Predicting Hospital Readmission for Diabetes Patients by Classical Approaches and Machine Learning Approaches

September 1, 2017

Student CO-PIs:

Chathurangi Heshani Pathiravasan
PhD Candidate,
Department of Mathematics,
Southern Illinois University,
1263 Lincoln Dr,
Carbondale, IL 62901
Project advisor's name & signature: Prof. David Olive

PhD advisor's name & signature: Prof. Bhaskar Bhattacharya

Gabrielle LaRosa
Undergraduate,
Mathematical Biology and Statistics,
University of Pittsburgh,
3725 Sutherland Drive,
Pittsburgh, PA, 15213
Advisor's name: Prof. G. Bard Ermentrout

Justine Stiftel
Undergraduate,
Bryn Mawr College,
101 N Merion Ave,
Bryn Mawr, PA 19010

R. W. M. A. Madushani
Post-Doctoral Associate (OPS),
Department of Medicine,
University of Florida,
1600 SW Archer Road,
Gainesville, FL 32610

Project Type: Multi-institution/ multi-disciplinary project

Total Funds Requested: $6000

1 Problem Statement

Hospital readmission is very expensive and it is an emerging indicator of quality of health care. On average, Medicare spends double the amount for an episode with one readmission. Moreover, diabetes is a chronic disease affecting people of all ages and is prevalent in a large amount of people in the US. It has been estimated that about 366 million people world wide have diabetes, and this number is likely to increase to 552 million by the year 2030 [4]. Hospital readmission is a major concern in diabetes care; over 250 million dollars was spent on treatment of readmitted diabetic patients in 2011. The recent cost analysis suggests that $252.76 million can be saved across 98,053 diabetic patient encounters by incorporating the cost sensitive analysis models [1]. Thus examining the historical patterns of diabetics care is very essential which might lead to improvements in patient safety and prevent future readmissions. This not only improves the quality of care but also reduces the medical expenses on readmission. There are a few studies on readmission rates among individuals with diabetes. It has been suggested that greater attention to diabetes reflected in HbA1c determination may improve patient outcomes and lower cost of inpatient care [7]. Different tools and models were build to predict the risk of all-cause readmission within 30 days among hospitalized patients with diabetes [6], [8]. In order to provide a valid assessment and to find future directions which might lead to improvements in patient safety, we must determine all the contributing factors for predicting readmission of diabetes patients. Indeed, we must closely study classical and machine learning approaches for predictive models and investigate the relationship of readmission rates with the predictors.

2 Proposed Activity

The data was obtained from the center for Machine Learning and Intelligence Systems at the University of California over a period of 10 years, from 1999 to 2008. It contains about 100,000 instances and 55 attributes such as patient number, race, gender, age, admission type, time in hospital, age etc. The main response variable is inpatient readmission which is a nominal variable. Originally it had three categories. We can define the readmission attribute as having two values for the main analysis: "readmitted", if the patient was readmitted with 30 days (short-term) and readmission after 30 days (long term) together, "otherwise" if no readmission at all. For other short term analysis or comparison of short term vs long term can be done by defining the readmission attribute in a proper manner. Non-relevant features can be left out as their presence does not affect the analysis, such as Patient ID, race and gender. The data set should be cleaned first before modeling and all the missing values needed to be handled prior to being fed into the model. Also data can be partitioned into training (60%) and validation (40%) datasets to provide a precise model assessment. For ease of analysis, the large number of levels in nominal variables can be grouped in a logical way and reduced to a smaller number of levels. Some of the quantitative variables may have high values of skewness and kurtosis. In order to reduce skew, the log transformation can be used. Also to provide a valid assessment, we would like to improve predictive models mainly by two ways.

First, we would like to improve the classical approaches for classification such as discriminant analysis and generalized logistic regression. Depending on the situation we must distinguish which methods are appropriate to use. When the true decision boundaries are linear then generalized logistic regression (GLM) or Linear Discriminant Analysis (LDA) will tend to perform well. When the boundaries are moderately nonlinear Quadratic Discriminant Analysis (QDA) may give better results. For more complicated decision boundaries, non-parametric approaches such as the K- Nearest Neighbors (KNN) method can be superior. To build a proper predictive model, we must decide which functions of predictors should be in the model. Indeed, we must figure out what predictors need to be transformed. A scatterplot matrix is used to examine the marginal relationship of the predictors and response. We can use backward elimination, stepwise and forward methods for variables selection. A GLM is a special case of a generalized additive model (GAM) which allows nonlinear functions of each predictor while maintaining additivity. The estimated additive predictor (EAP) in GAM is given by equation (1).

$$EAP = \hat{\alpha} + \sum_{j=1}^{q} \hat{S}_j(x_j) \tag{1}$$

where q is the number of predictors , α is constant, and S_j is a unknown function of the predictor x_j. The

estimated sufficient predictor (ESP) for a GLM is of the form $ESP = \hat{\alpha} + \hat{\beta}^T \mathbf{x}$. The nonlinear fits of a GAM can potentially make more accurate predictions for the response. In order to fit a more valid GAM, we will need to use smoothing splines rather than natural splines [5]. Standard ROC and area under the curve can be used to measure the predictability of the models. New response graphs (plot of EAP vs response Y) can be used to check the goodness of the fit of GAM. Especially, interesting plots such as EE plots (plot of EAP versus ESP) are useful for detecting cases with high leverage and clusters of cases.

