Predicting Diabetes Related Readmission

Robert P. Yerex, Ph.D

December, 2014

UVA Medical Center Accountable Care Organization

**Objectives**. Develop of a statistical learning algorithm to predict readmission of patients for diabetes related complications within 30 days of discharge.

**Data Sources.** 70,000 patient records from the *Health Facts* [[1]](#footnote-1) database

**Statistical Learning / Predictive Model.** A range of tree/forest based algorithms, including Random Forest, Ada Boost, as well as several multidimensional transformation algorithms such as SVM, and NN, were applied

**Model Performance.** The final model, an ADABoost tree model, achieved 95% accuracy with a sensitivity of 82%, and a specificity of 97%

# Background

In fiscal 2011, for Medicaid patients, the three conditions with the largest number of 30-day all-cause readmissions were mood disorders (41,600 readmissions), schizophrenia (35,800 readmissions), and diabetes (23,700 readmissions). These conditions resulted in about $839 million in incremental hospitalization costs[1]. Meanwhile the rate increase in the occurrence of diabetes (type II) continues to accelerate [2] such that it is likely that within a decade or so, the cost of diabetes related readmission will eclipse mood disorders and schizophrenia. Given these trends, predictive models that can differentiate the risk of readmission among diabetic patients would be of great value. Strack et. al. [3] studied the relationship between HbA1c Measurement and hospital readmission rate. While this study did not develop a predictive model, it did illustrate the usefulness of a well curated data set that I am making use of in order to test various algorithm and metric development approaches for predicting diabetic related readmissions.

# Data

The data set consisted of approximately 70,000 patient records from the *Health Facts* [[2]](#footnote-2) database, a national data warehouse of clinical records gathered from US hospitals that use the Cerner Electronic Health Record System. A simplified data dictionary of the original data set is presented in Appendix A. An in-depth description of the curation process can be found in Strack et. al. [3].

# Modeling

Target outcome for all models was the binary variable 30 day readmission (TRUE / FALSE) which was derived from the *readmitted* variable of the original data set. This variable was transformed from {‘<30’, ‘>30’, ‘No’} to {TRUE, FALSE}. In addition several transformations were applied to the original data including:

* Recoding of *age* from a set of nominal non-overlapping ranges to an ordinal integer variable: {[0,10), [10,20),…,[100,110)} to {10, 20,…,110}
* Recoding of *weight* from a set of nominal non-overlapping ranges to an ordinal integer variable: {[0,25), [25,50),…,[200,225)} to {25, 50,…,225}

## Algorithm Selection / Feature Development

Given the combination of continuous, binary, categorical, and ordinal data types, initial focus was on tree/forest based algorithms, including Random Forest, Ada Boost, and multidimensional transformation algorithms such as SVM, and NN.

## Model 1: Ada Boost

The first complete model makes use of an AdaBoost algorithm, the VIMP chart is shown in Figure 1. Cross validation diagnostic statistics are shown in Table 1.

