ADS 503 - Applied Predictive Modeling

Summer 2024 - Week 1

Dave Hurst

University of San Diego
SHILEY-MARCOS SCHOOL OF ENGINEERING

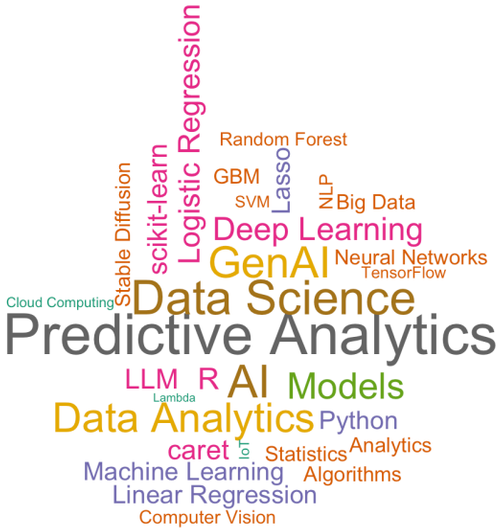
Start Recording!

University of San Diego
SHILEY-MARCOS SCHOOL OF ENGINEERING

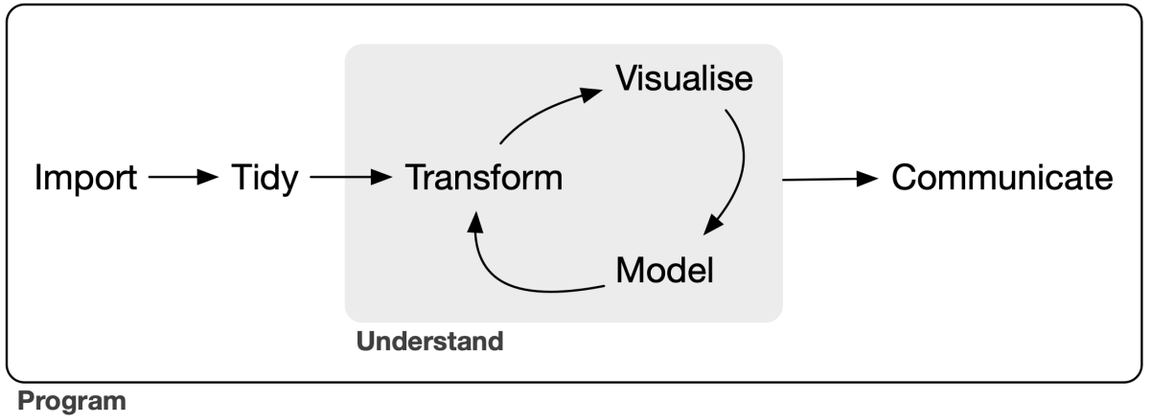
Agenda

- Course Introduction
- R / caret / Tidyverse
- Assignments
  - RStudio / Posit.cloud
  - Quarto
- Discussions
- Lab 1

Predictive Modeling



Model of the steps of a typical data science project[1]



$$\text{DataScience} = \text{DataAnalytics} + \text{Modeling}$$

$$\text{Modeling} \equiv \text{PredictiveModeling} \equiv \text{PredictiveAnaly}$$

$$\text{AI} = \text{TrainedModel}$$

1. https://r4ds.hadley.nz/intro

University of San Diego
SHILEY-MARCOS SCHOOL OF ENGINEERING

Models

Fundamentals

- Linear Regression
- Logistic Regression
- Penalized Regression (Lasso, etc.)

Big Data (data lakes)

- Deep Learning
  - Image recognition
  - LLM - etc.

