Final Report: House Prices Advanced Regression Techniques

Adesh Valecha, Sushruti Acharya

April 30, 2018

Introduction

The main objective of this project is to analyze, clean the data and choose strongest predictors and build the model to predict the Sale Price. We will be predicting the Sale Price of individual residential property as described in the Kaggle dataset. The data has been split into 50% train and 50% test sets. The testing dataset consists of 1459 observations with 80 variables and the training dataset consists of 1460 observations each with 81 variables including the Sale Price. The dataset consists of 2919 observations in all, consisting of 14 discrete, 23 nominal, 20 continuous and 23 ordinal variables (not including Sales Price). The training dataset is used for training the model and the testing dataset is used for predicting the Sales Price and evaluating the model performance.

Data modeling and cleaning

After performing exploratory data analysis, it was noted that NA represented a category for variables - Alley, PoolQC, Fence, MiscFeature, BsmtQual, BsmtCond, BsmtExposure, BsmtFinType1, FireplaceQu, GarageType, GarageFinish, GarageQual and GarageCond, therefore it was replaced with “Unknown” so that the system would not consider it as garbage value. There were some variables with missing values (2 category variables – MasVnrType and Electrical and 3 continuous variables – LotFrontage, MAsVnrArea and GarageYrBlt). For the categorical variables, NA was replaced with “Unknown” and assigned a level. For the Continuous variables, the NA value was replaced by their mean values.

Model and model development

We are using five modelling methods in our report to find the strongest predictors. These methods are Linear regression, Ridge model, Lasso model, Ridge/Lasso mixed model and Correlation. We selected the best predictors from each model and built a baseline model (Linear model) to get the strongest predictors across all models. Below table displays the strongest predictors for each model.

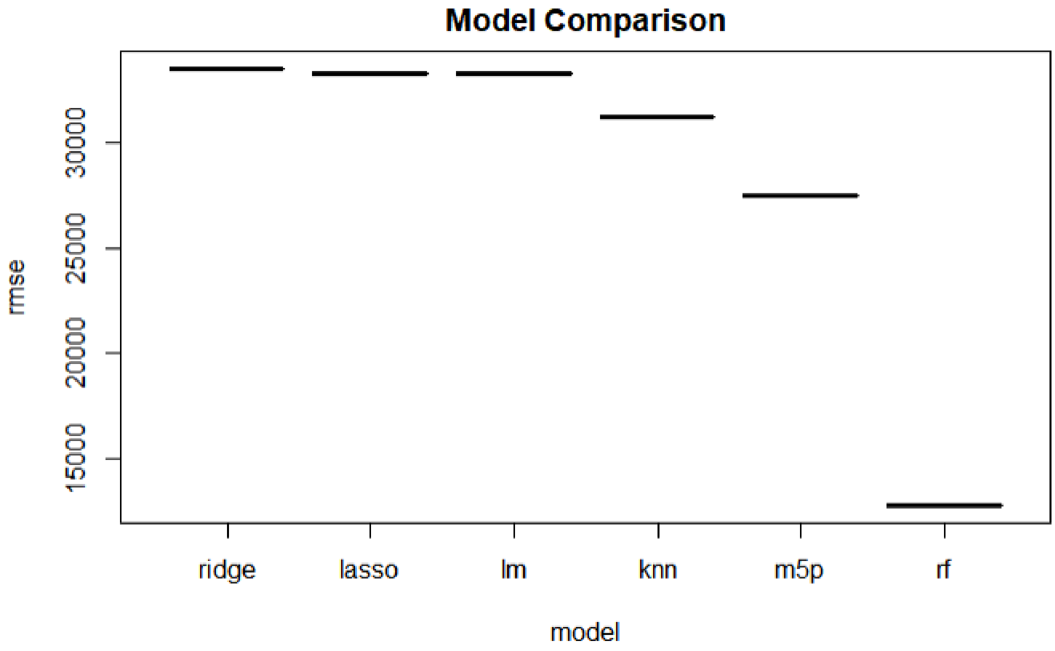
Linear Model	Ridge Model	Lasso Model	R/L Mix. Model	Correlation
Pool quality	OverallQual	GrLivArea	OverallQual	OverallQual
Utilities	GrLivArea	OverallQual	GrLivArea	GrLivArea
Street	X1stFlrSF	Pool quality	X1stFlrSF	GarageCars
GarageCars	BsmtQual	PoolArea	BsmtQual	ExterQual
KitchenAbvGr	KitchenQual	GarageCars	KitchenQual	GarageArea
OverallQual	PoolQC	BsmtQual	Pool quality	BsmtQual
ExternalQual	X2ndFlrSF	KitchenQual	GarageCars	TotalBsmtSF
Condition2	GarageCars	YearBuilt	X2ndFlrSF	X1stFlrSF
KitchenQual	ExterQual	ExterQual	ExterQual	KitchenQual
BsmtQual	PoolArea	MasVnrArea	PoolArea	FullBath
BsmtFullBath	TotRmsAbvGrd	TotRmsAbvGrd	TotRmsAbvGrd	GarageFinish
RoofMatl	MasVnrArea	OverallCond	MasVnrArea	TotRmsAbvGrd
LandSlope	OverallCond	MSSubClass	OverallCond	YearBuilt
OverallCond	MSSubClass	LotArea	MSSubClass	MasVnrArea
MasVnrType	YearBuilt	BsmtExposure	BsmtExposure	YearRemodAdd

