State and Region Based Housing Price Prediction Report

Members: Justin Nichols, Andrew Rodrigue, Zeke Abshire

CSC 3730 Final Project

Instructor: Mingxuan Sun

2 December 2021

Table of Contents

Preface

The ultimate goal of this project is to predict the housing prices of a certain region or state within the United States. The user will be able to enter their income and desired house age, number of rooms, number of bedrooms, and population of the area they are prospecting. The program will run both the KNN and Linear Regression predictive analysis models and output the predicted housing price according to the inputs from the user. The data used within this predictive application contains USA housing data from USA_Housing.

Motivation

The motivation behind this project is to provide users the ability to obtain housing prices for any region or state in the United States to help them better understand market prices. Whether someone is looking to move to a particular part of the country or is just inquisitive about the various housing prices around the United States, this housing price predictive model enables them to obtain accurate results based on the specific information they provide.

As of late 2021, the housing market within the United States has become an unlikely beneficiary of the COVID-19 pandemic. Presently, the housing market is booming, with properties being sold in record time. With this, home prices have climbed at a fast pace. According to CNBC, the median price for resold homes has reached over $363,000 as of June 2021. Due to a number of factors such as a boost in demand and low mortgage rates, supply is shrinking and prices are up the most they have been since the 1970s (CNBC).

The motivating factor behind the development of this housing predictive model was to provide prospective homebuyers transparency in housing prices around the country. By providing users the ability to pick a region or state and compare home prices based on their personal information, this model enables them to quickly perform accurate research and make educated decisions on where they want to live.

Problem Definition

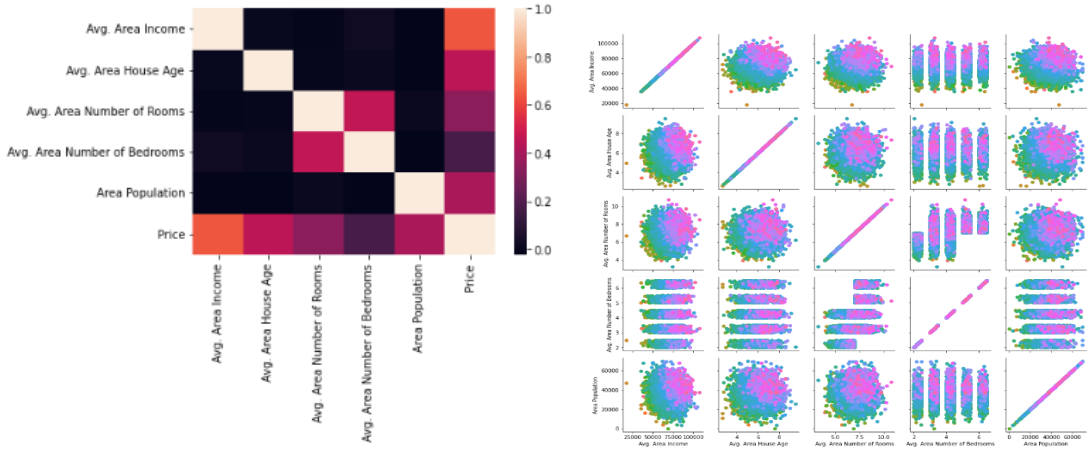
The housing market within the United States has drastically changed over the past year and with this, the prices of homes have been on the uptrend. Prospective homebuyers have been left in the dark on current house pricing trends and some have been subjected to overpaying on homes in certain areas. To solve this problem, this housing price predictive model calculates the price of a home in a particular area by allowing homebuyers to input an array of information. By using the KNN and Linear Regression predictive models, this model removes any misinterpretation of housing prices around the country.

Dataset

The dataset contains housing data and was obtained from Kaggle using the link USA_Housing. It is stored in a CSV formatted file. The dataset contains 5000 unique entries. The attributes of the dataset include average area income, average area house age, average area number of rooms, average area number of bedrooms, area population, price, and address. It is important to note that for privacy reasons, all data is not authentic but instead based on real addresses.

	Avg. Area Income	Avg. Area House Age	Avg. Area Number of Rooms	Avg. Area Number of Bedrooms	Area Population	Price	Address
0	79545.458574	5.682861	7.009188	4.09	23086.800503	1.059034e+06	208 Michael Ferry Apt. 674\nLaurabury, NE 3701...
1	79248.642455	6.002900	6.730821	3.09	40173.072174	1.505891e+06	188 Johnson Views Suite 079\nLake Kathleen, CA...
2	61287.067179	5.865890	8.512727	5.13	36882.159400	1.058988e+06	9127 Elizabeth Stravenue\nDanieltown, WI 06482...
3	63345.240046	7.188236	5.586729	3.26	34310.242831	1.260617e+06	USS Barnett\nFPO AP 44820
4	59982.197226	5.040555	7.839388	4.23	26354.109472	6.309435e+05	USNS Raymond\nFPO AE 09386

Attribute Correlations



Data Analysis Algorithms

The two predictive models used include K-Nearest Neighbor and Linear Regression. K-Nearest Neighbor, or KNN, is a classification prediction method that groups a value based on the group of the data points nearest to it. Linear Regression models the relationship between two variables by fitting a linear equation to the observed data.

