Home Depot Product Search Relevance

Mitali Bharali

OBJECTIVE

To predict
a relevance score for the provided combinations of search terms and products

Relevance is a number between 1 (not relevant) to 3 (highly relevant).

For example, a search for "AA battery" would be considered highly relevant to a pack of size AA batteries (relevance = 3), mildly relevant to a cordless drill battery (relevance = 2), and not relevant to a snow shovel (relevance = 1).

About this project

Dataset consisted of:

- train.csv (74.1k x 5) - the training set, contains products, searches, and relevance scores
- test.csv (167k x 4) - the test set, contains products and searches. We are predicting the relevance score for these pairs
- product_descriptions.csv (124k x 2) - contains a text description of each product. We merged this table to the test set via the product_uid
- attributes.csv (2.04m x 3) - provides extended information about a subset of the products (typically representing detailed technical specifications)
- sample_submission.csv (167k x 2) - a file showing the correct submission format
- relevance_instructions.docx - the instructions provided to human raters

About the dataset

Training set

53489 products

11795 search query

54667 product_uid

Testing set

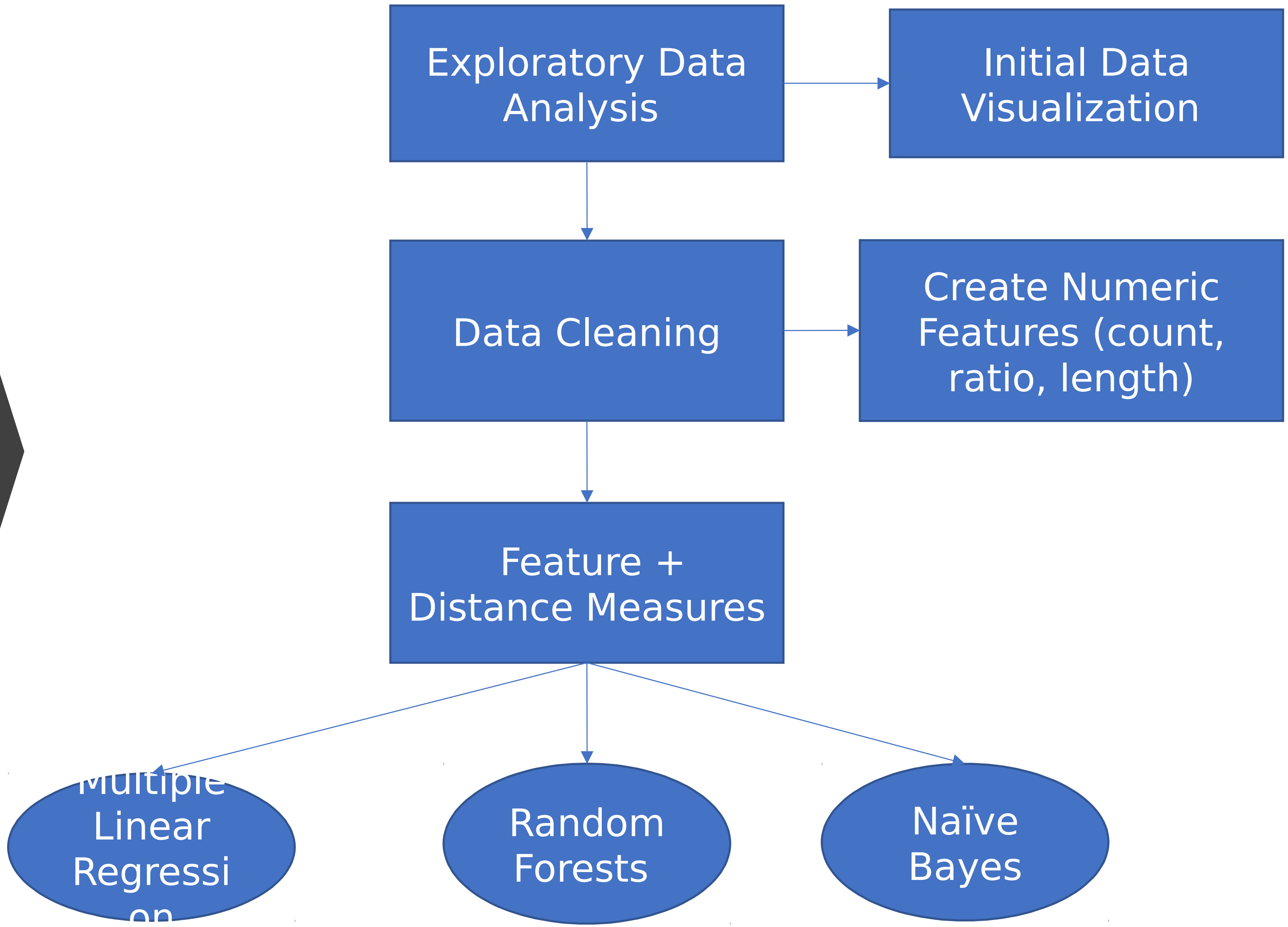
94731 products

22427 search query

97460 product_uid

STRATEGY/WORK FLOW

Work Flow


Exploratory Data Analysis
Initial Data Visualization
Data Cleaning
Create Numeric Features (count, ratio, length)
Feature + Distance Measures
Multiple Linear Regressi on
Random Forests
Naïve Bayes

