

Home Depot – Calculating Search Relevance

Farha Mohsin, Satishraju Rajendran, Mohmad El-Rifai,, Nitish Bahadur

WPI

Case Study 4

Abstract

We study Home Depot Kaggle dataset to calculate the search term result relevance between 1(irrelevant) and 3(perfect relevance) given a search term and product title, along with product attributes and product description. Along with exploratory analysis, regression, using different regression techniques was explored.

Keywords: Regression, Random Forrest, XGBoost, Supervised Learning;

Introduction

The Home Depot¹ (HD) is an American retailer of home improvement and construction products and services. It operates many big-box format stores across the United States. The company is the largest home improvement retailer in the United States. The store operates out of large warehouse-style buildings averaging 105000 ft². The domain homedepot.com attracted at least ²120 million visitors annually.

The key³ to future e-commerce growth for HD is to link the chain retailer's 2,200 stores to its web and mobile properties. Making the best use of existing assets, especially stores, is how Home Depot plans to grow online in 2016. Its total web sales of \$3.76 billion came from its buy online, pickup in store and buy online ship to-store program, the company says. Home Depot online inventory exceeds 1 million products, compared with about 35,000 in a typical Home Depot store. E-commerce continues to account for a bigger slice of total sales. To maintain growth online, the company plans to continue to give web shoppers what they want most: more assortment, quicker delivery and even more convenient ways to shop.

Ubiquitous Internet, mobile computing revolution, and the plethora of how to do it yourself web sites have changed the shopping habits of potential customers that visit homedepot.com. Not only has this put pressure on product marketing but also put additional demand on online presence. A deeper study of the homedepot.com traffic indicates⁴ a decrease in visitors, as illustrated in the picture below.

¹ https://en.wikipedia.org/wiki/The_Home_Depot

² Compete.com survey

³ <https://www.internetretailer.com/2016/01/20/home-depot-hammers-away-online-growth>

⁴ <http://www.alexa.com/siteinfo/homedepot.com>



Further, the prospective customer engagement statistics for home depot is weaker than lowes.com, its closest competitor, as indicated below.

Bounce Rate	homedepot.com	Daily Pageviews per Visitor	Daily Time on Site
28.80% ▼ 1.00%		5.80 ▼ 0.68%	5:56 ▼ 2.00%
Bounce Rate	lowes.com	Daily Pageviews per Visitor	Daily Time on Site
18.90% ▼ 1.00%		5.81 ▲ 2.10%	6:10

Based on this preliminary research we want to evaluate ways on improving customer experience on homedepot.com such that the bounce rate decreases, daily page view increases, daily time spent on site increases and consequently not only customer satisfaction increases but also product sales increases.

Problem 1: Business Use Case

1.1 Business Problem

To determine to what extent a search result⁵ matches the search query that it is paired with. Given a search query, product title, product image, and optional product URL, we need to determine the intent and relevancy of the search result. Relevancy is numeric rating between 1(Irrelevant) and 3(Perfect Match); a rating of 2 is partially or somewhat relevant.

1.2 Why the problem is important to solve?

The Home Depot needs to improve online sales growth to improve its profit margin. A prospective customer starts its engagement by searching about the tools and techniques of how to do it yourself projects. An intelligent online search query engine will not only improve customer experience, but also lead to shorter sales cycle.

1.3 What is your idea to solve the problem?

Build a search result relevancy rating engine that rates the search result based on brand, functionality, and intention. Additional, information regarding product attributes, product description will be used to buttress the search relevance.

1.4 What differences you could make with your data science approach?

Text classification techniques, formalizing the search algorithms using supervised learning techniques, and additional data science discipline such as cross validation and bootstrapping will help improve search result relevance on out of sample data sets.

⁵ <https://www.kaggle.com/c/home-depot-product-search-relevance>

1.5 Why do you believe the idea deserves the investment of the "sharks"?

To continue year over year growth in online sales, the home depot has to solidify its brand superiority by investing in easy to use, customer driven, online tools. Search query relevance is one of the initial prospective customer and home depot engagement points. Making the customer experience engaging, will lead to happier customers and improve home depot sales.

Moreover, online search result relevancy engine will reduce the time a sales associate has to spend with each customer per sale, regardless of the cost of the product. One of the by-products of the better relevancy rating is that same number of sales associate can help more number of customers, reducing the selling and general administrative cost and improving net income.


