**A group of people sitting around a table with icons

AI-generated content may be incorrect.**

2025

**Index**

**1.Introduction………………………………………………………………………….................5**

**2. Data Description……………………………………………………………………………….6**

**2.1 Dataset Source………………………………………………………………………..6**

**2.2 Brief Description of the Data………………………………………………………..6**

**3. Data Preprocessing and Engineering……………………………………………...................9**

**3.1 Data Loading and Initial Inspection………………………………………………..9**

**3.2 Data Cleaning and Standardization………………………………………...............9**

**3.3 Outlier and Missing Value Mitigation……………………………………...............10**

**3.4 Feature Engineering………………………………………………………………...11**

**3.5 Final Preparation for Modeling……………………………………………………12**

**4. Project Research Questions………………………………………………………….............13**

**4.1. Descriptive Analysis Questions……………………………………………………13**

**4.2. Predictive Modeling Questions……………………………………………………13**

**5. Exploratory Data Analysis (EDA)…………………………………………………………...14**

**5.1 Feature Distribution and Correlation Analysis……………………………………14**

**5.2 Target Audience Performance Benchmarking……………………………………18**

**5.3 Categorical Feature Performance Benchmarking………………………..............19**

**5.4 Continuous Metric Relationships and Correlation Analysis……………..............21**

**6. Methodology………………………………………………………………………….............24**

**6.1 Tools and Technologies……………………………………………………………..24**

**6.2 Data Preprocessing and Feature Engineering…………………………….............25**

**6.3 Exploratory Data Analysis (EDA) and Data Preparation………………..............25**

**6.4 Descriptive Analysis (Q1 and Q3)………………………………………….............26**

**6.4.1 Clustering Parameters……………………………………………………26**

**6.4.2 Enhanced Feature Space…………………………………………………27**

**6.4.3 Elite Cluster Identification (Composite Score)…………………………27**

**6.4.4 Advanced Blueprint Characterization (Detailed)……………................27**

**6.5 Predictive and Optimization Modeling……………………………………............28**

**7. Results & Interpretation…………………………………………………………………......29**

**7.1 Q1: Which Channel and Strategy Yield the Highest ROI?....................................29**

**7.2 Q3: How Does Audience Conversion Rate Vary by Budget Tier on Twitter?......30**

**7.3 Q2 Results: Elite Campaign Clustering Approaches……………………………..31**

**7.3.1 1️⃣ First Approach: K-Means with Mode-Based Elite Blueprint………31**

**7.3.2 2️⃣Second Approach: K-Means with Quantile-Based Elite Identification…………………………………………………………………....32**

**7.3.3 3️⃣ Third Approach: K-Means with Composite Score & Mode……….35**

**7.3.4 Result Analysis…………………………………………………………....38**

**7.4 Q4. Predictive Modeling Approaches and Results………………………………..38**

**7.4.1 Results for All Approaches……………………………………………….41**

**7.5 Q5. Acqusition Cost Prediction……………………………………………………43**

**7.5.1 First Approach: Baseline Campaign Features………………………….43**

**7.5.2 Second Approach: Enhanced Feature Set………………………………45**

**7.6 Bonus Finding: RFECV Feature Selection and Model Performance Results….46**

**8. Conclusion……………………………………………………………………………………47**

**9. Recommendations……………………………………………………………………………48**

**10. Acknowledgements And References………………………………………………………49**

**Executive Summary of Project Updates (Post-Feedback)**

|  |  |  |
| --- | --- | --- |
| Section | Revision Description | Key Changes |
| Section 4.1 | **Revised Descriptive Analysis** | Shifted the focus of the first **descriptive analysis question** from **Conversion Rate** **to Return on Investment** (ROI), now identifying the highest ROI strategy as recommended. |
| Section 5.1 | **Expanded Exploratory Data Analysis (EDA)** | Added a comprehensive **distribution analysis** for key continuous features, including **Clicks**, **Impressions**, and **Engagement Score**. |
| Section 5.1 | **Correlation & Modeling Justification** | Included **Pearson correlation analysis** to explicitly clarify relationships between key metrics and **ROI**. Detailed interpretations were added, emphasizing **cost efficiency** and engagement metrics for ROI, which in turn justify the use of **clustering** and **tree-based models**. |
| Sections 5.2 & 5.3 | **Target Audience & Categorical Feature Benchmarking** | Benchmarked campaign performance by **Target Audience**, **Channel**, and other categorical features. New tables and visualizations were included to present performance metrics like **ROI** and **Conversion Rate** across these categories. |
| Section 5.4 | **Continuous Metric Relationships** | Added visual aids, specifically **correlation heatmaps** and **scatter plots**, to better visualize and understand the relationships between **ROI**, cost, and other continuous metrics. |
| Section 7.3 | **Refined Clustering for Elite Campaigns** | Multiple **K-Means clustering strategies** were implemented, including quantile-based and composite scoring. Advanced statistical validation using **Chi-square** and **Cramér’s V** was added. **Most importantly, the methodology was refined so that elite campaign characteristics are now identified directly from the original dataset points, rather than relying solely on cluster-level summaries.** This ensures findings are more actionable and accurate. |
| Section 7.4 | **Predictive Modeling Enhancements**  **(Q4, Q5)** | Expanded the **feature sets** for predictive modeling approaches to include additional **business** and **engagement metrics** and tested several model types and feature combinations to assess if predictive performance could be improved. Despite these enhancements, the results across all approaches and models were nearly **identical**, with **no meaningful improvement** in predictive accuracy or ability to distinguish successful campaigns. |

**1. Introduction**

Social media advertising is now a cornerstone of contemporary marketing, with platforms like Facebook, Instagram, and Twitter accounting for a substantial share of marketing expenditures. As organizations increasingly depend on these channels to connect with and influence their audiences, relying solely on basic Key Performance Indicators (KPIs) is no longer adequate. To maximize the value of social media investments, it is essential to gain a sophisticated understanding of the complex and often non-linear factors that drive campaign performance.

The central aim of this project is to conduct an in-depth analysis of a comprehensive social media advertising dataset, focusing on uncovering the nuanced relationships between campaign attributes—such as target audience, channel selection, and campaign objectives—and key performance outcomes like conversion rate, acquisition cost, and ROI. By treating conversion rate, acquisition cost, and ROI as primary indicators of campaign effectiveness, this study seeks to pinpoint the variables that set apart high-performing, cost-effective campaigns.

