CSML1030: Machine Learning Capstone

Machine Learning Certificate, York University

May 10, 2020

Hospital Mortality Reduction: Machine Learning Approaches

Author:
Peter Gray[a]

Author:
Pratik Chandwani[b]

Author:
Conrard Giresse Tetsassi Feugmo[c]

Collaborator:
David Rotenberg[d]

Professor:
Cecilia Ying[e]

[a] Student, CSML1030
[b] Student, CSML1030
[c] Student, CSML1030
[d] Krembil Centre for Neuroinformatics at CAMH (Center for Mental Health and Addiction)
[e] York University School of Continuing Studies

This work aims to discover possible approaches to improving hospital care by using machine learning methodologies on both demographic and clinical patient data. A step-by-step analysis will be conducted, and factors that affect deaths or suicide attempts among patients will be identified.

I. Introduction

A. Background

More than 165,000 Canadians die in hospitals every year [1]. This is approximately 60% of all deaths in Canada. There are wide variations in hospital mortality. Much of this variation remains unexplained and may reflect quality of care. Among the causes of death in Canadian hospitals, the top six are [2]:

- Stroke
- Chronic obstructive pulmonary disease (COPD)
- Sepsis
- Heart attack
- Pneumonia
- Heart failure

According to the Canadian Mental Health Association, there is a fundamental link between mental health and physical health [3]. People with serious mental illness are at higher risk of chronic physical conditions, and those suffering from chronic physical conditions are at higher risk of anxiety and depression.

Frequent visits to emergency rooms may indicate that people are not getting access to the services or the support they need in the community for help with mental health and/or addictions. According to the Canadian Institute for Health information, many people visit an emergency room at least 4 times a year [4]. Among these people 50% are under age 35 (including youth and children under age 20), 32% have poor mental health conditions (anxiety, mood disorder) and 18% have addiction to substances (usually alcohol).

The death rate at hospitals constitutes an indicator of health care quality. By analyzing clinical and demographic data, patterns that are correlated with a higher risk can be discovered. An understanding of these patterns could be applied to reducing patient deaths and improving the quality of care.

B. Objectives

The broad objective of this research is to enable a reduction in hospital mortality by using machine learning methodologies on datasets that are available to the public.

More specifically, the objective is to identify data features that can be used to predict an increased risk of patient mortality. This knowledge can be used to improve information collection practices to ensure that data that would be the most useful for machine learning applications is collected as accurately as possible.

And in general, the objective is not to build a precise tool that is useful in the field. We put primary emphasis on interetability and explainability so that our results can be used to inform decision making on other Machine Learning projects, as well as information collection policies.

These objectives were improvised on short notice in response to COVID-19 closures, which began exactly when this project did. As a result, bridging the gap between CAMH's business problems, and publicly available data, has been the defining essence of this project and its objectives.

C. Datasets

After examining many datasets for their suitability to our improvised objective, two datasets were chosen based on their broad range of features, adequate size, and their relevance to the task of reducing hospital mortality.

The first dataset is from Kaggle's Women In Data Science (WiDS) 2020 Datathon[5]. This dataset has been used in a competition to predict hospital mortality based on clinical data collected within the first 24 hours after patient admission.

The second dataset was provided to our team by CAMH, and contains data from Mexico regarding hospital mortality, including demographic and clinical data. This dataset also has data about suicide related hospital mortality, and as such is especially relevant to CAMH's healthcare challenges.

II. Exploratory Data Analysis

It took us a lot of effort to explore the features. We explored each and every features' distribution and their correlation to the target. We found medical reasoning for the correlation of some features and that gave us some idea of how to impute missing values smartly, and what kind of features we should extract. We did adversarial validation (accelerated by SparkBeyond) to help us decide on dropping some features manually, notably the hospital & ICU ID columns, as well as the hospital and encounter IDs.

III. Feature Imputation Approaches

- Categorization based on domain knowledge (with a dummy missing category)- in medicine missing values are often very informative! E.g., a patient not doing a test may often mean that the attending staff did not think they needed it
- Deducing probable values from other variables (based on domain knowledge) - for example surgery or ventilation status
- Imputation based on domain knowledge - e.g. BMI
- Predictive Model imputation - using XGBoost to predict the variable, then using predictions for the imputation. (Used for the apache death risk variables. We used a subset of the input data for this submodel, with columns removed according to their percentage of missing values, so as to more closely match the distribution of instances missing the target variables)
- Statistical imputation - mean/max/mode/average , or imputation of variables with a value greater than the maximum or less than the minimum (to help distinguish the missing variables when using tree based models). Many - of the models used supported missing values natively, and these were also used as part of the ensemble. (e.g. LGBM, XGBoost, Catboost).

IV. Feature Engineering

Based on domain knowledge or manual expertise:

- Creating a "golden" feature called mostlydead, which groups insights from different biological functions or measures that are not compatible with life (and provided good outcome separation)
- Evaluating the perception of patient severity by the medical staff based on number of measures taken in the first hour type of tests (for example whether an arterial blood was tested or not)
- Grouping variables with the same meaning, for example values measured both in an invasive and non-invasive manner
- Creating categorical variables that convey the severity of a specific physiological or laboratory measure by creating categories of extreme values from both ends of the normal range
- Deltas/Diff Variables that measure fluctuations of physiological and laboratory measurements by subtracting the maximal and minimal values, e.g. the value range of a specific test in the first hour, first day, first day vs first hour, and whether a test was started after the first hour
- Aggregated features - e.g. counting how many multiple chronic conditions or subgroups of chronic conditions a patient had, the total number of tests performed, summation of tests or diagnostics with missing values,
- Grouping medical diagnoses codes into semantic categories based on the hierarchical nature of the medical coding system
- Grouping of GCS measures to a final GCS score
- Computing renal function based on relevant lab test and diagnostic codes
- Computing pulse pressure variables by subtracting the systolic and diastolic BP measures
- Measures of anomalousness - Isolation forest derived anomaly score (computed using +/- columns), counts of row level missing values and zeroes.
- LSH/Hashed "profile" features - based on rounding the age and weight + chronic diseases so as to define "clusters" of patients. We also added some derived features based on these demographic clusters (e.g. mean apache score).
- Computing a variable that deduces non-elective surgery
- 'Hospitalid' had small overlapping values between train and test , so we dropped disjoint values between test-train lists.

Based on automated feature engineering using SparkBeyond)(SB) for feature exploration insights (that were then manually implemented)

- Non linear feature interactions - these were then captured as interaction features. E.g. ratios between features/lab tests, the max or min of multiple different values/tests (such as: the maximum of the hospital and icu apache death prob risk score, d1lactatemin /diasbpmin, d1heartratemin / d1sysbpmin, d1resprate div mbpmin)
- Model agnostic ranking and selection of features by SB - this was not used for the actual models due to the rules, but was highly conducive to insights and ideas for further exploration
- Semantic features were examined by using Sparkbeyond's engine to connect external knowledgebases (such as wikipedia) to the textual descriptions of the medical diagnosis codes (acquired via the codebook shared by the competition organizers). These features were insightful but did not improve the models performance, presumably due to the low cardinality of the structured codes making high level semantics less beneficial in this use case, so we did not use them.
- A few such concepts from the diagnosis codes associated with higher risk of death includes relations to: endarterectomy, intensive care medicine, pulmonology, respiration & lungs, infectious diseases (e.g. sepsis, infections).
- Concepts with a lower risk of death included: coronary artery bypass (e.g. open heart surgery), medical specialties (e.g. surgeons), drugs (e.g. overdose or drug withdrawals) crime.

V. Feature Selection

We used several different techniques for features selection. We checked these techniques separately and made different experiments whilst iteratively dropping features within a fast LGBM pipeline, according to what improved the local validation or Leaderboard score. We dropped features using a mix of the following methods:

- Features with a high percentage of missing values (75% and up)
- Collinear (highly correlated) features - with threshold above 0.99
- Features with zero standard deviation (contains the same value, only 'readmitionstatus')
- Features with high ranking in RFE (linear regression model, like 'paco2forphapache').
- Features with zero importance and zero influence with permutation in ELI5 (like 'hepaticfailure').
- Features with a different distribution between train and test - according to adversarial validation models and K-S tests ('icuid' for example).
- Aggressive and highly important features ('apache4ahospitaldeathprob', 'apache4aicudeathprob', and derived features) were dropped in most of our models. We found that this improved our validation performance, likely due to using the same features to calculate this score and the bias in how the score is calculated by humans - leading the model to overfit with it.
- Adversarially selected features with near perfect train/test seperation - with no coverage over the test set (i.e the ID columns noted above)

VI. Modeling

Modeling approaches that result in high interpretability and explainability were favoured, so that results would be transferable to CAMH's data collection policies and future development of machine learning applications. We used:

- Boosting trees (Catboost, Xgboost, Lightgbm, h2o GBM)
- Scikit learn (KNN, Logistic Regression, Random Forest, Extremely Randomized Trees, Histogram-Based Gradient Boosting)
- Deep learning models (Pytorch and Tensorflow) - not that good results but diverse and contributed to the model
- Ensembling We used 3 different methods for a strong and robust ensemble of models:
  - Diverse data sets - imputation techniques, features, and data generation
  - Stacking the models using 3 layers stacknet. Normalize the models probabilities.
  - Time Augmentation (TTA). TTA is a novel technique that aims to increase accuracy and decrease unfair bias. It works by by generating new similar test sets with mild changes. For example, we changed the age, gender, and ethnic values. Then we ensemble them together. A research paper describing the method will be released in the future.