Tabular Modeling (databases)

- SVM
- Specialized Models
  - eg. Quantile Regression

Open Source Modeling Frameworks

Open Source Frameworks

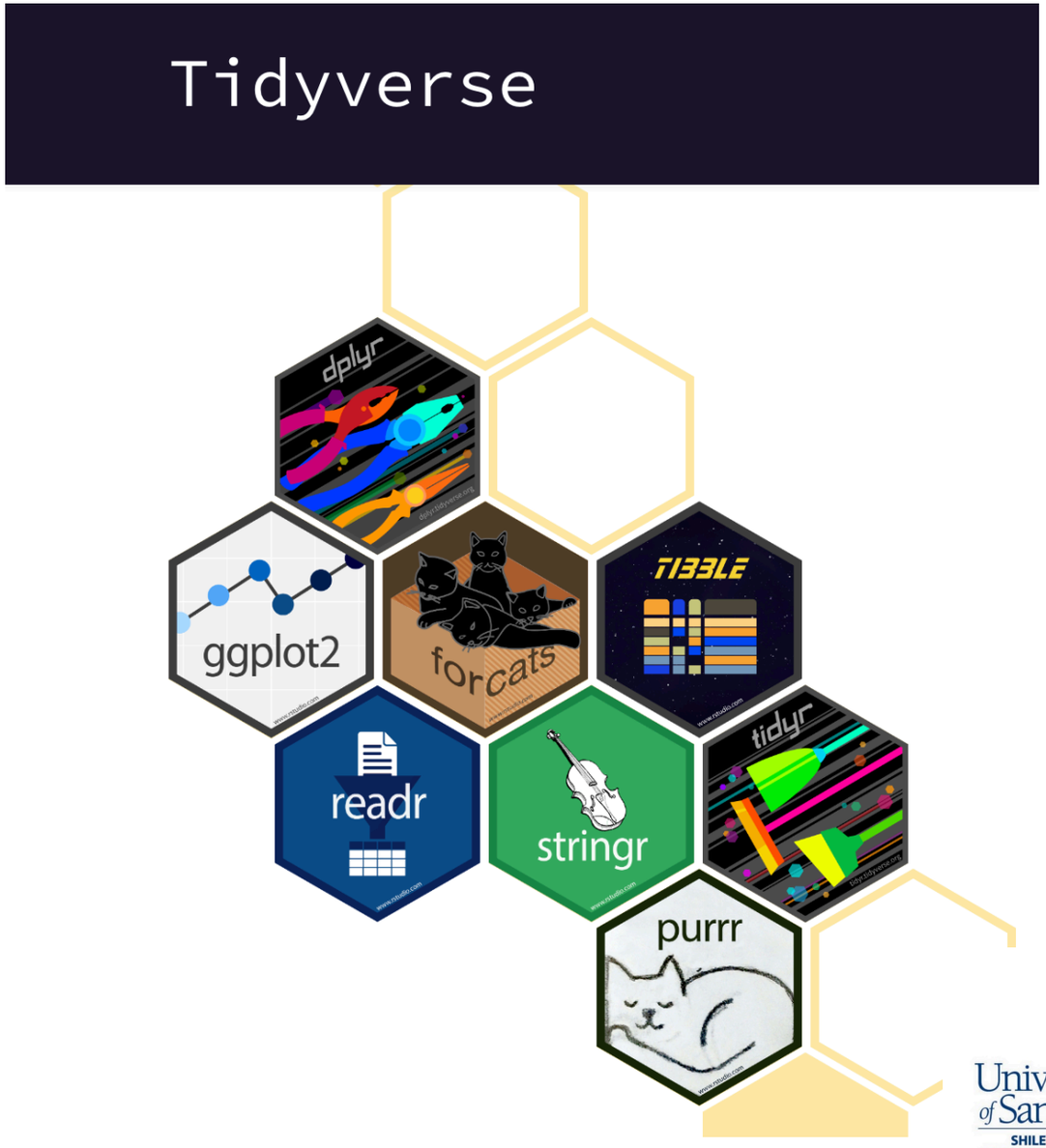
- caret - R based framework featured in Applied Predictive Modeling
- tidymodels - R based framework from the authors of caret
- scikit-learn - Python based framework
- H2O.ai - Python/R
- Tensorflow & PyTorch - Deep learning
- SparkML - Spark
- (Knime, Dataiku …… many many more)

Proprietary Frameworks

- SAS
- Mathmatica
- …

Thoughts about R

- User $\Longleftrightarrow$ Developer Spectrum[1]
- Scientific Domains (R)
  vs Computer Science (Python)
- Base R vs Tidyverse



1. (ref Roger Peng)

