The Florida Housing Project

By Kyle Morris

DSC 630 Predictive Analytics

Spring 2020

Milestone 5

# Executive Summary

The Florida Housing Project is a total of two predictive analytics models that take data provided by the US Census Bureau in their American Community Survey, 2018, and, using the power of predictive analytics, predicts the resulting value of either the Monthly Rent (rent model) or the Property Value (property value model.

Using some 100,000+ entries available to us, our models were trained by example to predict the results. The output of that model was then compared to the actual values for those homes, to see how the models stacked up.

Instead of utilizing every single input for every single observation, we were able to refine our model to use a smaller subset of inputs to make our predictions. This keeps costs and computational times low while sacrificing very little overall accuracy. This results in our models being easier to deploy and quicker to adapt to changes.

Our final predictions were compared to the actual values. For each home, the inputs were put into the models and the resultant dollar amount was compared to the value recorded for that home. In the vast majority of cases, the model’s prediction was within $500, for the Rent model, and a single order of magnitude for the Property Value model. This shows that our model has some ability to predict the resulting output.

While the data is Florida focused, this technique can be adapted to using data from other states. The specific details of how the model was developed are in the technical report to follow.

# Abstract

Given the attributes of a home, that is to say, the number of bedrooms, amenities offered, and location, we should be able to estimate the cost of the home. In this case, we would take the monthly rent to estimate the total cost of a 30-year fixed-rate mortgage. The data came from the Census bureau and classified housing units as either undervalued, overvalued, or a good fit.

# Intro/background of the problem

The Florida Housing Project will be using information collected on housing in Florida to predict housing costs. Given these inputs, we should be able to tell whether a house is a good deal or not based on the predicted price. I will be the only one working on this project.

There are many different styles of houses available, and while I would like to focus on any single-family domiciles (while ignoring dorms, student housing, etc.) it may eventually fall to us having to focus on a single style of home. I’m hoping that we can keep all types in, however.

# Background

Ultimately, the final sale price of a home is based on what the market will bear. That being said, several tangible, measurable indicators can affect the price of a home. While some intangibles do exist, generally similar homes within the same region will sell for a similar price range. Set a price too high, and the house will sit on the market for months.

We do have some value indicators available to us in our data set. Mortgage payments and appraised value for property taxes stand out.

# Problem Statement

Through analysis of publicly available housing records in the state of Florida, the goal is to narrow down fields that have the most effect on market price. Once the 230+ fields of data have been pared down to a more manageable set, the hope is to build a linear model that can predict the price of a home and, given that value, indicate whether a home is priced as expected, priced too low, or priced too high.

# Scope

While data is available for all 50 states, the District of Columbia, and Puerto Rico, the variance between the different regions due to the cost of living, availability of housing, etc. can result in wildly different values. To minimize this regional variation, we will be focusing only on data from the state of Florida. Data is available over both five years and a single year span, and we will be focusing on a single year, the most recent data available.

Although the information that is provided can be used to compare costs between multiple years, we will only be focusing on a single year’s worth of data.

# Methods

Our data will be read in the CSV format using Python. From there, we will transform and clean up the data. This outputted data will then be read into R which is what the rest of the analysis will be conducted in. We will split the data into a training set and a test set, using our model to predict the value for the test set.

As we currently have nearly 100,000 observations in our data set, we shouldn’t need fear not having enough data for analysis. However, certain programs are better suited to larger databases than others, so if we begin to run into problems, we may need to adjust our approach (or cut down our data).

Data Sources

The primary requirement for data for this project will be the housing data. We will be using data provided by the US Census. The ACS Public Use Microdata Sample files are a sample of the actual responses to the American Community Survey and include most population and housing characteristics. This information is freely provided in CSV format by the Census Bureau, and we will be using the 2018 housing data for the state of Florida.

The data was imported into Python and initial cleanup was performed. The dataset was not complete and had several columns with NA values. The following columns: ['BDSP', 'HUGCL', 'NPP', 'NR', 'NRC', 'PARTNER', 'PSF', 'R18', 'R60', 'R65', 'SRNT', 'SSMC', 'SVAL'] were identified as having both NAN values and values in which 0 was a valid response. In this case, a value of -1 was selected to provide a numerical result for the column while still differentiating the NA case from the 0 cases.

