Ames House Prices Prediction Models

Background Information

Pricing house is a common problem in the real estate industry. On average, home sellers pay around more than $10,000 on realtor fees, from advising house prices to advertising the properties on multiple platforms. As real estate agencies rarely reveal how they develop their valuation framework and collect housing data and research, it is not always clear what factors influence the worth of a house. The Boston Housing Dataset in 1978 provides data about houses and sales in Boston and this dataset has been popular as a regression modeling challenge. However, this dataset has become outdated and the number of observations was around 500. A new dataset, Ames housing data provided by Professor Dean De Cock from Truman State University, is updated with more variables and observations. Ames housing price contains the sale of individual residential property in Ames, Iowa from 2006 to 2010. The data set contains 2930 observations and a variety of predictors (23 nominal, 23 ordinal, 14 discrete, and 20 continuous).

Motivation

Without the experience in pricing properties and the help of realtors, how can a person get a good estimation of the house price? What factors are of most importance in determining the worth of the house?

In this report, I would investigate this dataset variable, narrow down a list of predictors for predictive modeling, employ a variety of regression models to fit this dataset and compare their performance based on the RMSE of the training set.

Exploratory Data Analysis (EDA)

I used the training dataset from Kaggle, which contains 1460 observations and 81 predictors. I investigated the numerical variables and examined their correlation against the response variables to narrow down the lists of significant numerical variables.

Numerical Variables

Out of 38 numerical variables, 3 variables have missing values in the dataset, which are LotFrontage (Linear Feet Street connected to the property), GarageYearBuilt, and MasVnrArea (Masonry Veneer Area). Examining the null value cases could hint that some houses lack certain features/rooms (for example garage, ...).

It is difficult to look at all 38 numerical variables at once since we are interested to get some variables with moderate correlation with the response variables to put into regression models, thus we examine the correlation matrix of numeric variables with the response variable (Table 1). We filter only variables that have correlation strength with response variables above 0.3 and generate distribution plots (Figure 1)

We prefer variables with close to a normal distribution, continuous values, and few zeroes. We narrow down the variables to 11 numerical predictors. From the pairwise of these variables, we remove

variables that are too highly correlated with another variable aside from the response variable (Figure 2).

We remove LotFrontage since it contains NA values and has a skewed distribution, YearBuilt and GarageYearBuilt are also removed due to high correlation with YearRemodAdd.

Categorical variables

Following a similar procedure as numerical variables, we look at 43 categorical predictors and look at their pairwise plots with SalePrice (pairwise plots can be found in the HTML file). From visual inspection, we select categorical predictors that can separate sale prices into specific price groups and do not have lots of NAs. We find that Alley, ExterQual, MSZoning, and KitchenQual meet these criteria.

Selected variables for Regression Modeling: 11 variables (details under appendix)

Data Transformation

Since some data have skewed distribution, including sale price. We applied the log transformation to these variables.
logBsmt: log(TotalBsmtSF + 1)
logLivArea: log(GrLivArea +1)
logSale: log(SalePrice +1)

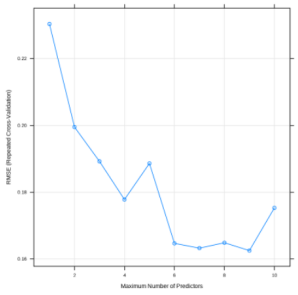
Training and Validation

For most regression methods, we would use repeated K-Fold cross-validation. We use 5 folds and repeated them 5 times. The goal is to produce stable results through repeated CV. We also set the seed to 05112022.

Model 1: Linear Regression

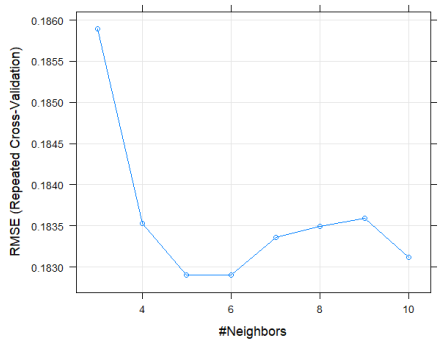
We use linear regression with stepwise selection given 10 predictors. The final model contains 9 variables. It has an RMSE of 0.1624781 and an R-squared of 83.50%.

