An Analysis of the Price Change in Online Mobile Shopping

Shubo Zhang

December 12, 2016

Capstone Project

Prepared for Data Science, Springboard

1.Introduction and Background

Online shopping has become one of the most prevailing shopping patterns in recent years. Dramatically different from in-store shopping, customers will not necessarily visit stores to select commodities, they are able to navigate their purchasing process online, as long as they have an electronic device and are accessible to one of the e-payment methods. This feature of online shopping provides extreme convenience to those are occupied with working or schooling stuff and living far from the target stores or in the rural areas. Meanwhile, customers possess more choices of commodities in online shopping. By exploring different online shopping websites, customers may find more options, such as more colors and brands, for a specific type of goods. Furthermore, by replacing the traditional in-store pattern, e-commerce significantly contributes to expenses curtailment. Producers may spend less on retail stores establishment and reduce the operation cost from the online shopping pattern.

However, the qualities of online products are invisible to some extent. Even though the descriptions, prices and the pictures of the goods are always available to check online, consumers sometimes feel difficult to fully know the products, and hence they need external assistance to make a purchasing decision. From my own perspective, there are three types of external assistance that play significant roles in this process. First, ratings and comments from the other users are primarily important. Lacking a sensory of the goods, consumers may review the comments first to ensure the quality of the product and regard the rating results as the reference before making a purchasing decision, even though he/she is starving for the item. If the consumer measures the quality of the product is lower than his/her expectation according to the ratings and comments, then the consumer may change his/her mind. Second, logistic affairs may significantly affect costumers' purchasing decision. Taobao, which is the largest e-commerce platform in China, creates a subsector in the rating system for logistic issues. Customers can rate their satisfaction with the logistics in terms of delivery speed, the service of the couriers, the intactness of the parcel. Many negative comments on that platform are due to logistic dissatisfaction. Therefore, the logistic issues may be a crucial measurement to evaluate the online purchasing experience from the costumers' perspective. Third, the online shopping platform may impose an important effect on the online shopping behaviors. In the U.S. or Canada, Amazon and eBay are the two well-known and common-used online shopping websites. Commodities being sold in those two websites may reach higher sales volumes. However, those two platforms are dominated by the local ones, such

as Taobao and Jingdong in China. Only a small portion of Chinese online customers choose Amazon and eBay, and hence they cannot exhibit such great sales performance there.

Customers unlikely bargain with the producers in the process of online shopping. The price as well as deals are set by the producers in conjunction with platforms and are given on the websites, customers may compare the price from one platform to the other, but they might not directly approach the producers for bargaining. Thus, customers become price takers in online shopping.

These features give rise to an atypical shopping pattern compared to the in-store one, and they potentially impose an enormous impact on the operational performance of the online shopping pattern.

On the other hand, providers and platforms are profit-seekers. The prices of commodities may be adjusted frequently for either clearance or profit making purposes. Normally, firms have their own schedules to make price adjustments according to the products' sales performance, users' feedbacks, rival's pricing strategies, etc. In this analysis, I am trying to provide a plausible model in predicting the price changes of the mobiles for the online consumers. The unique features of online shopping will be taken into consideration. In terms of the variables represents the price, I will count the prices in two time periods, which are the initial time period and the next time when the price information is recorded, to trace whether the prices rise or not. Besides comparing the prices of the same product across multiple online shopping websites, being able to fully realize the trends of price change of the 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 last stage.

Meanwhile, by investigating the effect of many factors along with the nature of the product itself, such as average rating of the products, whether having free shipping service, brands and colors 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.

2.Data Description and Methodology

I will use one dataset published by Kaggle called price change prediction of electronics in Online shopping <https://inclass.kaggle.com/c/price-change-prediction-of-electronics-in-onlineshopping/data>, which includes the information of different electronic items, which are mobiles and cameras sperately, on Indian online shopping websites for several months from 2011 to 2012.The dataset has 7765 obersavations and includes the brands, color, shipping methods, stock status, rating, websites where sold, category and price information of the products.

2.1 Data Wrangling

Before conducting the empirical analysis, I need to work with the raw data to make them usable.

