Risk Stratification for Hospital Readmission of Heart Failure Patients: A Machine Learning Approach

Chun Pan Hon
Center for Data Science
Institute of Technology
University of Washington
Tacoma
darrencp@uw.edu

Mayana Pereira
Center for Data Science
Institute of Technology
University of Washington
Tacoma
mayanaw@gmail.com

Shanu Sushmita
Center for Data Science
Institute of Technology
University of Washington
Tacoma
sshanu@uw.edu

Ankur Teredesai
Center for Data Science
Institute of Technology
University of Washington
Tacoma
ankurt@uw.edu

Martine De Cock*
Center for Data Science
Institute of Technology
University of Washington
Tacoma
mdecock@uw.edu

ABSTRACT

Being able to stratify patients according to 30-day hospital readmission risk, anticipated length and cost of stay can guide clinicians in discharge planning and intervention recommendation, leading to an increase of quality of care, and a decrease of healthcare cost. In this contribution, we present a comparative performance of decision trees, boosted decision trees and logistic regression models that can flag, at the time of discharge, patients with an anticipated early, lengthy and expensive readmission. We validate our models using discharge records of 500K congestive heart failure patients from California-licensed hospitals.

1. INTRODUCTION

There are about 34 million hospital admissions annually in the U.S..[1] One in five patients is readmitted to the hospital within 30 days of being discharged. Congestive heart failure (CHF) is one of the leading causes of hospitalization, especially for adults older than 65 years of age [3]. Many of these readmissions could be avoided by proper interventions. 30-day readmission, cost, and length of stay are commonly understood as healthcare quality measures and cost drivers in the U.S. [2]. The ability to predict them accurately provides many benefits for accountable care, now a global issue and foundation for the U.S. government mandate under the Affordable Care Act.

While predicting 30-day readmission has been identified as one of the key problems for the healthcare domain, not many solutions are known to be effective [4]. In fact, to improve the clinical process of heart failure patients, healthcare organizations still leverage the proven best-practices, called "*Get With The Guidelines*" by the American Heart Association. Furthermore, uncertainty in length of hospital stay is a major deterrent to effective scheduling for admission of elective patients. A model to predict the Length of Stay (LOS) for hospitalized patients can be an effective tool for healthcare providers, as it will enable early interventions to prevent complications, among other things [1]. However, the ability to risk stratify for LOS based on patient admission and hospital characteristics is limited, and more challenging for CHF patients. Readmissions and prolonged hospital stay act as substantial contributors to rising healthcare costs [3]. Alongside of predicting 30-day readmission and LOS, in this contribution we also investigate algorithm performance for forecasting the cost of CHF hospital admissions. Previously proposed cost prediction models were primarily focused on 'general' healthcare costs as opposed to hospital admissions, and were often rule based and regression models. The development of healthcare cost prediction models using machine learning methods has been more recent (e.g. [7]).

In recent years, state and federal governments are starting to make increasing amounts of healthcare data publicly available. Analytics solutions that leverage this data for reducing cost growth are central to improving accountability in care. Though the availability of large volumes and variety of data sources has improved significantly in recent years, many state-of-the-art machine learning approaches remain unexplored so far in the healthcare analytics domain. In this study, we leverage longitudinal inpatient data made available by the California Office of Statewide Health Planning and Development (OSHPD) to train and test machine learning models for predicting, at the time of discharge, *(1) whether the next admission of the patient will be within 30 days, (2) whether the hospital stay of the next admission will be long, i.e. more than 6 days*, and *(3) whether the cost of the next admission will be high, i.e. above $95.7K*. For each of the classification tasks under study, we build logistic regression models, decision trees and boosted decision trees. We investigate the use of different demographic, administrative and clinical features, and observe how adding more feature groups improves the performance of the models. To the best of our knowledge, our work is the first effort to build and validate machine learning models for risk stratification of CHF patients in the OSHPD data.

2. METHODS AND RESULTS

We requested non public data for the years 2009-2013 from the California Office of Statewide Health Planning and Development (OSHPD). The dataset is a collection of records in tabular format with each row corresponding to one hospital discharge record of one patient. After performing a series of data preprocessing steps (see [5] for a description of similar steps), we extracted all the records of all patients who have CHF as a primary or secondary

*Guest professor at Ghent University

[1]http://www.aha.org/research/rc/stat-studies/fast-facts.shtml, accessed on Jun 11, 2016

BCB '16 October 02-05, 2016, Seattle, WA, USA

ACM ISBN 978-1-4503-4225-4/16/10.

DOI: http://dx.doi.org/10.1145/2975167.2985648[dx.doi.org]

diagnosis in at least one of their records.[2] Since we are interested in predicting the cost and length of stay of future hospitalizations, we omitted patients from the study who had only one hospital admission, bringing the total number of unique patients to 497,697 and the total number of records to 2,451,412. Table 1 shows an overview of all the features used. Some of them are taken directly from the raw OSHPD data[3], while others are constructed. In particular, we constructed the Charlson comorbidity features by converting all non-primary diagnosis codes into comorbidities using the mapping defined in [6].

Feature	Feature Group
Age Gender Race	F_1
Cost of Current Admission (Total Charges) Length of Stay of Current Admission	F_2
Type of Admission Source of Admission Source of Admission - Route MS-DRG Severity Code Type of Care Same Day Discharge Disposition	F_3
17 Charlson comorbidities	F_4
# of previous admissions # of ED visits in past 6 months # of distinct Charlson comorbidities so far # of distinct diagnoses so far	F_5

Table 1: Overview of feature groups

We trained and tested machine learning models for three binary classification problems at the time of discharge: next admission within 30 days, length of stay more than 6 days, and cost of next admission to be high, i.e. above $95.7K. A threshold of 30 days was chosen because the 30-day-readmission rate of patients is a commonly used quality metric in the U.S.. The threshold for length of stay stems from the discretization used in the popular LACE index [8]. Finally, $95.7K for the cost threshold is derived directly from the OSHPD data, splitting off the highest quartile.

For each of the classification tasks mentioned above, we built logistic regression (LR) models, decision trees (DT) and boosted decision trees (ADA). All algorithms are trained and tested using R software in a 5-fold cross-validation setup. Since the data is imbalanced (see Table 2), in each fold we undersampled the training data by randomly selecting instances from the majority class to match with the number of samples of the minority class. After undersampling both classes have the same number of instances in the training data. No undersampling was done on the test data.

The results for the three binary classification tasks are presented in terms of AUC in Table 3. Overall, three key observations can be made: (1) the AUC scores from all three methods show promise in accurately predicting early, lengthy and costly readmissions (above 60%); (2) the performance of the logistic regression models is at par with non-linear methods like decision and boosted decision trees; (3) adding more information (more features) improves the overall performance of the models. The highest AUC scores are observed when all features ($F_1 + F_2 + F_3 + F_4 + F_5$) were used. In particular, for the 30-day readmission problem, the machine learning models outperform the regularly used LACE index which yields and AUC of 0.6025 on the CHF OSHPD data.

[2]Using the ICD-9-CM codes for CHF: 398.91 and 428.XX.
[3]http://www.oshpd.ca.gov/HID/Data_Request_Center/documents/PDD_Nonpublic_DataDictionary.pdf

Task	Positive Class	Negative Class
30-Day	34.21%	65.78%
LOS (> 6 days)	29.82%	70.17%
Cost (> 95.7K)	25.08%	74.91%

Table 2: Class distributions in the dataset

Feature Combination	LR	DT	ADA
Readmission within 30-Day			
F1	0.5443	0.5379	0.5443
F1+F2	0.5631	0.5734	0.5785
F1+F2+F3	0.5958	0.5929	0.5995
F1+F2+F3+F4	0.6061	0.5980	0.6086
F1+F2+F3+F4+F5	**0.6565**	0.6411	0.6461
LOS > 6 days			
F1	0.5158	0.5276	0.5290
F1+F2	0.6057	0.6100	0.6146
F1+F2+F3	0.6077	0.6128	0.6216
F1+F2+F3+F4	0.6109	0.6159	0.6240
F1+F2+F3+F4+F5	0.6150	0.6122	**0.6285**
Cost > 95.7K			
F1	0.5300	0.5526	0.5551
F1+F2	0.6174	0.6242	0.6368
F1+F2+F3	0.6160	0.6254	0.6391
F1+F2+F3+F4	0.6180	0.6263	0.6413
F1+F2+F3+F4+F5	0.6222	0.6263	**0.6473**

Table 3: AUC results for different combinations of feature groups from Table 1. LR = Logistic Regression, DT = Decision Tree, and ADA = AdaBoost. The best results are highlighted in bold.

3. CONCLUSION

We presented a comparative performance of decision trees, boosted decision trees and logistic regression models that can flag high risk CHF patients at the time of discharge. Preliminary results show promise in using these methods for accurately stratifying patients according to 30-day readmission risk, anticipated length and cost of hospital stay. In our future research, we aim to investigate additional state-of-the-art machine learning methods, as well as improve our existing models through feature engineering.

4. REFERENCES

[1] P. R. Hachesu, M. Ahmadi, S. Alizadeh, and F. Sadoughi. Use of data mining techniques to determine and predict length of stay of cardiac patients. *Healthc Inform Res*, 19(2):121–129, Jun 2013.

[2] A. Hines, M. Barrett, H. Jiang, and C. Steiner. Conditions with the largest number of adult hospital readmissions by payer. *HCUP Statistical Brief*, 172, 2011.

[3] S. F. Jencks, M. V. Williams, and E. A. Coleman. Rehospitalizations among patients in the Medicare fee-for-service program. *New England Journal of Medicine*, 360(14):1418–1428, 2009.

[4] K. Ottenbacher, P. Smith, S. Illig, R. Linn, R. Fiedler, and C. Granger. Comparison of logistic regression and neural networks to predict rehospitalization in patients with stroke. *Journal of Clinical Epidemiology*, 54(11):1159–1165, 2001.

[5] M. Pereira, V. Singh, C. P. Hon, T. G. McKelvey, S. Sushmita, and M. De Cock. Predicting future frequent users of emergency departments in California state. In *Proceedings of the 1st Workshop on Methods and Applications for Healthcare Analytics (MAHA) in conjunction with ACM BCB 2016*, 2016.

[6] H. Quan, V. Sundararajan, P. Halfon, A. Fong, B. Burnand, J.-C. Luthi, L. D. Saunders, C. A. Beck, T. E. Feasby, and W. A. Ghali. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Medical Care*, pages 1130–1139, 2005.

[7] S. Sushmita, G. Khulbe, A. Hasan, S. Newman, P. Ravindra, S. B. Roy, M. De Cock, and A. Teredesai. Predicting 30-day risk and cost of "all-cause" hospital readmissions. In *Workshops at the Thirtieth AAAI Conference on AI*, 2016.

[8] B. Zheng, J. Zhang, S. W. Yoon, S. S. Lam, M. Khasawneh, and S. Poranki. Predictive modeling of hospital readmissions using metaheuristics and data mining. *Expert Systems with Applications*, 42(20):7110–7120, 2015.