Bank Churn Modeling Project Report

Introduction

This project aims to analyze and model customer churn behavior using a bank's customer dataset. The primary objective is to identify patterns and build predictive models to anticipate customer churn, which can help the bank develop strategies to retain valuable customers.

Data Overview

I chose a data set from kaggle. It is a synthetic dataset containing 10,000 entries. I found this a sufficient size to run off a local computer for learning purposes. The dataset used in this project, "Churn Modeling.csv," contains various features related to customers' demographics, account information, and churn status. The key variables include:

- **CreditScore**: Customer's credit score.
- **Geography**: Customer's location.
- **Gender**: Customer's gender.
- **Age**: Customer's age.
- **Tenure**: Number of years the customer has been with the bank.
- **Balance**: Customer's account balance.
- **NumOfProducts**: Number of bank products the customer uses.
- **HasCrCard**: Whether the customer has a credit card.
- **IsActiveMember**: Whether the customer is an active member.
- **EstimatedSalary**: Customer's estimated salary.
- **Exited**: Whether the customer has left (1) or stayed (0).
- https://www.kaggle.com/datasets/santoshd3/bank-customers/code

Data Exploration and Preprocessing

Initial Exploration

- **Null Values**: The dataset was checked for missing values, ensuring data completeness.

- **Descriptive Statistics**: Key summary statistics of the dataset were calculated to understand the data distribution and variability.
- **Visual**: The data was graphed to look at the distribution curves

Data Transformation

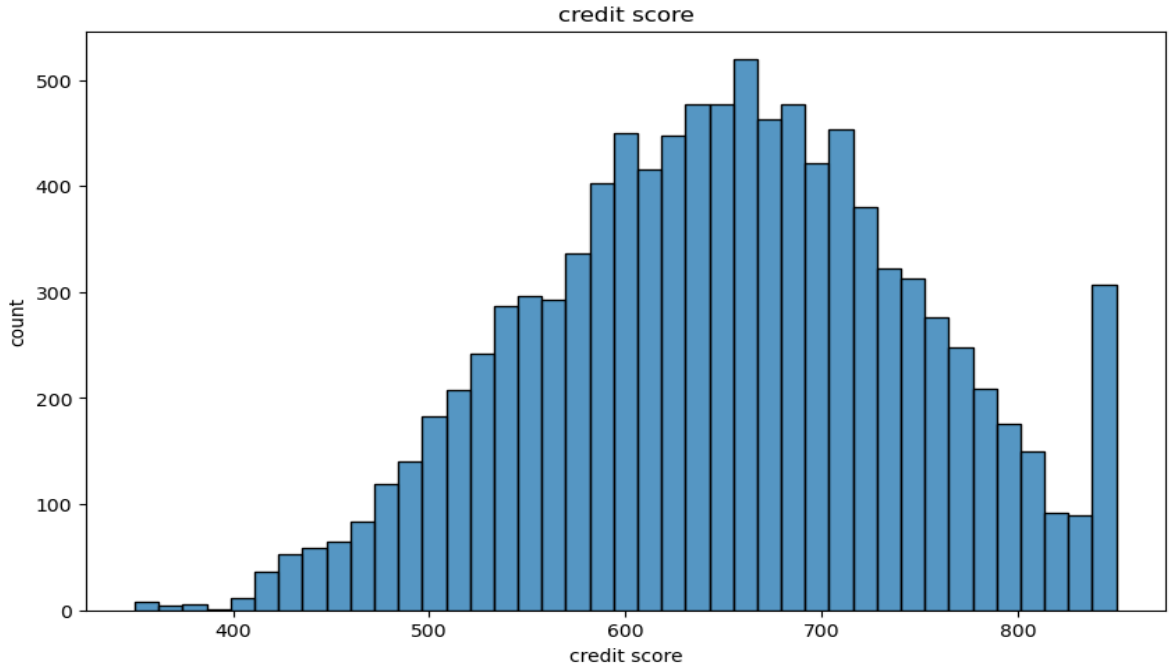
To facilitate analysis, the dataset underwent the following transformations:

- **Binning Credit Scores**: The credit scores were categorized into bins such as 'Very bad,' 'Bad,' 'OK,' 'Good,' and 'Excellent' to simplify analysis and visualization.
- **Salary_balance ratio**: a new column of a ratio between customer salary and balance in the bank was created to get an idea how much customers are saving
- **Correlation matrix**: the features were plotted to help visualize which features have the strongest effect on churn rates

Data Visualization

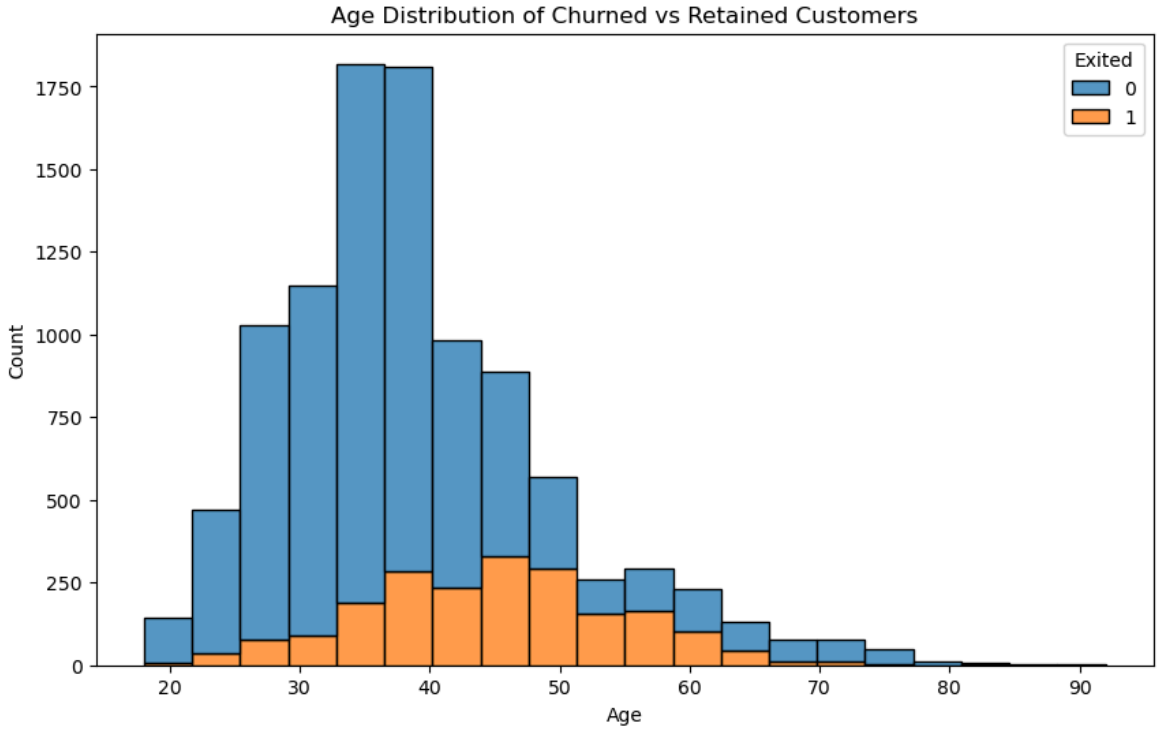
Visualizations were created to understand the distribution of key features and their relationship with churn:

- **Credit Score Distribution**: A histogram of credit scores was plotted, showing most customers have scores between 600 and 700.

