iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** XII **Month of publication:** December 2023

DOI: https://doi.org/10.22214/ijraset.2023.57625

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 XII Dec 2023- Available at www.ijraset.com

A Comprehensive Survey on Prediction Models and the Impact of XGBoost

Yashkumar Burnwal[1], Dr. R.C. Jaiswal[2]
[1]Student, [2]Associate Prof., Department of Electronics and Telecommunication Engineering, SCTR's Pune Institute of Computer Technology, Pune, India.

I. INTRODUCTION

Prediction models are essential to decision-making in many different fields, such as marketing, finance, healthcare, and sports. They are essential resources for deriving valuable insights from intricate information, supporting both experts and scholars in making rational decisions. With the goal of improving precision and efficiency, the Extreme Gradient Boosting (XGBoost) algorithm has become a prominent performer and attracted a lot of interest from the Data Science community. The aim of this survey paper is to provide a comprehensive analysis of widely used prediction models. Next, let's look at XGBoost, which has an outstanding performance record. The survey describes the basics of current prediction models and how XGBoost works in real-world scenarios. Since we are talking about predictive models in general and XGBoost, this survey is a useful tool to highlight the impact of XGBoost in this broad area.

II. PREDICTION MODELS

1) *Linear Regression:* A fundamental method for predictive modeling is linear regression. It establishes a linear connection between the target variable and the input features. It is widely used in many different sectors due to its interpretability resulting from coefficients indicating the impact of each attribute. However, complex, nonlinear patterns in data are difficult to capture for linear regression, limiting its applicability in some situations.
2) *Decision Trees:* Decision trees, a common predictive modeling technique, create a tree-like structure to make decisions based on input features. They are exceptionally good at recognizing non-linear relationships in data and ensuring interpretability through the use of open decision-making processes. Still, over-adjusting complexity in deep trees can result in overly complex models that underperform on fresh, untested data.
3) *Random Forest:* Random Forest is an approach that uses collaborative learning. To reduce overfitting and improve prediction accuracy, multiple decision trees are used. Although it is used extensively in many different areas, it can become computationally complex, especially as the number of trees in the ensemble increases. Despite this flaw, Random Forest is still a useful tool for predictive modeling because it offers better accuracy and stability.
4) *Support Vector Machines (SVM):* SVMs are great for prediction and classification because they can find the best hyperplanes to partition classes in difficult input spaces. However, there are problems with their computing power, especially when processing large amounts of data. Despite this disadvantage, SVMs are well suited to dealing with complex decision boundaries in many areas.

III. WHAT IS XGBOOST

XGBoost stands out among the many prediction models as a stable and adaptable algorithm that works well with structured tabular data. It is a fast and efficient implementation of gradient boosting decision trees. In XGBoost, a loss function, a regularization term and the sum of each decision tree are combined. The following is an expression for the overall objective function of the jth iteration:

$$Obj^{(j)} = \sum_{i=1}^{n} loss\left(y_i, \hat{y}_i^{(j-1)}\right) + \sum_{k=1}^{j} \Omega(f_x)$$

Here,

- $\Omega(f_k)$ the regularization term of the kth tree, penalizes complex models to avoid overfitting
- The sum of the loss function and regularization term is optimized during the training process.

The prediction of the final model is given by the sum of predictions from all individual trees:

$$\hat{y}^i = \sum_{k=1}^{K} f_x(x_i)$$

Where:

- Result is the predicted output for the ith instance.
- $f_k(x_i)$ is the prediction of the kth tree for the input features x_i.

IV. COMPARISON OF LINEAR REGRESSION, DECISION TREE, AND XGBOOST

It involves assessing their strengths and weaknesses across different criteria such as accuracy, interpretability, and computational efficiency. Below is a detailed comparison of these three algorithms:

Table 1: Comparison of Linear Regression, Decision Tree, and XGBoost

Criteria	Linear Regression	Decision Tree	XGBoost
Accuracy	The premise of linear regression is that the target variable and the input features have a linear relationship. If the underlying relationship is approximately linear, it works well. However, problems can arise when dealing with non-linear patterns in the data.	Compared to linear regression, decision trees are more versatile because they can capture nonlinear relationships in the data. However, they are prone to overfitting, especially when the trees become too complicated and too deep.	In terms of accuracy, XGBoost excels. It is possible to lessen overfitting and capture complex relationships by combining boosting and regularization strategies with a collection of decision trees. This is the recommended choice in many situations since it frequently outperforms linear regression and single decision trees.
Interpretability	Linear regression models are easy to interpret. The coefficients assigned to each characteristic provide information about the strength and direction of their influence on the target variable.	Decision trees provide interpretability to a certain extent. Users can understand the decision-making process thanks to the tree structure, although this interpretability may be limited with deep trees.	Although XGBoost provides feature importance values, it is less interpretable than linear regression. Interpreting individual tree contributions is challenging due to the complicated and all-encompassing nature of the boosting process.
Computational Efficiency	Linear regression is computationally efficient and works well for large data sets. The training process involves solving a closed-form equation, making it faster than iterative algorithms used by Decision Trees and XGBoost.	Training a decision tree can be computationally expensive, especially as the tree grows larger. However, once the tree is built, predictions are made quite quickly.	The main goal of XGBoost is efficiency. Compared to traditional gradient boosting implementations, shorter training times are achieved by using parallelization, regularization, and early stopping mechanisms. The ensemble structure also results in efficient predictions.
Handling Non-linearity	Because linear regression assumes a linear relationship, it may not be as effective at capturing nonlinear patterns.	Decision trees naturally handle nonlinear relationships and are therefore suitable for complex data structures. However, they tend to overfit.	The ensemble of decision trees in XGBoost is designed to efficiently capture nonlinear relationships. Regularization techniques are a reliable option for dealing with complex patterns as they help prevent overfitting.

V. KEY FEATURES OF XGBOOST

Regularization to prevent overfitting, gradient boosting for ensemble learning, hyperparameter tuning for optimization, feature importance metrics to improve interpretability, reliable handling of missing data, and effective parallelization for scalability are some of the key features of the algorithm.

1) .*Regularization and Control Overfitting:* XGBoost employs both L1 (Lasso) and L2 (Ridge) regularization in its objective function to penalize model complexity, mitigating overfitting. By preventing the model from overfitting the training set, these regularization strategies help the model become more general and perform better on untested data.
2) *Gradient Boosting and Ensemble Learning:* Gradient boosting is fundamental to XGBoost, sequentially combining weak learners (individual decision trees) to form a powerful ensemble model. Through the use of boosting, ensemble learning improves overall accuracy and predictability by leveraging the strengths of multiple models to compensate for individual weaknesses.
3) *Hyperparameter Tuning:* XGBoost offers various hyperparameters such as learning rate, tree depth and boosting rounds. Optimizing these hyperparameters is critical to maximizing model accuracy and generalization. The learning rate controls the step size during optimization, the tree depth influences the model complexity, and the number of boosting rounds determines the overall strength of the ensemble.
4) *Feature Importance and Interpretability:* XGBoost provides feature importance metrics to facilitate model interpretation. This data facilitates feature selection and improves understanding of the model's decision-making process. The frequency with which a feature appears in trees and the role it plays in partitioning decisions are factors that affect feature importance..
5) *Handling Missing Data:* The robustness of XGBoost also extends to dealing with missing data in training. The algorithm handles missing values well and can therefore be applied to real data sets with potentially incomplete information. This feature makes the algorithm more useful in different situations.
6) *Parallelization and Scalability:* XGBoost is characterized by scalability as it can easily process large amounts of data. To take advantage of multicore architectures and maximize computational efficiency, the algorithm is parallelized. XGBoost is therefore suitable for applications that process large amounts of data and where parallel processing significantly speeds up model training.

VI. IMPACT OF XGBOOST ON PREDICTION MODELS

1) *XGBoost's Performance Comparison:* XGBoost's impact on prediction models is profound, especially when compared to traditional models. Numerous studies have conducted head-to-head comparisons, demonstrating its superior performance in terms of predictive accuracy and generalization. The ensemble nature of XGBoost often outshines standalone models like Linear Regression and even other ensemble techniques like Random Forest.
2) *XGBoost in Real-world Applications:* Practical applications and case studies also underline the noticeable impact of XGBoost in various areas. In finance, it is used for credit scoring, fraud detection and portfolio optimization.

