IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Ad Clicks Deception Finding Using Supervised Machine Learning Techniques

A Ravi Kumar, K Bhanu Prakash, A Venkata Suresh Reddy, Ch Vasanth Reddy. K Shashi Vardhan
Dept of Information Technology, Sreenidhi Institute of Science and Technology (Autonomous), Hyderabad

Abstract
Machine learning classifiers were evaluated for their effectiveness in detecting fraudulent ad clicks. Among the models tested, the Random Forest algorithm demonstrated the strongest performance. Incorporating predictive models not only improved the click-through rate but also enabled the system to learn from data autonomously, eliminating the need for manual domain expertise. However, a significant challenge in this field is the limited availability of well-labeled, publicly accessible datasets, which restricts the scope of experimentation and analysis. Fraudulent ad clicks can be artificially generated and mixed with genuine user activity on websites, complicating detection efforts. In a comparative analysis, the Support Vector Machine (SVM) achieved the highest standalone accuracy at 93.3%, while other machine learning models ranged between 84% and 92%. In ensemble approaches, combining XGBoost with Random Forest also yielded a 93.3% accuracy rate, whereas the combination of SVM and XGBoost resulted in a lower accuracy of 76.6%. Overall, both SVM and the ensemble of XGBoost and Random Forest demonstrated the best predictive performance.

Keywords: Advertise, fraud click, Ensemble machine learning, prediction, user, legitimate, link.

I INTRODUCTION

Online advertising has become a major target for fraudulent activity, particularly in the business sector. As digital media and web technologies continue to evolve, advertising has shifted away from traditional platforms like newspapers and television toward online channels and mobile applications to reach new audiences. Major internet companies such as Google, Facebook, and Yahoo generate a substantial portion of their income from digital advertisements [1]. These platforms serve as intermediaries, connecting advertisers with publishers and charging fees based on user engagement metrics, such as clicks. Publishers, in turn, receive compensation for driving user traffic to advertisers' websites.

However, this compensation model has opened the door to a serious security challenge: click fraud. This type of fraud distorts advertising metrics and leads to significant financial losses. Some advertising platforms may experience artificially inflated click rates, often caused by automated systems or malicious users who generate meaningless clicks. As online advertising has become a critical source of income for many websites, attackers have exploited this structure to generate fraudulent income by continuously clicking on paid links [2].

The financial incentives of digital marketing have made it particularly vulnerable to manipulation. Click fraud is the most prevalent form of deceptive activity in performance-based advertising. In such schemes, malicious actors repeatedly click on ads without genuine interest, driving up costs for advertisers. These fraudsters may also manipulate the system by adopting fake identities or deploying bots to simulate legitimate user behavior [3]. The goal is often either to gain financial rewards or to exhaust a competitor's advertising budget. This form of exploitation undermines the reliability of digital marketing, disrupts auctions, and negatively impacts content quality and advertiser trust. Since many websites rely heavily on advertising revenue, the issue of click fraud threatens their sustainability.

Addressing click fraud is both resource-intensive and technically challenging. Fraud bots are constantly evolving, adapting to detection mechanisms and becoming more sophisticated over time [4]. Various prediction-based techniques have been developed to detect such fraud, including statistical models that can trace suspicious IP addresses involved in fraudulent activity [5]. However, most current detection systems operate offline and are not equipped for real-time intervention.

This study aims to explore techniques for filtering out unreliable users, identifying valuable customers, and improving repayment predictions—an approach that could enhance transparency for both advertisers and consumers. Detecting click fraud involves solving a binary classification problem: determining whether a user will or will not engage in fraudulent behavior.

Given the imbalanced nature of most datasets in this context, precision and recall are often prioritized over overall accuracy. Logistic regression, for instance, may be favored due to its performance on false negatives as measured by confusion matrix metrics. Fraud detection often requires the combination of high-dimensional feature sets and complex machine learning algorithms. Additionally, the volume of data can be overwhelming—large advertising platforms may process hundreds of millions of clicks each day. This scale makes training and deploying real-time models both computationally expensive and technically demanding. Consequently, many fraud incidents remain undetected at the time they occur. Real-time detection would significantly improve advertisers' and networks' ability to respond promptly and mitigate the effects of fraudulent activity.

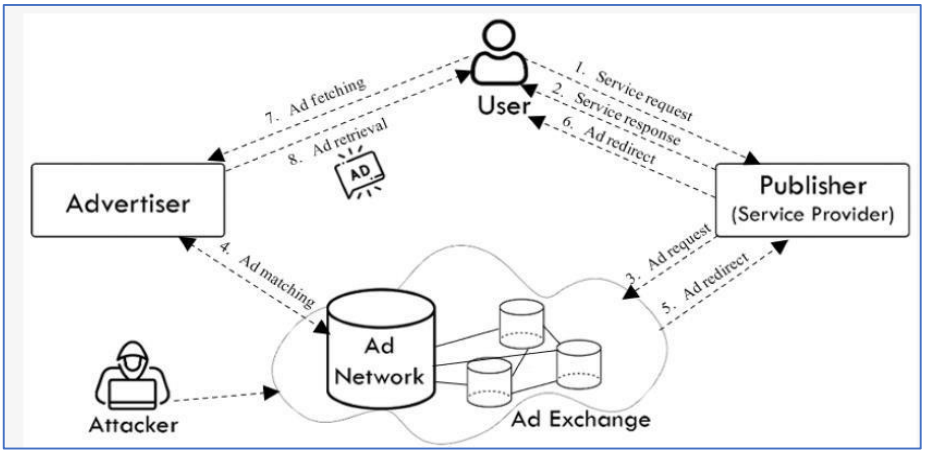

Figure 1: Architecture of online advertising [15]

II BACKGROUND ANALYSIS

Brand advertising and performance advertising are the two main categories of digital marketing. The key goal of brand advertising is to increase awareness by exposing the brand to as many people as possible. Before the rise of digital marketing, brand advertising was the dominant strategy, with television, newspapers, billboards, and other outdoor media being the primary channels. In this form of advertising, marketers focus on how many people they can reach in relation to their budget. Brand ads often operate on a pay-per-impression model, where advertisers are charged based on how many times their ad is viewed. Facebook is an example of a platform designed to help businesses boost visibility.

With declining ad revenue, content creators are increasingly exploring alternative monetization strategies. Ads remain a key source of income for websites. By showing targeted advertisements and encouraging user interaction, website owners can earn revenue while offering value to their visitors. Advertisers, ad networks, and publishers usually base payments on metrics like page views, clicks, or completed forms. Google is known for its pay-per-click advertising, but similar models are also available through Yahoo and Bing. These search engines deliver relevant ads based on users' queries and charge advertisers when users engage with them.

Paid advertising placements and cost-per-click strategies aren't unique to Google; search engines like Yahoo and Bing also support these models. These platforms are responsible for displaying contextually relevant ads and track user engagement to charge advertisers accordingly. However, both brand and performance marketing are vulnerable to ad fraud.

Fraud—especially impression-based fraud—is widespread in digital advertising. Platforms like Google Display Network and Facebook Audience Network (FAN) are commonly targeted. Ad networks often partner with third-party publishers, but some dishonest publishers exploit these relationships to promote fake content or generate false ad impressions. The most frequent type of fraud in this space is click fraud, where bad actors simulate ad clicks to mislead the ad platform into misbilling advertisers. These

fraudsters continually seek new ways to exploit the system and divert revenue for personal gain. Click fraud, which may involve bots or people generating invalid clicks, is used to either earn undeserved money or exhaust a competitor's advertising budget. Over time, this undermines the digital advertising ecosystem and threatens the viability of ad-funded content, possibly leading advertisers to abandon these platforms.

