Predicting North Carolina High School Outcomes

Ethan Shen

May 02, 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 better understand these relationships to support underperforming schools?

1.1 Data

I first found three separate datasets, each including data from 2018-2019 school year: one detailing each school's growth index, one containing the percent of students in each subgroup (racial and socioeconomic) 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 Initial Challenges

The initial datasets presented many challenges. In each dataset, there are multiple rows associated with each school. Each row has a distinct *subgroup*, for example gender, race, or socioeconomic status and *subject*, which is associated with some standardized test like the ACT. Each distinct combination of *subgroup* and *subject* within a school then had variable *total_pct*, indiciating the percentage of students within that specific *subgroup* and *subject* combination that met the standard for that test.

However, my goal was to have one dataset where each school formed one observation. For the first two datasets, I thus filtered the data for *subgroup* values of *All Students*, which would provide information about overall school performance. Below is a snapshot of the third dataset, which required the most manipulation.

school_name	subgroup	subject	den	total_pct
State of North Carolina	ALL	ACT	101541	55.8
State of North Carolina	FEM	ACT	51510	59
State of North Carolina	MALE	ACT	50031	52.5

Each school is repeated several times to show the percentage of students who meet benchmark standards on the ACT (*total_pct*) by various subgroups. The *den* variable indicates the total size of each of the subgroups amongst those students who have taken the ACT. Because all juniors in North Carolina are mandated to take the ACT,the *den* variable corresponding to the subgroup *ALL* represents the size of the junior class, while the *den* variable corresponding to other subgroups (such as *MALE*) represents the size of that particular subgroup in the junior class.

I wrote a function that calculates the percentage of juniors that fall into each of the following subgroups: white, male, economically disadvantaged, and disabled. Thus, I was able to extract the composition of each school's junior class as it pertained to these four categories. I used these percentages as a proxy for the school-wide composition, as this information was not publicly available.

The first three observations in the final dataset are below.

school_code	school_name	percent_white	percent_male	percent_eds	percent_swd
010324	Eastern Alamance High	57.746479	55.63380	33.80282	9.15493
010348	Graham High	20.155039	44.18605	66.66667	10.07752
010360	Hugh M Cummings High	4.761905	43.80952	72.38095	10.95238

title_1	percent_ACTEN_meeting	percent_ACTMA_meeting	percent_ACTRD_meeting	percent_ACTSC_meeting
N	35.2	21.1	26.4	21.8
N	12.2	7.7	13.2	8.5
N	5.7	5.2	6.7	5.0

percent_ACTWR_meeting	percent_ACT_meeting	missed_school_days	region	growth_status	growth_index_score
26.8	51.1	1	Piedmont-Triad	Not Met	-3.09
8.5	31.8	1	Piedmont-Triad	Met	-0.93
8.2	18.1	1	Piedmont-Triad	Met	0.70

1.3 Missing Data

I wanted to ensure I addressed missing data as well. Since there were only approximately 400 high schools in the datasets, I was already starting with a relatively small sample size. The final dataset had 2% missing data, distributed in a way where I had 396 complete cases but only 6 covariates with complete information. From the plots, there seemed to be missingness dependence, meaning that the data could be MCAR. Thus, I utilized MICE to deal with the missing data.

1.4 Assumptions

1. Benchmark extremes: Values below the 5% benchmark were categorized as "<5%" and those above 95% were categorized as ">95%" in all the 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 status, in the hopes that I can gain a better understanding of which factors contribute to successful performance in North Carolina high schools, and how policymakers can make tangible and meaningful changes to the state's education system.

When moving forward with the modeling, I excluded three variables. The first is *growth_index_score*, which is a direct predictor of each school's *growth_status*. This would cause the predictive results to be extremely high, which is not conducive to bringing about meaningful change. I also removed the *school_code* and *school_name*, as neither of these variables are helpful.

