Churn Modelling

Churn modelling dataset contains features of bank customers that have either cancelled their accounts or stayed with the bank. It solves an important binary classification problem which is to identify customer retention based on previous historical data. The dataset is publicly available and it is obtained from Kaggle[1]. The dataset contains a total of 10,000 instances and 13 features. These features include row number, customer id, surname, credit score, geography, gender, age, tenure, balance, number of products, has credit card, is active member, and estimated salary. The target variable is 0 if the customer retained with the bank and 1 if the customer has closed their account with the bank. After loading the dataset, the next step was to perform exploratory data analysis. The dataset contains categorical and numerical variables, and it also does not have any missing values. After performing data visualization, it was discovered that the dataset is highly imbalanced as it contains 7963 samples belonging to class 0 and 2037 instances belonging to class 1. To balance the dataset only 2037 samples belonging to class 0 were selected. Feature selection was also performed to remove all features that are not relevant such as row number, customer id, surname, geography, and gender. There were two experiments performed, the first one included the gender and geography features and in the second experiment these features were dropped. In the first experiment one hot encoding was performed on geography and gender but including these features does not seem to improve the models performance therefore these features were removed in the second experiment. The dataset was shuffled and split into training and test set with a split ratio of 70:30. Next, the decision tree classifier model was created with criterion equals entropy and the model was trained. It produced a test accuracy of 68% and the most important features were analyzed using the .feature_importances_ function with age being the most important feature with an information gain of 0.25190. For this particular dataset, the decision tree classifier was not the best model, however hyperparameter tuning like GridSearchCV can be used to obtain a better performing model.

References:

[1] Shruti_Iyyer, “Churn modelling,” *Kaggle*, 03-Apr-2019. Available: https://www.kaggle.com/datasets/shrutimechlearn/churn-modelling.