PREDICTING SALES ON EBAY

Individuals selling used items are often faced with a difficult choice -- should they try to sell the items through a yard/estate sale, a consignment shop, an auction, or some other means? Often, this choice will come down to the convenience of selling the items, the price the items can fetch, and the speed with which the items can be sold.

To determine whether analytics can be used to help make this choice, we will look at whether data from previous auctions on eBay, a major online auction and shopping site, can be used to predict whether a new item will be sold at some target price. We will limit our attention to Christian Louboutin shoes, using data from nearly 4,000 auctions from late 2014. In this analysis, the dependent variable will be the binary outcome variable **sold**, which takes value 1 if the item was sold and 0 if it was not sold. We also include **saleprice**, which is the price the shoe sold at (NA for shoes that did not sell). For each item, the file ebay.csv contains the following independent variables:

- **biddable**: Whether this is an auction (biddable=1) or a sale with a fixed price (biddable=0)
- **startprice**: The start price (in US Dollars) for the auction (if biddable=1) or the sale price (if biddable=0)
- **condition**: The condition of the shoe (New with box, New with defects, New without box, or Pre-owned)
- **size**: The size of the shoe (converted to US shoe sizes)
- **heel**: The size of the heel (Flat, Low, Medium, High)
- **style**: The style of the shoe (Open Toe, Platform, Pump, Slingback, Stiletto, or Other/Missing)
- **color**: The color of the shoe (Beige, Black, Brown, Red, or Other/Missing)
- **material**: The material of the shoe (Leather, Patent Leather, Satin, Snakeskin, Suede, or Other/Missing)
- **snippit**: A short snippit of text describing the shoe
- **description**: A long text description describing the shoe

PROBLEM 1 - LOADING THE DATA (1/1 point)

Use the read.csv function to load the contents of ebay.csv into a data frame called eBay, using stringsAsFactors=FALSE. What proportion of all shoes were sold?

0.2104847

Answer: 0.21

EXPLANATION

This can be computed with the table or mean functions.

Final Check Save Hide Answer *You have used 1 of 2 submissions*

PROBLEM 2 - MISSING VALUES (1 point possible)

Which of the numerical variables has at least one missing value?

☐ biddable

☐ sold

☑ startprice

☑ size ✓

EXPLANATION

This can be read from the output of the summary function.

Hide Answer *You have used 1 of 1 submissions*

PROBLEM 3 - MOST COMMON SHOE SIZE (1/1 point)

What is the most common shoe size in the dataset?

○ 4

○ 5

○ 6

○ 7

● 8 ✓

○ 9

○ 10

○ 11

○ 12

EXPLANATION

This can be determined with the table function.

Hide Answer *You have used 1 of 1 submissions*

PROBLEM 4 - CONVERTING VARIABLES TO FACTORS (1 point possible)

Convert the following variables to factors using the as.factor function:

- sold

- condition

- heel

- style

- color

- material

Which of the following methods requires the dependent variable be stored as a factor variable when training a model for classification?

- ● Logistic regression (glm) ✗
- ○ CART (rpart)
- ○ Random forest (randomForest) ✓

EXPLANATION

We convert the outcome variable to a factor for the randomForest() method.

Hide Answer *You have used 1 of 1 submissions*

PROBLEM 5 - SPLITTING INTO A TRAINING AND TESTING SET (1 point possible)

Obtain a random training/testing set split with:

set.seed(144)

library(caTools)

spl = sample.split(eBay$sold, 0.7)

Split articles into a training data frame called "training" using the observations for which spl is TRUE and a testing data frame called "testing" using the observations for which spl is FALSE.

EXPLANATION

Use the subset function to put the TRUE observations in the training set, and the FALSE observations in the test set.

Why do we use the sample.split() function to split into a training and testing set?

- ○ It is the most convenient way to randomly split the data
- ● It balances the independent variables between the training and testing sets ✗
- ○ It balances the dependent variable between the training and testing sets ✓

Hide Answer *You have used 1 of 1 submissions*

PROBLEM 6 - TRAINING A LOGISTIC REGRESSION MODEL (1/1 point)

Train a logistic regression model using independent variables "biddable", "startprice", "condition", "heel", "style", "color", and "material", using the training set to obtain the model.

Which of the following characteristics of a shoe are statistically significantly (p < 0.05, aka at least a * in the regression summary) associated with a lower chance of an item being sold?

