Exploring Data Science Techniques to Identify at Risk Students

Parthvi Bhutani (pb1967), Jenna Eubank (jke261), Shaan Khosla (sk8520),

Mingzhou (Randy) Zhu (mz1770)

Business Understanding

The Problem

In the past fifteen years high school dropout rates have largely decreased in the United States, however, there continues to be a significant disparity among demographic groups. Black and Hispanic students are less likely to graduate than their white counterparts and students in urban centers are significantly less likely to graduate than their suburban neighbors, regardless of race [1]. For example, in 2009 only 38% of high school freshmen living in the city of Cleveland would go on to graduate within four years compared to the surrounding suburbs, which had a graduation rate of 80%[2].

Two key factors recognized as early indicators of dropout before graduation are a student having a 9th grade attendance rate below 80% and/or 9th grade GPA below 2.0 [3]. While attendance rates can be tracked in more real time and are already collected and monitored by high schools, we looked to find a model that could predict a failing 9th grade GPA before it happens and could trigger additional preventative measures.

While there is significant research from academic and government institutions, models have not reached school districts at scale for ongoing management of student populations at the high school level [4]. Predictive models on student success are more commonly implemented in higher education, partially because of the financial incentives for student retention through a two or four year program.

As of fall 2020 there are a projected 56.4 million students enrolled in elementary, middle, and high schools in the United States [5] and with the growing complexity of remote learning it is more important than ever to keep at risk students engaged with education. A high school education has a meaningful impact on long-term opportunities. According to the Census Bureau adults without a high school diploma earned less annually and were in worse health than adults with higher education levels [6]. An improvement in educational outcomes can have a significant impact on both society as a whole and quality of life for individuals on a large scale.

Data Understanding

Our focus is on identifying at risk students, specifically by finding early indicators of a failing 9th grade GPA in the hopes of early intervention. We selected two datasets with student demographic information and school performance results. Dataset 1 (referred to as "existing school dataset") was acquired from Kaggle [7] and has math, reading, and writing scores five features for 1,000 students. The second dataset (referred to as "surveyed dataset") [8], sourced from a university study and made available through UC Irvine, includes twenty-two features with an increased focus on student background and personal activities. This data differs in that it includes features that were collected by questionnaires and could not be obtained by the school without participation from the students. While this does introduce self selection bias, we proceeded under the assumption that a thorough survey process in a school environment would net a high response rate.

We began our analysis with just the existing school dataset because it used features that would be available to any school without student engagement. This would allow schools to implement the model with statistics they have on hand, or could generate independently. The surveyed dataset was more complex so was identified later to improve model results. However, since these data were not from the same student population and had different features they were used to compare the model performance and understand the value that fielding a student survey might bring.

Target variables

Both datasets were selected since they included scores for final grades that could be used as target variables. The existing school dataset math, reading, and writing scores were evaluated as target variables but also as features in potentially predicting a different score (e.g. using the writing score to predict reading outcomes). The raw scores were rounded to bins of 10 (e.g. an 84 was encoded as an 80). For the surveyed dataset the final grade, G3, was ranged 0 - 20. Since our aim is to anticipate student failures and the score range was smaller we converted raw scores to binary variables of pass (1 where the score >= 12) or fail (0 where the score < 12) based on the criteria outlined in the dataset.

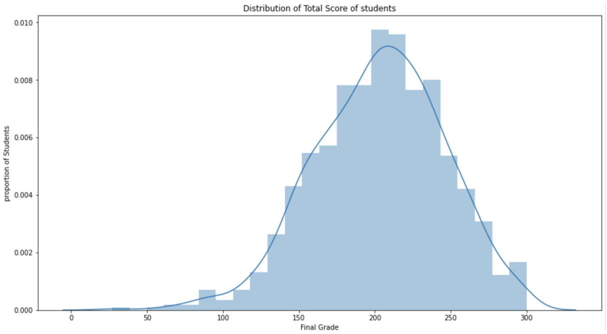
Data Preparation and Exploratory Data Analysis

