2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Time Series Data Modelling of COVID-19 Positive Data applying popular Ensemble and Deep Learning Algorithm

. Shrirang A. Kulkarni, Varadraj P. Gurupur, *Sr. Member, IEEE*, Christian King

***Abstract*— Modelling of COVID-19 as time series data for machine learning problems for greater prediction accuracy is a challenging problem. This problem is further compounded when we attempt to extrapolate the data. In the present work popular ensemble strategies are used to accurately model Florida State COVID-19 positive cases as a time series data. Experimentally it was found that the optimized XGBoost model was superior to Random Forest in terms of Root Mean Square Error (RMSE) which indicated data around best fit, by 60.97% and to CatBoost by 50.69%. In terms of Mean Absolute Error (MAE) which predicted the average of deviation from true values, XGBoost outperformed Random Forest by 61.5% and CatBoost by 51.25%. XGBoost fared relatively poorer for a single step extrapolation and it deviated from the observed value as a regression error by 22.01%. However it was found that the Multi-Layer Perceptron (MLP) as a Deep Learning algorithm outperformed XGBoost in terms of extrapolation regression error by 6.629% and indicated more promising results. This clearly indicates us to use the Deep Learning Model for generalization and extrapolation of time series values.**

***Clinical Relevance*— This modeling of healthcare time series data is useful to extrapolate COVID 19 infections as time series data and prepare emergency medications and health services.**

I. INTRODUCTION

Random Forest is a popular bagging method that builds an ensemble of decision trees accurately [1]. However XGBoost [2] one of the powerful ensemble technique based on gradient-boosting learning has emerged as a popular technique with high accuracy of prediction, due to its feature regularization capability. CatBoost [3] another gradient boosting strategy provides accurate results in a short time.

II. METHODS AND RESULTS

The Figure 1 illustrates hyperparameter tuning of the ensemble models and it was observed that Random Forest performed optimally for 750 trees and XGBoost and CatBoost performed well for 1250 trees. The dataset used was COVID-19 Florida State Positive cases from [4].The dataset consisted of 347 records and was converted to supervised time series data for application to ensemble methods. The ensemble models experimental results are illustrated in Table 1. From Table 1 in terms of MAE XGBoost is superior to Random Forest by 61.5% and CatBoost by 51.25%. In terms of RMSE XGBoost outperformed Random Forest by 60.97% and CatBoost by 50.69%. Deep Learning Multilayer Perceptron (MLP) is a type of feed-forward neural network [5] which was applied to single step extrapolation of data and the results were compared with XGBoost.

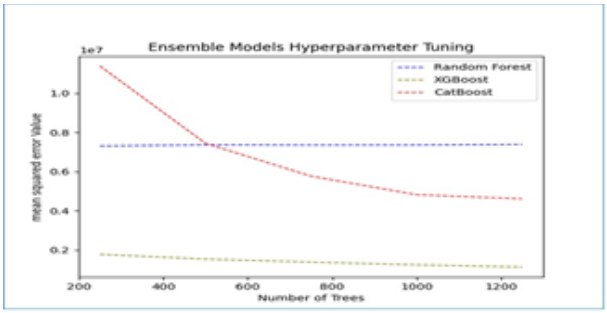

Figure 1. Hyperparamter tuning of Random Forest, XGBoost and CatBoost.

TABLE I. RANDOM FOREST, XGBOOST AND CATBOOST APPLIED TO COVID-19 POSITIVE TIME SERIES DATA

Ensemble Model	MAE	RMSE
Random Forest	2694.311	2710.014
XGBoost	1037.274	1057.702
CatBoost	2127.951	2145.368

The actual positive cases data of day *n* was 2520, Deep Learning MLP predicted it as 2699 and XGBoost predicted it as 3243. Therefore, the regression error was minimal for Deep Learning MLP at 6.63% and was higher for XGBoost since it was 22.29%. This clearly indicated that Deep Learning MLP can be more effective in predicting the outcome of positive COVID-19 cases for time series data with more accuracy when compared to XGBoost, Random Forest, and CatBoost.

REFERENCES

[1] L Breiman, "Random Forests", *Machine Learning*, vol.45, 2001, pp. 5–32

[2] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System", *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016 pp. 785–794.

[3] .L. Prokhorenkova, G.Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features", *In Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS'18)*, Curran Associates Inc., Red Hook, NY, USA, 2018, pp. 6639–6649.

[4] The Data – The Covid Tracking Project, Creative Commons CC BY 4.0, https://covidtracking.com/data,

[5] H. I. Fawaz, G. Forestier, J. Weber, L. Idoumghar & P-A. Muller , "Deep learning for time series classification: a review", *Data Mining and Knowledge Discovery*, vol.33, 2019, pp.917–963

Shrirang A. Kulkarni is with School of Global Health Management and Informatics at University of Central Florida, Orlando, 32816, USA, phone: 407-979-6289; e-mail: sakulkarni@ucf.edu .