Course Length
2 days

Course Description
Data Science Projects with Python will help you get comfortable with using the Python environment for data science. This course will start you on your journey to mastering topics within machine learning. These skills will help you deliver the kind of state-of-the-art predictive models that are being used to deliver value to businesses across industries.

Overview
Data Science Projects with Python is designed to give you practical guidance on industry-standard data analysis and machine learning tools in Python, with the help of realistic data. The course will help you understand how you can use pandas and Matplotlib to critically examine a dataset with summary statistics and graphs and extract the insights you seek to derive. You will continue to build on your knowledge as you learn how to prepare data and feed it to machine learning algorithms, such as regularized logistic regression and random forest, using the scikit-learn package. You'll discover how to tune the algorithms to provide the best predictions on new and, unseen data. As you delve into later chapters, you'll be able to understand the working and output of these algorithms and gain insight into not only the predictive capabilities of the models but also their reasons for making these predictions.

By the end of this course, you will have the skills you need to confidently use various machine learning algorithms to perform detailed data analysis and extract meaningful insights from data.

Scope
Data Science Projects with Python takes a case study approach to simulate the working conditions you will experience when applying data science and machine learning concepts. You will be presented with a problem and a data set and walked through the steps of defining an answerable question, deciding what analysis methods to use, and implementing all of this in Python to create a deliverable.

Target Audience
If you are a data analyst, data scientist, or a business analyst who wants to get started with using Python and machine learning techniques to analyse data and predict outcomes, this course is for you. Basic knowledge of computer programming and data analytics is a must. Familiarity with mathematical concepts such as algebra and basic statistics will be useful.

Technical Requirements

Hardware:

For an optimal student experience, we recommend the following hardware configuration:

- Processor: Intel Core i5 or equivalent
- Memory: 4GB RAM (8 GB Preferred)
- Storage: 35 GB available space

Software:

You'll also need the following software installed in advance:

- OS: Windows 7 SP1 64-bit, Windows 8.1 64-bit or Windows 10 64-bit, Ubuntu Linux, or the latest version of OS X
- Browser: Google Chrome/Mozilla Firefox Latest Version
- Notepad++/Sublime Text as IDE (Optional, as you can practice everything using Jupyter notecourse on your browser)
- Python 3.4+ (latest is Python 3.7) installed (from https://python.org)
- Python libraries as needed (Jupyter, Numpy, Pandas, Matplotlib, BeautifulSoup4, and so on)

Installation and Setup

Before you start this course, make sure you have installed the Anaconda environment as we will be using the Anaconda distribution of Python.

Course Outline

Lesson 1: Data Exploration and Cleaning

- Python and the Anaconda Package Management System
- Different Types of Data Science Problems
- Loading the Case Study Data with Jupyter and pandas
- Data Quality Assurance and Exploration
- Exploring the Financial History Features in the Dataset
- Activity 1: Exploring Remaining Financial Features in the Dataset

Lesson 2: Introduction to Scikit-Learn and Model Evaluation

- Introduction
- Model Performance Metrics for Binary Classification
- Activity 2: Performing Logistic Regression with a New Feature and Creating a Precision-Recall Curve

Lesson 3: Details of Logistic Regression and Feature Exploration

- Introduction
- Examining the Relationships between Features and the Response
- Univariate Feature Selection: What It Does and Doesn't Do
- Building Cloud-Native Applications
- Activity 3: Fitting a Logistic Regression Model and Directly Using the Coefficients

Lesson 4: The Bias-Variance Trade-off

- Introduction
- Estimating the Coefficients and Intercepts of Logistic Regression
- Cross Validation: Choosing the Regularization Parameter and Other Hyperparameters
- Activity 4: Cross-Validation and Feature Engineering with the Case Study Data

Lesson 5: Decision Trees and Random Forests

- Introduction
- Decision trees
- Random Forests: Ensembles of Decision Trees
- Activity 5: Cross-Validation Grid Search with Random Forest

Lesson 6: Imputation of Missing Data, Financial Analysis, and Delivery to Client

- Introduction
- Review of Modeling Results
- Dealing with Missing Data: Imputation Strategies
- Activity 6: Deriving Financial Insights
- Final Thoughts on Delivering the Predictive Model to the Client