DS7333 Quantifying the World: Case Study 2

Rick Fontenot, Feby Thomas Cheruvathoor

1. **Introduction**

Discuss goals and importance of predictions

1. **Methods**

Data description and preparation

Insert summary of features available, size of dataset

Missing Values

Summarize features with missing values. Discuss imputation of each column, ie.e drop weight fill NA with “missing” for others and why we chose that

Duplicate Observations

Have not checked this yet, we should and discuss

Categorical Variables

Discuss category level reductions on each features and which were treated as ordinal like age

Standardizing data

In order to determine which features are most important to predicting the critical temperature, we scaled the data so that the regression weights can be compared. We scaled all features to a mean of zero scaled to unit variance. Scikit learn StandardScaler function has been used to scale the data points.

Target Variable

Summarize Imbalance in classes of target

Discuss using stratified splits for both training and validation

Discuss using SMOTE resampling

Cross Validation

In order to compare the models for overfitting, we first set aside 10% of the data into a validation set (shuffled data to get random observations). After our model hyperparameters are tuned on the training set, we will compare performance metrics of the model on this unseen data.

For the training data, we used 10-fold cross validation and the mean accuracy across the folds for model tuning.In addition to accuracy, we will also summarize the precision and recall for each class, for comparison.

1. **Results**

Discuss regularization explored but did not improve model.

Initial model showed dispositions for expired patients as most important. Removed those records and re-ran. Minimal loss in classification performance and better look at important features

1. **Conclusions**

Show summary table of metrics for training and validation to compare original model to re-sampled (SMOTE) model.

Important to discuss that while overall accuracy dropped with SMOTE resampling, the poor recall for Readmitted<30 days drastically improved with resampling

Most important feature ranking is very similar for original and resampled models

Discuss reasoning on why specific features may be most important

**Appendix**

1. **Code**

A rendered notebook containing code for this analysis can be accessed at: