Salary Prediction

Course project for STAT 684- Professional Internship

Linhan Hu

Abstract

In this project, I analyzed the new bank salary dataset from ecampus. It is a multivariate dataset aiming to find the most important features within 8 different features to predict an employee's current salary. I choose to use four different methods to determine the most important features: Lmg, Random Forest Trees, Boosting and Lasso. According to these four methods, Starting Salary; Education; Jobcategory; Age play the most important role in the prediction of current salary. This report documented in detail about exploratory data analysis; transforming the data and fitting the model.

Introduction

Employers typically adjust their market data when determining how much to pay a specific employee to do the job. After they determine the value of the position by researching the data on pay practices for comparable jobs at comparable companies, they adjust the data to reflect the employee's background and experience. The goal of this project is to find a model with important features that can predict employee’s salary properly.

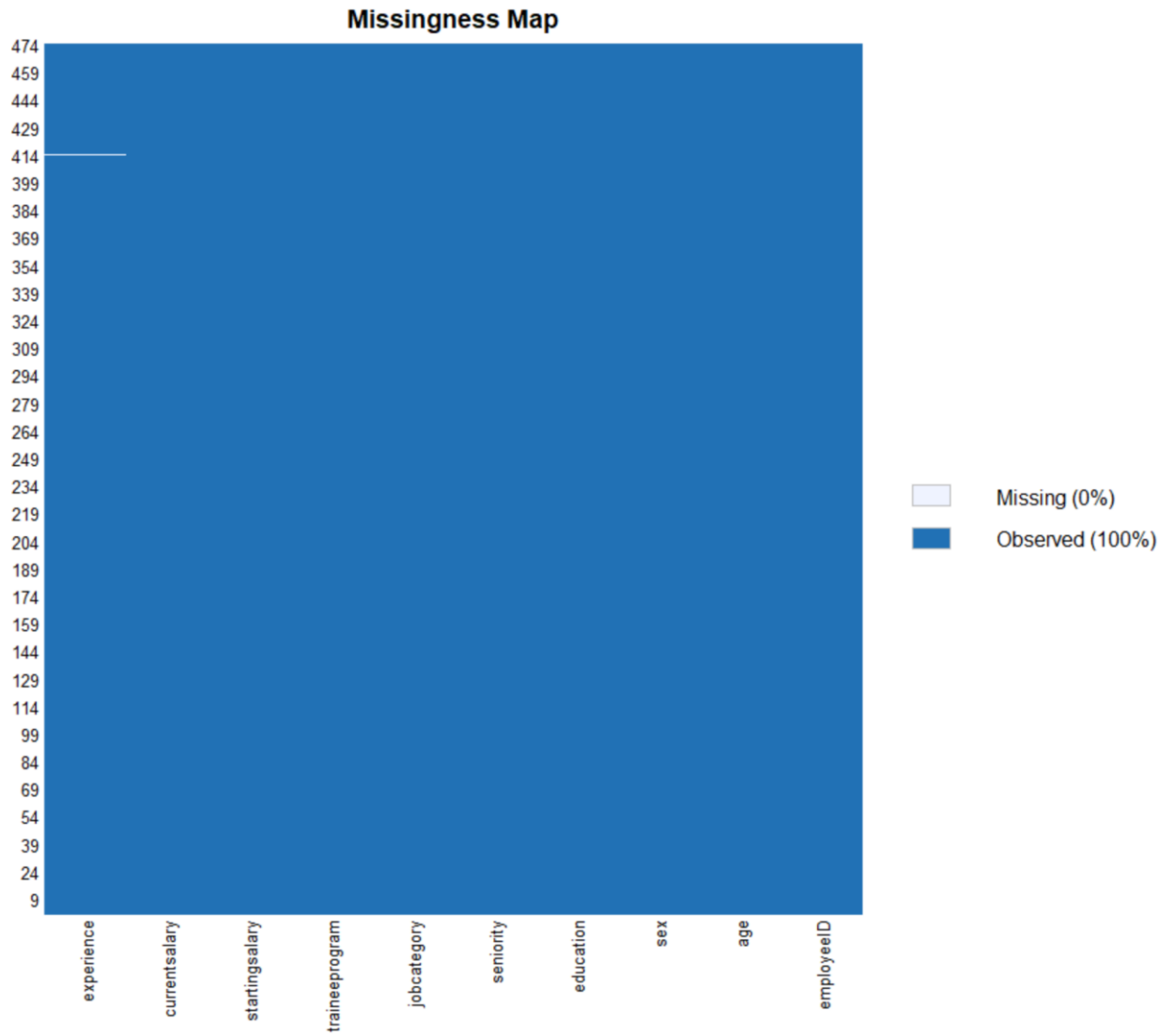
To achieve the goal of this research, this project used machine learning models such as Lasso, Random Forest Tree, and Boosting Tree. The reason to use these methods is that they often yield a relatively accurate prediction and are easy to interpret the result at the end. The result of feature importance can be directly used by the employers and employees to provide a general guideline on how much they should be paid.