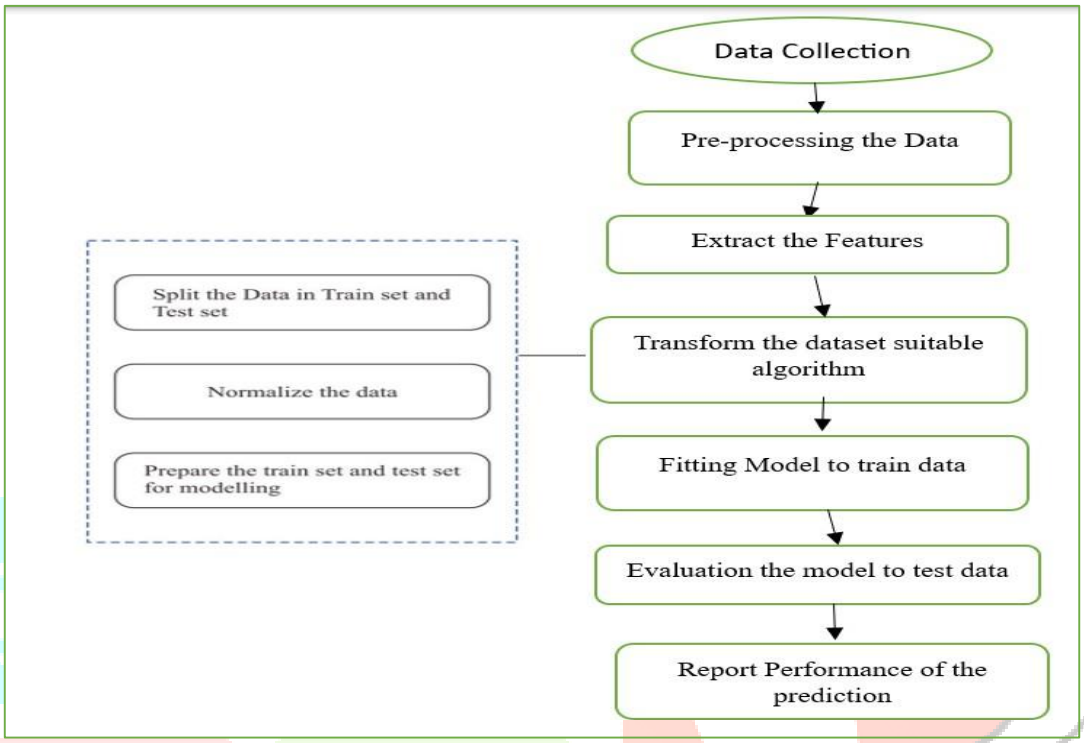

Figure 2: Proposed architecture

III RESULTS AND ANALYSIS

We will leverage multiple machine learning models built using Python to develop an ensemble model aimed at predicting whether a customer made a purchase. The dataset for this project has been sourced from Kaggle. Since this is a classification task, we will evaluate model performance using appropriate classification metrics. The prediction will be based on a set of selected features included in our project.

1. User ID - Customer Unique ID
2. Gender - Gender of a customer - M/F
3. Age - Age of a customer
4. Estimated Salary - Estimated salary of a customer

The upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enable it. Saving Social_Network_Ads.csv to Social_Network_Ads (1).csv

Table 1: Dataset

	User ID	Gender	Age	EstimatedSalary	Purchased
0	15624510	Male	19	19000	0
1	15810944	Male	35	20000	0
2	15668575	Female	26	43000	0
3	15603246	Female	27	57000	0
4	15804002	Male	19	76000	0

Table 2: Dataset Exploration

	User ID	Gender	Age	EstimatedSalary	Purchased
0	15624510	Male	19	19000	0
1	15810944	Male	35	20000	0
2	15668575	Female	26	43000	0
3	15603246	Female	27	57000	0
4	15804002	Male	19	76000	0
...	...	...	...	...	...
395	15691863	Female	46	41000	1
396	15706071	Male	51	23000	1
397	15654296	Female	50	20000	1
398	15755018	Male	36	33000	0
399	15594041	Female	49	36000	1

400 rows × 5 columns

The following table 2 describes the dataset exploration. As the User ID is based on the customer ID and is unique by each customer, we will drop the User ID

Table 3: Dataset with drop user ID

	Gender	Age	EstimatedSalary	Purchased
0	Male	19	19000	0
1	Male	35	20000	0
2	Female	26	43000	0
3	Female	27	57000	0
4	Male	19	76000	0
...	...	...	...	...
395	Female	46	41000	1
396	Male	51	23000	1
397	Female	50	20000	1
398	Male	36	33000	0
399	Female	49	36000	1

400 rows × 4 columns

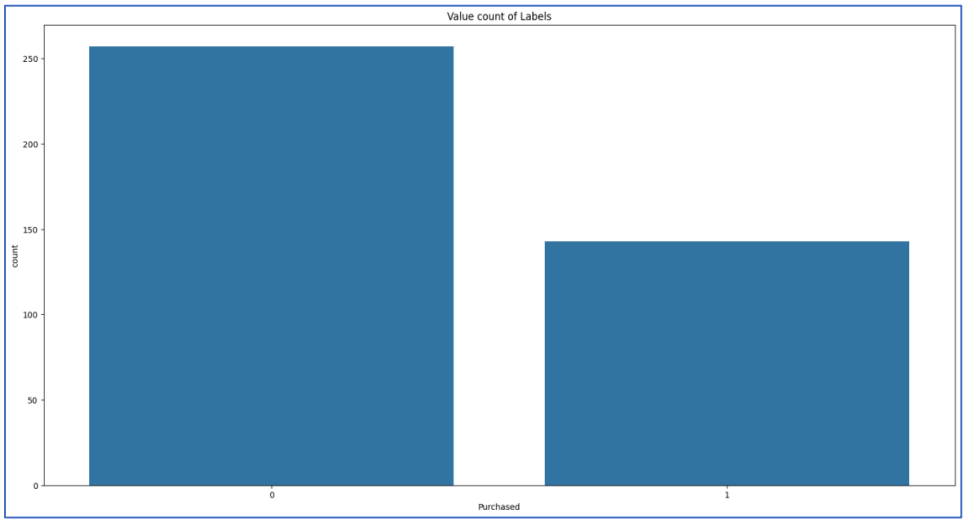

Figure 3: Value count of Labels

As we can see, the data is imbalanced. Value count of labels based on much number of people who click the ads.

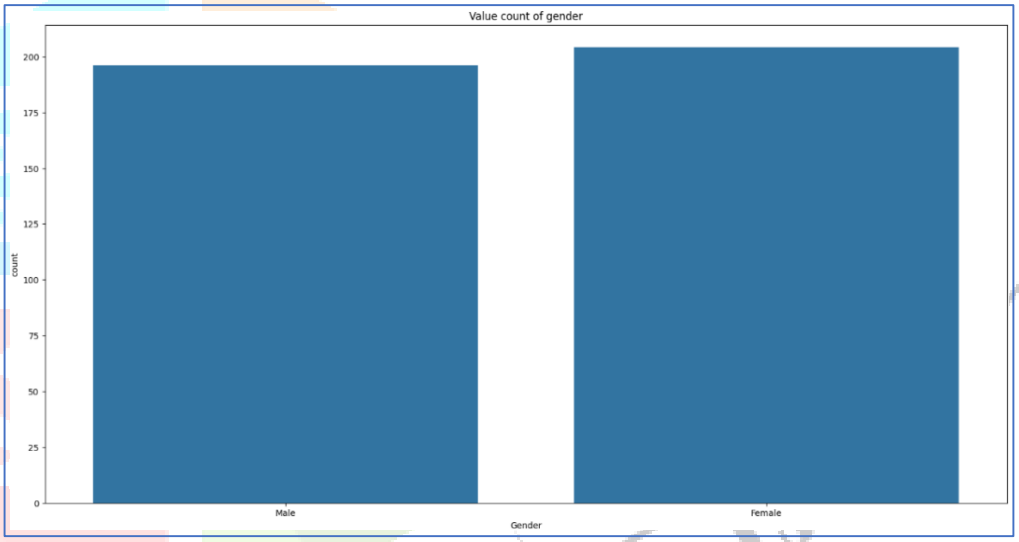

Figure 4: Value count of Gender

Figure 4 describes the gender-based ad clicks.

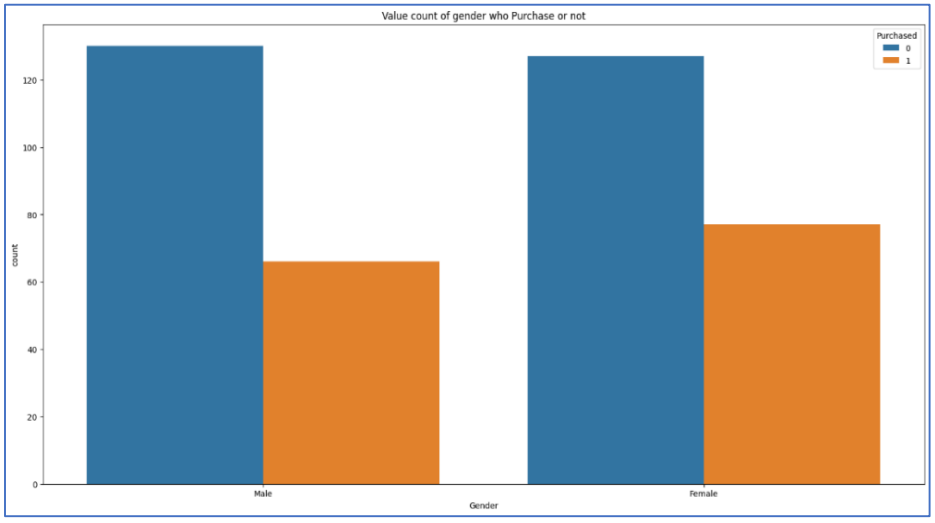

Figure 5: Value count of Gender, who purchased or not

Figure 5 shows the gender of the purchasers of items.

