Krishna Goel
3/13/2021

Optimizing Promotion Success for the Starbucks Mobile App

Background

Marketing has been a major focus of organizations for a long time. With the growing access to technology and the accompanying creation of data, corporations can better market their products and services (Kolathayil). The adoption of machine learning and artificial intelligence by marketing organizations is staggering. According to Forbes 84% of marketing organizations were either implementing or expanding AI capabilities in 2018. This is fueled by the success of AI backed marketing's empirical success, with Forbes stating that 75% of organizations see over a 10% boost to sales. Marketing isn't just about improving the company's bottom line. In fact 57% corporate executives believe AI will improve customer experience (Columbus). Afterall, ultimately better marketing means customers will get the products they desire.

Marketing is defined as "the process of determining customer wants and needs and then providing customers with goods and services that meet or exceed their expectations" (A Brief History). Part of this is being able to determine what kind of promotions consumers are likely to respond to. A company has many tools available to it for marketing. With a variety of consumers, it is essential that a company figure out what tool will be effective with which consumer. This way the corporation will be able to cost-effectively spend its marketing money and consumers will be able to enjoy their desired products at a preferable price.

Setup

Problem Statement

Starbucks has provided through Udacity simulated data for a single product from its promotion app. Through this application Starbucks sends the users a variety of different promotions, including Buy One Get One (Bogo) and Discounts. In addition to the types of promotions that Starbucks, they have provided some demographic information about the users and all the "interactions" that user has had with Starbucks. Such as when the received, viewed, and completed a transaction and when they bought something at Starbucks. After playing around with the data, I noticed that not all of Starbucks promotions translated into a user shopping at Starbucks. This raises a natural question of value to the company, what promotions are successful in influencing which users to shop at Starbucks?

It would be valuable if Starbucks could send a given user exactly the promotion, he wants to see in order to shop at Starbucks. This way Starbucks would potentially waste less money handing out ultimately futile promotions and make the rewards app more valuable to the consumer, giving them exactly what they desire.

My approach to discovering which promotions are more successful starts with identifying when a promotion "influenced" a user to shop at Starbucks. I will compute this influence by seeing when a user received a promotion, viewed it, and within the validity period

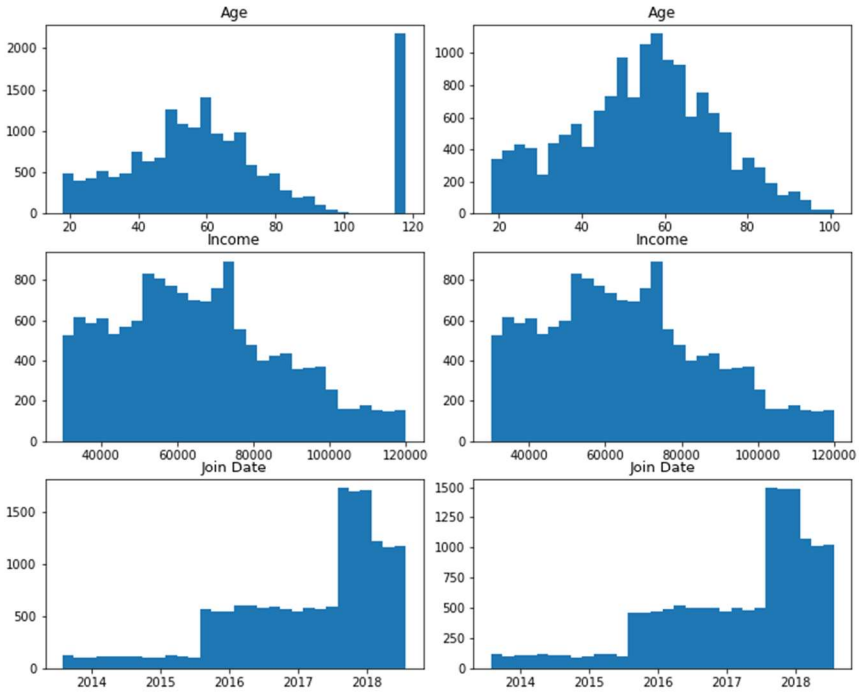
of the promotion shopped at Starbucks. Conversely, if someone received a promotion but didn't follow the above step – then the corresponding promotion was unsuccessful. With this binary indicator in place, I will use the information provided about the user and the promotion to create a model that can predict the success of a promotion.