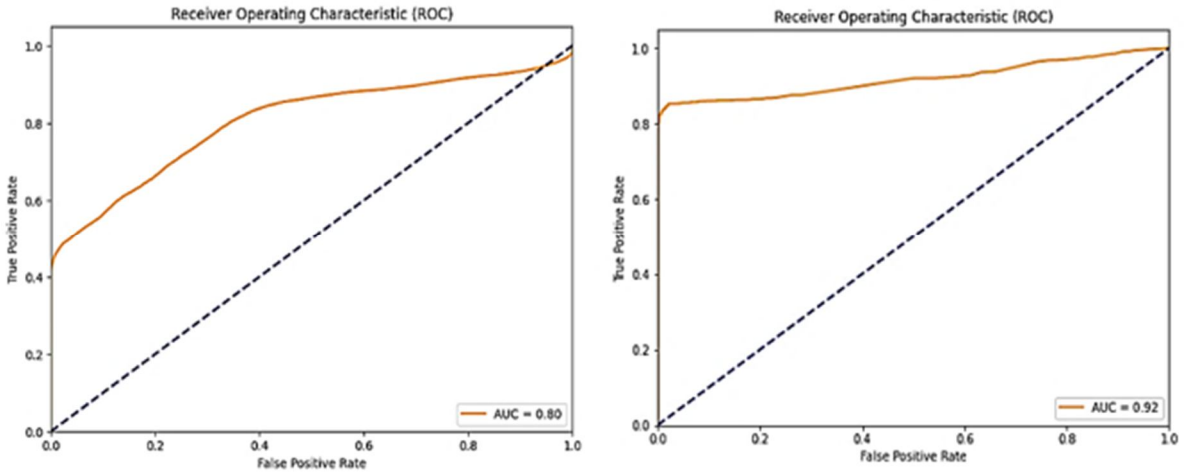


Fig. a. ROC curve for regression model (left) vs. XGBoost (right)

In Fig. a, is an example comparing the performance of XGBoost and regression model on a sample data used for portfolio optimization, compared to the regression model in portfolio optimization. The model predicts the likelihood of a customer taking out a personal loan based on customer demographics and various data from financial bureaus.

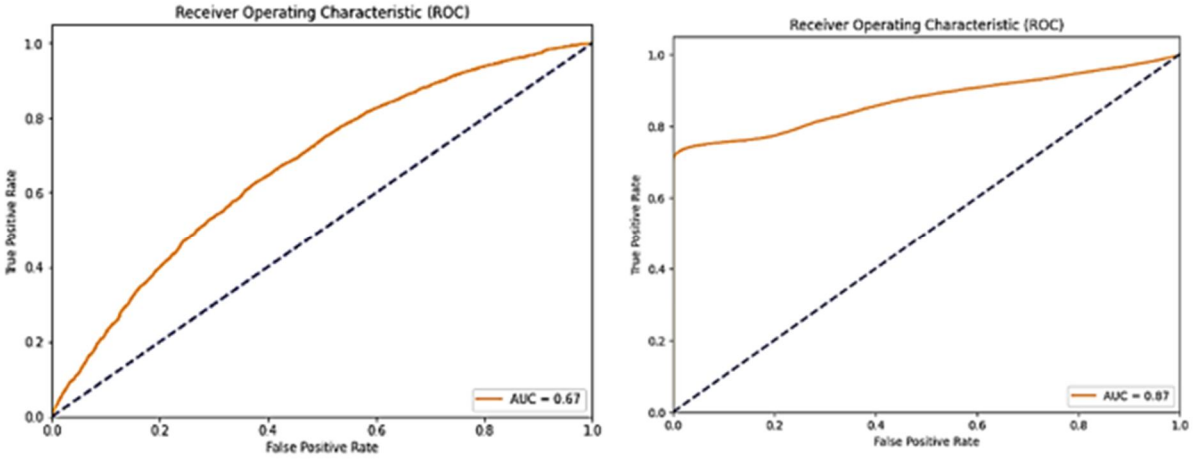


Fig b. ROC curve for random forest (left) Vs XGBoost (right)

Fig. b is another example where performance XGBoost and random forest is compared on a data set used to predict the likelihood of a customer filing a promotional complaint based on call center campaign data, customer demographics and recent financial products used. In both examples, XGBoost outperforms the regression model and random forest.

VII. XGBOOST'S CHALLENGES AND LIMITATIONS

Although XGBoost has shown incredible success, there are still challenges that needs to be overcome. A major disadvantage is that overfitting can occur, especially if hyperparameters are not adjusted properly. The process of hyperparameter tuning itself is difficult and requires skill to achieve the best results. Although XGBoost can handle missing data, it can be problematic when the data is very sparse or irregularly distributed.

VIII. FUTURE DIRECTIONS AND RESEARCH OPPORTUNITIES

1) *AutoML and Automated Hyperparameter Tuning:* There's a lot to discover about XGBoost's integration with AutoML (automated machine learning) frameworks. By using methods such as genetic algorithms or Bayesian optimization, automated hyperparameter optimization can simplify the model development process and increase the usability of XGBoost for users of different skill levels.
2) *Addressing Imbalance and Biases*: Ongoing research could refine XGBoost's mechanisms for dealing with imbalanced data and account for specific biases in different application areas. Improving the algorithm's robustness to imbalanced datasets will contribute to its applicability in areas where class imbalances are prevalent.