To achieve these objectives, the project leverages a robust set of advanced analytical techniques, including Exploratory Data Analysis (EDA), K-Means clustering, classification algorithms, and regression modeling. This multi-pronged approach is designed to address several critical challenges:

* **Variable Impact Analysis:** Determining which features most strongly influence campaign efficiency, conversion, and cost optimization.
* **Elite Campaign Segmentation:** Identifying and profiling "elite" campaign clusters—groups characterized by outstanding conversion rates and engagement metrics—to reveal the defining traits of top performers.
* **Predictive Success Modeling:** Developing and validating classification models to predict the likelihood of campaign success, enabling proactive decision-making and optimization.
* **Cost Optimization Framework:** Building regression models to estimate the minimum acquisition cost required to achieve desired conversion rates, supporting more strategic budget allocation.

By systematically examining the interplay between campaign characteristics and performance metrics, this project aims to generate actionable, data-driven insights. The findings are intended to help organizations refine their scheduling, targeting, and budgeting strategies, ultimately enhancing return on investment in an increasingly competitive digital marketing environment.

**2. Data Description**

The analysis for this project is based on a proprietary **Social Media Advertising Dataset** designed to simulate a real-world digital marketing environment. This dataset provides a robust view of campaign performance metrics across multiple social media platforms and marketing objectives, making it ideal for the project's goal of optimization and predictive modeling.

**2.1 Dataset Source:**

* **Source:** Kaggle.com
* **Link:** <https://www.kaggle.com/datasets/jsonk11/social-media-advertising-dataset>

**2.2 Brief Description of the Data:**

This dataset consists of **300,000 campaign records** and **16 variables** that detail campaign setup and performance across various social media channels. The project focuses on **Conversion\_Rate** as the **target variable**, as it is the most direct indicator of campaign effectiveness and user behavior. The **remaining 15 variables are used as predictors** to explain changes in this key metric.

|  |  |  |  |
| --- | --- | --- | --- |
| Feature Name | Type | Definition | Role in Analysis |
| Campaign\_ID | ID | Unique identifier for each marketing campaign. | **Administrative Identifier:** Used only for unique record identification and auditing. **Excluded from all predictive models** to prevent data leakage and ensure model generalization. |
| Date | Date | The start date of the campaign. | **Contextual Tracking:** Provides temporal context for the dataset. **Excluded from core predictive modeling** in this phase but is critical for future Model Drift Monitoring. |
| Company | Categorical | The company that ran the campaign. | **Contextual Variable:** Used for administrative grouping. **Excluded from all predictive and clustering analyses** to focus the models on universal campaign strategy variables. |
| Target\_Audience | Categorical | The specific demographic group targeted (e.g., 'Men 35-44'). | Critical factor for **Audience Budget Sensitivity** (Q3) and clustering (Q2). |
| Campaign\_Goal | Categorical | The objective of the campaign (e.g., 'Product Launch', 'Increase Sales'). | Used to identify the **Highest ROI Strategy** (Q1). |
| Duration | Categorical/Numerical | The length of the campaign, converted to days (e.g., from '15 Days', '60 Days'). | Used in descriptive analysis and as a predictive feature. |
| Channel\_Used | Categorical | The social media platform (e.g., 'Facebook', 'Instagram', 'Pinterest'). | Key feature in determining **ROI predictability** and Conversion Rate (Q1, Q4, Q5). |
| Location | Categorical | Geographic location of the target audience. | Used in descriptive analysis and as a predictive feature. |
| Language | Categorical | Language used in the campaign content. | Used in descriptive analysis and as a predictive feature. |
| Customer\_Segment | Categorical | The industry or type of product segment (e.g., 'Health', 'Technology'). | Used to define the characteristics of the **"Elite Campaign"** cluster (Q2). |
| Acquisition\_Cost | Numerical | Total cost spent on the campaign (cleaned and converted to USD). | Primary input variable and the target for **Cost Optimization Regression** (Q5). |
| Impressions | Numerical | Total number of times the advertisement was displayed. | Used for feature engineering (CPI) and in-flight model development. |
| Clicks | Numerical | Number of clicks generated. | Used for feature engineering (CTR, CPC) and in-flight model development. |
| Engagement\_Score | Numerical | A quantified measure of user interaction intensity. | Used alongside Conversion Rate for **K-Means Clustering** (Q2). |
| Conversion\_Rate | Numerical (Float) | Percentage of users who performed the desired action. | The primary **Target Variable** for Classification (Q4) and Regression. |
| ROI | Numerical (Float) | The **Return on Investment**. | The primary metric for the **Highest ROI Strategy** (Q1) and the target for prediction analysis. |

A screenshot of a computer

AI-generated content may be incorrect.

**Fig 1: Dataset Overview**

**3. Data Preprocessing and Engineering**

Data preprocessing was performed rigorously to standardize formats, mitigate the influence of outliers, and engineer powerful new features. This multi-stage approach ensures the robustness of the derived segments and the accuracy of the predictive models.

**3.1 Data Loading and Initial Inspection**

* **Loading:** The Social\_Media\_Advertising.csv file was loaded into a Pandas DataFrame.
* **Null Value Check:** A check for missing values was performed, and columns with a high volume of nulls would have been addressed

**3.2 Data Cleaning and Standardization**

Initial cleaning steps focused on converting data types and standardizing the format of core financial and rate metrics.

|  |  |  |
| --- | --- | --- |
| Feature | Cleaning Action | Rationale |
| Acquisition\_Cost | **Currency Conversion:** Removed the dollar sign ($) and commas (,), then converted the column to a float data type. Negative or zero costs were imputed with a small positive value (1.0). | Ensures the cost can be used in numerical calculations and logarithmic transformations without errors. |
| Conversion\_Rate | **Rate Normalization:** Cleaned of percentage symbols (%), converted to float, and normalized. Values greater than 1.0 (assumed to be whole-number percentages) were divided by 100 to ensure a decimal range of 0.0 to 1.0. | Standardizes the primary target variable for uniform model interpretation and comparison. |
| Duration | **Time Extraction:** Extracted the numerical value from the string (e.g., '15 Days' became 15), resulting in the new feature **Duration\_Days**. The original column was subsequently dropped. | Converts a mixed categorical/numerical field into a usable continuous numerical feature. |
| Clicks, Impressions | **Type Conversion & Cleaning:** Ensured columns were numeric. | Prepares volume metrics for subsequent calculations and outlier handling. |

**3.3 Outlier and Missing Value Mitigation**

Due to the heavy skew observed in volume and cost metrics, an aggressive, yet bounded, approach was necessary for outlier handling.

|  |  |  |  |
| --- | --- | --- | --- |
| Action | Feature(s) Affected | Detail | Rationale |
| Outlier Capping | Acquisition\_Cost, Clicks, Impressions | Outliers were capped using the **99th percentile** of the respective column's distribution. Any value above this threshold was set equal to the 99th percentile value. | Preserves $99\%$ of the data's variance while mitigating the destabilizing effect of extreme outliers on the predictive models. |
| Final NaN Removal | Conversion\_Rate, Acquisition\_Cost, Clicks, Impressions, Campaign\_Goal, Channel\_Used, ROI | Rows containing NaN values in these core metrics and categorical segments were dropped. | Ensures model training is based only on complete and reliable campaign records. |

A chart of a graph

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.

