Predicting Readmission Probability for Diabetes Inpatients

STAT 471, Qiaochu Guo

Nov 18th, 2018

Table of Contents

# Executive Summary

## Motivation

The percentage of adult with Diabetes has risen from 4.7% in 1980 to 8.5% in 2014, and the WHO expects the rate to continue rising. Currently, it is costing the US about $245 billion each year and has thus become a significant social as well as economic burden to the country. In 2012, the Centers for Medicare and Medicaid Services announced that they would no longer reimburse hospitals for services rendered if a patient was readmitted with complications within 30 days of discharge. This policy change provides extra incentive for us, the healthcare provider, to figure out how to decrease admission rate, hopefully by offering sufficient treatment to all patients when they are admitted into the hospital for the first time.

Diabetes is a chronic condition and can be managed in most cases (type II diabetes). With proper control through medications and lifestyle changes, diabetes onsets can be mostly prevented and thus avoiding the complications brought along with the onsets, which result in costly treatments required through the hospital system. Therefore, active human effort can be made to meaningfully improve the re-admission rate, and it is our goal in this report to find out what factors to watch out for and which procedures prove to work.

## Data Overview

This report uses a cleaned up version of data from the Center for Clinical and Translational Research at Virginia Commonwealth University. It documented over 100,000 unique hospital admissions about ~70,000 patients who stayed for inpatient visits in 130 US hospitals over the 10 years from 1999 to 2008. The dataset records mainly 5 categories of information about the patients: 1) demographic information such as race, gender and age; 2) hospital visit record about number and types of past visits such as emergency vs outpatient, as well as how the patient is admitted and discharged; 3) clinical information such as symptom diagnoses and number of procedures; 4) lab test results of diabetes indicators such as glucose serum level 5) medication prescription type and dosage changes;

After initial EDA, we retained most of the observations in readmission.csv, only excluding the 3 gender Invalid/Unknown cases and thus resulting in a dataset of size 101,763 entries. We treated the NA’s in race and tertiary diagnose as an additional level of categorical variables and log transformed five variables with extremely skewed distribution (detail see EDA section below).

## Methods and Models

With the 28 variables available to be used for prediction (31 features in the dataset, 2 of them are IDs and 1 is the response variable), we first explored their relationships and added three interaction terms. Then separated the dataset into three portions train, validation, and test data following a 8:1:1 proportion.

Using the training data, we conducted backward elimination and created Model.backward, with 14 variables all at 0.05 significance. We also conducted lasso model selection and created Model.lasso, with 7 variables. After comparing them by AUC and misclassification error (with Bayes rule 2:1 penalty) on the validation set, we decided to use the lasso model as the final model. It has slightly lower AUC than Model.backward, yet it has lower misclassification error which we cares more about in the final model used for future predictions. More importantly, it is much sparser than the Model.backward and all of its coefficients are significant at 0.01 level, which means lower chance of overfitting, more likely to generalize well for new data coming in for prediction and also, greater ease of use and lower cost– the hospital needs to collect much less data on a patient to make a decision.

## Results

The Modle.lasso’s 7 variables tell us among the 5 categories of data listed above, hospital visit record, medication prescribtions and diagnoses are especially important, while demographics and lab test results are not. Below are three pieces of key takeaways from the analysis of the coefficients:

1. We should pay special attention to patients with a large number of emergency and inpatient visits, and those who are taking diabetes Medication or many distinct medications in general. We do not believe that these factors “cause” readmission, rather, they are indicators that predictively shows one patient to be of high-risk and needs extra attention. For example, we could set up priority service for these patients to set up meetups with doctors and nurses, and we could follow up with them regularly through phone after discharge to monitor their recovery from onsets and prevent <30 day rebounce.
2. Patients discharged to home (without Home Health Service) has a significantly lower probability to come back within 30 days than patients discharged to home with Home Health Service or SNF and other facilities. This could be explained by that patients discharted to home simply has less severe conditions, but it could also be a very worrying warning sign that our hospital is shifting the responsibility to other health care providers (HHS and SNF) by pre-maturally discharging patients over, only to let their conditions exacerbate and then the patients need to come back for a re-admission within 30 days. It is not only costly with the new policy, but also signals a failure to provide responsible service to let our patients get well as soon as possible. We should re-examine our criteria for discharging to these other facilities and potentially raising the bar of discharge.
3. Patients diagnosed with Diabetes with neurological manifestations (ICD 250.6), occlusion of cerebral arteries (ICD 434) or diabetic coma (ICD 250.2) have a higher chance of readmission. It is possible that it is the case only because these conditions are chronic and physiologically more likely to re-occur, but it is also possible that we are not treating these conditions properly. The hospital should still be careful and could set up special committees to investigate if the treatments to these three conditionc could be improved or if the patients with these conditions could be better advised on how to manage their symptoms.

## Concerns

The final model (Model.lasso) yields a 0.22 MCE when penalizing twice as much for false negatives as for false positives, thresholding at 0.33. However, worryingly, there is a large disparity between specifiticity (99.20%) and sensitivity (3.90%), meaning that the model is not as useful in pinpointing which patient is going to be re-admitted. We did no re-adjust our cost Bayes rule result, so that we are not subject to data-snooping.

Thus, our model is not very strong at pinpointing which patients are goint to be re-admits in a binary way. However, the model is still statistically significant and can inform us to better understand what correlates with readmission and therefore take actions in general to reduce the rate, instead of on specific predicted positive patients.

# EDA and Data Preparation

## Data Cleaning

The dataset contains 101,766 observations and 31 variables, of which 2 are patient IDs and are excluded as a feature, 1 is the response variable and the rest 28 are the initial consideration thought for the logistic model. When examing the levels of all the categorical variables, we found that all levels of all categorical variables are significant except one – gender has 3 Invalid/Unknown entries. Since there are so few of them, we cannot make meaningful inference by setting it as another level, so we drop these 3 observations. Besides that, all categorical variables have at least 150 observations in all levels and we are comfortable with that. (Exhibit 1) Also note that we treated the “?” in race and tertiary diagnosis as an extra layer instead of dropping those observations. Here we assume that “?” is caused by either by lost of record in which case we assume the population to be a balanced mix of all other levels, or they are lost for a specific reason. Either way, we can treat those “?” as homogenuous within group and therefore constitutes as a level.

We also conducted pairwise histogram to examine the distribution of re-admits and non-readmits across all levels for each categorical data. However, no interesting discovery was made. All categorical variable’s distribution does not exhibit noticable difference controlled for re-admits and non-readmits. For reference, we show the histogram for Tertiary Diagnoses to demonstrate the method (Exhibit 2)

# Appendix

If you don’t want to print the R code you wrote (but want to run it, and want to show the results), use a chunk declaration like this: {r, echo=F} • If you don’t want to show the results of the R code or the original code, use a chunk declaration like: {r, include=F} • If you don’t want to show the results, but show the original code, use a chunk declaration like: {r, results=‘hide’}. • Ifyoudon’twanttoruntheRcodeatalluse{r, eval = F}.

Exhibit 1: Categorical Data Levels

for (i in names(re\_data)) {  
 if (is.factor(re\_data[, i])) {  
 print(i)  
 print(summary(re\_data[, i]))  
 }  
}

[1] "race"  
 ? AfricanAmerican Asian Caucasian   
 2273 19210 641 76099   
 Hispanic Other   
 2037 1506   
 [1] "gender"  
 Female Male Unknown/Invalid   
 54708 47055 3   
 [1] "max\_glu\_serum"  
 >200 >300 None Norm   
 1485 1264 96420 2597   
 [1] "A1Cresult"  
 >7 >8 None Norm   
 3812 8216 84748 4990   
 [1] "metformin"  
 Down No Steady Up   
 575 81778 18346 1067   
 [1] "glimepiride"  
 Down No Steady Up   
 194 96575 4670 327   
 [1] "glipizide"  
 Down No Steady Up   
 560 89080 11356 770   
 [1] "glyburide"  
 Down No Steady Up   
 564 91116 9274 812   
 [1] "pioglitazone"  
 Down No Steady Up   
 118 94438 6976 234   
 [1] "rosiglitazone"  
 Down No Steady Up   
 87 95401 6100 178   
 [1] "insulin"  
 Down No Steady Up   
 12218 47383 30849 11316   
 [1] "change"  
 Ch No   
 47011 54755   
 [1] "diabetesMed"  
 No Yes   
 23403 78363   
 [1] "disch\_disp\_modified"  
 Discharged to home   
 60234   
 Discharged to home with Home Health Service   
 12902   
 Discharged/Transferred to SNF   
 13954   
 Other   
 14676   
 [1] "adm\_src\_mod"  
 Emergency Room Other   
 57494 7926   
 Physician Referral Transfer from Home Health   
 29565 6781   
 [1] "adm\_typ\_mod"  
 Elective Emergency Other Urgent   
 18869 53990 10427 18480   
 [1] "age\_mod"  
 0-19 20-59 60-79 80+   
 852 32373 48551 19990   
 [1] "diag1\_mod"  
 250.6 250.8 276 38 410 414 427 428 434 435 486 491   
 1183 1680 1889 1688 3614 6581 2766 6862 2028 1016 3508 2275   
 493 518 577 584 599 682 715 780 786 820 996 Other   
 1056 1115 1057 1520 1595 2042 2151 2019 4016 1082 1967 47056   
 [1] "diag2\_mod"  
 250 250.01 250.02 276 285 401 403 411 413 414   
 6071 1523 2074 6752 1520 3736 2823 2566 1042 2650   
 424 425 427 428 486 491 496 518 584 585   
 1071 1434 5036 6662 1379 1545 3305 1355 1649 1871   
 599 682 707 780 Other   
 3288 1433 1999 1491 37491   
 [1] "diag3\_mod"  
 ? 250 250.02 250.6 272 276 285 401 403 414   
 1423 11555 1369 1080 1969 5175 1200 8289 2357 3664   
 424 425 427 428 496 585 599 707 780 Other   
 1063 1136 3955 4577 2605 1992 1941 1360 1334 42333   
 V45   
 1389   
 [1] "readmitted"  
 <30 >30 NO   
 11357 35545 54864

Exhibit 2: Tertiary Diagnosis Distribution for non-readmits vs. readmits

![](data:application/pdf;base64,)