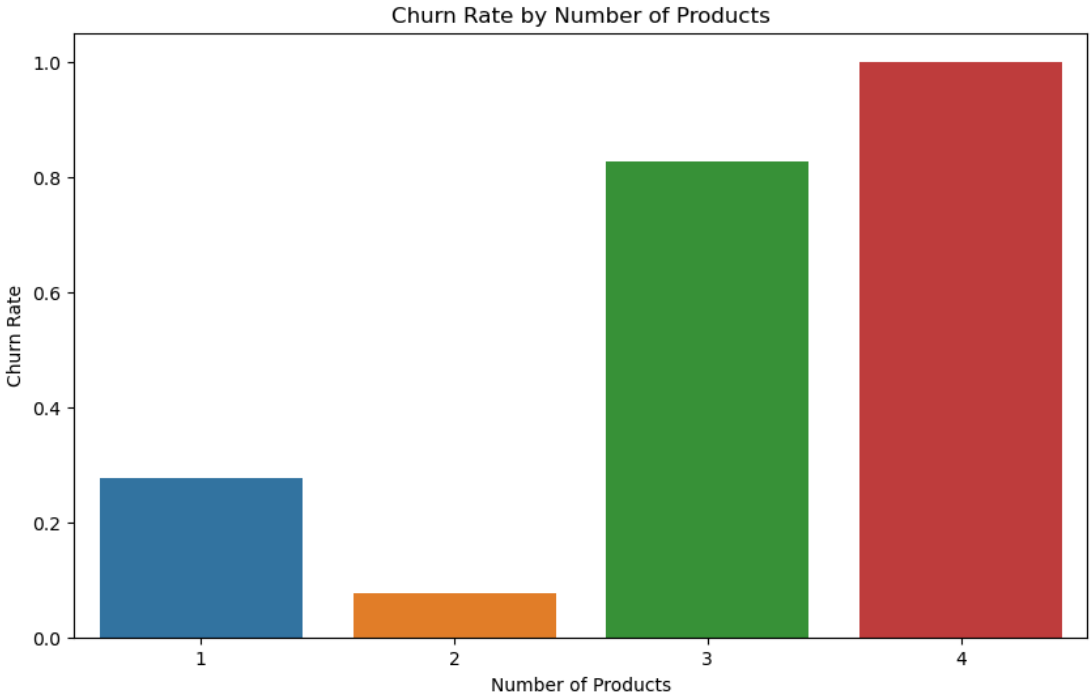


-

- **Age Distribution**: An age distribution histogram indicated that most customers are in their late 20s to late 40s.

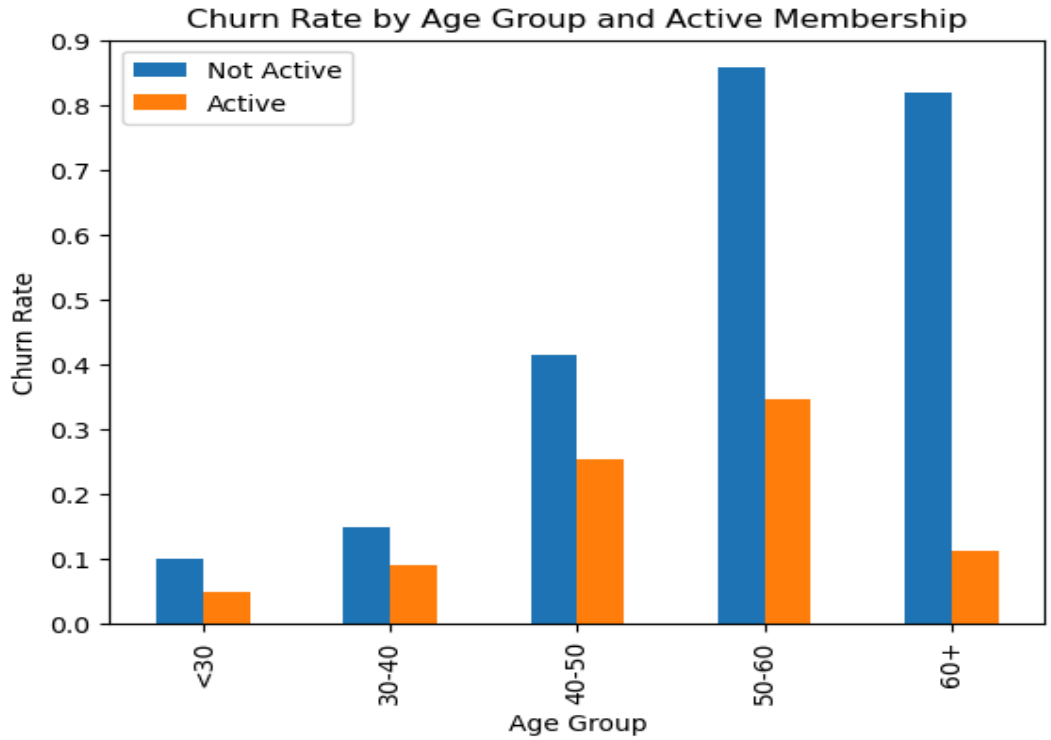


- **Number of products**: customers with more than 2 products are more likely to leave

