Predicting the Length of Stay in ICU using MIMIC II Dataset

Xinyue Qiu, Ji Yoon Lee, Weiyi Li

1. Abstract

Background and Aim. Long stays in the Intensive Care Units (ICUs) are an economic burden to individuals and the economy as a whole. The main goal of hospitals is to decrease the length of stay(LOS). In this study, we identify key risk factors of ICU stay with four different models. *Materials and Method*. 12 explanatory variables (clinical and demographic) and 2 outcome variables (length stay and length of stay group) of 2,234 patient records were used to build Elastic Net regression model, Regression Tree model, Random Forest regression, and Artificial Neural Network (ANN) models. *Results*. The test MSE of the elastic net regression model is 52.0; the regression tree model had a test MSE = 53.4; the random forest regression had a test MSE = 59.7; the ANN model for regression had a test MSE = 57.93 and the ANN model for classification had a model accuracy = 76.6%. Based on the models, we identified the important factors related with prediction of LOS as: SAPS II, age, gender, religion, marital status, weight, admission source, care unit. *Conclusion*. Our models indicate that SAPS II is associated with longer LOS, and weight, admission age, and ICU stay care unit being MICU is associated with shorter LOS.

2. Introduction and Background

Intensive Care Units (ICUs) are special treatment departments within a hospital for patients with severe or life-threatening illness and injuries(Toptas, 2018). Patients in the ICU require close monitoring, constant care, and rapid intervention by highly trained physicians and nurses. The high cost of ICU stays not only influences patients and their families, but also impacts the economy as a whole. ICUs have been a major driver of hospital costs and one of the goals of the United States hospitals is to optimize the number of ICU beds(Gruenber, 2006; Toptas, 2018; Oliveira, 2010).

The patient's length of stay (LOS) at the ICU is an important metric for the US healthcare system. Studies have shown that LOS is associated with certain demographic and clinical variables, but there have been a limited number of studies that predict LOS using machine learning models(Toptas, 2018; Moyer, 1994; Gardner, 2006). Knowing factors impacting LOS will significantly contribute to prediction of LOS and accordingly, provides insights into ways of optimizing the number of ICU beds. Additionally, insights of patients LOS enables hospitals to better regularize the ICU when immense needs of ICU beds arise. In this study, lasso regression, regression tree and random forest regression, and neural network models were developed to understand the risk factors of prolonged ICU LOS.

3. Methods

3.1 Data

The datasets for this study were extracted from the MIMIC II original ICU database. The data was collected over a period of 2001 and 2008, from medical, surgical, coronary care, and neonatal ICU care units. There were 37 tables of 26,870 samples related to patients and admissions.

3.1.1 Database Creation

After downloading the full dataset from Kaggle, the csv files were uploaded to Github for further interaction. The tables were joined by the key: *hadm_id* (Hospital admission ID).

By locating all desired variables, 5 tables were created from the original dataset using *sqlite3* library in Python. A relational database *mimic2.db* was created based on the tables. Figure 1 shows the Entity Relationship (ER) diagram representing the relationship between the admission ID (primary key) and other independent variables identified in the tables.

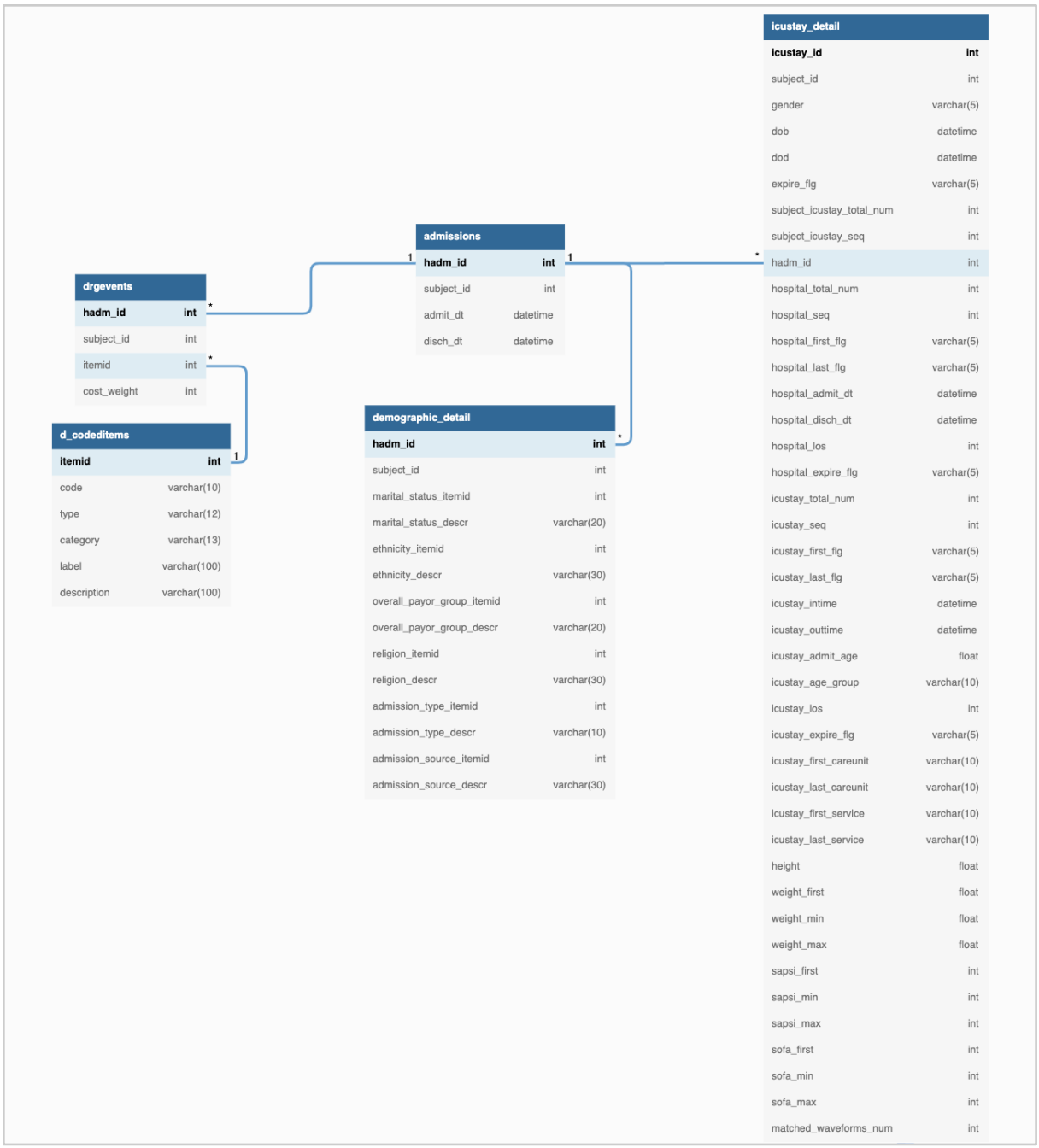


Figure 1: ER diagram of final dataset(mimic2.db)

3.1.2 Variables

12 demographic and clinical variables were extracted. The variables *length_stay* and *los_group* were created as outcome variables (*length_stay = discharge time – admit time*; *los_group = 1*: 0-6 days, *2*: 7-13 days, *3*: 14-20 days, *4*: 21+ days)(Moitra, 2016). The variable types and descriptions are provided in Table 1.

Variable Name	Type	Description
marital_status_descr	Explanatory	Description of patient's marital status
ethnicity_descr	Explanatory	Description of patient's ethnicity
religion_descr	Explanatory	Description of patient's religion
admission_type_descr	Explanatory	Description of patient's admission type to hospital
admission_source_descr	Explanatory	Description of patient's admission source
description	Explanatory	Basic description of patient's disease
gender	Explanatory	Patient's gender
weight_min	Explanatory	Minimum weight of patient's weight when in ICU
sapsi_max	Explanatory	Maximum SAPSI
sofa_max	Explanatory	Maximum SOFA
icustay_admit_age	Explanatory	Patient's age when admitted to ICU
icustay_first_careunit	Explanatory	First care unit type
length_stay	Outcome	ICU length of stay (days) Formula: icustay_outtime - icustay_intime
los_group	Outcome	ICU length of stay group classified based on length_stay 1: 0-6 days 2: 7-13 days 3: 14-20 days 4: 21+ days

Table 1: Descriptions of Explanatory and Outcome variables

3.1.3. Data Preprocessing

Missing data

Our dataset contains missing entries represented as “unknown”, “other” and “unobtainable”. Missing entries were converted into NA values and dropped from the dataset. The final table contained 2,234 complete patient admission records.

Variable Recode

The variable *ethnicity_descr* contains entries with repeated information like ‘White-Russian’ and ‘White’. For simpler interpretation, we converted the variable into generalized groups like ‘Black’, ‘Asian’, and ‘White’.

Correlation Check

To avoid multicollinearity, a heatmap was created to check the correlation between the numeric explanatory variables.

