ECO 395M Final Project

Hannah Jones & Joey Herrera

5/10/2021

Abstract

Improving student success via better grades is the focus of every school since better grades afford students a higher chance at getting into college, having better job opportunities, and ultimately having the ability to choose the best path for them. One factor associated with good grades is high parent and family involvement in a students academic journey. Thus, many schools emphasize increasing parent and family involvement to facilitate better student outcomes. This paper seeks to use data from the NHES's parent and family involvement survey to understand which demographic barriers are limiting parent involvement, which sorts of parent involvement have the highest return on investment, and finally, how this information can equip schools to provide equitable programs. This data will be evaluated using principal component analysis and fitting a random forest model to extract meaningful summaries of relevant features. We also look at clustering analysis to see how the students in the dataset cluster based on demographics and experiences.

Using principle components analysis to run a random forest, we are able to build a model that predicts student scores with about 65% accuracy. We take a deeper look at the highly impactful principle components to understand which variables are highly predictive regarding student grades. We find that impactful principle components are highly loaded by parent

race variables, student disability, and school communication.

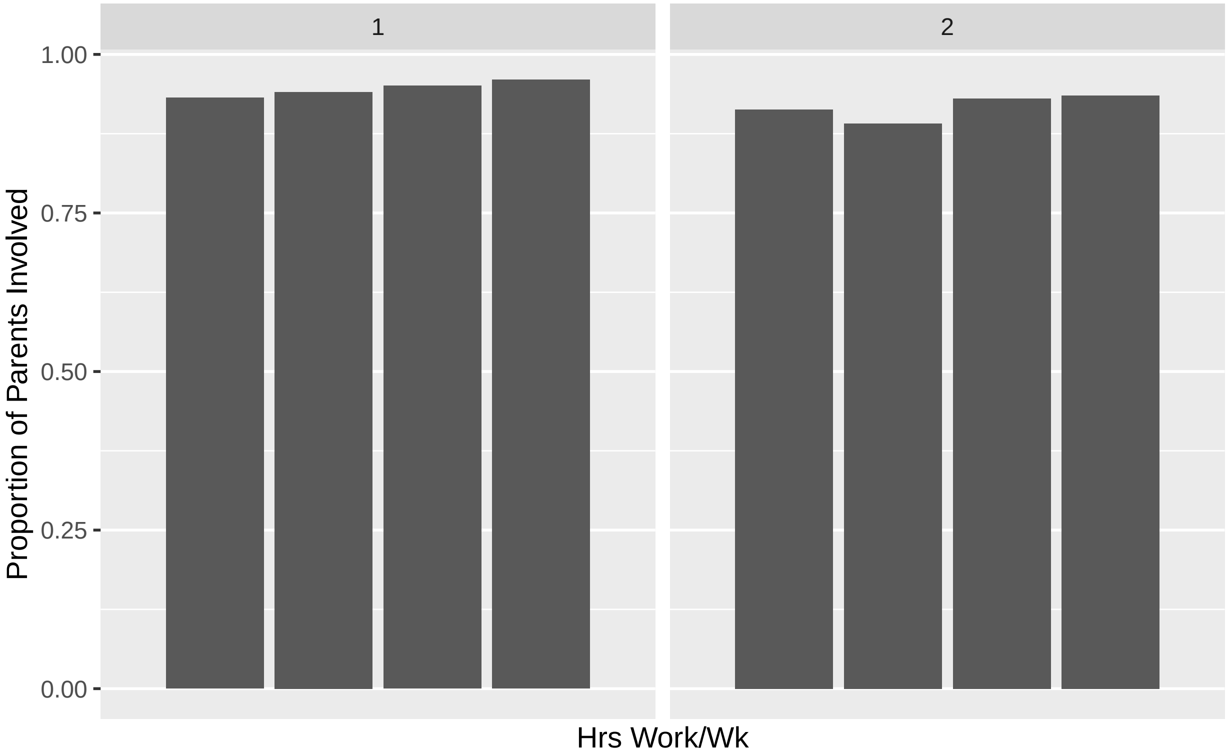
Introduction

