Machine Learning: Prediction of House Sales in King County Washington, USA

Sherien Hassan

December 16, 2020

1 Introduction

We all know there are various factors that go into the price of a home ranging from sqaure footage to the homes grade (the construction quality of improvements). To be able to grasp the true affects on the cost of house sales will help buyers and sellers understand how the market works and lead them to making better choices overall on home purchases.

In this project, we will develop and evaluate the performance and predictive ability of a model trained and tested on data collected from House Sales in King County Washington, USA. The objective is to find the best fit from various machine learning models with an end goal of determining the best amongst them to predict future house sales using regression modeling. Regression analysis within machine learning allows us to predict a continuous outcome variable based on the values of predictor variables. For this data set, we will explore Multivariate Linear Regression, Ride Regression, Random Forest Regression, Decision Tree Regression, and Support Vector Regression. To determine which model is the best fit, we will look at the metrics Root Mean Squared

Error(RMSE), Adjusted R Squared(R^2) and Mean Absolute Error(MAE).

2 Hypothesis

Housing sale prices data will correlate well with living space. This hypothesis was made after analysis on the data was done and the obvious observation that columns such as bedroom and bathroom quantity would depend on the size of the house. Meaning, the larger the house, the more likely these values would increase. Hence, if we analyze house sale prices in relation to the size of the home, we could come to a conclusion about how future values could fluctuate.

3 Data Description

The data used in this project is from Kaggle, a data analysis community site, and describes House Sales in King County, USA. The data set ranges from 2014 to 2015 with 21613 observations in home sales as well as 19 columns, plus house price. The columns given in this data set are as follows:
id: number associated to a certain home
date: Date house was sold
price: Price of the sold house
bedrooms: Number of bedrooms
bathrooms: Number of bathrooms
`sqft_living`: Square footage of the living space
`sqft_lot`: Square footage of the lot
floors: Total floors in the house
waterfront: Whether the house is on a waterfront(1: yes, 0: no)
view: Special view from the house
condition: Condition of the house
grade: Represents the construction quality of improvements. Grades run from

grade 1 to 13.

`sqft_above`: Square footage of house apart from basement

`sqft_basement`: Square footage of the basement

`sqft_built`: Built year `sqft_renovated`: Year when the house was renovated

zipcode: Zip code of the house

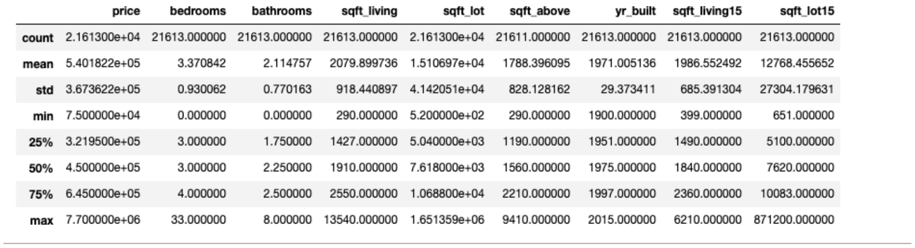
lat: Latitude coordinate of the house

long Longitude coordinate of the house

`sqft_living15`: Living room area in 2015(implies some renovations)

`sqft_lot15`: Lot area in 2015(implies some renovations)

