Influential Factors on Affordability

Group Name: Invictus

Members: Hong Lyu, Ziwei Zhou, Jiashu Miao, Langkun Guo

UCLA Stats Department

Professor Almohalwas, Stats 101C: lecture 2

Dec 13, 2018

Abstract

Affordability is the most decisive factor in buyers' decision-making process and imaginably, there has been interests in predicting the affordability of houses accurately. In this paper, we tried to find out the most influential explanatory variables in predicting the affordability among 79 potentially variables and the most effective model by applying different classification methods including Logistic Regression, K-Nearest Neighbors Method, and Random Forest. There were 3,500 houses to be considered as training dataset and 1,500 houses to be considered as testing dataset. The models are trained using the training dataset after data cleaning and used to predicted on the testing dataset. Based on our results, Random Forest is the most accurate model with a 98.51% prediction rate and we intend to have future research about other possible classification models.

I. Introduction

The 2008 economic crisis had a huge impact on many aspects of the lives of Americans. Most of the American families have experienced some negative effects from the economic downturn over the last couple of years. This particular recession has been characterized by the falling of real estate values. The economic crisis showed us that real estate market is indeed susceptible to decline just like other businesses. Even though it has been ten years since the recession, the catastrophic loss in the real estate market has left families wondering what attributes of a property make it more or less desirable in the market and if the properties on the market are affordable.

In order to understand the inherent characteristics of real estate market in the United State and what are the features of houses that would significantly affect the prices, we decided to use the Ames Housing dataset provided by Dean De Coke to analysis the real estate market in the United States. The data set has 79 explanatory variables that describe almost every aspect of residential homes in Ames, Iowa. When people plan to buy houses, they usually consider the number of bedroom and the number of bathroom as the elements that would influence the house prices. However, the data set

shows that there are so many more factors that could contribute to the fluctuation of the negotiation prices. We want to use the variables in the Ames Housing dataset to predict the affordability of each home in Ames, Iowa.

We used different classification methods such as logistic regression, K-Nearest Neighbors method, and Random Forest method and use our model to predict the affordability of the houses and compare to the actual data.

II. Data Exploration

Before we get into more advanced data cleaning, we would like to explore the data first to get a basic understanding of the variables. We first select out the numerical predictors and plot their respective density plots by their affordability.

There are in general two categories in the output plots:

1. If the Affordable Curve (Red Curve) and Unaffordable Curve (Green Curve) do not overlap each other too much, then this predictor is efficient in differentiating the affordability.
2. If the two curves overlap each other mostly, then this predictor is inefficient in differentiating the affordability and should not be included in our classification model.

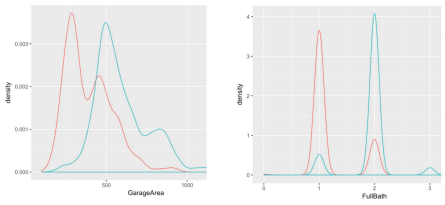

Figure 1: Example of predictors efficient in separating the classes.

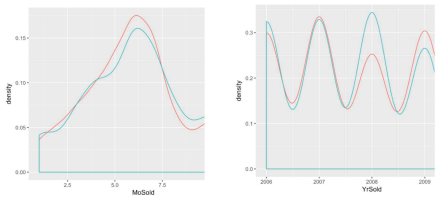

Figure 2: Example of predictors inefficient in classifying

III. Data Cleaning

The original housing dataset has 3500 observations and 81 variables (include the variables observation number and affordability). Based on the first glance of the training data, we see a lot of missing values (NA) and 0s. For example, out of the 3500 observations, the variable “Alley” has 3273 missing values and the variable LowQualFinSF has more than 300 zeros.

According to the practical meaning of the variables and our perspectives on how those variable would affect the housing prices, we decide to delete the variables with more than 1500 missing values, of which are Alley, PoolQC, MiscFeature, Fence, FireplaceQu and the variables with more than 3000 zeros which are LowQualFinSF, 3SsnPorch, ScreenPorch, PoolArea, MiscVal. Since the variables with too many missing and “meaningless”

observations are muted from the data, the accuracy of our model could be effectively increased.

IV. Data Imputation

As for the variables that contain not many NAs or 0s but still some, simply removing the variables is not enough. We did data imputation so that missing values and zeros would be replaced by either specific values or values from certain distributions or algorithms case to case. Usually the process starts with general replacement, with which sample means, modes, medians or random values from the data would be used to replace the missing and meaningless values. This works pretty well when the model is relatively simple and straightforward with relative low requirement for high accuracies. Sometimes when the model becomes advanced and complex, the general replacement method performs surprisingly better due to its light impact and little manipulations to avoid biased results on original dataset. A more statistically idea way to process the missing values is predicting the missing values with certain distribution, for examples, linear regression model which contains

This works well when the we are dealing with continuous variables, but categorical variables are ideally imputed by replacing with most frequency value or "none".

We could use k-Nearest Neighbors approach to impute missing values. This method does the following: for every observation to be imputed, it identifies 'k' closest observations based on the Euclidean distance and computes the weighted average (weighted based on distance) of these 'k' observations. The advantage is that you could impute all the missing values in all variables without considering the variables' types and run the result in one function.

