House Prices: Advanced Regression Techniques

MAP 553-Regression

Aksel Kaastrup Rasmussen

20/12/2019

1 Introduction.

We are given a dataset which consists of information of a large number of residential homes in Ames, Iowa. Our goal is not only to estimate the sale price of the homes as accurately as possible, but also to gain information about the given variables; what really gets people's wallet out of their pocket when considering buying a house? We will face the challenge of dimensionality - we have a large dataset with lots of variables, and we will face the problem of choosing a good model for which the assumptions of the data is accurate. The problem here is rather how to make the data satisfy the assumptions without compromising the information. Our research hypothesis is that we with a linear model and data modelling can describe the sale price of the homes reasonably well. Furthermore we believe we will find that more advanced models will perform better, although simple models are easier to interpret.

2 Exploratory Data Analysis/Inital Modeling.

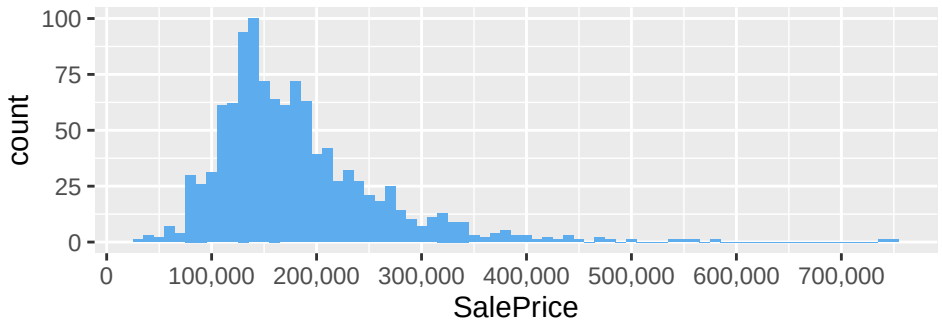
Our traning and test dataset consists of 1095 and 365 observations respectively with both 67 features. It is clear both cleaning and basic handling of missing values has already taken place, and we will not explore such techniques further here. There has already been taking care of a great deal of ordinal variables using ordinal encoding and normalizing. We will continue this work.

The dataset consists of both numerical, nominal and ordinal variables, and right from the start we will label encode ordinal variables into numbers. This arguably holds for *Street* and *PavedDrive*, since the levels have an order. With this ordinal encoding we ensure the variables keeps the ordered structure. In addition there are lots of categorical *quality* variables, which indeed are ordinal.

We have chosen to remove the variable *Utilities*, since this variable is constant for all training observations, and finally we have split the data into categorical and numerical variables with dimension 36 and 30 respectively. As a start of the exploratory analysis, let us look into the response variable.

2.1 The response variable

We notice the response variable *SalePrice* is right skewed. This is due to the fact that more people can afford a relatively cheap house compared to an expensive house.



This skewness will justify a Box-Cox transform in the early modelling phase in order to satisfy the assumption that the variables are normally distributed. Let us now take a look into the response variables relationship with the numerical variables. Consider figure 1, where we have compared all numerical variables with correlation coefficient above $\frac{1}{2}$ with *SalePrice*.