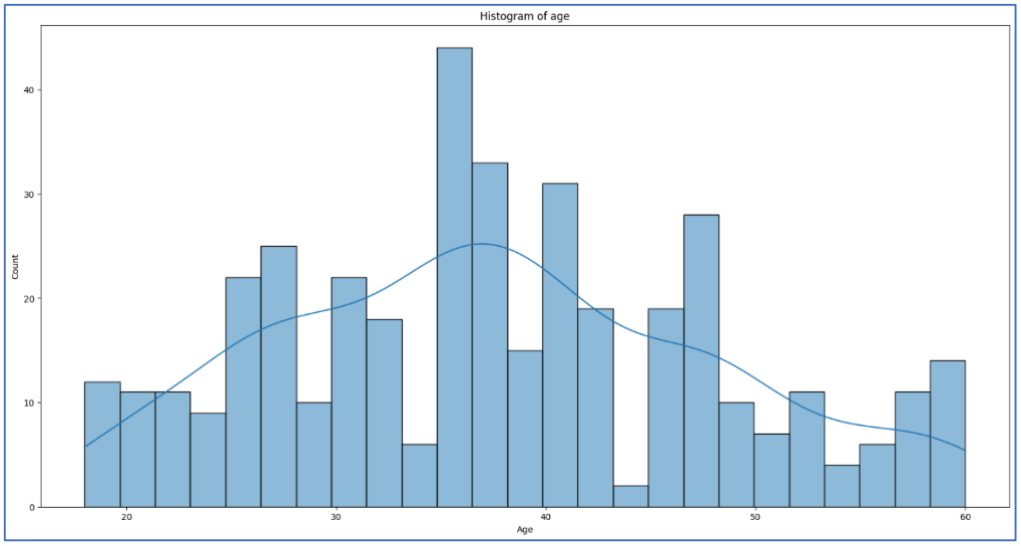

Figure 6: Histogram of age

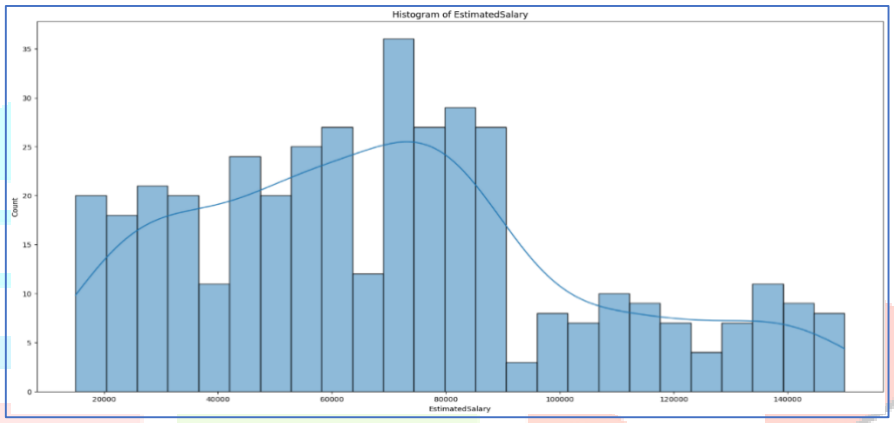

Figure 7: Histogram of Estimated Salary

The following Figures 6 and 7 show the age and corresponding estimated salary of customers.

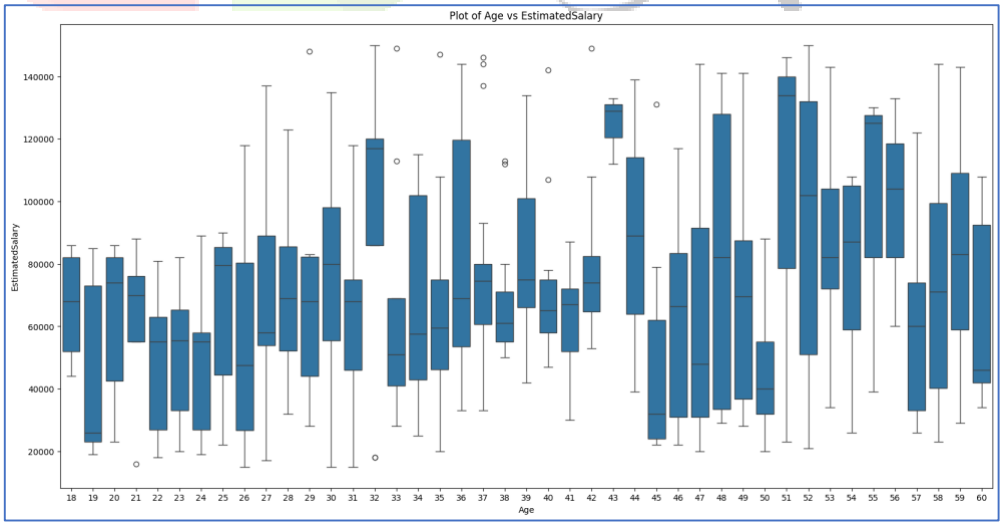

Figure 8: Plot of Age Vs Estimated Salary

Figure 8 shows the age versus estimated salary.