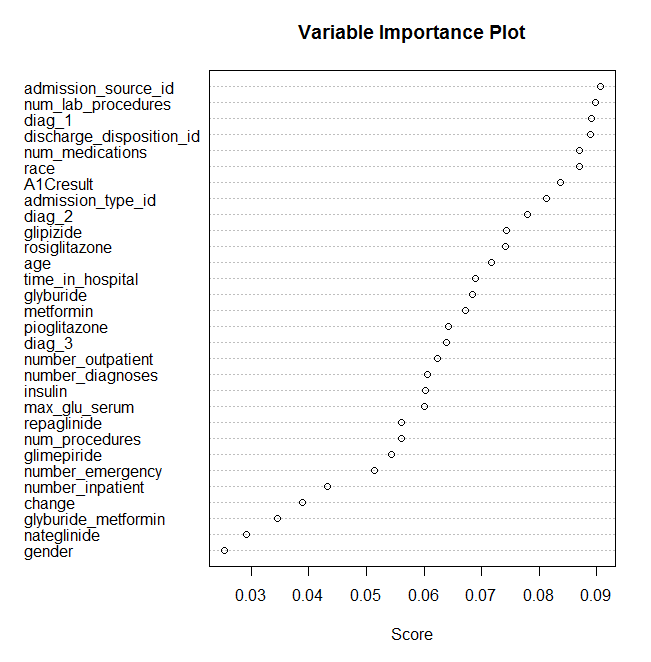


Figure 1: Model 1 - VIMP

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |
|  | Predicted | | |  | Prevalence | 12% |
| Actual |  | FALSE | TRUE |  | Sensitivity | 50% |
| FALSE | 0.88 | 0.01 |  | Specificity | 99% |
| TRUE | 0.06 | 0.06 |  | + Pred | 86% |
|  |  |  |  |  | - Pred | 94% |
|  |  |  |  |  | Accuracy | 93% |

Table 1: Model 1 - Accuracy

## Model 2

In the first iteration, each of the diagnoses was handled as a separate feature. There is likely information loss in this approach as certain combinations of diagnoses are likely more closely associated with 30 day readmission than others. There are a number of ways to group multiple features into a single constructed feature. The simplest is to determine all possible unordered combinations and code accordingly. For the model 2, a number of inconsequential variables were removed and the three separate diagnoses codes were replaced with a single code representing the index of the unordered combination.

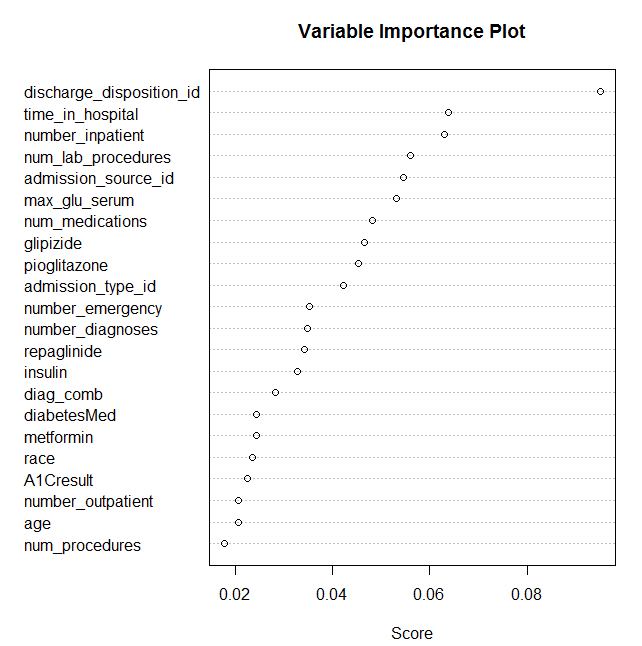


Figure 2: Model 2 – VIMP

The update VIMP chart is shown in Figure 2. Note that a number of the remaining features moved up or down the list. This is related to the way in which VIMP is calculated and the presence of interactions among the features. The updated diagnostics are shown in Table 2

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Predicted | | |  | Prevalence | 11% |
| Actual |  | FALSE | TRUE |  | Sensitivity | 55% |
| FALSE | 0.88 | 0.01 |  | Specificity | 99% |
| TRUE | 0.05 | 0.06 |  | + Pred | 86% |
|  |  |  |  | 1.000 | - Pred | 95% |
|  |  |  |  |  | Accuracy | 94% |

Table 2: Model 2 - Accuracy

## Model 3

With the understanding that the order of the diagnoses does matter in that the ICD9 Code recorded in the diag\_1 field is the primary diagnosis whereas diag\_3 is a secondary diagnosis. One way to code combinations where order does matter is to use permutations. However this is a constrained permutations with the value for diag\_1 being “anchored”.

1. An undirected acyclic graph of each patients three diagnoses codes was generated, with diag\_1 as the primary vertex.
2. An algorithm for storing such graphs in a common format was used to generate a hash value as the index code for the combination. This hash value was used as the coded value for the feature *diag\_*perm.