The data consists of 474 observations from New bank salary.

Data and Preprocessing

The dataset is the salary dataset from Ecampus with 474 observations, it is composed of 5 numerical variables: Age; Education; Experience; Seniority and Starting Salary. And 3 categorical variables: Sex; Jobcategory and Traineeprogram. To fit this data into our model, the categorical variables are dummy-recoded into separated indicator variables.

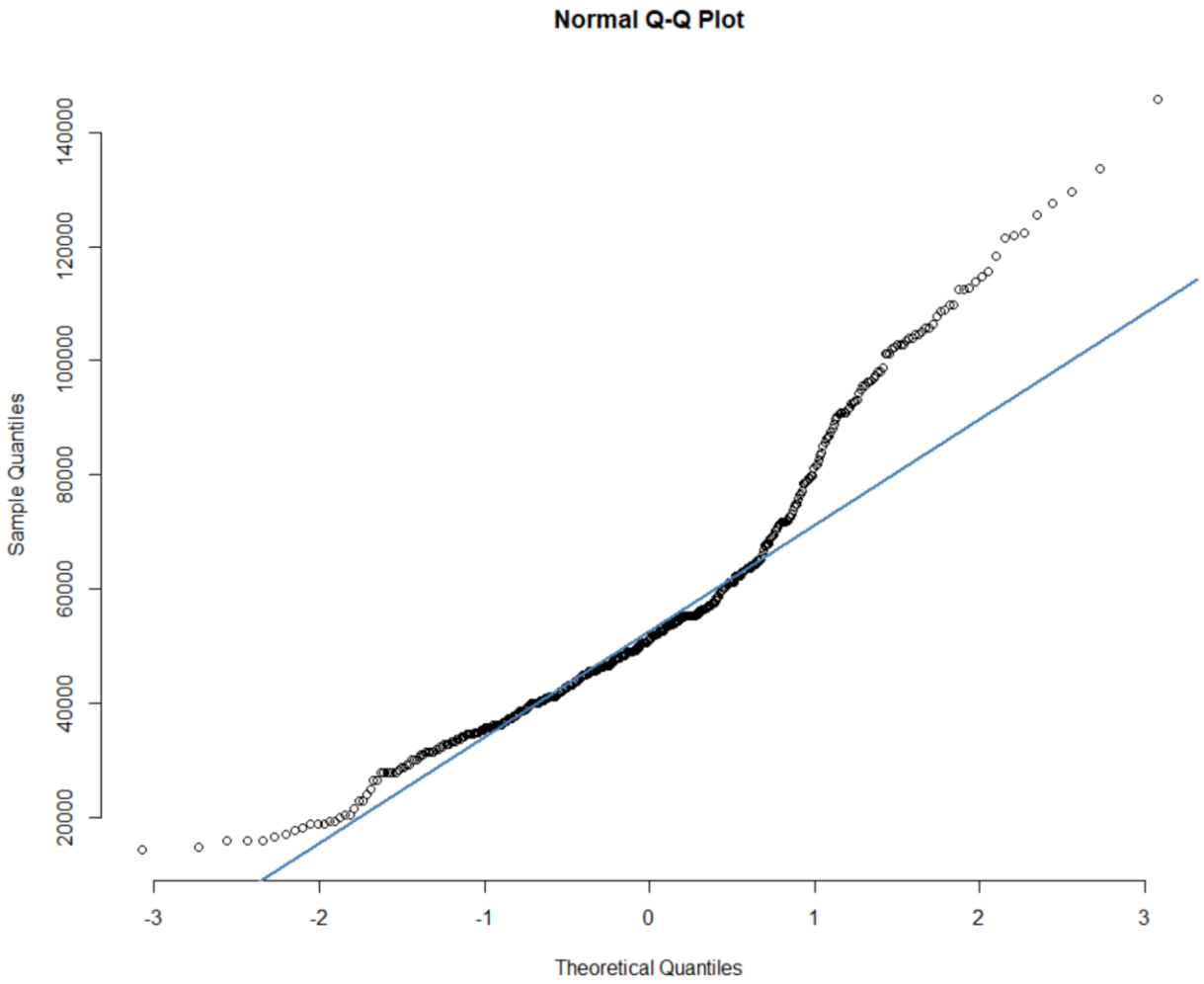
First, we visualize the whole dataset to check for the missing values where white area indicates the missing value:



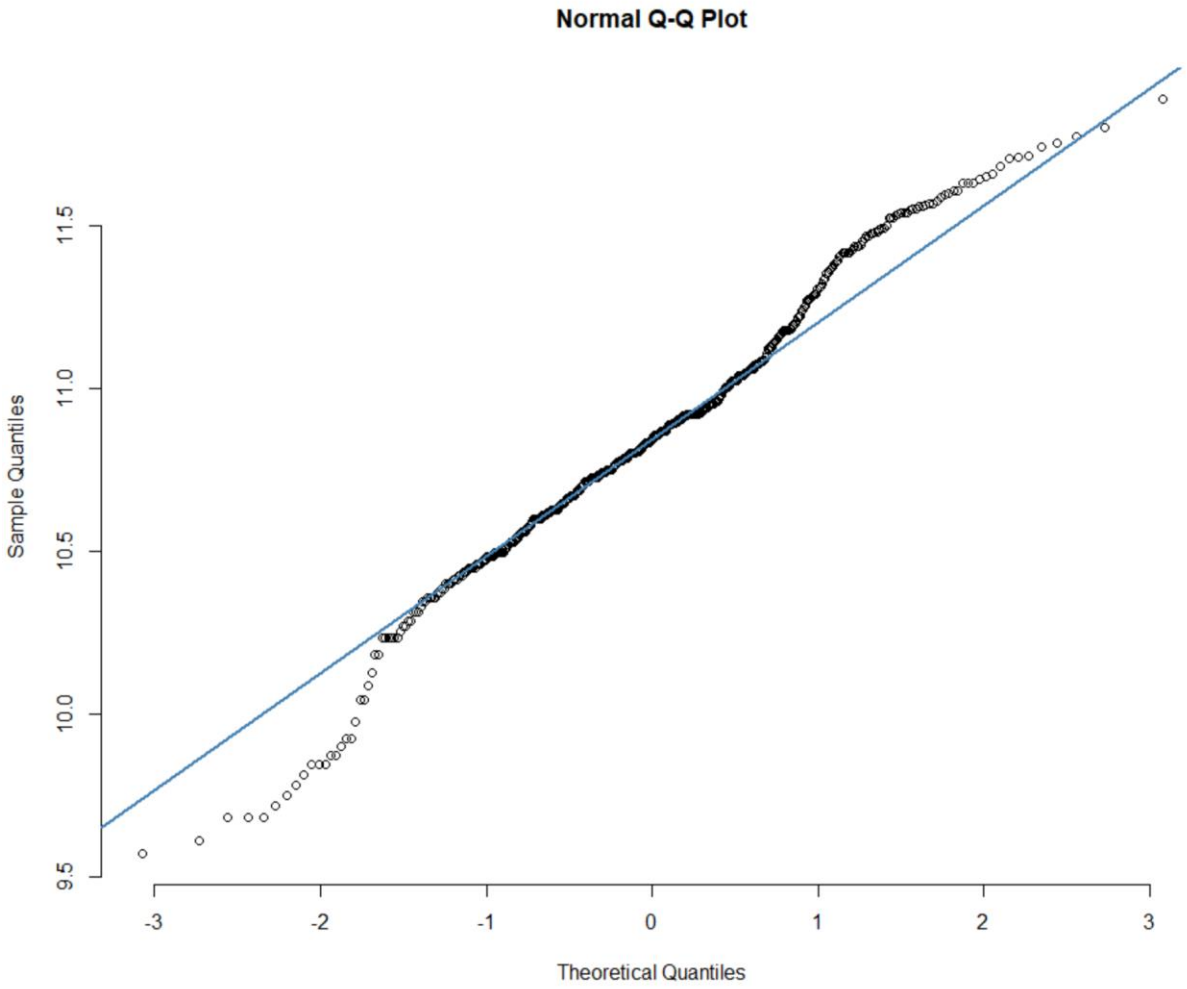
So, the dataset is complete with only one missing value which is perfect for the analysis part.

