Hospitals Readmissions

Introduction

The expense of hospital readmissions account for an enormous portion of hospital inpatient services spending. It was reported that Diabetes is one of the top ten causes of death and the costliest disease in the United States. Hospital readmission is a major concern for diabetic patients since the need for being readmitted indicates that inadequate care was provided to the patient at the of first admission. Therefore, reducing the readmission rates leads to more precise treatment and lesser amount spent in hospital expenses. Hence, the objective of this assignment is to predict the likelihood of a diabetic patient being readmitted.

Data Set Description:

-Source: https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008

-Data is collected from 130 hospitals in the U.S. from 1999-2008

-Contains 101,766 observations and 50 features

-Names of features

Feature Name	Type
Encounter ID	Numeric
Patient Number	Numeric
Race	Nominal
Gender	Nominal
Age	Nominal
Weight	Numeric
Admission type	Numeric

Feature Name	Type
Discharge Disposition	Numeric
Admission Source	Numeric
Time in Hospital	Numeric
Payer code	Nominal
Medical Specialty	Nominal
Number of lab procedures	Numeric
Number of procedures	Numeric

Feature Name	Type
Number of medications	Numeric
Number of outpatient visits	Numeric
Number of emergency visits	Numeric
Number of inpatient visits	Numeric
Diagnosis 1 (1st Diagnose)	Nominal
Diagnosis 2 (2nd Diagnose)	Nominal
Diagnosis 3 (3rd Diagnose)	Nominal

Feature Name	Type
Number of diagnosis	Numeric
Glucose serum test result	Nominal
A1c test result	Nominal
Change of medications	Nominal
Diabetes medications	Nominal
23 features for medications [found in the link below]	Nominal
Readmitted	Nominal

-More detailed description of features can be found this Research Article:
https://www.hindawi.com/journals/bmri/2014/781670/tab1/

-Target variable is Readmitted, indicating the days to inpatient readmission.

- “<30” if the patient was readmitted in less than 30 days (**Class 1**: 11357 observations)
- ”>30” if the patient was readmitted in more than 30 days (**Class 2**: 35545 observation)
- ”No” for no record of readmission (**Class 3:** 54864 observations)

Details of Data Manipulation:

- **Data cleaning :**
  - Removing the duplicates based on *patient_nbr* since it a unique number given to each unique patient. (71518 rows in data set now)
    - Class 1: 6293 observations 2
    - Class 2: 22240 observations 1
    - Class 3: 42985 observations 0
  - Removed uninformative features(21)due to either, a huge amount of missing sample values (>50%), or since some of the features are not relevant to classify the data towards our target (Like encounter ID), or if the feature is completely unbalanced (>95% of data points have the same value for the feature).
  - Hence, the new data set has 71518 rows and 33 columns

- Balancing the Classes using SMOTE:
  - Since the Class 1 is noticably lower then the others. We will be using SMOTE. **SMOTE** stands for *Synthetic Minority Oversampling Technique*. This is a statistical technique for increasing the number of cases in your dataset in a balanced way.
  - After SMOTE:
    - Class 1: 42985 observations 2
    - Class 2: 22240 observations 1
    - Class 3: 42985 observations 0

- Split the Data Set into Train and Test set:
  - Train -> 80%
    - Size: 86568 rows and 33 columns
    - Class 1: 34399 observations
    - Class 2: 17854 observations
    - Class 3: 34315 observations
  - Test -> 20%
    - Size: 21642 rows and 33 columns

- Class 1: 8586 observations
- Class 2: 4386 observations
- Class 3: 8670 observations

Random Forest:

Decision Trees are used to determine a course of action, where each branch of the tree represents a possible decision, occurrence or reaction. Random Forest is an ensemble machine learning algorithm that develops by aggregating multiple decision trees to construct a prediction model.

Below are the basic principles of generating a Random Forest:

1. Randomly select "n" features from total "m" features (where n is much less than m)
2. Calculate the node "d" using the best split point among the "n" features.
   a. The highest Gini index is the best for splitting the nodes.

$$Gini\ (P) = \sum_{i=1}^{n} p_i (1 - p_i\ ) = 1 - \sum_{i=1}^{n} (p_i\ )^2$$

   b.
      i. Where P = (p1, p2, ... pn) and pi is the probability of an object that is being classified to a particular class
      ii.Note: feature with the least Gini index is preferred as root node

3. Split the node into child nodes using the best split point
4. Steps 1 to 3 are repeated until the "K" number of nodes has been reached.
5. Build forest by repeating steps 1-4 for "x" number of times to create "x" number of trees

Afterwards, perform prediction using the Trained Random Forests by

1. Passing the test features though rules of each randomly created trees
2. Calculating the votes for each predicted outcome
3. The highest voted predicted target will be considered as the final prediction for the random forest algorithm.

Model Building: We will use Random Forest and apply run various numbers of tree as parameter to measure its performance. The following classifiers are used:

Number of Trees	**Training Accuracy**	**Testing Accuracy**	**OOB Accuracy**
1	**84.9%**	**59.37%**	**46.81%**
2	**85.7%**	**62.72%**	**52.24%**
4	**94.1%**	**66.43%**	**58.19%**
8	**98.1%**	**69.36%**	**63.53%**
16	**99.7%**	**71.11%**	**67.84%**
32	**99.99%**	**72.58%**	**70.44%**
64	**99.99%**	**73.14%**	**72.06%**
100	**99.99%**	**73.53%**	**72.27%**
200	**99.99%**	**73.72%**	**73.50%**

As seen from the table above, we notice an upward trend in the accuracies as the number of trees increase. From the table above, the best tree was number of trees=200 resulting in Test Accuracy of 73.72% and OOB Accuracy of 73.50%

Confusion Matrix on the Training Set

	Class 3	Class 2	Class 1	Class Error %
Class 3	34314	0	1	0.0029%
Class 2	0	17854	0	0%
Class 1	1	0	34398	0.0029%

Class 1: Readmitted <30 days

Class 2: Readmitted >30 days

Class 3: No record of readmission

From the table above, we get very small class errors because we are testing the data on the training set. This is expected. The highlighted values represent the true prediction of the classes.

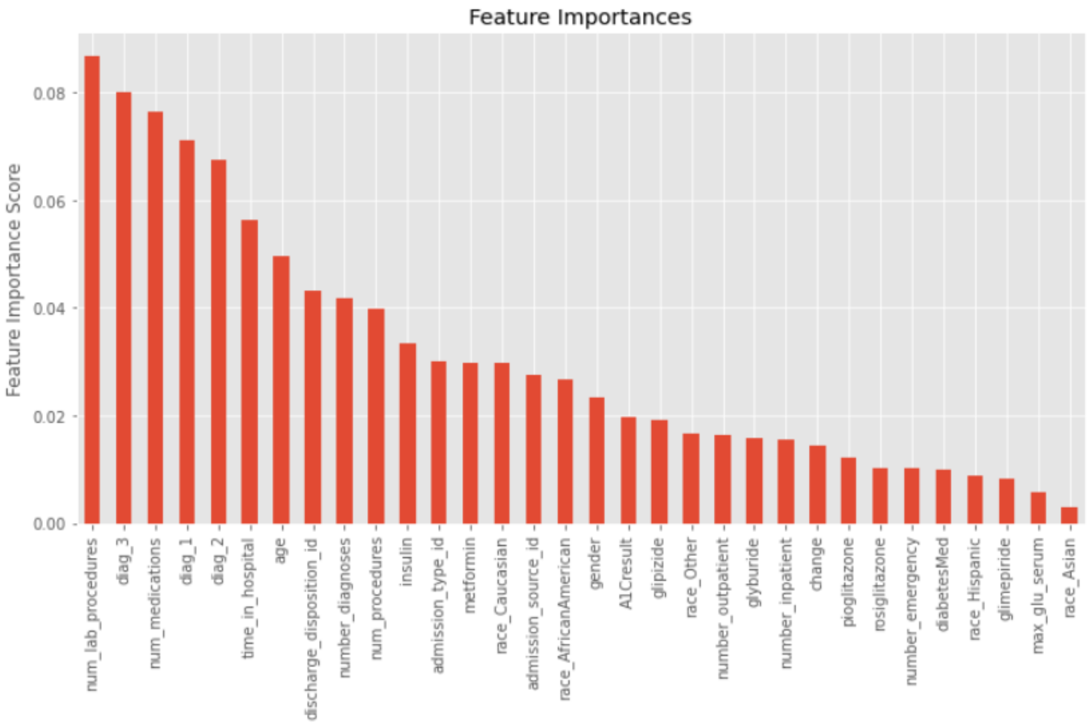
Confusion Matrix on the Test Set

	Class 3	Class 2	Class 1	Class Error %
Class 3	7740	612	318	11%
Class 2	3549	654	183	85%
Class 1	943	140	7503	13%

From the table above, Class 2 has the highest class error perctange. This is due to the fact that earlier when we balanced the classes the Class 2 stood out compared to Class 1 and Class 3. To prevent this, there needs to be better rebalancing of the classes. The error rate for Class 3 is 11% meaning that the model correctly classfied “No record of readmission” 89% of the time. The error rate for Class 2 is 85% meaning that the model correctly classified 15% of the time when patients were actually Readmited more than 30 days. The error rate for Class 1 is 13% meaning that the model correctly classified 87% of the time when patients were actually Readmited less than 30 days.

Importance Features

Using the importance() function, we can view the importance of each variable. This function is the extractor function for variable importance measures as produced by randomForest. The first measure is computed from permuting OOB data: For each tree the error rate for classification is recorded, Then the same is done after permuting each feature. The difference between the two are then averaged over all trees and normalized by the standard deviation of the differences. The second measure is the total decrease in node impurities from splitting on the variable, averaged over all trees. For our classification tree the node impurity is measured by the Gini index.



As seen from the graph number of lab procedures, type of diagnose(diag_3: 3rd diagnose, diag_2: 2nd diagnose, diag_1:1st diagnose), and number of medications are the key factors that affect readmission rate. This is reasonable since the more times a patient gets diagnosed, the more likely he is to be readmitted. Hence, the number of lab procedures is the most significant factor.