Starbucks Promotional Response Prediction: Project Report

1. Introduction

Digital loyalty ecosystems have transformed how companies engage with customers. Starbucks, through its highly successful mobile rewards program, delivers personalized offers such as Buy-One-Get-One (BOGO), percentage discounts, and informational promotions. However, not all customers respond the same way to every offer. Sending the wrong offer can waste marketing budget, reduce customer satisfaction, and distort revenue attribution.

This project applies machine learning to predict **whether a customer will respond to a Starbucks promotional offer**, using a realistic simulated dataset that captures complex customer behavior and multi-step event sequences. The final deliverable includes:

- A clean labeled dataset of customer–offer interactions
- Exploratory Data Analysis across demographics, offer types, and behavioral patterns
- Multiple machine learning models (Logistic Regression, Random Forest, AutoGluon Tabular)
- End-to-end inference using the best model
- Business insights identifying which offers should be targeted, avoided, or optimized

This report follows the required structure of the Udacity Capstone project:
Domain Background → Problem Statement → Data Exploration → Methodology → Modeling → Evaluation → Conclusions.

---

2. Domain Background

Loyalty programs generate massive volumes of behavioral and transactional data. Starbucks' digital ecosystem is a strong example of data-driven marketing where customer engagement is influenced by:

- Offer types (BOGO, discount, informational)
- Customer demographics (age, income, gender, tenure)
- Event sequences (received → viewed → completed)
- Timing windows (offer validity, delayed actions, early engagement)

The dataset used for this project is a simulated dataset reflecting **realistic patterns**, not random synthetic noise. Key behavioral complexities include:

- Customers who complete offers without viewing (not influenced)
- Customers who make purchases without receiving any offer
- Customers who repeatedly ignore offers
- Offers that expire before being viewed or completed
- Multi-event pathways for the same customer

These characteristics mirror real-world marketing constraints, making this problem both challenging and relevant.

---

3. Problem Statement

Starbucks sends various promotional offers to app users, but customer behavior varies widely. Some users:

- Respond positively to certain offers
- Ignore or reject offers
- Make purchases regardless of promotions
- Prefer NOT receiving offers at all

The business goal is to **predict whether a customer will respond to a specific offer** based on their profile, behavior, and offer attributes.

Formal ML Problem

Given a customer–offer instance, predict:

Responded={1,if viewed AND completed within validity window0,otherwise\textbf{Responded} = \begin{cases} 1, & \text{if viewed AND completed within validity window} \\ 0, & \text{otherwise} \end{cases}Responded={1,0,if viewed AND completed within validity windowotherwise

This is a **supervised binary classification** task.

Correct predictions support key decisions:

- Who should receive a BOGO vs. discount offer
- Which customers reliably convert without offers
- Which segments should not receive promotions
- How to reduce wasted marketing spend

---

4. Datasets and Inputs

The project uses three JSON files from Starbucks:

1. portfolio.json — Offer Details

- Offer ID
- Offer type (bogo, discount, informational)
- Difficulty (minimum spend)
- Reward
- Duration (days)
- Channels (email, mobile, web, social)

2. profile.json — Customer Metadata

- Customer ID
- Age
- Gender
- Annual income
- Membership start date
- Missing ages (118) replaced with median/mode logic

3. transcript.json — Event Logs

Contains four event types:

- offer received
- offer viewed
- offer completed
- transaction

Each record includes:

- Person (customer ID)
- Time (hours from start)
- Value dict (offer_id, amount, reward)

This file captures the **temporal flow** of customer–offer interactions.

---

5. Data Exploration & Labeling

5.1 Cleaning Steps

- Flattened nested `value` JSON fields
- Standardized timestamps to offer windows
- Fixed missing values and demographic outliers
- Merged portfolio + profile + transcript

5.2 Creating Offer Instances

For every `offer received` event:

- Compute offer_start and offer_end
- Search for `offer viewed` within window
- Search for `offer completed` within window
- Label `responded = 1` ONLY if viewed → completed

This eliminates noise where customers completed without being influenced.

5.3 Key Visual Insights

(**Insert your charts here in Google Docs**)

Include:

- Overall response rate
- Response rate by offer type
- Response rate by age group
- Response rate by income group
- Heatmap of income × offer type
- Customer-specific response rate distribution

These plots are critical evidence supporting the insights & conclusions section.

6. Insights From EDA

Your EDA produced several actionable findings:

1. Overall Response Rate

~36% respond, ~64% do not.
This confirms **class imbalance**, important for model evaluation.

