BA631 Visual Reporting and Communication

DECODING THE PROPENSITY OF HOSPITAL ADMISSION RATES OF DIABETIC CUSTOMERS BY ANALYZING A CLINICAL DATABASE OF 99,000 PATIENT RECORDS

Ravi Teja Reddy, Bandi
Shitanshu Pokhriyal

ABSTRACT OF BUSINESS PROBLEM

A large portion of hospital inpatient management and expense is because of hospital readmission rates. And Diabetes is one of the top ten leading causes of death, which is also considered the most expensive disease in the United States. Patients with diabetes are at high risk of getting readmitted than those patients who are not diabetic. Therefore, understanding who/what kind of patients are readmitting and trying to minimize readmission rates will help patients to reduce medical costs.

The objective of this study is to predict the likelihood of a diabetic patient being readmitted.

DATA SOURCE

The data was submitted on behalf of the Center for Clinical and Translational Research, Virginia Commonwealth University and the de-identified abstract data has been stored and collected from UCI Machine learning repository[1] originally originated from the globally used Cerner EMR systems for patient care management.

The data set[2] contains instances of nearly 100,000 and over 50 variables of 70000 patients which are collected over a time of 10 years from 130 US hospitals. The data set contains information regarding,

1. Inpatient hospital admission.
2. Diabetic encounter, diagnosis of which kind of diabetes was entered into the system.
3. Length of stay with min 1 day and max 14 days.
4. Laboratory tests performed during the encounter.
5. Medications administered during the encounter.

Based on the above criteria the set contains attributes patient number, race, gender, age, admission type, time in the hospital, medical specialty of admitting physician, number of lab test performed, HbA1c test result, diagnosis, number of medications, diabetic medications, number of outpatients, inpatient, and emergency visits in the previous year before hospitalization, etc.

The stakeholders in this study are Government, the Insurance industry, Educational institutions, and patients.

[1] https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008#
[2] https://archive.ics.uci.edu/ml/machine-learning-databases/00296/dataset_diabetes.zip

DATA DICTIONARY

Out of 50 available variables, the following variables have been used for our business question. The dictionary of variables used for Data analysis can be found in the appendix.

Variable	Data type	Description
encounter_id	Numeric	Unique identifier of an encounter
patient_nbr	Numeric	Unique identifier of a patient
race	Nominal	Values: Caucasian, Asian, African American, Hispanic, and other
gender	Nominal	Values: male, female, and unknown/invalid
age	Nominal	Grouped in 10-year intervals: [0, 10), [10, 20), . . ., [90,100)
weight	Numeric	Weight in pounds **Masked**
admission_type_id	Nominal	Integer identifier corresponding to 9 distinct values, for example, emergency, urgent, elective, newborn, and not available
discharge_disposition_id	Nominal	The integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available
admission_source_id	Nominal	Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital
time_in_hospital	Numeric	The integer number of days between admission and discharge
payer_code	Nominal	Integer identifier corresponding to 23 distinct values, for example, Blue Cross\Blue Shield, Medicare, and self-pay **Masked**
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
num_lab_procedures	Numeric	Number of lab tests performed during the encounter
num_procedures	Numeric	Number of procedures (other than lab tests) performed during the encounter
num_medications	Numeric	Number of distinct generic names administered during the encounter
number_outpatient	Numeric	Number of outpatient visits of the patient in the year preceding the encounter
number_emergency	Numeric	Number of emergency visits of the patient in the year preceding the encounter
number_inpatient	Numeric	Number of inpatient visits of the patient in the year preceding the encounter
diag_1	Nominal	The primary diagnosis (coded as first three digits of ICD9); 848 distinct values

diag_2	Nominal	Secondary diagnosis (coded as first three digits of ICD9); 923 distinct values
diag_3	Nominal	Additional secondary diagnosis (coded as first three digits of ICD9); 954 distinct
number_diagnoses	Numeric	Number of diagnoses entered into the system
max_glu_serum	Nominal	Indicates the range of the result or if the test was not taken. Values: “>200” “>300”, “normal”, and “none” if not measured
A1Cresult	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.
change	Nominal	Indicates if there was a change in diabetic medications (either dosage or generic name). Values: “change” and “no change”
diabetes med	Nominal	Indicates if there was any diabetic medication prescribed. Values: “yes” and “no”
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, 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
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.

DESCRIPTIVE ANALYSIS AND CLEANING

DATA DIMENSION (VARIABLES AND OBSERVATIONS)

As mentioned earlier the dataset has over one hundred thousand instances and 50 variables, out of which 13 are numeric (continuous) and 37 are factor variables (non-continuous)

PREPROCESSING AND CHALLENGES

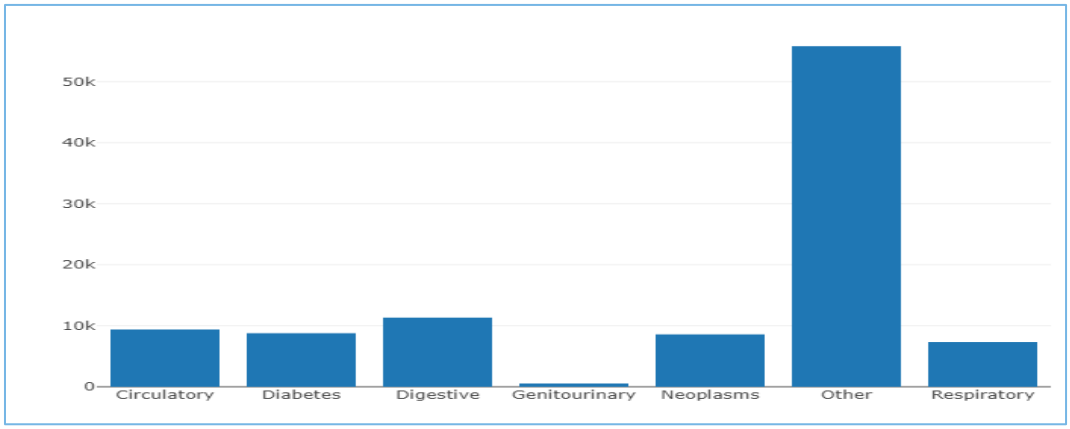
1. Many variables have '?' and 'Unknown/Invalid' in them as values, where there are missing or masked values for privacy purposes. So, during the cleaning process, we have replaced them with NA.
2. Some variables which have only NA or monotonous or masked for privacy are not useful for our analysis such as encounter_id, patient_nbr, weight, payer_code, medical specialty - the specialty of the doctor who referred the patient not statistically significant in R and patient other conditions are captured in diagnosis are dropped.
3. We have used data heat maps in R to identify missing values. Once we remove all the variables not useful for our analysis and fix all other variables values, we can implement a correlation matrix and other visualizations.

DIMENSIONALITY AND DATA REDUCTION

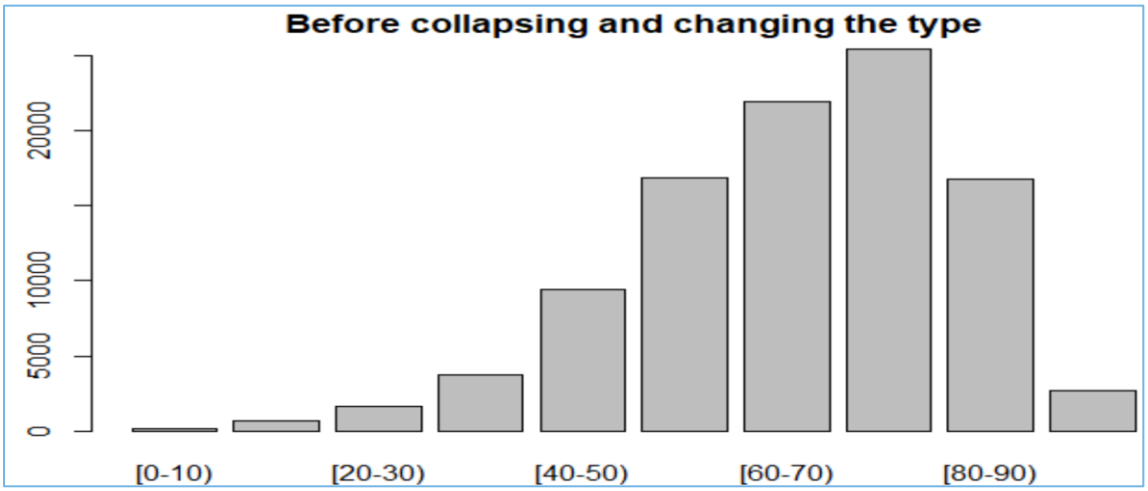
We have clubbed some of the classes of categorical variables for convenience and better analysis, also will remove some outliers in numerical variables for analysis in R as it will help with the accuracy of our model.

CATEGORY REDUCTION

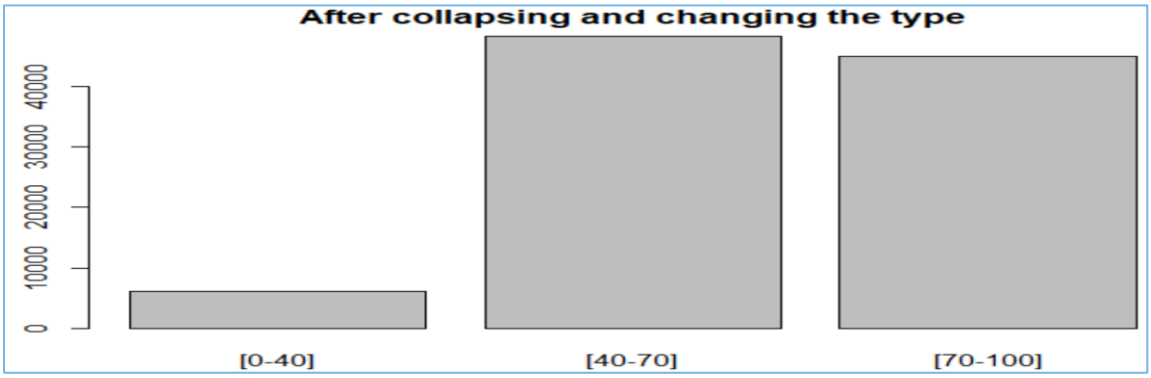
- We have taken the variable diag_1 and plotted the medical codes according to CDC to understand what kind of problems the patient is having. The same has been done for diag_2 and diag_3



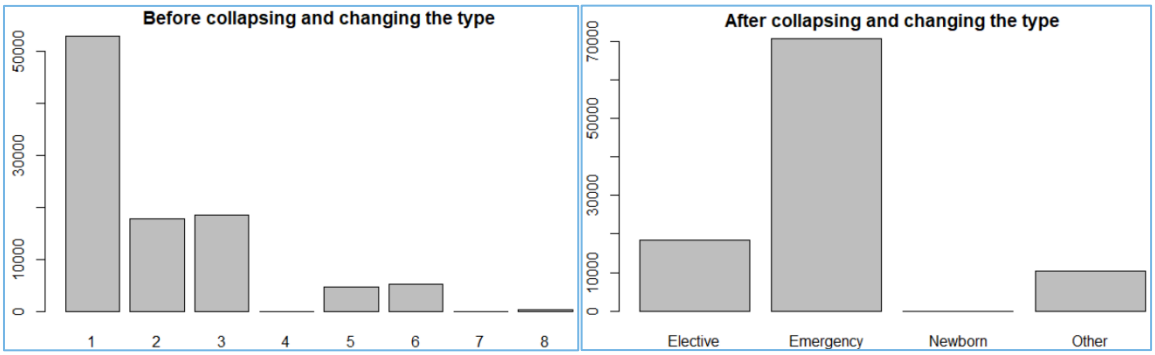
- We have taken the age variable and it has different classes of age. These were kept as is for Dashboard Visualizations.



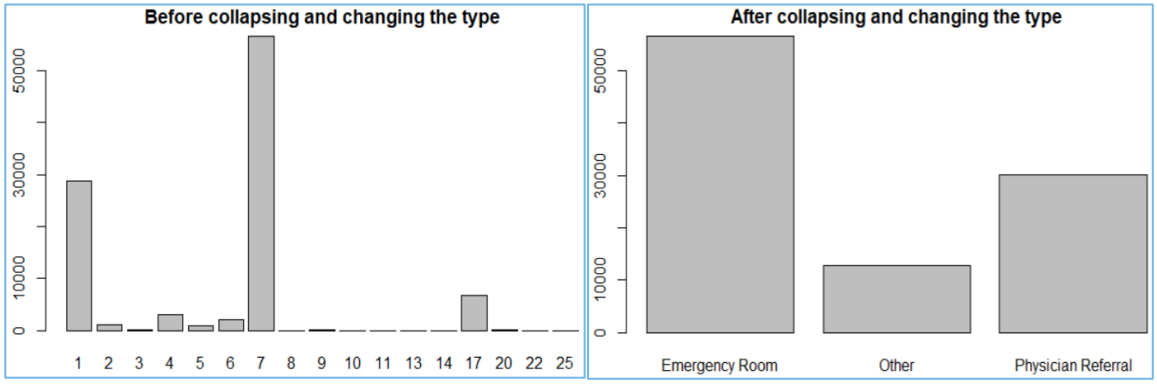
- To analyze the data in R in order to get accurate results, we have collapsed 10 classes to 3 classes as the data is left-skewed. The new three classes in the age variable are 0-40, 40-70 & 70-100.



