Contents

Executive Summary

Hospital readmission for diabetic patients is a major concern in the United States. Over $250 million dollars was spent on treatment of readmitted diabetic inpatients in 2011. This disease is chronic and does not have any specific cure. Hospital readmissions are expensive as hospitals face penalties if their readmission rate is higher than expected and reflects the inadequacies in health care system. Most hospitals can agree that their main goals are centered around improving outcomes, creating more satisfied patients, and better value. For these reasons, it is important for the hospitals to improve focus on reducing readmission rates.

This report attempts to identify the key factors that influence readmission for diabetes and to predict the probability of patient readmission. The dataset used is from UCI Machine learning website and represents 10 years (1999-2008) of clinical care at 130 U.S. hospitals. It contains over 50 (33 character and 17 numeric) variables with attributes such as patient number, demographic characteristics, various type of hospital visits, length of stay, diabetic medications, number of outpatient, inpatient, emergency visits, medications administered and readmitted status. We have trained and tested a range of models to identify the best performing model. We used Decision tress, Random Forest, and Logistic Regression, to discover key factors that explain readmission are number of lab procedures, number of impatient visits, number of medications, time spent in hospital and discharge disposition id. We built a shiny app to predict the probability of readmission. This app may be used as a tool by physician to get a sense of readmission for a patient. As expected, the number of inpatient visits in last one year is the most predictive factor of readmission, but the accuracy of the model is improved as we add more information, especially the historical data, Admission type, admission source and discharge disposition. The results obtained by random forest and decision tree on the actual imbalanced data set are the same as previous research on this data set, as evidence towards the predictors available are not encompassing the true cause of readmission. However, we used synthetic data balancing techniques and improved model accuracy by 12%.

1. Introduction

As per the Center for Disease Control and Prevention, diabetes is the 7th leading cause of death in the United States. 9.3% of Americans are diagnosed with diabetes every year. Diagnosis for diabetes has quadrupled in the past 25 years in the US. The world health organization predicts that the disease will steadily increase worldwide over the next thirty years. Diabetes presently costs the US $245 billion every year with the hospital inpatient care being the greatest expense at $105 billion (Figure 1).

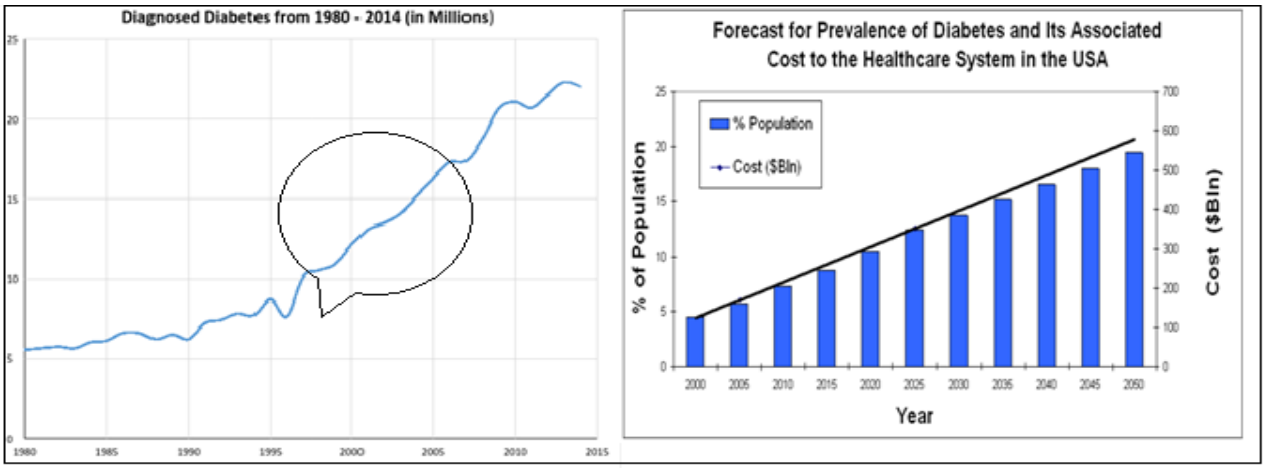


Figure 1. Diabetes Forecast

Hospitals that contract with Medicare to provide inpatient care, agree to accept predetermined payments for that care. The payment rates are standardized by the patient's clinical condition and adjusted by local market conditions.

Two programs under Obamacare penalize hospitals for poor performance. The Hospital-Acquired Condition (HAC) Reduction Program penalizes hospitals in the worst-performing quartile by paying 99% of the normal payment. The Readmissions Reduction Program (HRRP) can penalize hospitals up to 3% of payments if expected readmission standards are not met. Since these rules may be determined arbitrarily, it is in the hospital's best interest to remain competitive in these metrics.

In this report, our objective is to obtain insight into UCI 10-year diabetes dataset by identifying the key factors associated with readmission. This report covers results obtained from exploratory data analysis and a supervised machine learning algorithms that were used to predict the probability of readmission as a function of specific variables.

Some of the questions our analysis and model answered:

- Can we improve accuracy(AUC) through data balancing techniques?
- Is H1Ac* result a good predictor of readmission as compared to glucose serum test?
- Does age factor into probability of readmission?
- Does number of procedures, medication, and lab procedures correlate with the readmission?
- Does admission type, admission source and discharge disposition have a significant impact on readmission?

*Also known as Hb1Ac and A1C

2. Literature Review

Numerous studies related to analyzing the risk factors that predict readmission rates for diabetic patients have been conducted by several scholars. Diabetes patient readmission rate has been a growing trend in the U.S. and researchers are eager to determine predictors for readmission of diabetic patients since hospital readmissions are expensive. In 2011, $250 million was spent amongst 23,700 total diabetes readmission patients [1]. This turns out to be about $10,500 per person that is readmitted into the hospital for diabetes. When a patient is readmitted, the hospitals must use additional resources, in addition, the hospital compromises its integrity by personifying inadequate care that threatens a patient's life. The papers below were used to provide our team direction into various modeling approaches we should consider when designing our model.

The [2] paper, proposes the idea to build a binary classification of the dataset in two manners for use in the model. The first classification method combines patients that were readmitted after 30 days, as it is presumed that they share similar characteristics to patients that were not readmitted (<30 versus (>30 & No)). The second classification combines all readmission patients versus patients not readmitted (<30 & >30) versus 'No'). When using a Precision-Recall curve, the second classification method performed better amongst different algorithms than the first classification. The different algorithms used included: Naives Bayes, Bayesian Networks, Random Forest, Adaboost, and Neural Networks. The area under the Precision-Recall curve performed best when using a Neural Network. Area under the curve was 65%.

The [3] paper, used the following models to determine which model provided the highest prediction accuracy: Knearest Neighbor Binary Classification (average accuracy prediction around 60%), Support Vector Machine Binary Classification on Readmission Rate (average prediction accuracy 61%), and Decision Tree Model (average prediction accuracy 54.6%). The Support Vector Machine performed the best. Additionally, this paper stated that the significant categorical features included the categorical variables: discharge disposition, admission source, and admission type.

A multi-label classification method approach was used in the [4] paper. The multi-labeled classification algorithm predicts the ranking of all labels. Based on results from this classification method, the study focused on learning about patients through the variables related to race and gender. Other features that deemed to be important included demographics, diagnoses, diabetic medication, number of visits, and payment details. The models tested were the following: Random Tree Model, Decision Tree, K Nearest Neighbor, Naives Bayes, Hoeffding Tree, and JRIP. JRIP performed the best with a prediction accuracy of 70%, followed by Hoeffding Tree, which had a prediction accuracy of 70%.

In a final report [5], according to the dataset, the HbA1c test, an administered test that states a person's average blood sugar, suggest that simply measuring a patients HbA1c was associated with a lower rate of readmission in individuals with diabetes, regardless of the outcome of the test. This test provides readings in percentages that ranges from 4% to 14% [6]. A reading between 5.7% and 6.4% mean you have a higher chance of getting diabetes, which explains why 55% patients in the dataset that were administered this test had a change in medication. Intuitively, the results help medical professional adjusted diabetes medication to be appropriate for the patient's blood sugar level provided by the test. Though HbA1c appears to be a good predictor in diabetic readmission, the paper states that the test is ordered infrequently and is considered as a limitation to the interpretation of the data. The model used for these conclusions was a multivariate logistic regression model.

Most papers provided ideas that we can use to implement and improve our model. In addition, they provided similar insight which reinforces variable contributions. Unanimously, all papers stated that the following variables had a large percentage of missing values and therefore the variables were removed from all the datasets: weight (97%), payer code (40%), and medical specialty (47%). Moreover, all papers used a dataset that covered a 10-year clinical period dating from 1999 to 2008 data. Ideally, the model we create will have a better accuracy than the previous studies done and using different techniques of data balancing using ROSE R package.

3. Motivation and Background

Hospital readmission for diabetic patients is a major concern in the United States. Over $250 million dollars was spent on treatment of readmitted diabetic in-patients in 2011 [2]. This disease is chronic and does not have any specific cure. Severity varies from person to person depending on the symptoms and levels of blood sugar in human body. This project seeks to identify key factors that influence readmission of diabetic in-patients to improve quality of care, and reduce threat to life to reduce medical expenses.

Hospital readmissions are expensive as hospitals face penalties if their readmission rate is higher than expected rate and reflect the inadequacies in health care system. Rates vary between 0% to 3% of the normal reimbursements rate for hospitalizations [7]. This is an interesting topic for our research because most hospitals can agree that their main goals are centered around improving outcomes, creating more satisfied patients and better value. However, re-admittance issues appear to be a large concern amongst not only the hospital, but for the patient and insurance providers as well. This is a real-life problem that can potentially help hospitals improve not only their bottom line but also reduce the chances of readmission because the well-being of a human's life is very important.

Over 30 million Americans have been diagnosed with diabetes. With a large and increasing number of diabetes patients, readmissions to hospitals becomes expensive for both the patient and hospital. Additionally, a large percentage of readmissions is reflected as an inadequacy in the healthcare system. Our research will aid hospitals in building a system that focuses on the reduction of preventable in-patient diabetic care patients. In addition, it will help reduce the medical expenses of readmissions.

4. Dataset

Our Dataset was taken from UCI Machine Learning Website. It is comprised of diabetic clinical care patient data collected over a period of ten years (1998 – 2008) from hospitals across the United States. The patient data is represented by 101,767 patient records that have been privacy preserved in accordance with the Health Insurance Portability and Accountability Act of 1996 to prevent patient information leakage.

The dataset contains fifty variables: thirty character and seventeen numeric. We selected the patient's readmission status as our dependent, character variable. Initially, readmission status had three levels – readmission in less than 30 days, readmission in greater than or equal to 30 days, and never readmitted. We converted three levels to two levels as focus of this report is to predict readmission (Figure 2). Technically readmission is defined as an admission within 30 days of a discharge.

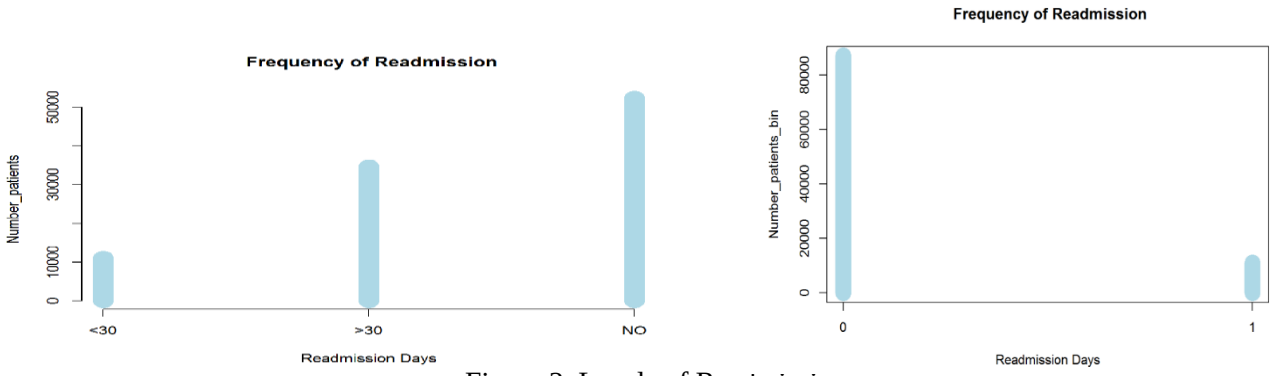


Figure 2. Levels of Readmission

4.1 Data Selection Rationale

The division of the response variable poses the problem of a slightly imbalanced dataset as there are approximately 55,000 patients that were not readmitted, and approximately 11,000 patients who were readmitted before 30 days, and 35,000 patients who were readmitted after 30 days (Figure 3). Much past research has been done utilizing this split of the dataset, however, we chose to form our hypotheses from the perspective of the Affordable Care Act's formula to calculate the readmission adjustment factor, in which hospital reimbursement from Medicare/Medicaid will be reduced with patients that are readmitted to the hospital within thirty days of the prior admission for diabetic related circumstances. In doing so, we re-split the dataset in to two levels – patients readmitted under 30 days, and all other patients, as the later will not be the subject of penalty. The new division of the patients has imbalanced the dataset even further – approximately 12,000 readmitted less than 30 days (11%) as compared to approximately 90,000 (89%) not readmitted in less than 30 days.

