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. 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'.

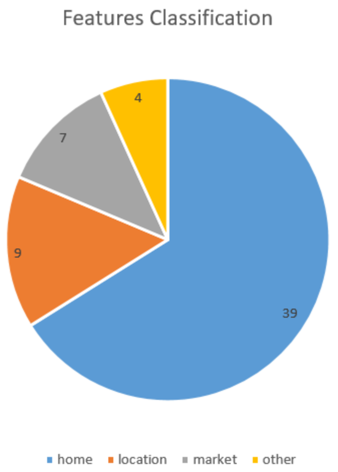

Figure 1. Classification of all features

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.

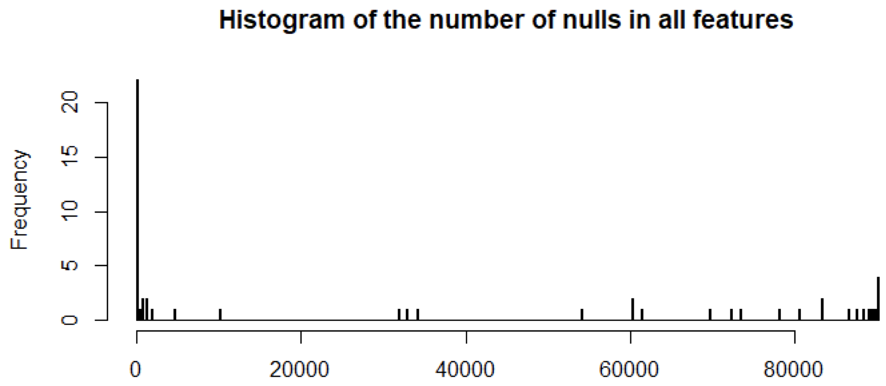

Figure 2. Number of nulls in features

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 Data Analysis and Answers to Questions

Row data contain 2 tables, one tables contain all the information of the house. The other contains the logerror calculated between Zestimate and real sale prices of the house. After merging 2 tables, there are 59 features and the outcome is logerror. In this process, many raws of data are deleted for only considering the data with same house id. After that, 90725 row data records indicating all the information of the real estimate.

Because the data set contains several categorical factors, first we consider using random forest model. Since random forest model cannot handle data set with null value, we delete all the data records with null values. After that, 74028 data entries with 25 features are remained.

Describe how you plan to avoid over (and under-)fitting, and how you will test the effectiveness of the models you develop?
For now, random forest is applied to build our model to fit the data set, so we do not need to concern that our model will be overfitting as the number of trees is set as a large value. Also, we are going to split out 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.

Run a few preliminary analyses on the data, including perhaps some regressions or other supervised models, describing how you chose which features (and transformations) to use.

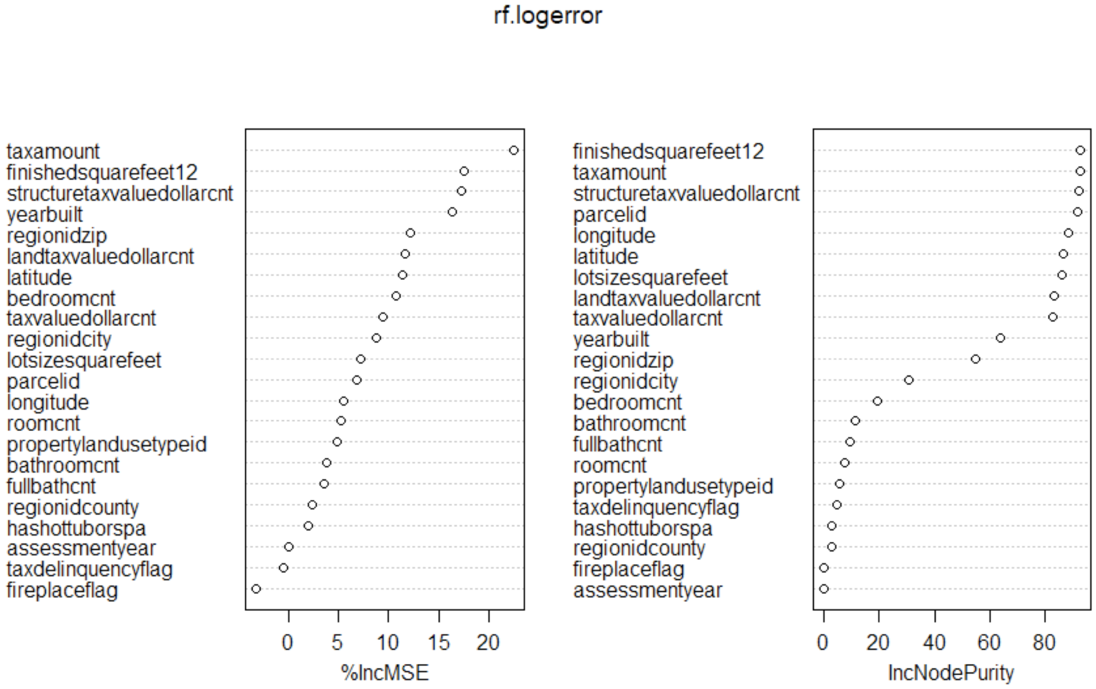

Figure 3. Weight of Features

The figure is about the importance all of the features play in our model. It can be concluded that the finishedsquarefeet12 is the most important factor till now. More researches are needed to do.

Explain what remains to be done, and how you plan to develop the project over the rest of the semester.

In the rest of the semester, we are trying to do regression model to fit our data, and try to adjust the parameters of our model to decrease the test error. Moreover, more deep analysis are needed to do to figure out the key factors that determine our prediction.