

**W. P. CAREY SCHOOL OF BUSINESS HONOR CODE MASTER’S PROGRAMS**

We, the students of the W.P. Carey Master’s programs, are committed to maintaining the highest levels of honesty, integrity, and professional courtesy for ourselves and amongst our classmates. We believe that ethical behavior is not only the cornerstone of a good academic program, but of good business as well. As such, we are steadfast in our commitment to the following principles:

* Academic Integrity – We view the W.P. Carey Academic Integrity Policy as a living, breathing document, not just concepts on a page. As such, we are resolute in ensuring that we abide by the principles contained within, both in letter and in spirit. Furthermore, we will hold our classmates to the same high standards to which we hold ourselves.

* Consideration of Others – As W. P. Carey students, we believe that we should treat others as we wish to be treated. To this end, we will be considerate of classmates, professors, and recruiters in all aspects. Furthermore, we will constructively add to the collaborative environment by helping and encouraging others.
* Professionalism – As W. P. Carey students, we believe that the behaviors we practice today will become habit in the future. To that end, we are committed to act in a professional manner in all school settings, including classes, networking events, and interviews. We will be cognizant of the fact that we are representing the program and our classmates at all times.

Full Name (printed): \_\_\_\_\_Ankita Tomar\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Signature\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Date \_\_4/24/2016\_\_\_\_\_\_\_\_\_\_\_

Full Name (printed): \_\_\_\_\_Harsh Pathak\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Signature\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Date \_\_4/24/2016\_\_\_\_\_\_\_\_\_\_\_

Full Name (printed): \_\_\_\_\_Srinivasa Bala Rudraraju\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Signature\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Date \_\_4/24/2016\_\_\_\_\_\_\_\_\_\_\_

Full Name (printed): \_\_\_\_\_Yuwei Zhu\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Signature\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Date \_\_4/24/2016\_\_\_\_\_\_\_\_\_\_\_

Full Name (printed): \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Signature\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Date \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Onsite MSBA Applied Project Report

Spring 2016

W.P. Carey, ASU

|  |  |
| --- | --- |
| Topic | Home Depot Product Search Relevance |
| Team | A-13 |
| Team Members | Ankita Tomar, Harsh Pathak, Srinivasa Rudraraju, Yuwei Zhu |
| Client information | Kaggle Competition – Home Depot |

# **Executive Summary**

The aim of the Capstone Project is to determine Home Depot Search relevance scores as accurately as possibly by using machine learning techniques. This model automates the existing, time taking process of manually calculating the search relevance scores using human raters. Updating the search algorithms will help search the products easily on the Home Depot website and therefore, be more customer friendly.

The project was initiated by diving deep into the data sets provided to see which information would be helpful in predicting the relevance scores. This process helped us understand the important aspects to be considered from the given datasets in the next phase of the project. The next essential step was cleaning the data to ensure we had no anomalies in the data. We applied text mining techniques to clean up all the text before it could be used for feature selection. As part of feature selection we either compared the search terms with product title, description and brand name to get the counts, ratios or used ‘TF-IDF Vectorizer’ with ‘ngram’ to extract important text features. The model building was done using Python and Azure ML. We proceeded with Python after we maximized our accuracy in Azure. We built models using various regression methodologies like Random Forest Regression, XGBoost and Support Vector Regression in Python. Finally Grid Search was used to optimally select the parameters to get the best possible accuracy with the models that were built during the project. We obtained the best results using Random Forest Regression.

Finally, we have a model built in Python, which can determine the product relevance score with respect to the search term. This should help the analytics team at Home Depot to automate the search relevance scoring process. Their teams can now improve the search algorithms further by running numerous iterations to improve search results for customers than spending a lot of time with current manual process. This has a direct impact on customer experience which leads to increased sales if the customers find what they want easily on the website. The model learnt on the dataset provided and hence, has its own limitations as to the accuracy we can expect with other datasets. We may see variations in accuracy if we get too many new search terms with a lot of spelling mistakes. Also, for bigger datasets the model may take longer time but will certainly be faster than the manual process. Handling the spelling mistakes automatically in the context of the search term, improving on performance of the algorithms by running on large data sets and trying deep learning methodologies will be additional areas that can be worked on as an extension to the current project. It was a tremendous learning experience for us as a team to work on a real world problem and enhance our knowledge on machine learning. We hope to build on the knowledge we gained and apply in future projects.