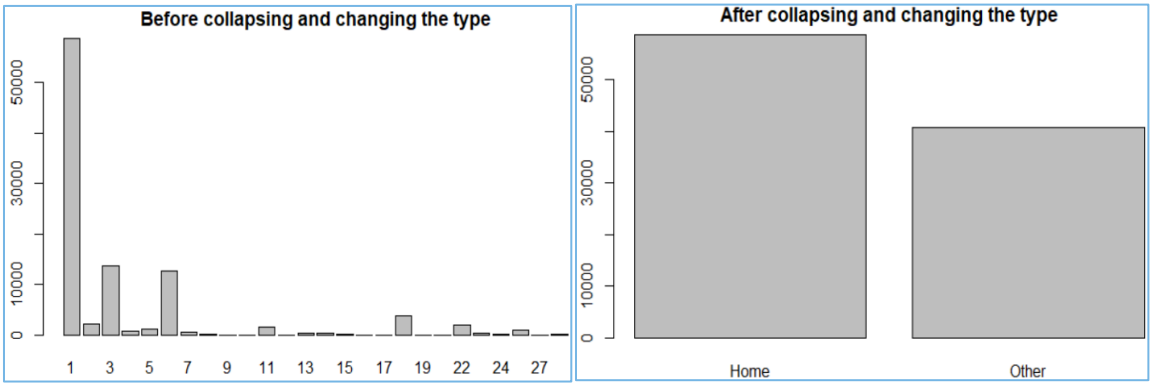
- In the variable admission type id, we have 8 classes with their ID codes; based on their similarities and necessities we have clubbed 8 classes into 4 and renamed variable and classes



- Like the above variable, we have referred to data dictionary and collapsed the classes accordingly for the admission source ID variable and renamed the classes and variable admission source. This variable will tell us from what source the patient has been admitted to the hospital.

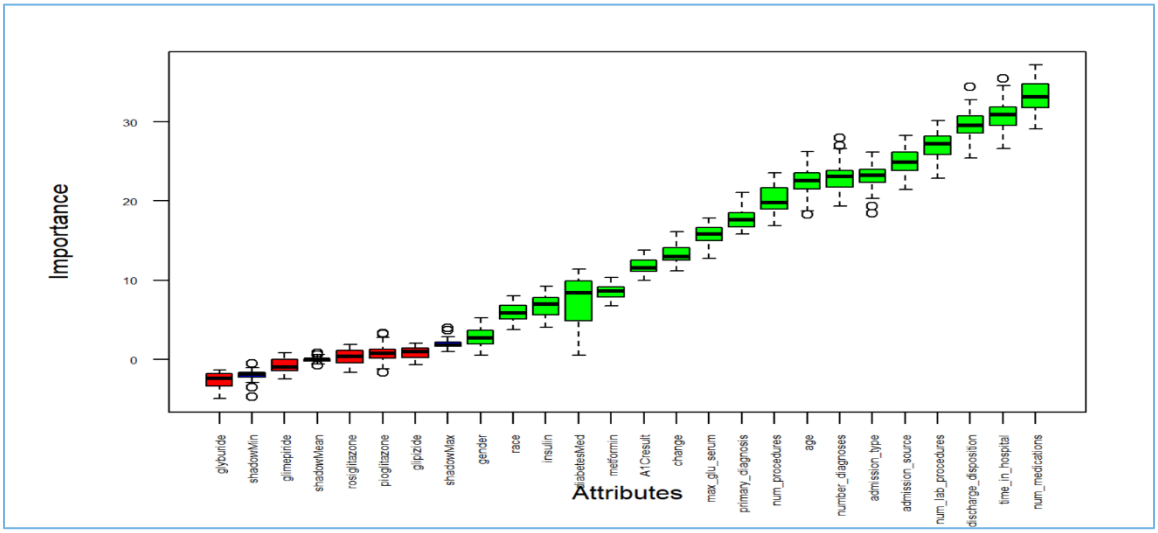


- Discharge disposition id tells us where the patient has been to after discharging like home, hospice, etc. we have collapsed 28 classes into 2 and renamed them accordingly and renamed the variable.

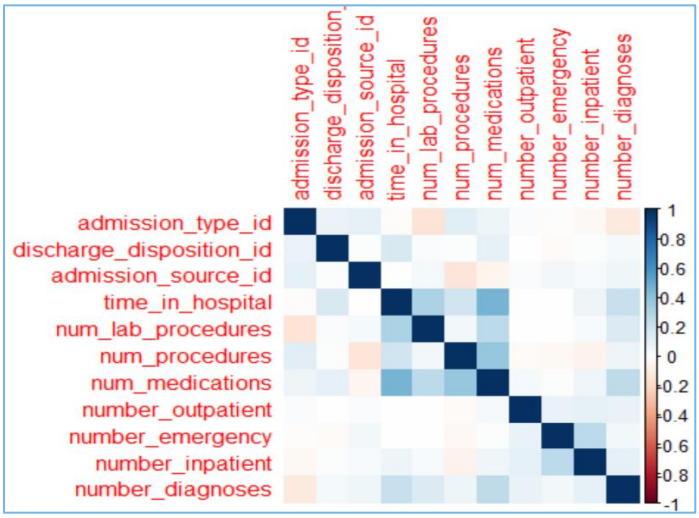


DATA AND DIMENSION REDUCTION (FOR DATA ANALYSIS IN R ONLY)

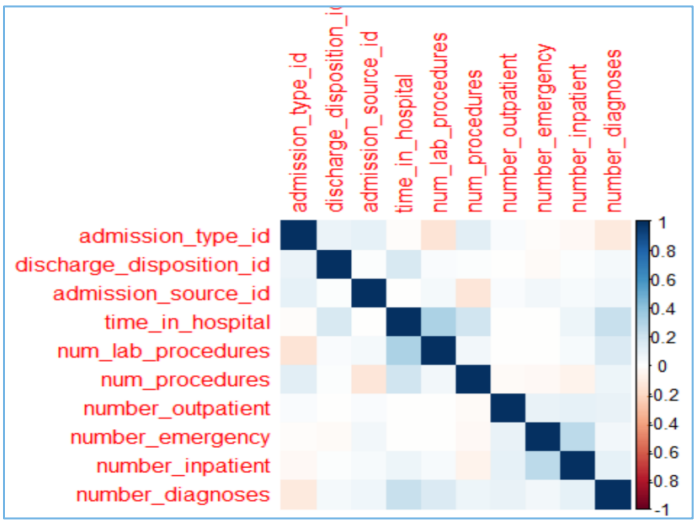
- Variables which has only NA or monotonous or are not useful for our analysis such as encounter_id, patient_nbr, weight, payer_code, medical specialty (diagnosis variables are similar to this) are dropped.
- All the instances with NA values are dropped.
- Now we have used the BORUTA function in R, "Boruta is an all-relevant feature selection wrapper algorithm, capable of working with any classification method that output variable importance measure (VIM); by default, *Boruta uses Random Forest. The method performs a top-down search for relevant features by comparing original attributes' importance with importance achievable at random, estimated using their permuted copies, and progressively eliminating irrelevant features to stabilize that test"*[3].
- In Boruta plots box plots for all the variables - blue boxes consist of minimum, maximum z scores of shadows features we created, green ones are most significant, red ones are unimportant, and yellow color code is given to the variable whose significance Boruta is not able to identify.



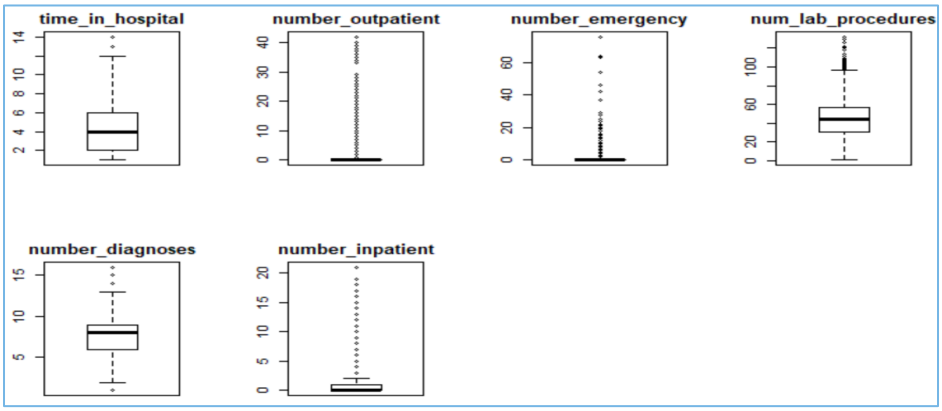
- We have implemented a correlation matrix to remove any variables with higher correlation. We can observe that number of medications and time spend in the hospital are correlated which makes sense.



- Hence, we have removed the num_medications variable and can observe the correlation matrix after the removal of the variable.

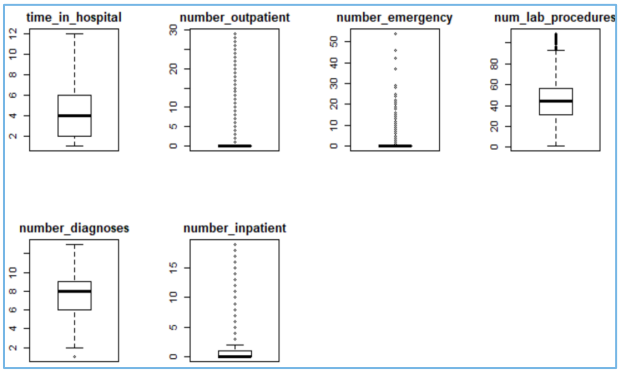


- Now in the data, we have some outliers in numerical variables



- We have removed only some extreme outliers that could be disruption to data.

```
time_in_hospital <= 12)
number_outpatient <= 30)
number_emergency <= 60)
num_lab_procedures <= 110)
number_diagnoses <= 13)
number_inpatient <= 20)
num_procedures <= 5)
```



DATA MINING (R)

As we are trying to find whether a patient is being readmitted and what are the factors, we will implement our data mining techniques on readmitted variables with other remaining variables as independent variables. Readmitted variables are factor variables with values 1 and 0, readmitted and not readmitted. As we are trying to implement a model that is going to predict between two classes our job is now classification.

We have created dummies for the remaining categorical variables, which got our variable count to 41 (image 23). And then we have split the 60 % data into training and 40% to validation sets (image 24).

DATA MINING – THREE TECHNIQUES APPLIED

After preprocessing the data and choosing the model, we are now going to run the model using three different techniques to compare the accuracy of the different techniques. And infer better results together.

[3] https://www.rdocumentation.org/packages/Boruta/versions/7.0.0/topics/Boruta

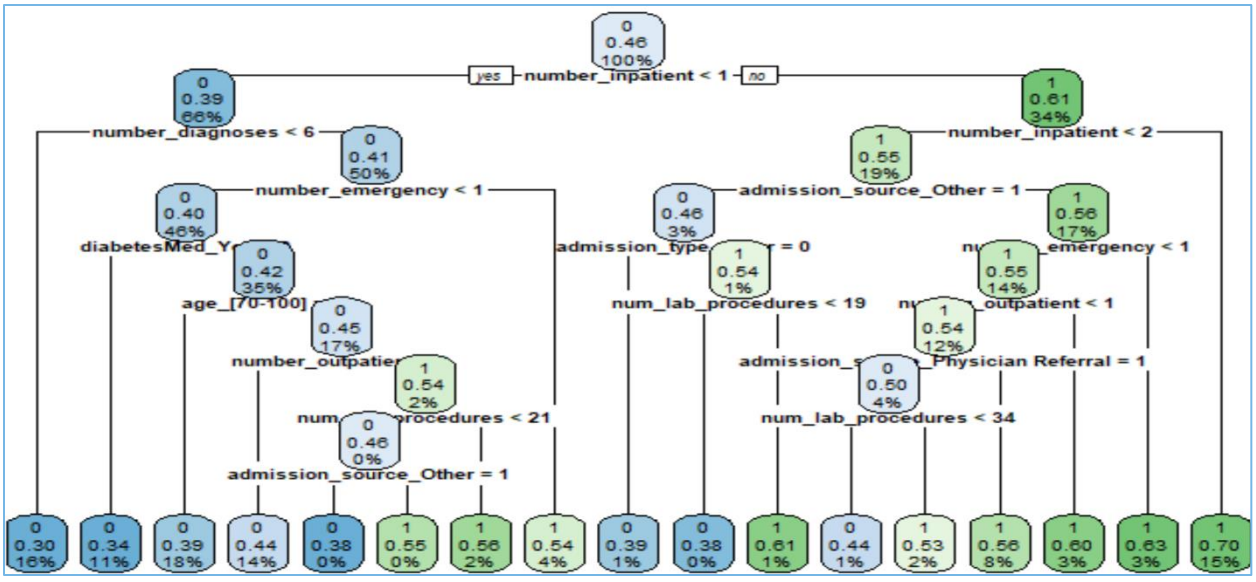
STEPWISE LOGISTIC REGRESSION

First, we have run the stepwise logistic regression, wherein each step variable is either added or removed from the model. When the minimum possible AIC value is achieved our regression stops. We can look at the variables in the final step and their estimates.

The relation between input and output function is approximate better. Makes smoother decisions with well-distributed estimations. Variables like higher age, inpatient are highly significant (image 25).

- This model suggests that patients who are diagnosed diabetic, age above 40, being Outpatient, the sugar level in hemoglobin more than 8 are more likely to readmit.
- People who are spending more time in the hospital, inpatient, taking a greater number of medications, a greater number of procedures are less likely to readmit to the hospital again.

CLASSIFICATION TREE



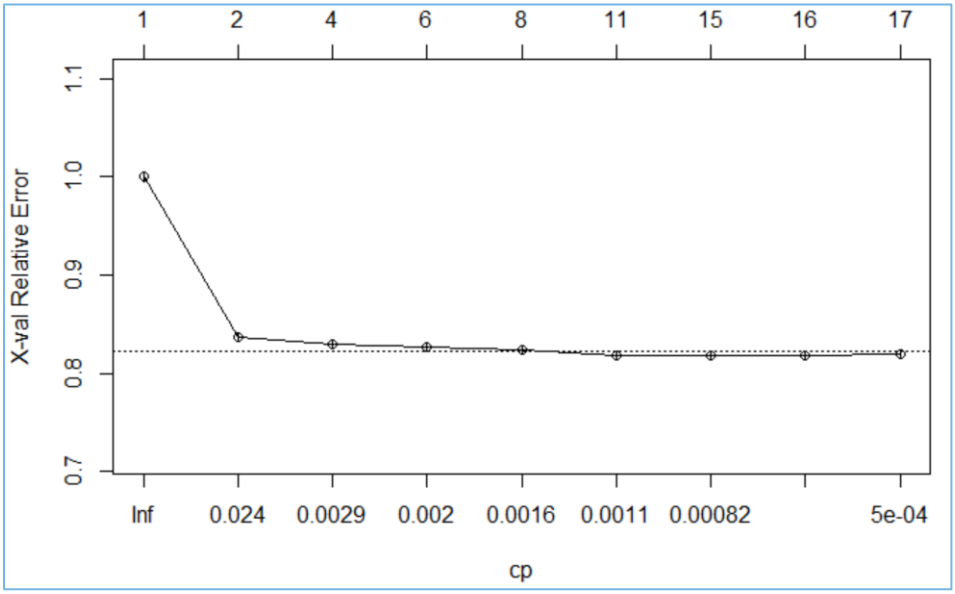
In the classification tree, the parent node is readmitted as it is our dependent variable and we can observe all other ate leaves with each of them having a weight.