A more advanced and mathematically involved way is using the "mice" package, which is short for Multivariate Imputation by Chained Equations, that create multiple imputations as compared to a single imputation (such as mean) takes care of uncertainty in missing values.

Under this method, we assume that the missing data are Missing at Random (MAR), meaning the probability that a value is missing depends only on observed value and can be predicted using them. It imputes data on a variable by variable basis by specifying an imputation model per variable. The case to case imputation for each the missing value of each variable effectively increases the accuracies in most conditions.

V. Model Selection

There are many different classification methods we can try to predict the affordability of the houses in the dataset. In order to get the highest prediction accuracy, we tried total of five models and compared the results. The models we tried are: K-Nearest Neighbors algorithm, Decision Tree, Logistic Regression and random forest. Out of the 23 trials we had with different models, we summarized the test accuracy results with range in Table 1.

Model	KNN	Tree	Logistic Regression	Random Forest
Test Accuracy	0.9161 +/- 0.0191	0.9343 +/- 0.0296	0.9417 +/- 0.0121	0.9727 +/- 0.0237

Table 1: Classifying models with accuracy estimates via cross validation.

After comparison, Random Forest model is chosen given the highest test accuracy. Cross validation is then conducted to for grid searching over "mtry" and "ntrees".

VI. Results

Grid Searching Results

After we decided to use Random Forest as the final classification method, we performed grid searching over hyper parameters "mtry" and "ntree" to tune the model.

a) mtry

For a fixed ntree = 200, we used tuneRF function. Given "stepFactor" = 2 and "improve" = 1e-5, the best value for mtry is 4, as according to figure 3.

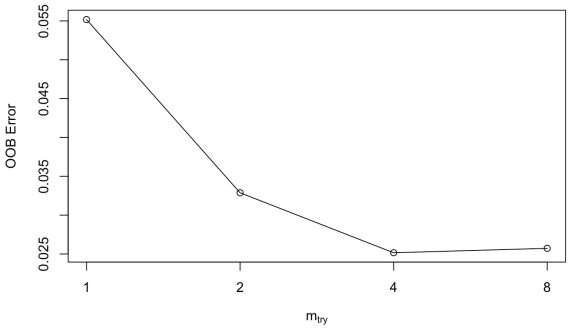

Figure 3: mtry vs OOB Error plot. mtry=4 is selected for our final model according to the most efficient decrease of OOB error.

b) ntree

Given the mtry acquired, we did a manual grid search over ntree using "caret" package, and ended up with best accuracy for ntree = 500.

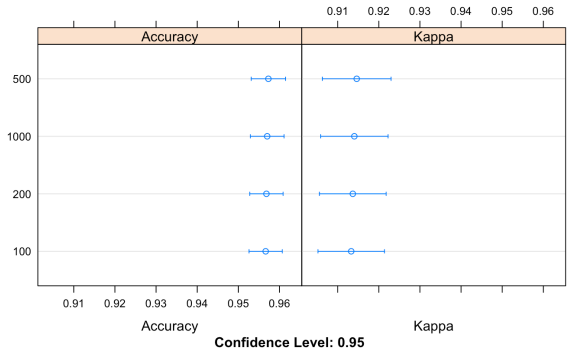

Figure 4: Accuracy of different tree sizes. Given the accuracy, we chose ntree=500 for our final model.

Final Output Method

After acquiring the final model with tuned hyper parameters, we noticed the fluctuation of accuracy is also large for different seeds. So to reduce the fluctuation, the final model is repeatedly trained for 100 times and used to predict the 1500 testing observations each time. For each of the 1500 observations, the majority of the 100 predictions win as the final output for the prediction of the observation.

Accuracy Results

With final model and repeated prediction, our accuracy in predicting the testing data is 0.9813 for private leaderboard, which contains 50% of the data; and 0.9851 over the whole set of data.

VII. Conclusion

Summary

Given the final model trained over the whole training data, we extracted the top 20 most

influential features sorted by Gini and accuracy respectively. The intersect of these two sets of features contains 10 features, namely:

OverallQual, Neighborhood, GrLivArea, Age, FullBath, TotalBsmtSF, X1stFlrSF, GarageArea, X2ndFlrSF, LotArea

We can conclude that these ten factors are the most important factors for the price of a house in Iowa.

Limitations and Potential Improvements

There are certain limits in our modeling process and we can improve it by devoting time into following areas.

i. **Investigation more into the provided features**

   a) We could have tried grouping categorical variables with too many levels, "Neighborhood" by geological relations for instance. Such variables could be grouped into fewer levels according to their reality meanings to improve our model.

   b) Data imputation based on the reality meaning of features

      In our best submission, we used data imputed via MICE. However, it is notable that some of the missing values have their own reality meaning (do not have such feature, for example) so we are overfitting the data in some sense. We could get better results if we impute data according to their reality meaning.

   c) Create more useful predictors like "Age". In our data cleaning and imputation process, we also created a predictor called "Age", which turns out to be a good predictor. More predictors like the total area of the house could definitely help improve the accuracy of model.

ii. Tuning the random forest model

   In our model, we only tuned two hyper parameters "mtry" and "ntree". Hyper parameters like "nodesize" and "maxnodes" were not tuned. We could expect a model with better performance if these hyperparameters are also tuned.

iii. Other models to be tried

   There are other models such as XGBoost that we did not try and combining predictions from other models can also improve the accuracy.