Pandemia: A COVID-19 Predictor

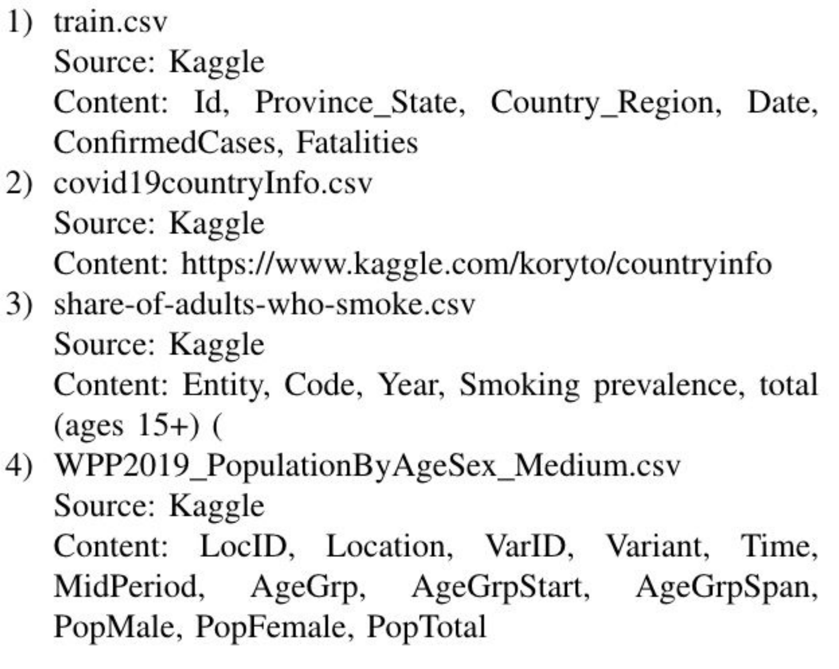
Team
Alaukika Diwanji: 014547013
Amit Garg: 014541072
Nachiket Trivedi: 014544933

Introduction

- Problem Statement:
  - Predicting the global turnout of COVID-19 pertaining to each and every country, specific to the parameters of confirmed cases and fatalities.
  - Predicting the same for each individual state and territory of the United States.

- Datasets:
  - Formulated 6 different datasets from various sources such as Kaggle, CDC, WHO, Johns Hopkins University, US Government Records etc.
  - The datasets range from the dates 22nd January to 17th April, 2020.
  - The first problem statement consists data of 150+ countries while the second one consists data of all 50 US states and 4 territories.