Problem 2: The Math Part

2.1 Problem formulation in Math

Real world search terms are fraught with incorrect spellings. Additionally, search terms have large number of features (thousands of different kinds of words). Moreover, because of English language nuances where each word can mean multiple things it is hard to capture the intention. In addition, search result relevance score should be explainable, as opposed to a black box approach. Given these criteria, we propose to use a decision tree based algorithm because decision trees⁶ are invariant under scaling and various other transformations of feature values, is robust to inclusion of irrelevant features, and produces inspectable models. Unfortunately, decision trees are seldom accurate.

To avoid overfitting their training sets, because decision trees have low bias, but very high variance we propose to use Random forests are a way of averaging multiple deep decision trees, trained on different parts of the same training set, with the goal of reducing the variance. Although Random forest comes at the expense of a small increase in the bias and some loss of interpretability, Random forests generally greatly boosts the performance of the final model.

The problem contains 3 data sets: training, product description, and product attributes. A sample is illustrated below:



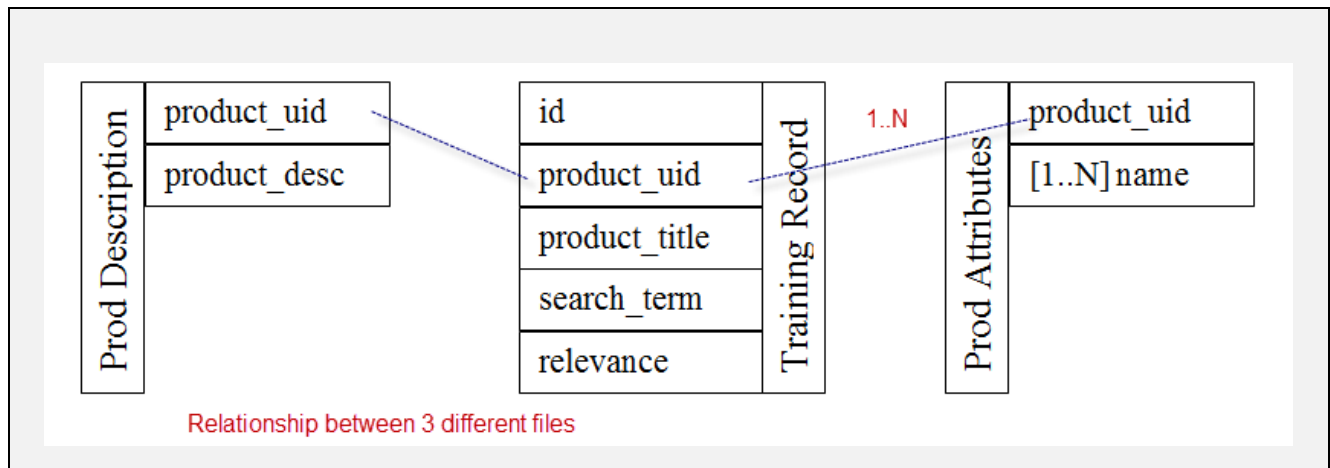
id	product_uid	product_title	search_term	relevance
2	100001	Simpson Strong-Tie 12-Gauge Angle	angle bracket	3
3	100001	Simpson Strong-Tie 12-Gauge Angle	l bracket	2.5
9	100002	BEHR Premium Textured DeckOver 1-gal. #SC-104 Tugboat Wood and Concrete Coating Deck Over		3
16	100005	Delta Vero 1-Handle Shower Only Faucet Trim Kit in Chrome (Valve Not Included)	rain shower head	2.33
17	100005	Delta Vero 1-Handle Shower Only Faucet Trim Kit in Chrome (Valve Not Included)	shower only faucet	2.67
18	100006	Whirlpool 1.9 cu. ft. Over the Range Convection Microwave in Stainless Steel with Sens convection otr		3
20	100006	Whirlpool 1.9 cu. ft. Over the Range Convection Microwave in Stainless Steel with Sens microwave over stove		2.67
21	100006	Whirlpool 1.9 cu. ft. Over the Range Convection Microwave in Stainless Steel with Sens microwaves		3
23	100007	Lithonia Lighting Quantum 2-Light Black LED Emergency Fixture Unit	emergency light	2.67
27	100009	House of Fara 3/4 in. x 3 in. x 8 ft. MDF Fluted Casing	mdf 3/4	3

