A red and white logo

Description automatically generated

ALY6140 Python and Analytics Systems Technology

Marketing Campaign Data Analysis

Final Project Report   
By  
Sahil Hemanth Deshmukh

**Introduction**

In the modern advertising landscape, optimizing marketing campaigns across various platforms is critical to achieving high conversion rates and maximizing returns on investment (ROI). Companies leverage multiple digital advertising platforms like Google Ads, Facebook Ads, and DV360 to promote their products, but understanding which campaigns will yield the best outcomes remains a challenging task. The aim of this report is to analyze marketing campaign data, build predictive models, and derive insights to help businesses make data-driven decisions.

The dataset contains over **72,000 rows** and **35 columns** representing various aspects of marketing campaigns, including:

* **Campaign Details**: Information about the campaign's duration (no\_of\_days), creative dimensions, media costs, approved budgets, and platforms used (ext\_service\_name).
* **Performance Metrics**: Data related to **impressions**, **clicks**, and **reach**.

This report seeks to answer critical business questions by applying machine learning models, to predict the performance of marketing campaigns and recommend strategies for optimization.

**Dataset Overview**

The dataset includes 35 columns that capture essential details about each marketing campaign. Key fields include:

* **no\_of\_days**: The number of days the campaign has been running.
* **ext\_service\_name**: The external service or platform used to run the campaign (e.g., Google Ads, Facebook Ads).
* **creative\_width** and **creative\_height**: The dimensions of the ad creatives.
* **media\_cost\_usd**: The total media cost spent for the campaign in USD.
* **impressions** and **clicks**: The number of times the campaign was displayed and clicked on.
* **unique\_reach** and **total\_reach**: Metrics representing the reach of the campaign, with unique reach focusing on distinct viewers and total reach representing the overall reach.

These fields provide a comprehensive view of each campaign's setup and performance, serving as the foundation for predictive modeling in this report.

**Rationale for Dataset**

The Marketing Campaign dataset was selected due to its ability to provide a granular and comprehensive view of digital advertising campaigns across various channels (social, search, mobile video) and platforms. Our team's experience in marketing analytics makes this dataset particularly valuable for delving into essential key performance indicators (KPIs) such as impressions, clicks, and media costs. Additionally, the dataset's temporal data enables time-series analysis, while the inclusion of financial metrics, like campaign budgets and media costs, allows for a detailed examination of the economic aspects of advertising campaigns. Overall, the dataset provides an excellent foundation for optimizing marketing strategies and improving campaign ROI.

**Data Loading and Initial Exploration:**

After loading, the initial exploration was performed to understand the structure and content of the data, including checking the first few rows and summarizing the dataset's information. This allowed us to confirm that the data loaded correctly and assess the variables included in the dataset.

**Renaming Columns:**

The dataset contains a variety of columns related to marketing campaigns, including campaign details, creative information, financial metrics, and performance indicators such as impressions and clicks.

To enhance the readability and usability of the dataset, we renamed several columns. This step ensured that column names were more descriptive and followed a consistent naming convention, making it easier to understand and work with the data during analysis.

**Dropping Irrelevant Columns:**

Certain columns such as 'Position in Content', 'Unique Reach', and 'Total Reach' were identified as irrelevant to the analysis. These columns were dropped to reduce noise in the dataset and focus on the variables that could provide more meaningful insights.

**Rounding Numerical Values:**

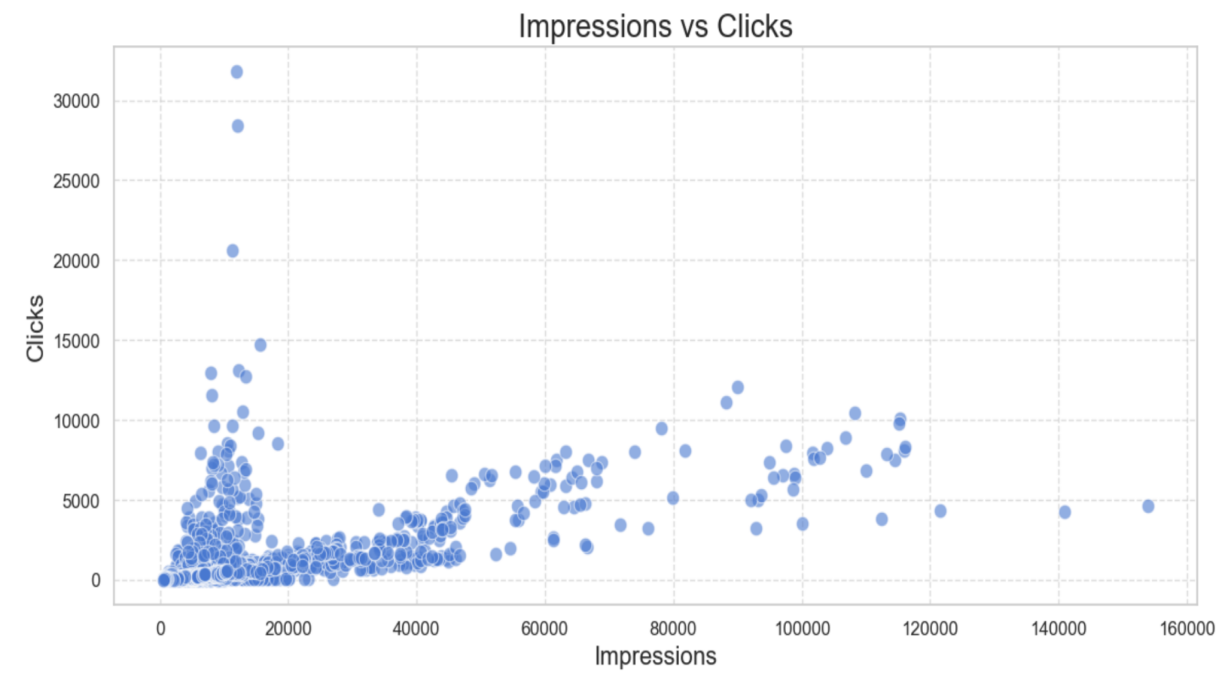
We standardized the precision of all numerical columns by rounding them to two decimal places. This step was essential to maintain consistency in the data, especially for columns that deal with financial metrics or measurements.

