Feature Ranking in Search Rank

Yilin Wei Instructor: Dr. German Creamer

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®
Business Intelligence & Analytics

Briefing

The goal of this project is to make data driven decision based on machine learning. Based on the top 4 important features in predicting search rank and observation of experimental design, I put forward 3 suggestions to improve search rank for Diapers.com, an e-commerce website of baby items.

Methodology

Data Collection

With the rising trend and large traffic of "baby formula", I scrape top 30 search results of keywords:

- Baby formula (head-term keyword)
- Best baby formula (long-tail and informational keyword)
- Buy baby formula (transactional keyword)

(Clarke, 2015) summarized three key principles of Google ranking: authority, trust and relevance, so I select the following features:

Feature	Meaning
Main_fb_share	Number of shares of main page on Facebook
Main_tw_count	Number of shares of main page on Twitter
Sub_fb_share	Number of shares of subpage on Facebook
Sub_tw_count	Number of shares of subpage on Twitter
Year of register	Number of of years exist
Word Rank	World rank of main page
Inbound link	Number of subpage inbound links
Page Speed	Score of load page speed
Img_count	Number of images
Desc_sim	Similarity between description for <meta> and keyword
H1_sim	Similarity between text for <h1> and keyword
H2_sim	Similarity between text for <h2> and keyword
H3_sim	Similarity between text for <h3> and keyword
Img_sim	Similarity between text for <img> and keyword

Preprocess

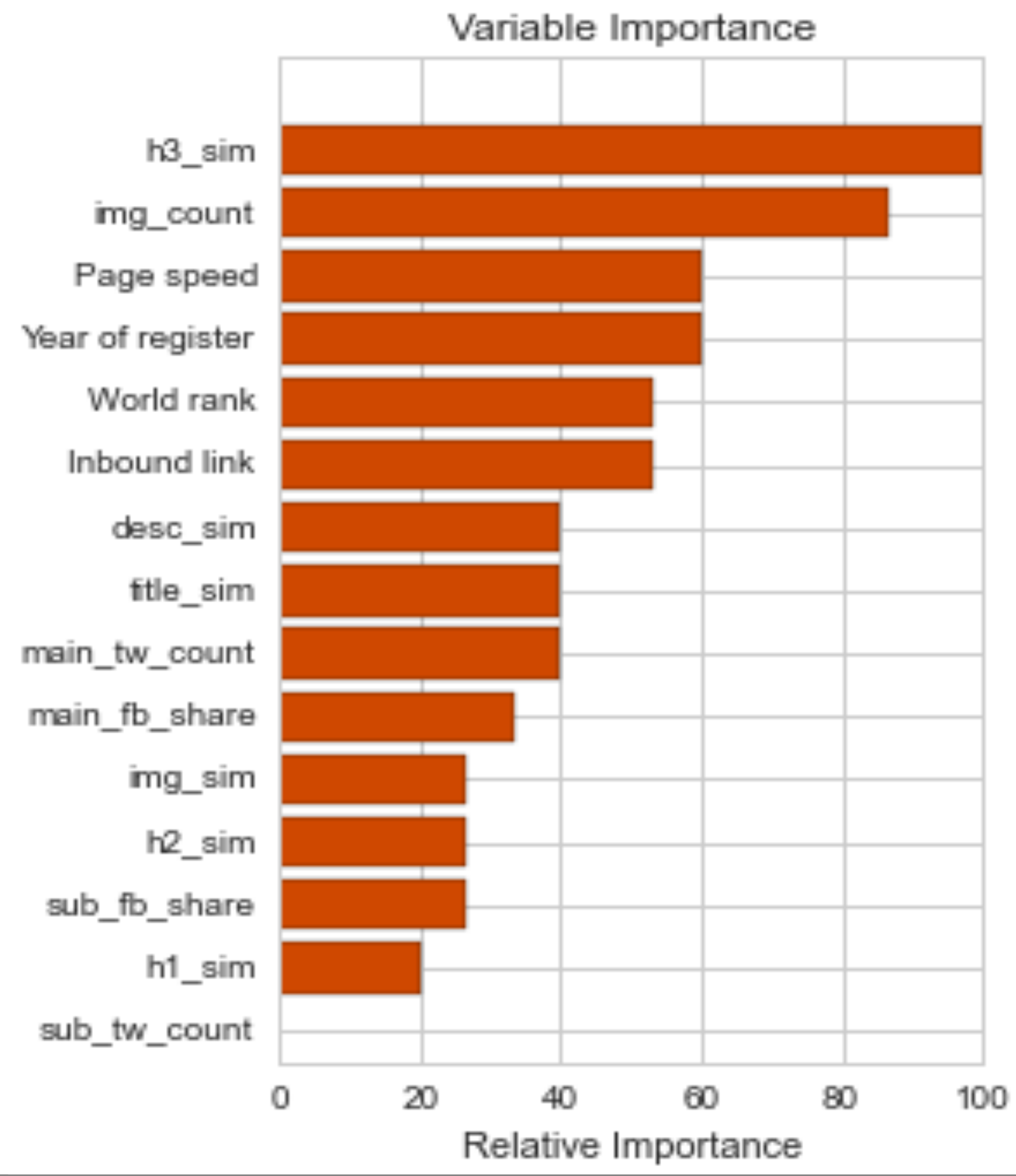
I replace the missing value with mean and normalize the data. The dataset is randomly divided into train (90%) and test(10%).

