Data Science with Machine Learning

Curriculum

NYC DATA SCIENCE
ACADEMY

Program Objective

Data science is a fast-evolving field and offers many employment opportunities for people with a robust operational analysis background. In recent years, technological development in data collection and storage and innovations in data science tools and methodologies have made it even more important to have properly trained data analysts and data scientists to perform data analyses to gain business insights.

NYC Data Science Academy designed the Data Science with Machine Learning bootcamp to provide accelerated training to fulfill the need for data science professionals in the employment market. The objective of the Data Science Bootcamp is to provide training in primary data science tools and methods that prepares students for employment opportunities across all industries as data science professionals.

Program Description

The Data Science with Machine Learning bootcamp is an advanced certificate program that is designed primarily for individuals who have earned a baccalaureate or higher degree and want to further their career in the field of data science. It is a very accelerated training program in which students learn the major tools and methods for performing data analyses and apply them to various projects typically found in the data science field.

At the foundation level of the program, students learn to employ R and Python for data analytics projects and for presenting research results effectively. Beyond the foundational level, students study machine learning with Python and carry out research projects that involve advanced data science methods and strategies. The program also exposes students to concepts and practices in deep learning and big data.

Prework

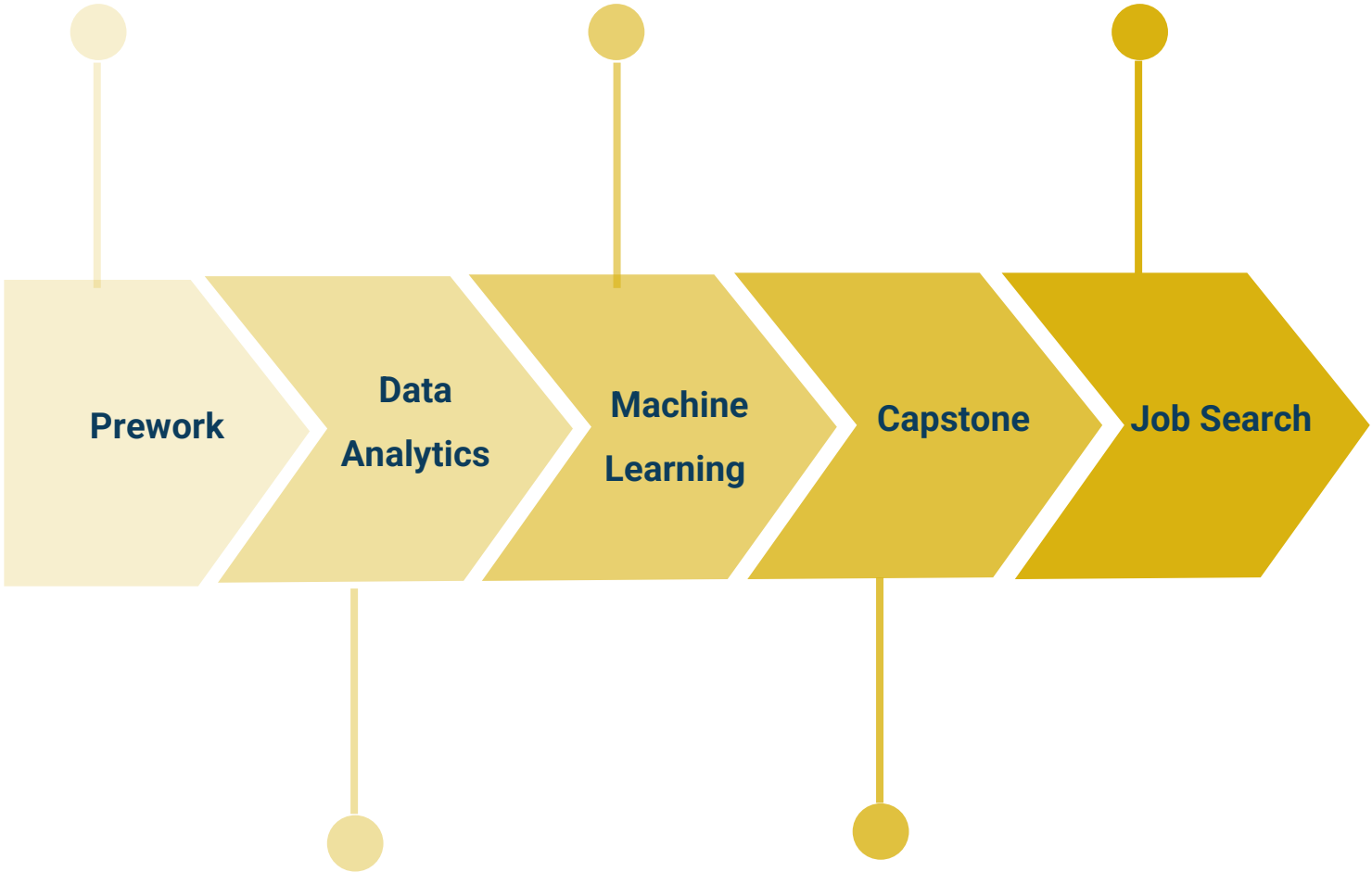
Access to online prework to prepare for an accelerated, immersive learning at NYC Data Science Academy.

