International Journal of Computer Science and Applications,
©Technomathematics Research Foundation
Vol. 11, No. 3, pp. 45 – 56, 2014

MACHINE LEARNING TECHNIQUES FOR PREDICTING HOSPITAL LENGTH OF STAY IN PENNSYLVANIA FEDERAL AND SPECIALTY HOSPITALS

PARAG C. PENDHARKAR

Information Systems, Penn State University at Harrisburg, 777 W. Harrisburg Pike, Middletown, PA 17057, United States of America
pxp19@psu.edu
http://www.personal.psu.edu/pxp19/

HITESH KHURANA

Psychiatry, Pt. B.D. Sharma PGIMS, 5/9J, Medical Enclave, Rohtak, Haryana 124001, India
doctorhitesh@rediffmail.com

In this paper, we compare three different machine learning techniques for predicting length of stay (LOS) in Pennsylvania Federal and Specialty hospitals. Using the real-world data on 88 hospitals, we compare the performances of three different machine learning techniques—Classification and Regression Tree (CART), Chi-Square Automatic Interaction Detection (CHAID) and Support Vector Regression (SVR)—and find that there is no significant difference in performances of these three techniques. However, CART provides a decision tree that is easy to understand and interpret. The results from CART indicate that psychiatric care hospitals typically have higher LOS than non-psychiatric care hospitals. For non-psychiatric care hospitals, the LOS depends on hospital capacity (beds staffed) with larger hospitals with beds staffed over 329 having average LOS of 13 weeks vs. smaller hospitals with average LOS of about 3 weeks.

Keywords: Keyword1; keyword2; keyword3.

1. Introduction

Resource planning plays an important role in providing quality healthcare at lower costs. For healthcare organizations, the key issue in effective management of resources is to manage tradeoffs between service quality and costs. There are several mechanisms used to manage these tradeoffs, and among these mechanisms are: improved work flows (Wang et al. (2013)], use of better scheduling systems for managing staff and admissions scheduling [Pérez et al. (2013); Maenhout and Vanhoucke (2013); Hulshof et al. (2013)], and capacity planning [Romero et al. (2013); Zhang and Puterman (2013)]. For Federal hospitals, in addition to service quality and costs, a related variable in the form of resource utilization plays a key role in healthcare administrators' decision-making. Federal hospitals may have agreed with the government to serve a certain number of patients and veterans [Hulshof et al. (2013)], and must serve these patients well while keeping the costs low and maximizing resource utilization. Lower resource utilization

means wasting tax payer dollars, and higher utilization sometimes results in delays and poor quality of health care due to lack of sufficient resources. Some government reimbursement programs have "pay for performance" incentives to encourage better resource utilization and increased efficiency [Azari et al. (2012)].

For impatient care units, two variables play an important role in determining hospital resource utilization. The first variable is predicting a patient's hospital length of stay (LOS), and second variable is predicting readmissions [Kelly et al. (2013)]. Ideally, a hospital must minimize both variables to provide high-quality healthcare and improve resource utilization. Predicting hospital LOS allows a hospital to predict discharge dates for a patient admitted to the hospital, which in turn allows improved scheduling of elective admissions leading to reduce variance in hospital bed occupancies [Robinson et al. (1966)]. Predicting LOS also allows a hospital to scale its capacity during its long-term strategic planning.

Hospital LOS plays an important role in predicting health care costs as well [Cosgrove (2006)]. As a result, for nearly fifty years, several studies have focused on predicting hospital LOS. For example, Robinson et al. [Robinson et al. (1966)] attempted to predict LOS for a hospital in California; Kelly et al. [Kelly et al. (2012)] investigated factors that help in predicting LOS for colorectal resection and radical prostatectomy patients [Kelly et al. (2013); Weisgerber et al. (2011)] attempted to predict LOS for infants with bronchiolitis. Studies on predicting LOS either attempt to predict LOS assuming "all diseases" hospital or specialty hospital treating only certain diseases. For a general model predicting LOS for any hospital, an improved approach would be to use the disease type variable as input to predict LOS because a certain disease type will have either higher or lower average LOS compared to other disease types.

In our study, we use three different machine learning approaches to predict LOS in Pennsylvania Federal and specialty hospitals. Specifically, we use classification and regression tree (CART), Chi-squared automatic interaction detection (CHAID) and support vector regression (SVR) techniques to learn an LOS prediction model. Additionally, we also average the predictions of all three techniques to create an ensemble prediction. Further, we validate our models using V-fold validation approach. To our knowledge, machine learning regression techniques are not used in predicting LOS in hospitals; and we propose a model that does not use data from a single hospital, but all the Federal and specialty hospitals in the state of Pennsylvania. Thus, the results of our study are highly generalizable for the state of Pennsylvania.

The rest of the paper is organized as follows: In section 2, we briefly describe the three machine learning regression techniques used in our study. In section 3, we describe our data and report the results of our experiments. In section 4, we conclude our paper with a summary and provide directions for future research.

2. Machine Learning Regression Approaches

We assume a regression function $y=f(\mathbf{x})$, where y is the dependent variable and $\mathbf{x} \in \Re^k$ is a vector of k independent variables. More specifically, we assume that training dataset

for all our regression approaches can be represented as $\{(\boldsymbol{x}_1,y_1),\ldots,(\boldsymbol{x}_p,y_p)\} \subset \Omega \times \Re$, where Ω denotes space of input patters (i.e., $\Omega = \Re^k$). We use two regression tree algorithms and one support vector regression model to learn regression function that maps independent variables onto the dependent variable. The two regression tree algorithms that we use are CART [Breiman et al. (1984)] and CHAID [Kass (1980)]. CART algorithm learns a non-parametric, non-linear regression tree for a given set of training data. CHAID is a statistical algorithm that can be used to derive regression trees. Regression tree models approximate the forecasting function in a staircase function. Figure 1 illustrates a regression tree (at the top), the actual forecasting function (as a bold curve) and the staircase forecasting function learnt by the regression tree (histograms using dotted lines).

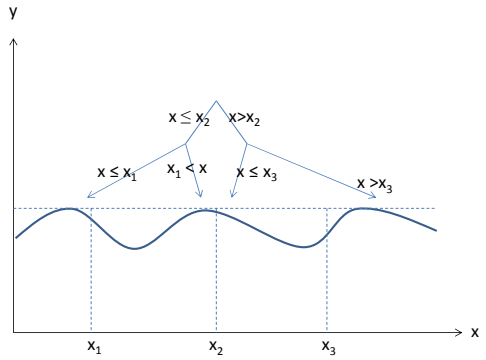

Fig. 1: The function $y=f(x)$ approximation using a regression tree

CART constructs a binary decision tree by splitting the data set in a way that the data in the descendant sub-sets are more *pure* than the data in the parent set. For example, in a regression problem, let $(\boldsymbol{x}_n,y_n)$ represent nth example, where $\boldsymbol{x}_n$ is the nth example vector on independent variables and y_n is the value of the dependent variable. If there are total of N examples then CART calculates a best split s^* so that following is maximized over all possible splits S:

$$\Delta R(s^*,t) = \text{argmax}_{s\in S}\,\Delta R(s,t). \qquad (1)$$

Where $\Delta R(s,t) = R(t) - R(t_L) - R(t_R)$ is improvement in resubstitution estimate for split s of t. The resubstitution estimate $R(t)$ is defined as follows:

$$R(t) = \frac{1}{N}\sum_{x_*\in t}\left(y_n - y(t)\right)^2. \qquad (2)$$

The variables t_L and t_R are left and right values for split t. The variable $y(t)$ is defined as follows:

$$y(t) = \frac{1}{N(t)} \sum_{x_* \in t} y_n. \quad (3)$$

Where *N(t)* is the total number of cases in *t*. The tree continues to grow until a node is reached such that no significant decrease in resubstitution estimate is possible. This node is the terminal node.

CHAID is a heuristic statistical method that examines the relationships between many categorical independent variables and a single categorical dependent variable. For continuous values, CHAID divides the continuous data into equal fixed (usually 10) categories, and then merges the categories that are judged to be statistically insignificant. When dependent variable is categorical, Chi-squared test is used, and when dependent variable is continuous, F test is used for determining statistical significance. More information on CHAID can be found in Kass [Kass (1980)].