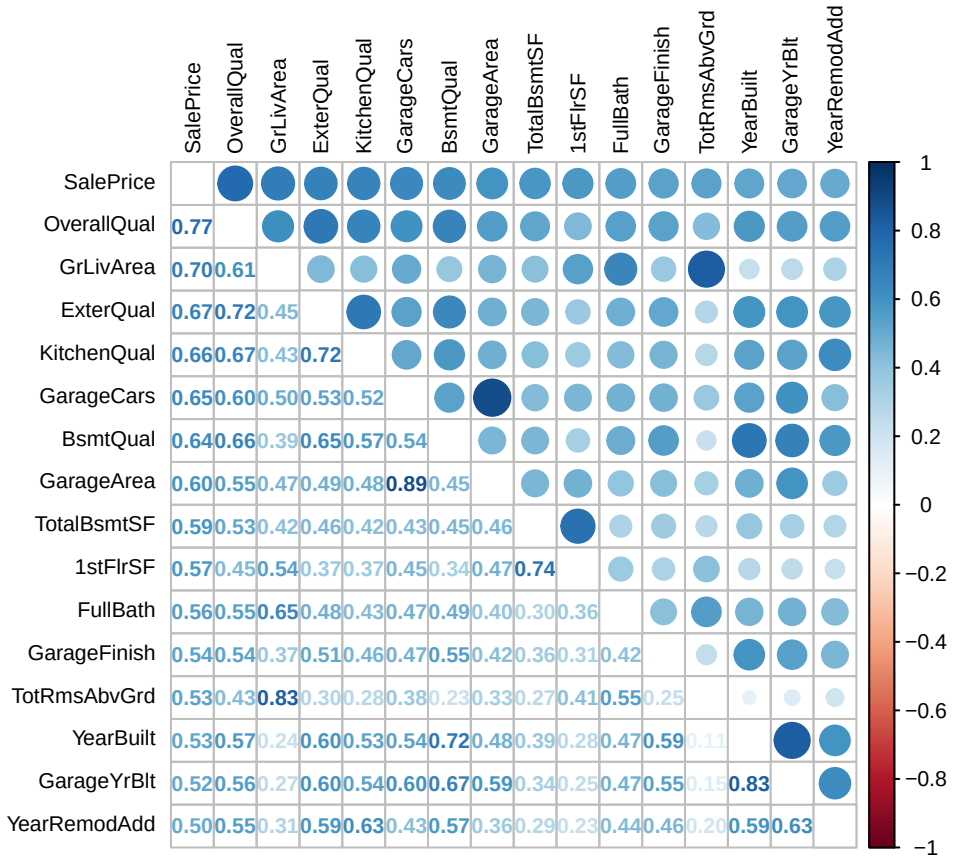


Figure 1: Correlation plot of numerical variables of high correlation with SalePrice

It turns out the overall quality of the house and *GrLivArea*, the above ground living area, are the variables of highest correlation with *SalePrice*. We also see there is high intercorrelation between some of this variables of interest. Not surprisingly are features like *GarageCars* and *GarageArea*, and features like *YearBuilt* and *GarageYrBlt* highly correlated. In general we observe that variables concerning the same area (ie. garage, basement and so on) are typically highly correlated. This fact will give reason to a model reduction in the modeling phase by removal of variables that are highly correlated to a more important variable.

2.1.1 Relationship with correlated features

Considering figure 2 there seems to be a growing trend to in SalePrice when increasing the overall quality. Also there does not seem to be any obvious outlier, perhaps apart from the somewhat expensive house with overall quality 4. Considering figure 3 there are no strong trend, but one could argue *SalePrice* is a bit higher for newer houses than old houses.

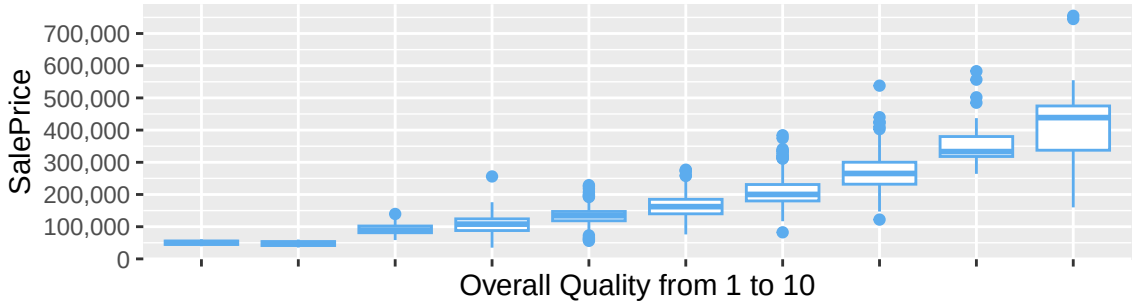


Figure 2: Boxplot of SalePrice divided by their overall quality ranging from 1 to 10.

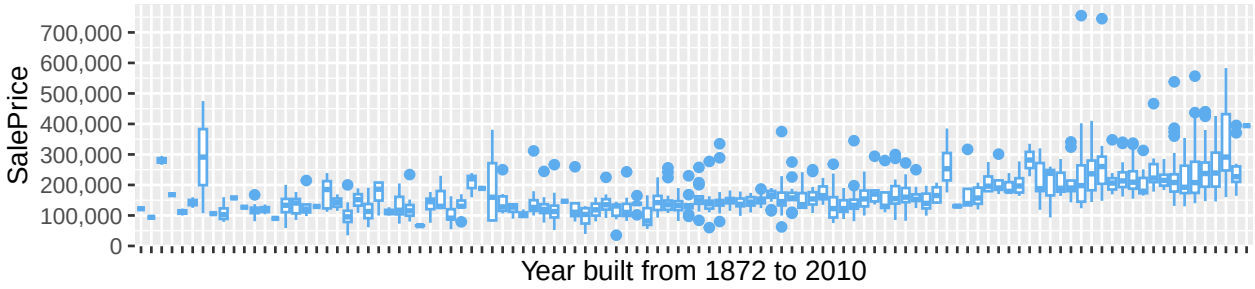


Figure 3: Boxplot of SalePrice divided by the years the houses are built from 1872 to 2010

