

Customer Segmentation & Behavior Analysis Using Data Mining Techniques

Problem Statement:-

Digital advertising platforms generate large volumes of performance data, but advertisers struggle to understand which factors drive conversions and ROI. This project applies data mining techniques to analyze ad performance, identify meaningful patterns, segment ads, and build predictive insights to support better ad-spend decisions.

Description of Dataset and the link for the dataset:-

Dataset: Facebook Ads Dataset (Kaggle)

<https://www.kaggle.com/datasets/madislemsalu/facebook-ad-campaign>

Dataset Overview

Attribute Category	Description
Dataset Name	Facebook Ads Campaign Dataset
Source	Kaggle
Number of Records	100,000+
Number of Features	10+
Data Type	Structured, numeric & categorical
Domain	Digital Advertising & Marketing
Privacy	No personal or sensitive data

Feature Description

Feature Name	Description
Impressions	Number of times an ad was displayed
Clicks	Number of user clicks on the ad
CTR	Click-Through Rate
CPC	Cost Per Click
Spent	Total advertising spend

Conversions	Total conversions recorded
Total_Conversion	All conversion events
Approved_Conversion	Validated conversion events
Age	Age group of target audience
Gender	Gender of target audience
Interest	Interest category identifier

Data Preprocessing Details:-

Step	Technique Used	Purpose
Missing Values	Median Imputation	Handle skewed conversion data
Outlier Treatment	IQR-based Capping	Reduce effect of extreme values
Standardization	Z-score normalization	Ensure feature comparability
Categorical Encoding	One-Hot Encoding	Convert age & gender
Feature Removal	ID column removal	Eliminate non-predictive features

Handling Missing Values

Only the conversion-related variables (`total_conversion` and `approved_conversion`) had missing values. The median, a reliable metric appropriate for skewed distributions, was used to impute these missing observations. The dataset had no missing values after imputation.

Removing or Correcting Outliers

Box plots were used to identify and quantify outliers in numerical variables (`interest1`, `interest2`, `interest3`, `impressions`, `clicks`, `spent`, and conversion metrics) using the **Interquartile Range (IQR)** method. To lessen the impact of extreme values while maintaining overall data integrity, values outside of $Q1 - 1.5 \times IQR$ and $Q3 + 1.5 \times IQR$ were capped to the corresponding bounds.

