Final Project Report
Capstone 2: When Should Students Take Paid Tutoring

Table of Contents

Introduction

Academic achievements are an important aspect for most people in their secondary school life as academics play a major role in determining which future career paths are available to them based on their grades in high school. As such, most students naturally would strive to aim for achieving high grades in school. However, every student's ability to learn is different, where some students may excel in subjects involving memorization like Science but struggle in subjects involving critical thinking like English. This may cause problems for students as universities or colleges may require grade requirements for certain subjects, and a student may be struggling in that subject to meet the grade threshold on their own. As such, in those situations, students may choose to seek out assistance to help improve their grades, and one of the common options is to seek out paid tutoring.

As tutoring can help students improve their grades on courses they are struggling with or on courses they need to achieve high grades to meet a requirement for, it can be a valuable option for students to seek. However, tutoring also costs money, with varying costs depending on the quality of service provided. As such, it is worth pondering if the results gained from getting tutoring outweigh the monetary costs. Some students may not need tutoring to obtain their desired grades, while for others, it could be just the help they need to achieve their desired academic goals.

Thus, this capstone project can help provide insight for students to determine whether tutoring would likely produce their desired results by analyzing the features and demographic data of existing students who have/haven't paid for tutoring and comparing the results.

Dataset Information

The data used for this capstone project is from the *Student Grade Prediction* dataset from Kaggle (https://www.kaggle.com/dipam7/student-grade-prediction) provided by Paulo Cortez. This dataset contains information regarding 395 students from two Portuguese secondary schools regarding the subjects of Mathematics and Portuguese (English class equivalent).

The dataset's 33 features include the following:

- school (student's school; GP or MS)
- sex (student's gender)
- age (student's age)
- address (urban or rural)
- famsize (student's family size; LE3 (less than or equal to 3) or GT3 (greater than 3)
- Pstatus (parent's cohabitation status; T (together) or A (apart)
- Medu (mother's education; scale of 1 to 5)
- Fedu (father's education; scale of 1 to 5)
- Mjob (mother's job)
- Fjob (father's job)
- reason (student's reason to choose this school)
- guardian (student's guardian; mother, father, or other)
- traveltime (travel time from home to school; scale of 1 to 5)
- studytime (weekly study time; scale of 1 to 5)
- failures (number of failed classes)
- schoolsup (if student is receiving academic assistance from school; yes or no)
- famsup (if student is receiving academic assistance from family; yes or no)
- paid (if student is receiving paid classes or tutoring; yes or no)
- activities (if student has extra-curricular activities; yes or no)
- nursery (if student attended nursery school; yes or no)
- higher (if student wants to take post-secondary education; yes or no)
- internet (if student has access to internet at home; yes or no)
- romantic (if student is in a romantic relationship; yes or no)
- famrel (quality of relationship with family; scale of 1 to 5)
- freetime (amount of free time outside of school; scale of 1 to 5)
- goout (amount of time spent with friends; scale of 1 to 5)
- Dalc (weekday alcohol consumption; scale of 1 to 5)
- Walc (weekend alcohol consumption; scale of 1 to 5)
- health (student's current health status; scale of 1 to 5)
- absences (number of absenses for student; scale of 1 to 5)
- G1 (first period grade; range of 0-20)
- G2 (second period grade; range of 0-20)
- G3 (third period grade; range of 0-20)

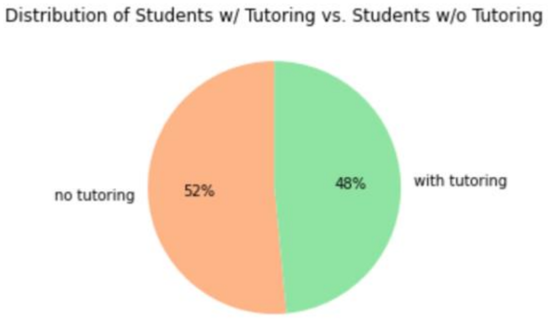
Our target variable is the paid feature as it represents if a student is taking paid tutoring.

