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

Recommended Citation

Staples, L, R. Rimby. Outcome Prediction in Intensive Care Unit Settings with Claims Data. Southern Data Science Conference Poster Submission, 2019.

Outcome Prediction in Intensive Care Unit Settings with Claims Data

KENNESAW STATE UNIVERSITY
ANALYTICS AND DATA SCIENCE INSTITUTE

Lauren Staples, Ryan Rimbey

Kennesaw State University

ABSTRACT

The MIMIC-III data comes from an Intensive Care Unit in Boston over a period of ~10 years. This data contains billing codes as well as lab and demographic data. This project predicts the outcome "death within 30 days of discharge" through the lense of a healthcare billing company, to see if healthcare companies can play a role in healthcare quality, by only using data that they would have access to (billing and demographic data). This project used a unique method of nominal data variable reduction specific to ICD-9 and CPT codes, and compared the performance of logistic regression and neural networks on the prediction of a balanced binary target variable (death within 30 days of discharge). Averaged cross validated accuracies of all methods were around 71%, which is 21% better than chance alone.

INTRODUCTION

The motivation for this project was to see how accurate (and operational) outcome prediction can be from **only using billing data and demographic data**, data which a health care company would likely be limited. 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 occurred in 5, 939 cases, and so the non-death cases were undersampled to create a balanced dataset.

Predictor 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 had to be performed. This creates a new binary variable for each level of the nominal variable. Since there are 12,000 ICD codes in our data, a simple transformation would produce 11,999 predictor variables. This procedure was used and this was called the High Dimension Model (see 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 new variables then had up to 12 levels (0-9 and V or E). One-hot encoding of these variables produced 41 (42 minus 1) 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 uses a cutoff of 0.5, 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 Factor Elimination, setting a retainer of 20 variables. 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.

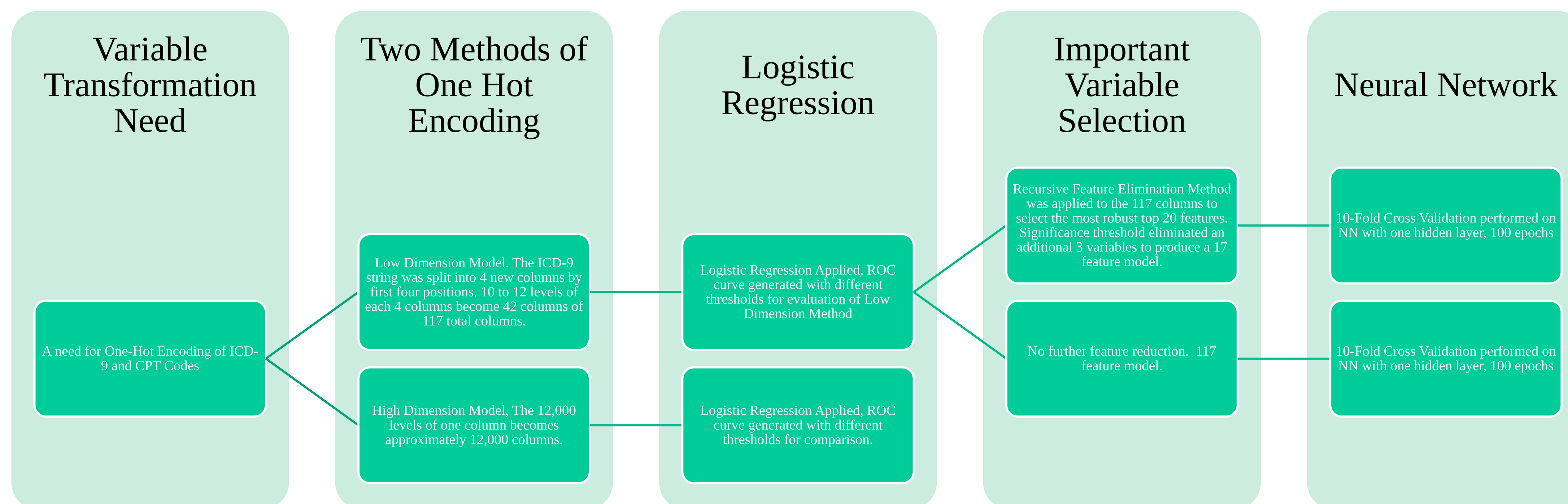

Figure 1- Analysis Flow Chart.

