Data Mining Project

Rutgers University

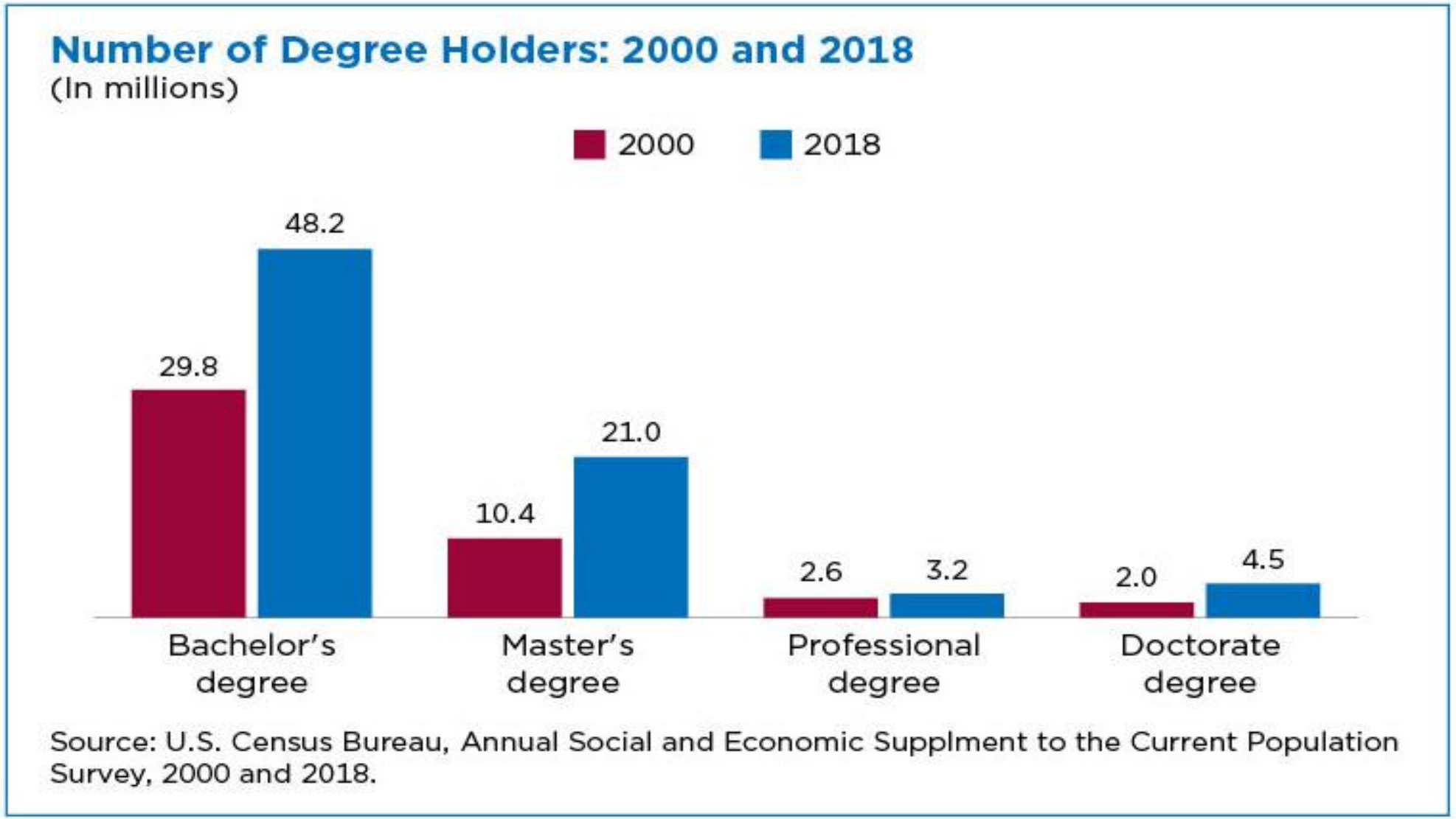
Data Mining -- 2019 Spring

Name: Weijun Zhu

Ji Wu

Dinglun Tan

Project Overview



https://www.census.gov/library/stories/2019/02/number-of-people-with-masters-and-phd-degrees-double-since-2000.html

Doctorate in Ivy?

- How does cumulative GPA impact your admission?
- which is evaluated the most:
  - GRE? TOEFL? GPA? Recommendation Letter? Research? Statement of Purpose?
- A strong recommendation letter = Big chance ?

Project Overview

- We use the R language to handle the Admission Dataset from Kaggle.
- Project is about student's advantage in their applying for graduate school.
- CGPA, TOEFL , GRE score and others attributes may affect the Chance of Admit.

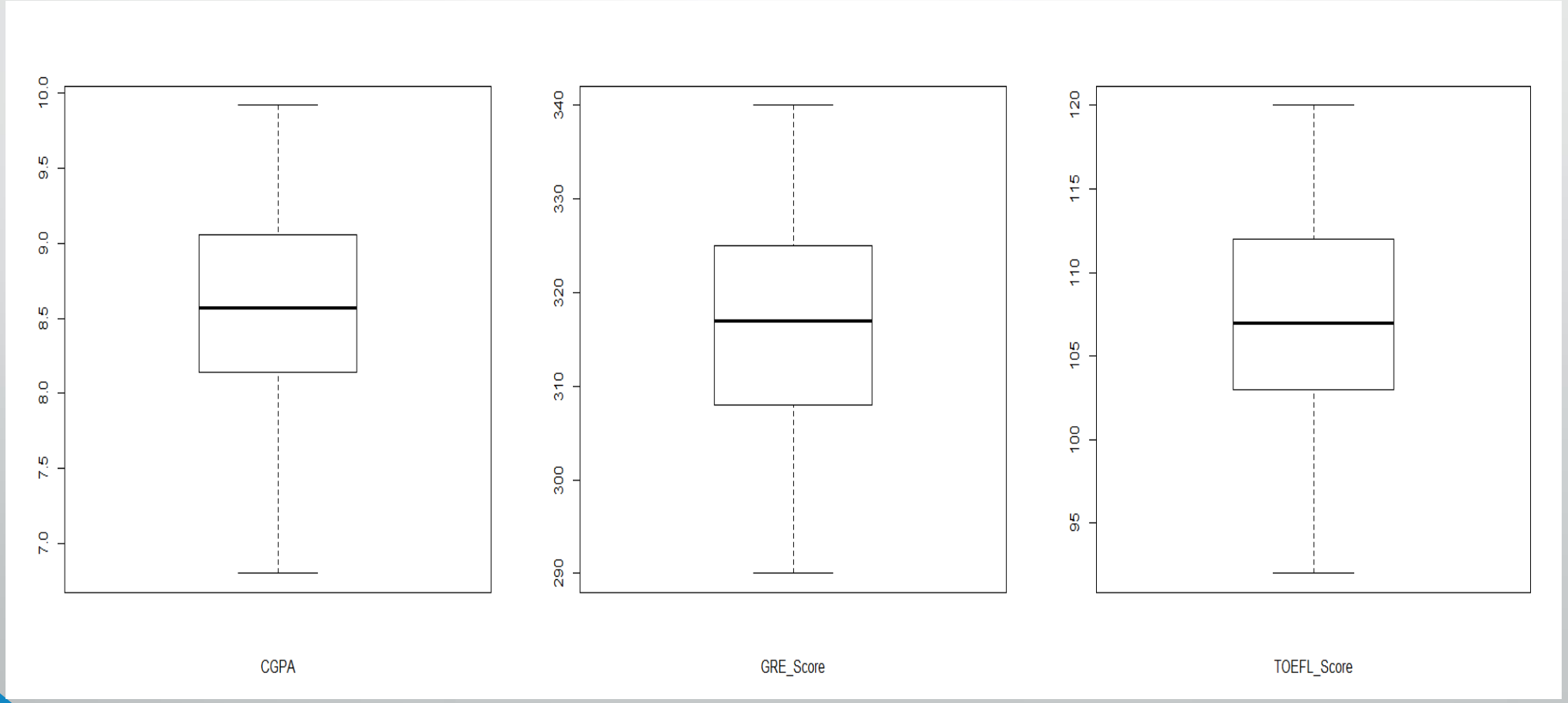
Preview of the Database

- We have data from almost 1000 different data.
- There are nine attributes in our datasets
- Introduction to each attributes:
  - CPGA: Cumulative GPA
  - SOP: Statement of Purpose (Rating:1-5, 5 means the best)
  - LOR: Letter of Recommendation (Rating: 1-5, 5 means the best)
  - Research: 0 --- No research ; 1--- Do research

	A	B	C	D	E	F	G	H	I	J
1	Serial No.	GRE Score	TOEFL Sco	University	SOP	LOR	CGPA	Research	Chance of Admit	
2	1	337	118	4	4.5	4.5	9.65	1	0.92	
3	2	324	107	4	4	4.5	8.87	1	0.76	
4	3	316	104	3	3	3.5	8	1	0.72	
5	4	322	110	3	3.5	2.5	8.67	1	0.8	
6	5	314	103	2	2	3	8.21	0	0.65	
7	6	330	115	5	4.5	3	9.34	1	0.9	
8	7	321	109	3	3	4	8.2	1	0.75	
9	8	308	101	2	3	4	7.9	0	0.68	
10	9	302	102	1	2	1.5	8	0	0.5	
11	10	323	108	3	3.5	3	8.6	0	0.45	
12	11	325	106	3	3.5	4	8.4	1	0.52	
13	12	327	111	4	4	4.5	9	1	0.84	
14	13	328	112	4	4	4.5	9.1	1	0.78	
15	14	307	109	3	4	3	8	1	0.62	
16	15	311	104	3	3.5	2	8.2	1	0.61	
17	16	314	105	3	3.5	2.5	8.3	0	0.54	
18	17	317	107	3	4	3	8.7	0	0.66	
19	18	319	106	3	4	3	8	1	0.65	
20	19	318	110	3	4	3	8.8	0	0.63	
21	20	303	102	3	3.5	3	8.5	0	0.62	
22	21	312	107	3	3	2	7.9	1	0.64	
23	22	325	114	4	3	2	8.4	0	0.7	
24	23	328	116	5	5	5	9.5	1	0.94	
25	24	334	119	5	5	4.5	9.7	1	0.95	
26	25	336	119	5	4	3.5	9.8	1	0.97	
27	26	340	120	5	4.5	4.5	9.6	1	0.94	
28	27	322	109	5	4.5	3.5	8.8	0	0.76	
29	28	298	98	2	1.5	2.5	7.5	1	0.44	
30	29	295	93	1	2	2	7.2	0	0.46	
31	30	310	99	2	1.5	2	7.3	0	0.54	
32	31	300	97	2	3	3	8.1	1	0.65	
33	32	327	103	3	4	4	8.3	1	0.74	
34	33	338	118	4	3	4.5	9.4	1	0.91	
35	34	340	114	5	4	4	9.6	1	0.9	
36	35	331	112	5	4	5	9.8	1	0.94	
37	36	320	110	5	5	5	9.2	1	0.88	

