Home Depot Product Search Relevance Prediction

Ling
Peng
Mohamed

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

Statistical Learning & Analytics
Spring, 2016

Introduction:

Key problem:

- Regression model to predict the relevance Score between the search query and the product
- Manually prediction vs. Model prediction

Business Understanding:

- Help improving the user experience

Data Understanding:

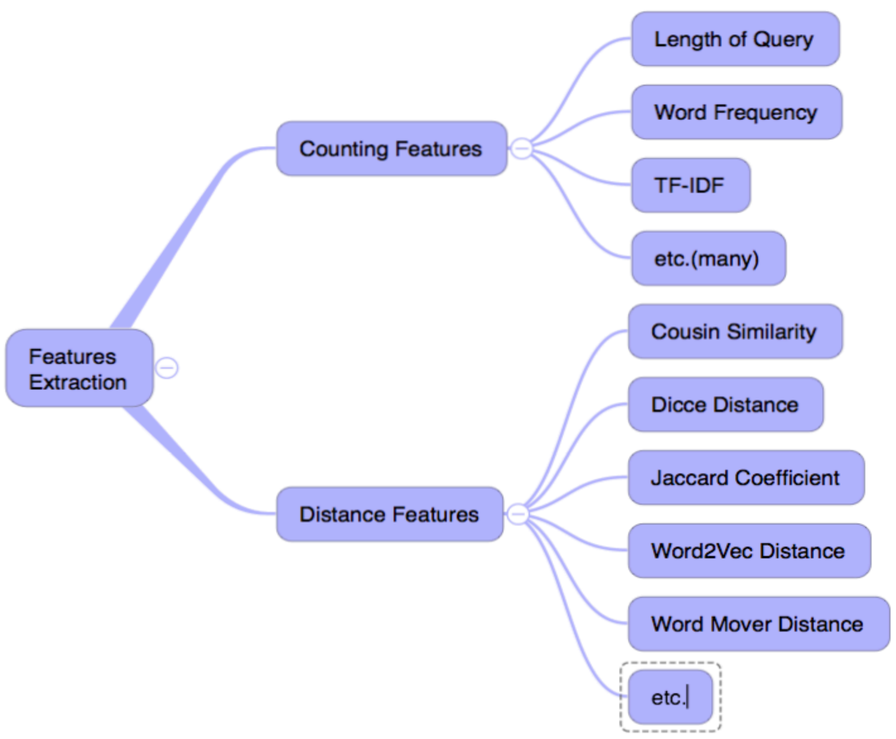
- Data source: kaggle.com
- Pure text: search query ; product title/description/attribute
- 15% miss-spelling in query
- Queries are quite short , 4 words in average
- 240760 rows(train + test)
- No features

Data Preparation:

Data Pre-Processing:

- Spelling checking
- Set stop words manually
- Stemming based on EDA
- Query expanding by word replacement with synonym and definition

Feature Extraction



Model:

Random Forest

Gradient Boosting with Bagging

Different kinds of linear regression

Support Vector Regression

Results & Evaluation

172 ↑902 Go duck! 0.46447 20 Mon, 25 Apr 2016 03:50:05

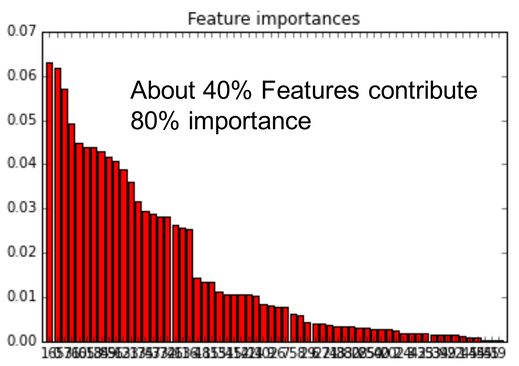
$40,000 2,147 teams

Home Depot Product Search Relevance

Rank:8.0%

Model Evaluation: Root Mean Square Error:

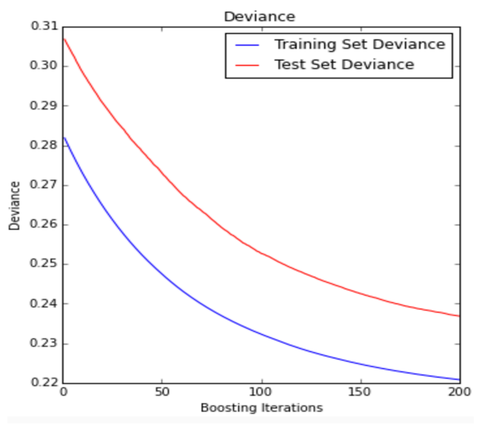
$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$



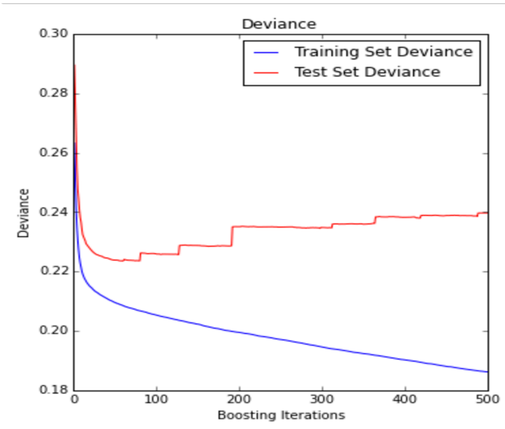
Base - Line RMSE
Linear Regression: 0.49

Single Decision tree:0.64
VS
Random Forest:0.4644

Just Boosting:0.48
VS
Boosting with Bagging:0.468



Not enough features
No sweet spot



After feature extraction

Insight

- Cross- validation is still important for random forest.
- When the features are not enough to capture main information , there is no need to tune parameters
- When tuning the model , some parameters are conflicting with each other: Using fitting curve and grid search to find the best best combination
- Feature extraction is the most important part in this project
- Because the query is quite short, we just considered the term frequency , because the sequence is not important.

DEMO