4 Data Analysis



	price	bedrooms	bathrooms	sqft_living	sqft_lot	sqft_above	yr_built	sqft_living15	sqft_lot15
count	2.161300e+04	21613.000000	21613.000000	21613.000000	2.161300e+04	21611.000000	21613.000000	21613.000000	21613.000000
mean	5.401822e+05	3.370842	2.114757	2079.899736	1.510697e+04	1788.396095	1971.005136	1986.552492	12768.455652
std	3.673622e+05	0.930062	0.770163	918.440897	4.142051e+04	828.128162	29.373411	685.391304	27304.179631
min	7.500000e+04	0.000000	0.000000	290.000000	5.200000e+02	290.000000	1900.000000	399.000000	651.000000
25%	3.219500e+05	3.000000	1.750000	1427.000000	5.040000e+03	1190.000000	1951.000000	1490.000000	5100.000000
50%	4.500000e+05	3.000000	2.250000	1910.000000	7.618000e+03	1560.000000	1975.000000	1840.000000	7620.000000
75%	6.450000e+05	4.000000	2.500000	2550.000000	1.068800e+04	2210.000000	1997.000000	2360.000000	10083.000000
max	7.700000e+06	33.000000	8.000000	13540.000000	1.651359e+06	9410.000000	2015.000000	6210.000000	871200.000000

Figure 1: Data Set Overview

Price will be the variable that we test amongst other columns to see which is the true factor for an increase in house sales, so it is important to note that prices do vary starting from $75K to $7.7M. Another notable column would be `sqft_living`, in which varies from 290 square feet to 13459 square feet. Let us take a closer look at what the correlation is between price and other columns. We will first see which columns are not needed and drop them using scatter plots.

As could be seen, there is no correlation between price and date, zip code or id. These three columns will be dropped for the rest of the project.

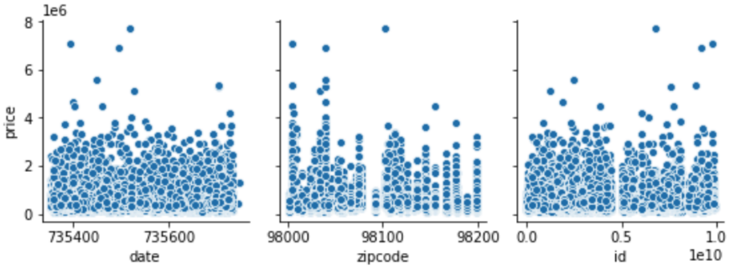
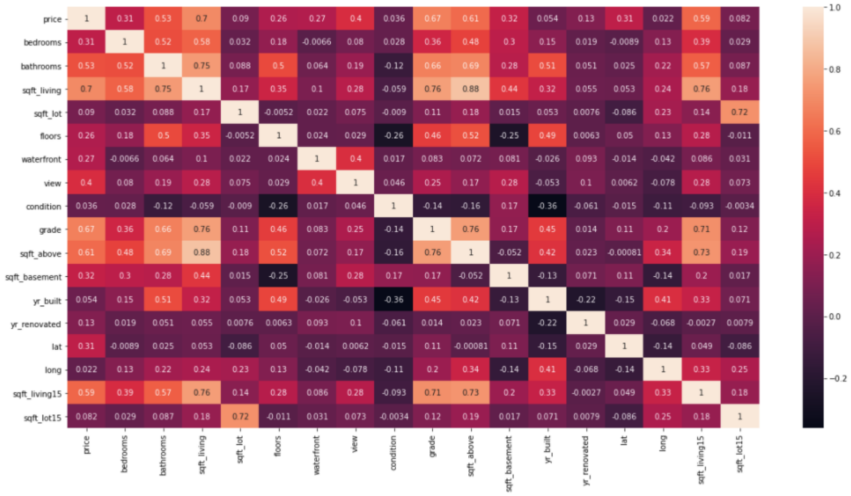


Figure 2: Correlation between Price with date, zip code and id

Now, we will see if there is any linearity relationships of some sort between variables, done through the use of a correlation heat map. The reason we even care to look and remove linearity is because it will increase the variance of coefficient estimates and make our estimates sensitive to minor change in the several models we will fit.