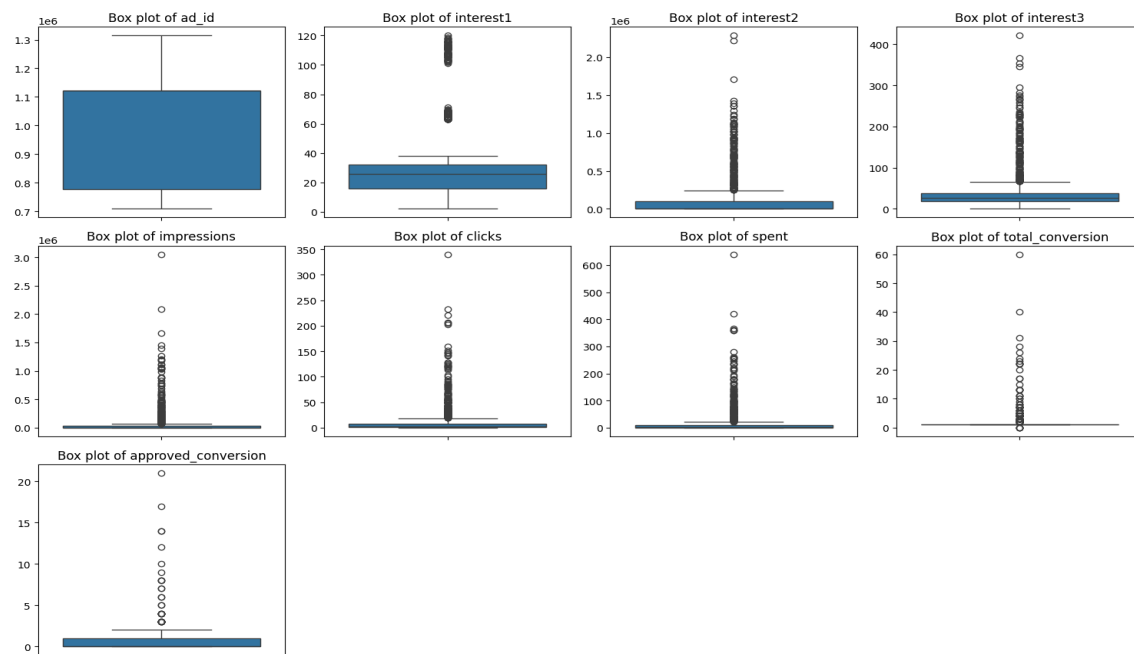
Normalization / Standardization

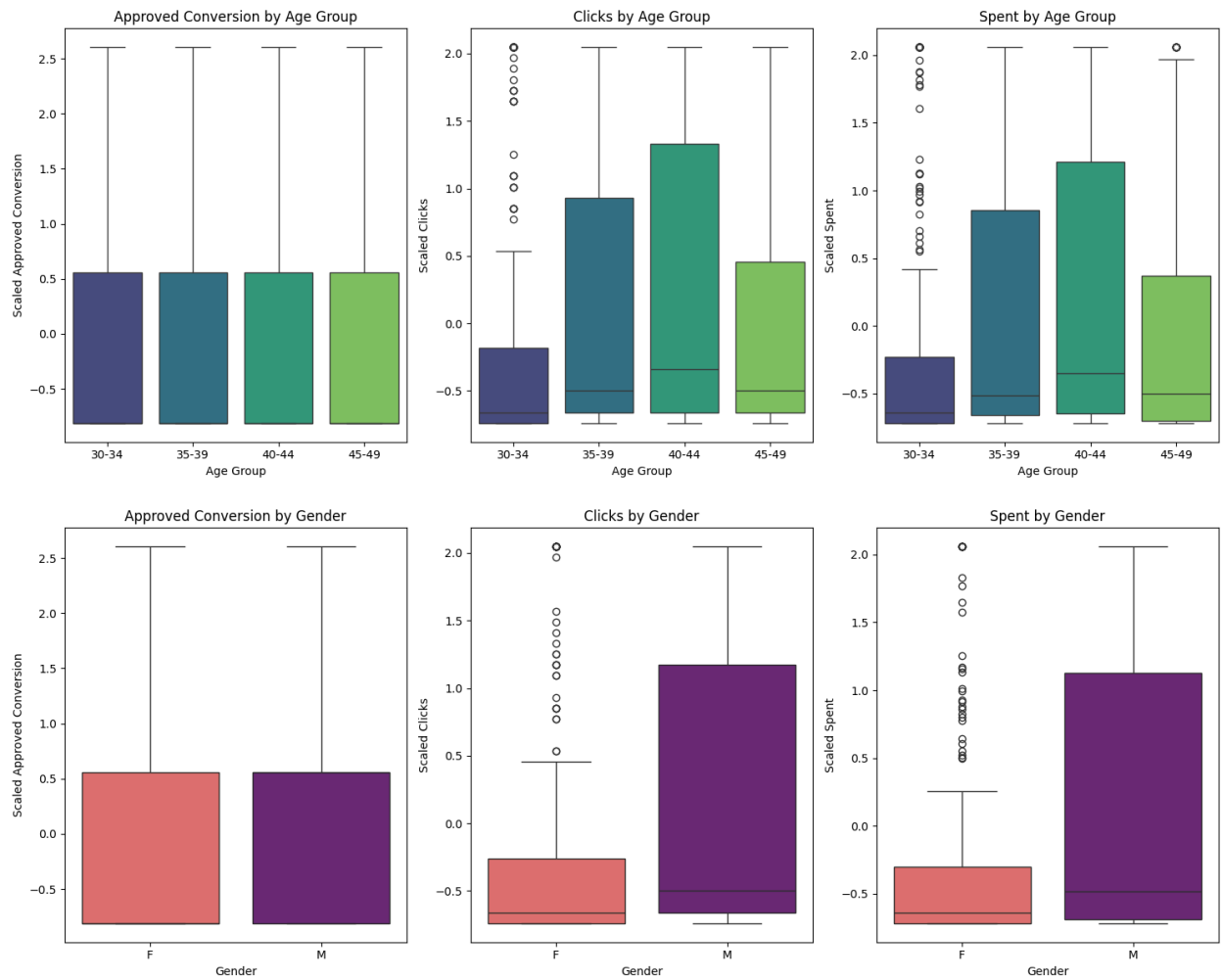
All numerical variables were standardized using z-score **z-score normalization** via `StandardScaler` to guarantee comparability across features and enhance model performance. This made the data appropriate for dimensionality reduction and

