Optimizing Advertising Spend with Data Science and Machine Learning

Adewale Adeyemo | December 2024

Executive Summary

This report outlines a data-driven approach to optimizing advertising spend for an online advertising firm, utilizing data science and machine learning to improve ROI while maintaining or improving engagement. The project involves data preprocessing, exploratory data analysis (EDA), statistical analysis, and machine learning model development to provide actionable insights and recommendations for budget allocation.

Problem Statement

In the current advertising landscape, the company faces challenges in optimizing its Return on Investment (ROI) for ad campaigns and placements. While Campaign 2 has the best Click-Through Rate (CTR), Campaign 1 delivers the highest ROI, highlighting a gap between engagement and profitability. Placement "ghi" generates the best yield but is also the most expensive, underscoring the need for cost-effective strategies.

Key Issues Identified:

1. Campaign efficiency varies significantly across metrics like CTR and ROI.
2. High-cost placements like "ghi" yield high returns but strain the budget.
3. Inefficient allocation of advertising spend reduces profitability.

Key Challenge: Efficiently allocate advertising spend to achieve higher ROI without compromising user engagement.

Project Objectives

1. Analyze marketing spend data to identify drivers of ROI.
2. Optimize budget allocation by identifying cost-effective campaigns and placements.
3. Build a predictive model to forecast ROI, enabling data-driven decision-making.

Initial Approach and Tools

Libraries Used:

- numpy: Numerical operations and data handling.
- pandas: Data manipulation and preprocessing.
- seaborn and matplotlib: Data visualization.
- sklearn: Machine learning model building and evaluation.
- joblib: Model serialization for deployment.

Data Preprocessing

Data Overview

Dataset Summary:

- **Total Entries:** 15,408
- **Columns:** 14 (e.g., month, day, cost, clicks, revenue, placement, etc.)
- **Irrelevant Columns Removed:** Unnamed: 12, Unnamed: 13
- **Null Values:** Handled for the placement column (413 missing values removed).
- **Duplicates:** Removed 5 duplicate records to ensure data integrity.

Feature Engineering: New features were created to derive insights:

- **CPC (Cost per Click):** Measures cost efficiency per click.
- **CTR (Click-Through Rate):** Measures engagement relative to impressions.
- **RPC (Revenue per Click):** Measures revenue efficiency per click.
- **ROI (Return on Investment):** Evaluates profitability of campaigns and placements.

Exploratory Data Analysis (EDA)

EDA was performed to identify actionable insights. Key visualizations and findings include:

Cost vs Revenue (Scatter Plot): Positive correlation, indicating higher costs lead to higher revenue, but with diminishing returns for some campaigns.

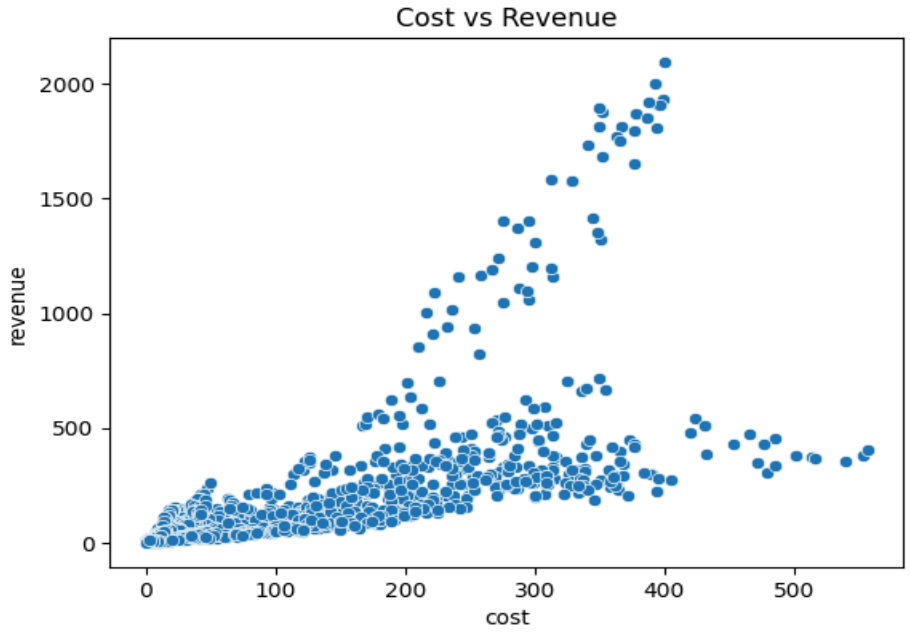

Fig 1. Cost vs Revenue

Statistical Analysis

1. **Inferential Statistics:**
    - Heatmaps revealed strong correlations between cost, displays, and ROI.
    - ROI shows strong dependence on placements and campaign type.
2. **Descriptive Statistics:**
    - Campaign 1 has the highest mean ROI, but with significant variability.
    - Campaign 2 has lower costs and better CTR but needs optimization to improve yield.