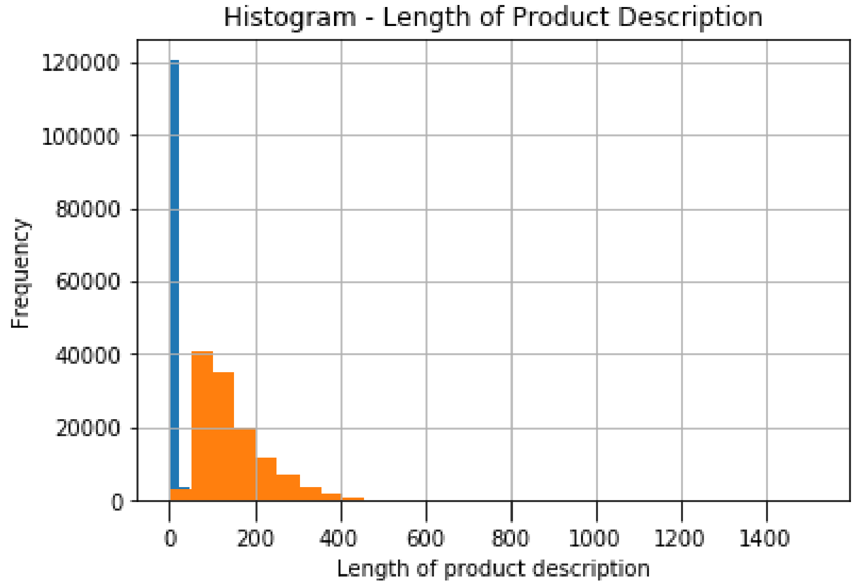
Exploratory Data Analysis

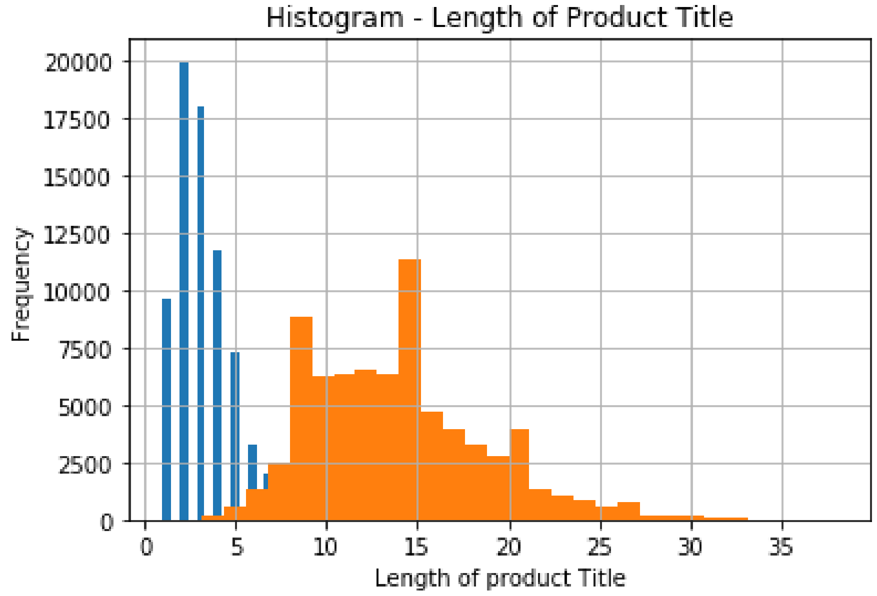
- Histogram 1: Length of Product Description

- Blue bars represent the frequency of digits
- Orange bars represent the frequency of alphabets