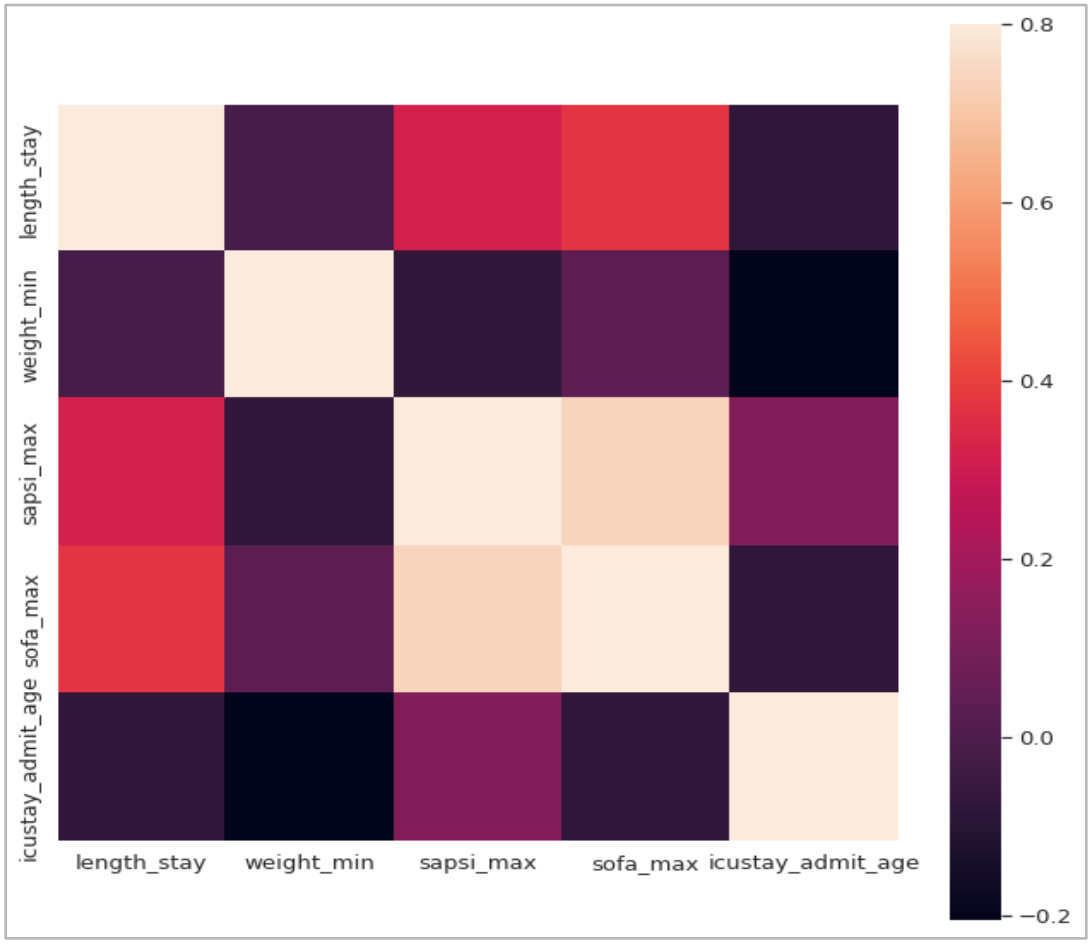


Figure 2: Heat map of numeric explanatory variables

Lighter color indicated that the two variables are more highly correlated. The above heat map has shown that *sapsi_max* and *sofa_max* have the highest correlation among all pairs. This is reasonable given they are both severity scores. *sofa_max* was thus removed from the variable set.

Variable	Value	Missing	Statistic
Marital Status, n (%)	DIVORCED	0	128 (5.7)
	MARRIED		1120 (50.2)
	SEPARATED		30 (1.3)
	SINGLE		414 (18.5)
	WIDOWED		541 (24.2)
Ethnicity, n (%)	AMERICAN INDIAN/ALASKA NATIVE	0	1 (0.0)
	ASIAN		19 (0.9)
	BLACK		227 (10.2)
	HISPANIC		38 (1.7)
	MULTI RACE ETHNICITY		1 (0.0)
	NATIVE HAWAIIAN OR OTHER PACIFIC ISLAND		2 (0.1)
	WHITE		1945 (87.1)
Religion, n (%)	7TH DAY ADVENTIST	0	3 (0.1)
	BAPTIST		6 (0.3)
	BUDDHIST		15 (0.7)
	CATHOLIC		1162 (52.0)
	CHRISTIAN SCIENTIST		15 (0.7)
	EPISCOPALIAN		41 (1.8)
	GREEK ORTHODOX		25 (1.1)
	HEBREW		2 (0.1)
	HINDU		1 (0.0)
	JEHOVAH'S WITNESS		12 (0.5)
	JEWISH		488 (21.9)
	LUTHERAN		1 (0.0)
	METHODIST		4 (0.2)

	MUSLIM		5 (0.2)
	PROTESTANT QUAKER		444 (19.9)
	ROMANIAN EAST. ORTH		4 (0.2)
	UNITARIAN-UNIVERSALIST		5 (0.2)
Admission Type, n (%)	ELECTIVE	0	216 (9.7)
	EMERGENCY		1955 (87.6)
	URGENT		62 (2.8)
Admission Source, n (%)	EMERGENCY ROOM ADMIT	0	1465 (65.6)
	REFERRAL		394 (17.6)
	TRANSFER		374 (16.7)
ICU Care Unit, n (%)	CCU	0	475 (21.3)
	CSRU		552 (24.7)
	FICU		224 (10.0)
	MICU		933 (41.8)
	SICU		49 (2.2)
Length of Stay Group, n (%)	0-6 days	0	1706 (76.4)
	14-20 days		97 (4.3)
	21+ days		128 (5.7)
	7-13 days		302 (13.5)
Gender, n (%)	F	0	1075 (48.1)
	M		1158 (51.9)
Weight, median [Q1,Q3]		0	73.0 [61.0,86.5]
SAPSI, median [Q1,Q3]		0	16.0 [13.0,20.0]
Age, median [Q1,Q3]		0	73.6 [61.7,82.3]
Length of Stay, mean (SD)		0	6.2 (9.6)

Table 2: Summary Statistics of Variables

3.2 Models

Four different machine learning models are developed to predict LOS. The models were trained on 70% of the data and validated on 30% of the remaining data. Test mean squared errors (MSE) were calculated to measure model performance.

Model 1: Multiple linear regression model with Elastic Net regularization

After converting all qualitative labels into dummy variables, a multiple linear regression model was implemented to predict the ICU LOS. Given the abundance of dummy variables, model feature space became sparse. To handle sparsity, we applied Elastic Net Regularization to our model. The regularization linearly combines the L1 and L2 penalties of the lasso and ridge methods and selects the meaningful features to model construction(Zou, 2005). By setting one parameter as the default value from the *ElasticNet* model in python, curves of coefficients were plotted to show how they decreased to 0. Cross validation was also performed accordingly for identification of the ideal model parameters for tuning the model. The final model was constructed using the *linear_model* package from the *scikit-learn* library in python.

Model 2: Regression Tree model

A regression tree model was developed to predict the ICU LOS by learning decision rules from the data(Loh, 2011). This method was used to visualize and interpret the decision factors of LOS and was implemented using the *DecisionTreeRegressor* function from the *scikit-learn* library in python. The number of instances parameter was tuned to minimize test MSE. The regression tree was built using the optimal number of instances and the variable importances were calculated.

