MIE 1624 Introduction to Data Science and Analytics – Fall 2021

Final Exam Project

Deadline: Wednesday, December 15^{th}, 11:59 PM

Background

The COVID-19 pandemic is the most serious public health crisis to have engulfed the world in the last two years. Forecasting the COVID-19 vaccination trend has become difficult. Numerous health professionals, statisticians, researchers, and programmers have been tracking the virus's spread in various parts of the world using a variety of methods. The proliferation of vaccines developed by talented scientists piqued our interest in learning more about ongoing vaccine programs, and our passion for deriving meaningful insights from data drew us to this particular endeavour. For the last year, we have been hoping for vaccines that would allow us to resume our normal lives. In this final exam project, we will examine the global vaccination drive.

One dataset that you can use for this analysis is the Our World in Data Covid-19 Vaccination dataset at the University of Oxford. This dataset contains daily data on the number of people who have been vaccinated or are fully vaccinated in 218 countries.

The objective of this project is to forecast vaccination rates using data science or machine learning algorithms in order to determine the impact of vaccination on our daily lives. Along with modelling and projecting COVID-19 immunization rates, you will examine correlations between vaccination and a secondary dataset of your choosing. For example, you can investigate the correlation between vaccination rates and the number of new COVID-19 cases or the number of hospitalizations to ascertain how vaccination effects on them. The project's overarching goal is to derive insights from the COVID-19 vaccination dataset in order to inform how our healthcare system, government, and industry can tackle this growing problem through increased immunization.

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. Understand and utilize data visualization and exploratory data analysis techniques to understand how the time series of COVID-19 Vaccination rates behaves and evolves.
3. Train and test time series data science or machine learning algorithms in order to model COVID-19 vaccination rates projections under multiple scenarios and gain insights or answer the overarching question you have chosen to pursue for this project.
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 via vaccination.

To do:

1. **Data Cleaning – [1 mark]**

   You are provided with a link to a .csv file containing the COVID-19 (coronavirus) vaccinations dataset. In this section, you need to clean up the dataset by dealing with missing data and filling them with appropriate values (not necessarily zero) or dropping the unnecessary data. You may need to modify or create your own data cleaning pipeline, depending on the algorithm you use to model and predict vaccination rates. Please keep in mind that this is time series data, so the number of vaccinations on any given day should be the total number of vaccinations. (Hint: to fill in the missing values, examine the correlation matrix between the features, perform statistical tests to determine the similarity between their distributions, and finally decide whether to fill them with zeros or other appropriate values.)

2. **Data Visualization and Exploratory Data Analysis – [4 marks]**

   Depending on how you wish to conduct your analysis of COVID-19 vaccination rates, present four graphical figures that illustrate various aspects or information contained in the data that will be explored further with your models. How might these trends be used to aid in the task of methodically extracting all relevant data and trends? Consider how accessing the data and creating these visualizations will inform the preprocessing and feeding of the data into your models. All graphs should be legible and presented in a readable format in the notebook. All axes must be labelled appropriately. Additionally, for data visualizations, if necessary, conduct exploratory data analysis in other forms.

3. **Model selection and fitting to data – [9 marks]**

   Select a model of your choice (you may select an ARIMA, SIR Model, optimization or Monte Carlo simulation modelling or any of the other models we have covered) that will allow you to project the time series of COVID-19 Vaccination rates into the future. This analysis should be conducted for a minimum of two countries (one of them must be Canada; select others based on your choice). You should generate three projections for each of the countries: one that assumes the worst-case scenario, another that assumes the best-case scenario, and a third that models a base-case scenario in between the best and worst-case scenarios. You must justify your algorithm selections and the approach you intend to take in generating the three projection cases. According to the forecasting you have done, how many people will be vaccinated in the next 50 days in each country? Additionally, you may choose to examine multiple models and report on their suitability for answering your overarching question about COVID-19 vaccination. You should also use the dataset provided to train the models you selected and discuss and interpret the findings of these models. Depending on the findings of your models and how you interpret them, you may also use this section to improve the model.

4. Relating COVID-19 Vaccination to a Second Dataset – [5 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 section of the project, you will examine and analyze a factor associated with COVID-19 vaccination rates using your second chosen dataset. For instance, this factor could be the number of new COVID-19 cases or hospitalizations in the locations you selected for Section 3 analysis (i.e., Canada and the others). Additionally, you can correlate vaccination rates with social distancing metrics, economic impacts, hard and soft lockdowns, local/global travel, herd immunity, contact tracing, and hospital capacity, to name a few. The possibilities here are limited only by the data that is available for the geographic locations you selected in Parts 1-3. As a result, it is recommended that you choose a secondary (or more, optional) location that has access to multiple datasets.

5. Deriving insights about the effect of vaccination and discussion – [5 marks]

Using the findings from your models in Sections 3 and 4 on the coronavirus vaccination rates, you are now tasked with discussing the effect of vaccination on our daily lives. Which of your chosen countries has the most effective vaccination program? From what aspects? Why? What discoveries have you made as a result of the dataset and your models? Use evidence-based insights derived about the disease from your model(s) and your data analysis to justify your findings.

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 (Kernel 3).
- 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. 22
- 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

- Covid19_vaccinations.csv / Complete_covid19_dataset.csv: These .csv files contain any information you need for COVID-19 including vaccination rates, confirmed cases and deaths, hospitalization and ICU admissions, and other variable of interest. You are free to

choose to analyse any factor, any country of your interest. 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 local version of this data file.

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 if available, 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_finalproject.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 - 20% penalty
- more than one day 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.