**Deliverable 3**

**Quora Question Pairs**

# **Overview**

Millions of online users visit Quora’s website daily to ask questions and seek answers. Making the process of finding the best answers quickly and accurately is the aim of the service. With the volume of users, similar questions create extraneous information that starts to obfuscate the best answers from being found. Quora is addressing the problem by delegating the process of analyzing, predicting, and identifying duplicate questions to machine learning models. Optimization of a model for performing this task is the key to making their service more viable for their users and the challenge posed to the data science community.

# **Folder layout:**

## Root**:**

**Main App:** quora.py

**Support files:** textcleaner.py**,** graphs.py

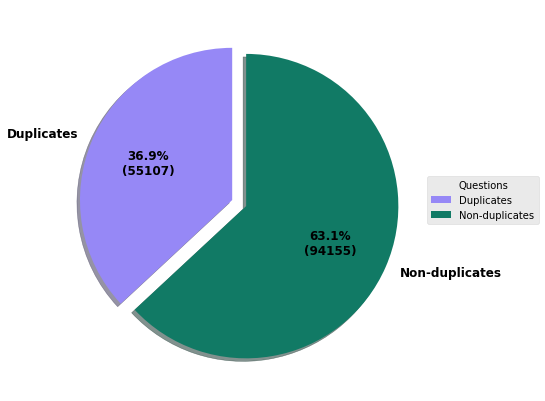
## Saved Notebooks:

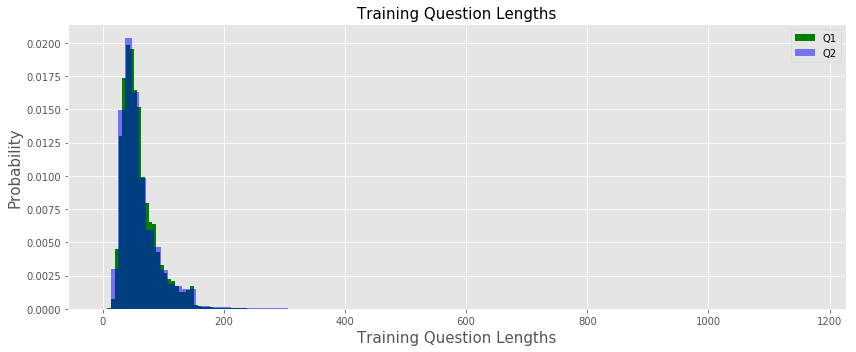
**Juypter format:** quora.ipynb – precision results (0.78% acc)

