Get a Grip on Graduation

Using Machine Learning to Help Virginia Educators Keep a Grip on Potential High School Dropouts

Richard Seifert
Department of Astronomy
UVA Graduate School of Arts & Sciences
Charlottesville, VA
rseifert@virginia.edu

Andy Taylor
Department of Astronomy
UVA Graduate School of Arts & Sciences
Charlottesville, VA
ajt9gx@virginia.edu

***Abstract*—Every year, Virginia high schools say an early goodbye to ∼7,000 of their students. While it is unclear why a student decides to leave school, what is clear is that—without a high school diploma—their career opportunities are diminished. Thanks to efforts by the Viriginia Department of Education (VDOE) and the Virginia Longitudinal Data System (VLDS), enough data has been amassed to robustly identify the most vulnerable students. Here, we attempt to determine which students are most at risk of dropping out and aim to identify the specific factors these students have in common that most strongly correlate with their decision to leave school. To do this, we explore a range of predictive models and train these models with publicly available data on graduating students as well as their schools and counties. We then use our best model to isolate the influence of individual factors on dropout rate, ultimately ranking these factors by their ability to influence the model's dropout predictions. We find that the factors most strongly connected with dropout rate are race, income, and English-language proficiency, and we discuss ways that Virginia educators can work to mitigate education inequality with regard to these factors.**

I. Introduction

A. Motivation

When a student chooses to drop out of high school, they dramatically alter the course of their life. Every year in Virginia, thousands of students drop out of high school [1], and many never return to continue pursuing a diploma. In today's world, "education inflation" lowers the value of high-school diplomas and college degrees relative to a generation ago, meaning that job candidates who lack a high-school diploma are becoming less and less hireable. To work toward solving this problem, we are studying the graduation and dropout rates of high school students in Virginia.

B. Project Description

We aim to create a model that is capable of predicting the dropout rates of Virginia high school students based on data reported for graduating classes. A final model capable of making accurate predictions will provide valuable information to state education officials about where to direct future efforts in order to increase student retention. Additionally, as new high schools are built, our model will provide an estimate for the new school's dropout rate, which could serve as a benchmark against which to assess the school's performance. Finally, our model may be easily generalized to a national population and provide similar information to education officials in other states and at the national level.

In section II, we discuss the data we use and ways we have augmented and manipulated that data to help us make better predictions of Virginia graduation rates. In section III, we discuss the models we have created, compare their performance, and discuss our final model. In section IV, we use our final model to explore underlying trends in the graduation data, and ultimately report on the features in our data that most strongly influence graduation outcome. Finally, in section V, we conclude and discuss potential ways that high dropout rates we observe could be mitigated in the future.

II. Preparing The Data

For this project, we are using data from the Virginia Department of Education (VDOE [2]; doe.virginia.gov), which we augment with data from the Public School Review (PSR [3]; publicschoolreview.com) and the US Census [4] (census.gov/quickfacts). The data we obtained from the VDOE spans 12 years, with aggregate data on over one million students from over 400 Virginia high schools. The VDOE provides a number of categorical features pertaining to the school (county name, school name) and to students (race/ethnicity, gender, disability and disadvantage status, English proficiency). For each unique combination of these features, the VDOE reports the total number of fourth-year students and the percentages of these students who are graduating and dropping out; these do not perfectly complement each other, since students planning to graduate late are not considered graduates *or* dropouts by the VDOE.

A. Missing Data

We encountered some issues with the VDOE dataset due to its format and due to the VDOE's censuring of personally identifiable information (which we agree with, but also attempt to work around). The data is formatted hierarchically with aggregate rows that represent sums over features as well as filled rows with all features reported. For the purposes of this project, we are not using the aggregate rows, and only use filled rows to train models. However this poses a problem because, to protect individuals' privacy, no data was reported for any subset of fewer than 10 students. For example, at Charlottesville High School in 2008, there were

only 14 Hispanic 4th-year students. Since stratifying these 14 students by other demographics (e.g. gender, disability, English proficiency) would reduce the size of each subset to fewer than 10 students, the VDOE does not report these demographics for this group of students, and as it stands we would not incorporate these students into our predictive models. It is important to note that this omission scheme systematically removes data on minority students, because subsets of these groups fall below the 10 student threshold more often than their counterparts belonging to racial/ethnic and gender majorities. Since our aim is to identify populations that are currently under-served by Virginia high schools, it is vitally important that we mitigate this problem as well as we can.

