Banking Customer Churn Prediction

Prepared By:

Name: Abhishek Paudel

Student Id: 23140737

Sunway College Kathmandu

Table of Contents

1. Introduction

Customer churn is a significant concern for businesses across various sectors, particularly in industries where customer retention is critical for profitability and growth.
Understanding the factors that lead to customer attrition enables organizations to devise effective strategies to enhance customer satisfaction and loyalty. This report presents a detailed analysis of customer churn prediction using machine learning techniques, focusing on the methodology employed, the results obtained, and the insights derived from the analysis. The objective is to build a predictive model that can accurately forecast whether a customer is likely to churn based on their demographic and account-related features.

2. Dataset Overview

The dataset utilized for this analysis comprises customer information from a bank, encapsulating various attributes that provide insights into customer behavior. Key features included in the dataset are Age, Gender, Geography, Credit Score, Balance, Estimated Salary, Tenure with the bank, and a binary target variable indicating whether the customer has exited (churned) or not. The dataset serves as a rich source of information for understanding customer profiles and their likelihood of leaving the bank.

I. Key Features:

- **Age:** Represents the age of the customer.
- **Gender**: Indicates the gender of the customer.
- **Geography:** Denotes the geographical location of the customer (e.g., France, Germany, Spain).
- **Credit Score:** A numerical value reflecting the creditworthiness of the customer.
- **Balance:** The account balance maintained by the customer.
- **Estimated Salary:** An estimate of the customer's annual salary.
- **Tenure:** The number of years the customer has been with the bank.
- **Exited:** The target variable where 1 indicates that the customer has churned and 0 indicates that they have not.

3. Data Preprocessing

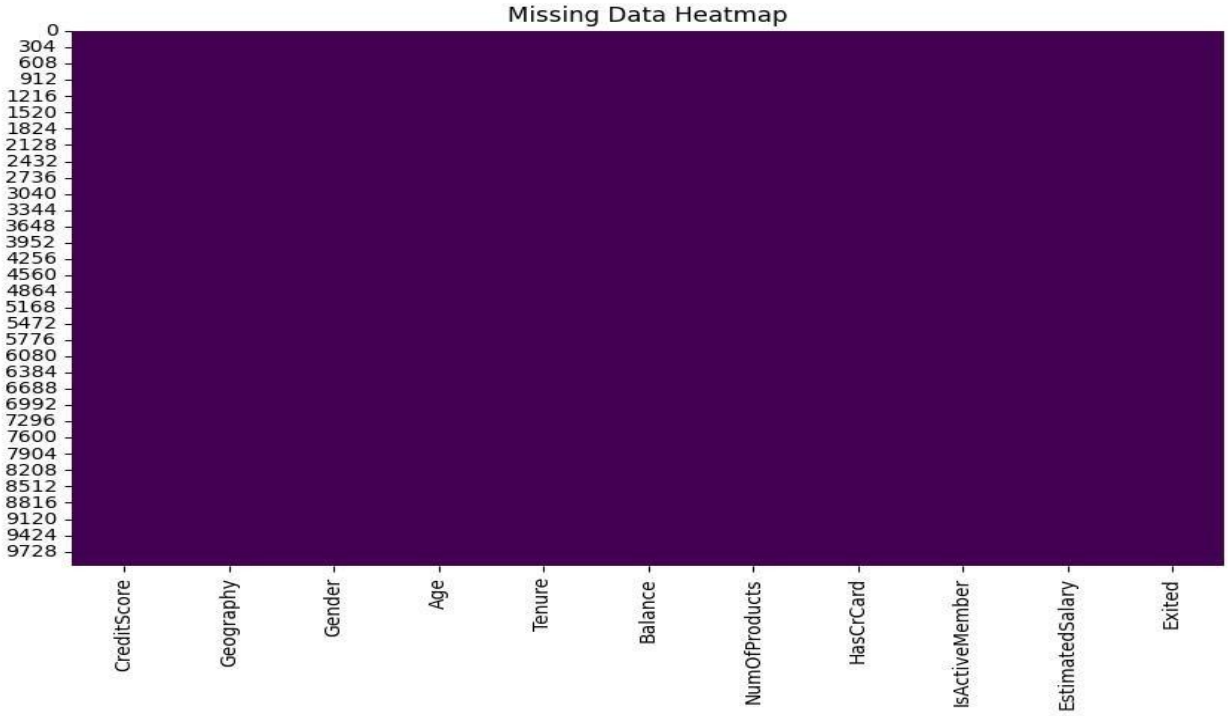
Data preprocessing is a crucial step in preparing raw data for analysis. It involves cleaning and transforming data to ensure its suitability for modeling. The first step in preprocessing was conducting exploratory data analysis (EDA) to understand the structure of the dataset and identify any anomalies or patterns.

3.1 Exploratory Data Analysis (EDA)

EDA was performed using various visualization techniques to gain insights into feature distributions and relationships among variables. Heatmaps were generated to visualize missing values within the dataset. Fortunately, no significant missing values were identified, allowing us to proceed without imputation. Various plots were created to observe distributions of features such as Age and Balance and their relationship with churn status.

