CAPSTONE PROJECT 1

PREDICTIVE MODELING OF D.C. RESIDENTIAL HOUSING PRICES

Julia Sheriff

Springboard | Data Science Career Track

1. INTRODUCTION

The purpose of this capstone project is to use predictive modeling to determine residential property values in Washington D.C. After acquiring the data and conducting exploratory data analysis, I proceeded to use linear regression and random forest regression models to find the top-performing model, which was a random forest regressor. "80% of the percentage errors show a maximum positive percentage error of 24% and a minimum negative percentage error of 20.27%, which measure excess and deficient worst-case scenarios for a price predicted by the best model." A timeseries model would be beneficial for improving precision and predicting future residential prices.

2. APPROACH

2.1 DATA ACQUISITION AND WRANGLING

The data is from a Kaggle competition, "D. C. Residential Properties", provided by Open Data DC at https://www.kaggle.com/christophercorrea/dc-residential-properties.

I removed many columns due to redundancy or limitations inherent in the data:

- 'UNNAMED', equivalent to the index
- 'STATE' and 'CITY', identical for each sale
- 'FULLADDRESS', values would be hard to fill accurately; 'NATIONALGRID', 'LATITUDE' and 'LONGITUDE' could function as a substitute
- 'X' and 'Y', analogous with 'LATITUTDE' and 'LONGITUDE'
- 'HALF_BATHROOM', combined with 'BATHROOM'

OUTLIERS:

I removed the following erroneous datapoints.

- 'STORIES': 250, 275, 826
- 'YR_RMDL' (year remodeled): 20
- 'STRUCT' (structure): 'default'

Other numerical features had skewed data, so I chose boundaries appropriate for the nature of the data.

- 'GBA', < Q1 - 1.5*IQR, > Q3 + 6* IQR
-'LIVING_GBA', < Q1 - 1.5*IQR, > Q3 + 6* IQR
-'LANDAREA', < Q1 - IQR, > Q3 + 10* IQR
-'PRICE', < Q1 – IQR, > Q3 + 8* IQR

All categories for categorical data seemed reasonable.

MISSING VALUES:

-Numerical data:

-Grouped features by neighborhood

-Imputed with a rolling mean: window of 500. No column had more than 1% of missing data after imputation.

-Dropped remaining rows.

-Categorical data:

-Grouped features by neighborhood.

-Replaced NaN values with the mode of that column in that neighborhood.

-PRICE:

-removed all observations with missing price.

2.2 STORYTELLING AND INFERENTIAL STATISTICS

STORYTELLING:

Residential housing prices in DC are dependent on the geographical location and sale date and housing costs are increasingly unaffordable for local residents. Between 1992 and 2018, prices one-bedroom units and three-bedroom units tripled in price (see Figure 01). Prices increased most rapidly between 1996-2006 and 2011-2016. Different geographical wards have had different property values and growth.