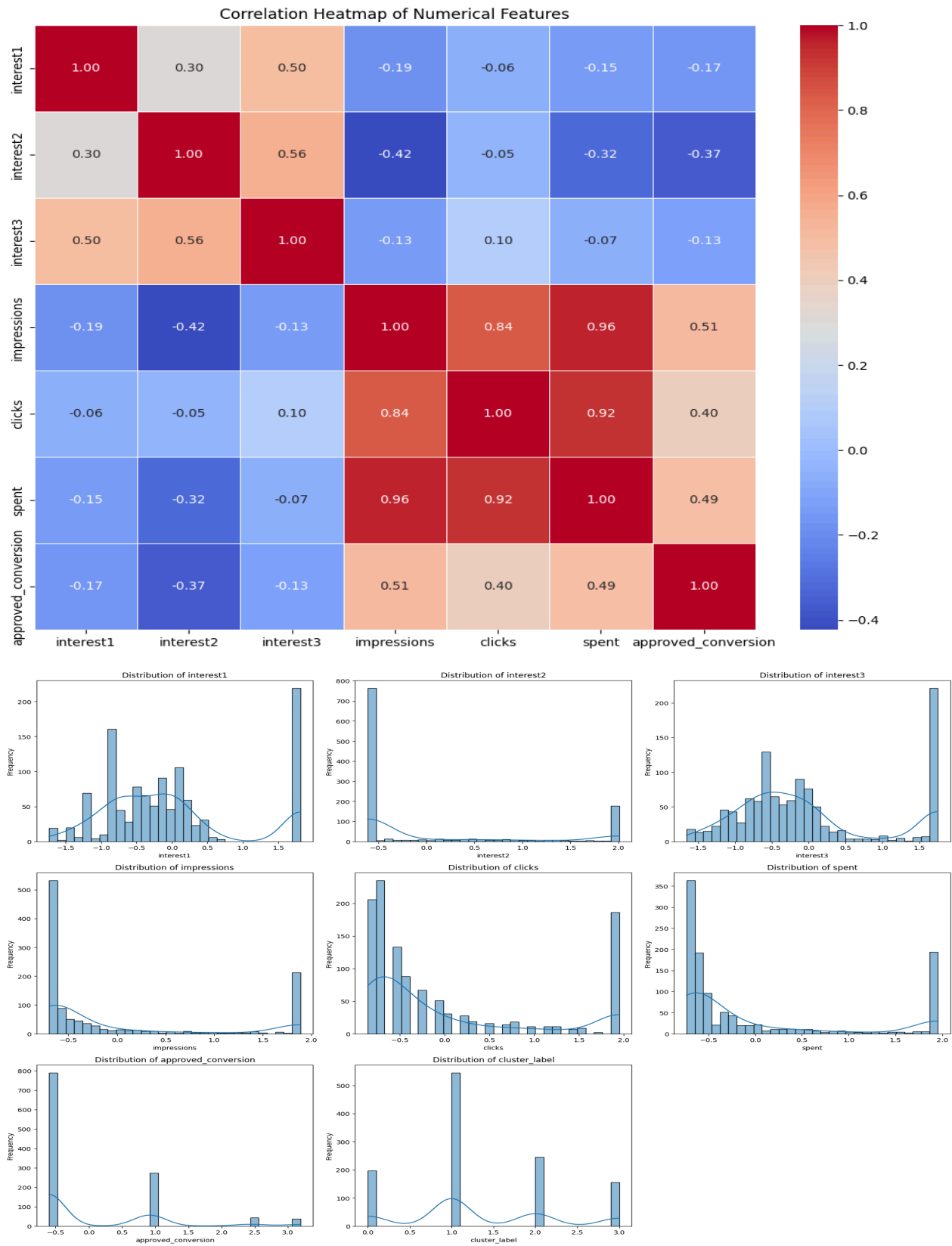
distance-based algorithms by transforming each feature to have a mean of zero and a standard deviation of one.

Feature Selection and Feature Engineering

Date variables were converted to datetime format, and a new feature representing campaign duration was engineered. Identifier columns (`ad_id`, `campaign_id`, `fb_campaign_id`) were removed as they do not provide predictive value. Categorical variables (`age` and `gender`) were transformed using **one-hot encoding** with `drop_first=True` to avoid multicollinearity, resulting in a structured feature set appropriate for modeling.