Support vector regression (SVR) technique assumes a regression function taking following form: $y = \boldsymbol{w}^T \boldsymbol{x} + b$, where $\boldsymbol{w} \in \Omega$ and $b \in \Re$. Vapnik's [Vapnik (1995)] SVR formulation can be stated as follows:

$$Minimize\ \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{p}(\xi_i + \xi_i^*), \quad (4)$$

Subject to:

$$y_i - \boldsymbol{w}^T \boldsymbol{x}_i - b \leq \varepsilon + \xi_i, \quad (5)$$

$$\boldsymbol{w}^T \boldsymbol{x}_i + b - y_i \leq \varepsilon + \xi_i^*, \text{ and} \quad (6)$$

$$\xi_i, \xi_i^* \geq 0. \quad (7)$$

The loss function for formulation (4)-(7) is ε-sensitive insensitive loss function, which is defined as $L(y_i, f(\mathbf{x}_i)) = (|f(\mathbf{x}_i) - y_i| - \varepsilon)_+$, $\varepsilon \geq 0$ [Moguerza and Munoz (2006)]. The constant $C > 0$ determines the tradeoff between flatness of function *f* and extent to which deviations larger than ε are tolerated [Smola and Scholkopf (2004)]. For non-linear SVM, the regression function takes a form $y = \sum_{i=1}^{p} \alpha_i \times K(\boldsymbol{x}_i, \boldsymbol{x}) + b$, where $K(\mathbf{x}_i, \mathbf{x})$ is a positive definite kernel function. The formulation (4)-(7) remains unchanged except the weight vector $\boldsymbol{w} = \sum_{i=1}^{p} \alpha_i \phi(\boldsymbol{x}_i)$ and ϕ is the mapping that defines the kernel function. In our research, we use the Gaussian kernel function $K(\boldsymbol{x}_i, \boldsymbol{x}) = e^{-\frac{\|\mathbf{x}_i - \mathbf{x}\|^2}{c}}$ and $\varepsilon = 0.1$.

3. Data, Experiments and Results

Data on various Federal and Specialty hospitals were obtained from the Pennsylvania Department of Health. The data set consisted of year 2013 information on 88 hospitals throughout Pennsylvania. Information about the following was collected from each hospital:

1. Hospital name
2. Type of service
3. Beds staffed

4. Admissions

5. Discharges

6. Average length of stay

A preliminary analysis was undertaken on the collected data. First, we provide distribution of type of service for different hospitals in Figure 2. Table 1 provides the description of type of service variables. Table 2 provides descriptive statistics for other variables. The average LOS was approximately 77 days with a minimum of 1 day and a maximum of approximately 1,589 days.

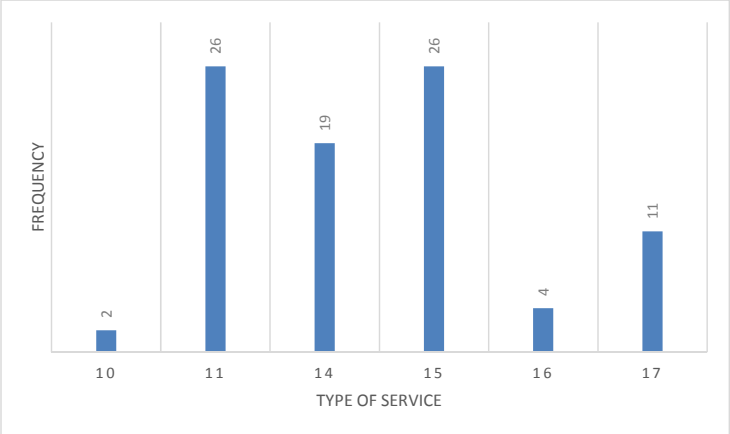

Fig. 2: Type of Service Distribution

Table 1: Type of Service Description

Type of Service Number	Description
10	General care
11	Psychiatric care
14	Rehabilitation
15	Long term acute care
16	Alcoholism/Drug Abuse
17	Other

Table 2: Descriptive Statistics

Variable	Mean	Std. Dev.	Minimum	Maximum
Beds Staffed	96.19	88.37	14	493
Admissions	2162.99	4285.60	61	28996
Discharges	2550.04	5694.19	68	37614
Average LOS	76.64	258.83	1	1588.35

Given the variables, the set of variables that predict the average LOS are the admissions (A), type of service (S), discharges (D) and beds staffed (B). In other words, the LOS may be represented in the following functional form: $LOS = f(A, S, D, B)$, where $f(.)$ is a mapping that maps the independent variables to the dependent variable LOS. The objective of this study is to learn and test this mapping using machine learning techniques. However, we first test the functional relationship between the independent and dependent variables using analysis of covariance (ANCOVA). Table 3 provides the summary of our ANCOVA analysis.

Table 3: The ANCOVA Summary Table

Source	Type III Sum of Sq.	df	Mean Sq.	F-Value	Sig.
Model	2963116.43	8	370389.55	10.21	0.000*
Intercept	196473.86	1	196473.86	5.42	0.023**
A	1070413.45	1	1070413.45	29.51	0.000*
D	30077.16	1	30077.16	0.83	0.365
B	2130880.56	1	2130880.56	58.75	0.000*
S	181749.80	5	36349.96	1.00	0.422
Error	2865335.54	79	36270.07		
Total	6345338.40	88			
Corrected Total	5828451.98	87			

*Significant at 99%, **significant at 95%; R-Squared=50.8%; Adjusted R-Squared 45.9%

The results indicate that two factors *B* and *A* play a major role in predicting LOS. The ANCOVA model tests linear relationships and the low *R*-squared value, and a significant intercept indicate that non-linear machine learning models may perform slightly better than linear models. We run CHAID and CART algorithm on entire dataset of 88 hospitals and generate CHAID and CART trees. Figures 3 and 4 illustrate the results of our experiments. The CART prediction tree indicates that approximately 90% of the hospitals have admissions greater than 126.5 and number of beds staffed less than or equal to 329. For these hospitals average LOS is approximately 3 weeks. Neither CHAID nor CART retained variables discharges and service type for their analysis. When combined with the results of ANCOVA, it appears that admissions and hospital capacity (i.e., number of beds staffed) plays a major role in determining average LOS. For an administrative standpoint, it is an important finding because it appears that for Pennsylvania Federal and specialty hospitals, comparisons can be made across different service offerings.

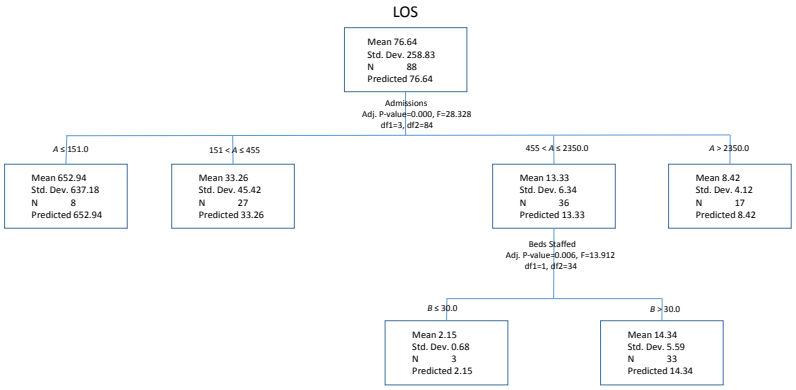

Fig. 3: The CHAID LOS Prediction Tree

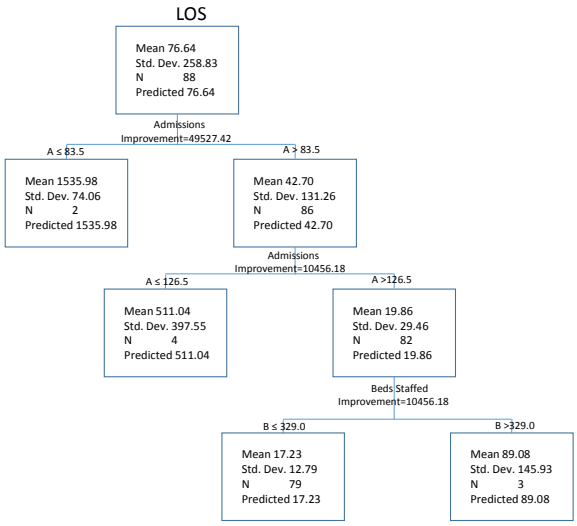

Fig. 4: The CART LOS Prediction Tree

We now compare predictive performances of three different machine learning algorithms using V-fold validation. In V-fold validation, original dataset is separated into five mutually exclusive and collectively exhaustive datasets. Next, each dataset is used as a holdout sample and remaining four are combined and used as training dataset to create five holdout sample tests. Since our original dataset contained 88 examples, our V-fold evaluation datasets contained four datasets with 17 examples and one dataset with 20 examples. Figures 5-9 illustrate the results of our experiments on five holdout samples. The ensemble prediction was average prediction of all three machine learning techniques.

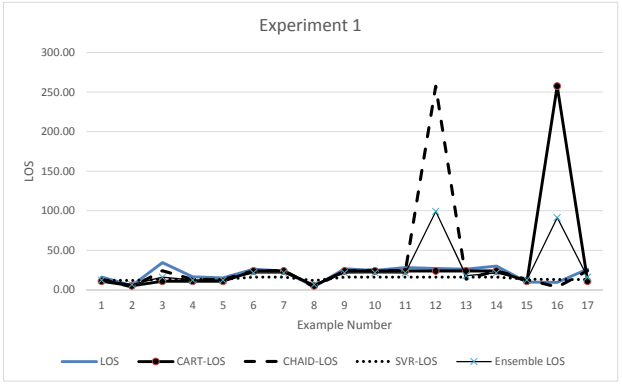

Fig. 5: Results of first holdout sample

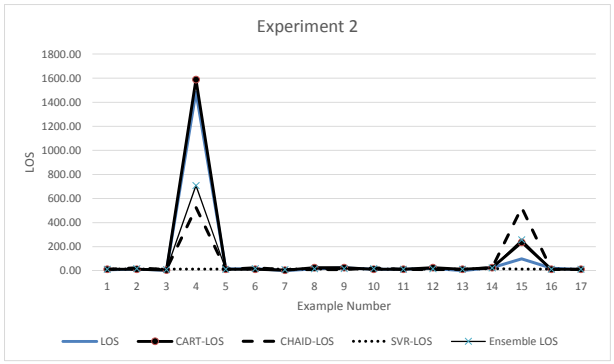

Fig. 6: Results of second holdout sample

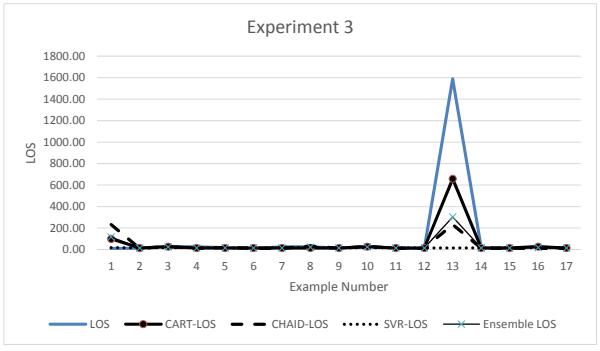

Fig. 7: Results of third holdout sample

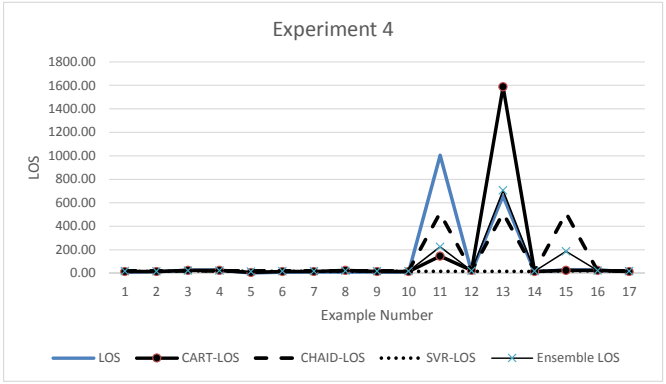

Fig. 8: Results of fourth holdout sample

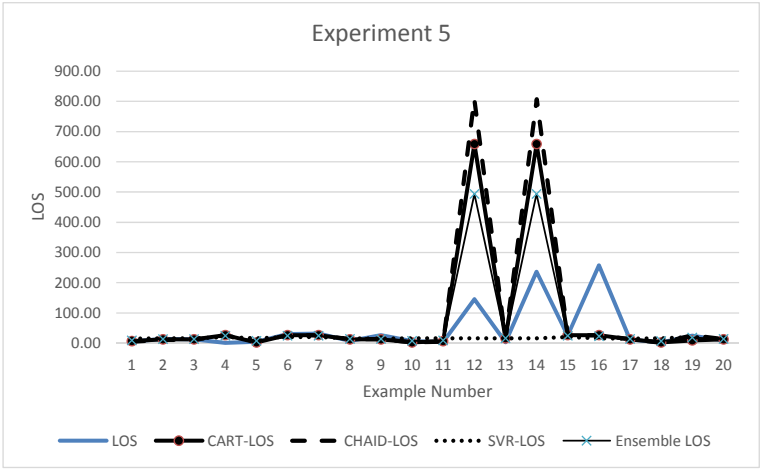

Fig. 9: Results of fifth holdout sample

Table 4 illustrates the root-mean-square (RMS) error between actual LOS and predicted LOS for different techniques. For different experiments different techniques perform better, but CART has lower mean RMS error. There was no significant difference in means for the RMS errors for different techniques indicating that no one technique had the best performance.

