TOPIC – COVID 19 Prediction

Business Problem:

Coronavirus disease 2019 is a contagious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The disease has since spread worldwide, leading to an ongoing pandemic. Symptoms of COVID-19 are variable, but often include fever, cough, headache, fatigue, breathing difficulties, and loss of smell and taste. Symptoms may begin one to fourteen days after exposure to the virus.

Background / History:

COVID-19 transmits when people breathe in air contaminated by droplets and small airborne particles containing the virus. The risk of breathing these in is highest when people are near, but they can be inhaled over longer distances, particularly indoors. Transmission can also occur if splashed or sprayed with contaminated fluids in the eyes, nose, or mouth, and, rarely, via contaminated surfaces. People remain contagious for up to 20 days and can spread the virus even if they do not develop symptoms. This project proposal mainly aims at exploring COVID-19 through data analysis and projections.

Data Explanation:

The datasets are fetched from several websites which are mentioned as follows:

1. John Hopins University - CSSEGISandData/COVID-19: Novel Coronavirus (COVID-19) Cases, provided by JHU CSSE (github.com).
2. Apple mobility - COVID-19 - Mobility Trends Reports - Apple
3. World Health Organization - Coronavirus disease (COVID-19) (who.int)
4. CDC - Coronavirus Disease 2019 (COVID-19) | CDC

After combining all data from various data sources, the dataset will have the following columns:

1. State – State name.
2. Country – Country name.
3. Last Update – Last updated date of the data.
4. Latitude – Latitude of the state.
5. Longitude – Longitude of the state.
6. Confirmed cases – Number cases confirmed at present.
7. Deaths – Number of deaths happened so far.
8. Recovered – Number of cases recovered.
9. Active cases – Number of present active cases.

Methods:

This project concentrates more on visualization. So, I'm planning to use the seaborn library for all the visualization. Seaborn is a library for making statistical graphics in Python. It builds on top of matplotlib and integrates closely with pandas data structures. Seaborn helps you explore and understand your data. Its plotting functions operate on data frames and arrays containing whole datasets and internally perform the necessary semantic mapping and statistical aggregation to produce informative plots. Its dataset-oriented, declarative API lets you focus on what the different elements of your plots mean, rather than on the details of how to draw them.

For the prediction, we can include Support Vector Machine, Polynomial Regression, and Bayesian Ridge regression. Support vector machines are a set of supervised learning methods used for classification, regression, and outliers detection. Polynomial regression, like linear regression, uses the relationship between the variables x and y to find the best way to draw a line through the data points.

Analysis:

The Covid-19 pandemic is the most important health disaster that has surrounded the world for the past two years. The dataset consists of weekly confirmed cases and weekly cumulative confirmed cases obtained from World Health Organization and other organizations. Then the distribution of the data was examined using the most up-to-date Covid-19 weekly case data and its parameters were obtained according to the statistical distributions.

Conclusion:

As of now, to conclude we haven’t changed any of the methods for the prediction. We will split the training dataset into train and test sets and we will use the train set to fit the model and generate a prediction for each element on the test set. Finally, we will track all observations in a list called history that is seeded with the training data and to which new observations are appended at each iteration.

Assumption:

Here we are going to implement various methods to build the model. But again, we are not sure that any new mutation in the virus can change drastically the situation which will make the predictions wrong. But, for this analysis, we are assuming that no more mutation will be there in the COVID virus. Apart from this assumption, there is no assumption made and the values which we are going to implement in this project are real-time values only.

Limitation:

As mentioned earlier, we don’t know the exact accuracy until we complete the coding part of this project. And there are some other factors like virus mutation, government actions like curfews, lockdown, travel ban, etc will affect the real-time data and the predictions which is not considered here. So, probably the accuracy of the results may be less in this approach which can’t be mentioned now.

Challenges:

With the resurgence of machine learning and artificial intelligence, never has it been easier to implement predictive algorithms both new and old. With just a few lines of code, state-of-the-art models can be readily accessible at the fingertips of the budding data enthusiast, ready to conquer whatever insurmountable digital task may lay at hand. But a little bit of knowledge can be a dangerous thing. While much of machine learning can be attributed to statistics and programming what is equally important, but often skipped over in favor of instant gratification, is domain knowledge. There are a few Challenges in using the Bayesian Regression algorithm which is as follows:

1. The inference of the model can be time-consuming.
2. If there is a large amount of data available for our dataset, the Bayesian approach is not worth it and the regular frequentist approach does a more efficient job.