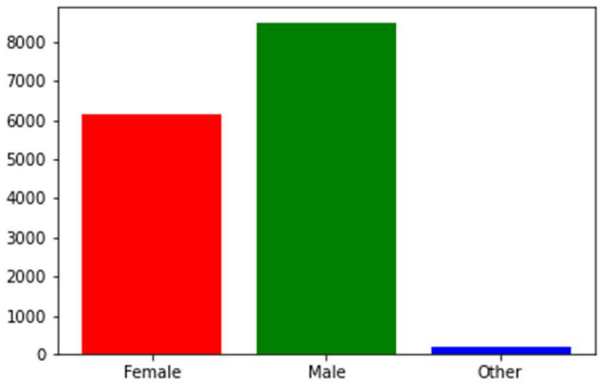
Metrics

The primary metric I will be using is the f1 score. The ideal situation for Starbucks is that it sends promotions to all customers who will shop because of them and no superfluous promotions. This requires a balance between precision and recall, something the f1 score evaluates very well. It is worth noting that in this case, Starbucks is sending promotions using its own online app, so it doesn't incur a huge cost of actually sending the promotion. Consequently, it would actually make sense in this scenario for Starbucks to give more weightage to recall, as it would be worse to miss out on a potential customer than the loss of sending the promotion to a customer who won't use it. As a result, the model I will ultimately use will have a very high recall in addition to the best f1 score. I still chose to go with the f1 score to provide a more general analysis for situations in which there is a substantial cost of advertising involved. All three will be presented for each model I use.

Data Exploration and Visualization

Profile: This table includes demographic information on customers. The given information is their customer's age, income, gender, and when the became a member of the Starbucks rewards app. I have included histograms of these below.