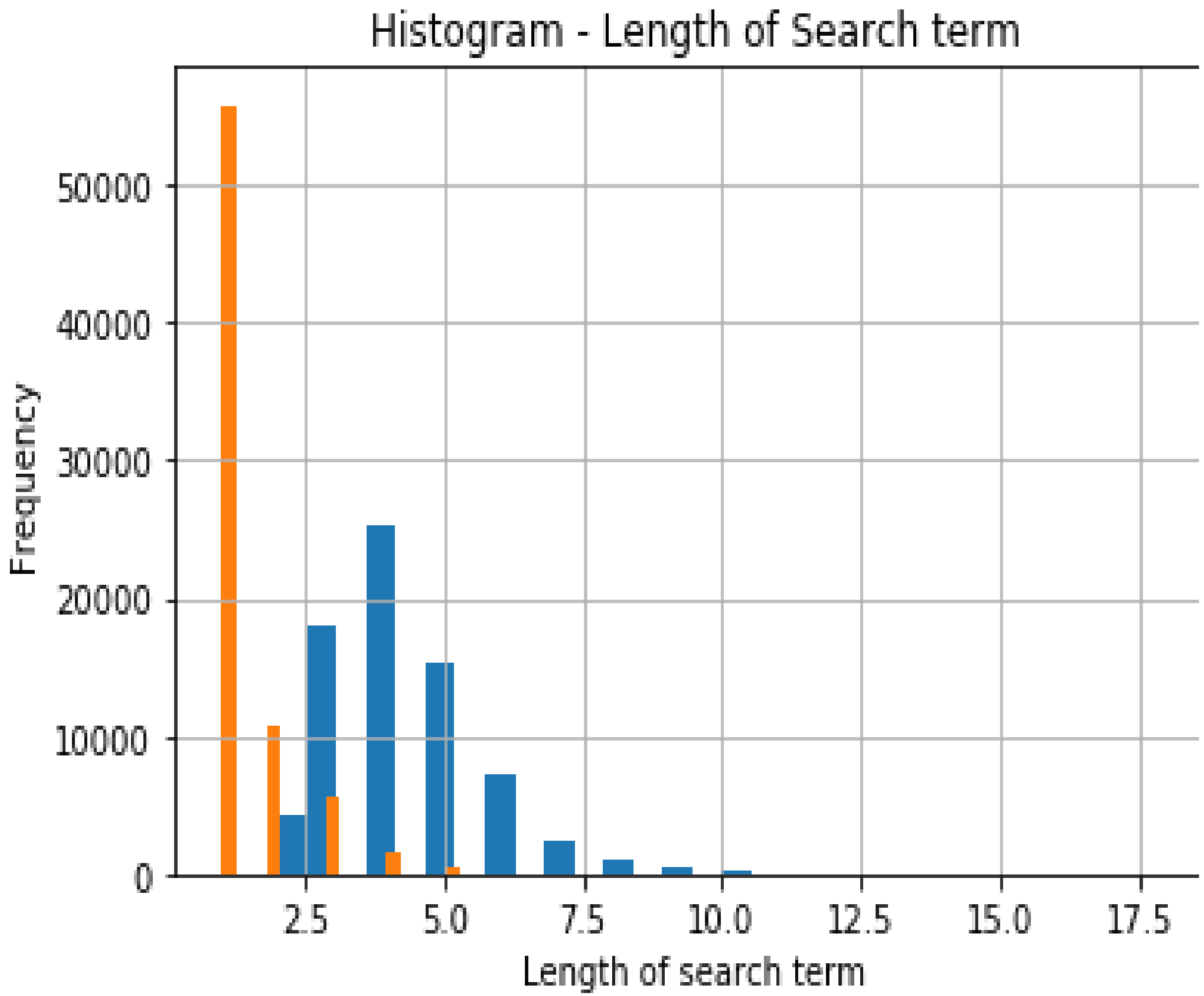
- Histogram 2: Length of Product Title

- Blue bars represent the frequency of digits
- Orange bars represent the frequency of alphabets

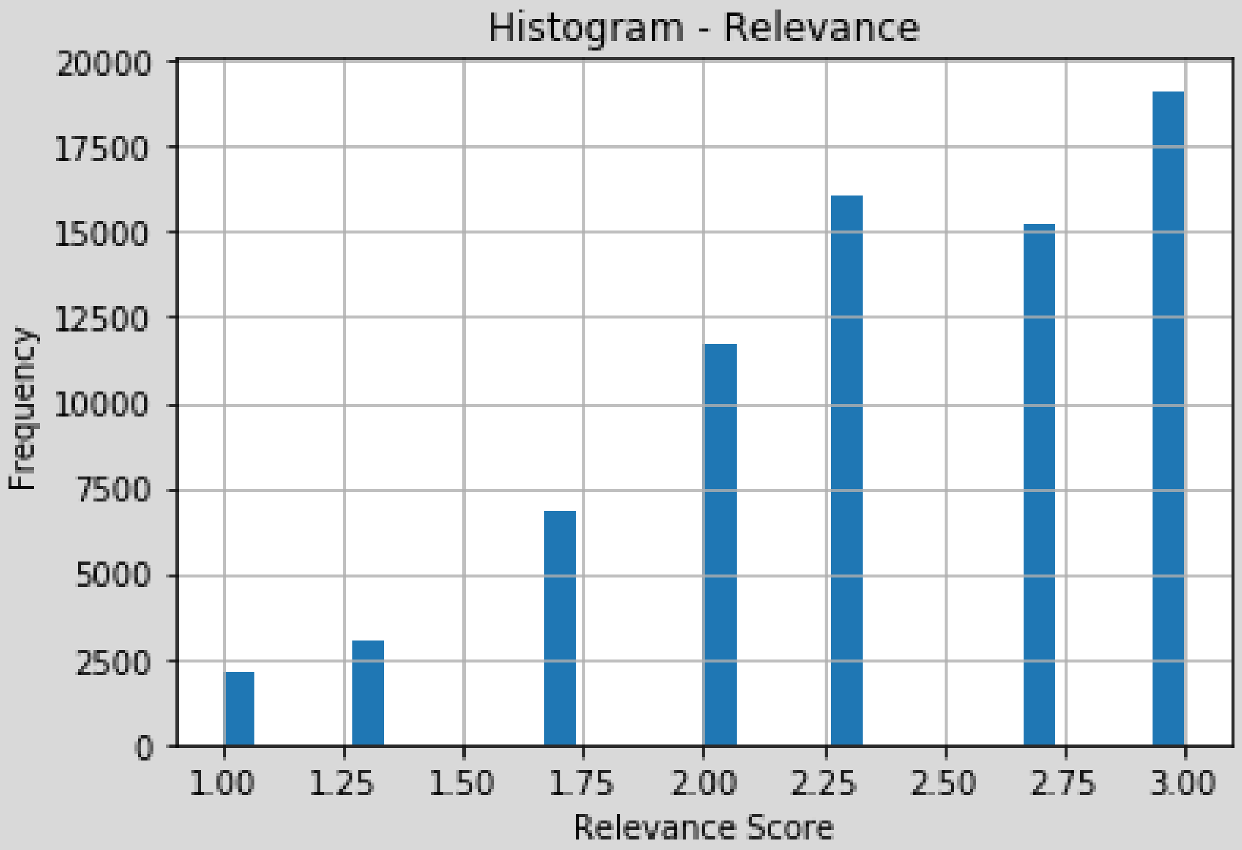




Exploratory Data Analysis



Exploratory Data Analysis




```
Out[16]: 3.00    19125
         2.33    16060
         2.67    15202
         2.00    11730
         1.67     6780
         1.33     3006
         1.00     2105
         2.50       19
         2.25       11
         2.75       11
         1.75        9
         1.50        5
         1.25        4
         Name: relevance, dtype: int64
```

