Proceedings of Kaggle Competition Stats 101C Final Assignment, UCLA, December 15th, 2018

**Data Mining on a Housing Database**

Kristee Song, Yoshimi Kina, Statistics Department UCLA, Los Angeles, CA, 90024, USA {[kristee.song@gmail.com](mailto:kristee.song@gmail.com), [yoshimikina@gmail.com](mailto:yoshimikina@gmail.com)}

**Abstract**

This project uses data mining techniques to predict affordability of houses in Ames, Iowa. The study accompanies a classroom curriculum in the course Statistics 101C “Intro to Statistical Models and Data Mining” at University of California, Los Angeles into the uses of machine learning to increase the accuracy of predictive modeling techniques. To this end, the study will use a nominal data set titled “Predicting Affordability of Housing”, and RStudio to code the data models. Various data mining algorithms are used against the Housing Database, such as decision trees, logistic regression and the k-nearest neighbor classifier. Finally, after deriving inadequate accuracy results of <97% on the competition website Kaggle, we suggest to upgrade our machine learning method to random forest in order to achieve our best accuracy rate (to our potential) of 97.066%.

**Introduction**

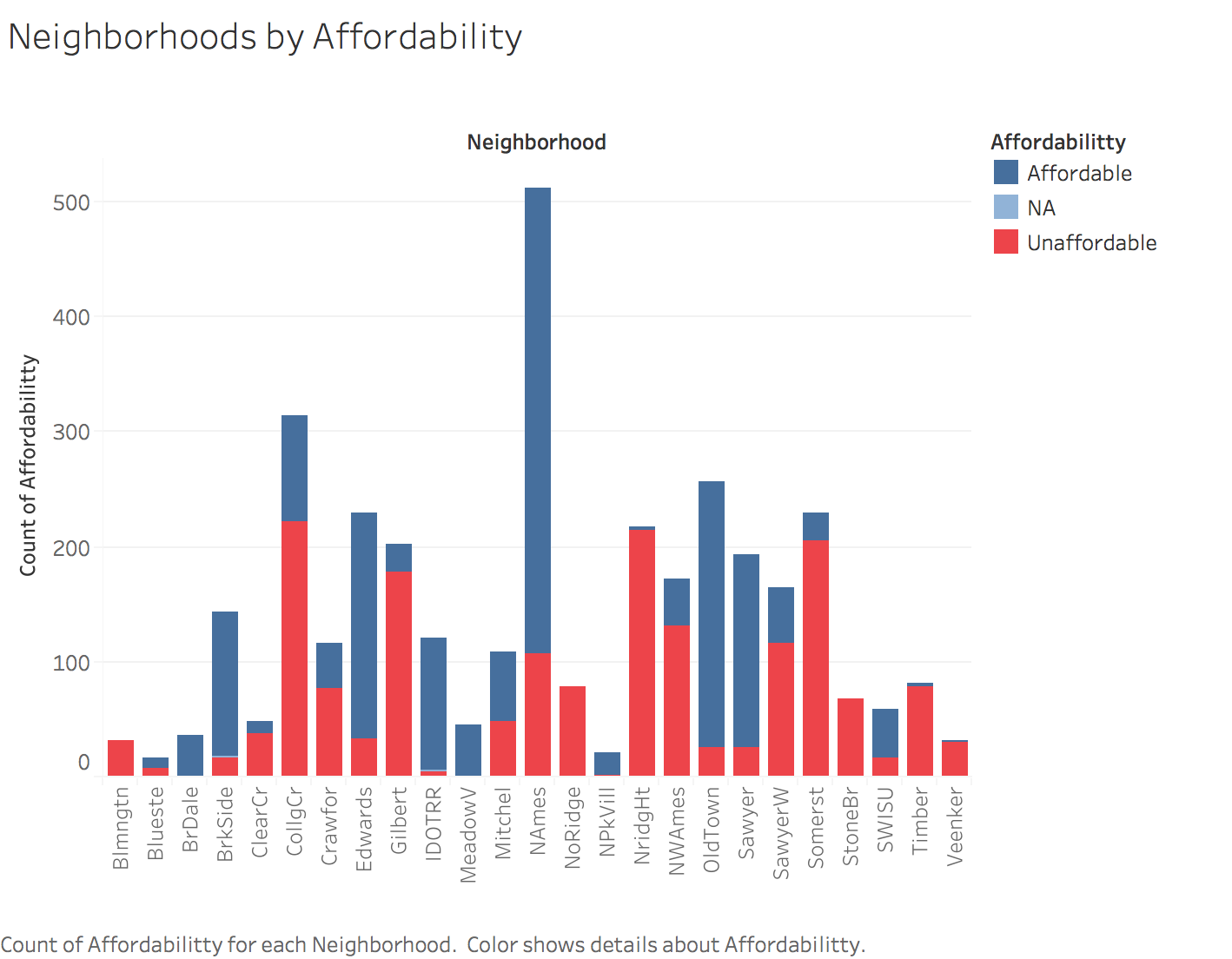
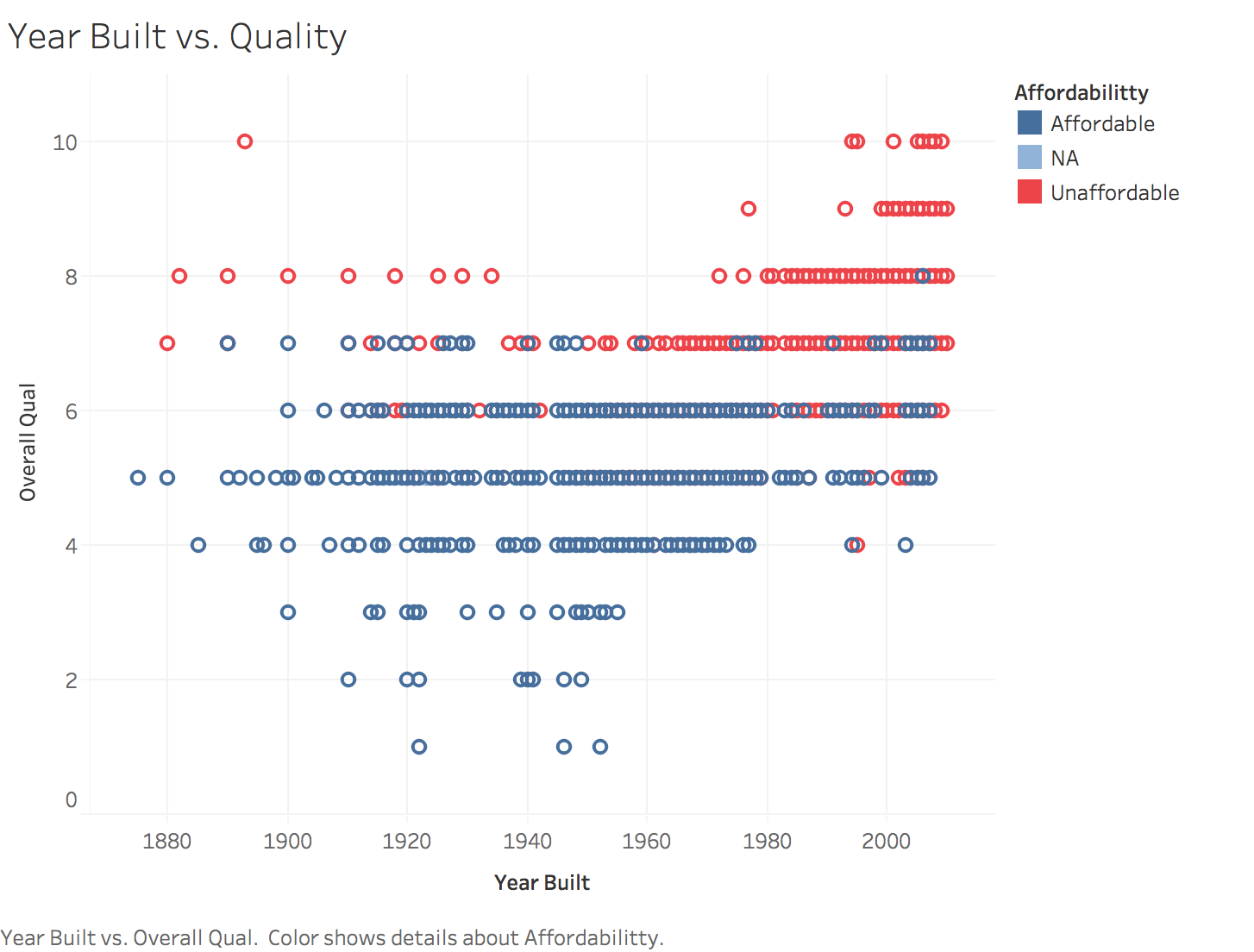
The data set and competition is provided by Dean De Cock for usage specifically in Data Science education. Although the purposes of this data mining project was for an educational conversation in machine learning methods, one may be able to use our findings to, given several attributes about a home in Ames, Iowa, categorize a residential home as “affordable” or “unaffordable”. If modeled accurately, these models may become useful for real estate agents or advertisers in Ames who would like to appeal to buyers looking in a specific price range. However, in our particular analysis, there is no significance to the specific city of Ames, Iowa itself. Our motive for this analysis is to predict housing affordability based on specific attributes, and the applications for this data is not exclusive to the residences in Ames. In this report we will explain our methods of data cleaning. We will then explain our usage of various algorithms, such as decision tree modeling, logistic regression and the k-nearest neighbor classifier, and submit them to Kaggle.com for evaluation in predicting the true affordability of each house. The computed accuracy rate from Kaggle.com is the Misclassification Error rate comparing the predicted affordability to the true affordability categorization of each house. Finally, for the purposes of this report, we will define the methods used and discuss what method is or would have been more effective for predicting the response variable.

