Improved Predictive Models for Readmission of Patients with Diabetes

Chathurangi Pathiravasan, R. W. M. A. Madushani, Gabrielle LaRosa

August 31, 2018

Team Interactions/Meetings:

We had multiple online team meetings (skype and zoom conference calls) and few in-person meetings after the proposal submission. We have also had numerous online chat interactions.

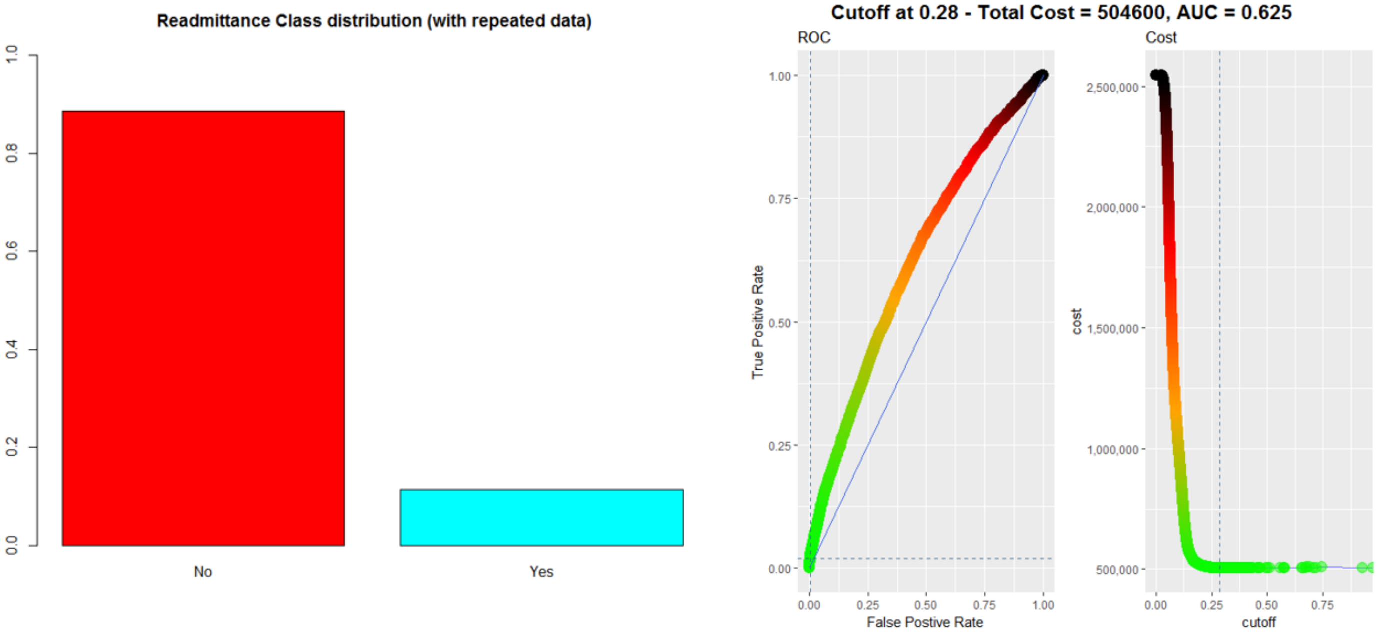
- Dates of Online Meetings:
  - September 30, 2017
  - October 15, 2017 & October 22, 2017
  - November 12, 2017 & November 26, 2017
  - December 10, 2017
  - January 21, 2018
  - February 4, 2018 & February 25, 2018
  - March 18, 2018 & March 25, 2018
  - April 15, 2018 & April 29, 2018
  - June 03, 2018 & June 17, 2018
  - July 01, 2018 & July 22, 2018
  - August 11, 2018 & August 26, 2018
- Dates of in-person Meetings:
  - We met at the Symposium on Data Science and Statistics (SDSS) in Reston, Virginia from May 15, 2018 - May 20, 2018 and discuss more details of our results and future directions.

Problem Background:

Hospital readmission rate is an important measurement of quality of health care and a major contributing factor of total medical expenditures. Diabetes is one of the chronic diseases associated with high hospital readmission rates. Examining the historical patterns of diabetics care is very essential which might lead to improvements in patient safety and prevent future readmissions. In this study, we use a dataset obtained from UCI machine learning repository that contains about 101,766 multiple readmission data of 71,518 unique diabetic patients with 55 attributes such as age, gender, race, admission type, number of inpatient visits, primary diagnosis, number of lab procedure, time in hospital etc. This dataset was originally used in a study that was aimed to find the impact of HbA1c measurement on hospital readmission rates of diabetic patients [1]. In our work, the main objective to build accurate predictive models for early hospital readmission (within 30 days after being discharged from the hospital) and to identify key risk factors contributing to readmission risk of diabetic patients.

After preparing the data set, we implemented different classical methods and machine learning techniques for classification to predict early readmission risk probabilities. In the preprocessing stage, we were able to remove all records that resulted in either discharge to hospice or patient death from the analysis to avoid from biasing. One of the major challenge was dealing with the highly imbalanced nature of the dataset. As far as readmission composition is considered, 90% of the data was no readmission (See

Figure 1a) within 30 days. Classical prediction confusion matrix based on 0.5 cutoff threshold and it is widely used for balanced data set. Sensitivity and specificity based on 0.5 cutoff may not be appropriate for evaluating perdition accuracy of imbalanced data. In fact, we may have to consider different cutoff (see Figure: 1b). Also, to provide valid assessment on the predicted readmission rate for imbalanced data, we can use the Receiver Operating Characteristic Curves (ROC) and Area Under the Curve (AUC). In particular, we evaluated the model performance by partitioning the data set into two parts: Training set (about 60%) and test set (about 40%) and computing the AUC on the test data. The other problem with the data set is that it contains multiple patient admission data (repeated patient data). Most of the classical models such as generalized linear models (GLM), generalized additive models (GAM), discriminant analysis assume the sample independence. Thus for those models, we used only the earliest available admission encounter data for each patient.



(a) Readmission Composition

(b) Choosing different cutoff for ROC

Figure 1: Imbalanced problem with the data set

Methods:

We have used some of the specific classical and machine learning approaches to deal with repeated data. Figure 2 summarize the models we have used in this study. In particular, we considered GLM, GAM, LDA, QDA without repeated data while marginal models and generalized linear mixed models with repeated data. On the other hand, we used random forest with repeated data out of other machine learning approaches. We considered Naïve Bayes, support vector machines (SVM) and gradient boosting methods without repeated data. We have calculated AUC for all these classical and machine learning approaches and compared the model performance. Also, based on Gini Index and accuracy we have found most important predictors in random forest models.

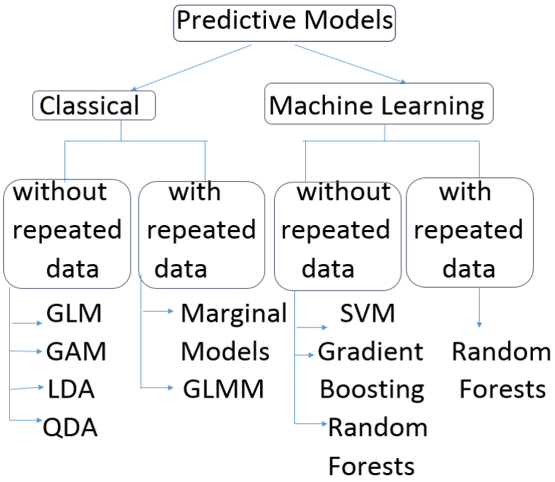


Figure 2: Imbalanced problem with the data set

Results:

We have considered three different GLM models. In fact, existing GLM model[1] is considered as the base model and AUC is computed. Most of the models we have considered include 19 predictors (8 continuous and 11 categorical predictors).

- Continuous Predictors:
  time in hospital, number of lab procedures, number of procedures, number of medications, number of outpatient, number of emergency, number of inpatients, number of diagnoses
- Categorical Predictors:
  race, gender,age, admission type, discharge disposition, admission source, medical specialty, primary diagnosis, max glucose serum, A1C result, medication change

It can be seen that GAM model has a better fit compared to GLM based on lower AIC value. Also, it has better predictive power than all GLM models. We have only considered the continuous predictors for LDA and QDA. It's predictive power is lower, compared to the base model. GLMM and marginal models have the highest predictive power compared to all the other classical models.

Classical Predictive Models	AUC
GLM 1 (Base Model, AIC= 24733.7)	0.6034
GLM 2 (with 19 predictors, AIC= 24509.1)	0.6255
GLM 3 (backward elimination sub model, AIC= 24491.3)	0.6239
GAM (all excluding number of procedure, AIC= 24482.4)	0.6291
LDA (Linear Discriminant Analysis with continuous predictors)	0.6016
QDA (Quadratic Discriminant Analysis with continuous predictors)	0.6018
MM (Marginal Models with repeated data)	0.6480
GLMM (Generalized Linear Mixed Model with repeated data)	0.6470

Table 1: Performance of the classical models

As far as machine learning approaches are considered, Random Forest has better predictability. Naïve Bayes performs well compared to base model. First, we used "without repeated data" for Random forest 1. To improve the random forest model performance and deal with imbalanced problem, we used over sampling and under sampling methods. In these situations, AUC are considerably higher. Further improvement of predictability can be obtained by including repeated data (RF3). SVM takes substantially longer processing time and it's predictive power is very low.

Top ten most important predictors for Random Forest model 3 (with repeated data) are depicted in the Figure: 3. Also, these predictors are highly significant in classical predictive models such as GLM and GAM.

Machine Learning Predictive Models	AUC
NB (Naïve Bayes)	0.6249
RF1 (Random Forest without repeated)	0.6360
RF2: Dealing with imbalanced (without repeated) :Over Sampling	0.6547
RF2: Dealing with imbalanced (without repeated) :Under Sampling	0.6527
RF3 (Random Forest with repeated)	0.6560
SVM (Support vector machine: 19 Predictors)	0.5416
EGB (Extreme gradient boosting (EGB):1000 runs)	0.5835

Table 2: Performance of the Machine Learning models

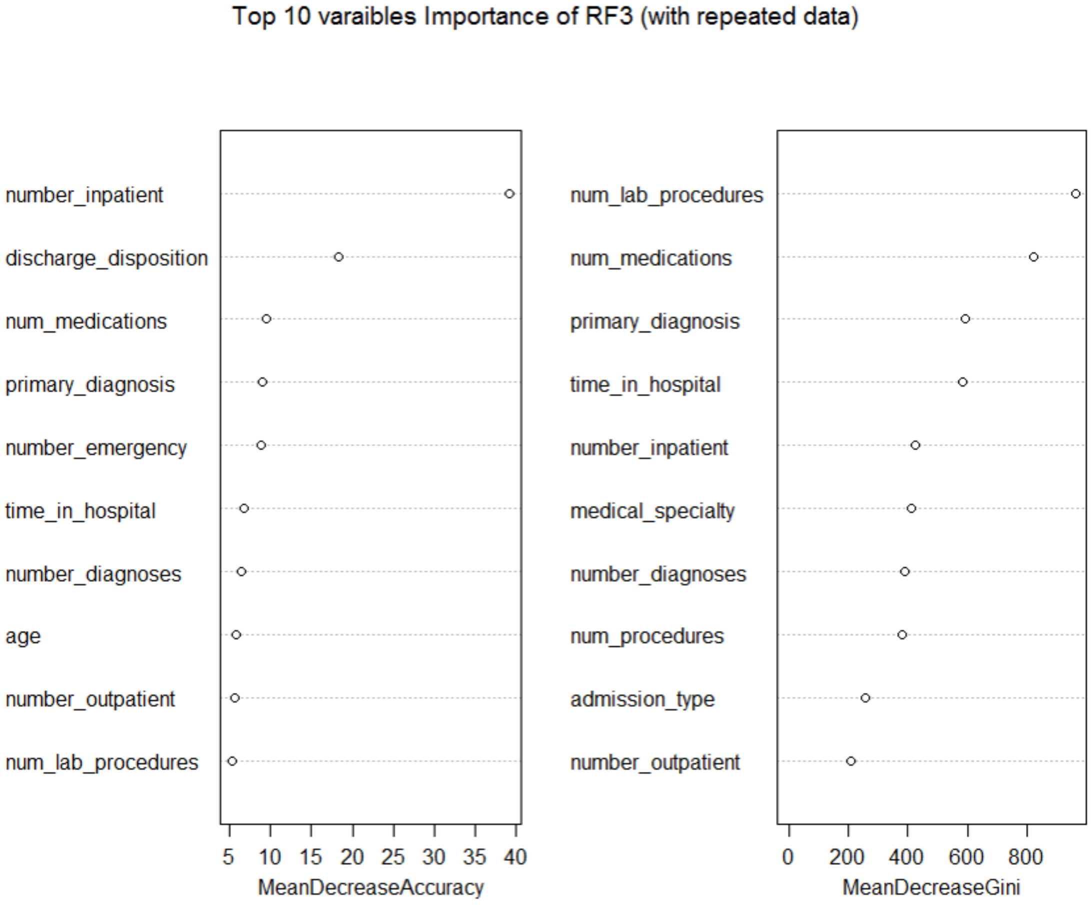


Figure 3: Variable Importance of Random Forest with Repeated Data

Conclusions:

We successfully addressed the imbalanced problem and repeated observations problems in the data set. Different predictive models can be used to predict short term readmission (30- days readmission) of patients with diabetics and AUC can be used to compare these classical and machine learning models. GLM model 2 has more predictability than base model. GAM model is a better fit (AIC lower) and also it has more predictive power than GLM models. Naïve Bayes model has better predictability compared to GLM. There are substantial improvements in the predictability when we consider marginal and mixed effect models. Random Forests performs well compared to other machine learning approaches. We found that Number of lab procedures, number of medications, time in hospital, number of diagnoses, number of procedure, primary diagnosis and medical specialty are the main contributing factors of readmission.

List of presentations, posters, conferences:

- *"Improved Predictive Models for Readmission of Patients with Diabetes"* at the Symposium on Data Science and Statistics (May 17, 2018)

  Abstract/e-Poster link:

  https://ww2.amstat.org/meetings/sdss/2018/onlineprogram/AbstractDetails.cfm?AbstractID=304678

- *"Predicting Hospital Readmission for Diabetes Patients by Classical and Machine Learning Approaches"* at the Vancouver Convention Center, Canada for Joint Statistical Meetings (July 30, 2018)

  Abstract link:

  https://ww2.amstat.org/meetings/jsm/2018/onlineprogram/AbstractDetails.cfm?abstractid=330117

Details of money spent:

We had $6000 fund and all the funding was spent on presenting results in two conferences (SDSS 2018 and JSM 2018).

Future Directions:

We are planning to consider deep learning methods for predicting the readmission rates and compare the other machine learning approaches. Also, initial data set is over a period of 10 years, from 1999 to 2008. We can improve the predictability of models using more data over a longer period of time (from 1999 - 2017). We have achieved all our research goals using short term readmission analysis. Also, we are planning compare short term readmission (less than 30 days) and long term readmission rates (greater than 30 days) using improved classical and machine learning approaches. Robustness of these models also can be checked in further analysis. Combining all these results, we are planning to publish our work in a peer reviewed journal.

References

[1] 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.