Capstone Project Xiaohan Liu

Data Scientist Nanodegree Jan 17, 2020

Optimizing App Offers with Starbucks

**Definition**

***Project Overview***

This project analyzes Starbuck customers’ behaviors to learn how people make their purchasing decisions and how those decisions are influenced by promotional offers. Customer profile, promotions and customer purchasing events datasets are available for this analysis.

***Project Statement***

There are three goals for this project:

1. Identify a cohort whose purchasing decision is not affected by the promotional offers
2. Build a classification model to identity cohort who will be most likely to purchase with promotional offers
3. Identify an efficient way to send promotional offer to customer

The following step are taken to achieve the goal:

1. Download and preprocess customer profile, promotions and customer purchasing events datasets
2. Conduct exploratory analysis to identify customers’ observable traits associated with purchasing behavior
3. Train a classifier that can determine if a customer will purchase with a promotional offer
4. Evaluate and refine the classifier

The analysis is expected to be useful for determining an efficient way to send promotional offers. The classifier is expected to identify a cohort whose purchasing event is affected by promotional offers.

***Metric***

Classification model accuracy =

This metric is used to evaluate how accurate a classifier can classify customers who use promotional offers.

**Analysis**

***Data Exploration***

The profile dataset contains 17000 Reward program users’ information with the following 5 fields:

1. Gender (categorical): male, female, other, null
2. Age (numeric): customer’s age; missing values are encoded as 118
3. Id (hash/string): customer id
4. became\_member\_on: when customer became a member with Starbuck
5. income (numeric): customer’s income

The portfolio dataset contains promotional offers sent during 30-day test period with the following 6 fields:

1. reward (numeric): money awarded for the amount spent
2. channels: offer delivered through web, email, mobile and social
3. difficulty (numeric): money required to be spent to receive reward
4. duration (numeric): time for offer to be open, in days
5. offer\_type (string): bogo (buy one get one), discounts informational
6. id (string/ hash): promotional offer id

The transcript datasets contain 306648 offer events with the following 6 fields:

1. person (string/hash): customer id
2. event (string): offer received, offer viewed, offer completed, transactions (not using offer)
3. amount(numeric): money spent in purchasing event
4. offer\_id(string/hash): promotional offer id
5. reward (numeric): money gained from “offer completed”
6. time (numeric): hours after start of test

Some data above is missing. Promotional offer id is stored as string/ hash format. We will drop the missing values and extract the offer id during the preprocessing step.

***Data Visualization***

Fig. 1 The plot shows the distribution of customer age in the Reward program. Customer age ranges from 18 to 100. The median customer age is around 60. Some outliers existed since missing values are encoded as 118.

Chart, box and whisker chart

Description automatically generated

Fig.1

Fig. 2 This plot shows the distribution of customer income in the Reward program. The range of income is wide, and the median income is around $65,000.

Chart

Description automatically generated

Fig.2

**Methodology**

***Data Preprocessing***

The data preprocessing steps are available in the “Starbucks Capstone Challenge” notebook and consists of the following steps:

1. drop missing value from profile datasets
2. extract promotional offer id from transcript “value” column

Text

Description automatically generated

1. join profile and portfolio dataset
2. calculate the following ratios for each customer and join the results with profile and portfolio dataset named customer\_profile
   1. transaction ratio = # of transactions / # of purchasing events
   2. completion ratio = # of offer completed / # of offer received
   3. view ratio = # of offer viewed / # of offer received
3. create dataset offer\_profile\_group1 by filtering on customer\_profile with completion ratio =0
4. create dataset offer\_profile\_group2 by filtering on customer\_profile with completion ratio >0

***Implementation***

Before implementing a classifier, we want to do some exploratory analysis to see if there are any differences between offer\_profile\_group1 and offer\_profile\_group2. Fig.3 shows the comparison of ages between the two groups. Blue represents offer\_profile\_group1 and orange represents offer\_profile\_group2. We can see the median age of the customer group that has completion ratio 0 is less than the median age of the customer groups that has completion ratio greater than 0. This suggests that younger people may tend to purchase no matter whether they are offered a promotional offer or not.

Chart, box and whisker chart

Description automatically generated

Fig.3

Fig.4 shows the comparison of income between the two groups. Blue represents offer\_profile\_group1 and orange represents offer\_profile\_group2. 75% of the customer group which complete 0 offers have income less than the median income of customer groups which complete 1 or more offers. This suggests that younger people with lower income are generally less sensitive to promotional offers. we focus on offer\_profile\_group2 to understand what causes customers to respond to an offer.

