**Analysis Plan for Data Analytics LACE Project**

**Introduction**

The final group project for the Data Analytics class requires analyzing a given dataset and predicting patient readmissions from LACE: Length of Stay, Acuity of the Admission, Comorbidity (Charlson comorbidity index score), and Emergency department use. Outcome of interest is a binary dependent variable: Readmit, which is calculated based on the individuals’ EHR admission dates. Each of the different independent variables are each individual LACE component.

The dataset provided is fabricated SQL database, putting no patient information at risk. Data have been collected organized with SQLite for R.

Barrett Campbell and Patty Langasek are the group members for our project.

**Known Information**

Length of Stay has been calculated from emergency department admission to discharge.

Acuity of Admission is a numerical score indicating admission to the emergency department for the individual.

Charlston Comorbidity index score is a summed score calculated based on diagnostic codes for the patient.

Emergency department use is aggregated admissions to the emergency department for the individual.

LACE is the summed score for all of the other variables. It’s a numeric value.

**Expected Data**

Length of Stay (Length\_of\_stay) will be a numeric value of whole numbers; 0 < LS < 30.

Acuity of Admission (Admission\_acuity) is a 2-level factor (0 or 3).

Charlston comorbidity score;

Emergency department use (ED\_visits\_score) is a factored numeric; 0 < ED\_visits\_score =< 4

LACE (LACE) is numeric summation of all of the other variables; 0 < LACE < ?? 100?

**Sensitivity/Specificity Tradeoffs**

False Positives will identify patients at risk for hospital readmission within 30 days of an emergency visit when they would otherwise not present. Consequences to the patient of falsely identifying risk include more testing and possible follow-up commitments. Consequences to the HCO include increased cost associated with increased monitoring of the patient, increased costs from increased tests, increased burden on health care providers.

False Negatives will fail to identify patients at risk for readmission within 30 days of an emergency visit. Consequences to the patient of not identifying at risk patients are increased emergency admissions, possible decreased quality of life, and increased risk of serious complications for not adequately treating diseases. Consequences to the HCO include the increased cost associated with readmissions, loss of insurance revenue, and loss of hospital prestige.

With the task of identifying patients at risk of readmission, and the more serious consequences of the need for emergency department readmission, we will be focusing on sensitivity as a higher priority over specificity.

**Plan**

First, a full EDA is in order to ensure the data behave and present as expected. We’ll look at each individual LACE component as well as the components needed to calculate specific scores (as necessary). We’ll evaluate the quantity of readmission factors, and the distributions of each of the component scores, look for missing or corrupted data, and ensure data quality.

Next, the data will need to be partitioned. The instructors indicate 2 partitions, however, we have been trained to partition into 3 to test overfitting/validation and get the prediction error. This is easily done by shifting the suggested numbers only slightly. Instead of an 80/20 split for training and testing, we can do a 70/15/15 split for training, validating, and testing respectively. 15% of the data will be initially preserved and untouched until we are done completely validating the model, and the predictions from that will be used to verify the validation errors. The remaining 85% of the data will be split into a training partition (70%) and a validating partition (15%). The validating partition will be isolated completely from the training partition and used to assess accuracy of the model.

*Building the models*

*Linear Model:*

Per project requirements, we are to use a linear model (glm()) after having selected 2 comorbidities in the ‘C’ portion of the LACE requirements. All comorbidities have been included in our data, increasing the likelihood of interactions complicating a simple linear model. As such, looking at just the comorbidities of patients that have been readmitted to evaluate for interactions would be crucial to ensure statistical confidence. Should some comorbidities prove to have strong interactions with length of stay or with each other, then that should be specifically factored into the model.

*Painful Model:*

The second part of the project requirement is to present a learning model to be able to predict readmissions, using any learning model our hearts desire. Because so many of the LACE variables are factorable (or are not continuous), this is a pretty easy case for a many-permutation decision tree algorithm (random forest). Examples given by the instructor were decision tree and neural network. Because of the few variables, a support vector machine (SVM) would also work well to classify/predict readmission risk (especially as it’s a binary classification). I think the comorbidities could be simplified once interactions and significance are accounted for, making SVM a decent option. In the interest of time and simplicity for explaining, I think a random forest is the best option, as we can also pull statistics about what accounts for the largest tree splits, giving us possibly actionable targets for readmission prevention in the future. Though, I may maverick my way through and also evaluate the results with SVM as a benchmark for both the linear model and Random Forest predictions.

**Presentation Deliverables**

We will need graphed results for a brief presentation for the class. We’ll use ggplot to show (1) the EDA readmission information, (2) an easy to understand plot for RF and SVN, and then (3) a prediction plot from the linear model as well as (4) the machine learning model that shows readmission predictions, and finally (5) how many patients (out of 1000) would be misclassified and how those misclassifications distribute (from a confusion matrix).

**Possible Considerations/Complications**

*SQL mistakes.* Unlikely as Barrett is highly skilled and comfortable with SQL and we’ll be working from his tables, but this would result in unreproducible predictions.

*Undiscovered interactions.* This would possibly indicate a stronger sense of importance on a variable than actually exists, which would show in bias training in our models.

*Difficult interactions.* We may come across very complex interactions that may be, for the sake of time, impossible to fully explore and account for, requiring us to reduce or reevaluate components in our overall models.