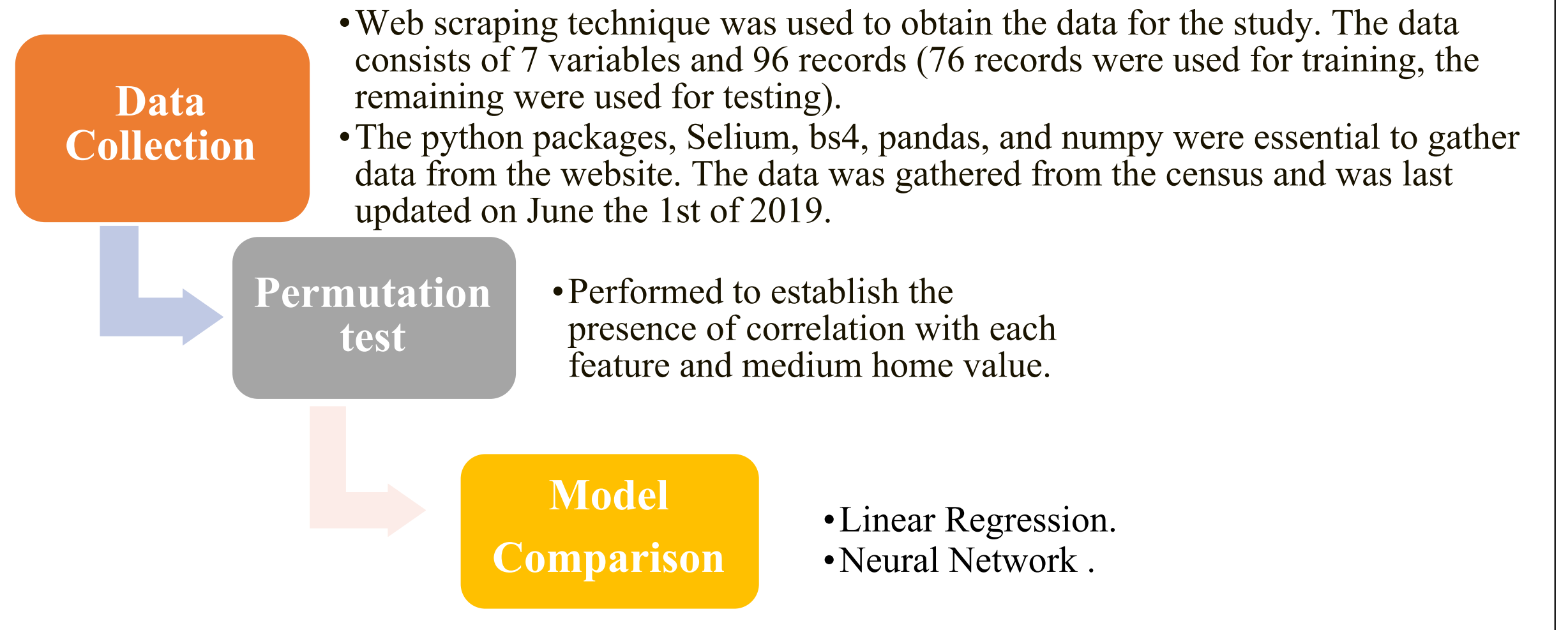
Neural Networks and Linear Regression to Predict Median Home Value

Elda Piedade
Faculty Advisor: Mitsue Nakamura,
Department of Mathematics and Statistics
University of Houston Downtown, TX.

Abstract

The objective of this research is to implement linear regression and neural network algorithms to predict the median home value for the houses in Houston zip codes. Data was gathered from 96 zip codes in Houston by utilizing python web scrapping resources to collect data from the Texas Hometown Locator website (owned by HTL, Inc.).With the dataset extracted and cleaned, exploratory data analysis and statistical analysis were performed to understand the relationship between the median home value and other variables, such as diversity index, per capita income, and average household size. Based on the analysis, data was modeled with linear regression and neural networks, with the goal of assessing which model yields the best results. Initial results from the linear regression model indicate that the diversity index and average household size are not significant predictors. On the other hand, the per capita income identifies as the best predictor. Additionally, although previous assumptions, the diversity index is only moderately negatively correlated with the median home value. In other words, the more diverse a zip code is, the slightly smaller is the median home value. Further, into this research, the population in focus will extend to other Texas cities, such as Austin. With more data, the relationship between variables and median home value can be better evaluated as well as the prediction capacity of each model.

