AWS Machine Learning Engineer Nanodegree

Capstone Project Report

Abstract

This project, undertaken as part of a Udacity Nanodegree course on SageMaker, aimed to analyze customer behavior within the Starbucks rewards mobile app. The primary objective was to predict the probability of a customer redeeming an offer after viewing it, based on their demographic details and past interactions. To achieve this, a LightGBM model was developed, achieving a strong performance with 68% accuracy and an impressive recall rate of 83%.

The model offered key insights into the factors influencing offer redemption. It revealed that customers with moderate rewards, long-term memberships, and engagement through social channels were more likely to complete offers. Additionally, middle-income individuals and those over 40 had a higher likelihood of redeeming offers, particularly when the reward value was increased. Simpler offers also saw higher completion rates.

Overall, this model proved effective in identifying a large proportion of completed offers, making it a valuable resource for enhancing marketing strategies and optimizing promotional campaigns. This project highlights the potential of machine learning, particularly using SageMaker, to improve customer engagement and support business growth.

Definition

Project Overview

This project focuses on analyzing customer behavior within the Starbucks rewards mobile app. The objective is to understand how customers respond to different promotional offers based on their demographics and past interactions. Gaining these insights will help Starbucks refine its marketing strategies, enhance customer engagement, and drive revenue growth.

The key challenge is identifying demographic groups that are more likely to respond positively to specific promotional offers. Since Starbucks provides a variety of promotions to its rewards app users, response rates can vary across different customer segments. This project aims to explore customer demographics and preferences, using this information to predict the likelihood of a user redeeming an offer after viewing it.

Proposed Solution

To address this challenge, a machine learning model will be designed to predict the likelihood of a user redeeming a promotional offer after viewing it. The model will leverage customer demographics and past interactions to identify which offer characteristics are most effective for specific customer segments.

As an initial benchmark, a heuristic approach will be used by analyzing historical response rates. While this rule-based method provides a general understanding of how different customer groups react to offers, it lacks the accuracy and sophistication of a machine learning model, which incorporates a broader range of features and individual user behaviors.

Project Methodology

The project methodology consists of several key steps. First, the data will undergo cleaning, preprocessing, and formatting to ensure consistency. Next, exploratory data analysis will be performed to uncover trends, correlations, and potential features for the machine learning model. Following this, feature engineering will be applied to improve the model's predictive accuracy.

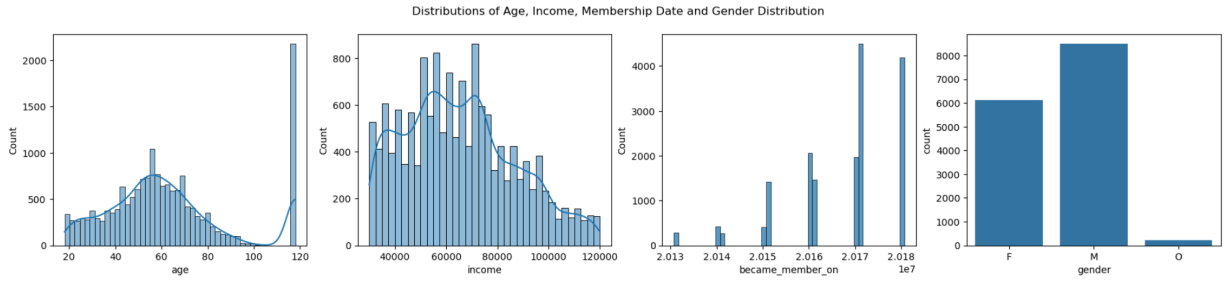
To automate model selection, SageMaker AutoGluon will be utilized to evaluate multiple algorithms on the validation set and determine the best-performing model. The chosen model will then be trained and fine-tuned using SageMaker's advanced capabilities. Its performance will be evaluated on the test set and compared against a baseline model. Finally, the results will be analyzed to provide insights into customer behavior and their responses to different promotional offers.

Data Exploration and Preprocessing

The data exploration and preprocessing phase involved analyzing and cleaning three datasets provided for this challenge. These datasets contained information on rewards program users, promotional offers, and user interaction events.

Profile_Dataset

The user profile dataset included demographic details for 17,000 Starbucks rewards program users, such as age, membership start date, and income. The average age of users was 62.53 years, while the mean income was $65,404. Membership initiation dates ranged from 2013 to 2018.

