Quora Duplicate Question Pairs (Kaggle)

By: Piyush Bhattad, Vaibhav Prasad Desai

Quora recently announced the first public dataset that they ever released. It includes 404351 question pairs with a label column indicating if they are duplicate or not.

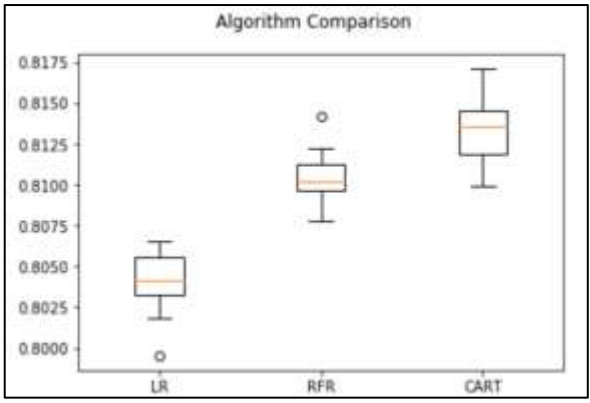
Issues: There are 255045 negative (non-duplicate) and 149306 positive (duplicate) instances. This induces a class imbalance however when you consider the nature of the problem, it seems reasonable to keep the same data bias with your ML model since negative instances are more expectable in a real-life scenario.

Exploratory data Analysis: When we analyze the data, the shortest question is 1 character long (which is stupid and useless for the task) and the longest question is 1169 character. The average length is 59.

We have removed any questions whose total character length is less than 9.Also instead of removing all stop words we have focused only on removal of words like ('the', 'a',an,).We have even stemmed the words to improve accuracy.

Also as the question set was very large to process we are focusing on sample (50k data points) to build the model.

The most important part in this dataset for semantic analysis was word representation in form of vectors. We are using simple bag of words concept to derive vectors for questions which are further used to build main features cosine similarity between question pair and Euclidean distance which are used for the relation with the response variable 'is_duplicate'.



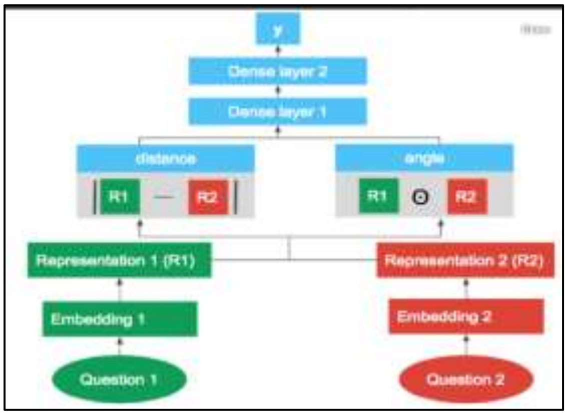
As the response variable is binary we start model training with logistic regression for classification and then append machine learning methods like "Decision Tree Classifier" and "Random Forest Classifier"

- LR: 0.803986 (0.002032) :: Logistic
- RFR: 0.810484 (0.001740) ::RandomForest
- CART: 0.813451 (0.002067) :: Decision Tree

To improve the score we will further use genism library Doc2vec function to represent sentences in vector form with over 300 dimensions which will be multiplied with TFIDF score and finally mean of summation of vectors will be performed.

On the above cosine similarity and instance function will be derived along with additional features like POS tagging and word match share.

These will be passed as inputs to neural network with two dense layers to predict the labels.



Reference:

https://engineering.quora.com/Semantic-Question-Matching-with-Deep-Learning

Quora competition discussion board: https://www.kaggle.com/c/quora-question-pairs/discussion