To do this, we represent the VDOE dataset with a graph structure, connecting filled rows with their aggregate row "parents". Then we can recover omitted students using their non-omitted "siblings". An illustration of this is shown in Table I. We show hypothetical data consisting of three flags (values either 0 or 1) and group size. The blank space in the first column of the first row indicates an aggregated feature, combining group size for both 0 and 1 flag values. We are able to recover omitted students by combining the aggregate "parent" row (18 students) with the filled "sibling" row (12 students) in order to infer a group size of 6 students for the omitted row. Following this scheme across the entire VDOE dataset, we were able to recover information on over 65% of all students with missing data. At the end, we were missing data on only 7% of all high school students in Virginia. Though we regret our inability to obtain data on all students, we consider our efforts successful (and potentially alarming, as it represents a breach of the VDOE's attempt at releasing non-identifiable data).

Flag 1 (0/1)	Flag 2 (0/1)	Flag 3 (0/1)	Group Size
	0	1	18
0	0	1	12
1	0	1	(Inferred 6)

TABLE I
ARTIFICIAL VDOE DEMOGRAPHIC DATA, DEMONSTRATING OUR SCHEME TO FILL IN INFORMATIONAL GAPS.

B. Feature Engineering

Before beginning our analysis, we wanted to bring in more information to augment the VDOE data that we believed would be helpful in predicting graduation rate. Having lengthy experience as students ourselves, we believed that having information about the schools (total enrollment, test scores, number of teachers, etc.) as well as location-specific information (e.g. median income) would be beneficial for predicting graduation rates. Using data scraped from Public School Review and from the US Census, we were able to add the following features about the high schools and their surrounding areas to our dataset: income per capita, median household income, median family income, size of school (students), student/teacher ratio, fraction of students who tested proficient in math, and fraction of students who tested proficient in reading. In total, our final dataset contains 13 features and roughly 72,000 rows, where each row corresponds to a group of students with matching demographic features.

Feature	Values/Range	Source
Gender	M/F	VDOE
Race Code	1, 2, 3, 4, 5, 99	VDOE
Disabled	Y/N	VDOE
Disadvantaged	Y/N	VDOE
Limited English Proficiency	Y/N	VDOE
Cohort Size	1 – 250 students	VDOE
Income per Capita	\$16,000 – \$58,000	USC
Median Household Income	\$22,000 – \$115,000	USC
Median Family Income	\$34,000 – \$142,000	USC
School Size	20 – 3,785 students	PSR
Student-Teacher-Ratio	6 – 21	PSR
Math Proficiency	15% – 99%	PSR
Reading Proficiency	49% – 99%	PSR

TABLE II
OVERVIEW OF FEATURES WE USE TO PREDICT DROPOUT OF VIRGINIA HIGH SCHOOL STUDENTS. DATA COMES FROM THE VIRGINIA DEPARTMENT OF EDUCATION (VDOE) [2], PUBLIC SCHOOL REVIEW (PSR) [3], AND THE US CENSUS (USC) [4].

III. PRODUCING THE MODEL

A. Classification Approach

Originally, our aim was to come up with a model which could take information on an individual student and predict whether or not that student would graduate; we wanted to treat this as a classification problem. Doing so required an inflation of the data so that each entry corresponded to an individual student instead of a cohort of students. So for each cohort of students in our original dataset, we used the reported dropout rate to determine the numbers of students graduating and dropping out. From here, we constucted our inflated dataset with each row corresponding to an individual student and replaced the dropout *rate* with a dropout *status* (either a 0 or 1 if that student did or did not drop out).

