edX (https://www.edx.org)

MITx: 15.071x The Analytics Edge

sandipan_dey (/dashboard) ▾

Courseware (/courses/MITx/15.071x/1T2014/courseware)

Course Info (/courses/MITx/15.071x/1T2014/info)

Discussion (/courses/MITx/15.071x/1T2014/discussion/forum)

Progress (/courses/MITx/15.071x/1T2014/progress)

Help

yllabus (/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/)

chedule (/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/)

PREDICTING THE POPULARITY OF NEWS STORIES

Newspapers and online news aggregators like Google News need to prioritize news stories to determine which will be the most popular. In this problem, you will predict the popularity of a set of New York Times articles containing the words "Google", "Microsoft", or "Yahoo" from the time period May 2012-December 2013. The dependent variable in this problem is the variable **popular**, which labels if an article had 100 or more comments in its online comment section. The independent variables consist of a number of pieces of article metadata available at the time of publication:

- **print**: 1 if an article appeared in the print edition, 0 if only online
- **type**: the type of the article, either "Blog," "News," or "Other"
- **snippet**: a text snippet from the article
- **headline**: the text headline of the article
- **word.count**: the number of words in the article

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

Load nytimes.csv (/c4x/MITx/15.071x/asset/nytimes.csv) into a data frame called articles, using the stringsAsFactors=FALSE option.

EXPLANATION

This can be done with the read.csv function.

What proportion of articles had at least 100 comments?

0.1079137

0.1079137

Answer: 0.1079137

EXPLANATION

If you use the table function, you can see that 105 of the 973 articles had the popular variable set to 1, for a proportion of 105/973=0.1079137.

Hide Answer

You have used 2 of 2 submissions

PROBLEM 2 - COMPUTING A CORRELATION (1/1 point)

What is the correlation between the number of characters in an article's headline and whether the popular flag is set?

-0.1126912

-0.1126912

Answer: -0.1126912

EXPLANATION

This can be obtained with the cor function.

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

PROBLEM 3 - CONVERTING VARIABLES TO FACTORS (1/1 point)

Convert the "popular" and "type" variables to be factor variables with the as.factor() function.

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 4 - SPLITTING INTO A TRAINING AND TESTING SET (1/1 point)

Set the random seed to 144 and then obtain a 70/30 training/testing split using the sample.split() function from the caTools package. Store the split variable in a variable called "spl", which we will use later on. Split articles into a training data frame called "train" and a testing data frame called "test".

EXPLANATION

The split can be accomplished by first using the sample.split fuction, giving as a first argument the dependent variable "popular", and then using 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 5 - TRAINING A LOGISTIC REGRESSION MODEL (1/1 point)

Train a logistic regression model (using the train data frame) to predict the "popular" outcome, using variables "print", "type", and "word.count".

Which of the following coefficients are significant at the p=0.05 level (at least one star)?

- ☑ print ✓
- ☑ typeNews ✓
- ☐ typeOther
- ☑ word.count ✓

EXPLANATION

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

All of the variables except typeOther are significant at the p=0.05 level.

Hide Answer *You have used 1 of 1 submissions*

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

Consider an article that was printed in the newspaper (print = 1) with type = "News" and a total word count of 682. What is the predicted probability of this observation being popular, according to this model?

0.09351271

0.09351271

Answer: 0.09351285

EXPLANATION

This observation has print=1, typeNews=1, typeOther=0, and word.count=682, so it has a logistic function parameter of -2.5075573-0.8468333+0.9055929+682*0.0002600 = -2.271478. 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 7 - INTERPRETING MODEL COEFFICIENTS (1 point possible)

What is the meaning of the coefficient on the print variable in the logistic regression model?

○ Articles from the print section of the newspaper are predicted to have 42.9% lower odds of being popular than other articles.

○ Articles from the print section of the newspaper are predicted to have 57.1% lower odds of being popular than other articles.

○ Articles from the print section of the newspaper are predicted to have 84.7% lower odds of being popular than other articles.

○ Articles from the print section of the newspaper are predicted to have 42.9% lower odds of being popular than an otherwise identical article not from the print section.

○ Articles from the print section of the newspaper are predicted to have 57.1% lower odds of being popular than an otherwise identical article not from the print section. ✓

● Articles from the print section of the newspaper are predicted to have 84.7% lower odds of being popular than an otherwise identical article not from the print section. ✗

EXPLANATION

The coefficients of the model are the log odds associated with that variable; so we see that the odds of being popular are exp(-0.8468333)=0.4287706 those of an otherwise identical non-print article. This means the print article is predicted to have 57.1% lower odds of being popular.

Hide Answer *You have used 1 of 1 submissions*

PROBLEM 8 - OBTAINING TEST SET PREDICTIONS (1/1 point)

Obtain test-set predictions for your logistic regression model. Using a probability threshold of 0.5, on how many observations does the logistic regression 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.

0

0

Answer: 0

EXPLANATION

We can obtain test-set predictions using the predict function. If you then summarize your predictions, you can see that the maximum predicted probability of being popular is 0.488, so no observations will be labeled as popular using a threshold of 0.5. As a result, the logistic regression predictions exactly coincide with the predictions of the naive baseline model.

