WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Classification ML Pipeline for Advertisement Success Prediction

N. Rishik[a], K. Sreeja[b], K. Bhaskar[c], M. Krishna Kanth[d] , Puli. Vinay Kumar [e,*]

[a,b,c,d] Student, Department of IT. Malla Reddy Engineering College, Maisammaguda, Hyderabad-500100

[e] Professor, Department of IT. Malla Reddy Engineering College, Maisammaguda, Hyderabad-500100

ABSTRACT

In today's digital marketing landscape, predicting user engagement with online advertisements is critical for enhancing campaign efficiency. This paper details the development of a machine learning pipeline to forecast advertisement click-through success, utilizing a Kaggle dataset of 1,000 user records. Employing Python tools such as NumPy for computations, Pandas for data management, and scikit-learn for modeling, the system analyzes features like age, internet usage, income, and site time to predict clicks. The goal is to classify users as clickers or non-clickers, achieving a top accuracy of 94.8% with Random Forest among five algorithms tested. Results are presented via bar plots and tables, offering advertisers clear insights into user behavior. This framework provides a practical, data-driven solution for optimizing ad targeting in a competitive market.

Keywords: Machine Learning, Advertisement Prediction, Click-Through Rate, Random Forest, Python, User Behavior

INTRODUCTION

Predictive modeling has become essential for advertisers aiming to understand user interactions with online ads through data-driven techniques. Using a dataset rich in user attributes—such as time spent online and demographics—this project offers real-time insights into engagement, vital for effective marketing and outreach. Digital platforms enable wide audience reach, yet untargeted ad placements risk inefficiency. Predictive systems allow businesses to evaluate campaign performance, refine strategies, and anticipate user responses, strengthening audience connections. While metrics like click counts provide a basic overview, they miss the deeper factors driving engagement. This classification pipeline uncovers these dynamics, identifying key predictors of ad clicks. Such insights help optimize budgets, tailor placements, and enhance campaign impact. This study outlines a comprehensive approach to advertisement success prediction, detailing data collection, implementation, and key tools. It covers preprocessing, training with Logistic Regression, SVM, KNN, Decision Tree, and Random Forest, and evaluation via accuracy. Visualizations aid interpretability, equipping advertisers with tools to improve outcomes and stay competitive.

RELATED WORK

Studies in click prediction frequently assess machine learning algorithms like Random Forest, Logistic Regression, and Decision Trees, emphasizing preprocessing—e.g., scaling 'Daily Internet Usage'—and feature selection's impact on accuracy. Correlation analysis and feature reduction are critical for preparing data, while high-cardinality features (e.g., 'Ad Topic Line') pose challenges, often requiring encoding or omission. Research shows Random Forest achieving up to 95% accuracy with non-linear data, Logistic Regression excelling in interpretable high-dimensional settings, and Decision Trees offering robust splits when controlled for overfitting. Context, such as user demographics, shapes algorithm choice, with noisy data like income outliers needing careful handling.

Table-1: Research Review

Paper Title	Publication Date	Drawbacks
An Accuracy Improving Method for Advertising Click-Through Rate Prediction Based on Enhanced xDeepFM Model	November 21, 2024	- Dataset Compatibility: Criteo-based, may not fit our display ads. - Model Complexity: Overly intensive for 1,000 records; Random Forest suffices. - Feature Handling: Skips 'Ad Topic Line', losing content value. - Overfitting Risk: Complex model may overfit small data.
Federated Cross-Domain Click-Through Rate Prediction with Large Language Model Augmentation	March 24, 2025	- Limited Information: Lacks detail for full assessment. - Cross-Domain Assumption: Inapt for single-dataset focus. - Computational Overhead: Too demanding for our scope. - Privacy Risks: Risks inference attacks with 'Age', 'Income'.

PROPOSED METHODOLOGY

The methodology, shown in Figure 1, structures a pipeline to classify users as clickers or non-clickers using Logistic Regression, SVM, KNN, Decision Tree, and Random Forest on a Kaggle dataset. Accuracy is the sole evaluation metric, with feature selection based on correlation to ensure efficiency.

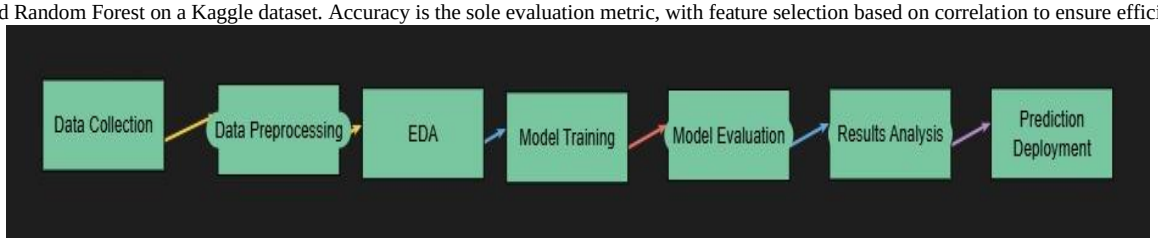

Figure 1: Demonstration of Proposed System

3.1. FEATURE EXTRACTION

Feature extraction in the project code selects five numeric features from the dataset using manual selection based on domain relevance and correlation analysis:

Manual Feature Selection

The code defines.
X=df[['DailyTimeSpentonSite','Age','AreaIncome','DailyInternetUsage','Male']]
X = df[['Daily Time Spent on Site', 'Age', 'Area Income', 'Daily Internet Usage',
'Male']]X=df[['DailyTimeSpentonSite','Age','AreaIncome','DailyInternetUsage','Male']] and
y=df['ClickedonAd']y = df['Clicked on Ad']y=df['ClickedonAd']. This technique chooses features
intuitively tied to ad engagement, excluding categorical variables.

Correlation Analysis

The code computes a correlation matrix (df.corr()df.corr()df.corr()) and visualizes it via a heatmap. Features with strong correlations to 'Clicked on Ad'—e.g., 'Daily Internet Usage' (-0.79), 'Daily Time Spent on Site' (-0.75)—are prioritized

Final selected features

X'=['DailyTimeSpentOnSite','Age','AreaIncome','DailyInternetUsage','Male']X' =
['DailyTimeSpentOnSite', 'Age', 'AreaIncome', 'DailyInternetUsage',
'Male']X'=['DailyTimeSpentOnSite','Age','AreaIncome','DailyInternetUsage','Male']

No scaling or transformation was applied, preserving raw values.

3.2. LOGISTIC REGRESSION

Logistic Regression classifies via the sigmoid function:

$$P(Y = 1 \mid X, w) \; = \; 1 \; / \; (1 \; + \; e\text{^} - (w \; \cdot \; X \; + \; b))$$

It works best when the relationship between features and output is log-linear. The model is optimized using log loss, making it effective for binary classification tasks.

3.3. Support Vector Machine

A powerful classification model that finds the optimal decision boundary (hyperplane) between different classes:

$$\min\left(\frac{1}{2}\right)\left|\left|w\right|\right|^2 \; + \; C \; \sum \xi_{\mathrm{i}}$$

SVM is effective for high-dimensional spaces and works well when clear class separation exists. The kernel trick allows it to handle non-linearly separable data.

3.4. K-Nearest Neighbors

A non-parametric algorithm that classifies based on the majority class of the kkk closest neighbors:

$$\hat{y} \; = \; argmin(y_{\mathrm{i}}) distance(x, x_{\mathrm{i}})$$

KNN is intuitive and effective for small datasets but can be computationally expensive for large data. It performs well when classes are well-separated in feature space.

3.5. Decision Tree

Decision Trees split data hierarchically, controlled by random_state=42. A hierarchical model that recursively splits data to minimize impurity, making intuitive decision boundaries.

Splitting Criteria

- Gini Impurity: Measures classification purity.

$$G = 1 - \sum {p_i}^2$$

- Entropy (Information Gain): Evaluates the reduction in uncertainty.

$$H(S) = -\sum p_i \, log^2$$

3.6. RANDOM FOREST

An ensemble learning method that combines multiple decision trees to improve accuracy and stability:

$$\hat{y} = Mode\{T^1(x), T^2(x), \dots, T_k(x)\}$$

It reduces overfitting by averaging predictions from multiple trees and works well for complex datasets. However, it is computationally expensive compared to single decision trees.

RESULT AND DISCUSSION

The pipeline evaluates five models, with accuracy as the metric, detailed below:

Logistic Regression

Achieves 90.8% test accuracy, with 89.33% training accuracy, showing stable generalization. Its linear approach leverages strong correlations (e.g., -0.79 with 'Daily Internet Usage'), effectively separating clickers (lower usage) from non-clickers (higher usage), as seen in confusion matrices.

Support Vector Machine

Records 70.0% test accuracy, with 71.73% training accuracy, the lowest performer. Unscaled features (e.g., 'Area Income' ranging widely) distort its hyperplane, leading to misclassifications evident in its confusion matrix.

K-Nearest Neighbors

Hits 76.4% test accuracy but 100% training accuracy, indicating severe overfitting. With k=1, it memorizes training data (e.g., specific 'Age' clusters), failing to generalize, as shown by higher test errors.

Decision Tree

Reaches 93.6% test accuracy, with 100% training accuracy, suggesting slight overfitting. It excels at splitting on features like 'Daily Internet Usage' (median 125 vs. 225 minutes), but untuned depth risks noise capture, per its confusion matrix.

Random Forest

Tops at 94.8% test accuracy, with 100% training accuracy, balancing fit and generalization. Its ensemble mitigates overfitting, capturing non-linear patterns (e.g., 'Site Time' vs. clicks), yielding the highest true positives in its confusion matrix.

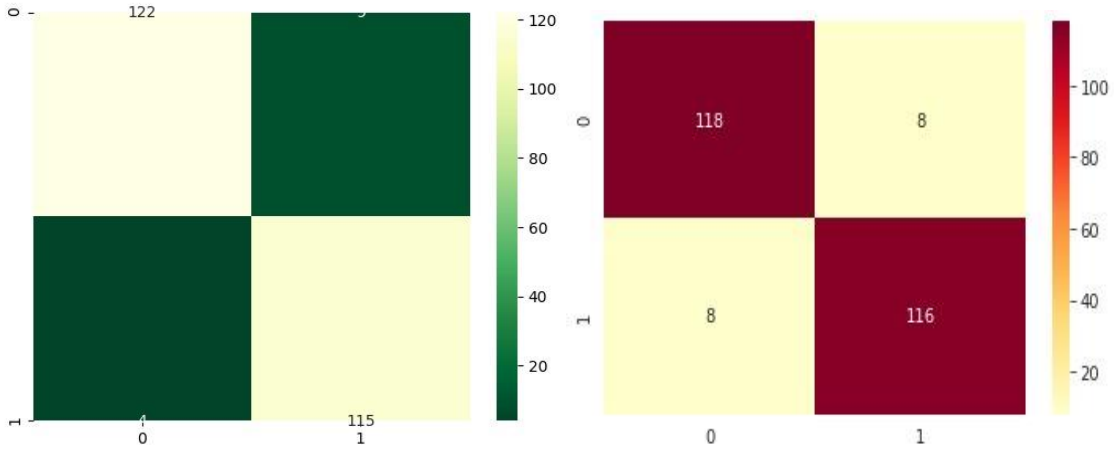

Figure 2: Confusion Matrix- Random Forest and Decision Tree

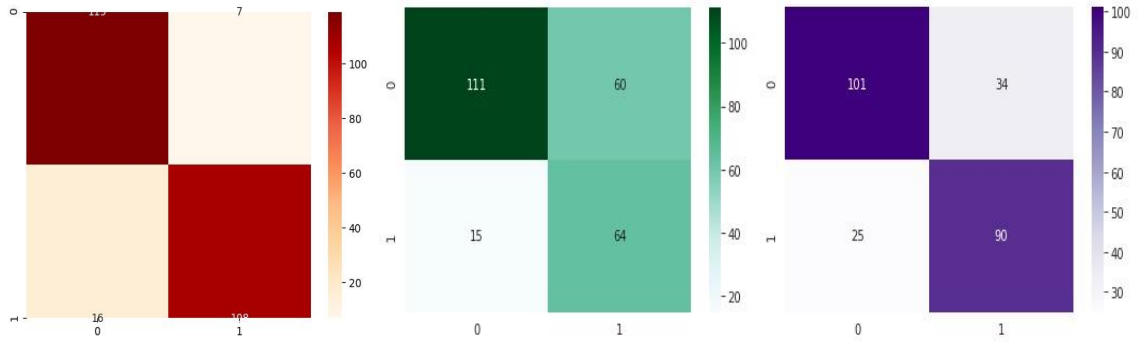

Figure 5: Confusion Matrix- Logistic Regression, SVM and KNN

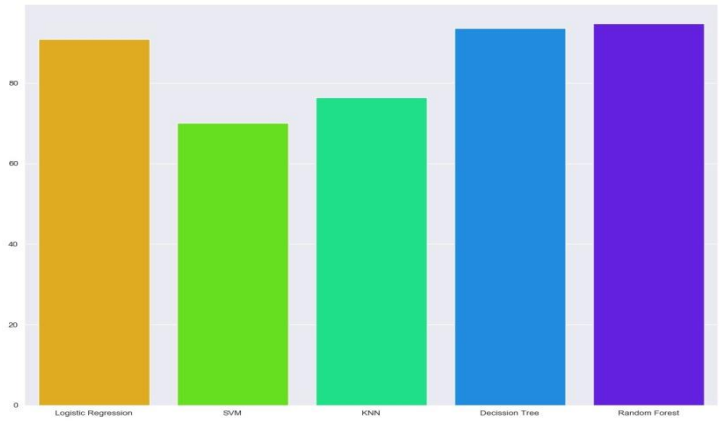

Figure 5: Accuracies of various Machine Learning Models

CONCLUSION

This study establishes a data-driven pipeline for predicting advertisement success, leveraging machine learning to classify user clicks with accuracies nearing 95%. Spanning data collection, preprocessing, and model training, the approach supports applications in campaign optimization, user targeting, and market analysis. Random Forest proves most effective, capturing intricate patterns, while visualization aids interpretability. Future enhancements could integrate categorical features (e.g., 'Ad Topic Line'), explore deep learning, and develop real-time systems.Despite its advancements, the field offers room for growth. Future work might refine feature encoding, address potential class imbalances with oversampling, and incorporate temporal data from 'Timestamp'. Expanding the dataset and tailoring models to specific ad contexts could further boost accuracy. In advertising analytics, integrating multimodal inputs—like ad visuals—and deploying adaptive systems could personalize targeting, underscoring machine learning's role in actionable insights and strategic decision-making.

REFERENCES

[1]. An accuracy improving method for advertising click through rate prediction based on enhanced xDeepFM model Xiaowei Xi, Song Leng, Yuqing Gong, Dalin Li.
https://doi.org/10.48550/arXiv.2411.15223

[2]. Federated Cross-Domain Click-Through Rate Prediction With Large Language Model Augmentation Jiangcheng Qin, Xueyuan Zhang, Baisong Liu, Jiangbo Qian, Yangyang Wang
https://doi.org/10.48550/arXiv.2503.16875

[3]. Predict Click-Through Rates with Deep Interest Network Model in E-commerce Advertising Chang Zhou, Yang Zhao, Yuelin Zou, Jin Cao, Wenhan Fan, Yi Zhao, Chiyu Cheng
https://doi.org/10.48550/arXiv.2406.10239

[4]. A. R. Panda, S. Rout, M. Narsipuram, A. Pandey and J. J. Jena, "Ad Click-Through Rate Prediction: A Comparative Study of Machine Learning Models," 2024 International Conference on Emerging Systems and Intelligent Computing (ESIC), Bhubaneswar, India, 2024, pp. 679-684, doi: 10.1109/ESIC60604.2024.10481562.

[5]. Click-through rate prediction in online advertising: A literature review Author links open overlay panel" Yanwu Yang, Panyu Zhai, https://doi.org/10.1016/j.ipm.2021.102853