TEXT CLEANING

Basics

Fix Casing :	Hammer > hammer
Remove Symbols:	ft. > ft
Remove Stop Words:	hammer for nails > hammer nails
POS Tagging:	hammer > [hammer,noun]
Lemmatization:	drills > drill
Stemming:	running > run

Advanced

Standardize Numbers :	Five > 5
Standardize Measurements:	2 feet by 4 inches > 2x4
Split Joined Words:	wiremesh > wire mesh
Correct Spelling:	insullation > insulation

Feature Engineering

1) *Create num columns based on text columns*
   - Count number of words from search query which appears both in product_title and product_description
   - Compute ***Edit Distance*** from search query which appears both in product_title and product_title
   - Compute the ***Cosine Similarity*** between search query, product_title and product_description
   - Compute the ***Jaccard Similarity*** between search query, product_title and product_description
   - Count number of words in the product description
   - Create new columns for each pair
2) *Remove all text columns*

As a result we will have vectors of numbers that suites well for the machine learning.

Distance Measures

EDIT DISTANCE:

The distance between the source string and the target string is the minimum number of edit operations (deletions, insertions, or substitutions) required to transform the source into the target.

COSINE DISTANCE:

Cosine similarity calculates similarity by measuring the cosine of angle between two vectors. With cosine similarity, we need to convert sentences into vectors.

JACCARD DISTANCE:

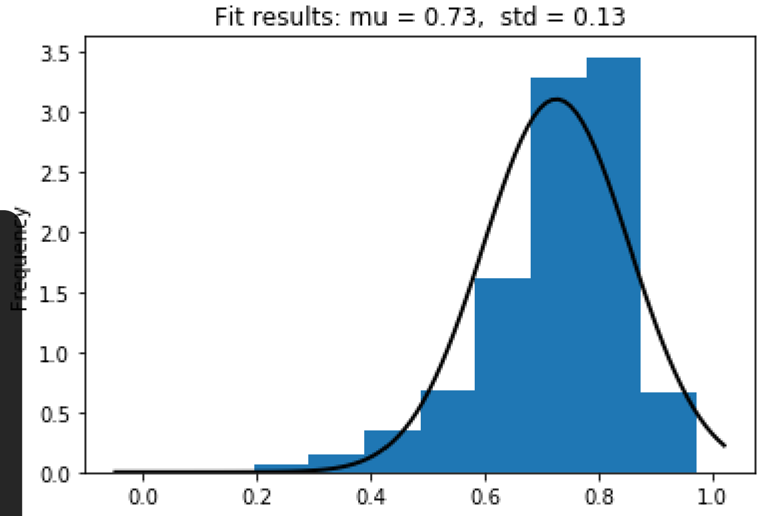
Jaccard Distance is a measure of how dissimilar two sets are. Lower the distance, more similar are the two strings.

Feature Creation

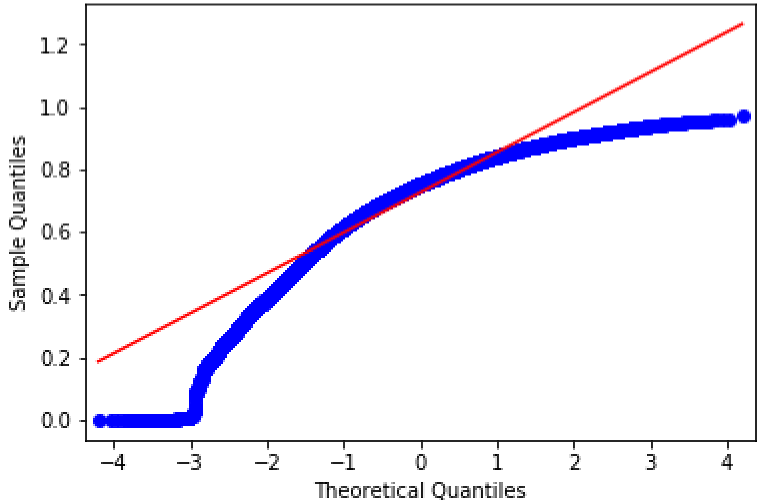
product_uid	product_title	search_term	relevance	product_description	search_term_tokens	product_title_tokens	product_description_toke
100001	simpson strongtie angle	angl bracket	3.0	angles make joints stronger also provide consi...	[angle, bracket]	[simpson, strong-tie, 12-gauge, angle]	[not, only, do, angles, make, joints, stronger...
100001	simpson strongtie angle	l bracket	2.5	angles make joints stronger also provide consi...	[l, bracket]	[simpson, strong-tie, 12-gauge, angle]	[not, only, do, angles, make, joints, stronger...
100002	behr premium textured deckover tugboat wood co...	deck over	3.0	behr premium textured deckover innovative soli...	[deck, over]	[behr, premium, textured, deckover, 1-gal., #s...	[behr, premium, textured deckover, is, an, in...

