Predicting Hospital Readmission for Patients with Diabetes – Sean Flanagan

# Introduction

A number of patients that are admitted to hospital with a diagnosis of diabetes are subsequently readmitted at a later date. Having patients readmitted to hospital is problematic for the patient and also adds cost to the health care system. Being able to correctly predict which patients are likely to be readmitted and identifying patterns associated with readmitted patients could serve to reduce patient suffering and save money for the health care system.

After the data are pre-processed and feature selection is performed, the data will be split into a training set and a test set, and a number of machine learning classification algorithms (Naïve Bayes, decision tree, random forest, and support vector machines) will be used to train the data. Additionally, an ensemble algorithm will be applied to the data to determine if that can improve the ability to predict hospital readmission. The performance of the various algorithms will be evaluated on the test data set. In each case, the efficiency and effectiveness will be measured.

In addition to the classification analysis, association rule learning will be applied using the Apriori algorithm to discern patterns in data that associate various attributes with hospital readmission.

### General Comments

This is a very complex dataset with many attributes and, for a number of those attributes, there are many levels. For example, the primary, secondary, and tertiary diagnoses can each contain up to almost 1000 different codes. While there are many ways to approach the analysis of a dataset set such as this, I chose primarily to reduce complexity by recategorizing the levels of attributes that had many categories. Obviously some information can be lost in doing so, and follow up work on these data could be done to tease out more nuanced information from the various attributes.

The other advantage of recategorizing some of the attribute levels is that it can be done such that the data will be more balanced. For some of the attributes I created a new level called “Uncommon” and reassigned observations to this level if the proportion of observations in the dataset for this level of the attribute was below a particular threshold.

# Literature Review

A research article by Strack et al. entitled “Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records”[[1]](#endnote-1) accessed data from a US national data warehouse called Health Facts that covered the period of 1999 to 2008 for 130 sites providing clinical care. The data used for their analysis was extracted from this larger database, focusing on 55 attributes of interest and was made available to the UCI Machine Learning Repository; this is the same dataset used for my analysis and also used by the two other articles described in the Literature Review section. The primary variable of interest in the Strack study was the measurement of hemoglobin A1c (HbA1c). Multivariate logistic regression was used to fit the relationship between HbA1c measurement and readmission while controlling for other covariates in the data. They found that, for those patients with a primary diagnosis of diabetes mellitus, the decision to measure HbA1c (regardless of the test result) was a significant predictor of hospital readmission.

A paper called “Identifying Diabetic Patients with High Risk of Readmission”[[2]](#endnote-2) conducted by Bhuvan et al. describes an analysis of the same dataset created by the Strack study. In the analysis, the data scientists performed classification using five classifiers (Naïve Bayes, Bayesian networks, random forests, Adaboost, and neural networks), feature analysis (using ablation and associative rule mining), and a cost sensitive analysis. They also conducted an evaluation of the performance of the various classifiers used. They concluded that the random forest algorithm performed optimally at predicting patients at high risk of readmission. The analysis identified specific features that were associated with increased likelihood of hospital readmission; in particular, they found that the number of inpatient visits, into whose care the patient was discharged, how the patient was admitted, and the number of lab tests performed were important in identifying hospital readmission. Finally, a cost analysis predicted a potential savings of just over $US250 million over the 98,000 patients by applying a “special diagnosis” to patients identified as high risk of being readmitted.

An article by Usman Raza entitled “How to use machine learning to predict hospital readmissions?”[[3]](#endnote-3) describes another data science analysis of the dataset created by Strack et al. This article describes in good detail the steps taking in the analysis including code snippets. The analysis treated the prediction variable for hospital readmission as a binary class (“readmitted” or “not readmitted”) rather than as a multivariate prediction variable (“readmitted under 30 days”, “readmitted over 30 days”, “not readmitted”). It was found that the random forest and decision tree models outperformed the logistic regression model. The three most important predictors for hospital readmission were found to be time spent in hospital, patient age, and into whose care the patient was discharged. The authors point out some limitations of the data such as the lack of information on patient access to care.

## Dataset

The dataset used for this analysis was obtained from the UCI Machine Learning Repository (<https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008>). The description of the dataset as given by the UCI Machine Learning Repository is as follows:

*The dataset represents 10 years (1999-2008) of clinical care at 130 US hospitals and integrated delivery networks. It includes over 50 features representing patient and hospital outcomes. Information was extracted from the database for encounters that satisfied the following criteria.*

1. *It is an inpatient encounter (a hospital admission).*
2. *It is a diabetic encounter, that is, one during which any kind of diabetes was entered to the system as a diagnosis.*
3. *The length of stay was at least 1 day and at most 14 days.*
4. *Laboratory tests were performed during the encounter.*
5. *Medications were administered during the encounter.*

*The data contains such attributes as patient number, race, gender, age, admission type, time in hospital, medical specialty of admitting physician, number of lab test performed, HbA1c test result, diagnosis, number of medication, diabetic medications, number of outpatient, inpatient, and emergency visits in the year before the hospitalization, etc.*

The dataset, in fact, includes exactly 50 attributes and 101,766 observations.