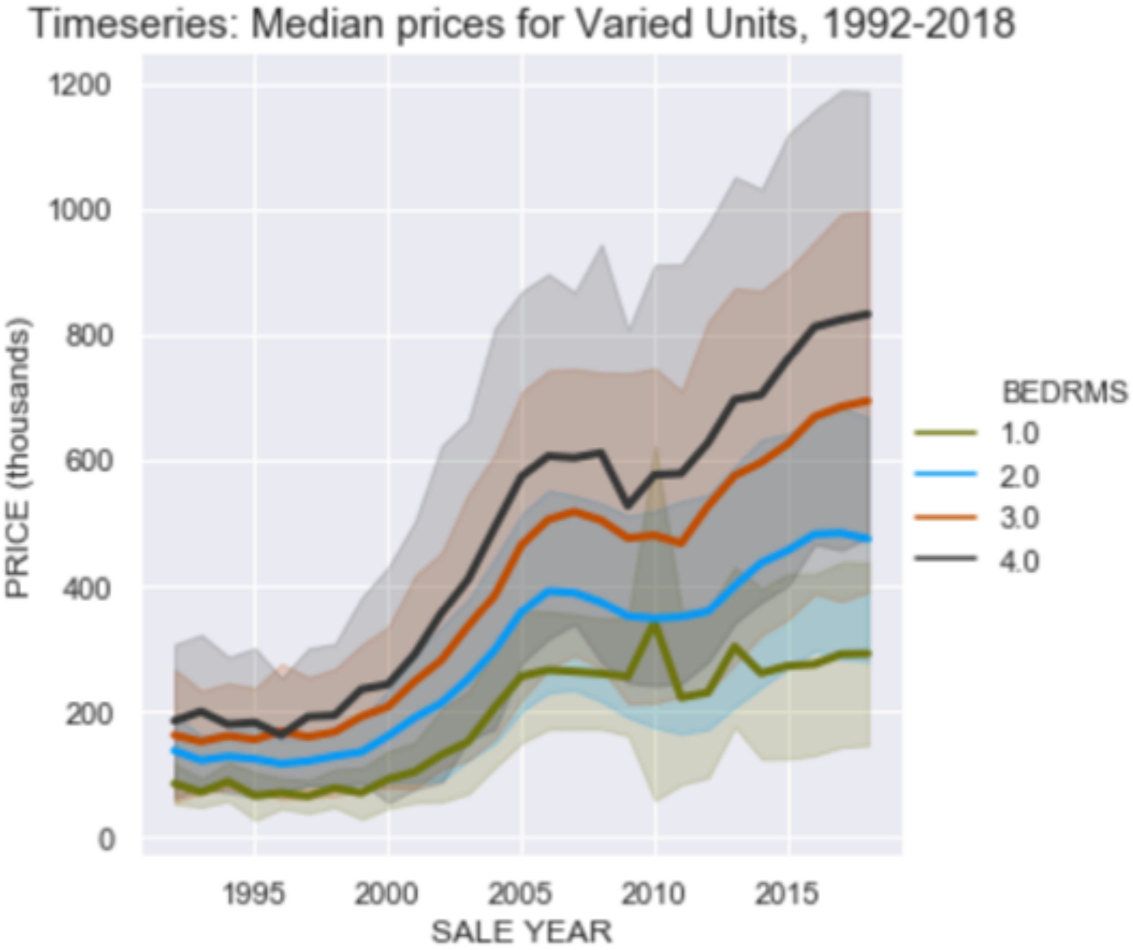


Figure 01

Wards two and three had the greatest range in property values. Ward six had the most consistent growth in property values from 2008-2018, and the highest number of sales. Wards seven and eight had the lowest property values and the lowest number of sales from 1992-2016, but had a dramatic increase in sales between 2016 and 2018. Housing is most affordable for local residents in the southeast, wards seven and eight (see Figure 02). Housing is somewhat more affordable for local residents in the northeast, wards two and three. Properties are consistently unaffordable in central DC.

