**Diabetes Readmission Rates:**

**Analysis of Clinical Database Patient Records**

**Team:Sherlock Holmes**

**Amol Kolhe, Felix Schirra, Yana Romalis, Ekaterina Marchenko**

**San Diego State University**

Professor: Dr. Aaron Elkins

Course: MIS 620 - E-Business and Big Data Infrastructure

Team Name - Sherlock Holmes

Date: 12/11/2018

INDEX

[Executive Summary 3](#_3znysh7)

Data Visualization 10

Data Separation and Control 14

Rebalancing Data 14

Decision Tree 24

Boosting Tree 24

Naive Bayes 25

Linear Discriminant Analysis (LDA) 26

Boosted Trees 28

Bagging 28

Appendix 33

# 

# **Executive Summary**

The objective of our analysis was to predict if a diabetes patient will be readmitted to the hospital after the first treatment and to provide results that will be useful for the hospitals to adjust their treatment.. We had 3 goals to work on : 1) Develop a metric that with 70% accuracy predict if the patient will be readmitted to the hospital, 2) Find the top three statistically significant predictors, 3) Interpret the results of the model and provide insights that would allow hospitals to manage their diabetes patients more efficiently.

The dataset we used represents 10 years (1999–2008) of clinical care at 130 hospitals and integrated delivery networks throughout the United States.

The dataset has 50 columns and around 100k records. The missing data was deleted and after data cleaning we were left with 69k and 26 variable. After creating the dummy variable the dataset had around 69k and 128 variables.

First, it was crucial to understand the context of each predictor so that we could properly understand the data. With the raw dataset, data preparation must be performed to provide insight on the accuracy of modeling predictions. The data processing involved initial screening of raw data, handling predictors having missing values, dummy variables and predictor transformation, and identifying correlated predictors. Without data preparation and validation the data does not provide appropriate informations.

Various data modeling techniques were applied to predict the readmitted patients. The initial hypothesis was to find if the dependent variable have a significant difference over the rest of the predictors. We had three classes in our dependent variable and we decided to build models on the three classes as well as convert it into two classes to see if this increases the performance of the models we have created. The model approaches we used to predict these were logistic regression to understand the most significant predictors with two classes. The models we created with three classes were LDA, Naive Bayes, Neural Networks and Decision Trees and we got 62.34% Accuray with neural networks. Being said that we also tried to merge the classes and build the models for the same. The models build were Decision tree, Decision tree with tuning parameters , Random Forest, Bagging Tree, Boosting trees and Neural Networks. We tried more tree models as we were getting got accuracy rate with the decision trees but we finally got a 63.3% Accuracy rate with neural networks.

Our models failed to reach the desired mark of 70% accuracy. But it did yield important insights as seen in the results below. We failed our first goal but completed our second goal, but still have the ability to achieve our third goal if we continue with the next step of our project. If we gather more data, as we discuss in our recommendations, and narrow our focus to certain important factors, we should be able to maximize the readmitted efficiency predictability. If we are successful, the potential value of our project could obtain great results for the hospitals so that they can be well prepared for the patients in the near future.

1. **Discovery**

**Introduction**

Diabetes is an increasingly pressing problem for the United States. The American Diabetes Association states that 2015 over 30 million of all Americans (9.4%) had diabetes and another 84 million (26%) had prediabetes, a bland form of diabetes that can evolve into diabetes. While these numbers are set to raise further, they already accounted for $237 billion in direct medical costs in 2014 (Strack, B, 2014).

Whereas the analytical detection of diabetes from medical data is very popular and extensively covered, we will study the readmission rate of diabetes patients. For hospitals the primary objective is to treat patients such that no short-term readmission is needed and therefore less costs are caused.   
More successful treatment requires deep insights into the mechanisms and the factors that affect readmissions. For this purpose we will analyze a dataset of diabetes patients who were treated in the hospital with given readmission information. The goal is to build a classification algorithm with sufficient predictive power for practical use.

**Data Description**

The start of our analysis is a dataset created from the Health Facts database which collects electronic clinical records of participating hospitals in the USA. It covers the years from 1999 to 2008 for 130 hospitals. Each observation is a hospital admission of diabetes patients with a hospital stay between 1 and 14 days that include laboratory tests and administered medications throughout the stay.

This extracted dataset was downloaded from kaggle.com/brandao/diabetes. It has 101766 encounter observations of 50 variables covering basic information like gender, age and race, information about the admission and discharge disposition and the time, treatment and diagnoses in the hospital. A readmission feature indicates whether the patient was not readmitted (“No”), readmitted in less than 30 days (“<30”) or in more than 30 days (“>30”) (Strack, B, 2014).

A detailed description of the original dataset with all features is attached in the appendix.

1. **Data Preparation**

**Data Preprocessing**

The original dataset had many columns with missing values and near-zero variance. First, we removed the unique encounter ID because it has no predictive value and then we dropped columns with a huge fraction of missing values: weight (97% missing), payer code (53% missing) and medical specialty (52%).

Secondly, features with near zero variance were removed. The medications (citoglipton, glyburide-metformin, tolazamide, acarbose, troglitazone, miglitol, pioglitazone, repaglinide, chlorpropamide, acetohexamide, metformin-pioglitazone, metformin-rosiglitazone, glipizide-metformin, glimepiride-pioglitazone, examide) were all dominated by one value (“No”) in at least 99% of all observations and therefore removed.

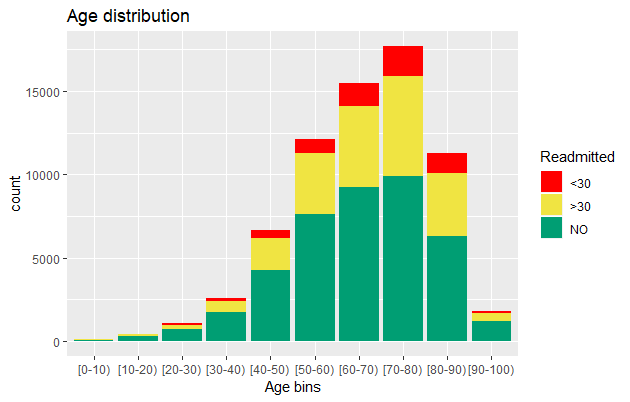
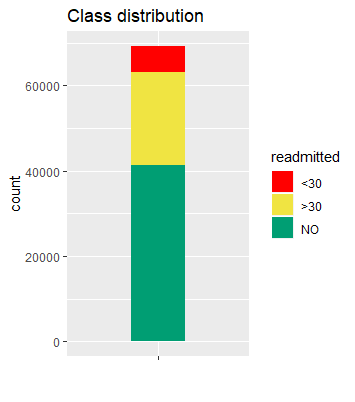
Since the dataset contained multiple encounters for patients, we had to remove all but the first encounters for each patient so that our observations are independent. This step removed around 30,000 observations and accounts for most of our removed observations.

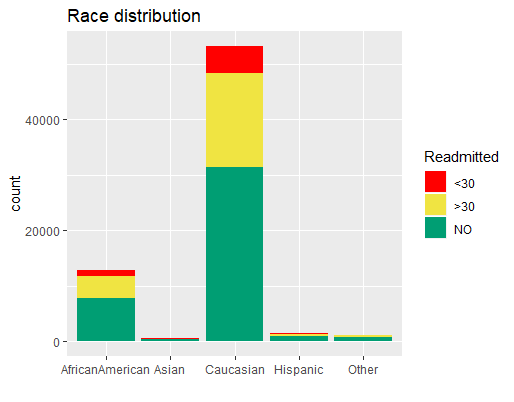
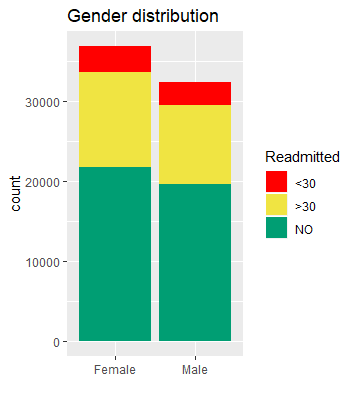
Then we filtered 3 observations with unknown gender, around 2000 observations with unknown race and around 400 observations with either unknown primary diagnosis (diag\_1) or secondary diagnosis (diag\_2). Afterwards we grouped the primary diagnosis with 692 distinct factors and the secondary diagnosis with 722 distinct factors into their 18 ICD-9 (international classification of diseases) categories, i.e. 1-139 is infectious and parasitic diseases, 140-239 are neoplasms and so forth. Given the rather good quality of the primary and secondary diagnosis and for the sake of dimensionality, we dropped the additional secondary diagnosis (diag\_3) which had around 1,500 missing values.

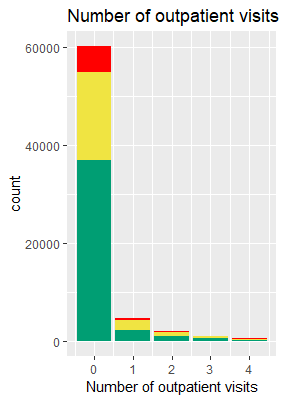
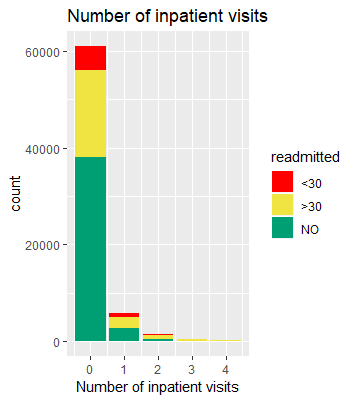
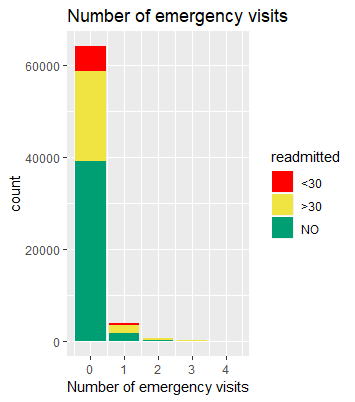
The cleaned dataset without missing values contains 69285 observations with 26 features that expand to 147 categories with dummy variables since most of our features are categorical.

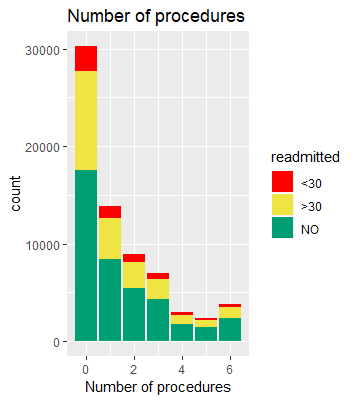
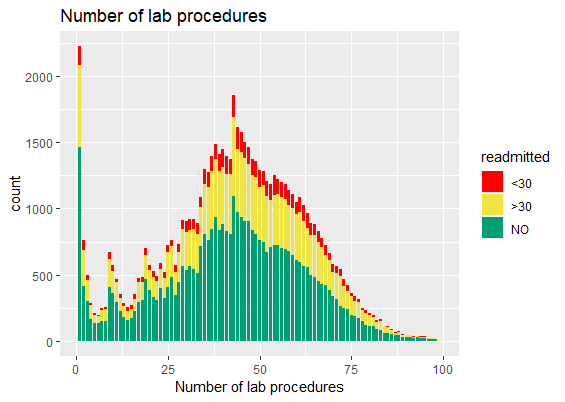
|  |  |  |
| --- | --- | --- |
| Feature | Type | Description |
| Race | Factor | Caucasian, Asian, African American, Hispanic, and other |
| Gender | Factor | male, female, and unknown/invalid |
| Age | Factor | Grouped in 10-year intervals: [0, 10), [10, 20), . . ., [90, 100) |
| Admission type | Factor | Integer identifier corresponding to 9 distinct values, for example, emergency, urgent, elective, newborn, and not available |
| Discharge disposition | Factor | Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available |
| Admission source | Factor | Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital |
| Time in hospital | int | Integer number of days between admission and discharge |
| Number of lab procedures | int | Number of lab tests performed during the encounter |
| Number of procedures | int | Number of procedures (other than lab tests) performed during the encounter |
| Number of medications | int | Number of distinct generic names administered during the encounter |
| Number of outpatient visits | int | Number of outpatient visits of the patient in the year preceding the encounter |
| Number of emergency visits | int | Number of emergency visits of the patient in the year preceding the encounter |
| Number of inpatient visits | int | Number of inpatient visits of the patient in the year preceding the encounter |
| Diagnosis 1 | Factor | The primary diagnosis (coded as first three digits of ICD9); 848 distinct values |
| Diagnosis 2 | Factor | Secondary diagnosis (coded as first three digits of ICD9); 923 distinct values |
| Number of diagnosises | int | Number of diagnoses entered to the system |
| Glucose serum test result | Factor | Indicates the range of the result or if the test was not taken.  Values: “>200,” “>300,” “normal,” and “none” if not measured |
| A1c test result | Factor | 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 of medications | Factor | Indicates if there was a change in diabetic medications (either dosage or generic name). Values: “change” and “no change” |
| Diabetes medications | Factor | Indicates if there was any diabetic medication prescribed.  Values: “yes” and “no” |
| 5 features for medications | Factor | metformin, glimepiride, glipizide, glyburide, insulin; 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 | Factor | 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. |

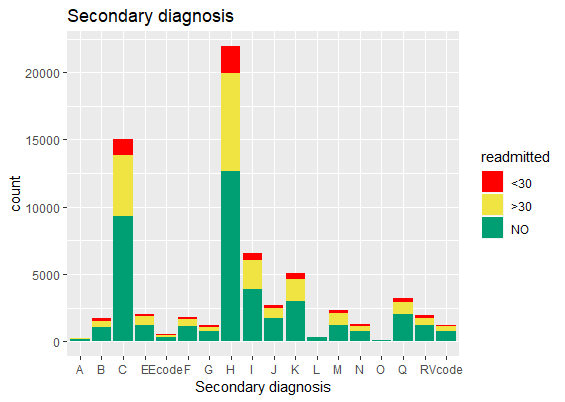
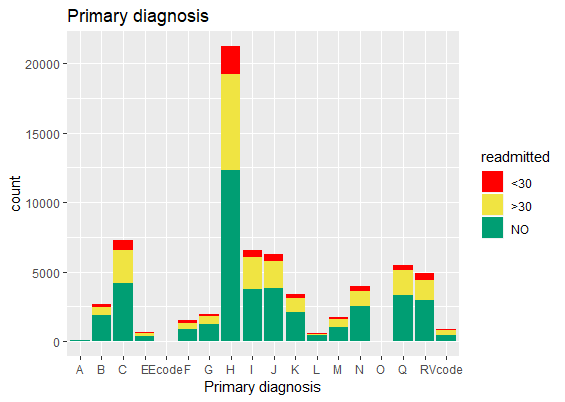
**Data Visualization**

