Technical Report – Housing Data Science Project

15 April 2023

Prepared By: Clark P. Necciai Jr.

INTRODUCTION

Our end goal in analyzing data relating to historical house prices was to provide a reliable predictive model with meaningful predictions in-line with PA Reality's business strategy to identify over/ underpriced homes. This report details our findings by the order in which we accomplished this. We first began with Exploratory Data Analysis through which we better understood the relationships and complexities within the data and tailored it to our needs. Next, we fit and evaluated models according to various performance metrics to determine which model would be our final, chosen model. Lastly, we summarize our results and subsequently provide key takeaways and future recommendations.

EXPLORATORY DATA ANALYSIS

Our primary focus in the exploratory data analysis phase is determining relationships, patterns, and any concerns within the given dataset. Beginning with our inspection of the dataset, we first imported the dataset then utilizing *glimpse()* to discover the varying **types** and **names** of variables, **dimensionality** of the dataset, and **format** of values. Using *is.na()*, we find that luckily there are no missing values in this dataset. After initial conversion of our character types to factors, we start with an investigation of our response variable, *price*. A density plot and *shapiro.test()(*used in determining normality), reveal that *price* is exceedingly right-skewed; most of our observations contain lower priced residencies. A non-normally distributed response could have significant impacts on our predictions. Therefore, we elected to perform a *log()(logarithmic)* transformation, afterwards converting our final predictions using *exp()(exponential function)*. Post transformation, two observations were determined to still be outliers in the response variable and so were subsequently removed after ensuring any rare values were not lost post removal.

Note: Functions will be denoted using a function name followed by '()'.

Inspecting the variables of *desc(house description)*, *exteriorfinish*, *rooftype,* and *location* revealed a few notable aspects of each. Most of our observations(85%) consisted of single-family homes. A mere two observations consisted of mobile homes. While this may give us better predictions for single-family residences, it may negatively impact the predictability of home prices for other types due to this blatant bias for a single home type. For our residencies' exteriors, we find most of our homes consisting of brick and frame types. With roof types, nearly 84% of our observations consisted of a single type, shingle. For the location of our homes, while the distribution of values was not *as* skewed, approximately 65% of our values came from outside the city(NotCity). We see this type of one-sidedness also occur in the *basement* variable, with approximately 94% of observations having a basement. We should be aware of this skewness as we build models, as variables with one-sidedness when it comes to values may not capture the full range of variation in the underlying dataset. Variables for the total number of rooms(*totalrooms*) and number of fireplaces(*fireplace*), while still not having an entirely normal distribution, contained a greater mix of values. Plotting price as a function of the total number of rooms separated by location revealed to us the reasoning as to why our initial upward trend of prices per number of rooms was not monotonic. When accounting for location, most observations for the number of rooms (3, 14, and 16 rooms) consisted of values falling far outside the intuitive range. Other variables were also dissected to reveal possible influences on price, but none had so much an obvious trend than with total number of rooms.

Afterwards, an oddity was discovered when inspecting the number of stories for residences. It was found that two rather unique, singular and strange values of 1.7 and 2.8 were found. Although at first glance it may appear as mis-input values, these observations did not warrant removal. Though without context, these may be after all, correct values and so were kept. A much more concerning oddity was found with the discovery of **twenty-nine** separate observations having a value of 0 for *lotarea*. Intuitively, zero should not be a possible value for the lot area, as this would imply a residence to not even exist. While we could have elected to impute values for these observations, we simply do not know the reason for their being there and should not unnecessarily tamper with these values without context. We also do not want to remove these observations due to this oddity due to the magnitude of observations with this occurrence costing us much information loss. For these reasons, these oddities were also chosen to remain as is and a part of the dataset.

As a last important point, we plotted a correlation and scatterplot matrix between our variables. These plots aided us in quickly visualizing the relationships we would find should we investigate each variables relationship to one another. It first appeared in the scatterplots that some of the most obvious trends apparent in the data occurred between our response, *price*, and the following variables: *totalrooms, bedrooms, bathrooms, fireplaces,* and *sqft.* However, a more statistically inclined approach using the correlation matrix revealed moderately strong relationships only between our response and *bedrooms, totalrooms,* and *fireplaces.*

METHODS OVERVIEW/DETAILS

With our Exploratory Data Analysis complete and a better understanding of our dataset, we proceeded with models with which we believe may have suitable interpretability for our problem and adequate predictive performance. As a primary performance metric, I decided upon using the root mean square error(**RMSE**) as way to compare model performance. RMSE has the advantage of being easily interpretable, as its value is in the same units as our response variable, (log)price. Because of this, it would be easy to explain the performance of whatever model we end up with to PA Reality. As a means of finding a good test error estimate, repeated 10-fold cross validation was used repeatedly 5 times per model. This approach yielded stable test error estimates. In the following section we will discuss each of the considered models' advantages and notable aspects encountered during the fitting process. The most important variables were found using the *varImp()* function for each model. This function is designed specifically to work with *train()* which is derived from the caret package. These variables are ordered from most to least important and only the top five are shown for brevity.

