Ad Click Guard: Enhanced Ad Click Fraud Detection in Mobile Advertising

DIC PROJECT PHASE 1:

Team Members :
Sujith Gollamudi
Bahalul Khan Pathan
Sai Krishna Aditya Gorthi

Problem Statement:

Click fraud is a prevalent issue in the online advertising industry, leading to misleading click data and wasted resources. TalkingData, China's largest independent big data service platform, handles an overwhelming volume of mobile ad clicks, a significant portion of which is potentially fraudulent. The current approach involves measuring user click journeys and flagging IP addresses with suspicious click behaviour. However, to stay ahead of fraudsters, TalkingData seeks a predictive model to forecast app downloads after mobile ad clicks.

Background:

With China being the largest mobile market globally, and TalkingData covering over 70% of active mobile devices, the scale of fraudulent traffic is enormous. This poses a significant challenge for app developers relying on accurate click data for advertising decisions. The existing preventive measures, such as IP and device blacklists, are effective but not fool proof. The challenge is to develop an advanced algorithm that can predict whether a user will download an app after clicking a mobile ad, thereby enhancing the effectiveness of fraud detection.

Objectives:

1. Build a predictive model to identify users likely to download an app after clicking a mobile ad.
2. Improve fraud detection efficiency beyond the current blacklist approach.
3. Enhance the accuracy and reliability of click data for app developers.

Significance:

Mitigate financial losses due to click fraud for app developers and advertisers.
Provide a proactive solution to stay ahead of evolving click fraud strategies.
Contribute to a more reliable and trustworthy online advertising ecosystem.

2. Data Sources

The dataset for this project has been sourced from **Kaggle's TalkingData AdTracking Fraud Detection competition**. The dataset consists of click records, including features such as IP address, app ID, device type, OS version, channel ID, click timestamp, attributed time (if app was downloaded), and the target variable indicating app downloads.

Dataset Source: TalkingData Ad Tracking Fraud Detection - https://www.kaggle.com/competitions/talkingdata-adtracking-fraud-detection/data

Data Cleaning/Processing:

In our project we performed various Data Cleaning or Processing Techniques

1. Balancing Data :
   There is a class imbalance in the dataset where out of 185 million records only 0.23% of data is fraudulent, so to handle this imbalance in the dataset we have done data resampling, This involves reducing the number of instances of the majority class, typically by randomly removing instances, It aims to balance the class distribution by preventing the model from being overwhelmed by the abundance of majority class examples.

2. Handling missing values :
   In our dataset, the column 'attributed_time' contains a significant number of missing values. To address this issue and ensure the integrity of our analysis, we opted to handle the missing values by excluding the 'attributed_time' column from our dataset. The rationale behind this decision is based on the observation that a substantial portion of 'attributed_time' entries is missing, and imputing these values may introduce bias or inaccuracies into our analysis. The decision to drop the 'attributed_time' column is justified by the recognition that imputation might introduce unintended biases, and the missing values in this context do not contribute significantly to the analysis. The resultant dataset is now more suitable for downstream modelling and exploratory data analysis.

3. Removing Unwanted Columns:

   In the pre-processing stage of our dataset, a strategic decision was made to enhance the clarity and efficiency of subsequent analyses by removing columns that contribute minimally or not at all to the analysis and modelling objectives. Unwanted columns, which were deemed irrelevant or redundant for the intended analysis, were systematically excluded from the dataset. This process of column removal aims to streamline the dataset, making it more focused and conducive to meaningful insights during exploratory data analysis and subsequent modelling.

4. Datatype Checking/Conversion:

   Since the dataset is from Kaggle competition , it was observed that categorical variables, including IP addresses, have been obfuscated by mapping them to numerical values. Consequently, the default datatypes in the dataset are integers. Given this mapping strategy and the inherent numerical representation of the categorical variables, there was no necessity to perform explicit datatype conversion. The dataset's natural representation aligns with the Kaggle competition specifications, eliminating the need for further conversion to numerical datatypes.

5. Feature Scaling:

   In the context of the Kaggle competition dataset, a crucial observation was made regarding the predominant categorical nature of the variables. As the majority of these variables are essentially categorical features mapped to numerical values, the conventional need for feature scaling is obviated. Feature scaling, which is often imperative in datasets with numerical variables of varying scales, becomes unnecessary here. Applying feature scaling to categorical variables could potentially distort the meaningful information encoded in the categorical mappings. Therefore, in the Kaggle dataset scenario, the decision was made to forego feature scaling, preserving the inherent categorical nature of the variables and avoiding unintended alterations to the data's semantics.

6. Dropping Duplicates:-

   In the process of refining our dataset, a critical step involved the identification and removal of duplicate records. Duplicate records, if present, can introduce redundancy and distort the accuracy of subsequent analyses. To enhance the effectiveness of the dataset and ensure the integrity of our findings, duplicates were systematically dropped. This strategic decision not only streamlines the dataset by eliminating redundant information but also promotes accuracy in downstream analyses.

Exploratory Data Analysis (EDA):

EDA is a critical step in our data analysis process, We have conducted various EDA operations on our dataset:

1. **Exploratory Data Analysis (EDA) Step 1: Count of Clicks per IP**

   In this EDA Univariate Technique, we sought to gain insights into the distribution of clicks across different IP addresses in the dataset. The Pandas library was employed to calculate the click counts for each unique IP address. The resulting counts were then structured into a tabular format using a Data Frame. This tabular presentation, displayed below, provides a clear overview of the frequency of clicks associated with each unique IP. The 'IP' column represents the unique IP addresses, and the corresponding 'Count' column indicates the number of clicks attributed to each IP.