Initially, I was only concerned with socio-economic factors that contribute to a school's success. However, a school's geogrpahical location may also be important in determining that school's success. For example, regions with a high average median family income usually have better schools than regions with a lower average median family income. I want to analyze the impact that geographical location has on determining *growth_status* by creating one set of classifiers without *region* and one set including *region*.

3. Procedure

3.1 Analysis excluding Geographical Information

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*, and *missed_school_days*.

I assigned the response 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 classification rate to conclude whether or not these models are capable of accurately predicting *growth_status* for North Carolina High Schools.

3.2 Analysis including Geographical Information

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*. Similarly, I assigned the response 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 to find the classification rate.

3.3 Comparison

I will first compare the classifiers without *region* by looking at their classification rates. In addition, I can conclude which model performs the best. Then I will analyze the classifiers with *region* and determine whether adding this feature will improve the accuracy of my classifiers.

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 the response variable, *growth_status*.

Table 1: Without Region

	Naive Bayes	Ordinal Logistic	Random Forest	SVM
Accuracy	0.4362	0.45	0.3949	0.45

Table 2: With Region

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

4.1 Confusion Matrix for Best Classifier

Table 3: Confusion Matrix for Radial SVM (Without Region)

	Actually Not Met	Actually Met	Actually Exceeded
Predicted Not Met	17	11	11
Predicted Met	15	26	14
Predicted Exceeded	0	4	2

Table 4: Confusion Matrix for Ordinal Logistic (Without Region)

	Actually Not Met	Actually Met	Actually Exceeded
Predicted Not Met	15	9	11
Predicted Met	14	22	8
Predicted Exceeded	3	10	8

Table 5: Confusion Matrix for Radial SVM (With Region)

	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 for classifiers without *region*, the radial SVM and ordinal logistic perform the best, predicting with an accuracy of 0.45, followed by the the Naive Bayes with an accuracy of 0.4362. For the classifiers with *region*, radial SVM performs the best, predicting with an accuracy of 0.47, followed by the Naive Bayes with an accuracy of 0.4362. Compared to the model without *region*, the model including *region* has a higher accuracy of 0.02 for the SVM, 0.0268 for the Random Forest, no change for the Naive

Bayes, and a lower accuracy of 0.02 for the ordinal logistic. It is interesting to note that only the ordinal logstic model was negatively affected by the addition of *region*.

Within both sets of classifiers, all three of the best performing models predict schools that **Met** standards with the highest accuracy, with the SVM including *region* having an accuracy of 0.5384 and the Ordinal Logistic excluding *region* having an accuracy of 0.5. The SVM excluding *region* has an accuracy of 0.473, which is worse than the two previous models.

4.2 Conclusion

By adding the *region* feature, our models will generally classify whether a high school exceeds, meets, or does not meet North Carolina standards better. The accuracy of the models with *region* generally increases for each respective classifier compared to the model without *region*. However, there is a notable outlier, the ordinal logistic, whose accuracy actually decreases with the inclusion of *region*.

In addition, two of the best classifiers were able to predict **Met** standards with over 50% accuracy, but no model was able to predict **Not Met** or **Exceeded** with over 50% accuracy.

4.3 Limitations

One of the biggest limitations in this project was the amount of data. First and foremost, the data only recorded information from the 2018-2019 school year and only contained roughly 400 observations, which is not adequately enough data to accurately predict a school's growth and may be why the accuracy of the classifiers are less than 50%. Having data from only one year's data also prevented me from analyzing whether a school's performance in previous years affects their current performance. Secondly, I had to make assumptions about the data that may have compromised the classifiers. A prime example of this was how I used junior-class composition percentages (such as percent white or percent economically disadvantaged) as a proxy for school-wide composition because I could not find this information anywhere else. Lastly, though the datasets provided the percentages of students who met ACT benchmarks for each high school, the average ACT score would be more beneficial, especially in helping determine which schools exceed standards. In continuing this research, I want to utilize more information and reach out to policymakers and officials who may have more comprehensive, and perhaps restricted, datasets related to school composition and performance.