Real Estate Prediction

Kasen Teoh, Chung En Pan, Jieru Bai

Summer Session I 2021

1 Introduction and Hypothesis

1.1 Introduction

Every day houses are being bid on and sold to the highest buyer. In many areas, such as Seattle and Silicon Valley, the population continues to grow rapidly with little surface area to accommodate, leading to an increasing house price. With the housing market constantly changing, it is easy for buyers to overpay or sellers to undersell properties. In this project, we aim to analyze house listings and use this data to identify which variables are most significant in predicting housing prices and to create a model to predict housing prices.

1.2 Data Set

Our data is sourced from a Kaggle Data Set (House Sales in King County USA, 2015). There is a total of 21 variables and 21613 observations from May 2014 to May 2015. All variables (sqft_living, bedrooms, bathrooms, etc.) in the data set are numerical variables with the exception of the condition, grade, and zip code. The view, condition, grade, and zip code are all categorical variables. Some factors that may affect the quality of the data and introduce noise are duplicate observations. Within a single year, houses may be bought and sold multiple times. Because we do not have access to the address, we are unable to determine whether observations have been duplicated. We believe that all the variables included are relevant to predicting the house price. However, as mentioned below in our hypothesis, we believe the variables of sqft_living, bedrooms, and view would create the best model in terms of accuracy and interpretability.

1.3 Hypothesis

We hypothesize that sqft_living, bedrooms, and view would be the most significant variables in affecting the house price because from our experience, the larger the house and the better the location, the more expensive the particular house is. We believe that multiple linear regression would out perform the neural network and lasso regression model because of the linear relationship between the predictors and outcome variable and because a neural network would be too complex and attempt to fit the noise in the data while a lasso would possibly not include a significant variable because it was highly correlated with another predictor.

2 Methods

2.1 Multiple Linear Regression $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$

For an initial approach, we had chosen to utilize multiple linear regression to analyze our data. Because we have multiple variables, included in our data set, we wish to include multiple predictors in our model. Additionally, we hypothesized that the larger a house property is with a better view, then the more the house is worth, i.e a linear relationship between square footage and view and price. We first used linear regression to identify the significant predictors and then fed these predictors into linear and lasso regression and the neural network for prediction.

2.2 Lasso $Y = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$

Lasso Regression takes all the predictors present in the data and aims to reduce the dimensions and predictors and perform linear regression on the resulting predictors. We believed that lasso regression would be useful in predicting housing prices because it mathematically reduces the predictors, unlike how we picked and chose predictors based on the correlation.

2.3 Neural Networks

The last model that we chose is a neural network. Neural Networks are a black box model that analyzes models through neurons that aim to identify any underlying pattern (even those unknown to us) throughout the data. Because neural nets seek to find any patterns, we believe this will be efficient in predicting housing prices because there may be patterns that are unknown to the human intuition. We included three hidden layers with 256, 512, and 512 neurons, respectively, and an activation function minimizing the rmse.

2.4 K-fold Cross-Validation

K-fold Cross-Validation is useful in model selection. The cross validation method shuffles then splits the data into k equal folds (or groups). Each iteration uses k-1 folds for training and uses the remaining one fold for testing. For our project, we utilized K-fold cross-validation twice to decide between two different multiple linear regression models and between neural networks and lasso regression. In each iteration, we fitted both models to ensure that they would be trained/tested on the same data and selected the model with the smallest mean RMSE and the greatest r-squared value. We had chosen to use K-fold over LOOCV (Leave One Out CV) because it may give a more accurate test set RMSE and finds a middle ground between bias and variance.

3 Results

From the correlation heatmap below, we see the top 5 features most correlated with price are sqft_living (0.7), grade (0.67), sqft_above (0.61), sqft_living15 (0.59), and bathrooms (0.53). We arbitrarily chose the threshold of multicollinearity to be 0.7, i.e if two predictors were correlated at least 0.7, then we would only choose one of the predictors. Because these five variables (sqft_living, sqft_living15, sqft_above, grade, and bathrooms) are

highly correlated with sqft_living (greater than our chosen threshold), if we include them all in the regression model, we will be violating the assumption of independence among the predictor variables. We chose to only include sqft_living because it is the most correlated with price among the five. Consequently, we chose to compare our hypothesis (a model with sqft_living, bedrooms, and view) with the top 5 variables most correlated with price: sqft_living, sqft_basement, bedrooms, view, and lat.

