ISSN: 2278-4632
Vol-11 Issue-01 2021

COMPARATIVE ANALYSIS OF RANDOM FOREST AND LINEAR REGRESSION FOR PREDICTION OF TEMPERATURE

Dr.K.Chitra Assistant Professor, New Horizon College, Bangalore :: chitranatarajan1979@yahoo.in

Abstract

Linear regression is one of the well understood algorithms in statistics and machine learning. Linear regression is used to model the relationship between two variables by fitting a linear equation. One of the variable is an explanatory variable, and the other variable is a dependent variable. A scatterplot is used to represent the relationship between two variables. The line equation for linear regression is Y=a+bx. Random forest consists of a large number of decision trees that operate as an ensemble. Each individual tree in the random forest finds out a prediction and the class with the best results becomes the model's prediction. Random forest is an ensemble of decision trees. Many trees are constructed in a certain random way to form a Random Forest. Each and every tree is created from a different sample of rows, and different sample of features are selected for splitting at each node. Each of the trees makes its own individual prediction. The average of these predictions are used to produce a single result. In this research paper, the results of Random forest and Linear regression are compared using temperature dataset.

Keywords :

SSE ,SSTO,SSE, n_estimators ,max_depth, MSE,MAE,RMSE,R2 and Adjusted R2

I.INTRODUCTION

There are two types of learning techniques in machine learning model – Supervised and Unsupervised learning. There are two types under supervised learning – Regression and classification. There are two types of Regression – Linear Regression and Logistic Regression. There are two types of classification techniques – Decision trees and Random Forest. Linear regression is commonly used type of predictive analysis. Random Forest is a supervised machine learning algorithm made up of decision trees; Random Forest is used for both classification and regression. In this research, the comparison of Random Forest Model with Linear regression has been done using temperature dataset

II.LITERATURE SURVEY

This section presents some of the existing research works related to Linear Regression and Random Forest. Yanjun Qi **[1] proposed** the Random Forest (RF) technique, which includes an ensemble of decision trees and incorporates feature selection and interactions naturally in the learning process, is a popular choice. It is nonparametric, interpretable, efficient, and has high prediction accuracy for many types of data. Chen, X., Liu, M [2] proposed domain-based models for protein interaction prediction and preliminary results have demonstrated their feasibility. Protein interactions are

ISSN: 2278-4632
Vol-11 Issue-01 2021

of biological interest because they orchestrate a number of cellular processes such as metabolic pathways and immunological recognition. Domains are the building blocks of proteins; therefore, proteins are assumed to interact as a result of their interacting domains. domain-based models for protein interaction prediction have been developed, and preliminary results have demonstrated their feasibility.

Murat Kayri et al.,[3] compared Multiple Linear Regression, Random Forest, and Artificial Neural Network methods. For comparison of these data mining techniques, the power production data from a Photovoltaic Module was used in the research. In this study, the model was constituted from seven variables. One of the variables is dependent (power) and the others are independent variables (global radiation, temperature, wind speed, wind direction, relative humidity, solar elevation angle). In this paper, the Mean Absolute Error and the correlation coefficient were used in order to compare the estimation performance of the mentioned data mining techniques. Haoyuan Hong et al.[4] compared the results of Logistic Regression and a Random Forest model for the construction of a landslide susceptibility map in the Wuyuan area, China. Thirteen landslide variables were analyzed, namely: lithology, soil, slope, aspect, altitude, topographic wetness index, stream power index, stream transport index, plan curvature, profile curvature, distance to roads, distance to rivers and distance to faults, while 255 sites classified as landslide and 255 sites classified as non-landslide were separated into a training dataset (70%) and a validation dataset (30%). The comparison and validation of the outcomes of each model were achieved using statistical evaluation measures, the receiving operating characteristic and the area under the success and prediction rate curves.

Daniel O. McInerney and Maarten Nieuwenhuis [5] tested two non-parametric estimation techniques in two study areas in Ireland. For each area, plot level estimates of standing volume per hectare and basal area per hectare were computed from the National Forest Inventory field data and combined with SPOT 4 XS satellite imagery and a digital elevation model to form a set of observations. These observations were then used to predict variables across the satellite image using *k*-Nearest Neighbour (*k*NN) estimation and a Random Forest algorithm. Comparisons between the two techniques were assessed based on the estimation errors primarily using the Root Mean Square Error (RMSE) and relative mean deviation (bias). In both study areas it was found that the RMSE was lower for *k*NN than for RF. **Hooman Latifi** & Barbara Koch [6] proposed the models based on spectral and 3D information extracted from airborne optical and laser scanner data. The survey was completed across two geographically adjacent temperate forest sites in southwestern Germany, using spatially and temporally comparable remote-sensing data collected by similar instruments. Samples from the auxiliary reference stands (called off-site samples) were combined with random, random stratified and systematically stratified samples from the target area for prediction of standing volume, above-ground biomass and stem count in the target area. A range of combinations was used for the modelling process, comprising the most similar neighbor (MSN) and random forest (RF) imputation methods, three sampling designs and two predictor subset sizes.