Data Pre-processing: Global Forecast

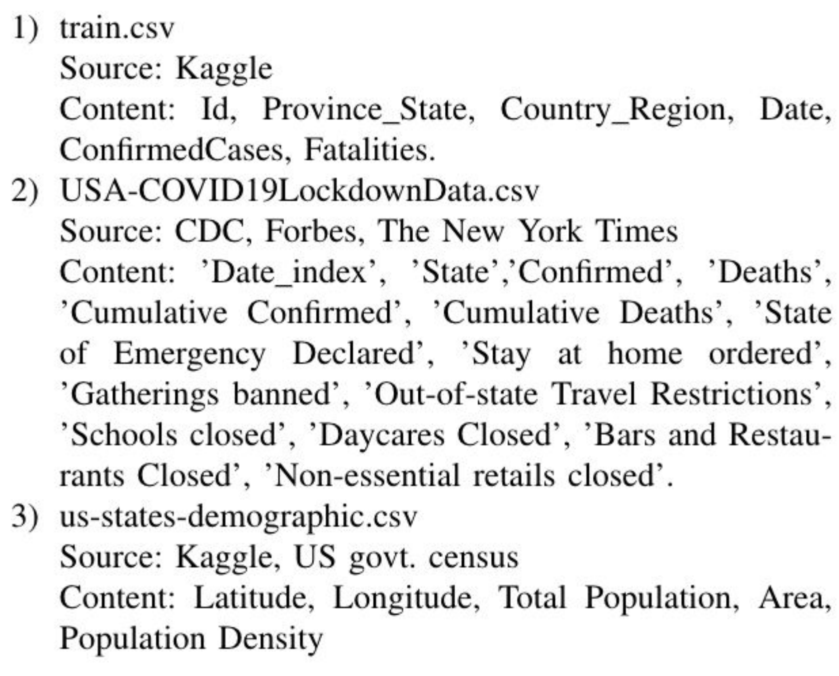


1) train.csv
   Source: Kaggle
   Content: Id, Province_State, Country_Region, Date, ConfirmedCases, Fatalities
2) covid19countryInfo.csv
   Source: Kaggle
   Content: https://www.kaggle.com/koryto/countryinfo
3) share-of-adults-who-smoke.csv
   Source: Kaggle
   Content: Entity, Code, Year, Smoking prevalence, total (ages 15+) (
4) WPP2019_PopulationByAgeSex_Medium.csv
   Source: Kaggle
   Content: LocID, Location, VarID, Variant, Time, MidPeriod, AgeGrp, AgeGrpStart, AgeGrpSpan, PopMale, PopFemale, PopTotal

- Resolving format conflicts. Eg: Country_Region(US, U.S., United States) etc.
- Resolving data inconsistency. Eg: train.csv(data till 17th April), covid19countryInfo.csv(data till 25th March) etc.

- Final dataset: *Date, Country, Confirmed Cases, Fatalities, quarantine, schools, hospibed(hospital bed availability per 1000 people), lung(death rate due to lung disease per 100k people), total pop, density(population density), age65+(proportion of people aged more than 65 years), smokersperc.*

Data Pre-processing: US States’ Forecast

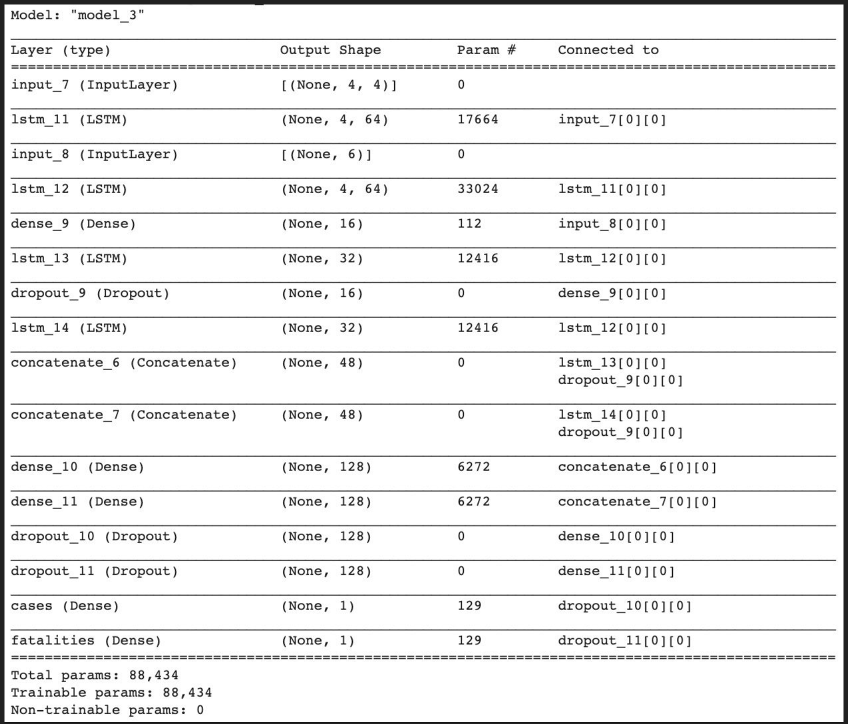


1) train.csv
   Source: Kaggle
   Content: Id, Province_State, Country_Region, Date, ConfirmedCases, Fatalities.
2) USA-COVID19LockdownData.csv
   Source: CDC, Forbes, The New York Times
   Content: ’Date_index’, ’State’,’Confirmed’, ’Deaths’, ’Cumulative Confirmed’, ’Cumulative Deaths’, ’State of Emergency Declared’, ’Stay at home ordered’, ’Gatherings banned’, ’Out-of-state Travel Restrictions’, ’Schools closed’, ’Daycares Closed’, ’Bars and Restaurants Closed’, ’Non-essential retails closed’.
3) us-states-demographic.csv
   Source: Kaggle, US govt. census
   Content: Latitude, Longitude, Total Population, Area, Population Density

