Predicting Diabetic Patient Hospital Readmission Rate: Analysis of 100,000 Clinical Database Patient Records

Presented By:

Kaushiki Chauhan

Data Science

Rowan University

Glassboro, New Jersey

Supervised By:

Dr. Umashanger Thayasivam

1. INTRODUCTION

Diabetes is a chronic disease where a person suffers from an extended level of blood glucose in the body. It may be affected by height, race, gender, age but a major reason is a sugar concentration. The present analysis of a large clinical database was undertaken to examine historical patterns in patients with diabetes admitted to a US hospital and to inform future directions which might lead to improvements in patient safety. The Health Facts database (Cerner Corporation, Kansas City, MO), a national data warehouse that collects comprehensive clinical records across hospitals throughout the United States.

Databases of clinical data contain valuable but heterogeneous and difficult data in terms of missing values, incomplete or inconsistent records, and high dimensionality understood not only by number of features but also their complexity. [8]. Additionally, analyzing external data is more challenging than analysis of results of a carefully designed experiment or trial because one has no impact on how and what type of information was collected. Nonetheless, it is important to utilize these huge amounts of data to find new information/knowledge that is possibly not available anywhere.

2. DATA DESCRIPTION

The Health Facts data we used was an extract representing 10 years (1999–2008) of clinical care at 130 hospitals and integrated delivery networks throughout the United States: Midwest (18 hospitals), Northeast (58), South (28), and West. Most of the hospitals (78) have bed size between 100 and 499, 38 hospitals have bed size less than 100, and bed size of 14 hospitals is greater than 500.

The data represents integrated delivery network health systems in addition to stand-alone hospitals, the data contains both inpatient and outpatient data, including emergency department, for the same group of patients. However, data from out-of-network providers is not captured.

The dataset was created in two steps. First, encounters of interest were extracted from the database with 55 attributes. This dataset is available as a Supplementary Material available online at http://dx.doi.org/10.1155/2014/781670 and is also in the process of submission to the UCI Machine Learning Repository [9] so that it is easily available toother researchers.

Second, preliminary analysis and preprocessing of the data were performed resulting in retaining only these features(attributes) and encounters that could be used in further analysis, that is, contain sufficient information.

101,766 encounters were identified to fulfill all of the above five inclusion criteria and were used in further analysis. From the information available in the database, we extracted 50 features describing the diabetic encounters, including demographics, diagnoses, diabetic medications, number of visits in the year preceding the encounter, and payer information. The full list of the features and their description is provided in Table 1.

Table 1: List of features and their descriptions in the initial dataset (the dataset is also available at the website of DataMining and Biomedical Informatics Lab at VCU (http://www.cioslab.vcu.edu/)).

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	Categorical	Values: Caucasian, Asian, African American, Hispanic, and other	2%
Gender	Categorical	Values: male, female, and unknown/invalid	0%
Age	Categorical	Grouped in 10-year intervals: [0, 10), [10, 20), . . ., [90, 100)	0%
Weight	Numeric	Weight in pounds	97%
Admission Type	Categorical	Integer identifier corresponding to 9 distinct values, for example, emergency, urgent, elective, newborn, and not available	0%
Discharge Disposition	Categorical	Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available	0%
Admission Source	Categorical	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	Categorical	Integer identifier corresponding to 23 distinct values, for example, Blue Cross\Blue Shield, Medicare, and self-pay	52%
Medical specialty	Categorical	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	Categorical	The primary diagnosis (coded as first three digits of ICD9); 848 distinct values	0%
Diagnosis 2	Categorical	Secondary diagnosis (coded as first three digits of ICD9); 923 distinct values	0%

Diagnosis 3	Categorical	Additional secondary diagnosis (coded as first three digits of ICD9); 954 distinct values	1%
Number of diagnosis	Numeric	Number of diagnoses entered to the system	0%
Glucose serum test result	Categorical	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	Categorical	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	Categorical	Indicates if there was a change in diabetic medications (either dosage or generic name). Values: “change” and “no change”	0%
Diabetes medications	Categorical	Indicates if there was any diabetic medication prescribed. Values: “yes” and “no”	0%
24 features of medications	Categorical	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, metformin-rosiglitazone, and metformin-pioglitazone, the feature indicates whether the drug was prescribed or there was a change in the dosage. Values: “up” if the dosage was increased during the encounter, “down” if the dosage was decreased, “steady” if the dosage did not change, and “no” if the drug was not prescribed	0%
Readmitted	Categorical	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%

