Term Project – Home Depot Product Search Relevance

Team: Venkatesh Duvvuri (VED14@pitt.edu); Haifa Alnasser (HIA11@pitt.edu); Gopi Tata (GKT3@pitt.edu)

A. Introduction:

This is the term project report for the Data Analytics course that explains the problem to be solved which is the Home Depot Product Search Relevance. The project report explains the problem, data that has been provided and to be used, the statistical model and methods to be used in the prediction process of the outcome. The report has been divided into various sections viz., Problem Defined, Competition Guidelines, Analysis of Data Provided, Project Methodology and Implementing the Solution. Finally in the Appedix section we have provided explanation of the Python packages, functions utilized, the plots obtained, the result and other supporting details.

B. Problem Defined:

Home Depot Product Search Relevance is a Kaggle Competition to predict relevance score for the provided combinations of products and search terms. All search terms and result are related to home improvement, gardening, construction and do-it-yourself projects.

In this task, we need to determine to what extent a search result matches the search query that it is paired with. The criticality in this task is about the intent and relevancy of the search query and not about the opinion of relevancy.

C. Competition Guidelines:

There are certain guidelines provided in the competition based on which the relevancy of the search and results are determined and categorized. Actual Search item is provided by a user and is processed by the search engine and should return the title, image, and may be for certain searches the URLs of the products. Each pair (Search Term and Product Retrieved) was evaluated by at least 3 Human raters and the average was recorded in the training dataset. The relevance of the product of a search item (or query) is categorized as Irrelevant, Partially or somewhat relevant, or Perfect match. The three categories are further defined as below.

Perfect match if –

a) The result matches the query intent perfectly.
b) The searched product is a part of a kit or included as a part or an item in the product shown.
c) The product is exactly what the customer was most likely trying to find.
d) The query was for a brand and the result is made by that brand.

Partially or somewhat relevant if:

a) The result generally matches the query, but there is ambiguity in either search terms or product details that prevents from categorizing as 'Perfect Match'.
b) Actual product is the same but differs in brand, dimensions, color or specifications.
c) Product specifications are not clear or product is a somewhat ambiguous synonym of the search term.
d) One of the many products searched in the query are shown.

Irrelevant (Paired with Search by Mistake) if:

a) The result is not at all related to the intent of the search term.
b) The result is an item that is used with search term such as tool, accessory, extra part, cover or case.
c) The result has a completely different meaning, even if there are words in common with the search term.
d) The result is a collection of items and the search term does not appear in the set.

D. Analysis of the Data Provided:

Home Depot had provided various data namely the train dataset, test dataset, Product Attributes and Product descriptions data of various home improvement related products in .csv format for this product search and relevance competition.

Product_Description.csv: In this file, product_description, there are about 124428 unique products with their descriptions.

Attributes.csv – this file contains the attributes of various products. There about 2044803 entries. The data provided is the name and value for various products. Apparently there are about 155 entries that are not tagged to any product ; in other words there are 155 entries without a product_id attached

to them and may have to be excluded while determining the relevancy of a product for a given search item.

Test.csv: There are 166693 entries in this dataset consisting of various search terms, products description for various products.

Train.csv: There are 74067 entries in train.csv dataset consisting of various search terms, products description for various products and search relevance score.

We further observed that there are 86264 unique products (product_uid) in attributes.csv, 54667 unique products (product_uid) in train.csv, and 97460 unique products (product_uid) in test.csv

E. Project Methodology:

a) Selecting the Statistical Model:

For any problem to be researched and intended to solve statistically the selection of the statistical model to be used is critical. Given the multiple statistical models available selecting the model depends upon the complexity and the nature of the problem. We started to evaluate the suitability or application of multiple statistical models studied during the course on after the other. Given the Linear Regression suits best or used to predict quantitative values certainly not a good candidate for this project.

Though Linear Discriminant analysis is used to predict qualitative values this particular project task not about merely categorizing the output as such and hence not strong candidate model either. The linear regression and linear discriminant analysis both are exclusively linear methods whereas the predicting the relevance of the search outcome is not linear. That being said the other choice is to explore the Classification and Regression Trees for the model section.