# **Background**

Customers depend on Home Depot’s product authority website to find, compare and buy the latest products and get timely solutions to their home improvement needs. Customers expect instantaneous and relevant search results with great accuracy and efficiency for their search terms on the website. Home Depot uses search relevancy as an implicit measure to gauge how quickly they can get right products to the customers. Search relevance is the relevance of the products returned as a result of search query and a score is assigned to each such Search term and Product pair.

Currently, the company uses human raters to evaluate the impact of potential changes to their search algorithms, which is a slow and subjective process. The search relevance score is a number between 1 and 3 with 1 – not relevant, 2 – mildly relevant and 3 – highly relevant. Each pair of product description and search term are evaluated by at least 3 human raters in order to come up with the average score which will be used while improvising the search algorithms in Home Depot. The human raters do not have access to the product attributes but they do have access to the product images. The entire process is very tedious and time consuming as it involves manual intervention.

More details can be found at the below Kaggle competition link:

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

# **Problem Statement**

The existing process of calculating Search Relevance Scores at Home Depot has a high dependency on human intervention for both, score prediction and iteration based algorithm - accuracy improvement. This is a resource intensive process and is not a scalable solution for Home Depot.

The main goal of the project is to use machine learning techniques to predict the search relevance scores. This would help reduce or eliminate the human intervention and help update the search algorithms as dynamically as possible. Building an automated process to predict search relevance scores will help customers get more accurate search results. This will in turn help the business in achieving more sales as the customers will be able to find what they want accurately and quickly.

The designed model for predicting the search relevance scores will be used by the Home Depot Analytics teams to update their search algorithms. The best models can potentially be deployed to achieve automation in the entire search relevance score prediction, which is of financial significance both in terms of reducing costs and increasing business, by better search results. The key assumptions while planning, executing and finalizing the results were that the search relevance score will be dependent on the match of the words in search term with product title, product description and attributes given as insights to develop the predictive model. The model does not have the privilege of comparing an image like the human raters do while they can see the image by going to the URL as well. The relevance score is a continuous variable between 1 and 3 can be predicted by encoding the numerical features we extract from the text features provided in the train, test, attributes and descriptions datasets provided to work on the project.

Our main aim with the assumptions we had in mind was to: –

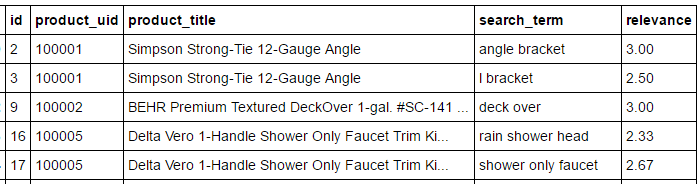
* Use Descriptive Statistics/Text Mining techniques to analyze the trends, patterns and exceptions in the data set which can be used to treat missing data, outliers, and skewness in some of the input variables and effective feature selection in the predictive analytics stage.
* Use Data Mining/ Machine Learning methods for regression as part of the predictive analytics step to train and test the data in order to determine the relevance score. The evaluation methodology of the accuracy is the RMSE (Root Mean Squared Error) of the predicted vs. actual relevance score in the test data set.

# **Methods**

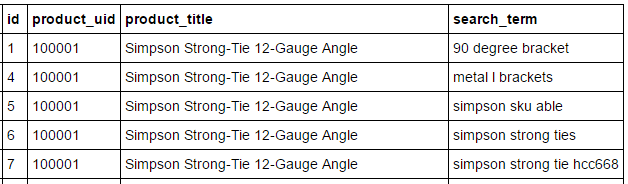
The overall project was executed in a sequential manner using the below sub-tasks with key decisions taken at each and every level to arrive at the final model. The best model with highest accuracy was an ensemble of 3 best Random Forest Models with weight of 0.5, 0.3 and 0.2. Data Exploration was done to bring in additional insights from attributes and description datasets. Text Data was processed to eliminate spelling mistakes, special characters, punctuation, special notation and finally stemming and removing stop words. Text Data was encoded into numerical features by taking the frequencies, counts, and ratios of the search term occurrence in product title, description and brand name. Additional features were encoded using ‘TF-IDF Vectorizer’ to create sparse matrix and encode the features of the columns considered. Finally Grid Search was used to optimize the parameters and obtain the best accuracy possible with this methodology. A detailed approach at each and every level is discussed in the sub-tasks below.

**Data Exploration**

The first step to execution was to do an end-to-end analysis of the datasets provided with the problem. The snapshot of the test and train sets is provided below.

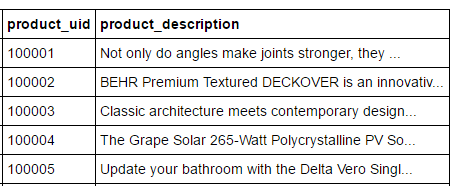


**Fig.1 - Train dataset**

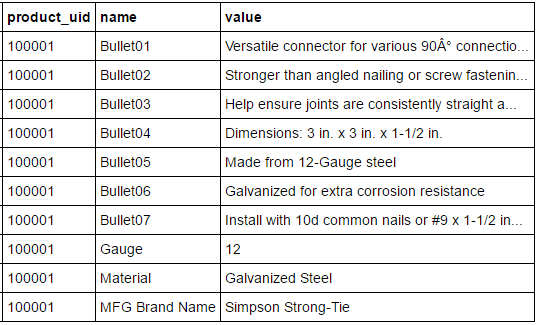


**Fig.2 - Test dataset**

Exploring the data sets provided in the problem, we see that the relevance score can be predicted based on the matching terms in search\_term and product\_title fields. We also needed to bring in additional insights from product description and attributes sheet, which provides more information while predicting relevance score. The descriptions dataset had the product\_description, which was a much broader and complete description of the products. We also explored the attributes dataset to see which attribute was frequently occurring while using that information in the overall dataset to predict the search relevance scores.

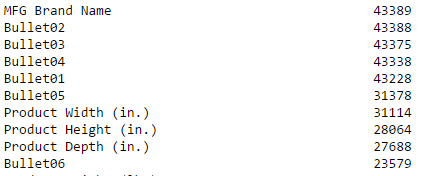


**Fig.3 - Description dataset**

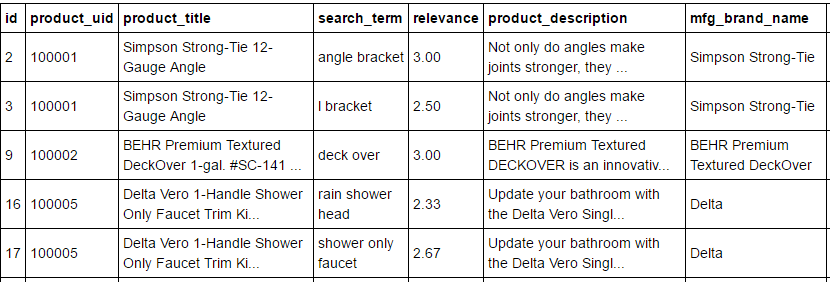


**Fig.4 - Attribute dataset**

The frequency of the attribute field was considered to pick the best attribute, which may help in predicting the relevance score. Hence we considered the brand name as another column, which will bring insights while comparing the search terms. The additional data we used before we did the data cleansing and feature extraction was to use the product\_description and brand\_name fields from the datasets provided for the project.



**Fig.5 Frequency of Attributes**



**Fig.6 - Training data with product\_description and mfg\_brand\_name**

**Data Cleansing**

The training and test data sets don’t have any missing data. However, when we integrated the description data and brand name from product description and attributes sheet, we see that the brand names for the data is missing which we replaced with “Unknown Values”.

Text mining techniques like converting text into lower case, eliminating punctuation, special characters, symbols, standardizing notations, retaining unique words, stemming, stop words and selected spelling correction methodologies had to be applied while cleaning the data using NLP techniques.

We explored the ‘pyenchant’ and autocorrect modules in python to correct the spelling mistakes on search term. Both methodologies take abnormally longer time to correct and did not correct the words in the context of their usage. Hence it did not help with improvement in the accuracy of the model.

Finally, we used data from Kaggle open forums to do spelling correction on search term. One of the methodology corrected individual spelling mistakes and another one corrected phrases that were misspelled in the search terms. Using both these methods improved the accuracy of the overall models.

The above steps were done at several stages of the project to do effective data cleaning before we did feature selection for performing the regression analysis.

**Feature Selection**

We used the length of the search\_term along with the word count of the matching words in product\_title, product\_description and mfg\_brand\_name as the essential features while providing the inputs to the regression algorithms in our initial analysis. We also used the count of the product\_uid as an additional feature while encoding the features in all our initial models.