On the other hand, we would like to improve a recent statistical and machine learning approach known as support vector machine (SVM). It is an extension of the support vector classifier that results from enlarging the feature space in a specific way, using kernels. We would like to develop an improved support vector machine algorithm in the R software to predict the readmission for diabetes patients. Stepwise regression, forward regression, LARS [2] (Least-angle regression) and LASSO (least absolute shrinkage and selection operator) can be used for variable selection techniques and prominent factors can be identified to determine readmission rates. In order to predict the readmission, different predictive models can be build based on Decision trees with variations in the splitting rule target criteria (Entropy, Gini), gradient boosting, stochastic gradient boosting [3], Random Forest, Master Boot Record (MBR) and SVM with different kernel function (linear kernel, polynomial kernel or radical kernel). Misclassification rates can used to compare different models and select the best model.

3 Goals and Outcomes

The main goal of this study is to predict the probability of diabetic patient being readmitted and to identify contributing factors. We would like to improve classical approaches for classification and machine learning approaches for predicting readmission of diabetes patients.

By comparing classical approaches for classification and support vector machines separately, we would like to analyze significant factors of readmission of diabetes patients and investigate the different relationships of readmission rates with predictors. Indeed, we would like to determine whether specific drugs or combinations of drugs indicate likelihood of readmission. We would like to investigate whether the number of procedures, medication, and lab procedures correlate with either the readmission probability or the likelihood that the HbA1c test is performed. In fact, we would be interested to Study the correlation between the HBA1C test result and the readmission rate controlling the covariates of the primary diagnosis result. Also, we would like to see whether categorical features of Admission Source, Admission Type, Discharge disposition and age have a significant impact on the readmission prediction. Moreover we would like to develop and validate a tool or model to predict short-term and long term readmission risk of diabetes patients. We would like to compare different improved prediction models and select the best model which gives us valid assessment on the readmission rate of diabetes patients. Providing such an assessment and finding future directions might lead to improvement in patient safety.

4 Proposed Work Statement

- Following is the tentative project schedule starting September 1, 2017.
  - For the first one to two months, we would like to carefully read relevant literature such as Discriminant analysis, GLM and GAM, support vector machines.
  - Next three months we will focus improving prediction models by classical approaches and come up with conclusions and future findings for diabetic patients' safety.
  - Then we would like to improve the support vector machine algorithm for predicting readmission rates of diabetic patients next couple of months.
  - Team meetings will consists of video conference sessions once every two weeks and meet in person at University of Pittsburgh, PA for big data analysis and to improve SVM algorithms.
  - All data and codes will be shared on the GitHub repository so that every team member can easily update the progress.
  - Team members will the attend conference to present our work (oral or poster presentation).

- Each team member will have a unique contribution to this project.
  - **Chathurangi Pathiravasan:** She is interested in improving classical approaches for classification. Indeed she will mainly focus on implementing GAM concepts and determine future directions for patient safety.
  - **Gabrielle LaRosa:** She is interested in improving support vector machine algorithm for predicting diabetes patients.
  - **Justine Stiftel:** She will focus on decision trees for finding contribution factors and different methods for building predictive models such as stochastic gradient boosting, random forest etc.
  - **Anusha Madushani:** She will help Justine's work and compare all our results and findings. Also, she will focus on predicting short-term and long term readmission rates of diabetes patients.

5 Diversity Statement

Women today are more likely than men to complete college and attend graduate school, and make up nearly half of the country's total workforce. Statistical data science is one of the most diverse field in the stem sciences and nowadays many statistics graduates are women. Indeed, all four of the project members are women and we are confident that our research project will interest people from diverse backgrounds.

6 Budget & Justification Section

As depicted in section 4 under the project schedule we are planning to meet in person (3 days in December 2017) for big data analysis at the University of Pittsburgh, PA. We have chosen Pennsylvania because two team members are from PA and it is the least expensive place to meet. Also each team member plans to attend a conference on mathematical biology and statistics in the future. A rough estimate of expenses associated the proposed activities is summarized as follows:

- Four team members attend and co-present project results (oral or poster) at an appropriate statistical conference = $4000.
  - $200 per student registration
  - $400 average flight
  - $400 per logging
- In person meeting at University of Pittsburgh, PA to improve SVM algorithms (3 nights, 2 working days at Pittsburgh) = $2000.
  - Justine (driving $200, hotel $100/night x 3 nights) = $500
  - Chathurangi (flight $450 + hotel $300) = $750
  - Madushani (flight $450 + hotel $300) = $750
  - Gabrielle (already at Pittsburgh)

References

[1] Malladihalli S. Bhuvan, Ankit Kumar, Adil Zafar, and Vinith Kishore. Identifying diabetic patients with high risk of readmission. *arXiv preprint arXiv:1602.04257*, 2016.

[2] Bradley Efron, Trevor Hastie, Iain Johnstone, Robert Tibshirani, et al. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004.

[3] Jerome H. Friedman. Stochastic gradient boosting. *Computational Statistics and Data Analysis*, 38(4):367–378, 2002.

[4] Erik Arango Gutierrez, Hemanshu Mundhada, Thomas Meier, Hartmut Duefel, Marco Bocola, and Ulrich Schwaneberg. Reengineered glucose oxidase for amperometric glucose determination in diabetes analytics. *Biosensors and Bioelectronics*, 50:84–90, 2013.

[5] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*. Springer, 2013.

[6] Daniel J. Rubin, Elizabeth A. Handorf, Sherita Hill Golden, Deborah B. Nelson, Marie E. McDonnell, and Huaqing Zhao. Development and validation of a novel tool to predict hospital readmission risk among patients with diabetes. *Endocrine Practice*, 22(10):1204–1215, 2016.

[7] Beata Strack, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura, Krzysztof J. Cios, and John N. Clore. Impact of hba1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. *BioMed research international*, 2014, 2014.

[8] Xing Yifan and Jai Sharma. Diabetes patient readmission prediction using big data analytic tools. http://www.jimxingyf.com/pdf/CSE4095.pdf, 2016. [Online; accessed 15-Aug-2017].