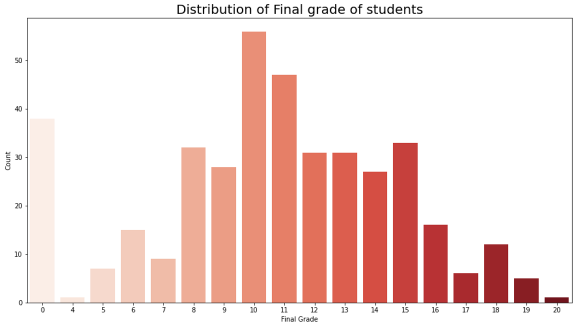
Cleaning and preparation

Both datasets included binary, categorical, and numeric features which needed to be prepared for modeling. Some categorical variables were already integer encoded but others needed conversion (e.g. rural = r, urban = u). There were only two numeric features in the surveyed dataset, absences and age, that were left as is. We also checked for null values and outliers in numeric variables, particularly absences, and needed no revisions for either factor.

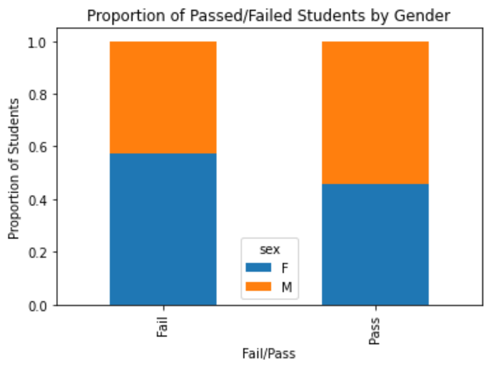
Exploratory data analysis

Since neither dataset was particularly large, we were able to easily perform extensive EDA on both. Our intention for evaluation was not to combine the two datasets but compare the results, so it was important to see if they had similar distributions for shared features and final scores. Results for the existing school dataset (left below) were more normally distributed while the surveyed data (right below) was skewed slightly right and had a lower relative mean score. The higher number of zero scores in the surveyed data would also have to be considered during analysis.



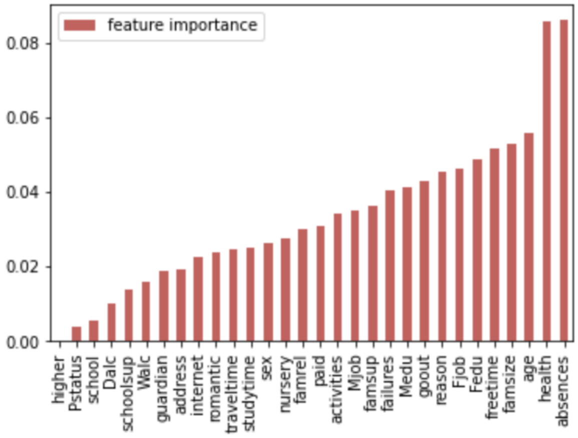


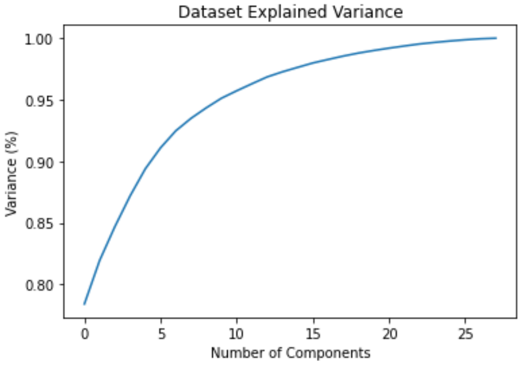
In addition to comparing the two datasets we looked to see that the results aligned with features stated by external studies to correlate with lower grades. Potential differences in achievement between female and male students have always been an interesting topic in educational research, as well as having policy and economic implications. The persistent trend of an overrepresentation of male students in the group of high-achievers in both mathematics and science is striking.[9] In our data 60% of all failing students were females. We found that in our data absences were highly related to failing grades (appendix figure 5) as was also supported by a study conducted in Oregon schools [3]. The common view is that more educated parents provide an environment, which improves their children's opportunities and decision processes. This assumption was, for example, the base of World Bank programmes to improve female education with evidence that more educated mothers have healthier children [10]. We observed a substantial increase in student grades as the mother's education improved (appendix figure 4).



