1．Data exploration

After importing the training dataset and test dataset, the columns which contains text and cannot be used to analyze directly and ‘Rating’ column which has nothing to do with salary are removed. There are no null values in both datasets , but the illegal values are found in the company age column and seniority column. As for the company age ,the outliers are also found but reasonable so not to be removed, so the illegal values can be replaced by the mean value of remain all data in company age. Since the proportion of ‘na’ ,the illegal value in seniority column is above 80%, we don’t use this column for data visualization and prediction. Finally ,jop\_simp column is transfer to several categorial variables and the categorial variable ,high\_salary is added into both datasets depending on avg\_salary.

2.Data Visualization

2.1 the distribution of avg\_salary from training data and test dataset

According to the figure ,we can find that the two distribution is similar and nearly symmetrical, the ascending branch of which are even coincident and the descending branch of which are slightly different especially when the avg\_salary is greater than 150. No matter in training data or test data ,the probabilities of high\_salary are obviously smaller than 0.5,which means the distribution of data between high salary and low salary are not symmetric. So, we cannot only concern the accuracy of prediction when measuring models.

2.2 explore the relationship between high salary and the kind of job.

For the training data, there are no significant differences in the medians of the various positions, except that the average salary of director is slightly higher than that of the other groups. So, the job\_simp may have little influence on average salary. Besides, the box plot of test data is different from the training data’s especially for the width of analyst and data scientist. Additionally , there are few managers and directors in the test dataset.

2.3 visualize categorial variables (same state, python, excel, Hadoop, spark, aws, tableau, big data)

For each category variable, if the share of certainty in the high wage group is significantly different from that in the low wage group, the variable is likely to influence high wage. If this rate in training data is higher than that in test data, the categorial variable has positive influence on the high variable, otherwise ,it may have negative influence. However ,it is impossible that commanding skills such as python ,excel causes low salary logically. For same state, hadoop, spark, tableau, big data, the training data and test data shows opposite results.

2.4 heck whether company age has influence on average salary.

According to the scatter plots ,most observations gather in the area whose company age is less than 100. Besides ,there are no obvious relationship between average salary and company age ,which means company age may have nothing to do with average salary.

3.Dimension reduction

Using the explained variance ratio to measure the principal comment components ,the ratio is higher than 0.8 when the number of principal components is higher than 11, which means no one variable is more important than the others conspicuously. The reason may be that the job\_simp variable is transfer to 5 categorical variables which should have the same influence on high salary. Overall ,the outcome of PCA is not favorable.

4.Regression and Prediction

Only using accuracy to measure models is not reasonable for unsymmetric data, because the accuracy can be high even predict all observations to 0.So for every model ,the accuracy and sensitivity should be calculated to compare.

Firstly，I use 4 random forest models and select the best three models according to cross validation. The n\_estimators of three models selected are 500,100,2 respectively. Then ,I use best SVM with linear kernel, best SVM with radial kernel, and classification tree to see whether there is improvement.

RandomForestClassifier with n\_estimators =500

Test data:

Accuracy = 0.75 sensitivity = 0.12

Training data : Accuracy = 0.9117647058823529

RandomForestClassifier with n\_estimators =100

Test data:

Accuracy = 0.7327586206896551 sensitivity = 0.12

Training data : Accuracy = 0.9117647058823529

RandomForestClassifier with n\_estimators =2

Test data:

Accuracy 0.7586206896551724 sensitivity = 0.04

Training data : Accuracy = 0.8445378151260504

Linear SVM with C = 0.01

Test data:

Accuracy = 0.7844827586206896 sensitivity = 0

Training data = Accuracy = 0.8004201680672269

Radial SVM with C = 0.01, the result is the same as Linear SVM.

Classification tree with max\_depth =2 ，the result is the same as Linear SVM and Radial SVM.

For the SVM and classification tree ,although accuracy is close to 0.80,the sensitivity is 0 which means these models predict all the observations to 0. So these models are useless and fail to predict. Considering accuracy and sensitivity, the random forest model with 500 trees performs best ,but there may be overfitting problem because the training data accuracy is higher than 0.9.Additionnaly ,the reason why the best model still performs worse relatively is that there are only 450 observations to train models which maybe not enough. If I can get more data from the resource website ,maybe the performance can be improved. However , the model selection may output different result.