Predicting the Risk of Sepsis using Machine Learning Techniques

Akash Gupta (CWID: 11705393)

1 Introduction

Sepsis is a *life-threatening organ dysfunction caused by a dysregulated host response to infection* [1]. Patients with sepsis are at considerable risk for severe complications and death. In-hospital mortality rates for sepsis patients range from 10% to 20% [2], and between 2007 and 2013, the number of hospital admissions due to sepsis increased nearly 49% to more than 352 per 100,000 persons per year [3]. At about $20 billion, or 5.2% of national hospital costs, sepsis is considered the most expensive condition treated in U.S. hospitals [4].

The physiological mechanism of sepsis is a complicated process. The complication involves the effect of individual variables as well as the random variation. Using the machine learning techniques, we determine hidden patterns and develop models to predict mortality. The developed models enable practitioners to take timely intervention.

Our problem is a supervised classification. We selected the relevant variables using L_1 and L_2 norm. Then we used following methods to develop predictive models to assess the risk of bad outcome. The performance of models were compared using area under receiver operating characteristic curve (AUROC).

1. Logistic regression
2. Decision tree
3. Gaussian process classifier
4. Neural network

2 Dataset

We used MIMIC-III dataset which comprises over 58,000 hospital admissions for 38,645 adults and 7,875 neonates [5]. The data spans June 2001 - October 2012. The population of sepsis was selected using ICD-9 codes of sepsis. In final data, we had 18,701 records and 41 variables. Table 1 lists the names of variable used for the analysis.

Table 1: Clinical signs extracted from dataset

Non-invasive blood pressure (diastolic)	Non-invasive blood pressure (mean)
Non-invasive blood pressure(systolic)	Temperature (Fahrenheit)
Alkaline phosphate	ALT
Anion gap	Arterial base excess
Arterial CO_2 pressure	Arterial O_2 pressure
AST	BUN
Calcium non-ionized	Chloride (serum)
Creatinine	Glucose (serum)
Glucose (finger stick)	HCO_3 (serum)
Hematocrit (serum)	Hemoglobin
INR	Lactic acid
Magnesium	pH arterial
Phosphorous	Platelet count
PTT	Sodium (serum)
TCO_2 calc arterial	Total bilirubin
WBC	Central venous pressure
GCS (eye opening)	GCS (motor)
GCS (verbal)	Heart rate
Inspired O_2 fraction	O_2 saturation pulse oxymetry
Respiratory rate	Discharge type (outcome variable)

3 Variable selection

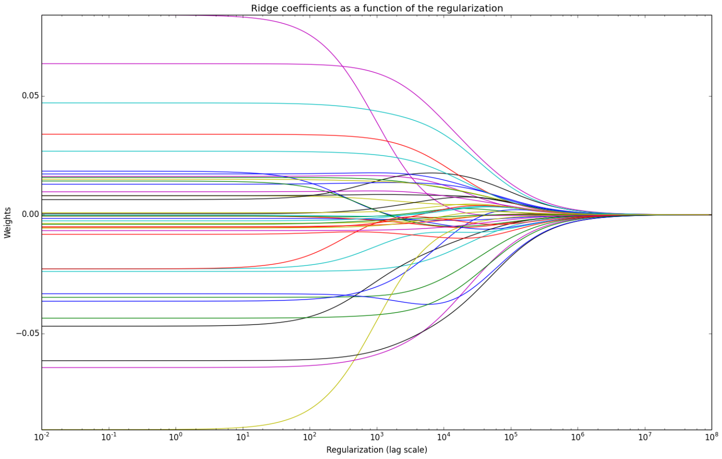

Figure 1: L_2 regularization path: plot the coefficients (or weight) as a function of the regularization parameter. As the value of regularization parameter (X-axis) increases, the model complexity reduces. The weights of each variable tends to zero as regularization parameter reach infinity