Considering the relationship between *SalePrice* and *GrLivArea*, the second most correlated variable, we notice in figure 4 that there is a somewhat strong positive trend in SalePrice as function of GrLivArea. One could suggest that house 596 and 199 are outliers, since they are very large properties with a relatively small sale price. Taking a harder look at these houses they have a maximum normalized score of 2.65 in overall quality, which gives further reason to classify them as outliers. We will keep these houses in mind when removing outliers.

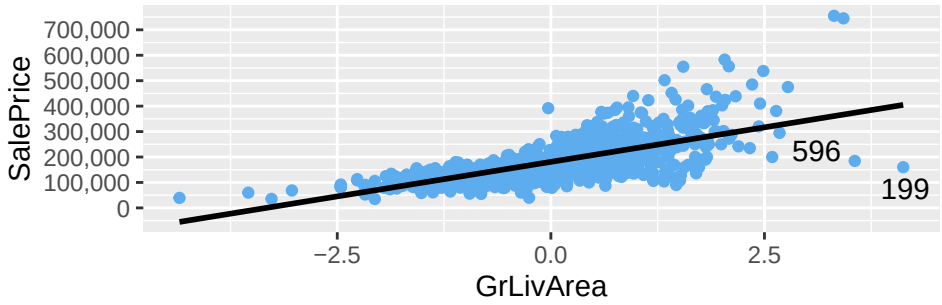


Figure 4: Scatterplot with trend of SalePrice as function of GrLivArea

2.2 Important predictors

Let us take a look at variables of high importance for the prediction. One can do this by building a multiple linear regression model using all the variables and look at the significance of each variable, but right now we

would like to get a feel of the data without building a huge model. To this end we will use random forests as inspired by (Bruin 2017), where we will refer to a classic reference for random forests (Breiman 2001). The technique computes the mean square error before and after permuting the features among the dataset to which we fit a random forest. The technique is implemented in the R-library `randomForest`.

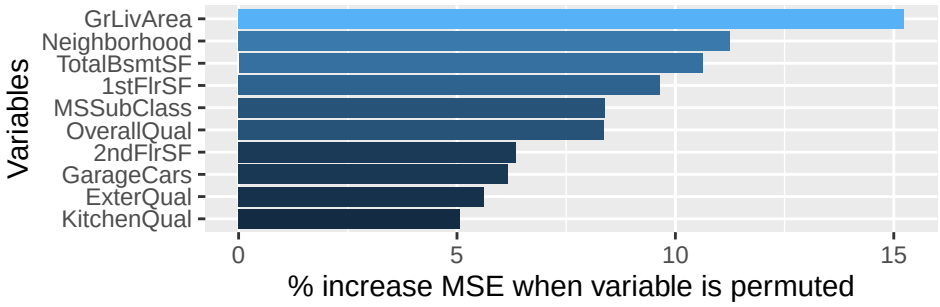


Figure 5: Most important features found by random forest technique

In figure 5 we see the 10 most important variables as ranked by the random forest model. We have already seen some of these variables' relationship with the response variable. Indeed it comes as no surprise that the above ground living area and overall quality are important, since houses scoring high in these variables are expected to be expensive. Let us now take a look on the distribution of predictors among the top 10, four of which are seen in figure 6. Some of the variables seem to have a central Gaussian distribution like *GrLivArea* or *1stFlrSF*, while distribution for other variables are skew like *OverallQual*, and yet other variables distribution are harder to determine since they have already undergone some transformation. We will keep this in mind when standardizing and and log-transforming in the modelling phase.

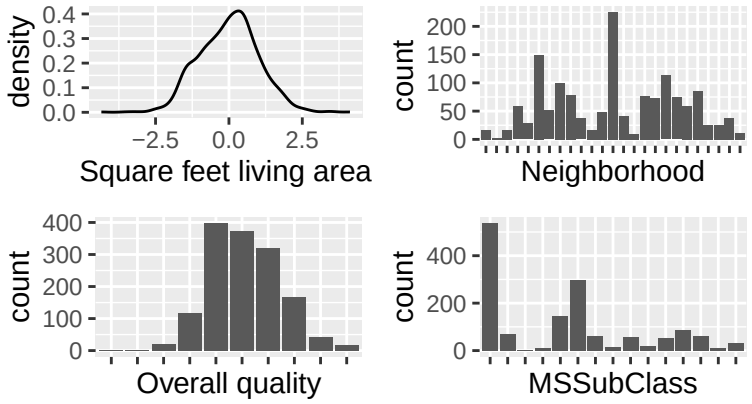


Figure 6: Distributions of important variables

One idea we could explore further is to merge the different neighborhoods in the neighborhood variable if the sale prices are similar enough. In this way we could reduce the dimensionality of this categorical variable.

