Building a Student Intervention System

Classification vs. Regression

The student intervention system is fundamentally a classification problem. We are attempting to identify at risk students and classify them as such in order to begin an intervention process. For this problem either the student is at risk or is not. We are not attempting to predict specific outcomes for each student, merely whether or not they are in need of intervention.

Exploring the Data

Some basic statistics of our dataset

- Total number of students: 395
- Number of students who passed: 265
- Number of students who failed: 130
- Number of features: 31
- Graduation rate of the class: 0.67%

Training and Evaluating Models

The three models tested, in order, are Logistic Regression, SVM, and K Nearest Neighbor.

- **Logistic Regression**

  The general applications for Logistic Regression are binary classifications based on the probability of an outcome. In this instance < .5 we would assign the label of "at risk" and > .5 we would assign the label "not at risk".

  Logistic regression is fast and good at defining two different classes (again the classification occurs based on the probability of an outcome.) Furthermore, logistic regression is easy to implement.

  Logistic regression is prone to overfitting if it is not properly regularized, choosing the proper regularization value can be tricky (however, most libraries that implement logistic regression automate much of this process.)

  Since logistic regression is very effective in making binary classifications based on the probability of the outcome it is an excellent model to use in order to classify a student as "at risk" and in need of intervention.

Table:

	Training Set Size		
	100	200	300
Training Time (secs)	0.002	0.003	0.004
Prediction Time (secs)	0.000	0.001	0.000
F1 Score Training Set	0.8740	0.86689	0.843267
F1 Score Test Set	0.78740	0.80315	0.794118

- **SVM**

The general applications of the SVM model are the binary classification of labeled data.

A strength of SVMs is that SVMs focus on correctly classifying data before any other optimizations. This allows SVMs to generalize fairly well.

SVMs, however, ignore outliers, which might be important when it comes to a student intervention system (and might be a weakness in this instance.) Furthermore, SVMs are complex and this complexity has an impact on performance. The complexity of SVMs can lead to issues when attempting to scale SVMs to larger datasets & if there are resource constraints (such as the constraints in computing costs for the intervention system.) As seen in the table below, SVMs also take more time to train and test (*notice the growth of the training time as the dataset increases in size in the table below.*) Two other potential weakness of SVMs are, 1) the dataset has to be linearly separable (although there are techniques that can be used to resolve this issue) and 2) the proper use of a kernel function. Kernel functions are a sufficiently complex topic that they will not be broached in this document.

This model was chosen because it prioritizes correct classification above all else. Given the desire to correctly classify students in need of intervention, the SVM model seemed to be a viable candidate.

	Training Set Size

	100	200	300
Training Time (secs)	0.002	0.003	0.006
Prediction Time (secs)	0.001	0.001	0.002
F1 Score Training Set	0.84768	0.8690	0.87212
F1 Score Test Set	0.76056	0.76712	0.73611

- **K Nearest Neighbors**

  The general applications for K Nearest Neighbor (KNN) are classification and regression.

  A strength of KNN is that the model scales well with larger datasets. KNN is also a simple algorithm, which can put less demand on resources.

  Picking k in KNN can be tricky. Using the correct k (or the number of nearest data points to compare to) can be a somewhat difficult process and using the incorrect k can lead to poor predictions. KNN also assumes that near points are similar; this might not be necessarily true and may lead to incorrect classifications. KNN also requires using the correct distance function for a given problem, not all distance functions will work well in various domains and choosing the wrong one can a have negative impact on classification.

  This model was chosen in order to see whether a simple classification model was effective enough to on the data. Given the constraints, using a simple model would be highly advantageous when scaling to larger datasets.

	Training Set Size		
	100	200	300
Training Time (secs)	0.001	0.001	0.001
Prediction Time (secs)	0.002	0.002	0.005
F1 Score Training Set	0.78049	0.81595	0.824268
F1 Score Test Set	0.76510	0.741259	0.732394

Best Model

The best model given the constraints is Logistic Regression. The chosen model has the highest F1 score while simultaneously having the lowest training/prediction times.

As briefly discussed above in the section “Training and Evaluating Models”, logistic regression makes binary classifications based on the probability of an outcome. What does this mean? Simply, the logistic regression model will take an arbitrary number of inputs and give the probability of an outcome. Logistic regression will then use the probability score to determine the appropriate label (“yes” or “no”, 0 or 1, “at risk” or “not at risk”, etc.)

During the training process, logistic regression will compare the predicted label against the actual label from the “training” data (or the subset of the dataset that we use to teach, or train, the model.) The model does this in order to find the best predictions. So logistic regression will iterate through the dataset constantly comparing it’s predictions with the actual label and make adjustments until the best model is found (the model with the best predictions accuracy.)

After the training process is complete we test the best logistic regression model we found against a test set that was removed from the dataset before training. The purpose of this is to make sure that we haven’t over fit our data (more simply, we haven’t fixed the outcomes to agree with the real labels.) If the test set produces a reasonable accuracy score we can begin the process of introducing new data and using the best logistic regression model to make real world predictions.