# It’s Fraud! Application of Machine Learning Techniques for Detection of Fraudulent Digital Advertising

**Snigdha Mathur**
Student, FORE School of Management, New Delhi.
E-mail: snigdha.gemini@gmail.com

**Sunita Daniel**
Associate Professor, Information and Technology Area, FORE School of Management, New Delhi.
E-mail: sunita@fsm.ac.in
https://www.fsm.ac.in/peoples/faculty/prof-sunita-daniel

*Received September 12, 2021; Accepted December 12, 2021*
*ISSN: 1735-188X*
*DOI: 10.14704/WEB/V19I1/WEB19166*

## Abstract

Due to the on-going COVID-19 pandemic, industries are heavily reliant on online meeting platforms. The pandemic has forced most MNCs to rely on online platforms such as Zoom or Microsoft Teams to hold daily business meetings and even conferences and other corporate events. Schools and other educational institutions have also been forced to conduct classes and all other events through these platforms. This increased unavoidable dependence on online platforms has resulted in an exponential increase in advertising fraud.

Advertising fraud is the application of any method or technology that hampers the proper delivery of advertisements to the proper audience or the proper place, or forcefully inserts advertisements at undesirable times or locations. This could take multiple forms and has become far more widespread with increased use of online platforms. Some common methods used for digital fraud can be the pay-per-click (PPC) model, domain spoofing or in the form of bots, but the main objective is to gain financial advantages from advertising transactions.

The primary objective of this study is to identify and understand the factors lying behind the presence of fraudulent activities on any online medium and to analyse the probability of downloading an application after coming across the online advertisement, and watching it. The study also aims to highlight how marketing agencies tend to float fraud advertisement just to gain more revenue from their end.

## Keywords

Conversion Fraud, CTR, Online Fraud, Marketing Agencies, Click Fraud.

## Introduction

Digital advertising in a wide spectrum can refer to the practice being followed by companies to deliver promotional content through various online and digital platforms. It uses online mediums like email, social media, search engines, mobile apps, and websites to grab the audience's attention. With the emergence and the upcoming trend of digital advertising, many companies have also seen a boom in advertising fraud. Advertising fraud occurs when companies and/or agencies pay to display ads on fraudulent websites or when advertising campaign data is influenced by click bots (Kanei & Fumihiro, 2020). Ad fraud is very hard to detect as differentiating between a legitimate customer and a bot is nearly impossible, without further analysis of information. A study on a survey conducted by Goggle Inc. in 2007 states that 10% of clicks on advertisements that has been floated over the internet in their AdWords program are not legitimate user clicks, which in turn translates to a loss of 1 billion USD yearly in terms of revenue for filtering out such fraudulent clicks (Zhu, Xingquan et al, 2017).

Through this study, we try to identify the crucial managerial factors underlying digital fraud and the factors that affect and further promote such activities over the internet. The proposed framework can predict whether a click generated is legitimate or fraudulent based on past data accumulated by the company – *Talking Data Ad Tracking Services.*

This research paper consists of 6 sections. The first section provides a literature review which helps in understanding the lack of previous literature pertaining to this particular problem and highlights the importance of the study. The second section describes the proposed system, leading to the third section describing the technique used for predicting fraud. The fourth section contains the results obtained, followed by the conclusion in the fifth section. The last section describes the limitations of the study due to the dataset and tools applied.

## Literature Review

Conversion fraud can be described as one of the main issues that has developed vastly over the last few years all over the internet, and is a major form of digital ad fraud. Between March 2019-2020 and March 2020-2021, a 28.32% increase was noted in terms of conversion fraud. A study that utilized bluff ads to combat click fraud found that fake ads help confirm the legitimacy of the individual generating the click, since they require more effort (Haddadi &Hamed, 2010). The most common type of fraud present in the online industry is click fraud, which refers to marketers or agencies displaying fake ad clicks to direct the potential customer to a fraud site.

A study exploring the relationship between the number of clicks and the conversion rates per advertisement was conducted by applying Heuristics and MLP structures. These techniques helped in better understanding the possible ways of online fraud and how to tackle it carefully. The research informs about the strong relationship between the clicks and conversion rates, which can be better deciphered using modern methods (Srivastava, 2020). Delving into the impact of various online media channels, such as email marketing, mobile phone marketing, optimization for search engines and corporate websites, and social media marketing, the effectiveness of online digital media advertising has revealed, that by finding the moderation and mediation elements that have a significant impact on the external and internal elements of online advertising, brand sustainability and repeat purchases can be improved.

An analytical framework using Recursive Feature Elimination (RFE) and classification through Hellinger Distance Decision Tree (HDDT) has been previously proposed to study advertisement fraud (Taneja, Mayank et al, 2015). An automated feature engineering-based approach to drastically reduce the number of false-positive in fraud prediction has also been presented, using *Deep Feature Synthesis* algorithm to derive behavioural features based on the historical data of the card associated with the online transaction (Wedge, Roy et al, 2018).

