Analyzing the Feature Importance of Different Variables on the Price of Ikea Products

Philip Krück, Johannes Pein

Hamburg School of Business Administration

B.Sc. Business Informatics (18A-BI)

Digital Toolbox: Data Business

Lecturer: Ulf Köther

Group Number: 7

Matriculation Numbers: 3938 (P.Krück), 4001 (J.Pein)

04.12.2020

Contents

List of Abbreviations

R . . . . Statistical Programming Language.

1
Introduction

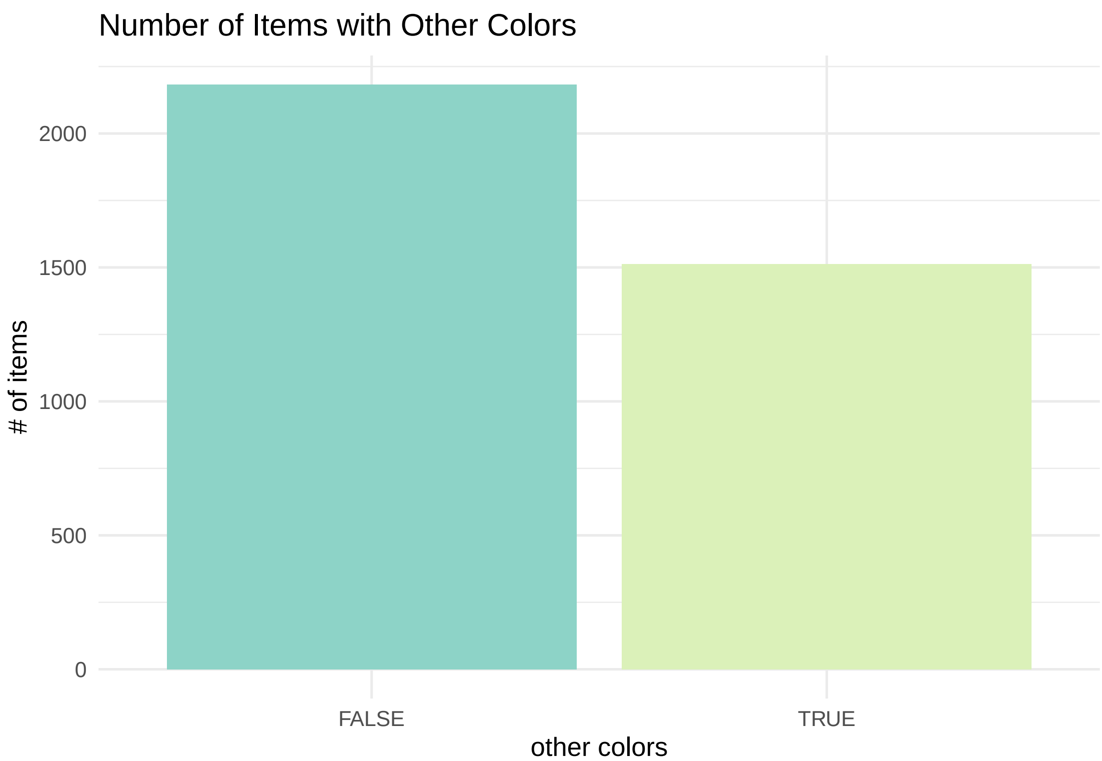

Figure 1.1: Sample plot

2
Theoretical Background & Research Question

2.1 Theoretical Background

2.1.1 Data Set

The data set was obtained by a kaggle.com user (Reem Abdulrahman) by the means of webscraping techniques from the Saudi Arabian Ikea website in the furniture category on the 20th of April 2020. Noteworthy features include the name, category, price in Saudi Riyals, the designer and dimensions (width, height and depth). The data set has 13 variables and 2962 observations.

2.1.2 RF Basics

In order to analyze the feature importance in relation to the price variable, a random forest regression model was chosen. A random forest consists of many decision trees, which make a prediction based on a majority decision process. In standard decision trees, each nodes is split to achieve the best performing model. In random forests however, the nodes are randomly split. Compared to linear regression models, the random forests model not only takes into account the mean and covariance structure of response, but also deeper deeper aspects of data[1] leading to a more advanced and robust model.

2.1.3 Feature Importance

- feature and predictor variable can be used interchangibly

2.2 Research Question

This paper explores the influence for different variables on the price in the given data set. The motivating forces for this research question are the possible implications for price determination of new items.

