Predictive Analytics Exam

Sample Project – Student Success

From: Steve Jones, Sharpened Consulting
To: You
Re: New Consulting Opportunity

We have just been presented a unique opportunity to work with School Wiz, a group dedicated to providing remedial education to troubled students. School Wiz has heard about our work and wants to explore using our services to advance their business goals. If we secure their business, we will be working with them for the next several months on data collection and analysis. However, they are not yet convinced that predictive analytics can help them. To earn their business, we need to demonstrate how we can use our tools to answer their major questions, which are:

1. How accurately can we predict which students will pass based on a variety of factors; and
2. Which factors are most important for predicting pass rates?

They have collected preliminary data and sent it to us. We can use this dataset to illustrate how we can answer these questions. I've enclosed their materials, which includes a brief description of the data, a data dictionary, and some code to get you started.

I am meeting with their representative in five hours. I know it is last minute, but I would like you build three models, two tree-based models and one generalized linear model to answer these questions and write a report for that meeting that includes the following:

1. **Executive Summary**: Report which models you used to answer the questions above and summarize the key results of the analysis.
2. **Data Exploration and Feature Selection**: Present and discuss key characteristics of the data. Describe and explain any data cleaning, data transformations, and feature creation.
3. **Model Selection and Validation**: Describe the model fitting and validation process used. State which models you selected and why they are preferable to other choices.
    a. Build two tree-based models (one decision tree and one random forest) and one GLM. The tree-based models can be used to inform your GLM work—make sure to comment about which tree-based model you use to inform your GLM work and why.
    b. State which model(s) you recommend and why.
4. **Findings and Recommendations**: Interpret the results of the selected model(s) and discuss additional steps that could be to be taken to improve the analysis had there been more time.

The first and fourth parts should be written for School Wiz and thus be nontechnical. The middle two sections should include technical details that our team can use should we win this contract.

Along with this report, provide the Rmd file that you used to build your models so that if they wish to adjust some things on their own they can do so. They had someone try to learn how to analyze the data on their own and gave up, so you can start with their Rmd file, which I've provided. It would be especially impressive if we can be specific with our report. Statements such as "Our model predicts the pass rate correctly in x% of cases," or "The variable y was significant in models we fit" would be helpful. Also remember to use clear language and justify your decisions and assumptions. Finally, if you have any reservations about using certain variables, make sure you state your reservations.

Thanks,

Steve

From: John Essen, School Wiz
To: Steve Jones
Re: Dataset

As per our recent correspondence, I've attached the preliminary dataset we've collected to jumpstart our initiative. Some key features of this data are:

- There are 29 predictor variables that were collected prior to students entering the formal school year.
- There are four variables that were only observed after the school year started:
    - absences – the number of absences;
    - G1 – the grade for the first trimester of the school year;
    - G2 – the grade for the second trimester; and
    - G3 – the grade for the final trimester.
- A student who receives a grade of 10 or more passes.

We had someone start working on some models to fit the data, but he didn't get very far before realizing it was too difficult. We did determine that G1, G2, and G3 were highly correlated, so we'd like you to just focus on building a model for G3. The Rmd file he started is attached. Some questions we weren't sure about that were never resolved:

- The grades are on a scale from 0 to 20, but we may want to model just the pass rates. If you can figure out a good way to build a model for whether someone will pass (G3 >= 10) instead of the overall grade, that would be great.
- We only want to use the 29 predictor variables known in advance in the model to predict G3 because we will know them before the school year starts, but the dataset we have includes G1, G2, and absences as well. (Note: they have been removed in the Rmd file provided.)
- We wonder if we could create some new variables based on those we have to improve accuracy. For example, we think the parent variables may be important; it may be worth it to play around with several combinations.
- Some of the data seems wrong. If you need to remove some observations that would be okay but explain why you are removing them.

I've included the data dictionary so you know exactly what each variable is. Good luck! I really hope you guys can blow us away at our meeting later.

Thanks,

John

Data Dictionary	
Attribute	**Description (Domain)**
school	student's school (binary: GP (Grand Pines) or MHS (Marble Hill School))
sex	student's sex (binary: female or male)
age	student's age (numeric: from 15 to 22)
address	student's home address type (binary: U (Urban) or R (Rural))
famsize	family size (binary: GT3 (>3) or LE3 (3≤)
Pstatus	parent's status (binary: A (Apart) or T (Together))
Medu	mother's education (numeric: from 0 to 4 [a])
Fedu	father's education (numeric: from 0 to 4 [a])
Mjob	mother's job (nominal [b])
Fjob	father's job (nominal [b])
reason	reason to choose school (nominal: home (close to home), reputation (school reputation), course (course preference), or other)
guardian	student's guardian (nominal: mother, father, or other)
traveltime	home to school travel time (numeric: 1 – <15 minutes, 2 – 15 to 30 minutes, 2 – 30 minutes to 1 hour or 4 – > 1 hour)
studytime	weekly study time (numeric: 1 – <2 hours, 2 – 2 to 5 hours, 3 – 5 to 10 hours or 4 – >10 hours
failures	number of past class failures (numeric: *n* if $0 \leq n < 3$, else 3)
schoolsup	extra educational support (binary: yes or no)
famsup	extra family supplement (binary: yes or no)
paid	extra paid classes (binary: yes or no)
activities	extra-curricular activities (binary: yes or no)
nursery	attended nursery school (binary: yes or no)
higher	wants to take higher education (binary: yes or no)
internet	internet access at home (binary: yes or no)
romantic	has a romantic relationship (binary: yes or no)
famrel	quality of family relationships (numeric: from 1- very bad to 5 – excellent)
freetime	free time after school (numeric: 1 – very low to 5 – very high)
goout	going out with friends (numeric: 1 – very low to 5 – very high)
Dalc	weekday alcohol consumption (numeric: from 1 – very low to 5 – very high)
Walc	weekend alcohol consumption (numeric: from 1 – very low to 5 – very high)
health	current health status (numeric: from 1 – very bad to 5 – very good)
absences	number of school absences (numeric: from 0 to 75)
G1	first trimester grade (numeric: from 0 to 20)
G2	second trimester grade (numeric: from 0 to 20)
G3 (target)	third trimester grade (numeric: from 0 to 20)

[a] 0 – none, 1 – primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education (high school) or 4 – higher education (college)

[b] teacher, health (health care related), services (civil services, administrative or police), at_home, or other