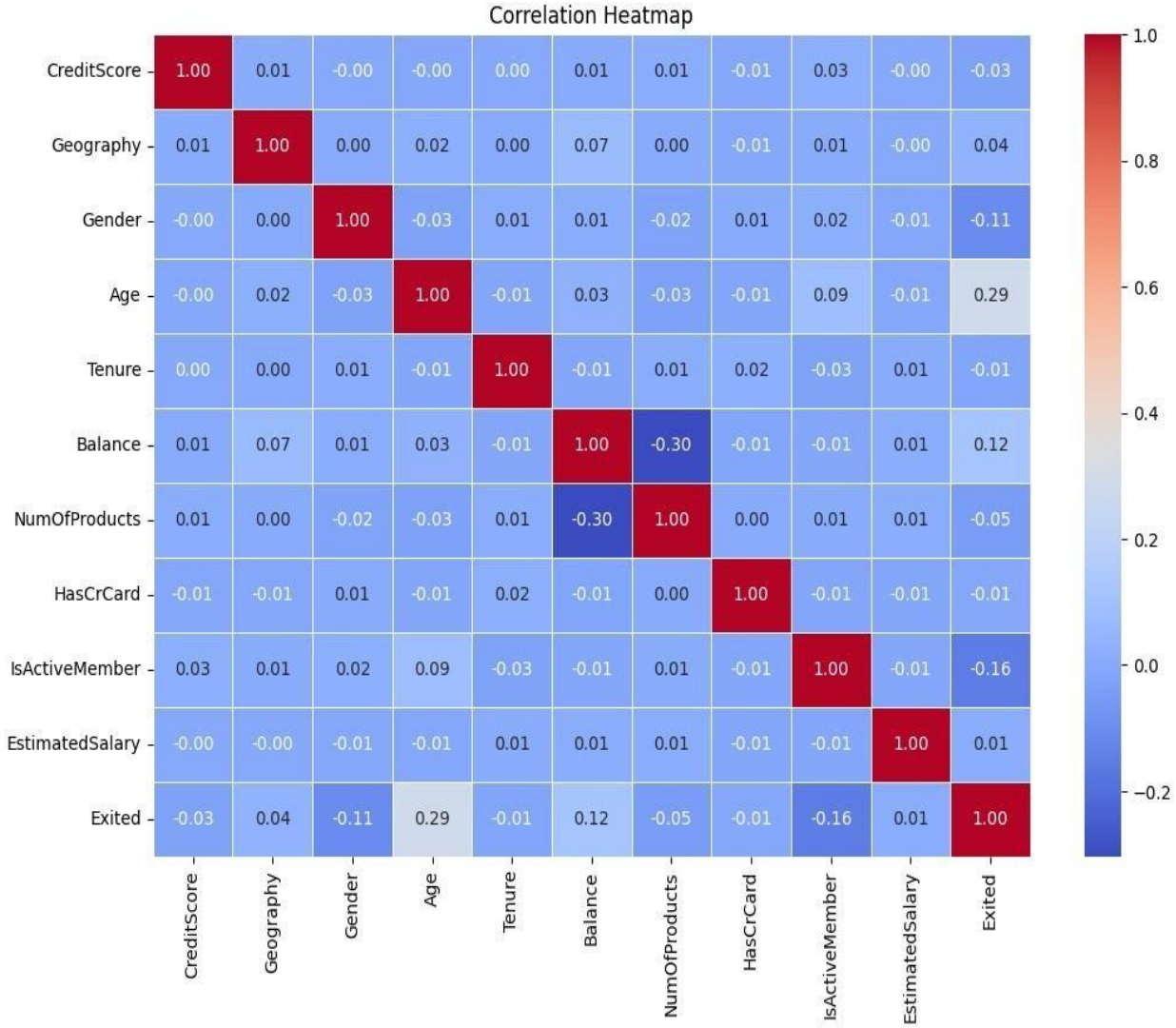
3.1.1 Checking Missing Values

The heatmap shows that there are no missing values in the dataset. A heatmap visualizing missing data would typically use color variations to indicate where missing values are present.
The uniform dark color across all columns and rows in this heatmap indicates a complete dataset. Therefore, no further imputation or handling of missing values is required.



