Global Advertising Performance

Project Title

Global Advertising Performance Analysis Using Machine Learning

Group Members and Roles

- **Namrata Bhoyar –** Preprocessing & Feature Engineering
- **Nihal Pujari –** Statistical Analyst
- **Anuj Kamble –** Noise Injection & Cleaning
- **Gourav Somanna –** Supervised Learning
- **Pramodkumar Shivanna –** Unsupervised learning

1. Introduction

Digital advertising platforms such as TikTok, Meta, and Google generate vast amounts of performance data daily. Businesses rely on this data to optimise campaign spending, maximise conversions, and improve profitability. However, campaign performance is influenced by multiple interacting factors including engagement metrics, platform type, industry category, and geographical region.

However, advertising performance data presents several analytical challenges. First, campaign outcomes are influenced by multiple interacting variables, including engagement metrics, platform dynamics, industry characteristics, and geographical factors. Second, financial variables such as ad spend and revenue are often skewed and contain extreme values, which complicates statistical modeling. Finally, real-world marketing data is rarely perfect; it may include measurement errors, reporting inconsistencies, and noisy observations.

To address these challenges, this project integrates statistical inference with machine learning techniques to develop a comprehensive analytical framework for advertising performance evaluation. By combining hypothesis testing, regression modeling, classification analysis, and clustering methods, we aim to move beyond descriptive analytics and toward predictive and strategic insights.

Specifically, this study seeks to:

- Statistically evaluate differences in platform performance,
- Predict future campaign conversions,
- Classify campaigns by profitability,
- Identify latent performance segments within the data.

The overall goal is to construct a robust, interpretable, and scalable framework that can support data-driven decision-making in digital advertising.

2. Dataset Description

The dataset used in this study is the **Global Ads Performance Dataset** obtained from Kaggle. It contains over 500+ campaign-level observations across 14 features.

2.1 Feature Types

Categorical Features

- Platform (TikTok, Meta, Google)
- Industry
- Country
- Campaign Type

Numerical Features:

- Impressions
- Clicks
- Ad Spend
- Conversions
- Revenue
- ROAS
- CTR
- CPC
- CPA

Temporal Feature:

- Date (used to extract month, week, and weekday)

The dataset is structured and tabular, suitable for both supervised and unsupervised machine learning tasks.

3. Problem Statements

The project addresses four primary objectives:

1. **Platform ROI Comparison**
   To test whether Return on Ad Spend (ROAS) differs significantly across advertising platforms.

2. **Conversion Forecasting (Regression)**
   To develop supervised learning models that predict campaign conversion counts.

3. **Profitability Classification**
   To classify campaigns as profitable (ROAS > 1) using engagement and campaign features.

4. **Market Segment Discovery (Unsupervised Learning)**
   To identify hidden campaign performance segments using dimensionality reduction and clustering.

4. Data Preprocessing & Feature Engineering

Data preprocessing was conducted to ensure model robustness and statistical validity.

4.1 Handling Missing Values

- Numerical features were imputed using the **median** to reduce the impact of skewed distributions.

- Categorical features were imputed using the **max**.

Median imputation was preferred over mean due to the presence of extreme values in financial variables such as ad spend and revenue.

4.2 Feature Engineering

- Extracted time-based features: Month, Week, and Day of Week.

- Recalculated CTR, CPC, and CPA to validate metric integrity.

- Applied **quantile clipping (1%–99%)** to limit extreme outliers.
- Standardised numerical features to improve model stability and clustering accuracy.

5. Noise Injection & Robustness Testing

To simulate real-world data imperfections, 3% multiplicative Gaussian noise was injected into financial and engagement features.

5.1 Purpose

- Simulate logging and tracking errors
- Evaluate model robustness
- Test effectiveness of cleaning methods

5.2 Cleaning Method

Quantile-based smoothing (winsorization) was applied to mitigate extreme variations introduced by noise.

Comparative analysis between original, noisy, and cleaned distributions showed that cleaning preserved central tendencies while reducing artificial variance.

6. Statistical Analysis

6.1 Descriptive Statistics

We computed:

- Mean
- Variance
- Skewness
- Correlation matrix

Findings:

- Financial metrics exhibited right-skewed distributions.
- Strong positive correlation was observed between ad spend and revenue.