Project Overview



Data Collection

- Web scraping technique was used to obtain the data for the study. The data consists of 7 variables and 96 records (76 records were used for training, the remaining were used for testing).
- The python packages, Selium, bs4, pandas, and numpy were essential to gather data from the website. The data was gathered from the census and was last updated on June the 1st of 2019.

Permutation test

- Performed to establish the presence of correlation with each feature and medium home value.

Model Comparison

- Linear Regression.
- Neural Network .

Variables

1. The Diversity Index is a scale of 0 to 100 that represents the likelihood that two persons, chosen at random from the same area, belong to different races or ethnic groups. If an area's entire population belongs to one race and one ethnic group, then the area has zero diversity. An area's diversity index increases to 100 when the population is evenly divided into two or more race/ethnic groups.
2. The Housing Affordability Index base is 100 and represents a balance point where a resident with a median household income can normally qualify to purchase a median price home. Values above 100 indicate increased affordability, while values below 100 indicate decreased affordability.
3. The outcome variable is the median home value (median price home).

$$\begin{aligned} x_1 &= Total.Population \\ x_2 &= Diversity.Index \\ x_3 &= Median.Household.Income \\ x_4 &= Per.Capita.Income \\ x_5 &= Total.Housing.Units \\ x_6 &= Average.Household.Size \\ x_7 &= Housing.Affordability.Index \end{aligned}$$

Methods

Permutation test of the correlation:

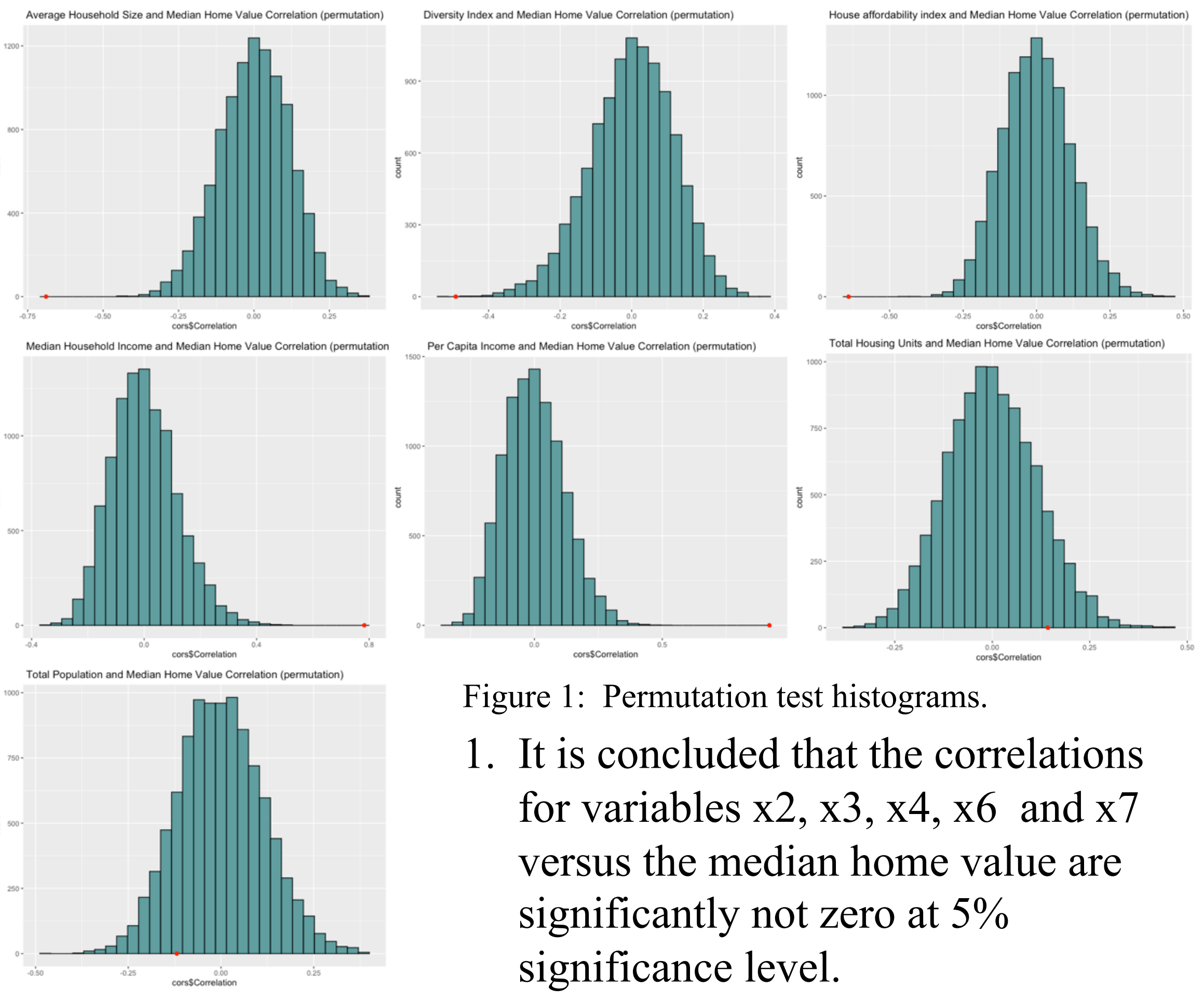


Figure 1: Permutation test histograms.

1. It is concluded that the correlations for variables x2, x3, x4, x6 and x7 versus the median home value are significantly not zero at 5% significance level.

Linear Regression Models:

Model 1:

$$(1.043e+05)+(1.120e-01)x_1-(4.427e+02)x_2+(9.801e-01)x_3+(5.573e+00)x_4$$
$$-(1.590e+00)x_5+(4.050e+04)x_6-(1.235e+03)x_7 = Median.Home.Value$$

Model 2 :

$$(1.187e+01)+(5.001e-07)x_1+(6.240e-03)x_2+(8.985e-06)x_3+(1.033e-05)x_4$$
$$-(4.255e-06)x_5-(8.499e-02)x_6-(5.618e-03)x_7 = log(Median.Home.Value)$$

