Machine-Learning Analysis:

I used linear regression techniques and random forest regressions to analyze residential housing prices. Models were trained and tested on prices between $35,000 and $2,500,000.

For linear models, log base 10 transformations with ridge regression improved the baseline model. Subnbhd was the only geographical variable in the baseline models. The best model with outliers had high dimensional geographical variables removed and one collinear variable removed. Removing high VIF variables did not improve the performance. The model without outliers performed the best with the same variables removed.

BEST LINEAR MODELS FOR GIVEN SCENARIOS:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Type | Variables removed | MAPE | RMSE | R^2 train | R^2 testing |
| Baseline | 'QUADRANT', 'WARD', 'NBHD', 'CENSUS\_BLOCK', 'ZIPCODE', 'SQUARE' | 41.41 | 194645.07 | 0.7067 | 0.7177 |
| Improved baseline  -ridge regression  -log(‘PRICE’) | 'QUADRANT', 'WARD', 'NBHD', 'CENSUS\_BLOCK', 'ZIPCODE', 'SQUARE' | 28.84 | 198741.51 | 0.7650 | 0.7517 |
| Best model with outliers  -ridge regression  -log(‘PRICE’) | 'KITCHENS', 'SUBNBHD', 'CENSUS\_BLOCK', ‘SQUARE’ | 27.48 | 191419.47 | 0.7816 | 0.7702 |
| Best model without outliers  -log10(‘PRICE’)  -Ridge Regression | 'KITCHENS', 'SUBNBHD', 'CENSUS\_BLOCK', 'SQUARE' | 23.73 | 148619.06 | 0.8370 | 0.8351 |

* MAPE- Mean Absolute Percentage Error
* RMSE- Root Mean Squared Error

**Outlier Descriptions, compared to original data (1.7% of data):**

* Lower or higher prices
* sold during different times
* lower or higher landarea
* more built in 1940s-1960s
* remodeled 2005-2010
* fewer rooms/bedrooms
* more in NE, Wards 2/6
* Columbia Heights, Petworth, Brookland, Dearwood, Chevy Chase, Mount Pleasant, Congress Heights

**Linear model assumptions:**

Best model with outliers:

-Linear

-Not normal distribution (fat tails)

-Heteroskedastic (small issue)

