Machine Learning Based Framework for Prediction of Diabetic Patient Readmission Rate in Hospitals

Chandan Srivastava
Microsoft, Data and Artificial Intelligence, Hyderabad, India
E-Mail: chandan.iitk@gmail.com; Tel: +91-9010626734.

ABSTRACT Diabetic Readmission Decision (DRD) model is imperative research contribution as hospitals indicator to avoid extra medical expenses and increase patient trust and care. Diabetes is a chronic medical disease, which may cause of high risk of readmission to hospital. However, clinical evidence says that glycemic control for inpatient and outpatient is less responsible for re- hospitalization rate. Current analysis over large patient data (demographic, medical & clinical) using machine learning methods support vector machine (SVM) for predicting readmission rate (risk score) and decision (Yes/No). Proposed model gives the low misclassification rate 0.27 with 72.42% sensitivity of identification for readmission. Overall analysis reflects that correct determination of HbA1c may reduce the risk of readmission and inpatient care cost.

KEYWORDS: Diabetic Readmission, Ensemble Approach, Support Vector Machine, Prediction and Risk Score.

I Received 2 September 2017 **II Revised** 5 June 2018 **II Accepted** 12 June 2018 **II In press** 8 August 2018 **II Online** 28 June 2018 **II**

TRANSACTIONS ON SCIENCE AND TECHNOLOGY

INTRODUCTION

Diabetes is a chronic condition concerning people of all ages and is common in around 25.8 million people in the United States (Strack *et al.*, 2014). Hospitals are bearing a significant proportion of costs for small percentage of patients, particularly those having chronic medical conditions. These costs are in a large fraction, because of repeated hospitalizations. However, diabetic patients in hospital may have higher risk of readmission than those without diabetes. As per review (Dungan, 2012), approximately 20% of all hospitalized medicare diabetic patients are readmitted within 30 days, and 34% are readmitted within 90 days of discharge anticipation of unexpected hospital diabetic readmission has then received large attention as one approach of reducing hospital costs. For example, the Medicare Payment Advisory Commission has recommended reduced reimbursement rates for patients having early rehospitalizations for congestive heart failure (CHF) (Dungan, 2012; Epstein, 2009).

The improvement in standardization of HbA1c measurement (Littlea & Sacksb, 2009) methods has considerably decreased among other methods. On the other hand, in terms of both the morbidity and mortality, management of hyperglycemia in hospitalized patients has a significant bearing on outcome (Strack *et al.*, 2014). However, there are few national assessments of diabetes care in the Hospitalized patient which could serve as a baseline for change (Lansang & Umpierrez, 2008). After literature review, we found that only a single model is able to clearly define and identify potentially avoidable readmissions (Halfon *et al.*, 2002). Most of the models are being applied currently in clinical, research or policy arenas. We found few models were more emphasizing in model design for the calculation of risk-standardized readmission rates for hospitals outcome comparison purpose. A few of them are clinical models which were representing to identify high-risk patients and the other models in both categories have poor predictive ability (Kansagara *et al.*, 2011; Billings *et al.*, 2006; Howell *et al.*, 2009). The statistical analysis has proven that the relationship between the

probability of readmission and the HbA1c measurement depends on the primary diagnosis (Strack *et al.*, 2014).

The present analysis is focused on a large clinical database which was undertaken to examine historical patterns of diabetes care in patients with diabetes who are admitted to hospitals in the United States and to inform future directions which might lead to improvements in patient safety. In particular, we consider measurement of HbA1c is associated with a reduction in readmission rates in individuals those are admitted to the hospital (Strack *et al.*, 2014). Thus, we build a robust diabetic readmission decision predictive model to predict the readmission decision and estimate the risk score for diabetic patients being readmitted using the patient demographic, hospital history and clinical data (Frank & Asuncion, 2010). The data set is used for model development, divided into training and validation for development of Diabetic Readmission Decision (DRD) models. Predictive model has been developed using statistics and advanced machine learning technique - Support Vector Machine (SVM) using open source analytics tool R-package, which is similar to boundary based methods. The review analysis suggested that DRD model compared to rest of the developed predictive models is giving good performance and having the lowest misclassification rate and parallel reporting with good model performance.