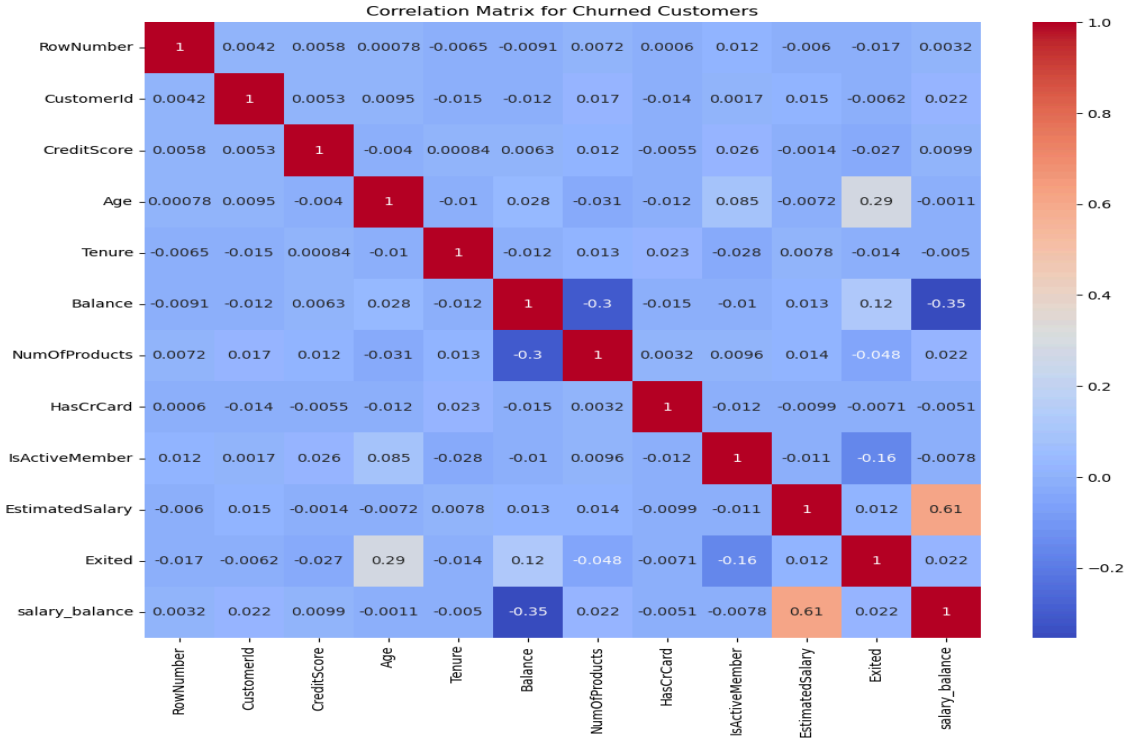


-

- Active membership: customers between the age of 40-60 years old are more likely to leave



- **Correlation matrix:** to help visualize which features bear the strongest relationship a correlation matrix was made to quickly identify features of interest



These visualizations help identify potential patterns and segments of customers more likely to churn.

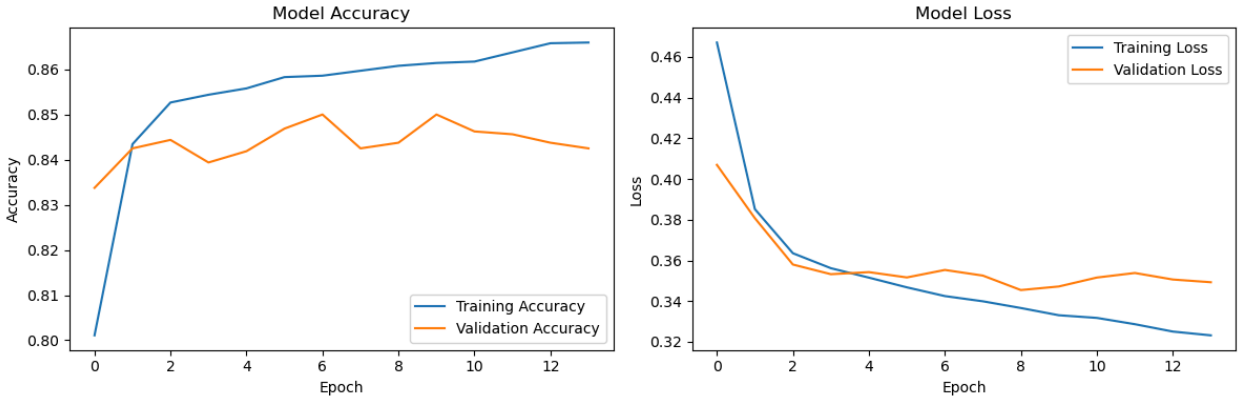
Modeling and Evaluation

The aim is to build a model that accurately predicts whether a customer will churn, allowing the bank to intervene and retain customers effectively.

- **Model Selection**: TensorFlow was used as the main model. This was done as a technical demonstration of my skills. Aside from this model Random Forest Classifier and Decision Tree Classifier offered similar results to our TensorFlow.
- **Model Training**: the data was split into the ‘exited’ feature vs the remaining features. The data was further divided into 80% training data vs 20% testing data for validation.
- **Evaluation**: A classification report was run on the results of the tensor flow model. It showed the model achieved an accuracy score of 86%. Accuracy scores from Random Forest Classifier also got an 86%.

Conclusion

The project provides insights into customer behavior, highlighting factors contributing to churn. By understanding these patterns, the bank can develop targeted retention strategies, improving customer satisfaction and reducing churn rates. We can see that the tensor flow model reached an accuracy of 86% on the training data and 85% on the validation data. After 2 epochs it looks like the model may start to over fit. Other models like Random Forest Classifier can achieve similar levels of accuracy which may be easier to use in a real world situation.



Future work could involve applying the model to real world data to see how well it can predict churn. The various models can be further tuned to work in real time for a real world application.