STAT 844 PROJECT REPORT

Sophia/Yunqing He
y352he@uwaterloo.ca

Friday 11th May, 2018

1 Introduction

The Length Of Stay (LOS) in Intensive Care Unit (ICU) is a crucial measurement in determining hospital resource allocation and evaluating health care quality. Since the intensive care unit focuses on treating and monitoring severely ill patients, it is equipped with a variety of advanced medical equipments and teams of physicians and nurses for close observation and monitoring purposes. Therefore, a single ICU stay can generate an extensive amount of data. [14] From a managerial prospective, LOS in ICU is a indication of how effective medical resources are allocated within the hospital. [5] Studies suggested that even a small proportion of prolonged LOS in the ICU contributes to a large proportion of hospital expenses. [8] Therefore, in order to reduce hospital expenses, it's best to start with providing productive care for both patients and hospital by accurately predicting the LOS in the ICU. Furthermore, studies have shown that prognosis is especially difficult for ICU physicians since symptoms and conditions are much more complex and less predictable for ICU patients. [13] However, recent studies have been mostly focusing on predicting mortality in the ICU, including developing sophisticated disease severity score for mortality prediction. [16] [3] [10] [7] Those serve as the motivation to develop a systematic methodology to predict the length of stay in the ICU using machine learning models.

In general, researchers tend to estimate the possible LOS in 2 ways: model LOS in ICU across all types of patients (with particular focus on assessing variable importance) [8] [12] [15] [5] or model it for a specific type of patients [11] [9]. A common purpose of both types of studies is benchmarking ICU care delivery. For models that focus on all type of diseases, there are 3 major aspects that need to be improved based on a systematic review conducted in 2017. [18] Firstly, many studies have concluded that the LOS in ICU is right-skewed, which indicates there are fewer ICU stays that have prolonged LOS while majority of them constitute relatively short LOS. Due to such feature of the data, many studies decided to eliminate those extreme values when training their models. Such approach can yield satisfactory prediction accuracy for majority of the patients. However, for benchmarking purpose, eliminating those extended ICU stay may overestimate the quality of care in the ICU and lead to biased evaluation. [18] [8] [12] [15] Secondly, ICU survivors and non-survivors may behave differently due to various physical and psychological reasons and it makes sense to develop separate models for predicting those ICU patients. More importantly, models developed based on non-survivors' data may have the ability to detect warning signs for fast-deteriorating ICU patients. [18] Lastly, many studies have adopted ordinary linear regression approach that can result in negative prediction values. A simple logarithm transformation can fix this problem. [18] There are fewer disease-specific studies that also focus on predicting LOS in ICU. [11] [9] Many tend to focus on patients who undergo cardiac surgery since "intensive care is a standard component of postoperative treatment for most patients who undergo cardiac surgery". [2]

Thus, based on the brief literature review summarized above, the goal of this project is to predict the length of first ICU Stay (LOS) for different types of surgical patients admitted in the (ICU) of the Beth Israel Deaconess Medical Center between 2001 and 2012. More specifically, I'm planning to utilize surgical patients' information only within their first 24 hours in the ICU to predict their total LOS. The reason for using those information as predictors is that since many information contained in the MIMIC III database (explained in the Data Collection section) are highly correlated with the response variable: LOS in ICU. For example, the total number of times blood pressure are measured is highly correlated with total LOS, because as a routine procedure, the longer a patient stays in the ICU, the more times his/her blood pressure would be measured. However, those information are not available if we want to predict a patient's LOS, since they (i.e. total number of times a procedure is practiced) are known only after the patient is discharged or expired/died.

As a general assumption, my reasoning for using such information is that the first 24 hours after transferred to the ICU is critical in determining his/her total LOS. Another assumption is that patients who went through invasive surgical procedures might have a longer LOS, compared to non-surgical patients. Therefore, I plan to exclusively focus on surgical patients who were transferred to the ICU during their hospital stays. More details about inclusion/exclusion criteria will be provided in the Data Collection section.

2 Data Collection

This section summarizes information about MIMIC III database and data extraction criteria used for this project.

2.1 Database

The datasets are extracted from a database called Medical Information Mart for Intensive Care III (MIMIC III). [1] It is "a large, freely-available database comprising deidentified health-related data associated with over forty thousand patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012".[1] It is a structured database that contains demographics, admission, chart events, diagnosis and treatment information about ICU patients. In total, there are 26 tables that outline information from a patient's initial contact with health providers until his final discharge from any departments within the hospital. [17] After excluding 6 item tables (that list codes and corresponding descriptions, e.g. code for stroke and its description), there are 20 tables that capture patients' full interactions with the health services.[17] A full list of those tables are provided, each with a brief description as the following (Tables that are used for this project are listed below & the rest are listed in the Appendix): [1]

- **Admission**: Define a patient's hospital admission
- **Patients**: Contains all charted data for all patients
- **Services**: Lists services that a patient was admitted/transferred under
- **ICUstays**: Outlines patient information about each ICU stay
- **ChartEvents**: Contains all charted data for all patient
- **LabEvents**: Contains all laboratory measurements for a given patient, including outpatient data

Note: Important identifier

1. SUBJECT_ID is unique to a patient, who may have multiple hospital admissions
2. HADM_ID is unique to a patient hospital admission; patients may have multiple hospital admission; each hospital admission may have one or multiple ICU stays
3. ICUSTAY_ID is unique to a patient ICU stay

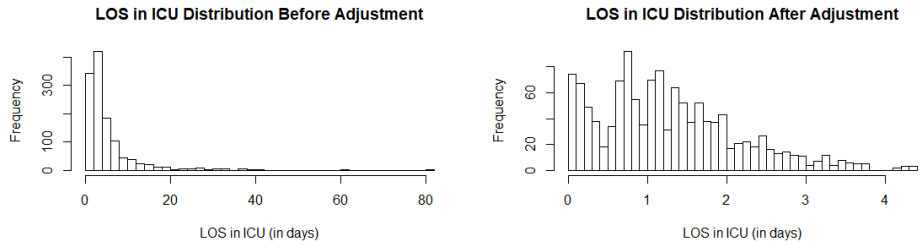
2.2 Data Extraction

Since the goal of this project is to predict the Length of Stay (LOS) for surgical patients admitted in the ICU using their first 24-hour information, data are extracted based on the following criteria:

- **Non-expired/Not-dead patients** who get discharged from ICU; Expired patients are excluded from the training set because their LOS are unstable (some with very short LOS); However, there might be potential for a separate project focusing characterizing features of expired patients from ICU that could be useful for reducing hospital ICU mortality rate
- **Admission Type Not under "Newborn"**: based on previous research, newborn patients in ICU behaves quite differently from other adults; MIMIC III database did not differentiate ICU with NICU, thus it could helpful to exclude those records in order to build a better predictive models for adult ICU patients
- **Only surgical patients**: in MIMIC III's Services Table, it lists services that a patient was admitted/transferred under. Instead of using locations/departments the patient is residing in, using Services Table can filter out non-surgical patients based on the actual service he/she received. The reason is that due to bed shortage and other possible concerns, the department a patient is residing in may not be the department that provided services for that patient. Types of surgical patients included: Cardiac Surgery, Neurological Surgery, Orthopedic Surgery, Plastic Surgery, General Surgery, Trauma (Surgical), Vascular Surgery and Obstetrics Surgery (related to childbirth).
- **LOS greater than 24 hours**: since only information within first 24 hours are used, it's only sensible to include LOS greater than 24 hours.

After determining the scope of patient records, 5 major datasets containing information about patients' first 24 hours ICU stay are constructed using SQL and codes are included in the codes folder. [6]

1. **ICU Stay Details**: contains basic ICU stay information, such as gender, age, ethnicity, admission type, hospital expire flag, hospital stay sequence, whether is first hospital stay, hospital in time, hospital out time, LOS in hospital, ICU admission time, ICU discharge time, LOS in ICU, ICU Stay sequence, whether is first ICU stay.
2. **Height Weight Output**: contains first, minimum and maximum heights and weights of ICU patients
3. **First 24-hour Vitals**: contains patients' vitals information that were measured during their ICU stays. For example, heart rate, blood pressure, glucose level, respiratory rate etc. All minimum, maximum and average values of those measurements are recorded



(a) LOS in ICU (days) Distribution Before Adjustment

(b) log(LOS) in ICU (days) Distribution After Adjustment

Figure 1: LOS in ICU (days) Distribution Before and After Adjustment

4. **First 24-hour Lab Results**: contains patients' lab results that were requested for each patient. Different patients may have missing values in certain measurements due to the fact that those particular lab procedures were not requested for that patient. Examples include hemoglobin level, chloride level, potassium levels etc.