In the next set of columns, ['FINCP', 'OCPIP', 'SMOCP'], it was determined that the mean was more appropriate as a replacement value. Finally, the rest of the columns did not use 0 as a value and so those were put in.

Once these transformations were complete, the data was outputted back into a CSV file and sent into R. We began to look at our target variables, GRNTP (Gross Rent Payments), RNTP (Rent Payments), and VALP (Property Value). Preliminary analysis of the data showed that both rent payments were mostly normally distributed, but with huge spikes at 0 (~76% of our data points), while Property Value was positively skewed. With options for rent payments to either replace with average or remove those values, we went with removing the values. We still have 30,000 points of data without them, and we can then produce a model that predicts rent payments for renters, and house value for purchasers.

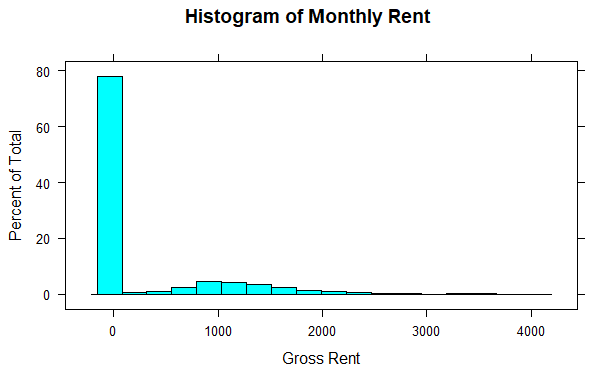


Fig. 1: GRNTP Histogram

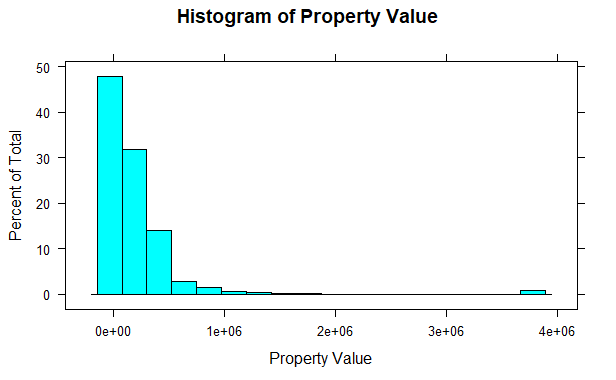


Fig. 2: VALP Histogram

A log transformation should be sufficient for the Property Value to normalize that dataset. From there, we can build the model. Our data was partitioned into 2 sets, a training set, and a testing set. For the monthly rent, this was done on our subset (rent > 0) data. For property value, the original dataset was used.

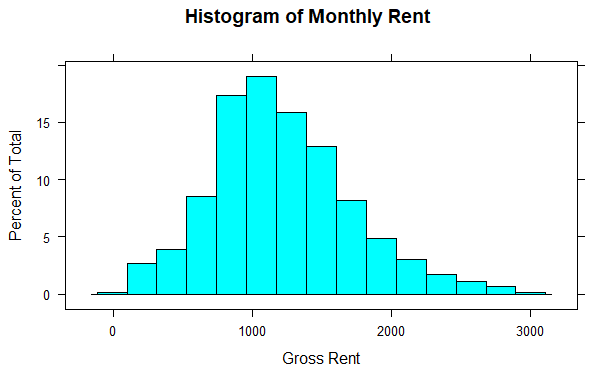


Fig. 3: GRNTP with values > 0

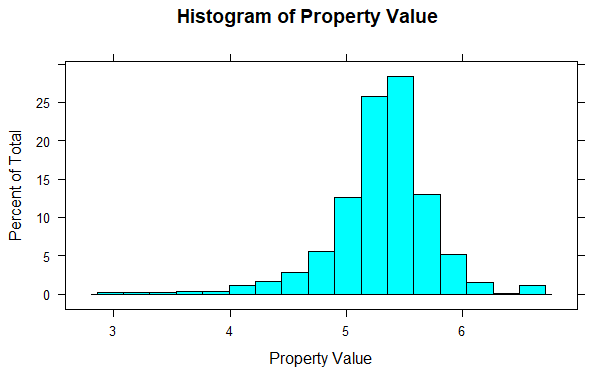


Fig. 4: VALP after log transformation.

