Machine Learning Engineer Nanodegree Program

Capstone Project Proposal

UDACITY

Starbucks Capstone Challenge

Yassine Elhallaoui
March-2020

Domain Background

Though predictive analytics has been around for decades, it's a technology whose time has come. More and more organizations are turning to predictive analytics to increase their bottom line and competitive advantage because of Growing volumes and types of data, and more interest in using data to produce valuable insights & tougher economic conditions and a need for competitive differentiation using Data Science and Machine Learning.

Organizations are turning to predictive analytics to help solve difficult problems and uncover new opportunities. Common uses in many industries and its a good opportunity for STARBUCKS to be one of the pioneers to implement this solution that will lead to :

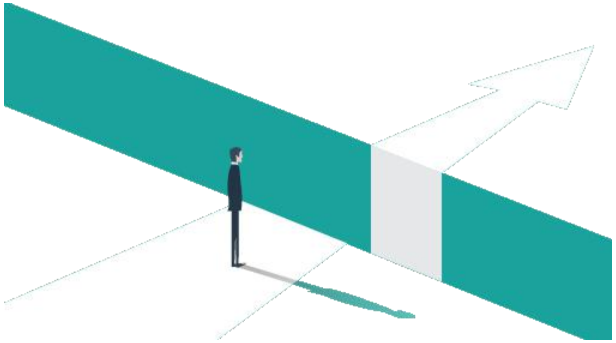
- Optimizing marketing campaigns.
- Improving operations.
- Reducing Business risk.

and studying the client's behaviors will lead to complete satisfaction and make him more comfortable and give the client the sensation of welcome and [we know how to serve you well].

Problem Statement

In another hand we can call it Challenges as with the ML methods it clearly the Solution :

- Combine transaction, demographic and offer data to determine which demographic groups respond best to which offer type. This data set is a simplified version of the real Starbucks app because the underlying simulator only has one product whereas Starbucks actually sells dozens of products.
- Every offer has a validity period before the offer expires. As an example, a BOGO offer might be valid for only 5 days. You'll see in the data set that informational offers have a validity period even though these ads are merely providing information about a product; for example, if an informational offer has 7 days of validity, you can assume the customer is feeling the influence of the offer for 7 days after receiving the advertisement.
- Transactional data showing user purchases made on the app including the timestamp of purchase and the amount of money spent on a purchase. This transactional data also has a record for each offer that a user receives as well as a record for when a user actually views the offer. There are also records for when a user completes an offer.



Datasets and Inputs

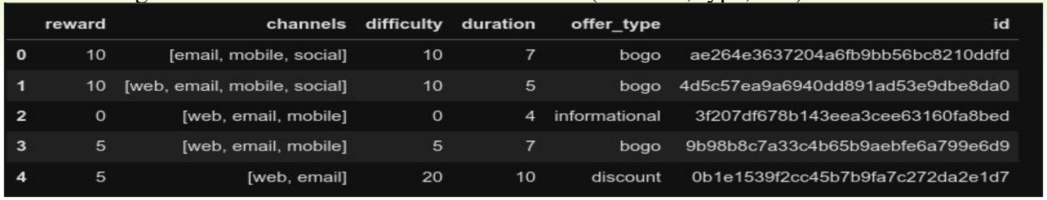
This data set contains simulated data that mimics customer behavior on the Starbucks rewards mobile app.

Once every few days, Starbucks sends out an offer to users of the mobile app. An offer can be merely an advertisement for a drink or an actual offer such as a discount or BOGO (buy one get one free).

Some users might not receive any offers during certain weeks. Not all users receive the same offer, and that is the challenge to solve with this data set.