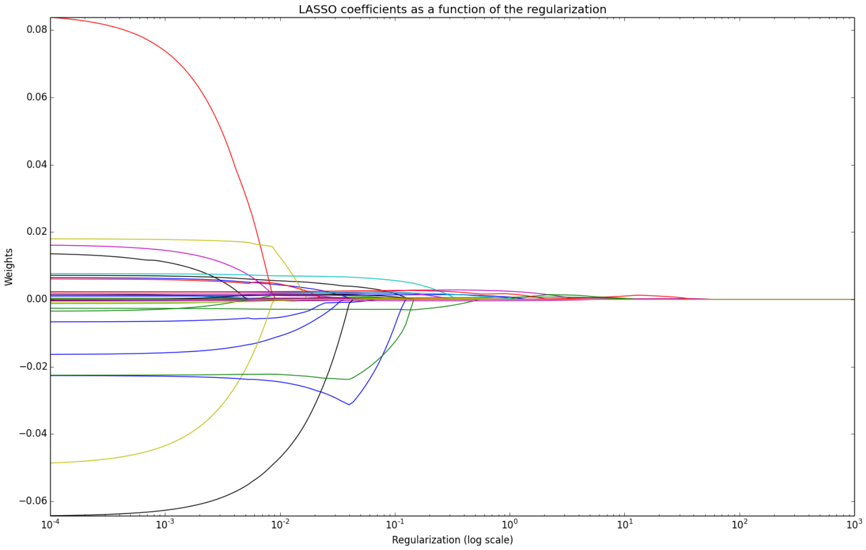

Figure 2: L_1 regularization path (without centering and scaling): plot of the weight of input variable as the function of regularization parameter. Few weights are exact zero due to inherent property of Laplacian norm. In L_1 norm weights are zero at specific value of regularization parameter while in L_2 norm, weights tends to zero as regularization parameter reach infinity. Due to different range of input variables, the algorithm fail to converge

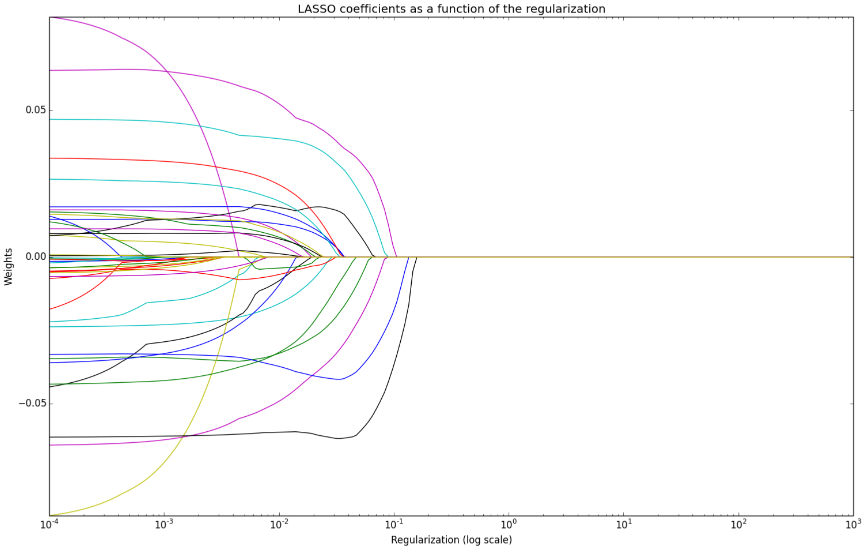

Figure 3: L_1 regularization path (with centering and scaling): The centering and scaling of each variable lead the convergence of algorithm. The plot shows change in weights with regularization parameter on data after centering and scaling. Optimal regularization parameter obtained using cross validation (CV) is 0.0073

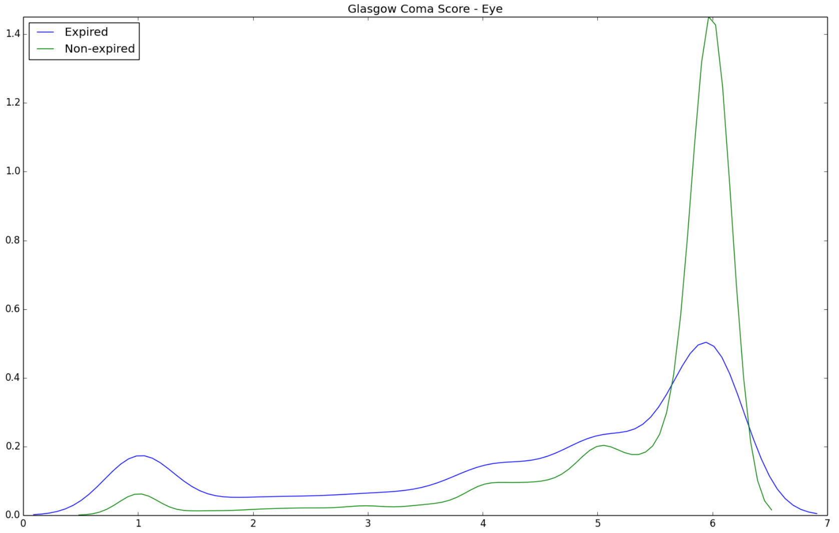

Figure 4: Likelihood function of the most important variable: The blue and green curve show the distribution of Glasgow Coma Scale - Eye given outcome as expired and non-expired, respectively. From visual inspection, we can comment that there exists a difference in distribution. Low Glasgow Coma Scale implies high chance of bad outcome and vice-versa

