Capstone 2: Milestone Report 2

Predicting Hospital Readmission Using Unstructured Clinical Notes

Initial Model

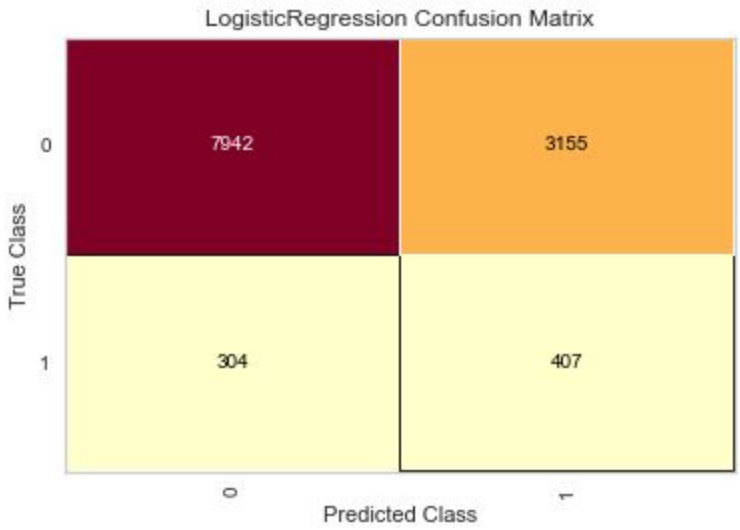
A bag-of-words representation and logistic regression will be used for the initial model predicting 30-day hospital readmission. Since this is an imbalanced dataset (with the positive class representing about 6% of the data), we will build models using random under sampling and support vector machine synthetic minority over-sampling technique (SVM-SMOTE). 33% of the training data was set aside for validation.

Random Under Sampling

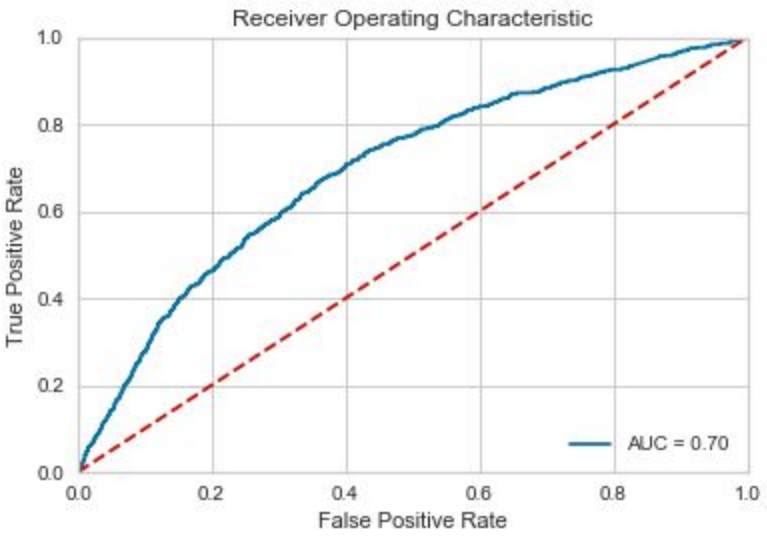
The CountVectorizer was initialized with the a maximum of 3000 features and standard English stopwords. The LogisticRegression was initialized with a regularization term (C) of 0.0001, a l2 penalty, and a lbfgs optimization algorithm. This produced the following results:

	precision	recall	specificity	f1	support
0	0.96	0.72	0.57	0.82	11097
1	0.11	0.57	0.72	0.19	711
avg / total	0.91	0.71	0.58	0.78	11808

A negative class specificity of 0.72 and precision of 0.96 indicates that our model is decent at predicting those cases where the patient will not be readmitted. However, a positive class recall of 0.57 shows that our model will only accurately predict a little over half of all readmission cases. A visual sense of the predictive power can be gained by examining the confusion matrix. It becomes obvious that our model is predicting a lot of false positives. This may not be bad,



however, as false positives in this case would likely mean improved care for patients who may not need it. In order to maximize hospital resources, however, future models should attempt to reduce the number of false positives and decrease the number of false negatives. In other words, future models will attempt to improve the ROC-AUC, which equals 0.70 with the current model.

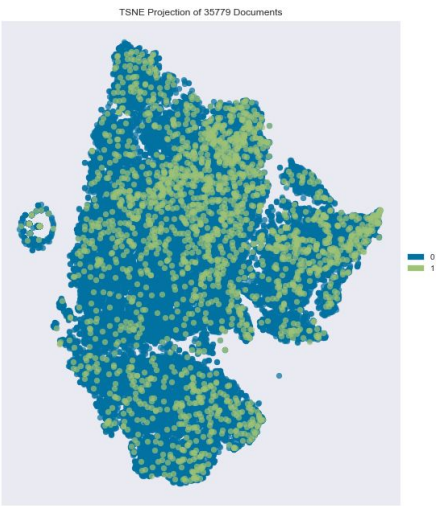


Over-sampling with SVM-SMOTE

The CountVectorizer was instantiated with the same parameters as the under-sampled model, but the maximum number of iterations was increased to 500 in order to allow for convergence. This produced the following results:

	precision	recall	specificity	f1	support
0	0.94	0.96	0.12	0.95	11097
1	0.16	0.12	0.96	0.14	711
avg / total	0.90	0.91	0.17	0.90	11808

Although SVM-SMOTE appears to improve the average recall score significantly, it appears that this is artificially inflated by biasing the negative class. This is what we want to avoid in an imbalanced class set since predicting no hospital readmissions would provide us with a model that is 94% accurate, but 0% accurate in achieving its goal of predicting readmission. This model is only able to accurately predict about 12% of hospital readmissions. This implies that for this particular problem, we may be better off with undersampling the negative class. The tSNE plot produced in exploration may have predicted this as the two classes were not apparently grouped when we reduced their dimensionality.



Examining the confusion matrix, we can see that model is largely predicting the negative class, generating a lot of false negatives and few false positives. While this may improve the model's precision, when predicting hospital readmission, it is probably better to sacrifice precision and accuracy for recall. Due to the high number of classification errors, this model produces an ROC-AUC score of 0.62.