*Julia Sheriff*

*Springboard | Data Science Career Track*

*capstone Project 1*

PREDICTIVE MODELING OF D.C. RESIDENTIAL HOUSING PRICES

**1. INTRODUCTION**

The purpose of this report is to use predictive modeling to determine residential property values in Washington D.C.. After acquiring the data and doing exploratory data analysis, I proceeded to use linear regression and random forest regression models to find the best predictive solution. The best model was a random forest regressor, which can predict 80% of housing prices between 80% and 125% of their value. A timeseries model would be beneficial for a 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 outliers which indicated erroneous data.

- ‘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. Prices increased most rapidly between 1996-2006 and 2011-2016. Different geographical wards have had different property values and growth. 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. Housing is somewhat more affordable for local residents in the northeast, wards two and three. Properties are consistently unaffordable in central DC.

STATISTICAL ANALYSIS:

INTRODUCTION:

This analysis focuses on finding variables that have a significant impact on real estate prices. The price data is not normally distributed, so I used graphs and applied nonparametric significance tests to find whether or not other variables could be beneficial in predicting price. Many variables showed an impact on price, and many variables were related to 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; the numbers don’t have a relational value. 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 the variety in median prices per area.

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 zipcode, 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.

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. When I applied Spearman’s correlation coefficient for a monotonic relationship across all observations, there was a moderate positive monotonic relationship.

OTHER VARIABLES:

Other significant variables with multiple categories included rooms, stories, and grade. These variables had stronger graphical evidence for a difference in distributions than quadrant. For binary categorical variables, I applied 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 regressions to analyze residential housing prices. I removed housing prices below $30,000 from the regressions because those properties were likely sold under the market price.

The linear regression baseline model used only ‘SUBNBHD’ as a geographical variable. This was done to reduce the number of dummy variables in the set. The training R^2 was .7040, the testing R^2 was .699, the root mean squared error (RMSE) was 227499.7660, and the mean absolute percentage error (MAPE) was 43.17%. The model was not overfitting significantly. Residual plots showed that this regression failed tests for linearity, normality, and homoskedasticity.

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 Training RMSE was 56190.45, the Testing RMSE was 144251.99, and the MAPE was 19.35%. The random forest model was overfitting significantly.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Type | Variables removed | MAPE | Training  RMSE | Testing  RMSE | R^2 |
| Linear Regression, Baseline | 'QUADRANT', 'WARD', 'NBHD', 'CENSUS\_BLOCK', 'ZIPCODE', 'SQUARE' | 43.17 | 227499.77 | 227499.77 | 0.6999 |
| Random Forest, Baseline | 'CENSUS\_BLOCK', 'SQUARE' | 19.35 | 56190.45 | 144251.99 | N/A |

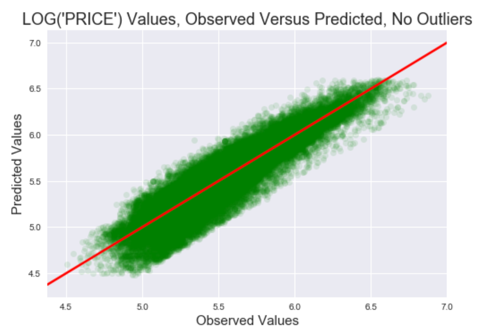
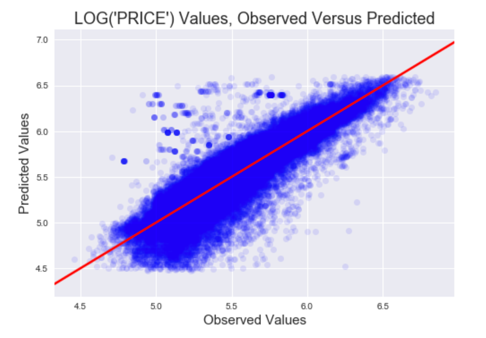
**2.4 EXTENDED MODELING**

For the linear regression model, a log base 10 transformation with ridge regression improved the baseline model. Testing R^2 improved from .70 to .77. The best model including outliers had high dimensional geographical variables removed and one collinear variable removed, ‘KITCHENS’. 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 heteroskedastic. The model without outliers (1.7% of the data) performed the best with the same variables removed, with testing R^2 at .8406. 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 | Testing  RMSE | Training R^2 | Testing  R^2 |
| Baseline | 'QUADRANT', 'WARD', 'NBHD', 'CENSUS\_BLOCK', 'ZIPCODE', 'SQUARE' | 43.17 | 227499.77 | 0.7040 | 0.6999 |
| Improved baseline  -ridge regression  -log(‘PRICE’) | 'QUADRANT', 'WARD', 'NBHD', 'CENSUS\_BLOCK', 'ZIPCODE', 'SQUARE' | 2.08,  \*29.21 in dollars | 0.16 | 0.7633 | 0.7587 |
| **Best model with outliers**  -ridge regression  -log(‘PRICE’) | 'KITCHENS', 'SUBNBHD', 'CENSUS\_BLOCK', 'SQUARE' | 2.01  \*28.00 in dollars | 0.16 | 0.7793 | 0.7745 |
| **Best model without outliers**  -log10(‘PRICE’)  -Ridge Regression | 'KITCHENS', 'SUBNBHD', 'CENSUS\_BLOCK', 'SQUARE' | 1.79  \*23.89 in dollars | 0.13 | 0.8409 | 0.8406 |

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



**Outlier Descriptions (1.7% of data):**

* higher prices
* sold during different times
* lower 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

RANDOM FOREST MODELS:

Predicting log base 10 of prices also yielded better results for the random forest model, as the log base 10 yielded better results for the linear model. Using parameter tuning further improved the 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.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Type | FEATURES REMOVED | | MAPE | Tr RMSE | Tst RMSE | Parameters |
| Baseline | 'CENSUS\_BLOCK', 'SQUARE' | | 19.35 | 56190.45 | 144251.99 | n\_estimators = 10 |
| Baseline  log | 'CENSUS\_BLOCK', 'SQUARE' | | 1.26,  \*17.64  in PRICE | 0.04 | 0.11 | n\_estimators = 10 |
|  | | -‘EXTWALL’ | 1.2553 | - | - | - |
|  | | -‘INTWALL | 1.2554 | - | - | - |
|  | | -‘SALE\_NUM’ | 1.2550 | - | - | - |
| **Optimized Baseline**  **log** | 'CENSUS\_BLOCK', 'SQUARE' | | 1.23 | 0.05 | 0.11 | |  | | --- | | bootstrap=True,  criterion='mse',  max\_depth=80, | | max\_features='auto',  max\_leaf\_nodes=None, | | min\_impurity\_decrease=0.0,  min\_impurity\_split=None, | | min\_samples\_leaf=2,  min\_samples\_split=2, | | min\_weight\_fraction\_leaf=0.0,  n\_estimators=110,  n\_jobs=None, | | oob\_score=False,  random\_state=None,  verbose=0,  warm\_start=False) | |
| AdaBoost  log | ''CENSUS\_BLOCK', 'SQUARE' | | 3.02 | 0.22 | 0.22 | base\_estimator=None,  learning\_rate=0.05,  loss='exponential',  n\_estimators=90,  random\_state=None |
| GradientBoost  log | ''CENSUS\_BLOCK', 'SQUARE' | | 1.51 | 0.11 | 0.13 | alpha=0.9,  criterion='friedman\_mse',  init=None,  learning\_rate=1, loss='ls',  max\_depth=3,  max\_features=None,  max\_leaf\_nodes=None,  min\_impurity\_decrease=0.0,  min\_impurity\_split=None,  min\_samples\_leaf=1,  min\_samples\_split=2,  min\_weight\_fraction\_leaf=0.0,  n\_estimators=150,  n\_iter\_no\_change=None,  presort='auto',  random\_state=None,  subsample=1.0, tol=0.0001,  validation\_fraction=0.1,  verbose=0,  warm\_start=False |

**2. 5 ANALYSIS OF RESULT**

The Random Forest Optimized Baseline Model performed the best at predicting the log(Price) with a 1.22 MAPE and 0.11 RMSE, slightly better than the Random Forest Baseline Log Model. The Random Forest Gradient Boosting Model performed third best with a 1.51 MAPE and 0.13 RMSE. While the Linear Log Ridge Regression had the best performance of all linear models, with 1.79 MAPE and 0.13 RMSE , it performed worse than the best random forest models,.

**3. CONCLUSION AND FUTURE WORK**

SUMMARY:



We our best model, we can predict price with a mean absolute percentage error of 17.27%. 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 43.17 MAPE, the model’s general accuracy is limited in precision. The table on the left shows percentage errors in price prediction, showing predictions too low with negative percentages, and too high with positive percentages.



The model predicts outlier prices higher than observed for low property values, and outlier predicts prices lower than observed for high property values. There are several outliers shown in this graph, and the graph includes all 96,772 observations.



80% of the time our error in price prediction is above or below 1.52% to 18.63% 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. The percentage error is more sensitive to errors in dollars for low properties because the dollar is compared to a lower observed value.

QUESTIONS:

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 report do not treat 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.

Type of sale:

-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.