Workshop – R for Machine Learning: A Hands-On Introduction

Details

- Course Time: Monday, June 17th, 2019
  - Schedule:
    - Workshop starts at 8:30am
    - Morning Coffee Break at 10:30am – 11:00am
    - Lunch provided at 12:30pm – 1:15pm
    - Afternoon Coffee Break at 3:00pm – 3:30pm
    - End of the Workshop: 4:30pm
    - Coffee breaks and lunch are included.

Overview

This class will be an introduction to the statistical programming language R. We'll explore data science use cases and use R for data wrangling, data mining, visualization and prediction. Throughout the class we will be approaching problems analytically and we'll use R to explore data, make better decisions and identify areas for improving performance with prediction. The combination of data analytics, R and the data science process will provide the foundation for using R for data science problems. Students ***must*** come prepared with an understanding of computer programming and a curiosity for data science.

Goals

This is a "short course" of one day. Classes will be given in a lab setting, with student exercises mixed with lectures. Students ***must*** bring a laptop to class. There will be a modest amount of exercises to be completed in class. Due to the focused nature of this course, there will be no individual class projects but the instructors will be available to help students who are applying R to their own work outside of class.

Syllabus

Unit 1: Data Science and R Intro

- BIG DATA
- Data Science
- Roles in Data Science
- Use Cases
- Data'isms
- Class Format overview
- R Background
- R Intro
- R Studio

Unit 2: Visualize

- Rules of the road with data viz
- Chart junk
- Chart terminology
- Clean chart
- Scaling data
- Data Viz framework
- Code plotting

Unit 3: Data Analysis

- How to begin your data journey?
- The human factor
- Business Understanding
- EDA – Exploratory Data Analysis
- Data Anomalies
- Data Statistics
- Key Business Analysis Takeaways
- Diamond data set exercise
- Hands on challenge with Bank Marketing

Unit 4: Introduction to Regression

- Regression Definition

- Examples of regression
- Formulize the formula
- Plotting
- Statistical definitions involved
- mtcars regression example
- Business use case with regression

Unit 5: Introduction to Machine Learning

- ML Concept
- Types of ML
- CRISP Model
- Modeling
- Evaluation
- Titanic Example
- Dplyr
- Decision Trees
- Feature Engineering

Unit 6: Strategy

- Data Driven Decision Making
- Data Science Strategy
- Strategy Fails
- Macroeconomic strategy
- Adapting
- Data Science Project
- Data Impact
- Project guide
- Opportunities for improvement
- Big Box Store Strategic Exercise

INTRODUCTION TO DATA SCIENCE OVERVIEW

Data science has become the central approach to tackling data-heavy problems in both business and academia. In this course, students learn how data science is done in the wild, with a focus on data acquisition, cleaning, and aggregation, exploratory data analysis and visualization, feature engineering, and model creation and validation. Students use the R statistical programming language to work through real-world examples that illustrate these concepts. Concurrently, students learn some of the statistical and mathematical foundations that power the data-scientific approach to problem solving.

WHO IS THIS COURSE FOR?

The R Bootcamp is for anyone with a basic understanding of data analysis techniques and those interested in improving their ability to tackle problems involving multi-dimensional data in a systematic, principled way. A familiarity with the R programming language is helpful, but unnecessary, if the pre-work for the course is completed (more on that below).

PREREQUISITES

Students should have some experience with programming and have some familiarity with basic statistical and linear algebraic concepts such as mean, median, mode, standard deviation, correlation, and the difference between a vector and a matrix. In R, it will be helpful to know basic data structures such as data frames and how to use R Studio.

Students should complete the following pre-work (approximately 2 hours) before the first day of class:

1. R Programming - https://www.rstudio.com/online-learning/#R
2. R Studio Essentials Programming 1: Writing Code https://www.rstudio.com/resources/webinars/rstudio-essentials-webinar-series-part-1/

OUTCOME

Upon completing the course, students have:

An understanding of data science problems solvable using R and an ability to articulate those use cases from a statistical perspective.
The ability to create data visualization output. Familiarity with the R data science ecosystem, strategizing and the various tools to continue developing as a data scientist

--

STIPULATIONS

The above agenda, course structure and content is subject to change given time restrictions allotted for the course. Outcomes are not guaranteed. Students should download all course content during the class time as the content may not be available at a future date.