Danny Nguyen

The code for this solution is at: https://github.com/dnguyen352/CS4400X

Similarity to the sample solution, I implemented data reading and EDA first. After that, I performed blocking to reduce number of pairs to be compared. Beside blocking on the attribute “brand”, I performed blocking on “modelno” attribute and checked other cases which could happen between the matching entities. For example, “modelno” of this item can be presented in other item’s title. I used pairs2LR function in the sample solution to combine the matching entities into a table. I also used feature engineering with 3 attributes instead of 5 attributes to obtain a feature matrix X_c for the candidate set. I also used Model training provided in the sample solution to handle the training data imbalance issue. Finally, I generated the output of predicted matching pair.