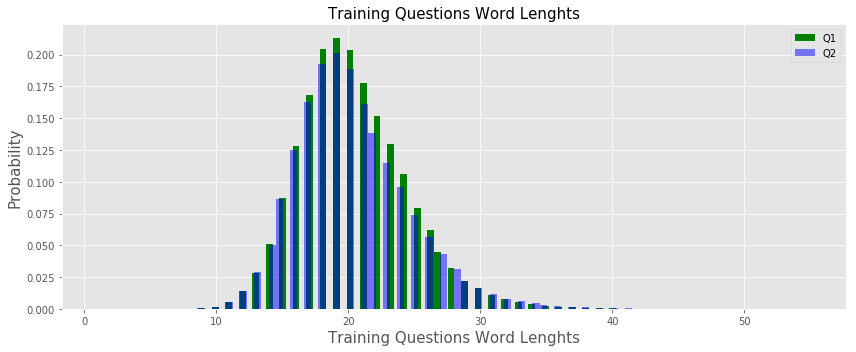
# **Data Exploration**

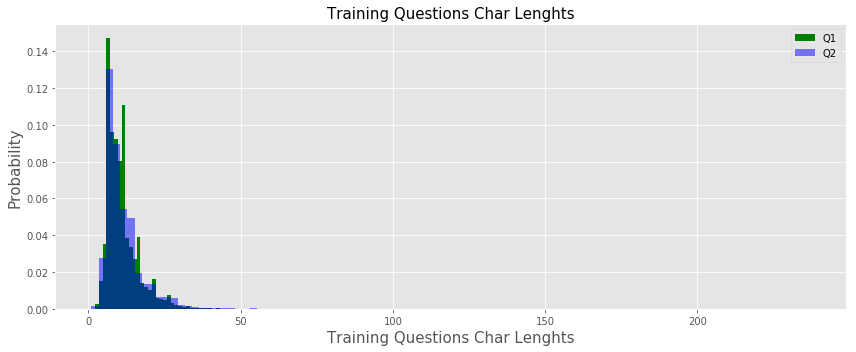
Using the raw training and test data there isn’t a lot of information to obtain without vectorizing and feature engineering. Below are the graphs showing relations between different fields in the raw dataset.

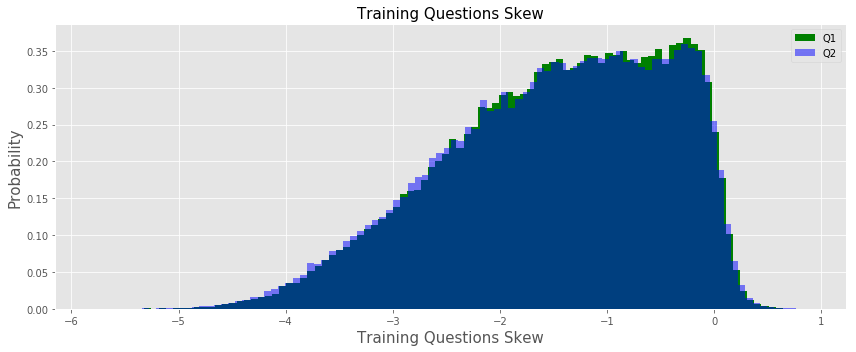
**The duplicate counts and distribution**

****



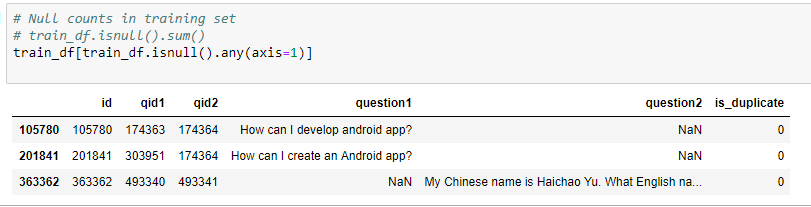
****

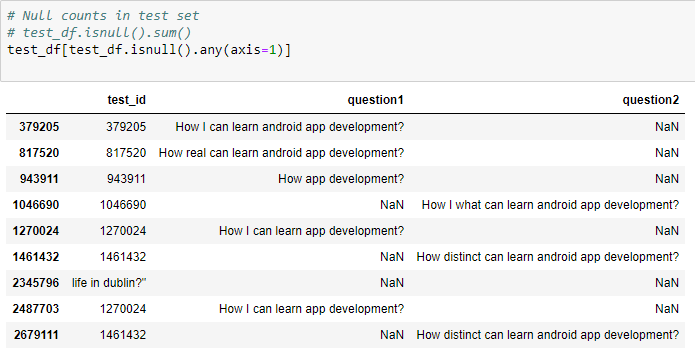
****

****

**Missing data**

There were several rows in both the raw training and test data files that needed to be dropped since they had ‘NaN’ values.

****

****

# Preprocessing

## **Text cleaning**

In order to achieve a higher degree of accuracy when comparing questions, the raw data should be scrubbed of any text (special characters, misspelled words, numbers, etc.) that will hinder the prediction results. Using the guidance from prior competitors referenced below I attempted to make fine adjustments to a function that uses regular expressions to correct misspelled words, replace certain special characters and number formats, convert contractions, but in the end it over manipulated the data to the point where it only hindered the prediction results since I was never able to achieve better accuracy results using it. The best results were obtained from cleaning only during the vectorization process with word2vec embeddings and I only removed stop words and non-alpha characters.