3. EXPLORATORY DATA ANALYSIS

Figure 1: Demographic data of a patients

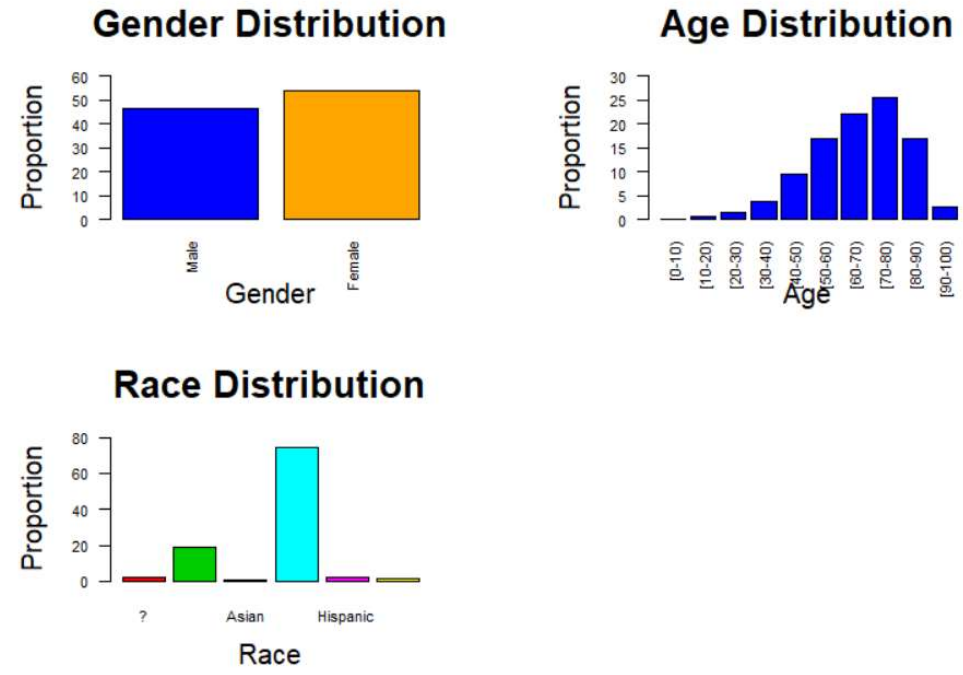


Figure 2: Medication Data of Patients

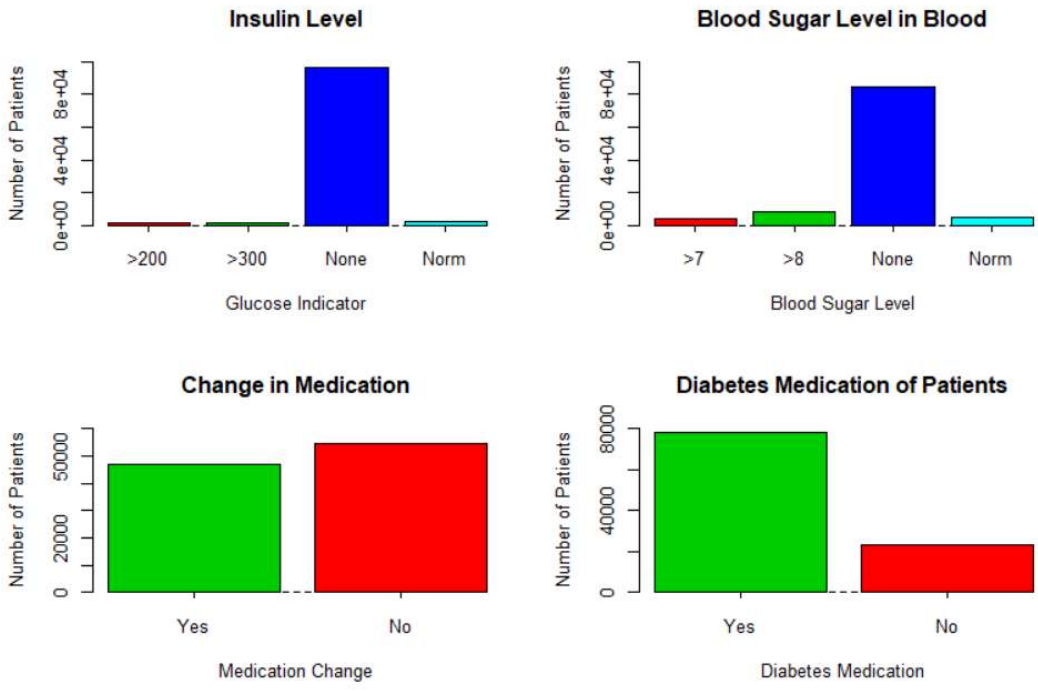


Figure 3: Demographic data of the patient based on Readmission:

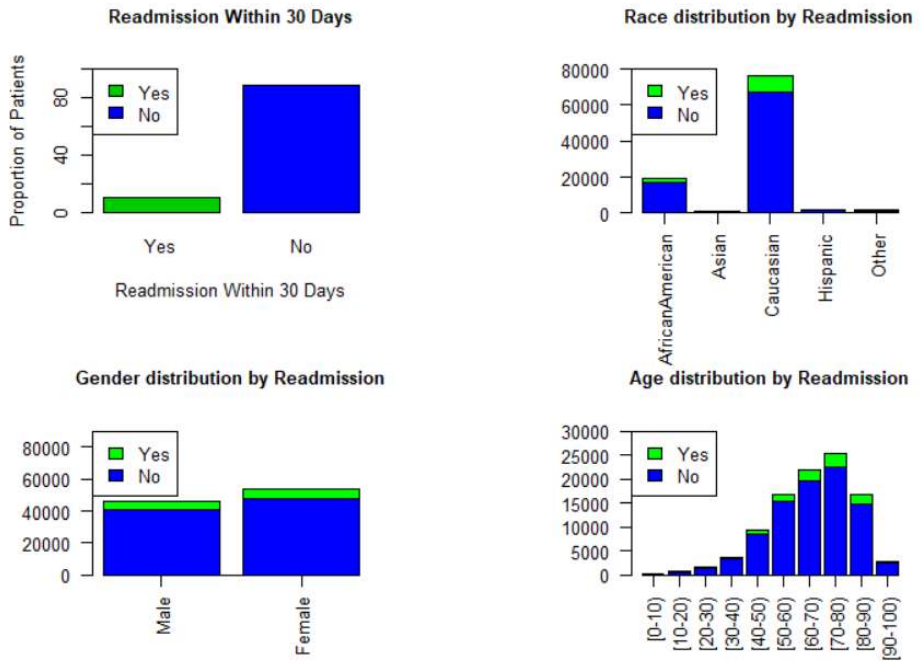


Figure 4: Number of admitted or readmitted patients spent time(in hours) in hospital