We understood the inherent hospitals challenges and have developed a robust readmission prediction decision support system to predict the readmission decision (Yes/NO) and determining the patient readmission rate. This will help the hospitals to reduce the unpredicted readmission, to improve the diagnosis protocol and diabetic-specific interventions to prevent readmission. The results from this research will help hospitals stratify the readmissions risk score for adhering readmission and patient engagement compliance. The research follows the dataset standards that contain no personally identifiable information.

METHODOLOGY

Data Used for Model Development

The used dataset belongs to the Health Facts database (Cerner Corporation, Kansas City, MO, http://dx.doi.org/10.1155/2014/781670), and obtained from Center for Machine Learning and Intelligent Systems at the University of California, Irvine (Lansang & Umpierrez, 2008). Description of raw data collection approach can be found in detail (Lansang & Umpierrez, 2008). The dataset D contains 101,766 instances $\{d_1, \ldots\ldots, d_n\}$ and 55 variables $F = \{f_1, \ldots\ldots, f_n\}$ such as insulin and length of stay, etc. and their binary labels $y_i \in R$ (where "1" stands for readmitted and "0" not readmitted), where $(1 < i \leq n)$, $n = |D||F|$ (the number of patient-predictor pairs).

Data Preprocessing

Databases of clinical data contains valuable but heterogeneous and difficult data in terms of missing values, incoherent values and dimension reduction by their complexity (Cios & Moore, 2002). Further, we removed all attributes having inappropriate/missing information and zero "0" for 90% of complete dataset and deal effectively with present uncertainty in datasets by using fundamental numerical analysis (Srivastava, 2013). After appropriate selection, we are having the final 18 attributes (Figure 1) dataset for model development and validation.

The statistical information on ordinal attributes corresponding to the above mentioned complete data set is displayed in (figure 1), as a box plot. Attributes (race, gender, A1Cresult, metformin,

glipizide, insulin, change and diabetesMed) have been transformed as their numeric variable and some of them which are clinical measurement attributes (admission_type_id, discharge_disposition_id, admission_source_id and medical_specialty) have been considered as given in original dataset. The diagram below entails a variety of different boxes, plots shapes and positions for 6 attributes. It shows the median, upper quartile (UQ), lower quartile (LQ), calculate interquartile range (IQR = UQ-LQ) values and outliers for each of chosen attributes of the data set. This provides a clear-cut insight of attribute's behavior and evidence concerning model performance.

