acy20zw_200206297_AS1

April 2021

1 Question 1

1.1 Best parameter

1. Random Forest Classifier

RFC	maxDepth	numTrees	maxBins	Accuracy	AUC
	10	30	10	0.7073	0.7056

2. Gradient boosting Classifier

GBT	maxIter	maxDepth	maxBins	Accuracy	AUC
	15	4	10	0.7089	0.7076

3. Neural networks Classifier

MPC	maxIter	layers	StepSize	Accuracy	AUC
	30	[28,2]	0.03	0.6387	0.6332

1.2 Results of the model with large data set

Model	Feature	5_core time	10_core time	accuracy	AUC
RFC	_c26, _c28, _c27	937.16s	748.13s	0.7228	0.7210
GBC	_c26, _c28, _c23	212.10s	156.81s	0.7152	0.7138
MPC		74.49s	57.48s	0.6362	0.6293

1.3 Observations

- The performance of each model is not very good may because our parameters' settings in cross-validation did not contain the best parameter. But we can notice that the Random Forest Classifier can achieve the best performance when the small dataset was used.

- When the larger dataset was used to training the model, the accuracy and AUC of gradient boosting and random forest classifier increased while neural network classifier's accuracy and AUC decreased. It might because the model is overfitted.
- While 5 core and 10 core be used to compare the performance, wen can find it will need more time while those task use 5 cores,

2 Question 2

2.1 Preprocessing

1. First, data which contain more than 2 null value will be dropped.

2. Second, the mode of each column will be used to fill the null value.

3. Third, based on the conclusion in Assignment1 in Machine Learning and Artificial Intelligence, several columns will be dropped.

4. Some columns will be convert to double directly and others will be use StringIndexer to convert to double

2.2 Linear Regression

Core_num	MSE	MAE	Time(s)
5 Core	98049.91	134.75	271.97
10 Core	97656.52	134.24	188.74

2.3 Combined model

Core_num	MSE	MAE	Time(s)
5 Core	74361.81	168.77	428.32
10 Core	73097.81	168.69	296.55

2.4 Observations

- During the preprocess part, the mode of some columns of data are 'Null', which means while we use the mode of each column to fill null value, we need to pay attention to it.
- The mean square error of linear regression is much higher than the combined model, which might because the linear regression model will predict a value for all data while the label of some data is 0 which might increase the mean square error of the linear regression model. But the combined

model classified which data should be the label 0 and then predict the label's value for other data.

- The mean absolute error of the combined model is higher than linear regression model. It might because of the training set in linear regression model contains data which label is 0, it might affect the model to predict data have a label which is near to zero. But the combined model has already classified those data with label 0 or 1 and only predict the data with label 1 which may have some label with a big data.
- It is clear to find that while use 10 core, the time is much lower than use 5 core.