Progress Report: Predicting Hospital Readmission for Diabetes Patients

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

March 1, 2018

Examining the historical patterns of diabetics care is very important which might lead to improvement in patient safety and prevent future readmissions. This could not only improve the quality of health care but could also reduce medical expenses. Thus the main goal of this study is to build accurate predictive models for hospital readmission of patients with diabetes and identifying key contributing factors of readmission.

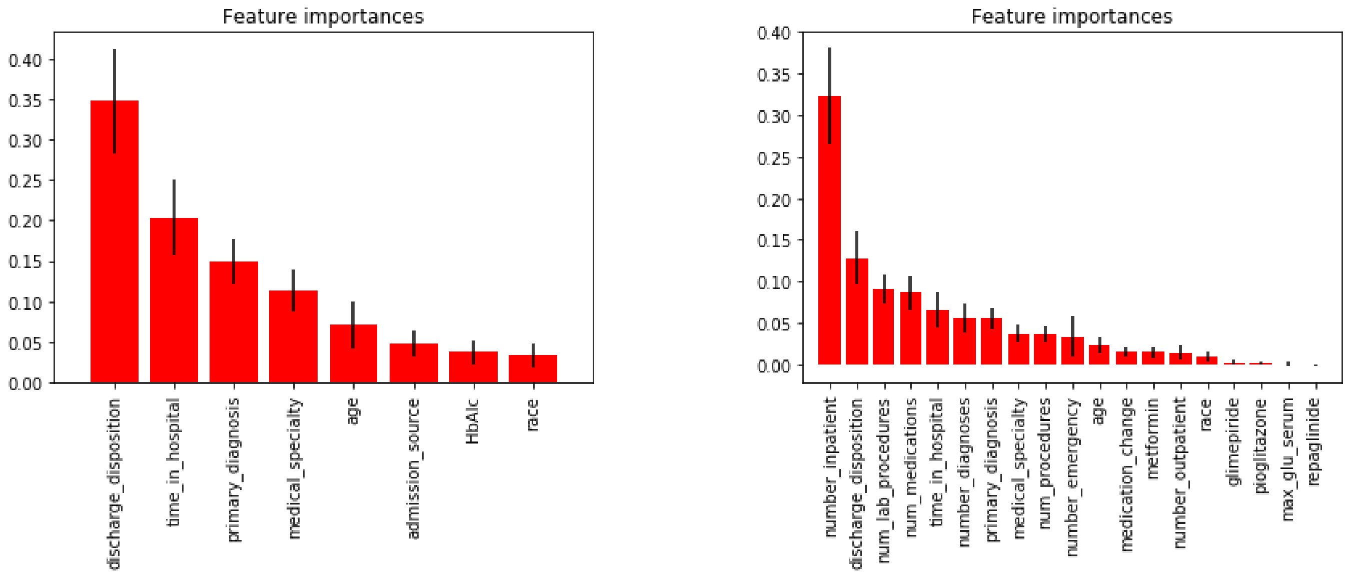
The data set was obtained from UCI Machine Learning Repository and for the ease of compare our results with existing study [1], first we started short term readmission analysis. In fact, we defined the readmission attribute as having two values: “readmitted” if the patient was remitted within 30 days or “not readmitted” as both readmission after 30 days and no readmission together. Most the classical approaches such as GLM requires statistical independence of the data. Since the preliminary data set contains multiple inpatient visits, we first considered only the earliest available encounter of each patient. Also, we removed all encounters that resulted in either discharge to hospice or patient death, to avoid from biasing. First we considered the existing GLM model in literature which includes few interaction terms [1]. We have investigated the predictability of the model by forming data set into two parts: Training set (about 60%) and test set (about 40%). Area under the curve of this model is 0.6034.

Then, we considered the GLM model including 19 different linear predictors: number inpatient, time in hospital, number of emergency, number of outpatient, number of diagnoses, number lab procedures, number of medications, race, gender, age, admission type, discharge disposition, admission source, medical specialty, primary diagnosis, max glucose serum result, A1Cresult, medication change. As far as ROC is considered, the area under the curve (AUC) is 0.6255, thus observing improved predictability. Non-linear terms of GAM allow us to capture nonlinear patterns in data and thereby making more accurate predictions. Hence, we fitted a GAM model including all aforementioned predictors which resulted in a slight increase of the predictive power with AUC around 0.6291. We also determined features that can be deleted in GLM without losing information using backward elimination. It turns out that number of outpatients, number of lab procedure, race, admission type, admission source, max glucose serum result, A1Cresult, medication change may not require for the model. However, the predictive power of the model is comparably lower (AUC = 0.6239) although its lower AIC value (24491.3) compared to previous GLM AIC value (24509.18) suggests that it is a better fit for the training set.

Recent studies show that random forests are one of the most accurate learning algorithms with the ability to balance error in class population for imbalanced data sets. Given that our data set is highly imbalanced (only 9% of early readmission), we fitted four different random forest models aiming to achive higher predictive accuracy. First, using the same 8 features used for the logistic model in literature [1], we trained a random forest model (RF model 1). The model parameters were tuned using a fivefold cross validation and the optimal model was selected by maximizing the average AUC values. The optimal model was fitted with 2300 trees, each fitted with maximum three features and with 130 minimum samples per leaf. The AUC obtained on the test data for RF model 1 is 0.604. Even though, the RF model 1 resulted in a comparable predictive power to that of the logistic model discussed in [1], it is important to note that the predictor “HbAlc” was not among the most predictive features from the random forest (RF model 1 in Figure: 1). This suggests that, even though the measurements of HbAlc test may be associated with reduction of hospital early readmission as concluded in [1], other predictors should also be taken into consideration. One such variable is discharge disposition.

There were several other variables available in the data set such as number of medications, number of inpatient visits, number of lab procedures etc., which were not considered in the logistic model given in the paper [1]. In order to see the effect on the readmission risk prediction from these variables, we incorporated all these variables into our random forest model (RF model 2). Altogether, model was built using 18 predictors except features of medications. AUC from the RF model 2 is 0.635. Together with higher predictive power, it was observed that, all the variables except number of emergency visits, race, number of outpatient visits and maximum glucose serum test results, are more predictive of early readmission rates than “HbAlc”. For comparison purpose, in our RF models 1 and 2, the variable “HbAlc” was created combining the original variables in the dataset: Alc test results with change of medication as described in the paper [1]. Since, there can be loss of information due to combining these variables, and given that random forest should be able to capture nonlinear interacting patterns in the data, next we modified our random forest model using Alc test results and change of medication as two separate variables (RF model 3). This only increased the AUC by a sight amount to 0.636. However, the predictive importance of these two features were maintained to be low similar to the previous RF models.

As mentioned earlier, one problem of using the logistic regression model is we are limited to use data from only one admission encounter for each patient to comply with the independent sample assumption of the logistic model. However, such assumptions are not violated when modeling with random forest. Having an imbalance data set, building accurate predictive models is quite challenging and therefore, limiting the data that can be used is not desired. Hence, enriching the data set by utilizing multiple admission data of the same patients can increase the predictive accuracy. Therefore, we trained a random forest model (RF model 4) using multiple admission data. The predictive power of the model increased considerably as expected (AUC is 0.656) and Figure 1 shows the feature important ranking from the (RF model 4).



(a) RF model 1

(b) RF model 4

Figure 1: Feature importance ranking from Random Forests model 1 and 4

In an ongoing analysis, we are implementing other learning algorithms such as discriminant analysis, gradient boosting and support vector machines. In order to provide valid assessment on the readmission rate of diabetes patients, these models will be compared to select the best model based on area under the receiver operating characteristic curves (AUC). Also, we are planning to extend our analysis using additional information in the response variable ’readmission’ to compare the short term and long term readmission rates.

List of conferences planning to attend

- E-poster presentation at Reston, Virginia Hyatt Regency Reston for Symposium on Data Science & Statistics. Will be held on May 16 - 19, 2018 (Abstract submitted).
- Poster presentation at Vancouver Convention Center, Canada for Joint Statistical Meetings (American Statistical Association). Will be held on July 28 - August 2, 2018 (Abstract submitted).

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.