Home Depot Competition

Zhaoxuan Ren Yumeng Lu Dan Sun Tao Li Yue Su

Motivation & Goal

■ Background

The Home Depot (referred to in some countries and often in colloquial speech simply as Home Depot) is an American retailer of home improvement and construction products and services. It operates many big-box format stores across the United States (including all 50 U.S. states, the District of Columbia, Puerto Rico, the United States Virgin Islands and Guam), all ten provinces of Canada, as well as Mexico. The company is headquartered at the Atlanta Store Support Center in Cobb County, Georgia, in Greater Atlanta.

■ Motivation

Search relevancy is an implicit measure Home Depot uses to gauge how quickly they can get customers to the right products. Currently, human raters evaluate the impact of potential changes to their search algorithms, which is a slow and subjective process. By removing or minimizing human input in search relevance evaluation, Home Depot hopes to increase the number of iterations their team can perform on the current search algorithms.

■ Goals

In this competition, Home Depot is asking Kagglers to help them improve their customers' shopping experience by developing a model that can accurately predict the relevance of search results. Thus the purpose is to build an appropriate model using certain methods and tools to predict the relevance of search results and improve the shopping experience of customers.

Methods

■ Dataset Description

- train.csv - the training set, contains products, searches, and relevance scores
- test.csv - the test set, contains products and searches. You must predict the relevance for these pairs.
- product_descriptions.csv - contains a text description of each product.
- attributes.csv - provides extended information about a subset of the products (typically representing detailed technical specifications). Not every product will have attributes

- product_uid - an id for the products
- product_title - the product title
- product_description - the text description of the product (may contain HTML content) search_term - the search query

- material - a kind of attributes
- brand - a kind of attributes
- relevance - the average of the relevance ratings for a given id

■ Dataset Processing

- Merge Dataset

```
test <- merge(test,productDescriptions, by.x = "product_uid", by.y = "product_uid", all.x = TRUE, all.y = FALSE)
test <- merge(test,MFGBrandName, by.x = "product_uid", by.y = "product_uid", all.x = TRUE, all.y = FALSE)
test <- merge(test,material, by.x = "product_uid", by.y = "product_uid", all.x = TRUE, all.y = FALSE)
```

- Clean up Data

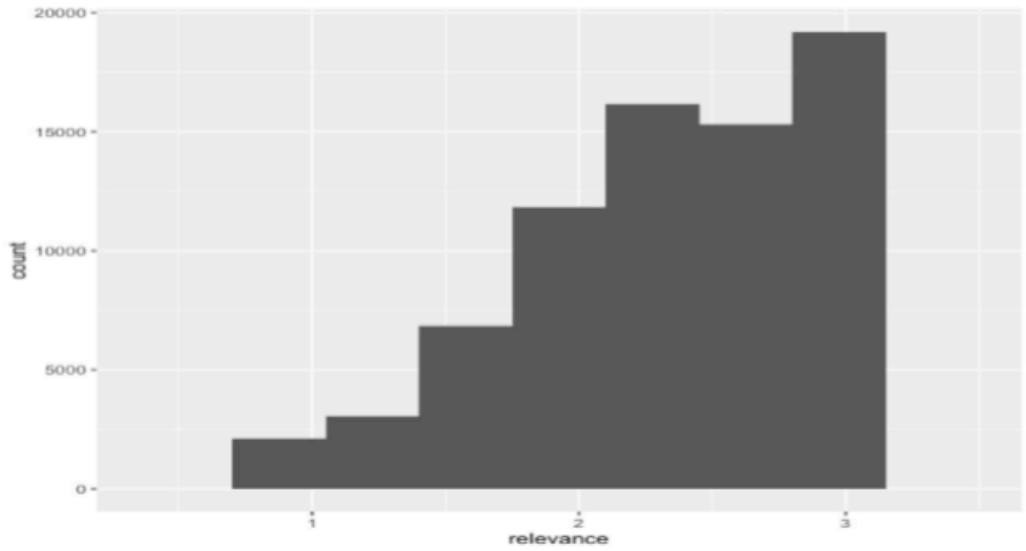
```
test$name.x = NULL
test$name.y = NULL
```

- Split Search Terms

```
#split search term by space
#   words <- unlist(strsplit(as.character(words)," "))
#   nwords <- length(words)
#   for(i in 1:length(words)){
#     pattern <- paste("\\w*\\s",words[i],"\\s\\w*",sep="")
#     n_title <- n_title + grepl(pattern,title,perl=TRUE,ignore.case=TRUE)
#     n_desc <- n_desc + grepl(pattern,desc,perl=TRUE,ignore.case=TRUE)
#     n_brand <- n_brand + grepl(pattern,brand,perl=TRUE,ignore.case=TRUE)
#                                 n_material      <-      n_brand      +
grepl(pattern,material,perl=TRUE,ignore.case=TRUE)
#   }
```

■ Relevance Distribution

The graph shown below indicates the relevance distribution of the dataset. We can find out that according to the increase of relevance, the count number continually increase.



- Dataset after Data Processing

```
> train = read.csv('./finalTrain(wordRoot).csv')
> train[1,]
  X.1 nwords n_title n_desc n_brand n_material X product_uid id                    product_title   search_term
1   1      2       1      1       0          0 1      100001  2 Simpson Strong-Tie 12-Gauge Angle angle bracket
  relevance
1         3

product_description
1 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 wher
e a structural connection is needed. Some can be bent (skewed) to match the project. For outdoor projects or th
ose where moisture is present, use our ZMAX zinc-coated connectors, which provide extra resistance against corr
osion (look for a "Z" at the end of the model number).Versatile connector for various 90 connections and home r
epair 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 resistanc
eInstall with 10d common nails or #9 x 1-1/2 in. Strong-Drive SD screws
                Brand         Material
1 Simpson Strong-Tie Galvanized Steel
```

- Method

Because of its relevance continuity,we decide to use regression analysis to deal with this problem.

- Regression Analysis

Regression analysis is widely used for prediction and forecasting, where its use has substantial overlap with the field of machine learning. Regression analysis is also used to understand which among the independent variables are related to the dependent variable, and to explore the forms of these relationships. In restricted circumstances, regression analysis can be used to infer causal relationships between the independent and dependent variables. However this can lead to illusions or false relationships, so caution is advisable; for example, correlation does not imply causation.

Many techniques for carrying out regression analysis have been developed. Familiar methods such as linear regression and ordinary least squares regression are parametric, in that the regression function is defined in terms of a finite number of unknown parameters that are estimated from the data. Nonparametric regression refers to techniques that allow the regression

function to lie in a specified set of functions, which may be infinite-dimensional.

- Conclusion

```
> weights <- chi.squared(relevance~., dataset)
> weights
             attr_importance
nwords            0.08322153
n_title           0.14742180
n_desc            0.10800737
n_brand           0.03715768
n_material        0.03926718
id                0.20753466
search_term       0.49586038
```

- Analysis
  - Model 1: Generalized Boosted Regression Models

RMSE: 0.49

  - Model 2: Random Forest

RMSE: 0.50