Predicting House Prices

November 2018

Kent Burgess, Esther Chang

Agenda

- Research Overview
- Background & Data Description
- Exploratory Data Analysis
- Modeling Process
- 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

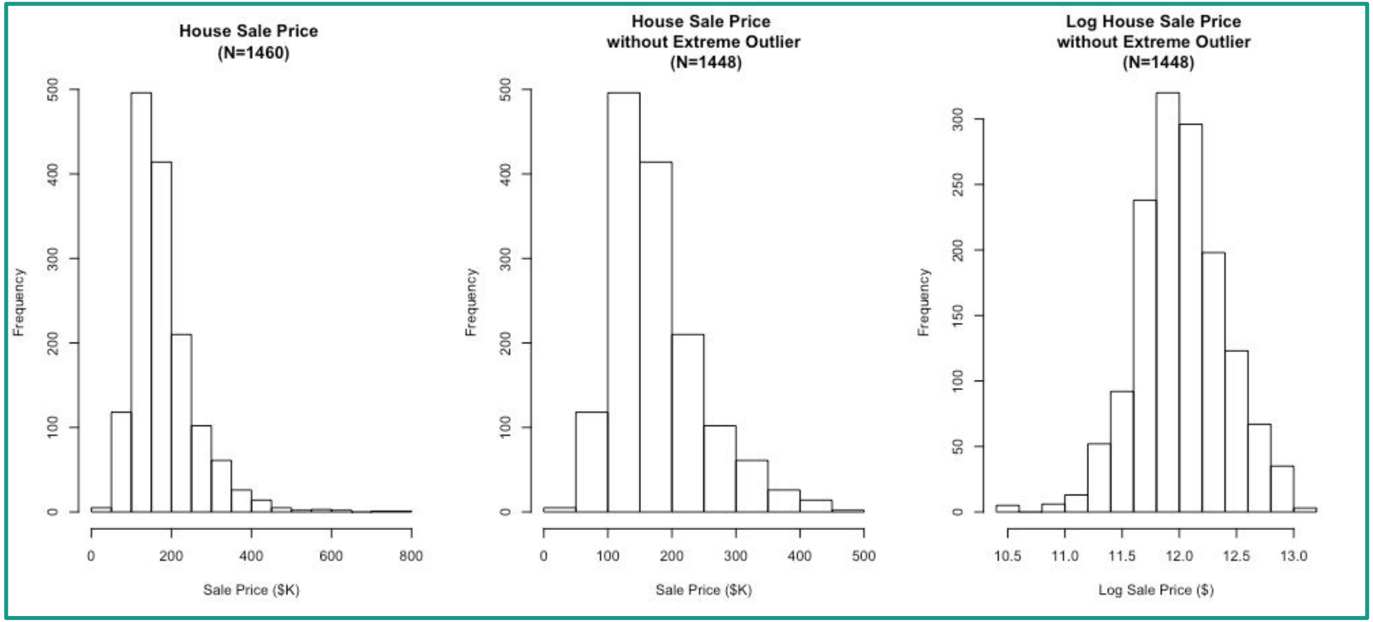
Research Questions:

- 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
    - 23 nominal - categorical feature identifying types of dwellings materials or environmental conditions
        - BldgType, HouseStyle, RoofStyle
    - 23 ordinal - a categorical feature providing ratings of the various items associated with the house
        - ExterCond, BsmtCond, KitchenQual
    - 14 discrete - quantifying the number of items included in the house
        - Fireplaces, FullBath
    - 20 continuous -related to the various area dimensions of the houses and lots, and specific rooms of the house
        - WoodDeckSF, LotArea
- 80 variables that focus on the quality and quantity of the physical features of the house

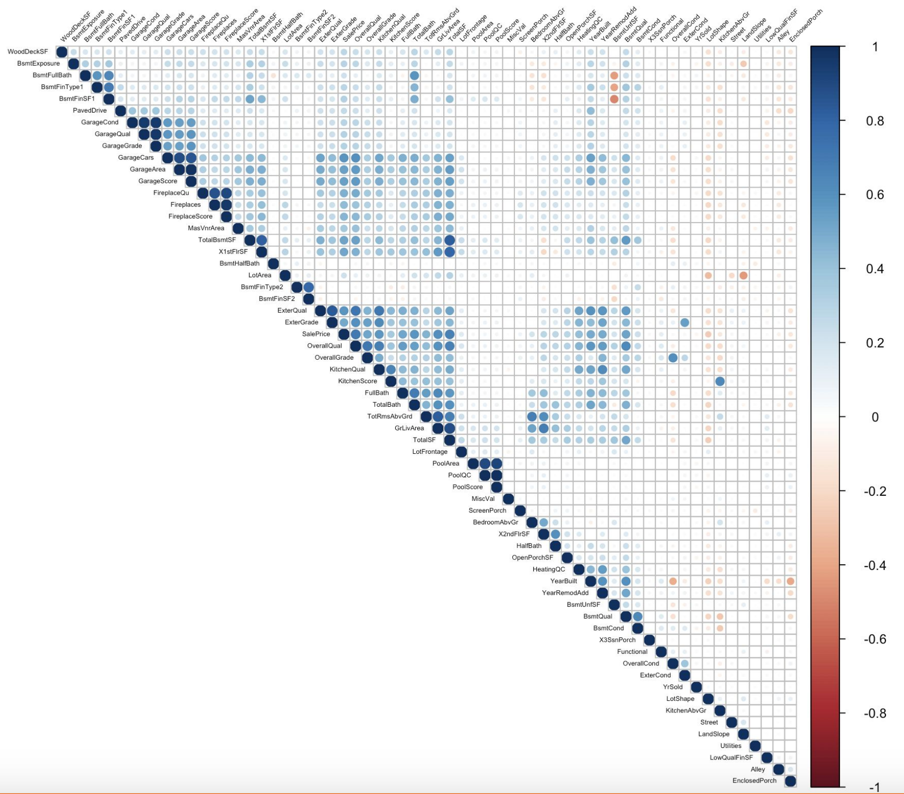
Exploratory Data Analysis: Sale Price



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
No. Outliers (Log)	28

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



Exploratory Data Analysis: Ordinal vs. Categorical

Treatment of 15 ordinal variables by comparing AIC - 13 assigned as categorical and 2 ordinal/numerical

Example: *ExterQual: Evaluates the quality of the material on the exterior*

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

Factor Model AIC		Numerical Model AIC
504.13		570.25

Exploratory Data Analysis: Missing & Outliers

After data cleaning and recoding, we were left with following variables:

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

Modeling Process: Variable Selection

I. Selection based on Stepwise Backward:

"Street" "LotConfig" "LandSlope" "Neighborhood" "Condition1" "Condition2" "BldgType" "RoofMatl" "ExterQual" "Foundation" "Heating" "HeatingQC" "CentralAir" "PoolQC" "SaleType"

II. Selection based on Correlation and VIF:

"OverallQual" "GrLivArea" "ExterQual" "TotalBath" "GarageScore" "X1stFlrSF" "FullBath"

III. Selection based on Complete Dataset (No Imputation):

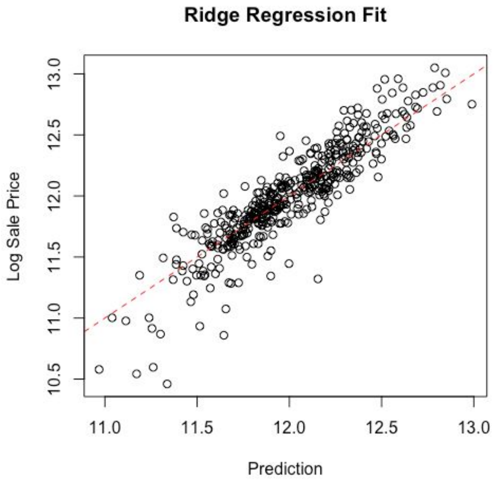
Modeling Process: Modeling Method

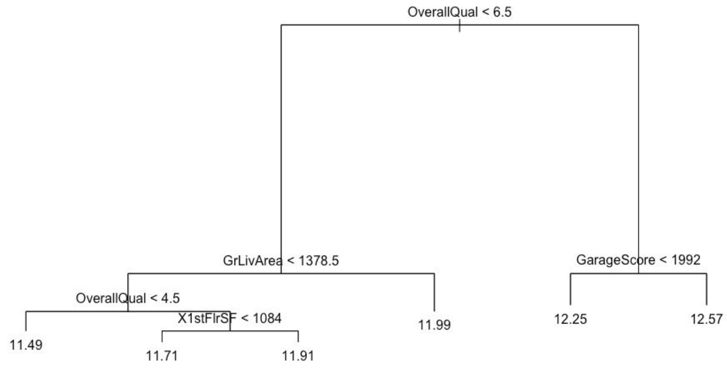
I. Ridge Regression
II. Lasso Regression
III. Regression Trees
IV. Bagging & Random Forest

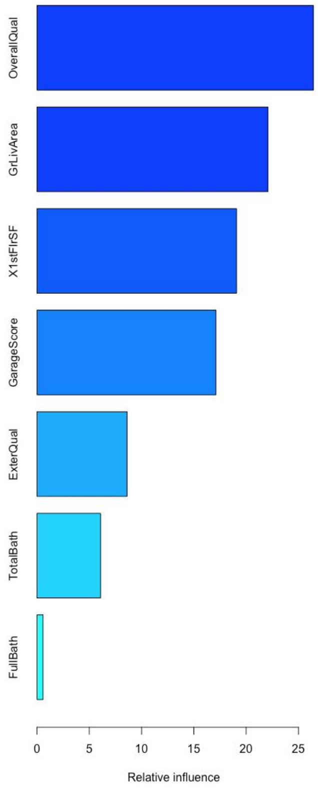
Results - Aggregation

Variable Selection	Model	Kaggle Score	Train Error (MSE)	Test Error (MSE)
I	**IV**	**13.54%**	**2.71%**	**5.51%**
II	III	22.54%	4.35%	6.25%
I	III	26.17%	5.80%	8.30%
I	II	31.58%	4.45%	4.98%
I	I	32.98%	3.79%	4.79%
II	II	52.13%	2.52%	2.61%
II	I	52.98%	2.43%	2.47%

Results - Models







Conclusion & Future Studies

- Random Forest Model performs the best in prediction
- Training/Test Error was lowest for regression models but yielded worst Kaggle score
- Variables like OverallQual, GrLivArea and ExterQual showed up the most among our models

- “Smart” imputation on missing values would improve the fit
- Automating the model fitting process with ranges of parameters and variable choices to come up with the best model would be ideal

Questions? Thank you!