3 Modeling and Diagnostics

Let us start the modelling by preparing our dataset. This preparation will consist of one-hot encoding of all categorical variables and a standardization of numerical variables. Our data combined now amounts to 1460 observations with 199 variables.

3.1 A multiple linear regression model

We have fitted our training data to a simple multiple linear regression model to the data of $p = 199$ variables and $n = 1095$,

$$Y = X^T \beta^* + \varepsilon,$$

where $\varepsilon \sim N(0, \sigma^2)$, $(X, Y) \in \mathbb{R}^p \times \mathbb{R}$ and $\beta^* \in \mathbb{R}^p$ is the unknown regression vector. We assume we have a sample of i.i.d. copies of $(X, Y) : (X_i, Y_i)_{i=1}^n$. We will explore this assumption further as well as the various assumptions regarding the residuals, see **P1** to **P4** in (Lounici 2019). Using our designmatrix of all our observations $\mathbb{X} = (X_1, \ldots, X_n)^\top = (X_{i,j})_{1<i<n, 1<j<p}$ and using our observed responses $\mathbb{Y} = (Y_1, Y_2, \ldots, Y_n)^T$, we want to estimate β^* by the following minimization problem

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\mathbb{Y} - \mathbb{X}\boldsymbol{\beta}\|^2.$$

Notice then that the residuals $\epsilon = \mathbb{Y} - \mathbb{X}\boldsymbol{\beta} \sim N(0, \sigma^2 I_n)$. This is done in `R` using the `lm`-function. The diagnostics for this regression can bee seen in figure 7.

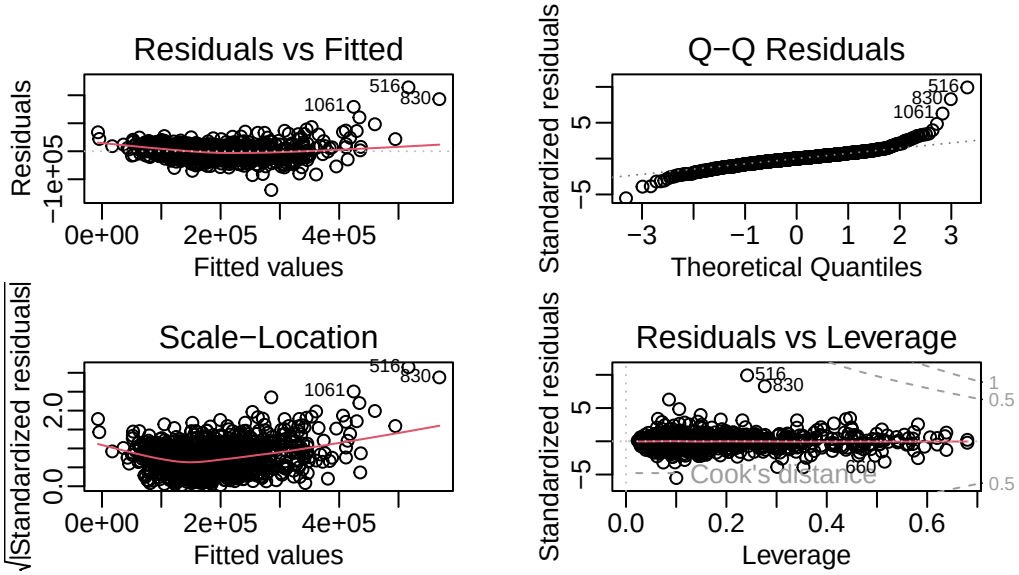


Figure 7: Diagnostics of full linear model

There are many different problems with this model. Not only does it use all 199 variables, many of which are unsignificant and many of which return parameters with N/A. We tackle this by reducing the number of dimensions in the variables. Here we merge neighborhoods of similar distribution in *SalePrice*. Indeed there is a clear top 3 in neighborhoods in terms of mean and median sale prices. Namely *NoRidge*, *Stonebr* and *NridgHt* seems to be more expensive neighborhoods to buy a house in. Furthermore we collect the bottom 4 neighborhoods in a cheaper category. We collect the rest in a middle neighborhood.