shared_words_mut	shared_words	edistance_sprot	edistance_sd	j_dis_sqt	j_dis_sqd	search_query_length	number_of_words_in_descr
4	24	20	589	0.2	0.0	12	71
3	24	20	592	0.0	0.0	9	71
21	62	53	850	0.0	0.0	9	111

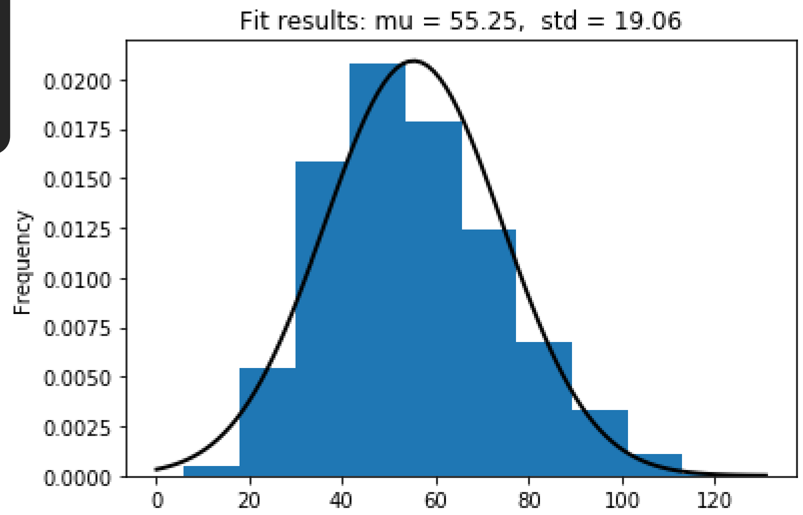
Features Analysis



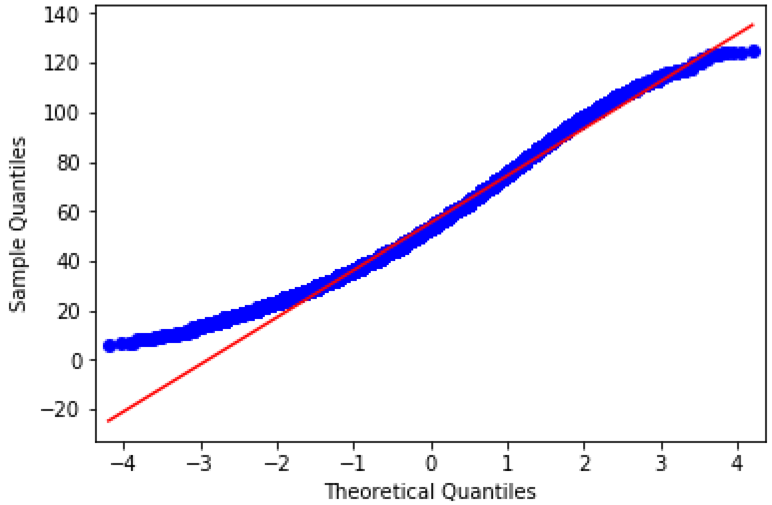
Histogram of Cosine Distance



qqplot of Cosine distance

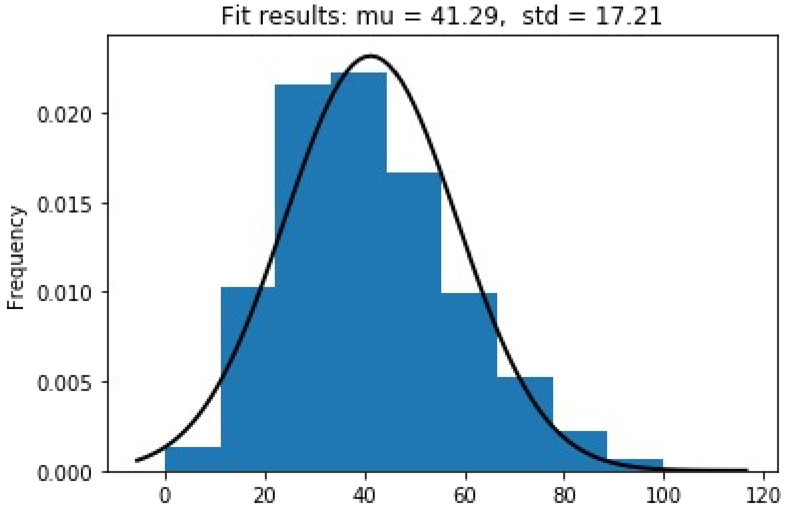


Histogram of Shared Words

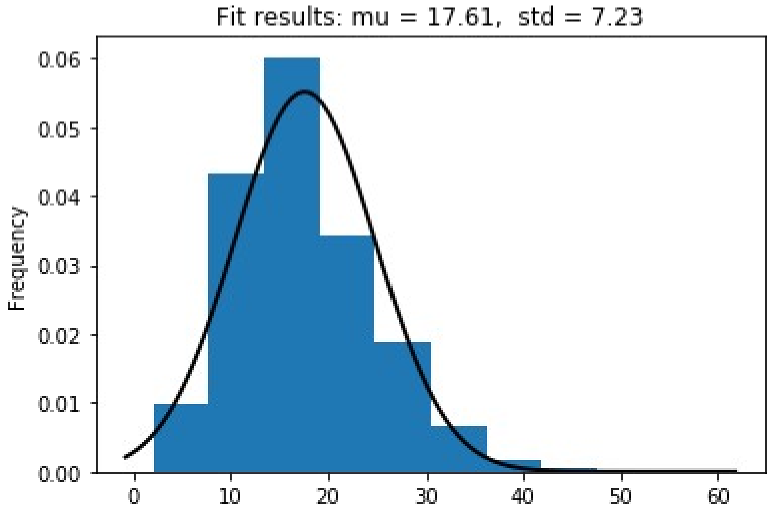


qqplot of Shared Words

Training Features Analysis

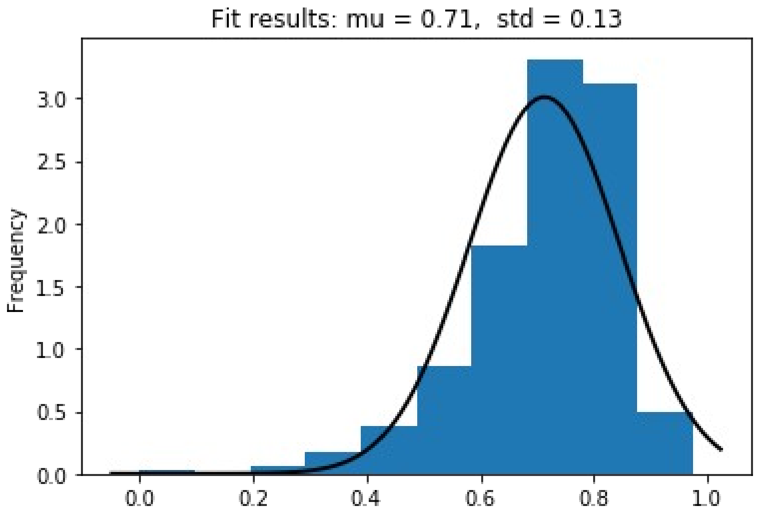


Histogram of Edit Distance (Search term, Product Title)

