ISSN No:-2456-2165

Ensemble Learning Approach to Improve Existing Models

Agrim Mehra[1], Priyansha Tripathy[2], Ashhad Faridi[3], Ayes Chinmay[4]
Department of Computer Science and Engineering
International Institute of Information Technology
Bhubaneswar, Odisha

Abstract:- Machine learning is becoming an exciting field because it can be applied to the different problems we face in our daily lives. An essential function of machine learning is predicting the possibility of any future event. Improving this accuracy of prediction has always been one of the most challenging aspects of Machine Learning. In this paper we have tried to improve the accuracy by Combining the Ensemble Learning [1] approach of Bagging [2] and Boosting [3] with Linear regression. The problem of extrapolation in case of Bagging techniques is also explored and corrected using Ensemble methods. The combination will yield better results when used in the case of Bagging techniques as compared to the results given by individual models. Simulations of these algorithms are achieved in R and are further demonstrated in future sections.

Keyword:- *Ensemble Learning, Bagging, Boosting, Multiple Linear Regression, RMSE, Random Forest, Extrapolation.*

I. INTRODUCTION

Machine learning is an artificial intelligence(AI) applica-tion which provides systems the capability to assimilate and improve instinctively from experience without explicit pro-gramming. It consists of many techniques for making accurate predictions, one of them being Ensemble Learning.

Ensemble learning is a technique of machine learning in which we combine multi - model decisions to improve overall performance. Ensemble Learning involves various techniques, such as Stacking, Bagging, Boosting, Blending. In this paper the main focus is on increasing the accuracy of Bagging and Boosting Ensemble Learning methods. In this paper, along with these Ensemble learning techniques, Linear Regression Machine Learning model is also applied. Difference in the accuracy of these models is shown individually and when used in combination using the Root mean Square error (RMSE) value.

II. APPROACH OVERVIEW

The following sections discuss important terms associated with the application of our approach:

A. Multiple Linear Regression

Multiple linear regression [6] (MLR), also known as multi-ple regression, is an analytical method which uses more than one explanatory variables to predict the value of our response variable. The whole basis of the Multiple Linear Regression is modelling the relationship between the explanatory variables and the dependent variable.

MLR is one of the most basic and effective Machine Learning technique used to predict the dependant variable.

$$y_i = {}_0 + {}_1x_{i1} + {}_2x_{i2}:::: {}_px_{ip} +$$

For i equals to n observations, where:
y_i is our response variable,

x_i are the independent variables of our model, ${}_0$ is the constant term (y-intercept),
${}_p$ defines the slope coefficients of each independent variable, is the error term of our model.

B. Bagging

Bagging technique is also known as Bootstrap aggregation. It is one of the methods used in Ensemble Learning. Bag-ging is, applying the Bootstrap procedure to a high-variance machine learning algorithm [7] such as Decision Tree etc. Random forest is a Machine Learning method that uses the concept of Bagging. When the predictions from the sub-models are uncorrelated or at best weakly correlated, the combination of predictions from multiple models result in the better working of ensembles. Random Forest modifies the algorithm to make the constituent sub-trees learn and have less correlation with each other.

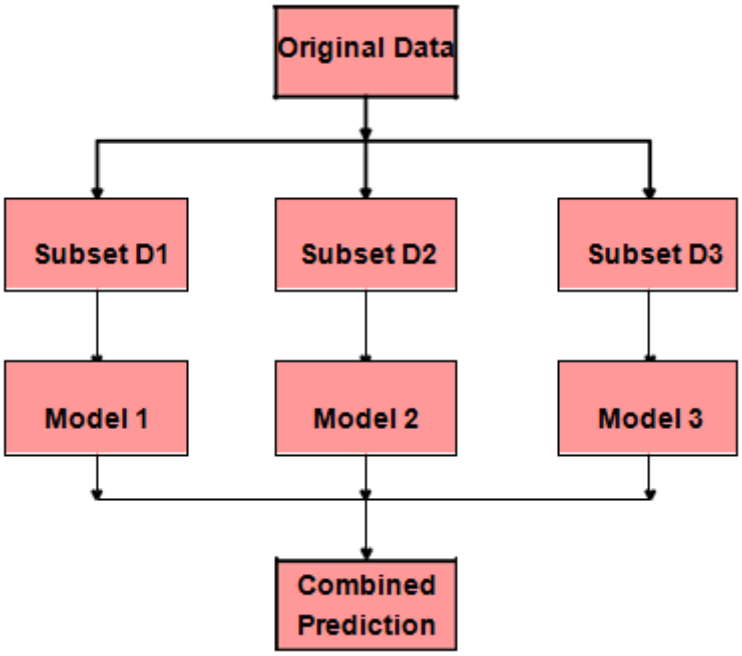

Fig 1:- Bagging Model Representation

C. Boosting

Boosting is a sequential process in which each subsequent model tries to correct the previous model's errors. The models that follow depend on the previous model.

Here, the models are defined as Learner and the prediction is done by all the models. Models with better accuracy are classified as Strong Learner while the Models with lower accuracy are classified as Weak Learner.

It combines weak learner's outputs and creates a strong learner that ultimately improves the model's accuracy of prediction. Boosting pays more attention to examples that by preceding weak rules are miss-classified or have higher errors. In this paper implementation of Boosting is done by applying XGBOOST [8] regression algorithm.

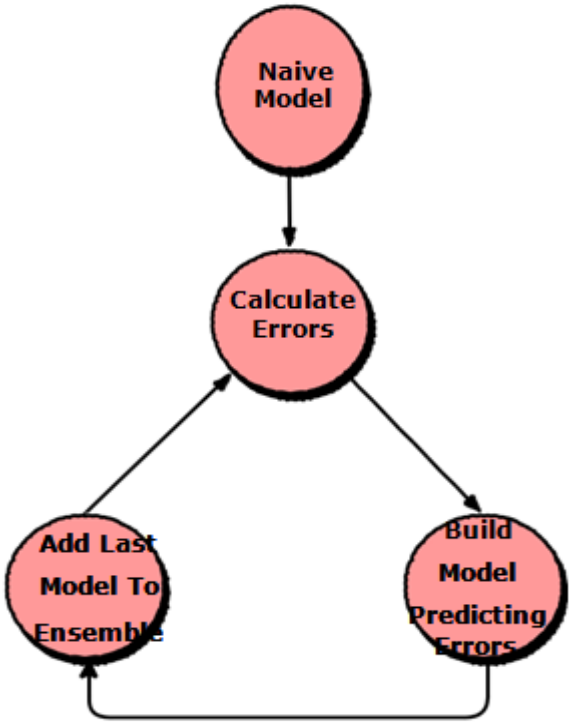

Fig 2:- XGBOOST Model Representation

D. Root Mean Square Error(RMSE)

$$RMSE_{f0} = \left[\sum_{i=1}^{N} (z_{fi} - z_{oi})^2 = N \right]^{1=2}$$

where
N = sample size.
z_{fi} is the predicted value
z_{oi} is the actual value

E. Support Vector Regression

In SVR [9] we try to fit the error within a certain threshold. Here the Red Line is the HYPER PLANE. It is the line with maximum number of points present on it.

We are now creating two error lines at + e and -e distance from our HYPER PLANE respectively.

SVR plane adjusts this value of [e].SVR calculates the minimum value of [e] for which maximum points come in between these two error lines which give degree of tolerance.

We will take only the points which are in between these two error lines.

Only those points that have the least error rate will be taken.

Thus resulting in a better fitting model.

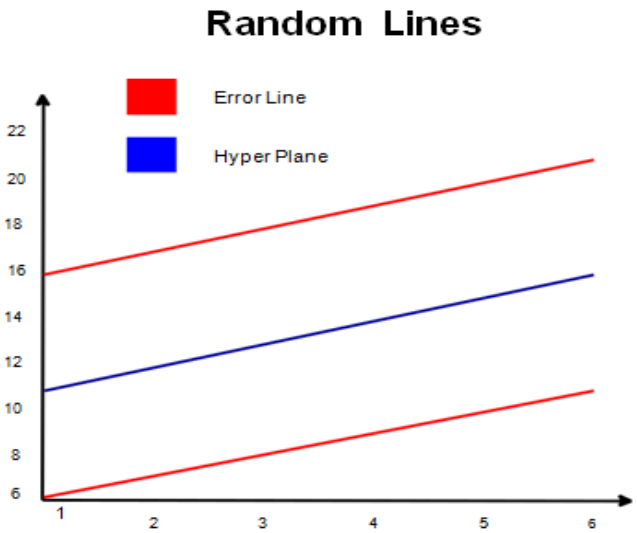

Fig 3:- Blue Line: Hyper Plane; Red Line: Error Line

F. Extrapolation

Estimating values of an unknown function f(x) in certain intervals [a, b] for specific arguments x when a number of observed values f(x) are available within that interval. If in any case you want to estimate f(x) value when x is outside [a, b], then the problem is known as extrapolation.

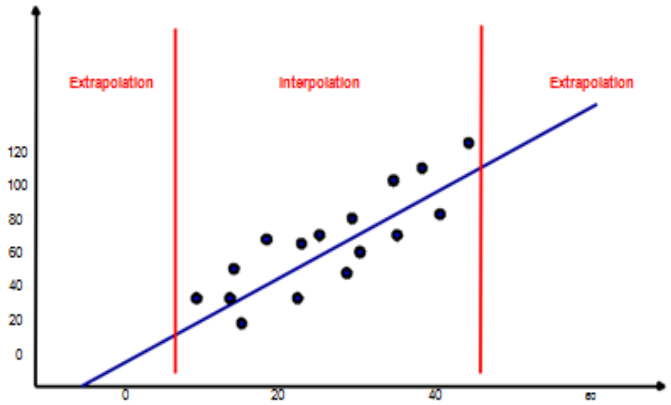

Fig 4:- Blue Dots: Dataset values; Red Line: Range[a,b]

III. IMPLEMENTATION

To implement our approach in regard to accuracy com-parison the Auto MPG dataset from UCI Machine Learning repository is used and for explaning Extrapolation concept and use of Ensemble Method [10] in it Annual Rainfall dataset is used from data.gov.in .

A. Dataset Description

MPG dataset is used for calculating the accuracy of Ensem-ble methods. Number of Instances are 398
Number of Attributes are 9 (includes the class attribute) We calculated the mpg value (Dependent variable) based on other Independent variable. Five models are applied to calculate the value of these dependent variable.

- Linear Regression Model
- Random forest Model
- XGBOOST regression Model
- Linear regression model stacked with Random Forest
- Linear regression model stacked with XGBOOST.

RMSE value of all these models are calculated to estimate their accuracy.

To check and highlight the problem of extrapolation in the case of tree based models(for eg. Random Forest) Annual Rainfall [05] dataset is used. Number of Instances: 12457

Number of Attributes: 20

Two models are applied to check the extrapolation problem in case of Bagging methods

- Only Random Forest models
- SVR Ensembled with Random Forest

ISSN No:-2456-2165

IV. OBSERVATION

A. Accuracy of Ensembled models

Graphs are plotted along with RMSE values for each of our approaches.

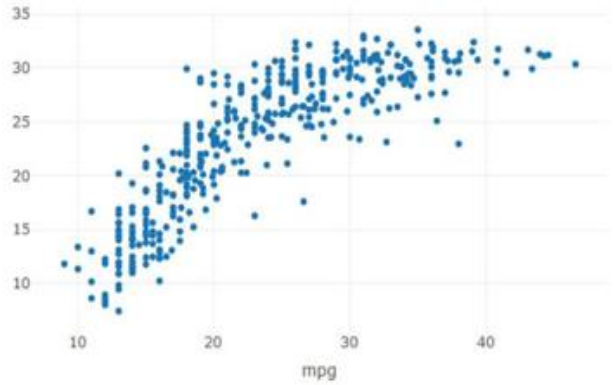

Fig 5:- Actual MPG Values vs Predicted MPG Values
Rmse Values: 4.269

- Linear Regression Model:
- Random Forest Regression:
- XGBOOST Regression:
- Linear Regression stacked with Random Forest:
- Linear Regression stacked with XGBOOST:

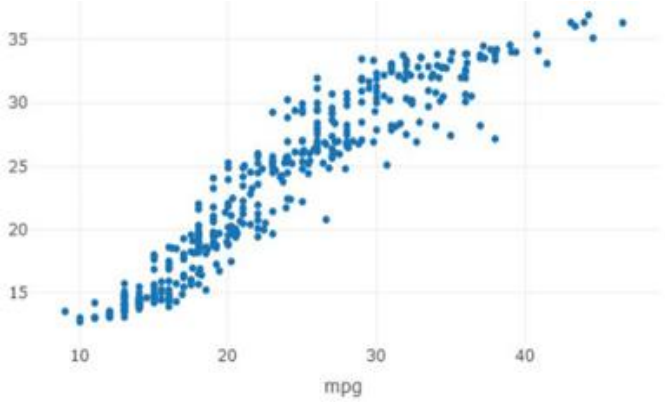

Fig 6:- Actual MPG Values vs Predicted MPG Values
Rmse Values: 2.624

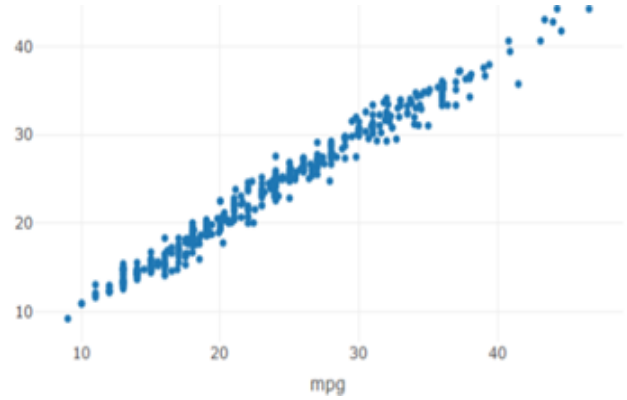

Fig 7:- Actual MPG Values vs Predicted MPG Values
Rmse Values: 1.185

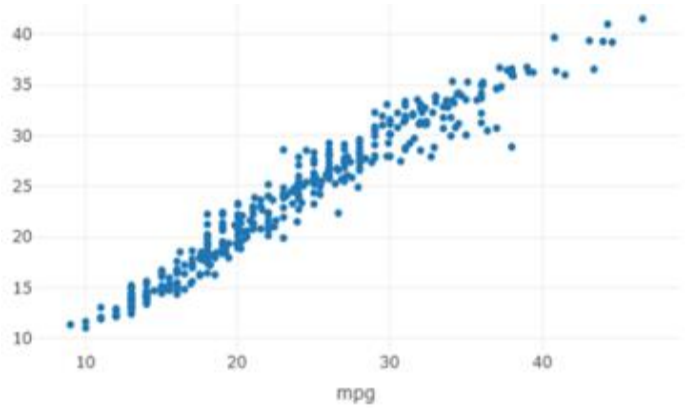

Fig 8:- Actual MPG Values vs Predicted MPG Values
Rmse Values: 1.778

B. Extrapolation of Random Forest versus Ensemble model

We have data from 1900 to 1980 for the rainfall. So here the range [a,b] is [1900,1980]. Years 1980-2015 are out of bound and not part of Data range for Training the model.

- Random Forest is applied to Annual Rainfall dataset.

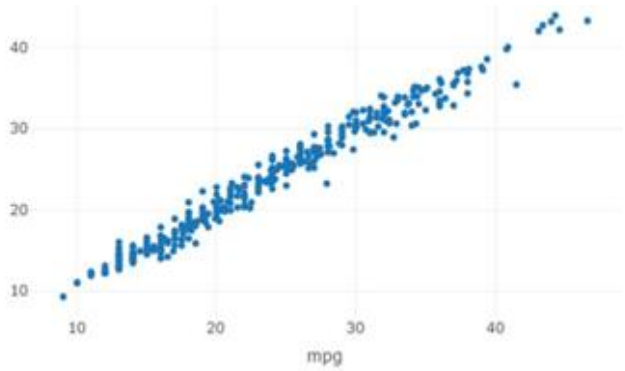

Fig 9:- Actual MPG Values vs Predicted MPG Values
Rmse Values: 1.249

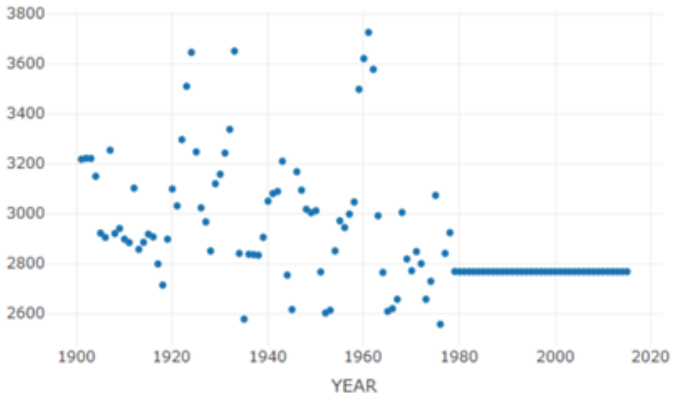

Fig 10

Observation clearly suggests that Extrapolation is not taken care by Random Forest and outside the Limit [a,b] the pre-dicted value becomes Constant. 2.Ensemble Method is applied to Annual Rainfall dataset. In this case

ISSN No:-2456-2165

Observation clearly suggests that Data outside of Limit [a,b] is also predicted by Ensemble Method .

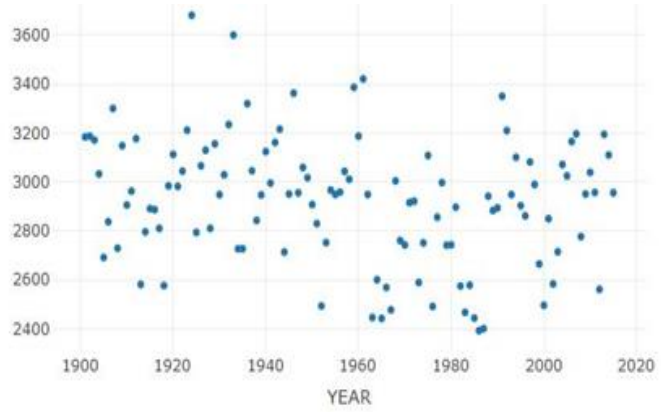

Fig 11

V. RESULTS

Accuracy order of various models observed are

- Bagging : Stacked Model>Individual Random forest>Linear Regression
- Boosting: Stacked Model~Individual XGBOOST>Linear Regression

Ensemble Methods also efficiently predicted the data Values out of bound [a,b].On the other hand,Random Forest failed to achieve any Logical Result out of the bound [a,b].

VI. CONCLUSION

Observations clearly indicate that in case of Random For-est stacked with Linear Model the accuracy is better than that of individual Random Forest Model.In case of XG-BOOST,stacking the model with Linear Regression has no profound effect on the Accuracy.Linear regression results were significantly improved when stacked with Bagging or Boosting methods.

Ensemble Method clearly improves the prediction capability of Random Forest in case of Extrapolated data.

REFERENCES

[1]. Zhi-Hua Zhou, “Ensemble Learning”, Encyclopedia of Biometrics, Springr, 2015.
[2]. Peter Buhlmanna, “Bagging, Boosting and Ensemble Methods”, Springer Handbooks of Computational Statistics, 2011.
[3]. Thomas G. Dietterich,”Ensemble Methods in Machine Learning”, In-ternational Workshop on Multiple Classifier Systems, Lecture Notes in Computer Science, vol. 1857, 2000.
[4]. T. Chai and R. R. Draxler,”Root mean square error (RMSE) or mean absolute error (MAE)?– Arguments against avoiding RMSE in the literature”, Geoscientific Model Development, vol. 7, issue 3, pp. 1247-1250, 2014.
[5]. Eesha Goel and Er. Abhilasha, “Random Forest: A Review”, IJARESSE, Volume 7, Issue 1, 2017.
[6]. Aleksandar Petrovski, Silvana Petruseva and Valentina Zileska Pancov-ska, “Multiple Linear regression model for predicting bidding price”, Technics Technologies Education Management, vol. 10, issue 3, pp. 386-393, 2015.
[7]. Zhengye Chen,”The Application of Tree-based model to Unbalanced German Credit Data Analysis”, MATEC Web of Conferences, 2018.
[8]. Tianqi Chen, Carlos Guestrin, ”XGBoost: A Scalable Tree Boosting System”, Machine Learning, Cornell University, 2016.
[9]. Durgesh Srivastava and Lekha Bhambhu, ”Data Classification using Sup-port Vector Machine”, Journal of Theoretical and Applied Information Technology, vol. 12, issue 1, pp. 1-7, 2010.
[10]. Chun-Xia Zhang and Jiang-She Zhang ,”A Survey of Selective Ensemble Learning Algorithms”, Chinese Journal of Computer, vol. 34, issue 8, pp. 1399-1410, 2011.