**HOUSE PRICE PREDICTION**

**Arindam Samanta**

Bellevue University

Bellevue, NE 68005, USA

**David Thampi Selvaraj**

Bellevue University

Bellevue, NE 68005, USA

**Rohan Valder**

Bellevue University

Bellevue, NE 68005, USA

# Abstract

2021 and 2022 housing market trends and forecasts shows an extraordinarily strong real estate market, with strong housing demand in every part of the country. The improved economy and millennials entering their peak home buying years are fueling a residential housing boom. Along with this, low mortgage rates, combined with an increase in working from home because of the pandemic, have spurred a surge in housing demand, particularly in lower-density suburbs. In 2020, more existing homes were sold than in any other years since 2006.

As a team, we were curious about estimating a house price based on the various features available in the dataset. Some of the fields did not make any difference in the dataset that we had to drop after a complete data analysis. We built various models using Python and R to make house price predictions.

Random Forest had an accuracy of 85.4% and Lasso Regression of 81%. Both models were built in Python. The Linear Regression Model was built using R had an accuracy of 77.8%.

# Intro/background of the problem

We took this topic for predicting house prices under such a housing boom so that people could make better decisions for home investments. Each succeeding research has generally improved the predictive power of the model by emphasizing the attributes of the property such as the housing site, housing quality, geographical location, and environment.

Our object is to discuss major factors which influence the housing price and make precise predictions for it. The model provides price prediction of similar comparable of people’s dream homes, such that the buyers and sellers can better negotiate the price as per the market trend. This predictive model is very relevant as the median existing home price for all housing types in May was $350,000, 23.6% up from last May ($283,500) as every region registered price increases.

For the selection of the prediction method, we compare and explore various prediction methods.

## Questions that we are trying to answer

The study answers the below research questions:

Question 1: Which machine learning algorithm performs better and has the most accurate result in predicting home prices? And why?

Question 2: What are the factors that affect the house price?

# Methods

The more the houses have to offer, the higher the price is. The dataset has a lot to offer, and there are many ways to go about it.

For this project, we will model various predictive process data related to house price datasets. The datasets come from Kaggle. We can achieve this by following the below steps:

* Loading Data
* Data preprocessing and Cleaning
* Data exploration and visualization
* Data modeling

Evaluating the model

# Results

# The dataset is interesting. There are lots to learn from the housing-related

# lessons. This dataset will be used for the model for house price prediction

# based on the location. Our goal for the predictive model is to have precise

# and accurate results. We will continue to focus on tuning the model as the

# course goes on and work on improving the process. The project needs to use

both R and Python.

# Discussion/Conclusion

In this case study, we will be using machine learning algorithms to predict the house prices. We are planning to use a data set of many houses with several parameters. We are thinking of using 70 percent of the data set to train the model and 30 percent to test the model. Expecting the results are truly accurate. And we will be testing it with different parameters also. We will be implementing the step-by-step procedure to analyze the dataset and finding the correlation between the parameters. So that, we can select the parameters which are not correlated to each other and are independent in nature. These feature set are then given as an input to four algorithms and a csv file will be generated consisting of predicted house prices. Hence, we calculate the performance of each model using different performance metrics and compare them based on these metrics. We are planning to use following model

* (Decision Tree) gives the highest accuracy of <>.
* (Lasso) gives the least accuracy of <>.
* (Logistic Regression) gives an accuracy of <>.
* (Support Vector Regression) gives an accuracy of <> respectively.

Thus, we conclude that we implemented classifiers to the problem of regression to check how well can classifier fit to regression problem.

# Acknowledgements

# Google Scholar and Google for being able to quickly identify valuable resources that can be read and referenced.

# Data Science bloggers, researchers, and contributors of knowledge.

# Professor Fadi Alsaleem for their development and delivery of the Predictive Analytics course.

# Group members for their involvement and great ideas to complete the study.

# References

Data Sources: Identified following sources for data, will be decided one of the sources based on the key features and data validation.

<https://www.kaggle.com/seif2018/house-price-prediction/comments>

<https://www.kaggle.com/c/zillow-prize-1/overview>

US Housing Market Forecast 2021-2022: Will It Crash or Boom?

https://www.noradarealestate.com/blog/housing-market-predictions/