IT Club Internship (2025)

Week 6

Project: Predicting Customer Churn with Machine Learning

Internee Name: Beenish Guluna

Internee ID: ITCP-25/ML-002

Student of Computer Systems Engineering at

University of Engineering and Technology

Email Address: beenisharif248@gmail.com

LinkedIn Account: https://www.linkedin.com/in/beenish-guluna-980896295/

GitHub Account: https://github.com/Beenish-Guluna/

Internship Starting Date: 27th, January 2025

Dataset Overview

The dataset used for this project comes from a **bank** and contains customer information such as **credit score, geography, gender, age, tenure, balance, number of products, credit card ownership, activity status, estimated salary, and whether they exited the bank**. This structured data helps in identifying patterns that lead to customer churn.

Dataset Source: https://www.kaggle.com/datasets/shrutimechlearn/churn-modelling/data

My Approach Towards Model Making:

1. **Cleaned and Prepared the Data**: Checked for missing values and created new features to better understand customer behaviour. For example:
   - Added a feature to check if a customer's balance was zero.
   - Grouped customers by age and tenure.
   - Created a ratio of balance to salary to see how finances might impact churn.
   - Combined product usage and active membership into a single feature.
2. **Trained a Random Forest Model**: Chose Random Forest because it's great at handling complex data and avoids overfitting.
3. **Optimized the Model**: Used random search to find the best hyperparameters, like the number of trees and their depth. The best setup was:
   - max_depth: 20
   - min_samples_leaf: 2
   - min_samples_split: 8
   - n_estimators: 150
4. **Evaluated the Model**: Tested it on unseen data and got:
   - **87% accuracy** (up from 85% before tuning).
   - A detailed look at feature importance showed that **balance-to-salary ratio**, **age groups**, and **active membership** were the most influential factors.

Findings:

- The model works well, with **87% accuracy**, but the improvement after tuning was small (just 2%). This means the initial model was already pretty good.
- The most important features were:

1. **Balance-to-Salary Ratio**: Customers with a high ratio are more likely to churn.
2. **Age Groups**: Younger and older customers behave differently.
3. **Active Membership**: Inactive customers are at higher risk of leaving.
4. **Number of Products**: Customers with fewer products are more likely to churn.

Function:

- The model can help identify customers at risk of leaving so the business can take action, like offering discounts or personalized deals.

Conclusion

- While a 2% improvement isn't huge, it shows the impact of tuning and feature selection. Exploring more features or different models could improve churn prediction further.