An Analysis of the Decision making in Online electronics shopping: evidence from Indian market

Introduction:

Online shopping has been one of the most prevailing shopping modes, especially for those who are always occupied with work and school stuff. Being different from the traditional consumption pattern, consumers are unlikely to bargain with the providers and hence they become the "price takers" in this process. Besides comparing the prices of the same product across multiple online shopping websites, being able to fully realize the trends of price change of target products is a prerequisite for making a wise purchasing decision. Once consumers notice that the price will go up in the next time period, then it will be better to purchase at the current stage. Therefore, this research will be able to provide a plausible model in predicting the price changes of the electronics for the online consumers.

Meanwhile, by investigating the effect of many factors other than the nature of the product itself, such as average rating of the products, whether having free shipping service, list prices on the websites, etc., on the price change in next time period, the electronics producers will be able to find out the driving forces of the price changes, and adjust the prices and additional services of their products strategically to reach their business goals based on the analysis. In addition, the dataset being used in this project includes the products' information from 36 brands. Generally, they are competitors in the online electronics market. By analyzing the data from other companies, one can clearly detect the driving forces for price changes in different firms, and make responses to others' behaviors in order to gain more revenue. Thus, the companies which involved in the electronics online shopping industries may need to conduct such an analysis.

Data:

I will use one dataset published by Kaggle called *price change prediction of electronics in Online shopping*, which includes the information of different electronic items on Indian online shopping websites for several months from 2011 to 2012.

(The link: https://inclass.kaggle.com/c/price-change-prediction-of-electronics-in-onlineshopping/data)

The dataset includes the brands, color, shipping methods, stock status, rating, websites where sold, category and price information of the products. Some of these variables can be selected to be the independent variables, in the mean time, whether the price goes up can be characterized as the dependent dummy variable in this analysis. Admittedly, this dataset provides valuable information with respect to the nature of the products and services being offered along with the products. However, there are still some omitted factors that are correlated with the independent variables and are the determinants of the price change, such as the edition of the products and the frequency of the new products release. The edition of the products is strongly correlated with the stock status and the average rating of one product, since a new edition of a product may attract more clients to order (like Iphone 7 recently), which may lead to a lower level of inventory stock; and may also cause an insufficient information regarding average rating. Therefore, these factors can be categorized as omitted variables in this project. The failure of including these factors in the dataset tends to give rise to bias in the estimation.

Methodology:

The logistic regression will be employed in this project. The original dataset can been divided into training and test set through a cross-validation way, which benefits measure the validity of the model.I may conduct the the logistic regression model based on the data in the training set, and then predict the price change by using the data in the test set. And then I may exert the accuracy, the sensitivity and the specificity analysis to evaluate the performance of the prediction model in this case. Also, according to the coefficients reported from the estimation, I may conclude the main driving forces for the price change in the next time period.

Data wrangling:

Two types of data wrangling are needed to be done before the estimation. First, all the missing values need to be deleted from the dataset or be converted to the median value of the variable. Because the variables that will be used in this projects are either dummy variables or categorical variables, I will apply the former method to move all the missing values out. Second, the name, the brand and the color of the products are strings. However, the variations of price change arising from these factors cannot be neglected, a fixed effect might be included in the model as well. Therefore, the strings are needed to be converted to factors.

Preliminary Results:

At this stage, I will only use whether having free shipping service and whether the product is in stock to predict the price change.

First, I divided the dataset into two subsets, the training set contains 2/3 samples of the original dataset, while the test set contains the rest 1/3.

```
index_train <- sample(1:nrow(CellPhone),2/3*nrow(CellPhone))
trainingset <- CellPhone[index_train,]
testset <- CellPhone[-index_train,]
```

Second, I executed the logistic regression and used whether the price increased in the next period as the dependent variable.

```
Results_pre <- glm(PriceUp ~ freeShipping+inStock,family="binomial", data=trainingset)
```

Third, I did the prediction based on the model above.

```
predictions <- predict(Results_pre,newdata=testset,type="response")
```

Forth, I summarized the data in predictions and found that the min and max values are 0.0064 and 0.00787, respectively, while the median is 0.248. Hence, I chose 0.03 as the cut-off value.

Last, I created the confusion matrix to show the performance of the prediction.

```
table(testset$PriceUp,predictions>=0.03)
```

The results are as follow:

```
  FALSE TRUE
0  1218 1312
1    17   42
```

However, the validity of this model is questionable, as the extremely low reported values from the prediction. I think I need to reconsider the variables that would be used in this prediction and make efforts to enhance the accuracy of the prediction in my further work.