[1] Grömping, "Variable Importance Assessment in Regression: Linear Regression versus Random Forest."

3
Methods

3.1 Data Cleaning and Transformatoin

To examine our data set properly, we first had to restructure and reformat it. This initial data cleaning step included type conversions, value mutation, addition of new calculated fields and the dropping of irrelevant columns. Concretely, we converted name, category and designer to categorical variables. In the designer column, we converted blank strings and values prefixed by "IKEA of Sweden" to missing values (`NA`). Furthermore, we converted both the price and old price to double values and changed the currency from Saudi Arabian Riyals to Euros based on the exchange rate from the time the data set was obtained by the author @ref(#theoretical_background). To better facilitate the comparison of the different sizes of furniture items, we calcuted the size in cubic meters based on the depth, width and height values. Finally, we selected only columns that could have a potential impact in our analysis (see Table 3.1 and 3.2).

tidy category -> duplicate filtering - there were some observations with the same item id. All other value were the same in these instances except for the category. - The authors considered multiple approaches to handle these data duplications - Option one -> mutate duplicates into one combination of categories (e.g. a and b to a, b), then the newly created categories would have a low count of observations

Table 3.1: Initial Data Set formatting.

X1	item_id	name	category	price	old_price	sellabl
0	90420332	FREKVENS	Bar furniture	2650	No old price	TRUE
1	368814	NORDVIKEN	Bar furniture	9950	No old price	FALSE
2	9333523	NORDVIKEN / NORDVIKEN	Bar furniture	20950	No old price	FALSE
3	80155205	STIG	Bar furniture	690	No old price	TRUE
4	30180504	NORBERG	Bar furniture	2250	No old price	TRUE
5	10122647	INGOLF	Bar furniture	3450	No old price	TRUE

Table 3.2: Data Set after cleaning process.

name	category	price_eur	old_price_eur	sellable_online
FREKVENS	Bar furniture	65.02	NA	TRUE
NORDVIKEN	Bar furniture	244.14	NA	FALSE
NORDVIKEN / NORDVIKEN	Bar furniture	514.05	NA	FALSE
STIG	Bar furniture	16.93	NA	TRUE
NORBERG	Bar furniture	55.21	NA	TRUE
INGOLF	Bar furniture	84.65	NA	TRUE

while the single categories would have a decreased count. We would then have an increased count in categories. Because the count of the individual categories is low, no definite conclusion can be drawn for the interaction on the price - Option two -> create multiple variables for each observed category for the same item id. The authors wanted to analyze the feature importance of category variable as such and not the multiple occurrences of the category variable - The authors chose option of selecting the observations where the category count occurred most frequent.

TODO: Format these tables

3.2 8 Step EDA (nice heading)

[1]. - Our EDA analysis is based on the protocol for data exploration proposed by ZUUR paper - This 8 step analysis is especially useful for regression models

3.2.1 Step 1: Outliers in Price and Independent Variables

- Assumption: outliers are not by chance or random (no observer error)

[1] Zuur, “A protocol for data exploration to avoid common statistical problems,” 33–35.

- Assumption seems to hold up when looking at individual outlier observations. These are congruent in themselves. Assume: web scraping technique is highly unlikely to have measurement error ocurrence
- Use outliers in model

3.2.2 Step 2: Homogeneity of Price

- To test the assumption of homogeneity of variance for price (homoscedasticity) by the means of conditional boxplotting
- As can be seen, price is homogoneous for all category variable except for beds where the variance differes widely (see plot xxx)
- price is heterogenous for all individual names by category (see plot xxx)
- Looking at both categories simultaneously -> homogenous
- Im Rahmen der Arbeit nicht möglich alle kategorischen Variablen Kombinationen anzuschauen. 5 kategorische Variablen -> 2ˆ5 = 32 Möglichkeiten

3.2.3 Step 3: Missing Value Trouble

- not normal verteilt
- ikea ist im Niedrigpreissegment angesiedelt
- see figure . . . .
- exponential decay in price, old price and size (eˆ-x)

3.2.4 Step 4: Zeros

- many missing values from old price -> assume those were not on sale
- size_m3 missing values because of calculation formula
- designer -> removed values containing digits (were clearly falsely scraped)

3.2.5 Step 5: Collinearity between Independent Variables

- high collinearity between old price and price
- relatively high collinearity between price and size
- as can be seen in table

3.2.6 Step 6: Relationship between Independent Variables and Price

- from eda_covariance.Rc (in Anhang + verweisen)
- strong relationship b/w price + old price & b/w price + size (see)
- other relationships aren't strong

3.2.7 Step 7: Interactions

- Coplotting designer and name works, while the two combination would not plot
- The linear model predicted infinite values and thus coplot the values properly for the other two options
- However, dropping all NA values and thus reducing the total data size to 354 observations would case for the combination coplot name and category while name + designer combination would not work
- The authors hypothesized that infinite values were caused by a division of zeros of the linear model since there occured 0 values in size
- This however proved to be wrong after applying respective filters
- The following interaction could be analyzed
- - There is probably no significant interaction between size, price, name & designer as can be seen in coplot -> lines are nearly parallel
- Based on the coplot of category and name with the 354 observations, inparallelity could be observed and thus could conclude a interaction. However, this could also be due to the small sample size
-

(footnote): the authors would highly appreciate any solutions on this matter[2]

[2] This is a footnote.

3.2.8 Step 8: Independence of Price

- Durch das tidying in (cross reference) duplicate removal -> Zuur paper Step 8 1. citation: meaning that information from any one observation should not provide information on another after the effects of other variables have been accounted for. This concept is best explained with examples.

3.3 Random Forest Regression Model

This analysis was conducted using the R randomForest package, which is based on the original Breiman and Cutler's Fortran code for random forest regression. To learn more about how random forests work or the randomForest package, see Liaw and Wiener[3] or #chapter2 #TODO. To reproduce the analysis conducted in this paper, the prepatory steps are described now. The following steps are based on an already cleaned ikea data frame which is described in 3.1. This data frame is then transformed further to be used with the randomForest package. First, the variable `old_price_eur` is removed from the data frame, due to a very high correlation and relationship to the price variable analyzed in 3.2.5 and 3.2.6 Then, the designers and names, which are not part of the 50 `designers` and 49 `names` with the highest number of occurences, are grouped in the `other` value. This is because the `randomForest` method does not allow categorical variables with more than 53 predictors. The last step is dealing with the missing values in the data. As described in 3.2.4, there are missing values in the `size_m3` and `designer` variables. To use the `randomForest` method of the randomForest package on the data, those missing values are dealt with using three different approaches. In the first approach the rows with missing values are deleted, reducing the total number of rows by aproximately 50%. In the second approach the missing values are dummy coded with a value of -1000. The third approach uses the `na.roughfix = na.omit` argument, which is the built in way of the randomForest package to deal with missing values. After

[3] "Classification and Regression by randomForest."

preparing the data, the `randomForest` method of the randomForest package is applied to the data with number of trees set to 2000 and importance set to `TRUE`.

```
randomForest(price_eur ~ ., rf_ikea, ntree=2000, keep.forest=FALSE,
importance=TRUE)
```

Then the `importance` method of the randomForest package is used to save the feature importances, which are computed by permuting feature importance, which is described in #chapter2 #TODO. The three different approaches of dealing with the missing values in the data set lead to different results, so the authors chose to calculate the mean result of the three approaches. The result of this analysis is presented in the following chapter.

- TODO: is normal distribution relevant for random forest (step 3) -> see article or cite something

4
Results

by J. Pein

4.1 TODO: set n_trees to 2000 before handing in

In this chapter, the results of the analysis of the feature importance of different predictor variables (features) on the response variable price of Ikea products are presented.

As described in 3.3, the feature importance was calculated using permuting feature importance of the randomForest R package. In this analysis, feature importance is derived from the percentage increase of the mean squared error (MSE) of the overall random forest regression model with the response variable `price_eur`. When the percentage increase of the MSE is higher, the feature is more important, accordingly, when the percentage increase of the MSE is lower, the feature is less important.

Thus, as can be seen in the figure above, the most important feature is `size_m3` with an increase of the MSE of 182%. The second, third and fourth most important features are `designer` with an increase of 120%, `name` with an increase of 114% and `category` with an increase of 105%. The fifth most important feature is `other_colors` with an increase of the MSE of 78% and the least important feature is `sellable_online` with a 9% increase.

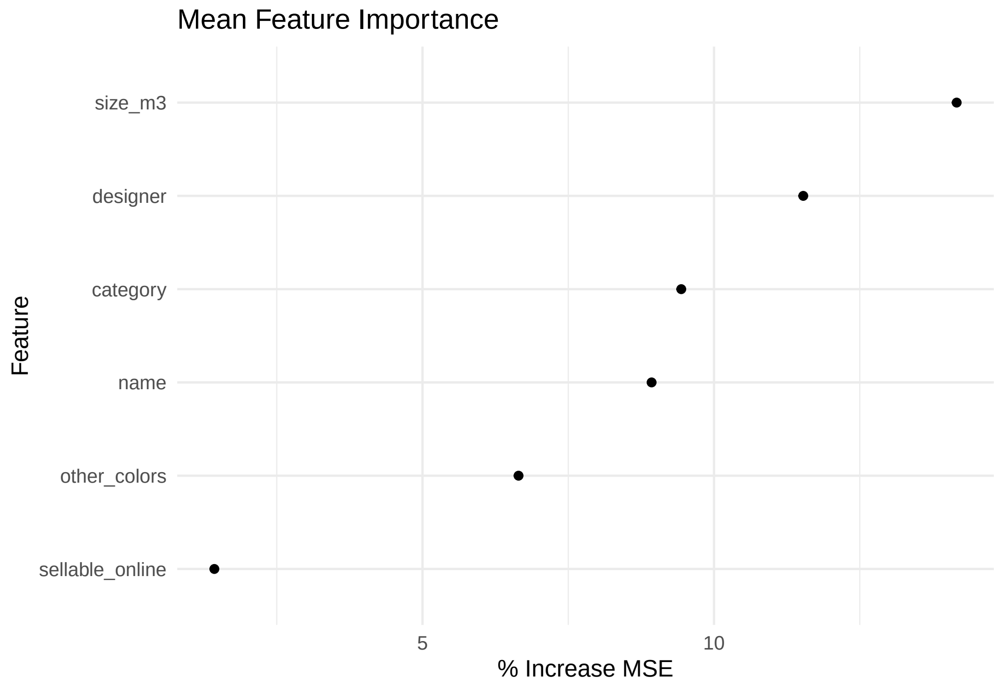

Figure 4.1: A nice image.

These results are further discussed in the following section.

5 Discussion

5.1 size_m3

- material cost
- big items more expensive than small ones
- high correlation to price

5.2 designer

- designers with high number of products produce products in wide price range : plot reference -> appendix

- many different combinations of designer (49-53)

- many designer-combinations with low number of products:

  - n combinations with occurrences<5

-> overfitting: cite scholarly article -> low generalization : model might perfom bad on other data

5.3 name

- overfitting

5.4 category

- beds are more expensive than chairs : some categories are more expensive than others (show plot)
- but, price range varies heavily in category (show plot)

5.5 other_colors

- products of every price range have both options : low importance
- mean price is higher if other_colors is true (show plot covariance price+other_colors)
- interquartile-range is smaller if other_colors is false

5.6 sellable_online

sellable true: price range too high sellable ntrue: very rare

5.7 Conclusion & Ausblick

general results might contain bias or variance errors, further investigating

- Further research:
  - analyze designer overfitting, recommend using overfitting techniques (e.g. . . .)
  - analyze same research question with other techniques (lm, name more) and compare
  - data scraping from other country-pages

6
Individual Statements

6.1 Philip Krück

6.2 Johannes Pein

Appendices

A
Appendix

A.1 Plot abc

A.2 Plot xyz

B
Bibliography

Grömping, Ulrike. "Variable Importance Assessment in Regression: Linear Regression versus Random Forest." *The American Statistician*, nos. Vol. 63, No. 4 (2009): 308–19. `https://prof.beuth-hochschule.de/fileadmin/prof/groemp/downloads/tast_2E2009_2E08199.pdf`.

Liaw, Andy, and Matthew Wiener. "Classification and Regression by randomForest." *R News*, nos. Vol. 2/3, December 2002 (2001). `https://www.researchgate.net/publication/228451484_Classification_and_Regression_by_RandomForest`.

Zuur, E.N. Ieno lain. "A protocol for data exploration to avoid common statistical problems." *British Ecological Society*, 2010. doi:10.1145/1738826.1738829.