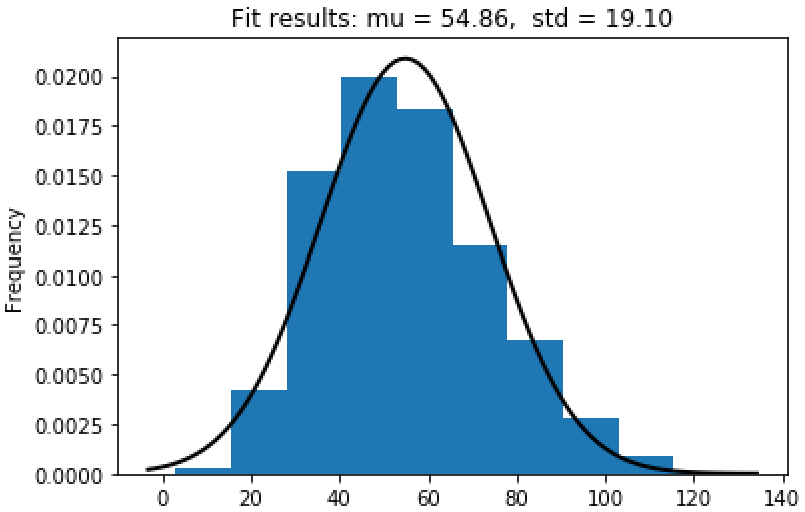


Histogram of Search Query Length

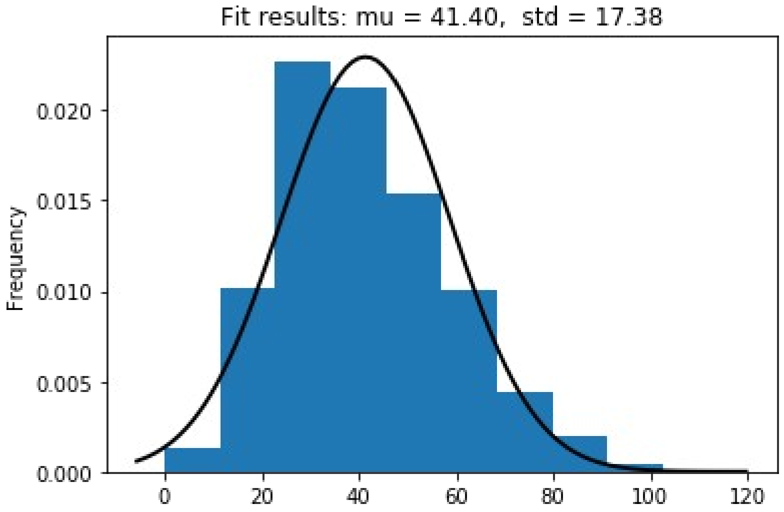
Testing Features Analysis



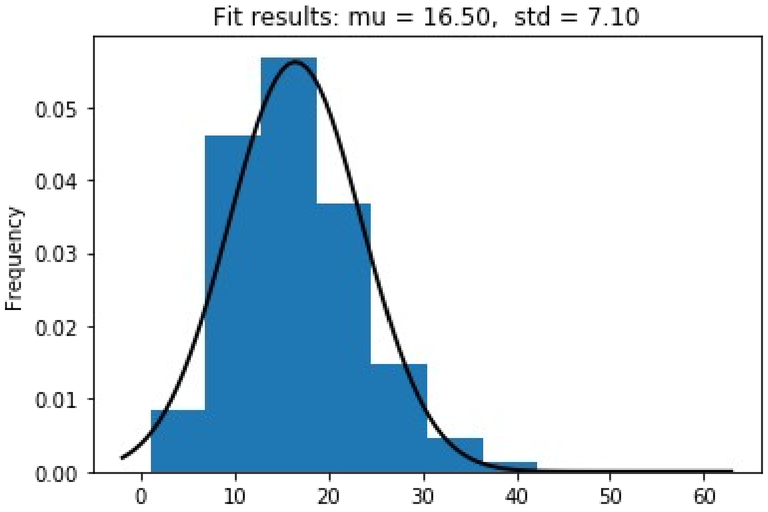