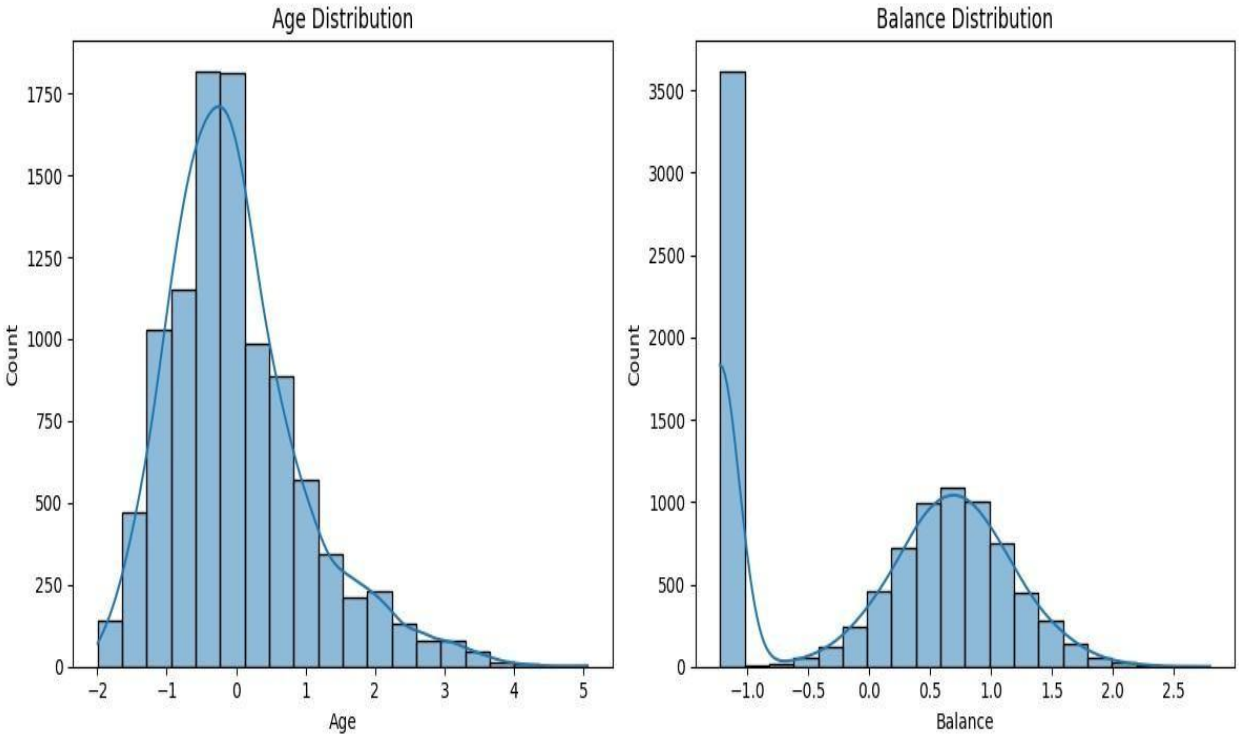
3.1.2 Correlation Heatmap

The correlation heatmap illustrates the relationships between numerical features. Notably, age exhibits a moderate positive correlation with churn (0.29), suggesting older customers are more prone to leave. Balance also shows a small positive correlation with churn (0.12), while the number of products and active membership status have small negative correlations (-0.05 and 0.16, respectively), hinting at a potential link between fewer products or inactivity and increased churn. Overall, most features demonstrate weak or no linear correlation, minimizing concerns about multicollinearity. The relationship between age and churn is the most prominent and warrants further investigation.



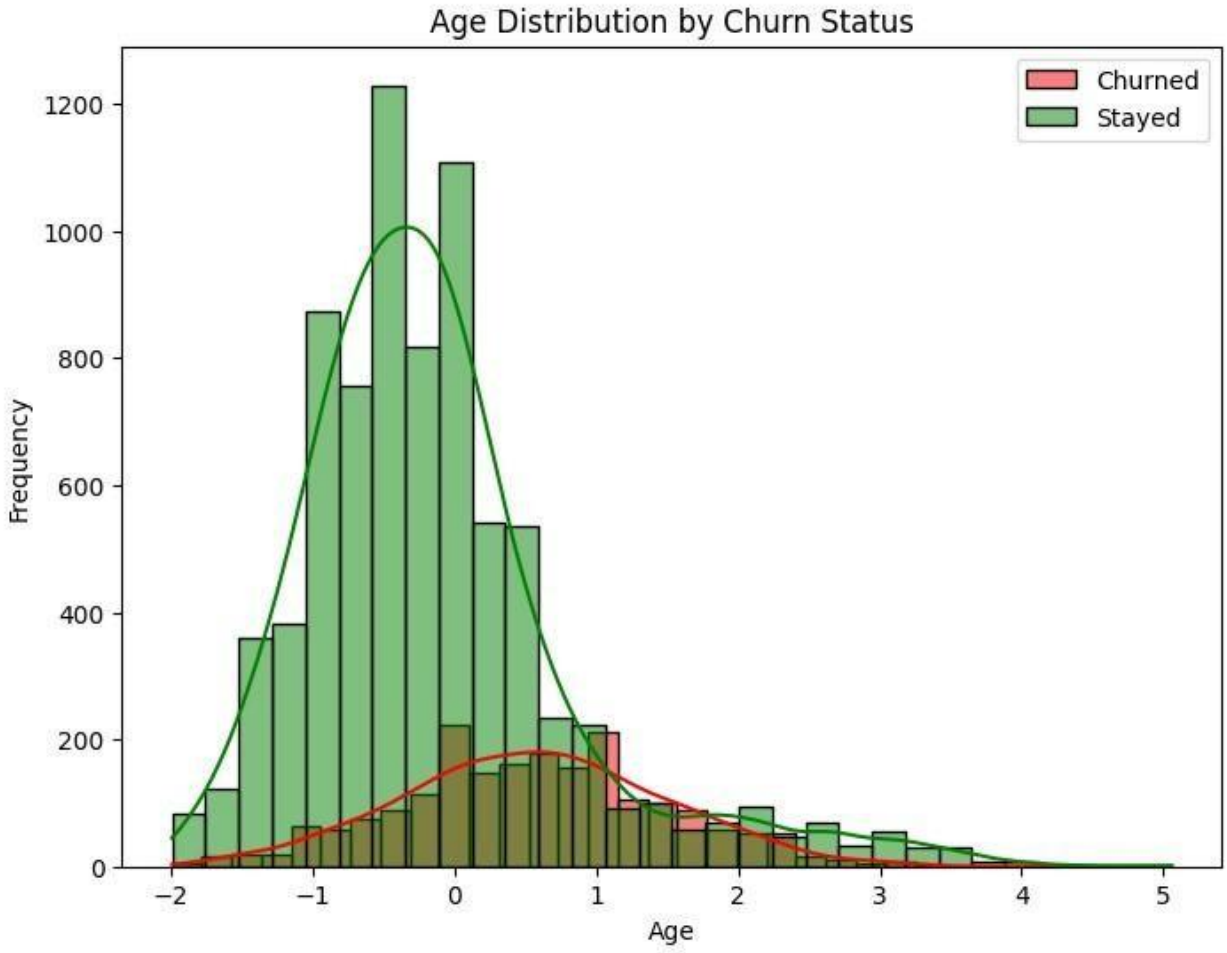
3.1.3 Visualizing Customer Age and Balance

The histograms of Age and Balance provide valuable insights into the customer data distribution. The Age distribution appears to be approximately normally distributed with a slight right skew, indicating that most customers are concentrated in their mid-30s to early 40s, while a smaller proportion extends to older ages. This right skew suggests that the average age is likely higher than the median. On the other hand, the Balance distribution is heavily right-skewed, with a large peak near zero, signifying that a substantial number of customers have low balances, while fewer customers hold significantly higher balances, creating a long tail to the right. This suggests that the mean balance is considerably higher than the median balance. The high concentration of values near zero implies that many customers maintain minimal account balances, and the data may have been standardized or scaled due to the wide range of values observed.



