2020

Using data mining to identify the most influential factors in training results

Xiaoqing Wu
CUNY New York City College of Technology

Daanial Ahmad
CUNY New York City College of Technology

Using Data Mining to Identify the Most Influential Factors in Training Results

Xiaoqing Wu, Daanial Ahmad, Dr. Nan Li, Dr. Lin Zhou

Department of Mathematics, New York City College of Technology

Abstract

Data Science is used as a tool to find hidden facts in the data. We want to find out what factors such as 'AGE', 'TAX', 'PUPIL-TEACHER RATIO', 'PER-CAPITA INCOME' contribute the most to housing prices. To answer this question, we studied the dataset of "Boston Houses Prices". By applying the Lasso Regression (a Data Mining Technique) on the data set of "Boston Houses Prices" we identified the influential factors in the linear model. As a conclusion we found that there were six inputs which contributed the most to the prices of houses and those inputs are as follow: (i) CRIM-per capita crime rate by town, (ii) ZN- proportion of residential land zoned for lots over 25000 sq. Ft, (iii) CHAS-Charles River Dummy Variable, (iv) RM- Average number of rooms per dwelling, (v) Black- proportion of black by town, and (vi) LSTAT-Lower status of population

Lasso Regression

- According to the basic linear regression/equation the line of best fit is Y=XB+C, where Y being a dependent variable is the output, X being the variable corresponds to input, B is the coefficient and constant C represents the Y-intercept which is a constant.
- Things in real life are not as simple as they seem to be. When we have tons and tons of data with multiple inputs (i.e. X's) than there are some inputs which have no relation to the output. But if we include those inputs in our model it will make noise in our model and make our model inaccurate.
- It is meaningful to find out which input is more influential to the output. Our project was to find out the most influential input in the '*Boston houses prices*' dataset.
- Lasso regression is a type of linear regression that uses shrinkage. Shrinkage is where it ignores all the least influential inputs which don't contribute to our output thus making our model simple and more accurate.

Methodology

- This research study used the Boston housing data whose origin is natural that is the data is original it has not been simulated.
- The Boston housing dataset contains 506 observations.
- In the data there are 14 different attributes, but our desired output is "*MEDV-Median Value of owner-occupied homes"* with other 13 attributes as input.
- The model is Y=XB+C, where X is a row vector with 13 entries and the column vector B is the corresponding coefficient vector. And constant C is the Y-intercept.
- Linear Regression and Lasso Regression have the same model, but their error functions are different.
- The error function of linear regression is as follow:

$$\sum_{i=1}^{n}(y_i - \sum_j x_{ij}\beta_j + C)^2$$

- The error function of lasso regression is as follow:

$$\sum_{i=1}^{n}(y_i - \sum_j x_{ij}\beta_j + C)^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

- In the above two equations n represents the number of observations in the data set.
- But for the second equation p represents the number of coefficients that is 13 for our dataset.
- But notice there is one new term that is lambda. It is this lambda which plays an important role in reducing the dimension of our model and holds great significance in selecting the right model with greater precision of accuracy. The right model is the one which has the most important input variable, and which primarily effects our output variable. **Now let us explain the power of lambda and how its value affect the linear model.**
- As we increase the value of lambda, we see that more and more coefficients become zero and our eliminated from the equation Y=XB + C and only the strong ones are left in the model which contributes the most.

a) For lambda=0 the model is the same as the linear regression model. In that case none of the coefficients (i.e. inputs Xi's) become zero.
b) For large lambda that is lambda= infinity all the coefficients (i.e. inputs Xi's) become zero.

