DSC 630 Predictive Analytics

Josh Gardner

28 May 2019

**Course Project: Predicting What Price Range I Should Look at When Buying a House**

Josh Gardner

Bellevue University Data Science 630: Predictive Analytics

Executive Summary

As a part of the American Dream, buying and owning a house is a goal many adults have and cherish, myself included. Over the past year, I have been researching the housing market to see if I could afford to buy a house, and if I could, how expensive of a house I could buy.

As a part of the coursework of Bellevue University’s DSC 630: Predictive Analytics class, we are challenged to find a project to work on using the methodologies we learned during the class. This gave me the opportunity to merge my personal research with the methodologies I was learning in class.

Using the data that was contained in the American Housing Survey, I decided to predict what an individual’s minimum and maximum spending limits were to buy a house. The way the minimum and maximum values were determined was by first grouping the data from the American Housing Survey together based upon how similar the different points of data were to each other. Once the data was grouped together, an individual’s data could be entered into the data to see which grouping of data the input data was closest to. The input data was then considered a part of the group it was closest to.

The group could then be analyzed to identify what the lowest priced house was for the individuals contained within the group. The same method could also be used to identify what the highest priced house was within the group. Because the data is grouped together based on similarities within the data, the minimum and maximum values can be applied to the input data that was also grouped into the same group. In this way, an estimate of what an individual’s minimum and maximum housing price range can be estimated.

Background

For the Course Project, I will be working on the data set from the American Housing Survey: Housing Affordability Data System. The data set can be downloaded from <https://www.huduser.gov/portal/datasets/hads/hads.html>. I will be looking at the ASCII Version of the 2013 data. If needed, I will also use the previous year’s data. Data is available for every other year beginning in 1985. The past data would let us explore how trends within the data have changed over time.

The American Housing Survey started as the Housing Affordability Data System and began in 1985. In 2002, the Housing Affordability Data System became the American Housing Survey. Both data systems categorize housing units by affordability and housing by the Adjusted Median Income, Fair Market Rent, and poverty income. They also include the housing cost burden for both the homeowner and renters. The purpose of the American Housing Survey is to measure the affordability of housing units and the housing cost burdens of households, relative to the area median incomes, Fair Market Rent, and the poverty income.

The Housing Affordability Data System compiled this data from 1985 through 2009. The American Housing Survey compiled this data from 2002 onward with the ability to link the data to the Housing Affordability Data System. The Housing Affordability Data System grew out of a project for the Millennial Housing Commission to provide similar tabulations of the affordability of housing units and the housing cost burdens of households for the years 1985, 1995, and 1999. The Affordability Data System incorporates over 20 years of housing data, offering great insights into housing affordability.

The data that is used in the American Housing Survey is collected from national sample microdata that is collected for the odd numbered years from 1985 to 2009. Poverty levels are gathered from and based on the Census Bureau’s officially released thresholds for poverty. Housing costs for homeowners is calculated based on the monthly mortgage costs, utilities, property insurance, and property tax. Housing costs for renters is defined as the contracted cost of rent plus the cost of utilities.

The Housing Cost Burden is calculated by dividing the household’s monthly housing costs by its monthly income. However, mortgage costs are not included in the household’s monthly housing cost.

In all, there are 99 variables that are captured within the data set. A few of the variables are only used during certain years and are marked accordingly within the data dictionary associated with the data set.

Problem Statement

Using the collected data, I would like to predict what y personal price range for buying a house is. I have begun looking to buy a house and would like to know my upper and lower limits to my house budget. To take the analysis further, I would like to predict what an individual’s price range for buying a house is based on that individual’s personal details. The predictive model inputs with consist of continual, discrete, and categorical values.

To produce the predictive model, I will examine the data that is included in the American Housing Survey, starting with the year 2013. I will be examining the data from after the Financial Crisis of 2008. Limiting the time frame to be the data after 2008 will give three data sets to explore, namely for the years 2009, 2011, and 2013. The data from the previous years will provide a model for how the different trends within the data change over time.

Methods

This project was split into 7 main sections, namely that of the Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, Deployment, and Summary / Conclusions. Each step was completed in order to maintain a set structure on the analysis.

Python and R were both used throughout the project. R was primarily use for data exploration and data cleaning while Python was primarily used to generate and test the different models. The data was saved in a .csv file to facilitate moving the data between the two different languages.

