PREDICTING HOSPITAL READMISSIONS IN THE MEDICARE POPULATION

SAJID ZAIDI*

1 INTRODUCTION

Avoidable hospital readmissions cost taxpayers billions of dollars each year. The Medicare Payment Advisory Commission has estimated that almost $12 billion is spent annually by Medicare on potentially preventable readmissions within 30 days of a patient's discharge from a hospital [1]. The Medicare program has begun to apply financial penalties to hospitals that have excessive risk-adjusted readmission rates. There is much interest in the health policy and medical communities in the ability to accurately predict which patients are at high risk of being readmitted. Not only are there strong financial reasons to avoid readmissions, readmission to the hospital can be a sign of poor clinical care and can indicate a worsening of a patient's condition [2]. If doctors and nurses were aware of which patients were at highest risk, they could focus their efforts on these patients and could improve coordination of care with post-acute providers and family physicians.

There has been some interest in this problem in the machine learning community as well. The Heritage Health Competition was a predictive modeling competition with the objective of predicting hospital readmissions, with a $3 million cash prize. However, the dataset used for that competition was highly de-identified and thus was missing much of the key information useful for predictions. It also had a low number of patients who were generally healthy[1]. In this paper, I will apply machine learning methods to a dataset of Medicare claims to predict which patients are at a high risk of being readmitted to the hospital. I will then compare my results to the performance of risk adjustment models currently used by the Medicare program to predict readmissions.

2 DATASET

I use a dataset consisting of all Medicare claims for the year 2013. Medicare is the government health insurance program for seniors over 65 years of age, and a claim is evidence of a health care service that contains information on

* *szaidi@stanford.edu*

1 This is likely why no team was able to surpass the performance threshold to claim the grand prize

diagnoses and procedures performed on a given date. I take all inpatient hospital claims during 2013, and construct indicators for whether or not the patient was readmitted to any hospital within 30 days (this is the same definition that Medicare uses for its financial penalties). Each observation is thus a patient's hospital stay. This dataset has 5,719,330 observations. However, this is too large for the computing resources I have available, so I took a 5% random sample, resulting in a dataset of 285, 967 observations. 13.2% of the observations have a readmission.

3 FEATURES AND PREPROCESSING

This is a classification problem, where the positive class is the 13.2% of patients who are readmitted. I construct features using the Hierarchical Condition Model (HCC model). This is a standard classification of diagnoses and illnesses used in medical research[3]. For each patient, I take all the claims occurring in the 6 months prior to the hospital admission and use these claims to construct binary variables that indicate whether a patient has a given condition. I also include demographic data such as the patient's age and gender, and whether the patient is enrolled in Medicaid[2] or is institutionalized in a nursing home. It is important to note that all the features I construct use information available at the time of admission to the hospital, so the model could be used to make real time predictions while the patient is in her initial hospital stay. After this preprocessing, I have 95 features.

I reserve a random 20% sample of the data as a pristine test set. I will only use this test set at the end to compare my model's performance to that of models in the literature. I split the data using stratified random sampling, to ensure that the training and test sets have the same proportion of positive classes.

4 MODELS

I will use five different machine learning algorithms. All these models will be trained on the 80% training set.

LOGISTIC: The first is L2-regularized logistic regression. I will use 10-fold cross validation in order to determine the penalty parameter.

GBM: The second algorithm I will use is gradient boosting with logistic regression. I tune some of the hyperparameters (such as the number of models in the ensemble) using 3-fold cross validation.

RANDOMFOREST: The third algorithm I will use is the random forest algorithm[4]. This is an ensemble learner that constructs a series of decision trees and averages the results.

SVM: The fourth algorithm I will use is L2-regularized SVM with a linear kernel. The data is too large to use a kernel on my computing hardware, and since almost all my features are binary indicators, a kernel may be less useful in any case. To determine the C penalty parameter, I use the heuristic C function available in the LiblineaR package.

ENSEMBLE: Finally, I will construct my own ensemble learner. I will use logistic regression to combine the predictions of the previous four models.

2 A health care program for the poor

The features in this logistic regression are the predictions output by the other four models.

I will report three performance metrics: F1 score, area under the receiver operating characteric curve (AUROC), and the classification error. AUROC will not be calculated for SVM since it is not probabilistic. For the other 3 methods, to determine classification for the F1 score and classification error, for each model I will use the probability threshold that maximizes the F1 score on the training dataset. In other words, I will not use a 0.5 probability threshold to predict a positive class, but rather I will do a grid search to find the threshold that maximizes F1 on the training set. Of course, the same threshold will be used for prediction on the test set. I use this method because I have skewed classes, and F1 score is the objective I want to maximize.

5 RESULTS AND DISCUSSION

The following table presents the results from these algorithms.

Table 1: Performance of Learning Algorithms

Model	Training F1	Test F1	Training AUROC	Test AUROC	Training Error	Test Error
GBM	0.2883	0.2826	0.6332	0.6262	0.3026	0.3059
Random Forest	0.5143	0.1989	0.6825	0.6021	0.0942	0.1771
L2-Regularized Logistic	0.2867	0.2802	0.6323	0.6256	0.3056	0.3086
SVM	0.0011	0.0003			0.1325	0.1326
Ensemble	0.5551	0.1785	0.7211	0.5146	0.0881	0.1849

In general, these models have performed poorly. GBM, Logistic regression, and SVM do not seem to have overfit the data, since their test set performance is almost exactly the same as their training set performance. This indicates that these models have high bias, and future efforts should focus on feature engineering. Random forest, on the other hand, seems to have overfit the training data by quite a bit, as evidenced by the difference in performance between the training and test sets. In the future, I would try lowering the number of decision trees used by the algorithm, and perhaps determine that hyperparameter through cross-validation. The ensemble learner also overfit, but that is certainly due to the random forest predictions.

One notices that the classification error for logistic regression and gradient boosted logistic regression is around 30%, even though positive values are only 13% of the data. This may seem strange, but this occurs because I did not use a probability threshold of 0.5. Instead, I used the threshold that maximized F1 score, which happens to be around 0.14. In effect, I lowered the probability threshold to increase recall, at the cost of somewhat lowering the precision. This has the effect of worsening the classification error. However, I believe F1 score is more useful in the case of skewed classes, so I think the tradeoff is worth making.

Comparing my results to the literature, Horwitz et al.[5] developed the official readmissions prediction model that is used by Medicare to determine each hospital's financial penalty (Medicare uses the model to calculate each hospital's excess readmission rate over and above the rate predicted by the

model). They developed 5 different models, for 5 different medical conditions, and they report area under the ROC curve ranging from 0.63 to 0.67. My results are very close to that range, albeit the bottom of the range. They do not report any other performance metrics.

6 CONCLUSIONS AND FUTURE WORK

There is a lot of work that can be done in the future. Based on the lack of overfitting for most of my models, it seems the most promising avenue is to construct more features using the claims data, such as indicators for prior hospitalizations or more fine grained diagnosis features. The claims data is very rich, and there are enormous possibilities for the features that could be constructed. If clinical notes and text data from Electronic Medical Records were added to the mix, the feature space could become very large. Since this is an area with an enormous amount of data and a vast feature space, new approaches such as deep learning may be valuable.

Finally, the fact that my models have performed poorly, and yet still come very close to the performance of the official Medicare model, raises questions about the accuracy of the prediction models used by the federal government. These models are used to calculate over $1 billion in financial penalties to American hospitals, so this is certainly an area of immense policy importance where machine learning experts can contribute a lot. Up to this point, most work in the health policy community has focused on using single logistic regressions. If the entire toolbox of machine learning were applied to this problem, I am sure that we could achieve far better performance than the current state of the art. I hope this paper has contributed to that effort.

REFERENCES

[1] MedPAC. Promoting greater efficiency in medicare, june 2007 report to congress.. *www.medpac.gov*, 2007.

[2] Mohsen Bayati, Mark Braverman, Michael Gillam, Karen Mack, George Ruiz, Mark Smith, and Eric Horvitz. Data-driven decisions for reducing readmissions for heart failure: General methodology and case study. *PLoS: One*, 2014.

[3] Gregory Pope et al. Evaluation of the cms-hcc risk adjustment model. *Centers for Medicare and Medicaid Services*, 2011.

[4] Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002.

[5] Leora Horwitz et al. Hospital-wide all-cause unplanned readmission measure. final technical report. *Centers for Medicare and Medicaid Services.*, 2012.