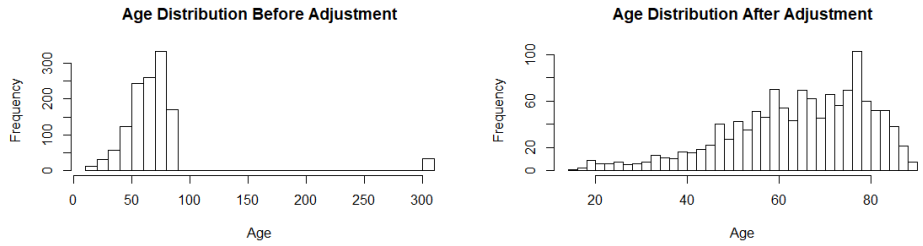
5. **First 24-hour GCS**: GCS or Glasgow Coma Score is "a neurological scale which aims to give a reliable and objective way of recording the conscious state of a person". Examples include GCS motor, GCS eyes, GCS verbal etc.

3 Method

3.1 Data Processing

The goal of this project is to construct models to effectively predict the length of ICU stay. As literature concluded, the LOS in ICU are usually right skewed, due to the fact that majority of the patients have shorter LOS in the ICU while fewer may have prolonged ICU stays. The distribution of LOS in ICU of this dataset is shown above (*Figure 1 Left*) and it suggests a strong right-skewed distribution. Therefore, **log transformation is used on the response variable** to correct such distribution (distribution of log LOS shown above in *Figure 1 Right*).

In the general ICU setting, patients' vital signs are routinely measured by health care professionals, while patients' lab results are more variable since patients with different diseases may require drastically different lab measurements. Therefore, there are much more missingness in the lab results compared to vital signs measurements. Information on demographics about patients, such as gender, age, admission type and discharge location etc, are generally complete for all patients involved in the training dataset. Initially, all demographic variables are selected into the training set, which include gender, age, ethnicity, admission type, first hospital stay, discharge location, language, religion and marital status. For first 24-hour vitals, average value of vital sign measurements are utilized, which include heart rate, systolic blood pressure, diastolic blood pressure, mean blood pressure, respiratory rate, spo2 (peripheral capillary oxygen saturation) and temperature in Celcius. For first 24-hour lab results, only variables that have less amount of missing values are included, which include mean glucose, max creatinine, max hematocrit, max hemoglobin, max potassium, max sodium and max blood urea nitrogen. For first 24-hour GCS, GCS Motor, GCS Verbal, GCS Eyes and EndoTrachFlag (whether intubated) are included in the training dataset.



(a) Age Distribution Before Adjustment (b) Age Distribution After Adjustment

Figure 2: Age Distribution Before and After Adjustment

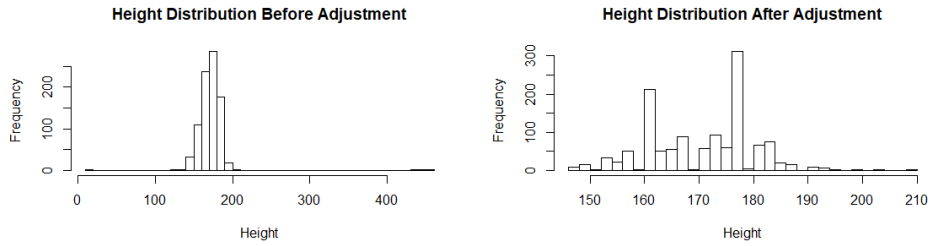
Overall, imputation is moderately utilize to fill the **missing values** in continuous variables. The reason is that in medical data, certain missingness doesn't mean it's useless information, but possibly represent certain physical features about the patients. Some summary statistics about the original dataset are shown below to illustrate the amount of missingness in vitals and lab results:

```
> summary(df1$heartrate_mean)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  49.82   77.67   85.92   87.04   95.40  141.70       7
> summary(df1$glucose_mean)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   70.0   117.1   132.9   139.3   153.0   352.7       5
> summary(df1$hemoglobin_max)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   7.20   10.80   12.10   12.14   13.40   21.00       3
```

In additional to missing values, there are also few **outliers** in the demographic features including age, height and weight that are simply due to input errors. As shown below and in *Figure 2 Left*, the summary statistics and distribution of age indicate that there are people with age over 300 years old.

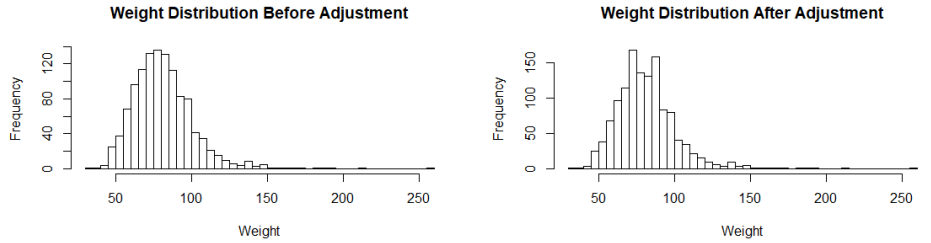
```
Age distribution before adjustment
> summary(df1$age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  15.21   54.67   66.33   69.73   76.72  306.20
Age distribution after adjustment
> summary(df1_final$age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  15.21   54.67   65.96   63.95   76.41   89.00
```

Due to the fact that those patients are admitted to the ICU and their physical conditions might be unique compared to the general population, a rather conservative approach of imputation using gender-specific mean value is used. For all female patients whose ages are greater than 100 years old, the average age of all female ICU patients (which is around 76.5 years old) is used. Similarly, male patients who ages are over 100 years old, the average age of all male ICU patients (which is around 65.8 years old) is used. For height measurement, average heights of female and male patients are used for patients whose heights are less than 145 meters (since children and newborn patients are excluded from the training set, it's considered impractical to have



(a) Height Distribution Before Adjustment (b) Height Distribution After Adjustment

Figure 3: Height Distribution Before and After Adjustment

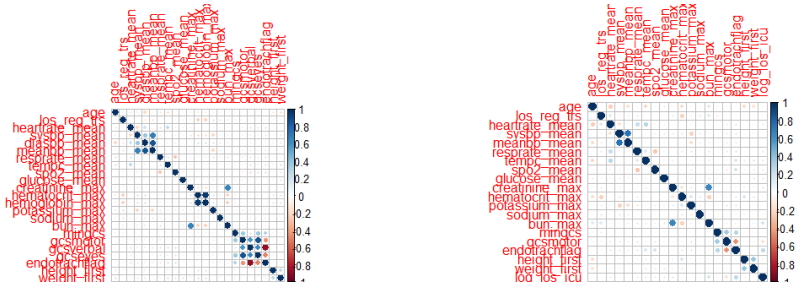


(a) Weight Distribution Before Adjustment (b) Weight Distribution After Adjustment

Figure 4: Weight Distribution Before and After Adjustment

height less than 145 meters for general adult patients) or greater than 220 meters for each gender type. Moreover, another major problem of height and weight is large amount of missing values (height has 395 NAs out of 1262 observations; weight has 81 NAs out of 1262 observations). This is solved also by using gender-specific average values. The summary statistics and distribution of those variables are provided below and in *Figure 3* and *Figure 4*:

