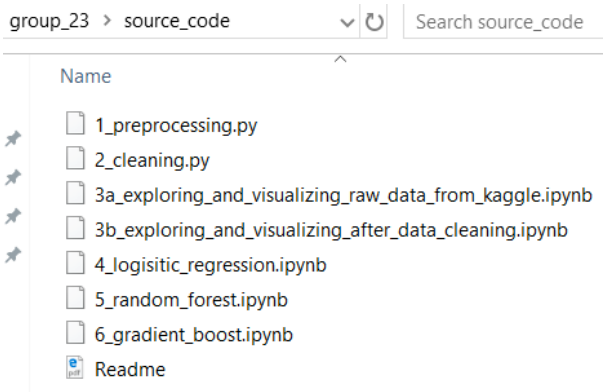
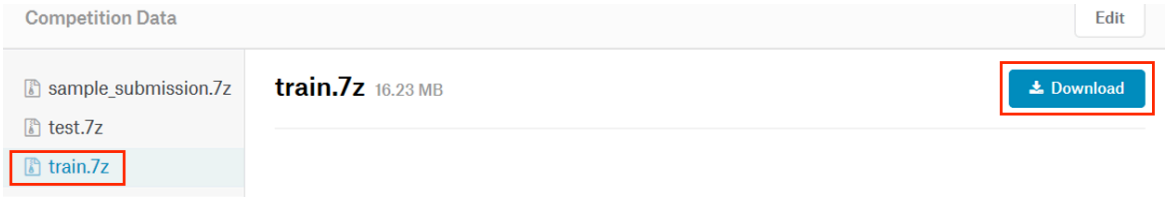
Source code structure:

All the code files needed to run the model is shown in the screenshot below:



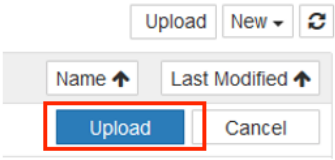
Steps to run the code:

1. Download data file "**train.csv**" from https://www.kaggle.com/c/porto-seguro-safe-driver-prediction/data. Select train.7z and then click Download button. Save the file in a directory of your choice, say "c:\ids"



2. Unzip this file and extract here to get a file named "**train.csv**". The path for our case should look like "c:\ids\train.csv"
3. Launch your Python IDE and then open file **1_preprocessing.py**.
    a. Search for "train = pd.read_csv" and replace the directory where your **train.csv** file exists ("c:\ids\train.csv" in our case)
    b. Search for "rows_0_1.to_csv" and replace the directory where you want "**train_u.csv**" file to exist (say "c:\ids\train_u.csv" in our case)
4. Run the file. After the code is run, you will find the file here: "c:\ids\train_u.csv". Close file **1_preprocessing.py**.
5. Open **2_cleaning.py**.
    a. Search for "train = pd.read_csv" and replace the directory where your "**train_u.csv**" file exists ("c:\ids\train_u.csv" in our case)
    b. Search for "train_final_df.to_csv" and replace the directory where you want "**train_final.csv**" file to exist (say "c:\ids\train_final.csv" in our case)
6. Run the file. After the code is run, you will find the file here: "c:\ids\train_final.csv". Close file **2_cleaning.py**.

7. Open your Jupyter Notebook. Click “Upload” on the top-right corner. Select file “**3a_exploring_and_visualizing_raw_data_from_kaggle.ipynb**” and open. Click Upload button again.



You will now find this file listed in your directory. Click on this file. Search for “train = pd.read_csv” and replace the directory where your **train.csv** file exists (“c:\ids\train.csv” in our case). Click on “Cell” on the menu bar and then click “Run All”. This will run the entire code set. You can browse through the outputs. This file details the feature drop analysis and missing data imputation.

8. For data visualization and analysis, upload file “**3b_exploring_and_visualizing_after_data_cleaning.ipynb**” in Jupyter Notebook as per #7 above. Click on this file. Search for “train = pd.read_csv” and replace the directory where your **train_final.csv** file exists (“c:\ids\train_final.csv” in our case). Click on “Cell” on the menu bar and then click “Run All”. This will run the entire code set. You can browse through the outputs. This file details visual data analysis.

9. For running Logistic Regression, upload file “**4_logisitic_regression.ipynb**” in Jupyter Notebook as per #7 above. Click on this file. Search for “train = pd.read_csv” and replace the directory where your **train_final.csv** file exists (“c:\ids\train_final.csv” in our case). Click on “Cell” on the menu bar and then click “Run All”. This will run the entire code set. You can browse through the outputs.

10. For running Random Forest Classifier, upload file “**5_random_forest.ipynb**” in Jupyter Notebook as per #7 above. Click on this file. Search for “train = pd.read_csv” and replace the directory where your **train_final.csv** file exists (“c:\ids\train_final.csv” in our case). Click on “Cell” on the menu bar and then click “Run All”. This will run the entire code set. You can browse through the outputs.

11. For running Gradient Boost Classifier, upload file “**6_gradient_boost.ipynb**” in Jupyter Notebook as per #7 above. Click on this file. Search for “train = pd.read_csv” and replace the directory where your **train_final.csv** file exists (“c:\ids\train_final.csv” in our case). Click on “Cell” on the menu bar and then click “Run All”. This will run the entire code set. You can browse through the outputs.