### Raw text

| **id** | **qid1** | **qid2** | **question1** | **question2** | **is\_duplicate** |
| --- | --- | --- | --- | --- | --- |
| **0** | 0 | 1 | 2 | What is the step by step guide to invest in sh... | What is the step by step guide to invest in sh... | 0 |
| **1** | 1 | 3 | 4 | What is the story of Kohinoor (Koh-i-Noor) Dia... | What would happen if the Indian government sto... | 0 |
| **2** | 2 | 5 | 6 | How can I increase the speed of my internet co... | How can Internet speed be increased by hacking... | 0 |
| **3** | 3 | 7 | 8 | Why am I mentally very lonely? How can I solve... | Find the remainder when [math]23^{24}[/math] i... | 0 |

### After applying text clean

| **id** | **qid1** | **qid2** | **question1** | **question2** | **is\_duplicate** |
| --- | --- | --- | --- | --- | --- |
| **0** | 0 | 1 | 2 | what is the step by step guide to invest in sh... | what is the step by step guide to invest in sh... | 0 |
| **1** | 1 | 3 | 4 | what is the story of kohinoor koh i noor diamond | what would happen if the indian government ste... | 0 |
| **2** | 2 | 5 | 6 | how can i increase the speed of my internet co... | how can internet speed be increase by hack thr... | 0 |

Ref**.** <https://www.kaggle.com/currie32/the-importance-of-cleaning-text>

## Vectorize

Working with large data sources such as Quora’s it’s necessary to use some sort of method to vectorize the texts to incorporate robust feature engineering and semantics to produce accurate prediction results within a reasonable timeframe. I used Gensim Word2Vec using Standford Glove pretrained word embeddings along with TFIDF vectorizer to compute distances between two texts and find their relations. The libraries that are useful for computing the distances are Fuzzywuzzy, Word Mover’s Distance (WMD), and the scipy spatial.distance library.

