Analysis of Home Price Prediction

Nick Chandler, Karla H., Finlay M., Alyssa H.

June 30, 2024

Abstract

In this project, we examined two datasets, real estate and income. The main goal of our analyses was to predict the price of a home given several features of the houses and to determine whether income data could be used as a surrogate for location data when access to observations in a specified location is limited. We used 3 models to do prediction and then did Principal Components Analysis to further analyze the feature space. Our results were mainly that the zip code yields very useful spatial information that cannot be derived from income alone. Additionally, we found that the relationship between the predictors and the response is non-linear in many cases.

1 Background & Significance

Our dataset concerns home prices from a real estate website and income in various regions. Specifically, we have joined the two tables on the zip code predictor. We perform regression on the price of homes given the following predictors: number of bedrooms, bathrooms, the size of the lot (in acres), the size of the house (in square feet), the mean income (in USD), the standard deviation of income (in USD), and zip code.
Our research questions consist of the following:

- How does location relate to home prices?
- How useful are various predictors in prediction of price?

Developing and analyzing a model to predict the price of homes serves as a real estate and development tool for understanding the value of a house. Additionally, we may obtain economic insight into how mean income and income disparities (standard deviation of income) are related to housing characteristics.

2 Methods

2.1 Linear Regression

We fit a linear regression model to the data in R as a baseline for our experimentation

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6$$

where,

$$\langle x_1, x_2, x_3, x_4, x_5, x_6 \rangle = \langle \text{Beds, Baths, House Size, Lot Size, Avg. Income, Std. Income} \rangle$$

In addition to the linear regression model fit on all the data, we also split the dataset on zip-code and fit models to each subset of the data with more than 100 observations. The histogram of the R^2 scores is in the results section.

2.2 Elastic-Net Regression

We also fit an elastic net regression model to the same predictors as the linear regression. We used $\alpha = 0.5$, which means the L_1 and L_2 regularization terms are mixed equally, therefore allowing variable elimination.

2.3 Neural Network

The neural network we implemented was a fully connected feed forward network. The Architecture is:

- Layer widths: $(512, 256, 128, 64, 32, 8, 1)$
- Loss Function: Mean Squared Error
- Optimizer: Adam
- Activation Function: ReLU

We used an ensemble of 10 separately trained neural networks, all with the same architecture. Additionally, we examined performance with and without zip codes passed into the model.

2.4 Random Forest

We fit Random Forest (RF) regression models to the data. We trained with and without spatial information. Since we had $6/7$ predictors, we chose $m = 2$. Finally, we used the elbow method to determine the number of trees to use.

2.5 Principal Components Analysis

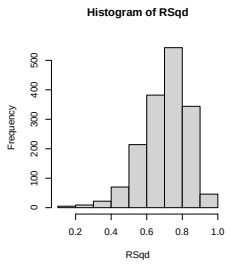
We fit a principal components model to the same predictors as the linear regression.

3 Results

3.1 Linear Regression

The adjusted R^2 score for the linear regression model is: 0.2455. Additionally, there was a statistically significant result indicating that the “Lot Size" variable should be dropped. Finally, the coefficient table is included in the appendix.

This histogram indicates that the majority of models had an R^2 larger than 0.5, indicating moderate performance. Still, there are others with much lower performance which could contribute to the R^2 of the model fit on all the data being low.



3.2 Elastic-Net Regression

The adjusted R^2 for the Elastic-Net model was: 0.2281. This method eliminated the “Lot Size" variable. The coefficient estimates of the elastic net regression are in the appendix.

3.3 Neural Network

The R^2 of the 10 ensembled neural networks with spatial information (zip code) was 0.6214 and 0.6003 for the arithmetic mean and median of the model predictions respectively. The R^2 of the 10 ensembled neural networks with no spatial information was 0.4487 (mean) and 0.4381 (median). The R^2 for each of the neural networks is included in the appendix.

3.4 Random Forest

After trying multiple different values of ntree, we decided that 50 was best although the elbow plot was noisy. The elbow plot is in the appendix. The R^2 for the model with spatial information was 0.6188. The R^2 for the model without spatial information was 0.6121.

3.5 Principal Components Analysis

The scree plot, cumulative proportion of variance explained plot, and biplots are included in the appendix. The principal component loading vectors are in Table 1.

	φ_1	φ_2	φ_3	φ_4
bed	0.4164	0.4576	-0.0006	0.4365
bath	0.4954	0.3477	-0.0038	0.2782
acre lot	-0.0032	0.0077	0.9999	0.0013
house size	0.3724	0.3480	-0.0001	-0.8556
avg income	0.4790	-0.5118	0.038	0.0051
std income	0.4617	-0.5354	0.0077	-0.0073
% of variance	0.3763	0.2526	0.1667	0.1139

Table 1: Principal Component Loading Vectors

Interpretations:

- φ_1: Positive relationship between all variables (except acre lot).
- φ_2: Contrast of house quality (size, # of bed, # of bath) with income.
- φ_3: Lot size component.
- φ_4: Contrast between room amenities and house size.
- 90.95% of variance explained with four components

4 Analysis & Future Work

First, given the meaningful interpretations of the PCA of the data, an avenue for future work could be to re-train all of the models we have but using the scores of each observation instead of the observations themselves. To compare the results of the prediction methods, we present the ranking of prediction accuracy based on R^2 (zip denotes those with spatial information):

$$\text{NN (zip)} > \text{RF (zip)} > \text{RF (no zip)} > \text{NN (no zip)} > \text{Linear Regression} > \text{Elastic-Net Regression}$$

$$\text{R-Squared: } 0.6214 > 0.6188 > 0.6121 > 0.4487 > 0.2455 > 0.2281$$

The performance of the random forest and neural network in comparison to the linear models indicates that there is a likely a non-linear relationship between the predictors and the response since the linear models performed far worse. Of the best models evaluated, neural networks and random forests, we find that including spatial information in the form of zip codes is useful in increasing prediction accuracy, though more so in the case of neural networks. The histogram of R^2 scores for linear models when split on zip code tends to have an R^2 above 0.5, higher than the R^2 for the single linear model trained on all the data. This indicates that spatial information is useful. Our future work is to determine how to encode the spatial information more effectively. In this direction, an approach may be to take the center latitude and longitude of each zip code as features which could prove beneficial for the prediction accuracy of the models. Finally, due to the output of both the linear regression and elastic-net regression, we conclude that the Lot Size is not useful in predicting price while all other variables have some use in this respect. A general conclusion one could draw from this analysis is that there are certain locations which have more easily predictable home prices given the predictors we conducted analysis on.

References

[1] *R Documentation-* https://www.rdocumentation.org/

[2] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. New York: Springer, 2013.

A Appendix

Note that for all methods, we utilized a dev-set approach to estimating the test error via R^2. That is, we held out approximately 20% of the data, trained on the remaining 80%, and ran predictions on the held-out 20% to obtain an estimate of the test R^2.

A.1 Linear Regression

```
Call:
lm(formula = data$price ~ data$bed + data$bath + data$house_size +
    data$acre_lot + data$mean_income + data$stdev_income)

Residuals:
      Min        1Q    Median        3Q       Max
-31205528   -223916    -64825    109539  77676238

Coefficients:
                    Estimate Std. Error  t value Pr(>|t|)
(Intercept)       -5.096e+05  3.534e+03 -144.181   <2e-16 ***
data$bed           2.346e+04  9.410e+02   24.936   <2e-16 ***
data$bath          1.802e+05  9.651e+02  186.755   <2e-16 ***
data$house_size    3.385e+01  5.036e-01   67.227   <2e-16 ***
data$acre_lot     -2.856e-01  1.217e+00   -0.235    0.814
data$mean_income   2.092e+00  4.650e-02   44.983   <2e-16 ***
data$stdev_income  5.659e+00  9.080e-02   62.328   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 570800 on 482623 degrees of freedom
Multiple R-squared:  0.2455,    Adjusted R-squared:  0.2454
F-statistic: 2.617e+04 on 6 and 482623 DF,  p-value: < 2.2e-16
```

Figure 1: Output of Linear Regression

A.2 Elastic-Net Regression

```
                              s1
(Intercept)   -1.580455e+05
bed            5.159915e+03
bath           1.465148e+05
acre_lot       0.000000e+00
house_size     1.044685e+01
mean_income    1.524002e+00
stdev_income   3.282441e+00
```

Figure 2: Elastic Net Regression Output

Note that we also tried to do Elastic-Net Regression on the principal component scores of our observations but obtained an unsatisfactory R^2 of approximately 0.0108. We decided it would not be worth including in the primary analysis.

