Student Intervention System

Classification vs Regression

Each high school student has to pass a final exam in order to graduate. The goal of this project is to help predict which students are having trouble and intervene to try to correct the problem. A dataset has been provided to train a model with. This will be a classification model not a regression model. It is a classification problem because there are two distinct groups. Since students either pass high school or they do not, there is no middle ground, thus making it a classification problem.

Exploring the Data

Some basic information about the dataset provided is shown below in Table 1.

Total number of students	395
Number of students who passed	265
Number of students who failed	130
Graduation rate of the class	67%
Number of features	31

Table 1 - Basic Dataset Information

Preparing the Data

The code had to be cleaned up and organized before giving it to the algorithm to create a model. The first columns were the features and the last column was the target that the model is trying to predict. The features and description are listed in appendix A. Some of the columns were categories or factors and not numerical values. The algorithms that were going to be used do not handle these well so these columns were encoded to numerical values. Finally, the dataset was split into a randomized training and testing set. 75% of the data (300 samples) was used to train the model and the remaining portion was used to test or validate the model on unseen data (95 samples).

Training and Evaluating Models

Three classification models were chosen to train on the dataset. These models were a gradient boosting classifier, a random forest, and a support vector machine (SVM). These were all available from from the scikit-learn.

The gradient boosting classifier and the random forest models was chosen to explore because it has a good track record from Kaggle.com competitions as a general classifier. This track record is a very good pro for both of these algorithms and a great place to start. These will be used as a benchmark. A con for both of the algorithms is the this dataset is a little small. These would perform better with larger datasets. Another con for both of these is that they can be computationally intensive and take a while to run.

The SVM was chosen because it is good at generalizing well (if tuned correctly) and be able to find the best way to separate the features. The SVM also requires little time to compute the model. It runs quickly and many models can be tried in a short amount of time. One con of the SVM model is that different kernels might have to be tried to get the best fit. This is normally just done by trial and error to see which kernel works the best.

All three of these models were tried with various training set sizes to see how this affected the results. Table 2, 3, and 4 show the results of training these models. These tables show the time that it took to train the model and the time need to predict with various training set sizes. The F1 score is also shown for the training and test sets. The training sets are not a good predictor for the performance of the model but the unseen test set is more representative of the actual performance.

	Training Set Size		
	100	**200**	**300**
Training Time (sec)	0.101	0.114	0.169
Prediction Time (sec)	0.002	0.002	0.005
F1 Score for Training Set	1.000	1.000	1.000
F1 Score for Test Set	0.693	0.752	0.785

Table 2 - Gradient Boosting Classifier Results

	Training Set Size		
	100	**200**	**300**
Training Time (sec)	3.682	3.889	4.100
Prediction Time (sec)	0.259	0.293	0.322
F1 Score for Training Set	1.000	1.000	1.000
F1 Score for Test Set	0.775	0.781	0.781

Table 3 - Random Forest Results

	Training Set Size		
	100	**200**	**300**
Training Time (sec)	0.006	0.007	0.014
Prediction Time (sec)	0.002	0.002	0.007
F1 Score for Training Set	0.870	0.847	0.843
F1 Score for Test Set	0.746	0.770	0.794

Table 4 - SVM Results

Choosing the Best Model

After splitting the data into a training set and a test set, three models were trained on the training set. The three models used were a gradient boosting model, random forest, and a support vector machine (SVM). These model's performance was determined based on the how it performed at classifying the test set. The SVM outperformed the other two models on classifying the test set, training time, and prediction time. For these reason it is recommended to move forward with the SVM model with an F1 score 0.794.

The SVM runs very quickly had a limited resource requirements. It can be run several times a year to update any databases of the students with little overhead. The SVM trained and predicted with similar or quicker times than the other models.

With a two dimensional problem the SVM is a model which tries to draw a curve between the different features to separate the outcomes. This is a multidimensional problem so the SVM is going to try to make a surface, instead of a single curve, between all of those dimensions that best separates the students that graduated and those that did not. The best surface or curve is the one that maximizes the distance between the different points of that feature with the different outcomes.

This model will be fined tuned to try get the best performance out. This will be done by using a grid search to try several different input parameters. The final model had a F1 score of 0.80. This is the model that will be used for the actual students.

Appendix A - Dataset Column Discriptions

- **school** - student's school (binary: "GP" or "MS")
- **sex** - student's sex (binary: "F" - female or "M" - male)
- **age** - student's age (numeric: from 15 to 22)
- **address** - student's home address type (binary: "U" - urban or "R" - rural)
- **famsize** - family size (binary: "LE3" - less or equal to 3 or "GT3" - greater than 3)
- **Pstatus** - parent's cohabitation status (binary: "T" - living together or "A" - apart)
- **Medu** - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
- **Fedu** - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
- **Mjob** - mother's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other")
- **Fjob** - father's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other")
- **reason** - reason to choose this school (nominal: close to "home", school "reputation", "course" preference or "other")
- **guardian** - student's guardian (nominal: "mother", "father" or "other")
- **traveltime** - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
- **studytime** - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
- **failures** - number of past class failures (numeric: n if 1<=n<3, else 4)
- **schoolsup** - extra educational support (binary: yes or no)
- **famsup** - family educational support (binary: yes or no)
- **paid** - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
- **activities** - extra-curricular activities (binary: yes or no)
- **nursery** - attended nursery school (binary: yes or no)
- **higher** - wants to take higher education (binary: yes or no)
- **internet** - Internet access at home (binary: yes or no)
- **romantic** - with a romantic relationship (binary: yes or no)
- **famrel** - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
- **freetime** - free time after school (numeric: from 1 - very low to 5 - very high)
- **goout** - going out with friends (numeric: from 1 - very low to 5 - very high)
- **Dalc** - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
- **Walc** - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
- **health** - current health status (numeric: from 1 - very bad to 5 - very good)
- **absences** - number of school absences (numeric: from 0 to 93)
- **passed** - did the student pass the final exam (binary: yes or no)