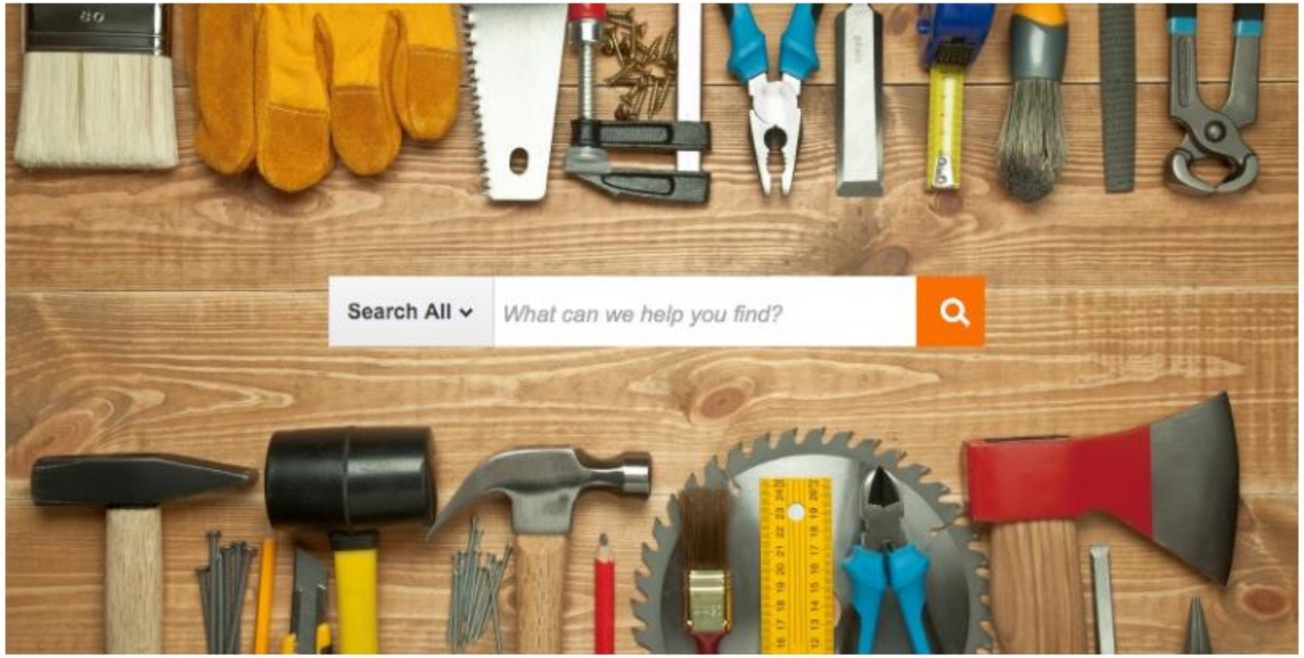
Home Depot

Product Search Relevance Prediction

Kedar Wagholikar - MS in Technology Management at University of Illinois, Urbana-Champaign

Problem Statement



Shoppers rely on Home Depot’s product authority to find and buy the latest products and to get timely solutions to their home improvement needs. From installing a new ceiling fan to remodeling an entire kitchen, with the click of a mouse or tap of the screen, customers expect the correct results to their queries – quickly. Speed, accuracy and delivering a frictionless customer experience are essential.

In this competition, Home Depot is asking Kagglers to help them improve their customers' shopping experience by developing a model that can accurately predict the relevance of search results.

Search relevancy is an implicit measure Home Depot uses to gauge how quickly they can get customers to the right products. Currently, human raters evaluate the impact of potential changes to their search algorithms, which is a slow and subjective process. By removing or minimizing human input in search relevance evaluation, Home Depot hopes to increase the number of iterations their team can perform on the current search algorithms.

