ISA 419: Data-Driven Security

14: Regression Analysis

Fadel M. Megahed, PhD

Endres Associate Professor
Farmer School of Business
Miami University

@FadelMegahed
fmegahed
fmegahed@miamioh.edu
Automated Scheduler for Office Hours

Spring 2024

Quick Refresher of Last Class

- ✅ Define clustering.
- ✅ Explain the $k-$means clustering algorithm for numeric datasets.
- ✅ Implement the $k-$means algorithm in Python using the `pycaret` package.
- ✅ Describe scenarios where other/advanced clustering algorithms are needed.

Learning Objectives for Today's Class

- Describe the basic concepts of regression analysis, including the roles of independent and dependent variables.
- Assess regression models using metrics like R-squared and Mean Squared Error and interpret the results.
- Describe the two modeling mindsets: explanatory and predictive.

What is regression analysis?

An Overview of Regression Models

- **Given:** $(x_1, y_1),\ (x_2, y_2),\ \ldots, (x_n, y_n)$.
- Learn a function $f(x)$ to predict y given x.
  - If y is a continuous variable, we do have a regression type problem.
  - In this class, we will **not** limit ourselves to simple and multiple linear regression models. Hence, we will also explore: (a) more advanced linear regression models (e.g., LASSO), (b) non-linear regression models (e.g., polynomial regression), and (c) machine learning models like decision trees, random forests, etc.
    - Simple linear regression: single predictor $x \rightarrow y$.
    - Multiple linear regression: multiple predictors $x_1, x_2, \ldots, x_p \rightarrow y$.

Simple Linear Regression

- Simple linear regression is a statistical method that allows us to summarize and study the relationship between a **response variable** and **one predictor variable**.

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

where:

- Y_i is the dependent variable for the i^{th} observation.
- X_i is the independent variable for the i^{th} observation.
- β_0 is the intercept, which is the value of Y when $X = 0$.
- β_1 is the slope, which is the change in Y for a one-unit change in X.
- ε_i is the error term, i.e., the difference between observed value of Y and its predicted value from the model.
    - It captures the effect of all other factors that influence the dependent variable.
    - We assume that the error term is $\sim N(0, \sigma^2)$.

Recall ISA 225: A Simple Linear Regression Exampl

10:00

- Using any software of your choice, analyze the zeroAccessUFO.csv dataset by answering the following questions:
  - Plot the relationship between the number of UFO sightings ($x=$ufo2010) and the number of zeroAccess infections ($y=$infections).
  - Use the plot to determine if a linear relationship exists between the two variables.
  - If a linear relationship exists, fit a simple linear regression model to the data.
  - Describe the goodness of fit of the model.
  - What is your conclusion?

"Miami Univeristy Links UFO Sightings to ZeroAcces Infections"

05:00

- **Question 1:** What happens if you: (a) add `pop` to your model to fit a multiple linear regression model? and (b) fit a single linear regression model using `pop` alone as the predictor?
- **Question 2:** Compute the correlation between `pop` and `ufo2010`.
- **Question 3:** Describe what you have learned from these two exercises.

- **Q1:** Edit me in your web browser.
- **Q2:** Edit me.
- **Q3:** Edit me.