- Got macro data like gatherings banned(10 After 15th March, 5 After 25th March) etc.
- Mapped these macro values to numerical values by assigning weights as per their significance.

- Final Dataset: *Date, State, Confirmed, Deaths, Cumulative Confirmed, Cumulative Deaths, State of Emergency Declared, Stay at home ordered, Gatherings banned, Out-of-state Travel Restrictions, Schools closed, Daycares Closed, Bars and Restaurants Closed, Non-essential retails closed, Lat, Long, Population, Area, Density.*

Algorithms Used: Global



```
Model: "model_3"
__________________________________________________________________________________________________
Layer (type)                    Output Shape         Param #     Connected to
==================================================================================================
input_7 (InputLayer)            [(None, 4, 4)]       0
__________________________________________________________________________________________________
lstm_11 (LSTM)                  (None, 4, 64)        17664       input_7[0][0]
__________________________________________________________________________________________________
input_8 (InputLayer)            [(None, 6)]          0
__________________________________________________________________________________________________
lstm_12 (LSTM)                  (None, 4, 64)        33024       lstm_11[0][0]
__________________________________________________________________________________________________
dense_9 (Dense)                 (None, 16)           112         input_8[0][0]
__________________________________________________________________________________________________
lstm_13 (LSTM)                  (None, 32)           12416       lstm_12[0][0]
__________________________________________________________________________________________________
dropout_9 (Dropout)             (None, 16)           0           dense_9[0][0]
__________________________________________________________________________________________________
lstm_14 (LSTM)                  (None, 32)           12416       lstm_12[0][0]
__________________________________________________________________________________________________
concatenate_6 (Concatenate)     (None, 48)           0           lstm_13[0][0]
                                                                 dropout_9[0][0]
__________________________________________________________________________________________________
concatenate_7 (Concatenate)     (None, 48)           0           lstm_14[0][0]
                                                                 dropout_9[0][0]
__________________________________________________________________________________________________
dense_10 (Dense)                (None, 128)          6272        concatenate_6[0][0]
__________________________________________________________________________________________________
dense_11 (Dense)                (None, 128)          6272        concatenate_7[0][0]
__________________________________________________________________________________________________
dropout_10 (Dropout)            (None, 128)          0           dense_10[0][0]
__________________________________________________________________________________________________
dropout_11 (Dropout)            (None, 128)          0           dense_11[0][0]
__________________________________________________________________________________________________
cases (Dense)                   (None, 1)            129         dropout_10[0][0]
__________________________________________________________________________________________________
fatalities (Dense)              (None, 1)            129         dropout_11[0][0]
==================================================================================================
Total params: 88,434
Trainable params: 88,434
Non-trainable params: 0
```

- Deep Learning
    - Deep Neural Networks
    - Recurrent Neural Networks (Long Short Term Memory Units)
- We used the dense neural network model for the demographic data for training such as: density, population etc.
- To preserve sequentiality and consider the past trends, we use the LSTM recurrent neural network.