There were many issues encountered with the programming of the models, which was primarily due to a lack of experience in building models in Python. These issues were addressed and overcome by research and experimenting with the models throughout the course of this project.

Business Understanding

The first step to the project was understanding what I was looking for within the data. The data contains information about the Fair Market Price, the Adjusted Median Income, and Poverty Levels, to name a few variables.

Before getting into the data, it is important to know that housing prices have been rising year over year. As housing costs continue to rise, the prospect of buying a house can become slimmer for individuals who are attempting to buy their first house. There are multiple governmental assistance programs available to help first time home buyers, but it can still be a daunting task to figure out what a person’s upper and lower limits are to their house buying price range.

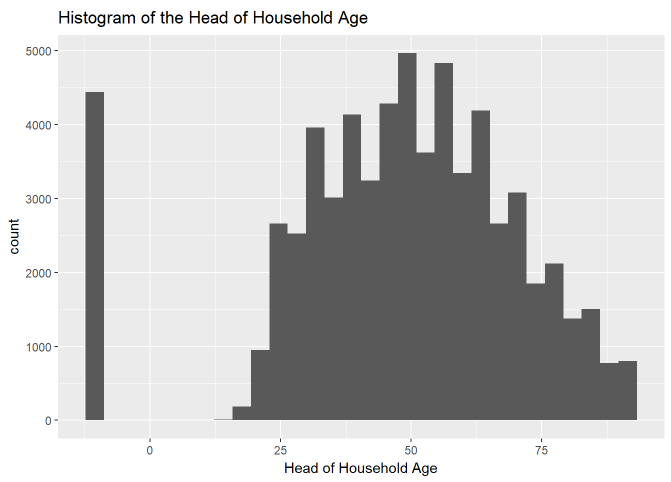
At the core of this project, I would like to identify if I would be able to afford to buy a house. However, to make this project to be of more use to others, I plan to develop a model where an individual would be able to enter in their personal details to see what housing price range others in a similar situation currently have spent on their house or rent.

My research for this project centers around grouping the gathered data together into separate clusters. Once the clusters are developed, an individual using the model would be able to enter their personal details into the model to see which cluster their information belongs to. The minimum and maximum housing costs associated with the individual’s cluster would then be reported as the minimum and maximum price range for the individual.

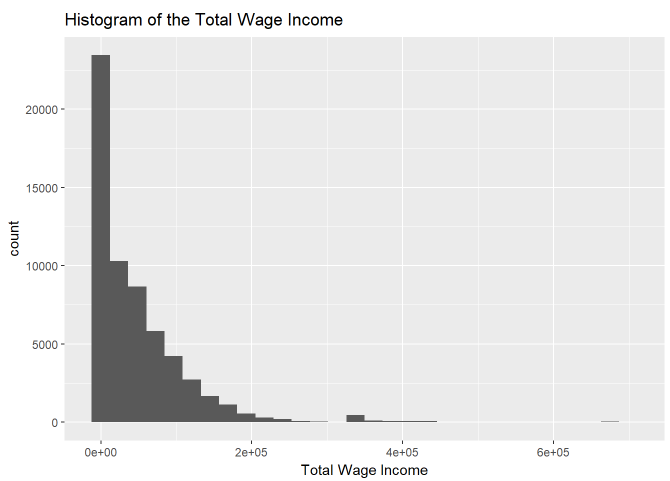
Data Understanding

Once the Business Understanding and what was being looked for was defined, I needed to understand what was contained within the collected data. The data from the American Housing Survey contained 99 variables with a few variables describing similar, if not duplicate items. The data contained 64,534 rows of data.

One of the first things I examined was the missing data that was contained within the data. The data from the American Housing Survey did not contain any NaN, N/A, null, or empty values. Rather, the missing values were given a specific value for each variable. For example, when the age of the head of household was missing a value, a -9 was entered as the value for the variable.

The missing values occurred when there was an error combining the data from the Housing Affordability Data System and the American Housing Survey. The errors were caused by an inability of the two different systems to communicate fully. When I examined the missing data, the missing values were from the Housing Affordability Data System. This means that the missing data was dated information, or the older housing information.

Unfortunately, I was not able to retrieve the missing data. Because the missing data was the dated information and because I was unable to retrieve it, I decided to exclude the missing data from the analysis.