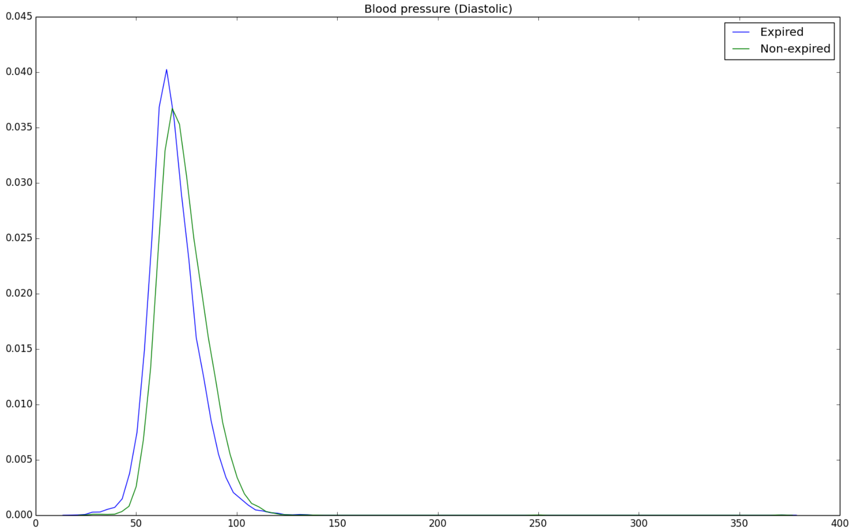

Figure 5: Likelihood function of the least important variable: The blue and green curve show the distribution of blood pressure given the outcome as expired and non-expired, respectively. From visual inspection, we can comment that it difficult to distinguish expired and non-expired using blood pressure as both centered more or less same point and have same shape

4 Models

4.1 Logistic regression

By experiments we find the best polynomial power to develop logistic regression (shown in Figure 6). Figure 7 shows the comparison of the performance of logistic regression using linear and Gaussian radial basis functions. We understand that not all variable contribute equally to predict the risk of mortality, therefore, we performed experiments to inspect the trade off between the number of variables and accuracy (shown in Figure 8). We observed that with 12 variables we can capture most of the variations that was captured using 26 variables.

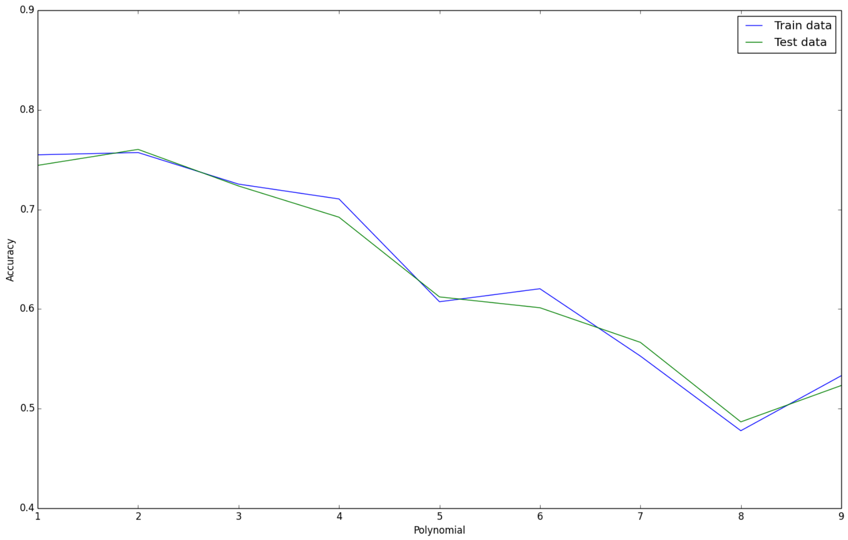

Figure 6: Selection of the best polynomial power for logistic regression: The plot shows the training and testing accuracy of logistic regression model with varying polynomial powers. The experiments show that linear combination of variables provide best results compare to higher polynomials

We attempted to reduce the dimensionality using principal component analysis. The fewer number of principal components facilitates visualization. Figure 9 shows that with two dimensions (principal components), we can retain most of the information as obtained using model with 26 dimensions.

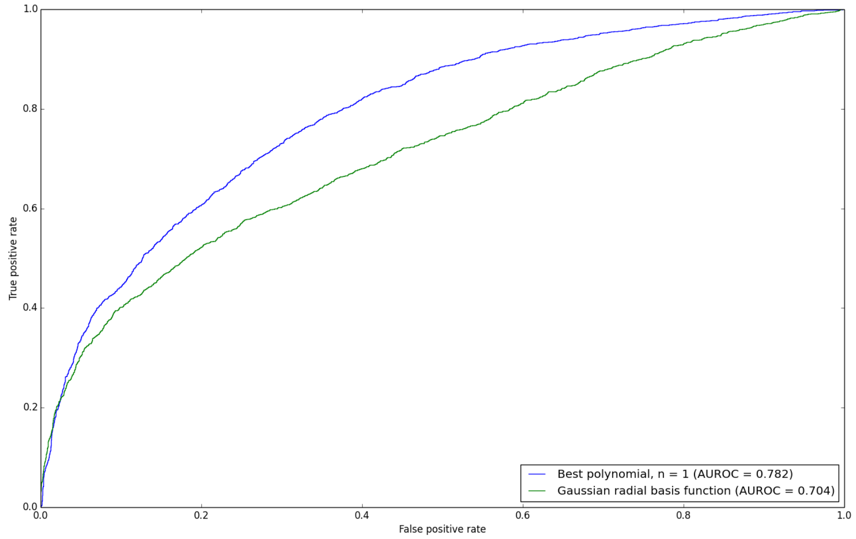

Figure 7: Receiver operating characteristic of logistic regression using polynomial and radial basis function: The plot shows that with AUCROC of 0.782, logistic regression with linear combination of variables performs better than logistic regression with gaussian radial basis function (AUROC = 0.704)

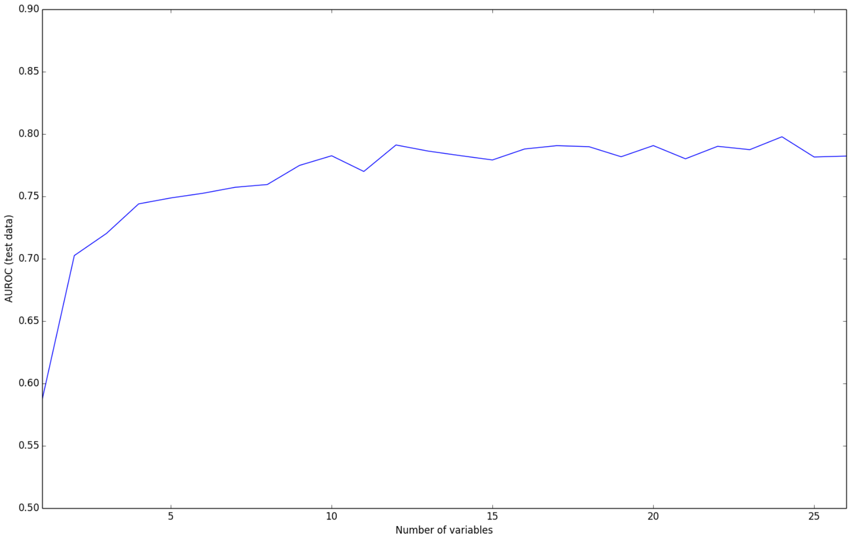

Figure 8: Recursive feature elimination and logistic regression: Using recursive feature elimination, the best set of i features ($i \in \{1, 2, \ldots, 26\}$) were found. The above plot shows that AUROC of the logistic regression with 12 numbers (AUROC = 0.781) is more or less same as AUROC with 26 variables (AUROC = 0.782 (with 10 fold cross validation)). Therefore, using only 12 clinical variables rather than 26 derived from laplasian prior (LASSAO regression), we can explain the most of the underline model.

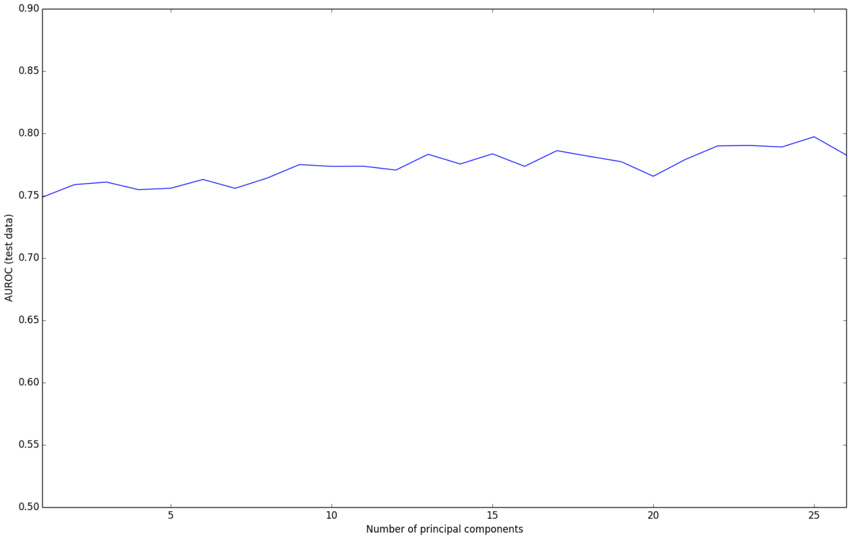

Figure 9: Principal components and logistic regression: We derived principal components from dataset with selected variables ($n = 26$). From above plot we can observe that with one principal component, we can explain most of the underlying model.