Future Uses:

COVID prediction aims to determine the number of persons affected in each country in the future. The accurate prediction of this scenario will lead to making decisions like lockdown, travel ban for each country to avoid the pandemic further worst. If the predictions came well with accuracy, we can implement it as a mobile app with a good User Interface for public use.

Recommendations:

As per websites like Kaggle and other data science websites, the recommended model for this project is Support Vector Machine, Polynomial Regression, and Bayesian Ridge regression. SVM is a supervised machine learning algorithm that can be used for both classification and regression challenges. In statistics, polynomial regression is a form of regression analysis in which the relationship between the independent variable x and the dependent variable y is modeled as an nth degree polynomial in x.

Implementation Plan:

We can use data from the above-mentioned sites to extract the data for the analysis. Then will create lots of visualizations to show the cases for each country. Then we will implement Support Vector Machine, Polynomial Regression, and Bayesian Ridge regression for the prediction. Based on the model accuracy we will conclude which model will be suited for predicting the COVID -19 cases and we can publish the results somewhere on the server.

Ethical Assessment:

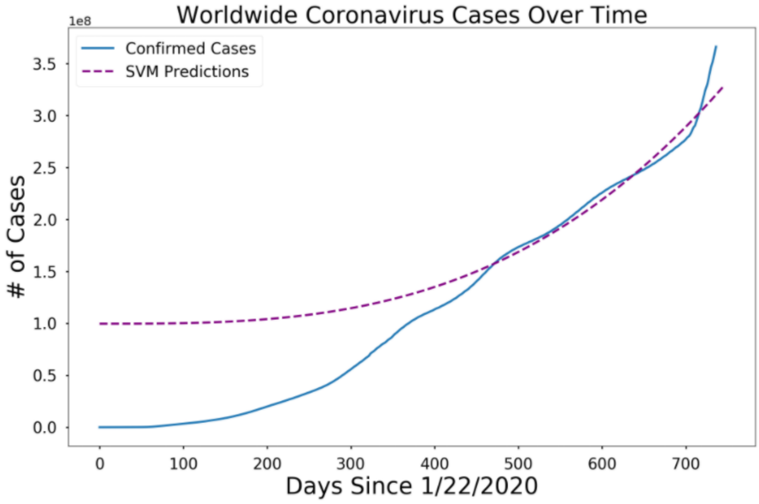
In this project, we are going to use the data which is available for public use from websites like apple mobility, John Hoppins University, World Health Organization, etc. So, there are no ethical issues in handling data. But the ethical issue might raise when we release the results of this project because it's experimental and we don’t know exactly how the model will behave for future predictions. So, there are some potential threats that the result may mislead the people and government of each country in the world. Thus, to avoid this kind of issue, we need to test the model thoroughly with all possible parameters.

Reference:

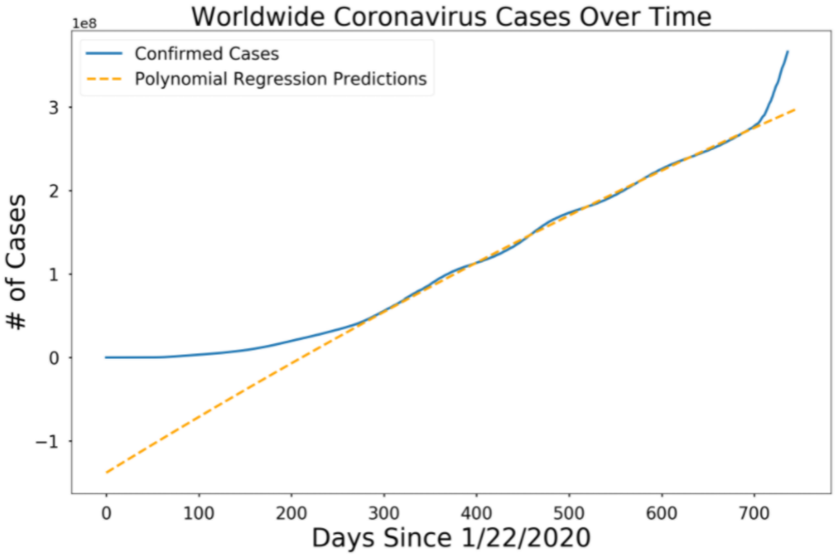
1. COVID-19 - Wikipedia
2. Coronavirus (COVID-19) Visualization & Prediction | Kaggle

