GBOLAHAN AKEEM AFUWAPE

FORECASTING HOPSITAL BED SPACES, VACCINATIONS TURNOUTS AND OTHER HEALTH FEATURES

BACKGROUND

According to the WHO, the world has recorded over 160 million cases of COVID-19 infections from 2020 till date. One of the greatest concerns across the world is the potential impact of surging infections on their health care systems. Governments went through the motion of effecting lockdowns, shutdowns, curfews etc. to help plan and mitigate the risk of a healthcare crisis. Also, vaccine wastage due to inadequate logistics is fast becoming a source of concern as public health units as they are to predict turnouts for vaccine doses.

This project aims to help governments and health boards manage the crisis by providing weekly data-driven insights to guide decisions or actions based on past patterns. Based on past data, forecasts might assist public health units can better plan for vaccination by using forecasts for the turnout, mitigating vaccine wastage. We approach this work considering the daily numbers aggregated by health organizations on various features for covid cases encountered. The project is handled as a Multivariate time series, using models that lend themselves to the successful forecasting of future trends by understanding the patterns in past data.

DATA ACQUISTION, QUALITY AND COMPLETENESS

The data used in this project were aggregated from Ontario's 34 Public Health Units through API connection, information on demographics of Ontario was sourced from Statistics Canada, and the daily vaccination turnouts were web-scrapped using Selenium.

The granularity of the data was a limitation as some data were aggregated at a Region level while others are at provincial level. All data used in the project are updated daily so we had a cutoff point of 1st of May 2021. Another challenge noted is the "labelling" of the various Public Health Units, this different for different data obtained from the same API. We solved this by creating a "working key" for all health units using the government assigned PHU ID.

Datasets used in this work include but are not limited to :

- Ontario covid cases by Public Health Unit:
- Daily vaccination doses administered and individuals per Public Health Units
- Daily aggregated regional hospitalization numbers, detailing the number of persons in intensive care units, and in need of ventilators for Ontario's Health Region. (Ontario's 34 Public health units are grouped into 5 Health regions)

As a summary, there are over 500, 000 covid infection cases in Ontario, with 18 % of those cases linked to outbreaks and 26% of those cases occurred in Toronto Health region. Most cases occurred with persons in the age group between 20 to 40 years old, with the highest mortality rate amongst the elderlies with a mortality rate of 25 % for age group in 90 and above.

MODELS

The interdependencies of the features in the aggregated data necessitated the use of Vector Autoregression (VAR) models and Recurrent Neural Networks (RNN). VAR models are dependent on the lag parameter and stationarity of the features. RNN models included stacked Long-Short Term Memory (LSTM) architectures, and hybrids containing layers of Convolutional Neural Networks. While CNNs are used working with spatial data, the inclusion of 1D CNNs helps extract information is the forward directions.

Models used in this work include:

- Vector Autoregression
- LSTM (layers: 100,100,100) with dense layer (21)
- CNN (filters: 64,32)-LSTM (layers: 100,100,100) with dense layer (21)
- CNN (filters: 64,32) - LSTM (layers: 64,32,16,8) with dense layer (21)

The trends in the past 30 days are used to train different models which are then evaluated by comparing forecasted “N” number of days with withheld data not used in training our models. Errors calculated include the root mean squared errors (RMSE), Mean Squared Error (MSE) and Mean Absolute Percentage Error (MAPE). The errors of the models varied with features forecasted from the different models. For example, the MAPE for ventilators was as high as 38% using the VAR model, 2.7% using the LSTM model. Overall, models presented in this project generalize trends observed in the data, with MAPE values less than 15 % excluding the spike in values for the VAR model when predicting ventilator demand. Models can be further improved by tuning the lag parameter of the VAR models, and hyperparameters for the neutral networks.

CONCLUSIONS

This work attempts to make forecasts using multivariate time series of various health related features. Data are augmented from different sources, cleaned, and processed to run on 4 different models. The models are evaluated based on the difference between the forecasted and test data. Model performances varied with the different features, but insights obtained from this work provide a background for future works.