Welcome to a GCDI Workshop!

Introduction to Predictive Modeling in R II
Yuxiao Luo & Rilquer Mascarenhas

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.

GC Digital Initiatives

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!

GC Digital Initiatives

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..)
- ...

Applications of ML

Model	Applications
Logistic Regression	Price prediction
Fully connected networks	Classification
Convolutional Neural Networks	Image processing
Recurrent Neural Networks	Voice recognition
Random Forest	Fraud Detection
Reinforcement Learning	Learning by trial and error
Generative Models	Image creation
K-means	Segmentation
k-Nearest Neighbors	Recommendation systems
Bayesian Classifiers	Spam and noise filtering

Terms in Machine Learning

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

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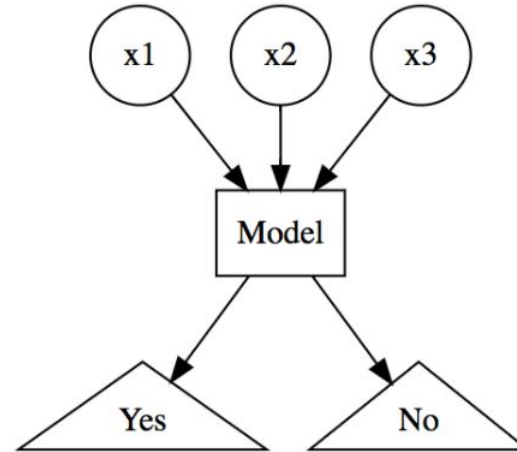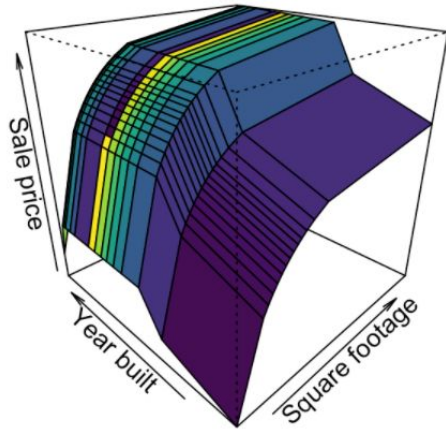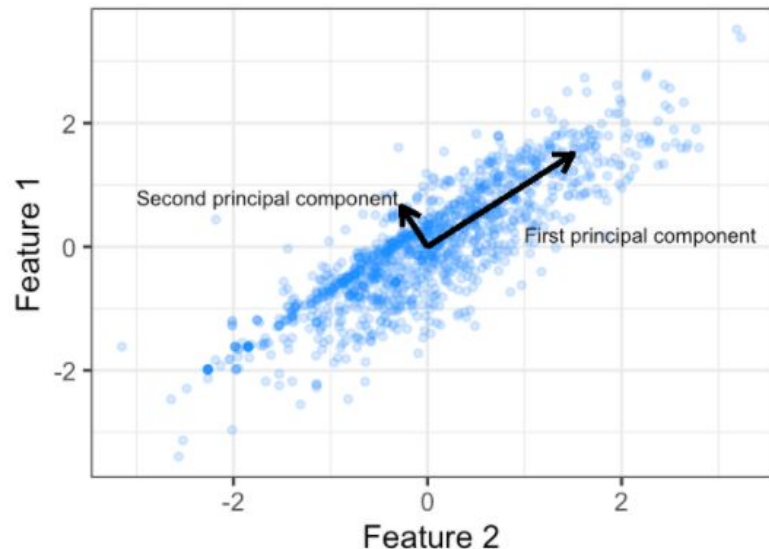
Supervised learning