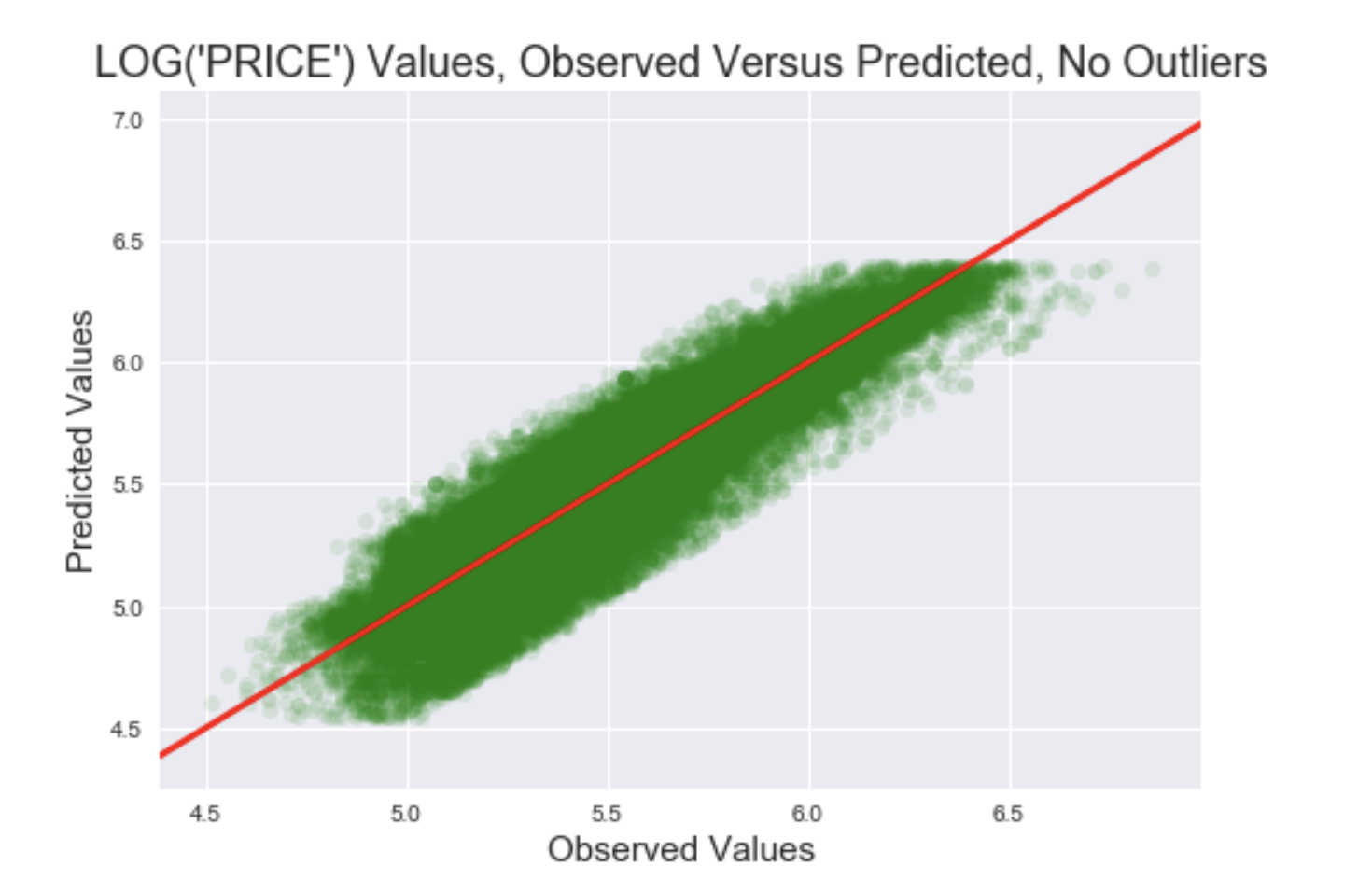
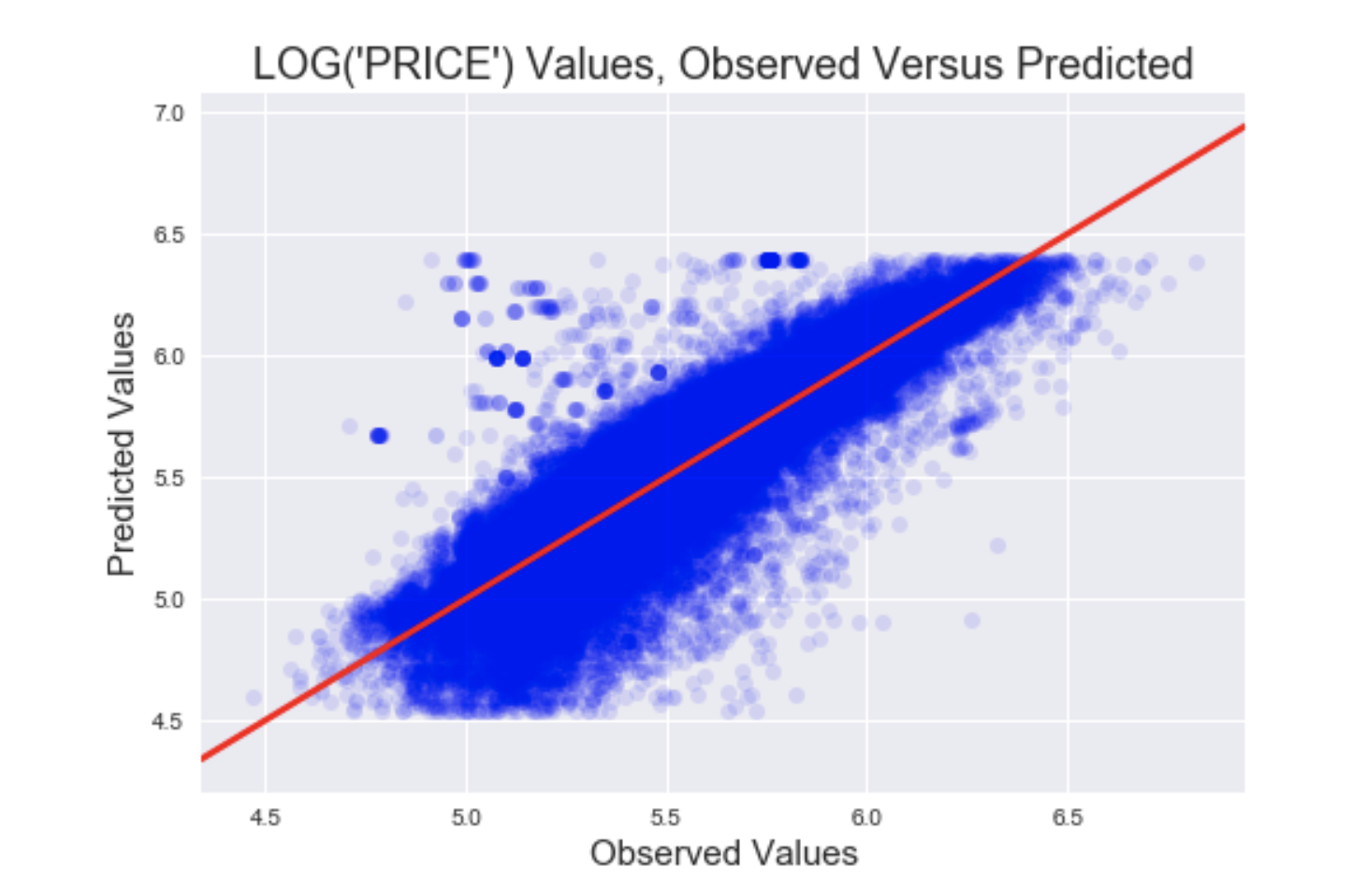
Best model without outliers:

-Linear (small issue)

-Closer to normal distribution

-Heteroskedastic (small issue)

Because our data doesn’t fit some assumptions for linear models, perhaps nonlinear models will perform better.



BEST RANDOM FOREST MODELS:

|  |  |  |
| --- | --- | --- |
| Type | MAPE | RMSE |
| Random Forest | 18.60 | 117445.79 |
| Adaptive Boosting | 63.25 | 259032.33 |
| Gradient Boosting | 25.48 | 136350.93 |
| Log(PRICE)  Random Forest | 17.01 | 121204.73 |
| Log(PRICE) AdaBoost | 41.56 | 281095.98 |
| Log(PRICE) GradientBoost | 20.55 | 152555.08 |

Least influential variables:

EXTWALL, INTWALL, EXTWALL, SALE\_NUM, STYLE, ROOF, HEAT, CNDTN

Removing the least influential variables decreased model performance slightly.

The Log(PRICE) Random Forest Model performed the best at predicting PRICE with a 17.01 mean absolute percentage error and 121204.73 RMSE. This model also performed better than linear models.

* We can predict price with a mean absolute percentage error of 17.01%.
* 80% of the time we can predict price between 80% and 125% of its value.

