The Workshops of the Thirtieth AAAI Conference on Artificial Intelligence
Expanding the Boundaries of Health Informatics Using AI:
Technical Report WS-16-08

Predicting 30-Day Risk and Cost of "All-Cause" Hospital Readmissions

Shanu Sushmita, Garima Khulbe, Aftab Hasan, Stacey Newman
Center for Data Science, Institute of Technology, University of Washington, Tacoma
sshanu@uw.edu, garima1@uw.edu, aftabh@uw.edu, newmsc@uw.edu

Padmashree Ravindra, Senjuti Basu Roy, Martine De Cock, Ankur Teredesai
Center for Data Science, Institute of Technology, University of Washington, Tacoma
padmashree.ravindra@gmail.com, senjutib@uw.edu, mdecock@uw.edu, ankurt@uw.edu

Abstract

The hospital readmission rate of patients within 30 days after discharge is broadly accepted as a healthcare quality measure and cost driver in the United States. The ability to estimate hospitalization costs alongside 30 day risk-stratification for such readmissions provides additional benefit for *accountable care*, now a global issue and foundation for the U.S. government mandate under the Affordable Care Act. Recent data mining efforts either predict healthcare costs or risk of hospital readmission, but not both. In this paper we present a dual predictive modeling effort that utilizes healthcare data to predict the risk and cost of any hospital readmission ("all-cause"). For this purpose, we explore machine learning algorithms to do accurate predictions of healthcare costs and risk of 30-day readmission. Results on risk prediction for "all-cause" readmission compared to the standardized readmission tool (LACE) are promising, and the proposed techniques for cost prediction consistently outperform baseline models and demonstrate substantially lower mean absolute error (MAE).

1 Introduction

Patients with chronic conditions repeatedly get admitted to a hospital for treatment and care. They are often discharged when their condition stabilizes only to get readmitted again, many times within just a few days. This process is termed as *hospital readmissions*. The readmission problem in the U.S. is severe: currently one in five (20%) Medicare patients are readmitted to a hospital within 30 days of discharge. Three quarters of these readmissions (75%) are actually considered avoidable (Jencks, Williams, and Coleman 2009). In addition to raising red flags about gaps in quality of care, hospital readmissions also place a huge financial burden on the health system. In 2011, there were approximately 3.3 million adult 30-day all-cause hospital readmissions in the United States, and they were associated with about $41.3 billion in hospital costs (Hines et al. 2011). Avoidable readmissions account for around $17 billion a year (Jencks, Williams, and Coleman 2009). In the U.S., the readmission rate of patients at a hospital is tracked as a proxy for measuring the overall quality of treatment a patient has received, and, under the Affordable Care Act, Medicare has started penalizing hospitals that have higher-than-expected rates of 30-day readmissions[1].

In this paper we tackle two related problems, namely *(1) predicting whether a patient is at risk of being readmitted to the hospital within 30 days after discharge*, and *(2) estimating the cost of that hospital readmission*. The ability to prioritize a care plan along both of these variables can enable hospital systems to more effectively allocate the limited human and budgetary resources available to the high-risk individuals (i.e., higher-cost, earlier readmissions). Potential care transition gaps and targeted interventions can be derived from such models with a more profound impact on overall population management.

Existing dedicated efforts for accurately predicting 30-day risk of readmission are mostly focused on a specific cohort[2], such as congestive heart failure patients (Balla, Malnick, and Schattner 2008), cancer patients (Francis et al. 2015), emergency readmissions (Shadmi et al. 2015), etc. While these models are very useful, there is a lot of value in having *all-cause* risk and cost of readmission models that are not tied to a specific disease. In addition to allowing to derive risk and cost scores for patients who do not belong to any of the well studied cohorts, these models can also be used for incoming patients for which we do not know (yet) which cohort they belong to. To the best of our knowledge, none of the recent efforts predict cost or risk for all-cause readmissions, which is a completely different medical and data mining problem involving large, heterogeneous patient population sizes compared to disease specific cohorts such as heart failure.

In this study, we evaluate state-of-the-art machine learning techniques for predicting 30-day risk and cost on admission data of patients provided by a large hospital chain in the Northwestern U.S. We treat the risk prediction problem as a binary classification task, namely predicting whether the next admission of a given patient will be within 30 days or not. The LACE index is often used in clinical practice for this purpose (Zheng et al. 2015). This index considers

[1]http://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/Readmissions-Reduction-Program.html, accessed on Oct 22, 2015

[2]A sub-group of a given population with similar characteristics (e.g., medical conditions), such as s group of diabetes patients.

four numerical variables, namely length of stay (L), acuity level of admission (A), comorbidity condition (C), and use of emergency rooms (E). The LACE score of a patient is obtained by summing up the values of these four variables at the time of discharge. A threshold (usually ≥ 10) is then set to determine which patients are at "high" readmission risk (Zheng et al. 2015). We use LACE as a baseline to compare the performance of the machine learning algorithms we investigate in this paper. We find that the use of machine learning techniques allows to achieve higher sensitivity (recall) without penalizing the specificity and precision too much. On the cost prediction side, we find that the simple baseline strategy of forecasting that the next admission of a patient will cost as much as the average of his previous admissions works reasonably well. In addition, a substantially lower mean absolute error (MAE) can be achieved with M5 model trees.

