# Starbucks Marketing Analysis

## Purpose

The purpose of this analysis was to create a model to predict whether a marketing initiative would work on a customer based on the characteristics of the offer and the customer. The datasets provided were:

* portfolio.json – contains metadata about each marketing offer
* profile.json – contains demographic data of the customers
* transcript.json – contains records of transactions, offers viewed, offers received, and offers completed

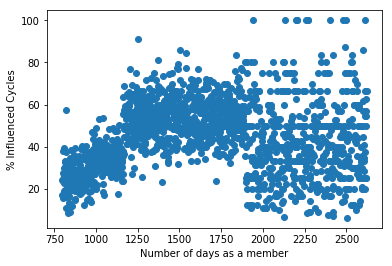
Some interesting information regarding general trends for the customers and the offers that they completed were found, and a decently well performing classification model was created to predict the success of a marketing campaign.

## Exploratory Discoveries

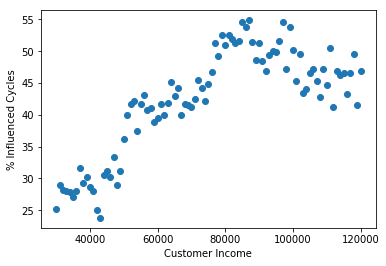
During the exploratory data analysis some interesting information was found about the characteristics of the customers and the offers they decided to complete.

Here the term ‘cycle’ is used to describe a customer receiving an offer, possibly viewing it, and then possibly completing the offer. If the offer was both viewed and completed it is considered to be a cycle influenced by the campaign, and not influenced if the offer was not viewed. If the offer was not completed, then it is considered not completed. The percentage of influenced cycles is the number of influenced cycles divided by the total number of cycles.

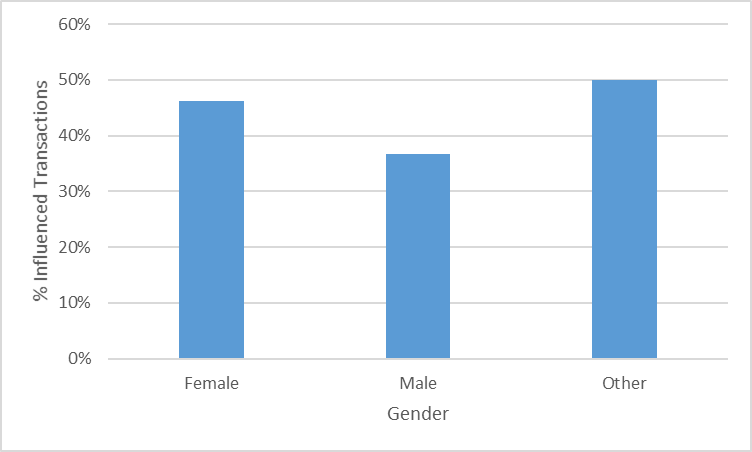
The number of days as a Starbucks member does seem to have a bit of a relationship with the percentage of cycles influenced by a marketing offer, although at around the 5-year mark the relationship decays. In general the linear relationship is not strong (corr=0.15).



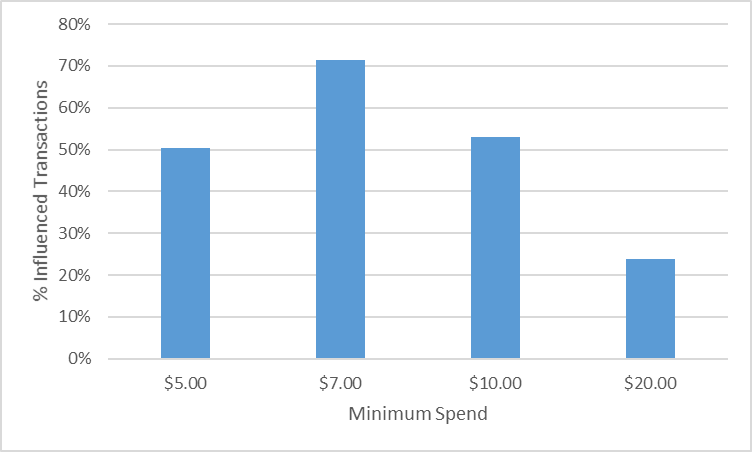
The income of a Starbucks member is generally positively correlated with the percentage of influenced cycles (corr=0.77).