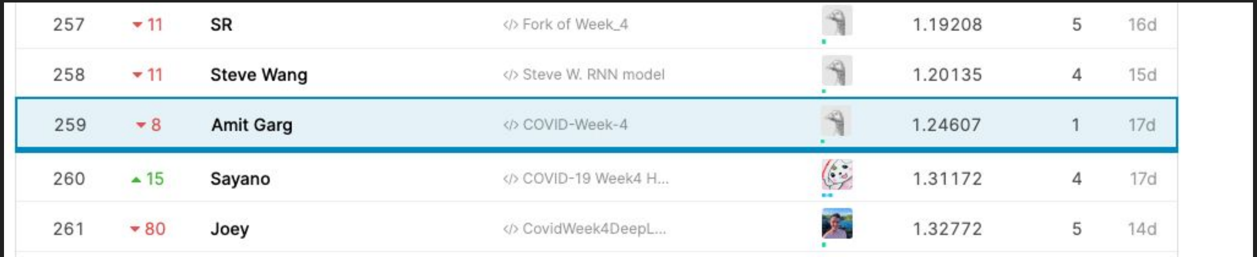
Algorithms Used: US States

- Considered past cumulative confirmed cases and deaths to predict for the specific day. For eg: for predicting data for 11th April, we took into consideration the data of 8th, 9th and 10th April as well.

- Used Multiple Linear Regression and Support Vector Regression for each state and territory, and used the model which gave the best results, individually.
  - Multiple Linear Regression: As here we have multiple fields such as state at home ordered, schools closed etc., we can aptly use this model, as it requires multi parametered x values.
  - SVR: As each data row doesn't vary to a larger extent, SVR implementation seems apt.

Global Prediction

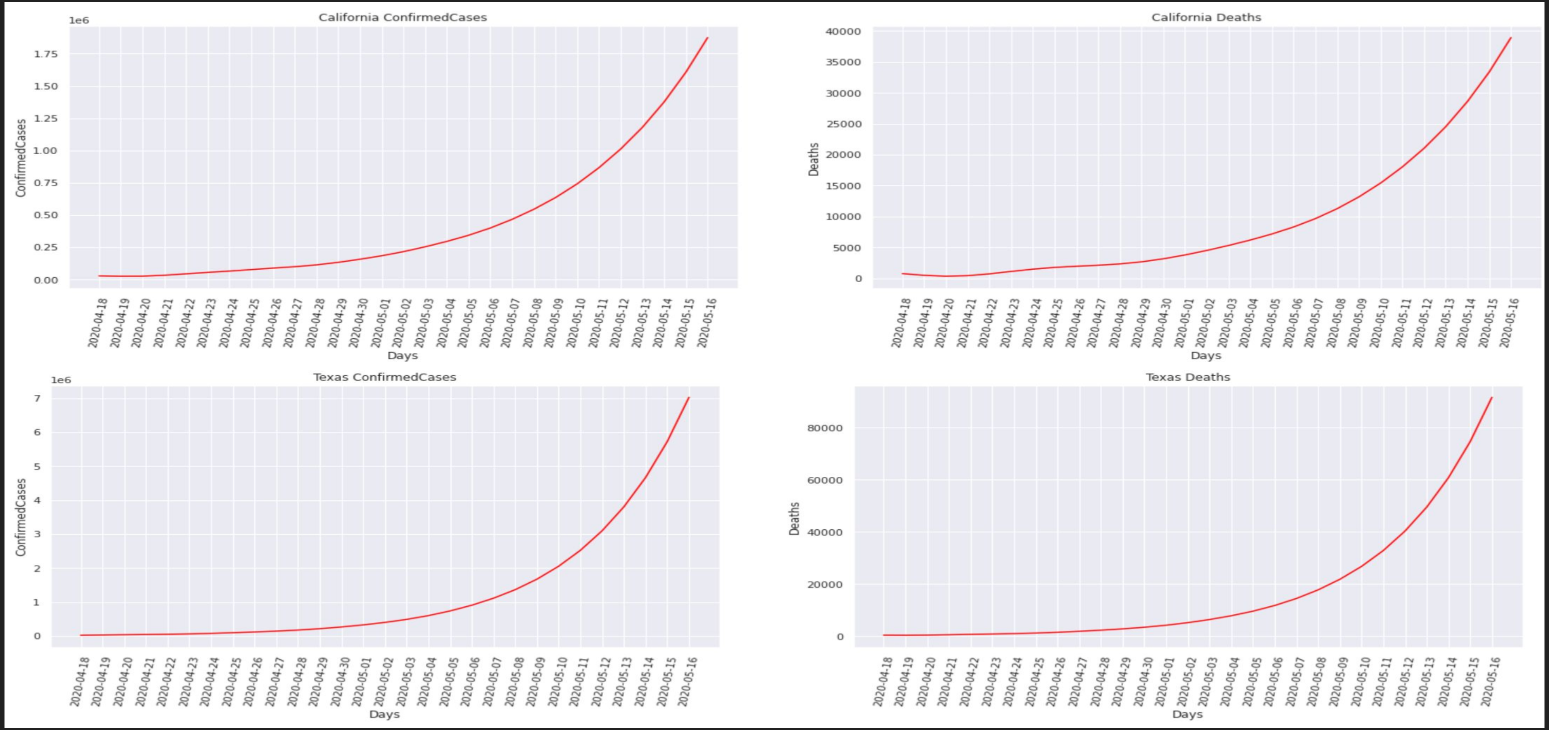
- The training parameters consist of the demographic data as well as temporal data, with 90% of that being our training data.
- We train our model for 250 epochs, for minimizing loss and thereby increasing the performance metric.
- We've also submitted our notebook in the Kaggle competition COVID 19 Global Forecasting.



257	▼11	SR	</> Fork of Week_4		1.19208	5	16d
258	▼11	Steve Wang	</> Steve W. RNN model		1.20135	4	15d
259	▼8	Amit Garg	</> COVID-Week-4		1.24607	1	17d
260	▲15	Sayano	</> COVID-19 Week4 H...		1.31172	4	17d
261	▼80	Joey	</> CovidWeek4DeepL...		1.32772	5	14d

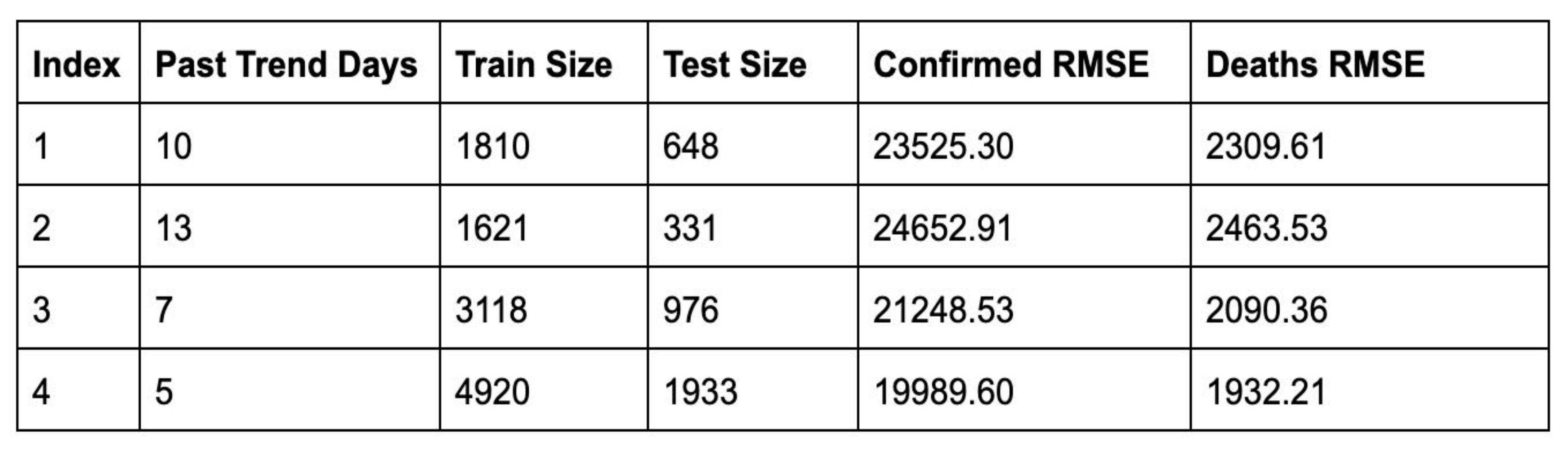
US State Prediction

- We’ve picked our best regression models based on the RMSE scores and following are the predictions of California and Texas from 17th April to 15th May.