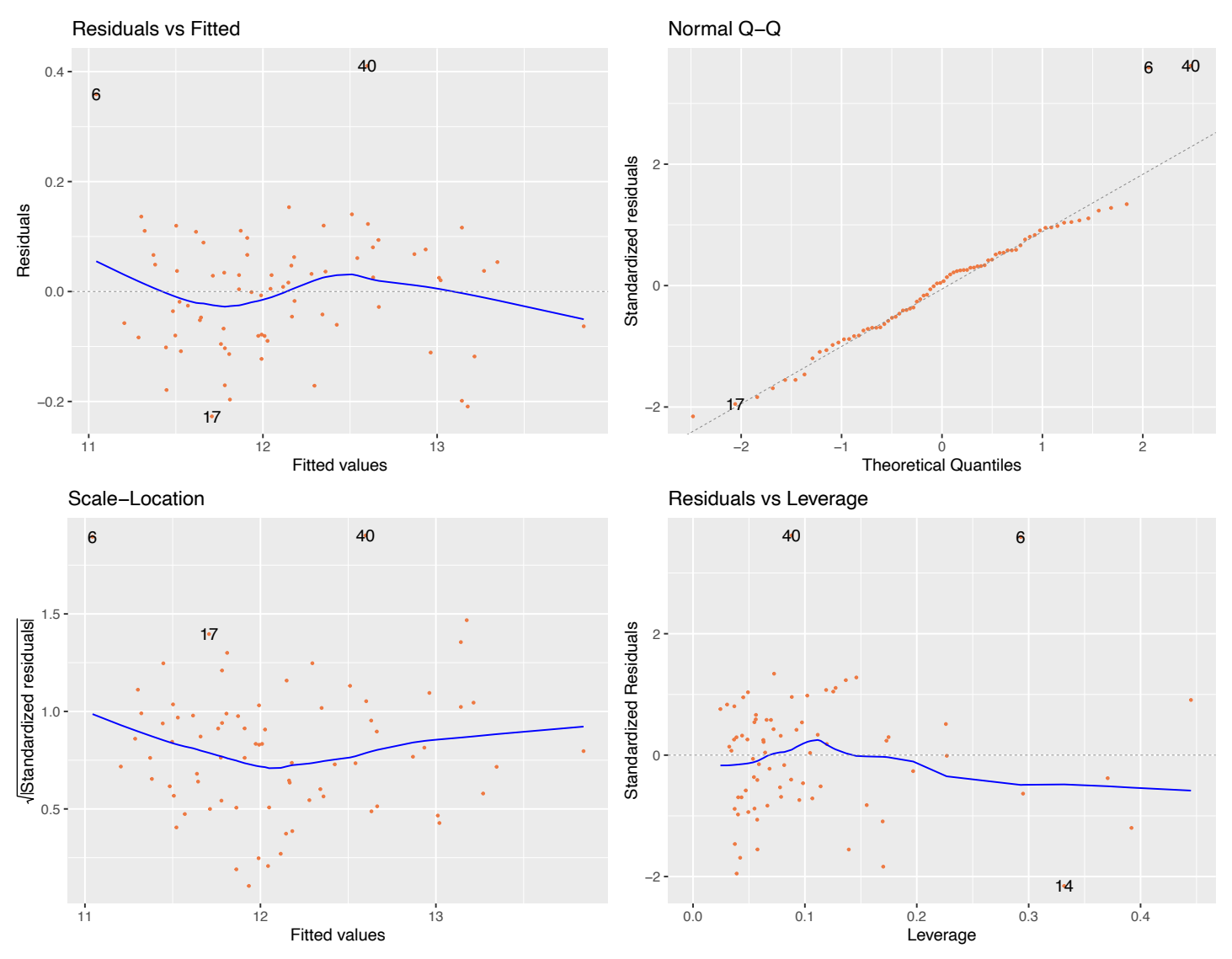


Figure 2: Model 2 residual plots.

Neural Network Model:

Layers:

- The input layer (7).
- Hidden layer(4).
- Output layer(1).

Model Compiler:

- ADAM optimizer.
- MSE loss function.

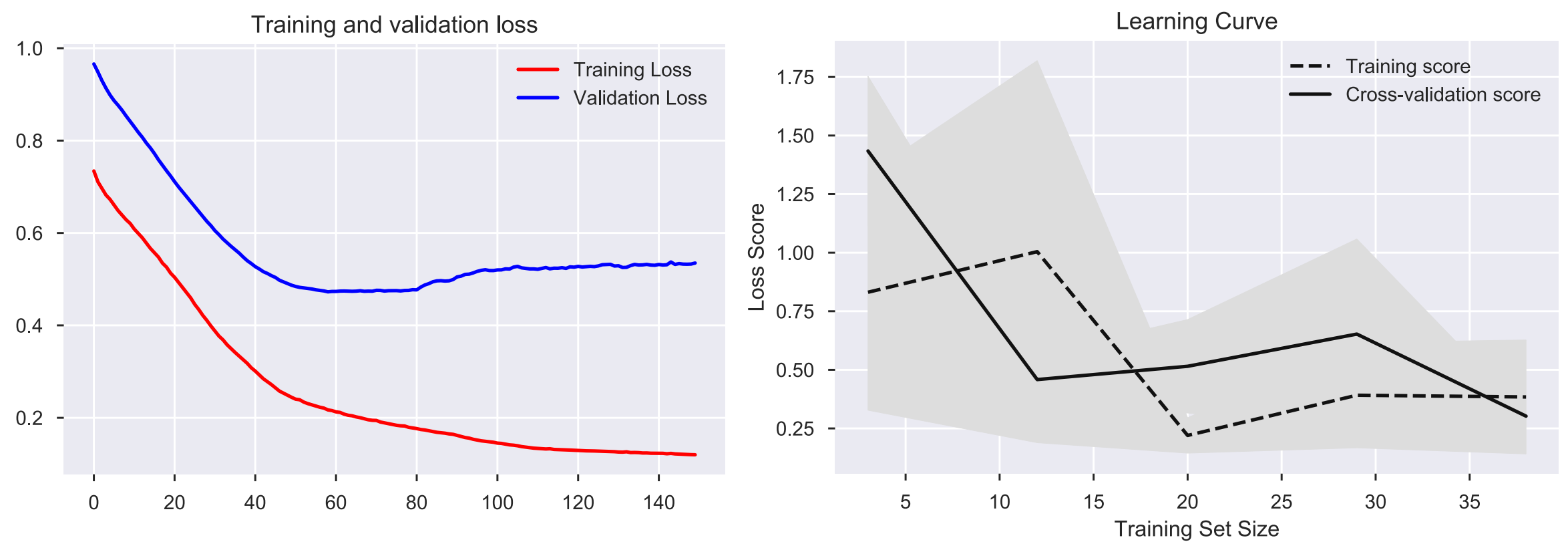
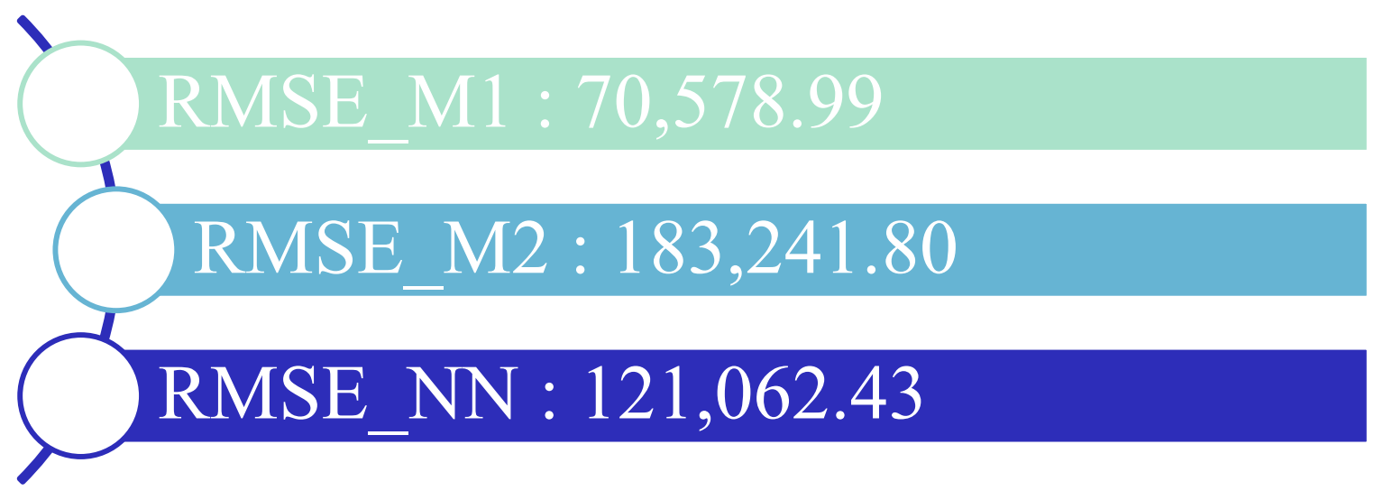


