**TEBONG ROLAND**

**DSC680-T301 Applied Data Science (2243-1)**

[**5.1 Project 2: Proposal and Data Selection/Milestone 1**](https://cyberactive.bellevue.edu/webapps/assignment/uploadAssignment?content_id=_15084584_1&course_id=_522998_1&group_id=&mode=view)

**Bank Term Deposit Subscription Prediction**

**Topic:**

"Machine Learning Strategies for Optimizing Revenue in the Portuguese Bank: Predictive Analytics for Customer Term Deposit Subscriptions"

**Context:**

Facing a decline in revenue due to a shortfall in long-term customer investments, the Portuguese Bank seeks to revitalize its financial standing. This project centers on the use of machine learning techniques to predict customer propensity to subscribe to term deposits. By identifying likely subscribers, the bank can channel its marketing resources more effectively, enhancing both customer engagement and its investment portfolio.

**Business Problem:**

The challenge lies in accurately predicting which customers are inclined to subscribe to term deposits. This predictive insight will enable the bank to tailor its marketing efforts, leading to increased subscription rates and, consequently, a boost in long-term investment revenue. The project will use historical customer data to build a model that not only predicts customer behavior but also sheds light on the factors influencing their decision to invest in term deposits.

**Data Description:**

The dataset encompasses a wide array of customer-related information[[1]](#footnote-1):

* Age - Numeric representation of the age of a person.
* Job - Categorical, nominal classification of a person's job type, such as 'admin.', 'blue-collar', 'entrepreneur', 'housemaid', 'management', 'retired', 'self-employed', 'services', 'student', 'technician', 'unemployed', 'unknown'.
* Marital - Categorical, nominal status indicating the marital situation of a person, options include 'divorced', 'married', 'single', 'unknown'; note: 'divorced' also includes widowed individuals.
* Education - Categorical, nominal classification of a person's education level, with options like 'basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'illiterate', 'professional.course', 'university.degree', 'unknown'.
* Default - Categorical, nominal indicator of whether a person has credit in default, with options 'no', 'yes', 'unknown'.
* Housing - Categorical, nominal status indicating if a person has a housing loan, with options 'no', 'yes', 'unknown'.
* Loan - Categorical, nominal status showing if a person has a personal loan, options include 'no', 'yes', 'unknown'.
* Contact - Categorical, nominal type of contact communication used, such as 'cellular' or 'telephone'.
* Month - Categorical, ordinal representation of the last contact month of the year, options range from 'jan' to 'dec'.
* Day\_of\_Week - Categorical, ordinal representation of the last contact day of the week, including 'mon', 'tue', 'wed', 'thu', 'fri'.
* Duration - Numeric value representing the last contact duration in seconds. Note: This attribute significantly impacts the output target, e.g., duration=0 leads to y='no'.
* Campaign - Numeric count of the number of contacts performed during the current campaign for this client, including the last contact.
* Pdays - Numeric count of the days passed since the client was last contacted from a previous campaign; 999 indicates the client was not previously contacted.
* Previous - Numeric count of the number of contacts performed before the current campaign for this client.
* Poutcome - Categorical, nominal outcome of the previous marketing campaign, with options 'failure', 'nonexistent', 'success'.
* Y (Target Variable) - Binary indication of whether the client has subscribed to a term deposit, with options 'yes', 'no'.

**Methodology:**

The project will undertake the following steps:

* Data Preprocessing: Cleaning data, handling missing values, encoding categorical variables, and normalizing numerical data.
* Exploratory Data Analysis (EDA): A thorough analysis to understand customer demographics, financial behaviors, and responses to previous marketing efforts.
* Feature Engineering: Identifying and creating relevant features that could influence a customer's decision to subscribe to a term deposit.
* Model Selection and Training: Experimenting with various machine learning algorithms such as, Decision Trees, Random Forest, and ANN. The focus will be on models that can handle binary classification effectively.
* Model Evaluation: Using metrics like accuracy, precision, recall, F1-score, and ROC-AUC curve to assess the performance of each model. Special attention will be paid to recall to minimize false negatives, as missing out on potential subscribers is critical for the bank.
* Model Optimization: Implementing techniques like cross-validation, hyper parameter tuning (using GridSearchCV), or ensemble methods to enhance model performance.

**Ethical Considerations:**

Utmost care will be taken to ensure ethical compliance:

* Adherence to GDPR and other relevant data privacy regulations.
* Ethical use of data, ensuring no discriminatory biases are present in the model.
* Transparency in model decisions, with an emphasis on explainable AI.

**Challenges/Issues:**

The project will likely encounter several challenges, including but not limited to:

* Dealing with imbalanced datasets where the number of subscribers is significantly lower than non-subscribers.
* Avoiding over fitting to ensure that the model generalizes well to new, unseen data.
* Navigating the dynamic nature of customer behavior which can be influenced by external economic factors.

**References:**

Etaiwi, W., Biltawi, M., & Naymat, G. (2017). Evaluation of classification algorithms for banking customer’s behavior under Apache Spark Data Processing System. *Procedia computer science*, *113*, 559-564.

Başarslan, M. S., & Argun, İ. D. (2018, April). Classification of a bank data set on various data mining platforms. In *2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT)* (pp. 1-4). IEEE.

Valarmathi, B., Chellatamilan, T., Mittal, H., Jagrit, J., & Shubham, S. (2019, May). Classification of imbalanced banking dataset using dimensionality reduction. In *2019 International Conference on Intelligent Computing and Control Systems (ICCS)* (pp. 1353-1357). IEEE.

Apampa, O. (2016). Evaluation of classification and ensemble algorithms for bank customer marketing response prediction. *Journal of International Technology and Information Management*, *25*(4), 6.

1. https://www.kaggle.com/datasets/rashmiranu/banking-dataset-classification/data [↑](#footnote-ref-1)