**Handling missing values:**

All missing values were handled appropriately, either through imputation with statistical measures or logical replacements.

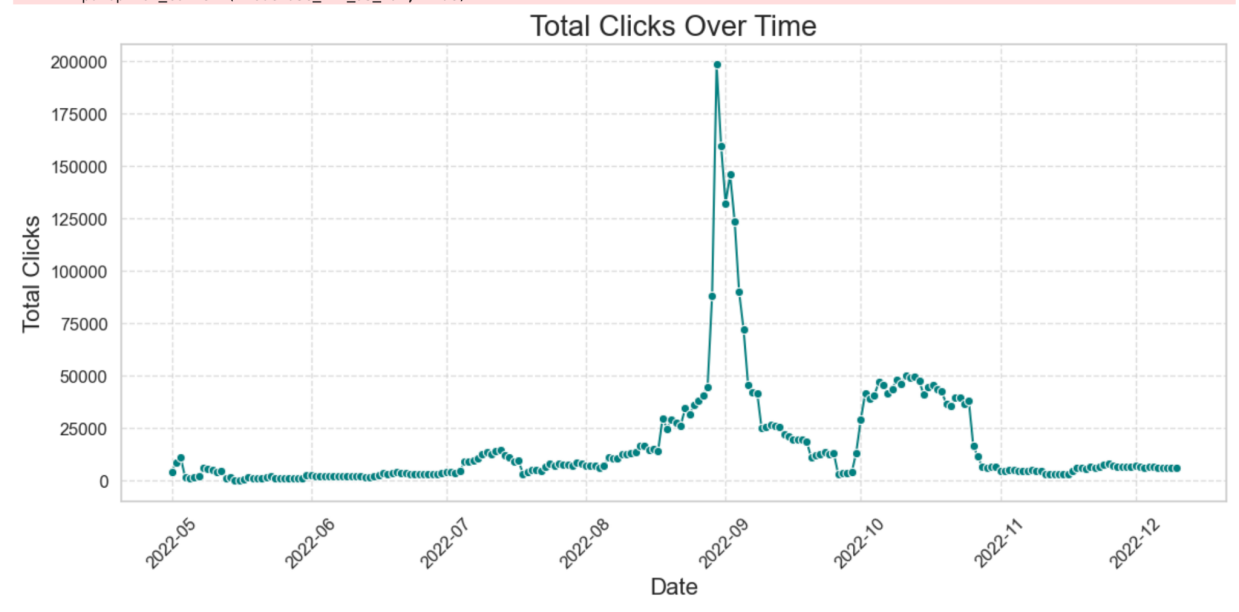
**Scatter plot of Impressions vs Clicks:**

The scatter plot illustrates the relationship between Impressions and Clicks in a marketing campaign. Each point represents a campaign, with Impressions on the x-axis and Clicks on the y-axis. The plot shows a positive trend, indicating that campaigns with more impressions tend to generate more clicks, though with noticeable variance.

****

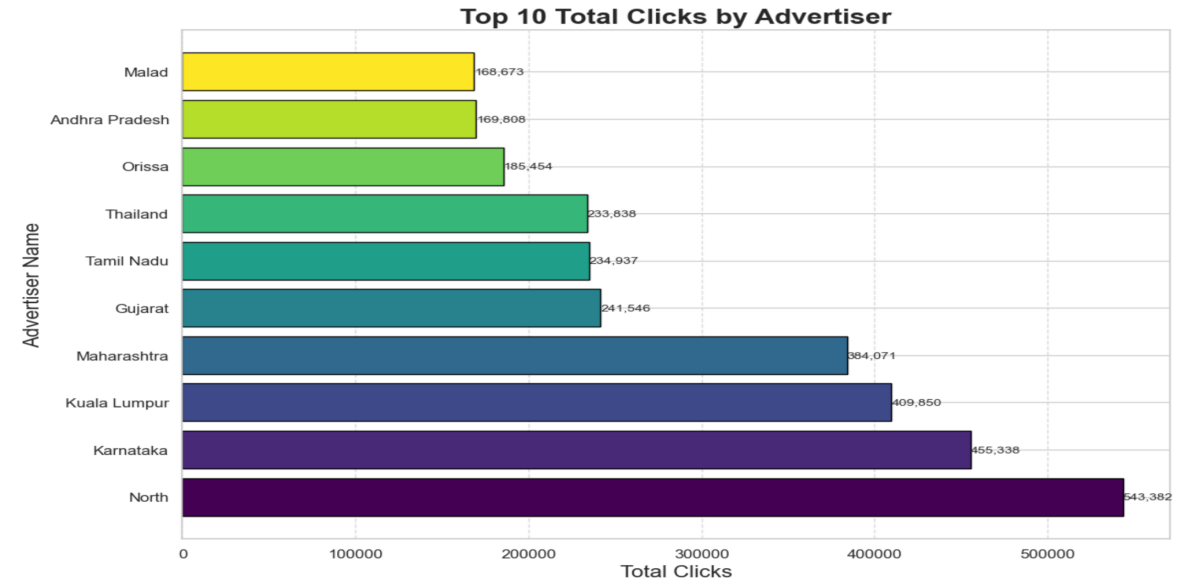
**Line plot of Clicks Over Time:**

The line plot represents the total number of clicks over time for marketing campaigns. It shows a time-based trend of campaign clicks, with a sharp peak in early September 2022, followed by a decline and subsequent smaller peaks. This suggests high campaign activity during certain periods, likely due to targeted marketing efforts or promotional events.



**Top 10 Advertisers by Total Clicks bar chart:**

The bar chart displays the top 10 advertisers based on total clicks. Each advertiser is represented by a horizontal bar, with the length of the bar indicating the total number of clicks. "North" leads with over 500,000 clicks, followed by "Karnataka" and "Kuala Lumpur." This visualization helps highlight the most effective advertisers in terms of user engagement.



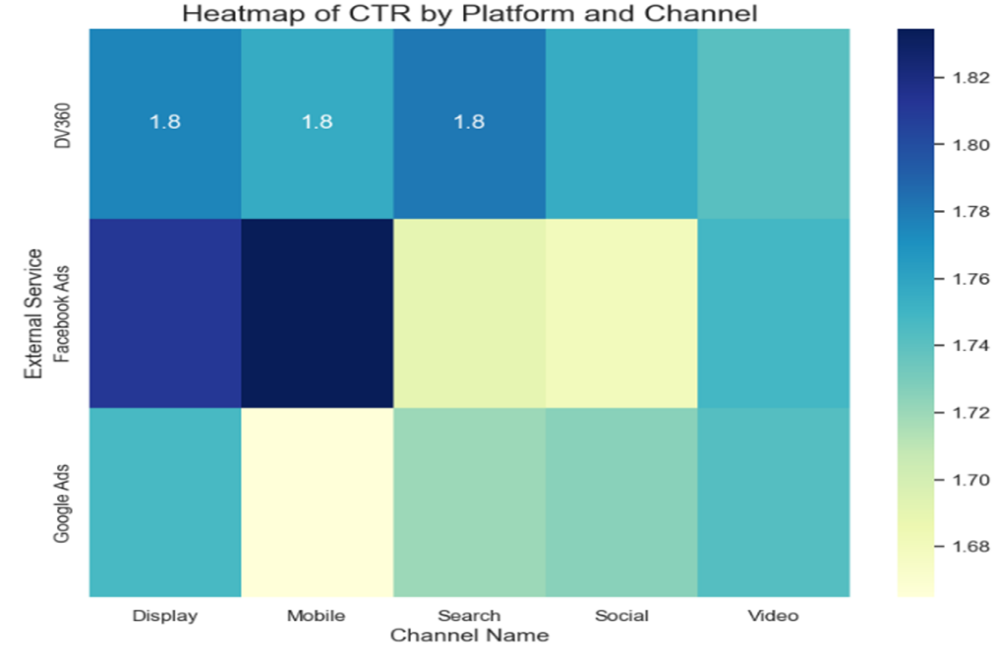
**BUSINESS PROBLEMS**

**Business Problem 1: Which platforms and channels generate the highest engagement (clicks) for the least cost? What factors influence ad clicks, and how can we optimize budget and spending to maximize click-through rates?**

**Initial Analysis**

Evaluating channel and platform performance by click-through rates (CTR)

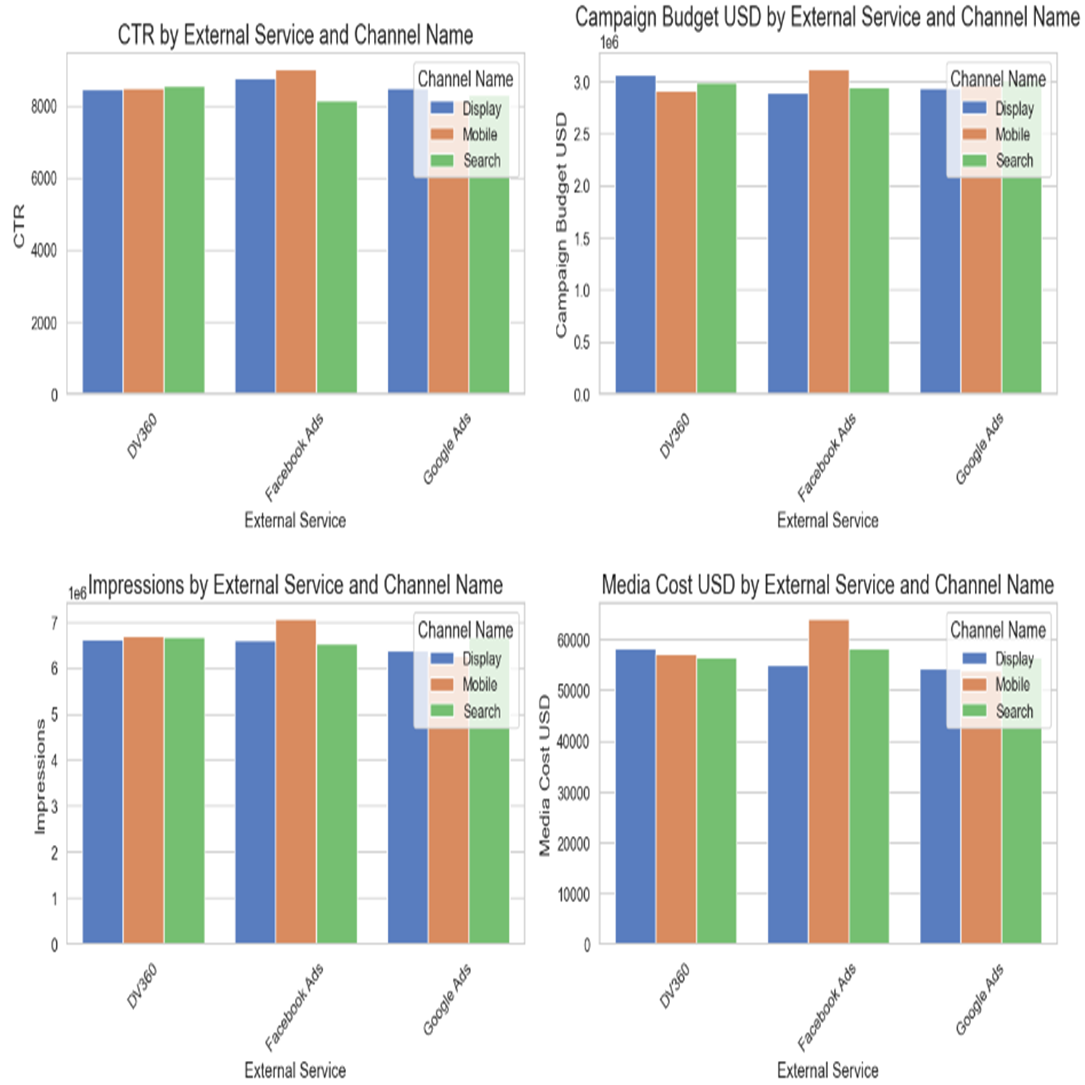
A heatmap was generated to visualize the performance of various channels (Display, Search, Social, etc.) and platforms (Facebook, Google Ads, etc.) in terms of their CTR (Click-Through Rate). Heatmaps provide a quick and intuitive way to understand how certain combinations of platforms and channels lead to higher engagement.



