Predicting the Effectiveness of Starbucks Offers Based on Demographic Groups

1. Definition

Domain Background

Starbucks, one of the largest coffeehouse chains globally, has a rewards mobile app that sends out offers to its customers periodically. These offers can vary from advertisements for a drink to discounts or buy-one-get-one-free (BOGO) deals. However, not all customers receive the same offers, and the challenge lies in determining which demographic groups respond best to each offer type. Understanding this relationship can help Starbucks tailor their offers and maximize customer satisfaction and sales. Predicting human behavior is one of the most challenging tasks, but if successful, it can yield tremendous benefits for a company's sales numbers. Therefore, there is a continuous need to improve algorithms and predictions to obtain accurate and reliable results, helping businesses make more informed decisions in their marketing strategies.

Problem Statement

The problem to be solved is determining which demographic groups respond best to different types of Starbucks offers. This involves analyzing customer transaction, demographic, and offer data to identify patterns and relationships. By solving this problem, Starbucks can improve its targeted marketing strategies and increase the effectiveness of its offers.

Datasets and Inputs

The data for this project consists of three JSON files:

1. portfolio.json: Contains offer IDs and metadata about each offer (duration, type, difficulty, reward, and channels).
2. profile.json: Contains demographic data for each customer (age, gender, income, and membership date).
3. transcript.json: Contains records for transactions, offers received, offers viewed, and offers completed.

The portfolio.json contains information about the offer types. There are three types of offers that can be sent:

1. buy-one-get-one (BOGO): if the recipient spends a certain amount the recipient gets a reward of equal amount
2. discount: the recipient receives a reward of a percentage of the amount spent
3. and informational: the recipient receives information about a product. There is no reward involved

profile.json
Rewards program users (17000 users x 5 fields)

- gender: (categorical) M, F, O, or null
- age: (numeric) missing value encoded as 118
- id: (string/hash)
- became_member_on: (date) format YYYYMMDD
- income: (numeric)

portfolio.json
Offers sent during 30-day test period (10 offers x 6 fields)

- reward: (numeric) money awarded for the amount spent
- channels: (list) web, email, mobile, social
- difficulty: (numeric) money required to be spent to receive reward
- duration: (numeric) time for offer to be open, in days
- offer_type: (string) bogo, discount, informational
- id: (string/hash)

transcript.json
Event log (306648 events x 4 fields)

- person: (string/hash)
- event: (string) offer received, offer viewed, transaction, offer completed
- value: (dictionary) different values depending on event type
    - offer id: (string/hash) not associated with any "transaction"
    - amount: (numeric) money spent in "transaction"
    - reward: (numeric) money gained from "offer completed"
- time: (numeric) hours after start of test

These datasets will be used to analyze customer responses to different offers and the impact of demographic factors on their behavior.

Evaluation Metrics

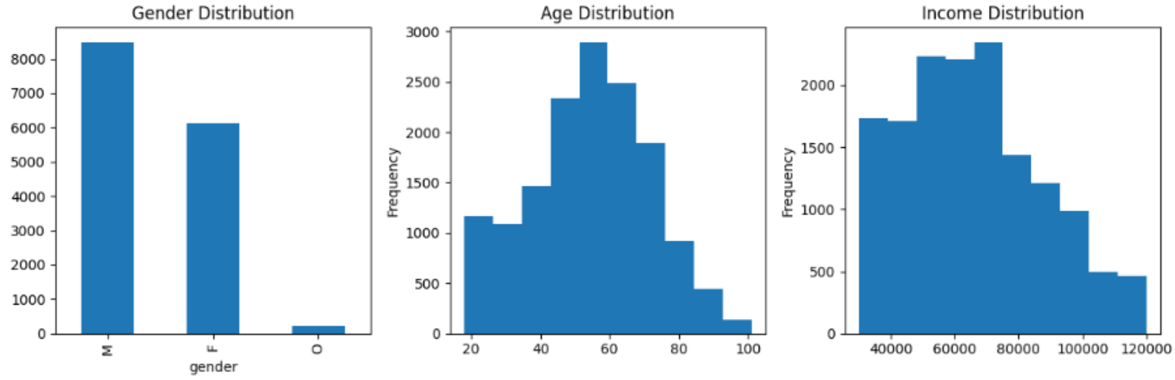
The evaluation metrics used in this project are:

1. Accuracy: Accuracy measures the proportion of correct predictions (both true positives and true negatives) out of all predictions made. It is a widely used metric that gives a general sense of how well the model performs. However, accuracy can be misleading if the dataset is imbalanced, as it might give high scores when the model only predicts the majority class well.
2. F1 Score: F1 score is the harmonic mean of precision and recall. It is used when we need to consider both false positives and false negatives equally. F1 score is particularly useful when dealing with imbalanced datasets, as it provides a balanced measure of the model's performance.
3. Recall: Recall (also known as sensitivity or true positive rate) measures the proportion of actual positive cases that the model correctly identifies as positive. It is important in situations where we want to minimize the number of false negatives (e.g., not identifying a customer who would respond well to an offer).
4. Precision: Precision measures the proportion of positive predictions that are actually correct. It is important in situations where we want to minimize the number of false positives (e.g., not targeting a customer who would not respond well to an offer).

2. Analysis

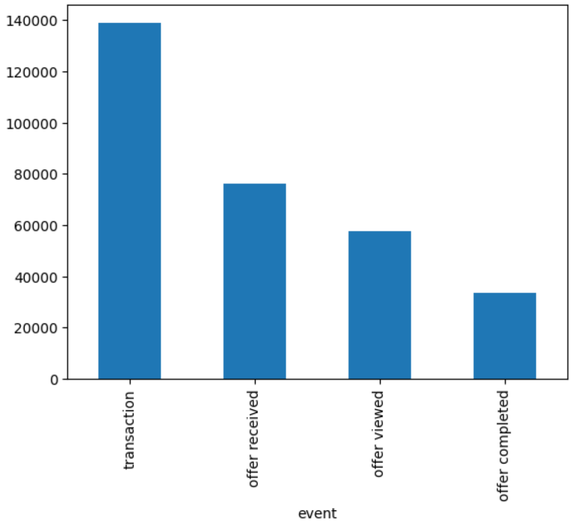
Data Exploration and Visualization

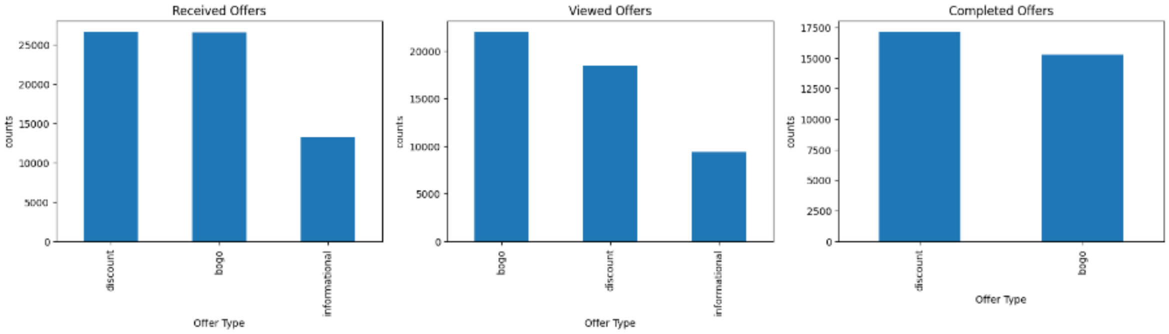
The final data set used for the machine learning models is composed of the portfolio, transcript and profile data sets. Missing Values were dropped to obtain a clean data set.
In the figure below the distribution of three demographic characteristics are displayed namely gender, age and income.



There are over 8,000 male customers, around 6,000 female customers and a minority from under 500 customers is assigned to another gender. The age is more or less normally distributed, same as the income.

Next, we look at the occurrences of the four different event types: transaction, offer received, offer viewed and offer completed. As to be expected the frequency drops from the number of offers received to the number of offers completed. Partially this is due to the fact, that informational offers cannot be completed as they are merely informative. But the more important aspect is that naturally not every received offer will be completed by a customer.





Algorithms and Techniques

To investigate if demographic characteristics can help predict if a customer will complete an offer or not we are going to use two machine learning models: Logistic Regression and LightGBM (Light Gradient Boosting Machine). These models were chosen based on their ability to handle classification problems and their performance on datasets with mixed data types and non-linear relationships. We will start with a baseline model using Logistic Regression and then move on to the more advanced LightGBM model.

Logistic Regression is a simple and widely used statistical method for analyzing a dataset with one or more independent variables that determine an outcome, which is a binary variable. It works by estimating the probability of the binary outcome based on the input features. Logistic Regression is chosen as the baseline model because it is easily interpretable, computationally efficient, and provides a good starting point to evaluate the performance of more complex models.

LightGBM is a gradient boosting framework that uses tree-based learning algorithms. It is designed to be efficient and scalable, handling large datasets with high accuracy and speed. The LightGBM model is chosen for its ability to handle categorical features, missing values, and complex interactions between features, making it suitable for our dataset with mixed data types and possible non-linear relationships between demographic features and offer completion. Additionally, LightGBM is known for its effectiveness in handling imbalanced datasets.

In both models, the input data will be preprocessed and transformed to be suitable for the algorithms. This includes encoding categorical variables, scaling numerical variables, and handling missing values. The models will then be trained and evaluated using cross-validation, and their performance will be compared using the evaluation metrics mentioned earlier (accuracy, F1 score, recall, and precision). This will help us determine the effectiveness of each model in predicting the completion of Starbucks offers based on demographic characteristics.

Benchmark

The benchmark model for this project is a simple Logistic Regression model.

3. Methodology

Data Preprocessing

Before training the models, the data from the three JSON files (portfolio, profile, and transcript) will be preprocessed and merged into a single dataset. This involves the following steps:

1. Encoding categorical variables: Channels, Offer Type, Gender and Event Type will be encoded using one-hot encoding, creating binary columns for each category.
2. Scaling numerical variables: Age and income will be scaled using StandardScaler to ensure they are on the same scale as other features and to improve the performance of the models.
3. Handling missing values: Rows with missing values in the demographic data (gender, age, and income) will be dropped to ensure a clean dataset.
4. Merging datasets: The datasets will be merged based on customer ID and offer ID, creating a final dataset with demographic information, offer details, and event data for each customer-offer pair.

Model Implementation and Refinement

After preprocessing the data, the Logistic Regression and LightGBM models will be implemented using their respective libraries in Python. The models will be trained using the prepared input features (demographic data and offer details) and the binary target variable (offer completed or not). Hyperparameter tuning for the LGBM will be performed using random search to find the optimal set of parameters for the model.

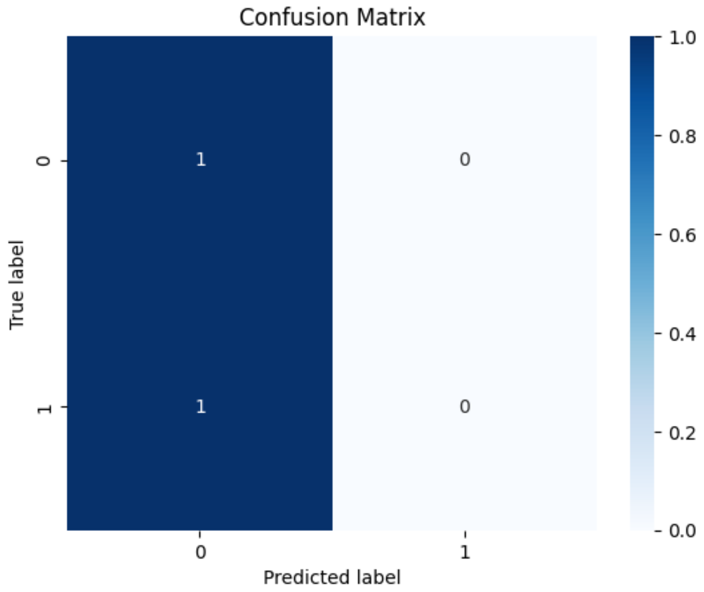
The performance of the Logistic Regression and LightGBM models will be evaluated using cross-validation. This involves splitting the dataset into training and validation sets multiple times and calculating the evaluation metric for each split. The average scores across all splits will be reported as the final performance of each model.

4. Results

The Logisitic Regression achieved the following scores for the previously defined evaluation metrics:

Accuracy: 0.782
F1 Score: 0.686
Recall: 0.782
Precision: 0.611

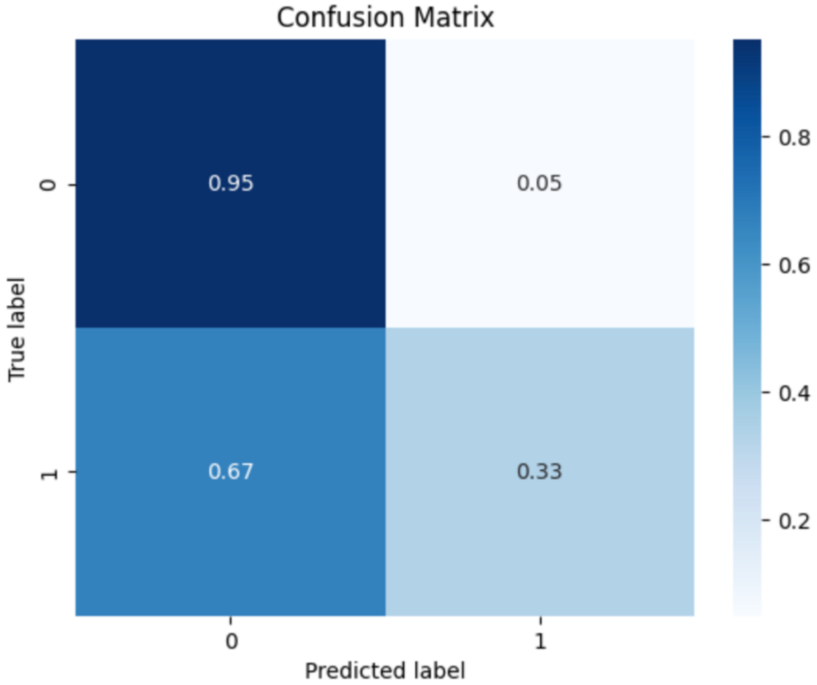
Even though the Accuracy score is not too bad this simple model fails to effectively predict if an offer is going to be completed or not. The Logistic Regression always predicts the majority class which only leads to such a high accuracy as the data set is imbalanced. The confusion matrix shown below demonstrates this fact:



The LGBM Model achieves significantly better results and is able to predict both classes. Even though the predictions for the minority class “offer completed” are not as good as those for the majority class the model is still able to predict a fair share correctly. The scores for the evaluation metrics are as follows:

Accuracy: 0.816
F1 Score: 0.792
Recall: 0.816
Precision: 0.796

The improvement is also visible in the confusion matrix:



Justification

The comparison of the evaluation metrics and confusion matrices for the Logistic Regression and LightGBM models reveals that the LightGBM model outperforms the baseline Logistic Regression model in all aspects. The improvement in accuracy, F1 score, recall, and precision indicates that the LightGBM model is more effective in predicting the completion of Starbucks offers based on demographic characteristics.

The superior performance of the LightGBM model can be attributed to its ability to handle categorical features, missing values, and complex interactions between features. Additionally, LightGBM is known for its effectiveness in handling imbalanced datasets, which is an issue faced in this project.

Based on these results, we can conclude that the LightGBM model is a significant improvement over the baseline Logistic Regression model and provides a valuable tool for Starbucks to optimize their targeted marketing strategies and increase the effectiveness of their offers.

5. Conclusion

Throughout this project, we have explored the problem of predicting the effectiveness of Starbucks offers based on demographic groups. We started by analyzing the data, exploring patterns and relationships between customer demographics and offer completion. We then implemented and compared two machine learning models, Logistic Regression and LightGBM, to predict offer completion.

The most interesting aspect of the project was observing the significant improvement in performance when using the LightGBM model compared to the baseline Logistic Regression model. This demonstrates the power of advanced machine learning techniques and their ability to handle complex datasets and relationships between features.

The most difficult aspect of the project was dealing with the imbalanced dataset and ensuring the models did not simply predict the majority class. This required careful consideration of the evaluation metrics and model selection.

The final LightGBM model fits our expectations for the problem and can be used in a general setting to predict the effectiveness of Starbucks offers based on demographic groups.

There are several ways in which the current implementation could be improved:

1. Model Selection: We could explore other machine learning models, such as neural networks or support vector machines, to see if they can further improve the prediction performance.
2. Hyperparameter Tuning: We could spend more time on hyperparameter tuning for the LightGBM model, using techniques like grid search or Bayesian optimization, to find the optimal set of parameters that yield the best results.
3. Ensemble Methods: We could combine multiple models using ensemble methods, such as stacking or bagging, to improve the overall prediction performance.

By implementing these improvements, it is possible that we could achieve even better results in predicting the effectiveness of Starbucks offers based on demographic groups. However, the current LightGBM model already shows a significant improvement over the baseline Logistic Regression model and provides a valuable tool for Starbucks to optimize their targeted marketing strategies.

1.