**TEBONG ROLAND**

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

[**7.1 Project 2: Draft/Milestone 2**](https://cyberactive.bellevue.edu/webapps/assignment/uploadAssignment?content_id=_15084585_1&course_id=_522998_1&group_id=&mode=view)

**Banking Dataset Classification**

**1. Business Problem**

Machine Learning is utilized in this scenario to analyze and predict customer behavior, specifically targeting those likely to subscribe to a term deposit. The decline in revenue due to a lack of long-term investments can be addressed by identifying patterns in customer data that correlate with the likelihood of subscribing to term deposits. Using ML algorithms can help in efficiently processing large volumes of customer data, uncovering insights that are not immediately apparent, and automating the prediction process to assist in making data-driven marketing and business decisions.

* The task is to predict whether a bank's customers will subscribe to a term deposit, which is a binary classification problem.
* The data consists of historical customer information from the bank, which includes demographic details, account information, and past interactions with the bank.
* Performance metrics could include accuracy, precision, recall, F1-score.

**2. Background/History**

There has been a revenue decline in the Portuguese Bank and they would like to know what actions to take. After investigation, they found that the root cause was that their customers are not investing enough for long term deposits. So the bank would like to identify existing customers that have a higher chance to subscribe for a long term deposit and focus marketing efforts on such customers.

**3. Data Explanation**

* **Data Loading:** The analysis begins with importing data from the banking\_data.csv' file using panda’s library. This dataset contains 32950 records across 16 columns.
* **Data Dictionary:** The dataset encompasses various attributes. The detailed data dictionary is given below.

1. **\*\*age:\*\*** Numeric - Age of a person

2. **\*\*job:\*\*** Categorical, nominal - Type of job ('admin.', 'blue-collar', 'entrepreneur', 'housemaid', 'management', 'retired', 'self-employed', 'services', 'student', 'technician', 'unemployed', 'unknown')

3. **\*\*marital:\*\*** Categorical, nominal - Marital status ('divorced', 'married', 'single', 'unknown'; note: 'divorced' means divorced or widowed)

4. **\*\*education:\*\*** Categorical, nominal - ('basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'illiterate', 'professional.course', 'university.degree', 'unknown')

5. **\*\*default:\*\*** Categorical, nominal - Has credit in default? ('no', 'yes', 'unknown')

6. **\*\*housing:\*\*** Categorical, nominal - Has housing loan? ('no', 'yes', 'unknown')

7. **\*\*loan:\*\*** Categorical, nominal - Has personal loan? ('no', 'yes', 'unknown')

8. **\*\*contact:\*\*** Categorical, nominal - Contact communication type ('cellular', 'telephone')

9. **\*\*month:\*\*** Categorical, ordinal - Last contact month of the year ('jan', 'feb', 'mar', …, 'nov', 'dec')

10. **\*\*day\_of\_week:\*\*** Categorical, ordinal - Last contact day of the week ('mon', 'tue', 'wed', 'thu', 'fri')

11. **\*\*duration:\*\*** Numeric - Last contact duration, in seconds. Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no')

12. **\*\*campaign:\*\*** Numeric - Number of contacts performed during this campaign and for this client (includes last contact)

13. **\*\*pdays:\*\*** Numeric - Number of days that passed by after the client was last contacted from a previous campaign (999 means the client was not previously contacted)

14. **\*\*previous:\*\*** Numeric - Number of contacts performed before this campaign and for this client

15. **\*\*poutcome:\*\*** Categorical, nominal - Outcome of the previous marketing campaign ('failure', 'nonexistent', 'success')

**\*\*Target Variable:\*\***

- **\*\*y:\*\*** Binary - Has the client subscribed to a term deposit? ('yes', 'no')

