به نام خدا

گزارش تمرین دوم

مبانی داده کاوی

دکتر ناظرفرد

سجاد اعظمی

۹۲۳۱۰۳۱

زمستان ۹۵

Grade Prediction and House Price Prediction and Comparison Using Three Different Models

Sajad Azami
Amirkabir University of Technology
sajjadaazami@gmail.com

Definitions

- **K-fold Cross-Validation** is a method for estimating risk of a model. Here we divide the data into *k* groups, often *k=10*. We omit one group of data and fit the models to the remaining data. We use the fitted model to predict the data in the group that was omitted. Then, we estimate the risk using our metric of interest, like RMSE in this project. This process is repeated for each of the *k* groups and the resulting risk estimates are averaged.

- **MSE, MAE and RMSE,** are evaluation metrics in order to compare models together.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|$$

$$RMSE = \sqrt{MSE}$$

  MSE corresponds to expected value of quadratic loss, it is always positive and values close to zero are better. The MSE is the second moment of the error, and thus incorporates both the variance of the estimator and its bias. For an unbiased estimator, the MSE is the variance of the estimator and RMSE is the standard deviation. MAE is a scale-dependent accuracy measure and is commonly used in Times Series Analysis.

- **Covariance, Correlation:** Let X and Y be random variables with means μ_x and μ_y and standard deviations σ_x and σ_y. Covariance and Correlation between X and Y is defined by:

$$Cov(X,Y) = E\left[(X - \mu_x)(Y - \mu_y)\right]$$

$$\rho_{x,y} = \frac{Cov(X,Y)}{\sigma_x \sigma_y}$$

  Covariance measures joint variability of two random variables. Mathematical concepts of correlation and covariance are very similar; both describe the degree to which two random variables or sets of random variables tend to deviate from their expected values in similar ways. Correlation is always between -1 and +1, so it is dimensionless while covariance is in units obtained by multiplying the units of the two variables.

- **Regression toward the Mean:** Regression to the mean occurs when the second measurements of a particular variable are less extreme than the first. It was derived from a biology paper by Sir Francis Galton who noticed that tall and short men tend to have sons with heights closer to the mean. The term "regression" is due to this work.

- **L1 norm, Lasso Regression:** A norm is a function that assigns a strictly positive length or size to each vector in a vector space, save for the zero vector, which is assigned a length of zero. One of the famous norm families are *p-norms.* Let $p \geq 1$ be a real number. We define *p-norm* by:

$$\left|\left|x\right|\right|_p = (\sum_{i=1}^{n} |x_i|^p)^{1/p}$$

  L1 norm or least absolute deviations (LAD), basically minimizes the sum of the absolute differences between the target value and the estimated values. So if we assign *p=1* we will get L1 norm. Lasso regression uses L1 norm as its cost function. Lasso reaches sparse and robust outputs, so this method finds application in many areas compared to method of least squares.

- **L2 norm, Ridge Regression:** L2 norm or least square deviations (LSD), minimizes the sum of the squared differences between the target value and the estimated values. We assign *p=2* to get L2 norm. Ridge regression uses L2 norm as its cost function. Respectively, Ridge reaches non-sparse or smooth outputs.

Part One, Grade Prediction

In this project, we are going to use a dataset of student grades.

Dataset Description

Table includes grades for seven courses of 240 students. We are going to learn various models for predicting the 7th grade using first six. Therefore, our train set will be a *240*6* matrix and target value will be 7th grade. We will divide this data set to *200*6* matrix for train set and *40*6* matrix for test set. We will use 10-fold cross validation using *RMSE* as evaluation metric for model comparison.

Scatter Plot:

In this section, target value is shown versus each feature. We will discuss correlations intuitively using scatter plots. Fig.1 shows scatter plot of data before imputation. Table.1 approximately describes correlation, range and amount of outliers in data.

