Risk feature assessment of readmission for diabetes

Qian Zhu, PhD, Anirudh Akkati, MS, Pornpoh Hongwattanakul
Department of Information Systems,
University of Maryland Baltimore County,
Baltimore, USA
{qianzhu, anirudh1, hongw1}@umbc.edu

***Abstract*—About 382 million people have Diabetes in 2013, and the International Diabetes Federation estimated that there are 4.9 million people died from Diabetes in 2014. Diabetes continues to be a chronic disease plagued by frequent hospital readmissions. In order to better understand the risk features impacting readmissions for future prevention and management, in this study, we programmatically analyzed a large clinical dataset containing more than 100,000 clinical records for diabetes patients from 130 US hospitals. Specifically, we developed three different machine learning algorithms, Logistic Regression, Random Forest and manipulated Random Forest to identify and prioritize the most significant risk features. By comparing the results generated by these three methods, the manipulated Random Forest illustrates greater capacity of generating a more complete and concrete list of readmission related risk features. Such method is generalizable and can be applied in other disease oriented studies.**

Keywords—Diabetes; Readmission; Risk features; Machine learning;

I. INTRODUCTION

Diabetes is a group of metabolic diseases showing that there are high blood sugar levels over a prolong periods, it can cause serious health complications including heart disease, blindness, and kidney failure. About 382 million people have Diabetes in 2013, and the International Diabetes Federation estimated that there are 4.9 million people died from Diabetes in 2014. [1] Estimates indicate an additional 2 million people 20 years and older are diagnosed with diabetes each year. Finally, estimates of people at risk for diabetes or people with pre-diabetes are approximately 79 million.[1] Despite the growth in scientific advances in management, diabetes continues to be a chronic disease plagued by frequent hospital readmissions. Patients with diabetes account for approximately 480,958 hospital in-patient stays per year with a 30-day readmission rate of 97,784, accounting for a 20.3% hospital readmission rate.[2] Given these statistics, it's not surprising that reducing readmission rates for patients with diabetes has become an important goal for hospitals and healthcare providers. And, given the number of people who will be diagnosed with diabetes in the future, emphasis on that goal is only likely to increase.

Readmission rate defines hospital readmission as patient admission to a hospital after being discharged from an earlier hospital stay. Generally, research and experiments usually specifics the time/length for the readmission rate. For example, with in 30 days or 3 months. Readmission rate is also considered as a quality and obligation in the hospital. Beginning in 2013, hospitals with high risk-standardized readmission rates will be subject to a Medicare reimbursement penalty[3]. Therefore, better assess the clinical features that have negative impact to the readmission rate will result in better preventing and managing health conditions of diabetes papers, ultimately improve their quality of life. Electronic health records (EHR) contain a large volume of longitudinal patients' medical records, which of the analysis of the large clinical database in diabetes patients has been developed to evaluate and examine the historical pattern of diabetes care in patients with diabetes admitted to enhance the future directions, which will lead to improvements of medical care. EHR data contains a great number of features.

In this study, we will access the dataset representing 10 years (1999-2008) of clinical care for diabetes patients at 130 US hospitals and integrated delivery networks, from the University of California Irvine, to systematically assess risk features impacting readmission by comparing three machine learning algorithms. We introduce the dataset applied in this study in the Materials section, and describe the methodology step-by-step in the Methods section. Finally we discuss the benefit obtained and lessons learnt in this study.

II. MATERIALS

In this research experiment, we used the health Facts database (Cerner Corporation, Kansas City, MO), a national data warehouse that store comprehensive clinical data across the United States hospitals. Health Fact is a voluntary program that gathers the clinical records from the organizations that use the Cerner Electronic Health Record System. The database collected the medical records from participating institutions. The data include encounter data (emergency, outpatient, and impatient), provider specialty, demographics (age, sex, and race), diagnoses and in-hospital procedures documented by ICD-9-CM codes, laboratory data, pharmacy data, in- hospital mortality, and hospital characteristics. The Health Facts data represents 10 years (1999 – 2008) of clinical records from 130 hospitals throughout the United States. The dataset has 55 attributes and available as supplementary resource online at http://dx.doi.org/10.1155/2014/781670 and at UCI machine-learning repository[4] online at

http://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008. 101,766 records were identified to satisfy all five-inclusion criteria below.

1. Inpatient encounters
2. Diabetes encounters as one of any kind of diabetes
3. The length of stay was between 1 to 14 days
4. Laboratory were performed during the encounter
5. Medication was administered during the encounter.

The dataset consists of hospital admissions of length between one and 14 days, and the dataset did not represent a patient death or discharge to a hospice. Each data records relevant to a unique patient diagnosed with diabetes. The dataset contains multiple inpatient records and the observations that could not be considered as statistical independent variables.[4]

III. METHODS

A. Data preperation.

The original data contains incomplete and noisy data as inevitable in any real-world data.

- There were multiple features with very high percentage of missing values. These features are *Weight* (97% values missing), *Payer code* (40% values missing), and *Medical specialty* (47% values missing), which are too sparse so we removed them from our experiment.

- *Encounter ID* and *Patient number* as indexes, along with other medication related attributes, such as *acetohexamide*, *glimepiride-pioglitazone*, *examide*, *citoglipton*, *metformin-rosiglitazone*, *metformin-pioglitazone*, *troglitazone*, *tolbutamide* and *glipizide-metformin* have less meaningful information, as most of the observations were recorded with same value, which renders no change or variation on target variable. Thus, we excluded them for further analysis.

- The rest of 41 attributes have been included for further analaysis.

B. Risk feature assessment for readmission.

To systematically assess clinical features that impact the readmission rate, we performed and compared three different machine learning algorithms, Logistic regression,[5] Random forest [6] and modified Random forest.[7] Out of many machine learning methods, the reason to chose these three are, 1) logistic regression gives basic overview of features impacting and binary output with probability response and p values; 2) Random forest is one of the advanced classification methods, which beats decision trees, SVM and other traditional classification techniques. 3) Modified Random Forest is the methodology where we sample and distribute each variable in our data to check the effect on readmission learning rate. We were using R to build these models for risk feature assessment.

- *Logistic Regression*

Logistic regression[8-10] is used to analyze relationships between categorical dependent variable and quantitative or qualitative independent variables. In our dataset, independent variables are both continuous and featureial variables. It is a special case of linear model. [11]

The logistic regression model can be represented by the below formular.

$$Ln[\frac{P}{P-1}] = \alpha + \beta 1X1 + \beta 2X2 + \cdots + \beta nXn + e$$

And the estimated probability is

$$P = 1/[1 + e^{(\alpha + \beta 1X1 + \beta 2X2 + \cdots + \beta nXn)}]$$

Where P is the probability that the event Y occurs, p(Y = 1) p/(1-p) is the odds ratio. ln[p/(1-p)] is the log of odds ratio also known as logit. Each X (independent variable) is given individual coefficient β which measures its weight contributing to the variation in dependent variable.

Logistic regression forms a best fitting equation or function using the maximum likelihood method, which maximizes the probability of classifying the observed data into the appropriate category given the regression coefficients. Logistic Regression does not make any assumptions of normality like linear regression, and homogeneity of variance of other variables. The model doesn't depend on variable correlations with each other like naïve Bayes.[12]

Using R, Logistic regression was performed on all variables excluded the one mentioned in the section A. The model generated results in many arguments like Standard error, Z-value and auto calculated P value using Z-value, AIC value.

- *Random Forest*

Random forest is one of the bagging algorithms,[13] which come under ensemble methods,[14] and the other is boosting algorithms. It is a conglomerate of many decision trees[15] with certain parameters, and the target variable with more votes from each tree is the final output. Random forest reduces variance of large number of complex modes with low bias. Each tree is grown with: N random cases from all the observations, these acts as training for model m number of variables are selected out of total M variables, where m << M in general m is taken as m <= sqrt (M). Each tree is left to grow without pruning. We calculated error rate by using OOB (out of bag error), which is a method of measuring the prediction error, and tuning is done using error vs. number of trees, which was applied to determine the number of trees to be generated.

- *Manipulated Random Forest.*

We used the above Random Forest algorithm as a wrapper class and extended its functionality. The functional requirement is to observe the change in accuracy when we randomized a feature and rank the features accordingly. The Figure 2 depicts the workflow of the manipulated Random Forest designed in this study.

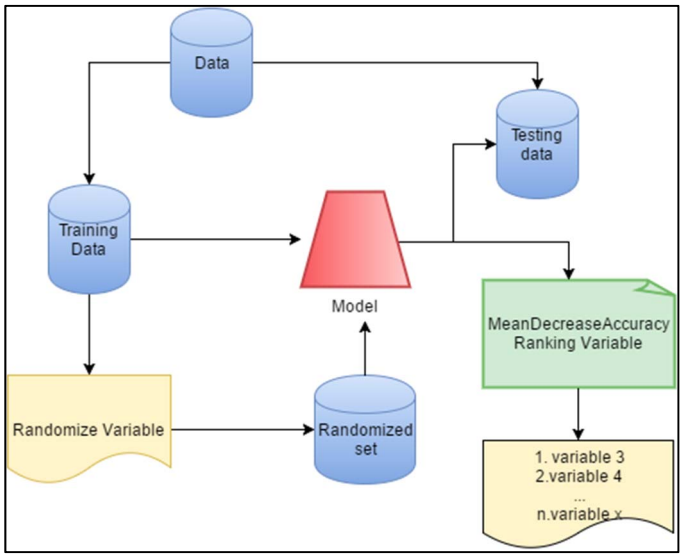

Figure 2. Workflow of the manipulated Random Forest.

• Train and test the model without any modification using Random forest.

• Add randomness[13] to each feature one by one by creating new data each time and test the data on the model created before and observe the meanDecreaseAccuracy.

• Repeat the same step for same feature n number of times and average the meanDecreaseAccuracy.

• After repeating the above two steps on every feature, we rank features according to the average taken from meanDecreaseAccuracy.

IV. RESULTS

A. Results based on Logistic regression.

P-value helps to determine the significance of each variable in way how it is related to the target value. Having p-value less than 0.05 implies strong significance and rejects null hypothesis, p- value greater than 0.05 indicates weak significance against null hypothesis and cannot reject null hypothesis, P-value which equals to 0.05 is a marginal value and the readers can draw their conclusions. Table 1 shows the selected features with p-values less than 0.05.

Table 1. A list of clinical features with high P-Value.

Features	P-Value
number_inpatient	0
number_diagnoses	0
number_emergency	0
number_outpatient	0
diabetesMedYes	0
discharge_disposition_id	0
num_lab_procedures	0
max_glu_serumNone	0
payer_codeSP	0.0001
payer_codeMD	0.0014
admission_type_id	0.002
payer_codeMC	0.0021
payer_codeCM	0.0054
insulinSteady	0.0164
time_in_hospital	0.0188
payer_codePO	0.0383
payer_codeDM	0.0481

Shown in the table 1, we get 17 features that are obviously leading to high readmission. For instance, Number of inpatient visits to the hospital in that year (number_inpatient); number of diagnoses (number_diagnoses); number of emergency visits (number_emergency); number of outpatient visits (number_outpatient); if the patient is prescribed for diabetes medications (diabetesMedYes); the way patient is discharged (discharge_disposition_id); number of lab procedures (num_lab_procedures); how the patient is admitted (admission_type_id); time spent at hospital (time_in_hospital) and payer code along with glucose and Insulin levels (max_glu_serumNone, insulinSteady).

B. Results based on Random Forest.

We calculated error rate shown in the Figure 3. Numbers of trees to be generated are decided using this plot, where knee point occurs is taken as optimum number of tree to create. In the Figure 3, the black line is an average of the red line and green line, where red line and green line shows error rate for the classification outputs YES and NO respectively. The black curve has its knee break point between 150 and 200, the error rate after 200 is constant till 1000 tree we generated before.

In order to identify important variables, same as splitting measure used in decision trees, Gini index[16] is used to calculate variable importance, Mean Gini index is calculated from all the trees generated and variables are ranked accordingly. It is shown to right in the Figure 4. Variable ranking is also done using the accuracy of the model, which is our main objective and shown to left in the Figure 4.

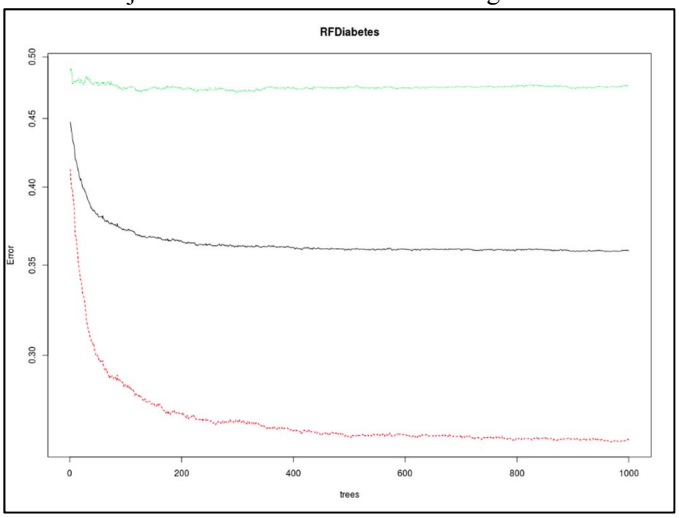

Figure 3. Tree generation for Random Forest Model.

C. Results based on manipulated Random Forest.

We plotted the importance vs variable plot which is in the Figure 5. There is huge contrast with top four variables and others, using automatic threshold limit we considered the green bar plot variable to be the impacting features.

D. Results generated by the three models.

Table 2 is derived from the results generated by the three machine learning methods we performed in this study. Each column contains features with their respective scores. The significant risk features impacting readmission can be identified from the Table 2. Most of the common features, such as number of inpatient, number of emergency have been generated by the three methods and are ranking on the top list.

By comparing the results generated by the three methods, the list of risk features generated by the modified Random Forest is more complete. Interestingly, two medications, Pioglitazone and Repaglinide are included in the list. In order to proof such finding, we conducted literature review. Pioglitazone shows risk to incresase readmission rate according to the published studies, for instance, [17] reported that Pioglitazone increased risk of bladder cancer in people with type 2 diabetes. Findings reported in [18] shows that the use of Pioglitazone increases the fracture risk in both men and women. For Repaglinide, we found that metformin, which is one of the top oral diabetes medication, usually use combination with Repaglinide. When diabetic patients take a combination Pioglitazone and Repaglinide, there are some severe effect such as Hypoglycemia, or low blood sugar. Symptoms include headache, hunger, weakness, sweating, tremors, irritability, trouble concentrating, rapid breathing, fast heartbeat, fainting, or seizure.[19] Therefore, we believed that those effects from Pioglitazone and Repaglinide could make hospitalized patients to readmit.

V. Discussion and Conclusion

In this study, we access a health facts database to programmatically identify risk features significantly impacting readmission for diabetes patients. We applied three different machine learning algorithms to prioritize risk features associated with readmission. Results shown in this experiment, besides the features commonly observed as highly impacting readmission, some interesting results have also been generated and proofed via literature review. As following, we will discuss the benefit we obtained and lessons we learnt from this study, and consequently the planned future work.

A. Modification of Random Forest.

We modified the traditional Random Forest algorithm by adding randomness to each feature and ranking features based on the average taken from meanDecreaseAccuracy. The main advantage of this modification is we can provide strong evidence without bias because of a number of iterations and randomizing features. It provides ranking for all features, which we can decide later how to use the most significant and least significant features. Disadvantage of this approach is time consumed by the process is much larger than original random forest because of large computations and iterations on each feature. However we can improve the computation time by deciding the number of iterations to be done on each feature beforehand. Modifying algorithm resulted in more confidence on impacting parameters.

To improve the accuracy by implementing the PCA (Principal component analysis), however, based on our preliminary experiment, we found that traditional PCA suffers from two important limitations. First, PCA have an assumption that the relationships between variables are linear. Second, PCA is only capable if all of the variables are scaled at the numeric level (interval or ratio scale of measurement). PCA may not be an appropriated method to do analysis on our dataset, which contains a lot of non-numeric and features data. To circumvent these limitations, an alternative, referred to as nonlinear principal components analysis, has been developed. In nonlinear PCA, categorical data will be assigned numeric values through a process called optimal quantification, which is also referred to as optimal scaling, or optimal scoring. Optimal quantification replaces the category labels with category quantifications in the method that as much as possible of the variance in the quantified variables is accounted for [20]. In nonlinear PCA, the optimal quantification process and the linear PCA model estimation are performed together and simultaneously. This is to get the best result for the data reduction. Therefore, our future work will apply the nonlinear principal components analysis in advent project experiment.

B. Risk feature identification.

According to our experiment, we were able to identify a number of risk features impacting readmission, such as, number of times that the patient visited hospital as inpatient in that particular year, emergency visits and Number of lab procedures (0-6) and type of disposition (29). Among the diagnosis features, diagnosis 1 (primary diagnoses) was more important than diagnosis 2 and diagnosis 3 (secondary diagnoses). Time spent in hospital has less effect on readmission. Among all the medication administered, repaglinide and pioglitazone were regarded to be more important, which is proofed based on the literature review. Insulin and Glucose serum test results were down the list in all three experiments, which means less impact on readmission rate. Additionally, our report has shown that A1C test didn't even cross threshold. Such findings might be useful to illustrate a direction for future prevention, however, we were focusing on one single dataset, which is not a longitudinal dataset and even contains incomplete data points (e.g., many attributes were disregarded because of less variance and missing value, described in the section of Data Preperation). In addition, the dataset dose not differentiate the patients with Type 1 Diabetes or Type 2 Diabetes, consequently, our results cannot show the difference between these two types of diabetes. Thus, in the future study, we will apply the same method designed in this study to analyze EHR data at VA, which is a national wide clinical resource and provides more data dimensions to validate and refine our designed method for readmission assessment.

REFERENCES

1. *Statistics about Diabetes*. [cited 2016 August 21]; Available from: http://www.diabetes.org/diabetes-basics/statistics/.
2. *Taking steps in the hospital to prevent diabetes-related readmissions*. [cited 2016 August 21]; Available from: https://americannursetoday.com/taking-steps-in-the-hospital-to-prevent-diabetes-related-readmissions/.
3. Hansen, L.O., et al., *Interventions to reduce 30-day rehospitalization: a systematic review.* Annals of internal medicine, 2011. **155**(8): p. 520-528.
4. Strack, B., et al., *Impact of HbA1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records.* BioMed research international, 2014. **2014**.
5. Shalizi, C., *Advanced data analysis from an elementary point of view*. 2013: Citeseer.
6. Breiman, L., *Random forests.* Machine learning, 2001. **45**(1): p. 5-32.
7. Ho, T.K., *The random subspace method for constructing decision forests.* IEEE transactions on pattern analysis and machine intelligence, 1998. **20**(8): p. 832-844.
8. Freedman, D.A., *Statistical models: theory and practice*. 2009: cambridge university press.
9. Walker, S.H. and D.B. Duncan, *Estimation of the probability of an event as a function of several independent variables.* Biometrika, 1967. **54**(1-2): p. 167-179.
10. Cox, D.R., *The regression analysis of binary sequences.* Journal of the Royal Statistical Society. Series B (Methodological), 1958: p. 215-242.
11. McCullagh, P. and J.A. Nelder, *Generalized linear models*. Vol. 37. 1989: CRC press.
12. Hosmer Jr, D.W., S. Lemeshow, and R.X. Sturdivant, *Applied logistic regression*. Vol. 398. 2013: John Wiley & Sons.
13. Breiman, L., *Bagging predictors.* Machine learning, 1996. **24**(2): p. 123-140.
14. Zhou, Z.-H., *Ensemble methods: foundations and algorithms*. 2012: CRC press.
15. Quinlan, J.R., *Simplifying decision trees.* International Journal of Human-Computer Studies, 1999. **51**(2): p. 497-510.
16. Gini, C., *Concentration and dependency ratios.* Rivista di Politica Economica, 1997. **87**: p. 769-792.
17. Tuccori, M., et al., *Pioglitazone use and risk of bladder cancer: population based cohort study.* bmj, 2016. **352**: p. i1541.
18. Aubert, R., et al., *Rosiglitazone and pioglitazone increase fracture risk in women and men with type 2 diabetes.* Diabetes, Obesity and Metabolism, 2010. **12**(8): p. 716-721.
19. *metformin and repaglinide*. [cited 2016 August 21]; Available from: https://http://www.drugs.com/mtm/metformin-and-repaglinide.html.
20. Linting, M., et al., *Nonlinear principal components analysis: introduction and application.* Psychological methods, 2007. **12**(3): p. 336.

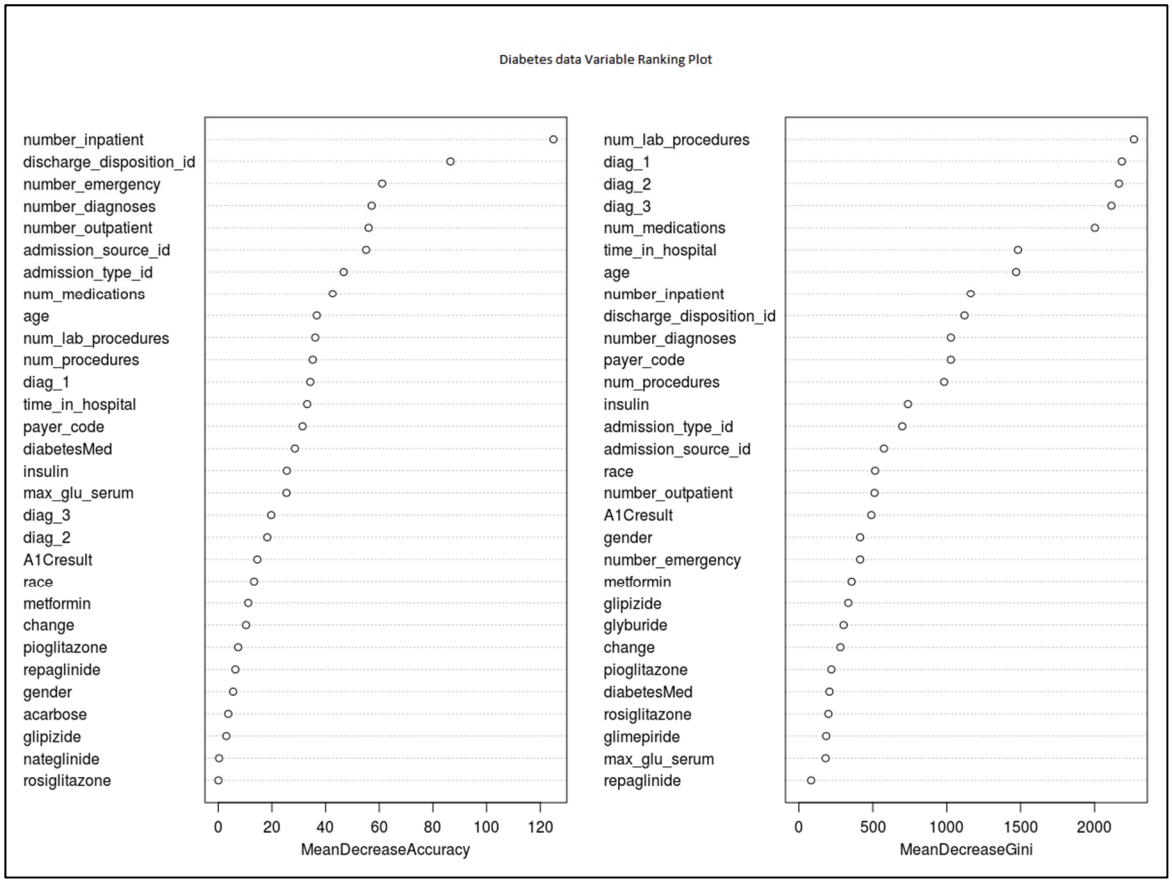

Figure 4. Important feature identification and comparison based on Random Forest models.

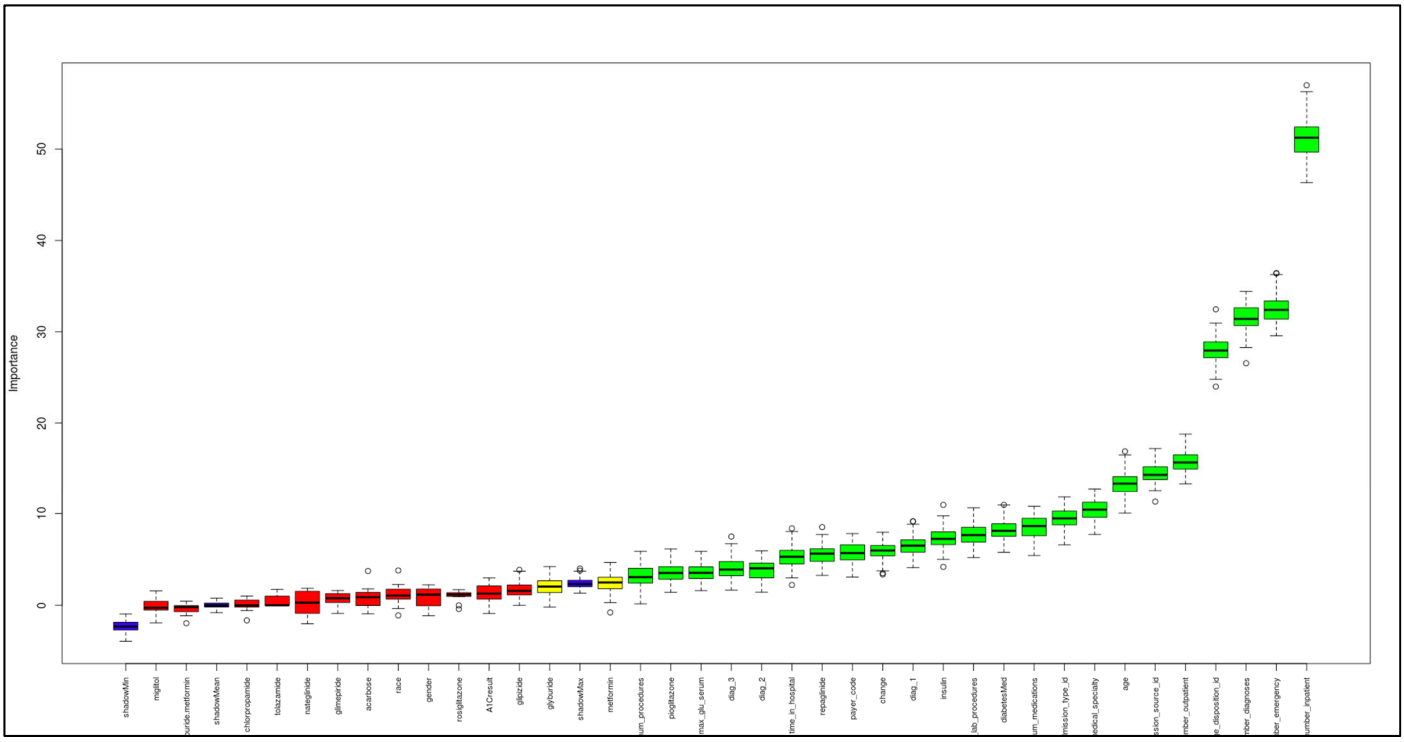

Figure 5. Important feature identification based on Manipulated Random Forest models.

Table 2. Results generated by three machine learning methods.

Logistic regression		Random forest		Modified Random Forest	
number_inpatient	0	number_inpatient	142.3995	number_inpatient	51.05605
number_diagnoses	0	discharge_disposition_id	82.28912	number_emergency	32.44755
number_emergency	0	number_emergency	64.45878	number_diagnoses	31.38095
number_outpatient	0	number_diagnoses	53.39914	discharge_disposition_id	28.0013
diabetesMedYes	0	number_outpatient	49.59221	number_outpatient	15.76404
discharge_disposition_id	0	admission_source_id	49.25453	admission_source_id	14.46891
num_lab_procedures	0	num_medications	38.35412	age	13.31888
max_glu_serumNone	0	age	35.05665	medical_specialty	10.35689
payer_codeSP	0.0001	admission_type_id	34.79478	admission_type_id	9.523341
payer_codeMD	0.0014	diag_1	34.02519	num_medications	8.530853
admission_type_id	0.002	payer_code	32.22958	diabetesMed	8.20103
payer_codeMC	0.0021	num_lab_procedures	31.0295	num_lab_procedures	7.725313
payer_codeCM	0.0054	diabetesMed	30.00931	insulin	7.308996
insulinSteady	0.0164	num_procedures	29.18252	diag_1	6.508905
time_in_hospital	0.0188	time_in_hospital	24.42465	change	5.929467
payer_codePO	0.0383	max_glu_serum	23.50781	payer_code	5.727688
payer_codeDM	0.0481	insulin	23.23505	repaglinide	5.552527
		diag_2	19.18177	time_in_hospital	5.261875
		diag_3	16.39058	diag_3	4.059605
				diag_2	3.859605
				pioglitazone	3.580348
				max_glu_serum	3.544507
				num_procedures	3.226567