International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 05 | May 2020 www.irjet.net p-ISSN: 2395-0072

Comparative Study and Proposed Approach for Multi-Variate Regression through Gradient Boosting

Siddharth Kekre

B.Tech. Computer Science and Engineering, Student: Medi-Caps University. India.

Abstract - *It is a well-known fact that E-Commerce platforms have a very tough competition and online retailers need to set the price of their goods carefully to maximize profits and their value. In this paper, an in-depth analysis of a regression model is shown, which is very close to the original algorithm. Results of this study show difference between the standard approach and a proposed approach for Regression problems. A model has been proposed in which machine learning is incorporated, it can increase profits for the online retailers by predicting the price they should sell their item for based on previous data so that the company can take the required action or anticipate the financials beforehand.*

Key Words: **Machine Learning, Price Prediction, Random Forest, Gradient Boosting Regression, Optimising Algorithm**

1. INTRODUCTION

One of the key goals of online retailers is to boost their sales while increase the desirability of their products. To achieve this, they opt for certain techniques such as offering discounts to the customers. Some of the key aspects associated with these kinds of promotions are offering discounts at the right time and as for the losses that ideally should be none but considering a more pragmatic approach, should be as low as possible while increasing customer base and boosting sales. Online shopping in India is an exponentially growing market. In fact, it is expected to a reach a sum of over 52 billion USD by 2022 [1]. With a market like this, predicting the most optimum price for the product is very important.

Through recent studies in disciplines such as Machine Learning and Statistics; researchers have proven that machine learning methods could predict more accurate diagnosis provide suggestions as compared to traditional statistical methods. A great prediction model will empower many online retailers to determine the most optimum price for the item, maximize their profits and boost seller reputation on e-commerce platforms.

This work is focused on designing such a desired prediction model. There is a pool of Machine Learning techniques and models out there to choose from, provided in this work is a comparative study of some of the most effective algorithms, namely Linear Regression, Random Forest Regression and Gradient Boosting Regression. Also, in this study a study is provided that tells which algorithm should be tuned in what way to obtain most optimum results. All conclusions are supported by obtained results.

2. DATASET

The Dataset used for performing analysis was obtained from a Machine Learning Competition [2] organized by HackerEarth from May 01, 2020 to May 21, 2020. The task was to predict price of Gift Items based on certain factors.

3. ALGORITHMS USED

There are three main techniques for Machine Learning, namely Supervised, Unsupervised and Reinforcement Machine Learning [3]. Adapting these techniques for the problem statement depends majorly upon on the type of dataset and operation to be performed. Upon careful study, the problem statement discussed in this finding falls under the principles of Supervised Machine Learning and to be precise is a case of Regression [4].

3.1 Linear Regression

Linear Regression is a common method to mostly used when value of a variable is to be predicted through values of one or more variables. One important condition is to be considered that these variables have a simple linear dependency among them that can be easily represented in the form as simple as $Y = a + bX$ where X is the explanatory variable (collection of dependent variables) and Y is the dependent variable (which we are trying to predict) [5].

3.2 Random Forest Regression

A random forest is a collection of randomized base regression trees that grow in randomly selected subspaces of data [6]. These random trees are combined to form one optimum prediction tree. Defining a random state in the regressor makes sure same trees are generated each time the algorithm is run [7]. This helps reduce ambiguity and provides more reliable results.

3.3 Gradient Boosting Regression

Like Random Forest, Gradient Boosting is a set of Decision Trees. In Boosting, each new tree (Tree 2) is a fit on a modified version of the original data set (Tree 1). Here, the idea is to improve upon the predictions of the first tree. Our new model is therefore Tree 1 + Tree 2 [8]. We then compute further from this new 2-tree model and grow a third tree to predict the revised residuals. Predictions of the final

ensemble model is therefore the weighted sum of the predictions made by the previous tree models.