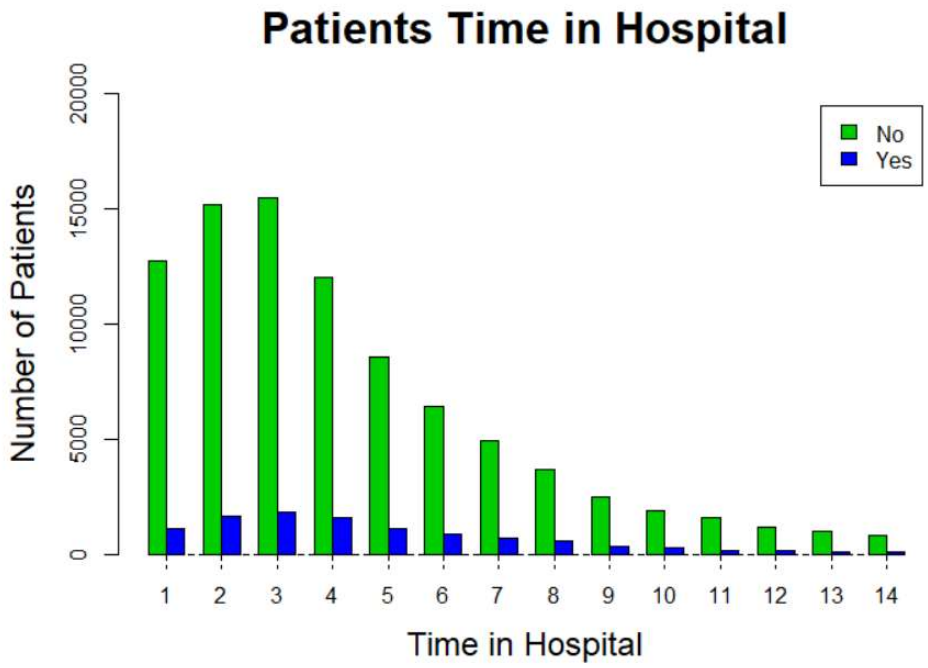


Figure 5 : Admission type of a patient

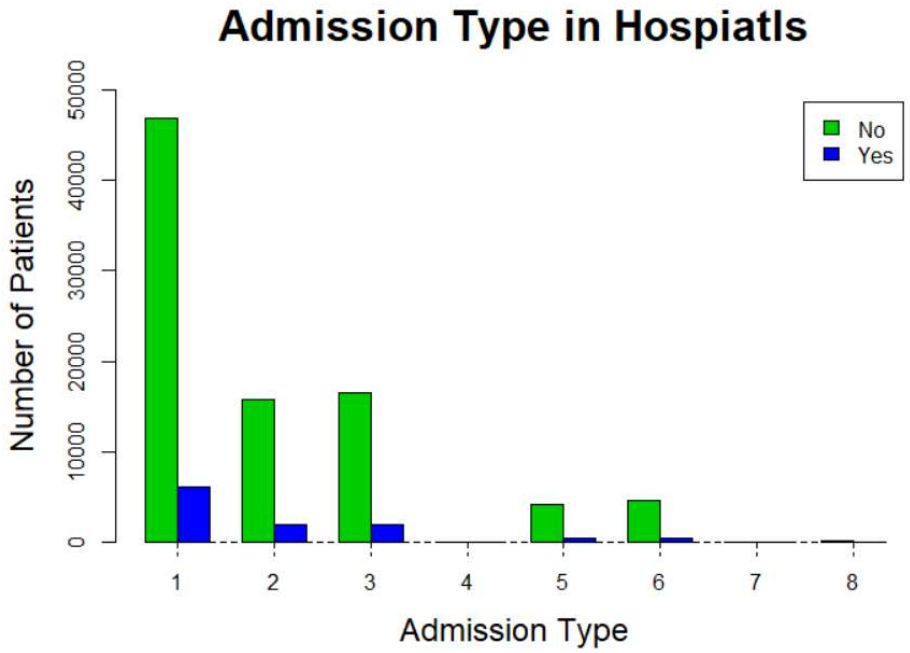
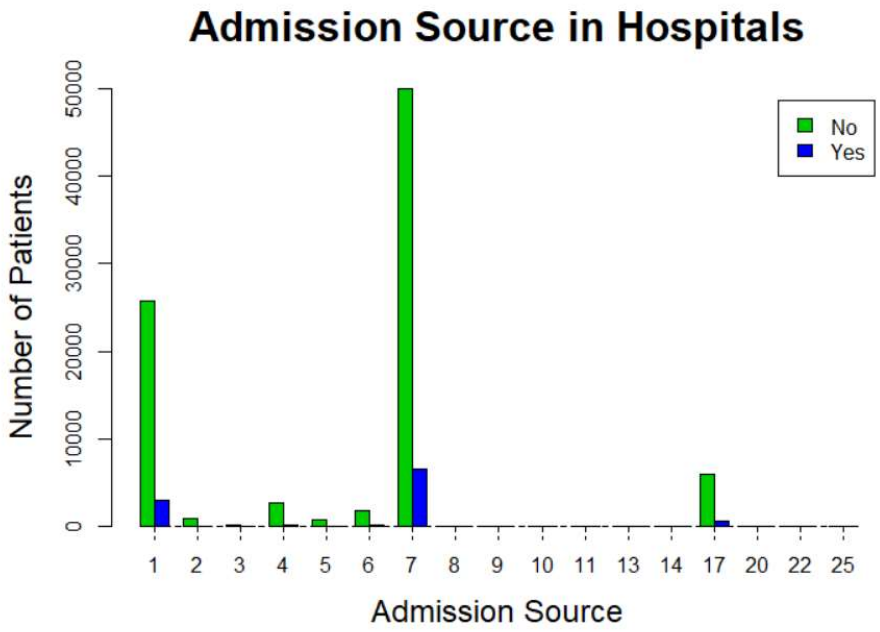


Figure 6: Admission source in Hospital



4. DATA PREPROCESSING

Table 2: Preprocessed categorical data to numeric value for further analysis

Attribute	**Initial Attribute Values**	**Attribute Values after Preprocessing**
race	AfricanAmerican, Asian,Caucasian,Hispanic,Other	1,2,3,4,5
gender	Male, Female	1,2
Age	[0-10),[10-20),[20-30), [30-40),[40-50),[50-60), [60-70),[70-80),[80-90), [90-100)	1,2,3,4,5,6,7,8,9,10
admission_type_id	Emergency, Urgent, Elective, Newborn, Not Available, NULL, Trauma Center, Not Mapped	1,2,3,4,5,6,7,8
discharge_disposition_id	Discharged to home, Discharged/transferred to another short-term hospital, Discharged/transferred to SNF, etc	1,2,3,4,5,6….29
admission_source_id	Physician Referral, Clinic Referral, HMO Referral, Transfer from a hospital, etc	1,2,3,4,…..26
max_glu_serum	>300 >200 normal or none	3 2 0
A1Cresult	>8 >7 None, Norm	3 2 0
23 diabetic medications	Down Up No, Steady	1 2 0
change	No Ch	0 1
diabetesMed	No Yes	0 1
readmitted	>30, NO <30	0 1

5. DIMENTIONALITY REDUCTION METHOD

Our dataset has 50 variables which is not possible to visualize, and the model built from these 50 variables may give worst accuracy. So, before proceeding to dimensionality reduction technique let us check the correlation between each variable.

Figure 7: Cluster Plot

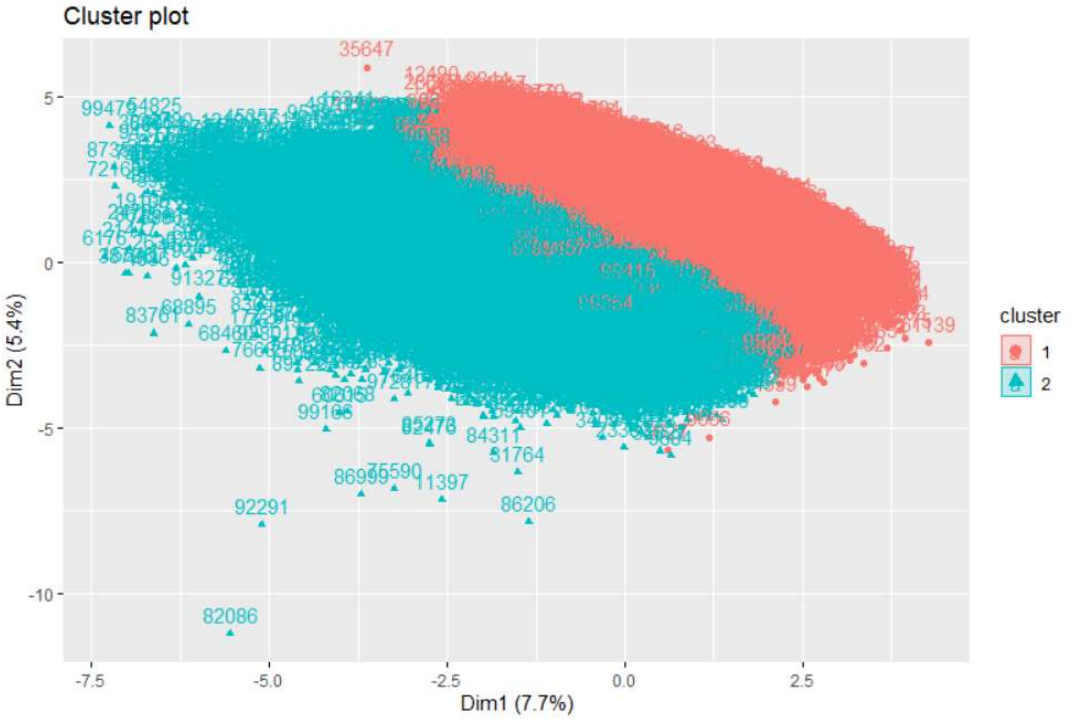
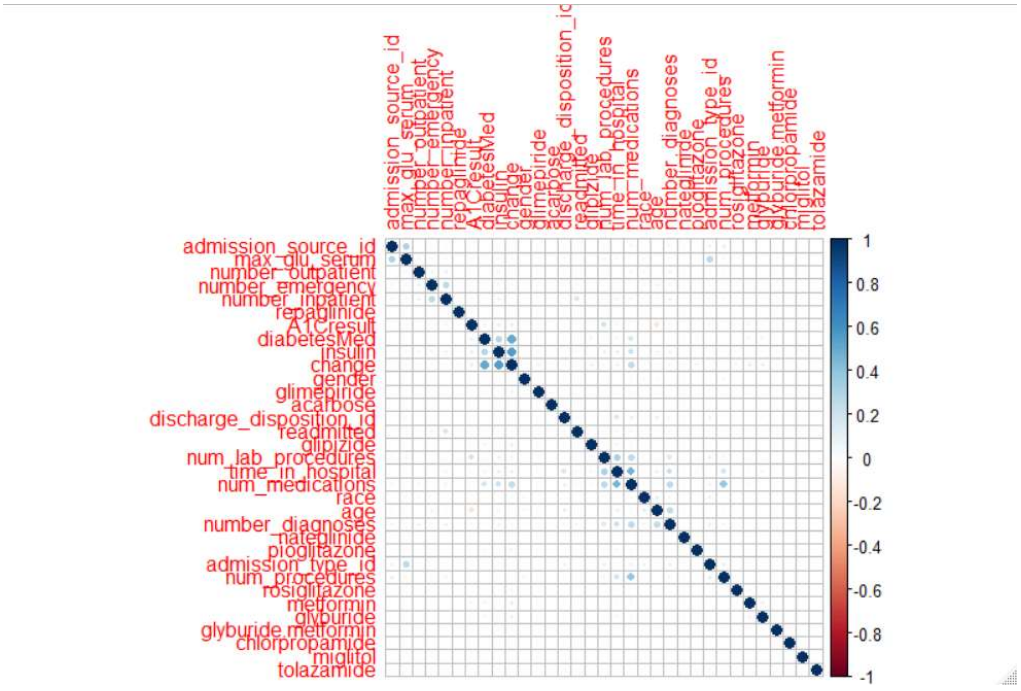


Figure 8: Correlation matrix before the dimensionality reduction



Applied dimensionality reduction technique to all 50 variables dataset. After using Principal Component Analysis on 50 variables, it concludes that 22 Principal Component(PC22) explains 80% of variability.

From Figure 9 we can conclude that the PC1 and PC2 explains 13.13% of variability. So, reducing the number of variables based on the loadings. The variables having higher loading will be taken into consideration for further analysis and model building.

The Scree Plot(Elbow Method shows that the elbow bends at PC2. Considering two PC for variable explanation. Figure 10 shows the variable with high variances in Biplot. Variables like change, insulin, diabetesMed, time_in_hospital, num_medications, number_diagnosis, A1Cresult, num_procedures, etc., has high variances.

High variances variables in PC1(Figure 11) and PC2(Figure 12) are shown in figures.

Figure 13 shows the correlation between variables after Dimensionality Reduction technique Principal Component Analysis

Figure 9: Scree plot

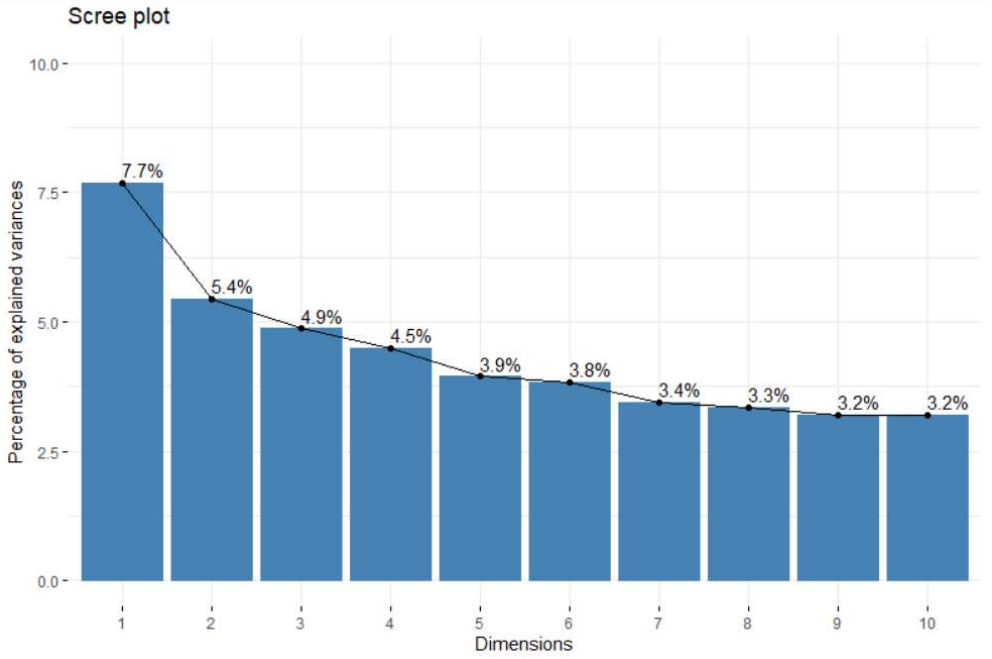


Figure 10: Biplot of PCA shows high variance variables.

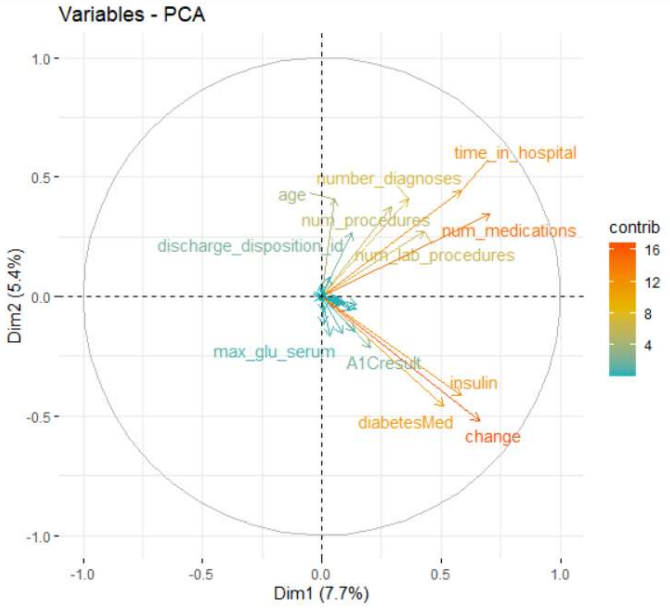


Figure 11: Variables from PC1 variance

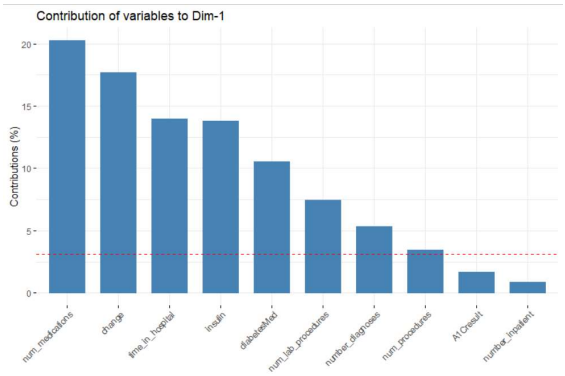


Figure 12: Variables from PC2 variance

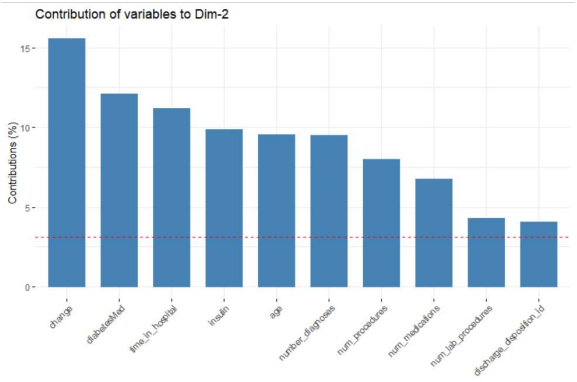
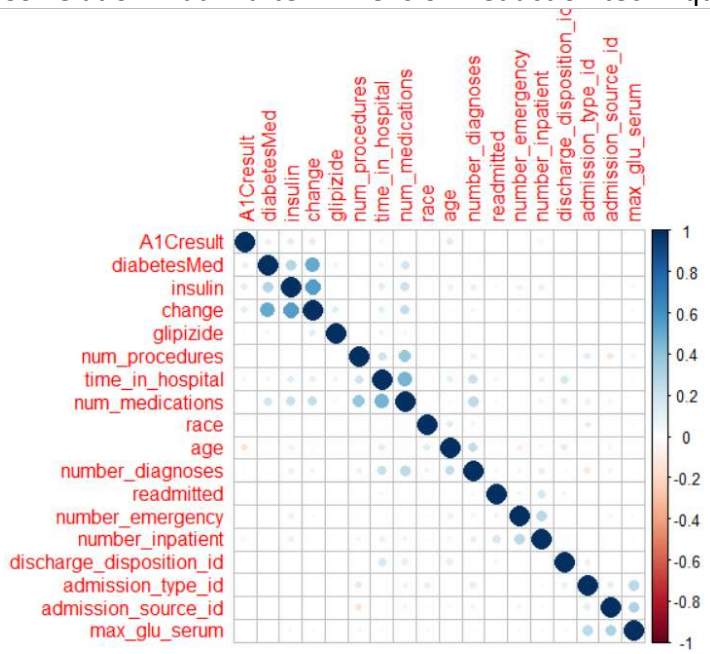


Figure 13: Correlation Matrix after Dimension Reduction technique



6. MODEL BUILDING

Before Model Creation:
Split our dataset of around 100,000 observations into training set and testing set of 70-30%. Training dataset consisting of 70% data has 69645 observations and 18 variables(after dimension reduction). Testing dataset consisting of 30% data has 29848 observations and 18 variables.

To create a relevant Logistic Regression, we must select relevant variables. Performed Logistic Regression on high variance variables after Principal Component

Analysis. Logistic Regression model was built to find each independent variable impact on response variable(i.e., Readmitted variable).

7. RESULTS AND DISCUSSION

Once the model with high significant variables is created, the model is trained on training dataset. The trained model is then validated on testing dataset.

To create a model, we performed Logistic Regression and removed a variable with the highest p-value(>0.05). The high significant variables such as age, discharge_disposition-id, time_in_hospital, number_inpatient, number_outpatient, number_diagnosis, max_glu_serum,insulin diabetesMed, etc., are selected. See Table 3.

Figure 14 is confusion matrix of logistic regression model. It has high True Positive value, but the False Negative values are high. The validation dataset has 88.76% Accuracy with high sensitivity and low specificity(Table 4). The trade-off between sensitivity(True Positive Rate) and specificity(1-False Positive Rate) is shown in figure 15. The Area Under Curve (AUC) is 64% which is good but not too good could have been better with some other model.

Table 3: Logistic Regression summary

```
glm(formula = readmitted ~ ., family = binomial(link = "logit"),
    data = training_set)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5061  -0.4956  -0.4375  -0.3933   2.4440

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -3.104007   0.084829 -36.591  < 2e-16 ***
race                      0.001971   0.014509   0.136 0.891951
age                       0.024944   0.008343   2.990 0.002792 **
admission_type_id        -0.013390   0.009184  -1.458 0.144855
discharge_disposition_id  0.025707   0.002110  12.185  < 2e-16 ***
admission_source_id      -0.004313   0.003276  -1.317 0.187975
time_in_hospital          0.016445   0.004543   3.620 0.000295 ***
num_procedures           -0.018317   0.008207  -2.232 0.025624 *
num_medications           0.003843   0.001879   2.046 0.040770 *
number_emergency          0.034415   0.010222   3.367 0.000761 ***
number_inpatient          0.263335   0.007796  33.779  < 2e-16 ***
number_diagnoses          0.046075   0.007331   6.285 3.28e-10 ***
max_glu_serum             0.082886   0.029748   2.786 0.005331 **
A1Cresult                -0.028045   0.014762  -1.900 0.057460 .
glipizide                 0.138520   0.059131   2.343 0.019150 *
insulin                   0.079520   0.021576   3.686 0.000228 ***
change                   -0.039786   0.033577  -1.185 0.236049
diabetesMed               0.125744   0.034859   3.607 0.000309 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 48919  on 69644  degrees of freedom
Residual deviance: 47134  on 69627  degrees of freedom
AIC: 47170

Number of Fisher Scoring iterations: 5
```

Table 4: Logistic Regression Confusion Matrix summary

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 26443  3300
         1    54    51

               Accuracy : 0.8876
                 95% CI : (0.884, 0.8912)
    No Information Rate : 0.8877
    P-Value [Acc > NIR] : 0.5265

                  Kappa : 0.0228

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.99796
            Specificity : 0.01522
         Pos Pred Value : 0.88905
         Neg Pred Value : 0.48571
             Prevalence : 0.88773
         Detection Rate : 0.88592
   Detection Prevalence : 0.99648
      Balanced Accuracy : 0.50659

       'Positive' Class : 0
```

Figure 14: Confusion Matrix of validation dataset

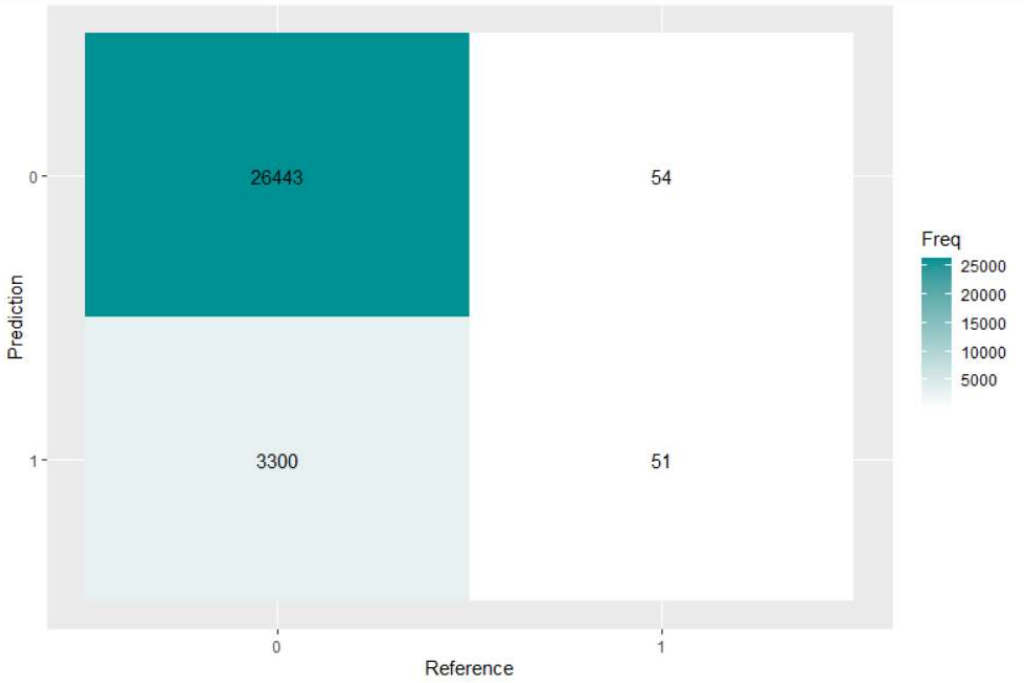
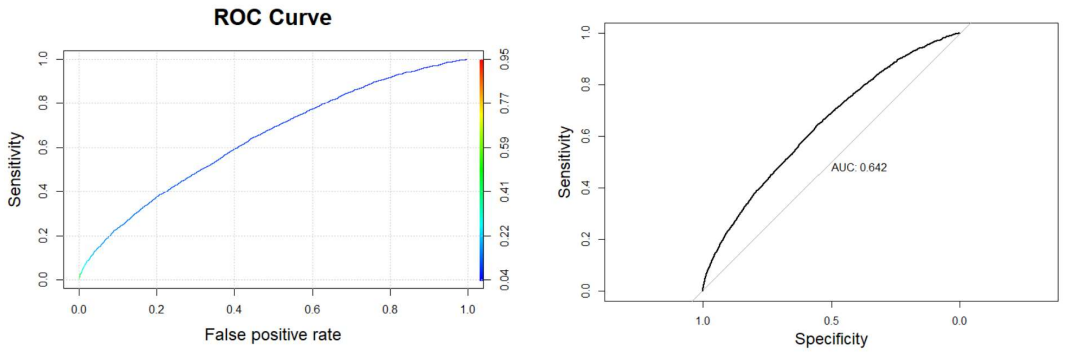


Figure 15: Receiver Operation Characteristics (ROC) Curve



8. CONCLUSION

In conclusion, instead of using all 50 variables for obtaining readmission rate of diabetic patient change study suggested to focus on patient's medicine, diabetes medication, age, insulin, number of diagnosis, lab procedure, discharge disposition. The most significant variables(p-value < 0.05) give readmission rate more accurate.

9. REFERENCES

[1] K. J. Cios and G. W. Moore, "Uniqueness of medical data mining," *Artificial Intelligence in Medicine*, vol. 26, no. 1-2, pp.1–24, 2002.

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

[3] R. M. Bergenstal, J. L. Fahrbach, S. R. Iorga, Y. Fan, and S. A. Foster, "Preadmission glycemic control and changes to diabetes mellitus treatment regimen after hospitalization," *Endocrine Practice*, vol. 18, no. 3, pp. 371–375, 2012.

[4] D. Baldwin, G.Villanueva, R.McNutt, and S. Bhatnagar, "Eliminating inpatient sliding-scale insulin: a reeducation project with medical house staff," *Diabetes Care*, vol. 28, no. 5, pp. 1008–1011, 2005.

[5] Stone J, Hoffman G. Medicare hospital readmissions: issues, policy options and PPACA. Vol 7–5700, R40972: Congressional Research Service, Penny Hill Press; 2010.

[6] (http://www.cioslab.vcu.edu/)