2019

Outcome Prediction in Intensive Care Unit Settings with Claims Data

Lauren Staples
lstaple6@students.kennesaw.edu

Ryan Rimby
Kennesaw State University, rrimbey@students.kennesaw.edu

Follow this and additional works at: https://digitalcommons.kennesaw.edu/dataphdgreylit

Part of the Statistics and Probability Commons

Recommended Citation

This Conference Proceeding is brought to you for free and open access by the Ph.D. in Analytics and Data Science Research Collections at DigitalCommons@Kennesaw State University. It has been accepted for inclusion in Grey Literature from PhD Candidates by an authorized administrator of DigitalCommons@Kennesaw State University. For more information, please contact digitalcommons@kennesaw.edu.
INTRODUCTION

The motivation for this project was to see how accurate (and operational) outcome prediction can be from only billing and demographic data. This was done to see if a company would likely be interested. The dataset selected for this project is the MIMIC-III database, which contains billing data, lab data, doctor and nurse notes, and demographic data (while remaining deidentified). This dataset was collected from Beth Israel Deacon Hospital from 2001-2012 [1]. Billing data includes procedural codes (CPT) and diagnoses codes (ICD-9), which are both nominal data types and require special variable transformation, as described in the Methods Section.

METHODS

A dependent or target variable was created to depict an outcome variable that describes death within 30 days of discharge. This variable was simply assigned as binary (1 if death occurred). Deaths were collected from 3,939 cases, and so the non-death cases were under-sampled to create a balanced dataset.

Predicting variables ICD-9 and CPT codes (defined in Section ICD-9 Codes Explained) are nominal in nature, and so a procedure called One Hot Encoding was used to create new binary variables for each level of the nominal variable. Since there are 12,000 ICD codes in our data, a simple transformation would produce 12,000 predictor variables. This procedure was used and this was called the High Dimension Model (Figure 1).

However, due to the hierarchical nature of ICD-9 codes (see section ICD-9 Codes Explained), a second method of variable transformation was employed. The first four characters of the ICD-9 variables were split by position into four new variables. These four new variables then had 12 levels (10 and V), or 48 One-hot encoding of these variables produced 4 x 48 (42 variables) binary variables. This method is called the Low Dimension Model in Figure 1.

Logistic Regression was applied with both the low and high dimension data sets. While Sklearn had a cutoff of 0.1, it has a function for creating a ROC curve in metrics that calculates fpr, tpr, thresholds, and thus two ROC curves were created for comparison in performance (See Results).

Important variables were determined by Recursive Feature Elimination, setting parameters for 10 fold cross validation. This was only performed for the Low Dimension Model, as shown in Figure 1. Collinearity and significance were examined as well. Lastly, a single layer Neural Network was implemented to see if it could outperform logistic regression.

RESULTS

Figure 2 shows Receiver Operator Characteristic curve for logistic regression both the low and high dimension model. The Area Under the Curve for each model is around 79% for both models, showing that there is no significant information lost when using the Low Dimension Model. Using a cutoff value of 0.2, both models have an averaged cross-validated accuracy of ~79% with logistic regression. With a balanced binary target, this is a 21% lift over chance.

Figure 4 shows the most important variables from the Low Dimension Model, found via recursive feature elimination (RFE). The significance and variance inflation factor of the 20 important variables retained were assessed, and 10 of the 20 were retained. Figure 3’s side bar highlights at least one interesting variable. With patients who have secondary diagnosis beginning with V, there is a lower probability of death within 30 days of discharge. This variable is in the group “factors influencing health status and contact with health services” and include exposure to communicable diseases, a need for vaccinations, and need for isolation.

Lastly, Figure 4 shows the results for the neural networks, which were performed with 10 fold cross validation on the Low Dimension Model in two ways, once with only the 17 variables after RFE, and once on the Low Dimension Model with no variables removed with RFE. The average results are not similar to one another, they are similar to the average accuracy found by logistic regression.

Ibid-9 Codes Explained

ICD-9 Diagnosis Codes are version 9 of the International Statistical Classification of Diseases and Related Health Problems. The Agency for Healthcare Research and Quality, an independent organization of the Health and Human Services. They are used worldwide for morbidity and mortality statistics, reimbursement systems, and automated decision support in health care.

The ICD-9 codes have a hierarchical structure, as seen in Figure 4. The highest level is the first character, followed by next level, and so on. The first character can have digits 0-9, or the letters A or V. Codes containing A in the second position show that it is in the ‘other heart disease’ category, the 8 in the third position shows that this disease is in the ‘other infectious or parasitic diseases’ category. For example, in the example to the right of Figure 4, the 4 shows that the diagnosis is in the heart category, the 2 in the second position shows that it is in the ‘other heart disease’ category, the 8 in the third position shows that this disease is in the ‘other infectious or parasitic diseases’ category. The hierarchy nature of these diagnosis codes were capitalized on in our feature reduction technique, described in the Methods Section.

SOURCES


6. The website for the data is: http://mimic.physionet.org/download/mimic3/.

7. https://www.mimic.ohsu.edu/ for more information on the dataset.