**Project 2 Predictive Modeling of world university ranking**

Weina Ma

Contents

[1. Abstract 3](#_Toc465071743)

[2. Data Preparation—Shanghai & Timesdata 3](#_Toc465071744)

[2.1 Class variables 4](#_Toc465071745)

[2.2 Features 5](#_Toc465071746)

[2.3 Final dataset 8](#_Toc465071747)

[3. Modeling -- Shanghai 9](#_Toc465071748)

[3.1 Fitting 7 Different Classification Models-- Shanghai 9](#_Toc465071749)

[3.2 Discuss the advantages of each model -- Shanghai 11](#_Toc465071750)

[3.3 Important features-- Shanghai 12](#_Toc465071751)

[3.4 Model evaluation--Shanghai 12](#_Toc465071752)

[3.5 Feature selection and simple model--Shanghai 13](#_Toc465071753)

[4. Modeling -- Timesdata 14](#_Toc465071754)

[4.1 Fitting 7 Different Classification Models-- Timesdata 14](#_Toc465071755)

[4.2 Important features-- Timesdata 16](#_Toc465071756)

[4.3 Model evaluation-- Timesdata 16](#_Toc465071757)

[4.4 Feature selection and simple model-- Timesdata 17](#_Toc465071758)

[5. Evaluation and Deployment - Shanghai & Timesdata 18](#_Toc465071759)

[5.1 Prediction – Shanghai 18](#_Toc465071760)

[5.2 Prediction – Timesdata 19](#_Toc465071761)

[5.3 Conclusion 19](#_Toc465071762)

[6. Acknowledgment 20](#_Toc465071763)

[7. References 20](#_Toc465071764)

# 1. Abstract

Due to the importance of university ranking in real world, this report is trying to figure out which university will improve over time in the ranking and which university will fall back by using 7 different algorithms, including Random forest, Conditional Inference Tree, C 4.5, Rule-based classifier, Linear Support Vector Machines, Artificial Neural Network and Gradient boosting decision tree. According to model evaluation, we figure out the most efficient model is random forest. Also, the most important features are Improved\_pub and award\_X2014 in shanghai dataset. The most important features are Improved\_teaching and Improved\_research in timesdata. By the end of our project, we made a 2016 rank prediction of University Southern California for stake holders based on the most efficient model.

*Key words: University ranking, model comparison, feature selection, prediction*

# 2. Data Preparation—Shanghai & Timesdata

In this report, we will use shanghai and times data to build different models. Since the number of observation of 2010 to 2015 are almost same in shanghai and timesdata each year, which make it easier to create rank differences with few missing values.

For shanghai dataset, there are 500 observations each year. 

For times dataset, there are 400 observations each year. 

|  |  |
| --- | --- |
|  | In this report, we will use data in 2010, 2011, 2014, 2015 of shanghai to predict whether rank will improve or not in 2016. |
|  | In this report, we will use data in 2013, 2014, 2015 of timesdata to predict whether rank will improve or not in 2016. |

While in cwurdata, we have 100 observations in 2013 and 1000 observations in 2014. Since the difference of rank 2013 and rank 2014 is 900, we get 900 missing values, which will case inaccuracy in building models. Therefore, we choose shanghai and timesdata to build the models.

## 2.1 Class variables

We are interest in which university will improve over time in the ranking and which university

will fall back. Thus, we would create a class variable called improved. That could be true or false. If it is true, the university will improve, if it is false, it doesn't improve.

So, we use recent rank improvement variable as class variables, which is rank of 2014 minus rank of 2015.



We would like to discretize the class variables.



|  |  |  |
| --- | --- | --- |
|  |  | **Shanghai** |
|  |  | **Times data** |

According to the figures above, we can conclude that the distribution of rank improvement between 2014 and 2015 in both shanghai and timesdata are reasonable. Because those distributions are normal distribution with few outliers. Also, most of ranks didn’t change much, which make sense.

In addition, because classification algorithms may show error dealing with missing values, so we imputed the final dataset.

For class variable, we choose delete the missing values because missing values of class variables are 40 and 30.



I have tried to discretize other features, but it seems useless. Numeric features is suitable for the model in this report.