The next thing I observed within the data were the outliers. Most of the variables contained a large variability and ranges. Because of this, most of the variables contained outliers. Perhaps it’s not surprising that the most notable outliers were contained in the Annual Salary variable. The majority of the reported Annual Salaries were below $200,000. This made the Annual Salaries that were over $1,000,000 outliers.

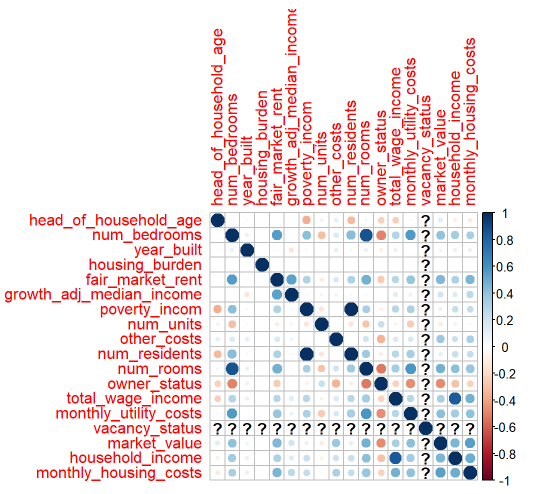
When evaluating the outliers, I opted to keep the outliers in the data. My reasoning is two-fold. First, I wanted to keep the variability within the data in the hope of being able to capture more distinct clusters when the data was clustered. My hope was that the variability within the different variables would allow better overall clusters to be developed.

The second reason I kept the outliers in the data was on the off chance that the individual user that would use the model after it was deployed would contain one of the outlier values. If the model was built using the outliers, the individual’s input data would have a more accurate group to be clustered into. For these individuals, keeping the outliers in the data would produce a more accurate housing price range.

The next thing I explored in the data was the one and two-dimensional relationships contained within the data. I plotted each variable’s histogram and boxplot to view each variable’s distribution. I also calculated the correlations between each of the variables to see which variables were highly correlated.

Data Preparation

Once I had a good understanding of what was contained within the data, I needed to prepare the data and get it into the correct format. This stage took the most time as it did include cleaning the data and performing the Exploratory Data Analysis, which results were described in the previous section.

I completed the data cleaning within R. The data was initially in a .txt format. I initially imported the data into a R Data Table. I quickly changed the Data Table into a Data Frame for the ease of working with Data Frames and my familiarity with Data Frames. I was able to verify that there were no null values by using the is.na() function in R. 

I also generated the various histograms and boxplots within R using the ggplot2 library to explore the different one-dimensional relationships contained within the data. When I was done exploring the one-dimensional data, I moved onto looking at the two-dimensional relationships within the data. I did this by looking at the different scatterplots and correlations between each of the variables.

Throughout the process of generating the different graphs and visualizations, I completed cleaning activities on the data as needed. For example, some of the numbers were imported as strings rather than either integers or floats. Looking at the initial histograms made these values clear and I was able to get all the data into the correct formats.

The missing values and outliers were treated as is outlined in the preceding section.

Once the data preparation was completed in R, the data was saved to a .csv file and imported into Python. The imported data was explored again in Python to ensure that all data was imported and formatted correctly. The data was placed into a Data Frame in Python.

The last step of the Data Preparation that I completed was performing Principal Component Analysis to reduce the number of features being examined. By using Principal Component Analysis, I was able to reduce the number of variables from 99 to 13.

Modeling

Once the data was cleaned and in the correct format, I started to build my models. I decided to build my models using the K-Means Clustering Algorithm. There were several reasons for this choice. The first reason was that there isn’t a target variable contained within the data set that I wanted to use. I could have used the Fair Market Value variable as my target variable. However, I could not easily get a price range using a continual variable as my target variable without adding in some personal bias. My personal definition of what is affordable may be the same as someone else’s.

The second reason I selected the K-Means Clustering Algorithm is because I wanted to find similar groups contained within the data. My aim is to group new inputs according to what they are similar to and then build an estimate for the housing price range based on the cluster the new data is grouped into.

Once it was decided to use K-Means Clustering for the model, I needed to decide what value of k to use. I used the Elbow Method to evaluate which value of k would be the best. Unfortunately, there was no one single value for k. The curve within the elbow method was a smooth curve without any distinct bends in the graph. I experimented with different values of k when building the model.

