Classifying Length of Stay Using Patient Information: Model Development and Validation on MIMIC-III Clinical Database

Yuxuan Chen, Limeng Liu, Qiaoxue Liu, Weizhou Qian

Introduction

Hospital beds, wards, and laboratories are important health care resources with limited availability. By estimating the length of the time for a patient to stay in the hospital, it will possibly help the hospital to better manage the resources and reduce costs in that the length of stay (LOS) was found able to explain up to 90% interpatient variation in hospital costs [1]. Also, LOS is a basic indicator to elevate the quality of health care services [2]. However, the prediction of LOS could be very complex due to numerous factors that could be related to the social or clinical components of patients that can affect their health care status. Age, payment classification, source of referral, specialty of the doctor, and ethnic group were reported to be associated with the mean value of LOS. [3] Additionally, comorbidities, main diagnosis or secondary diagnosis, marital status, testing, and treatment stage strategies could also affect LOS or LOS in the emergency department.[4-6]. To overcome the challenges, many previous works, which were based on machine-learning algorithms, were developed.[2][7][8]

MIMIC-III is a database integrating comprehensive clinical data of patients who were admitted to the Beth Israel Deaconess Medical Center in Boston. The database was constructed with data from several sources, which include archives from critical care information systems, hospital electronic health records, and the SSA Death Master File.[9]

In this project, we utilize the MIMIC-III database and implement two basic machine learning methods with pre-selected and transformed features based on a literature review to provide a robust and accurate prediction of the LOS for patients admitted to Beth Israel Deaconess Medical Center in Boston. Except for the age, diagnosis, ethnic group, marital status, admission type, which are reported as main factors affecting the length of stay, patients were dead or live is considered as it limits the maximum LOS for an individual. We also add the proportion of abnormality in lab tests to our model because LOS is reported to depend on test strategy and comorbidities, which could be represented by the proportion of abnormalities if patients experienced at least a similar testing process for the same diseases...

Method

DATA PROCESSING STEPS

The entire MIMIC-III dataset has over 700 million rows in all 40 tables performing different functionalities and linking with each other as shown in Figure 4. Because of the large amount of data, it is difficult to perform data processing, analyzing, and modeling in a local computing machine. Thus, we plan to use the following steps described below for both high efficiency and promising accuracy: (1) download the demo data to local machine; (2) process the demo data and save processing scripts; (3) use the query builder[10] to extract a relatively large dataset; (4) test scripts in the large local dataset, fix any issues and optimize the code to get the highest efficiency; (5) use the biostatistics cluster or google cloud to access the entire MIMIC-III dataset, which is approximately 7GB in total; (6) implement a bash script to unzip the data folder and subfolders to find the tables needed; (7) implement the script in server for extracting and processing; (8) use

cloud computing to implement several machine learning models on a dataset; (9) tuning hyperparameters of machine learning models; (10) return the results and metrics; (11) compare the accuracy and other metrics between different models.

Some key variable processing includes: changing the person whose "age" column is over 89 to 90 since the dataset hides the age for a patient who is over 89 years old; selecting the one ICD9 to diagnose result from the list based on the priority; calculating the abnormal rate among all lab test results; determine the threshold to categorized the length of days; etc.

VARIABLE SELECTION

The response variable is the Length of Stay(LOS) in the hospital, regarding the location, marked as "los" in the dataset. The calculation of LOS is based on the differences between discharge time and admission time, formatting by days. Some potential predictors are selected from different tables, including patient's age, patient's gender, ethnicity/race, marital status, death, type of admission (whether emergency or not), times enter ICU, ICD9 priority diagnosis, number of ICD9 diagnoses, percentage of abnormal lab tests.

MACHINE LEARNING METHODS

Several machine learning methods are implemented in classifying the length of stay at the hospital. Since the length of stay is calculated into numerics, a second step encoding is conducted. The threshold of the length of stays is set as five days in comparison with other literature. Categorical variables, including ethnicity/race, gender, marital status, admission type, priority diagnosis, death, have the processing of one-hot to avoid integer variation in working on machine learning models rather than using factorization (assign numbers for different categories will produce bias for machine learning training since the larger number will be treated as more important). Observations that contain missing will be assumed as "Missing Completely at Random" and eliminated before fitting the models.

Support Vector Machines & Random Forest (RF)

Support vector machines (SVMs) are a set of supervised learning methods used for classification, regression, and outlier detection [11]. In our project, SVM is used for classifying the length of stays as long (greater than 5 days) and short (less or equal than 5 days) with basic parameters. A random forest is a meta estimator that fits several decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting [12] and is also conducted in this project.

