Data Scientist Salary Prediction

Le Thi Thanh Tam

9/5/2022

Data Scienctist Salary Prediction Overview

- I started my career path as a Data Scietist and working as this role can be intellectual challenging. I wonder whether a Data Scientist get paid well ? How much a Data Scientist salary ? What skills are needed to improve ? ... so It would be great to go through the data and answer my questions by myself. This is my motivation to start this project.

Modelling Process

- First, I use **Multiple Linear Regression** as a based line model for Regression problem.
  - First to improve this linear model for better prediction accuracy and model interpretability, I use the **forward stepwise selection** method for selecting subsets of predictors. This method is known as among our various predictors, we believe just a subset of those be really related to the response (Salary).
  - Then use the **validation set approach** to test and select the best model for this data and run the multiple linear regression.
  - The result from Multiple Linear Regression with MSE is 995.8831. Then I consider the **heteroscedasticity** phenomenon from that result, that is the reason why I use the Lasso
- The **Lasso** : hope the coefficiet estimates can significantly reduce their variance for a more accurate prediction.
  - First, split the sample data into a training set and a test set in order to estimate the test error of the lasso.
  - Then perform **cross-validation** and compute the associated test error.
  - As expected, the lasso regression perform better than multiple linear regression compared MSE = 762.544 (mean square errors) in predicting the salary.
- **Random Forest**
  - Next consider even more general non-linear model tree-based.
  - The Random Forest model far outperformed the other approaches on the test and validation sets.

