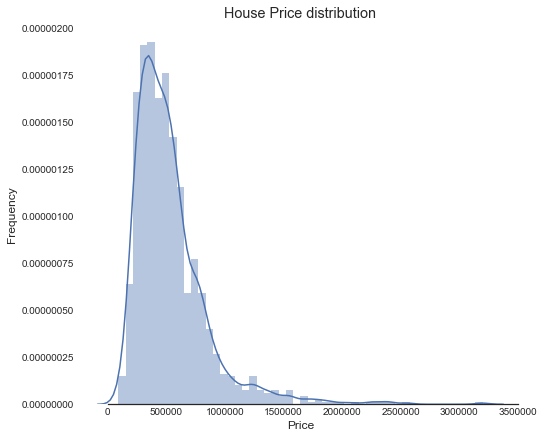
II. Regularization Data Preparation

This dataset is part of the data in real estate market from May to July 2014. The data frame has 1085 observations of house price on 18 variables, including at street names, number of bedrooms, numbers of bathrooms, price and so on. We will use 10 of them to evaluate the model.

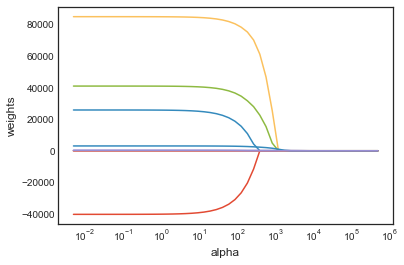
|  |  |
| --- | --- |
| **Variable** | **Definition** |
| date | The date when the house was sold |
| price | The property's sale price in dollars |
| bedrooms | Numbers of bedrooms in the house |
| bathrooms | Numbers of bathrooms in the house |
| sqft\_living | Living size in square feet |
| sqft\_lot | Lot size in square feet |
| floors | Numbers of floors in the house |
| condition | The living condition of the house |
| sqft\_above | Lot size in square feet above basement level |
| sqft\_basement | Basement size in square feet |
| yr | The house ages |
| renovated yr | The house age since last renovation |

The dependent variable is the price of house and the distribution plot is as follows. It could be found that the distribution is right-skewed with extreme large values. Next, the data was separated into training sample and testing sample by 70% and 30% in order to evaluate the overfitting problem and also the bias-variance tradeoff.



2.1. Lasso

Firstly, in order to choose the optimal scale of the regularization, we draw the plot of the relationship of coefficients weights and alpha. It could be found that when alpha is larger than 1000, the penalty is too large, and coefficients would lose efficacy.

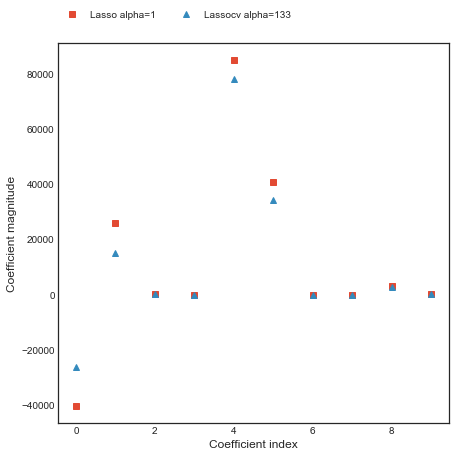


Then, use the general Lasso method with penalty equals 1. The training score is about 0.59 and testing score is about 0.52, and mean square error is about 53089250448.1.

In order to evaluate the performance of Lasso model, we try the cross validation and the optimal penalty is about 133. Then training score is about 0.59 and testing score is about 0.51, which the difference is increased by CV method. The MSE also increases slightly to 53089339669.1. It's worth mentioning that the model includes 9 variables, instead of 10 variables as before. Because the scores are similar, there’s little overfitting problem.

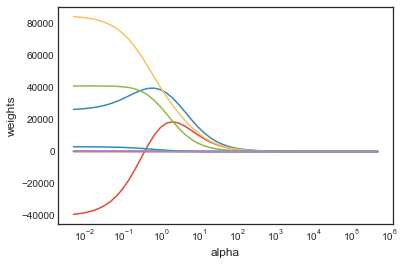
|  |  |  |  |
| --- | --- | --- | --- |
|  | **training accuracy** | **testing accuracy** | **MSE** |
| Lasso | 0.589820589 | 0.517274748 | 53089250448 |
| LassoCV | 0.58783877 | 0.513972497 | 53089339669 |

From the plot below, we could find the differences between two models with different penalties and independent variables, like Lasso include the sqft\_basement information into the model and Lasso with cross validation exclude them.



2.2. Ridge

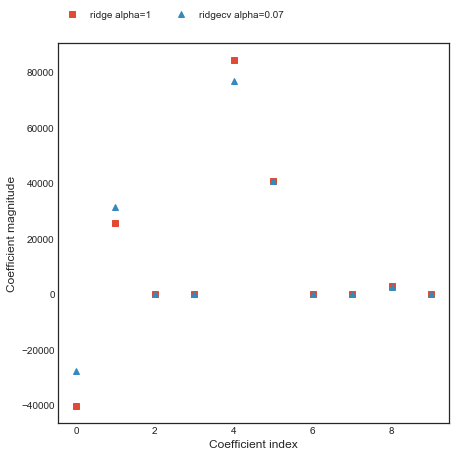
Similar procedure to check the penalty scale, and the coefficients would lose efficacy with alpha is about 100. Use the general Ridge method with penalty equals 1. The training score is about 0.59 and testing score is about 0.52, and mean square error is about 53079738090.8, which is similar with Lasso method.



