Welcome to a GCDI Workshop!

Introduction to Predictive Modeling in R

Yuxiao Luo & Connor French

GC Digital Initiatives

THE GRADUATE CENTER
CITY UNIVERSITY OF NEW YORK

What is GCDI?

GCDI offers different types of ***support for digital scholarship*** such as evening **workshops**, **one-on-one consultations** for faculty and students, **working groups** based around common tools or data sources, special **events** such as our annual Digital Showcase and Sound Series, and **online resources**.

Whether you have participated in our workshops before or your idea of the perfect software is a paperback edition, there is something for you! Our offerings are open to scholars at all levels of digital experience and at all stages of graduate research. Whether you are digitally driven, curious, or defiant, we are prepared to help.

OK, so how can I get involved with GCDI?

- Join the **GCDI Group** on the CUNY Academic Commons for updates on how to be involved. **cuny.is/group-gcdi**
- Check the Event Calendar for upcoming **events** and **workshops**. **cuny.is/workshops**
- Request a **one-on-one consultation**. Faculty: **cuny.is/gcdfconsults**
- Join the **R Users Group (RUG)** at **cuny.is/rug**
- Join the **GIS Working Group** at **cuny.is/group-gis-working-group**
- Join the **Data Visualization Group** at **: cuny.is/dvg**
- Join the **Python Users Group (PUG)** at **cuny.is/pug**

GC Digital Initiatives

How else can I keep in touch with GCDI?

- Follow GCDI on Twitter: **@cunygcdi**
- Follow the GC Digital Fellows on Twitter: **@digital_fellows**
- Follow the **#digitalGC** hashtag on Twitter
- Follow the GC Digital Fellows blog, Tagging the Tower: **digitalfellows.commons.gc.cuny.edu**
- Contact the fellows at **digitalfellows.commons.gc.cuny.edu/contact-us/**
- Drop-in to the Digital Scholarship Lab **Open House** hosted once per semester, or attend PUG, Events, or Workshops!

Today's workshop

Introduction to Predictive Modeling in R

GC Digital Initiatives

Plan

- Introduction
- Build a modeling framework
- Predict the dataset
- Other resources

What is Machine Learning?

- There is no an agreed-upon definition

- Usually it means algorithms that can “learn” from data and address issues in different scenario

- These algorithms come from methods/models that have been developed in last 50 years from statisticians and computer scientists...

Scenarios of Using ML

- Predicting coupon redemption rates for a given marketing campaign
- Predicting default rate of a loan approved for different customers
- Segmenting customers based on collected attributes or historically purchasing behavior for targeted marketing
- Determining whether a photo is of a cat or a dog

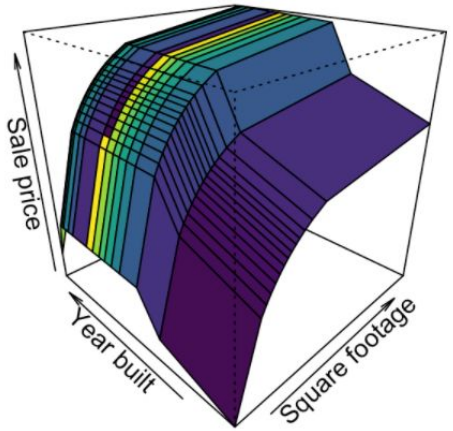
Business systems incorporated with ML:

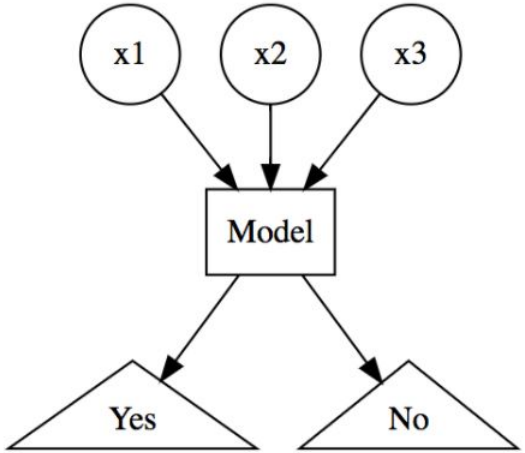
- Decision support system for managerial roles
- Operational intelligence and analytical system (data monitoring, IoT, etc..)
- ...

Terms in Machine Learning

Supervised learning: predictive models (predict an outcome given variables)

Unsupervised learning: descriptive models (clustering, dimension reduction..)



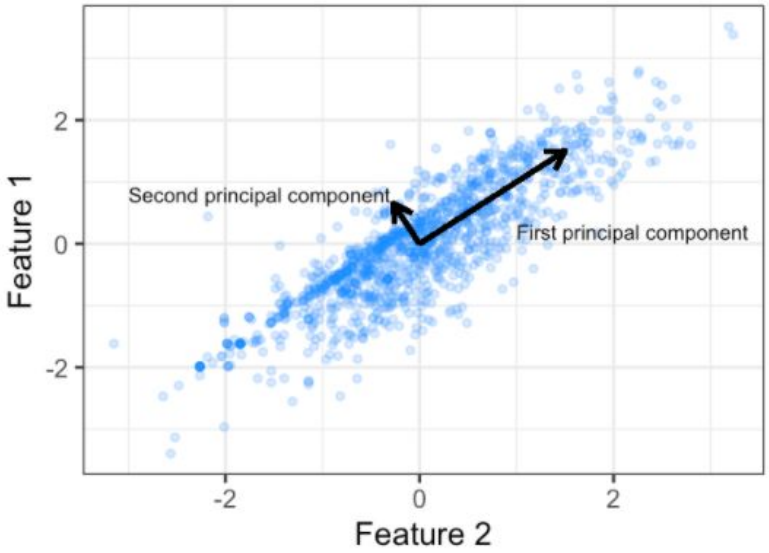


Supervised learning

- Supervised learning: predictive models
  - Predict a target given a bunch of predictors
  - Two types:
    - Regression: predicting a continuous outcome (usual linear regression)
    - Classification: predicting a categorical outcome (logistic/multinomial regression)

Unsupervised learning

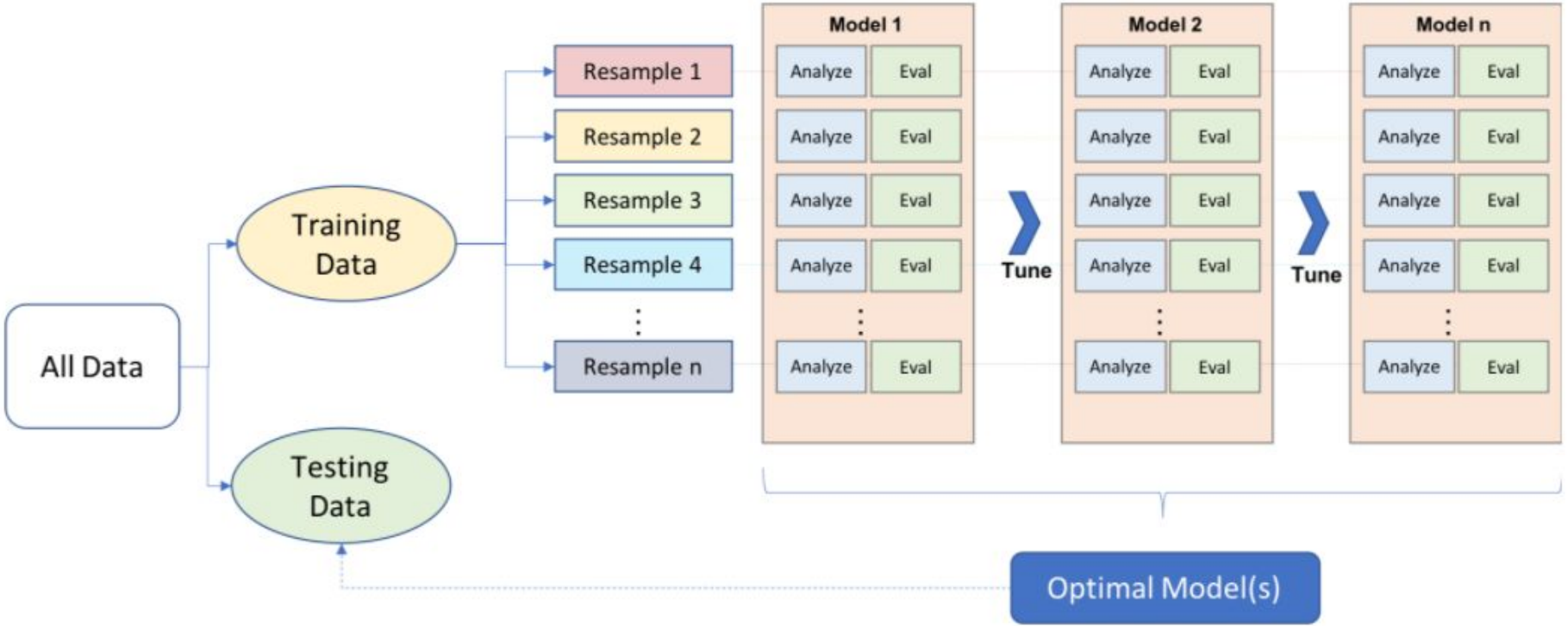
- Descriptive analysis of the dataset without target values
- Usually, unsupervised learning algorithms are
  - Cluster algorithms: creating groups of interesting observations or variables
  - Dimensionality reduction algorithms: reduce the dimensionality of a large dataset to something that we can visualize, try to lose as little information as possible [The goal is usually to find the “best” 2- or 3-dimensional representation of a dataset with many variables]



Modeling process in supervised learning

- The modeling pipeline:
  a. Data pre-processing: read in the data, data cleaning, transform/create new variables
  b. Model fitting: run model(s), estimate/ tune its parameters
  c. Model performance assessment: in ML, there are different measurement methods, we will introduce the most popular one here

Modeling process in supervised learning



Assess predictive performance

Split the dataset into 2 separate sets:

1. Training set: we use this dataset to look at the data and try out different models. We find out model(s) that fit the data well with the model goodness measures

2. Test set: we assess the predictive ability of the model(s) selected from the training set by evaluating the measure in predicting the test set (which were not used in step 1)

Assessing model performance in test set

- Regression (ex., numerical & continuous)
  - Mean squared error (MSE), mean absolute error, etc.

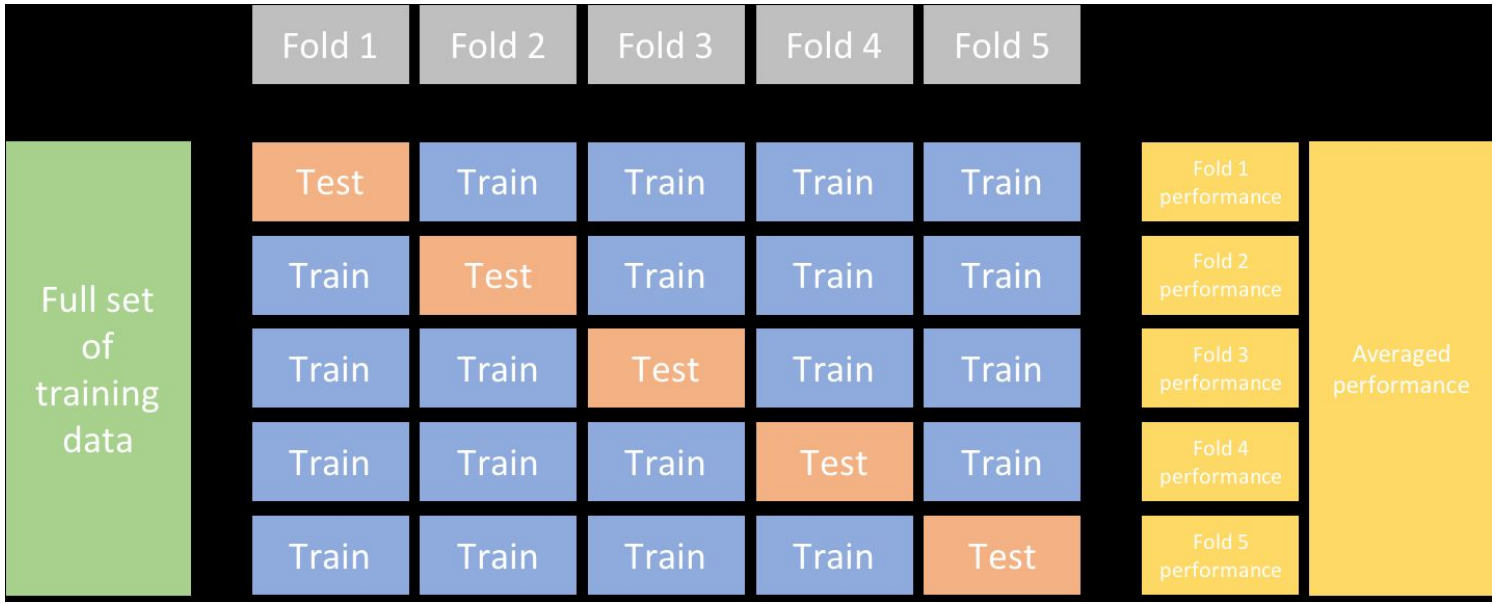
- Classification (ex., categorical)
  - Accuracy (% of observations that are well classified), precision, sensitivity, specificity, area under the curve (AUC))

Regularization & Cross Validation

- The predictive performance of linear regression models suffers when the predictors are highly correlated
  - Standard errors are inflated, giving rise to intervals that are wide
  - Small changes in predictors lead to big changes in predictions. Predictions are less robust.
- The predictive performance of the model suffers when there are overfitting issues
  - Overfitting is a modeling error that occurs when a function is too closely fit to a limited set of data points
  - The noise in the training data is captured and learned as concepts by the model

Without getting into technical details:

- Regularized regression adds additional parameters that reduce the variability in the estimation of the regression coefficient
- How those new parameters estimated (ex., lamda)? K-fold cross validation.



Implement regularized regression

- Regularized regression: lasso, ridge, elastic nets, etc…

- We will implement elastic nets using R package *glmnet* and tune the hyperparameter using R package *caret.*

- For more technical details, you can check out Chapter 6 of Hands on Machine Learning with R.

What is classification

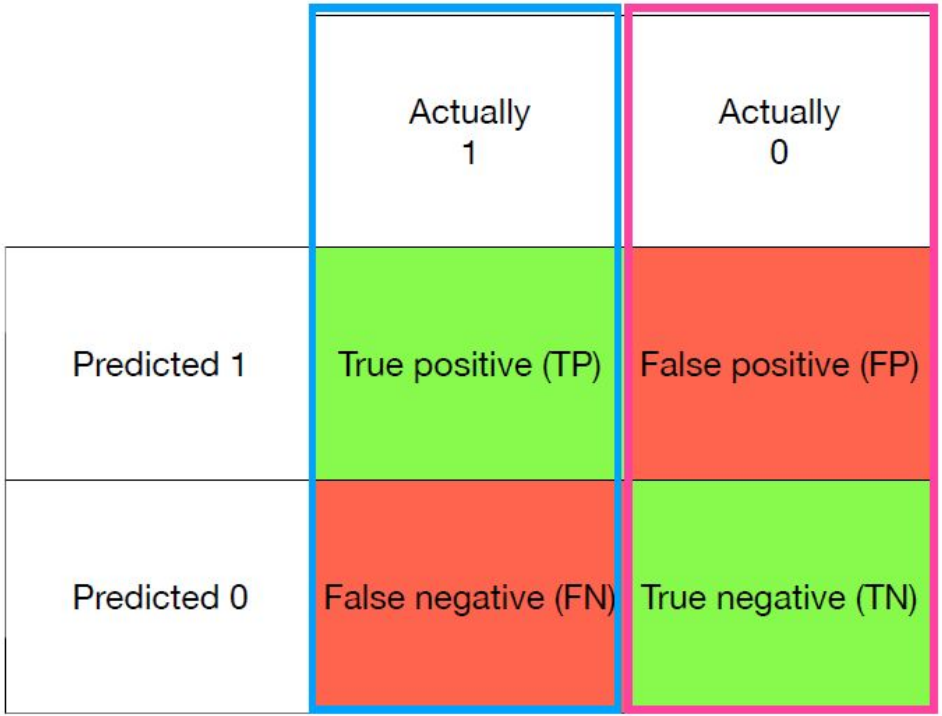
- Predicting a categorical outcome given a set of predictors
- Logistic and regularized logistic regression are commonly used classification methods (using here)
- There are other classification methods: tree-based model, k-means, etc…
- We will use a Titanic dataset today.
  - **Goal:** predict whether a passenger survived (1 = survived, 0 = didn't) using predictors (age, sex, class)
  - Split data into training and testing set
  - Train a model using training set
  - Predict survival in test set