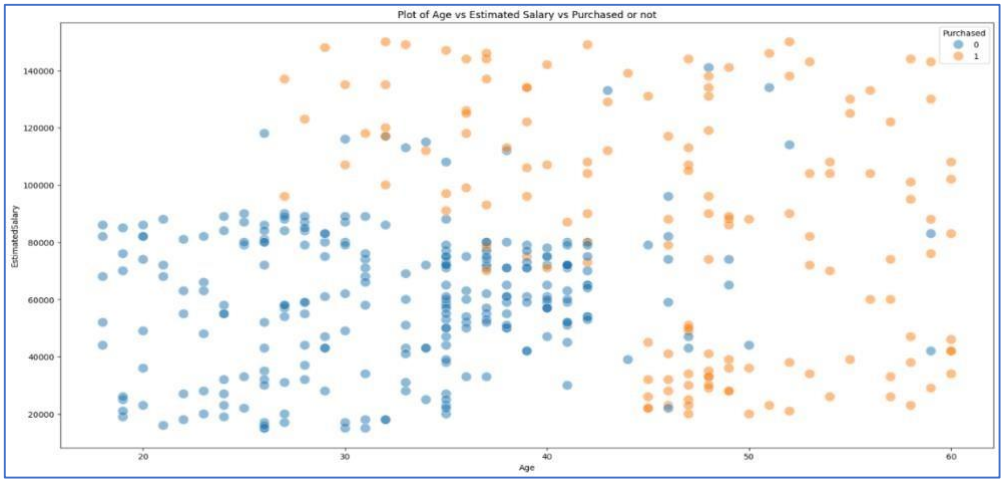

Figure 9: Plot of Age vs. estimated Salary Vs Purchase or not

Figure 9 illustrates the relationship between age, estimated salary, and whether or not a purchase was made. A heatmap provides a visual way to represent data values, making it easier to identify patterns and trends. These visual tools are especially useful for understanding user interaction on a webpage, such as which sections draw the most attention or fail to engage users. To examine the relationships between variables, two commonly used statistical methods are the Pearson and Spearman correlation coefficients. The Pearson coefficient measures how strongly two variables are linearly related, whereas the Spearman coefficient is used to assess the strength and direction of a monotonic relationship.

