STAT3612 STATISTICAL MACHINE LEARNING

Group Project Presentation

30-day All-Cause Hospital Readmission Prediction

Group(8) members:

Zixuan Yao (3035845148)

Ziyu Wang (3035777547)

Jiahe Fang (3035772482)

Yining Huang (3035662522)

Zixun Huang (3035844522)

https://github.com/zxyao5148/STAT3612_2023_1A_GroupProject

CONTENT

PART 1

DATA EXPLORATION

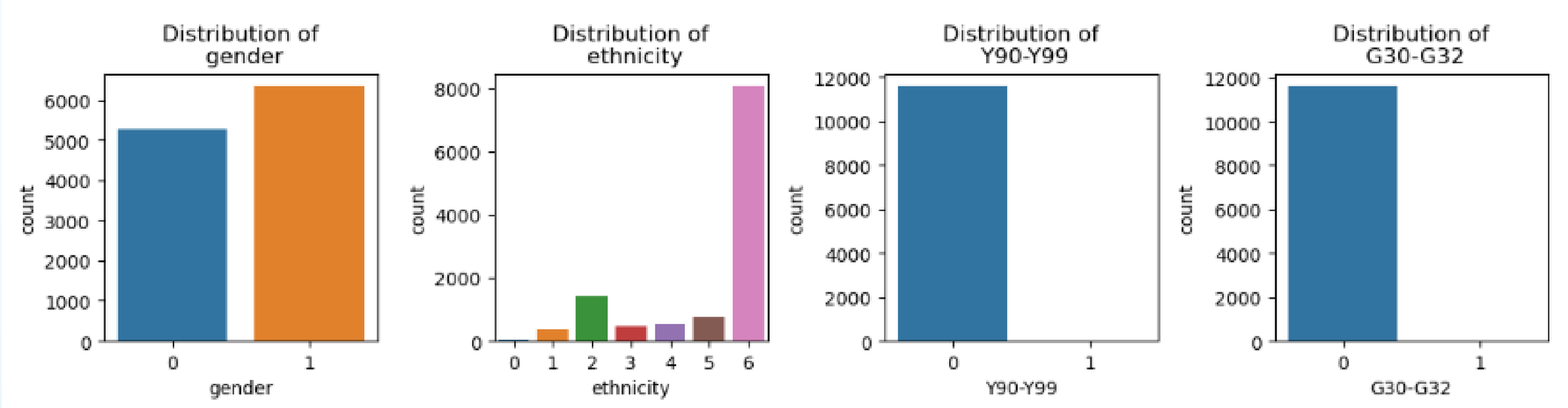
Data Exploration

- **Check for missing values**
- **Descriptive statistics**
- **Feature Engineering**:
  - Identify and extract statistics to summarize the attributes of patients' records.
    - Mean/mode
    - Latest
    - Std, Quartiles (Q1, Q3), IQR & range
    - Max/min
    - Kurtosis & Skewness

Data Exploration

Identify patterns and insights - Distribution (categorical data)

- E.g.



Observation: features with constant values (constant variance) -> drop

Data Exploration

Identify patterns and insights - Class Imbalance

Observation: more labels of 0s than 1s could result in a biased model (poor in predicting 1s)

Solution: SMOTE (Synthetic Minority Over-sampling Technique) and weighting adjustment

Data Exploration

Identify patterns and insights - Correlation

- **Observation:**
- 0.3 < corr < 0.4:
  - pO2 Blood mean/mode & latest, pH Blood latest
- 0.2 < corr < 0.3:
  - Lactate Blood mean/mode, std & latest, pH Blood mean/mode, pCO2 Blood mean/mode, std & latest
- 0.1 < corr < 0.2:
  - ELECT/CALORIC/H2O mean/mode, ANTIBIOTICS mean/mode, etc.

