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

MITx: 15.071x The Analytics Edge

zhushun0008 (/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)**

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

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

Help

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 point possible)

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

What proportion of articles had at least 100 comments?

Check Save *You have used 0 of 2 submissions*

PROBLEM 2 - COMPUTING A CORRELATION (1 point possible)

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

=

Check Save *You have used 0 of 2 submissions*

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

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)

Final Check Save *You have used 0 of 1 submissions*

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

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".

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

Final Check Save *You have used 0 of 1 submissions*

PROBLEM 5 - TRAINING A LOGISTIC REGRESSION MODEL (1 point possible)

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

Final Check Save *You have used 0 of 1 submissions*

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

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?

Check Save *You have used 0 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.

Final Check | Save | *You have used 0 of 1 submissions*

PROBLEM 8 - 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 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.

Check | Save | *You have used 0 of 2 submissions*

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

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

Check | Save | *You have used 0 of 2 submissions*

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

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

Final Check Save *You have used 0 of 1 submissions*

PROBLEM 11 - ROC CURVES (1 point possible)

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

Final Check Save *You have used 0 of 1 submissions*

PROBLEM 12 - READING ROC CURVES (1 point possible)

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

Final Check Save *You have used 0 of 1 submissions*

PROBLEM 13 - 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

Final Check Save *You have used 0 of 1 submissions*

PROBLEM 14 - CROSS-VALIDATION FOR A CART MODEL (1 point possible)

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?

Check Save *You have used 0 of 2 submissions*

PROBLEM 15 - TRAIN CART MODEL (1 point possible)

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

Check Save *You have used 0 of 2 submissions*

PROBLEM 16 - BUILDING A CORPUS FROM ARTICLE SNIPPETS (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 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?

Check Save *You have used 0 of 2 submissions*

PROBLEM 17 - REMOVING SPARSE TERMS (1 point possible)

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?

Check Save *You have used 0 of 2 submissions*

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

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

Check Save *You have used 0 of 2 submissions*

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

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?

Check Save *You have used 0 of 2 submissions*

PROBLEM 20 - 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 snippet text are significant at the p=0.05 level?

Check Save *You have used 0 of 2 submissions*

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

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

Check Save *You have used 0 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.

Final Check Save *You have used 0 of 1 submissions*