Data Preprocessing

- Data of each attribute distributes good
- No outliers & No noises
- Distribute centralized
- Don’t need to normalization and zero-centered

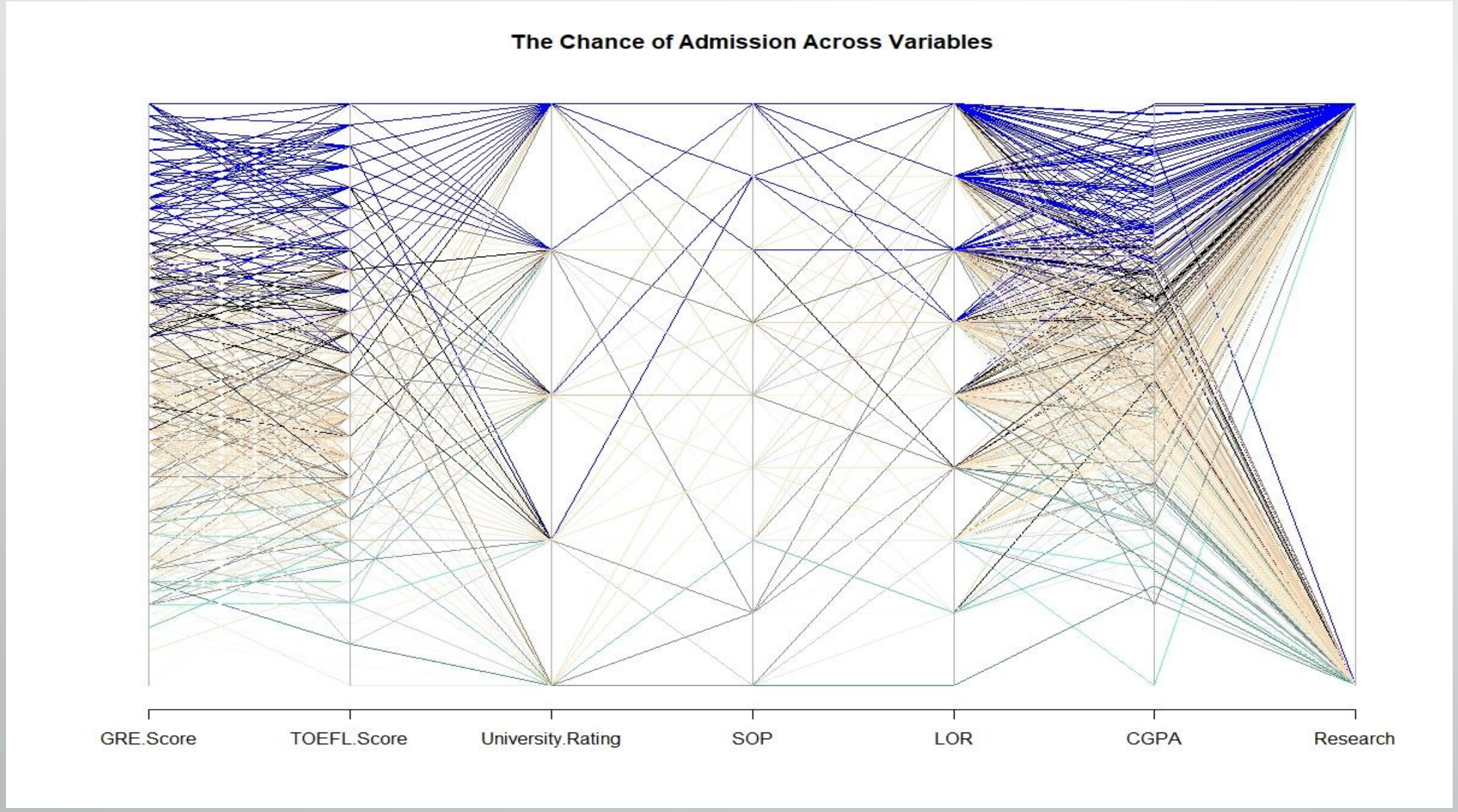


Data Preprocessing

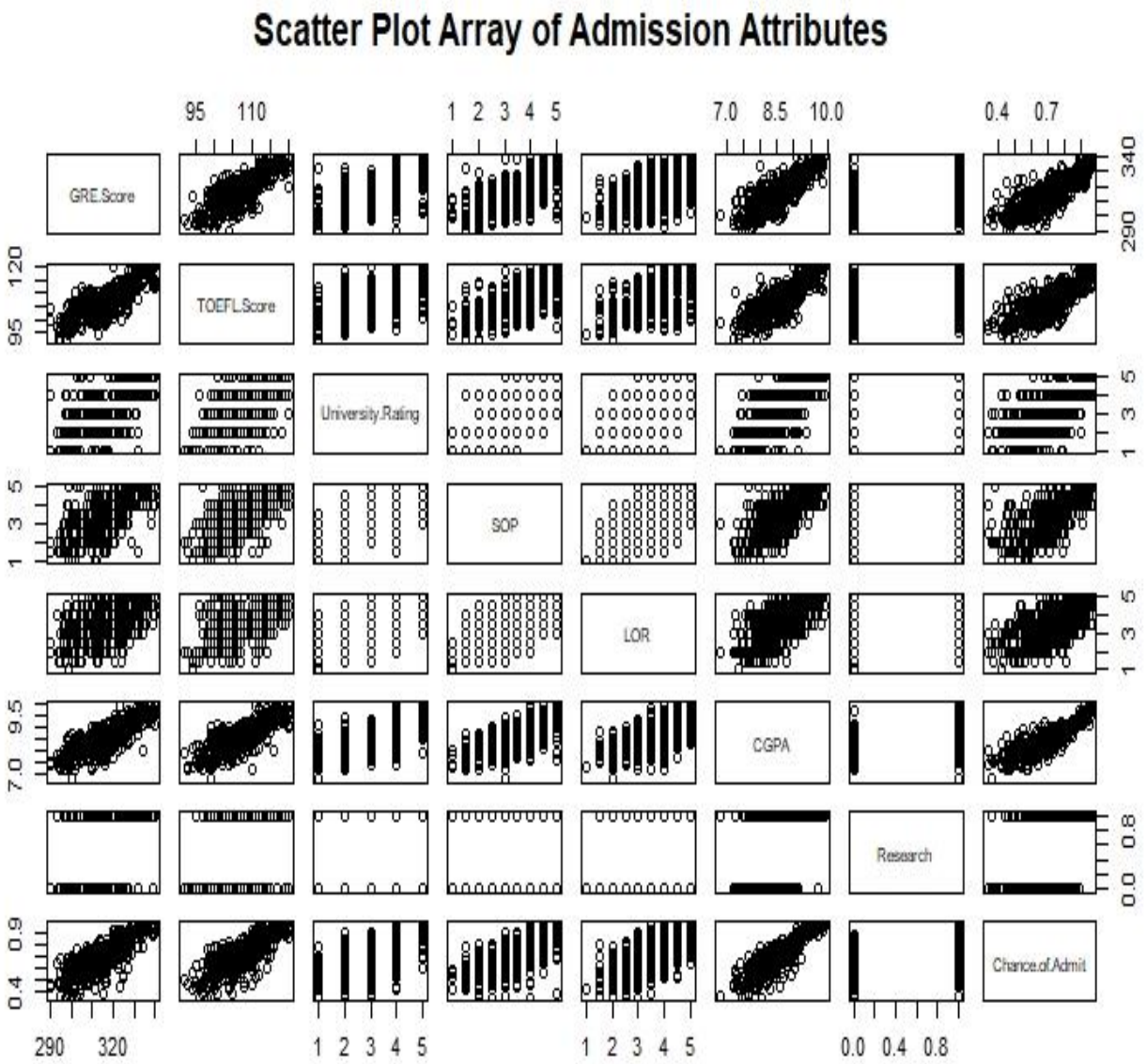
- Delete the attributes we don’t need
- Delete the attribute: Serial.No
- Handling the NA value the dataset : Omit the NA value

