**EXECUTIVE SUMMARY**

Discharge of a patient from any ICU of medical institution is regarded as an important part of transferring care to medical service provider, which can be a patient’s family member, an employed caregiver nurse, a friend, or a normal ward.[15] Due to the complexity of medical and non-medical discharge standards, many patients and care providers are still facing difficulties in the discharge process. Although the discharge policy differs among medical providers, one necessary criterion needs to be added is to ensure the safety of patients through discharge and transfer to the next responsible party.[17] The risk of discharging from ICU, including readmission and even mortality, is caused by premature discharge as intensivists make wrong judgement on patients’ health condition and permit them to leave the ICU. On the other hand, late discharge will induce over bed occupancy, unnecessary waste of medical resources, and iatrogenic harm. Indeed, statistics show that post-ICU mortality is high and this situation is nonnegligible.

In this program, machine learning techniques will be implemented to design, construct, and evaluated ICU discharge prediction classifier. The classifier would learn patterns from historical dataset and to predict outcomes in future data. Based on laboratory indicators, medical checks, and patients’ characters, the classifier will give the result to indicate whether the patients are ready to discharge or not. And this would assist intensivists to make the judgement and help medical system to mitigate post-ICU mortality and readmissions, and reach a more efficient status.

The following points are key deliverables:

* A literature review of previous work on machine learning application in ICU discharge
* ICU discharge classifiers with a Random Forest and a Neural Network algorithm
* A comparison analysis between two algorithms with FAT toolkit and data in the dataset

Two machine leaning techniques will be applied to construct the classifiers: an interpretable random forest algorithm which can explain the significance of each variable; a neural network algorithm which is an

algorithmic mathematical model that imitates the behavioral characteristics of animal neural networks and performs distributed parallel information processing.

Electronic healthcare data from MIMIC-III is used as database to train the classifier.[18] MIMIC is a publicly available dataset developed by the computational physiology laboratory of MIT that contains deidentified health data related with approximately sixty thousand intensive care unit admissions.

FAT Forensics which is a Python package is going to be used to evaluate Fairness, Accountability and Transparency of the classifier. Two algorithms, random forest and neural network, can then be compared and analyzed with the FAT Forensics evaluation results. Also, some MIMIC data which has not been trained will be run by the classifier and then evaluated by checking whether the outcomes are consistent with reality.

**INTRODUCTION**

**Background**

An ICU provides the crucial care, including specific therapies for the disease, physiological and life support monitoring, and comprehensive medical decision making, for patients who are suffering from acute illness and injury. Discharging patients from intensive care unit (ICU) is a daily task for intensivists, a vital event for patients and families, and also a significant issue for health care system over the world [1].

Discharging patients from ICU is a complex process and the decision of whether they are ready for discharge is the first as well as the most essential step of the process. For critically ill patients in ICU, the care requirements are quite complex and the discharge transition indicates the degradation of medical environment. Transition from a resource-rich environment to a relatively resource-poor environment means patients have to face a vulnerable period of time. If the discharge occurs prematurely, patients will be readmitted or even face the risk of exacerbating the illness. On the other hand, late discharge would cause unnecessary waste of medical resources, delay admissions of new patients, and the risk of iatrogenic harm.

According to Oxford Textbook of Critical Care, severity of disease for patients is mainly measured by degree of acute illness, physiological reserve of patients, the concurrent level of cure, and the organ life system support [2]. These factors are very difficult to quantify and the patients’ health conditions mainly have to be judged by doctors. There are also scoring systems which are based on large numbers of patient databases helping to inform discharge safely as well as timely. Although scoring systems can acquire more information of patients than a single doctor can obtain in a lifetime, the doctor acknowledges more detailed situation of the individual patient. Therefore, the scoring systems could only act as an assistant character rather than a decision maker.

However, several studies showed that a salient number of patients passed away after discharging from ICU. GiViTI study indicates that post-ICU mortality accounts for 26% of all deaths [2], the Portuguess study shows this proportion is 23% [3], and the North Thames study shows this proportion is 27% [4]. L. Smith et al. suggest that high mortality after ICU discharge is perhaps caused by discharging prematurely [5]. These evidences indicate that the current discharge process is not working efficiently and there need more efforts on minimizing readmissions, patients and family dissatisfaction, unnecessary waste of medical resources, and avoidable morbidity and mortality.

Machine learning refers to the process of using some algorithms to guide the computer to use the known data to get an appropriate model and use this model to predict the new situation. In data analytics field, machine learning is widely used to enhance the decision-making performance [6]. It is practical to implement machine learning methods in ICU discharge process.

In this project, ICU discharge classifiers will be designed, constructed, and tested based on the implementation of machine learning techniques to optimize the problem of ICU discharge problem.

**Aims and Objectives**

The general aim of this project is to build ICU discharge classifiers with a novel machine learning technique to assist the ICU discharge process. The classifier can analyze medical indicators of ICU patients and give the conclusion that whether patients are ready to discharge or not.

The purpose of the project is to applicate a machine learning technique which is Random Forest from a novel perspective to mitigate the ICU discharge problem. To be specific, the aim is to improve the prediction of patient discharge outcome based on attributes including laboratory indicators, medical checks, and patients’ characters.

To accomplish this aim, electronic healthcare dataset from MIMIC-III will be used. And the outcome will be evaluated by a Python toolkit which is Fairness, Accountability, and Transparency (FAT) Forensics.

The following points are project objectives:

* Finish a machine learning online course to acknowledge the foundation of machine learning.
* Be familiar with Random Forest algorithm and a way to get path for a single prediction made by a Random Forest.
* Construct ICU discharge classifiers based on the MIMIC-III electronic healthcare database.
* Apply Neural Network algorithm and a special Random Forest algorithm which can be get the path for a single prediction.
* Be familiar with FAT Forensics
* Evaluate the outcome with FAT toolkit and data in the dataset.