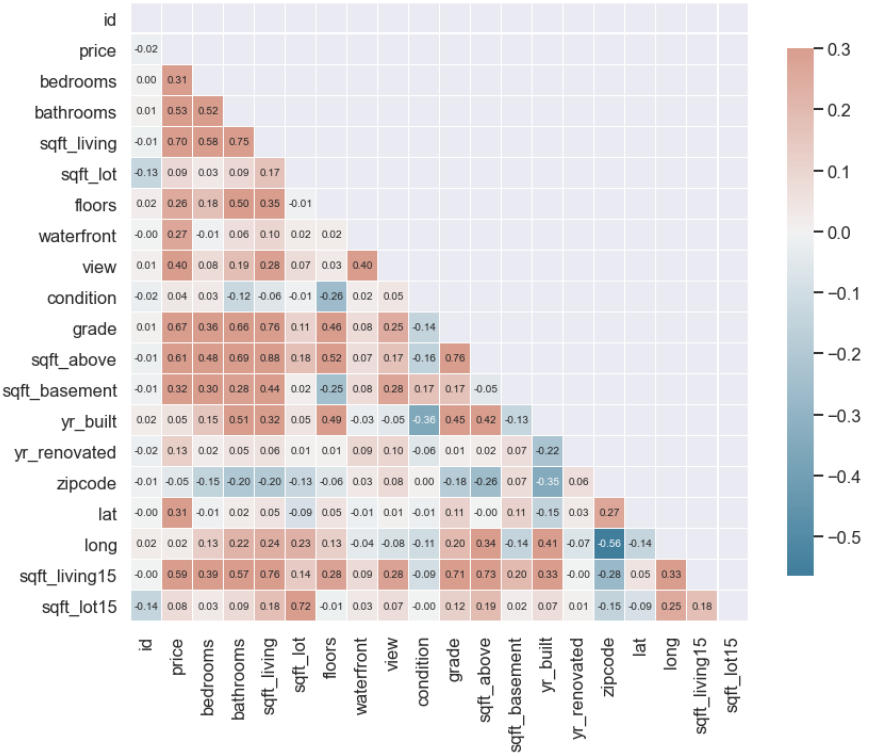

Figure 1: Correlation matrix of all the features included in the data set

Below, we see the distributions of the 5 predictors plus the outcome variable. The distribution of sqft_living seems to be somewhat normal with it being skewed a little bit to the right. Additionally, for the variable of view, we see almost all of the observations in the data have a view of 0. The distribution of price is extremely skewed to the right. Hence we log transform the data so that the distribution is more normal.

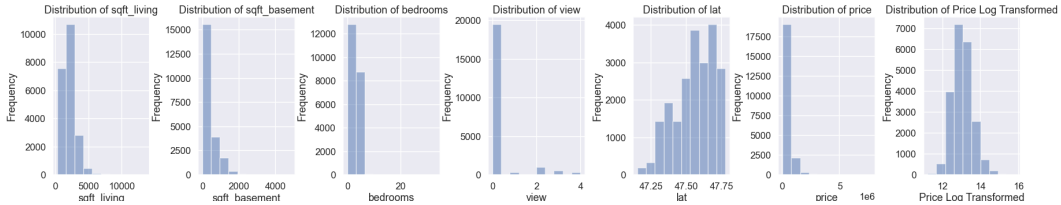

Figure 2: Distributions of the variables that we are going to test

Below, we see the K-fold cross validation results for the standardized data. The first model is the model with sqft_living, view, and bedrooms (our hypothesis) and the second

model consists of sqft_living, sqft_basement, view, bedrooms, and lat (highest correlation with price). We implement the second model because it results in a lower rmse and greater r-squared value because it is more complex and flexible than the first model.

Model	1st Iteration	2nd Iteration	3rd Iteration	4th Iteration	5th Iteration	Mean RMSE
First Model	0.703(0.511)	0.7(0.517)	0.689(0.522)	0.701(0.502)	0.7(0.507)	0.699
Second Model	0.565(0.684)	0.566(0.684)	0.559(0.685)	0.561(0.681)	0.559(0.686)	0.562

Table 1: K-Fold Cross-Validation RMSE and R-squared (R-squared in parentheses)

Below are the results of the multiple linear regression trained on 80% of the data. All the predictors are are significant, accounting for 68.58% of the data's variance. Holding all other variables constant, one unit increase in sqft_living (most significant predictor) results in a 0.67 increase in price_log transform's standard deviation.