This analysis serves as a fundamental exploration into the click patterns across IP addresses, offering valuable information regarding potential sources of click activity. Such insights are crucial for understanding the distribution of user engagement and identifying patterns that may influence the likelihood of app downloads after clicks.

2. **Exploratory Data Analysis (EDA) Step 2: App Distribution Analysis**

   As part of our comprehensive Exploratory Data Analysis (EDA), we delved into the distribution of mobile applications ('app') within the balanced dataset. Utilizing Pandas, we computed the value counts for each unique 'app' ID, providing insights into the frequency of each application. Subsequently, we transformed these counts into a structured Data Frame, 'app_counts_table,' presenting a tabular view with columns 'App ID' and 'Count.' To facilitate a thorough examination, we utilized Pandas's option context to display all rows of the table. This analysis offers a nuanced understanding of the distribution of app occurrences, a crucial step in unravelling patterns and characteristics within our dataset.

3. **Exploratory Data Analysis (EDA) Step 3 : Device Counts**

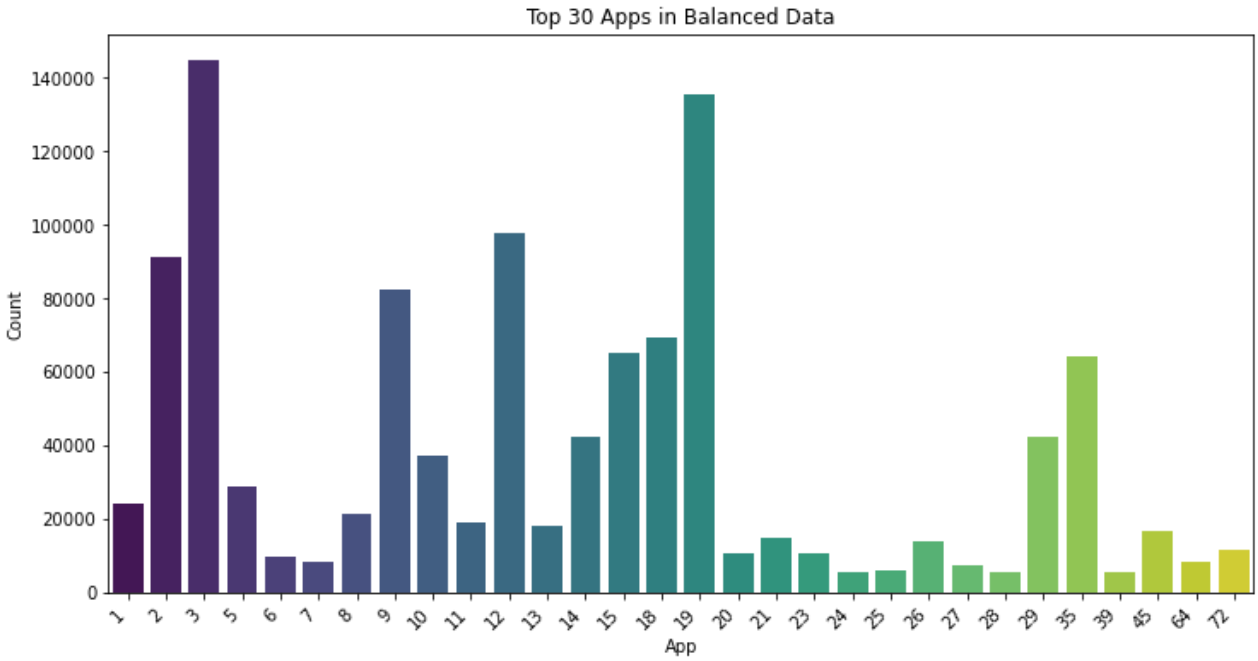
   This EDA step provides crucial insights into the diversity and prevalence of different mobile devices interacting with the ad platform. The tabulated device counts reveal the frequency at which each unique device ID appears in the dataset, shedding light on potential patterns, anomalies, or trends related to device engagement. This information becomes instrumental in understanding the landscape of user interactions, guiding subsequent modelling decisions, and aiding in the identification of potential correlations between device types and app downloads. The presented tabular view encapsulates a comprehensive summary of device distribution, empowering stakeholders to make informed decisions for effective fraud detection and prevention strategies.

4. **Exploratory Data Analysis (EDA) Step 4: OS Counts**

   The detailed examination of the 'os' column through value counts and tabular representation is a crucial step in our Exploratory Data Analysis (EDA). By quantifying the frequency distribution of operating system IDs, we gain valuable insights into the dataset's composition. This analysis provides a clear understanding of the prevalence of different operating systems, enabling us to identify dominant OS types and potential outliers. The tabular presentation offers a concise summary, aiding in the identification of patterns or irregularities within the dataset. This EDA step serves as a foundation for informed decision-making during subsequent modelling and analysis phases, ensuring that our understanding of the data is comprehensive and nuanced.

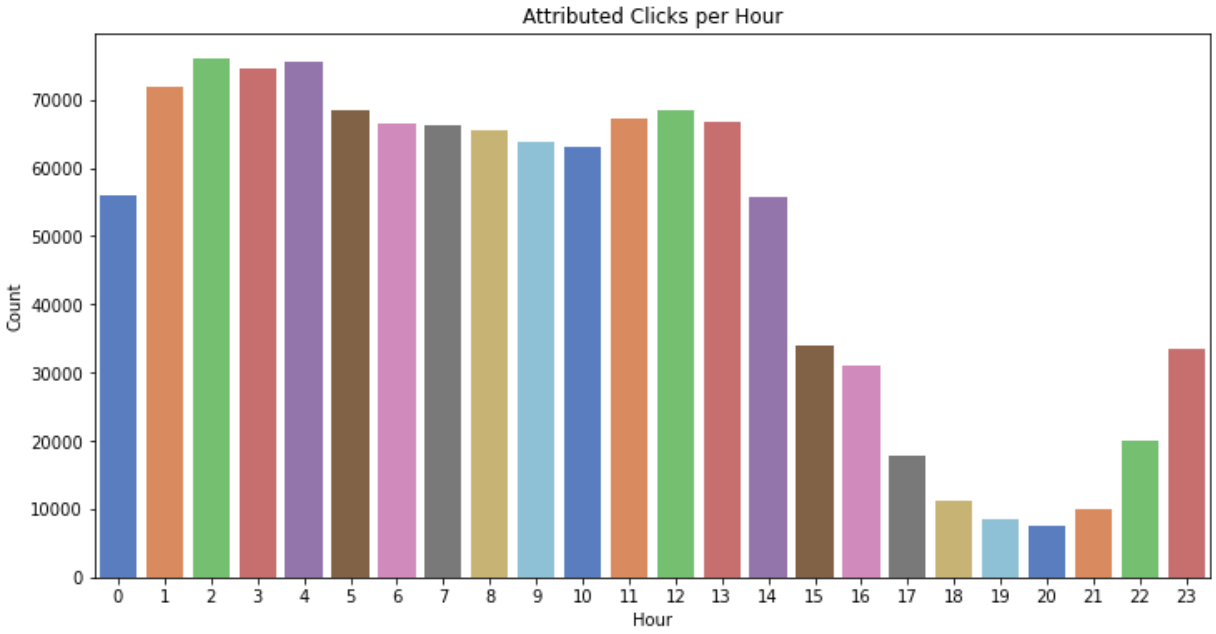
5. **Exploratory Data Analysis (EDA) Step 5: Uncovering App Distribution**

In this EDA step, we analyse the distribution of app occurrences within our balanced dataset. The bar plot vividly illustrates the top 30 most frequently encountered apps, providing insights into their prevalence. This analysis is instrumental in identifying potential patterns, trends, or anomalies related to app usage. The significance lies in guiding subsequent modelling efforts, allowing us to prioritize influential apps and comprehend their impact on user interactions. By spotlighting the prominent apps, this EDA aids in shaping a more informed approach towards feature engineering and model optimization.



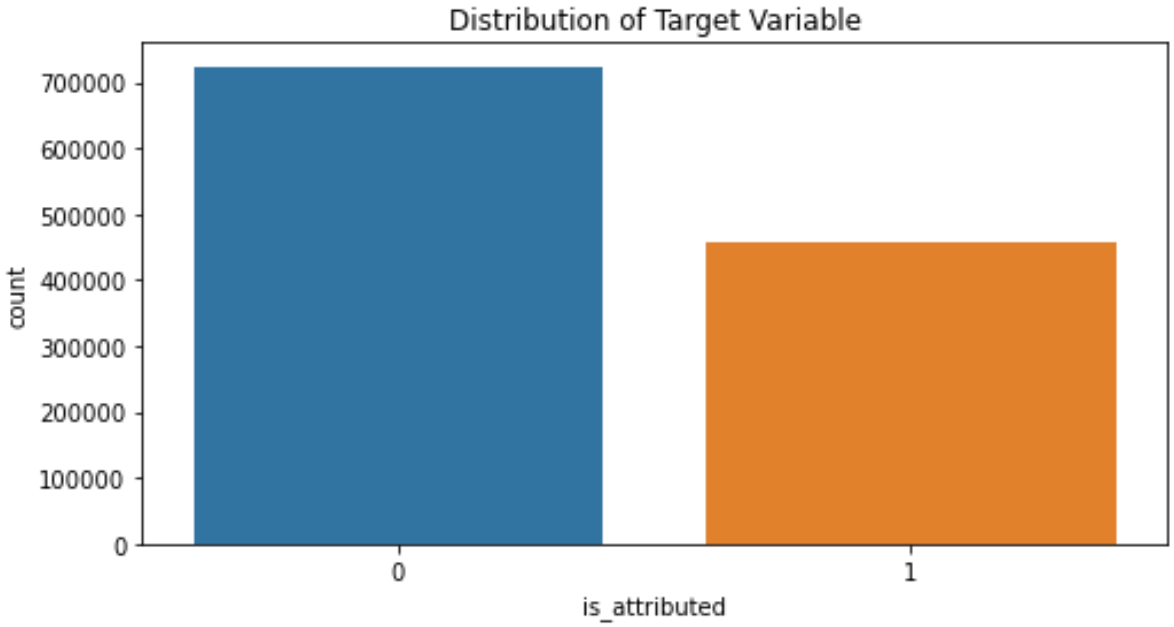
6. **Exploratory Data Analysis (EDA) Step 6: Unveiling Temporal Patterns - Clicks per Hour**

Understanding the temporal dynamics of click behavior is pivotal in unraveling patterns crucial for fraud detection and prevention. In this EDA step, we meticulously examined the distribution of attributed clicks across different hours of the day. The resulting plot reveals distinctive patterns, shedding light on peak hours of app downloads post-ad clicks. This temporal insight is invaluable for optimizing ad campaigns, enabling advertisers to strategically target audiences during periods of heightened download activity. The visual representation enhances our understanding of user engagement, informing future modelling decisions for effective fraud prediction.



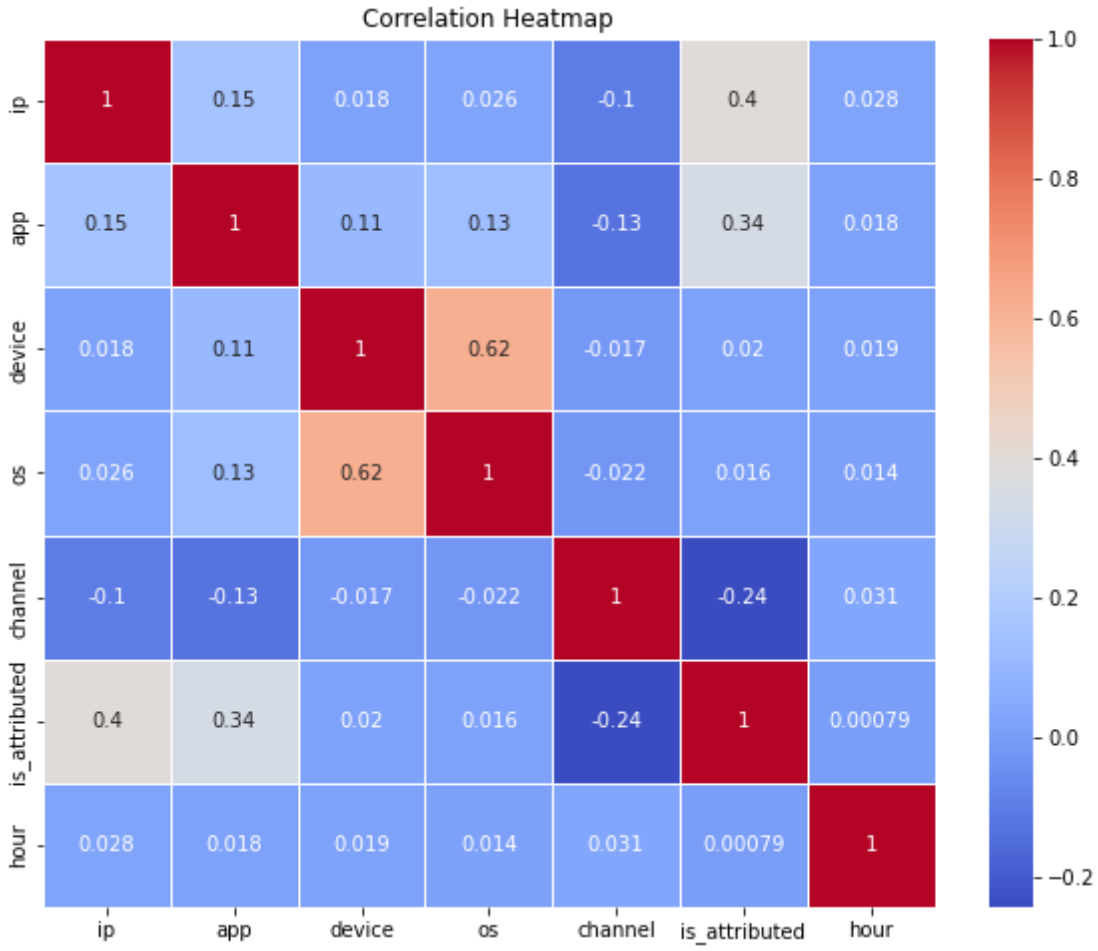
7. Exploratory Data Analysis Step 7 : Understanding Target Variable Distribution

This step is important for comprehending the imbalance within the target variable, 'is_attributed.' This visualization provides a clear depiction of the distribution of app downloads after ad clicks, offering insights into the prevalence of fraudulent and non-fraudulent activities. The plot, showcasing a binary distribution, serves as a foundational understanding for subsequent analyses. This visualization is crucial for informing the modelling strategy, as the class imbalance can significantly impact the model's ability to discern patterns associated with app downloads. Identifying and addressing this class imbalance is essential for robust and accurate predictive modelling.



8. **Exploratory Data Analysis Step 8 : Correlation Analysis - Unveiling Patterns in Balanced Data**

