# TeleMatics

Dangerous driving predictions

This Document contains how to run and get Predictions of the TeleMatics Dataset.

In case you do not need a detailed instruction you can do the Installation Instructions and then directly jump into **How to get the predictions on test data section which is the last section.**

**Installation Instructions.**

1. Please clone the github repo on your local Machine.Below is the command

**git clone https://github.com/kartheek23/TeleMatics.git**

2. Please ensure you have python 3.6 or above in your machine. Ignore this step if python is already installed.

**https://www.python.org/downloads/**

3. Install Python Virtual Environment by the below command if you do not have it installed. Otherwise please ignore this step.

**pip install virtualenv**

4.To activate Virtual Environment for this project on Windows and Mac. Please key in the following commands at your local machine cloned repo from step 1.

**Windows.**

**test\_env\Scripts\activate**

**Mac.**

**source test\_env/bin/activate**

Further details can be found at **https://uoa-eresearch.github.io/eresearch-cookbook/recipe/2014/11/26/python-virtual-env/**

5. Once done with step 4, please install the required packages for the project by entering this command. This completes the setup of the project.

**pip install -r requirements.txt**

***About Project:***

The project has the following folders

**Root Folder:** Regression Model

**Sub Folders:** DataExploration,DataSets, config, processor and trained\_models

***Explanation of the sub folders and their python files***

1. **DataExploration:** Data Exploration folder has three python Notebooks.
2. **CombineDatasets.ipynb**: This file combines all the datasets in the features folders under

RegressionModel/DataSets/features and also undersamples the majority class and then writes a .csv file into MergedDataSet folder in Datasets directory. Please note that you have

To add the files in RegressionModel/DataSets/features folder in order to make it work as git is not allowing such huge files to be uploaded.

**Note: In order to make this file work make sure all the test data is copied into RegressionModel/DataSets/features folder.**

1. **DataExploration.ipynb**: This file primarily analyses if there are any missing data present

In the data,checking for outliers, class imbalance and other things that might be useful

to tackle as part of data preprocessing.

1. **FeatureImportance.ipynb**: This is where the model is built and the pickle file is built to give predictions. The output of this file is a pickle file which is under trained\_models

**with name model\_v1**.pkl

2.**Datasets folder contains the below folders.**

* 1. **MergedDataSet:** When the CombineDatasets.ipynb file is run then a file test.csv is generated in this folder.
  2. **Predictions**: Two files would fall into this folder which would be discussed in the below sections. The first file is a .csv file containing predictions and the other one is is a .csv file which contains the probabilities of the predictions.
  3. **TestFileUpload**: This is the folder in which you have to upload the .csv file for running the model.
  4. **features**: This folder where you have the initial datasets given to us for building the model, however I did not add them in Git because of its size. Once downloading the repo you can place them here.
  5. **labels**: The labels containing each of the test set’s target variable.

**3.config Folder**

a. **config.py:** All the important paths that the framework uses are written here.

**4.processor Folder**

a. **dataManagemnt.py:** This file contains functions to load the dataset, save the pipeline and load the pipeline which would help us in making the predictions.

**5. trained\_models**

a. **pipeline.py:** This file contains a file to build a model leveraging SKlearn pipeline. The pipeline model which is XGBoost has been run on cloud to reach optimal parameters as the data is large.

b. **model\_v1.pkl** is a file that has been generated to give the predictions. This file can be generated

by running train\_pipeline.py or FeatureImportance.ipynb files

c. **train\_pipeline.py** is a python file that helps us in building the model\_v1.pkl file on the test data.

d. **predict.py** is a file that gives us the predictions on the test data that is placed in Predictions folder of the datasets folder.

**How to get the predictions on test data.**

Please complete all the steps (1 to 5) in the Installation directory and once it is done, please get it into the command line/terminal and then cd into trained\_models directory and then execute the below command. Once the below command is executed you can see the predictions in the predictions folder in the DataSets directory. You can use these files to calculate metrics that you need like roc\_auc and other metrics. As I do not have the test data target data I cannot calculate the predictions.

**python predict.py**

**Please Ignore any warning messages once the above command is executed.**

**Note: Please make sure you put the .csv file on which you want to build the predictions in the TestFileUpload folder in DataSets folder in order to execute the above command.**