Term Project

Due: 29/12/2022 – 23:30

Presentation: 30/12/2022

Instructions:

You should make an end to end analytics project that includes the following steps.

You will use Covid-19 dataset which is a collection of the COVID-19 data maintained by *Our World in Data*. It is updated daily and includes data on confirmed cases, deaths and testing [1].

Purpose of this project is to predict the next day's new COVID-19 cases.

Firstly, you should setup MongoDB and PyCharm to your computer. You can use the instructions given to you for setup. You will use PyCharm to make ETL from dataset URL to MongoDB. You are expected to fill the ETL functions that are provided to you. (You should install pyspark and pymongo libraries with pip install commands)

Secondly, you should setup Anaconda to your computer. Anaconda will enable you to use Jupyter Notebook for analysis and modeling purposes. You are expected to perform:

1- Explanatory data analysis
2- Feature extractions
3- Modeling trials
4- Visualizations

in Jupyter Notebook with a bulk data that you read from MongoDB. You should save your final model to a pickle object for reproducibility. You can use Modeling Template provided to you.

After, you are expected to create a modeling pipeline for the best model you choose in PyCharm. Modeling pipeline ensures that the sequence of data preparation operations performed is reproducible. You should write a script that reads data from MongoDB, perform data preparation steps for your chosen model and scores the model data by using model pickle you saved.

Finally, when the whole scoring pipeline is ready, you should trigger the process daily to score daily data extracted from database URL. Write your predictions to MongoDB. You may test your prediction results next day with real data.

Modeling Task:

The problem given to you is a regression problem with a time series data which means that the target value to be estimated is autoregressive. You may generate combination of past values of the target variable (eg. average new cases in the last week etc.)

According to your analysis, clearly state your approach to this problem.

Use data until 01/08/2022 for training purpose. You may test your model performances from 01/04/2022 to 01/11/2022.

Look at R2, Mean Absolute Percentage Error (MAPE) and Mean Absolute Error (MAE) for model performances. Check error distributions by using box plot.

Finally, forecast the new cases in Turkey by using your model for the first week of November 2022 (1-7th of November) and write your predictions to MongoDB.

Notes will be given to jupyter notebook, presentation of results and final scripts written in pycharm.

Task	**Points**
ETL to MongoDB	25
EDA - Visualization	10
Feature Extraction	15
Modeling	15
Final scoring	20
Presentation	15

Reference

[1] Hannah Ritchie – "Coronavirus Source Data". Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/coronavirus-source-data' [Online Resource]