SMU Data Science Review

Volume 2 | Number 1 Article 22

2019

Identifying High Risk Patients for Hospital Readmission

Ethan Graham
Southern Methodist University, ethang@smu.edu

Asha Saxena
Southern Methodist University, ashas@smu.edu

Heather Kirby
Frederick Regional Health System, HKirby@fmh.org

Recommended Citation

Graham, Ethan; Saxena, Asha; and Kirby, Heather (2019) "Identifying High Risk Patients for Hospital Readmission," *SMU Data Science Review*: Vol. 2: No. 1, Article 22.
Available at: https://scholar.smu.edu/datasciencereview/vol2/iss1/22

Identifying High Risk Patients for Hospital Readmission

Ethan Graham[1], Asha Saxena[1], Heather Kirby[2]
[1] Master of Science in Data Science, Southern Methodist University,
Dallas, TX 75275 USA
[2] Frederick Regional Health System, 400 West Seventh Street,
Frederick, MD 21701 USA
{ethang, ashas}@smu.edu
HKirby@fmh.org

Abstract. The Affordable Care Act (ACA), passed in 2010, set forth a framework for healthcare providers to have a vested interest in better patient outcomes and to reduce the Total Cost of Care (TCOC) for patients. A large portion of TCOC comes from patients who make multiple unscheduled hospital visits for the same underlying pathology: a hospital readmission. In this paper, we tackle the difficulty of identifying risk markers for diabetes patients' hospital readmission. Using data from the Health Facts Database, we use logistic regression and support vector machines to identify the risk that a diabetes patient has of a hospital readmission.

1 Introduction

The healthcare industry is facing new challenges that will affect its entire operation and execution. One of the biggest challenges is the shift from fee for service (healthcare organizations are paid only for service rendered) to value-based care (healthcare organizations are paid to keep their population healthy). As healthcare organizations shift focus to high quality and care, managing cost has been challenging. According to Brian Morrissey, current president of the Board of Managers for Acuitas Health, "The biggest challenge in value based health care, population health and care management is identifying the right patients and engaging them so that the resources invested achieve the greatest improvement in outcomes" [1]. Another major challenge is managing hospital readmissions. Predicting and lowering the hospital readmission rate is the low hanging fruit for healthcare organizations seeking increase quality of service and decrease healthcare spending. Nearly 20% of Medicare beneficiaries are hospitalized within 30 days after discharge, at an annual cost of $17 billion. The Affordable Care Act (ACA) therefore created a financial penalty for "excessive" readmissions at hospitals.

Common care management steps, when implemented during the first 30 days after discharge, can help health systems reduce avoidable hospital inpatient readmission rates. Our research will focus on finding the high-risk patients who are prone to coming back and getting readmitted for less than 30-day readmission. We will review the data set to understand the patterns of those patients who are readmitted and recommend solutions to fix the problem. The medical industry is undergoing Our objective is to review the accuracy that is achievable when using logistic regression and support vector machines (SVM).

Our dataset is the diabetes dataset of 130 US hospitals for the years from 1999-2008 from the UCI Machine learning repository. The diabetic patient population offers a large data set, and a large opportunity for healthcare cost savings with proper care.
In 2015, 30.3 million Americans, or 9.4% of the population, were living with diabetes. Of those, 1.25 million are children, and 7.2 million were undiagnosed. Every year, another 1.5 million Americans are diagnosed with diabetes, and in 2015, 84.1 million Americans minors had prediabetes. Also in 2015, diabetes was the 7th most common cause of death in the United [2].

With these large numbers of patients and mortality rates, comes large healthcare costs. In 2017, diabetes racked up $237 billion in direct medical costs, $327 billion total costs, and $90 billion in reduced productivity. The average medical spending for diabetes patients is 2.3 times more than similar patients without diabetes [2].