On the Difference between R and Python

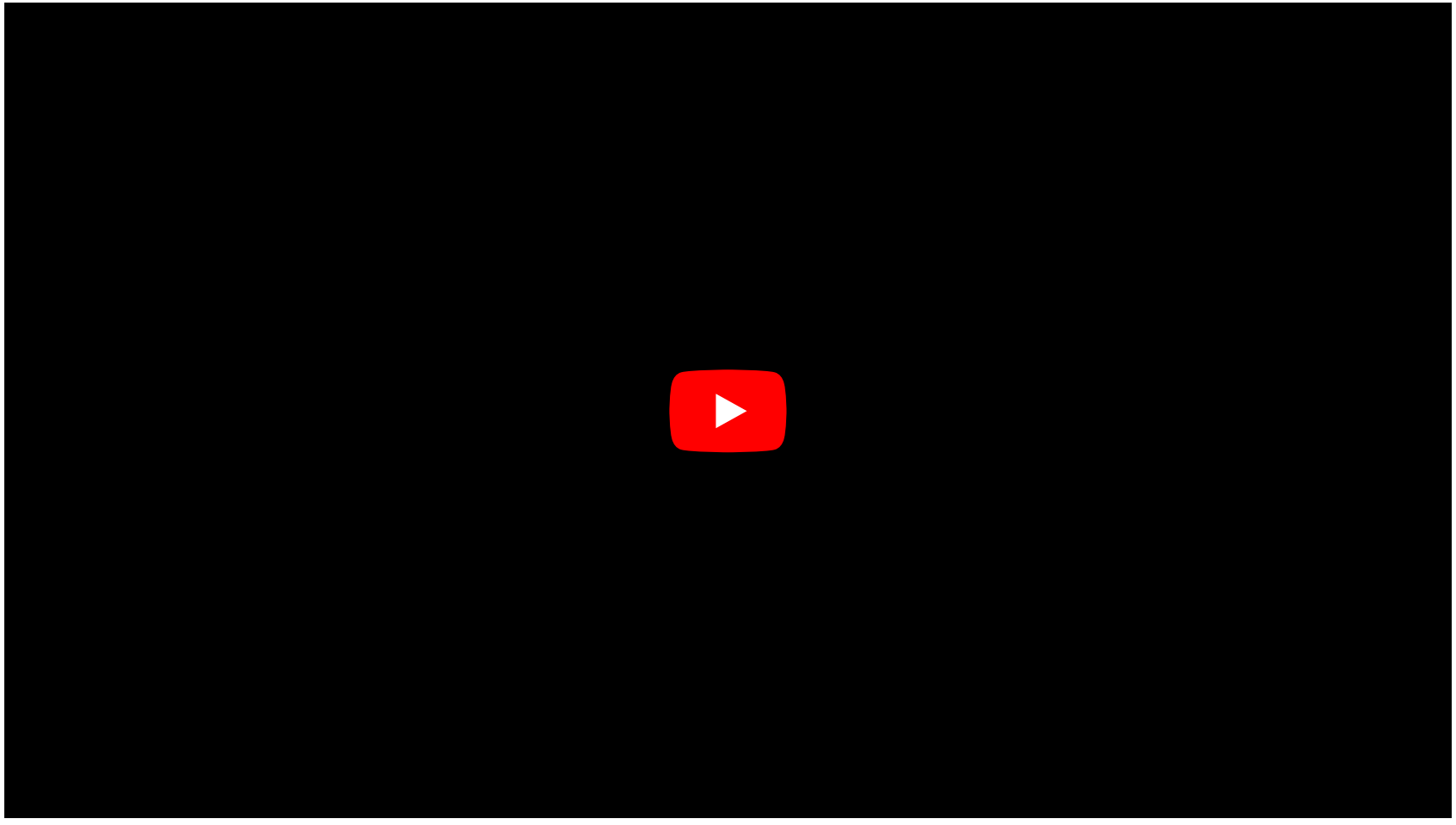
R and Python Implementations: A Brief Comparison

Aspect	R (using the 'stats' package)	Python (using 'PyCaret')
Primary Library	`stats`	`PyCaret`
Modeling Mindset	Explanatory (focus on understanding the relationships between variables)	Predictive (focus on accurately predicting outcomes)
Data Handling	Models often fit on entire dataset without explicit splitting (though splitting can be done manually)	Automatic splitting of data into training and testing sets, with cross-validation implemented on the training set
Implications	Emphasis on the quality of the model fit and understanding variable relationships	Emphasis on the model's predictive performance on unseen data
Model Extraction	Easier to extract and interpret models, facilitating detailed analysis of coefficients and statistical tests	More complex to extract specific model details due to a focus on prediction; models are often part of a larger pipeline
Model Evaluation	Emphasis on statistical tests and model fit statistics (e.g., R-squared, p-values)	Emphasis on predictive performance metrics (e.g., RMSE, R-squared, MAE)