This expansion increased the size of our dataset from roughly 72,000 rows to over a million rows, where now each row corresponded to a single student that either did or did not drop out. We trained a binary classifier on this dataset and compared the results of the `predict_proba()` function on the testing data with the actual dropout rate, and we found our results disappointing. In our inflation of the data, we had not introduced any new student-specific information, so each student in a cohort was essentially identical in the model's eyes except that some arbitrarily dropped out while others arbitrarily graduated. Having no other predictive features, the model was unable to properly learn what caused two students with identical feature vectors to have different outcomes, and so the model performed poorly. In light of the disappointing performance and the questionable technical correctness of this methodology, we ultimately discarded the classification approach to solving this problem.

B. Regression Approach

Instead, we decided to treat this like a regression problem, using cohort information to predict the *percentage* of students choosing to drop out. Taking this approach solved the problem of our models receiving identical data points with different labels, but we quickly encountered a different obstacle. When we examined our data, we saw a strong trend between dropout rate and cohort size and it seemed that small cohorts were doomed to have large dropout rates. While this trend is technically real, it is also extremely misleading, because small

cohorts are much more susceptible to dropout rate inflation due to Poisson noise. For example, it is much more likely that 3 students drop out in a 5 student cohort than that 300 students drop out in a 500 student cohort, but both represent a 60% dropout rate. The problem is actually exacerbated by our efforts to infer cohorts of fewer than 10 students. This strongly nonlinear trend posed a problem for our preliminary models, and because it is the dominant trend in our data, the models we trained basically only learned "small cohort size = large dropout rate". To get around this, we chose to recast our target variable from dropout rate to dropout *number*. This change of variables is illustrated in Figure 1.

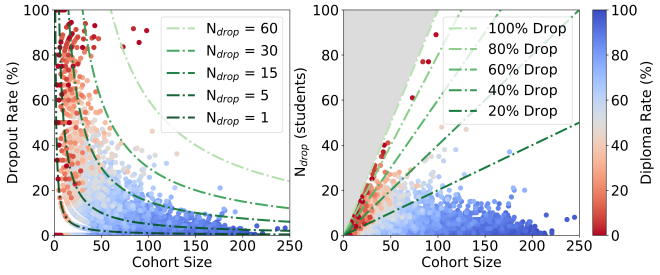

Fig. 1. Left: Percentage of students dropping out as a function of the cohort size. The dominant trend is well explained by small-number variability, as shown by green contours of constant *number* of dropout students. Right: Number of students dropping out as a function of cohort size, with green contours of constant dropout rate.

C. Early Models

Having decided on a methodology (regression), we began with a simple model to get a baseline score which more complicated models strove to out-perform. Using scikit-learn's LinearRegression model, we were able to achieve an RMSE of 1.88 students when predicting dropout number (see Figure 2, upper left). This problem is clearly more complicated than a linear regression model is able to account for, but we still find it useful for obtaining a baseline RMSE.

Next, we trained scikit-learn's RandomForestRegressor models. After tuning hyperparameters using a randomized grid search, we were able to achieve an RMSE of 1.4 students when predicting dropout number (see Figure 2, upper right). As expected, the more complex model performed better than the simple LinearRegression model, but upon further inspection, it was still clear that the random forest model wasn't able to adequately reproduce many of the high-dropout points.

Finally, we performed some preliminary modelling using a sequential neural net from the Keras package in the TensorFlow library. Without tuning hyperparameters, the model was able to achieve an RMSE of 1.38 students when predicting dropout number (see Figure 2, lower left). The performance of this un-tuned neural network was promising, because without extensive tweaking, it was already outperforming our fully-tuned random forest regression model.

D. Final Model

We continued to refine our neural net model. Our goals were to achieve better test RMSE by refining the architecture of our model and to find a workaround for enabling the model to fit better to large cohort and large dropout datapoints. To get lower RMSE from our model we wished to vary hyperparameters of the neural net (e.g. number of layers, number of nodes at each layer, activation function) in order to achieve better validation RMSE, which should predict better test RMSE. This proved difficult in practice, because each model we tested required a large amount of time to train. This made it impractical to do a robust, randomized search over hyperparameters.