Confusion matrix

	Actually survived (1)	Actually died (0)
Predicted to survive (1)	23	15
Predicted to die (0)	25	44

- Compare predicted values against real values
- Green: correct classifications
- Red: wrong classifications
- **Accuracy:**
  - Proportion of observations that are classified correctly
  - (23+44)/(23+44+15+25) ~ 0.626

Other metrics



- **Accuracy**:

  (TP + TN)/(TP+FP+FN+TN)

- Sensitivity: proportion of actual 1s that are classified correctly. TP/(TP+FN)

- Specificity: proportion of actual 0s that are classified correctly. TN/(TN+FP)

Example

	Actually survived (1)	Actually died (0)
Predicted to survive (1)	23	15
Predicted to die (0)	25	44

- **Accuracy**: proportions observations classified correctly

  (23+44)/(23+44+15+25) ~ 0.626

- Sensitivity: proportions of actual 1s that are classified correctly

  23/(23+25) ~ 0.479

- Specificity: proportions of actual 0s that are classified correctly

  44/(44+15) ~ 0.746

Conclusion

	Actually survived (1)	Actually died (0)
Predicted to survive (1)	23	15
Predicted to die (0)	25	44

Accuracy: proportions of observations classified correctly

~ .626

Sensitivity: proportions of actual 1s that are classified correctly ~ .479

Specificity: proportions of actual 0s that are classified correctly ~ .746

Our model is better at detecting deaths than survivals

Comparing models

- In practice, we might be considering more than one model (for example, we might fit a logistic regression and a regularized one)
- We can use metrics such as **accuracy**, sensitivity, and specificity to compare them
- Suppose we have 2 models
  - Model A: 0.725 accuracy, 0.5 sensitivity, 0.95 specificity
  - Model B: 0.675 accuracy, 0.85 sensitivity, 0.5 specificity

Which one is better? It depends..

Resources (all open-access)

- **Datacamp Free Course** (Introduction to Machine Learning for Every One)
- **Hands-On Machine Learning with R** by Bradley Boehmke & Brandon Greenwell
- **R Programming for Data Science** by Roger D. Peng
- **R for Data Science** by Hadley Wickham

Connect with the R community!

cuny.is/rug

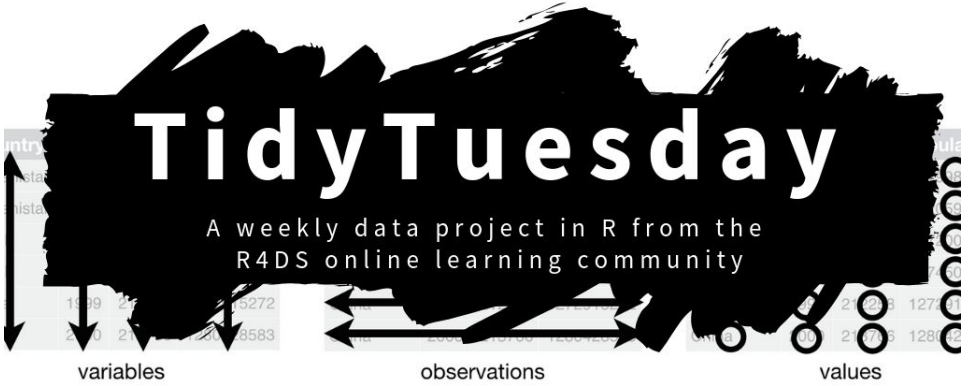
cuny.is/dvg

#rstats

#tidyverse

#r4ds

#rspatial



https://github.com/rfordatascience/tidytuesday

https://rladies.org/

https://www.rfordatasci.com/

Learn more!

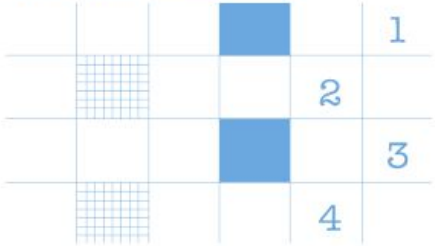
RStudio Primers

Learn data science basics with the interactive tutorials below.

The Basics

Start here to learn the skills that you will rely on in every analysis (and every primer that follows): how to inspect, visualize, subset, and transform your data, as well as how to run code.

Work with Data

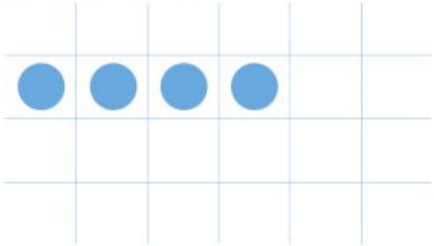


Learn the most important data handling skills in R: how to extract values from a table, subset tables, calculate summary statistics, and derive new variables.

Visualize Data

Learn how to use ggplot2 to make any type of plot with your data. Then learn the best ways to visualize patterns within values and relationships between variables.

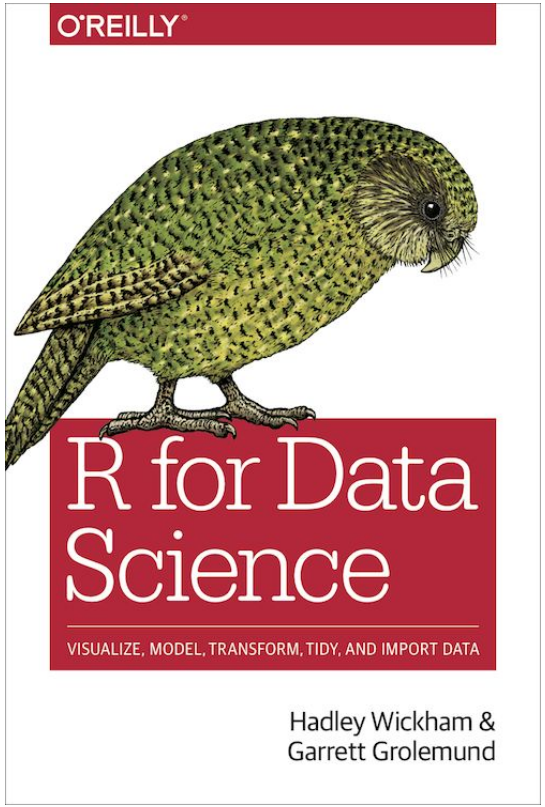
Tidy Your Data



Unlock the tidyverse by learning how to make and use tidy data, the data format designed for R.

https://rstudio.cloud/learn/primers

Learn more!



https://r4ds.had.co.nz/

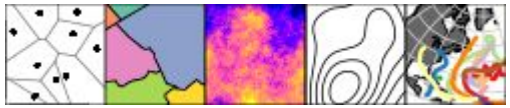
GC Digital Initiatives

http://cuny.is/gcdi

R Studio Community

https://community.rstudio.com/

r-spatial



https://www.r-spatial.org/

stackoverflow

https://stackoverflow.com/

What next?

Before you leave, **please fill out a workshop evaluation** so we can improve our programming! **cuny.is/gcdi-webevals**

- Need more support about what you just learned? Request follow-up individual consultations at **cuny.is/gcdi-consults**
- Drop-in to Office Hours for support with digital projects. For current dates, visit our website.
- Join the GCDI Group on the CUNY Academic Commons for all GCDI-related updates! **cuny.is/group-gcdi**

Thank you for attending and being involved! The #*DigitalGC* is you!

GC Digital Initiatives