Grade Predictor and Intervention Suggestion Model

Dataset

Student Grade Prediction

The data is taken from a Kaggle project proposal but I hope to design this project to mirror the structure of many Student Information Systems. This process and data structure should seem plausible to many admissions, counseling, and learning professionals.

Problem

The era of benchmarking learner potential by a final grade may very soon be coming to an end (reference and reference). However, numerical models to inform educator, coach, and counselor intervention will continue to be an domain of research, discovery, and innovation (reference). This project aims to investigate the potential of estimating learner performance from factors extrinsic to the curriculum.

> **Can we accurately forecast student performance through extracurricular and socioeconomic factors?**

Client

The client of this project will be the educators, coaches, counselors, and administrators who strive to provide the most meaningful and data-driven experience to their learners. With timely statistics and data summaries, these members of a learner's support system can intervene with more purpose and precision. The vast majority of individuals working within the education system will agree that they operate within an information rich environment but that data is largely siloed to the individual who gathers the data. They may also concede that the majority of that data is misunderstood and underrepresented in the process of supporting the *individual learner*. High level data analysis continues to be a driving force in budgetary and staffing decisions but those business innovations now need to trickle down and start informing individual student progress.

Project Overview

This project will begin by assessing the statistical significance of utilizing environmental and socioeconomic factors to estimate student progress. This stage of the investigation will inform the building of a intervention model based on the most influential factors from this dataset. This process would be analogous to one taken by a technology or data specialist within an organization; to properly assess the significance of their data collection and analysis.

This project and process should be viewed as a microcosm of a potential application within a larger education system and vastly larger dataset.

Deliverables

This project will deliver code and the promise of a data-driven process to support timely learner intervention.

An extension of this project could be the creation of a web portal that is accessible by learners and the members of their educational support system.

Dataset Cleaning

The major cleaning task with this dataset is understanding and handling the zero values in the final grade column. It was found, through exploratory data analysis (EDA) that these zero values were held exclusively by students who also had zeros indicated for their second period (second trimester) grade. This leads me to believe that the student had sufficient data to be included in the dataset. Namely, these missing final grades were records that did include a period one (first trimester) grade. I attribute these zero final grades to student attrition from the schools and therefore removed them from the dataset. These zero values would greatly influence the training of a model and weigh strongly on summary statistics.

Another major task for data transformation is the encoding of all categorical variables which is handled in the machine learning pipeline, using the OneHotEncoder from the Scikit-learn Preprocessing module.

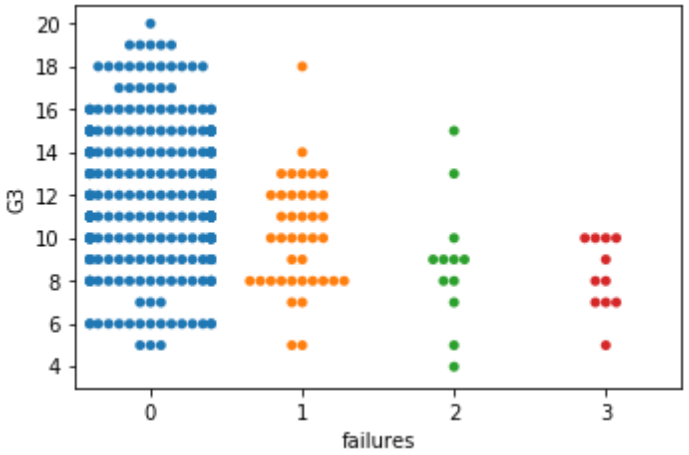
Two features are engineered. The first is the “passed” variable, which is a binary representation of the final grades. A threshold is placed at ten marks out of twenty. The second is the “improved” variable, which is also binary, where a value of 1 indicates that the student improved their grade from period one to period two and a value of 0 indicates that they did not improve.

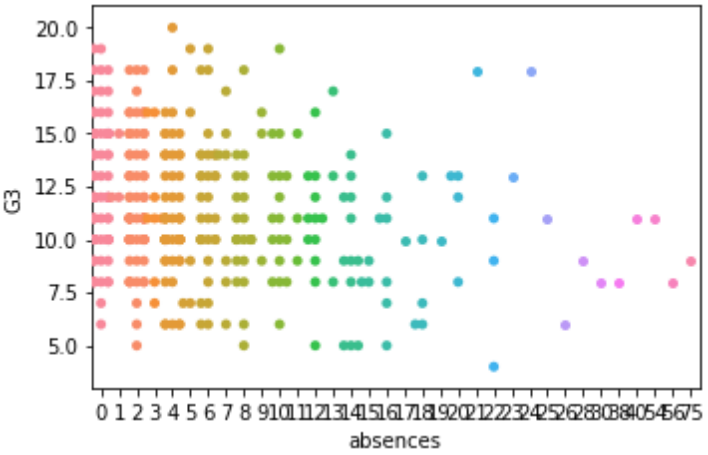
The “passed” variable will be used in a logistic regression model to produce the predicted probability of receiving a grade of 10 or higher on the final based on only external factors.

The “improved” variable will be used in conjunction with the period one and two grades to generate a grade prediction model. This model will be used after the period two grades are registered.

Statistics and Visualizations

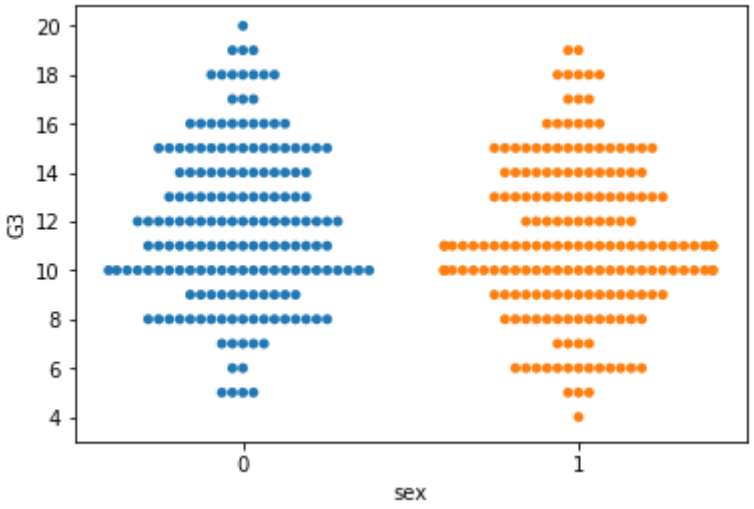
While performing EDA, all variables were explored for potential correlation with final grades (G3).





Features including, the number of failures and absences showed a negative correlation with final grades, which is to be expected. Failures had the strongest negative correlation with a pearson coefficient of -0.29.

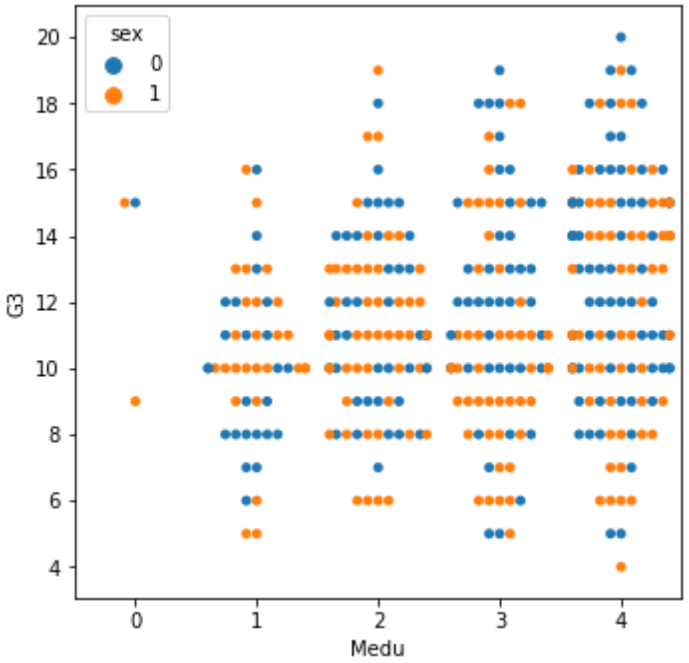
During EDA there was an emergent story.