GridSearchCV

To improve the accuracy of the classification results, a method named GridSearchCV helps tune hyperparameters in the machine learning models. By listing all candidates for each hyperparameter and fitting with all combinations, it will allow the selection of the best parameters automatically with evaluation results for each parameter candidate.

Cross-Validation

Cross-validation is a resampling method that uses different portions of the data to test and train a model on different iterations. Using the 5-fold cross-validation and calculation of the average accuracy score will help to get unbiased validation and evaluation results.

Results

DATA DESCRIPTION

The MIMIC-III is a restrictedly de-identify clinical database that is consists of 26 comma-separated-value(CSV) files, including tables such as admission and discharge timestamp of patients, specimens, and their outcome of patients, the corresponding diagnosis, and related demographic data. Among these files, 7 tables that contain the interesting columns are selected. Specifically, we choose ADMISSIONS.csv with 58,976 observations, LABEVENTS.csv with 27,854,055 rows, D_ICD_DIAGNOSES.csv with 14,567 rows,

D_ITEMS.csv with 12,487 rows. DIAGNOSES_ICD.csv with 651,047 rows, ICUSTAYS.csv with 61,532 rows, and PATIENTS with 46,520 rows.

Table1: Summary Statistics

Variables	N(%)	Variables	N(%)	Variables	N(%)	Variables	Means(SD)
Ethnicity		**Admission Type**		**Gender**		**Total LOS in ICU**	5.3 (10.2)
American native	54 (<0.1%)	Elective	6,874 (12%)	F	24,883 (44%)		
Asian	1,906 (3.4%)	Emergency	40,820 (72%)	M	31,560 (56%)	**Admission Age**	55 (27)
Black	5,550 (9.8%)	Newborn	7,531 (13%)	**ICU Times**			
Caribbean	9 (<0.1%)	Urgent	1,218 (2.2%)	1	53,242 (94%)	**Abnormal Test Percentage**	0.33 (0.10)
Hispanic	2,054 (3.6%)	**Marital Status**		2	2,816 (5.0%)		
Middle Eastern	42 (<0.1%)	Divorced	3,080 (5.5%)	3	318 (0.6%)	**Hospital Expired**	
Multi Race	124 (0.2%)	Life Partner	15 (<0.1%)	4	53 (<0.1%)	Expired	5,725 (10%)
Native Hawaiian	17 (<0.1%)	Married	23,122 (41%)	5	9 (<0.1%)	Not Expired	50,718 (90%)
Portuguese	61 (0.1%)	Separated	543 (1.0%)	6	3 (<0.1%)	**Length of Stay**	35060(62%)
South american	7 (<0.1%)	Single	12,722 (23%)	7	2 (<0.1%)	> 5 days	35,060 (62%)
White	39,493 (70%)	Widowed	6,929 (12%)			<= 5days	21,383 (38%)
Other	1,461 (2.6%)	Unknown	10032(18%)				
unknown	5,665 (10%)						

To answer the question on how to use the machine learning model in classifying the length of stay, 23 columns were selected or derived from the original data file. While some of the columns(such as patient ID) are selected for merging the data, predictors of interest include length-of-stay in hospital and ICU, type of admission, ethnicity of the patients, their marital status, age, diagnosis. In specific, the length-of-stay is the difference between admission time and discharge time, and the age is calculated by subtracting the birth date from the admission time.

Both the continuous and categorical variables in the dataset are examined. While examining the correlation between continuous predictors, all continuous predictors have a weak correlation indicating independence as shown. Regarding the size of the zipped data, which is 7GB, the data were uploaded and merged on the biostatistics cluster using Rscript. Having the data merged and variable selected among all tables matched with the patient's admission identification, there are a total of 58,976 observations in the final dataset. While having all missing values removed, 47,403 observations remain.

