iPad on eBay, sold or not sold

Predictive Modeling---Group 1

Tiffany Sung, Olivia Pan, Lining Jiang,

Tammy Huang, Jireh Zhou

# Problem Set

By using data of iPads on eBay, we want to see which factors could affect the probability of whether that iPad could be sold or not.

# Dataset

We have 1,861 observations in our dataset. We have 9 variables to predict the probability of selling the iPad on eBay including: description, biddable, startprice, condition, cellular, carrier, color, storage and productline. For which description (we converted it into words count in description) and startprice are continuous variables and the others 7 are categorical ones.

We also found some interesting findings during summarizing the data. Firstly, for the storage, most iPads in our sample are 16 GB storage while 128G has the fewest. However, those 128G are not proportionately shown on the sold and unsold bar chart we made, they seemed to be hard to sell. Second, for the condition, most iPads sold on eBay are used. However, used iPads tends to be sold more on eBay.

Furthermore, while most price of iPads on eBay are set as biddable, our sample shows that there are much more proportions of biddable ones be sold than unbiddable ones.

# Methodology

We separated our dataset into 1500 observations for training set and the remain 361 observations for testing set.

## Logistic regression model/ K-fold Cross Validation

Firstly, we constructed 9 different logistic regression models. The ninth models were constructed by using all variables and according to the results, we then constructed the other eight models by different combinations of the 4 most significant variables: biddable, startprice, productline and storage.

Glm7: sold ~ biddable + startprice + productline

Glm8: sold ~ biddable + startprice + productline + storage

Glm9: sold~ all variables we have

By using AIC and CV-error we got to do comparison, we choose glm7 as our final logistic regression model.

Predict an iPad can be sold if the prediction value is greater than 0.5

The accuracy of our test dataset is about 81%

## Classification Tree

We first built a big tree which has 183 nodes. After cross validation, we found that the tree with only 5 nodes actually had the smallest deviance, of 0.1499. Then we prune the tree to 5 nodes and calculate the accuracy to be around 80%, which is similar to the results from logistic regression model.

In the dataset that we used, there are 9 variables in total. However, only 3 variables are left in this tree model which includes biddable, startprice, and productline. This means that these 3 variables are relatively more significant than the other 6 variables.

## Random Forests/ Boosting

In addition, we used Random Forests, and found that 45 trees could reduce the residual mean deviance from the origin 0.1499 to 0.1494. The result again shows that biddable, startprice, and productlone are the most important factors. The boosting also validates this conclusion.

# Results

|  |  |  |  |
| --- | --- | --- | --- |
|  | Logistic Model  (glm7) | Classification Tree  (5 nodes) | Random Forest  (45 trees) |
| Variables Used | **Biddable, StartPrice,**  **Product line** | **Biddable, StartPrice,**  **Product line** | **Unknown** |
| Accuracy | **81.11%** | **80.03%** | **86.00%** |