4. METHODOLOGY

4.1 Standard Approach

- Pre-Processing
  - Handling Null Values (One of Following)
    - Dropping Null Values
    - Replacing Null Values with Mean
    - Replacing Null Values with Median
  - Handling Outliers (One of the Following)
    - Dropping Outlier Data
    - Leaving Outliers Untouched
    - Normalising Outlier using Z-Score
- Selecting Feature Columns
- Split the Data [10] for Training and Testing the Model (Usually in 70:30 ratio for Training and Testing respectively, no specific Random State for splitting is considered)
- Apply standard regression algorithms (Linear Regression [11] / Random Forest Regression [7] / Gradient Boosting Regression [12])
- Optimize the Model to increase Testing Accuracy.
- Prediction Results Obtained using R2 Scores [13]. (refer Table-1)

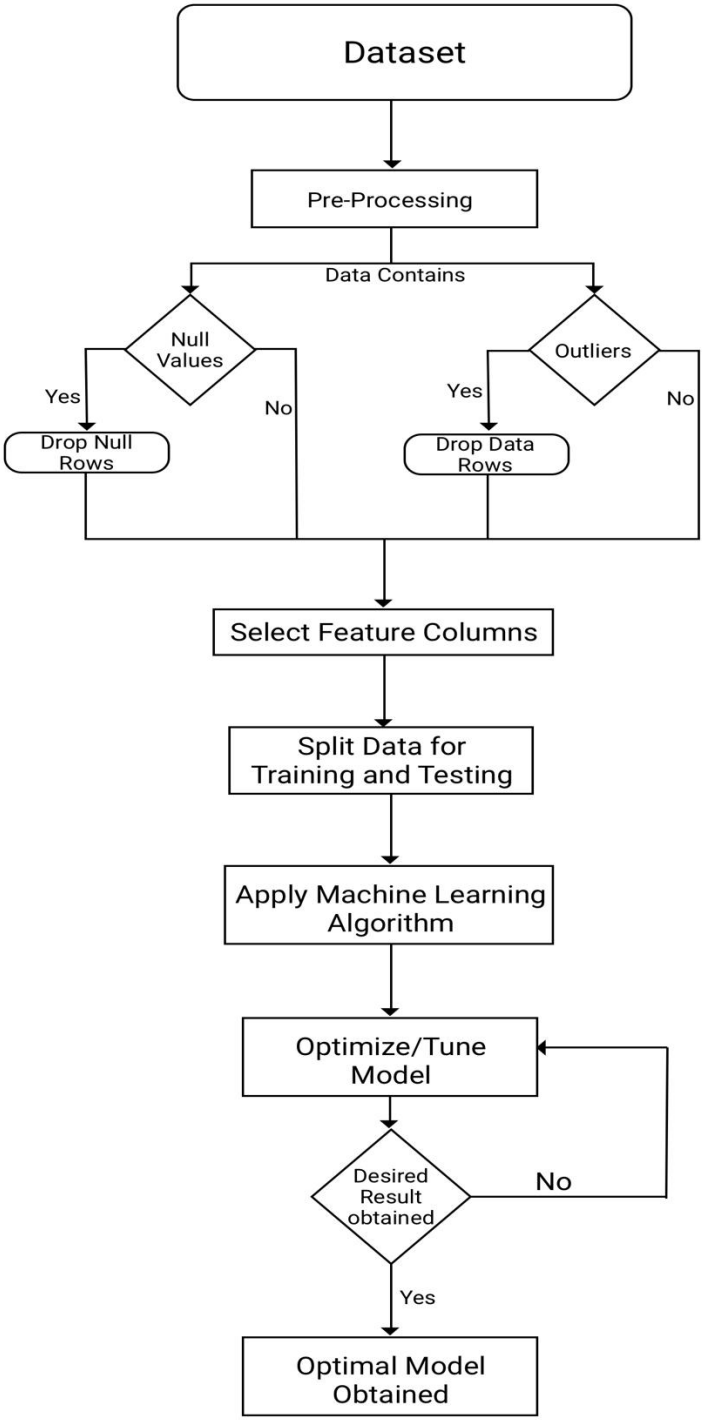

Fig-1: Standard Approach

4.2 Proposed Approach

- Pre-Processing

  (Predict the Null Values and Imputing them in Dataset)
  - Target Variable is a Feature.

- Outliers are Normalized using Z-Score.
- Data is split [10] into 70:30 ratio for Training and Testing respectively with Random State set to '0' (Integer Zero).
- Apply Gradient Boosting Regression [12]. (refer Table III for the reason)
- Model is optimized to increase Testing Accuracy. (refer Table-2)
- Prediction Results Obtained using R2 Scores [13] (refer Table-3)
- Predictions are imputed in Original Data Set.