The following points are key deliverables:

* A literature review of previous work on machine learning application in ICU discharge
* ICU discharge classifiers with a Random Forest and a Neural Network algorithm
* A comparison analysis between two algorithms with FAT toolkit and data in the dataset

Currently, clinical judgement is the primary factor in deciding whether patients are ready to discharge from ICU or not and high post-ICU mortality suggests that such way is not working efficiently. On the other hand, late discharge will cause high ICU bed occupancy with unnecessary medical resources waste. Building a classifier with machine learning techniques can help a lot with clinical judgement to discharge patients at right time and help the healthcare system work more efficiently with lower mortality and readmissions.

**LITERATURE REVIEW**

Machine learning technology is applied in the area of medical science widely and effectively. For example, there is a study investigating the determinants which are affecting the length of stay of patients and improving the efficiency of medical resources usage.[7] Moreover, machine learning approaches can notably play a role in the aspect of measuring patient discharge. Because on one hand, the health resource is finite, which requires discharging patients in due course so that those successors can obtain medical care. However, on the other hand, this process is complicated. How to determine a patient should be discharged becomes a crucial issue and merely depending on doctors is a waste of time and full of loopholes. Hence, the application of machine learning technology provides clinicians a crucial reference as well as simplifies the burden of decision-making. But focusing on the practice, it still needs to be optimized and there are some aspects being disputed.

**Criteria**

There are two sorts of obstacles during the process of discharging a patient, resolving into clinical and nonclinical. [8] On the clinical aspect, clinicians’ subjective decision can not avoid frequent faults, especially with such heavy pressure, which will lead to the wrong estimation of the patients’ condition.[9] Focusing on the other aspect, in face of growing need for health facilities and morbid administrative and clinical missions, the best timing for discharging patients can not be in control correctly. The consequence is the violation of the rule of resource optimization. These facts show the practical tendency for the application of machining learning by data analysis to accomplish complex tasks that are out of men’s ability. However, concerning the criteria of measuring the patient discharge, selection and evaluation are significant issues.

First of all, clinical criteria is the core reference for machine to make final decision. In *Patient Discharge Classification Using Machine Learning Techniques*, the authors pick seven important aspects as the basis of subsequent algorithm. It includes patient’s internal and surface temperature, blood pressure with these three’s stability and oxygen condition. Another work about this subject quotes the nurse-led discharge (NLD) put forward by Knight in 2003. It extends the test range to 15 names as well as fractionizes the item, with five respiratory items, two cardiovascular, one pain, one temperature, one central nervous system and five blood covered.[10] Besides for clinical one, patient’s reaction to discharge is also listed as one vital no-clinical criteria, which is called Ready-for-discharge (RFD) and not-ready-for-discharge (NRFD). Therefore, patient discharge becomes bidirectional result. By setting specific standard for every item, machine learning makes a yes-no decision of each variable, and then elicits overall conclusion. However, the problem is that rare patients can meet all norms so that they are allowed to discharge. In this project, scoring system will be adopted: assuming that there are some items rigid that a patient must achieve, then about the rest, scoring mechanism will be imported according to different conditions. Ultimately, the final value is calculated by a weighted average. If it reaches the given score, a man is permitted to discharge. Another issue is the database of the tester. Universally, the origin of data is basically from the university or institutions database, which collecting from hospitals. Anthony’s work adopts the data from the University of California Irvine (UCI), mainly monitoring on 90 cases, while *Towards a decision support tool for intensive care discharge: machine learning algorithm development using electronic healthcare data from MIMICIII and Bristol, UK* extends the number of cases with two kinds of database: Bristol Royal Infirmary general intensive care unit and Medical Information Mart for Intensive Care-III, totally 9402 patients are considered. Another Kyan’s work innovatively analyzes the cases with updated data for every 24 hours.[11] But in a word, the source of the data is not the most important thing, the universality, reliability and representativeness must be taken into account.

**Algorithm**

The successful application of machine learning depends on accurate computing methods. Algorithm is the key element that determines the quality of the prediction model and there are some methods have been implemented and compared.

手机屏幕截图

描述已自动生成

Figure 1 Basics illustration of discharging a patient

Anthony et al. apply three machine learning algorithms which are Decision Trees, Random Forest and Bayesian Network. Decision Trees construct a model with three key elements including a root, a subset, and a leaf. The root is the beginning place where the original dataset is split into two or more subsets. This process will be repeated until a leaf node is found and the depth of the tree is affected by the criteria of the algorithm.[12] Figure 1 shows the fundamental explanation of Decision Trees in discharging a patient. Random Forest algorithm can be seen as a forest consisted with Decision Trees. A cluster of trees is called ensembles and the result of prediction is the outcome of various trees’ majority vote. Bayesian Network is a part of probabilistic graphical models which are applied to illustrate knowledge when facing an uncertain domain. Directed Acyclic Graph is one of its representation to learn the properties of a group pf random variables and multiple groups conditional probability distributions. For instance, the Bayesian Network can be used to express the probabilistic relationship between disease and its related symptoms, if a certain symptom is known, the Bayesian Network can be used to calculate the probability of occurrence of various diseases. Within the range of three ways: Decision Trees, Random Forest and Bayesian Network, compared with two others, Random Forest reveals the highest error rates among three. And Bayesian Network owns the lowest error rate as well as the highest precision rate. But on the whole, machine learning algorithm shows positive tendency for the practical application in the area of patients discharge. Notably, their decisions about remain in the hospital is nearly accurate and stable, but on the aspect of discharge, it remains to be observed because of the shortage of the subjects.[13]

Research of Chris et al. implements two machine learning methods: Random Forest and Logistic Classifier. Logistic Classifier implies logistic regression method which is used to model the probability of a certain class. The result is that although on Bristol Royal Infirmary general intensive care unit (GICU) Logistic Classifier outperforms Random Forest, on another database Medical Information Mart for Intensive Care (MIMIC)-III, Random Forest is the better one and overall, Random Forest shows more positive prospect.

