Project Report

**Covid-19 Predictive modeling dashboard**

**1. Problem Statement:**

Coronavirus disease (COVID-19) is an infectious disease caused by the SARS-CoV-2 virus. The first known infections from SARS-CoV-2 were discovered in Wuhan, China. Since then, the virus has spread over multiple companies and evolved into a pandemic. The economic and social disruption caused by the pandemic is devastating. Millions of enterprises face an existential threat. Nearly half of the world’s 3.3 billion global workforce are at risk of losing their livelihoods.

Travel and tours industry is one of the worst hit industries that had a major revenue fall after the outbreak of Covid 19 pandemic. Having a system that can monitor the current situation of Covid around the world provides better control and improve confidence at a time of uncertainty. We are interested in developing the following:

* An ELT pipeline to fetch daily covid data into data lake and perform necessary transformations on the data
* A machine learning model that can make projections on the anticipated number of covid cases for the upcoming 3-month period based on past data and current trends.
* Dashboard with different visualizations for analysing the covid related data and projections on number of active cases, hospitalizations, recovery rate and deaths.

**2. Data**

* [COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University](https://github.com/CSSEGISandData/COVID-19) This is the data repository for the 2019 Novel Coronavirus Visual Dashboard operated by the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE). Also, Supported by ESRI Living Atlas Team and the Johns Hopkins University Applied Physics Lab (JHU APL).
* [Canada population by province](https://www.statista.com/statistics/481509/canada-population-projection-by-province/) This dataset contains the population data of various provinces in Canada

## 3. Method

## 3.1 Data Ingestion using pipelines

Covid – 19 data on number of confirmed cases, hospital admissions, deaths etc. are constantly changing. The primary dataset used here is the GitHub Data Repository by the Center for Systems Science and Engineering of CSSE at Johns Hopkins University.

To make up-to-date predictions and live dashboard, our dataset must update at regular intervals. To achieve this, we are creating ELT pipelines using Azure Data Factory. Azure blob storage is used for storing population data. A pipeline is created for ingesting this data into Azure Data Lake Gen 2. The architecture is shown below.

A picture containing graphical user interface

Description automatically generated

Diagram

Description automatically generated

Another pipeline is created for ingesting daily covid data. The pipeline is scheduled to run every day at a specific time using triggers. The architecture is shown below.

Graphical user interface, application

Description automatically generated

The final data is available in the data warehouse. The Tableau and Power BI dashboard, and ML models fetch the data from this dashboard.

## 3.1 Data Wrangling

[Data Wrangling Report](https://github.com/VargheseTresa/projects/blob/main/Covid%20EDA%20and%20forecasting/project/notebooks/01_data_wrangling.ipynb)

The original dataset includes confirmed cases and deaths data from all over the world since Jan 22, 2020, to the current date. We are specifically interested in only data pertaining to different provinces in Canada. Hence, we first filter out all other data.

**Handling of missing values:**

Null values are present only for Repatriated Travellers. This row will eventually get eliminated when we filter for top 10 provinces based on number of deaths. So no treatment of null values needed.

**Preprocessing and feature engineering:**

For each province, a separate data frame is created. In each province data frame, we convert it into a time series data by using pivot operation. We begin by melting wide data frame is into long data frame. Then the following features are extracted.

* The cases and deaths reported are running totals. From this data, number of daily deaths and confirmed cases are extracted.
* To assist with further analysis, the 7 day and 30-day moving average of confirmed cases and deaths are calculated
* To filter out noise and to identify trends, exponential weighted moving average is calculated for confirmed cases and deaths.
* Mortality rate is calculated as ratio of number of deaths to number of confirmed cases
* Few errors in number of cases and deaths reported (present day number lower than previous day number) led to negative values for number of new cases. These are replaced with zero as number of new cases/deaths for that day.
* The time series data created is written as csv file for further EDA and creation of dashboard in Tableau.

## 3.2 EDA

[Tableau Dashboard](https://public.tableau.com/app/profile/tresa6898/viz/CovidDashboardCanada/Trends-Deaths)

**Trends in number of cases and deaths:**

First, overall trend in number of confirmed cases and deaths since the beginning of reporting are plotted. To get better insights, along with daily new cases, 7-day moving average and 30-day moving average values were also calculated and plotted.

|  |
| --- |
|  |
|  |

Chart, line chart

Description automatically generated

The period in December sees an exponential growth in number of confirmed cases; however, number of deaths doesn’t seem to grow in a similar rate. To confirm that, lets further dig deep into period starting from October 2021. This time, we have a look at number of daily new confirmed cases and deaths, 7-day moving average and exponential weighted moving average. A plot on daily mortality rate is also created.

Chart, line chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Shape

Description automatically generated

Further interactive analysis on various provinces is done using Tableau dashboard. Our final observations from the EDA are:

* Number of confirmed cases are again increasing exponentially in January; however, the number of deaths is not increasing at the same rate based on our initial inspection
* However, the 7 day and 30-day moving average shows an alarming trend in number of deaths.
* To make more accurate analysis and to identify trends, we filter out noise by computing exponential weighted moving average. This also confirms an alrming increase in the number of deaths.
* The mortality rate is reduced significantly when compared to the beginning months when the disease started spreading out.
* Province wide analysis reveals the top states with highest number of confirmed cases and death in the following order
  1. Quebec
  2. Ontario
  3. Alberta
  4. British Columbia
  5. Alberta

**Stationarity check for time series data:**

Graphical user interface

Description automatically generated with medium confidence

Based on the results from Dickey-Fuller Test, first difference method is applied to make the data stationary.

Chart, histogram

Description automatically generated

## 3.3 Predictive Modeling

[Modeling Notebook](https://github.com/VargheseTresa/SpringBoard/blob/main/CAP2/CAP2_modelling.ipynb)

**LSTM**

Long Short-Term Memory is a kind of recurrent neural network (RNN) architecture. The RNN are mainly used in processing sequential data (text, natural language or image captioning) and in time series forecasting. Their main difference with feedforward or convolutional networks is the fact that they have some sort of ‘memory’. RNNs feed the output back as an input, making the output dependent on prior events.

**Why LSTM?**

The idea behind RNN was to build a NN that was able to learn to use past information. When the useful information is close in time, RNN can do the job. But If we need to go further back in time RNN fails, and here is where LSTM comes into play. LSTMs are capable of keeping the important information, doesn’t mind of back in time it is, and forget the useless one.

**Hyper parameter tuning:**

In every machine learning algorithm, we have the parameters and the hyperparameters. Parameters are learnt by the algorithm itself by training. On the other hand, hyperparameters are set manually by the user.

For Neural Networks, there is no clear guidelines or any formal procedure to design and choose the neural network hyperparameters. So, usually trial and error techniques and intuition are used.

The parameters we need to tune in our LSTM deep learning model are

1. Number of training epochs
2. Number of batches

**4. Takeaways**

* EDA revealed that customers who are on month-to-month contract have high probability to churn compared to those on 2-year contract.
* Churn Score is a good predictor of customer churn. The difference between average churn score of those who churn and do not churn is between 31.74 and 33.01.
* Among the customers who churned, 70% of them had Fibre Optic Internet Service with the company. Analysis of churn reason also reveals that 33% of the customers who left the company were offered with better internet or devices by competitors.
* Among the customers who left the company, 37% were not satisfied with the services provided by the company.

**5. Future Extensions**

Availability of transaction data of customers can further assist in finding time sensitive trends in customer behavior.

More research on competitors in the market and the services they offer can reveal any attention needed on the services and packages offered by the company. It also helps in understanding where the company stands in the competitive market.

**6. References**

Not all the work in this notebook is original. Some parts were borrowed from online resources. I take no credit for parts that are not mine. They were solely used for illustration purposes.

1. https://www.statisticshowto.com/welchs-test-for-unequal-variances/

2. https://medium.com/@sosterburg/mapping-data-with-folium-356f0d6f88a9