SEYBOLD Report
ISSN: 1533 - 9211
DOI: 10.5281/zenodo.6876034

PREDICTION OF SURGICAL PATIENTS LOS USING CART ANALYSIS AND COMPARISON WITH KNN AND RANDOM FOREST TECHNIQUES

P. Suneel Kumar
Professor, Department of Electronics and Communication Engineering, Sridevi Women's Engineering College, Hyderabad, India, psunilkumar.ece@gmail.com

Morla Rachana
U.G. Student, Department of Electronics and Communication Engineering, Sridevi Women's Engineering College, Hyderabad, India, rachanamorla@gmail.com

Bhagavatula Sai Malavika
U.G. Student, Department of Electronics and Communication Engineering, Sridevi Women's Engineering College, Hyderabad, India, malubhagavatula642@gmail.com

Nittala Sai Shridula
U.G. Student, Department of Electronics and Communication Engineering, Sridevi Women's Engineering College, Hyderabad, India, sai.shridula@gmail.com

ABSTRACT
The In-hospital length of stay (LOS) is expected to increase as disease complexity increases and the population ages. This will affect healthcare systems, especially with the current situation of decreased bed capacity and increasing costs. Therefore healthcare, accurately predicting LOS would have a positive on metrics. The length of stay (LOS) is an important indicator of the efficiency of hospital management. The purpose of this study was to determine which factors are associated with length of hospital stay, based on electronic health records, to manage the hospital, stay more efficiently our aim of the project is to predict the length of stay of the patients using the CART algorithm Classification and Regression Trees (CART) is only a modern term for what is otherwise known as Decision Trees. In this paper, we devise a two-stage classification model to classify patients into resource user groups with lower variability by using a digital record of the patient's health. It is possible to use a variety of statistical methods to divide patients into groups with lower levels of variability in their use of resources.
Key Words: Random Forest, CART Analysis, Length of stay, KNN

Introduction
Hospitals are reservoirs of critical resources and knowledge. The length of time patients spend in hospital beds is known to be a good representation of the number of resources utilized, for example, bed capacity, staffing, and equipment. When resources are limited and demand exceeds supply, allocation becomes a problem. Demand forecasts are essential for management. The demand for medical care is more complex than the demand for many other goods. Length of stay (LOS) is the number of days that an in-patient will remain in the hospital The Los for

SEYBOLD Report
ISSN: 1533 - 9211
DOI: 10.5281/zenodo.6876034

the same diagnosis may vary from 2 to 50+ days between patients. This variation can be due to several factors such as a patient's characteristics, social circumstances, or treatment complexity. Patient hospital length of stay (LOS) can be defined as the number of days that an in-patient will remain in the hospital during a single admission event The primary thing of hospital managers is to establish appropriate healthcare planning by allocating facilities and necessary human resources required a for efficient hospital operation by patient needs. The goal of this project is to create a model that can predict the length of stay for patients upon admission to a hospital.

Existing System

Harper developed Apollo, a statistical analysis program, in which he incorporated CART analysis to classify patients into similar resource user groups.[2] He classified patients according to their surgery times and found that the patient's age and surgery type were the main explanatory variables. Ting et al utilized time series models for predicting the number of discharges. But the disadvantages of these existing systems are the data taken was not that accurate to predict the length of stay of the patients.

Proposed System

Classifying patients according to their surgery type is not enough as the remaining variability in patients' Los is significantly large for efficient planning. Hence, we were required to explore other covariates that could help explain the remaining variability. Moreover, a major cause of patient flow failure is the long-stay patients. When many long-stay patients show up together, they use a significant number of resources for a long time, causing cancellations and misplacements. Our aim is to in patients into the short-stay, medium-stay, and long long-staying groups in each surgical department. To manage patient flow in a surgical suite efficiently, we need to have effective strategic patient flow management (SFM) and operational patient flow management (OFM) policies. SFM deals with long-term decision-making, such as designing a master surgery schedule (MSS) whereas OFM focuses on efficient management of patient flow every day. We can improve both SFM and OFM policies, if we can predict patients' resource requirements or Loss accurately. This is because we can schedule the elective patients' arrivals according to the resource availability.

