Training Report & LSTM research  
Date 5th April 2020 || Muhammad Farid

**Table of Content**

1. Training Objective
2. Training subject
3. Training Methods and Activities
4. Key findings and future recommendation on ML implementation
5. Algorithm used / Models / data-source
6. Sample Project
7. LSTM Forecasting

**1. Training objective**

Training objective is to learn and gain knowledge in python, libraries such as (Pandas,Numpy,scikit-learn and machine learning) and understand how to choose the correct charts / models to predict value based on data model

**2. Training subject**

* Basic and intermediate python
* Python Library ( Numpy , Pandas, Matplotlib )
* Visualization (matplotlib, Seaborn)

**3. Training method and activities**

* W3School
* Python Documentation
* DataQuest & Datacamp
* NICF ( Machine Learning ) -> Course attended during the weekends

**4. Training content**

* Python fundamental & Intermediate python
  + Variables, data types, list, dictionary, array, casting, loops, functions, OOP
* Basic Statistic calculation
  + Confusion matrix, precision matrix
* Python Library
  + Pandas , Numpy , Scipy , Scikit-learn, seaborn
* Performance Evaluation
  + PCA ( Principle Component Analysis

**IDE used**

* Anaconda (Jupyter notebook )
* Sublime text 3

**4. Key findings**

* Understanding python syntax
* We can use scikit-learn library to do prediction
* Basic matrix mathematics and confusion matrix to find F1 Score
* Finding the right fit model depending on datasets and output value
  + Model Algorithm and matrix
  + If output is of 2 options , e.g (Yes | no) , (True | false ) -> we can use classification algorithm
  + If output is knowing , will require clustering algorithm (Unsupervised)
* Data Manipulation with Numpy and pandas
* Converting raw data to dataframe with selected values
* Proper Steps of modelling datasets (1 -> Data Processing , 2 -> Build Model , 3-> Validation )

**5. Algorithm and scikit-Learn Applications**

* Classification
  + KNN & Logistic regression
* Regression
  + Linear Regression
  + Elastic Regression
* Clustering
  + KMean Clustering
  + Hierarchical Clustering
* Dimension Deduction
  + PCA (Principle Component Analysis)

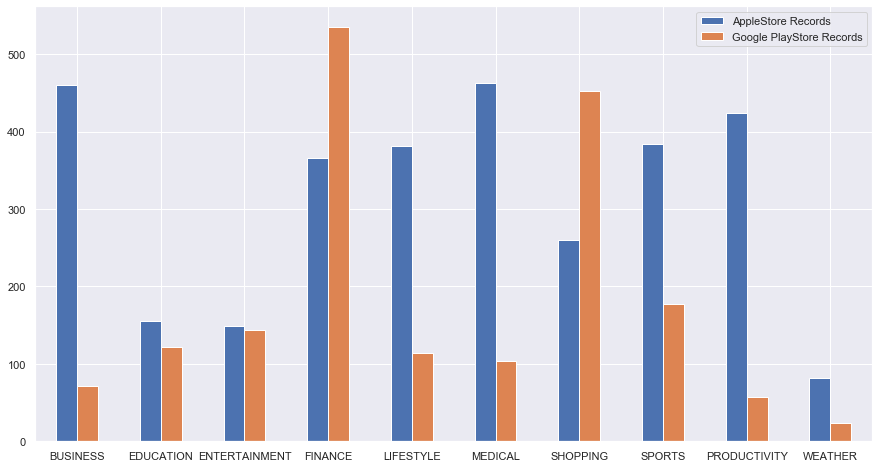
**6. Sample Project**

Project 1 | Number of application available in Google play and apple store :   
Technology and library used : Python , Pandas, Numpy , Scipy stats , Seaborn & Matplotlab

**Project 1 objective:**

1. To display the number of applications exist in apple store and google play.
2. Steps taken: Created 2 data frames for both Apple store and Google play
   1. Created Function to extract data (Import data via csv file)
   2. Created function to count frequency (Count the data that is within the category ( games , lifestyle , etc etc )
   3. Convert data to list using numpy to only get the common category in both datasets
   4. Convert datasets to pandas and plot bar chart for visualization

Output



I have uploaded my project folder on github for your reference: File( Project 1) <https://github.com/mdfarid/python-project/blob/master/project_1.ipynb>