Figure 3: RMSE of suggested stepwise models



Model 2: KNN Regression

We use KNN regression with k from 3 to 10. At k = 5, the model produces the lowest RMSE of 0.1829005 and R-squared of 79.10%



Model 3: Regularized Model (ElasticNet)

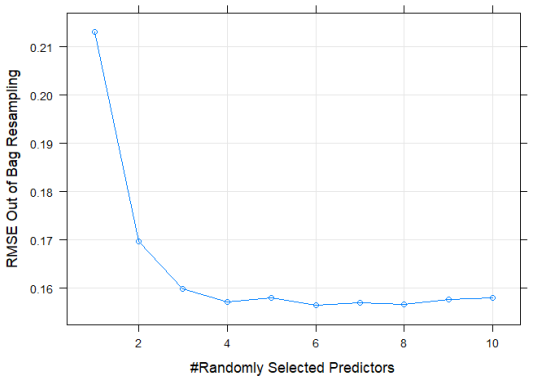
We choose ElasticNet because we suspect that our variables contain both useful and less useful variables, thus using a hybrid model of LASSO and Ridge would be effective. For tuning parameters, we set alpha to 1 and lambda between 0 and 0.5 with a step of 0.05. The recommended model suggested lambda = 0. This model would be close to a full linear regression with RMSE of 0.1586 and R-squared of 84.27%

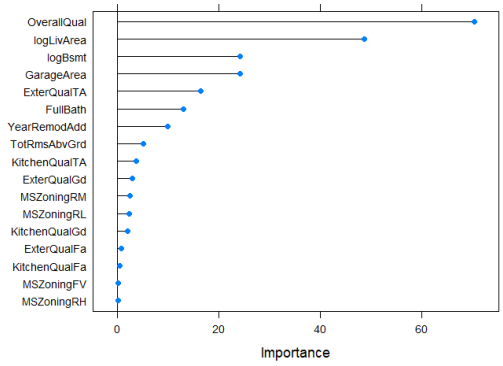
	s1
(Intercept)	12.024057395
YearRemodAdd	0.044916380
TotRmsAbvGrd	-0.018473992
OverallQual	0.117531593
GarageArea	0.064558386
FullBath	0.002136259
ExterQualFa	-0.010500773
ExterQualGd	-0.027015644
ExterQualTA	-0.042702136
MSZoningFV	0.088270577
MSZoningRH	0.041230718
MSZoningRL	0.198396452
MSZoningRM	0.112572098
KitchenQualFa	-0.027973571
KitchenQualGd	-0.044866999
KitchenQualTA	-0.063542861
logBsmt	0.037998077
logLivArea	0.144790886

Model 4: Regression Tree - Random Forest

We try regression tree and random forest as the ensemble method for this dataset. Since the variables have weak to moderate correlation with each other, we want to use the random forests to reduce variability and decorrelate trees. We use Out-of-Bag estimation error to train the model and use mtry from 1 to 10. The method suggests that mtry = 6 would be the optimal value, and random forest would result in RMSE of 0.1564. Based on the variable importance plot, the overall quality is the most

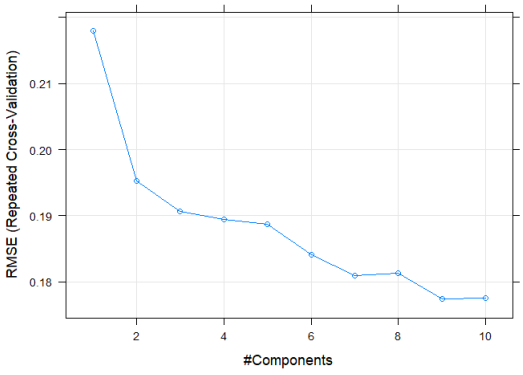
important variable, and then other numerical variables related to the area. The zoning type has the least impact among other variables.





Model 5: Dimension Reduction Model - PCR

We fit PCR model with ncomp (number of components) from 1 to 10. The best PCR model has 9 components an RMSE of 0.177 and an R-squared of 80.35%



	.outcome
YearRemodAdd	0.0276787265
TotRmsAbvGrd	0.0489266834
OverallQual	0.0741778970
GarageArea	0.1030416921
FullBath	0.0366170687
ExterQualFa	-0.0243179692
ExterQualGd	0.0142817086
ExterQualTA	-0.0272112407
MSZoningFV	0.0008638928
MSZoningRH	-0.0023426043
MSZoningRL	0.0276806578
MSZoningRM	-0.0302165404
KitchenQualFa	-0.0112623894
KitchenQualGd	0.0007362477
KitchenQualTA	-0.0229243030
logBsmt	0.0601329159
logLivArea	0.0639208052

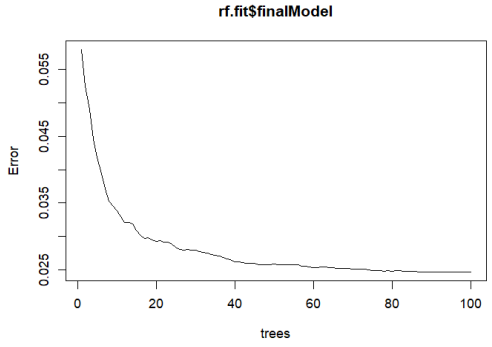
Model Comparison

Model	Test MSE
Stepwise regression	0.1624781
KNN regression	0.1829005
ElasticNet	0.1586696
Random Forest	0.1564387
PCR	0.1774054

Random Forest is the most efficient method for this dataset with an RMSE of 0.1564 and Elastic Net is the 2nd most efficient with an RMSE of 0.1586.
Random forest performs well with this dataset as this dataset includes both categorical and numerical predictors. This method generalizes well and is able to split using the more significant variables at first. The Out-Of-Bag estimation also would help Random Forest produce efficient prediction error.

