Classification of Body Performance

Team 5

Anni Wang
Tianhao Wu
Yueting Wu
Zhiyong Jiang
Zihe Yan

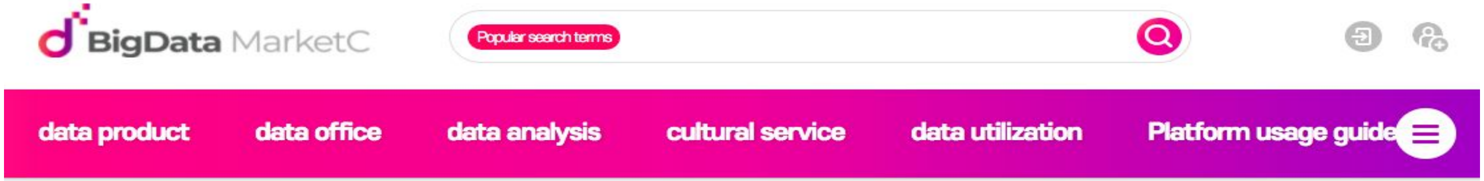
BerkeleyMEng

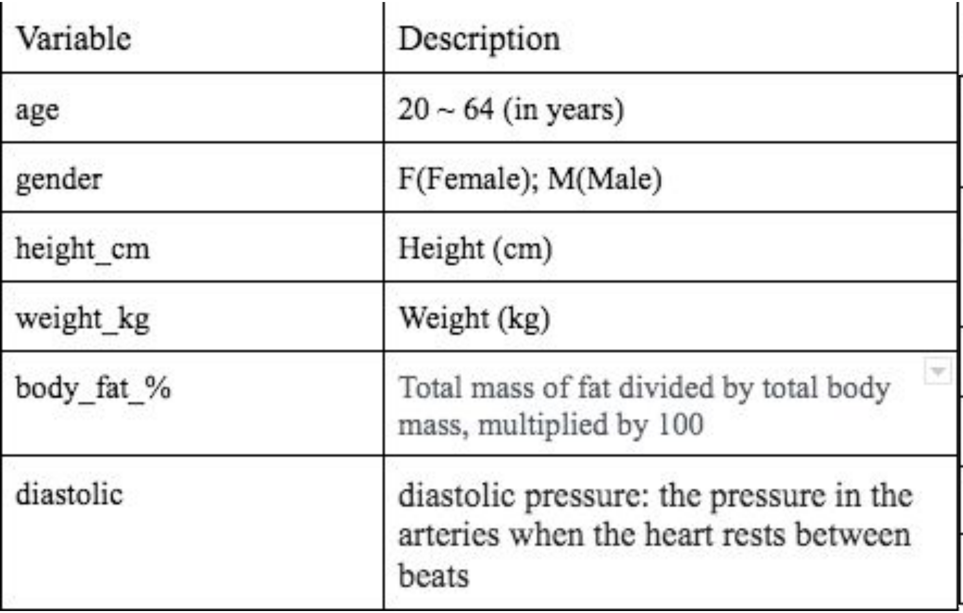
Motivation

- What: Our project focuses on the assessment of body performance based on some physical indicators and exercise performance data.
- Why: Body health is closely related to our life. Better body health represents more energy helping us accomplish goals, greater pride in ourselves, better emotions and so on.

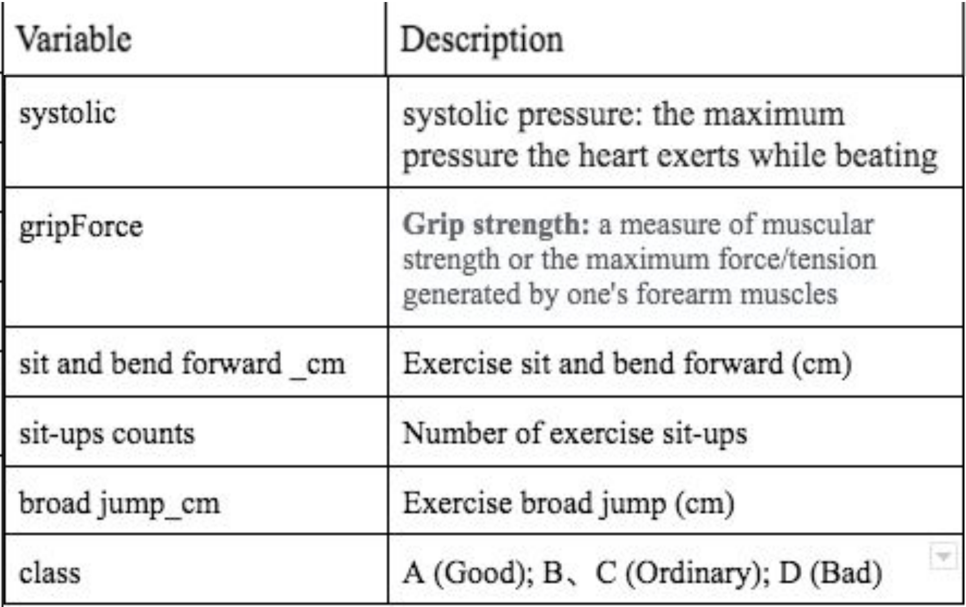
Data Description

Data Source (Sports Promotion Foundation)





Variable	Description
age	20 ~ 64 (in years)
gender	F(Female); M(Male)
height_cm	Height (cm)
weight_kg	Weight (kg)
body_fat_%	Total mass of fat divided by total body mass, multiplied by 100
diastolic	diastolic pressure: the pressure in the arteries when the heart rests between beats