The various methods of Classification and Regression Trees models like Bagging, Boosting and Random Forests are being explored and used for this non-linear classification. Also as we learned during the course that the ensemble technique that combines multiple models in order to achieve an improved performance of prediction we decided to use Bagging and Random Forest methods of Regression Trees models.

b) Data Cleansing:

For any statistical process the data being used in the models plays a key role. Using a valid clean data is certainly the key for various test runs of a model in the predictive analytics. This very fundamental data cleansing necessity prompted us to explore various methods or tools to have the data as clean as possible.

We found that scripting language Python as a rich collection of functions and packages related text mining like stemming in order to cleanse the data and remove any misspelled words, special characters etc. to make the test data as unambiguous as possible. We have a choice between Porter and Snowball stemming with the Snowball Stemming being the latest and the most advanced Stemming algorithm available out there so we would be using the snowball stemming by Natural Language Toolkit (NLTK) Package in Python.

c) Data Validation:

Data validation is typical foremost task of any statistical or software testing in order to eliminate bad data that otherwise would affect the testing results negatively. We would validate data and may exclude any product attributes that do not belong to any products or product descriptions that have blank product ids etc.

d) Creating Data Structures:

We intended to keep all the available data as one big data source against which we can run-through the search item in order to make the prediction as opposed to creating relational data model. We would be using the Data Frame data structure in this project as we believe that is appropriate for the merged datasets.

F. Implementing the Solution:

To begin with we loaded both the Train and Test Data This task was followed by merging the Product Descriptions data set into a concatenated Data Frame of the Train and Test data sets. With that we have got all the data related to the Product, Product Search terms into a single data structure. Then, we included the data from attributes.csv file in the data frame generated above.

Then as part of data validation and cleansing we created few methods for correcting typos in the provided search query by the user to reduce the uncertainty in the prediction values of the search algorithm being used for suggesting the products. Then we used stemming to increase the uniformity

of the search again. In this way our prediction assures that we don't leave out the granularity of the search algorithm being used by Home Depot.

We had a choice between Porter and Snowball stemming; the Snowball Stemming being the latest and the most advanced Stemming algorithm available out there we have used the same provided by Natural Language Toolkit (NLTK) Package in Python.

Next, we found the number of matching words between the search query in Product Title and Product Description individually giving us an overall relevance of the search query and the product.

Finally we used Random Forest Regression and bagging to fit the Training model and predict the relevance between search query and products in the test data.

G. Appedix:

a) Various Python Packages used:

i. Numpy – (np) - *import numpy as np*. This statement will allow us to access NumPy objects using np.X instead of numpy.X.

ii. Pandas (pd) - *import pandas as pd. Pandas* is a Python package providing fast, flexible, and expressive data structures designed to make working with "relational" or "labeled" data both easy and intuitive. Pandas is well suited for many different kinds of data: Tabular data with heterogeneously-typed columns, as in an SQL table or Excel spreadsheet

iii. Pandas.read – to *Read* data file like .csv into Data Frame.

iv. Pandas.concat - The *concat* function (in the main pandas namespace) does all of the heavy lifting of performing concatenation operations along an axis

v. Pandas.merge – with this *merge* statement, pandas has full-featured, high performance in-memory join operations idiomatically very similar to relational databases like SQL.

vi. DataFrame - *DataFrame* is one of the data structures available in Python.It is a two dimensional labeled data structure with columns of potentially different types. It like a spreadsheet or SQL table and accepts many different kinds of input. It is generally the most commonly used pandas object

vii. Stemming Process - *Stemming* is part of a composite process of extracting words from text and turning them into index terms. The idea of stemming is to improve information retrieval performance

generally by bringing under one heading variant forms of a word which share a common meaning.

viii. Random Forest Regressor - A *Random Forest* is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and use averaging to improve the predictive accuracy and control over-fitting.

ix. BaggingRegressor - A *Bagging Regressor* is an ensemble meta-estimator that fits base regressors each on random subsets of the original dataset and then aggregate their individual predictions (either by voting or by averaging) to form a final prediction.

x. SnowballStemmer - *Snowball* is a small string processing language designed for creating stemming algorithms for use in Information Retrieval. Snowball is a language in which stemming algorithms can be easily represented.