HomeDepot Search Relevancy

OR 568 Final Project

- Divya Aigal

Project Description

- Predict the relevance of search results on homedepot.com
- Search relevancy is an implicit measure to gauge how quickly customers get to the right products
- Develop a model to accurately predict the relevance of search results
- Increase number of iterations their team performs on current search algorithms

Reference: https://www.kaggle.com/c/home-depot-product-search-relevance

Data Sources

- Data Source - https://www.kaggle.com/c/home-depot-product-search-relevance/data

Data files descriptions:

- **attributes.csv** – product_uid, name, value
- **product_description.csv** - product_uid, description
- **train.csv** – id, product_uid, product_title, search_term, relevance – **74067 instances**
- **test.csv** – id, product_uid, product_title, search_term – **166693 instances –** (Not using this test set as there are no values for response variable)

Data Format

- Training data:

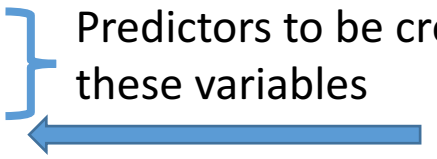
id	product_uid	product_title	search_term	relevance
2	100001	Simpson Strong-Tie 12-Gauge Angle	angle bracket	3.00
3	100001	Simpson Strong-Tie 12-Gauge Angle	l bracket	2.50
9	100002	BEHR Premium Textured DeckOver 1-gal. #SC-141 T...	deck over	3.00
16	100005	Delta Vero 1-Handle Shower Only Faucet Trim Kit in C...	rain shower head	2.33
17	100005	Delta Vero 1-Handle Shower Only Faucet Trim Kit in C...	shower only faucet	2.67
18	100006	Whirlpool 1.9 cu. ft. Over the Range Convection Micro...	convection otr	3.00
20	100006	Whirlpool 1.9 cu. ft. Over the Range Convection Micro...	microwave over stove	2.67
21	100006	Whirlpool 1.9 cu. ft. Over the Range Convection Micro...	microwaves	3.00

Product Description file:

product_uid	product_description
100001	Not only do angles make joints stronger, they also pr...
100002	BEHR Premium Textured DECKOVER is an innovative s...
100003	Classic architecture meets contemporary design in th...
100004	The Grape Solar 265-Watt Polycrystalline PV Solar Pan...
100005	Update your bathroom with the Delta Vero Single-Ha...
100006	Achieving delicious results is almost effortless with th...
100007	The Quantum Adjustable 2-Light LED Black Emergenc...
100008	The Teks #10 x 1-1/2 in. Zinc-Plated Steel Washer-H...

- Data types of the variables:

Variables	Data type
id	Integer
product_uid	Integer
product_title	Character
search_term	Character
Product_desc	Character
relevance	Integer

