MIE 1624 Introduction to Data Science and Analytics – Fall 2020

Final Exam Project

Deadline: Tuesday, December 15th, 11:59 PM

Background

Coronavirus disease (COVID-19) is an infectious disease caused by a newly discovered coronavirus. Most people infected with the COVID-19 virus experience mild to moderate respiratory illness and recover without requiring special treatment. Older people, and those with underlying medical problems like cardiovascular disease, diabetes, chronic respiratory disease, and cancer are more likely to develop serious illness. The COVID-19 virus spreads primarily through droplets of saliva or discharge from the nose when an infected person coughs or sneezes. In this final exam project we will model spread of the coronavirus and study how we can project and plan for best and worst case scenarios as we are in the midst of the 2nd wave.

The dataset we will be using to accomplish this is the COVID-19 Data Repository from the Center for Systems Science and Engineering (CSSE) at John Hoskins University. This dataset has daily level information on the number of affected cases, deaths and recovery from 2019 novel coronavirus.

The goal of this project is to develop data science or machine learning models that forecast COVID-19 cases with the aim of helping policymakers plan for the days and months ahead, and take action to change the course of the pandemic for the better. In addition to modelling and projecting COVID-19 cases you will also use a secondary dataset, of your choice, to understand how case numbers correlate to that dataset. Just one example of this is that we could look at the correlation of case numbers to the issuance and compliance with a universal mask mandate to explore the effectiveness of masks. As an output of this comparison you will be required to outline some evidence-based insights, policies and guidelines that would help minimize the case numbers and related adverse effects.

Learning Objectives

1. Implement a data pre-processing pipeline to clean and wrangle time series data in order to prepare it for time series modelling.
2. Train and test time series data science or machine learning algorithms in order to model COVID-19 case projections under multiple scenarios.
3. Understand and utilize data visualization and exploratory data analysis techniques to understand how the time series of COVID-19 cases behaves and evolves.
4. Improve on skills and competencies required to collate and present domain specific, evidence-based insights. Particularly, in this case to gain insights and guide the fight against the COVID-19 pandemic.

To do:

1. **Data Cleaning – [5 Marks]**

   The COVID-19 CSSE John Hopkins dataset is provided to you as a url to .csv file, you will need to extract the case counts and clean the data so that you have a time series of COVID-19 cases from your chosen location. Here it would be prudent to choose a location that has detailed data available, for example a province or municipality in Canada or a state in the USA (*Read Part 4 for more on this). Depending on the algorithm you choose to use to model and find projections of the cases you will require a different format of data or approach and thus you will design your data cleaning pipeline accordingly. Please note that this is a time series data and so the number of cases on any given day is the cumulative number.

2. **Data Visualization and Exploratory Data Analysis – [10 Marks]**

   Depending on how you want to conduct your analysis of COVID-19 cases present 3 graphical figures that visualize aspects or information in the data that you will further explore with your models. How could these trends be used to help with the task of methodically extracting all information and trends of this type? Consider how accessing the data and creating these visualizations will inform how the data will need to be pre-processed and fed into your models. All graphs should be readable and presented in the notebook. All axes must be appropriately labeled.

3. **Model selection and fitting to data – [35 Marks]**

   Select a model of your choice (for example you may select an SIR Model, ARIMA, optimization or MC simulation modelling or any of the other models we have covered) that will allow you to project the time series of COVID-19 Cases into the future. You should output three projections, one that assumes worst case spread, another that assumes best case spread and a third that models a base-case in between best and worst spread. You must justify your algorithm choices and the approach you will use to generate the 3 cases of projections. You may also choose to study multiple models and report on the suitability of each in developing projections for COVID-19. You should also use the dataset provided to fit/train model(s) selected and discuss and interpret the findings of these model(s). You may also use this section to improve the model depending on the findings of your model(s) and how you interpret them.

