Badass working title

Ece Caliskan, Hyunah Blue?, Nicolai Sprenger, Felix Wolff

Document objective, taken from assignment

You will document, explain and justify your methodology, experiments, and results in a term paper. The main part of term paper including relevant graphs and tables excluding the reference list should not exceed 16 pages. You are required to complete the task in a group of 3-4 students for which you must register on moodle. For the assignment, you are highly encouraged to go beyond the standard methods taught in class as well as make use of the scientific literature and conduct and document your own experiments with the data. Make sure to consider all stages of a typical modeling process: - research the relevant technical and task-related knowledge in the literature - gather, clean and preprocess the relevant data - select the best model and model parameters - deploy and assess the model in terms of performance and plausibility with possibly revision of any step or the whole process.

Why is there another document in this directory?

(Felix) I want to propose another document style, which would be written in LaTeX. I have written my bachelor thesis with the classicthesis theme and prefer its clean look over the barebones Rmd look. It's just looks but that counts with the people reading it, too. Unfortunately the Rmd PDF generation is hardly customizeable. Let's discuss this on Jan 10.

Paper outline

1. Abstract: What our solution achieves, anad what it is made of
2. (debateable) survey of state-of-the-art approaches for binary classification
3. Data cleansing (20%) and feature engineering (80%)

a. where are the NULL values
b. which values are suspicious? e.g. birthdates etc
c. distributions that might be bad for some models
d. a few words on used imputation strategies
e. newly engineered features and how leakage was adressed with subsampling

2. Evaluating and building models

a. Agree on metric to measure model performance with - PCC, AUC, Brier...
b. Evaluated models and their performance
c. How was subsampling done uniformly across the different models
d. How was hyperparameter tuning done?
e. A short note on collaboration

3. Model stacking

a. Individual model weaknesses, where does which model perform badly?
b. Stacking logistic regression
c. Resulting architecture
d. Performance comparison

4. Closing remarks

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

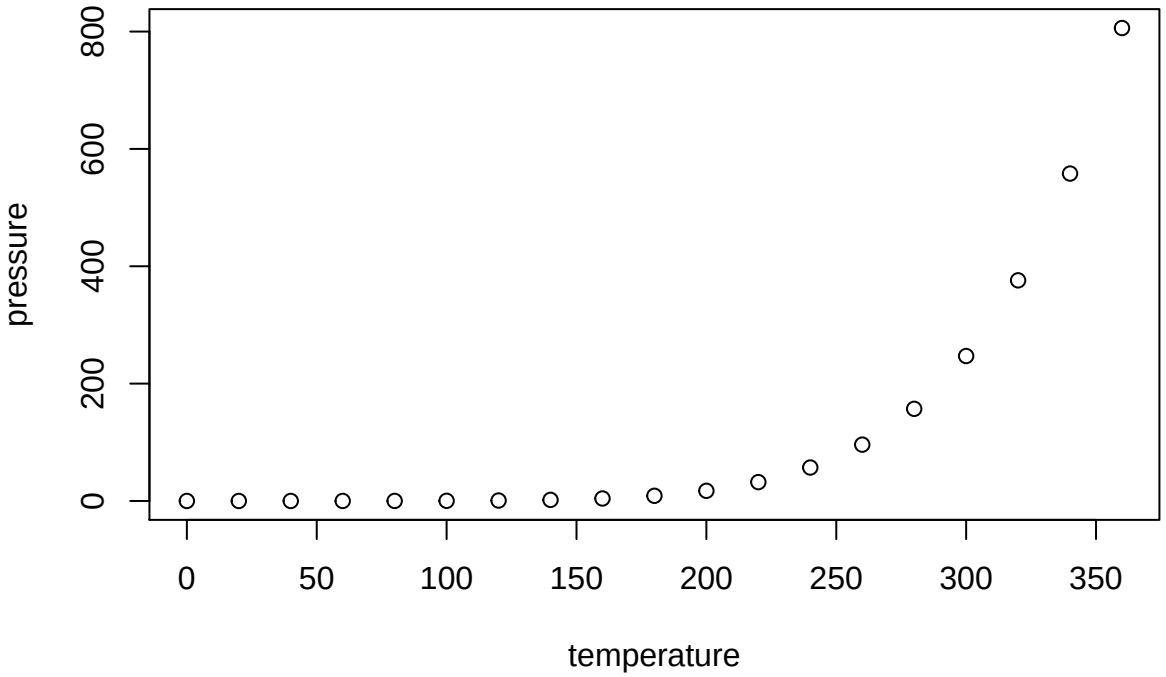
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed           dist
##  Min.   : 4.0   Min.   :  2.00
##  1st Qu.:12.0   1st Qu.: 26.00
##  Median :15.0   Median : 36.00
##  Mean   :15.4   Mean   : 42.98
##  3rd Qu.:19.0   3rd Qu.: 56.00
##  Max.   :25.0   Max.   :120.00
```

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.