⁶ https://en.wikipedia.org/wiki/Random_forest

product_uid	product_description
100001	Not only do angles make joints stronger, they also provide more consistent, straight corners. Simpson Strong-Tie offers a wide variety of angles in various sizes and thicknesses to handle light-duty jobs or projects where a structural connection is needed. Some can be bent (skewed) to match the project. For outdoor projects or those where moisture is present, use our ZMAX zinc-coated connectors, which provide extra resistance against corrosion (look for a "Z" at the end of the model number).Versatile connector for various 90 connections and home repair projectsStronger than angled nailing or screw fastening aloneHelp ensure joints are consistently straight and strongDimensions: 3 in. x 3 in. x 1-1/2 in.Made from 12-Gauge steelGalvanized for extra corrosion resistanceInstall with 10d common nails or #9 x 1-1/2 in. Strong-Drive SD screws
100002	BEHR Premium Textured DECKOVER is an innovative solid color coating. It will bring your old, weathered wood or concrete back to life. The advanced 100% acrylic resin formula creates a durable coating for your tired and worn out deck, rejuvenating to a whole new look. For the best results, be sure to properly prepare the surface using other appropriate BEHR products before applying DECKOVER. Coverage: 100 sq. ft. per gallon. Proposition 65 information: Revives wood and composite decks, railings, porches and boat docks, also great for concrete pool decks, patios and sidewalks100% acrylic solid color coatingResists cracking and peeling and conceals splinters and cracks up to 1/4 in.Provides a durable, mildew resistant finishCovers up to 75 sq. ft. in 2 coats per gallonCreates a textured, slip-resistant finishFor best results, prepare with the appropriate BEHR product for your wood or concrete surfaceActual paint colors may vary from on-screen and printer representationsColors available to be tinted in most storesOnline Price includes Paint Care fee in the following states: CA, CO, CT, ME, MN, OR, RI, VT
100003	Classic architecture meets contemporary design in the Ensemble Curve series, made of solid Vikrell material, blending sleek, clean lines with gentle curves. Corner shelving is perfect for storing bath accessories. Modular design allows it to be moved around corners and through doorways with ease. Curve wall with a smooth, contemporary finishFeaturing integrated storage shelves, subtly narrower for tighter spacesDesigned for use in 18 in.

product_uid	name	value
100001	Bullet01	Versatile connector for various 90° connections and home repair projects
100001	Bullet02	Stronger than angled nailing or screw fastening alone
100001	Bullet03	Help ensure joints are consistently straight and strong
100001	Bullet04	Dimensions: 3 in. x 3 in. x 1-1/2 in.
100001	Bullet05	Made from 12-Gauge steel
100001	Bullet06	Galvanized for extra corrosion resistance
100001	Bullet07	Install with 10d common nails or #9 x 1-1/2 in. Strong-Drive SD screws
100001	Gauge	
100001	Material	Galvanized Steel
100001	MFG Brand Name	Simpson Strong-Tie
100001	Number of Pieces	
100001	Product Depth (in.)	
100001	Product Height (in.)	
100001	Product Weight (lb.)	
100001	Product Width (in.)	
100002	Application Method	Brush,Roller,Spray
100002	Assembled Depth (in.)	6.63 in
100002	Assembled Height (in.)	7.76 in
100002	Assembled Width (in.)	6.63 in
		Revives wood and composite decks, railings, porches and boat docks, also great for

The relational model between the 3 data files is indicated below:



2.2 Math Solution

Using the following algorithm⁷, we will use Random Forest Regression:

- 1) For $b = 1$ to B , the number of trees to average over
 - a) Draw a bootstrap sample \mathbf{Z}^* of size N from the training data.
 - b) Grow a random-forest tree T_b to the bootstrapped data, by re-cursively repeating the following steps for each terminal node of the tree, until the minimum node size n_{min} is reached.
 - i) Select m variables at random from the p variables.
 - ii) Pick the best variable/split-point among the m .
 - iii) Split the node into two daughter nodes.
- 2) Output the ensemble of trees $\{T_b\}^B$
- 3) To make a new regression prediction at point x :

$$\hat{f}_{rfr}^B(x) = \frac{1}{B} \sum_{b=1}^B T_b(x)$$

2.3 Implementation of the Solution

To implement our solution we use `sklearn.ensemble RandomForestRegressor`, `BaggingRegressor` and `GradientBoostClassifier`

- After reading the training, testing, product description file, attribute files we leveraged data frames to extract features; lambda functional program allowed us to do compact in-line transformations and feature extractions.
- Mean squared error and mean scorer was used as metrics. These metrics were imported from `sklearn.metrics` package.
- Pipeline was built using the regressor and `GridSearch` method was used to search the best model parameters such as `n_estimators` and `depth`.

⁷ Element of Statistical Learning, Chapter 15

Problem 3: The Hacking Part

3.1 Data Collection

The HD datasets are provided by HD as part of the HD Kaggle competition.

