

Predicting 30 day hospital readmission likelihood for diabetes patients

Problem:

Insurance companies and hospitals are increasingly focused on decreasing the hospital readmission rate for patients. In addition to reducing costs, the Centers for Medicare and Medicaid Services (CMS) Hospital Readmissions Reduction Program (HRRP) imposes penalties on institutions with excessive readmission rates.

Client and Decision Desired:

Note: This is a capstone project carried out as part of Springboard's Data Science Career Track Course. The client and client scenarios described are fictitious.

The client is a hospital seeking to reduce costs by lowering their readmission rate. The hospital has a pilot program to provide additional support to individuals at high risk of being readmitted to the hospital within 30 days. However, because of a tight budget, the hospital must target the program to patients at high risk of being readmitted. The hospital has already decided they will target patients who have diabetes as this group is unusually likely to be readmitted. The hospital will use this analysis to decide which patients to enroll in the pilot program. They would also like to understand which factors are useful in predicting which individuals are at high risk of readmission.

Data:

I will use the 'Diabetes 130-US hospitals for years 1999-2008 Data Set' which is an open dataset that is obtainable at:

<https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008#>

The dataset includes records from more than 100K patient encounters involving a diabetes diagnosis over 10 years of clinical care at 130 US hospitals and integrated delivery networks. There are over 50 variables including information on patient age, medications, hospital visits, diagnoses, medical tests, and medical procedures and whether or not the patient was readmitted within 30 days of discharge from hospital.

One paper written using this data came to the conclusion that "... measurement of HbA1c is associated with a reduction in readmission rates in individuals admitted to the hospital."

Beata Strack, Jonathan P. DeShazo, Chris Gennings, et al., "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. doi:10.1155/2014/781670
<https://www.hindawi.com/journals/bmri/2014/781670/>

Plan to solve the problem:

1. Initial overview of the data: Look through the data and do some simple visualizations to see the condition of the various factors, including which are in need of cleaning because of missing or bad data and which are extraneous to the analysis and should be filtered out.

2. Data cleaning and other preparation of the data: Based on #1, clean and prepare the dataset for analysis. Segregate an appropriate amount of the data for model preparation, reserving a subset for testing of the model.
3. Exploratory Data Analysis: Obtain an initial understanding of the data through basic summary statistics and visualization. Use this as a guideline for which factors to focus on for further analysis.
4. Data modeling: Based on the EDA (#3) determine which model(s) is an appropriate one to use for predictive analysis. Determine a threshold of significance to evaluate the model.
5. Model evaluation: Based on the data modeling, determine which factor(s) are most important in targeting patients for the pilot program. If this were an actual engagement it would be important to obtain pre and post pilot program information on the hospital client to determine whether the recommendation resulted in the desired outcome (lower readmission rates for diabetes patients).
6. Report: Write up the findings appropriately for both technical and business audiences.

Deliverables:

1. A Jupyter notebook containing the technical steps listed above with runnable code, visualizations, data tables, statistical analysis, data models, and background context.
2. Slide deck of the findings to be used to present to client. Actually, this would be used to present to Springboard audience.
3. Git repository containing above data and analysis.