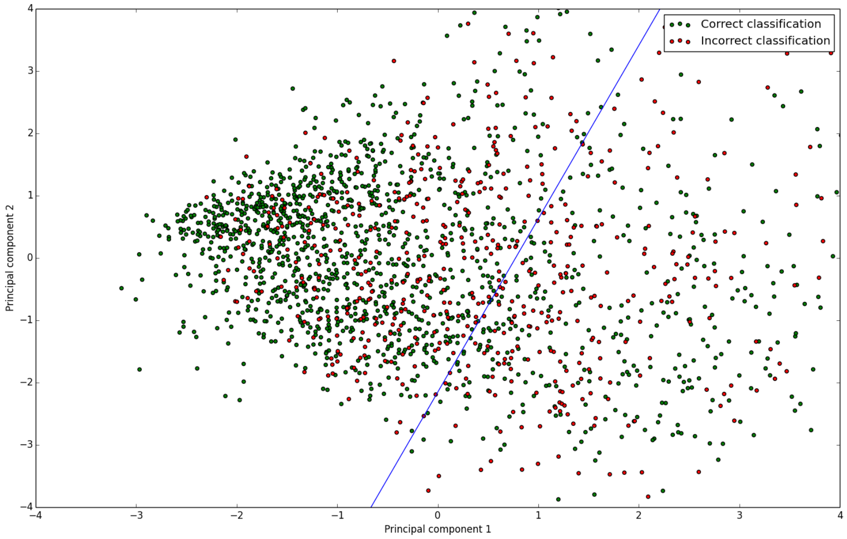

Figure 10: Linear discriminant with combination of principal component analysis and logistic regression: Two principal components were derived from original data (26 variables) for the visualization purpose. We used a sample of 2000 data points here. The blue line is a logit discriminant. As move further from the discriminant, the density of correctly classified points increases (green dots). There exists many incorrect classified points near discriminant.

4.2 Decision tree

A decision tree can easily over-fit the training data set by growing the decision tree to a large number of leaves. Therefore, we performed experiments to find the optimal number of leaves (shown in Figure 11). The decision were constructed for both Ginni and Entropy splitting criteria shown in Figure 11 and 12, respectively. Figure 13 shows the receiver operating characteristic curve for decision tree.

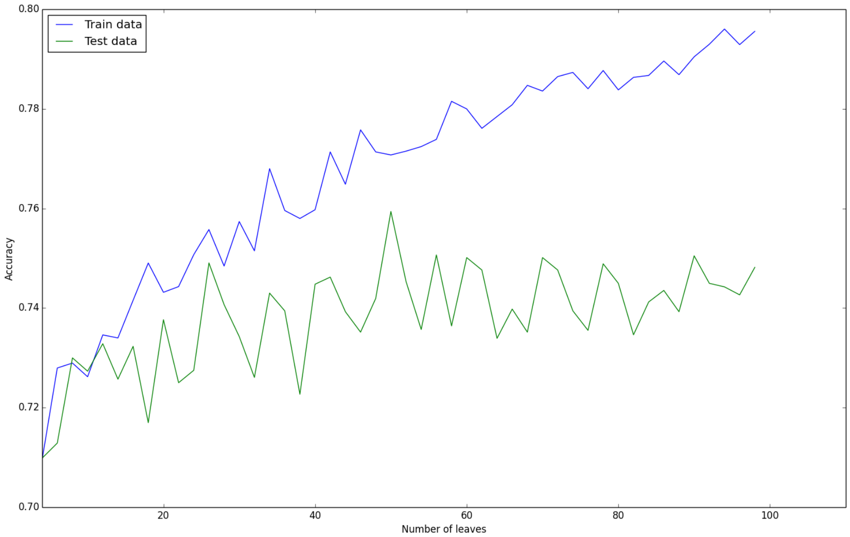

Figure 11: Selecting optimal number of leaves (splitting criterion = Ginni): The plot shows that training accuracy increases monotonically with number of leaves. While the test data accuracy becomes more or less constant as number of leaves increases

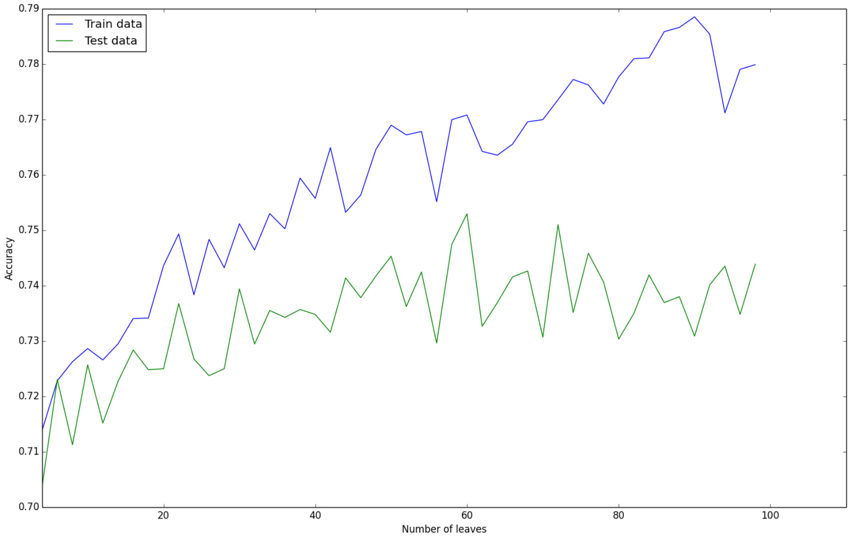

Figure 12: Selecting optimal number of leaves (splitting criterion = Entropy)