Student success, which is traditionally measured through standardized test scores and class grades, is the universal measuring stick for enhanced opportunity in the United States. One significant indicator in predicting student success is parent and family involvement. Since schools only facilitate learning to students throughout the day, a burden is placed on parents and legal guardians to continue cultivating a learning environment outside of school. To increase family engagement, schools have created programs and events to get parents and families involved. These programs and events seem to work well in improving the family involvement for students whose parents work traditional 9-5 jobs. Most families of color and low-income households are not afforded this same opportunity due to work hours, language barriers, or other obligations.

```
##   HourBins P1HRSWK_bins avg_grades avgparentinvolv avg_cultural_exp
## 1        0            0   1.692624       0.9266192        0.7967054
## 2     0-20            1   1.546144       0.9292035        0.8268015
## 3    20-40            2   1.568722       0.9451975        0.8201428
## 4      40+            3   1.495515       0.9559663        0.8496874
##   avg_fam_time meancreateop mean_enjoy
## 1    0.9475852    0.8936728   1.810558
## 2    0.9582807    0.8925411   1.782554
## 3    0.9376184    0.8922898   1.793470
## 4    0.9296004    0.9051373   1.798315
```

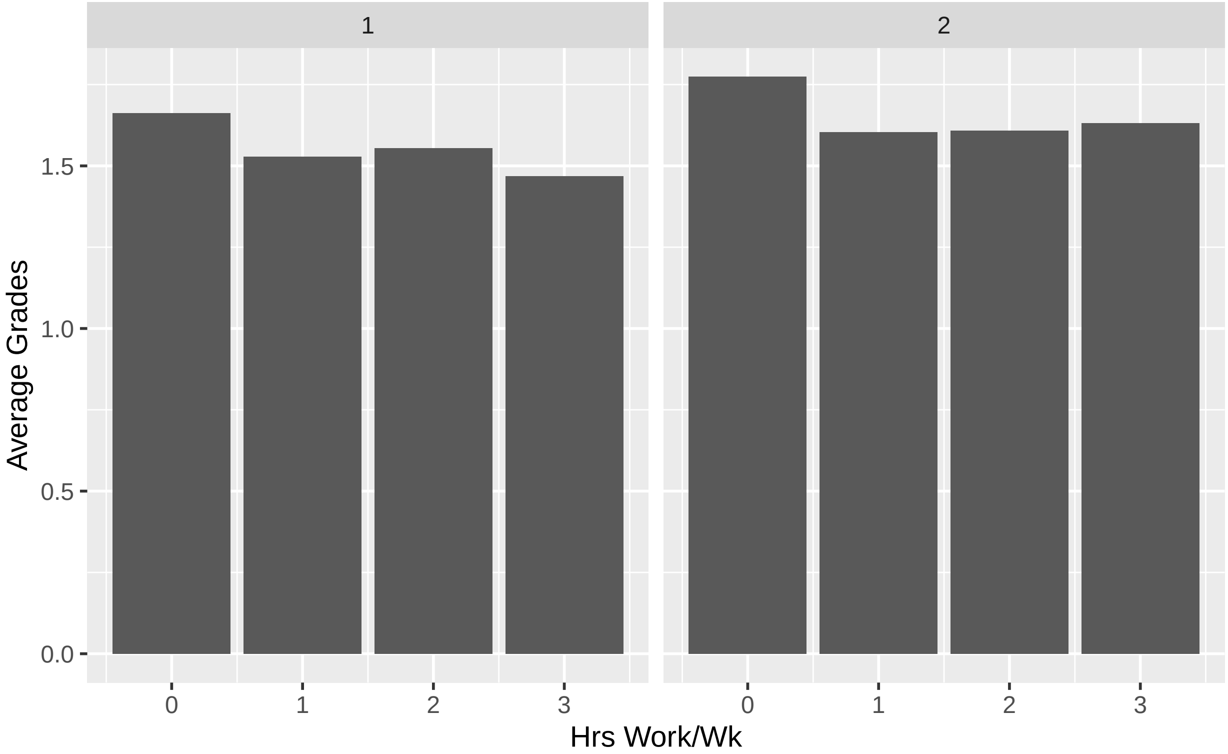
Table 1: Hours Parent Work per Week versus Outcomes

FIgure 1: Proportion of Parents Involved at School



Facets: 1=white, 2=non−white; Y−axis: 1 is A's, 2 is B's; x−axis: increasing from 0 as work hrs increase

