Data Scientist Nano Degree

Starbucks Capstone Challenge

Yusheng Yang
Apr 19th, 2020

Project Overview

This project is based on the data of user behavior on the simulated Starbucks rewards mobile app. Starbucks will send some push to app users. The push may be just an ad for a drink or a coupon or BOGO. Different push types may have different push validity period, i.e. touch frequency. If a message push is valid for 7 days, you can think that the customer may be affected by the push in these 7 days.

The purpose of this project is to combine transaction data, demographic data and push data to determine which kind of people will be affected by a push?

This project will train a machine learning model. The project will apply xgboost model to build the forecast. Xgboost is a boosting integration algorithm. Its main idea is to combine hundreds of tree models into an efficient model, which can generate new trees through continuous iteration. Take the challenge at Kaggle, the machine learning competition website. Of the 29 challenge winning solutions released by kaggle blog in 2015, 17 used xgboost. Among these solutions, eight use xgboost alone to train models, while most others combine xgboost with neural networks to produce results.

Problem Statement

The goal of the project is to study the impact of each push on consumers. According to the experimental results, we can decide which push strategy is more suitable for different business goals. Possible business objectives include:

1. Which offer can make consumers come back faster?
2. Which offer can maximize the amount of each consumption?
3. Predict the possible average consumption amount of each consumer.

Metrics

1- Return rate

When measuring the return speed of consumers, the return rate is used:

$$r = \frac{X}{N}$$

X: Number of consumers who purchase more than or equal to 2 times in the experimental period;
N: Number of consumers with consumption in the experimental period;
r: The return rate of comsunmers;

2-RMSPE

RMSPE is used as the evaluation index, which represents the difference between the final prediction result and the actual value. The project will evaluate the prediction model and optimize the model parameters. The goal of the project will be to minimize RMSPE.

$$\mathrm{RMSPE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\frac{y_i - \hat{y}_i}{y_i})^2}$$

3-R2 Score

R2 can be used to measure the effect of the model, which measures the improvement of the effect of the model compared with the average index. The closer R2 is to 1, the better the effect of the model is.

$$\mathrm{R2} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

Data Exploration

The data of the project takes Starbucks' preferential push information as the research object, including data transaction data, demographic data and push data. This data set is simplified from the real data of Starbucks app. Because the simulator below only produces one drink, in fact, there are dozens of Starbucks drinks.

The three data sets contain 15 variables in total. The specific information is as follows:

	table	var	data type	meaning
1	portfolio	id	(string)	offer id
2	portfolio	offer_type	(string)	type of offer ie BOGO, discount, informational
3	portfolio	difficulty	(int)	minimum required spend to complete an offer
4	portfolio	reward	(int)	reward given for completing an offer
5	portfolio	duration	(int)	time for offer to be open, in days
6	portfolio	channel	(list of	combination of four push channels

		s	strings)	
7	profile	age	(int)	age of the customer
8	profile	became_member_on	(int)	date when customer created an app account
9	profile	gender	(str)	gender of the customer (note some entries contain 'O' for other rather than M or F)
10	profile	id	(str)	customer id
11	profile	income	(float)	customer's income
12	transcript	event	(str)	record description (ie transaction, offer received, offer viewed, etc.)
13	transcript	person	(str)	customer id
14	transcript	time	(int)	time in hours since start of test. The data begins at time t=0
15	transcript	value	(dict of strings)	either an offer id or transaction amount depending on the record

Table 1: the variables

Assess

After analysis, problems of each data table are found:

portfolio

- channels should be split into 4 columns 'web', 'email', 'mobile', 'social'；
- offer_type should be encoded with one-hot;
- id is too complex, should be reencoded，and change the name to 'promotion_id'；

profile

- gender has nan values；
- age has nan values(118)；
- became_member_on is date, the data format is not right；
- income has nan values；
- id is too complex, should be reencoded，and change the name to 'customer_id'；

transcript

- person and offer_id should be transformed into 'customer_id' and 'promotion_id'；
- Split data stored in dictionary format；

- 'event' contains 4 events, which should be split into 4 tables--'offer received', 'offer viewed', 'transaction', 'offer completed'；

Visualization

Q1: How many users are there? How many users have received push messages? How many users have viewed the push information? How many users have purchase records? How many users have completed the push offer?

