Midterm Report

Wei Zou (wz299) Jiahe Xu (jx266)

October 27, 2017

1 Data Filtration

In order to predict the log-error between their Zestimate and the actual sale price, effective model needs to be generated to learn different features. However, the data set provides 59 features and many of them have no effect on the result of log-error. So our group should choose some most related features to learn and then decide whether the model is underfitting or overfitting.

Firstly, it is necessary to figure out the definition of Zestimate and sale price. Zestimate is the estimation of home values based on 7.5 million statistical and machine learning models that analyze hundreds of data points on each property. And Sale price is the real transaction price of one home.

One point in our project we want to clarify is that we are NOT going to predict the real sale price or estimated price by Zillow. Our data records do not contain any information about sale price. Instead, all we got is the log error calculated from real sale price and estimated price. Logerror = log(sale price) - log(estimated price by Zillow)
In such case, as we mentioned, our group would pay more attention to finding the key factors that determine the prediction.

To find the possible relationship between features and some depend variables of log-error, the class of all features are divided into 4 groups including home, location, market and other. For example, 'total area' and 'Type of home heating' are assigned into 'home' group because they are kind of features to depict the basic parameter of house. 'latitude' and 'longitude' are typical features of 'location' and 'total tax assessed value' is 'market' element. There are also some nonnumerical or meaningless features which are defined as 'other'.

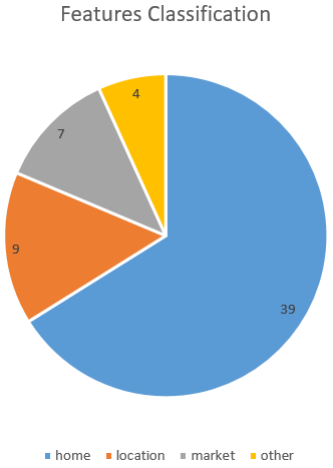

Figure 1. Classification of all features

One of important problems of project's data set is that many features have large amounts of missing values, when may leads lack of training data for specific features. Thus, the number of nulls in all features are counted and the features with more than 30,000 are abandoned. After this step, only 30 features are left.

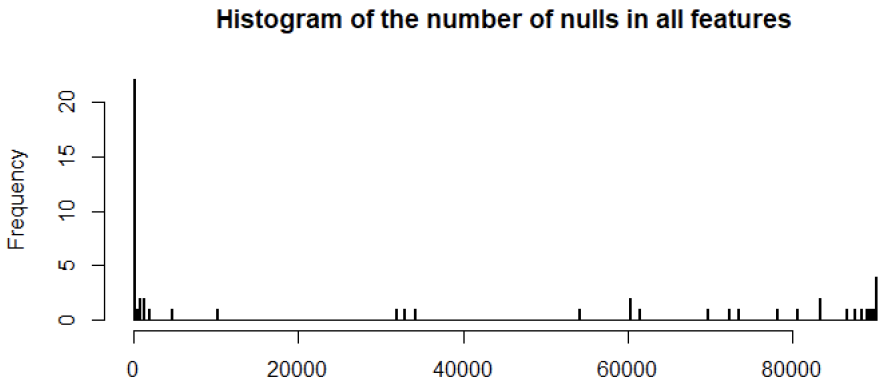

Figure 2. Number of nulls in features

To avoid duplicated and useless data, the whole data set are checked and repeated features(with different name but same data) and factors such as flips (Federal Information Processing Standard code) are deleted. For now, 25 features are remained and our model is planned to learn from them. 2 features belongs to 'other', 7 of 25 is in 'market', 8 features are related to 'location' and the rest are assigned in 'home' group.

2 Dealing with Data Records with Null Value

Row data contain 2 tables, one tables contain all the information of the house, while the other contains the logerror calculated from Zestimate and real sale prices of the house. After merging 2 tables, there are 59 features and the outcome is logerror. We have 90725 row data records indicating all the information of the real estimate.

As for data cleaning, we first plot the histogram of the number of nulls in all features and delete the columns in which the number of nulls is above 30000. Also, we delete some duplicate factors and meaningless factors such as flips (Federal Information Processing Standard code). Because the dataset contains several categorical factors, first we consider using random forest model. Since random forest model cannot handle dataset with null value, we delete all the data records with null values. After that, 74028 data entries are remained.

3 Fitting Models and Preliminary Analysis

Currently, we only use random forest model to fit our dataset, so we do not need to concern that our model will be overfitting when we set the number of trees as a large value (of course it is not too large). Also, we are going to split out our data points into training set (2/3 of all data records) and test data set(1/3 of all data records) to test the performance, specifically the test error, of our model. The updated MSE of our prediction is 0.0257.

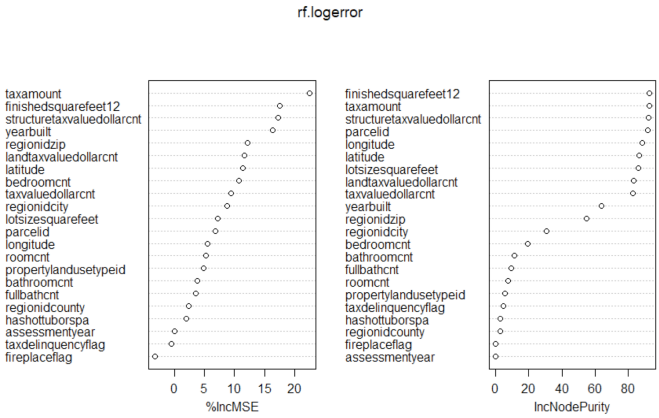

Figure 3. Importance of factors

The figure is about the importance all of the features play in our model. From both of the figures, it can be concluded that these three factors finishedsquarefeet12 (Finished living area), taxamount (The total property tax assessed for that assessment year) and structuretaxvaluedollarcun (The assessed value of the build structure on the parcel) are the most important factors in our prediction.

We take factor finishedsquarefeet12 (Finished living area) as an example to plot out as figure 4. However, when we try to find their correlation, it turn out that the Pearson coefficient is about 0.04160, indicating that there is almost no linear relationship between them. Our group will try to clarity how much importance of this factor and quantize it.

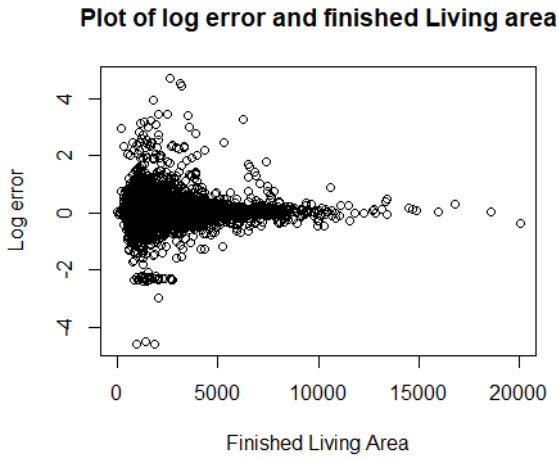

Figure 4. Plot of log error and finished living area

4 Future Plan

In the rest of the semester, we are trying to do regression model (Lasso, ridge and logistic regression) to fit our data and to measure their performance in the test dataset. Moreover, our analysis is a little bit crude and need more adjustment to the parameters in our model to decrease the test error. Also, more deep analysis are needed to do to figure out the key factors that determine our prediction.