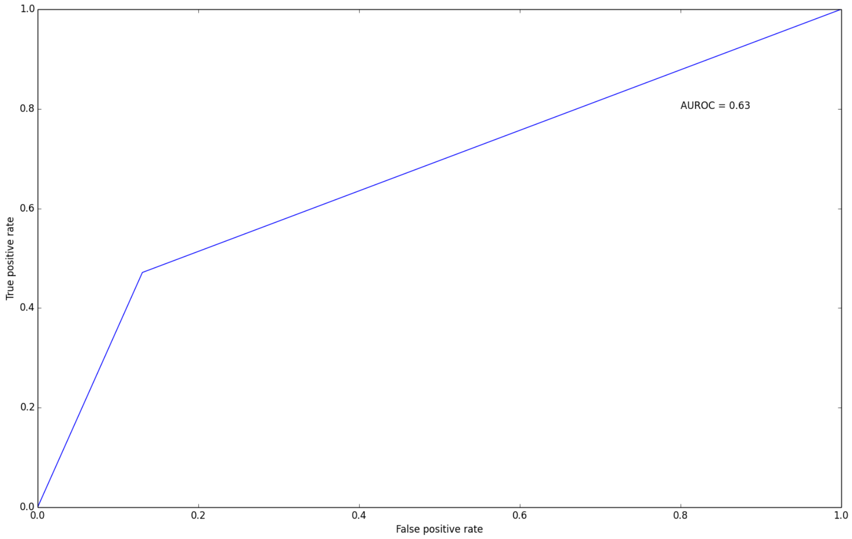

Figure 13: Receiver operating characteristic curve for decision tree. The AUROC is 0.63

4.3 Gaussian process classifier

We used Gaussian process classifier with radial basis functions kernel. The hyper-parameter was obtained using experiments over varying hyper-parameters shown in Figure 14. Using the best hyper-parameter of 3.8, the AUROC was 0.85 for Gaussian classifier. Figure 15 shows the receiver operating characteristic curve.

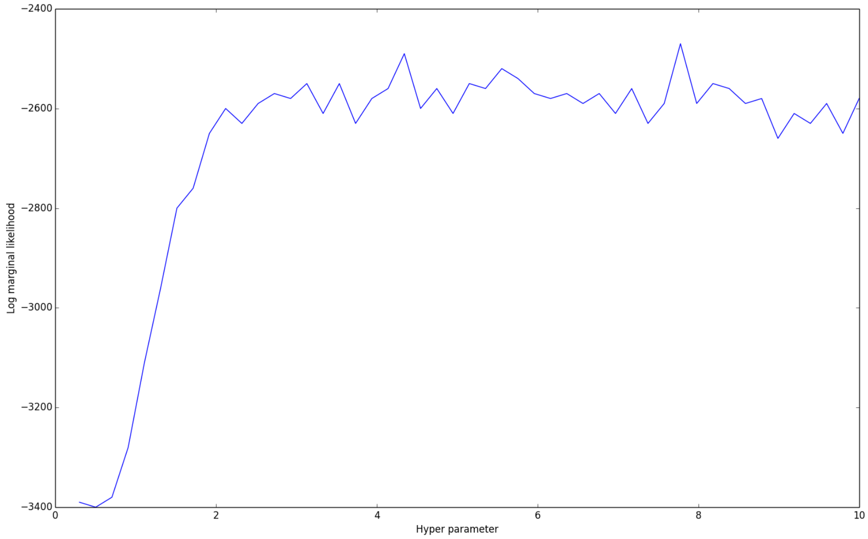

Figure 14: Selecting optimal value of hyper parameter: the plot shows the performance of Gaussian process classifier with radial basis kernel over varying hyper-parameter. The hyper parameter beyond 3.8 does not improve the log marginal likelihood.

The experiments were also performed using absolute exponential kernel. Interestingly, the performance of Gaussian process classifier with absolute exponential kernel is significantly than with radial basis function.