Figure 2: Average Grades by Race and Parent Hours of Work



Facets: 1=white, 2=non−white; Y−axis: 1 is A's, 2 is B's; x−axis: increasing from 0 as work hrs increase

```
##            Language english_speakinghh avg_grades avgparentinvolv
## 1 Other non-english                  0   1.600932       0.9029503
## 2           English                  1   1.568900       0.9476955
##   avg_cultural_exp avg_fam_time meancreateop mean_enjoy
## 1        0.7857143    0.9534161    0.8594720   1.670031
## 2        0.8276703    0.9370772    0.8996382   1.810288
```

Table 2: Household Language versus Outcomes

The Tables 1 and 2 above show a summary of grades, proportion of parents involved, proportion of students engaging in cultural experiences (such as plays, zoos, library visits), proportion of students engaging in consistent family time, and average student enjoyment of school. Please note that the grade coding is such that 1 is mostly A's, 2 is mostly B's, 3 is mostly C's and 4 is mostly D's. Likewise, student enjoyment is ranked with 1 as most enjoyment and 4 as least enjoyment of school. Figures 1 and 2 show that at each level of parent work hours, white parents tend to be more involved in school activities than non-white parents. Similarly, grades for students of white parents are better for every level of parent work. This suggests that there is some sort of breakdown between schools and parents of color when it comes to getting involved and helping students succeed. The second chart shows the gap in student grades, parent involvement, cultural exposure, creative pursuits and enjoyment between households that are English speaking and households that are not English speaking. This finding emphasizes the gap in experiences afforded to white, English speaking students, and non-white students who do not live in English-speaking households

Creating an equitable array of programs to increase parent and family involvement for families of different races, ethnicities, and socioeconomic statuses is a growing issue. In response, the National Household Education Surveys (NHES) program created a national survey with several hundred questions focusing on different aspects of parent and family engagement, This survey is sent to households across the United States, and encourages parents fill it out and return their responses. The NHES typically receives around 50 million survey responses each time the survey is sent out.

This paper seeks to provide basic program recommendations to schools by evaluating the impact of indicators derived from the NHES parent and family involvement (ppfi) survey on student grades. Using a 1% sample of the full data set, this paper will use a principal component analysis (PCA) and random forest forest technique to illustrate the effects of different survey questions (or features) on student grades. We hope to see which demographic characteristics are highly predictive in success or lack of success, and how this information can be used to target students at risk of falling behind.

The following sections of this paper will address the empirical methodology, results, and concluding remarks associated with the research problem.

Methods

As mentioned in the prior section, the data used to evaluate student grades comes from a sample of the NHES's 2019 parent and family involvement survey. The data extract includes approximately 17,000 observations and over 800 different features. Each observation is associate with an individual survey response that has been given a unique identifier using a mix of numbers and letters to avoid providing researchers with sensitive information. Each feature corresponds to either a question on the ppfi survey or an additional variable added for NHES's weighting purposes. Every question on the survey allows respondents to bubble in answers. Thus, each corresponding feature is a factor variable, and every answer is represented as a a unique number within the question/ feature.

In order to find the optimal number of principal components and fit a random forest model to the data, the data must be cleaned and morphed into a format that PCA and random forest functions accept. The first step of the cleaning process involves filtering out any variables that are irrelevant the research problem. In this case, variables regarding home schooling and virtual learning. This project focuses specifically on recommendations of schools (either public, magnet, or charter) and their in-person classes. Next, all observations containing a "-1" were recoded. If an observation included a "-1" for an answer to a question, then they were deemed by the NHES to have a valid reason for skipping the question. These observations were recoded to "0" for simplicity. Additionally, we engineered features to compress like-variables into one overall variable. For example, instead of having separate variables for brothers and sisters, we created one variable for siblings. We also combined like variables around extracurricular activity involvement, parent involvement in school activities, and other cultural activities. It is important to note that the NHES serves as a filter for potentially corrupt data because they censor sensitive data and do not include incorrectly filled in survey responses in their data. For instance, a family's student cannot be in two different grades at the same time. Observation such as the one described above were removed from the data before the sample was extracted.

We run a principle component analysis to simplify our dataset of ~230 variables into a more manageable set of explanatory variables that we can study more closely. To run the PCA, we remove our eventual outcome variable, grades, in order to make this exercise unsupervised. We then add the grades back in to the data and use the principle components in building a random forest model on a training set of data, to then test on a test set. This model allows us to predict student outcomes, and analyze which principle components are most influential for prediction. Taking a closer look at the principle components gives us insight into the groups of variables that are highly predictive for student outcomes. This will enable us to test our hypothesis of parent involvement's importance in student success, and the various barriers to student success.