Approach

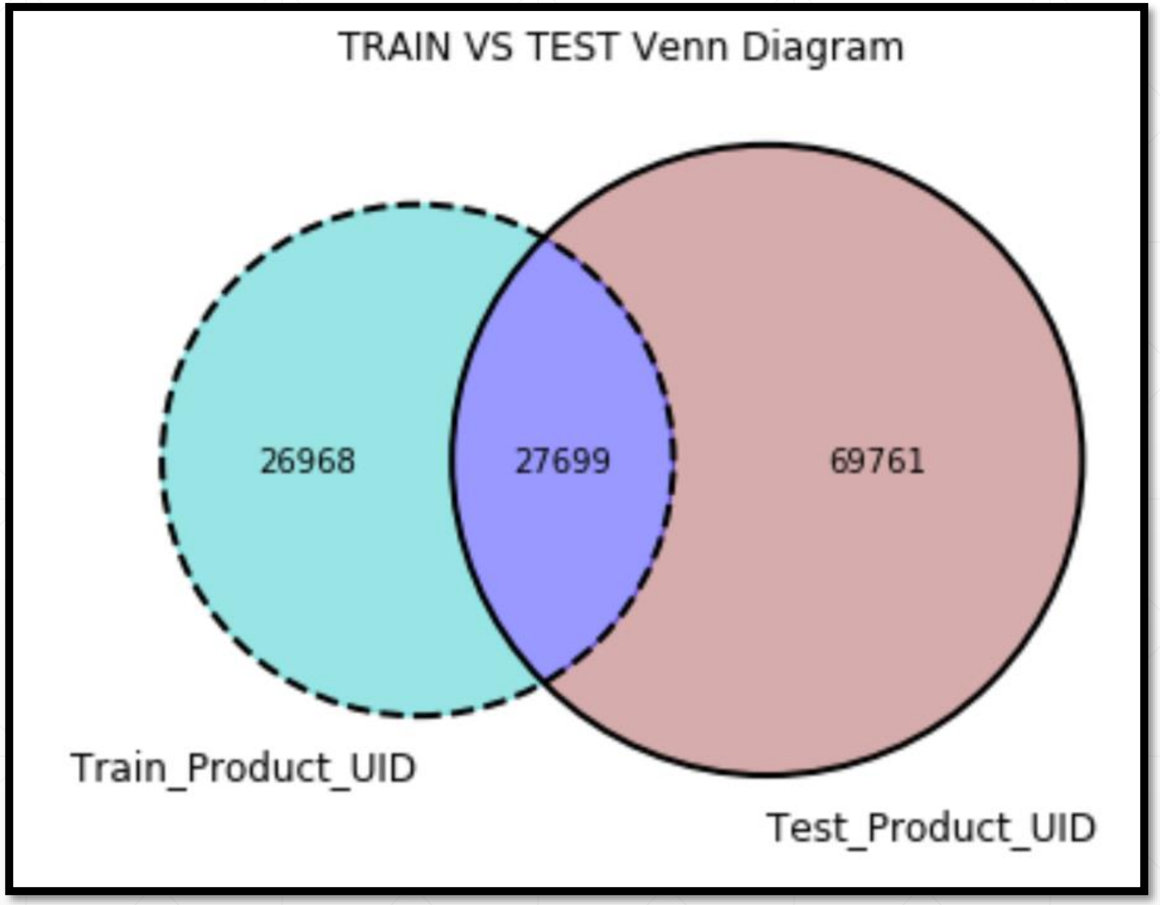
1. Exploring the Data Corpus
2. Cleaning the data
3. Feature Engineering
4. Dimensionality Reduction
5. K-Fold Cross Validation
6. Machine Learning Model