Experiments: Global

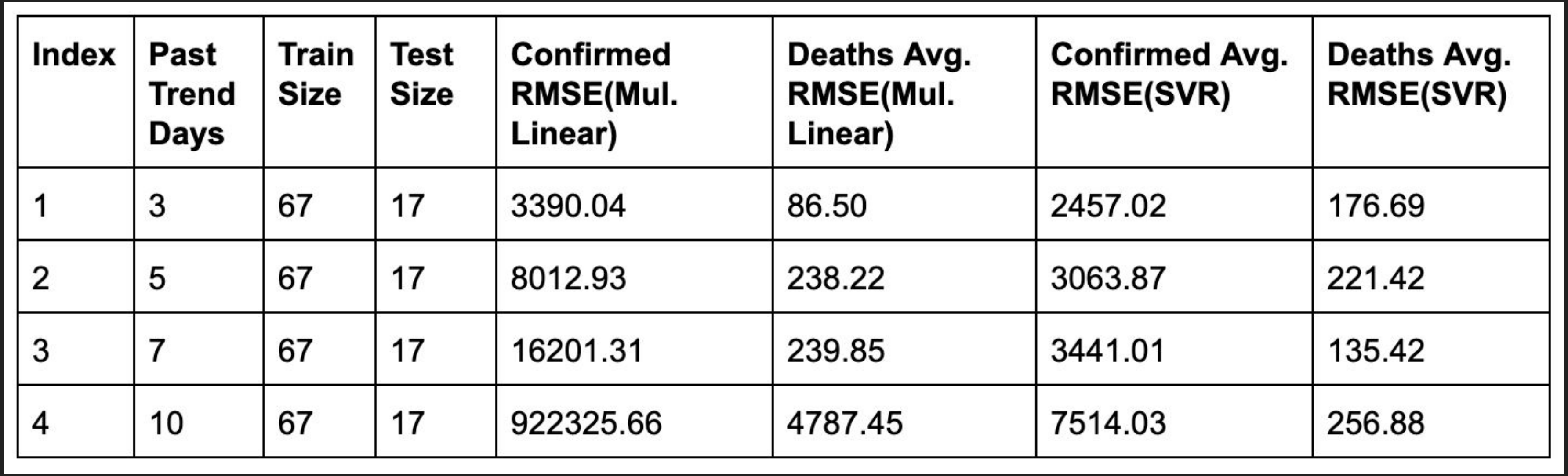
- We considered and ran our algorithm for 5, 7,10, 13 days and calculated the RMSE for Confirmed cases and fatalities.



Index	Past Trend Days	Train Size	Test Size	Confirmed RMSE	Deaths RMSE
1	10	1810	648	23525.30	2309.61
2	13	1621	331	24652.91	2463.53
3	7	3118	976	21248.53	2090.36
4	5	4920	1933	19989.60	1932.21

Experiments: US States

- We considered and ran both our models for 3, 5, 7, 10 days and calculated the RMSE for confirmed cases as well as fatalities, and compared the results of both multiple linear regression and SVR.

