Machine Learning Engineer Nanodegree

Capstone Proposal

Alfred Lyle

July 31, 2017

Domain Background

My capstone project will focus on the problems 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].

Problem Statement

While the CDC has looked at a variety of demographic characteristics related to inpatient hospital deaths, they do not have statistics for how race and gender affects patient mortality. Other studies have shown that race and gender leads to increased mortality rates in coronary artery bypass surgery[3]. The authors of that study supposed the reasons for increased mortality rates in certain populations may be due to differences in hospital access and quality of care those hospitals provide or due to genetic differences in race/ethnicity. This is an area of healthcare knowledge that could use further study.

[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.

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[4]. 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 want to use the database for my project as it contains a diverse and large population of intensive care unit patients. The information is also from only one hospital; therefore any differences in outcomes should be due to differences in access or quality of care. This database contains patient demographic information such as language, ethnicity, marital status, and gender. This information can be crosslinked to patient outcomes such as length of stay and mortality. With the information in this database it should be possible to assess how patient gender and ethnicity affects medical outcomes.

Solution Statement

Using a supervised learning algorithm it will possible to find which combinations of race and gender have higher mortality rates. Since the MIMIC database has so many patients, it will also be possible to add features such as marital status and spoken language to see if they have any effect on patient mortality. Our classifier should be able to find and rank the most important demographic features for patient mortality. However, as a majority of hospital ICU admissions generally do not result in patient death, it may be more useful to create an algorithm that looks at increased patient length of stay (LOS) while controlling for the same initial diagnosis and patient age. The algorithm should be able to predict test patient LOS and mortality based on our input features with 80% accuracy to show it is a reliable algorithm.

Benchmark Model

The current model is a study of 9,518 ICU patients in 35 California hospitals from the years 2001-2004[5]. It used linear regression models to determine the effect of race/ethnicity on patient mortality and LOS. This study controlled for age, gender, Acute Physiology Score, DNR status, Socio-economic status, and expected source of income. The outcome of the study was they found hospital mortality and ICU LOS did not differ by race or ethnicity.

Evaluation Metrics

[4] 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

[5] 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.

In this study it will be far more important to correctly determine if a patient dies (true positive) rather than incorrectly stating a patient will die (false positive). For this reason a $F_{0.5}$ score should be used to evaluate the quality of the algorithm's predictions. The algorithm that gives the highest $F_{0.5}$ score can then be evaluated for the most important features.

Project Design

For this project it will be necessary to access the PostgreSQL database and then use classified learning algorithms in Python to analyze the data. Using psycopg2 it is possible to query PostgreSQL databases from Python. It is then possible to pull information from the tables inside the MIMIC-III database.

```
curr.execute("SELECT LOS from ICUSTAYS")
```

These tables can be cross-linked to the PATIENTS and ADMISSIONS tables using the SUBJECT_ID and HADM_ID data. The relevant features can all be merged into a new dataframe.

```
mimicdf = pd.merge(patients, los, on='SUBJECT_ID')
```

With this dataframe in place, it will be necessary to remove data points missing LOS, ADMITTIME, DISCHTIME, DEATHTIME and other important values from the dataset. The target variables will then be stored and removed the dataframe.

```
los = mimicdf['LOS']
mimicdf_features = mimicdf.drop('LOS', axis=1)
```

After removing the target variables from the features, using a test_train split to create test data will be 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 $F_{0.5}$ score. The feature_importances can determine which features are most important in predicting outcomes. Further analysis can be done by controlling for age, and diagnosis by creating new datasets all of a certain age or illness and then fitting the data to a new model.