Gwen Rino
Dhiraj Khanna, mentor
October 2017

Predicting Student Outcomes

Can we predict

who is at high risk of failing?

UCI

Machine Learning Repository

Center for Machine Learning and Intelligent Systems

Student Performance Data Set

Download: Data Folder, Data Set Description

Abstract: Predict student performance in secondary education (high school).

Data Set Characteristics:	Multivariate	**Number of Instances:**	649	**Area:**	Social
Attribute Characteristics:	Integer	**Number of Attributes:**	33	**Date Donated**	2014-11-27
Associated Tasks:	Classification, Regression	**Missing Values?**	N/A	**Number of Web Hits:**	217797

Source:

Paulo Cortez, University of Minho, GuimarÃ£es, Portugal, http://www3.dsi.uminho.pt/pcortez

Data wrangling

- Data was very clean, no missing values, in tidy format
- Two different files, one Portuguese grades, one math grades
- Combined the datasets into one with a new variable, “course”
- Also created a second version of the dataset in which G3 was recast from a numeric to a factor variable “outcome” with two levels, “pass” and “fail”

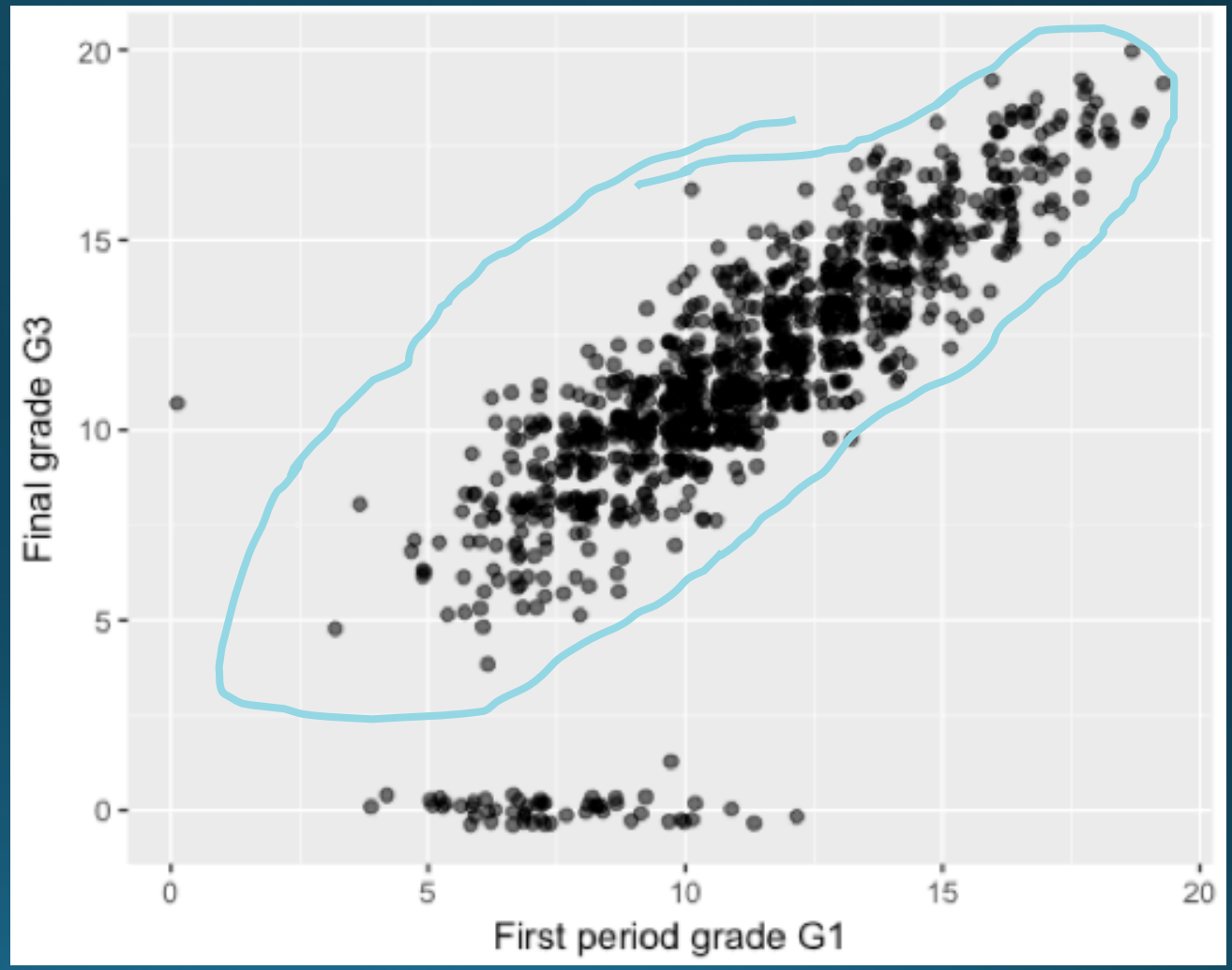
Can we predict who is at risk of failing

using

midterm grades

as predictors?