Jianxin Wu [7] proposed algorithms for large-scale support vector machines (SVM) classification and other tasks using additive kernels. First, a linear regression SVM framework for general nonlinear kernel

ISSN: 2278-4632
Vol-11 Issue-01 2021

is proposed using linear regression to approximate gradient computations in the learning process. Second, they proposed a power mean SVM (PmSVM) algorithm for all additive kernels using nonsymmetric explanatory variable functions. Ming Yuan et al, [8] introduced a general formulation for dimension reduction and coefficient estimation in the multivariate linear model. The method proposed can be formulated as a novel penalized least squares estimate. The penalty that they employed is the coefficient matrix's Ky Fan norm. Such a penalty encourages the sparsity among singular values and at the same time gives shrinkage coefficient estimates and thus conducts dimension reduction and coefficient estimation simultaneously in the multivariate linear model**.**

III. LINEAR REGRESSION MODEL

Linear regression is used for finding linear relationship between target variable and one or more input variables .There are two types of Linear regression models – Simple Linear Regression , Multiple linear regression.

Simple linear regression is used to find relationship between two continuous variables. One is independent variable and another one is dependent variable. **Multiple linear regression** is a model that is used to predict the value of one dependent variable based on two or more independent variables.

Metrics for model evaluation in Linear Regression

1.R-Squared value

This value ranges from 0 to 1. Value '1' indicates predictor accounts for all the variation in Y. Value '0' indicates that predictor 'x' accounts for no variation in 'y'.

2. Regression sum of squares (SSR)

This parameter gives information about how far estimated regression line is from the horizontal 'no relationship' line.

3. Sum of Squared error (SSE)

$$\text{Error} = \sum_{i=1}^{n} (\text{Predicted\_output} - \text{average\_of\_actual\_output})\text{^}2$$

How much the target value varies around the regression line (predicted value).

$$\text{Error} = \sum_{i=1}^{n} (\textit{Actual\_output} - \textit{predicted\_output})\text{^}2$$

4. Total sum of squares (SSTO)

This parameter is used to tell how much the data point move around the mean.

ISSN: 2278-4632
Vol-11 Issue-01 2021

$$\text{Error} = \sum_{i=1}^{n}(\text{Actual\_output} - \text{average\_of\_actual\_output})^2$$

$$R^2 = 1 - (SSE/SSTO)$$

The following are different heuristics **for different preparation of data in Linear Regression**

1. **Linear Assumption**. The relationship between input and output variable should be linear If the relationship is not linear, we may need to transform the data to make the relationship linear.
2. **Remove Noise**. Data cleaning is necessary and the outliers in the output variable should be removed.
3. **Remove Collinearity**. If highly correlated input variables are there, then the Linear regression will over-fit for the data. We should remove the most correlated input variables.
4. **Rescale Inputs**: If the input variables are rescaled using standardization or normalization, then Linear regression will give more reliable predictions

IV.Random Forest Model

Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble.. Random forest is a supervised learning algorithm. The "forest" it builds, is an collection of decision trees, usually trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result. Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.

Hyper-parameters of Random Forest

1. **n_estimators**
   A random forest has multiple trees and It is possible to set the number of trees in random forest.
2. **max_features**
   “**max_features**” is used to select the number of features at each node.