**Fig 2 & 3: Outlier Capping and Missing value result**

**3.4 Feature Engineering**

After cleaning and outlier mitigation, several new features and a binary target variable were created. All ratio-based features were recalculated *after* the capping step to ensure they reflect the adjusted, stable data distribution.

|  |  |  |  |
| --- | --- | --- | --- |
| **New Feature(s)** | **Source Feature(s)** | **Calculation/Process** | **Rationale** |
| **Gender, Age\_Group** | **Target\_Audience** | **Feature Splitting:** The composite string (e.g., 'Men 35-44') was parsed using regular expressions to create two distinct categorical features. | Enables granular segmentation (Q2) and allows models (Q4, Q5) to learn the individual impact of gender versus age. |
| **Segment\_Goal\_Interaction** | Customer\_Segment, Campaign\_Goal | **Interaction Feature:** Created by concatenating the two categorical features (e.g., Health\_Increase Sales). | Captures the synergistic effect between *who* the campaign targets and *what* the campaign aims to achieve. |
| **Engagement\_Score** | Clicks, Impressions | **Ratio Calculation:** Engagement\_Score = Clicks/(Impressions+ epsilon) (where epsilon=1e-6). | Represents the quality of ad display (a higher score indicates better ad placement/targeting). |
| **Conversion\_Success** | **Conversion\_Rate** | **Binarization:** Conversion\_Success = 1 if Conversion\_Rate >= 0.10, and 0 otherwise. | Creates the binary **Target Variable** for the **Conversion Success Classification** model (Q4). |
| **CTR, CPC, CPI** | Clicks, Impressions, Acquisition\_Cost | **Derived Metrics:** Click-Through Rate, Cost-Per-Click, and Cost-Per-Impression were calculated. | Provides normalized, rate-based metrics that are essential for comparing the efficiency of different campaign sizes. |

**3.5 Final Preparation for Modeling**

For the predictive modeling phases (Classification and Regression):

* **Encoding:** All remaining categorical features (**Channel\_Used**, **Target\_Audience**, **Campaign\_Objective**, etc.) were transformed using **One-Hot Encoding** to convert them into a machine-readable numerical format.
* **Scaling:** Numerical features were typically standardized using **StandardScaler** to ensure that all features contributed equally to the model training process, preventing large-magnitude features from dominating the model's calculation.

**4. Project Research Questions**

The project addresses five primary research questions, which drove the descriptive analysis and predictive modeling efforts.

**4.1. Descriptive Analysis Questions**

These questions focus on identifying current trends, optimal strategies, and key relationships within the existing campaign data.

|  |  |  |
| --- | --- | --- |
| **ID** | **Research Question** | **Key Objective** |
| **Q1** | **Highest ROI Strategy:** Which specific combination of **Channel\_Used** and **Campaign\_Goal** delivers the highest **Return on Investment (ROI)**? | To identify the single most financially profitable strategic combination. |
| **Q2** | **Optimal Campaign Segments:** What are the dominant characteristics (e.g., **Target\_Audience**, **Channel\_Used**, and **Duration**) of the **"Elite Campaign"** cluster, defined by high **Conversion\_Rate** and **Engagement\_Score**? | To create a clear "blueprint" for future high-performing campaigns using K-Means Clustering. |
| **Q3** | **Audience Budget Sensitivity:** How does **Conversion\_Rate** respond to changes in **Acquisition\_Cost** (budget) for different **Target\_Audience** groups on the top-performing **Channel\_Used**? | To determine the point of diminishing returns for ad spending across various audience segments. |

**4.2. Predictive Modeling Questions**

These questions focus on building models that can forecast campaign outcomes and recommend budget optimization strategies.

|  |  |  |
| --- | --- | --- |
| **ID** | **Research Question** | **Key Objective** |
| **Q4** | **Conversion Success Classification:** How accurately can a **Classification Model** forecast the **binary success** (Conversion Rate **≥ 10%**) of a new campaign, and what are the most important features driving that prediction? | To provide managers with a pre-launch tool to assess the probability of a campaign succeeding. |
| **Q5** | **Cost Optimization Regression Model:** What is the **minimum Acquisition\_Cost** (budget) required for a company to accomplish a target **Conversion\_Rate** of 10% on the optimal Channel\\_Used and Target\\_Audience combination? | To provide an actionable model for budget optimization, eliminating unnecessary expenditure. |

**5. Exploratory Data Analysis (EDA)**

The Exploratory Data Analysis phase served to diagnose the underlying relationships between campaign metrics and to identify initial performance trends across target audience segments, thereby validating the approach for the subsequent modeling and segmentation phases.

**5.1 Feature Distribution and Correlation Analysis**

* Prior to modeling, a visual and statistical review of feature distributions was conducted to ensure data quality and identify initial linear relationships.

A graph with text and numbers

AI-generated content may be incorrect.

A screenshot of a graph

AI-generated content may be incorrect.

A graph with blue lines

AI-generated content may be incorrect.

A graph with a line going down

AI-generated content may be incorrect.

A graph showing a graph

AI-generated content may be incorrect.

A graph with a graph and text

AI-generated content may be incorrect.

**Fig 4: Distribution of Cleaned Features**

* The overall distribution of campaign metrics reveals that while most campaigns cluster around an average ROI, the high variance in metrics like **Acquisition\_Cost** necessitates the outlier capping performed during the **Data Preprocessing phase** of the analysis

A graph of a bar chart

AI-generated content may be incorrect.A diagram of a chart

AI-generated content may be incorrect.