Machine Learning I and II with Python

Foundation of statistics, regressions, classifications, model selections, unsupervised learning, and Machine Learning Project in a group.

Job Search

Career development strategies and post-bootcamp activities to seek employment in the data science field.



Data Analytics

Data analytics using R, Python, and other tools and methods; two application projects.

Capstone

Work with team members to design and complete a culminating project that uses real-world datasets.

Prework

Once students are enrolled in the bootcamp, they are granted access to our online, self-paced pre-work materials to get prepared in linear algebra, statistics, and some foundational work in coding.

Enrolled bootcamp students can also choose to take part-time, beginner-level courses hosted at our NYC campus for preparation. Tuition paid for such courses will be credited as part of bootcamp tuition.

Curriculum Topic Outline

Data Science Toolkit

- Linux system
    - Operating systems and Linux
    - File system and file operations
    - Text-processing commands
    - Other useful commands
- Git
    - What is version control and Git?
    - Installing Git
    - Getting started with Git
    - Git tips
    - Undoing changes
    - What is GitHub?
    - Working with remotes
- SQL: Part 1
    - What is SQL?
    - What does SQL do?
    - Basic terminology
    - Working with SQL hosted in a Docker container through GUI
    - Basic SELECT statements
    - Operators
    - Case statements
    - String operations
    - Date and time data
- SQL: Part 2
    - LIMIT statements
    - WHERE statements
    - Aggregating functions
    - GROUP BY statements
    - HAVING statements
    - ORDEWR BY statements
- SQL: Part 3
    - UNION and UNION ALL
    - Normalization
    - JOIN operations
    - Subqueries

- SQL: Part 4
    - Window functions
    - Comparison with Grouping and Aggregating
    - Application
- SQL: Part 5
    - Data manipulation
    - Loading data from Les
    - Data integrity

Data Analytics with R

- Introduction to R: Part I
    - Introduction to R
    - Introduction to RStudio
    - R objects
    - Functional programming: apply
- Introduction to R: Part II
    - More data types
    - Control statements
    - Functions
    - Data transformations
- Manipulating Data with dplyr and tidyr
    - Introduction to dplyr
    - Built-in functions
    - Join data sets
    - Groupwise operations
    - Reshape the layout
    - Split/Combine cells
- Data Visualization with "ggplot2"
    - Why ggplot2?
    - The “Grammar of Graphics”
    - Constructing a ggplot2 plot
    - Scatterplots
    - Bar charts
    - Histograms
    - Visualizing big data
    - Saving graphs
- Advanced ggplot2
    - Customized graphics
    - Titles
    - Coordinate systems

    - Scales
    - Themes
    - Axis labels
    - Legends
    - Other visualizations
- Introduction to Shiny
    - A quick introduction to shiny
    - Building a Shiny App from scratch
    - Improving your Shiny App
- Shiny Topics
    - GoogleVis
    - Leaflet
    - Shiny dashboard
- Foundations of Statistics
    - Descriptive statistics
    - Introduction to inferential statistics
    - Introduction to machine learning
- Advanced Statistics
    - Distributions
    - Going through codes

Project 1: Exploratory Visualization & Shiny

Data Analytics with Python

- Object Oriented Programming in Python
    - Object oriented programming
    - Portability and readability
    - Case study and examples
- Data Structures and Control Flows
    - Common data structures “out of the box”
    - Iterated and vectorized operations
    - Error handling
    - Case study and examples
- Advanced Topics
    - More class
    - Iterators and generators
    - Function arguments and unpacking
    - Case study and examples
- Basic Python Roundup
    - ETL case study
        - NumPy

        - An example
        - Getting started
        - Items/spider/pipelines/settings.py
        - In class lab
- NumPy
    - NumPy overview
    - Ndarray
    - Vectorized/Element-wise Operations
    - Subscripting and Slicing
    - Intro to Matrix Multiplication and Inversion
    - Random number generation
    - Case study
- Data Manipulations with Pandas
    - Series and common operations
    - Data frames and common operations
    - Data manipulation
    - Time series
- Data Visualization in the NumPy Stack
    - Matplotlib
    - Seaborn
    - Case study
- SciPy and Data Analysis Roundup
    - Introduction to SciPy
    - Case study

Project 2: Data Analytics with Web Scraping

Business Cases in Data Science

- Role of Data Science and DS Professionals
    - Data-driven decision making in the business world
    - The data professional's ecosystem
    - Data science as a strategic asset for businesses across industries
    - The roles of a data analyst and data scientist
    - Survey of storytelling with data
- Data Science Solutions to Business Problems
    - Key business operational concepts
    - Typical business problems
    - Role of data to solve business problems
    - The process of data analytics
    - Data science modeling - supervised and unsupervised
    - Anatomy of a business case with data analysis for marketing
- Data Science Solutions to Business Problems

