Optimizing Offer Completion Using Machine Learning

A Data-Driven Approach to Marketing Optimization

Coskun Erden

11/27/2024

Executive Summary

This report details the findings and results of the Starbucks Capstone Challenge, a machine learning project aimed at predicting offer completion rates and optimizing marketing strategies. The project leverages customer demographic and behavioral data along with offer-specific attributes to identify patterns and actionable insights for improving promotional offer performance.

Project Overview

The Starbucks Capstone Challenge simulates how people make purchasing decisions and how these decisions are influenced by promotional offers. The dataset includes three key components:

1. Profile Data (profile.json): Contains information about 17,000 users, including demographics such as gender, age, income, and membership start date.
2. Portfolio Data (portfolio.json): Includes details on 10 promotional offers, categorized as buy-one-get-one (BOGO), discount, or informational. Each offer specifies a reward, difficulty, duration, and delivery channels (web, email, mobile, social).
3. Transcript Data (transcript.json): Captures 306,648 event logs representing interactions such as offers received, offers viewed, offers completed, and transactions, along with corresponding timestamps, user IDs, and monetary values.

The basic task was to identify which groups of people are most responsive to each type of offer and to determine how best to present these offers for maximum engagement and effectiveness.

Key Findings

1. Gradient Boosting Performance:
    - Gradient Boosting emerged as the best-performing model among five algorithms tested (Logistic Regression, Random Forest, Support Vector Machine, XGBoost, and Gradient Boosting).
    - It achieved the highest overall accuracy of 91% and a well-rounded F1-score of 0.61 for the minority class, indicating a strong ability to balance class imbalances while maintaining predictive performance.

2. Feature Importance Analysis:
   - Behavioral features like time (when offers are presented) and duration (how long offers remain valid) had the greatest impact on offer completion rates.
   - Simpler offers with moderate difficulty and reasonable rewards were more likely to succeed.
   - Demographic features such as age and gender played a minimal role in predicting offer completions.
3. Offer Type Responsiveness:
   - BOGO Offers: These performed best when targeted at high-income groups with longer membership durations, as they require higher spending thresholds.
   - Discount Offers: Showed effectiveness across broader user groups, offering flexible engagement with moderate rewards.
   - Informational Offers: Benefited from strategic timing and multi-channel communication to increase visibility and awareness, despite not offering monetary rewards.

Actionable Strategies

Based on the analysis, several actionable recommendations were derived to optimize promotional campaigns:

1. Optimize Offer Timing: Prioritize sending offers during peak activity periods to maximize engagement. Timing is critical and outweighs static demographic features in influencing behavior.
2. Simplify Offers: Reduce complexity by lowering the difficulty levels required to claim rewards, while maintaining reasonable durations for better completion rates.
3. Enhance Rewards: While rewards are not the most significant factor, improving their attractiveness can still positively impact completion rates.
4. Leverage Behavioral Targeting: Shift focus from demographic targeting to behavioral patterns, such as activity and engagement levels.
5. Channel Optimization: Utilize multiple communication channels (e.g., web, mobile, email) to maximize the visibility of offers.
6. Tailor Offers for High-Value Customers: Design customized offers for customers with high incomes or longer membership durations, as they are more likely to complete offers.
7. Refine Campaigns Based on Offer Types: Use insights to craft tailored promotions for each type of offer, ensuring they resonate with specific customer preferences and behaviors.

Conclusion

The Starbucks Capstone Challenge demonstrates the value of integrating machine learning into marketing strategies to derive actionable insights and enhance customer engagement. By leveraging Gradient Boosting and focusing on key behavioral features, Starbucks can design data-driven promotional strategies to optimize offer completion rates, improve customer satisfaction, and maximize ROI. Future work may include exploring advanced techniques like time-series analysis and improved sampling methods to address remaining challenges and refine predictions further.

2. Introduction

Promotional offers are an integral part of marketing campaigns, driving customer engagement and influencing purchasing behavior. However, despite their significance, businesses often face challenges in designing and delivering offers that achieve high completion rates. By leveraging data-driven insights and machine learning models, organizations can better understand customer behavior and optimize promotional strategies to maximize returns on investment (ROI). This report focuses on addressing these challenges using a dataset provided by the Starbucks Capstone Challenge.

