Case Study

Data Scientist at PartnerRe

PartnerRe

Full position - North America

This case study simulates typical tasks that PartnerRe's Data Scientists would solve in their day to day work. That is, it is designed to give you an impression of the activities on the job. At the same time, we will get a feel for how you perform in these situations.

We hope you enjoy the work on the case problems and find the process instructive!

Please solve the case study following these guidelines:

- Use a recent version of R (e.g., 3.6.x) and R packages to produce the results.
- Use R Markdown for your analysis and write brief comments documenting your work and outcomes. You will be asked to submit the PDF from the markdown.
- Where appropriate, focus on using the `tidyverse` packages (e.g. `dplyr` and `tidyr` for data manipulation and `ggplot2` for plotting).
- Write generalized code. That is, your code should not depend on knowing the content of the data such as scheme names in advance.
- `for` loops should not be required. Please avoid them entirely.
- Where you modify the data (e.g. because there are `NA`s), please be prepared to describe the steps you took and your thinking behind the choices you made.

For the presentation, please

- expect us to walk through your code together,
- be prepared to answer questions about your solution,
- be prepared to analyse the data interactively during the session,
- You will be presenting from a virtual session (sharing screen from your own computer with your statistics environment on it),

Please solve the following problems.

Problem 1 – Data handling, analysis and plotting

The first problem of the case study builds on the data in the files `p01-02_portfolio.csv` and `p01-02_rates.csv`. One file contains membership information for a Group Life portfolio and one has information on the rates which should be charged.

Question a.

Read the data from the two files into R's memory. The rates are applicable to each individual in the portfolio, depending on that individual's age and gender. Combine the two datasets into a single table by looking up the rate for each line of the portfolio.

Question b.

Group the Industry field into common-sense based groupings and determine the mean, standard deviation and quintiles of DeathSI for each of your industry groups.

Question c.

The following code performs a Monte Carlo simulation on the data you have loaded and combined in Question a.:

```
set.seed(1234)
nsim <- 1000
res <- lapply(1:nsim, function(i,...) {
  x <- ifelse(
    runif(dim(combined_data)[1]) < combined_data$Rate / 1000,
    combined_data$DeathSI,
    0
  );
  list(cost = sum(x), count = length(x[x > 0]))
})
```

Apply this simulation to each scheme in the dataset you were provided, running 1'000 simulations per scheme. Produce a plot of the simulated outcomes ("cost"). Your plot should show

- a separate histogram per scheme;
- all 5 histograms below each other so that they can be easily compared;
- vertical lines in each graph indicating the median, mean and 99.5th percentile of each distribution.

Problem 2 – Statistical learning

The following code loads the data used in Problem 2. The data is stored in two `.csv` files containing a data set and a scoring set, respectively. These data are NOT related to Problem 1.

```
library(tidyverse)
library(here)

dta     <- read_csv(
  here("data/p02re_data.csv"),
  col_types = cols()
)
scoring <- read_csv(
  here("data/p02re_scoring.csv"),
  col_types = cols()
)
```

The goal of this problem is to predict the value of the variable `outcome` in the dataset defined above. The measure of goodness of fit is the area under the receiver operating characteristic curve (AUC) and will be measured on the scoring set.

Question a.

Describe the statistical learning problem. Explain the difficulties in training a model with the above dataset.

Question b.

The variable `group` has different factor levels in the data and scoring set. How will this difference impact your predictive model? What solutions might you offer?

Question c.

Train a model to predict `outcome`. You are free to use any package in R. Explain the logical progression that leads you to your final solution. Be prepared to interrogate relationships within the model (i.e., how each variable impacts the model and their relationship to the `outcome`)

Question d.

Give an estimate of the AUC you expect on the scoring set. Discuss the choice of AUC as the performance measure (e.g. instead of accuracy).