6.2 Hypothesis Testing

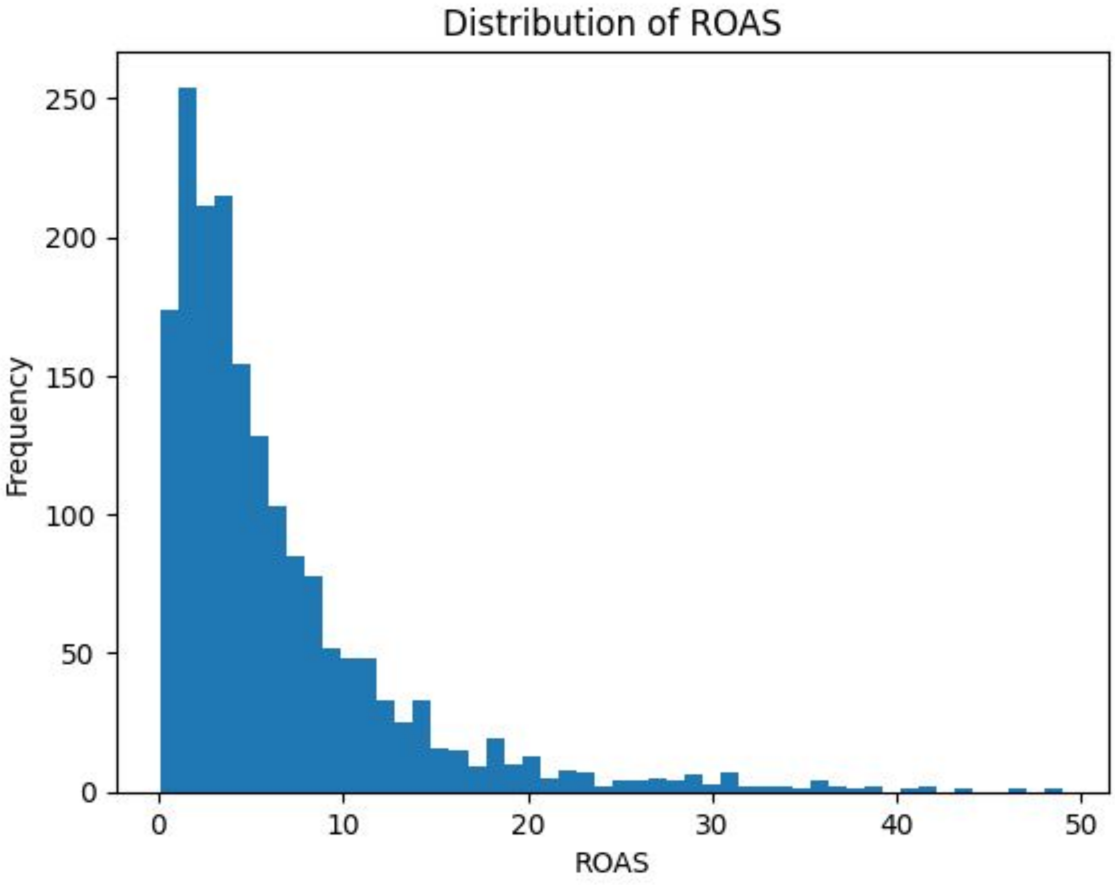
A **Kruskal–Wallis H-test** was applied to evaluate whether ROAS differed significantly across platforms.

- Null Hypothesis: ROAS distributions are identical across platforms.
- Result: $p < 0.05$
- Conclusion: Significant differences exist between platforms.

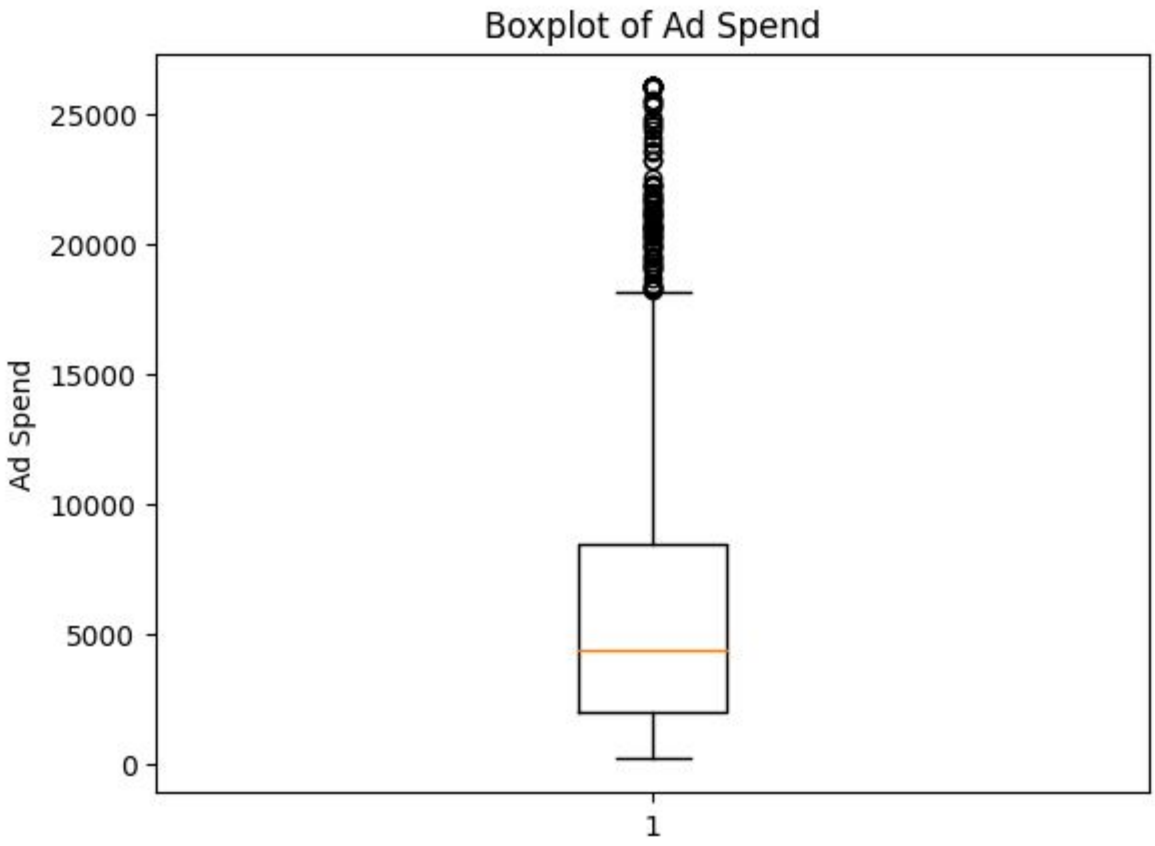
This validates that platform selection influences return efficiency.