2. Discount offers outperform BOGO

Discount offers showed the highest engagement rates.

3. Age drives responsiveness

- Under 25: lowest response
- 35–65: strongest responders

4. Income strongly correlates with response

- <40K: very low response
- 70K–150K: highest response
- 150K+: moderate (likely self-purchasers)

5. Informational offers have near-zero response

This is critical:
Customers do not act on non-incentivized offers.

6. Customer-specific response distribution

- Large cluster of always-non-responders
- Some consistent responders

- Many mid-range, depending on offer type
  This segmentation directly supports targeting strategy.

These insights feed the final business recommendations.

---

7. Feature Engineering

Key engineered features:

Customer Demographics

- Age buckets
- Income groups
- Gender one-hot encoding
- Membership tenure in days

Offer Metadata

- Offer type one-hot encoded
- Difficulty & reward
- Channels encoded as binary features

Behavioral Features

- Prior views
- Prior transactions
- Prior completions
- Customer-specific response rate (optional)

Target Variable

`responded` (binary)

These features align with real-world marketing science literature.

---

8. Modeling Approach

Train/Test Split

70% training, 30% test.

Models Evaluated

1. **Logistic Regression (baseline)**
    - Interpretable
    - Establishes benchmark
    - Good for linear relationships
2. **Random Forest**
    - Captures non-linearities
    - Handles categorical + numeric well
3. **AutoGluon Tabular (best model)**
    - Ensembling of multiple models
    - Automatic hyperparameter optimization
    - Gradient boosting + deep learning + stacking
    - Achieved highest ROC-AUC

AutoGluon training was executed using:

```
predictor = TabularPredictor(
    label="responded",
    problem_type="binary",
    eval_metric="roc_auc"
).fit(
    train_data=train_ag,
    tuning_data=val_ag,
    presets="best_quality",
    time_limit=1800,
    use_bag_holdout=True
)
```

---

9. Model Performance

(Insert your table formatted in Google Docs)

Model	Accuracy	ROC-AUC
Logistic Regression	0.75	0.839
Random Forest	0.78	0.858
AutoGluon (Best Ensemble)	**0.80**	**0.882**

Interpretation

- ROC-AUC is the preferred metric due to class imbalance
- AutoGluon significantly outperforms traditional models
- Ensemble learning helps capture complex customer–offer dynamics

What the model learned

Top predictive features include:

- offer_type

- income group
- age group
- membership tenure
- difficulty
- reward

These align perfectly with the EDA insights — validating both the data and the approach.

---

10. Inference Example

A small inference test demonstrates how the model scores new customer–offer pairs:

```
sample = test_df.iloc[:5].copy()
preds = predictor.predict(sample)
probs = predictor.predict_proba(sample)[1]
```

The output shows:

- Predicted probability of response
- Final classification label

This mirrors real-world deployment where marketers score “next best offer” candidates.

---

11. Business Insights & Recommendations

1. Discount offers are the most effective

Consistently higher response rates across all major demographics.

2. Avoid sending informational offers

They produce near-zero conversion and waste customer attention.

3. Target higher-income customers (70K–150K)

These customers respond well to incentives.

4. Younger customers (<25) are poor responders

Shift budget away from this segment.

5. Middle-aged customers (35–65) are ideal offer recipients

They show the highest and most stable engagement.

6. Some customers will convert WITHOUT offers

Identifiable via:

- high transaction frequency
- low view rates
- moderate–high income
  These customers should be excluded from promotions to save cost.

7. Personalized offer assignment is feasible

The model achieves high ROC-AUC (0.882), making real-time targeting viable.

---

12. Conclusion

This project successfully implemented an end-to-end machine learning pipeline for predicting Starbucks promotional offer response. Key achievements include:

- Cleaned and labeled a complex event-driven dataset
- Performed rich EDA revealing strong demographic and behavioral drivers

- Engineered features aligned with marketing science
- Trained multiple ML models, with AutoGluon achieving best performance
- Extracted business insights that directly improve marketing efficiency

By identifying customers most likely (and least likely) to respond, Starbucks can:

- Reduce unnecessary promotional costs
- Improve customer satisfaction with relevant offers
- Optimize revenue impact of campaigns
- Personalize digital engagement at scale

The project demonstrates the value of combining rigorous data processing, modern ML modeling, and actionable business interpretation — all essential skills for real-world data science practice.