

# Quora Question Pair Similarity Problem.

Problem Analysis.

Problem Statement: Quora is an internet platform where users from across the world post questions that they want to be answered and other users volunteer to answer them.

The Problem here is there can be multiple users asking the same question in different ways of word usage and two questions may mean to ask for the answer.

Due to the sheer scale of the number of questions being asked from across the world, it is desirable to find if one question is duplicate of the other and merge the questions. Furthermore, it improves the customer experience to get answer immediately if a question they intend to ask has already been answered, possibly due to another similarly “worded” question.

Main Objectives.

If the model misclassifies a question that is not a duplicate for being duplicate, It can be costly as the consequence would be to lose the questions and there by affecting main objective of the problem. Hence model accuracy is not negligible.

Evaluation Metric:

Instead of having a binary answer whether is question is similar or not, it is beneficial to know how the similar a question is to other, Therefore probability of questions being similar can be the ideal way to evaluate the metric as it gives a chance to set higher expectations of similarity in the future.

It is trivial to know if the model is interpretable because the end goal is to find if the questions asked are duplicates and not how they became duplicates.

A binary classification problem evaluating using **log-loss** metric – Because we optimize the model based on probability of question being a duplicate or not.

Also, we inspect the binary confusion to monitor the model performance.

Data Literacy:

The data consists of **404290** rows when downloaded which will further be subjected to pre-processing. Data attributes being "id","qid1","qid2","question1","question2","is\_duplicate".

Id: Unique ID of each row in the dataset

Qid1, Qid2: Question ID of questions 1 and 2 respectively

Question1 and Question2 : The Questions asked in text

Is\_duplicate? : The flag indicating if both the questions are duplicates or not.

Splitting the data into train and test in the ratio 80:20, by random.

Extra thoughts: The decision of model and actual result can differ due to the temporal nature of the data trained and tested. It would be sensible to train the model on the old data and test it on the new data. because the questions though bring duplicate can a temporal perspective. For example Narendra modi can be the prime minister now but may not be in the future, therefore a question asked can be mean to refer to other situation. Therefore a timestamp along with the questions would greatly contributed to the accuracy of the model in real life scenario