### Attributes

The target variable will be a nominal attribute called “readmitted” that takes values “<30” for patients readmitted to the hospital within 30 days, “>30” for those readmitted after 30 days, and “NO” if the patient was not readmitted. The attributes from the dataset used as part of the analysis are summarized in the following table:

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature name** | **Description and values** | **Modification** | **Variable Name** |
| Encounter ID | Unique identifier of an encounter |  | encounter\_id |
| Patient number | Unique identifier of a patient |  | patient\_nbr |
| Race | Values: Caucasian, Asian, African American, Hispanic, and other |  | race |
| Gender | Values: male, female, and unknown/invalid |  | gender |
| Age | Grouped in 10-year intervals: 0, 10), 10, 20), …, 90, 100) | Used midpoint of range | age |
| Admission type | Integer identifier corresponding to 9 distinct values, for example, emergency, urgent, elective, newborn, and not available | Combined “Not Available”, “NULL”, “Not Mapped” to “Unknown” | admission\_type\_id |
| Discharge disposition | Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available | Combined “Unknown/invalid”, “NULL”, “Not Mapped” to “Unknown”. | discharge\_disposition\_id |
| Admission source | Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital | Combined “Not Available”, “NULL”, “Not Mapped” to “Unknown”. Combined all types relating to childbirth as “Babies”. | admission\_source\_id |
| Time in hospital | Integer number of days between admission and discharge |  | time\_in\_hospital |
| Payer code | Integer identifier corresponding to 23 distinct values, for example, Blue Cross/Blue Shield, Medicare, and self-pay |  | payer\_code |
| Medical specialty | Integer identifier of specialty of admitting physician, corresponding to 84 distinct values (cardiology, internal medicine, family/general practice, surgeon, etc.) | Combined specialties for which there were 500 or fewer admissions to “Uncommon”. | medical\_specialty |
| Number of lab procedures | Number of lab tests performed during the encounter |  | num\_lab\_procedures |
| Number of procedures | Number of procedures (other than lab tests) performed during the encounter |  | num\_procedures |
| Number of medications | Number of distinct generic names administered during the encounter |  | num\_medications |
| Number of outpatient visits | Number of outpatient visits of the patient in the year preceding the encounter |  | number\_outpatient |
| Number of emergency visits | Number of emergency visits of the patient in the year preceding the encounter |  | number\_emergency |
| Number of inpatient visits | Number of inpatient visits of the patient in the year preceding the encounter |  | number\_inpatient |
| Diagnosis 1 | The primary diagnosis (first 3 digits of ICD9); 848 distinct values | Recategorized into 9 broad groups of diagnosis type. | diag\_1 |
| Diagnosis 2 | Secondary diagnosis (first 3 digits of ICD9); 923 distinct values | Recategorized into 9 broad groups of diagnosis type. | diag\_2 |
| Diagnosis 3 | Additional secondary diagnosis (first 3 digits of ICD9); 954 distinct values | Recategorized into 9 broad groups of diagnosis type. | diag\_3 |
| Number of diagnoses | Number of diagnoses entered to the system |  | number\_diagnoses |
| Glucose serum test result | Indicates the range of the result or if the test was not taken. Values: “>200,” “>300,” “normal,” and “none” if not measured | Relabeled as “Missing”, “Normal”, “High”, “Very High”. | max\_glu\_serum |
| A1c test result | 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. | Relabeled as “Missing”, “Normal”, “High”, “Very High”. | A1Cresult |
| 24 features for medications | A list of 24 diabetes medications. The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: “up” if the dosage was increased, “down” if decreased, “steady” if no change, and “no” if not prescribed | Dropped all medication attributes except for insulin due to sparseness | Medication name is used as variable name |
| Change of medications | Indicates if change in diabetic medications. Values: “change” and “no change” |  | Change |
| Diabetes medications | Indicates if there was any diabetic medication prescribed. Values: “yes” and “no” |  | diabetesMed |
| Readmitted | 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. |  | Readmitted |

### Descriptive Statistics

#### Univariate Analysis

For the categorical variables, frequency tables were created and the distribution of the levels examined. Re-categorization of levels was done in cases where too many levels were present and / or some levels were very sparse. For example, there are 73 different medical specialties in the original dataset, most of which interacted with a small number of patients. By combining those medical specialties into a new level called “Uncommon” it is possible to determine the predictive value of patients being seen by doctors with an uncommon specialty on hospital readmission.

Numerical variables were summarized in terms of minimum, maximum, median, mean, Q1, and Q3.

# Approach

The following block diagram shows the steps of the approach to this analysis of the hospital readmission data.

The analysis was done in R. The code can be found at <https://github.com/SeanPatrickFlanagan/CKME136-Project/blob/master/Analysis%20of%20Readmission%20Data.Rmd>

## Step 1: Download Data and Import the CSV File into R

I downloaded the hospital readmission data from <https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008>. Included in the zipped folder is a CSV file IDs\_mapping that contains information on three of the attributes in the main dataset; specifically, the meaning of the various ID codes for admission type, discharge disposition, and admission source.

The hospital readmission data was read into R as a CSV file and stored in a dataframe called readmission\_data\_original. The data was examined in its raw form.

## Step 2: Descriptive Statistics Before Data Cleaning

I created summary tables of categorical and numerical variables.