- **Remark:** features with high pairwise correlation were removed

Data Exploration

More on Feature Engineering & Feature Selection

- **More on Feature Engineering:**
  - Polynomial, interaction, and deviation terms (not quite useful, though)

- **Feature Selection:**
  - ANOVA F-test

- **Remark:** Feature selection or not is subject to an individual model’s attribute

PART 2

MODEL TRAINING

Model Description

- **Logistic regression**: easy to interpret
- Discriminant Analysis
  - Linear Discriminant Analysis
  - Quadratic Discriminant Analysis
- Naive Bayes
- K-Nearest Neighbours
- Support Vector Machine
- **Tree-based methods**: interpretable and accurate
  - Classification Tree
  - Bagging, Random Forests, Extra Trees
  - Boosting: XGBoost, Gradient Boosting, and AdaBoost
- Neural Networks
  - Transformer
  - Recurrent Neural Network

In total **16** models to be evaluated...

Model Fitting

Step1: Preliminary round of model fitting

- fit all models with default parameters on the training set with latest feature values
- evaluate the performance of each model using cross-validated AUC
- select the most promising model and proceed to feature selection

Preliminary Model Evaluation		
Name	CV AUC	Test AUC
Logistic Regression	0.7377	0.7231
LDA	0.7424	0.7418
QDA	0.6880	0.7047
Naive Bayes	0.6699	0.6818
KNN	0.6153	0.5988
SVM	0.7108	0.7229
Decision Tree	0.6196	0.6132
Bagging	0.7253	0.7284
Random Forest	0.7674	0.7763
Extra Trees	0.7695	0.7699
XGBoost	0.7549	0.7593
Gradient Boosting	0.7734	0.7578
Transformer	0.7834	0.7750
Bidirectional LSTM	0.7614	0.7351

Note: LDA, QDA, Naive Bayes, KNN, Decision Tree have applied feature selections

QDA, Naive Bayes, KNN, Decision Tree:
low AUC; few hyperparameter to be tuned
-> dropped

Transformer, RNN:
hard to tune hyperparameter
-> dropped

Logistic Regression, LDA, Bagging, Random Forest, Extra Trees, XGBoost, Gradient Boosting:
relatively high AUC; more hyperparameter
-> proceed to next step

Model Fitting

Step2: Adjusting class imbalance

- bootstrap oversampling for logistic regression etc.
- adjusting class weighting for tree-based methods
  - oversampling leads to **0.5** test AUC when incorporating parameter tuning...

Test AUC Comparison with Imbalance Adjustment Methods				
Name	Origin AUC	Smote-Adjusted AUC	Class Weight-Adjusted AUC	Changes
Logistic Regression	0.7319	0.7366	0.7343	
LDA	0.7292	0.7212		
Bagging	0.7323	0.7304		
Random Forest	0.772	0.7786	0.7807	
Extra Trees	0.7717	0.7716	0.7713	
XGBoost	0.7679	0.7552		
Gradient Boosting	0.7644	0.7709		
Note: LDA, QDA, Naive Bayes, KNN, Decision Tree have applied feature selections				

Model Fitting

Step3: Best subset feature selection

- train XGBoost with features being different combinations of descriptive statistics
- determine the best subset of features that consistently yields the highest AUC (test)
  - the best subset of features turns out to be **latest+mean+sd+min+max**

Combinations	AUC (test)
latest	0.7809
latest+mean	0.8060
latest+mean+sd	0.8077
latest+mean+sd+min+max	**0.8110**
latest+mean+sd+min+max+Q1+Q3	0.8050
latest+mean+sd+min+max+kurtosis+skewness	0.8058
...	...

Future opportunities

more systematic method for feature selection:

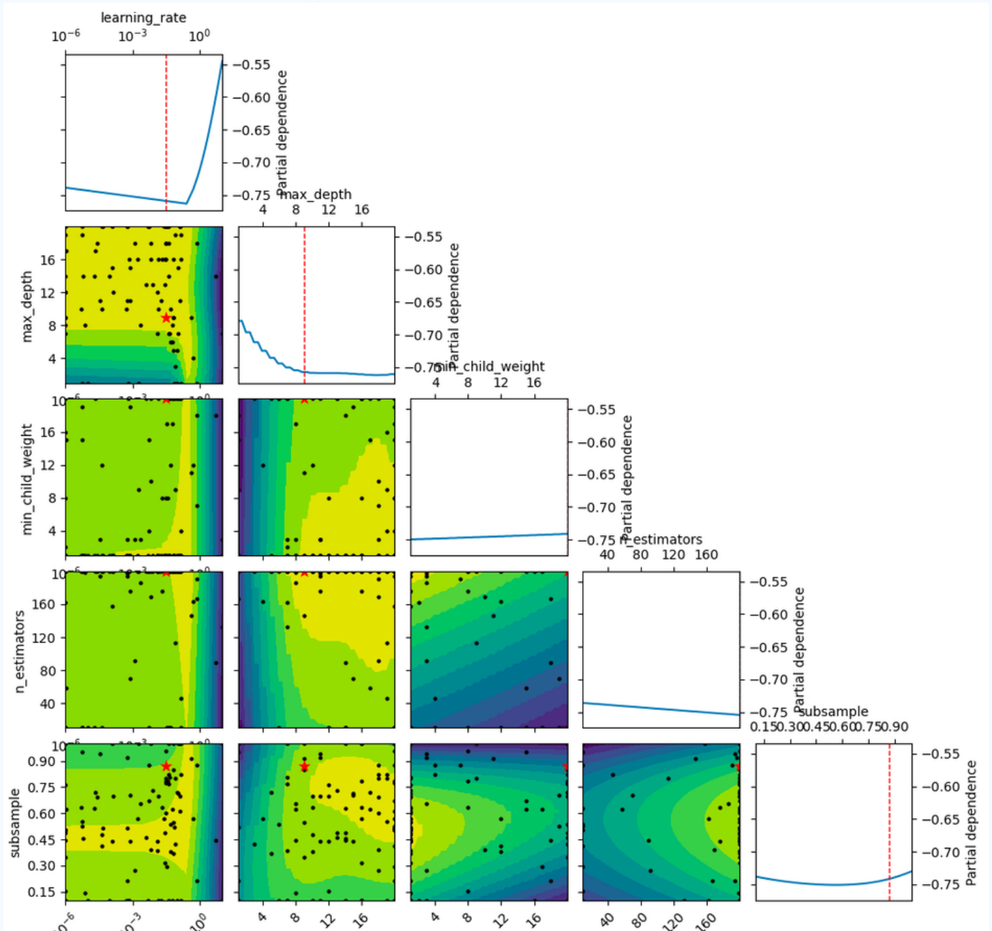
- best subset selection
- stepwise selection
- genetic algorithm
- ...