Illustrations:

1. **SVM Model prediction:**

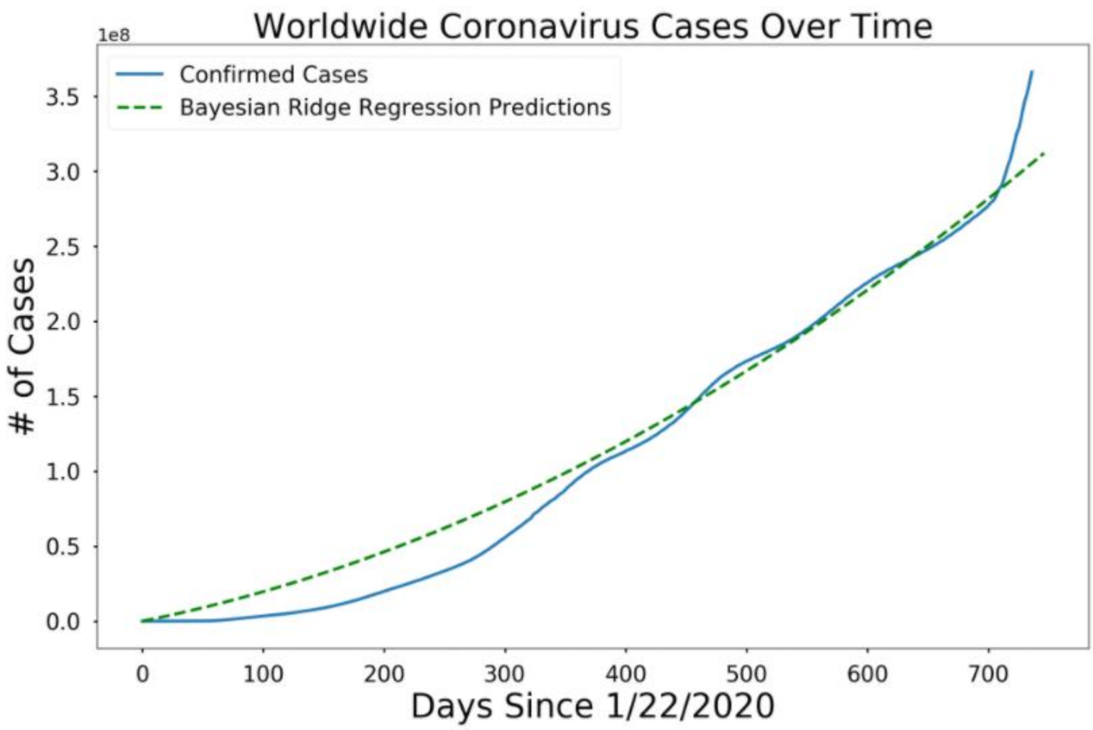


2. **Polynomial Regression Prediction:**

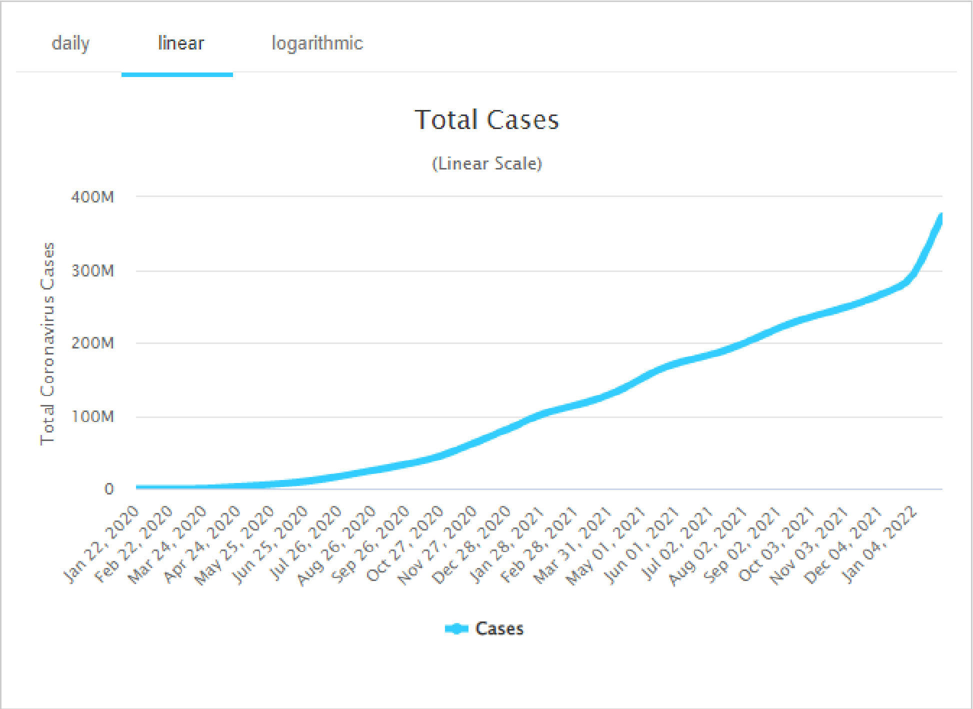


3. Bayesian Ridge Regression Prediction:

Show hidden code



4. Actual Data:



Appendix:

1. Bayesian Ridge Regression Prediction – In statistics, Bayesian linear regression is an approach to linear regression in which the statistical analysis is undertaken within the context of Bayesian inference. When the regression model has errors that have a normal distribution, and if a particular form of the prior distribution is assumed, explicit results are available for the posterior probability distributions of the model's parameters.
2. SVM – In machine learning, support vector machines are supervised learning models with associated learning algorithms that analyze data for classification and regression analysis. Developed at AT&T Bell Laboratories by Vladimir Vapnik with colleagues, SVMs are one of the most robust prediction methods, being based on statistical learning frameworks or VC theory proposed by Vapnik and Chervonenkis. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. SVM maps training examples to points in space to maximize the width of the gap between the two categories. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.
3. Polynomial Regression – In statistics, polynomial regression is a form of regression analysis in which the relationship between the independent variable x and the dependent variable y is modeled as an nth degree polynomial in x. Polynomial regression fits a nonlinear relationship between the value of x and the corresponding conditional mean of y, denoted E(y |x). Although polynomial regression fits a nonlinear model to the data, as a statistical estimation problem it is linear, in the sense that the regression function E(y | x) is linear in the unknown parameters that are estimated from the data. For this reason, polynomial regression is considered to be a special case of multiple linear regression.