2 A Primer on Hospital Care Models

Hospitals have operated on a fee for service model for many decades. Fee for service means hospitals and emergency rooms are paid by number of visits and tests ordered [3]. In the interest of increasing the quality of care, the healthcare industry is moving to a value-based care model, where payment is based on the value of care received. Value-based payments are still being perfected, and most are based on a shared savings model. Shared savings incentivize providers to increase cost savings for a patient population with direct kickbacks of the cost savings to the provider. This shift affects all aspects of the medical billing industry, from accounting, quality measures, physician performance metrics, and patient outcomes. To track these new performance metrics, powerful analytics for each patient population are used. This allows care providers real-time monitoring of their performance, reimbursements, as well as the ability to identify which providers are high performing. A key aspect for performance measurement is tracking the 30-day readmissions for a provider.

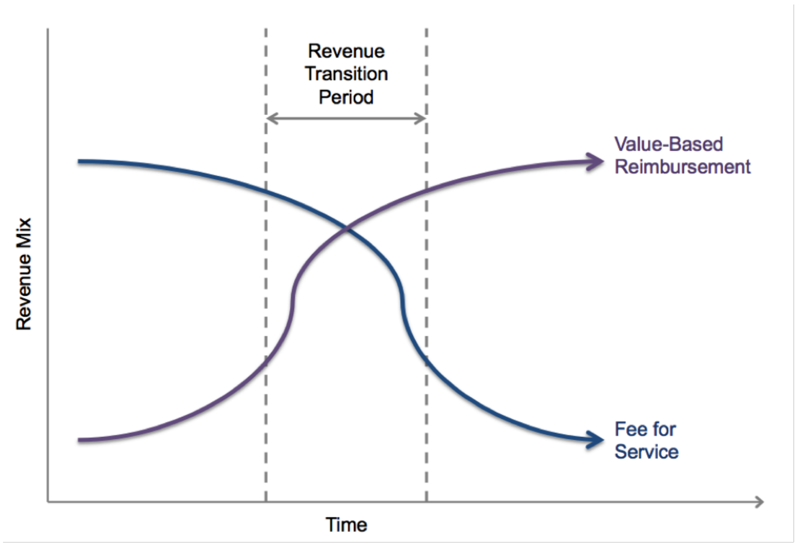

Fig. 1. A conceptual illustration of the changing revenue models for healtcare providers [3].

Avoiding hospital readmissions is important to keeping costs down. The ACA added section 1886(q) to the Social Security Act, which established the Hospital Readmissions Reduction Program (HRRP) in 2012 [4]. This reduces payments to hospitals with excess readmissions. The financial incentive structure of the HRRP makes the first 30 days after patient discharge a critical point in care management. If readmission rates are lowered, this will impact patient welfare, quality of care, and the bottom line of providers. Under the HRRP, healthcare providers are also penalized for unacceptably high readmission rates among Medicare and Medicaid patients [5]. Medicare levies penalties for exceeding readmission benchmarks [4]. From 2010 to 2015, through various patient intervention efforts, hospitals have reduced readmissions by an estimated total of 565,000 patients. Even with this good work, there's still much room for improvement: the federal government estimates that readmissions cost $26 billion per year, and that 65% of those readmissions are avoidable.

2.1. Risk Scores

Risk scoring is part of healthcare analytics that attempts to quantify some aspect of a patient's health [6]. Typically, patients only receive one risk score. With the introduction of more big data analytics into healthcare, risk scores by disease state. The introduction of granular risk scores allows healthcare providers to deliver care to the right patient at the right time, and even at a lower cost. Risk scoring is used by insurance providers as well as healthcare providers, and the models that they create. These risk scores inform premiums for covered individuals, and the coverage options available to prospective covered populations. All types of insurance rely on some type of risk scoring strategy for their insured populations, and the accuracy of risk scores is paramount to ensuring the best outcomes for the insurance providers as well as the insurance policy holders.

