Predicting Academic Success Using Multiple Classifier Models

Amit Parikh
University of North Carolina at Chapel Hill
asparikh@live.unc.edu

Aryaman Agrawal
University of North Carolina at Chapel Hill
aryamana@live.unc.edu

Ernest Ermongkonchai
University of North Carolina at Chapel Hill
erneste@email.unc.edu

Kaan Nymaan
University of North Carolina at Chapel Hill
nymanka@email.unc.edu

Luke Schmidt
University of North Carolina at Chapel Hill
lukeant@ad.unc.edu

December 9, 2022

Abstract

Classification is one of the fields at the forefront of machine learning research. Our daily lives already use advanced data processing models, whether for identifying spam emails or classifying patient diagnoses. This project utilizes a dataset of students of diverse backgrounds containing both qualitative and quantitative aspects. In this paper, we construct and compare the accuracy to classify students falling above or below average in grades with logistic regression, random forest classifier, and KNN classifier models.

I. Introduction

i. The Data

The data set used for our research comes from Kaggle and contains information on 650 students in regards to their personal lives and their grades in their Portuguese class. For our model we decided to use a Feature Selection algorithm and ended up using the following features in our model: 'School', 'Sex', 'Address', 'Mother's Education', 'Father's Education', 'Mother's Job', 'Reason for Choosing School', 'Travel Time to School', 'Study Time', 'Number of Failed Classes', 'Amount of Higher Education Classes', 'Internet Access', 'Schoolday Alcohol Consumption', 'Weekend Alcohol Consumption, and Absences'.

The data set was created by Paulo Cortez in the UCI Machine Learning Repository. Each student is binarily classified as an above-average student or a below-average student.

ii. Implications

Our research can be applied to many different school districts and classes across Portugal, providing insight into the feature importance in Portuguese culture. Teachers in these districts can use this model to identify students who may be at risk of under-performance, allowing them to proactively provide the necessary support.

II. Framework

i. Preparing and Exploring the Data

The general framework of steps that we took are listed below, separated under three primary categories: Preparing & Exploring the Data, Building the Classifier Models, and Evaluating Model Results.

1. We downloaded the Portuguese.csv file from Kaggle and read it into our notebook
2. Calculated the median final grade among students in our dataset
3. Used this median to create a binary target variable representing if a student earns an above or below average grade
4. Encoded all categorical features as numeric using one-hot encoding
5. Selected the top 15 features by their ANOVA F-value, eliminating features that were not relevant to the model
6. Visualized the distributions and covariance matrix of our selected features to ensure the input data was statistically sound and had low multi-collinearity
7. Split the data into training and testing sets using a 75/25 split. It is industry standard, and is a good split to avoid over-fitting

ii. Building the Classifier Models

These general steps were applied to each of the three classification models we compared:

1. Create classification model object (Logistic Regression, Random Forest, or KNN)
2. Build a grid of hyper-parameters and tune these during Cross Validation
3. Fit the model using the X and y training data

iii. Evaluating Model Results

These general steps were applied to each of the three classification models we compared:

1. Make predictions using each of the trained models
2. Calculate accuracy of predictions using the Jaccard similarity index for in and out of sample data

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

3. Model predictions into a confusion matrix to better understand the strengths and weaknesses of our model's predictive power
4. Find feature importance of our finalized model to analyze and interpret the meaning from our model

III. Building the Model

Once our Kaggle dataset is loaded, we utilize several classifier models to predict students' performance in the classroom. We will test three different models in order to compare the accuracy of each and find a better solution for our prediction.

After training each of our models with the full dataset containing 30 features (including birth, sex, housing, etc.), we realized that this would cause our final prediction to overfit our training data. This is because we saw that some variables were highly correlated with each other from calculating a covariance matrix with all 30 features. With this information at hand, we were able to use a feature selection algorithm that removed half of the unneeded correlated features. This was done by taking the top 15 features by their ANOVA F-value. Making this adjustment significantly improved our prediction accuracy as indicated by our confusion matrix.

i. Logistic Regression

The purpose of using the logistic regression model is to have a baseline classification model which is simple to implement and an effective way to check whether other models meet the performance and accuracy expectations.

A logistic model takes the multiple features and uses a sigmoid function to squash the results into the [0, 1] range. This is an example of binary logistic regression which will be able to predict an above-average grade (1) and a below-average grade (0), where the parameters are:

$$p(y|\mathbf{x}, \beta_0, \beta) = \frac{1}{1+exp\{-y(\beta_0+\mathbf{x}^T\beta)\}}$$

From cross validation we found that the optimal hyperparameters are C = 1 and a penalty = none. C tells the model how much weight to give to the training data and the penalty hurts models with more variables if present.

ii. Random Forrest Classifier

Random Forest is one of the most popular and powerful classification algorithms because of its similarity to human-like thinking which leads to high intuition and interpretability.