- ***Train.csv***: contains four attributes *product_uid*, *product_title*, *search_term* and *relevance* with 74067 records.
- ***Test.csv***: contains four attributes *product_uid*, *product_title*, *search_term* with 166694 records.
- ***Attributes.csv***: contains three attributes *product_uid*, *name*, *value* with 1048486 records.
- ***Product_description.csv***: contains two attributes *product_uid*, *product_description* with 124429 records.

3.2 Implementation

3.2.1 DATA ISSUES:

Implementing a model that will be able to predict the relevance of a search term to a product, by brand, functionality and intention is not an easy task, but with the given HD datasets the task gets harder. Typographical errors, missing values, inconsistent data, outliers and issues with data formats are a subset of the issues we faced while working with HD datasets.

- **Examples of Typos:**

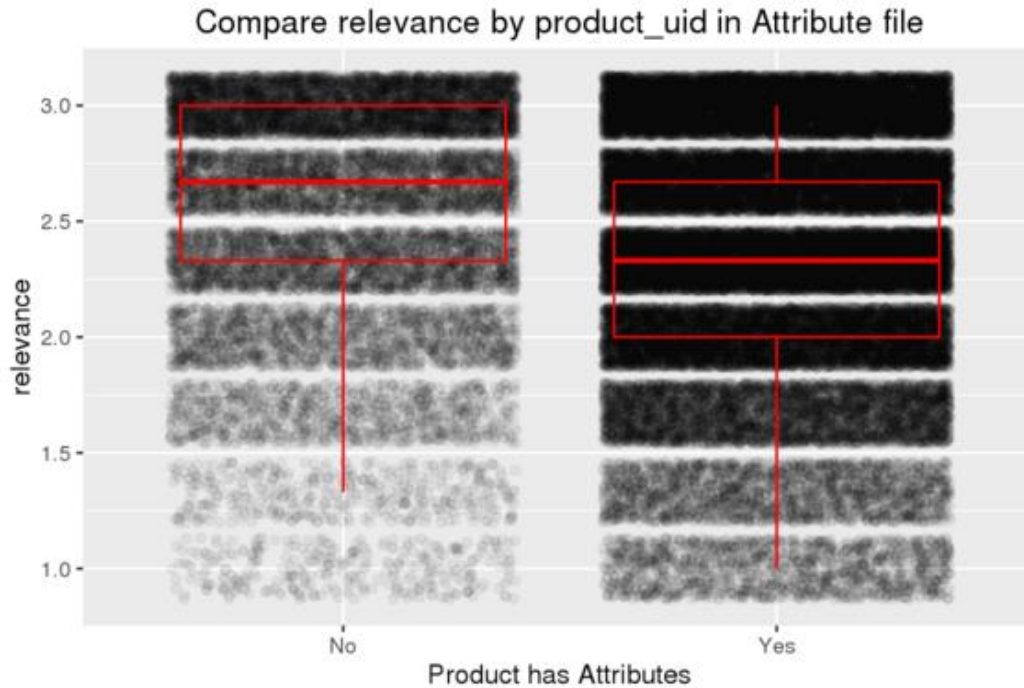
Across all the datasets and especially in product description and attributes datasets we found many instances where the words concatenated together . For example, sidewalks100%, surfaceActual, environmentsCertified

- **Examples of Misspelling:**

Mowe, sprkinler, keorsene, lanterun

- **Example of inconsistent data:**

In the attributes.csv file there are 5343 different categories and majority of the items don't share the same set of attributes



List of most frequent Attributes between the items

Attribute.name	Freq.
mfg brand name	86,250
bullet02	86,248
bullet03	86,226
bullet04	86,174
bullet01	85,940
product width (in.)	61,137
bullet05	60,529
product height (in.)	54,698
product depth (in.)	53,652
product weight (lb.)	45,175

3.2.2 DATA PROCESSING:

Given the data issues, cleaning the data was a pre-requisite to build a robust model.

Steps	
1.	<p>Fixing search terms incorrect spellings:</p> <p>We used a dictionary to fix misspells in home improvement, gardening, construction and do-it-yourself context. While a regular English dictionary failed to fix the misspelled terms in the HD context, the custom dictionary built by using the set of words from product descriptions fixed majority of misspelled words.</p> <p>For example, the misspelled word “mowe” will be fixed by regular dictionary as “more” but for HD context it should be “mower”.</p>
2.	<p>Cleaning Data:</p> <p>Redundant words made the computation of relevance rating more complex and slower. To improve processing times, Python stop words list was used as a preliminary step. Words that were not contributing to the model such as alpha numeric terms were removed to improve model efficiency.</p> <p>Additionally, regular expression was used to fine grained tokenizing. The fine grained tokenizer helped us segregate composite words that were separated by hyphen, which a regular tokenizer failed to parse correctly.</p>
3.	<p>Stemming the words:</p> <p>To avoid mismatching between the search term and the words in the title, description and attributes we used Python stem functionality to unify the words, so that we avoid mismatches,</p> <p>For example, without using stem function, a comparison between write and</p>

	writing will result with a mismatch, but with using stem function it is going to convert writing to write (original word) and stem write to write because it is the original word, so that the comparison will result in a positive match.
--	--