The Two Modeling Mindsets: Explanatory Vs. Predictive

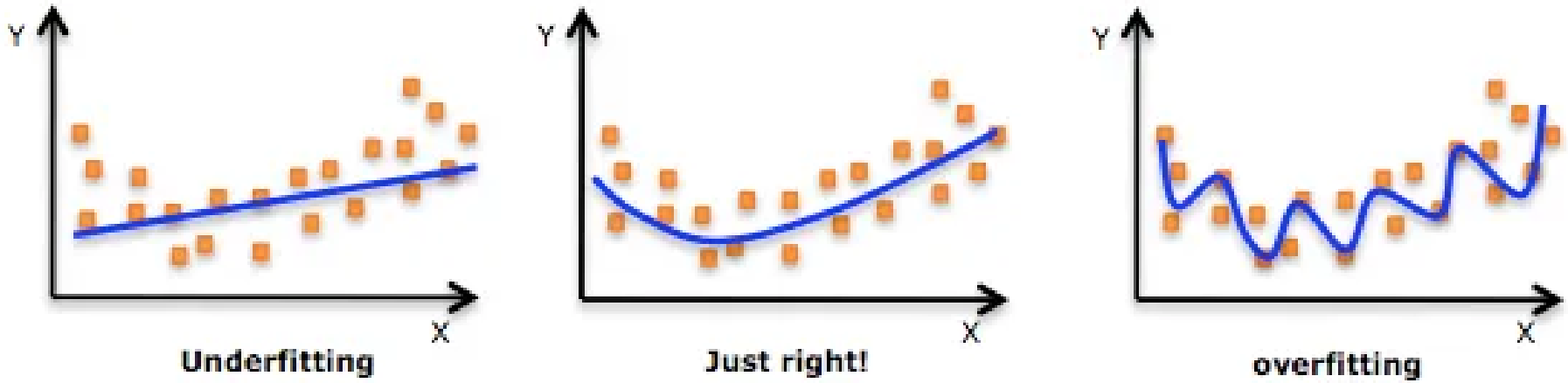
- **Explanatory Modeling**:
  - Focuses on understanding the relationships between variables.
  - Emphasizes the quality of the model fit and the understanding of variable relationships.
  - Easier to extract and interpret models, facilitating detailed analysis of coefficients and statistical tests.
  - Emphasis on statistical tests and model fit statistics (e.g., R-squared, p-values).
- **Predictive Modeling**:
  - Focuses on accurately predicting outcomes.
  - Emphasizes the model's predictive performance on unseen data.
  - More complex to extract specific model details due to a focus on prediction; models are often part of a larger pipeline.
  - Emphasis on predictive performance metrics (e.g., RMSE, R-squared, MAE).