In the next phase, we explored more features like ratio of search term words in title, description and brand names as additional features to the regression models. Also, the brand name was encoded as a numerical feature to the models in this phase.

In the last phase, we used the ‘TF-IDF Vectorizer’ with ‘ngram’ of words in pipeline to select more features using this method and to encode more features for the regression models.

**Train and Test Algorithms on ML Platforms**

We narrowed down our options to Python and Azure ML in the next phase when we wanted to train our models. The features that were selected in the feature selection phase were used to train the models and predict the relevance scores on the test data sets.

The models we initially trained using simple random forest initially in Python to check on the accuracy of the model. The reason we chose random forest regression was the inherent quality of the algorithm to take care of any overfitting and the way it corrects the final predicted variable by taking a mean of all the regression trees it builds in order to correct the misclassification or the overfitting which may be a result of considering the limited features we had at disposal initially. Linear Regression or Decision Tree Regression seemed to be straight forward models which may over fit the data.

In the later phases of the project, when we were able to encode more features, we also considered Support Vector Regression as it fits both linear and non-linear models. XGBoost was also tried at this stage but we did not see a lot of improvement using these 2 algorithms. We also carried out ensemble of the best random forest models and a combination of Random Forest, XGBoost and SVR.

**Parameter Tuning**

After all the data exploration, cleaning, refining of feature selection and training/testing algorithms on multiple platforms/algorithms, we had to work on tuning parameters and selecting the parameters that best fit the model and predicted the relevance score as closely as possible. Grid Search was used at this stage to tune the different parameters like number of trees, tree depth and maximum number of features we selected from TF-IDF Vectorizer.

We improved the accuracy of the predictions by a long way from the point where we started. However, Grid Search methodology to select the optimal parameters is a time consuming process among the steps. We also did 2-10 fold cross validation as part of Grid Search to make the model predict as generically as possible.

# **Results and Conclusions**

The key results from the execution of the project are the prediction of search relevance score using the machine learning model we built using Random Forest Regression. This will help automate the process of prediction of relevance score in Home Depot. Home Depot can update their search algorithms process almost instantaneously instead of going through the existing manual process. This will help them spend more time on running more iterations to better help customers to get very accurate search items they are trying on the Home Depot website.

|  |  |  |  |
| --- | --- | --- | --- |
| **Submission#** | **Model** | **RMSE** | **Weights** |
| 55 | Random Forest Regression-1 | 0.47208 | 0.3 |
| 57 | Random Forest Regression-2 | 0.47238 | 0.2 |
| 58 | Random Forest Regression-3 | 0.47045 | 0.5 |
| 59 | Ensemble Of Random Forest Regression | 0.46920 |  |

**Fig.7 – Results of the ensemble of the best 3 models**

The solution can be deployed by the analytics division at Home Depot to modify the search algorithms. We did our best to make the error from the actual prediction as minimal as possible on the test data, which had both existing terms in train data, and new terms that were additionally used. However, due to performance limitations we could not pick an ideal spell corrector that would automatically correct the spellings on search terms. Any new search terms outside the data provided may be predict the search relevance scores with less accuracy than this model as we don’t have the automatic spell corrector feature.

The future scope of the project can be extended to get an automated spell checker in place which will ensure any new search terms the customers used are accurately predicted so that the search algorithms are updated in a timely manner. The scope can also be extended to include any additional insights to segment the products which will further improve the accuracy of the model. The project can also be extended to explore more algorithms for regression to see if we have better algorithms, which may accurately predict the relevance scores by improving the performance also at the same time.

# **References**

1. <http://scikit-learn.org/stable/modules/pipeline.html>
2. <https://www.kaggle.com/briantc/home-depot-product-search-relevance/homedepot-first-dataexploreation-k>
3. <https://codetidy.com/8197/>
4. <https://www.kaggle.com/buinyi/home-depot-product-search-relevance/disclosing-external-data>
5. <https://www.kaggle.com/the1owl/home-depot-product-search-relevance/rfr-features-0-47326/discussion>

# **Appendix – Reproduction of Results**

**Data Sources**: Download all the below datasets in the same folder as the scripts which will be executed.

Test Dataset: <https://myasucourses.asu.edu/courses/1/2016SpringDYN-T-Laub/groups/_313739_1//_16542570_1/test.csv>