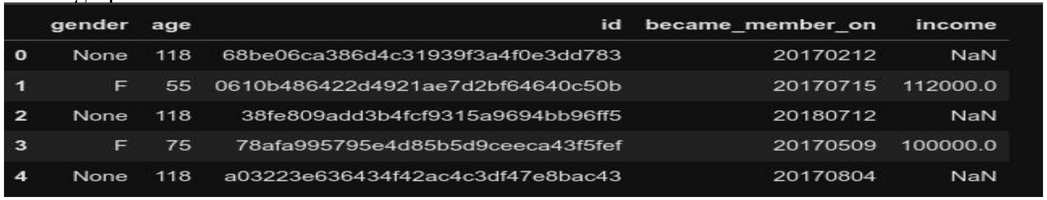
There are 3 files that store these datasets: :

A. portfolio.json - containing offer IDs and metadata about each offer (duration, type, etc.).



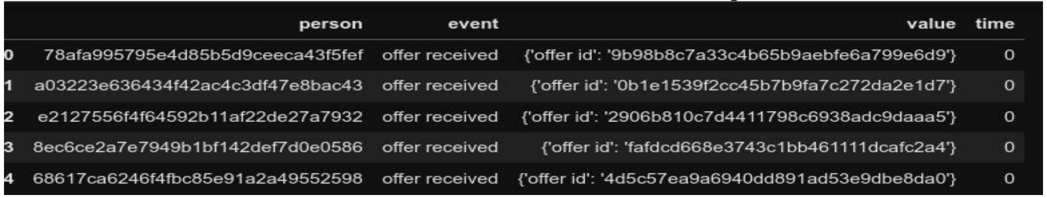
	reward	channels	difficulty	duration	offer_type	id
0	10	[email, mobile, social]	10	7	bogo	ae264e3637204a6fb9bb56bc8210ddfd
1	10	[web, email, mobile, social]	10	5	bogo	4d5c57ea9a6940dd891ad53e9dbe8da0
2	0	[web, email, mobile]	0	4	informational	3f207df678b143eea3cee63160fa8bed
3	5	[web, email, mobile]	5	7	bogo	9b98b8c7a33c4b65b9aebfe6a799e6d9
4	5	[web, email]	20	10	discount	0b1e1539f2cc45b7b9fa7c272da2e1d7

B. profile.json - demographic data for each customer.



	gender	age	id	became_member_on	income
0	None	118	68be06ca386d4c31939f3a4f0e3dd783	20170212	NaN
1	F	55	0610b486422d4921ae7d2bf64640c50b	20170715	112000.0
2	None	118	38fe809add3b4fcf9315a9694bb96ff5	20180712	NaN
3	F	75	78afa995795e4d85b5d9ceeca43f5fef	20170509	100000.0
4	None	118	a03223e636434f42ac4c3df47e8bac43	20170804	NaN

