Predicting House Prices

November 2018

Kent Burgess, Esther Chang

Agenda

- Research Overview
- Background & Data Description
- Data Engineering
- Exploratory Data Analysis
- Models
- Results
- Conclusion & Future Studies

Research Overview

Objective: To predict the sale price given number of house features

Tools: Machine Learning algorithms in R

Data: Two sets of data - train data (with sale price) and test data (without) - provided by Kaggle

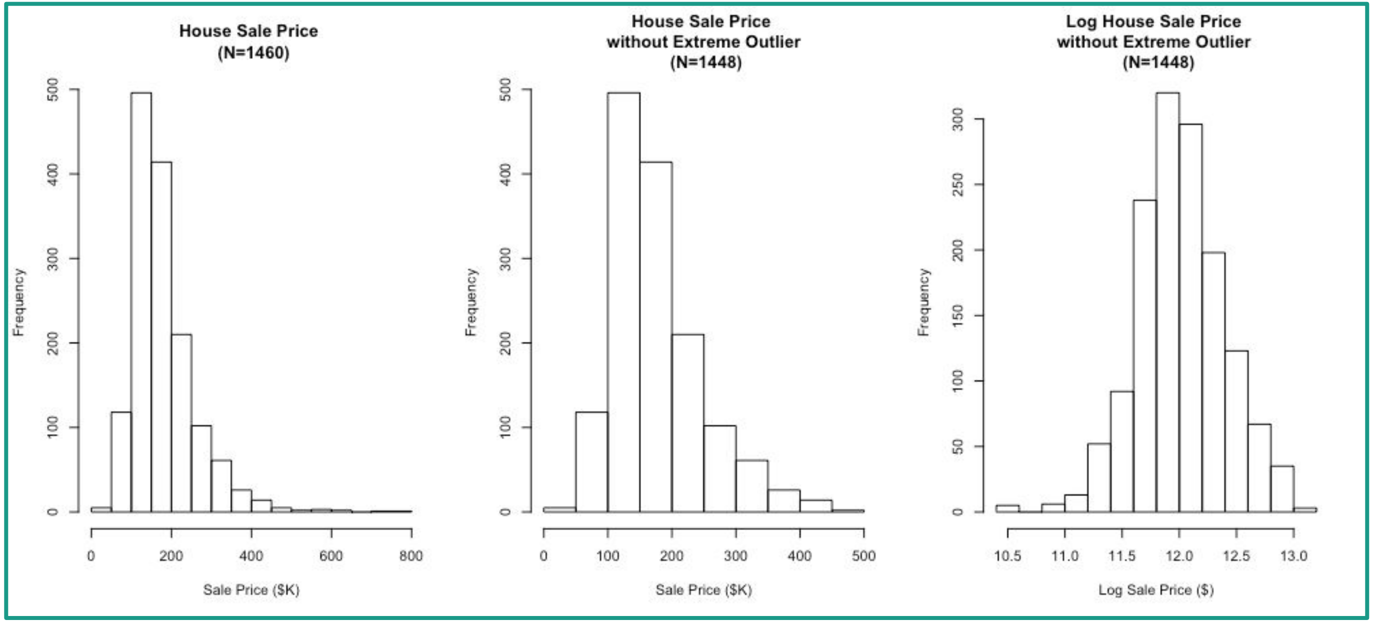
Research Questions:

- Which model scored the best in predicting sale price for the test data?
- How similar or different was the test scores compared to the training/test error in cross validation?
- Was there any recurring predictors in fitted models?

Background & Data Description

- Ames, Iowa dataset contains 2930 residential property observations between 2006 and 2010
- 80 variables that focus on the quality and quantity of the physical features of the house
- Added 9 additional variables:
    - OverallGrade, GarageGrade, ExterGrade, KitchenScore, FireplaceScore, GarageScore, PoolScore, TotalBath, TotalSF
- 15 variables were "ordinal"

Data Engineering: Sale Price



Mean	180,921
SD	79,442.5
Min	34,900
1Q	129,975
Median	163,000
3Q	214,000
Max	755,000
No. Outliers	61
No. Extr Outliers	12

Data Engineering: Ordinal vs. Categorical

Treatment of 15 ordinal variables by comparing AIC

Example:

PoolQC: Pool quality
Ex Excellent - 4
Gd Good - 3
TA Average/Typical - 2
Fa Fair - 1
NA No Pool - 0?