A.3 Neural Network

```
Beginning the ensemble
1509/1509 [==============================] - 3s 2ms/step
Model number: 0
Validation set Mean Squared Error: 1450751142656.0
Test set R2 score: 0.5977850290067221
1509/1509 [==============================] - 2s 961us/step
Model number: 1
Validation set Mean Squared Error: 158503845888.0
Test set R2 score: 0.5822953162670708
1509/1509 [==============================] - 2s 1ms/step
Model number: 2
Validation set Mean Squared Error: 1854635704032.0
Test set R2 score: 0.4463143531140425 5
1509/1509 [==============================] - 2s 1ms/step
Model number: 3
Validation set Mean Squared Error: 1510803701760.0
Test set R2 score: 0.6197190382873 86
1509/1509 [==============================] - 2s 990us/step
Model number: 4
Validation set Mean Squared Error: 147096338432.0
Test set R2 score: 0.6247605252975033
1509/1509 [==============================] - 2s 2ms/step
Model number: 5
Validation set Mean Squared Error: 145172856832.0
Test set R2 score: 0.5948213597882537
1509/1509 [==============================] - 2s 1ms/step
Model number: 6
Validation set Mean Squared Error: 149032960000.0
Test set R2 score: 0.6018149715406054
1509/1509 [==============================] - 2s 962us/step
Model number: 7
Validation set Mean Squared Error: 143173943296.0
Test set R2 score: 0.5987282471705409
1509/1509 [==============================] - 2s 1ms/step
Model number: 8
Validation set Mean Squared Error: 142982971392.0
Test set R2 score: 0.6247868210559897
1509/1509 [==============================] - 3s 2ms/step
Model number: 9
Validation set Mean Squared Error: 145152966656.0
Test set R2 score: 0.61557577 12642653
Average Validation MSE: 1512734496576.0
Average R-Squared: 0.5906601432792381
Ensembled R-Squared (Arithmetic Mean): 0.62143934296619 51
Ensembled R-Squared (Median): 0.6002716093555731
```

```
Model number: 0
Validation set Mean Squared Error: 231268630528.0
Test set R2 score: 0.4391479330021162
1509/1509 [==============================] - 1s 948us/step
Model number: 1
Validation set Mean Squared Error: 266148593664.0
Test set R2 score: 0.4471184968771049
1509/1509 [==============================] - 2s 1ms/step
Model number: 2
Validation set Mean Squared Error: 229789745152.0
Test set R2 score: 0.44997360663717056
1509/1509 [==============================] - 2s 1ms/step
Model number: 3
Validation set Mean Squared Error: 233937240064.0
Test set R2 score: 0.4369616739456339
1509/1509 [==============================] - 2s 1ms/step
Model number: 4
Validation set Mean Squared Error: 2345842 76992.0
Test set R2 score: 0.43410246863202 24
1509/1509 [==============================] - 2s 1ms/step
Model number: 5
Validation set Mean Squared Error: 233755820032.0
Test set R2 score: 0.43555757084039115
1509/1509 [==============================] - 2s 1ms/step
Model number: 6
Validation set Mean Squared Error: 231176880128.0
Test set R2 score: 0.43191852917252427
1509/1509 [==============================] - 3s 2ms/step
Model number: 7
Validation set Mean Squared Error: 281248956416.0
Test set R2 score: 0.33903891043827095
1509/1509 [==============================] - 2s 1ms/step
Model number: 8
Validation set Mean Squared Error: 236476694528.0
Test set R2 score: 0.43948988836155 08
1509/1509 [==============================] - 4s 2ms/step
Model number: 9
Validation set Mean Squared Error: 236035981312.0
Test set R2 score: 0.4398774603926 5363
Average Validation MSE: 2414422281881.6
Average R-Squared: 0.42931864482994386
Ensembled R-Squared (Arithmetic Mean): 0.4486949053392
Ensembled R-Squared (Median): 0.4380548034738 7504
```

Figure 3: Output of Ensemble - Spatial Info (Left), No Spatial Info (Right)

