iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** I **Month of publication:** January 2023

DOI: https://doi.org/10.22214/ijraset.2023.48773

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue I Jan 2023- Available at www.ijraset.com

Visualization and Prediction of the Company's Revenue Using Machine Learning and Data Analysis

Vibhor Harit[1], Shubham Yadav[2], Saurabh Prakash[3], Sumit Kumar[4], Taksh Chanana[5], Vikrant[6]
[1]*Assistant Professor,* [2, 3, 4, 5, 6]*Student, Department of Computer Science and Engineering ,IMS Engineering College Ghaziabad , UttarPradesh* – 201015

Abstract: ***Revenue decides the future of the company and the growth in upcoming years so it is vital for companies to know about it so that any changes in time can be made to make the business profitable. Most business organization largely depends on their sales, demand, and sales trend for their growth. This paper analyzes and predicts company sales and demands using machine learning and data analysis. The dataset Ocean Cafe available on Kaggle has been used for analysis and prediction purposes. The model used for prediction is XGBoost in which R-square, MAE(Mean Absolute Error), MSE(Mean Squared Error), and RMSE(Root Mean Squared Error) are calculated.***
For visualization purposes, various pie-chart, line graphs, and bar graphs are drawn making use of the python libraries and functions.
Keywords: Data analysis, Mean squared error, XGBoost, Prediction, Machine learning, Visualization, Graphs, cross validation, gradient boosting.

I. INTRODUCTION

One of the major Objectives of this research work is to find out the requirements and basicity of the company and to predict the revenue using machine learning algorithms to achieve the best accuracy. Today's business has been so compact and dealing with such a large set of data is a herculean task. The volume of data is expected to grow exponentially in the further year. Prediction gives a way to achieve the best possible need of the company. Revenue forecasting and prediction give insight into how a company should manage its workforce and resources.

Paper aims to provide a feasible solution to small business corporates to look after their business demand and sales and to give shape to business analytics that is hard to contextualize via big numbers. Nowadays small corporates have to face many issues related to the business they have close their business due to the unavailability of good guidance and visualizations. this provides the correct way to utilize things in the right direction. this gives an overall view of all products being sold per day and gives a correct analysis of the products and their revenue

The prediction model will help to analyze the relationship among various attributes of the dataset used in the modeling. The dataset extracted from Kaggle (Ocean Cafe) is verified data and it is used for training and prediction. It consists of 100000+ entries specifically datewise and categorized data which is available online.

The prediction model built will provide a prediction based on the different attributes available in the dataset i.e. Date, BillNumber, Itemdesc, Time, Quantity, Rate, Tax, Discount, Total, etc. The prediction model is implemented based on the XGBoost Machine learning algorithm it usually calculates MAE(Mean Absolute Error), MSE(Mean Squared Error), RMSE(Root Mean Squared Error), and R-Square values to evaluate the accuracy of the model.

The paper further walks through various sections. Section I gives an introduction to the problem, Section II illustrates the prior research done in this field, and Section III contains the data set description and methodology of the research. Section IV provides the Results and analysis of the various values of the model and the last section contains the conclusion.

II. LITERATURE SURVEY

During the initial phase of our project, we face so many difficulties related to the execution of the data set and other factors like accuracy so we go through another similar project which is already made on the same problem. During the literature survey, we find other more efficient technologies which gave good results for our project.

Amruta Aher, Dr.K.Rajeswari, and Prof.Shushma [1] have proposed a prediction model to analyze the customer's past spending and predict the future spending of the customer. The dataset referred to is Black Friday Sales Dataset which is available on Kaggle. It is used for analysis and prediction purposes. They have machine learning models such as Linear Regression, Ridge Regression, Lasso Regression, Decision Regression, and Random Forest. The performance evaluation measure Root Mean Squared Error (RMSE) is used to evaluate the models used. The MSE rate of the Random Forest Regressor is 3062.72 and hence it is more suitable for the prediction model to be implemented.