	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqft_above	sqft_basement	yr_built	yr_renovated	lat	long	sqft_living15	sqft_lot15
price	1	0.31	0.53	0.7	0.09	0.26	0.27	0.4	0.036	0.67	0.61	0.32	0.054	0.13	0.31	0.022	0.59	0.082
bedrooms	0.31	1	0.52	0.58	0.032	0.18	-0.0066	0.08	0.028	0.36	0.48	0.3	0.15	0.019	-0.0089	0.13	0.39	0.029
bathrooms	0.53	0.52	1	0.75	0.088	0.5	0.064	0.19	-0.12	0.66	0.69	0.28	0.51	0.051	0.025	0.22	0.57	0.087
sqft_living	0.7	0.58	0.75	1	0.17	0.35	0.1	0.28	-0.059	0.76	0.88	0.44	0.32	0.055	0.053	0.24	0.76	0.18
sqft_lot	0.09	0.032	0.088	0.17	1	-0.0052	0.022	0.075	-0.009	0.11	0.18	0.015	0.053	0.0076	-0.086	0.23	0.14	0.72
floors	0.26	0.18	0.5	0.35	-0.0052	1	0.024	0.029	-0.26	0.46	0.52	-0.25	0.49	0.0063	0.05	0.13	0.28	-0.011
waterfront	0.27	-0.0066	0.064	0.1	0.022	0.024	1	0.4	0.017	0.083	0.072	0.081	-0.026	0.093	-0.014	-0.042	0.086	0.031
view	0.4	0.08	0.19	0.28	0.075	0.029	0.4	1	0.046	0.25	0.17	0.28	-0.053	0.1	0.0062	-0.078	0.28	0.073
condition	0.036	0.028	-0.12	-0.059	-0.009	-0.26	0.017	0.046	1	-0.14	-0.16	0.17	-0.36	-0.061	-0.015	-0.11	-0.093	-0.0034
grade	0.67	0.36	0.66	0.76	0.11	0.46	0.083	0.25	-0.14	1	0.76	0.17	0.45	0.014	0.11	0.2	0.71	0.12
sqft_above	0.61	0.48	0.69	0.88	0.18	0.52	0.072	0.17	-0.16	0.76	1	-0.052	0.42	0.023	-0.00081	0.34	0.73	0.19
sqft_basement	0.32	0.3	0.28	0.44	0.015	-0.25	0.081	0.28	0.17	0.17	-0.052	1	-0.13	0.071	0.11	-0.14	0.2	0.017
yr_built	0.054	0.15	0.51	0.32	0.053	0.49	-0.026	-0.053	-0.36	0.45	0.42	-0.13	1	-0.22	-0.15	0.41	0.33	0.071
yr_renovated	0.13	0.019	0.051	0.055	0.0076	0.0063	0.093	0.1	-0.061	0.014	0.023	0.071	-0.22	1	0.029	-0.068	-0.0027	0.0079
lat	0.31	-0.0089	0.025	0.053	-0.086	0.05	-0.014	0.0062	-0.015	0.11	-0.00081	0.11	-0.15	0.029	1	-0.14	0.049	-0.086
long	0.022	0.13	0.22	0.24	0.23	0.13	-0.042	-0.078	-0.11	0.2	0.34	-0.14	0.41	-0.068	-0.14	1	0.33	0.25
sqft_living15	0.59	0.39	0.57	0.76	0.14	0.28	0.086	0.28	-0.093	0.71	0.73	0.2	0.33	-0.0027	0.049	0.33	1	0.18
sqft_lot15	0.082	0.029	0.087	0.18	0.72	-0.011	0.031	0.073	-0.0034	0.12	0.19	0.017	0.071	0.0079	-0.086	0.25	0.18	1

Figure 3: Correlation Heat Map between price and other columns

Here, we can see that `sqft_living`, grade, `sqft_above`, `sqft_living15`,`sqft_loft15`, and `yr_built` are all correlated so they will be combined and dropped, and when we start our regression modeling, we will see it come back again as our features.

One last thing to note is, there were various attempts in which to scale my data using Min-Max Scaling, Standardized scaling, Robust scaling as well as normalization(not depicted). We could see that the only upgrade in scaling was after robust scaling, although, not by much. Hence, these methods will not be used.