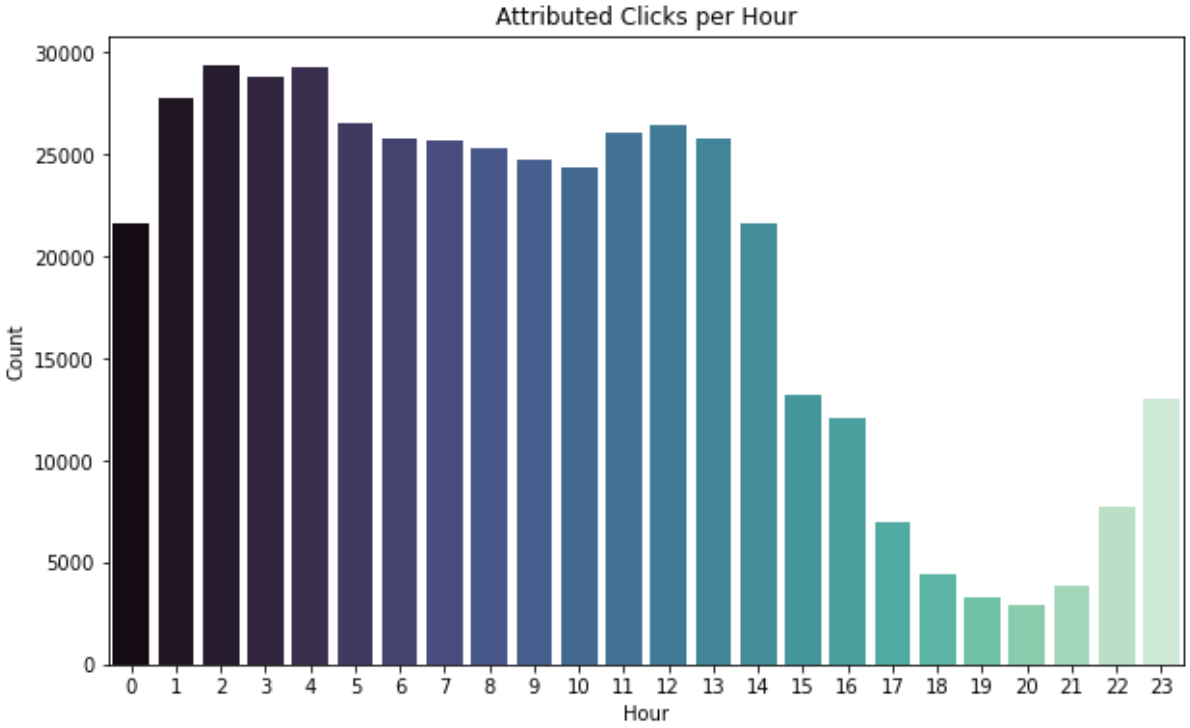
   The "Correlation Heatmap" visually depicts the pairwise correlations between different features. This heatmap serves as a critical diagnostic tool, allowing us to discern patterns and dependencies among variables. The degree and direction of correlation are color-coded, aiding in the identification of potential associations. This step is instrumental in understanding the interplay between variables, guiding feature selection, and providing insights into potential multicollinearity. The heatmap, with annotated correlation coefficients, enhances our comprehension of feature relationships, enriching subsequent modelling endeavours.



9. **Exploratory Data Analysis Step 9 : Attributed Clicks per Hour**

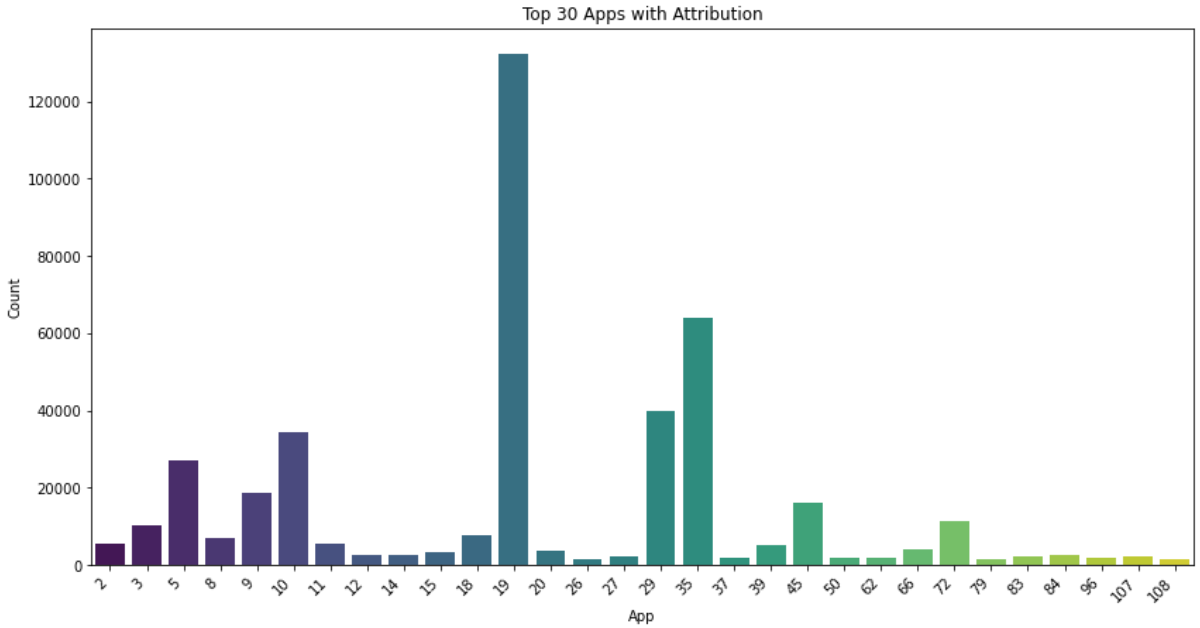
   This analysis unveils patterns in user behavior, highlighting peak hours for app downloads. Advertisers can strategically time ad placements during periods of heightened engagement, optimizing campaigns for effectiveness. Identifying peak engagement periods empowers marketers to tailor strategies, ensuring optimal resource allocation. The analysis serves as a foundation for more effective decision-making in ad placement, contributing to the overarching goal of predicting app

downloads. This insight guides advertisers in making informed decisions, ultimately enhancing the success of ad campaigns. By optimizing ad timings, advertisers can maximize visibility, engagement, and, consequently, the impact of their advertising efforts.