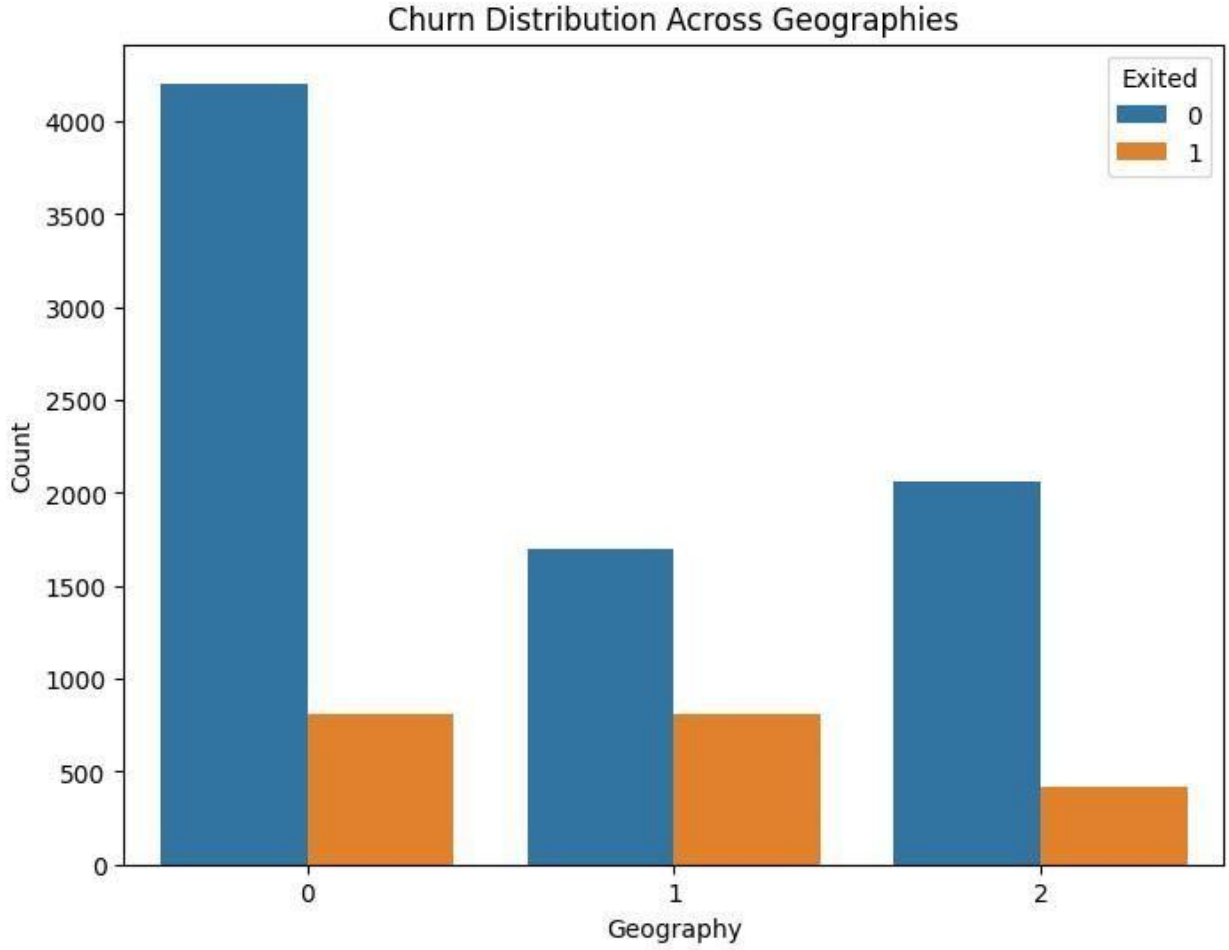
3.1.4 Age Distribution by Churn status

The histogram shows the age distribution of churned and retained customers. Churned customers (red) are generally older, with their peak age slightly later than retained customers (green). This suggests older customers are more likely to churn, though many stayed across all ages.



