CSE 519 -Homework 3 Report

Amol Damare
adamare@cs.stonybrook.edu
Sbu Id : 107914028

Punit Mehta
Punit.Mehta@stonybrook.edu
Sbu Id: 111461860

November 1, 2017

1 Goal

The goal of this homework assignment is to improve upon the models we built in homework 2 for Zillow's kaggle competition. In this assignment we will learn about how to systematically approach the data science problem and employ various statistical tools to improve upon a baseline model. In this assignment we have improved upon our previous models. This report will explain our approach and reasoning behind all the tasks that we did. In first section we will give a brief overview of what tasks we did in homework 2. In next section we will present the our ideas for improvement and reasoning behind these ideas. Lastly we will present our results and evaluation of our models

2 Previous work

In last homework we designed a simple linear model and used data given by Zillow to predict the "logerror". In the assignment we hand picked the features removed any missing values and used this basic cleansed data to fit the model. We used residual graphs to evaluate this model. Table 2 presents the results of this model. This model also scored 0.0674213 on Zillow's competition. As we can see from the score we obtained there is still room to improve this model. In this assignment we will attempt improve this model.

Data Set	Training	Testing
Mean Square Error	0.0285	0.0275

Table 1: Mean square errors for linear regression model in homework 2

Before we discuss the improvement methods we used, let's discuss the issues with our previous model.

2.1 Issues with previous model

In the previous model we did not do any extensive data cleaning. We neither normalized or scaled the data. Only cleaning task we did was to remove missing entries. Other main issue with previous assignment we did was we used features that we thought would work well. There was no statistical reasoning behind this selection. If we could rectify these issues then we think we will be able to improve our model.The next section will present the ideas we used to improve the model and our reasoning behind these ideas.

3 Improvement Tasks

One of the first things we thought of improvement was selection features based some statistical evidence.

3.1 Feature Selection

Zillow data has 57 features for a given property. We want to select the best features among these to consider for our models. One of the ways we can do this is by using feature selection algorithms. There are various selection algorithms that can be used. We can eliminate the features which has very low variance, as low variance means that this feature doesn't convey a lot of information. And feature has similar values for most of the data points. Another way we can eliminate the features is using univariate feature selection algorithms which uses univariate statistical tests such as X^2.
In this experiment, however, we are going to use sklearn's SelectFromModel feature selection algorithm. It uses an estimator and selects the features having weights or coefficients only above a specified threshold. Thresholds can be specified numerically or they can be determined heuristically by sklearn library. We are going to use Lasso regression for selecting the features. But before we could do that, we have to prepare our data to be used for regression

3.2 Data cleaning and preparation

To get good results from feature selection algorithms it is recommended that all variables should be normalized and scaled. Since we are using lasso regression for feature selection we can skip scaling of the data. Following are the steps we used to get clean and prepare the data for feature selection.

1. Encode all the categorical variables. This step is necessary to perform in regression.
2. Impute missing values in the data. In this step we just replaced missing values with mean. (We did try removing them but results were not that good).
3. Normalize the data set.

After these steps we applied feature selection algorithm to get the best possible features we can use for our model. We got 17 features out of 57 as a result of feature selection algorithm. We will now present our experiment and evaluation of our models.

4 Experiment and Evaluation

We used 2 models in this experiment. We wanted to know if we have improved our baseline linear regression model by using feature selection. And we also wanted to test a complex model, so we choose random forest regression. Random forest regression is an ensemble learning method in which multiple decision trees are created at training time and a value is predicted using these decision trees. We think this model will work best with the features we have chosen since the features have very less covariance between them and all of the features are normalized so we all the trees in random forest will have similar weights. We trained these 2 models using the Zillow data we obtained after applying the cleaning steps we performed. In next section we will present the evaluations of these models.

4.1 Evaluation

As expected linear model improved after performing the data cleaning and feature selection. Table **??** shows the training and testing errors for the linear model. This model has r^2 score of 0.04 which indicates that model is good. Figure 1 shows the residual graph of our linear

Data Set	Training	Testing
Mean Square Error	0.0264	0.0238

Table 2: Mean square errors for linear regression model in homework 3

regression model. We also performed permutations test for the linear model. Figure 2 shows the result of permutation test for linear model. As our $pvalue < 0.005$, we can reject the null hypothesis which is that our model is as good as random prediction. So we can say that our linear model is a sound model and does a fairly well job of predicting the logerror. We have submitted this model's result to zillow competition on kaggle. Linear model worked well and improved our previous model on kaggle. We got score of 0.0655138 which was improvement over our previous homework's submission which scored 0.0659056. We will now present the

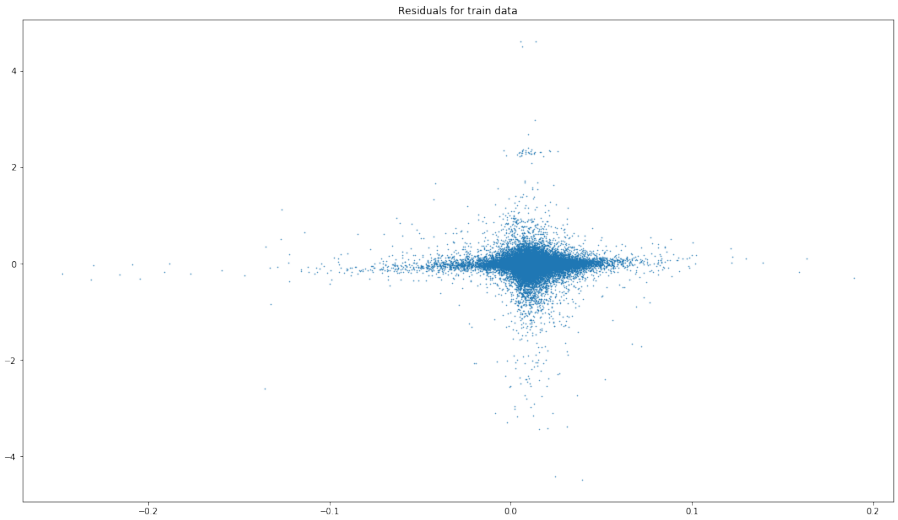

Figure 1: Residual plot for linear regression model

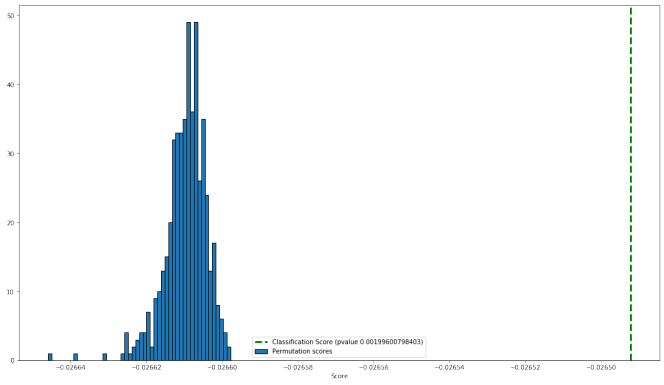

Figure 2: Permutation test for linear regression model

evaluation of our random forest model. Table 3 gives the mean error values of this model. This model gives us similar results to that of linear regression. Figure 3 shows result of permutation test for random forest model. As you can see from permutation tests this model is worse. P-value is 0.67 so half the time it works same as giving a noise output for prediction. It seems from the p-test that the above model might not be best one considering the features we took to solve this problem.

Data Set	Training	Testing
Mean Square Error	0.0265	0.0239

Table 3: Mean square errors for random forest model in homework 3

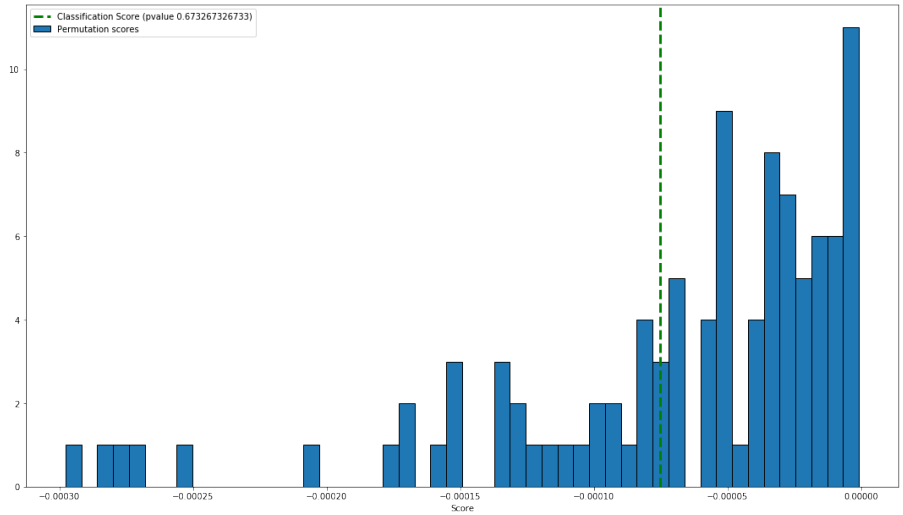

Figure 3: Permutation test for random forest model

5 Conclusion

In this assignment we learned various techniques we can apply to improve our models. And it takes careful evaluation of choices we make in data cleaning that leads to improvement of a model. We also learned about use of permutation test in evaluation of the models

References

[1] Steven Skiena ,*The Data Science Design Manual* ,2017,Springer International Publishing.

[2] Kaggle Zillow Challenge, *https://www.kaggle.com/c/zillow-prize-1*

[3] Linear Regression Wikipedia article, *https://en.wikipedia.org/wiki/Linear_regression*

[4] Sci-kit learn package for Python, *http://scikit-learn.org/stable/*

[5] Pandas package for python,*http://pandas.pydata.org/*

[6] Sci-kit learn SVM tutorial, *http://scikit-learn.org/stable/modules/svm.html*

List of Figures

List of Tables