The result of the study focusing on the point of view of online marketing professionals towards Online Advertising Fraud (OAF), through interviews, suggested that effective fraud detection is often hampered by ineffective measurement techniques and unreasonable customer expectations (Dornyei, Krisztina Rita 2020). A study has also described the level of fraud that is associated with illegitimate online clicks on pay-per-click (PPC) advertisements (Kshetri, Nir& Jeffery, 2019). Since Google and Facebook are often accused of using a secretive technique to detect invalid clicks and thus identifying the probable fraud that can occur, a third party working independently is often hired to accurately measure the online advertising delivery as warranted.

Although multiple studies have previously focused on identifying and alleviating digital fraud, none of them have implemented machine learning techniques to derive managerial insights from the dataset of conversion fraud in digital advertising. Hence our study proposes a novel framework comprising of machine learning techniques along with graphical representation to efficiently identify and predict fraudulent advertisements from the dataset.

## Proposed Framework

We have developed a framework for the prediction of conversion fraud in digital advertisements. The framework consists of Machine Leaning techniques like Random Forest, XGBoost, and Neural Networks with activation function models to predict conversion fraud. Along with machine learning, data visualization techniques have also been applied to explore the data thoroughly and recognize the underlying factors in digital fraud.

## Data Visualization

Data visualization is the process of extracting valuable insights from the dataset in a graphical or pictorial manner. Commonly used graphs like bar, column, and line help in depicting the trends that are followed in the dataset. The representation of the data in a given dataset or any information in the form of a graph, chart or any other visual format, helps in analysing trends and patterns.

The various types of data visualisation commonly used are:

- Heatmap – uses a graph with numerical data points highlighted in light and warm colour patterns to signify whether the variables in the data are highly correlated or low correlated with the target variable.
- Histogram – used to measure frequencies or distributions of the numerical data over a range of values.
- Bar Plot – This plot shows the relationship between a numeric and a categorical variable. Each entity of the categorical variable is represented as a bar. The size of the bar represents its numeric value.
- Line plot - A line plot displays the data on the graph as points on a number line usually depicting the frequency of each data point.
- Box plot - The box plot can help in identifying the presence of outliers in the dataset. Outliers in the dataset can hamper the accuracy levels achieved in the further machine learning algorithms.

## Machine Leaning Techniques

Machine Learning algorithms are used to predict patterns in a dataset and test the accuracy of the predicted data. To test the accuracy of the data, the data is split into 2 parts: training data and test data. The training data is used to train the dataset so that the predictive model can be built, and the test data is used to validate the model by comparing the predicted target variable to the trained dataset. The accuracy of predicted dataset is checked and calculated. The function *test_train_split*is used for splitting the data in random ratios in order to present a fair accuracy reading (Tan, Jimin et al, 2021).

Machine Learning comprises of vast types of techniques and algorithms used for various purposes on the dataset. One such algorithm, Logistic Regression has been applied on the fraud detection dataset. Logistic Regression is a type of supervised learning classification algorithm that is used to predict the probability of the target variable in the data. The nature of target variable in logistic regression is dichotomous (only 2 possible outcomes in the variable) (Hosmer, David W & Stanley, 1980).In fraud detection dataset, the target variable *is_attributed* comprises of 0s and 1s depicting if the app has been downloaded or not.

Decision Tree is a type of supervised machine learning algorithm with its main goal to predict the target variable from the dataset supplied (Bonaccorso Giuseppe, 2017). For prediction, the decision tree uses a tree like representation with parent and leaf nodes corresponding to class label and the attributes associated with it. It is a flowchart-like overall structure in which each internal node of the tree depicts a ‘test’ on a particular feature whereas each leaf node depicts a class label (a decision taken that needs to be performed after computing all the features in the data).

Relating to our dataset, a decision tree would be able to divide the data based on the various features like IP address or app address relating all these features to the target variable i.e., *is_attributed*. The accuracy using a decision tree is calculated based on the features and the division of data into respective nodes and leaves. The accuracy of a decision tree can be increased by either disregarding the missing and the outlier values or by using more techniques like feature engineering to gather and assemble the data into a sub-set based on the features been identified by the model.

Random Forest Classifier is a collection of large number of decision trees which operate as an individual in order to yield a better prediction (Bonaccorso Giuseppe, 2017). In random forest classifier, the main variable is to identify the number of estimators (the number of distinct decision trees to be considered for prediction). Every decision tree formed in a random forest algorithm determines a class prediction and the class with the highest votes in terms of target variable becomes the overall prediction of the model.

The performance of proposed system is evaluated using accuracy as popular metric that refers to the ability of system to correctly predict the class label of new or unseen data. Accuracy is computed using the following formula:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (1)$$

True positives (TP) = number of correct classifications predicted as yes (or positive).
True negatives (TN) = number of correct classifications predicted as no (or negative)
False positive (FP) = number of data points that are incorrectly predicted as yes (or positive) when it is actually no (negative).
False negative (FN) = number of data points that are incorrectly predicted as no when it is actually yes

A Neural Network can be described as a web of interconnected nodes where every node is solely responsible for computations. A neural network can be compared to a human brain consisting of neurons (Albawi, Saad, Tareq Abed Mohammed &Saad Al-Zawi, 2017). These neurons assign the algorithm in a way that they are responsible for making accurate decisions without needing external data for verification, whereas in other Machine learning algorithms, decisions are primarily made according to the past data that the model encountered.