All of the top predictors from the different models were then tested under various permutations and combinations in order to find the strongest model. Comparing the output of all the above models and techniques we reached to a conclusion that the strongest predictors which can be used to develop a model are as below:

Strongest Predictors:

1. Pool quality
2. GrLivArea
3. YearBuilt
4. GarageCars
5. OverallQual
6. KitchenQual
7. BsmtQual
8. OverallCond
9. X1stFlrSF
10. MSSubClass
11. PoolArea
12. ExterQual
13. MasVnrArea
14. LotArea
15. TotRmsAbvGrd
16. GarageArea
17. TotalBsmtSF
18. FullBath

Different modeling methods were then applied on the above predictors to find the most accurate model. These methods included Linear, Ridge, Lasso, KNN, Random forest and M5P. The below plot compares the RMSEs for all the models and shows the minimum and maximum RMSE of the models.



Looking at the RMSEs of the above models, we can say that the random forest and M5P method has the lowest RMSE.

Model Performance

The random forest and M5P models were then tested for in-sample and out-of-sample performance. Below are the results:

Random Forest model:
Out of sample RMSE: 31363.79 and R-squared:0.8393970
In-sample RMSE: 12794.9 and R-squared: 0.87
Kaggle score: 0.15217

M5P model:
In sample RMSE: 27532.66 and R-squared:0.879926
Kaggle score: 0.14030
Kaggle Rank: 2147

