INFSCI 2725: Term Project Progress Report

By Rebecca Ly, Don Stuart and JoAnn Dwornicki

Due 3/14

What is Kaggle?

Kaggle is a community of data scientists who compete in various competitions to solve some of the most complex business problems. Topics for these competitions range from the life sciences to engineering to information technology and more. Kaggle works as a place for individuals and experts to compete and collaborate in competitions hosted by businesses where businesses seek solutions from data scientists.

These competitions solve real-life problems and Kaggle provides data-driven results to companies in order to solve real-world problems. Teams compete against one another in a constant competition to solve real-world problems and comprehensively implement solutions.

Kaggle is a way for students from all over the world to meet one another to collaborate and discuss methods and ideas from various fields ranging from computer science, statistics, finance and more.

Home Depot Problem

Home Depot wants to improve their search relevancy when their customers enter search queries on their website looking for the most up-to-date tools and equipment. Participants in this competition would need to develop a model that can accurately predict the relevance of the search results between 1 (poor) and 3 (accurate).

Search relevancy refers to how accurate results are compared to the search term that is entered. An example can be looking up restaurants in the area but getting only coffee shops within the area. Individuals will get frustrated when they don’t receive the results they are immediately looking for and often abandon the search. Home Depot is hoping to avoid that type of problem and would like Kagglers to predict the search relevancy of their current model so that they can identify the gaps and improve upon it. Home Depot uses search relevancy to predict how quickly customers are able to look up the products they are searching for.

What We’ve Done

Over the past week, we’ve looked into studying more about R and Python to see which would be most suitable for this type of competition. Both R and Python have their pros and cons, some of which include R being more suitable for analyzing data and Python for manipulating data.

We have done extensive research regarding the structure of Kaggle and the Home Depot competition itself. We wanted to make sure we had a full understanding of the project data and the task at hand. We looked at the testing and training sets of the data as well as the sample set in order to gain more insight on how to best utilize all of it. We’ve looked into the sample scripts posted online on the Kaggle website to help generate ideas. There appeared to be plenty of information posted by fellow competition participants, which contained helpful advice for Kaggle beginners such as ourselves. Some links pointed to website called DataCamp that has endless online interactive data training courses which have proved to be very lucrative thus far.  Aside from looking on Kaggle and DataCamp, we have also been watching video tutorials on Youtube to help further solidify our understanding of both Python and R.

The Main Goal and How to Accomplish It

After thorough examination of the datasets and forum posts on Kaggle, it appears the main goal of this competition not to predict the true relevancy scores of the search item/query pairs, but to actually replicate the human ratings and improve upon the efficiency of the current system that Home Depot has in place. There is some sample search item/query pairs listed complete with the average human relevancy scores in the training dataset. Our goal will be to replicate this system, but with even more efficient and accurate ratings that of which are unaffected by constantly changing algorithms.

Our plan of attack so to speak will start with meticulous combing through of the training csv file to try and seek out any kind of patterns of note. This expedition may include the use of statistical analysis and other similar tools found in the Microsoft Excel program. Python and R will most likely also be used for analysis of this file. We will place importance on the item pairs that have perfect relevancy scores of 3. We will try to figure out why they have perfect scores and compare/contrast them with those that have irrelevant scores of 1 or close to it. The model or algorithms used will probably have a close resemblance to the exhaustive grid search model and may also contain bits and pieces from the search engine method described in the R Bloggers link that was given to us on Courseweb. Once these above tactics are further researched by us, and then eventually implemented, we believe they will provide a more than optimal solution to this business problem.