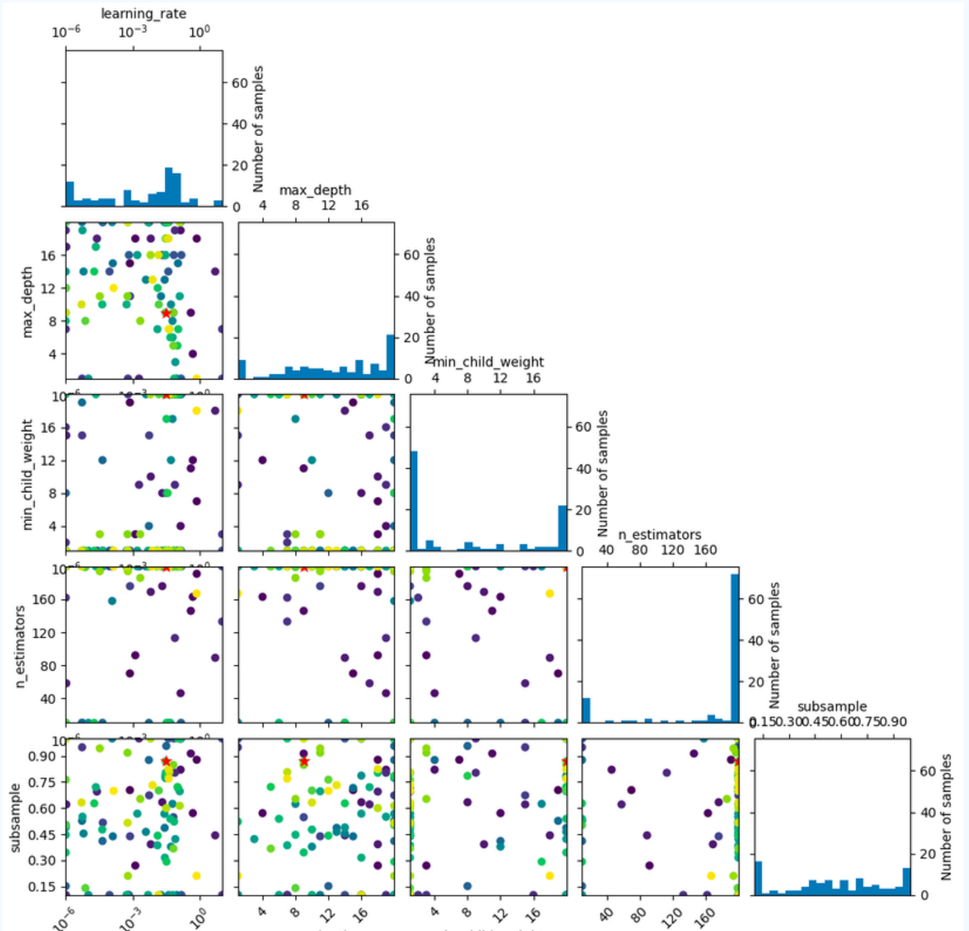
Model Fitting

Step4: Intensive hyperparameter tuning

- use Bayesian Optimization and Random Search to detect important hyperparameters and their potential optimal range
- employ Grid Search to fine tune the models that showed promising results in preliminary round of parameter tuning

XGBoost





Model Comparison

MODEL	Logistic Regression	Linear Discriminant Analysis	Bagging	Random Forest	Extra Trees	Gradient Boosting	AdaBoost	XGBoost
AUC (valid)	0.8029	0.8009	0.7934	0.80000	0.8017	0.8014	0.7865	**0.8068**
AUC (test)	0.7886	0.7829	0.8003	0.8059	**0.8093**	0.8019	0.7897	**0.8130**
Accuracy (valid)	0.7329	0.7302	0.8690	0.8703	0.8735	0.8737	0.8616	0.8740
F1-score (valid)	0.7321	0.7292	0.6798	0.7197	0.7158	0.7163	0.6612	0.7196

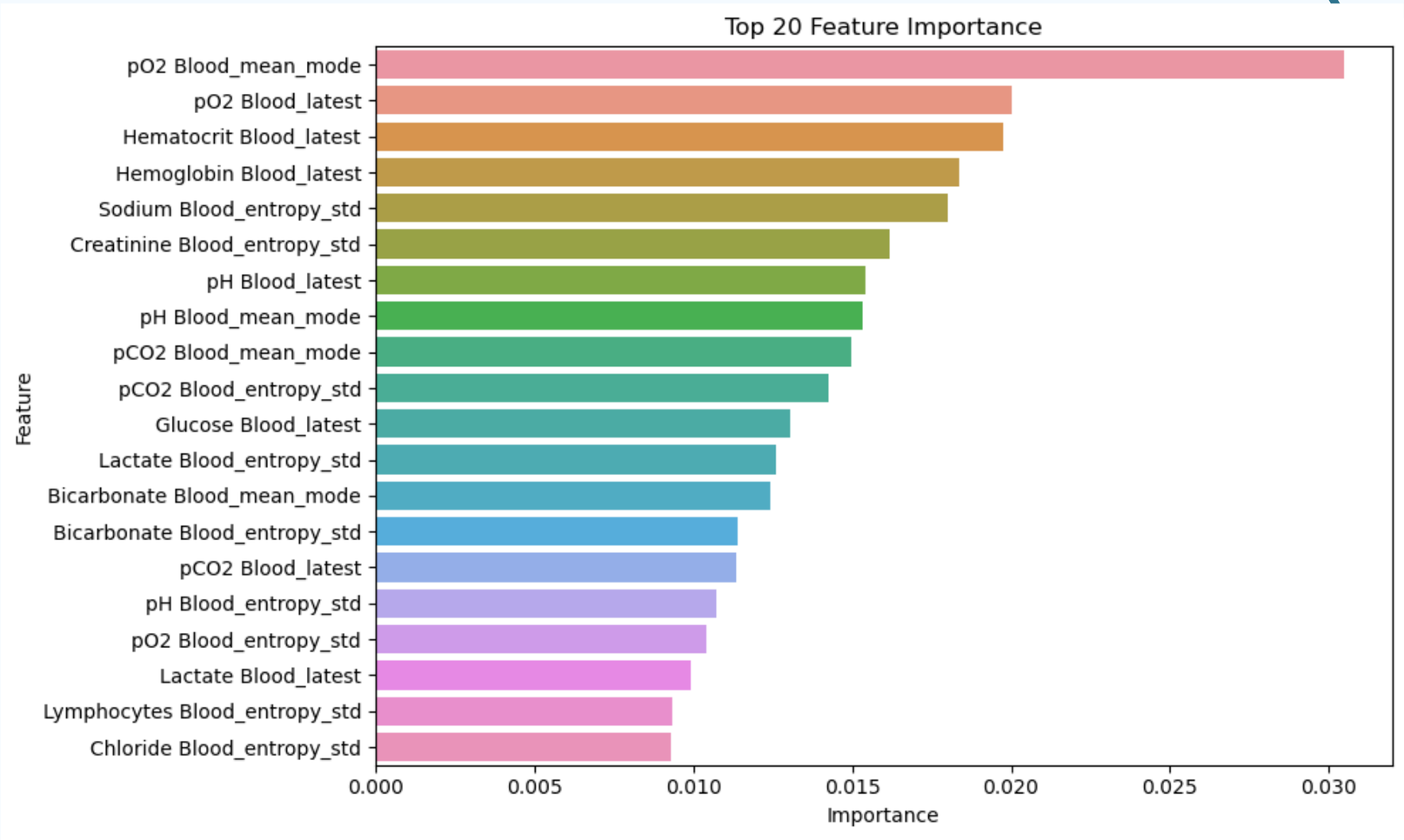
PART 3

INTERPRETATION

Extra trees

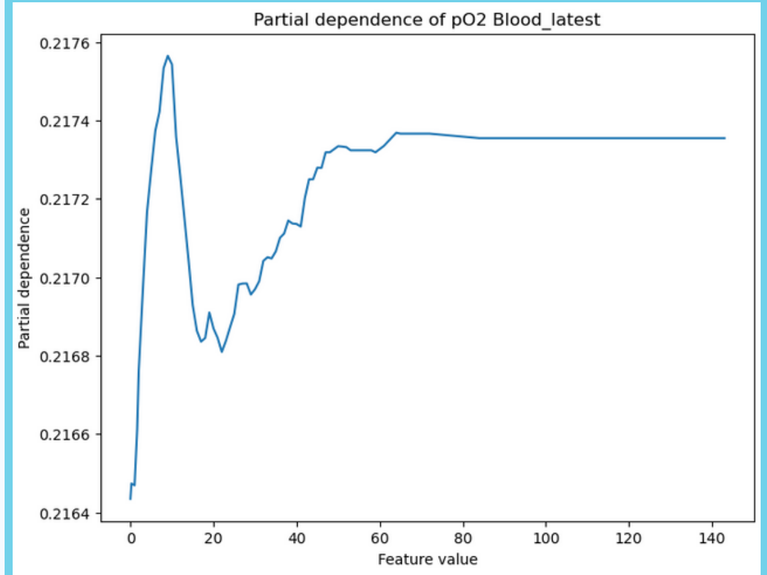
- **Importance**
  Identify significant features

