Assignment 1: Analytic Plan – Week 3

Merrimack College

Machine Learning

Amol Gote

20th Sept 2020

Readmission of the patient to the hospital is important predictor from business perspective. The data set contains “Readmitted” categorical variable which indicates following

1. No Readmission
2. Readmitted within 30 days
3. Readmitted in more than 30 days

From the business point of view for various hospitals, No readmission is good case, readmitted greater than 30 days is ok, this could be because of patients’ overall immune system, but patient admitted within thirty days is not a good sign as it is an indicator that prior treatment was not effective. So, in summary the problem statement would be “Will the patient be readmitted and if will it be within 30 days or greater.”

Along with above prediction other informational insights can be produced as well this include

1. The types of Patient’s are readmitted these could be demographic characteristics like Race, Gender, Age and Weight.
2. Length of stay of the patients which are readmitted within 30 days and greater than 30 days. If patient is readmitted within days, its kind of financial penalty to the hospital so the stay needs to be minimized.

The data available for analysis contains records from year 1999-2008 at 130 US hospitals and delivery networks. It contains 50 variables; the data set contains 71518 records. Readmittance distribution is as following

* Readmittance within 30days are 6293.
* Readmittance greater than 30 days is 22240.
* No readmittance is 42985.

Variables of interest are

1. Demographic- age. Distribution of records across various age varies, so it is considered. In case of race caucasian are in majority, gender data is equally distributed, and weight has lot of NA’s so they are not considered.
2. time\_in\_hospital. Data for this variable show normal distribution and could be of significance since there could be a correlation between readmission and number of days spent is hospital.
3. Procedures - num\_lab\_procedures, num\_procedures, num\_medications, number\_outpatient, number\_emergency, number\_inpatient. These could be precursor signs of readmittance hence all the procedure ralted variables are considered.
4. Diagnosis - diag\_1, diag\_2, diag\_3, number\_diagnoses. These variables decide the patient’s treatment care. Certain diagnosis patterns could be causing readmittance to the hospital hence all diagnosis parameters are picked.
5. change – change of medication this could lead to readmission.
6. max\_glu\_serum
7. A1Cresult
8. From Diabetics medication following variables are of interest
   1. Metformin, repaglinide, nateglinide, chlorpropamide, glimepiride, glipizide, glyburide, pioglitazone, rosiglitazone, acarbose, miglitol, insulin
9. From 24 features for medication only glyburide.metformin is of interest, remaining variables for medication in this category have value of not prescribed which is indicator that data is skewed.
10. medical\_specialty. – This variable indicates the specialty of person who treated, data for this evenly distributed, but there are lot of NA’s so the variable is not considered.
11. Here is the final list of variables

age, time\_in\_hospital, num\_lab\_procedures, num\_procedures, num\_medications, number\_outpatient, number\_emergency, number\_inpatient, diag\_1, diag\_2, diag\_3, number\_diagnoses, change, max\_glu\_serum, A1Cresult, Metformin, repaglinide, nateglinide, chlorpropamide, glimepiride, glipizide, glyburide, pioglitazone, rosiglitazone, acarbose, miglitol, insulin, glyburide.metformin

From the above variables only 3 variables have missing values diag\_1, diag\_2, diag\_3 which are less number compared to the dataset size, so dropping those records from the dataset. Also, these variables are categorical variable so numerical imputation like mean or median cannot be applied. Alternate. Dropping NA values drops the dataset size to 70233 from 71518. Also, the certain columns had lot of NA values like weight, medical\_speciality, hence those variables were not chosen. Variable payer\_code also has lot of missing values, but it was not of statistical significance.

* What are the business/research questions to be answered? What are the expected informational outcomes to be produced?
* What data (in terms of source and content) is available? What specific variables are of interest and why?
* What are the expected data due diligence and feature engineering steps? How will missing values be treated? How will categorical variables be encoded?
* The proposed problem should be a supervised classification problem. For this, what is the target of prediction?
* Choose at least two machine learning algorithms (e.g., neural networks, random forest, etc.) that will be used and explain why they were chosen. Note, random forest must be included in the chosen machine learning algorithms.
* What are the expected analytic outcomes (e.g., a supervised classification analysis will produce record/customer specific probabilities)?
* What are the proposed result validation approaches? Specific evaluation methods should be discussed. Note, ROC curves and AUC must be included in the chosen result validation approaches.