**4. Methods & Analysis**

**4.1 Exploratory Data Analysis (EDA):**

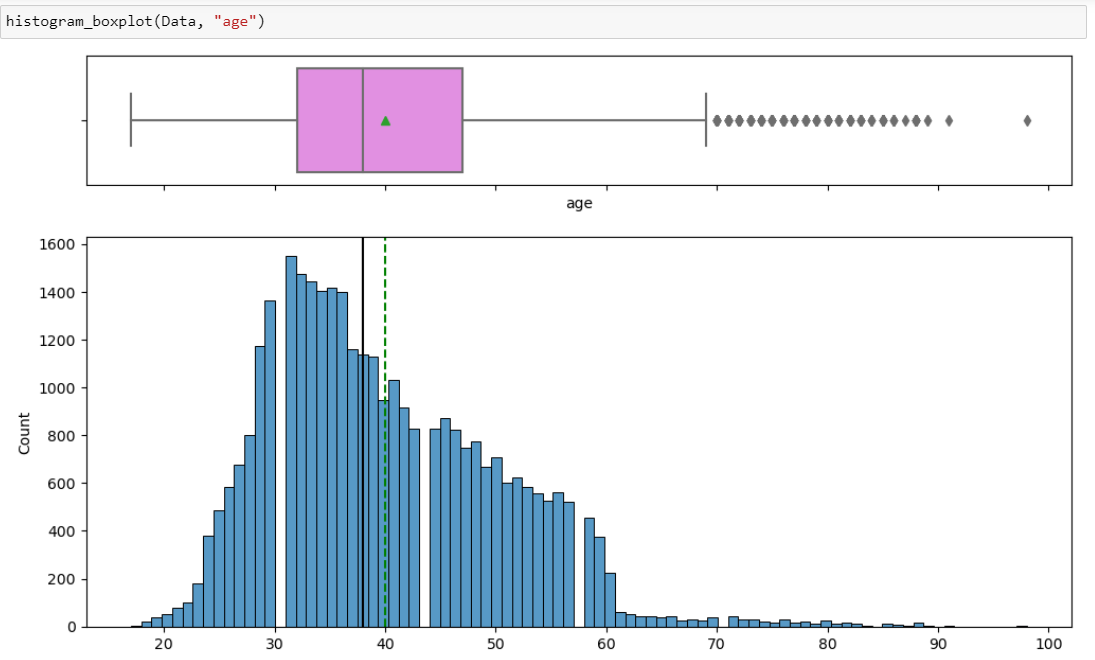
Let's check the statistical summary of the data.

* Age: The average age of the bank's clients in this dataset is around 40 years old, with the youngest being 17 and the oldest 98. Most clients are in their 30s to 40s, which is typical for a working and economically active population.
* Duration: The average duration of the last contact with the clients was about 258 seconds, but there's a big range, from just a few seconds (10 seconds) to over an hour (4918 seconds). This suggests there's a lot of variation in how long the bank's employees are talking to clients.
* Campaign: On average, a client was contacted about 2.5 times during the current campaign, with a maximum of 56 times. This could imply that while some clients need more persuasion, the bank doesn't usually go beyond a handful of contacts.
* Pdays: The '999' value indicates that many clients were not previously contacted, which aligns with the 'previous' column average being close to zero. This means the current campaign might be the first direct marketing interaction for most clients.

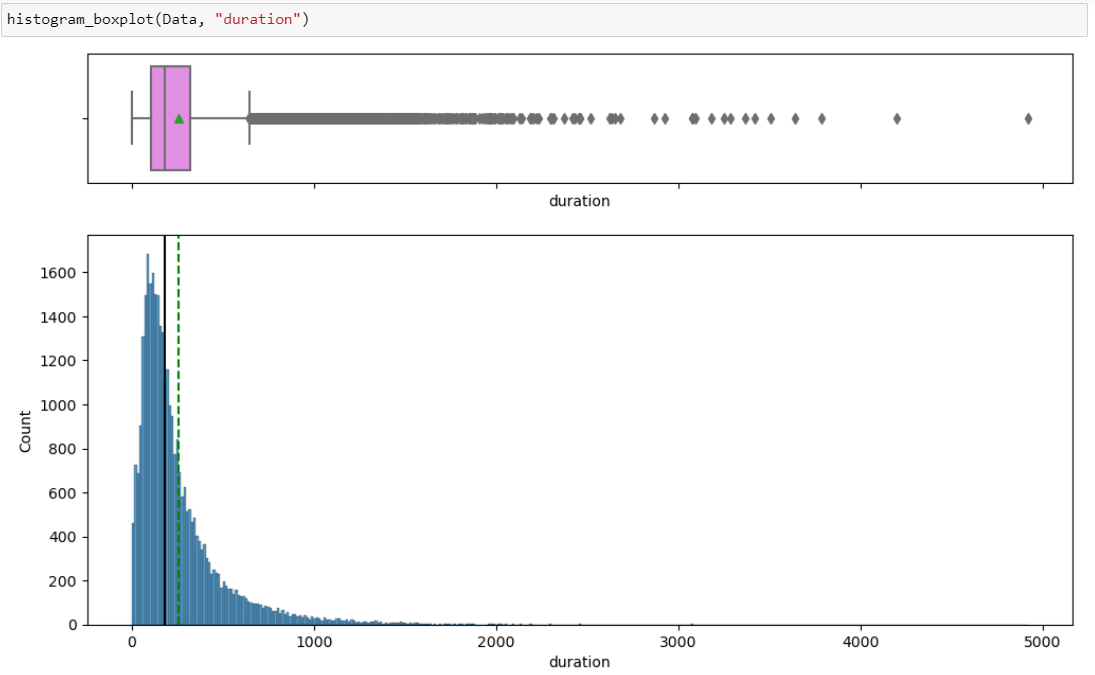
Let’s check unique values in the dataset.