ExterQual: Evaluates the quality of the material on the exterior (no missing)
Ex Excellent -5
Gd Good - 4
TA Average/Typical - 3
Fa Fair - 2
Po Poor - 1

Factor Model AIC	Numerical Model AIC
504.13	570.25

BsmtQual: Evaluates the height of the basement
Ex Excellent (100+ inches) - 5
Gd Good (90-99 inches) - 4
TA Typical (80-89 inches) - 3
Fa Fair (70-79 inches) - 2
Po Poor (<70 inches - 1
NA No Basement - 0

Factor Model AIC	Numerical Model AIC
595.76	598.58

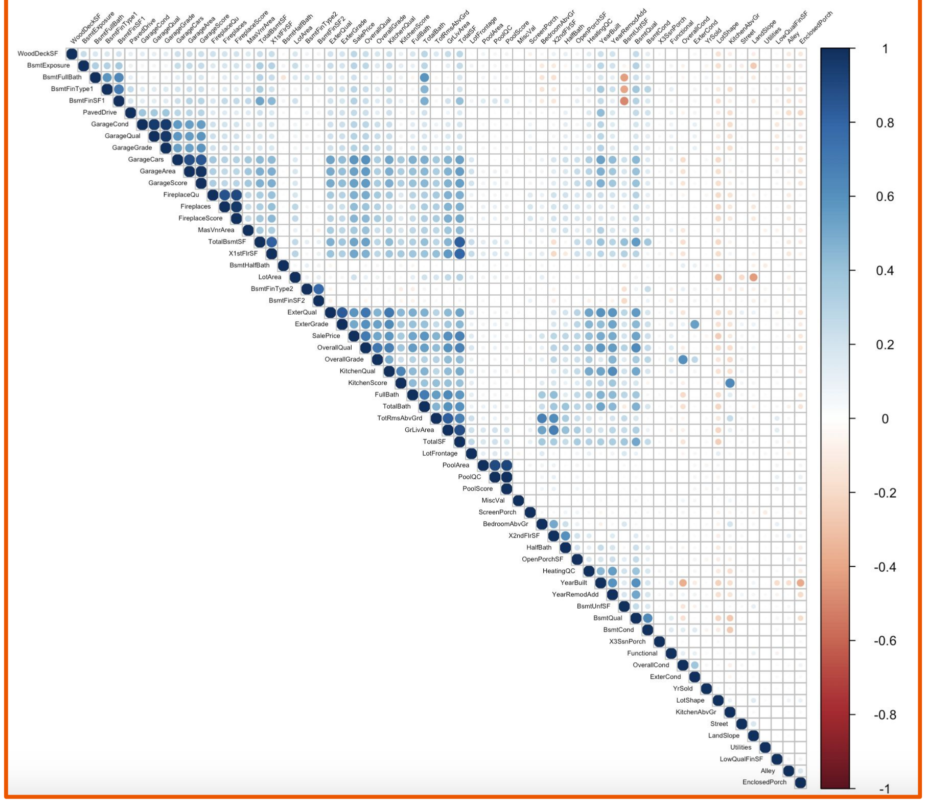
More Data Description: Missing & Outliers

After some initial and intuitive recoding, we had:

Train				Test			
Variable	**No. NAs**	**Variable**	**No. Extreme Outliers**	**Variable**	**No. NAs**	**Variable**	**No. Extreme Outliers**
GarageYrBlt	81	EnclosedPorch	208	FireplaceScore	730	EnclosedPorch	251
Electrical	1	GarageGrade	169	GarageGrade	78	ScreenPorch	140
		BsmtFinSF2	167	GarageScore	78	KitchenAbvGr	66
		ScreenPorch	116	GarageYrBlt	78	MiscVal	51
		BsmtHalfBath	82	MSZoning	4	MasVnrArea	29
		KitchenAbvGr	68	BsmtFullBath	2	LotArea	19
		MiscVal	52	BsmtHalfBath	2	LowQualFinSF	14
		LotArea	34	Functional	2	OpenPorchSF	14
		MasVnrArea	28	TotalBath	2	X3SsnPorch	13
		...		...		...	

Exploratory Data Analysis: Correlation

Top 10 Variables against log Sale Price wo outliers		AIC	R Squared (adjusted)
1	**OverallQual**	-182.18	0.65
2	**Neighborhood**	185.34	0.56
3	**GrLivArea**	456.35	0.46
4	**GarageCars**	474.86	0.45
5	**ExterQual**	504.13	0.44
6	**BsmtQual**	518.41	0.44
7	**KitchenQual**	542.87	0.43
8	**GarageArea**	581.58	0.41
9	**GarageFinish**	647.55	0.38
10	**GarageYrBlt**	652.27	0.29



Modeling Process: Variable Selection

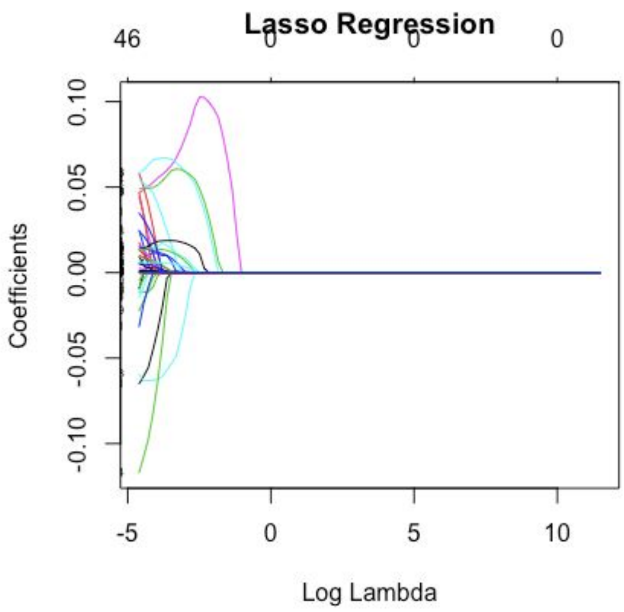
I. Full Model - Kitchen Sink
II. Selection based on Stepwise Model - backward from Kitchen Sink
III. Selection based on Stepwise Model - backward from non-missing variables: "Street" "LotConfig" "LandSlope" "Neighborhood" "Condition1" "Condition2" "BldgType" "RoofMatl" "ExterQual" "Foundation" "Heating" "HeatingQC" "CentralAir" "PoolQC" "SaleType"
IV. Selection based on Correlation and VIF: "OverallQual" "GrLivArea" "ExterQual" "TotalBath" "GarageScore" "X1stFlrSF" "FullBath"

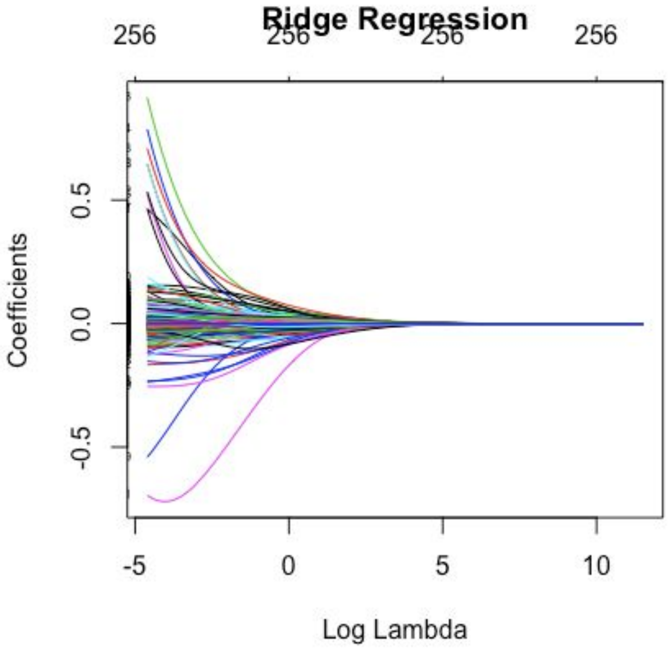
Modeling Process: Modeling Method

I. Ridge & Lasso & Elastic Net Regression (with varying hyperparameter)
II. Regression Trees
III. Random Forest
IV. Gradient Boosting

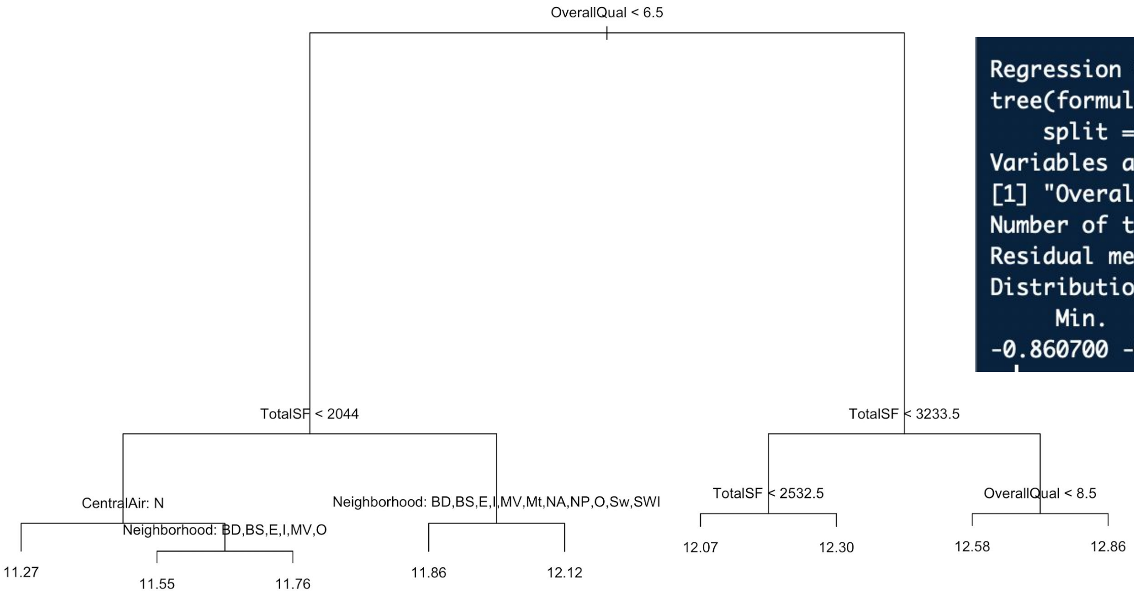
Lasso and Ridge Models

-lambda =.3054





Regression Tree Model




```
Regression tree:
tree(formula = SalePrice ~ ., data = houses, subset = treetrain,
    split = "deviance")
Variables actually used in tree construction:
[1] "OverallQual"  "TotalSF"      "CentralAir"   "Neighborhood"
Number of terminal nodes:  9
Residual mean deviance:  0.03521 = 35.63 / 1012
Distribution of residuals:
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-0.860700 -0.106900  0.006208  0.000000  0.113300  0.657500
```

Results - Aggregation

Model	Kaggle Score	Model RMSE
Gradient Boosting (alpha=0.001)	**0.13536**	**0.1646208**
Regularized Regression (alpha = 0.1)	0.14698	0.1587
Regression Tree Model	0.22540	0.2190
Ridge Ordinal Model	0.23260	0.2013
Lasso Ordinal Model	0.26168	0.2189

Conclusion & Future Studies

- Random Forest Model performs the best in prediction
- RSME was the lowest for regression models but yielded worst Kaggle score
- Despite the concern of multicolinearity, most of top correlated variables were in the best models

- Improve scores using stacking and other machine learning optimization methods
- “Smart” imputation on missing values would improve the fit
- More automation of the model fitting process with ranges of hyperparameters and variable selection iterations to come up with the best model would be ideal

Questions? Thank you!