Flip 01 Project Report

Baobao Song

Hunan University, China

Introduction

This is a correlation prediction problem. Home Depot wants to improve their customers' shopping experience by developing a model that can accurately predict the relevance of search results of Home Depot product. Many information are given about product description and relevance ratings. And the goal is to predict the relevance for each pair that contains products and searches listed in the test set.The raw datasets contain five files, which attributes are shown below.

Name	Attribute
product_descriptions.csv	product_uid,product_description
attributes.csv	product_id,name,value
train.csv	id, product_id, product_title,search_term, relevance
test.csv	id, product_id, product_title,search_term
sample_submission.csv	id,relevance

Test Preprocessing

- Merge test.csv and train.csv datasets and concat product description(product_descriptions),name and value(attributes.csv)
- Convert to lowercase
- Remove punctuation
- Remove stopwords
- Stemming by SnowballStemmer

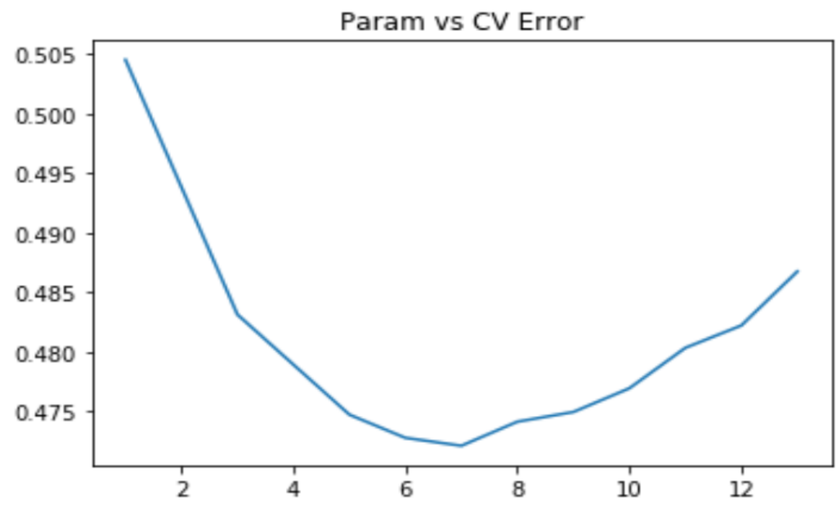
Text features

- Fundamental feature1:the length of search term
- Fundamental feature2:the number of same keywords between search terms and product title
- Fundamental feature3:the number of same keywords between search terms and product description
- Levenshtein feature1:Levenshtein text similarity between search terms and product title
- Levenshtein feature2:Levenshtein text similarity between search terms and product description
- TF-IDF feature1:Cosine similarity between search terms and product title
- TF-IDF feature2:Cosine similarity between search terms and product description
- Word2Vec feature1:Cosine similarity between search terms and product title
- Word2Vec feature2:Cosine similarity between search terms and product description

Different Models and Parameters Adjustment

Using different models to predict the relevance and choose the best one. In each method, adjusting parameters include max depth and iterations. CV error is used to judge the model performance and its scoring is MSE. The models and the process of parameters adjustment(such as g max depth of Gradient Boosting Regressor) are below.

- Bagging Regression
- Gradient Boost Regression
- Random Forest Regression



Results

- Model performance:Gradient Boost Regression>Random Forest Regression>Bagging Regression
- Parameter(Gradient Boost Regression):max_depth:7,iteration:30
- Score:0.47101(RMSE)
- Rank:347/2123

Conclusion

Compare to midterm presentation, MSE decreased from 0.48032 to 0.4710. The reason mainly is more features and different modeling. Different features that use different method can help the final model predict well. And by adjusting four models' parameters, Gradient Boost Regression perform best. In the next step, using more information about product and fine-tuned parameters may lead a better result.