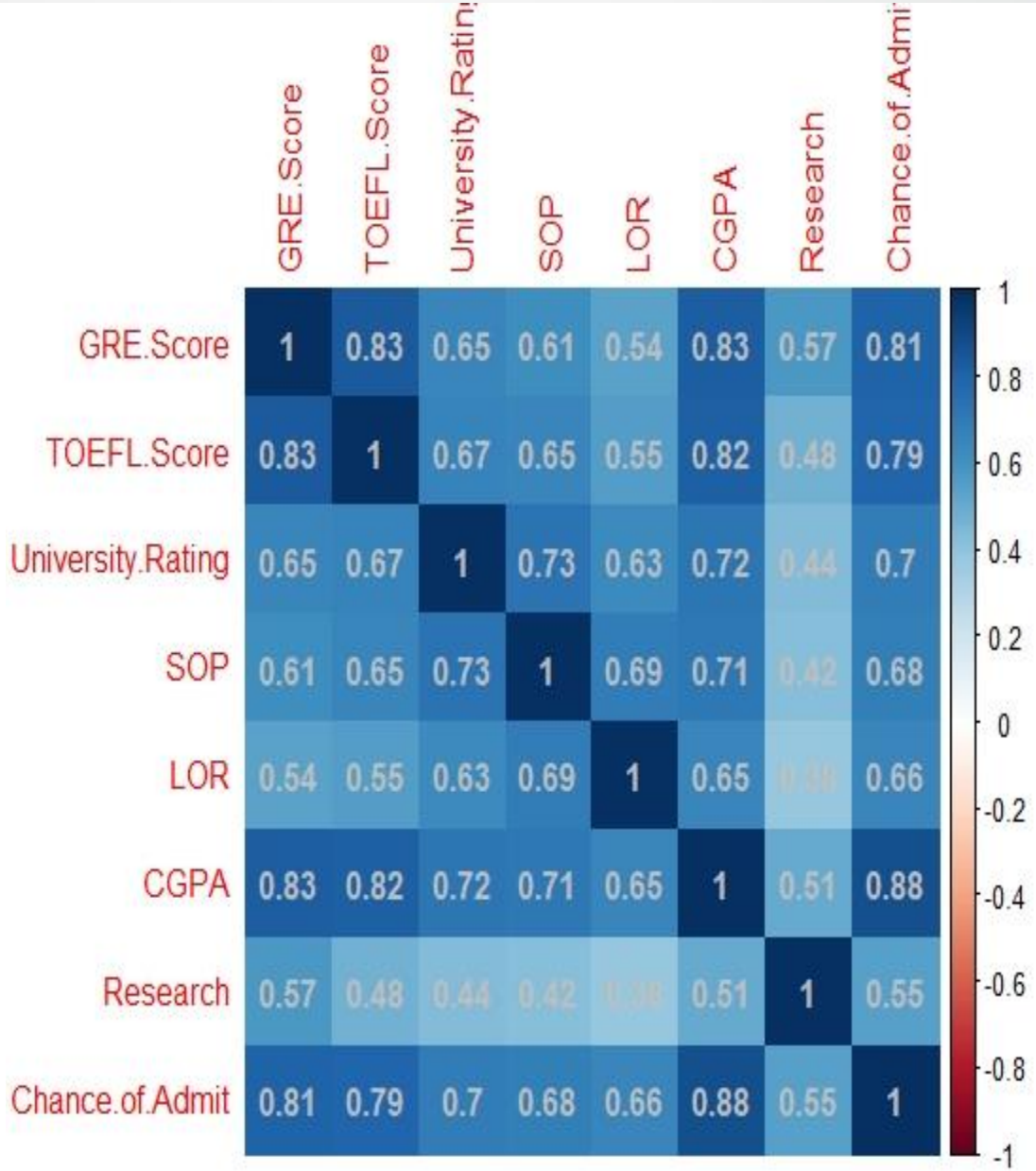
```
### Delete the Attribute
Admission <- dplyr::select(Admission, -Serial.No.)
### Handle the NA value
Admission <- na.omit(Admission)
```

Parallel Coordinates plot



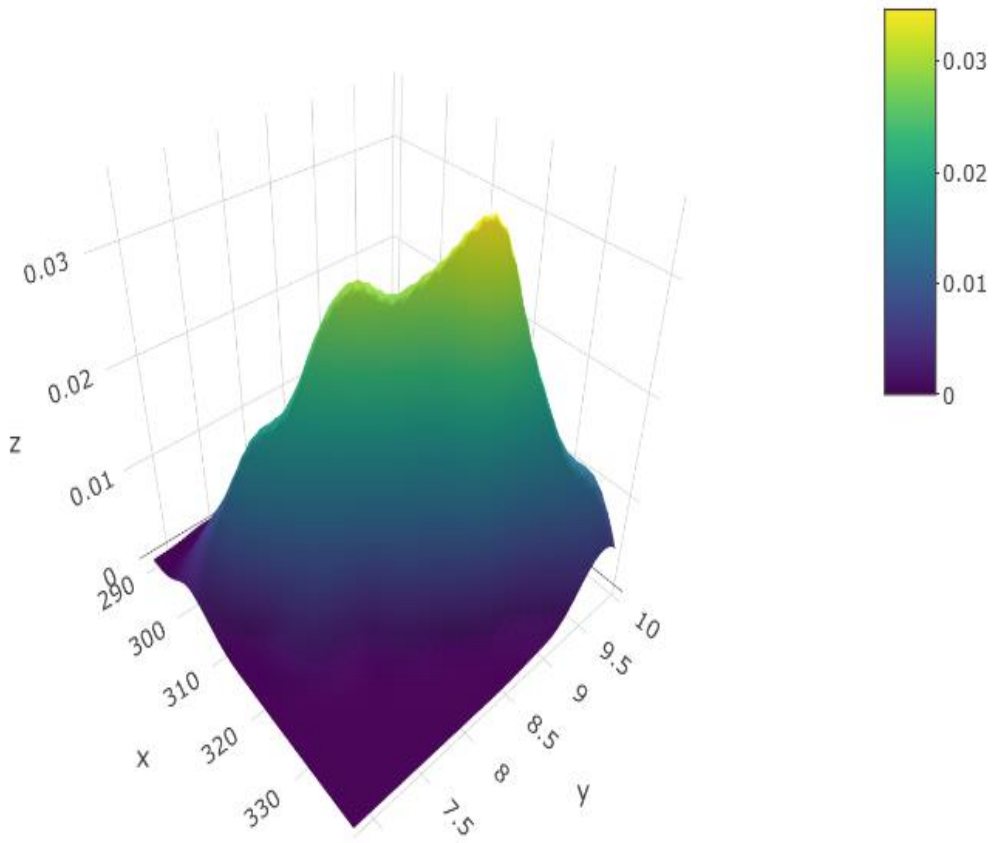
Correlation of Each Attribute



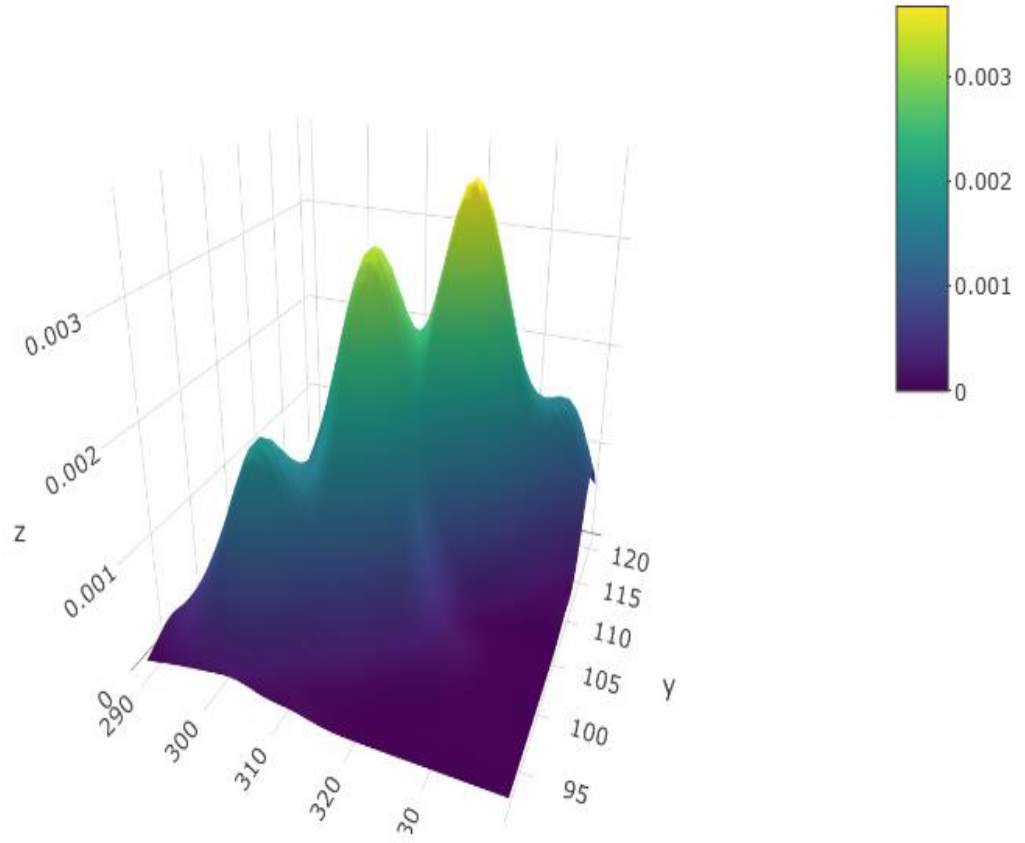


3D-Graph

GRE & CGPA



GRE & TOEFL



How to Define the Classifier?