### Overview of Results of Basic Descriptive Stats

Initially, before any pre-processing of the data, there were 101 766 records in the dataset. There were 37 categorical attributes and 13 numerical attributes.

#### Categorical Variables

* Race is not evenly represented in the data. Caucasian is by far the largest category
* Gender is split quite evenly between male and female
* Distribution of ages appeared somewhat bell-shaped with the category [70-80) being the most frequent
* Data were entered as “?” for 97% of weight observations
* The payer code most often unknown, but Medicaid was the next most common code indicated
* There were 73 distinct medical specialties and almost 1000 diagnoses in the dataset making analysis difficult for these two attributes
* Most often, the glucose serum and hemoglobin A1C tests were not performed
* With the exception of insulin, most of the 24 diabetes medications were not administered often
* There was a reasonably even split between patients who had a change in their medication during their encounter at the hospital and those who did not
* Most of the patients did receive some form of diabetes medication
* A little more than a tenth of patients were readmitted within 30 days, a third after 30 days, and more than half were not readmitted at all

#### Numerical Variables

* Attributes admission\_type\_id, discharge\_disposition\_id, and admission\_source\_id were numeric variables (integer) in the original dataset although these are really categorical attributes
* Summary of mean, standard deviation, and normality of numerical variables

|  |  |  |  |
| --- | --- | --- | --- |
| **Attribute** | **Mean** | **Standard Deviation** | **Appearance of Normality** |
| Time\_in\_hospital | 4.4 days | 3 days | Skewed right |
| Num\_lab\_procedures | 43 | 19.7 | Somewhat normal. Large number of zeroes |
| Num\_procedures | 1.3 | 1.7 | Not normal |
| Num\_medications | 16 | 8.1 | Appears normal (right skew) |
| Number\_outpatient | 0.4 | 1.3 | Not normal |
| Number\_emergency | 0.2 | 0.9 | Not normal |
| Number\_inpatient | 0.6 | 1.3 | Not normal |
| Number\_diagnoses | 7.4 | 1.9 | Not normal |

Most of the values for num\_procedures, number\_outpatient, number\_emergency, and number\_inpatient are zero. The data for these attributes are clearly not normal and attempts at transforming the data to approach normality were unsuccessful. These attributes will be converted to categorical variables with values “None” and “At least one”.

## Step 3: Data Cleaning and Basic Feature Selection

Some records and attributes from the data file were removed.

The “weight” attribute from the original dataset was not used here because 98% of the values were missing.

Patient records were removed from the dataset based on the following criteria:

* All three diagnoses were missing, or
* “Gender” is missing or invalid (since there were only three records in this situation), or
* The patient died or was transferred to hospice care (since they have no chance of being readmitted), or
* The encounter was not the first for the patient (to avoid biasing the results with patients with multiple encounters)

There are almost 1000 distinct diagnoses codes in the diag1, diag2, and diag3 attributes. Analysis based on this many distinct values would make interpretation of the results difficult. Therefore, these diagnoses were recategorized into nine categories as was done in the original analysis of the data (Strack et al.).

For consistency and to allow for a more meaningful analysis, I renamed values “Not Mapped”, “?”, “Unknown/Invalid”, “NULL”, “Not Available” to “Unknown”.

Any admission source that refers to babies, infants, or childbirth were combined to “Babies” as these represent a very small fraction of the records.

Doctors that received patients came from 73 medical specialties. Most of these medical specialties represented a small number of records. Medical specialties that saw 500 or fewer patients were reclassified to a new category called “Uncommon”. This reduced the number of medical specialty attributes from 73 down to 15.

For a given patient, most of the medication attributes are “No” leading to sparse data among 24 attributes. The only exception was insulin which was more balanced across the various possible levels. A new attribute was created called num\_meds that represents the total number of medications each patient was administered during the encounter. Insulin was kept as an attribute, but the other medication attributes will be dropped.

## Step 4: More Detailed Descriptive Statistics

The univariate analysis was repeated after the basic pre-processing of the data was done.

#### Categorical Variables

* Race is not evenly represented in the data. Caucasian is by far the largest category
* Gender is split quite evenly between male and female
* The type of admission was most often characterized as an emergency
* The majority of patients were released to their own home
* Most admissions were done through emergency room visits
* The payer code most often unknown, but Medicaid was the next most common code indicated
* Medical specialty was most often unknown
* For both the primary and secondary diagnoses the circulatory system was indicated most often with “other” as the second most common category of diagnosis. For the tertiary diagnosis, “other” was just slightly more often specified than circulatory
* Most often, the glucose serum and hemoglobin A1C tests were not performed
* Insulin was administered in about half of the patient encounters
* There was a reasonably even split between patients who had a change in their medication during their encounter at the hospital and those who did not
* Most of the patients did receive some form of diabetes medication
* A little less than a tenth of patients were readmitted within 30 days, a bit less than a third after 30 days, and about 60% were not readmitted at all

#### Numerical Variables

* Summary of mean, standard deviation, and normality of numerical variables

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute** | **Mean** | **St. Dev.** | **Normality** | **Skewness** | **Kurtosis** |
| Age | 65.4 yr | 16 yr | Somewhat normal. Skewed left. | -0.63 | 3.4 |
| Time\_in\_hospital | 4.3 days | 2.9 days | Skewed right | 1.2 | 4 |
| Num\_lab\_procedures | 42.9 | 19.9 | Somewhat normal. Many zeroes. | -0.22 | 2.7 |
| Num\_procedures | 1.4 | 1.8 | Not normal | - | - |
| Num\_medications | 15.7 | 8.3 | Appears normal (right skew) | 1.4 | 6.9 |
| Number\_outpatient | 0.3 | 1.1 | Not normal | - | - |
| Number\_emergency | 0.1 | 0.5 | Not normal | - | - |
| Number\_inpatient | 0.2 | 0.6 | Not normal | - | - |
| Number\_diagnoses | 7.2 | 2 | Not normal | - | - |

An attempt was made to transform the attribute number\_diagnoses using Tukey’s Ladder of Powers. Even after transformation, this attribute was clearly not normal. Consequently, it was converted into a categorical variable with two levels, “low” and “high”, to balance the distribution.

#### Outliers

Attributes with the largest number of outliers (more than 1.5 x IQR) are num\_procedures (3845, 5.49%), number\_outpatient (9122, 13.03%), number\_emergency (5101, 7.29%), and number\_inpatient (8195, 11.71%). In each case, there are a large number of records with a value of zero so that the IQR is skewed down. There is no reason to believe these outliers are the result of bad data and removing them may reduce the predictive power of the machine learning models.

#### Correlation

A correlation matrix was constructed for pairs of numeric variables and no significant correlations were found. Comparing correlations among categorical variable using a G-test, however, showed that most of the pairs of attributes were highly correlation. The exceptions were the following pairs of attributes:

|  |  |  |
| --- | --- | --- |
| **Attribute 1** | **Attribute 2** | **p-value** |
| gender | admission\_source\_id | 0.3954 |
| gender | number\_inpatient | 0.6899 |
| gender | number\_diagnoses | 0.6262 |
| gender | max\_glu\_serum | 0.6584 |
| gender | insulin | 0.1408 |
| admission\_source\_id | diabetesMed | 0.8533 |
| num\_procedures | number\_outpatient | 0.879 |
| num\_procedures | number\_inpatient | 0.4662 |
| num\_procedures | insulin | 0.1381 |
| num\_procedures | change | 0.0507 |
| number\_inpatient | max\_glu\_serum | 0.3055 |
| number\_inpatient | change | 0.1731 |

From this list, the following pairs of independent triplets were found:

gender, number\_inpatient, max\_glu\_serum

num\_procedures, number\_inpatient, change

## Step 5: Exploratory Analysis

The numerical variables were scaled use the following formula in order to ensure that the support vector machine algorithm, which uses distance measures to calculate hyperplanes, would not be skewed by different scales for the numerical attributes.

## Step 6: Association Rules

I used the Apriori algorithm to examine association rules to induce new knowledge about the relationships among the hospital data attributes with respect to the class variable “readmitted”. I experimented with various levels of support and confidence to generate a reasonable number of rules that had a lift of more than 1 (a lift of 1 would indicate that the rule provides no more confidence in predicting the class variable than simply guessing the class variable).

***Rules for patients readmitted in less than 30 days***

lhs rhs support confidence lift count

[1] {discharge\_disposition\_id=Rehab,

admission\_source\_id=ER,

medical\_specialty=Unknown,

number\_diagnoses=Low} => {readmitted=<30} 0.001 0.5 6 82

[2] {admission\_type\_id=Emergency,

discharge\_disposition\_id=Rehab,

medical\_specialty=Unknown,

number\_diagnoses=Low} => {readmitted=<30} 0.001 0.5 6 83

We would be six times more likely (lift = 6) to correctly predict a patient would be readmitted in less than 30 days if the patient was discharged to Rehab, the medical specialty was unknown, the number of diagnoses was low, and the patient was either admitted through the ER or the admission type was considered an emergency.

***Rules for patients readmitted in more than 30 days***

lhs rhs support confidence lift count

[1] {discharge\_disposition\_id=Home,

payer\_code=Unknown,

number\_inpatient=At least one,

diabetesMed=Yes} => {readmitted=>30} 0.02 0.5 2 1157

[2] {race=Caucasian,

discharge\_disposition\_id=Home,

payer\_code=Unknown,

number\_inpatient=At least one} => {readmitted=>30} 0.02 0.5 2 1111

[3] {discharge\_disposition\_id=Home,

payer\_code=Unknown,

number\_inpatient=At least one} => {readmitted=>30} 0.02 0.5 2 1458

[4] {discharge\_disposition\_id=Home,

num\_procedures=None,

number\_inpatient=At least one} => {readmitted=>30} 0.02 0.5 2 1089

[5] {discharge\_disposition\_id=Home,

payer\_code=Unknown,

number\_inpatient=At least one,

max\_glu\_serum=None} => {readmitted=>30} 0.02 0.5 2 1326

[6] {discharge\_disposition\_id=Home,

payer\_code=Unknown,

number\_outpatient=None,

number\_inpatient=At least one} => {readmitted=>30} 0.02 0.5 2 1151

The most common attributes associated with patients being readmitted to hospital after 30 days are

* Discharging patients to their home
* Having an unknown payer code
* Having had at least one inpatient visit in the last year

Other attributes that seem to be associated with hospital readmission in more than 30 days are:

* A diabetes medication was prescribed
* The patient was Caucasian
* No additional (non-lab) procedures were performed
* Maximum glucose serum test was not administered
* Number of outpatient visits in last year was none

***Rules for patients not readmitted to hospital***

lhs rhs support confidence lift count

[1] {age=(25,35],

admission\_type\_id=Elective,

diag\_1=Other,

change=No} => {readmitted=NO} 0.002 0.9 2 160

[2] {age=(25,35],

admission\_type\_id=Elective,

admission\_source\_id=Physician Referral,

diabetesMed=No} => {readmitted=NO} 0.002 0.9 2 165

[3] {time\_in\_hospital=[1,4.25],

admission\_type\_id=Uncommon,

payer\_code=Uncommon,

diabetesMed=No} => {readmitted=NO} 0.002 0.9 2 147

[4] {age=(25,35],

gender=Female,

admission\_type\_id=Elective,

diabetesMed=No} => {readmitted=NO} 0.002 0.9 2 145

[5] {admission\_type\_id=Uncommon,

payer\_code=Uncommon,

number\_emergency=None,

diabetesMed=No} => {readmitted=NO} 0.002 0.9 2 151

[6] {age=(25,35],

admission\_type\_id=Elective,

diag\_1=Other,

diag\_2=Other} => {readmitted=NO} 0.002 0.9 2 141

We are twice as likely (lift = 2) to correctly predict a patient will not be readmitted to hospital if they are in the age range of 25 to 35, had an elective admission, and either

* Their primary diagnosis was “Other” and either
  + There was no change in the medication during the encounter, or
  + The secondary diagnosis was also “Other”
* No diabetes medications administered and either
  + They were female, or
  + They were admitted via a physician referral

Other patients who are equally unlikely to be readmitted are those who were not given diabetes medication, both the admission type and payer code were “Uncommon”, and either

* They spent fewer than five days in hospital
* They had no emergency visits to the hospital in the last year

***Rules for patients readmitted to hospital over any time period***

I was interested to know what rules could be derived if the class variable was binary rather than having three classes; specifically, readmitted=YES and readmitted=NO. (To do this I replaced any occurrences of   
“<30” and “>30” with “YES” in the dataset). Running the apriori algorithm on this dataset yields the following rules with support at least 0.04 and confidence at least 0.6 and lift more than 1 (note that no rules were generated for the readmitted = “NO” class with lift more than 1):

lhs rhs support confidence lift count

[1] {num\_diab\_meds=[0,3],

payer\_code=Unknown,

number\_inpatient=At least one} => {readmitted=YES} 0.04 0.6 2 2944

[2] {payer\_code=Unknown,

number\_inpatient=At least one} => {readmitted=YES} 0.04 0.6 2 2965

[3] {admission\_source\_id=ER,

number\_inpatient=At least one} => {readmitted=YES} 0.04 0.6 2 2828

[4] {num\_diab\_meds=[0,3],

discharge\_disposition\_id=Home,

number\_inpatient=At least one} => {readmitted=YES} 0.04 0.6 2 2912

[5] {discharge\_disposition\_id=Home,

number\_inpatient=At least one} => {readmitted=YES} 0.04 0.6 2 2943

[6] {num\_diab\_meds=[0,3],

race=Caucasian,

number\_inpatient=At least one,

diabetesMed=Yes} => {readmitted=YES} 0.04 0.6 2 3094

In all four rules the patient had at least one inpatient visit in the last year. Other attributes combined with the inpatient attribute that show an association for hospital readmission are

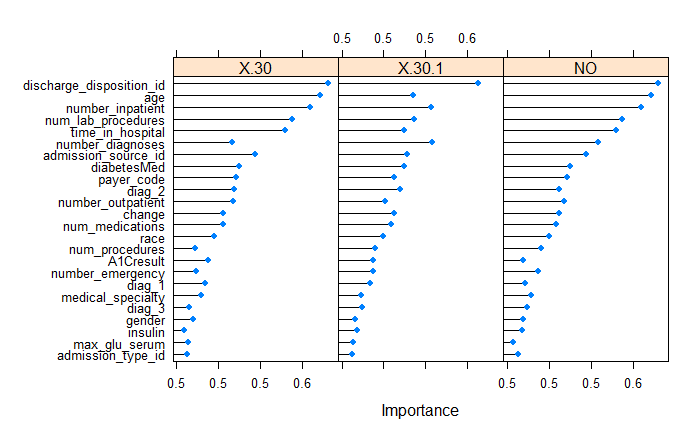
* Unknown payer code
* Admitted to hospital through the emergency room
* Patient discharged to home
* Caucasians who were prescribed at least one diabetes medication
* Three or fewer diabetes medications were prescribed

## Step 7: Dimensionality Reduction

I used the R functions FSelector and learning vector quantization (LVQ) to identify attributes that are more important predictors of the target variable “readmitted”.

The following plot was shows the importance of features across classes as determined by the LVQ algorithm:

Figure - LVQ Importance of Attributes Across Classes



The following attributes appeared in the top 50% of both the FSelector algorithm (based on information gain) and the LVQ algorithm results (Note that there was not a large difference between the highest LVQ importance score (0.572) and the lowest (0.503) out of 20 attributes reported):

number\_inpatient, discharge\_disposition\_id, age, num\_lab\_procedures, number\_diagnoses, time\_in\_hospital, diag\_2, diabetesMed, payer\_code, number\_outpatient

These attributes were used to construct a reduced attribute dataset that will be trained on various models.

## Step 8: Experimental Design

I split the data into training (80%) and test set (20%). The models were developed using cross-validation on the training dataset. Once the final model(s) were selected, the test dataset was used to determine the performance.

Treatment for imbalance in the target variable, *readmitted*, was also considered.

## Step 9: Modeling and Evaluation

I applied various classification techniques (Naïve Bayes, decision tree, random forest, and support vector machines) to the data. Models were trained on 80% of the training set using 10-fold cross-validation (the other 20% was withheld as a test dataset until the final model(s) were determined). Confusion matrices were computed for each fold and summed to give a total confusion matrix within the cross-validation phase and the micro-accuracy and micro-F1 scores was calculated. Each model was then validated using the validation set and a new confusion matrix generated along with the micro-accuracy and micro-F1 scores. Run time for each model was also measured.

For the support vector machine algorithm, I experimented with different values for the parameters cost (5, 100), gamma (0.05, 0.2), and kernel (linear, radial, polynomial). The accuracy and F1 scores were almost identical for all values, with the only difference being that the linear kernel had the fastest runtime. Future work could involve using a grid search on the cost, gamma, and kernel parameters to find the optimum values.

I trained each of the models beginning with the full set of attributes and then repeated for the reduced feature selected data (number\_inpatient, discharge\_disposition\_id, number\_diagnoses, num\_lab\_procedures, time\_in\_hospital, diag\_2, number\_outpatient, age, payer\_code, diabetesMed, and readmitted). The following table shows a summary of the results.

Table - Model results using multi-class variable (readmitted under 30, over 30, or not at all)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | All Attributes | | | Reduced Attribute Set | | |
| Model | Accuracy | F1 Score | Run Time (s) | Accuracy | F1 Score | Run Time (s) |
| Naïve Bayes | 0.725 | 0.588 | 28.98 | 0.727 | 0.591 | 14.16 |
| Decision Tree | 0.716 | 0.574 | 15.02 | 0.73 | 0.595 | 25.09 |
| Random Forest | 0.735 | 0.603 | 121.11 | 0.735 | 0.602 | 44.14 |
| Support Vector Machine | Aborted after many hours | | | 0.733 | 0.599 | 3163.05 |

There is minimal difference in the accuracy or F1 scores for any of the models, whether using the full set of attributes or the reduced set. The major difference is in the runtime. Since there is no real difference in the accuracy and F1 scores for the random forest algorithm, the best model for these data would be random forest on the reduced attribute set due to the shorter runtime.

There is an imbalance in the class variable. Initially, I combined the “<30” and “>30” classes (since both represent readmission to hospital) to a single “YES” class to address this imbalance.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table - Distribution of multi-class attribute "readmitted"   |  |  |  |  | | --- | --- | --- | --- | | **Value** | **<30** | **>30** | **NO** | | **Frequency** | 6285 | 22226 | 41476 | | **Proportion** | 0.09 | 0.32 | 0.59 | |  | Table - Distribution of binary class attribute "readmitted"   |  |  |  | | --- | --- | --- | | **Value** | **YES** | **NO** | | **Frequency** | 28511 | 41476 | | **Proportion** | 0.4 | 0.6 | |

The models were run again on the reduced attribute set with a binary classification (readmitted = “YES” or readmitted = “NO”). Given that the support vector machine algorithm didn’t produce better performance and the run time was significantly longer, I decided to omit it from this analysis. The results are summarized in the following table.

Table - Model results using binary class variable (readmitted or not)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | All Attributes | | | Reduced Attribute Set | | |
| Model | Accuracy | F1 Score | Run Time (s) | Accuracy | F1 Score | Run Time (s) |
| Naïve Bayes | 0.615 | 0.698 | 25.67 | 0.607 | 0.704 | 13.53 |
| Decision Tree | 0.605 | 0.695 | 13.16 | 0.617 | 0.726 | 13.72 |
| Random Forest | 0.628 | 0.725 | 83.53 | 0.619 | 0.732 | 33.81 |

The accuracy is not as good for the binary class model than they were for the multiclass model, however the F1 scores have improved. If a hospital was primarily concerned about reducing the number of false negatives for predicting readmission over any time frame (i.e. those patients who were predicted to not be readmitted but actually were), then using decision trees or random forests for the binary class dataset would be most appropriate. Since, in the United States, hospital funding is related to readmission under 30 days or not, this metric may not be of the most importance to the hospital administration. The differences between all of the models are minimal.

#### Optimizing the number of folds for cross-validation

Given that the models are performing similarly well, and that the runtimes for Naïve Bayes and decision trees are the lowest, I used these two models to optimize the number of folds for cross-validation. Regardless of the number of folds used in the cross-validation for both the Naïve Bayes and decision tree algorithms, the accuracy and F1 scores remained virtually the same.

#### Undersampling the Majority Classes on Multiclass Dataset

In order to balance the class variable for the multiclass dataset, I did a random undersampling of the majority classes for the training set. I included all observations from the minority class (readmitted <30).

The various models were run on this balanced data set using both the full attribute set and the reduced attribute set.

Table - Model results using balanced dataset (undersampling majority classes)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | All Attributes | | | Reduced Attribute Set | | |
| Model | Accuracy | F1 Score | Run Time (s) | Accuracy | F1 Score | Run Time (s) |
| Naïve Bayes | 0.621 | 0.431 | 7.39 | 0.618 | 0.426 | 3.73 |
| Decision Tree | 0.601 | 0.401 | 2.09 | 0.595 | 0.392 | 2.63 |
| Random Forest | 0.619 | 0.429 | 31.6 | 0.605 | 0.407 | 15.12 |

Both the accuracy and the F1 scores are worse when the data were undersampled from the majority classes using random samples.

#### Undersampling the Majority Class on Binary Class Dataset

I once again converted the dataset to a binary class dataset, however this time I considered readmission in less than 30 days being the “YES” class, and more than 30 days or not at all being the “NO” class. The data were balanced by random undersampling the majority class for the training set. I included all observations from the minority class (readmitted <30). The various models were run on this balanced data set using both the full attribute set and the reduced attribute set.

Table - Model results using balanced dataset with binary class (undersampling majority class)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | All Attributes | | | Reduced Attribute Set | | |
| Model | Accuracy | F1 Score | Run Time (s) | Accuracy | F1 Score | Run Time (s) |
| Naïve Bayes | 0.603 | 0.596 | 4.93 | 0.608 | 0.595 | 2.49 |
| Decision Tree | 0.581 | 0.572 | 1.16 | 0.584 | 0.567 | 0.86 |
| Random Forest | 0.608 | 0.603 | 12.86 | 0.585 | 0.574 | 7.76 |

The accuracy for these models was lower in each case compared to the models run with the multiclass data. The F1 scores were marginally improved but still very low.

#### Synthetic Oversampling of Minority Class on Binary Class Dataset

There are limited options for oversampling minority classes for multiclass datasets. While it is possible to use a one-vs-all approach with, say, SMOTE, “random oversampling does not help the classification and suffers from overfitting”[[4]](#endnote-4) in the context of a multiclass problem. Instead, I decided to convert the dataset to a binary class model with classes “<30” and “NOT” for the readmitted attribute and then run SMOTE on this dataset (which was highly imbalanced 10% to 90%). The results of running the models on this synthetically oversampled dataset are shown in Table 7.

Table - Model results using balanced dataset (synthetic oversampling)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | All Attributes | | | Reduced Attribute Set | | |
| Model | Accuracy | F1 Score | Run Time (s) | Accuracy | F1 Score | Run Time (s) |
| Naïve Bayes | 0.753 | 0.751 | 46.37 | 0.699 | 0.696 | 22.7 |
| Decision Tree | 0.902 | 0.895 | 20.88 | 0.905 | 0.897 | 21.95 |
| Random Forest | 0.919 | 0.917 | 116.37 | 0.903 | 0.898 | 53.7 |

The accuracy and F1 scores are dramatically improved for decision trees and random forests (for both the full attribute dataset and the reduced attribute dataset) using the binary class data balanced using synthetic oversampling (SMOTE). I suspect that the oversampling of the minority class is introducing significant bias given the large increase in performance. If runtime is not an issue, then the best model would be random forest on the full attribute set. Otherwise, the decision tree model on the reduced attribute set would work well.

#### Ensemble Techniques

The random forest algorithm that was used as one of the models is a bagging or bootstrapping type of ensemble algorithm and yielded results that were only marginally better than Naïve Bayes or decision trees for the most part.

I used the Adaboost package in R to create boost ensemble models. Adaboost uses decision trees as weak classifiers. Since Adaboost is designed for binary class data, the algorithm was run on the binary class dataset and the balanced binary class datasets (both the dataset created by undersampling the majority class and the dataset created by synthetic oversampling of the minority class). I experimented with different values for the number of iterations for the boosting algorithm and there was minimal improvement in accuracy and F1 score after about 7 iterations.

Table - Model results using Adaboost ensemble algorithm

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **All Attributes** | | | **Reduced Attribute Set** | | |
| **Binary Dataset Used** | **Accuracy** | **F1 Score** | **Run Time (s)** | **Accuracy** | **F1 Score** | **Run Time (s)** |
| Unbalanced | 0.901 | 0.0316 | 36.99 | 0.899 | 0.039 | 16.89 |
| Balanced (undersampling) | 0.566 | 0.551 | 4.11 | 0.567 | 0.446 | 2.14 |
| Balanced (SMOTE) | 0.918 | 0.915 | 74.33 | 0.915 | 0.911 | 34.01 |

The Adaboost ensemble algorithm performs very poorly on all datasets except for those data that were balanced using SMOTE.

#### Run Best Models on Test Data

The following table shows the results of running the best performing models on the test data (20% of original dataset that was partitioned off).

Table - Best models run on test data

|  |  |  |  |
| --- | --- | --- | --- |
| **Model Tested** | **Accuracy** | **F1 Score** | **Run Time (s)** |
| Random forest on multiclass, unbalanced, reduced attribute set | 0.735 | 0.603 | 12.51 |
| Random forest on binary class, SMOTE balanced, full attribute set | 0.817 | 0.187 | 39.67 |
| Decision tree on binary class, SMOTE balanced, reduced attribute set | 0.889 | 0.091 | 8.6 |
| Adaboost on binary class, SMOTE balanced, reduced attribute set | 0.833 | 0.151 | 96.46 |

The F1 scores for the models that relied on synthetic oversampling of the minority class are abysmal. The best model was the random forest on the multiclass, unbalanced dataset with all attributes included.

# Results

Based on the data mining using association rules, the main conditions associated with patients being readmitted in less than 30 days are

* Being discharged to a rehabilitation facility
* Being treated by a physician whose medical specialty is listed as unknown
* Having a low number of overall medical diagnoses

In addition to these conditions, if the patient was either admitted through the ER or the admission type was considered an emergency they are more likely to be readmitted in under 30 days.

Overall, the following attributes appear to be most significant in predicting patient readmission:

|  |  |
| --- | --- |
| * number of inpatient visits | * to where discharged |
| * number of lab procedures | * time in hospital |
| * number outpatient visits | * age |
| * diabetes medication prescribed | * secondary diagnosis |
| * number of diagnoses | * payer code |

However, there is a high degree of correlation among the patient variables so it is difficult to determine if these variables are in fact ultimately responsible for patient readmission.

Models that used this reduced set of attributes performed almost as well as models that used all of the attributes. For example, the following table shows the accuracy and F1 scores for the random forest algorithm on the both the full attribute set and the reduced attribute set. The only real difference is the runtime.

Table - Comparison of random forest performance on full and reduced attribute datasets

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Full Attribute Set** | | | **Reduced Attribute Set** | | |
| **Model** | **Accuracy** | **F1** | **Time (s)** | **Accuracy** | **F1** | **Time (s)** |
| Random Forest | 0.735 | 0.603 | 121.11 | 0.735 | 0.602 | 44.14 |

The data were imbalanced with respect to the class variable “readmitted” as shown in Table 2 (reproduced here for convenience):

Table - Distribution of multi-class attribute "readmitted"

|  |  |  |  |
| --- | --- | --- | --- |
| **Value** | **<30** | **>30** | **NO** |
| **Frequency** | 6285 | 22226 | 41476 |
| **Proportion** | 0.09 | 0.32 | 0.59 |

Balancing of the data was done in the following ways:

1. Create a “readmitted or NOT” dataset by combining “<30” and “>30” classes (this created a 40/60 split)
2. Undersampling the majority classes (“>30” and “NO”) to have same number as minority class (“<30”)
3. Create a “<30 versus NOT” binary class dataset by combining “>30” and “NOT” classes, then undersampling the majority class
4. Use SMOTE to synthetically oversample the minority class in the binary class dataset described in 3 above.

Attempts at balancing the classes either had minimal effect or, in some cases, dramatically reduced the performance of the models when predictions were made on the test data.

Ultimately, the model chosen as the best was the random forest algorithm run on the multiclass, reduced attribute dataset. This model had an accuracy of 0.735 and an F1 score of 0.602 when run on the test data.

## Step 10: Conclusions

The association rules discovered in this analysis may point to areas where a change in procedure during the encounter or interventions after the encounter has ended could be beneficial in preventing readmission. If successful, this would not only reduce health care costs, but also increase patient well-being.

The random forest model developed here could be used to predict whether a new patient entering the hospital is likely to be readmitted in under 30 days, over 30 days, or not at all. Of course, in order have the required information to make that prediction the patient’s encounter would have necessarily ended. This means that it would be too late to perform any interventions that could have possibly prevented readmission to hospital. However, follow-up protocols could be developed for those patients deemed to be at higher risk of readmission to hospital. This would be advantageous from a cost perspective if the protocols were less expensive than having the patient be readmitted to the hospital.

This analysis shows that patient readmission to hospital can be predicted with a modest amount of accuracy. Hospitals could benefit from examining procedures or aspects of the patient encounter that relate to certain patient attributes (e.g. the medical specialty of the treating physician).

1. Strack B., DeShazo J., Gennings C., Olmo J. L., Ventura S., Cios K. J., Clore J. N. *Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records*. Retrieved October 6, 2018, from https://www.hindawi.com/journals/bmri/2014/781670/. [↑](#endnote-ref-1)
2. Bhuvan M. S., Ankit K., Adil Z., Vinith K. *Identifying Diabetic Patients with High Risk of Readmission.* Retrieved October 6, 2018, from https://arxiv.org/pdf/1602.04257.pdf. [↑](#endnote-ref-2)
3. Raza, U. *How to use machine learning to predict hospital readmissions?* Retrieved October 6, 2018, from https://medium.com/@uraza/how-to-use-machine-learning-to-predict-hospital-readmissions-part-1-bd137cbdba07. [↑](#endnote-ref-3)
4. Wang S., Yao X. *Multi-Class Imbalance Problems: Analysis and Potential Solutions*. Retrieved November 21, 2018, from https://www.cs.bham.ac.uk/~xin/papers/Shuo\_SMCb-V3.pdf. [↑](#endnote-ref-4)