Kyan et al. develop a predictive model to aid discharge processes with a multiplayer perceptron feedforward Neural Network approach. The discharge process is related with numbers of clinical and nonclinical variables interacted in a complex and extremely nonlinear way, and Neural Network is suitable for dealing such situation. [14] In addition, online learning can be realized with Neural Network model, so that the model would continuously update itself with new data available online which could then improve the prediction accuracy. The outcome is indicated as a continuous score from 0 to 1 representing the possibility of discharging the patient in 2 hours. The result shows the Neural Network model significantly performs better than the baseline model from sensitivity (52.5% vs 56.6%) and specificity (51.7% vs 82.6%) perspectives.[14] The model is also used to identify discharge barriers by toggling the variables and recalculating the score.

**PROPOSED APPROACH**

**Criteria**

The importance of criteria can not be neglected, reasonable criterion lays the solid foundation for success in ultimate outcome, which clinicians and patients both approve. Once some key criteria are omitted or ignored, the reliability of the outcome will be low and the process instead triggers the loss the sense of its existence. This project uses Bristol Royal Infirmary general intensive care unit (GICU) dataset with twenty three medical indices, including creatinine, partial pressure of oxygen, fraction of inspired oxygen, partial pressure of carbon dioxide, minimum and maximum blood pressure, pain level, potassium, minimum and maximum heart rate, minimum and maximum Glasgow Coma Scale, blood urea nitrogen,bicarbonate, airway, minimum and maximum respiratory rate, haemoglobin, minimum and maximum peripheral capillary oxygen saturation, minimum and maximum temperature, and sadium.

**Algorithm**

Logistic Regression, Random Forest which can be get the path for a single prediction, and Neural Network are three algorithms going to be applied and compared in this project.

**Logistic Regression**

Basically, Logistic Regression (LR) is a machine learning method for solving binary (0 or 1) problems, which is used to estimate the possibility of incidents. The hypothetical function of LR is as follows:

The curve of Sigmoid function (g(z)) is as follows:

图片包含 截图, 游戏机

描述已自动生成

It can be seen that Sigmoid function is an S-shaped curve with value between 0 and 1 and this trait is very important to solve binary classification problems as the outcome value is constrained in 0 and 1. Then, a cost function would be used as measurement criteria to train vector which represents model parameters so that the output values of the model can be as consistent as possible with the real ones. The LR cost function is as follows:

Different algorithms can be implemented to find the smallest value of cost function and gradient decent algorithm is a relatively simple and effective one. Gradient decent algorithm:

Finally, the model would be built with best parameter vector found.

**Random Forest**

A random forest is based on decision tree algorithm. Decision tree is a very command classification or regression method of supervised learning which means the classification results have been accessible. Decision tree is a kind of tree structure, in which each internal node represents a judgment on an attribute, each branch represents the output of a judgment result, and finally each leaf node represents a classification result. A random forest model is consisted of a multitude of decision trees and the prediction result is made by voting or mean prediction of these decision trees [rf 12]. Compared with decision tree, the excellent performance of random forest mainly depends on random sampling, which makes it more resistant to overfitting, and ensemble algorithm, which equips it with higher accuracy [rf 3].

Typically, Random Forest is regarded as a black box which is viewed in terms of inputs and outputs without the perception of internal working process. As a matter of fact, a forest is consisted of a great number of decision trees and each decision tree is trained with data using features which are randomly selected. Therefore, it is infeasible to gain a complete understanding of the internal decision process by checking each decision tree. Furthermore, a decision tree with 10 depth could contain thousands of nodes which makes it almost unrealistic to build an interpretive model.

One way to get an insight into random forest is to calculate the importance of features by arranging the value of each feature one by one and checking how it changes the performance of the model, or calculating the number of ‘impurities’ which is usually depending on the variance of the regression trees and Gini coefficient or entropy of classification trees. Each feature is deleted when it is used in a node. Both methods are useful, but they are both crude and static because they have little insight in understanding individual decisions on actual data.

There is a method to turn random forest which is the black box into a white box by identifying its decision paths. Within a decision tree, there is a path or paths from the root to the leaf or leaves when a decision is made. Each decision is formed by a series of decisions, which is protected by specific features, and each decision makes contribution to the final prediction outcome. Suppose a decision tree has M leaves and the feature space is divided into M regions (Rm, 1<=m<=M). Classically, a tree’s prediction function is defined as

M is the number of leaves of the tree and Rm represents the region in the feature space which is corresponding to leaf m. Cm is a constant relating with region m and its value is determined during the training period of the tree. In case of a regression tree, the value corresponds to the average value (or ration in the case of a classification tree) of the response variables of samples belonging to the Rm region. Finally, *I* is the indicator function. If x belongs to region Rm, *I* returns 1, otherwise returns 0. This is a concise definition and could gain the meaning of tree with underlying method that the value will be returned when correct leaf is reached. However, the operational part of the decision tree is ignored which indicates that the decision path and available information are not taking effect.

In practice, the prediction model can be interpreted in a more operational way by a sequence of regions corresponding to each node of the tree. On the account that each decision will add or subtract the value which is inherited from parent node, the prediction outcome can be calculated as the summation of the feature contribution plus the deviation.

K indicates the number of features and C full is the value at root node. For a decision tree, each feature’s contribution is not a fixed predetermined value. The rest of the feature vectors that determine the decision path through the tree would affect the value. Therefore, the contrib(x,k) is calculated as contribution of each feature in the feature vector x.

Then, it is straightforward to move from decision tree to random forest as the prediction of a forest can be viewed as the average of predictions of trees.

J is the number of trees in the forest.

This is a unique way to interpret random forest internal working process and it would be worthwhile to implement it to see how each variable works in discharge patients from ICU.