Neural networks are broadly divided into 3 types:

- ANN (Artificial Neural Networks)
- CNN (Convolution Neural Networks)
- RNN (Recurrent Neural Networks)

ANN can be defined as a group of multiple perceptron or neurons present together at each layer. It processes data only in the forward direction and consists of 3 layers: Input, Hidden and Output. Input layer is responsible for accepting the input, the hidden layer assesses the input and the output layer produces the final output (Fig. 1).

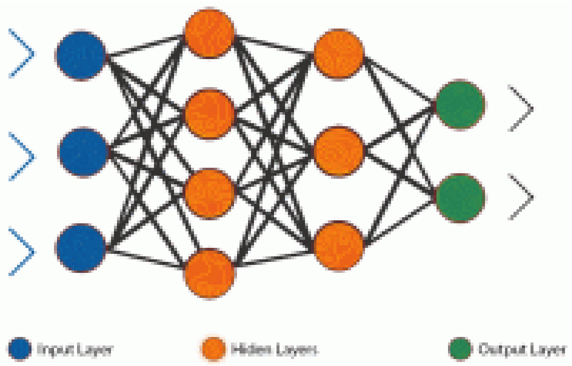


**Fig. 1 The figure explains the ArtificialNeural Network [Ijaz Khan, 2021]**

CNN is a popular method of deep learning, as it is used in image extraction and video processing projects. In this, kernels are used to extract the relatable features of the audio or video from the input and results are filter to create a feature map.

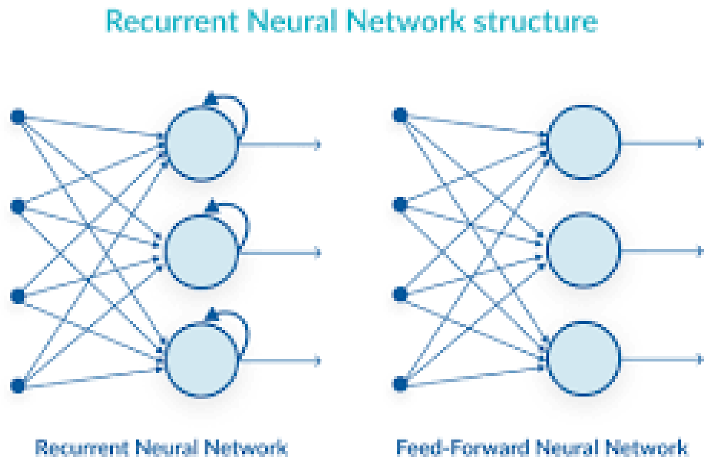

Recurrent Neural Network structure
Recurrent Neural Network
Feed-Forward Neural Network

**Fig. 2 The figure explains the Recurrent Neural Network (Ijaz Khan, 2021)**

RNN is similar to ANN but has a recurrent connection in the hidden layer of the network. This loop at the hidden layer ensures that the information captured is sequential in its nature as in the input layer (Fig.2). RNN is most used in time-series, test and audio data.

In our study for predicting conversion fraud, we have used LSTM (Long-Short-Term Memory) network as this method is used to train the time series data. This network has a similar flow as a RNN as both process data by passing the data only in forward direction. The only difference created is between the cell of LSTM as they allow the LSTM to keep or forget the information based on the prediction been made by the layers.

Other activation functions which are responsible for transforming the overall average input from the starting mode into the output mode are: *ReLu* and *Sigmoid*.

*ReLu* (Rectified Linear activation function) is a linear model that either displays the output in positive if passed directly or negative otherwise. It is a commonly used activation function because it's easy to train and results in more precise accuracy of the data / layers in neural networks. Mathematically, *ReLu* is defined as

$$f(x) = \max(0, x) = \begin{cases} x_i \; if \; x_i > 0 \\ 0 \; if \; x_i < 0 \end{cases} \qquad (2)$$

The main aim of the *ReLu* activation function is to undergo a threshold operation to every input in the neural network where all values less than 0 are assigned as 0.

*Sigmoid* activation function is a non-linear function and transforms the input values between the range of 0 to 1. Sigmoid function produces a ‘S’ shaped curve and is mathematically given by the formula

$$f(x) = \left(\frac{1}{1+ exp^{-x}}\right) \qquad (3)$$

## Experimental Method

### Dataset Description

The dataset was collected from the *TalkingData* company- a large independent Big Data platform that includes industry leading solutions ranging from mobile apps and gaming analytics to mobile ad tracking. The dataset is real-time data gathered throughout the month of November in 2020 and can be found in Kaggle: https://www.kaggle.com/c/talkingdata-adtracking-fraud-detection

The dataset contains over 1,00,000 entries with 8 columns (details in Table 1). *Is_attributed* is the target variable used in prediction of conversion fraud while all other variables are used in data visualization to extract key insights from any past trends observed in the dataset.