Feature selection

The existing school dataset only had five features so there was no need to remove any for the analysis, however the surveyed dataset included twenty-two so required exploration. We first looked at feature correlations (appendix figures 2 and 3) and ran the data through a decision tree to get feature importance.





We then applied PCA on the dataset to extract important features. Ten principal components explain more than 95% of total variance. We chose the first 10 principal components to be our final features to be trained on. Mother's job and father's job were also dropped because they were strongly correlated with mother's education and father's education.

Modeling and Evaluation

Baseline models and problems with the existing school dataset

For the existing school dataset we first broke the target variable into multiple bins: based on initial testing predicting the exact score would not be possible. We then tried two classification methods: a Decision Tree and Logistic Regression. Even when attempting to predict a ten point grade range these methods performed very poorly with an AUC score of .044 for the total score prediction and scores of .208 - .232 for predicting the individual math, reading and writing scores with a Decision Tree. We believe this was due to an insufficient number of features and that all features are categorical. As a result we deemed this data would not support further analysis and that we should seek a more complex set of features. We selected our secondary dataset at this stage, the surveyed dataset. Although it has fewer students we thought it would still be more suitable for model building.

On this new dataset, we decided to use a Decision Tree as a baseline model to compare our other models with. The AUC of this baseline model was .6061. We also decided to use AUC, which is base rate invariant, as a metric for evaluation due to the class imbalances in the dataset. This is because using accuracy would give a biased view of how well our classifiers are actually performing.

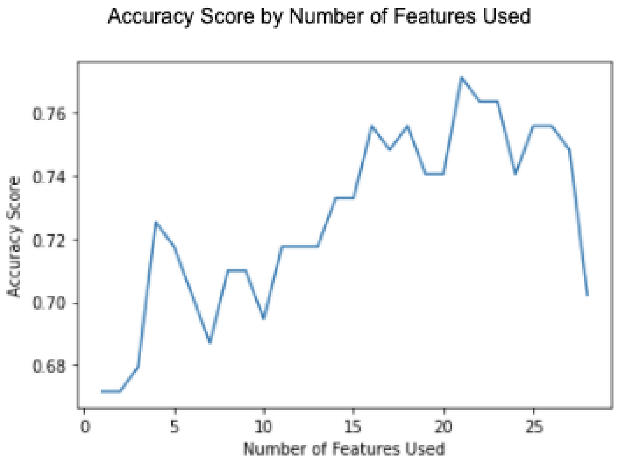
Model 1: Logistic Regression

We hypothesized logistic regression would be particularly well suited for this dataset since it is robust with small sample sizes and imbalanced classes. Since our dataset only contains 395 students with roughly ⅔ of the instances belonging to a single class it fits well within logistic regressions strengths.

The hyperparameters we tested were mostly default aside from C. C is the inverse of regularization strength and we decided this was a necessary hyperparameter to test a range of values in the grid search for a number of reasons. We used the range of [.01, .1, .5, 1] as the C values in our grid search. Primarily, we knew that since our dataset was small, we would be susceptible to large amounts of variance. To counteract this, regularization can specify that the norm of the weights can’t grow past a certain point, thereby reducing estimation variance.

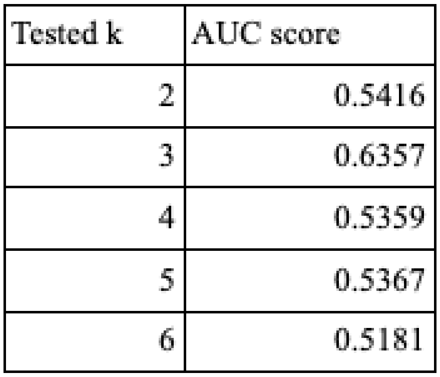
Model 2: Random Forest

The next model we hypothesized would work well on our dataset and for our business problem was a Random Forest. The primary reason was because of the high degree of variance that comes with a smaller sample size. Since Random Forests use bagging to reduce variance, we knew that this would outperform our baseline model of a Decision Tree. We tested various n_estimators as the main parameter in our cross validation to optimize this reduction. We also hypothesized that it is better to have too many trees rather than too little, and to test the out of sample error. For this reason, the upper limit of n_estimators was 750 while the lower limit was 10. Within this model we ran a Forward Stepwise Selection algorithm, iteratively adding one feature at a time, depending on which one improved the out-of-sample AUC the most. Ultimately, we did not use this type of feature selection and we chose to use PCA due to the better results with KNN.