There are **17000** customers in total;
16994 customers received the coupons, **99.96%** of total customers;
16834 customers viewed the coupons, **99.06%** of total received customers;
16578 customers used the coupons, **98.48%** of total viewed customer;
12774 customers completed the coupons, **77.05%** of total transaction customers;

As calculated above for each step of funnel conversion, the conversion rate of each step is very good, but the last step needs to be improved;

Q2: What is the average number of consumption per user?

The average consumption times is **8.38;**
The largest consumption times is **36**;

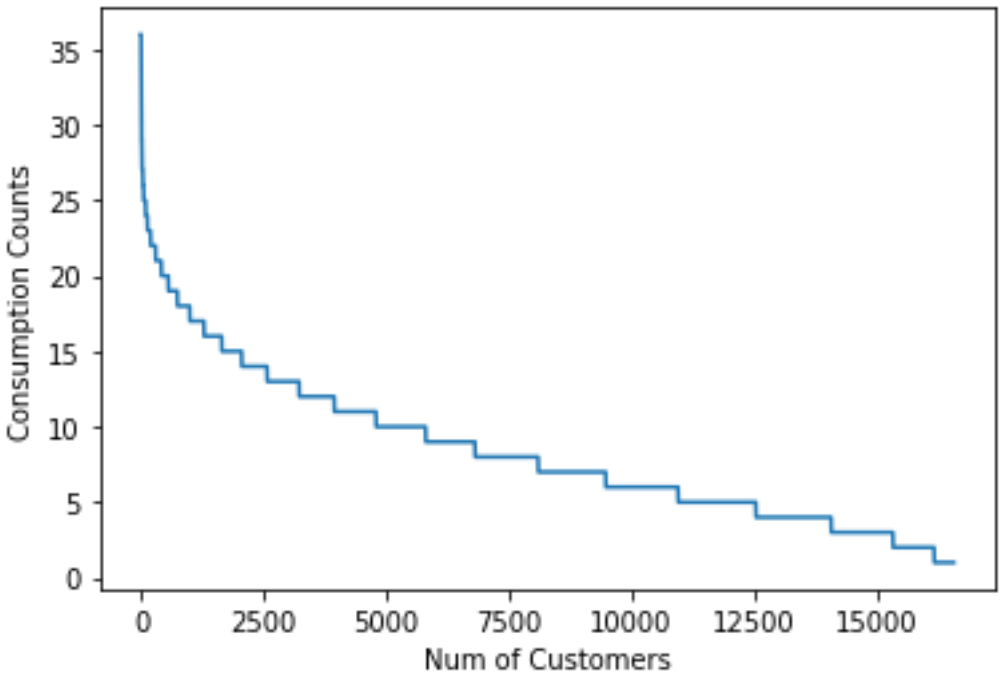

Fig 1: the distribution of consumption times

Q3: What is the return rate of consumers? What are the return rates for different offer coupons?

the average return rate of users is **97.58%,** that means 97.58% of consumers will come back for consumption during the experiment. But under different offer strategies, the proportion will change:

	promotion_id	return_rate
0	4	97.31
2	10	97.36
7	3	97.36
1	5	97.61
4	2	97.65
3	7	97.68
9	8	97.70
6	6	97.78
5	9	97.81
8	1	97.84

Table2: Details of Return Rate of Each Offer

Obviously, the first strategy has the highest return rate if Starbucks wants customers to come back faster. This strategy is undoubtedly a better way.

Let's take a look at the first strategy:

	reward	difficulty	duration	offer_type	web	email	mobile	social	promotion_id
0	10	10	7	bogo	0	1	1	1	1
1	10	10	5	bogo	1	1	1	1	2
2	0	0	4	informational	1	1	1	0	3
3	5	5	7	bogo	1	1	1	0	4
4	5	20	10	discount	1	1	0	0	5
5	3	7	7	discount	1	1	1	1	6
6	2	10	10	discount	1	1	1	1	7
7	0	0	3	informational	0	1	1	1	8
8	5	5	5	bogo	1	1	1	1	9
9	2	10	7	discount	1	1	1	0	10

Table3: Details of Each Offer

So, it’s a BOGO strategy and use ‘e-mail, mobile and social’ to touch users.

Q4: What is the average consumption level of each voucher?

As shown below, the mean GMV of promotion 5 is the most largest one, and it's 13 which

means under this strategy, the average consumption per consumer is $13, which is nearly $0.50 higher than the minimum of $12.54. Considering the user scale of Starbucks, this strategy will greatly increase the sales amount of the company.

	promotion_id	mean_gmv
4	2	12.54
6	6	12.63
3	7	12.70
0	4	12.73
8	1	12.75
5	9	12.77
2	10	12.82
7	3	12.82
9	8	12.94
1	5	13.00

Table4: Details of Average GMV of Each Offer

Q5: With the change of age, gender and membership time of users, what is the change of return rate and consumption level?

1-as AGE changes:

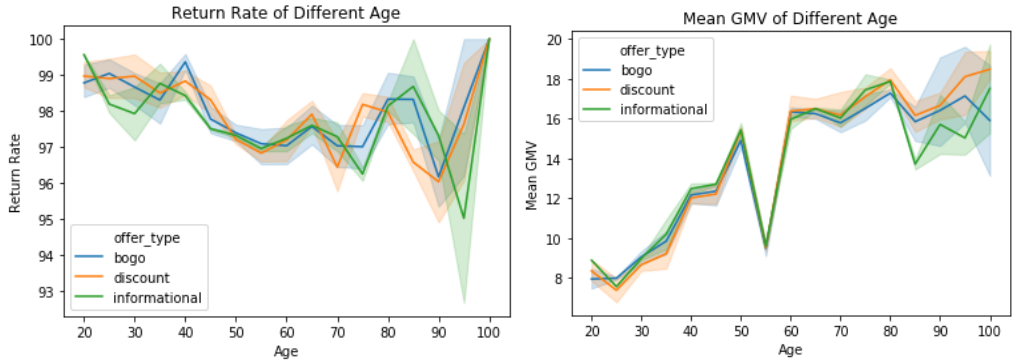

Fig 2: the changing of Return Rate and Mean GMV with different ages

With the increasing of age, the return rate of consumers begins to decline, but the consumption quota will increase each time(in addition to the 55 year old data, the missing age was filled with an average of 55 years, which led to this situation), and the consumption quota of single consumption is the highest at the age of 80 or so. There is not much difference in the reflow rate of consumption among different age groups;

2-as Gender changes:

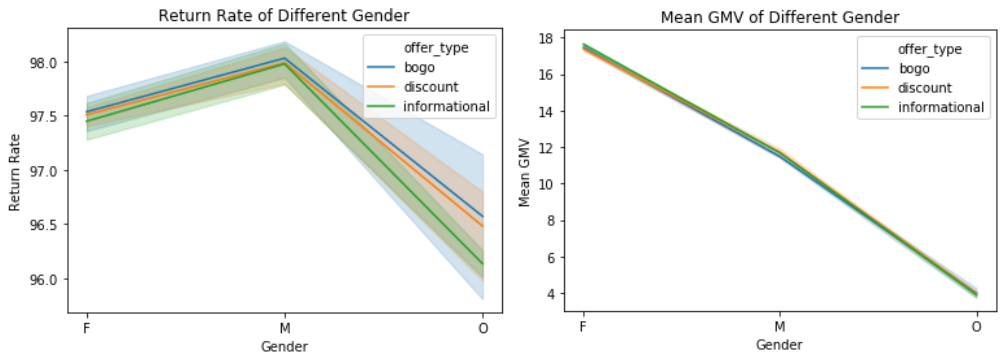

Fig 3: the changing of Return Rate and Mean GMV with different gender

Male consumers have higher reflow speed, while female consumers have higher single average consumption;