1. Multiple Linear Regression

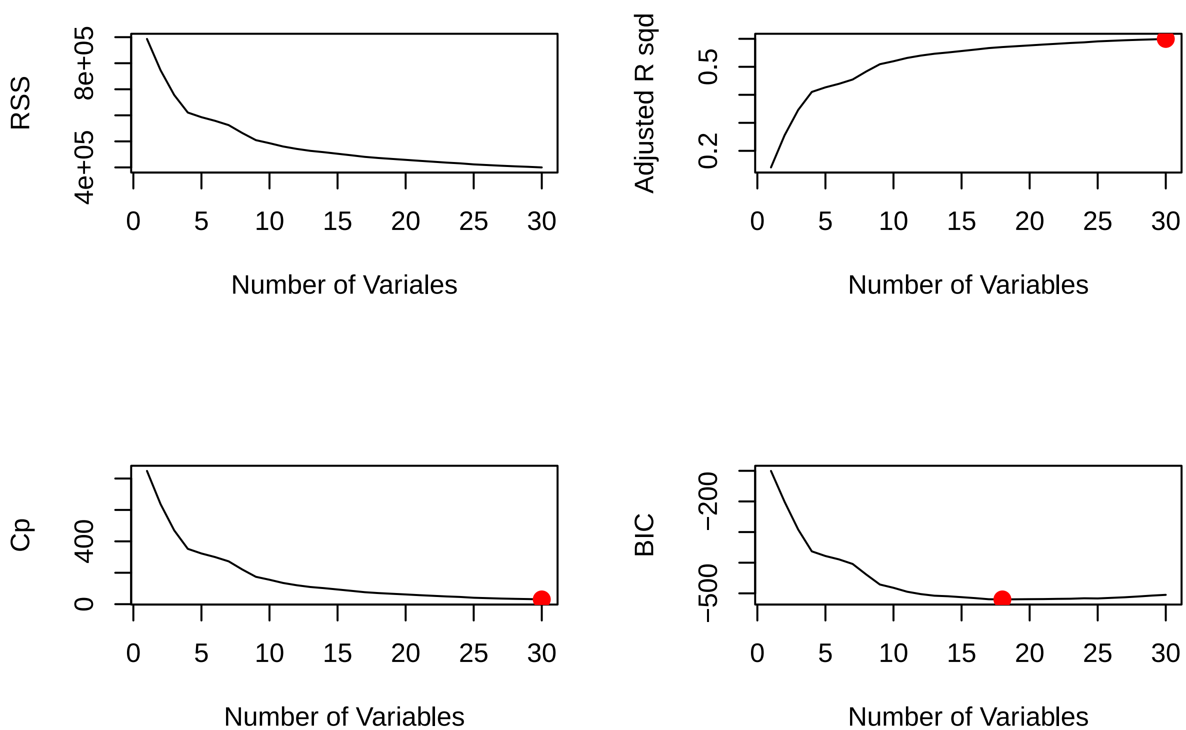
1.1. Forward Stepwise Selection

- Method for selecting subsets of predictors.
- Forward stepwise selection begins with a model containing no predictors, and then adds predictors to the model, one at a time until all of the predictors are in the model.
- This is the result of R squared statistics by using this approach :

```
##  [1] 0.1419651 0.2574784 0.3487904 0.4135071 0.4305654 0.4439358 0.4598250
##  [8] 0.4889180 0.5152634 0.5264122 0.5386571 0.5475069 0.5546571 0.5596790
## [15] 0.5653519 0.5709860 0.5769100 0.5809774 0.5844411 0.5878818 0.5913969
## [22] 0.5946584 0.5979972 0.6008059 0.6046067 0.6071140 0.6095831 0.6116869
## [29] 0.6135834 0.6158089
```

- we see that the R2 statistic increases from 14 %, when only one variable is included in the model, to almost 62 %, when all variables are included. As expected, the R2 statistic increases monotonically as more variables are included.

Plot RSS, Adjusted r squared, Cp and BIC for all of the models



- We see the best model with the highest Adjusted R squared and lowest Cp is 30-variable model.

- The best two-variabe :

```
##         (Intercept)      job_title_simDA seniority_by_titleSR
##            98.70957            -39.95864             27.88535
```

- For this data, the best one-variable model contains only **job_title_sim** for Data Analyst position, the best two-variable model additionally includes **seniority_by_title** with Senior level.