VII. Conclusions—Most Important Features

A. The Most Important Data Features for Predicting Hospital Patient Mortality with Machine Learning

1. Basic Decision Tree

Impurity-based feature importance as ranked by a basic decision tree:

1) Lowest systolic blood pressure
2) Lowest peripheral oxygen saturation
3) The motor component of the Glasgow Coma Scale
4) Lowest lactate concentration
5) Lowest heart rate
6) Highest lactate concentration
7) Lowest mean blood pressure
8) Lowest body temperature
9) Lowest arterial pH
10) Temperature resulting in highest APACHE III score
11) Highest body temperature
12) The eye opening component of the Glasgow Coma Scale
13) Blood Urea Concentration resulting in the highest APACHE III score
14) Lowest bicarbonate concentration

2. Light Gradient Boosting Machine

Impurity-based feature importance as ranked by a Light GBM:

1) Age
2) Lowest heart rate
3) Length of hospital stay before ICU admission
4) Blood urea concentration resulting in highest APACHE III score
5) Highest platelet count
6) Highest body temperature
7) Lowest lactate concentration
8) Lowest body temperature
9) Highest heart rate
10) Lowest respiration rate
11) Lowest systolic blood pressure
12) Highest systolic blood pressure
13) Lowest glucose concentration
14) Highest sodium concentration

B. The Most Important Data Features for Predicting Suicide as Cause of Death with Machine Learning

1. Random Forest

Most important features as ranked by a random forest:

1) Age
2) On medication
3) Population (size of the city)
4) Marital status - Single/not-single
5) Violence in the family
6) Education level
7) Industrial machinery operators
8) Urban/Rural Area

2. Light Gradient Boosting Machine

Most important features as ranked by a Light GBM:

1) Age
2) Violence in the Family
3) Gender - Female
4) Urban/Rural Area
5) On medication
6) Size of the city
7) No schooling
8) Unemployed
9) Speaks indigenous language

VIII. Analysis—Explainability and Interpretability

We achieved an ROC AUC of 0.93 with an LGBM on the Hospital Mortality data, and 0.95 with AdaBoost on the Cause of Death data. However, the objective of this project was not to train a workable model, but to gain an interpretable understanding of which data features contribute the most to the performance of machine learning models.

To gain insights into how our models made their decisions with the features identified, we output visualizations of the tree. The two hand-labeled visualizations included below show several paths to a positive prediction of mortality (indicated by a blue box), and allow one to see the decision flow and how the various features factor into how the prediction is made:

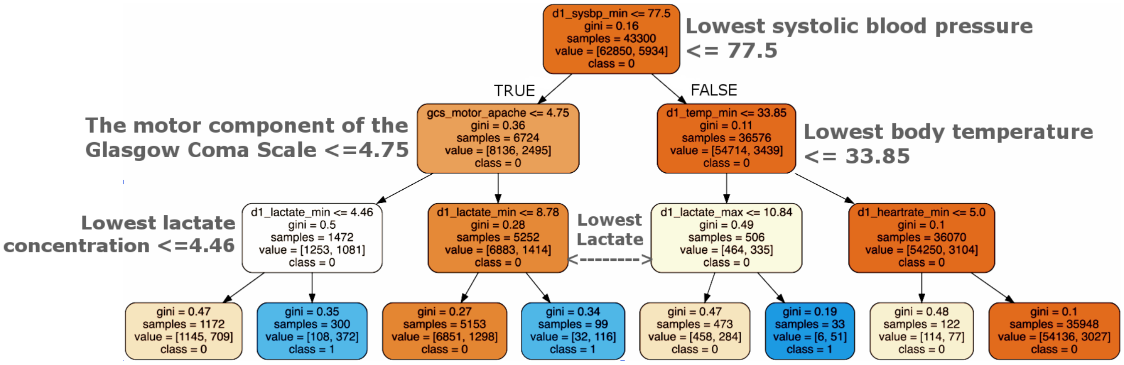

Figure 1—Visualizing a decision tree for predicting hospital mortality. In this case, the first decision is based on blood pressure. If the lowest systolic blood pressure is low enough, the motor component of the Glasgow Coma Scale is considered next. Following that, the lowest lactate concentration provides the final clue the model needs to make a prediction.

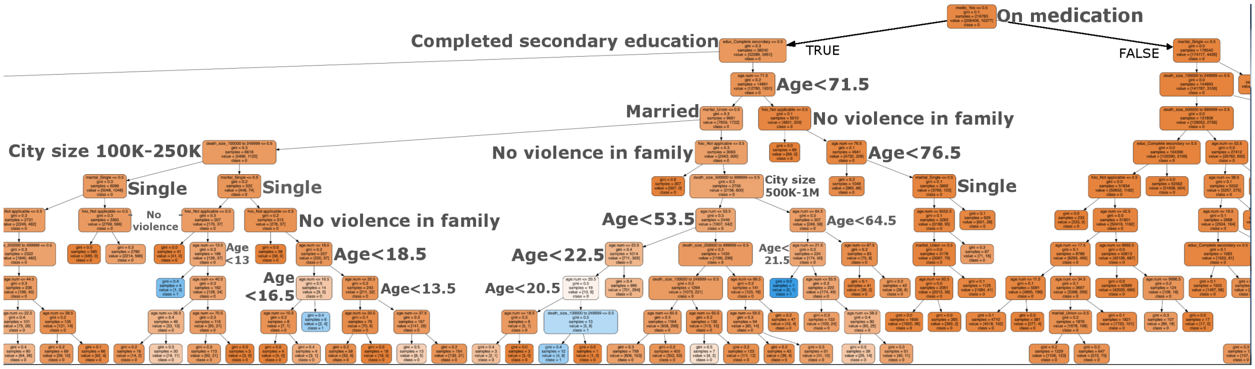

Figure 2—Visualizing a section of a decision tree for predicting Suicide as Cause of Death. First, if a patient is on medication, then, has not completed secondary education, then is less than 71.5 years old, then is not married, and then has violence in their family, they are leading to a part of the decision tree that is rife with positive results. Following further along this logical pathway, the patient's age provides the indicators needed for prediction.

IX. Recommendations & Future Work

This project has its roots in future work. The datasets that were chosen, the machine learning approaches that were applied, and the way the results have been presented, are all intended to guide improvements to the collection of data for future research.

The number of features we have identified, and the number of pieces of patient data that can be collected, are very finite. Some, like blood pressure, are simple. Some, like substance abuse problems, are multidimensional. Because of the sensitivity of each individual feature, we feel the most meaningful way to begin to use our results for the improvement of data collection would involve interpretation by people with domain knowledge.

By comparing the important features as determined by this research with current data collection policies, and with how accurately that data has been collected up until now, persons with knowledge of how this data is collected could make meaningful suggestions for improvement.

Ultimately, using this optimized data to train Machine Learning applications could enable the development of tools that provide more penetrating diagnostic insights.

Acknowledgments

Our group would like to acknowledge the help of David Rotenberg, Operations Director for the Krembil Centre for Neuroinformatics at CAMH. We would also like to thank our professor from the York University School of Continuing Studies, Cecilia Ying, for all her assistance throughout this project.

References

[1] "Statistics Canada. (2018). Table 13-10-0715-01 Deaths, by place of death (hospital or non-hospital)," https://doi.org/10.25318/1310071501-eng, 2018. Accessed: 2020-03-25.

[2] "Hospital Deaths (HSMR)," https://yourhealthsystem.cihi.ca/hsp/inbrief?lang=en#!/indicators/005/hospital-deaths-hsmr/;mapC1;mapLevel2;/, 2017. Accessed: 2020-03-25.

[3] "The Relationship between Mental Health, Mental Illness and Chronic Physical Conditions," https://ontario.cmha.ca/documents/the-relationship-between-mental-health-mental-illness-and-chronic-physical-conditions/, 2008.

[4] "Frequent Emergency Room Visits for Help With Mental Health and/or Addictions," https://yourhealthsystem.cihi.ca/hsp/inbrief?lang=en#!/indicators/005/hospital-deaths-hsmr/;mapC1;mapLevel2;/, 2018. Accessed: 2020-03-25.

[5] "Women in Data Science (WiDS) Datathon 2020," https://www.kaggle.com/c/widsdatathon2020/overview, 2020.