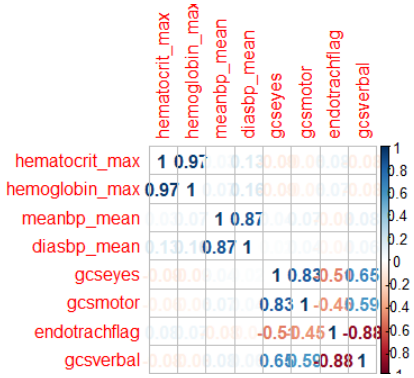
```
Height Before Adjustment (in meters)
> summary(df1$height_first)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   12.7   162.6   170.2   171.0   177.8   454.7     395
Height After Adjustment (in meters)
> summary(df1_final$height_first)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  147.3   160.9   172.7   170.4   176.4   208.3
Weight Before Adjustment (in pounds)
> summary(df1$weight_first)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  34.40   68.00   79.50   81.81   91.50  257.00      81
Weight After Adjustment (in pounds)
> summary(df1_final$weight_first)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  34.40   69.53   79.95   81.75   90.70  257.00
```



(a) Correlation of All Continuous Variables Before Adjustment

(b) Correlation of Continuous Variables After Adjustment

Figure 5: Correlation of Continuous Variables Before and After Adjustment



(a) Correlated Continuous Variables

Figure 6: Correlation of Continuous Variables

Another issue that is addressed during the data preprocessing step is correlation between explanatory variables. With basic amount of subject knowledge, certain correlated variables were excluded during the initially stage of data extraction. *Figure 5a* illustrates the correlation plot between all continuous variables after imputation and outliers adjustment. As we can see, there are many off-diagonal dots which indicated strong positive or negative correlations between those variables. Variables with absolute (Pearson) correlation coefficient greater than 0.80 are specified in *Figure 6*. Within each pair of correlated continuous variables, the ones that is more correlated with the response variable is removed. Finally, all final continuous variables that are included in the final dataset show no obvious correlation with the response variable: Log LOS in ICU (Last row in *Figure 5b*).

For categorical variables, pairwise correlations are evaluated using Chi-square tests and the results show that most categorical variables have little or no obvious correlation between them (most p-values are significantly smaller than 0.05). Since tree-based methods and generalized linear model via penalized maximum likelihood approaches (discussed in the next section) are used for building the predictive model, correlation problem is addressed when using those modeling techniques. Thus, no categorical variables is removed or adjusted.

There is one derived variable that is included in the final dataset: length of time from the time when patient was admitted to the hospital to the time when patient was transferred to the ICU. This variable captures the certain aspect of patient's illness severity. If such length of time is

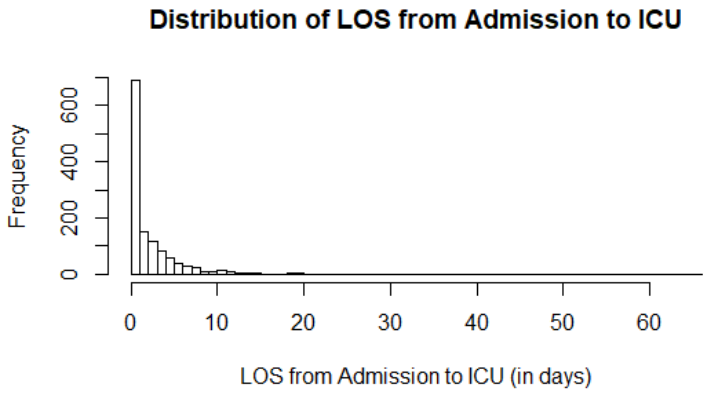


Figure 7: Distribution of LOS from Hospital Admission to ICU Admission

rather short, it means that the patient is deteriorating relatively fast. The summary statistics and distribution of this variable (called "los_reg_trs") are shown below. Similar to the total LOS in the ICU, the length of time between hospital admission and ICU admission is also right-skewed.

```
> summary(df1_final$los_reg_trs)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
 0.0000   0.0013   0.8130   2.3560   2.9400  65.4900
```

3.2 Analytical Framework

The 2-layer analytical framework used for this project is demonstrated by *Figure 8*. Multiple ensemble methods are used to optimize the prediction accuracy on the testing dataset. On the bottom layer, 5 machine learning models are used to construct 5 predictive models.

1. **Smoothing**: In general, smoothing spline fit the data more locally. The extreme case is to have knots on every distinct X values. Due to the high dimensionality of the data, multivariate adaptive regression spline (or MARS, using *earch* package/functions) is implemented.

   Generally speaking, MARS is an adaptive method that is build on the idea of piecewise linear regression. From the textbook of this course (Element of Statistical Learning), piecewise linear basis functions are of the following forms:

$$\begin{aligned}(x-t)_+ &= x-t \text{ if } x > t\\ &= 0 \text{ otherwise}\\ (t-x)_+ &= t-x \text{ if } x < t\\ &= 0 \text{ otherwise}\end{aligned}$$

   Thus, the collection of basis functions is $C = \{(X_j - t)_+, (t - X_j)_+\}, t \in \{x_{1j}, ..., x_{Nj}\}, j = 1, 2, ..., p$ and stepwise linear regression technique is used to estimate the coefficients:

$$f(X) = \beta_0 + \sum_{m=1}^{M} \beta_m h_m(X) \tag{1}$$

where each $h_m(X)$ is a function in C

2. **Random Forest algorithm**: The second basic model, namely Random Forest model, utilizes one of the ensemble methods of bagging to prevent high instability from a single tree-based model. As explained in lecture, a major disadvantage of a decision-tree method is its high variance due to minor changes in the data. With slight changes in the training data, the regions and boundaries of partition can be change drastically, which causes unstable model predictions. In Random Forest algorithm, B bootstraps samples are constructed from the training dataset and a proportion of the explanatory variables is used to fit B tree-based models. The predictions resulted from each models are averaged to produced the final predictions on the testing set. The reason that only a proportion of explanatory variables is used for each bootstrap sample is to avoid highly correlated bootstrap samples and (in terms) correlated predictions. In summary, Random Forest algorithm is a type of tree-based method that incorporates the ensemble method of bagging to improve prediction accuracy.

3. **Generalized Boosted Regression Models**: the third basic model/learner implements gradient boosting method to improve the prediction accuracy from multiple regression models. Compared to bagging which is used in Random Forest algorithm, boosting method is to produce the final prediction by a linear combination of previous learners, instead of simply taking the average (like in Random Forest). Such ensemble method is particularly concentrating on reducing biases. Namely, each updated model is focusing on the incorrectly predicted observations and the algorithm adjusts the model to minimize those residuals.

4. **Penalized Linear Model**: While simple linear model can easily overfit the data, generalized linear model with penalty terms can address this problem more efficiently. There are 3 popular penalized regression models: LASSO regression, Ridge regression and elastic net. They are all trying to solve the following optimization problem:

$$min_{\beta_0,\beta} \frac{1}{N} \sum_{i=1}^{N} w_i l(y_i, \beta_0 + \beta^T x_i) + \lambda[(1-\alpha)||\beta||_2^2/2 + \alpha||\beta||_1] \tag{2}$$

where $l(y_i, \beta_0 + \beta^T x_i)$ is the negative log-likelihood function. For LASSO regression, $\alpha = 1$ and the optimization problem is simplified to $min_{\beta_0,\beta} \frac{1}{N} \sum_{i=1}^{N} w_i l(y_i, \beta_0 + \beta^T x_i) + \lambda[\alpha||\beta||_1]$. For Ridge regression, $\alpha = 0$ and the optimization problem is simplified to $min_{\beta_0,\beta} \frac{1}{N} \sum_{i=1}^{N} w_i l(y_i, \beta_0 + \beta^T x_i) + \lambda[(1-\alpha)||\beta||_2^2/2]$. For elastic net, it's controlled by the value of α which is between 0 and 1.

5. **Support Vector Machine**: support vector machine (SVM) algorithm is a powerful and well-understood machine learning algorithm. It's largely used in solving classification problems, but also very useful in regression setting. The fundamental idea behind this algorithm is to establish hyperplane that separate distinct data points. While for linearly separable data, there exists many possible hyperplanes, SVM tends to find the "best" hyperplane that maximizes the margin between support vectors. In this way, only a small fraction of the data are used to locate the "best" hyperplane and therefore reduce the computational costs of running such algorithm.