KNN

In the housing predictive model, KNN was used to predict a house price based on user-defined input. First, the relevant data from a Pandas DataFrame was partitioned into price interval

categories of $150,000. Then, the data was normalized and split into training and testing data. Using the normalized data, the various values for KFold and KNN were tested to find the optimal accuracy for the model. The best KFold and KNN value was used for the prediction. The model outputs a predicted price range and the classification number to the user along with plots describing the best K for KFold and the best K for KNN.

Linear Regression

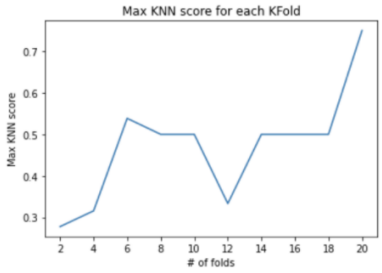
In the housing predictive model, Linear Regression was used to predict a house price based on user-defined input. First, the relevant data from a Pandas DataFrame was used for the prediction. Then, the data was split into training and testing data. The model outputs a predicted price to the user along with a plot comparing the actual price versus the predicted price.

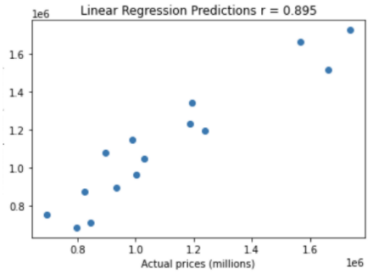
Experiments

Experiment 1

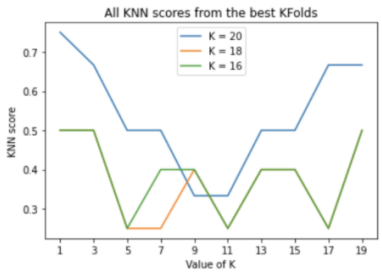
For Louisiana,

```
Obtain housing price predictions.
Enter income: 50000
Enter house age: 8
Enter number of rooms: 5
Enter number of bedrooms: 1
Enter population: 10000
Select an Option:
1.U.S.
2.State
3.Region
2
Input a state abbreviation: la
House Prediction for LA:
```






```
near Regression: $530000
```



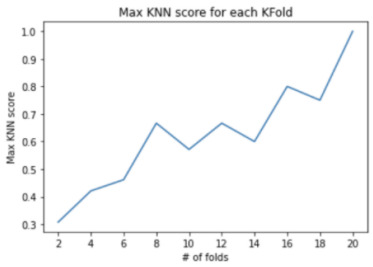

```
KNN: $1650000.0 - $1800000.0 | Classifier output: 11.0
```

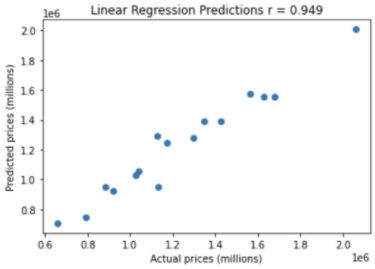
Note, the data used is not based on real prices for the area.

Experiment 2

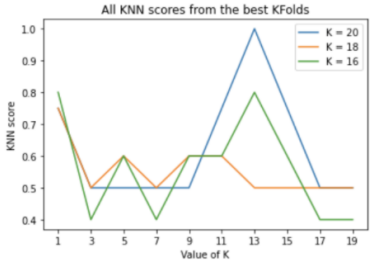
For California,

```
Obtain housing price predictions.
Enter income: 120000
Enter house age: 3
Enter number of rooms: 8
Enter number of bedrooms: 3
Enter population: 50000
Select an Option:
1.U.S.
2.State
3.Region
2
Input a state abbreviation: ca
House Prediction for CA:
```






```
Linear Regression: $2300000
```



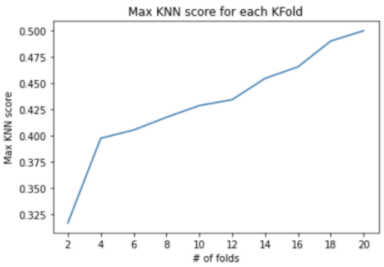
Note, the data used is not based on real prices for the area.