Histogram of Cosine distance



Histogram of Shared words

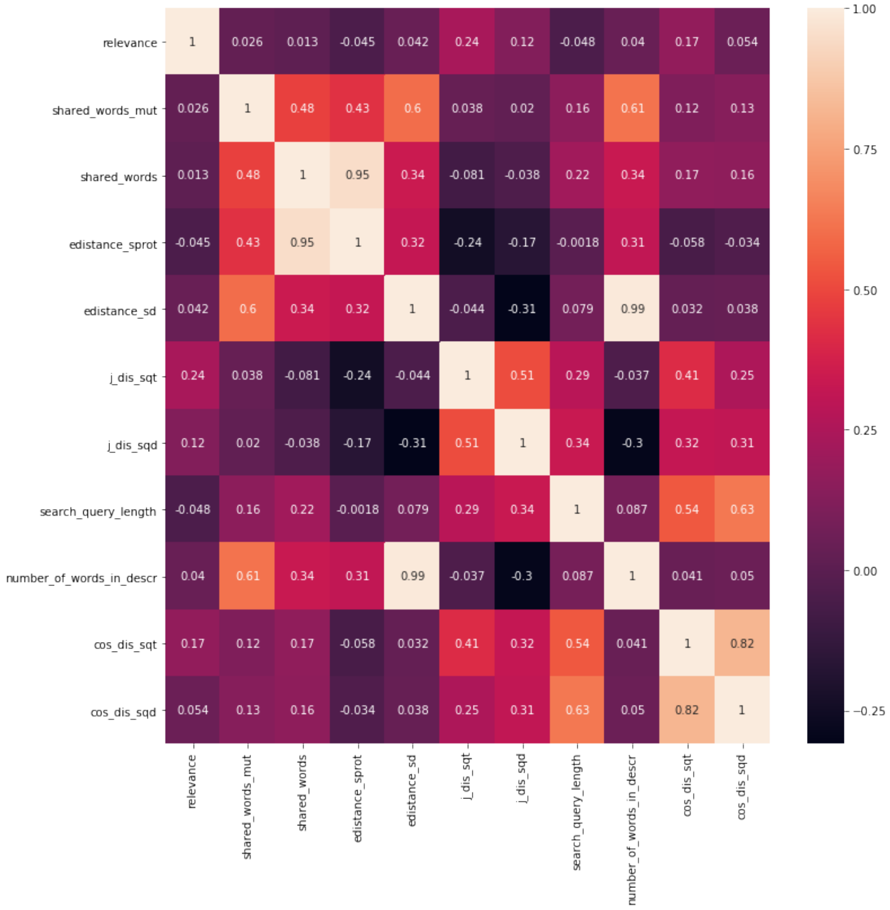


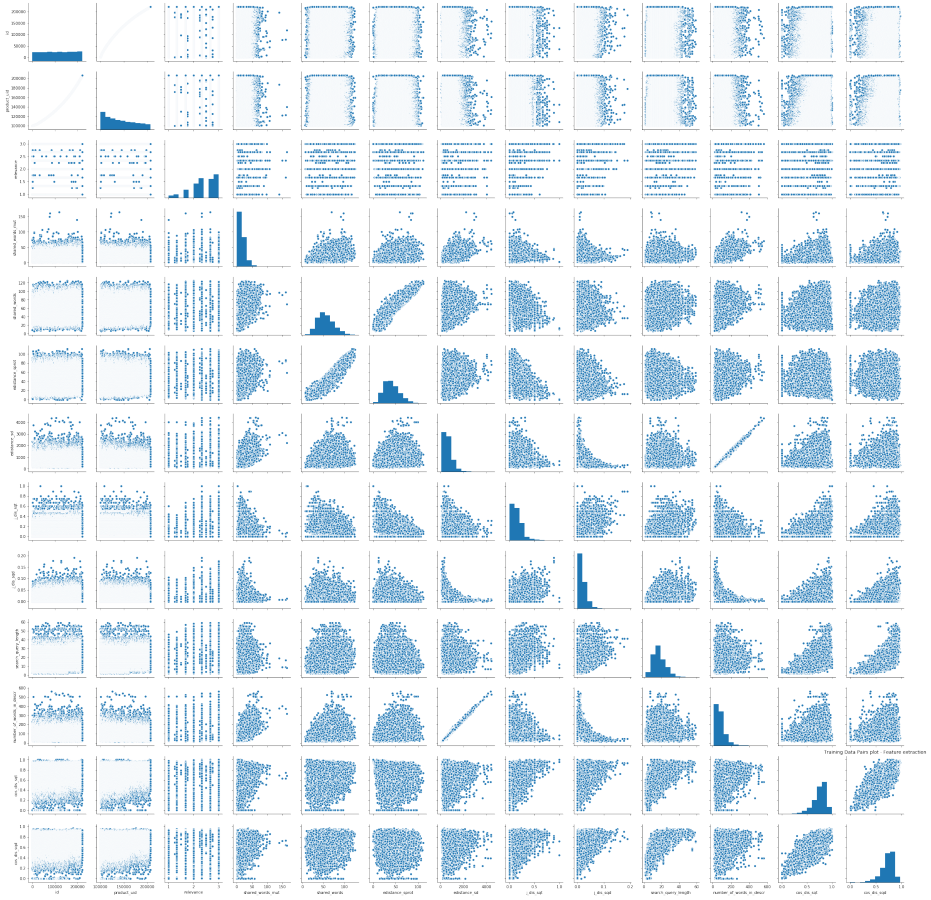
Histogram of Edit Distance (Search Term Vs Product Title)



