An Analysis of the Price Change in Online Mobile Shopping

Shubo Zhang

November 15, 2016

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$color[Electronics$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
Eleinuse <- na.omit(Electronics)
```

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.

```
Eleinuse$inStock[Eleinuse$inStock=="2"]<- "0"
Eleinuse$inStock<-as.factor(Eleinuse$inStock)
Eleinuse$freeShipping[Eleinuse$freeShipping=="2"]<- "0"
Eleinuse$freeShipping<-as.factor(Eleinuse$freeShipping)
```

2.2 Methodology

Since the dependent variable, which is whether the product price goes up on the next time period, is a dummy variable, a generalized linear model with binomial distribution instead of a simple linear one may lead to a more precise estimation. Meanwhile, as a practical project, the model is expected to predict the price change with high level of accuracy, so a logistic regression model is utilized in this case. 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.

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

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

- **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
- α - constant term

- **freeShipping** - it is a dummy variable and represents whether this product comes with free shipping serive
- **inStock** - it is a dummy variable and represents whether this product is in stock
- **avRating** - it is a continous variable in the domain [0,5] and represents the avarage rating of the product
- **brand** - it represents the brand of the product
- **color** - it represents the color of the product
- **siteName** - it is a categorical variable and represents the name of the platforms where the product sold
- ξ - 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 on the dependent variable at the first stage.

```
my_result_1 <- glm(PriceUp ~ avRating, family="binomial", data=trainingset)
coef(summary(my_result_1))
```

```
##              Estimate Std. Error   z value   Pr(>|z|)
## (Intercept) -1.541650  0.7103677 -2.170214 0.02999065
## avRating    -0.423861  0.1666024 -2.544147 0.01095449
```

```
## Call:
## logitmfx(formula = PriceUp ~ avRating, data = trainingset)
##
```

```
## Marginal Effects:
##               dF/dx  Std. Err.       z    P>|z|
## avRating -0.0131168  0.0050744 -2.5849 0.009741 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the reported coefficient and marginal effect, the average rating is statistically significant at 1% significance level. However, it imposes a detrimental impact on price increase. As one unit raise in the average rating, the likelihood of increasing the product's price is lower by 1.3%. Definitely, the estimation at this stage might be biased, owing to the failure of inclusion other omitted variables, which are correlated with the average rating and are the determinants of the price change as well, such as the logistic affairs. Thus, the variables that are the proxies of logistic affairs are included in the second stage.

```
my_result_2 <- glm(PriceUp ~ avRating+freeShipping, family="binomial", data=trainingset)
coef(summary(my_result_2))
```

```
##                 Estimate Std. Error   z value    Pr(>|z|)
## (Intercept)   -2.5736547  0.9502846 -2.708299 0.006762907
## avRating      -0.4381099  0.1570774 -2.789134 0.005284920
## freeShipping1  1.2280089  0.7355388  1.669537 0.095011067
```

```
## Call:
## logitmfx(formula = PriceUp ~ avRating + freeShipping, data = trainingset)
##
## Marginal Effects:
##                   dF/dx Std. Err.       z    P>|z|
## avRating      -0.012818  0.004673 -2.7431 0.006087 **
## freeShipping1  0.025427  0.010069  2.5254 0.011558 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## dF/dx is for discrete change for the following variables:
##
## [1] "freeShipping1"
```

By including the variable that whether free delivery service is supported, the marginal effect of the average rating on the price change stays unchanged, while having free delivery service may result in a 2.5% increase in the probability of price raise. 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+freeShipping+siteName, family="binomial", data=trainingset)
coef(summary(my_result_3))
```

```
##                        Estimate  Std. Error     z value     Pr(>|z|)
## (Intercept)         -1.8335247   1.3721189 -1.33627246 1.814602e-01
## avRating             0.1079572   0.3838939  0.28121613 7.785446e-01
## freeShipping1        0.4262141   0.7926483  0.53770894 5.907780e-01
## siteNameHomeShop18 -15.1587578 758.8039193 -0.01997717 9.840616e-01
## siteNameInfibeam    -2.9305145   0.5867593 -4.99440624 5.901705e-07
```