Table 4: The RMS error of prediction for different techniques

Experiment Number	CART	CHAID	SVR	Ensemble
1	60.60	56.13	9.86	27.30
2	42.16	254.66	357.02	191.90
3	226.66	333.17	381.66	313.07
4	307.03	171.28	285.60	193.26
5	157.52	202.25	79.16	110.44
Mean	158.80	203.50	222.66	167.19
Standard Dev.	111.61	102.70	168.21	106.49

In a situation where all techniques are equal, we pick a technique that is easy to understand and implement and, based on Figure 4, CART tree appears to be simple to understand and implement. The CART tree in Figure 4 indicates that admissions less than 83.5 and beds staffed greater than 329 are outliers as only five hospitals satisfy these categories. Vast majority of 83 hospitals had either admissions greater than 83.5 or staffed beds less than 329. For these hospitals, admissions play a key role in determining LOS. If admissions are greater than 126.5, then LOS drops significantly to about 3 weeks. However if admissions are between 83.5 and 126.5, then length of stay increases to over one year on average. There were only four hospitals falling in this range, and the service type for all these four hospitals was psychiatric care. We performed further data analysis on mean LOS for psychiatric hospitals and non-psychiatric hospitals and found that mean LOS for psychiatric hospitals was unusually high at 745 days and mean LOS for non-psychiatric hospitals was approximately 19 days. Given high LOS for psychiatric care hospitals, we also noticed that these hospitals have lower admissions' rates. Figure 10 illustrates the CART tree from Figure 4 with psychiatric hospital leaves shaded with red color. When we ignore psychiatric hospitals, CART tree becomes very simple to understand where LOS for non-psychiatric hospitals comes down to one rule: *IF Admissions >126.5 AND Beds staffed ≤ 329 THEN LOS= 17.23 ELSE LOS=89.08*. We believe that this rule is really easy to understand and implement in non-psychiatric care hospitals.

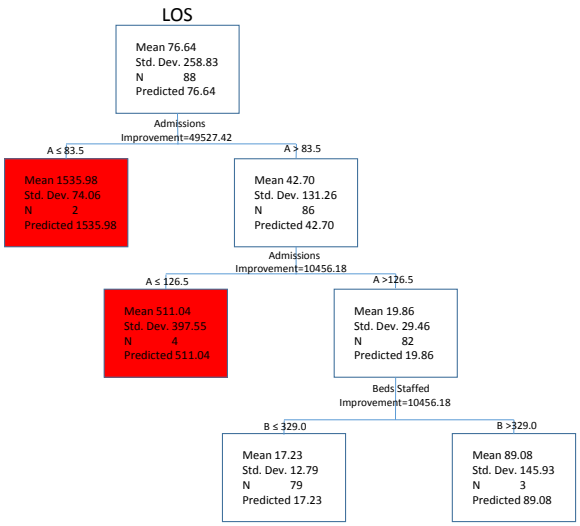

Fig. 10: CART Tree with psychiatric hospital colored leaves

4. Summary, Conclusions and Directions for Future Work

In this paper, we investigated the factors that impact LOS for Federal and Specialty care hospitals in Pennsylvania. Using the real-world dataset from 88 hospitals, we tested performance of three different machine learning techniques. The results indicate that LOS primarily depends on two factors: admissions and beds staffed. While service type generally does not play a role, we found that psychiatric care hospitals have higher LOS. Since there were only a few such hospitals, service type factor was found statistically insignificant. However, non-parametric greedy search machine learning technique, CART, detect the higher LOS for psychiatric hospitals. Given higher LOS for psychiatric care hospitals, admissions play a major role for these hospitals. As admissions increase in these hospitals, more resources should typically be allocated to psychiatric care hospitals due to high LOS. For non-psychiatric care hospitals, LOS depends primarily on beds staffed. For large-size hospitals with beds staffed greater than 329, the expected LOS is about 13 weeks. Whereas, for small-size hospitals with beds staffed equal to 329 or less, the expected LOS is about 3 weeks.