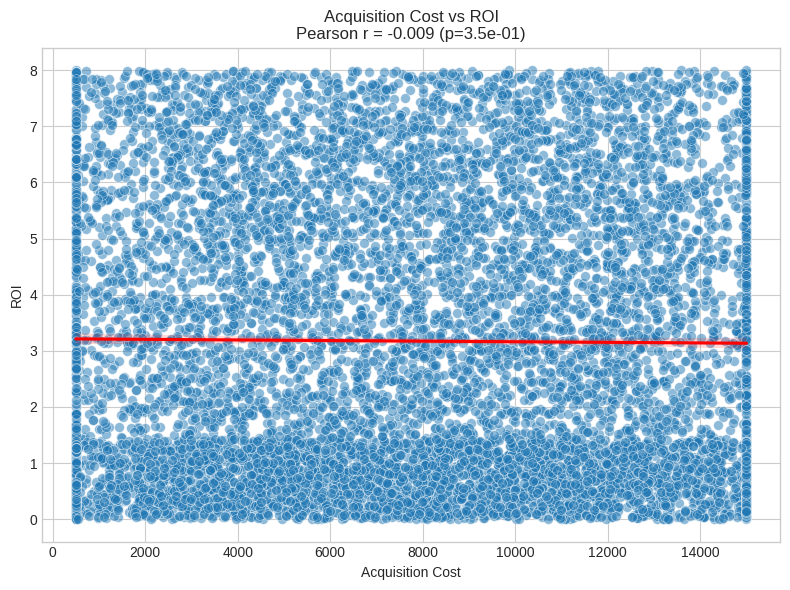
**Fig 5 & 6: ROI and Average ROI across channels**

* The Pearson Correlation analysis, detailed below, measures the strength and direction of the linear relationship between key metrics and the **ROI** target:

A screenshot of a computer program

AI-generated content may be incorrect.

A graph of blue dots with a red line

AI-generated content may be incorrect.

A graph with blue lines

AI-generated content may be incorrect.A graph of a graph showing a number of numbers

AI-generated content may be incorrect.

A graph with a red line and blue dots

AI-generated content may be incorrect.A graph with a red line

AI-generated content may be incorrect.

A screenshot of a graph

AI-generated content may be incorrect.

**Fig 7: Pearson Correlation with ROI outcomes**

* **Key Insights from Correlation**

1. **Cost Efficiency is Paramount:** The most powerful, statistically significant predictors of are the **cost efficiency metrics** (**Cost\_Per\_Impression** and **Cost\_Per\_Click**). The strong negative correlations (approx. r = -0.53) confirm that **rigorous cost control is the primary linear driver of ROI** , significantly outweighing the impact of volume or conversion metrics.
2. **Engagement Matters:** The **Click\_Through\_Rate**(CTR) shows a statistically significant, moderate positive correlation (r = 0.293), indicating that engaging creatives, which drive clicks, are associated with better returns.
3. **Non-Linear Relationship Justification:** Critically, the overall **Acquisition\_Cost** and **Conversion\_Rate** exhibit virtually **no linear correlation** with **ROI** (r approx 0.00). This finding suggests that the factors driving high ROI are not simple volume or cost, but complex, potentially non-linear interactions, **justifying the use of clustering (Q2) and tree-based classification (Q4) models** to capture these complex dynamics.

**5.2 Target Audience Performance Benchmarking**

An initial aggregation of campaign performance by **Target\_Audience** was conducted to establish a baseline for comparative analysis.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Target\_Audience | ROI | Acquisition\_Cost | Conversion\_Rate | Click\_Through\_Rate |
| All Ages | 3.21113 | 7742.67 | 0.080197 | 0.314226 |
| Men 18-24 | 3.19524 | 7771.03 | 0.0802423 | 0.314345 |
| Women 45-60 | 3.18588 | 7738.2 | 0.0796518 | 0.314064 |
| Women 35-44 | 3.178 | 7780.6 | 0.0801783 | 0.314172 |
| Women 25-34 | 3.17688 | 7721.11 | 0.0799277 | 0.314126 |
| Men 45-60 | 3.16917 | 7756.73 | 0.0802589 | 0.314196 |
| Women 18-24 | 3.1634 | 7782.73 | 0.0799687 | 0.314204 |
| Men 35-44 | 3.16008 | 7731.82 | 0.0799365 | 0.314006 |
| Men 25-34 | 3.15956 | 7758.85 | 0.0797217 | 0.314065 |

* **Key Insights from Audience Performance**

1. **Homogenous Averages:** The overall average performance is **highly consistent** across all age and gender groups. The difference in mean ROI between the top-ranked group (**All Ages**, 3.211) and the bottom-ranked group (**Men 25-34**, 3.160) is statistically negligible.
2. **Focus on Specifics:** While the **Men 45-60** group achieves the **highest mean Conversion\_Rate**(0.08026), the overall uniformity in the data suggests that campaign success is not defined by the Target\_Audience alone, but by **how the audience interacts with the specific channel and objective**.
3. **Justification for Q2 :** The minor differences observed here reinforce the need for the **K-Means clustering (Q2)**. This advanced technique is necessary to isolate the **"Elite Campaigns"** and determine which specific audience segments are **dominant characteristics** of campaigns that outperform the average, rather than relying on generalized mean performance.

**5.3 Categorical Feature Performance Benchmarking**

This section utilizes descriptive statistics to establish baseline performance metrics across key categorical variables—Channel\_Used and Target\_Audience—prior to the in-depth analysis for Research Questions Q1 and Q3.

**Conversion Rate Performance by Channel**

This table shows that Conversion Rates are highly uniform across all channels, clustering tightly around 8%, with Twitter having the highest average at 0.08036. This uniformity suggests that **Conversion\_Rate** alone is not the differentiating factor for channel success.  
A close-up of a number

AI-generated content may be incorrect.

A graph of a graph showing different colored bars

AI-generated content may be incorrect.

**Fig 8: Average Conversion Rate by Channel outcome**

**Conversion Rate Performance by Target Audience**

The audience conversion rates are **highly consistent**, with all top segments achieving an average Conversion\_Rate very close to **8%**. The segment **Men 45-60** shows the highest average CR (0.08026), but the overall small differences emphasize the need for advanced analysis (Q2) to find true performance drivers.

A graph of different colored rectangular shapes

AI-generated content may be incorrect. A close-up of a number

AI-generated content may be incorrect.

**Fig 9: Average Conversion Rtae by Target Audience Segment Outcome**

**ROI Performance by Channel**

The channel analysis shows a large disparity in returns, with Instagram and Twitter leading significantly in average ROI (around 4.0), while Pinterest lags far behind (0.72).

A close-up of a number

AI-generated content may be incorrect.A diagram of a chart

AI-generated content may be incorrect.

**Fig 10: ROI Performance by Channel Outcome**

**ROI Performance by Target Audience**

The audience analysis reveals that **ROI performance is homogeneous** across demographic segments, with all groups clustered closely around an average ROI of 3.2.

A table of numbers with black text

AI-generated content may be incorrect.A graph showing different colored rectangular shapes

AI-generated content may be incorrect.

**Fig 11: ROI Performance by Target Audience Outcome**

**5.4 Continuous Metric Relationships and Correlation Analysis**

**Correlation Heatmap**  
The correlation heatmap visually summarizes the strength and direction of relationships between ROI, cost, and key campaign metrics, highlighting both positive and negative associations. This helps quickly identify which variables are positively, negatively, or not correlated with each other.