As a way of reducing the dimensionality of categorical variables that are sparse we will consider Tukey's test to determine if the mean difference between specific pairs of group in the categorical variable are statistically significant. As an example consider the variable *BldgType*.

From the insignificant differences we group the variables in two thereby reducing dimensionality and ensuring non-sparsity in the future dummy variables. We do something similar for *RoofStyle*.

In order to explain more of the variance of the residuals we see in the diagnostics plot, *Residuals vs Fitted*, we construct 2 new variables based on the old ones: *Age* and *TotalBath*, which counts the total number of bathrooms in the house. We will consider creating a model of higher order at a later point.

The diagnostics in Figure 7 of the model reveals issues with the assumptions of the linear model. In the scale-location plot, we see a clear trend. This implies there is problem with **homoscedasticity** of residuals;

the price of expensive houses varies more than cheap houses. This also brings us to the skewness of the response variable as well as the predictors, which contradicts the assumption of **multivariate normality**. In order to fix this issues we will log-transform skew predictors. We do a **Box-Cox** transform of the response variable, since this gave better results compared to a log-transform. This will also improve the spread location of our residuals.

Now upon one-hot encoding categorical variables, we might be worried about the sparsity in certain levels of the variables. This could lead to akward situations where the level is nonzero in the training set, but not in the testing set or vice versa. We fix this by simply removing levels with fewer than 10 observations, since *treaments* with fewer observations will hardly explain the variance in the response variable. This reduces the number of variables from 199 to 124.

As a final note to the diagnostics Figure 7 we would like to **remove outliers**. The "Residuals vs Leverage" plot reveals potential leverage points 830 and 516, which also seems to be regression outliers in the "Scale-Location" plot. We will not remove these just yet, since the model assumptions are not satisfied. In addition the leverage analysis locates certain observations of Cook's distance 1. We suspect this is due to $h_{ii} = 1$ for some observation i, hence $P_X Y = Y$, where P_X is the usual projection. This could just be due to sparsity. Lastly we remove the outliers 596 and 199, partly because of the reasons already stated: very low sale price compared to area and overall quality, and partly since they have the highest Cook's distance for the new linear model.

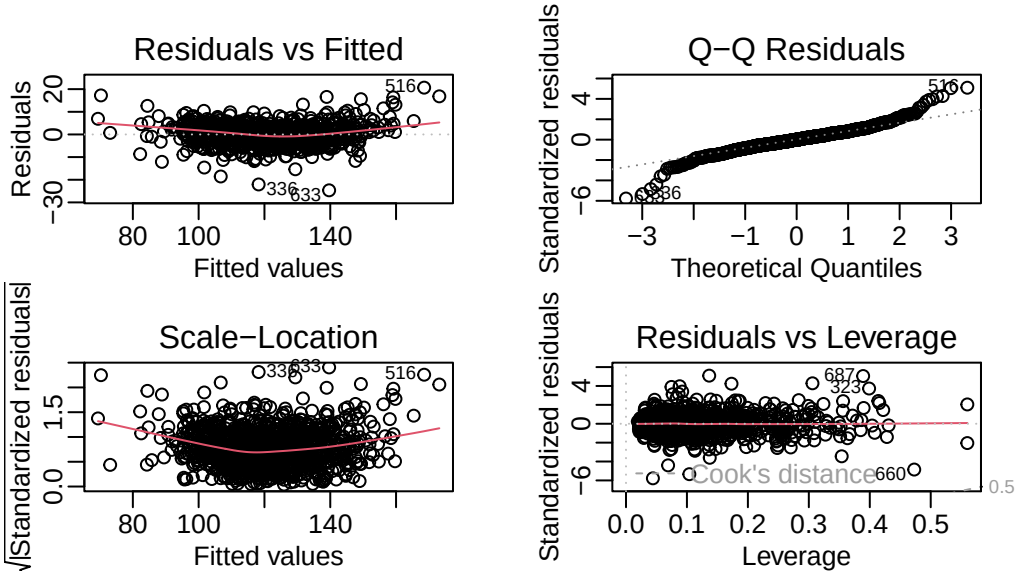


Figure 8: Diagnostics of linear model with modifications

To comment our diagnostics of our new improved linear model, Figure 8, it seems there is still unexplained variance in the residuals. The residual seems normally distributed and while the "Scale-Location" plot does reveal a trend, it is better. Indeed now there are a pretty similar variance for the most of the residuals, but for more extreme fitted values the variance is higher. Finally we seem to have no leverage points. Without showing a plot we assure the reader that indeed the autocorrelation assumption of the residuals is also satisfied. We note for this model we have `RMSE` $= 24659.24$ and $R^2 = 0.91$ evaluated on the test data, which can be improved considerably. See the section "Comparison" for a comparison with our other models.