First, since the mobile industry is the only area of interest in this analysis, the objects indicate camera will be neglected. Thus, I drop all the camera variables from the dataset by converting `camera` to missing values.

```
Electronics<- read.csv("/Users/KingsShubo/Desktop/R/train.csv")
Electronics$category[Electronics$category=="Cameras"] <- NA
```

Second, I fill in the blanks with `NA`, and then employ the `na.omit` command to remove all the missing values out of the dataset.

```
Electronics$category[Electronics$category=="Cameras"] <- NA
Electronics$color[Electronics$color==""]<- NA
Electronics$freeShipping[Electronics$freeShipping==""] <- NA
Electronics$inStock[Electronics$inStock==""]<-NA
Electronics$avRating[Electronics$avRating==""]<- NA
Electronics$reviewCount[Electronics$reviewCount==""]<- NA
Electronics$listPrice[Electronics$listPrice==""]<-NA
Electronics$shippingPeriod[Electronics$shippingPeriod==""] <- NA
Electronics$PriceUp[Electronics$PriceUp==""]<- NA
Electronics <- na.omit(Electronics)
Electronics$inStock[Electronics$inStock=="2"]<- "0"
Electronics$inStock<-as.factor(Electronics$inStock)
Electronics$freeShipping[Electronics$freeShipping=="2"]<- "0"
Electronics$freeShipping<-as.factor(Electronics$freeShipping)
```

Third, as instructed on the websites, if the product comes with free shipping service, then freeShipping will be recorded as one, two otherwise. Similarly, if the product is still in stock, then the variable inStock will be shown as one, two otherwise. In order to include these two variables as dummy variables in the analysis, I need to convert `two` to `zero` and transform them to the format of factor.

Finally, the effects of the outliers and the unrepresentative values are non-negligible in an analysis, as they may result in bias in estimations. I need to remove the outliers and some "minorities" in the dataset to ensure the validity of the estimation.

```
data.frame(summary(Electronics$color))
Electronics$color[Electronics$color=="Brown"]<- NA
Electronics$color[Electronics$color=="Dark Tarnish"]<- NA
Electronics$color[Electronics$color=="Stealth Black"]<- NA
```

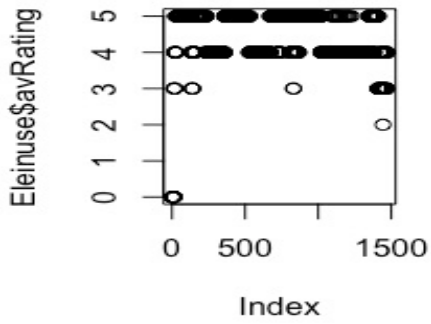
Table 1: Summary of Mobile Phone Color

Black	521
Blue	42
Brown	1
Dark Tarnish	3
Green	61
Grey	239
Orange	39
Red	15
Silver	33
Stealth Black	4
White	449
Yellow	64

Through summarizing the frequency of each color, I find that the mobile phones with color brown, dark tarnish and stealth black show low frequency in this dataset, which is less than 10 times in each case. Variables with low frequency are not representative and hence may mislead the analysis. Simultaneously, color will be treated as a categorical data in the estimation, in order to avoid dummy trap and perfect multicolinearity problems, one category of the categorical data will be omitted. When interpreting the results, the coefficients of the other categories illustrates the effect of the specific categories on the dependent variable compared to the omitted one. Thus, in case the low-frequency categories, which may lead to unfaithful results, are chosen as the omitted one, I remove the categories whose frequency are lower than ten. As well, this policy applies to other variables in this project.

```
plot(Eleinuse$avRating)
Electronics$avRating[Electronics$avRating=="2"] <- NA
```

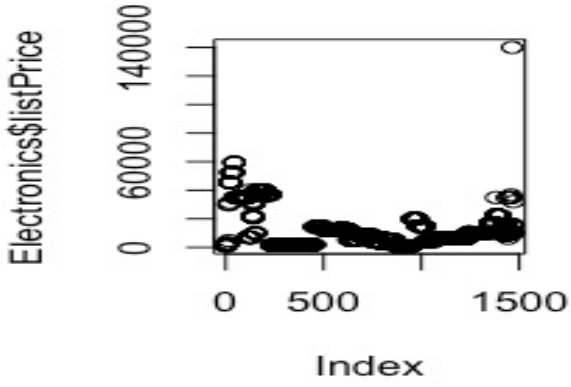
Figure 1: Average Rating of Products



As for the average rating of products, R treats it as a continuous variable with values lying in the domain of [0,5]. However, it is a categorical variable indeed. After plotting this variable, I found that there are five categories within it, which are integers from zero to five, except one. Also, there is only mobile's average rating is equal to two. Applying the policy I mentioned above, which is removing outliers and "minorities", I drop the observation where the average rating is two. Then, I convert the average rating of products from a continuous variable to a categorical one for the programming sake.

```
summary(Electronics$listPrice)
plot(Electronics$listPrice)
```

Figure 2: List Price of Products



Electronics$listPrice[Electronics$listPrice=="14000"]<-NA

data.frame(summary(Electronics$shippingPeriod))

