Project Proposal

Abstract:

General description and justification for your project. What are you trying to do?

Design and build a streaming application that provides a model, predictions, and visualization of COVID19 virus data in a Shiny dashboard. The objective of our dashboard is implementing a predictive model to identify which factors have the greatest impact on the spread and strength of the COVID-19 virus in a specific geographic area.

Project objective:

- Build a functional dashboard in Shiny presenting COVID19 virus predictions.
- Stream COVID data using Kafka or Spark Streaming.
- Visualization of predictive analytics model and data.
- Predictions or time series analysis based on COVID19, environmental, and economic factors.

Current Methods:

How is it done today, and what are the limits of current practice? ***This may be answered from other sources and literature in the domain of study.***

Today, the most widely used predictive modeling techniques are decision trees, regression and neural networks. Regression (linear and logistic) is one of the most popular methods for statistical analysis. Another method used today is regression analysis which estimates relationships among variables. With models of the number of confirmed cases beginning to be released to the public, analysis on this data has become a top priority for Universities as well as Government Agencies. The limitations of the models available don't quantify the relationship between demographics in a populous (Other than identifying risk factors on an individual basis) and government actions on the extent and severity of the virus.

Our Approach:

What's new in your approach and why do you think it will be successful?

Most models relate only to the number of cases per geographic location or the number of hospitalization and deaths in a given area. There are also only general descriptive statistics on the risk factors related to the virus. This project would attempt to quantify the effect of government action, demographic data, healthcare resources and the population's compliance with government orders on the rate of spread and severity of health impact. Rather than simply stating risk factors as descriptive statistics, this analysis

would attempt to provide predictive models based on the risk factors identified, as they appear in a given population.

Impact:

Who cares? If you're successful, what difference will it make?

Governments, healthcare workers, the general public, in short everyone has a stake in this information. The more we understand about this virus, the better we can be prepared if and when it comes back. With COVID19 and Coronavirus dominating the news and media, streaming virus data will provide a predictive analytics tool for tracking the impact of the pandemic. Creating awareness of the spread of the virus could help us better understand its impact. These results could help save lives.

Risks:

What are the risks and the payoffs?

The primary risks include incorrectly analyzing the data and providing misleading results. Incorrect conclusions could lead to damaging actions. Another risk is if states inaccurately identify confirmed cases or understate the total number of deaths. A payoff is correctly identifying the results that lead to taking actions that mitigate the adverse effects of the virus. If we can identify which actions lead to increased mitigation, then results could provide more hope and encouragement than fear.

Costs:

How much will it cost?

The data is publicly available and the tools being used in the analysis are open source including: Spark, R, Python, etc.

Checkpoints:

What is the midterm and final "exams" to check for success?

- **Midpoint Checkpoint**: Feature and model selection, tuning, and validation.
- **Final Exam Checkpoint**: Dashboard functionality includes: visualization, model results, and streaming data.

Initial data gathering and analysis to determine a statistically significant relationship will take a week. Attempts to most accurately model the relationship at a more granular level will take a week. Midterm determination would involve identifying a significant relationship and final examination would be determined by quantifying and modeling that given relationship with more granular detail. The Final Exam Checkpoint will focus primarily on testing and application deployment.

Project Schedule:

How long will it take?

Our project schedule is set for a 4 week deadline where we will implement a shortened agile process of 2, 1 week iterations, followed by 1 week of testing, and the final week is dashboard deployment. Our statistical analysis will rely on if we are able to reject a null hypothesis for a given factor will determine which factors, if any, have a statistically significant relationship with the effect of the virus in a specific geographic area.

Team:

Ishaan Khurana – **Product Manager**

Jacqueline Connolly – **Pipeline**

Kyle W. Brown – **Visualization/Dashboard**

Derick Karolak – **Machine Learning**

Nischitha Manjunanth – **Storage/Research**

Data Source:

Link to public API's:
https://rapidapi.com/collection/coronavirus-covid-19

Or

Link to public description of data source:
https://github.com/CSSEGISandData/COVID-19

Link to public of data source:
https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_daily_reports