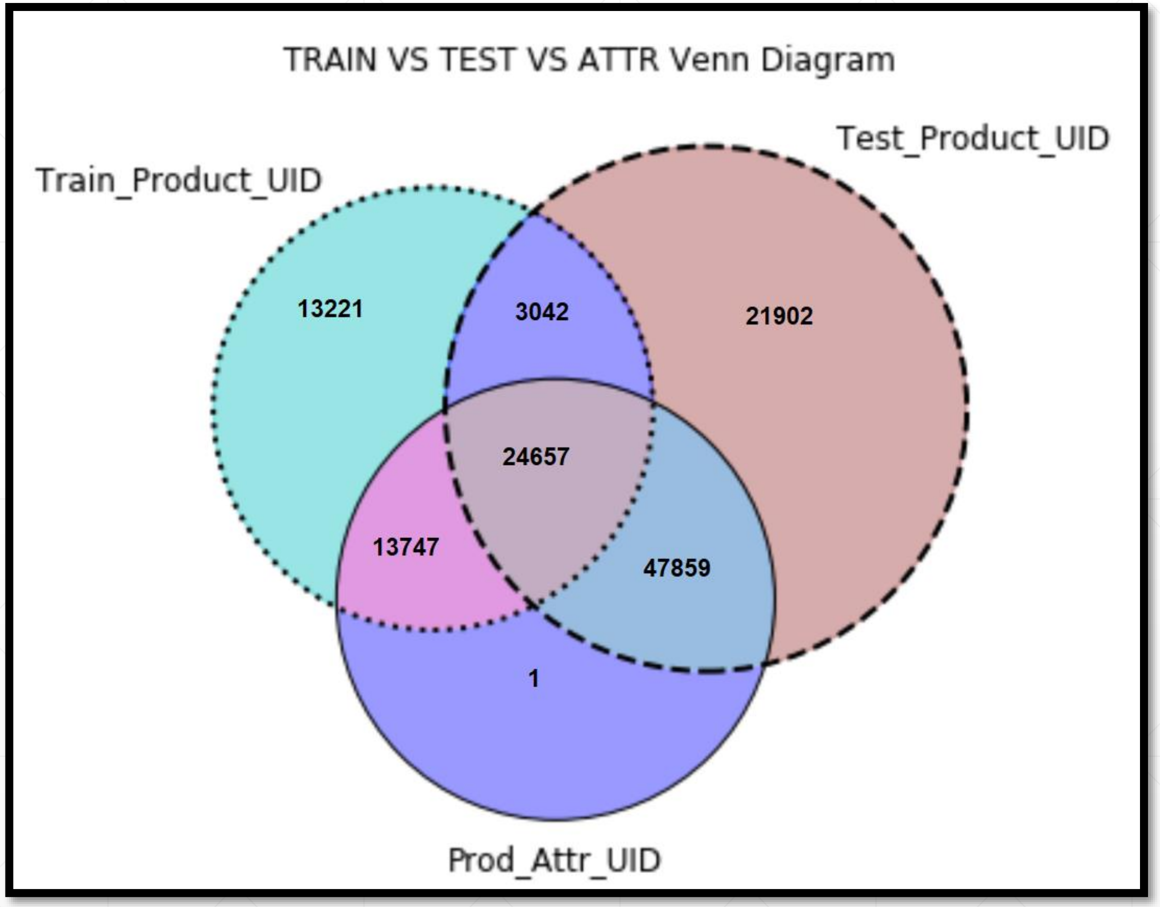
Exploring the Data Corpus

- **Datasets (Format *.csv)**
  - **Product** - There are 124428 products. Each product has a unique ID and a text description
  - **Attributes** - Most products are given extended information like bullet points of product functionality descriptions, size, color, brand, material, etc.
  - **Train Dataset** - Contains 74067 search/product pairs and their corresponding relevance scores
  - **Test Dataset** - Contains 166694 search/product pairs
  - **Relevance Scores -** Score 3 (perfect match), Score 2 (partially or somewhat relevant) and Score 1 (not relevant).
- **Evaluation -** The metric to evaluate the prediction errors is Root Mean Square Error (RMSE)