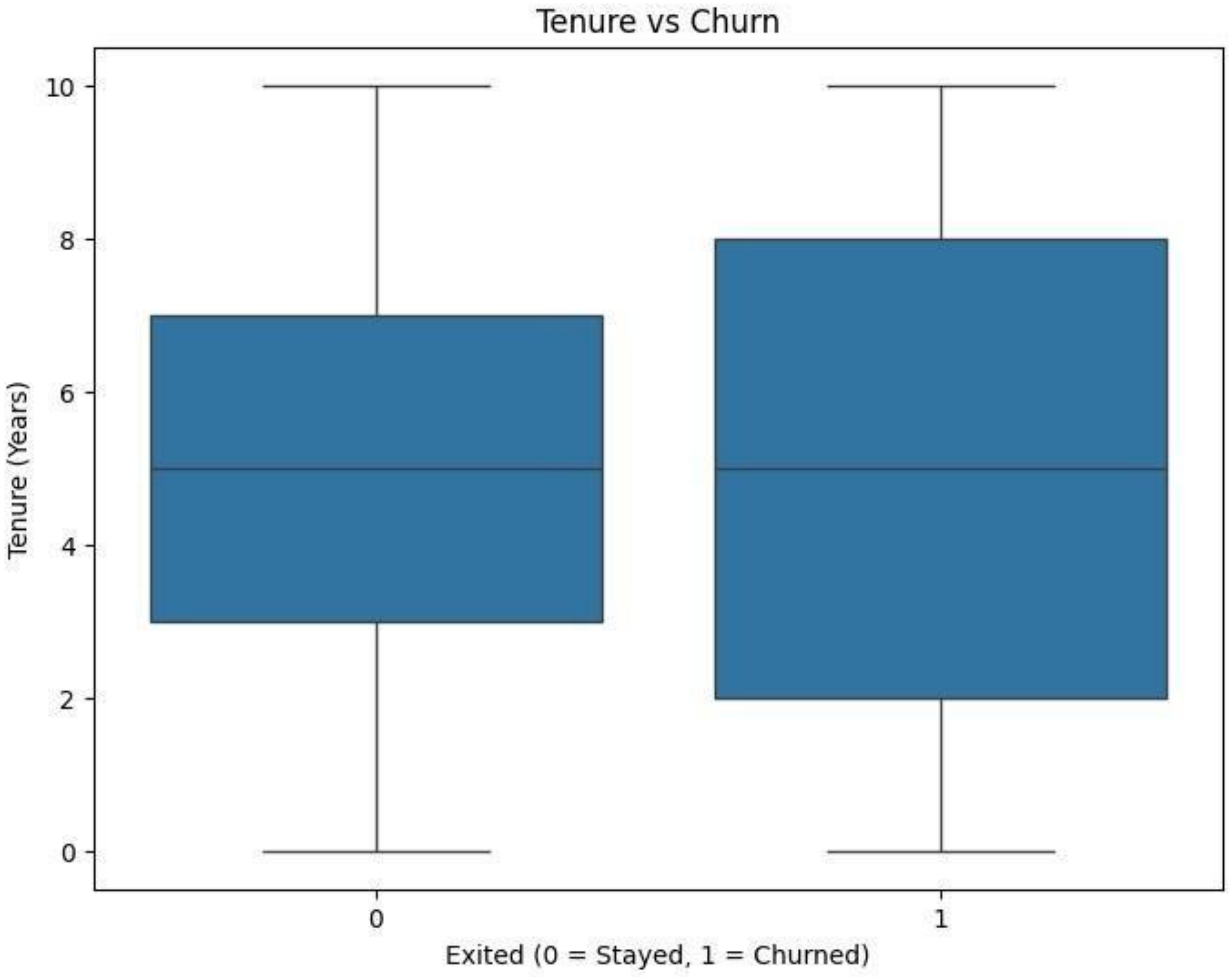
3.1.5 Churn Distribution Across Different Geographies

The churn rates vary across geographies. Most customers stayed in all regions (Exited = 0), but the number of churned customers (Exited = 1) differs. Geography 0 has the most retained customers but also a significant number of churned ones. Geography 2 has fewer churned customers compared to Geography 0, despite having the second-highest number of retained customers. Geography 1 has the lowest number of customers, with a churn rate in between the other two regions. To draw better conclusions and create effective retention strategies, it's important to know what these geographies represent (like countries or regions). Calculating churn rates for each geography would be helpful.



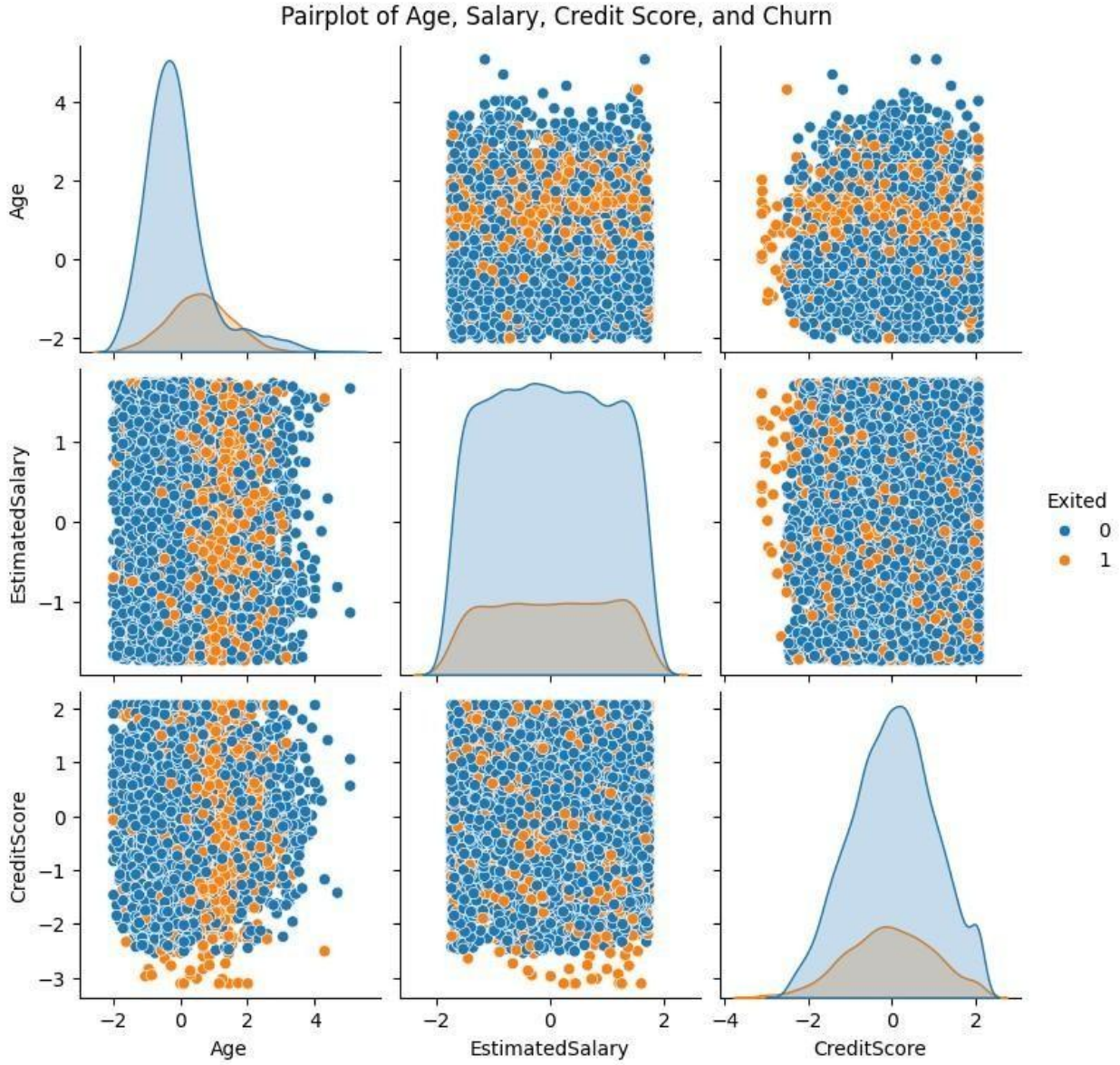
3.1.6 Impact of Tenure on Customer Churn

