Predicting Chronic Absenteeism in Secondary Schools

Shunkai Ding[a], Dominic Matriccino[a]

[a]*Department of Statistics; University of Virginia,*

Abstract

The project investigates the relationship between social and demographic factors and chronic absenteeism to inform school policies and programs aimed at improving educational outcomes for underprivileged students. The project explores multiple resampling methods and data mining models, including random forest and penalized logistic regression. The top-performing model is a penalized logistic regression model trained on an upsampled training dataset. The most important features to predict chronic absenteeism are grades, schools, paying for extra tutoring, travel time, and study time. Therefore, to encourage attendance and active participation, schools could provide additional support for students who face difficulties coming to school or paying for academic materials.

Keywords: Classification, Resampling Methods, Random Forest, Penalized Logistic Regression

1. Introduction

1.1. General Background

School performance and attendance is influenced by various social and demographic factors. In the United States, students who perform well in school typically have parents or grandparents who have attended college and thus understand the importance of obtaining a quality education. In addition, they have access to educational resources and opportunities that make it easier for them to succeed. On the other hand, students from underprivileged backgrounds, who lack access to quality education and face life difficulties such as poverty, food insecurity, and homelessness, are more likely to encounter obstacles that affect their academic performance and attendance. As a result, these students tend to be chronically absent from school and perform poorly.

The project aims to explore the student performance data donated to the UCI Machine Learning Repository in 2014 by researchers who were interested in examining the impact of demographic and social information on the academic performance of students in Portuguese secondary schools [1]. The data was collected during the 2005-2006 school year from two public schools in Portugal. The data was obtained from school reports and questionnaires directed towards students and their families. By analyzing this data, the project aims to gain insights into the relationship between various social and demographic factors and student attendance, which could inform school policies and programs to improve the educational outcomes of underprivileged students.

1.2. Objectives

To gain a better understanding about the chronic absenteeism and inform programs designing, the project covers the following two objectives:

(1) Develop a model to predict chronic absenteeism.

(2) Identify the factors that are associated with chronic absenteeism.

2. Data

2.1. Data Structure

The data is stored in two separate files, with one file for each class: Portuguese and Math. In total, the dataset contains 1033 observations and 33 variables. Each observation in the dataset represents a single student who participated in the study. The variables include information such as the education background of students' parents, their interest in higher education, and lifestyle choices.

For the purpose of our project, we create two additional variables. The first variable is a categorical variable called "Subject," which indicates the class from which the data was obtained. The second variable is a binary categorical variable called "CA," which indicates whether a student was chronically absent. Chronic absenteeism is defined by the U.S Department of Education as missing at least 15 days of school in a year [2], and we use this standard to identify whether a student was chronically absent. The complete data dictionary is in Appendix A.

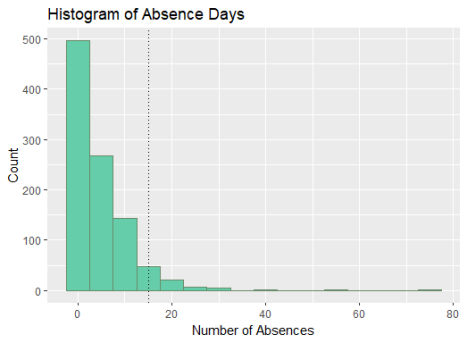

Figure 1: Histogram of Absence Days

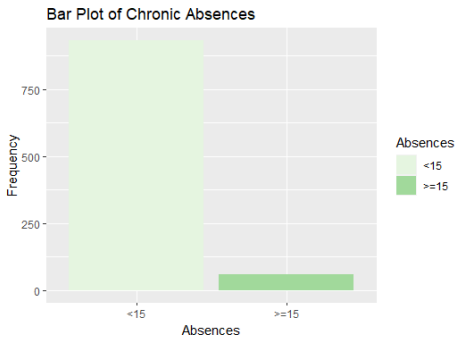

Figure 2: Bar Plot of Chronic Absences

2.2. Exploratory Data Analysis

Our variables of interest "absences" and "CA" can be visualized with a histogram. We can see that the distribution of absences are heavily left skewed. Most students have between 0 and 5 absences with the median value being 2. As absences increase the number of students who experience relatively high absences decreases exponentially. As mentioned, any student that has more than 15 absences is considered chronically absent. The dotted line on the histogram indicates 15 absences. We also make a bar plot of the binary variable "CA". Based on the bar plot, there exists a strong imbalance. There are only 59 students who we would consider to be chronically absent and 931 students who are not considered chronically absent. Students that are not chronically absent make up 94.04% of the dataset and chronically absent students make up approximately 5.96% of the dataset.