**Project 2: LSTM Weather Forecast prediction – Method 1**

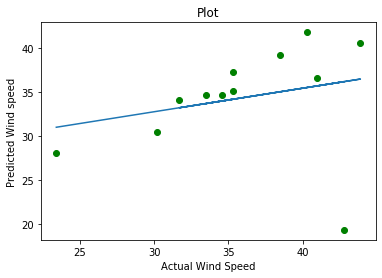
* Data is from extracted is from nea website for February
* Forecast created is for Changi Area, this can be change to other area based on the data that can be downloaded in nea website <http://www.weather.gov.sg/climate-historical-daily/>

**Datasets:**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Daily Rainfall Total (mm) | Highest 30 Min Rainfall (mm) | Highest 60 Min Rainfall (mm) | Highest 120 Min Rainfall (mm) | Mean Temperature (∞C) | Maximum Temperature (∞C) | Minimum Temperature (∞C) | Mean Wind Speed (km/h) | Max Wind Speed (km/h) |
| 0 | 0 | 0 | 0 | 28.3 | 32 | 26.2 | 17.6 | 38.5 |
| 0 | 0 | 0 | 0 | 27.9 | 31.8 | 26.6 | 16.9 | 40.3 |
| 0 | 0 | 0 | 0 | 28 | 31.5 | 25.6 | 13 | 34.6 |
| 0 | 0 | 0 | 0 | 27.5 | 31.5 | 24.6 | 11.2 | 31 |
| 0 | 0 | 0 | 0 | 27.8 | 32.3 | 24.5 | 12.2 | 33.8 |
| 0 | 0 | 0 | 0 | 28.1 | 32.2 | 25.6 | 14.4 | 35.3 |
| 0 | 0 | 0 | 0 | 28 | 31.3 | 25.9 | 14.4 | 41 |
| 0 | 0 | 0 | 0 | 27.7 | 31.4 | 24.7 | 9.7 | 30.2 |

Steps taken to do the prediction

* Processing data (data extraction) via csv
* Build model -> Linear Regression
  + Model is trained at 30% (Using train\_test\_split) with scikit library
* Prediction

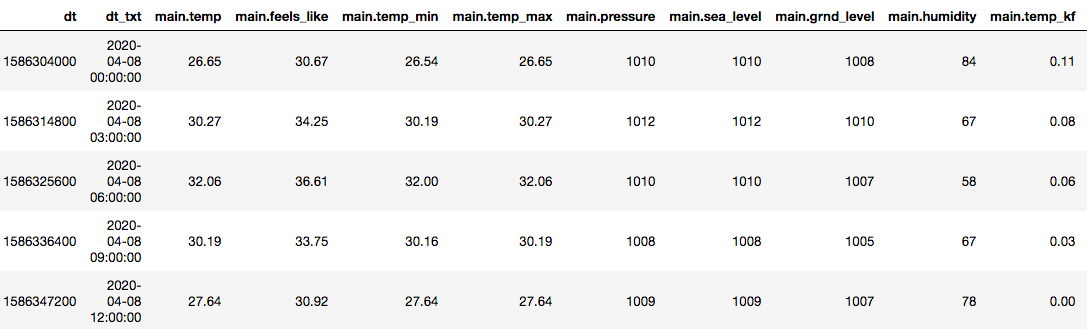
Data is randomized

|  |  |
| --- | --- |
| Actual Data | Predicted |
| 40.3 | 28.1 |
| 34.6 | 30.48 |
| 35.3 | 37.2 |
| 41.0 | 34.7 |
| 35.3 | 19.3 |
| 31.7 | 35.13 |
| 23.4 | 34.1 |

Project file: <https://github.com/mdfarid/python-project/blob/master/Lstm_neuron.ipynb>

**Project 2: LSTM Weather Forecast prediction – Method 2**

Objective : To explore with Tensorflow library to get predicted data

Resource : www.openweathermap.org  
Sample datasets

Project status : In progress

**Remarks**

Challenges faced on LSTM –

* Data might not be accurate and it still need to be tune
* Implemented the 3 steps to get data -> data processing , modeling , validation(prediction )
* Still unsure if it is the right model
* Data was not available in csv format for the past 10 days, data is only available in csv for the month before

Remarks – Development in progress

* can use RNN and Tensorflow based on my research to do lstm weather forecasting