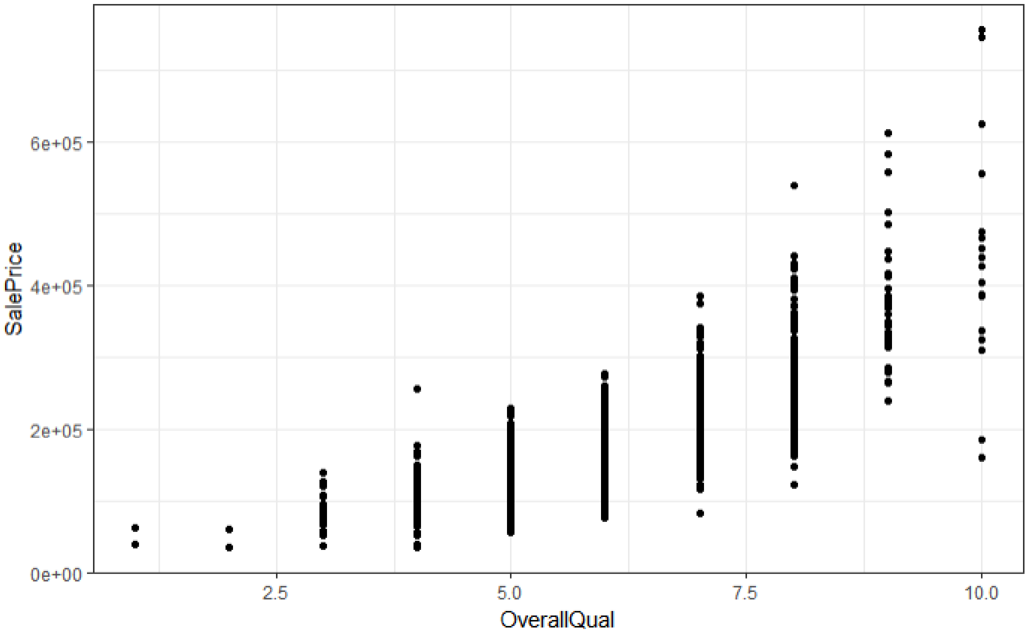
On checking the RMSE score on Kaggle for Random Forest and M5P models, we observed that M5P is performing better than Random Forest. So, M5P is the best model.

Visualizations

Analyzing Data and effect of different most Significant Predictors (3) that were commonly identified through LM, KNN, GLMNET, Random Forest and correlation with ggplot.

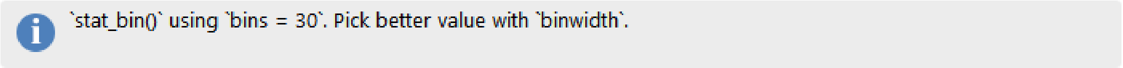
Impact of the most significant factor OverallQual on Sale Price

The coefficient for overall quality is .79 which seems to be the highest from rest of all predictors. The graph clearly shows that better the quality higher is the sales price.