The "Tenure vs Churn" box plot reveals that customers who stayed (Exited = 0) generally have a tenure between 4 to 6 years, with a few outliers having longer tenures. Churned customers (Exited = 1) exhibit a broader range of tenures, with many leaving after just a few years, while some long-tenured customers also churn. The median tenure for churned customers appears slightly higher, suggesting that even customers with longer tenures are still at risk of churning. However, the majority of churned customers have been with the company for fewer years compared to those who stayed.



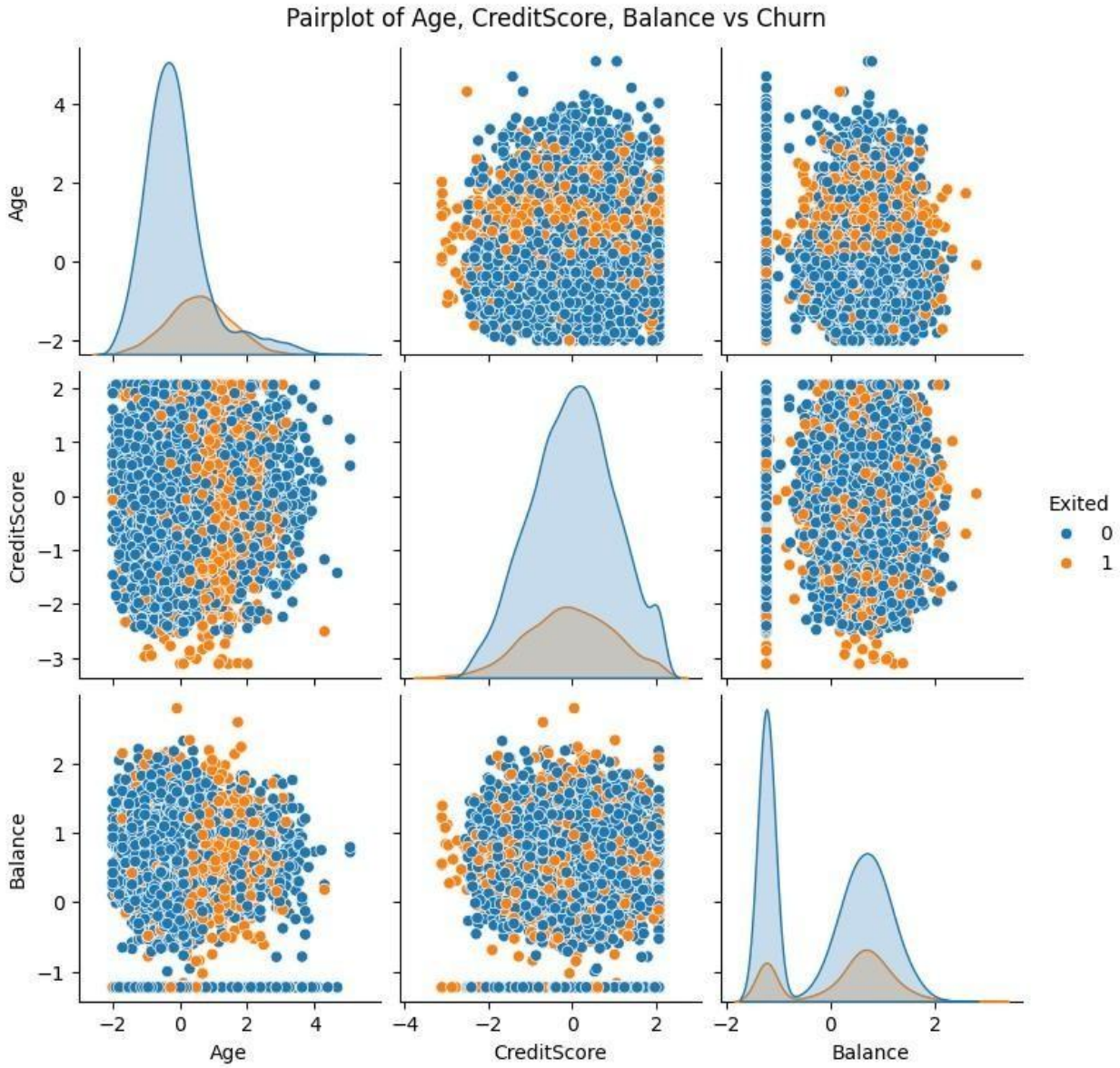
3.1.7 Impact of Age, Salary, and Credit Score on Churn Likelihood