Inferential Analysis

Yield by Placement (Bar Plot): Placement "ghi" yields the highest ROI, but it's also the most expensive.

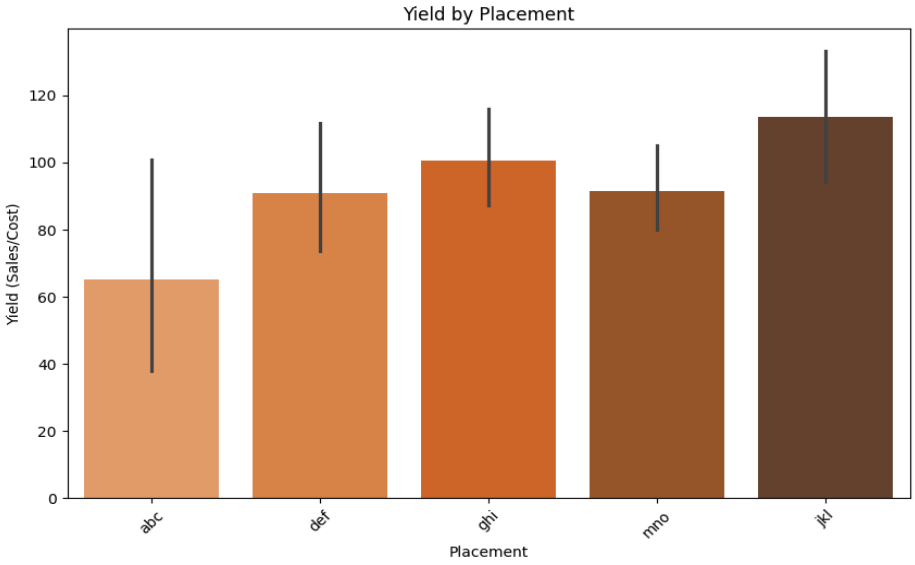

Fig 2. Yield vs Placement

Yield by Campaign: Campaign 1 outperforms others, providing 3x more yield than Campaign 2.

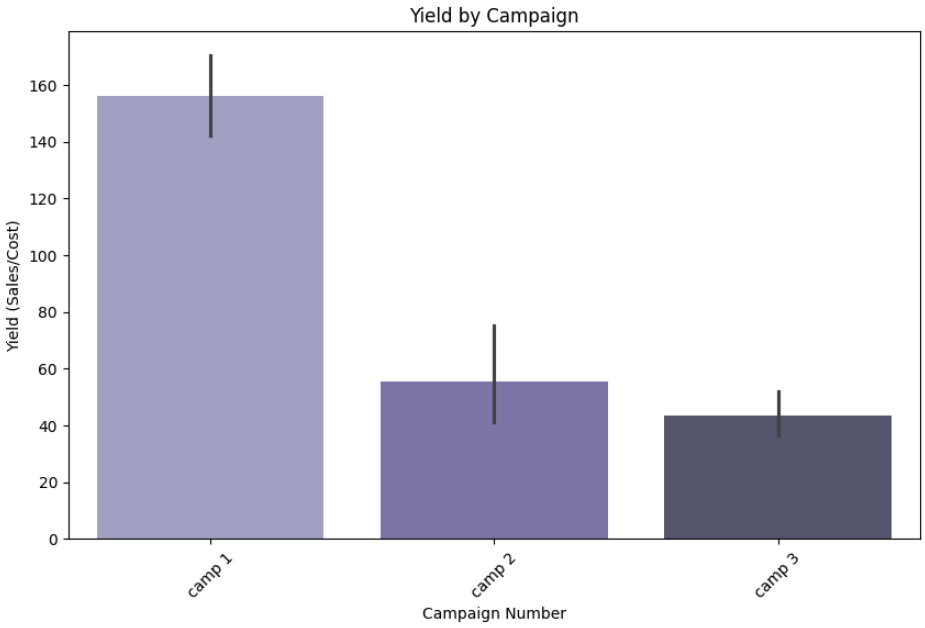

Fig 3. Yield vs Campaign

Yield by Placement (Bar Plot): Placement "ghi" is the most expensive; "abc" is the cheapest, suggesting areas for cost optimization.