**Neural Network**

Neural networks are invented by imitating neurons or neural networks in the brain. Human brain is full of neurons and figure below shows the structure of a single neuron cell [nn wiki]. Neurons have a certain number of input nerves which are called dendrities. They can be regarded as input wire and intended to receive information from other neurons. The output nerves of neurons are called axons which are used to transmit signals or information to other neurons. In short, a neuron is a computing unit that receives information from the input nerve and do some calculations, and then transmit the results through its axons to other nodes or other neurons in the brain.

图片包含 游戏机

描述已自动生成

The neural network model is based on many neurons and each neuron is a learning model. These neurons, which are also known as activation units, take some features as input and provide an output with their own calculation. The figure below is an example of neural network model. x1, x2, x3 are input units to which the raw data is put. a1, a2, a3 are intermediate units which are responsible for processing the data and passing it to the nest layer. Finally, the output unit is responsible for calculating . Neural network model is a network of many logic units organized according to different levels. The output variables of each layer are the input variables of the next layer. In the figure below, layer 1 is called input layer, thee middle layer which is layer 2 is called hidden layer, and layer 3 is called output layer. The hidden layer is between the input layer and the output layer. It is called hidden layer because the value generated in it is not directly visible as the sample matrix X used in the input layer or the label matrix y used in the output layer.

图片包含 物体, 钟表

描述已自动生成

Each layer of the neural network contains several neurons, and the neurons between the layers are connected by the weight matrix . An information transmission process can be described as follows:

* Layer J’s neurons receive upper afferent stimulation which is also known as nerve impulses.
* After the stimulation is acted by the activation function g, an activation vector is generated. is the activation value obtained by the i-th neuron in layer J

Stimulation is transmitted from the previous layer to the net layer, so it is called forward propagation.

For non-linear classification problems, logistic regression will use polynomial expansion features, resulting in the emergence of feature vectors with huge dimensions. In neural network, the dimension of features will not be increased, which means the scale of input layer of neural network will not be expanded. Instead, the features will be continuously optimized by increasing hidden layer and correcting weights in hidden layer. The activation value which is produced by forward propagation process can be regarded as the optimized features.

图片包含 游戏机, 物体, 手表, 钟表

描述已自动生成forward propagation

Cost function of neural network is more complex than logistic regression’s because there is only one dependent variable and one output variable. However, in the neural network, there could be many output variables. is a vector with dimension k and the dependent variable in training set is also a vector of the same dimension.

Cost function:

* *L = the total number of layers of the neural network*
* *= the number of neurons of layer l*
* *= the number of neurons in output layer (the number of categories)*

The neural network allows multiple hidden layers and the neurons in each layer can predict the output. Therefore, it is not reasonable to directly use the gradient descent method of the traditional regression problems to minimize . It is necessary to consider each layer’s prediction error and then optimize them. Back propagation, which is going to firstly calculate the error of last layer and calculate the error of each layer in reverse direction until the penultimate layer.

Back propagation:

图片包含 游戏机, 物体, 手表, 钟表

描述已自动生成back propagation

With all the errors () available, the partial derivative of cost function can be calculated as follows:

In all, there are seven steps to build a neural network model:

* Selecting model structure including input layer, hidden layer, and output layer.
* Randomly initializing the parameters. If all the initial parameters are 0 or a same number, it means all the activations units in the second layer will have same value. Therefore, the initial parameters should be generated with random function.
* Calculating each layer’s with forward propagation method
* Calculating cost function
* Calculating all the partial derivatives with back propagation method
* Checkout the partial derivatives with gradient descent method. A method called numerical gradient checking is used to check whether the partial derivates are conformed with requirements. When the parameter is a real number, the derivative of the cost function at a certain point is . The surrogate values of and (is a very small value, usually 0.001) and their average value are going to be calculated. It is considered to be rational if .
* Use optimization algorithms to minimize the cost function

**Model Construction**

A GitHub repository is built to record the process of code writing which is available at <https://github.com/benjin1998/finalproject>.

**Preprocessing**

The data is read in with pandas and the empty properties are filled with mean value of corresponding feature. Then, the dataset is split into train set, which accounts for eighty percent of whole dataset, and test set, which accounts for twenty percent of dataset.

Standard scale method is implemented in the logistic regression and neural network models building. The function of standard scale is to remove the mean and normalize the variance and it is for each feature dimension, not for the sample. With standard scale, the processed data will conform to the standard normal distribution, which means the mean value is 0 and the standard deviation is 1. Standard scale mainly has two advantages: it will speed up the speed of finding the optimal solution by gradient descent algorithm, otherwise, the model will be difficult to converge or even fail to converge; it is possible to improve the accuracy (wiki). Probability model, such as decision tree and random forest, do not need to implement standard scale, because the distribution of variables and the conditional probability between variables is relevant with model’s construction and outcome instead of the value of variables. However, standard scale method is important for the optimization problems with algorithms such as AdaBoost, SVM, logistic regression, KNN, neural networks (wiki). Therefore, this project implements standard scale method in logistic regression and neural network models and not in random forest model.

**Gr****idSearchCV Method**

GridSearchCV method can be divided into two parts which are grid search and cross validation.

In machine learning models, the parameters that need to be selected manually are called hyper parameters. For example, the number of decision trees in a random forest, the number of hidden layers and nodes in each layer in the neural network models, all need to be specified in advance. If the hyper parameters are not selected properly, the problem of under fitting or over fitting will appear. Grid search is a method of parameter tuning. A parameter grid is the place where the range of each hyper parameter set manually. Then, grid search method will train each combination of hyper parameters in the parameter grid through loop traversal and the hyper parameters which has best performance will be the final result.

Cross validation is a common method in machine learning to build models and verify model parameters. Cross validation will use data repeatedly, the obtained sample data are divided into different training sets, which is used to train the model, and test sets, which is used to evaluate the quality of the model prediction. On this basis, different training set and test sets can be obtained and a sample in training set could become a sample in test set next time. The function of cross validation is to reduce over fitting to a certain extent and get as much valid information as possible from the limited data [wiki].

