# **Digital Marketing Conversion Prediction Analysis**

Mario Paulin Vega

## Executive Summary

This project developed and implemented an XGBoost machine learning model to predict customer conversion probabilities for digital marketing campaigns. Using a dataset of 10,000 customer records, the analysis provides actionable insights for campaign optimization and cost reduction.

## Data Description

The data is publicly available at Kaggle https://www.kaggle.com/datasets/rabieelkharoua/predict-conversion-in-digital-marketing-dataset/data

Dataset Specifications:

* Total Sample Size: 10,000 customer records
* Data Split: 70/30 train/test split (7,000 training, 3,000 testing)
* Time Period: Cross Sectional Marketing campaign data.
* Target Variable: Binary conversion (0 = no conversion, 1 = conversion)

Main Variables:

* Customer Data: Age, Income, Gender, Loyalty Points
* Campaign Attributes: Campaign Type (Brand Awareness, Conversion, Retention), Campaign Channel (Email, PPC, SEO, Social Media, Referral)
* Financial Metrics: Ad Spend
* Behavioral Data: Return Customer (New or returning customer)
* Engineered Features: 31 total features including interaction terms and one-hot encoded categoricals

Challenge: Unbalanced target (12% non-convertion)

## Business Questions and Objectives

The analysis addressed three marketing questions:

1. Do campaigns targeting conversion do reflect on higher conversion rates and probabilities?
2. What are the most efficient and cost effective channels for conversion campaings?
3. Based on probability of conversion, are the campaigns targeted correctly, or can we improve the expected value by removing some low conversion customers from the target population?

## Methodology

Model Development:

* Primary Algorithm: XGBoost (Extreme Gradient Boosting)
* Feature Engineering: 31 features including interaction terms and categorical encodings
* Hyperparameter Tuning: Grid search optimization
* Validation: Stratified train/test split with cross-validation

Comparison Models:

* Baseline Logistic Regression
* Un-tuned XGBoost

### Key Business Insights

1. Campaign Effectiveness Analysis:

* While average campaign spending by Type is similar and around $5,000, Conversion campaigns do result in an increased probability of conversion of around 12% as per the model.
* Most Effective channel: As per the model PPC campaigns with 66.2% predicted conversion probability, an increase of around 2% over the average channel.
* Most Cost-Efficient: SEO campaigns at $4,832 cost per conversion vs average of $5,325

2. Cost-Effectiveness:

* The model simulations indicate that, considering a customer conversion value of $4,000, we cannot improve the cost benefit of the campaigns by dropping low conversion probability customers from the target population.
* Model indicates a cost of $441 average cost per 1% probability increase in conversion.

## Conclusion

The XGBoost model identifies customers with varying conversion probabilities, enabling data-driven budget allocation decisions.

Next Steps:

1. Additional tuning of model hyperparamenters and re-balancing strategies
2. Gathering additional features and feature engineering
3. Understanding timeline of the data. Some of the features were excluded from the model as it is not clear if they are obtained as intermediate outcomes of the campaigns or if the data could be obtained earlier.