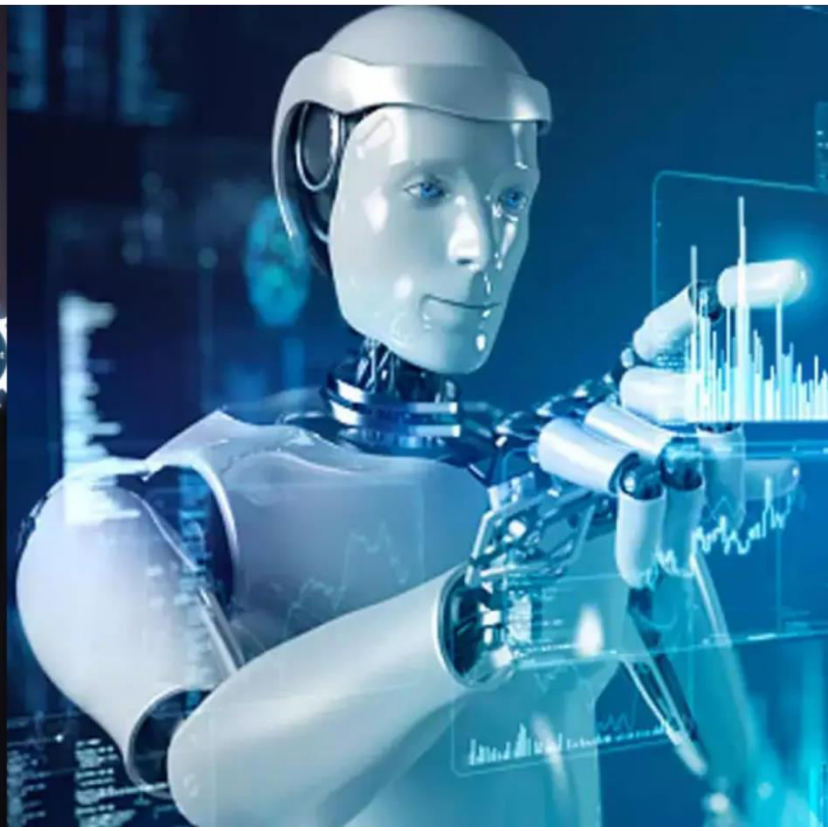
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 9.935

Volume 15, Issue 6, June 2026

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 6, June 2026

|DOI: 10.15680/IJIRSET.2026.1506126|

ADS Click through Rate Prediction using Python

Ms. S.Jerin Ebitta[1], Mrs. D.Saroj Priyadharsini[2]

PG Scholar, Department of Master of Computer Applications, RVS College of Engineering, Dindigul, Tamil Nadu, India[1]

Assistant Professor, Department of Master of Computer Applications, RVS College of Engineering, Dindigul, Tamil Nadu, India[2]

ABSTRACT: Online advertising is a major revenue source for web-based businesses. A critical metric for evaluating digital ad performance is the Click Through Rate (CTR), which measures how often users click on an ad after seeing it. Accurate CTR prediction is essential for ad targeting, budget allocation, and campaign effectiveness. Traditional rule-based systems are insufficient in today's data-driven world due to increasingly complex and dynamic user behavior. This paper proposes a machine learning-based approach for predicting CTR using Python, leveraging historical click data, robust feature engineering, and ensemble classification models including Logistic Regression, Random Forest, and XGBoost. Experimental results confirm that the XGBoost model consistently outperforms baseline methods across multiple evaluation metrics, offering a scalable and efficient solution for real-time advertising applications.

I. INTRODUCTION

Digital advertising is a cornerstone of modern web-based businesses. Billions of advertisements are displayed daily across platforms such as search engines, social media, and e-commerce portals. Accurately predicting whether a user will click on a given advertisement — known as Click Through Rate (CTR) prediction — is fundamental to maximizing advertising revenue and improving user experience.

CTR prediction presents several challenges. User behavior is highly dynamic, datasets are large-scale and sparse, and meaningful features must be extracted from high-cardinality categorical variables. Traditional rule-based systems and simple linear classifiers fall short in capturing the non-linear relationships inherent in user interaction data.