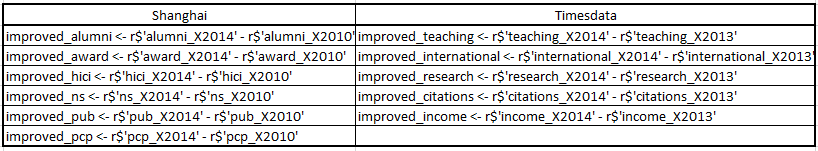
Thus, we can conclude that the class variables are good with enough quantity and good quality.

## 2.2 Features

First, we remain features of 2010 and 2014 of shanghai data in the final dataset. Also, we remain features of 2013 and 2014 of timesdata in the final dataset.

|  |  |
| --- | --- |
| Shanghai | Timesdata |
|  |  |

Second, because we are concerned about whether the rank will improve or not next year, so improvement of features in the past may affect the rank. So, we add most feature improvement as variables. The codes are shown below.



From the figures above, we find out the distribution of all the improved variables are normal distribution, which means our dataset is good and reasonable. What’s more, the majority didn’t change much according to the pictures.

|  |  |
| --- | --- |
| Shanghai | Timesdata |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

In addition, aggregate all features may be one important variables. So add all features up.



|  |  |
| --- | --- |
|  |  |

Because number of students is much greater than other features. For instance, the range of income is 27.9 to 100 while the range of number of students is 1283 to 120986. So, we made a log transformation in both 2013 and 2014. Part results are shown below.

|  |  |  |
| --- | --- | --- |
|  |  | Timesdata |

Because missing values of numeric data in the final datasets are few, we choose replace missing values with mean.



We have tried to import external data, but the result is not good . So we just use the data transformation and manipulation within those two datasets.

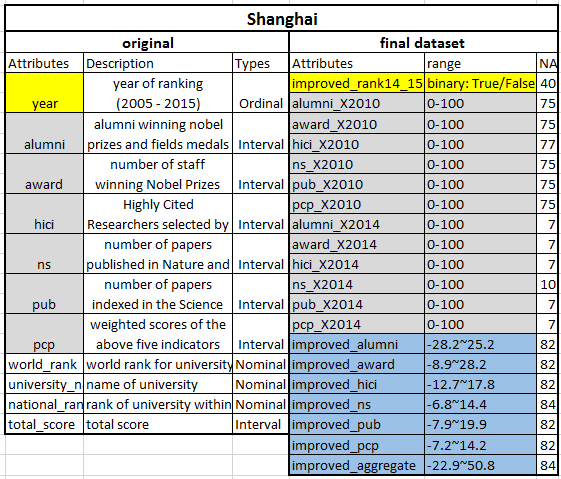
In addition, we delete some original variables, like university name, total score, year, world and national rank.

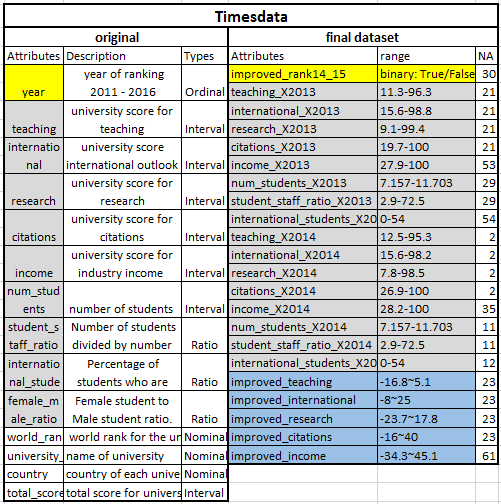
Remaining university name is bad because it looks at the university name, then made models based on university name that prediction the university will improve or not. It is not what we want, because if we have a university that is not in our training dataset, the tree will not know the name and won't tell us the answer. So, we could use id to replace the name.

We get rid of World Rank, because the decision tree is weird, the tree tells us just look at the rank, if your 2015 World Rank is greater than some value, then you will not improve, kind of mean and this is not make sense. It means no matter what kind of effort you make, like improve research, if your rank is greater than some specific value, then you will not improve.

Because total score decide rank. So, remove score. Finally, we got a good dataset.

## 2.3 Final dataset





# 3. Modeling -- Shanghai

## 3.1 Fitting 7 Different Classification Models-- Shanghai

--(Random Forest, C45,Support Vector Machines, Rule-Based Classifier, Neural Network, Conditional Inference Tree (Decision Tree), Gradient boosting)

Create fixed sampling scheme (10-folds) so we can compare the models.



The outputs of 7 Different Classification Models— Random Forest, C45,Support Vector Machines, Rule-Based Classifier, Neural Network, Conditional Inference Tree (Decision Tree), Gradient boosting are shown below.

|  |  |
| --- | --- |
| **Random Forest**      From the table above, we can conclude that random forest fit good model because of high accuracy and Kappa value. | From this figure, we can conclude that the most important features are Improved\_pub, award\_X2014, ns\_X2014 and pcp\_X2010 for random forest. |
| **C4.5-like trees**    The fist variable that decision tree thinks it is the information gain, it split first. |  |
| **Rule-Based Classifier** |  |
| **Support Vector Machines** |  |
| **Neural Network** |  |
| **Conditional Inference Tree** |  |
| **Gradient boosting**    For gradient boosting, the best cross-validation iteration is 1320, which won’t cause Under and Overfitting in 1320. | From this figure, we can conclude that the most important features are Improved\_pub > improved\_ns> award\_X2014 > ns\_X2014. |

## 3.2 Discuss the advantages of each model -- Shanghai

|  |  |
| --- | --- |
| Algorithm | Advantages |
| Random Forest | Random Forests train each tree independently, using a random sample of the data. This randomness helps to make the model more robust than a single decision tree, and less likely to overfit on the training data. There are typically two parameters in RF - number of trees and no. of features to be selected at each node. |
| C45 | Simple depth-first construction. Trees are pruned.  Uses Information Gain (improvement in Entropy).  Handling both continuous and discrete attributes (cont. attributes are split at threshold).  Needs entire data to fit in memory (unsuitable for large datasets). |
| Support Vector Machines | High accuracy, nice theoretical guarantees regarding overfitting, and with an appropriate kernel they can work well even if you're data isn't linearly separable in the base feature space. |
| Rule-Based Classifier | As highly expressive as decision trees  Easy to interpret, Easy to generate  Can classify new instances rapidly  Performance comparable to decision trees |
| Neural Network | Efficient and achieve better results for complex applications with a huge amount of data.  Neural networks are quite simple to implement.  No free lunch theorem. Neural networks cannot be retrained. |
| Conditional Inference Tree (Decision Tree) | Inexpensive to construct  Extremely fast at classifying unknown records  Easy to interpret for small-sized trees  Accuracy is comparable to other classification  techniques for many simple data sets |
| Gradient boosting | Boosting is based on weak learners (high bias, low variance).  Improve the stability and often also the accuracy of classifiers.  Reduces variance in the prediction, Reduces overfitting |

## 3.3 Important features-- Shanghai

According to the outputs of seven models, we figure out the most important features in each model are similar, but not same. The most important variables are Improved\_pub and award\_X2014.

Improved\_pub means improvement of number of papers indexed in the Science Citation Index-Expanded and Social Science Citation Index between 2010 and 2014.

Award\_X2014 means number of staff winning Nobel Prizes and fields medals in 2014.

|  |  |
| --- | --- |
| Algorithm | Important features |
| Random Forest | Improved\_pub > award\_X2014 > ns\_X2014 > pcp\_X2010 |
| C45 | Improved\_pub > award\_X2014 > ns\_X2010 > hici\_X2010 |
| Rule-Based Classifier | Improved\_pub > award\_X2014 > pub\_X2014 > pcp\_X2010 |
| Conditional Inference Tree (Decision Tree) | Improved\_pub > alumni\_X2010 > award\_X2014 > improved\_alumni |
| Gradient boosting | Improved\_pub > improved\_ns> award\_X2014 > ns\_X2014 |

## 3.4 Model evaluation--Shanghai

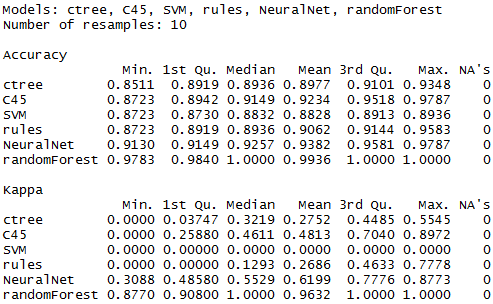
Since we created fixed sampling scheme (10-folds) so we can compare the models now.

Because all models are used 10-folds cross validation. Within the dataset, it will shuffle the data, partition data into 10 disjoint subsets, repeat 10 times, train on 9 partitions, test on the remaining one, average the results. The accuracy of each model is believable.