C. transcript.json - records for transactions, offers received, offers viewed, and completed offers.

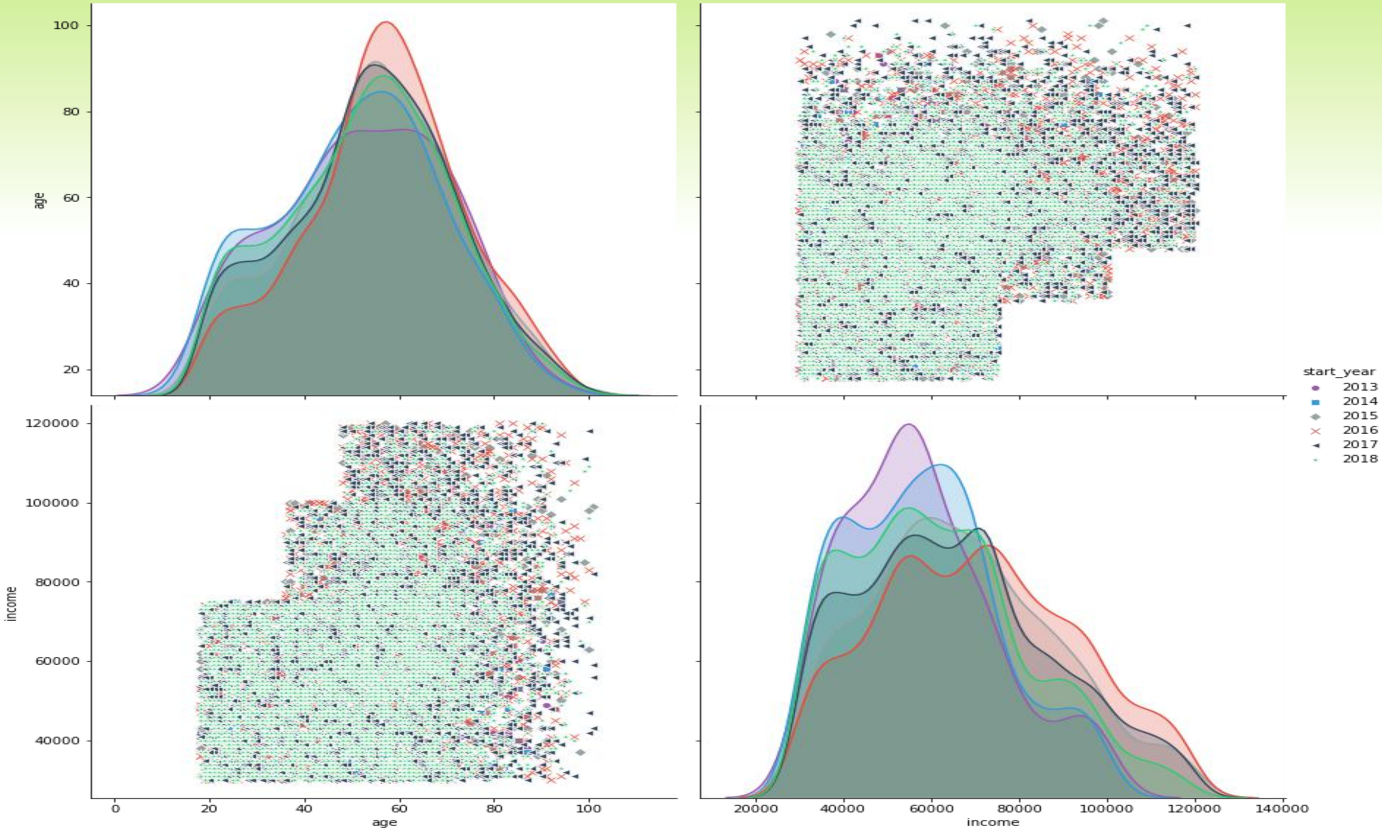


	person	event	value	time
0	78afa995795e4d85b5d9ceeca43f5fef	offer received	{'offer id': '9b98b8c7a33c4b65b9aebfe6a799e6d9'}	0
1	a03223e636434f42ac4c3df47e8bac43	offer received	{'offer id': '0b1e1539f2cc45b7b9fa7c272da2e1d7'}	0
2	e2127556f4f64592b11af22de27a7932	offer received	{'offer id': '2906b810c7d4411798c6938adc9daaa5'}	0
3	8ec6ce2a7e7949b1bf142def7d0e0586	offer received	{'offer id': 'fafdcd668e3743c1bb461111dcafc2a4'}	0
4	68617ca6246f4fbc85e91a2a49552598	offer received	{'offer id': '4d5c57ea9a6940dd891ad53e9dbe8da0'}	0

Datasets and Inputs

Data Visualization :

