Predicting the “Re-admission possibility of a patient into the hospital” & Pattern Extraction

Business Case:

Every business wants to be at top in its own domain but the issue with health care industry mainly with hospitals is that patient’s readmission with in 30days due to which the hospitals image is at stake.

Problem statement:

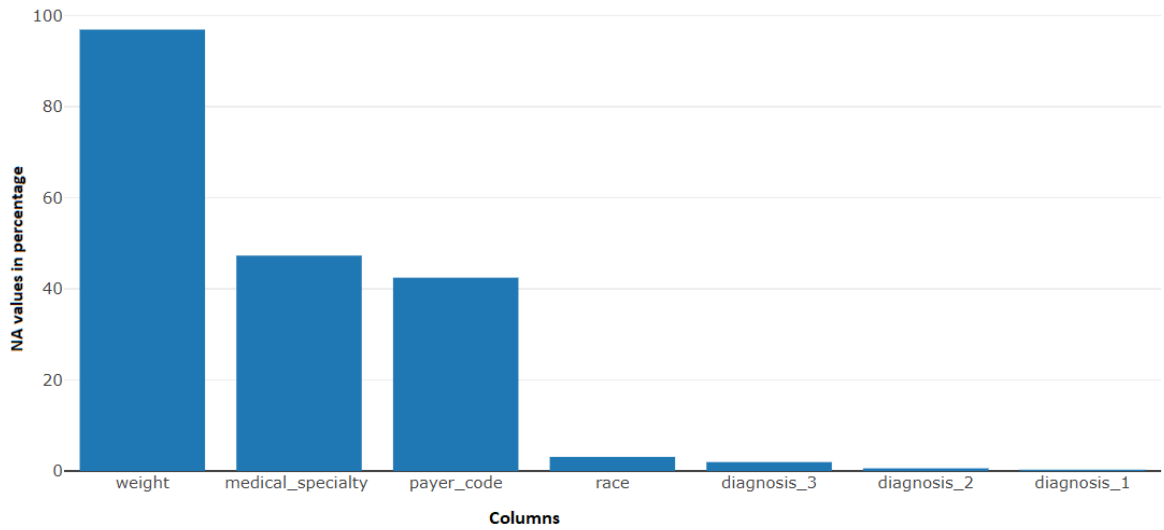
A leading hospital in the US is suddenly seeing increase in the patient readmission in less than 30 days. This is serious concern for the hospital as it may indicate insufficient treatment or diagnosis when the patient was admitted first and later released under clean bill of health. Not only the image of hospital as healthcare provider is compromised, this is also increased cost to the entire medicare ecosystem in form of increased insurance claims. So it is in Hospital’s interest to support their diagnosis by a better predictive model which you are going to build.

Objective:

Classify the patients treated by hospital into two primary categories:

- Readmitted within 30 days
- Not readmitted

NA values in Data:

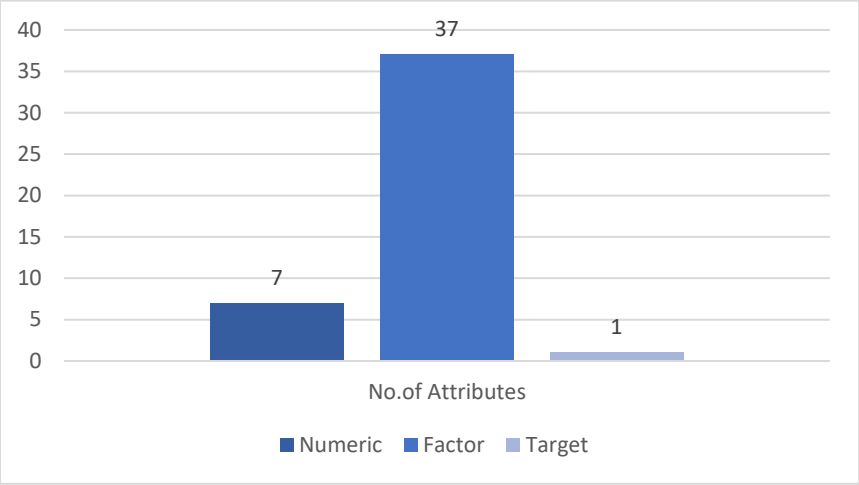


Attributes distribution:

Total number of different types of attributes present in the data are:

- Numeric : 7
- Factor : 37
- Target : 1

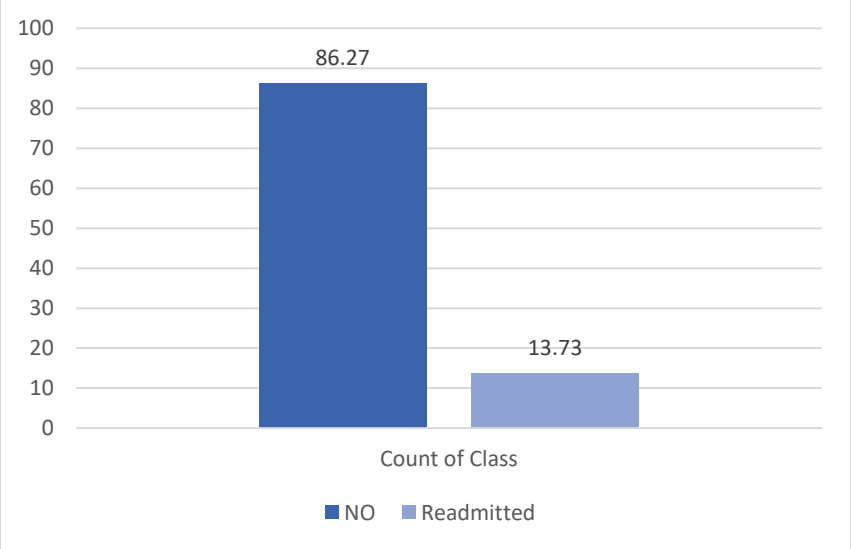
Where Target is binary class.



Target distribution:

How target class is distributed:

- No : 86.27%
- Readmitted: 13.73%



Data Cleaning Process:

Data cleaning is commonly defined as the process of detecting and correcting corrupt or inaccurate records from a dataset, table, or database.

- Data quality is an important component in any data mining efforts. For this reason, many data scientists spend most of their time preparing and cleaning their data before it can be mined for insights.
- There are four broad categories of data quality problems:
    1. missing data
    2. abnormal data (outliers)
    3. departure from models
    4. goodness-of-fit
- For this project we will be using some resampling techniques to handle the imbalance data

Feature Engineering and Variables Extraction:

Feature engineering is the process of using domain knowledge of the data to create features that make machine learning algorithms work. Feature engineering is fundamental to the application of machine learning and is both difficult and expensive.

As the drug variables which are not prescribed most of the time due to which there is a huge loss of information.

Most of the time the class No is prevailing in all the variables due to which the information which has to be extracted to predict the readmission.

Drugs clubbed:

- **Sulfonylurea:** Chlorpropamide, Glimepiride, Acetohexamide, Glipizide, Glyburide, Tolbutamide, Tolazamide, Glyburide Metformin, Glipizide Metformin.

- **Meglitinides:** Repaglinide, Met glinide.

- **Thiazolidinediones:** Pioglitazone, Rosiglitazone, Troglitazone, Metformin Rosiglitazone, Metformin Pioglitazone.

- **Biguanide:** Metformin, Glyburide Metformin, Glipizide Metformin, Metformin Rosiglitazone, Metformin Pioglitazone.

- **Glucosides:** Acarbose, Miglitol.

- **Insulin:** Insulin.

We will be just clubbing all this variable under one common molecule name and just break them into two class whether it was prescribed or not.

No of Days Stayed and Month:

This variable was being extracted based on the admission date and discharge date.

Stayed= discharge – admission

Month from the Admission Date.

Releveling Attributes:

We have various attributes which were having huge number of levels. In order to avoid the huge levels, we will be using some techniques to overcome this.

admission_type_id: As this had 8 levels we made it down to 2 levels such as casual and emergency were Emergency, Urgent and Trauma Centre into one class and others too Casual.

discharge_disposition_id: As this had 29 levels and out of which 4 levels were discarded and others were clubbed into 14 levels.

admission_source_id: As this had 26 levels which were clubbed into 12 levels.