Exploratory Analysis Findings:-

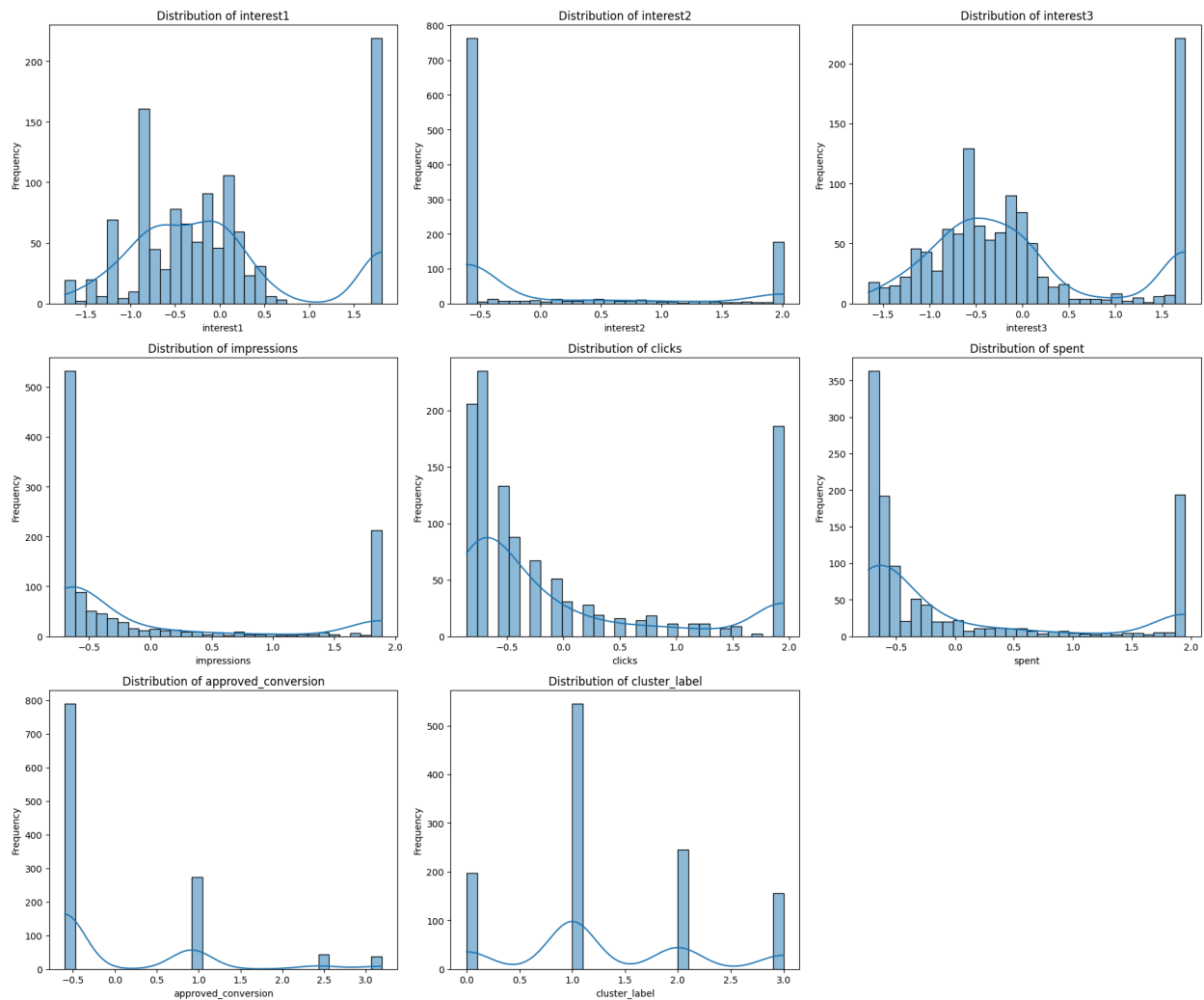
Exploratory data analysis revealed that key advertising performance metrics—**impressions, clicks, and spend**—exhibit strong right-skewed distributions, indicating that most advertisements generate low engagement while a small subset accounts for disproportionately high activity. This pattern persisted even after outlier capping and standardization, reflecting the inherent structure of real-world advertising data.

The **Interquartile Range (IQR)** method for outlier detection found extreme values in a variety of numerical features. By stabilizing the distributions and lowering variance, capping these values improved their suitability for clustering and predictive modeling. However, following imputation and outlier handling the **total_conversion** total_conversion feature became constant, highlighting a preprocessing-induced limitation and making it useless for further analysis.

Strong positive correlations between impressions, clicks, and spend were revealed by correlation analysis, suggesting high multicollinearity. These metrics showed moderately positive correlations with approved conversions, indicating that exposure and engagement are significant factors influencing conversion outcomes. Interest-related characteristics, on the other hand, showed weak linear correlations with performance metrics, indicating little direct impact on overall ad performance.

Analysis of categorical variables identified initial data quality issues in the age and gender fields, which were corrected by filtering invalid values. After cleaning, the **30–34 age group** and **male users** were the most represented. Engagement metrics such as clicks and spend varied across age groups and genders, with higher values observed among certain demographics, while **approved conversion rates remained relatively consistent**, indicating that demographic effects are stronger in early engagement stages than in final conversion outcomes.

Finally, unsupervised clustering revealed **distinct advertisement segments**, confirming the presence of meaningful structure within the data. These findings directly informed feature selection, preprocessing decisions, and the choice of modeling techniques applied in later stages of the analysis.



Aspect	Observation
Distribution Shape	Strong right skew in impressions, clicks, spend
Outliers	Present across multiple numeric features
Correlation	Strong correlation among impressions, clicks, spend
Demographics	30–34 age group & males most represented
Interest Features	Weak linear relationship with conversions
Clustering	Clear performance-based ad segments identified

Methods Used (Algorithms + justification)

This project applied a combination of **unsupervised and supervised data mining techniques** to analyze digital advertising performance, in alignment with the objectives of segmentation, pattern discovery, and prediction.

I. Clustering (K-Means)

Based on encoded demographic features and standardized performance metrics, different groups of ads were identified using K-Means clustering. This algorithm was chosen because it is effective and easy to understand when dealing with big numerical datasets. By grouping ads with similar engagement and spending patterns, clustering enabled the identification of meaningful performance segments such as low-engagement, high-spend, and relatively efficient ads. The optimal number of clusters was determined using the **Elbow Method** and validated with **Silhouette Score**, ensuring cluster quality and separation.

II. Dimensionality Reduction (Principal Component Analysis)

Principal Component Analysis (PCA) was applied to reduce the dimensionality of the feature space while preserving the majority of variance in the data. Given the high number of features resulting from one-hot encoding, PCA helped mitigate multicollinearity and improve computational efficiency. The resulting principal components also facilitated visualization of cluster separation and supported interpretability of underlying performance patterns.

III. Regression Modeling (Random Forest and Gradient Boosting)

Two ensemble-based regression models, **RandomForestRegressor** and **GradientBoostingRegressor**, were used to forecast authorized conversions. These models were selected due to their capacity to capture feature interactions and non-linear relationships that are frequently found in marketing data. Random Forest provided a strong baseline with robustness to noise and overfitting, while Gradient Boosting was selected to assess whether sequential error correction could improve predictive performance. Using two models enabled a comparative evaluation of ensemble approaches under identical preprocessing and data splits.

Results & Evaluation

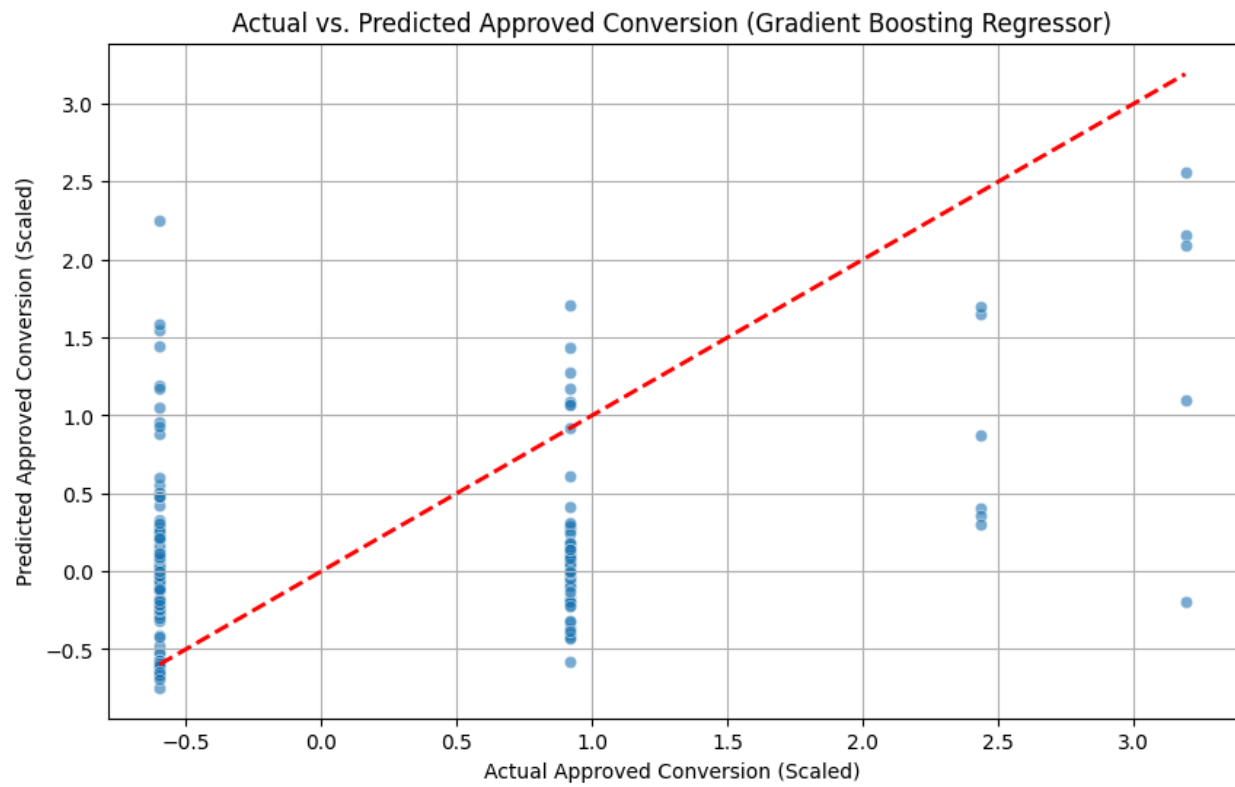
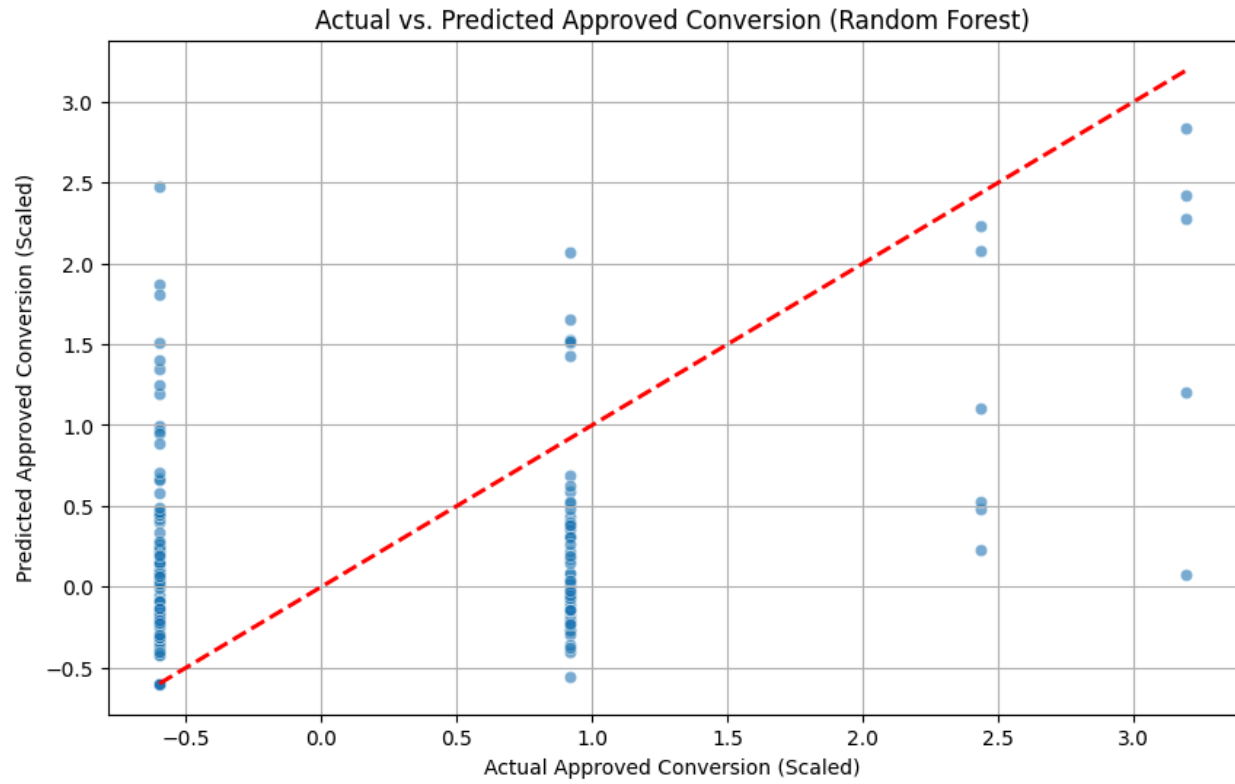
Model	MAE	MSE	R ²
Random Forest Regressor	0.6021	0.7118	0.1554
Gradient Boosting Regressor	0.5999	0.6829	0.1897

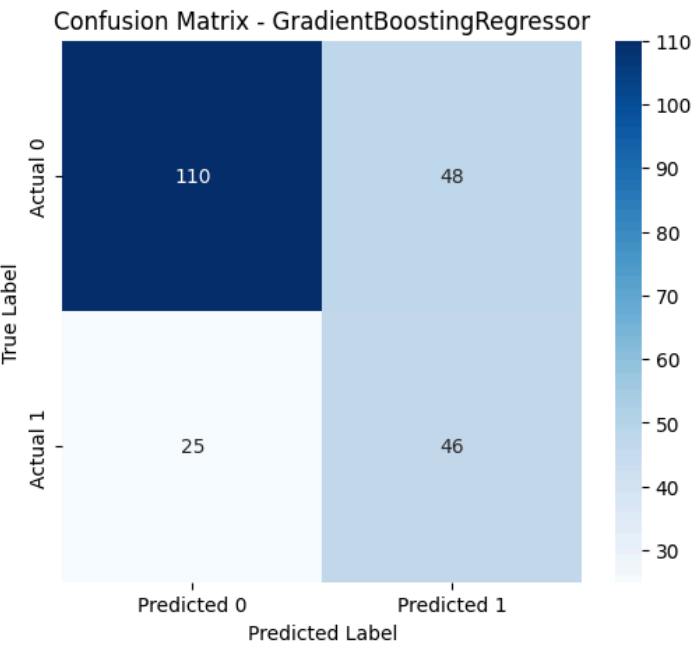
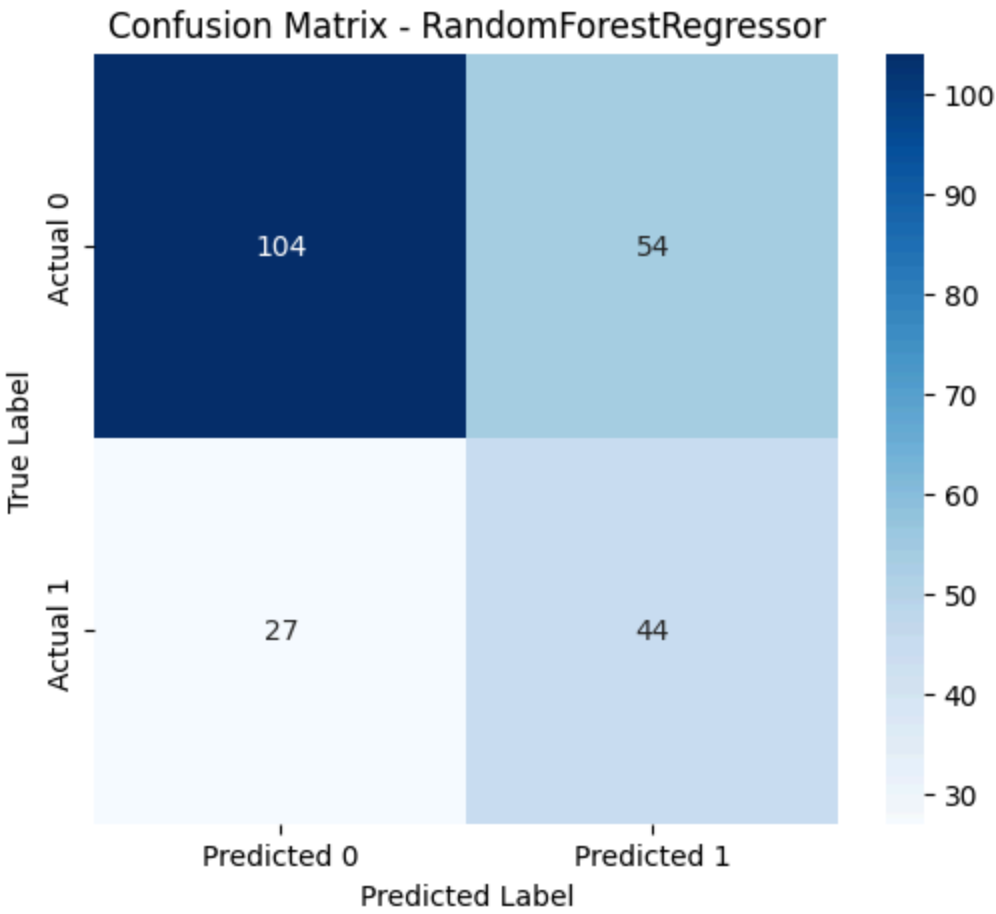
Two regression models—**RandomForestRegressor** and **GradientBoostingRegressor**—were trained to predict the scaled **approved_conversion** target variable using an 80/20 train–test split. Model performance was evaluated using **Mean Absolute Error (MAE)**, **Mean Squared Error (MSE)**, and **R-squared (R²)**, along with visual inspection of actual versus predicted values.

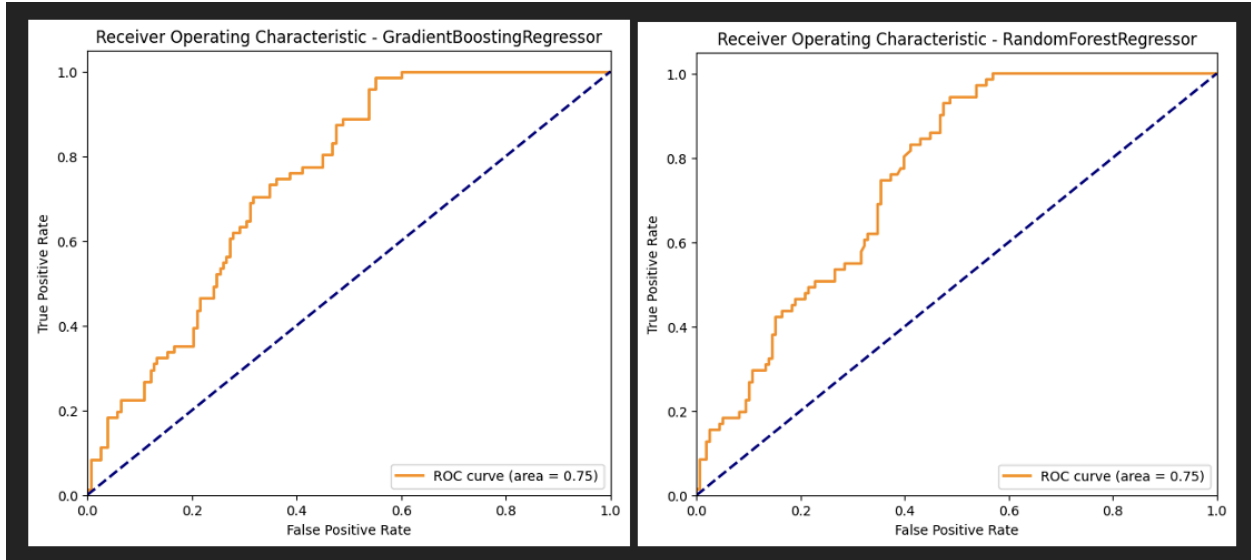
The RandomForestRegressor achieved an MAE of **0.6021**, an MSE of **0.7118**, and an R² of **0.1554**, indicating that approximately 15.5% of the variance in approved conversions was explained by the model. The GradientBoostingRegressor showed a marginal improvement, with an MAE of **0.5999**, an MSE of **0.6829**, and an R² of **0.1897**, explaining roughly 19% of the variance. While Gradient Boosting outperformed Random Forest across all metrics, the improvement was modest.

The actual versus predicted plots for both models revealed a consistent pattern: predicted values were heavily clustered around lower conversion levels, while higher actual conversion values were frequently underestimated. This indicates limited ability of both models to capture the full variability of the target variable, particularly for higher conversion outcomes.