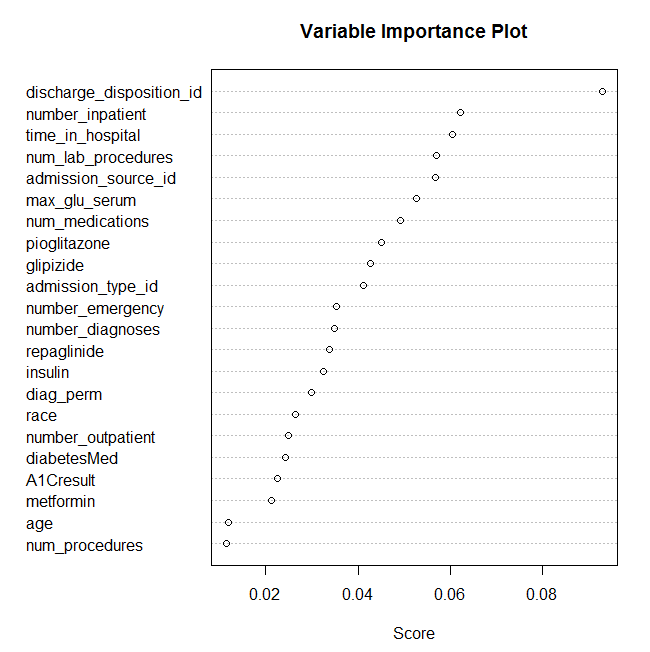


Figure 3: Model 3 - VIMP

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Predicted | | |  | Prevalence | 11% |
| Actual |  | FALSE | TRUE |  | Sensitivity | 64% |
| FALSE | 0.88 | 0.01 |  | Specificity | 99% |
| TRUE | 0.04 | 0.07 |  | + Pred | 88% |
|  |  |  |  |  | - Pred | 96% |
|  |  |  |  |  | Accuracy | 95% |

Table 3: Model 3 - Accuracy



Figure 4: Positive Predictive Accuracy Parameter Space

Cross validation was performed using the Bootstrap .632 Method [4]. From these results, shown in Figure 4 (this is actually positive predictive value, not overall accuracy), we can see that a maximum tree depth of five performed best. As far as the number of trees used, 50 and 200 performed comparably. It should also be noted that this was using an equally weighted error penalization matrix.

In terms of accuracy, shown in Table 4, the final model was no more accurate overall. This is in part due to a reweighting of the error cost matrix where a false negative was more strongly penalized than a false positive. This also impacted the measures of positive and negative predictive strength. It must be noted that when tested on the full data set (earlier models were tested on a subset of the data; the prevalence of the underlying true outcome (30 day readmission) was lower, 10.6% vs. 11.4%.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Predicted | | |  | Prevalence | 11% |
| Actual |  | FALSE | TRUE |  | Sensitivity | 82% |
| FALSE | 0.86 | 0.03 |  | Specificity | 97% |
| TRUE | 0.02 | 0.09 |  | + Pred | 75% |
|  |  |  |  |  | - Pred | 98% |
|  |  |  |  |  | Accuracy | 95% |

Table 4: Final Model Accuracy

# Citations

[1] A. Hines L., M. Barnett L., J. Jiang H., and C. Steiner A., “Conditions With the Largest Number of Adult Hospital Readmissions by Payer, 2011,” Healthcare Cost and Utilization Project, Statistical Brief #!72, Apr. 2014.

[2] E. W. Gregg, X. Zhuo, Y. J. Cheng, A. L. Albright, K. M. V. Narayan, and T. J. Thompson, “Trends in lifetime risk and years of life lost due to diabetes in the USA, 1985–2011: a modelling study,” *Lancet Diabetes Endocrinol.*, vol. 2, no. 11, pp. 867–874, Nov. 2014.

[3] B. Strack, J. P. DeShazo, C. Gennings, J. L. Olmo, S. Ventura, K. J. Cios, and J. N. Clore, “Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records,” *BioMed Res. Int.*, vol. 2014, p. e781670, Apr. 2014.

[4] B. Efron and R. Tibshirani, “Improvements on Cross-Validation: The .632+ Bootstrap Method,” *J. Am. Stat. Assoc.*, vol. 92, no. 438, pp. 548–560, Jun. 1997.

# Appendix A: Data Dictionary



1. Cerner Corporation, Kansas City, MO [↑](#footnote-ref-1)
2. Cerner Corporation, Kansas City, MO [↑](#footnote-ref-2)