DSC630 – Predictive Analytics Project Proposal

Thip Rattanavilay

Bellevue University

**Introduction**

Blood donation is an integral and essential part of the healthcare system. Without blood banks, many of the medical procedures that we otherwise take for granted could not take place. The modern lifestyle, ever-increasing mobility and accompanying higher accident rates, and incidences of natural and human-made disasters (such as wars, earthquakes, etc.) have led to an ever-rising demand for blood transfusions. A constant supply of blood is needed to help ensure that hospitals have access to enough blood to meet their current and future needs. One of the most important factors for a stable supply is the retention of donors, as return donors allow blood banks to not spend additional efforts and resources looking for new ones.

**Background**

Since blood centers in most countries typically rely on volunteer donors to meet the hospitals' needs, donor retention is critical for blood banks. Identifying regular donors is critical for the advance planning of blood banks to guarantee a stable blood supply. In this research, donors' data was collected from a blood bank from 2017.

**Problem Statement**

Every day in the world’s poorest countries, people die because of a shortage of blood. Given our mission, we're interested in predicting if a blood donor will donate within a given time window. Use information about each donor's history Months since Last Donation: this is the number of months his since this donor's most recent donation. Number of Donations: this is the total number of donations that the donor has made. Total Volume Donated: this is the total amount of blood that the donor has donated in cubic centimeters. Months since First Donation: this is the number of months since the donor's first donation. Submission format This competition uses log loss as its evaluation metric, so the predictions you submit are the probability that a donor made a donation in 2017.

**Scope**

This project we're interested in predicting if a blood donor will donate within a given time window. Machine learning algorithms such as logistic regression (LG), random forest (RF) and support vector classifier (SVC) were applied to develop and evaluate models for classifying blood donors as return and nonreturn donors. The natural imbalance of the donors' distribution required extra attention and considerations to produce classifiers with good performance.

**Preliminary Requirements**

A blood bank dataset that includes sufficient variables or features is needed to proceed. Presence of both demographic and clinical variables is preferred, and a standardized process for hospital collecting and reporting of the data needs to be in place. We will assume that the dataset has been collected in the most methodical and comprehensive manner as possible because a missing entry in either the index hospitalization information, which may come from another data source, or the blood donor returning itself would cause a true blood donor returning to not show up in our data set.

**Technical Approach**

The dataset from the UCI Machine Learning Databases repository will be used. This dataset is comprised of 10 years of care at 130 US hospitals and integrated delivery networks between 1999 and 2008 (Beata, 2014). Over 50 features are present representing blood donor and hospital outcomes including, but not limited to, age, gender, race, admission type, discharge type, time in hospital, and whether the blood donor was returning or not. Records are indexed by unique encounter id and blood donors are indexed by a unique blood donor id. Encounters that are present in the dataset satisfied several criteria. First, the encounter had to be a hospital admission. Second, the encounter needed to have diabetes included in the diagnosis. Third, the length of stay had to be between one and 14 days. Finally, lab tests and medications had to have been given during the encounter. In all, over 100,000 records are included in the dataset. The target variable, which tracks whether the blood donor was returning, is a categorical variable including the values of no, greater than 30 days, or less than 30 days. Considering the last two variables as yes would make the target variable binary.

**Analysis**

Working through the dataset we will look for possible patterns for those blood donors that were re-admitted within the 30-day timeframe. We intend to look at age, gender, race, diagnosis type, and all the different types of medication to check for any possible relationships. We will need to put together tests that will look at each value and help us determine whether there is a relationship at a sample level, and if that passes, we can move forward to working with the larger data.

**Requirement Development**

The primary requirements for development of our classification model that predicts blood donation and is a dataset containing our target variable and possible predictor variables. From this we will obtain the features we believe will best train our model and predict our target variable. The variables in the original dataset will need to be evaluated for outliers, missing data, and correlation. Outliers will need to be handled as to not affect our model in unexpected ways. Missing data will need to be examined to see if there is too much missing data in a variable to have it be useful in predicting the target variable. Correlation between the variables and the target variable will be used to determine the best candidates for model features. Finally, candidate features will need to be cleaned in a way that makes them usable as inputs to the applicable model we will be fitting.

**Model Deployment, Testing, and Evaluation**

Since this is classification problem, model selection will be based on such. First, the data will be randomly split into three sets: training, testing, and validation. The training set will be used to train the models, the testing data set will be used to test the performance of the models, and the validation data set will be used to do a final test on the models after tuning model hyperparameters. Models applicable to classification problems that will be explored are K-nearest neighbors, logistic regression, Naïve Bayes, Decision tree, and Random forest.

Evaluation of the models will be completed by calculating key performance indicators such as accuracy, precision, recall, and f1 score from a confusion matrix. Since failing to predict a blood donation, or a type 2 error, is much worse for the hospital, insurance company, and blood donor than predicting a blood bank when there wouldn’t have been one, we will be focused on tuning our model to reduce type 2 errors specifically. This means we will be focusing on obtaining a model with high recall. Furthermore, since the prevalence of blood bank is much lower than not being returning in our dataset, we won’t have as much of an issue where our model could always predict one outcome and have positive results. If our model always predicted the most common outcome, not being returning, our accuracy would be high, but our focus of type 2 errors would also be high, and our recall would be 0.

Once our preliminary models have been evaluated, we will begin tuning our model hyperparameters in applicable models. For example, in our decision tree models, we can tune the hyper parameters to specify how deep we want to go in our tree. Then, for K-nearest neighbors, we can adjust the number of nearest neighbors to use to calculate probability for a given record’s result. We can continue to tune these parameters for each model until we have optimal results. Finally, once the models are tuned to the best of our ability, we can test generalization of the models with the validation dataset. This step will be critical to ensure our model is not overfit to the test set.

**Expected Results**

Blood donations occur 11.16% of the time in the dataset we have obtained. That means, a model that simply predicts the most common outcome, that the blood donor will not be returning within 30 days, would be 89% accurate. However, that would give us 0% recall which would not improve upon the business problem. On the other hand, if we were to always mark a record that the blood donor would be returning, we would have 100% recall, but only 11.16% accuracy, which also wouldn’t be optimal or useful since the hospital cannot keep as stock full of blood. Finally, if we were to randomly select a blood donor to donate, we would catch half of the blood donors that were to be returning. In other words, our model would have 50% recall. This model would also have 50% accuracy. Therefore, we would want results that had at least 50% recall while improving on the accuracy of 50%. Otherwise, it would be no better than simply picking a result at random.

**Execution and Management of Project**

**Project Plan**

Currently, we are in the late second, early third week of the term. We have found our main dataset and will start working with it by loading it into Python and then working to clean it up by looking for outliers, duplicates, missing information, and performing simple visualizations to identify potential patterns. These initial steps of cleaning the data may be able to split up between the three group members using a primary key in the dataset, encounter\_id, which is a unique identifier for each record. While one person is examining one group of variables, another person can be working on a separate group. If any records are deleted in one of the group members datasets, as a means to handle outliers for example, we would be able to keep the datasets in sync by bringing the datasets back together and merging on the common primary key.

Once the datasets are merged, decisions on the feature variables to be used in the modeling step can be made by examining correlations between the potential feature variables and the target variable. When the features are selected, we can create a new feature dataset that can be split up between train, test, and validation. These three data sets can then be copied so each of the three team members can work on model training and tuning separately. Finally, the results of all the models on the validation dataset, which was not used to tune the model, can be combined and compared. It is from this comparison that we can choose the optimal model for the business case.

Communication will be completed in the Teams group chat and Teams file for uploading the documents. Word documents will be worked on individually in a Teams shared drive. R and Python code files and notebooks will be worked on individually, and upload to Teams, or Blackboard. If needed, we will explore using a Github repository for code. Each week I will discuss the tasks for upcoming Milestones and work appropriately through my task, discussions on Teams.

**Project Risk**

One of the biggest risks with this project is that we either find no variable combination meets our threshold or too many variable combinations exceed it. The threshold can be adjusted lower if the data does not indicate a clear combination, but if too many meet, we may have to raise it or re-check the data.

A second potential risk is that the social determinants of health may not be the same between two people who look the same on paper. For example, a black man living in an urban setting and a black man living in a rural setting might look the same on paper, but the differentiating factors between the two, or the social determinants of health, are actually *income and* *educational level*, not race. So, their risk for blood donor returning would not necessarily be the same. Therefore, if race is used as a feature variable, it may just be a proxy variable for the true predictors of our target variable. In other words, if we are using proxy variables, we risk our models being less than optimal.

**References**

Beata Strack, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura,

Krzysztof J. Cios, and John N. Clore, “Impact of HbA1c Measurement on Hospital

Blood donor returning Rates: Analysis of 70,000 Clinical Database Blood donor Records,” BioMed Research International, vol. 2014, Article ID 781670, 11 pages, 2014.

Blood donationss Reduction Program (HRRP). (n.d.). Retrieved September 11, 2020, from

https://www.cms.gov/Medicare/Medicare-Fee-for-Service-

Payment/AcuteInblood donorPPS/Blood donor returnings-Reduction-Program

Leonhardt, M. (2019, November 04). Rising health-care costs stall Americans' dreams of buying

homes, building families and saving for retirement. Retrieved from

https://www.cnbc.com/2019/11/04/health-care-costs-are-preventing-many-americans-

from-hitting-life-milestones.html