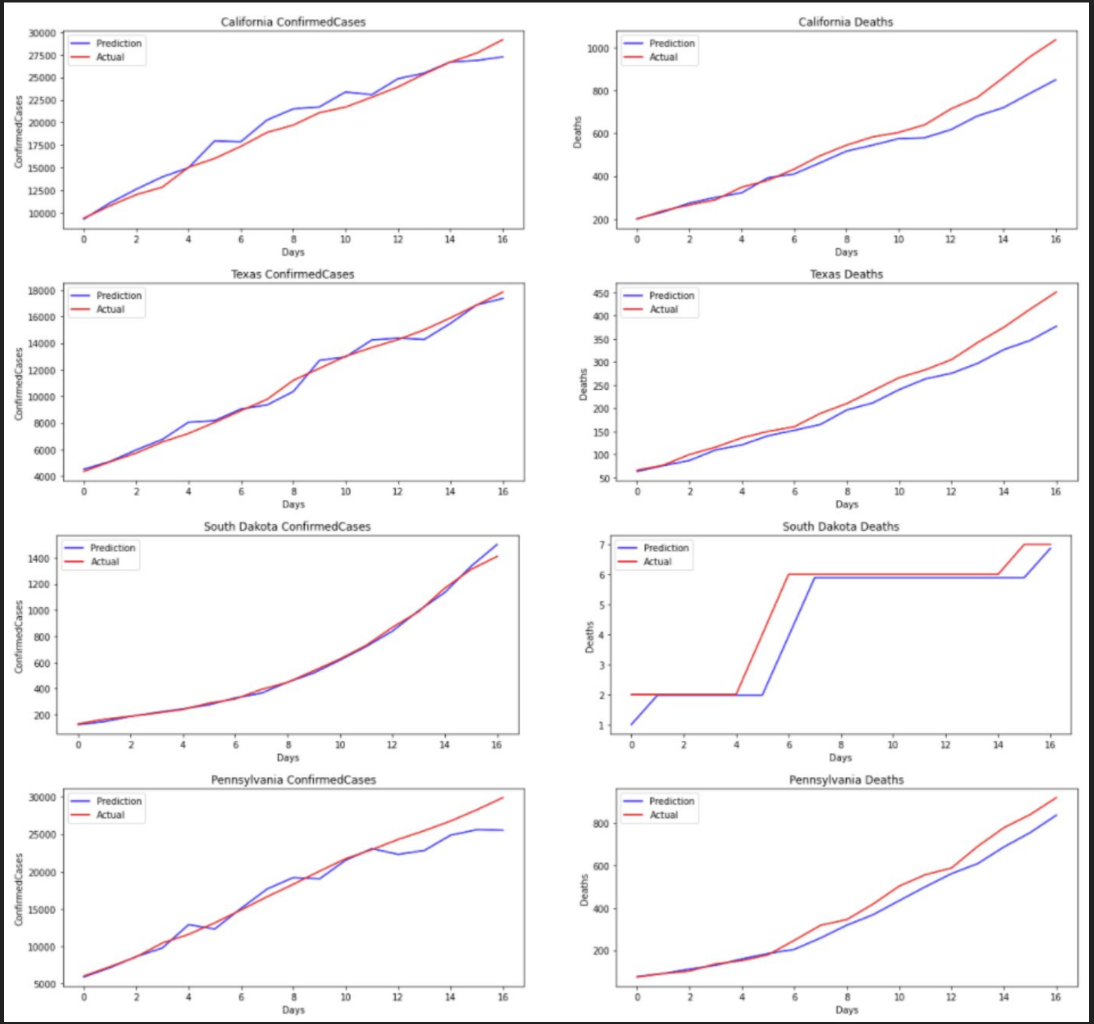


Index	Past Trend Days	Train Size	Test Size	Confirmed RMSE(Mul. Linear)	Deaths Avg. RMSE(Mul. Linear)	Confirmed Avg. RMSE(SVR)	Deaths Avg. RMSE(SVR)
1	3	67	17	3390.04	86.50	2457.02	176.69
2	5	67	17	8012.93	238.22	3063.87	221.42
3	7	67	17	16201.31	239.85	3441.01	135.42
4	10	67	17	922325.66	4787.45	7514.03	256.88

Evaluation

- We evaluated our model on validation data from 1st April to 17th April.

- These are our top 4 charts based on RMSE.



Conclusion

- We addressed two major problem statements: COVID-19 global predictor and US states' predictor, and provided thorough experimentation for the same.
- Things that didn't work:
  - In the state-wise predictor, the available data was very less (only the states and territories). Hence couldn't apply deep learning.
  - Random Forest and Decision Trees didn't display a decent convergence of loss, hence used multiple linear regression and SVR
- Future work: We intend to apply diligent domain knowledge to gather a large amount of data (specific to each US states and territories), and thereby apply deep learning models.

Thank you