CWRU DSCI-451: 451SemProj-Overview

Roger H. French, JiQi Liu

September 19, 2018

Contents

2 DSCI451: Data Science Semester Research Project

A Data Analysis/Prediction & Modeling Project

2.1 Purpose of Semester Project Assignment

In the Semester Project, for students enrolled in DSCI451,

- you will take a four-part approach
  - to doing a data analysis through EDA and Insights
  - for a topic from your area of research interest.
- If you are able to do some modeling and statistical learning, all the better.

2.2 Basic steps we use to construct a data analysis.

Modified from Jeff Leek's slides

- (available in your repo in 17f-dsci351-451-prof/3-readings/)

2.2.1 SemProj. Part a) Define Question

- Background on the research area and critical issues
- Define the question

- Define the ideal data set
- Determine what data you can access
- Define critical capabilities and identify packages you will draw upon
- Obtain the data, define you target data structure
- Clean and tidy the data

2.2.2 SemProj Part b) Cleaning and EDA

- Write you databook, defining variables, units and data structures
- Data visualization and exploratory data analysis
- Observations of trends and functional forms
- Power transformations
- Validate with reference to domain knowledge
- Evaluate the types of Modeling Approaches to take

2.2.3 SemProj Part c) Modeling and Statistical Learning

- Types of modeling to try
- Statistical prediction/modeling
- Model selection
- Cross-validation, Predictive R2
- Interpret results
- Challenge results

2.2.4 SemProj Part d) Present your final models and learnings

- Present your results
- Present reproducible code
- Comparison to other modeling approaches in the literature

You will use .R scripts and

- do reports and presentations in .Rmd files,
- so that they are interactive, reproducible, open-science presentations.

Organize and store you code in your repo.

You will turn in

- all code (.R and .Rmd) and
- and dataframes (save them to disk as *.Rda files)
  - save(foo,file="data.Rda")
- as part of the project.

In each class we will have a discussion section during Practicum,

- to discuss your progress, experiences and questions.

2.3 Final Semester Project Report Structure and Format

For DSCI352-DSCI452, the final data science research report should be written like a scientific paper

- and have the following types of sections.

- Title
- Author
- Author Affilication
- License: ideally CC-BY-SA 4.0 (but a license choice is yours)
- Abstract
- IntroductionModeling
- Data Science Methods
- Exploratory Data Analysis
- Statistical Learning: Modeling & Prediction(if appropriate)
- Discussion
- Conclusions
- Acknowledgements
- References, Citations

2.3.1 Abstract

Summary of the nature, finding and meaning of your data analysis project.

2.3.2 Introduction

Background and motivation of the Data Science question

2.3.3 Data Science Methods

To be applied (such as image processing, time-series analysis, spectral analysis etc

2.3.4 Exploratory Data Analysis

Results and steps in the data analysis

2.3.5 Statistical Learning: Modeling & Prediction

If your analysis can accomplish some modeling, include it here.

2.3.6 Discussion

Discussion of the answers to the data science questions framed in the introduction

2.3.7 Conclusions

2.3.8 Acknowledgements

2.3.9 References

2.4 How to make your report

The report is done as an Rmarkdown document, which can be run/compiled to produce two versions of the report as a pdf.

One shows your R code and figures, and the other doesn't show R code, just your figures.

You'll then turn in a zip file (and leave a copy in your repo), with the dataset (if its not to huge, if it is large, can you make a smaller dataset), Rmd file that works, and the two pdf reports.

Just choose to do a pdf report, instead of a set of presentation slides.

The license choice of CC-BY-SA 4.0 is suggested so that others can use and build on your codes, in an open-source manner.

With more restrictive licenses, others won't be able to use your code in the future.