Domain Background

The coffee retail giant company Starbucks owns a consumer mobile application, which sent out rewards and offers for customers occasionally. The business team in Starbucks want to find out the pattern of a customer that takes the offer, providing us the data sets that include customer profiles, transaction history and offer type data.

This project is designed to analyze the application data and designed and develop a machine learning model to predict the probability that a particular group of customers will take the offer. The model should test properly by using the testing data after deployed the machine learning model.

Datasets and Inputs

The data is contained in three files:

- portfolio.json - containing offer ids and meta data about each offer (duration, type, etc.)
- profile.json - demographic data for each customer
- transcript.json - records for transactions, offers received, offers viewed, and offers completed

Here is the schema and explanation of each variable in the files:

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

The data is coming from Starbucks application log.

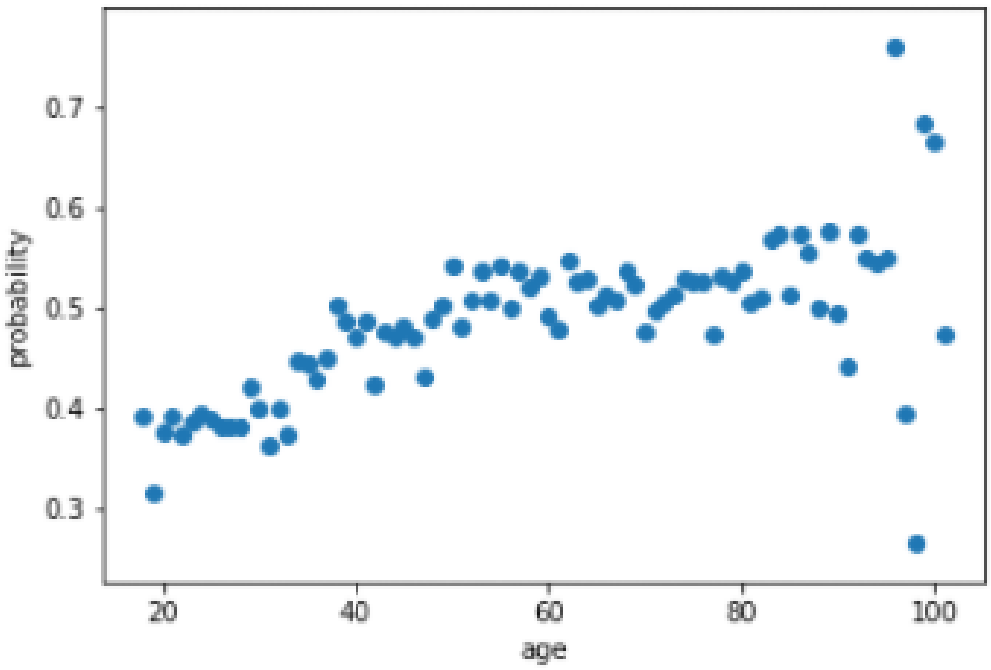
Problem Statement

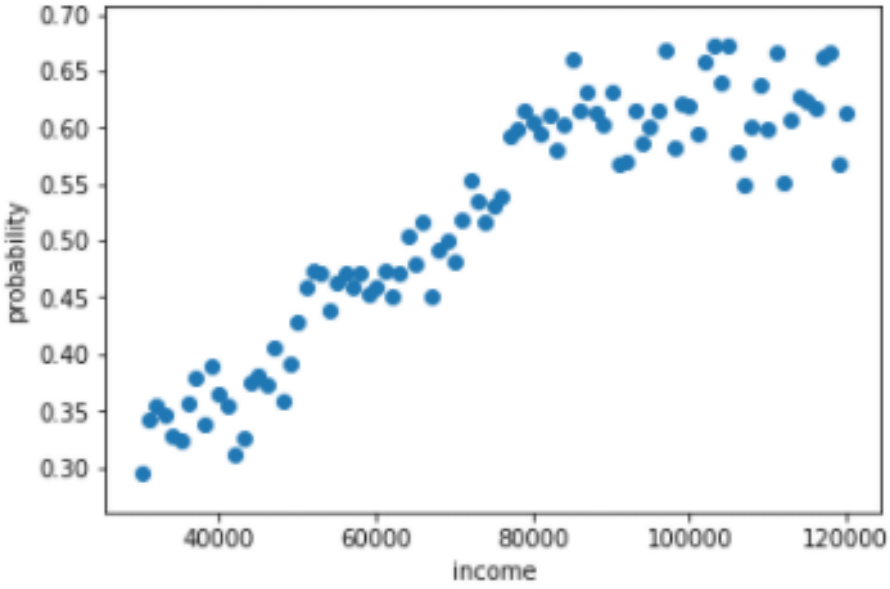
The business problem that we can tell from the data set is different group of customers responded differently to the offers. See below data:

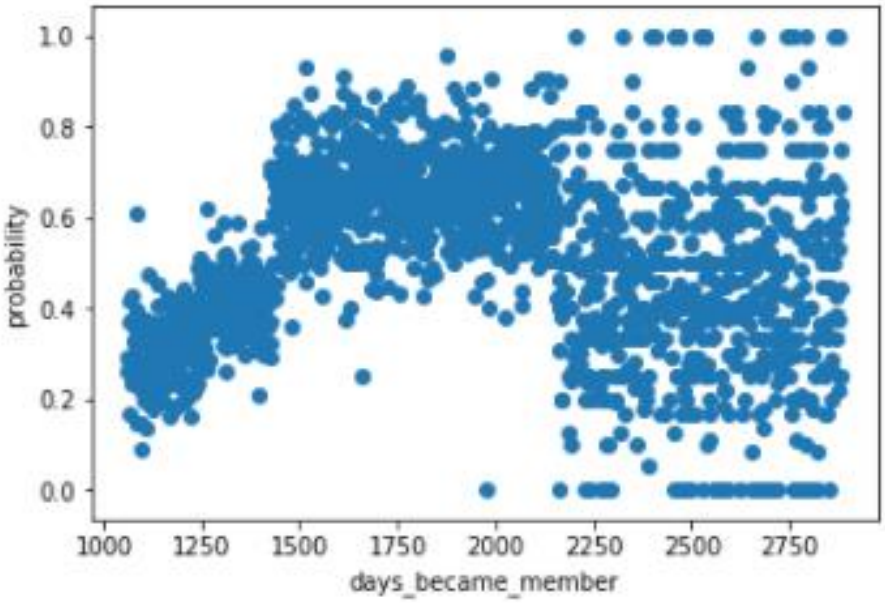
:

	age	became_member_on	gender	person	income	probability	days_became_member
1	55	20170715	F	0610b486422d4921ae7d2bf64640c50b	112000.0	0.500000	1438.0
3	75	20170509	F	78afa995795e4d85b5d9ceeca43f5fef	100000.0	0.750000	1505.0
5	68	20180426	M	e2127556f4f64592b11af22de27a7932	70000.0	0.500000	1153.0
8	65	20180209	M	389bc3fa690240e798340f5a15918d5c	53000.0	0.833333	1229.0
12	58	20171111	M	2eeac8d8feae4a8cad5a6af0499a211d	51000.0	0.333333	1319.0
13	61	20170911	F	aa4862eba776480b8bb9c68455b8c2e1	57000.0	0.600000	1380.0
14	26	20140213	M	e12aeaf2d47d42479ea1c4ac3d8286c6	46000.0	0.166667	2686.0
15	62	20160211	F	31dda685af34476cad5bc968bdb01c53	71000.0	0.666667	1958.0

The field “probability” describes the probability that a customer is likely to take the offer, and it is calculated by joining profile table with the transaction table, divide the count of offer complete and count of offer received data. By further look into the correlation between the individual column and probability:







gender	probability
F	0.563502
M	0.431158
O	0.549057

From the chart, fields like gender, income and age might be a good candidate to forecast the probability.

Solution Statement and Project Design

For this business problem, we could train a model by using AWS Sagemaker. The following step would be taken throughout the workflow:

1. Transform the data into a Pandas data frame with the fields gender, age, income and probability.
2. Split the data into training, validation and testing data set, and then output these data to csv files.
3. Upload the csv files to AWS S3.
4. Call Sagemaker estimator object to train the model.
5. Call transform method to predict the test data, display the scatter plot for prediction performance.
6. Deploy the model to an end point and clean up the resources.

Model Selection and Benchmark Evaluation

For this problem, the solution model would be linear regression, as from the chart we could see the relationship is linear and the Y variable is continuous. The benchmark model would be logistic regression model in this case. After training these two models I will run a scatter plot to compare the performance between the solution model and benchmark model.