Ramraj Sa Nishant Uzirb Sunil Rb and ShatadeepBanerjeeb [2] have proposed a Prediction and Classification of Different Datasets. they used different data sets NASA dataset available on NASA's official site handles the dataset name Airfoil Self-Noise Dataset, and another dataset is Banknote Authentication Dataset which is used for Data extracted from images that were taken from real and forged specimens. They used algorithms like the AdaBoost algorithm was then modified into a set of other statistical algorithms called ARCing algorithms. These algorithms used a process called "arcing". Arcing is an acronym for Adaptive Reweighting and Combining. Then came a new algorithm called eXtreme Gradient Boosting (XGBoost) which changed the way gradient
boosting was done. In XGBoost, individual trees are created using multiple cores, and data is organized to minimize the lookup times. Then, in the case of GB, the mean squared error is 7.96. In the case of the banknote authentication dataset, for XGBoost, the accuracy is 98.45% in the case of the training and testing set evaluation method.

Tianqi Chen and Carlos Guestrin [3] have proposed a scalable end-to-end tree-boosting system called XGBoost. They have used various algorithms to train models like SPLIT FINDING ALGORITHMS, Approximate Algorithms, Exact Greedy Algorithms for Split Finding, etc. The exact greedy algorithm is very powerful since it enumerates all possible splitting points greedily. However, it is impossible to efficiently do so when the data does not fit entirely into memory. they used four datasets Allstate, Higgs Boson, Event classification Yahoo LTRC and Criteo. The first dataset we use is the Allstate insurance claim dataset. The task is to predict the likelihood and cost of an insurance claim given different risk factors. They learned when building XGBoost, a scalable tree-boosting system that is widely used by data scientists and provides state-of-the-art results on many problems. These lessons can be applied to other machine-learning systems as well. By combining these insights, XGBoost can solve real-world scale problems using a minimal amount of resources.

Lawrence O Hall, Nitesh Chawla and Kevin W Bowy [4] proposed a method Combining Decision Trees Learned in Parallel .it mainly focuses on very large dataset sets that may be utilized for visualization To focus attention on the salient regions of the data set being visualized it is useful to have information on the interesting regions of data. the main thing they did in the project set our goal to have a single decision system after learning is done independently on n disjoint subsets of data The independent learners can be viewed as agents learning a little about a domain with the knowledge of each agent to be combined into one knowledge base. Towards this end, the independent decision trees might be combined into a single decision tree. They used two data sets The Iris data (Fisher 1936 Merz Murphy) which has continuous-valued attributes and classifies examples as one of the classes of Iris plant and The second dataset the Pima Indians Diabetes data set
(Merz& Murphy) which has 8 numeric attributes and classifies 768 examples into one of 2 classes. The classification accuracy when generating rules from the unpruned and pruned trees for the 2 processor simulation with the Iris data is shown on the first row of results and compared with the accuracy when one decision tree is generated from each fold. The accuracy is slightly better than that of the decision trees for both the pruned and unpruned trees On this data set the pruned and unpruned rules are the same.

Odegua Rising [5] proposed a method called Applied Machine Learning for Supermarket Sales Prediction . where he explore that The goal of every supermarket is to make a profit. This is achieved when more goods are sold and the turnover is high. A major challenge to increasing sales in a supermarket lies in the ability of the manager to forecast sales patterns and know readily beforehand when to order and replenish inventories as well as plan for manpower and staff. The amount of sales data has steadily been on the increase in recent years and the ability to leverage this gold of data separates high-performing supermarkets from others.to find a better solution he uses various algorithms the first one is K-Nearest Neighbour Algorithm. The idea behind KNN is that given a sample of instances in a sample space, a new instance is similar if it belongs to the same class as the already existing sample. The idea is to first select k nearest neighbor to the sample whose class we want to predict. In that sense, KNN does not need any training and is seen as a memorization-based technique. KNNs are good and fast for small data sets but become less efficient when the data set increases. And the second algorithm he used was Gradient Boosting, Boosting is a popular machine-learning algorithm that falls under the umbrella of ensembles. Boosting was introduced in answer to the question of whether a “weak learner” could be made better by using some form of modification. This was discovered to be possible and the first boosting algorithm Adaptive Boosting (AdaBoost) was created.