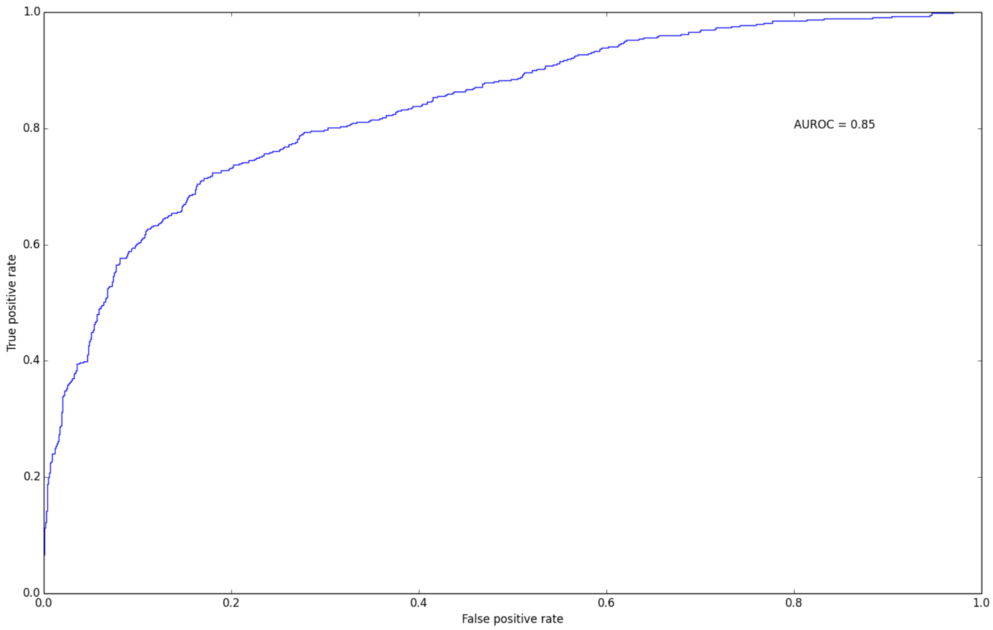

Figure 15: Receiver operating characteristic curve for Gaussian process classifier with radial basis kernel. The AUROC is 0.85

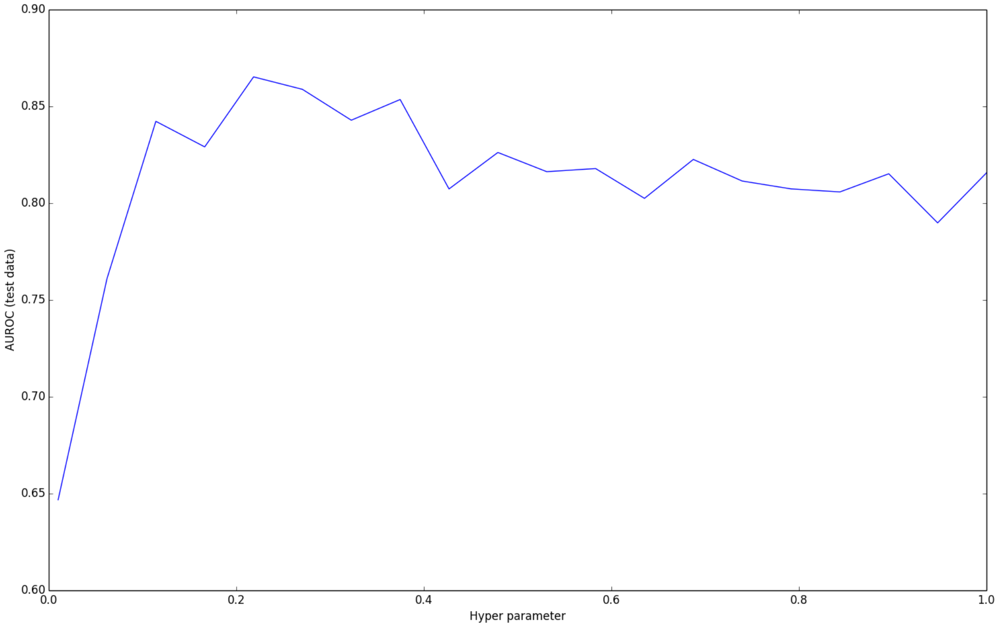

Figure 16: Gaussian process classifier with absolute exponential kernel: The plot shows the change in AUROC with hyper parameter of absolute exponential kernel. The optimal hyper parameter is 0.218 and corresponding test data AUROC is 0.865

4.4 Deep learning

A multilayer perceptron network, with two hidden layers (100 and 50 neurons) was constructed. Figure 17 the architecture of neural network. We expected better performance with deep learning. However, the AUROC of deep learning was very low 0.58. We plan to refine results of deep learning as a future work.