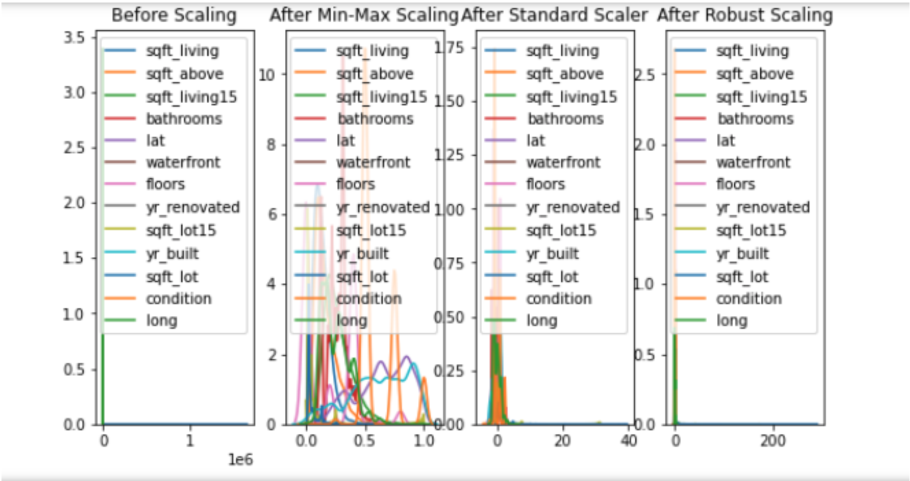


Figure 4: Different Scaler Effects

5 Methods

5.1 Model 1: Multivariate Linear Regression

Multivariate Linear Regression is an extension of simple linear regression and helps us predict the value of a variables based on various other different variables. First, lets look at the RMSE value of 0.334345 in which was calculated. One must note that RMSE has the same value as our dependent variable, so in order to decide whether this is a good RMSE we need to see the parameter set for that. As could be observed, this is a very low RMSE which is good for that would indicate a better fit. Next, looking at our R^2 value, we see a value of

0.593061 which is our number one way to pick between two models. Hence, this indicates an accuracy of 59 percent which is very low. The way in which this could be increased is if predictors are added to the model which might or might not fully help, depending on our predictors added. Lastly, looking at our MAE value of 0.259024 which in comparison to RMSE, is less(as it normally should be). Now, to analyze the scatter plot, we can see that there are some outliers which could be avoided in the future with an analysis of a box plot on the original date, but nonetheless, this has left us with a positive linear model.

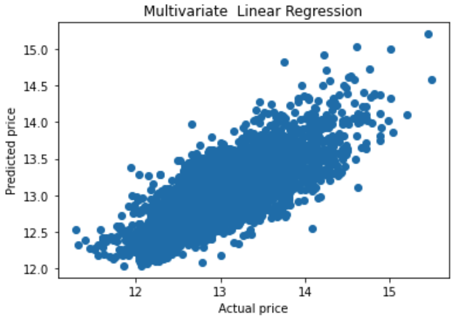


Figure 5: Multivariate Linear Regression Scatter Plot

5.2 Model 2: Ridge Regression

Next, we shall observe our Ridge Regression model. Here, we had the values of 0.335820, 0.595266, 0.260357 which were RMSE, R^2 and MAE, respectively. We know that Ridge Regression is used to solve the issue of multicollinearity, meaning, when independent variables are highly correlated. But, earlier in our data analysis section, we found that `sqft_living`, grade, `sqft_above`, `sqft_living15`,`sqft_loft15`, and `yr_built` were all highly correlated so we took care of them so that they do not effect the results of our fitted model. Although generally, going into this project, I started with Linear Regression knowing that it is usually a base model to see how our model fits before trying