**Table 1 Details of features explored from the attributes in the dataset**

| S No. | Attribute | Features |
|---|---|---|
| 1. | IP address | VarIP<br>Ip/click |
| 2. | Channel Id | VarChannel<br>Channel/click |
| 3. | Os Id | VarOS<br>Os/click |
| 4. | App Id | VarApp<br>App/click |
| 5. | Device Id | VarDevice<br>Device/click |
| 6. | Click_time | Avg Click Per Hour<br>Avg Click Per Day<br>Most Clicks |
| 7. | Attributed_time | Time Of Click<br>Majority Of Clicks |
| 8. | Is_attributed | Yes/No (download) |

## Experimental Framework

The target variable (*is_attributed*) was found to be highly imbalanced in a ratio of 99760:227 where the number of fraudulent activities occurred was practically one-fourth of the non-fraudulent activities. Since an imbalanced dataset can further lead to false and inaccurate predictions while executing machine learning algorithms on the predictor and the target variables, the dataset needed to be balanced. SMOTE (Synthetic Minority Oversampling Technique) method was applied to balance the dataset. Using SMOTE, the value count of the target variable was distributed equally with 69840 in each category. This then allowed for extraction of the precise features and factors behind conversion fraud. Further, machine learning algorithms were applied to the target variable by dividing the dataset in a ratio of 70% for training and 30% for testing.

Random Forest using Feature Importance was used to identify the importance of each feature in deriving the accuracy of the prediction and then the features were ranked.

All machine learning and neural network techniques were coded in Jupyter platform using Python 3 programming language.

## Results and Discussions

The data exploration and visualisation of the dataset enabled to derive many valuable insights which can be implemented in a managerial aspect related to online fraud.

Fig. 3 shows the correlation graph using a heatmap to depict the levels of correlation between the variables in the dataset.

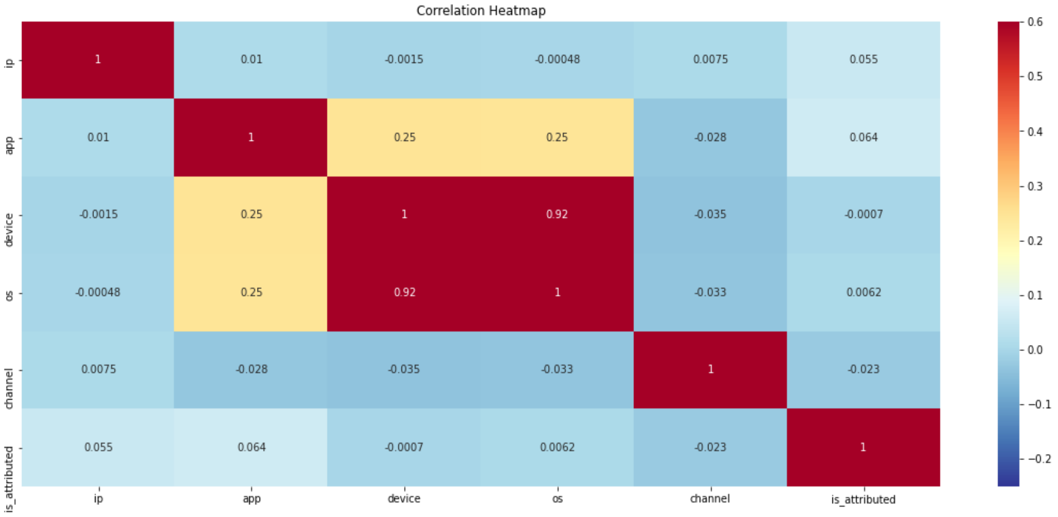


**Fig. 3 Correlation heatmap for the dataset**

We can see that the variable '*Device*' and '*OS*' have a very strong positive correlation (0.92) as the OS version of any electronic gadget actually depends on the device being used by the customer. Apart from this, '*App id*' has also shown a positive correlation with 'device' and 'os' (Fig 3).

This means that varying the "*device*" or the "*OS id*" of the gadget of the customer, would most likely change the "*app id"* assigned to the customer. The target variable (*is_attributed*) is mostly dependent on the variable "app id" and the "*IP address*" of the customer. This implies that the agency majorly targets the IP address of the customer in order to increase the probability of the app being downloaded by the customer.