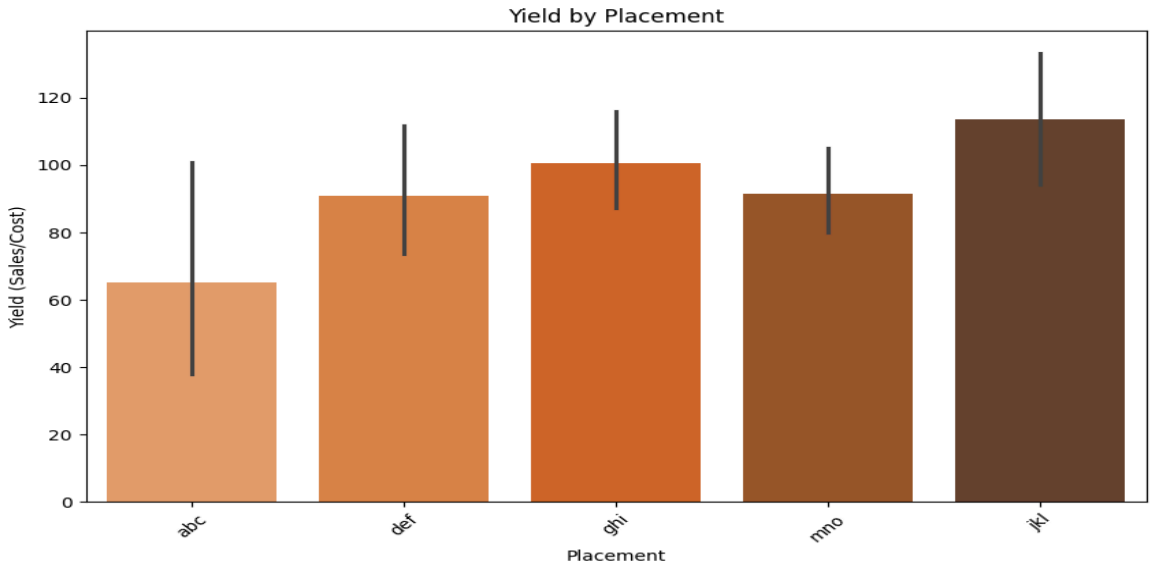

Fig 4. Yield vs Placement

The correlation matrix highlights significant relationships between numerical features. For example, post_click_sales_amount shows a strong positive correlation with revenue (0.89) and ROI (0.38), indicating that higher sales lead to increased profitability. Conversely, features like CPC have weak or negative correlations with ROI, suggesting that lowering costs per click could improve returns. These insights guide feature prioritization for the model.