Results

Running the principle components analysis on the dataset yields a total of 235 principle components. Using 60 principle components explains ~75% of the variance in the data, so we will move forward building the random forest model with the 60 principle components predicting the student grades.

```
##                    Accuracy    Outcome
## 1 Total Predicted Correct 2277.00000
## 2         Percent Correct   65.05714
```

Table 3: Accuracy of Random Forest Model

Table 3 above reflect the total predicted correctly out of the test set, and the percent correct. Figure 3 below is the variable importance plot for the principle components. I have only shown the first 10 principle components.

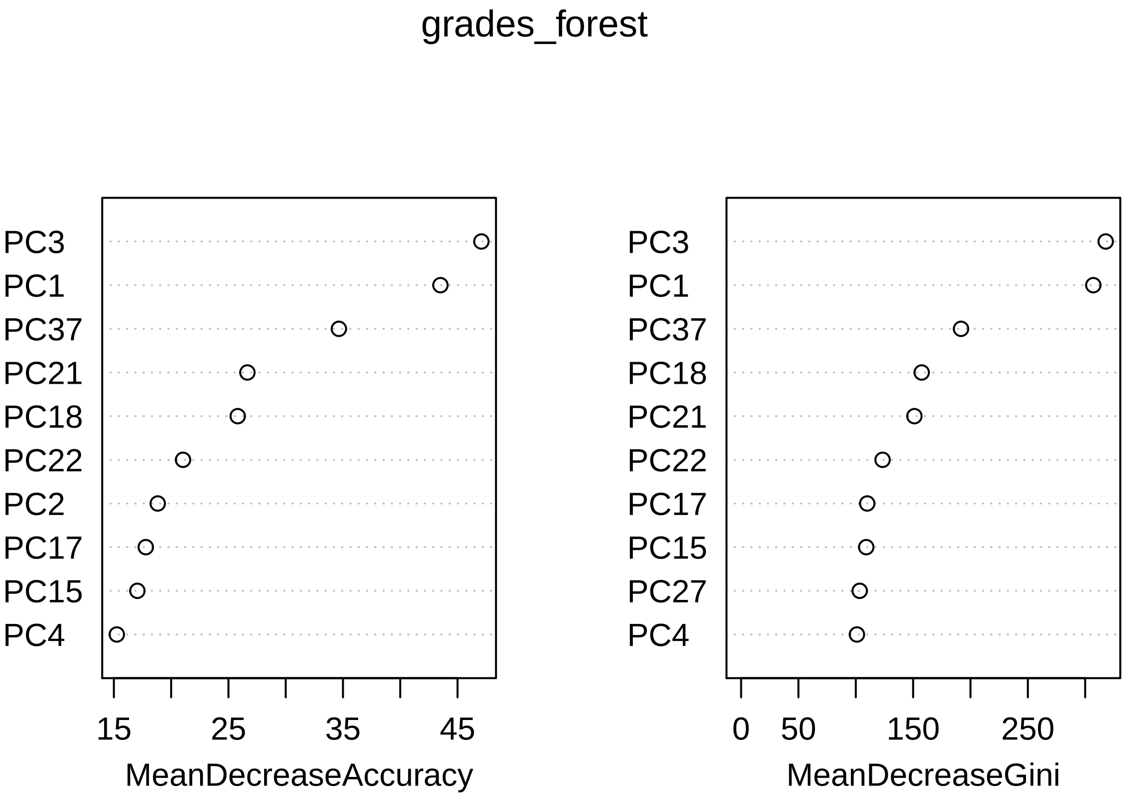


Figure 3: Variable Importance Plot for Random Forest

Principle component 1 is by and large the most impactful component, followed by PC 3, 37, 21, and 18. We will look at the implications of this in the conclusion.

