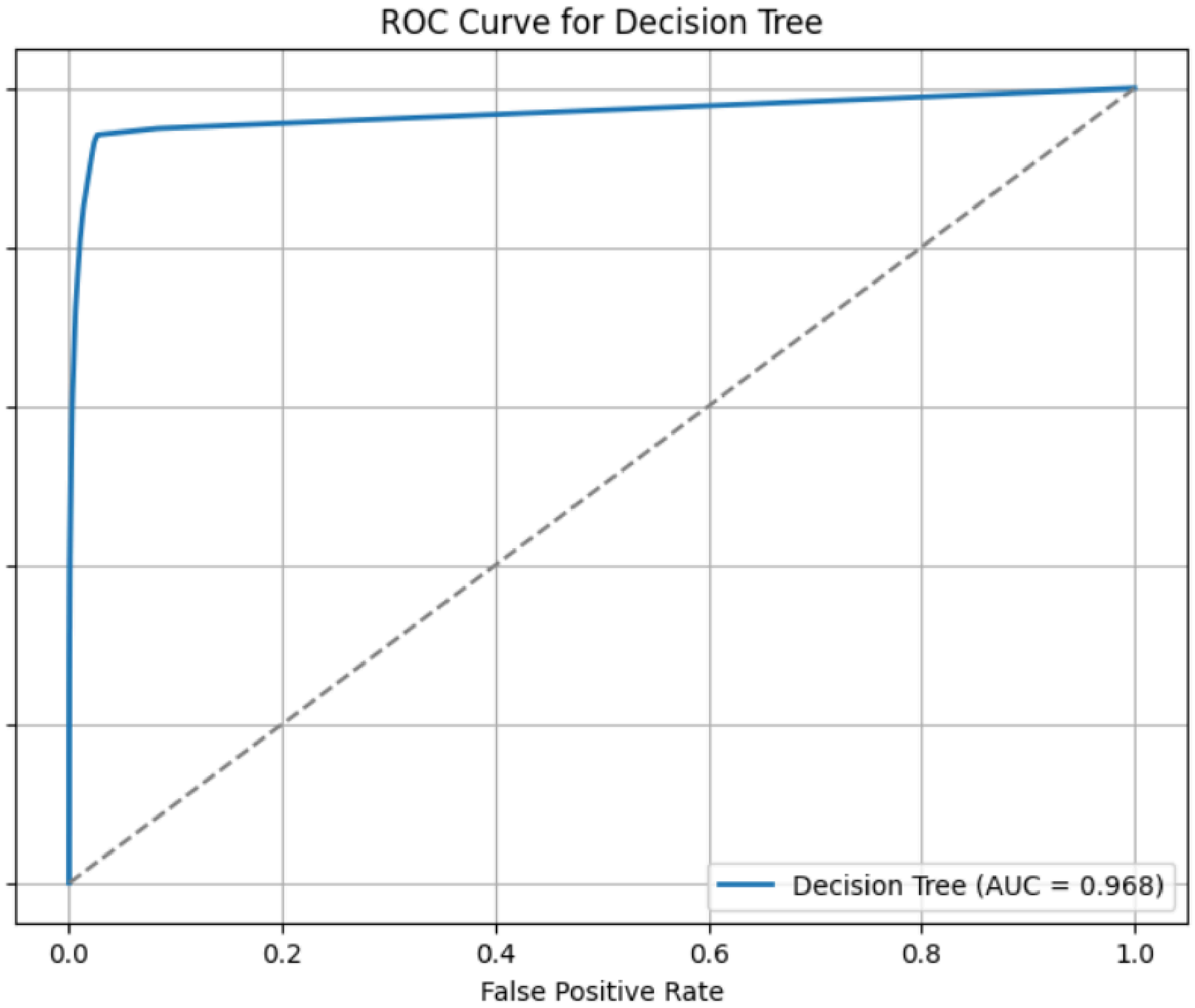
Models Used:

1. **Logistic Regression:** A simple baseline model to understand the linear relationship between features and churn.
2. **Decision Tree Classifier**: A tree-based model that captures non-linear relationships and interactions between features.
3. **Random Forest Classifier**: An ensemble model that uses multiple decision trees to improve performance and reduce overfitting.
4. **Weighted Naive Bayes:** A probabilistic model where feature contributions were adjusted using calculated weights based on feature importance