Figure 3: Neural network loss and learning curves.

Conclusion

- Model 1 was not a good fit for the data because it violated the homoscedasticity, normality of residual and linear relationship assumptions.
- Model 2 is a model that consists of the log transformation of the outcome variable. This model did not violate any assumptions; thus it was considered the best model for significance testing.
- The neural network model produced a similar root mean squared error. In other words, it did not perform better than the previous models due to the underrepresentation of the data.

Results



RMSE_M1 : 70,578.99

RMSE_M2 : 183,241.80

RMSE_NN : 121,062.43

- Artificial neural networks uses randomness while being fit on a dataset, such as random initial weights, due to this randomness the mean RMSE was reported for better assessment.

The most significant variables according to Model 2 are:

- Diversity Index
- Median Household Income
- Per Capita Income
- Housing Affordability Index

R-squared:

- 0.9657
- 96% of the median home value variability is explained by the model.

Future Plans

- Generate a larger dataset to further access the model for prediction purposes.
- Perform variable selection for the best linear regression model.
- Modify neural network parameters to get better results.

References

1. Kassambara, A. (2018). *Machine Learning Essentials: practical guide in r*. S.l.: CREATESPACE PUBLISHING.
2. Chihara, L., & Hersterberg, T. (2011). *Mathematical statistics with resampling and R & Probability*. Hoboken, NJ: Wiley.
3. Houston, Texas (TX) ZIP Code Map. (n.d.). Retrieved from https://texas.hometownlocator.com/zipcodes/zipcodes,city,houston.cfm.
4. Getting started with the Keras Sequential model. (n.d.). Retrieved from https://keras.io/getting-started/sequential-model-guide/.

Acknowledgments

Professor Mitsue Nakamura, Dr. Mary Jo. Parker, Dr. Katherine Shoemaker UHD Scholars Academy and the of the Governor Wagner-Peyser workforce grant award administered by the Texas Workforce Commission (2819WPB002).