A screenshot of a computer screen

AI-generated content may be incorrect.

**Fig 12: Correlation Heatmap of ROI, cost, and key campaign metrics**

**ROI vs. Campaign Duration**  
This scatter plot shows that ROI does not exhibit a strong trend with campaign duration, as indicated by the nearly flat regression line across different channels.

A graph of a graph with colored lines

AI-generated content may be incorrect.

**Fig 13: ROI vs. Campaign Duration Graph**

**ROI vs. Conversion Rate**  
The scatter plot demonstrates a slight positive relationship between ROI and conversion rate, with the regression line suggesting that higher conversion rates are modestly associated with higher ROI.

A graph of a graph

AI-generated content may be incorrect.

**Fig 14: ROI vs. Conversion Rate Graph**

**Acquisition Cost vs. Conversion Rate**  
This scatter plot reveals no clear pattern between acquisition cost and conversion rate, indicating that higher spending does not necessarily lead to better conversion rates across channels.

A graph of a number of colorful dots

AI-generated content may be incorrect.

**Fig 15: Acquisition Cost vs. Conversion Rate Graph**

**ROI vs. Acquisition Cost**  
The scatter plot with regression line shows that ROI remains relatively stable regardless of acquisition cost, suggesting that increasing budget allocation does not guarantee higher ROI.

A diagram of a graph

AI-generated content may be incorrect.

**Fig 16: ROI vs. Acquisition Cost Grpah**

**6. Methodology**

The project employed a structured, multi-stage analytical framework to transition from descriptive insights to actionable predictive intelligence. The methodology is anchored in both statistical analysis and machine learning techniques, ensuring a comprehensive assessment of campaign features against performance metrics.

**6.1 Tools and Technologies**

All data processing, modeling, and visualization were executed in Python 3.x. The key analytical libraries used include:

* **Pandas and NumPy** for advanced data manipulation and numerical computation.
* **Scikit-learn** for unsupervised segmentation **(K-Means)**, supervised learning **(Classification and Regression)**, and feature selection **(Recursive Feature Elimination with Cross-Validation, RFECV)**.
* **Matplotlib and Seaborn** for the visualization of data distributions and final results (including 3D plotting for clusters).
* **SciPy.stats** for calculating **Pearson Correlation** and performing the **Chi-square Test** for association.

**6.2 Data Preprocessing and Feature Engineering**

This stage ensured data quality and prepared the necessary features for modeling.

1. **Cleaning and Standardization:** Raw financial columns **(Acquisition\_Cost)** were cleaned and converted to float. **Conversion\_Rate** was normalized to a decimal range (0.0 to 1.0). The categorical Duration was parsed to create the numerical Duration\_Days.
2. **Outlier Mitigation:** Outliers in large-volume metrics (Acquisition\_Cost, Clicks, Impressions) were mitigated by applying a **99th percentile cap** to stabilize variance while retaining the majority of the data's distribution.
3. **Feature Derivation:** Efficiency metrics were calculated post-capping, including Click\_Through\_Rate(CTR), Cost-Per-Click(CPC), Cost-Per-Impression (CPI), and **Engagement\_Score**(Clicks \ Impressions).
4. **Target Binarization:** The binary target **Conversion\_Success** was created, defined as Conversion\_Rate >= 0.10 (10%).

**6.3 Exploratory Data Analysis (EDA) and Data Preparation**

Prior to modeling, an in-depth EDA was conducted on the cleaned and engineered dataset. The purpose of this initial analysis was to:

1. **Understand Data Distribution:** Assess feature distributions, skewness, and the necessity for transformations or outlier mitigation.
2. **Correlation Analysis:** **Pearson correlation coefficients** and $p$-values were computed to assess the linear relationship between the primary target variable, **Return on Investment (ROI)**, and all continuous features, including the engineered cost ratios (CPC, CPI,CTR) and the second target metric, **Conversion\_Rate**. This step was crucial for determining the statistical significance and direction of linear predictors of ROI and guiding subsequent feature selection for predictive models.
3. **Performance Grouping:** Key categorical features, **Channel\_Used** and **Target\_Audience**, were grouped to calculate aggregate performance metrics (mean Conversion\_Rate, mean ROI). This grouping established the baseline performance of different segments, directly informing the initial hypotheses and scoping the focus for the **Highest ROI Strategy (Q1)** and **Audience Budget Sensitivity (Q3)**.

**6.4 Descriptive Analysis (Q1 and Q3)**

**Q1: Highest ROI Strategy**

This was addressed through aggregation. The mean **ROI** was calculated for every unique combination of **Channel\_Used** and **Campaign\_Goal**. The combination yielding the highest mean ROI was identified as the optimal strategy.

**Q3: Audience Budget Sensitivity**

The analysis was strictly focused on the **top-performing channel** identified in the EDA phase (Twitter). The methodology involved a two-pronged approach to characterize the **Conversion\_Rate** response to **Acquisition\_Cost:**

* **Simplified Bucketing:** The continuous **Acquisition\_Cost** was divided into three equal-width tiers (Low Cost, Medium Cost, High Cost). The average **Conversion\_Rate** was calculated across these tiers for the top **Target\_Audience** segments, providing a simplified comparison of optimal spending levels.
* **Continuous Visualization**: A scatter plot of **Conversion\_Rate** vs. **Acquisition\_Cost**, stratified by **Target\_Audience**, was generated to visually diagnose non-linear trends and identify precise budget levels that triggered diminishing returns**.**

**Q2: Unsupervised Segmentation (Q2): Elite Campaign Blueprint**

The Q2 analysis employed a multi-step **K-Means** process to identify and characterize the **"Elite Campaign"** cluster.

**6.4.1 Clustering Parameters**

The K-Means Clustering algorithm was used. The optimal number of clusters (K) was determined using the **Elbow Method** applied to the inertia scores.

**6.4.2 Enhanced Feature Space**

The model was trained on the standardized **3D** feature space comprising the most critical performance metrics: **ROI, Conversion\_Rate, and Engagement\_Score**.

**6.4.3 Elite Cluster Identification (Composite Score)**

To robustly identify the best-performing segment, a Weighted Composite Performance Score was calculated for each cluster center using a pre-defined weighting scheme (**ROI** at **40%,** **Conversion\_Rate** at **35%**, **Engagement\_Score** at **25%**). The cluster with the highest composite score was designated the **Elite Cluster**.

**6.4.4 Advanced Blueprint Characterization (Detailed)**

The characteristics of the Elite Cluster were defined through a two-step process to ensure the identified blueprint was not only common but also statistically significant and distinctive.