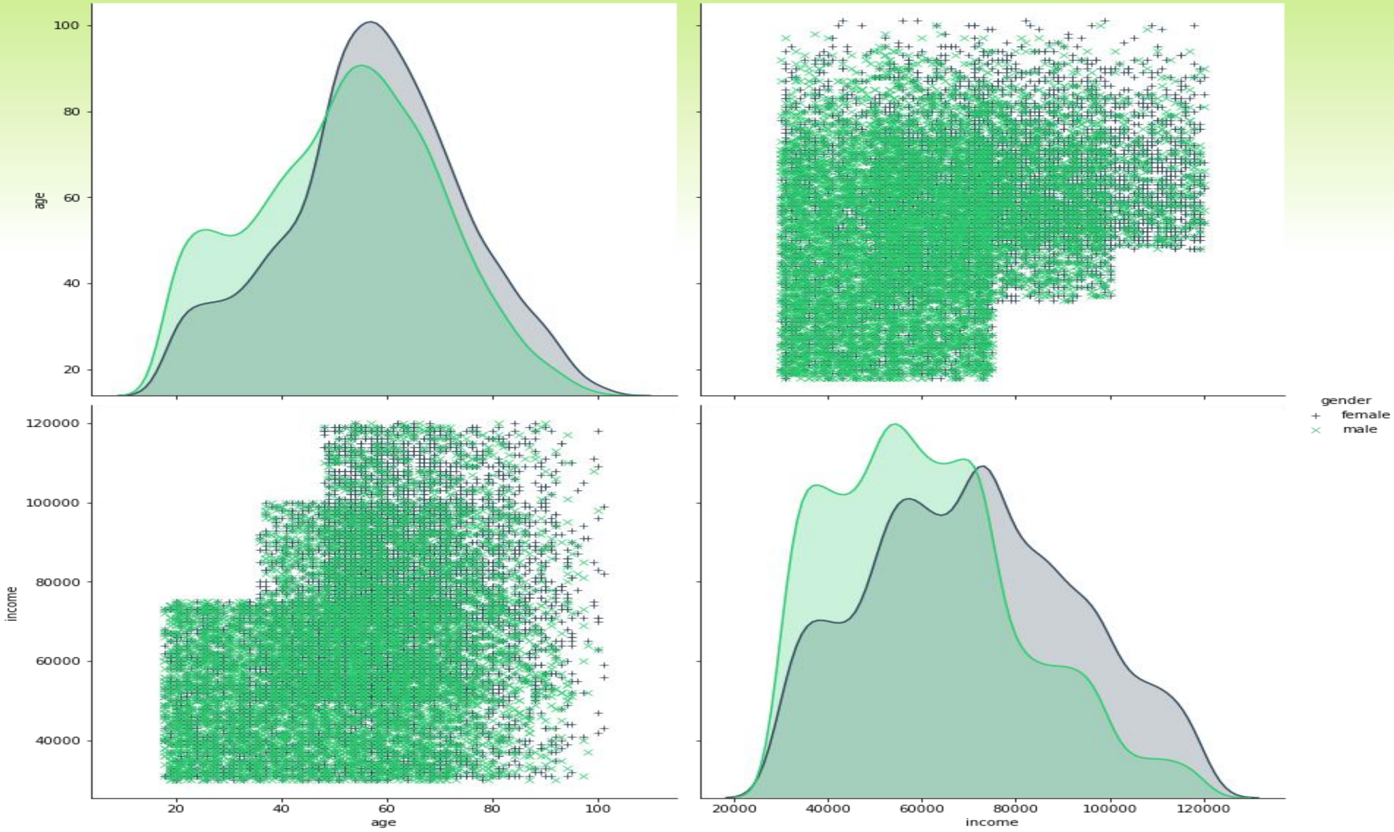
- Customers Global Data Pattern Visualization [Age, Income,] per year: Age distribution plot depicts that the median age of a customer is 60 and most of our customers belong to an age range between 40 to 70. Income distribution plot shows that the number of customers whose average salary is less than 70K is high than the other side considering 70K to be median of the income distribution. Membership distribution has interesting results - 2017 has the highest registered customers than any starting from 2013. The plot also shows that there is an increasing trend in the number of registrations except for 2017:



Datasets and Inputs

Data Visualization :

- plots below conclude that minimum and maximum income for both males and females are approximately the same but the count of males customers in low-income level is slightly higher than that of females customers. considering all highest income are from age od 40 above.



