Quora Question Pairs

An edWisor.com’s project

Submitted by: Sandhya Parkar

Table of Contents

Contents

Quora Question Pair

1.Problem Statement:

To build a binary classification model to predict which of the provided pair of questions contain two questions with the same meaning.

2.Business understanding:

Quora is a platform where over 100 million people visit every month, so it’s no surprise that many people ask similar qustions, Multiple questions with the same intent can cause seekers to spend more time finding the best answer to their question, and make writers feel they need to answer multiple versions of the same questions.

3.Data Understanding:

We have two sets of data of which one is train set and another is test set.

The train data set contains 404290 observations and 6 variables namely:

a. id: id of training set question pairs
b. q1id: Unique id of question1
c. q2id: Unique id of question2
d. question1: Full text of question 1
e. question2: Full text of question 2
f. is_duplicate: The target variable, it is set to 1 if both the questions have the same meaning and set to 0 if not

The test data set contains 2345796 observations and 3 variables namely:

a. test_id: id of the test set question pair
b. question1: Full text of question1
c. question2: Full text of question2

```
> str(train)
'data.frame':	404290 obs. of  6 variables:
 $ id          : int  0 1 2 3 4 5 6 7 8 9 ...
 $ qid1        : int  1 3 5 7 9 11 13 15 17 19 ...
 $ qid2        : int  2 4 6 8 10 12 14 16 18 20 ...
 $ question1   : Factor w/ 290457 levels "\177what is the average cost you pa
id for a day-trading course & how long was the course & do you find that it h
as seriously he"| __truncated__,..: 220871 220936 38667 263584 255694 6184 12
9722 33453 238925 125480 ...
 $ question2   : Factor w/ 299175 levels "","\177how to minimize the loss fun
ction? \177",..: 228833 245080 45973 29348 257402 97697 233519 238101 246943
60624 ...
 $ is_duplicate: int  0 0 0 0 0 1 0 1 0 0 ...
```

4. Data Preprocessing:

Our data have two variables question1 and question2 which contains text data, so we have to perform text mining on both question1 and question2. We have defined custom functions as delete_html(), remove_linebreaks(), clean_string(), and rm_words() to clean our text data. We used regular expressions in these function to clean our text data.

delete_html() delete all html tags from our text data.

Remove_linebreaks() removes line breaks from our text data.

clean_string() converts the text into lower case, remove numbers, new line characters, special characters and punctuation from our text data.

rm_words() removes specified words (i.e. stop words) from our text data so as to clean it.

There are certain specified operations which we need to perform on textual data for text mining which are fulfilled by above mentioned functions. These operations are described below:

a. **Case Folding**: For case folding purpose we have defined a function clean_string(), which converts all text data into lower case using a predefined function tolower() to provide consistency between similar words.

b. **Removing HTML tags**: There is no significance of html tags in text mining so we clean it to improve the performance of our system by using delete_html() function.

c. **Remove Numbers:** Removing all numbers from our text data using regular expression in clean_string() function.

d. **Remove Line breaks:** Line breaks and spaces is of no meaning to our data so we clean it using function remove_linebreaks().

e. **Remove Punctuation:** Punctuation marks like full stops, commas etc do not bring any special meaning to our data so we clean it using clean_string() function.

Apart from this we extracted the important variables only by removing q1id and q2id variables from our test and train set.

5. Feature Engineering:

The predictive power of a model can be increased significantly if supplied with the right feature engineered variable inputs. In Quora question pairs project, we are supplied with textual data as inputs which in itself cannot help much in prediction. We have created some new features to improve the predictive power of our model. Each of these variables are mentioned in detail below. We have used text2vec package in our analysis.

a. **Combing variables of same importance:** We have combined question1 and question2 in a single variable question.

b. **Calculating distance based score:** We tried to calculate distance between question1 and question2 using cosine similarity but it has not created any significant impact on our analysis, so we dropped the variable.

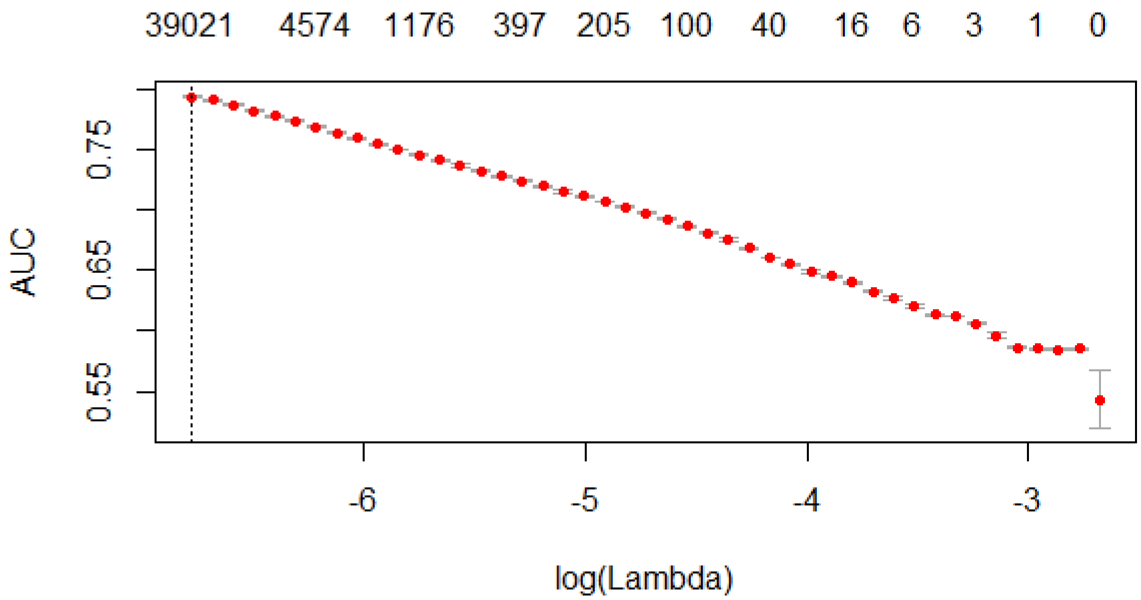
c. **Calculating iterators:** We have created iterator over tokens using itoken() function. It helps data to process in a memory friendly chunks.