Age: As it was in ordinal form in the given data we converted them into numeric by just taking mean into account.

A1Cresult and Ma Gluc serum: Just did a simple change of converting the old levels into normal, abnormal and not tested.

ICD code Levelling:

The three variables which were diagnosis 1, diagnosis 2, diagnosis 3 where above 600 levels are been clubbed into 13 levels based on ICD 9 codes.

Handling Data Imbalance

What is Imbalanced Classification?

Imbalanced classification is a supervised learning problem where one class outnumbers other class by a large proportion. This problem is faced more frequently in binary classification problems than multi-level classification problems.

What are the methods to deal with imbalanced data sets?

The methods are widely known as 'Sampling Methods'. Generally, these methods aim to modify an imbalanced data into balanced distribution using some mechanism. The modification occurs by altering the size of original data set and provide the same proportion of balance.

Below are the methods used to treat imbalanced datasets:

1. Under sampling
2. Oversampling
3. Synthetic Data Generation

1. Under sampling

This method works with majority class. It reduces the number of observations from majority class to make the data set balanced. This method is best to use when the data set is huge and reducing the number of training samples helps to improve run time and storage troubles.

2. Oversampling

This method works with minority class. It replicates the observations from minority class to balance the data. It is also known as *up sampling*.

3. Synthetic Data Generation

In simple words, instead of replicating and adding the observations from the minority class, it overcome imbalances by generates artificial data. It is also a type of oversampling technique.

Model Building and Comparison

Based on the data which where been cleaned and further variables extraction models where been created.

Two Main data frames where been created:

1. Data with all variables.

2. Data with drugs dropped.

The models which where been chosen to predict the readmitted class are:

1. Logistic Regression
2. SVM
3. Naïve Bayes
4. Decision Tree
5. Ada Boost

Logistic Regression:

Logistic Regression is a classification algorithm. It is used to predict a binary outcome (1 / 0, Yes / No, True / False) given a set of independent variables.

To represent binary / categorical outcome, we use dummy variables. You can also think of logistic regression as a special case of linear regression when the outcome variable is categorical, where we are using log of odds as dependent variable. In simple words, it predicts the probability of occurrence of an event by fitting data to a logit function.

SVM:

Support vector machine algorithm is to find a hyperplane in an N-dimensional space(N — the number of features) that distinctly classifies the data points.

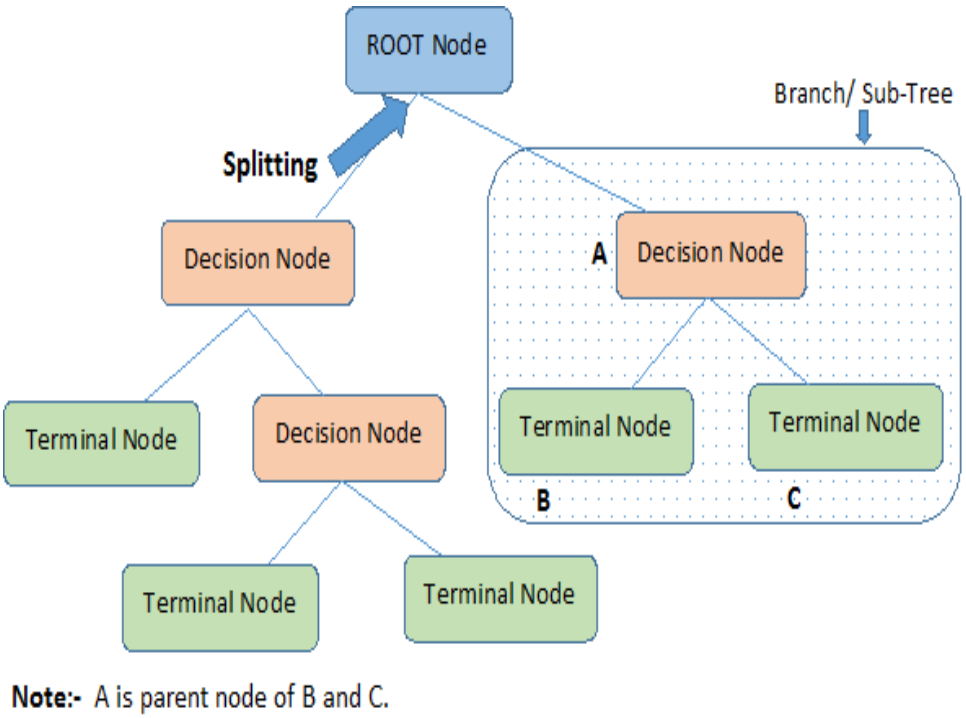
To separate the two classes of data points, there are many possible hyperplanes that could be chosen. Our objective is to find a plane that has the maximum margin, i.e the maximum distance between data points of both classes. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence.