There are different cross validation methods, such as k-fold cross validation, leave one out cross validation and so on. K-fold cross validation is used to verify models in this project. It divides the original data into k-fold groups. A validation set is made for each subset data and takes the reaming k-1 subset data as the training set, so as to get k models. The k models are evaluated in the validation set and the mean squared error will be calculated and get the cross validation error.

GridSearchCV method combines grid search and cross validation to train the model with each group of hyper parameters, and the hyperparameters with minimum error of validation set will be selected. The main disadvantage of the method is that it could be time consuming. The more hyper parameters, the more candidate values, and result in the longer time consumption. For example, suppose there were n hyper parameters and each hyperparameter has m candidate values, the GridSearchCV will train models which could cost a large amount of time if m and n are large numbers.

RandomizedSearchCV is a method to solve such delimma. Compared with GridSearchCV, not all candidate hyper parameter settings are going to be tried out, but rather a fixed number of hyper parameter settings is sampled from the specified distributions. Furthermore, Bergstra and Bengio’s [2012] research indicates that RandomizedSearchCV method is more efficient than GridSearchCV method.

**Logistic Regression Algorithm**

Firstly, dataset is imported and preprocessed with standard scale method and split into train dataset and test dataset. A pipeline function named PolynomialLogisticRegression is built to control the polynomial degree of features.

图片包含 游戏机, 刀, 桌子

描述已自动生成

Then, an original logistic regression classifier is built with logistic regression model from sklearn library. All the hyper parameters are default and polynomial degree is 1. The original classifier is fit on the train dataset and the result shows that the accuracy is 0.761 which is verified on test dataset.

Next, the other classifier, which is named GSCV\_LR classifier, is built with GridSearchCV method. The tuned hyper parameters are penalty, which is used to specify the penalization norm, and C, which is the inverse of regularization strength. K-fold cross validation method is implemented with StratifiedKFold method from sklearn library. The train dataset is divided into ten folds and shuffle each group’s samples before splitting into batches. The GSCV\_LR classifier is trained on train dataset and result suggests the best hyper parameter setting is C=0.1, penalty:’l1’, and degree 1 of feature polynomial. The accuracy of GSCV\_LR classifier is 0.778.

**图片包含 游戏机

描述已自动生成**

**Random Forest Algorithm**

Firstly, an original random forest classifier is built with RandomForestClassifier method imported from sklearn library. All the hyper parameters are default and train the classifier on train dataset. The accuracy of original classifier is 0.788.

Then, the other classifier, which is named GSCV\_RF classifier, is built with RandomizedSearchCV method. The tuned hyper parameters are n\_estimators, which represents the number of decision trees in the forest, max\_depth, which indicates the maximum depth of the tree, min\_samples\_split, which is the minimum number of samples that are required to split an internal node, min\_samples\_leaf, which is the minimum number of samples required to be at a leaf node, max\_features, which suggests the number of features to consider when looking for the best split, and bootstrap, which shows whether bootstrap samples are used when building trees. Then, the GSCV\_RF classifier is trained on train dataset and result shows that the best hyper parameter setting is n\_estimators = 1200, min\_samples\_split = 2, min\_samples\_leaf = 4, max\_features: ‘auto’, max\_depth: None, bootstrap: Flase. The accuracy of GSCV\_RF classifier is 0.779.

A package for interpreting scikit-learn’s decision tree and random forest predictions, which is named treeinterpreter, is imported to decompose each prediction into bias and feature contribution components. A function named rf\_explain is designed to explain a patient case with treeinterpreter package and store the output information including prediction, bias, and contribution in a data frame.

图片包含 游戏机, 截图

描述已自动生成

Then, two functions, which are visualiee\_1\_patient and compare\_two\_patients are designed to visualize the interpretation outcome. Visualize\_1\_patient function has a data frame incoming variable, which could be derived from rf\_explain function, containing the output information with treeinterpreter package. Then the effects of features, which are medical indicators, will be visualized in the format of line chart. Compare\_2\_patients function can be seen as visualizing the output information with treeinterpreter package of two patients at the same time so that the condition of two patients can be compared.

The aim of two functions is to inform the doctors how each feature is affecting the patients’ possibility to discharge. Compared with pure numbers, visualizing with line chart could help doctors acknowledge health condition of patients more intuitively and then make further treatment plans to help them discharge from ICU.

图片包含 游戏机, 文字, 地图

描述已自动生成

Figure illustrates the effects of features of an example patient selected from test dataset. The classifier prediction result is negative with prediction value of 0.09 which means fail to discharge. It can be seen that gcs and airway are two most significant factors that encumber patient’s discharge. Doctors could analyze health condition of patient according to the visualized information and make further treatment plans to improve disadvantage medical performance emphatically.

图片包含 文字, 游戏机, 地图

描述已自动生成

In this figure, the other example patient, who is predicted able to discharge with prediction value of 0.95, is introduced and analytical outcome of two patients are plotted at the same time. The prediction results of two patients are extremely different and the figure could help doctors analyze the significance of medical indicators that are influencing discharge of two patients. For this example, it can be estimated that gcs, airway, hr, and haemoglobin are the main different between two patients and finally lead to the huge gap between their opposite prediction result.

One potential implication of such method is that doctors could implement compre\_2\_patients function to investigate the performance of different treatment strategies on two ICU patients with similar symptoms. For instance, the whole world is suffering from COVID-19 at present and there is no known effective treatment strategy. Doctors could adopt different medical strategies to different COVID-19 patients in ICU. After a period of time, compre\_2\_patients function could be employed to analyze the health condition of patients and evaluate the effectiveness of corresponding treatment strategies. In this way, compare\_2\_patients function would be contributing tool to support medical works.

**N****eural Network Algorithm**

Dataset is imported and preprocessed with standard scale method and split into train dataset and test dataset.

The original classifier is firstly built with MLPClassifier method which is imported from sklearn library. All the hyper parameters are default and the classifier is fitted on train dataset. The accuracy of original classifier is 0.763.

