CS 244 Machine Learning
Professor Tjaden
Flora Mukako and Adriana Castillo

Banking Customer Churn Prediction

Banking is a fast-evolving industry shaped by political shifts, global events, and changing societal norms. These factors influence how people interact with money and drive the development of new technologies, tools, and features that aim to make financial services more intuitive and user-friendly. Our project focuses on the customer experience side of banking, specifically predicting customer churn. By analyzing patterns in user behavior across different types of customers, we aim to uncover trends that can inform how banks design their products and services. This offers a valuable opportunity to see how data can directly support better decision-making in product development and customer engagement.

The existing work about bank customer churn has focused on predicting which customers are likely to stop using certain banking services through performing deep data analysis on decision-making processes within the banking sector. Supervised learning methods like random forests, SVM, or neural networks have been used to identify factors influencing customer behavior and utilization fluctuations. Some studies have combined various machine learning methods with sampling techniques across churn datasets to improve prediction accuracy. For instance, one study found that ensemble methods, which combine multiple classifiers, provide more robust results by capturing patterns in customer data. Findings suggest that through extensive feature extraction and predictive modeling, machine learning can inform customer churn management strategies.

We obtained our dataset from Kaggle while searching for data related to customer behavior in financial services, including credit scores, loans, and investments. The dataset includes information on 10,000 users, with features such as credit score, geography (Germany, Spain, or France), gender, age, tenure, balance, number of products, credit card ownership, estimated salary, and whether the customer exited the bank. We removed unnecessary columns like customer identification and names to prepare the data for our machine learning models. We also converted non-numerical variables into numerical formats, for example, by encoding the geography column and converting gender into binary values.

To build a fair and reliable model, we first analyzed the distribution of key features in the dataset. This helped us understand its overall structure and diversity, and ensured that no major imbalances were present. We found that gender was fairly balanced, with 5,457 male and 4,543 female users. We also examined credit scores, using a threshold of 650 to compare groups: 5,063 users had scores above 650, while 4,937 fell below.

We used a 70% training, 15% validation, and 15% testing split to ensure an unbiased representation of the data and to properly evaluate model performance. To determine which model to use, we trained and tested three different machine learning algorithms: logistic regression, random forest, and K-nearest neighbors (KNN).

For logistic regression, we scaled the training data to prevent features with different numerical ranges, such as credit score and estimated salary, from disproportionately influencing the model. We also applied the same scaling transformation to the prediction inputs to maintain consistency and avoid bias from mismatched feature magnitudes. In contrast, we did not scale the data for our random forest model, as

tree-based algorithms are not sensitive to feature scale. During implementation, we also performed hyperparameter tuning to improve model performance. For logistic regression, we tested different values of the regularization parameter C using powers of 10 to evaluate the effect of varying magnitudes. For KNN, we used odd values of k to avoid classification ties. We chose not to use the F1 score for evaluation because our classes were already balanced, instead, we used overall accuracy, which was sufficient for our analysis.

After training all three models, we compared their performance on the validation and test sets. The random forest model outperformed the others, achieving validation accuracy of 0.86 and test accuracy of 0.88. By comparison, the logistic regression model achieved validation accuracy of 0.696 and test accuracy of 0.698, while the KNN model reached 0.786 on validation and 0.796 on the test set. Based on these results, we selected the random forest model for our final analysis. To better understand which features most influenced the predictions, we conducted a weight extraction. This helped guide the creation of example user profiles and informed us which features we kept fixed versus which we varied to observe changes in prediction. The most influential features included age (0.249), balance (0.146), estimated salary (0.139), and credit score (0.133).

During implementation, we also performed hyperparameter tuning to improve model performance. For logistic regression, we tested different values of the regularization parameter C using powers of 10 to evaluate the effect of varying magnitudes. For KNN, we used odd values of k to avoid classification ties. We chose not to use F1 score as our evaluation metric because our classes were already balanced; instead, we used overall accuracy, which was sufficient for our analysis.

In our analysis of customer churn using the bank dataset, we started by understanding the influence of each feature on the results. We computed the feature weights of both logistic regression and random forest models. One of our initial assumptions was that a customer’s geography might significantly influence their likelihood to churn, potentially due to regional financial behaviors or policy differences. However, after analyzing the feature weights and importance scores from both models, we found that geography had minimal impact on churn prediction. This insight led us to pivot our project’s direction. Instead of forming hypotheses around geographical behavior, we chose to focus on more influential features like age and salary that influenced the model's decision-making the most.

To better understand the model’s behavior, we manually created a smaller test dataset with fixed features such as geography (France), tenure (5 years), credit score (500), and gender (male), with a set of pre-selected features that had a strong predictive influence on the model. We worked with values for age (20, 45, 70, 95), balance (30,000, 60,000, 120,000), number of products (1, 4) and 10 varying salary values, to have a fairly smaller dataset for us to identify patterns in the data and understand how these specific changes influenced the model’s predictions. The random forest model consistently predicted that users enrolled in four products were more likely to churn, regardless of whether they had high or low balances and salaries. This was unexpected, as we initially assumed that users with higher product engagement would be less likely to exit. However, we also recognize that our preselected values were relatively narrow in scope, which may have influenced the model's predictions.

Although we scaled our data and examined feature distributions to promote fairness, there are still limitations on the model’s interpretations. The data only includes users from three Western European

countries (Germany, France, and Spain), which limits representation and makes it difficult to draw broad conclusions. Additionally, in our predictive examples, we fixed certain feature values, such as gender being set to male, because of had minimal influence on model output (e.g., gender was weighted 0.249, while balance was much higher at 0.146). Although this helped simplify our analysis, it could introduce one-sided assumptions and limit inclusivity. The low impact of certain features does not eliminate their relevance, and the lack of diversity in both data and features creates bias.

Reflecting on our experience working on the bank user churn analysis project, we encountered different challenges and thought about ideas for future improvements. One of the challenges was working with a large, pre-defined dataset. Even though it was advantageous to have a variety of set features, it was hard for us to understand the nuances of the dataset and identify patterns that could have explained our model's behavior. The size and complexity of the dataset limited our ability to intuitively interpret feature interactions, which led us to create a smaller dataset that we can easily interpret. This smaller dataset allowed us to use specific features and better observe how those changes influenced predictions. Regarding machine learning techniques employed, we focused on random forest. If we had more time, we could have experimented with models like linear regression to predict other variables that inform us better on why certain users churn, like tenure, to estimate how long different types of users typically stay with the bank. In terms of future work, we think it would help to have more informative features like customer complaints, transaction frequencies, so that there are more insights about why customers decide to churn or not, instead of just having static features. Additionally, having longitudinal data, such as annual or monthly balances, and how they fluctuate over time, would provide a more comprehensive context on users' behavior.