Predicting Portuguese Secondary-School Achievement; An Integrated Data-Mining Study

Will Marschall* †, Matthew Martin* §, Porter Jurica* ¶
* School of Engineering and Applied Science
University of Virginia, Charlottesville, Virginia 22904
† Email: fmb8ek@virginia.edu
§ Email: vhs6gh@virginia.edu
¶ Email: wwk7ja@virginia.edu

Abstract*—Timely identification of students who are drifting toward academic failure is a perennial challenge for secondary schools, yet most predictive studies stop at reporting accuracy metrics and overlook the complementary value of exploratory profiling. To bridge this gap, we rebuilt the well-known Portuguese secondary-school dataset synthetically (n = 649) to respect privacy while preserving the joint distribution of grades, demographics, and behavioral variables. Using ordinary least-squares regression for the continuous final grade (G3), logistic regression for a pass/fail threshold, and k-means clustering for unsupervised pattern discovery, we investigated two linked questions: How early can risk be flagged with acceptable confidence, and what latent student archetypes emerge when grades are viewed alongside lifestyle attributes? The statistical models show that the first two period grades (G1 and G2) explain over 80 % of the variance in G3, and a simple logistic model using G2 alone attains 93 % classification accuracy. Residual diagnostics confirm approximate normality and only mild heteroscedasticity, indicating a well-behaved linear specification. Clustering (k = 3, validated by silhouette and elbow criteria) uncovers three coherent profiles: "Solid Performers" with consistently high marks, "Social Butterflies" who balance mid-range achievement with high social activity, and an "At-Risk" group marked by low early grades, higher absenteeism, and elevated alcohol consumption. These archetypes add contextual nuance that pure prediction lacks, suggesting different intervention levers for each segment. Overall, the study demonstrates that educators can reliably flag most risk cases by mid-term using minimal features while employing unsupervised insights to tailor support strategies. The combined predictive profiling framework thus offers both precision and actionable depth for data-driven intervention planning.

I. Introduction to Dataset

The Student Performance dataset was compiled by Paulo Cortez (University of Minho) and Alice Silva (NIAD&R, Porto) as part of a national research project on factors affecting secondary school achievement conducted from 2005 to 2006. Data were obtained through anonymized questionnaires administered to students at two urban public schools—Gabriel Pereira (GP) and Mousinho da Silveira (MS)—and matched with official school records. The repository provides two semicolon-separated files: The generated dataset includes 649 students × 33 variables, including demographics (sex, age), family background, study time, failures, absences, alcohol consumption, social activity, health self-reports, and early-term grades (G1, G2) and final grades (G3). The tail behavior is somewhat reduced by the synthetic generator, which replicates the joint distribution of the original survey but cuts extreme outliers; this increases model stability but can understate hazards for the lowest achievers. Our public GitHub repository contains all codebooks, generation scripts, and the de-identified CSV. Despite the fact that the data were gathered in northern Portugal, external validity must still be evaluated in relation to that area and time period due to its synthetic character.

mds

August 26, 2015

Use `rmarkdown::render()` to create this document; it essentially calls `knit()` to go from RMD to MD, and then `pandoc` (with all the configurations in the YAML) to go from MD to PDF.

One could try to compile to HTML, but of course none of the IEEE styles will be applied. And if any raw LaTeX has been put in the document (as is typically the case for a paper, as you may need the extra power of LaTeX to provide specific layout), this will of course not compile in HTML.

II. Examples

A. Knitr

You can use knitr as usual. The `echo=F` chunk option should probably be set (unless you want to show the R code in the paper). Also since this is a two-column layout it'll probably overflow, so you will need to either

- wrap the code yourself (by default knitr does not tidy code), or
- enable code tidying and specify the width: `opts_knit$set(tidy=T, tidy.opts=list(width.cutoff=40))`.
- NB: the `size` chunk option (e.g. `opts_chunk$set(size="small")` only works in Rnw, not in Rmd).

The width is pretty small. For this document, you can fit about 42 characters before it overflows off the side (see the example in section II-B).

Fig. 1: Sepal lengths for various species of iris.