4 Final models

Our current linear model faces the challenge of reduction and is still lacking in explaining variance in the residuals. Because of the colinearity and sparsity of many variables t-tests is not necesarrily a good way of estimating variables significance. Fisher tests seems a bit infeasiable for this amount of variables, so we must

turn to other methods. Thus we will start by gaining some knowledge from a Lasso regression. It is a linear regression model utilizing a ℓ_1 penalty, i.e. estimates the parameters of the linear model by minimizing the Lagrangian

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ \|\mathbb{Y} - \mathbb{X}\boldsymbol{\beta}\|^2 + \lambda\|\beta\|_1 \right\},$$

where λ is a parameter we optimally obtain by 10-fold cross-validation. For sufficiently large values of λ this penalty produces parsimonious models.

4.1 Lasso

We find Lasso estimators for 2 models: one of first order and one of second order interactions of our variables. The first model throws away 51 of the 124 variables, whereas the second model throws away 7409 out of 7626 variables, which is a quite remarkable reduction in both cases.

Using instead the λ which gives the most regularized model such that error is within one standard error, we find among the important variables: *GrLivArea*, *OverallQual*, *SaleTypeNew*, *KitchenQual* in and *LotArea* as top 5 in the **first model**. Not far away from our results from random forests, Figure 5. In this model we throw away 89 and 7549 variables respectively. The importance of these variables make a lot of sense: the above ground living area, the overall quality and kitchen quality are just very important factors for people.

As important variables in the **second model** we find *GrLivArea*, *OverallQual*, *PavedDrive:Exterior1stBrkFace*, *KitchenQual* and *LotArea* as top 5, which are indeed very close to model 1, but also rather surprising. Somehow the interaction between a paved drive and the exterior covering on the house is quite relevant for the house. We also find the interaction *Neighbor:Condition2Norm* of surprisingly high importance: it is the combination of being in the right neighborhood and having a normal proximity to conditions like railroads.

We note RMSE is 25231.37 and 21482 for the first and second model respectively, while R^2 is 0.911 and 0.937 when used on the testing data.

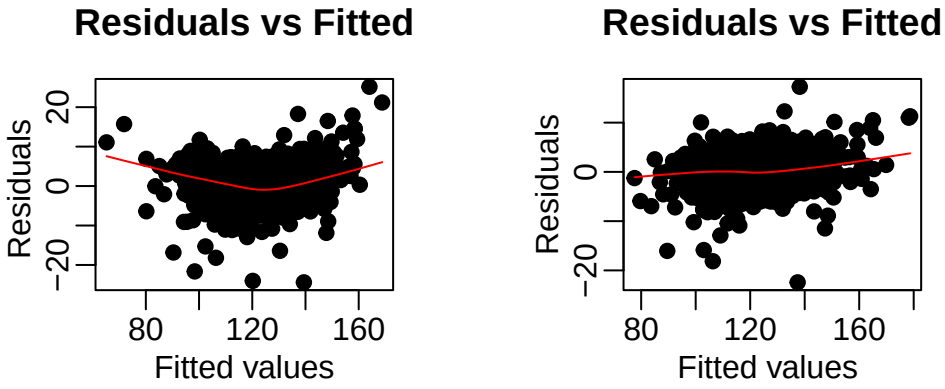


Figure 9: Centrality of residuals with respect to the fitted values in the first and second Lasso model

As a final note we have produced a simple linear regression model including the top 50 important variables as found by the Lasso estimator. We then use the Akaike information criterion (**AIC**) to select a model using forward, backward and both ways selection, gaining the the same model with backward and both ways selection, yielding only a 6 variable reduction. Doing an **ANOVA** on this remaining model of 44 variables suggests indeed all variables are significant, except for one, which has a p-value of 0.10. The estimates, standard errors, confidence intervals, and p-values of the parameters for this model can be found in the **Appendix**.

4.2 Gradient Boosting

As a final model we will use gradient boosting as implemented in the Xgboost library. We will not comment the method here, rather refer to the paper (Chen, Tianqi and Guestrin, Carlos 2016), and comment the final

result in the next section.

