Capstone Three-Project Proposal

# Introduction

The Home Depot, Inc. is largest home improvement retailer in the United States. It Supplies tools, construction products and services. It operates so many products that shoppers rely on Home Depot’s products and to get timely solutions for their home improvement needs. It does not easy to set up such a huge search database to service the clients. The customers need a fast, accuracy and reliable search results when they type some particular terms in the Home Depot website. Based on the train, test and product descriptions tables, we want to develop a model that can accurately predict the relevance of search results.

# Problem

Given a set of data: train, test and product descriptions, there are only 11 columns which is not big for data scientist. But more importantly, except the relevance score column, all other useful columns are text, which is string in python. We need some NLP technology to transfer those columns into numerical columns, therefore, we can process those information for machine learning models.

# Input

the input is 4 columns which are the ‘distances between search term and title & description’, cosine similarity between search term and title & description’

# Output

The output is the predict relevance score which range is 0~3(0 is no relevant, 3 is most relevant)

# ML Technique

We use two NLP tech: Levenshtein distance and TF-IDF to create some new features. Then in the model part, we use Random Forest, Gradient Boosting and XGB to do the regression. Finally, we compare the mean suqared error(MSE) , one of an estimator measures the average of the squares of errors—that is, the average squared difference between the estimated values and the actual values.

# dataset

The detailled description and the dataset itself can be found under the following

https://www.kaggle.com/c/home-depot-product-search-relevance/data