Model 3 K-Nearest Neighbors

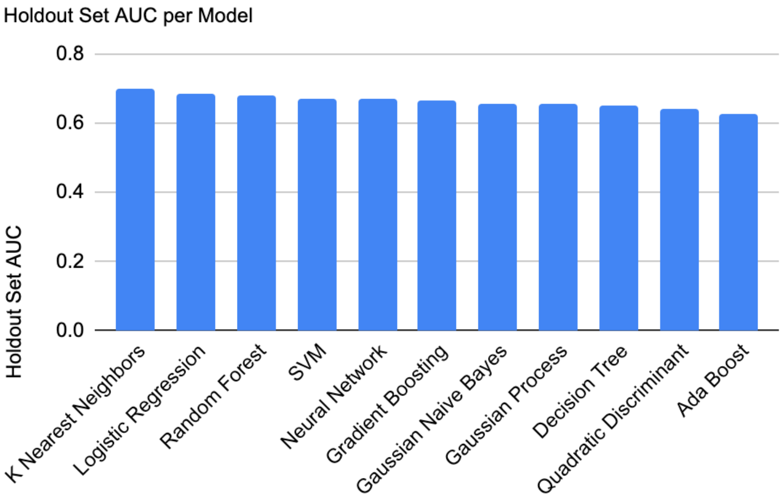
Our final candidate model was K-Nearest Neighbors, selected based on the assumption that students that have similar characteristics, likely will perform similarly on exams. We also appreciated the fact that KNN can easily give us probability estimates. Some downsides of KNN include that dimensionality reduction or feature selection is critical as the curse of dimensionality makes it so that all points are equally as far away. This extra step of dimensionality reduction adds complexity to our model, but this was needed in this case. Another downside is that scoring a new instance, or making a prediction, can be a slower algorithm to execute when finding the distance between every instance in the dataset. This, however, was a lower priority for our problem as there is no need for instant predictions. The most important parameter for KNN is k so we ran an extensive grid search to ensure that this metric was fully optimized. Too many nearest neighbors will result in a simple average of a large part of the dataset, while too few neighbors gives us poor generalizability. Additionally, we tested several distance metrics since we had categorical features that were label encoded, so a Euclidean distance metric would not make much sense here. The algorithms we tried as parameters were “auto”, “ball_tree”, “kd_brute”, and “brute”. The last parameter we adjusted was distance weights, which significantly outperformed uniform weights.



Tested k	AUC score
2	0.5416
3	0.6357
4	0.5359
5	0.5367
6	0.5181

Grid Search

Finally, we utilized SciKit Learn’s GridSearchCV package to run a grid search of eight models in addition to our selected list (full list of models and performance metrics available in the appendix; figure 1). These other models included Neural Nets, AdaBoost, and Gradient Boosting Classifier to name just a few. Although we tested these, we were hesitant about their success due to adding too much model complexity for such a small dataset. Since our dataset was so small, we were fortunate in not being heavily constrained by model train and test time. To give some reference, we were



able to use the GridSearch package to investigate a total of 11 models and a set of their respective hyperparameters in under a few minutes.

One downside of using this method to training and testing our models was that cross validation results in an even smaller sample size to train the model on. This created model instability but overall we did not observe any serious negative consequences. For clarification, the final model that we will put into production for grade predictions will be trained on the entire dataset that we have available, which should only increase the predictive power of our model. This is because although more data will not affect the bias of our model, it will likely reduce the variance.

Final model selection

Model	**Holdout Set AUC**
K Nearest Neighbors	.702
Logistic Regression	.687
Random Forest	.679

Out of the 3 closely evaluated models, we would recommend selecting KNN. There are two reasons: first, KNN has the highest AUC and the lowest model complexity. It outperforms Logistic Regression and Random Forest with a 70.2% AUC. Secondly, KNN is easy to interpret. Essentially, KNN decides if a student would pass the exam based on the results of a certain number of students with similar backgrounds.