- The Chance of Admit is the classifier
- We need to find a suitable value to separate the classifier of Chance of Admit by two parts.
- We use the Linear Regression and boxplot to find a suitable value to separate the Chance of Admission.
- AIC Index to eliminate some attributes

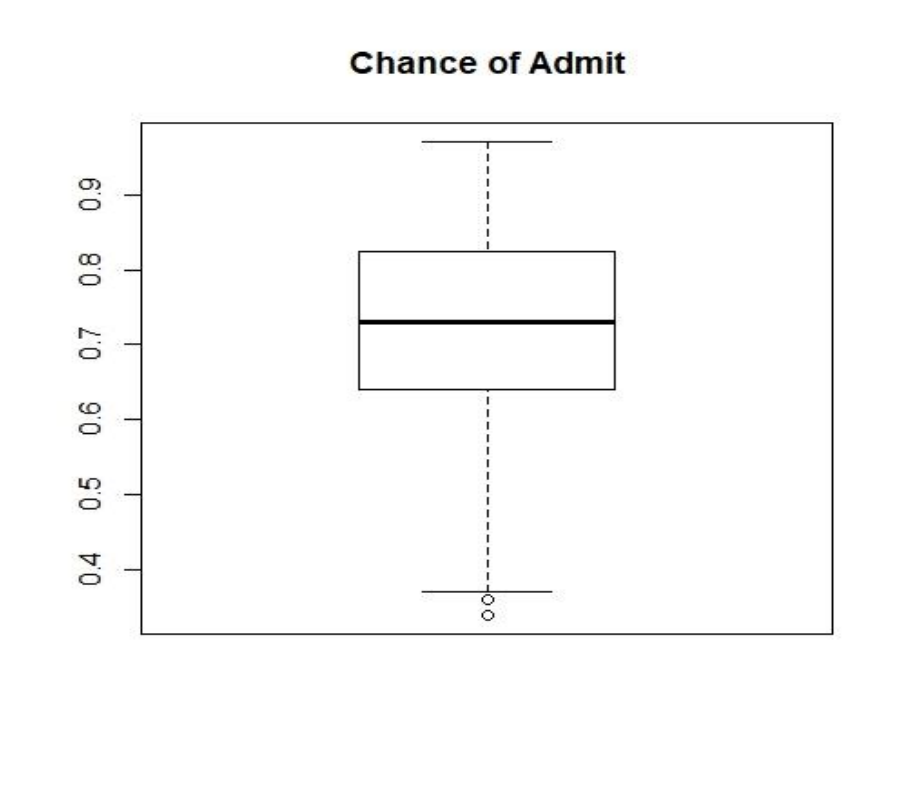
Linear Regression & AIC to determine Classifier?

```
Start:  AIC=-5013.14
Chance.of.Admit ~ GRE.Score + TOEFL.Score + University.Rating +
    SOP + LOR + CGPA + Research

                    Df Sum of Sq    RSS     AIC
- SOP                1   0.00004 3.3685 -5015.1
<none>                           3.3685 -5013.1
- University.Rating  1   0.01506 3.3835 -5011.1
- TOEFL.Score        1   0.06491 3.4334 -4998.0
- GRE.Score          1   0.08431 3.4528 -4992.9
- Research           1   0.08671 3.4552 -4992.3
- LOR                1   0.12317 3.4917 -4982.8
- CGPA               1   0.92418 4.2927 -4796.9

Step:  AIC=-5015.13
Chance.of.Admit ~ GRE.Score + TOEFL.Score + University.Rating +
    LOR + CGPA + Research

                    Df Sum of Sq    RSS     AIC
<none>                           3.3685 -5015.1
- University.Rating  1   0.01644 3.3850 -5012.7
- TOEFL.Score        1   0.06531 3.4338 -4999.8
- GRE.Score          1   0.08472 3.4532 -4994.8
- Research           1   0.08671 3.4552 -4994.3
- LOR                1   0.13758 3.5061 -4981.1
- CGPA               1   0.94836 4.3169 -4793.9
```

```
Call:
lm(formula = Chance.of.Admit ~ GRE.Score + TOEFL.Score + University.Rating +
    LOR + CGPA + Research, data = Admission)

Residuals:
      Min        1Q    Median        3Q       Max
-0.265193 -0.022474  0.009969  0.034720  0.157668

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -1.2682451  0.0793138 -15.990  < 2e-16 ***
GRE.Score          0.0018145  0.0003829   4.739 2.50e-06 ***
TOEFL.Score        0.0028115  0.0006757   4.161 3.48e-05 ***
University.Rating  0.0058151  0.0027858   2.087   0.0371 *
LOR                0.0188039  0.0031136   6.039 2.27e-09 ***
CGPA               0.1185885  0.0074791  15.856  < 2e-16 ***
Research           0.0242723  0.0050625   4.795 1.91e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06142 on 893 degrees of freedom
Multiple R-squared:  0.8134,    Adjusted R-squared:  0.8122
F-statistic:   649 on 6 and 893 DF,  p-value: < 2.2e-16
```

Chance of Admit = -1.2682451 +0.0018145*GRE+0.0028115*TOEFL+0.0058151*University_Rating+0.0188039*LOR+0.1185885*CGPA+0.0242723*Research

The Value to Separate the Classifier



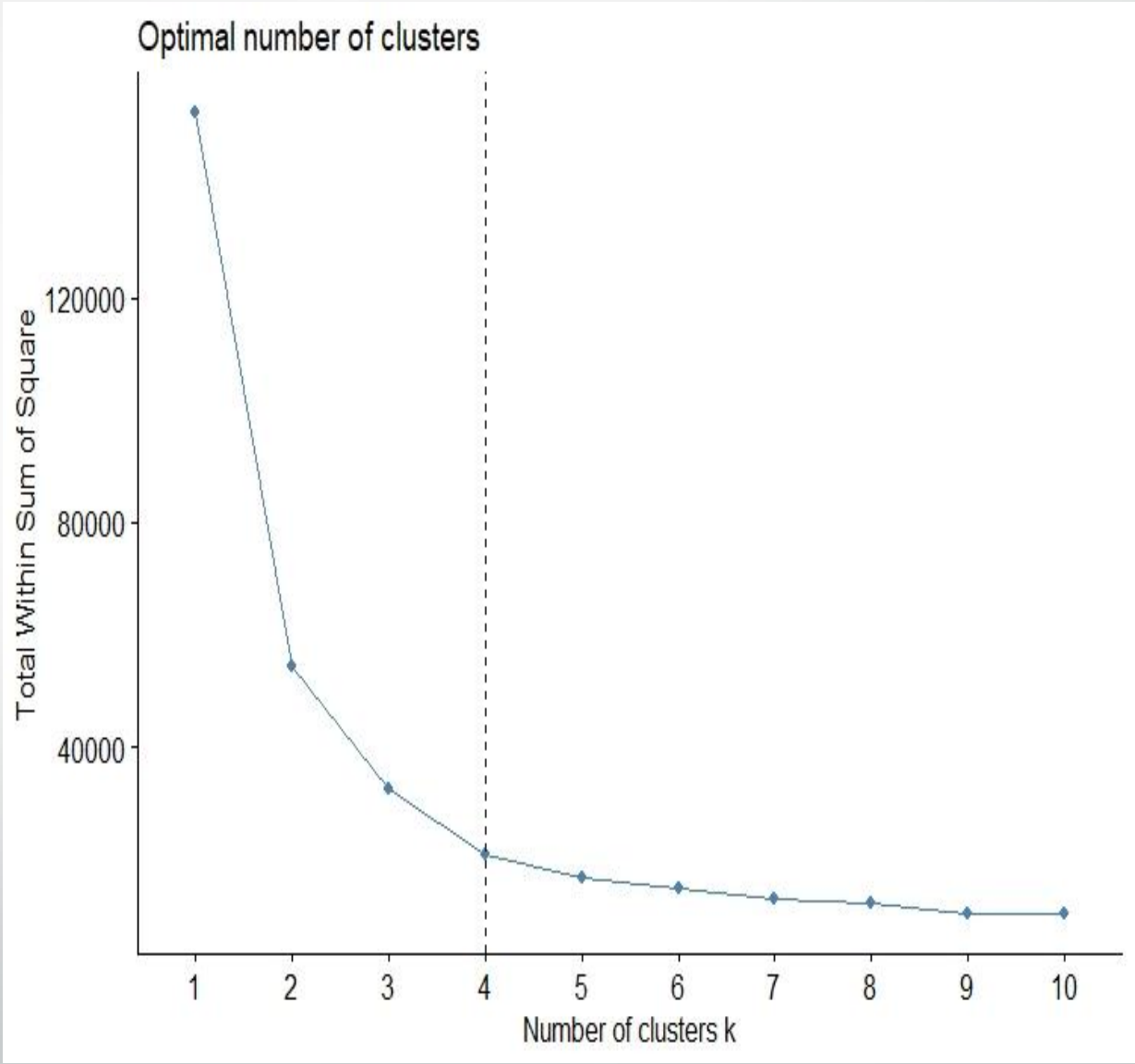
- We compute the mean of the attribute
- Let these means of values into the Linear Regression that we computed.
- Then we get the Chance of Admit=0.72

Classification:

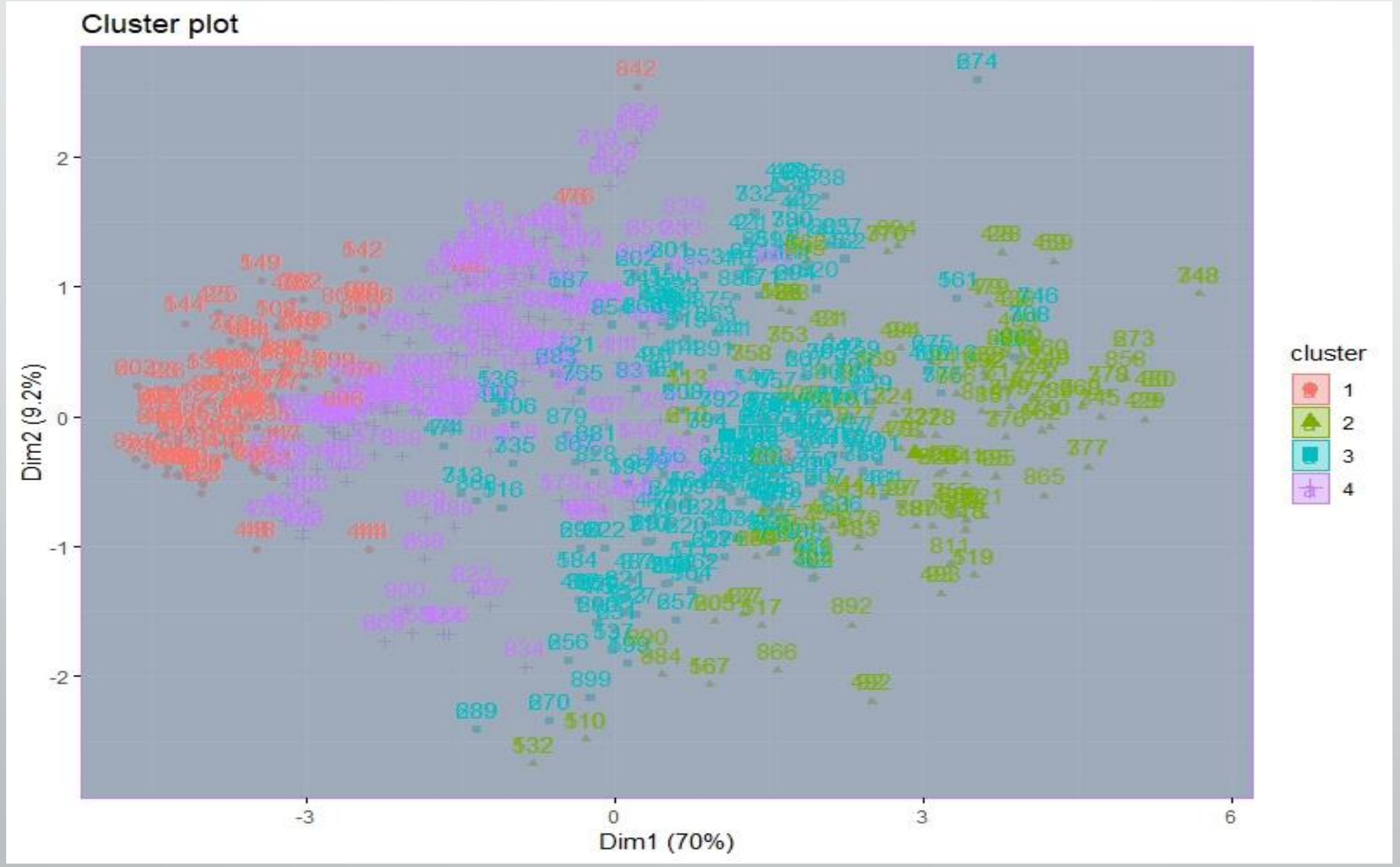
- The Chance of Admit > 0.72 --- Big Chance
- The Chance of Admit <= 0.72 --- Small Chance