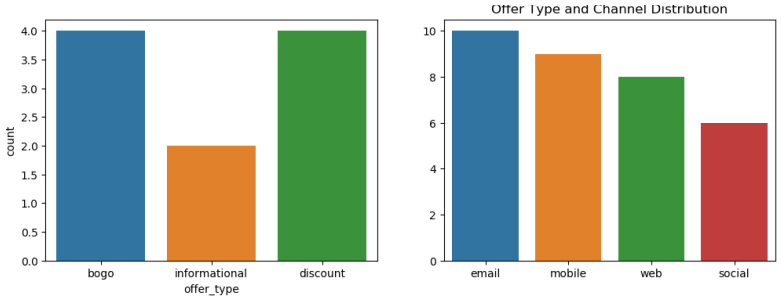


Upon reviewing the user profile dataset, we noticed a significant number of customers had their age recorded as '118', which we interpreted as a default placeholder. These values were replaced with NULL. Additionally, we found that when one of the key demographic predictors (age, income, gender) was missing, all of them were absent, so we removed these entries, which made up 11% of the dataset.

For feature engineering, the 'became_member_on' column was transformed into 'membership_days', representing the number of days since a customer joined the program. This modification helped simplify the model. The 'gender' column was one-hot encoded to make it compatible with models that do not accept categorical data. The 'age' feature was divided into six categories: ['18-24', '25-34', '35-44', '45-54', '55-64', '65+'], in order to prevent overfitting during model training.

Portfolio_Dataset

The offer portfolio dataset contained details about 10 different offers sent during a 30-day trial period, including information on rewards and offer types, such as BOGO (Buy One Get One), discounts, and informational offers.



In preparation for prediction, the 'channel' column was transformed from an array format into multiple boolean columns, each representing a distinct channel. This allowed for a better analysis of the predictive influence of each channel. Similarly, the 'offer type' columns were converted into boolean dummy columns.

Transcript Dataset
The user interaction transcript dataset recorded 306,534 user interactions, including timestamps, transaction amounts, and rewards. The average transaction amount was $12.78, with a standard deviation of $30.25. Rewards ranged from 2 to 10, with an average of 4.9 and a standard deviation of 2.89.

From this dataset, two key metrics were defined for further analysis:

1. **Likelihood of viewing an offer**: This metric assesses the probability that a customer will view the promotional offers sent to them.
2. **Likelihood of completing an offer, given it was viewed**: This metric focuses on customers who have already viewed an offer and measures the likelihood of them completing it, helping to evaluate the effectiveness of the offer post-viewing.

During this exploration, it was found that some customers received the same offer multiple times, with up to 17% of occurrences being duplicates. To address this, the dataset was deduplicated at the customer-offer level by taking the minimum time and the maximum transaction amount and reward value. The customer and offer features were then merged with the target variable into a single dataset, resulting in one row per customer-offer combination. The mean likelihood of viewing an offer was 90%, while the likelihood of completing it after viewing was 37%, indicating an imbalanced dataset.

Ultimately, we decided to focus on the second metric, "Likelihood of completing an offer, given it was viewed," as it provides a more accurate assessment of offer effectiveness. By concentrating on offer performance post-viewing, we can directly measure their impact and refine Starbucks' marketing strategies.

Implementation
This section covers two critical stages of our machine learning project: model selection and hyperparameter tuning.

Model Selection
Model selection is a pivotal step in identifying the most appropriate model that can reliably predict unseen data. In this project, our goal is to choose a model that can predict customer responses to promotional offers based on demographic data and historical interactions.

For automated model selection, we utilized SageMaker AutoGluon, a tool for automated machine learning (AutoML) that simplifies the process of selecting the best-performing model for a given dataset. AutoGluon tests a variety of machine learning algorithms with different hyperparameters and architectures, then compares their performance on a validation set. Algorithms tested include Neural Networks, Decision Trees, Random Forests, Gradient Boosting Machines, and Support Vector Machines.

Prior to model fitting, several preprocessing steps were carried out, including:

- Removing NaN values from the target column ('offer_completed_after_view') and converting it to a boolean type.
- Dropping features unsuitable for training, such as 'became_member_on', 'age_group', 'person', 'offer_id', and 'offer_viewed'.
- Splitting the dataset into training, validation, and test sets, with a ratio of 70%, 15%, and 15%, respectively. These datasets were uploaded to S3 using the `sess.upload_data()` function.

In our AutoGluon setup, we defined the task as a binary classification problem, with 'offer_completed_after_view' as the target label and 'average_precision' as the evaluation metric. The training time was limited to 30 minutes. After training, AutoGluon evaluated 14 models, with the top three being:

- **WeightedEnsemble_L2**: Validation score of 0.64 and prediction time of 4.22 seconds.
- **NeuralNetFastAI_BAG_L1**: Validation score of 0.63 and prediction time of 2.10 seconds.
- **LightGBMXT_BAG_L1**: Validation score of 0.63 and prediction time of 1.07 seconds.

To prioritize simplicity and interpretability, we chose the **LightGBMXT_BAG_L1** model, as it offered a good balance of performance and faster prediction time. Additionally, LightGBM is efficient, widely available in SageMaker, and offers insights into feature importance, making it a convenient and valuable tool for this project.

Model Training and Hyperparameter Tuning
For model training and hyperparameter tuning, we used Amazon SageMaker's built-in capabilities. The **LightGBM** classification model, a gradient boosting framework based on tree-based learning algorithms, was selected for this task.

We retrieved the necessary Docker image, training script, and pre-trained model tarball for LightGBM. These components were then fine-tuned for our specific task. The hyperparameters explored included:

- **Learning Rate**: Ranging from 0.01 to 0.2.
- **Number of Leaves**: Ranging from 2 to 50.
- **Feature Fraction**: Ranging from 0.5 to 1.
- **Bagging Fraction**: Ranging from 0.5 to 1.
- **Bagging Frequency**: Ranging from 1 to 10.
- **Maximum Depth**: Ranging from 1 to 10.
- **Minimum Data in Leaf**: Ranging from 1 to 30.
- **Extra Trees**: Set to both True and False.

A **SageMaker Estimator** instance was created with the retrieved model and hyperparameters, and a hyperparameter tuning job was initiated with the Bayesian optimization strategy. This strategy was chosen because it uses past results to guide future hyperparameter evaluations, making the process more efficient.

The tuner was allowed to run a maximum of 20 jobs, with 3 jobs running in parallel. After the tuning process was completed, we analyzed the first 10 hyperparameter combinations from the logs to evaluate if simpler configurations could offer similar performance.

Rank	Average Precision	Validation Time (s)	Learning Rate	Number of Leaves	Feature Fraction	Bagging Fraction	Bagging Frequency	Max Depth	Min Data in Leaf	Extra Trees
1	0.67	31.48	0.11	50	1	1	7	7	2	1
2	0.66	81.71	0.12	50	0.5	0.51	7	9	27	1
3	0.66	31.4	0.04	48	1	0.91	7	6	7	1
4	0.66	65.69	0.03	37	1	0.98	7	6	26	1
5	0.65	31.57	0.02	24	1	0.85	7	9	26	1
6	0.65	31.53	0.09	47	0.96	0.84	6	4	6	1
7	0.64	31.5	0.09	18	0.89	0.53	7	7	30	1
8	0.64	31.41	0.01	22	0.81	0.5	2	5	4	1
9	0.63	31.61	0.05	11	0.97	0.5	7	7	15	1
10	0.61	31.51	0.05	50	0.93	0.96	6	1	28	1

The selected model was configured with the following hyperparameters: a bagging fraction and feature fraction of about 1.0, bagging frequency of 7, a learning rate of roughly 0.109, maximum depth of 7, minimum data in a leaf of 2, and number of leaves of 50. The model also employed the 'gbdt' boosting type, had early stopping rounds set to 30, and used the extra trees method. This configuration yielded an average precision of approximately 66.7%, the highest .

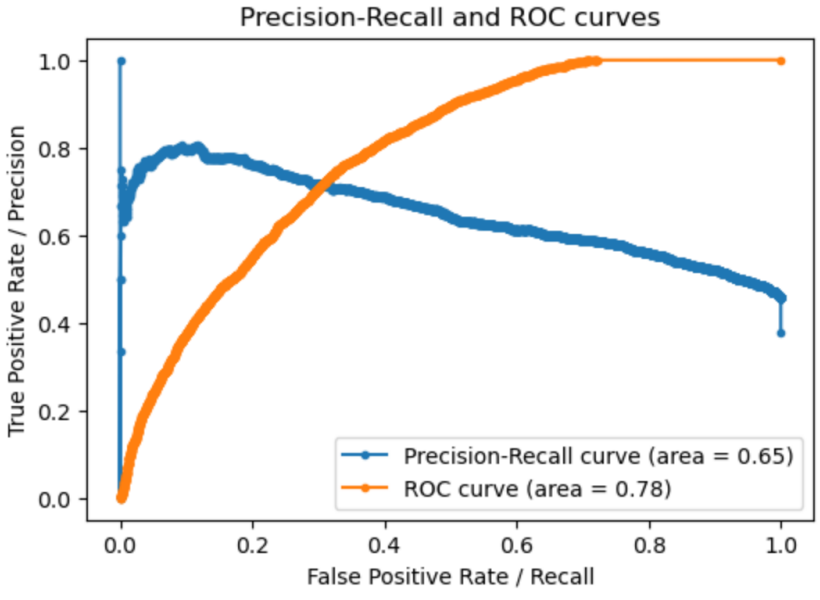
Evaluation

In this final section, we assess the performance of our predictive model, focusing on its ability to predict whether a customer will complete an offer after viewing it. The model's performance is compared to a baseline model to evaluate its relative effectiveness. We also examine the characteristics of the best-performing model, its potential biases, and areas for improvement. Additionally, we investigate the features that play a significant role in predicting offer completion. This comprehensive evaluation provides a deeper understanding of the model's strengths and weaknesses, while offering valuable insights for future enhancements.

Model Performance
In this section, we present the model's performance, evaluated without a fixed threshold, and compare it to the baseline model.

Threshold Agnostic
This approach allows us to assess the model's performance across various thresholds, offering a complete picture of its predictive power. We focus on two key metrics:

- **Precision-Recall Area Under the Curve (PR AUC)**: This metric measures the area beneath the precision-recall curve and is particularly useful for evaluating performance in imbalanced datasets. A high PR AUC score indicates that the model is good at identifying customers who complete offers (high recall) and that the customers predicted to complete offers are likely to do so (high precision).
- **Receiver Operating Characteristic Area Under the Curve (ROC AUC)**: This metric evaluates the area under the ROC curve, balancing the trade-off between the true positive rate (TPR) and false positive rate (FPR). A high ROC AUC score suggests the model is effective at distinguishing between customers who will complete an offer and those who won't.



Our model achieved a **PR AUC score** of approximately 0.644 and an **ROC AUC score** of about 0.780. These results indicate a reasonable performance in predicting customers who will complete an offer after viewing it, with a stronger capability to distinguish between customers who completed the offer and those who did not.

Fixed Threshold
The choice of threshold plays a significant role in performance metrics for imbalanced classification problems. While there is no single "optimal" threshold, selecting one allows us to compute important metrics such as precision, recall, and F1-score, and to generate a confusion matrix, which offers more detailed insights into the model's performance.

A common approach is to choose the threshold that maximizes the **F1-score**, which balances both precision and recall. However, the best threshold depends on the business objectives and the relative importance of precision versus recall. For this project, the threshold that maximized the **F1-score** was approximately **0.35**. Using this threshold, we obtained the following results:

Metric	Offer Not Completed	Offer Completed
Precision	0.85	0.55
Recall	0.58	0.83
F1-score	0.69	0.66

The results show our model has higher precision for non-completing customers but higher recall for completing ones. This indicates the model is adept at identifying non-completing customers but may misclassify some as completing. However, it accurately identifies a large portion of completing customers. The model's overall accuracy is 0.68, correctly classifying 68% of instances. This balance can be adjusted depending on whether precision or recall is prioritized.

Comparison with baseline

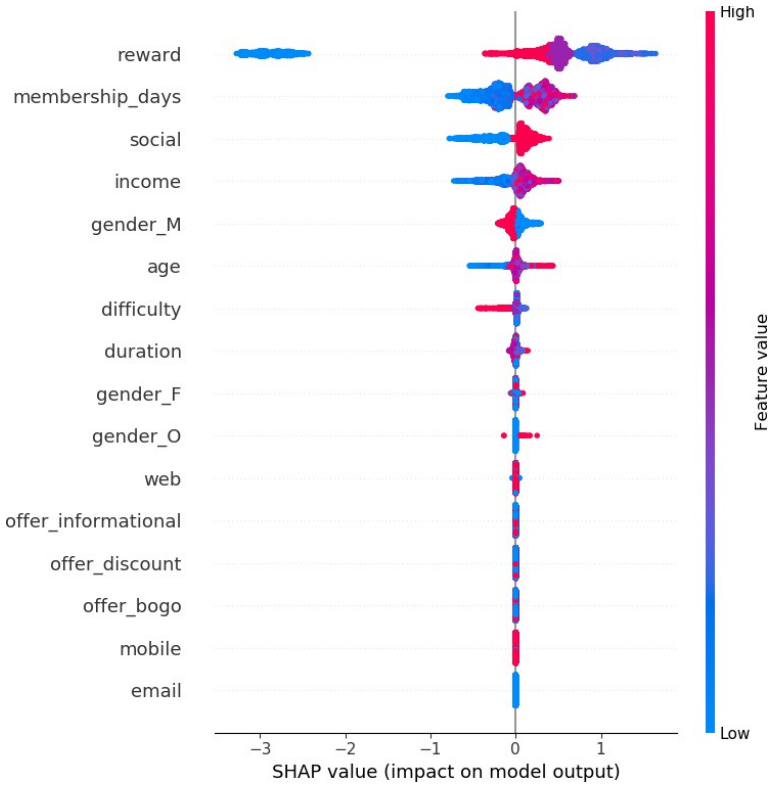
Baseline models serve as a reference point to gauge a model's performance. We're comparing our LightGBM model to a baseline model that predicts the majority class, using McNemar's test to determine if the performance difference is statistically significant. The McNemar's test confusion matrix reveals:

- 456 customers didn't complete the offer and were accurately predicted by both models.
- 2213 customers completed the offer, correctly predicted by the LightGBM model but not the baseline.

- 1821 customers completed the offer, correctly predicted by the baseline but not the LightGBM model.
- 2551 customers completed the offer and were accurately predicted by both models. .

With a p-value of 0.000 from McNemar's test, the performance difference between the models is statistically significant. The LightGBM model correctly predicted more customers who completed the offer than the baseline model, indicating its superior performance. However, the business context and the costs of false positives and negatives should be considered when selecting a model.

Model Interpretability



In this section, we will utilize the **SHAP (SHapley Additive exPlanations)** summary plot to gain insights into the features of the model and their influence on predicting offer completion. SHAP is a powerful tool that helps visualize the contribution of each feature to the model's predictions. In the summary plot, each dot represents the effect of a particular feature on the prediction for an instance. The y-axis lists the features, while the x-axis shows the SHAP value, which indicates how much a feature influences the model's output. The color of the dot corresponds to the feature's actual value, with warmer colors representing higher values and cooler colors representing lower values.

The SHAP feature importance summary plot reveals several key insights into the factors that influence offer completion, along with detailed analyses of specific feature interactions. These insights can inform targeted marketing strategies and help optimize promotional offers.

Key Insights from SHAP Feature Importance:

1. **Reward**: The relationship between reward and offer completion probability is non-linear. Initially, as the reward increases, so does the likelihood of completion. However, after a certain threshold, further increases in reward can decrease the probability of completion. This suggests diminishing returns or a perception of extreme rewards being unrealistic.
2. **Membership Days**: A positive correlation between the number of membership days and offer completion suggests that long-term members are more likely to complete offers, possibly due to stronger loyalty.
3. **Social Channel**: The social channel is the most effective medium for sharing offers, outperforming web, mobile, and email. This reflects the viral nature of social media, where users can easily share and engage with offers.
4. **Income**: Higher income levels are associated with a higher probability of offer completion. This is likely because wealthier customers have more disposable income to spend, making them more inclined to complete offers.
5. **Gender**: Male users are more likely to not complete the offer compared to female and non-binary users, suggesting potential differences in offer engagement by gender.
6. **Age**: Older customers tend to have a higher likelihood of completing offers, suggesting that age plays a role in engagement, possibly linked to greater financial stability or brand loyalty.
7. **Difficulty**: There is a negative correlation between the difficulty of the offer and its completion probability. Easier offers tend to have higher completion rates.
8. **Duration and Offer Type**: These features showed no significant importance in predicting offer completion, suggesting that other factors may play a more dominant role.

These insights are valuable for designing more effective marketing strategies, targeting specific customer segments, and optimizing channels for sharing offers. However, it's essential to remain cautious about fairness and avoid discriminatory practices, especially when tailoring offers based on gender or age.

Detailed Analysis of Specific Feature Interactions:

Reward and Membership Days:

- **No Reward (0)**: Long-term members are less likely to complete offers without rewards.
- **Low Rewards (2-3)**: Moderate rewards tend to increase completion rates, especially for long-term members.
- **Medium Reward (5)**: A consistent impact on offer completion across different membership durations.
- **High Reward (10)**: While high rewards encourage offer completion, their effectiveness diminishes for long-term members.

These patterns suggest that reward strategies should be tailored to different membership durations. Generally, lower rewards yield higher chances of completion, especially among long-term members.

Income and Reward Amount:

- **Low Income (30k-50k)**: Customers in this income range are less likely to complete offers, especially as reward amounts increase.

- **Middle Income (50k-100k)**: Customers in this group show a higher likelihood of completing offers, with completion rates increasing as reward amounts rise.
- **High Income (>100k)**: High-income individuals, although fewer in number, show a reduced likelihood of completion, particularly when the reward is less than 8.

This highlights the importance of aligning reward amounts with income brackets to maximize offer completion.

Gender and Reward Amount:

- **Male**: Male customers tend to have slightly negative SHAP values, indicating a lower likelihood of offer completion, particularly for higher rewards.
- **Non-Male (Female or Other)**: Non-male customers show slightly positive SHAP values, suggesting a higher likelihood of completion, especially when higher rewards are offered.

While these patterns can inform marketing strategies, they must be implemented in a way that avoids reinforcing stereotypes or discrimination. Gender-based targeting should be fair and inclusive.

Age and Reward Amount:

- **Younger Age (18-40)**: Younger customers are less likely to complete offers, particularly for high rewards. This could be due to limited disposable income or more selective shopping habits.
- **Older Age (>40)**: Older customers are more likely to complete offers, especially for higher rewards. This may be due to greater financial stability and stronger brand loyalty.

While these insights are useful, any age-based marketing strategies should be applied ethically, ensuring that no age group is unfairly excluded or targeted.

Conclusions

The **LightGBM model** developed for predicting whether a customer will complete an offer has shown strong effectiveness. By setting the threshold around **0.35**, optimized for maximizing the **F1-score**, the model correctly identifies **83%** of customers who complete an offer, with a high **recall** of **0.83**. However, its **precision** is **0.55**, meaning that when the model predicts a customer will complete an offer, it is correct **55%** of the time. The overall **accuracy** of the model stands at **0.68**.

The **Precision-Recall** and **ROC** curves further reinforce the model's performance, evaluating its ability to differentiate between customers who will and will not complete an offer. The model achieved a **PR AUC** score of **0.644** and an **ROC AUC** score of **0.780**, suggesting it is proficient at distinguishing between customers who complete offers and those who do not.

Key insights from the model's feature importance include:

- **Reward**: Moderate rewards boost the likelihood of offer completion, especially among long-term members. Conversely, very low or high rewards decrease this likelihood.
- **Membership Days**: Long-term members are more likely to complete offers.
- **Social Channel**: The social channel outperforms web, mobile, and email for offer sharing.
- **Income**: Middle-income customers are more likely to complete offers, particularly with higher rewards. Both low- and high-income customers show a decreased likelihood.

- **Difficulty**: Easier offers tend to have higher completion rates.
- **Duration and Offer Type**: These features did not show a significant impact on offer completion.
- **Gender**: Non-male customers are more likely to complete offers, especially those with higher rewards, while male customers are less likely to complete them.
- **Age**: Customers over 40 are more likely to complete offers, particularly those with higher rewards. Younger customers are less likely to do so.

These insights can guide marketing strategies, but it is essential to ensure fairness and avoid discrimination based on gender or age. Ethical considerations and business objectives should be taken into account when tailoring offers using these findings.

In summary, the model is highly effective in capturing offer completions, even at the expense of some false positives. Its results indicate that it can be a valuable tool for optimizing promotional offers and improving marketing strategies.

Next Steps

The next steps should involve retraining a new **LightGBM model**, focusing solely on the features that demonstrated strong predictive power. These features include reward, membership days, income, social channel, difficulty, and duration. By excluding gender and age-related features, the model's interpretability could improve, noise may be reduced, and performance might be enhanced. If performance doesn't improve, it will be important to reconsider whether including gender and age is aligned with the business goals.

Another promising approach would be to develop a **two-step model**. In this framework, the first model would predict whether a customer will view an offer. If the customer views the offer, the second model would predict whether they will complete the offer. The target population for the second model would consist of customers who have viewed the offer. Analyzing the features influencing both the likelihood of viewing and completing an offer in this two-stage process could enhance the model's accuracy and interpretability. By splitting the problem into two phases, we might uncover unique patterns for each stage, potentially leading to more accurate and insightful predictions.