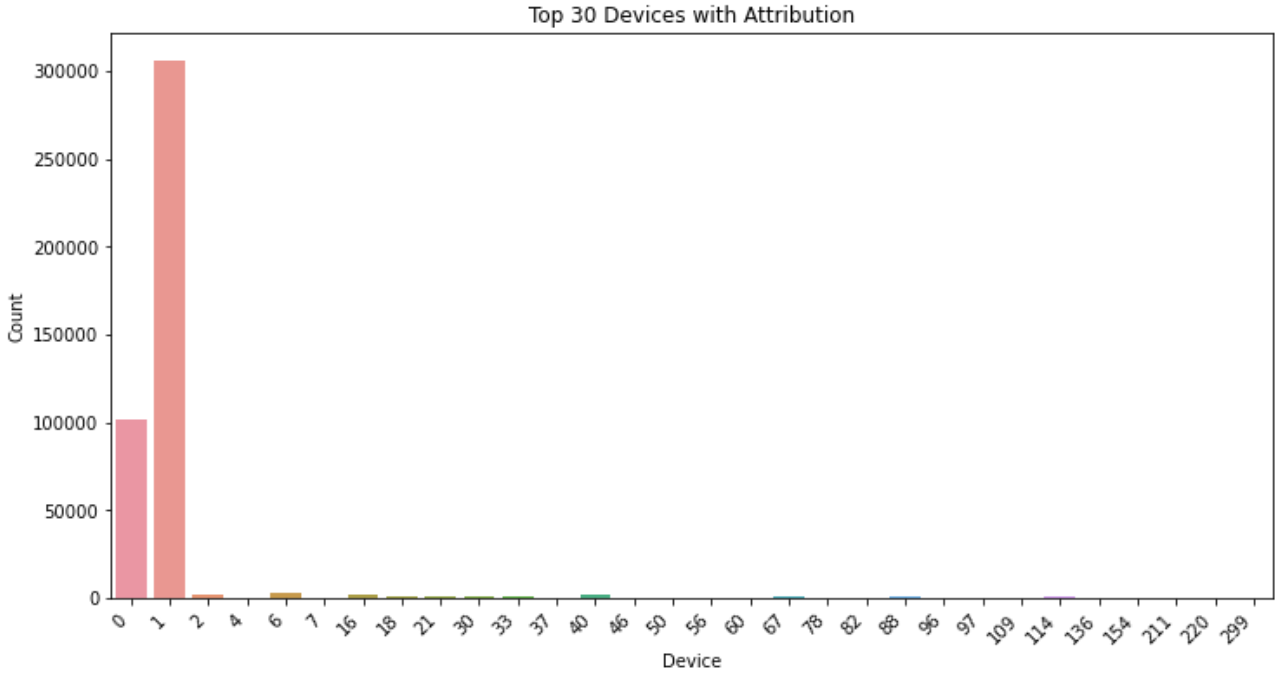
10. Exploratory Data Analysis Step 10 : Identifying Top Apps with Attribution

This EDA step is highly significant as it uncovers specific apps that significantly contribute to app downloads following ad clicks. By identifying the top apps with attribution, we gain valuable insights into the apps that excel in driving app downloads. This information is instrumental for app marketing strategies, as it directs focus on optimizing campaigns and resources for the most successful apps.



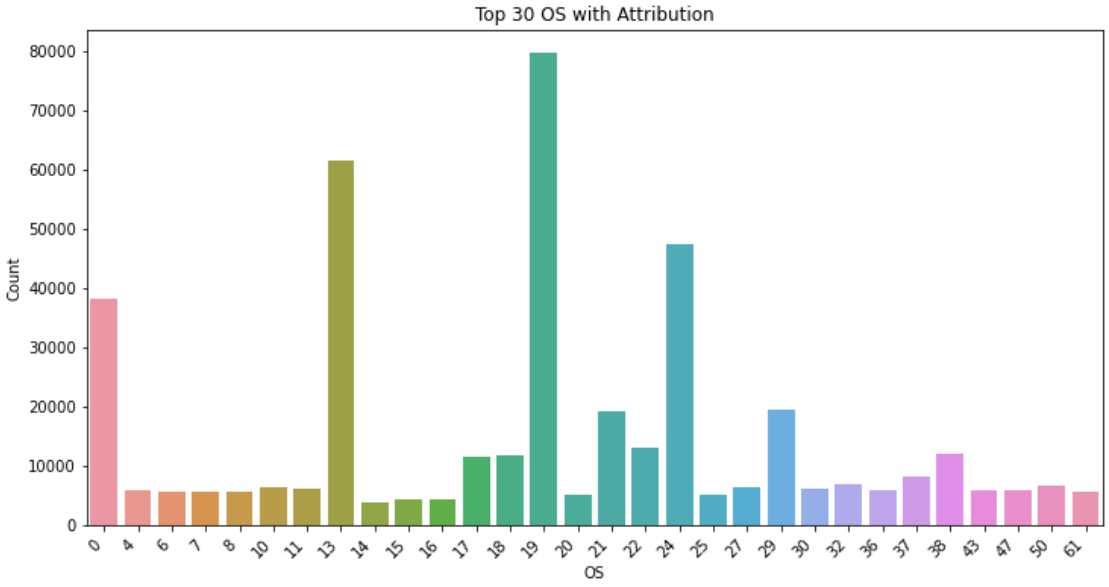
11. Exploratory Data Analysis Step 11 : Identifying top devices with attribution

This EDA step holds significance as it uncovers specific device types that significantly contribute to app downloads following ad clicks. By identifying the top devices and OS versions with attribution, we gain a deeper understanding of the technology platforms that excel in driving app downloads. This information is invaluable for optimizing app marketing strategies, ensuring compatibility with popular devices, and guiding decisions related to app development and marketing campaigns.



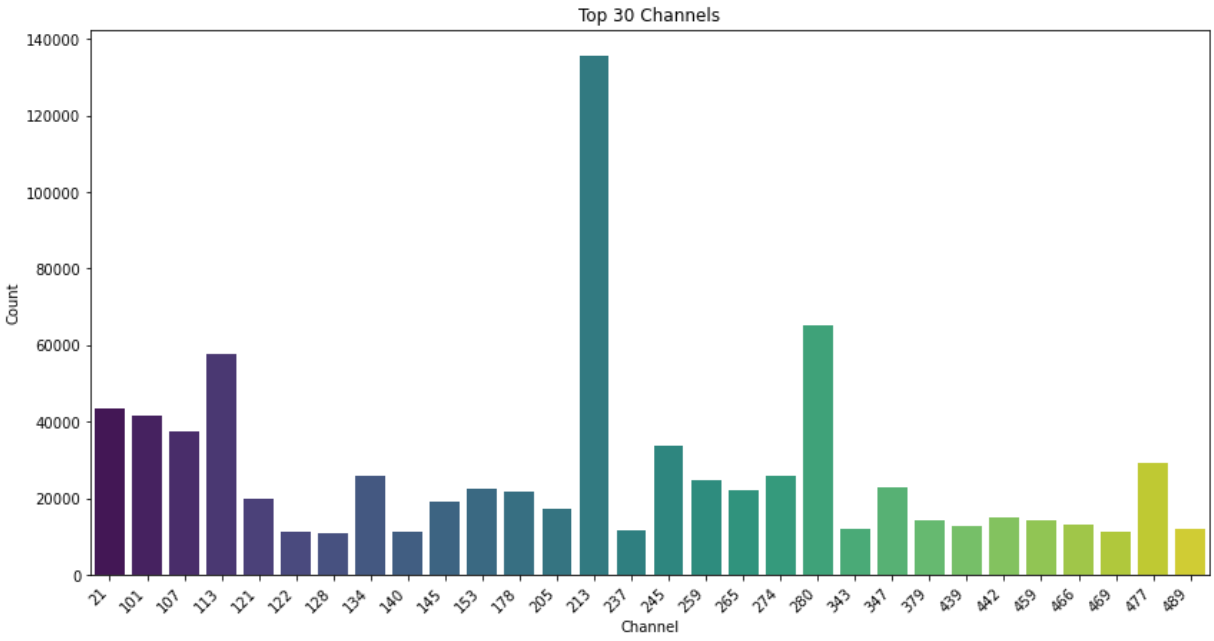
12. Exploratory Data Analysis Step 12 : Top 30 OS with Attribution EDA

By visualizing the click-to-download ratio for each OS, we gain insights into the OS preferences of users who are more likely to download apps after clicking on mobile ads. This information is invaluable for advertisers and app developers in tailoring their strategies to target specific OS environments effectively. Identifying the dominant OS platforms associated with app downloads enables precise targeting, optimizing marketing efforts, and potentially reducing click fraud risks on non-contributing OS versions. Additionally, this EDA step aids in feature selection and guides the development of predictive models, enhancing the overall effectiveness of fraud detection and app download prediction strategies.



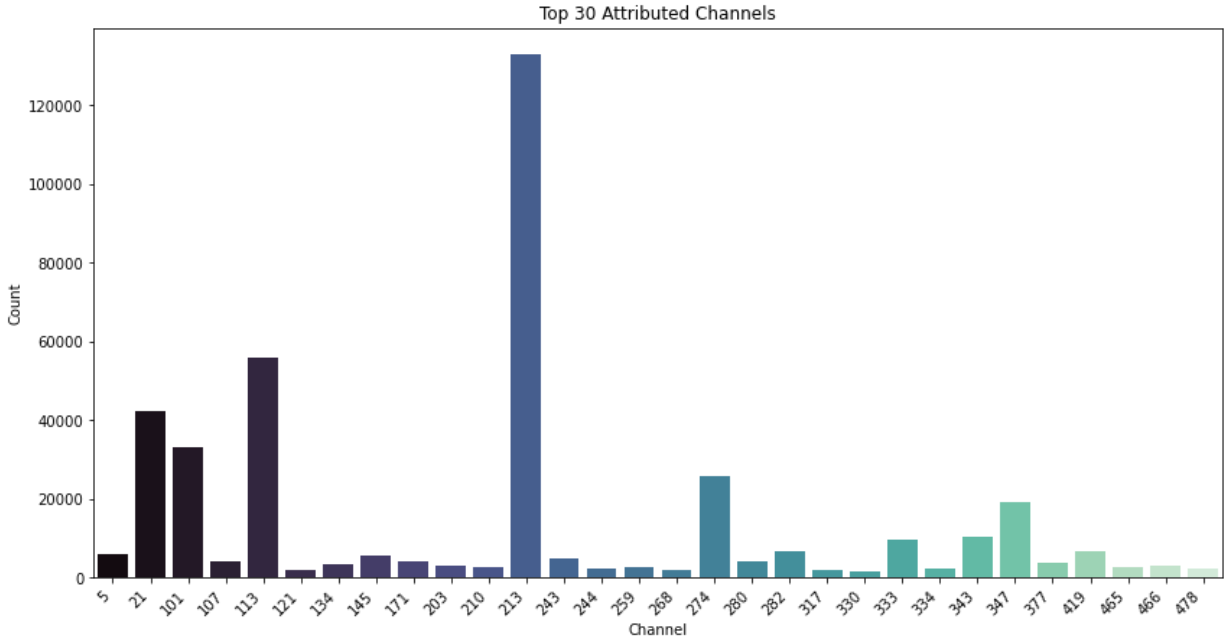
13. Exploratory Data Analysis Step 13 : Identifying Top Channels