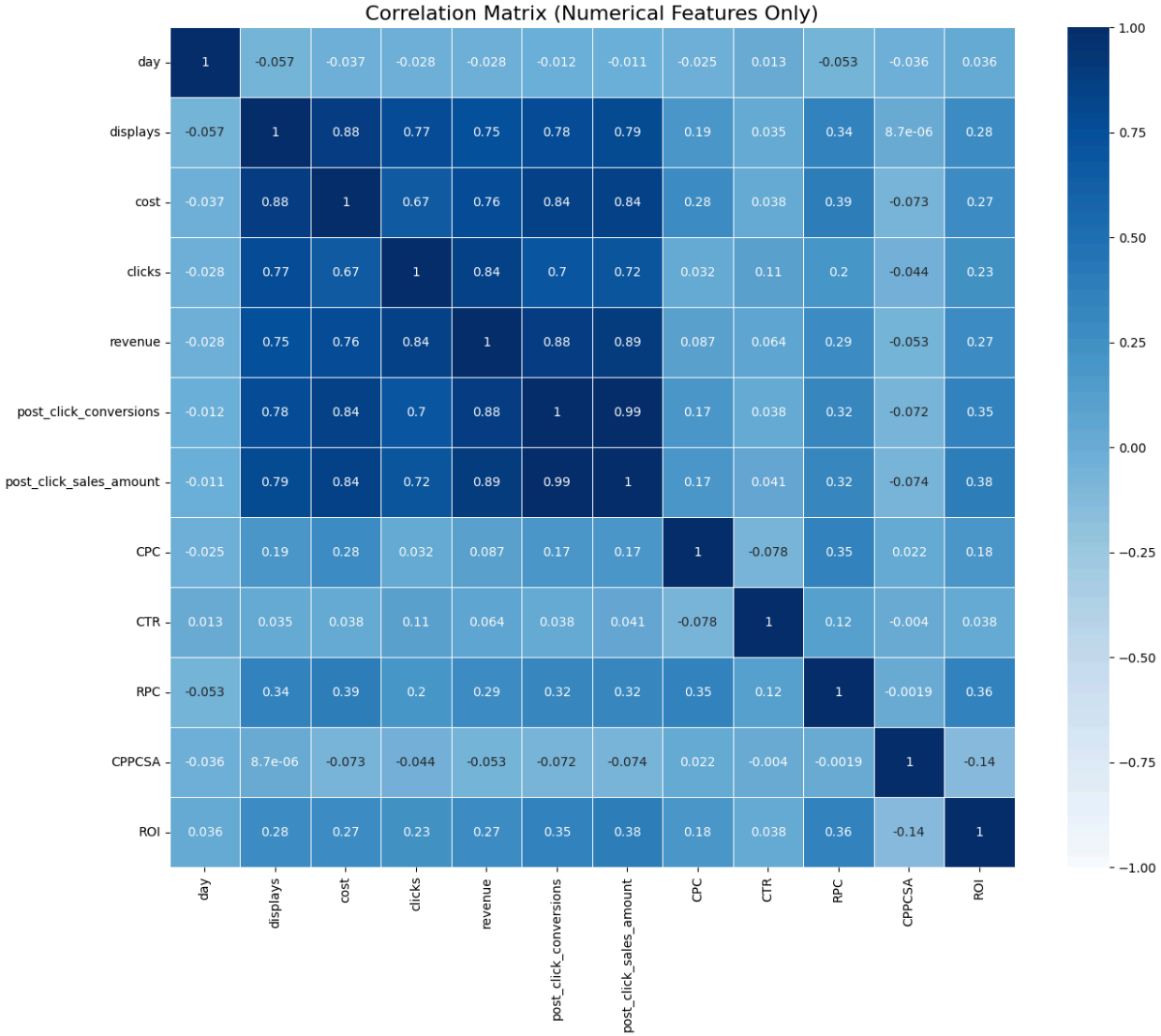

Fig 5. Correlation Matrix

Descriptive Statistics Analysis

Engagement vs Clicks (Bar Plot): High engagement consistently drives more clicks.

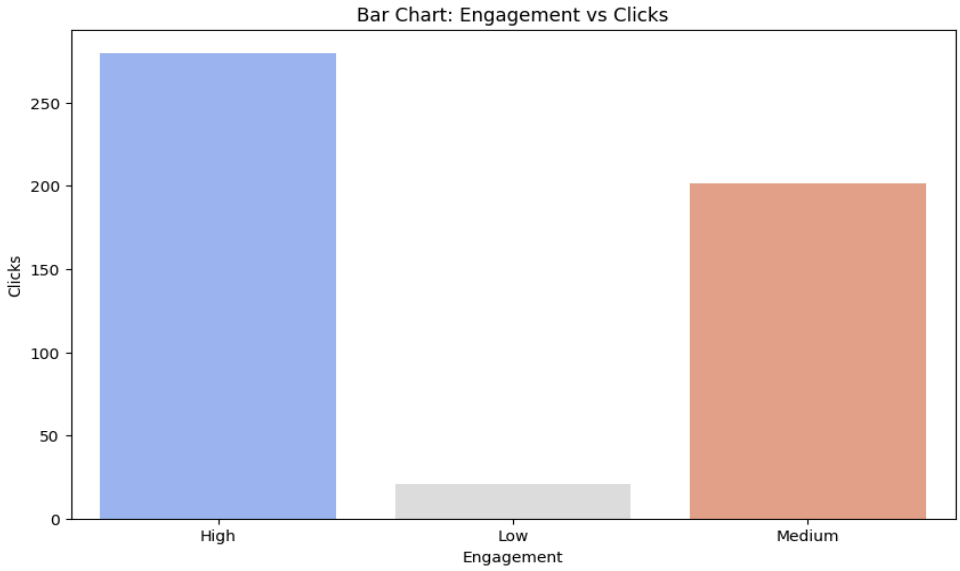

Fig 6. Engagement vs Clicks

CTR by Campaign (Bar Plot): Campaign 2 has the highest CTR, indicating strong engagement potential.

