1 code

The code for this solution is at: https://github.com/wurenzhi/CS4401X-Spring2021-Project

2 Solution outline

The solution includes five steps: (1) Data reading and EDA, (2) Blocking, (3) Feature engineering, (4) Model training and (5) Generating output.

2.1 Data reading and EDA

In this step, we read the left table and right table, as well as the training set. We explore the dataset to get some ideas of designing the solution. For example, we found that the left table has 2554 rows and the right table has 22074 rows, so there are 2554*22074=56376996 pairs. Examining every pair is very inefficient, so we will need a blocking step to reduced the number of pairs that we will work on.

2.2 Blocking

We perform blocking on the attribute "brand", generating a candidate set of id pairs where the two ids in each pair share the same brand. This is based on the intuition that two products with different brand are unlikely to be the same entity. Our blocking method reduces the number of pairs from 56376996 to 256606.

2.3 Feature engineering

For each pair in the candidate set, we generate a feature vector of 10 dimensions by obtaining the jaccard similarity and levenshtein distance of pair on the five attributes. In this way, we obtain a feature matrix X_c for the candidate set. We do the same to the pairs in the training set to obtain a feature matrix X_t. The labels for the training set is denoted as y_t.

2.4 Model training

We use a random forest classifier. We train the model on (X_t, y_t). Since the number of non-matches is much more than the number of matches in the training set, we set class_weight="balanced" in random forest to handle this training data imbalance problem. We perform prediction on X_c to get predicted labels y_c for the candidate set.

2.5 Generating output

The pairs with $y_c = 1$ are our predicted matching pairs M. We remove the matching pairs already in the training set from M to obtain M^-. Finally, we save M^- to output.csv