- **Fatal feature**
  PO2_Blood_mean_mode



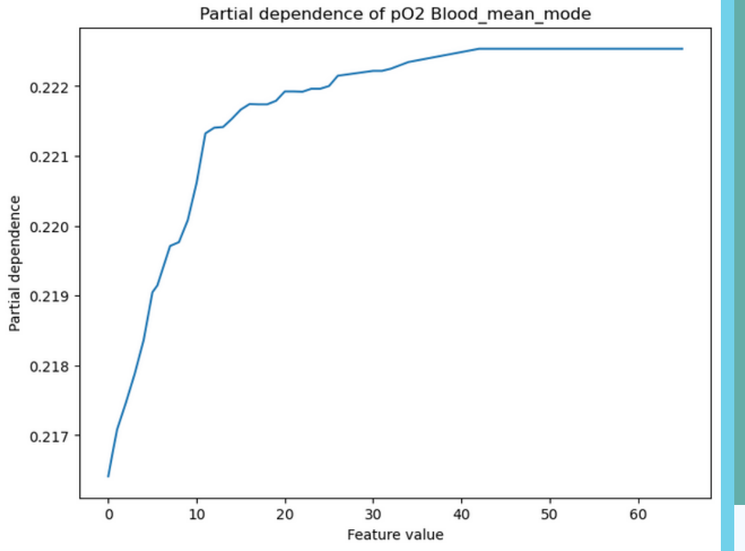
One-way Partial dependence

2

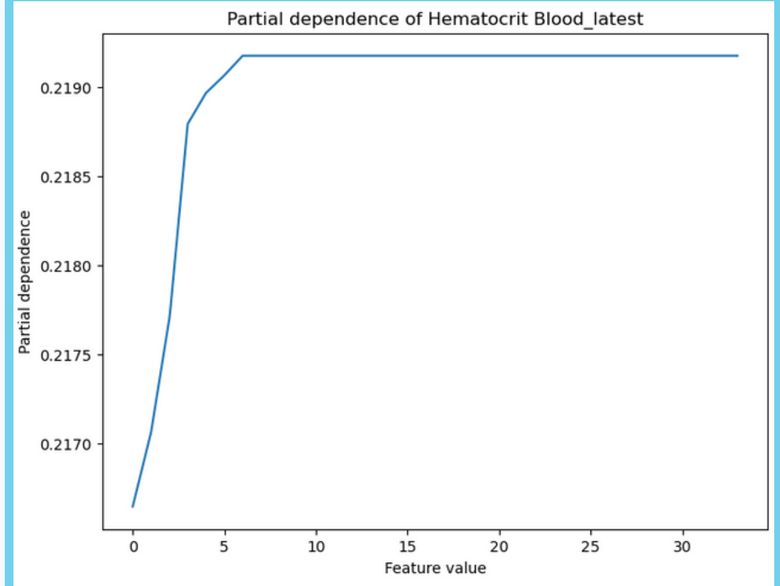


- **Investigate interaction**
  x: individual feature
  y: predicted response

1



3



PART 4

LIMITATIONS & SCOPES OF IMPROVEMENT

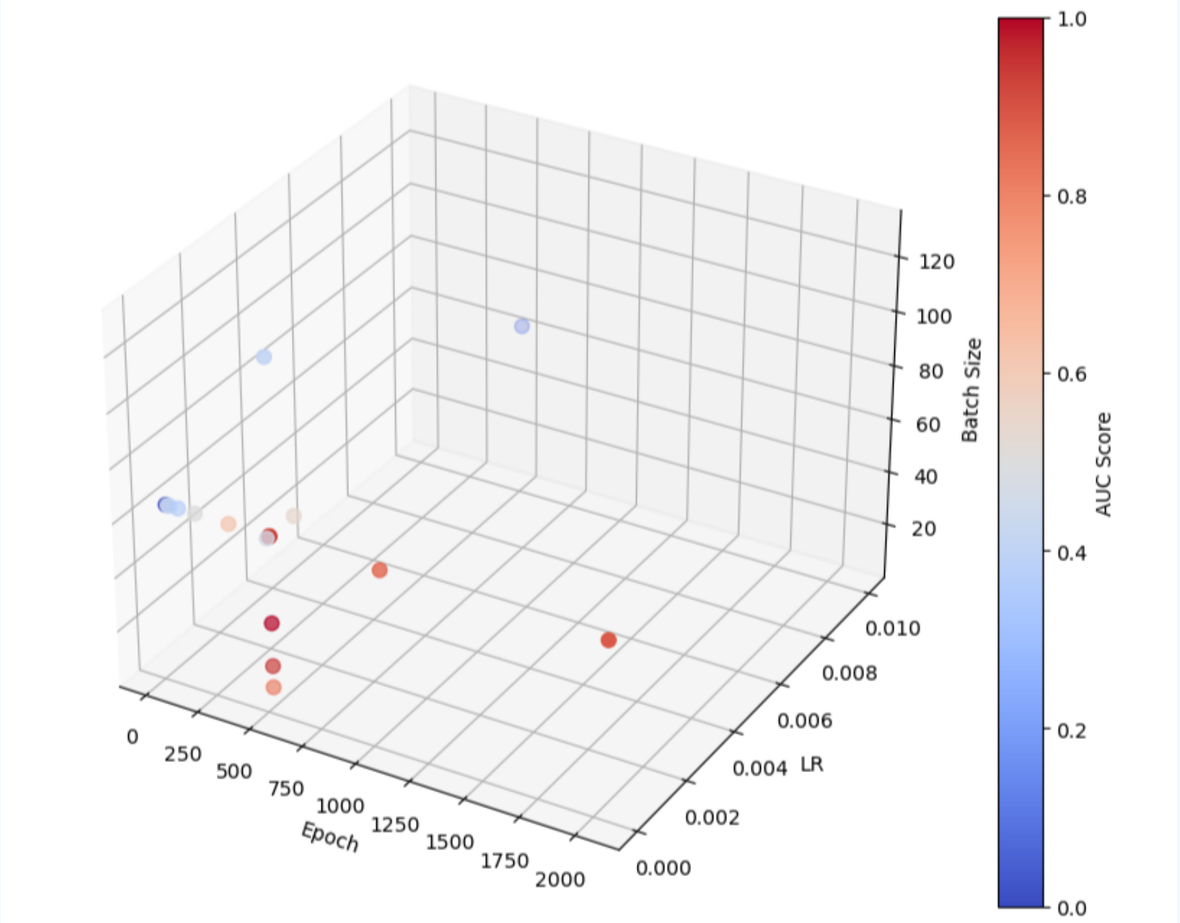
FNN and RNN

Step 1: Construct FNN model

Step 2: Manually tune the parameters

Observation: In the FNN model, the number of hidden layers does not significantly impact the AUC, while the model fitting is very sensitive to parameter changes

Solution: Manually adjust the parameters and record, optimizing in the most feasible direction