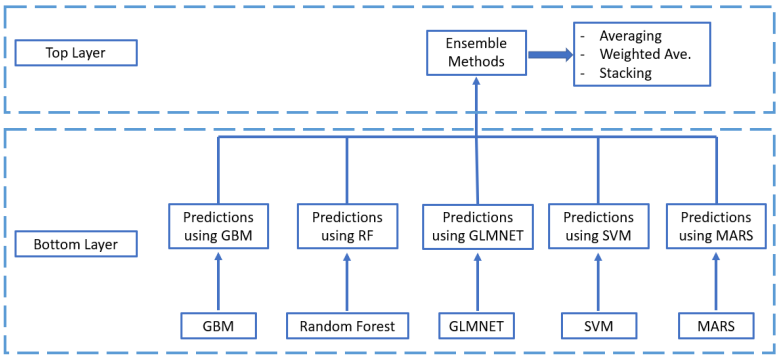


Figure 8: Analytical Framework

After predictions are made from those 5 basic learners at the bottom layer, several ensemble methods are used on the top layers. Namely, averaging method, weighted average method and stacking method.

1. **Averaging Method**: it's one of the simplest ensemble methods. As 5 sets of predictions are generated from those 5 basic models, averaging method will take the average of those 5 sets of predictions and use it as the final predictions on the testing set.

2. **Weighted Average Method**: similar to the averaging method, weighted average generally assign higher weights to individual models that have higher prediction accuracy during the training stage. For example, if GBM model has a lower cross-validation prediction error rate, I would assign higher weights (e.g. 0.4) to predictions produced by the GBM model and relatively smaller weights to the other models (e.g. 0.2 to the remaining 4 models).

3. **Stacking Method**: stacking method is a more sophisticated ensemble method compared to averaging or weighted averaging method. It takes 5 sets of prediction values as input variables and fit a machine learning algorithm to produce the final prediction. Stacking method works best if the basic models are less correlated with each other. In other words, each basic learners captures different aspects of the behavior of the true model.

4 Results

The followings are the test set prediction errors (cross validation errors and out-of-sample error) for each basic models:

Cross Validation Error (MSE)

Model	MSE
MARS	0.575
Random Forest	0.591
SVM	0.609
GLMNET	0.648
GBM_Normal	0.766
GBM_Huber	0.780

Out-of-Sample Prediction Error (MSE)

Model	MSE
GBM_Huber	0.596
Random Forest	0.611
GBM_Normal	0.612
MARS	0.624
SVM	0.627
GLMNET	0.640

Note: the MSE is calculated using log(LOS)

GLMNET refers to the LASSO model; GBM_Normal refers to the stochastic gradient boosting model with Gaussian distribution (i.e. loss function); GBM_Huber refers to the stochastic gradient boosting model with Huber loss function.

For individual basic models, the cross-validation error is lowest when using MARS model (MSE = 0.575), followed by Random Forest (MSE = 0.591) and SVM (MSE = 0.609). In terms of prediction accuracy on the testing, it is highest using the GBM with Huber loss function (testing MSE = 0.596), followed by Random Forest and GBM with Gaussian loss function (0.611 & 0.612). Note: most of those MSE can be reproduced by the codes submitted by setting seed, with some small variations due to random sampling within the modeling algorithm.

4.1 Smoothing: Multivariate Adaptive Regression Spline (MARS)

For MARS model with only main effect variable (no interaction term), it produces the lowest out-of-sample prediction error (MSE = 0.624). Its model summary information are provided in *Figure 9*. Those predictions are used in the top layer for ensemble method. The reason is that even though out-of-sample prediction error may underestimate the true MSE, but the goal of the top layer model is to improve the prediction accuracy on the testing set and therefore those results are utilized. For MARS model that allowed 2-way interactions, it produces the generalized cross-validation error (MSE = 0.539). Its model summary information are provided in *Figure 10*.

For both models, the top 5 most important variables contributing to the changes in log of LOS in ICU are: length of time from admission in hospital to transfer to ICU, endotrach_flag (whether intubated), discharge location to home and home with health care services and the average body temperature.

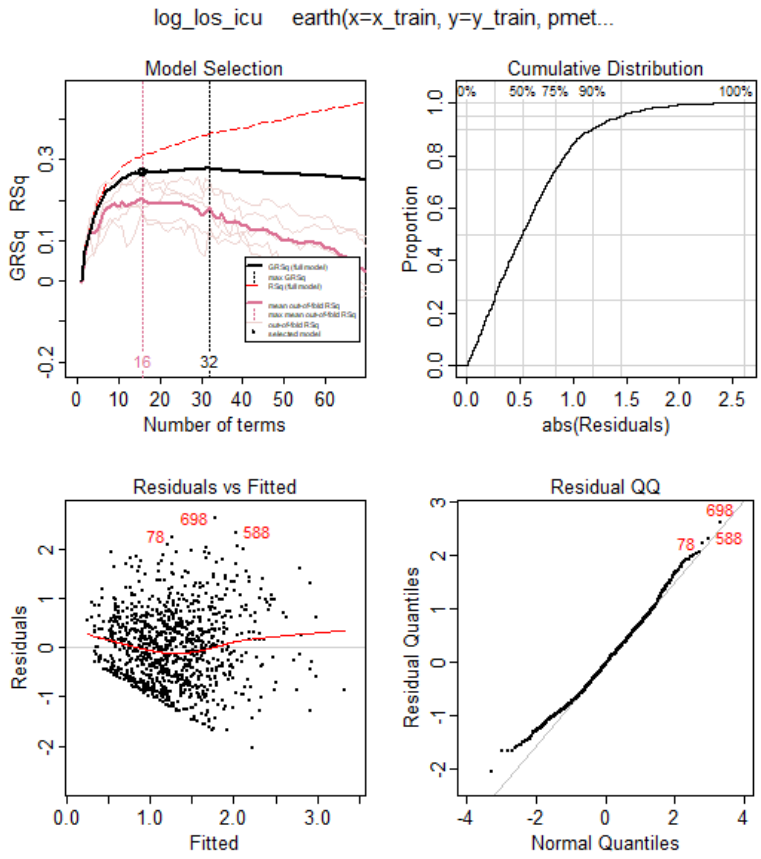


Figure 9: MARS model with lowest out-of-sample prediction error

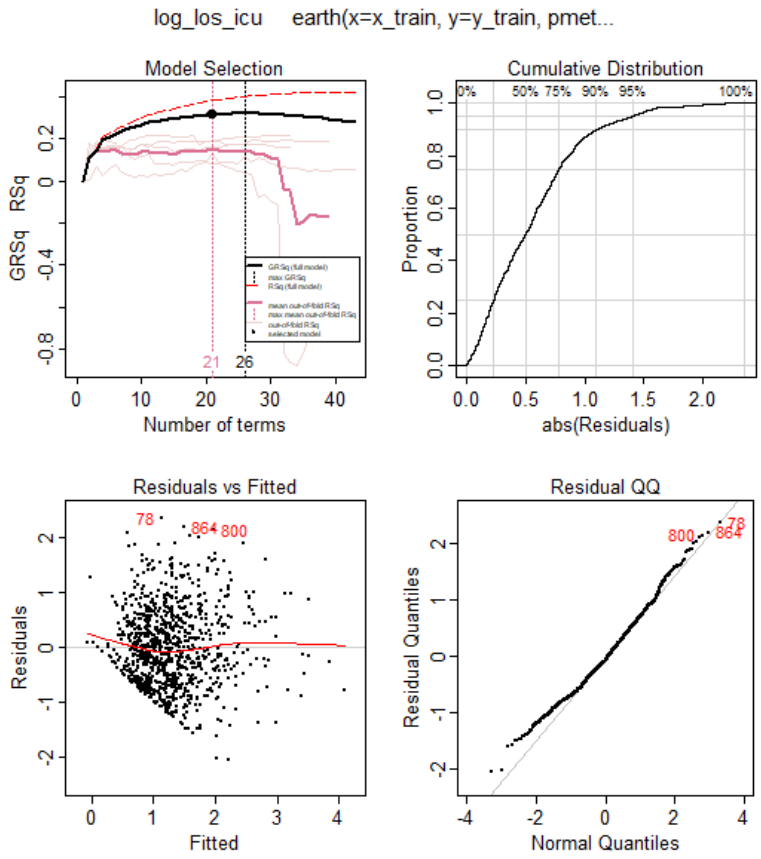


Figure 10: MARS model with lowest CV prediction error

4.2 Random Forest

A Random Forest model is trained with 100 trees and is tuned using random grid search method with 5-fold cross validation. Random grid search method is a random search of all the combinations of the hyperparameters. The search criteria is to evaluate the model performance for mtry equals 1 to 15 (by 1) with the stopping tolerance of change in MSE of 0.001.

Based on the final predictive Random Forest model, the top 5 variables that contribute the most to the variations in the response (log(LOS)) are listed as the following with their corresponding percentages:

1. Discharge Location: 10.5%
2. Religion 6.4%
3. Respiratory Rate 5.8%
4. Endotrach_flag (whether intubated) 5.8%
5. Max Hematocrit 5.4%

In the h2o package, the variable importance score "is determined by calculating the relative influence of each variable: whether that variable was selected during splitting in the tree building process and how much the squared error (over all trees) improved as a result".

Most of them are sensible and help us to better interpret the model. The most common discharge locations in the training set are Home Health Care (33.9%), Rehab Facilities (24.6%), Home (17.4%) and Skilled Nursing Facilities (16.8%). As some of the most important vital signs for ICU patients, respiratory rate, whether the patient is intubated (Endotrach_flag) and maximum hematocrit (which is the ratio of the volume of red blood cells to the total volume of blood; Low hematocrit can be caused by blood loss; High hematocrit may be caused by dehydration or other disorders) are all reasonable predictors of the total length of stay in the ICU. As for religion of the patients is considered as an important predictor of the ICU LOS, it poses an interesting indication of the training set. However, I believe due to the small size of the training data, this predictor may just be causing a random noise in the response and the model is unable to detect it. But it might be helpful to further investigate the influence of religion of patients on their ICU LOS. The complete model summary are provided in the Appendix.

4.3 Gradient Boosting Method (GBM)

Two GBM models are trained and the only difference is one uses a Gaussian loss function and the other uses a Huber loss function. As we can see from the out-of-sample prediction error rate (MSE), GBM model using Huber loss function (0.596) slightly outperform the one using Gaussian loss function (0.612). Grid search with 5-fold cross validation technique is used for both models, with 10000 trees, 0.05 learning rate and stopping tolerance of 0.0001. First round of grid search is aimed to determine the minimum and maximum depth of the model, while the second round is utilized to tune the other parameters such as sample rate, column sampling rates per split, column sampling rates per tree etc. (Details are commented in the GBM code). The final prediction is generated using the average of the predictions made by the top 10 cross-validation models.

Based on the GBM model with Gaussian and Huber loss function, the top 5 variables that contribute the most to the variations in the response (log(LOS)) are listed as the following with their corresponding percentages:

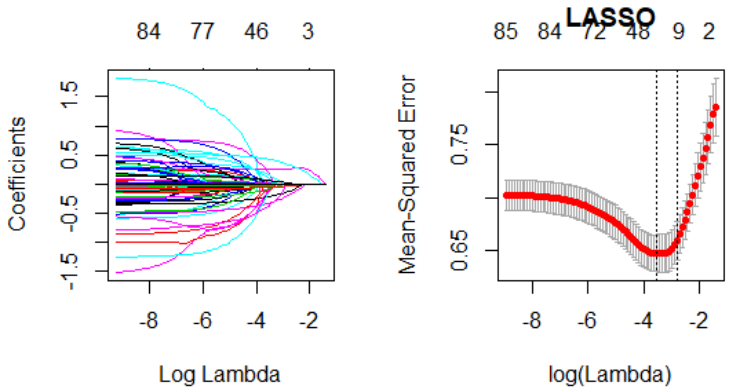


Figure 11: LASSO model

1. Discharge Location: 8.5%
2. Religion 6.6%
3. Respiratory Rate 5.8%
4. LOS from Admission to Transfer to ICU 5.0%
5. Average SpO2 4.9%

The variable importance for GBM model with Huber loss function:

1. Discharge Location: 6.8%
2. Religion 6.6%
3. Respiratory Rate 5.6%
4. Max Hematocrit 5.4%
5. Average Blood Pressure 4.8%

4.4 GLMNET

In the penalized linear model, LASSO, ridge regression and elastic net models along with models with alpha value ranging from 0.1 to 0.9 are evaluated using 5-fold cross validation. Two different types of models are constructed: one using lambda value that is produced by cross-validation which reaches the first standard error and the other using lambda from the cross-validation model with minimum error/MSE. The difference between those 2 models is that the first model (using lambda.1se) is within 1 standard error of the best model so that the model is rather simpler compared to the model using lambda value with minimum CV error (using lambda.min).

Figure 11, 12 and 13 are the cross validation results on determining the value of lambda for LASSO model (alpha = 1), Ridge Regression model (alpha = 0) and elastic net model (alpha = 0.5).

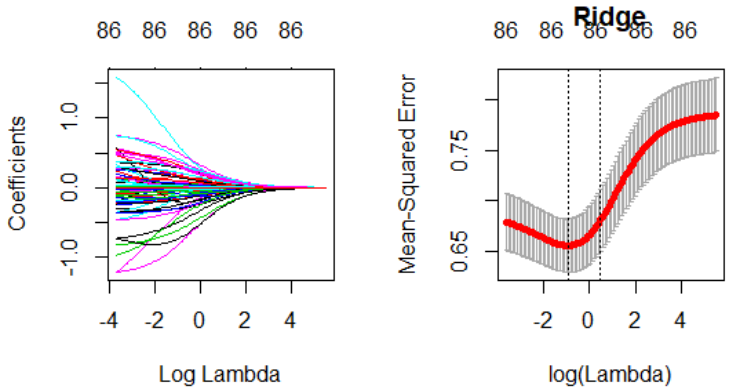


Figure 12: Ridge Regression model

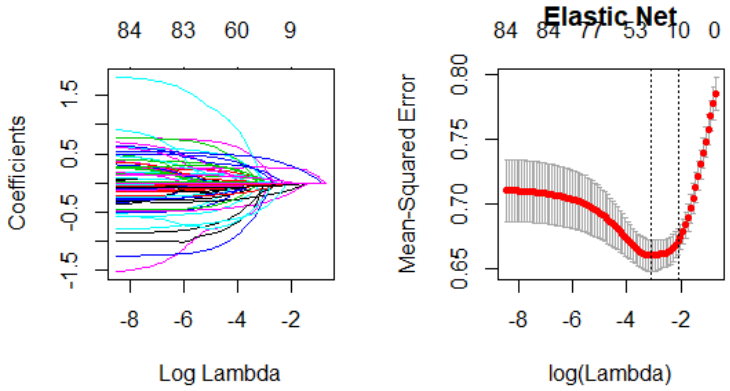


Figure 13: Elastic Net model

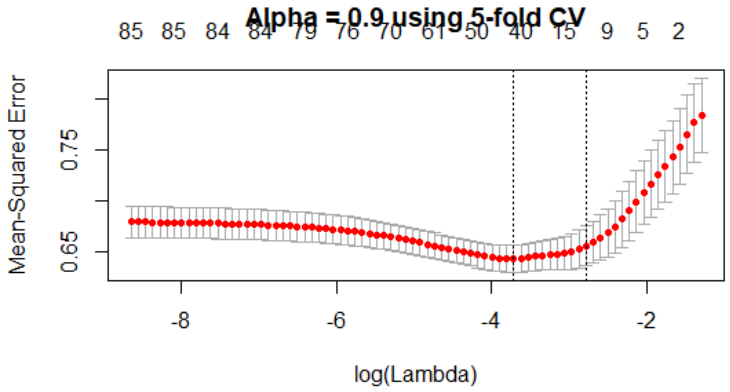


Figure 14: Using Lambda.1se

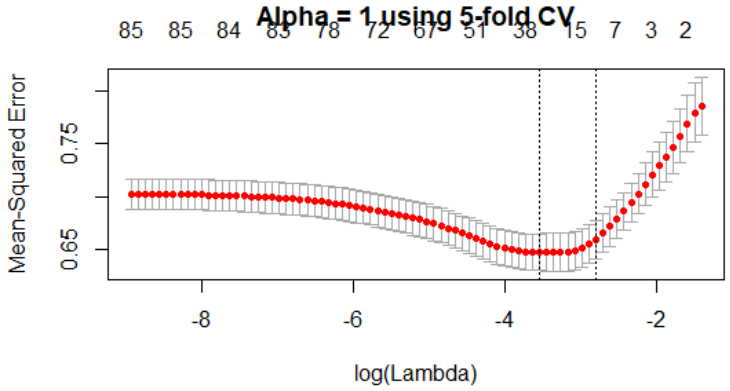


Figure 15: Using Lambda.min

The next tunning procedure is to train the penalized regression models using different values of alpha, ranging from 0 to 1 by 0.1 and select alpha value based on the smallest MSE using 5-fold cross validation. As explained previously, lambda.1se and lambda.min are used in predicting log of ICU LOS on the testing set. Results are shown in Figure 14 and 15. When using lambda.1se, the best penalized regression model is when alpha = 0.9. When using lambda.min, the best mode is when alpha = 1.