A Random Forest classifier is an ensembling technique consisting of many decision trees. It builds an independent set of trees which classify inputs by voting on outcomes - an approach that is more accurate than individual trees.

Cross validation yielded an optimal max depth of 3 and 166 for n_estimators. The max depth represents the number of splits the model can make. N_estimators is the number of trees to build before averaging the predictions.

iii. KNN Classifier

The purpose of using the KNN model, just like the Logistic Regression, is for its simplicity and accuracy and is an effective way to check the other models.

A KNN classifier is a supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of individual data points. This model works off the assumption that similar points are in close proximity to each other.

$$D(X, Y) = (\sum_{i=1}^{n} |x_i - y_i|^p)^{\frac{1}{p}}$$

Using the Minkowski distance formula used above for the data points and cross validation, the optimal number of nearest neighbors to use was found to be 14.

IV. Results

The random forest classifier was found to be the most accurate machine learning model we trained, achieving an accuracy rate of 74.8% on our validation set and 76.5% on our training set. 74.8% was the highest out-of-sample accuracy rate in comparison to other models such as logistic regression (72.3%), and k nearest neighbors (68.7%).

The results of the confusion matrix showed that while the random forest classifier was able to accurately measure and predict positive cases, it was not as successful in accurately predicting and measuring negative cases. This suggests that further fine-tuning and optimization of the model may be necessary in order to improve its accuracy in negative cases as the false positives is as high as 35.

The most important factor in predicting student performance, as determined by the mean decrease in impurity (MDI) method, was found to be failing a prior class. This was followed by seeking higher education, mother's education and job, school, weekly

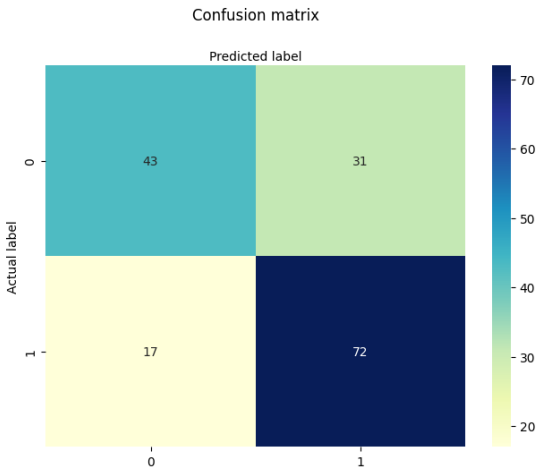

Figure 1: *Confusion matrix for Random Forest Model*

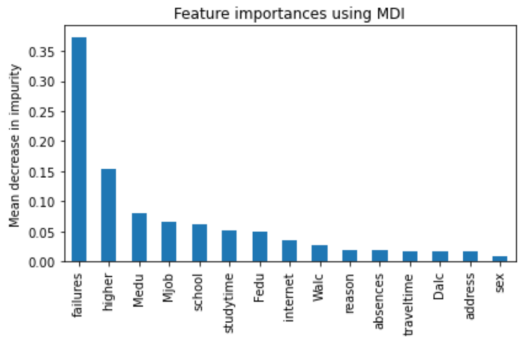

Figure 2: *Feature Importances Using MDI*

study time, father's education, and internet access, among other factors. Collectively, these were judged to be the most influential components of student performance.

V. Conclusion

Even though our results were fairly accurate, it could be significantly improved through the use of a larger data set or a more robust set of features. While our general framework could be applied to other schools and countries, we cannot use the same set of features for just any country or subject. This is because academic performance is also reliant on factors such as culture and the importance placed on academics in other countries. Our study can be enhanced by using the same framework in multiple countries and schools to gain a better understanding of what factors influence academic success.

References

[1] LynnSpeiserPersonEnvelope, Jaime, and Michael E.MillerEnvelope. "A Comparison of Random Forest Variable Selection Methods for Classification Prediction Modeling." Expert Systems with Applications, Pergamon, 23 May 2019, https://www.sciencedirect.com

[2] Rao, Medabalimi. "Texture Classification Using Minkowski Distance Measure-Based Clustering for Feature Selection.", https://www.spiedigitallibrary.org/journal-of-electronic-imaging. Accessed 9 Dec. 2022.

[3] "Alcohol Effects On Study | Kaggle." Kaggle: Your Machine Learning and Data Science Community, https://www.kaggle.com/datasets/whenamancodes/alcohol-effects-on-study Accessed 9 Dec. 2022.