- However, to obtain the accuracy of that, we need to perform on the test set

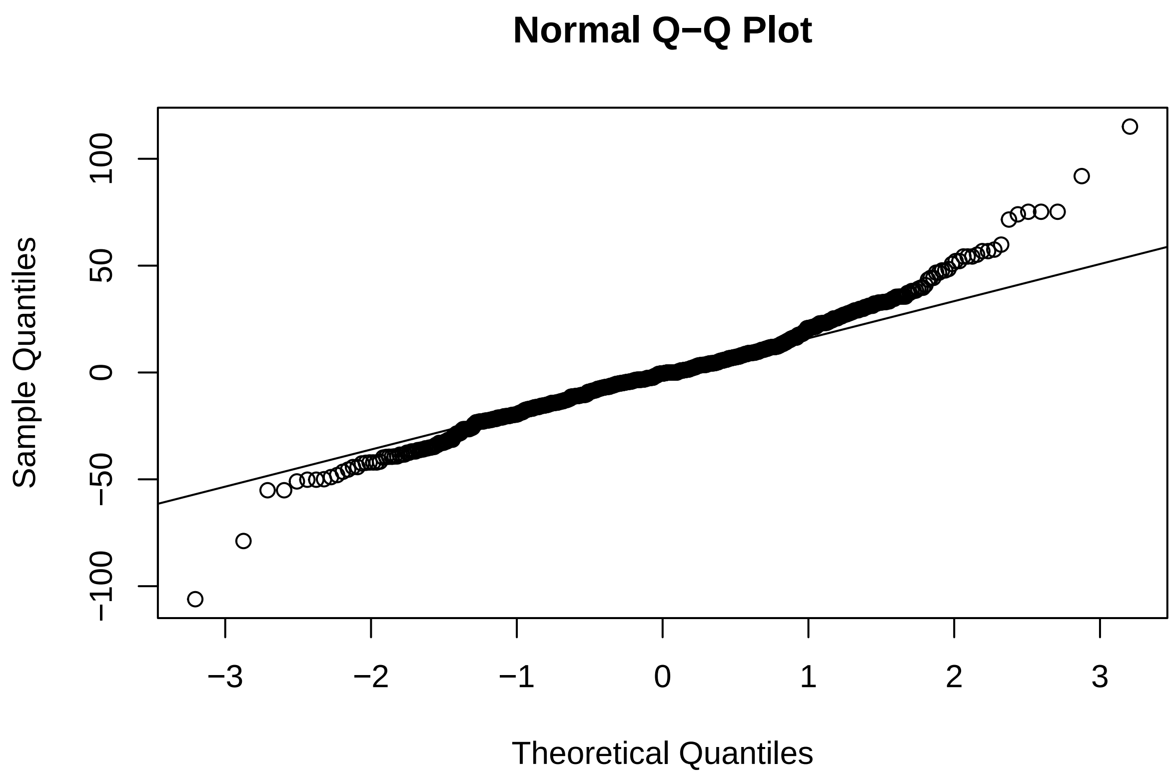
1.2. The validation set approach

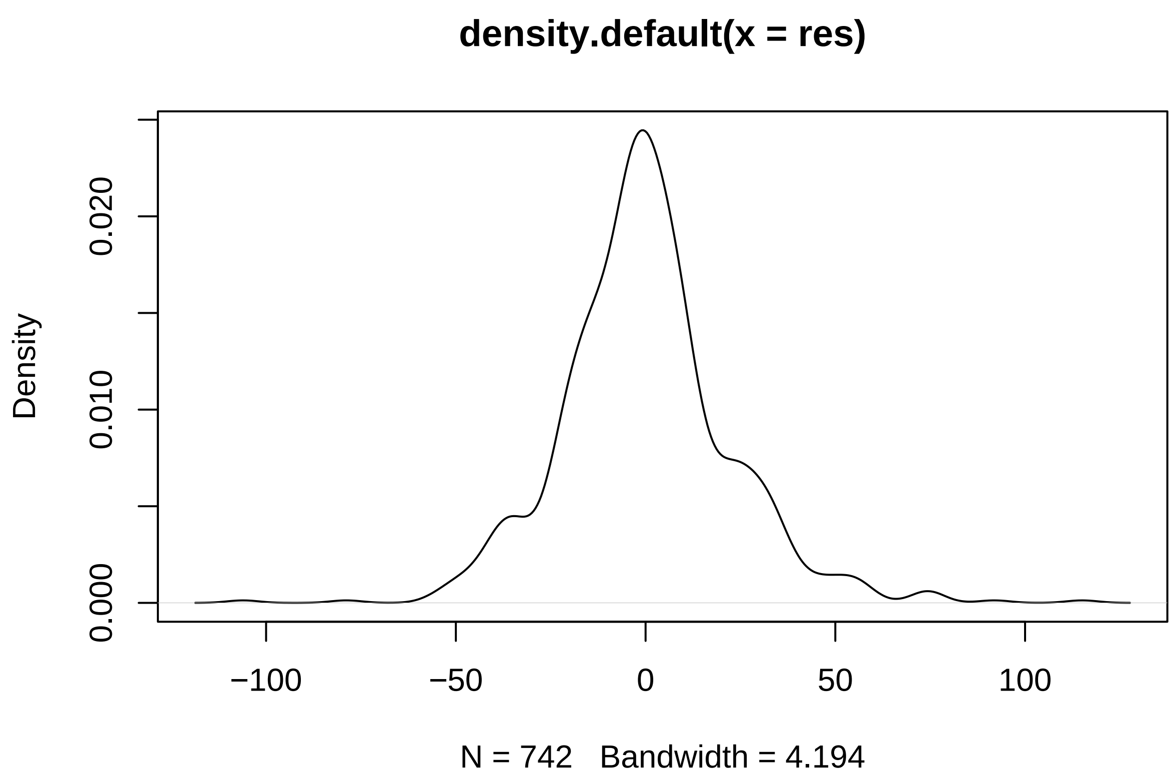
- Next step I will use the the validation set approach to test and select the best model for this data and run the multiple linear regression

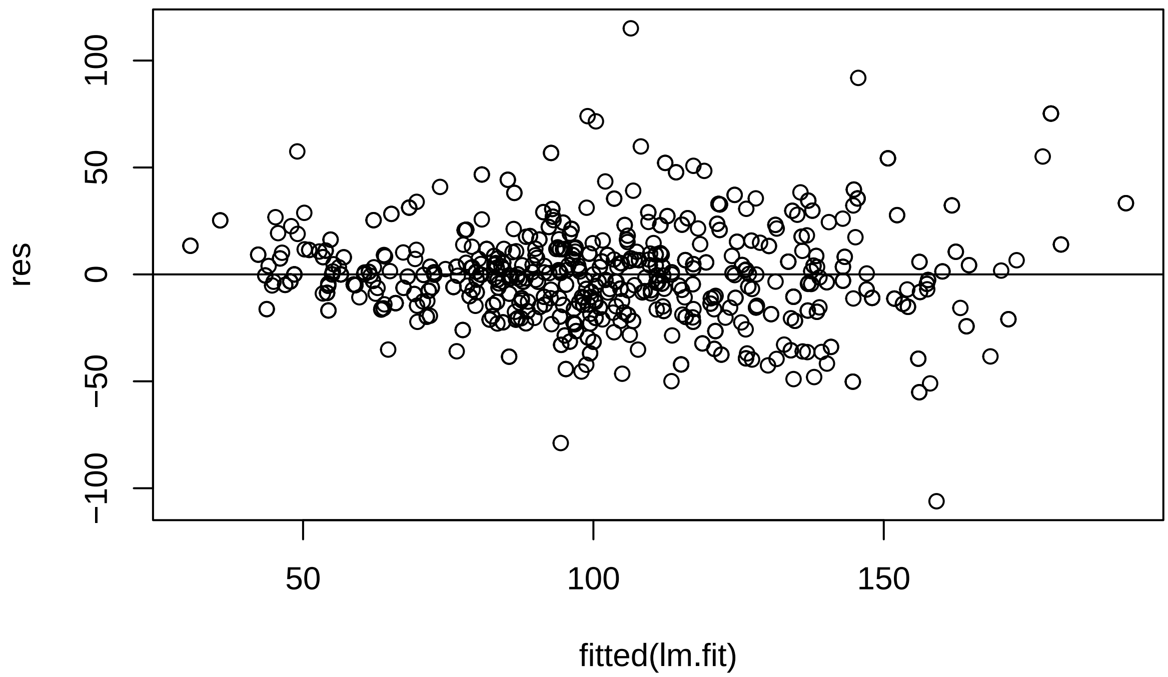
```
##              (Intercept) type.of.ownershipNonprofit
##             150.17574160               -15.78661700
##                  sector4                    sector5
##             -35.49952098                11.76052117
##                  sector7                   sector13
##             -46.11832206                 8.61749668
##                 sector18                   sector19
##              17.72221970               -19.03778925
##                   hourly          employer.provided
##             -11.74302734                42.71252574
##           job.locationAZ             job.locationCA
##             -18.08492549                22.98298272
##           job.locationCT             job.locationFL
##             -23.57062976               -16.32591917
##           job.locationGA             job.locationNM
##             -28.87810490               -40.88225180
##           job.locationTN                        age
##             -18.96916595                 0.06431561
##                   python                        sas
##               6.88794567                 9.18734423
##                    keras                    pytorch
##              17.74438133                -8.38286690
##          job_title_simDA            job_title_simDE
##            -100.59759389               -67.23090906
##          job_title_simDS             job_title_simM
##             -60.25438536               -80.60934710
##         job_title_simMLE             job_title_simN
##             -47.52067898               -82.92112034
##     seniority_by_titleSR                   num_comp
##              25.92271992                 1.39228862
```

- Our final best model after using validation set approach including 29 variables and the MSE is 995.8831