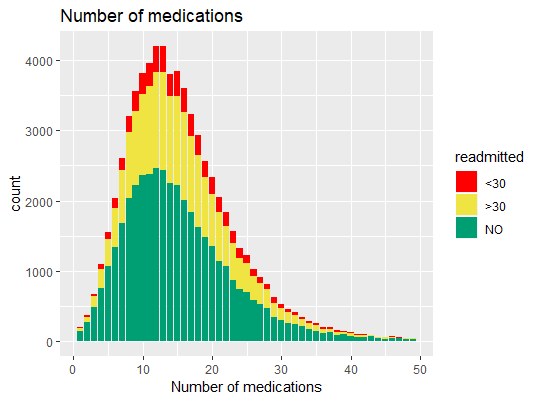
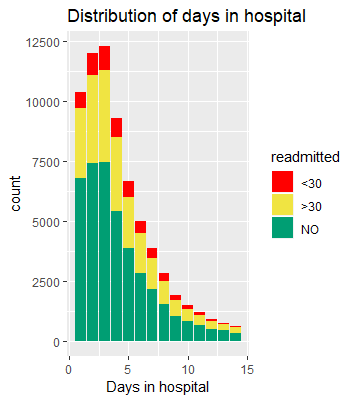
The first step was to visualize the overall class distribution. Our three classes are imbalanced with 9% being readmitted in less than 30 days, 31% readmitted after 30 days and 60% without readmission. For our model planning and model building we need to take this imbalance into account. 

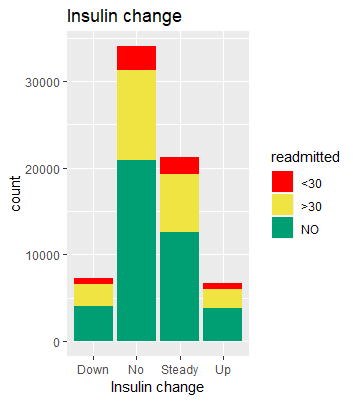
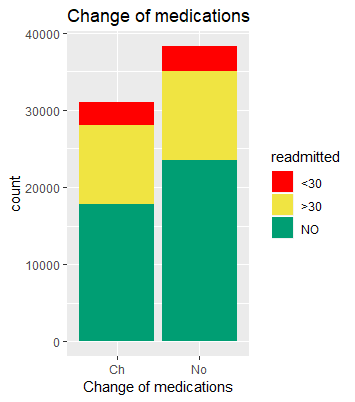
Most of the patients are between 40 and 90 years old with 60 to 70 and 70 to 80 being the most frequent age groups. This makes sense since older people are more likely to have diabetes and be treated in a hospital. Considering the gender, there are slightly more females than males. The majority race is Caucasian with African-American forming a minority while Asian, Hispanic and others are hardly represented. The readmission classes are evenly distributed within each age group, gender and race.

The great majority of patients had no inpatient, outpatient or emergency encounter in the year before their recorded encounter with a small fraction having one or more of those previous encounters.

The number of lab procedures varies around 43 but has another peak at one lab procedure. Half of the patients had no other procedure next to lab procedures with the other half having up to three other procedures and a few having more than three.

The majority of patients was diagnosed within category H, corresponding to diseases of the circulatory system. The second most frequent diagnosis is within category C, i.e. endocrine, nutritional and metabolic diseases, and immunity disorders.

The distribution regarding the time that each patient spent in the hospital is as expected. Most of the patients stay just up to three days and just a few percent had to stay longer than a week. In their time in the hospital they received on average 16 medications.

The medication was changed for 43% of the patients while 57% received only one medication. The insulin feature indicates whether insulin was administered and if so, if the prescription was kept steady or went up or down. 

The data visualization process has shown the imbalance of the readmission class that is present in the overall dataset and in every feature distribution.

**Data Separation and Control**

It is critical to separate the preprocessed data set into training and test subsets. The training set will be used to find the model with best training data performance, and the models will also be verified using the test data set. The best model will have both low bias and low variance. CaTool was used in the code to split the whole data set into 80% training set values and 20% test data values. Train control was also set up before the training of models for resampling and saving predictions for each resample fold. Our analysis used 10-fold cross validation.

**Rebalancing Data**

In many cases it happens that there the number of records we have of each class is very different.

Many models don’t perform well in such cases and they tend to predict the larger classes at the expense of classes with fewer records. In some cases the imbalanced data makes sense: for example the chance to default on a loan is much smaller than to pay it back on time. Therefore in this case it is expected to have less records from “default” class. However ,in other cases the proportion between number of records in each class doesn’t represent the ground truth and might be due ,for instance, to too small training set. In any case, when the difference between the number of records between the largest class and the smallest class is big, it’s recommended to balance the data before building the models. There two main resampling approaches: oversampling and undersampling. In oversampling approach, more records are created from the minor class. One example of oversampling method is SMOTE. The undersampling approach reduces the number of the major class. In our project, we used SMOTE to balance the data set before running the Naive Bayes model. In R SMOTE only balances two classes. Since we had three different classes in the dataset, we applied SMOTE twice to get equal number of records for each one of our three classes.

1. **Model Planning & Model Building**

In our research for model planning and building was used Logistic regression, Decision Tree, Boosting Trees, Naive Bayes, Linear Discriminant Analysis (LDA), Bagging Tree. The performance of usage these methods are shown below.

**Logistic regression**

Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.  
We created the logistic regression model for our model and found that the A1C test result and the number of diagnoses the most significant predictors are related to the class. The following are the results for the same.

**Decision Tree**

Decision trees partition the feature space of the given inputs into disjunct regions. In case of simple decisions with a constant threshold they will form rectangles. Depending on the type of the variable to predict one can grow different types of trees: classification trees for a discrete variable and regression trees for a continuous variable. Since we want to classify with a factor we will focus on growing a classification tree.Every decision tree consists of a root, parent nodes, their child nodes, the final leaves and a decision at each branch. For every branch one has to select a split variable and a threshold value. Every observation satisfying the condition goes left, the others go right. The splitting continues until the node is pure i.e. consists of a single class or a stopping criterion like minimum node size steps in indicating that further splitting has neglectable effect. While large trees tend to overfit when splitting to pure nodes, small trees might underfit the data using too little information. Feature selection is based on a splitting rule determining the "best" pair of feature and threshold i.e. minimizing the node impurity. The advantage of a single decision tree is its easy interpretation with every class assignment being easy to verify. On the other hand decision trees are high variance estimators caused by the hierarchical structure of the tree and therefore the predictions can change drastically by just a small change in inputs

**Boosting Trees**

Adaboost or adaptive boosting is an algorithm that generates weak learners iteratively that learn to classify more of the examples. This model is an extension for the decision tree to improve the performance of the models. Boosted decision trees can be done without the dummy variables as the model takes care of it but we are using the dummy codecs because this function does it internally and it is better to handle it yourself (i.e., less error prone).

**Naive Bayes**

Naive Bayes is a classification algorithm that predicts the class based on the attributes of the object. This algorithm relies on the Bayes Rule for its predictions. It’s called “Naive” because it naively assumes that all the features are conditionally independent from each other. Despite its simplicity, Naive Bayes tends to perform well on certain data. It is used often in text classification problems.   
In R, Naive Bayes accepts both categorical and numeric variables. For numeric variables a normal distribution of a variable within each class is assumed and used to calculate the class conditional probability for that variable. The class conditional probability for categorical variable is calculated based on the number of observations of that variable in the specified class.

We used the Naive Bayes algorithm both with only categorical variables (by discretizing numeric variables into categories) and with mixed types of variables- numeric and categorical.

In both cases SMOTE balance technique was applied to the training set.

**Linear Discriminant Analysis (LDA)**

LDA is a classification method. It works the best in cases when the distribution of predictors within each class is approximately normal.Also in cases when there are more than 2 classes .In our data we had 3 classes and even though we didn’t check the distribution of each predictor in each class,we decided to run LDA and see if we get a decent results.

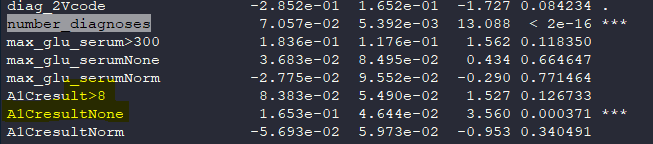
We run LDA twice: first time on the data in its original unbalanced form and second time on SMOTE balanced data.

**Bagging Tree**

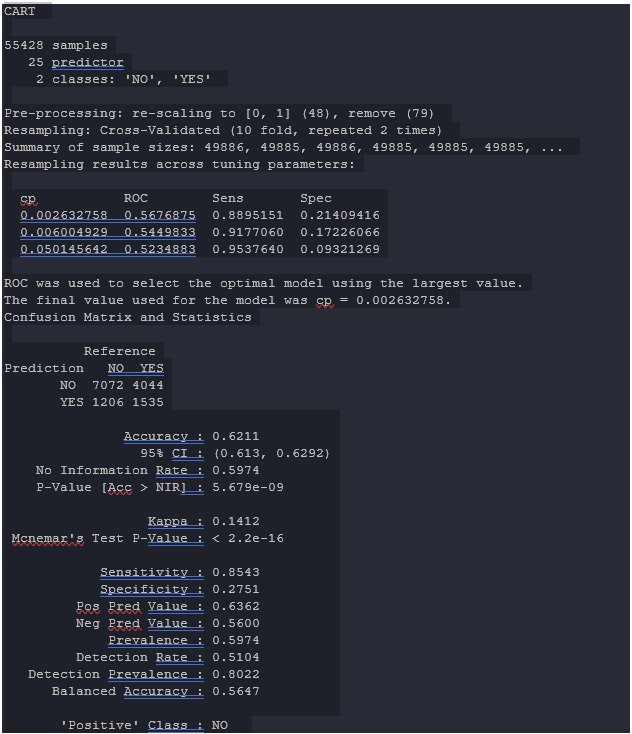
Bagging is used as bootstrapping to create many training sets and simple models are trained on each and combined. It creates many small decision trees to improve the performance.