The analysis reveals that DV360 consistently achieves the highest CTR across all channels, making it ideal for multi-channel campaigns. Google Ads excels in Display and Video, while External Service shines in Mobile and Search. Facebook Ads shows competitive performance in Display, Mobile, and Social. To maximize effectiveness, prioritize DV360 for broad campaigns, utilize External Service for Mobile and Search, leverage Google Ads for Display and Video, and employ Facebook Ads for Social media initiatives. Continuous monitoring will help optimize performance and ROI.

Analyzing Performance Metrics Across Advertising Channels

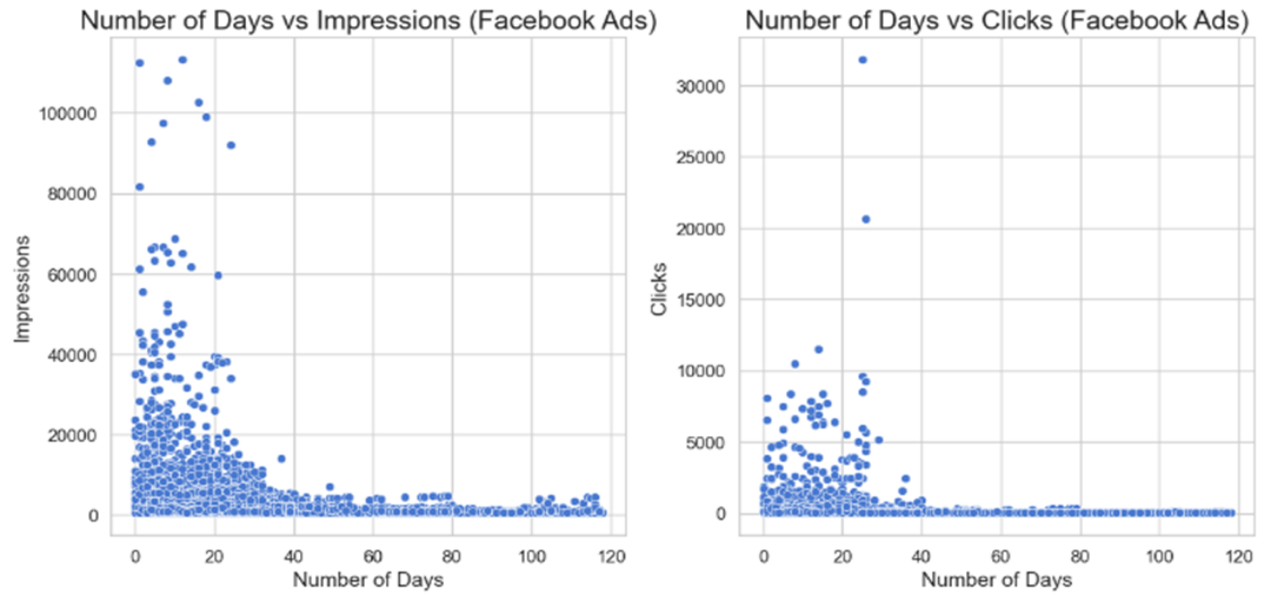
The grouped bar charts visualize the performance metrics (CTR, Campaign Budget USD, Impressions, and Media Cost USD) for three external services (DV360, Facebook Ads, Google Ads) across different channels (Display, Mobile, Search).

****

Key findings suggest that DV360 outperforms both Facebook Ads and Google Ads in click-through rate (CTR), impressions, and budget allocation, making it the most effective platform overall. Facebook Ads excels in the Mobile channel, while Google Ads demonstrates balanced performance across channels. Recommendations include prioritizing DV360 for multi-channel campaigns, using Facebook Ads for mobile-focused efforts, and leveraging Google Ads for a comprehensive approach, all while continuously monitoring performance for optimization.

Impact of Campaign Duration on Ad Performance

The scatter plots visualize the relationship between the number of days a Facebook ad campaign has been running and its impressions and clicks.