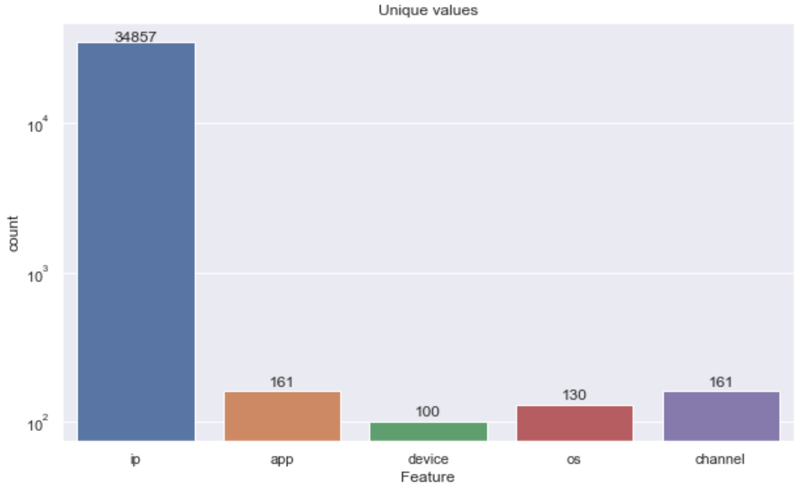


**Fig. 4 Bar graph depicting the unique values in dataset**

Fig. 4 shows the count of unique values present in all variable columns in the dataset. The largest count of unique values is in the column of *"IP address"* because every household or Wi-Fi Router has a separate IP address attached to it. The variables "*app"* and "*channel"* have 161 unique values whereas there are only 100 unique values for "*device"*. From this graph we get to know that the variable *"device"* is the most redundant variable as it only has 100 unique values present in the dataset.

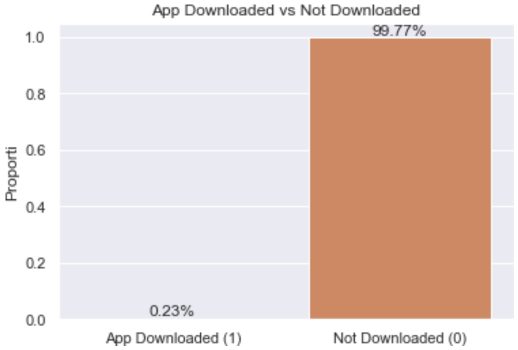
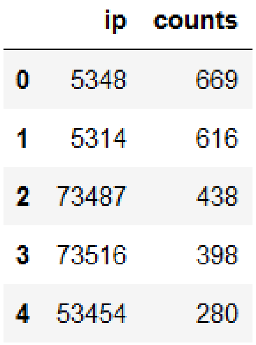


**Fig. 5 Bar graph for percentage of app downloaded**

The bar graph in Fig. 5 shows that only 0.23% of the online advertisements were considered as legitimate with no fraudulent activity involved, as only 0.23% of the population in the dataset have downloaded the app. 99.77% of the times, the app was not downloaded (fraud website or fraud advertisement). This depicts that the majority of customers tend to think that online advertisement involves fraud sites as only about 0.2% have clicked on the website and downloaded the app. The main reason behind this can be the massive increase in spam messages and emails being floated by marketing agencies hence making the customers distrust such advertisements.



|   | ip | counts |
|---|---|---|
| 0 | 5348 | 669 |
| 1 | 5314 | 616 |
| 2 | 73487 | 438 |
| 3 | 73516 | 398 |
| 4 | 53454 | 280 |

**Fig. 6 Table for IP addresses with their corresponding counts**

One main trend identified, was that each marketing agency has an individual IP address which is used for sending out advertisements en masse. The IP address in most online fraud activities was found to be similar i.e., only a specific set of IP addresses belong to each marketing agencies as changing IP address for each advertisement floated in the internet would lead to high operational cost to the agency. In our dataset, most of advertisements belonged to the IP address of '5348' with a count of 669 and most of the ads from these IP addresses were considered to be fraud. (Fig. 6)

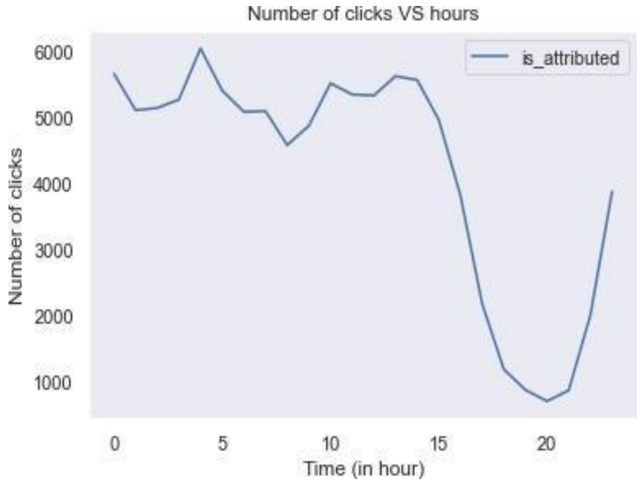


**Fig. 7 Line graph for depicting the hour of the day with number of clicks**

The other observation made during the data analysis (Fig.7) was that the time of receiving the advertisement via mail or message played an important role in determining whetherit was fraudulent or not. This is because a majority of agencies target customers during their free time (conventionally during office and lunch breaks).

