COVID-19 FORECASTING THROUGH LENS OF MACHINE LEARNING & DEEP NEURAL NETWORKS

SANJEEV.
Michigan State University
East Lansing, MI
{thenkara}@msu.edu

Abstract

This Project encompasses implementation of Stacked LSTM (RNN based DNN) , XGBoosting (ML- Boosting Regressor model) ARIMA (Time Series Model) under same setting to juxtapose performance of prediction of the Confirmed Cases Fatalaties accurately with least Root Mean-Squared_Error

1. Introduction

Corona Virus ,widely known as COVID-19 is an ongoing global pandemic disease that causes illness in the respiratory system in humans, thereby posing a devastating effect in our day-to-day life. It has affected millions of peoples, who are either sick or are being killed due to the spread of this disease.Thus, detection of spread of COVID-19 would be of cardinal importance to aid in recognizing qualitative decision making The recent groundbreaking advancements in the domain of Artificial Intelligence, particularly in the area of Natural Language processing Computer Vison employing efficient Machine Learning Deep Learning techniques by Researchers have made it feasible to solve complex problems with ease.

Many researchers, including data scientists, have been working intensely to determine ways to eradicate this disease. Data scientists can effectively contribute to the research by designing prediction models that highlight the probable activities of this virus, which can further help in accurately predicting the spread of this virus. Hence, deep learning (DL) models are regarded as accurate tools which can help in developing prediction models. Though many neural networks (NNs) have been described in the past, the recurrent neural network (RNN) and the long short-term memory (LSTM) are investigated in the forecasting of COVID-19 as they can use temporal data

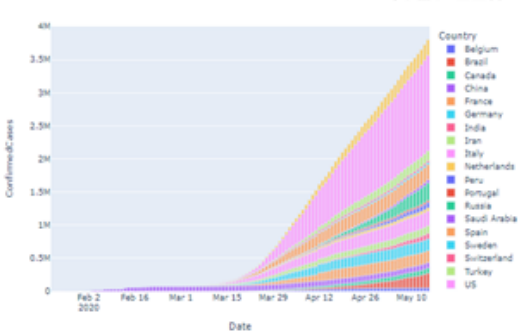
This work aims to comprehensively study and analyze rising confirmed cases fatalities due to outbreak based on Deep Learning , Machine Learning,Time Series models that extrapolate or forecast or predict such cases in future. Since Ensemble(bagging boosting model),DNN (LSTM its derivatives) Time series models(ARIMA its derivatives) have shown considerable success in forecasting temporal data, thus selecting these models to accurately predict future trends.

2. Database Description

The primary dataset employed for implementation of COVID-19 Forecast is provided by Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE), which is frequently updated to keep track of live confirmed cases fatalities. It is located at the github repository CSSEGISandDataCOVID-19 for reference. The dataset consists of following columns namely ID,Province state,Country_region,Date, ConfirmedCases,Fatalities. The number of registered deaths by Covid-19 accounting to total of 35995 training_set 13456 test set as of 15th May'20. Initially in the Exploratory Analysis of Data (EDA) phase, data wrangling or preprocessing has been performed to cleanse, scale the dataset to be devoid of missing values, outliers, imbalanced, distribution post which data deems to be fit to enable gain crucial insights of rise in global covid cases and fatalities. During this period ,plots pertaining to global forecast that of top 10 affected countries that witnessed spike in the frequency of cases have been laid to analyze the drift. Following which new columns were engineered that corresponds to lag trend based features to arrive at an estimated prediction based on past/previous metrics. Furthermore demographic features pertaining to population count can also be employed to further deep dive to aid accurate estimation.

3. Model Architecture

A. Xgboost Regressor modelXg_boost stands for Extreme Gradient Boosting.It falls under the family of Ensemble model employing decision trees. These techniques differs based on the build and traversals i.e Bagging



Country	ConfirmedCases
US	39749076.0
Italy	9386990.0
Spain	9322515.0
China	7946795.0
Germany	7080489.0
France	6749906.0
United Kingdom	6360259.0
Iran	4358781.0
Turkey	4179317.0
Russia	3965789.0
Brazil	3290280.0