4.3 Comparison

In this section we compare our models predictions on the test data. Note we have only *trained* and estimated parameters on the training data. In the case of the Lasso and Xgboost we trained the models using cross validation, hence this adds another test layer.

Table 1: Testing metrics of our final models. Lasso1 and Lasso2 uses 72 and 217 variables resp, where as the simple linear model uses 124. The AIC selected linear model uses only 45

	Ensemble	Xgboost	Lasso1	Lasso2	Linear.model	Linear.model.AIC
RMSE	2.21e+04	2.42e+04	2.52e+04	2.15e+04	2.47e+04	2.59e+04
Normalized RMSE	1.23e-01	1.35e-01	1.40e-01	1.19e-01	1.37e-01	1.44e-01
R^2	9.32e-01	9.15e-01	9.11e-01	9.37e-01	9.13e-01	9.01e-01

Considering Table 1 we conclude Lasso2 gets the best result. Even combined with the Xgboost model the ensemble method does not perform better. Interestingly there is not much difference in performance between the simple linear model using 124 variables and the linear model only using 45 variables.

5 Discussion.

We conclude our hypothesis somewhat corresponds to our results. We have seen a fairly accurate description of the sale price with a Lasso model using 217 variables which performs best on the test data with a normalized MSE of 0.1193. We have seen this model performs well on the diagnostics as well. We have considered the most important variables as found by different methods: random trees, Lasso, AIC, anova, and thereby combating the colinearity of the many variables. Finally we have seen that a large reduction of variables is possible, while still maintaining a certain level of accuracy.

With that in mind there are some limitations of our analysis. For example the data cleaning could be better, as there is a lot of information lost in the process. We put a lot of trust in our Lasso model, however it is not very precise, since it is a biased estimator due to the regularization. This also means it does not make sense to estimate the variability of the parameters making it harder to interpret the model. The Xgboost model needs more training in order to perform better, but this is also an opportunity for further work on our predictions. This includes hypertuning of the parameters as well.

Future directions of our research could include an improvement of our reduced linear model by picking a number of important variables suggested by an ensemble model of Xgboost, random trees and Lasso and then finally reducing this my clever model selection tools as glmulti or simulated annealing. Then using a more rigorous MANCOVA we could gain more insights in the variables. Finally a lot more work could be put into the interpretation of the important variables and their parameter estimations - what weighs positively and what weighs negatively?

6 References

Breiman, Leo. 2001. “Random Forests.” https://doi.org/10.1023/A:1010933404324.

Bruin, Erik. 2017. “House Prices: Lasso, XGBoost, and a Detailed EDA.” Kaggle. https://www.kaggle.com/erikbruin/house-prices-lasso-xgboost-and-a-detailed-eda.

Chen, Tianqi and Guestrin, Carlos. 2016. “Xgboost: A Scalable Tree Boosting System.”

Lounici, Karim. 2019. “Visual Diagnostics and Model Validation.”

7 Appendix

	Estimate	Std. Error	t value	Pr(>|t|)	2.5 %	97.5 %
(Intercept)	117.4725323	0.3418569	343.630675	0.0000000	116.8017303	118.1433343
GrLivArea	4.4996001	0.2019986	22.275402	0.0000000	4.1032323	4.8959678
OverallQual	2.6066024	0.2150891	12.118708	0.0000000	2.1845480	3.0286568
TotalBsmtSF	2.4535041	0.2181293	11.247933	0.0000000	2.0254842	2.8815241
KitchenQual	1.1652296	0.1863364	6.253367	0.0000000	0.7995948	1.5308644
GarageCars	1.2809297	0.1689609	7.581218	0.0000000	0.9493894	1.6124699
YearRemodAdd	0.6718153	0.1795683	3.741281	0.0001930	0.3194611	1.0241696
LandContourHLS:Exterior1stMetalSd	7.4171040	1.8584033	3.991116	0.0000703	3.7704890	11.0637191
LandContourLow:SaleConditionPartial	-7.8006963	4.0085519	-1.946014	0.0519201	-15.6663979	0.0650052
SaleConditionFamily:MasVnrArea	-3.9465944	0.9347129	-4.222253	0.0000263	-5.7807162	-2.1124726
LotShape:‘MSSubClass2,5 story all ages‘	-3.9303086	1.1173403	-3.517557	0.0004542	-6.1227875	-1.7378297
SaleConditionFamily:ExterCond	-3.3011713	0.8939185	-3.692922	0.0002331	-5.0552452	-1.5470974
‘Exterior1stWd Sdng‘:GarageTypeBasment	-3.3769534	2.4297826	-1.389817	0.1648796	-8.1447461	1.3908393
SaleConditionFamily:Exterior1stPlywood	-3.9721290	1.8673351	-2.127165	0.0336394	-7.6362702	-0.3079878
Exterior1stBrkFace:LotArea	2.9842400	0.8167452	3.653820	0.0002712	1.3815980	4.5868820
LotArea:Condition2Norm	1.3136520	0.1592834	8.247262	0.0000000	1.0011013	1.6262027
SaleConditionPartial:RoofStyleHip	2.6877422	1.0137573	2.651268	0.0081400	0.6985171	4.6769673
Exterior1stBrkFace:CentralAirY	3.5645220	0.7857574	4.536415	0.0000064	2.0226850	5.1063589
Condition2Norm:Neighbor	1.0652316	0.1696516	6.278935	0.0000000	0.7323361	1.3981272
‘MSSubClass2 story 1945-‘:MSZoningRL	3.7805837	0.8943229	4.227314	0.0000257	2.0257164	5.5354511
‘MSSubClass1 story 1945-‘:HeatingQCTA	-3.0144743	1.1138707	-2.706305	0.0069142	-5.2001451	-0.8288036
Exterior1stMetalSd:HeatingQCFa	-3.3645044	1.3678302	-2.459738	0.0140644	-6.0485021	-0.6805067
LandContourLow:Exterior1stPlywood	6.8434875	1.5266292	4.482744	0.0000082	3.8478896	9.8390854
GarageTypeBasment:LandSlopeMod	-5.5458607	2.8116581	-1.972452	0.0488206	-11.0629811	-0.0287404
Functional:HouseStyle2Story	1.7335691	0.2116855	8.189360	0.0000000	1.3181933	2.1489448
BsmtQual:MSZoningFV	2.8880272	0.6471089	4.462969	0.0000090	1.6182506	4.1578039
GarageCond:Exterior2ndBrkFace	4.3706309	0.9939941	4.397039	0.0000121	2.4201858	6.3210761
RoofMatlCompShg:YearBuilt	1.8371430	0.2305579	7.968249	0.0000000	1.3847353	2.2895507
SaleConditionPartial:WoodDeckSF	1.5947952	0.4473474	3.565003	0.0003803	0.7169965	2.4725939
LotArea:‘Exterior2ndWd Shng‘	2.3658154	0.6533003	3.621329	0.0003071	1.0838899	3.6477410
OverallCond:SaleTypeWD	1.7029311	0.1663389	10.237721	0.0000000	1.3765359	2.0293263
GrLivArea:LandSlopeMod	2.6277319	0.5065932	5.187065	0.0000003	1.6336795	3.6217844
YearRemodAdd:MSZoningRH	3.5493398	1.0364812	3.424413	0.0006400	1.5155251	5.5831545
CentralAirY:Condition1Norm	1.0919893	0.3539748	3.084935	0.0020893	0.3974092	1.7865694
TotalBsmtSF:FoundationPConc	0.7956444	0.3569229	2.229177	0.0260135	0.0952795	1.4960093
GrLivArea:‘MSSubClass2 story 1946+‘	2.0395504	0.4733799	4.308485	0.0000180	1.1106701	2.9684307
LotShape:RoofStyleHip	-0.9344473	0.2784021	-3.356466	0.0008178	-1.4807364	-0.3881582
KitchenQual:LandContourHLS	1.7483825	0.6301371	2.774607	0.0056251	0.5119085	2.9848565
HalfBath:‘MSSubClass1,5 story fin‘	1.4368036	0.4153695	3.459098	0.0005638	0.6217531	2.2518541
Neighbor:Condition1Feedr	1.3939713	0.6266083	2.224630	0.0263185	0.1644217	2.6235210
TotalBsmtSF:BsmtQual	1.0441610	0.1812167	5.761947	0.0000000	0.6885721	1.3997498
HeatingQCTA:YearBuilt	0.7536541	0.3430978	2.196616	0.0282663	0.0804174	1.4268909
HeatingQCTA:OverallCond	0.8334018	0.2421111	3.442228	0.0005997	0.3583240	1.3084796
OverallQual:KitchenQual	1.3298101	0.1259975	10.554258	0.0000000	1.0825740	1.5770462
‘MSSubClass2 story 1946+‘:FullBath	0.9324564	0.5126660	1.818838	0.0692214	-0.0735122	1.9384250