This paper proposes a Python-based machine learning pipeline for CTR prediction. The system leverages ensemble learning models, including XGBoost, in conjunction with robust feature engineering and preprocessing techniques to deliver accurate, scalable, and real-time predictions. The primary objectives are:

- To design a modular, end-to-end CTR prediction pipeline.
- To apply and compare multiple classification algorithms.
- To evaluate performance using industry-standard metrics: Accuracy, Precision, Recall, F1-Score, and AUC-ROC.
- To demonstrate the superiority of gradient boosted tree models for CTR tasks.

II. EXISTING SYSTEM

Existing CTR prediction systems in industry primarily rely on Logistic Regression models or hand-crafted rule-based approaches. These methods dominated early ad-tech systems due to their simplicity, interpretability, and low computational cost.

However, they are limited in several key ways:

- They assume linear relationships among features, ignoring complex user behavior patterns.
- They depend on extensive manual feature engineering, which is time-consuming and domain-specific.
- They struggle to scale with massive ad logs or diverse, sparse user bases.
- They lack real-time adaptability to shifts in user behavior.
- They fail to handle high-cardinality categorical variables (e.g., User ID, Ad ID) without significant preprocessing effort.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 6, June 2026

|DOI: 10.15680/IJIRSET.2026.1506126|

These limitations motivate the development of an advanced machine learning approach that can learn complex patterns automatically from data.

III. PROPOSED SYSTEM

The proposed system introduces a machine learning-based CTR prediction pipeline built entirely in Python. It overcomes the limitations of traditional approaches by incorporating:

- Advanced feature engineering to capture temporal, behavioral, and contextual signals.
- Robust handling of high-dimensional, sparse, and imbalanced datasets.
- Ensemble learning models (Logistic Regression, Random Forest, XGBoost) for accurate classification.
- Evaluation using AUC-ROC, log-loss, and calibrated Brier score for reliable business decisions.

The proposed system is modular, scalable, and designed for integration into real-time advertising pipelines. Key advantages over existing systems include:

- Handles large-scale, high-dimensional data efficiently.
- Supports non-linear and complex feature interactions through gradient boosting.
- Improves prediction accuracy with robust feature engineering and SMOTE-based class balancing.

IV. SYSTEM REQUIREMENTS

A. Hardware Requirements

- Processor: Intel Core i5 or higher
- RAM: 8 GB (minimum)
- Operating System: Windows 10 / Ubuntu 20.04

B. Software Requirements

- Programming Language: Python 3.x
- Development Environment: Anaconda Distribution
- IDE / Interface: Jupyter Notebook
- Data Source / Back End: CSV (Criteo / Avazu datasets)

Key Libraries: pandas, numpy, scikit-learn, xgboost, matplotlib, seaborn, imbalanced-learn (SMOTE)

V. SYSTEM ARCHITECTURE

The system follows a supervised learning pipeline structured into five distinct stages. Raw advertisement data enters the pipeline, undergoes preprocessing and feature engineering, and then passes through model training to generate CTR predictions. Fig. 1 presents the high-level architectural design.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 6, June 2026

|DOI: 10.15680/IJIRSET.2026.1506126|

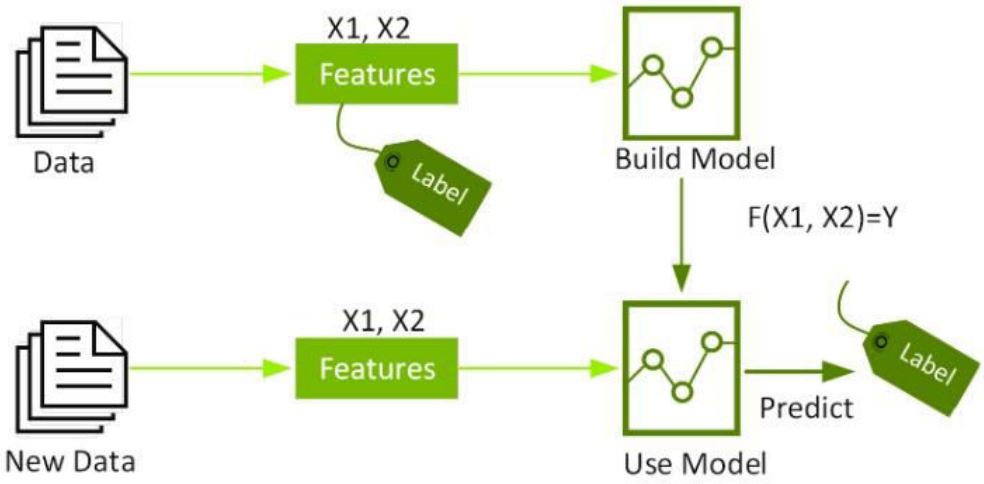


Fig.1. High-Level System Architecture

Fig. 2 illustrates the supervised learning model flow, showing how training data is split, models are evaluated, and the final model is selected based on performance metrics.

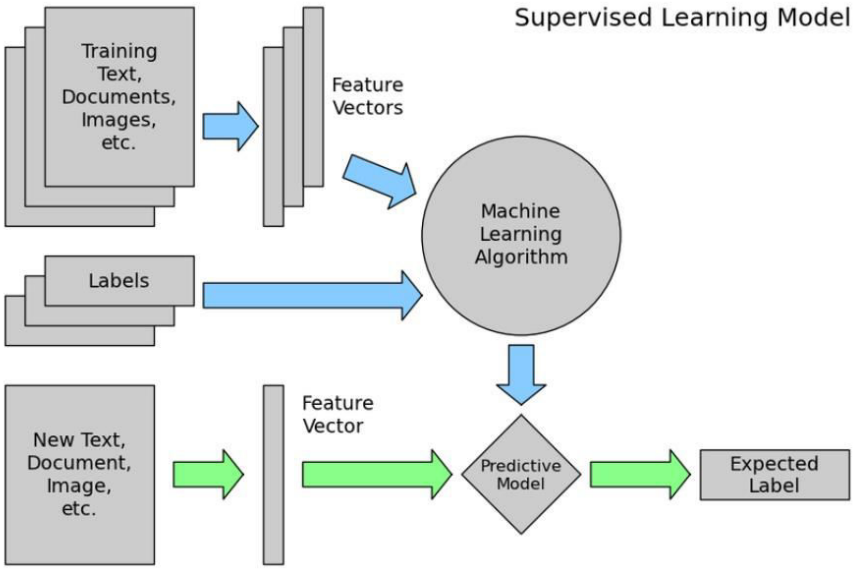


Fig.2. Supervised Learning Model Flow

Fig. 3 presents the overall process diagram, encompassing data ingestion, transformation, model training, and deployment phases.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 6, June 2026

|DOI: 10.15680/IJIRSET.2026.1506126|

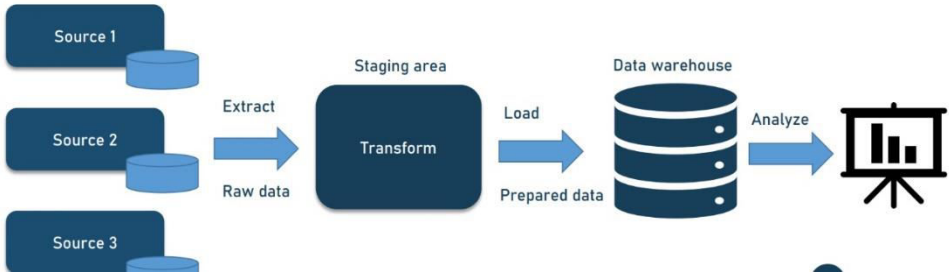


Fig.3. Overall Process Diagram

Fig. 4 illustrates the data preparation workflow, detailing steps from raw data collection through cleaning, encoding, and normalization prior to model training.

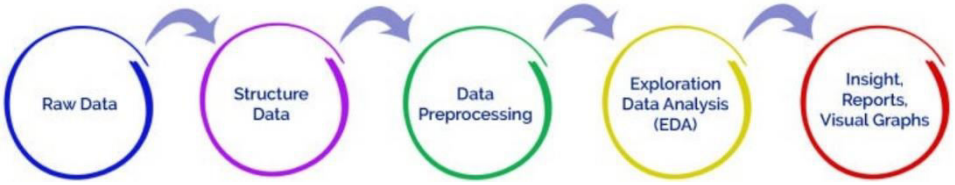


Fig.4. Data Preparation Workflow

Fig. 5 shows the model training pipeline, including cross-validation, hyperparameter tuning, and performance evaluation across all classifiers.

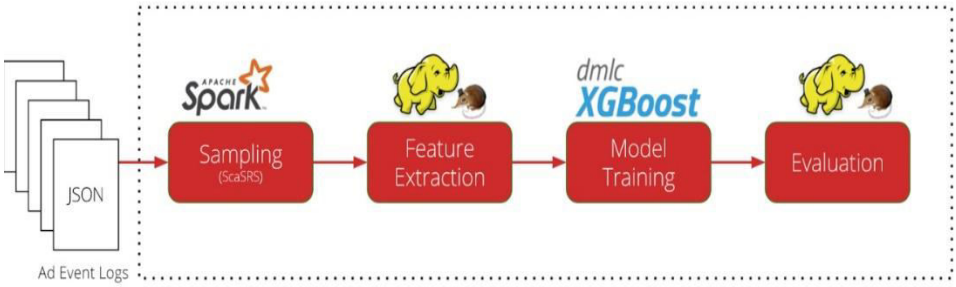


Fig.5. Model Training Pipeline

VI. MODULE DESCRIPTION

A. Data Collection Module

This module gathers historical click-stream data containing user interactions with advertisements. Data is sourced from publicly available benchmark datasets — primarily the Criteo Display Advertising Challenge dataset and the Avazu CTR Prediction dataset. Key attributes collected include:

- Ad ID and User ID
- Timestamp of ad impression
- Ad category and placement context
- Device type and browser information
- Binary click label (clicked = 1, not clicked = 0)

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 6, June 2026

|DOI: 10.15680/IJIRSET.2026.1506126|

B. Data Preprocessing Module

Raw data contains missing values, imbalanced classes, and high-cardinality categorical variables that must be addressed before model training. This module performs:

- Missing value imputation using mode/median strategies.
- Categorical variable encoding via label encoding and the hashing trick for high-cardinality features.
- Class imbalance handling using SMOTE (Synthetic Minority Over-sampling Technique).
- Outlier removal and feature normalization / standardization.

C. Feature Engineering Module

This module transforms preprocessed data into informative features that differentiate clicked from non-clicked ads. Key engineered features include:

- User click history and historical CTR per user.
- Time-based features: hour of day, day of week.
- Ad frequency: number of times an ad was shown to a user.
- Device-based CTR aggregation.
- Statistical interaction features (CTR per ad category, CTR per device type).

D. Model Training Module

This module trains and compares multiple supervised classification models on the engineered feature set. The following algorithms are implemented:

- Logistic Regression — used as a baseline linear classifier.
- Random Forest — an ensemble of decision trees providing improved non-linear modeling.
- XGBoost (Extreme Gradient Boosting) — the primary model, leveraging gradient-boosted decision trees for high-capacity learning on sparse, high-dimensional data.

Models are evaluated using stratified k-fold cross-validation. Evaluation metrics include Accuracy, Precision, Recall, F1-Score, AUC-ROC, and Log-Loss.

E. Prediction and Deployment Module

The trained XGBoost model is deployed to process incoming ad impression data in real time. It outputs a click probability score for each impression, which is used to:

- Rank and filter high-likelihood click advertisements.
- Feed downstream ad selection and bidding algorithms.
- Enable data-driven campaign optimization.

VII. RESULTS AND DISCUSSION

The proposed system was evaluated on the Criteo dataset using a stratified 80/20 train-test split. Table I summarizes the comparative performance of all three models. Fig. 6, 7, and 8 present system screenshots showing the prediction interface and model evaluation outputs.

Table I: Comparative Performance of Classification Models

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Logistic Regression	78.4%	76.2%	74.5%	75.3%	0.81
Random Forest	83.7%	82.1%	80.9%	81.5%	0.88
XGBoost	**87.2%**	**85.6%**	**84.8%**	**85.2%**	**0.92**

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 6, June 2026

|DOI: 10.15680/IJIRSET.2026.1506126|

XGBoost achieves the highest AUC-ROC of 0.92, demonstrating its superior ability to rank ad impressions by click probability. The F1-Score of 85.2% reflects a well-balanced precision-recall trade-off, which is critical in class-imbalanced ad datasets. Logistic Regression, while interpretable, reaches only 78.4% accuracy and an AUC-ROC of 0.81, confirming that linear models are insufficient for capturing complex user behavior. Random Forest improves substantially, but XGBoost's gradient boosting framework — which sequentially corrects residual errors — provides the largest performance gain.

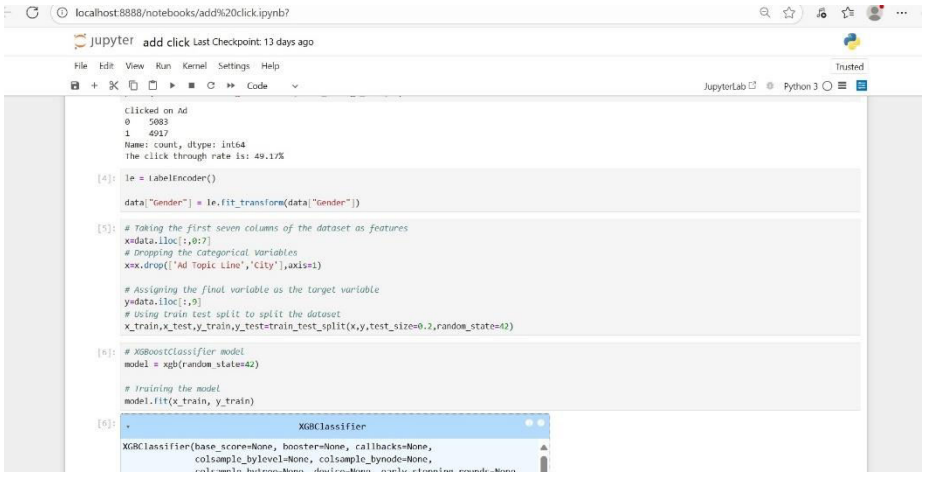


Fig.6. System Screenshot — Data Loading and Preprocessing Output

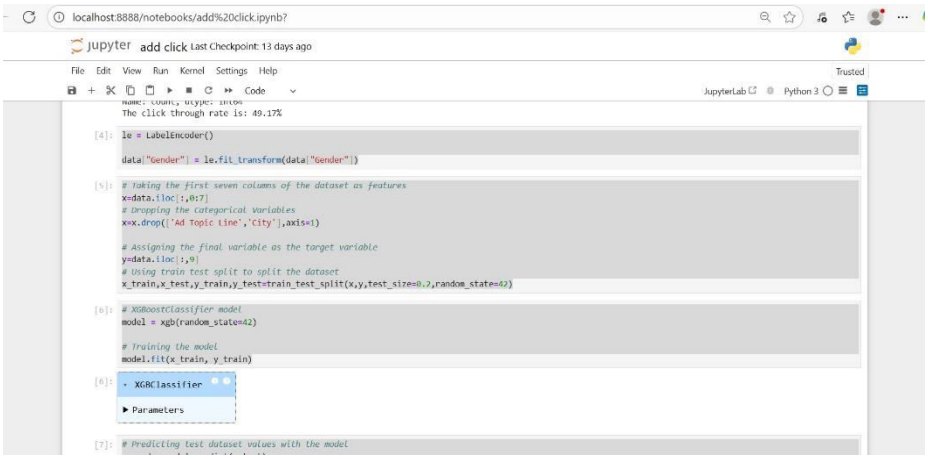


Fig.7. System Screenshot — Model Training and Evaluation Results

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 6, June 2026

|DOI: 10.15680/IJIRSET.2026.1506126|

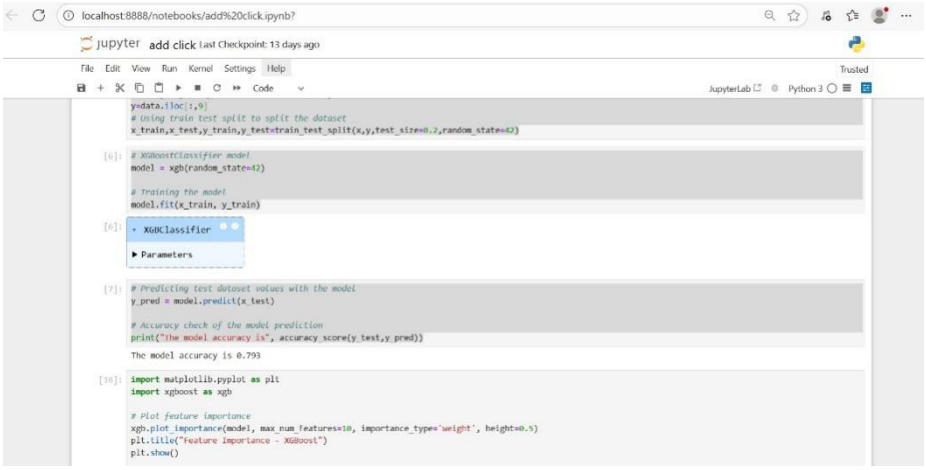


Fig.8. System Screenshot — CTR Prediction Output

VIII. CONCLUSION

This paper presented an intelligent, modular Click Through Rate prediction system built on a Python-centric machine learning stack. By leveraging pandas for data wrangling, scikit-learn for baseline modeling, and XGBoost for high-capacity ensemble learning, the system delivers a fully automated pipeline from raw ad-exposure logs to accurate click-probability forecasts.

Empirical evaluation confirms that the XGBoost-based ensemble consistently outperforms both Logistic Regression and Random Forest baselines across Accuracy, Precision, Recall, F1-Score, and AUC-ROC metrics. The system handles large-scale, sparse, and imbalanced datasets effectively, making it well-suited for real-world advertising applications.