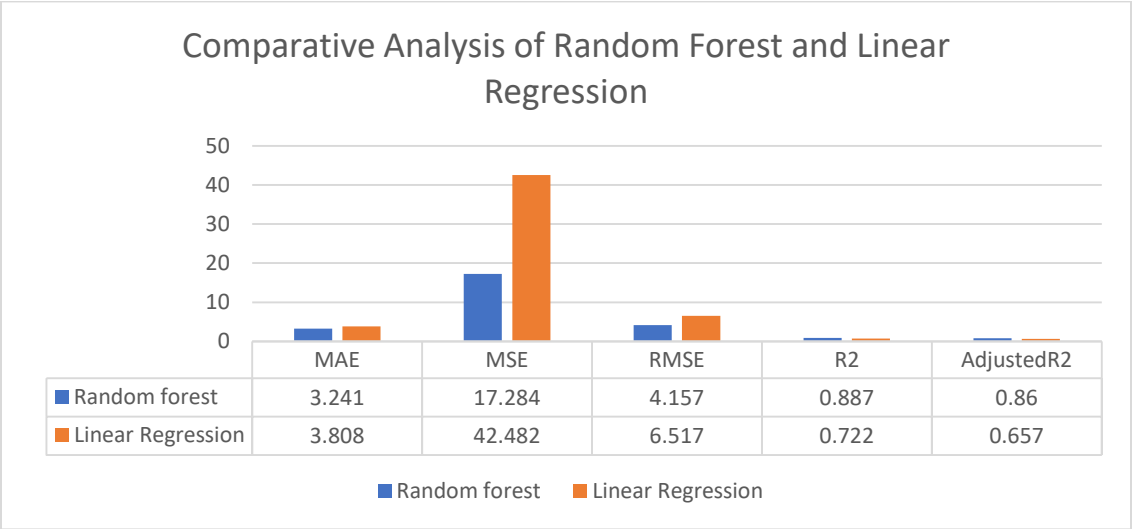
3. **max_depth**
   It is used to set the number of levels in tree
4. **min_samples_split**
   It is used to set the minimum number of samples required to split a node and the minimum number of samples at the leaf node.
5. **min_samples_leaf**
   It is used to set number of samples in leaf node.

IV.PERFORMANCE ANALYSIS

ISSN: 2278-4632
Vol-11 Issue-01 2021

This section presents the performance analysis of Random Forest algorithm and Linear regression algorithm using temperature dataset. The metrics like Mean Squared Error(MSE), Mean Absolute Error(MAE), Root Mean Squared Error(RMSE) , R2 and Adjusted R2 are considered for performance analysis. The main objectives of the study is to compare the error rates in Random Forest Model and Linear Regression Model using temperature dataset. Eight variables were considered, namely: day, month, year, Week, temp1, temp2, average, actual. The dataset were separated into a testing dataset (70%) and a validation dataset (30%). The temperature is predicted in the dataset. The comparison of Random Forest Model and Linear Regression model using the metrics MAE,MSE,RMSE,R2,Adjusted R2 is depicted below:

Model /Metrics	MAE	MSE	RMSE	R2	AdjustedR2
Random forest	3.241	17.284	4.157	0.887	0.86
Linear Regression	3.808	42.482	6.517	0.722	0.657



V.CONCLUSION

A random forest is a meta estimator that collects a number of decision tree classifiers on various sub-samples of the dataset and used to improve the predictive accuracy and It controls over-fitting. Linear regression is used to model the relationship between continuous variables.

The main objective of this research work is to compare the performance of Random forest and Linear regression by using some of the metrics like MSE,MAE,RMSE,R2 and Adjusted R2 in temperature dataset. The error rates in random forest is less comparing to Linear regression in the temperature dataset. The Random forest model will be more effective than Liner regression in prediction of temperature in this temperature dataset.

VI.REFERENCES

[1] Yanjun Qi **,"**Random Forest for Bioinformatics**",** Ensemble Machine Learning pp 307-323,0 F**irst Online:** 19 January 2012.

[2] Chen, X., Liu, M: Prediction of protein–protein interactions using random decision forest framework. Bioinformatics **21**(24), 4394 (2005).

[3] Murat Kayri et al.," The performance comparison of Multiple Linear Regression, Random Forest and Artificial Neural Network by using photovoltaic and atmospheric data", 14th International Conference on Engineering of Modern Electric Systems (EMES),1-2 June 2017.

[4] Haoyuan Hong et al.,"Comparing the Performance of a Logistic Regression and a Random Forest Model in Landslide Susceptibility Assessments. the Case of Wuyaun Area, China ", Workshop on World Landslide Forum WLF 11 June 2017: Advancing Culture of Living with Landslides pp 1043-1050.

[5] Daniel O. McInerney and Maarten Nieuwenhuis "A comparative analysis of *k*NN and decision tree methods for the Irish National Forest Inventory",Pages 4937-4955 | Published online: 22 Sep 2009.

[6] **Hooman Latifi** & Barbara Koch, "Evaluation of most similar neighbour and random forest methods for imputing forest inventory variables using data from target and auxiliary stands"
Pages 6668-6694 | Received 16 Feb 2011, Accepted 17 Feb 2012, Published online: 06 Jun 2012,https://doi.org/10.1080/01431161.2012.693969.

[7] Jianxin Wu,Linear Regression-Based Efficient SVM Learning for Large-Scale Classification, IEEE Transactions on Neural Networks and Learning Systems , Volume: 26, Issue: 10, Oct. 2015),**Page(s):** 2357 – 2369, 06 January 2015.

[8] Ming Yuan et al, "Dimension reduction and coefficient estimation in multivariate linear regression 22 May 2007,**https://doi.org/10.1111/j.1467-9868.2007.00591.x**