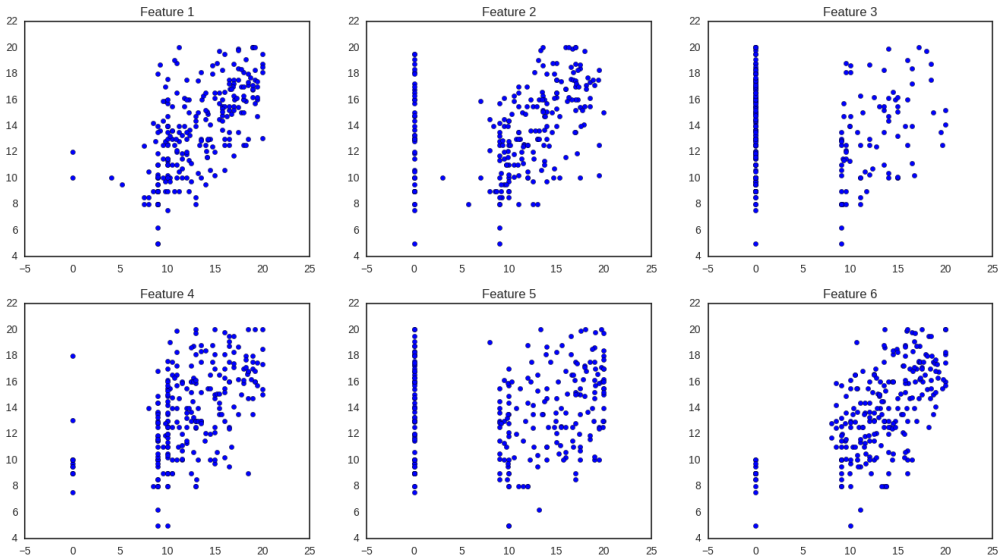

Figure 1 - Scatter Plot before Imputation

	Correlation Abs	Correlation Sign	Outlier
Feature 1	Strong	+	Moderate
Feature 2	Moderate	+	High
Feature 3	Very Weak	+	Very High
Feature 4	Moderate	+	Moderate
Feature 5	Very Weak	+	Very High
Feature 6	Very Strong	+	Very Low

Table 1 - Data Description

As we are dealing with a dataset of grades, we expect correlations to be positive, and we can see that they are.

Imputation

They are many methods to fill missing values like using mean or median or just simply drop them. Considering that they are many NA values, dropping option will not be a good choice. Using median and mean can be good, but we have better options since we know the data well. We are going to use Gaussian Noise. First, we compute target student's average and target course's average. Missing values are going to be imputed by:

$$new\ value = 0.5 * Normal(mean\ of\ row, 1) + 0.5 * Normal(mean\ of\ column, 1)$$

Resulting scatter plot is shown in Fig.2.