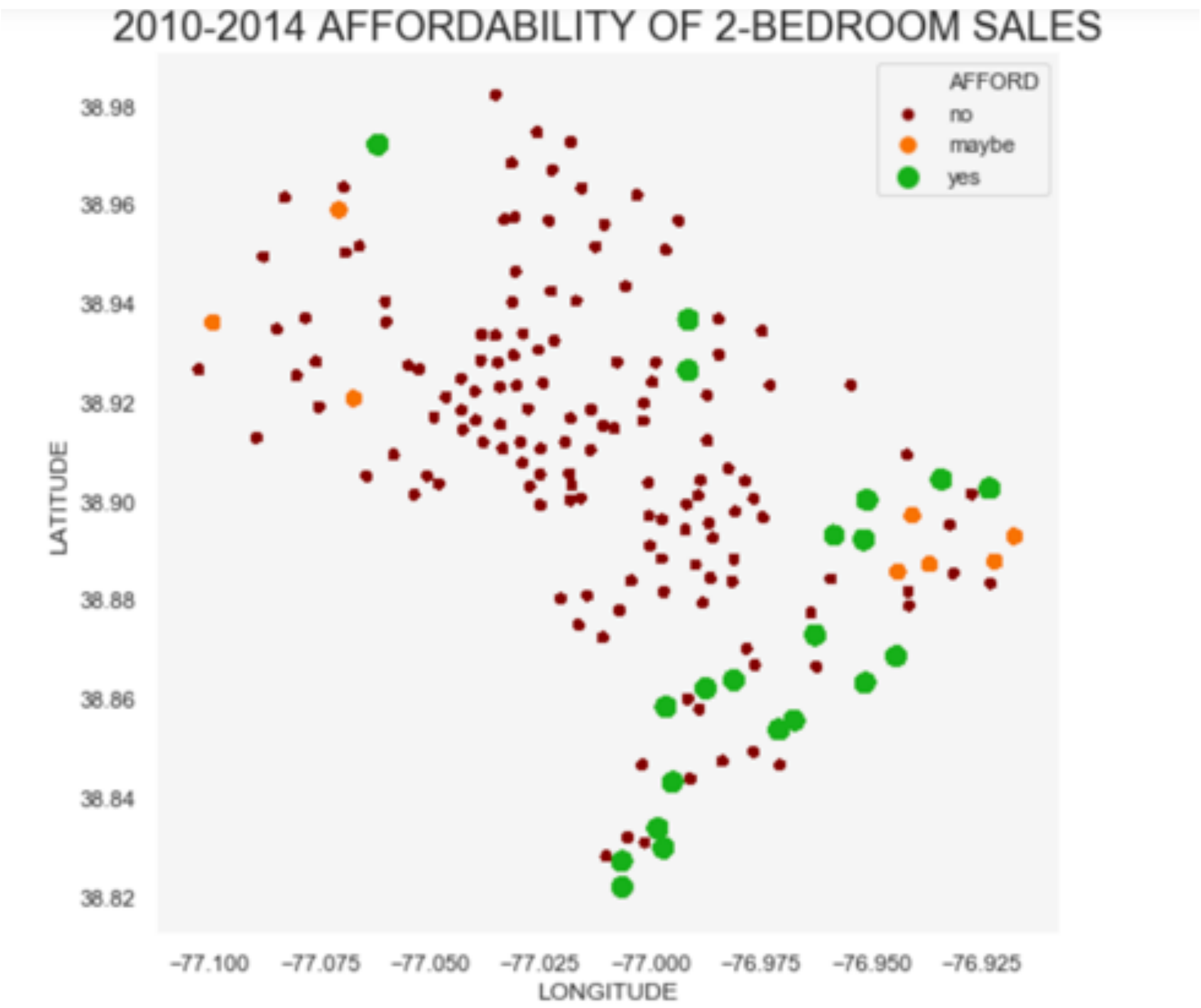


Figure 02

INFERENTIAL STATISTICS:

INTRODUCTION:

This analysis focuses on exploring whether variables have a significant correlation with real estate prices. The price data is not normally distributed, so I used graphs and applied nonparametric significance tests to quantify correlation with price. Several variables were correlated with price, and others were correlated with geographical location.

GEOGRAPHICAL VARIABLES:

Variables that were challenging to work with included census block and square. These are nominal categorical variables with thousands of categories. While the values are numbered, the numbers don't have a consistent correlation to the geographical space they represent. These geographical variables do have an impact on prices similar to quadrant, ward, neighborhood, and sub-neighborhood. I graphed median prices per census block and square to get an idea of their distribution per area (see Figure 03)

The broadest categorical geographical variable is quadrant. After viewing violin and box plots for Quadrant, I applied the Kruskal Wallis test and found that the prices in Quadrant were from different theoretical distributions, so price is affected by quadrant. Ward, neighborhood, sub-neighborhood, census tract, and zip code, which are approximate geographical subdivisions of quadrant, graphically appear to also be significant in predicting price. The difference in prices are more extreme for smaller geographical categories, so these differences may be more significant in predicting price.