3-Member Duration:

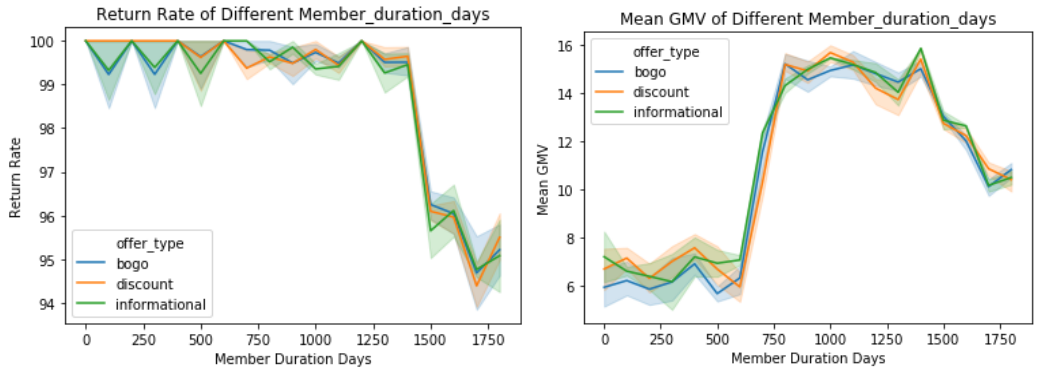

Fig 4: the changing of Return Rate and Mean GMV with different member duration

The speed of backflow decreased significantly after becoming a member about 1400 days; the average consumption level of members from 750 days to 1400 days was the highest;

Data Preprocessing steps

1-outliers removing

Abnormal value of income and consumption quantity:

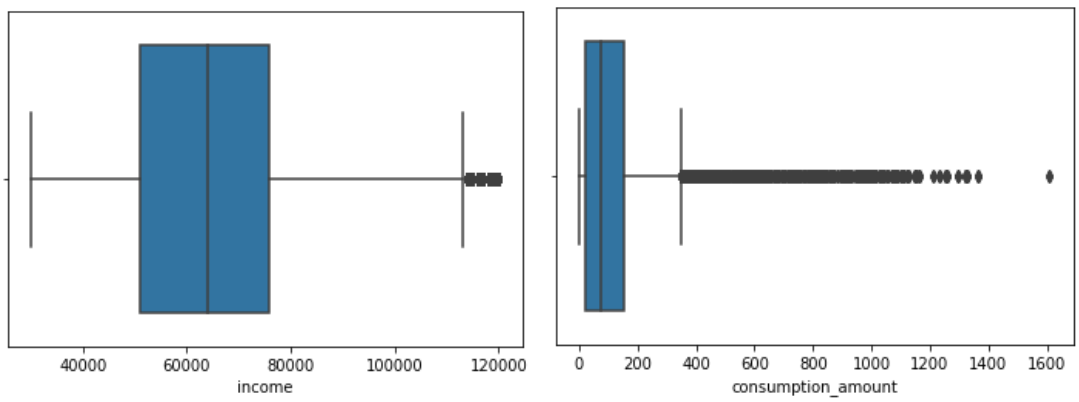

Fig 5: the outliers of Income and Consumption Amount

Using IQR standard to eliminate the abnormal of two variables. And we only consider the customers with positive consumptions which means we will ignore customers who don't consume.

2-distrbution of average consumption amount

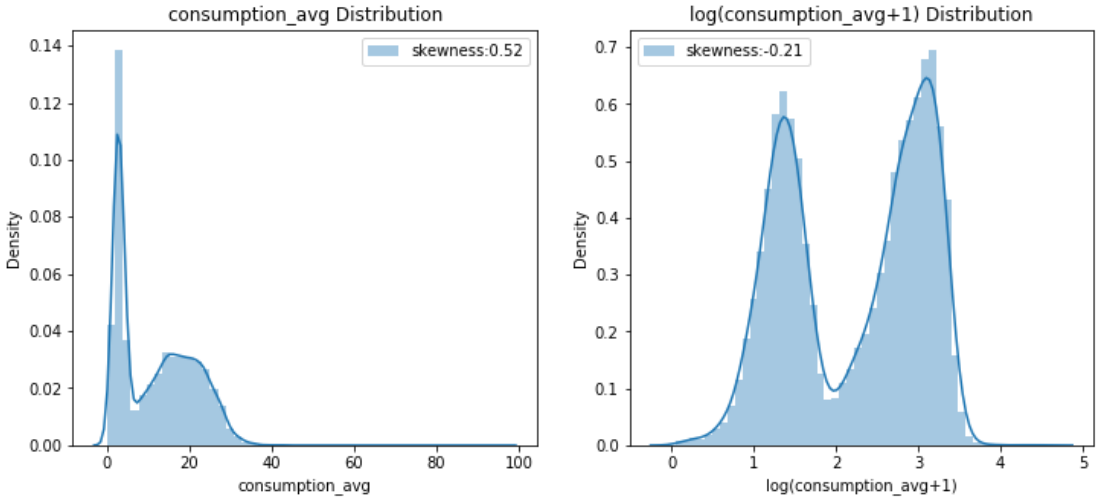

Fig 6: the distribution of log consumption_average

Obviously, there are two different average consumption patterns， and if we use log1p method , we can reduce the skewness of average consumption.

3- Feature selection and feature transformation

First, we need to combine consumer personal information, offer information and transaction information, then handling features in different ways.

- Using one hot coding technology to deal with gender and offer type features
- Convert the time of membership to days, that is, the time of membership minus the minimum membership date in the data set.

The final selected features are: **['age', 'income', 'reward', 'difficulty', 'duration', 'web', 'email', 'mobile', 'social', 'received_cnt', 'viewed_cnt', 'gender_F', 'gender_M',**

'offer_type_bogo', 'offer_type_discount', 'offer_type_informational', 'member_duration_days', 'consumption_avg'].

Implementation

1- Training benchmark model

We use linear model and random forest training as the benchmark:

Model	test_data RMSPE	R2
LinearRegression	1.47	0.61
RamdomForestRegressor	1.16	0.86

Table 5: the benchmark model results

The importance ranking of features can be obtained from the random forest model:

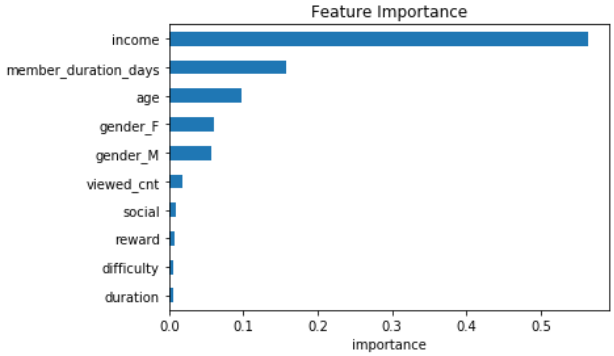

Fig 7: the 10 most important features

It can be concluded from the importance of features that the most important feature is the income level of consumers, followed by the length of time to become a member. For offer information, the number of views is the most important. **It implies that we should reach users as many times as possible, and of course, we should not let users get bored**

After the training set and test set are divided, the initial xgboost model is trained respectively, and the parameters are adjusted based on the initial model. All training results are compared with the results of the benchmark model, and the results are as follows:

Model	test_data RMSPE	R2
XGB-learning_rate:0.03	1.04	0.77
XGB-learning_rate:0.025	1.036	0.73
XGB-learning_rate:0.02	1.017	0.66

Table 6: the XGB model results

With the decrease of learning rate, although rmspe will decrease, R2 score will also decrease rapidly. So at the 0.03 learning rate may be a good choice which has a higher rmspe score but the r2 score is not too bad.