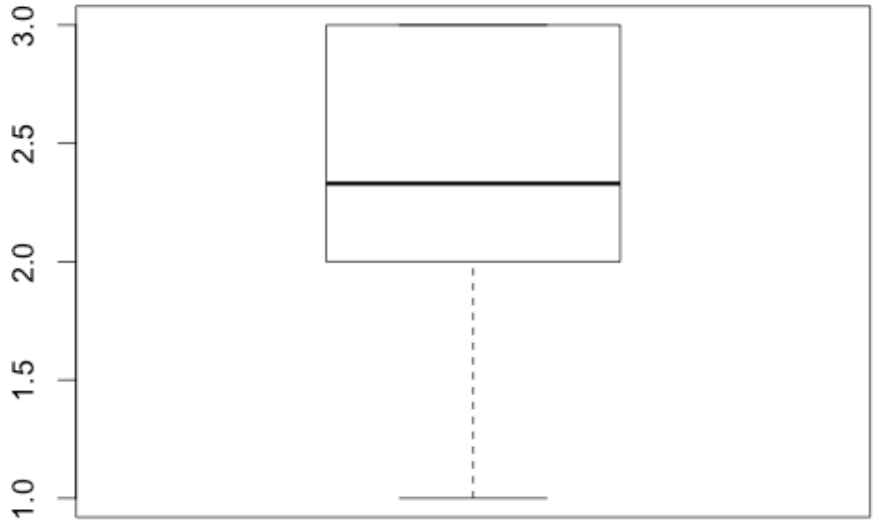


Predictors to be created from these variables

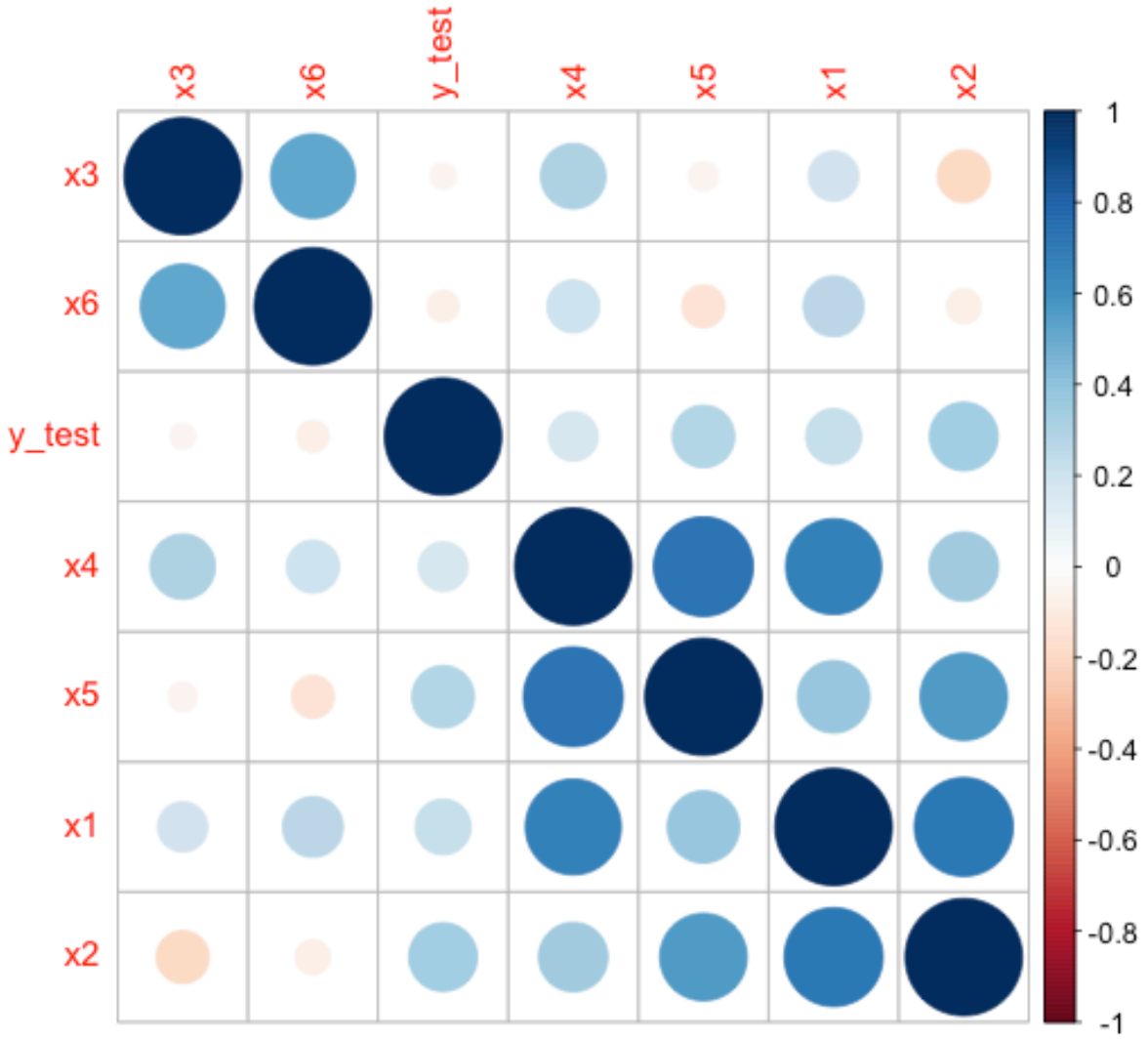
Response Variable

Data Quality

- No missing values in the data
- Figure shows the box-plot for the relevance (output variable) for the training set
- Figure depicts that most of the class labels for the response variable lie in the range from 2.0 to 3.0



Correlation among predictors



- From the correlation plots, we can see that there is a higher correlation between predictors x1, x2, x4 and x5
- Predictors x3 and x6 also show a fair amount of correlation

Approach to this problem

- **Step 1: Create aggregated training set**

  Aggregate product_description.csv and train.csv files to create a larger training file with id, product_uid, product_title, search_term, product_description and relevance as variables

- **Step 2: Data cleaning and predictor creation**

  Relevance is a numeric response variable, hence variables (search term, product description and title) be converted into numeric predictors, as they are in text format

- **Step 3: Divide data into training and test sets**

  Data is divided as, 75% data – Training set and 25% - Test set

Approach to this problem

(… continued from last slide)

- **Step 4: Apply various algorithms and check performance**
  - – Check performance of various models on the training set on basis of two predictors groupings :
    - Group 1 – First 3 predictors
    - Group 2 – All the predictors
  - – Apply transformations (centering and scaling) on the data-sets
  - – Predict and tune model performance on the test set
  - – Note the RMSE and RSquared values for both groups
- **Algorithms applied**:

  Linear Regression, PLS, Ridge, Lasso, SVM, MARS

  RMSE is the value on which the model selection will be based

More on Step 2 – Data Cleaning

- Problem pertains to text mining, hence data cleaning is a very important step

- The objective is to create a word vector from the Character variables

- The Character (Text) variables that require cleaning are:
  - search_term
  - product_title
  - product_description

More on Step 2 – Data Cleaning

(… continued from last slide)

- Data cleaning steps used in this project are:

1. **Convert the text to lowercase** - This makes sure that all the text is in the same case and will eliminate redundant vectors arising due to difference in cases, e.g. water, Water and WATER are same words, but could be mis-judged as different words when computing word vectors.
2. **Break down the variables into vectors** – The character variables need to be broken down into a vector of words which is further utilized for analysis. This is done with help of the **unlist** command in R. The unlist command when used with the string split command of R where space is used as the splitter works to vectorize the character variable.
3. **Remove punctuations–** It is very important to remove punctuations from the vectors, as they do not add in much information and remove them will help the vector to remain simple enough.

More on Step 2 – Data Cleaning

(… continued from last slide)

4. **Stripping off blank spaces** – When vectorizing the character variable, there could have been blank spaces which could also have been vectorized. These need to be stripped off from the word vector as they do not add meaning to the analysis.
5. **Stemming the words** – Stemming is a function of deriving the root word for a particular word. A word can have different forms, e.g. take, taking, taken, took are different forms of the original word take. In this project we have used the porter stemmer, which stems a word to its root word. So, for our example, the root word for take would be “tak”. This helps in minimizing redundancy in the word vector due to the same word repeating itself in different forms.

All these steps help in creating a better word vector and also minimize computational time while using different models.

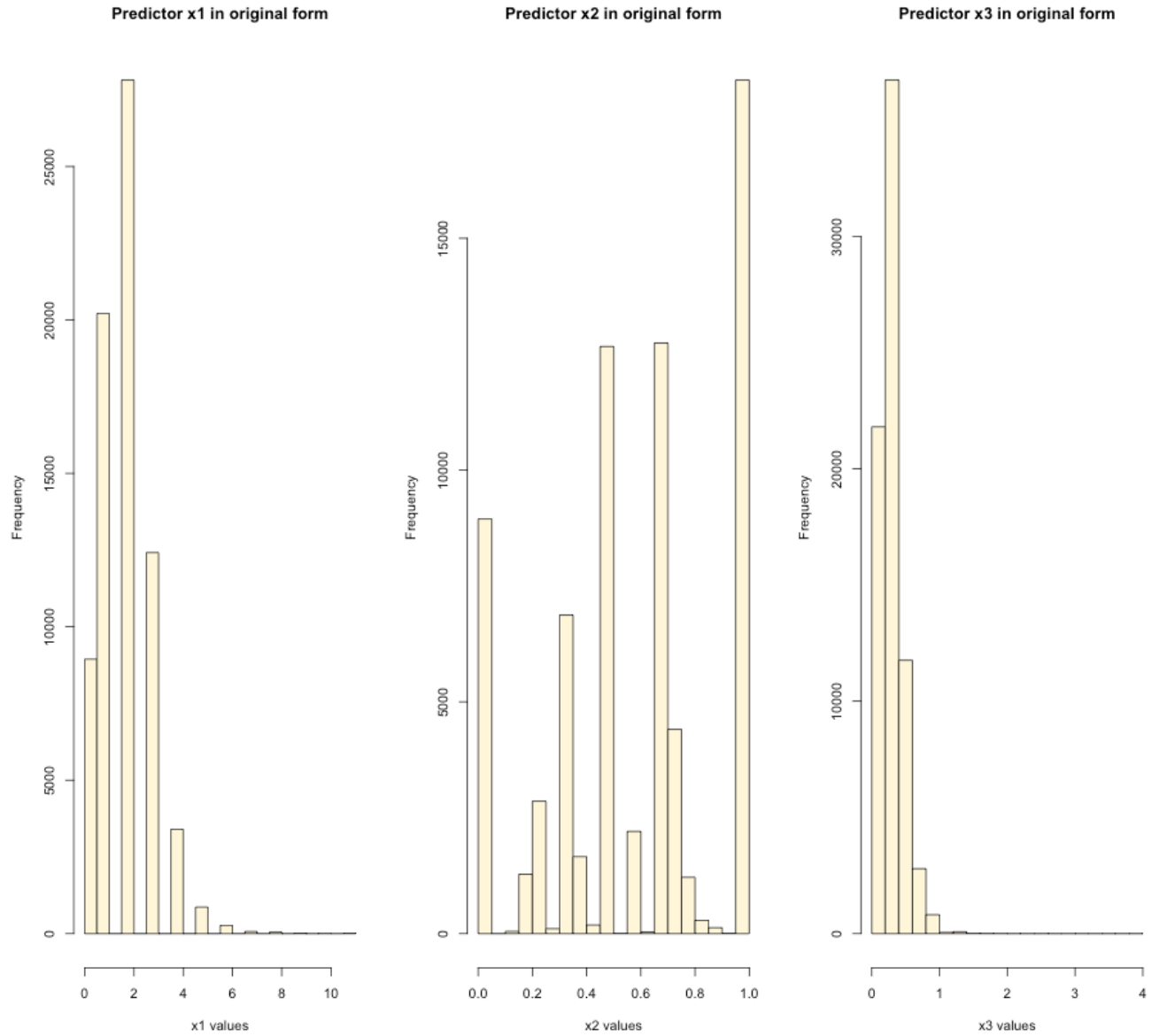
Creating Numeric Predictors:

