Prediction of Shopper Loyalty - Solution for Kaggle's Acquiring Valued Shoppers Challenge

Madhumathi K IMT2012020
Tanmayee Narendra IMT2012046
Tarun Tater IMT2012047

December 16, 2015

Abstract

In this paper, we present our approach for predicting which shoppers will become loyal to a particular product, given their prior transaction history for a year.

1 Introduction

Consumer brands often offer discounts to attract new shoppers to buy their products. The most valuable customers are those who return after this initial incented purchase. With enough purchase history, it is possible to predict which shoppers, when presented an offer, will buy a new item. However, identifying the shopper who will become a loyal buyer is a more challenging task.

The *Acquire Valued Shoppers Challenge* asked participants to predict which shoppers are most likely to repeat purchase. This International challenge on the popular internet platform Kaggle[1] continued for 95 days from 10th April 2014 to 14th July 2014, and attracted 952 active participants. Area under receiver operating curve (AUC) was used for the evaluation.

Complete, basket-level, pre-offer shopping history for a large set of shoppers who were targeted for an acquisition campaign was provided. The incentive offered to that shopper and their post-incentive behavior was also provided.

The data captured the process of offering incentives (a.k.a. coupons) to a large number of customers and forecasting those who will become loyal to the product. Let's say 100 customers are offered a discount to purchase two bottles of water. Of the 100 customers, 60 choose to redeem the offer. These 60 customers were the focus of this competition. The task was to predict which of the 60 will return (during or after the promotional period) to purchase the same item again.[1]

[1] https://www.kaggle.com

2 Description of Dataset

Four files in Comma Separated Values format were provided.

- `transactions.csv` - transaction history for all customers for a period of at least 1 year prior to their offered incentive
- `trainHistory.csv` - the incentive offered to each customer and information about the behavioral response to the offer
- `testHistory.csv` - the incentive offered to each customer but does not include their response (we are required to predict the repeater column for each id in this file)
- `offers.csv` - information about the offers

The following tables describe the fields in the dataset.

2.1 History

The `trainHistory.csv` file (991 kB) has the following features.

Attribute	**Description**
id	A unique id representing a customer
chain	An integer representing a store chain
offer	An id representing a certain offer
market	An id representing a geographical region
repeattrips	The number of times the customer made a repeat purchase
repeater	A boolean, equal to repeattrips >0
offerdate	The date a customer received the offer

The `testHistory.csv` file has the all features in `trainHistory.csv` except repeattrips and repeater.

2.2 Transactions

The `transactions.csv` file (22 GB) has the following features.

Attribute	**Description**
id	A unique id representing a customer
chain	An integer representing a store chain
dept	An aggregate grouping of the catogory (Eg water)
category	The product catogory (Eg sparkling water)
company	An id of the company that sells the item
brand	An id of the brand to which the item belongs
date	The date of purchase
productsize	The amount of the product purchase (e.g. 16 oz of water)
productmeasure	The units of the product purchase (e.g. ounces)
purchasequantity	The number of units purchased
purchaseamount	The dollar amount of the purchase

2.3 Offers

The `offers.csv` file (431 B) has the following features.

Attribute	Description
offer	An id representing a certain offer
category	The product category (Eg sparkling water)
quantity	The number of units one must purchase to get the discount
company	An id of the company that sells the item
offervalue	The dollar value of the offer
brand	An id of the brand to which the item belongs

There are totally 37 offers.

3 Analysis of the Dataset

Essentially, the challenge is a binary classification problem – once a shopper has redeemed a particular offer, we are required to predict whether or not the shopper will buy the same product again (irrespective of whether the offer is still valid or not). There are two classes – those who buy the same product again, and those who don't. This feature is captured by the `repeater` feature of `trainHistory.csv`. If the shopper has bought the same product several times, the `repeatTrips` column will be greater than 1. Consequently, `repeater` colunm will have a value of 1. (`repeater` column is of boolean type)

To create this prediction, a minimum of a year of shopping history prior to each customer's incentive was given, as well as the purchase histories of many other shoppers (some of whom received the same offer). The transaction history contains all available items purchased, not just items related to the offer. **Only one offer per customer is included in the data**. The training set is comprised of offers issued from 1st March to 30th April 2013. The test set is offers issued from 1st May to 31st July 2013. (See Figure 1)

This challenge provided 349,655,789 rows of completely anonymised transactional data (data stream) from over 300,000 shoppers: 160,057 shoppers in the training set (among them 43455 or about 27.15% are loyal) and 151,484 shoppers in the test set.

`Category`, `company` and `brand` represent the most important characteristics in the transaction data. We defined each 'product' as a combination of these three features. Basically, current information is encoded in `trainHistory.csv`. The transactions history file should be used to derive additional features based on the shopper's past transactions.

Every record in `testHistory.csv` and `trainHistory.csv` was augmented the details of the offer (company, brand and category of the offer) from the offers.csv file. In addition, we generated the following secondary features.

1. numCompany – number of times shopper has bought from that company
2. numCategory – number of times shopper has bought from that category

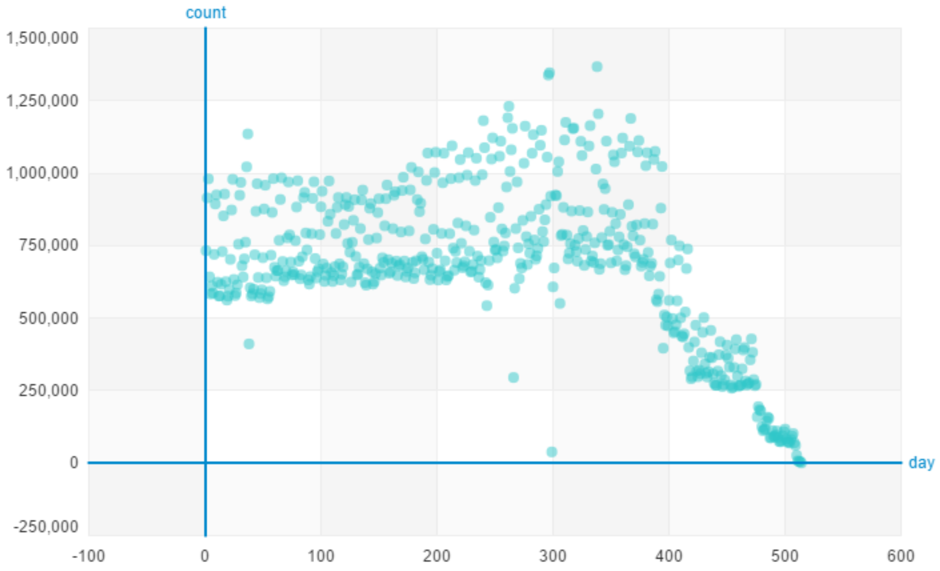

Figure 1: Number of transactions versus time

3. numBrand – number of times shopper has bought from that brand
4. quanCompany – quantity that the shopper has bought from that company
5. quanCategory – quantity that the shopper has bought from that category
6. quanBrand – quantity that the shopper has bought from that brand
7. amountCompany – amount that the shopper has bought from that company
8. amountCategory – amount that the shopper has bought from that category
9. amountBrand – amount that the shopper has bought from that brand
10. hasNeverBoughtCompany – boolean value signifying if shopper has previously never bought from particular company
11. hasNeverBoughtBrand – boolean value signifying if shopper has previously never bought from particular brand
12. hasNeverBoughtCategory – boolean value signifying if shopper has previously never bought from particular category

13. hasBoughtCompanyBrand – boolean value signifying if shopper has previosuly bought from both the company and brand
14. hasBoughtCategoryBrand – boolean value signifying if shopper has previosuly bought from both the category and brand
15. hasBoughtCompanyCategory – boolean value signifying if shopper has previosuly bought from both the category and company
16. hasBoughtCompanyBrandCategory – boolean value signifying if shopper has previosuly bought from category, company and brand
17. category180 – number of times shopper has bought from that category in the past 180 days
18. category150 – number of times shopper has bought from that category in the past 150 days
19. category120 – number of times shopper has bought from that category in the past 120 days
20. category90 – number of times shopper has bought from that category in the past 90 days
21. category60 – number of times shopper has bought from that category in the past 60 days
22. category30 – number of times shopper has bought from that category in the past 30 days
23. category15 – number of times shopper has bought from that category in the past 15 days
24. brand180 – number of times shopper has bought from that brand in the past 180 days
25. brand150 – number of times shopper has bought from that brand in the past 150 days
26. brand120 – number of times shopper has bought from that brand in the past 120 days
27. brand90 – number of times shopper has bought from that brand in the past 90 days
28. brand60 – number of times shopper has bought from that brand in the past 60 days
29. brand30 – number of times shopper has bought from that brand in the past 30 days

30. brand15 – number of times shopper has bought from that brand in the past 15 days
31. company180 – number of times shopper has bought from that company in the past 180 days
32. company150 – number of times shopper has bought from that company in the past 150 days
33. company120 – number of times shopper has bought from that company in the past 120 days
34. company90 – number of times shopper has bought from that company in the past 90 days
35. company60 – number of times shopper has bought from that company in the past 60 days
36. company30 – number of times shopper has bought from that company in the past 30 days
37. company15 – number of times shopper has bought from that company in the past 15 days
38. hasNotBoughtAny180 – boolean value signifying if shopper has not bought from either company, brand or category in the past 180 days
39. hasBoughtAll15 – boolean value signifying if shopper has bought from company, brand and category in the past 15 days
40. categoryLoyalty – percentage of transactions made from this category
41. companyLoyalty – percentage of transactions made from this company
42. brandLoyalty – percentage of transactions made from this brand

4 Results

4.1 Evaluation Scheme

For this competetion, the method of evaluation was area under receiver operating curve (AUC). A receiver operating characteristic (ROC), or ROC curve, is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied. The curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. When using normalized units, the area under the curve (often referred to as simply the AUC, or AUROC) is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one (assuming ‘positive’ ranks higher than ‘negative’).[2]

For each id in the testHistory.csv file, we were required to predict a probability that the customer repeat purchased the product from the offer that they received.

4.2 Results

The training data was augmented with the secondary features that were computed. As we said before, 27.15% of shoppers are loyal. This means that there is some imbalance between the two classes.

Initially, we used Support Vector Machines (SVM) to train our classifier. However, the score for this model was low, as SVM is highly sensitive to class imbalance.

We also used Logistic Regression to train our classifier, and the accuracy that was obtained is 0.42948.

Last, we used random forests to create a model from the augmented training data. Random forests are known to be resilient against class imbalance, as we can specify the skewness in the training data before the model is built. The accuracy that was obtained was 0.56121. (The highest accuracy that was obtained in the challenge is 0.62703) For all the above approaches, we used the *graphlab* library in python.

5 Conclusion

The primary challenge that we faced was the huge size of the dataset. Since the transactions history file was almost 22 GB, it took us some time to optimise our methods so that we could extract our secondary features in a feasible amount of time. Also, we realised that the RAM of the computer played a significant role in influencing the time taken to extract features. For example, on a machine with 12 GB RAM, the time taken for extraction of features was 160 minutes, whereas a machine with 6 GB RAM took 10-11 hours. Since we did not have unlimited access to the machine with 12 GB RAM, we had to make do with our own laptops, which has significantly lesser RAM. This greatly increased the time taken for our computation.

The accuracy of the winning solution is 0.62703, which is not very much higher than the accuracy that we were able to achieve, considering the limited resources at our disposal.

6 Acknowledgements

We would like to thank our friends for generously lending their laptops, without whom we wouldn't have been able to extract features as quickly.

References

[1] "Acquire valued shoppers challenge - predict which shoppers will become repeat buyers." https://www.kaggle.com/c/acquire-valued-shoppers-challenge. Accessed: 2015-12-14.

[2] "Reciever operating characteristic." https://en.wikipedia.org/wiki/Receiver_operating_characteristic. Accessed: 2015-12-14.

[3] V. Nikulin, "On the method for data streams aggregation to predict shoppers loyalty," in *Neural Networks (IJCNN), 2015 International Joint Conference on*, pp. 1–8, July 2015.