Also try the cross validation and get the optimal penalty is about 0.07. Then training score is about 0.59 and testing score is about 0.51, slightly increase the difference by CV method. Both models include all the 10 variables and are not much overfitted.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **training accuracy** | **testing accuracy** | **MSE** |
| Ridge | 0.58981974 | 0.517361242 | 53079738091 |
| RidgeCV | 0.587334326 | 0.510201355 | 53136549734 |

The plot below shows the differences coefficients with different penalties and independent variables, like Ridge magnifies the effects of number of hits and home runs, and they are omitted by the model with larger penalty.



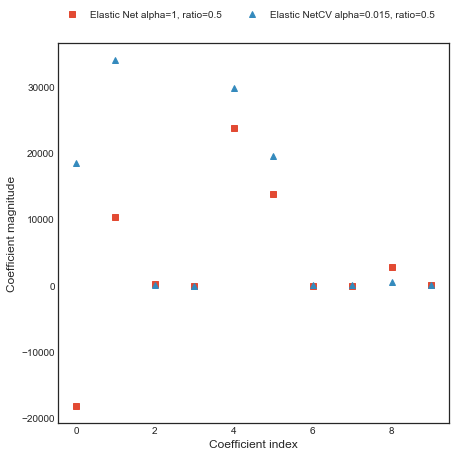
2.3. Elastic Net

Use the general Elastic Net method with penalty equals 1, l1 ratio equals 0.5. The training score is about 0.58 and testing score is about 0.51, and mean square error is about 53370627815.

Also try the cross validation and get the optimal alpha is about 0.005 and l1 ratio is 0.5. Then training score is about 0.43 and testing score is about 0.37. Both models include all the 10 variables and little overfitting problem.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **training accuracy** | **testing accuracy** | **MSE** |
| Elastic Net | 0.578 | 0.5147 | 53370627815 |
| Elastic Net CV | 0.4346 | 0.3653 | 69805689542 |

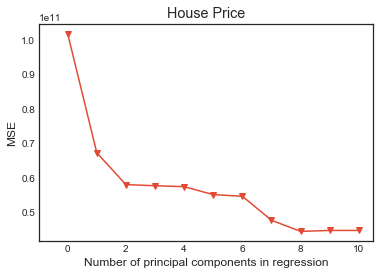
The plot below shows the coefficients magnitude with different penalties, like Elastic Net magnifies most of the effects of independent variables than Elastic Net CV method does.



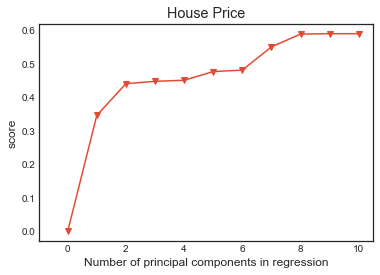
2.4. PCA

Firstly, with 10-Fold Cross Validation, choose the number of principal components in the regression.

Calculate Mean Square Error using CV for the 10 principle components by adding one component at the time, and the result is shown below.



The calculated regression scores with different number of predictors, shown as below. Combining the two indicators, MSE and score, we observe that even though the model with 2 components has the lowest MSE, its score is not good enough. Comparatively, for the model with 8 independent variables, there is dramatical decrease in MSE and significant increase in model score. As a result, we choose 8 predictors to do the following PCA regression.



For the final regression with PCA, in training sample, MSE is 44482920464 and accuracy score is 0.59; and in testing sample, MSE is 54464675225 and accuracy score is 0.53. This means there is no overfitting problem and bias-variance trade-off is satisfied, even if the model is not accurate enough.

|  |  |  |
| --- | --- | --- |
|  | **training accuracy** | **testing accuracy** |
| MSE | 44482920464 | 54464675225 |
| Accuracy | 0.58982059 | 0.534450859 |

2.5. SVM

In Support Vector Machine regression, we applied four types of kernel, which are linear kernel, polynomial kernel, sigmoid kernel and RBF kernel.

|  |  |  |
| --- | --- | --- |
|  | **training accuracy** | **testing accuracy** |
| Linear | 0.54043661 | 0.484202599 |
| Polynomial | 0.508338139 | 0.462796007 |
| RBF | -0.048108431 | -0.058428986 |
| Sigmoid | -0.048112605 | -0.058428986 |

According to the accuracy scores of three methods, the SVM regression is also overfitting. Comparatively, polynomial SVM is more appropriate in this project.

Conclusion

According to the model evaluations and bias-variance tradeoff, principal components analysis shows the best performance among the six models, for the highest testing score and the smallest gap between training and testing scores. Therefore, the PCA method is the optimal model for the baseball player salary prediction without overfitting problem.

III. Reference

https://www.kaggle.com/shree1992/housedata