Then, the other classifier, which is named GSCV\_NN classifier, implements RandonmizedSearchCV method is constructed. There are five tuned hyper parameters that are hidden\_layer\_sizes, which describes the structure of hidden layers, activation, which is the activation function for the hidden layer, solver, which indicates the solver for weight optimization, alpha, which represents L2 penalty regularization term parameter, and learning rate, which suggests the learning rate schedule for weight updates. The classifier is fitted on train dataset and the outcome shows the best hyper parameter setting is solver: ‘adam’, which refers to a stochastic gradient-based optimizer, learning rate: ‘constant’, hidden layer sizes: (50, 50, 50), alpha=0.0556, activation: ‘logistic’. The accuracy of GSCV\_NN classifier is 0.797.

**P****erformance Evaluation**

Apart from accuracy, this project also uses also criteria to evaluate classifiers’ performance.

**Accuracy, Precision and Recall**

The prediction outcome could be shown as following table.

|  |  |  |
| --- | --- | --- |
|  | Truth: Positive | Truth: Negative |
| Prediction: Positive | True Positive | False Positive |
| Prediction: Negative | False Negative | True Negative |

Precision is defined as positive predictive value which can be calculated as the number of true positive divided by the sum of number of true positive and number of false positive [wiki].

Recall is defined as true positive rate or sensitivity which can be calculated as the number of true positive divided by the sum of number of true positive and number of false negative [wiki].

F1 score, which is also called balanced F score, is a criterion to evaluate the accuracy with precision and recall. The F1 score is the harmonic mean of precision and recall and the best F1 score reaches 1 with perfect precision value of 1 and recall value of 1. The F1 score is calculated as:

The accuracy, precision, recall, and F1 score performance is integrated in the table below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Accuracy | Precision | Recall | F1 Score |
| Original LR | 0.761 | 0.74 | 0.77 | 0.755 |
| GSCV LR | 0.778 | 0.744 | 0.815 | 0.778 |
| Original RF | 0.788 | 0.741 | 0.816 | 0.777 |
| GSCV RF | 0.779 | 0.728 | 0.816 | 0.769 |
| Original NN | 0.763 | 0.719 | 0.756 | 0.737 |
| GSCV NN | 0.797 | 0.76 | 0.788 | 0.774 |

As implemented in ICU discharge field, precision means the proportion of the number of patients who are predicted correctly to discharge to the number of patients who are predicted to discharge, and recall means the proportion of patients who are predicted correctly to discharge to the total number of people who should be discharged in the test dataset. Therefore, the classifier with higher precision suggests more patients’ health condition would be stable after discharge and the lower the post ICU mortality would be. The classifier with higher recall indicates more patients who are healthy enough to discharge would be discharged timely so that medical resources could be distributed more efficiently.

In the built classifiers, neural network algorithm with RandomizedSearchCV method has the best accuracy – 0.797 and precision – 0.76 performance, random forest algorithm works best in recall, and logistic regression with GridSearchCV method has the highest F1 score – 0.778.

**ROC Curve and AUC Score**

The ROC curve, which is the abbreviation of receiver operating characteristic, explains the diagnostic ability of a binary classifier [wiki]. The ROC curve is drawn by plotting the true positive rate as y axis and the false positive rate as x axis with different threshold values.

True positive rate indicates the proportion of the number of true positive to the number of all positive samples and the calculation formula is:

False negative rate is the proportion of the number of false positive to the number of all negative samples and the calculation formula is:

The AUC score, which is shortened from area under curve, is the area under ROC curve. The AUC could be regarded as a probability value. When a positive and a negative sample are randomly selected, the probability that the classifier would rank the positive sample in front of the negative sample according to the calculated score is the AUC value. The larger the AUC value, the more likely the classifier would rank the positive sample in front of the negative sample, so as to be a better classifier.

The ROC curve and AUC score of each algorithm is shown in below figures:

图片包含 游戏机, 地图, 文字

描述已自动生成LR

图片包含 游戏机, 文字, 地图

描述已自动生成RF

图片包含 游戏机, 地图

描述已自动生成NN

Each figure above show ROC and AUC of three classifiers including original classifier on test dataset, original classifier on train dataset, and GridSearchCV classifier on test dataset of corresponding algorithms. For all three algorithms, original classifiers on train dataset gain the best performance especially for random forest and neural network algorithms, the AUC of corresponding classifiers reach 1 which means all the prediction are correct. The phenomenon that original classifiers on train dataset perform much better than original classifiers on test dataset reveals that overfitting problem generally exists for all three algorithms. As explained above, GridSearchCV and RandomizedSearchCV methods are helpful to mitigate the overfitting problem and also, it could be inferred from the figures that the classifiers with GridSearchCV or RandomizedSearchCV method have better ROC performance than the original classifiers which indicates the methods function to find the best hyper parameter settings and improve classifiers’ performance.

Overall, the neural network classifier with RandomizedSearchCV method reaches the highest AUC of 0.88 and the logistic regression classifier with GridSearchCV methods and random forest classifier with RandomizedSearchCV methods have the same AUC of 0.87.

**FAT Forensics**

FAT Forensics is a Python toolkit used for evaluating Fairness, Accountability and Transparency of artificial intelligence systems. The FAT Forensics can act as a part of a data processing pipeline to analyze a machine learning system in pre-production. Figure 2 shows the structure and main techniques implemented by FAT Forensics.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Fairness | Accountability | Transparency |
| Data/Features | Systematic Bias (disparate treatment labelling)  Sample size disparity (e.g. class imbalance) | Sampling bias  Data Density Checker | Data description |
| Models | Group-based fairness (disparate impact) | Systematic performance bias | Partial dependency  Individual conditional expectation |
| Predictions | Counterfactual fairness (disparate treatment) |  | Counterfactuals  Tabular by LIME |

Figure 2 Illustration of FAT Forensics

Data Transparency

Dataset is grouped depending on the outcome, which in this project is that 0 - unable to discharge and 1 - qualified to discharge and then, the transparency of dataset is inspected. With fatf.transparency.data.describe\_functions function, the distribution of all features could be displayed as numeric statistic.