From this table, we can figure out the accuracy of each models don’t have much different. The range is from 0.85 to 0.978. However, Kappa are significantly different. Since kappa compare the accuracy of the classifier with a random classifier. Also, only Random Forest had a relative good Kappa value, which is 0.877. That means this model don’t have imbalance problem.

In addition, the most accurate algorithm is Random Forest, which is 0.978.

So, we conclude that Random Forest is the most efficient algorithm for shanghai dataset.



## 3.5 Feature selection and simple model--Shanghai

Decision trees implicitly select features for splitting, but we can also select features manually.

Also, we cannot collect enough data sometimes or we want to collect the most important variables first. We could use feature selection to get 5 best features and build a simple model.

From the outputs of this simple model, we can know the 5 best features are “award\_X2010", "improved\_pub","award\_X2014","pcp\_X2014" and "alumni\_X2010".

|  |  |
| --- | --- |
|  |  |

|  |
| --- |
|  |
|  |

# 4. Modeling -- Timesdata

## 4.1 Fitting 7 Different Classification Models-- Timesdata

— (Random Forest, C45,Support Vector Machines, Rule-Based Classifier, Neural Network, Conditional Inference Tree (Decision Tree), Gradient boosting)

Create fixed sampling scheme (10-folds) so we can compare the models.



The outputs of 7 Different Classification Models— Random Forest, C45,Support Vector Machines, Rule-Based Classifier, Neural Network, Conditional Inference Tree (Decision Tree), Gradient boosting are shown below.

|  |  |
| --- | --- |
| **Random Forest**    From the table above, we can conclude that random forest fit good model because of high accuracy and Kappa value. | From this figure, we can conclude that the most important features are Improved\_teaching > Improved\_reasearch>international\_students\_X2013 > num\_students\_X2013 |
| **C4.5-like trees** |  |
| **Rule-Based Classifier** |  |
| **Support Vector Machines** |  |
| **Neural Network** |  |
| **Conditional Inference Tree** |  |
| **Gradient boosting**    For gradient boosting, the best cross-validation iteration is 280, which won’t cause Under and Overfitting in 280. | From this figure, we can conclude that the most important features are Improved\_teaching > international\_students\_X2013 > Improved\_reasearch> improved\_income |

## 4.2 Important features-- Timesdata

According to the outputs of seven models, we figured out the most important features in each model are similar, but not same. The most important variables are Improved\_teaching and Improved\_reasearch.

Improved\_teaching means university score for teaching.

Improved\_reasearch means university score for research.

|  |  |
| --- | --- |
| Algorithm | Important features |
| Random Forest | Improved\_teaching > Improved\_reasearch> international\_students\_X2013 > num\_students\_X2013 |
| C45 | Improved\_teaching > Improved\_citations > international\_students\_X2013 |
| Rule-Based Classifier | international\_students\_X2013 > Improved\_teaching > Improved\_citations > improved\_income |
| Conditional Inference Tree (Decision Tree) | Improved\_reasearch> Improved\_teaching > improved\_international > Students\_staff\_ratio\_X2013 |
| Gradient boosting | Improved\_teaching > international\_students\_X2013 > Improved\_reasearch> improved\_income |

## 4.3 Model evaluation-- Timesdata

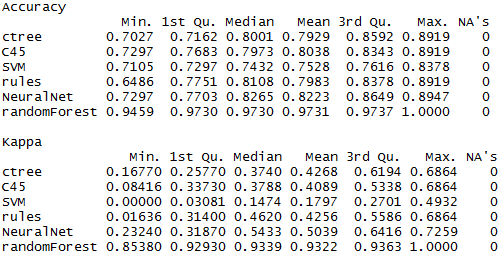
Since we created fixed sampling scheme (10-folds) so we can compare the models now.

Because all models are used 10-folds cross validation. Within the dataset, it will shuffle the data, partition data into 10 disjoint subsets, repeat 10 times, train on 9 partitions, test on the remaining one, average the results. The accuracy of each model is believable.

From this table, we can figure out the accuracy of each models are different. The range is from 0.70 to 0.94. However, Kappa are significantly different. Since kappa compare the accuracy of the classifier with a random classifier. Also, only Random Forest had a relative good Kappa value, which is 0.85. That means this model don’t have imbalance problem.

In addition, the most accurate algorithm is Random Forest, which is 0.9459.

So, we conclude that Random Forest is the most efficient algorithm for timesdataset.