The hyperparameter selection step tells us the best K in this case is 3. As a result, whether a student would pass the exam is decided by 3 other students with the most similar backgrounds. This fits common sense: students whose parents are well educated and from wealthy families are more possible to succeed and vice versa. We discarded the other models because we only gain marginal increases in performance but have to deal with much more model complexity. A table with more specific results can be found in the appendix.

Deployment and Considerations

Model Deployment

Use and optimization of this model faces a few challenges including data collection and a long feedback cycle. Since KNN is an instance based model, we would have to worry about computational costs of predicting new instances due to having to calculate the distance of a new instance to every instance in our dataset. However, in this business case that is not a relevant constraint since there is no time constraint on making predictions.

Due to the specificity of some of the features, it is unlikely that any new participating school would have the data already available. Records on lifestyle factors such as family and romantic relationships, health status, alcohol consumption, etc. are likely not captured and kept in a structured way for all students. As a result, there would be no way to test the efficacy of the model in a new school system.

Additionally, collecting information on some of the features in the model would require student participation and consent. In this scenario there is a high potential for self selection bias. Students who opt to participate in the study may have higher engagement with the school and may be more likely to graduate than students who do not participate. Although, this new information will keep our dataset up to date with any new changes.

A full test of the model’s predictions in a real life test with a new cohort would take more than four years from data collection to study completion when the students graduate. With graduation only occurring once a year it also gives fewer opportunities for model optimization.

Ethical Implications

Although identifying students at risk of failing can help identify a need for additional support, it can cause potential problems for both high risk and low risk students. Classification may lead to different treatment from teachers, administration, or guardians, both in and out of the classroom. For example, placing a student in a remedial classroom based solely on demographic characteristics would be entirely unethical. Conversely, if the school does not offer the same support to students identified as lower risk it could exclude them from opportunities. The type and method of additional assistance would have to be carefully crafted, potentially on a feature-by-feature basis for each school environment.

Conclusion and future work

This exploration does indicate that machine learning might be able to aid in the identification of at risk students but still faces some challenges. The features evaluated here can be part of future model development that could improve over current methods. An interesting avenue of exploration could include the development of data collection processes that would make these models more scalable across multiple schools and an increased number of students. While privacy and ethics concerns should not be ignored, models of this nature would have the potential to assist in the identification and support of at risk students.

We noticed that all 3 candidate models don't perform perfectly and we think there are mainly two reasons: lack of data and singleness of algorithm. Our dataset has less than 1000 observations. In the future if we can obtain more data to be trained on, the final accuracy would inevitably increase. In addition, we can try bagging or boosting of multiple models instead of a single model. A combination of multiple models may lead to a much more accurate result. More powerful algorithms such as neural networks are also a good choice to look into.

References

[1] The Condition of Education - Preprimary, Elementary, and Secondary Education - High School Completion - Status Dropout Rates - Indicator May (2020). Accessed December 5, 2020. https://nces.ed.gov/programs/coe/indicator_coj.asp.

[2] Dillon, Sam. “Large Urban-Suburban Gap Seen in Graduation Rates.” The New York Times. The New York Times, April 22, 2009. https://www.nytimes.com/2009/04/22/education/22dropout.html.

[3] Burke, A. (2015). Early identification of high school graduation outcomes in Oregon Leadership Network schools (REL 2015–079). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Northwest. Retrieved from http://ies.ed.gov/ncee/edlabs.

[4] Dutt, Ashish, Maizatul Akmar Ismail, and Tutut Herawan. "A systematic review on educational data mining." *IEEE Access* 5 (2017): 15991-16005.

[5] Snyder, Thomas D., and Sally A. Dillow. *Digest of education statistics 2011*. National Center for Education Statistics, 2012.

[6] Chapman, Chris, Jennifer Laird, Nicole Ifill, and Angelina KewalRamani. "Trends in High School Dropout and Completion Rates in the United States: 1972-2009. Compendium Report. NCES 2012-006." *National Center for Education Statistics* (2011).

[7] Jakki Seshapanpu. (2018, November). Students Performance in Exams, Version 1. Retrieved October 5, 2020 from https://www.kaggle.com/spscientist/students-performance-in-exams.

[8] P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.