Train Dataset: <https://myasucourses.asu.edu/courses/1/2016SpringDYN-T-Laub/groups/_313739_1//_16542562_1/train.csv>

Descriptions Dataset: <https://myasucourses.asu.edu/courses/1/2016SpringDYN-T-Laub/groups/_313739_1//_16542679_1/product_descriptions.csv>

Attribute Dataset: <https://myasucourses.asu.edu/courses/1/2016SpringDYN-T-Laub/groups/_313739_1//_16542755_1/attributes.csv>

Test Data Best Model: <https://myasucourses.asu.edu/courses/1/2016SpringDYN-T-Laub/groups/_313739_1//_16542898_1/test_new_scext.csv>

Train Data Best Model: <https://myasucourses.asu.edu/courses/1/2016SpringDYN-T-Laub/groups/_313739_1//_16542863_1/train_new_scext.csv>

Train and Test with Spell Check: <https://myasucourses.asu.edu/courses/1/2016SpringDYN-T-Laub/groups/_313739_1//_16543299_1/train_new.zip>

**Tools**:

Python and Azure ML

**Scripts and Execution Instructions**:

1. Basic Data Exploration:

Install Anaconda 3.5 -> Open “cmd” -> type “jupyter notebook” + Enter

<https://myasucourses.asu.edu/courses/1/2016SpringDYN-T-Laub/groups/_313739_1//_16542576_1/2.2%2C%202.3%2C%202.4%20and%202.5%20Data%20Exploration%2C%20Data%20Cleansing%2C%20Relationships%20and%20Feature%20Selection.ipynb>

1. First Model with Random Forest Algorithm

Run Spyder from Anaconda prompt and ensure the datasets specified in the above folder are in the same folder.

<https://myasucourses.asu.edu/courses/1/2016SpringDYN-T-Laub/groups/_313739_1//_16542587_1/RF%20Initial%20Version.py>

1. Mid-term Scripts for Random Forest, XGBoost and Support Vector Regression. Unzip the datasets in “Train and Test Spell Check” zip file. Use Spyder editor in Anaconda to run the below scripts.

RF:

<https://myasucourses.asu.edu/courses/1/2016SpringDYN-T-Laub/groups/_313739_1//_16542596_1/Home%20Depot%20RFR%20Submission_32.py>

XGBoost: This script can be run only on Mac or Linux machine after XGBoost module using “pip install xgboost” in the terminal.

<https://myasucourses.asu.edu/courses/1/2016SpringDYN-T-Laub/groups/_313739_1//_16542599_1/Home%20Depot%20XGB%20Submission_34.py>

SVR:

<https://myasucourses.asu.edu/courses/1/2016SpringDYN-T-Laub/groups/_313739_1//_16542611_1/Home%20Depot%20SVM%20Submission_33.py>

1. Azure ML execution document. The azure ML implementation is shown in the document below.

<https://myasucourses.asu.edu/courses/1/2016SpringDYN-T-Laub/groups/_313739_1//_16542740_1/Azure%20ML%20implementation%20Using%20Feature%20Hashing.docx>

1. Final Version of top-3 Models using Random Forest, and also XGBoost and Support Vector Regression scripts. Use the best train and test datasets for the below scripts. Use Spyder editor in Anaconda to run the below python scripts. The best 3 RF models are chosen for ensembling as per the weights in the results section to get the best accuracy.

**RFR Best Moel-1:**

<https://myasucourses.asu.edu/courses/1/2016SpringDYN-T-Laub/groups/_313739_1//_16542618_1/BestScript_HomeDepot_6.py>

**RFR Best Model-2:**

<https://myasucourses.asu.edu/courses/1/2016SpringDYN-T-Laub/groups/_313739_1//_16542621_1/BestScript_HomeDepot_8.py>

**RFR Best Model-3:**

<https://myasucourses.asu.edu/courses/1/2016SpringDYN-T-Laub/groups/_313739_1//_16542623_1/BestScript_HomeDepot_9.py>

SVR Model:

<https://myasucourses.asu.edu/courses/1/2016SpringDYN-T-Laub/groups/_313739_1//_16542638_1/BestScript_HomeDepot_11.py>

XGBoost Model:

<https://myasucourses.asu.edu/courses/1/2016SpringDYN-T-Laub/groups/_313739_1//_16542629_1/BestScript_HomeDepot_10.py>