Build Model

Empirical researches show that gradient boosted regression trees is very suitable for web search ranking, so I use this algorithm to build model. I apply grid search on loss function, max_depth, n_extimators and max_leaf_nodes. The R Square of final model is 0.4630 and MSE is 46.8352.

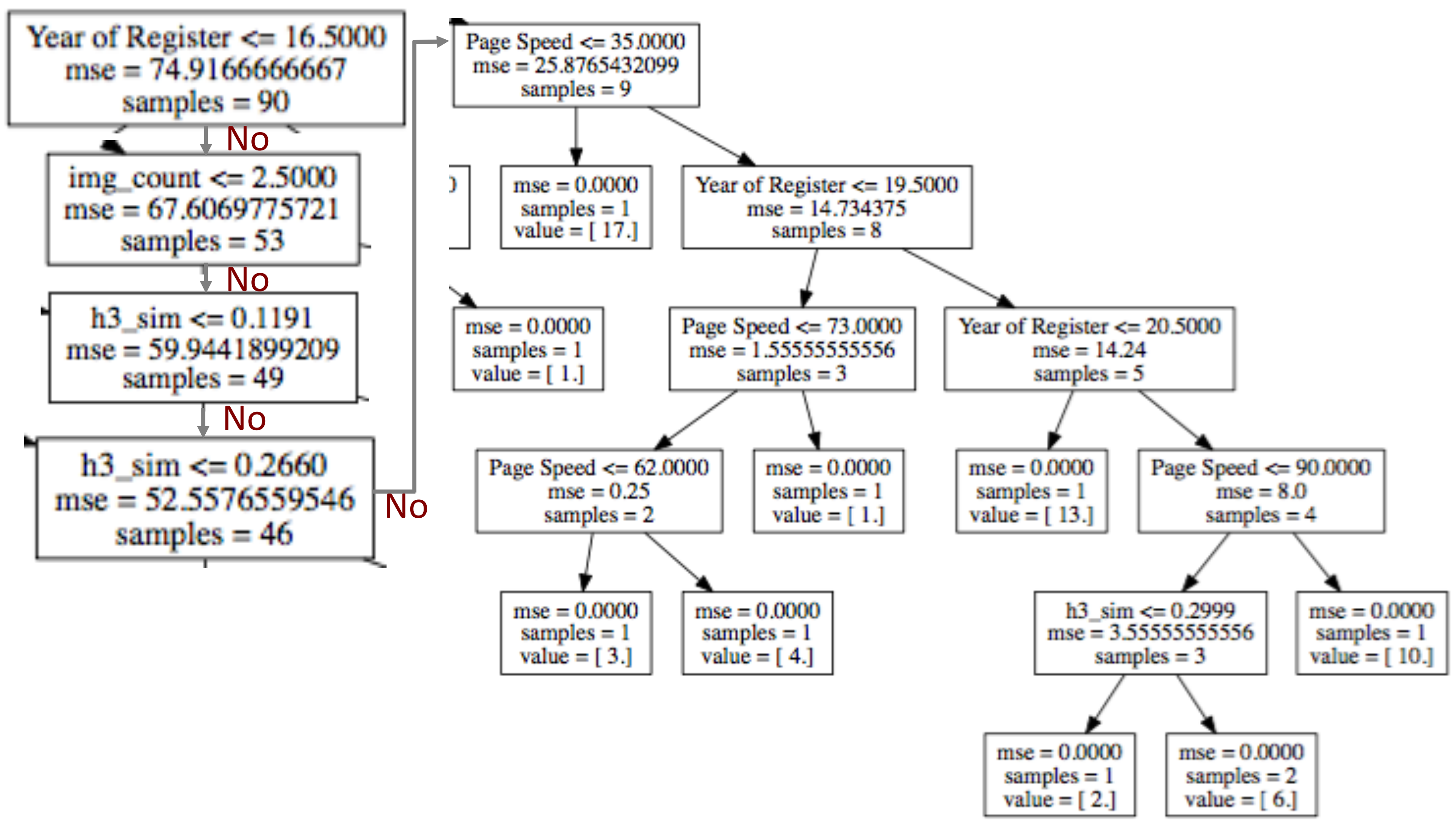
Findings

- H3_sim is more important than other similarity factors; social media features of main page is more important than that of subpage; all factors related to trusty are important