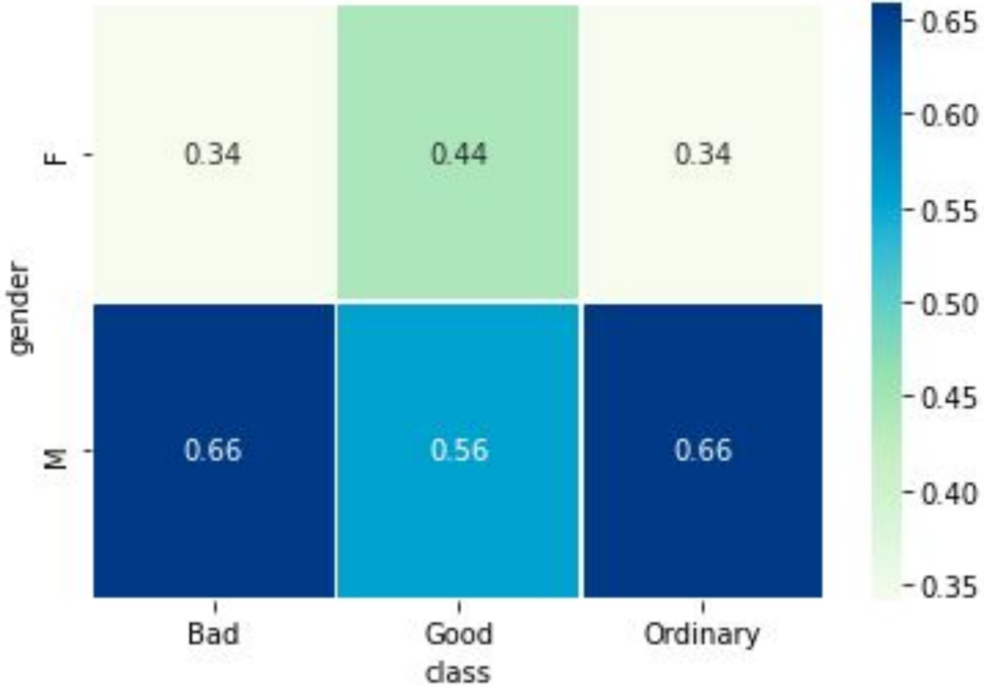
Variable	Description
systolic	systolic pressure: the maximum pressure the heart exerts while beating
gripForce	**Grip strength:** a measure of muscular strength or the maximum force/tension generated by one's forearm muscles
sit and bend forward _cm	Exercise sit and bend forward (cm)
sit-ups counts	Number of exercise sit-ups
broad jump_cm	Exercise broad jump (cm)
class	A (Good); B、C (Ordinary); D (Bad)

Data shape (13393,12)

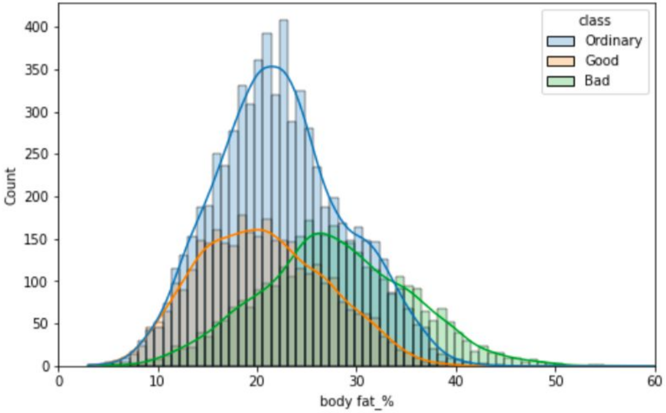
Exploratory Data Analysis

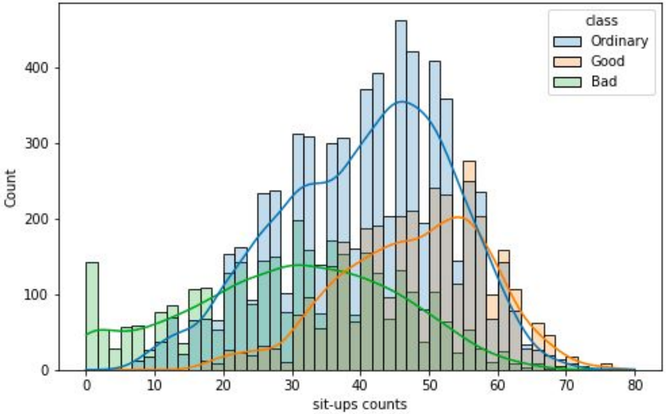
- Headmap of gender distribution of different classes.

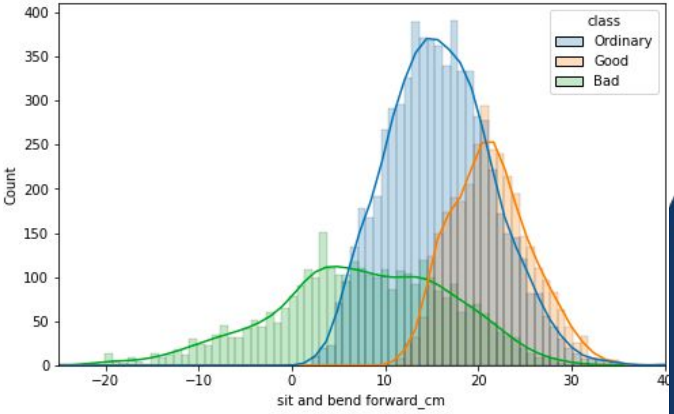
  Male have higher percentage in Class ‘Bad’ and ‘Ordinary’.
  In general, female outperform male in this body performance assessment.



