

Patient Re-admittance Analysis

AUTHOR
MARC NAVIA

Executive summary

Introduction

Our analytics team was given four datasets from a large hospital system on various data about patients, hospitals, and geography. The hospital system is having a problem with patients being readmitted back into their system after being discharged and is looking for insights on what factors could be causing this. Throughout this process, the analytics team will conduct Explanatory data analysis to get a deeper insight into the data. From those insights, multiple supervised machine learning algorithms such as logistic regression and decision trees will be used to predict the probability of a patient being readmitted.

Summary of Problem

The hospital is looking for a way to find the probability of a patient of being readmitted back into their system after being discharged. With the passage of the Affordable Care Act in 2012 came with the Hospital Readmission Reduction Program (HRRP). This program financially penalized hospital with a higher than expected risk-standardized 30-day readmission rates. This penalty will continue to grow with every passing year so reducing patient readmittance is a priority.

Data and Analysis

Before we begin any analysis, we should understand the data that we are working with. We were given four datasets that have information about the demographics of the patient, the condition of the patient in the hospital and information about the initial hospital stay. There also a dataset that has information about the geographical location of the patient.

From this data, we will have to do data preprocessing to ensure the quality and integrity of the data. With the data finally clean and joined together, analysis can be done. Some of the modeling techniques that will be used are logistic regressions and decision trees.

To gain further insight we will use visual analysis on the data using Tableau. This can help visualize our data and create a story for the insights that we found when exploring the datasets.

Findings/Recommendations

Limitations and Next Steps

Some of the limitations that could affect our modeling is the amount of data that we have received. While 160,000 record may seem like a lot to some, having a larger data set can allow the model to learn a lot better. There could be also important variables that have strong explanatory value for our target variable that has been left out of the dataset. The next steps for improving our analysis is to go out and gather more data. Through a limitation that can occur from this is that if we have got too much data we might not have enough computing power to make a proper analysis without investing more capital in improving our assets.

Task:

What level of the data should the analysis be conducted?

To begin with, descriptive analytics should be done to get a better understanding of the data. This will allow for a better modeling process, helping to gain insights into the data. This will give valuable insight into the variables we are working with and help find relevant data to use.

Diagnostic analytics will be very important in this process. This will help see why patients are being re-admitted back into the hospitals. Once we have a better understanding of why the problem is happening, predicting becomes a much easier process.

Lastly, predictive analytics will be the cream of the crop for the hospital. Predicting which patients are a high probability of being readmitted will help the hospital greatly. The hospital will be able to efficiently allocate resource to riskier patients.

What are the most prevalent tools (software) being applied by Data Scientist today?

Some prevalent technology that data scientist is using today is open-source software. Some of these include programming language such as Python and R. For the R language many of the packages are support by academia with some of the most cutting-edge models being offered. Python like R has a huge community supporting multiple packages with very relevant tools that use many state of the art models. Packages such as Tensorflow, Keras, and Pytorch give data scientist tools to implement state of the art deep learning models. All of these tools are free as they are open-source and supported by the data science community.

Some other tools include those that focus specifically on visualization such as Power BI and Tableau.

Other tools being used are those such as Hadoop and Apache Spark which help manage big data by rapidly analyzing, transforming, and queries data

What techniques are emerging in the field of Data Science?





Some techniques that are emerging in the field of Data Science are those in supervises machine learning models such as Support Vector Machines and Random Forest. These techniques have become more popular as GPU has become exponentially better. While these machine learning algorithms are not new, they have only become popular as they can finally be computed quickly on a very large scale. As mention before with the GPU the increasing computing power has led to a revolution in deep learning and artificial neural networks. These datasets only gain extreme value with large datasets compared to other supervised and unsupervised machine learning techniques since they are much more computationally heavy to perform.

What is the business value of the hospital's challenge being resolved? What type of monetary savings might be revealed?

Significant business value can be gained if the hospital challenge is resolved. Having the ability to accurately predict a probability that a patient who enters the hospital and then is discharged and readmitted provides significant financial value. Having a reduce patient readmittance rate can decrease the amount that the hospital system is fined by the national government for violating the re-admittance benchmark. Having a much lower re-admittance rate can also make their hospital system more attractive to consumers. People who have to get riskier surgery might make the choice based on that factor, bringing in a significantly more capital to the hospital. While trying to build the model, additional insights into the data can bring significant financial value to the hospital. We could possibly find certain doctors or nurses are associated with higher patient re-admittance or certain hospital who are performing poorly. This can help management make a better decision on allocating resources, leading to lower expenses and a stronger bottom line.

Appendix A. Screenshots of Data Prep

Raw file upload into SAS Studio

-  readmit_condition.csv
-  readmit_demographic.csv
-  readmit_geomap.tab
-  readmit_hospital.txt

Raw file conversion to SAS datasets

```
%web_drop_table(DATAPREP.readmit_conditions);
```

```
FILENAME REFFILE '/home/manavia0/Data Prep/readmit_condition.csv';
```

```
PROC IMPORT DATAFILE=REFFILE  
  DBMS=CSV  
  OUT=DATAPREP.readmit_conditions;  
  GETNAMES=YES;  
  DATAROW=2;  
  GUESSINGROWS=50;  
RUN;
```

```
PROC CONTENTS DATA=DATAPREP.readmit_conditions; RUN;
```

```
%web_open_table(DATAPREP.readmit_conditions);
```

```
%web_drop_table(DATAPREP.readmit_geomap);
```

```
FILENAME REFFILE '/home/manavia0/Data Prep/readmit_geomap.tab';
```

```
PROC IMPORT DATAFILE=REFFILE  
  DBMS=TAB  
  OUT=DATAPREP.readmit_geomap;  
  GETNAMES=YES;  
RUN;
```

```
PROC CONTENTS DATA=DATAPREP.readmit_geomap; RUN;
```

```
%web_open_table(DATAPREP.readmit_geomap);
```

```
%web_drop_table(DATAPREP.readmit_hospital);
```

```
FILENAME REFFILE '/home/manavia0/Data Prep/readmit_hospital.txt';
```

```
PROC IMPORT DATAFILE=REFFILE  
  DBMS=CSV  
  OUT=DATAPREP.readmit_hospital;  
  GETNAMES=YES;  
  DATAROW=2;  
  GUESSINGROWS=50;
```

```
RUN;
```

```
PROC CONTENTS DATA=DATAPREP.readmit_hospital; RUN;
```

```
%web_open_table(DATAPREP.readmit_hospital);
```

```
%web_drop_table(DATAPREP.readmit_demographic);
```





```
FILENAME REFFILE '/home/manavia0/Data Prep/readmit_demographic.csv';
```

```
PROC IMPORT DATAFILE=REFFILE  
  DBMS=CSV  
  OUT=DATAPREP.readmit_demographic;  
  GETNAMES=YES;  
  DATAROW=2;  
  GUESSINGROWS=50;
```

```
RUN;
```

```
PROC CONTENTS DATA=DATAPREP.readmit_demographic; RUN;
```

```
%web_open_table(DATAPREP.readmit_demographic);
```

- ▷  READMIT_CONDITION
- ▷  READMIT_DEMOGRAPHIC
- ▷  READMIT_GEOMAP
- ▷  READMIT_HOSPITAL

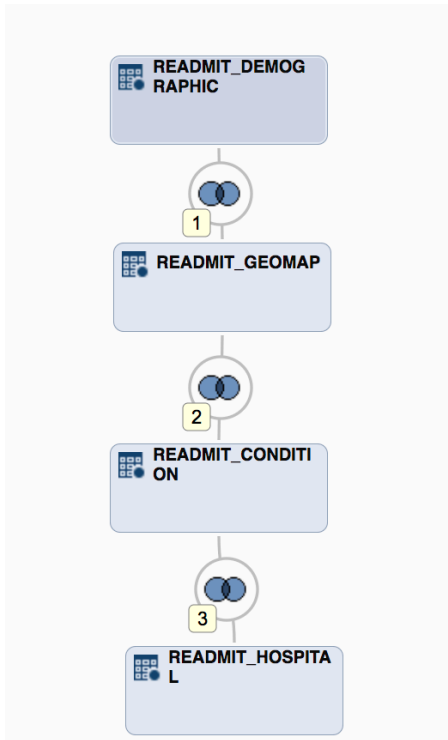
Merging of the SAS datasets into one final file

```
%web_drop_table(DATAPREP.MergedData);

/* Query code generated for SAS Studio by Common Query Services */

PROC SQL;
CREATE TABLE DATAPREP.MergedData
AS
SELECT READMIT_DEMOGRAPHIC.gender, READMIT_DEMOGRAPHIC.race_cd, READMIT_DEMOGRAPHIC.Patient_Age
FROM DATAPREP.READMIT_DEMOGRAPHIC READMIT_DEMOGRAPHIC
INNER JOIN DATAPREP.READMIT_GEOMAP READMIT_GEOMAP
ON
  ( READMIT_DEMOGRAPHIC.ENCOUNTER_KEY = READMIT_GEOMAP.ENCOUNTER_KEY )
INNER JOIN DATAPREP.READMIT_CONDITION READMIT_CONDITION
ON
  ( READMIT_GEOMAP.ENCOUNTER_KEY = READMIT_CONDITION.ENCOUNTER_KEY )
INNER JOIN DATAPREP.READMIT_HOSPITAL READMIT_HOSPITAL
ON
  ( READMIT_CONDITION.ENCOUNTER_KEY = READMIT_HOSPITAL.ENCOUNTER_KEY ) ;
QUIT;

%web_open_table(DATAPREP.MergedData);
```



Code for data cleansing of needed files

```
1 data HOSPITAL.mergedcorrect;
2   set HOSPITAL.mergedcorrect; /* The changes will be done to mergedcorrect data set */
3   if DAYS_ICU < 0 then days_ICU = 0; /* removing all the negative values */
4   if NUMBER_CHRONIC_COND = '*' then NUMBER_CHRONIC_COND = ''; /* changing to a missing values
5   that sas can read in */
6   number_chronic_cond = input(number_chronic_cond, 3.); /* chaning the number_chronic_cond
7   from string to numeric */
8   if Department = '??' then Department = ''; /* changing to a missing value that sas can
9   read in */
10  if order_total_charges < 0 then order_total_charges = .; /* Removing negative values */
11  Run;
12
```

10% sampling process

The screenshot displays the SAS Studio interface. On the left, the 'DATA' pane shows the dataset 'HOSPITAL.MERGEDCORRECT'. Below it, the 'ROLES' section shows 'readmit_number' as the stratification variable. The 'OUTPUT DATA SET' section shows the output dataset name 'HOSPITAL.HOSPITAL_READMIT' and the option 'Include all variables in output data set' checked. On the right, the 'CODE' pane shows the following SAS code:

```
1 /*
2  *
3  * Task code generated by SAS Studio 3.7
4  *
5  * Generated on '9/10/18, 11:21 AM'
6  * Generated by 'manavia0'
7  * Generated on server 'ODAWS04.ODA.SAS.COM'
8  * Generated on SAS platform 'Linux LIN X64 3.10.0-693.21.1.el7.x86
9  * Generated on SAS version '9.04.01M5P09132017'
10 * Generated on browser 'Mozilla/5.0 (Macintosh; Intel Mac OS X 10_
11 * Generated on web client 'https://odamid.oda.sas.com/SASStudio/ma
12 *
13 */
14
15 proc sort data=HOSPITAL.MERGEDCORRECT out=WORK.SORTTempTableSorted;
16   by readmit_number;
17 run;
18
19 proc surveyselect data=WORK.SORTTempTableSorted
20   out=HOSPITAL.HOSPITAL_READMIT TO_SCORE_10PCT method=srs sam
21   strata readmit_number / alloc=prop;
22 run;
23
24 proc delete data=WORK.SORTTempTableSorted;
25 run;
```

Deduplication code of 10% records from the initial file

The screenshot displays the SAS Studio interface with the 'CODE' pane showing the following SAS code:

```
1 data HOSPITAL.readmit_to_model_90pct;
2   merge HOSPITAL.mergedcorrect (in=a) HOSPITAL.hospital_readmit_to_score_10pct (in=b);
3   if a and b then delete;
4 run;
```


Final datasets (90% & 10%) for data mining & predictive modeling and scoring

View: Column names Filter: (none)

Columns Total rows: 144960 Total columns: 42 Rows 1-100

Property	Value
Label	
Name	
Length	
Type	
Format	
Informat	

	ENCOUNTER_KEY	PATIENT_NUMBER	Diagnosis_Group	DIAGNOSIS_LONG_DESC
1	105256108	9921916108	CHF	ACUTE ON CHRONIC DIASTOLIC
2	105256109	9921916109	CHF	ACUTE ON CHRONIC DIASTOLIC
3	105256110	9921916110	CHF	ACUTE ON CHRONIC SYSTOLIC
4	105256111	9921916111	CHF	ACUTE ON CHRONIC SYSTOLIC
5	105256112	9921916112	CHF	ACUTE ON CHRONIC SYSTOLIC
6	105256113	9921916113	CHF	ACUTE ON CHRONIC SYSTOLIC
7	105256114	9921916114	CHF	CONGESTIVE HEART FAILURE UN
8	105256115	9921916115	CHF	CONGESTIVE HEART FAILURE UN
9	105256116	9921916116	CHF	ACUTE MYOCARDIAL INFARCTIC
10	105256117	9921916117	CHF	ACUTE MYOCARDIAL INFARCTIC
11	105256118	9921916118	CHF	CONGESTIVE HEART FAILURE UN
12	105256119	9921916119	CHF	CONGESTIVE HEART FAILURE UN
13	105256120	9921916120	CHF	CONGESTIVE HEART FAILURE UN
14	105256121	9921916121	CHF	ACUTE ON CHRONIC SYSTOLIC
15	105256122	9921916122	CHF	ACUTE ON CHRONIC SYSTOLIC
16	105256123	9921916123	CHF	ACUTE ON CHRONIC SYSTOLIC
17	105256124	9921916124	CHF	ACUTE ON CHRONIC SYSTOLIC
18	105256125	9921916125	AMI	PNEUMONIA ORGANISM UNSPE
19	105256126	9921916126	AMI	BACTERIAL PNEUMONIA UNSPE