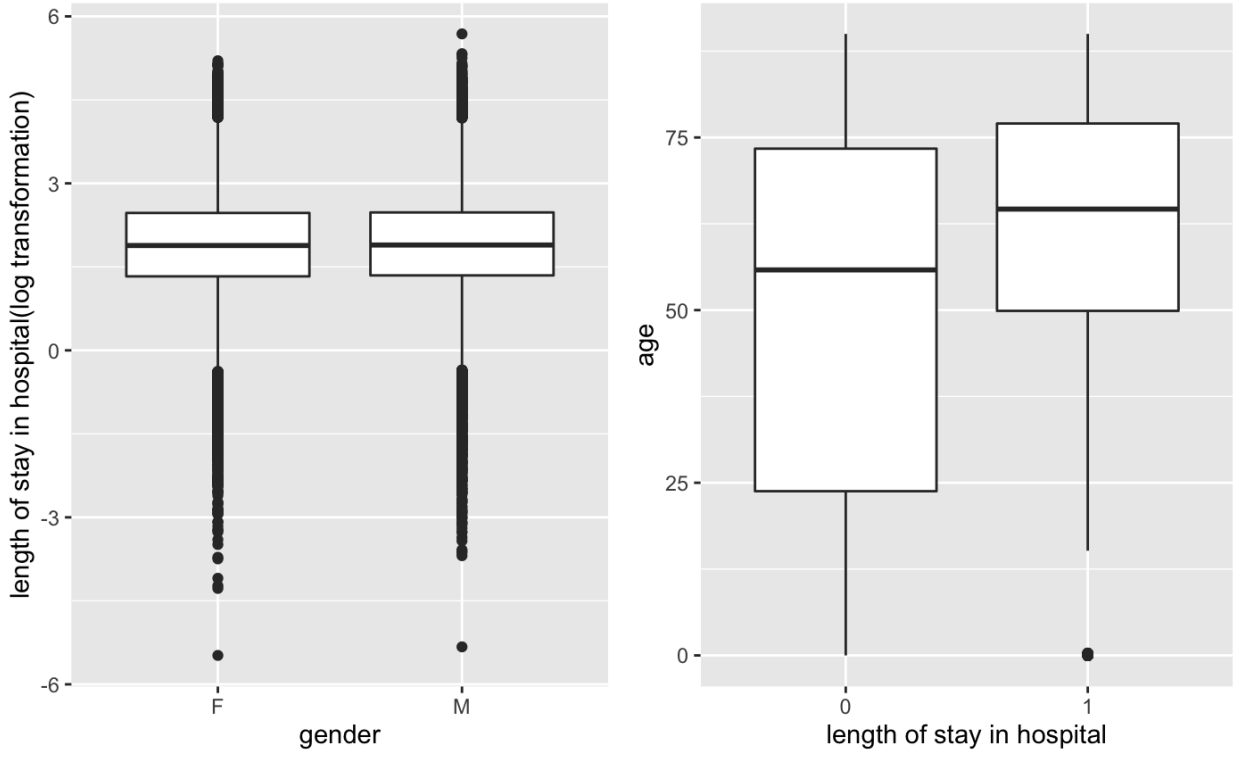

Figure 1: Boxplot

MACHINE LEARNING RESULTS

Having all variables being standardized and processing one-hot methods to categorical variables, there are a total of 33 variables excluding diagnosis and 14566 variables including diagnosis (more than 10,000 unique variables in diagnosis. The smaller data was performed in both Support Vector Machines(SVM) and Random Forest(RF) models; however, since SVM takes much more time for processing data and cannot use GPU instances, a dataset with diagnosis is only trained and tested in RF models.

Support Vector Machines

The pipeline with both standard scaler and SVC, which is the classification model of SVM, is used for fitting. The train and test data were split by test size equal to 0.2, which means that one-fifth of the data is test data, and the rest is training data. For the SVC parameters, gamma is set to be automatically generated. Using the cross-validation with 5 folds and calculating the mean accuracy rate across the 5 times training, the average score is around 0.69658. The Area Under Curve metric(AUC) is also used for identifying the ability of binary classification performance. AUC can be expressed by the Wilcoxon-Mann-Whitney statistic, with a return of decimal numbers between 0 and 1, where higher means the better ability of classification. For this SVC model, the calculated pipeline AUC is approximately 0.71. A Receiver Operating Characteristic Curve (ROC) is also produced to determine the ability of model performance, as shown in Figure 2.1. The x-axis of the curve represents a false positive rate, where smaller indicates a better model. The y-axis of the curve represents a true positive rate, where closer to 1 means a better model. In such a setting, an expected model should have a sharp ROC curve that tends to the top left corner. However, in this baseline SVC model, the curve is slightly concave down.