Machine learning and ensemble techniques used in this research had a similar statistical performance. This allows a decision-maker to pick a technique that is easy to interpret and implement. In our case, we found that CART tree was simplest and translates into one simple decision-making rule. The rule appears logical as well where it encourages higher resource allocations for large-capacity hospitals and lower resource allocation for small-capacity hospitals.

One of the limitations of our study was limited sample size and variables available for analysis. At the state level, sample size cannot be increased because we used data from all the hospitals from the state. However, the availability of capacity, financial and service quality variables could be increased to improve understanding of other factors

affecting LOS. Additionally, sample size could be improved by including data from other states. Doing this will improve the generalizability of the study and its results. Future research needs to address this issue.

References

Azari, A.; Janeja, V. P.; Mohseni, A. (2012): Predicting Hospital Length of Stay (PHLOS): A Multi-tiered Data Mining Approach. 2012 IEEE 12th International Conference on Data Mining Workshops (ICDMW), pp. 17–24.

Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. (1984): Classification and Regression Trees, Wadsworth International Group, Belmont, CA.

Cosgrove, S. E. (2006): The Relationship between Antimicrobial Resistance and Patient Outcomes: Mortality, Length of Hospital Stay, and Health Care Costs. Clinical Infectious Diseases, **42**(Supplement 2), pp. S82–S89.

Hulshof, P. J. H.; Boucherie, R. J.; Hans, E. W.; Hurink, J. L. (2013): Tactical resource allocation and elective patient admission planning in care processes, Health Care Management Science, **16**(2), pp. 152–166.

Kass, G. (1980): An exploratory technique for investigating large quantities of categorical data. Applied Statistics, **29**(2), pp. 119–127.

Kelly, M.; Sharp, L.; Dwane, F.; Kelleher, T.; Comber, H. (2012): Factors predicting hospital length-of-stay and readmission after colorectal resection: a population-based study of elective and emergency admissions, BMC Health Services Research, **12**(1), pp. 77-77.

Kelly, M.; Sharp, L.; Dwane, F.; Kelleher, T.; Drummond, F. J.; Comber, H. (2013): Factors predicting hospital length-of-stay after radical prostatectomy: a population-based study, BMC Health Services Research, **13**(1), 244-244.

Maenhout, B.; Vanhoucke, M. (2013). Analyzing the nursing organizational structure and process from a scheduling perspective, Health Care Management Science, **16**(3), pp. 177–196.

Moguerza, J. M.; & Munoz, A. (2006): Support vector machines with applications, Statistical Science, **21**(3), pp. 322–336.

Pérez, E.; Ntaimo, L.; Malavé, C. O.; Bailey, C.; McCormack, P. (2013): Stochastic online appointment scheduling of multi-step sequential procedures in nuclear medicine, Health Care Management Science, **16**(4), pp. 281–299.

Robinson, G. H.; Davis, L. E.; & Leifer, R. P. (1966): Prediction of Hospital Length of Stay, Health Services Research, **1**(3), pp. 287–300.

Romero, H. L.; Dellaert, N. P.; Geer, S.; Frunt, M.; Jansen-Vullers, M. H.; Krekels, G. A. M. (2013): Admission and capacity planning for the implementation of one-stop-shop in skin cancer treatment using simulation-based optimization. Health Care Management Science, **16**(1), pp. 75–86.

Smola, A. J.; Scholkopf, B. (2004): A tutorial on support vector regression, Statistics and Computing, **14**, pp. 199–222.

Vapnik, V. (1995): The Nature of Statistical Learning Theory, Springer, New York, NY.

Wang, J.; Li, J.; Howard, P. K. (2013): A system model of work flow in the patient room of hospital emergency department, Health Care Management Science, **16**(4), pp. 341–351.

Weisgerber, M. C.; Lye, P. S.; Li, S.-H.; Bakalarski, D.; Gedeit, R.; Simpson, P.; Gorelick, M. H. (2011): Factors predicting prolonged hospital stay for infants with bronchiolitis, Journal of Hospital Medicine: An Official Publication of the Society of Hospital Medicine, **6**(5), pp. 264–270.

Zhang, Y.; Puterman, M. L. (2013): Developing an adaptive policy for long-term care capacity planning, Health Care Management Science, **16**(3), pp. 271–279.