Consider the result from multiple linear regression



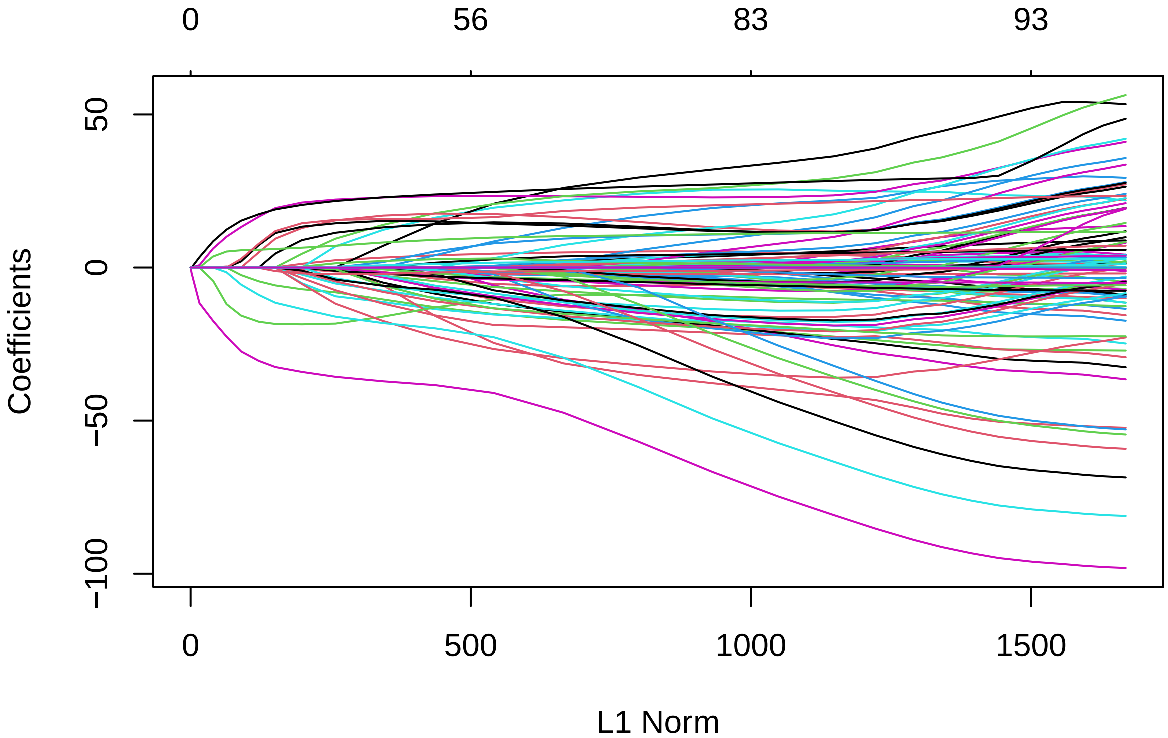




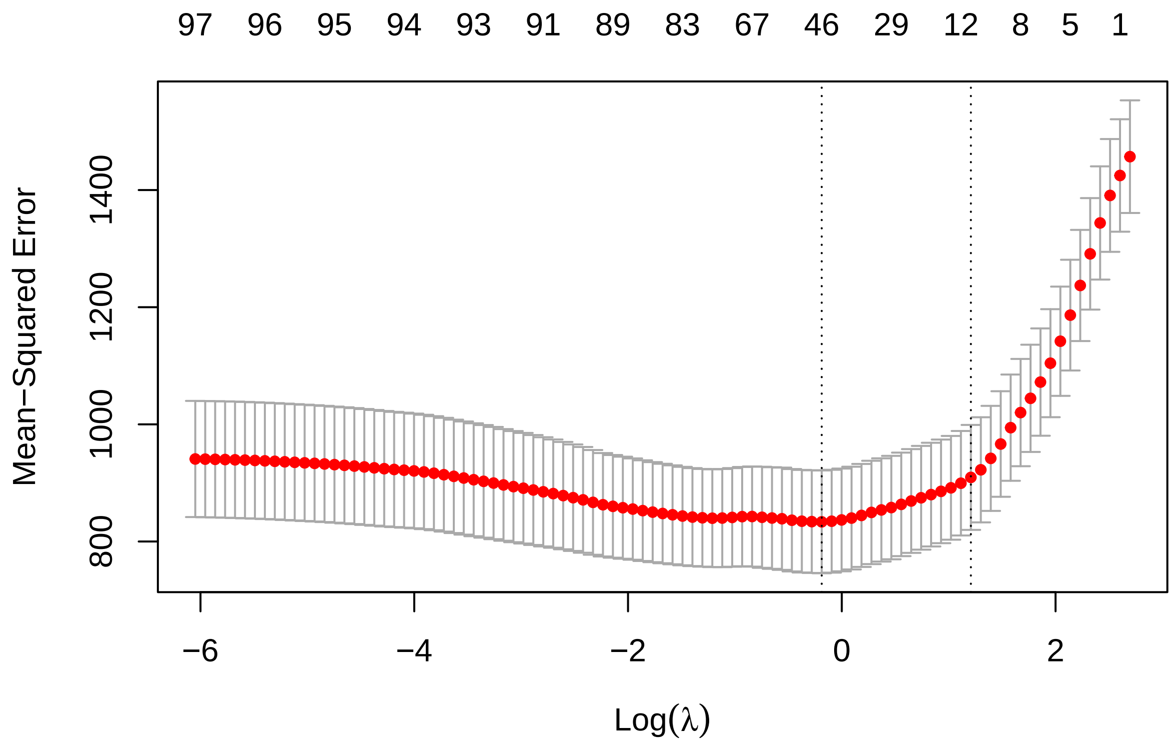
- we can see the data showing heteroscedasticity. the residuals are observed to have unequal variance