## 4.4 Feature selection and simple model-- Timesdata

Decision trees implicitly select features for splitting, but we can also select features manually.

Also, we cannot collect enough data sometimes or we want to collect the most important variables first. We could use feature selection to get 5 best features and build a simple model.

> subset

[1] "improved\_research" "improved\_teaching"

[3] "international\_students\_X2013" "international\_students\_X2014"

[5] "teaching\_X2013"

From the outputs of this simple model, we can know the 5 best features are "improved\_research", "improved\_teaching", "international\_students\_X2013", "international\_students\_X2014" and "teaching\_X2013".

|  |
| --- |
|  |

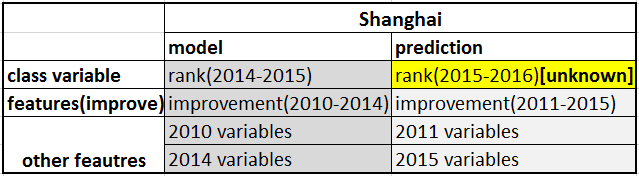
|  |
| --- |
|  |
|  |

# 5. Evaluation and Deployment - Shanghai & Timesdata

## 5.1 Prediction – Shanghai

After getting relative accurate models, we would like to make a prediction based on this model.

So, we made a test data based on the table below. We use features in 2011 and 2015 to fit the data and get the prediction whether the university rank will improve or not in 2016, which is unknown.



|  |
| --- |
|  |
|  |

The output of “f.predict.test” is part of prediction that we want to get in 2016. Next, we would like to choose an university at random to see whether the rank will improve or not.

|  |  |
| --- | --- |
|  |  |

First, we want to predict the rank of Southern Methodist University. Unfortunately, there are no SMU in the dataset.





Then we choose University of Southern California. Using which statement, we could know USC in the position of 508. Correspondingly, we search the result of 508, which is true. It means the rank of USC will improve in 2016.





## 5.2 Prediction – Timesdata

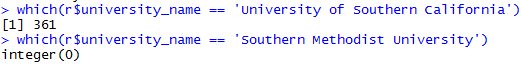
After getting relative accurate models, we would like to make a prediction based on this model.

So, we made a test data based on the table below. We use features in 2014 and 2015 to fit the data and get the prediction whether the university rank will improve or not in 2016, which is unknown.

|  |  |
| --- | --- |
|  |  |

First, we want to predict the rank of Southern Methodist University. Unfortunately, there are no SMU in the dataset.

Then we choose University of Southern California. Using which statement, we could know USC in the position of 361. Correspondingly, we search the result of 361, which is true. It means the rank of USC will not improve in 2016.



## 5.3 Conclusion

Generally speaking, stake holders, like students and university administrators could use this model to predict any university rank they would like to know in a specific world ranking dataset, which is meaningful. For instance, rank of USC will improve in 2016 of shanghai dataset while rank of USC will not improve in 2016 of timesdata. The reason is the criteria for the ranking is different.

Also, university administrators could know the most important features for ranking of shanghai dataset are Improved\_pub and award\_X2014. Therefore, university administrators could improve in those 2 aspects first.

On the other hand, university administrators could know the most important features for ranking of timesdataset are Improved\_teaching and Improved\_research.

Thus, university administrators could improve in teaching and research first to improve rank of timesdata.

To improve number of papers indexed in the Science Citation Index-Expanded and Social Science Citation Index, number of staff winning Nobel Prizes and fields medals, teaching score and research score. The university could increase funds for research, adjusting curriculum arrangement etc.

# 6. Acknowledgment

Specially thanks for [Dr. Michael Hahsler](http://michael.hahsler.net/)’s great help and mentoring on this project

# 7. References

Source:

1. ARWU (Shanghai Ranking) Methodology:

<http://www.shanghairanking.com/ARWUMethodology2016.Html>

1. WUR (Times Ranking) Methodology:

https://www.timeshighereducation.com/news/rankingmethodology2016

1. Dataset

https://www.kaggle.com/mylesoneill/world-university-rankings

1. R code

<http://michael.hahsler.net/SMU/EMIS7332/>

1. R code

<https://emudrak.github.io/2015-01-15-cornell/lessons/R/03-data-frames.html>

1. R package

<http://stackoverflow.com/questions/17376939/problems-when-trying-to-load-a-package-in-r-due-to-rjava>