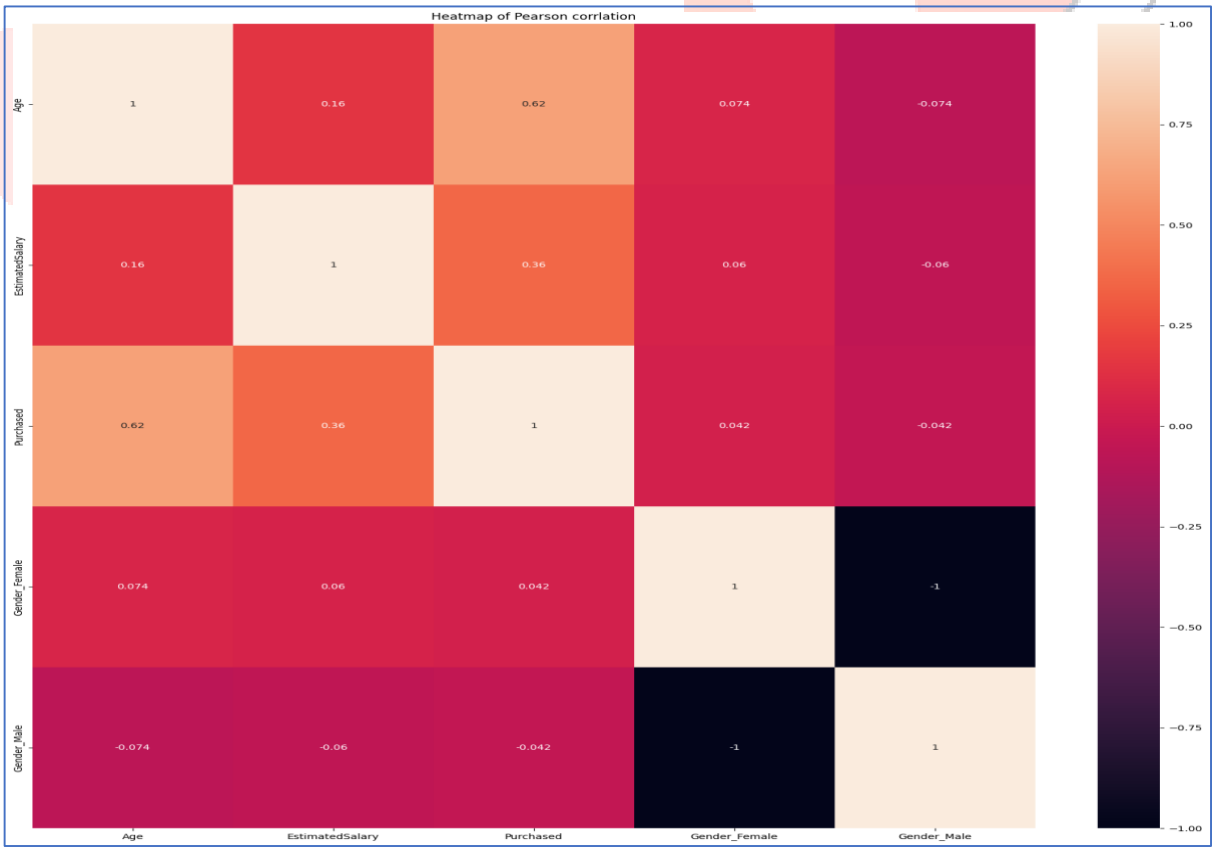

Figure 9: Heat map of Pearson Correlation

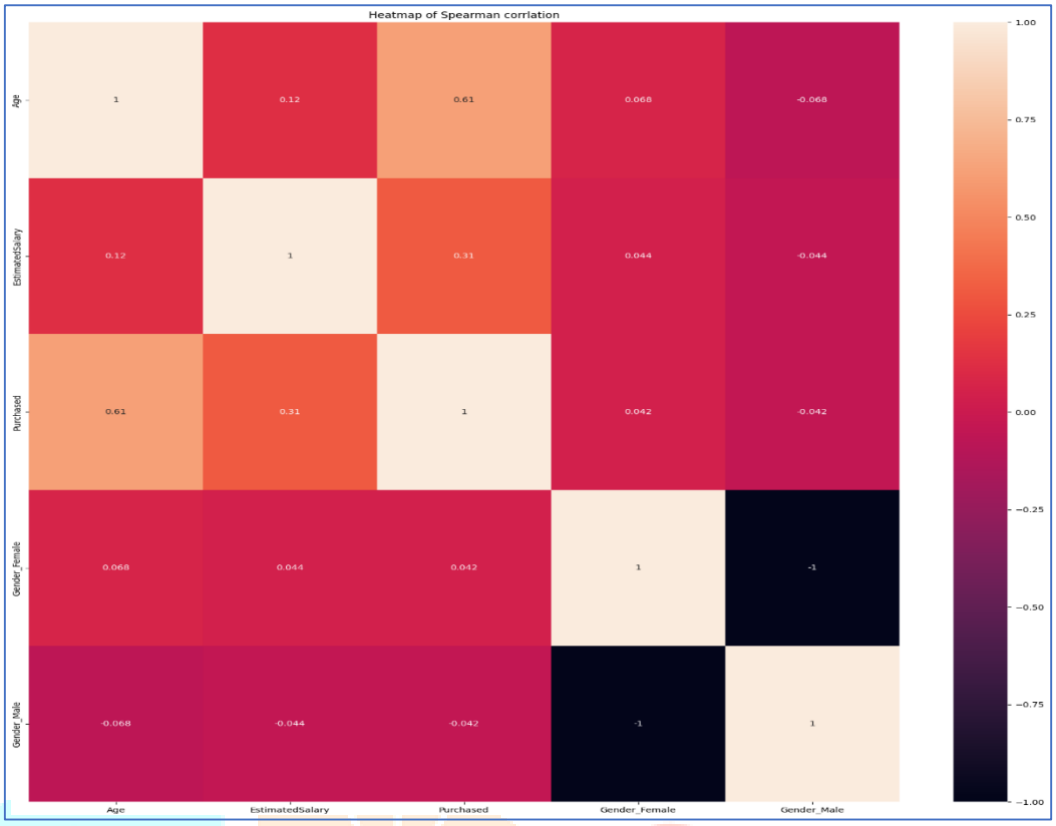

Figure 9: Heat map of Pearson Correlation

3.1 Modelling

In our research, we have to use multiple machine-learning models. The following Table 4 shows the score or accuracy of different models.

Table 4: Score of different models

	Score
DecisionTreeClassifier	0.841667
LogisticRegression	0.858333
XGBClassifier	0.866667
RandomForestClassifier	0.900000
AdaBoostClassifier	0.908333
KNeighborsClassifier	0.916667
GradientBoostingClassifier	0.916667
LGBMClassifier	0.916667
SVC	0.933333
XGBRFClassifier	0.933333

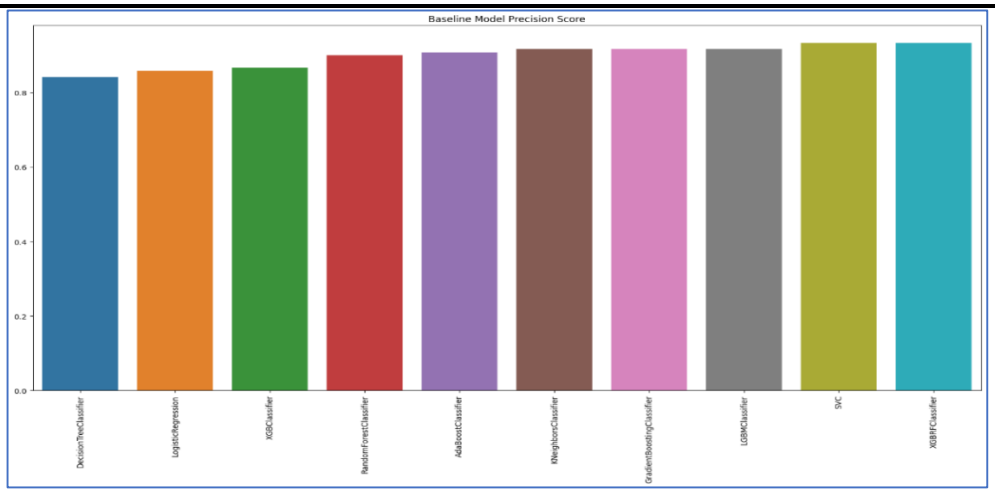

Figure 10: Baseline model precision score of models

From the baseline modeling, we can see that the top models are support vector machine and XGBoost, and the Random Forest ensemble learning generate the same type of score, 93.3%.

Table 5: Model comparison

Classifier	Score
SVC	93.3%
XGB and RF	93.3%

We can try tuning the hyperparameters to check if the model improves. As the data is imbalanced, we will use the F1 scores for the scoring.