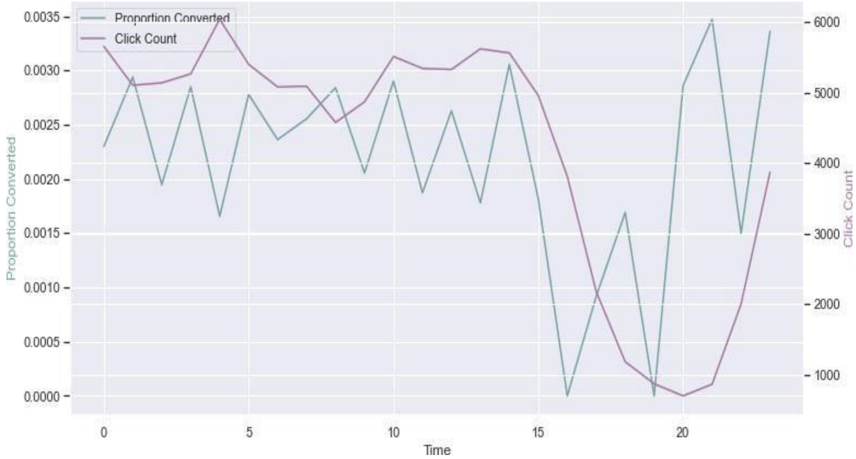
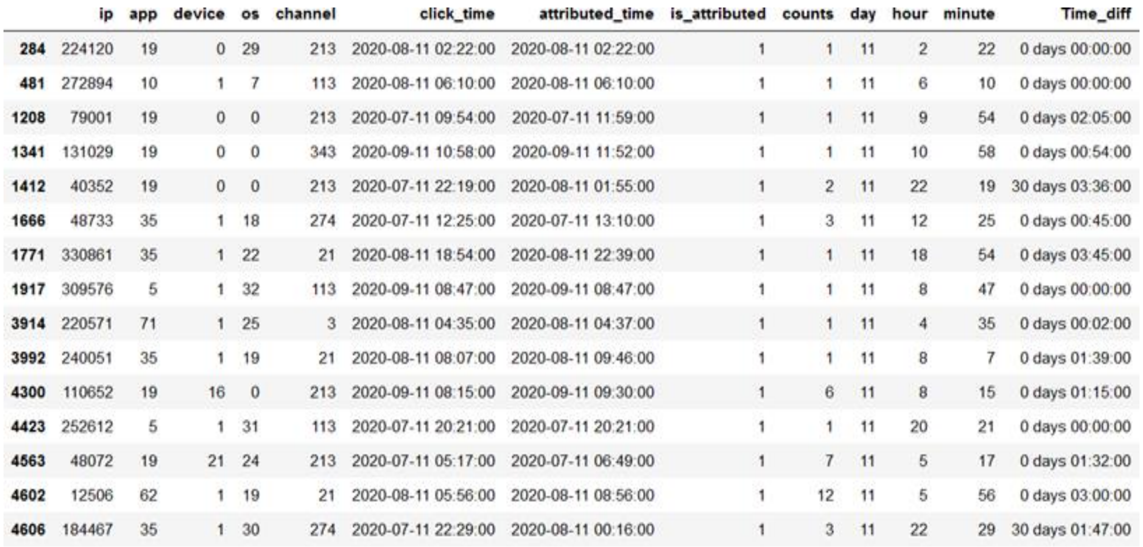


**Figure 8 Combined graph for hourly conversion ratio and hourly click frequency**

It was also found that the hourly conversion ratio and hourly click frequency follow a similar pattern (Fig. 8). This indicates that the conversion ratio for any application is dependent on the clicks per app, as conversion ratio increases when the click frequency per advertisement increases.