This EDA step holds significance as it unveils advertising channels that play a crucial role in generating ad clicks. By identifying the top channels, we gain insights into sources or platforms that attract a significant amount of user attention and engagement. Understanding the characteristics and behavior associated with these channels can inform targeted marketing strategies, budget allocation, and resource utilization. It aids in optimizing ad campaigns by focusing on channels that exhibit the most substantial interaction with the ads.



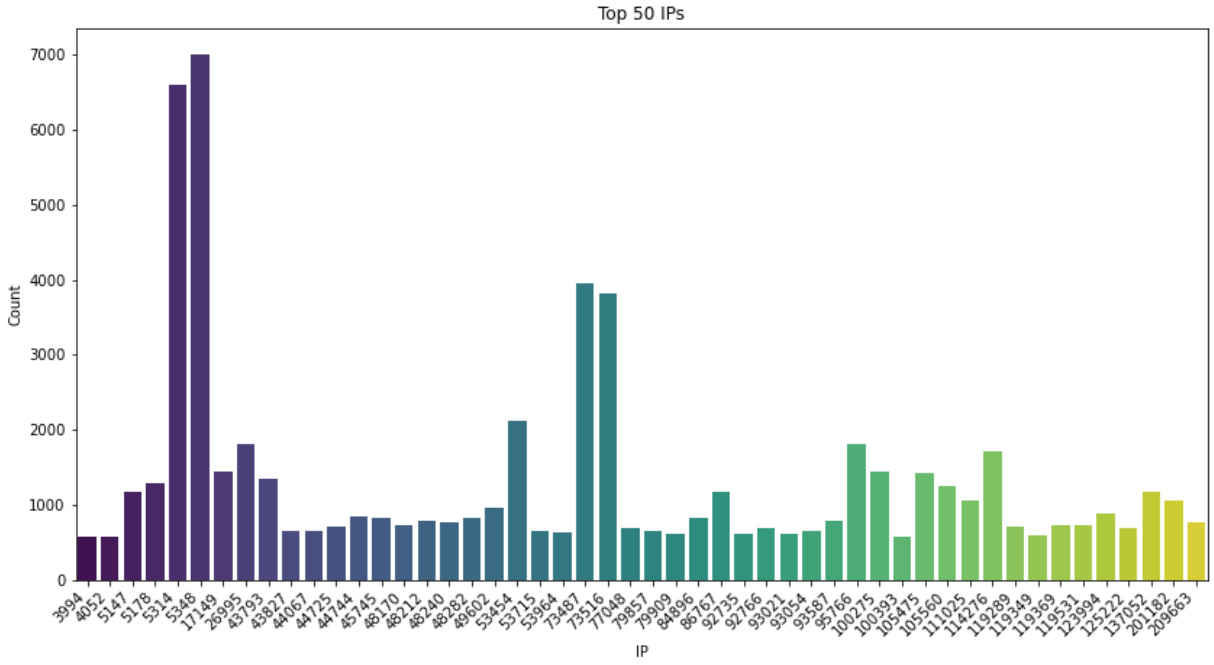
14. Exploratory Data Analysis Step 14 : Identifying Top attributed channels

The exploration of attributed clicks across channels is pivotal for identifying fraud hotspots and optimizing resources. Pinpointing the top 30 attributed channels aids in strategic budget allocation, refining fraud prevention strategies, and enhancing decision-making for advertisers. This analysis ensures efficient resource utilization and promotes continuous improvement in the dynamic landscape of mobile advertising.



15. Exploratory Data Analysis Step 15 : Identifying Top Clicked Ip's

The exploration of the top 50 IPs with the highest click counts in the dataset holds significant importance in click fraud detection. This analysis reveals potential click hotspots and helps identify IPs exhibiting unusually high click activity, which may signify fraudulent behaviour. The insights gained contribute to the refinement of IP blacklisting strategies, offering a proactive defence against fraudulent activity. Additionally, advertisers can optimize resource allocation based on IP click patterns, ensuring efficient ad spend and targeted engagement. EDA on top IPs also provides valuable insights into user behaviour, aiding in strategic decision-making for refining ad targeting strategies and enhancing the overall reliability of click data.



16. Exploratory Data Analysis Step 16 : Identifying Top Attributed IP's

This EDA step is important as it helps us recognize IP addresses that play a substantial role in driving app downloads through ad clicks. By identifying the top attributed IPs, we gain valuable insights into potential hotspots for user engagement. Further analysis and modeling can now be directed towards understanding the underlying factors that contribute to the success of these specific IPs in driving app downloads. This information is invaluable for optimizing ad campaigns and resource allocation in mobile app marketing strategies.