Quora-Question-Pair-Similarity

Business Problem:

Description:

Quora is a place to gain and share knowledge—about anything. It’s a platform to ask questions and connect with people who contribute unique insights and quality answers. This empowers people to learn from each other and to better understand the world. Over 100 million people visit Quora every month, so it's no surprise that many people ask similarly worded questions. 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 question. Quora values canonical questions because they provide a better experience to active seekers and writers and offer more value to both groups in the long term.

Problem Statement:

- Identify which questions asked on Quora are duplicates of questions that have already been asked.
- This could be useful to instantly provide answers to questions that have already been answered.
- We are tasked with predicting whether a pair of questions are duplicates or not.

Real world/Business Objectives and Constraints:

1. The cost of a misclassification can be very high.
2. You would want a probability of a pair of questions to be duplicates so that you can choose any threshold of choice.
3. No strict latency concerns.
4. Interpretability is partially important.

Type of Machine Learning Problem:

It is a binary classification problem, for a given pair of questions we need to predict if they are duplicate or not.

Performance Metric:

1. log-loss
2. Binary Confusion Matrix

Train and Test Construction:

We build, train and test by randomly splitting in the ratio of 70:30 or 80:20 whatever we choose as we have sufficient points to work with.

Sources/Useful Links:

- Source: https://www.kaggle.com/c/quora-question-pairs
- Discussions: https://www.kaggle.com/anokas/data-analysis-xgboost-starter-0-35460-lb/comments
- Kaggle Winning Solution and other approaches on this Quora question pair analysis: https://www.dropbox.com/sh/93968nfnrzh8bp5/AACZdtsApc1QSTQc7X0H3QZ5a?dl=0
- Blog 1: https://engineering.quora.com/Semantic-Question-Matching-with-Deep-Learning
- Blog 2: https://towardsdatascience.com/identifying-duplicate-questions-on-quora-top-12-on-kaggle-4c1cf93f1c30