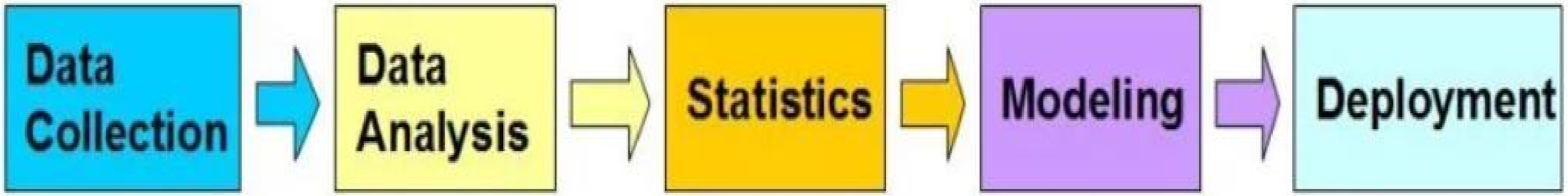
Solution Statement

The feature importance given by all 3 models was that the tenure of a member is the biggest predictor of the effectiveness of an offer. Further study would be able to indicate what average tenure days would result in an effective BOGO offer, My strategy for solving this problem has mainly two steps.

First, I combined offer portfolio, customer profile, and transaction data.

Second, I assessed the accuracy and F1-score of a naive model that assumes all offers were successful. Third, I compared the performance of logistic regression and random forest models. This analysis suggests that the random forest model has the best training data accuracy and F1-score. Analysis suggests that random forest model has a training data accuracy of 0.762 and an F1-score of 0.753. The test data set accuracy of 0.740 and F1-score of 0.730 suggests that the random forest model I constructed did not over-fit the training data.

The Process of our Solution will be as on the below Flow :



Benchmark Model

- As its a Classification problem then the Logistic Regression & Random Forest Classifier will be the Benchmark.
- Accuracy Score and F1_Score will be used to evaluate the performances of the Model against a Naive Predictor as a reference to compare the model's resuts.

Evaluation Metrics

using different statistic measures to end up with the best accuracy :

Precision / Recall : F1-Score

Naive Predictor :

- Naive predictor accuracy: 0.471
- Naive predictor f1-score: 0.640

Logistic Regression Model Results :

Accuracy

Naive predictor: 0.471

Logistic regression: 0.693

F1-score

Naive predictor: 0.640

Logistic regression: 0.805

Random Forest Classifier Model Results :

Accuracy

Naive predictor: 0.471

Random forest: 0.761

F1-score

Naive predictor: 0.640

Random forest: 0.751

?:

Logistic Regression : is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

F1-score : that combines the two measures [Precision & Recall] that tuning the model to grow the precision results in a smaller recall and vice-versa.

Random Forest Classifier : The random forest is a supervised learning algorithm that randomly creates and merges multiple decision trees into one "forest." The goal is not to rely on a single learning model, but rather a collection of decision models to improve accuracy. The primary difference between this approach and the standard decision tree algorithms is that the root nodes feature splitting nodes are generated randomly.

Project Design

First, there is the data preparation step:

review the data sources, understand their content and cleaning the data by recreating the customer journey (from the received offer to the relative transaction) through the transcript dataset. Moreover, we have to join all the different pieces of information coming from the 3 data sources.

Finally, creating the target variable, which is the base of all our analyses.

The next step is data exploration. analyzing the newly formed datasets to understand the distributions of the features and their relationship, especially with the target, investigate possible missing values, data skewness and categorical features with too many categories.

Then, tackle the data preprocessing part. After analyzing the data, and transforming the original dataset through different stages: missing imputation, categories encoding, data standardization.

Second, the development of the model:

Creating 2 different Machine Learning models, one to predict the BOGO propensity, the other for the Discount counterpart. For each model, by trying different algorithms, such as Logistic Regression**,** Random Forest Classifier**. Then, comparing the models and choose the best one for each type of offer with the 2 best models and combine the results in order to obtain a single type of offer to give to each customer.**

Finally, the step to measure the performances of the built process and compare them with the current benchmark, to understand if the proposed solution is viable to implement the current offer attribution process.

Project Design

- **Programming Languages, tools, and libraries:**

1.Jupyter-Lab.
2.sklearn Library.
3.searborn Library for plotting.
4.Pandas Library.
5.tqdm Library.
6.jsone Library

- **Models Used:**

1.Logistic Regression :
2.Random Forest Regressor

Yassine Elhallaoui
March-2020