Machine Learning Engineer Nanodegree

Capstone Proposal

Alfred Lyle

August 22, 2017

Domain Background

My capstone project will focus on a problem in the domain of America's healthcare system. More specifically it will focus on how hospital treatment leads to different outcomes for different sections of the population. As America's population continues to grow and as improvements in healthcare techniques allow people to live longer lives, we would expect to see the number of hospital visits increase. The Centers for Disease Control and Protection (CDC) reports an 11% increase in hospitalizations, from 31.7 million in 2000 to 35.1 million in 2010[1]. Over this same period of time, inpatient hospital deaths declined by 8%, from 776,000 in 2000 to 715,000 in 2010[1].

The CDC has looked at the demographics of inpatient hospital deaths. In 2010, female inpatient hospital deaths at 364,000 are slightly higher than male deaths at 351,000[1]. Breaking down deaths by age, 73% of patients who died in the hospital were aged 65 and over[1]. The CDC also found that patients who died had an on-average longer length of stay in the hospital. This primarily due to the increase chance of hospital acquired infections leading to increased mortality[2]. Other studies have shown that race and gender may lead to increased mortality rates in coronary artery bypass surgery[3]. However, The higher mortality rates seen in women of all races as well as increased mortality in black males has been observed in some studies, but has also been deemed not statistically significant by some researchers when controlling for certain risk factors[4].

---

[1] Margaret Jean Hall, Ph.D.; Shaleah Levant, M.P.H.; and Carol J. DeFrances, Ph.D., **"**Trends in Inpatient Hospital Deaths: National Hospital Discharge Survey, 2000–2010", *Centers for Disease Control and Prevention*, March 2013, https://www.cdc.gov/nchs/data/databriefs/db118.htm#x2013;2010%3C/a%3E%20

2 Rosman, Maya et al. "Prolonged Patients' In-Hospital Waiting Period after Discharge Eligibility Is Associated with Increased Risk of Infection, Morbidity and Mortality: A Retrospective Cohort Analysis." BMC Health Services Research 15 (2015): 246. PMC. Web. 31 July 2017.

[3] Becker, Edmund R., and Ali Rahimi. "Disparities in Race/ethnicity and Gender in in-Hospital Mortality Rates for Coronary Artery Bypass Surgery Patients." *Journal of the National Medical Association* 98.11 (2006): 1729–1739. Print.

[4] Erickson, Sara E. et al. "The Effect of Race and Ethnicity on Outcomes Among Patients in the Intensive Care Unit: A Comprehensive Study Involving Socioeconomic Status and Resuscitation Preferences." *Critical care medicine* 39.3 (2011): 429–435. *PMC*. Web. 1 Aug. 2017.

Problem Statement

The problem hospitals face is determining which patients have a high mortality risk. Patient age and length of stay (LOS) are known risk factors for patient mortality; however, there is no consensus in what other features may be risk factors. Furthermore, demographic features such as marital status, ethnicity, religion, and language are not always filled out in admissions reports. Thus, do these additional features help predict patient mortality, and if so, what is the minimum number of features that can be included without reducing the predictive power of the model?

Datasets and Inputs

The datasets to be used to address this problem will be the MIMIC-III (**M**edical **I**nformation **M**art for **I**ntensive **C**are III) database[5]. This database is comprised of deidentified health-related data associated with over 40,000 patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012. Requesting access to this database can be done at https://mimic.physionet.org/gettingstarted/access/. To be granted access to the database, it is necessary to complete the CITI "Data or Specimens Only Research " course.

I will use this database for my project as it contains 7 patient demographic features including age, gender, ethnicity, marital status, religion, language and type of insurance as well as patient LOS. As the entire patient data comes from a single hospital, any differences in outcomes can be assumed to come from patient features rather than quality of care the patient receives. Before removing NaN and 'Unknown' values, there are 46476 data points with 15737 (33.86%) patient deaths. After removing all NaN and 'Unknown' values, there are 21241 data points with 6622 (31.18%) of patient deaths.

Solution Statement

The solution will be to use a supervised learning algorithm with the MIMIC-III data to create different models using different combinations of features. Using the data, it will be possible to create an initial model using known risk factors of patient age and LOS. This model can then be compared to features without any missing data, gender and type of insurance. Models can then be built using different combinations of the remaining features (ethnicity, marital status, religion, and language) and ranked and compared to the first 2 models. Finally these models will be ranked it will be determined which demographic features hospitals should ensure to get from their patients to best determine high risk patients.

Benchmark Model

[5] MIMIC-III, a freely accessible critical care database. Johnson AEW, Pollard TJ, Shen L, Lehman L, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, and Mark RG. Scientific Data (2016). DOI: 10.1038/sdata.2016.35. Available from: http://www.nature.com/articles/sdata201635

The benchmark will be a model using solely patient age and LOS, which should be better than a model than predicts all patients survive. A similar study on predicting 30 day hospital readmissions was hosted on Kaggle. This competition contained data including patient age, gender, and length of stay. The top team created a model with an AUC score of 0.68469 and the 10 place team had an AUC score of 0.64823[6]. The final models should have an AUC score in this range.

Evaluation Metrics

The model will be evaluated using sklearn.metrics.roc_auc_score function to compare the true patient mortality values versus the predicted values. Models including gender and type of insurance will be compared against the benchmark model. The best model (highest average AUC score over 10 random seeds) will then have the remaining features (ethnicity, marital status, religion, and language) added. These models with added features will be compared against the best model. Finally, it will be seen whether adding these features produced a noticeable difference by comparing average AUC scores.

Project Design

In this project, the MIMIC-III data is spread across a variety of PostgreSQL tables. First the tables will be accessed and column data will be inserted into a Pandas DataFrame. The DataFrames will be merged on SUBJECT_ID and HADM_ID. As there are multiple hospital admissions per patient, and it can be assumed patients survived the earlier admissions, duplicate patient records will be dropped from the DataFrame. Furthermore, rows missing feature data will also be removed. New columns like patient age will be created by subtracting 'admittime' from 'dob'. Groups can be controlled for by only looking at patients over a certain age (ex. >65 years old) and length of stay (ex. >10 days). Finally, the target variable will be stored and dropped from the DataFrame and initial model based on age and LOS will be created.

```
expire_flag = df['expire_flag']
df_features = df[['age', 'los']]
```

After removing the target variables from the features, a test_train split will be used to save data necessary to test the reliability of the model. A classification model can then be fit to the data (ex. tree.DecisionTreeClassifier). A grid search function can be used to tune the hyper-parameters with the scoring_function set to the sklearn.metrics.roc_auc_score score. With the initial model created, further models will be created using more features. These models will need to be run multiple times to ensure the average AUC score is higher. Finally these models will be ranked and compared by AUC score. This information will be visually graphed using matplotlib to convey the conclusions of the project.

---

6 "Private Leaderboard - Predicting 30 Day Hospital Readmissions", *Kaggle Inc.*, April 2015, https://inclass.kaggle.com/c/predicting-30-day-hospital-readmissions/leaderboard/private