Analytical Process

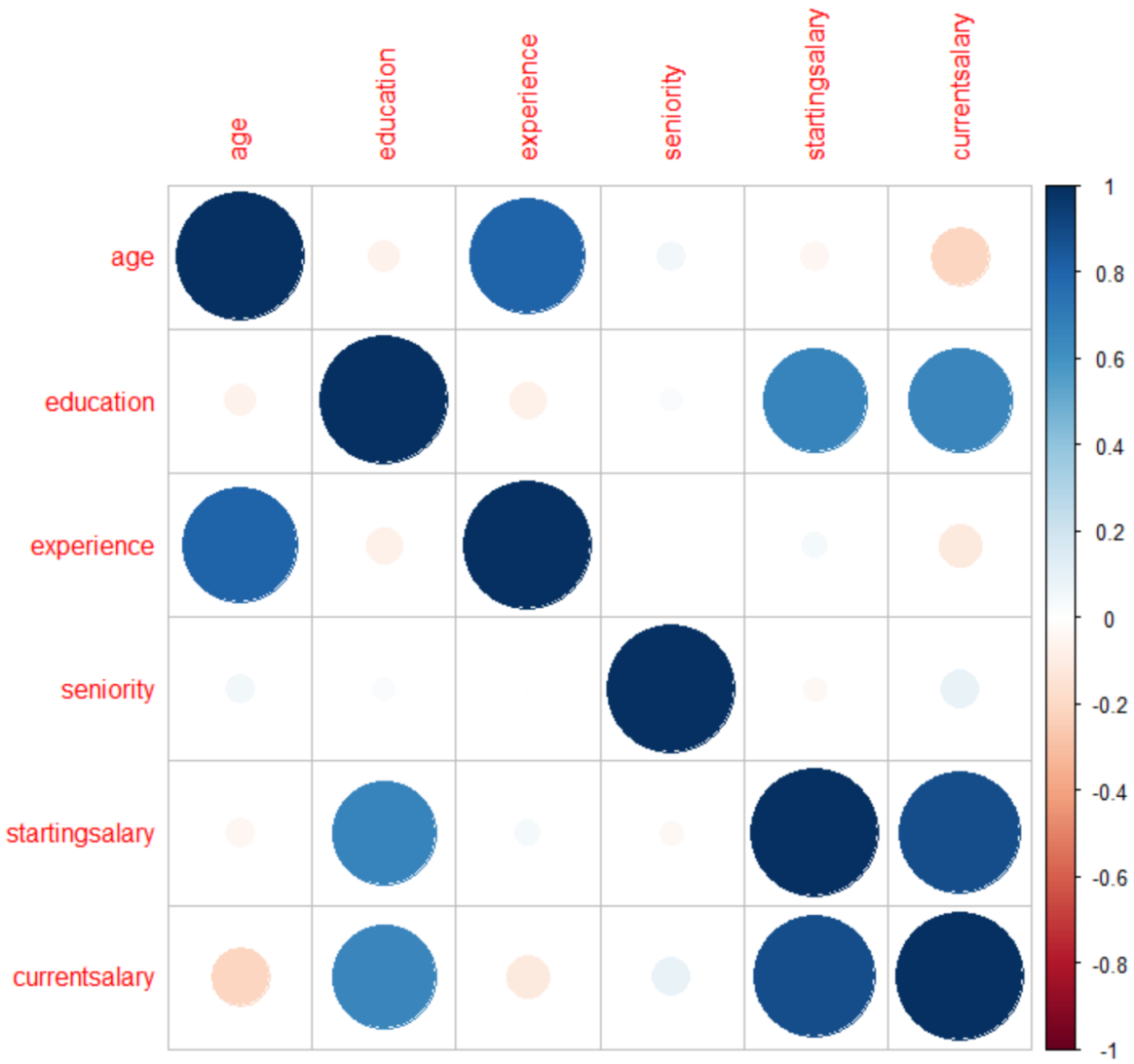
To meet the assumption of the model, the dependent variable, Current Salary should be normally distributed, but it is right skewed with skewness of 0.99 as follows:



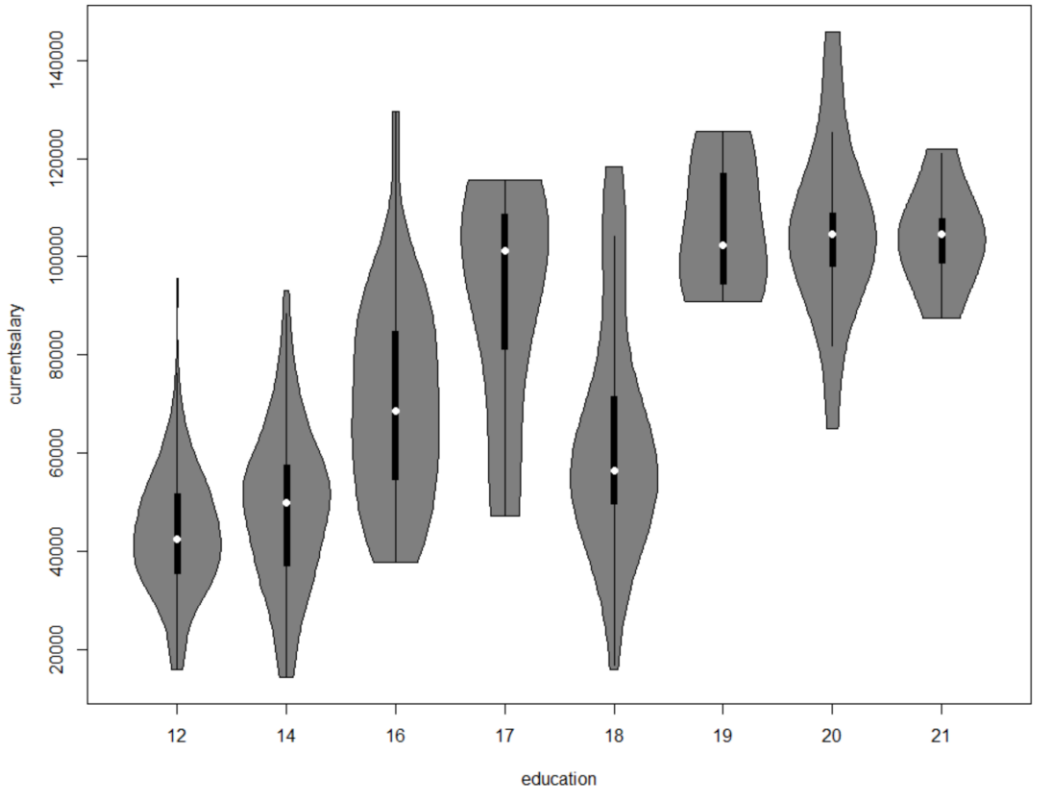
To better fit it into the model, I log-transformed the dependent variable to get an approximately normal look:

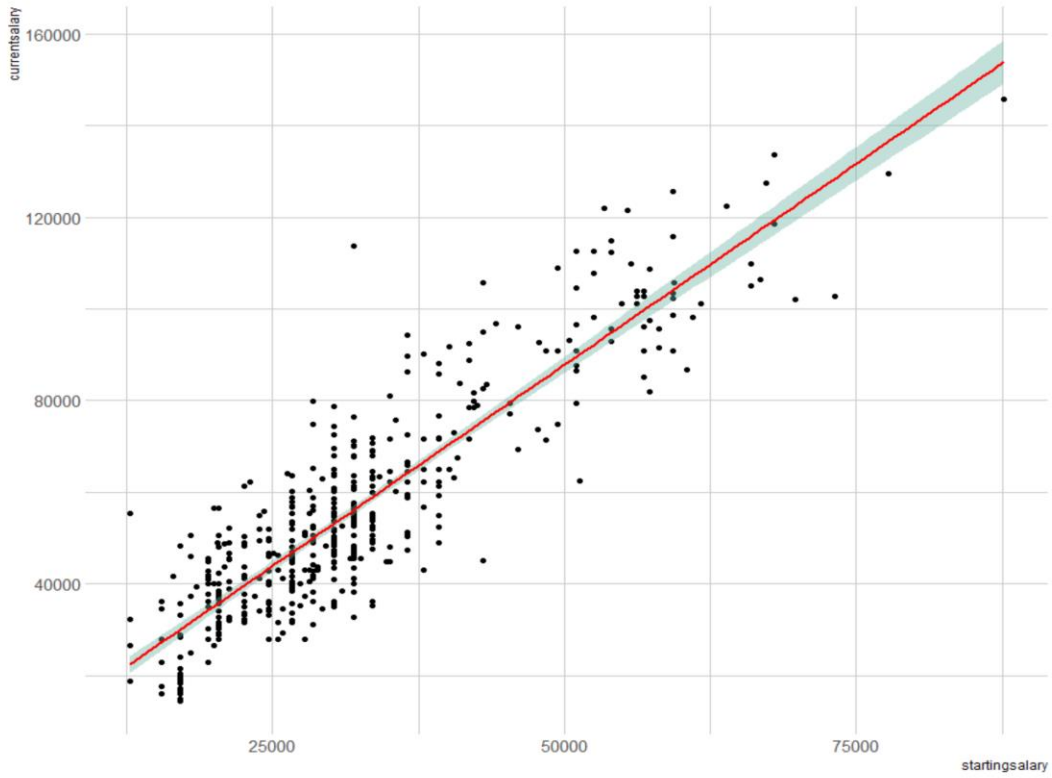


In order to better understand the relationship between the variables, I conducted a correlation filtering process to check the Pearson correlations between all the variables. The following plot shows the correlation matrix:



We can see that education and starting salary have the highest correlation with current salary, so I visualized their relations as follows:


currentsalary
20000
40000
60000
80000
100000
120000
140000
12
14
16
17
18
19
20
21
education


currentsalary
160000
120000
80000
40000
25000
50000
75000
startingsalary

As we can see, although there could be some potential outliers, the relationship between starting salary and current salary is approximately linear.

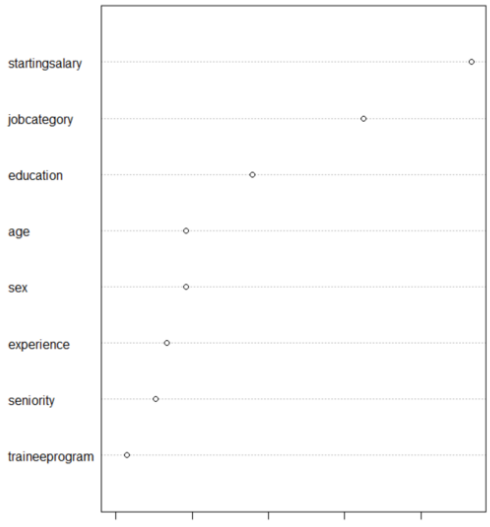
To get more understanding of the qualitative variables, I first used methods called lmg which essentially calculate the proportion of R^2 each predictor contributed in the model, where R^2 represents the proportion of variance explained by a set of predictors. And lmg yields the following feature importance:

```
Relative importance metrics:

                              lmg
x.age                0.0648357659
x.sexM               0.1256706244
x.education          0.1273958113
x.experience         0.0181503856
x.seniority          0.0179495588
x.jobcategoryExecutive  0.0662241352
x.jobcategoryFinance    0.0124920743
x.jobcategoryIT         0.0649947258
x.jobcategoryManagement 0.0259016886
x.jobcategorySecurity   0.0145357100
x.jobcategoryTeller     0.0426267675
x.traineeprogram1       0.0103720009
x.traineeprogram0       0.0002271963
x.startingsalary        0.4086235553
```

Starting Salary; Education; Sex are the most important variables in predicting Current Salary.