Some variables are undervalued. Some probability estimations are constant. Inpatient, admission source, admission type are significant variables here.

As we are implementing this technique on dummy variables, we can have a look at what kind of patients are being readmitted and who is not. If we look at the left to the tree an inpatient who was diagnosed diabetic and joined the hospital in an emergency will likely get readmitted 56% of the time.

Or if we look at the right side of the tree a person who is an outpatient is likely to get readmitted with a chance of 15%. Similarly, we can interpret the tree and understand the who and through what causes are getting readmitted. Our classification tree accuracy is 53% appx.

We can get more nodes to understand the above tree by adjusting the complexity of the tree which will add more variables to the tree. As we increase the complexity the error rate decreases.



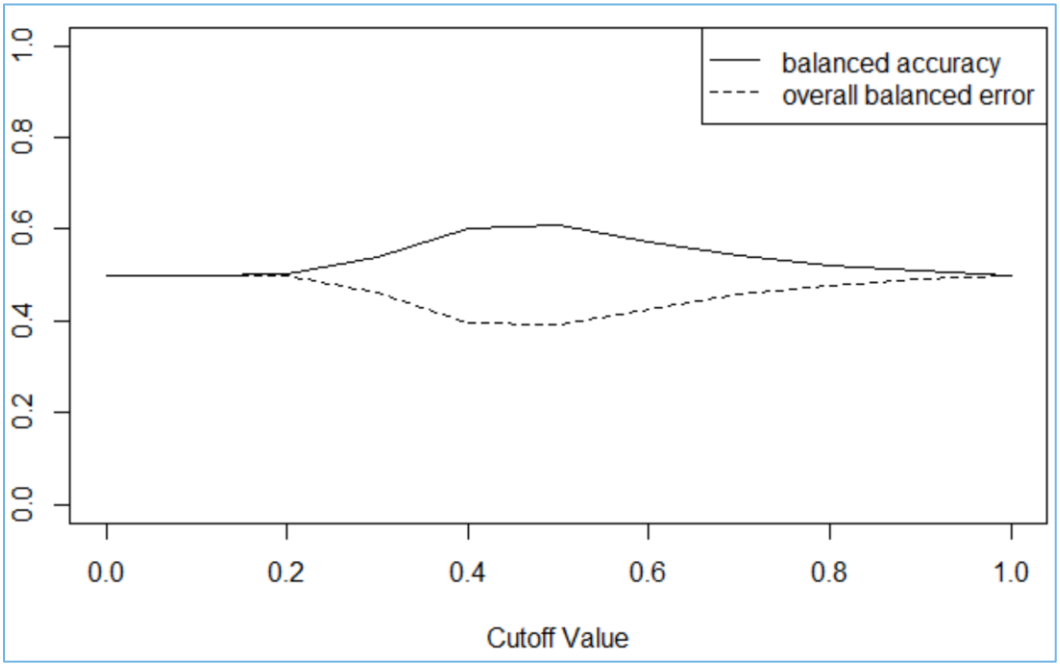
NAÏVE BAYES

Like the above techniques, we have implemented Naïve Bayes where we have observed the accuracy of the model to be 51.6%. (image 29)

EVALUATION AND RESULTS

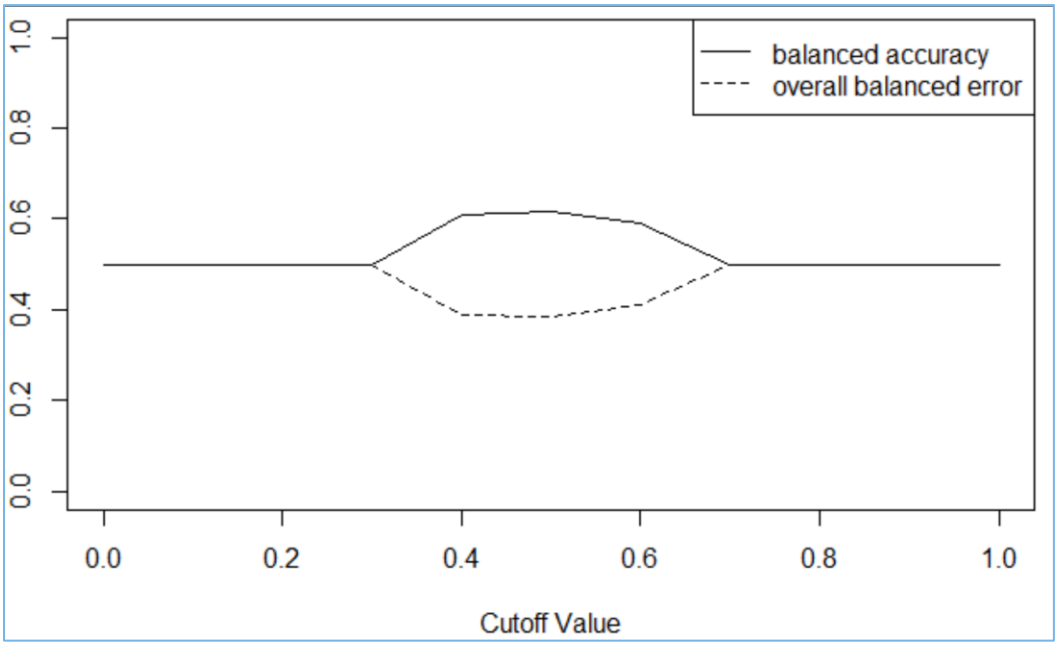
Model	Accuracy
Stepwise Logistic	0.6201
Classification tree	0.5298
Naïve Bayes	0.5163

As we have done a classification task, we have considered balanced accuracy as my primary evaluation metric. Normal accuracy is sometimes misleading when the classes are not balanced. So, we are comparing the balanced accuracy of all three techniques.

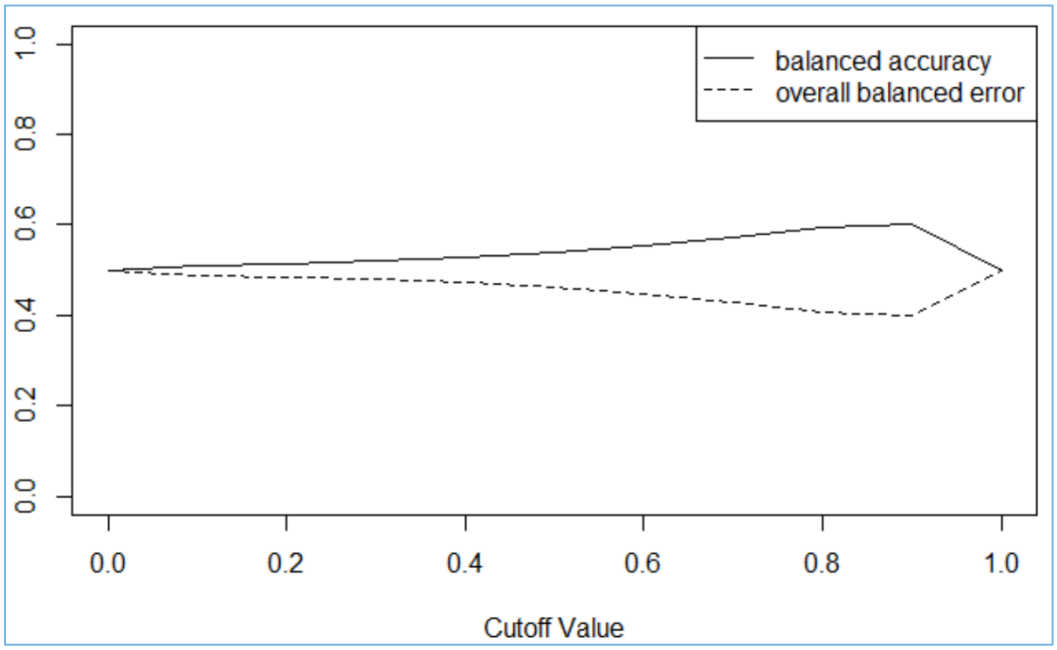


In the above graph, we are comparing the performance of the model with the cutoff value, the stepwise logistic regression model performed well with a cutoff of 0.5. We can achieve balanced accuracy of above 60%.

In the below graph we have a balanced accuracy plot for the classification tree, and as same as for the logistic regression model we can achieve maximum balanced accuracy near 60% with a cutoff value of 0.5.



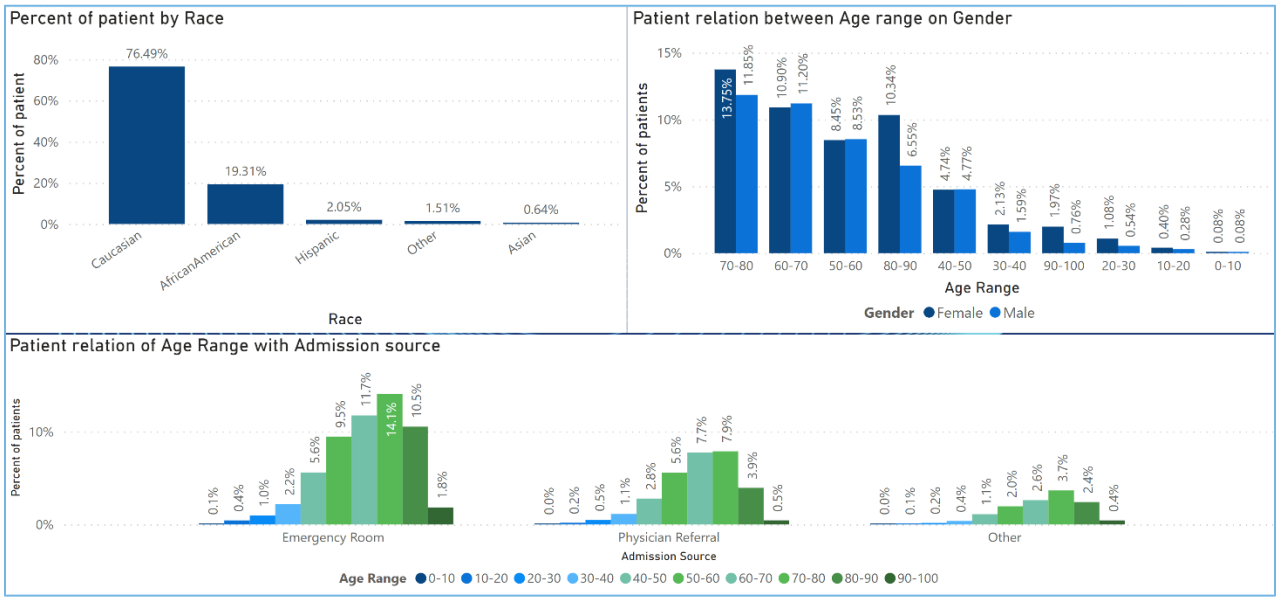
In the below plot balanced accuracy plot for Navies Bayes, and we are getting maximum accuracy as below 60% at a cutoff value of 0.9



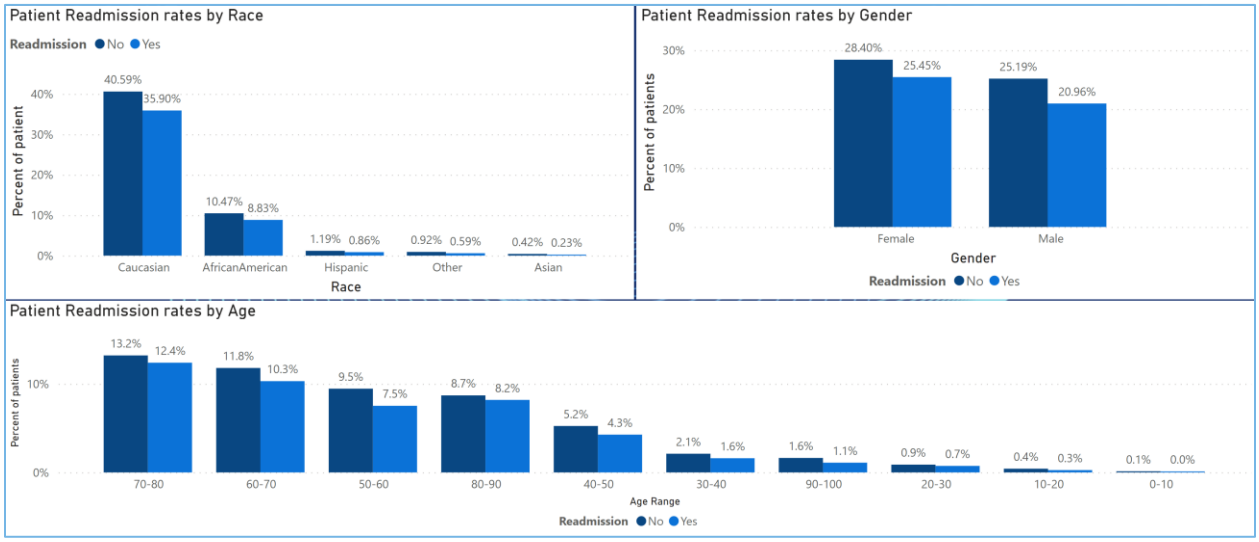
VISUALIZATION ANALYSIS (DASHBOARD)

PATIENT DEMOGRAPHIC ANALYSIS

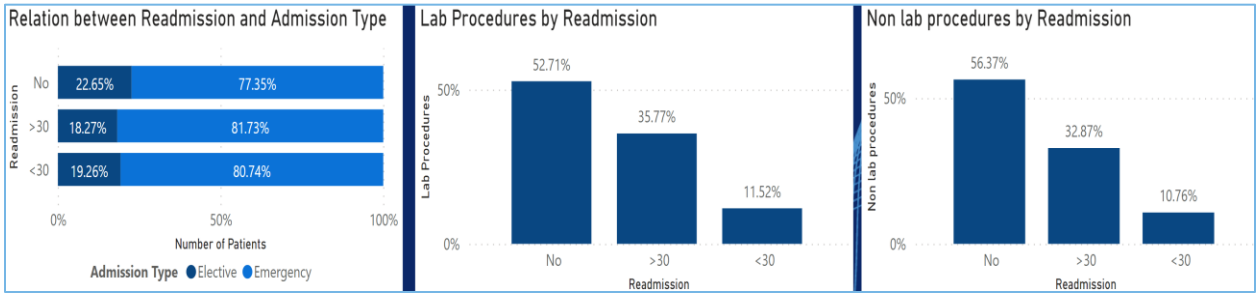
- Database of 99k Diabetic Patients obtained for a hospital in Ohio with information from the Serner system, used globally in developed markets for patient history management
- 76% of the patients are Caucasian and 19% African Americans; 77% of the base over 50 years age; Data evenly distributed with 53% of the patient's females and 47% Males
- Higher Rate of Patients getting admitted in the Emergency scenario than a planned admission/physician referral - 45% customers getting admitted in an emergency are >50 years
- Caucasian patients have a high readmission rate of 35.9% and females contribute to over 25% of readmissions.



- Higher age bands of over 50 years contribute to over 43% of readmissions.



- Readmissions are high (>80%) via the Emergency route vis a vis Elective booking made through the doctor
- Lab Vs. Non-Lab procedures both impact readmission by ~46%



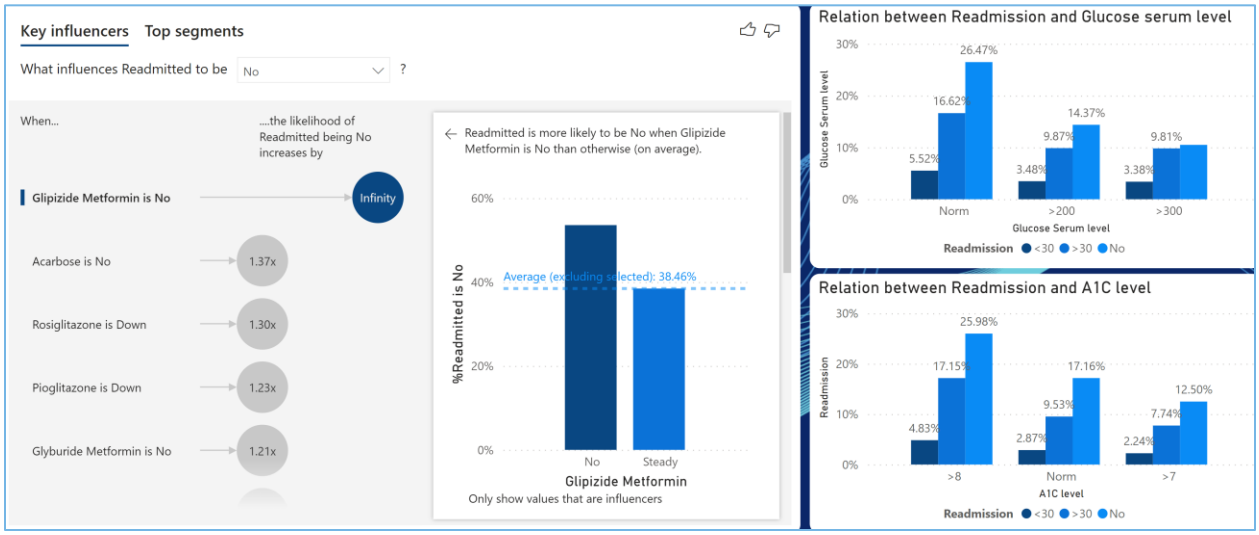
- 16.5% of the patients (16,500) made OPD visits of which 58% patients were readmitted within the year
- 33.6% were Inpatient visits (33681) of which 60% were readmitted within one year

Patient Readmission analysis

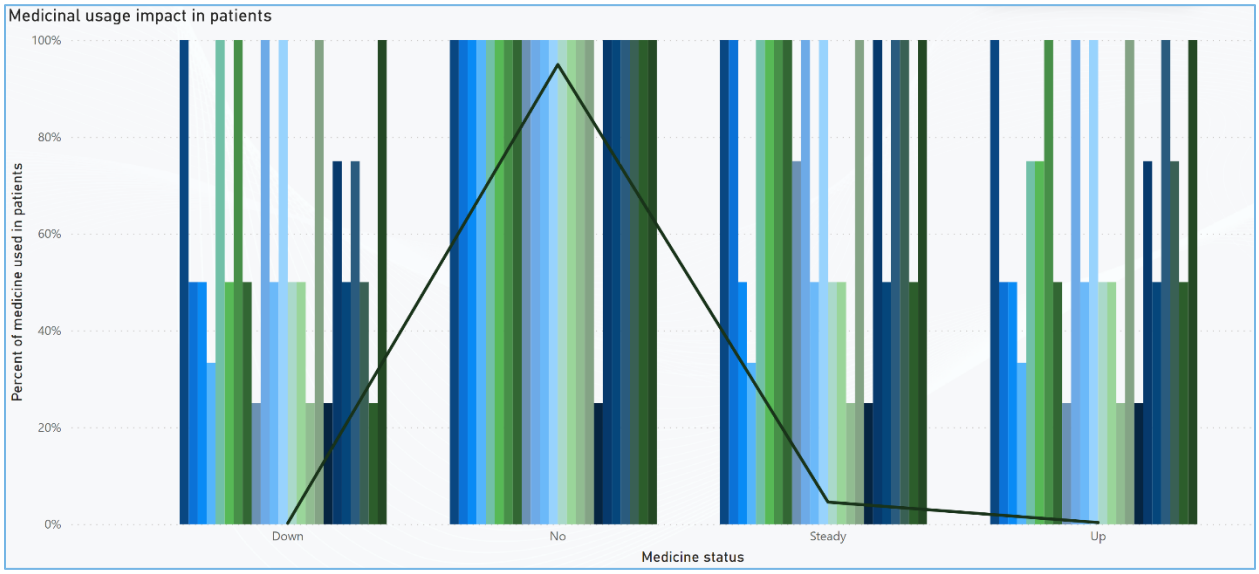
Age Range	Patients	OPDs Visitors	With in 30days	After 30days	No Readmit	Inpatient Visitors	With in 30days	After 30days	No Readmit
0-10	160	2	0.01%		0.01%	12	0.00%	0.01%	0.02%
10-20	682	60	0.04%	0.16%	0.17%	180	0.06%	0.27%	0.20%
20-30	1611	251	0.38%	0.58%	0.56%	591	0.52%	0.78%	0.45%
30-40	3699	511	0.44%	1.27%	1.39%	1214	0.77%	1.72%	1.12%
40-50	9465	1431	1.35%	3.76%	3.56%	3115	1.67%	4.51%	3.07%
50-60	16895	2676	1.92%	7.35%	6.95%	5410	2.57%	7.57%	5.92%
60-70	21988	3874	3.21%	10.71%	9.56%	7331	3.61%	9.94%	8.22%
70-80	25468	4372	3.62%	12.23%	10.65%	8863	4.22%	11.34%	10.76%
80-90	16800	2944	2.42%	7.99%	7.42%	6054	2.78%	7.65%	7.54%
90-100	2724	379	0.31%	0.79%	1.20%	911	0.38%	0.95%	1.37%
Total	**99492**	**16500**	**13.68%**	**44.84%**	**41.47%**	**33681**	**16.59%**	**44.73%**	**38.67%**

SERUM, HBA1C AND MEDICINES IMPACT

- Serum level refers to the effect of the prolonged medication administered and obtained by suitable blood tests.The analysis shows a relationship between Glucose Serum Levels and Readmission. In scenarios of high glucose serum levels, the chance of readmission is seen to be higher.
- HbA1c tests reflect the average blood sugar levels in a prolonged 2-3-month period. Higher HbA1c has a direct relationship with readmissions.
- The interactive influencers charts can be used in power BI to understand the impact of all the medications used. The dashboard can be accessed online and included in the folder.



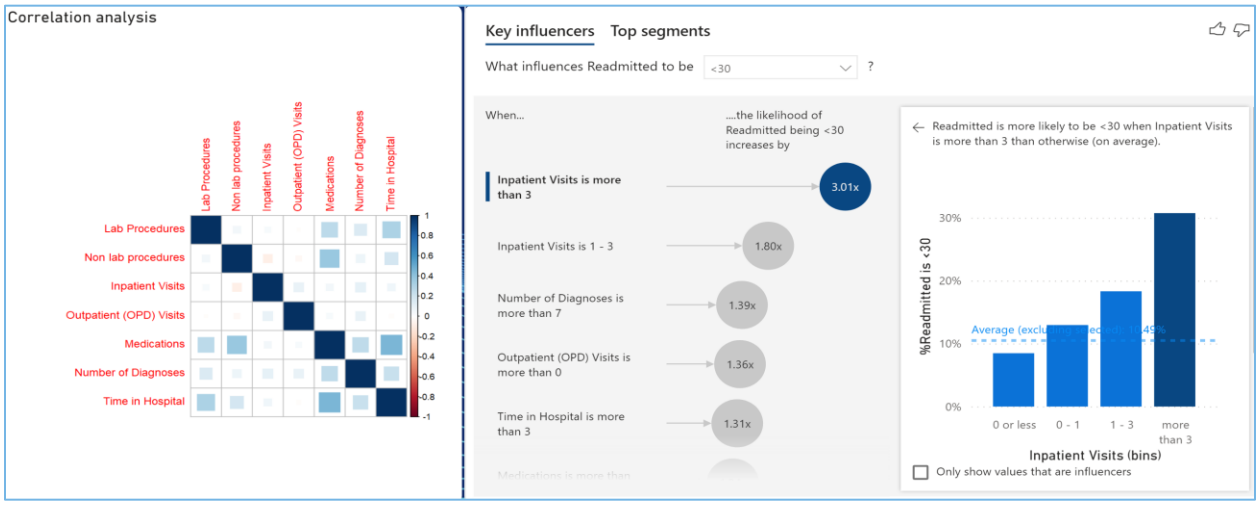
- We analyzed the impact of over 24 medications that were administered to patients by the physician.



- The analysis on the impact of readmission was done at levels of dose reduction, Steady dose, and increasing the dose

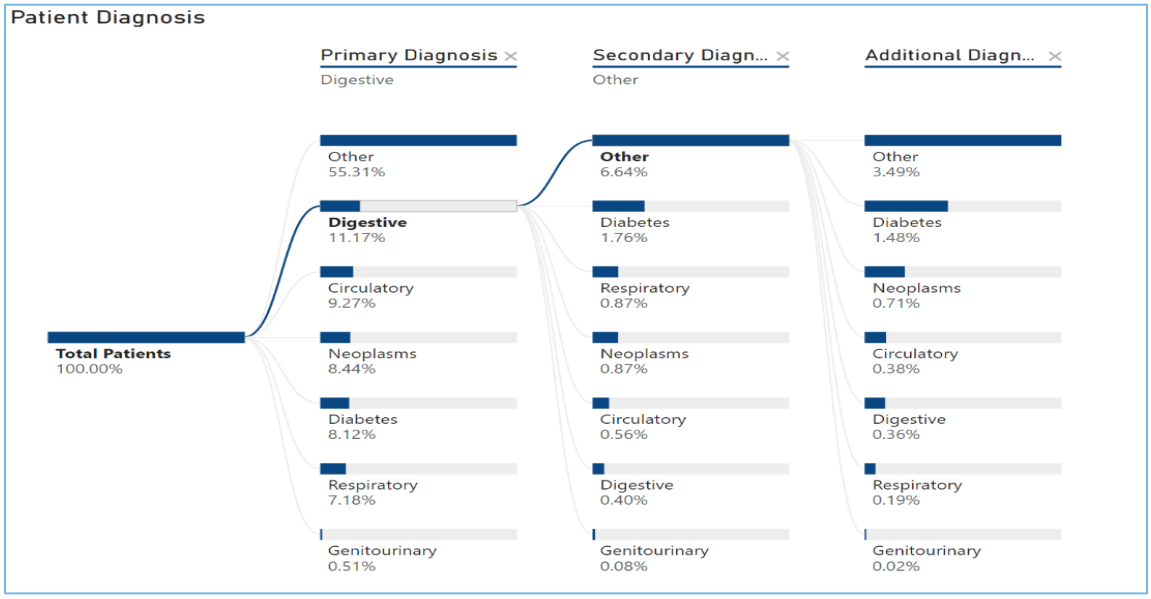
CORRELATION ANALYSIS AND ADMINISTRATIVE INTELLIGENCE

- Lab Procedures have a direct correlation with medications, diagnosis, and time in Hospital.
- Medications are also observed to have a high correlation with Time spent in Hospital and non-lab procedures.
- An intelligent parameter-driven decision system was created to analyze the impact of key influencers on readmission rates.



DIAGNOSTIC MODELLING & DECISION ANALYSIS

- Captured 3 levels of diagnosis and created an information-based decision tree model to understand if multiple diagnostics are required to provide optimum patient analysis.
- Diabetes is diagnosed in the primary or secondary diagnosis has shown that other three levels of diagnosis such as circulatory/neoplasms/respiratory are minimized.
- This is an interactive chart using which we can identify different diagnostic patterns which can help doctors understanding patient's condition better.



CONCLUSION

- *Our data analysis model helps in making recommendations which would help a diabetic patient to stay healthy and avoid readmission into a hospital.*

- *Our model and visualization analysis find similar patterns in readmitted and non-readmitted patients.*

- *We suggest that patients who are diagnosed as diabetic, aged above 40, being Outpatient, sugar level in hemoglobin more than 8, admitted in an emergency, spent less number of days in the hospital for treatment, has low insulin level in the body, used less medication (in different combinations of these factors) are more likely to get readmitted.*

- *People who are spending more time in the hospital, being inpatient, taking a greater number of medications, a greater number of procedures, with age less than 40, (in different combinations of these factors) are less likely to readmit to the hospital again.*

APPENDIX

IMAGE 1

```
-- Data Summary ------------------------
                           Values
Name                       mData
Number of rows             101766
Number of columns          50
___________________________
Column type frequency:
  factor                   37
  numeric                  13
___________________________
Group variables            None
```

IMAGE 2

```
-- Variable type: numeric -------------------------------------------------------
# A tibble: 13 x 11
   skim_variable            n_missing complete_rate          mean            sd
 * <chr>                        <int>         <dbl>         <dbl>         <dbl>
 1 encounter_id                     0             1 165201646.    102640296.
 2 patient_nbr                      0             1  54330401.     38696359.
 3 admission_type_id                0             1         2.02          1.45
 4 discharge_disposition_id         0             1         3.72          5.28
 5 admission_source_id              0             1         5.75          4.06
 6 time_in_hospital                 0             1         4.40          2.99
 7 num_lab_procedures               0             1        43.1          19.7
 8 num_procedures                   0             1         1.34          1.71
 9 num_medications                  0             1        16.0           8.13
10 number_outpatient                0             1         0.369         1.27
11 number_emergency                 0             1         0.198         0.930
12 number_inpatient                 0             1         0.636         1.26
13 number_diagnoses                 0             1         7.42          1.93
```

IMAGE 3

```
-- Variable type: factor ------------------------------------------------------
# A tibble: 37 x 6
   skim_variable            n_missing complete_rate ordered n_unique
 * <chr>                        <int>         <dbl> <lgl>      <int>
 1 race                             0             1 FALSE          6
 2 gender                           0             1 FALSE          3
 3 age                              0             1 FALSE         10
 4 weight                           0             1 FALSE         10
 5 payer_code                       0             1 FALSE         18
 6 medical_specialty                0             1 FALSE         73
 7 diag_1                           0             1 FALSE        717
 8 diag_2                           0             1 FALSE        749
 9 diag_3                           0             1 FALSE        790
10 max_glu_serum                    0             1 FALSE          4
11 A1Cresult                        0             1 FALSE          4
12 metformin                        0             1 FALSE          4
13 repaglinide                      0             1 FALSE          4
14 nateglinide                      0             1 FALSE          4
15 chlorpropamide                   0             1 FALSE          4
16 glimepiride                      0             1 FALSE          4
17 acetohexamide                    0             1 FALSE          2
18 glipizide                        0             1 FALSE          4
19 glyburide                        0             1 FALSE          4
20 tolbutamide                      0             1 FALSE          2
21 pioglitazone                     0             1 FALSE          4
22 rosiglitazone                    0             1 FALSE          4
23 acarbose                         0             1 FALSE          4
24 miglitol                         0             1 FALSE          4
25 troglitazone                     0             1 FALSE          2
26 tolazamide                       0             1 FALSE          3
27 examide                          0             1 FALSE          1
28 citoglipton                      0             1 FALSE          1
29 insulin                          0             1 FALSE          4
30 glyburide.metformin              0             1 FALSE          4
31 glipizide.metformin              0             1 FALSE          2
32 glimepiride.pioglitazone         0             1 FALSE          2
33 metformin.rosiglitazone          0             1 FALSE          2
34 metformin.pioglitazone           0             1 FALSE          2
35 change                           0             1 FALSE          2
36 diabetesMed                      0             1 FALSE          2
37 readmitted                       0             1 FALSE          3
```

IMAGE 4

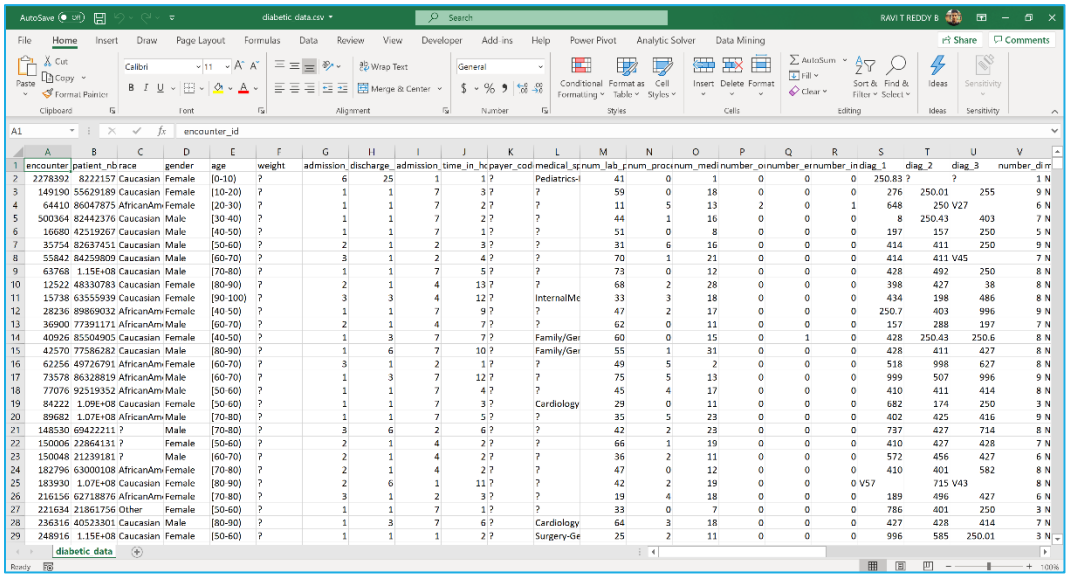


IMAGE 5

```
114882984 48330783 63555939 89869032 ...
 $ race                    : Factor w/ 6 levels "?","AfricanAmeric
2 ...
 $ gender                  : Factor w/ 3 levels "Female","Male",..
 $ age                     : Factor w/ 10 levels "[0-10)","[10-20)
5 ...
 $ admission_type_id       : Factor w/ 8 levels "1","2","3","4",..
 $ discharge_disposition_id: Factor w/ 26 levels "1","2","3","4",.
...
 $ admission_source_id     : Factor w/ 17 levels "1","2","3","4",.
...
 $ time_in_hospital        : int  3 2 2 1 3 4 5 13 12 9 ...
 $ num_lab_procedures      : int  59 11 44 51 31 70 73 68 33 47 ..
 $ num_procedures          : int  0 5 1 0 6 1 0 2 3 2 ...
 $ num_medications         : int  18 13 16 8 16 21 12 28 18 17 ...
 $ number_outpatient       : int  0 2 0 0 0 0 0 0 0 0 ...
 $ number_emergency        : int  0 0 0 0 0 0 0 0 0 0 ...
 $ number_inpatient        : int  0 1 0 0 0 0 0 0 0 0 ...
 $ diag_1                  : Factor w/ 717 levels "?","10","11",..
278 254 284 122 ...
 $ diag_2                  : Factor w/ 749 levels "?","11","110",.
316 262 48 243 ...
 $ diag_3                  : Factor w/ 790 levels "?","11","110",.
88 231 319 668 ...
```