To get the values for the estimated housing price range, an individual’s information was entered into the model. The lower limit for the housing price range was set to be the minimum value for the Fair Market Value that was contained within the cluster the individual’s data was clustered into. The upper limit for the housing price range was set to the be maximum value for the Fair Market Value that was contained within the identified cluster. Combining the minimum and maximum values then gave a range of housing prices that incorporated the observations that were the most similar to the individual’s input data.

Evaluation

Once the housing price ranges were estimated and the models were built, it was time to evaluate the models. The first round of models were built using 10 clusters. The developed models successfully identified 10 distinct clusters. However, there was too much overlap between the estimated housing price ranges. Some of the estimated price ranges were encompassed by another cluster’s price range. This was problematic as it didn’t give a unique answer for each cluster. Based solely on the estimated price range, an individual could be associated with several different clusters.

The second round of models were built using a value of 15 for k. Once again, the model identified distinct clusters within the data. When the different housing price ranges were extracted and examined, there was some overlap between the different price ranges, but the different ranges had much better separation than the k = 10 model.

All the developed models that used different values of k contained different amounts of overlap between the different price ranges. I ended up selecting the k = 15 model. The k = 15 model did not have the most distinct price ranges, but price ranges that seemed to be the most natural. What I mean by that is that the k = 15 model did not contain a cluster that had a price range from $0 to $1,500,000 or other price ranges that would seem excessive.

Models that contained a higher value of k restricted some of the price ranges to be non-representative of someone looking to buy a house. In the k = 22 model, one of the cluster’s estimated price range was from $75,000 to $80,000. This price range felt too restrictive to truly be representative of those looking to buy a house.

The k = 15 model struck a good balance between the estimated price ranges being too lenient and the models being too strict.

Out of curiosity, I also spoke with my personal bank’s representatives to see what their estimated housing price range was for me using my personal information. To my amusement, the k = 15 model had a maximum price range amount that was $10,000 higher than my bank’s estimate.

Deployment

The developed model was not designed to be deployed and so it was not designed to be user friendly. The model was designed to explore the American House Survey and draw conclusions based on the data and to address the defined business question. All these activities are incorporated into my Course Project for DSC 630: Predictive Analytics.

If the model was being developed to be deployed for consumers, the final model would have been designed and deployed differently. If the model was to be for consumer use, I would incorporate the model into a smart phone application. The application would ask the user to enter in their information. The user’s input would be passed to the model, which would group the user’s inputs into the nearest cluster, and return the cluster’s minimum and maximum values.

In addition to placing the model into an application, I would make other changes to the model as well. The current model does not incorporate an individual’s FICO score, geographical location, or the current tax rates. Prior to deploying this model to consumers, I would want to build another model that incorporated these values. I would also want to build a model that incorporated multiple types of machine learning to estimate the individual’s price range.

Summary and Conclusions

In conclusion, based on the model, I am unable to afford a house in my current area. Outside of the model, I may be able to afford to buy a house if I either find an excellent deal or move to a neighborhood that I don’t want to move to, but the model did not consider either of these factors.

Acknowledgments

I would like to thank everyone who helped me with this project. First and foremost, Professor Deitenbeck, thank you for all of you instruction and help you gave to us. Second, to all my fellow classmates. I want to thank all of you for all your feedback and suggestions. This project would not be where it is now without your help. Last, to my family. Your understanding and support was invaluable. Giving me the time to focus and complete this project help a ton. Thank you!

References

Office of Policy Development and Research. “American Housing Survey: Housing Affordability Data System.” Retrieved from <https://www.huduser.gov/portal/datasets/hads/hads.html>

Zillow. “Utah Home Prices & Values.” Retrieved from <https://www.zillow.com/ut/home-values/>

Hoffower, H. (18 Aug. 2018) Business Insider. “The most expensive and affordable states to buy a house, ranked.” Retrieved from <https://www.businessinsider.com/cost-to-buy-a-house-in-every-state-ranked-2018-8>

Albon, C. (2018) *O’Reilly Media, Inc.* “Machine learning with Python cookbook.”

Abbott, D. (2014) *John Wiley & Sons, Inc.* “Applied predictive analytics.”

O’Neil, C., Schutt, R. (2014) *O’Reilly Media, Inc.* “Doing Data Science.”