STAT 1361: Final Project Technical Report

Following housing bubble that eventually led up to the Great Recession in 2007, paired alongside the tumultuous times we find ourselves in, the demand for affordable housing has never been higher. The typical standard to search for price estimates of houses is Zillow. With this in mind, a rival company, PA-VA realty seeks to do a similar task to Zillow but specialized for houses in Virginia and Pennsylvania. Generally, each column describes a specific aspect of a home in Virginia or Pennsylvania. Upon closer analysis on each variable, there are quite a bit of oddities. For one thing, the variable, fireplaces consist of missing values. As a result, running any sort of statistic such as mean or standard deviation wouldn't be possible. Thus, to overcome this problem, any values that were NA in the fireplace's column were replaced with the mean of the values. Replacing all the values with 0 would not make sense, as it would imply that there are zero fireplaces for a given home, whereas in reality, there could actually be one or two. It could just be that in the data collected, the number of fireplaces for a given home is unknown. Thus, normalizing the missing values with the mean would make more sense. One thing to note is that using the mean to replace the missing values may yield a decimal number. As a step further, I rounded the mean up so that the values would make intuitive sense, as 0.5552 fireplaces doesn't make any sense. Additionally, upon looking at a summary of each variable, as seen below:

```
       id          price                desc        numstories      yearbuilt     exteriorfinish    rooftype       basement
PA10001:   1   Min.   :  35847   CONDOMINIUM  :  69   Min.   :1.000   Min.   :1805   Brick   :858   METAL  :234   Min.   :0.0000
PA10002:   1   1st Qu.: 139575   MOBILE HOME  :   1   1st Qu.:1.000   1st Qu.:1922   Concrete:  4   ROLL   : 56   1st Qu.:0.0000
PA10003:   1   Median : 267905   MULTI-FAMILY :  42   Median :2.000   Median :1940   Frame   :388   SHINGLE:727   Median :1.0000
PA10004:   1   Mean   : 343143   ROWHOUSE     :  16   Mean   :1.691   Mean   :1944   Log     :  3   SLATE  :383   Mean   :0.5607
PA10005:   1   3rd Qu.: 426548   SINGLE FAMILY:1272   3rd Qu.:2.000   3rd Qu.:1960   Stone   : 52                 3rd Qu.:1.0000
PA10006:   1   Max.   :3990701                        Max.   :3.000   Max.   :2017   Stucco  : 95                 Max.   :1.0000
(Other):1394
  totalrooms       bedrooms       bathrooms       fireplaces         sqft           lotarea          state         zipcode        AvgIncome
Min.   : 3.000   Min.   : 1.00   Min.   : 1.000   Min.   :0.0000   Min.   :  475   Min.   :     0   PA:713   Min.   :15003   Min.   :11306
1st Qu.: 6.000   1st Qu.: 3.00   1st Qu.: 1.500   1st Qu.:0.0000   1st Qu.: 1349   1st Qu.:  3794   VA:687   1st Qu.:15216   1st Qu.:29570
Median : 7.000   Median : 3.00   Median : 2.000   Median :0.0000   Median : 1831   Median :  7744            Median :15332   Median :39943
Mean   : 7.515   Mean   : 3.36   Mean   : 2.233   Mean   :0.6101   Mean   : 2201   Mean   : 32528            Mean   :19120   Mean   :43911
3rd Qu.: 9.000   3rd Qu.: 4.00   3rd Qu.: 2.500   3rd Qu.:1.0000   3rd Qu.: 2573   3rd Qu.: 16166            3rd Qu.:23222   3rd Qu.:55239
Max.   :23.000   Max.   :12.00   Max.   :10.000   Max.   :5.0000   Max.   :15872   Max.   :3820212           Max.   :23235   Max.   :95289
                                                  NA's   :687
```

The encoding of zip code does not make sense, as it seems that zip code is encoded quantitatively. However, taking statistics like mean of zip code such as 15213 and 23235 would be 19224, which in the context of zip code has no meaning. Instead of encoding zip code as a quantitative variable, I encoded it as a categorical variable. Additionally, looking at each of the descriptions of the variables, I noticed it is not even necessary to keep zip code, as the variable AvgIncome and State are locational and in particular, AvgIncome focuses in on the average income for a given zip code. Another detail in the data collected for each house is the id, which uniquely identifies a house. In predicting the cost of a house, something identifying a house, like street address is far less important, as variables like state and average income already cover locational factors.

Data aside, a large part was model and variable selection. First, after splitting the data into training and test data, I computed a baseline as the MSE between the test and train set. Then, I considered the following models: simple linear regression, stepwise regression, backward stepwise regression, forward stepwise regression, ridge regression, lasso regression, random forests, bagged trees, trees, pruned trees, gradient boosted trees, PLS, and PCR. The intuition behind this was to take the simplest regression model, and progressively use more complex models, and observe the changes. In particular, with methods like stepwise, and random forests, the goal was to determine the most important variables and to rerun the models again to see if there were any notable improvements. Initially, I had considered kNN and GAMs, however due to problems with how I had encoded the variables, there were problems with running the respective methods, and I decided to drop the results, as they were far too unreliable to trust. Additionally, in running certain methods such as lasso and random forests, I used cross-validation to find the “best” tuning parameters.