IX. CONCLUSION

This comprehensive survey examined the predictive modeling landscape and examined the impact of XGBoost on data science. Predictive modeling has evolved over time, moving from classical linear regression to ensemble techniques such as Random Forest and SVM, leading to the increasing popularity of XGBoost.

XGBoost's success is due to its strong architecture, effective management of structured table data, and a wealth of features that enhance its customizability. The algorithm's influence is visible in many areas, as it regularly outperforms traditional models and proves its effectiveness in practical applications.

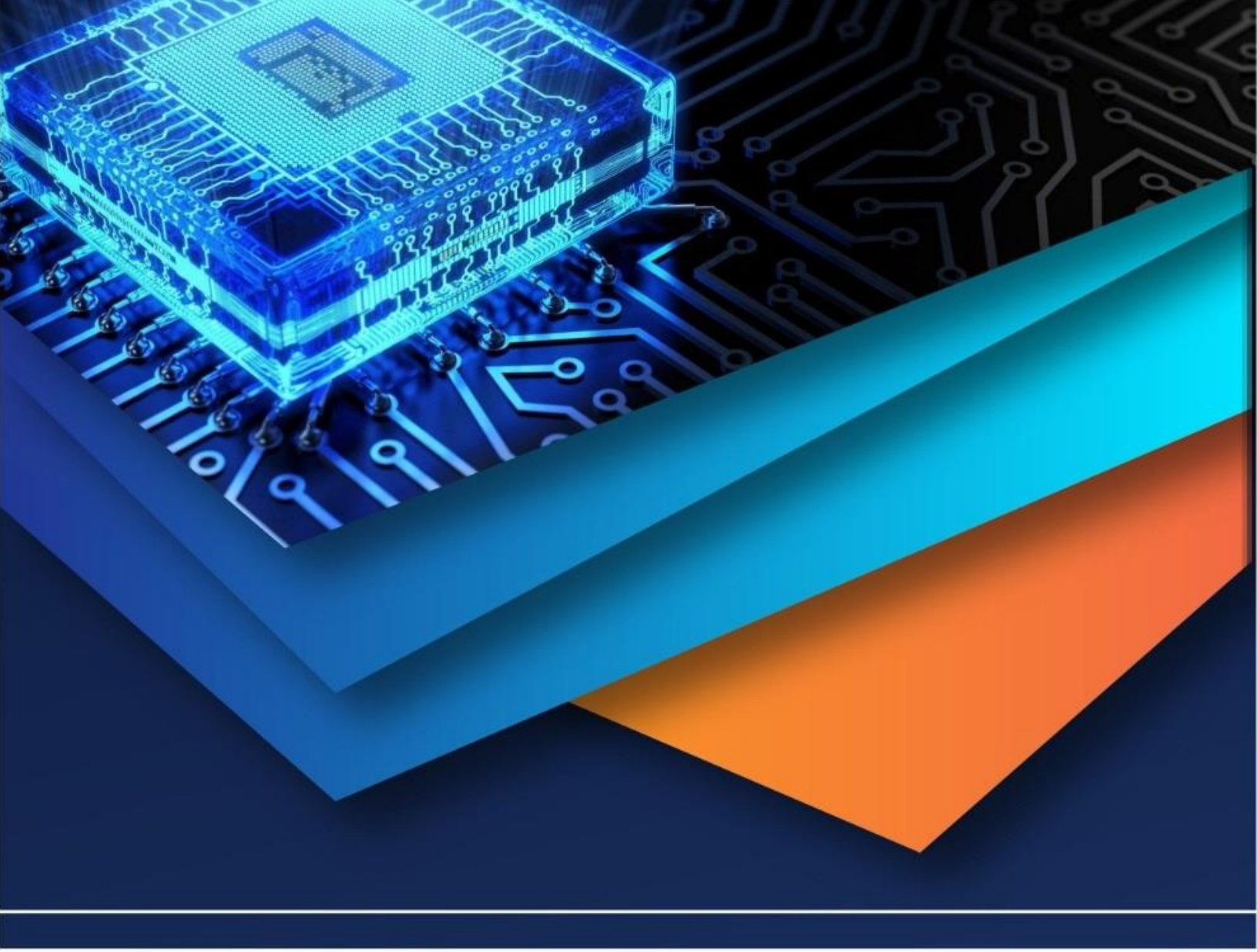
Future research and development will likely focus on improving XGBoost's interpretability, addressing specific issues, and smoothly integrating it into cutting-edge trends such as AutoML and hybrid modeling. XGBoost, now a leader in predictive analytics, is well-positioned to continue reshaping data science by providing insightful analysis and insightful predictions in a world increasingly reliant on data.