2.1 Problem Statement

Offer completion rates play a pivotal role in the success of marketing campaigns. Accurately predicting which customers are most likely to complete offers is critical for optimizing marketing strategies. However, achieving this goal is challenging due to inherent issues in the data, including:

Class Imbalance: Offer completions represent a small subset of the overall dataset, making it difficult for machine learning models to learn patterns effectively.

Complexity of Behavioral Data: Customer responses to offers are influenced by a combination of demographic traits, behavioral patterns, and offer attributes, making it essential to identify the right combination of features to predict outcomes.

Diversity of Offer Types: The presence of multiple offer types (e.g., BOGO, discount, informational) and delivery channels adds another layer of complexity to understanding customer preferences and designing targeted strategies.

2.2 Objectives

To address these challenges, this project focuses on the following key objectives:

Predict Offer Completion: Develop and evaluate machine learning models to accurately predict offer completions, balancing predictive accuracy and recall, particularly for the minority class (completed offers).

Identify Key Influencing Factors: Analyze the relative importance of features such as timing, duration, reward, and demographic attributes in influencing offer completion rates.

Propose Actionable Strategies: Provide data-driven recommendations to enhance promotional campaigns, targeting specific customer segments with optimized offers.

2.3 Business Relevance

The ability to improve offer completion rates has direct implications for businesses, particularly in terms of customer engagement and revenue generation. By addressing the challenges outlined above, organizations can achieve the following:

Enhanced Customer Engagement: Personalized and well-timed offers increase customer satisfaction and loyalty, driving repeat interactions.

Optimized Marketing Spend: Accurate predictions allow businesses to allocate resources more effectively, ensuring marketing budgets are spent on strategies with the highest ROI.

Improved Campaign Effectiveness: Understanding which groups of customers respond best to specific offers enables businesses to refine their promotional strategies and maximize impact.

Data-Driven Decision-Making: Leveraging machine learning and advanced analytics ensures that marketing strategies are guided by insights derived from real-world data, reducing reliance on intuition or trial-and-error approaches.

By integrating predictive modeling with actionable insights, this project aims to demonstrate the potential of machine learning in transforming traditional marketing campaigns into data-driven, highly targeted initiatives that deliver measurable results.

3. Data Overview

The analysis relies on three key datasets from the Starbucks Capstone Challenge, each providing crucial information on customers, promotional offers, and interactions. This section outlines the dataset structures, preprocessing steps undertaken to prepare the data, and exploratory data analysis to derive initial insights.

3.1 Datasets Used

Portfolio Dataset:

id: Unique identifier for the offer.

offer_type: Type of the offer: BOGO, discount, or informational.

reward: Monetary reward received upon completing the offer.

difficulty: Spending threshold required to qualify for the reward.

duration: Offer availability period (in days).

channels: Channels through which the offer was delivered (web, email, mobile, social).

Profile Dataset:

id: Unique identifier for the user.

gender: User's gender (M, F, O, or null).

age: Age of the user; missing values encoded as 118.

income: User's annual income in USD.

became_member_on: Date when the user joined the rewards program.

Transcript Dataset:

person: Unique identifier for the user.

event: Type of event (offer received, viewed, completed, or transaction).

value: Dictionary containing details such as the offer ID, amount spent, or reward earned.

time: Time (in hours) since the start of the test period.

3.2 Preprocessing Steps

To ensure the data is suitable for analysis and modeling, the following preprocessing steps were performed:

Handling Missing Values:

Gender: Missing values were replaced with "Unknown."

Age: Imputed missing values (encoded as 118) with the median age of the dataset.

Income: Missing income values were left unchanged to analyze their impact.

Reward Earned: Imputed as 0 for events where no reward was recorded.

Feature Engineering:

Binary Columns for Channels: Created separate columns (web, email, mobile, social) with binary values (1/0) to indicate the presence of each channel for an offer.

Offer Completed: Added a binary column to indicate whether an offer was successfully completed (1) or not (0).

Data Transformation:

Age Categories: Grouped ages into categories such as 'young adult,' 'mid-career,' 'senior,' etc., for better interpretability.

Offer Types: Encoded as categorical variables for use in machine learning models.

Dataset Consolidation:

Merged datasets using user and offer IDs to create a unified structure for analysis and modeling.

Retained relevant records such as offer-related events (received, viewed, completed) and excluded redundant transactions.

3.3 Exploratory Data Analysis

Initial exploration was conducted to uncover patterns and relationships in the dataset. Key analyses include:

Heatmap of Feature Correlations:

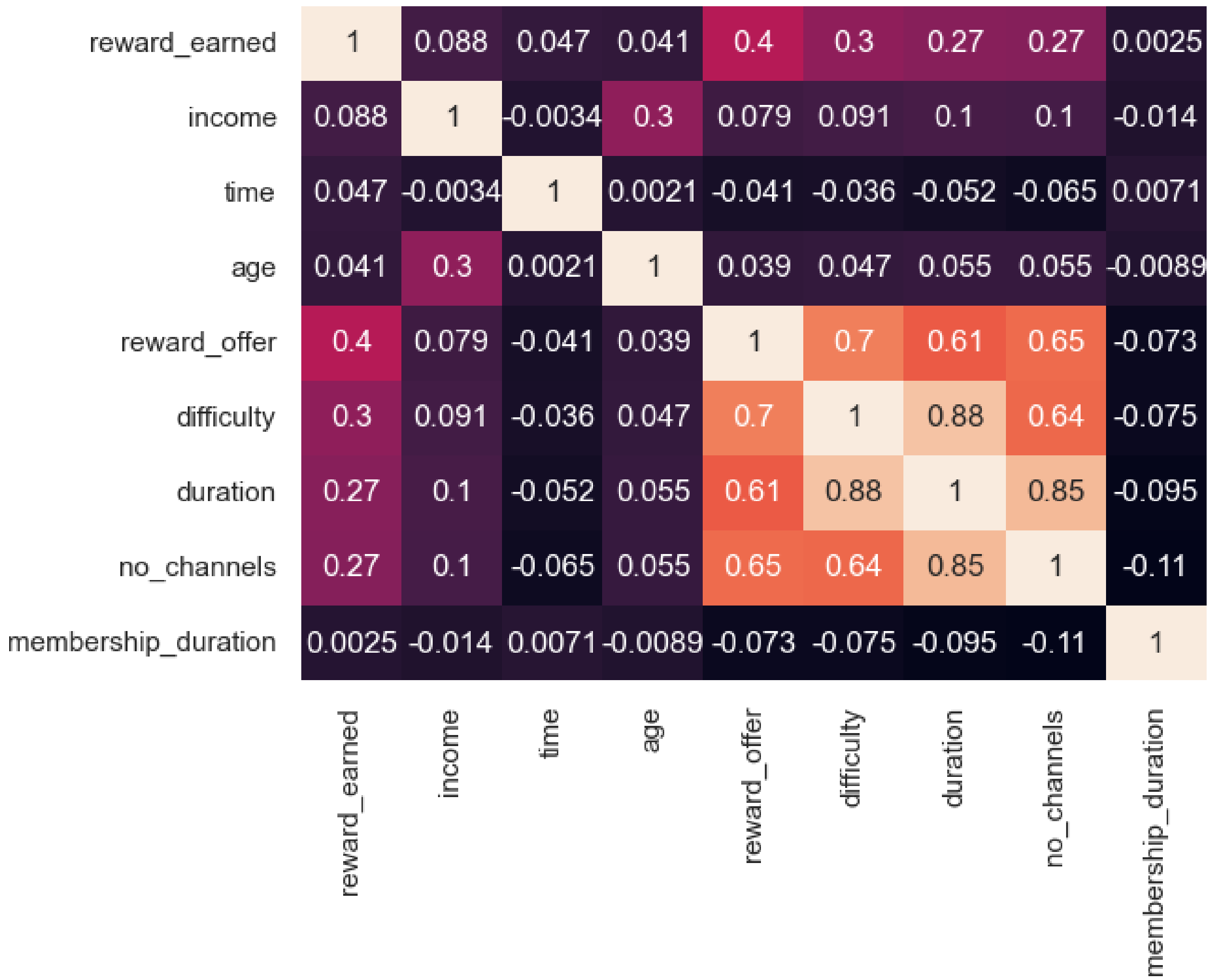
A correlation heatmap (Figure 1) was generated to identify relationships between numerical features such as reward, duration, income, and time. This helped prioritize features for predictive modeling.

Key Insights:

Strong positive correlation between difficulty, duration, and reward, indicating that higher rewards require longer durations and greater spending thresholds.

Time had a moderate correlation with offer completion rates, suggesting that the timing of offer delivery plays a critical role.

Figure 1: Correlation Heatmap of Key Features



Distribution of Events:

A breakdown of event types revealed the proportion of offers received, viewed, and completed, highlighting the significant drop-off between viewing and completing offers.

Demographic Analysis:

Examined the distribution of gender, age, and income across the dataset. For example:

The majority of users fell into the 'mid-career' age group, with a higher propensity to complete offers.

Gender distribution was nearly even, but males exhibited a slightly higher completion rate for discount offers.

Channel Effectiveness:

Analysis of delivery channels (web, email, mobile, social) showed that multi-channel offers had higher completion rates compared to single-channel offers.

4. Methodology

This section outlines the methodology used to predict offer completion rates, detailing the feature selection process, machine learning models applied, and the evaluation metrics utilized for performance comparison.

4.1 Feature Selection

Feature selection was based on both exploratory data analysis and domain knowledge to ensure the inclusion of variables most relevant to predicting offer completion. The selected features are grouped into continuous and categorical variables:

Continuous Features:

Time: The time (in hours) when the offer was received or acted upon, relative to the start of the test period.

Duration: The number of days for which the offer was active.

Difficulty: The minimum amount of spending required to qualify for the offer reward.

Income: The annual income of the user.

Categorical Features:

Offer Type: Type of offer (BOGO, discount, informational).

Age Categories (Age_Cat): Age groups categorized into bins (e.g., young adult, mid-career, senior).

Gender: Gender of the user (M, F, O, or Unknown).

Channels: Binary features indicating whether the offer was delivered via web, email, mobile, or social channels.

Feature engineering was performed to transform raw data into meaningful inputs for the machine learning models. For instance:

Binary encoding was used for channel presence.

One-hot encoding was applied to categorical variables like gender, age categories, and offer type.

Scaling of continuous variables was done using StandardScaler to normalize their range.

4.2 Models Used

To ensure robust predictions and account for the inherent class imbalance in the dataset, a variety of machine learning models were implemented and compared. Each model was tuned using hyperparameter optimization to maximize performance:

Logistic Regression:

A baseline linear model to predict offer completion probabilities.

Class weights were adjusted to handle imbalanced data.

Random Forest:

An ensemble learning method that uses multiple decision trees to improve predictive accuracy and reduce overfitting.

Hyperparameters tuned included the number of trees, maximum depth, and minimum samples per split.

Gradient Boosting:

A sequential ensemble method that builds weak learners (decision trees) iteratively, focusing on errors from previous iterations.

Achieved the best overall performance in terms of balancing precision, recall, and F1-score.

Support Vector Machine (SVM):

A non-linear model using an RBF kernel for classification.

Struggled with class imbalance, leading to poor recall for the minority class.

XGBoost:

An optimized implementation of gradient boosting with scalability and performance improvements.

Applied scale position weighting to address class imbalance effectively.

Each model was trained on 80% of the data and tested on the remaining 20% to evaluate its predictive performance.

4.3 Evaluation Metrics

To assess the models' effectiveness and account for class imbalance, multiple evaluation metrics were used:

Accuracy:

Measures the proportion of correct predictions out of the total predictions.

While easy to interpret, accuracy can be misleading in the presence of class imbalance, as it may over-represent the majority class.

F1-Score:

The harmonic mean of precision and recall, providing a single metric that balances false positives and false negatives.

Particularly important for this problem, where both types of errors have significant implications.

Precision:

Measures the proportion of true positives among all predicted positives.

Useful to evaluate how well the model minimizes false positives (e.g., predicting offer completions that don't actually occur).

Recall:

Measures the proportion of true positives among all actual positives.

Reflects the model's ability to identify all instances of offer completions, even at the expense of some false positives.

Confusion Matrix:

Provides a detailed breakdown of true positives, true negatives, false positives, and false negatives, offering deeper insight into model performance.

Macro-Averaged Metrics:

Used to evaluate model performance across classes by equally weighting each class, regardless of class imbalance.

Metrics such as macro-averaged precision, recall, and F1-score were calculated.

These metrics, combined with hyperparameter tuning and careful model evaluation, allowed for a thorough comparison of each model's ability to predict offer completions accurately and reliably.

5. Results

This section provides a detailed analysis of model performance based on the evaluation metrics, highlighting the strengths and weaknesses of each model and identifying the most suitable one for predicting offer completion.

5.1 Model Performance Comparison

Figure 2: Summary of the performance of the five machine-learning models evaluated in this project:

Model Performance Summary

Model	Precision (Minority Class)	Recall (Minority Class)	F1 Score (Minority Class)	Overall Accuracy
Logistic Regression	0.26	0.99	0.41	0.67
Random Forest	0.32	0.66	0.44	0.80
Gradient Boosting	0.64	0.58	0.61	0.91
SVM	1.00	0.00	0.00	0.88
XGBoost	0.49	0.91	0.64	0.88

5.2 Analysis of Model Performance

Each model's performance is evaluated based on its ability to balance precision, recall, and F1-Score for the minority class while maintaining high overall accuracy. The key observations are as follows:

1. **Logistic Regression**:
   - **Strengths**:
     - Achieved the **highest recall** for the minority class (0.99), indicating it can identify almost all actual offer completions.
   - **Weaknesses**:
     - Very low precision (0.26), meaning many false positives.
     - Low F1-Score (0.41) and accuracy (0.67), reflecting an overall imbalance in performance.
2. **Random Forest**:
   - **Strengths**:
     - Balanced performance between precision (0.32) and recall (0.66) for the minority class.
     - Reasonable overall accuracy (0.80).
   - **Weaknesses**:
     - Moderate F1-Score (0.44), indicating room for improvement in balancing precision and recall.
3. **Gradient Boosting**:
   - **Strengths**:
     - Delivered the **highest overall accuracy** (0.91), demonstrating robust performance across all classes.

        - Achieved a balanced F1-Score (0.61) for the minority class, indicating a good trade-off between precision (0.64) and recall (0.58).
    - **Weaknesses**:
        - Recall for the minority class could be improved, as it is slightly lower than other models like Logistic Regression or XGBoost.
4. **Support Vector Machine (SVM)**:
    - **Strengths**:
        - Perfect precision (1.00) for the minority class, meaning no false positives.
    - **Weaknesses**:
        - **Failed to recall any minority class instances** (recall = 0.00), leading to an F1-Score of 0.00.
        - This indicates the model is unsuitable for imbalanced datasets where minority class prediction is critical.
5. **XGBoost**:
    - **Strengths**:
        - Achieved a **strong recall** for the minority class (0.91), second only to Logistic Regression.
        - Balanced performance with an F1-Score of 0.64 and high accuracy (0.88).
    - **Weaknesses**:
        - Precision for the minority class (0.49) is moderate, indicating potential for improvement with further tuning.

5.3 Key Insights

1. **Gradient Boosting Emerges as the Best Model**:
    - Gradient Boosting outperforms other models with its highest overall accuracy (0.91) and a well-balanced F1-Score (0.61), making it the most reliable choice for predicting offer completions.
2. **XGBoost as a Strong Contender**:
    - XGBoost delivers strong recall (0.91) for the minority class and balanced performance, indicating it could be a viable alternative if further hyperparameter tuning is applied.
3. **Challenges with SVM**:

    - The SVM model highlights the challenges of handling class imbalance, achieving no recall for the minority class despite its high accuracy.
4. **Trade-offs in Logistic Regression and Random Forest**:
    - Logistic Regression prioritizes recall but sacrifices precision, resulting in poor F1-Score.
    - Random Forest provides balanced performance but does not achieve the same level of accuracy or F1-Score as Gradient Boosting or XGBoost.

This detailed analysis demonstrates the strengths and limitations of each model, supporting the decision to prioritize Gradient Boosting for deployment while considering refinements to XGBoost for further optimization.

6. Discussion and Insights

This section delves into the critical observations, challenges, and implications of the project findings, focusing on the performance of models and the business relevance of feature importance analysis.

6.1 Key Observations

Gradient Boosting as the Most Balanced Model:

Gradient Boosting emerged as the top-performing model, achieving the highest overall accuracy (0.91) and a well-balanced F1-Score (0.61) for the minority class.

The model demonstrates a good trade-off between precision (0.64) and recall (0.58) for minority class predictions, making it suitable for handling imbalanced datasets.

Feature Importance Analysis:

Time is the most influential factor, highlighting the critical role of offer timing in driving completions. Offers sent during peak customer activity periods are more likely to succeed.

Duration also plays a significant role, suggesting that longer offer windows allow customers more time to act, increasing the likelihood of completion.

Other moderately impactful features, such as difficulty and reward amount, underscore the importance of simplifying offers and enhancing their perceived value.

Behavioral Insights over Demographics:

Features like age categories and gender showed minimal importance, indicating that demographic factors play a lesser role in offer completion rates.

The findings support a shift towards behavioral targeting, focusing on customer activity patterns and preferences rather than demographic profiles.

XGBoost as a Strong Alternative:

XGBoost delivered strong recall (0.91) for the minority class and a high F1-Score (0.64), indicating its potential for deployment with additional tuning.

SVM Challenges:

Despite achieving perfect precision (1.00) for the majority class, SVM failed to recall any minority class instances, reflecting its limitations in handling imbalanced datasets.

6.2 Limitations

Class Imbalance:

The inherent class imbalance in the dataset remains a significant challenge. Although techniques like class weighting and hyperparameter tuning helped improve performance, fully addressing this issue requires advanced methods like oversampling (e.g., SMOTE) or ensemble techniques.

SVM Model Limitations:

The SVM model completely failed to identify minority class instances, achieving a recall of 0.00 for the minority class. This underscores the difficulty of using SVM in scenarios with highly imbalanced data.

Simplistic Behavioral Assumptions:

The dataset focuses on transactional and event-based data, which may oversimplify real-world behavioral patterns. Factors like seasonality, customer preferences, and market trends were not considered.

Limited Scope of Offer Types:

The analysis is constrained by the three offer types (BOGO, discount, informational) present in the dataset. Additional offer variations or promotional strategies might yield deeper insights.

Dependence on Simulated Data:

The dataset is a simulation of real-world events, meaning that certain customer behaviors or external factors influencing offer completions may not be accurately reflected.

6.3 Strategic Implications

Optimizing Offer Timing:

Insights from feature importance analysis suggest prioritizing offers during periods of peak customer activity (e.g., weekends, holidays, or specific times of day).

Simplifying Offers:

Reducing offer difficulty and increasing duration can significantly enhance completion rates, as customers are more likely to engage with less complex promotional structures.

Refining Targeting Approaches:

Behavioral targeting, such as analyzing purchase history and activity patterns, should take precedence over traditional demographic segmentation.

Enhancing Minority Class Recall:

Future models should focus on improving recall for the minority class to ensure better identification of successful offer completions.

Exploring Advanced Techniques:

Incorporating advanced sampling techniques, such as SMOTE or cost-sensitive learning, could help mitigate class imbalance and enhance model performance.

6.4 Opportunities for Future Work

Integration of External Data:

Including external data sources, such as market trends or customer feedback, could improve the robustness of predictions.

Dynamic Offer Personalization:

Developing dynamic recommendation systems based on real-time customer activity and preferences could significantly improve offer relevance and completion rates.

In-depth Customer Segmentation:

A deeper analysis of customer segments, beyond basic demographics, could provide actionable insights for tailoring promotional strategies.

Model Ensemble Approaches:

Combining the strengths of Gradient Boosting and XGBoost through ensemble methods might yield superior results, particularly for minority class predictions.

By addressing these limitations and leveraging the identified opportunities, the proposed strategies can be further refined to maximize their impact on business outcomes.

7. Recommendations

This section provides actionable suggestions categorized into short-term actions for immediate implementation and long-term strategies to enhance future marketing campaigns and improve offer completion rates.

7.1 Short-Term Actions

Deploy Gradient Boosting Predictions:

Leverage the predictions from the Gradient Boosting model for the upcoming campaign. Its balanced performance in handling the minority class ensures better identification of potential customers who are likely to complete offers.

Optimize Offer Timing:

Implement timing strategies based on customer activity patterns (e.g., time of day, weekdays vs. weekends).

Use feature importance insights to prioritize sending offers during high-activity periods, such as lunch breaks, evenings, or seasonal sales events, to maximize completion rates.

Test Simpler Offers:

Simplify current offers by lowering difficulty levels and extending their duration. Customers may find it easier to engage with and complete offers that are less complex and provide sufficient time for consideration.

Target High-Value Customers:

Utilize insights about membership duration to focus on loyal customers who are more likely to engage with offers. Tailored campaigns for these groups could yield higher returns.

7.2 Long-Term Strategies

Refine Models Using Advanced Sampling Techniques:

Address class imbalance by applying advanced techniques such as Synthetic Minority Oversampling Technique (SMOTE), ADASYN, or cost-sensitive learning to improve recall and precision for the minority class.

Incorporate ensemble models combining Gradient Boosting and XGBoost to capitalize on their respective strengths.

Explore Temporal Patterns (e.g., Seasonal Trends):

Conduct an in-depth analysis of seasonal trends and market conditions to identify cyclical patterns influencing offer completion rates.

Use time-series modeling to align campaigns with periods of heightened customer activity, such as holidays, back-to-school seasons, or end-of-year sales.

Develop Dynamic Campaign Strategies:

Integrate dynamic recommendation systems that use real-time customer activity data to optimize offer delivery. For example:

Adjust offers based on recent transaction histories or engagement behaviors.

Employ A/B testing to fine-tune offer types, timing, and channels.

Behavioral Targeting and Multi-Channel Campaigns:

Shift focus from demographics to behavioral insights, such as shopping habits and preferred channels (e.g., email, social media).

Design multi-channel campaigns that engage customers across multiple touchpoints to increase visibility and engagement.

Invest in Data Enrichment:

Enhance the dataset by integrating external data sources such as:

Customer feedback and survey results.

Market trends and competitor analysis.

Social media sentiment analysis to gauge brand perception.

Monitor and Iterate:

Establish a feedback loop for continuous improvement. After deploying campaigns, analyze performance metrics and refine models based on results.

Regularly retrain models with updated data to maintain relevance and predictive accuracy.

7.3 Additional Recommendations

Implement Personalized Offers:

Utilize machine learning models to personalize offer types and rewards for different customer segments, ensuring higher engagement rates.

For example, loyal customers could receive BOGO offers, while new users might respond better to discounts.

Optimize Channels for Delivery:

Analyze the effectiveness of each channel (e.g., web, email, mobile) in driving offer completions and focus resources on high-performing platforms.

Experiment with cross-channel strategies, such as combining email notifications with targeted ads, to maximize reach.

Regular Performance Reviews:

Conduct periodic model evaluation sessions to ensure optimal performance.

Adjust hyperparameters, retrain models, and experiment with new algorithms to stay ahead of changing customer behaviors.

By implementing these short-term actions and long-term strategies, the organization can improve its marketing ROI, increase customer engagement, and achieve higher offer completion rates.

8. Conclusion

This study highlights the critical role of machine learning in enhancing marketing strategies by predicting offer completion rates and identifying key influencing factors. Through rigorous analysis and model evaluation, Gradient Boosting emerged as the most effective approach for handling class imbalance and predicting offer completions, delivering the best balance between precision and recall for the minority class.

Key findings include:

Gradient Boosting's Effectiveness: Its superior performance in balancing metrics underscores its suitability for predicting offer completions in imbalanced datasets.

Behavioral Factors Over Demographics: Time and offer duration were identified as the most influential features, emphasizing the need to focus on customer behavior rather than demographic characteristics.

Actionable Strategies: Recommendations such as optimizing offer timing, simplifying offers, and tailoring campaigns for high-value customers provide a clear roadmap for improving engagement and ROI.

While significant progress was made, challenges such as class imbalance and limited precision for the minority class remain. Addressing these limitations through advanced sampling techniques, temporal analysis, and dynamic campaign strategies will further enhance predictive accuracy and marketing impact.

In conclusion, this project demonstrates the power of data-driven decision-making in marketing, providing valuable insights to optimize promotional strategies and drive business growth. Future work should focus on refining models and leveraging additional data to sustain and build upon these improvements.

9. References

Scikit-learn documentation: https://scikit-learn.org/

Dataset Source: https://github.com/CoskunErden/Udacity_DS_Capstone_Project