3.2.3 MODELS AND FEATURES

After processing and cleaning the datasets, we merged train, product_description and attributes dataset on the product_id and we start looking for features that can help us build a good prediction model for relevancy. We tried several feature engineering approaches. We report the word count and word proportion approaches below:

WORD COUNT FEATURES APPROACH:

Given the search term, we find the number of occurrences of its words in the product title, description and attributes. The result was a three new attributes that we used in order to predict the relevancy of the search term to a given item.

product_info	word_in_title	word_in_description	brand_name	attr	word_in_brand	word_in_attr
angl bracket\tsimpson strong-ti 12-gaug anglt...	1	1	simpson strong-ti	12 galvan steel 1 1.5 3 0.26 3	0	0
I bracket\tsimpson strong-ti 12-gaug angltnot...	1	1	simpson strong-ti	12 galvan steel 1 1.5 3 0.26 3	0	1
deck over\tbehr premium textur deckov 1-gal. #...	1	1	behr premium textur deckov	brush,roller,spray 6.63 in 7.76 in 6.63 in soa...	1	0

WORD PROPORTION APPROACH:

The second approach we tried is to find the proportion of the search term across product title, product description and attributes. In this approach we calculate the percentage of the search term words that exists in each of the attributes mentioned.

	id	relevance	prop_in_title	prop_in_description	prop_in_attr	len_of_query
0	2	3	0.50	0.50	0.50	2
1	3	2.5	0.50	0.50	0.50	2
2	9	3	0.50	0.50	1.00	2
3	16	2.33	0.33	0.33	1.00	3
4	17	2.67	1.00	0.67	0.67	3

RANDOM FOREST MODEL:

Using the sci-kit pipeline processing with cross validation, we tried random forest tree model to predict the relevance result; RMSE was used as a metric. The model was run twice: one with no spelling correction and then with spelling corrections. Both experiment results are presented below:

Model	Pre-processing	Training Obs	Test Obs.	CV	RMSE
RFR	No spelling correction	55550	18516	Yes	0.49
RFR	Spelling and typo corrections using product description dictionary	55550	18516	Yes	0.47
LR	Spelling and typo corrections using product description dictionary	55550	18516	No	0.49
BR	Spelling and typo corrections using product description	55550	18516	No	0.48

	dictionary				
--	------------	--	--	--	--

The grid search algorithm reported the following best parameters:

- rfr_n_estimators: 25
- rfr_max_depth: 6

A max depth of 6 was satisfactory because decision trees need to be readable too.

Large depth makes the decision trees less readable.

LINEAR REGRESSION MODEL:

Model	Pre-processing	Training Obs	Test Obs.	CV	RMS E
LR	Spelling and typo corrections using product description dictionary	55550	18516	No	0.49

BAGGING REGRESSION MODEL:

Model	Pre-processing	Training Obs	Test Obs.	CV	RMS E
BR	Spelling and typo corrections using product description dictionary	55550	18516	No	0.48

Conclusion

The key to future e-commerce growth for HD is to link the chain retailer's 2,200 stores to its web and mobile properties. Making the best use of existing assets, especially stores, is how Home Depot plans to grow online in 2016. To help fuel this growth and improve shareholders return HD needs to reduce its bounce rate, improve both daily page view per visitor and daily time on site. Improving these metrics will potentially lead to higher web sales, a rapid growing segment for HD.

Relevance of search results plays a key part in improving customer experience and bolstering customer satisfaction. To improve the search relevance better quality search results are required.

In this case study we find that improving search results is hard on real world data sets because of incorrect spellings and determining the intention of the search term. Preliminary experiments are extremely promising and the use of data science techniques such as random forest regression, with and without spelling correction, pre-processing with custom tokenizers help build better search algorithms.

However, given more resources and time, we feel our RMSE of 0.47 can be further improved by exploring other algorithms such as Xtreme Random Boost and better feature engineering techniques, such as limiting the feature scores are calculated by maximizing the cumulative sum of brand, functionality, and intention, such that each of these are equally weighted.