ROI prediction in Social Media Advertising Campaigns

Rishav Mondal

Dhruv Pancholi

Shubham Gosavi

Professor: **Hamidreza Ahady Dolatsara**

Introduction

In today’s digital era, social media platforms have become a cornerstone for businesses to connect with their target audience and promote products or services. However, ensuring a high Return on Investment (ROI) from social media advertising campaigns requires a deep understanding of campaign performance metrics and the factors influencing them. This project focuses on leveraging advanced data analytics and machine learning techniques to predict ROI for social media advertising campaigns, offering actionable insights to optimize marketing strategies.

The project utilizes a rich dataset containing details of advertising campaigns, including metrics such as conversion rate, acquisition cost, engagement score, and channel usage. By analyzing and modeling this data, we aim to uncover the key factors driving ROI and provide recommendations to improve the effectiveness of advertising efforts. Through the application of state-of-the-art regression models, including Random Forest Regressor, LSTM, and Neural Networks, this project highlights the power of data-driven decision-making in marketing.

Ultimately, the goal of this project is not only to predict ROI but also to offer a framework that businesses can use to make informed, strategic decisions about their social media investments. By focusing on insights derived from real-world data, the study provides valuable guidance for maximizing the impact of advertising campaigns while minimizing costs.

Dataset Description

The dataset comprises 300,000 records detailing advertising campaigns across various platforms. Key attributes include:

- **Campaign Information:** Campaign ID, Target Audience, Campaign Goal, Duration, Channel Used.
- **Performance Metrics:** Conversion Rate, Acquisition Cost, ROI (target variable).
- **Engagement Details**: Clicks, Impressions, Engagement Score.
- **Other Attributes:** Location, Language, Customer Segment, Company.

Preprocessing Steps:

1. **Cleaning:** Removed dollar signs from `Acquisition_Cost` and converted it to numeric.
2. **Encoding:** Categorical variables like `Channel_Used`, `Location`, and `Target_Audience` were encoded for machine learning models.
3. **Normalization:** Features such as `Clicks` and `Impressions` were normalized to ensure consistent scaling.

Exploratory Data Analysis (EDA)

Key Insights:

1. **ROI Statistics:** Average ROI by Channel: Pinterest showed the highest average ROI, while Facebook was the least effective.
2. **Top Campaigns:** The highest ROI was observed for campaigns targeting the 'Technology' segment.
3. **Engagement and Conversions:** Conversion Rate is strongly correlated with ROI. Impressions and Clicks provide moderate predictive value.
4. **Visualizations:** Various bar plots, scatter plots, and heatmaps were created to analyze campaign performance.

