**CSED 426 Big Data**

**Assignment 4**

****

**Submitted by- Sajan Maharjan**

**POVIS id-** [**thesajan@postech.ac.kr**](mailto:thesajan@postech.ac.kr)

**Registration Number- 20182095**

Assignment 4 Report

# San Francisco Crime Classification Using Random Forest Classifier

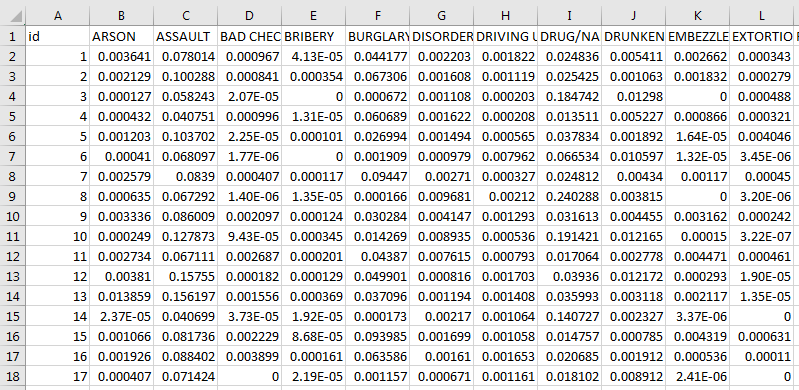
**Command to execute:**

|  |
| --- |
| ***python sanfrancisco\_cc\_RF.py*** |

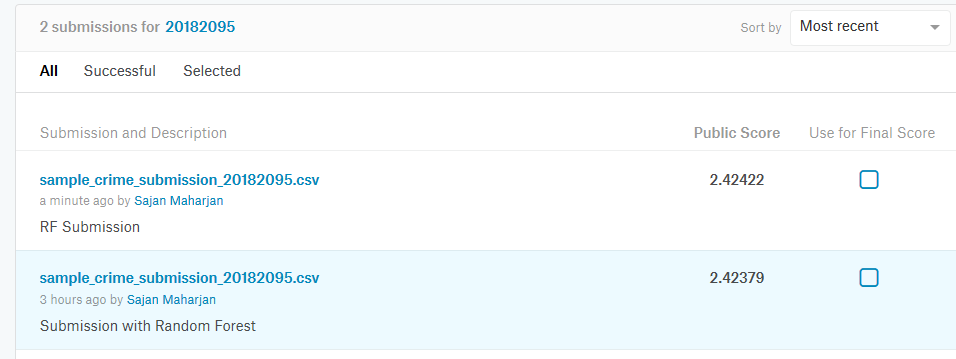
The above python program uses ***RandomForestClassifier*** to make multi-class crime predicitons. Broadly speaking, we first perform feature engineering of the *Dates* column from given CSV files for training and test data sets. This provides us with information such as year, month, day, hour, day of week, time of the day (night or not) which is useful is predicting crimes. Also, we apply feature engineering on the ***Address*** column to obtain information such as whether the crime scene is an intersection of two streets or a block address of a particular street. Next, the ***X*** and ***Y*** columns are converted to two significant places. Additionally, new columns are added representing the polar co-ordinates with angles at 30, 45 and 60 degrees. We use label encoder on ***PdDistrict*** column to represent the nominal data as numerical data. During training, we need to predict Category of the crime committed; however, ***Category*** of crime committed is represented in nominal data. So we convert them to numerical data using label encoder. We do not consider the columns Resolution, Descript and Address for our analysis.

The program takes about 5 to 10 minutes to complete execution and output of the program is written to the file ***sample\_crime\_submission\_20182095.csv.* Also the given training and test files should be located in the folder where the program script is being run**

**Sample Output:**



**Scoring on Kaggle Competition:**



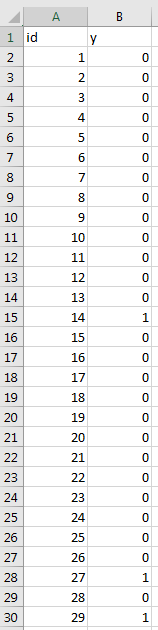
# Binary Classification On Bank Data (Client Subscription to Term Deposit)

**Command to execute:**

|  |
| --- |
| ***python bank\_classifier\_RF.py*** |

The above python program uses RandomForestClassifier to make predictions on the bank data determining if the client will subscribe to a term deposit or not. We use label encoder to transform data on the columns- Job, Marital, Education, Contact, Month and Poutcome. For columns- loan, housing and default that contain Boolean values yes or no, we use if/else conditions and transform them to value 0 or 1. We predict the goodness of the classification using f1\_score and output of the program is written to file ***sample\_bank\_submission\_RF.csv***. **Also the given training and test files should be located in the folder where the program script is being run**

**Sample Output:**



**Scoring on Kaggle Competition:**