Literature Survey

A) Clustering patient length of stay using mixtures of Gaussian models and phase type distributions

Gaussian mixture distributions and Coxian phase-type distributions have been popular choices for model-based clustering of patients' length of stay data [3] This paper compares these models and presents an idea for a mixture distribution comprising components of both of the above distributions. Also, a mixed distribution survival tree is presented. A stroke dataset available from the English Hospital Episode Statistics database is used as a running example.

SEYBOLD Report
ISSN: 1533 - 9211
DOI: 10.5281/zenodo.6876034

B) Time-Series Approaches for Forecasting the Number of Hospital Daily Discharged Inpatients

This research compares three models: a model combining seasonal regression and ARIMA, a multiplicative seasonal ARIMA (MSARIMA) model, and a combinatorial model based on MSARIMA and weighted Markov Chain models in generating forecasts of daily discharges [1]. The models are applied to three years of discharge data of an entire hospital. Several performance measures like the direction of the symmetry value, normalized mean squared error and mean absolute percentage error is utilized to capture the under- and overprediction in model selection.[4] The findings indicate that daily discharges can be forecast by using the proposed models.

C)Can Cluster-Boosted Regression Improve Prediction: Death and Length of Stay in the ICU? we review relevant literature on predicting values in healthcare data, and on predicting death in ICUs in particular BACKGROUND Being able to predict and manage the length of stay (LOS) in the ICU is important for efficient utilization of resources Prediction of who is likely to die in the ICU, or after leaving the ICU, is also important, both in preventing premature death and in allocating resources (such as rapid response teams)[7] more efficiently Prediction of death in the ICU has been a focus of research for some time. For instance, the APACHE2 scoring system was developed for assessing the risk of death in intensive care Much of the subsequent work on data mining and modeling of intensive care outcomes was summarized by Ohno-Machado Examples include a logistic regression model to predict the risk of death for children under age 16 in the ICU and univariate and multivariate logistic regression analysis to find independent predictor variables of death in patients with acute type A aortic dissection More recently, Hug developed a real-time acuity score suitable for continuous risk assessment of ICU patients

D)A Framework for Operational Modelling of Hospital Resources

For hospitals where decisions regarding acceptable rates of elective admissions are made in advance based on expected available bed capacity and emergency requests, accurate predictions of inpatient bed capacity are especially useful for capacity reservation purposes

E) Understanding the current state of patient flow in a hospital:

Daily bed shortages are mostly influenced by the timing of arrival and [5] discharge of patients with a short length of stay. Patients who stay for longer than one to two days contribute most significantly to the observed weekly bed availability problem

F) Modeling hospital length of stay by Coxian phase-type regression with heterogeneity

Variability and unpredictability are typical features of emergency departments (EDs) where patients randomly arrive with diverse conditions. Patient length of stay (LOS) represents the consumption level of hospital resources,[3] and it is positively skewed and heterogeneous. Both accurate modelings of patient ED LOS and analysis of potential blocking causes are especially useful for patient scheduling and resource management. To tackle the uncertainty of ED LOS, this paper introduces two methods: statistical modeling and distribution fitting. The models are applied to 894 respiratory disease patients' data in the year 2014 from Ethe D of a Chinese public tertiary hospital. Covariates recorded include patient region, gender, age, arrival time, arrival mode, triage category, and treatment area. A Coxian phase-type (PH) distribution model

SEYBOLD Report
ISSN: 1533 - 9211
DOI: 10.5281/zenodo.6876034

with covariates is proposed as an alternative method for modeling ED LOS [4]

G) Classification trees. A possible method for iso-resource grouping in intensive care

