Predicting North Carolina High School Outcomes

Ethan Shen

May 01 2020

1. Introduction

North Carolina is one of the best states to attend college in; it ranks 12th for higher education. However, this same report found that the state ranks a disappointing 28th for Pre K-12 education. I decided to explore a key metric for student success in North Carolina - the growth index score, calculated by the NC Department of Public Instruction's Education Value-Added Assessment System (EVAAS). This motivated me to ask the question: what performance and demographic factors of high schools in NC are most related to success, and how can policymakers understand these relationships to support underperforming schools?

1.1 Data and Initial Challenges

I first found three separate datasets: one detailing each school's growth index, one containing the percent of students in each subgroup meeting the ACT benchmark standards and one detailing the region in which each high school is located, its Title I status, number of days missed due to Hurricane Florence, and information about the school's performance on various standardized tests and assessments based on different racial and socioeconomic subgroups.

The data has two response variables: the *growth_index_score* (GIS), a numerical value ranging from -17 to 13 and *growth_status*, a categorical variable determining if a school did not meet, met, or exceeded state standards. The GIS was calculated used state-administered exams to determine current achievement compared to prior achievements and scores below -2 have "Not Met," between -2 and 2 have "Met," and above 2 have "Exceeded" standards. In this project, I will focus on classifying the *growth_status*.

1.2 Assumptions

1. Benchmark extremes: Values below the 5% benchmark were categorized as "<5%" and those above 95% were categorized as ">95%" in all our original data. I converted them to 5 and 95, respectively.
2. Junior class composition: As all juniors in North Carolina are mandated to take the ACT, I manipulated the dataset to obtain information about the junior class composition of each high school.
3. Race/ethnicity: For the percent_white variable, if schools had no information about white students, I assumed 1 - (non-white racial percentages) equaled percent_white.
4. Gender: I am unfortunately assuming a gender binary and regarding 1 - (female percentage) as male percentage.

2. Modeling

The focus of this project aims to identify various academic and composition related factors of a high school that are relevant to its growth index score and status. This project aims to achieve the goal of inference, rather than predicting the growth index score of each high school.

SVM ordinal logistic

then within each one category perform hierarchal modeling, gam

3. Procedure

3.1 Analysis

First I split the data into a 75-25 training-testing split, resulting in approximately 300 schools in the training set. I ran the multi-level classifiers on the training data containing the following features: *percent_white*, *percent_male*, *percent_eds*, *percent_swd*, *title_1*, *percent_ACTEN_meeting*, *percent_ACTMA_meeting*, *percent_ACTRD_meeting*, *percent_ACTSC_meeting*, *percent_ACTWR_meeting*, *percent_ACT_meeting*, *missed_school_days*, and *region*.

I assigned the resopnse variable of *growth_status* of "Not Met", "Met", and "Exceeded" to be -1, 0 and 1. Then I will compare the prediction on the test data (the remaining 100 schools) to find the misclassification rate to conclude whether or not these models are capable of accurately predicting *growth_status* for North Carolina High Schools.

In addition, to determine which SVM will perform the best, I performed cross-validation to conclude which *type* and *kernel* resulted in the highest accuracy. For the Random Forest, I performed 10-fold cross validation and 7-fold cross validation with the Naive Bayes. This was done to maximize the accuracy rate across the different classifiers.

3.2 Comparison

I will compare the four classifiers by looking at the misclassification rates and then recognize which model performs the best.

4. Results and Conclusion

The classifiers I used are Naive Bayes, Random Foreset, SVM with a radial kernel, and Ordinal Logistic Regression. All four models are multiclass as there are three levels in our response variable, *growth_status*.

	Naive Bayes	Ordinal Logistic	Random Forest	SVM
Accuracy	0.4362	0.43	0.4217	0.47

4.1 Confusion Matrix for Best Classifier

The best classifier was the SVM. Below is the confusion matrix.

	Actually Not Met	Actually Met	Actually Exceeded
Predicted Not Met	19	12	11
Predicted Met	9	21	9
Predicted Exceeded	4	8	7

From these models, I found that the radial SVM performs the best, predicting with an accuracy of 0.47, followed by the Naive Bayes with an accuracy of 0.4362. I also display the confusion matrix for the radial SVM.