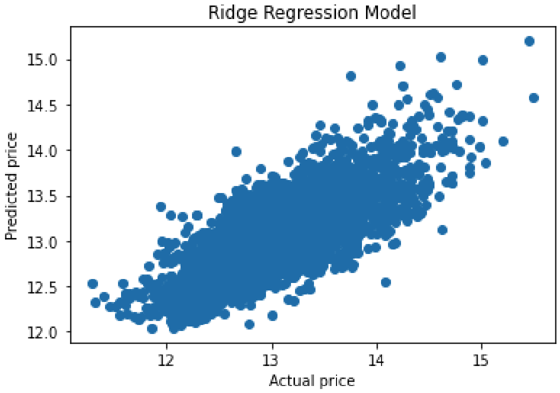
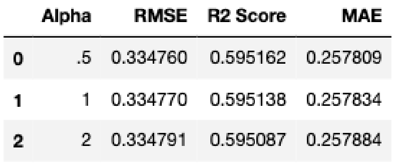


Figure 6: Ridge Regression Regression Scatter Plot

the more complicated methods. This was also the case for Ridge Regression, despite its ability to take care of multicollinear columns, it is merely a model to understand how our data both trained and tested act in a given scenario. One odd thing to note is, I did not use normalizing feature scaling (normalize=True) and this is because when I did, the R^2 values did decrease by 10 percent. The reason I say this is odd is because scaling is usually necessary for this model since the regularization term in its cost function uses feature weights. I did, try to change the values of alpha to see if this would improve the model as seen in figure 7 although, there was no huge change worth discussing.



	Alpha	RMSE	R2 Score	MAE
0	.5	0.334760	0.595162	0.257809
1	1	0.334770	0.595138	0.257834
2	2	0.334791	0.595087	0.257884

Figure 7: Variations of Alpha on Ridge Regression Model

5.3 Model 3: Random Forest

Random Forest Regression is a model in which deals with regression and classification tasks which allows us to train each decision tree on a different data sample where sampling is done with replacement, called bagging or bootstrap aggregation. This was indeed our winning model that will be talked about in comparison to the other models in our Results section. Not only did this model have low values for RMSE: 0.21147 and MSE:0.15316, its accuracy rate was 84 percent. I did not include the number of trees in the forest in the codes parameters, letting Python pick its default 100 trees. I also had criterion MSE which again, if was not specified as a parameter would be the default python uses. This measures the quality of the split which is one of the main reasons this model is better than Decision Tree and others which will be discussed later. Hence, no hyperparameters are used from the default ones which means no regurlization for Random Forest was needed

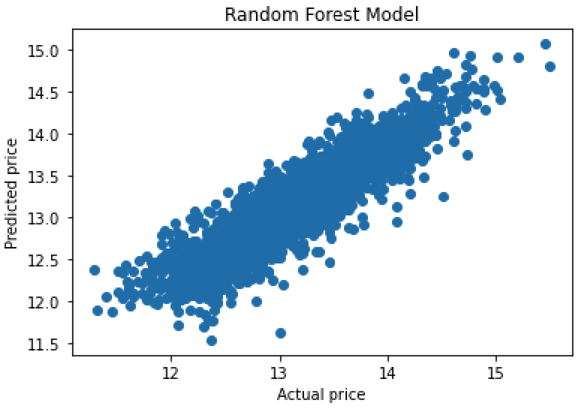


Figure 8: Random Forest Regression Scatter Plot

Looking at scatter plot, we see positive correlation as is a pattern in all models we have seen thus far. Some reasons for why this model could have worked so well is because it adds additional randomness to the model while it grows and it does not search for the most important feature, it searches for the best which gives us a better model overall.