The Analysis indicate a trend of decreasing impressions as campaign duration increases, suggesting potential ad fatigue. While clicks show a less clear relationship with time, some campaigns maintain consistent performance. To address these issues, consider refreshing ad creatives, experimenting with campaign durations, and analyzing high-performing outliers. Monitoring impressions and utilizing A/B testing can help optimize your Facebook ad strategies for better effectiveness.

Top Keywords in Facebook Ads by Click-Through Rate

The analysis reveals that "Fall jewelry" has the highest click-through rate (CTR), making it the most effective keyword, followed closely by "Body jewelry" and "Pearl jewelry." Using specific keywords, especially related to particular jewelry types, tends to yield better CTRs than generic terms, while "Tassel earrings" has the lowest performance. It's recommended to focus on targeted keywords, consider seasonal trends, test variations, and regularly monitor keyword effectiveness to enhance Facebook ad campaigns and boost overall performance.

Top Campaign Templates by Click-Through Rate

Template 1.0 has the highest CTR, indicating it's the most effective for driving clicks, while Templates 23.0 and 90.0 perform moderately. Templates 92.0 and 93.0 have lower CTRs and need optimization. To enhance campaign performance, prioritize high-performing templates, optimize low-performing ones, conduct A/B tests to identify effective designs, and regularly monitor performance using predictive models that visualize actual versus predicted clicks.

**Using ML Models for Analysis**

##### Linear Regression

Linear regression was applied as a baseline model to predict clicks using features such as Campaign Budget, Impressions, and Media Cost.

Model Performance: MSE: 58,787.90, R-squared: 0.518 (explains 51.8% of the variance in clicks), RMSE: 242.46

Interpretation: Linear regression captured the overall trends but missed key interactions, indicating that a more complex model might be required to fully understand how features interact to influence clicks.

Ridge Regression

Ridge regression introduced regularization to penalize large coefficients and reduce overfitting. This improved the model's performance slightly.

Model Performance: MSE: 58,253.91, R-squared: 0.522, RMSE: 241.36

Interpretation: Regularization helped reduce overfitting but still did not fully capture the complex relationships between features and clicks.

##### Decision Tree

A decision tree was used to model non-linear relationships between features, leading to improved predictions but with some overfitting.

Model Performance: MSE: 69,426.04, R-squared: 0.436, RMSE: 263.49

Interpretation: The decision tree captured non-linear relationships but tended to overfit the data, resulting in a higher MSE and RMSE compared to other models.

##### Random Forest

Random forest, an ensemble method, significantly improved performance by averaging multiple decision trees, reducing overfitting, and providing more accurate predictions.

Model Performance: MSE: 53,256.12, R-squared: 0.573, RMSE: 230.75

Interpretation: Random forest outperformed the decision tree model by reducing variance and capturing complex relationships in the data, resulting in better overall performance.

##### Gradient Boosting

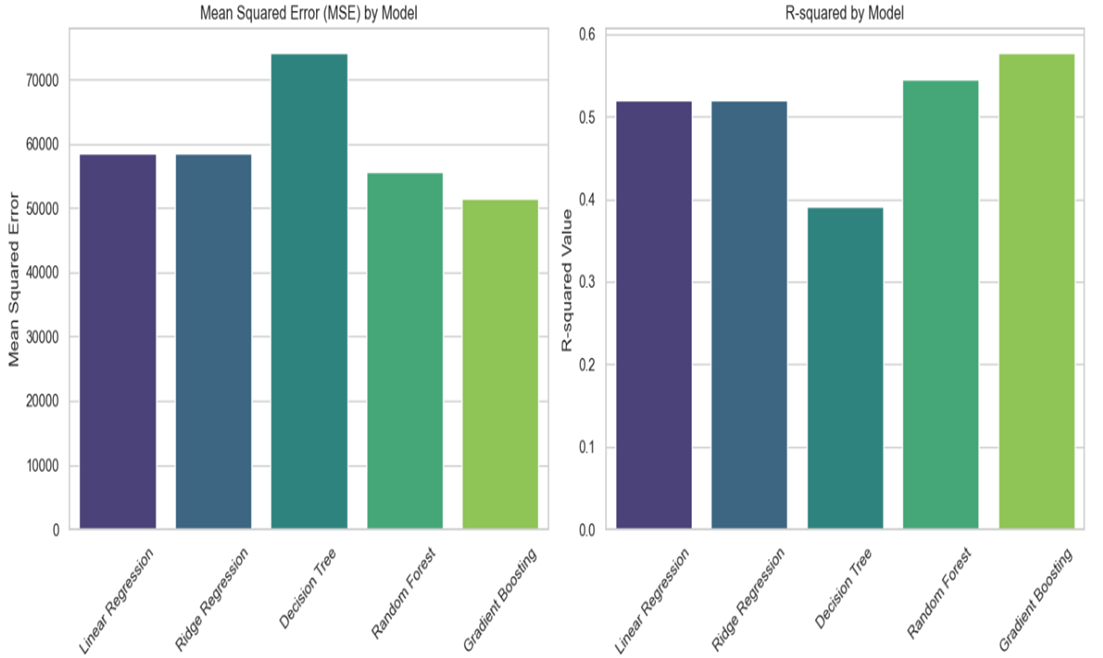
Gradient boosting further refined predictions by iteratively improving the model's performance, leading to the best results among all the models tested.

Model Performance: MSE: 51,247.56, R-squared: 0.598, RMSE: 226.39

Interpretation: Gradient boosting provided the most accurate predictions, effectively modeling non-linear interactions between features and consistently reducing the error rates.

## Model Performance Comparison

The charts visualize the Mean Squared Error (MSE) and R-squared values for five regression models: Linear Regression, Ridge Regression, Decision Tree, Random Forest, and Gradient Boosting.