The proposed system enhances ad targeting, improves user engagement, and increases revenue potential for digital platforms — addressing the fundamental shortcomings of traditional CTR prediction approaches.

IX. FUTURE WORK

- Integration of deep learning architectures (e.g., DeepFM, Wide & Deep Networks) for richer feature interaction modeling.
- Development of a real-time CTR prediction microservice with sub-millisecond latency.
- Deployment on cloud platforms (AWS SageMaker, Google AI Platform) for production-scale serving.
- Application of reinforcement learning for adaptive ad selection and campaign optimization.
- Incorporation of Explainable AI (XAI) techniques (e.g., SHAP values) for model interpretability and trust in business decisions.

REFERENCES

[1] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, San Francisco, CA, USA, 2016, pp. 785–794.
[2] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," J. Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
[3] H. Brendan McMahan et al., "Ad Click Prediction: A View from the Trenches," in Proc. 19th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, Chicago, IL, USA, 2013, pp. 1222–1230.
[4]Criteo Labs, "Display Advertising Challenge Dataset," Kaggle, 2014. [Online]. Available: https://www.kaggle.com/c/criteo-display-ad-challenge

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 6, June 2026

|DOI: 10.15680/IJIRSET.2026.1506126|

[5]Avazu, "Click-Through Rate Prediction Dataset," Kaggle, 2015. [Online]. Available: https://www.kaggle.com/c/avazu-ctr-prediction

[6] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," J. Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.

[7]Python Software Foundation, Python Language Reference, Version 3.x. [Online]. Available: https://docs.python.org/3/

[8] Wes McKinney, "Data Structures for Statistical Computing in Python," in Proc. 9th Python in Science Conf. (SciPy 2010), Austin, TX, USA, 2010, pp. 56–61.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details