The rest of the paper is organized as follows: after giving an overview of related work in Section 2, we formalize the risk and cost prediction problems in Section 3. The machine learning algorithms applied in this paper for risk and cost predictions are explained in Section 4. The dataset and features are described in Section 5. In Section 6 we discuss the performance of the algorithms. Finally, in Section 7 we conclude with our overall findings.

2 Related Work

In this section, we give a brief overview of research efforts done independently along each of the two dimensions: readmission risk prediction and healthcare cost prediction. To the best of our knowledge, there is no existing work that studies risk and cost prediction problems in a combined way.

Healthcare Cost Prediction

Previously proposed cost prediction models often used rule-based methods and linear regression models. A challenge with the rule-based methods (e.g. (Kronick et al. 2002)) is that they require substantial domain knowledge which is not easily available and is often expensive. Linear regression models on the other hand are challenged by the skewed nature of healthcare data. Healthcare cost data typically features a spike at zero, and a strongly skewed distribution with a heavy right-hand tail (Jones 2010). As a result, the prediction models are posed with the challenge of an extreme value situation. This phenomenon is also observed in the dataset used in this study (see Figure 1). Consequently, several advanced statistical methods (in-sample estimation) have been proposed to overcome the skewness issue, such as General Linear Models (GLM) (Manning, Basu, and Mullahy 2005), mixture models (Mullahy 1997), etc. For a comprehensive comparison of previously proposed statistical methods for healthcare cost prediction, we refer to the review paper (Mihaylova et al. 2011). The development of healthcare cost prediction models using machine learning methods has been more recent (e.g., (Lahiri and Agarwal 2014; Sushmita et al. 2015)). (Lahiri and Agarwal 2014) investigate classification algorithms to predict whether an individual is going to incur higher or lower healthcare expenditure. (Sushmita et al. 2015) use three machine learning algorithms for cost prediction – regression tree, M5 model tree and random forest, and observe improved performance when compared to traditional methods. In this paper, we also investigate these algorithms for the task of predicting cost of hospital readmission. To the best of our knowledge, their utility for predicting the costs of hospital readmissions specifically (as opposed to predicting general healthcare costs) has not been investigated before.

Hospital Readmission Prediction

In 2011, there were approximately 3.3 million adult 30-day all-cause hospital readmissions in the United States, and they were associated with about \$41.3 billion in hospital costs (Hines et al. 2011). Many of these hospitalizations are readmissions of the same patient within a short period of time. These readmissions act as a substantial contributor to rising healthcare costs (Jencks, Williams, and Coleman 2009). Readmission rates are also used as a screening tool for monitoring the quality of service and efficiency of care provided by healthcare providers (Balla, Malnick, and Schattner 2008). While predicting risk-of-readmission has been identified as one of the key problems for the healthcare domain, not many solutions are known to be effective (Krumholz et al. 2007; Ottenbacher et al. 2001). In fact, to improve the clinical process of heart failure patients for instance, healthcare organizations still leverage the proven best-practices, called "*Get With The Guidelines*" by the American Heart Association. In general, related work on risk-of-readmission prediction has primarily attempted to study cohort specific readmission risk, such as, heart failure, pneumonia, stroke, and asthma, but the effort of designing large scale machine learning algorithms for all-cause readmission is still at a rather rudimentary stage.

Despite several years of continued research efforts in modeling risk of readmission and healthcare cost, a dual predictive tool that utilizes healthcare data to predict risk and cost of hospital readmission has not been explored before. This study makes the first step in that research direction.

3 Problem Description

The goal of this study is to predict a patient's 30-day **risk** of hospital readmission and the associated **cost** of that readmission. We assume that the learning task at hand is a combination of a supervised classification problem (risk prediction) and a regression problem for predicting the cost (in dollars) of the readmission. The feature vector $X_i = (x_{i1}, x_{i2}, ..., x_{iM})$ of an instance i includes information about general demographics such as age and gender of the patient, as well as specific clinical and cost information at the time of discharge from the hospital. The goal is to produce an output vector $Y_i = (y_{i1}, y_{i2})$ consisting of a label y_{i1} that indicates whether the next admission of the patient will be in 30 days ("yes") or not ("no"), and the cost y_{i2} of the next admission. Let us use $\mathcal{X}$ to denote the set of all instances (feature vectors), and let $\mathcal{Y} = \{\text{yes}, \text{no}\} \times \mathbb{R}^+$ be the set of all dual labels. Given training examples of the form (X_i, Y_i) with $X_i \in \mathcal{X}$ and $Y_i \in \mathcal{Y}$, the aim is to

X (Input Vector)		Y (Output Vector)	
Admission		Next Admission	
Unique Identifier	Features (X)	Risk (y_1)	Cost (y_2)
id_1	$x_{11}, x_{12}, \ldots . x_{1M}$	yes	$45,132
id_1	$x_{21}, x_{22}, \ldots . x_{2M}$	yes	$41,305
id_1	$x_{31}, x_{32}, \ldots . x_{3M}$	no	$17,809
id_2	$x_{41}, x_{42}, \ldots . x_{4M}$	yes	$21,305
id_2	$x_{51}, x_{52}, \ldots . x_{5M}$	no	$55,809
id_3 **(test case)**	$x_{61}, x_{62}, \ldots . x_{6M}$	?	?

