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 7, 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 about 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 they went to, Sex, Address, Mothers education, Fathers education, Mother's job, Reason for choosing this school, Travel time to school, Study time, Number of failed classes, Amount of higher education classes, Time on internet, School-day alcohol consumption, Weekend alcohol consumption, and Absences. The data set was created by Paulo Cortez in the UCI Machine Learning Repository. Each student can only be classified as an above-average student or a below-average student.

ii. Implications

Since our data has simple information about the students' lives in Portugal, our research can be applied to many different school districts and classes across Portugal. We can especially use our research to find feature importance according to Portuguese culture. Teachers in these school districts could use this model to keep an eye out for students who are potentially going to be below average.

II. Methods

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 overfitting

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
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

ii. Random Forrest Classifier

iii. KNN Classifier

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. 748% 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 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

Our study utilized a confined and niche dataset, but the algorithm and results are widely applicable. While our general framework could be applied to other scenarios, 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 importance placed on academics in other countries. Our study can be enhanced by coming up with new feature sets and using the results to compare and contrast different cultures and the emphasis they place on academics. This allows us to gain a better understanding of how academic importance varies throughout the world.

Table 1: *Example table*

Name		
First name	Last Name	Grade
John	Doe	7.5
Richard	Miles	2

References

[Figueredo and Wolf, 2009] Figueredo, A. J. and Wolf, P. S. A. (2009). Assortative pairing and life history strategy - a cross-cultural study. *Human Nature*, 20:317–330.

VI. Results

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

$$e = mc^2 \tag{1}$$

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem.

Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

VII. Discussion

i. Subsection One

A statement requiring citation [Figueredo and Wolf, 2009]. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

ii. Subsection Two

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.