- In this project, we used the R computer programing language to conduct all the computations and statistical analysis. And, we conduct the data analysis as follows:

1) we divide our data into 50% of training data and 50% of testing data. For technical purpose we need to normalize the data.
2) We used the Lars program to run the LASSO Regression model in the R programing language, with fixed lambda (λ) values ranging from -3 to 1.5 on the horizontal axis of graph 1. Then, we used the coef function, a build-in function in the R language, with fixed lambda to compute all the β coefficients (for crim, zn, indus, chas, nox, rm, age, dis, rad, tax, ptratio, black, lstat variable) from the training and testing data.
3) We calculated predicted values from multiplying all 13 column vectors from training and testing data by individual β coefficients. Then, we computed training errors using the Root-Mean-Square Errors (RMSE) function with inputs of predicted values and medv (dependent variables). Similar to the calculation for training errors, the testing error was calculated with the use of the RMSE method.
4) We draw the diagram to illustrate the relation. In this diagram(which is figure 1), the x axis is lambda, the y axis is the error, the red point is the test error and green point is the training error.

Graph and Table

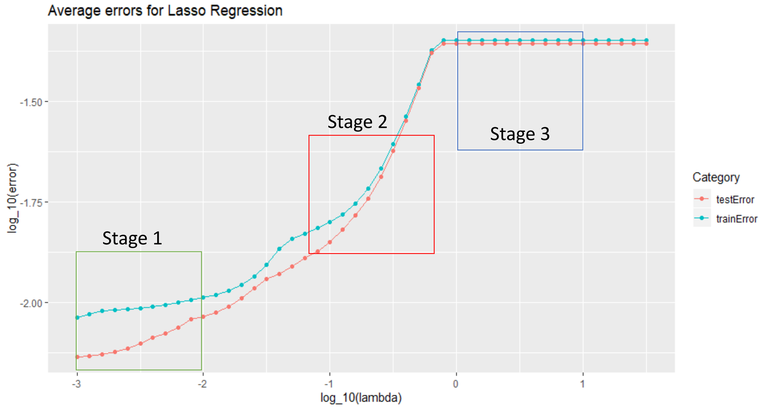

Figure 1: Testing and Training Errors

Log10(λ)	Constant C	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat
-3	0.019234173	-0.04	0.0424	-0.036	0.0344	-0.149	1.4012	0	-0.188	0.0741	-0.114	-0.393	0.126	-0.275
-2.9	0.007129949	-0.039	0.0444	-0.041	0.0345	-0.06	1.4529	0	-0.169	0.0487	-0.09	-0.279	0.1301	-0.275
-2.8	0	-0.038	0.0458	-0.047	0.0347	0	1.462	0	-0.157	0.0226	-0.062	-0.191	0.1253	-0.279
-2.7	0	-0.036	0.0461	-0.05	0.0355	0	1.4228	0	-0.157	0	-0.029	-0.141	0.1103	-0.287
-2.6	0	-0.036	0.0471	-0.047	0.0362	0	1.3808	0	-0.153	0	-0.034	-0.081	0.0962	-0.295
-2.5	0	-0.037	0.0482	-0.043	0.037	0	1.3279	0	-0.147	0	-0.04	-0.005	0.0783	-0.305
-2.4	0	-0.037	0.0442	-0.034	0.037	0	1.2996	0	-0.124	0	-0.032	0	0.0682	-0.307
-2.3	0	-0.037	0.0389	-0.021	0.0368	0	1.2664	0	-0.094	0	-0.02	0	0.0563	-0.308
-2.2	0	-0.038	0.0321	-0.005	0.0366	0	1.2247	0	-0.056	0	-0.006	0	0.0413	-0.31
-2.1	0	-0.036	0.0221	0	0.0369	0	1.1907	0	-0.016	0	0	0	0.0232	-0.3
-2	0	-0.035	0.019	0	0.0358	0	1.1678	0	0	0	0	0	0.0159	-0.282
-1.9	0	-0.035	0.0211	0	0.0333	0	1.1456	0	0	0	0	0	0.0159	-0.26
-1.8	0	-0.034	0.0237	0	0.0302	0	1.1175	0	0	0	0	0	0.0159	-0.232
-1.7	0	-0.034	0.027	0	0.0263	0	1.0822	0	0	0	0	0	0.0159	-0.196
-1.6	0	-0.034	0.0312	0	0.0213	0	1.0377	0	0	0	0	0	0.0159	-0.152
-1.5	0	-0.033	0.0364	0	0.0151	0	0.9817	0	0	0	0	0	0.0159	-0.096
-1.4	0	-0.033	0.043	0	0.0072	0	0.9113	0	0	0	0	0	0.0159	-0.025
-1.3	0	-0.019	0.0404	0	0	0	0.8647	0	0	0	0	0	0.0255	0
-1.2	0	0	0.0315	0	0	0	0.8342	0	0	0	0	0	0.0348	0
-1.1	0	0	0.0153	0	0	0	0.8259	0	0	0	0	0	0.027	0
-1	0	0	0	0	0	0	0.8143	0	0	0	0	0	0.016	0
-0.9	0	0	0	0	0	0	0.7932	0	0	0	0	0	0	0
-0.8	0	0	0	0	0	0	0.7473	0	0	0	0	0	0	0
-0.7	0	0	0	0	0	0	0.6895	0	0	0	0	0	0	0
-0.6	0	0	0	0	0	0	0.6168	0	0	0	0	0	0	0
-0.5	0	0	0	0	0	0	0.5252	0	0	0	0	0	0	0
-0.4	0	0	0	0	0	0	0.41	0	0	0	0	0	0	0
-0.3	0	0	0	0	0	0	0.2649	0	0	0	0	0	0	0
-0.2	0	0	0	0	0	0	0.0822	0	0	0	0	0	0	0
-0.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 1: β Coefficient