图片包含 游戏机, 截图

描述已自动生成

Figure is shown as an example to display distribution information of feature creatinine. Therefore, there are 3957 data points in the applied GICU dataset and creatinine’s value is ranging from 11 to 988. The mean of feature creatinine is 89.2 while the median is 69. Next, the distribution information of each feature in the group with outcome 0 and in the group with outcome 1 could be displayed with same method.

图片包含 游戏机, 截图

描述已自动生成

图片包含 游戏机, 截图

描述已自动生成

The figures indicate the distribution of feature creatinine grouped on prediction outcome 0 and 1. For patients who fail to discharge, the creatinine ranges from 11 and 695, and for patients who successfully discharge, the creatinine ranges from 13 and 988, which suggests that overlap exists between them. The mean of group with outcome 0 is 94.2 and it is larger than the mean of group with outcome 1 which is 83.3.

The data description method is adopted to all the features and the mean values of each feature with prediction outcome 0 and 1 is visualized in the figure below.

图片包含 游戏机, 地图

描述已自动生成

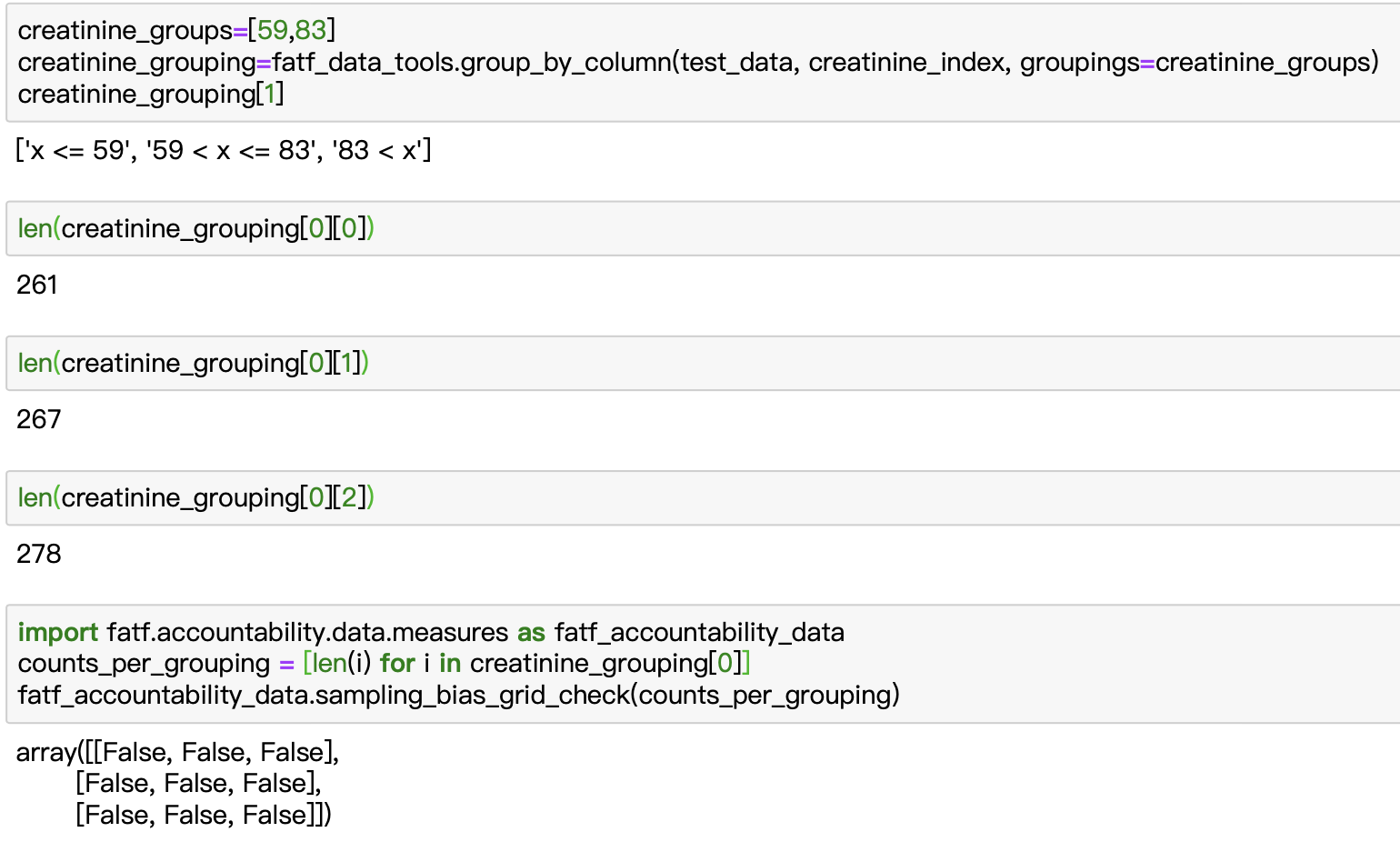
It can be seen from the figure that the gap between the mean of each feature with outcome 0 and 1 is very small, however, there are also some features existing evident difference between two groups including creatinine, fio2, hr, gcs, resp, and haemoglobin. The mean statistics could be provided to doctors and help them to judge whether the features with significant gap between two groups are the factors that impacting discharge of patients.

Fairness of Data and Model --- gender

Robustness of Data and Models

…

First step of examining the robustness of data and models is to group the dataset regarding features. Feature creatinine is taken as the example to demonstrate how the process works. The test dataset is split into three groups with split values of 59 and 83 so that the sizes of three groups are close to each other. The code is shown in the figure below and the size of group whose creatinine value is smaller than 59 is 261, the size of group whose creatine value is between 59 and 83 is 267, and the size of group whose creatinine value is larger than 83 is 278.