IMAGE 6

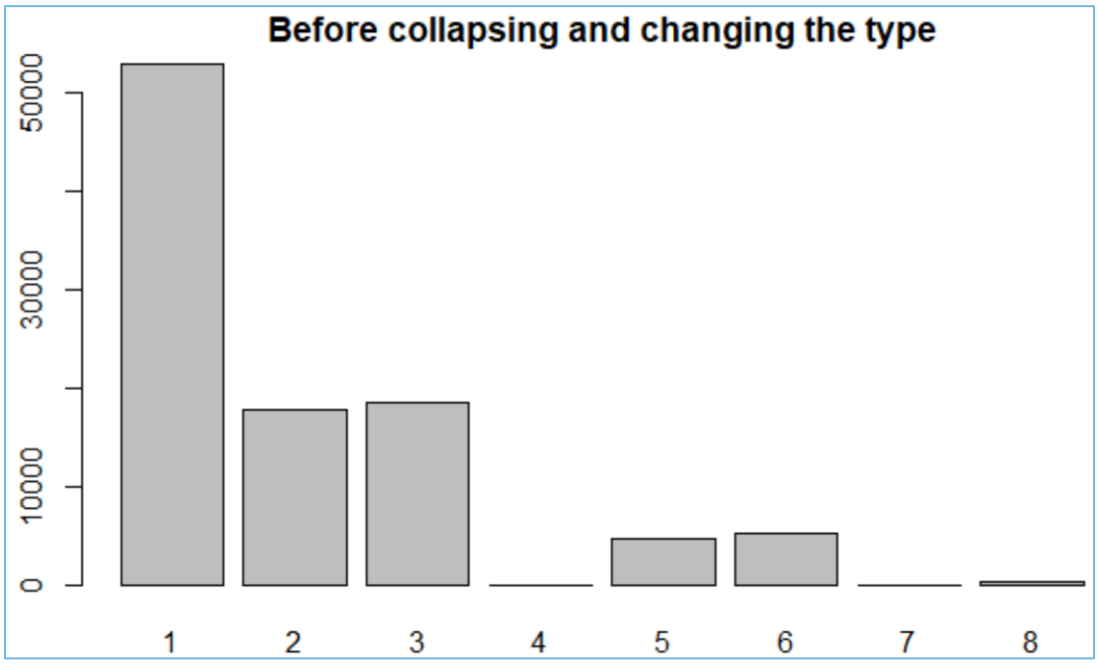


IMAGE 7

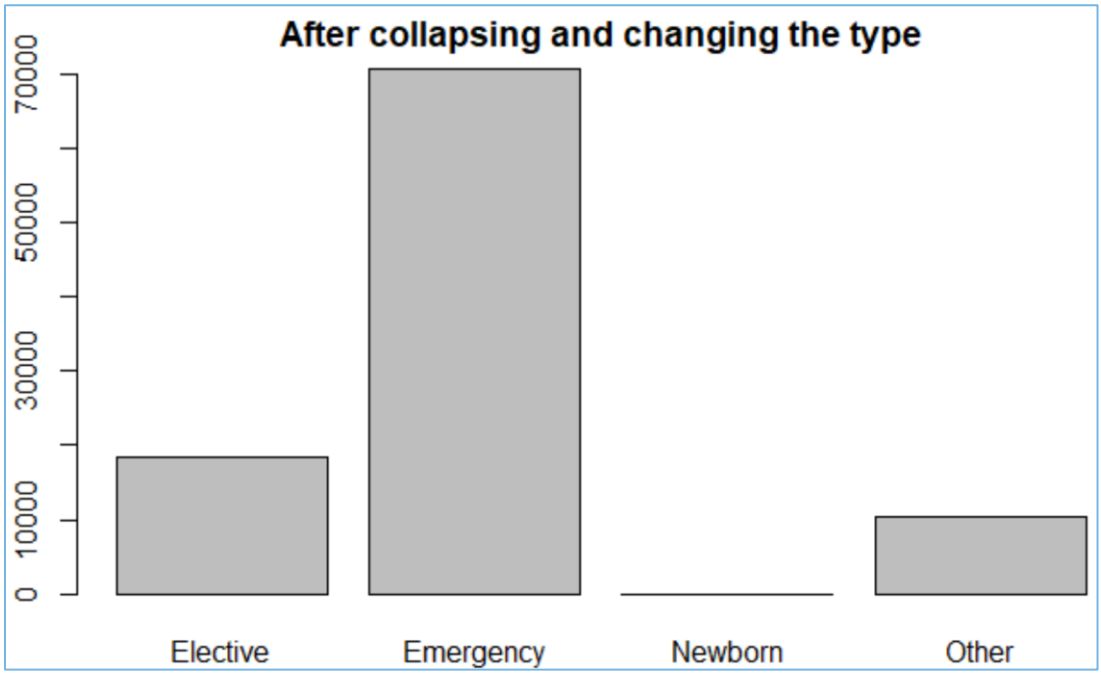


IMAGE 8

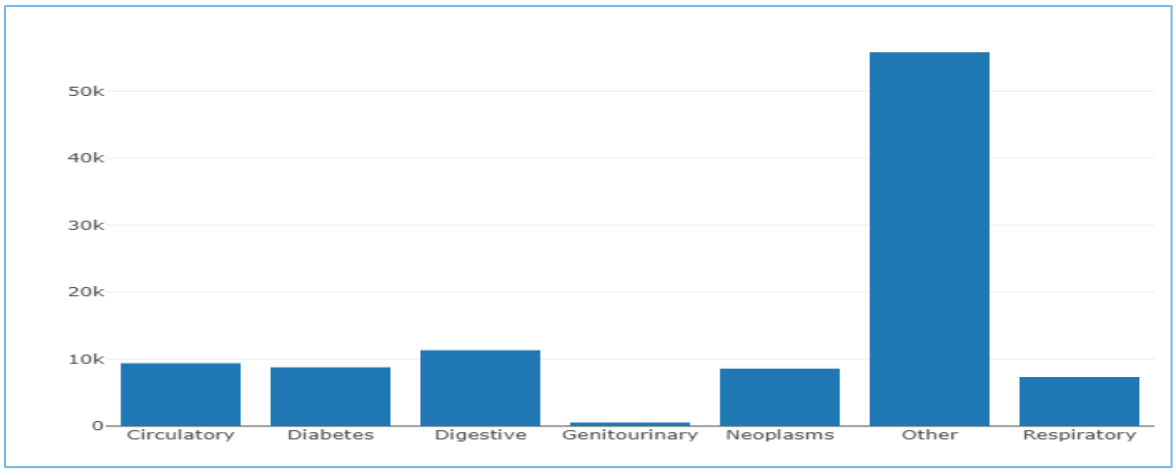


IMAGE 9

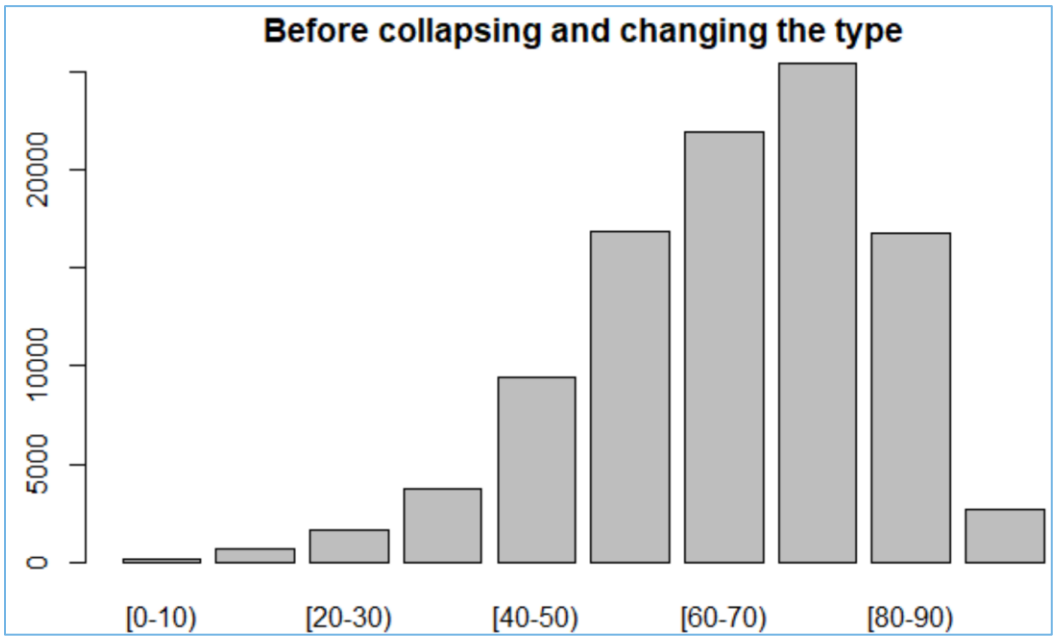


IMAGE 10

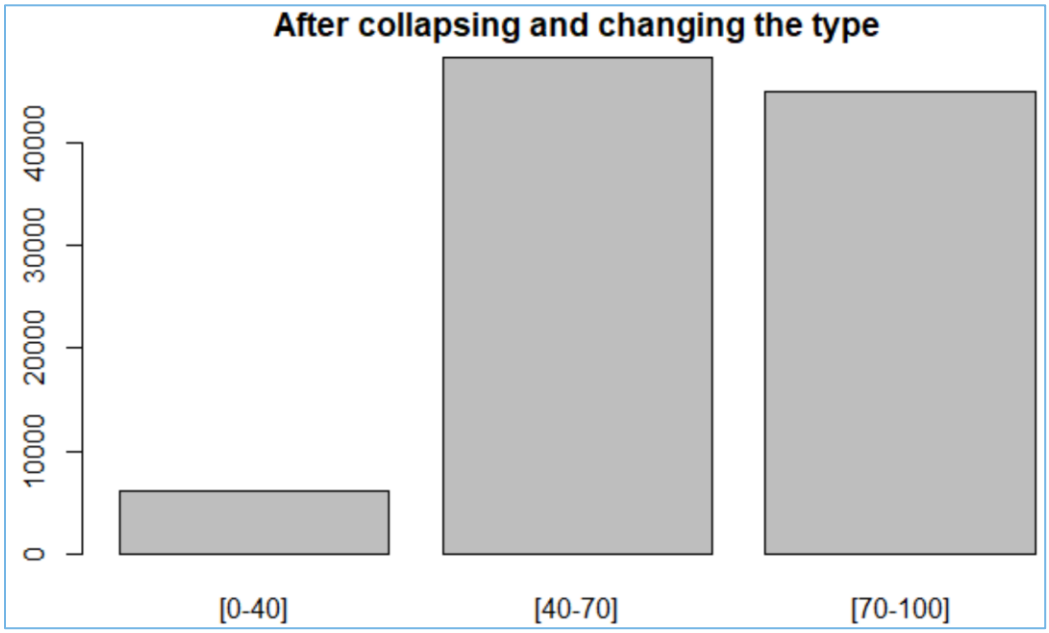


IMAGE 11

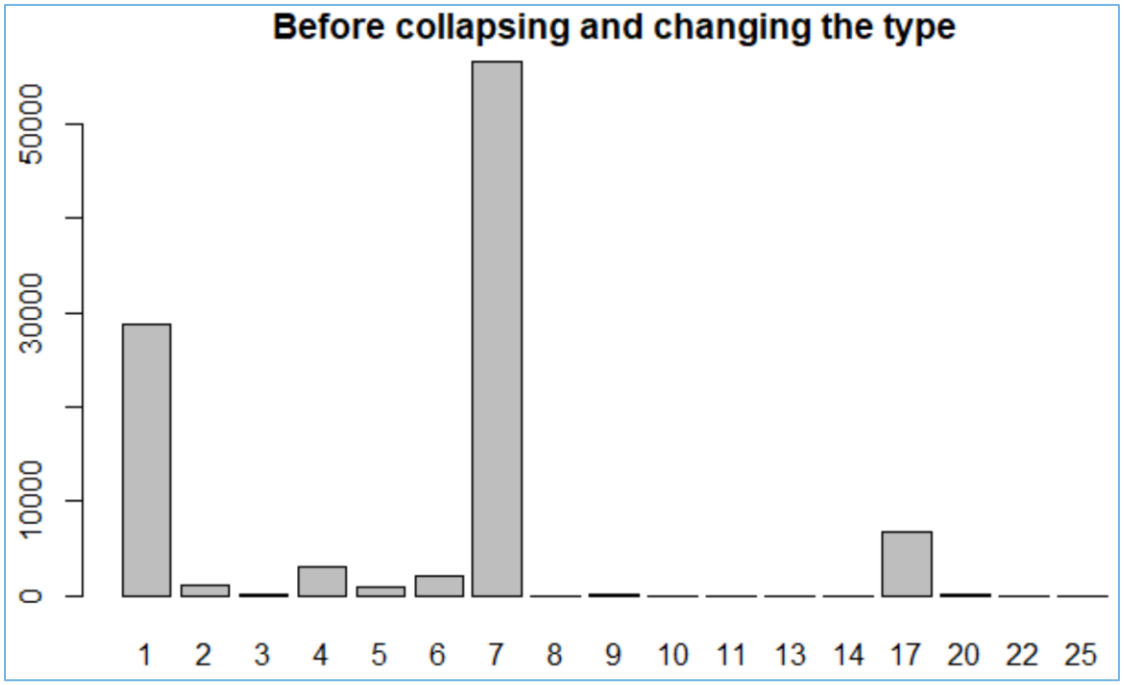


IMAGE 12

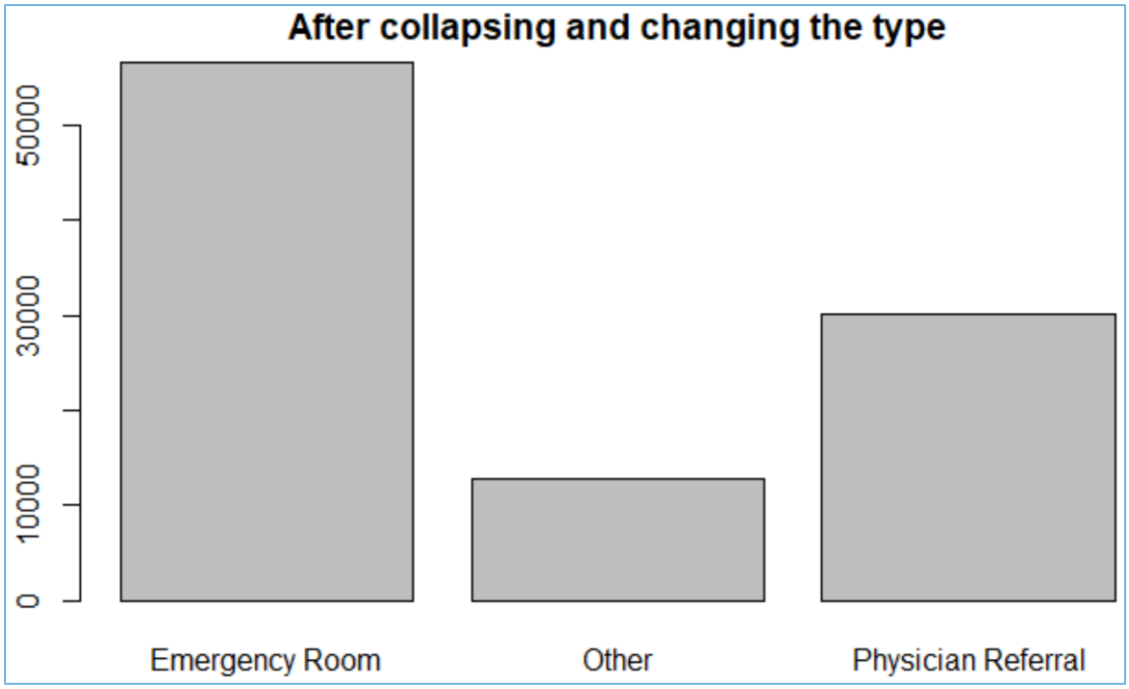


IMAGE 13

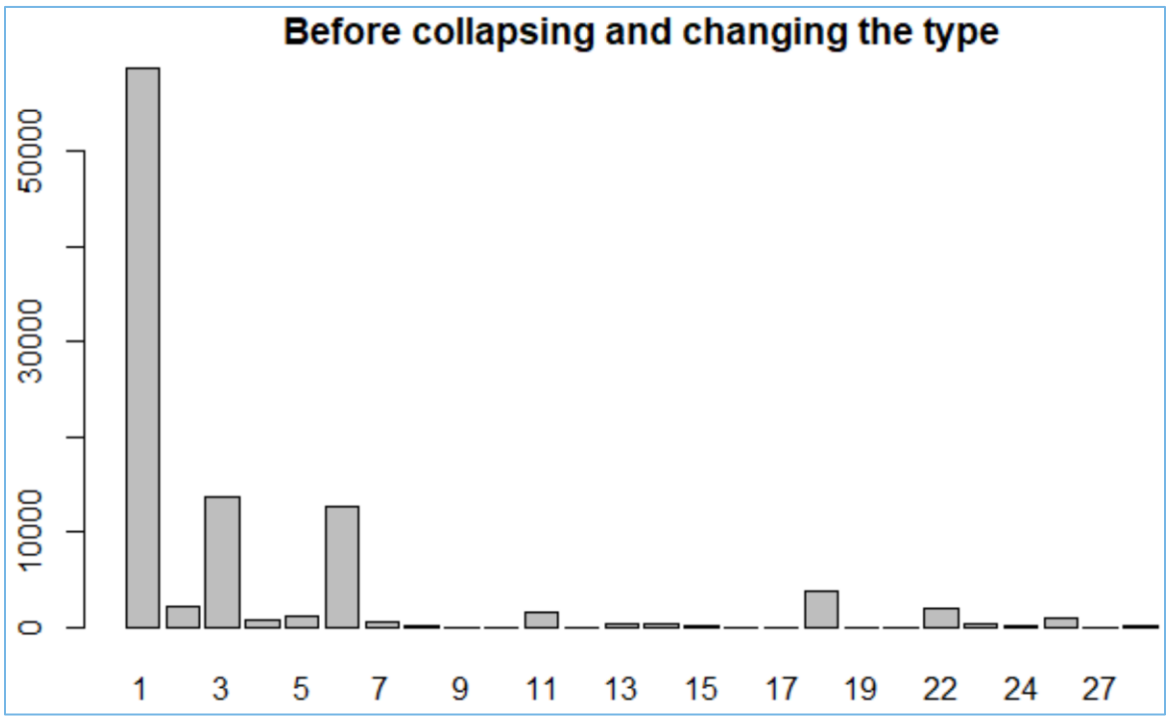


IMAGE 14

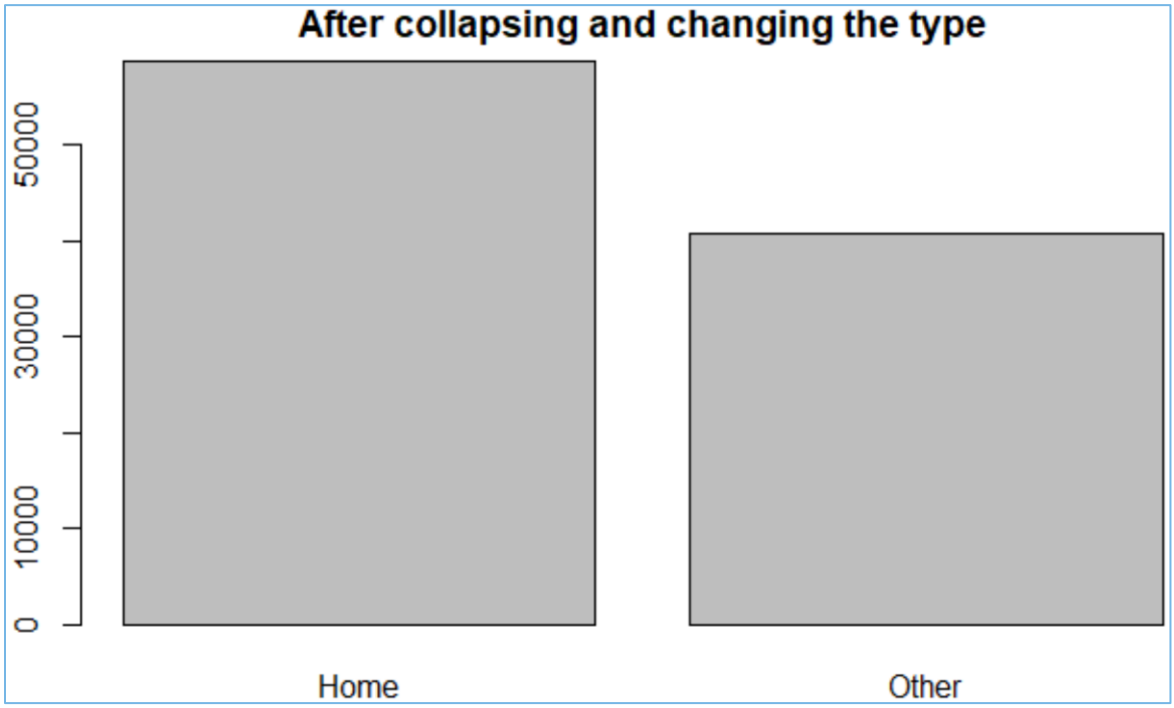


IMAGE 15

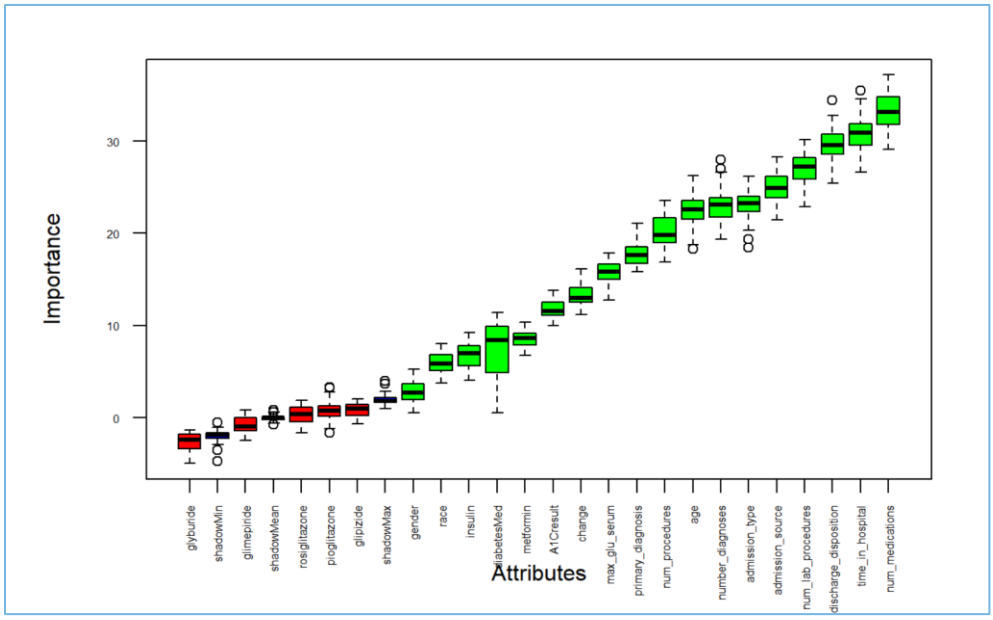


IMAGE 16

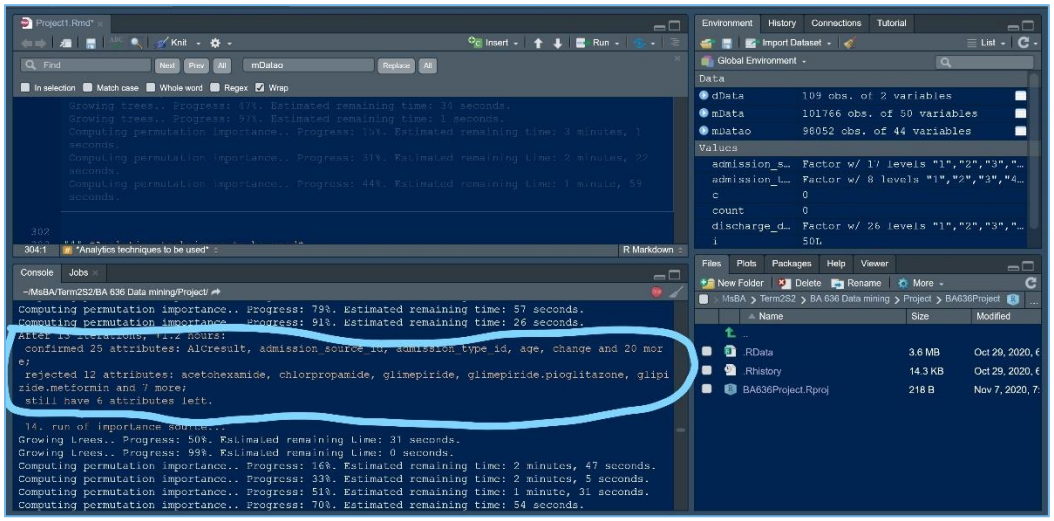


IMAGE 17

```
race 0
gender 0
age 0
admission_type_id 0
discharge_disposition_id 0
admission_source_id 0
time_in_hospital 0
num_lab_procedures 0
num_procedures 0
num_medications 0
number_outpatient 0
number_emergency 0
number_inpatient 0
number_diagnoses 0
max_glu_serum 0
A1Cresult 0
metformin 0
insulin 0
change 0
diabetesMed 0
readmitted 0
primary_diagnosis 0
```

IMAGE 18

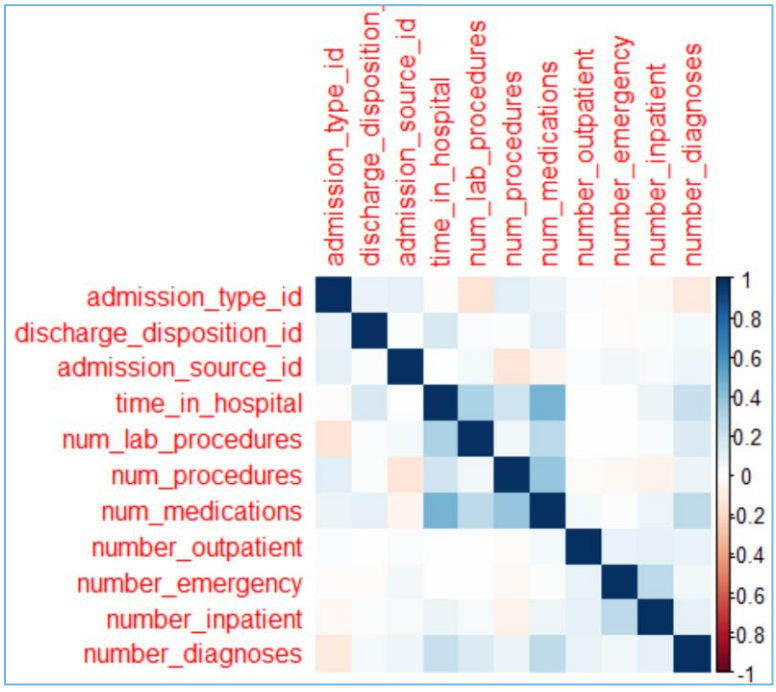


IMAGE 19

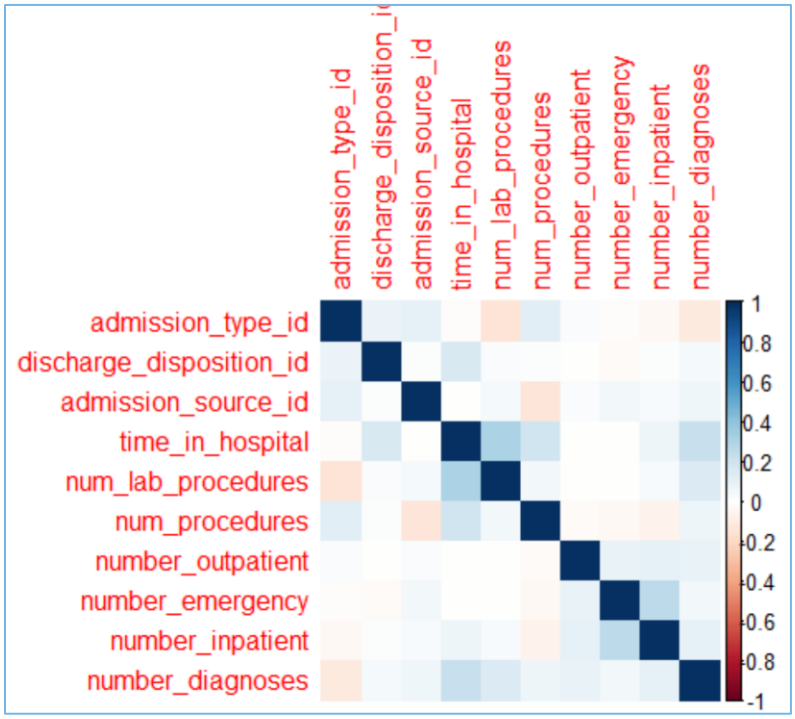


IMAGE 20

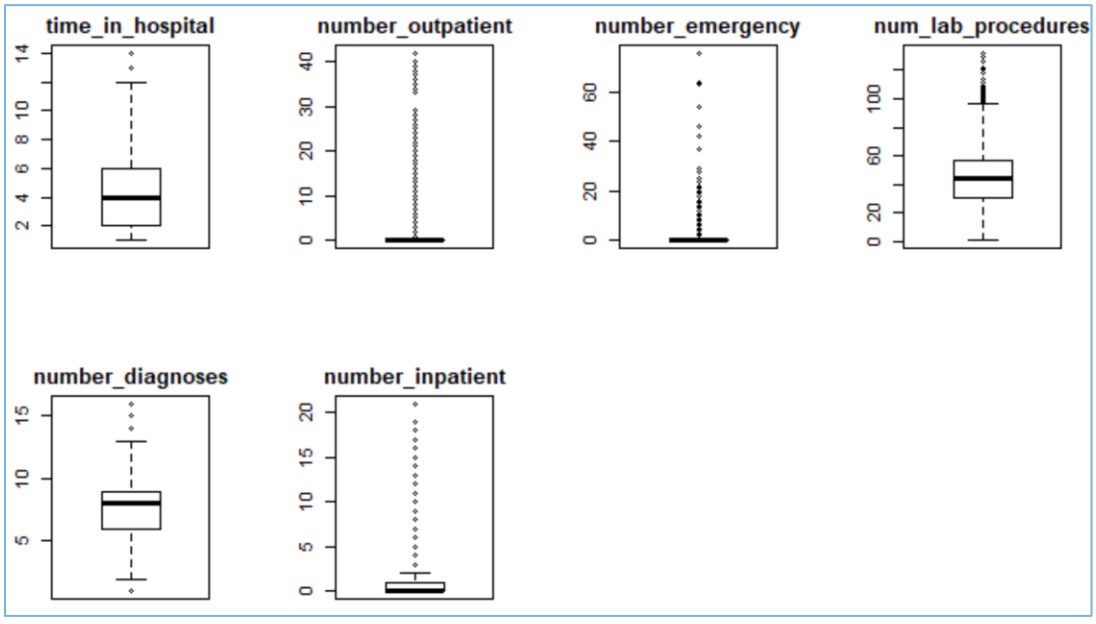


IMAGE 21

```
time_in_hospital <= 12)
number_outpatient <= 30)
number_emergency <= 60)
num_lab_procedures <= 110)
number_diagnoses <= 13)
number_inpatient <= 20)
num_procedures <= 5)
```

IMAGE 22

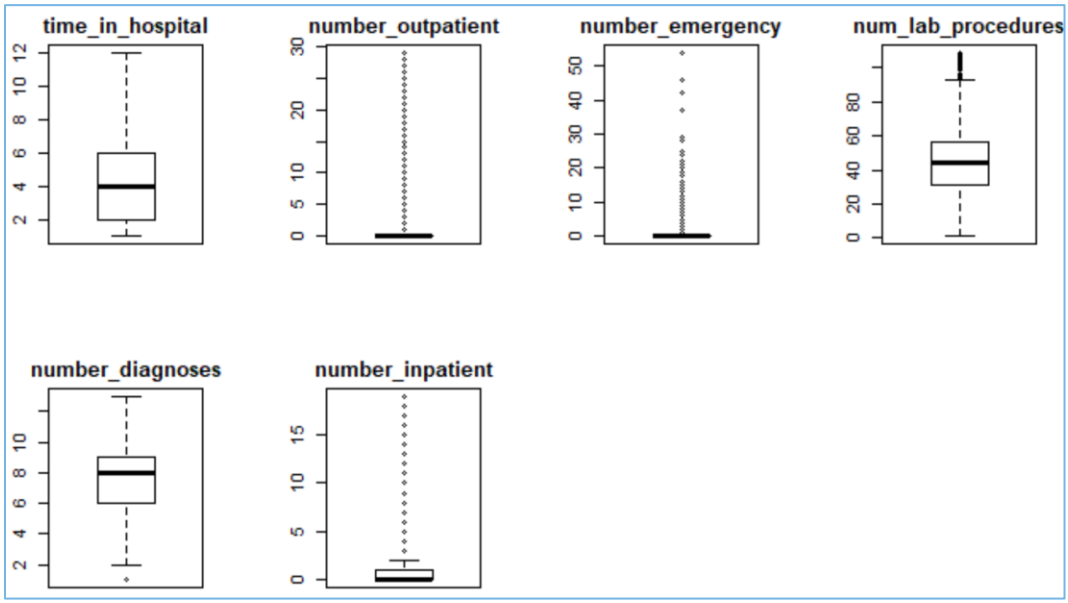


IMAGE 23

```
select_columns = c(
    "race", "gender", "age", "admission_type",
    "discharge_disposition", "admission_source",
    "max_glu_serum", "A1Cresult", "metformin",
    "insulin", "diabetesMed", "primary_diagnosis"
),
```

IMAGE 24

```
set.seed(32)
samp <- sample(1:nrow(mData_dummies),round(0.60*nrow(mData_dummies)))
DF.training <- mData_dummies[samp,]
DF.val <- mData_dummies[-samp,]
```

IMAGE 25

```
Coefficients:
                                                Estimate
(Intercept)                                   -1.6946915
time_in_hospital                               0.0267869
num_lab_procedures                             0.0007589
num_procedures                                -0.0478967
number_outpatient                              0.0672648
number_emergency                               0.2140451
number_inpatient                               0.3564435
number_diagnoses                               0.0705051
changeNo                                      -0.0588191
race_Asian                                    -0.3795661
race_Caucasian                                 0.0358515
race_Hispanic                                 -0.1500674
race_Other                                    -0.2383021
gender_Male                                   -0.0793945
`age_[40-70]`                                  0.2093167
`age_[70-100]`                                 0.2752967
admission_type_Emergency                       0.0649551
admission_type_Newborn                        -9.5847594
admission_type_Other                           0.4965651
discharge_disposition_Other                   -0.0877113
admission_source_Other                        -0.4906249
`admission_source_Physician Referral`         -0.0795829
`A1Cresult_>8`                                 0.1165904
A1Cresult_None                                 0.1009190
metformin_No                                   0.1751332
insulin_Steady                                -0.1719727
insulin_Up                                    -0.0997649
diabetesMed_Yes                                0.3427795
primary_diagnosis_Diabetes                     0.2033131
primary_diagnosis_Genitourinary               -0.5582656
primary_diagnosis_Other                        0.0844913
```

IMAGE 26

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 15476  9920
         1  4158  7507

               Accuracy : 0.6201
                 95% CI : (0.6152, 0.6251)
    No Information Rate : 0.5298
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.2231

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.4308
            Specificity : 0.7882
         Pos Pred Value : 0.6435
         Neg Pred Value : 0.6094
             Prevalence : 0.4702
         Detection Rate : 0.2026
   Detection Prevalence : 0.3148
      Balanced Accuracy : 0.6095

       'Positive' Class : 1
```

IMAGE 27

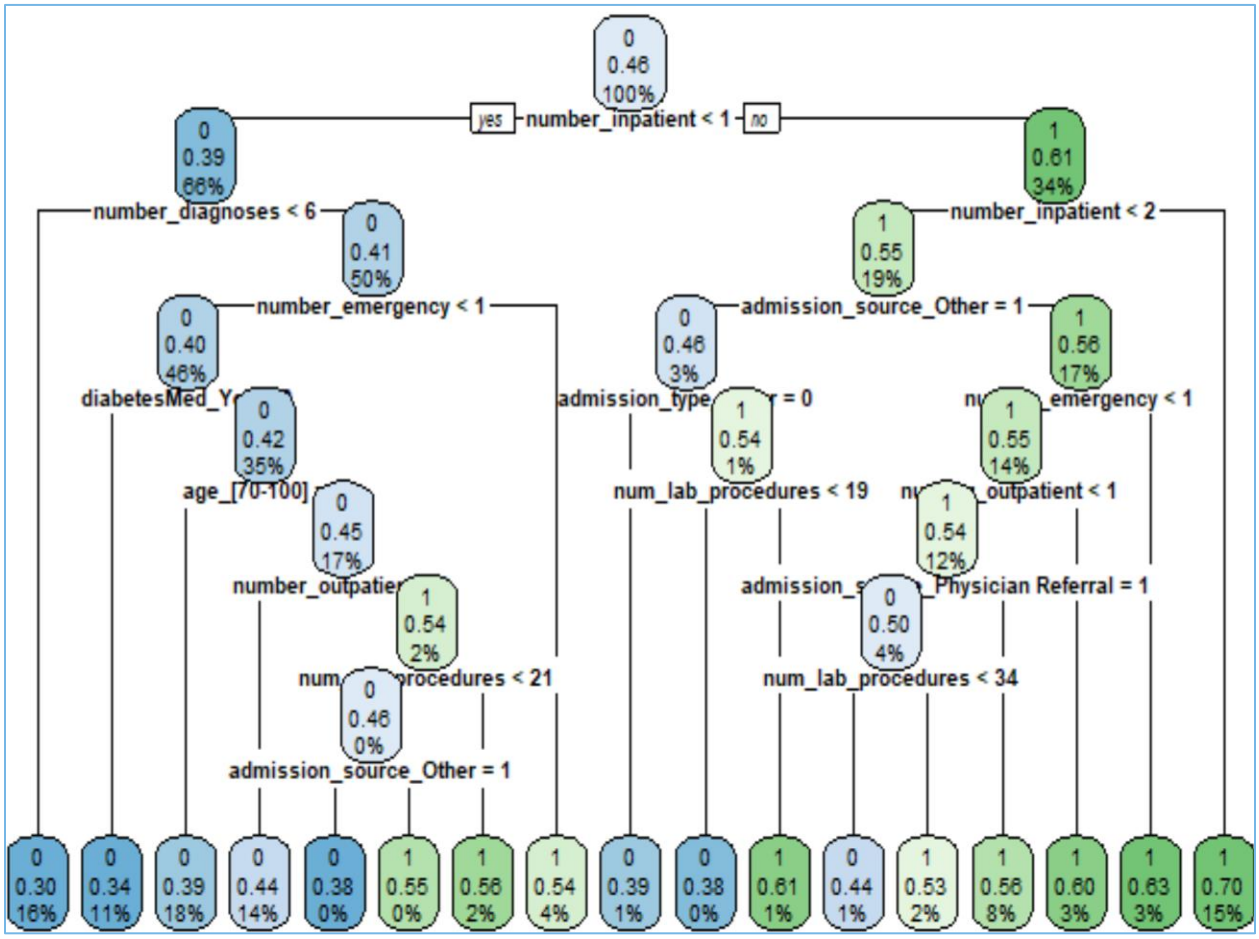


IMAGE 28

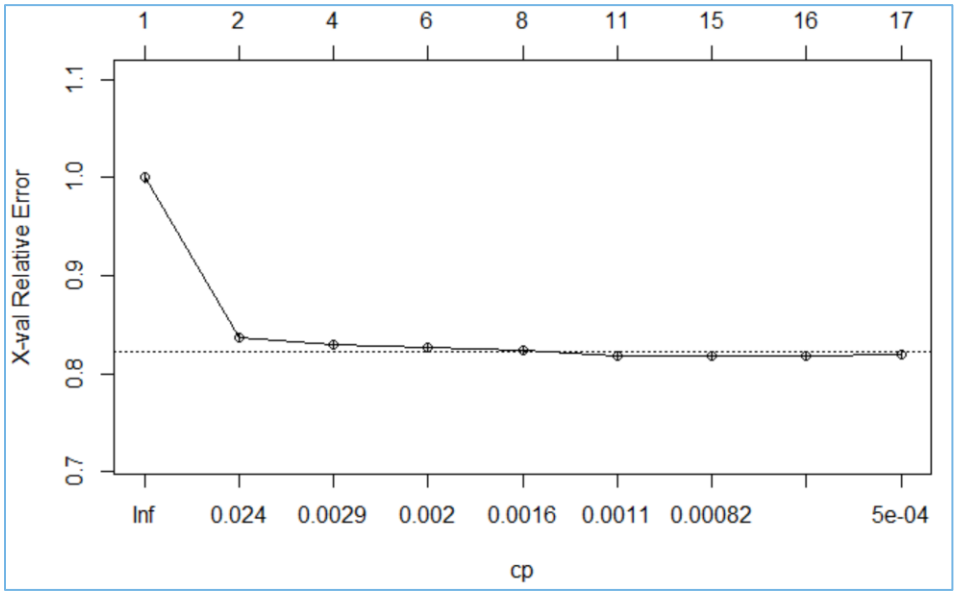


IMAGE 29

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 19634 17427
         1     0     0

               Accuracy : 0.5298
                 95% CI : (0.5247, 0.5349)
    No Information Rate : 0.5298
    P-Value [Acc > NIR] : 0.5021

                  Kappa : 0

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

            Sensitivity : 0.0000
            Specificity : 1.0000
         Pos Pred Value :    NaN
         Neg Pred Value : 0.5298
             Prevalence : 0.4702
         Detection Rate : 0.0000
   Detection Prevalence : 0.0000
      Balanced Accuracy : 0.5000

       'Positive' Class : 1
```

IMAGE 30

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0  3301  1593
         1 16333 15834

               Accuracy : 0.5163
                 95% CI : (0.5112, 0.5214)
    No Information Rate : 0.5298
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0732

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

            Sensitivity : 0.16813
            Specificity : 0.90859
         Pos Pred Value : 0.67450
         Neg Pred Value : 0.49224
             Prevalence : 0.52978
         Detection Rate : 0.08907
   Detection Prevalence : 0.13205
      Balanced Accuracy : 0.53836

       'Positive' Class : 0
```

IMAGE 31

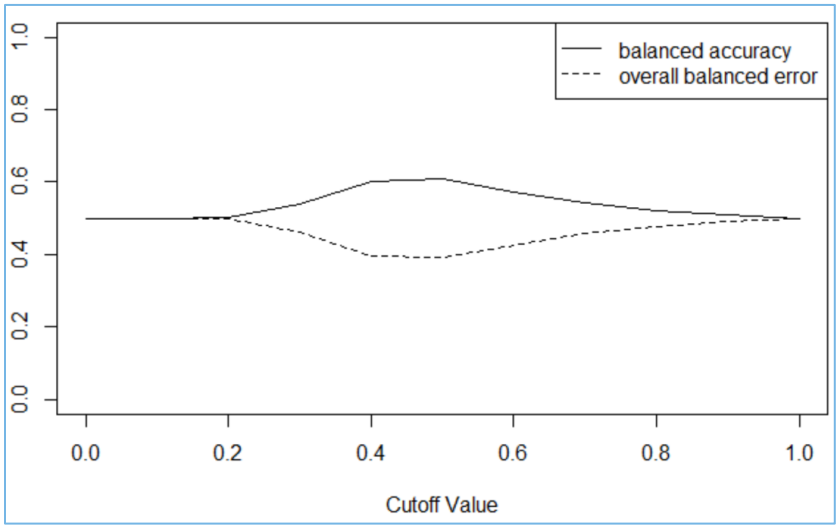


IMAGE 32

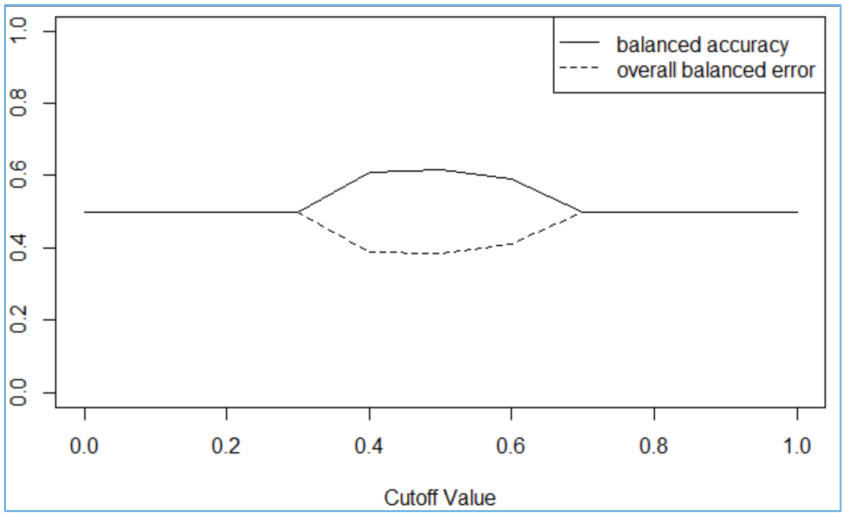
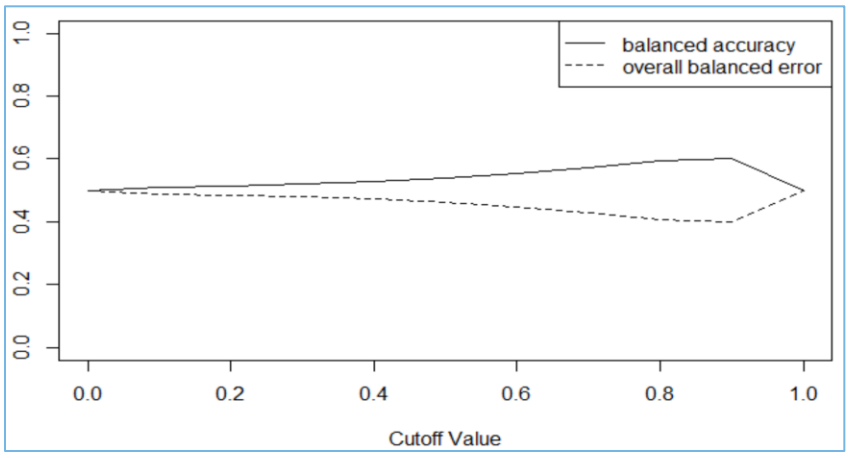


Image 33



DATA DICTIONARY

Variable	Data type	Description
race	Factor	Race of the patient
Gender	Factor	Gender of patient
age	Factor	Age of patient
admission_type	Factor	Emergency, Elective, Newborn, Other
discharge_disposition	Factor	Discharged to home or Other location
admission_source	Factor	Physician Referral, Other, Emergency Room
time_in_hospital	Numeric	Time spent in hospital in months
num_lab_procedures	Numeric	Number of lab procedures
num_procedures	Numeric	Non-lab procedures
num_medications	Numeric	Number of medications
number_emergency	Numeric	Number of times admitted as an emergency
number_outpatient	Numeric	Number of times admitted as an outpatient
number_inpatient	Numeric	Number of times admitted as an inpatient
number_diagnoses	Numeric	Number of diagnoses done
max_glu_serum	Factor	Glucose serum test result
A1Cresult	Factor	Hb A1C or hemoglobin A1c (shows sugar level in blood)
metformin	Factor	One of the administered medicinal drugs
insulin	Factor	One of the administered medicinal drugs
change	Factor	Change of medication
diabetes med	Factor	Diabetes medications
readmitted	Factor	Readmission to hospital
primary_diagnosis	Factor	Type of disease diagnosed (Circulatory, Diabetes, Digestive, Genitourinary, Neoplasms, Other, Respiratory)