Exploratory Data Analysis







Analytic Models

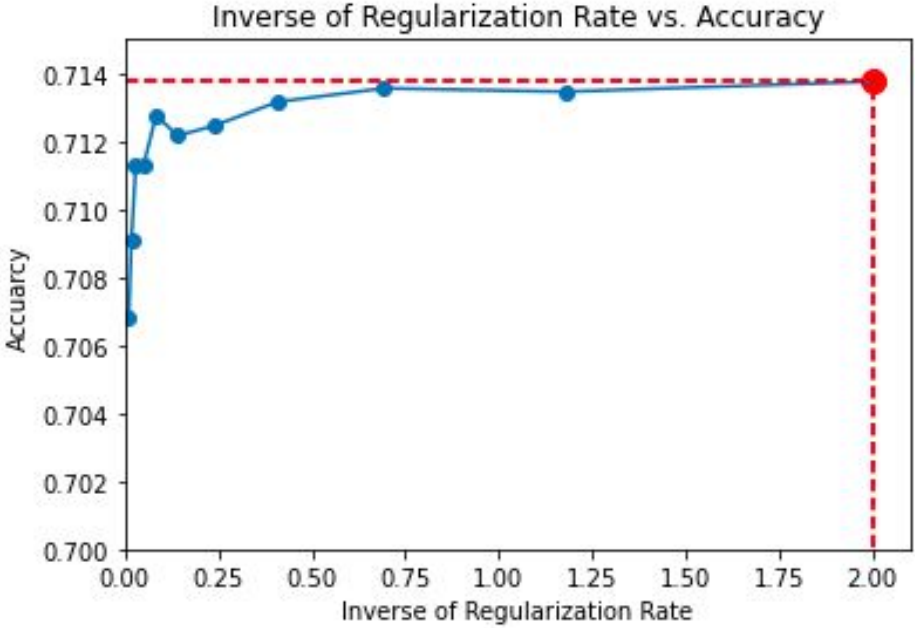
Models we use:

1. Baseline model
2. Logistic Regression
3. Random Forest
4. Bagging
5. Gradient Boosting
6. Neural Network

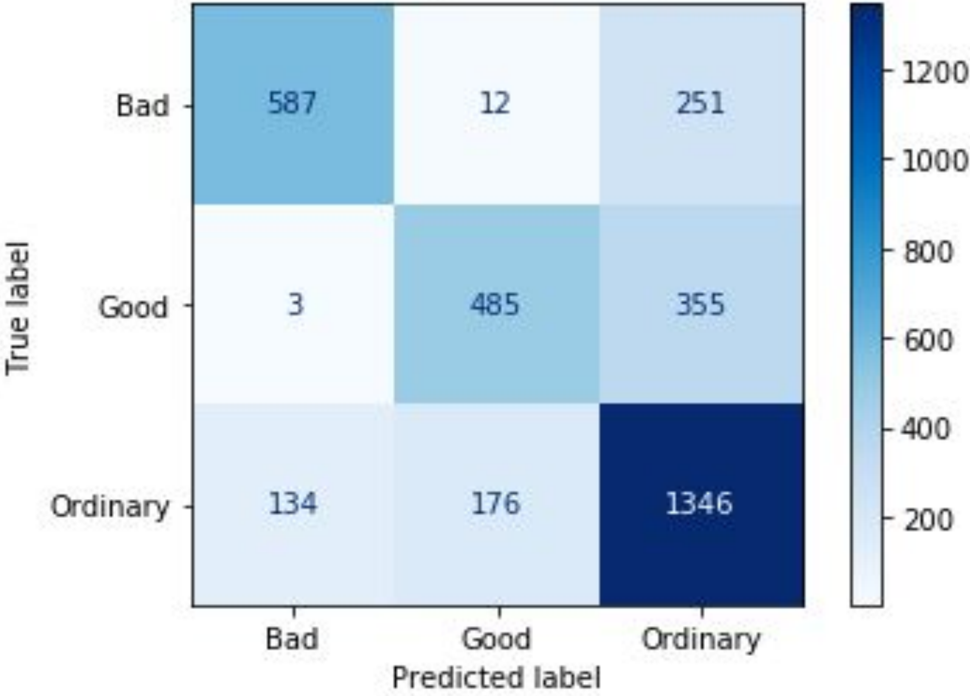
Models -- Logistic Regression

- Finding Inverse of Regularization rate = 2 as our best parameter to run logistic regression model

```
Best accuarcy: 0.7137592487194O8
Best C: 2.0000000000000004
```



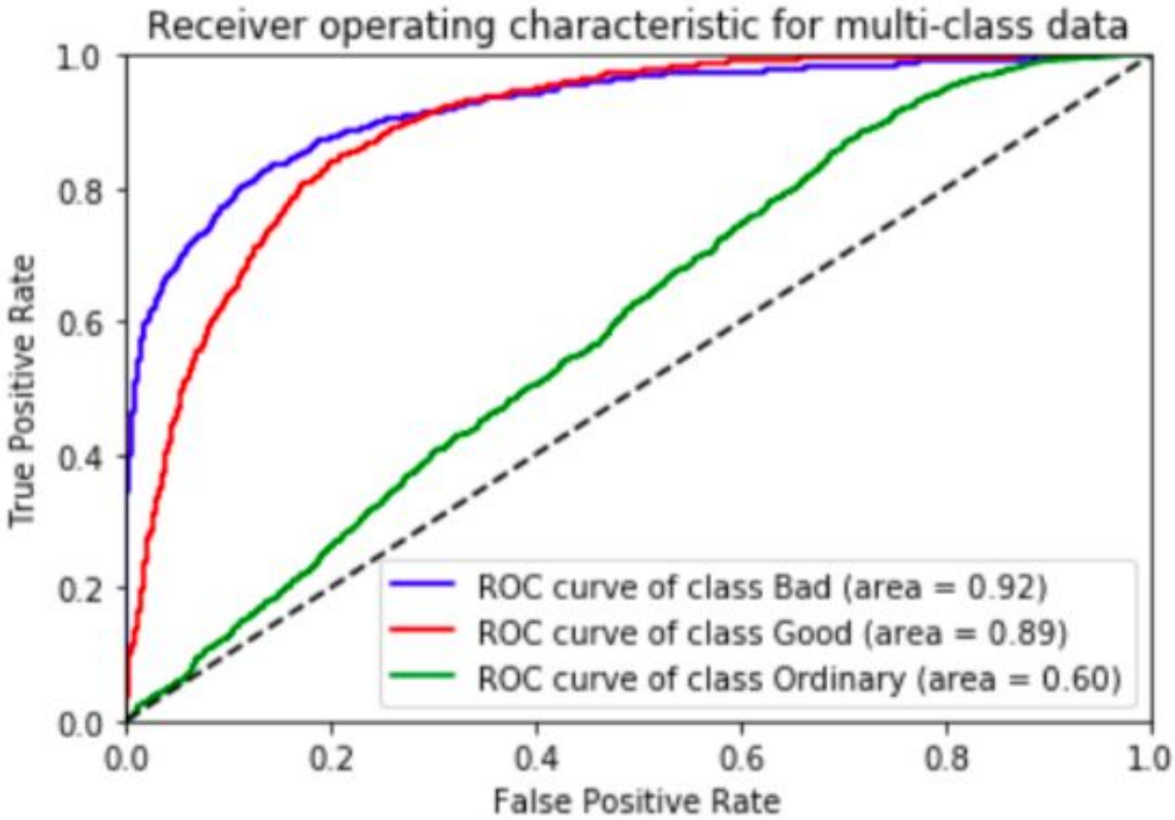
- Confusion Matrix



True label \ Predicted label	Bad	Good	Ordinary
Bad	587	12	251
Good	3	485	355
Ordinary	134	176	1346

Models -- Logistic Regression

Below is the ROC curve for different classes prediction performance.
Logistic Regression performs better on predicting Class ‘Bad’ and ‘Good’.

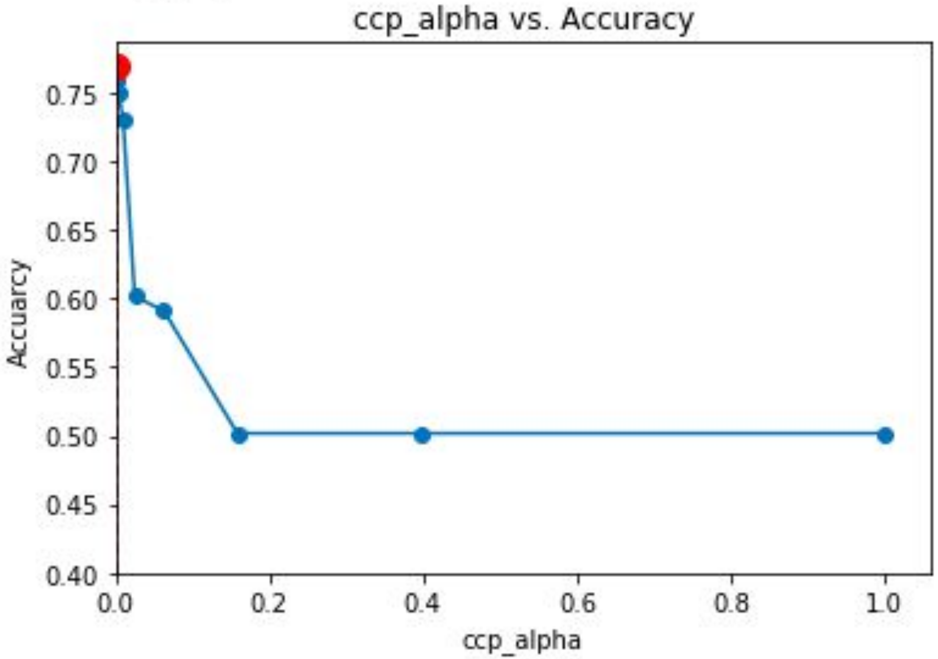


BerkeleyMEng

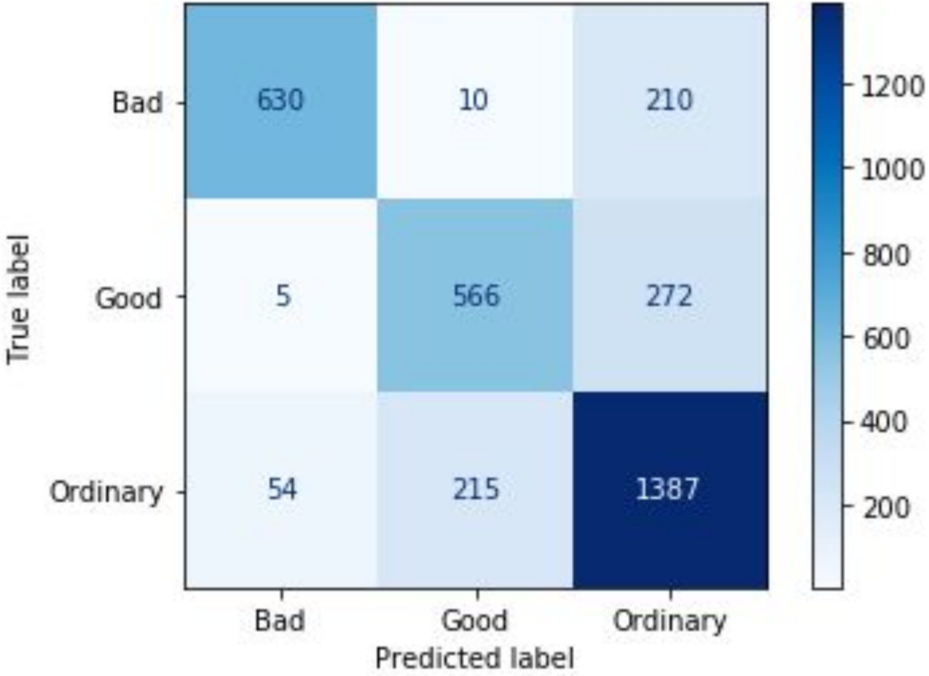
Models -- Random Forest

- Finding Best ccp_alpha = 0.0001 to run random forest model

```
Best accuarcy: 0.7684189573215351
Best ccp_alpha: 0.0001
```



