Classification

1.General Approach

In this section, we will attempt to construct an ensemble of models that will help us determine whether an assessment will pass an external audit check (i.e., whether it will be flagged for review or assigned to a clinical expert for follow-up analysis). As in the previous section, we first expand on our assumptions about the predictors we assess, how we split the data, and the quantitative performance metrics we prioritize.

With respect to predictors, we focused on the country in which the assessment was conducted, the patient's membership in the treatment or control group, the date of the visit, and the total PANSS score on that visit. For some methods, country cannot be used as a predictor because in Study E, there were some assessments in which the United Kingdom was the country of assessment, while in other studies this country was not available. Some statistical learning techniques (e.g., logistic regression) cannot accommodate previously unknown (i.e., not observed in the training set) classification values; other methods (e.g., Naive Bayes) do not have this limitation.

Using the following approach, we divided the data into a training set, a development set, and a test set. We initially selected study E as the test set, i.e., the dataset for which we will make predictions on the LeadStatus variable. We randomly selected 75% of the remaining observations (consisting of data from studies A-D) as the training set (used to create our model) and 25% as a separate development set (used to evaluate the performance of different models). In particular, when examining the performance of a model on the development (dev) set, we will focus on the area under the curve (AUC) of the ROC curve and on measures of Cross-entropy (or log loss), as these do not depend on the probability threshold chosen to assess "pass" or "fail". We feel this is the best approach because we are ultimately responsible for estimating the likelihood that an evaluation will be emphasized, not the binary outcome ("pass" or "fail") itself.

2. Naïve Bayes

Similar to the naive predictions in the forecasting section, we will now consider what is the most straightforward classification method. The first approach is naive Bayesian classifier. The naive Bayes classifier mainly calculates the probability of a response via Bayes' theorem, provided that the predictor variables are conditionally independent of each other (thus greatly facilitating the calculation). The method is called "naive" because this assumption almost certainly does not hold for some subset of predictor variables in the dataset. Nevertheless, the naive Bayesian classifier can serve as a useful "baseline" prediction by which we compare the performance of our other models. We see from Figure 15 that the naive Bayes classifier performs well on the dataset, with an AUC and log loss of 0.7698 and 0.4722, respectively. Note that this performance is not reflected in the test set (public leaderboard), where the log loss is 0.70123. This returns to the observation at the end of Section 3, where we note that Study E differs markedly from the patients in the other studies. at least in terms of their initial visit scores. We believe we see this bias here because the prior in the naive Bayesian classifier was computed exclusively based on the patients in Studies A-D.

图表, 折线图

描述已自动生成

Figure 15: The ROC curve and associated AUC and log loss for the naïve Bayes classifier (measured using the development set)

3. Logistic Regression

We first evaluate a simple logistic regression on treatment group, visit day, and overall PANSS score (for the reasons described in the previous subsection). Figure 16 depicts the resultant ROC curve, in which we specify the development AUC and log loss as 0.6149 and 0.449, respectively. Note that this model's log loss on the test set (the public Kaggle leaderboard) is 0.61847 after being trained on the whole of the data from experiments A-D. While this is unquestionably an improvement over the Naive Bayes test error, we tried to further reduce it by including all of the individual PANSS values into the model.

图表, 折线图

描述已自动生成

Figure 16: The ROC curve and associated AUC and log loss for logistic regression using only the total PANSS score (Development set)

When evaluating each individual PANSS score (and eliminating the entire PANSS score), we find an AUC and log loss of 0.6770 and 0.5224 respectively (Fig. 17). The test set log loss matching to the public Kaggle leaderboard is 0.67744. This indicates the possibility of overfitting, since the apparent progress on the training set disappears when tested on the test set. Even though none of the development set was utilized for training, our model has a small degree of bias due to the fact that the training observations and development set observations only come from studies A to D, and the test set only includes patients from study E. Consequently, our model contains a source of bias that we cannot directly address.

The next move in our analysis was to consider the logistic regression model that takes into account all of the individual PANSS scores, but to do feature selection by lasso to finally reduce the model variance (which should improve the test set score). In Figure 18, the ideal value of, the shrinkage value in lasso, is determined using cross-validation (10-folds). The optimal value of, according to the one standard error criterion, is so that only 24 predictors are evaluated. We omit P1, P2, P6, G1, G6, G7, and G11 as predictors. Figure 19 shows the ROC curve for logistic regression with lasso and this specific value of = 0.00298 along with the AUC and log loss on the dev set. While log loss of 0.5238 is bigger than that observed in Fig. 17, the log loss for the test set was actually 0.66589, which is an improvement over the model that contains all of the individual PANSS scores. In this instance, the increase in bias caused by removing the variables stated above from our regression model resulted in a variance reduction sufficient to reduce the overall test error rate. It should be noted, however, that the original logistic regression model beats the lasso-based model by a significant margin (test set log loss of 0.61847 vs. 0.66589). This shows that limiting variation should be our priority if we intend to further improve our test set log loss.

图表, 散点图

描述已自动生成

Figure 17: The ROC curve and associated AUC and log loss for logistic regression using all of the individual PANSS scores (Development set)

图表, 箱线图

描述已自动生成Figure 18: Determining the optimal shrinkage value for λ lasso in the context of our logistic regression on all individual PANSS scores.

图表, 折线图

描述已自动生成Figure 19: The ROC curve and associated AUC and log loss for logistic regression utilizing lasso (Development set)

4. Linear Discriminant Analysis

Next, we evaluated linear discriminant analysis as a classifier (LDA). Considering just treatment group, visit day, and PANSS total score as variables results in a model with little variation (which, given the findings of the previous section, seems to be the best way to minimize our test error rate.) Not only is LDA effective for reducing our optimum test set error, but it may also assist us in comprehending the nature of the underlying data. This is accomplished by comparing the performance of LDA to that of other techniques and inferring the resulting influence on the dataset under the assumptions of multiple models. Figure 20 depicts the ROC curve for LDA, where the AUC and log loss on the design set are 0.6136 and 0.5454, respectively. Training this model on the whole data from Study A-D provides a log loss of 0.6187 for the test set. Consequently, LDA performs almost equally to our logistic regression using simply the complete PANSS score. Logistic regression and linear discriminant analysis (LDA) both create a linear decision boundary; provided the underlying Bayes decision boundary is also linear, we anticipate both approaches to perform quite well. Given that LDA assumes the observations in both classes are drawn from a Gaussian distribution, the equivalence in performance between the two approaches indicates that this is a reasonable assumption for this data set.图表, 折线图

描述已自动生成Figure 20: The ROC curve and associated AUC and log loss for linear discriminant analysis (Development set)

5. Quadratic Discriminant Analysis

Consideration now given to quadratic discriminant analysis (QDA). If the decision boundary is somewhat nonlinear, we anticipate that QDA will perform better than LDA. Again, just the three predictions of treatment group, visit day, and PANSS total score are considered. The AUC and log loss for the dev set are shown in Figure 21 as 0.6568 and 0.5560, respectively. The test set's log loss (obtained from the Kaggle leaderboard) is 0.63167. Consequently, QDA does not perform badly (relative to the majority of previously studied approaches), but it is marginally inferior to LDA and our first logistic regression. Recall that QDA is distinct from LDA in that each class now has its own covariance matrix - we no longer assume they are comparable. The decrease in bias, however, is accompanied with an increase in variance. Clearly, this trade-off results in a net drop in our classifier's performance, which makes sense given that the best classifiers to until have been rather rigid.图表, 折线图

描述已自动生成Figure 21: The ROC curve and associated AUC and log loss for quadratic discriminant analysis (Development set)

6. The gradient boosting method and random forests

Gradient boosting method (GBM) and random forests were the final two models we investigated. Both approaches seek to directly address the variance issue in terms of the bias-variance tradeoff, thus we anticipate that they will both function well with this dataset. GBM is accomplished by considering a collection of "weak" learners, which are often short trees (commonly stumps) that only permit fitting residuals at a slower pace. Random forests, on the other hand, reduce model variance by only examining a subset of the entire number of predictors per split in a given tree for a particular set of learners.

Figure 22 illustrates the effectiveness of the gradient boosting method (using XGBoost modified for random discrete grid search) when predictors including treatment group, visit day, and total PANSS score are used. Notably, the development set has an AUC of 0.7132 and a log-loss value of 0.5075, which is fairly competitive compared to the other classifiers evaluated in this section. In spite of this, the log loss of the test set was determined to be 0.67320, exceeding only logistic regression and naive Bayes classifiers based on all individual PANSS values. At this time, our team noticed that we may not be using the GBM's potential since it only utilizes three predictors. While previously introducing a single PANSS score decreased test set performance (due to greater model variance), we predict that GBM will not experience similar setback because to its concentration on decreasing variance. Figure 23 depicts the ROC curves, related AUC, and log loss for GBM utilizing all PANSS values individually. The model obtained an AUC of 0.8194 and a log loss of 0.4201, indicating that the development set has increased significantly. However, as previously, this is not indicative of the test set's performance. This model's test set log loss was determined to be 0.73258. By analyzing the relative significance of the different predictors in the modified GBM (Figure 24), we may acquire insight into the reasons behind this phenomenon. Note that the relative relevance, as determined by the h2o.varimp\_plot() function, largely takes into account the degree to which each predictor decreases the MSE at each step and the frequency with which a predictor is utilized to produce a split.图表, 折线图

描述已自动生成

Figure 22: The ROC curve and associated AUC and log loss for the gradient boosting method using the total PANSS score (Development set)

图表, 折线图

描述已自动生成

Figure 23: The ROC curve and associated AUC and log loss for the gradient boosting method using all individual PANSS scores (Development set).

图表, 漏斗图

描述已自动生成Figure 24: The relative variable importance for the gradient boosting method using all individuals PANSS score.

Predictors are used to determine the frequency of tree splits. The majority of PANSS scores fall within the positive and general symptom categories, as seen in the graph. In reality, just one of the first 10 factors is linked to unfavorable symptoms (G15). Recalling Section 3, patients in Study E scored considerably differently than other patients on the first principal component, which was mostly associated to positive and general symptoms. On the test set, this bias was again demonstrated as a decline in GBM performance.

The random forest approach was the last method we investigated. On the development set, even a little overshooting time (again using random discrete grid search) is noteworthy (Figure 25). We observe the optimal AUC and log loss for the development set to be 0.8353 and 0.4108, respectively. Nonetheless, the performance on the test set is unimpressive, with a log loss value of 0.70370. We feel this pertains to the same issue as the gradient boosting approaches outlined before. For these methods to perform equally well on the test set, sophisticated methods for incorporating information from study E into the training set or manually adjusting the training set to more closely resemble the test set would need to be developed.图表, 折线图

描述已自动生成

Figure 25: The ROC curve and associated AUC and log loss for the random forest method using all individual PANSS scores (Development set).

7. Discussion

The log loss values for the test set varied from 0.61847 (best result on the Kaggle public leaderboard) to 0.7332. In general, it seems that the least flexible approaches, such as logistic regression and LDA, fared the best. We ascribe this remarkable performance to the fact that patients in Study E scored considerably differently than patients in previous trials on the positive and general PANSS categories (as shown in Section 3). Since Study E patients were not included in the training set for the classification problem, we anticipated that the differences would provide a challenge for making predictions on the test set (i.e., all of Study E).

Attempting to tackle this issue using support vector machines (SVMs) might provide more insight into why LDA and logistic regression seem to be superior to other approaches. In a preliminary examination of the use of SVMs to this dataset, we plotted observations from both categories on a plane spanned by visit day and PANSS total scores (Figure 26). This chart explains why the SVM did so badly overall (and why it is not covered in this report): it illustrates why the SVM fared so poorly overall. This graphic clearly demonstrates that the data (at least when seen on this plane) are not segregated in any way. Therefore, we anticipate that the SVM will struggle to identify a correct hyperplane; all the "X" in the picture are support vectors (and the number of support vectors seems to be about the same as the total number of data points). In contrast, we discover that although logistic regression performs poorly with well-separated categories (i.e., its parameter values are unstable), it has no issues with continuous data. Even though the Bayesian error rate would be higher in this instance, logistic regression (and the closely related LDA) may still do well.图表, 散点图

描述已自动生成

Figure 26: Observations visualized in the plane. Color denotes the class identity (whether the audit was passed or not) and "X"s denote support vectors for a linear support vector machine.