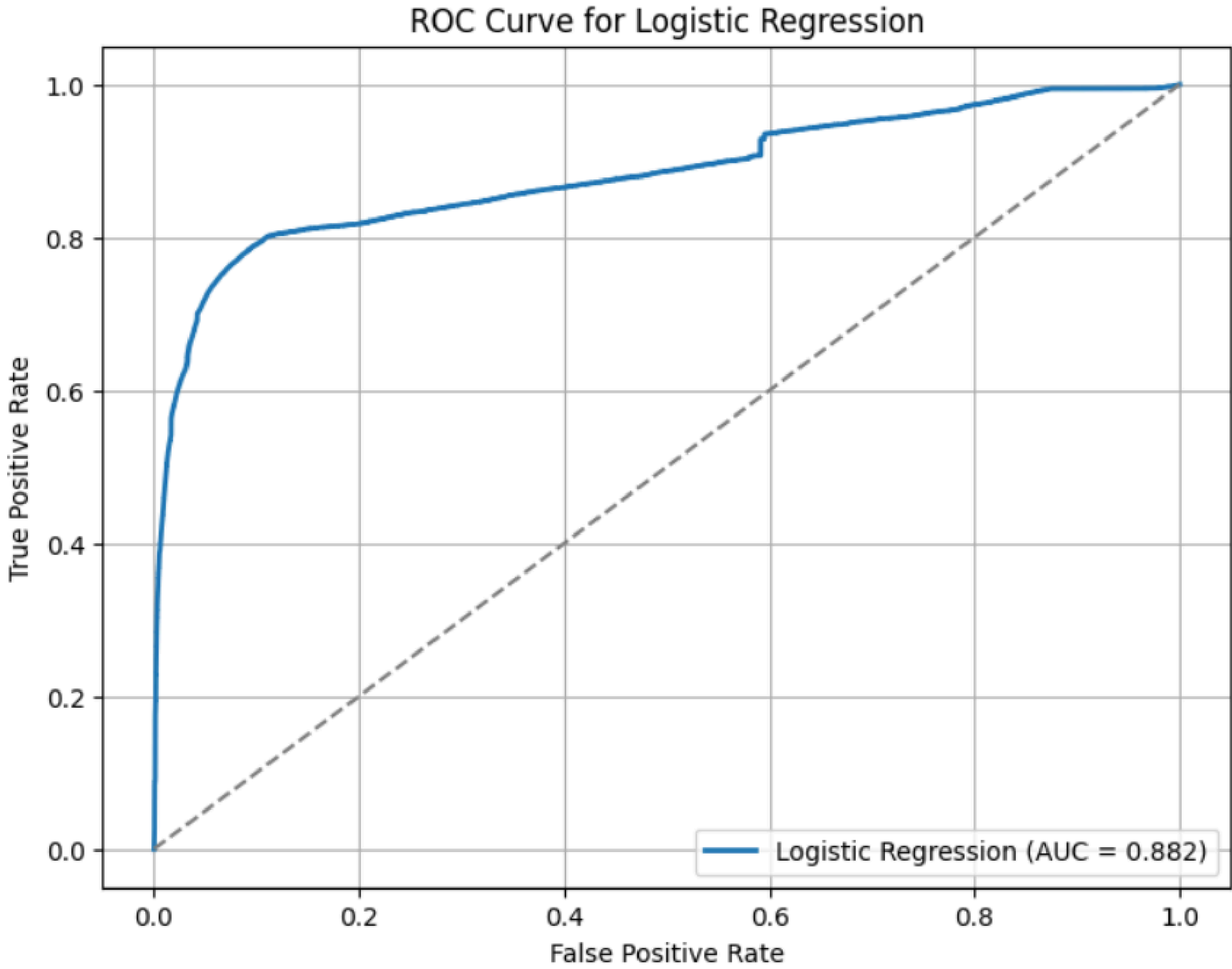
Performance Metrics

	Decision Tree	Logistic Regression	Random Forest	Naive Bayes
Accuracy	97%	93%	97%	97%
Weighted Avg Precision	0.97	0.94	0.98	0.97
Weighted Avg Recall	0.97	0.93	0.97	0.97
Weighted Avg F1 Score	0.97	0.94	0.97	0.97
Support	189338			