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

PROBLEM 9 - COMPUTING TEST SET AUC (1/1 point)

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

0.7853598

0.7853598

Answer: 0.7853598

EXPLANATION

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

Hide Answer *You have used 2 of 2 submissions*

PROBLEM 10 - 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 popular and a randomly selected non-popular article ✓
- ○ The proportion of the time the model correctly identifies whether or not an article is popular
- ○ The relative strength of the model compared to the naive baseline model

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 11 - ROC CURVES (1/1 point)

Which cutoffs (or thresholds) are plotted on an ROC curve for a logistic regression model?

- ○ No cutoffs
- ○ Only the cutoff 0.5
- ○ Only the cutoff yielding the maximum training set accuracy
- ○ Only the cutoff yielding the maximum testing set accuracy
- ● All cutoffs between 0 and 1 ✓

EXPLANATION

The ROC curve plots the true and false positive rates for all cutoffs between 0 and 1.

Hide Answer *You have used 1 of 1 submissions*

PROBLEM 12 - READING ROC CURVES (1/1 point)

Plot the colorized ROC curve for the logistic regression model.

At roughly which logistic regression cutoff does the model achieve a true positive rate of 0.39 and a false positive rate of 0.04?

- ○ 0.02
- ● 0.22 ✓
- ○ 0.42
- ○ 0.62
- ○ 0.82

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 green color, corresponding to cutoff 0.22, is associated with a true positive rate of 0.39 and false positive rate of 0.04.

Hide Answer *You have used 1 of 1 submissions*

PROBLEM 13 - CROSS-VALIDATION TO SELECT PARAMETERS (1/1 point)

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 14 - 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 "print", "type", and "word.count". Select the cp value from a grid consisting of the 50 values 0.01, 0.02, ..., 0.5.

How many of the 50 parameter values achieve the maximum cross-validation accuracy?

50

50

Answer: 50

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, all 50 parameters yielded the exact same accuracy.

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

PROBLEM 15 - TRAIN CART MODEL (1/1 point)

Build and plot the CART model trained with cp=0.01. How many variables are used as splits in this tree?

0

0

Answer: 0

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.

It turns out that there is no tree plotted, because there are no splits.

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

PROBLEM 16 - BUILDING A CORPUS FROM ARTICLE SNIPPETS (1/1 point)

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

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

1) Convert all words to lowercase

2) Remove punctuation

3) Remove English stop words. As in the Text Analytics week, if you have a non-standard set of English-language stop words, please load the stopwords stored in stopwords.txt (/c4x/MITx/15.071x/asset/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?

3926

3926

Answer: 3926

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 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 17 - REMOVING SPARSE TERMS (1/1 point)

Remove all terms that don't appear in at least 5% of documents in the corpus, storing the result in a new document term matrix called spdtm.

How many unique terms are in spdtm?

17

17

Answer: 17

EXPLANATION

This can be accomplished with the removeSparseTerms function.

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

PROBLEM 18 - EVALUATING WORD FREQUENCIES IN A CORPUS (1/1 point)

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

compani

Answer: compani

EXPLANATION

articleText 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 compani has the highest average frequency, meaning it appears the most frequently across all snippets.

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

PROBLEM 19 - ADDING DATA FROM ORIGINAL DATA FRAME (1/1 point)

Copy the following variables from the articles data frame into articleText:

1) print

2) type

3) word.count

4) popular

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

How many variables are in testText?

21

21

Answer: 21

EXPLANATION

These steps can be accomplished by setting articleText$print equal to articles$print, articleText$type equal to articles$type, articleText$word.count equal to articles$word.count, and articleText$popular equal to articles$popular.

Then you can use the subset function to create trainText and testText. From str(testText), the data frame has 21 variables.

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

PROBLEM 20 - TRAINING ANOTHER LOGISTIC REGRESSION MODEL (1/1 point)

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 snippet text are significant at the p=0.05 level?

0

0

Answer: 0

EXPLANATION

The new model can be trained with the glm function and summarized with the summary function. The only significant terms are the intercept, print, typeNews, and word.count, none of which are word frequencies from the snippet text.

Hide Answer *You have used 2 of 2 submissions*

PROBLEM 21 - TEST SET AUC OF NEW LOGISTIC REGRESSION MODEL (1/1 point)

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

0.6852357

0.6852357

Answer: 0.6852357

EXPLANATION

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

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

PROBLEM 22 - ASSESSING OVERFITTING OF NEW MODEL (1 point possible)

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 significantly worse test-set performance (AUC of 0.685 vs. 0.785). Therefore, removing variables would improve the test-set performance (e.g. removing all word frequencies would improve test-set AUC by 0.100).

Hide Answer *You have used 1 of 1 submissions*

About (https://www.edx.org/about-us) Jobs (https://www.edx.org/jobs)
Press (https://www.edx.org/press) FAQ (https://www.edx.org/student-faq)
Contact (https://www.edx.org/contact)

EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.

(http://www.meetup.com/edX-Global-Community/)

(http://www.facebook.com/EdxOnline)

(https://twitter.com/edXOnline)

(https://plus.google.com/108235383044095082

(http://youtube.com/user/edxonline)

Terms of Service and Honor Code -
Privacy Policy (https://www.edx.org/edx-privacy-policy)