Methodology

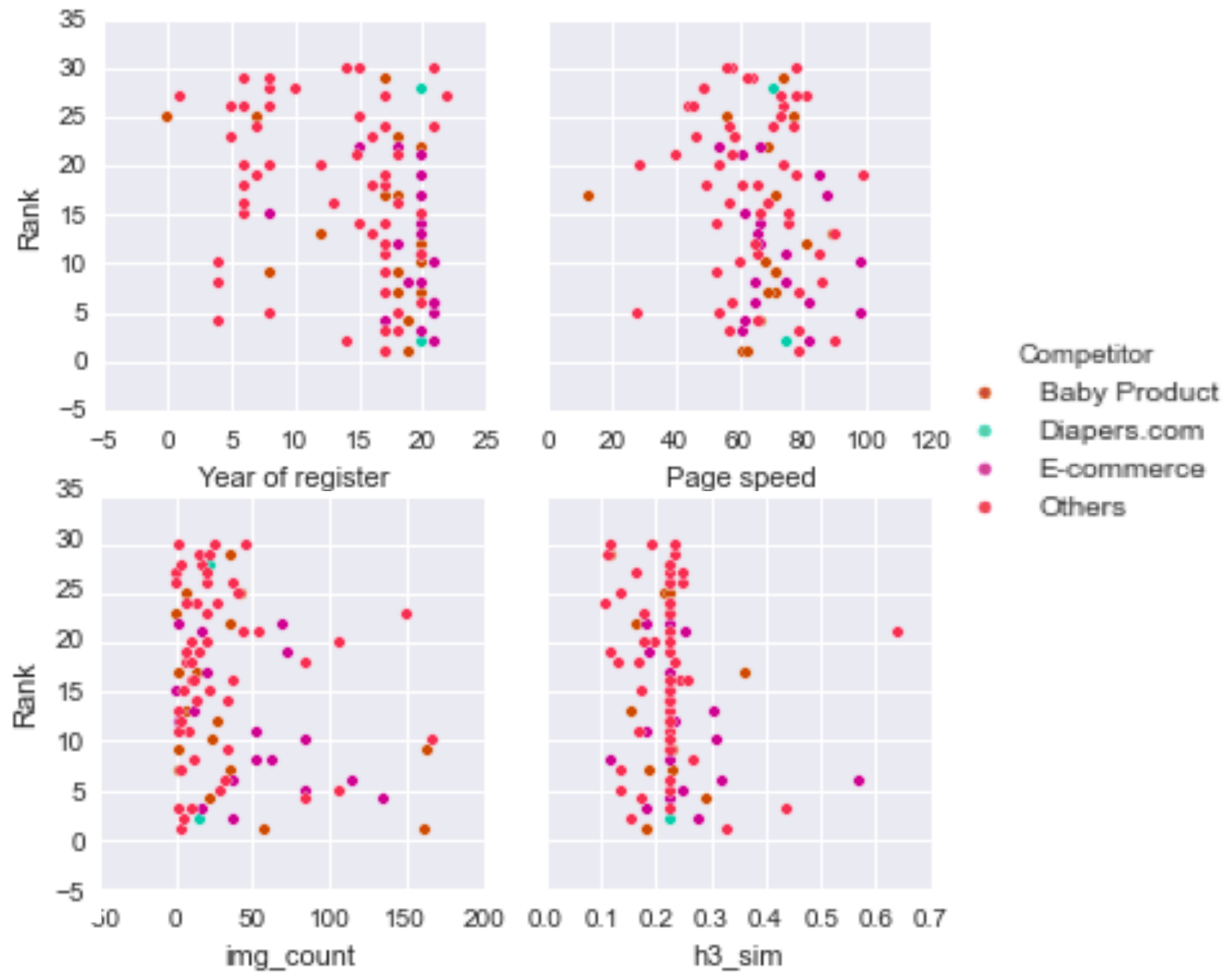
- Decision tree shows when the **year of register is larger than 19.5, img_count is larger than 2.5, h3_sim is larger than 0.2660 and page speed is larger than 35**, the webpage is likely to rank in top 4



Decision Making

Observation

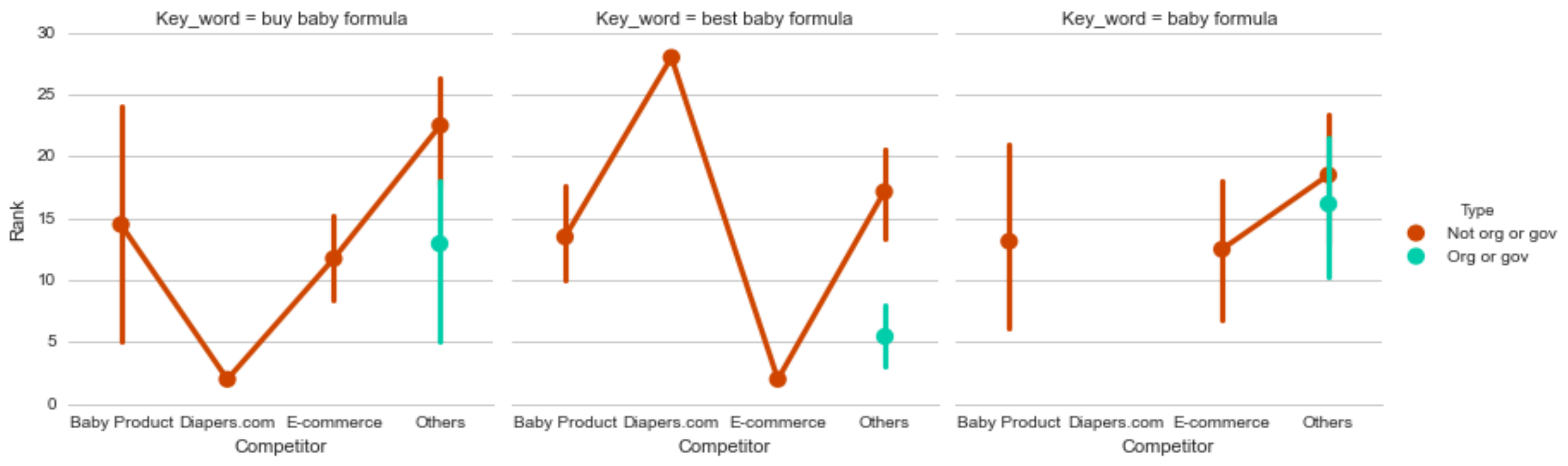
Position of Diapers.com in top 4 features



•A subpage of Diapers.com having lower page speed ranks lower

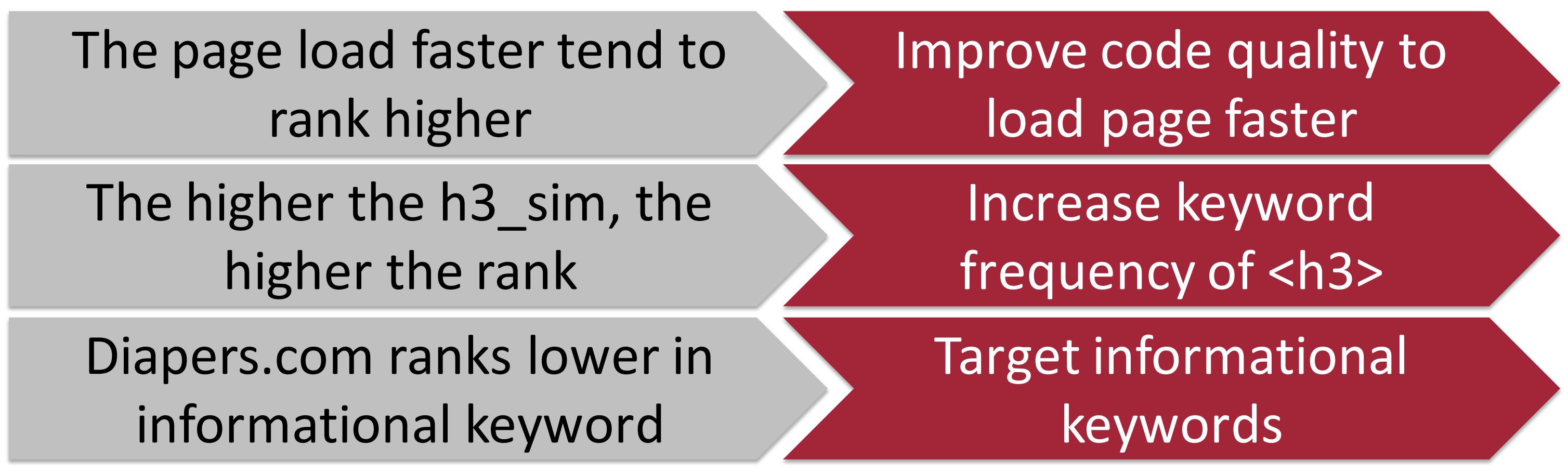
•Diapers.com performs below average in h3_sim

Experimental design of websites



- Authoritative websites (.org or .gov) rank higher
- General e-commerce websites rank higher in informational keyword; Diapers.com ranks higher in transactional keyword

Action



The page load faster tend to rank higher

Improve code quality to load page faster

The higher the h3_sim, the higher the rank

Increase keyword frequency of <h3>

Diapers.com ranks lower in informational keyword

Target informational keywords

Reference:
Clarke, A. (2015). *Search engine optimization 2016: Learn SEO with smart internet marketing strategies.* CreateSpace Independent Publishing Platform.

Portfolio: bit.ly/yilinW