Results

- We obtain our result based on the observation and analysis of the above diagram.

1) From figure 1, it shows three stages of progression from the average errors for Lasso Regression model. The lambda (λ) values of stage 1 ranges from -3 to -2, from -1.25 to -0.25 for stage 2 and from 0 to 1 for stage 3. From stage 1, we observed slow beginning for both test error and training error with increasing lambda values. As to stage 2, both test error and training error steeply progresses and approaches saturation as lambda values increases. However, both test error and training error reach plateau with increasing lambda values in stage 3.
2) Additionally, from stage 1 of figure 1, the gap between testing error and training error is wider than in other stages. Thus, it reflects that the regression model associated with stage 1 is not accurate or have less relevant factors and more noises. This phenomenon is known as overfitting. From table 1, when Log_10(λ)< -2 (within the range of stage 1), we observed seven variable (including indus, nox, age, dis, rad, tax, and ptratio) their become zero. It reflects and confirms our previous findings on regression model of stage 1, which means the model is not accurate and applicable.
3) From stage 2 of figure 1, the gap between the two errors is getting narrower comparing to stage 1. From table 1, as the values of Log_10(λ) increases, we observed that some coefficients values become zero, which indicates that those zero-coefficients are weakly-related coefficients and have small or no impact to the medv values (dependent variables). Therefore, those coefficients will be dropped in this process. However, the none-zero coefficients are more related or correlated to the medv values. For example, when Log_10(λ) within the range of [-1.9, -0.9] in table 1, there are 6 variable (including crim, zn, chas, rm, black, and lstat) that are none-zero, which shows that they are the most influential factors to the medv values. Furthermore, it supports that the regression model of stage 2 is a good and accurate model.
4) From stage 3 in figure 1, we observed the testing error and training error is approximately overlapped with each other, and the error are both very large. it is much larger than the error in stage 2 and 3 from figure 1. From table 1, as the values of Log_10(λ) increases, we observed that more coefficients values become zero comparing to stage 2. Besides, when Log_10(λ) larger and larger which indicates we will have higher error. Thus, the regression of stage 3 is not a good model.

Conclusions

- Overall, for this project, we established a statistical framework to explore the most significant factors that affects housing price. By choosing the Log_10(λ) =-1.8 in table 1, we successfully identified six coefficients (including crim, zn, chas, rm, black, and lstat) are the most significant factor to the medv value.

References

- The lecture note from Dr. Nan Li.

Acknowledgements

- The authors would like to thank City Tech's Emerging Scholars program, under the leadership of Prof. Hamidreza Norouzi, for supporting this research.