K-Means

- Optimal number of clusters
  K = 4
- Choose the point which the slop of the line tends to be steady.
- That point is the optimal number of clusters.



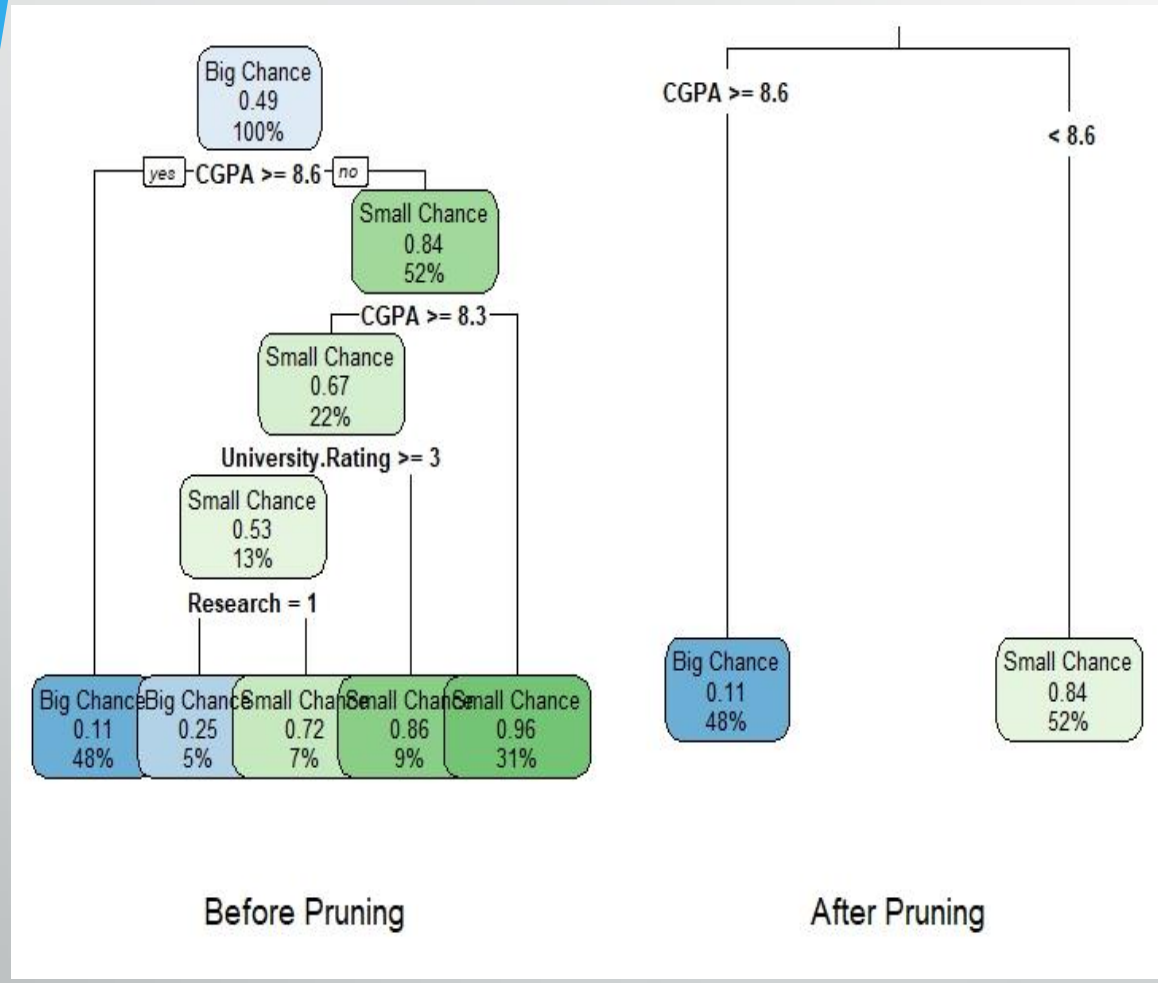
K-Means

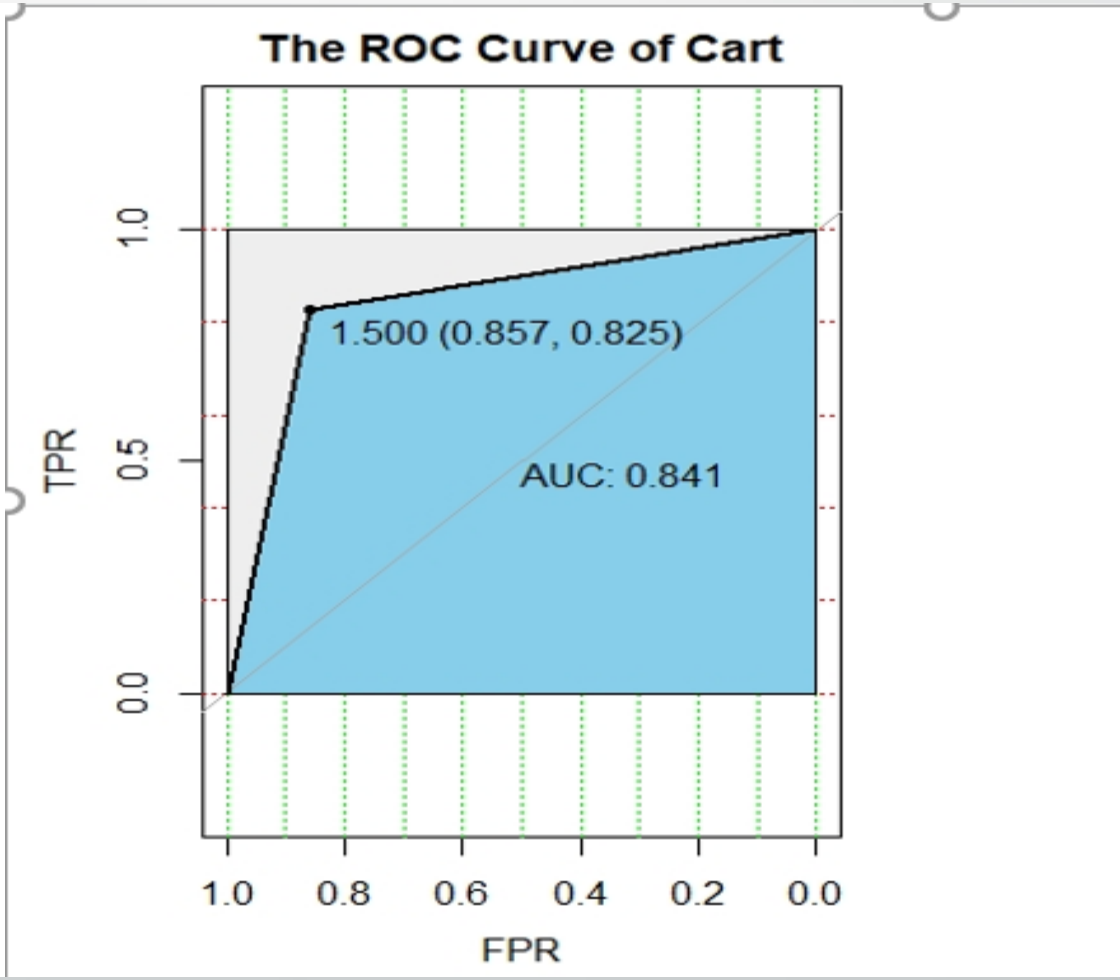


Data Mining

- Cart Algorithm
- C4.5 Algorithm
- Support Vector Machine(SVM) Algorithm
- K-NN Algorithm
- Adaboost Algorithm
- RIPPER Algorithm

CART Algorithm (Prune by Gini Index)