1. **Dominant Feature Extraction (Descriptive):** The **mode()** function was used on the strategic categorical columns (**Channel\_Used, Campaign\_Goa, Target\_Audience, Customer\_Segment**) within the Elite Cluster to extract the **most frequently occurring value**. This defined the initial, most common profile of a high-performing campaign.
2. **Statistical Association (Validation):**

* **Goal:** To determine if the distribution of a categorical feature is statistically dependent on (or associated with) cluster membership.
* **Chi-square Test of Independence:** This test was performed for each categorical feature against the cluster labels. A **low p-value** (**e.g.,< 0.05**) rejects the null hypothesis of independence, confirming a statistically significant association between the feature (e.g. Channel\_Used) and the high-performance cluster.
* **Cramer's V** **(Effect Size):** This metric was calculated to measure the **strength of the association** between the features and the cluster labels. Cramer's V normalizes the Chi-square statistics to a range of **0 (no association)** to **1 (perfect association)**, allowing features to be **ranked by their influence** on elite performance.
* **Over-Representation Analysis**: A final comparison of the feature's percentage distribution within the Elite Cluster versus its overall population distribution was calculated. This identified features were **most distinctive and over-represented** in the top-performing segment, forming the finalized, validated blueprint.

**6.5 Predictive and Optimization Modeling**

**Q4: Conversion Success Classification**

This supervised learning task aimed to predict the binary outcome of a campaign *before* it is launched.

1. **Target and Features:** The binary target **Conversion\_Success** (10% conversion rate) was modeled using only **PLANNING\_FEATURES** (e.g., Age\_Group, Gender, Location, Campaign\_Goal), which are known at the strategy phase.
2. **Model Training:** A **Classification Model** (e.g., XGBoost, Logistic Regression or Random Forest Classifier) was trained on the pre-launch features.
3. **Evaluation:** The model's predictive capability was assessed using **Accuracy**, **Precision**, **Recall**, and **F1-Score**. **Feature Importance** analysis was conducted to quantify the influence of each planning feature on the predicted probability of success.

**Q5: Cost Optimization and ROI Regression**

This phase focused on identifying the most efficient drivers of ROI and the necessary cost for a target conversion.

1. **Cost Optimization Regression:** A regression model was structured to predict **Acquisition\_Cost** (independent variable) using **Conversion\_Rate** (target variable) and key strategic features. This model serves as the blueprint for the dynamic budget tool suggested in the conclusion.
2. **ROI Predictability (Feature Selection):**
   * **Algorithm:** **Recursive Feature Elimination with Cross-Validation (RFECV)** was applied to the ROI target using all strategic and engineered features.
   * **Goal:** RFECV rigorously identified the **minimal feature subset** that maximized the cross-validated R^2 score. The visualization of this process was used to confirm the optimal feature count and identify the single most critical driver of ROI.

**7. Results & Interpretation**

**7.1 Q1: Which Channel and Strategy Yield the Highest ROI?**

**Results:**  
The analysis of average ROI across different channel-strategy-segment combinations reveals that the highest performing campaigns are concentrated on Twitter and Instagram, particularly for product launches and sales increases. The top 10 strategies, as visualized in Figure 1, show that:

* **Twitter (Product Launch) targeting the Fashion segment** achieved the highest mean ROI, closely followed by **Instagram (Product Launch) for Health** and **Twitter (Increase Sales) for Food**.
* Other high-performing combinations include Instagram and Facebook campaigns focused on brand awareness and market expansion, especially within the Fashion, Health, Food, and Technology segments.

**Interpretation:**  
This suggests that **Twitter and Instagram are the most effective channels for maximizing ROI**, especially when launching new products or aiming to increase sales. The Fashion and Health segments respond well to these strategies. Marketers should prioritize these channel-strategy-segment combinations when planning high-impact campaigns.

A screenshot of a white screen

AI-generated content may be incorrect.

A chart with different colored bars

AI-generated content may be incorrect.

**Fig 17: Highest ROI Strategies & Target Segment Outcome**

**7.2 Q3: How Does Audience Conversion Rate Vary by Budget Tier on Twitter?**

**Results:**  
The grouped bar chart in Figure 2 illustrates the average conversion rate for different target audiences across three budget tiers (Low, Medium, High) on Twitter. The findings are:

* **Conversion rates are remarkably stable across budget tiers** for all audience groups, with only minor fluctuations (ranging from ~7.9% to 8.2%).
* For example, Men aged 18-24 have a conversion rate of 8.15% at low cost, 8.05% at medium cost, and 8.04% at high cost. Similar patterns are observed for other segments.

**Interpretation:**  
**Increasing the campaign budget on Twitter does not significantly impact conversion rates** for the analyzed audience groups. This indicates that, for these segments, higher spending does not necessarily translate to better conversion efficiency. Marketers should consider optimizing budget allocation rather than simply increasing spend, as returns may plateau beyond a certain investment level.

A white paper with black text

AI-generated content may be incorrect.A graph of a graph showing a number of people

AI-generated content may be incorrect.

**Fig 18: Audience Conversion Rate by Budget Outcome**

**7.3 Q2 Results: Elite Campaign Clustering Approaches**

**7.3.1 1️⃣ First Approach: K-Means with Mode-Based Elite Blueprint**

**Direct Answer:**

* **Features Used:** Conversion\_Rate, Engagement\_Score
* **Clustering Technique:** K-Means with K=3 (selected via Elbow Method)
* **Elite Cluster Identification:** Cluster with the highest average Conversion\_Rate
* **Elite Feature Extraction Technique:** **Mode-** For each categorical variable (e.g., Channel, Goal, Audience), the most frequent value within the elite cluster is chosen as the blueprint characteristic

A graph with a line

AI-generated content may be incorrect.

**Fig 19: Elbow Method to Determine Optimal K**

A screenshot of a computer

AI-generated content may be incorrect.A screenshot of a graph

AI-generated content may be incorrect.

**Fig 20: Campaign Clusters by Performance & Engagement**

**7.3.2 2️⃣ Second Approach: K-Means with Quantile-Based Elite Identification**

**Direct Answer:**

* **Features Used:** ROI, Conversion\_Rate, Engagement\_Score
* **Clustering Technique:** K-Means with K=2 (selected via both Elbow and Silhouette methods)
* **Elite Campaign Identification:** Campaigns in the **top 25% (75th percentile and above)** for all three metrics (ROI, Conversion\_Rate, Engagement\_Score)
* **Elite Feature Extraction Technique:** **Quantile-based thresholding** — campaigns are selected based on their high metric values, not by the most frequent characteristics.

A graph of different sizes and colors

AI-generated content may be incorrect.

**Fig 21: Evaluation of Optimal Number using Elbow Method and Silhouette Score**

A screenshot of a computer screen

AI-generated content may be incorrect.