- Confusion Matrix



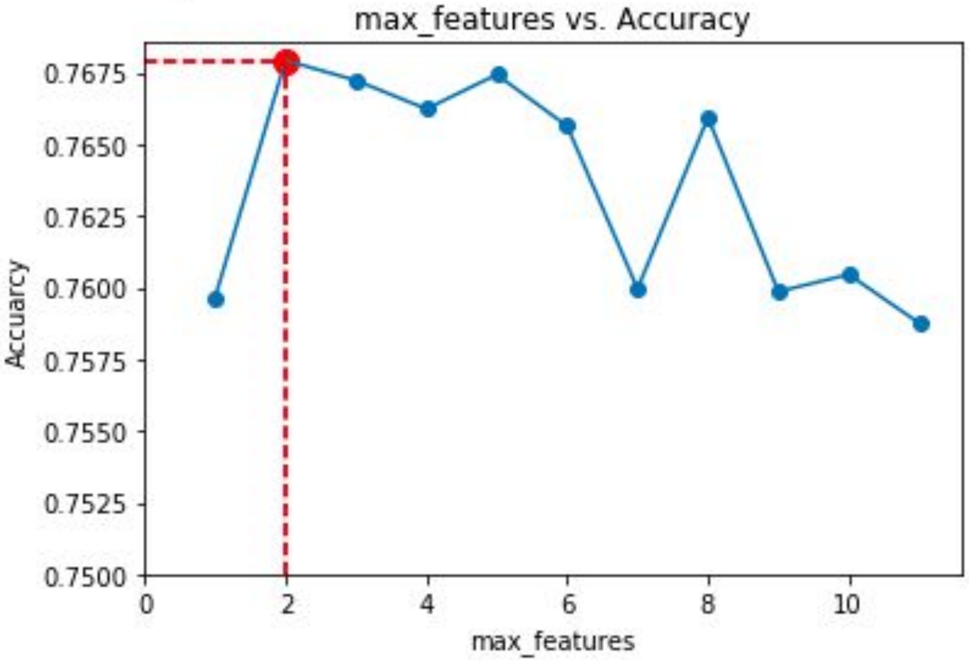
	Bad	Good	Ordinary
Bad	630	10	210
Good	5	566	272
Ordinary	54	215	1387

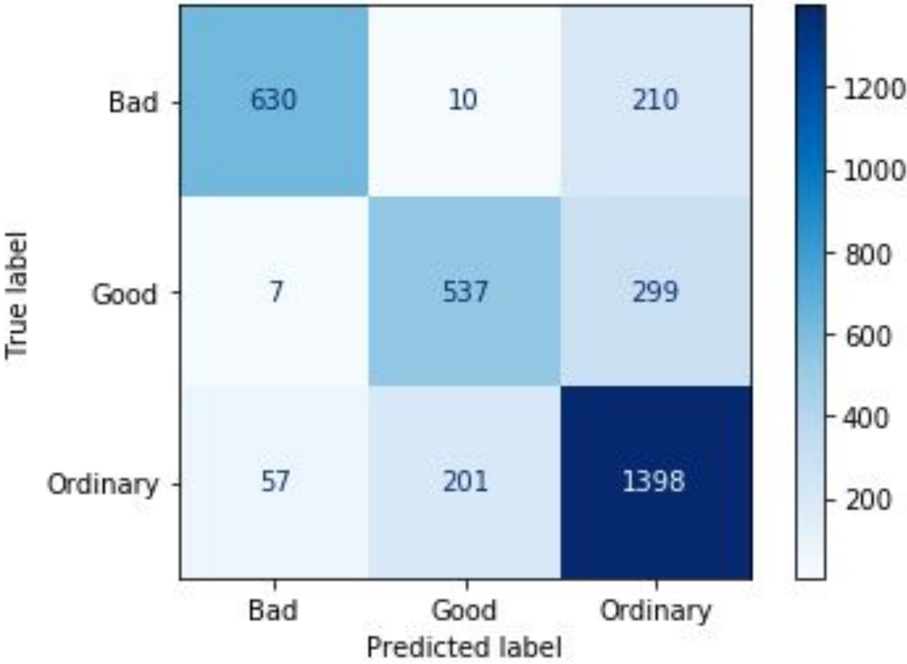
Models -- Bagging

- Finding Best max_features = 2 to run bagging model

- Confusion Matrix

```
Best accuarcy: 0.7679206023095274
Best max_features: 2
```

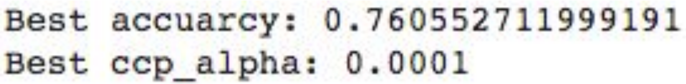




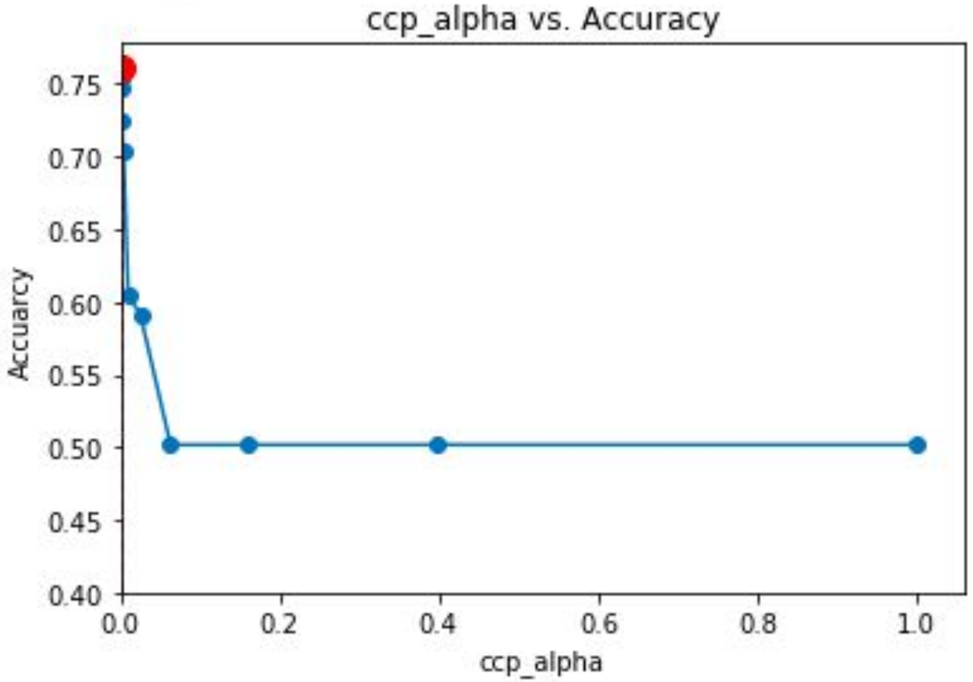
	Bad	Good	Ordinary
Bad	630	10	210
Good	7	537	299
Ordinary	57	201	1398

Models -- Gradient Boosting

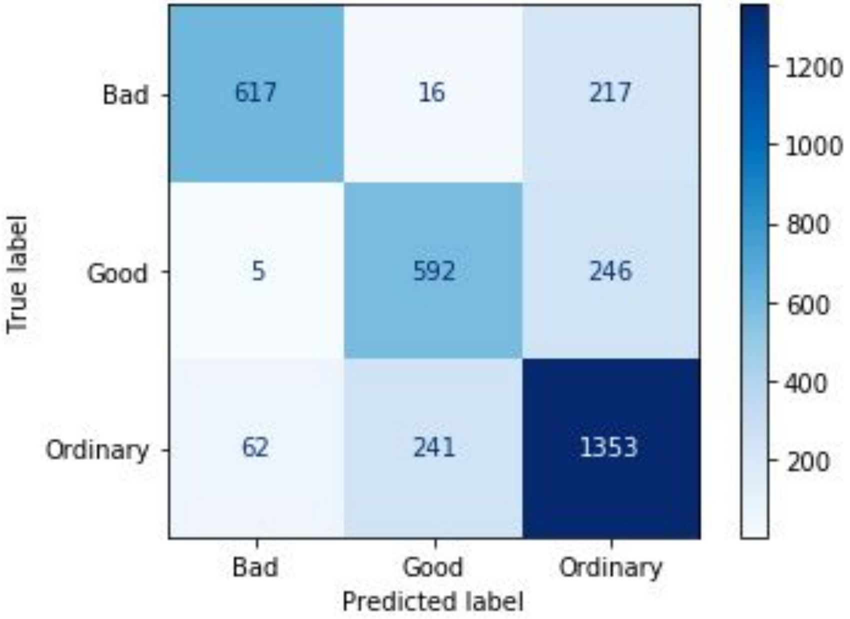
- Finding Best ccp_alpha = 0.0001 to run gradient boosting model



```
Best accuarcy: 0.760552711999191
Best ccp_alpha: 0.0001
```

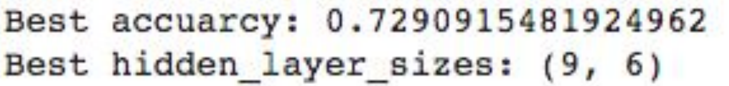


- Confusion Matrix

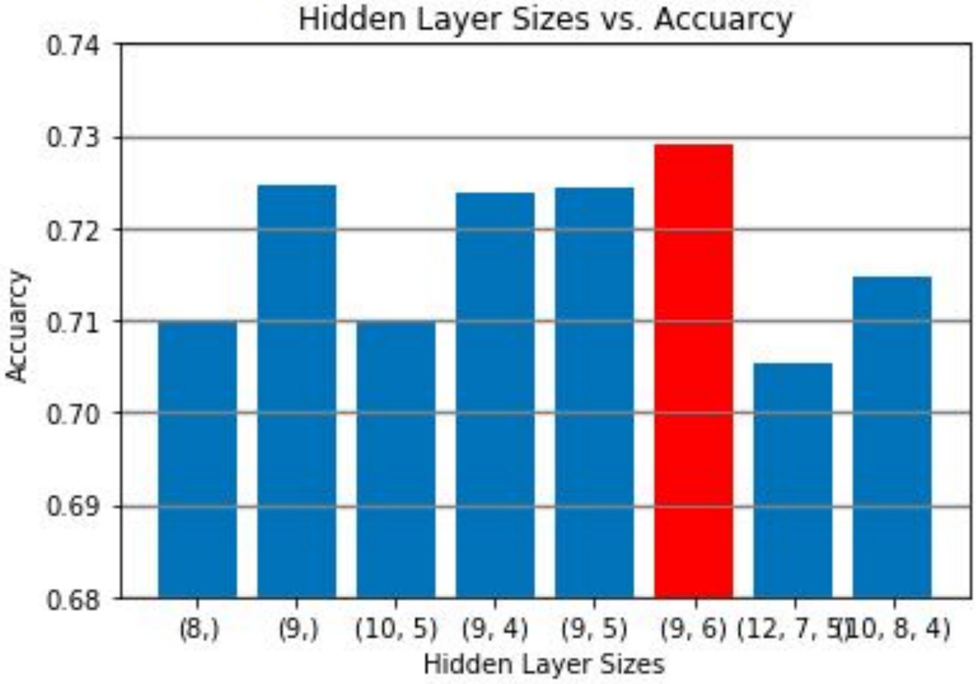


Models -- Neural Network

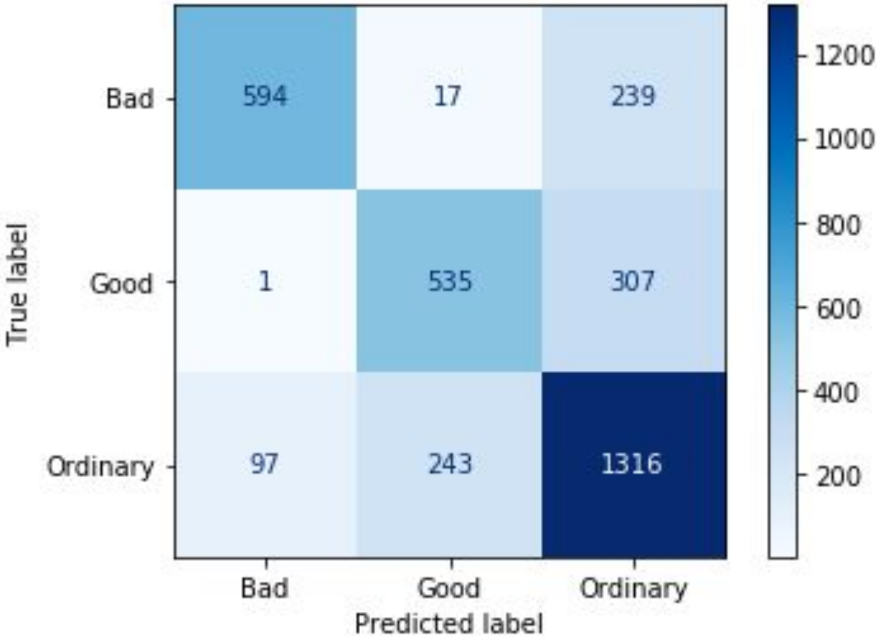
- Finding Best hidden_layer_sizes : (9,6) to run neural network model



```
Best accuarcy: 0.7290915481924962
Best hidden_layer_sizes: (9, 6)
```



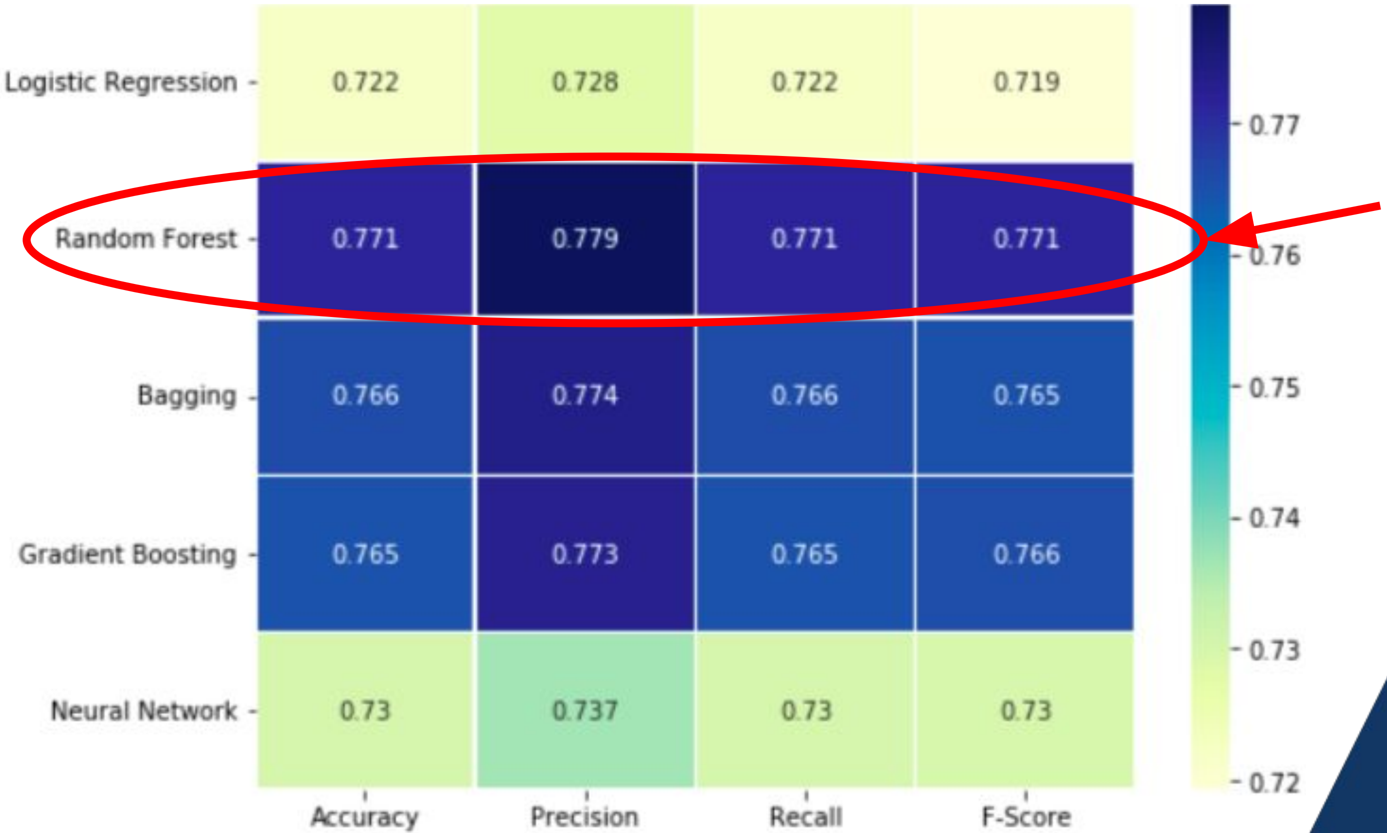
- Confusion Matrix



True label \ Predicted label	Bad	Good	Ordinary
Bad	594	17	239
Good	1	535	307
Ordinary	97	243	1316

