USING REGRESSION TO PREDICT HOUSE SALES

BY DAVID MARMOR

PROBLEM STATEMENT

- Can we use house data to predict the price a house will sell for in our city of Ames?
- People who want to sell their homes can have an idea of how much they will get from the sale.
- People who want to buy a home will have an idea of what kind of offer they should make on a house to get a fair deal.
- We will have two models. One model will mostly be used for prediction. The other can also use inference to figure out how specific features effect the house price.
- Regression and R-squared. Regularization.

DATA

- 2,051 houses sold from 2006 -2010
- 81 features
- Target feature: Sale Price
- Variety of variable types

DATA CLEANING

- Create dummy variables
- Fill in NAs
- Run correlation scores of independent variables
- Remove variables with less than a 0.4 correlation with Sale Price.

CORRELATION PLOT

FEATURE CHECKING

- Of the 14 features that aren't sale price 8 made it into the model.
- 5 were removed for multicollinearity
- One was removed for not having a linear relationship with sale price.

VARIABLES IN MODEL

- Overall Quality – ordinal variable. Material and finish of the house rated 1-10.
- Gr Liv Area – above grade ground living area in square feet
- Garage Area – Garage size in square feet
- Total Bsmt SF – basement area in square feet
- Mas Vnr Area – Masonry Veneer area in square feet
- Fireplaces – number of fireplaces in house
- NridgHt – is the house in Northridge Heights area
- BsmtFin SF – finished area of basement in square feet. Not correlated with total bsmt sf.

PREDICTIVE LINEAR REGRESSION

R squared score on training data

0.852

R squared score on testing data

0.837

Slightly overfit. Overall model is explaining most of the variation.

RIDGE AND LASSO MODELS

Ridge training score

0.852

Ridge testing score

0.837

Lasso training score

0.852

Lasso testing score

0.837

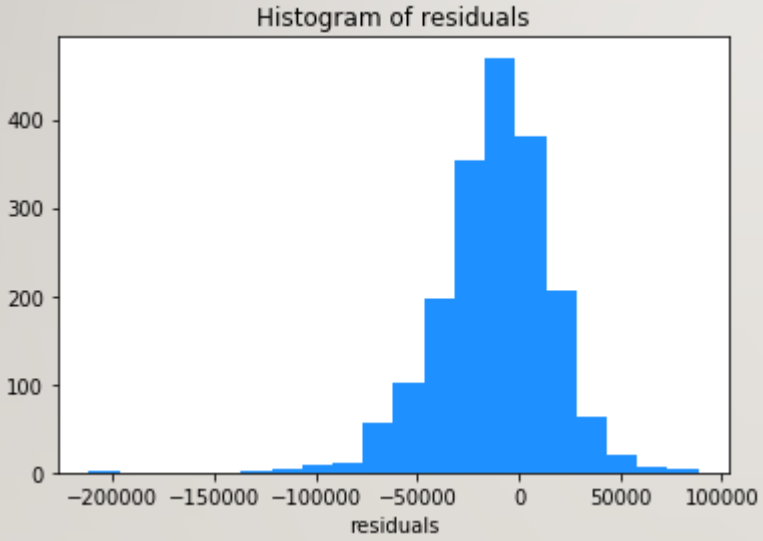
Almost identical. Regularization did not help.

INFERENTIAL MODEL

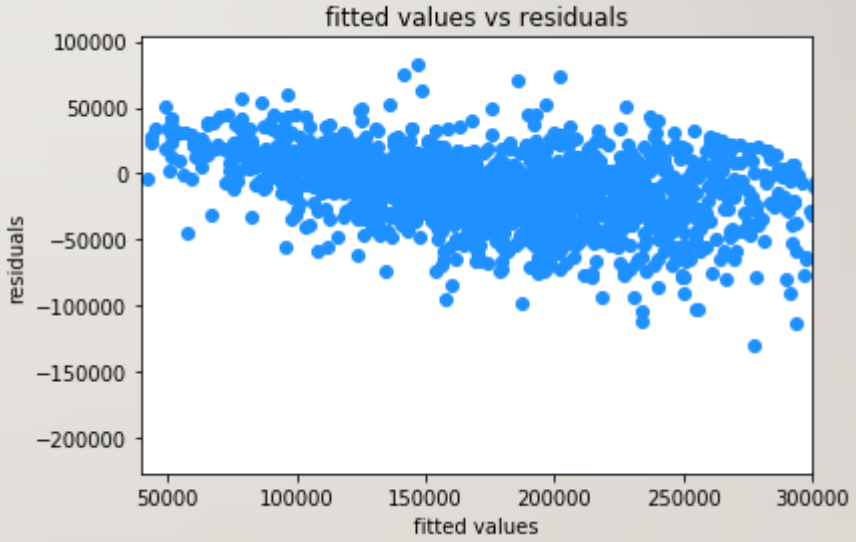
- The predictive model did not meet line conditions.
- Only for houses that sold for under $300,000
- Not as predictive as the predictive model.
- But it meets line condition.
- No multicollinearity.
- Linear relationship between the variables.
- Independence

LINE ASSUMPTIONS

- Residuals normal.



Residuals vs fitted values



COEFFICIENTS

```
[('Overall Qual', 23228.7819230184),
 ('Gr Liv Area', 16496.41856123732),
 ('Garage Area', 8804.292856093094),
 ('Total Bsmt SF', 7197.802898641518),
 ('Mas Vnr Area', -442.31857707145264),
 ('Fireplaces', 4257.509230580838),
 ('NridgHt', 2905.440282195942),
 ('BsmtFin SF 1', 7990.760919016816)]
```

Variables are scaled so coefficients represent the change in dollars all else being constant for an increase in one standard deviation in that variable

Overall Quality and ground living area have the biggest coefficients.

CONCLUSIONS

- We are able to model house sale prices in Ames.
- We can explain most of the variation in house sale prices.
- The inferential model allows us to know what factors impact home sales for houses under $300,000.
- Overall quality of the house and square feet are the most important factors.
- Provide the models as resources for home sellers and buyers in the area.

FURTHER WORK

- Inferential Model for houses that sell for over $300,000
- Model taking into account more factors particularly time.
- Can model be generalized for other cities? If not other cities may need to create their own models.

THANK YOU

- Any questions?