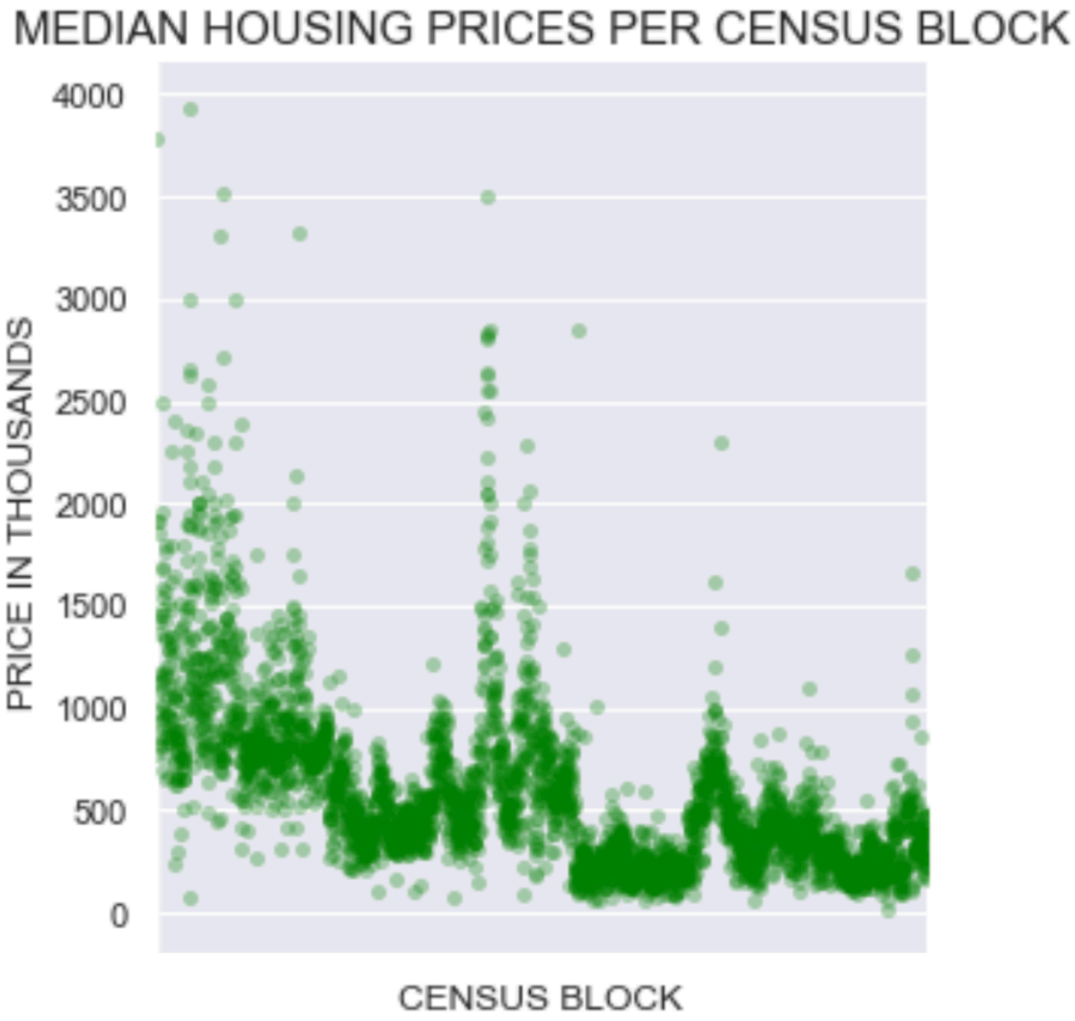


Figure 03

SALE YEAR:

Sale year also had a strong impact on sale price. When graphing the median sale price per year, the positive correlation between sale date and price was clearest (see Figure 04). When I applied Spearman's correlation coefficient for a monotonic relationship across all observations, there was a moderate positive monotonic relationship.

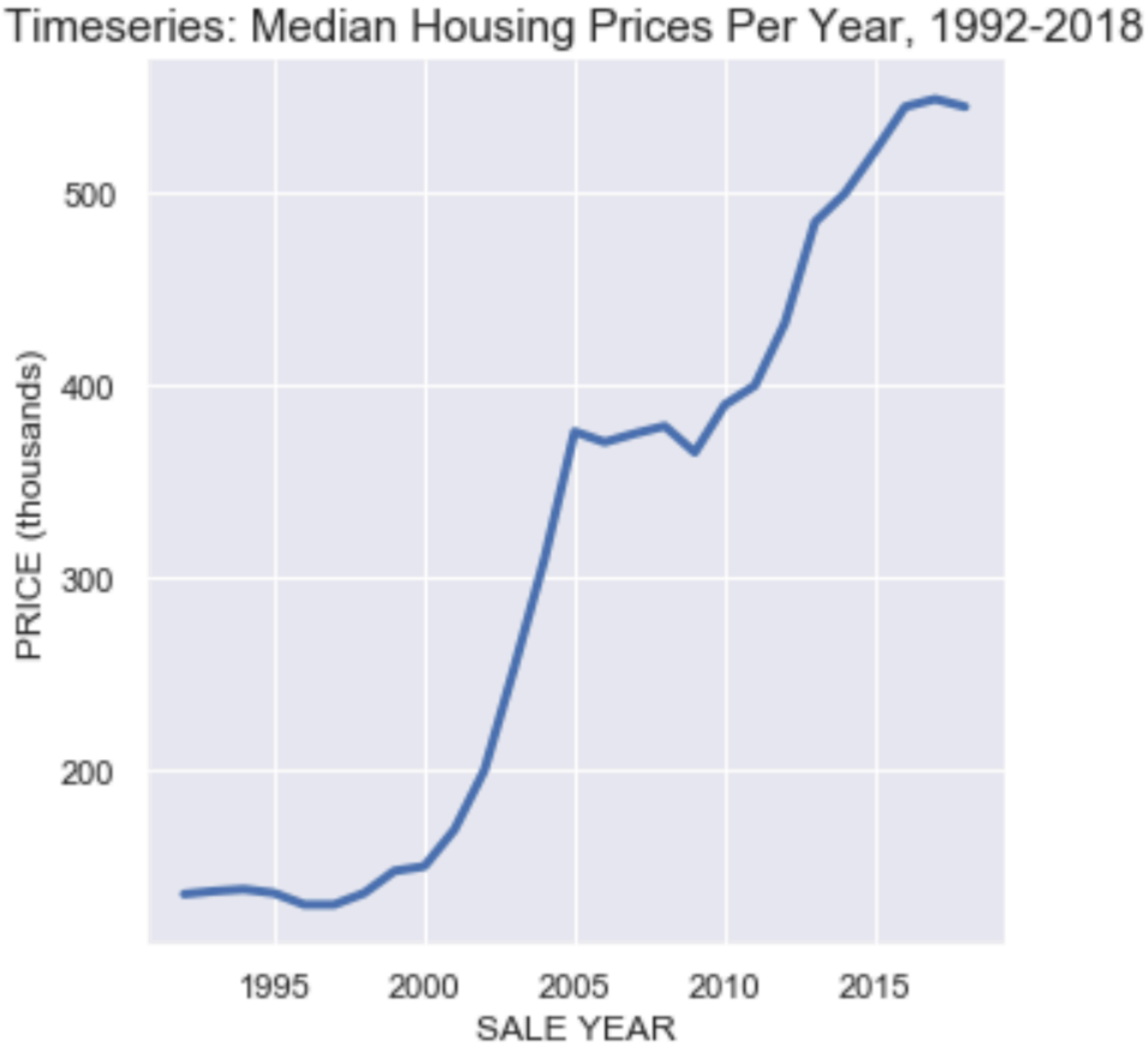


Figure 04

OTHER VARIABLES:

Other significant variables with multiple categories included rooms, stories, and grade. There was a difference in prices between qualified and unqualified buyers according to the Mann-Whitney-Wilcoxon test. I found that the prices for qualified and unqualified buyers were from different theoretical price distributions.

2.3 BASELINE MODELING

I used linear regression techniques and random forest regressors to predict residential housing prices as a function of the features in the dataset. I removed the most extreme values from prices, 1.13% of the dataset, with prices below $35,000 and above $2,500,000.

The linear regression baseline model used only 'SUBNBHD' as a geographical variable. R^2 was .72, the root mean squared error (RMSE) was 194645.07, and the mean absolute percentage error (MAPE) was 41.41%. Residual plots showed that this regression failed tests for linearity, normality, and homoskedasticity, suggesting that we can find a better model

The random forest regression baseline model used 'QUADRANT', 'WARD', 'NBHD', 'SUBNBHD', and 'ZIPCODE' as the geographical variables. I left more geographical variables in this baseline model, since fine-tuning on the linear regression model indicated that including lower-dimension categorical geographical variables improved performance. The Testing RMSE was 117445.79, and the MAPE was 18.60%.

Type	Variables removed	MAPE	RMSE	R^2
Linear Regression, Baseline	'QUADRANT', 'WARD', 'NBHD', 'CENSUS_BLOCK', 'ZIPCODE', 'SQUARE'	41.41	194645.07	0.7177
Random Forest, Baseline	'CENSUS_BLOCK', 'SQUARE'	18.60	117445.79	N/A

Figure 05

2.4 EXTENDED MODELING

For the linear regression model, a log base 10 transformation with ridge regression improved the baseline model. Testing R^2 improved from .72 to .75. The best model including outliers had high dimensional geographical variables removed and one collinear variable removed, ‘KITCHENS’ (see Figure 05). Removing high variance inflation factor (VIF) variables did not improve the performance. The best model including outliers passed the linear model assumptions for linearity. Residuals had fatter tails than a normal distribution, and the model was a bit heteroscedastic. The model without outliers (1.7% of the data) performed the best with the same variables removed, with testing R^2 at .84 (see Figure 06). This model’s residuals were closer to a normal distribution. Because these models didn’t pass the assumptions for linear regression, this suggested that a nonlinear model may yield better results.

LINEAR MODELS:

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

Figure 06

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

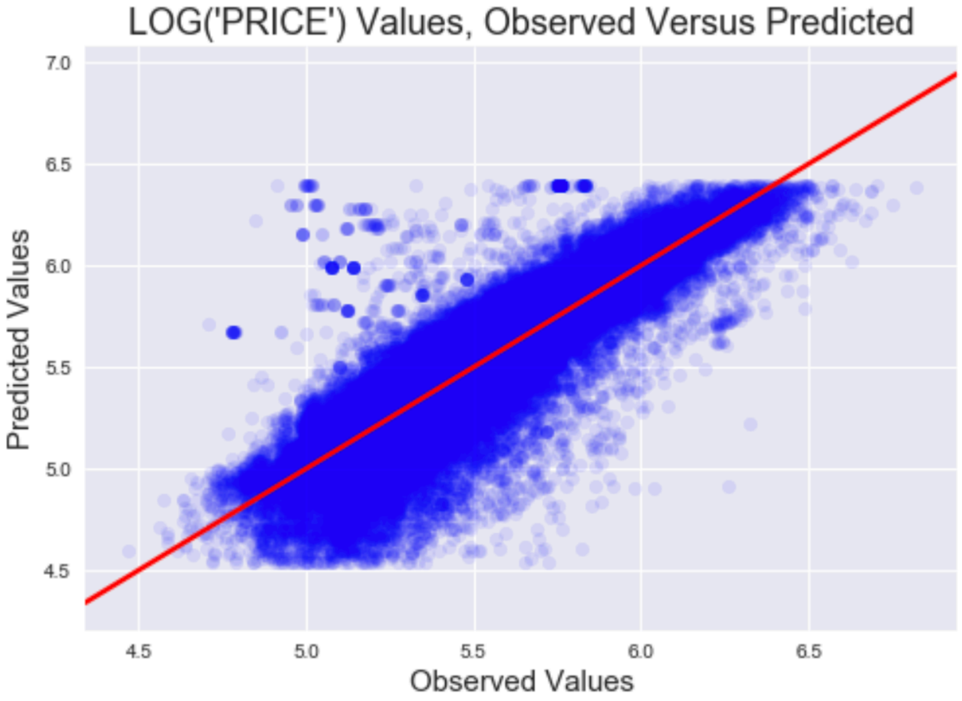


Figure 07

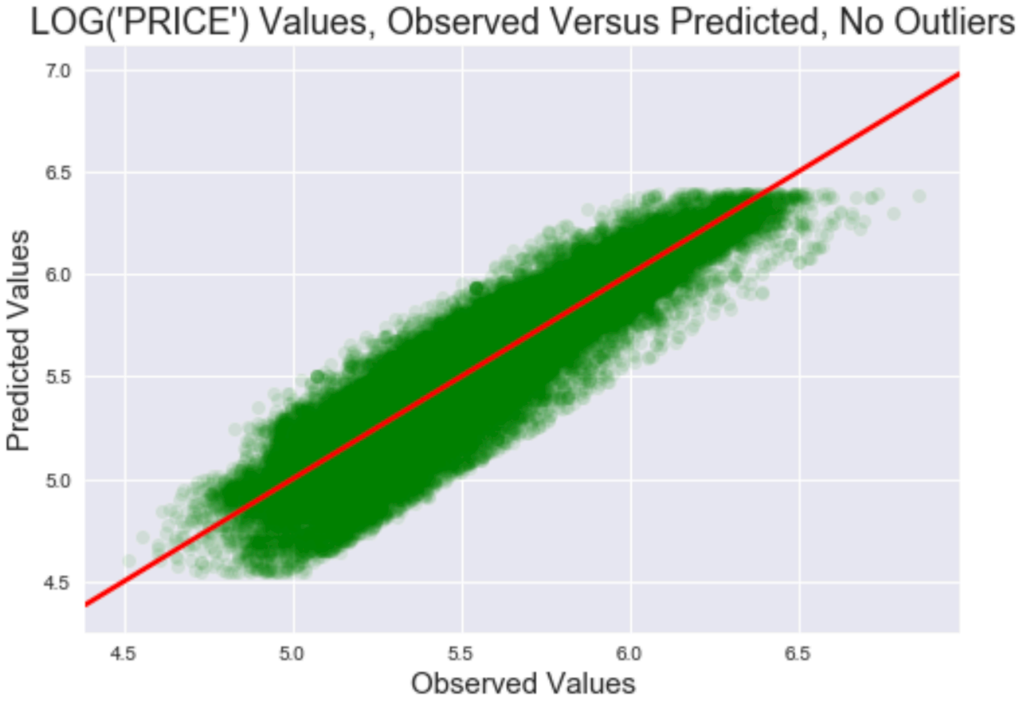


Figure 08

Outlier Descriptions (1.7% of data):

- lower or higher prices
- sold during different times
- lower or higher landarea
- many 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

RANDOM FOREST MODELS:

Predicting log base 10 of prices also yielded better results for the random forest model. Removing low influence features (EXTWALL, INTWALL, EXTWALL, SALE_NUM, STYLE, ROOF, HEAT, CNDTN), did not improve the performance of the model. These features had a minimal effect on the regression model , but removing them did not improve the model. Ada boosting and gradient boosting with parameter tuning did not improve the model. The models below show the best performance after parameter tuning.

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

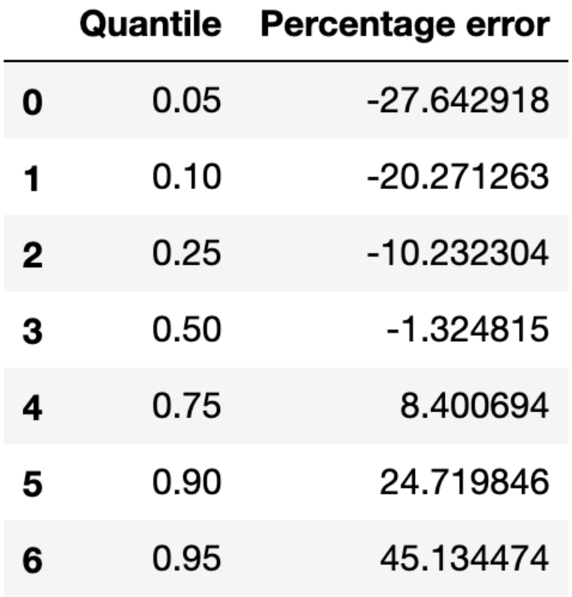
Figure 09

2. 5 ANALYSIS OF RESULT

The Log(Price) Random Forest Model performed the best at predicting the log(Price) with a 17.01 MAPE and 121204.73 RMSE, slightly better than the Price Random Forest Model. The Log(PRICE) Random Forest Gradient Boosting Model performed third best with a 20.55 MAPE and 152555.08 RMSE. While the Linear Log(PRICE) Ridge Regression had the best performance of all linear models, with 27.48 MAPE and 191419.47 RMSE , it performed worse than the best random forest models,.