Results

1-Return rate and mean GMV

BOGO and discount can improve reflow speed, but they have no significant effect on the average consumption; BOGO has more advantages in promoting consumer return;

If the goal of the business is to increase the probability of consumer return, BOGO is a more correct choice;

2-Features affecting the average consumption level of consumers

Observe the importance of the characteristics of the final model output, and find that the important characteristics can be divided into income, membership, gender, offer. Combined with the previous data exploration。

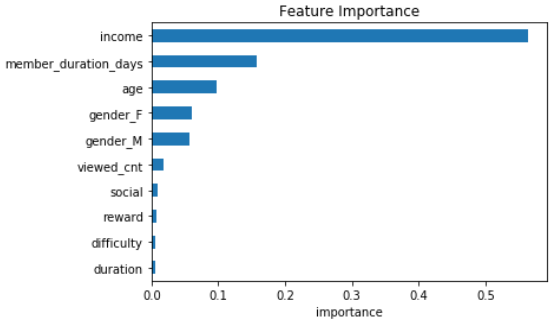

Fig 8: the 10 most important features

we can suggest different stores to find the best strategy according to their own reasonable experiments. **To find high-income people and make them become members**, or to open stores near **high-end business places**, will be a good strategy to try.
At the same time, we should pay attention to **female consumers**, because their average consumption level is higher.

3- Model assessment

By comparison, the most improved scheme of rmspe is to use all features, and set the learning rate to 0.02, and the rmspe on the test set is 1.017. But the r2 score is too low, so finally I choose the XGB with learning rate at 0.03, because the rmspe is not bad and r2 score is higer.

Model	RMSPE	R2
LinearRegression	1.47	0.61
RamdomForestRegressor	1.16	0.86
XGB-learning_rate:0.03	**1.04**	**0.77**
XGB-learning_rate:0.025	1.036	0.73
XGB-learning_rate:0.02	1.017	0.66

Table 7: the model results

Obviously, the model can not predict the personal consumption level of consumers very well, but the importance of the characteristics obtained from the model can give a lot of important reference information.

Future Improvements

1-ROI

Back to the original question, we need to find a better strategy through experiments. We also need to see how much it **costs for each offer**, then we can calculate ROI for each offer. ROI is also a good standard for selecting strategies. If we can collect these data, we can better answer which strategy is better

2-More features

Different stores face different **competition situations**, and the information of competitive products is not considered in this project, and the distance and preferential strength of competitive products are very important, so the promotion means of competitors should be paid enough attention to.

From the results of the benchmark model of random forest, we can see that the gap between the benchmark model and the final model is not particularly "large".

This project only uses the original data set, but **external data** will also have an impact on the sales volume of the store, such as: **weather, new competitive products, online shopping and other factors**. This project has not been expanded.

Reference

[1] Research and analysis of integration tree algorithm based on boosting
http://xueshu.baidu.com/usercenter/paper/show?paperid=f573a4a57b8b565fdaf16d61eee932fe&site=xueshu_se

[2] Data smoothing（log1p and exmp1）
https://blog.csdn.net/u012735708/article/details/84067992

[3] Brief introduction of xgboost algorithm principle and parameter adjustment
https://www.analyticsvidhya.com/blog/2016/03/complete-guide-parameter-tuning-xgboost-with-codes-python/

[4]XGBoost introduction
https://xgboost.readthedocs.io/en/latest/tutorials/model.html