4. **Relating COVID-19 Projections to a Second Dataset [25 Marks]**

   Select another dataset of your choice (You can look through and choose from the many available here or you can use any other dataset you may be able to find). In this part of the project you will use your 2nd chosen dataset to examine and analyze a factor that is related to COVID-19. Just as an example this factor could be looking at how and when mask mandates and mask-wearing policies, in the location you chose to analyze projections in Part 3, affected COVID-19 case counts. Other factors you could choose to

correlate the number of cases to include social distancing metrics, economic impacts, hard and soft lockdowns, local/global travel, herd immunity, contact tracing, hospital capacity just to name a few. The possibilities here are only limited by what you can find data around for the geographic location you chose in Parts 1-3. Thus, it is suggested that you choose a location that has multiple datasets available for it.

5. **Deriving insights about policy and guidance to tackle the outbreak based on model findings – [25 Marks]**

   Using the findings from your models in Part 3 and 4 on the coronavirus you are now tasked with discussing and proposing how scientists, doctors, nurses, healthcare professionals, industry and governments can best use the insights from your analysis to assist in the fight against the COVID-19 pandemic. Use evidence-based insights derived about the disease from your model(s) and your data analysis to justify proposed policies or action items.

The order laid out here does not need to be strictly followed. A significant number of marks in each section are allocated to discussion. Use markdown cells in Jupyter notebook as needed to explain your reasoning for the steps that you take.

Programming Tools:

- Software
  - Python version 3.X is required for this project. Python version 2.7 is not allowed.
  - Your code should run on the Google Colab Virtual Environment or CognitiveClass Virtual Lab (Python 3 kernel).
  - All Python libraries and built-ins are allowed but here is a list of the major libraries you might consider: numpy, Scipy, Scikit, Matplotlib, Pandas, NLTK.
  - No other tool or software besides Python and its component libraries can be used to touch the data files. For instance, using Microsoft Excel to clean the data is not allowed.
- Required data files
  - **time_series_covid19_confirmed_US.csv / time_series_covid19_confirmed_global.csv**: These .csv files contain case counts from the US and globally. Notice that for within the US and some other countries the case counts are further broken down into sub-regions. You are free to choose to analyze an entire country, a province/state or a municipality/region. Please ensure that you choose a place that has other ample datasets so that you can conduct the analysis for Part 4.
  - The data file cannot be altered by any means. The IPython Notebooks will be run using a local version of this data file. Please restrict the last date in the timeseries to be December 10th and delete any data from days after this date.

What to Submit:

1. Submit via Quercus portal an IPython notebook containing your implementation and motivation for all the steps of the analysis with the following naming convention:

   lastname_studentnumber_finalproject.ipynb

   Comment out any data retrieval processes (e.g. downloading your own additional data, etc.) in your code and replace it with code for reading the corresponding data from files. (**Submit all those data files together with your Jupyter notebook**).

   Make sure that you **comment** your code appropriately and describe each step in sufficient detail. Respect the above convention when naming your file, making sure that all letters are lowercase and underscores are used as shown. A program that cannot be evaluated because it varies from specifications will receive zero marks.

2. Submit 5 slides in PowerPoint and PDF formats describing your findings from exploratory analysis, model feature importance, model results and visualizations. Use the following naming conventions **lastname_studentnumber_finalproject.pptx** and **lastname_studentnumber_finalproject.pdf**

3. Submit a video in MP4 describing the findings in your code and report. Use the following naming conventions lastname_studentnumber_assignment1.mp4. Make sure your video is no longer than 3-minute – if it is, it may not be graded.

Late Submissions:

- up to 2 hours late - no penalty
- one day late - 15% penalty
- two days late - 30% penalty
- more than two days late - 0 mark

Tips:

1. You have a lot of freedom with however you want to approach each step and which library or function you want to use. As open-ended as the problem seems, the emphasis of the project is for you to be able to explain the reasoning behind every step.
2. While some suggestions have been made in certain steps to give you some direction, you are not required to follow them. Doing them, however, guarantees full marks if implemented and explained correctly.