3. Method

3.1. Resampling Methods

In 2.2, we learn that the data is imbalanced. Using the imbalanced data to train classification models are likely to develop a model biased towards the majority class [3]. At the same time, we are more interested in the minority class. Therefore, we would like to try different resampling methods to deal with the imbalanced data problem. We first divide the data into 70% of the training data (693 observations) and 30% (297 observations) of the testing data. We perform the following resampling methods only on the training data, and use the processed data to train our models and compare the results.

3.1.1. Upsampling

Upsampling is the process of increasing the number of observations in the minority class by randomly duplicating observations in the minority class [4]. After upsampling, there are 649 observations in the class of chronic absences and 649 observations in the class of no chronic absences.

3.1.2. Downsampling

Downsampling is the process of decreasing the number of observations in the majority class by randomly removing observations in the majority class [4]. After downsampling, there are 44 observations in the class of chronic absences and 44 observations in the class of no chronic absences.

3.1.3. ROSE (Random Over-Sampling Examples)

ROSE is a resampling technique that utilizes bootstrapping to produce synthetic examples from a conditional density estimate of both classes [5][6]. After ROSE, there are 371 observations in the class of chronic absences and 322 observations in the class of no chronic absences.

3.1.4. SMOTE (Synthetic Minority Over-sampling Technique)

SMOTE is a resampling technique that generates synthetic examples of the minority class using K nearest neighbors [7]. For the categorical variables in our data, we choose the most common category along neighbors. After SMOTE, there are 649 observations in the class of chronic absences and 649 observations in the class of no chronic absences.

3.2. Model Selection

In addition to the imbalanced data problem, we expect there may be multicollinearity problem. Therefore, we decide to train a random forest model and a penalized logistic regression model as they are robust to the multicollinearity problem. For the random forest model, we use the 5 times repeated out-of-bag (OOB) to choose the best number of predictors that are evaluated for each split and the number of trees. For the penalized logistic regression model, we use 10-fold cross-validation (CV) to choose the best tuning parameters α and λ. In order to pick the best combination of resampling method and model, we use the original data and the data generated by different resampling methods to fit the two models three times each and take the average to make comparisons.

4. Results

4.1. Model Evaluation

We use the fitted model to predict the unprocessed testing data. Table 1 shows the model evaluation statistics of the random forest models using data generated by different resampling methods. Table 2 shows the model evaluation statistics of the penalized logistic regression models. Sensitivity indicates the probability of predicting a student to be not chronically absent, conditioned on they are not. Specificity indicates the probability of predicting a student to be chronically absent, conditioned on they are.

	Original	Upsampling	Downsampling	ROSE	SMOTE
Accuracy	0.9473	0.9428	0.6487	0.8911	0.9439
Specificity	0.9965	0.9906	0.6529	0.9233	0.9870
Sensitivity	0	0.0227	0.5682	0.2727	0.1136

Table 1: Model Evaluation Statistics of Random Forest Models

	Original	Upsampling	Downsampling	ROSE	SMOTE
Accuracy	0.9461	0.8339	0.7329	0.7845	0.8855
Specificity	0.9917	0.8477	0.7355	0.7934	0.9103
Sensitivity	0.0682	0.5682	0.6818	0.6136	0.4091

Table 2: Model Evaluation Statistics of Penalized Logistic Regression Models

For both random forest models and penalized logistic regression models, the original data always leads to models with highest accuracy for our data. However, this is mainly because it trains the models to always predict the majority class. We can see that the sensitivity of the models using the original data is very close to 0, meaning that the models almost never correctly predict a student who is chronically absent. On the other hand, downsampling always leads to the models with highest sensitivity, but they all have relatively low overall accuracy. The rest of the three resampling methods lie in the middle, with relatively high overall accuracy and better sensitivity than the models trained with the original data.