Gradient Boosting is the most accurate model, with the lowest Mean Squared Error (MSE) and highest R-squared, while Decision Tree has the highest MSE and lowest R-squared, indicating it may underfit the data. Linear and Ridge Regression show similar performance, suggesting regularization had little impact. Recommendations include focusing on Gradient Boosting, fine-tuning hyperparameters, exploring feature engineering, and evaluating the model on new data for better generalization.

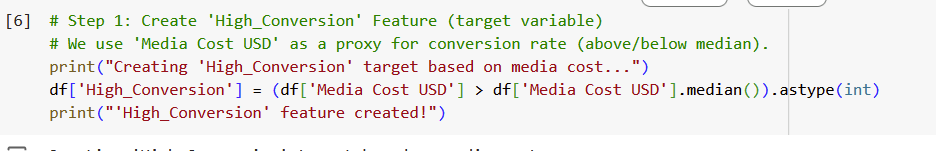
### **Business Problem 2:**

**Can we predict whether a campaign will achieve a high conversion rate based on factors like media cost, external service, and ad creative features?**

We aim to build a predictive model using **Logistic Regression** to classify campaigns into high and low conversion rate categories based on key features like **media cost**, **external service**, and **clicks**.

### **Step 1: Data Preparation**

The code begins by creating a new target variable, **High\_Conversion**, which categorizes campaigns based on whether their **Media Cost USD** is above or below the median value.



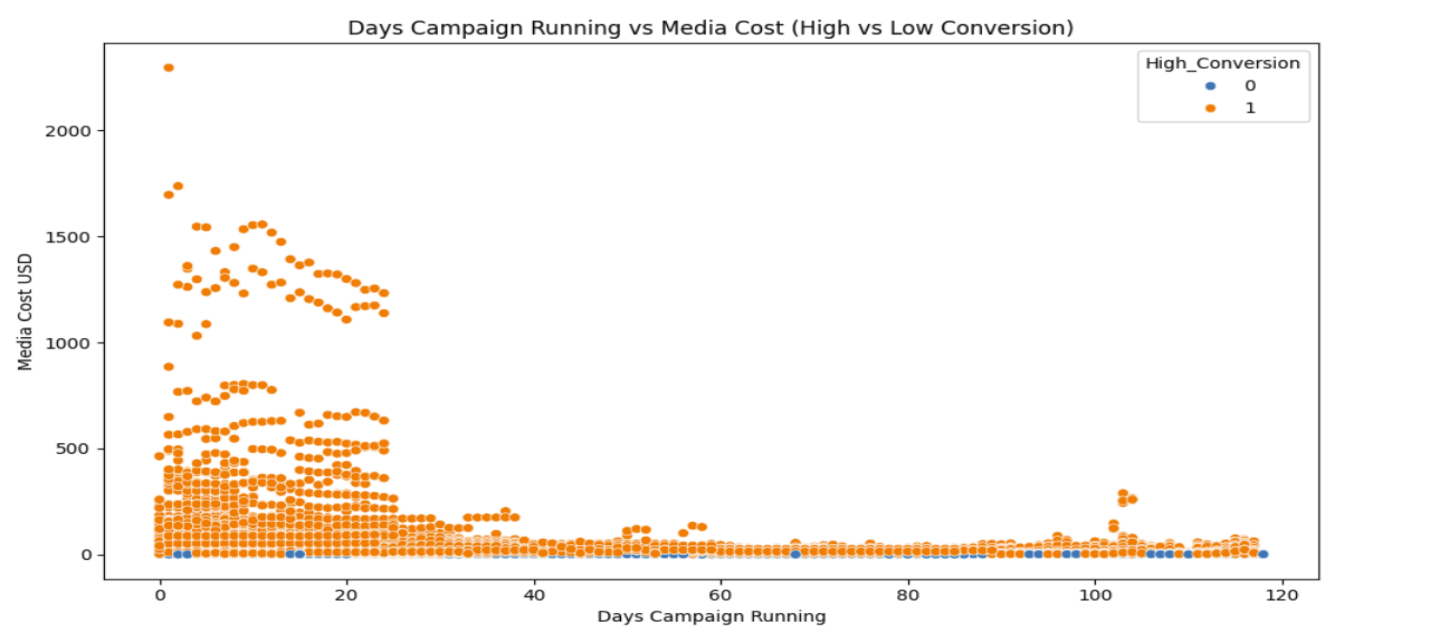
Here, we calculate the median of **Media Cost USD** and create a binary feature:

* **1**: Indicates a high-conversion campaign (media cost above the median).
* **0**: Indicates a low-conversion campaign (media cost below the median).

**Analysis & Interpretation:** By using **Media Cost USD** as a proxy for conversion rate, we are setting up the target variable for our classification model. This allows us to differentiate between high- and low-performing campaigns in terms of financial investment.

### **Visualization: Days Campaign Running vs Media Cost**

The scatterplot visualizes the relationship between **Days Campaign Running** and **Media Cost USD**, with the campaigns color-coded by their conversion category (high vs. low).

**Visualization Output:**

**Analysis & Interpretation:**

* We observe that campaigns with **higher media costs** are often associated with **high conversion rates** (orange dots).
* Campaigns that run for fewer days with higher media costs tend to have higher conversion rates, whereas campaigns with low media costs are more likely to have low conversion rates (blue dots).
* This suggests that both **media cost** and **duration** (days running) are influential in predicting the success of a campaign.

**Recommendation:** Businesses should focus on optimizing the duration and media cost for their campaigns. Spending more may lead to higher conversion rates, but the length of the campaign also plays a role. A balance between the right investment and optimal running time could lead to better conversion rates.