Visualizations and Their Significance

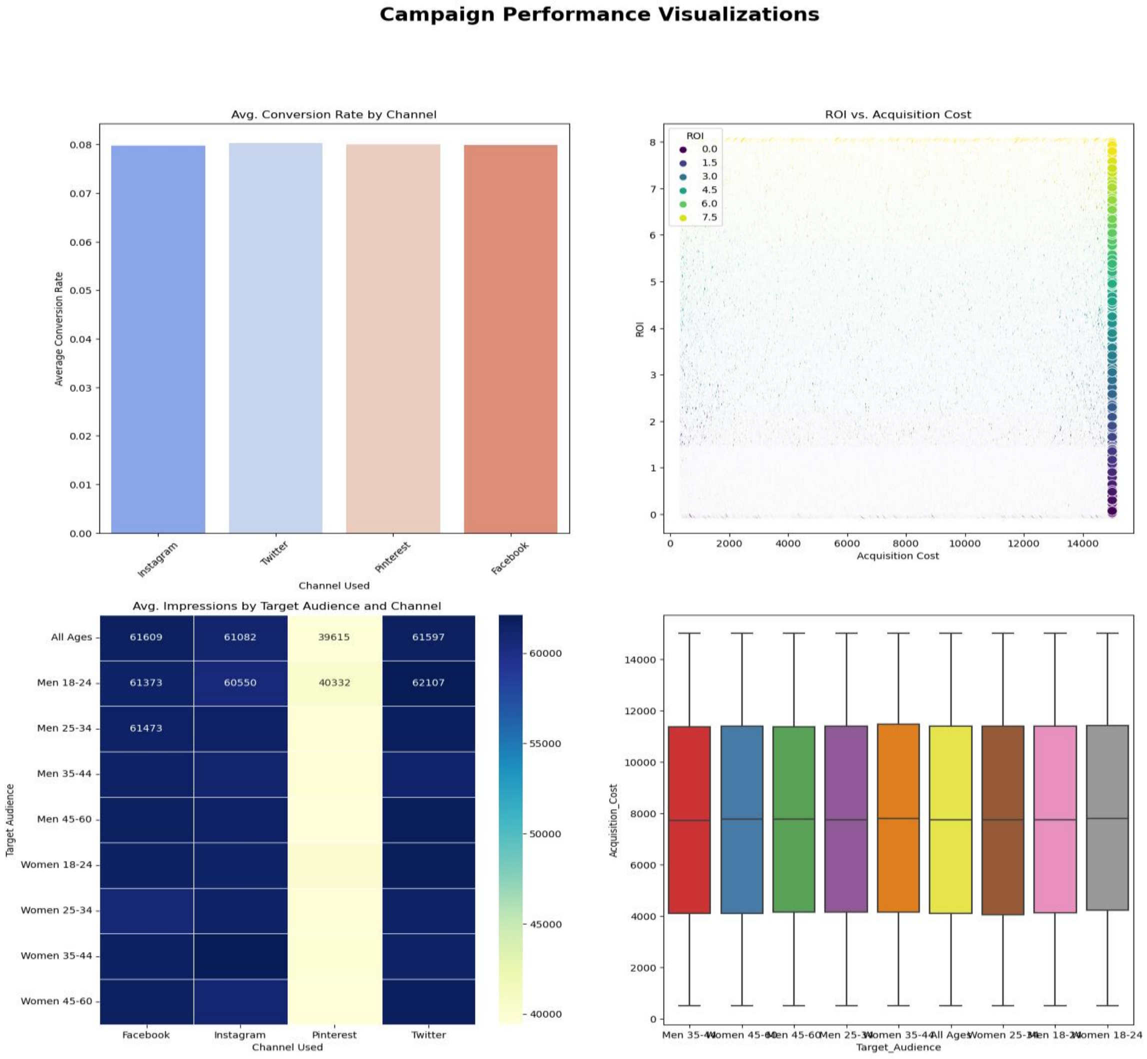

Fig 1: Campaign Performance Visualization

The above plot shows the performance of the various campaigns run by the company, in the various channels and the ROI and conversion rate that these campaigns produced. We can clearly see that the **Twitter** performed slightly better than the other channels.

The **ROI** vs **Acquisition Cost,** is quite dense in the region of 8 and between the are of 2-3, which indicates that most of the ROI is either 8 times the acquisition cost or between 2 to 3 time is the acquisition cost.

We can see from the plots that the average impressions with respect to channels shows that **Pinterest** generates the least impressions while **Facebook** generates the highest.