Country	Fatalities
US	2122579.0
Italy	1211456.0
Spain	995560.0
United Kingdom	933303.0
France	899572.0
China	333639.0
Iran	273392.0
Belgium	261746.0
Germany	229882.0
Brazil	218965.0

models(ex-Random Forest model)builds the tree in parallel fashion , whereas Boosting models tend to iteratively learn through errors by combining with each other to generate a collectively strong model, thereby optimize for better generalizability in sequential fashion . Key_traits encompass parallization, distritbuted computing cache Optimization, Simultaneously, the model shows the ranking of input features. Additionally, the eXtreme Gradient Boosting (XGBoost) model is a flexible machine learning method capable of dealing with the non-linearity of time series through its strong self-learning ability. Moreover, can help

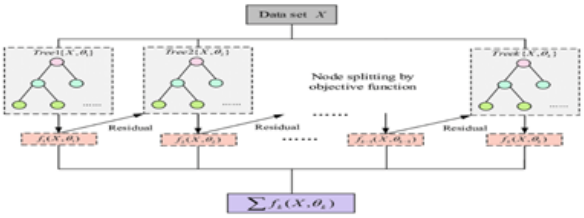


Figure 2

us obtain a stronger classifier from other classifiers and has other perks, such as avoiding overfitting, effectively dealing with missing values and reducing running time by parallel and distributed calculation.

B). LSTM model :Long Short-Term Memory Networks (LSTM) is a variantof RNN, it creates several gates to each single RNN cell, adding more flexibility and efficiency to the model [3]. Those gates help the model to memorize information over long input sequences. The

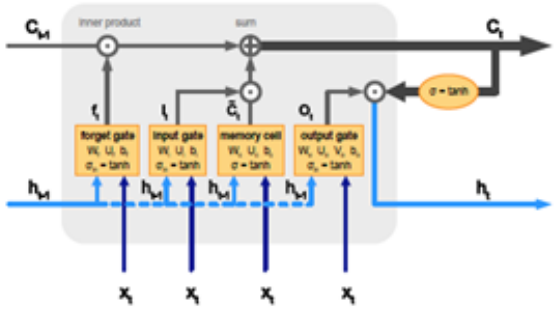


Figure 3

Stacked LSTM is an extension to this model that has multiple hidden LSTM layers where each layer contains multiple memory cells. This enriches the models to create more complex feature representation when tackling greater model complexity.

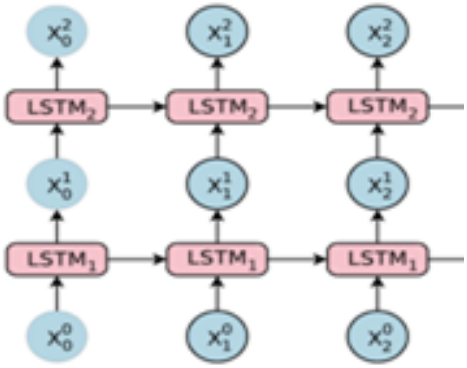


Figure 4

c) ARIMA:[Auto-Regressor integrated Moving Average]. It entails 3 components- where Auto-Regressive -implies that the target variable is regressed on its own previous value;Integrated denotes a differencing step applied to the data,recursively untilo it attains the state of Non-Seasonality ;Moving Average indicates that Mu(mean) of given model centers its forecast around the mean of Y.
Formula AR-> Y = B0 + B1∗Y_lag1 + B2∗Y_lag2 + ... + Bn∗Y_lagn
MA-> Y = Mu + B1∗E_lag1 + B2∗E_lag2 + ... + Bn∗E_lagn
I-> number of differences captured

Optimal parameters of Arima model(p,d,q) is chosen

based on Partial autocorrelation function(PACF) ,count of differences(d) Autocorrelation function(ACF), where PACF,ACF plots of AR,MA are captured resepectively . ARIMA model is trained on the distribution post achieving the state of Stationarity(i.e by employing Augmented Dickey Fuller-Test(ADF test).

4. Experimental Results & Graphical Plots

The system & software specification employed to conduct the experimental plots results are Cpu: Model name: Intel(R) Xeon(R) CPU @ 2.00GHz,Gpu: Tesla P100-PCIE-16GB Software:Linux,TensorFlow.

a). Xgboost _Plots

Optimal parameters chosen: optimal_chosen(n_estimator=1000,early_stopping_rounds =50,objective='reg:linear',max_depth=3),

varied-(boostertype=Dart,(boostertype=gb_linear)

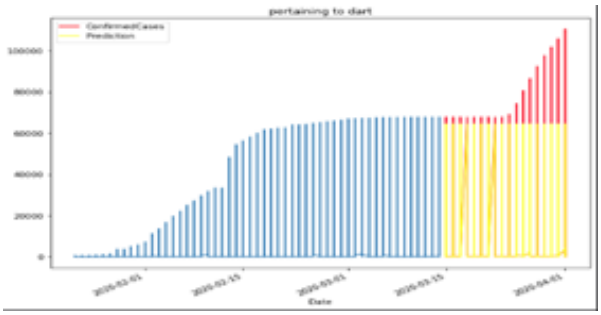


Figure 5. boost-typeDart rmse_score=**2148.574**

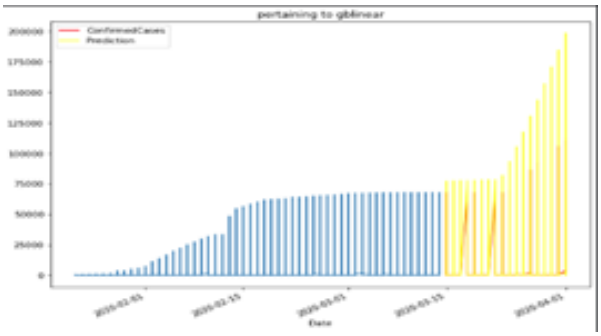
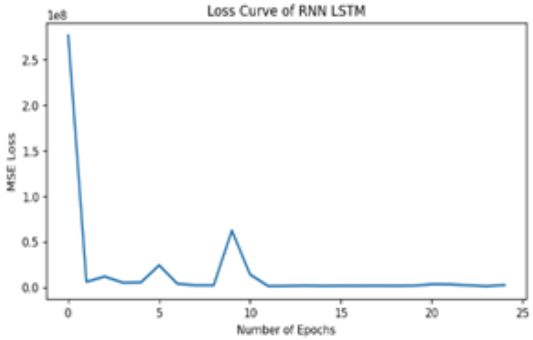


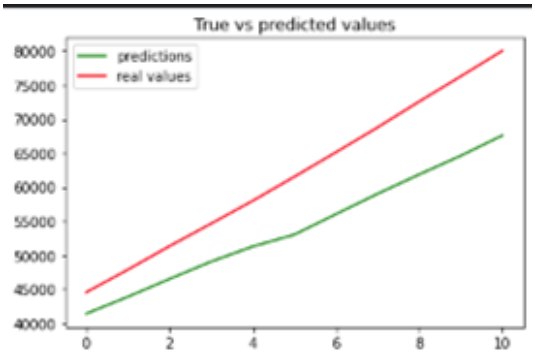
Figure 6. boost-typeGb_line rmse_score=**1910.570**

b). Stacked LSTM_Plots

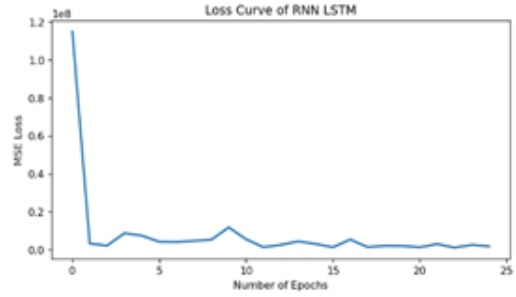
Optimal parameters chosen: optimal_chosen (activation=Relu,Optimizer=Adam,loss=mse,epochs=



(a) Rmse_Score-**7176.6403**



(b) LSTM- Prediction vs Ground_truth

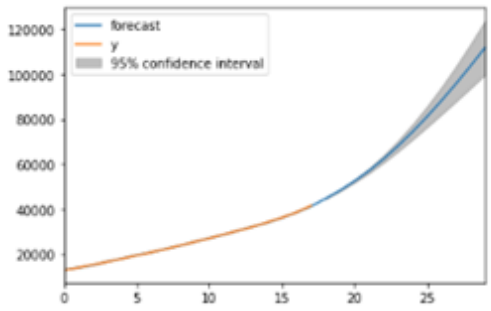


(a) Rmse_=**1338.6635**

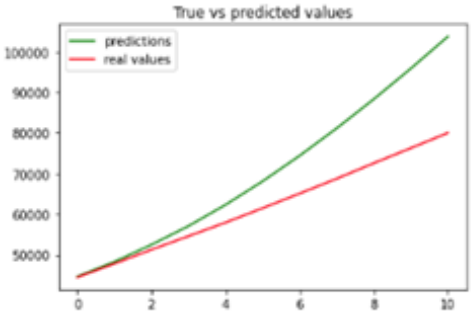
(b) Stacked_LSTM- Prediction vs Ground_truth

25,early_stopping_rounds=50,layers = 3 **Lstm** layers,objective='reg:linear',max_depth = 3), varied-(hidden layers_size ,Dropout,return sequence=True/False)

c). ARIMA_Plots Optimal parameters chosen: Optimal_chosen varied prameters-(autolag based on different



(a) Rmse_ Score=**11700.5685**



(b) Prediction vs Ground_truth

countries in ADF test)

5. SUMMARY

In summary we have compared the prediction of Covid-19 pandemic with boosting , Recurrent LSTM ARIMA models under same setting of lag period of 7-14 days . Overall performance of stacked_LSTM based model outperformed other models for long range estimation. Furthermore, if the number of hidden layers are increased ,it loses its capacity to generalize over the test dataset. Thus additional exogenous features must be extracted to more accurately predict the forecast.An important aspect found in the process of recursive training over the similar span of lag/delay periods(7-14)(2 weeks) is that prediction results employing ARIMA model tend to be consistent, unfortunately it lags behind other DNN ML Regressor models. This research can be further expanded to forecasting problems for multivariate and Seasonal time series for lag /delay period greater than one month .Other factors of interests such as other states and tuning hyper-parameters could be further investigated.

The future scope of work could be extended by integrating these models in conglomeration with optimal hyperparameter tuning to achieve minimum error.

References

[1] A. A. Adebiyi, A. O. Adewumi, C. K. Ayo, "Stock Price Prediction Usingthe ARIMA Model," in UKSim-AMSS 16th International Conference on Computer Modeling and Simulation., 2014.

[2] A. M. Alonso, C. Garcia-Martos, "Time Series Analysis - Forecastingwith ARIMA models," Universidad Carlos III de Madrid, UniversidadPolitecnica de Madrid. 2012.

[3] P. Baldi, S. Brunak, P. Frasconi, G. Soda, and G. Pollastri, "Exploitingthe past and the future in protein secondary structure prediction," Bioinformatics, 15(11), 1999.

[4] J. Brownlee, "How to Create an ARIMA Model for Time SeriesForecasting with Python," https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/, 2017.

[5] J. Brownlee, "Time Series Prediction with LSTM Recurrent NeuralNetworks in Python with Keras," https://machinelearningmastery.com/time-series-prediction-lstm-recurrent-neural-networks-python-keras/,2016.

[6] G. Box, G. Jenkins, Time Series Analysis: Forecasting and Control, SanFrancisco: Holden-Day, 1970.

[7] W. Langel, "Extrapolation of infection data for the covid-19 virus and estimate of the pandemic time scale."medRxiv, 2020. [Online].

[8] Wiersinga WJ, Rhodes A, Cheng AC, et al. Pathophysiology,transmission, diagnosis, and treatment of coronavirus disease 2019(COVID-19). JAMA 2020;324:782–93.

[9] Singh S, Parmar KS, Kumar J, et al. Development of new hybridmodel of discrete wavelet decomposition and autoregressiveintegrated moving average (ARIMA) models in application to onemonth forecast the casualties cases of COVID-19. Chaos SolitonsFractals 2020;135:109866.

[10] CDC COVID-19 Response Team. Geographic Differences inCOVID-19 Cases, Deaths, and Incidence - United States, February12-April 7, 2020. MMWR Morb Mortal Wkly Rep 2020;69:465–71.

[11] Aslam S, Adler E, Mekeel K, et al. Clinical effectiveness of COVID-19 vaccination in solid organ transplant recipients. Transplant Infectious Disease 2021;23.

[12] Yengil E, Onlen Y, Ozer C, et al. Effectiveness of booster measlesmumps-rubella vaccination in lower COVID-19 infection rates:a retrospective cohort study in Turkish adults. Int J Gen Med 2021;14:1757–62.

[13] Centers for Disease Control and Prevention. Trends in number of COVID-19 vaccinations in the US. COVID