- Selecting Feature Columns
- Split the Data [10] in 70:30 ratio for Training and Testing respectively with Random State set to '1'.
- Apply Gradient Boosting Regression [12]. (refer Table-4 for the reason)
- Optimize the Model to increase Testing Accuracy. (refer Table-5)
- Prediction Results Obtained using R2 Scores [13]. (refer Table-6)

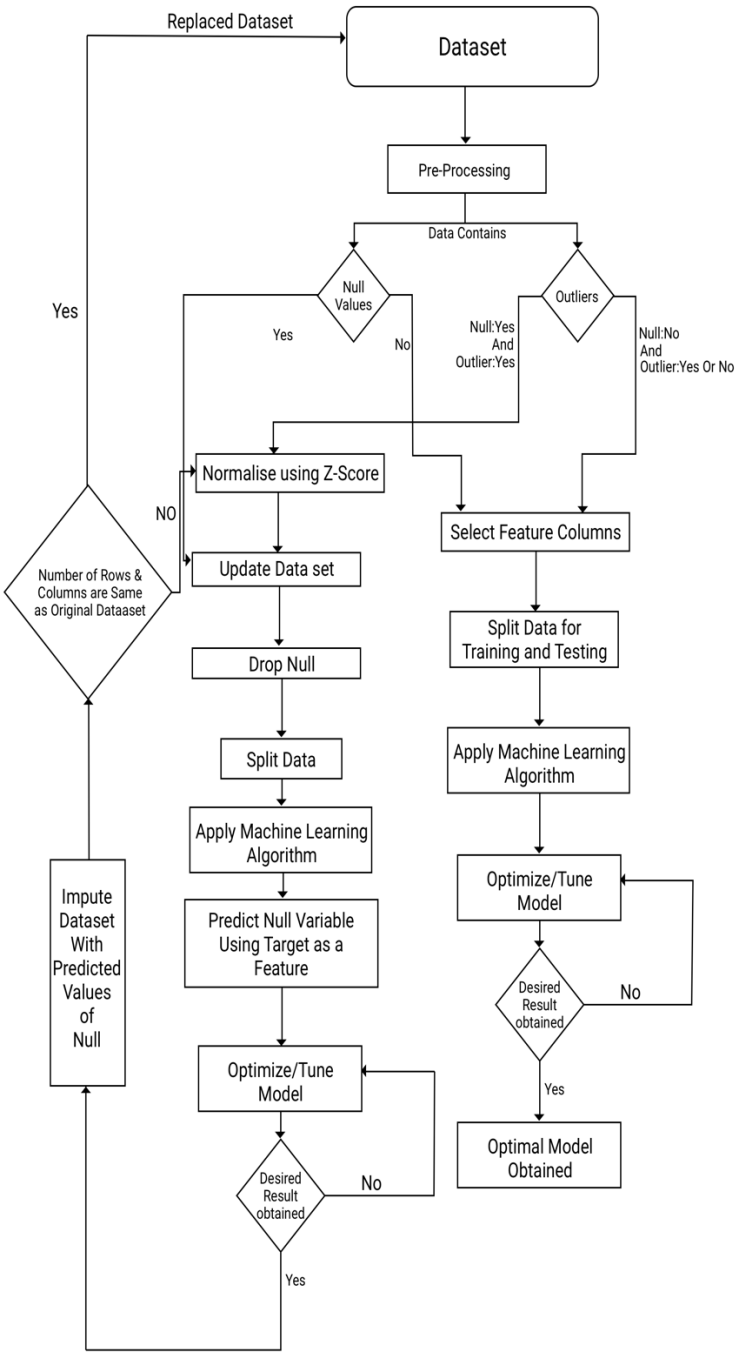

Fig-2: Proposed Approach

5. MODELS AND RESULTS

5.1 Standard Approach

TABLE-1: TARGET PREDICTION FROM DATASET

Testing Accuracy (%)				
Outliers	Method	Null Dropped	Null Replaced with Mean	Null Replaced with Median
Untouched	Linear Regression	14.16	35.388	35.933
	Random Forest Regressor	28.917	51.856	51.856
	Gradient Boosting Regressor	67.07	75.864	74.148
Data Rows Dropped	Linear Regression	14.16	35.388	35.933
	Random Forest Regressor	28.917	51.856	51.856
	Gradient Boosting Regressor	75.204	74.311	74.218