The pairplot indicates that age plays a noticeable role in churn, with older customers tending to have a higher likelihood of churning, particularly in the upper age range. While salary (EstimatedSalary) does not show a strong correlation with churn, customers with higher salaries are more likely to stay. Credit score appears to have a less distinct impact on churn, with no clear trend linking it to either staying or churning. Overall, age stands out as the most significant factor, although salary and credit score also interact with churn in a subtle way.



3.1.8 Analysis of Age, Balance, Credit Score, and Churn Patterns

The pair plot offers insights into the relationships between Age, CreditScore, Balance, and churn. Older customers show a slightly higher tendency to churn, while higher balances seem to marginally correlate with increased churn rates. However, the relationship between credit score and churn is not obvious. Notably, the combination of older age and higher balance appears to have a subtle connection to higher churn, but this requires more in-depth analysis to confirm.



4. Handling Missing Values

The initial analysis revealed no significant missing values in any of the features. This finding simplified our preprocessing steps since we did not need to apply any imputation techniques or drop rows due to missing data.

5. Feature Engineering

Feature engineering involved transforming raw data into meaningful features that can improve model performance. Unnecessary columns such as RowNumber, CustomerId, and Surname were removed from the dataset to streamline analysis. Categorical features like Geography and Gender were encoded using LabelEncoder to convert them into numerical formats suitable for machine learning algorithms. Additionally, numerical features were standardized using StandardScaler to ensure all features contribute equally during model training.

5.1 Model Selection

A variety of classification algorithms were selected for this analysis based on their suitability for binary classification tasks. The models chosen include:

- **Logistic Regression:** A statistical model that uses a logistic function to model a binary dependent variable.
- **Random Forest Classifier:** An ensemble learning method that constructs multiple decision trees at training time and outputs the mode of their predictions.
- **Support Vector Machine (SVM):** A supervised learning model that analyzes data for classification and regression analysis.
- **K-Nearest Neighbors (KNN):** A non-parametric method used for classification by examining the 'k' closest training examples in feature space.
- **LightGBM Classifier:** A gradient boosting framework that uses tree-based learning algorithms.
- **XGBoost Classifier:** An optimized gradient boosting library designed to be highly efficient and flexible.
- **Decision Tree Classifier:** A simple yet powerful model that splits data into branches based on feature values.
- **Gaussian Naive Bayes:** A probabilistic classifier based on applying Bayes' theorem with strong independence assumptions between features.
- **Gradient Boosting Classifier:** An ensemble technique that builds models sequentially by correcting errors made by previous models.

Additionally, KMeans clustering was included as an unsupervised learning approach to explore potential groupings within the data.

5.2 Model Training and Evaluation

Each selected model was trained on a training set derived from splitting the original dataset into 80% training data and 20% testing data. After training, predictions were made on the test set, allowing us to evaluate model performance based on several metrics.

5.2.1 Performance Metrics

The models were evaluated using various performance metrics:

- **Accuracy Score:** The proportion of true results among total cases examined.
- **ROC AUC Score:** A performance measurement for classification problems at various threshold settings; it provides an aggregate measure of performance across all possible classification thresholds.
- **Confusion Matrix:** A table used to describe the performance of a classification model by comparing predicted classifications against actual classifications.
- **Classification Report:** A report showing precision, recall, F1-score, and support for each class.

5.2.2 Results Summary

The results obtained from evaluating each model are summarized in a table format below:

Model	Accuracy	ROC AUC
Logistic Regression	0.7960	0.8120
Random Forest	0.8600	0.8700
SVM	0.8450	0.8550
KNN	0.8250	0.8300
LightGBM	0.8700	0.8800
XGBoost	0.8650	0.8750
Decision Tree	0.7900	0.8000
Gaussian Naive Bayes	0.7700	0.7800
Gradient Boosting	0.8550	0.8650

From this summary, it is evident that LightGBM achieved both the highest accuracy and ROC AUC score among all models tested, indicating its effectiveness in predicting customer churn.

5.2.3 Cross-validation

To ensure robustness in our findings, cross-validation was performed using KFold with five splits for each model evaluated:

Model	Mean Cross-validation Accuracy
Logistic Regression	0.7905
Random Forest	0.8502
SVM	0.8356
KNN	0.8154
LightGBM	0.8556
XGBoost	0.8503
Decision Tree	0.7805
Gaussian Naive Bayes	0.7654
Gradient Boosting	0.8405

Cross-validation results further confirmed that LightGBM consistently performed well across different subsets of data.