Overall, the results were pleasing. I used MSE as a measure of predictive accuracy. With the baseline model, the MSE was 198,916,733,473, and upon running a linear regression, a MSE of 13,391,182,261 was attained. With stepwise regression, the variable totalrooms was deemed to be insignificant. The results of stepwise regression, along with the correlations can be seen below:

```
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
price ~ desc + numstories + yearbuilt + exteriorfinish + rooftype +
    basement + totalrooms + bedrooms + bathrooms + fireplaces +
    sqft + lotarea + state + AvgIncome

Final Model:
price ~ desc + numstories + yearbuilt + exteriorfinish + rooftype +
    basement + bedrooms + bathrooms + fireplaces + sqft + lotarea +
    state + AvgIncome


         Step Df    Deviance Resid. Df   Resid. Dev      AIC
1                                 816 1.721611e+13 19992.51
2 - totalrooms  1 10595652850       817 1.722671e+13 19991.03
```

Correlations

Variable	Zero Order	Partial	Part
descMOBILE HOME	-0.003	-0.076	-0.032
descMULTI-FAMILY	-0.034	-0.115	-0.049
descROWHOUSE	-0.067	0.031	0.013
descSINGLE FAMILY	0.104	-0.030	-0.013
numstories	0.312	-0.081	-0.035
yearbuilt	0.045	-0.084	-0.036
exteriorfinishConcrete	0.002	0.008	0.003
exteriorfinishFrame	-0.190	-0.019	-0.008
exteriorfinishLog	0.017	-0.037	-0.016
exteriorfinishStone	0.112	-0.106	-0.045
exteriorfinishStucco	-0.024	-0.153	-0.066
rooftypeROLL	-0.063	0.026	0.011
rooftypeSHINGLE	-0.347	-0.053	-0.023
rooftypeSLATE	0.368	0.134	0.057
basement	-0.080	0.067	0.028
bedrooms	0.554	-0.104	-0.044
bathrooms	0.793	0.374	0.171
fireplaces	0.374	-0.177	-0.076
sqft	0.814	0.596	0.314
lotarea	0.155	0.114	0.049
stateVA	0.222	0.329	0.148
AvgIncome	0.154	0.119	0.051

Although the MSE from stepwise was not too much better than simple linear regression, it does aid in providing an inkling as to what variables should be eliminated. Variable selection was determined based on correlations to the response, as well as through significance deemed by random forests and stepwise regression. Upon looking at the zero-order correlations, the desc, yearbuilt, basement, and exteriorfinish may be adding noise to the regression, and can be potentially omitted. The MSEs can be seen in the table below:

```
      [,1]               [,2]
 [1,] "baseline"         "7339036454.80731"
 [2,] "198916733472.549" "bagged"
 [3,] "stepwise"         "7897522920.27762"
 [4,] "13395361071.1164" "tree"
 [5,] "backward"         "171475158116.507"
 [6,] "13395361071.1164" "pruned_tree"
 [7,] "forward"          "168708295039.562"
 [8,] "174596213074.38"  "boosted_tree"
 [9,] "ridge"            "183479391828.194"
[10,] "12775975203.9865" "pcr"
[11,] "lasso"            "12543737980.4564"
[12,] "11271700413.3676" "pls"
[13,] "random_forests"   "12571242583.32"
```

Here, we can see that random forests performed the best in terms of MSE, with lasso, ridge, backward, stepwise, PLS and PCR relatively close by. With random forests, the variable importance yields the following:

	% Inc MSE <dbl>
sqft	29.211441
rooftype	21.096443
bathrooms	20.053139
state	20.040095
lotarea	15.322565
AvgIncome	12.527564
exteriorfinish	11.774185
yearbuilt	11.466053
fireplaces	9.839329
basement	8.704895
totalrooms	7.206775
desc	4.534944
bedrooms	3.596273
numstories	2.558916

which indicates that numstories may be another variable to eliminate.

Generally, I believe it is safe to say that in considering the important variables to predict house prices in Virginia and Pennsylvania, it is not necessary to consider what's contributes to the outside, but more on what's on the inside. In fact, when the totalrooms variable was eliminated, the quality of the regression improved. With key variables like bathrooms, and bedrooms covering what most prospective home owners look for in a house, it would make sense to not include totalrooms. Additionally, I believe it's safe to assume that getting rid of zip code is a good idea, as there are variables already accounting for location, as well as the average income variable encompassing the impact of zip code to a certain extent. I believe that the MSEs are high due to the sparseness of the data. If there was more data at hand, perhaps maybe thousands of rows of data, the predictive accuracy of the model would drastically improve. Every model would perform much better, as we're able to get a better idea of trends and more variables are not necessary to keep in the model. Models like PCR, PLS and trees would perform much better. Potentially, pursuing more non-linear models such as GAMs, and perhaps considering interaction terms would be a good option to keep in mind for the future.