Table 2: Shipping Period of Products

0	9
1-3 working days	53
2-3 business days	1178
3-4 business days	77
3-4 working days	22
3-5 business days	21
4-5 business days	66
5-7 business days	27
5-7 working days	4
6-7 Working Days	10
7-8 Working Days	4

```
Electronics$shippingPeriod[Electronics$shippingPeriod=="0"] <- NA
Electronics$shippingPeriod[Electronics$shippingPeriod=="5-7 working days"] <- NA
Electronics$shippingPeriod[Electronics$shippingPeriod=="7-8 Working Days"] <- NA
Electronics <- na.omit(Electronics)
Electronics$avRating_dummies <-table(1:length(Electronics$avRating),as.factor(Electronics$avRating))
```

When it comes to the list price of the products, I conclude the highest value, which is 14000, as the outlier after plotting. Thus, I drop the extreme value. In terms of shipping period, the frequency of zero shipping day, 5-7 working days and 7-8 working days are lower than 10. As per my data cleansing strategy, I remove these three categories. Till now, I finalize my data wrangling and update my dataset ***Eletronics***, which contains 1445 observations. During this process, more

than 6000 observations are lost. However, it is due to the nature of the dataset, since it possesses too many missing values and outliers. I choose to sacrifice the number of observations in exchange for the precision of marginal effects and predictions.

2.2 Exploratory Analysis

In this project, I aim to solve two main problems: Estimating the effects of independent variables on the price change in the next sales period of mobile phones, at the same time, predicting the whether the price will increase or not. Base on the two objectives, ordinary least squares (OLS) and logistic regression model may both work. However, Since the dependent variable, which is whether the product price goes up on the next time period, is a dummy variable, logistic regression is more appropriate in this case. OLS is commonly used when the dependent variable is continuous, it is convenient for users to interpret the marginal effect of each independent variable on the outcome. However, when the dependent variable is binary, which is similar to ascertain the probability of price increase in my case, OLS may result in meaningless results. As the linear property of OLS, it may predict values less than zero or greater than one under some scenarios. This will make the estimation imprecise, because the probability must lie in [0,1]. Hence, logistic regression is preferred. The whole dataset is supposed to be randomly divided into a training set and a test set respectively. The training set contains two thirds of the data, while the test set is composed of the remaining one third. Accordingly, the training set has 980 observations and the test set constrains 491observations.

```
set.seed(345)
index_train <- sample(1:nrow(Electronics),2/3*nrow(Electronics))
trainingset <- Electronics[index_train,]
testset <- Electronics[-index_train,]
```

As discussed in the first section, rating and comments from other users, logistic affairs regarding the online shopping and the platforms are the three core factors that affect online consumers' purchasing decisions, and these decisions are associated with producers' decisions in price change, so these factors impose latent influence on price change. Based on this logic, the proxies of these features are incorporated into the estimation model. The model is as follow:

$$Logit(PriceUp) = \alpha + freeShipping + inStock + avRating + brand + reviewCount + color + siteName + ListPrice + shippingPeriod + \xi$$

$\boldsymbol{PriceUp}$— it is the dependent variable, which indicates whether the price of the product goes up on the next time when the price is recorded

$\boldsymbol{\alpha}$— constant term

$\boldsymbol{freeShipping}$— it is a dummy variable and represents whether this product comes with free shipping service

$\boldsymbol{inStock}$—it is a dummy variable and represents whether this product is in stock

$\boldsymbol{avRating}$—it is a categorical variable in the domain [0,5] and represents the average rating of the product

$\boldsymbol{brand}$—it represents the brand of the product

$\boldsymbol{reviewCount}$—it represents the number of users who gave rates of the products

$\boldsymbol{color}$—it represents the color of the product

$\boldsymbol{siteName}$—it is a categorical variable and represents the name of the platforms where the product sold

$\boldsymbol{ListPrice}$—it is the price of the products at the time recorded

$\boldsymbol{shippingPeriod}$—it represents the shipping period of the product

$\boldsymbol{\xi}$ -- error term

3. Data Analysis and Results

3.1 Marginal effects

Besides predicting the price change of the mobiles, the marginal effects of the three factors in the price change are expected to be investigated as well. Producers will be able to know the marginal effects of each variable and formulate corresponding service promotions to acquire a higher product price and hence more profits.

Based on the ideas of conducting marginal effects analysis, I would take the three features into consideration progressively, and make efforts to figure out the validity of these in-theory significant variables. Therefore, I estimate the effect of the average rating and the number of rating on the dependent variable at the first stage.

```
my_result_1 <- glm(PriceUp ~ avRating_dummies+reviewCount, family="binomial", data=trainingset)
coef(summary(my_result_1))
logitmfx(formula=PriceUp ~ avRating_dummies+reviewCount,data=trainingset)
```

Table 3: Marginal Effects of the First Stage

	dF/dx	Std.Err.	z	P>|z|	
avRating_dummies0	-0.015225	0.0058973	-2.5817	0.009832	**
avRating_dummies3	0.0998444	0.5371116	0.1859	0.85253	
avRating_dummies4	0.0121084	0.0719779	0.1682	0.866407	
reviewCount	-0.0030831	0.0183031	-0.1684	0.866233	

In order to avoid dummy trap, the variable represents the average rating is 5 is omitted in this regression. Compared to the products with the highest level average rating, those are rated zero on average have a 1.5% lower probability of increasing the products' prices. The effects of average rating of the products from 3 to 5 on the price change exhibit no differences at this point, as the marginal effects reported on the second and third variables in this regression are statistically insignificant. In terms of the number of rating, it imposes no impact on the price change.

```
my_result_2 <- glm(PriceUp ~ avRating_dummies+freeShipping+reviewCount+shippingPeriod,
family="binomial", data=trainingset)
coef(summary(my_result_2))
logitmfx(formula=PriceUp ~ avRating_dummies+freeShipping+reviewCount+shippingPeriod,data=trainingset)
```

Table 4: Marginal Effects of the Second Stage

	dF/dx	Std.Err.	z	P>|z|
avRating_dummies0	-0.0024168	0.31140543	-0.0078	0.9938
avRating_dummies3	-0.000975	0.12829457	-0.0076	0.9939
avRating_dummies4	-0.0010198	0.13405618	-0.0076	0.9939
reviewCount	-0.0001294	0.01700716	-0.0076	0.9939
freeShipping1	0.00010383	0.01376446	0.0075	0.994
shippingPeriod2-3 business days	-0.0215515	2.77727815	-0.0078	0.9938
shippingPeriod3-4 business days	-0.0067214	0.67424335	-0.01	0.992
shippingPeriod3-4 working days	0.00046792	0.06151463	0.0076	0.9939
shippingPeriod3-5 business days	-0.0029406	0.36427431	-0.0081	0.9936
shippingPeriod4-5 business days	-0.0046877	0.52045171	-0.009	0.9928
shippingPeriod5-7 business days	-0.0031666	0.38610952	-0.0082	0.9935

At this stage, the proxies of logistic affairs are incorporated into the regression, which are whether the product comes with free shipping service and the shipping period of the product. In addition to the highest average rating, the category that represents 1-3 working days in the shipping period variable is omitted. According to the reported z-values none of these variables is statistically significant after including the variables of logistic issues. It illustrates rating of the products and the shipping methods are not determinants of the price increase in the next period. Furthermore,

siteName is a proxy of the online shopping platforms. The third regression illustrates how the three unique features of online shopping affect the price change in the Indian Market.

```
my_result_3 <- glm(PriceUp ~ avRating_dummies+reviewCount+freeShipping+shippingPeriod+siteName,
family="binomial", data=trainingset)
coef(summary(my_result_3))
logitmfx(formula=PriceUp ~
avRating_dummies+reviewCount+freeShipping+shippingPeriod+siteName,data=trainingset)
```

However, the results after including the **siteName** are identical to the ones reported in Table 4, which means **siteName** is highly correlated to the existing variables of rating and logistic stuff. Due to perfect multicolinearity, **siteName** is omitted in this regression. Thus, I may conclude even though the rating system, logistic affairs and the platform where the product sold are potentially affect the sales of the product, they are not the driving forces of price change in the next time period. In the next regression, I would include all the variables provided in the dataset to find out the factors that significantly affect the price increase.

```
my_result_4 <- glm(PriceUp ~
avRating_dummies+reviewCount+freeShipping+shippingPeriod+siteName+inStock+brand+color+listPrice,
family="binomial", data=trainingset)
coef(summary(my_result_4))
logitmfx(formula=PriceUp ~
avRating_dummies+reviewCount+freeShipping+shippingPeriod+siteName+inStock+brand+color+listPrice,data=
trainingset)
```

Table 5: Marginal Effects of the Full Regression

	dF/dx	Std.Err.	z	P>|z|	
avRating_dummies0	-4.13E-05	1.43E-02	-0.0029	0.997695	
avRating_dummies3	-3.16E-05	1.10E-02	-0.0029	0.997715	
avRating_dummies4	-1.61E-06	5.63E-04	-0.0029	0.99772	
reviewCount	-1.30E-06	4.53E-04	-0.0029	0.997715	
freeShipping1	-2.11E-04	7.37E-02	-0.0029	0.997714	
shippingPeriod2-3 business days	-5.65E-04	1.97E-01	-0.0029	0.997713	
shippingPeriod3-4 business days	-1.27E-04	4.10E-02	-0.0031	0.99752	
shippingPeriod3-4 working days	-3.64E-05	1.27E-02	-0.0029	0.997715	
shippingPeriod3-5 business days	-4.98E-05	1.71E-02	-0.0029	0.997675	
shippingPeriod4-5 business days	-8.45E-05	2.80E-02	-0.003	0.997588	
shippingPeriod5-7 business days	-3.67E-05	2.03E-02	-0.0018	0.998556	
inStock1	2.63E-05	9.19E-03	0.0029	0.997715	
brandArise	1.00E+00	4.27E-01	2.342	0.019182	*
brandBlackberry	1.00E+00	5.16E-02	19.3856	2.20E-16	***
brandBlackBerry	1.00E+00	3.25E-01	3.0807	0.002065	**
brandHTC	1.00E+00	3.20E-01	3.1278	0.001761	**
brandIBall	9.99E-01	4.76E+00	0.2098	0.83379	
brandKarbonn	9.99E-01	1.92E+00	0.5192	0.603599	
brandLG	1.00E+00	2.57E+00	0.3894	0.696954	
brandMicromax	5.17E-04	2.16E+00	0.0002	0.999809	
brandMotorola	4.13E-04	2.12E+00	0.0002	0.999844	
brandNokia	1.00E+00	1.18E+00	0.8443	0.398477	
brandSamsung	9.83E-01	5.50E+01	0.0179	0.985746	
brandSony	1.00E+00	4.96E-01	2.0155	0.04385	*
colorBlue	-1.99E-05	6.94E-03	-0.0029	0.997715	
colorGreen	5.19E-04	2.60E+00	0.0002	0.999841	
colorGrey	-3.48E-05	1.22E-02	-0.0029	0.997715	
colorOrange	6.23E-04	3.45E+00	0.0002	0.999856	
colorRed	-4.43E-05	1.53E-02	-0.0029	0.997686	
colorSilver	4.99E-04	1.74E-01	0.0029	0.997714	
colorWhite	3.46E-05	1.21E-02	0.0029	0.997715	
colorYellow	9.95E-01	7.55E+00	0.1318	0.895174	
listPrice	2.36E-09	8.24E-07	0.0029	0.997715	

Definitely, the estimation only include three important factors might be biased, owing to the failure of inclusion other omitted variables, which are correlated with the independent variables and are the determinants of the price change as well. Thus, control variables are essential to be imported and the model at this stage is preferred in marginal effects analysis, as it reduces the bias arising from the omitted variables. The nature of the products like brands and colors may affect the pricing process via the brand loyalty and color preferences coming from the customers. Likewise, the products stock status and list price matters as well. Being out of stock may be due to a high sales volume, and the price of that product is more likely to raise in the next sales period; if the list price is low, then the likelihood of increasing the price is high generally. Based on these logics, I incorporated the four variables in the regression. The results are shown above, AirTyme under brand and Black under color are omitted, because of statistical policies. This model is the

most comprehensive one in this analysis, rating from users, logistic stuff and platforms of the products still perform insignificantly, while the new-added variables of stock status and list price of the products are apparently not the driving forces of the price increase either. Compared to the black products, products with other colors fail to reveal statistical differences. However, some brands possess a higher likelihood of increasing prices than the rest, such as Arise, Blackberry[1], HTC and Sony. Compared to AirTyme, the probabilities of price increase of these four brands are almost 100%. Some vital variables that significantly affect the price change might be missed in this dataset. A more comprehensive dataset is required for further studies.

3.2 Predictions

Throughout the marginal effects analysis, both customers and producers are about to realize the how each of the variables affect the pricing affairs. However, in order to make wise decisions in mobile purchase, customers need to know the likelihood of having a higher price of the target product in the next sale period. 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. In this section, I try to predict the price information based on the logistic regressions developed above. Because the first stage regression contains too few variables and the third regression reports identical results as the second one, I only employ the second stage regression, which has the variables represents three important features summarized in the previous parts, and the forth regression, which is composed of the most variables.

```
predictions_2 <- predict(my_result_2,newdata=testset,type="response")
predictions_4 <- predict(my_result_4,newdata=testset,type="response")
```

3.2.1 Choice of cut-offs

Keeping a high level of accuracy is the priority of a prediction. Thus, choosing an appropriate cut-off is an essential step in the predicting process. Because the forth regression in last sector is the most comprehensive one in terms of the inclusion of variables, I try to generate 101 cut-offs according to the quantile of its predictions and then choose the most suitable one.

[1] In the dataset, there are two categories of blackberry. However, there is no explanations about the differences or the similarities of these two categories on the official website. Therefore, I choose to keep both in this analysis.

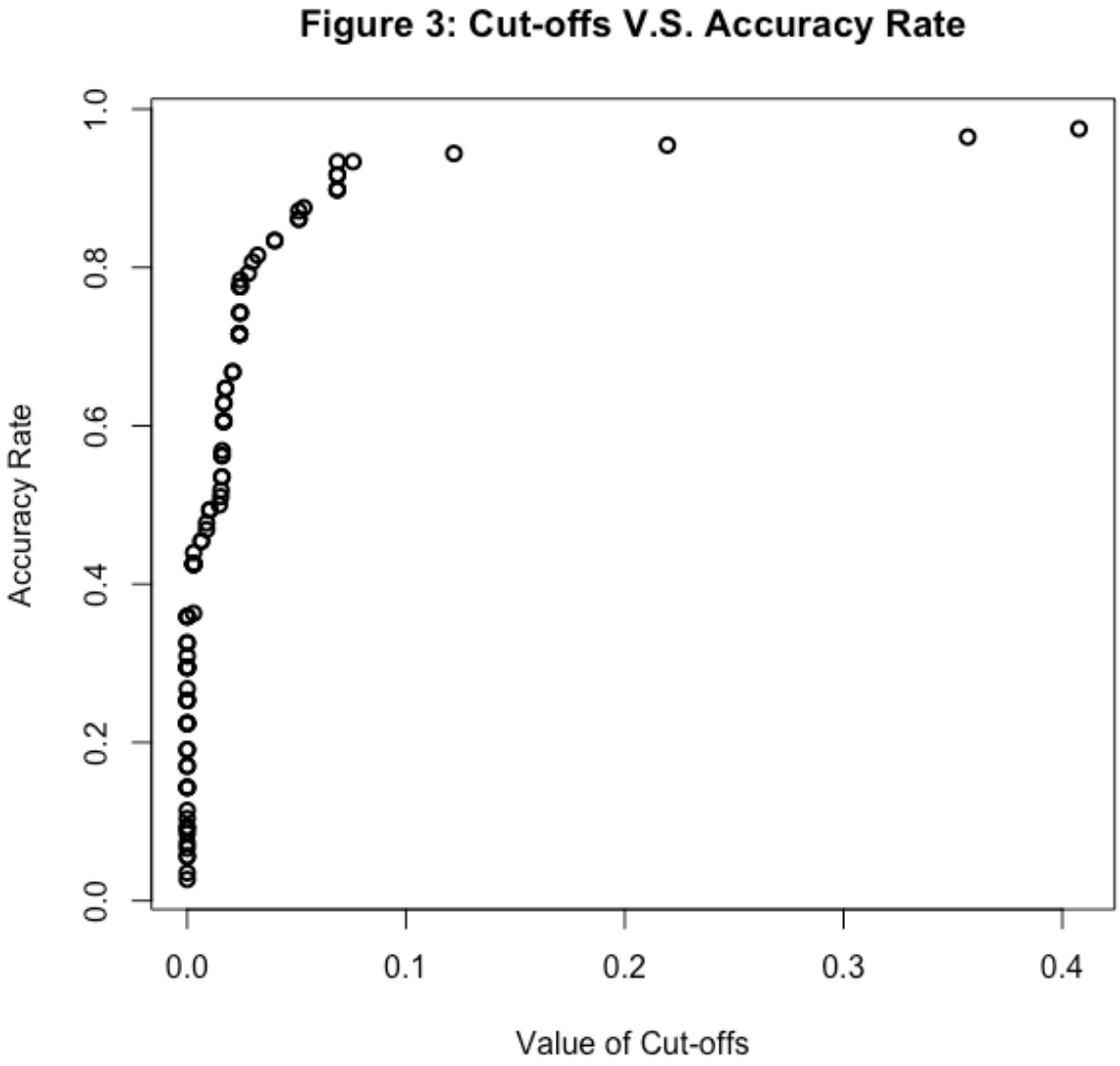
```
cutoffs<- quantile(predictions_4,seq(0,1,by=0.01))
M <- length(cutoffs)
pred <- matrix(data=NA, nrow=M, ncol=length(predictions_4))
m=1
for (m in 1:M){pred[m,] <- ifelse(predictions_4>cutoffs[m],1,0)
m=m+1}
my_table <- vector("list",101)
for (m in 1:M) { my_table[[m]] <- cbind(rep(NA,2), NA)}
for (m in 1:M){my_table[[m]] <- table(testset$PriceUp,pred[m,]) }
accuracy <- length(M)
for (m in 1:M) {accuracy[m] <- sum(diag(my_table[[m]])/nrow(testset))}
summary(predictions_4)
```

Table 6: Summary of the forth prediction

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	0	0.01471	0.02488	0.02416	0.4076

```
plot(cutoffs,accuracy,xlab= "Value of Cut-offs",
    ylab="Accuracy Rate",lwd=2, main="Figure 3: Cut-offs V.S. Accuracy Rate")
```

Figure 3: Cut-offs and Accuracy Rate



Shown in Figure 3, when the values of cut-offs move from 0 to 0.028, the accuracy rate raises sharply from 0 to 80 percent. Going beyond 0.028, the growth rate of accuracy rate tends to be gentle and gradually approach to 100 percent. In terms of the cut-offs lie in the domain of (0.068,0.7], the corresponding accuracy rates are always higher than 90 percent. The summary statistics of **predictions_4** illustrates the distribution of cut-offs contains big jumps in the right tail. With the objectives of keeping high accuracy rate, I choose the 91th quantile (corresponds to 0.068) as the cut-off. Figure 3 reveals a positive relationship between the cut-offs and the accuracy rates,

which means choosing a higher quantile will definitely lead to an incremental improvement in the accuracy rate, but a higher quantile may also give rise to a fewer number of true positive values, and hence makes the prediction meaningless. Thus, the 91th quantile of the prediction is chosen as the cut-off in this analysis, and the corresponding accuracy rate is 90 percent. Likewise, I create confusion matrices for the other three logistic regressions in the last sector by regarding 91th quantile of the respective predictions as the cut-offs, and store them in **pred_2**. Furthermore, in order to seek out the better fitted prediction model in these two, it requires to generate the Receiver Operating Characteristic (ROC) cuves and the associated Areas under the curves (AUCs) based on the **PriceUp** column in the testset and the predicted results of price change from the two regressions.

```
cutoff2 <- quantile(predictions_2,0.91)
pred_2 <- ifelse(predictions_2 > cutoff2,1,0)
Table_2 <- table(testset$PriceUp,pred_2)
accuracy_2 <- sum(diag(Table_2)/nrow(testset))
```

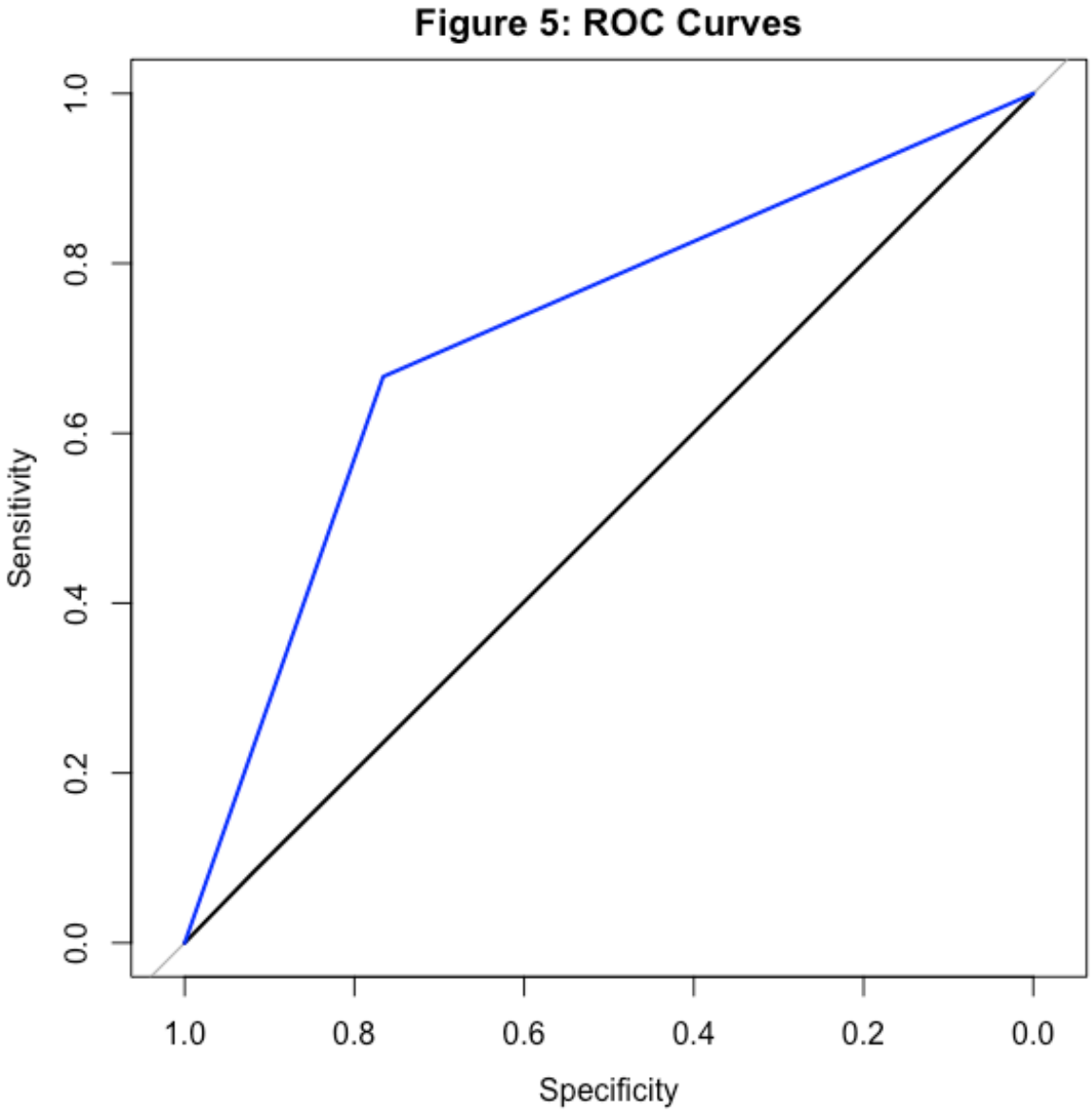
3.2.2 ROC and AUC

```
###install packages for ROC###
installed.packages("pROC")
library(pROC)
ROC_pred_4 <- roc(testset$PriceUp,pred[92,])
ROC_pred_2 <- roc(testset$PriceUp,pred_2)
plot(ROC_pred_4,main="Figure 5: ROC Curves")
lines(ROC_pred_2,col="Blue")
AUC_2 <- auc(ROC_pred_2)
AUC_2
```