3. CONCLUSION AND FUTURE WORK

CONCLUSIONS:



	Quantile	Percentage error
0	0.05	-27.642918
1	0.10	-20.271263
2	0.25	-10.232304
3	0.50	-1.324815
4	0.75	8.400694
5	0.90	24.719846
6	0.95	45.134474

Figure 10

We our best model, 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. While this is a significant gain over baseline models with a 41.41 MAPE, the model's general accuracy is limited in precision. The table above shows percentage errors in price prediction, showing predictions too low with negative percentages, and too high with positive percentages.

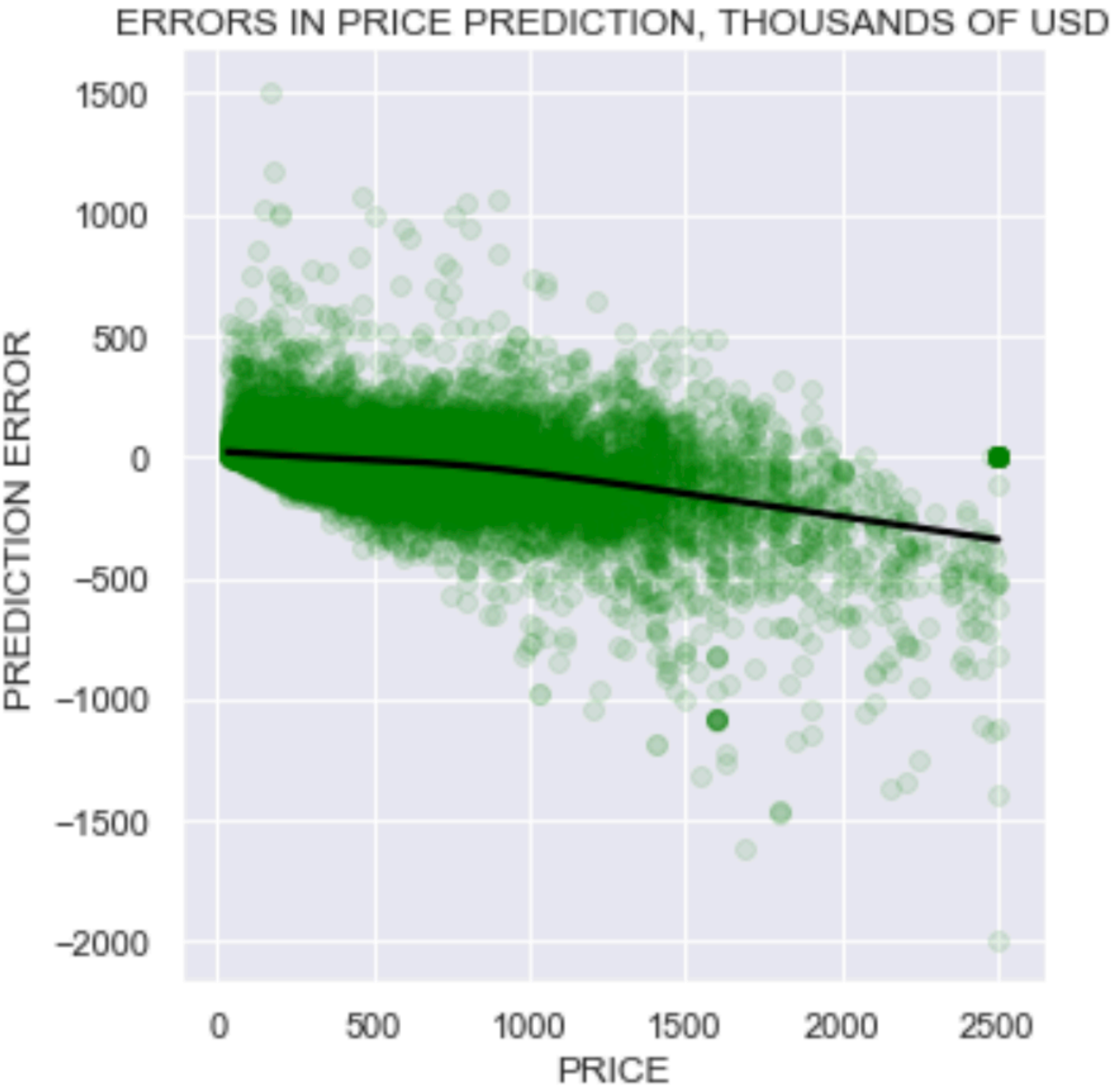


Figure 07

When considering observations with the high prediction error, the model predicted prices higher than observed for low property values, and lower than observed for high property values. There are several outliers shown in this graph, and the graph includes all observations (see Figure 07).