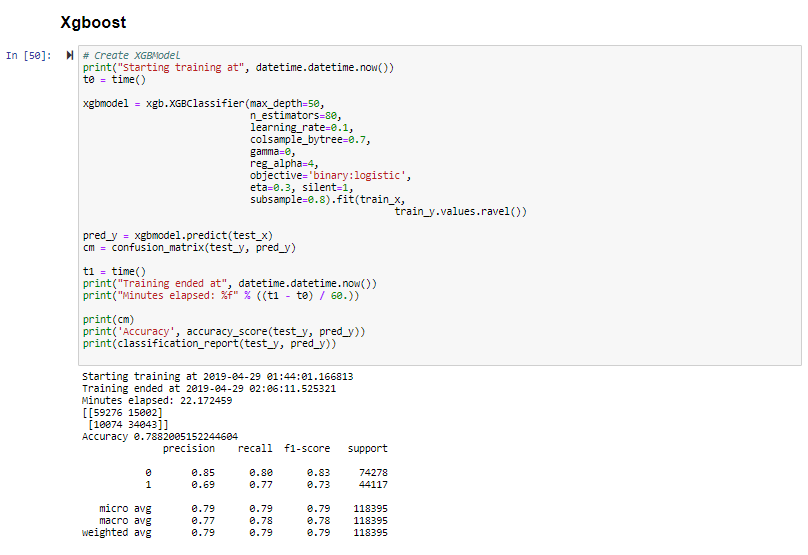
## Feature Engineering

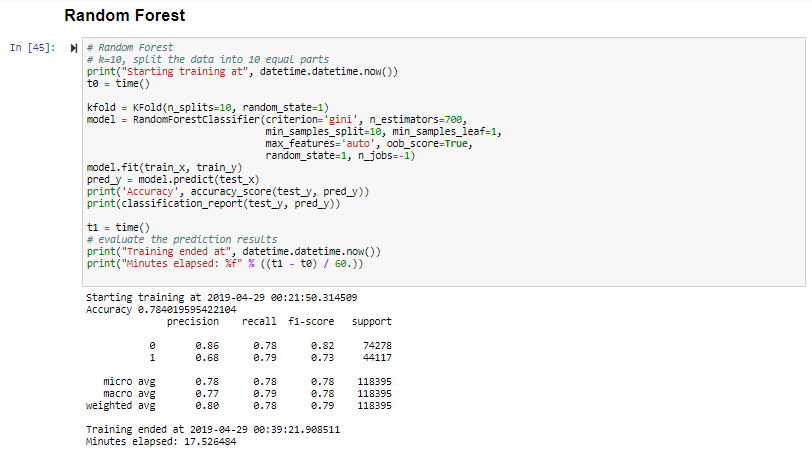
I incorporated 29 features into the data set using a variety of basic calculations and more advanced utilizing vectorization.

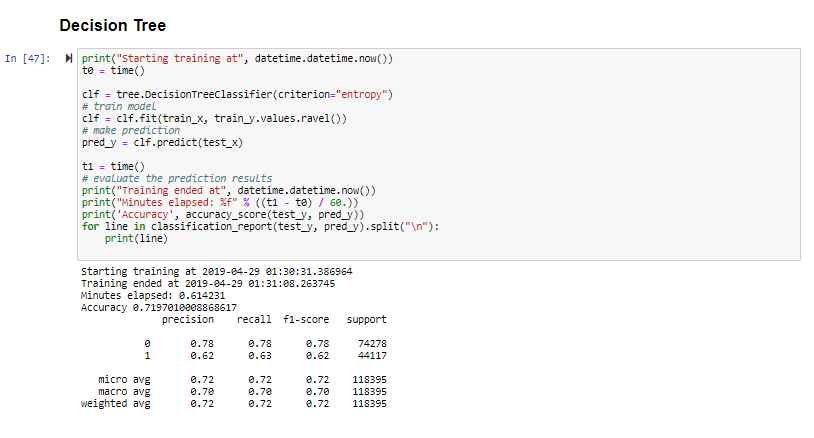


## **EVALUATE MODEL**

Xgboost obtained the best accuracy results, the other models below were also tried.