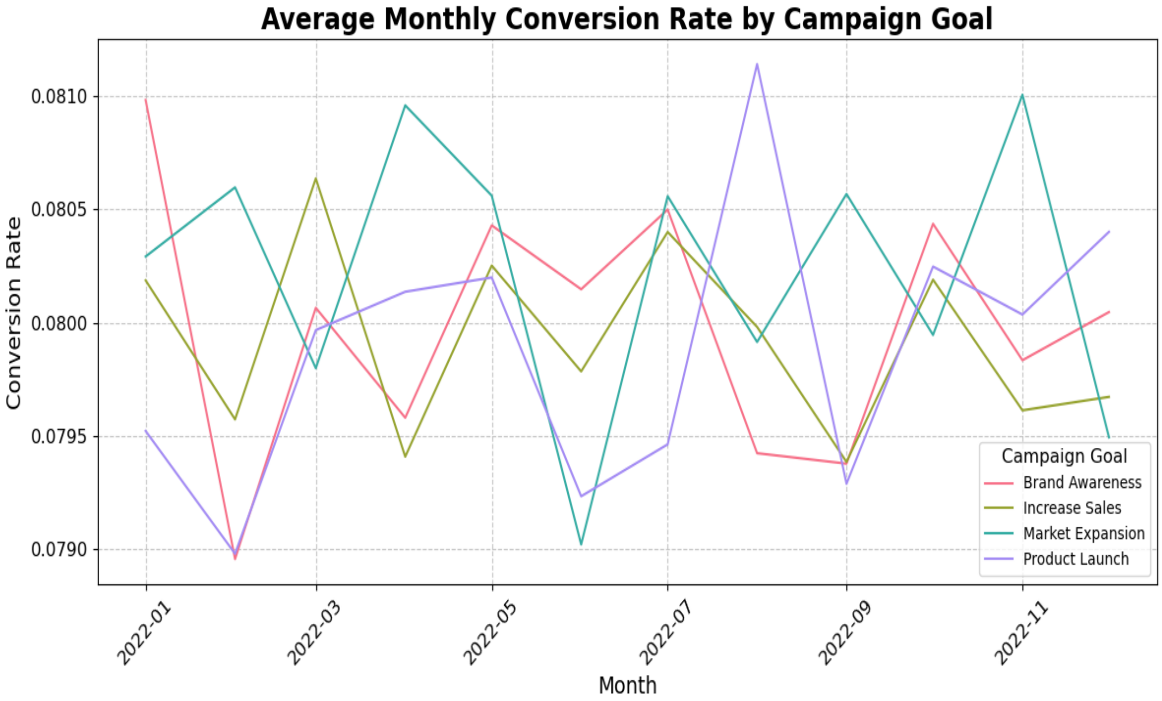

Fig. 2: Average Monthly Conversion Rate by Campaign Goal

We can clearly see there are exactly 4 campaigns that are running, the ‘Brand Awareness’, ‘Increase Sales’, ‘Market Expansion’ and ‘Product Launch’ campaigns. The **product launch** campaign has the highest conversion rate in the months of June and July in 2022, and the least in the months of January and February, meaning the product is something the people like the best during summer. The campaign of **Increase Sales** delivers at an average rate all throughout the year.

The plot below tells us that the marketing campaign does not always return favorable results. As in this case most of the campaigns yield a ROI between 0 - 1.5 which isn’t a significant ROI.

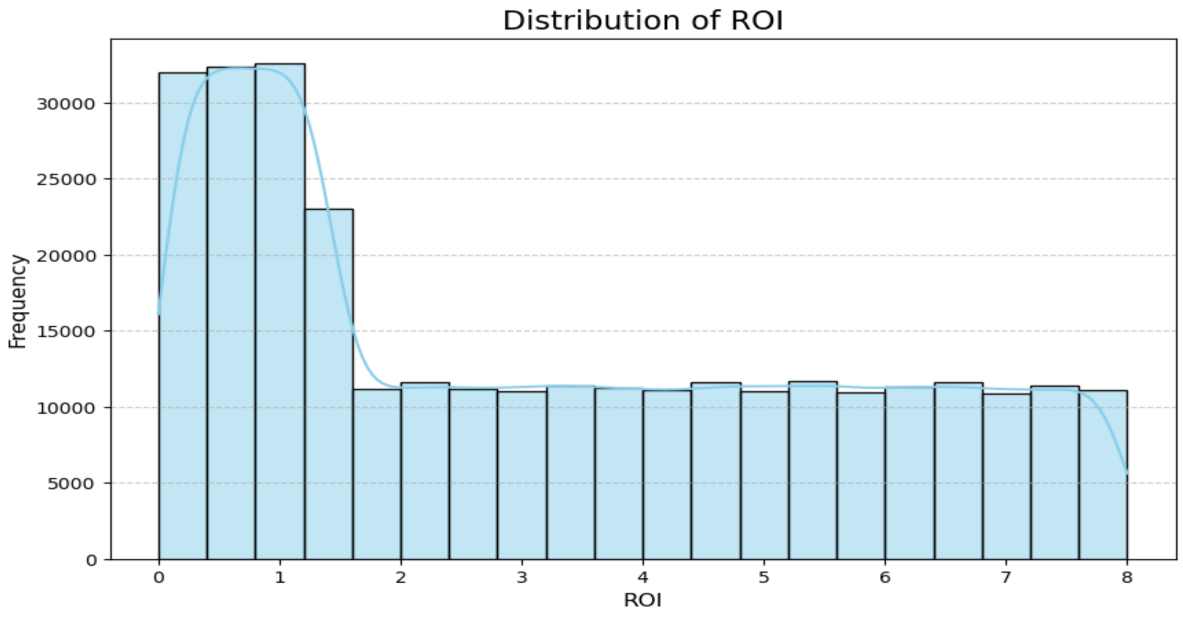

Figure 2: Distribution of ROI

The ROI also differs with respect to the channel used. Although **Instagram**, **Twitter**, **Facebook** have similar average ROI, **Pinterest** has significantly less average ROI, meaning that people use that platform significantly less compared to the others, and the company can do without using that platform for advertising.

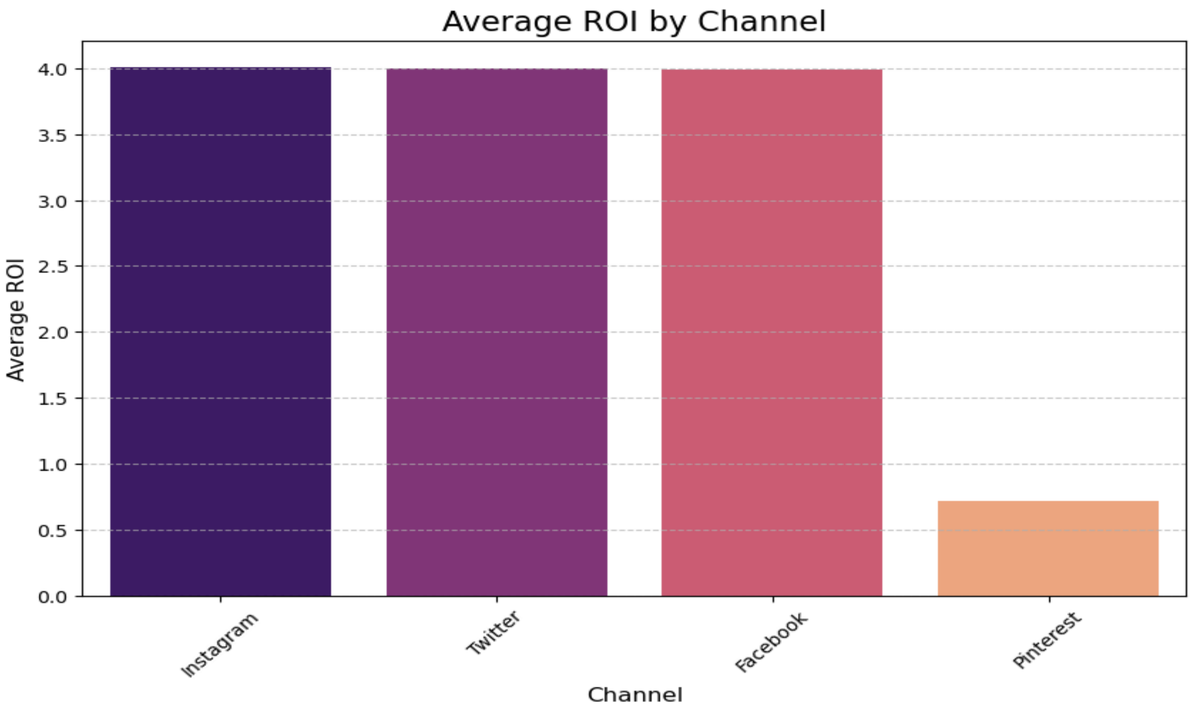

Fig. 4: Average ROI by Channel

Software and Tools

The project leveraged the following tools and technologies:

- **PySpark** for large-scale data processing and machine learning pipelines.
- **Pandas** and **NumPy** for data manipulation.
- **Seaborn** and **Matplotlib** for visualization.
- **TensorFlow/Keras** for LSTM implementation.

Predictive Modeling

Multiple machine learning models were implemented to predict ROI. These include:

- **Random Forest Regressor**:
  - A robust ensemble learning method capable of handling non-linear relationships and feature importance.
  - Implemented using PySpark's machine learning library.
- **LSTM (Long Short-Term Memory)**:
  - A deep learning model designed to capture sequential dependencies in time-series data.
  - Useful for identifying temporal patterns in advertising performance.
- **Linear Regression**:
  - A classic regression model that assumes a linear relationship between features and the target variable.
  - Implemented using PySpark, providing an interpretable baseline for model performance.

Results and Evaluation

The performance of each model was evaluated using the following metrics:

- **Root Mean Squared Error (RMSE)**: Measures the average magnitude of prediction errors.
- **Mean Absolute Error (MAE)**: Quantifies the average absolute differences between predicted and actual ROI.
- **R^2 (Coefficient of Determination)**: Indicates the proportion of variance in ROI explained by the models.

Model	**RMSE**	**MAE**	**R^2 Score**
Random Forest	2.00	1.59	0.33
LSTM	2.10	1.45	N/A
Linear Regression	2.17	1.82	0.22

Table 1: Model Performances

According to Table 1, we can see the **Random Forest** has the least error at 2.0, which in this case means that the predictions are correct **75%** of the times, that is the accuracy is 75%.

Conclusions and Recommendations

Key Insights:

- Channels and campaign goals significantly impact ROI.
- Cost-effective targeting strategies are critical for better ROI.

Recommendations:

- Focus on high-performing channels like Pinterest and Instagram.
- Use Random Forest for ROI predictions due to its robustness.
- Investigate deeper sequential patterns to justify LSTM usage further.

References

1. **Libraries and Frameworks**:
   - Apache Spark Project. "PySpark Documentation." https://spark.apache.org/docs/latest/api/python/
   - Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). "Scikit-learn: Machine Learning in Python." Journal of Machine Learning Research, 12, 2825–2830. https://scikit-learn.org/
   - Chollet, F., et al. (2015). "Keras: The Python Deep Learning Library." https://keras.io/
2. **Visualization Tools**:
   - Hunter, J. D. (2007). "Matplotlib: A 2D Graphics Environment." Computing in Science & Engineering, 9(3), 90-95. https://matplotlib.org/
   - Waskom, M. L. (2021). "Seaborn: Statistical Data Visualization." Journal of Open Source Software, 6(60), 3021. https://seaborn.pydata.org/
3. **Methodological References**:
   - Breiman, L. (2001). "Random Forests." Machine Learning, 45(1), 5-32.
   - Hochreiter, S., & Schmidhuber, J. (1997). "Long Short-Term Memory." Neural Computation, 9(8), 1735-1780.
4. **Dataset Source**:
   - The dataset used in this project was obtained from synthetic/generated data reflecting social media advertising campaigns for academic purposes.
5. **General References**:
   - Géron, A. (2019). "Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow." O'Reilly Media, Inc.
   - Goodfellow, I., Bengio, Y., & Courville, A. (2016). "Deep Learning." MIT Press.

.