The Predictive Modeling Mindset



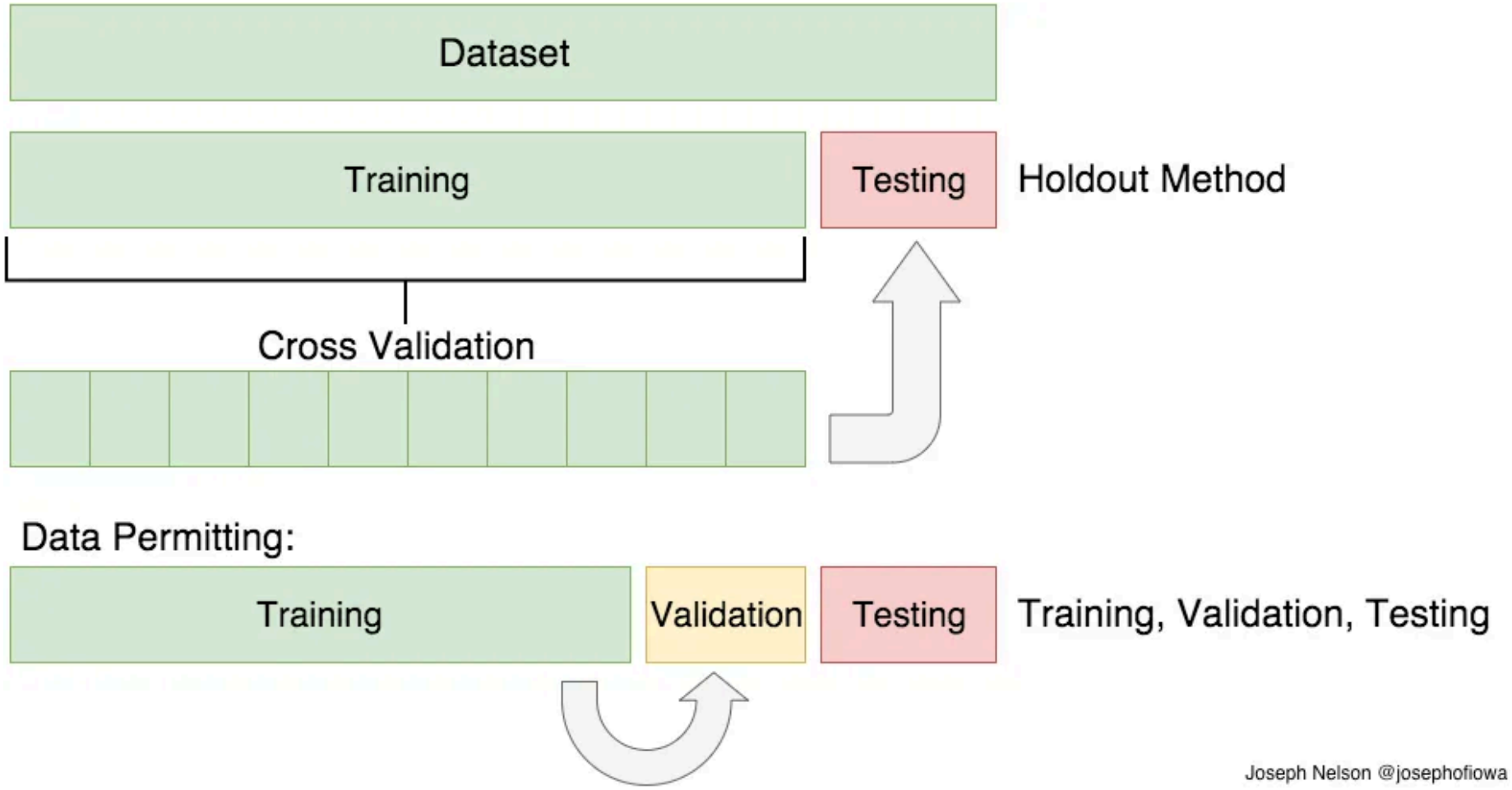
Note: While this video provides a classification example (i.e., a dichotomous dependent variable), the same steps apply to regression ML models. Hence, I am sharing it with you to illustrate the predictive modeling mindset.