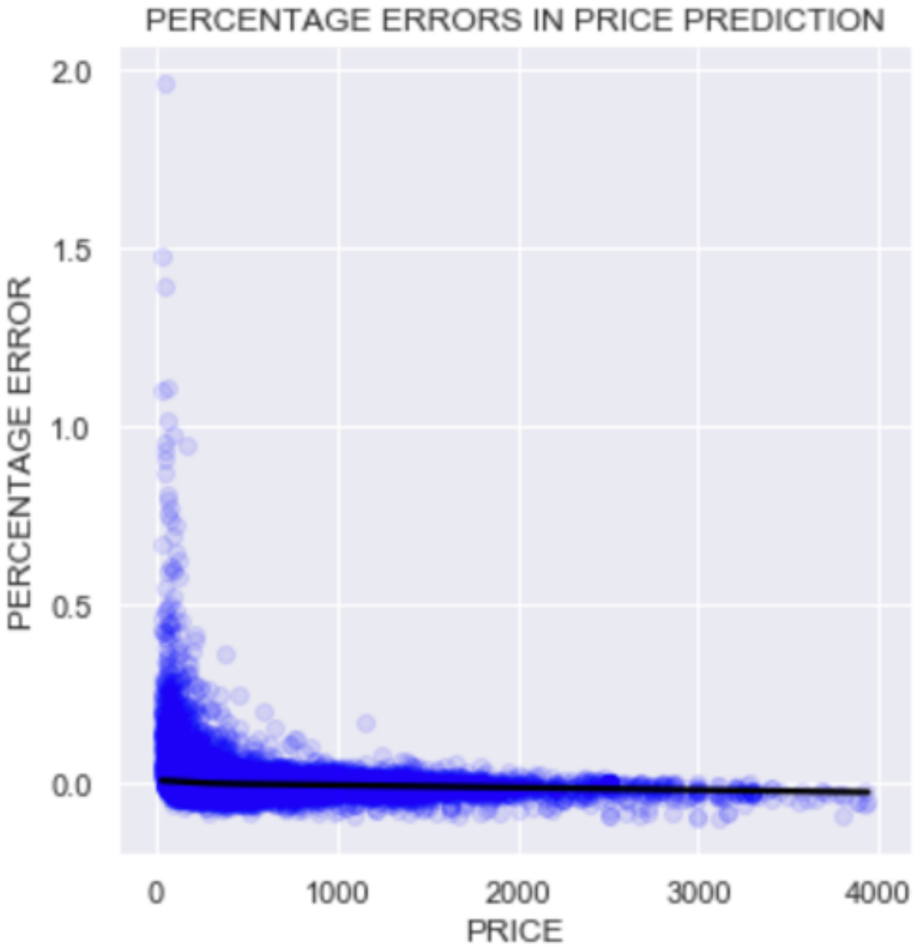


Figure 08

90% of the time our error in price prediction is less than 33.63% of the observed value. The percentage error in price prediction is highest for low property values, where the model tends to predict prices higher than observed (see Figure 08). The percentage error is more sensitive to errors in dollars for low properties because the dollar is compared to a lower observed value.

FUTURE WORK:

The data we have includes information on the sale price of residential properties. This data does not indicate the exact market price, as properties are sometimes sold under value for personal reasons, atypical of the market's general behavior. It is difficult to classify which observations were sold approximately at market price, and which were sold under value.

The statistical and machine learning applications used in this project do not model the data as a timeseries. Model performance could be improved by treating the data as such. To do this, we would need to create models for various subsets of time. We could also build models for specific geographical locations and compare data. These are more specific models which reflect the natural variation in the data. These applications would likely improve model variance and predictive accuracy. They could also provide further insight on residential housing prices in various neighborhoods and timeframes.

RECOMMENDATIONS:

Collecting data about the type of sale could help identify new subsets of data and improve the predictive performance of the model. For instance,

- Standard Sale
- Bank Owned Sales (REOs)
  - pre-foreclosure
  - foreclosure
  - post-foreclosure
- Short Sale

Collecting data about whether a real estate agent was involved with the sale (Real Estate Agent, No Real Estate Agent) could also help identify private sales, which could indicate more variance in price.

When applying this model to new data, one can approximate price by substituting the observed sale date with the latest sale date observed in the model. However, for more accurate predictions, we would need to build a timeseries model. The current model is best suited to approximate sale prices between 1992-2018.