[9] Meinck, S., Brese, F. Trends in gender gaps: using 20 years of evidence from TIMSS. *Large-scale Assess Educ* 7, 8 (2019). https://doi.org/10.1186/s40536-019-0076-3

[10] Chevalier, Arnaud. "Parental education and child's education: A natural experiment." (2004).

Appendix

Team Contributions

Group Member	Contributions
Shaan Khosla	● Grid Search with cross validation of 11 models ● Random Forest Forward Stepwise Feature Selection ● Wrote analysis for model 1,2, and 3 in paper ● Final model evaluation and write up
Jenna Eubank	● Data cleaning and preparation ● Decision Trees ● Business Understanding ● Data Understanding ● Model Deployment
Mingzhou (Randy) Zhu	● Clean and preprocess of dataset 2 ● PCA on dataset 2 ● Train and tune KNN & Logistic Regression models ● Part 6 of the report
Parthvi Bhutani	● Data Understanding ● Data Preparation ● Exploratory Data Analysis ● Data Source Documentation

Supplementary Data and Visualizations

Figure 1: Grid Search Results

Model	**Holdout Set AUC**
K Nearest Neighbors	.702
Logistic Regression	.687
Random Forest	.679
SVM	.672
Neural Network	.672
Gradient Boosting	.664
Gaussian Naive Bayes	.656
Gaussian Process	.656
Decision Tree	.649
Quadratic Discriminant Analysis	.641
Ada Boost	.626