After running the first model, it was discovered that the RNTP variable was the most significant factor. As that column does tie to GRNTP, it doesn’t tell us anything. The model was run again without this particular section.

The data currently has 231 columns worth of data – for our model to be useful from a computational standpoint, we’ll want to narrow this down to a more manageable number. The function Varimp can assist with this. We fed all 230 variables and our target variable into the system for both the rent model and the property value and analyzed the columns that were most important to the final calculation.

The top 7 or so columns were included. Most of these had a value of importance that was at least 25% or higher. Some variables were included below that threshold – notably, Monthly Gas Price. Since Electric and Water costs were included, it made sense to include the final utility.

After adjustment, 7 variables were identified that had both importance to determining gross rent: BDSP (number of bedrooms), ELEP (electricity cost per year), YBL (Year built), WATP (Water cost per year), BLD (building type), ACR (acreage of the lot) and GASP (Gas price per year). Many of the other factors were reports on the family that lives there instead of the home itself – while knowing multimillionaires live at a home, it doesn’t necessarily translate to a better home. These 7 are more focused on the home itself.

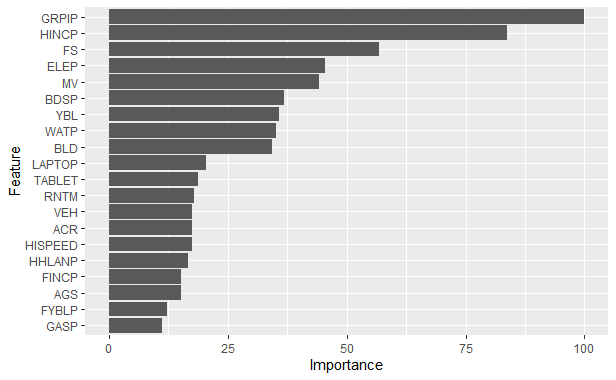


Fig. 5: Feature Importance before selection.

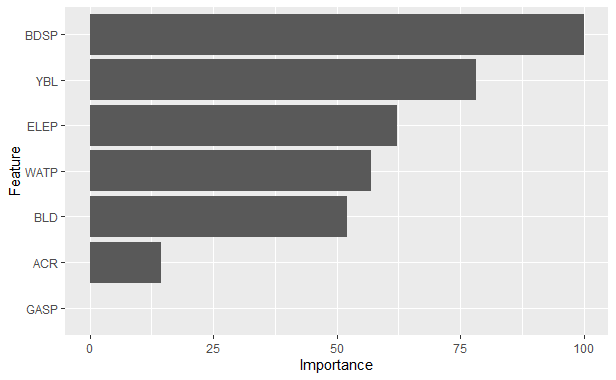


Fig. 6: Feature importance after selection.

For the Property Value, some other factors were considered. SVAL (Specified Owner Unit), BLD (building type), MHP (Mobile Home Costs), BDSP (Number of Bedrooms, YBL (Year built), CONP (Condo Fee), ELEP (Electricity Cost per Year), WATP (Water cost per year), and GASP(gas cost per year) were considered. While some of these are the same, they were not as important as in the rent and not necessarily in the same order.

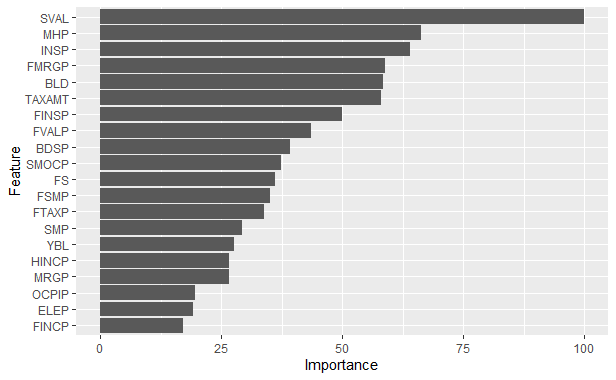


Fig. 7: Feature Importance before selection.