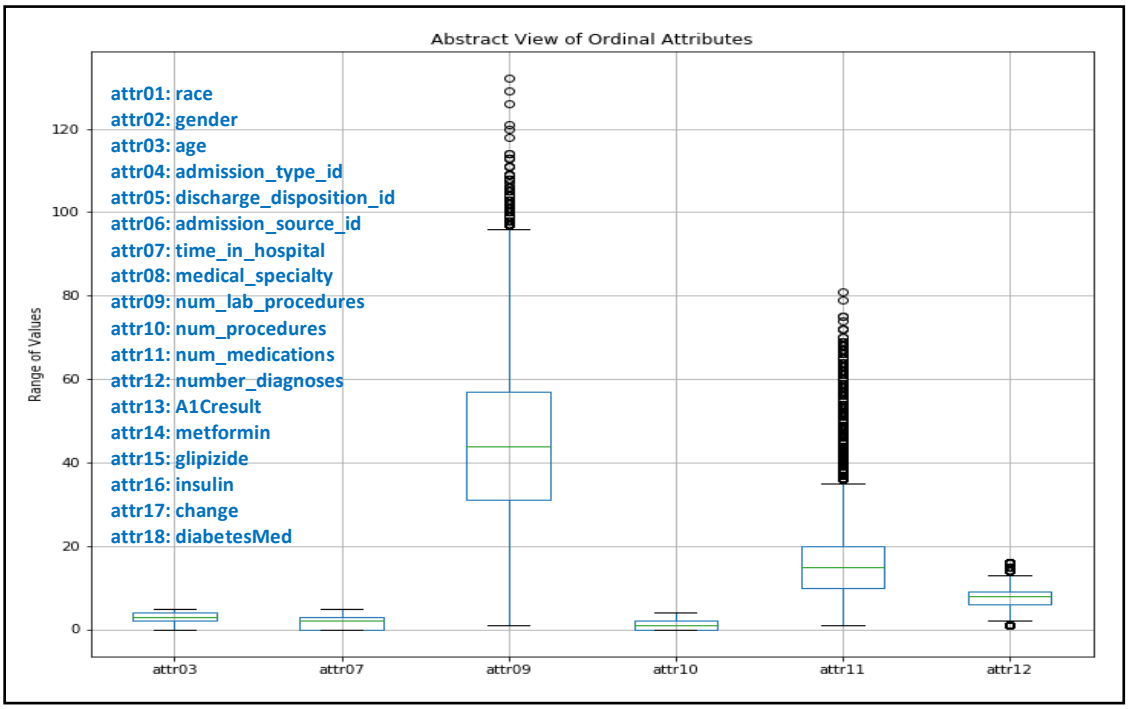

Figure 1. Data Visualization - an abstract view of Ordinal dataset for 6 descriptive attributes.

Model Development and Validation

The dataset was split into training (~59%) and validation (~41%) of entire dataset to provide a model assessment. We have developed a model using machine learning technique Support Vector Machine (SVM) in open source R (CRAN-e1071) package ("https://cran.r-project.org/web/packages/-e1071/index.html "). Train model with ~59% (60K) dataset N with $\{y_k, x_k\}_{k=1}^{N}$ having 18+1(output pattern y_k) set of variables $x_k \epsilon R^n$, where 19th variable y_k has binary class outcomes. The rest ~ 41% (~ 41K) set used for model validation and assessment. Also, calculate the train model performance using 10-fold cross validation and calculate average performance of the model after 500 random iteration over a complete dataset, reported in table 1.

Binary Classifiers

We have built three different classifiers (SVM_1, SVM_2, SVM_3) using same machine learning method - Support Vector Machine (SVM) (Chapelle *et al.*,1999) over three different data samples of training dataset N = (N_1=20K, N_2=20K, N_3=20K). Concerning the SVM for binary class classification, we have studied about two classes of data pattern classification. We tested and found that SVM classifier, classifying by constructing an optimal separating hyperplane between two classes. The rule of optimization is to increase the margin of separation $\|2/w\|$ and decrease the upper bound of classification error. As binary class classifier, first choose a non-linear mapping and then map input

vectors into high-dimensional feature space for constructing the optimal hyperplane. Thus, we can state that SVM as a quadratic optimization problem:

$$L(w,\xi) = \frac{1}{2} w^T w + C\sum_{i=1}^{N} \xi_i \quad (1)$$

with the following constraints: $y_i\ (w^T\ \phi\ (x_i)+b \geq 1-\ \xi_i)$ and $\xi_i \geq 0$. Whereas optimal hyperplane is denoted by w, the regularization parameter is by C (where, $C > 0$ is the penalty parameter of the error term) and the mapping function by ϕ whose dot product forms the kernel function. Furthermore, radial basis function (RBF) is used as the kernel

$$K(x_i, x_j) \equiv \phi(x_i)^T \phi(x_j).$$

$$\text{RBF: } K(x_i, x_j) = exp(-\gamma\ \|x_i - x_j\|^2),\ \gamma>0 \quad (2)$$

Here, γ are the kernel parameter.