Comments: Preparing the Data for Modeling



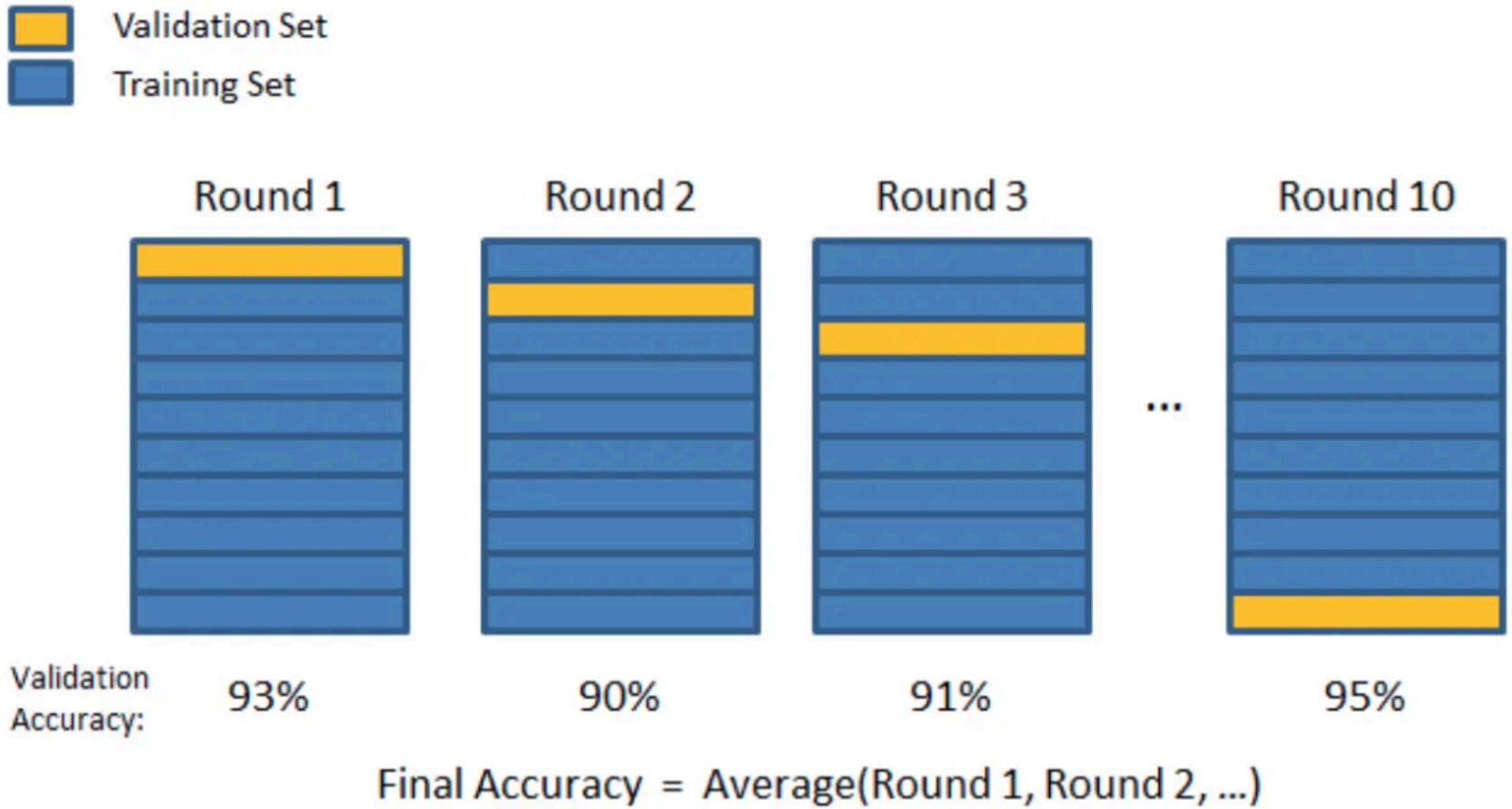
Source: Adi Bronshtein. (2017). Train/Test Split and Cross Validation in Python.

Comments: Preparing the Data for Modeling



Source: Adi Bronshtein. (2017). Train/Test Split and Cross Validation in Python. Figure is by Joseph Nelson.

Comments: Preparing the Data for Modeling



Source: Adi Bronshtein. (2017). Train/Test Split and Cross Validation in Python. Figure is by Joseph Nelson.

Recap

Summary of Main Points

By now, you should be able to do the following:

- Describe the basic concepts of regression analysis, including the roles of independent and dependent variables.
- Assess regression models using metrics like R-squared and Mean Squared Error and interpret the results.
- Describe the two modeling mindsets: explanatory and predictive.

 Review and Clarification 

- **Class Notes**: Take some time to revisit your class notes for key insights and concepts.
- **Zoom Recording**: The recording of today's class will be made available on Canvas approximately 3-4 hours after the end of class.
- **Questions**: Please don't hesitate to ask for clarification on any topics discussed in class. It's crucial not to let questions accumulate.