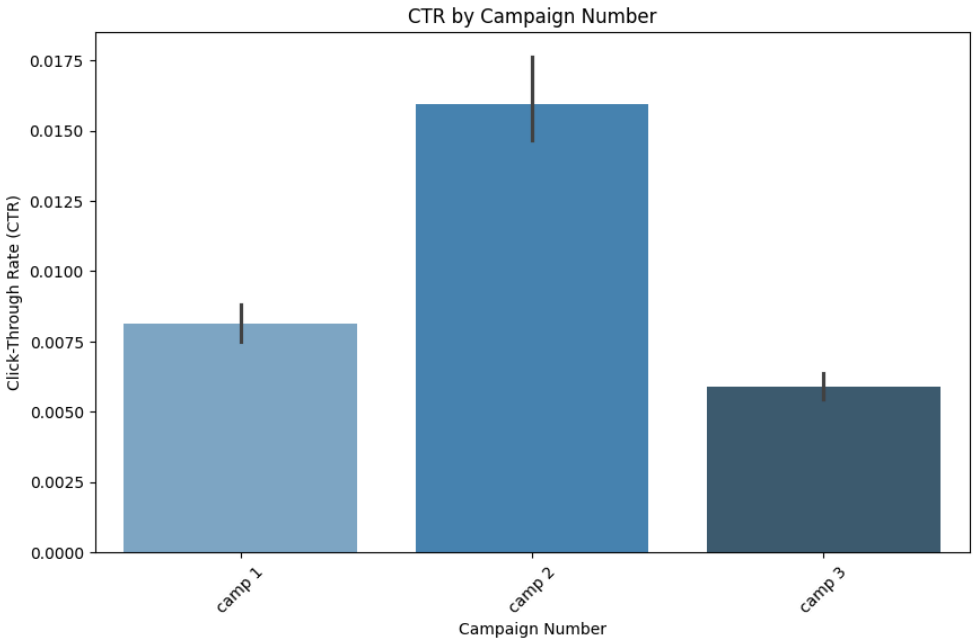

Fig 7. CTR by Campaigns

Machine Learning Model

Model Selection and Building:

Model Used: Random Forest Regressor

- Chosen for its ability to capture complex, non-linear relationships.
- Features: cost, clicks, displays, user_engagement, campaign_encoded, placement_encoded.
- Target: ROI

Steps Taken:

1. Encoded categorical features like placement, campaign_number, and user_engagement.
2. Split data into training (80%) and testing (20%) sets.
3. Trained the Random Forest model with default parameters (100 estimators).

Initial Model Evaluation:

- **RMSE:** 83.07
- **R^2 Score:** 0.37
- Insights: The model explains 37% of variance in ROI, highlighting room for improvement.

Model Tuning and Improvement:

1. **Hyperparameter Tuning:** Used RandomizedSearchCV to optimize parameters like n_estimators and max_depth.
2. **Cross-Validation:** Ensured generalizability and reduced overfitting.
3. **Feature Importance Analysis:** Identified cost, displays, and clicks as the most significant predictors of ROI.

Improved Model Performance:

- **RMSE:** 69.27
- **R^2 Score:** 0.56
- Insights: The model now explains 56% of the variance in ROI, a significant improvement.

Insights and Recommendations

Key Insights:

1. **Campaign Optimization:** Campaign 1 is the most profitable; reallocate budgets to this campaign for higher ROI.
2. **Placement Strategy:** Reduce spending on "ghi" and focus on cost-effective placements like "abc".

3. **Engagement Analysis:** High engagement directly correlates with clicks and revenue; prioritize campaigns that drive engagement.

Actionable Recommendations:

1. Use the ROI prediction model to guide budget allocation decisions.
2. Optimize Campaign 2 by leveraging its high CTR to improve profitability.
3. Regularly update the model with new data to ensure relevance.

Future Possibilities

1. **Incorporating Seasonality:** Analyze how monthly trends affect campaign performance.
2. **Granular Campaign Insights:** Evaluate individual ad creatives within campaigns for further optimization.
3. **Integration with Real-Time Data:** Deploy the model for live ROI predictions during campaigns.

Model Deployment

1. **Deploy on AWS:** Use AWS Lambda and SageMaker for scalable model serving.
2. **Create a Dashboard:** Provide stakeholders with an intuitive dashboard for ROI predictions and recommendations.
3. **Regular Model Maintenance:** Monitor model performance and retrain periodically with new data.

Conclusion

The analysis and model development provide a robust framework for optimizing marketing spends and maximizing ROI. By reallocating budgets and focusing on cost-effective strategies, the company can achieve higher profitability and efficiency. The ROI prediction model is a valuable tool for data-driven decision-making and continuous improvement in advertising performance.

Appendices

1. Sample Visualizations:

- Scatter plot of Cost vs Revenue
- Bar plot of Yield by Campaign

2. Model Performance Metrics:

- RMSE: 69.27
- R^2 Score: 0.56

3. Code Snippets:

- Feature engineering code
- Hyperparameter tuning implementation

4. Dataset Summary:

- Pre-processed data structure and key statistics