Statistical plots used

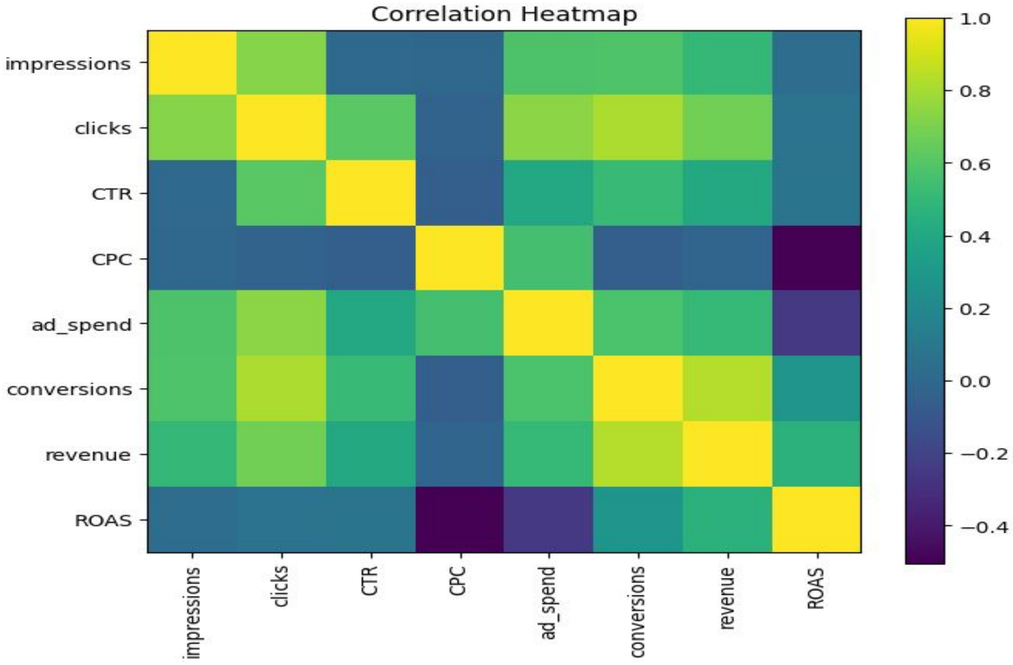
1. **Histogram of ROAS**



2 Boxplot of AD Spend



3 Correlation heatmap of numerical features



7. Supervised Learning

7.1 Regression – Conversion Forecasting

Objective:

Predict campaign conversion counts.

Models Compared:

- Ridge Regression
- Random Forest Regressor

Evaluation Metrics:

- MAE
- RMSE
- R^2
- 5-fold Cross-Validation

Results:

Random Forest outperformed Ridge Regression, achieving approximately 93% predictive accuracy (based on R^2). The model captured nonlinear relationships between engagement metrics and conversion outcomes.

7.2 Classification – Profitability Prediction

Objective:

Predict whether ROAS > 1.

Model Used:

- Random Forest Classifier

Evaluation Metrics:

- Accuracy
- F1-score
- ROC-AUC

Results:

The classifier demonstrated high ROC-AUC and maintained strong performance even when tested on noisy data, confirming robustness.

8. Unsupervised Learning

8.1 Dimensionality Reduction

Principal Component Analysis (PCA) reduced multi-dimensional features into two principal components for visualisation and clustering stability.