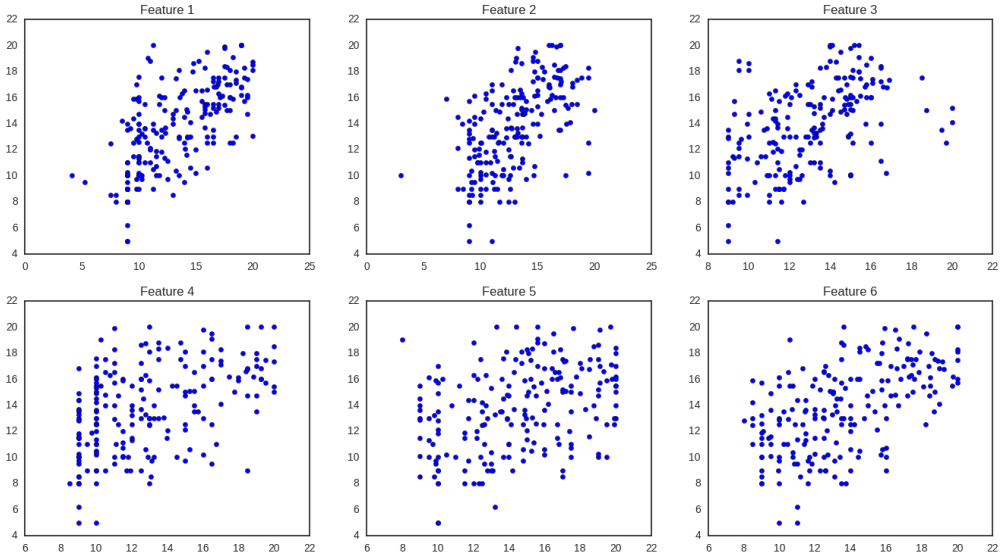

Figure 2 - Scatter Plot after Imputation

Models

Linear Regression

All models implemented use all features, in order to compare only model performance, therefore feature engineering will be done after.

First implemented model is Linear Regression. 10-fold cross validation risk with RMSE metric is used to estimate the risk of models in all 4 models implemented. Fitted line is plotted versus real values in Fig.3. Evaluation results is reported below.

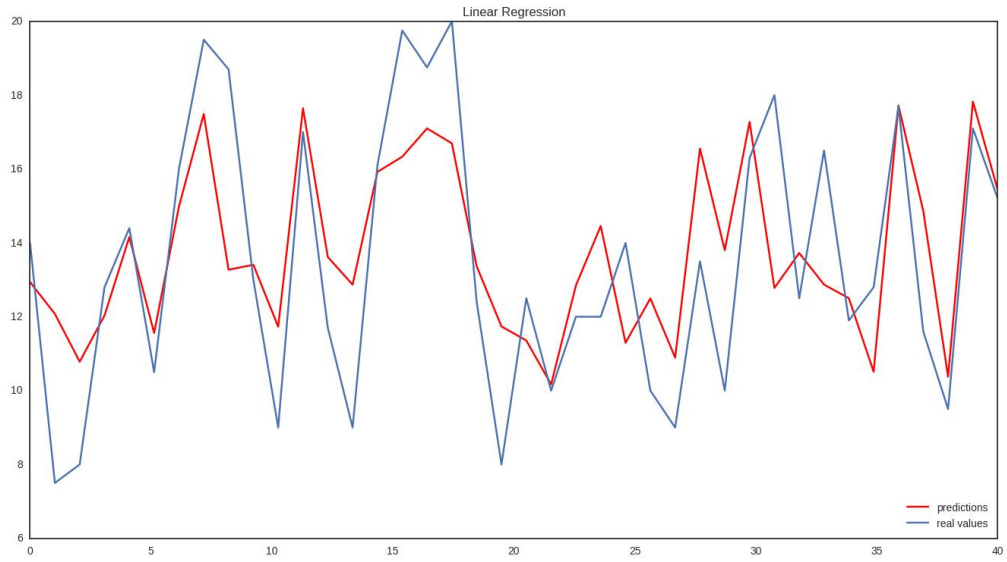

Figure 3 - Linear Regression Line

```
Linear Regression
Test RSS: 238.994624595
10-fold CV with RMSE: 2.452154386928338
```

Lasso

Now we are going to fit a line using Lasso. First we take $\lambda=0.1$, fitted line is plotted versus real values in Fig.4. Evaluation results is reported below.

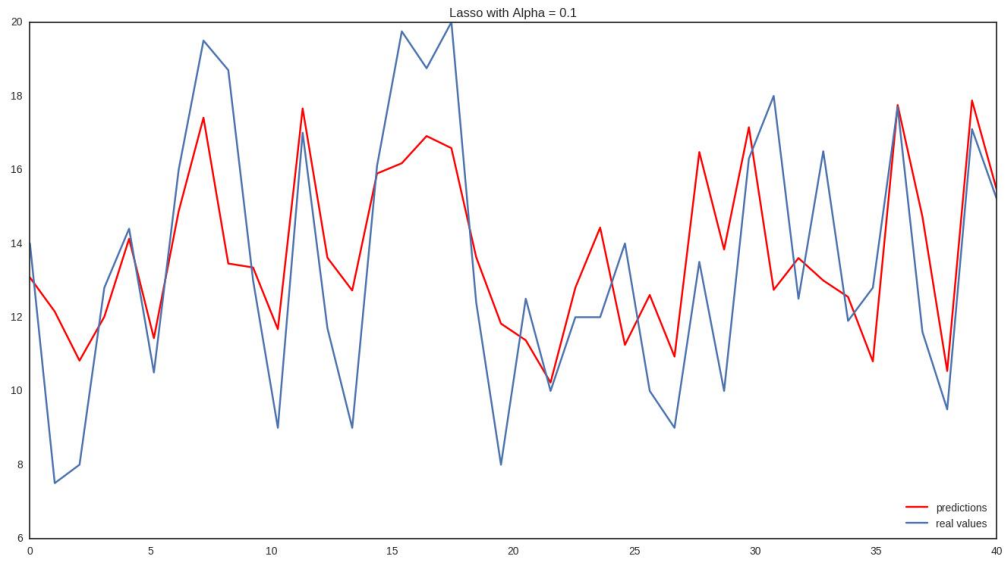

Figure 4 - Lasso with Lambda 0.1

```
Lasso with Lambda 0.1
Test RSS: 239.379645289
10-fold CV with RMSE: 2.4397165003943537
```

Now we are going to test Lasso with different λ. We are expecting to get simpler lines as we increase λ, since it acts as a multiplier for cost function's regularization objective. As shown in fig.5, increasing λ simplifies model and increases error, concretely, if we take lambda so small, it will act like previous model, Linear Regression without regularization.

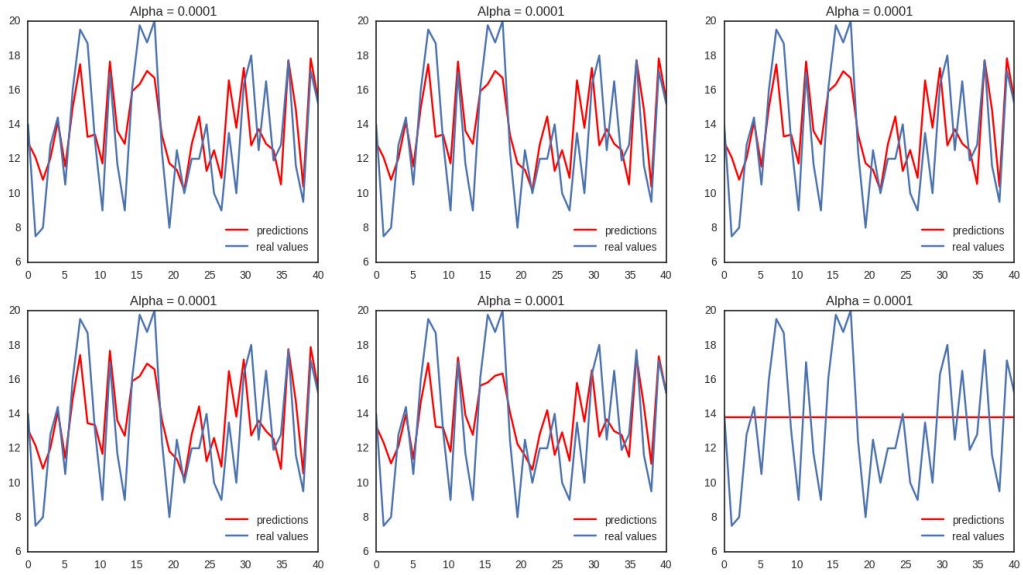

Figure 5 - Lasso with different lambdas

Gradient Boosting

Next, a Gradient Boosting model is implemented. Like previous models, 10-fold cross validation risk with RMSE metric is used to estimate the risk of models and results are reported below. Fitted line versus real values is plotted in Fig.6.

```
Gradient Boosting
Test RSS: 380.095152136
10-fold CV with RMSE: 2.472729431538572
```

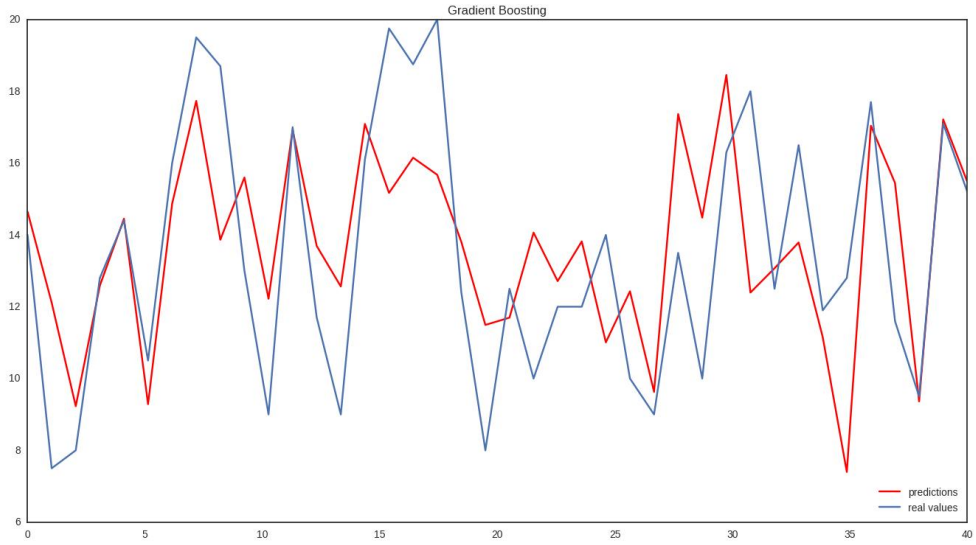

Figure 6 - Gradient Boosting model

SVR with RBF

Finally, a Support Vector Regression model is fitted on data using RBF kernel, *gamma* value of 0.1. Fitted line and results are shown below.

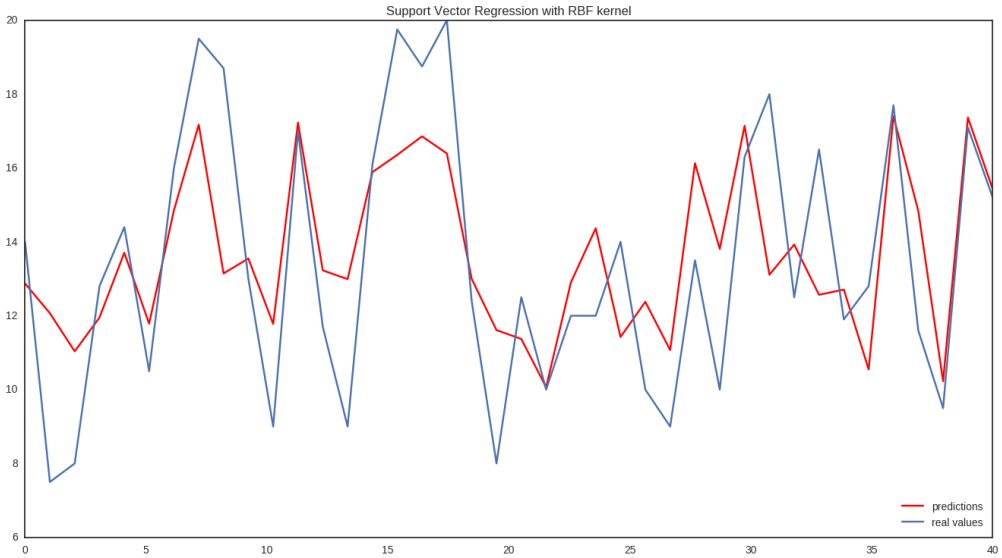

Figure 7 - SVR with RBF

```
Support Vector Regression with RBF kernel
Test RSS: 241.936574933
10-fold CV with RMSE: 2.4887298207357578
```

Conclusion

As described before, Lasso has found many application in real life scenarios, due to its good performance. Cross validations risks shows us the same thing in this problem. So we are going to use Lasso with lambda value of 0.1 to predict unlabeled data, results are attached to this report named under prediction.csv.

Part Two, House Price Prediction

In this project, we are going to use house prices of Ames, Iowa dataset provided by Kaggle.

Dataset Description

This dataset, includes 79 features and about 1500 record of data. We are going to perform previous section's actions on this dataset and check results using Kaggle's Leaderboard.

Preprocessing

First, we are going to impute data. Since there are not so many missing values in this dataset, and knowing that it has large number of features, we are not going to take much time on this part. There are many possible ideas to perform imputation; we are going to keep it simple by using median.
Second, we should encode categorical data to numeric types. This process can be done using encode_field function.

Models

Linear Regression

Using simple linear regression without any feature selection, performs poor since we have 79 features containing 53 categorical features. You can see prediction result in fig.8 and output of program below.

```
Linear Regression
Test RSS: 703577517323.0
10-fold CV with RMSE: 31168.61932524135
```

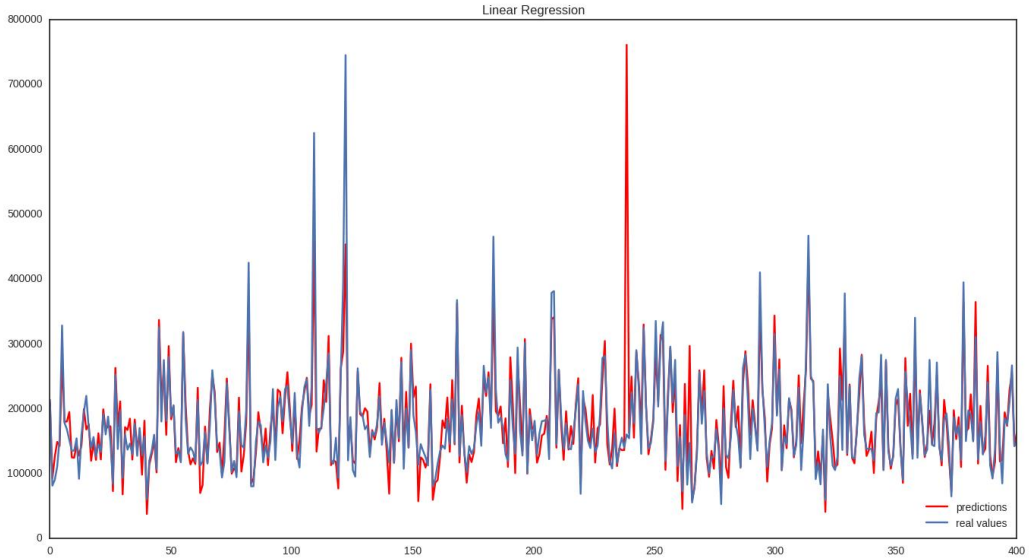

Figure 8 - Linear Regression

Random Forest

Random forest model is trained with parameters like previous section. It has max depth of 40 and 250 estimators with entropy as criterion. We expect random forest to perform well on this kind of data, but to keep fairness among models; we will not train it with complex parameters and many estimators. The result is like below.

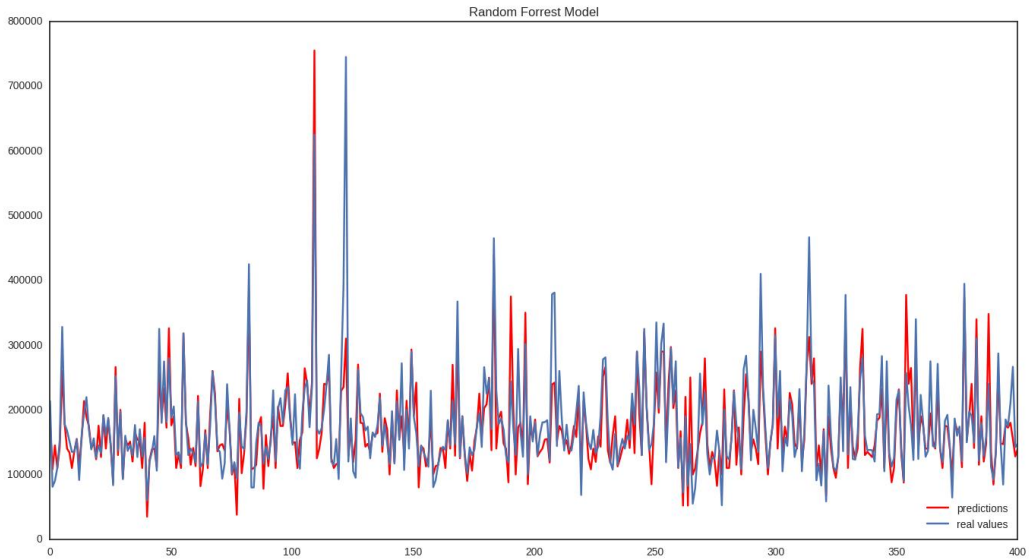

Figure 9 - Random Forest

```
Random Forrest
Best score: 0.016981132075471698
Test RSS: 477969802194
```

Lasso

Lasso model performs well on this data. We have categorical data converted to numerical types in preprocessing part, knowing this we expect to have poor performance using regression models, but lasso performs better than other methods in this family. Results are like below.

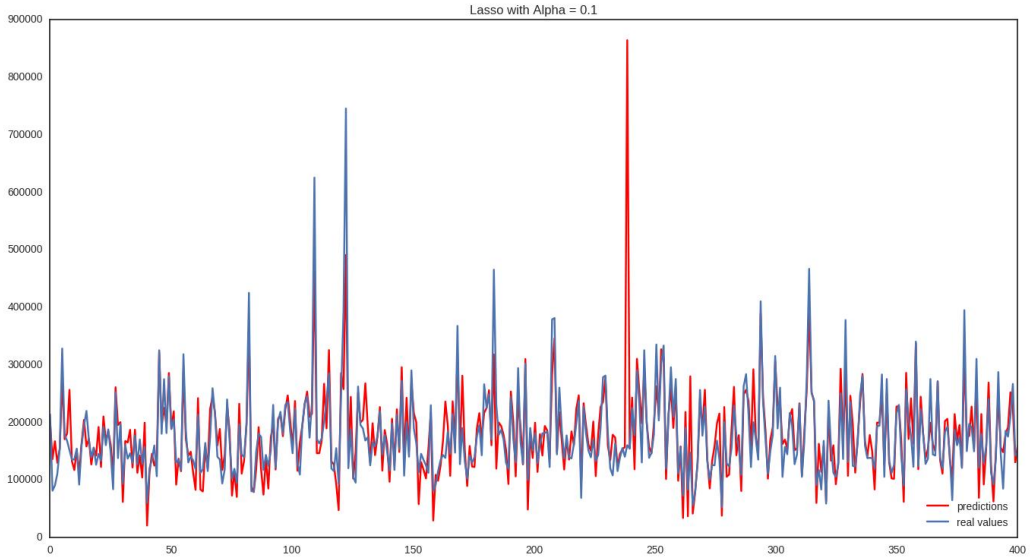

Figure 10 - Lasso Model

```
Lasso with Lambda 0.1
Test RSS: 703583632759.0
10-fold CV with RMSE: 31168.285192335647
```

Conclusion

According to results in previous part, we used Lasso with regularization for final prediction. This lead to score of 0.16 which is not good enough. Using subsets of features were tested using LassoCV method and it reduced features from 79 to 13, but it increased error, so for final submission we didn't use feature selection by this method. A deep analysis of features would help to choose good features and decrease error.