FLIPO1 PRESETANTION REPORT

BAOBAO SONG

Abstract. This report mainly explains, analyses and solve the problem of predicting the relevance score with the database which is given.The database contains some information includes relevance score of products information and search items. The report contains e parts. The first part decribes the problem, interprets the data and illustrates evaluation criteria. Moreover,the report do some test preprocessing, which includes remove punctuation and stopwords and others. Third, select some text features to construct feature matrix. Then, ~~RF~~ different models are used and parameters are adjusted to find the best result. The last one is conclusion.

Contents

Date: 2021-01-15.

1. Introduction

1.1. **Problem Describtion.** The Home Depot wants to solve the problem that search result matches the search query. They gives some data about the relevance scores which users rated and we need to predict the scores between search words and product id. The raw dataset contains train set with 74067 samples and test set with 166693 samples which need to write relevance score. Some other information such as product title also provide.

1.2. **Dataset Interpretation.** Here's the data in the dataset.

Table 1:Data

Name	Attribute or description
train.csv	id, product_id, product_title,search_term, relevance
test.csv	id, product_id, product_title,search_term
product_descriptions.csv	product_uid,product_description
attributes.csv	product_uid,name,value
relevance_instructions.doc	rate the relevance of a search result
sample_submission.csv	id,relevance

1.3. **Evaluation Criteria.** Home Depot have the score ranking of the test set which users have done before. Therefore, by predicting the relevanca score, MSE methods are used to evaluate the model performance. Before experiment, the evaluation methods to assess the model performance is very important, usually it has the MSE methods to evaluate.

2. Text Preprocessing

2.1. **Punctuation and Stopword Removment.** Before removing puctuation and stopwords, the test set and train set need to merge and concat the column of product description. Then the punctuations are removed besides ””, ”/”, -”, ” % ” because they may have some useful purpose. Next ppercase is convertes to lowercase and Arabic numerals arechanged. Finally, stopwords are removed. The results are shown in the figure 1.

2.2. **Stemming.** Using SnowballStemmer to extract stem of text and results are shown in the figure 2.

3. Text Feature

3.1. **Fundamental Feature.** Fundamental Features can directly get in the dataset nd it is the visual representation of relevance. Fundamental features are shown below and the code results is in figure 3.

- the length of search term
- the number of same keywords between search terms and product title
- the number of same keywords between search terms and product description

	id	product_uid	product_title	search_term	relevance	is_train	product_description	search_term_transform	product_title_transform	pro_des_trans
0	2	100001	Simpson Strong-Tie 12-Gauge Angle	angle bracket	3.00	1	Not only do angles make joints stronger, they ...	angle bracket	simpson strong - tie 12-gauge angle	angles make joints stronger also provide consi...
1	3	100001	Simpson Strong-Tie 12-Gauge Angle	l bracket	2.50	1	Not only do angles make joints stronger, they ...	l bracket	simpson strong - tie 12-gauge angle	angles make joints stronger also provide consi...
2	9	100002	BEHR Premium Textured DeckOver 1-gal. #SC-141 ...	deck over	3.00	1	BEHR Premium Textured DECKOVER is an innovativ...	deck	behr premium textured deck 1- gal . sc -141 tu...	behr premium textured deckover innovative soli...
3	16	100005	Delta Vero 1-Handle Shower Only Faucet Trim Ki...	rain shower head	2.33	1	Update your bathroom with the Delta Vero Singl...	rain shower head	delta vero 1- handle shower faucet trim kit ch...	update bathroom delta vero single - handle sho...
4	17	100005	Delta Vero 1-Handle Shower Only Faucet Trim Ki...	shower only faucet	2.67	1	Update your bathroom with the Delta Vero Singl...	shower faucet	delta vero 1- handle shower faucet trim kit ch...	update bathroom delta vero single - handle sho...
...	...	...	...	...	...	...	...	...	...	...
240755	240756	224424	stufurhome Norma 24 in. W x 16 in. D x 34 in. ...	24 whtie storage cabinet	NaN	0	Create a neat yet stylish storage space for or...	24 whtie storage cabinet	stufurhome norma 24 . w x 16 . x 34 . h linen ...	create neat yet stylish storage space organizi...
240756	240757	224425	Home Decorators Collection 49 in. D Alessandro...	adirondeck cusion	NaN	0	Our Bullnose Adirondack Chair Cushions fit Adi...	adirondeck cusion	home decorators collection 49 . alessandro spi...	bullnose adirondack chair cushions fit adirond...
240757	240758	224426	Simpson Strong-Tie HB 3-1/2 x 14 in. Top Flang...	hb	NaN	0	Joist hangers are designed to provide support ...	hb	simpson strong - tie hb 3-1/2 x 14 . top flang...	joist hangers designed provide support underne...
240758	240759	224427	1/4 in. -20 tpi x 1-1/2 in. Stainless Steel Bu...	hex sockets	NaN	0	These socket cap screws are ideal for applicat...	hex sockets	1/4 . -20 tpi x 1-1/2 . stainless steel button...	socket cap screws ideal applications require w...
240759	240760	224428	Bosch 4 in. Bi-Metal Hole Saw	4 inch hole saw	NaN	0	The Bosch quick change bi-metal hole saws feat...	4 inch hole saw	bosch 4 . bi - metal hole saw	bosch quick change bi - metal hole saws featur...

240760 rows × 10 columns

FIGURE 1. Punctuations and Stopwords are removed in the dataset

	id	product_uid	product_title	search_term	relevance	product_description	search_term_transform	product_title_transform	pro_des_trans
0	2	100001	simpson strong - tie 12- gaug angl	angl bracket	3.00	angl make joint stronger also provid consist s...	angle bracket	simpson strong - tie 12-gauge angle	angles make joints stronger also provide consi...
1	3	100001	simpson strong - tie 12- gaug angl	l bracket	2.50	angl make joint stronger also provid consist s...	l bracket	simpson strong - tie 12-gauge angle	angles make joints stronger also provide consi...
2	9	100002	behr premium textur deck 1- gal . sc -141 tugb...	deck	3.00	behr premium textur deckov innov solid color c...	deck	behr premium textured deck 1- gal . sc -141 tu...	behr premium textured deckover innovative soli...
3	16	100005	delta vero 1- handl shower faucet trim kit chr...	rain shower head	2.33	updat bathroom delta vero singl - handl shower...	rain shower head	delta vero 1- handle shower faucet trim kit ch...	update bathroom delta vero single - handle sho...
4	17	100005	delta vero 1- handl shower faucet trim kit chr...	shower faucet	2.67	updat bathroom delta vero singl - handl shower...	shower faucet	delta vero 1- handle shower faucet trim kit ch...	update bathroom delta vero single - handle sho...
...	...	...	...	...	...	...	...	...	...
240755	240756	224424	stufurhom norma 24 . w x 16 . x 34 . h linen s...	24 whtie storag cabinet	NaN	creat neat yet stylish storag space organ bath...	24 whtie storage cabinet	stufurhome norma 24 . w x 16 . x 34 . h linen ...	create neat yet stylish storage space organizi...
240756	240757	224425	home decor collect 49 . alessandro spiceberri ...	adirondeck cusion	NaN	bullnos adirondack chair cushion fit adirondac...	adirondeck cusion	home decorators collection 49 . alessandro spi...	bullnose adirondack chair cushions fit adirond...
240757	240758	224426	simpson strong - tie hb 3-1/2 x 14 . top flang...	hb	NaN	joist hanger design provid support underneath ...	hb	simpson strong - tie hb 3-1/2 x 14 . top flang...	joist hangers designed provide support underne...
240758	240759	224427	1/4 . -20 tpi x 1-1/2 . stainless steel button...	hex socket	NaN	socket cap screw ideal applic requir well tool...	hex sockets	1/4 . -20 tpi x 1-1/2 . stainless steel button...	socket cap screws ideal applications require w...
240759	240760	224428	bosch 4 . bi - metal hole saw	4 inch hole saw	NaN	bosch quick chang bi - metal hole saw featur p...	4 inch hole saw	bosch 4 . bi - metal hole saw	bosch quick change bi - metal hole saws featur...

240760 rows × 9 columns

FIGURE 2. Usage of SnowballStemmer

3.2. **Levenshtein Distance.** Levenshtein Distance means the times that one string change to another string. Therefore, Levenshtein Distance is used to measure the similarity between search words and other information about product.The features

	id	product_uid	product_title	search_term	relevance	pro_des_trans	len_of_query	commons_in_title	commons_in_desc
0	2	100001	simpson strong - tie 12-gaug angl	angl bracket	3.00	angles make joints stronger also provide consi...	2	1	1
1	3	100001	simpson strong - tie 12-gaug angl	l bracket	2.50	angles make joints stronger also provide consi...	2	1	1
2	9	100002	behr premium textur deck 1-gal . sc -141 tugb...	deck	3.00	behr premium textured deckover innovative soli...	1	1	1
3	16	100005	delta vero 1-handl shower faucet trim kit chr...	rain shower head	2.33	update bathroom delta vero single - handle sho...	3	1	1
4	17	100005	delta vero 1-handl shower faucet trim kit chr...	shower faucet	2.67	update bathroom delta vero single - handle sho...	2	2	2
...	...	...	...	...	...	...	...	...	...
240755	240756	224424	stufurhom norma 24 . w x 16 . x 34 . h linen s...	24 whtie storag cabinet	NaN	create neat yet stylish storage space organizi...	4	3	3
240756	240757	224425	home decor collect 49 . alessandro spiceberri ...	adirondeck cusion	NaN	bullnose adirondack chair cushions fit adirond...	2	0	0
240757	240758	224426	simpson strong - tie hb 3-1/2 x 14 . top flang...	hb	NaN	joist hangers designed provide support underne...	1	1	1
240758	240759	224427	1/4 . -20 tpi x 1-1/2 . stainless steel button...	hex socket	NaN	socket cap screws ideal applications require w...	2	2	2
240759	240760	224428	bosch 4 . bi - metal hole saw	4 inch hole saw	NaN	bosch quick change bi - metal hole saws featur...	4	3	2

240760 rows × 9 columns

FIGURE 3. Fundamental Feature in Database

of Levenshtein distance are shown below and the code results is in figure 4.

- Levenshtein text similarity between search term and product title
- Levenshtein text similarity between search term and product

id	product_uid	product_title	search_term	relevance	product_description	len_of_query	commons_in_title	commons_in_desc	dist_in_title	dist_in_desc
2	100001	simpson strong - tie 12- gaug angl	angl bracket	3.00	angl make joint stronger also provid consist s...	2	1	1	0.173913	0.039539
3	100001	simpson strong - tie 12- gaug angl	l bracket	2.50	angl make joint stronger also provid consist s...	2	1	1	0.139535	0.029801
9	100002	behr premium textur deck 1-gal . sc -141 tugb...	deck	3.00	behr premium textur deckov innov solid color c...	1	1	1	0.112676	0.010323
16	100005	delta vero 1-handl shower faucet trim kit chr...	rain shower head	2.33	updat bathroom delta vero singl - handl shower...	3	1	1	0.389610	0.064257
17	100005	delta vero 1-handl shower faucet trim kit chr...	shower faucet	2.67	updat bathroom delta vero singl - handl shower...	2	2	2	0.351351	0.052525
...	...	...	...	...	...	...	...	...	...	...
240756	224424	stufurhom norma 24 . w x 16 . x 34 . h linen s...	24 whtie storag cabinet	NaN	creat neat yet stylish storag space organ bath...	4	3	3	0.468085	0.076795
240757	224425	home decor collect 49 . alessandro spiceberri ...	adirondeck cusion	NaN	bullnos adirondack chair cushion fit adirondac...	2	0	0	0.268908	0.056478

FIGURE 4. Levenshtein Distance features

3.3. **TF-IDF.** First we set a new column named all_text which includes all words in dataset and make a corpus. Then, bag-of-words is used to change words into vectors. Finally, we should calculate the similarity of different sentence. And expand the sentence size to dictionary size and put zero in the position which don't

have strings before. In this part, two new feature are created at the result is shown in the figure 5.

- Cosine similarity between search terms and product title
- Cosine similarity between search terms and product description

_train	product_description	len_of_query	commons_in_title	commons_in_desc	dist_in_title	dist_in_desc	all_texts	tfidf_cos_sim_in_title	tfidf_cos_sim_in_desc
1	angl make joint stronger also provid consist s...	2	1	1	0.173913	0.039539	simpson strong - tie 12- gaug angl.angl make j...	0.270386	0.191815
1	angl make joint stronger also provid consist s...	2	1	1	0.139535	0.029801	simpson strong - tie 12- gaug angl.angl make j...	0.000000	0.000000
1	behr premium textur deckov innov solid color c...	1	1	1	0.112676	0.010323	behr premium textur deck 1- gal . sc -141 tugb...	0.194095	0.212764
1	updat bathroom delta vero singl - handl shower...	3	1	1	0.389610	0.064257	delta vero 1- handl shower faucet trim kit chr...	0.137359	0.051747
1	updat bathroom delta vero singl - handl shower...	2	2	2	0.351351	0.052525	delta vero 1- handl shower faucet trim kit chr...	0.355233	0.133826
...	...	...	...	...	...	...	...	...	...
0	creat neat yet stylish storag space organ bath...	4	3	3	0.468085	0.076795	stufurhom norma 24 . w x 16 . x 34 . h linen s...	0.214326	0.136015

FIGURE 5. TF-IDF features

3.4. **Word2Vec.** Word2Vec assigns a value to each word according to sentences and the context. So, we need to split the sentences from all_test and then split the word. Finally, the value of words we get and then calculate average vector according to sentences, and gain cosine similarity of two sentences.Two new feature are created at the result is shown in the figure 6.

- Cosine similarity between search terms and product title
- Cosine similarity between search terms and product description

mons_in_title	commons_in_desc	dist_in_title	dist_in_desc	all_texts	tfidf_cos_sim_in_title	tfidf_cos_sim_in_desc	w2v_cos_sim_in_title	w2v_cos_sim_in_desc
1	1	0.173913	0.039539	simpson strong - tie 12- gaug angl.angl make j...	0.270386	0.191815	0.393470	0.476581
1	1	0.139535	0.029801	simpson strong - tie 12- gaug angl.angl make j...	0.000000	0.000000	0.280904	0.275490
1	1	0.112676	0.010323	behr premium textur deck 1- gal . sc -141 tugb...	0.194095	0.212764	0.493544	0.358379
1	1	0.389610	0.064257	delta vero 1- handl shower faucet trim kit chr...	0.137359	0.051747	0.543286	0.367758
2	2	0.351351	0.052525	delta vero 1- handl shower faucet trim kit chr...	0.355233	0.133826	0.799418	0.565387
...	...	...	...	...	...	...	...	...
3	3	0.468085	0.076795	stufurhom norma 24 . w x 16 . x 34 . h linen s	0.214326	0.136015	0.604527	0.580506

FIGURE 6. Word2Vec features

4. Modeling and Forecasting

Using different models to predict the relevance and choose the best one. Bagging Regression, Gradient Boost Regression, Random Forest Regression ~~to predict the relevance.~~ and XGB Regression are and compared.In each method, adjusting parameters include max depth and iterations. CV error is used to judge the model performance and its scoring is MSE. At the end, Gradient Boost Regression performance is best, Random Forest Regression is second and Bagging Regression is last. Parameters adjustment processes of Gradient Boost Regression are shown in figure 7 and 8.

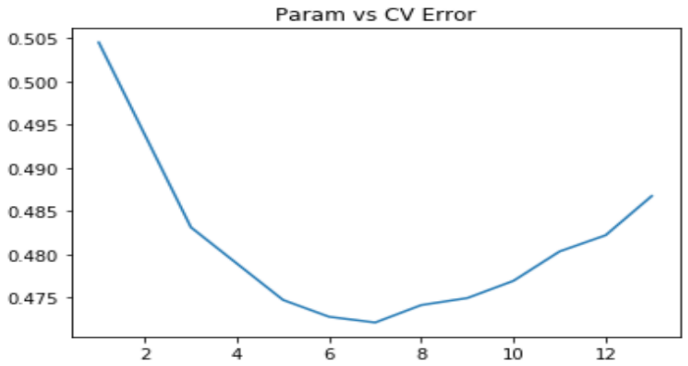

Figure 7. Adjusting max depth of ~~Random Forest Regression~~ Gradient Boosting Regressor

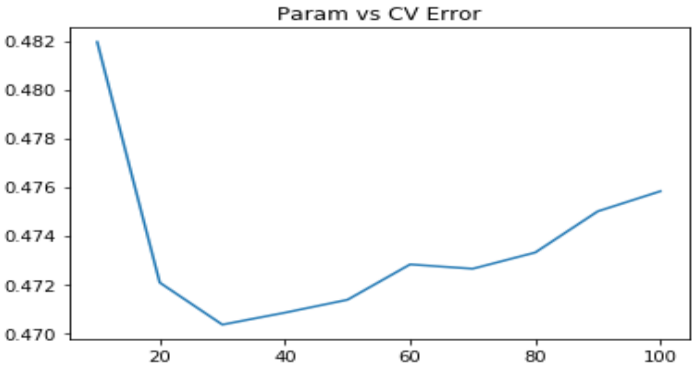

Figure 8. Adjusting iterations of Gradient Boosting Regressor

The result:

- score:~~0.48032~~ 0.47101
- rank:347/2123

5. Conclusions

- ~~More features need to choose and more models need to compare~~In this competition, text preprocessing and text feature are important and I know the procedure of dealing NLP problem. Different features that use different method can help the final model predict well.
- Compare to midterm presentation, MSE decreased from 0.48032 to 0.47101. The reason mainly is more features and different modeling. By adding four features of Word2vec and TFIDF, more information are consideres. Besides, Gradient Boosting Regressor is choosen because of its best performance which means that this model is more suitable for the problem.
- There are some shortcomings. The first team in this competiton is 0.43 and there exists gap. From my point of view, the reason mainly is that some data about product information don't use, and the parameters need to fine-tuned.