Fall 2019 DATA 621 Final Project

Corey Arnouts[1], Adam Douglas[1], Jason Givens-Doyle[1], & Michael Silva[1]

[1] CUNY School of Professional Studies

Author Note

MS in Data Science Students

Correspondence concerning this article should be addressed to Corey Arnouts, 119 W 31st St., New York, NY 10001. E-mail: Corey.Arnouts@spsmail.cuny.edu

Abstract

One or two sentences providing a **basic introduction** to the field, comprehensible to a scientist in any discipline.

Two to three sentences of **more detailed background**, comprehensible to scientists in related disciplines.

One sentence clearly stating the **general problem** being addressed by this particular study.

One sentence summarizing the main result (with the words "**here we show**" or their equivalent).

Two or three sentences explaining what the **main result** reveals in direct comparison to what was thought to be the case previously, or how the main result adds to previous knowledge.

One or two sentences to put the results into a more **general context**.

Two or three sentences to provide a **broader perspective**, readily comprehensible to a scientist in any discipline.

Keywords: CoIL Challenge, Logistic Regression

Word count: X

Fall 2019 DATA 621 Final Project

Introduction

Busineses use data science to extract insight from data. It has many practical business applications. The CoIL Challenge showcases the power data science can bring to bear on fundamental business problems.

The CoIL Challenge was a datamining competition organized by the the Computational Intelligence and Learning Cluster. It was held in the period of March-May 2000. The COIL Challenge had two tasks:

- Predict which customers are potentially interested in a caravan insurance policy; and
- Describe the actual or potential customers; and possibly explain why these customers buy a caravan policy.

In total 43 solutions were submitted. The winners of the challenge were Charles Elkan for the prediction task and Nick Street and YongSeog Kim for the description task.

In this paper we set out to complete the COIL Challenge ourselves using a logistic regression classifier. **SUMARISE FINDINGS**

Literature Review

Discuss how other researchers have addressed similar problems, what their achievements are, and what the advantage and drawbacks of each reviewed approach are. Explain how your investigation is similar or different to the state-of-theart. Please cite the relevant papers where appropriate.

Methodology

Discuss the key aspects of your problem, data set and regression model(s). Given that you are working on real-world data, explain at a high-level your exploratory data analysis, how you prepared the data for regression modeling, your process for building regression models, and your model selection.

We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study.

Experimentation and Results

Describe the specifics of what you did (data exploration, data preparation, model building, model selection, model evaluation, etc.), and what you found out (statistical analyses, interpretation and discussion of the results, etc.).

We used R (Version 3.6.1; R Core Team, 2019) and the R-package *papaja* (Version 0.1.0.9842; Aust & Barth, 2018) for all our analyses.

Discussion and Conclusions

Conclude your findings, limitations, and suggest areas for future work.

References

Aust, F., & Barth, M. (2018). *papaja: Create APA manuscripts with R Markdown*. Retrieved from https://github.com/crsh/papaja

R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Appendices

- Supplemental tables and/or figures.
- R statistical programming code.