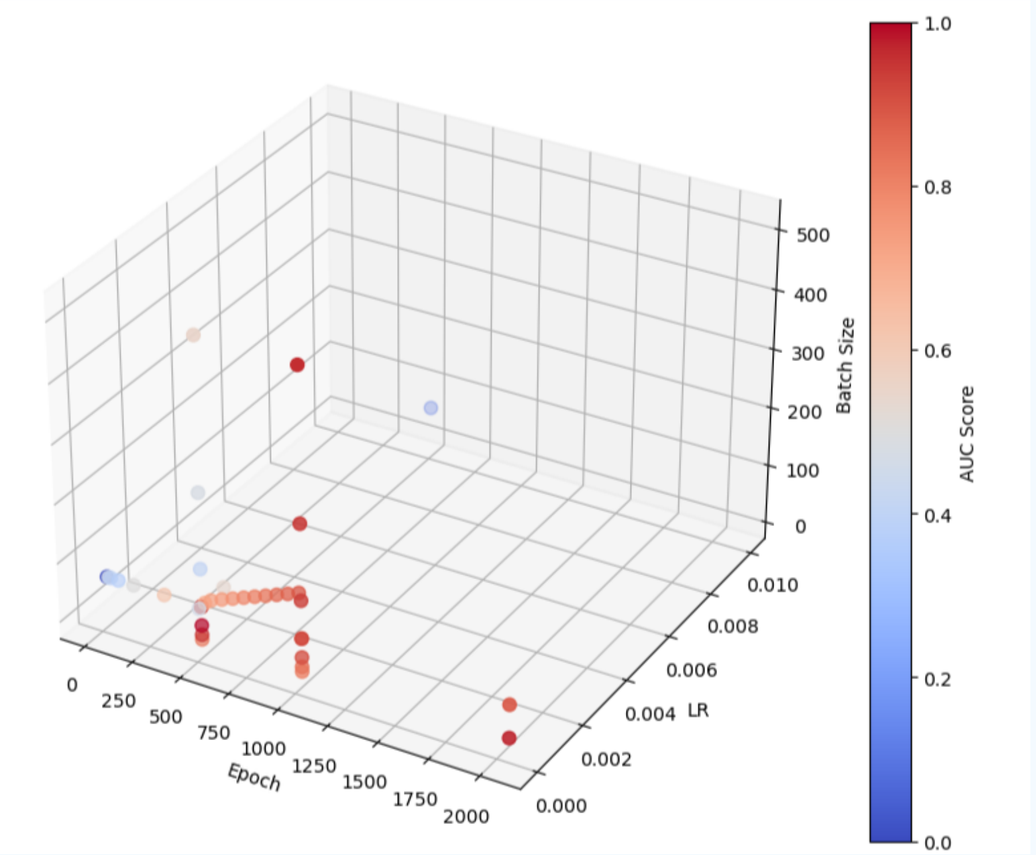
FNN and RNN

Step 1: Construct RNN model (introducing self attention)

Step 2: Hyperparameter tuning (Grid Search)

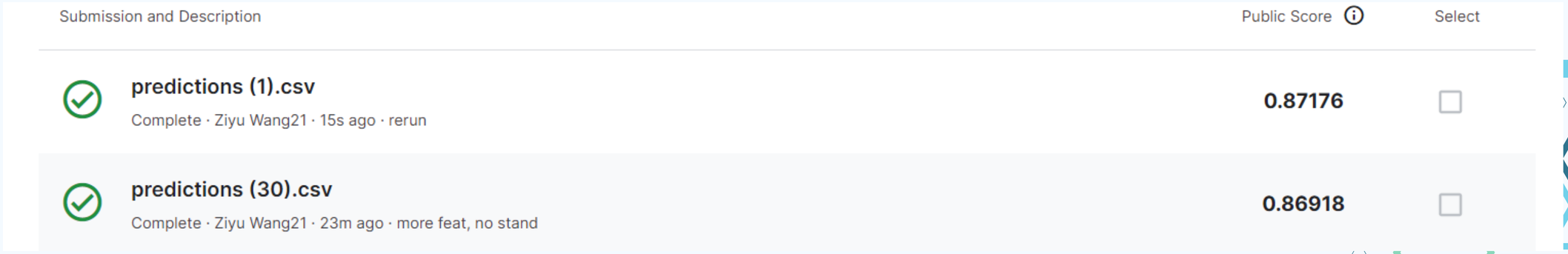
Observation: When manually tuning parameters, the RNN does not perform better than the FNN.

Solution: Tried CUDA acceleration and better GPU for calculation.



Feature Engineering

- Time-related features, such as admit time, death time, and **previous times of admission**, are significant in predicting readmission within 30 days
- **Length of stay** in hospital can be inferred by admit time and death time. By common sense, the longer the stay in the hospital, the more severe the illness, and therefore the higher the probability of readmission.



REFERENCES

Elreedy, D., & Atiya, A. F. (2019). A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance. *Information Sciences, 505*, 32-64.

Guo, C., Lu, M., & Chen, J. (2020). An evaluation of time series summary statistics as features for clinical prediction tasks. *BMC medical informatics and decision making, 20*(1), 1-20.

Johnson, A. E., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., ... & Mark, R. G. (2023). MIMIC-IV, a freely accessible electronic health record dataset. *Scientific data, 10*(1), 1.

THANK YOU

Q & A