- □ The item having "biddable" value 1 (aka item offered in an auction instead of as a fixed-price sale)

☑ The item having a high starting price ✓

EXPLANATION

The model can be trained with the glm function (remember the argument family="binomial") and summarized with the summary function.

Hide Answer *You have used 1 of 1 submissions*

PROBLEM 7 - PREDICTING USING A LOGISTIC REGRESSION MODEL (1/1 point)

Consider a shoe that is not for auction (biddable=0), that has start price $100, that is in condition "Pre-owned", that has "High" heels, that has style "Open Toe", that has color "Black", and that has material "Satin". What is the predicted probability that this shoe will be sold according to the logistic regression model?

0.2491443 **Answer:** 0.249

EXPLANATION

The observation has biddable=0, startprice=100, condition="Pre-owned", heel="High", style="Open Toe", color="Black", and material="Satin". Therefore, the prediction has logistic function value 0.5990788 + 100*-0.0044423 - 0.4952981 + 0.1224260 + 0.2226547 - 1.1078098 = -1.103178. Then you need to plug this into the logistic response function to get the predicted probability.

Final Check Save Hide Answer *You have used 1 of 2 submissions*

PROBLEM 8 - INTERPRETING MODEL COEFFICIENTS (1 point possible)

What is the meaning of the coefficient labeled "styleStiletto" in the logistic regression summary output?

○ Stilettos are predicted to have 83.3% higher odds of being sold than an average shoe in the dataset.

● Stilettos are predicted to have 83.3% higher odds of being sold than an otherwise identical open-toed shoe. ✗

○ Stilettos are predicted to have 129.9% higher odds of being sold than an average shoe in the dataset.

○ Stilettos are predicted to have 129.9% higher odds of being sold than an otherwise identical open-toed shoe. ✓

EXPLANATION

The coefficients of the model are the log odds associated with that variable; so we see that the odds of being sold are exp(0.8325406)=2.299153 those of an otherwise identical shoe in the baseline category for the style variable (which is "Open Toe"). This means the stiletto is predicted to have 129.9% higher odds of being sold.

Hide Answer *You have used 1 of 1 submissions*

PROBLEM 9 - OBTAINING TEST SET PREDICTIONS (1 point possible)

Obtain test-set predictions for your logistic regression model. Using a probability threshold of 0.5, on how many observations does the logistic regression model make a different prediction than the naive baseline model? Remember that the naive baseline model always predicts the most frequent outcome in the training set.

204

Answer: 80

EXPLANATION

Obtain test-set predictions with the predict function, remembering to pass type="response". Using table, you can see that there are 80 test-set predictions with probability 0.5 or greater.

Hide Answer *You have used 2 of 2 submissions*

PROBLEM 10 - COMPUTING TEST-SET AUC (1/1 point)

What is the test-set AUC of the logistic regression model?

0.7444244

Answer: 0.7444244

EXPLANATION

The test-set AUC can be obtained by loading the ROCR package, and then using the prediction and performance functions.

Final Check **Save** Hide Answer *You have used 1 of 2 submissions*

PROBLEM 11 - COMPUTING TEST-SET AUC (1/1 point)

What is the meaning of the AUC?

● The proportion of the time the model can differentiate between a randomly selected shoe that was sold and a randomly selected shoe that was not sold ✓

○ The proportion of the time the model correctly identifies whether or not a shoe will be sold

EXPLANATION

The AUC is the proportion of time the model can differentiate between a randomly selected true positive and true negative.

Hide Answer *You have used 1 of 1 submissions*

PROBLEM 12 - ROC CURVES (1 point possible)

Which logistic regression threshold is associated with the upper-right corner of the ROC plot (true positive rate 1 and false positive rate 1)?

○ 0 ✓

○ 0.5

● 1 ✗

EXPLANATION

A model with threshold 0 predicts 1 for all observations, yielding a 100% true positive rate and a 100% false positive rate.

Hide Answer *You have used 1 of 1 submissions*

PROBLEM 13 - ROC CURVES (1/1 point)

Plot the colorized ROC curve for the logistic regression model's performance on the test set.

At roughly which logistic regression cutoff does the model achieve a true positive rate of 80% and a false positive rate of 50%?

○ 0

● 0.16 ✓

○ 0.32

○ 0.48

○ 0.64

○ 0.81

EXPLANATION

You can plot the colorized curve by using the plot function, and adding the argument colorize=TRUE.

From the colorized curve, we can see that the light blue color, corresponding to cutoff 0.16, is associated with a true positive rate of 0.8 and false positive rate of 0.5.

Hide Answer *You have used 1 of 1 submissions*

PROBLEM 14 - CROSS-VALIDATION TO SELECT PARAMETERS (1 point possible)

Which of the following best describes how 10-fold cross-validation works when selecting between 3 different parameter values?

○ 3 models are trained on subsets of the training set and evaluated on a portion of the training set

○ 10 models are trained on subsets of the training set and evaluated on a portion of the training set

○ 30 models are trained on subsets of the training set and evaluated on a portion of the training set ✓

○ 3 models are trained on subsets of the training set and evaluated on the testing set

● 10 models are trained on subsets of the training set and evaluated on the testing set ✗

○ 30 models are trained on subsets of the training set and evaluated on the testing set

EXPLANATION

In 10-fold cross validation, the model with each parameter setting will be trained on 10 90% subsets of the training set. Hence, a total of 30 models will be trained. The models are evaluated in each case on the last 10% of the training set (not on the testing set).

Hide Answer *You have used 1 of 1 submissions*

PROBLEM 15 - CROSS-VALIDATION FOR A CART MODEL (1/1 point)

Set the random seed to 144 (even though you have already done so earlier in the problem). Then use the caret package and the train function to perform 10-fold cross validation with the data set train to select the best cp value for a CART model that predicts the dependent variable using "biddable", "startprice", "condition", "heel", "style", "color", and "material". Select the cp value from a grid consisting of the 50 values 0.001, 0.002, ..., 0.05.

What cp value maximizes the cross-validation accuracy?

0.005 **Answer:** 0.005

EXPLANATION

The cross-validation can be run by first setting the grid of cp values with the expand.grid function and setting the number of folds with the trainControl function. Then you want to use the train function to run the cross-validation.

From the output of the train function, parameter value 0.005 yields the highest cross-validation accuracy.

Final Check **Save** Hide Answer *You have used 1 of 2 submissions*

PROBLEM 16 - TRAIN CART MODEL (1/1 point)

Build and plot the CART model trained with the parameter identified in Problem 15, again predicting the dependent variable using "biddable", "startprice", "condition", "heel", "style", "color", and "material" . What variable is used most frequently as a split in the tree?

○ biddable
● startprice ✓
○ condition
○ heel
○ style
○ color
○ material

EXPLANATION

The CART model can be trained and plotted by first loading the "rpart" and "rpart.plot" packages, and then using the rpart function to build the model and the prp function to plot the tree.

Hide Answer *You have used 1 of 1 submissions*

PROBLEM 17 - BUILDING A CORPUS FROM ITEM DESCRIPTIONS (1 point possible)

In the last part of this problem, we will determine if text analytics can be used to improve the quality of predictions of which shoes will be sold.

Build a corpus called "corpus" using the "description" variable from the full data frame "eBay". Using the tm_map() function, perform the following pre-processing steps on the corpus:

1) Convert all words to lowercase, remembering to convert the corpus back to the PlainTextDocument type if necessary.

2) Remove punctuation.

3) Remove English stop words. As in the Text Analytics unit, if you have a non-standard set of English-language stop words, please load the stopwords stored in stopwords.txt and use variable sw instead of stopwords("english") when removing the stopwords.

4) Stem the document.

Build a document-term matrix called "dtm" from the preprocessed corpus. How many unique word stems are in dtm?

5994

Answer: 10530

EXPLANATION

To build a corpus, you want to use the Corpus and VectorSource functions.

You can convert all words to lowercase by using "tolower" as the second argument to the tm_map function. You may need to then convert all values in the corpus back to a PlainTextDocument by passing "PlainTextDocument" as the second argument to the tm_map function.

You can remove punctuation by using "removePunctuation" as the second argument to the tm_map function.

You can remove English stop words by using "removeWords" as the second argument to the tm_map function, and adding stopwords("english") as a third argument.

You can stem the documents by using "stemDocument" as the second argument to the tm_map function.

Lastly, you can build a document-term matrix called dtm with the DocumentTermMatrix function, and you can output the number of words by just typing dtm in your console.