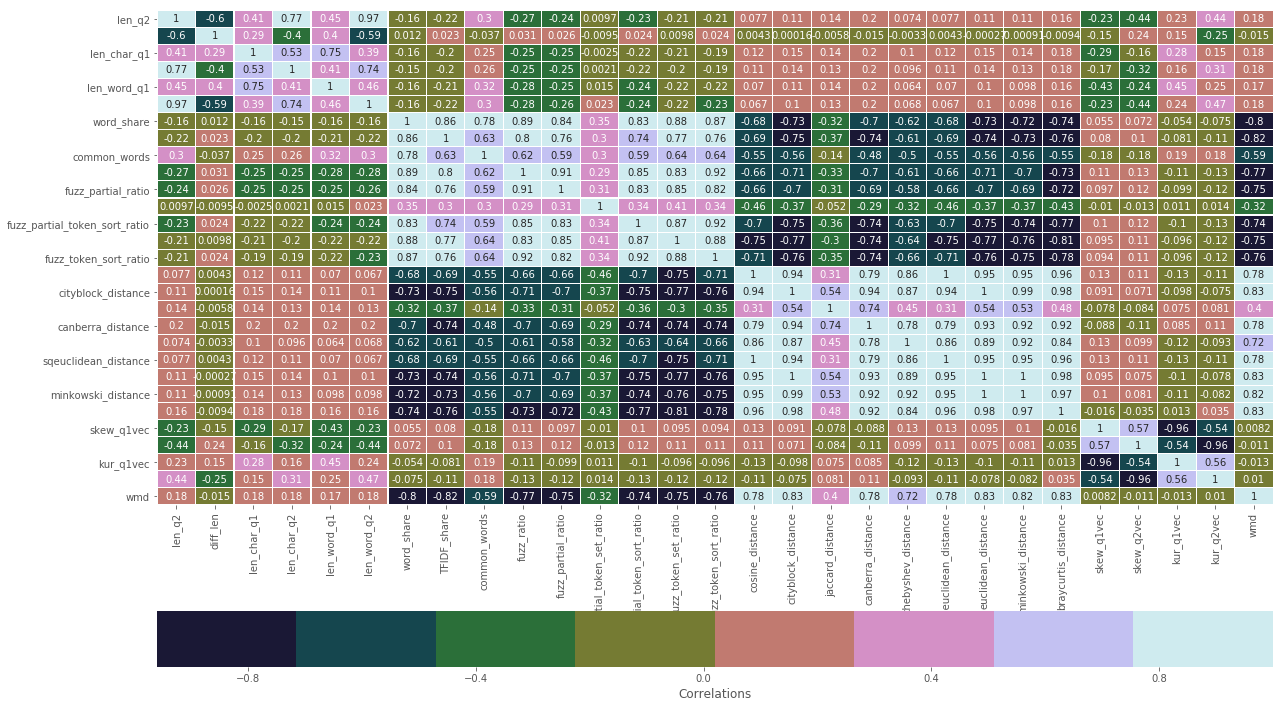




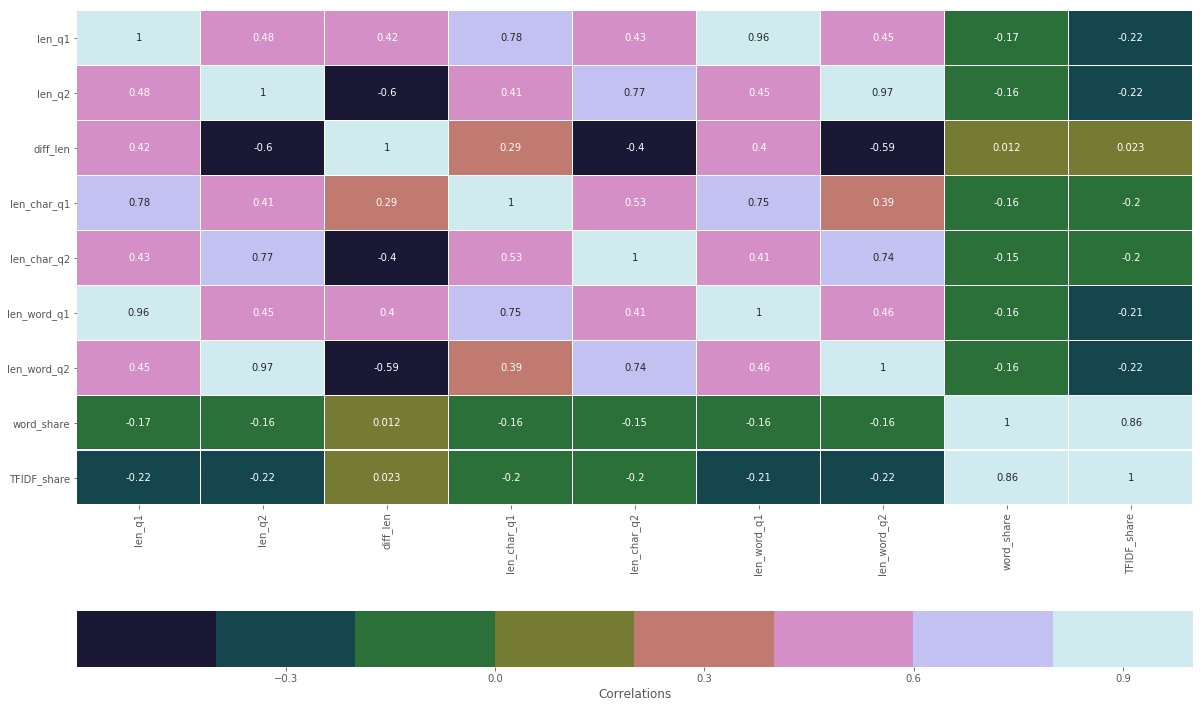
## **Recommendations**

In the future I will continue to try and improve results using several tools that came to light after finishing this project. The Seaborn heatmap graphs below shows how well the features selected correlate with one another. Using this I will remove certain features with a bad correlation value and reevaluate the models.