The concept of boosting is to correct the mistakes made by previous learners and improve in various areas. the data set he used was provided by Data Science Nigeria, a Data Science, and Artificial Intelligence Hub as part of their machine learning competitions. The data consist of numerous supermarket variables like the opening year, product prices, supermarket location, etc.
The data set contains a sample of 4990 instances with 13 features/variables. Taking the average prediction of the 10-fold cross-validation, we observe that the Random Forest algorithm does best among all three with an MAE of 0.409178. The Gradient Boosting model has a close MAE to the KNN but with a much lower standard deviation.

III. RESEARCH METHODOLOGY

A. Data Collection

Kaggle has more than 50 thousand public datasets available on its website. For this research, the publicly available dataset from Kaggle was used. The public dataset of the beachside restaurant has sales transactions with 1,45,830 entries. The dataset has 10 attributes which are bill number, item description, quantity, rate, tax, discount, total, category, and date & time when the item was ordered.

B. Data Preprocessing

The dataset received from Kaggle was not suitable for the machine learning algorithm because it has redundant entries, missing values, and unnecessary data. These redundant entries and unnecessary data like bill numbers were removed and the dataset size was reduced significantly. Missing values were filled by taking the mean of the column or row. Working Machine learning algorithm is based on mathematics (or numbers). So, the categorical data like the item's description, date, time, and item's category are needed to be encoded into numbers.
Item category such as FOOD, BEVERAGE, TOBACCO, and LIQUOR was converted to numerical values.

C. Data Splitting

Training our model with a dataset and testing it with another dataset will result in difficulty in understanding the correlation between features by our machine learning model and this will cause a decrease in performance.
So, to train and test the machine learning model with aim of getting good performance (or better predictability), it is required to divide the dataset into two parts in a standard ratio of 70:30. 70% of the dataset will be used to train the machine learning model. 30% of the dataset will be used to test our machine learning model which our model predicts.

D. Training

An effective open-source implementation of Gradient Boosting is XG Boost which is scalable and highly accurate. It is selected because it is designed to be fast and provide good performance. XG Boost also predicts with very high accuracy as compared to other implementations of Gradient Boosting. The training dataset was applied to the XG Boost to get the model trained.

E. Prediction And Calculation Of Error Values

The testing dataset was applied to the trained model to predict future sales data. Predicted future sales data is used for evaluation of the accuracy of our model. Error values such as Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R Squared are calculated.
By averaging all observations, Mean Absolute Error (MAE) calculates the absolute distance between the observations (its dataset entries) and the regression predictions. How closely a regression line resembles a set of data points is determined by the Mean Squared Error (MSE). The transformation between predicted values by our model and actual values is calculated by Root Mean Squared Error (RMSE). A statistical measure that represents the goodness of fit of a regression model is called R-Squared.
Calculated error values are mentioned in Table 3 and 4.

Table 1: Total monthly sales

Month	Total Sales
January	2436182.46
February	2204094.16
March	2494197.98

April	2602584.93
May	2942256.00
June	2483845.63
July	2698502.20
August	2649848.84
September	3212965.05
October	3339323.65
November	2810639.79
December	2779920.62

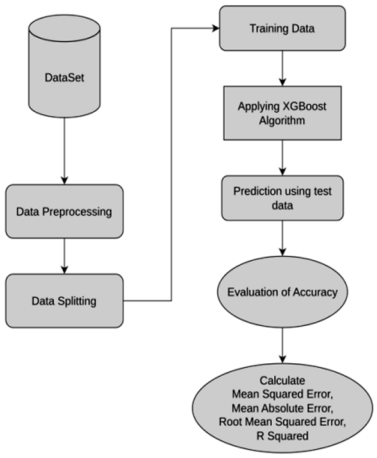


Figure 1. Flowchart of the proposed system

IV. RESULTS AND ANALYSIS

The goal of our project is to predict and analyze the company's Revenue using Machine learning and Data analysis.
To achieve the project's goal we have used Python programming language and its packages and for the prediction, we have used XGBoost(eXtreme Gradient Boosting) machine learning algorithm. Firstly, we collected the dataset from Kaggle and then we prepare the dataset for modeling.

A. Dataset Analysis

The dimension of the dataset is (145830, 10).