Classification and grouping of clinical data into defined categories or hierarchies is difficult in intensive care practice. Diagnosis-related groups are used to categorize patients based on the diagnosis. However, this approach may not apply to intensive care [5] where there is wide heterogeneity within diagnostic groups. Classification tree analysis uses selected independent variables to group patients according to a dependent variable in a way that reduces variation

Related Work

1) SOFTWARE DEVELOPMENT LIFE CYCLE (SDLC):

Software development life cycle is a cycle that has several stages such as requirement gathering, analysis, designing, coding, testing, and maintenance The SDLC aims to produce high-quality software that meets or exceeds customer expectations,

2) CART ANALYSIS:

Classification and Regression Trees (CART) is only a modern term for what is otherwise known as Decision Trees. Decision Trees have been around for a very long time and are important for predictive modeling in Machine Learning. The decision tree algorithm falls under the category of supervised learning. They can be used to solve both regression and classification problems. We can represent any Boolean function on discrete attributes using the decision tree. The CART model is used to find out the relationship between defective transactions and “amount,” “channel,” “service type,” “customer category” and “department involved.” After building the model, the Cp value is checked across the levels of the tree to find out the optimum level at which the relative error is minimum.

3) KTH NEAREST NEIGHBOUR (KNN):

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on the Supervised Learning technique. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a good suite category by using K- NN algorithm. K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for Classification problems.

4) RANDOM FOREST ALGORITHM:

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML As the name suggests, "Random Forest is a classifier that contains several decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, it predicts the final output.

SEYBOLD Report
ISSN: 1533 - 9211
DOI: 10.5281/zenodo.6876034

Implementation

The implementation phase is less creative than the system design. It is primarily concerned with user training, and file conversion. The system may be requiring extensive user training. The initial parameters of the system should be modified as a result of programming. A simple operating procedure is provided so that the user can understand the different functions clearly and quickly.

1) System Testing:
Testing has become an integral part of any system or project e, especially in the field of information technology. The importance of testing is a method of justifying if one is ready to move further and if one is capable to withstand the rigors of a particular situation cannot be underplayed and that is why testing before development is so critical.

2) Module Testing:
To locate errors, each module is tested individually. This enables us to detect error and correct it without affecting any other modules. Whenever the program is not satisfying the required function, it must be corrected to get the required result.

3) Integration Testing:
After the module testing, the integration testing is applied. When linking the modules there may be a chance for errors to occur, these errors are corrected by using this testing. In this system, all modules are connected and tested.

Results And Discussion
To implement this project, we have designed the following modules:
1) Upload Patient Stay Dataset: Using this module we will upload the dataset to the application.
2) Dataset Pre-processing: Using this module we will read the dataset and then replace missing values and then encode all non-numeric data to numeric values and then the lit dataset into train and test.
3) Run Two-State CART Algorithm: The processed train dataset will be trained with Random Forest, KNN, and CART algorithm by using classification in STAGE 1 and then train all 3 algorithms in STAGE 2 using clustering and regression algorithm. All these algorithms get trained using a cross-validation algorithm which trains each algorithm multiple times and then finds the average error rate.
4) Predict Length of Stay: Using this module we will upload patient test data and then the CART algorithm will analyze patient test data and predict the length of stay.

5) CV Error Graph: using this module we will plot the misclassification error rate for each

SEYBOLD Report
ISSN: 1533 - 9211
DOI: 10.5281/zenodo.6876034

algorithm.

The results show that the CART analysis is a useful tool for clustering patients, and it can perform feature selection even when there are a large number of predictor variables. We can fit the probability distributions to each partition obtained from the CART analysis which is useful for developing robust tactical and operational planning policies. By using the CART analysis, we were able to reduce the relative CV error in predictions up to 58:69% from the stage one predictor variables. By using the novel approach, we have developed, we were able to reduce the relative CV error in predictions further up to 9:0%. We compared the performance of the CART with that of the k-nearest neighbor regression (known) and the random forest (RF).