### **Step 2: Splitting and Scaling the Data**

The features selected are **Days Campaign Running**, **External Service ID**, and **Clicks**, The dataset is then split into training and testing sets, and the features are scaled to ensure proper model performance.

* **Data Splitting**: 70% of the data is used for training the model, and 30% is reserved for testing. This ensures that the model is evaluated on unseen data.
* **Scaling**: Standardization (scaling) is applied to the features so that all have a similar range, which is essential for Logistic Regression.

**Analysis & Interpretation:** Scaling is critical for models like **Logistic Regression** as it relies on the distances between data points. By scaling the data, we ensure that no single feature disproportionately affects the model's predictions.

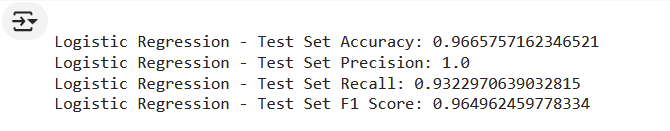
### **Step 3: Training the Logistic Regression Model**

The Logistic Regression model is trained on the scaled training data to predict whether a campaign will have a high conversion rate.

* The **Logistic Regression** model is used to classify campaigns as high or low conversion based on the selected features.
* **Training** involves finding the best-fit model for the training data, where the model learns the relationship between the features and the target.

### **Step 4: Model Evaluation**

After training, the model is evaluated on the test set. The key metrics include **accuracy**, **precision**, **recall**, and **F1 score**.



**Analysis & Interpretation:**

* **Accuracy** of 96.66% indicates that the model correctly classified campaigns most of the time.
* **Precision** of 100% means that when the model predicted a high conversion rate, it was always correct.
* **Recall** of 93.23% indicates that the model correctly identified most of the high conversion campaigns.
* The **F1 score** balances precision and recall and indicates strong overall performance.

**Recommendation:** This model can effectively predict whether a campaign will have a high conversion rate, based on external service, and clicks. Businesses can rely on this model to forecast campaign performance, helping them make data-driven decisions on where to allocate their marketing budget.

### **Business Problem 3:**

**How can advertisers optimize their campaign budget to maximize impressions, clicks, and reach while minimizing media costs?**

The goal is to optimize the campaign budget to maximize impressions, clicks, and reach while minimizing media costs. By predicting the Click-Through Rate (CTR), we can forecast which campaigns will yield the best results in terms of clicks relative to the amount spent.

CTR (Click-Through Rate): Calculated as *(Clicks/Impressions)×100*. This metric measures the effectiveness of the ad in generating clicks relative to the number of impressions. A higher CTR indicates better ad performance.

**Model 1: Ridge Regression for Predicting CTR**

The first model uses Ridge Regression to predict the Click-Through Rate (CTR) based on various factors, such as:

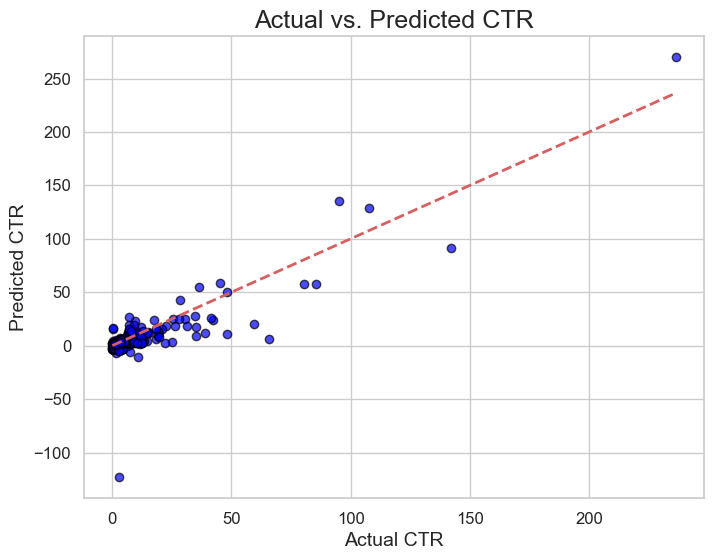
1. Media Cost USD: The total expenditure on media for the campaign. This is crucial for understanding how much budget is allocated and its relation to performance metrics.
2. Campaign Budget USD: The total budget allocated for the entire campaign. This helps in assessing the efficiency of budget usage across different campaigns.
3. Clicks\_Impressions: The product of clicks and impressions. This interaction term captures the combined effect of both metrics, providing insight into how clicks relate to impressions in driving CTR.
4. Cost\_Impressions: The product of media cost and impressions. This term highlights how much is spent per impression, which can help in understanding cost efficiency in reaching audiences.
5. Cost\_Clicks: The product of media cost and clicks. This reveals how much is spent per click, which is essential for evaluating the cost-effectiveness of campaigns.
6. spent\_budget\_per\_day: Calculated as the total campaign budget divided by the number of days the campaign is running. This indicates the daily spending rate and helps in understanding budget pacing throughout the campaign duration.
7. CPC (Cost Per Click): This is calculated as the media cost divided by the total clicks. It helps advertisers understand how much they are spending for each click, which is critical for budget optimization.
8. CPI (Cost Per Impression): This is calculated as the media cost divided by the total impressions. This metric indicates the cost incurred for each impression served, assisting in evaluating the cost efficiency of reaching potential customers.

The model uses polynomial features to account for non-linear relationships between the independent variables and the CTR, capturing more complex patterns in the data.