**Data Cleaning**

*Per particular residence in Ames, Iowa, we are given 79 variables regarding comprehensive attributes such as Zoning, Lot Area, existence of an alleyway, existence of a masonry veneer… all the way to the year the house was built, and the houses’ sale condition. We clean and redefine variables to gather only comprehensive data and consolidate to reduce collinearity between the variables.*

To begin, we reviewed the entire data set for Predicting Affordability of Housing and found that it had 3500 observations and 81 variables, including affordability. We noticed that many of the variables were either redundant or had too many missing values to be considered in our model. First we removed any variable that had a significant amount of missing values, this included variables such as Fence, PoolQC, Alley and others. These variables had well over 2500 missing values. Later we looked at the categorical variables, if majority of the observations were a member of one category we removed that variable as well. This included variables like LandSlope with 3311 observations corresponding to Gtl. From there, we looked for variables that were highly correlated with one another, for example there are many garage and basement variables that can all tell the same story. We chose one or two from each that best describes the data instead of keeping 7 variables that describe the garage.

After removing redundant variables, we researched houses in Ames, Iowa through the website Trulia. Trulia provided what the most important predictors were or what people look for when buying a house. This includes lot space, number of rooms, parking spaces, and so on. We kept all predictors that we considered important. From 81 predictors, we reduced the list to 7, excluding the response variable affordability. Using tableau, we used descriptive analysis to pick a handful of predictors to use for modelling.



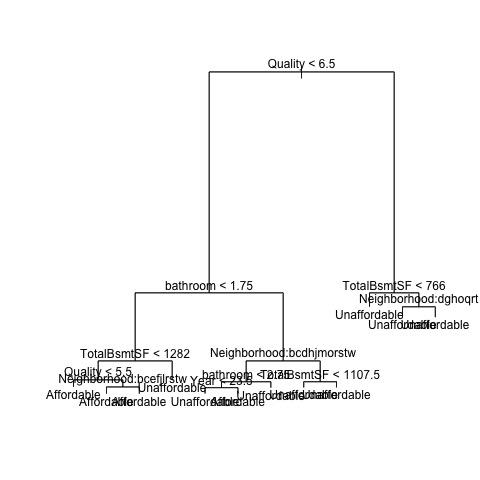
After choosing which predictors we would keep, we decided to adjust some of the variables to better predict affordability. Our final predictors are as follows:

|  |  |
| --- | --- |
| Neighborhood | What neighborhood they belong to |
| Neighbor | Percentage of unaffordable homes in that neighborhood |
| Quality | Overall material and finish quality |
| Year | How old the house is |
| TotalBsmntSF | Total square feet of basement area |
| Bathroom | # of bathrooms + .5(# of half bathrooms) |
| GarageCars | Size of garage in car capacity |

