Black Friday Sales Prediction

Rui Cao
University of British Columbia
rui.cao@alumni.ubc.ca
UBC ID: 99168981

Li Wang
University of British Columbia
lwang@math.ubc.ca
UBC ID: 87075140

Abstract

Customer behaviour is one of the most popular academic topics in Marketing and it is widely used in industry as well to improve decision making process. According to the sales data, people's purchase intentions reach the peak on Black Friday, largely because of the social norms and low prices. Hence, it is significantly crucial to predict customer behaviours during Black Friday sales to largely improve companies profits. We got Black Friday sales data from Kaggle and it made 537577 observations for customer behaviours on Black Friday in a retail store. We developed different methods to predict customer age and the kinds of products that customers will purchase. The performance of the selected models are relatively good.

1 Introduction

Understanding customer behaviour to improve decision making process has been a popular and crucial topic to industry and academy. Large amount of books and articles analyze customer behaviour from different aspects, and a book example can be [1]. In this book, the MIT behavioural economist Dan Ariely refutes the most common assumption in marketing that people behave in fundamentally rational ways. In this book, he explained how seemingly illogical forces skew our reasoning abilities. Similarly, [6] used the "disordered" customer behaviours on Black Friday to study the social and cultural values. According to [6], the social norms, expectations and many other factors boost customers' purchase intentions. Indeed, the crazy shopping phenomenon gets more severe every year and during 2018 Thanksgiving, US consumers reached the highest U.S. e-commerce sales day in history with $\$7.9$ billion in revenue [4]. Because of the huge amount of sales, it is significantly important for sellers to understand customer behaviour during Black Friday, and take action based on the predictions they made for those "crazy shoppers" to have even more sales.

As firms scrutinize their marketing strategies and outcomes, they began to use the capacity of data mining [7]. Also, there are more and more researchers use machine learning methods to understand customer behaviour in general. For example, [9] used clustering to improve prediction accuracy and reduce computational cost. [2] used non-parametric regression for short-term load prediction. More examples will be given in the next section.

Based on the importance of customer behaviour and the large sales potential on Black Friday, we use several machine learning models and different factors to predict the customer behaviour on

Preprint. Work in progress.

Black Friday. Our data comes from Kaggle, which has 12 customer features including ages, years of residence in the city, job position, etc. We would like to explore the correlations among the above features, so from where we can use machine learning models to make predictions. Specifically, in this research, we focus on the prediction of customer age and product category.

Based on our acknowledge, the sales prediction on Black Friday is seldom studied in previous literature. Our work can be an insight for companies to improve their Black Friday advertisements. We believe that, improve the marketing strategy on Black Friday will yield the most powerful effects compared with on other time, given that Black Friday can stimulate the highest customer purchase intention [4][6]. Combined with other customer behaviour studies using different research aspects, such as online reviews, web browser history and social media "like" records, this study can contribute to the field using machine learning prediction.

This paper is organized as follow. Section 2 talks about related work. In the Data Analysis section, we describe our data and illustrate the models we tried to make the prediction. It also talks about the optimal model we chose. Section 4 discusses future work.

2 Related Work

There are some related works applying machine learning methods in previous literature. [5] shows that data mining techniques can extract respectable knowledge from the customer's database and this article examined how to analyze customer behaviour to improve business performance. The model proposed by [5] helps to enhance the customer satisfaction and interaction more easily for the company. Thus, we hope our prediction would help the retail businesses to improve their performance by analyzing their black Friday sales data. [3] uses PCA and BP neural network to implement customer relationship management, and gives a set of customer segmentation index system by analyzing retail-business consumer behaviour. The customer segmentation [3] consists of three components: the basic situation of consumers, payment method, and attitude. [8] uses decision tree and multi-layer neural network to analyze click-streams of e-customers and extract information, then makes prediction on a digit market. By collecting the data about the customers' movements and their demographic information, [8] extract the online customers' behaviour patterns, and found that category of product is more important to men than women. The above two references provide us some clues about the choice of machine learning models.

3 Data Analysis

The data set has 537577 data in total. And there are 12 features corresponding to each data, which are user ID, product ID, gender, age groups, occupation, city category, years of residence in current city, marital status, three product categories, and total purchase amounts.

In order to do data analysis, we first did data cleaning. User ID and product ID are combinations of random numbers and we did not need them, so we deleted these two columns. Gender is written as "F" and "M". We changed "F" to 0 and "M" to 1. Age has 6 groups with different age ranges. For example, one group is written as $0 \sim 17$ meaning that this person's age is between 0 to 17. We changed the group from specific ages to numbers. So the exampled person is transferred from group $0 \sim 17$ to group 1. Occupation is a categorical variable that is written by numbers, and different number implies different jobs. City category is written as "A, B, C" etc. to represent different cities, and we changed it to numbers using $1, 2, 3 \ldots$. The residence in the current city explains how many years does the person live in current city. In this category, groups are named by $1, 2, 3, 4+$ and we changed $4+$ to 4. In marital status category, 0 is for not married and 1 is for married. Then we have 3 columns representing product category $1, 2$ and 3. Inside each category, the data uses different numbers to represent different products that customers bought. Notice that the number 4 item in category 1 is not the same with number 4 item in category 2. The final column is the total purchasing value for each customer.

Then we replaced the sparse spots with number 0. The sparsity is large in product category 2 and 3. And for each customers, it is possible to purchase different items in one category. So one customer may appear several times in the data set. Consider that, there are in total 5891 customers.

3.1 Data Visualization

We first tried to visualize data by compressing data using both PCA and dimensional ISOMAP. Because of the sparsity of product category 2 and 3. We removed these two features in our data. By randomly choosing the 10000 data from our data set, the PCA variance of the first two components are $9.99997580e-01$ and $1.72858695e-06$, which we can see in the figure that the vector is pointing one directions. If we apply the ISOMAP 2 components and K-means clusters. From the results of our clustering, we expect that the clustering models wouldn't be a good candidate for our data.

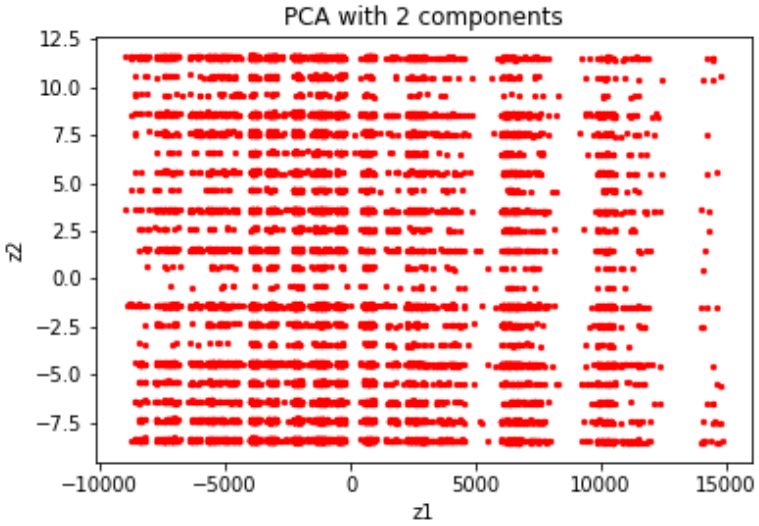

Figure 1: PCA with 2 components

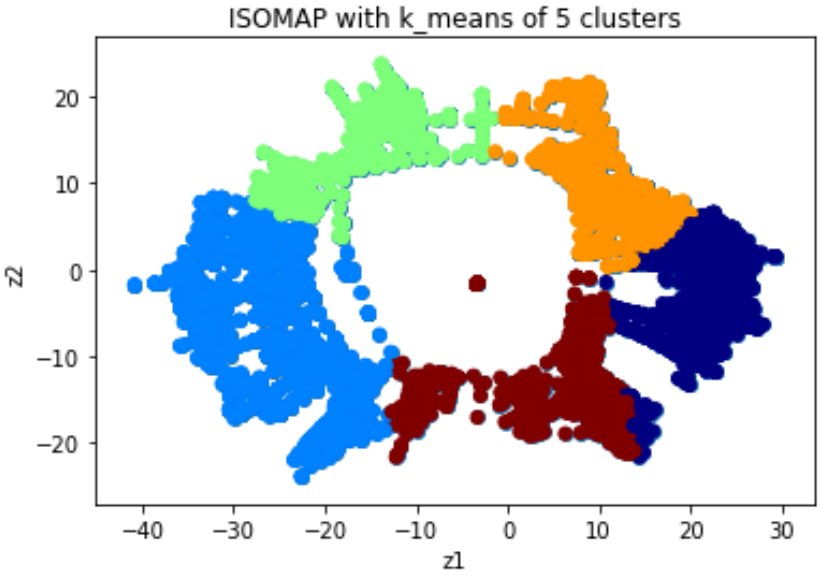

Figure 2: ISOMAP with K-means of 5 clusters

3.2 Prediction

In order to train model and make prediction, we divided our data into training set and validation set using scikit-learn build-in function. The test size we used was 0.2. Hence, we have 107516 testing examples and 430061 training examples. And we had different X matrix and y matrix based on different model purposes.

3.2.1 Product Category 1

To understand which product in category 1 are mostly likely bought by the customers based on their other information, we tried to predict the product category 1 by first dropping the feature of category 2 and 3, and we trained the data based on the other features of data. We used 5 folds cross validation here. We have tried the following models: decision tree, random forests, KNN, and softmax. Decision tree and random forests do a better job than the other twos. According to the decision tree results shown in Figure 3, we chose decision tree with depth 11, and we had the training error as 0.127 and the validation error was 0.131. Notice that in this figure, "testing error" is "validation error". Random forest with 5 features and 5 trees has training error of 0.015, and the validation error is 0.154. However, KNN with metric of cosine is not a good model for our case, which has training error 0.363, and validation error 0.496. Softmax has Training error of 0.651 and Validation error of 0.652. We also tried polynomial kernel, and Gaussian kernel, both does not give us satisfying results. Our prediction provides information to the sellers about what kinds of products in category 1 should they recommend to the customers based on customers' demographic, social and economic information.

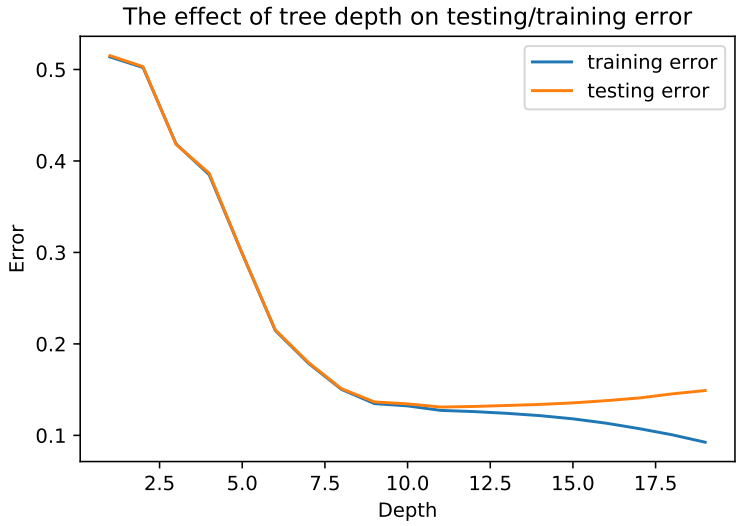

Figure 3: Effect of Tree Depth on Validation and Training Error on product prediction

3.2.2 Age

Secondly, we turned our eyes on age factor, which is the only continuous feature for customers. In the original data, the age is divided into 6 groups, with group 1 being people with age $0 \sim 17$, group 2: $18 \sim 25$, group 3: $26 \sim 35$, group 4: $36 \sim 45$, group 5: $46 \sim 54$ and group 6 includes people above 55 years old.

In order to predict the age, we need to decide what model to use. The first model we tried was KNN, which works effectively with large set of data. We tried 3 different k values which are $3, 5$, and 10. We found that the training error is the lowest with $k = 3$, but the validation error is relatively low with 0.669. And with $k = 10$, we have a good training error but low validation error. Next, we tried random forest using information gain. We first set the number of trees to be 10. And we got the result that training error is 0.037 and validation error is 0.412. Then we changed the value of this

hyper-parameter to be 50 and we got 0.008 training error and 0.410 validation error. With this sign of overfitting, we changed the number of trees to be 30, and we got a higher validation error but the validation error did not change. We figured out that this seemed like to be the most optimal result of random forest model and we stopped here.
Furthermore, we switched our strategy to use decision tree. Similar to the prediction on Product category, we tried different depths of the tree trying to find the most optimal one. And the result is shown on Figure 4. Notice that in this figure, "testing error" is "validation error".

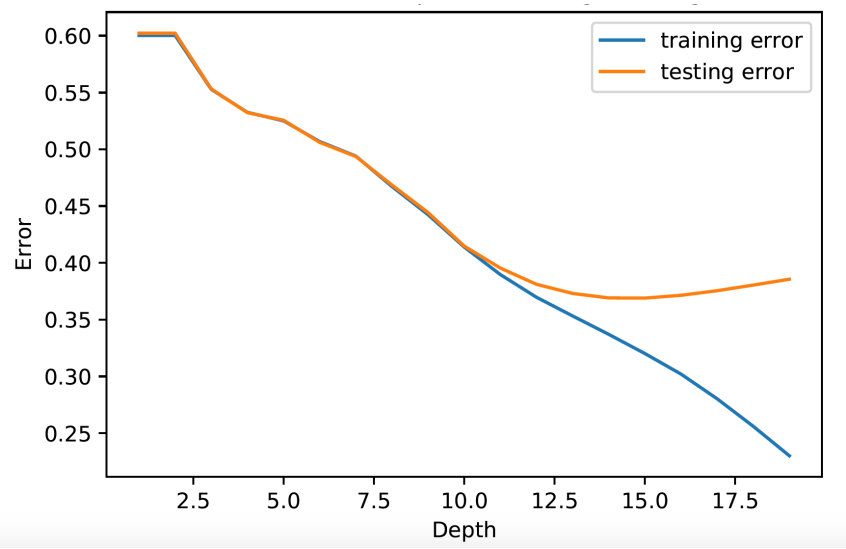

Figure 4: Effect of Tree Depth on validation and Training Error on age prediction

From this figure, we can see that the training error continuously to decrease as the depth increases, and the validation error decreases until depth 13 then increases. Therefore, here we chose depth 13 and this model has 0.352 testing error and 0.373 validation error.

Compared among KNN, decision tree and random forest, we chose decision tree with depth 14 to be our model to predict age, and the predictors are all the remaining factors except user ID and product ID.

3.2.3 Purchase

We further tried to predict purchase values using linear and polynomial model. However, the RMSE value for linear model is over 4000. To evaluate the performance of polynomial model, we calculate $\frac{\sqrt{(\text{testing error} - \text{validation error})^2}}{\text{testing error}}$ for each data. And the model overall yields more than 170% errors for both training and testing errors. We can see here that both models made bad predictions. So for this data, the purchase values cannot be predicted by linear model and polynomial model.

4 Future Work

We made prediction on what kind of product in the product category 1 customers would purchase on Black Friday by knowing their ages, occupations, and residence of the city, and their budget on shopping for Black Friday. In addition, we are able to make prediction on the customers ages based on their shopping behaviours, and occupations, and residence status of the city. Since there are total 12 unique products in the category 1. We have got a relative good prediction on the result of predicting the products, which is around 90% accuracy. However, one of the weaknesses is that we did not provide several possible products that a customer would like to buy on Black Friday. Because from the perspective of the retail business, in order to have the best revenue, they usually would recommend several products to their customer.

For future work, we would like study how the total purchase is related to the customers segmentation, which is based on behaviour,demographic, and lifestyle variable. Our current data does not provide us the lifestyle variable and behaviours. So with more features, we expect that we have better prediction on the customers purchase power.

References

[1] Dan Ariely. Predictably irrational: The hidden forces that shape our decisions. 2007.

[2] W. Charytoniuk, M. S. Chen, and P. Van Olinda. Nonparametric regression based short-term load forecasting. *IEEE Transactions on Power Systems*, 13(3):725–730, Aug 1998.

[3] M. Han. Customer segmentation model based on retail consumer behavior analysis. In *2008 International Symposium on Intelligent Information Technology Application Workshops*, pages 914–917, Dec 2008.

[4] Marcia Kaplan. Sales report: 2018 thanksgiving, black friday, cyber monday. *PracticalEcommerce*, 11 2018. https://www.practicalecommerce.com/sales-report-2018-thanksgiving-black-friday-cyber-monday.

[5] A. A. Mudimigh, F. Saleem, and Z. Ullah. Efficient implementation of data mining: Improve customer's behaviour. In *2009 IEEE/ACS International Conference on Computer Systems and Applications*, pages 7–10, May 2009.

[6] Thomas Raymen and Oliver Smith. What's deviance got to do with it? black friday sales, violence and hyper-conformity. *The British Journal of Criminology*, 56(2):389–405, 2016.

[7] Esa Rinta-Runsala. Bringing data mining to customer relationship management of every company. *VTT Technical Research Centre of Finland*, 2004.

[8] G. Sılahtaroğlu and H. Dönertaşli. Analysis and prediction of amp;x0395;-customers' behavior by mining clickstream data. In *2015 IEEE International Conference on Big Data (Big Data)*, pages 1466–1472, Oct 2015.

[9] X. Wang, M. Zhang, and F. Ren. Learning customer behavior for effective load forecasting. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–1, 2018.