Comparison of Different Regression Models to Predict the Sales Price of IKEA Furniture

Introduction

The goal of my project is to try and compare different regression methods to predict the sales prices of IKEA furniture. While this exact goal might be more interesting from an academic point of view, a big warehouse company might apply the same techniques in combination with different predictors e.g. to predict optimal sales prices of the same item to different customers. These methods can therefore extract a lot of economic potential from the right set of data.

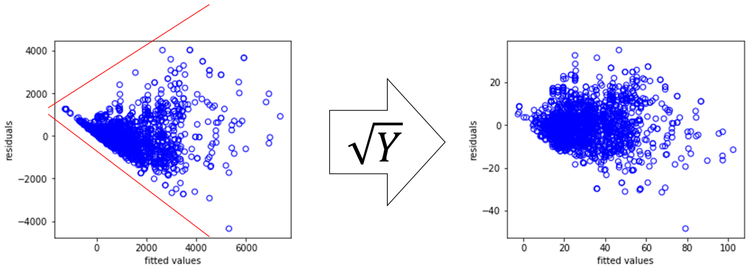

Figure 1. Residual values plotted over the predicted response

Starting with quantitative predictors only

Besides the sales price the dataset contains 3 quantitative variables (height, width, depth) and a few more qualitative variables for every item. A first model with those 3 quantitative predictors only yielded already good results, however the data showed a problem with heteroscedacity (Figure 1). To get rid of this problem I transformed the response to $\sqrt{Y}$, which also increased R^2 from around 0.71 to 0.76.

Adding qualitative variables to the model

I then went on and introduced dummy variables for the category the furniture belongs to (beds, bookshelves…). The original dataset contained 17 categories, which would mean to introduce 17 dummy variables. I used Forward step selection to find only the most relevant ones. The results were compared by Validation Model selection (Figure 2) and other indicators for the Test MSE like AIC and BIC.

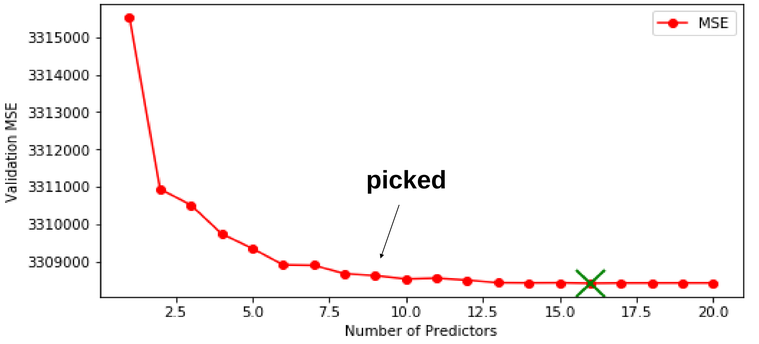

Figure 2. Estimated Test MSE by Validation Model Selection

Looking at the distribution of predictors and adding interaction terms

Studying the distribution of the quantitative predictors height, width and depth with respect to the furniture category (Figure 3) one quickly notices that sofas have an unusual high depth (indicated by the size of the data points) compared to the other categories. This indicates potential to better model the final sales price of a sofa by adding an interaction term. Trying different combinations of interaction terms I found out that the model gets better when having a negative interaction term "sofas*depth" in the model. This makes sense because their large depth leads to an overestimation of the price of sofas – an effect that is partly compensated by the added interaction term.

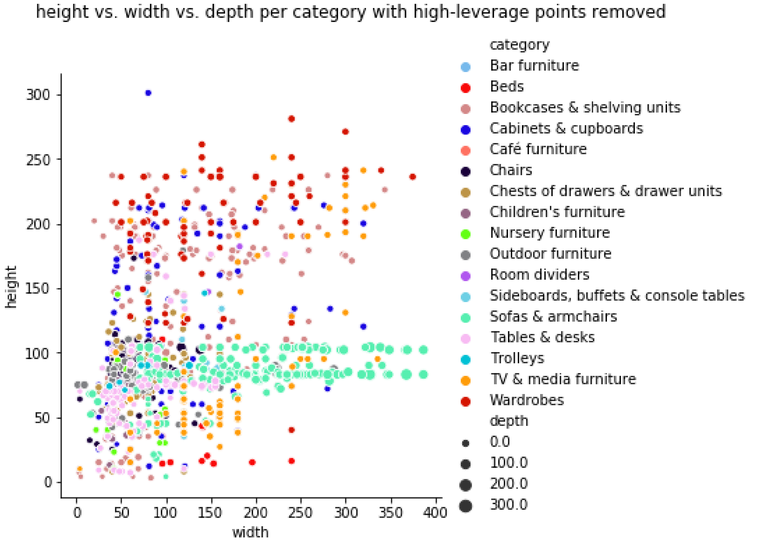

Figure 3. Distribution of predictors in the different furniture categories

Conclusions and Comments

I also tried a General Additive Model (GAM) with this dataset to get more flexibility in my prediction. Looking at the results, however, I don't feel like giving up the linear assumption is worth the additional complexity for this dataset.

In the end I got a linear model with 12 predictors and an R^2 of around 0.83 that I feel predicts the price of IKEA furniture quite well, without being overfitted to this specific dataset.

To optimize the model even further I would suggest to take another look at the interaction terms and run a best subset selection including all promising interaction terms.