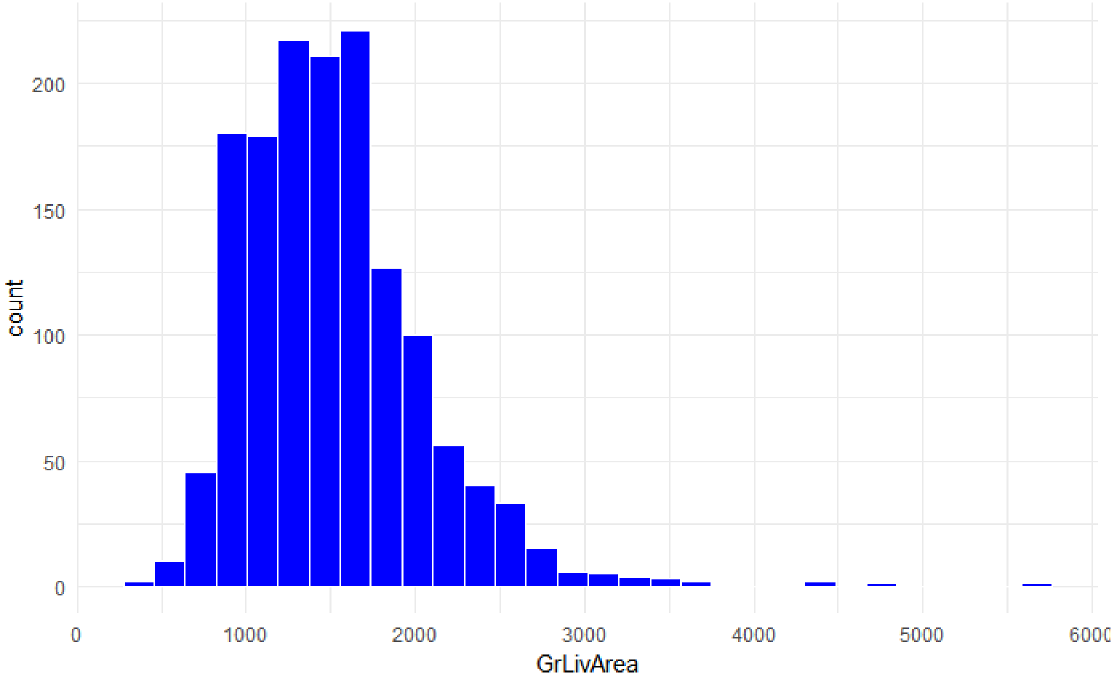


GrLivingArea Impact on Housing Price

In the below graph it can be seen that the count is pretty high for around houses having ground living area between 1000 to 2000 sqft, hence making it a great predictor influencing the price.

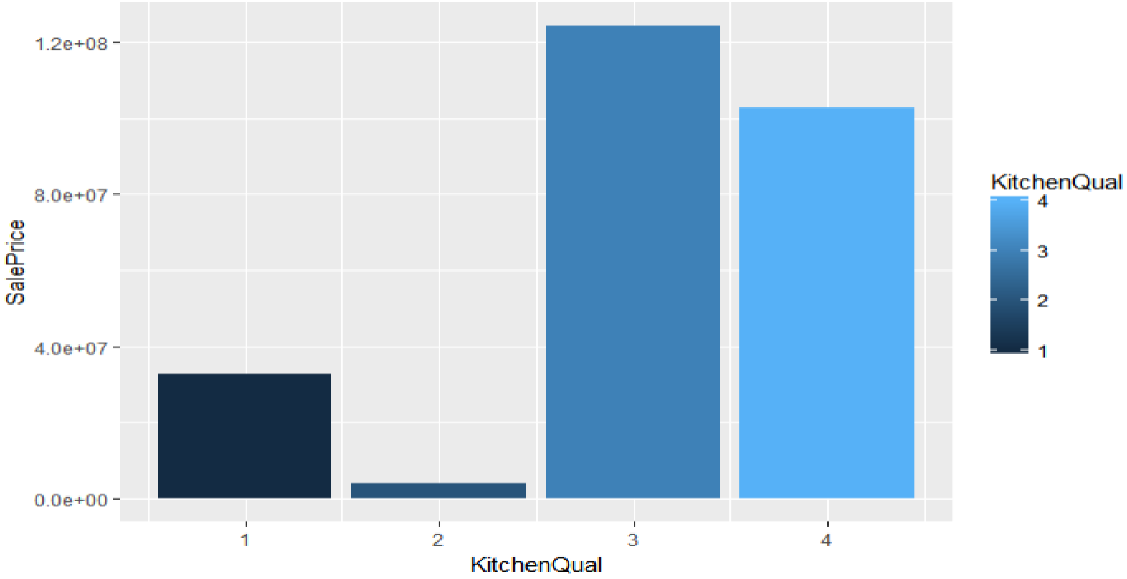


```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Kitchen Quality impact on Sale Price

As we can see in the below graph as well quality is segregated into four labels 1 being the least and 4 the best, the sales price is highly correlated.



Statistical Inference

Based on the analysis we did, we found that housing prices in Ames, Iowa can be strongly determined by the factors- Pool quality, Above grade (ground) living area square feet, Original construction date, Size of garage in car capacity, Overall material and finish quality, Overall condition rating, Kitchen quality, Height of the basement, First Floor square feet, The building class, Pool area in square feet, Exterior material quality, Masonry veneer area in square feet, Lot size in square feet, Total rooms above grade (does not include bathrooms), Size of garage in square feet, Total square feet of basement area and Full bathrooms above grade.

Kaggle Result

Based on different model comparison, we found that M5P model performance was best on out of sample data.

Kaggle score: **0.14030**
Kaggle Rank: **2147**