Machine Learning Engineer Nanodegree

Capstone Proposal

Alfred Lyle

August 14, 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 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.

Problem Statement

There are many known risk factors for increased patient mortality; however, there is no consensus on what impact race and gender have on patient mortality. The higher mortality rates seen in women of all

---

[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.

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]. It is important to determine what effect race and gender have on patient mortality in order to determine future public health policy to address this issue. Using hospital data and machine learning algorithms, can it be shown that including patient's gender and race increases the algorithm's accuracy in predicting patient mortality?

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 a variety of patient demographics as well as patient outcomes. 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. Using this dataset, it is possible to control for certain risk factors such as patient age and length of stay. Patient features including gender, ethnicity, marital status, religion and type of insurance can be used as inputs for the learning algorithm.

Solution Statement

The solution will be to use a supervised learning algorithm with the MIMIC-III data to show determine what effect patient features have on patient mortality. Using the data, it will be possible to create groups of high risk adults, those aged over 65 years and whose length of stay (LOS) is greater than 10 days. The algorithm will then predict patient mortality using a variety of patient features including gender and race. Once the algorithm has been trained, it is possible to determine which features have been given the highest weight values in predicting patient mortality. If race and gender have the top 2 highest weights, it will be shown that race and gender play a role in patient mortality.

Benchmark Model

[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.

[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 that predicts all patients survive, which would give an accuracy of 66.14% and an Area Under the Curve (AUC) score of 0.5. 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].

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. This function will give blind guessing an AUC score of 0.5 and a model that correctly predicts every value a score of 1.0. Thus, the algorithm should have an AUC score greater than 0.5 and should be similar to the top Kaggle submissions with an AUC score of ~0.65.

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.

```
expire_flag = df['expire_flag']
df_features = df.drop('expire_flag', axis=1)
```

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. Once the model is better than the benchmark models, the feature_importances can determine which features are most important in predicting outcomes. This information can 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