Chart, box and whisker chart

Description automatically generated

Fig.4

A random forest classification model is fitted and trained on the preprocessed training data (offer\_profile). Detailed process can be found in the Jupyter notebook, “Starbucks Capstone Challenge”.

1. Get dummy variables for categorical data like “gender”. Transfer completion ratio column. If completion ratio is greater than 0, then assign 1 to replace the original value.
2. Split datasets to x and y
   1. X includes age, income, gender, became\_member\_on, number of transactions completed, and number of offers viewed
   2. Y is completion ratio
3. Split dataset to train and test data sets
4. Create random forest classifier and fit it on the training dataset
5. Use the trained model to predict whether a customer will complete an offer and compared the prediction with the true actions to calculate the accuracy score

Table

Description automatically generated

Text

Description automatically generated with low confidence

Text

Description automatically generated

***Refinement***

**Model**

Since the accuracy score of the random forest classifier is 0.917, I used different classification methods, Logistic Regression and K-Nearest Neighbor to fit training datasets. The best score comes from random forest classification model, 0.917.

Text

Description automatically generated

Text

Description automatically generated

**Parameters**

Perform grid search to search for the best parameters of random forest tree classifier and K-Nearest Neighbor. The best score, 0.932, comes from random forest classifier with n\_estimators=9 and max\_depth = 7

Random Forest:

Graphical user interface, text

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

K-Nearest Neighbor:

Text

Description automatically generated

Graphical user interface, text, application

Description automatically generated with medium confidence

**Results**

***Model Evaluation and Validation***

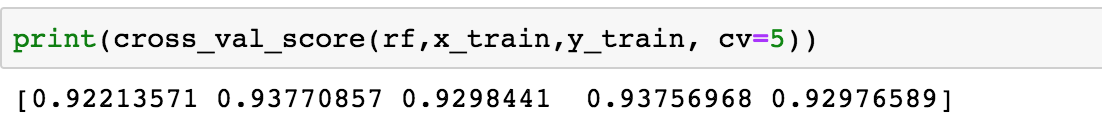
Optimal model: random forest classifier

Optimal parameter: max\_depth = 7 and n\_estimators=9

Accuracy score: 0.932

In order to validate the random forest classification model, cross-validation and classification report are used.

Below are the results for 5-fold cross-validation.



Below is the classification report for the random forest. The overall precision rate is 0.93 and the recall is 0.93. The model can be used to predict whether customer will complete an offer.

Table

Description automatically generated

***Justification***

The best classification model is the random forest classification model with max\_depth = 7 and n\_estimators=9. This model works better than the logistic regression and k-nearest neighbor model in this case since the random forest classifier adds in extra randomness to the model and the ensemble technique further improves the model.

**Conclusion**

***Reflection***

There are three goals for this project.

1. **identify a cohort whose purchasing decision is not affected by the promotional offers**.

The cohort with completion ratio zero is the group of people whose purchasing decision is not affected by promotional offers. Even though those people receive and view the offers, they never complete the offers. I did the explanatory analysis to identify some observable traits for this cohort. Fig.3 shows the comparison of age between cohorts using/not using promotional offers. Fig.4 shows the comparison of income between cohorts using/not using promotional offers. Based on the two boxplots, we can conclude that the cohort, not using promotional offers, are generally younger and have income less than $67,000 which is the median income of the cohort using promotional offers. Since their purchase behaviors are not affected by promotional offers, I suggest not to offer promotions to this cohort.

1. **Build a classification model to identity cohort who most likely to purchase with promotional offers**

A random forest classifier was built to predict whether a customer will complete an offer. The classifier was fit on the cohort with income greater than $67,000 and can be used to predict whether customer will complete an offer given his gender, age, income and purchasing history with 93% accuracy.

1. **Identify an efficient way to send promotional offer to customers**

Fig.5. shows the completion ratio for different offer type. Fig.6. shows the completion ratio for rewards and Fig.7 shows the completion ratio by offer type and gender. Discount with reward 3 has the highest completion rate. Female has relatively higher completion rate than male for both bogo and discount. I recommend offering discount with reward 3 or 2 more often to female customer group.



Fig.5

Fig.6 Fig.7

***Improvement***

Most analysis in this project focused on the observable traits of customers such as age, gender and income. However, more information like, occupation, can be collected to better understand the story behind customer’s purchasing behaviors. some hidden traits of customer may be more useful to build a classification model. Since the precision of the random forest classifier is 93%, the model can be further improved by using data that can truly reflect customer’s decisions.

For example, if occupation and purchasing channel data are available, we could better understand what impact a 22-year-old student’s purchasing decision. If we are seeing people with similar occupations or in the same demographic location have similar purchasing behaviors, we can identify hidden traits and use those to build a model. Different model may be needed for people from different locations.