- A linear model predicts most but not all students' final grades.
- ***Can a non-linear model do any better?***



Explorations in classification

CHALLENGES & SOLUTIONS

- Which variables should I consider? Forward and backward stepwise selection (package "leaps")
- How can I balance the dataset? Over- and undersampling, make synthetic minority class data (package "ROSE")
- How can I compare the accuracy of different models? AICs, AURC, confusion matrixes

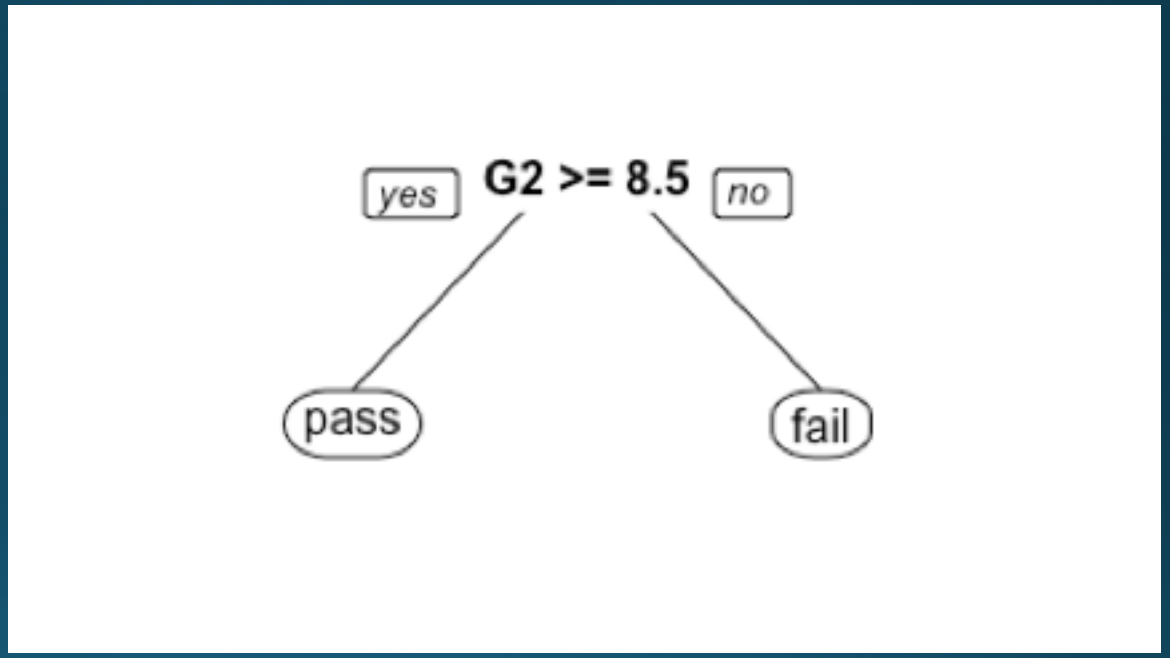
TYPES OF MODELS

- Logistic regression
- Decision trees (package "rpart" and "rpart.plot")
- Random forest (package "randomForest")

Two effective classification models

MODEL	INPUT	"FAIL" CORRECTLY PREDICTED	"PASS" CORRECTLY PREDICTED	CONSIDERATIONS
Logistic regression Data balanced by undersampling	G1 and G2	96.0%	93.2%	High accuracy. Output is opaque for most users; requires professional interpretation.
Decision tree Data balanced by both over- and undersampling	G1 and G2	96.0%	93.2%	High accuracy. Output is intuitive and easy to understand.

The best predictive models use only G1 and G2 as predictors. All the other variables have no impact on the effectiveness of the models.



Can we predict who is at risk of failing

without using midterm grades as predictors?

NO

Can we predict who is at risk of failing

without using second midterm grades as predictors?

Predicting student outcomes without second midterm grades

MODEL	INPUT	"FAIL" CORRECTLY PREDICTED	"PASS" CORRECTLY PREDICTED	CONSIDERATIONS
Logistic regression Data balanced by undersampling	G1 # of previous failures course (math or Portuguese)	92.0%	84.7%	In exchange for an earlier prediction, we lose accuracy overall and we overpredict failure. Some students will receive interventions even though they would pass without them.

Summary

With knowledge of students' midterm grades, failure is predictable

It is possible to make meaningful predictions early in the term

Demographic and social information in this dataset is not predictive of final grades

Formal data analysis isn't required for educators to know which students are most in need of interventions

More "pass" students will be misclassified and so will receive unnecessary interventions

Consider other factors?

Family income, food security, incarcerated parent, drug use, pregnancy/parenting, home language…

Thank you!