Model 3: Random Forest regression model

A random forest regression model is developed to predict the ICU LOS by using an ensemble learning method for regression. This model was used to further prevent overfitting and thus increase model performance on the test data. The out-of-bag error (OOB) error was used to quantify error. This method was implemented using the *ensemble.RandomForestRegressor* from the *scikit-learn* library in python. The number of trees was tuned to minimize OOB error. The random forest regression was built using the optimal number of trees and the variable importances calculated.

Model 4: Neural Network Model

Out of the consideration of non-linearities of ICU LOS as well as the complex interconnected relationships between explanatory variables, we chose to use feedforward artificial neural network models(ANN) to predict LOS(Brownlee, 2022). The model construction was performed using *Keras* and *TensorFlow* packages in python. Given the difficulty in determining the ideal network size, we tried training multiple ANN models using different hyperparameters(Bebis, 1994). The optimization algorithm was chosen to be ‘stochastic backward propagation’, loss function being ‘mean absolute error’, and metrics being ‘mean squared error’. The trained

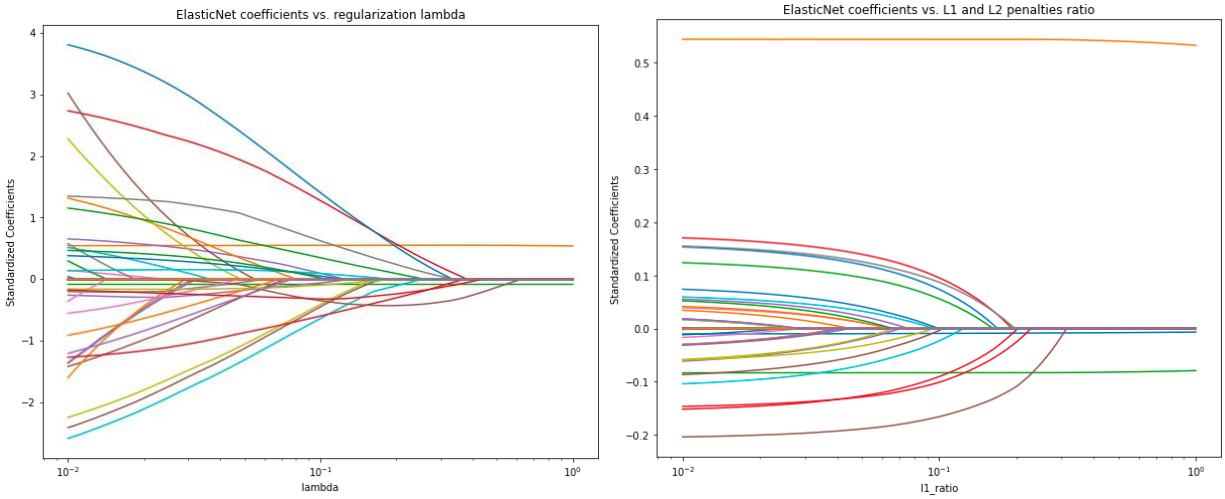
models were validated using a test dataset, with test MSE calculated accordingly. The model with lowest test MSE was identified as the best model.

To better interpret the performance of the ANN model, we considered using LOS groups(LOS_group) as the outcome variable and trained a multi-class classification model(Moitra, 2016). We adjusted the loss function to be 'categorical cross entropy', optimizer to be 'adam', and the metrics to be 'accuracy'(Zhang, 2018). The same procedure of training and validating ANN models using test dataset was repeated. The model with highest accuracy was identified as the best model.

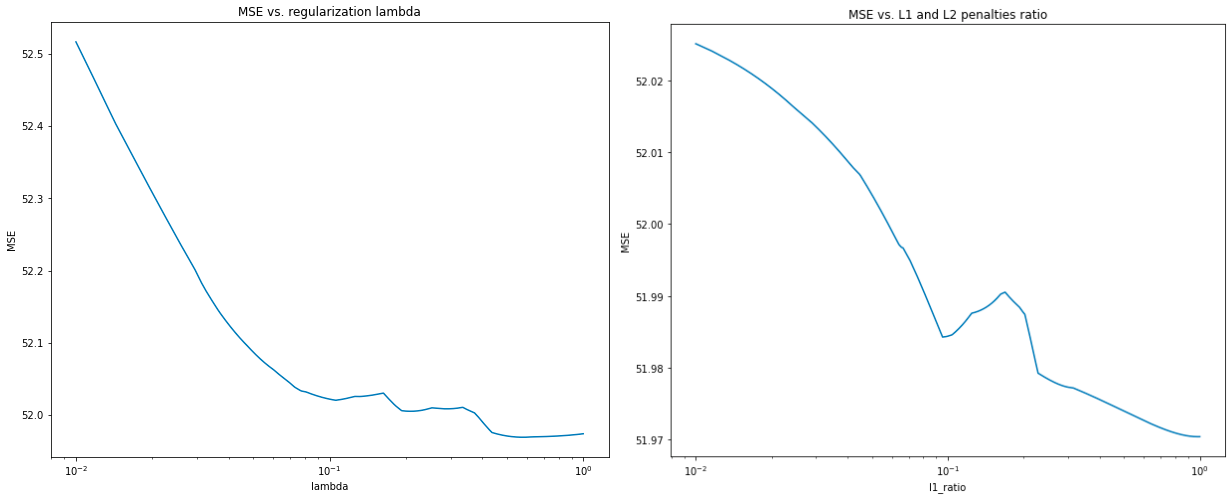
4. Results

Model 1: Multiple Linear Regression Model with Elastic Net Regularization

The multiple linear regression model has the lowest test MSE (MSE = 53.3). Many regression coefficients in the model are close to 0. Figure 3 depicted how coefficients and MSE converged during Elastic Net Regularization.



A) Coefficients Corresponding to Explanatory Variables over Elastic Net Parameters



B) Test MSE over Elastic Net Parameters

Figure 3: A) Plots of coefficients based on lambda or l1_ratio B) MSE curves based on lambda or l1_ratio

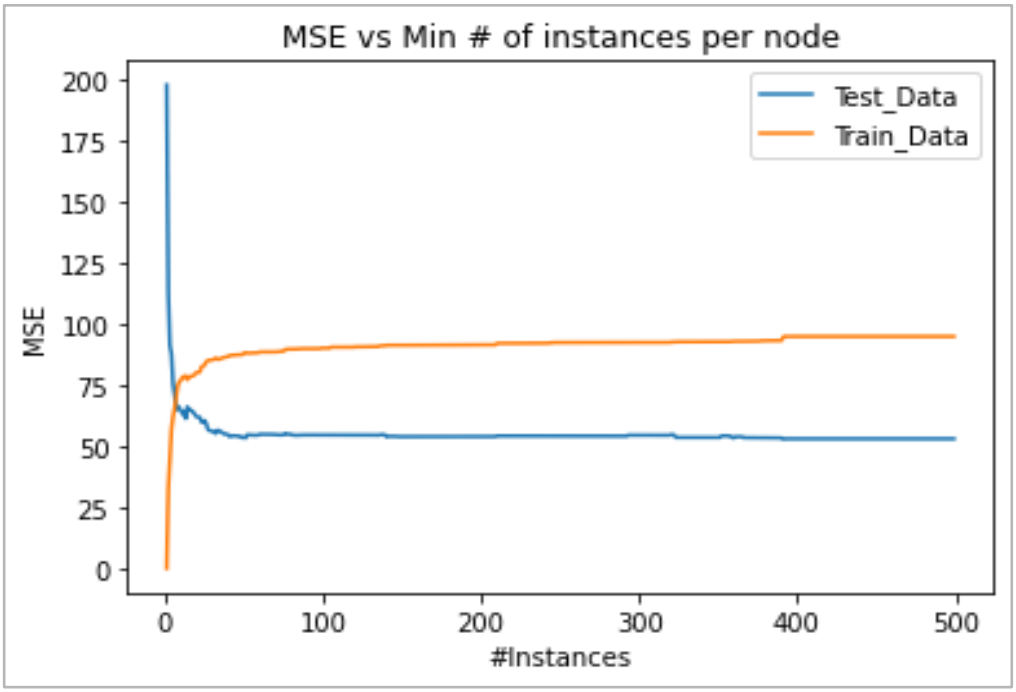
Optimal lamba and L1 ratio obtained from cross validation were 0.4256 and 0.5, respectively. The model's test MSE decreased (MSE = 52.0). The variables utilized for model constructions and their associated coefficients are summarized in Table 3.

Variable	Coefficient
weight_min	-0.00875038
sapsi_max	0.55251875
icustay_admit_age	-0.08333721
admission_source_descr_EMERGENCY ROOM ADMIT	-0.22448856
icustay_first_careunit_MICU	-0.0098205

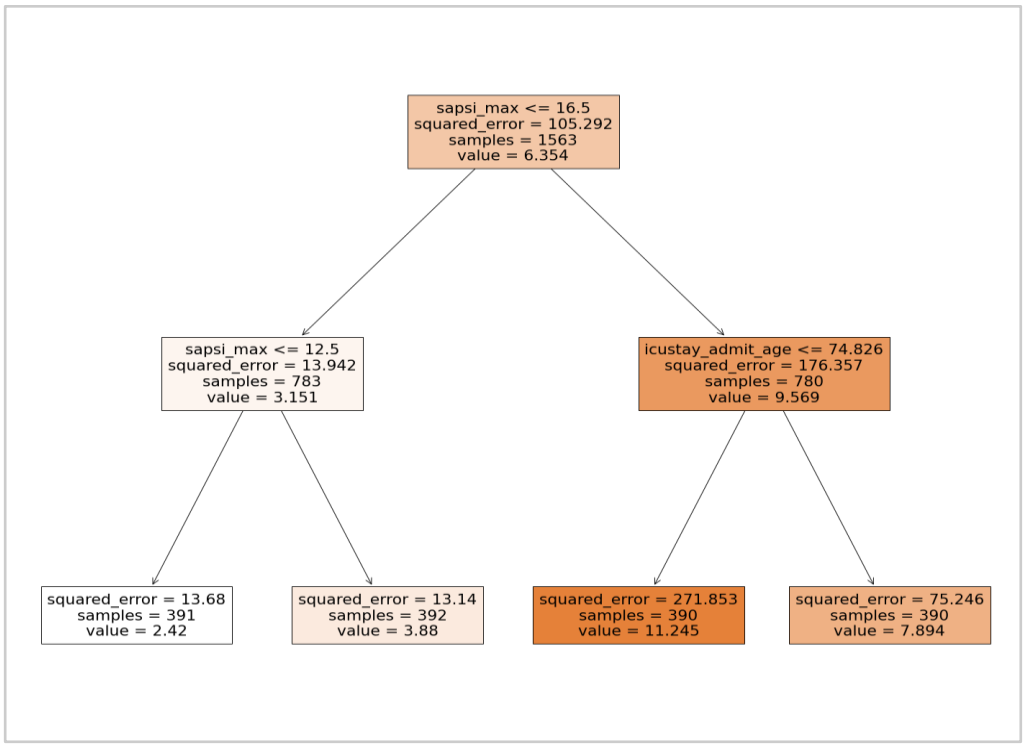
Table 3: Variables with Non-zero Coefficients in Elastic Net Regression

Model 2: Regression Tree Model

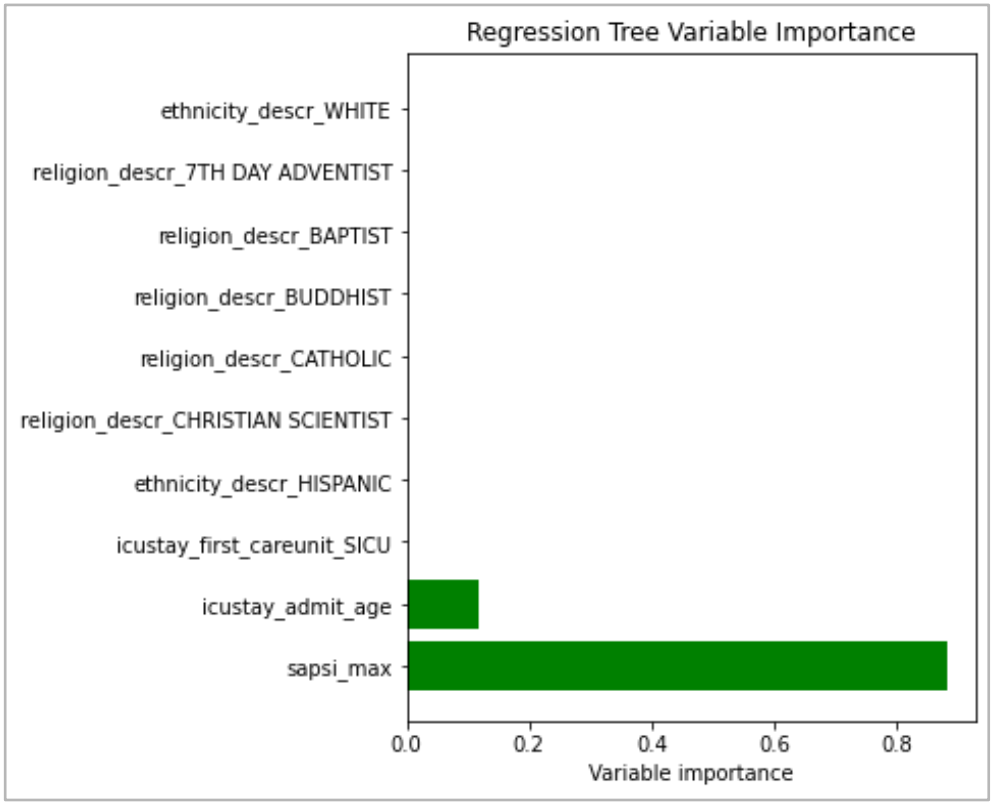
The regression tree model with the lowest test MSE (MSE=53.4) was obtained with the minimum number of instances per node set to 290. The two variables that predicted LOS are SAPS II (variable importance = 0.88) and age at ICU admissions (variable importance = 0.12). The model has identified four intervals of LOS based on explanatory variables: 390 samples with SAPS II greater or equal to 16.5 and ICU admissions age greater or equal to 75 (predicted LOS of 7.894); 390 samples with SAPS II greater or equal to 16.5 and ICU admissions age less than 75 (predicted LOS of 11.245); 392 samples with SAPS II greater or equal to 12.5 and smaller than 16.5 (predicted LOS of 3.88); 391 samples with SAPS II smaller than 12.5 (predicted LOS of 2.42).



