

## Elements of Machine Learning With Spark and Python

### Course Summary

#### Description

This course is taught using Spark and Python.

#### Objectives

After taking this course, students will be able to:

- Understand popular machine learning algorithms, their applicability, and limitations
- Practice the application of these methods in the Spark machine learning environment
- Understand practical use cases and limitations of algorithms

#### Topics

- Machine Learning (ML) Overview
- Machine Learning in Python and Spark
- Machine Learning Concepts
- Feature Engineering (FE)
- Linear Regression

#### Audience

This course is designed for data scientists and software engineers.

#### Prerequisites

Before taking this course, you should have a working knowledge of Apache Spark. If students are new to Apache Spark, we can offer one day of 'Introduction to Spark' training. Students need a programming background. Familiarity with Python would be a plus, but not required. No machine learning knowledge is assumed.

#### Duration

Three days

## Elements of Machine Learning With Spark and Python

### Course Outline

- I. **Machine Learning (ML) Overview**
  - A. Machine Learning landscape
  - B. Machine Learning applications
  - C. Understanding ML algorithms & models (supervised and unsupervised)
  
- II. **Machine Learning in Python and Spark**
  - A. Spark ML Overview
  - B. Introduction to Jupyter notebooks

**Lab: Working with Jupyter + Python + Spark**  
**Lab: Spark ML utilities**
  
- III. **Machine Learning Concepts**
  - A. Statistics Primer
  - B. Covariance, Correlation, Covariance Matrix
  - C. Errors, Residuals
  - D. Overfitting / Underfitting
  - E. Cross-validation, bootstrapping
  - F. Confusion Matrix
  - G. ROC curve, Area Under Curve (AUC)

**Lab: Basic stats**
  
- IV. **Feature Engineering (FE)**
  - A. Preparing data for ML
  - B. Extracting features, enhancing data
  - C. Data cleanup
  - D. Visualizing Data

**Lab: data cleanup**  
**Lab: visualizing data**
  
- V. **Linear regression**
  - A. Simple Linear Regression
  - B. Multiple Linear Regression
  - C. Running LR
  - D. Evaluating LR model performance

**Lab**  
E. Use case: House price estimates