Assignments

- Original PDFs are available in Canvas
- Updated Templates will be made available in posit.cloud prior to each module
  - also distributed on Slack
  - using templates are not required but are highly recommended
- Up to 10% of the Assignment points rely on well formatted submissions
  - no data dumps
  - no excessive error lists
  - well labeled tables and plots
  - text and code should be selectable (do not paste images of text/code)
- Answer all questions.
- GenAI - policy
  - Understand what you submit
  - Make attributions where appropriate
    (link to private thread in code comments is adequate)

RStudio / posit.cloud

- RStudio can be downloaded and run locally …
- or via posit.cloud
  - Recommendation
    - Maintain a free account on posit.cloud
    - Sign up for the ADS 503 Space ($5/mo) (Institution = “University of San Diego”)
    - Class Space will be discontinued at the end of the course
    - Jupyter is also supported
    - If you plan to use R regularly, you should also run RStudio locally
    - Use `github` and `renv`
- Please rename your projects using the format: S{section_number}-{initials}-ADS503-Mn.
  Eg. S1-DAH-ADS503-M1

University of San Diego
SHILEY-MARCOS SCHOOL OF ENGINEERING

Quarto

- Quarto is the next generation of literate programing from the authors of R Markdown
- see Quarto.org
- Renaming a .Rmd to .qmd should work for starters
- Unlike R Markdown, Quarto can use Python natively and Jupyter in place of knitr
- (this presentation is made with Quarto)

Discussion 1.1

Original

Prior to completing this discussion forum, read IBM Pitched its Watson Supercomputer as a Revolution in Cancer Care. It's Nowhere Close

…

Replacement

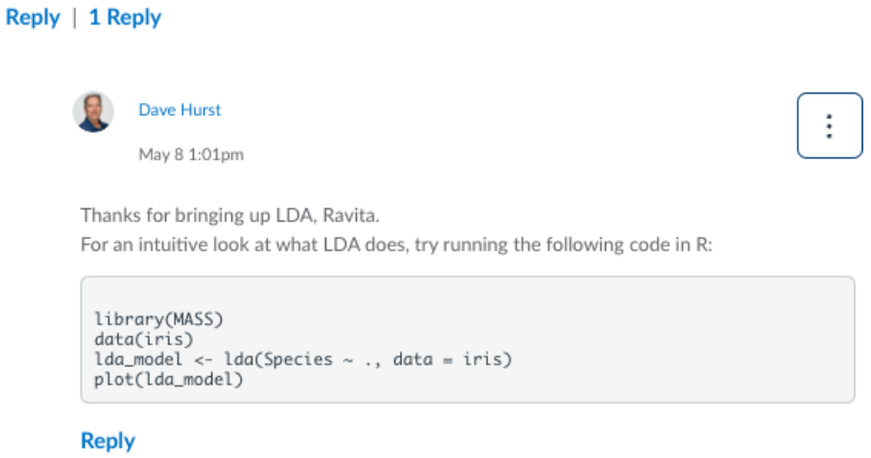
Generative AI (GenAI) and Large Language Models (LLM) in particular have changed the the data analytics landscape tremendously within the last 2 years. After performing some independent online research on your own, discuss what you think the major impacts for data analytics are, and how they might (or might not) extend into ***predictive*** analytics. You may choose your own topic, or dive into ***one*** of the following:

- Hallucinations are known problem with LLMs. How can companies take advantage of LLMs for predictive analytics while avoiding this pitfall.
- What do you expect the role of open source code languages such as R or Python will change as LLMs become more capable?
- There are several different types of Machine Learning frameworks. How would you categorize thes how do you think these categories will be affected by GenAI?

Discussions

Guidelines

- Be concise
- Use LLMs sparingly (grammar / fact checks are okay. Content generation is not.)
- Post meaningful responses … expound, challenge (politely), etc.



- Please DM us with ideas for future discussion topics.

University of San Diego
SHILEY-MARCOS SCHOOL OF ENGINEERING

Lab 1

University of San Diego
SHILEY-MARCOS SCHOOL OF ENGINEERING

Q&A

University of San Diego®
SHILEY-MARCOS SCHOOL OF ENGINEERING