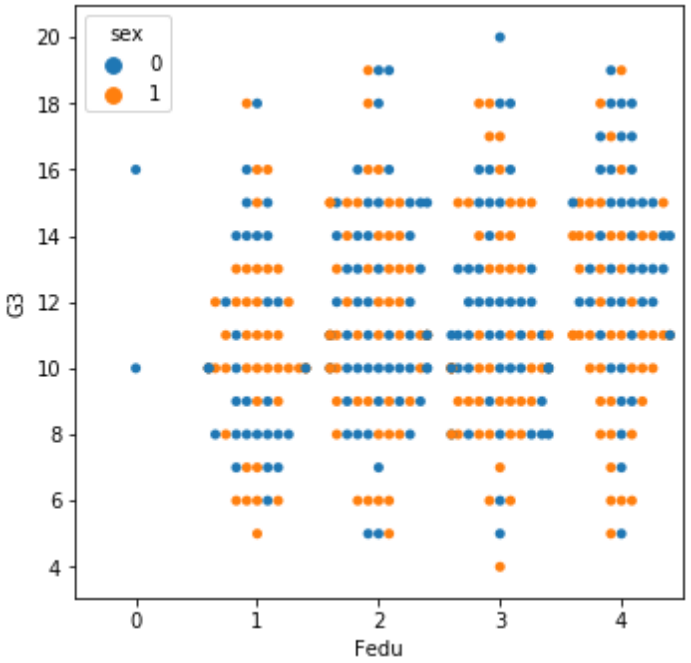


The mean final grade for female students is 11.21

The mean final grade for male students is 11.87

When these young women grow up, get jobs, and have children of their own; we find a correlation between the education of the mothers (Medu) and their children's performance in school. We will also plot the father's education level (Fedu):





pearson correlation coefficient for MOTHER'S EDUCATION LEVEL 0.19

If you want to improve the final grades of both males and females, ensure that girls are performing as well or better than the boys and ensure women are successful in the workforce.

Machine Learning

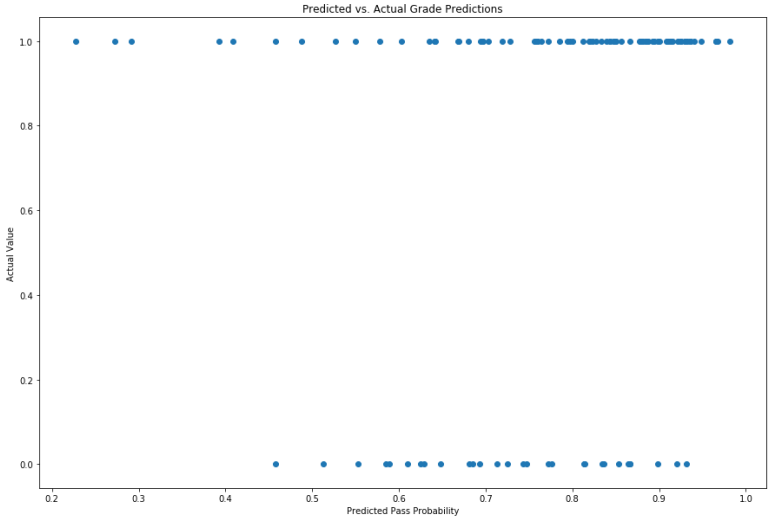
An initial search for possible models and hyperparameter values was performed using the HyperOpt package. This process provided a suggestion of a AdaBoostClassifier model. This model did not perform very well and I began to rethink the potential use of a grad prediction model. The final deliverable to the schools will be a system of three models.

The First Model - Intervention Prediction

This model will use logistic regression, which provides a probability of passing (> 0.5) based on socioeconomic and extracurricular factors.

Tuned Logistic Regression	
Metric	**Value**
Area Under the Curve	0.647

Top Five Feature Odds	
Feature	**Odds**
failures_0	2.12
absences_2	1.72
go_out_weekdays_2	1.67
failures_1	1.54
health_1	1.54