A) Test and Train MSE with respect to the number of instances per node



B) Regression Tree Model Output

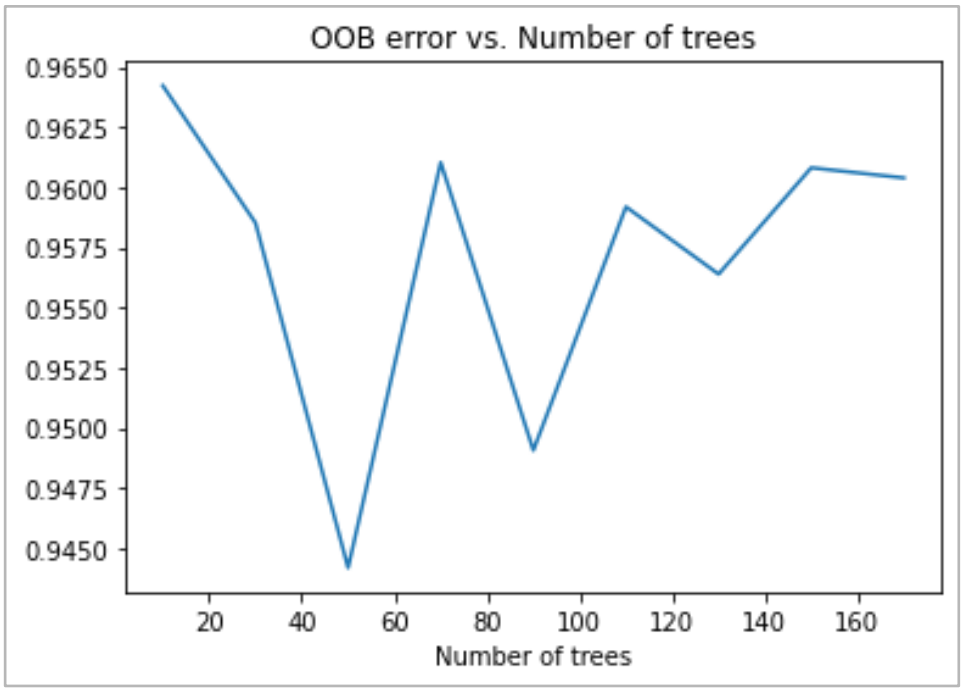


C) Regression Tree Variable Importance

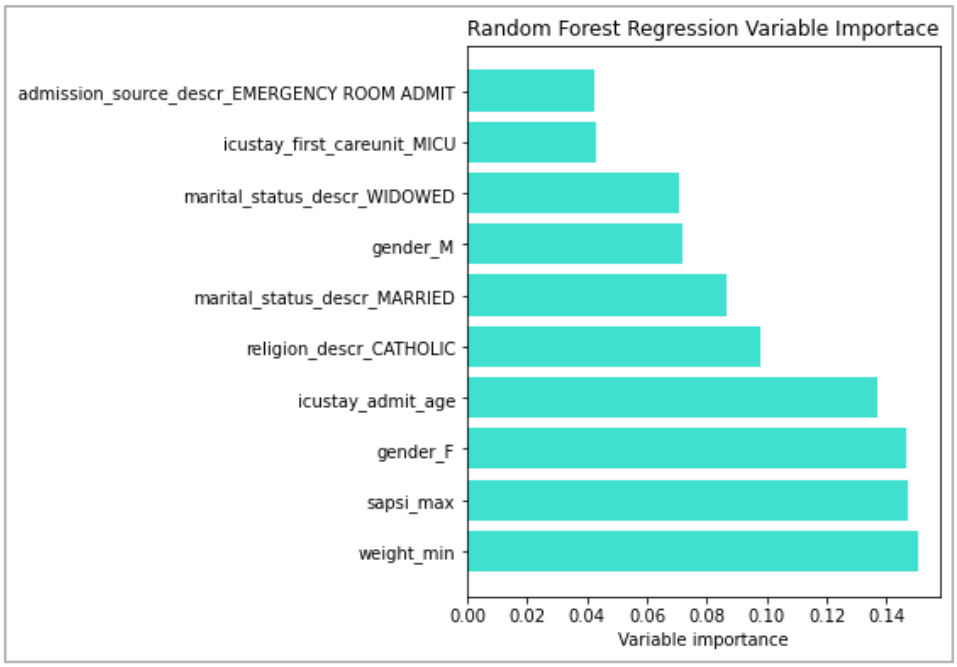
Figure 4. A) Plot showing test and train MSE with respect to the minimum number of instances per nodes; B) A regression tree model built with the minimum test MSE; test MSE = 53.4, #Instances = 290; C) Plot of the top 10 important variables for the regression tree with their relative importances

Model 3: Random Forest Regression Model

The random forest regression model with the lowest Out-Of-Bag (OOB) error was obtained when the number of trees was set to 50 (MSE = 59.79). The most important variables in growing the forests were weight (variable importance = 0.15), SAPS II (variable importance = 0.147), gender (variable importance = 0.147), ICU admissions age (variable importance = 0.137), religion (variable importance = 0.098), and ICU care unit(variable importance = 0.04), admission source(variable importance = 0.04).



A)OOB (Out-Of-Bag) error with respect to the number of trees



B)Random Forests Regression Variable Importance

Figure 5. A) Plot showing OOB error with respect to number of trees; B) Plot of the top 10 important variables for the random forest regression with their relative importances.

Model 4: Neural Network Model

The best regression model identified using artificial neural networks has the lowest test MSE (MSE = 57.93). The best multi-class classification model identified has the highest test accuracy (accuracy = 76.6%). Figure 6 shows the model summaries of the regression and classification model.

Layer (type)	Output Shape	Param #
dense_108 (Dense)	(None, 128)	38784
dense_109 (Dense)	(None, 256)	33024
dropout_64 (Dropout)	(None, 256)	0
dense_110 (Dense)	(None, 256)	65792
dense_111 (Dense)	(None, 1)	257

Total params: 137,857
Trainable params: 137,857
Non-trainable params: 0

A)Model summary of Artificial Neural Network Regression model

```
 Layer (type)                 Output Shape              Param #
=================================================================
 dense_100 (Dense)            (None, 128)               38784

 dense_101 (Dense)            (None, 256)               33024

 dropout_58 (Dropout)         (None, 256)               0

 dense_102 (Dense)            (None, 256)               65792

 dropout_59 (Dropout)         (None, 256)               0

 dense_103 (Dense)            (None, 256)               65792

 dropout_60 (Dropout)         (None, 256)               0

 dense_104 (Dense)            (None, 256)               65792

 dropout_61 (Dropout)         (None, 256)               0

 dense_105 (Dense)            (None, 256)               65792

 dropout_62 (Dropout)         (None, 256)               0

 dense_106 (Dense)            (None, 256)               65792

 dropout_63 (Dropout)         (None, 256)               0

 dense_107 (Dense)            (None, 4)                 1028

=================================================================
Total params: 401,796
Trainable params: 401,796
Non-trainable params: 0
```

B)Model Summary of Artificial Neural Network Classification model

Figure 6. A) model summary of Artificial Neural Network regression model. The model contains 3 hidden layers, 1 dropout layer, 1 input and 1 output layer; B) model summary of Artificial Neural Network classification model. The model contains 3 hidden layers, 3 dropout layers, 1 input and 1 output layer.

Based on the multi-class classification model, we were able to interpret the model outcome based on the *predict* function from *tensorflow*. When provided with basic patient information that aligns with our independent variable, the model is capable of yielding probabilities that indicate which LOS group the patient would belong to.