4. COVID - Coronavirus disease 2019 is a contagious disease caused by severe acute respiratory syndrome coronavirus 2. Symptoms of COVID-19 are variable, but often include fever, cough, headache, fatigue, breathing difficulties, loss of smell, and loss of taste. Symptoms may begin one to fourteen days after exposure to the virus. At least a third of people who are infected do not develop noticeable symptoms.
5. Seaborn – Seaborn is a library that uses Matplotlib underneath to plot graphs. It will be used to visualize random distributions.
6. Normal Distribution – The Normal Distribution is one of the most important distributions. It is also called the Gaussian Distribution after the German mathematician Carl Friedrich Gauss. It fits the probability distribution of many events, eg. IQ Scores, Heartbeat, etc.
7. Linear regression - In statistics, linear regression is a linear approach for modeling the relationship between a scalar response and one or more explanatory variables. The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression. This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.
8. Kaggle – Kaggle, a subsidiary of Google LLC, is an online community of data scientists and machine learning practitioners. Kaggle allows users to find and publish data sets, explore and build models in a web-based data-science environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges.

Questions:

1. What kind of visualizations are required for this project and what's the plan to implement it?
2. What's the reason for using Bayesian Ridge Regression Prediction?
3. What’s the reason for using Polynomial Regression?
4. How to integrate it with other applications like the web, android, iOS, and other platforms?
5. Is there any particular reason for using seaborn for visualization?
6. Is there any other models suit for this prediction?
7. What is SVM?
8. Do you have any plan to include future mutations of the COVID virus to predict the model?
9. How accurate is the model?
10. What are the assumptions made for this project?