The gender of the Starbucks member has an effect on the percentage of influenced cycles as well. It seems that male customers are the least likely to complete an offer.

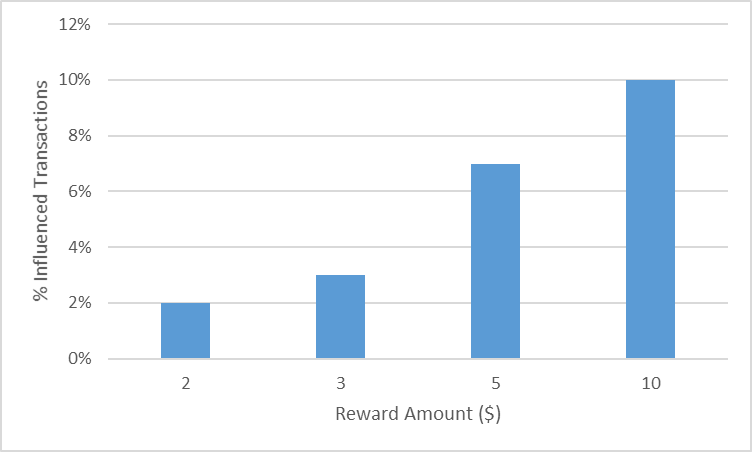


What is particularly strange is the difference between the minimum spend required to claim an offer, and the percentage of influenced cycles:

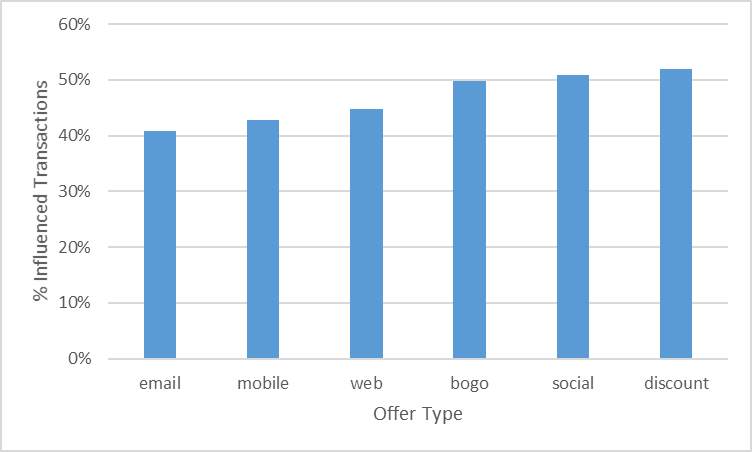


Offers that required a minimum spend of $7 were the most likely to influence a cycle.

Interestingly, rewards that offered only $3 were the most likely to influence a cycle. This may have to do with higher rewards requiring higher minimum purchase amounts.



In general the type of offer didn’t seem to have much effect on whether or not a cycle would be influenced. A further A-B test could confirm, but from the graph and numbers only there is not much difference.



## Model

To predict whether a marketing campaign would influence a certain type of customer, classification models were used. Logistic regression, SVM, and XGBoost were all tried. Logistic regression (accuracy=73.7%) and SVM (accuracy=76.2%) benefitted greatly from input scaling, but were still vastly outperformed in accuracy by the final XGBoost model.

After running k-fold cross validation with 10 splits on 4 different hyperparameters, the XGBoost model performed at an accuracy of 78.4%.

## Going Forward

In the future in order to increase the accuracy of the prediction model the following could be implemented as input variables:

* minimum gain = reward amount - minimum amount
* a separate variable to indicate if an informational offer influenced a cycle
* Using information not provided in the three datasets mentioned above:
  + average spend for each member
  + time since last purchase
  + frequency of purchases

Additionally with more computing power and time, the XGBoost model could be further tuned for greater accuracy. As it was, even introducing a parameter in the model to use GPU to speed up the process, model fitting took over 20 minutes. With more hyperparameter options and more folds a better model could be produced.