Table 1: Example input and output scenario for the risk and cost prediction task. Here, the first column indicates a unique identifier for each patient in the dataset.

learn a model $\mathcal{H} : \mathcal{X} \rightarrow \mathcal{Y}$ that can label new, unseen instances from $\mathcal{X}$ with a dual label from $\mathcal{Y}$ in an accurate way. We address this multi-label prediction learning problem in a manner similar to binary relevance (Tsoumakas and Katakis 2007), by learning a model for each label:

- **Risk of 30-Day Readmission – Classification Task:** For given training examples of the form (X_i, y_{i1}), where $X_i \in \mathcal{X}$ and $y_{i1} \in \{\text{yes, no}\}$, the goal is to learn a model $\mathcal{H}_1 : \mathcal{X} \rightarrow \{\text{yes, no}\}$ that accurately predicts whether the next admission of a patient will be within 30 days.
- **Cost of Readmission – Regression Task:** For given training examples of the form (X_i, y_{i2}), where $X_i \in \mathcal{X}$ and $y_{i2} \in \mathbb{R}^+$ (cost in dollars), the goal is to learn a model $\mathcal{H}_2 : \mathcal{X} \rightarrow \mathbb{R}^+$ that accurately predicts the cost of the next admission.

The combined model is then obtained as $\mathcal{H}(X) = (\mathcal{H}_1(X), \mathcal{H}_2(X))$, for X in $\mathcal{X}$.

We also tried other techniques for multi-label prediction like – Label Powerset (Cherman, Monard, and Metz 2011) and Chain Classifier (Read et al. 2009), but initial evaluation results with these methods were not as good as those obtained with the binary relevance approach, so we omit them from this paper.

Table 1 illustrates the input and output representations for the risk and cost prediction problem. Let us assume that the patient with id_1 was admitted to the hospital four times (say on Jan 14, Feb 2, Feb 28 and Apr 15). That means that he had three *re*admissions, two within 30 days (high risk) with cost being $45,132 and $41,305 respectively, and one after 30 days (low risk) with cost being $17,809. These response features are constructed based on attributes from the original, raw data (shown in Table 2). During the training phase, data of patient id_1 and id_2 will be used to train binary classifiers (to predict risk) and regression models (to predict cost); during the test phase the models will be used to predict the risk and the cost of patient id_3's next encounter.

We evaluate the accuracy of the learned models in several ways. Accuracy is traditionally measured as the percentage of instances that are classified correctly. It has been emphasized that the use of accuracy as an evaluation measure for data where there is an imbalance between positive and negative classes can yield misleading conclusions (Fatourechi et al. 2008; He and Garcia 2009). Readmission data is typically imbalanced. As Table 2 shows, approximately 27% of the admissions in our study are within 30 days (i.e. 27% positive instances), while the remaining 73% happen after 30 days (i.e., 73% negative instances). In addition to accuracy, we therefore also evaluate the binary classification models in terms of sensitivity (recall), specificity (true negative rate), and precision. Recall that sensitivity is TP/(TP + FN), specificity is TN/(TN+FP), and precision is TP/(TP+FP), with TP, FP, TN, and FN respectively denoting the number of true positives, false positives, true negatives and false negatives. The performance of the cost prediction algorithms is evaluated using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), with a lower error indicating a better performance.

4 Methods

In this section we give an overview of the machine learning algorithms used in this study. For risk-of-readmission prediction we used state-of-the-art classification techniques, while for cost prediction we used regression techniques.

Risk Prediction Methods

Support Vector Machine: A Support Vector Machine (SVM) is a statistical learning method for training classifiers based on different types of kernel functions – polynomial functions, radial basis functions, etc. An SVM learns a linear separating hyperplane by maximizing the margin between the classes (Drucker et al. 1996). The decision boundary is maximised with respect to the data points from each class (known as support vectors) that are closest to the decision boundary. For this study, we tested SVM with linear and radial kernel, and we report the results for radial in Table 3 because its performance was better than the linear kernel settings. We also tested for different regularization parameters ($C = 1, 5, 10, 15$), but the overall results did not change.

Logistic Regression: Logistic Regression is an example of a discriminative classifier that models the posterior $p(y_1|X)$ directly given the input features. That is, it learns to map the input (X) vector directly to the output class label y_1 (risk in our case). When the response is a binary (dichotomous) variable, logistic regression fits a logistic curve to the relationship between X and y_1 (Ng and Jordan 2001). The class decision for the given probability is then made based on a threshold value. The threshold is often set to 0.5, i.e. if $p(y_1|X) \geq 0.5$, then we predict that the next readmission of the patient will be within 30 days, and otherwise not. We tested with multiple threshold values to make the class decision.

Decision Trees: An alternative approach to linear classification is to partition the space into smaller regions, where the interactions are more manageable. Like for the other methods described in this section, the goal of a classification tree is to predict a response y_1 (risk in our case) from inputs X. This is done by growing a binary tree. At each internal node in the tree, a test is applied to one of the inputs, and depending on the outcome, the left or the right subbranch of the tree is then selected. Eventually a leaf node is reached, where the prediction is made. For this study, we used an implementation of the classification and regression tree algorithm (CART) (Breiman et al. 1984) in R. We tested the performance of classification trees using different complexity parameters (cp = 0.01, 0.001, 0.0005). In Table 3 we report the results of the best performing tree with cp set to 0.01 value.

Random Forest: Random forest regression is an ensemble learning method that operates by constructing a multitude of regression trees at training time and outputting the mean prediction of the individual trees for new observations. Each tree is constructed using a random sample of the observation and feature space from the original dataset. This has the effect of correcting the tendency of individual regression trees to overfit the training data (Breiman 2001).

Generalised Boosted Modeling (GBM): Boosting is an approach to machine learning based on the idea of creating a highly accurate predictive model ensemble by combining many relatively weak and inaccurate models (Freund and Schapire 1997). In other words, boosting is an optimization technique that minimizes the loss function by adding, at each step, a new model that best reduces the loss function. It is often used to grow an ensemble of classification trees, like we do in this paper. In this study we use the gbm implementation of AdaBoost in R, which is an implementation of extensions to Freund and Schapire's AdaBoost algorithm and Friedman's gradient boosting machine[3].

All the models are trained and tested using R[4]. Additionally, we also set the output of each model to be the class probability (prob= TRUE), instead of the class labels. This was done in order to test for different decision threshold values ($0.0 - 1.0$) to find the optimal balance between different evaluation measures. We report results for thresholds between $0.2 - 0.52$ in Figures 2-5.

Cost Prediction Methods

Linear Regression: Linear regression is used extensively in the literature on healthcare cost prediction, so, even though it has its limitations, it can not be ignored in this study. We use a linear regression model to predict cost using an M-dimensional vector of predictive variables (see Table 2).

M5 Model Tree: M5 model trees are a generalization of the CART model (Breiman et al. 1984). The structure of an M5 model tree follows that of a decision tree, but has multiple linear regression models at the leaf nodes, making the model a combination of piecewise linear functions. The algorithm for the training of a model tree breaks the input space of the training data through a recursive partitioning process similar to the one used in CART. After partitioning, linear regression models can be fit on the leaf nodes, making the resulting regression model locally accurate.

In addition to the linear regression and M5 Model Tree methods, **decision trees** and **generalised boosted modeling (GBM)** as described for risk prediction task were also used for predicting the cost.

5 Dataset and Features

The study in this paper includes admission data of patients provided by a large hospital chain in the Northwestern United States. Each admission record includes demographic information (e.g., gender, ethnic group), clinical information (e.g., primary diagnosis), care provider details, administrative data (e.g., length of stay) and billing information (e.g., charge amount). First, we performed data filtering as part of data pre-processing. Of the available ∼221K admission records, we excluded instances of admissions for which the patient died before discharge, or was transferred to another acute care facility within the hospital chain, or left against medical advice. Additionally, we excluded records where the next admission date is unknown, since they cannot be used to evaluate the correctness of cost and risk of readmission prediction. We also excluded hospitalizations with unspecified primary diagnosis.

Next, we performed several feature engineering steps. There were 214 features in the raw data. We used a forward stepwise regression approach (Derksen and Keselman 1992) to select a subset of this feature set. This subset is shown in Table 2. Most of the features from Table 2 correspond directly to features from the raw data; others have been constructed based on previous history of the patient. That is, most of the features are drawn from individual admission records, but some are aggregated across multiple admission records of the same patient. The features from the latter category are:

- **Number of Comorbidities:** this is the total number of unique comorbidities[5] that were registered for a patient up to the time of discharge. We used the Elixhauser comorbidity (Elixhauser et al. 1998) information of a patient to identify all comorbidities associated to that patient. Comorbidity is associated with worse health outcomes, increased healthcare costs and is known to impact prediction of risk of readmission (Donze et al. 2013). Therefore, we use it as one of the predictor variables.

- **Number of Existing conditions:** this is is the total number of unique diagnoses registered for this patient up to this point, including during previous admissions. The list of existing conditions of a patient is represented using ICD9-CM codes in the raw data (∼4K distinct values). We grouped the ICD9-CM codes using Clinical Classification Software (CCS)[6], and included the count of distinct CCS codes per patient as a feature.

- **Number of Previous Admissions:** this is the total number of hospital admissions registered for this patient up to this point. Here, the assumption is that a patient with a history of several hospital admissions is more likely to be readmitted again.

In this paper we use frequency counts (e.g. Number of Comorbidities) to overcome the limitation of significant sparseness in this dataset, for future research, we aim to explore statistical methods to overcome this limitation. Finally, we randomly sampled ∼10K instances with the feature set shown in Table 2 to train and test our models.

[3]http://cran.r-project.org/web/packages/gbm/gbm.pdf

[4]http://www.r-project.org

[5]Two or more coexisting medical conditions or disease processes that are additional to an initial diagnosis.

[6]https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp

Feature	Type	Distribution
Gender	Categorical	Female (5,818) Male (4,176)
Adult	Boolean	Yes (9,792) No (202)
Age $\geq$ 65	Boolean	Yes (4,801) No (5,193)
Ethnic Group	Categorical	Caucasian (8,303) African American (669) Hispanic/Latino (257) American Indian (185) Asian (172) Pacific Islander (157) Multi-Racial (83) Non-Hispanic (25) Middle Eastern (18) Eskimo (4) Other (121)
Marital Status	Categorical	Single (2,387) Married (4,148) Widowed (1,939) Divorced (1,084) Separated (211) Significant Other (192) Legally Separated (23) Domestic Partner (4) Other (2) Unknown (4)
Admit Type	Categorical	Emergency (7,350) Elective (2,519) Urgent (86) Trauma Center (39)
Financial Class	Categorical	Medicare (5,595) Medicaid (924) Self-pay (319) Other (3,156)
Care Type	Categorical	Acute (9,975) Geropsychiatric (19)
No. of Comorbidities	Numeric	Mean: 6.43 (SD = 4.25)
No. of Existing Conditions	Numeric	Mean: 10.67 (SD = 8.05)
Length of Stay (Days)	Numeric	Mean: 4.28 (SD = 4.41)
Same Day Discharge	Boolean	Yes (105) No (9,889)
Blood Pressure at Discharge	Categorical	80-89 or 120-139 (4,189) $<$80 and $<$ 120 (3,529) 90-99 or 140-159 (1,557) $>$ 99 or $>$ 159 (719)
No. of Previous Admissions	Numeric	Mean: 1.55 (SD = 2.87)
Next Admission $<$ 30 Days **(Response Variable)**	Categorical	Yes (2,697), 27% No (7,297), 73%
Additional Features for Cost Prediction		
Current Admit Cost ($)	Numeric	Mean: 53,530 (SD = 72,888)
Current Bed Charge ($)	Numeric	Mean: 10,120 (SD = 14,458)
Cost of Next Readmission ($) **(Response Variable)**	Numeric	Mean: 54,140 (SD = 74,400)

Table 2: Overview of the feature set used in the prediction of "risk" and "cost" of readmission

6 Result Analysis

We evaluated the machine learning methods from Section 4 for the risk and cost of hospital readmission prediction problems described in Section 3 on the dataset described in Section 5. In this section we present the results and discuss the key observations.

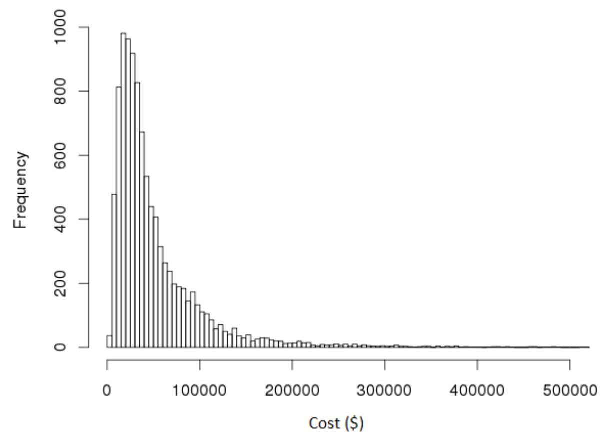

Figure 1: Distribution of hospital readmission cost in the readmission dataset

Risk Prediction

The results of the five machine learning algorithms as well as the LACE baseline, are presented in Figures 2-5 and Table 3. Among the existing risk prediction tools, the LACE index is regularly used in hospitals (Zheng et al. 2015). This index considers four numerical variables, namely length of stay (L), acuity level of admission (A), comorbidity condition (C), and use of emergency rooms (E). A LACE score is obtained by summing up the values of these four variables. A threshold (usually $\geq$ 10) is then set to determine patients with "high" readmission risk (Zheng et al. 2015). We use LACE as a baseline to compare the performance of the machine learning algorithms we investigate in this paper.

We evaluated the models developed with all five machine learning methods using 10-fold cross-validation across different threshold values (see Figures 2-5). This was done so that a threshold value that would give the highest possible sensitivity, but at the same time also have comparable specificity to that of the LACE tool can be identified. It should be noted that for the 30-day risk of readmission prediction task, higher sensitivity is more desirable. This is because correctly identifying the "high risk" patients who are likely to be readmitted within 30 days is more crucial than correctly identifying low risk patients (discussed in Section 1). Overall results corresponding to the best threshold values for all the models are shown in Table 3, and Figure 6 shows the trade-off between sensitivity and specificity for all the models.

There are three key observations to be made from these results. First, the results in Table 3 suggest that most machine learning methods show promising results when compared to the baseline LACE method. Not only was it possible to achieve higher sensitivity than LACE, but this was done without penalizing the specificity and precision too much. More in detail, with 3 out of the 5 machine learning methods we achieved a sensitivity of over 80%, while the results for specificity and precision remained comparable to that of LACE (sensitivity = 76%). The sensitivity results for the other two methods, namely SVM and decision tree, were

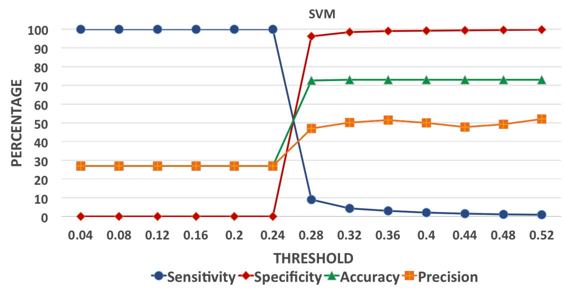

Figure 2: Risk prediction performance results of SVM

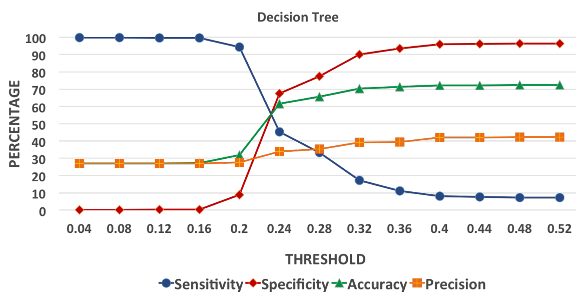

Figure 3: Risk prediction performance results of Decision Trees

also very high (sensitivity $\geq 94\%$) and the precision score was comparable, but the proportion of "low risk" instances which were correctly identified was very low (specificity $\leq 9\%$).

Second, the rate of change in sensitivity and specificity slightly differs across different machine learning methods. For instance, as the threshold values increase, the sensitivity and specificity in the decision trees, logistic regression, and generalised boosted models exhibit sigmoid curves (see Figure 3, and 5), characterized by a small progression in the beginning and then accelerating and converging over larger threshold values. For random forest models, the change is almost linear (Figure 4). For SVM a steep drop in sensitivity and rise in specificity is observed between 0.24 to 0.28 threshold values (Figure 2). Further investigation of the SVM results showed that there was big increase in the number of true negatives and false negatives around these threshold values, illustrating that SVM is less robust than the other methods, and that its good performance depends on fine tuning of the cutoff threshold.

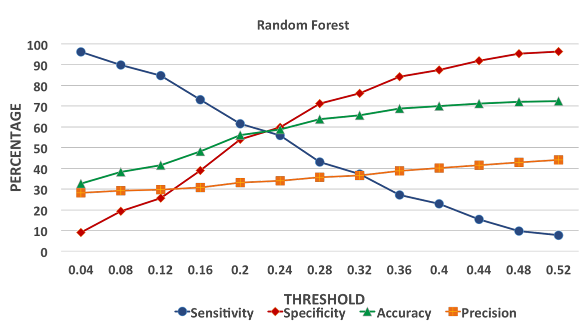

Figure 4: Risk prediction performance results of Random Forest

Algorithm	**Sensitivity (%)**	**Specificity (%)**	**Precision (%)**
LACE	76.42	38.95	31.63
SVM	98.11	1.84	26.98
Decision Trees	94.07	9.04	27.65
Random Forest	84.76	25.60	29.63
Logistic Regression	92.47	13.24	28.26
GBM	90.43	18.24	29.02

Table 3: Performance comparison of different machine learning methods for the task of predicting whether the next hospital admission of a patient will be within 30 days. The results are based on 10-fold cross-validation.

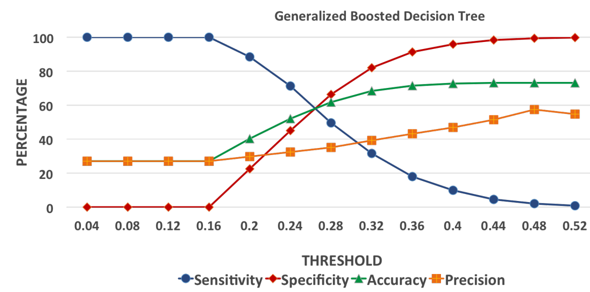

Figure 5: Risk prediction performance results of GBM

The third key observation is that, across all methods, 50-60% is the maximum score that can be achieved when a "perfect" balance across all measures (sensitivity, specificity, accuracy and precision) is desired. This is a meaningful result because it shows that the machine learning methods can give good performance for the majority ($> 50\%$) of instances across all measures. It is interesting to observe in Figures 2-5 that this optimal point of balance emerges within the same small range of threshold values across all different machine learning methods.

Overall, for the risk prediction task, the results for most machine learning methods for any type of readmission ("all-cause") are promising when compared to a standardized risk prediction tool (LACE). It was possible to achieve higher sensitivity (recall) without penalizing the specificity and precision too much. Improving the precision and specificity further will be a task to explore in future.

Cost Prediction

We measured the performance of the methods for cost prediction using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). A lower error indicates that the predicted dollar amount is closer to the actual cost. As for risk prediction, we evaluated all models using 10-fold cross-validation. An overview of MAE and RMSE results is presented in Table 4. We compared the results of four machine

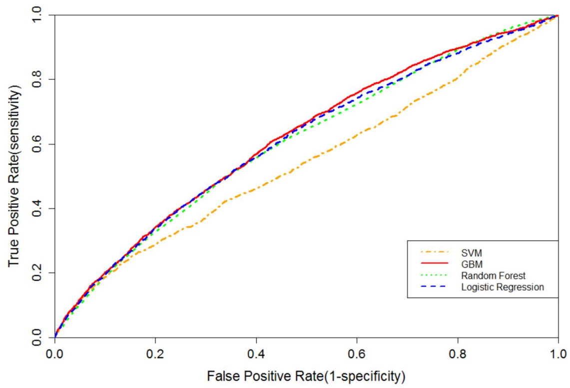

Figure 6: ROC curve comparing risk prediction performance results. It shows the trade-off between sensitivity and specificity. It can be seen that GBM is the best classifer, while SVM is the worst.