Data Wrangling

From performing data wrangling on the dataset, there are no null values in any of the features. However, there are values in the G2 and G3 features that are 0, which seem to be a substitute for null values as there are no values of 0 in G1, the first period grade. So, it would be infeasible for a student's grade to drop to 0 in the second or third period when it wasn't 0 in the first period, and thus, all rows containing a value of 0 in the G2 or G3 columns are removed. The dataset now contains 357 rows after being cleaned compared to the original 395 rows.

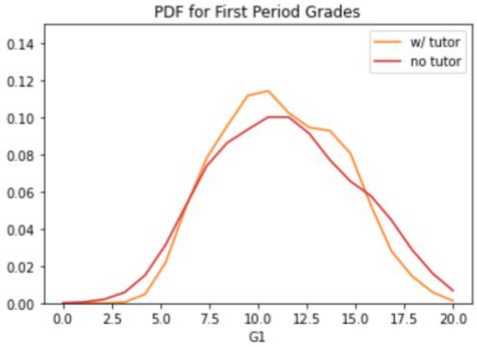
Exploratory Data Analysis

After wrangling and filtering the rows of the dataset, we can see from the pie chart below that the distribution of students with tutoring and without tutoring are approximately evenly distributed, with 48% of students (173 out of 357) in the dataset taking paid tutoring and 52% of students (184 out of 357) not taking any paid tutoring.



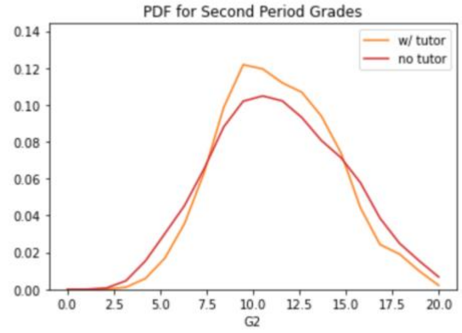
This shows that both sides have a nearly equal distribution of students, and so comparisons of the data can be made to find any differences in academic results between those taking paid tutoring vs those not taking paid tutoring.

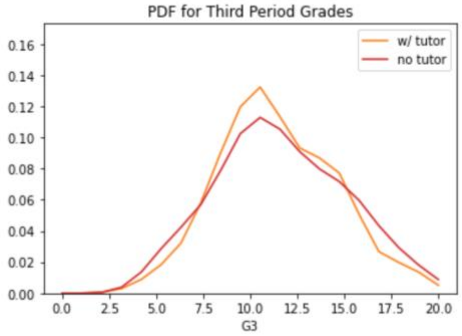
From creating PDFs comparing the likelihood of obtain grades between those taking paid tutoring vs those not taking paid tutoring in the first period, we obtain the following result:



From the PDF distribution, we can see that students with tutoring are more likely to achieve a grade within the 7.0 to 16.0 grade range (which ranges from average to above average) than students who have not taken paid tutoring. We can also see that students without tutoring are more likely to achieve a grade lower than 7.0 (below average) than students with tutoring. From those trends, we can infer that students with tutoring are more likely to perform better on average than students without tutoring. However, over the 16.0 grade range, it appears that students without tutoring are more likely to perform better than students with tutoring, which may contradict our inference. But that can be explained by the fact that students achieving those high grades may not need or seek out tutoring.

Looking at the PDF distribution for the second and third periods, we get the following:





We can see that these results have similar trends to the first period PDF, and so it can be inferred students with tutoring are overall more likely to perform better on average than students without tutoring, outside of those who are already performing spectacularly.

Pre-processing and Training

Overall, 12 features have been dropped prior to the train/test split in order to prevent overfitting because they are deemed to be irrelevant factors in determining whether a student is taking tutoring in this dataset due to either having an imbalanced distribution or having little to no feature importance.