3.2 Ensemble models

A. SVC and XGBoost

```
from sklearn.ensemble import VotingClassifier

svc=SVC(C=0.6, gamma=0.007000000000000001,kernel="rbf")
XGBRFClassifier=XGBClassifier(n_estimators=10,learning_rate=0.8,gamma=0.30000000000000004)
ensemble_model = VotingClassifier(estimators=[('svc', svc), ('XGBRFClassifier', XGBRFClassifier)], voting='hard')

ensemble_model.fit(X_train, y_train)

# Make predictions on the testing data
predictions = ensemble_model.predict(X_test)

# Calculate accuracy
accuracy = accuracy_score(y_test, predictions)
print(f"Accuracy: {accuracy}")

Accuracy: 0.7666666666666667
```

3.3 Comparative study

In this study, various machine learning algorithms were employed to predict fraudulent ad clicks. Additionally, ensemble learning techniques were utilized to enhance prediction performance. Table 6 presents a comparative analysis of the results. Among all the models tested, the Support Vector Machine (SVM) achieved the highest accuracy of 93.3%. The other individual models demonstrated accuracy levels ranging between 84% and 92%. For the ensemble methods, two different model combinations were

explored. The pairing of XGBoost and Random Forest also achieved an accuracy of 93.3%, while the combination of Support Vector Classifier (SVC) and XGBoost yielded a lower accuracy of 76.6%. Overall, the highest accuracy was observed with both the SVM model and the ensemble of XGBoost and Random Forest.

Table 6: Comparative study

S. No.	Name of Classifier	Score / Accuracy
1	SVC	93.3%
2	XGB and RF	93.3%
3	SVC and XGBoost	76.6%
4	Decision Tree	84.1%
5	Logistic Regression	85.8%
6	XGBoost	86.6%
7	Random Forest	90 %
8	AdaBoost	90.8%
9	KNN	91.6%
10	Gradient Boosting	91.6%
11	LGBM classifier	91.6%

4. CONCLUSION

The findings indicate that detecting ad click fraud in real time is achievable using basic machine learning classifiers, even when dealing with imbalanced datasets. Various machine learning models were evaluated, with Random Forest emerging as the most effective. The use of predictive modeling has not only improved click-through rates but also offers the advantage of learning patterns directly from the data, eliminating the need for manual feature engineering or domain expertise. One major challenge in analyzing ad click fraud is the scarcity of publicly available, accurately labeled datasets. This limitation restricts the scope of experimentation and research. Simulated ad click fraud, blended with genuine user interactions, can serve as a useful benchmark for testing detection strategies. Such simulations are especially relevant for fraud scenarios involving botnets or human click farms. Additionally, this process can expose a wide range of potentially sensitive user information, including HTTP headers, referrer data, and browser cookies.

REFERENCES

1. Reem A Alzaharani et al.,” AI-Based Techniques for Ad Click Fraud Detection and Prevention: Review and Research Directions”, J. Sens. Actuator Netw. 2023, *12*(1), 4;

2. Neeraja, Anupam, Sriram, Subhani Shaik, and V. Kakulapati,” Fraud Detection of AD Clicks Using Machine Learning Techniques”, Journal of Scientific Research and Reports, Volume 29, Issue 7, Page 84-89, June 2023.

3.Ch. Shravya, Pravallika, and Subhani Shaik,” Breast cancer prediction using machine learning Techniques”, International Journal of Innovative Technology and Exploring Engineering, Vol. 8, Issue 6, 2019.

4.P. Santosh and Subhani Shaik," Heart disease prediction with PCA and SRP", International Journal of Engineering and Advanced Technology, Volume-8, Issue-4, 2019.

5. Shiva Keertan J and Subhani Shaik," Machine Learning Algorithms for Oil Price Prediction", International Journal of Innovative Technology and Exploring Engineering, Volume-8 Issue-8, 2019.

6. KP Surya Teja, Vigneswara Reddy, and Subhani Shaik," Flight Delay Prediction Using Machine Learning Algorithm XGBoost", Journal of Adv Research in Dynamical & Control Systems, Vol. 11, No. 5, 2019.

7. Dr. Sunil Bhutada and Subhani Shaik, "IPL Match Prediction using Machine Learning", IJAST, Vol.. 29, Issue 5, April 2020.

8. Dr. R. Vijaya Kumar Reddy, Shaik Subhani, Dr. G. Rajesh Chandra, Dr. B. Srinivasa Rao," Breast Cancer Prediction using Classification Techniques", International Journal of Emerging Trends in Engineering Research, Vol. 8, No.9,2020.

9. Mr. Sujan Reddy, Ms. Renu Sri, and Subhani Shaik," Sentimental Analysis using Logistic Regression", International Journal of Engineering Research and Applications (IJERA), Vol. 11, Series-2, July 2021.

10. Ms. Mamatha, Srinivasa Datta, and Subhani Shaik," Fake Profile Identification using Machine Learning Algorithms", International Journal of Engineering Research and Applications (IJERA), Vol 11, Series-2, July 2021.