Figure 2: Final score feature correlation

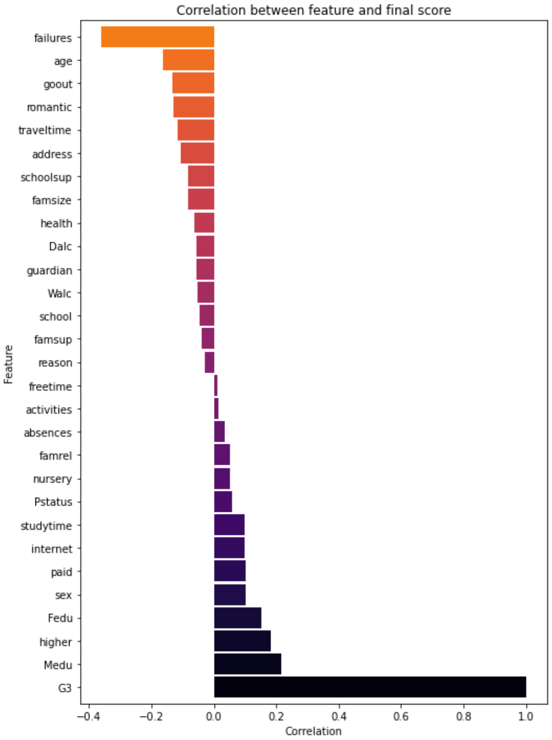


Figure 3: Feature correlation heatmap for survey dataset

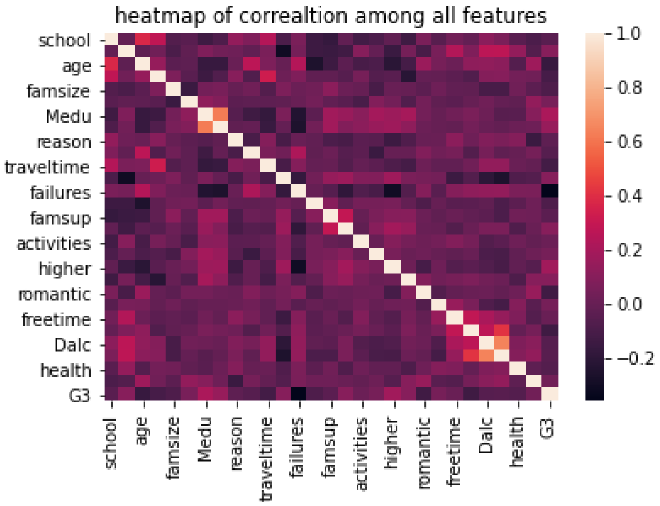


Figure 4: Final grade by mother's education

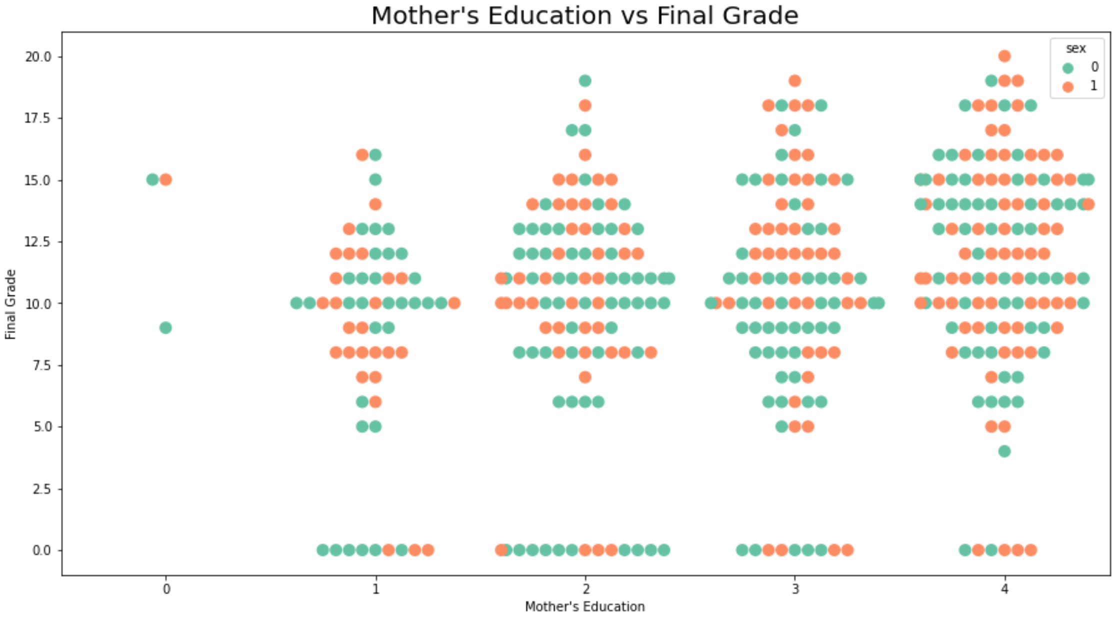
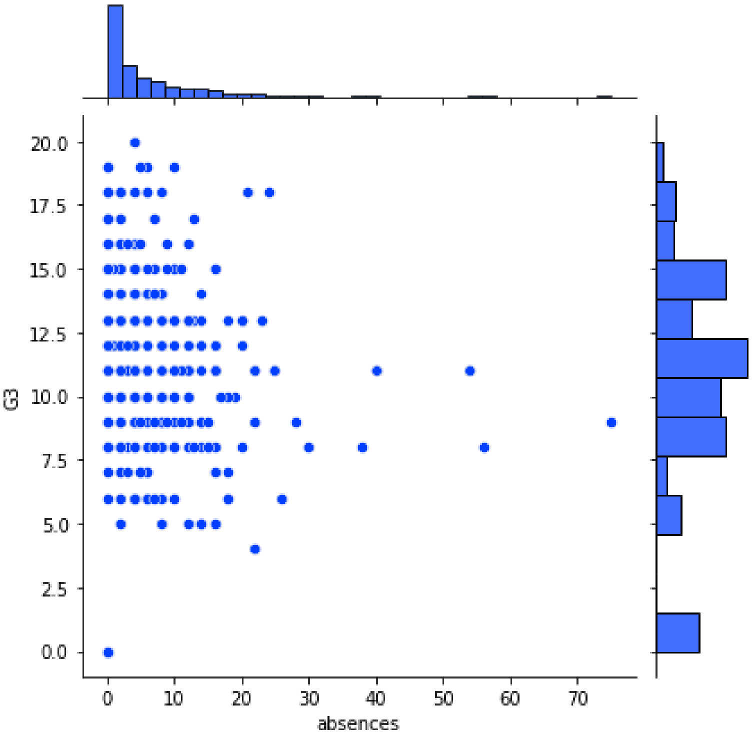


Figure 5: Final Grade Distribution by the number of absences



Code used for this research can be found in our https://github.com/mingzhouzhu22/Risk-Students-Score-Prediction