    - General business problems in the healthcare and pharmaceutical industries
    - Characteristics of data science projects in these industries
    - Anatomy of a business case wherein data analytics offered a solution to the problem
- Data Science Solutions to Business Problems
    - General business programs in the financial industry
    - Characteristics of data science projects in the financial industry
    - Anatomy of a business case in which data science offered an effective solution to a business problem
- Conceptualization of a Data Analytics Problem
    - Planning process for a data analytics project
    - Understand the business problem
    - Data available for the project
    - Review of dataset(s) to see what knowledge and insights can be extracted
    - Establish project objective(s)
    - Preparing data for analysis
    - Determine tools and methods
    - Envision the approach
    - Identify deliverables
    - Tell the story to stakeholders

Machine Learning I

- Simple Linear Regression
    - Overview of Machine Learning
    - Residuals, RSS, and the Coefficient of Determination.
    - Assumptions of Simple Linear Regression.
    - Model coefficients
        - Interpretation
        - Standard Errors
- Multiple Linear Regression
    - Assumptions of Multiple (General) Linear Regression
    - Coefficients of Continuous Features
    - Multicollinearity and Overfitting
    - Dummifying Categorical features
        - Coefficient Interpretation
    - Case Study
- Penalized Linear Regression
    - The Bias-Variance tradeoff
    - Euclidean (L2) vs Manhattan (L1) Distances

  - Ridge, Lasso, and ElasticNet Penalties
  - Hyperparameter grid search and the model coefficient shrinkages

- Logistic Regression and Gradient Descent
  - The failure of Linear Regression Models for Classification
  - Log-odds and Sigmoid function
  - The concept of Maximum Likelihood Estimation
  - Negative Log-Likelihood as Loss
  - Gradient Descent and Stochastic Gradient Descent

- Model Selection
  - Model assessment in practice
    - Grid search
    - Train-test split
  - Model comparison
    - Tuning
    - Various model types
  - Thoughts on Feature Engineering
    - Different features make different models
  - Case study

- Discriminant Analysis and Naïve Bayes Model
  - Conditional Probability and Bayes Theorem
  - Univariate and Multivariate Gaussian Distributions.
  - Bayes classifiers
  - Linear and Quadratic discriminant analysis and the curse of dimensionality
  - Naive Bayes models
    - Gaussian
    - Bernoulli
    - Multinomial
  - Case study

- Time Series
  - Introduction – the nature of time series analysis
  - Univariate time series data
    - Decomposition
    - Correlation and Autocorrelation
    - Stationarity
    - AR/MA/ARIMA models
  - Visit to Multivariate
  - Case study

Machine Learning II

- Tree-Based Models
    - Decision Trees
    - Bagging Trees
    - Random Forests
    - Case study
- Gradient Boosting
    - First step to increase stability
    - OOB Errors and assessment
    - Comparison to the Decision Tree
- Random Forests
    - Big picture behind Gradient Boosting
    - Gradient Boosting with Trees
    - Case study
- Support Vector Machines
    - Maximum Margin Classifiers
    - Support Vector Classifiers
    - Support Vector Regressors
    - Bias/Variance with SVM
    - Case study
- Supervised Learning Roundup
    - Regressors
    - Classifiers
    - Scalability and further topics
- Unsupervised 1: Clustering
    - KMeans
    - Hierarchical models
    - Case study
- Unsupervised 2: Matrix Factorization
    - Principal component analysis
    - The other LDA: Latent Dirichlet Analysis
    - Case studies
- Machine Learning Roundup
    - Shoutout to those forgotten
    - Working together: where unsupervised and supervised cooperate
    - Case study
- Machine Learning Project

Project 3: Machine learning on housing price

Data Science: Advanced Topics

- Neural Network Foundations

    - Revisit Logistics Regression
    - Introduction to the Multilayer Perception
    - Comparison of where Classical ML and DL perform well
    - Introduction to keras and Building the First MLP for Classification
- Neural Network Architectures
    - Dealing with overfitting
    - CNN
    - RNN
    - Auto encoders
    - GAN
- Neural Networks Beyond
    - Transfer learning
    - Case study
- Scalability in Operation
    - From desktop to database to distributed
        - Hadoop
    - Big data on a small scale
        - Down-sampling
- Cloud Computing and Deployment
    - To Cloud or Server?
    - Current Cloud platforms and offerings
        - AWS
        - Google
    - Survey of deployment and pipelines

Data Science Capstone Project

The capstone project is designed for students to employ the major data science concepts, tools, and methods they have learned in the program to solve a business operational problem with real data sets from a real business entity. Students are presented data sets and potential problems to solve. Students are then required to form project teams, develop a project proposal for instructor review and approval, and execute the project. When the project is completed, each project team is required to present the project findings and share the business insights obtained from the research.