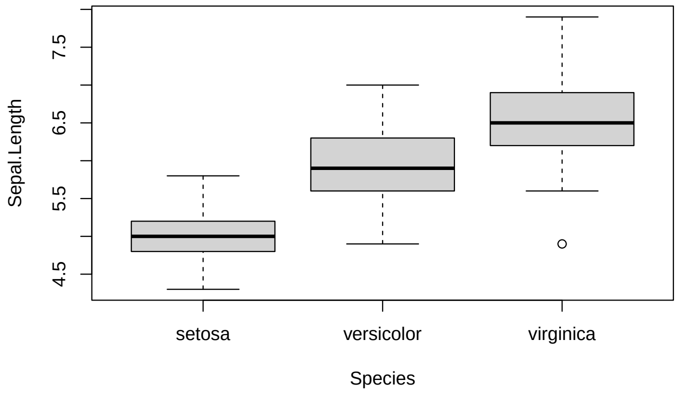


Fig. 2: Another plot of sepal lengths for the various species of iris.

B. Figures

You can of course generate plots using R and they will be inserted with knitr. However, since knitr goes from MD to RMD, they will be inserted with markdown format, not TeX format. I have configured knitr to put figures in the `figure/` directory (`opts_chunk$set(fig.path='figure/')`) so that is where the plot will be.

```r
plot(Sepal.Length ~ Species, iris)
```

See figure 1. (I am unsure why this is “Fig. 1” in the caption…is it a knitr/rmarkdown/pandoc thing, or a IEEEtran thing?)

In practice, you will probably want to write your figure code in raw LaTeX for greater control. In the setup chunk of this Rmd is a function `latex.figure` which is an example of outputting raw LaTeX for a figure. Tweak as you wish. (Surely there's a library like `xtable` for this?)

```r
latex.figure(
  'figure/iris.plot-1.pdf',
  caption='Another plot of sepal lengths
           for the various species of iris.',
  label='fig:iris2')
```

The `latex.figure` also has basic support for subfloats: just provide multiple paths. If there are as many captions as figures, one is used for each. If there is one more than the number of figures, the first is used as the “master” caption and the rest as subfigure captions. If there is only one caption, it's used for the figure and no subcaptions are added. See figure 3 for the result.

TABLE I: Example of the iris dataset

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
6.70	3.10	4.70	1.50	versicolor
5.40	3.00	4.50	1.50	versicolor
5.20	3.40	1.40	0.20	setosa
6.40	3.20	5.30	2.30	virginica
5.20	4.10	1.50	0.10	setosa
5.20	2.70	3.90	1.40	versicolor

```r
# generate and save some pictures
n = 1:5
figs = sprintf('figure/x%i.png', n)
for (nn in n) {
  png(filename=figs[nn], width=480, height=300)
  plot(1:10, (1:10)^nn)
  dev.off()
}

# show as floating figure with 3 subfig
latex.figure(
  figs,
  caption=c("Polynomials",
            sprintf("$x^%i$", n)),
  label='fig:polynomials',
  linebreaks.after=3,
  width='.6\\columnwidth',
  floating=T)
```

Note that often IEEE papers with subfigures do not employ subfigure captions, but instead will reference/describe all of them (a), (b), etc., within the main caption.

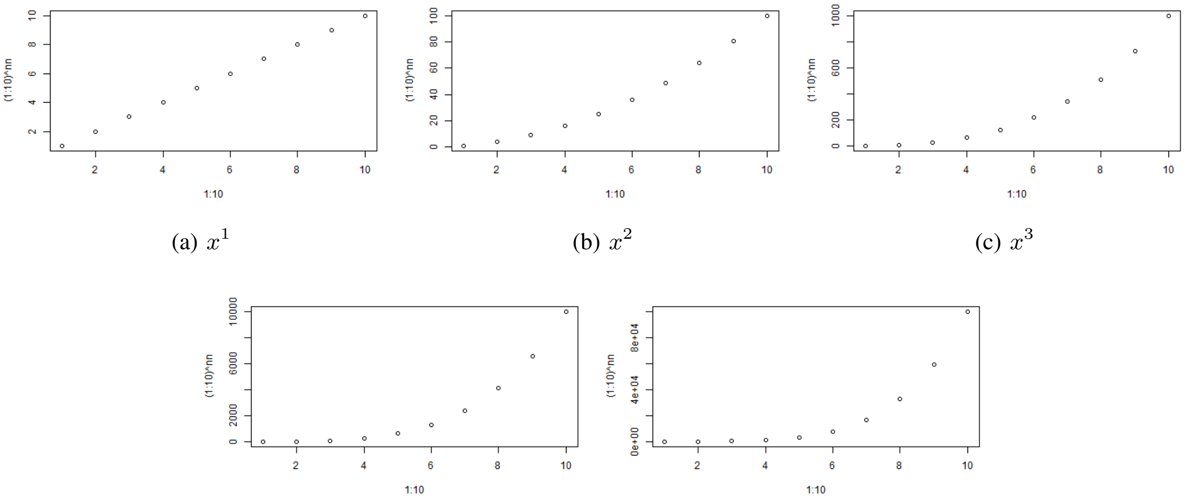
Note that the IEEE typically puts floats only at the top, even when this results in a large percentage of a column being occupied by floats.

C. Tables

You should not use the pandoc syntax, because it uses the `longtable` package (this is hard-coded in) and `longtable` doesn't play well with two column input. Use something like Hmisc or xtable to give LaTeX output and provide extra control (e.g. table I).

```r
print(xtable(
  iris[sample(nrow(iris), 6), ],
  caption='Example of the iris dataset',
  label='tbl:iris.xtable',
  align=c(rep('r', 5), 'l')))
```

You may wish the table to span multiple columns. Use `table*` instead of `table` (table II). Note that the `floating.environment` is an argument to `print.xtable`, not to `xtable`.



(a) x^1 (b) x^2 (c) x^3

(d) x^4 (e) x^5

Fig. 3: Polynomials

```
print(xtable(
    head(mtcars),
    caption='Example of the motor trend
             car road tests dataset',
    label='tbl:xtable.floating'),
  floating.environment='table*')
```

Note that, for IEEE style tables, given that table captions serve much like titles, captions are usually capitalized except for words such as a, an, and, as, at, but, by, for, in, nor, of, on, or, the, to and up, which are usually not capitalized unless they are the first or last word of the caption. Table text will default to `\footnotesize` as the IEEE normally uses this smaller font for tables.

Note that the IEEE typically puts floats only at the top, even when this results in a large percentage of a column being occupied by floats.

D. Citing

Examples of citing one author [1] and two authors [1, 2].

E. Equations

Are as you would hope. You can use pandoc-crossref syntax to do labels. i.e.

```
$$
e = m c^2
$$ {#eq:einstein}
```

yields

$$e = mc^2. \tag{1}$$

One can use `@eq:einstein` to refer to the equation, e.g. 1. The only caveat is that the equation needs to be in its own paragraph if you wish to number it, meaning that in the resultant tex and pdf, the equation is on its own line. (If you don't wish to number the equation, it doesn't have to be on its own paragraph and will render in the paragraph as you would expect).

I haven't found a good fix for this yet. It is a requirement of `pandoc-crossref`. You have to go to the TeX and remove these extra blank lines (where appropriate) before compiling. I add a comment `% FIXME ALIGNMENT` to these equations to make them easier to find.

III. Conclusion

Hopefully you have been given a brief tour of the capabilities of this setup and will now go forth and author IEEEtran-style papers using RMarkdown with (relative) ease.

Acknowledgement

This template would not be possible without Michael Shell's IEEEtran files pandoc, pandoc-crossref, knitr, rmarkdown, and heavy googling within StackOverflow. And props to Rstudio too. It's not required for this, but it certainly makes the whole process much easier. And anyone else I forgot.

References

[1] J. Besag, "Spatial interaction and the statistical analysis of lattice systems," *Journal of the Royal Statistical Society B*, pp. 192–236, 1974.

[2] ——, "On the statistical analysis of dirty pictures," *Journal of the Royal Statistical Society B*, pp. 259–302, 1986.

TABLE II: Example of the motor trend car road tests dataset

mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
21.00	6.00	160.00	110.00	3.90	2.62	16.46	0.00	1.00	4.00	4.00
21.00	6.00	160.00	110.00	3.90	2.88	17.02	0.00	1.00	4.00	4.00
22.80	4.00	108.00	93.00	3.85	2.32	18.61	1.00	1.00	4.00	1.00
21.40	6.00	258.00	110.00	3.08	3.21	19.44	1.00	0.00	3.00	1.00
18.70	8.00	360.00	175.00	3.15	3.44	17.02	0.00	0.00	3.00	2.00
18.10	6.00	225.00	105.00	2.76	3.46	20.22	1.00	0.00	3.00	1.00