After finding all the predictors we would use, we had to deal with the missing values in the predictors we had chosen. For predictors like GarageCars or TotalBsmntSF, after looking at all other predictors that also described garages and basements, we concluded that NA’s corresponded to homes that did not have a garage or a basement and input 0 for any NA. For other variables like number of rooms and things every house should include, we impute the mean. And for all other NA values we simply omitted them.

**Machine Learning Algorithms**

1. Trees

After cleaning up the data, we decided to run a tree model to see how well it would classify our data. We split the data into 50/50 training/testing set and ran the tree model on the first half. Our resulting tree was as follows. 

We considered pruning but the full model gave us the best misclassification error. Although it was not an optimal classification rate, it was a good start to our project that even told us which variables were very significant.



According to our confidence interval, our misclassification rate for our training data is ~10.41%, and an accuracy rate of ~89%.

1. Logistic Regression

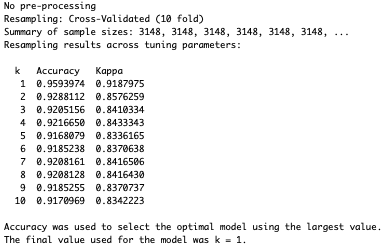
Our second attempt was a logistic regression model. We made a full model using all of the 7 predictors. Out of all the predictors, all but one (Year) were significant in predicting the affordability of a home.



According to our confusion matrix our misclassification rate is ~7.89%, or ~ 92.11% accuracy rate.

1. KNN

The last attempt was a KNN model. We began by extracting the numeric predictors from our model and scaling them for the method to perform more accurately. First using a 70/30 training/testing approach our accuracy raised to about 94-96% compared to ~92% with logistic regression. After running the model with a 10-K cross validation approach, we noticed using K=1, on average a cross validation KNN approach would give us about 96% accuracy rate. Even after modifying the predictors used, the accuracy would not change too drastically, thus 96% was the maximum accuracy we would receive using this technique.



**Results**

According to Kaggle, our results revealed that the methods of Logistic Regression and KNN were not to par with >96% accuracy. According to the testing accuracy rates we calculated through our models, the accuracy rates we scored on kaggle similar to what we expected. Although they are a bit lower, trees received the lowest score followed by logistic regression and KNN as expected. These methods were successful in classifying affordability of homes in Ames Iowa with an exceptional accuracy rate of over 90%, and with KNN leading with ~95%.

|  |  |  |
| --- | --- | --- |
| Method | Public Accuracy | Private Accuracy |
| Trees | 0.92000 | 0.90571 |
| Logistic Regression | 0.93111 | 0.91428 |
| KNN | 0.96666 | 0.94380 |

**Limitations and Opportunities for Improvement**

Although our most effective algorithm of a single nearest neighbor KNN gave us a high accuracy rate of 96.66%, our limitations in achieving a higher score may be from using clustering rather than a random forest technique. Depending on the type of data, there will be better techniques to run. It seems ~96% is the best a KNN model can achieve, but a much higher accuracy can be obtained by trying different methods. Thus, although the competition is closed, we explore using Random Forest with the entire training dataset to product a highest accuracy rate of 97%.



Due to limited time, there was no validation technique used to split the training data into parts to prevent overfitting, but the results still show drastic improvement than the KNN method. If we had submitted to Kaggle in time, our public accuracy rate would have increased by 1%. With more time and in-depth research, an ideal model can be created to accurately classify the affordability of a home by creating several cross validation sets and optimize the random forest algorithm.

Further research about the topic can also help in increasing accuracy rate. Choosing the most significant predictors and creating predictors based the attributes of homes can help depending on if how much they relate to the affordability of the home. Typically, the attributes of a house do a good job depicting how affordable a house will be. The factors that attract a person to buy a house can also change the pricing in houses.