**Evaluation Metrics:**

This model was evaluated using several metrics to assess its performance:

1. Mean Squared Error (MSE): 3.10  
   This metric gives us an idea of how far off our model’s predictions are from the actual values. In this case, an MSE of 3.10 means that on average, the model's predictions are about 3.10 units away from the true CTR values. Lower MSE values are better, as they indicate fewer prediction errors.
2. Mean Absolute Error (MAE): 0.50  
   MAE is a simpler measure of how much the model’s predictions differ from the actual values, on average. With a value of 0.50, it tells us that, on average, the model is off by 0.50 percentage points when predicting the CTR.
3. Root Mean Squared Error (RMSE): 1.76  
   RMSE is the square root of MSE, which gives us an error measure that’s in the same units as the CTR percentages. With an RMSE of 1.76, it means that the typical error in predicting CTR is about 1.76 percentage points.
4. R-squared (R²): 72.59%  
   R-squared tells us how well the model explains the variance in the CTR data. In this case, an R-squared of 72.59% means that our model explains about 72.6% of the variation in the CTR, which is a good sign that it’s capturing some important patterns.
5. Mean Absolute Percentage Error (MAPE): 42.37%  
   MAPE shows the average percentage error in the model’s predictions compared to the actual values. A MAPE of 42.37% indicates that, on average, the model’s predictions are off by around 42.4% of the true CTR values. While this is a bit higher than ideal, it still offers valuable insights into how well the model performs, particularly for campaigns where smaller errors can lead to big differences in performance.



These evaluation metrics give a comprehensive view of how well the model is performing. While the model has some room for improvement, especially with the MAPE, the relatively strong R-squared value indicates that it’s capturing a significant portion of the trends in the data. The MAE and RMSE provide additional insights into the magnitude of errors, which are helpful for understanding where the model’s predictions can be more accurate.

**Model 2: Multi-Armed Bandit (Thompson Sampling) for Campaign Optimization**

Thompson Sampling is another approach to optimize the allocation of campaign budgets by identifying which campaign has the highest estimated CTR. This model allows for dynamic decision-making, adjusting the budget allocation based on real-time performance feedback (successes and failures).

Incorporating Thompson Sampling adds significant value to the solution by making campaign optimization more dynamic, adaptable, and efficient. While Model 1 gives a solid foundation for predicting CTR, Thompson Sampling enhances it by enabling real-time, data-driven decision-making on how to best allocate the advertising budget across campaigns. This helps advertisers not only maximize impressions, clicks, and reach but also minimize media costs through strategic, informed adjustments to budget allocation.

1. Initial Setup:
   1. Each campaign starts with some initial belief about its success rate. This is represented as a Beta distribution with two parameters: α (alpha) and β (beta), where α represents the number of successes (clicks) and β represents the number of failures (no clicks).
   2. For simplicity, we start with uniform distributions (i.e., all campaigns are assumed to have an equal chance of success at the beginning).
2. Thompson Sampling Process:
   1. Sampling the Success Rate: In each trial (iteration), Thompson Sampling samples a value from each campaign's Beta distribution. This sampled value represents the estimated success rate (Click-Through Rate or CTR) for each campaign.
   2. Selecting the Best Campaign: The campaign with the highest sampled CTR is chosen for the next budget allocation.
   3. Recording Success or Failure: After allocating the budget and running the campaign, we record whether the campaign was a success (a click occurred) or failure (no click occurred).
   4. Updating the Distribution: Based on the result (success or failure), the parameters of the Beta distribution for that campaign are updated:
      1. If it was a success, the α parameter is incremented by 1.
      2. If it was a failure, the β parameter is incremented by 1.
   5. This process is repeated for a predefined number of trials (n\_trials), continuously refining the estimated CTR for each campaign.
3. End Result: After running all the trials, Thompson Sampling will have a refined estimate of the CTR for each campaign. The campaigns that have shown consistently higher CTRs will receive more budget in future iterations, maximizing overall performance.

After running the Thompson Sampling algorithm for 1,000 trials (simulating the dynamic budget allocation process over time), the estimated CTRs for each campaign were computed:

After running the Thompson Sampling algorithm, the estimated CTRs for each campaign were:

1. Campaign 1: 3.23%
2. Campaign 2: 6.24%
3. Campaign 3: 9.14%

#### Interpretation of Results:

1. Campaign 3 is estimated to have the highest CTR (9.14%), followed by Campaign 2 (6.24%), and Campaign 1 (3.23%). This suggests that Campaign 3 should receive the majority of the budget allocation.
2. The estimated CTRs provide a dynamic way to allocate resources based on actual campaign performance over time, making it easier to focus spending on high-performing campaigns.

**Conclusion**

The analysis of the three business problems yields actionable insights for marketers. DV360 consistently delivers the highest click-through rates (CTR) across Display, Search, and Video channels, making it ideal for multi-channel campaigns. Facebook Ads perform exceptionally well in mobile formats, while Google Ads provide balanced performance across various channels. Marketers should prioritize DV360 for multi-channel efforts, focus Facebook Ads on mobile campaigns, and utilize Google Ads for a versatile approach.

In predicting high conversion campaigns, the Logistic Regression model demonstrated 96.66% accuracy, effectively distinguishing between high and low conversion rates. This model serves as a valuable tool for marketers to make informed decisions on campaign spending and targeting. Additionally, the use of Ridge Regression and Thompson Sampling for budget optimization highlighted key factors such as cost per click (CPC) and daily budget pacing. By implementing a dynamic budget allocation strategy, marketers can continually optimize spending, focusing resources on high-performing campaigns while making real-time adjustments based on performance feedback**.**

**References**

Chavan, R. (2023, June 28). *Marketing campaign dataset*. Kaggle. <https://www.kaggle.com/datasets/rahulchavan99/marketing-campaign-dataset?resource=download>