Table 2: Features of dataset

Sr. NO.	Features
1	Data
2	Bill Number
3	Item Desc
4	Time
5	Quantity
6	Rate
7	Tax
8	Discount
9	Total
10	Category

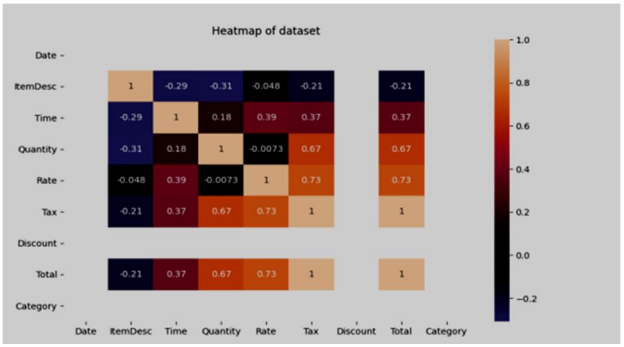


Figure-2. Heat Map

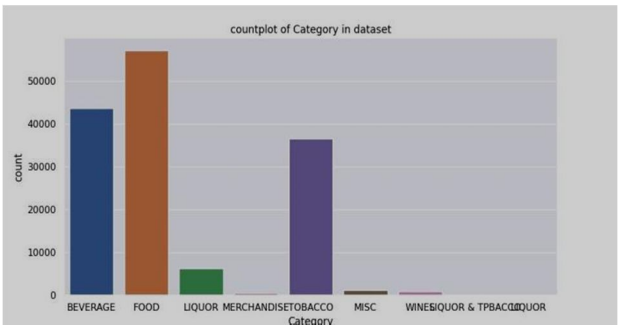


Figure-3. Category distribution plot

Numerical features in the dataset are Rate, Tax, Discount, Total, and Quantity. Categorical features in the dataset are Date, Bill Number, Item Desc, Time, and Category. In the preprocessing step we find the missing value and fill it with accurate value, then performed label encoding before using the dataset for the model. After that, we have done a data analysis of our dataset and found a whole insight into the dataset through graphs, plots, etc. To achieve this we used Python libraries seaborn, and matplotlib.

B. Prediction Model Analysis

In this project, for the predictive analysis, we have used an *XGBoost* regressor. This regressor learns from the dataset and finds the pattern in data after that it predicts the result based on learning/patterns.

First, we figured out the target variable/feature in our dataset and also the features that affect the target variable. According to this, we split our dataset into training and testing parts in the ratio of 70%-30%, after that we train the model, and then we used the model for the prediction

C. XGBoost

XGBoost(eXtreme Gradient Boosting) is a gradient boosting framework that is scalable, portable, and works as parallel tree boosting. It uses regularisation to avoid overfitting in the model, and parallel tree building and uses a depth-first approach for tree pruning.

Table 3: Training performance metrics

MODEL	R-square (R^2)	MAE	MSE	RMSE
XGBoost	0.9999	0.1377	0.2472	0.4972

Table 4: Testing performance metrics

MODEL	R-square (R^2)	MAE	MSE	RMSE
XGBoost	0.9985	0.2608	37.6146	6.130

V. FUTURE SCOPE

We have used XGBoost to achieve the project's goal that performing well on model evaluation metrics and finding out accurate values for the attributes.

Research area in machine learning is growing day by day which is why there are lots of alternatives for XGBoost hence about time, in the future, we have a model that performs well than XGBoost in the field of prediction that provides more accurate results.

VI. CONCLUSION

The ultimate goal of this project is to predict and analyze the company's Revenue using Machine learning and Data analysis to achieve this goal we have used Python language and its module for dataset analysis and prediction of revenue by regressor.

Based on model evaluation metrics, Xgboost performed well and provided the accurate result of metrics that are mentioned in Table 3 and Table 4 at the time of training and testing.

Hence we have a model that learns the pattern from the dataset and uses those patterns to find the revenue of the company. After this company has a whole insight into their product by graph and plot, and by using a model, the company is capable of finding their profit/loss in the financial calendar of the company. they have all theinsight to predict the revenue in the upcoming quarter period of the company. According to the result, they can take important decisions to maximize the growth of the company.