Model Comparison

Model comparison based on 4 key metric -- Accuracy, Precision, Recall, F-score (except baseline model)

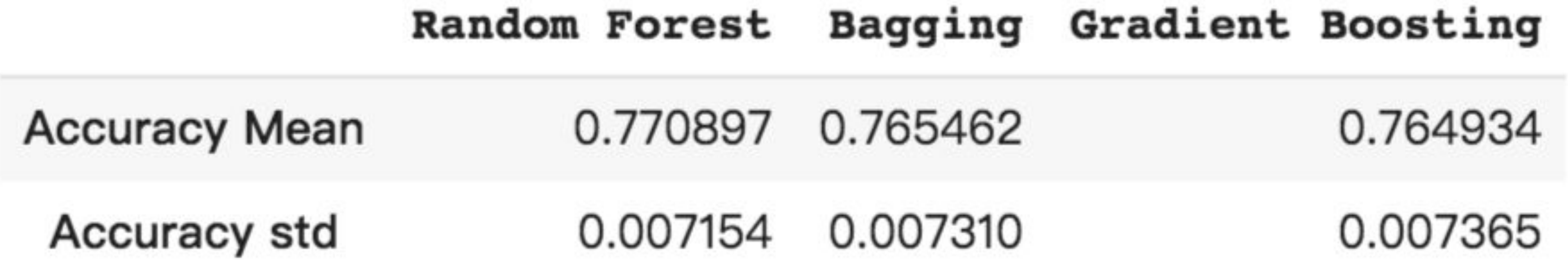


	Accuracy	Precision	Recall	F-Score
Logistic Regression	0.722	0.728	0.722	0.719
Random Forest	0.771	0.779	0.771	0.771
Bagging	0.766	0.774	0.766	0.765
Gradient Boosting	0.765	0.773	0.765	0.766
Neural Network	0.73	0.737	0.73	0.73

Perform best

Model Evaluation

Model evaluation through Boostrap to carefully find which model performs best.
Below is the performance_table presenting Accuracy mean and Accuracy std between model Random Forest, Bagging, Gradient Boosting

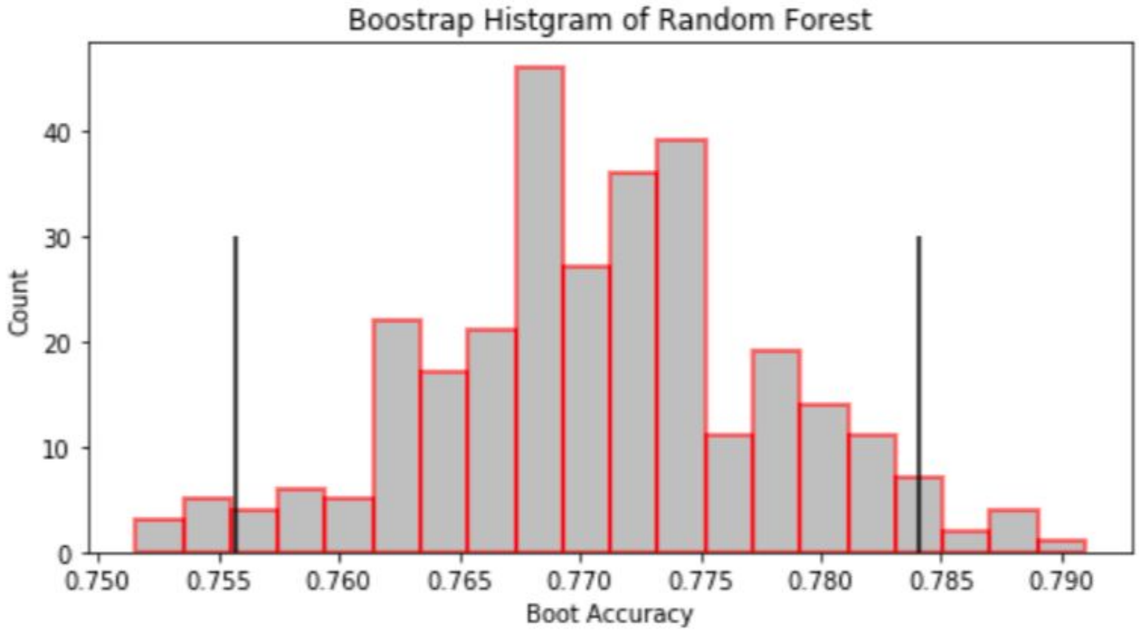


	Random Forest	Bagging	Gradient Boosting
Accuracy Mean	0.770897	0.765462	0.764934
Accuracy std	0.007154	0.007310	0.007365

Model Evaluation

Upon comparison, **Random Forest** perform best among all six models.
For our chosen model, we again use boostrap to construct a confidence interval for its accuracy.

95-percent CI of accuracy is [0.75573305 0.78411466]



Impact

1. What is the (potential) impact of your work with regard to the problem that you

are trying to solve?

People can use this model to assess their body performance by plugging some required data , such as gender, height, weight, body_fat_% and so on, to see which class they will fall into.

2. How might you expand the scope of your analysis to improve its impact

even more?

We can add more physical indicators such as BMI, number of cigarettes per day, blood glucose level or other exercise performance data like long jump in place, 800m long run to our independent variables to improve the accuracy of our analysis.