One of the biggest issues we faced with our early models was that they predicted well only for very small cohorts (LinearRegression was an extreme example of this; first panel of Figure 2). At the heart of this issue is the fact that the majority of the datapoints our models are trained on are small cohorts of fewer than 20 students (again, this is something made more severe after we inferred small cohorts from the VDOE data in Section II-A). To resolve this issue, we added weights during model training to give more importance to cohorts with more students. We believe weighting in this way gives each student equal influence on the model during training; weighting the training data according to the number of students belonging to each group should produce a model which predicts most accurately for the most students.

Using this weighted training scheme and some of the intuition we gained during our failed attempts at a random hyperparameter search, we were able to build and train a final neural network model which achieves an RMSE of 1.29 students with our test dataset (Figure 2, lower right). We believe that this represents nearly the best achievable RMSE with the data we are using.

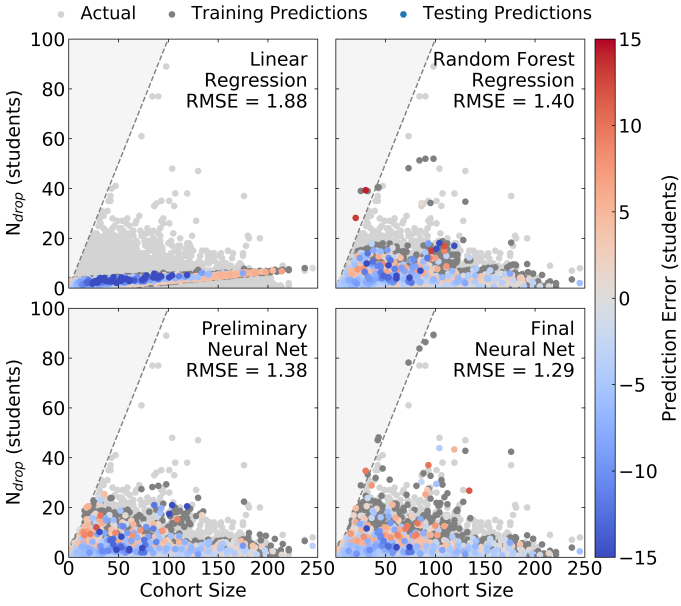

Fig. 2. Comparisons between our models. Shown above, we use our models to replicate the right panel of Figure 1 and also report each model's RMSE. Light gray points are true dropout counts, dark gray are predictions made with training data, and colored points are predictions made with testing data, colored by the error in the prediction.

IV. Data Inferences from the Model

With a reasonably good model in hand, we can now accurately predict dropout rates of Virginia high school students.

In addition to this, we believe our model has successfully encapsulated the underlying trends present in the data it was trained on, and therefore can be leveraged to gain insight about the factors influencing Virginia graduation outcomes. To extract this insight, we used our final model to make predictions on artificial data, varying individual features to demonstrate their influence on the predicted graduation rate. We show in Figure 3 a 5×4 grid illustrating how each feature affects the model's dropout rate predictions.

Before we discuss the correlations and importance of each feature, it is important to describe carefully how we used our model to extract the trends we present. In each of the panels of Figure 3, we evaluate our model varying only two features (one numerical and one categorical), however each prediction is also a function of the remaining eleven features our model was trained on. To make our predictions, we wanted to ensure that we were choosing these non-varying features to be representative of our data. A simple solution is just to use the global mean value from the data for each non-varying feature; however, we worried that this approach would overlook behavior in our model at the edges of the feature space (noteably, this is where minority students are represented). Instead, we chose to evaluate the model for a range of values for each non-varying feature and average them together to get our final prediction. To ensure that the values we chose for each non-varying feature were representative of the data, we sampled points from the full dataset and used these to get values for the non-varying features. The trends we present in Figure 3 represent averages over 35 randomly sampled values for each non-varying feature.