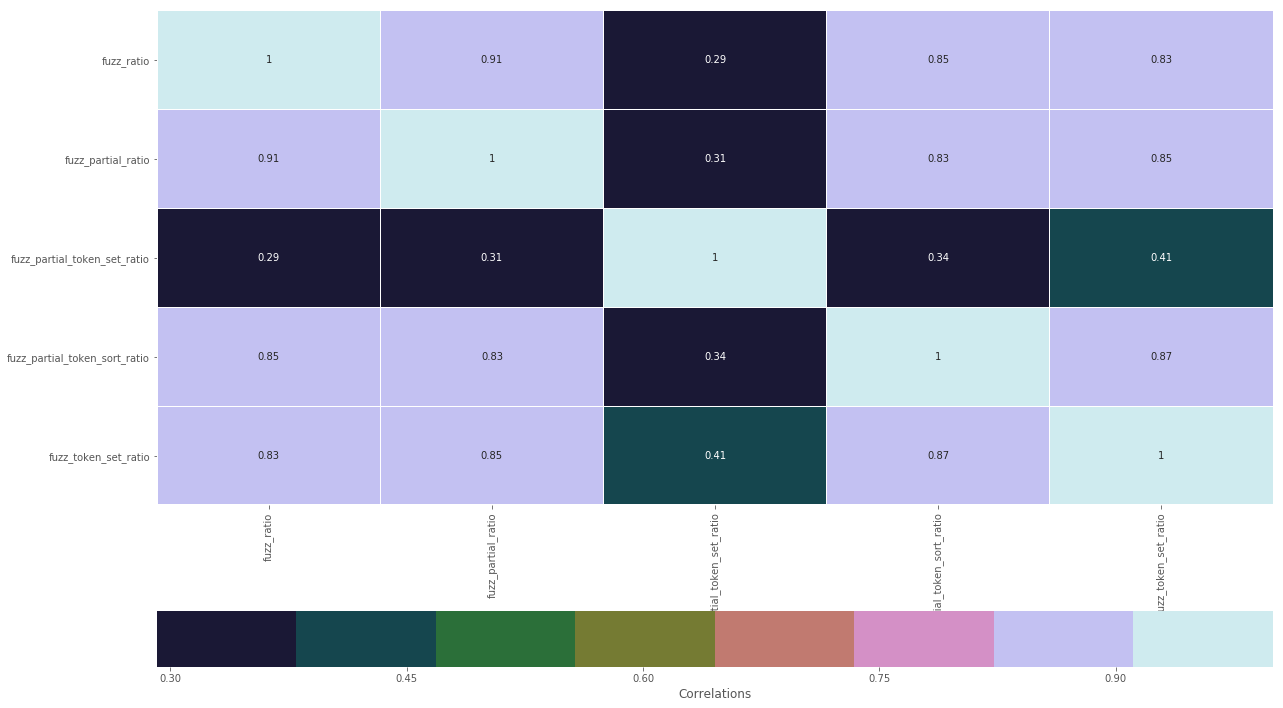
**All Features**

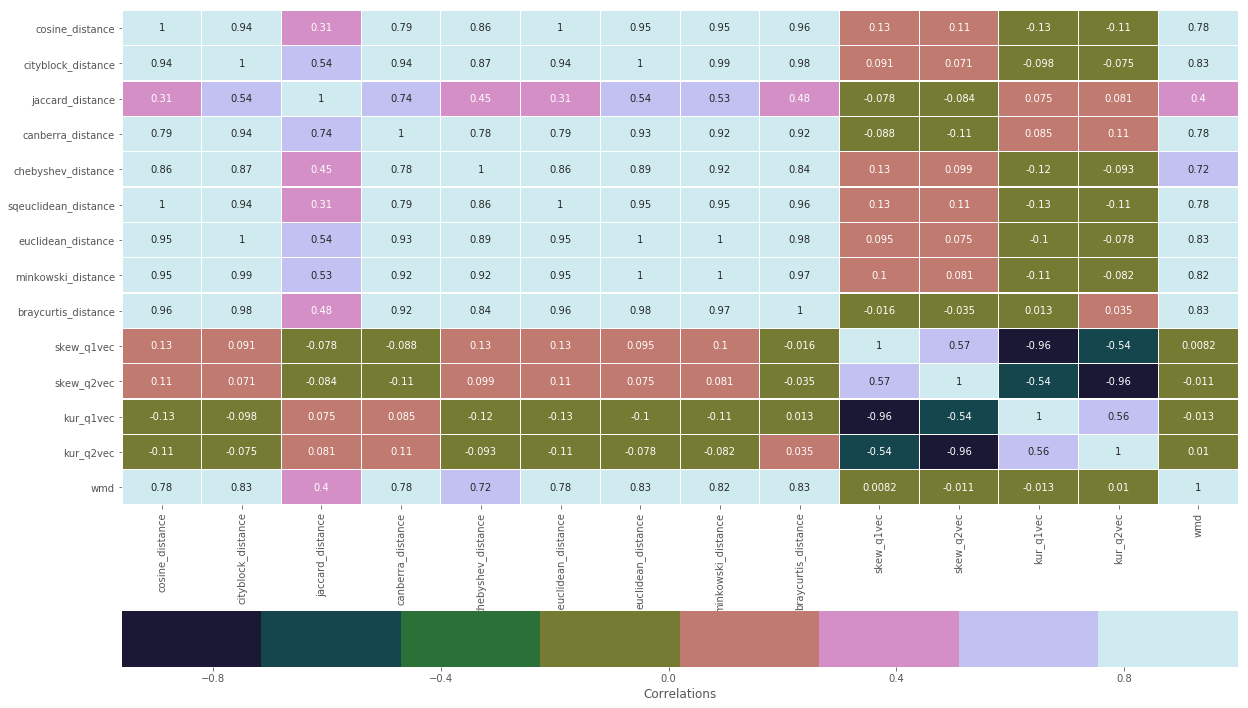


**Basic Features**