The 12 dropped features include the following:

- sex
- guardian
- Pstatus
- school
- address
- nursery
- famsize
- Mjob
- Fjob
- higher
- internet
- traveltime

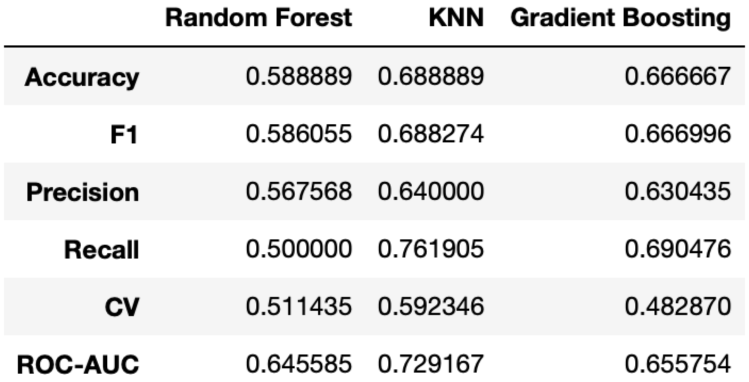
The remaining 21 features will be used in the train/test split to build a model for predicting the target variable, *paid*.

Modelling

Since this is a classification problem, the three models that will be used for this project are the Random Forest Classifier, the K-Nearest Neighbors Classifier, and the Gradient Boosting Classifier.

In order to make a selection on which of the three models is the best performing, they will be judged based on their accuracy, f1, precision, recall, cv, and roc-auc scores.

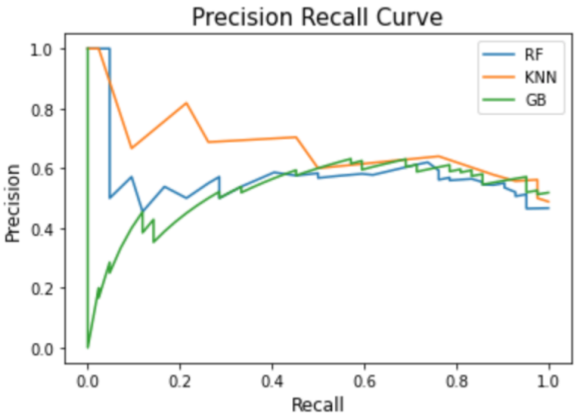
The results provided by each of the three classifier models are the following:



	Random Forest	KNN	Gradient Boosting
Accuracy	0.588889	0.688889	0.666667
F1	0.586055	0.688274	0.666996
Precision	0.567568	0.640000	0.630435
Recall	0.500000	0.761905	0.690476
CV	0.511435	0.592346	0.482870
ROC-AUC	0.645585	0.729167	0.655754

It is evident that the K-Nearest Neighbors classifier is by far the best performing model as it has the highest scores in every category.

From looking at the precision-recall curve:



it is clear that KNN is outperforming the other two models.

From these results, we can safely conclude that KNN is the best performing model among the three and select it as the model of choice to predict the target variable, *paid*.

Main Takeaways and Future Improvements

Overall, the K-Nearest Neighbors classifier was the best performing model, producing an accuracy score of ~0.6888, which is approximately 43% more accurate than randomly guessing if a student is taking paid tutoring or not. However, this also shows that the model still has room for improvement as an accuracy score of ~0.69 is not optimal.

An area of improvement could be expanding the sample size by including more subjects, more schools, and expanding the region such that we end up with a much larger sample size of students than the current size of 357.

Another area of improvement could include expanding the original dataset by collecting more relevant and detailed data to create more in-depth features to help fit the data better, such as collecting the exact time period that a student started taking paid tutoring like before first period grades were handed out or after a later period.

Furthermore, another improvement that could be made is implementing more yes or no questions in surveys to attempt to capture more of the individuality and personality of the students in order to help incorporate more relevant binary features to the model.