In addition to averaging over non-varying features in each tile of Figure 3, we also averaged over a range of cohort sizes. Cohort size is an somewhat perplexing feature in our dataset, because while it is purely an artifact of the VDOE's choice of data format (their measure to protect individuals' privacy by reporting on groups of students with common demographics, rather than each individual student), it is still strongly associated with other features in our data. In general (and almost by definition), minority students are better-represented in small cohorts, and majority students are better-represented in large ones, and this is an association that went into our model during training. With this in mind, we chose to combine predictions from our model at cohort sizes of 5, 10, 15, 50, and 150. We converted each predicted number of dropouts into a dropout rate and averaged the predictions together to obtain the dropout rates we report in Figure 3. As a result of our averaging scheme, the values we report for dropout rate are somewhat fictitious, but we believe that differences between predicted dropout rates are still valid for comparing the influence of different features.

Now that we have explained how Figure 3 was generated, let us discuss the implications presented therein. Below we examine the influence each feature has on the predicted dropout rate. We summarize our findings in Table III.

Race: By far we find that the most influential feature predicting dropout rate is race. We see that Hispanic and African-American students consistently have the highest dropout rates, often dropping out at anywhere from 2-5 times the rate of their white student counterparts. It is interesting to note the upper left panel of Figure 3, showing trends with race and median household income. We find that the discrepancy with race is mitigated as median household income increases, suggesting that the discrepancy is, in part, socioeconomic in nature.

Feature	Dropout Rate Variation	Predictive Power
Race	10 – 20%	H
Regional Wealth	5 – 15%	H
Reading Proficiency	4 – 12%	H
English Proficiency	2 – 8%	H
Gender	2 – 5%	M
Math Proficiency	2 – 5%	L-M
Student-Teacher Ratio	5 – 12%	L
Disability Status	0 – 4%	L
Disadvantage Status	0 – 4%	L

TABLE III
FEATURES AND THEIR PREDICTIVE POWER OVER HIGH SCHOOL DROPOUT RATES. FEATURES ARE CLASSIFIED AS EITHER HIGH (H), MEDIUM (M), OR LOW (L) PREDICTIVE POWER.

Regional Wealth: We see that the predicted dropout rate always decreases as the wealth of the county increases, and this can boost dropout rates by as much as 2 times in the poorest counties, compared to the wealthiest counties.

Reading Proficiency: We find that a low reading proficiency in schools is correlated with larger dropout rates, with students at low-reading-proficiency schools being up to 3 times more likely to dropout, compared to schools with high reading proficiency.

English Proficiency: We note that there is also a strong relationship between dropout rate and English proficiency, with low English proficiency (LEP) students being up to $\sim$2 times more likely to drop out.

Gender: We see a surprisingly consistent discrepancy between male and female students, with male dropout rates being $\sim$3% higher than that of female students across the board.

Math Proficiency: The percentage of students testing proficient in math does not have a very significant affect on dropout rates, at most boosting the dropout rate by $\sim$5%. What we found odd is that, unlike reading proficiency, schools with higher math proficiency see *more* dropouts, on average. Compared to reading proficiency, the range of scores is larger for math proficiency, with the lowest recorded math proficiency score in our dataset being 15%, compared to 49% for reading proficiency. There are relatively very few schools with math proficiencies at or below 50%, so we believe our model predictions may be unreliable for these low math proficiency scores. Looking only at 50% and greater math proficiency, we see that the trend of decreasing dropout rate with increasing math/reading proficiency becomes more clear.

Student-Teacher Ratio: We see a similar phenomenon occurring with the student-teacher ratio, where lower student-teacher ratio (seemingly beneficial) is consistently predicting larger dropout rates. In this situation though, what may be happening is that for student-teacher ratios $<$14, the model is trained mostly on small cohorts. Smaller schools tend to have lower student-teacher ratios and smaller graduating classes. This means that the majority of our low student-teacher ratio points are also small cohorts, and are therefore more susceptible to dropout rate inflation mentioned in section III-B. We believe that this dropout rate inflation was learned by our