Conclusion

The prediction model discussed above runs at about 65-70% accuracy. The lack of accuracy may be due to students who are atypical based on their demographics and parent involvement. The survey is extensive but may also suffer from bias as all information is self-reported. Taking a closer look at the top principle components against student grades in Table 4 gives an idea of how the principle components relate to grades. PC1 and 3 are highly negatively correlated with students making C's and D's. PC3, PC21 and PC18 all seem to be positively correlated with good grades (A's). PC37 is slightly negatively correlated with A's.

```
##       Grades ppfi_feateng.SEGRADES     avgpc1     avgpc3    avgpc37    avgpc21
## 1 Mostly A's                     1  0.8971885  0.9478610 -0.1466356  0.1834275
## 2 Mostly B's                     2 -0.7195359 -0.6333339  0.1323043 -0.1336767
## 3 Mostly C's                     3 -2.1529400 -2.5405975  0.3437575 -0.4466325
## 4 Mostly D's                     4 -3.4354434 -4.4526047  0.3498385 -0.9005014
##      avgpc18
## 1  0.1592221
## 2 -0.1023534
## 3 -0.4516128
## 4 -0.6903633
```

Table 4: Top 5 principle components and student grades

Looking more closely at the principle components, we can see what are the primary components of the principle components.

```
##    headPC1   headPC3          headPC37 headPC21 headPC18
## 1   P2PACI    DSBLTY            HDPDDX   CLIVYN   FHHOME
## 2  P2AMIND    HDADDX         HDAUTISMX SEGBEHAV FSSPCOUR
## 3 P2HISPRM  HDLEARNX          SEGRADEQ  SEGWORK FSSPROLE
## 4   P2ENRL SEFUTUREX   internet_access  FHWKHRS   FSSPHW
## 5  P2BLACK HDSPEECHX school_choice_pos     CSEX HDDELAYX
```

Table 5: Head Loadings of Principle Components

An important note about the data in this survey is that apart from the engineered variables, all yes/no questions code a 1 for yes and a 2 for no, instead of the typical 0 for no and 1 for yes. Therefore, all associations are a bit counter intuitive, since the larger value for many variables implies the absence of the trait. Taking this into account, Table 5 shows the head variables in PC1 revolve around the second parent's race. This is an important finding as it relates to providing extra support for students who are in a single-parent household, and extra attention for non-white households. The head variables for PC3 revolve around students with disabilities. PC37 also involves students with disabilities, as well qualitative work description, internet access, and school choice empowerment. PC21 involves around parent contact for good student work, time spent doing homework, and the child's sex. Finally PC18 involves time doing homework, and school communication about coursework, parent roles, and homework help resources.

These findings suggest a number of interesting possible interventions to help parents help their children. These variables confirm that non-white households, disabilities, child behavior, and school communication are all predictive in student success. This finding bolsters the earlier discussion around providing extra support to families who are not white and english speaking. As shown earlier, these families are likely to be less involved, which may hurt student grades. The findings around school communication about parent roles, and homework help provides an action step for schools to start bolstering family involvement. Additionally, schools should provide extra support for families of students with learning disabilities to account for the large impact of these disabilities on student success. Finally, school communications around student successes may not be causal of further student success, and may just be a product of success, but the findings around school communication and parent involvement suggest this is a helpful practice.

Given the importance of family involvement in student success as asserted by previous research and findings within the data, schools have put resources into creating programs that address the specific needs based on the demographic and socioeconomic information of student families. In particular, low-income and minority parents are less likely to be involved with their students' education in a traditional sense. Various structural factors including working multiple jobs or obstructive job hours. To provide these demographics with alternate opportunities schools have provided multiple opportunities for involvement, value different perspectives, and reach out to families in culturally appropriate ways (Auerbach, 2007). Schools must continue to adapt to their individual demographics because traditional concepts of family involvement limit the opportunity for low-income and minority-families to be involved. Thus, those students are likely to have less academic success than their counterparts. As our findings suggest, race and disabilities are highly impactful on student success and should be greatly considered when schools design programs to involve students and their families in meaningful ways.

An efficient way to provide recommendations to schools and policymakers is through the involvement of machine learning techniques. These data science skills have the power to delineate trends from large data sets, such as data revolving around the number of students in a school, in an interpretable manner. Besides predicting student success, there are a variety of places that machine learning can enhance the education sector. For instance, creating adaptive learning for people with different learning styles, improving efficiency in the learning environment, and assessment evaluation are just a few ways machine learning can postivly impact the education sector (Ryan, 2020).

References

Auerbach, Susan. “Visioning Parent Engagement in Urban Schools.” Journal of School Leadership, vol. 17, no. 6, Nov. 2007, pp. 699–734, doi:10.1177/105268460701700602.

Ryan, Jacob. “The Importance of Machine Learning in the Education Sector.” NewGenApps, 20 Dec. 2020, www.newgenapps.com/blog/the-importance-of-machine-learning-in-the-education-sector/