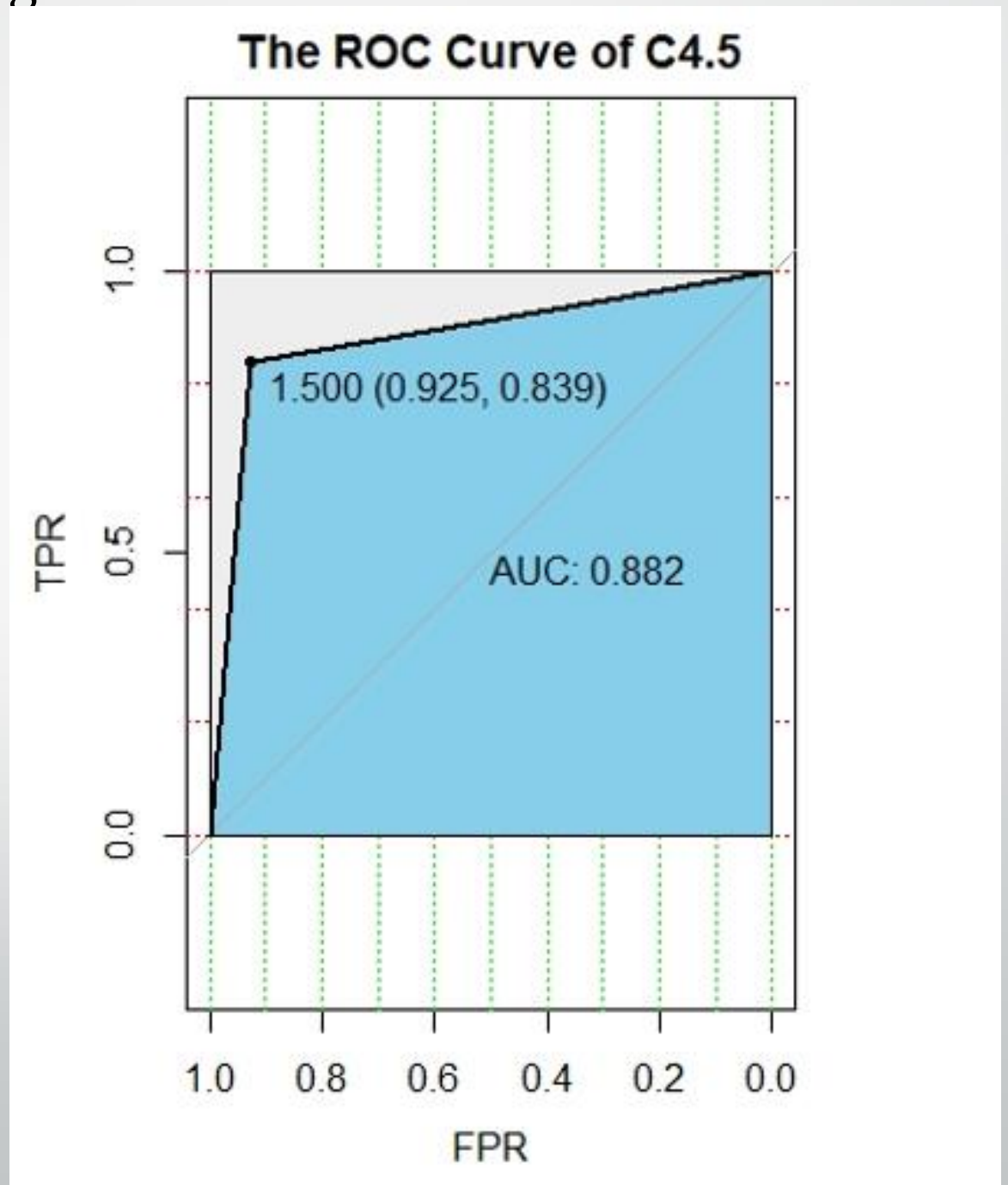
CART Algorithm—Confusion Matrix

	Big Chance	Small Chance
Big Chance	114	19
Small	24	113

- Accuracy = 0.841
- Recall = 0.857
- Precision = 0.825
- F-Measure = 0.8412

C4.5 Algorithm

```
CGPA <= 8.64
|   CGPA <= 8.1: Small Chance (143.0/1.0)
|   CGPA > 8.1
|   |   GRE.Score <= 319
|   |   |   GRE.Score <= 305: Small Chance (35.0)
|   |   |   GRE.Score > 305
|   |   |   |   SOP <= 2.5
|   |   |   |   |   University.Rating <= 2: Small Chance (38.0)
|   |   |   |   |   University.Rating > 2
|   |   |   |   |   |   Research <= 0
|   |   |   |   |   |   |   CGPA <= 8.42: Small Chance (10.0)
|   |   |   |   |   |   |   CGPA > 8.42: Big Chance (2.0)
|   |   |   |   |   |   Research > 0: Big Chance (3.0)
|   |   |   |   SOP > 2.5
|   |   |   |   |   Research <= 0
|   |   |   |   |   |   LOR <= 3.5
|   |   |   |   |   |   |   LOR <= 2.5
|   |   |   |   |   |   |   |   University.Rating <= 2: Big Chance (5.0/1.0)
|   |   |   |   |   |   |   |   University.Rating > 2: Small Chance (4.0)
|   |   |   |   |   |   |   LOR > 2.5: Small Chance (26.0/2.0)
|   |   |   |   |   |   LOR > 3.5
|   |   |   |   |   |   |   University.Rating <= 2: Small Chance (3.0)
|   |   |   |   |   |   |   University.Rating > 2: Big Chance (11.0/2.0)
|   |   |   |   |   Research > 0
|   |   |   |   |   |   TOEFL.Score <= 104: Small Chance (5.0)
|   |   |   |   |   |   TOEFL.Score > 104
|   |   |   |   |   |   |   SOP <= 3: Big Chance (9.0/1.0)
|   |   |   |   |   |   |   SOP > 3
|   |   |   |   |   |   |   |   TOEFL.Score <= 106: Big Chance (7.0/1.0)
|   |   |   |   |   |   |   |   TOEFL.Score > 106: Small Chance (6.0)
|   |   GRE.Score > 319
|   |   |   Research <= 0: Small Chance (6.0/1.0)
|   |   |   Research > 0
|   |   |   |   University.Rating <= 2
|   |   |   |   |   GRE.Score <= 321: Big Chance (2.0)
|   |   |   |   |   GRE.Score > 321: Small Chance (3.0)
|   |   |   |   University.Rating > 2: Big Chance (23.0/3.0)
CGPA > 8.64
|   Research <= 0
|   |   CGPA <= 9.01
|   |   |   SOP <= 4
|   |   |   |   SOP <= 2.5: Small Chance (6.0)
|   |   |   |   SOP > 2.5
|   |   |   |   |   LOR <= 3.5
|   |   |   |   |   |   TOEFL.Score <= 109
|   |   |   |   |   |   |   CGPA <= 8.74
|   |   |   |   |   |   |   |   CGPA <= 8.68: Big Chance (2.0)
|   |   |   |   |   |   |   |   CGPA > 8.68: Small Chance (3.0)
|   |   |   |   |   |   |   CGPA > 8.74: Big Chance (5.0)
|   |   |   |   |   |   TOEFL.Score > 109: Small Chance (10.0)
|   |   |   |   |   LOR > 3.5
|   |   |   |   |   |   GRE.Score <= 312: Small Chance (2.0)
|   |   |   |   |   |   GRE.Score > 312: Big Chance (9.0/1.0)
|   |   |   SOP > 4: Big Chance (6.0)
|   |   CGPA > 9.01: Big Chance (16.0)
|   Research > 0: Big Chance (230.0/6.0)

Number of Leaves  :     29

Size of the tree :      57
```