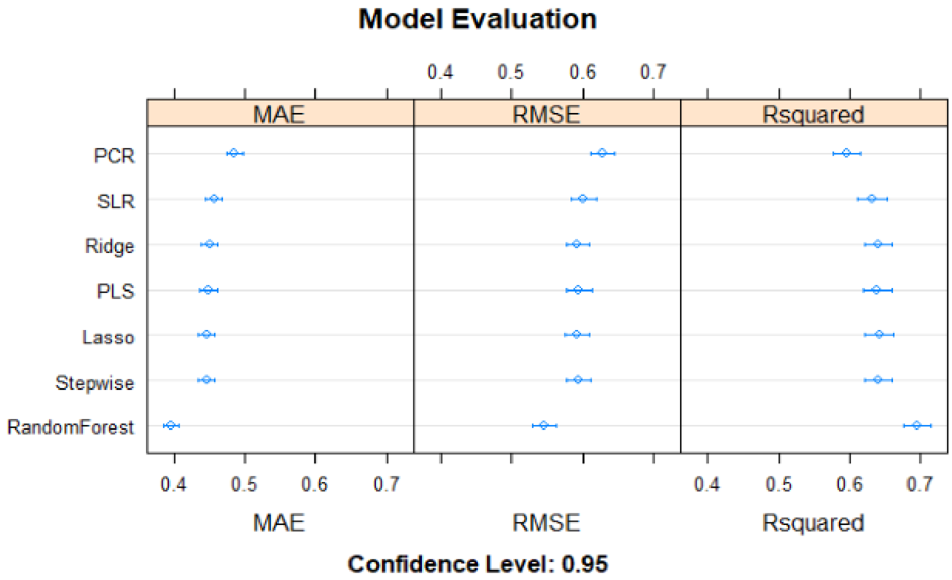
<u>Models Considered:</u>

- **Simple Linear Regression**

    - Upon first fit, we inspected the model for evidence of high variance inflation values(VIF) and found that *rooftype* had values well into the 70s. After the removal of rooftype, *totalrooms* and *location* were also removed due to their high VIF.
    - Top five most important variables: *yearbuilt, sqft, bathrooms, bedrooms, avgincome*
- **Ridge & Lasso Regression**
    - Ridge and Lasso models have the distinct advantage of preventing overfitting by pulling coefficient estimates towards zero. These techniques had higher predictive accuracy over the standard regression fit, implying the previous model may have suffered from overfitting.
    - Top five most important Ridge variables: *sqft, yearbuilt, bathrooms, bedrooms, avgincome*
    - Top five most important Lasso variables: *sqft, yearbuilt, bathrooms, bedrooms, rooftypeSLATE*
- **Stepwise Linear Regression**
    - Stepwise(forward & backward selection) modeling has the additional step of model comparison with AIC which prioritizes smaller models. This may further help reduce overfitting and yielded even better predictive accuracy over the previous three model fits.
    - Top five most important variables: *sqft, bathrooms, totalrooms, bedrooms, lotarea*
- **PCA & PLS**
    - Some of our most important variables have wide variances. PCA and PLS were considered due to their focus on high variance and dimensionality reduction capability.
    - Top five most important PCA variables: *sqft, bathrooms, totalrooms, bedrooms, lotarea*
    - Top five most important PLS variables: *bathrooms, sqft, totalrooms, bedrooms, fireplaces*
- **Random Forest**
    - Due to their generally high predictive capability and relative ease of interpretation, a random forest model was considered. Random forests, being a non-parametric approach, was a suitable fit for the types of high variance variables we have within our dataset.
    - Top five most variables: *sqft, bathrooms, lotarea, yearbuilt, totalrooms*

SUMMARY OF RESULTS

The random forest outperformed all other models in terms of not only our primary performance metric, **RMSE**, but also of **Rsquared** and **Mean Absolute Error(MAE)**. All other models performed within one standard deviation error of each performance metric relative to one another. When considering all models, the most important variables in my opinion were those that appeared in every model fit: ***sqft, bathrooms*****, and** ***bedrooms*****.**



CONCLUSIONS/TAKEAWAYS

This most challenging aspect of this housing dataset included an abundance of one-sidedness when it came to some values distributions for our variables. These highly skewed distributions no doubt hurt our overall predictive capability. These were mitigated by the decision to fit simplistic, not-over-complicated models to the data to prevent overfitting. This resulted in all our models performing well and our *best* model, the random forest model, having an even smaller test error than the others. I would be confident in random forest model's predictive capability reflecting my job performance. As a final recommendation to improve model fit, further collected data needs to have proportional levels of values across variables. Some of values of our variables were borderline *unique* due to their small counts. These types of variables had unreliable distributions and more than likely also hindered model fit. A larger sample size with a focus on mitigating the skewedness of variables would improve future modeling.