4.5 Support Vector Machine

In using support vector regression (SVR) model, 2 sets of hyperparameters are tuned through 5-fold cross validation. Namely, epsilon and cost parameters. Epsilon is the penalty in the loss function and when training the regression function, predictions that are within epsilon distance from the actual/true values will be given no penalty. For predictions that are greater than epsilon distance from the true value, cost parameter will determine the amount of penalty those

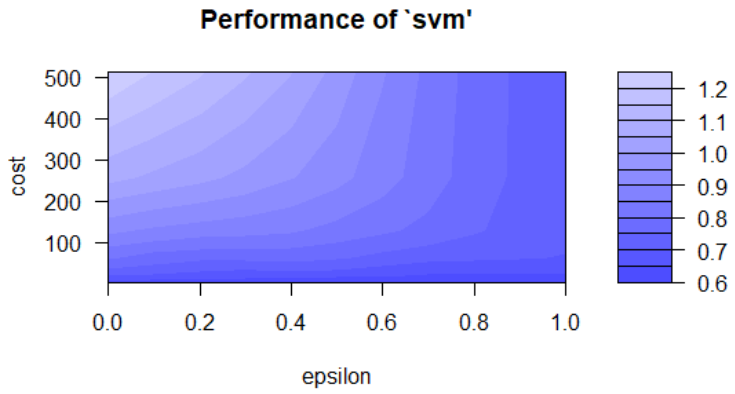


Figure 16: Grid Search with 5-fold Cross Validation Results

predicted points receive. Thus, the epsilon parameter controls the range of penalty enforced in the model training and the cost parameter determines the amount of penalty those penalized predicted values receive. A grid search along with 5-fold cross validation is implemented, with epsilon from 0 to 1 by 0.1 and cost from 2^2 to 2^9. In *Figure 16*, the 5-fold cross validation results are plotted (the darker the region, the closer the value of MSE to zero and in terms the better the model). And the epsilon and cost values for the best CV model is summarized below.

```
Parameter tuning of svm:

- sampling method: 5-fold cross validation

- best parameters:
 epsilon cost
     0.9    4

- best performance: 0.6093066
```

4.6 Ensemble Methods

Three types of ensemble methods are implemented: averaging, weighted average and stacking. The first 2 methods are relatively simple but less effective in improving the prediction accuracy on the testing set. The averaging method is taking a simple average of the estimated responses $\hat{Y}_{test}$ from all 6 basic models and use it as the final prediction on the testing set. This method yields a prediction error (MSE) of **0.905** which is higher than the prediction error of any individual basic models. The second method of weighted average is to take a empirical weighted average of the estimated responses $\hat{Y}_{test}$ from all 6 basic models, which yields a prediction error (MSE) of **0.592**. However, the weights are empirically determined based on the prediction errors (MSE) for each basic model. In other words, basic models that have lower out-of-sample prediction errors will be given higher weights.

For the stacking method, after training all 6 basic models (MARS, GLMNET, RF, GBM (Gaussian and Huber loss functions), SVM), the estimated responses for both training set ($\hat{Y}_{train}$) and testing set ($\hat{Y}_{test}$) are saved. As illustrated in the diagram of the analytical framework, on the

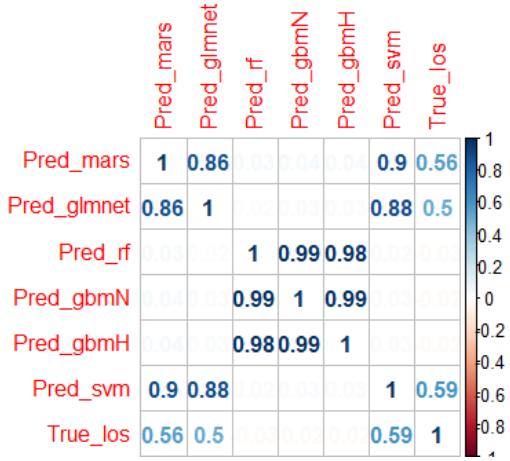


Figure 17: Correlation between predictions from each basic model

top layer, the true responses for the training set Y_{train} are fitted on the estimated responses for training set ($\hat{Y}_{train}$). After the top-layer model is trained, it's applied on the estimated responses $\hat{Y}_{test}$ in order to improve the prediction accuracy on the testing set.

However, one point to note when using stacking method is that the predictions produced by each basic model should be less correlated with each other. In this way, the stacked model can be most effective in improving the prediction accuracy on the testing set. In *Figure 17*, the correlation plot containing predictions from all six basic models indicates that there are high correlations between predictions from MARS model and SVM model, between GLMNET mode and SVM model, between RF model and GBM_Normal and GBM_Huber model and clearly between GBM_Normal and GBM_Huber model. Thus, the predictions that are used in the stacking methods are from SVM model and GBM_Huber model. The correlation between them are illustrated in *Figure 18*. The prediction error obtained from the stacked model is **0.624**. Recall, the original out-of-sample prediction errors for each basic model are:

- GBM_Huber: 0.596
- Random Forest: 0.611
- GBM_Normal: 0.612
- MARS: 0.624
- SVM: 0.627
- GLMNET: 0.640

Thus, the stacked model outperforms MARS, SVM and GLMNET models but still underperforms GBM_Huber, Random Forest and GBM_Normal models.

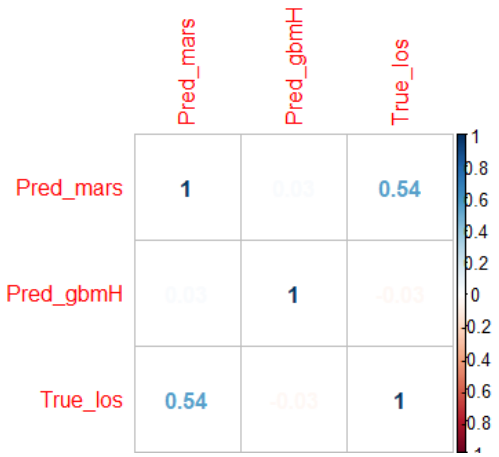


Figure 18: Correlation between predictions from 2 basic models

5 Conclusion & Future Work

5.1 Statistical Conclusion

According to the generalized cross validation error/MSE, the best model is produced by the multivariate adaptive regression spline, which yields a lowest CV error of 0.575. Both random forest and support vector machine models performs relatively similar, which have CV errors of 0.591 and 0.609 respectively. All of these 3 basic models are arguably easy to interpret. MARS model essentially fit a local regression that resemble stepwise linear regression. Random forest model is a tree-based model that can be easily understood using the analogy of decision trees. Support vector regression can be interpreted as a recursive binary partition of the sample space and the mean value within each partitioned region is used as the predicted response. The gradient boosting method with both Gaussian and Huber loss functions did not produce satisfactory results. However, compared to the tunning procedures for all other basic models, GBM model can be most flexible and there are plenty of hyperparameters that can be adjusted. Due to the relatively small data size, a highly sensitive GBM model may overfit the data or model the noise within the data rather than the true underlying trend. All of the basic model implemented have relatively short computation time. The first obvious reason is the small sample size and another reason is *caret* and *h2o* packages in R are all computationally efficient.