Regardless of model selection, the overall findings imply that predictive performance is constrained. The low R² values show that a significant amount of the variance in authorized conversions is still unaccounted for. This restriction is probably caused by the target variable's highly skewed distribution, the predominance of low conversion values, and possible information loss from preprocessing (such as outlier capping). Although ensemble approaches offered respectable baseline performance, more advancements are anticipated to necessitate improved feature engineering, different modeling approaches, or problem reformulation (e.g., classification instead of regression).







Discussion of Insights

Several significant insights into ad performance, user engagement, and the limitations of predictive modeling in this context were uncovered by the analysis of digital advertising data. A small number of ads account for a disproportionate share of impressions, clicks, and conversions, according to exploratory analysis, which revealed highly skewed advertising outcomes. This concentration, which reflects typical real-world marketing dynamics, indicates that the majority of campaigns have little effect while only a small percentage produce significant results.

Clustering analysis revealed discrete ad segments with various performance profiles, emphasizing the existence of both more economical engagement-oriented ads and ineffective high-spend ads. According to these segments, advertisers may find it advantageous to shift their budgets from clusters that consistently perform poorly to those that show greater engagement efficiency.

Several significant insights into ad performance, user engagement, and the limitations of predictive modeling in this context were uncovered by the analysis of digital advertising data. A small number of ads account for a disproportionate share of impressions, clicks, and conversions, according to exploratory analysis, which revealed highly skewed advertising outcomes. This concentration, which reflects typical real-world marketing dynamics, indicates that the majority of campaigns have little effect while only a small percentage produce significant results.

Clustering analysis revealed discrete ad segments with various performance profiles, emphasizing the existence of both more economical engagement-oriented ads and ineffective high-spend ads. According to these segments, advertisers may find it advantageous to shift their budgets from clusters that consistently perform poorly to those that show greater engagement efficiency.

Overall, these insights emphasize that advertising performance is driven by complex, multi-dimensional factors. While quantitative metrics and demographic attributes provide valuable signals, they are insufficient on their own to fully explain conversion behavior. Effective advertising optimization therefore requires combining performance data with richer contextual and behavioral information.

Limitations & Future Work

Although this study offers valuable insights into the effectiveness of digital advertising, a number of limitations should be noted. First, ad creatives, messaging, placement, and timing are not included in the dataset; instead, it is restricted to aggregated campaign-level metrics. These variables may account for the regression models' poor predictive power since they are known to have a major impact on user behavior and conversion rates.

Second, there was a significant concentration of low or zero values in the highly skewed distribution of the target variable, approved conversions. Preprocessing techniques like outlier capping and standardization may have decreased significant variance, further restricting model performance even though they were required for model stability. This demonstrates how difficult it is to use regression-based methods on sparse and unbalanced conversion data.

Third, a static snapshot of past advertising data was used for the analysis. It was difficult to model changes in performance over time because temporal dynamics like seasonality, learning effects, and budget pacing were not recorded. Furthermore, only age and gender were included in the demographic variables, giving only a partial picture of the characteristics of the audience.

These limitations could be addressed in a number of ways in future work. Predictive performance would probably be enhanced by adding richer feature sets, such as creative attributes, platform signals, and time-based variables. The structure of conversion data might be better captured by alternative modeling techniques like time-series analysis, zero-inflated regression, or classification models. Lastly, expanding the analysis to incorporate frameworks for budget optimization or causal inference may improve the findings' practical applicability and give decision-makers more useful information.

Conclusion

This project used a real-world dataset to analyze the performance of digital advertising using a full data mining workflow. Important patterns and performance segments were found using data preprocessing, exploratory analysis, clustering, dimensionality reduction, and regression modeling. The results showed that advertising outcomes are highly skewed, with a small number of campaigns accounting for most engagement and conversions. Unsupervised methods revealed distinct advertisement segments, while supervised models demonstrated limited predictive power for approved conversions. These findings indicate that conversion behavior is influenced by factors not fully captured in standard performance and demographic features. Overall, the project highlights both the value and limitations of data mining techniques for supporting data-driven decision-making in digital marketing.

References & Tools Used

References

1. Kaggle. *Facebook Ads Campaign Dataset*.
<https://www.kaggle.com/datasets/madislemsalu/facebook-ad-campaign>
2. Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann.
3. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning* (2nd ed.). Springer.
4. scikit-learn Developers. *scikit-learn: Machine Learning in Python*.
<https://scikit-learn.org/stable/>
5. McKinney, W. (2010). *Data Structures for Statistical Computing in Python*. Proceedings of the 9th Python in Science Conference.

Tools Used

- **Python** – Primary programming language used for data preprocessing, analysis, and modeling
- **Jupyter Notebook** – Interactive environment for code development and reproducibility
- **Pandas** – Data manipulation, cleaning, and feature engineering
- **NumPy** – Numerical computing and array operations
- **Matplotlib & Seaborn** – Data visualization for EDA and result interpretation
- **scikit-learn** – Implementation of machine learning algorithms (K-Means, PCA, Random Forest, Gradient Boosting), preprocessing, and evaluation metrics