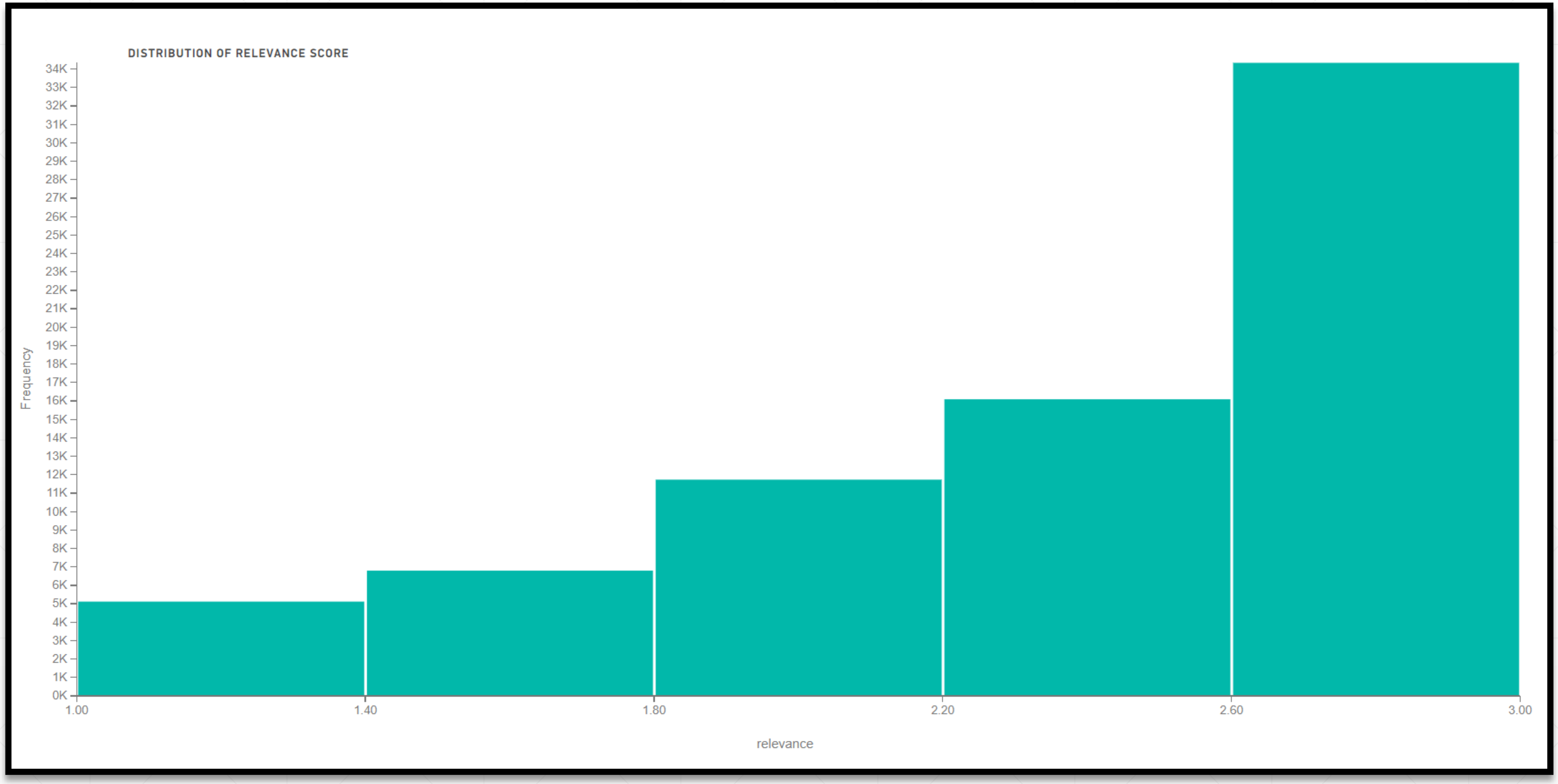
$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_{actual} - \hat{y}_{predicted}\right)^2}$$

Product Intersection between Datasets

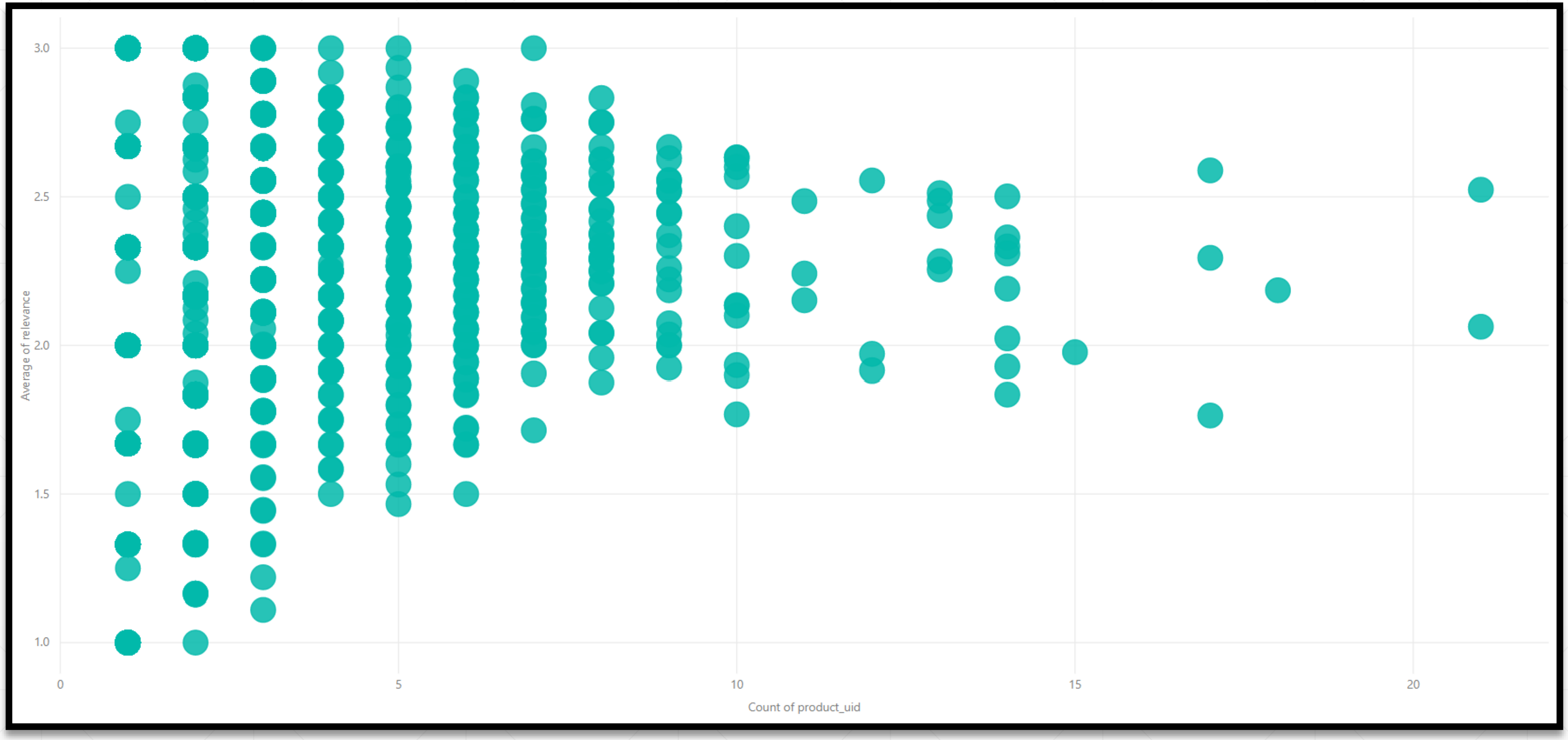




Distribution of Relevance Scores - Skewed

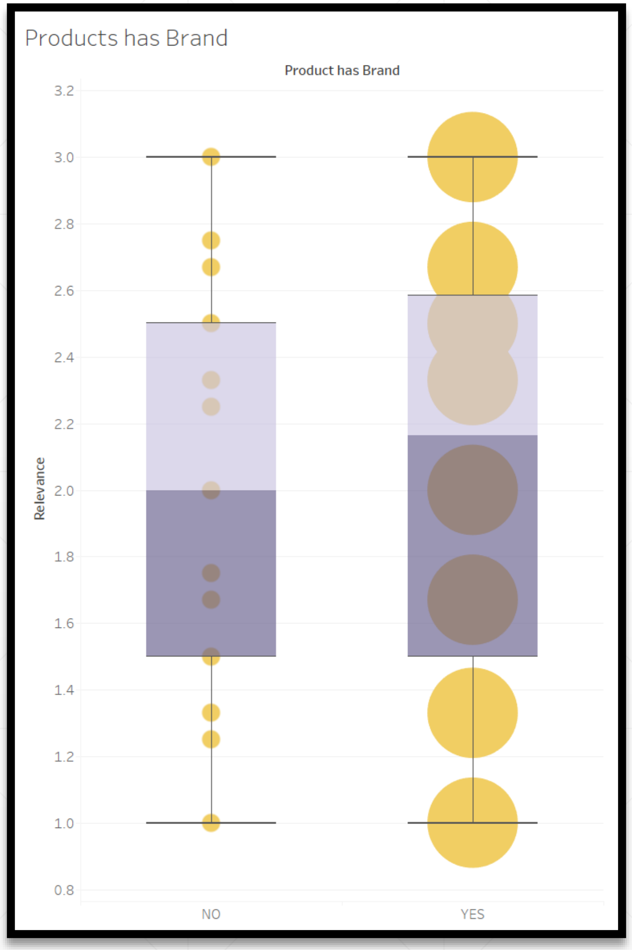


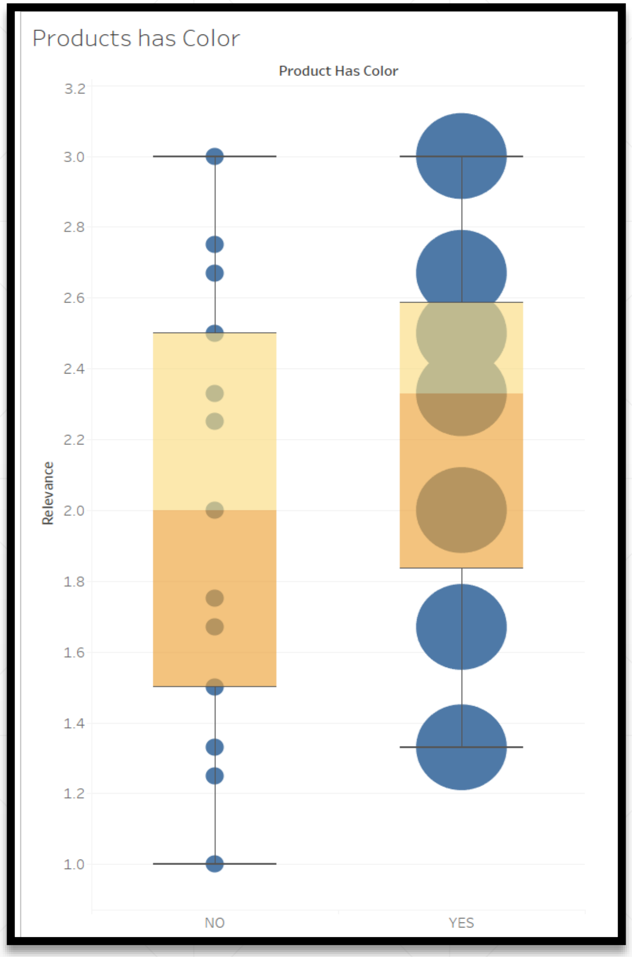
Distribution of Relevance Scores – <u>Normalized</u>

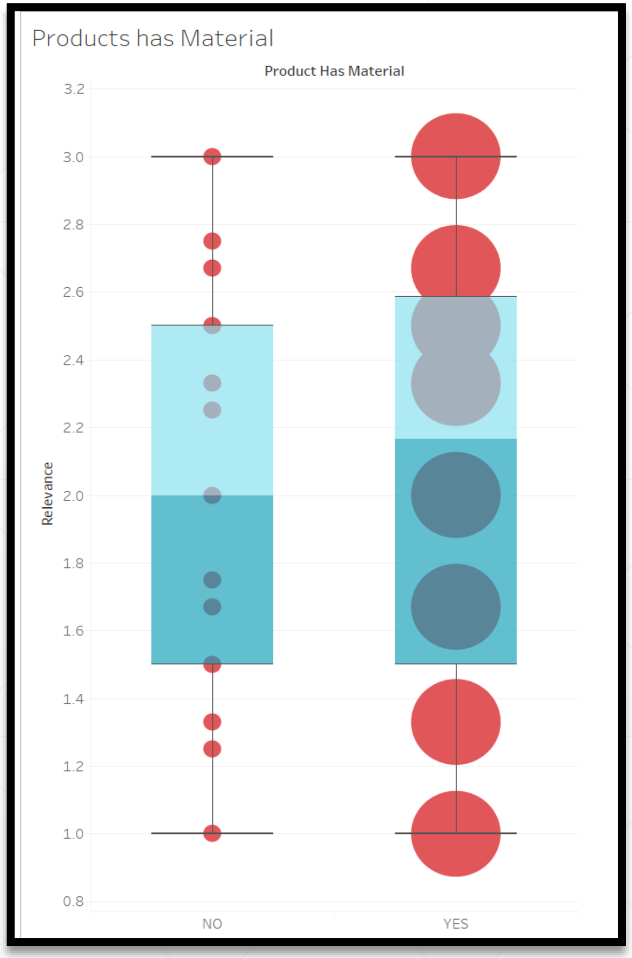


- As expected the variation of the **relevance** decreases as the number of times a *product_uid* occurs in the dataset. The **relevance** score is examined in the context of the number of parameters (words) in a **search_term**.

Include Product Attributes – Brand, Color, Material







Feature Selection

- Combine Train and Test dataset
- Add product description
- Append the product attributes of brand, color and material.

	C	D	E	F	G	H	I	J
1	relevance	product_title	product_uid	search_term	product_description	brand	color	material
2	3	simpson strong tie 12 gaug angl	100001	angl bracket	not onli do angl make joint stronger they also provid more consist straight corner sin	simpson strong tie	nan	galvan steel
3	2.5	simpson strong tie 12 gaug angl	100001	l bracket	not onli do angl make joint stronger they also provid more consist straight corner sin	simpson strong tie	nan	galvan steel
4	3	behr premium textur deckov 1gal	100002	deckover	behr premium textur deckov is an innov solid color coat it will bring your old weathe	behr premium textur deckov	brown tan tugboat	nan
5	2.33	delta vero 1 handl shower onli fa	100005	rain shower head	updat your bathroom with the delta vero singl handl shower faucet trim kit in chrom	delta	chrome chrome	nan
6	2.67	delta vero 1 handl shower onli fa	100005	shower onli faucet	updat your bathroom with the delta vero singl handl shower faucet trim kit in chrom	delta	chrome chrome	nan
7	3	whirlpool 1.9cu.ft. over the rang	100006	convect otr	achiev delici result is almost effortless with thi whirlpool over the rang microwav ho	whirlpool	stainless steel stainless	nan

Cleaning the Data

- Character Case Conversion
- Tokenization
- Remove stop words
- Exclude Punctuation
- Deduct Special Characters
- Stemming
- Lemmatization
- Spell Check Dictionary

Feature Engineering

Using Natural Language Processing, to transform text corpus into computational format.

- Length
  - Likelihood of two document* like search term and product title with disparate length being similar is low compared to document with similar length.
- Common words
  - Likelihood of a better relevance score for more common words between documents like search term and product attributes/title/description is higher than with less common terms.
- Cosine Similarity
  - Tokenizes → Computes the semantic importance of the tokenized word in the document (TF-IDF) → Vectorizes the TF-IDF weights → computes a cosine similarity vector product.
- Mean Similarity
- Fuzzy Wuzzy Ratios
  - Includes spelling mistakes, to understand customers intent.

* Document – set of words like in a search term.

Dimensionality Reduction

By using statistical representation that can describe most but not all of the variance within the data, thereby retaining the relevant information, while reducing the amount of information necessary to represent it.

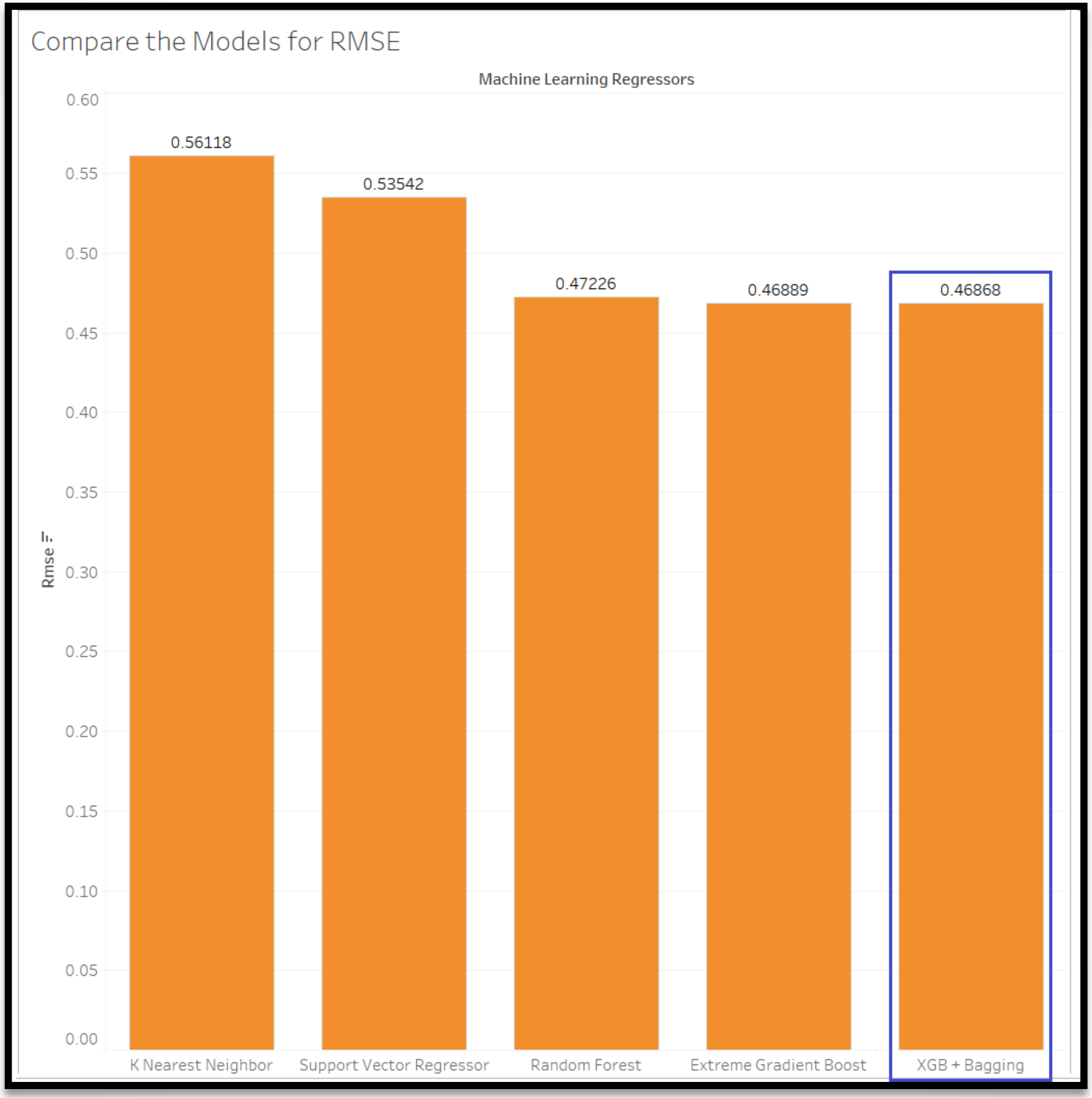
- Total extracted explanatory features are 38.
- Response feature is relevance score
- Objective is to penalize those explanatory features that add noise to the variance of relevance score
- Use LASSO,
  - **L**asso **A**bsolute **S**hrinkage & **S**election **O**perator
  - Regression Analysis method
  - Performs variable selection
  - Enhances prediction accuracy
  - Improves Interpretability of the statistical model.
- Post LASSO, the improved model contains 27 explanatory features.

K-Fold Cross Validation

Achieve generalized predictive model

- To avoid underfitting or overfitting of the prediction curve
- Approach is to fit the model for 'K' number of times*****
  1. Randomly split the train dataset in 80% train and 20% test
  2. Train the model with different regressor algorithms
  3. Test the model
  4. Validate the prediction by computing the RMSE, for i = 1
  5. Repeat steps i to iv, for i = K times
  6. Compute the Mean of RMSEs to find the optimal model
  7. ***Compare the mean RMSEs of regressors (next slide)***
  8. Select the algorithm producing the minimum root mean square error.
- The best regressor predicts the relevance score for the Kaggle's test data with minimum prediction error.

* Optimal K = N/(0.20*N), N is size of Train Dataset



Results

eXtreme Gradient Boosting + Bagging Regressor

Key Takeaways

- The prediction error is high, making the model impractical in business use
- The high error also means that there are other explanatory features that influence the product search relevance scores derived by the customer
- Future enhancements to improve the prediction accuracy of the model is to involve NLP packages like word2vec embeddings, Latent Semantic Analysis for sophisticated semantic and syntactic features engineering.

Appendix

- Scripting Language – Python and R
- Visualizations – Tableau, Power BI, matplotlib and ggplot2 packages
- Packages – numpy, pandas, nltk, sklearn and fuzzywuzzy
- Competition - https://www.kaggle.com/c/home-depot-product-search-relevance
- GitHub – https://github.com/kedarwagh
- LinkedIn – https://www.linkedin.com/in/kedarwagholikar
- Kaggle - https://www.kaggle.com/kdanalytics/kernels