Table 7: AUC

	Three Features	Full Regression
AUC	0.7163	0.5012

Figure 5: ROC Curves



AUC reveals the areas under the ROC curse, which is commonly used to determine which of models predicts the classes with the highest precisions. The models with higher AUCs are preferred. The black line in Figure 5 represents the ROC curve for the forth (full) regression, while the blue line represents the ROC curves for the regression particularly designed for the three features.

Table 6 summarizes the AUCs in the two cases: the AUC calculated for the full regression is only slightly higher than 0.5, but the one comes from the second regression shows a much higher

value, which is 0.7163. Hence, the regression only contains the proxies of consumer rating and logistic affairs where the products sold is strongly preferred in terms of predicting the price change.

4.Conclusion and recommendations

4.1 Business Perspective

As an atypical shopping pattern, online shopping is substantively affected by three factors beyond the nature of products: public rating system and comments, logistic affairs involvement and the choices of platforms. Since the pricing strategy of products comes from the market mechanisms. In other words, the price of a product is determined by the supply and demand in the goods market. From the producers' perspective, they are curious about the driving forces that lead to higher prices. According to the marginal effects analysis, the three features mentioned above fail to explain the price change, even though they may be the determinants of from the sales promotion perspective. Similarly, stock status, color and list prices of the products fail to explain the dependent variable. However, the brand of products is highly associated with the price increase in the next period. In summary, mobile devices produced by Arise, Blackberry, Sony and HTC are more likely to increase their prices. Hence, customers who are looking for products from these four companies will be better off, if they make purchasing decision at the current time period. Unfortunately, this analysis provides no guidance for producers on how to improve the chance of charging higher prices other than the brand itself.

In terms of the prediction, through controlling for obtaining a 90% accuracy rate and comparing the prediction performance from two different regressions, I will recommend consumers and producers predict the price change with a relatively high precision via the available data in public rating stuff and logistic affairs. Accordingly, the average rating, number of users who rated the products, whether having free delivery service and time spent on shipping are necessary to be taken into consideration. As the AUC reported in this scenario (0.7163) is much higher than the one from the full regression.

4.2 Technical Perspective

The existence of omitted variables of this analysis threatens the accuracy of the estimations and the predictions. First, the omitted variable bias problem result in a mis-estimation. The edition of the products and the frequency of the new products release are two typical omitted variables in this analysis. These two variables are strongly correlated with the stock status and the average rating of one product, since a new edition of a product with a low frequency of new version release

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. The failure of including these factors in the dataset tends to give rise to bias in the estimation. Second, the values of AUCs are not very high, which reveals the lack of other valuable variables in the predictions. Third, the internal competition within the mobile industry is missed too. The competition among the producers may result in different pricing strategies chosen by various firms against other existing strategies in the mobile industry, and thus imposes an impact on the price change beyond the variables included. However, there is no explicit information about the competition in the dataset. Hence, more explainable variables are required to provide an enhanced level of prediction.

When it comes to the problems of the chosen model for prediction, zero significant explainable variables seem to be a big concern. Admittedly, this problem is one of the drawbacks of this prediction, however, considering the goal of a prediction model, which is predicting the dependent variables with high precision based on the explainable variables, AUC instead of z-values should be our primary focus. Also, the nature of the regression, which contains dummy and categorical variables, allows us to interpret the coefficients by including "benchmarks". In the second regression, the coefficients reported are interpreted as the effects on the price change compared to the omitted variables, which are average rating of five and 1-3 working days shipping. Thus, insignificance of the variables is not equivalent to zero effects on price change to some extent and hence this is an imperfect but acceptable model.

The problems raised above are needed to be addressed with further endeavors and assistance. If the producers expect both of the marginal effects analysis and predictions with higher levels of accuracy, they need to collect a more comprehensive dataset to conduct the analysis.