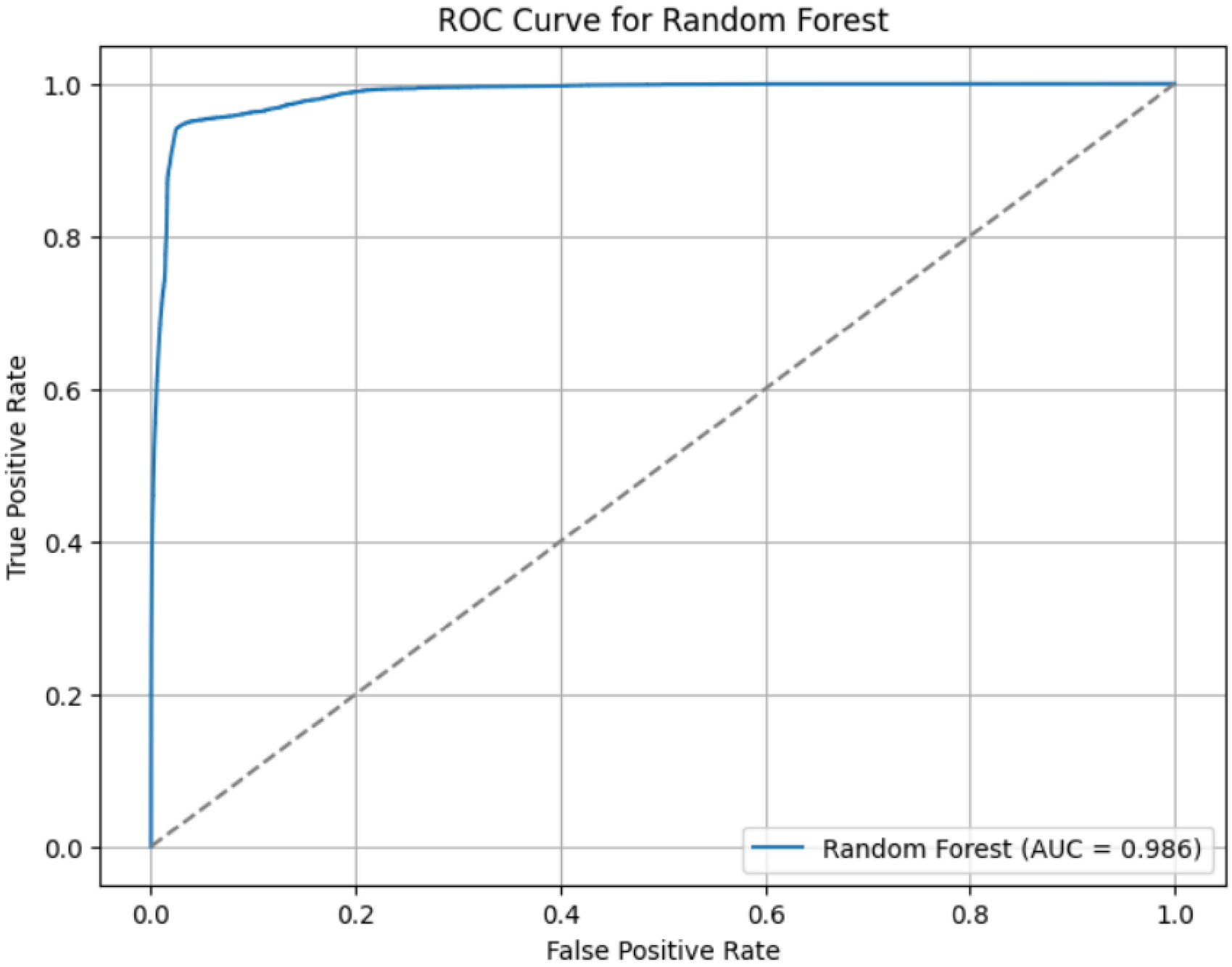
Metrics - ROC Curves



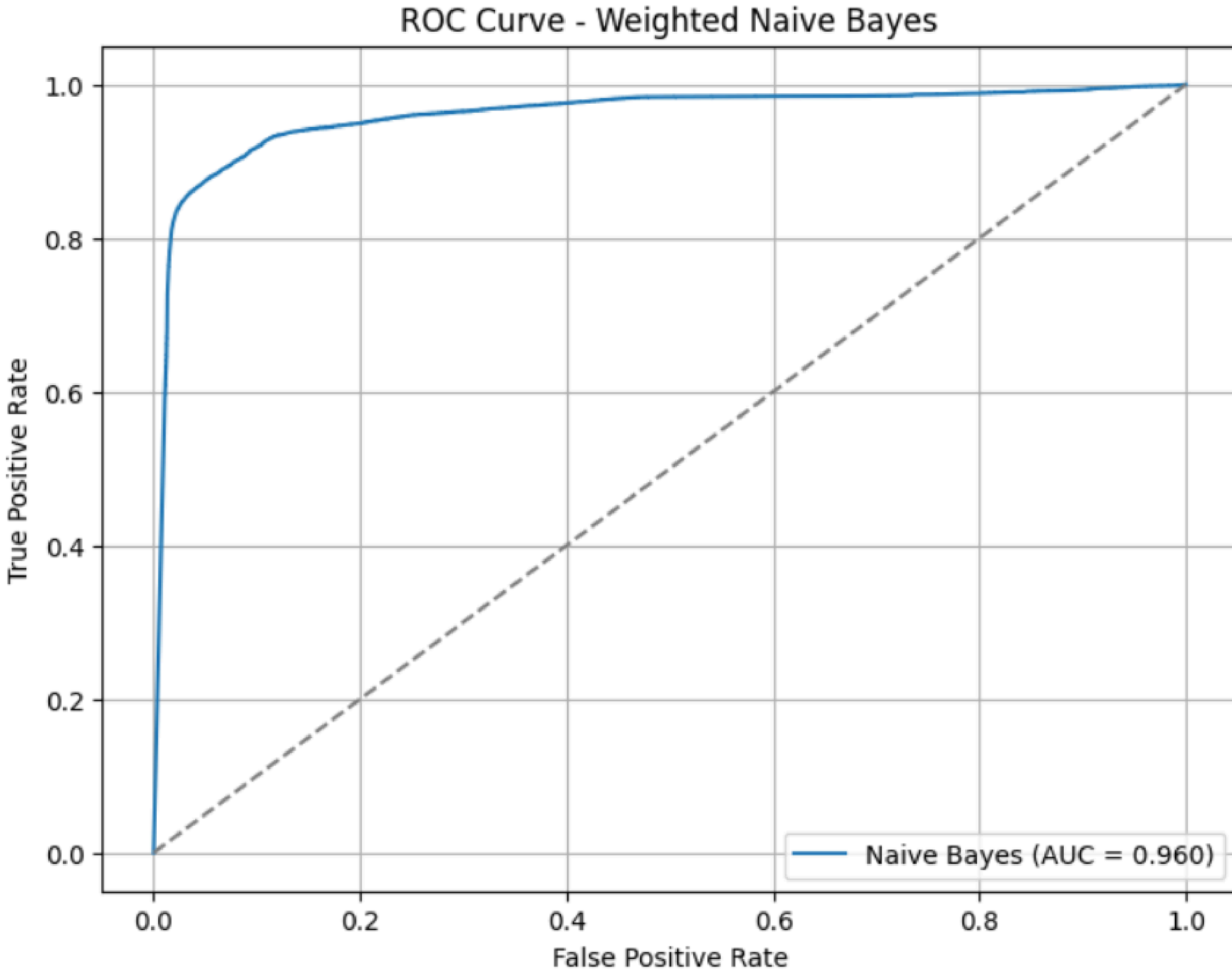
Metrics - ROC Curves



Metrics - ROC Curves



Metrics - ROC Curves



Testing

- Used random forest model for testing.
- Performed left join on the test dataset containing user IDs with the dataset containing user data.
- Saved results in a submission file to be displayed on dashboard.

Predicted Churners: 1036 out of 24274

Predicted Churn Percentage: 4.27%

Model Deployment Process

Step 1: Model Training and Saving

Trained a Random Forest Classifier

Saved the trained model using Joblib as random_forest_churn.pkl

Step 2: Setting up FastAPI Server

Created a FastAPI application to serve the model

Ensured input order matches training features exactly

Step 3: Local API Deployment

Deployed the model locally using Uvicorn server.
Made the API accessible at http://127.0.0.1:8000/predict

Step 4: Built a Streamlit Client

Created a user-friendly form in Streamlit to input all required model features manually.

Connected the Streamlit app to the deployed FastAPI server at http://127.0.0.1:8000/predict.

🎯 Time to See It in Action!

Let’s jump into our Churn Prediction Dashboard and experience real-time predictions with Streamlit !!!

Model Deployment on Streamlit: Deployment

Churn Prediction Dashboard

Churn Prediction Dashboard On Streamlit

Our dashboard consists of 4 main tabs:

1) Data Exploration:
Understand the dataset and uncover key patterns in user behavior.

2) Feature Engineering:
Dive deeper into the engineered features and the insights discovered from them.

3) Model Comparison:
Compare the performance of different models we trained and evaluated.

4) Predict Churn:
View predictions on Kaggle test data and explore ideas for future improvements.

Conclusion

After building and evaluating multiple machine learning models to predict customer churn, we found that the **Random Forest Classifier** performed the best in terms of accuracy, F1-score, and ROC AUC. This indicates that the model is capable of reliably distinguishing between users who will churn and those who will stay, based on the engineered features.

Thank You