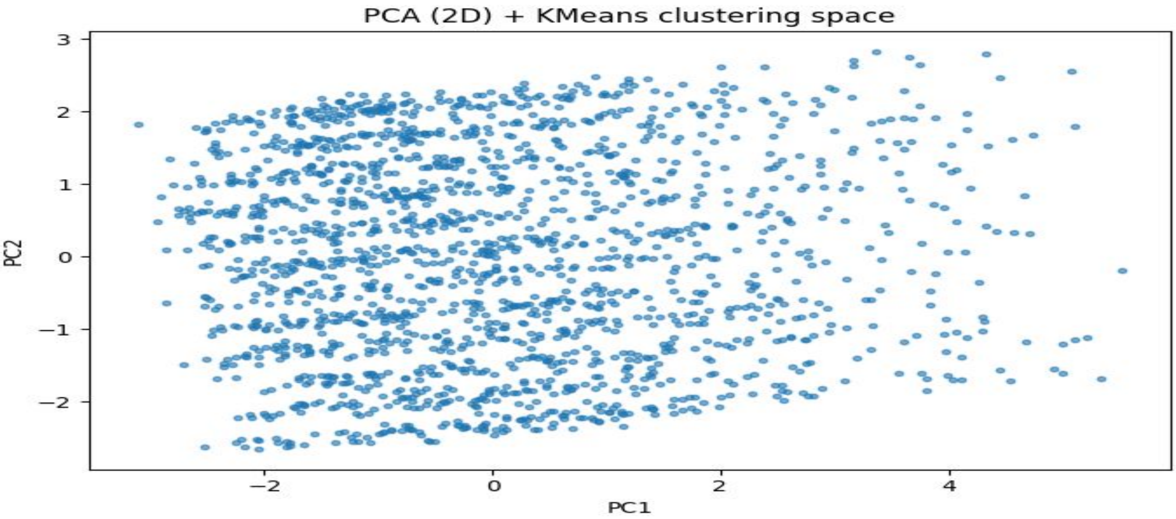
8.2 K-Means Clustering

Silhouette scores were used to determine the optimal number of clusters.

Cluster Interpretation

- **Cluster 0:** High-efficiency, low-spend campaigns (viral potential)
- **Cluster 1:** High-spend, large-scale campaigns
- **Cluster 2:** Underperforming campaigns requiring optimisation

This segmentation supports targeted strategy refinement.



9. Justification of Methodology

The methodology used in this study was selected based on the structure of the dataset and the nature of the research objectives.

First, a non-parametric statistical test (Kruskal–Wallis) was chosen to evaluate differences in ROAS across platforms. Financial metrics in advertising are often right-skewed and do not follow normal distributions. Therefore, non-parametric methods provide more reliable inference than traditional ANOVA in this context.

For predictive modeling, both regression and classification approaches were implemented to address different business questions. Regression was used to predict conversion counts, a continuous outcome variable essential for forecasting and budget planning. Ridge Regression served as a linear baseline model, while Random Forest was selected for its ability to capture nonlinear relationships and complex feature interactions. Advertising data often contains nonlinear patterns, making ensemble tree-based methods particularly suitable.

For profitability prediction, a Random Forest classifier was used due to its robustness, resistance to overfitting, and effectiveness in handling mixed data types. The classifier was evaluated using ROC-AUC and F1-score to ensure reliable performance beyond simple accuracy measures.

Dimensionality reduction using Principal Component Analysis (PCA) was applied before clustering to address the high dimensionality introduced by feature engineering and encoding. PCA improves clustering stability and enhances visualisation. KMeans clustering was then employed due to its interpretability and efficiency in identifying segment structures within standardised numerical data. The optimal number of clusters was selected using silhouette analysis to ensure well-separated groupings.

Additionally, controlled noise injection was introduced to simulate real-world measurement variability. This robustness testing demonstrates that the developed models maintain stability under moderate data perturbations, enhancing the practical reliability of the framework.

Overall, the selected methods balance statistical rigor, computational efficiency, interpretability, and real-world applicability.

10. Key Findings

- Platform choice significantly affects ROAS.
- Engagement metrics strongly predict conversions.
- Profitability can be predicted using pre-revenue indicators.
- Campaigns naturally cluster into distinct performance archetypes.
- The machine learning framework remains stable under moderate noise conditions.

11. Conclusion

This study integrates statistical testing and machine learning to analyse global advertising performance. The combination of regression, classification, and clustering provides both predictive and strategic insights.

The framework can support automated budget allocation, platform optimisation, and campaign evaluation. Future work may include real-time API integration and advanced ensemble methods for further performance improvements.

References

- Kaggle: Global Ads Performance Dataset
- Scikit-learn Documentation
- Statistical Learning Literature