Lasso

- Now I will perform the lasso _ the techniques for shirinking the regression coefficients towards zero. By using this technique, hope the coefficiet estimates can significantly reduce their variance for a more accurate prediction.
- Now I will perform the lasso in order to predict salary on this data



- we now see some of the coefficients will be exactly equal to zero
- perform cross-validation and compute the associated test error

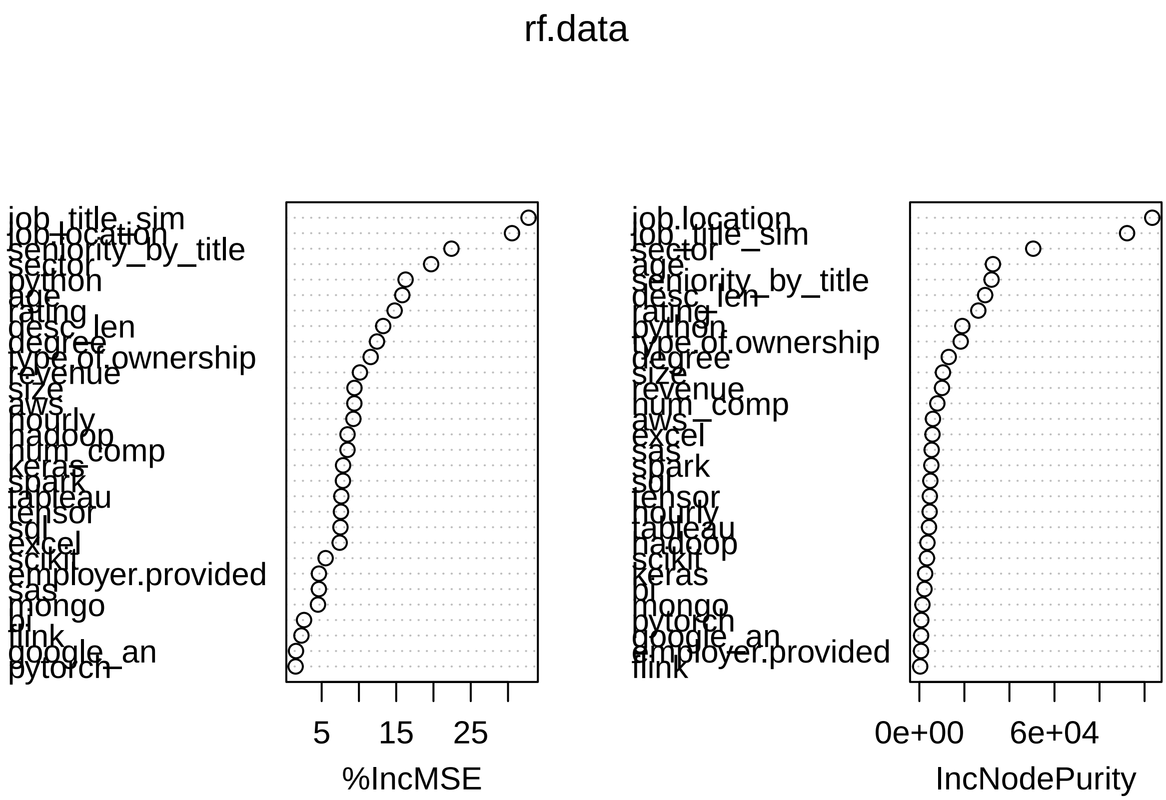


- As expected, the lasso regression perform better than multiple linear regression compared MSE = 762.544 (mean square errors) in predicting the salary.

RandomForest

- Again I split the data into the train and test set then compute the MSE

Plot of the importance measures



- The results indicate that across all of the trees considerd in the random forest, the wealth level of the job title (job_title_sim) and job_location are by far the most important variables.
- The Random Forest model far outperformed the other approaches on the test and validation sets with MSE = 575.7547

For coding review, please check my github page. Thank you