The number of data points for three groups are roughly equal. For completeness, the fatf.accountability.data.measures function is implemented to check whether sampling bias exists. The figure above shows how the function is used and the outcome indicates that none of the group violates the sampling bias criterion. The conclusion could then be made that creatine is a robust feature.

Following step is to evaluate the robustness of model regarding the creatine feature. The three groups, which are split into three similar size with creatine value, are continued to be used. Confusion matrices, which are tables to describe the performance of groups, for each group are trained with fatf.tils.metrics.tools. Then, the accuracy could be calculated with information provided by confusion matrices. The figure below shows the code to obtain the confusion matrices and accuracy of each group.



The result indicates that the accuracies of three groups are 0.82, 0.79, 0.79 respectively and they seem to be comparable with each other. For completeness, the systematic\_performance\_bias\_grid function of fatf\_accountability.model.measures is used to further check the existence of systematic performance bias. As shown in the figure above, there is no systematic performance bias for these three groups with RF classifier. Till this stage, the conclusion could be made that the chosen classifier is robust regarding the feature creatinine.

The same robustness examination process is implemented to all the features. For each feature, the model is split into three groups with comparable size and then, the systematic performance bias verification function of FAT toolkit is implemented to check the existence of bias. The results show that all the features pass the test and it could be confirmedly stated that the chosen model, which is the classifier with random forest algorithm implementing RandomizedSearchCV method, is robust without systematic performance bias from the accountability perspective.

Model Explaining: ICE and PD

Individual Conditional Expectation (ICE) and Partial Dependence (PD) tools are used

Model Explaining: LIME and Counterfactuals

**Optional Extension**

Admission to ICU is also an essential task in medical system. Similar machine learning techniques can be applied to construct an ICU admission prediction classifier. Such prediction classifier can be implemented to monitor the patients in the normal ward and admitted ones whose health conditions are predicted to worsen and require intensive care. Timely recognition and treatment of pejorative illness may help to avoid mortality and improvement medical outcome.

1. Adhikari NK, Fowler RA, Bhagwanjee S, and Rubenfeld GD. (2010). Critical care and the global burden of critical illness in adults. *Lancet*, **376**(9749),1339–46.
2. Apolone G, D'Amico R, Bertolini G et al. (1996) The performance of SAPS II in a cohort of patients admitted in 99 Italian ICUs: results from the GiVi-TI. Intensive Care Med 22: 1368±1378
3. Moreno R, Morais P (1997) Outcome prediction in intensive care: results of a prospective, multicentre, Portuguese study. Intensive Care Med 23: 177±186
4. Goldhill DR, Sumner A (1998) Outcome of intensive care patients in a group of British intensive care units. Crit Care Med 26: 1337±1345
5. Smith L, Orts CM, O'Neil I, Batchelor AM, Gascoigne AD, Baudouin SV (1999) TISS and mortality after discharge from intensive care. Intensive Care Med 25: 1061±1065
6. Town P, Thabtah F (2019) Data analytics tools: a user perspective. J Inf Knowl Manag 18(1):1950002
7. Baek H, Cho M, Kim S, Hwang H, Song M, Yoo S (2018) Analysis of length of hospital stay usingelectronic health records: a statistical and data mining approach. PLoS ONE 13(4):1–20
8. c. Cain, E. Neuwirth, J. Bellows, C. Zuber, J. Green(2012) *Patient experiences of transitioning from hospital to home: an ethnographic quality improvement projec*t. J Hosp Med 7(5):382–386. https://doi.org/10.1002/jhm.1918
9. G. Anthony, T. Fadi, A. Neda, R. Kumar Sayan(2019), *Patient Discharge Classification Using Machine Learning, Techniques*, Springer-Verlag GmbH Germany, part of Springer Nature 2019
10. CJ McWilliams, DJ Lawson, R SantosRodriguez, et al. *Towards a decision support tool for intensive care discharge: machine learning algorithm development using electronic healthcare data from MIMICIII and Bristol, UK*. BMJ Open 2019;9:e025925. doi:10.1136/bmjopen-2018-025925
11. KC Safavi, T Khaniyev, M Copenhaver, et al. *Development and Validation of a Machine Learning Model to Aid Discharge Processes for Inpatient Surgical Care*. JAMA Netw Open. 2019;2(12):e1917221. doi:10.1001/jamanetworkopen.2019.17221
12. Smith C (2017) Decision trees and random forests: a visual introduction for beginners, 1st edn. Blue Windmill Media, Legazpi
13. G. Anthony, T. Fadi, A. Neda, R. Kumar Sayan(2019), *Patient Discharge Classification Using Machine Learning, Techniques*, Springer-Verlag GmbH Germany, part of Springer Nature 2019
14. Breiman L. Random forests. Mach Learn. 2001;45(1):5-32. doi:10.1023/A:1010933404324
15. Ubbink D, Tump E,Koenders J, Kleiterp S,Goslings JC, Brolmann F (2014) Which reasons do doctors,nurses, and patients have for hospital discharge?Amixed-methods study. PLoSONE9(3):1–13. <https://doi.org/10.1371/journal.pone.0091333>
16. Cain C, Neuwirth E, Bellows J, Zuber C, Green J (2012) Patient experiences of transitioning fromhospital to home: an ethnographic quality improvement project. J Hosp Med 7(5):382–386. <https://doi.org/10.1002/jhm.1918>
17. Kocic S, Vasiljevic D, Radovanovic S, Radevic S, Vukomanovic I, Mihailovic N (2016) Possible usesof data from hospital discharge reports. Serb J f Exp Clin Res 18(2):163–167. <https://doi.org/10.1515/sjecr-2016-0023>
18. MIMIC-III, a freely accessible critical care database. Johnson AEW, Pollard TJ, Shen L, Lehman L, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, and Mark RG. Scientific Data (2016). DOI:[*10.1038/sdata.2016.35*](http://dx.doi.org/10.1038/sdata.2016.35). Available from: <http://www.nature.com/articles/sdata201635>