X. ACKNOWLEDGEMENT

I would like to sincerely thank Dr. R.C. Jaiswal, my mentor, for his tremendous help and guidance during the entire study process. His deep knowledge and understanding were helpful in improving this paper's quality and bringing it to a level of impact and presentation that I am pleased of. Furthermore, the counsel, expertise, and steadfast support of Dr. R.C. Jaiswal have been an important source of direction and have played a vital role in this endeavor's success.

REFERENCES

[1] Chen, T. & Guestrin C. (2016). XGBoost: A Scalable Tree Boosting System. arXiv preprint arXiv:1603.02754. (https://arxiv.org/pdf/1603.02754.pdf)
[2] Serdar Gündoğdu, (2023). Efficient prediction of early-stage diabetes using XGBoost classifier with random forest feature selection technique.
[3] Breiman L (2001) Random forests. Mach Learn 45:5–32.
[4] Yang Guang. (2021). Generalized XGBoost Method.
[5] Yongshi Deng, Thomas Lumley. (2021). Multiple Imputation Through XGBoost.
[6] Thomas Bartz-Beielstein, Sowmya Chandrasekaran & Frederik Rehbach. (2023). Tuning of Gradient Boosting
[7] Google Developers, Oct 2018, "Descending into ML: Linear Regression", Google LLC
[8] Jason Brownlee, March 2016, "Linear Regression for machine learning", Machine learning mastery, viewed on December 2018
[9] Jaiswal R.C. and Lokhande S.D., A. Ahmed, P. Mahajan, "Performance Evaluation of Clustering Algorithms for IP Traffic Recognition", International Journal of Science and Research (IJSR), volume-4, Issue-5, May-2015, pp. 2786-2792. (ISSN (Online):2319-7064, Index Copernicus Value (2013): 6.14|Impact Factor (2013):4.438
[10] Jaiswal R.C. and Lokhande S.D, "Comparative Analysis using Bagging, Logit Boost and Rotation Forest Machine Learning Algorithms for Real Time Internet Traffic Classification", IMCIP-International Multi Conference on Information Processing –ICDMW- International Conference on Data Mining and Warehousing-2014, PP113-124, ISBN: 9789351072539, University Visvesvaraya College of Engg. Department of Computer Science and Engineering Bangalore University, Bangalore.
[11] Jaiswal R.C. and Lokhande S.D, "Statistical Features Processing Based Real Time Internet Traffic Recognition and Comparative
[12] Study of Six Machine Learning Techniques", IMCIP- International Multi Conference on Information Processing- (ICCNInternational Conference on Communication Networks-2014, PP-120-129, ISBN: 9789351072515, University Visvesvaraya College of Engg. Department of Computer Science and Engineering Bangalore University, Bangalore.
[13] Jaiswal R.C. and Lokhande S.D, "Analysis of Early Traffic Processing and Comparison of Machine Learning Algorithms for Real Time Internet Traffic Identification Using Statistical Approach ", ICACNI-2014- International Conference on Advanced Computing, Networking, and Informatics), Kolkata, India, DOI: 10.1007/978-3-319-07350-7_64, Volume 28 of the book series Smart Innovation, Systems and Technologies (SIST),Page:577-587
[14] Jaiswal R. C. and Taher Saraf, " Stock Price Prediction using Machine Learning", Journal of Emerging Technologies and Innovative Research (JETIR), Open Access, Peer Reviewed and refereed Journal, Indexed in Google Scholar, Microsoft Academic, CiteSeerX, Thomson Reuters, Mendeley : reference manager, ISSN-2349- 5162, Impact Factor:7.95, Volume 9, Issue 9 pp. e33-e41, September 2022.
[15] Jaiswal R. C., Tejveer Pratap and Yashkumar Burnwal, " Multiparametric Monitoring of Vital Signs in Clinical and Home Settings for Patients ", Journal of Emerging Technologies and Innovative Research (JETIR), Open Access, Peer Reviewed and refereed Journal, Indexed in Google Scholar, Microsoft Academic, CiteSeerX, Thomson Reuters, Mendeley : reference manager, ISSN-2349-5162, Impact Factor:7.95, Volume 9, Issue 5 pp. a701-a705, May 2022.
[16] Jaiswal R. C. and Prajwal Pitlehra, "Credit Analysis Using K-Nearest Neighbours' Model", Journal of Emerging Technologies and Innovative Research (JETIR), Open Access, Peer Reviewed and refereed Journal, ISSN-2349- 5162, Impact Factor:7.95, Volume 8, Issue 5, pp. 504-511, May 2021.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY