Section 1: Project Definition

Overview:
This project performs data analysis on Starbucks promotional dataset in order to discover which customers will be receptive to a promotion. The purpose is to optimize Starbucks promotional strategy to encourage customers to spend more. Customer receptiveness was gauged based on a combination of user and promotional data.

Dataset:
There are three files:

1. Portfolio.json - table of promotions
2. Profile.json - User data for each customer
3. Transcript.json - Transaction records

Schemas:
portfolio.json

- id (string) - offer id
- offer_type (string) - type of offer ie BOGO, discount, informational
- difficulty (int) - minimum required spend to complete an offer
- reward (int) - reward given for completing an offer
- duration (int) - time for offer to be open, in days
- channels (list of strings)

profile.json

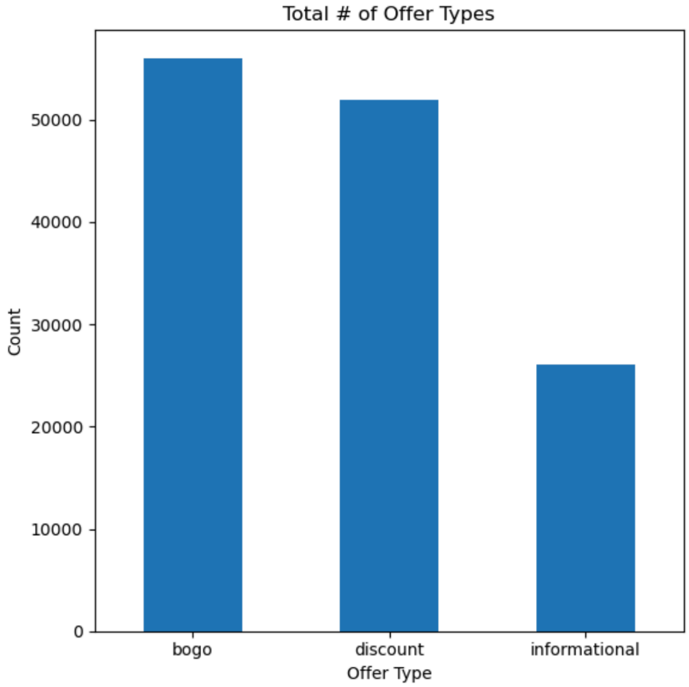
- age (int) - age of the customer
- became_member_on (int) - date when customer created an app account
- gender (str) - gender of the customer (note some entries contain 'O' for other rather than M or F)
- id (str) - customer id
- income (float) - customer's income

transcript.json

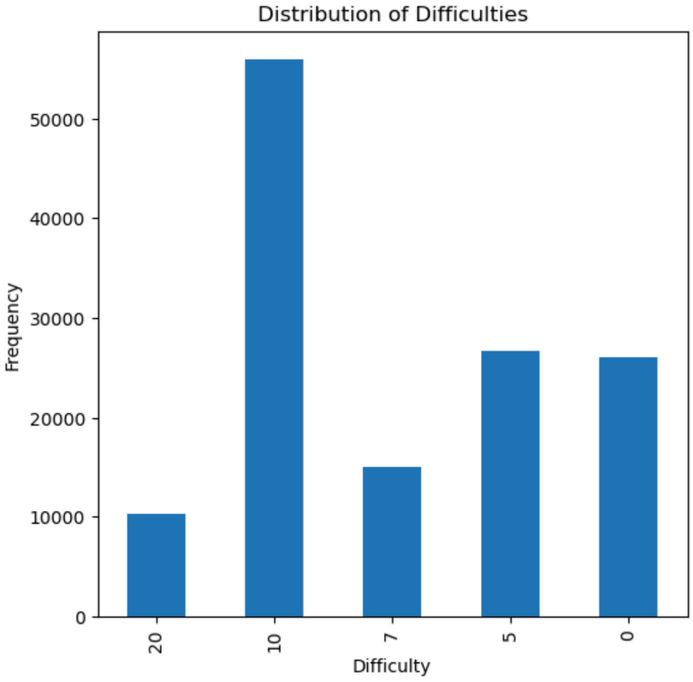
- event (str) - record description (ie transaction, offer received, offer viewed, etc.)
- person (str) - customer id
- time (int) - time in hours since start of test. The data begins at time t=0
- value - (dict of strings) - either an offer id or transaction amount depending on the record

All three datasets were eventually merged and passed into a random forest pipeline.

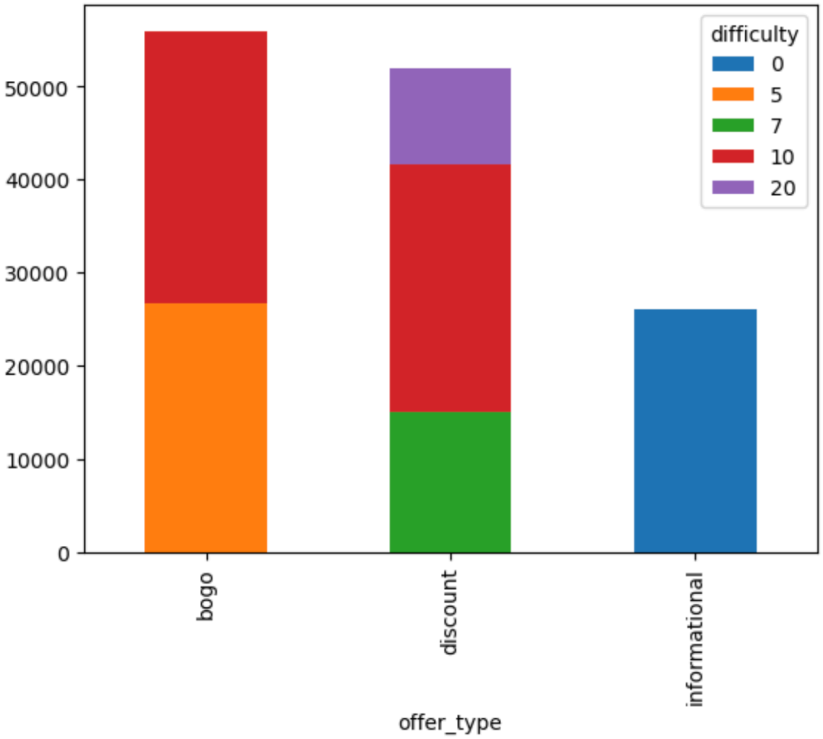
Data Visualization



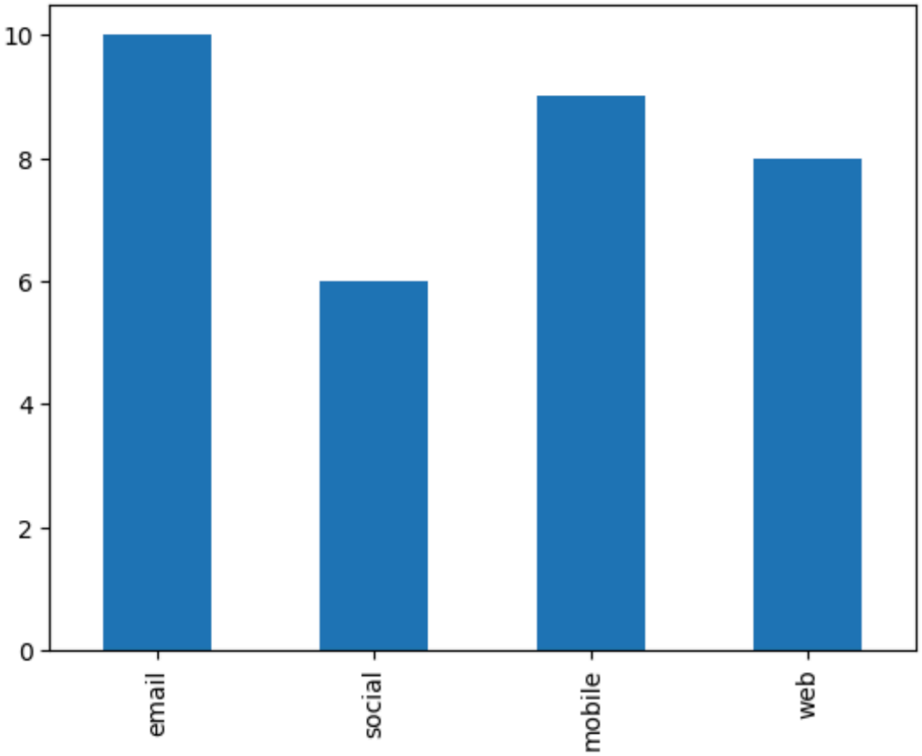
It seems the least offered type of discount is informational, followed by discount and bogo. There appears to be a preference towards discount strategies.



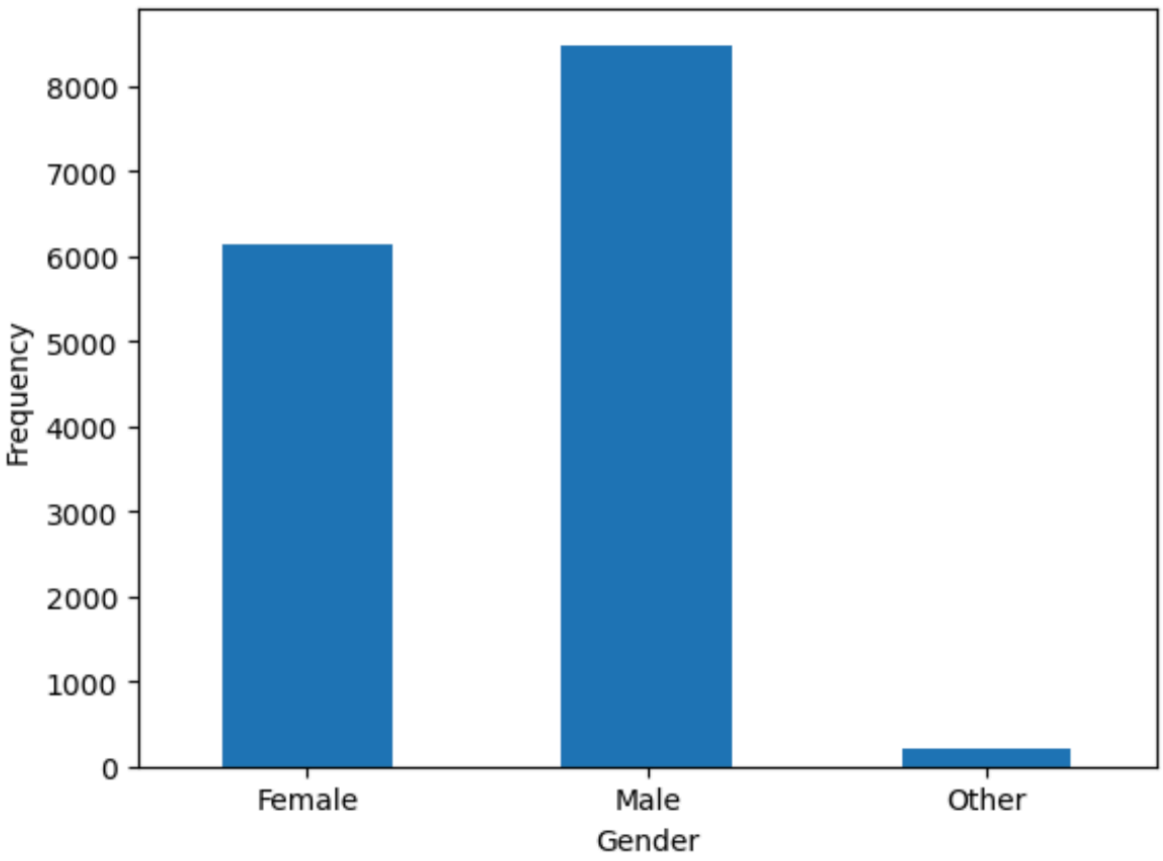
According to the chart, promotions with higher minimum purchase thresholds are more frequent.



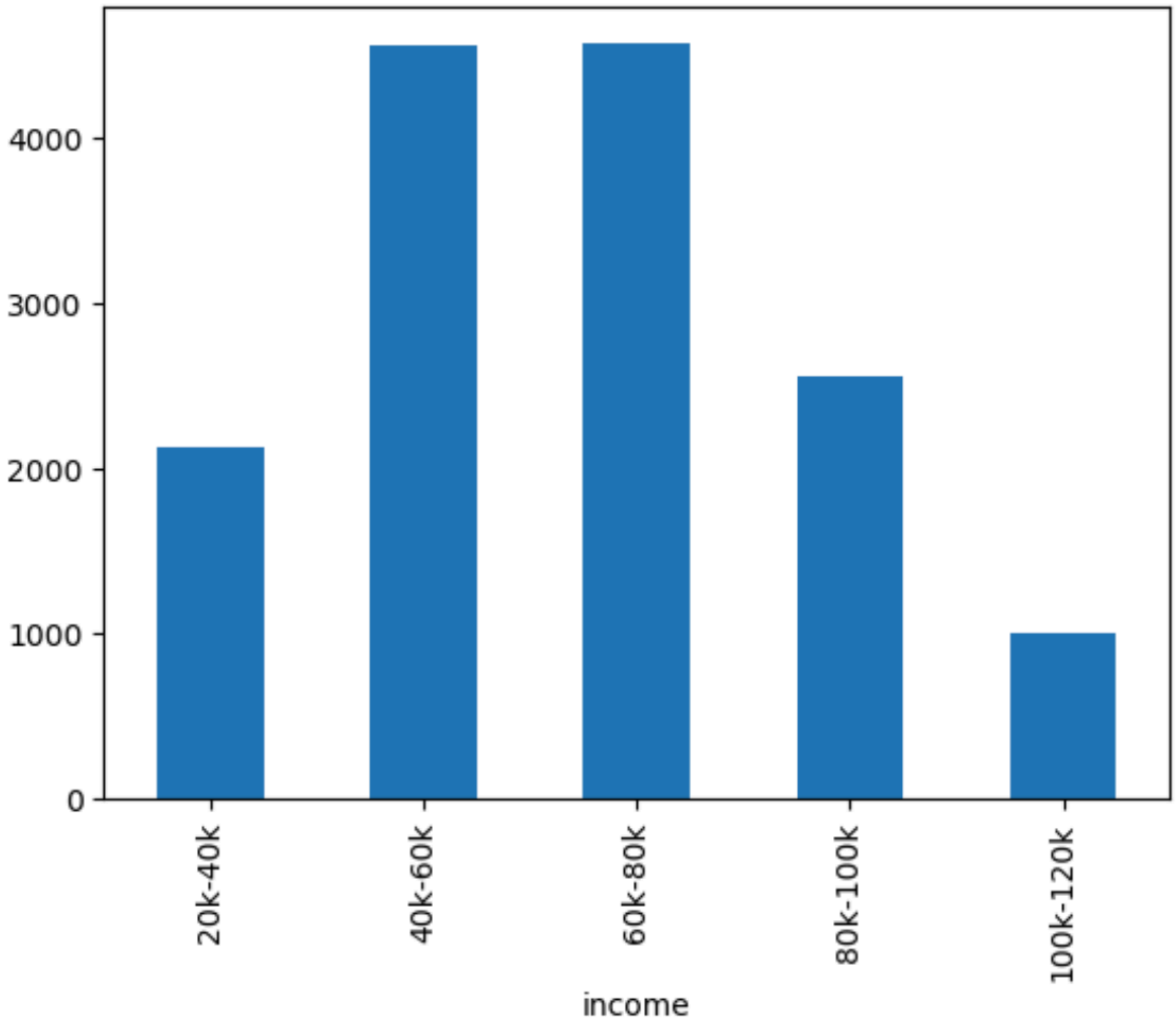
Combining these two categories reveals that between the ‘bogo’ and ‘discount’ type promotions, ‘bogo’ is on average less difficult to redeem than traditional discounts.



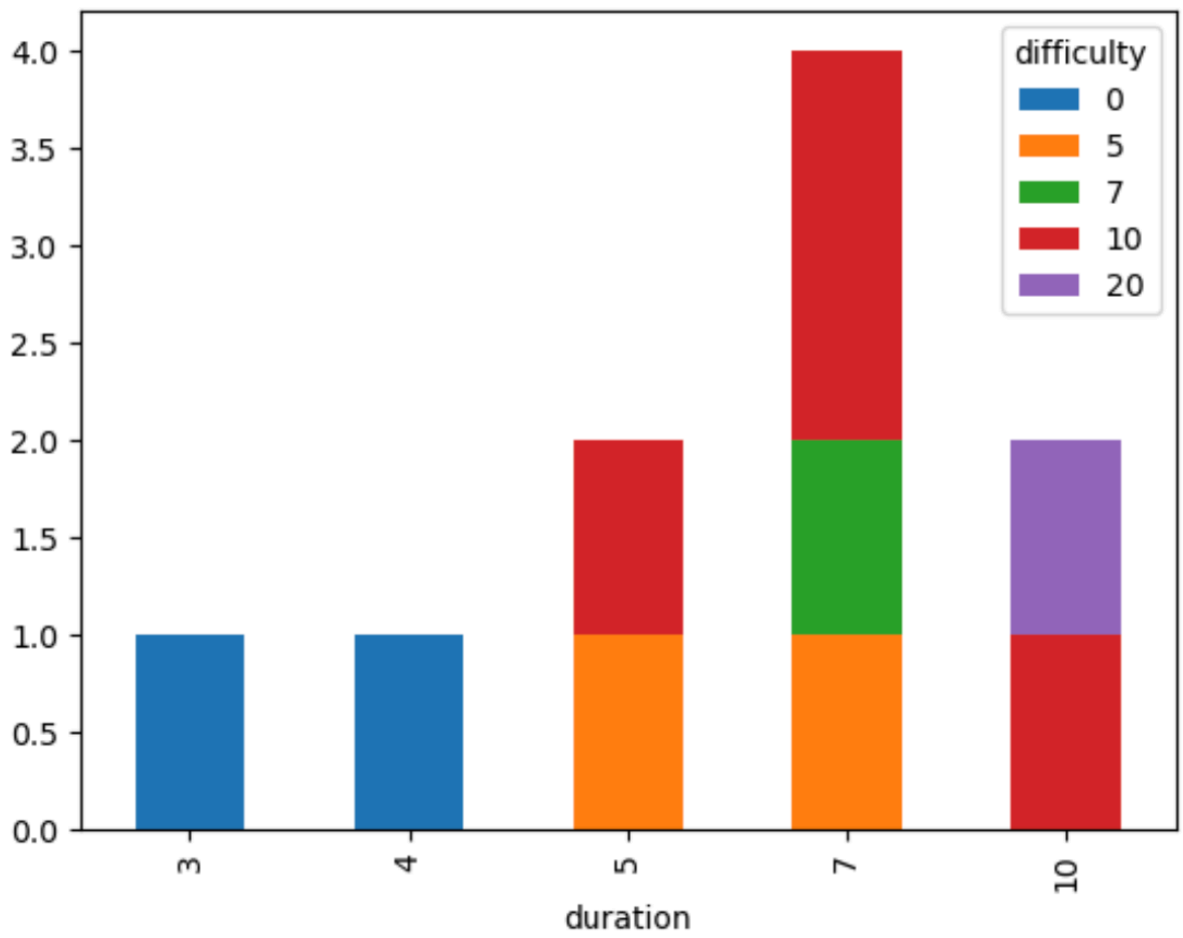
Despite the small sample size, it seems that all four channels are evenly represented.



There appear to be significantly more male customers represented than female.



Most customers appear to be within the 40-80k range. This may be attributed to the population mean income at the time the data was collected.



Shorter-dated promotions are strictly informational, while Starbucks gives customers between 5-10 days to redeem all other types.

Section 3: Methodology

I first started by defining the problem statement: Predict which customers are going to positively react to the ads.

First, the data was cleaned of problematic nulls. For example, there were some user accounts who listed users as 118 years old. The data was then split into training and test sets and trained on a ML pipeline using a random forest classifier. Random Forest was selected due to its ease of use and good performance when it comes to capturing non-linear relationships between discrete features. Precision and recall were used as the main performance metrics.

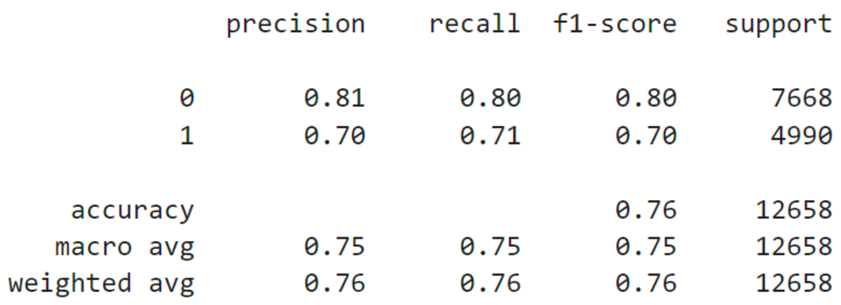
Cleaning:

I noticed that there were three main values for gender: Female, Male, and Other. There were also some marked as Null. Further inspection revealed the Null rows also had invalid values for income and age, so they were removed.

The value column was a dict, which forced me to separate out the individual keys into columns on their own. I extracted the offer_id, and noticed that some keys were ‘offer id’ without the underscore, so I made sure to include those in the final column.

I then defined what would be considered a success or failure. If a promotion is successful, the customer will have received, read, and redeemed the promotion. Otherwise, it would not be possible to determine that the customers actions were directly influenced by the ad.

Conclusion:



```
              precision    recall  f1-score   support

           0       0.81      0.80      0.80      7668
           1       0.70      0.71      0.70      4990

    accuracy                           0.76     12658
   macro avg       0.75      0.75      0.75     12658
weighted avg       0.76      0.76      0.76     12658
```

The Random Forest classifier shows moderate performance with a weighted precision and recall of 0.76, so on average the model accurately identifies which customers are receptive to ads ~76% of the time. Given this performance, Random Forest would be an appropriate classifier for fulfilling the problem statement.

Further tuning:

Considering that informational promotions cannot be accurately measured since they are not redeemable, it may be appropriate to remove this data entirely. It would also make sense to use gridsearchCV to test a combination of other ML algorithms and labels.