To choose the best model, we set a standard based on the accuracy and sensitivity. We hope to select a model with an accuracy above 0.8 and a specificity above 0.5, which means the model predicts the status of chronic absenteeism correct more than 80% of the times, and predicts correctly more

than 50% of the times when the student is actually chronically absent. The only model that satisfies the standard is the penalized logistic regression model trained with the upsampling training data set. The optimized α value is 0.85. Therefore, we select the model as our best model for prediction.

4.2. Important Features and Implications

We use permutation after fitting to determine the importance of each feature for predicting which students will be chronically absent. The top five important predictors except their final scores are: "G1", "school", "paid", "traveltime", and "studytime". The students who got a higher grade in the first period are more likely to be chronically absent. There were two schools in our dataset and the students in the school labeled as MS are less likely to be chronically absent than the students in the school labeled as GP. Students that paid for extra tutoring were less likely to be chronically absent. Travel time was a variable that gave a fairly obvious conclusion. Students that had shorter travel times to school were less likely to be chronically absent. Lastly, students that studied more were less likely to be chronically absent.

In general, students that are more engaged academically seem to be less likely to be chronically absent. Students that are not as engaged academically struggle to show up to class on a regular basis. Although we cannot say that there is a direct causal relationship between some of these variables and chronic absences. There are patterns that can give us a more complete picture what may related to students that are chronically absent. The school may consider providing additional supports for students who have difficulties in coming to school or paying for academic materials to encourage them to actively participate in academic activities and show up for classes.

5. Conclusion

This project develops a model to predict chronic absenteeism using student performance data. Various resampling techniques are applied to the training data to address the imbalanced data problem. Ultimately, the most successful model is the penalized logistic regression model trained on an upsampled dataset. The findings give insights into how the socioeconomic factors contributing to chronic absenteeism to inform school policies. Further research could explore additional models such as neural networks or boosting methods and investigate what resampling methods perform better under the models with other imbalanced datasets.

Appendix A. Data Dictionary

Attribute	Description
school	Student's school ("GP" or "MS")
sex	Student's sex
age	Student's age
address	Student's home address type (urban or rural)
famsize	Family size
Pstatus	Parent's cohabitation status
Medu	Mother's education
Fedu	Father's education
Mjob	Mother's job
Fjob	Father's job
reason	Reason to choose this school
guardian	Student's guardian
traveltime	Home to school travel time
studytime	Weekly study time
failures	Number of past class failures
schoolsup	Extra educational support
famsup	Family educational support
paid	Extra paid classes within the course subject
activities	Extra-curricular activities
nursery	Attended nursery school
higher	Wants to take higher education
internet	Internet access at home
romantic	With a romantic relationship
famrel	Quality of family relationships
freetime	Free time after school
goout	Going out with friends
Dalc	Workday alcohol consumption
Walc	Weekend alcohol consumption
health	Current health status
G1	First period grade
G2	Second period grade
G3	Final grade
Subject	The class from which the data was obtained
CA	Chronically absent

References

[1] P. Cortez, A. Silva, Using data mining to predict secondary school student performance, EUROSIS (01 2008).

[2] U.S. Department of Education (2019). [link].
URL https://www2.ed.gov/datastory/chronicabsenteeism.html#intro

[3] B. Krawczyk, Learning from imbalanced data: open challenges and future directions, Progress in Artificial Intelligence 5 (4) (2016) 221–232. doi:10.1007/s13748-016-0094-0.
URL https://doi.org/10.1007/s13748-016-0094-0

[4] H. He, E. A. Garcia, Learning from imbalanced data, IEEE Transactions on Knowledge and Data Engineering 21 (9) (2009) 1263–1284. doi:10.1109/TKDE.2008.239.

[5] G. Menardi, N. Torelli, Training and assessing classification rules with imbalanced data, Data Mining and Knowledge Discovery 28 (1) (2014) 92–122. doi:10.1007/s10618-012-0295-5.
URL https://doi.org/10.1007/s10618-012-0295-5

[6] N. Lunardon, G. Menardi, N. Torelli, Rose: a package for binary imbalanced learning, R Journal 6 (2014) 79–89. doi:10.32614/RJ-2014-008.

[7] N. Chawla, K. Bowyer, L. Hall, W. Kegelmeyer, Smote: Synthetic minority over-sampling technique, J. Artif. Intell. Res. (JAIR) 16 (2002) 321–357. doi:10.1613/jair.953.