CS3904 Internship
Executive Summary:
Data Science Churn Analysis for Credit Union Members

Introduction

This executive summary presents the findings of a comprehensive data science churn analysis conducted for IT consortium Open Technology Solutions' (OTS), specifically for their constituent member Bellco credit union. The primary objective of this analysis was to predict whether credit union members would close their current accounts or remain with the credit union. The analysis utilized a lookback window of January 1, 2022, to August 1, 2022, for account data, with a careful consideration of the preceding 30-days for closed accounts, only, to avoid capturing future information. Associated account metadata was examined from September 1, 2021, to August 1, 2022, (an additional 90-days) to capture peripheral information relevant to the analysis.

Methodology

The analysis was carried out through a series of Jupyter Notebooks, encompassing several steps, including data extraction from SQL databases in Snowflake, exploratory data analysis, feature engineering, and modeling. The credit union's financial system data was harnessed through SQL queries within Snowflake, which required establishing authentication via a YAML configuration file. The analysis encompassed both closed and open accounts, with a focus on feature engineering for effective utilization with a Random Forest classification model.

Data Exploration and Feature Engineering

An initial step involved reviewing existing database tables to identify account status changes and unique identifiers for closed accounts. A systematic approach was subsequently taken to identify salient information among over 1,500 database tables. A protracted and focused effort

was spent on identifying important information. Key features were selected, such as non-sufficient funds fees, ATM withdrawals, account lockout flags, average account balance, account age, primary account owner's age, associated zip code, and account type.

Modeling and Evaluation

Feature engineering was followed by the construction of a Pandas DataFrame to hold the selected features. The RandomForestClassifier algorithm from Sklearn was chosen due to its minimal preprocessing requirements and reputation for accuracy. The model exhibited notable success in predicting account closure, with high precision, recall, and F1-Score for both closed and open accounts. The overall accuracy of the model stood at an impressive 88%, with balanced precision and recall. The Receiver Operating Characteristic (ROC) analysis yielded an area under the curve of 0.94, further underscoring the robustness of this model for churn analysis.

Key Findings

The Random Forest algorithm provided insights into the significance of various features for predicting account closure. Notably, account type such as consumer loan and checking account emerged as pivotal factors. The count of transactions and the occurrence of ATM withdrawals also held considerable importance in the model's predictive capacity. This analysis reaffirmed the efficacy of the chosen algorithm in deriving meaningful insights from the data.

Conclusion

This data science churn analysis successfully predicted whether credit union members would close their accounts or remain active, leveraging Snowflake for data access and Random Forest for modeling. The comprehensive approach to feature selection and engineering, coupled with a robust model evaluation, led to high predictive capabilities. The analysis reaffirmed this intern's aptitude for distributed computing and functional programming, while also showcasing the value of prior credit union experience. The successful execution of this analysis solidified this intern's choice of major and demonstrated the potential of machine learning within finance.

Appendix

Jupyter Notebook Files

File	Description
closed accounts_2.0_written_to_table.ipynb	Logic for closed accounts is determined and implemented. A Snowflake connection is first established using a separate YAML configuration file.
open_accounts_written_to_table.ipynb	Logic for open accounts is determined and implemented. The set difference between closed accounts and active accounts with at least 30-processing transactions for the given lookback period is taken.
table_view_exploration _feature_engineering.ipynb	Represents a cursory exploration of 1,500+ database tables held within OTS' Snowflake instance including preliminary feature mapping heuristics. Code blocks containing PII have been curated to remove PII per request.
account_metadata.ipynb	This is a fine-tuned feature mapping which culminates in a final feature set. The final feature set is created as a Pandas DataFrame and is pickled. Code blocks containing PII have been curated to remove PII per request.
model_and_evaluation.ipynb	Takes previously determined finalized feature set and creates a RandomForestClassifier model. A model evaluation subsequently takes place with a confusion matrix, classification report, and ROC curve.