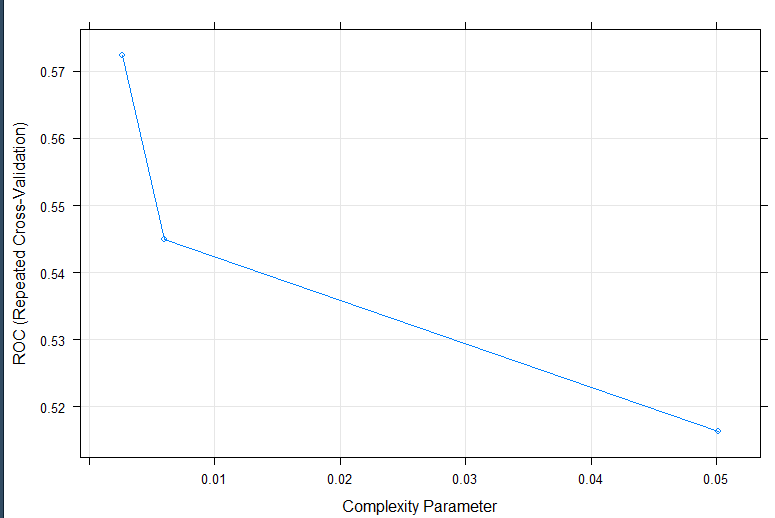
1. **Results and Performance**

**Logistic Regression**

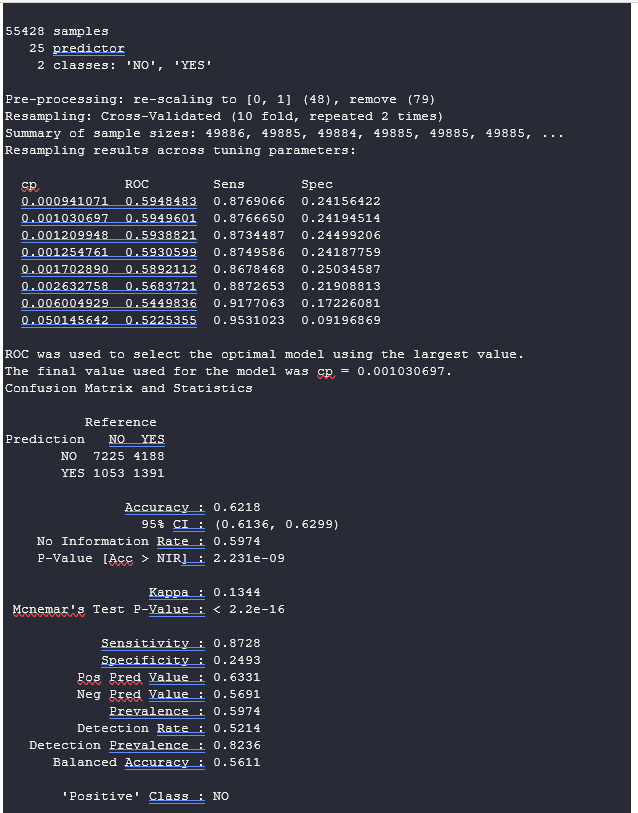


**Decision Trees**

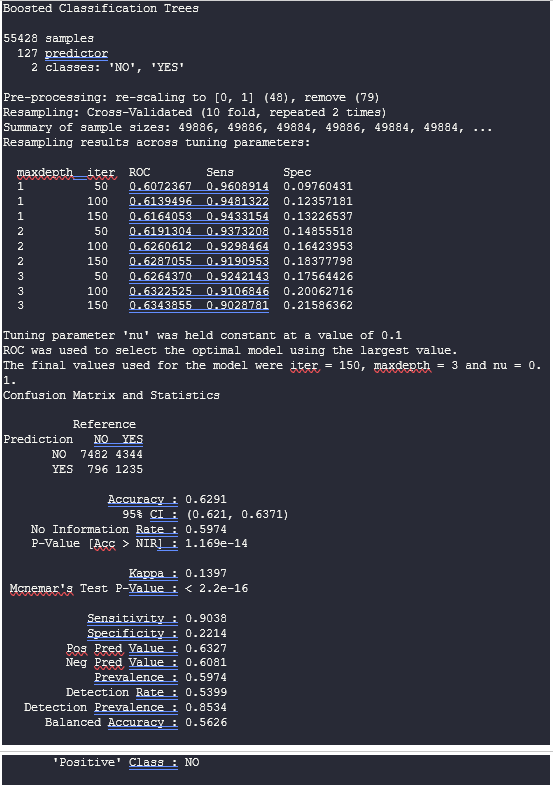
****

****

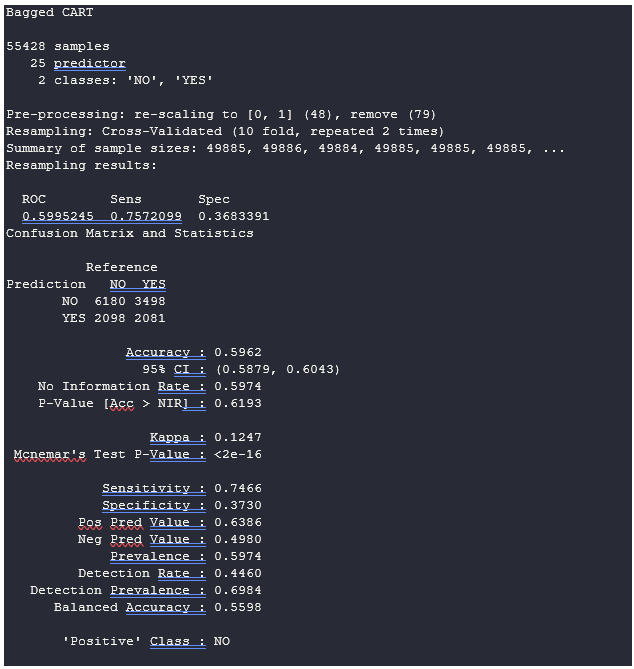
We also tried to tune the decision tree we tried with tune length as 8 and got the following results:



**Boosted Trees**



**Bagging**



**Naïve Bayes**

We conductedNaive Bayes on our SMOTE balanced clean data ,without converting numeric values into factors.The results are the following:

Prediction <30 >30 NO

<30 10 30 46

>30 354 1211 1386

NO 864 3110 6846

Accuracy : 0.5822

Class: <30 Class: >30 Class: NO

Sensitivity 0.0081433 0.27833 0.8270

Specificity 0.9939821 0.81696 0.2877

We applied the LDA model to the testing data and received the following result:

The results for unbalanced data are:

Confusion Matrix:

|  |  |  |  |
| --- | --- | --- | --- |
| Prediction | <30 | >30 | NO |
| <30 | 17 | 21 | 6 |
| >30 | 147 | 441 | 375 |
| NO | 1095 | 3986 | 8215 |

Overall Statistics

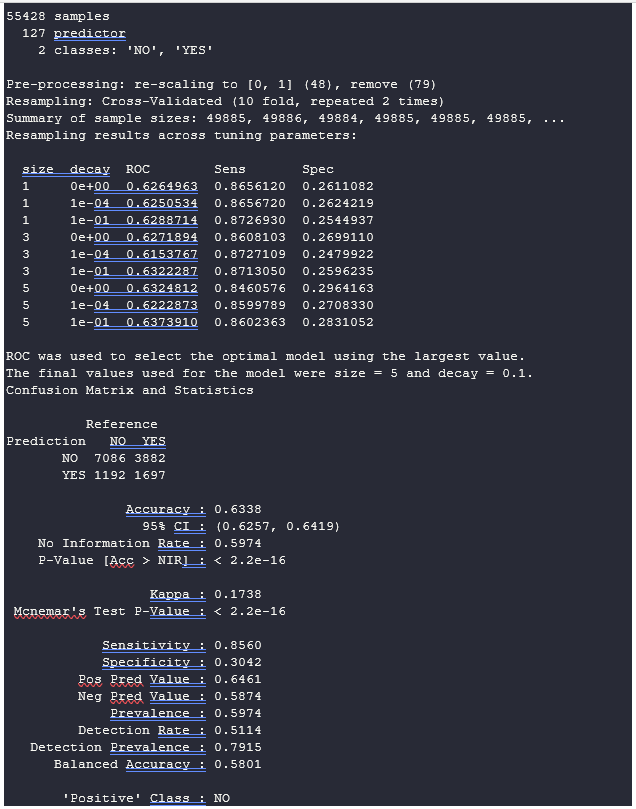
During our research we received that:

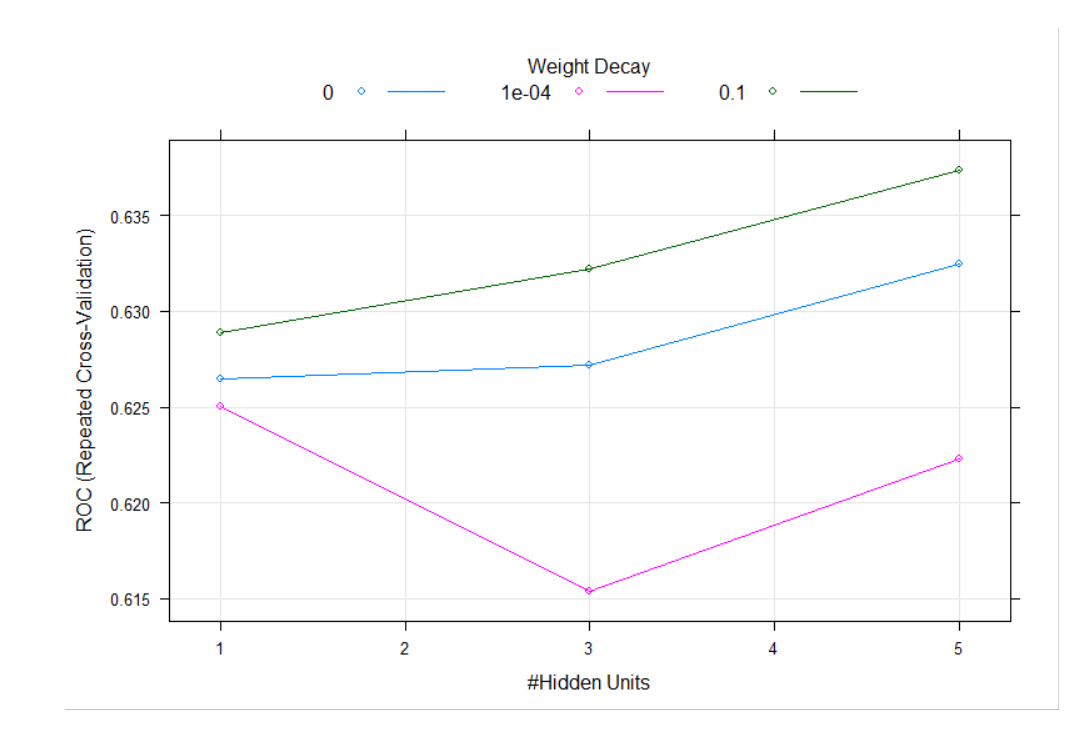
* Accuracy : 0.6064
* 95% CI : (0.5983, 0.6144)
* No Information Rate : 0.601
* P-Value [Acc > NIR] : 0.09567
* Kappa : 0.063
* Mcnemar's Test P-Value : < 2e-16

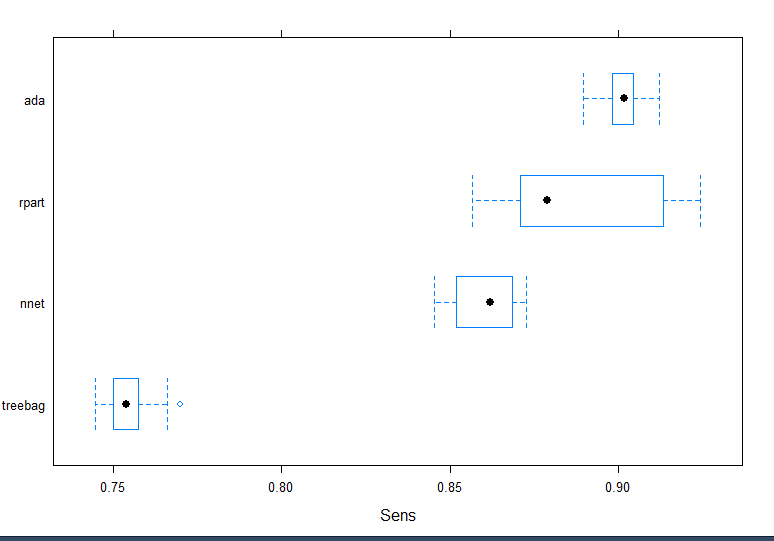
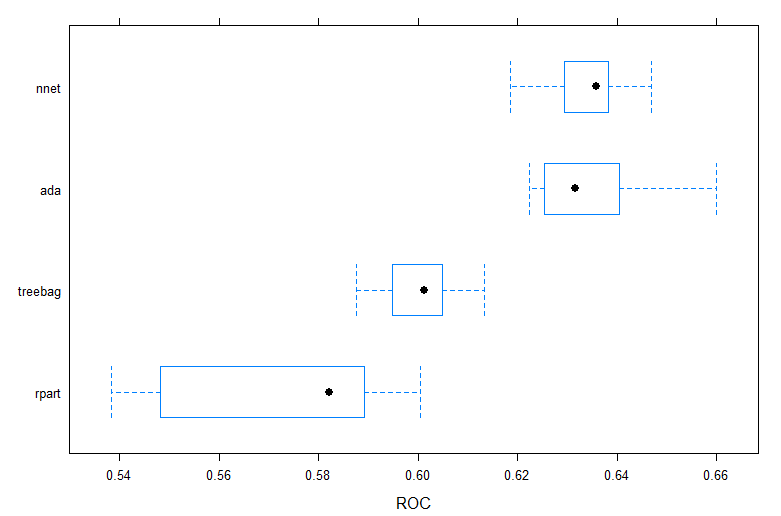
|  |  |  |  |
| --- | --- | --- | --- |
|  | Class: <30 | Class: >30 | Class: NO |
| Sensitivity | 0.013503 | 0.09915 | 0.9557 |
| Specificity | 0.997930 | 0.94703 | 0.1097 |
| Balanced Accuracy | 0.505716 | 0.52309 | 0.5327 |

**Neural Network for Two Classes**

We performed the neural networks model for two variables as well to check if we can increase the performance of the models and got the following results. We managed to improve the performance after combining the predictor to the two classes and got the following results:



****

1. **Discussion and Recommendations**

**Conclusion**

Based upon the result above, we can conclude that 63.33% of readmitted are identified correctly in this model. Also, 63.33% of the readmitted patients who can revisit the hospital are labeled correctly. Overall, we found that the Neural network model provides the best performance in terms of identifying readmitted patients correctly.

The purpose of this project is to prove precise tool for determining readmitted patients to the hospital taking into consideration various medication factors and diagnosis which would have the highest weights in determining the likelihood of a patient getting readmitted to the hospital in the future.

The main recommendations:

* Collect more data weight, BMI etc.
* Advanced preprocessing Convert categoricals to meaningful numerics
* Hire a domain expert from the medical field feature selection

**References**

Strack, B., DeShazo, J. P., Gennings, C., Olmo, J. L., Ventura, S., Cios, K. J., & Clore, J. N. (2014). Impact of HbA1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. *BioMed research international*, *2014*.

An Introduction to Statistical Learning with Applications in R, 1st edition, 2015, Springer-Verlag, New York, ISBN: 978-1-4614-7138-7 (ISL) Available for download at author’s website: http://www-bcf.usc.edu/~gareth/ISL/

Ledlter, J., & Ebrary. (2013). Business analytics and data mining with R. Hoboken, NJ: Wiley.

An Introduction to R Notes on R: A Programming Environment for Data Analysis and Graphics Version 3.5.0 (2018-04-23).

Mercer, J. D. (2014, May 22). *Does Winning Correlate with Franchise Value?* Retrieved from http://thefieldsofgreen.com/2014/05/22/does-winning-correlate-with-franchise-value/

**APPENDIX**

**Additional Visualization**