5.2 General Conclusion & Future work

Based on the variable importance measures from all basic models, 3 variables seems to be very influential on the length of stay in the ICU. Namely, discharge location, whether the patients have been intubated and the length of time between admission to hospital to transfer to the ICU. Generally speaking, the type of discharge location is sometime predictive in patient's physical conditions. For instance, patients with severe chronic diseases might be commonly discharged to specialized facilities (skilled nursing facilities (SNF)) or home with health care services. If those type of patients ended up in the ICU, their conditions may deteriorate much faster compared to other patients. Intubation is a commonly used procedure in the ICU, which is usually conducted on seriously ill patients. Study has also suggested that such procedure also poses great risk of having airway emergency in the ICU [4]. Therefore, it's also reasonable that multiple models

considered it as an important variable. Last but not least, the derived variable of length of time between admission to hospital to transfer to the ICU also appeared to be predictive on patients' LOS in the ICU. The reason could be if such time period is relatively short, it's a indicator that the patient's disease is at its advanced stage (e.g. stage IV cancer vs. stage II cancer) and could deteriorates fast. Longer time between hospital admission and ICU transfer allows the patient to get more chances to be treated with more moderate approaches.

Overall, patient's vital signs and lab results play less influential roles in predicting his/her LOS in the ICU. However, this could be caused by only using summarized measurements (e.g. average blood pressure during the first 24 hours of his ICU stay), instead of chronological measures (e.g. blood pressure at individual time points during the first 24 hours of his ICU stay). This could serve as one of the future works that could be conducted. Additionally, more diverse basic models and more complex stacking algorithm can be used to further improve the prediction accuracy. During the stacking step, other features from the training set can be used to train the stacked model. For the data extraction and processing stage, larger training dataset could be collected and model estimation can focus on different types of surgical patients, as patients who went through more complex or risky surgical procedure may behave differently from patients who went through relatively mild procedures. Lastly, since the MIMIC III database contains information on all types of patients, separate models for non-surgical patients could also be estimated in order to reduce the ICU LOS.

6 Appendix

Unused tables from MIMICIII database:

- **Callout**: Provides information when a patient was READY for discharge from the ICU, and when the patient was actually discharged from the ICU
- **Caregivers**: Defines the role of caregivers
- **cptevents**: Contains current procedural terminology (CPT) codes, which facilitate billing for procedures performed on patients
- **DatetimeEvents**: Contains all date formatted data
- **diagnoses_icd**: Contains ICD diagnoses for patients, most notably ICD-9 diagnoses
- **inputevents_cv**: CareVue ICU databases: Input data for patients
- **inputeevents_mv**: Metavision ICU databases: Input data for patients
- **MicrobiologyEvents**: Contains microbiology information, including tests performed and sensitivities
- **NoteEvents**: Contains all notes for patients
- **OutputEvents**: Output data for patients
- **Prescriptions**: Contains medication related order entries, i.e. prescriptions
- **ProcedureEvents_mv**: Metavision ICU database: Contains procedures for patients
- **Procedures_icd**: Contains ICD procedures for patients, most notably ICD-9 procedures
- **transfers**: Physical locations for patients throughout their hospital stay

The following R output are model summary statistics that may take longer time to train. All the following outputs and additional model information can be generated from corresponding codes.

6.1 Random Forest

Complete model summary for final Random Forest model:

```
Model Details:
==============

H2ORegressionModel: drf
Model Key:  mygrid_model_0
Model Summary:
  number_of_trees number_of_internal_trees model_size_in_bytes min_depth max_depth mean_depth min_leaves max_leaves mean_leaves
1             100                      100                670984        18        20    19.91000        419        498    465.60000
```

```
H2ORegressionMetrics: drf
** Reported on training data. **
** Metrics reported on Out-Of-Bag training samples **

MSE:  0.5970724
RMSE:  0.7727046
MAE:  0.6198545
RMSLE:  0.3584729
Mean Residual Deviance :  0.5970724


H2ORegressionMetrics: drf
** Reported on validation data. **

MSE:  0.6511805
RMSE:  0.8069576
MAE:  0.6436294
RMSLE:  0.347778
Mean Residual Deviance :  0.6511805


H2ORegressionMetrics: drf
** Reported on cross-validation data. **
** 5-fold cross-validation on training data (Metrics computed for combined holdout pr

MSE:  0.5902199
RMSE:  0.7682577
MAE:  0.6185527
RMSLE:  0.3562588
Mean Residual Deviance :  0.5902199


Cross-Validation Metrics Summary:
                        mean           sd cv_1_valid cv_2_valid cv_3_valid cv_4_valid
mae                  0.6185133 0.015480896  0.6089592  0.6358921  0.6260065 0.58060694
0.6411017
mse                 0.59050554  0.01370894  0.5985512  0.5881169 0.59274924 0.55678385
0.6163266
r2                  0.20320927 0.022755845 0.16176638 0.24228182 0.23418537 0.20572068
residual_deviance   0.59050554  0.01370894  0.5985512  0.5881169 0.59274924 0.55678385
0.6163266
rmse                  0.768339  0.00896473  0.7736609  0.7668878  0.7699021
0.7461795  0.7850647
rmsle               0.35626876 0.009842637 0.35652632 0.34271267 0.34935236 0.35005474

Scoring History:
             timestamp    duration number_of_trees training_rmse training_mae training_
1  2018-03-30 15:47:51   7.211 sec                0
```

```
2 2018-03-30 15:47:52   8.487 sec              100        0.77270        0.61985
0.59707           0.80696
  validation_mae validation_deviance
1
2        0.64363              0.65118

Variable Importances: (Extract with 'h2o.varimp')
=================================================

Variable Importances:
             variable relative_importance scaled_importance percentage
1 discharge_location          3887.361816           1.000000   0.105312
2            religion          2353.844727           0.605512   0.063768
3       resprate_mean          2156.193115           0.554667   0.058413
4       endotrachflag          2129.545654           0.547813   0.057691
5      hematocrit_max          1982.207642           0.509911   0.053700

___
           variable relative_importance scaled_importance percentage
22         language           964.327332           0.248067   0.026125
23        ethnicity           600.091064           0.154370   0.016257
24            mingcs           351.244049           0.090355   0.009516
25  admission_type             258.037415           0.066379   0.006990
26 first_hosp_stay             186.567307           0.047993   0.005054
27           gender           179.493637           0.046174   0.004863
```

6.2 GBM

```
GBM with Normal distribution
Variable Importances:
             variable relative_importance scaled_importance percentage
1  discharge_location           791.581421           1.000000   0.085777
2            religion           613.663330           0.775237   0.066498
3       resprate_mean           536.683716           0.677989   0.058156
4         los_reg_trs           458.299805           0.578967   0.049662
5           spo2_mean           449.035065           0.567263   0.048658
6           tempc_mean           432.092468           0.545860   0.046822
7         glucose_mean           428.021484           0.540717   0.046381
8       potassium_max            412.944916           0.521671   0.044747
9         meanbp_mean            407.920319           0.515323   0.044203
10     hematocrit_max            407.049835           0.514224   0.044109

GBM with Huber distribution
Variable Importances:
             variable relative_importance scaled_importance percentage
1  discharge_location           413.145111           1.000000   0.067799
2            religion           403.225494           0.975990   0.066171
3       resprate_mean           343.891724           0.832375   0.056435
```

4	hematocrit_max	328.869537	0.796015	0.053969
5	meanbp_mean	290.473297	0.703078	0.047668
6	spo2_mean	286.330414	0.693050	0.046988
7	tempc_mean	283.576630	0.686385	0.046536
8	sysbp_mean	279.122528	0.675604	0.045806
9	glucose_mean	260.558228	0.630670	0.042759
10	endotrachflag	257.691162	0.623730	0.042289