- *Convert variables "product-title", "product_description" , "search_term" into predictors*
- **Match search_term in product_title**:
  - Number of matched words – 1st predictor,
  - Percentage of matched words in the search term – 2nd predictor,
  - Proportion of matched words from search term in product title – 3rd predictor
- **Match search_term in product_description**:
  - Number of matched words – 4th predictor,
  - Percentage of matched words in the search term – 5th predictor,
  - Proportion of matched words from search term in product description– 6th predictor

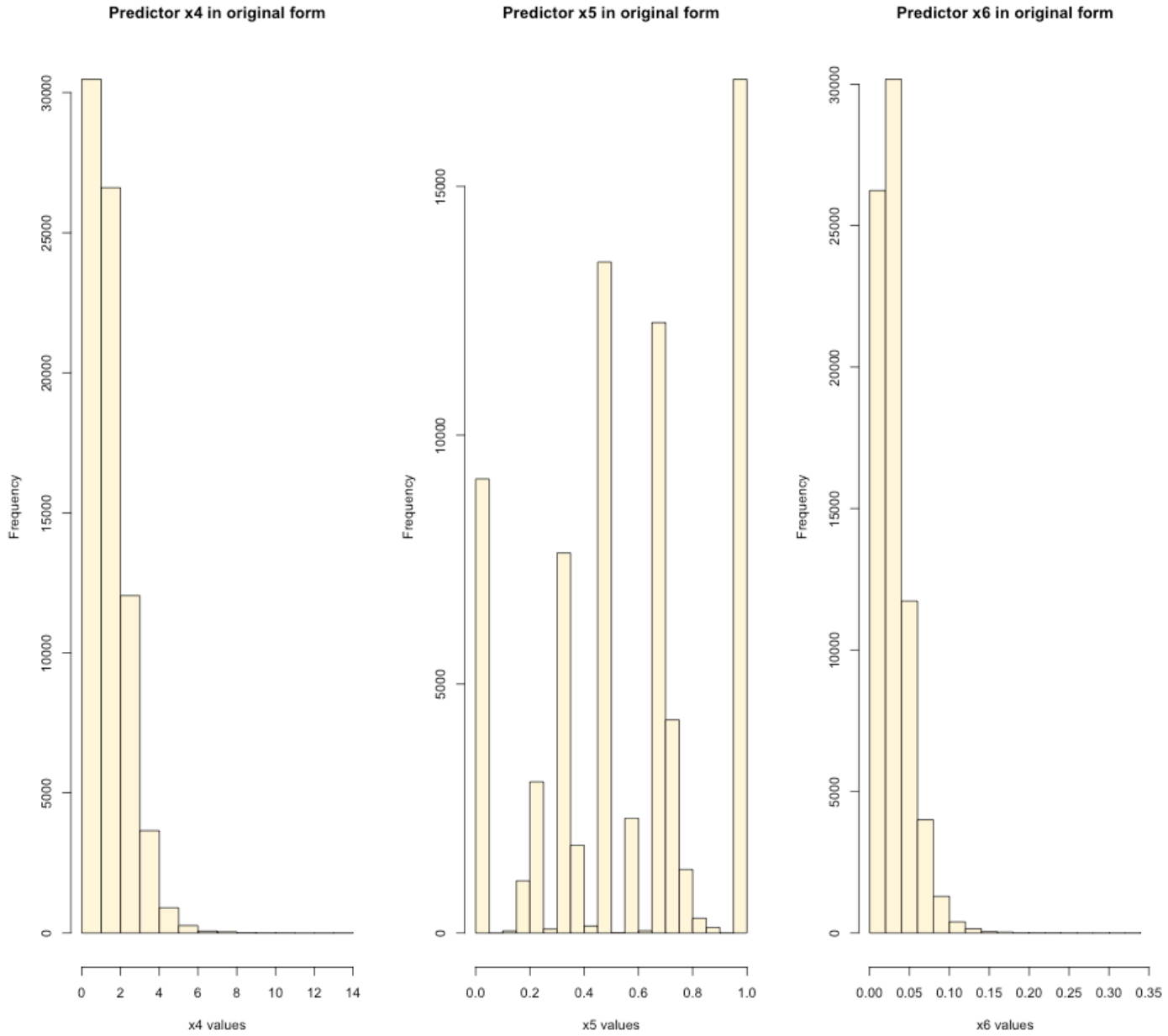
Data set with all predictors and response

	x1	x2	x3	x4	x5	x6	V7
1	1	0.5000000	0.50000000	1	0.5000000	0.015503876	3.00
2	0	0.0000000	0.50000000	0	0.0000000	0.015503876	2.50
3	0	0.0000000	0.18181818	1	0.5000000	0.011976048	3.00
4	1	0.3333333	0.23076923	1	0.3333333	0.028846154	2.33
5	3	1.0000000	0.23076923	3	1.0000000	0.028846154	2.67
6	1	0.5000000	0.13333333	1	0.5000000	0.004089980	3.00
7	2	0.6666667	0.20000000	2	0.6666667	0.006134969	2.67
8	1	1.0000000	0.06666667	1	1.0000000	0.002044990	3.00
9	2	1.0000000	0.22222222	2	1.0000000	0.016666667	2.67
10	2	1.0000000	0.14285714	2	1.0000000	0.024390244	3.00
11	1	0.5000000	0.33333333	2	1.0000000	0.016528926	2.67
12	4	0.8000000	0.29411765	5	1.0000000	0.021645022	3.00
13	1	0.5000000	0.11764706	1	0.5000000	0.008658009	3.00
14	1	0.5000000	0.11764706	1	0.5000000	0.008658009	2.00
15	3	0.5000000	0.40000000	3	0.5000000	0.015625000	2.67

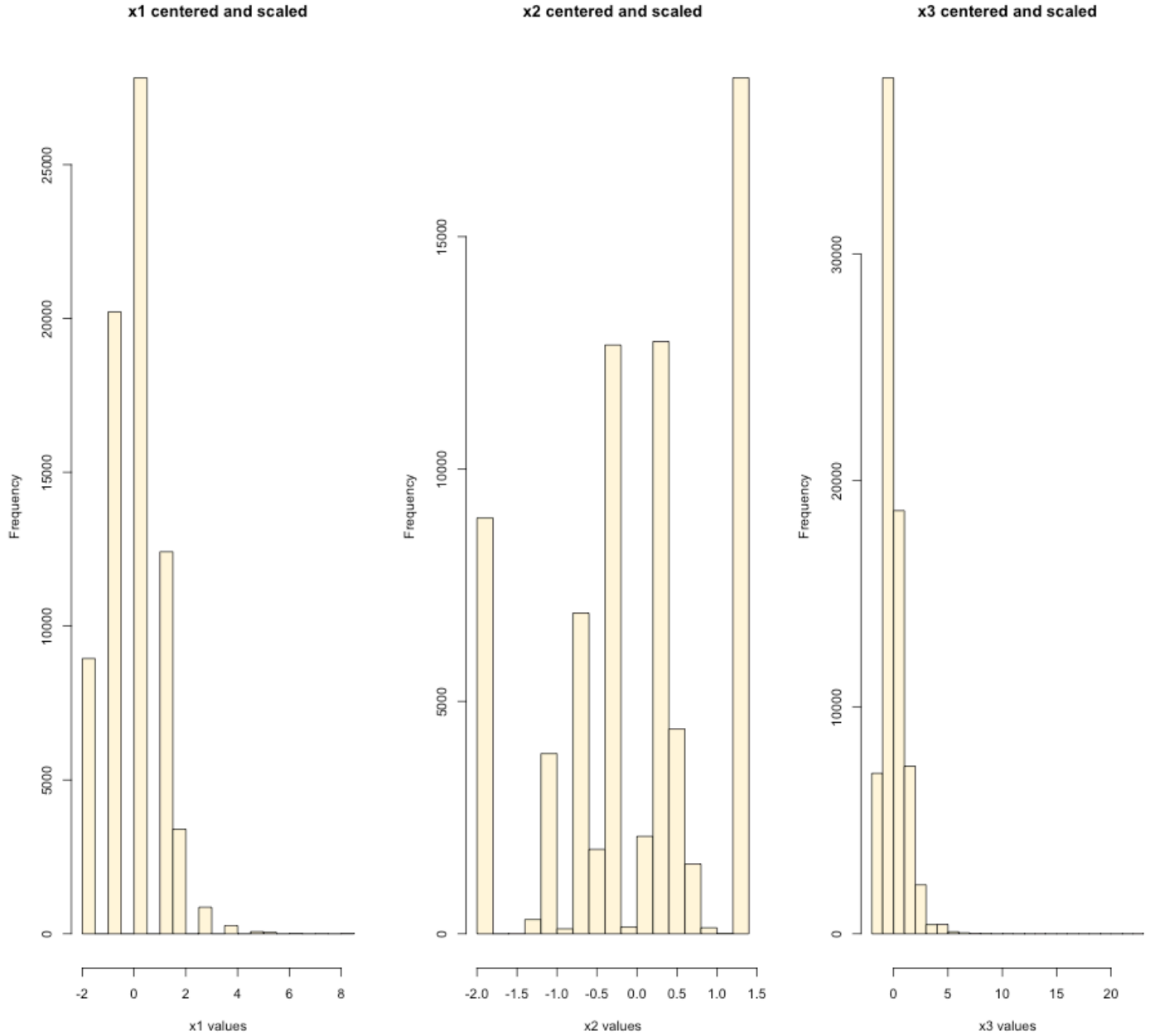
Histogram for original predictors 1, 2 and 3



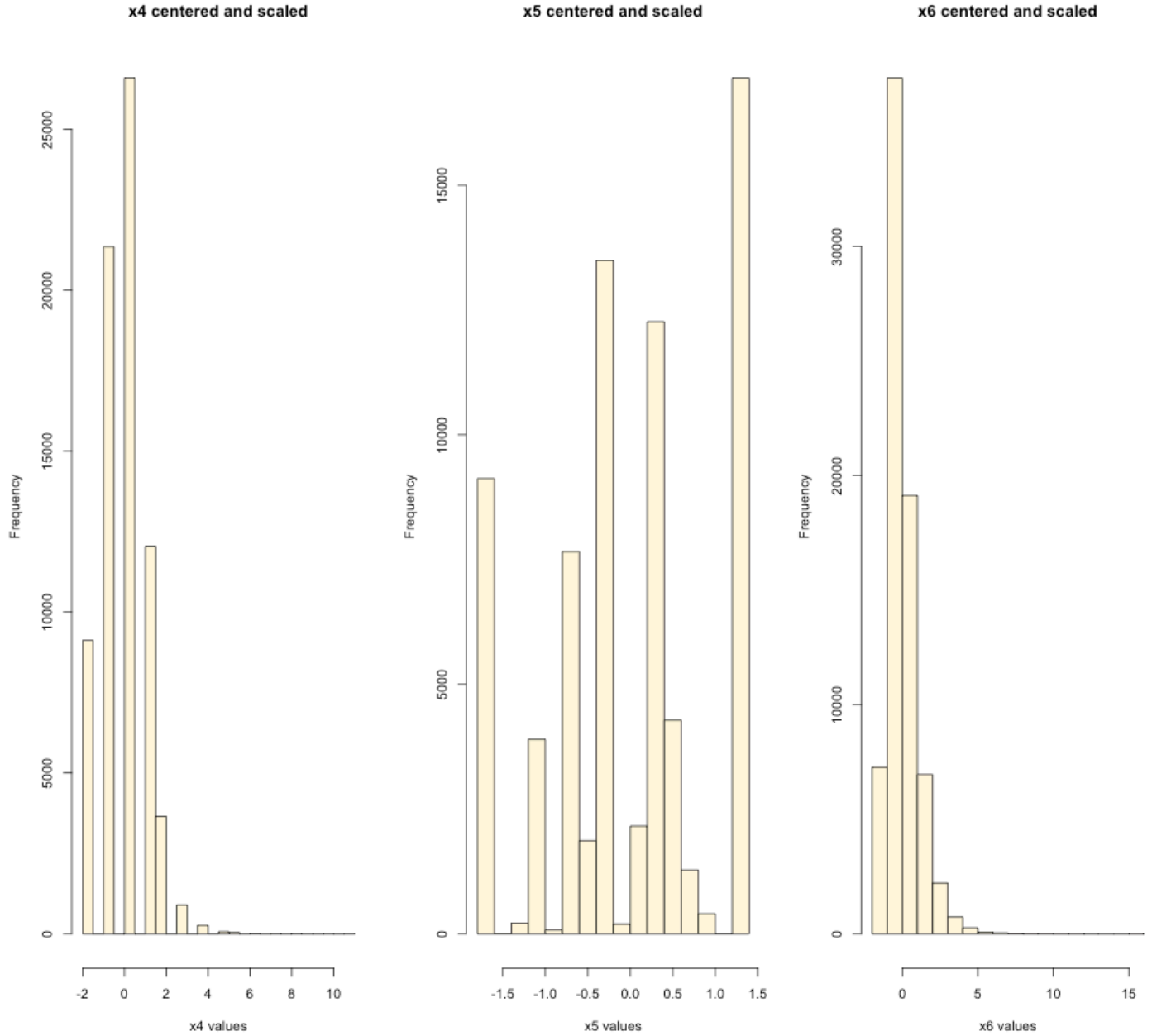
Histogram for original predictors 4, 5 and 6



Pre-processed & Transformed predictors 1, 2 and 3



Pre-processed & Transformed predictors 4, 5 and 6



Pre-processed data set with response variable

	x1	x2	x3	x4	x5	x6	V7
1	-0.7060854	-0.26103185	1.15002961	-0.6843346	-0.2249272	-0.75154123	3.00
2	-1.5822099	-1.81717023	1.15002961	-1.5535956	-1.7955210	-0.75154123	2.50
3	-1.5822099	-1.81717023	-0.76509697	-0.6843346	-0.2249272	-0.92849840	3.00
4	-0.7060854	-0.77974464	-0.47046211	-0.6843346	-0.7484585	-0.08228759	2.33
5	1.0461637	1.29510654	-0.47046211	1.0541874	1.3456667	-0.08228759	2.67
6	-0.7060854	-0.26103185	-1.05692579	-0.6843346	-0.2249272	-1.32406654	3.00
7	0.1700392	0.25768095	-0.65566117	0.1849264	0.2986041	-1.22148909	2.67
8	-0.7060854	1.29510654	-1.45819040	-0.6843346	1.3456667	-1.42664399	3.00
9	0.1700392	1.29510654	-0.52190630	0.1849264	1.3456667	-0.69321522	2.67
10	0.1700392	1.29510654	-0.99960227	0.1849264	1.3456667	-0.30579770	3.00
11	-0.7060854	-0.26103185	0.14686807	0.1849264	1.3456667	-0.70012436	2.67
12	1.9222883	0.67265118	-0.08916994	2.7927094	1.3456667	-0.44349907	3.00
13	-0.7060854	-0.26103185	-1.15134099	-0.6843346	-0.2249272	-1.09493249	3.00
14	-0.7060854	-0.26103185	-1.15134099	-0.6843346	-0.2249272	-1.09493249	2.00
15	1.0461637	-0.26103185	0.54813268	1.0541874	-0.2249272	-0.74546561	2.67

Linear Regression

- The linear regression model will help to predict how the relevance is based on the predictors and how to predict the relevance given the predictors
- We apply centering and scaling to all the predictor
- RepeatedCV with 5 repeats is used for the resampling
- Linear regression model does not have any tuning parameter

Performance with 3 and 6 predictors is shown below:

	3 predictors					6 predictors			
	RMSE	**Rsquared**	**RMSE SD**	**Rsquared SD**		**RMSE**	**Rsquared**	**RMSE SD**	**Rsquared SD**
Training set	0.5335015	0.000363464	0.003881727	0.000502107		0.4964688	0.1341702	0.00401085	0.007647651
Test set	0.535296492	0.000232057	-	-		0.4991941	0.1305645	-	-

Best model

PLS

- The PLS model combines the predictors to form components which would best summarize the predictors so as to maximize correlation of the predictor with the response
- Centering and scaling was applied to all the predictors
- 10 fold cross validation is performed on the model
- The tuning parameter number of components (ncomp) is 2 for both set of predictors

Performance with 3 and 6 predictors is shown below:

	3 predictors				6 predictors			
	RMSE	**Rsquared**	**RMSE SD**	**Rsquared SD**	**RMSE**	**Rsquared**	**RMSE SD**	**Rsquared SD**
Training set	0.5335082	0.000322717	0.003553881	0.000291357	0.4986622	0.1266111	0.00440339	0.01009714
Test set	0.535253343	0.000464307	-	-	0.5020126	0.120716	-	-

Best model

Ridge Regression

- The Ridge Regression model shrinks the coefficients of the correlated predictors so that it gives equal weightage to the correlated predictors.
- For this project, the predictors were centered and scaled.
- 10 fold cross validation is performed on the model
- The tuning parameter lambda had values 0.075 and 0 with 3 & 6 predictors respectively.

Performance with 3 and 6 predictors is shown below:

	3 predictors (lambda = 0.075)					6 predictors (lambda = 0)			
	RMSE	**Rsquared**	**RMSE SD**	**Rsquared SD**		**RMSE**	**Rsquared**	**RMSE SD**	**Rsquared SD**
Training set	0.5335082	0.000322717	0.003553881	0.000291357		0.4986622	0.1266111	0.00440339	0.01009714
Test set	0.535253343	0.000464307	-	-		0.5020126	0.120716	-	-

Best model

Lasso Regression

- In case of many correlated predictors, the Lasso model chooses only one of the correlated predictors.
- The predictors are centered and scaled.
- 10 fold cross validation is performed on the model
- The tuning parameter fraction = 0.2535714, lambda = 0 for 3 predictors & fraction = 1, lambda = 0 for 6 predictors.

Performance with 3 and 6 predictors is shown below:

	3 predictors (fraction = 0.2535714 and lambda = 0)				**6 predictors (fraction = 1 and lambda = 0)**			
	RMSE	**Rsquared**	**RMSE SD**	**Rsquared SD**	**RMSE**	**Rsquared**	**RMSE SD**	**Rsquared SD**
Training set	0.5335043	0.000312325	0.003627841	0.000283273	0.4964634	0.1342102	0.005447756	0.008621013
Test set	0.535298578	0.000716373	-	-	0.4991941	0.1305645	-	-

Best model

SVM

- The Support Vector Machine (SVM) model is a non-linear model that improves the RMSE values by adding in a penalty factor.
- For this project, the SVM model uses centering and scaling of the predictors.
- RepeatedCV with 5 repeats is used for the resampling
- The SVMLinear model was used here with tuneLength = 5 as the computation time was huge with the radial kernal.

Performance with 3 and 6 predictors is shown below:

	3 predictors					6 predictors			
	RMSE	**Rsquared**	**RMSE SD**	**Rsquared SD**		**RMSE**	**Rsquared**	**RMSE SD**	**Rsquared SD**
Training set	0.5338877	8.92E-05	0.003725288	9.53E-05		0.505021	0.1330588	0.003619638	0.00769219
Test set	0.534230551	0.001099311				0.5088778	0.1296484		

Best model

MARS

- The MARS model creates a piece-wise linear model by dividing a predictor into two.
- Centering and scaling is applied on all predictors
- 10 fold cross validation was repeated 5 times on the model
- The tuning parameters for MARS is the degree = 1 and the number of features to be pruned = 5 for 6 predictors (for 3 predictors, even after running model several times and tuning model parameters, Rsquared had NaN values)

Performance with 3 and 6 predictors is shown below:

	3 predictors (degree =1, nprune = 5)				6 predictors (degree =1, nprune = 5)			
	RMSE	**Rsquared**	**RMSE SD**	**Rsquared SD**	**RMSE**	**Rsquared**	**RMSE SD**	**Rsquared SD**
Training set	0.5335112	NaN	0.003889081	NA	0.4951717	0.1387348	0.004214384	0.008591895
Test set	0.5353576	NA	-	-	0.4977136	0.1357187	-	-

Best model

Performance of all algorithms

	Training set					Test set	
	RMSE	**Rsquared**	**RMSE SD**	**Rsquared SD**		**RMSE**	**Rsquared**
Linear Regression	0.4964688	0.1341702	0.00401085	0.007647651		0.4991941	0.1305645
PLS	0.4986622	0.1266111	0.00440339	0.01009714		0.5020126	0.120716
Ridge	0.4964928	0.1341315	0.004156315	0.008971832		0.4991941	0.1305645
Lasso	0.4964634	0.1342102	0.005447756	0.008621013		0.4991941	0.1305645
SVM	0.505021	0.1330588	0.003619638	0.00769219		0.5088778	0.1296484
MARS	0.4951717	0.1387348	0.004214384	0.008591895		0.4977136	0.1357187

Variable Importance for all models

Linear Regression	Overall	PLS	Overall	Lasso/Ridge	Overall	MARS	Overall
x2	100	x2	100	x2	100	x2	100
x5	56.914	x5	80.85	x5	72.165	x5	51.7
x6	43.849	x1	58.43	x1	41.451	x3	0
x3	35.793	x4	44.98	x4	26.009	x6	0
x4	7.239	x6	10.18	x6	3.031	x1	0
x1	0	x3	0	x3	0	x4	0

Predictors x2 and x5 are the best predictors as seen from the variable importance ranking for different models (not considered SVM in this case)

Interpretations and Conclusions

- The data set used for this project contained correlated predictors.
- Goal was to find how the model performs for the test set. Regression models perform better in our case, which was visible from the results.
- Results were based on least RMSE values for the models.
- From among all the models tested, **MARS** performs the best with degree = 1 and nprune value = 5 giving **0.4977136** as RMSE value.
- All models give x2 and x5 as the best predictors

References

- https://www.kaggle.com/c/home-depot-product-search-relevance

- Lecture notes from Dr. Xu