* The duration has the most unique value following the age column with 75 unique values. It can be observed the y column, which is the target variable has 2 unique values as it suggests it's a binary classification problem.
* Job: The most common job among clients is 'blue-collar', followed by 'admin.' and 'technician'. There are a few clients with 'unknown' job types.
* Marital Status: Most clients are married, then single, with a small number marked as 'unknown'.
* Education: A large portion of clients have a 'university.degree', and quite a few have completed 'high.school'. There are very few 'illiterate' clients.
* Credit in Default: Most clients do not have credit in default, and a significant number have 'unknown' status in this regard.
* Housing Loan: More clients have a housing loan than those who do not.
* Personal Loan: The majority of clients do not have a personal loan.
* Contact: 'Cellular' seems to be the preferred way of contacting clients, more so than 'telephone'.
* Month of Last Contact: 'May' is the most common month for last contact with clients, followed by 'July', 'August', and 'June'. The winter months like 'December' see very little activity.
* Day of the Week: The distribution of the last contact days is quite even across the weekdays, with 'Thursday' being slightly higher.
* Outcome of the Previous Marketing Campaign: Most clients have not been contacted before ('nonexistent'), and there are more 'failures' than 'successes'.

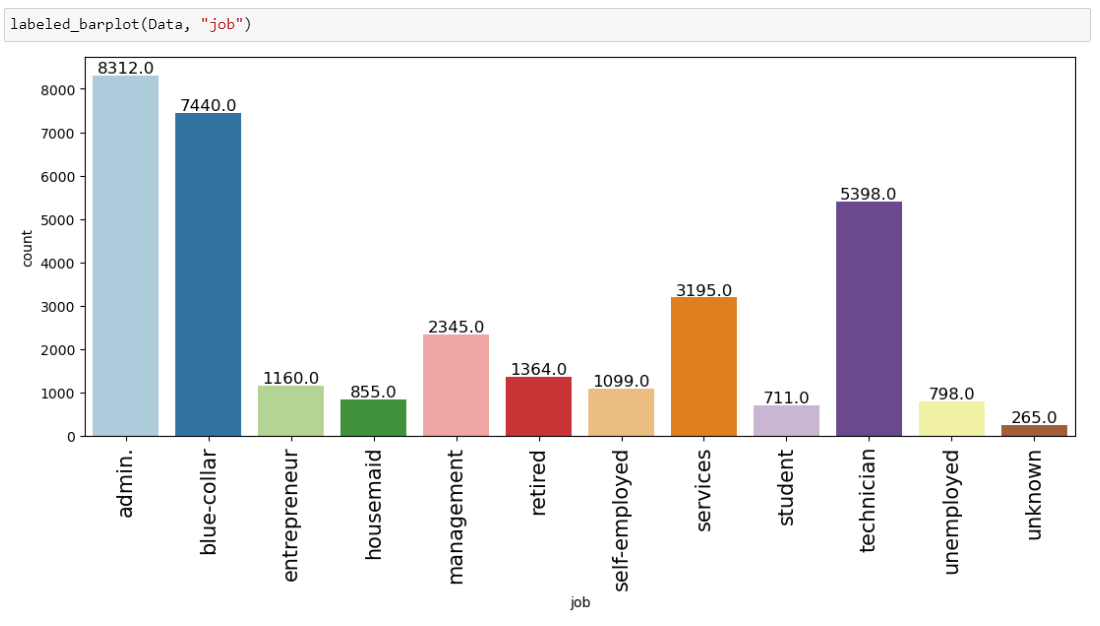
Let’s analyze key variables using univariate analysis:



* It can be observed that most clients are in their 30s and 40s, with fewer clients as the age increases. The boxplot suggests that the middle half of the client ages are tightly packed around the median, indicating a consistent age range among most clients.

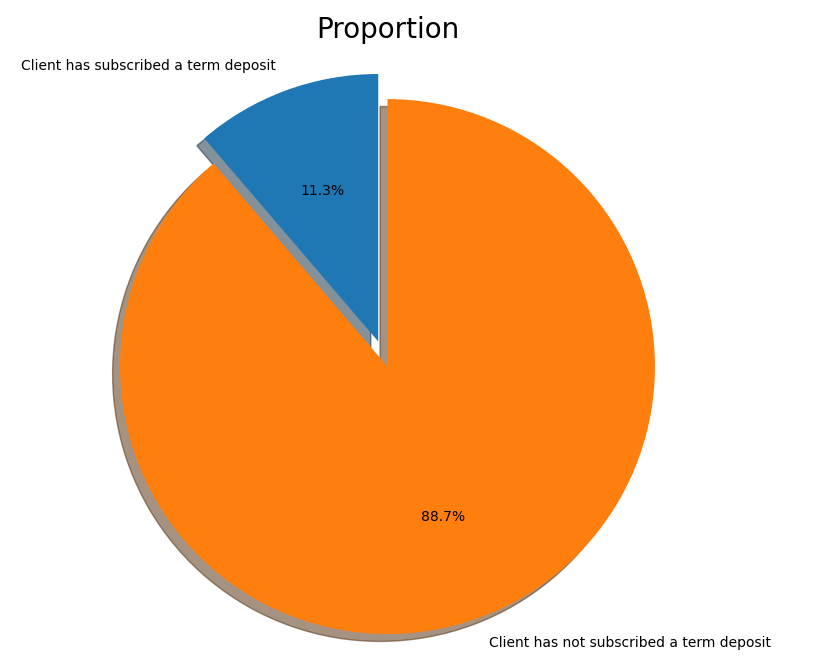


* The graph depicts the length of phone calls with bank clients. Most calls were short, with a steep drop-off as duration increases. There are some very long calls, but these are rare, as shown by the few dots far to the right on the boxplot, which represent outliers.



* The bar chart shows the job categories of the bank's clients. 'Admin.' and 'blue-collar' are the most common jobs, with a lot of clients working in these roles. 'Student', 'unemployed', and 'unknown' categories have the fewest numbers, indicating they make up a smaller portion of the bank's clientele.

Let’s check the target variable distribution:



* This pie chart shows that the actual distribution of classes is itself imbalanced for the target variable.
* Only ~11% of the customers in this dataset have actually subscribed a term deposit.
* Hence, this dataset and problem statement represent an example of Imbalanced Classification, which has unique challenges in comparison to performing classification over balanced target variables.

**4.2 Data Preprocessing:**

Let’s preprocess the data:  


Now let’s Encoding Categorical features.



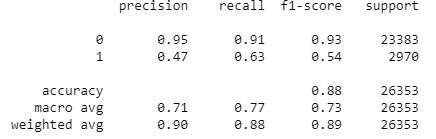
Now let’s split that data into training and test.



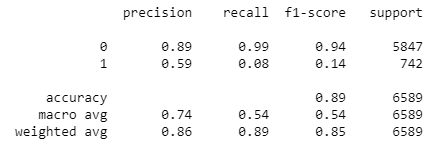
**4.3 Modeling**

Let’s build Artificial Neural Network, Decision Tree classifier and Random forest Classifier.

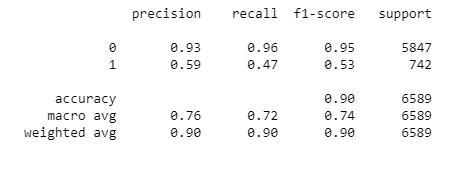
**Artificial Neural Network:** After fitting the training data into the ANN model the model has been tested on the test dataset, which was splitted above in the data preprocessing section. After testing the model the results are quite promising as shown below.



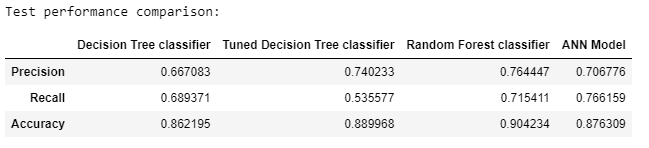
**Decision Tree Classifier:** After fitting the training data into the Decision tree classifier model the model has been tested on the test dataset, which was splitted above in the data preprocessing section. After testing the model the results are quite promising as shown below.



**Random forest Classifier:** After fitting the training data into the random forest classifier model the model has been tested on the test dataset, which was splitted above in the data preprocessing section. After testing the model the results are quite promising as shown below.



Let’s compare the models performance as shown below.



* It can be observed that the **Random Forest classifier** is outperforming other models with an accuracy of **90%**.

**5. Assumptions**

In the context of data collection, it is assumed that the information gathered from the Portuguese Bank's historical records is a representative sample of the broader customer population. This assumption hinges on the belief that the dataset encapsulates the diversity and characteristics of the bank's clientele accurately. However, it is crucial to acknowledge that any inherent biases or limitations in the data could influence the model's ability to generalize to the entire customer base. Additionally, in the broader business context, the assumption is made that the revenue decline is primarily attributed to customers' hesitancy in investing in long-term deposits, neglecting other potential factors that might contribute to the bank's financial challenges.

**6. Limitations**

The analysis faces several limitations that warrant consideration. Firstly, the findings heavily depend on the quality and completeness of the data contained in the 'banking\_data.csv' file. Incomplete or inaccurate information may compromise the accuracy and reliability of the models developed. Furthermore, the models' predictions are based on historical patterns, assuming that future customer behavior will follow similar trends. However, external factors not captured in the dataset, such as economic changes or regulatory shifts, may significantly impact the accuracy of the models in predicting subscription decisions.

**7. Challenges**

The primary challenges encountered in this analysis revolve around the imbalanced distribution of the target variable 'y,' indicating whether a client subscribed to a term deposit. The imbalanced classes, with a considerably lower percentage of positive cases, pose challenges in model training and may lead to a bias towards predicting the majority class. Additionally, the use of complex models, particularly the Artificial Neural Network, introduces the challenge of interpretability, making it difficult to provide transparent explanations for specific predictions.

**8. Future Uses/Additional Applications**

Looking ahead, the developed models can be leveraged for various purposes beyond the binary classification task. One potential application is customer segmentation, where the bank can categorize clients based on specific characteristics and behaviors identified by the models. This segmentation would allow for more targeted and personalized marketing strategies, enhancing the bank's ability to tailor its services to different customer segments. Moreover, continuous monitoring and updating of the model with new data could facilitate dynamic predictions, enabling the bank to adapt quickly to evolving customer trends and behaviors.

**9. Recommendations**

To enhance the effectiveness of the predictive models and the overall decision-making process, it is recommended to conduct a thorough analysis of feature importance. Understanding which factors significantly influence customer decisions can guide targeted marketing efforts and strategic planning. Additionally, implementing a feedback mechanism to gather insights from customers who both subscribed and did not subscribe to term deposits can provide valuable information for refining the models. This feedback loop can aid in uncovering customer preferences and improving the models' predictive capabilities over time.

**10. Implementation Plan**

The implementation plan for deploying the selected model into the operational systems of the bank involves careful consideration of integration and ongoing monitoring. The seamless integration of the model ensures that predictions are effectively incorporated into the decision-making processes of the bank. Furthermore, providing training to bank staff involved in interpreting and utilizing the model's predictions is essential. This training equips staff with the necessary skills to understand and act upon the model's outputs, fostering effective and informed decision-making within the organization.

**11. Ethical Assessment**

The ethical assessment encompasses various considerations, beginning with privacy concerns. It is imperative to ensure compliance with data protection regulations and implement robust measures to safeguard customer privacy throughout the analysis and deployment phases. Addressing fairness and bias is crucial to preventing discriminatory outcomes in model predictions. Regular assessments and adjustments should be made to mitigate biases and ensure equitable treatment of all customer segments. Transparency in communication with customers regarding the use of their data and the purpose of predictive models is essential for building trust and maintaining ethical standards throughout the implementation process.

**Questions:**

1. What is the primary business problem being addressed using machine learning in this scenario?
2. What is the target variable in the classification problem addressed in this report?
3. How many records and columns are there in the 'banking\_data.csv' dataset used for analysis?
4. What are some key attributes included in the data dictionary for the features used in the classification problem?
5. What are the assumptions made regarding the data collection process and its representativeness?
6. What limitations are highlighted in the analysis, specifically concerning data quality and model predictions?
7. What challenges are discussed, particularly related to imbalanced classes and the interpretability of complex models?
8. What future uses or additional applications are suggested for the developed predictive models?
9. What key recommendations are provided for enhancing the effectiveness of the predictive models?
10. How is the performance of different models, including Artificial Neural Network, Decision Tree, and Random Forest Classifier, compared in terms of accuracy?