DSC680-T301 Applied Data Science
4/14/2023
Joshua Greenert
Project 2: Customer Churn Analysis

Topic

The goal of this project will be to use machine learning to accurately predict customer churn for two separate businesses.

Business Problem

The banking and telecommunication industries are notable as having a high rate of churn, which can be costly for companies in terms of lost revenue and acquisition for new customers. The business problem that this project aims to solve is two-fold: depending on the business, there may be unique solutions to each business respectively, and a machine learning model can be created that can accurately predict which customers are most likely to churn based on the industry type. Additionally, this project aims to identify the factors that contribute to the churn for each industry type as well.

Datasets

The datasets used will be sourced from Kaggle — a massive dataset supplier — and can be found directly at the links in the references section below as well as in the following bullet points.

- **Telco Customer Churn Dataset**
  - https://www.kaggle.com/datasets/blastchar/telco-customer-churn
  - This dataset contains a total of 21 columns and 7043 rows of data. The data within contains information related to customers along with a target column — whether the customer has churned or not — that will be used to determine the significance of each feature for churn evaluation.

- **Bank Customer Churn Dataset**
  - https://www.kaggle.com/datasets/gauravtopre/bank-customer-churn-dataset
  - This dataset contains a total of 12 columns and 10000 rows of data. The data within contain features related to the customer such as credit score, gender, age, etc. along with a churn column to determine whether the customer is still with the respective bank.

Methods

The analysis methods used for this project include exploratory data analysis to understand the patterns and relationship within the data, feature engineering to select and process the relevant variables, and the application of machine learning models for predicting whether the customer is likely to churn. The main models that will be utilized in this process include logistic regression, decision trees, and random forest regression. Each model will be evaluated for its consistency and accuracy to determine if there is a "best" model approach based on a given industry.

Ethical Considerations

One potential ethical concern related to the process of predicting customer churn is the use of customer data without the user's explicit consent. To mitigate this ethical consideration, the project will ensure that all data used is publicly available and anonymized. Any findings or models will be shared in a transparent and accessible manner, with clear explanations of the methods used.

Challenges/Issues

The main challenge of this project will be the choice of what features to include in each model as this choice will be critical to determine the features with the strongest predictive

power. Moreover, any missing or incomplete data within these datasets may lessen the value of the accuracy of each respective model attempted.

References

BlastChar. (2019). Telco Customer Churn. Kaggle. https://www.kaggle.com/blastchar/telco-customer-churn

Topre, G. (2021). Bank Customer Churn Dataset. Kaggle. https://www.kaggle.com/gauravtopre/bank-customer-churn-dataset