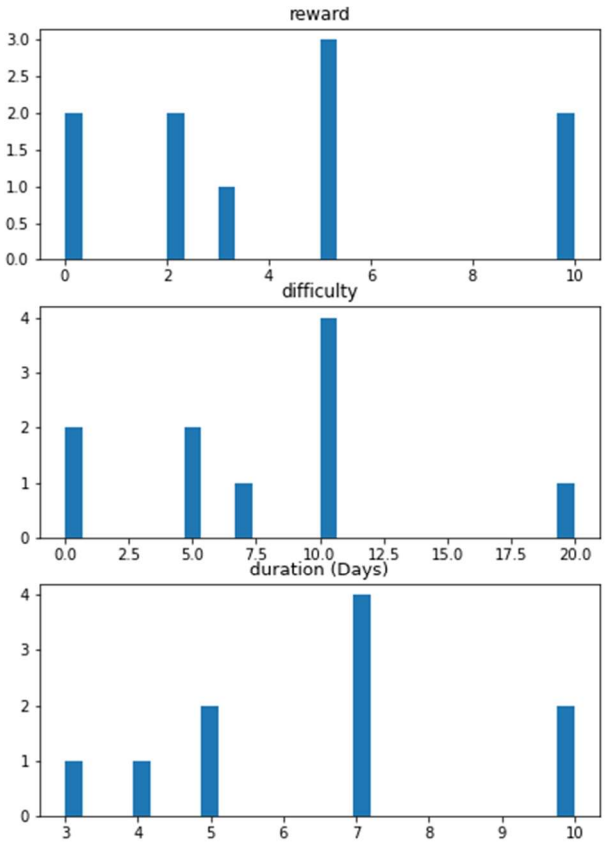


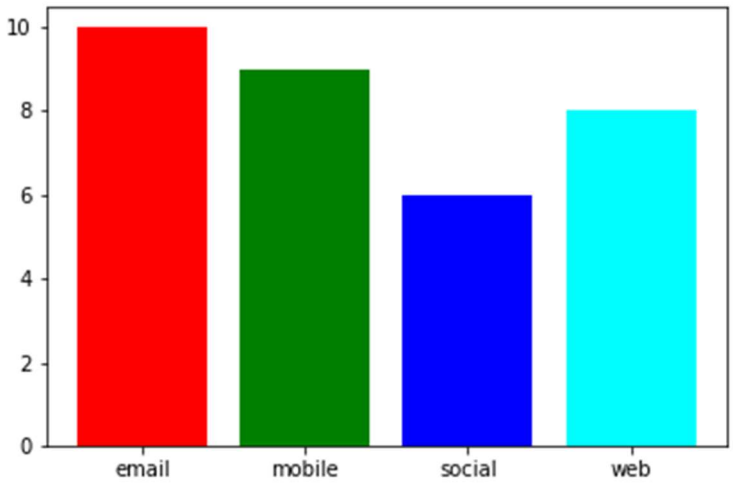


On the left, looking at the histograms of age, income, and join date. The first thing that jumps out is the cluster of people aged 118. After investigating these data points, it turned out that for these individuals we lacked data for income and gender. It looks like that for some users we simply did not have any demographic information. In fact, investigating missing data I found that the only missing data in the "profile" table had age as 118. Of the 17000 individuals given here 2175 had missing data (12.79%).

Before putting the data through models to predict promotion success I excluded any data relating to these individuals. I did this because there because the individuals missing one entry are missing all and there isn't really a natural default for any of the fields. Given the percentage of missing data, taking the mean would simply introduce significant noise.

Portfolio: This table included information about the kinds of promotions Starbucks sent to its app users. There were a total of 10 promotions – 4 Bogo, 4 Discount, and 2 informational. These promotions had additional descriptions such as the reward of the promotion in dollars, its validity period in days, the difficulty of the promotion in dollars (how much the customer would have to spend to get the reward), and the mediums through which Starbucks shared the promotion. Below I have shown histograms for the data in this table.

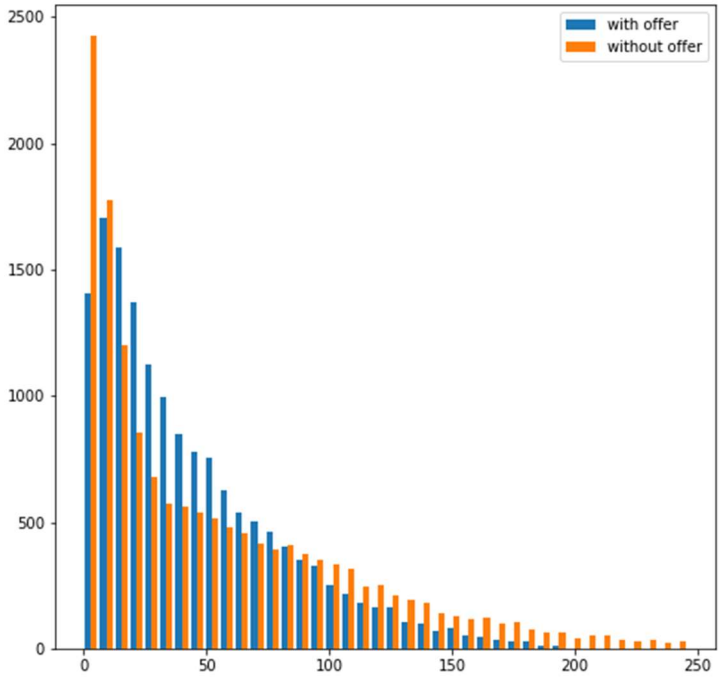




Looking at the above histograms nothing seems substantially odd. We see a nice distribution on the properties of the promotions such as rewards, difficulty and duration – all are roughly equally spread out. The fact that email was a medium for all promotions will cause some models to use the added degree of freedom to overfit, consequently I exclude before running models in the future.

Krishna Goel
3/13/2021

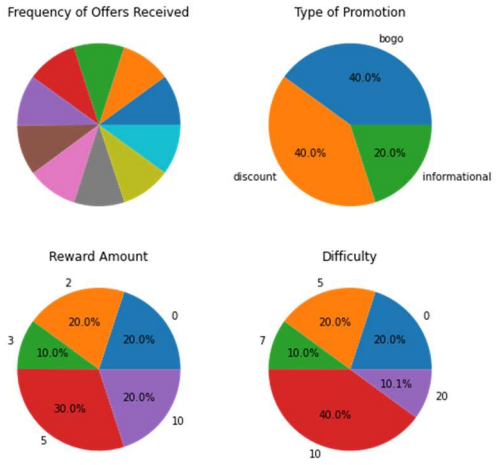
Transcript: This table had data on when customers viewed, received, and completed promotions since the start of the trial (in hours) as well as when they bought something at Starbucks. I used this table to mark transactions as either being “with promotion” or “without promotion”. This was done by seeing whether the user had seen a promotion and, in its validity, shopped at Starbucks. If they did so, then we say the transaction was “with promotion” and we attribute all promotions that were valid at this time (and viewed) as having contributed to the user’s decision to shop at Starbucks. To see if promotions were even useful in boosting Starbuck’s revenue, I plotted a histogram of how much individuals spend at Starbucks with and without promotion. I have included this below.