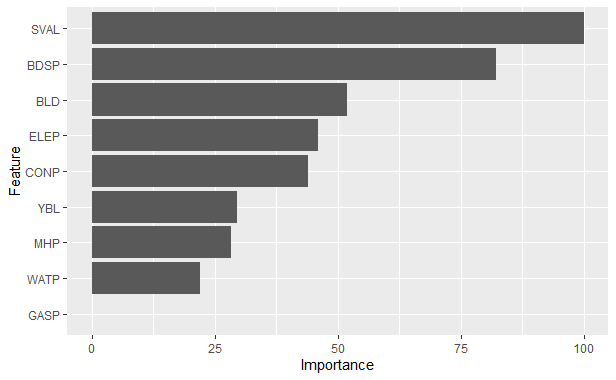


Fig. 8: Feature Importance after selection.

Essentially, at this point, our model had the outline. We now had to determine the specific method used to fit our data to our model. In this case, we chose Linear Regression for several reasons – a: the data seemed to pass the smell test for being linearly related, and b: linear regression is easy to interpret and predict how different inputs affect the model.

There are four assumptions associated with a linear regression model:

1. Linearity. The relationship between X and Y is linear.
   1. Covered in part a.
2. Homoscedasticity: The variance of residual is the same for any value of X.
3. Independence: Observations are independent of each other
   1. Covered by the method used to collect data – adding another bedroom to my home doesn’t directly influence the cost of your home, only indirectly via property values going up.
4. Normality – for any fixed value of X, Y is normally distributed.
   1. If you’ll recall, we applied a few transformations to our data in the cleaning step to make sure our data was normally distributed.

Thus, as we have established we met the conditions for linear regression, the model was trained on our data set. We used 10-fold cross-validation, which means we split the data into ten subsets, train the model on all but one of the subsets, and then evaluate the model on the subset not used for training. This allows us to repeatedly train the model, resulting in a more robust model overall.

# Results

With the final models completed utilizing the features deemed most important, the test set was then fed into both models and a predictive output was established. This new variable contained a predicted value for every single input in the testing set. It was now

The resulting model values were plotted against the test set values. In this case, we plotted the actual rent payment minus the predicted rent payment, and the actual property value minus the predicted value. The difference between the listed value and the predicted values is as shown:

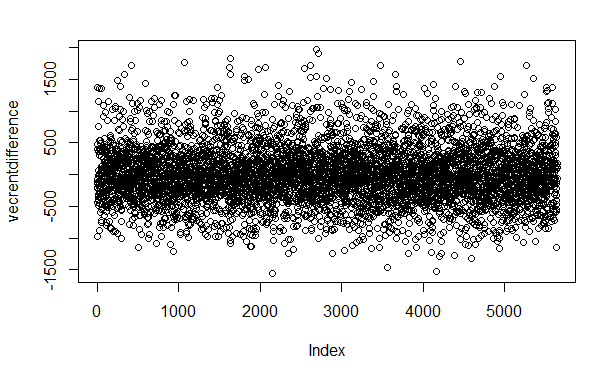


Fig. 9: GRNTP – Predicted Value.

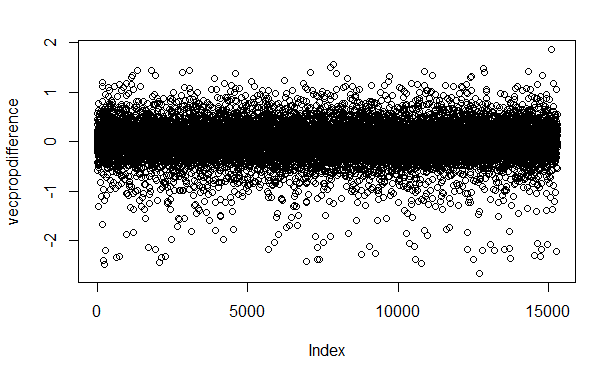


Fig. 10: VALP – Predicted Value

# Discussion/conclusion

As you can see, the vast majority of our predicted values fall within $500 of the actual values. With property values, those values are fairly close as well.

As stated in our introduction, while these values are by no means gospel, it can give you a better idea of which properties are overvalued and which are undervalued.

The results of this project could easily be adapted for use in other states. While this one was Florida focused, we could run the analysis again on another state’s properties. I suspect that the Florida Model would not perform as well on other states and it would be important to tweak things for each population of data given. Hopefully, the insight gleaned in running this project would translate well to my house search later this year.

# Acknowledgments

The Census Bureau, for making their data easily accessible. It cannot be overstated how important to this project that having a good codex in place made things so much easier to interpret the data.

# References

US Census Bureau