Model Result Summary

Model name	**Metrics**	**Type of model**
Elastic Net Regression Model	52.0 (Test MSE)	Regression
Regression Tree Model	53.4 (Test MSE)	Regression
Artificial Neural Network	57.9 (Test MSE)	Regression
Artificial Neural Network	76.6% (Test Accuracy)	Classification

Table 4: Model Result Summary

Based on the model output, we identified Elastic Net regression model as the best model with lower MSE of 52.0. The artificial neural network classification model has a highest test accuracy of 76.6%.

5. Conclusion

This paper explored the relationship between length of stay in ICU and basic clinical variables using a dataset from ICU bedside station, covering over the period from 2001 to 2008. Our model successfully identifies possible factors that contribute to longer stay at ICU. Based on the Elastic Net regression model, several factors that tend to decrease and increase the risk of longer ICU stays were identified. We have found that a higher severity score (SAPS II) contributes to longer stay in ICU. Having a higher weight, being older, and staying at MICU contribute to shorter stay in ICU. This finding aligns with our findings from the regression tree model. Based on our regression tree model, we have identified the factors contributing to ICU stay as: admission source, icu stay care unit, marital status, icu admission age, gender, severity score(SAPS II) and weight minimum.

6. Limitations and Future Directions

There are a few constraints on our analysis. The first challenge comes from missing entries in patient records. After pre-processing and dropping out the missing values, the sample size of our data decreases from 26,870 to 2,234. Machine or patient disconnections, transmission and recording errors, or human omissions all lead to missing data(Scott, 2013). Another challenge we faced was incompleteness of information. Given the nature of clinical data, we are aware of the fact that some important events may go unobserved. We have also found that the datasets obtained do not include all variables identified in MIMIC II User Guide. Important variables like date of death and mortality were masked due to sensitivity in the clinical dataset and thus we failed to incorporate them into our model.

Other constraints come from sample size. In our ANN classification model, we found the model performance not as ideal, given its loss converges within the first few epochs. The learning plot from the neural network model shows a steep increasing trend in early epochs and remains stable afterwards. This is indicative that our training dataset sample size is not large enough.

In the future, we would like to incorporate larger clinical datasets inside our model. Future research could also focus on processing and utilizing more variables related to vital signs. We would also like to validate our model with other clinical datasets, and test if our model is applicable in different hospitals' ICU settings.

7. Bibliography

1. SOUSA, S., MARTINS, F., ALVIMFERRAZ, M., & PEREIRA, M. (2007). Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations. *Environmental Modelling &Amp; Software*, *22*(1), 97-103. https://doi.org/10.1016/j.envsoft.2005.12.002
2. Bebis, G., & Georgiopoulos, M. (1994). Feed-forward neural networks. *IEEE Potentials*, *13*(4), 27-31. https://doi.org/10.1109/45.329294
3. Moitra, V., Guerra, C., Linde-Zwirble, W., & Wunsch, H. (2016). Relationship Between ICU Length of Stay and Long-Term Mortality for Elderly ICU Survivors*. *Critical Care Medicine*, *44*(4), 655-662. https://doi.org/10.1097/ccm.0000000000001480
4. Zhang, Z., & Sabuncu, M. (2018). Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, *31*.
5. Brownlee, J. (2022). *Dropout Regularization in Deep Learning Models With Keras*. Machine Learning Mastery. Retrieved 28 April 2022, from https://machinelearningmastery.com/dropout-regularization-deep-learning-models-keras/.
6. Scott, D., Lee, J., Silva, I., Park, S., Moody, G., Celi, L. and Mark, R., 2013. *Accessing the public MIMIC-II intensive care relational database for clinical research.* BMC Medical Informatics and Decision Making, 13(1).
7. *MIMIC2 original ICU*. Kaggle.com. (2022). Retrieved 1 May 2022, from https://www.kaggle.com/datasets/drscarlat/mimic2-original-icu.
8. Gruenberg, D., Shelton, W., Rose, S., Rutter, A., Socaris, S., & McGee, G. (2006). Factors Influencing Length of Stay in the Intensive Care Unit. *American Journal Of Critical Care*, *15*(5), 502-509. https://doi.org/10.4037/ajcc2006.15.5.502
9. Toptas, M., Sengul Samanci, N., Akkoc, İ., Yucetas, E., Cebeci, E., & Sen, O. et al. (2018). Factors Affecting the Length of Stay in the Intensive Care Unit: Our Clinical Experience. *Biomed Research International*, *2018*, 1-4. doi: 10.1155/2018/9438046
10. Oliveira, A., Dias, O., Mello, M., Araújo, S., Dragosavac, D., Nucci, A., & Falcão, A. (2010). Fatores associados à maior mortalidade e tempo de internação prolongado em uma unidade de terapia intensiva de adultos. *Revista Brasileira De Terapia Intensiva*, 22(3), 250-256. doi: 10.1590/s0103-507x2010000300006
11. Moyer, J. (1994). Factors Related to Length of ICU Stay for CABG Patients. *Dimensions Of Critical Care Nursing*, *13*(4), 194-199. doi: 10.1097/00003465-199407000-00004
12. Gardner, R., Sarkar, U., Maselli, J., & Gonzales, R. (2007). Factors associated with longer ED lengths of stay. *The American Journal Of Emergency Medicine*, *25*(6), 643-650. doi: 10.1016/j.ajem.2006.11.037
13. Loh, W. (2011). Classification and regression trees. *Wires Data Mining And Knowledge Discovery*, *1*(1), 14-23. doi: 10.1002/widm.8
14. Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal Of The Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301-320. doi: 10.1111/j.1467-9868.2005.00503.x