# Machine Learning With Apache Spark – 3 days

Machine Learning (ML) is changing the world. To use ML effectively, one needs to understand the algorithms and how to utilize them. This course provides an introduction into the most popular machine learning concepts. For each machine learning concept, we first discuss the foundations, its applicability and limitations. Then we explain the implementation/use, and specific use cases. This is achieved through a combination of about 50% lecture, 50% lab work to maintain an optimal balance of theory and practice.

For the lab work, we will utilize Apache Spark and Amazon SageMaker (a fully managed machine learning platform offered by Amazon).

**Please note the following:**

* This is an introductory to intermediate course.   
  Students interested in gaining in depth knowledge about data science and machine learning should inquire about the 6 month training program that is being run in parallel.
* In-depth coverage of Math / Stats is behind Machine Learning is beyond the scope of this course.
* The emphasis of this course is on ML, not SageMaker or Spark.
* This course is taught using Spark & Python.

# Objectives:

* Learn popular machine learning algorithms
* Practice the application of these algorithms using Apache Spark and Amazon Sagemaker

# What you will learn

* ML Concepts
* Regressions
  + Linear Regression
  + Logistic Regressions
* Classifications
  + Naïve Bayes
  + SVM
  + Decision Trees
  + Random Forest
* Clustering (K-Means)
* Principal Component Analysis (PCA)

# Duration

3 days

# Audience

Data analysts, Software Engineers.

# Skill Level

Beginner to Intermediate

# Prerequisites

* Good programming background
* familiarity with Python would be a plus, but not required
* No machine learning knowledge is required
* No Spark knowledge is required

# Lab environment:

Working Spark environment in the cloud will be provided for students.

Students would run all the labs in the provided cloud environment.

# What to Bring

* A modern laptop
* The class would need open access to Internet.   
  Highly restrictive firewalls/VPNs might prevent the participant from connecting to the cloud environment.
* SSH client : Putty for Windows, Mac/Linux already has ssh clients installed

# Detailed Course Outline:

**Section 1: Machine Learning (ML) Overview**

* Machine Learning landscape
* Machine Learning applications
* Understanding ML algorithms & models (supervised and unsupervised)

**Section 2: ML in Python and Spark**

* Spark ML Overview
* Introduction to Jupyter notebooks
* Lab: Working with Jupyter + Python + Spark
* Lab: Spark ML utilities

**Section 3: Machine Learning Concepts**

* Statistics Primer
* Covariance, Correlation, Covariance Matrix
* Errors, Residuals
* Overfitting / Underfitting
* Cross validation, bootstrapping
* Confusion Matrix
* ROC curve, Area Under Curve (AUC)
* Lab: stats

**Section 4: Feature Engineering (FE)**

* Preparing data for ML
* Extracting features, enhancing data
* Data cleanup
* Visualizing Data
* Lab: data cleanup
* Lab: visualizing data

**Section 5: Linear regression**

* Simple Linear Regression
* Multiple Linear Regression
* Running LR
* Evaluating LR model performance
* Labs
* Use case : House price estimates

**Section 6: Logistic Regression**

* Understanding Logistic Regression
* Calculating Logistic Regression
* Evaluating model performance
* Labs
* Use case: credit card application, college admissions

**Section 7: Classification: SVM (Supervised Vector Machines)**

* SVM concepts and theory
* SVM with kernel
* Labs
* Use case: Customer churn data

**Section 8: Classification: Decision Trees & Random Forests**

* Classification and Regression Trees (CART) introduction
* Decision Tree concepts
* Pruning trees
* Gini index
* Bias Variance Tradeoff
* Random Forest concepts
* Random Forests features and examples
* Labs
* Use case: predicting loan defaults, estimating election contributions

**Section 9: Classification: Naive Bayes**

* Naïve Bayes theory
* Example walkthrough
* limitations
* Lab
* Use case: spam filtering

**Section 10: Clustering (K-Means)**

* Theory behind K-Means
* Running K-Means algorithm
* Estimating the performance
* Lab
* Use case: grouping uber trips data, grouping shopping data

**Section 11: Introduction to Amazon SageMaker**

* Sagemaker architecture, design
* Running algorithms on SageMaker