Cameron Moore

4/25/2021

CS 4400 FINAL PROJECT ML

Public Github Repository: https://github.com/camjm21/CS4400-FinalProjML

Solution:

I have had a frustrating, but ultimately gratifying experience with learning the basics of Machine Learning. I tried several solutions, but kept coming back to this one. In this solution, I mainly used the import module: py_entitymatching.

Through this module, I was able to split up my training data into a training set and testing set. This helped me verify that I had a higher precision and F1 score than the sample solution. I used a blocking entity to block by brand and this was fairly simple through the module. The module also helped me determine which matching learner I wanted to use by running multiple learners and comparing their F1 scores. I decided on the Random Forest matcher as it consistently gave me the best results with regards to precision, recall, and F1. Also, through this module, I discovered several helpful features and was able to pile them into a dataframe to run the Random Forest matcher on.

Overall, through testing, I had an F1 score of .47 which is higher than the sample solution. Applying the Random Forest matcher to the candidate set, I came away with an output around 350 rows most of the time. The one I am submitting has 372 rows. Manually checking through them I have some hits but also some misses, but that is what ML is about.

In terms of similarity to the sample solution, I used the same way to output my csv file with the predicted matches. I also used the pairs2LR to set up the training dataframe. Other than that, I used guides and trial and error to work with the rest of the code.