Conclusion

In this report, we start with data exploration to fitting the regression model and compare. We find that random forest is the most useful method for this dataset. We also find another model with a close performance to the random forest is ElasticNet. Since random forest compiles results of a handful of trees, we could not plot this tree. However, we can determine that predictors like overall quality, living room area, basement area, external quality, and a number of bathrooms have the most influence on sale prices. We find that prediction error decreases significantly after fitting 25 trees



Appendix

Table 1: Moderate to strong correlation between numerical predictors and sale price

row	column	cor
<chr>	<chr>	<dbl>
LotFrontage	SalePrice	0.3442698
OverallQual	SalePrice	0.7978807
YearBuilt	SalePrice	0.5253936
YearRemodAdd	SalePrice	0.5212533
MasVnrArea	SalePrice	0.4886582
BsmtFinSF1	SalePrice	0.3903005
TotalBsmtSF	SalePrice	0.6156122
X1stFlrSF	SalePrice	0.6079691
X2ndFlrSF	SalePrice	0.3068790
GrLivArea	SalePrice	0.7051536
FullBath	SalePrice	0.5666274
TotRmsAbvGrd	SalePrice	0.5470674
Fireplaces	SalePrice	0.4618727
GarageYrBlt	SalePrice	0.5047530
GarageCars	SalePrice	0.6470336
GarageArea	SalePrice	0.6193296
WoodDeckSF	SalePrice	0.3368551
OpenPorchSF	SalePrice	0.3433538

We look at the distributions of these variables to see whether they should be transformed

Figure 1: Distribution of highly correlated numerical predictors

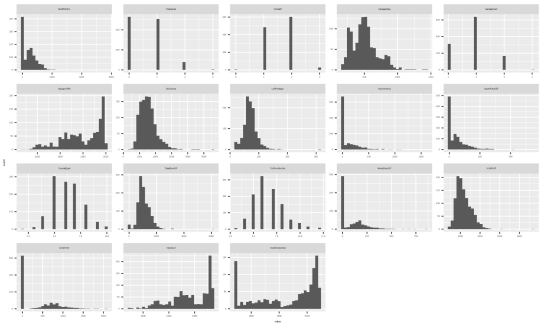
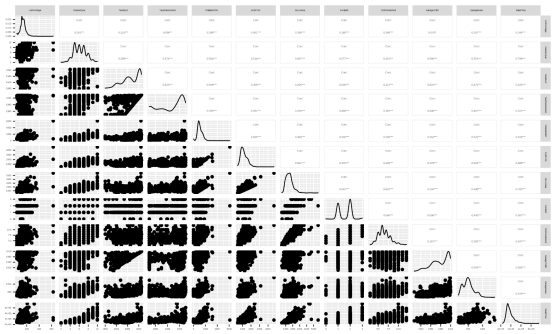


Figure 2: Pairwise plots of selected predictor variables



YearRemodAdd: Remodel date (same as construction date if no remodeling or additions)
TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)

TotalBsmtSF: Total square feet of the basement area
OverallQual: Rates the overall material and finish of the house (1: Very Poor - 10: Very Excellent)
GrLivArea: Above grade (ground) living area square feet
GarageArea: Size of garage in square feet
FullBath: Full bathrooms above grade
ExterQual: Evaluates the quality of the material on the exterior

Ex	Excellent
Gd	Good
TA	Average/Typical
Fa	Fair
Po	Poor

MSZoning: Identifies the general zoning classification of the sale.

A	Agriculture
C	Commercial
FV	Floating Village Residential
I	Industrial
RH	Residential High Density
RL	Residential Low Density
RP	Residential Low Density Park
RM	Residential Medium Density

KitchenQual: Kitchen quality

Ex	Excellent
Gd	Good
TA	Typical/Average
Fa	Fair
Po	Poor

Table 2: Coefficients of Linear Regression Model

(Intercept)	12.0240574
YearRemodAdd	0.0516127
OverallQual	0.1347468
GarageArea	0.0656879
MSZoningRL	0.0609307
KitchenQualFa	-0.0322005
KitchenQualGd	-0.0468608
KitchenQualTA	-0.0734401
logBsmt	0.0363315