As the field of machine learning is growing day by day, there is a lot of models available for prediction purpose example:

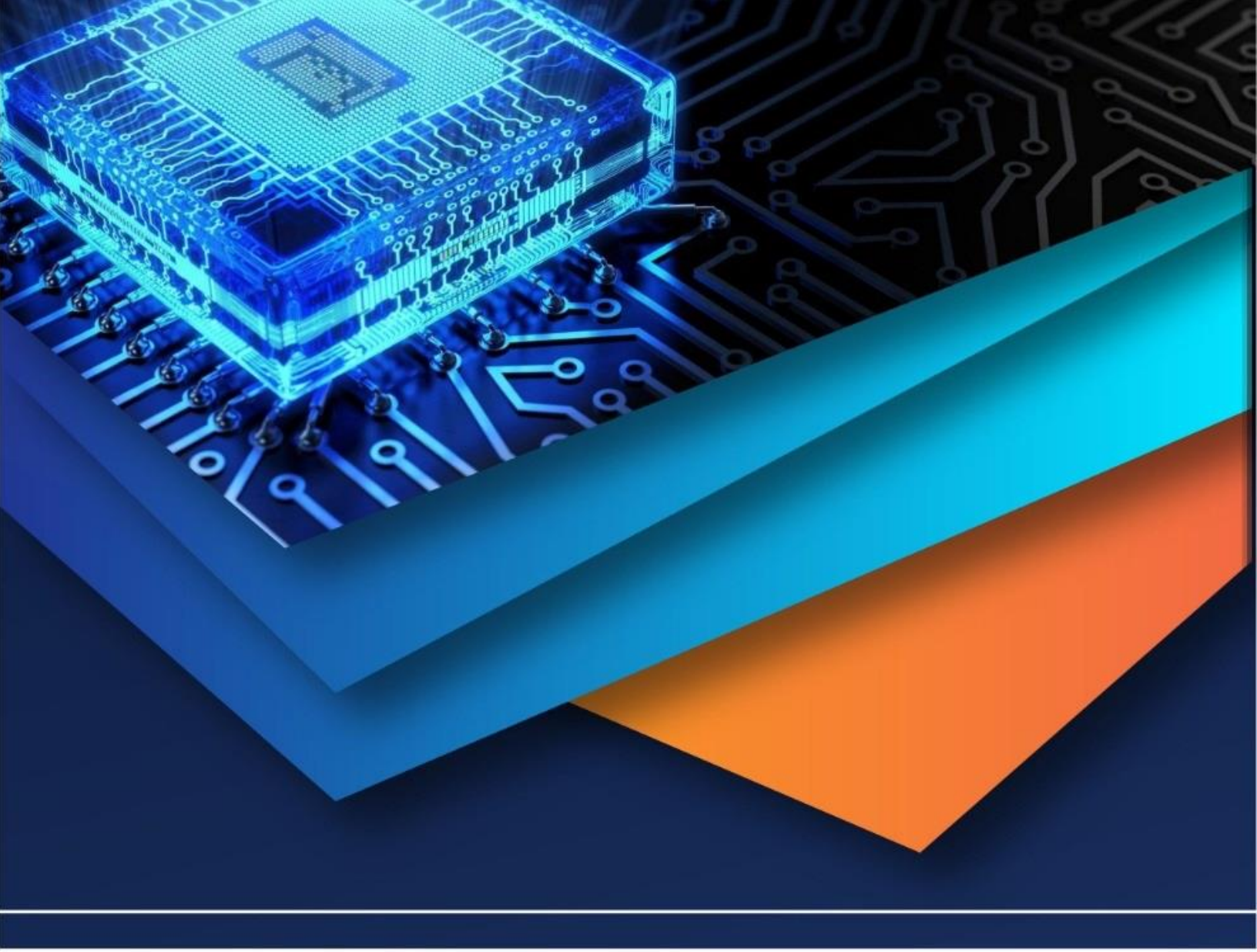
1) Random Forest
2) Gradient Boosting
3) XGBoost
4) Prophet

All of above the model is generally used for predictive analysis. We have used XGBoost in our project for the prediction of revenue. Xgboost has more advantages than other models it uses the concept of creating a decision tree parallelly, it uses regularization for removing overfitting problems. It can handle missing values in the dataset and it has inbuilt cross-validation.