Risk Score

Let x_i^+ represents the positive class feature vector, x_i^- represents the negative class feature vector and $f(x_i)$ generates a value between [0 & 1], which reflects the importance of x_i in its own class. Here $f(x_i)$ is based on the distance d_i^{cen}, where $d_i^{cen} = \|x_i - \bar{x}\|^{1/2}$ is the Euclidean distance to x_i from its own class center $\bar{x}$. The feature vectors which are closer to the center are treated as low risk and assigned higher $f(x_i)$ values, the feature vectors are far away from the center are treated as higher risk and assigned lower $f(x_i)$ values. The feature vectors in a moderate range from the center are treated as moderate risk and are assigned in the range [0.4 - 0.6] $f(x_i)$ values. Here for each i one decaying function of d_i^{cen} has been used to define the corresponding $f(x_i)$, which is represented by $f^{cen}(x_i)$ as follows:

$$f^{cen}(x_i) = 1 - d_i^{cen}/(\max(d_i^{cen}) + \beta), \quad (3)$$

where β is a small positive value used to avoid the case when $f(x_i)$ becomes zero..

Application Platform

A user friendly web application (Dashboard: "http://www.primetgi.com/predictive-analytics/") have been built for hospital care team using open source language and deploy Diabetic Readmission Model (DRD) model *.Rexec for showcasing model outcome and giving a real-time opportunity to the Hospitals (care team) to insert own data and test in/out patient status.

RESULTS AND DISCUSSIPON

As we analyze, the associated factors in patient demographic information, medical and clinical procedures for example age, gender, a number of lab procedure, a number of medications, a number of diagnosis, HbA1c etc. in the dataset for prediction are the key factors for identifying the significant readmission. Using these highly influence factors, we have provided an artificial intelligence model and found that the single classifier (SVM) based model was unable to perform well on class unbalanced dataset (~90% belongs to "0" not-readmitted patient class and ~ 10% belongs to "1" readmitted patient class) and performing as earlier models which have been reported.

Further, we designed a solution in such a way that building three different classifiers (SVM_1, SVM_2, SVM_3) for different 1/3 part (20K) of training dataset (60K) and test with remain ~41K validation set. Since, we realize that all single classifiers were unable to perform well, a consensus approach was applied ("no-lose" method: if any local model says "Yes", final prediction will be "Yes") to derive a final conclusion. Hence, we can deal well with variability (Strack *et al.*, 2014) of all key attributes to achieve an improved sensitivity.

The average performance of all the above three classifiers after k-fold cross validation in the best composition of the training set is 60% and model validation performance is with less mistakes (27%) and sensitivity 72.42% to predict the correct readmission, reported in table 1. All three classifiers, we treated as one class (readmission) classification model for prediction. We used grid search approach to optimize the models (radial basis function kernel) parameters (Cost & Gamma) and always keeping in mind the over-fitting situation.

Additionally, DRD model is providing risk score to each inpatient record in dataset and able to provide a score to outpatient as well, saying that this particular patient is having more chances for risk (%) of readmission. We used distance-based criteria over vector space using same machine learning (SVM) methods. Hence, We have analyzed the association between risk categories (high/ moderate/low) and actual readmission rates. The DRD model outcome (Table 1) suggests that the relationship between the readmission decision, assigned risk score and the HbA1c measurement significantly depends on the primary diagnosis (diabetes is the secondary diagnoses). Explicitly, rightly identify patient profile of readmission in DRD model space is clearly distinguished by the healthy candidates.

Table 1. Average Performance of the Model, using 18+1 Variables for Model Development.

Input: Train data is 60K Total Test data is: 41766	Average performance of the model
Cost	90
Gamma	0.07
Tp	3253
Tn	13962
Fp	23312
Fn	1239
Sensitivity	72.42%
Tp+Fn Total 1's	4492
Tn+Fp Total 0's	37274
% Misclassificatiom For 1's	27.58%
Accuracy for 1's	72.42%

The present study provides a striking cross-sectional view of inpatient diabetes care for more than 100K admissions in 54 hospitals in the USA.

- The results from this research will help hospitals stratify the readmissions risk score for adhering readmission and patient engagement compliance.
- Customize decision support system to the hospital Care Manger.
- Improve patient outcomes and lower cost of inpatient care.
- Attention to HbA1c measurement and diabetic medications in the hospital.
- Help Hospitals & patient to provide evidence-based information.
- Fulfillment and conformity to effective US standards.

CONCLUSION

In this study, we have developed a risk prediction and decision (Yes/No) model for reducing diabetic patient readmission and care cost in Hospital. Our evidence based analysis clearly shows that the profile of high risk patients is different than the moderate and low risk patients and direct

hospitals care to focus on expected inpatient care more with limited clinical observation. Moreover, we anticipate that further analysis is required over clinical data, patient report and a broad variety of factor interpretations and utilizations for the development of general predictive models for other chronic disease readmission risk predictions, with an effect of avoidable readmissions in US health systems.

ACKNOWLEDGEMENTS

This research supported by the Prime Technology Group, LLC, PA, USA.

REFRRENCES

[1] Billings, J., Dixon, J., Mijanovich, T. & Wennberg, D. (2006). Case finding for patients at risk of readmission to hospital: development of algorithm to identify high risk patients. *BMJ*, doi:10.1136/bmj.38870.657917.AE

[2] Cios, K. J. & Moore, G. W. (2002) Uniqueness of medical data mining. *Artificial Intelligence in Medicine*, **26**(1-2), 1–24.

[3] Chapelle, O., Haffner, P., & Vapnik, V. (1999) Support vector machines for histogram-based image classification. *IEEE Trans. on Neural Networks*, **10**, 1055–1064.

[4] Dungan, K. M. (2012). The Effect of Diabetes on Hospital Readmissions. *Journal of Diabetes Science and Technology*, **6**(5), 1045-52.

[5] Epstein, A. M. (2009). Revisiting readmissions--changing the incentives for shared accountability. *The New England Journal of Medicine*, **360**(14), 1457–9.

[6] Frank, A. & Asuncion, A. (2010) *UCI Machine Learning Repository*. University of California, School of Information and Computer Science.

[7] Halfon, P., Eggli, Y., Melle, V. G., Chevalier, J., Wasserfallen, J. B. & Burnand, B. (2002). Measuring potentially avoidable hospital readmissions. *J Clin Epidemiol*, **55**(6), 573-587.

[8] Howell, S., Coory, M., Martin, J. & Duckett, S. (2009). Using routine inpatient data to identify patients at risk of hospital readmission. *BMC Health Services Research*, 2009, 9:96.

[9] Kansagara, D., Englander, H., Salanitro, A., Kagen, D., Theobald, C., Freeman, M. & Kripalani, S. (2011). Risk Prediction Models for Hospital Readmission A Systematic Review. *Clinical Review*, **306**(15), 1688-1698.

[10] Littlea, R. R. & Sacksb, D. B. (2009). HbA1c: how do we measure it and what does it mean?, Current Opinion in Endocrinology. *Diabetes & Obesity*, **16**, 113–118, DOI:10.1097/MED.0b013e328327728d.

[11] Lansang, M.C. & Umpierrez, G. E. (2008) Management of Inpatient Hyperglycemia in Noncritically Ill Patients. *Diabetes Spectrum*, 21(4):248-255, DOI:10.2337/diaspect.21.4.248.

[12] Strack, B., DeShazo, J. P., Gennings, C., Olmo, J. L., Ventura, S., Cios, K. J. & Clore, J. N. (2014). Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records. *BioMed Research International*, 2014, Article ID 781670, http://dx.doi.org/10.1155/2014/781670.

[13] Srivastava, C. (2013) Biological Data Analysis: Error and Uncertainty. *World Journal of Computer Application and Technology*, **1**(3), 67 - 74.