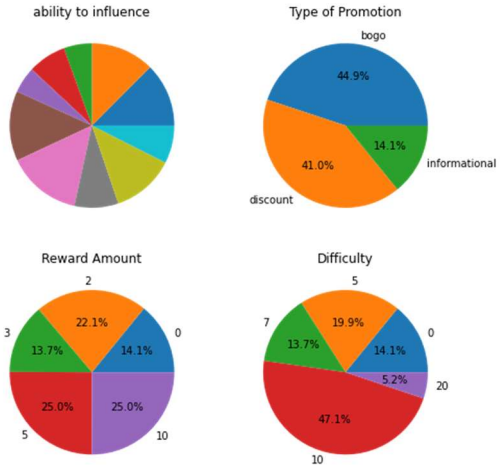
We clearly see from this histogram that customers indeed appear to spend more money while “with offer”. This seems to be especially true for those who spend fewer than a 100$ at Starbucks as we see the blue bars higher than the orange in this range. It is likely that customer’s who spend a lot at Starbucks, simply didn’t have enough promotions to cover all their spending.

Promotion Influenced Purchase Decisions: In order to determine if certain promotions or certain individual characteristics had any bearing on the success of a promotion, I first computed when offers “influenced” someone and when they didn’t. For this I used the transcript table to get how many times a user received a given promotion. Next, I did the same for influence for which I used what I had computed previously. That is if someone had a transaction that was “with offer” and if so with which offer. Using these two tables, I determine if someone had been influenced by a given promotion and received it that was the individual and promotion combination works, alternatively if the individual received the promotion but wasn’t influenced then it doesn’t work. Below are the PI charts showing these distributions.

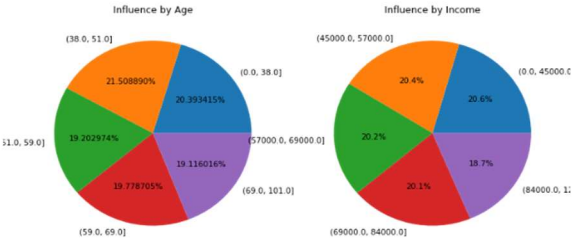
Krishna Goel
3/13/2021



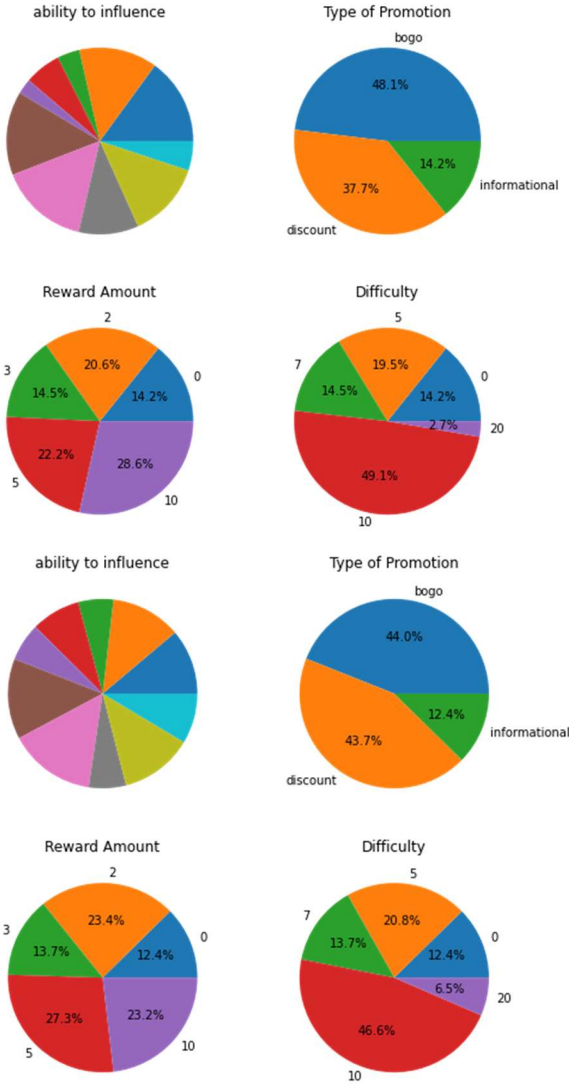
On the left we first see the raw distribution. That is the distribution of offers received by individuals and the characteristics of these offers. We see that offers were given at the same rate (top left) and other characteristics are distributed by how often they are in an offer. For example, if a reward was in 30% of the offers then it was received by individuals 30% of the time.