Comparison to Random Forest Classifier

Tuned Random Forest	
Metric	**Value**
Area Under the Curve	0.645

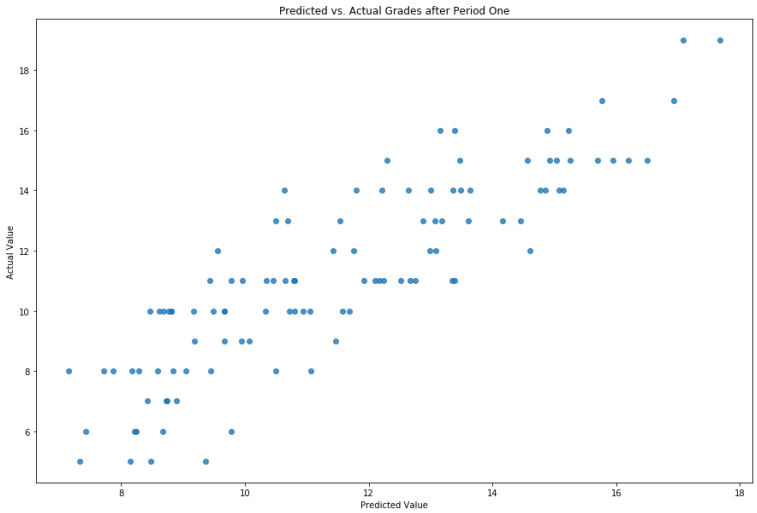
Top Five Feature Importances	
Feature	**Importance**
failures_0	0.028
failures_2	0.024
health_3	0.015
go_out_weekdays_4	0.014
reason_for_selection_course	0.013

The Second Model - Final Grade Prediction (after first period)

This model will use a Random Forest Regressor algorithm, which predicts the final grades based on the period one grades

Tuned Random Forest	
Metric	**Value**
Mean Squared Error	-2.50
Root Mean Square Error	1.58

Top Five Feature Importances	
Feature	**Importance**
period_one_grade_16	0.127
period_one_grade_17	0.094
period_one_grade_18	0.088
period_one_grade_15	0.076
period_one_grade_19	0.067

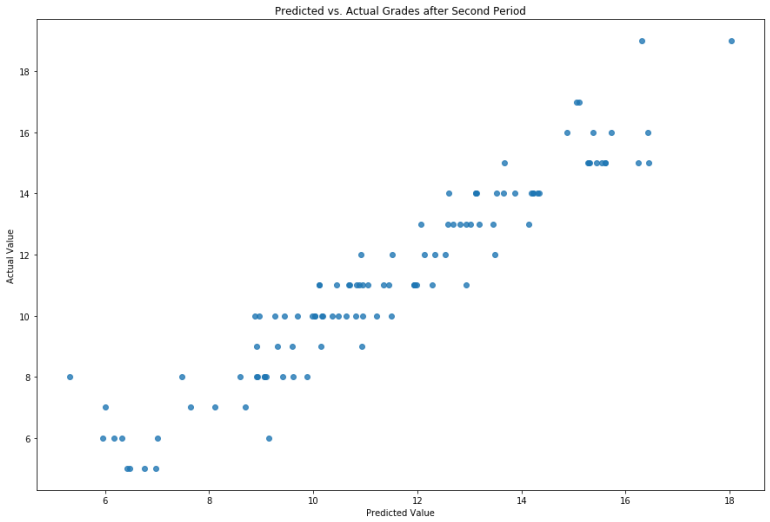


The Third Model - Final Grade Prediction (after second period)

This model will also use the Random Forest Regressor, which predicts the final grades based on the period one grades, period two grades, and whether there was improvement between the periods.

Tuned Random Forest	
Metric	**Value**
Mean Squared Error	-0.99
Root Mean Square Error	0.996

Top Five Feature Importances	
Feature	**Importance**
period_two_grade_18	0.175
period_two_grade_15	0.158
period_two_grade_16	0.087
period_two_grade_14	0.065
period_two_grade_13	0.060



With progressively more correlated data, the models perform significantly better towards the end of the year.

Machine Learning Conclusion

Using only extracurricular and socioeconomic data with the first logistic regression model we are not able to accurately predict the final grades of the students. However, we are able to gain significant insight into the variables that play a significant role in placing a student at risk of failure. We will explore those findings here.

The strongest variable that informs stakeholders of future student success is past success in other courses. I found a feature importance of 0.03 which represents the strongest decrease in model error over all other variables. When using a logistic regression model the odds of having a passing grade given no past failures is 2.12.

Other influential variables that increased model performance (decreased model error) are the number of absences (logistic odds are 1.71 with only two absences), the level of student health (feature importance 0.014 and logistic odds are 1.54), and the amount of time the student spends 'out' (logistic odds are 1.67).

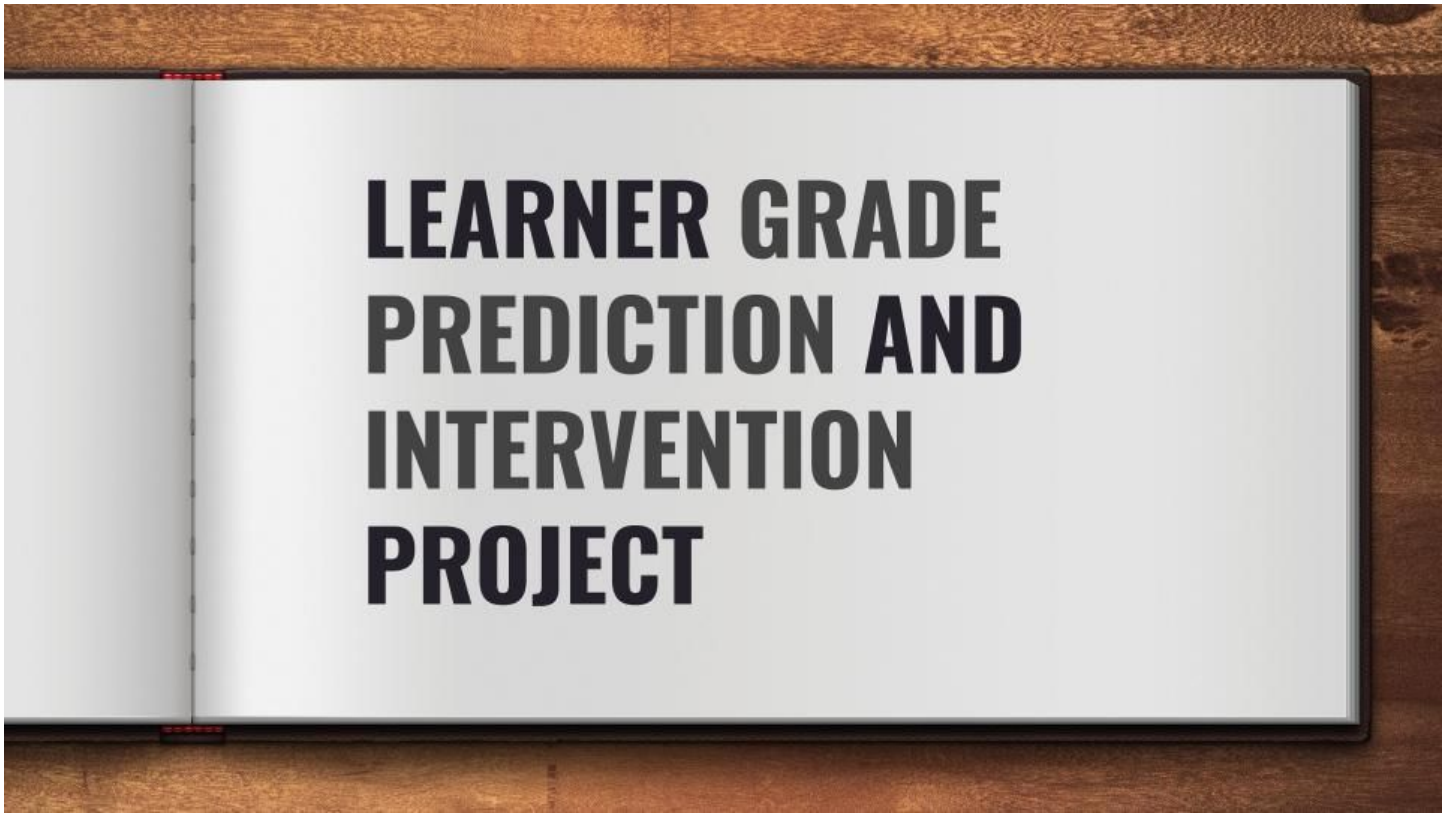
My recommendation would be to use the following metrics at the beginning of the school term to assess the potential risk of failure in the program; past success in other courses, number of historic absences from classes, and the student's general level of health during the course.

After the first and second semesters are finished the second and third models can be used respectively, to include the student's grades on those assessments. With each new standardized data point in the system the predictive power of the model increases in accuracy.

The first model is meant to inform stakeholders of 'at risk' learners and promote the generation of special programs and interventions. The second and third models are intended to provide additional insight into learner performance.

With the addition of significantly more data, I feel that the first logistic regression model would find new predictive power. Currently there is relatively little data in the top grade percentages. With less than 10 students achieving grades of 19 or 20 in the courses.

Presentation



Additional Reading and Resources

Solving Educational Data Interoperability Challenges Using the Ed-Fi Data Standard