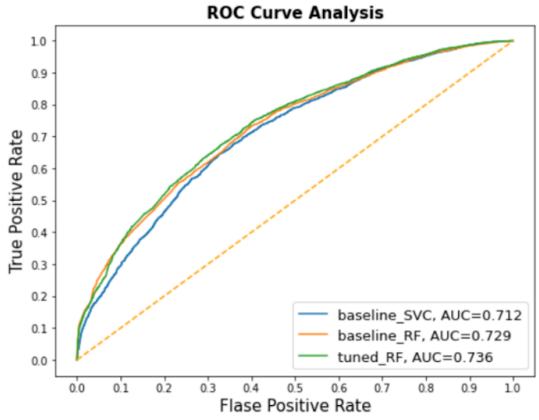

Figure 2.1: ROC Curve Analysis

	precision	recall	f1-score	support
0	0.65	0.27	0.39	3164
1	0.72	0.93	0.81	6317
accuracy			0.71	9481
macro avg	0.69	0.60	0.60	9481
weighted avg	0.70	0.71	0.67	9481

Figure 2.2: Classification Report of tuned RF

Figure 2: Model Results

Random Forest

Using the same subset data which exclude the diagnosis, a baseline random forest(RF) model with parameters that have a maximum depth equal to 5 and a random state equal to 0 is also performed. Using the same metrics as mentioned above, the 5-fold cross-validation returns a mean accuracy score of 0.69637, AUC score of 0.73, and the ROC plot shown in Figure 2.1. To compare the baseline SVC and RF model, the cross-validation score is quite close with a higher AUC. However, in considering the efficiency of model training and fitting, SVC is more time-consuming and does not perform more significantly accurately than RF models. Therefore, GirdSearchCV for tuning the hyperparameter is only conducted for RF models. By adding the parameter candidates, there are 24 candidates of combinations and totaling 120 fits for 5 folds. The best parameters shown as Gini for criterion, maximum depth sets to 10 and automatically generated maximum features, can thus improve the accuracy to 0.71 and increase AUC to 0.74 (Figure 2.1). A detailed classification report can be found in Figure 2.2. While adding the diagnosis as predictors as well and transforming into one-hot. GridSearchCV found a different combination of parameters: criterion of Gini, maximum depth of 2, and automatically maximum features. Although a more significant predictor is added,

the accuracy score does not increase (mean accuracy = 0.65, AUC = 0.71).

Discussion

It is surprising that the diagnosis included in the dataset of training did not improve the accuracy, but lowered the efficiency. One possible reason is that we have more than 10,000 unique values in diagnosis which means that they have added more than 10 thousand columns into the training data using one-hot methods. An adequate amount of predictors can help in improving the accuracy but can also have drawbacks especially because most of the predictors are one-hot transformed columns. In the future analysis, we may need to find the parent category of each icd9 diagnosis and categorize a smaller number of labels to decrease the number of columns in the training dataset.

To perform the analysis, we made some assumptions and thus brought some limitations. First, we consider using the admission id as a unique identifier for each observation. However, a patient may come to the hospital multiple times, this condition will be shown as the same subject id, but different admission ids in the dataset. Therefore, treating each admission id as an independent person is not reliable, and observations will thus be correlated. Secondly, we calculate the percentage of abnormal lab tests, to make hypotheses that higher abnormal rates may indicate worse health conditions and thus increase the length of stay. However, to form the hypothesis, we need to assume that for the same disease/initial diagnosis, all doctors will order the same lab tests, which have not been proved. Therefore, whether the percentage of abnormality can determine the servers or not is questionable. Thirdly, we categorized age because we cannot have detailed information about patients over 89. But age is significant information to impact the length of stay, and the categorization may obscure the true association. Lastly, the MIMIC dataset was collected only in one hospital in Boston, location will not be a variable in our project, but a potentially significant cause. Therefore, our project's result may not be suitable for predicting people from elsewhere other than Boston.

Contribution

Yuxuan Chen: Data cleaning and preparation; readme write up; **Limeng Liu**: Method and discussion section; ML models programming and implemented **Qiaoxue Liu**: Data description; Data preparation on the cluster; readme write up **Weizhou Qian**: Literature review, introduction, and objective, "readme", visualization

Reference*

[1] Rapoport, John, et al. "Length of Stay Data as a Guide to Hospital Economic Performance for ICU Patients." Medical Care, vol. 41, no. 3, Mar. 2003, pp. 386–397., https://doi.org/10.1097/01.mlr.0000053021.93198.96.

[2] Mekhaldi R.N., Caulier P., Chaabane S., Chraibi A., Piechowiak S. (2020) Using Machine Learning Models to Predict the Length of Stay in a Hospital Setting. In: Rocha Á., Adeli H., Reis L., Costanzo S., Orovic I., Moreira F. (eds) Trends and Innovations in Information Systems and Technologies. WorldCIST 2020. Advances in Intelligent Systems and Computing, vol 1159. Springer, Cham. https://doi.org/10.1007/978-3-030-45688-7_21

[3] Liu Yingxin , Phillips Mike Codde Jim (2001) Factors influencing patients' length of stay. Australian Health Review 24, 63-70., https://doi.org/10.1071/AH010063

* Full reference see GitHub Repository.