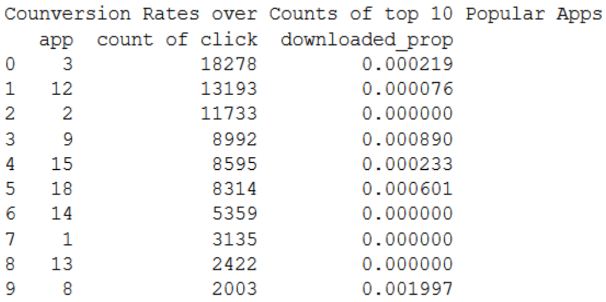


| | ip | app | device | os | channel | click_time | attributed_time | is_attributed | counts | day | hour | minute | Time_diff |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **284** | 224120 | 19 | 0 | 29 | 213 | 2020-08-11 02:22:00 | 2020-08-11 02:22:00 | 1 | 1 | 11 | 2 | 22 | 0 days 00:00:00 |
| **481** | 272894 | 10 | 1 | 7 | 113 | 2020-08-11 06:10:00 | 2020-08-11 06:10:00 | 1 | 1 | 11 | 6 | 10 | 0 days 00:00:00 |
| **1208** | 79001 | 19 | 0 | 0 | 213 | 2020-07-11 09:54:00 | 2020-07-11 11:59:00 | 1 | 1 | 11 | 9 | 54 | 0 days 02:05:00 |
| **1341** | 131029 | 19 | 0 | 0 | 343 | 2020-09-11 10:58:00 | 2020-09-11 11:52:00 | 1 | 1 | 11 | 10 | 58 | 0 days 00:54:00 |
| **1412** | 40352 | 19 | 0 | 0 | 213 | 2020-07-11 22:19:00 | 2020-08-11 01:55:00 | 1 | 2 | 11 | 22 | 19 | 30 days 03:36:00 |
| **1666** | 48733 | 35 | 1 | 18 | 274 | 2020-07-11 12:25:00 | 2020-07-11 13:10:00 | 1 | 3 | 11 | 12 | 25 | 0 days 00:45:00 |
| **1771** | 330861 | 35 | 1 | 22 | 21 | 2020-08-11 18:54:00 | 2020-08-11 22:39:00 | 1 | 1 | 11 | 18 | 54 | 0 days 03:45:00 |
| **1917** | 309576 | 5 | 1 | 32 | 113 | 2020-09-11 08:47:00 | 2020-09-11 08:47:00 | 1 | 1 | 11 | 8 | 47 | 0 days 00:00:00 |
| **3914** | 220571 | 71 | 1 | 25 | 3 | 2020-08-11 04:35:00 | 2020-08-11 04:37:00 | 1 | 1 | 11 | 4 | 35 | 0 days 00:02:00 |
| **3992** | 240051 | 35 | 1 | 19 | 21 | 2020-08-11 08:07:00 | 2020-08-11 09:46:00 | 1 | 1 | 11 | 8 | 7 | 0 days 01:39:00 |
| **4300** | 110652 | 19 | 16 | 0 | 213 | 2020-09-11 08:15:00 | 2020-09-11 09:30:00 | 1 | 6 | 11 | 8 | 15 | 0 days 01:15:00 |
| **4423** | 252612 | 5 | 1 | 31 | 113 | 2020-07-11 20:21:00 | 2020-07-11 20:21:00 | 1 | 1 | 11 | 20 | 21 | 0 days 00:00:00 |
| **4563** | 48072 | 19 | 21 | 24 | 213 | 2020-07-11 05:17:00 | 2020-07-11 06:49:00 | 1 | 7 | 11 | 5 | 17 | 0 days 01:32:00 |
| **4602** | 12506 | 62 | 1 | 19 | 21 | 2020-08-11 05:56:00 | 2020-08-11 08:56:00 | 1 | 12 | 11 | 5 | 56 | 0 days 03:00:00 |
| **4606** | 184467 | 35 | 1 | 30 | 274 | 2020-07-11 22:29:00 | 2020-08-11 00:16:00 | 1 | 3 | 11 | 22 | 29 | 30 days 01:47:00 |

**Fig. 9 Table showing the time difference between clicking and downloading the app**

On an average a customer took around 24 hours to download the app after going through the online advertisement link whereas some of the users downloaded the app even after 30 days of clicking on the website link (Fig. 9).

This indicates that a customer probably makes an informed decision on whether to download the app or not by taking the time and to learn and know about the app and its features, rather than downloading it on sudden impulse. The inference from the above observation can be that marketers should concentrate more towards highlighting the special attributes of the application as these help the customer in making an informed decision, increasing their probability of downloading the app.

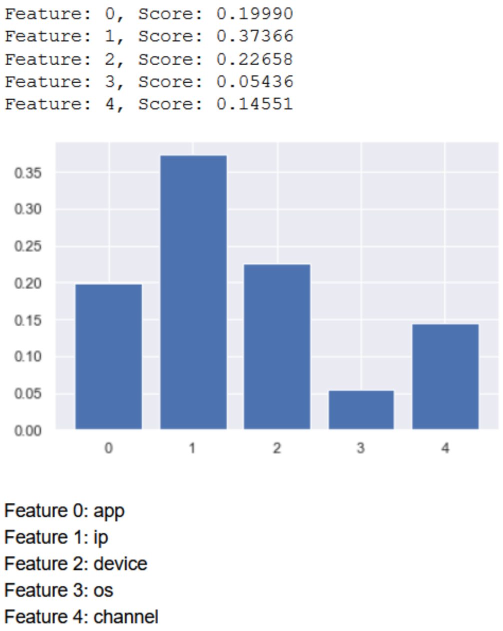


```
Counversion Rates over Counts of top 10 Popular Apps
   app  count of click  downloaded_prop
0    3           18278         0.000219
1   12           13193         0.000076
2    2           11733         0.000000
3    9            8992         0.000890
4   15            8595         0.000233
5   18            8314         0.000601
6   14            5359         0.000000
7    1            3135         0.000000
8   13            2422         0.000000
9    8            2003         0.001997
```

**Fig. 10 Conversion rates over counts of 10 most popular apps**

There is a significant difference in clicks per app (Fig. 10), with minimum of one click on an app and maximum at almost 19000. Apps with higher number of clicks also have the highest download proportion. The proportion fluctuates more as the counts go down, since each additional click has larger impact on the proportion value.

The machine learning technique – Random Forest along with feature importance produced the following result.



```
Feature: 0, Score: 0.19990
Feature: 1, Score: 0.37366
Feature: 2, Score: 0.22658
Feature: 3, Score: 0.05436
Feature: 4, Score: 0.14551
```

Feature 0: app
Feature 1: ip
Feature 2: device
Feature 3: os
Feature 4: channel

**Fig. 11 Feature Importance in Random Forest Classifier**

Through this method, it was found out that the variable IP address of the advertisement showed the highest score (Fig. 11). This means that IP address variable holds the highest importance while determining the accuracy of the model. The results obtained from the feature importance confirms the data visualisation results which also depicted that IP address is the most crucial factor in determining the causes of data fraud.