5.4 Model 4: Decision Tree Regression

Decision tree regression breaks down our data set into smaller subsets while simultaneously, associated a decision tree to the break down. It is important to understand that this is different than Random forest, which did much better than this model, because one has a collection of decision trees while the other is one single tree broken down, which may cause over fitting. The only reason one may like this model is due to its transparency but its accuracy rate of 70 is still quite low, despite being the second best model. Although still, with an RMSE: .288 and MAE: .209 value increased from Random Forest, this is not the more suitable model for our data. In terms of our scatter plot, we are still seeing the same pattern of a positive correlation with outliers. Again, in terms of feature scaling or any non-default parameters, decision tree, and most tree-based models, don't need one defined or used to improve the accuracy of the model.

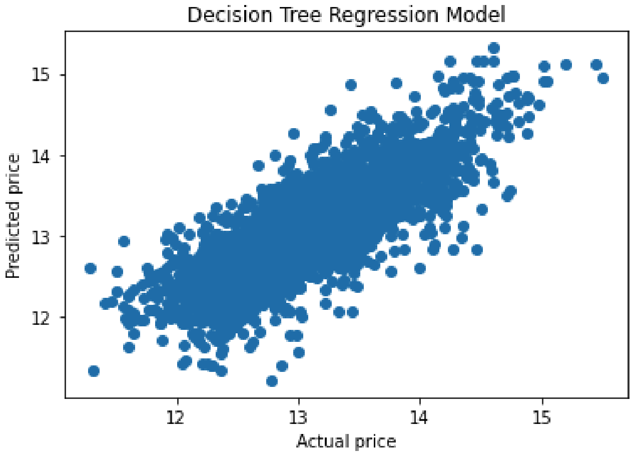


Figure 9: Decision Tree Regression Scatter Plot

5.5 Model 5: Support Vector Regression

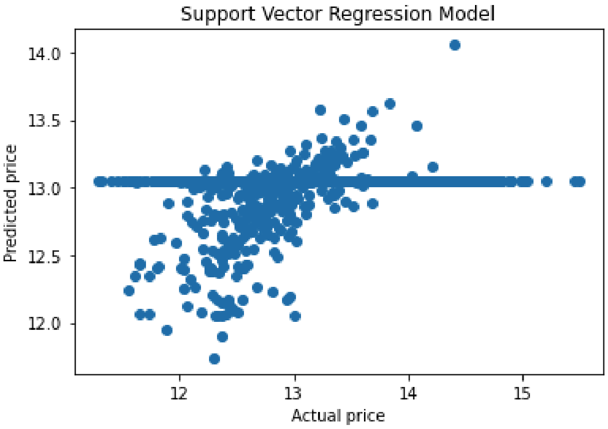
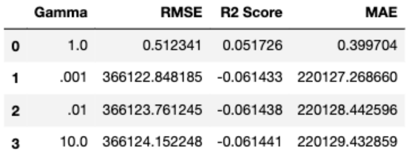


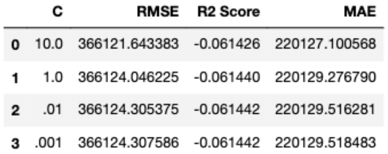
Figure 10: Support Vector Regression Scatter Plot

Support Vector Regression modeling helps one deal with limitations that come from distributional properties of underlying variables, geometry of the data and over fitting. Now, this model was the worst model across all metrics used to test best fit. Its accuracy rate was 5 percent, it had a higher RMSE value of .51 in comparison to the other models and a low MAE value of .20 which doesn't help much with the model. Also, our final models scatter plot is quite odd and so there were various attempts to fix this model. Recall, data has been logged, so I attempted Simple Random Regression on our original data but this only made the model worse by putting out R^2 value into the negatives. This was with a gamma value of .0001. I then attempted to change the gamma value to a different extreme, being, 100.00 an this too did not improve the results. For the sake of conciseness of the report, the paper has left out graphs of these variations but they are left in jupyter notebook incase needed for reference. I came to the conclusion have the data log was not doing much, so I tested different gamma and C values against one another, and there was barley any improvement, shown in Figures 11 and 12.