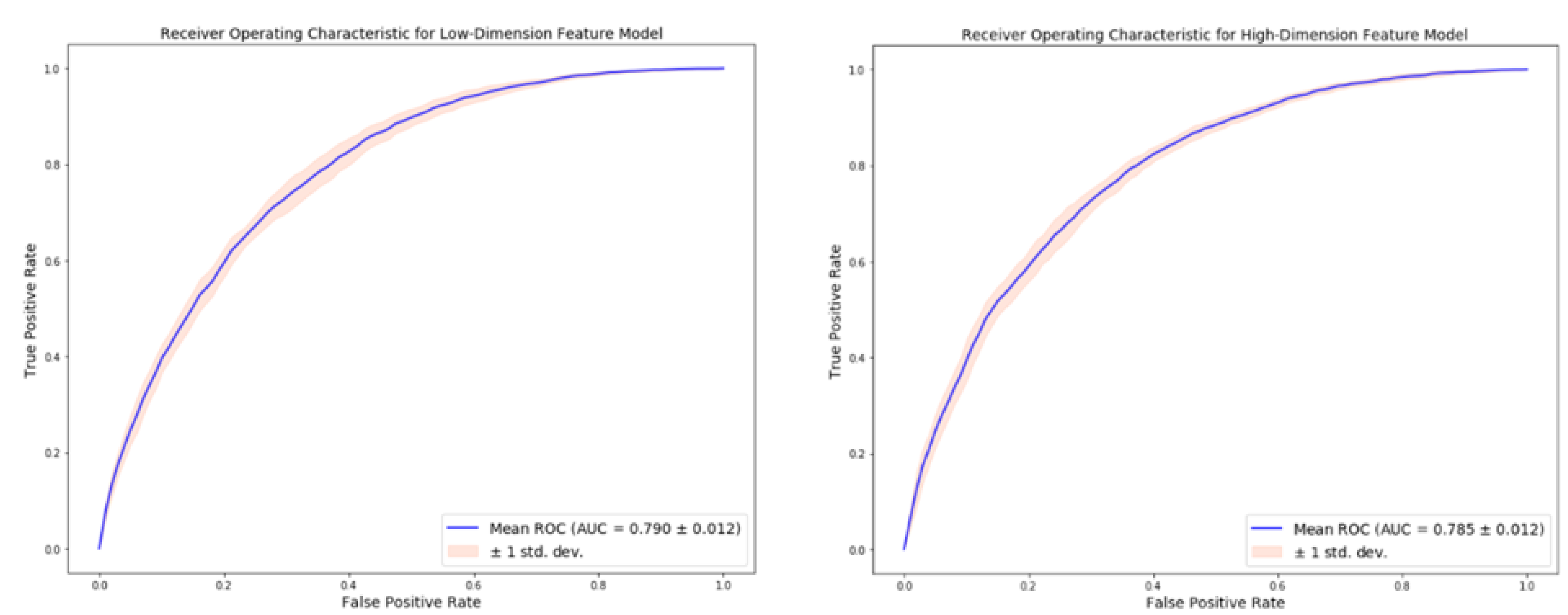

Figure 2 – Left: ROC for Low Dimension Model. Right: ROC for High Dimension Model.

Parameter	Coefficient	P>|z|
CPT_pos2_0	-2.2282	0.0354
intercept	-1.8461	0
ADMISSION_TYPE_ELECTIVE	-1.6273	0
CPT_pos2_1	-1.5242	0.022
ICD_sec_pos1_V	-1.2208	0.0003
ICD_sec_pos1_E	-1.0123	0
ICD_sec_pos1_6	-0.9419	0.0168
ICD_prim_pos3_4	-0.8862	0
CPT_pos1_3	-0.7563	0.0084
CPT_pos3_2	-0.6446	0
ICD_sec_pos1_1	0.4771	0
ICD_prim_pos4_	0.6272	0
CPT_pos3_0	0.7834	0
ICD_prim_pos1_0	0.8943	0
ICD_prim_pos1_1	1.0935	0
ICD_prim_pos1_V	1.1812	0.0016
age_at_admit	2.7255	0

Patients with secondary diagnoses in the ICD-9 group beginning with 'V' have a lower probability of death within 30 days of discharge. ICD-9 category V are 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.

Figure 3 – Important Features by Recursive Feature Elimination (Low Dimension Model).

Model	10-Fold Avg. Accuracy	Standard Deviation
Low Dimension Model, with Recursive Feature Elimination	69.47%	1.42%
Low Dimension Model, NO FURTHER Recursive Feature Elimination	71.90%	1.16%

Figure 4 – Neural Network Cross Validated Accuracy/Standard Deviation.

ICD-9 Codes Explained

ICD-9 Diagnosis Codes are version 9 of the International Statistical Classification of Diseases and Related Health Problems. They are published by the World Health Organization. They are used worldwide for morbidity and mortality statistics, reimbursement systems, and automated decision support in health care.

The ICD-9 codes have a hierarchal structure, as seen in Figure 4, below. This structure shows how each position has a meaning. The first position can have digits 0-9, or the letters E or V. Codes containing a 0 in the first position are infectious or parasitic diseases, 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 heart failure category. The hierarchal nature of these diagnoses codes were capitalized on in our feature reduction technique, described in the Methods Section.

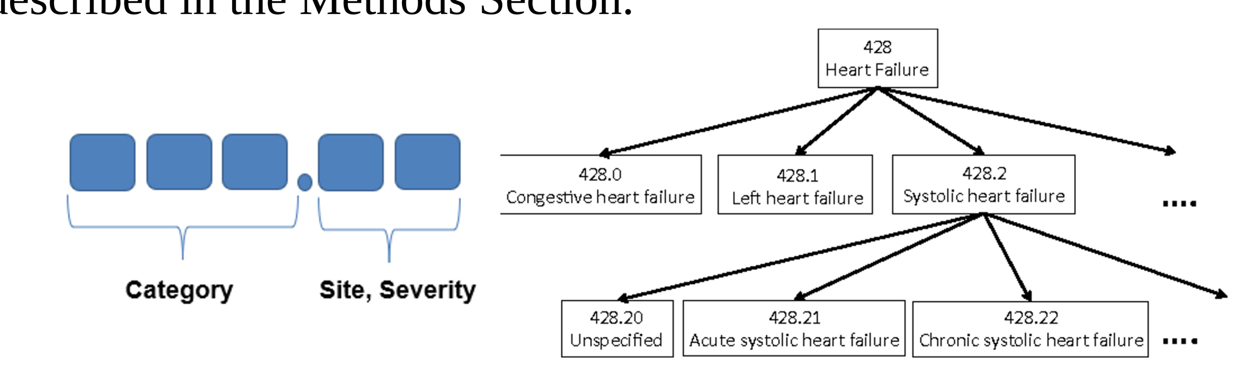

Figure 5- Hierarchical Nature of ICD-9 Codes

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 loss when using the Low Dimension Model. Using a cutoff value of 0.5, both models have an averaged cross-validated accuracy of ~71% with logistic regression. With a balanced binary target, this is a 21% lift over chance.

Figure 3 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 attained from RFE were assessed,, and 17 of the 20 were retained. Figure 3's side bar highlights at least one interesting variable: patients with a secondary diagnosis beginning with V have a lower probability of death within 30 days of discharge. This code category is in the group "factors influencing health status and contact with health services" and include exposure to communicable diseases and a need for vaccinations and isolation [7].

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 only similar to one another, they are similar to the average accuracy found by logistic regression.

SOURCES

1. 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.
2. Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.
3. Receiver Operating Charateristic, webpage tutorial, https://scikit-learn.org/stable/auto_examples/model_selection/plot_roc.html.
4. Kang, Mingon. Course Notes, Machine Learning (CS-7267, Fall 2018, Kennesaw State University).
5. Singh, Anima et al. "Leveraging hierarchy in medical codes for predictive modeling." BCB (2014).
6. https://en.wikipedia.org/wiki/List_of_ICD-9_codes
7. https://www.webmd.com/heart-disease/atherosclerosis-and-coronary-artery-disease#1