Hide Answer *You have used 2 of 2 submissions*

PROBLEM 18 - REMOVING SPARSE TERMS (1/1 point)

Remove all terms that don't appear in at least 10% of documents in the corpus, storing the result in a new document term matrix called spdtm.

How many unique terms are in spdtm?

145

Answer: 145

EXPLANATION

This can be accomplished with the removeSparseTerms function.

Hide Answer *You have used 2 of 2 submissions*

PROBLEM 19 - EVALUATING WORD FREQUENCIES IN A CORPUS (1 point possible)

Convert spdtm to a data frame called descriptionText. Which word stem appears the most frequently across all descriptions?

item

Answer: ship

EXPLANATION

descriptionText can be obtained by using as.data.frame, run on as.matrix, run on spdtm.

From using the summary function or the colSums function, we can see that the word stem ship has the highest average frequency, meaning it appears the most frequently across all descriptions.

Hide Answer *You have used 2 of 2 submissions*

PROBLEM 20 - ADDING DATA FROM ORIGINAL DATA FRAME (1/1 point)

Run the following code to add a "D" in front of all the variable names in descriptionText:

names(descriptionText) = paste0("D", names(descriptionText))

Copy the following variables from the eBay data frame into descriptionText:

1) sold

2) biddable

3) startprice

4) condition

5) heel

6) style

7) color

8) material

Then, split descriptionText into a training set called trainText and a testing set called testText using the variable "spl" that was earlier used to split eBay into train and test.

How many variables are in testText?

148

Answer: 153

EXPLANATION

These steps can be accomplished by setting descriptionText$sold equal to eBay$sold, descriptionText$biddable equal to eBay$biddable, and so on.

Then you can use the subset function to create trainText and testText. From str(testText), the data frame has 153 variables.

Hide Answer *You have used 2 of 2 submissions*

PROBLEM 21 - TRAINING ANOTHER LOGISTIC REGRESSION MODEL (1 point possible)

Using trainText, train a logistic regression model called glmText to predict the dependent variable using all other variables in the data frame.

How many of the word frequencies from the description text (variables beginning with the letter "D") are significant at the p=0.05 level? (This is the total number of independent variables beginning with the letter "D" that have at least one star.)

20

Answer: 13

EXPLANATION

The new model can be trained with the glm function and summarized with the summary function.

Hide Answer *You have used 2 of 2 submissions*

PROBLEM 22 - TEST SET AUC OF NEW LOGISTIC REGRESSION MODEL (2/2 points)

What is the training-set AUC of the new logistic regression model?

0.8190665

Answer: 0.819

What is the test-set AUC of the new logistic regression model?

0.7337875

Answer: 0.734

EXPLANATION

The training-set and test-set predictions can be computed with the predict function, and the AUCs can be computed with the prediction and performance functions from the ROCR package.

Hide Answer *You have used 2 of 2 submissions*

PROBLEM 23 - ASSESSING OVERFITTING OF NEW MODEL (1/1 point)

What is the most accurate description of the new logistic regression model?

○ glmText is not overfitted, and removing variables would not improve its test-set performance.

○ glmText is not overfitted, but removing variables would improve its test-set performance.

○ glmText is overfitted, but removing variables would not improve its test-set performance.

● glmText is overfitted, and removing variables would improve its test-set performance. ✓

EXPLANATION

glmText has more variables than the base logistic regression model, but it exhibits worse test-set performance (AUC of 0.734 vs. 0.744). Therefore, it is overfitted and removing variables would improve the test-set performance.

Hide Answer *You have used 1 of 1 submissions*

‹ ›

EdX offers interactive online classes and MOOCs from the world's best universities. Online courses from MITx, HarvardX, BerkeleyX, UTx and many other universities. Topics include biology, business, chemistry, computer science, economics, finance, electronics, engineering, food and nutrition, history, humanities, law, literature, math, medicine, music, philosophy, physics, science, statistics and more. EdX is a non-profit online initiative created by founding partners Harvard and MIT.

Terms of Service and Honor Code

Privacy Policy (Revised 10/22/2014)

POWERED BY OPENedX

About edX

About

News

Contact

FAQ

edX Blog

Donate to edX

Jobs at edX

Follow Us

Facebook

Twitter

LinkedIn

Google+

Tumblr

Meetup

Reddit

Youtube