3 Data Collection

Our data set is publicly available on UC Irvine's machine learning repository website. The data were submitted to UC Irvine on behalf of the Center for Clinical and Translational Research, Virginia Commonwealth University. This dataset is compliant with the Health Insurance Portability and Accountability Act (HIPAA), and is a de-identified abstract of the Health Facts database [7]. The dataset contains data from 1999 to 2008 of clinical care information at 130 US healthcare providers throughout the United States: Midwest (18 hospitals), Northeast (58), South (28), and West (16). 78 of the hospitals have bed counts between 100 and 499, 38 hospitals have bed counts less than 100, and 14 hospitals have more than 500 beds. Over 50 variables are recorded for each entry, representing a patient's current state when they were admitted to the healthcare provider. The variables present cover a wide range of data points about a patient: such as race, gender, age, admission type, time in hospital, number of lab test performed, number of medications prescribed, diabetic medications, emergency visits in the year before the hospitalization, etc.

The Health Facts database (Cerner Corporation, Kansas City, MO) is a national data warehouse that keeps comprehensive clinical records on healthcare providers enrolled in the voluntary Health Facts program, offered through organizations which use the Cerner Electronic Health Record System. The database is comprised of 41 tables with a total of 117 features. The tables are stored in a fact-dimension schema. The database describes 74,036,643 unique visits, 17,880,231 unique patients, and 2,889,571 providers. This data was captured for inpatients and outpatients of individual hospitals as well as integrated delivery network health systems. However, data from out-of-network providers is not captured. Our data set is a simplified view on this database, specific to readmission risk amongst patients with diabetes.

The data was explored and cleaned using Python 3.6.0.

3.1. Data Cleaning

The data was downloaded in CSV format as a file around 20MB. The file is read into the Python computing environment. Our dataset has missing values for both race and gender, so we drop the missing race values and recode the 3 unknown gender values to 'female', the most frequent record. As there are so few unknown records recoding them to the categorical attribute's mode is sufficient. The "age" variable is stored as text in the format "[0-10)" to indicate a range from 0 to 10 years. We re-key the "age" variable to a categorical numeric values 1 for less than 10 years old, 2 for less than 20 years old, etc. We do a similar change for the "race" variable, mapping it to another categorical numeric with 1 for "Asian", 2 for "AfricanAmerican", etc. The "insulin" variable is also mapped from the text based values "No", "Up", "Down", "Steady" to 0, 1, 2, 3 respectively. The dataset "readmitted" variable is a text based value with "NO", "<30" for a readmission in less than 30 days, and ">30" for a readmission in 30 or more days. The "readmitted" variable is mapped to 0, 1, and 0 respectively. Notice that ">30" is encoded to 0, the same value we encoded for "NO". This is done because for the purposes of this study we define readmission only for events within the first 30 days. We also create the "isMale" and "ONdiabetesMed" indicator variables based on the existing data in the dataset. With these new columns, we no longer need the original "gender" and "diabetesMed" columns and we drop them from any further analysis.

Table 1. Example variables and the variable type used for further analysis.

Variable	Type
age	interval
race	categorical
discharge_disposition_id	categorical
insulin	categorical
readmitted	ordinal, independent
admission_type_id	categorical
admission_source_id	categorical
time_in_hospital	categorical
number_diagnoses	numeric
gender	categorical
insulin	ordinal

Table 2. A sample of the uncleaned dataset.

patient_nbr	race	gender	age	weight	diabetesMed	readmitted
8222157	Caucasian	Female	[0-10)	No	No	NO
55629189	Caucasian	Female	[10-20)	Ch	Yes	>30
86047875	AfricanAmerican	Female	[20-30)	No	Yes	NO
82442376	Caucasian	Male	[30-40)	Ch	Yes	NO

Table 3. A sample of the cleaned dataset.

age	race	readmitted	diabetesMed	IsMale	ONdiabetesMed	…
1	3	1	0	0	0	…
2	3	0	1	0	1	…
3	2	1	1	0	1	…
4	3	0	1	1	1	…
5	3	0	1	1	1	…

After removing rows with missing data we can't recover, we are left with 99,493 rows total. While the data is in this state, we find that we can achieve a 96% compression compared to the size of the source dataset. This compressed format is useful for storage and transfer. There are still several remaining categorical columns that need to be one-hot encoded, then interleaved back into the data set, leaving one categorical dummy column out to avoid multicollinearity. Finally, we end up with a 61 column table to train our classifier on.

5 Machine Learning Models and Training Methods

5.1. Logistic Regression

We use the scikitLearn's ShuffleSplit package to do cross validation with 5 iterations [8]. A logistic regression model is selected as we are performing a binary classification: high risk, not high risk. The dataset is very unbalanced, with a ratio of readmitted patients to non-readmitted patients of 0.126. To compensate for this imbalanced data, we use stratified cross validation, which ensures that the proportion of each class is preserved in each sample used in a cross-validation fold. We then use recursive feature elimination to pare down the number of variables used in analysis to only include the most impactful variables:

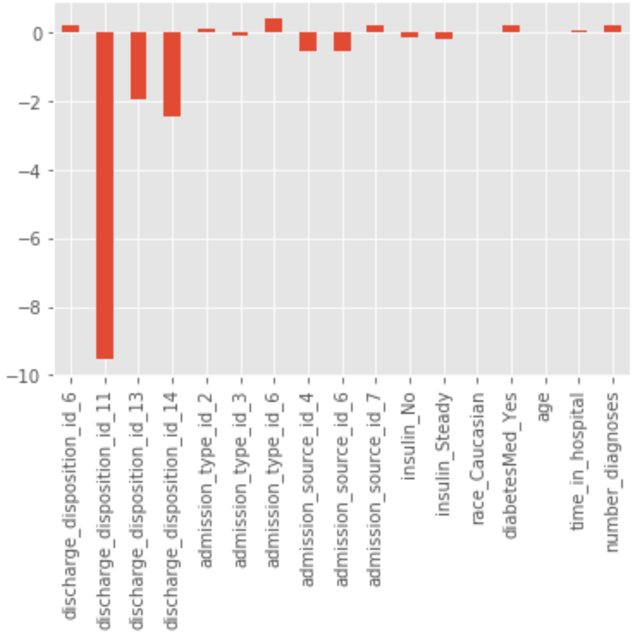

Fig. 2. The reduced variables and associated weights

Table 4. Confusion matrix from logistic regression with cross validation.

Iteration	Confusion Matrix		
		No	Yes
0	No	17630	35
	Yes	2230	4
		No	Yes
1	No	17664	1
	Yes	2234	0
		No	Yes
2	No	17661	4
	Yes	2230	4
		No	Yes
3	No	17658	7
	Yes	2231	3
		No	Yes
4	No	17655	9
	Yes	2231	2

Figure 2 illustrates that the discharge disposition has an outsized influence in the classification of high risk patents. The discharge disposition refers to where the patient was sent after their hospital stay: home, hospice, addiction center, etc. The accuracy of the model is promising at over 88% for each iteration, but the sensitivity is low, at below 0.5 on average. Specificity, or recall, refers to the true positives identified by the model, calculated as

$$s = \frac{(True\ positives)}{(True\ positives) + (False\ Negatives)} \tag{1}$$

$$CI_s = s \pm z\sqrt{\frac{(1-s)\times s}{(True\ positives) + (False\ Negatives)}} \tag{2}$$

Table 2 shows the confusion matrices of one of the analyses. A confusion matrix shows how many times the model correctly and incorrectly categorized each patient in the data set. This is a concise way to assess the quality of a model. By cherry picking the confusion matrix in iteration 2, which has the highest specificity, we calculate the 95% confidence interval of the specificity of this model as (0.323, 0.676). In addition, the false positive rate, α is 1 less the specificity. This indicates we can identify some of the high risk patients, but there are also a number of non-high risk patients that are classified as high risk. Due to the imbalanced nature of this data, we can expect that the total number of false positives will be greater than the total correctly classified patients over time. False positives arising from this classifier will cause a loss of effective patient intervention and a suboptimal return on investment for any such intervention. This suggests that a new patient risk score may have more utility than binary classification for these patients.

5.2. Support Vector Machine

We use scikitlearn's LinearSVC package to create a model the risk score associated to a patient's chance of hospital readmission. Using grid search to tune hyperparameters and different scoring schemes results in an optimal model using "neg_median_absolute_error" scoring function, and other hyperparameters specific to the LinearSVC model. Again applying recursive feature elimination to reduce the feature space results in a similar set of important features. We can use these feature importance's to generate a lift chart to determine expected ROI on for different patient intervention campaigns.

6 Results and Analysis

Generating a risk score for patients, along with feature importance for the reduced feature set, allows us to generate a lift chart that will predict how effective patient intervention schemes will be, given a specific budget:

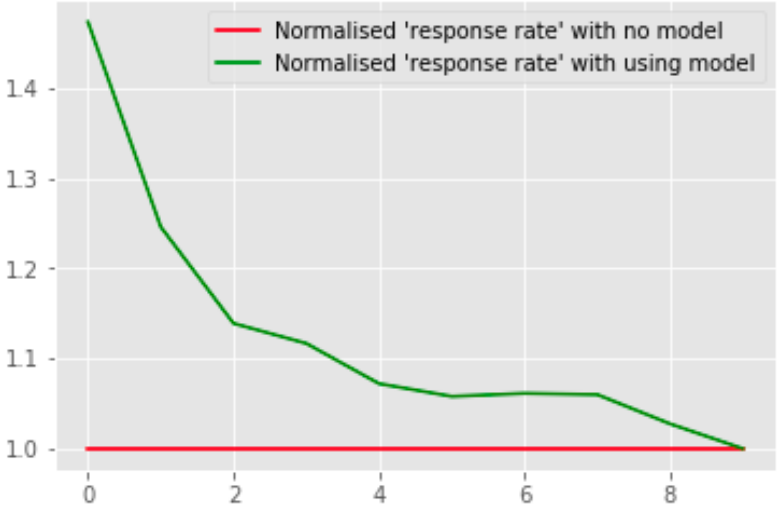

Figure 3. Lift chart of predicted response rate vs. random response rate by proportion of the population.

By focusing on patients with the highest risk score, we estimate a 480% increase in effectiveness over a random intervention strategy. According to SpeechMed, the average cost of hospital readmission for Medicare patients is $13,800 [9]. Assuming that the intervention strategy is 90% effective, healthcare providers could invest $2205 on the high risk patients and expect a positive ROI. Of course, the final numbers would need to be evaluated based on the effectiveness of the interventions.

This approach offers positive outcomes for some, but not all patients of interest. High risk patients that are not identified by this approach will unfortunately not receive extra care or consideration. However further improvements to this approach may be able to capture these patients and ensure they receive the extra care needed to prevent readmission.

7 Ethics

The IEEE Code of Ethics protects the safety, health, and welfare of the public [10]. This code of ethics is an essential guide for all undertakings in data science. It is an important guide for how to collect and protect data, as well as conduct and present research. Currently, no formal body exists to accredit or sanction data scientists specifically, international legislation is now starting to enforce these best practices for handling data.

Medical data comes with its own set of considerations. When dealing with medical data, there are specific and essential processes to follow to ensure all parties are protected. Ethics practices are meant to prevent harm, and for medical data used in research we must consider preventing harm to doctors, patients, researchers, research funding institutions, and their respective reputations. The favored mechanism to achieve this is confidentiality [11]. Confidentiality is used throughout the medical profession to create an environment of trust, respect, and privacy. The duty of confidentiality impacts what can be shared to researchers. In order to maintain confidentiality, a process known as anonymization or de-identification happens to data sets before they can be used for most research. De-identified data removes sensitive information such as names, social security numbers, insurance plan numbers, addresses, and anything that can reasonably be used to directly identify the person whose data was captured in the data set. In some cases, further de-identification, such as date shifting, is used [12]. Date shifting takes a date such as admission date or birthdate, and changes it to some other date in the future or past. The same change is applied to all dates in the data set, and thus the relationship between dates is preserved and is still useful for most types of analysis.

The dataset used for this analysis is HIPAA compliant, and anonymized, thus basic ethical standards are met. The General Data Protection Regulation (GDPR) European regulation may also impact the future availability of the data used for this study. The GDPR obligates data collectors of European citizens to erase any stored data upon request. The implications of this legislation have yet to be fully realized, but it may end up necessitating a similar response as has happened to direct surveys over the past few years. Direct surveys have seen an increasing incidence of non-response from the population, with a concordant non-response error adjustment needed for all such research. This adjustment may be used for to accommodate the future GDPR claims by individuals. The full impact of these types of regulations is still to be seen, but the analysis made in this research is legitimate none the less.

8 Conclusion

Avoiding readmissions is critical towards providing quality of care and reducing cost. Readmission penalties have been a major cause of distress for many healthcare provides. Medicare penalized about 2,500 hospitals last year by withholding more than half a billion dollars in payments. In efforts to reduce readmission rates, healthcare organizations are trying to understand what factors affect readmissions and how to control it.

In our capstone project we are analyzing the data set for patients who are readmitted for the diagnosis of diabetes and understand the factors that impact the readmissions. The data was provided by University of California, Irvine. We identified 11 major variables (age, race, discharge disposition id, insulin, readmitted, admission type id, admission source id, time hospital, number diagnoses, ismale, on diabetes med), these variables were identified as the primary variables that may have impacted the readmission rate for the patient population under consideration.

During our data preparation, we normalized our data for the purpose of running our models and reviewing the results. In this phase of our study we started our analysis by visualizing the data set we have at hand. The initial results indicated that the readmission risk was for the patients who were on diabetes medication and had larger number of various diagnoses in addition to diabetes. We also noticed that older patient population had the higher risk of readmission versus younger population. The results are not confirmed as we do see some variability in the charts as we ran our models. We further need more work and statistical analysis techniques be employed to gain additional insight. We intend to continue or analysis and understand what insights the data set provides to identify high risk indicators.

References

[1] "Acuitas Health Selects Health Catalyst to Support Physicians' Transition to Value-Based Care in New York's Capital Region," *Health Catalyst*. [Online]. Available: https://www.healthcatalyst.com/news/acuitas-health-selects-health-catalyst-to-support-physicians-transition-to-value-based-care. [Accessed: 12-Jul-2018].

[2] A. D. A. 2451 C. Drive, S. 900 Arlington, and Va 22202 1-800-Diabetes, "Statistics About Diabetes," *American Diabetes Association*. [Online]. Available: http://www.diabetes.org/diabetes-basics/statistics/. [Accessed: 11-Jul-2018].

[3] B. Bobbi and C. Jared, "The Key to Transitioning from Fee-for-Service to Value-Based Reimbursement," *HealthCatalyst*, 03 2018.

[4] C. for Medicare, M. S. 7500 S. B. Baltimore, and M. Usa, "Readmissions-Reduction-Program," 27-Apr-2018. [Online]. Available: https://www.cms.gov/medicare/medicare-fee-for-service-payment/acuteinpatientpps/readmissions-reduction-program.html. [Accessed: 09-Jul-2018].

[5] K. Marder, "Ten Essential Steps for Your Readmission Reduction Program (Executive Report)."

[6] "Risk Scoring: Big Data and Advanced Analytics Further Evolve the Healthcare Model," *Knowledgent*.

[7] "UCI Machine Learning Repository: Diabetes 130-US hospitals for years 1999-2008 Data Set." [Online]. Available: https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008. [Accessed: 14-Jul-2018].

[8] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[9] G. Olivia, "The Cost of a Hospital Readmission," Dec. 2016.

[10] "IEEE Code of Ethics." [Online]. Available: http://www.ieee.org/about/corporate/governance/p7-8.html. [Accessed: 14-Sep-2018].

[11] "Confidentiality: Ethical Topic in Medicine." [Online]. Available: https://depts.washington.edu/bioethx/topics/confiden.html. [Accessed: 14-Sep-2018].

[12] A. Long, "Clinical natural language processing for predicting hospital readmission," *Insight Data*, 14-Jun-2018. [Online]. Available: https://blog.insightdatascience.com/introduction-to-clinical-natural-language-processing-c563b773053f. [Accessed: 14-Sep-2018].