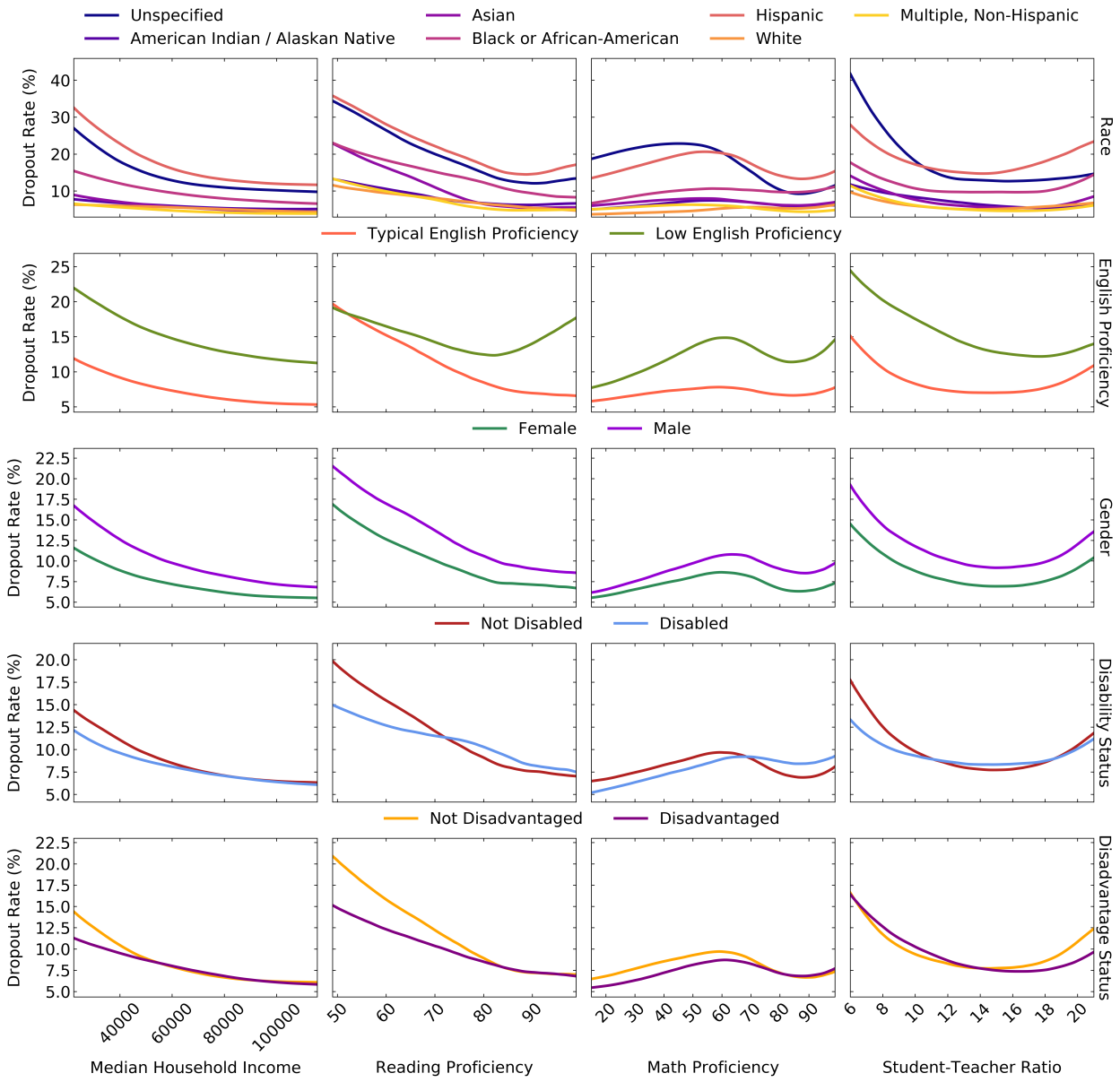

Fig. 3. Predicted dropout rates from our best-performing model across the full parameter space of our data. Each panel shows model predictions varying a given numerical feature (x-axis; same down a column) and a given categorical feature (colored lines; same across a row). Each line represents an average over remaining features as well as an average over cohort size, with the aim being to isolate the effects of varying only the specified features.

model, and is what gave rise to the perplexing trend between student-teacher ratio and dropout rate. Beyond student-teacher ratio of 14, the trend is as expected; smaller class sizes correlate with fewer dropouts.

