**Multi-Week Assignment#2 – Classification**

**Banking Marketing**

**Context:**

The Banking Dataset of different customers is related to direct marketing campaigns (phone calls) of a Portuguese banking institution . The classification goal is to predict if the customer will subscribe (yes/no) a term deposit (variable y).

Term deposits are a major source of income for a bank. Your money is invested for an agreed rate of interest over a fixed amount of time, or term. The bank has various outreach plans to sell term deposits to their customers such as email marketing, advertisements, telephonic marketing, and digital marketing. Telephonic marketing campaigns still remain one of the most effective way to reach out to people. However, they require huge investment as large call centers are hired to actually execute these campaigns. Hence, it is crucial to identify the customers most likely to convert beforehand so that they can be specifically targeted via call.

**Assignment:**

**You may work independently or work with classmates (up to 3 students in a group), though each classmate must submit their own HTML/PDF file**. If you work with classmates, please list the names of the classmates you worked with at the top of your submitted homework.  Submitted Jupyter Notebooks **will not be graded.**

You are also encouraged to reach out to your instructor to discuss other options that make sense to the project as well.

**Part 1. Exploratory Data Analysis (45 points + 10 bonus points)**

**- Due by 9pm Sunday March 14th**

Notes:

1. Use Python basics, Data Science libraries (Numpy, Panda), and Data Visualization Libraries(Matplotlib, Seaborn, Plotly, etc.);
2. Please use the 10 steps in the Section 1 from Multi-Week Assignment#1 as a general guidance; You can modify the steps to improve the analysis if needed.
3. A summary of your exploratory data analysis and findings with bullet points at the beginning of the Jupyter Notebook;
4. Perform descriptive analytics on all variables to help understand the data, the distributions and basic info (e.g.: value\_counts(), .describe() for numeric variables: Mean, Standard Deviation, Median, Min, and Max, etc.);
5. Choose at least one proper plot type to perform data visualization on each of all the variables;
6. Organize your coding work step-by-step with comments.

**Part 2. Supervised Learning Classification (120 points + 20 bonus points)**

**- Submission in Jupyter Notebook PDF or HTML file**

**- Due by 9pm Sunday, March 20th**

Please use the similar modeling processing from Multi-Week Assignment#1 as a general guidance by using K-Nearest Neighbors, Logistic Regression, Decision Tree, and Random Forest Classification Algorithms; You can modify the steps to improve the analysis if you see properly.

1. Data Pre-Processing including import libraries, import dataset(s), encode categorical data fields, split data into train and test datasets (30 points)
2. Feature selection (rank the features) (10 points)
3. Train the following 4 models, predict the Test set results, evaluate the model performance, compute the accuracy with k-Fold cross validation

#### K-Nearest Neighbors model (15 points)

#### Logistic Regression model (15 points)

#### Decision Tree model (15 points)

#### Random Forest model (15 points)

1. Compare the 4 models and select the best model with data driven support (20 points)
2. Any additional data driven insights you discover (up to 20 bonus points)

**Part 3: A Final Summary Report (3-5 Presentation slides) (20 points)**

**- Submission in PPTX or PPT file**

**- Due by 9pm Sunday, March 20th**

Share analytic result summary, insights, conclusions, learning, and future improvements, etc.

***Important Note:*** Changes may occur to this multi-week assignment at the instructor's discretion. When changes are made, students will be notified via email and in-class announcement.