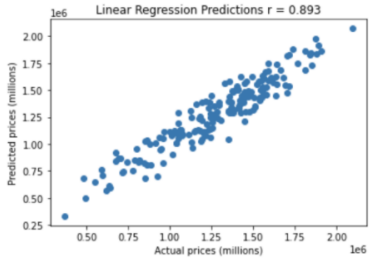
```
KNN: $1500000.0 - $1650000.0 | Classifier output: 10.0
```

Experiment 3

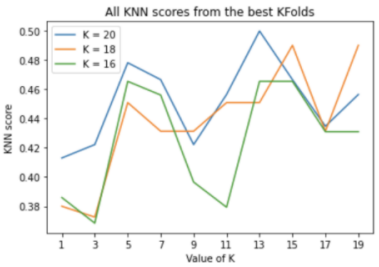
For Northeast,

```
Obtain housing price predictions.
Enter income: 80000
Enter house age: 6
Enter number of rooms: 7
Enter number of bedrooms: 3
Enter population: 35000
Select an Option:
1.U.S.
2.State
3.Region
3
Input a region (northeast, midwest, south, west):
northeast
House Prediction for northeast:
```






```
Linear Regression: $1460000
```



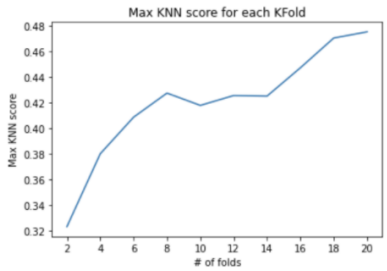
Note, the data used is not based on real prices for the area.

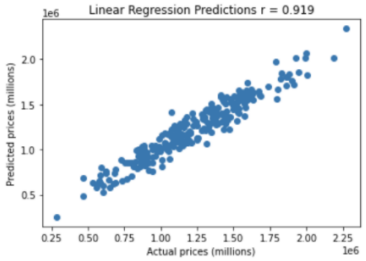
```
KNN: $1800000.0 - $1950000.0 | Classifier output: 12.0
```

Experiment 4

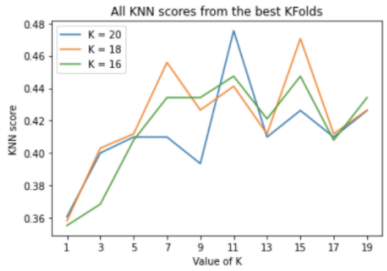
For South,

```
Obtain housing price predictions.
Enter income: 80000
Enter house age: 6
Enter number of rooms: 7
Enter number of bedrooms: 3
Enter population: 35000
Select an Option:
1.U.S.
2.State
3.Region
3
Input a region (northeast, midwest, south, west):
south
House Prediction for south:
```






```
Linear Regression: $1470000
```



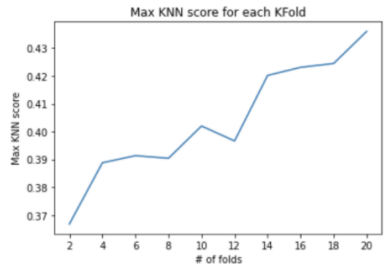
Note, the data used is not based on real prices for the area.

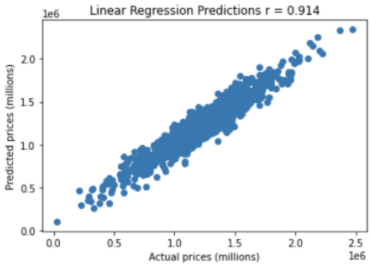
```
KNN: $1800000.0 - $1950000.0 | Classifier output: 12.0
```

Experiment 5

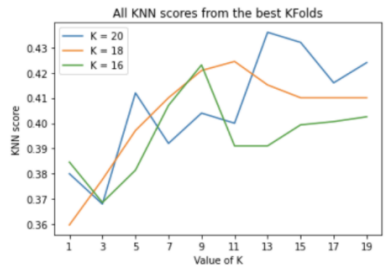
For U.S.,

```
Obtain housing price predictions.
Enter income: 80000
Enter house age: 6
Enter number of rooms: 7
Enter number of bedrooms: 3
Enter population: 35000
Select an Option:
1.U.S.
2.State
3.Region
1
House Price Prediction for US:
```






```
Linear Regression: $1460000
```



Note, the data used is not based on real prices for the area.

```
KNN: $2100000.0 - $2250000.0 | Classifier output: 14.0
```

Conclusion

Conclusively, for the housing predictive model, Linear Regression was found to be more accurate in predicting the housing price for a certain area than KNN. For all experiments, the accuracy score for KNN held around the low to mid 40 percent, whereas the accuracy score for Linear Regression held around the high 80 to low 90 percent.

The insights made from the experiments helped our group to conclude that some states or regions have on average higher home prices. We also learned that for this particular predictive model (housing prices), Linear Regression is a more accurate model to use over KNN.

How to Use the Predictive Model

To use the housing price predictive model, go to Google Colab and upload the Jupyter notebook (i.e. housing_prediction.ipynb). Click "Runtime" from the navigation bar and hit "Run all" to execute all the cells within the notebook. The last cell contains a function call to obtain user input. Within this cell, the user will first be asked to input their income along with the house age, number of rooms, number of bedrooms, and the surrounding area's population of the house they are looking for. Then, the user can obtain predicted house prices for either the entire U.S., a particular state, or a certain region. After giving their desired specifications, the program will output the housing price predictions.

Sources

https://www.kaggle.com/vedavyasv/usa-housing.

https://www.cnbc.com/2021/09/02/heres-why-experts-believe-the-us-is-in-a-housing-boom-not-a-bubble.html