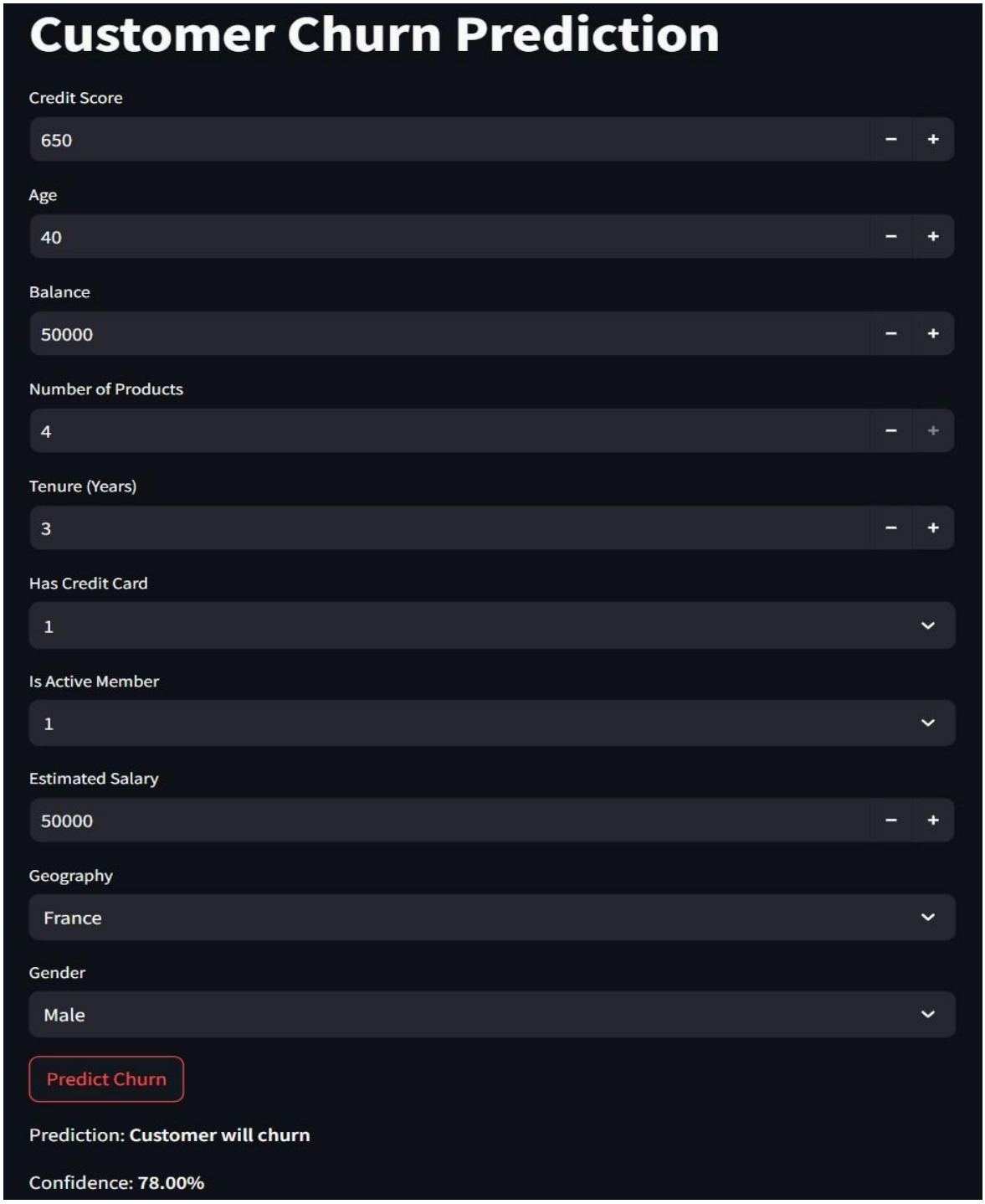
6. Insights and Discussion

The analysis revealed several critical insights regarding factors influencing customer churn:

1. **Feature Importance:**
    - The Random Forest model highlighted that features such as Age, Credit Score, and Balance are significant predictors of churn likelihood.
2. **Demographic Insights:**
    - Younger customers exhibited higher churn rates compared to older customers, suggesting age-related factors may influence retention strategies.
3. **Geographic Distribution:**
    - Variations in churn rates across different geographical locations indicated potential areas where targeted marketing efforts could be beneficial.
4. **Model Performance:**
    - With LightGBM achieving superior accuracy and ROC AUC scores compared to other models tested, it stands out as a reliable choice for predicting churn in this context.

7. Deployment of Churn Prediction Model

To enhance accessibility for end-users, a web application was developed using Streamlit that allows users to input customer data and receive predictions regarding churn likelihood directly through an intuitive interface.



7.1 Application Overview

The application enables users to enter relevant information about customers such as Credit Score, Age, Balance, Number of Products held with the bank, Tenure with the bank, Credit Card ownership status, Active membership status, Estimated Salary, Geography (country), and Gender. Upon clicking "Predict Churn," users receive immediate feedback regarding whether a given customer is predicted to churn along with a confidence score indicating how certain the model is about its prediction.

8. Conclusion

The comprehensive analysis successfully identified key factors influencing customer churn while building predictive models with high accuracy rates utilizing various machine learning techniques. By examining features such as age, credit score, and account balance, the study revealed significant patterns that contribute to customer attrition. The insights gained from this analysis can inform targeted strategies aimed at improving retention rates within organizations facing challenges related to customer attrition.

Moreover, the development of a user-friendly Streamlit application allows stakeholders to input customer data and receive real-time predictions on churn likelihood, thereby enabling proactive measures to enhance customer loyalty and reduce churn. This holistic approach not only aids in understanding customer behavior but also equips businesses with actionable strategies to foster long-term relationships with their clientele

9. Future Work

Future work could involve further tuning hyperparameters for optimal performance across selected models or exploring additional datasets containing more diverse demographic information or behavioral patterns relevant to churn prediction. Moreover, integrating real-time data streams could enhance predictive capabilities by allowing organizations to respond proactively rather than reactively when customers exhibit signs of potential churn.