Table 2 shows the accuracy rate for all the machine learning algorithms used in our study. The comparison table helps in understanding the accuracy score for all the machine learning techniques applied at the dataset.

**Table 2 Comparison of accuracy obtained by ML algorithms**

| **ML Algorithm** | **Accuracy** |
|---|---|
| Logistic Regression | 73% |
| Decision Tree | 92% |
| Random Forest (with feature importance) | 97% |
| Neural Network using Relu and Sigmoid (before feature importance) | 94.01% |
| Neural Network using Relu and Sigmoid (after feature importance) | 99.74% |
| Neural Network RNN using LSTM (after feature importance) | 91.85% |

The machine learning techniques displayed a high accuracy level in all the models. The highest accuracy level was achieved by the neural networks using *ReLu* and *Sigmoid* model 99.74% followed by the random forest classifier with a percentage of 97%. The accuracy found using the LSTM model was very less compared to the other models.

Implementing the feature importance in the random forest classifier was unable to increase the efficiency of the models. This would mean that all the variables (Ip address, App Id, Channel Id, Device Id and the time) played a significant role in predicting the probability of conversion fraud, and therefore all had to be considered to obtain an accurate model.

## Conclusion

Through this study on the underpinning factors of conversion fraud in digital advertising, one gets to know how agencies and marketers try to deceive the customers into downloading their application via online advertising. This study also helps to draw insight into the perspective behind online fraud and could identify the major factors that marketing agencies use in order to gain the number of clicks per advertisement. The idea behind the pay-per-click model in advertisement was understood through the research. Fraudulent activities and downloads in return help marketing agencies to earn an increased income as they are able to portray that people are invested in the advertisement and want to buy the product being advertised.

In conclusion, it can be stated that through this method and analysis, assumptions regarding online fraud were studied. Through the dataset provided, the probability of downloading the app was studied to identify whether the presence of online advertisement on any digital media platform. The main objective of the project was to identify the critical variables that are responsible for conversion fraud.

## Limitations

The limitation with respect to the dataset is the lack of analysing and gathering data on a larger scale. The existing dataset comprises of entries from a very limited and closed circle of audience. Collecting the exact data from a wide horizon of audience varying in terms of area, age and profession can help in understanding the exact motivations behind marketing agencies and how they progress forward in targeting the audience.

Another limitation with the research conducted could be the inability to perform detailed analysis using other SPSS techniques for validating the data and its consistency.

## Acknowledgments

The infrastructural support provided by FORE School of Management, New Delhi in completing this paper is gratefully acknowledged.

## References

Ahmed, R.R., Streimikiene, D., Berchtold, G., Vveinhardt, J., Channar, Z.A., & Soomro, R.H. (2019). Effectiveness of online digital media advertising as a strategic tool for building brand sustainability: Evidence from FMCGs and services sectors of Pakistan. *Sustainability, 11*(12).

Albawi, S., Mohammed, T.A., & Al-Zawi, S. (2017). Understanding of a convolutional neural network. *In International Conference on Engineering and Technology (ICET),* 1-6.

Bonaccorso, G. (2017). *Machine learning algorithms.* Packt Publishing Ltd.

Dörnyei, K.R. (2020). Marketing Professionals' Views on Online Advertising Fraud. *Journal of Current Issues & Research in Advertising,* 1-19.

Khan, I. (2021). *From ANNs (Artificial Neural Networks) to RNNs (Recurrent- Neural Networks).*

Haddadi, H. (2010). Fighting online click-fraud using bluff ads. *ACM SIGCOMM Computer Communication Review, 40*(2), 21-25.

Hosmer, D.W., & Lemesbow, S. (1980). Goodness of fit tests for the multiple logistic regression model. *Communications in statistics-Theory and Methods, 9*(10), 1043-1069.

Kanei, F., Chiba, D., Hato, K., Yoshioka, K., Matsumoto, T., & Akiyama, M. (2020). Detecting and Understanding Online Advertising Fraud in the Wild. *IEICE Transactions on Information and Systems, 103*(7), 1512-1523.

Kshetri, N., & Voas, J. (2019). Online advertising fraud. *Computer, 52*(1), 58-61.

Srivastava, A. (2020). Real-time Ad Click Fraud Detection.

Tan, J., Yang, J., Wu, S., Chen, G., & Zhao, J. (2021). *A critical look at the current train/test split in machine learning.* arXiv preprint arXiv:2106.04525.

Taneja, M., Garg, K., Purwar, A., & Sharma, S. (2015). Prediction of click frauds in mobile advertising. *In Eighth International Conference on Contemporary Computing (IC3)*, 162-166.

Wedge, R., Kanter, J.M., Veeramachaneni, K., Rubio, S.M., & Perez, S.I. (2018). Solving the false positives problem in fraud prediction using automated feature engineering. *In Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 372-388.

Zhu, X., Tao, H., Wu, Z., Cao, J., Kalish, K., & Kayne, J. (2017). Ad fraud categorization and detection methods. *Fraud Prevention in Online Digital Advertising*, 25-38.