CITI PROGRAM

Completion Date 19-Aug-2020
Expiration Date 19-Aug-2023
Record ID 37950294

This is to certify that:

JINGYAO ZHU

Not valid for renewal of certification through CME. Do not use for TransCelerate mutual recognition (see Completion Report).

Has completed the following CITI Program course:

Human Research (Curriculum Group)
Data or Specimens Only Research (Course Learner Group)
1 - Basic Course (Stage)

Under requirements set by:

Massachusetts Institute of Technology Affiliates

CITI
Collaborative Institutional Training Initiative

Verify at www.citiprogram.org/verify/?w4901bf2a-65e0-4ccd-a064-6a6402bb29ec-37950294

2.(b) Table 2

Metric	Deceased patients	Alive patients	Function to complete
Event Count 1. Average Event Count 2. Max Event Count 3. Min Event Count	1. 982.014 2. 8635 3. 1	1. 498.118 2. 12627 3. 1	event count metrics
Encounter Count 1. Average Encounter Count 2. Max Encounter Count 3. Min Encounter Count	1. 23.038 2. 203 3. 1	1. 15.452 2. 391 3. 1	encounter count metrics
Record Length 4. Average Record Length 5. Max Record Length 6. Min Record Length	1. 127.532 2. 1972 3. 0	1. 159.2 2. 2914 3. 0	record length metrics

4.1(b) Table 3

Model	Accuracy	AUC	Precision	Recall	F-Score
Logistic Regression	0.95454545	0.96138858	0.89880952	0.98692810	0.94080997
SVM	0.99401914	0.99309426	0.99702381	0.98820059	0.99259259
Decision Tree	0.77631579	0.78076001	0.60119048	0.79215686	0.68358714

4.1(c) Table 4

Model	Accuracy	AUC	Precision	Recall	F-Score
Logistic Regression	0.73809524	0.73401150	0.73333333	0.68041237	0.70588235
SVM	0.73809524	0.73478023	0.74444444	0.67676768	0.70899471
Decision Tree	0.67142857	0.66378394	0.55555556	0.63291139	0.59171598

4.1(d)

Here are some potential strategies to improve model performance

1. Removing outliers: outliers will affect model regression in either slowing regression or causing departure. So, remove or replace some “reasonable” outliers could help improve
2. Collecting more data: generally speaking, more data could help model learn more about its tendency. Therefore, increase model accuracy or decrease model loss.
3. Feature engineering: dummy some features or transfer some features could use somehow increase information gain. Finally, it will make model perform better
4. Model parameter tuning: tuning the model parameter to optimal value could improve the model accuracy
5. Model algorithm selection: The better algorithm selection, the lower model loss is because right algorithm could gather right information from features.

4.2(b) Table 5

CV strategy	Accuracy	AUC
K-Fold	0.72370972	0.71160783
Randomized	0.73571429	0.73084615

4.3(b)

In my model, I finally selected Gradient Boosting Classifier and Extra Trees Classifier together using VotingClassifier. The reason I select Gradient Boosting Classifier is because it is a boosting model and it will learn from previous model and make the best possible in the next one. Picking Extra Trees Classifier is that algorithm fits each tree on the whole training data. it does not like random forest or bagging algorithm. RidgeClassifier is better to work on correlated features. So, it may work for this project. And MLPClassifier stands for Multi-layer Perceptron classifier which is a kind of Neural Network. It will help convert the a few inputs into a more useful output. I do believe that the best estimation is to combine those four models.

Below is my results of cross validation and AUC score. Compared to the previous results, it actually did a great job.

CV strategy	Accuracy	AUC
K-Fold	0.72970488	0.72523097
Randomized	0.73540670	0.73603130

Search

Overview Data Notebooks Discussion Leaderboard Rules Team My Submissions **Submit Predictions**

You may select up to 2 submissions to be used to count towards your final leaderboard score. If 2 submissions are not selected, they will be automatically chosen based on your best submission scores on the public leaderboard. In the event that automatic selection is not suitable, manual selection instructions will be provided in the competition rules or by official forum announcement.

Your final score may not be based on the same exact subset of data as the public leaderboard, but rather a different private data subset of your full submission — your public score is only a rough indication of what your final score is.

You should thus choose submissions that will most likely be best overall, and not necessarily on the public subset.

29 submissions for Jingyao Zhu

Sort by Most recent

All Successful **Selected**

Submission and Description	Public Score	Use for Final Score
my_predictions.csv 31 minutes ago by Jingyao Zhu add submission details	0.72646	☑

No more submissions to show