Now on the left we have the distributions but based on how often offers influenced individuals. We see that discount and Bogo were more likely to influence an individual's decision to buy relative to the raw distribution. Furthermore, we notice that rewards of 2, 3, and 10 entice people to buy more. Similarly, middle difficulties were also more successful in getting people to buy, likely indicating a sweet spot for difficulty/reward trade off.



Now we see distribution of offers influence broken by age and income. Here, we notice something different. That is in aggregate offers do not have different influences based on the users age or income. However to investigate if demographics have no predictability I also made plots of offer type where the data had been subset to certain income groups.



Here we see the distribution of offer influence but only for individual earning the bottom third of salaries (less than 40,000$). We see this is different from the distribution of influence above. For instance Bogo now is far more successful of a promotion and we see other trends even more pronounced. Such as reward 10 being successful far more often.

Here we see the same distributions but for individuals earning in the top third of salaries (80,000 – 120,000$). Here we the opposite store, where relative to aggregate influences, Discount is a more successful promotion type and reward 10 is relatively less successful.

Data Exploration Takeaway: We see that combining the promotion data with the demographic data there is definitely some relationships that can be exploited to increase success of promotions.

Algorithms and Benchmark

Benchmark: I ran the multinomial naïve-bayes. This model is an efficient algorithm that is often used in binary classification tasks (Ortner). The Naïve-Bayes classifier achieved an F1 score of 0.544, precision of 0.538, and a recall of 0.551.

Algorithms:

Linear Learner: We have a binary classification task with a rich set of predictors. On the other hand, due to the binary nature of a lot of the dummy variables a close form

solution to regression will be highly sensitive to errors in the data due to collinearity. Consequently, AWS's implementation of logistic regression will offer the benefits of logistic regression in our classification task while using a gradient based numerical optimization algorithm. Furthermore, the AWS linear learner can optimize while reaching a target recall or precision. Because of the importance of recall in this task, I have decided to try Linear learner with different target recall values.

XGBoost: XGBoost is an immensely powerful algorithm that utilizes weak decision trees to create an ensemble scoring for a learning task. This capability of XGBoost will allow us to create decision boundaries that combine information about the person's demographic and the promotion characteristics. That is, for instance, a higher income person plus discount gives more information that the person alone. XGBoost's flexibility will allow us to have such combinations taken into account. We will be running XGBoost to maximize f1 score and tune its hyperparemters

- Max Depth: I used trees of depth 3-8 as we have 9 input variables (counting the dummy variables as 1). Best = 4
- Eta: I set eta from 0.05 to 0.5, we allowed the wide range to prevent overfitting but at the same time understanding that there is very nuanced information in the data so the small eta values might give the best model. Best = 0.073
- Alpha: This is L1 Regularization, I set range to 0 to 2. Again, not wanting to penalize variables too much due to relatively small amount of information they each carry. Best = 1.88
- Min Child Weight: I set the range to 1 to 10, Here I thought this was a good balance on creating new leaves between having the information in different variables combined and having shallower models. Best = 10

Methodology

Data Preprocessing: In addition to calculating the "influence" metric as described above in the data exploration section, I also had to discard any data related to just over 12 percent of individuals. I did this as these individuals lacked any demographic data, which was critical to answering my question.

Furthermore, we had a few categorical variables. This included promotion type, channel for promotion dissemination, and gender. I created dummy variables for each of these variables. Before using the data for model building I discarded the dummy variable related with the email channel because it was utilized in all promotions.

Finally, we had a datetime variable. This was when everyone joined the Starbucks rewards program. In order to translate this variable into a numerical quantity I found the earliest join date and translated every individuals join date in terms of days since that date.