- 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
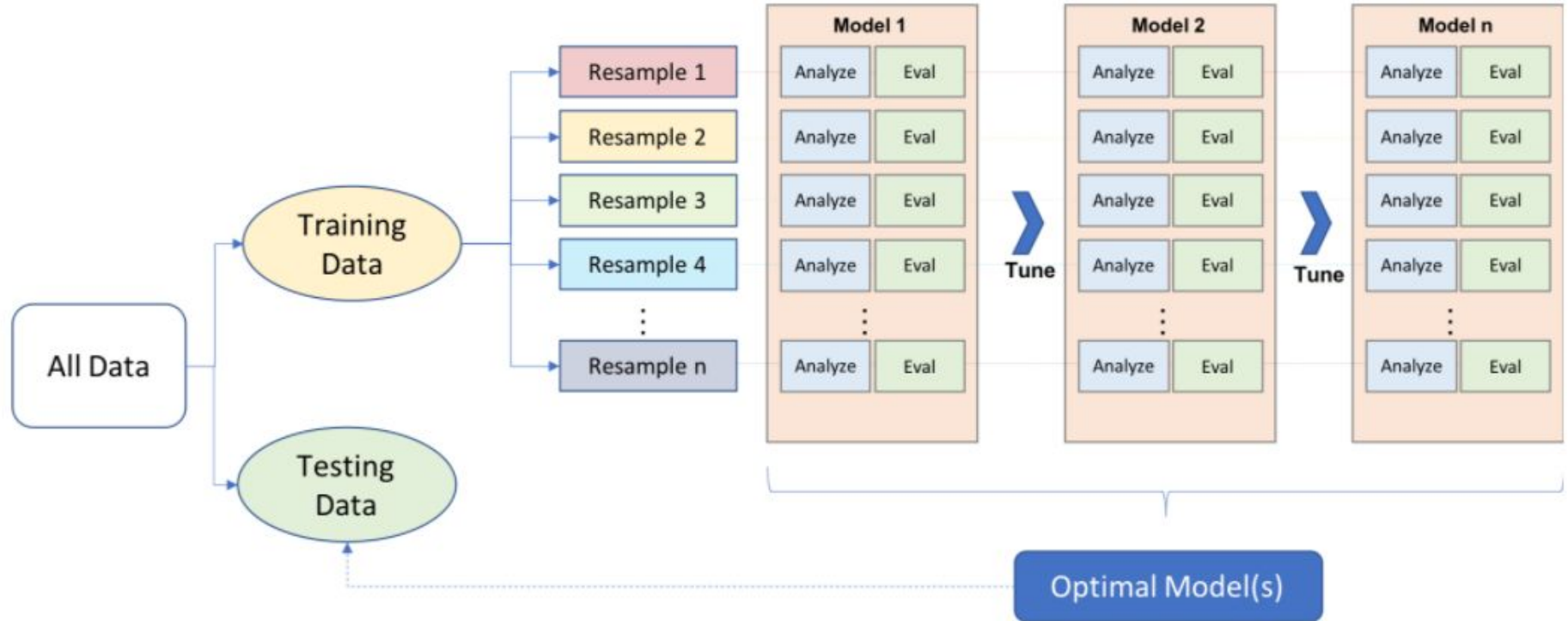
Assess predictive performance

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)
  - Root mean squared error(RMSE), 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))

Model Diagnostics

- 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

Overfitting Solutions

1. Regularization
   a. In the cost function, combining a penalty term and a tuning parameter that controls the relative impact of penalty term
   b. The idea is that too many predictors lowers the score in model selection
2. Cross validation: subsampling technique
3. Dimension reduction
   a. This approach involves projecting the p predictors into a M-dimensional subspace, where M < p. This is achieved by computing M different linear combinations, or projections, of the variables. Then these M projections are used as predictors to fit a linear regression model by least squares.
   b. Ex., principal component analysis (PCA), principal component regression (PCR)
4. Subset selection
   a. This approach involves identifying a subset of the p predictors that we believe to be related to the response. We then fit a model using least squares on the reduced set of variables.
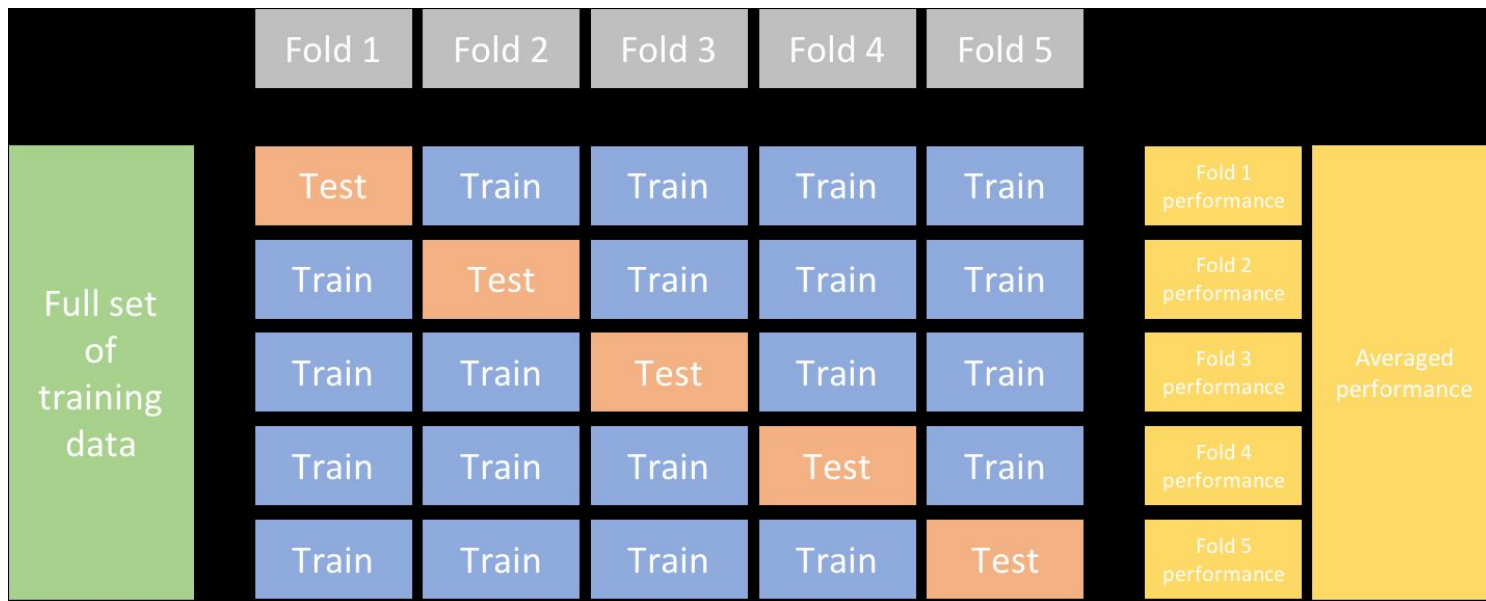
Reference: https://www.coursera.org/lecture/machine-learning/regularized-linear-regression-QrMXd

Regularization & Cross Validation

1. Obtains more "robust" estimates and often outperforms linear regression in certain prediction tasks when there are highly correlated predictors
    a. In linear regression, you might encounter multicollinearity problem and need to drop those factors
2. It can also handle datasets where there are more predictors than there are observations (with theoretical guarantees, under some conditions)
    a. In such situation, X^tX becomes singular matrix and its inverse doesn't exist for linear regression
    b. It leads to overfitting

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.

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5		
Full set of training data	Test	Train	Train	Train	Train	Fold 1 performance	
	Train	Test	Train	Train	Train	Fold 2 performance	
	Train	Train	Test	Train	Train	Fold 3 performance	Averaged performance
	Train	Train	Train	Test	Train	Fold 4 performance	
	Train	Train	Train	Train	Test	Fold 5 performance	

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.

Dataset

- Training Dataset: the sample of data used to fit the model.
- Validation Dataset: the sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyperparameters.
  - The validation dataset may also play a role in other forms of model preparation, such as feature selection.
  - Sometimes you don't see this because the practitioner choose to tune the model using the training set
- Test Dataset: the sample of data used to provide an unbiased evaluation of a final model fit on the training dataset.
  - The final model could be fit on the aggregate of the training and validation datasets.
  - Reference to the "test dataset" too may disappear if the cross-validation of model hyperparameters using the training dataset is nested within a broader cross-validation of the model.

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
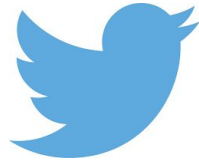
- **R for Data Science** by Hadley Wickham

Connect with the R community!
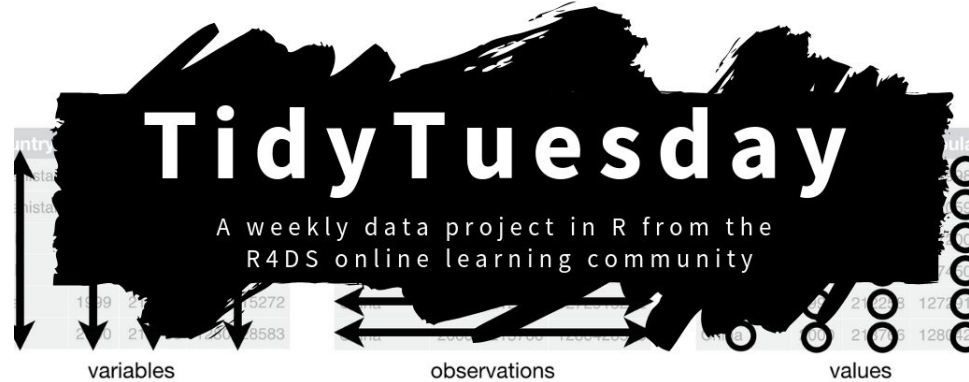

cuny.is/rug

DVG

cuny.is/dvg

#rstats

#tidyverse

#r4ds

#rspatial

variables   observations   values

https://github.com/rfordatascience/tidytuesday

Ladies

https://rladies.org/

r4ds
an online learning community
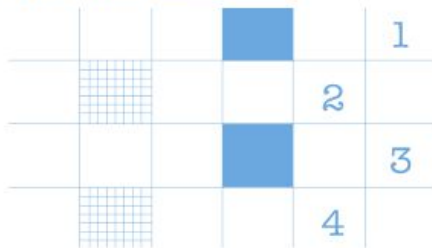
https://www.rfordatasci.com/

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.
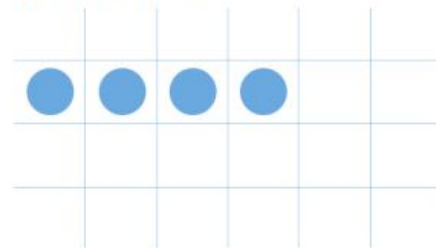
Work with Data

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.
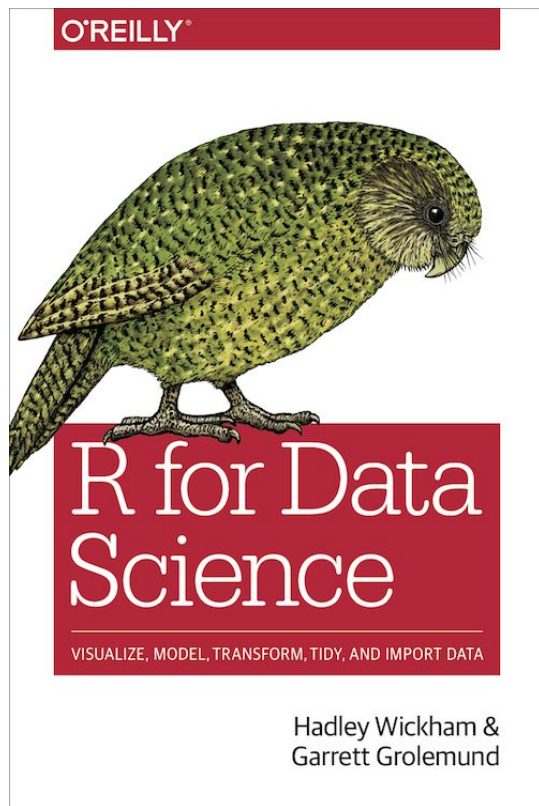
Tidy Your Data

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/

stack overflow

https://stackoverflow.com/
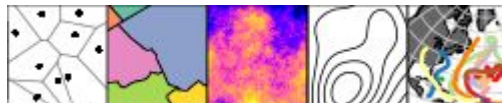
What next?

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