`Adjusted R-Squared: 0.686  RMSE: 0.558`

Parameters	Coefficients	Standard Error	T-Statistics	P-Values	95% Confidence Interval
Intercept	-0.056	0.005	-12.424	0.0	[-0.065 -0.047]
View = 1	0.496	0.036	13.889	0.0	[0.426 0.566]
View = 2	0.402	0.021	18.959	0.0	[0.361 0.444]
View = 3	0.592	0.029	20.75	0.0	[0.536 0.648]
View = 4	0.974	0.036	26.742	0.0	[0.903 1.045]
sqft_living	0.67	0.006	116.292	0.0	[0.658 0.681]
sqft_basement	-0.063	0.005	-12.875	0.0	[-0.046 -0.024]
bedrooms	-0.035	0.006	-6.352	0.0	[0.411 0.428]
lat	0.419	0.004	97.073	0.0	[0.411 0.428]

Table 2: Multiple Linear Regression Coefficients

Below, we plot the residuals vs the predicted prices (log transformed) and the distribution of the errors. The distribution of the errors is normal and centered around 0 and the residuals are random, i.e we have not violated any multiple linear regression assumptions.

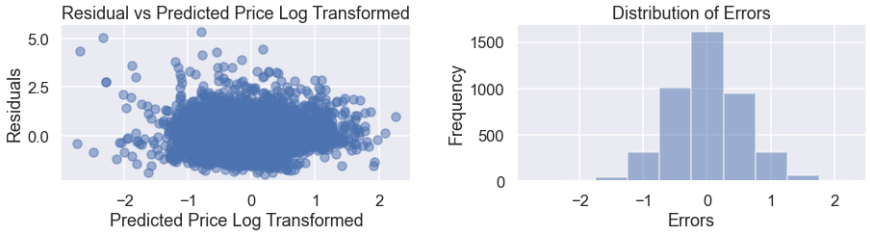

Figure 3: Distribution of the multiple linear regression errors

Model	1st Iteration	2nd Iteration	3rd Iteration	4th Iteration	5th Iteration	Mean RMSE
First Model	0.369	0.367	0.362	0.367	0.361	0.3652
Second Model	0.431	0.455	0.45	0.462	0.445	0.4486
Third Model	0.984	0.993	1.013	0.993	1.016	0.9998

Table 3: K-Fold Cross-Validation RMSE for Neural Net and Lasso Reg

Proceeding with prediction, above is the 5-fold cross validation between neural nets and lasso regression. The first model is a neural net with all the predictors and the second is a neural net with predictors from our linear regression interpretation (sqft_living, sqft_basement,

bedrooms, lat, and view), and the third is a lasso regression. Seen above, a neural net with all the predictors results in the lowest rmse. When trained on 80% of the data, the model is able to account for 86.9% of the data's variance.

```
Adjusted R-Squared: 0.869     RMSE: 0.362
```

The neural network, while more accurate, is a black box model and hence more complex than our original linear regression model. We trained the model in batches of 32 and validation split of 0.2 for 30 epochs to avoid overfitting. Our model did not show signs of overfitting as our test set rmse was around equal to the training.

4 Discussion

In our project, we sought to identify the most important variables related to predicting housing prices in Seattle and to utilize these variables to predict house prices. Through multiple linear regression, we identified sqft_living and latitude as the top two most significant predictors, proving our hypothesis to be half correct. Bedrooms, which we hypothesized to be among the top predictors, was the least significant in our regression model.

As for prediction, we utilized multiple linear regression, lasso regression, and neural networks to predict housing prices. Initially, we believed that multiple linear regression would be the best predictor because of the linear relationship between the predictors and the outcome (house price). However, with multiple linear regression, we calculated an r-squared value of 0.686 and an rmse of 0.558, i.e the model accounted for 68.58% of the data's variance.

We then used a 5-fold cross validation between neural nets with all the predictors and with the predictors from linear regression and lasso regression, identifying the better model. We found that the neural net with all the predictors resulted in a lower rmse across all 5 iterations. Fitting the neural network on 80% of the data, we resulted in an r-squared of 0.896 with an rmse of 0.362, i.e we were able to account for 89.6% of the data's variance, leading us to conclude that neural networks were the best in predicting housing prices, i.e proving our hypothesis wrong again.

In our project, there were many limitations. For instance, the scope of the data set. The data that we analyzed was taken between 2014 and 2015. We found that sqft_living was the most significant predictor, however, in the recent 6 years, the prices may be more reliant on other factors such as the view or the location. Another limitation is the size of the data. While we did have 21000+ observations, neural networks require large amounts of data, especially with the model and number of neurons we utilized. If we were to perform this project again, if possible, we would find a data set with data from more recent years minimizing any changes that happens with society and possibly combining multiple data sets, so our conclusions are not only limited to Seattle.