Dataset Size	Algorithm Name	Stage1 Error	Stage2 Error
1000	Random Forest	0.528	0.5229999999999999
1000	KNN	0.906	0.906
1000	Two-Stage CART	0.5730000000000001	0.55

FIG 2: OUTPUT. HTML

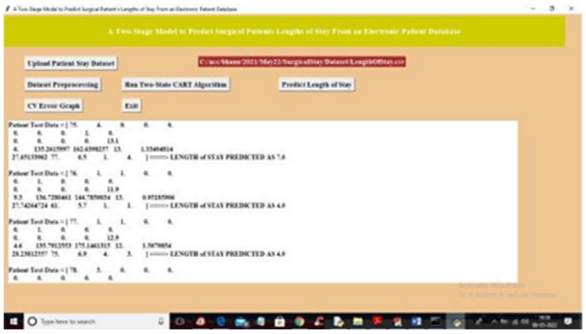

FIG 3: LENGTH OF STAY IS PREDICTED

FIG 4: CV ERROR GRAPH

Conclusion And Future Scope

The ability to predict LOS can provide a clinical indicator of the health status of a patient as well as assist in predicting the level of care that is required. It also aids hospital staff with improved prediction of bed and ward utilization. LOS varies concerning many factors including the severity of illness, diagnosis, and a variety of patient factors. This paper provides a review of LOS prediction methods, their respective shortcomings as well as the types of data and features that have been used in the literature. Despite the continuing efforts to predict and

DOI: 10.5281/zenodo.6876034

reduce the Los of patients, current research in this domain remains ad-hoc; as such, the model tuning and data pre-processing steps are too specific and result in a large proportion of the current prediction mechanisms being restricted to the hospital that they were employed in. Additionally, several studies focus on hospitals that are contained within very densely populated areas.

References

[1] AIHW, "Australia's health series no. 15. Cat. no. AUS 199," Canberra: AIHW, 2015," [Online; accessed20January2018]". n[Online]. Available: https://www.aihw.gov.au/reports/australias-health/ Australia's-health-2016/contents/summary

[2] P. R. Harper, "A framework for operational modeling of hospital resources," Health Care Manag. Sci., vol. 5, no.3, pp. 165–173, 2002.

[3] M. Faddy, N. Graves, and A. Pettitt, "Modeling length of stay in hospital and other right-skewed data: Comparison of phase-type, gamma, and lognormal distributions," Value Heal., vol. 12, no. 2, pp. 309–314, 2009.

[4] A. H. Marshall and S. I. McClean, "Using coxian9 phase-type distributions to identify patient characteristics for duration of stay in the hospital," Health Care Manag. Sci., vol. 7, no. 4, pp. 285–289, 2004.

[5] G. W. Harrison and G. J. Escobar, "Length of stay and imminent discharge probability distributions from multistage models: Variation by diagnosis, severity of illness, and hospital," Health Care Manag. Sci., vol. 13, no.3, pp. 268–279, 2010.

[6] L. Garg, S. McClean, B. Meenan, E. El-Darzi, and P. Millard, "Clustering patient length of stay using mixtures of Gaussian models and phase type distributions," in 2009 22nd IEEE Int. Symp. Comput. Med. Syst. IEEE, Aug 2009, pp. 1–7. [Online]. Available: http://ieeexplore.ieee.org/document/5255245/

[7] E.-D. Elia, A. Revlin, V. Christos, G. Florin, G. Marina, and M. Peter, Intelligent Patient Management, ser. Studies in Computational Intelligence, S. McClean, P. Millard, E. El-Darzi, and C. Nugent, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, vol. 189. [Online]. Available: http://link.springer.com/10.1007/978-3-642-00179-6

[8] X. Tang, Z. Luo, and J. C. Gardiner, "Modeling hospital length of stay by Coxian phase-type regression with heterogeneity," Stat. Med., vol. 31, no. 14, pp. 1502–1516, 2012.