Disability and Disadvantage Status: We do not see a strong impact on the predicted dropout rate due to disability and disadvantage flags. For both features, our model does not consistently predict higher dropout rates for students who are disadvantaged/disabled vs. those who are not, and the largest discrepancies in predicted dropout rate do not exceed ∼4%.

V. Conclusion

We find that the strongest factor influencing dropout rates is race, followed by regional wealth, and then by reading and English proficiency which we will jointly call language proficiency. A report by Virginia Performs [5] suggests that dropout discrepancies with race could be attributed to other factors such as, e.g., parental education, but that likely the dominant factor is household income. While our model does not find household income to be as predictive as race is, this is potentially due to the fact that we only had access to *median* household incomes on a county-by-county basis, smoothing out variations in household incomes within each county. Had our model been trained on cohort-specific median household incomes instead, it may have been able to rely more heavily on that feature, and household income may have shown up as a more influential feature in our analysis.

While some of the discrepancies with race may be at-

tributable to factors like household income, we believe our analysis indicates that language proficiency—which the Virginia Performs report does not mention—has a very notable impact as well. In addition to the direct discrepancies we observe between LEP students and those with typical English proficiency, the trend is corroborated by the fact that our model consistently predicts Hispanic students to have the highest dropout rate, sometimes as much as 2 times larger than their African-American peers, and 4 times larger than their white peers. While household income is a factor that high schools have no control over, students' ability to comprehend their coursework rests profoundly in the hands of Virginia educators, and we believe the situation can be improved through targeted efforts such as improved English as a Second Language (ESL) courses, or by offering courses taught in Spanish. It is estimated that over 10% of Americans speak Spanish at home [6] (37.6 million people in 2013), but our public education systems currently do not reflect this.

In the future, this type of analysis could be extended to look for graduation trends over time. The VDOE reports improved dropout rates over the past 10 years [1], however it is unclear if all populations of students are seeing a similar increased retention. Additionally, the type of analysis we conducted would be greatly improved in the future using data on individual students provided by the Virginia Longitudinal Data System (VLDS) [7], which is not available to the public, but can be accessed with permission for research purposes. With data on individual students, a classification model could be trained to predict the graduation status of individual students (see section III-A), rather than the regression approach we settled on for predicting cohort dropout rates.

Acknowledgements: Thanks to the support of the VDOE, PSR, and US Census for providing publicly available data, and special thanks to Tod Massa from the VLDS for helpful tips and advice.

REFERENCES

[1] The Virginia Department of Education, "Annual Dropout Statistics for Grades 9-12 by School Division & Ethnicity," *The Virginia Department of Education*. [Online]. Available: doe.virginia.gov/statistics_reports/graduation_completion/dropout_statistics

[2] The Virginia Department of Education, "On-Time Graduation Rate and Cohort Dropout Rate," *The Virginia Department of Education*. [Online]. Available: doe.virginia.gov/statistics_reports/research_data

[3] Public School Review, "Learn about public schools, find schools, analyze data and discuss public school issues," *Public School Review*. [Online]. Available: publicschoolreview.com

[4] US Census, "Quickfacts: Access Local Data," *United States Census Bureau*. [Online]. Available: census.gov/quickfacts

[5] Virginia Performs, "High School Dropout," *The Commonwealth of Virginia*. [Online]. Available: vaperforms.virginia.gov/Education_hsDropout.cfm

[6] Ana Gonzalez-Barrera and Mark Hugo Lopez, "Spanish is the most spoken non-English language in U.S. homes, even among non-Hispanics," *Pew Research Center*. [Online]. Available: www.pewresearch.org/fact-tank/

[7] Virginia Longitudinal Data System, "Cost-effective tool for extracting and analyzing insightful education data within a secure environment," *Virginia Longitudinal Data System*. [Online]. Available: vlds.virginia.gov/