6.3 MARS

Cross-validation mode with lowest CV error (degree = 2 penalty= 3):

```
> summary(mars3)
Call: earth(x=x_train, y=y_train, pmethod="cv", trace=3, degree=2, nfold=5)

                                                                                coefficient
(Intercept)
1.7277110
discharge_locationHOME
-0.4978550
discharge_locationHOMEHEALTHCARE
-0.3470001
h(tempc_mean-37.2381)
0.2871957
h(91-bun_max)
-0.0065395
discharge_locationREHAB/DISTINCTPARTHOSP * religionEPISCOPALIAN
1.2008066
discharge_locationREHAB/DISTINCTPARTHOSP * marital_statusSEPARATED
2.2990188
discharge_locationREHAB/DISTINCTPARTHOSP * marital_statusUNKNOWN(DEFAULT)
-1.2678165
discharge_locationREHAB/DISTINCTPARTHOSP * h(13-mingcs)
0.0802119
languageENGL * h(gcsmotor-5)
-0.3075883
languageRUSS * h(resprate_mean-24.16)
0.7643770
h(0.958183-los_reg_trs) * endotrachflag
0.8446303
h(los_reg_trs-0.958183) * h(sysbp_mean-134.6)
-0.0071489
h(los_reg_trs-0.958183) * h(138-sodium_max)
0.0073147
h(107.167-sysbp_mean) * h(91-bun_max)
0.0006902
h(80.5676-meanbp_mean) * h(97.24-spo2_mean)
0.0120025
h(meanbp_mean-80.5676) * h(97.24-spo2_mean)
0.0191689
```

```
h(76.5−meanbp_mean) * h(91−bun_max)
−0.0003978
h(tempc_mean−37.3769) * h(5−gcsmotor)
0.2663620
h(spo2_mean−97.24) * h(0.5−creatinine_max)
−3.6264873
h(spo2_mean−97.24) * h(53.2−weight_first)
0.0368831

Selected 21 of 43 terms, and 21 of 174 predictors using pmethod="cv"
Termination condition: RSq changed by less than 0.001 at 43 terms
Importance: los_reg_trs, endotrachflag, discharge_locationHOME, discharge_locationHOME
Number of terms at each degree of interaction: 1 4 16
GRSq 0.3147454  RSq 0.3815578  mean.oof.RSq 0.1500017 (sd 0.0439)

pmethod="backward" would have selected:
    26 terms 22 preds,  GRSq 0.3201573  RSq 0.402482  mean.oof.RSq 0.1413511
```

Cross-validation model with lowest out-of-sample prediction error (degree = 1 penalty= 2):

```
> summary(mars4)
Call: earth(x=x_train, y=y_train, pmethod="cv", trace=3, degree=1, nfold=5)

                                          coefficients
(Intercept)                                0.92196038
discharge_locationHOME                    −0.41788704
discharge_locationHOMEHEALTHCARE          −0.22556501
discharge_locationREHAB/DISTINCTPARTHOSP    0.22705038
languageENGL                              −0.17981084
endotrachflag                              0.46939940
h(0.958183−los_reg_trs)                    0.42288215
h(107.293−sysbp_mean)                      0.03756458
h(meanbp_mean−84.3889)                     0.01613580
h(resprate_mean−24.16)                     0.11586673
h(tempc_mean−37.2381)                      0.48914393
h(97.24−spo2_mean)                         0.09085830
h(bun_max−8)                               0.00545856
h(13−mingcs)                               0.05148476
h(height_first −160.02)                   −0.26399992
h(height_first −160.866)                   0.26600192

Selected 16 of 96 terms, and 14 of 174 predictors using pmethod="cv"
Termination condition: RSq changed by less than 0.001 at 96 terms
Importance: los_reg_trs, discharge_locationHOME, discharge_locationHOMEHEALTHCARE, er
Number of terms at each degree of interaction: 1 15 (additive model)
GRSq 0.2684035  RSq 0.3116408  mean.oof.RSq 0.2029307 (sd 0.0417)

pmethod="backward" would have selected:
    32 terms 20 preds,  GRSq 0.2789237  RSq 0.3655654  mean.oof.RSq 0.1820723
```

References

[1] Johnson AEW, Pollard TJ, Shen L, Lehman L, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, and Mark RG. Mimic-iii, a freely accessible critical care database. *Nature*, 2016.

[2] T Bardell, JF Legare, KJ Buth, GM Hirsch, and IS Ali. Icu readmission after cardiac surgery. *European journal of cardio-thoracic surgery*, 23(3):354–359, 2003.

[3] Lynda Copeland-Fields, Terry Griffin, Trisha Jenkins, Maggie Buckley, and Lowell C Wise. Comparison of outcome predictions made by physicians, by nurses, and by using the mortality prediction model. *American Journal of Critical Care*, 10(5):313, 2001.

[4] Jigeeshu V Divatia, Parvez U Khan, and Sheila N Myatra. Tracheal intubation in the icu: Life saving or life threatening? *Indian journal of anaesthesia*, 55(5):470, 2011.

[5] Laleh Gharacheh, Amin Torabipour, Farzad Faraji Khiavi, Amal Saki Malehi, and Maryam Haddadzadeh. Comparison of statistical models of predict the factors affecting the length of stay (los) in the intensive care unit (icu) of a teaching hospital. *Materia socio-medica*, 29(2):88, 2017.

[6] Johnson, Alistair EW, David J. Stone, Leo A. Celi, and Tom J. Pollard. The mimic code repository: enabling reproducibility in critical care research. *Journal of the American Medical Informatics Association*, 2017.

[7] William A Knaus, Douglas P Wagner, Elizabeth A Draper, Jack E Zimmerman, Marilyn Bergner, Paulo G Bastos, Carl A Sirio, Donald J Murphy, Ted Lotring, Anne Damiano, et al. The apache iii prognostic system: risk prediction of hospital mortality for critically iii hospitalized adults. *Chest*, 100(6):1619–1636, 1991.

[8] Andrew A Kramer and Jack E Zimmerman. A predictive model for the early identification of patients at risk for a prolonged intensive care unit length of stay. *BMC medical informatics and decision making*, 10(1):27, 2010.

[9] Rocco J LaFaro, Suryanarayana Pothula, Keshar Paul Kubal, Mario Emil Inchiosa, Venu M Pothula, Stanley C Yuan, David A Maerz, Lucresia Montes, Stephen M Oleszkiewicz, Albert Yusupov, et al. Neural network prediction of icu length of stay following cardiac surgery based on pre-incision variables. *PloS one*, 10(12):e0145395, 2015.

[10] Jean-Roger Le Gall, Stanley Lemeshow, and Fabienne Saulnier. A new simplified acute physiology score (saps ii) based on a european/north american multicenter study. *Jama*, 270(24):2957–2963, 1993.

[11] Joon Lee, Sapna Govindan, Leo A Celi, Kamal R Khabbaz, and Balachundhar Subramaniam. Customized prediction of short length of stay following elective cardiac surgery in elderly patients using a genetic algorithm. *World journal of cardiovascular surgery*, 3(5):163, 2013.

[12] John L Moran and Patricia J Solomon. A review of statistical estimators for risk-adjusted length of stay: analysis of the australian and new zealand intensive care adult patient data-base, 2008–2009. *BMC medical research methodology*, 12(1):68, 2012.

[13] Antonio Paulo Nassar Jr and Pedro Caruso. Icu physicians are unable to accurately predict length of stay at admission: a prospective study. *International Journal for Quality in Health Care*, 28(1):99–103, 2015.

[14] Alramzana Nujum Navaz, Elfadil Mohammed, Mohamed Adel Serhani, and Nazar Zaki. The use of data mining techniques to predict mortality and length of stay in an icu. In *Innovations in Information Technology (IIT), 2016 12th International Conference on*, pages 1–5. IEEE, 2016.

[15] Minna Niskanen, Matti Reinikainen, and Ville Pettilä. Case-mix-adjusted length of stay and mortality in 23 finnish icus. *Intensive care medicine*, 35(6):1060–1067, 2009.

[16] Adriana Pérez, Wenyaw Chan, and Rodolfo J Dennis. Predicting the length of stay of patients admitted for intensive care using a first step analysis. *Health Services and Outcomes Research Methodology*, 6(3-4):127–138, 2006.

[17] T. J. Pollard and A. E. W. Johnson. The mimic-iii clinical database, 2016.

[18] Ilona Willempje Maria Verburg, Alireza Atashi, Saeid Eslami, Rebecca Holman, Ameen Abu-Hanna, Everet de Jonge, Niels Peek, and Nicolette Fransisca de Keizer. Which models can i use to predict adult icu length of stay? a systematic review. *Critical care medicine*, 45(2):e222–e231, 2017.