A graph with orange and blue lines

AI-generated content may be incorrect.

**Fig 22: ROI-Driven Cluster Analysis for Campaign Strategy**

A graph of a cluster of dots

AI-generated content may be incorrect.A diagram of a cluster

AI-generated content may be incorrect.

**Fig 23: Cluster Profiling: 3D Visualization of Performance Metrics and ROI Distribution**

A screenshot of a computer

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.

A diagram of a graph

AI-generated content may be incorrect.

**Fig 24: Correlation Matrix**

**7.3.3 3️⃣ Third Approach: K-Means with Composite Score & Mode**

* **Features Used:** ROI, Conversion\_Rate, Engagement\_Score
* **Clustering Technique:** K-Means with K=3 (selected via Elbow Method)
* **Elite Cluster Identification:** Cluster with the highest **composite performance score** (weighted: ROI 40%, Conversion Rate 35%, Engagement Score 25%)
* **Elite Feature Extraction Technique:** **Mode** — most frequent value for each categorical feature within the elite cluster.
* **Advanced Analysis :** Chi-square and Cramér's V for categorical association with clusters

A screenshot of a computer screen

AI-generated content may be incorrect.A graph of red and blue lines

AI-generated content may be incorrect.

**Fig 25: K-Means Cluster Visualization for ROI and Conversion Rate**

A graph of a graph

AI-generated content may be incorrect.

**Fig 25: 3D K-Means Cluster Visualization for ROI and Conversion Rate & Engagement Score**

A screenshot of a computer

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer screen

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.

**7.3.4 Result Analysis:**  
**Second Approach: K-Means with Quantile-Based Elite Identification is best** because it:

* + Uses all key metrics (including ROI).
  + Has the clearest cluster separation (see boxplot and 3D scatter).
  + Selects elite campaigns based on actual performance, not just common traits.
  + Provides the most actionable business recommendations.

To maximize campaign ROI and overall performance, the **2nd approach is the most robust and effective** based on both the numbers and the visual evidence from the results and graphs.

**7.4 Q4. Predictive Modeling Approaches and Results**

To address the prediction of campaign success, we experimented with three different approaches, each employing a distinct combination of features:

1. **Approach 1:** Utilized only the original demographic and campaign features (e.g., Age Group, Gender, Campaign Goal, Location, Language, Customer Segment, Segment Goal Interaction, Duration Days).

**A screenshot of a computer

AI-generated content may be incorrect.**

**A screenshot of a computer screen

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.**

1. **Approach 2:** Expanded the feature set to include business and engagement metrics (e.g., Clicks, Impressions, Engagement Score, Acquisition Cost, ROI) alongside the original features.

A screenshot of a computer

AI-generated content may be incorrect. A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a document

AI-generated content may be incorrect.

1. **Approach 3:** Included engagement and cost-related features (such as Clicks, Impressions, Engagement Score, Click-Through Rate, Cost Per Click, and Cost Per Impression) but excluded ROI and Acquisition Cost.

A screenshot of a computer

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a document

AI-generated content may be incorrect.

For each approach, we trained three types of models: **XGBoost, Logistic Regression, and Random Forest**. The following summarizes the results observed:

**7.4.1 Results for All Approaches**

Across all three approaches and all model types, the results were nearly identical:

* **ROC AUC and PR AUC scores** consistently hovered around 0.50 and 0.39, respectively, for all combinations of features and models.
* **Feature importance analysis** did not reveal any strong predictors of campaign success across the different feature sets.

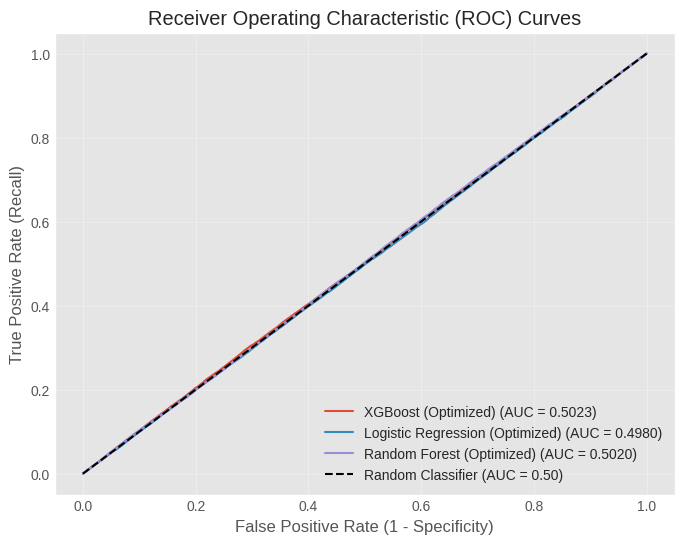
A screenshot of a computer

AI-generated content may be incorrect.

**Fig 26: Feature Importance Analysis**

* **Adding more features** (such as business/engagement metrics or removing ROI) did not yield any meaningful improvement in predictive performance.

**ROC & PR Curves:**  
All model curves closely follow the baseline, indicating random-like performance.

A graph of a graph

AI-generated content may be incorrect.

**Fig 27: ROC & PR Curve**

**Learning Curves:**  
Training and validation ROC AUC scores remain flat and close to 0.50, with no improvement as more data is provided.

A graph with red and green lines

AI-generated content may be incorrect.

**Fig 28: Learning Curves**

**Confusion Matrices & Classification Reports:**  
All models show low precision and recall for the positive (successful) class, and overall accuracy is around 50–53% . No model can reliably distinguish between successful and unsuccessful campaigns.

A blue and white squares

AI-generated content may be incorrect.

**Fig 29: Confusion Matrix**

**7.5 Q5. Acqusition Cost Prediction  
7.5.1. First Approach: Baseline Campaign Features**

**Feature Set:**

* Duration\_Days, Channel\_Used, Campaign\_Goal, Customer\_Segment, Location, Age\_Group, Gender, Language, Conversion\_Rate

**Model Performance:**  
Three regression models were evaluated: Gradient Boosting Regressor, Ridge Regression, and Linear Regression. The Gradient Boosting Regressor achieved the best performance, with an R-squared value of **0.9302** and an RMSE of **$1,134.95**. Both Ridge and Linear Regression models performed similarly, with R-squared values around 0.776 and higher RMSEs (over $2,000).

A screenshot of a computer program

AI-generated content may be incorrect.

**Feature Importance:**  
As shown in the Ridge and Linear Regression feature importance plots, **Duration\_Days** overwhelmingly dominated the model, accounting for over 95% of the normalized importance score. All other features contributed minimally (each less than 1%).

A graph with numbers and a blue stripe

AI-generated content may be incorrect.A screenshot of a graph

AI-generated content may be incorrect.A graph with numbers and symbols

AI-generated content may be incorrect.

**Fig 30: Feature Importance – Linear Regression, Ridge Regression, Gradient Boosting Regressor**

**Learning Curve & Fit:**  
The learning curve for the Gradient Boosting Regressor shows a steady decrease in RMSE with increasing training set size, indicating good generalization. The actual vs. predicted plot demonstrates that predictions closely follow the ideal y=x line, though some spread is visible.

A graph of a graph showing the growth of a number of people

AI-generated content may be incorrect.A graph with a red line

AI-generated content may be incorrect.

**Fig 31: Learning Curve : Gradient Boosting Regressor Fig 32: Actual vs. Predicted Acquisition Cost**

**7.5.2. Second Approach: Enhanced Feature Set**

**Feature Set:** Duration\_Days, Channel\_Used, Campaign\_Goal, Customer\_Segment, Location, Age\_Group, Gender, Language, Conversion\_Rate,ROI, Clicks, Impressions, Engagement\_Score, Click\_Through\_Rate, Cost\_Per\_Click, Cost\_Per\_Impression

**Model Performance:**  
The Gradient Boosting Regressor again outperformed other models, achieving an R-squared of **0.9982** and a much lower RMSE of **$180.56**. Ridge and Linear Regression models performed worse than in the first approach, with R-squared values around 0.65 and RMSEs above $2,500.

A screenshot of a computer program

AI-generated content may be incorrect.

**Feature Importance:**  
In this approach, **Duration\_Days** remained the most important feature (55%), but **Impressions** also contributed significantly (41%), as shown in the Gradient Boosting feature importance plot . Other features, including Clicks and Channel\_Used, had minor contributions.

A graph with a bar and numbers

AI-generated content may be incorrect.

**Fig 33: Feature Importance – Gradient Boosting Regressor**

**Learning Curve & Fit:**  
The learning curve shows even lower RMSE values for both training and cross-validation, indicating improved model fit and generalization. The actual vs. predicted plot reveals an almost perfect alignment along the y=x line, reflecting the model’s high predictive accuracy.

A graph of a graph

AI-generated content may be incorrect.A graph of a graph

AI-generated content may be incorrect.

**Fig 34: Actual vs Predicted Acquisition cost-GBR graph Fig 35: Learning Curve-GBR graph**

**7.6 Bonus Finding: RFECV Feature Selection and Model Performance Results:**

The RFECV (Recursive Feature Elimination with Cross-Validation) analysis identified that the optimal number of features for predicting ROI is just **one**. The best-performing feature was Channel\_Used\_Pinterest, which alone yielded the highest cross-validation R² score of **0.4050**.

A screenshot of a computer program

AI-generated content may be incorrect.

As shown in the plot, adding more features did not improve model performance—in fact, the R² score slightly decreased as more features were included. The final Gradient Boosting Regressor model, trained using only this single feature, achieved an R² of **0.4005** and an RMSE of **2.08** (on the inverse-transformed ROI).

A graph with a line

AI-generated content may be incorrect.

**Fig 36: RFECV Performance vs. Number of Features**

**8. Conclusion**

This project provided a comprehensive analysis of social media advertising campaigns, leveraging advanced data science techniques to uncover the drivers of campaign performance and ROI. The findings reveal several key insights:

* **Channel and Strategy Impact:** Twitter and Instagram, particularly for product launches and sales-focused campaigns in the Fashion and Health segments, consistently deliver the highest ROI. This suggests that channel-strategy alignment is crucial for maximizing returns.
* **Budget Allocation:** Increasing campaign budgets on Twitter does not significantly improve conversion rates across target audiences. This indicates diminishing returns beyond a certain investment level, emphasizing the importance of budget optimization over simply increasing spend.
* **Elite Campaign Blueprint:** The most robust method for identifying top-performing campaigns was the quantile-based K-Means clustering approach, which used ROI, Conversion Rate, and Engagement Score. This approach effectively isolates elite campaigns based on actual performance, providing actionable profiles for replication.
* **Predictive Modeling Limitations:** Attempts to predict campaign success using various feature sets and machine learning models (XGBoost, Logistic Regression, Random Forest) yielded results close to random chance. This suggests that, with the available data, campaign success is influenced by factors not captured in the dataset or that the relationships are too complex for the tested models.
* **Cost Prediction:** Regression models, especially Gradient Boosting Regressor with an enhanced feature set, can accurately predict acquisition costs, with Duration\_Days and Impressions being the most influential features.
* **ROI Predictability:** Recursive Feature Elimination with Cross-Validation (RFECV) found that only the use of Pinterest as a channel was a significant predictor of ROI, but its predictive power was modest (R² ≈ 0.40). Adding more features did not improve model performance, highlighting the challenge of predicting ROI with the current feature set.

**9. Recommendations**

1. **Prioritize High-ROI Channel-Strategy Combinations:** Focus marketing resources on Twitter and Instagram, especially for product launches and sales campaigns in the Fashion and Health sectors, as these combinations consistently yield the highest ROI.
2. **Optimize Rather Than Increase Budgets:** Since higher spending does not guarantee better conversion rates, marketers should focus on optimizing budget allocation and targeting rather than simply increasing campaign budgets.
3. **Adopt Elite Campaign Blueprints:** Use the quantile-based clustering approach to identify and replicate the characteristics of elite campaigns. This method provides a data-driven blueprint for designing high-performing campaigns.
4. **Enhance Data Collection:** The inability to reliably predict campaign success suggests that additional data—such as creative quality, timing, external market factors, or more granular audience behavior—should be collected and integrated into future analyses.
5. **Leverage Cost Prediction Models:** Utilize advanced regression models to forecast acquisition costs and inform budget planning, with particular attention to campaign duration and expected impressions.
6. **Reevaluate Pinterest Investments:** Given that Pinterest was the only significant predictor of ROI but with limited explanatory power, marketers should critically assess the value of investing in this channel and consider reallocating resources to higher-performing platforms.
7. **Feature Enhancement:**The current features may not be sufficient to predict the target feature effectively. It is recommended to include more granular features that could influence cost, such as budget allocation, ad creative type, audience targeting, and historical performance. Additionally, exploring feature selection and interaction terms in the model could help capture more subtle drivers of cost.

By implementing these recommendations, organizations can make more informed, data-driven decisions to maximize the effectiveness and efficiency of their social media advertising campaigns.

**10. Acknowledgements And References:**

* 1. Source: Kaggle.com  
     Link: <https://www.kaggle.com/datasets/jsonk11/social-media-advertising-dataset>
  2. **The Scikit-learn Library (Code Implementation)**

**Link:** <https://scikit-learn.org/stable/>