DSC 520 – Project-Predicting Hospital Readmission of Diabetic patients
Aditya Sumbaraju

**Section 3:**

**Why did you choose this research idea and the dataset?**

I was interested in this dataset because of the large amount of subjects and potential features. There was a good mix of numerical and categorical data, and there are many other avenues still worth exploring. Some parts I found limiting about this data set were the categorical variables like weight and age that could have easily been numerical. In the future, I would like to consolidate many of the drug features into their respective therapeutic drug classes for more meaningful conclusions about drug therapy.

**Summarize the problem statement you addressed.**

The dataset represents ten 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:

- Hospital admission rate using inpatient variable.
- It is a diabetic encounter, that is, one during which any kind of diabetes was entered into the system as a diagnosis.
  Medications were administered during the encounter. The data contains such attributes as the
  - ✓ patient number
  - ✓ race
  - ✓ gender
  - ✓ age
  - ✓ admission type
  - ✓ time in the hospital
  - ✓ medical specialty of admitting physician
  - ✓ number of lab test performed
  - ✓ HbA1c test result, diagnosis
  - ✓ number of medication
  - ✓ diabetic medications
  - ✓ number of outpatients
- The length of stay was at least 1 day and at most 14 days.
- Laboratory tests were performed during the encounter.

**Summarize the methodology employed. Why do you think your method is appropriate? You should refer to the literature.**

In the project, I build the logistic regression model with LASSCO in classifications to identify best model variables. And I have used the Bayes' rule for model evaluation to pick the best model and set the risk ratio at a01/a10 = 0.1 to consider the loss. Because for patients who are readmitted within 30 days, the hospital won't get paid from the social insurance system. In other words, the cost is more considerable for a false negative prediction. I find that the length of a hospital stay, the frequency of the inpatient visit, and the discharge location are the most important predictors for the readmission status.

**Summarize the interesting insights that your analysis provided.**

***Readmission indicator***

About 11.1% of observations have a record of readmission within 30 days.

Patient Demographics

Some patients have more than one record in our dataset. However, a patient's latest medical record can be really different from the previous ones. These duplicate data don't affect the demographic features much if we show them as percentages.

From the table, the composition of race is similar if we compare the proportions between two levels of readmission status. For gender, the composition is similar as well. As for the age distribution, the mean Age is slightly higher for the patients who have a record of ever being readmitted within 30 days.

***Patient Medical History***

As for labp, proc, and nmed, each shares similar distributions in the meas and spreads under two levels of the readmission status. For hosp and diag, the spread of the length of stay in the hospital is larger and the average number of diagnosis is higher for a patient being readmitted within 30 days.

***Patient Medical History***

For nout, ninp, and emer, there are a bunch of 0 in the cells and I compare the distributionS at the tail. Figure shows that for patients readmitted within 30 days, they tend to have lower numbers of outpaitient visits and emergency visites in the previous year.

***Clinical Results***

The correlation between mglu and A1Cr is -0.046, which is not significant.

The diagnosis of diabetes is based on the patients' blood glucose level, i.e. mglu. And the A1c test (A1Cr) is an important measure of how well a person with diabetes is controlling their blood glucose level. The American Diabetes Association recommends a goal of less than 7.0% A1c.

***Medication Details***

To briefly exam the hidden relationship between medicines, I recode the four levels of the dosage (“No” = 0, “Down” = 1, “Steady” = 2, “Up” = 3). I find that the correlations among metf, glim, glip, glyb, piog, rosi, insu are not significant.

**Summarize the implications to the consumer (target audience) of your analysis.**

One of the implications of the analysis was that, though we were not observing the time between follow ups, it showed affect on the readmission rate directly. One may understand without even looking at the data that if a patient does regular and frequent follow-up visits, he can surely control his vitals, thus having a positive effect on his survivability. This was also shown when we tried generating the models and time between follow-ups as a significant parameter.

Another such parameter was age, which one may anyways understand without even analysis that as the Age increases, the risk of readmission increases. So, although we plotted and saw it, it implies that as age will increase, there is always increased readmission risk.

**Overall, write a coherent narrative that tells a story with the data as you complete this section**

To predict the readmission status for patients with diabetes, the length of stay in the hospital, the number of inpatient visits, and where the patient was discharged to after treatment are the most important predictors. Clinical results and medication details may help physicians in the diagnosis in some way. Still, they may provide redundant information we needed for prediction as we already have the health service records. The Key variables are identified as  num_procedures, num_medications, number_emergency, number_inpatient, A1Cresult, metformin, glimepiride, insulin, diabetesMed, disch_disp_modified, adm_src_mod, age_mod, diag1_mod, diag2_mod, diag3_mod. The Bayes Rule Classification Threshold using the risk ratio of 2:1 was 1/3. This means that if the predicted probability of readmission exceeds 1/3, we will predict that that individual gets readmitted. Our misclassification error was about 22%.

**Discuss your analysis's limitations and how you, or someone else, could improve or build on it.**

There is no guarantee that the prediction will be the 'truth.' The model can only provide information in the past, but not necessarily the past trend will continue in the future. Therefore, it would be better to update the model from time to time. Moreover, I suggest taking physicans' views seriously and using the model as a useful reference.

**References:**

Authors of this Dataset:

Data obtained from: [Beata Strack, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura, Krzysztof J. Cios, and John N. Clore, “Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records,” BioMed Research International, vol. 2014, Article ID 781670, 11 pages, 2014.] (https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008)

Discover Statistics Using R, Andy Field | Jeremy Miles | Zoe Field

R for Everyone, Jared P Lander

Think Stats, Allen B Downey