learning methods, namely linear regression, M5 model tree, generalised boosted model and decision tree, with those of two baseline methods:

- **Average Baseline (AB):** the Average Baseline (AB) measure is the overall mean cost μ of individual average encounter costs for all the beneficiaries within the training set prior to the current encounter for which we are predicting the cost. This mean (μ) score is then used as the baseline predicted cost for all patient-encounter pairs in the test set.
- **Current Admit Cost (CAC):** the Current Admit Cost (CAC) baseline model is a linear regression model fitted using only the current admission cost during the training period as a predictor variable, with next readmission cost being the response variable. Note that the difference between this current admit cost baseline and the competing linear regression model is that all features (as shown Table 2) from the readmission dataset were used to train the linear regression model, while only the 'current admit cost' variable was used in the CAC baseline model.

Algorithm	MAE ($)	RMSE ($)
Average Baseline (AB)	21,609	27,176
Current Admit Cost (CAC)	20,882	26,458
Linear Regression (LM)	20,232	26,124
M5 Model Tree (M5)	18,263	24,824
Generalised Boosted Model (GBM)	20,065	26,388
Decision Tree (DT) (cp=0.01)	20,512	26,328

Table 4: Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) in dollars for the cost prediction task

As can be seen in Table 4, for all-cause readmissions, our data mining models exhibit lower prediction error compared to the Average Baseline (AB) method in terms of both MAE and RMSE. Within that, M5 model tree has the lowest prediction error. The errors for the Current Admit Cost (CAC)

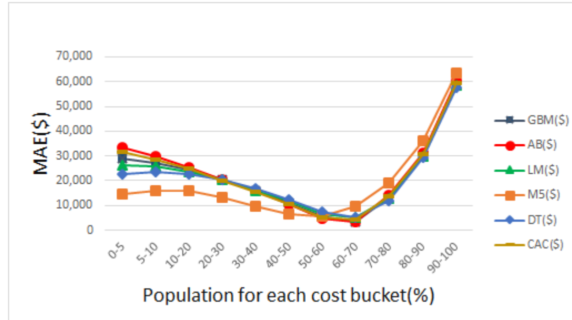

Figure 7: Comparison of average Mean Absolute Error (MAE) across different cost buckets.

baseline method were interestingly enough comparable to the errors of several of the more sophisticated methods.

Overall, two key observations can be made from the performance results shown in Table 4. First, current admit cost and average cost are strong baseline models, and therefore current hospital admission cost or average cost alone can be a good indicator for the next readmission cost provided it is available to the care provider. Second, among all machine learning algorithms, M5 model tree performed best and achieved a substantially lower MAE than strong baseline models for predicting next readmission cost.

The knowledge that the cost distribution in our dataset is highly skewed (see Figure 1), as is known to be the case with healthcare costs in general, inspired us to delve deeper into investigating for which fraction of the patients our models can predict costs with error margins that are reasonably bounded. To this end, we divided the population into 11 different cost buckets, shown on the horizontal axis in Figure 7. The cost buckets range from the 5% lowest cost patients (subpopulation 0-5%) to the 10% highest cost patients (subpopulation 90-100%). Next, for each of our cost prediction methods, we measured the average MAE over all patients within each subpopulation. The results are shown in Figure 7.

It is interesting to observe that all methods display a similar behavior: the predictions across all models are most accurate within the middle of the range, i.e. for moderate cost patients (40-70%). For the low cost patients (0-40%), the machine learning techniques clearly outperform the baseline models. This is especially the case for the M5 model tree. For the high-cost patients (70-100%) it is interestingly enough the other way around, although the difference in error between the different techniques is relatively small compared to the size of the actual healthcare costs in this case. Still, the results in Figure 7 indicate that it might be beneficial to train a hierarchical model that first predicts a cost bucket and then uses a model trained specifically for that cost bucket to arrive at a final prediction in dollars.

7 Conclusion

The rate of hospital readmissions of patients is a key measure that is tracked for numerous reasons. Consequently, risk stratification of a population and readmission models are becoming increasingly popular. Recent data mining efforts either predict healthcare costs or risk of hospital readmission, but not both. The goal of this study was a dual predictive modeling effort that utilizes healthcare data to predict the risk and cost of any hospital readmission ("all-cause"). For this purpose, we explored machine learning algorithms to do accurate predictions for risk and cost of 30-day readmission. For the task of risk prediction, results for most machine learning methods for any type of readmission ("all-cause") were promising when compared to a standardized risk prediction tool (LACE). It was possible to achieve higher sensitivity (recall) without penalizing the specificity and precision too much. On the cost prediction side, two key observations were made from the performance results of the machine learning methods. First, average admission cost and current admission cost are strong predictors, and therefore they alone can be a good indicator for the next readmission cost. Second, among the four machine learning algorithms, M5 model tree consistently performed better and achieved a substantially lower MAE than strong baseline models for predicting next readmission cost.