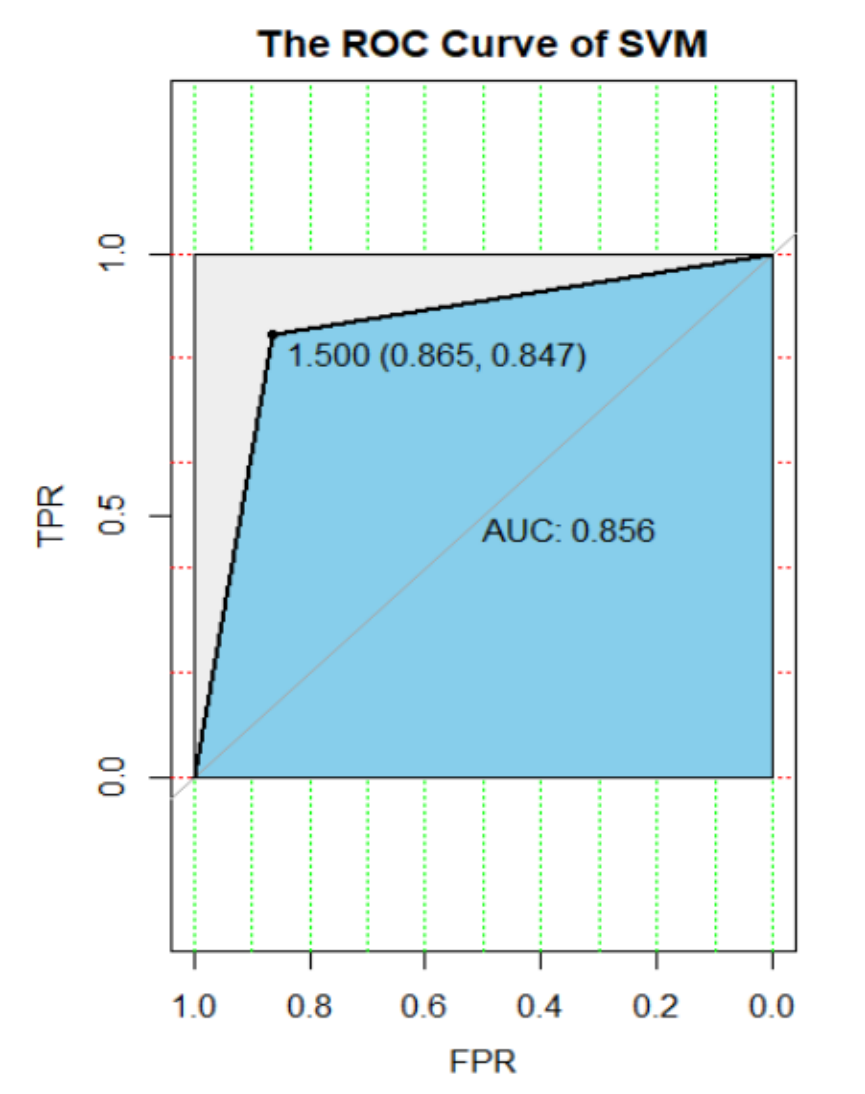


C4.5 --- Confusion Matrix

	Big Chance	Small Chance
Big Chance	123	10
Small Chance	22	115

- Accuracy = 0.882
- Recall = 0.925
- Precision = 0.839
- F-Measure = 0.885

SVM Algorithm



	Big Chance	Small Chance
Big Chance	115	18
Small Chance	21	116

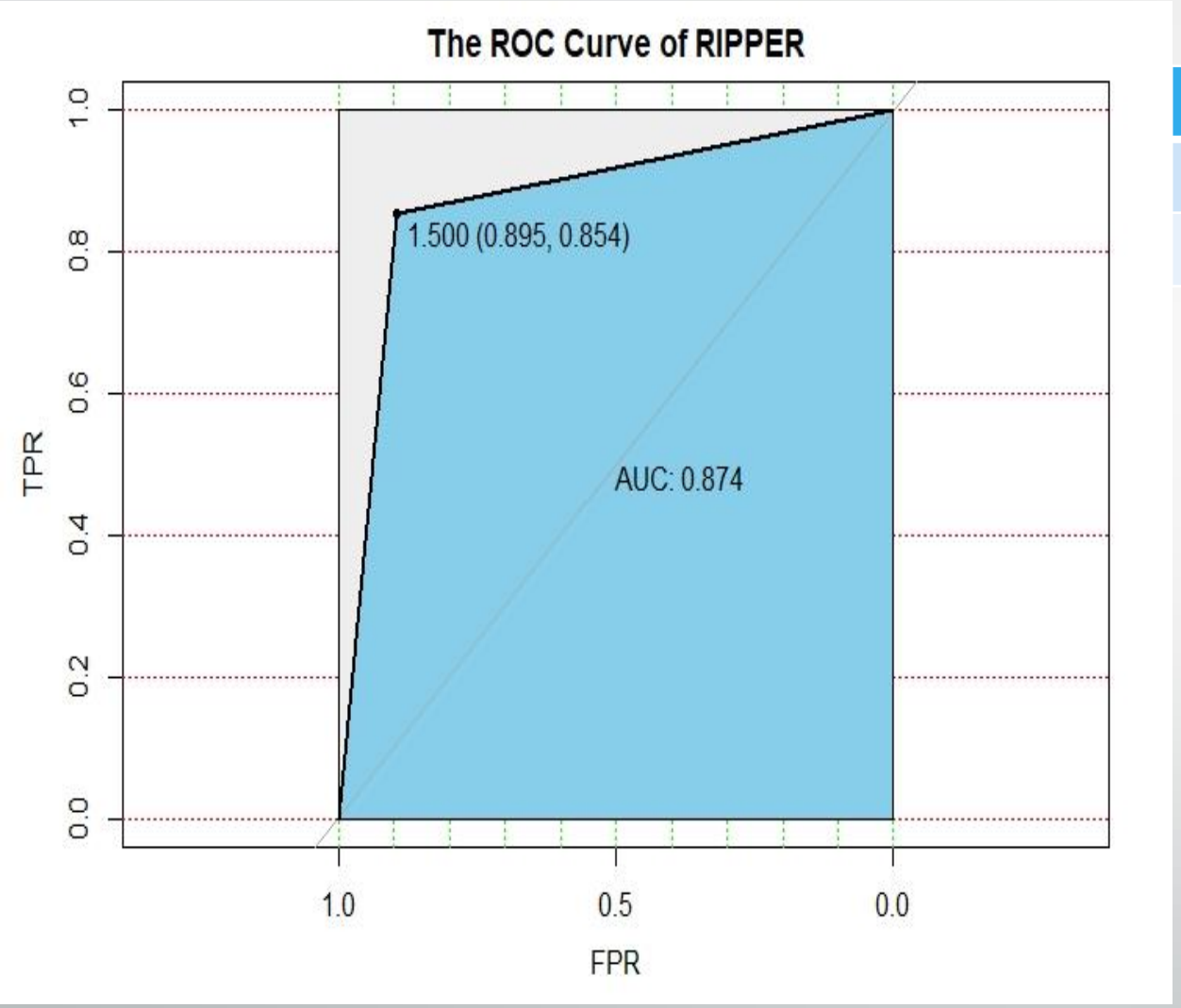
- Accuracy = 0.856
- Recall = 0.865
- Precision = 0.847
- F-Measure = 0.855

K-NN

	Big Chance	Small Chance
Big Chance	122	11
Small Chance	20	117

- Accuracy = 0.885
- Recall = 0.917
- Precision = 0.859
- F-Measure = 0.887

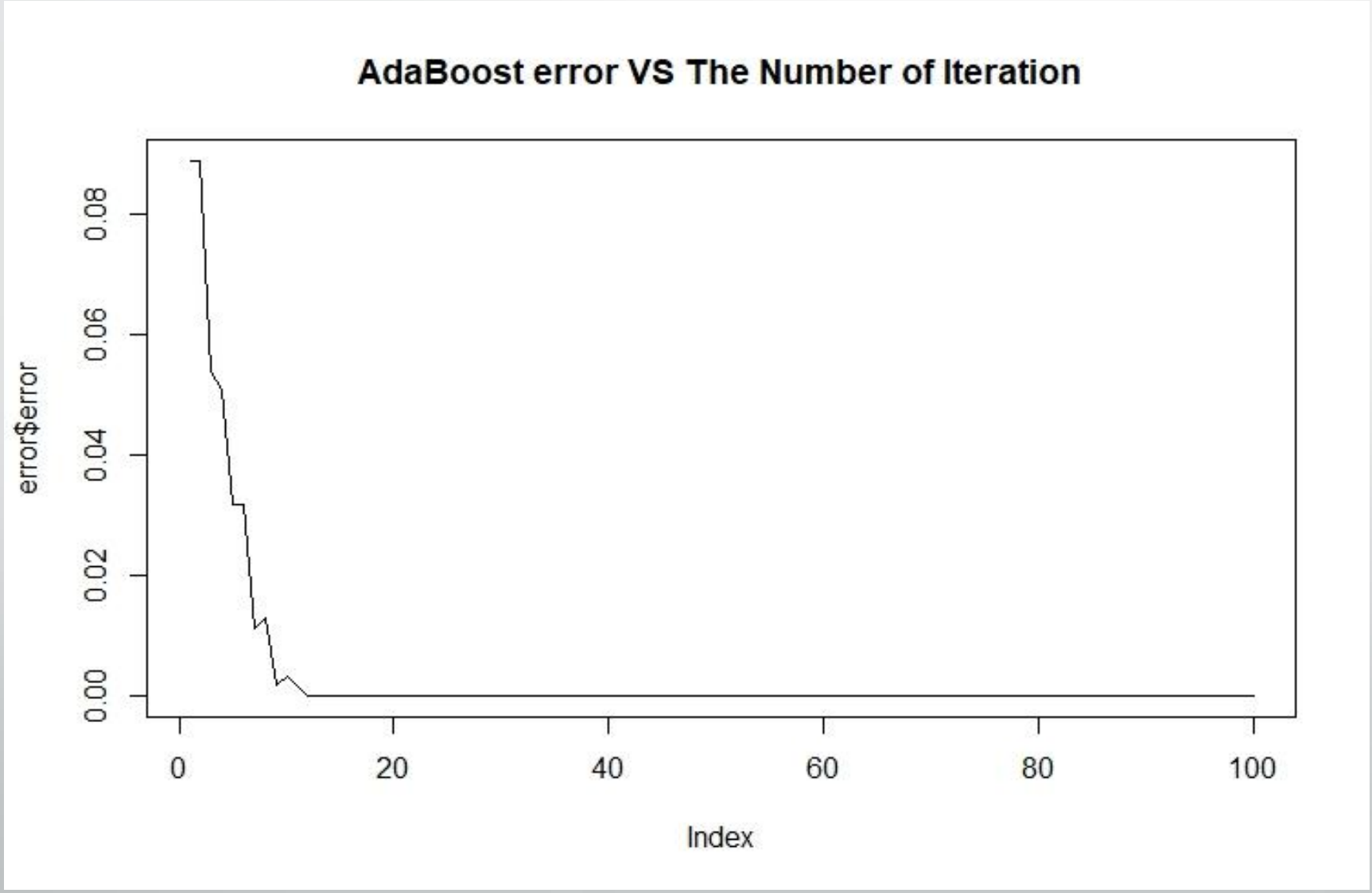
RIPPER



	Big Chance	Small Chance
Big Chance	119	14
Small Chance	20	117

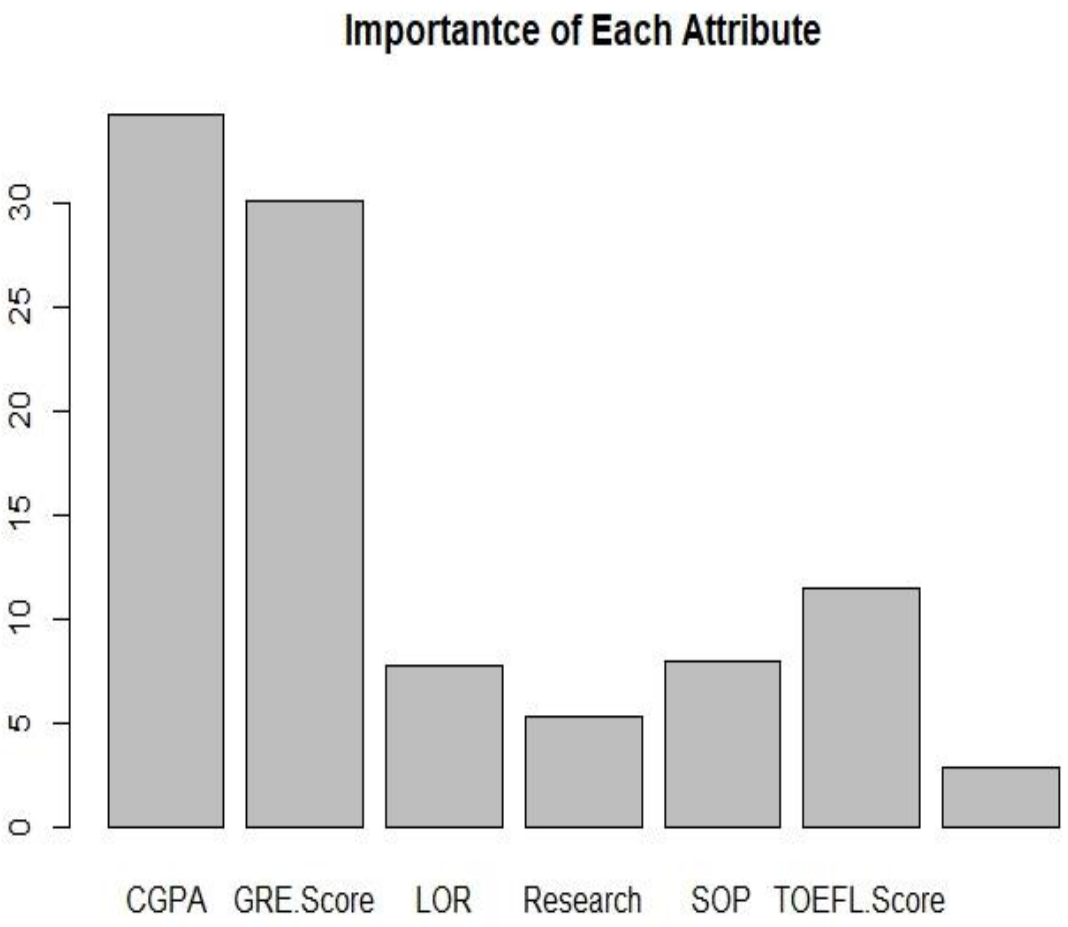
- Accuracy = 0.874
- Recall = 0.895
- Precision = 0.854
- F-Measure = 0.875

Adaboost

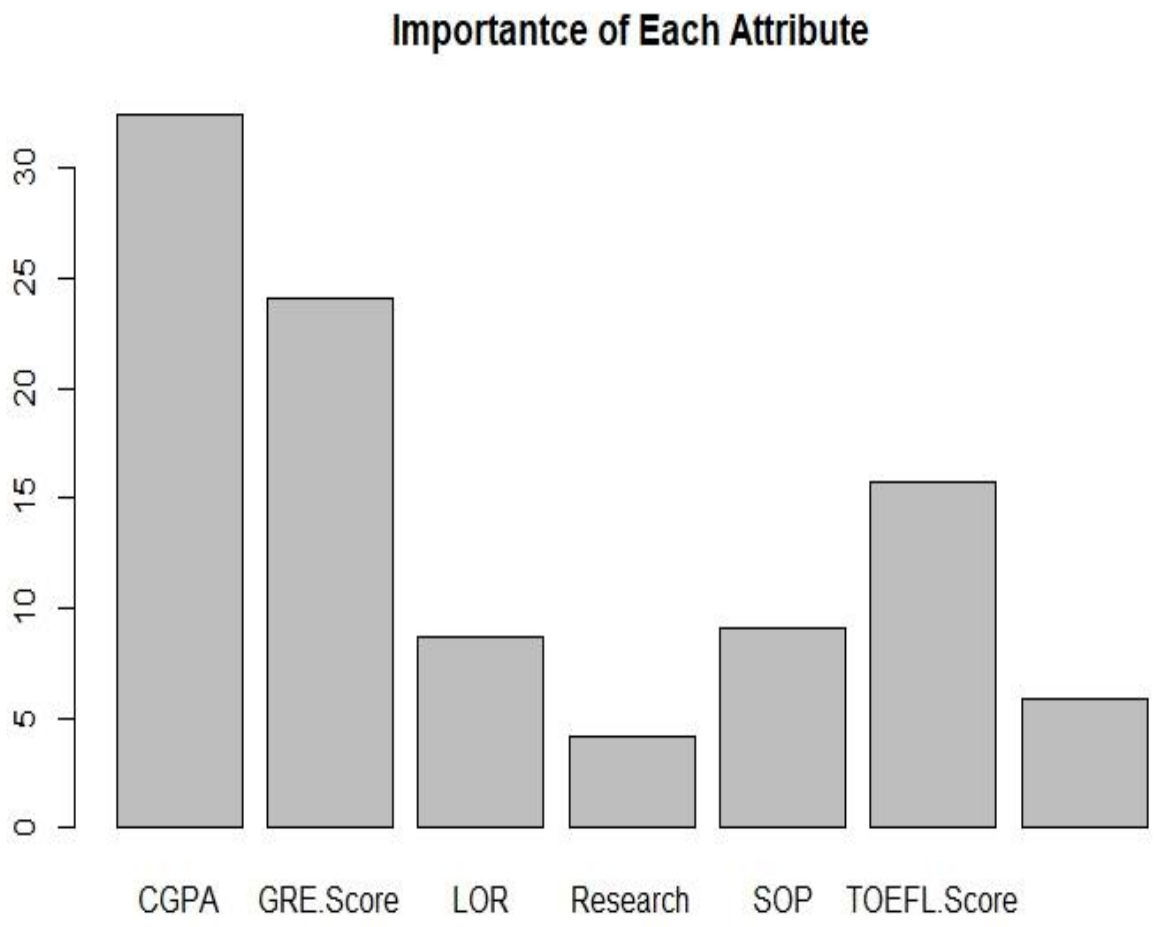

AdaBoost error VS The Number of Iteration
error$error
0.00
0.02
0.04
0.06
0.08
0
20
40
60
80
100
Index

Adaboost

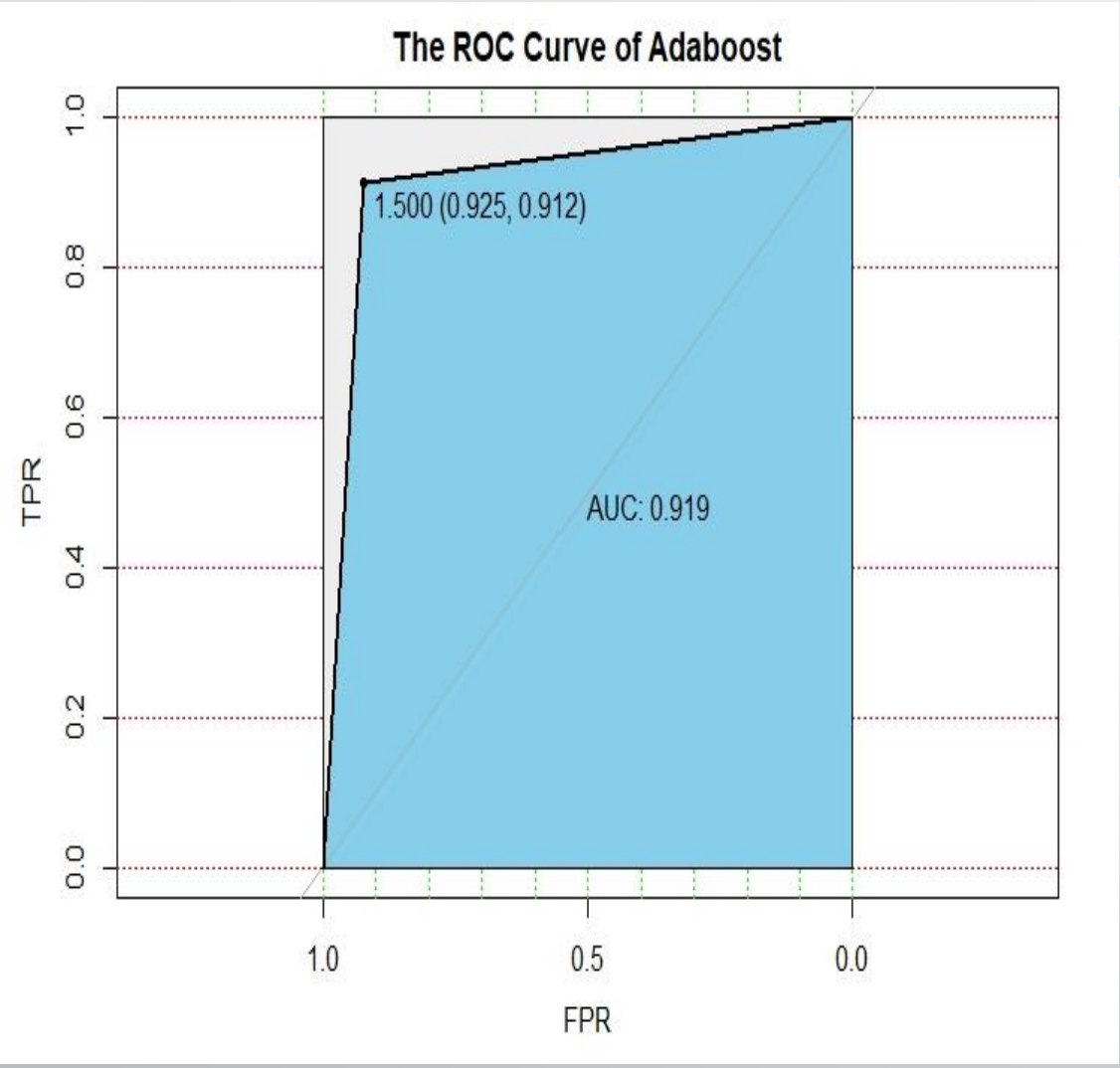
Iteration: 20 times



Iteration: 100 times



Adaboost



	Big Chance	Small Chance
Big Chance	123	10
Small Chance	12	125

- Accuracy = 0.919
- Recall = 0.925
- Precision = 0.912
- F-Measure = 0.918

Compare All Algorithms

	Accuracy	Recall	Precision	F-Measure
Cart	0.841	0.857	0.826	0.8412
C4.5	0.882	0.925	0.848	0.885
SVM	0.856	0.865	0.846	0.855
K-NN	0.855	0.917	0.859	0.887
Adaboost	0.919	0.925	0.911	0.918
RIPPER	0.874	0.895	0.856	0.875

Summary

- GPA and GRE are the most important factors in admission.
- Research , Recommendation letter and Statement of Purpose slightly affect the chance of admission.
- Prediction of Chance of Admit?
  - Use Linear Regression to estimate the chance of admission.
  - Adaboost Algorithm is the best algorithm to simulate the model.

Improve your GPA

Thanks !