	Gamma	RMSE	R2 Score	MAE
0	1.0	0.512341	0.051726	0.399704
1	.001	366122.848185	-0.061433	220127.268660
2	.01	366123.761245	-0.061438	220128.442596
3	10.0	366124.152248	-0.061441	220129.432859

Figure 11: Variations in Alpha for Support Vector Regression



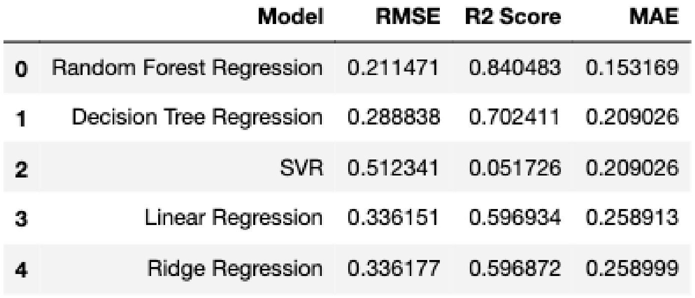
	C	RMSE	R2 Score	MAE
0	10.0	366121.643383	-0.061426	220127.100568
1	1.0	366124.046225	-0.061440	220129.276790
2	.01	366124.305375	-0.061442	220129.516281
3	.001	366124.307586	-0.061442	220129.518483

Figure 12: Variations in Gamma for Support Vector Regression

This has led me to the conclusion that Support Vector Regression is not a model in which should be used on this data set.

6 Results

The results were as follows:



	Model	RMSE	R2 Score	MAE
0	Random Forest Regression	0.211471	0.840483	0.153169
1	Decision Tree Regression	0.288838	0.702411	0.209026
2	SVR	0.512341	0.051726	0.209026
3	Linear Regression	0.336151	0.596934	0.258913
4	Ridge Regression	0.336177	0.596872	0.258999

Figure 13: Metric Comparison Between Models

To conclude, we use machine learning models, namely, Multivariate Linear Regression, Ride Regression, Random Forest Regression, Decision Tree Regression, and Support Vector Regression to determine which model is the best fit with the metrics Root Mean Squared Error(RMSE), R Squared(R^2) and Mean Absolute Error(MAE). As could be seen, our Random Forest Model was the best fit for this data set. Despite values like RMSE for SVR being better for other models, it is clear that the accuracy determined by $R2$ for the model in Random Forest Regression was much greater, 85 percent, than others. Our second best model was Decision Trees which makes sense since it is a variation of Random Forest, but not one that gives us much flexibility and randomness to work on splitting the best tree. Support vector regression was by far the worst model, despite constant variation in gamma and C taken into account and state of the original date, nothing would cause this model to increase its R^2 value. Models like Multivariate linear regression and ridge regression were not as bad compared to Support vector regression but nonetheless, were used as base models to see how this data works in different fittings.

7 Improvements

In the future, I would like to analyze more ways to improve my various R Squared value. This would include applying more transformations to the data before starting the fitting process. It seemed as though transformation after the model was in tact did not help, so it is the raw data that needs to be better analyzed and accounted for with its different variations. Data clean up is a vital process in the model fitting process and I noticed when I applied log, all of my models increased its accuracy rate by 30 percent(of course, except for SRV but that is because this model was not a good fit for the data from the start). Not only is data clean up important but also, having more ways to determine the best fit. Although in most regression models, the metrics I used are best to determine best fit, it is interesting to see if there are any slight changes that could be made if we knew more about a models fittings result. Lastly, after doing research of house sales data, a lot of people suggested the use of XGBoost, which is an implementation of gradient boosted decision trees that tends to give better results. It would be interesting to learn more about this model and it could truly help such a data set. In terms of our hypothesis Housing sale prices data will correlate well with living space. Hence, in the future, a more controlled machine learning project could be done to only analyze these two factors against other columns in the data. Lastly, one must remember that ”All models are wrong, but some are useful” as statistician George Box once said.