Acknowledgments

The authors would like to thank MultiCare[7] for providing access to the readmission dataset used in this study. This research was made possible thanks to generous support of Edifecs Inc. and a Microsoft Azure for Research grant.

References

Balla, U.; Malnick, S.; and Schattner, A. 2008. Early readmissions to the department of medicine as a screening tool for monitoring quality of care problems. *Medicine* 87(5):294–300.

Breiman, L.; Friedman, J. H.; Olshen, R. A.; and Stone, C. J. 1984. *Classification and Regression Trees*. Wadsworth Publishing Company.

Breiman, L. 2001. Random forests. *Machine Learning* 45(1):5–32.

Cherman, E. A.; Monard, M. C.; and Metz, J. 2011. Multi-label problem transformation methods: a case study. *CLEI Electron. J.* 14(1).

Derksen, S., and Keselman, H. J. 1992. Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology* 45(2):265–282.

Donze, J.; Aujesky, D.; Williams, D.; and Schnipper, J. L. 2013. Potentially avoidable 30-day hospital readmissions in medical patients: derivation and validation of a prediction model. *JAMA Internal Medicine* 173(8):632–638.

Drucker, H.; Burges, C. J. C.; Kaufman, L.; Smola, A. J.; and Vapnik, V. 1996. Support vector regression machines. In *Advances in Neural Information Processing Systems 9, Proceedings of the 1996 NIPS conference*, 155–161.

[7]http://www.multicare.org

Elixhauser, A.; Steiner, C.; Harris, D. R.; and Coffey, R. M. 1998. Comorbidity measures for use with administrative data. *Medical care* 36(1):8–27.

Fatourechi, M.; Ward, R. K.; Mason, S. G.; Huggins, J.; Schlögl, A.; and Birch, G. E. 2008. Comparison of evaluation metrics in classification applications with imbalanced datasets. In *7th International Conference on Machine Learning and Applications (ICMLA)*, 777–782.

Francis, N. K.; Mason, J.; Salib, E.; Allanby, L.; Messenger, D.; Allison, A. S.; Smart, N. J.; and Ockrim, J. B. 2015. Factors predicting 30-day readmission after laparoscopic colorectal cancer surgery within an enhanced recovery programme. *Colorectal Disease* 17(7):148–154.

Freund, Y., and Schapire, R. E. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer System Science* 55(1):119–139.

He, H., and Garcia, E. A. 2009. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering* 21(9):1263–1284.

Hines, A. L.; Barrett, M. L.; Jiang, J.; and Steiner, C. A. 2011. Conditions with the largest number of adult hospital readmissions by payer. Technical report, Healthcare Cost and Utilization Project (HCUP).

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

Jones, A. 2010. Models For Health Care. Technical report, HEDG, c/o Department of Economics, University of York.

Kronick, R.; Gilmer, T. P.; Dreyfus, T.; and Ganiats, T. G. 2002. CDPS-Medicare: The chronic illness and disability payment system modified to predict expenditures for Medicare beneficiaries. Technical report.

Krumholz, H. M.; Normand, S. L. T.; Keenan, P. S.; Lin, Z. Q.; Drye, E. E.; Bhat, K. R.; Wang, Y. F.; Ross, J. S.; Schuur, J. D.; and Stauffer, B. D. 2007. *Hospital 30-day heart failure readmission measure methodology. Report prepared for the Centers for Medicare & Medicaid Services*.

Lahiri, C. B., and Agarwal, N. 2014. Predicting healthcare expenditure increase for an individual from Medicare data. In *Proceedings of the ACM SIGKDD Workshop on Health Informatics*.

Manning, W. G.; Basu, A.; and Mullahy, J. 2005. Generalized modeling approaches to risk adjustment of skewed outcomes data. *Journal of Health Economics* 24(3):465–488.

Mihaylova, B.; Briggs, A.; O'Hagan, A.; and Thompson, S. G. 2011. Review of statistical methods for analysing healthcare resources and costs. *Health Economics* 20(8):897–916.

Mullahy, J. 1997. Heterogeneity, excess zeros, and the structure of count data models. *Journal of Applied Econometrics* 12(3):337–50.

Ng, A. Y., and Jordan, M. I. 2001. On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. In *Advances in Neural Information Processing Systems 14, Proceedings of the 2001 NIPS conference*, 841–848. MIT Press.

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

Read, J.; Pfahringer, B.; Holmes, G.; and Frank, E. 2009. Classifier chains for multi-label classification. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases: Part II*, 254–269.

Shadmi, E.; Flaks-Manov, N.; Hoshen, M.; Goldman, O.; Bitterman, H.; and Balicer, R. D. 2015. Predicting 30-day readmissions with preadmission electronic health record data. *Med Care* 53(3):283–289.

Sushmita, S.; Newman, S.; Marquardt, J.; Ram, P.; Prasad, V.; De Cock, M.; and Teredesai, A. 2015. Population cost prediction on public healthcare datasets. In *Proceedings of ACM Digital Health 2015 (5th International Conference on Digital Health)*, 87–94.

Tsoumakas, G., and Katakis, I. 2007. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining* 3(1):1–13.

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