CPT_S 575 - DATA SCIENCE - PROJECT REPORT

STARBUCKS CUSTOMER REWARDS PROGRAM DATA ANALYSIS

Pallavi Arivukkarasu and Ramyasai Sanjita Bhavirisetty

Department of EECS

Washington State University

ABSTRACT

This project analyzes how each special offer made by Starbucks affects different segments of their customer base, clustered by age, gender, and income; classifies and measures transaction volumes across these variables and groups; and uses predictive modeling to determine which offers are likely to succeed or not, allowing for a simulation playground using the Starbucks Customer Rewards Program Open Data.

1. INTRODUCTION

Starbucks is a data-driven corporation that obtains a 360-degree consumer picture by analyzing datasets containing customer data, special offers, and transactions. The research and search for offerings that successfully engage the company's existing consumers and attract new ones is the foundation of this project. Investing in a thorough marketing strategy requires the approval of various stakeholders and money and effort. As a result, a predictive model that can determine whether launching a specific offer for a particular target group is worthwhile should be a strategic asset for any organization.

Starbucks can better target its offers to clients who are more likely to take advantage of them from a business standpoint. They will maximize income from offer take-up to the correct clients while also saving money on marketing and promotional costs by sending more relevant offers. Customers will have a better experience with Starbucks and the application since they will receive individualized, relevant offers based on their preferences. This paper shows results on cleaning the dataset, analyzing various features, and finally using modeling concepts to show predictions on whether an offer is successful for a particular person belonging to a specified age group, gender, and income. We also compare the metrics between different models, and thus, it shows that SVC is the efficient model. These results help companies take the extra step and determine the best method to generate more revenue.[1], [2], [3], [4]

2. DATASET AND SOFTWARE USED

Dataset:

Name: Starbucks app - customer reward program data

Source: Kaggle [7]

Description:

- In the portfolio.json file, we have the data variables:

  reward (int), difficulty (int), duration (int) as continuous values channels (list of strings), offertype (str), and id (str) as categorical data.

- In the profile.json file, we have the data variables:

  age (int), became member_on (int), income (float) as continuous values, gender (str), id (str) as categorical data.

- In the transcript.json file:

  time (int) as a continuous value, person (str), event (str), and value (dict of strings) as categorical data. [7]

Investigation on information in the dataset - income field

The personal data, like the income, age, and membership, are taken from the customers who have applied for the Starbucks Rewards Visa credit card - Chase Bank (Starbucks Rewards Credit Card | Chase.com).

Background information: Use of the Starbucks credit card:

Benefits of card: Receive 4500 stars in the app if the user makes a total purchase of $500 in the first three months.

Conclusion: This dataset focuses on the customers who buy Starbucks coffee regularly.

Software/Libraries Used:

Jupyter Notebook, Pandas/Numpy, Matplotlib, Seaborn, Scikit

3. IMPLEMENTATION AND ANALYSIS

3.1. Data Exploration and Pre-processing

Starbucks sends out an offer to app users once every few days, according to the background information on the mobile app. Some customers may not receive any deals during certain weeks, and not all users will receive the

same offer. The properties of the given data, including the data type and the number of unique values, have been appropriately checked. After considering the observations, we determined a method for preparing the data for modeling.

Before preparing the data for the model, we went over the goal once again. After completing some fundamental data analysis, we had to reevaluate how to clean and prepare the data for the models we planned to develop. We must first define an "effective" offer to establish the primary drivers of a compelling offer within the Starbucks app. As a result, we investigated the datasets and their potential interactions.

We investigated the types of events that could occur for each offer type. There are four types of events: completed offer, received the offer, viewed offer, and transaction. However, because offer ids are not stored in the transcript event data, our data shows that we do not have any offer ids linked with transactions. As a result, the initial goal of data preparation is to establish a system for assigning offer ids to specific transactions.

Furthermore, we know that when BOGO and discount deals are completed, an offer completed event is triggered. On the other hand, this event is not related to informational offerings. As a result, we additionally outline the following approach for defining an effective offer:

- An effective BOGO and discount offer would be characterized if the following events were recorded in the correct order in time:

  offer received -> offer viewed -> transaction -> offer completed

- Meanwhile, for an informational offer, since there offer completed event associated with it, we will have to define transactions as a conversion to effective offer:

  offer received -> offer viewed -> transaction

Assigning transaction ids to offers:

Following the definition of the technique above, we must now investigate strategies for assigning offer ids to individual transactions. One of the things to think about is defining the following primary client groups:

1. Effective offers: People who are influenced and effectively convert:

offer received-> offer viewed -> transaction -> offer completed (*BOGO/discount offers*)

offer received -> offer viewed -> transaction (*informational offers - must be within the validity period of an offer*)

2. Ineffective offers: People who got and viewed an offer but did not convert:

offer received -> offer viewed

3. People who buy/complete deals despite not being aware of them:

offer received -> transaction -> offer completed -> offer viewed

transaction -> offer received -> offer completed -> offer viewed

offer received -> transaction -> offer viewed -> offer completed

offer received -> transaction (*informational offers*)

offer received -> transaction -> offer viewed (*informational offers*)

4. Those who received offers but did not act:

offer received

Meanwhile, before we can tag the effective offers column for informational offers, we must examine another factor: the integrity of the offer. The conversion event is a transaction, not an offer completion event.

The duration of the offer might be considered the duration of the influence in the case of informational offers. As a result, we can assume that an offer is only effective if used within the offer's time limit.

Meanwhile, we may presume that if a conversion/offer finished event occurs for BOGO and discount offers, it would occur during the timeframe of the offer, as it makes no sense for an offer to be completed after its validity period has expired.

We have label encoded the categorical variables -

```
from sklearn import preprocessing
encoder = preprocessing.LabelEncoder()
modeldata['offer_type'] = encoder.fit_transform(modeldata['offer_type'])
modeldata['gender'] = encoder.fit_transform(modeldata['gender'])
modeldata['status'] = encoder.fit_transform(modeldata['status'])
```

Figure 1: Label Encoding the Categorical Variables

Before:

```
modeldata
```

	offer_type	reward	difficulty	email	mobile	social	web	gender	age	income	status
0	bogo	5	5	1	1	1	1	F	24	60000.0	success
1	discount	3	7	1	1	1	1	F	24	60000.0	success
2	bogo	5	5	1	1	0	1	F	24	60000.0	success
3	bogo	5	5	1	1	1	1	F	55	74000.0	success
4	discount	2	10	1	1	1	1	F	55	74000.0	success

Figure 2: Before Cleaning the Dataset

After:

```
modeldata
```

	offer_type	reward	difficulty	email	mobile	social	web	gender	age	income	status
0	0	5	5	1	1	1	1	0	24	60000.0	1
1	1	3	7	1	1	1	1	0	24	60000.0	1
2	0	5	5	1	1	0	1	0	24	60000.0	1
3	0	5	5	1	1	1	1	0	55	74000.0	1
4	1	2	10	1	1	1	1	0	55	74000.0	1

Figure 3: After Cleaning the Dataset

3.2. Exploratory Data Analysis (EDA)

The fields like income, age, gender, and membership status are being compared and analyzed with the help of graphs and plots using the seaborn library.

The offer information like the type of offer, duration, and difficulty in achieving the offer are analyzed using the seaborn library's distribution plots.

Key Insights of EDA

1. Females spend more money at Starbucks on average than males.
2. Users who do not complete their membership profile spend less, implying that they would not benefit from "handcrafted drink" incentives.
3. Offers that include a discount are more likely to be fulfilled than buy-one-get-one deals.
4. Include social media in the distribution strategy for optimum completion. Except for the fact that the most successful and second-to-least successful offers were almost identical in every way, the most successful one was spread via social media while the other was not.

5. We used machine learning to prove that social media impacted whether or not an offer was completed.

6. Other crucial considerations included the duration of the offer, the age of the user, and the value of the incentive. Email, the length of time the user had been a member, and gender were unimportant variables.

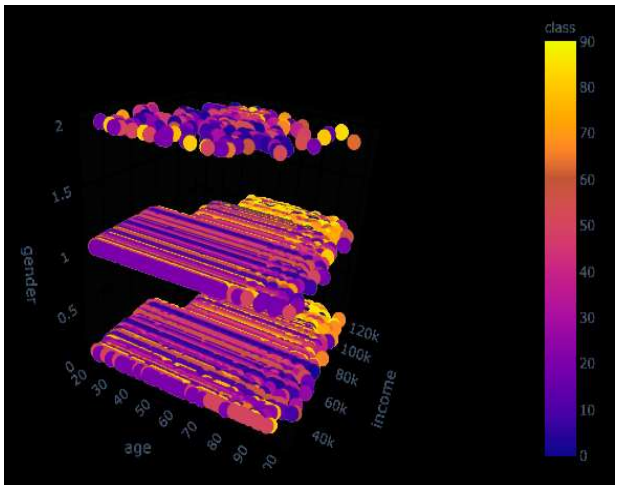

Figure 4: Gender Vs. Age Vs. Income Plot

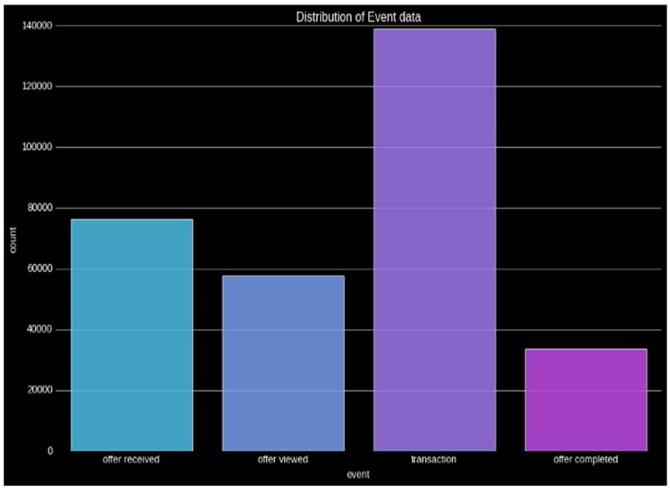

Figure 5: Distribution of Event Data

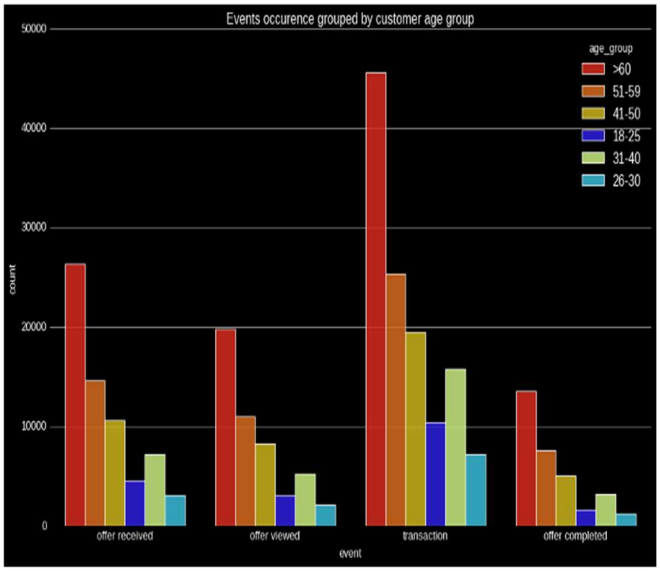

Figure 6: Event occurrence grouped by customer age groups

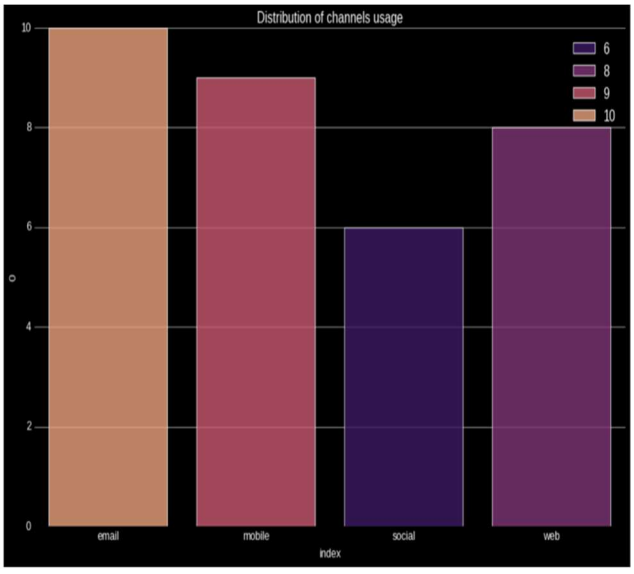

Figure 7: Distribution of channel usage

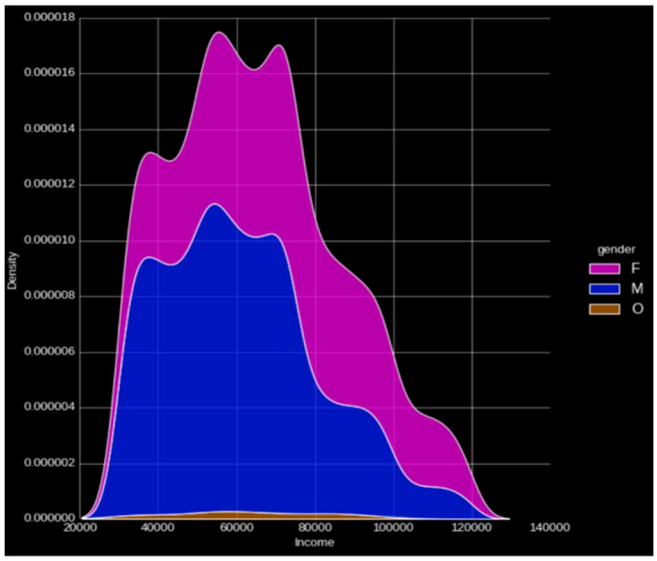

Figure 8: Income Vs. Density according to gender

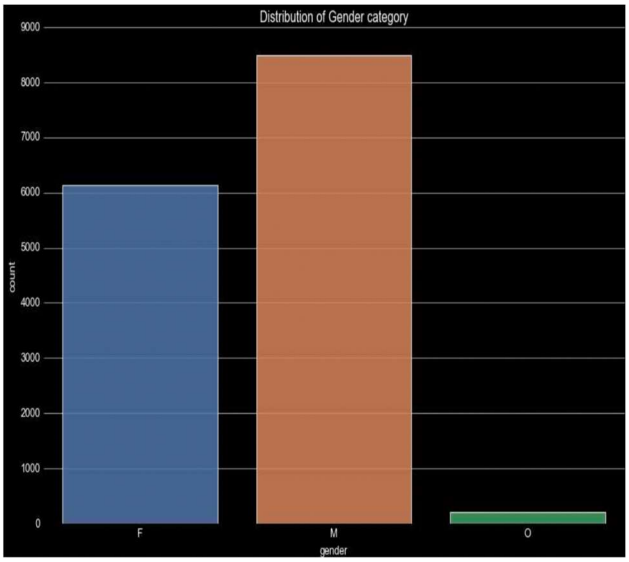

Figure 9: Distribution of Gender Category

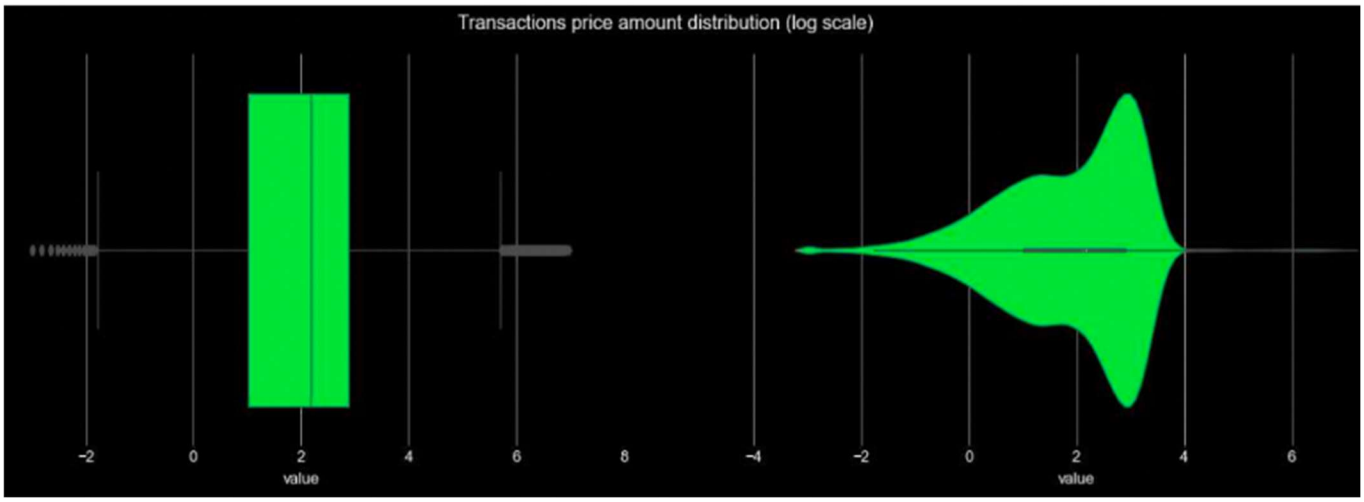

Figure 10: Transactions price amount distribution

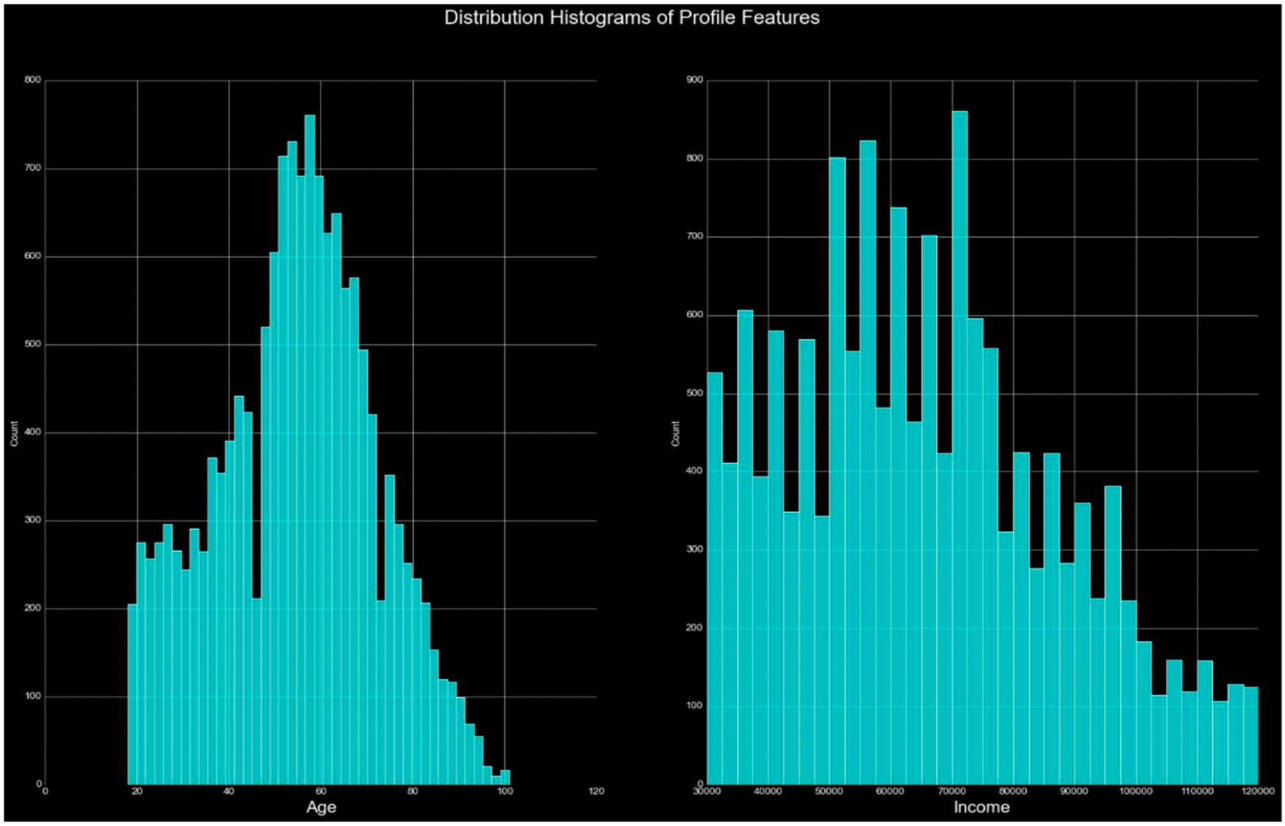

Figure 11: Histogram of Profile Features Distribution

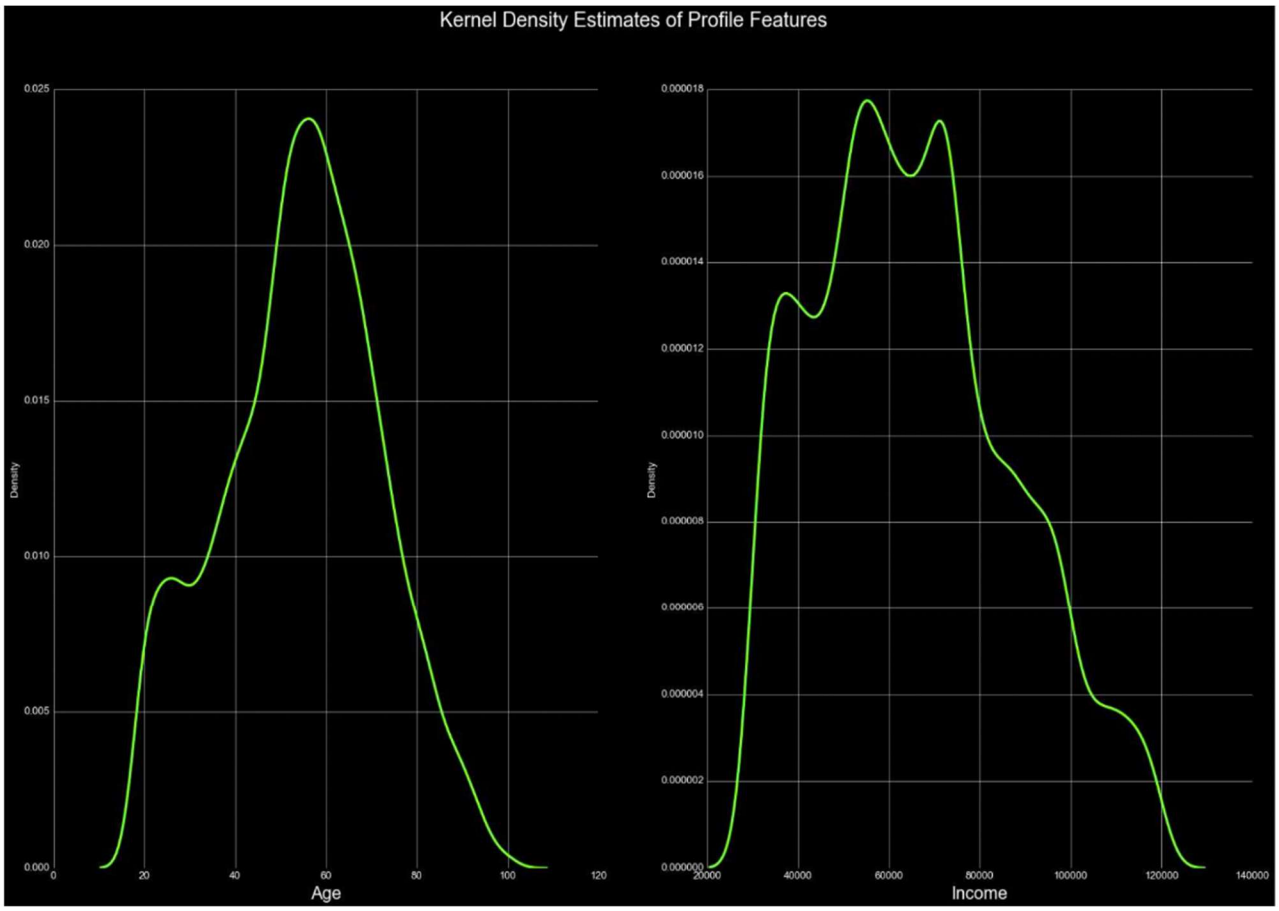

Figure 12: Kernel Density Estimates of Profile Features

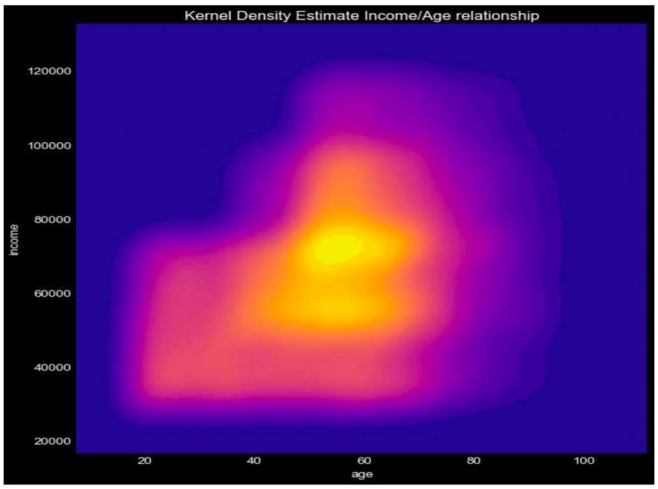

Figure 13: Kernel Density Estimate of Income/Age Relationship

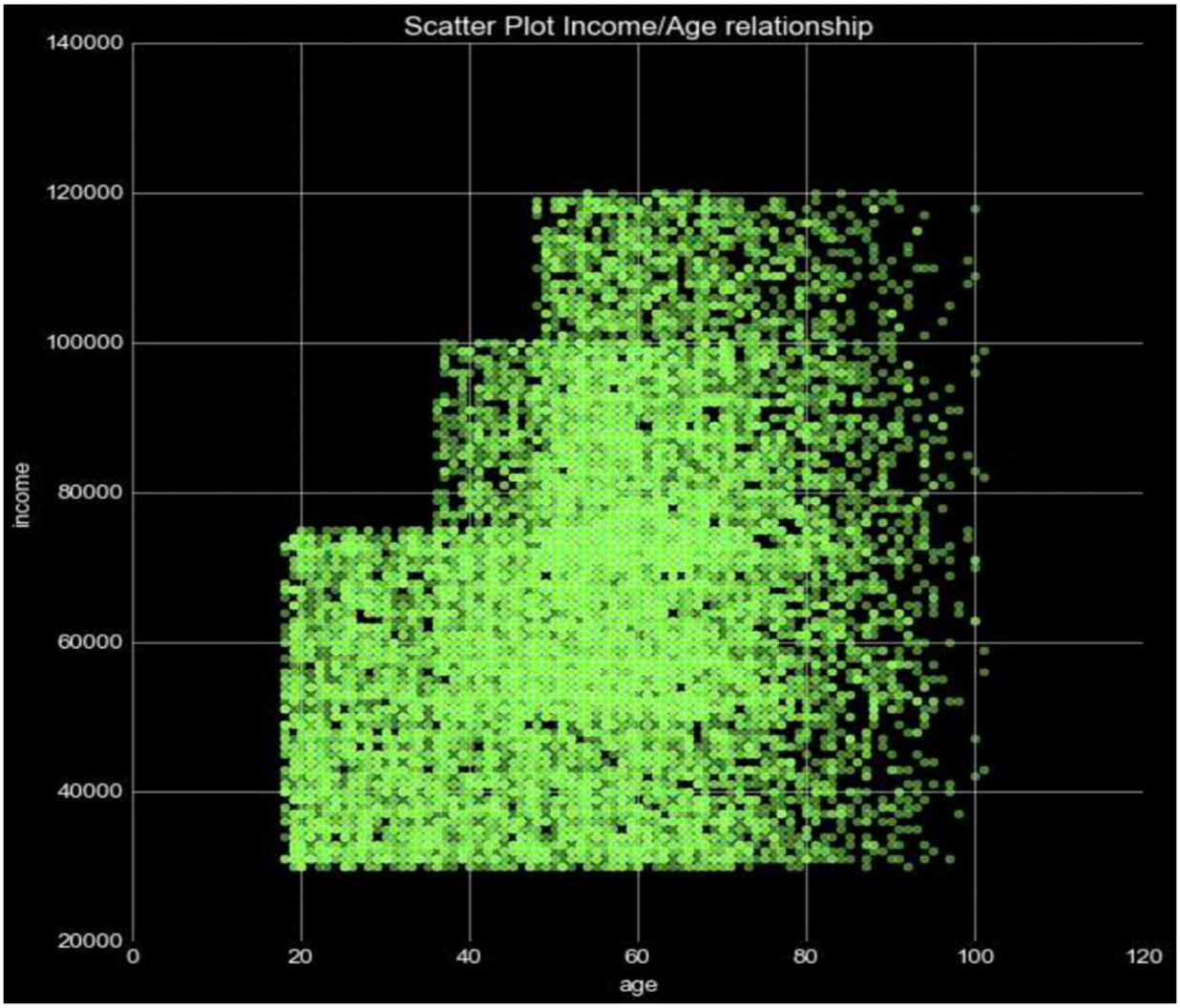

Figure 14: Scatterplot of Age/Income Relationship

3.3. Data Modelling

During cleaning or other data preprocessing steps, invalid rows are deleted, resulting in a massive imbalance between classes.

```
modeldata['status'].value_counts()/len(modeldata)
```

```
1    0.9599
0    0.0401
Name: status, dtype: float64
```

The cleaned dataset is an unbalanced class. For better prediction accuracy, the training set needs to undergo oversampling.

```
from imblearn.over_sampling import SMOTE
sm = SMOTE(sampling_strategy='minority', random_state=42)
X_train_oversampled, y_train_oversampled = sm.fit_resample(X_train,y_train)
X_test_oversampled, y_test_oversampled = sm.fit_resample(X_test,y_test)
```

Recalling our goal, we are developing two models to forecast the success of each sort of offer based on offer features and user demographics.

- K-Nearest Neighbors
- Support Vector Machine

We followed the standard modeling procedures:

- Define the variables for the target and feature
- Divide data for training and testing
- Use the feature scaler

As we develop further, systems will make recommendations based on real-time analysis of a customer's skin tone, facial traits, and emotions to tailor what to recommend or avoid. Top conglomerates use GPS technology and company applications to trigger relevant in-app offers when customers approach a store. Predictive analytics is a branch of statistics concerned with gathering data and applying it to forecast trends and behavior patterns. Predictive web analytics improves the calculation of statistical probabilities of future events on the internet. Data modeling, machine learning, artificial intelligence, deep learning algorithms, and data mining are statistical techniques used in predictive analytics.

4. EVALUATION AND RESULTS:

The metrics utilized in this project are the same ones used to calculate any classification model's performance. The number of test records wrongly and adequately predicted by a classification model is used to evaluate its performance.

TP, TN, FP, and FN are the four results based on their categorization. Each utilized model will be found in the project's code on GitHub (link at the bottom of the article).

The following are the metrics developed as a result of these findings:

- Accuracy
- Precision
- Recall
- F1-score

The most common categorization evaluation parameter is accuracy. In well-balanced datasets, it works well. The percentage of cases accurately classified by the model is called accuracy. However, accuracy might be deceiving. It can make a bad model appear to be good. When classes are uneven, the accuracy metric does not work correctly.

Precision, recall, and F1 is substantially superior for situations with unbalanced classes. These measures provide a more accurate picture of the model's quality. Precision indicates how well a machine learning model performs in categorization tasks. This measure would answer the following query in our project: recall informs us about the quantity that the machine learning model is capable of identifying. This statistic would answer the following question in our project: The F1-Score is a single score that combines the precision and recall metrics. This idealogy is helpful since it makes comparing the mixed results easier.

SVC shows balanced evaluation metrics making it a reasonably good model.

Table 1: Metrics Table

	Precision	**Recall**	**F1-score**	**Accuracy**
KNN (K=3)	90	67	77	73
SVC	80	78	79	79

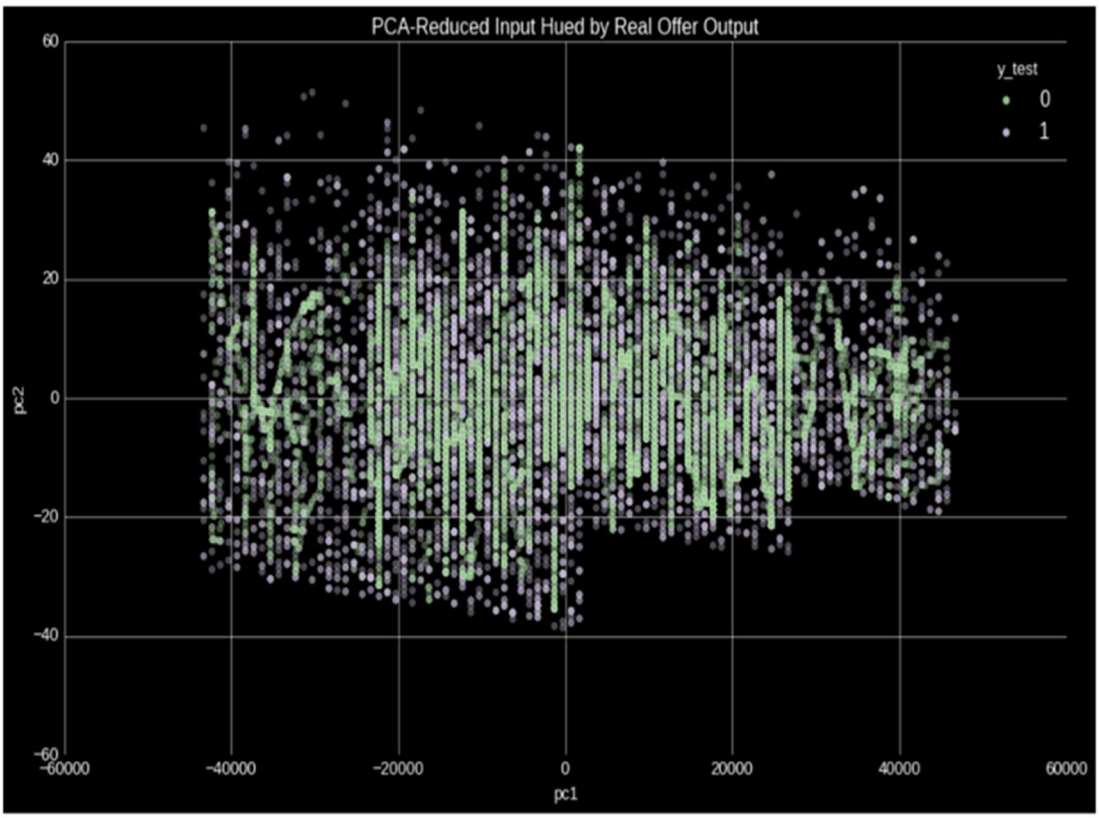

Figure 15: PCA-Reduced Input Hued by Real Offer Output

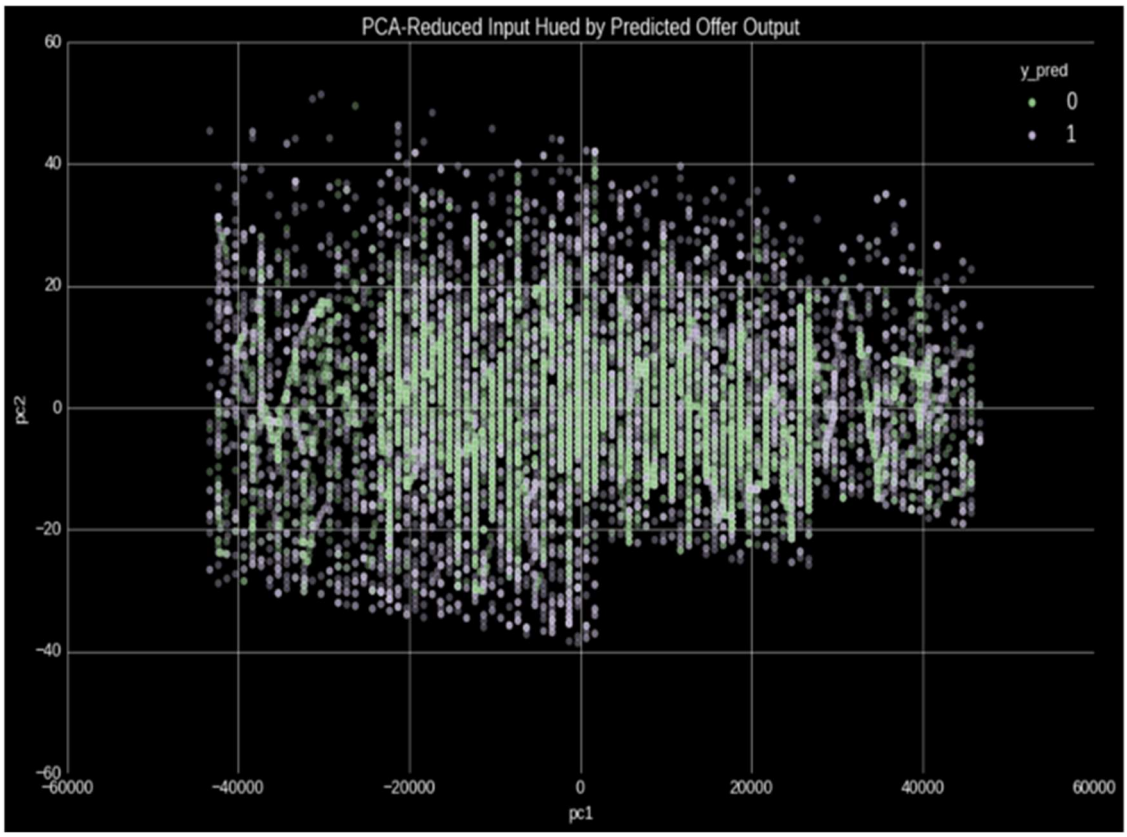

Figure 16: PCA-Reduced Input Hued by Predicted Offer

5. MODEL USE CASE:

Let us say our Starbucks manager assigned us the responsibility of creating a cost-effective offer for middle-aged women, as well as a report on the offer's expectations before launching it. First, it is helpful to understand what has previously worked.

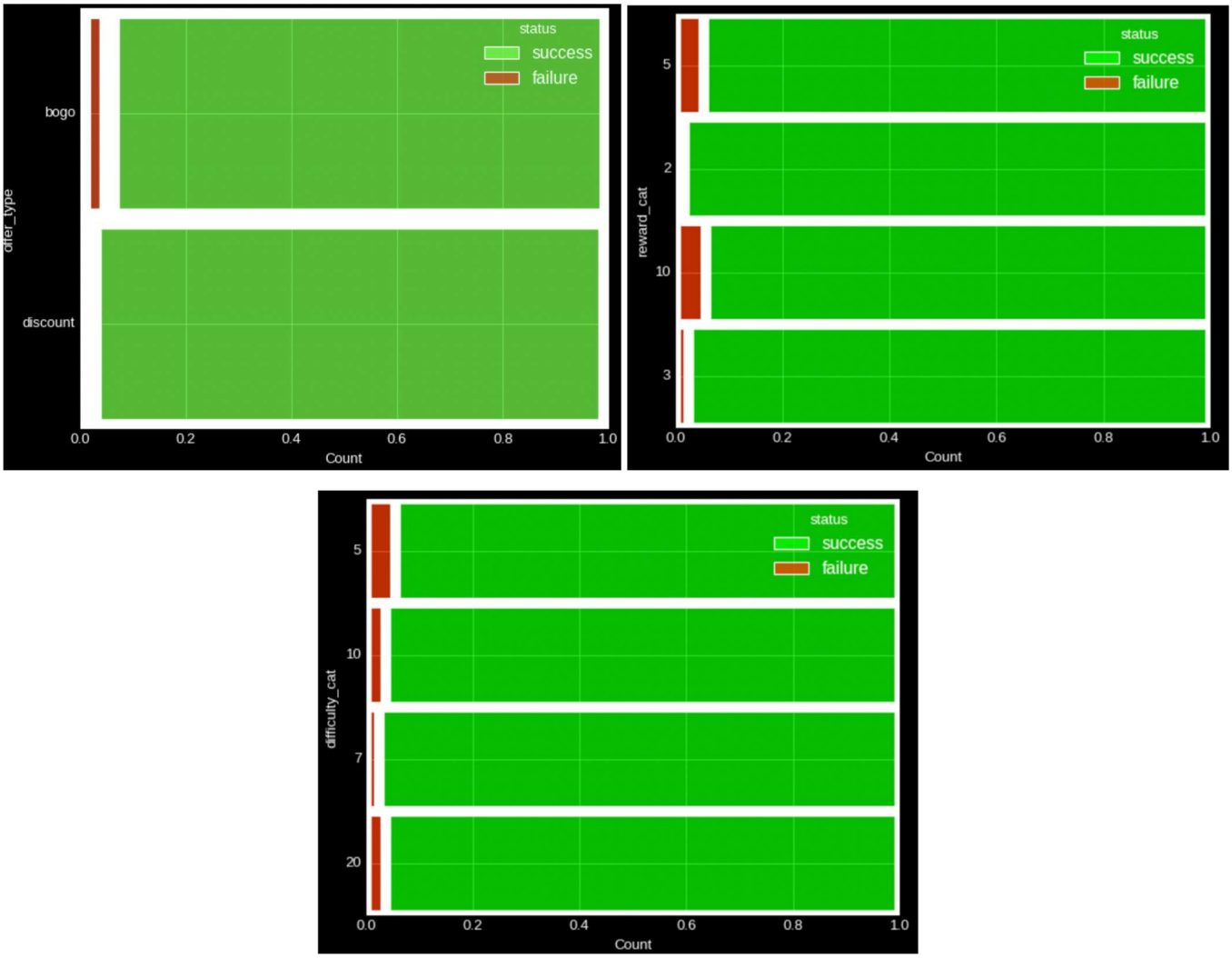

Figure 17: Predicting offer success and failure

Everything appears to convert well with this group, but a discount offer type with a reward of 2 and a difficulty of 7 would be an excellent offer to maximize probabilities.

```
In [588…  maw_predict = svm.predict(middle_aged_woman_final)

In [593…  middle_aged_woman_final['predicted'] = maw_predict

In [599…  middle_aged_woman_final['predicted'].value_counts()

Out[599…  0    989
          1     11
          Name: predicted, dtype: int64
```

When given these offer parameters, the categorization model projected that this offer would be accepted by 989 of 1000 middle-aged women and rejected by 11. As a result, this is a decent deal that would most likely produce similar results in the actual world. This model can be used to simulate and develop relevant offers that are likely to succeed in real life, and it can be used for various forms of testing and fine-tuning.

An application tailored to the user's needs: Rather than general offer simulation-testing, the ideal application for this model would probably be for user-specific prediction, by optimizing specific offers specifically for that user whom the algorithm knows has a high probability of success. If we go deep and work directly on the project, several multi-lateral implementations are worth investigating.

6. RELATED WORK

Predictive learning on subject lines for large-scale email marketing, as well as an assessment of Starbucks' company structure and the future consequences of its current business tactics, are two examples of relevant work in this sector.

The researchers look at a comprehensive overview of the specialty coffee business and Starbucks' involvement. Starbucks operates in a fast-growing market with a solid overall position. Starbucks has successfully navigated organizational and managerial challenges, serving as a great model for worldwide firms. The researcher will determine the company's strong and weak business strategies by looking at strategic imperatives such as expanding internationally and comprehending the international backdrop. Following that, the researcher will make strategic and execution recommendations for how Starbucks might expand internationally.

Email marketing is used to sell billions of dollars in services and commodities. Consumers frequently open emails based on the subject line, so subject lines significantly impact available rates. The email subject lines are traditionally generated based on the human editors' best judgment. We present an approach for assisting editors by forecasting subject line open rates using information from previous subject lines. Based on keywords, prior subject line performance, and grammar, the approach extracts various characteristics from subject lines.

Furthermore, we employ an iterative process called Attribution Scoring to assess the importance of individual subject-line keywords to overall open rates and use this information to enhance forecasts. A random forest-based model is built to incorporate these features to forecast performance. A dataset of over a hundred thousand different subject lines with several billions of impressions was used to train and test. For both new and old subject lines, the suggested technique improves prediction accuracy over the baselines. [5], [6]

7. CONCLUSION AND FUTURE WORK

The transcript dataset's sequential, event-based form posed the most difficulty; however, this is prevalent in in-app data and many digital analytics datasets that incorporate transactional data, such as the web. As a result, we found this challenge to be satisfying.

Given more variety of data on the types of products available and product costs, this study may have easily been expanded. It might be developed into a more in-depth customer segmentation exercise to identify customers who favor specific offer kinds or products and anticipate the relevant price band for such offers.

However, data preparation and cleaning would have been far more complex, so we are grateful that the data provided was more constrained and regulated in this case. Within this scope, we are pleased with the performance of the models.

8. CODE

https://github.com/brsanjita/Analysis-of-Starbucks-Customer-Rewards-Program-Data.git

REFERENCES:

1. Syuen Loh, "Using Starbucks app user data to predict effective offers," Towards Data Science, 2018
2. Jordi Lucas, "Starbucks Capstone Challenge: Offer Analysis and Success Prediction," Medium, 2018
3. "Analyzing and Predicting Starbucks App Offers," FoodNewsNews
4. Zixing Wang, "Use Machine Learning to Individualize Offers for Starbucks Customers," LinkedIn, 2020
5. Lauren Roby, "An Analysis of Starbucks as a Company and an International Business"
6. Raju Balakrishnan, Rajesh Parekh, "Learning to Predict Subject-Line Opens for Large-Scale Email Marketing"
7. https://www.kaggle.com/blacktile/starbucks-app-customer-reward-program-data