A.4 Random Forest

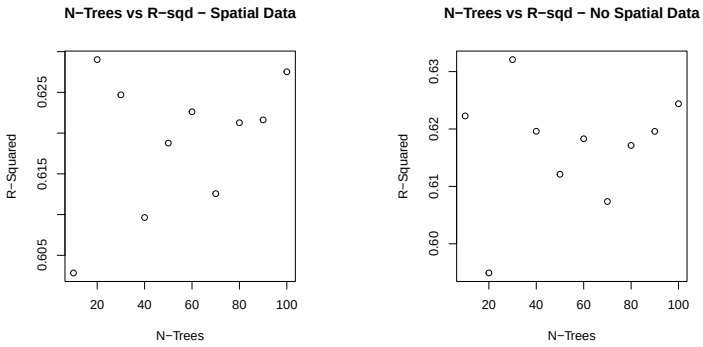


Figure 4: R-sqd vs. N-Trees

A.5 Principal Components Analysis

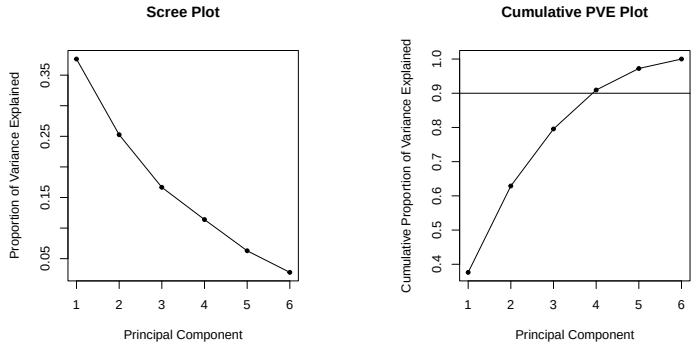


Figure 5: Scree and Cumulative Prop. of Var. Plots

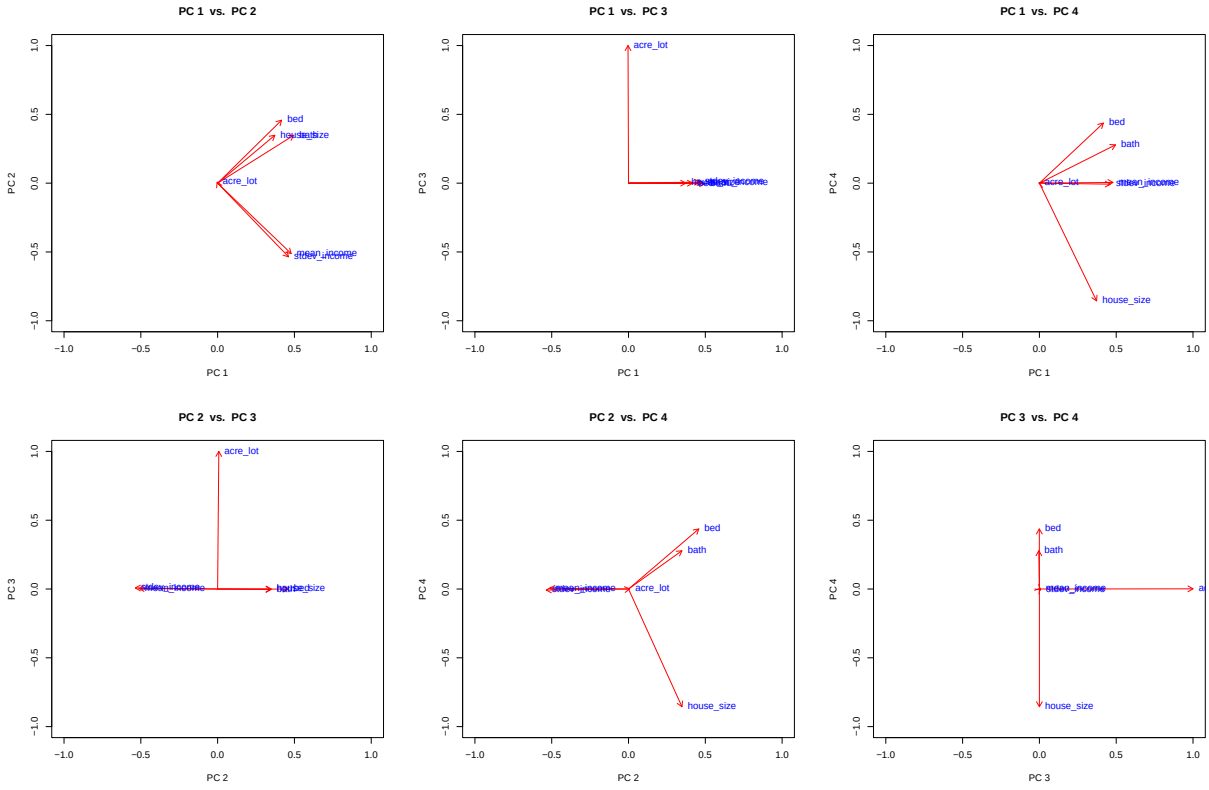


Figure 6: Biplots - Only Loading Vectors

We attempted to use the biplot() function in R but since we had too much data, the points would not plot properly. Therefore, we implemented a function that plots the appropriate vectors without observations manually.