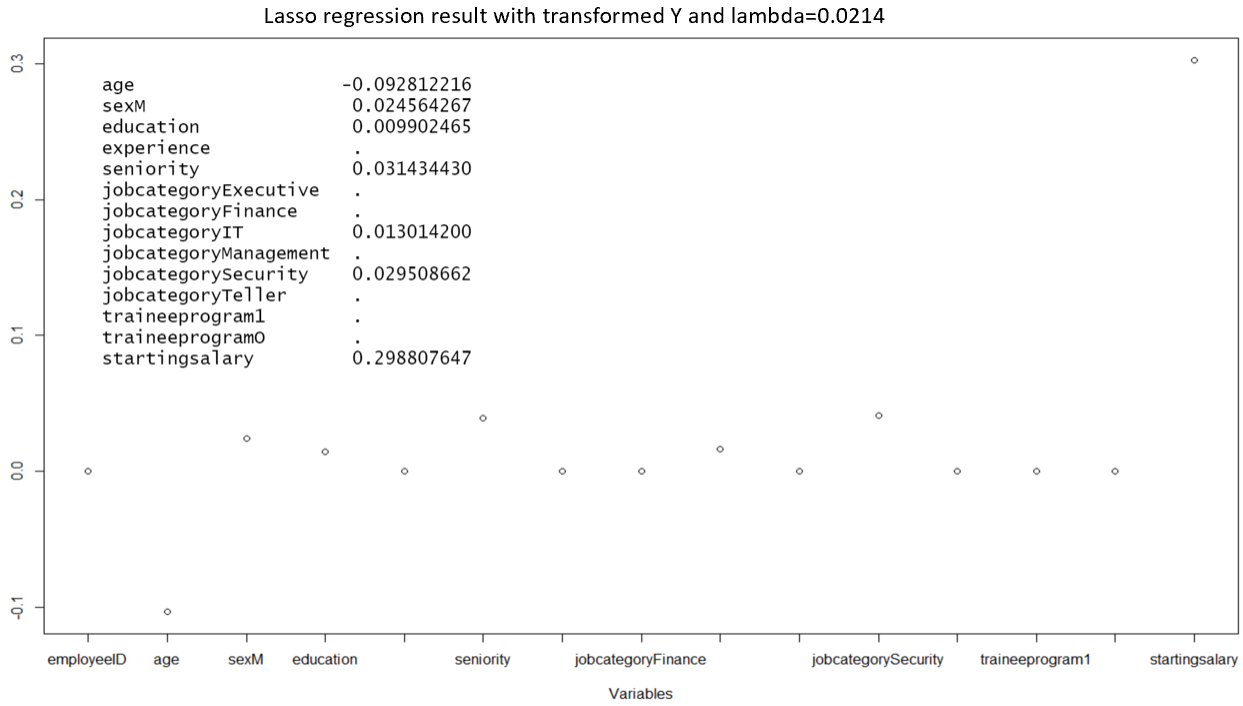
I fit a Random Forest model in Predicting Current Salary. We could utilize a randomized permutation test to view the importance of different features, including qualitative variables. And the random Forest yields the following feature importance:



	IncNodePurity
age	19212902574
sex	18721731795
education	34878161666
experience	13720062859
seniority	10379392764
jobcategory	60938854163
traineeprogram	2844437777
startingsalary	96488159122

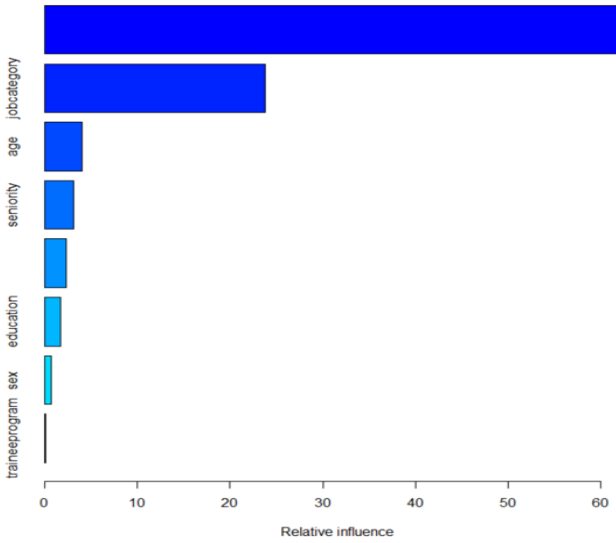
Starting Salary; Jobcategory; Education are the most important variables in predicting Current Salary.

Also I conducted lasso regression. First I used cross validation methods to find the best lambda value corresponding to 1 standard error above the minimum MSE so I can get usually a more conservative (fewer variables) solution than the minimum MSE. The result is as follows:



Starting Salary; Age; Seniority are the most important variables in predicting Current Salary.

Also, Boosting trees provide accurate predictions as well. So my last model is boosting trees which give us the following result:



startingsalary	63.9991472
jobcategory	23.8311240
age	4.0254098
seniority	3.1863997
experience	2.3660725
education	1.7450979
sex	0.7404364
traineeprogram	0.1063126

Starting Salary; Jobcategory; Age are the most important variables in predicting Current Salary.

Conclusion

In this project, I used four different models to fit the dataset in order to find important features in predicting one's current salary. All four methods agree that starting salary is the most important predictor, and Age, Education, Jobcategory follows.