5.2 Proposed Approach

TABLE-2: OPTIMIZE MODEL TO PREDICT NULL VALUES

Parameter	*Default Value*	*Set Value*
n_estimators	100	120
random_state	None	2
learning_rate	0.2	0.378
max_depth	3	5

TABLE-3: PREDICTING NULL VALUES

Training Accuracy (%)				
Outliers	Method	Null Dropped	Null Replaced with Mean	Null Replaced with Median
Untouched	Linear Regression	11.819	3.162	8.452
	Random Forest Regressor	13.819	5.162	10.452
	Gradient Boosting Regressor	93.628	82.404	85.233
Normalized using Z-Score	Linear Regression	18.819	10.162	15.452
	Random Forest Regressor	76.065	71.856	71.856
	Gradient Boosting Regressor	96.628	85.404	88.233

When comparing all suitable possible combinations of Pre-Processing the data that include handling Null Values and Outliers, and to choose an algorithm from three prominent algorithms to predict Null Values, the above results support the Proposed Approach.

TABLE-4: PREDICTING TARGET VARIABLE AFTER NULL VALUES HAVE BEEN PREDICTED (WITHOUT OPTIMIZATION)

Testing Accuracy (%)		
Method	Outliers Untouched	Outliers Normalized
Linear Regression	40.671	40.671
Random Forest Regressor	55.151	55.151

Gradient Boosting Regressor	80.352	80.445

When comparing all suitable possible combinations of predicting the target variable after null values have been predicted and imputed in the dataset, the above results support the Proposed Approach.

TABLE-5: OPTIMIZE MODEL TO PREDICT TARGET AFTER NULL VALUES HAVE BEEN PREDICTED

Parameter	*Default Value*	*Set Value*
n_estimators	100	135
random_state	None	18
learning_rate	0.2	0.359
max_depth	3	3

TABLE-6: TARGET PREDICTION AFTER MODEL IS OPTIMIZED

Testing Accuracy (%)		
Algorithm	Standard Approach	Proposed Approach
Gradient Boosting Regression	80.445	94.041

6. CONCLUSIONS

The Proposed Approach presents Testing Accuracy of **94.041%** whereas the Standard Approach provides Testing Accuracy of **80.445%** which is a significant increase by **16.90%**.

After analyzing all the results, it is clear that the Proposed Approach performs significantly better than the Traditional Approach and hence, can be applied for seminal real-life scenarios.

REFERENCES

[1] Akhil Zacharia, Arjun P C, Gokul Vilson, Soumya Varma, "Sales Prediction for Online Shopping", International Journal of Innovative Science and Research Technology, Volume 4, Issue 5, May – 2019.

[2] HackerEarth, "Machine Learning Challange Predict Price Good Friday Gifts", May 01, 2020 - May 20, 2020.

[3] https://towardsdatascience.com/what-are-the-types-of-machine-learning-e2b9e5d1756f

[4] Gulden Kaya Uyanik, Nese Guler, "A study on Multiple Learneer Regression Analysis". Elsevier, Procedia-Social and Behavioral Sciences 106(2013) 234–240.

[5] http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm

[6] Gerard Biau, "Analysis of Random Forest Models", Journal of Machine Learning Research 13 (2012), 1063-1095.

[7] "3.2.4.3.2. sklearn.ensemble.RandomForestRegressor", https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

[8] https://towardsdatascience.com/understanding-gradient-boosting-machines-9be756fe76ab

[9] https://intellipaat.com/blog/what-is-linear-regression/

[10] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html

[11] "sklearn.linear_model.LinearRegression", https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

[12] "3.2.4.3.6. sklearn.ensemble.GradientBoostingRegressor", https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html

[13] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html

[14] Sally Jo Cunningham and Geoffrey Holmes, "Developing innovative applications in agriculture using data mining". Department of Computer Science University of Waikato Hamilton, New Zealand.