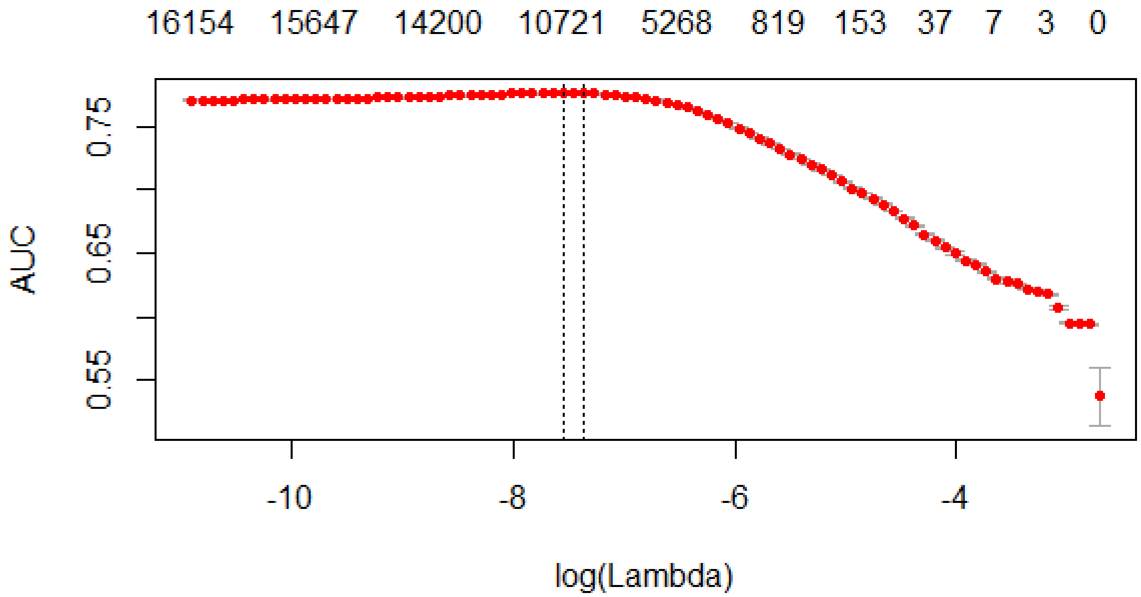
d. **Creating Vocabulary:** We built vocabulary with create_vocablury() function.

e. **Tf-idf transformation :** tf-idf transformation is done using text2vec package. The tf-idf are then provided as an input to our predictive model.

6. Visualization:

We performed visualization using plot() function on different steps.





7. Predictive Modelling:

For Modelling we considered binomial logistic regression using glmnet package with 4 fold cross validation for faster model processing. We tried various modelling techniques of which Binomial Logistic Regression Model was finalized due to highest accuracy.

Binomial Logistic Regression: Our model was able to predict target variable with accuracy between 85-90 %. We performed various steps for achieving highest accuracy.

The steps are described below:

a. We first created document term matrix for both train and test data and used 4 fold cross validation for model building.
b. We applied same preprocessing techniques and feature engineering variables on our test data and predicted model performance on test data. In this way we achieved approx. 72% accuracy.
c. To reduce training time and also to improve performance we used pruned vocabulary and created vectorizer and tested our model using this on our test data. Here we achieved an accuracy of 78% and also reduced training time.

d. Further we used n gram to predict our model. We used upto 2 grams. We achieved an accuracy of 81%.
e. We also used feature hashing but training time was reduce significantly but a small decrease in accuracy occurred.
f. Then we used tf-idf transformation and achieved a highest accuracy of 90%.

We finalized our model using tf-idf transformation.

7. Conclusion:

To predict whether two questions are duplicate or not we performed various techniques on our data. We have performed preprocessing and feature engineering on our data. We tested various models for our data to provide better accuracy and our finalized model is Binomial logistic regression using tf-idf transformation.

Applying this model we obtained a mean accuracy of 85%.

Thank You