```
## Call:
## logitmfx(formula = PriceUp ~ avRating + freeShipping + siteName,
##     data = trainingset)
##
## Marginal Effects:
##                           dF/dx  Std. Err.       z     P>|z|
## avRating             0.0021337  0.0175757  0.1214    0.9034
## freeShipping1        0.0073684  0.0562687  0.1309    0.8958
## siteNameHomeShop18 -0.0234662  0.0052099 -4.5041 6.665e-06 ***
## siteNameInfibeam   -0.2233783  1.2893317 -0.1733    0.8625
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## dF/dx is for discrete change for the following variables:
##
## [1] "freeShipping1"      "siteNameHomeShop18" "siteNameInfibeam"
```

In the third regression, the average rating and free delivery service are no longer statistically significant,

although both of them exhibit positive effect on the price change. Instead, the platform of online shopping does matter. In order to avoid dummy trap, the website called `BuyThePrice` is omitted. Compared to this one, products sold on `HomeShop18` have a 2.3% lower likelihood of have price increase, while for those sold on `Infibeam` have insignificant impact.

At the same time, 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 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. The results are shown below, after including the variables regarding the nature of the products and the stock status, none of the variables are reported as statistically significant. This illustrates neither brand nor color is a driving force for price change. Similarly, under this model with less likelihood of having omitted variable bias than the ones above, the three features are not the core determinants of the price change. Some vital variables that significantly affect the price change might be missed in this dataset. A more comprehensive dataset is required for further studies.

```
my_result_4 <- glm(PriceUp ~ freeShipping+avRating+siteName+brand+color+inStock,
                   family="binomial", data=trainingset)
coef(summary(my_result_4))
```

```
##                         Estimate   Std. Error       z value    Pr(>|z|)
## (Intercept)        -14.65696312 2.688505e+03 -5.451716e-03 0.995650182
## freeShipping1       -1.85557013 1.382115e+00 -1.342558e+00 0.179415155
## avRating             0.06419929 5.580605e-01  1.150400e-01 0.908413407
## siteNameHomeShop18 -19.78926526 3.368675e+03 -5.874495e-03 0.995312858
## siteNameInfibeam    -4.01059757 1.253757e+00 -3.198864e+00 0.001379701
## brandApple           0.68780706 3.603185e+03  1.908887e-04 0.999847693
## brandArise          17.00347643 2.688504e+03  6.324511e-03 0.994953804
## brandBlackberry     14.60384011 2.688505e+03  5.431956e-03 0.995665948
## brandBlackBerry     15.18027099 2.688505e+03  5.646362e-03 0.995494879
## brandHTC            14.39864453 2.688505e+03  5.355633e-03 0.995726844
## brandHuawei         16.25573608 2.688505e+03  6.046385e-03 0.995175712
## brandIBall          14.21393713 2.688504e+03  5.286932e-03 0.995781658
## brandKarbonn        15.41349934 2.688504e+03  5.733113e-03 0.995425663
## brandLG              0.67400093 8.065510e+03  8.356582e-05 0.999933324
```

```
## brandMicromax       16.09369897 2.688504e+03  5.986116e-03 0.995223799
## brandMotorola        0.67400093 3.104417e+03  2.171103e-04 0.999826771
## brandNokia          14.85573858 2.688505e+03  5.525651e-03 0.995591191
## brandSamsung        14.01159909 2.688505e+03  5.211670e-03 0.995841708
## brandSony           16.35036387 2.688504e+03  6.081584e-03 0.995147628
## colorBlue            0.82224681 1.420453e+00  5.788623e-01 0.562682083
## colorBrown         -17.60684259 1.075401e+04 -1.637235e-03 0.998693676
## colorGreen          -1.10548254 3.145052e+03 -3.514990e-04 0.999719544
## colorGrey            2.12652033 1.268342e+00  1.676615e+00 0.093617863
## colorOrange         -0.98449961 3.396605e+03 -2.898481e-04 0.999768735
## colorRed           -14.70759362 3.094311e+03 -4.753107e-03 0.996207583
## colorSilver         -1.15530498 1.790902e+00 -6.450967e-01 0.518864498
## colorStealth Black   1.89502664 2.012098e+00  9.418164e-01 0.346286664
## colorWhite           0.80764182 8.886152e-01  9.088769e-01 0.363415129
## colorYellow          0.91766740 1.279231e+00  7.173585e-01 0.473152935
## inStock1            -0.03793509 8.806334e-01 -4.307705e-02 0.965640116
```

Nevertheless, even though I may conclude that none of the variables are the determinants of the pricing issue based on the forth results above, I will still recommend producers to offer free delivery service and not to sell products via `HomeShop18` in accordance with the results in the first three regressions.

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.

```
predictions_1 <- predict(my_result_1,newdata=testset,type="response")
predictions_2 <- predict(my_result_2,newdata=testset,type="response")
predictions_3 <- predict(my_result_3,newdata=testset,type="response")
predictions_4 <- predict(my_result_4,newdata=testset,type="response")
```

3.2.1 Choice of cut-offs

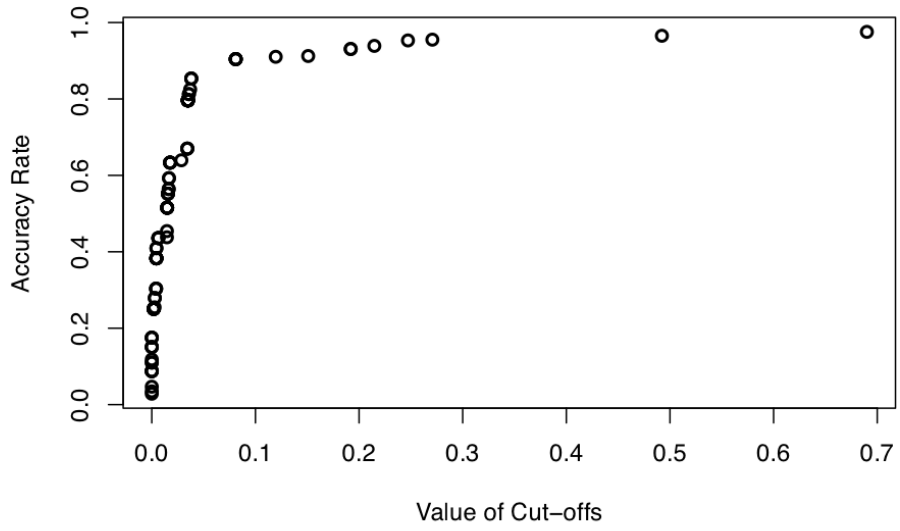
Keeping a high level of accuracy is the priority of a prediction. Thus, choosing an appropriate cut-off is an essential step in the predicting process. Because the forth regression in last sector is the most comprehensive one in terms of the inclusion of variables, I try to generate 101 cut-offs according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

```
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)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.000000 0.002896 0.014820 0.039280 0.034820 0.690000
```

```
plot(cutoffs,accuracy,xlab= "Value of Cut-offs",
     ylab="Accuracy Rate",lwd=2, main="Figure 1: Cut-offs V.S. Accuracy Rate")
```

Figure 1: Cut–offs V.S. Accuracy Rate



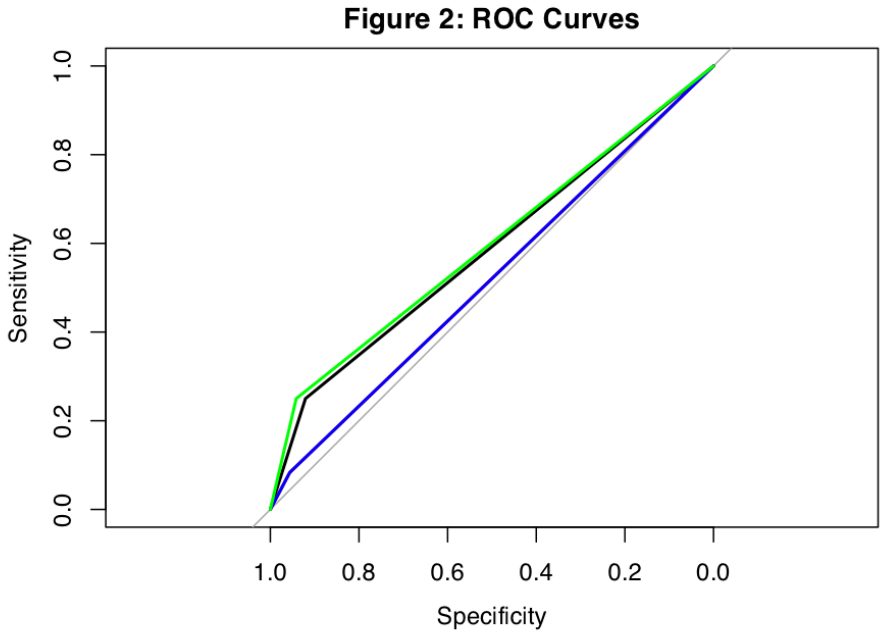
Shown in Figure 1, when the values of cut-offs move from 0 to 0.08, the accuracy rate raises sharply from 0 to more than 80 percent. Going beyond 0.08, the growth rate of accuracy rate tends to be gentle and gradually approach to 100 percent. In terms of the cut-offs lie in the domain of (0.08,0.7], 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 87th quantile as the cut-off. By checking the distribution of generated accuracy rates, I find that from the 86th quantile of the prediction to the 87th, there exists a big jump in the accuracy rate, which is from 85 percent to 90 percent. Figure 1 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 87th quantile of the prediction is chosen as the cut-off in this analysis, and the corresponding accruracy rate is 90.4 percent. Likewise, I create confusion matrices for the other three logitic regressions in the last sector by regarding 87th quantile of the respective predictions as the cut-offs, and store them in `pred_1`,`pred_2` and `pred_3`, seperately. Furthermore, in order to seek out the best fitted prediction model among those four, it requires to generate the Receiver Operating Characteristic (ROC) cures 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 four regressions.

3.2.2 ROC and AUC

```
ROC_pred_4 <- roc(testset$PriceUp,pred[88,])
ROC_pred_1 <- roc(testset$PriceUp,pred_1)
ROC_pred_2 <- roc(testset$PriceUp,pred_2)
ROC_pred_3 <- roc(testset$PriceUp,pred_3)
plot(ROC_pred_4,main="Figure 2: ROC Curves")
```

```
##
## Call:
## roc.default(response = testset$PriceUp, predictor = pred[88,     ])
##
## Data: pred[88, ] in 479 controls (testset$PriceUp 0) < 12 cases (testset$PriceUp 1).
## Area under the curve: 0.5853
```

```
lines(ROC_pred_1,col="Red")
lines(ROC_pred_2,col="Blue")
lines(ROC_pred_3,col="Green")
```




```
AUC_1 <- auc(ROC_pred_1)
AUC_1
```

```
## Area under the curve: 0.5197
```

```
AUC_2 <- auc(ROC_pred_2)
AUC_2
```

```
## Area under the curve: 0.5197
```

```
AUC_3 <- auc(ROC_pred_3)
AUC_3
```

```
## Area under the curve: 0.5958
```

The black line in Figure 2 represents the ROC curve for the forth regression, while the red, blue and green lines represent the ROC curves for the first three regressions respectively. The first two models coincide on

the figure, which reveals the inclusion of the proxy of logistic affairs technically has no contributions to the prediction. This statement is confirmed by turning out that the values of AUCs from those two regressions are identical.

AUC is commonly used to determine which of models predicts the classes with the highest precisions. The models with higher AUCs are preferred. Among these four reported values of AUCs, even though they are close, the third regression possesses the highest value and hence is the best fitted model in this analysis with an AUC of 0.5958. In the meantime, this conclusion illustrates the three unique characteristics, especially the average rating and the platforms, are the principal factors in predicting the price change, while the nature of the products and the stock status fail to.

4.Conclusion and recommendations

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, I conclude that including free delivery service within the mobile sales will raise the probability of reaching a higher price in the next sale period, while selling mobiles via `HomeShop18` drives to a negative effect on the price increase, compared to other online sales platforms. On the other hand, from the customers' point of view, they intend to know the trends of price change of the products so that they are able to make an efficient purchasing decision. Based on the predictions models above, the model exclusively contains the three important factors in online shopping is the best-fitted one to conduct predictions. Thus, customers can predict the price change with a relatively high precision through the available data in average rating, whether having free delivery service and the platforms where the products sold.

However, 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 misestimation. 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 all less than 0.6, 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. 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.