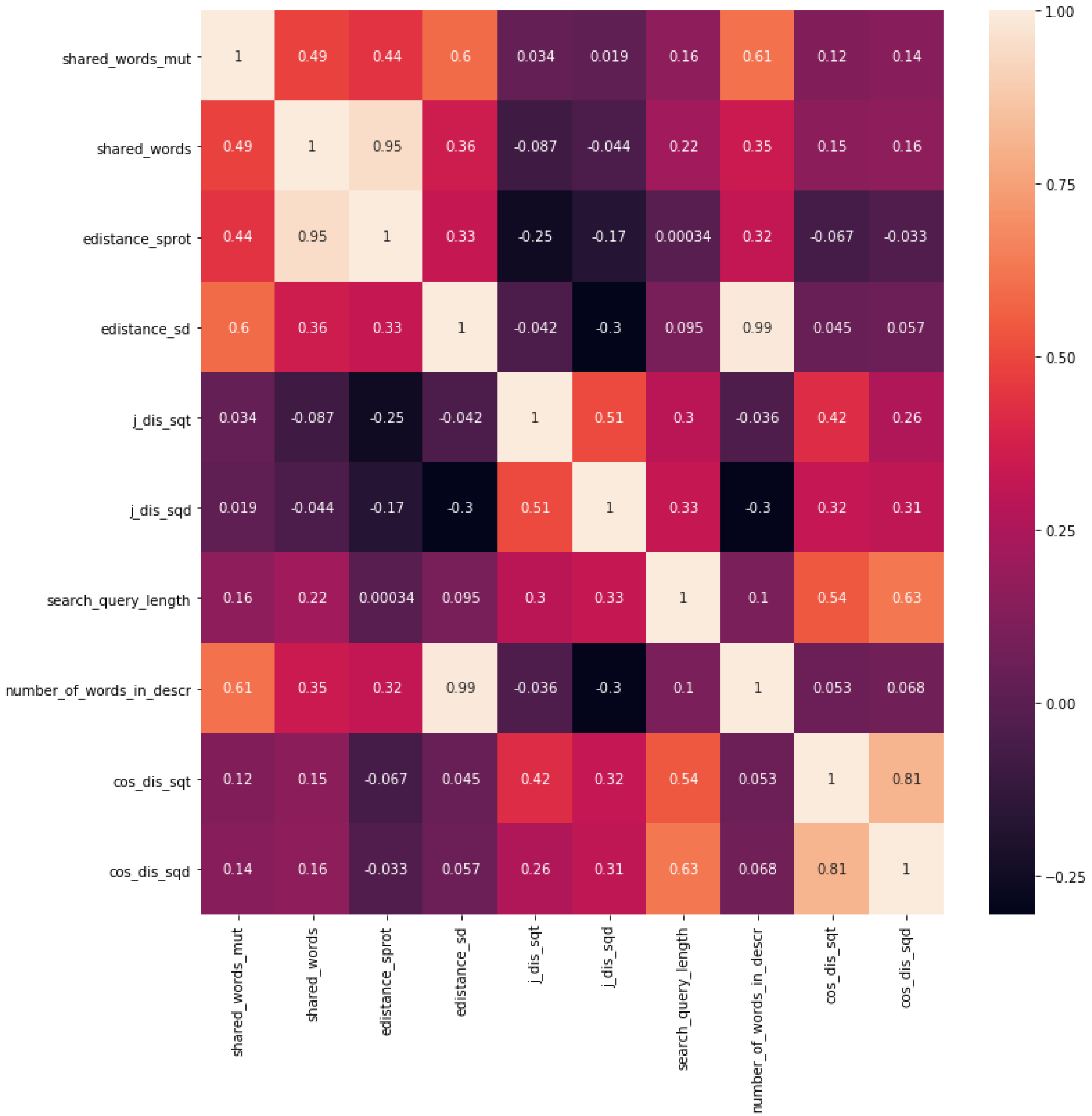
Histogram of Search query

Training set – Heat map





Testing Set – Heat map



MODELS USED

✓ The best regressor predicts the relevance score for the Kaggle's test data with minimum prediction error.

Comparing the mean RMSE :

Algorithm	RMSE
Linear Regression	0.49692
Random Forest	0.57807
Naïve Bayes	0.49694

- Some times Random Forest will over fit more easily than a linear regression
- In our case, Naïve Bayes and Linear Regression provides similar result
- Multiple Linear Regression provides more interpretability

Future Scope

- As we can see, the prediction error is high
- The high error also means that there are other explanatory features that influence the product search relevance scores derived by the customer
- Can also take polarity of words into consideration
- Use of deep learning / Xgboost with Bagging

Thank you