Naïve Bayes:

It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as 'Naive'.

Decision Tree:

Decision tree is a type of supervised learning algorithm (having a pre-defined target variable) that is mostly used in classification problems. It works for both categorical and continuous input and output variables. In this technique, we split the population or sample into two or more homogeneous sets (or sub-populations) based on most significant splitter / differentiator in input variables.



Decision Tree Split

Adaboost:

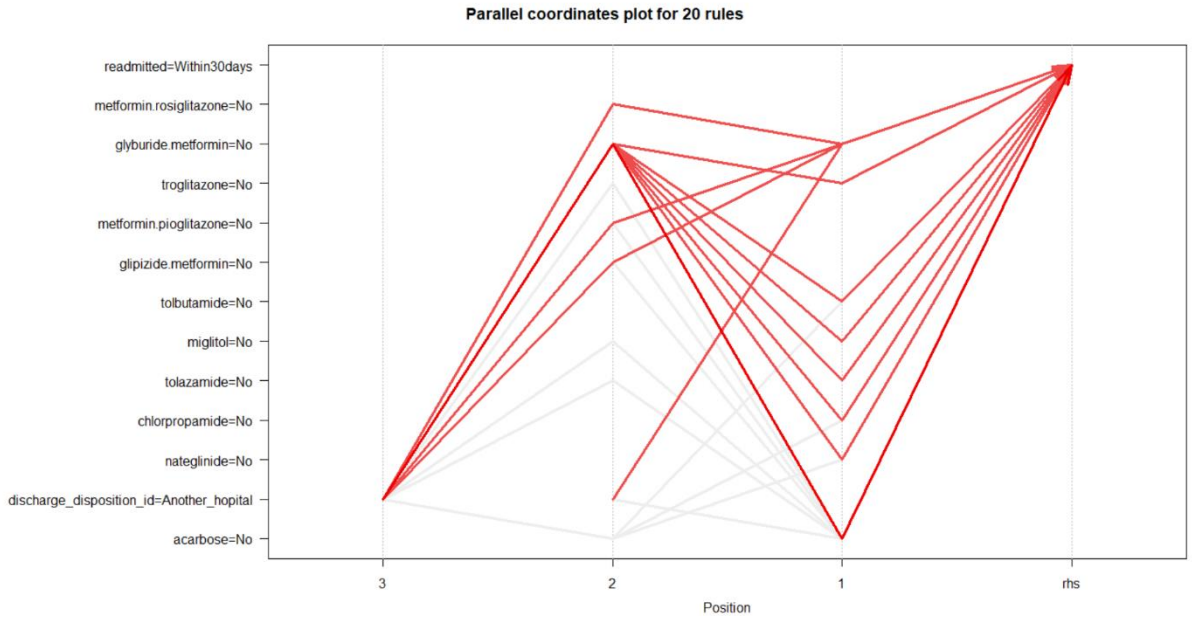
Adaptive Boosting, is a machine learning meta algorithm formulated by Yoav Freund and Robert Schapire, who won the 2003 Gödel Prize for their work. It can be used in conjunction with many other types of learning algorithms to improve performance. The output of the other learning algorithms ('weak learners') is combined into a weighted sum that represents the final output of the boosted classifier.

AdaBoost is adaptive in the sense that subsequent weak learners are tweaked in favour of those instances misclassified by previous classifiers. AdaBoost is sensitive to noisy data and outliers. In some problems it can be less susceptible to the overfitting problem than other learning algorithms. The individual learners can be weak, but as long as the performance of each one is slightly better than random guessing, the final model can be proven to converge to a strong learner.

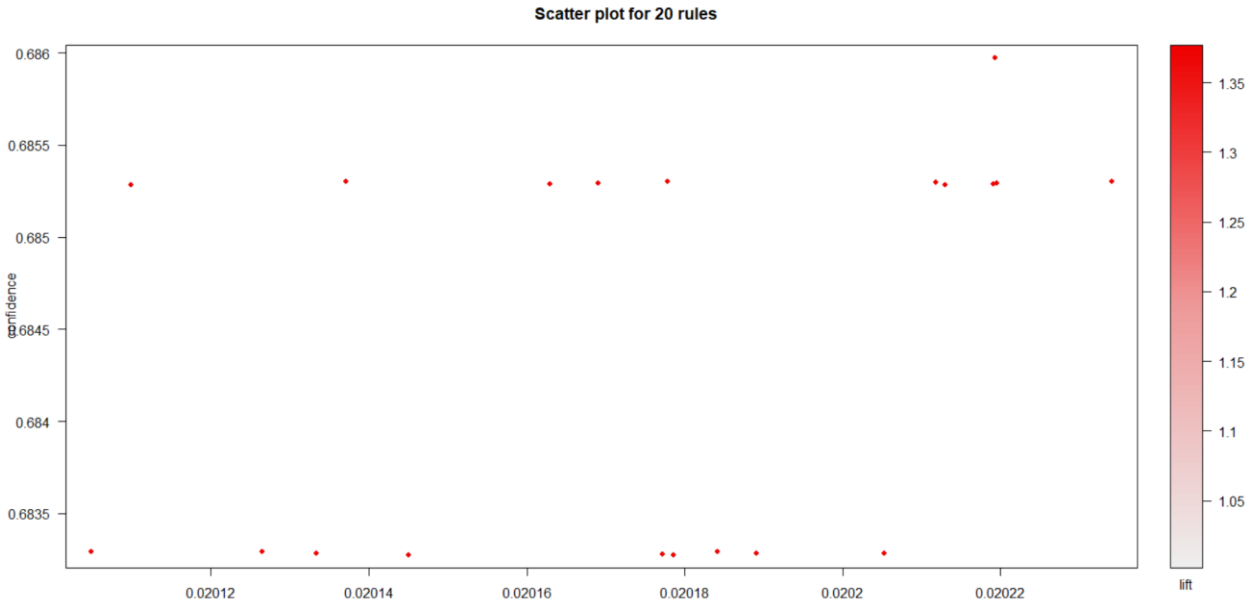
Pattern Extraction using the Association Rules Model:

rules	support	confidence	lift	count
{acarbose=No,glyburide.metformin=No,discharge_disposition_id=Another_hopital} => {readmitted=Within30days}	0.020173	0.685967	1.37471	699
{glyburide.metformin=No,discharge_disposition_id=Another_hopital} => {readmitted=Within30days}	0.020173	0.685294	1.373363	699
{nateglinide=No,glyburide.metformin=No,discharge_disposition_id=Another_hopital} => {readmitted=Within30days}	0.020173	0.685294	1.373363	699
{chlorpropamide=No,glyburide.metformin=No,discharge_disposition_id=Another_hopital} => {readmitted=Within30days}	0.020173	0.685294	1.373363	699
{tolazamide=No,glyburide.metformin=No,discharge_disposition_id=Another_hopital} => {readmitted=Within30days}	0.020173	0.685294	1.373363	699
{miglitol=No,glyburide.metformin=No,discharge_disposition_id=Another_hopital} => {readmitted=Within30days}	0.020173	0.685294	1.373363	699
{tolbutamide=No,glyburide.metformin=No,discharge_disposition_id=Another_hopital} => {readmitted=Within30days}	0.020173	0.685294	1.373363	699
{glyburide.metformin=No,glipizide.metformin=No,discharge_disposition_id=Another_hopital} => {readmitted=Within30days}	0.020173	0.685294	1.373363	699
{glyburide.metformin=No,metformin.pioglitazone=No,discharge_disposition_id=Another_hopital} => {readmitted=Within30days}	0.020173	0.685294	1.373363	699
{troglitazone=No,glyburide.metformin=No,discharge_disposition_id=Another_hopital} => {readmitted=Within30days}	0.020173	0.685294	1.373363	699
{glyburide.metformin=No,metformin.rosiglitazone=No,discharge_disposition_id=Another_hopital} => {readmitted=Within30days}	0.020173	0.685294	1.373363	699
{acarbose=No,discharge_disposition_id=Another_hopital} => {readmitted=Within30days}	0.020173	0.683284	1.369335	699
{nateglinide=No,acarbose=No,discharge_disposition_id=Another_hopital} => {readmitted=Within30days}	0.020173	0.683284	1.369335	699
{chlorpropamide=No,acarbose=No,discharge_disposition_id=Another_hopital} => {readmitted=Within30days}	0.020173	0.683284	1.369335	699
{acarbose=No,tolazamide=No,discharge_disposition_id=Another_hopital} => {readmitted=Within30days}	0.020173	0.683284	1.369335	699
{acarbose=No,miglitol=No,discharge_disposition_id=Another_hopital} => {readmitted=Within30days}	0.020173	0.683284	1.369335	699
{tolbutamide=No,acarbose=No,discharge_disposition_id=Another_hopital} => {readmitted=Within30days}	0.020173	0.683284	1.369335	699
{acarbose=No,glipizide.metformin=No,discharge_disposition_id=Another_hopital} => {readmitted=Within30days}	0.020173	0.683284	1.369335	699
{acarbose=No,metformin.pioglitazone=No,discharge_disposition_id=Another_hopital} => {readmitted=Within30days}	0.020173	0.683284	1.369335	699
{acarbose=No,troglitazone=No,discharge_disposition_id=Another_hopital} => {readmitted=Within30days}	0.020173	0.683284	1.369335	699

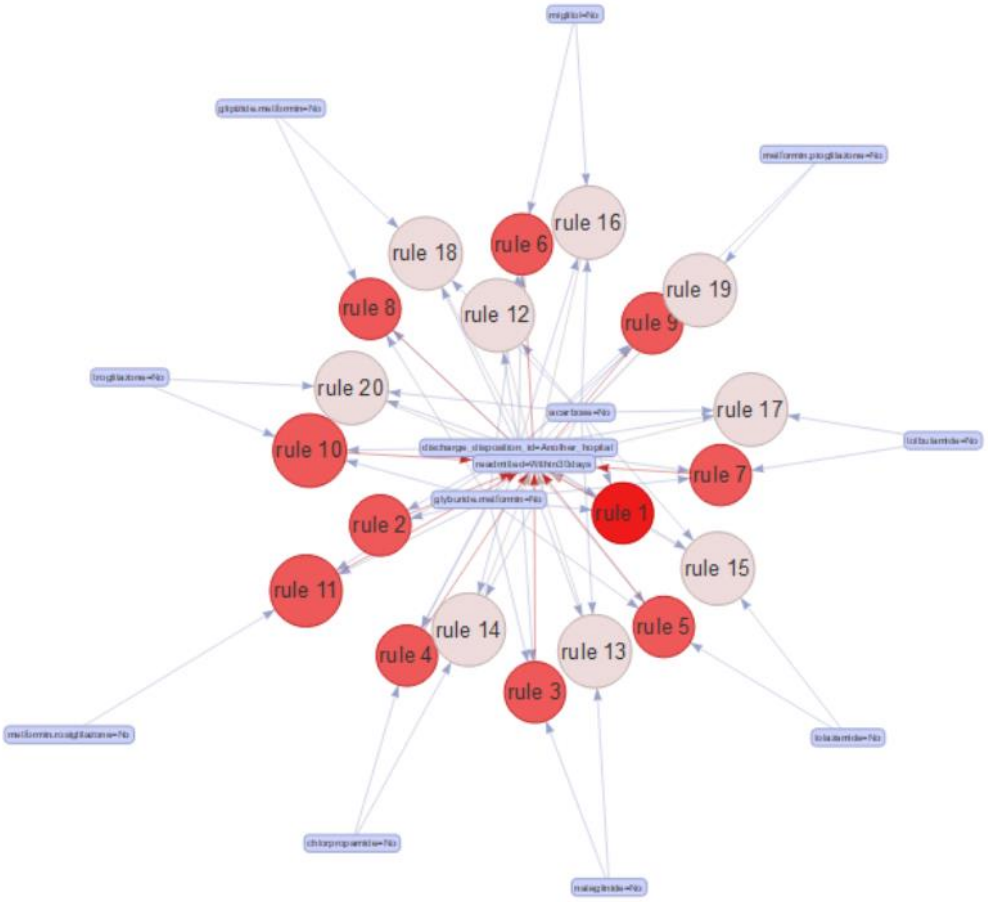
Parallel Coordinates plot for 20 Rules:



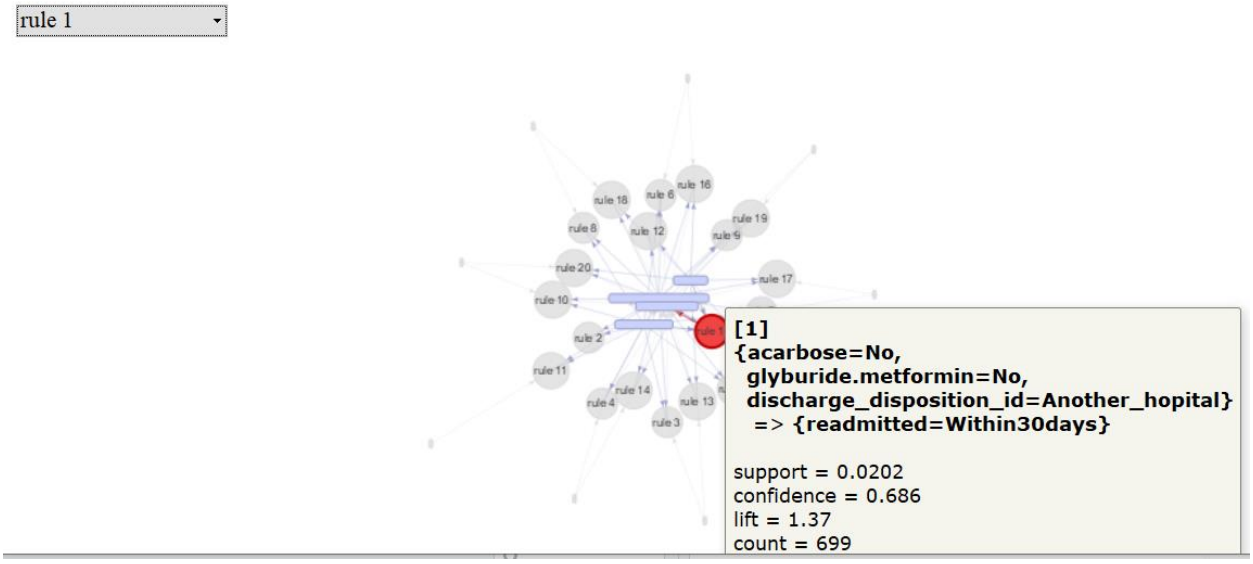
Scatter Plot for 20 Rules:



Rplot for 20 Rules:



Checking support, confidence, lift and count for rule 1. Same way we can see the rest of the rule details in R.



Conclusion:

Below variables are driving the decision most:

- Number of days stayed
- Number of Procedures
- Number of medications
- Number of Diagnosis
- Age
- Discharge Disposition ID

As the constraint was to predict the people who will join the hospital with in 30 days so that a proper action can be taken in order to avoid.

As there is strong urge for the interpretability then we will be going for decision tree as it gives more interpretability when compared to other models.

If Metric is important rather than interpretability then we will be going for the Logistic Regression