Readmission - captured in data			Readmission - based on Obama care	
Readmission in Days	Number of encounters	% of Encounters	Criteria based on Obama care	Readmission - Yes/NO
<30	11,357	11%	YES	11%
>30	35,545	35%	NO	89%
NO	54,864	54%		

Figure 3. Classification Slicing

4.2 Dataset Description

We divided the dataset's 50 variables into 6 clusters based on the description of the data entered at the point of patient' admission: (1) patient information – demographics such as race, gender, age, etc., (2) number of procedures performed – procedures were done in a laboratory setting or they were not, (3) previous 12 month history – number of outpatient, inpatient and emergency room visits the patient had throughout the year before the initial admission, (4) Medical Diagnosis – there were three separate ICD-9 (International Classification of Disease Diagnosis) diagnosis codes given to a patient; the primary, secondary and tertiary, (5) Tests – there are two main tests that inform the patient whether or not they have diabetes, and (6) Medication features – did the patient's dosage levels of current medications go up, down, stay steady, etc. (Figure 4 & 5)(Appendix B).

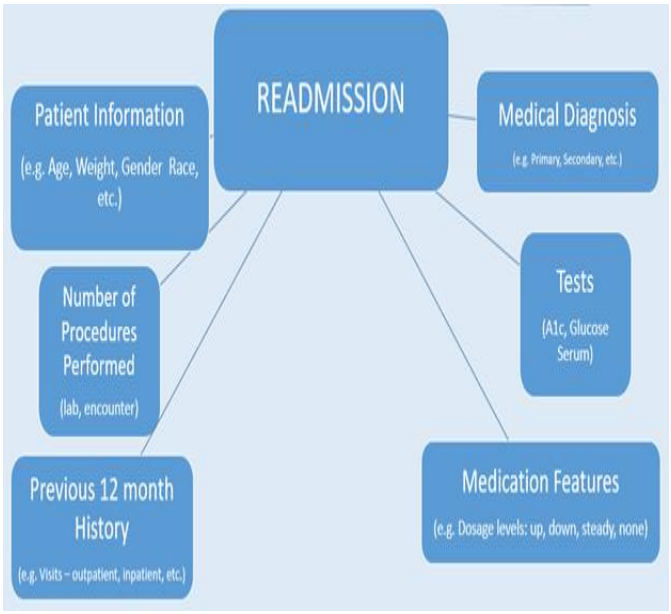


Figure 4. Key Categories of Variables

List of features and their descriptions in the initial dataset			
Feature name	Type	Description and values	% missing
Encounter ID	Numeric	Unique identifier of an encounter	0%
Patient number	Numeric	Unique identifier of a patient	0%
Race	Nominal	Values: Caucasian, Asian, African American, Hispanic, and other	2%
Gender	Nominal	Values: male, female, and unknown/invalid	0%
Age	Nominal	Grouped in 10-year intervals: 0, 10), 10, 20), ..., 90, 100)	0%
Weight	Numeric	Weight in pounds.	97%
Admission type	Nominal	Integer identifier corresponding to 9 distinct values, for example, emergency, urgent, elective, newborn, and not available	0%
Discharge disposition	Nominal	Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available	0%
Admission source	Nominal	Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital	0%
Time in hospital	Numeric	Integer number of days between admission and discharge	0%
Payer code	Nominal	Integer identifier corresponding to 23 distinct values, for example, Blue Cross/Blue Shield, Medicare, and self-pay	52%
Medical specialty	Nominal	Integer identifier of a specialty of the admitting physician, corresponding to 84 distinct values, for example, cardiology, internal medicine, family/general practice, and surgeon	53%
Number of lab procedures	Numeric	Number of lab tests performed during the encounter	0%
Number of procedures	Numeric	Number of procedures (other than lab tests) performed during the encounter	0%
Number of medications	Numeric	Number of distinct generic names administered during the encounter	0%
Number of outpatient visits	Numeric	Number of outpatient visits of the patient in the year preceding the encounter	0%
Number of emergency visits	Numeric	Number of emergency visits of the patient in the year preceding the encounter	0%
Number of inpatient visits	Numeric	Number of inpatient visits of the patient in the year preceding the encounter	0%
Diagnosis 1	Nominal	The primary diagnosis (coded as first three digits of ICD9); 848 distinct values	0%
Diagnosis 2	Nominal	Secondary diagnosis (coded as first three digits of ICD9); 923 distinct values	0%
Diagnosis 3	Nominal	Additional secondary diagnosis (coded as first three digits of ICD9); 954 distinct values	1%
Number of diagnoses	Numeric	Number of diagnoses entered to the system	0%
Glucose serum test result	Nominal	Indicates the range of the result or if the test was not taken. Values: ">200," ">300," "normal," and "none" if not measured	0%
A1c test result	Nominal	Indicates the range of the result or if the test was not taken. Values: ">8" if the result was greater than 8%, ">7" if the result was greater than 7% but less than 8%, "normal" if the result was less than 7%, and "none" if not measured.	0%
Change of medications	Nominal	Indicates if there was a change in diabetic medications (either dosage or generic name). Values: "change" and "no change"	0%
Diabetes medications	Nominal	Indicates if there was any diabetic medication prescribed. Values: "yes" and "no"	0%
24 features for medications	Nominal	For the generic names: metformin, repaglinide, nateglinide, chlorpropamide, glimepiride, acetohexamide, glipizide, glyburide, tolbutamide, pioglitazone, rosiglitazone, acarbose, miglitol, troglitazone, tolazamide, examide, sitagliptin, insulin, glyburide-metformin, glipizide-metformin, glimepiride-pioglitazone,	0%
Readmitted	Nominal	Days to inpatient readmission. Values: "<30" if the patient was readmitted in less than 30 days, ">30" if the patient was readmitted in more than 30 days, and "No" for no record of readmission.	0%

Figure 5. Variable Description

4.3 Dataset Cleaning

We encountered many data challenges while cleaning the data. One of the main challenges was the high vertical and horizontal dimensionality of the dataset. The first step we took in reducing these dimensions was to ignore the non-essential variables (patient number, encounter, ID., etc.) in the design of our model, as the information contained would hold zero significance in the probability of a patient's readmission status. Next, we focused on analyzing the impact of missing data points from the variables. Some missing data was represented by a question mark in the dataset. We chose to reduce horizontal dimensionality of the dataset first. Weight was missing 97% of its records, and therefore it was removed from our model. (Table 1). We were disappointed to take this course of action, because there are many cases that knowing a patient's weight can help either reverse diabetic conditions or play an important factor in preventing the onset of certain types of diabetes. The National Institute of Diabetes and Digestive and Kidney Disease has published a report stating that A1C tests can result in a different diagnosis than the blood glucose tests. In some people, an A1C test may render a diagnosis of diabetes when the blood glucose test does not, and vice versa. Individuals with conflicting test results may be in an early stage of the diabetes, where their blood glucose levels are still too low to show up on every test. Sometimes, making tiny alterations in régime, such as losing a small amount of weight and increasing physical activity, can assist people in the early stage reverse diabetes or postpone its onset. The next attribute removed was medical specialty with 49% of its data missing. Lastly, we removed the payer code variable, as 40% of the records were missing.

Data	% Missing	Action
Weight	97%	Removed Attribute
Medical Specialty	49%	Removed Attribute
Payer Code	40%	Removed Attribute
Race	2%	Deleted 2273 Missing Records
Diagnosis 3	1%	Deleted 1423 Missing Records
Diagnosis 2	0%	Deleted 358 Missing Records
Diagnosis 1	0%	Deleted 21 Missing Records

Table 1. Variables with Large Number of Missing Observations

Vertical dimensionality was quite ludicrous for numerous variables in the dataset. We had to be extremely careful when deciding whether to clean, recode, or remove the variable so that we didn't lose any levels or interactions that may affect the probability of readmission. We used logistic regression, Decision Tree, and Random Forest which allowed for the addition of both multi-level character and numeric variables into our model. The Diagnosis 3, 2, and 1 contained 954, 923, and 848 levels respectively and contained only the first three digits of the patient's diagnosed ICD-9 code, which is the category of the diagnosis, and such a vague interpretation of the actual diagnosis that we could not use in our analysis due to high number of levels (Table 2). Another high vertical dimensionality challenge was payer code, also turned into a coding challenge. The variable payer code, which was missing 40% of its data, also had no key for us to refer to in order for us to understand what any of the codes represented, so we did not use in the model.

DATA	NUMBER OF LEVELS	TYPE OF VARIABLE
Diagnosis 3	954	Character
Diagnosis 2	923	Character
Diagnosis 1	848	Character
Medical Specialty	84	Character
Discharge Disposition	29	Character
Payer Code	23	Character
Admission Source	21	Character
Age	10	Numeric
Admission Type	9	Character

Table 2. Vertical dimensionality of Variables

5. Exploratory Data Analysis

Exploratory data analysis was used to summarize main characteristics that could be of use to understand the readmission attributes. We used Table function in R to analyze the data.

5.1 Preliminary Data Overview

Figure 6 illustrates the number of encounters by age. We observed that data is skewed but since it is a categorical variable we didn't standardized or normalized it.

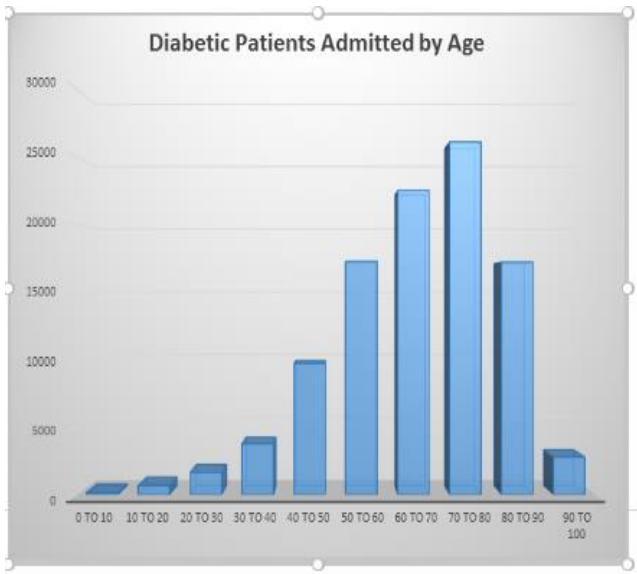


Figure 6: Encounters by age

In absolute numbers

	[0-10)	[10-20)	[20-30)	[30-40)	[40-50)	[50-60)	[60-70)	[70-80)	[80-90)	[90-100)
No Readmission	64	435	1265	3140	8265	15059	19349	22302	14690	2418
Readmission	1	31	213	408	1000	1638	2460	3004	2012	299

In percentage terms

	[0-10)	[10-20)	[20-30)	[30-40)	[40-50)	[50-60)	[60-70)	[70-80)	[80-90)	[90-100)
No Readmission	98.5%	93.3%	85.6%	88.5%	89.2%	90.2%	88.7%	88.1%	88.0%	89.0%
Readmission	2%	7%	14%	11%	11%	10%	11%	12%	12%	11%

Table 3. Distribution of Readmission by Age

We observed that the distribution of readmission and no readmission is not uniform for age. Surprisingly, it is quite high for the age group 20-30. (Table 3)

5.2 Distribution of Readmission by Race

In absolute numbers

	African American	Asian	Caucasian	Hispanic	Other
No Readmission	16,741	560	66,567	1,777	1,342
Readmission	2,140	65	8,512	207	142

In Percentage

	African American	Asian	Caucasian	Hispanic	Other
No Readmission	89%	90%	89%	90%	90%
Readmission	11%	10%	11%	10%	10%

Table 4. Distribution of Readmission by Race

We observed that the distribution of readmission and no readmission is uniform across all races.

5.3 Distribution of Readmission by Gender

In absolute numbers

	Female	Male
No Readmission	46831	40155
Readmission	6002	5064

In Percentage

	Female	Male
No Readmission	88.6%	88.8%
Readmission	11.4%	11.2%

Table 5. Distribution of Readmission by Gender

5.3 Distribution of Readmission by number of days spent in hospital

The graph below indicates that readmission rate increase with time spent in hospital. (Figure 7)

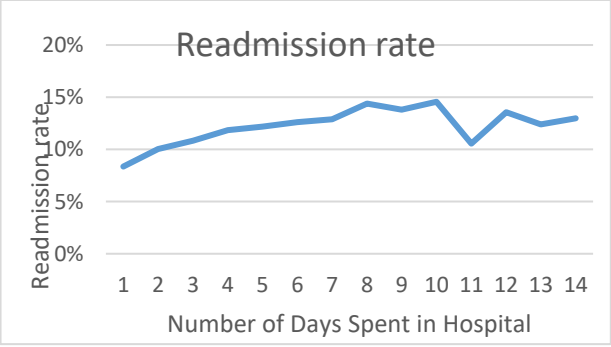


Figure 7. Distribution of Readmission by Number of Days Spent in Hospital

6. Data Analysis and Ensemble Methods

We explored ensemble-based classifiers with the assumption that these groupings of weak learners would help improve the accuracy metrics. Ensembles create a set of classifiers and then classify a test data point by taking a weighted average of the individual "weak" models predictions. We used random forest and gradient boosted decision trees.

Since the data is imbalanced, it affects the model performance considerably, as the threshold for class prediction can no longer remain at the default value of 50%. Two ways by which we explored an appropriate threshold for this exercise are:
a) Manually inspection of various values of threshold for a good mix of accuracy and recall/ sensitivity.
b) Use the cost penalty matrix to determine a good threshold (Appendix C)

6.1 Interpreting Unbalanced Data Result

Most classification algorithms calculate accuracy based on the percentage of observations correctly classified. With imbalanced data, the results are highly deceiving since minority classes hold minimum effect on overall accuracy. The most frequently used metrics are Accuracy & Error Rate.

$$\text{Accuracy: } (TP + TN)/(TP+TN+FP+FN)$$
$$\text{Error Rate} = 1 - \text{Accuracy}$$

In medical diagnosis, test sensitivity is the ability of a test to correctly identify those with the disease (true positive rate), whereas test specificity is the ability of the test to correctly identify those without the disease (true negative rate).

Sensitivity is defined as:
$$\text{Sensitivity} = TP/ (TP + FN)$$

TP is the number of true positives and FN is the number of false negatives. Sensitivity is the fraction of readmissions that are actually positive that were predicted as positive by the model. Since the more positive-skewed the dataset is the higher the sensitivity, we also need to look at specificity, defined as:

$$\text{Specificity} = TN/ (TN + FP)$$

where TN is the number of true negatives and FP is the number of false positives. Specificity is the fraction of readmission status that are actually negative, that were predicted as negative by the model.

Recall or Sensitivity: It is a measure of actual observations which are labeled (predicted) correctly i.e. how many observations of positive class are labeled correctly. It is also known as 'Sensitivity'.

$$\text{Recall} = TP / (TP + FN).$$

Technique	Overall Accuracy for three levels (<30, >30, and no readmission	Overall Accuracy with two levels
Decision Tree	58%	90%
Random Forest	58%	89%

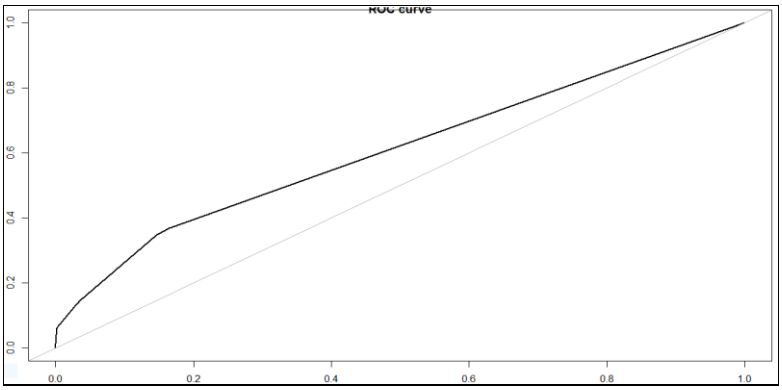
Table 6. Result based on accuracy metrics

The above table indicate the accuracy of 90%, but it is not a right indicator as sensitivity is very low.(Table 6)

		Prediction		
		0(Negative)	1(Positive)	
Actual	0(Negative)	17,364(TN)	34(FP)	
	1(Positive)	2077(FN)	137(TP)	
Accuracy	89.2%			
Specificity	99%			
Sensitivity	6.6%			

Table 7. Confusion matrix

These metrics may provide deceiving results and are highly sensitive to changes in data. Further, various metrics can be derived from confusion matrix (Table 7). Fortunately, we have a ROC (Receiver Operating Characteristics) curve to measure the accuracy of a classification prediction. We examined ROC Curve by plotting TP rate (Sensitivity) and FP rate (Specificity). (Figure 8)



The low accuracy rate (60.4%) of the model signify either of the two reasons: -

a) The model is not accurately predicting due to imbalanced data.
b) There are missing factors in the data such as socio economic status

Figure 8. ROC Curve Imbalanced Data

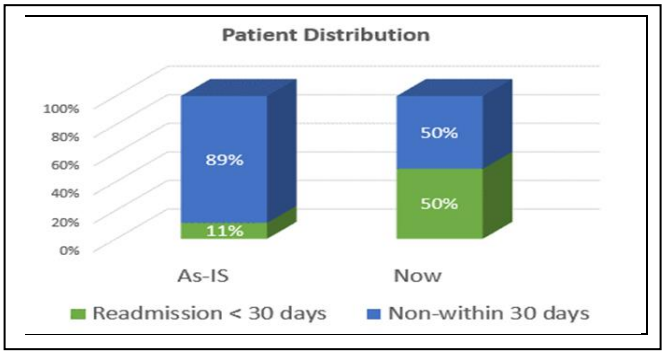
6.2 Approach to handle Imbalanced data

The term imbalanced refers to the disparity encountered in the dependent (response) variable. Therefore, an imbalanced classification problem is one in which the dependent variable has imbalanced proportion of classes. With imbalanced data sets, an algorithm doesn't get the necessary information about the minority class to make an accurate prediction. In this diabetes data set we have only 11% readmissions under

30 days and about 89% readmissions were after 30 days or were no readmissions. This binary classification made data highly imbalanced (Figure 9).

We used four data balancing techniques using the "ROSE" package to balance the data:

1. Under sampling
2. Oversampling
3. Balanced sampling
4. Synthetic Data balancing



FFigure 9. Dependent Variable Data Balancing

"ROSE" package provides functions to deal with binary classification problems in the presence of imbalanced classes. Artificial balanced samples are generated per a smoothed bootstrap approach and allow for aiding both the phases of estimation and accuracy evaluation of a binary classifier in the presence of a rare class.

1. Under sampling- This method works with majority class and decreases the number of observations from majority class to make the data set balanced. Apparently, removing observations may cause the training data to lose important information pertaining to majority class. With under sampling AUC under ROC curve increased from 60%(imbalanced sample) to 62%. (Balanced sample)

2. Over sampling- This method works with minority class and repeats the observations from minority class to balance the data. An benefit of using this method is that it leads to no information loss. The disadvantage of using this method is that, since oversampling simply adds replicated observations in original data set, it ends up adding multiple observations of several types, thus leading to overfitting. With over sampling AUC under ROC curve increased from 60%(imbalanced sample) to 63%. (balanced sample)

3. Both under and over sampling-With both technique method AUC under ROC curve increased from 60%(balanced sample) to 62%. (balanced sample)

4. Synthetic Data balancing- SMOTE algorithm creates artificial data based on feature space (rather than data space) similarities from minority samples. To generate artificial data, it uses bootstrapping and k-nearest neighbors.

ROC Curves for Comparison of different data balancing approach.

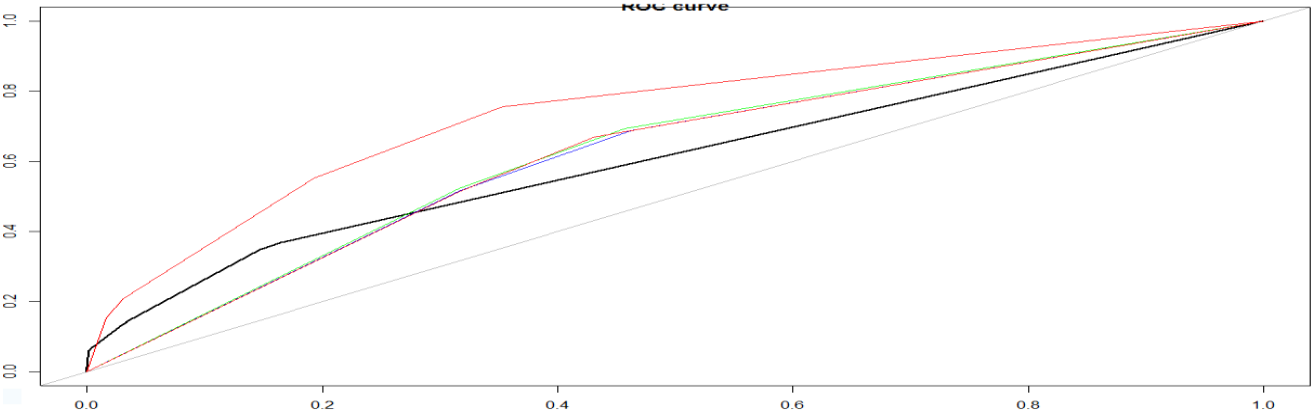


Figure 10: ROC Curves for Comparisons

Rose package	No of observations	AUC	Color
Synthetic Balancing using ROSE	9848	73%	Red
Both	9847	62%	Light red
Under sampling	2214	62%	Blue
Oversampling	17398	63%	Green
Unbalanced	19612	60%	Black

Table 8. AUC for Different Balancing Methods

The average accuracy rate (72.4%) of the model even on a balanced data set can signify that there are missing factors in the data (such as socio economic status) (Table 8).

6.3 Key Predictors

Based on Random Forest results, the mean decrease accuracy and mean decrease Gini, the top 5 key predictors for readmission for our dataset are: number of lab procedures, number of impatient visits, number of medications, time spent in hospital, and discharge disposition id. (Figure 11)

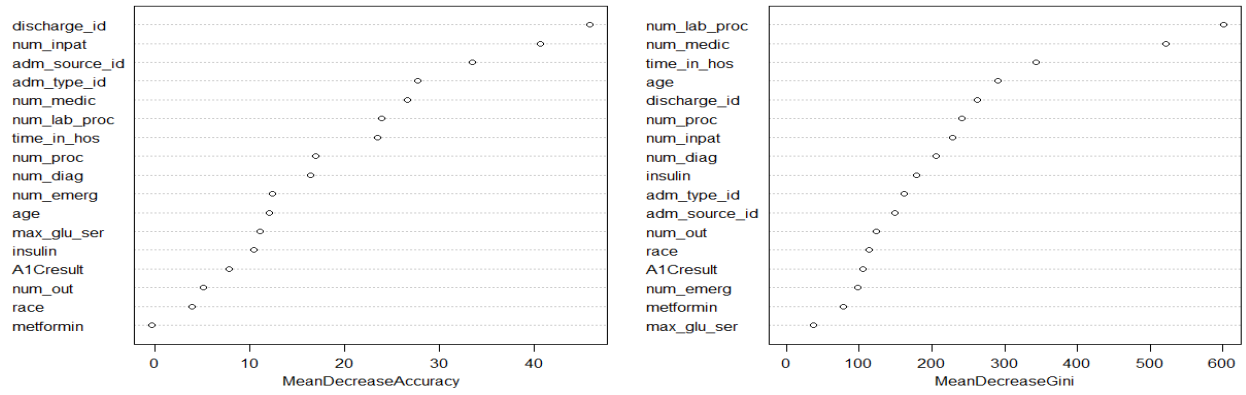


Figure 11. Random Forest Variable Selection

The following plot shows the error rate (Figure 12)

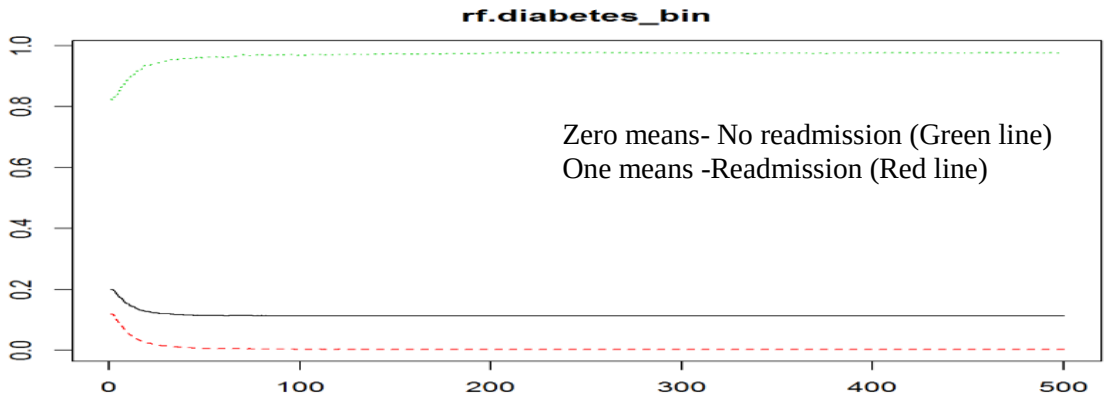


Figure 12. Error Rate

7. Key Insights

In our exploratory data section, we posed numerous questions about possible key predictors that would benefit our model. As per our model, we discovered that number of inpatient visits was the root node for the decision tree, see Appendix A. It makes sense that if the patient has a history of chronic health issues that would require readmission they might be susceptible to it for diabetes as well. However, it would be interesting to know what the inpatient visits were for each patient and if those reasons were correlated to a diagnosis of diabetes. Number of patient' procedures was also a key predictor in determining the probability of being readmitted. Patient procedures were a significant contributor of readmissions. Previously, we questioned the importance of the two tests for diabetes, A1C and blood glucose. When considering these tests, our model confirmed that the A1C test is also a possible explanatory variable compared to blood glucose test. The A1C test result is reported as a percentage - the higher a person's blood glucose levels have been, the higher the percentage. A normal A1C level is below 5.7 percent (Table 9). It was also determined that age and gender were not good predictors.

Diagnosis	A1C Level
Normal	Below 5.7 percent
Diabetes	6.5 percent or above
Prediabetes	5.7 to 6.4 percent

Table 9. A1C Levels

8. Deployment

We built a Shiny app that provides percentage probability of readmission. The app is geared for hospital personnel to determine if they should spend extra resources, such as additional tests or observations after discharge, on a patient based on that patient's percentage of readmission that the app calculated. The individual hospitals will have to determine what cutoff percentage is adequate to pursue additional resources on a patient.

We hope the results of the app will prevent future readmission and reduce penalties. The hospital personnel inputs patient information on drop down menu items located on the left side panel, as shown in Figure 13. Two attributes, discharge ID and admission source ID, had a large number of inputs with long explanations and therefore separate tabs were created for them to be able to view what each code represents, see Figure 14. The attributes to select on the app are the variables used in the Decision Tree model.

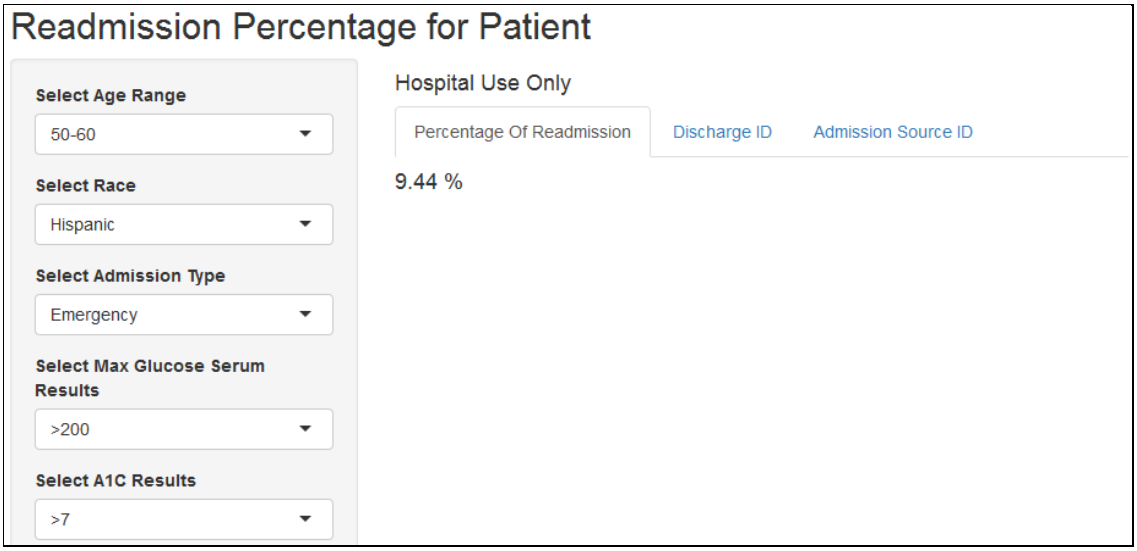


Figure 13. Shiny App: Percentage of Readmission

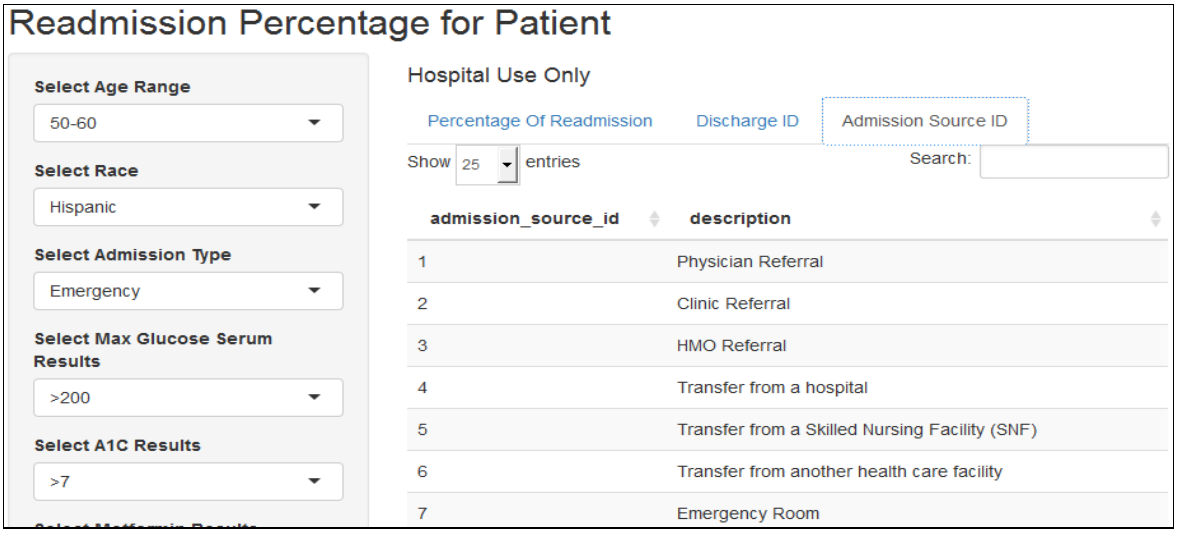


Figure 14. Shiny App: Admission ID tab

This app can be expanded to include average cost of a patient if that patient exceeds the percentage threshold set by the individual hospital's so they can have a better judgement of the individual contributions to the existing hospital budget.

9. Key Recommendations

Based on the results of supervised leaning techniques including Decision trees and Random Forest, we have following recommendations:

1. Granular historical data collection- Hospitals should collect granular data points including number of previous impatient visits, emergency visits etc. as these data points are key good predictors of readmission
2. Real time data collection- Hospitals should also preserve real time data points like admission source, admission type, weight of patient at the time of admission etc.
3. Hospitals are advised to not only include inpatient treatment but also continue care after discharge to better monitor patients and prevent possible readmission.
4. Include demographic and socio-economic data as predictors as these can cause an impact in model and emphasis on weight data

10. Reference

[1] B. Herman. The Costs of 10 Top Medicaid Readmission Conditions http://www.beckershospitalreview.com/finance/the-costs-of-10-top-medicaid-readmission-conditions.html *Becker's Healthcare.*2014

[2] B. M S. A. Zafar, V. Kishore. Identifying Diabetic Patients with High Risk of Readmission. *National Institute of Technology Karnataka, Surathkal, India.* 2016

[3] X. Yifan, J. Sharma. Diabetes Patient Readmission Prediction Using Big Data Analytic Tools.

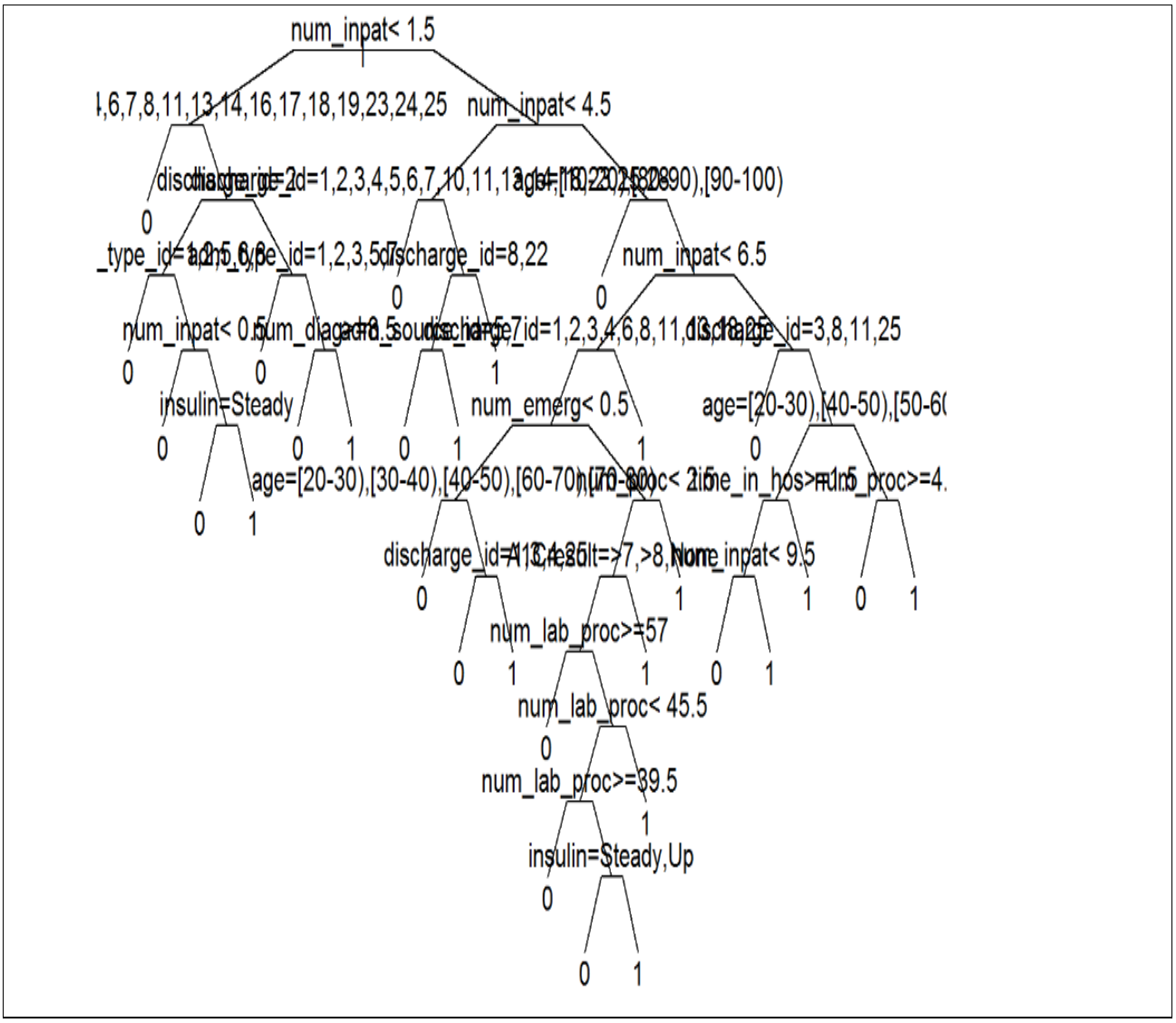
[4] N. Cotha, M. Sokolova. Multi-labeled Classification of Demographic Attributes of Patients: A case Study of Diabetic Patients. *University of Ottawa, Institute for Big Data Analytics*.

[5] B. Strack, J. DeShazo, C. Gennings, J. Olmo, S. Ventura, K. Cios, J. Clore. Impact of HbA1c measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records.*BioMed Research International.* 2014

[6] WebMD. Hemoglobin A1c (HbA1c) Test for Diabetes, http://www.webmd.com/diabetes/guide/glycated-hemoglobin-test-hba1c. *WebMD.* 2016

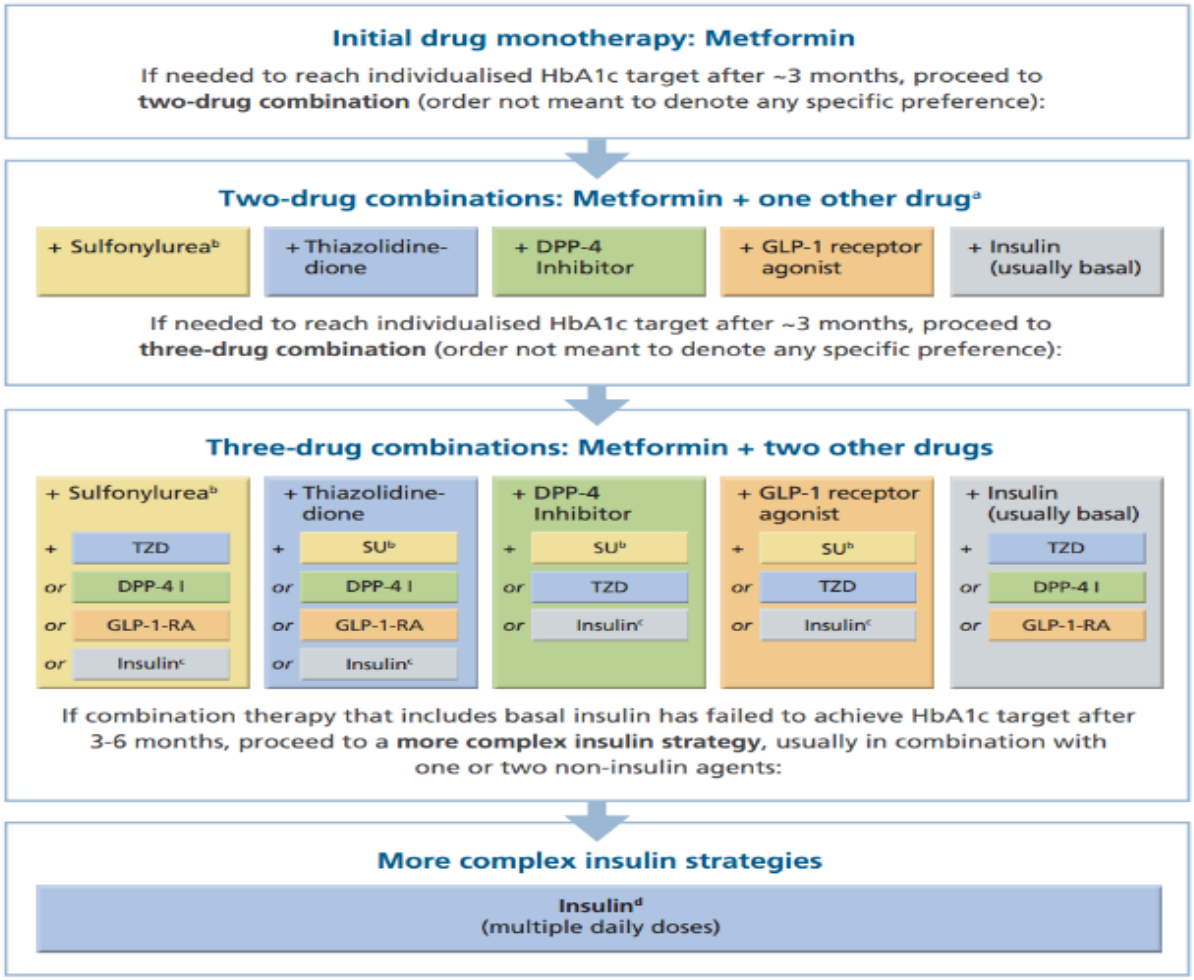
[7] Q.Phillips. Preventing Hospital Readmissions. http://www.diabetesselfmanagement.com/blog/preventing-hospital-readmissions. *Diabetes Self Management*. 2014

APPENDIX A: Tree Structure using R-part



APPENDIX B: Diabetes Drugs

Table 1: Treatment Approach



APPENDIX C: Cost Matrix

COST PENALTY MATRIX		PREDICTED CLASS	
		Readmission	**No Readmission**
ACTUAL CLASS	**Readmission**	1. Benefit of predicting a right readmission	2. Cost incurred by hospital when the model predicts 'no readmission' but a patient is readmitted
	No Readmission	3. Cost incurred when a patient isn't readmitted even though the model predicts s/he will	4. Benefit of predicting a right 'no readmission'