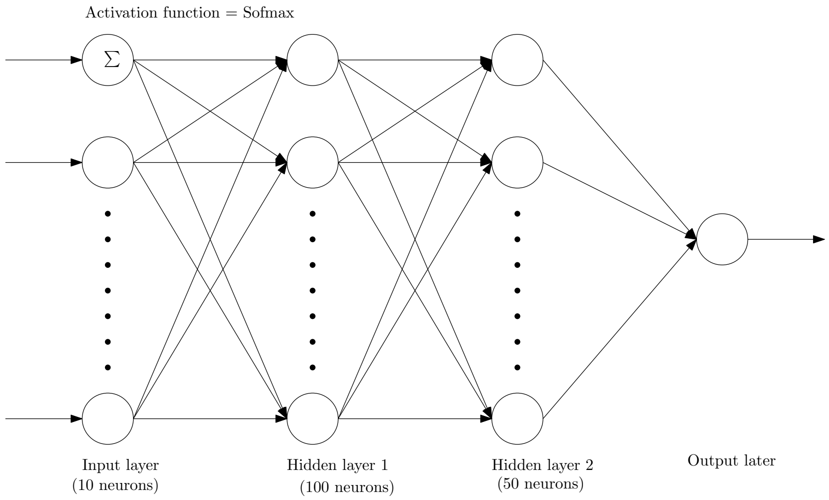

Figure 17: Architecture of applied network

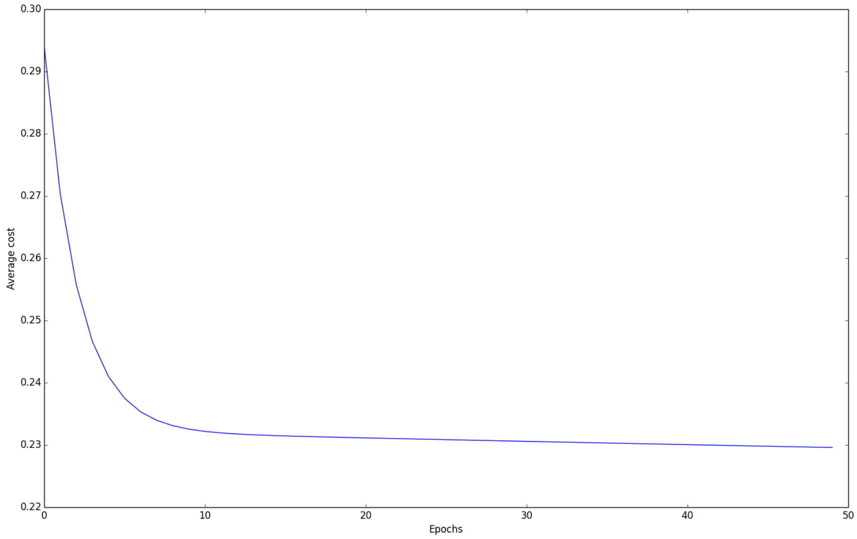

Figure 18: Training multi-layer perceptron: The plot shows decrease in mean square error with epoch. As expected, there is no change in objective cost beyond a certain number of epochs

5 Conclusion

In this project, we compared the performance of four different modeling techniques (Logistic regression, decision tree, gaussian process classifier, neural network) to predict the risk of mortality. From Table 2, we conclude that Gaussian process classifier with absolute exponential kernel fits best on the given data.

Table 2: Comparison of AUROC of developed models

Model	AUROC
Logistic regression (Polynomial function)	0.78
Logistic regression (Gaussian radian basis function)	0.70
Decision tree	0.63
Gaussian process classifier (radial basis kernel)	0.85
Gaussian process classifier (absolute exponential kernel)	0.87

References

[1] Mervyn Singer, Clifford S Deutschman, Christopher Warren Seymour, Manu Shankar-Hari, Djillali Annane, Michael Bauer, Rinaldo Bellomo, Gordon R Bernard, Jean-Daniel Chiche, Craig M Coopersmith, et al. The third international consensus definitions for sepsis and septic shock (sepsis-3). *Jama*, 315(8):801–810, 2016.

[2] Yonathan Freund, Najla Lemachatti, Evguenia Krastinova, Marie Van Laer, Yann-Erick Claessens, Aurélie Avondo, Céline Occelli, Anne-Laure Feral-Pierssens, Jennifer Truchot, Mar Ortega, et al. Prognostic accuracy of sepsis-3 criteria for in-hospital mortality among patients with suspected infection presenting to the emergency department. *Jama*, 317(3):301–308, 2017.

[3] Carolin Fleischmann, André Scherag, Neill KJ Adhikari, Christiane S Hartog, Thomas Tsaganos, Peter Schlattmann, Derek C Angus, and Konrad Reinhart. Assessment of global incidence and mortality of hospital-treated sepsis. current estimates and limitations. *American journal of respiratory and critical care medicine*, 193(3):259–272, 2016.

[4] Celeste M Torio and Roxanne M Andrews. National inpatient hospital costs: the most expensive conditions by payer, 2011: statistical brief# 160, 2006.

[5] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3, 2016.