Implementation: Before using Sagemaker, I used several scikit learn models to get a feel for the data. One thing that immediately popped out was that almost all the models achieved an accuracy of 77% regardless of the hyperparameters. Investigating this I found out the models were all predicting that promotions will be unsuccessful as in the overall data 23% of the actual

promotions were successful in influencing consumers. Consequently, modes that couldn't improve on this likely found the highest accuracy in predicting all labels the same. To get around this I created a new data set which had the labels in equal frequency. This helped see which models were able to better predict the success of a promotion.

Other interesting coding problems included calculating the influence metric on the data as this was an extensive computation and as a result was error prone. While doing this I constructed tests to ensure that my algorithm was giving the results I expected. I constructed tests at other parts as well, such as getting which promotions were received/viewed by which individual as this was also a computationally extensive task and somewhat prone to bugs.

Refinement: As described somewhat in the algorithms sections. I utilized AWS's linear learner and XGBoost to come up with a model that would help Starbucks determine what promotions are working. I doing so I utilized Linear Learner at different target recalls and XGBoost with hyper parameter tunning with the ranges justified in the algorithms section. Below is a table of various success measures along with our chosen f1 score for the best XGBoost model, the best recall value, and the raw linear learner.

Model	F1	Recall	Precision	Accuracy
Linear Learner (No target Recall)	0.620	0.655	0.589	0.598
Linear Learner (Recall target = 0.9)	0.667	0.901	0.530	0.550
Best XGBoost	0.621	0.661	0.587	0.597

We see in the above table that not only did the linear learner at 0.9 target recall have a higher F1 score, it achieved a remarkable recall on the test set. This could be very valuable to a company like Starbucks, as now it can identify which individuals to send its promotions to!

Best Model Evaluation

Robustness: To evaluate the robustness of my linear learner model with target recall of 0.9, I ran 5-fold cross validation using this model. Below I have a table of the results of the cross-validation on all four metrics, f1, recall, precision, and accuracy.

Fold Number	F1	Recall	Precision	Accuracy
1	0.679	0.900	0.545	0.563
2	0.677	0.899	0.542	0.564
3	0.657	0.896	0.518	0.545
4	0.654	0.899	0.514	0.540
5	0.677	0.908	0.539	0.559

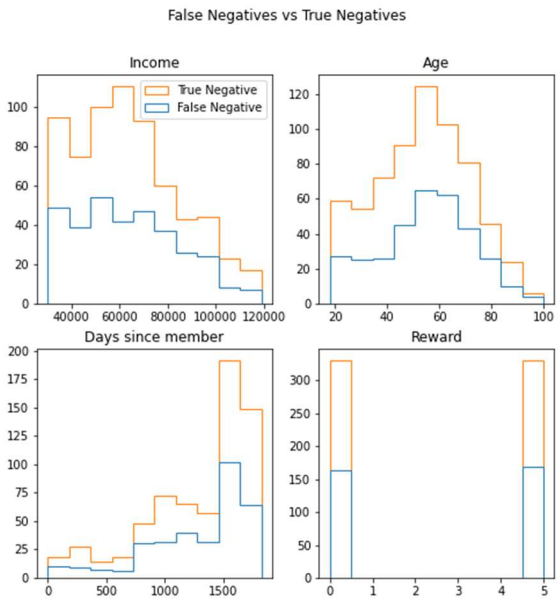
As we observe, the linear learner model is very robust with the evaluation statistics being roughly equivalent across every evaluation criterion. The most substantial variation is in precision, where we see a 0.3 drop in folds 3 and 4 relative to the other 3. This results in these folds also having lower f1 and accuracy scores. Nevertheless, overall the results suggest the

model is robust. The standard deviation of the F1 statistic is 0.01 and the range is 0.025 (model 1 and 4). This is a relatively small spread – substantiating that indeed the models performance doesn't vary a lot based on the test data.

Comparison to Benchmark: We see the f1 score of the linear learner is 0.667 while it was 0.554 for the benchmark. This is over a 0.1 improvement, which is certainly significant when also taking into account the robustness of the linear learner model. Furthermore, the linear learner has a far better recall than the benchmark model (0.901 vs 0.551) while having a relatively similar precision (0.530 vs 0.538). This is especially notable for the case of Starbucks due to the relative low cost of sending promotions. That is, Starbucks doesn't pay a lot to send a given promotion to a user so they would ideally want to send the promotions to as many people as would be influenced by them. Hence the fact the linear learner has such as high recall while having a high f1 demonstrates that it is indeed very useful for this use case.

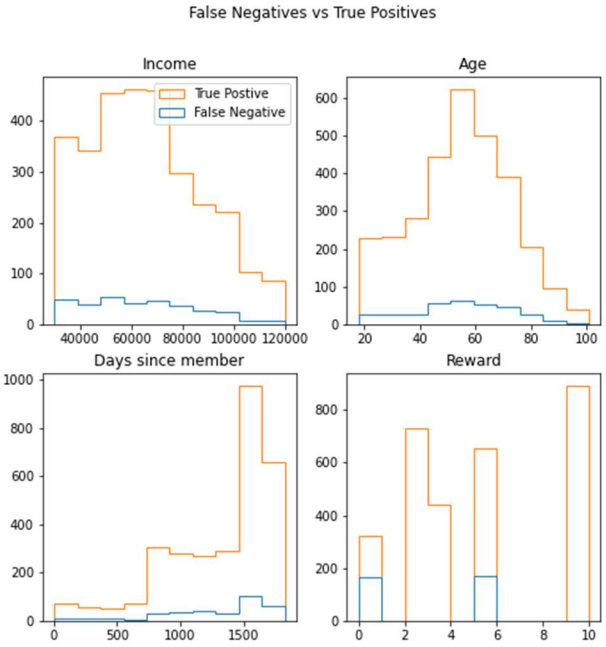
I would like to point out that Starbuck's original data would constitute sending out a promotion to every individual, so in the space of the data we have, Starbucks has a recall of 1, albeit a precision of around 0.23 due to 23% of those promotions influencing people. This is simply the nature of any useful data we would get for this task, it will be sent promotions and will have this high recall value. However, if Starbucks utilized our model it would be able to, as long as our sample is representative of the population of Starbucks customers, send promotion to around 90% of all individuals who will be influenced by. This could be a win for Starbucks as we saw in the data exploration section, that indeed people spend more at Starbucks when in possession of a promotion.

Investigating False Negatives: On the right I have shown some distributions of promotion and demographic characteristics in the case of false negatives, and true negatives. As expected, these, distributions are similar in density and we see that there are certainly cases where individuals will be influenced by promotions even though characteristics about them and promotion might suggest otherwise. That is we see false negatives have similar distribution to true negatives.



Krishna Goel
3/13/2021

Here on the right, we see the same variables but instead we are plotting true positives against the false negatives. We see that indeed true positives have quite a different distribution. For instance older people and middle income individuals appear more likely to use promotions. On the other hand, how long an individual has been a member appears to have no bearing on whether they will use the rewards app.



Conclusion:

I believe that the linear learner offers a good model to judge which individuals a certain promotion should be sent to. Nevertheless, I certainly think we can do better given the low accuracy of our model. I would imagine time series data, that is past patterns of a consumer would offer very good insight into their likelihood of respondonding to a future promotion. The current dataset doesn't have a lot of data for single individuals to offer significant training/test data in this direction. However, with more long term data on individual behavior I believe we would be able to create strong models at predicting promotion success.

Krishna Goel
3/13/2021

Bibliography

A BRIEF HISTORY OF MARKETING. Valencia College http://faculty.valenciacollege.edu/srusso/chapter13.htm. Accessed 10 Mar. 2021

Louis Columbus. "10 ways Machine Learning is Revolutionizing Marketing." Forbes, 25 Feb. 2018, https://www.forbes.com/sites/louiscolumbus/2018/02/25/10-ways-machine-learning-is-revolutionizing-marketing/?sh=622a3ddc5bb6

Kolathayil, Yasim. "Machine Learning for Marketers." *Medium*, Towards Data Science, 28 Feb. 2019, https://towardsdatascience.com/machine-learning-for-marketers-78bff070cbd6.

Ortner, Alex. "Top 10 Binary Classification Algorithms [a Beginner's Guide]". *Medium*, 28 May. 2020, https://medium.com/@alex.ortner.1982/top-10-binary-classification-algorithms-a-beginners-guide-feeacbd7a3e2