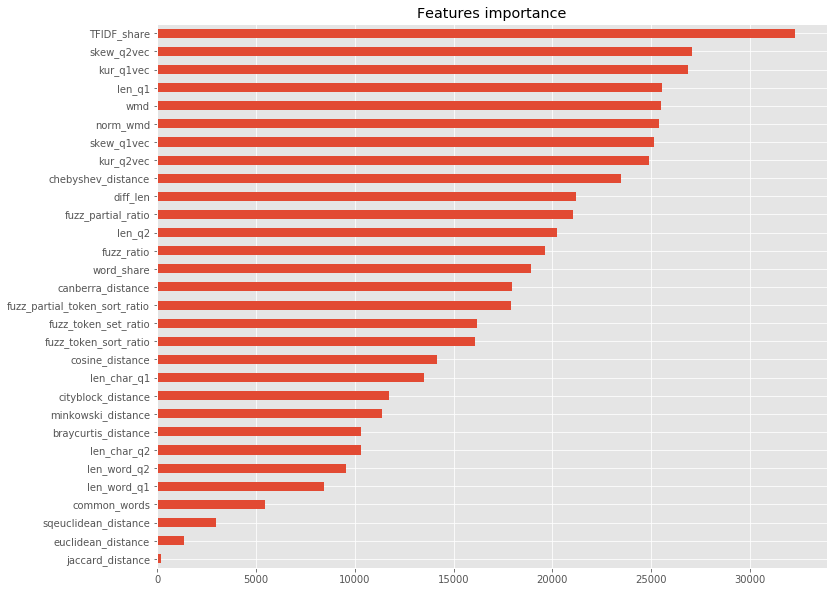


**Fuzzy Ratio Features**



**Distance Features**

**Feature Importance with XGBOOST**

****