REFERENCES

[1] Amruta Aher , Rajeswari Kannan , Sushma Vispute, 2021, Data Analysis and Price Prediction of Black Friday Sales using Machine Learning Techniques, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 10, Issue 07 (July 2021)
[2] Santhanam, Ramraj & Uzir, Nishant & Raman, Sunil & Banerjee, Shatadeep. (2017).Experimenting XGBoost Algorithm for Prediction and Classification of Different Datasets.
[3] Chen, Tianqi & Guestrin, Carlos. (2016). XGBoost: A Scalable Tree Boosting System. 785-794. 10.1145/2939672.2939785. Tree boosting is a highly effective and widely used machine learning method. In this paper, we describe a scalable end-to-end tree boosting system called XGBoost.
[4] Lawrence O Hall, Nitesh Chawla and Kevin W Bowy Combining Decision Trees Learned in Parallel ,Distributed Data Mining Workshop at International Conference of Knowledge Discovery and Data Mining ,1998
[5] Odegua Rising. (2020). Applied Machine Learning for Supermarket Sales Prediction.
[6] C. M. Wu, P. Patil and S. Gunaseelan, "Comparison of Different Machine Learning Algorithms for Multiple Regression on Black Friday Sales Data," 2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS), 2018, pp. 16-20, doi: 10.1109/ICSESS.2018.8663760.
[7] Cross-Validation of a model https://aws.amazon.com/machine-learning/
[8] An End-to-End Guide to Understand the Math behindXGBoost https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost
[9] Linear regression analysis is used to predict the value of a variable based on the value of another variable https://www.ibm.com/in-en/topics/linear-regression#:~:text=Resources-,What%20is%20linear%20regression%3F,is%20called%20the%20independent%20variable
[10] Regression a statistical method https://www.investopedia.com/terms/r/regression.asp#:~:text=Investopedia%20%2F%20Joules%20Garcia-,What%20Is%20a%20Regression%3F,(known%20as%20independent%20variables)
[11] Diaz-Uriarte and S.A. de Andres. Gene selection and classification of microarray data using random forest. BMC Bioinformatics, 7:1471–2105, 2006.

[12] O. Chapelle and Y. Chang. Yahoo! Learning to Rank Challenge Overview. Journal of Machine Learning Research - W & CP, 14:1–24, 2011.
[13] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting. Annals of Statistics, 28(2):337–407, 2000.
[14] Nari Sivanandam Arunraj and Diane Ahrens. "A hybrid seasonal autoregressive integrated moving average and quantile regression for daily food sales forecasting". In: International Journal Production Economics 170 (2015), pp. 321–335.
[15] V. Svetnik, A. Liaw, C. Tong, J. Culberson, R. Sheridan, and B. Feuston. Random forest: A classification and regression tool for compound classification and QSAR modeling. Journal of Chemical Information and Computer Sciences, 43:1947–1958, 2003.
[16] The k-nearest neighbors (KNN) algorithm https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761
[17] Gradient boosting is a method https://www.analyticsvidhya.com/blog/2021/09/gradient-boosting-algorithm-a-complete-guide-for-beginners/
[18] Random forest is a Supervised Machine Learning Algorithm https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/
[19] Lasso regression is a type of linear regression https://www.statisticshowto.com/lasso-regression//
[20] National Institute of Diabetes and Digestive and Kidney Diseases , Pima Indians Diabetes Data Set, https://archive.ics.uci.edu/ml/support/diabetes
[21] Lasso and Ridge Regression https://www.datacamp.com/tutorial/tutorial-lasso-ridgeregression
[22] L1 and L2 Regularization Methods https://towardsdatascience.com/l1-andregularization-methods-ce25e7fc831c
[23] Decision Tree Regressor https://scikitlearn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html
[24] A Gentle Introduction to k-fold Cross-Validation https://machinelearningmastery.com/k-fold-cross-validation/
[25] What is Regression and Classification in Machine Learning https://www.geeksforgeeks.org/regression-classification-supervised-machine-learning/
[26] XGBoost stands for "Extreme Gradient Boosting" https://xgboost.readthedocs.io/en/stable/tutorials/model.html



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY