EDVANCER
EDUVENTURES

Certified Data Scientist with Python course curriculum

Topic	What does it mean?
Introduction to Data Science • What is data science & why is it so important? • Applications of data science across industries • Data science methodology • Data Scientist Toolbox • Tool of choice- Python: what & why? • Course Components	In this section we shall provide you an overview into the world of data science & machine learning. You will learn about the various applications of data science, how companies from all sort of domains are solving their day to day to long term business problems. We'll learn about required skill sets of a data scientist which make them capable of filling up this vital role. Once the stage is set and we understand where we are heading we discuss why Python is the tool of choice in data science.
Python Training	
Introduction to Python • Installation of Python framework and packages: Anaconda and pip • Writing/Running python programs using Spyder, Command Prompt • Working with Jupyter Notebooks • Creating Python variables: Numeric, string and logical operations • Basic Data containers: Lists, Dictionaries, Tuples & sets • Practice assignment	Python is one of the most popular & powerful languages for data science used by most top companies like Facebook, Amazon, Google, Yahoo etc. It is free and open source. This module is all about learning how to start working with Python. We shall teach you how to use the Python language to work with data.
Iterative Operations & Functions in Python • Writing for loops in Python • List & Dictionary Comprehension • While loops and conditional blocks • List/Dictionary comprehensions with loops • Writing your own functions in Python • Writing your own classes and functions as class objects • Practice assignment	This is where we move beyond simple data containers and learn about amazing possibilities and functionalities hidden in various associated operators. We get introduced to wonderful world of loops, list and dictionary comprehensions. In addition to already existing functions and classes we learn to write our own custom functions and classes. This module sets the stage for handling data and ML algorithm implementation in python.
Data Summary; Numerical and Visual in Python • Need for data summary • Summarising numeric data in pandas • Summarising categorical data • Group wise summary of mixed data • Need for visual summary • Introduction to ggplot & Seaborn • Visual summary of different data combinations • Practice Exercise	Data summary is extremely important to understand what the data is saying and gain insights in just one glance. Visualization of data is a strong point of the Python software using the latest ggplot package using much celebrated grammar of graphics. We also introduce you another powerful package seaborn in additional material section.
Data Handling in Python using NumPy & Pandas • Introduction to NumPy arrays, functions &properties • Introduction to pandas • Dataframe functions and properties • Reading and writing external data • Manipulating Data Columns	Python is a very versatile language and in this module we expand on its capabilities related to data handling. Focusing on packages numpy and pandas we learn how to manipulate data which will be eventually useful in converting raw data suitable for machine learning algorithms.
Data Science & Machine Learning in Python	

Introduction to Machine Learning • Business Problems to Data Problems • Broad Categories of Business Problems • Supervised and Unsupervised Machine Learning Algorithm • Drivers of ML algos • Cost Functions • Brief introduction to Gradient Descent • Importance of Model Validation • Methods of Model Validation • Introduction to Cross Validation and Average Error	In this module we understand how we can transform our business problems to data problems so that we can use machine learning algos to solve them. We will further get into discovering what all categories of business problems and subsequently machine learning altos are there. We'll learn what is the ultimate goal of any machine learning algorithm and go through a brief description of the mother of many modern optimisation methods- Gradient Descent. We'll wrap up this module with discussion on importance and methods of validation of our results.
Generalised Linear Models in Python • Linear Regression • Limitation of simple linear models and need of regularisation • Ridge and Lasso Regression (L1 & L2 Penalties) • Introduction to Classification with Logistic Regression • Methods of threshold determination and performance measures for classification score models • Case Studies	We start with implementing machine learning algorithms in this module. We also get exposed to some important concepts related to regression and classification which we will be using in the later modules as well. Also this is where we get introduced to scikit-learn, the legendary python library famous for its machine learning prowess. **Case Studies:** 1. **Automate lender & borrower matching through prediction of loan interest rates** - In this case study, we try to automate the process of lender and borrower matching for a fintech company by predicting interest rates offered. 2. **Classify customers based on revenue potential for a wealth management firm**- In this classification case study, we help a financial institution to predict which one of their customers are going to fall in high revenue grid so that they can be given selective discounts for customer acquisition in a highly competitive industry of wealth management.
Tree Models using Python • Introduction to decision trees • Tuning tree size with cross validation • Introduction to bagging algorithm • Random Forests • Grid search and randomized grid search • ExtraTrees (Extremely Randomised Trees) • Partial Dependence Plots • Case Studies • Home exercises	In this module we will learn a very popular class of machine learning models, rule based tree structures also known as Decision Trees. We'll examine their biased nature and learn how to use bagging methodologies to arrive at a new technique known as Random Forest to analyse data. We'll further extend the idea of randomness to decrease bias in ExtraTrees algorithm. In addition, we learn about powerful tools used with all kind of machine learning algorithms, gridSearchCV and RandomizedSearchCV. **Case Studies:** In the class we continue with the case studies taken in previous module of simple linear models and see how the tree based models compare in terms of performance in comparison to the linear models. In **take home exercises** we have two case studies: 1. **Capture risks associated with micro loans:** In the 1st exercise you will work on micro loans. Its inherently risky to hand out micro loans because of

	lack of checks in the natural process of micro loans. and in this case study we try to capture risk associated with these micro loans. 2. **How do the tech specifications of a vehicle impact its emissions**? In the 2nd case study we find out effect of technical design specification of a vehicle on average emission and thus its environmental impact.
Boosting Algorithms using Python • Concept of weak learners • Introduction to boosting algorithms • Adaptive Boosting • Extreme Gradient Boosting (XGBoost) • Case study • Home exercise	Want to win a data science contest on Kaggle or data hackathons or be known as a top data scientist? Then learning boosting algorithms is a must as they provide a very powerful way of analysing data and solving hard to crack problems. **Case Studies:** **1. Save lives by predicting health issues in diabetics**: A health care system in a state is struggling with poor detection of severity of health issues in diabetic people. This results in need for re-hospitalisation and many unfortunately not in time. Find out if boosting algos can save lives! **2. Predicting annual income based on census data:** In the take home exercise, find out whether someone is going to have annual income higher than a certain amount just by simple census data and thus identifying potential fraud cases when it comes to filing their taxes.
Support Vector Machines (SVM) and KNN in Python • Introduction to idea of observation based learning • Distances and Similarities • K Nearest Neighbours (KNN) for classification • Introduction to SVM for classification • Regression with KNN and SVM • Case study • Home exercises	We step in a powerful world of “observation based algorithms” which can capture patterns in the data which otherwise go undetected. We start this discussion with KNN which is fairly simple. After that we move to SVM which is very powerful at capturing non-linear patterns in the data. **Case Study:** Since KNN and SVM can take a lot of processing time, we have kept the class discussion case study simple. In this case we see how we can use student’s height and shoe size to determine which grade do they belong in. Same implementation steps can be used to work on any complex business problem as well.
Unsupervised learning in Python • Need for dimensionality reduction • Introduction to Principal Component Analysis (PCA) • Difference between PCAs and Latent Factors • Introduction to Factor Analysis • Patterns in the data in absence of a target • Segmentation with Hierarchical Clustering and K-means • Measure of goodness of clusters • Limitations of K-means • Introduction to density based clustering (DBSCAN)	Many machine learning algos become difficult to work with when dealing with many variables in the data. In comes to rescue PCA which solves problems arising from data which has highly correlated variables. The same idea can be extended to find out hidden factors in our data with Factor Analysis which is used extensively in surveys and marketing analytics. We also learn about two very important segmentation algos; K-means and DBSCAN and understand their differences and strengths. **Case Studies:**

	1. **Understanding impact of cash assistance programs in New York:** To understand PCA, we take up data of cash assistance programs in New York. This has more than 60 variables. We'll see how can we reduce the size of the data. 2. **Car Survey Data:** We take up car survey data which contains technical & price detail of vehicles through 11 numeric variables. We'll see if these 11 variables represent any hidden factors representing different properties of a vehicle. 3. **Pricing wines based on chemical properties:** For K-Means we take data containing chemical properties of 4000+ white wines and examine whether we can find segments of wines based on their chemical compositions. 4. **Customer spend data at a retail chain**: For DBSCAN we see how DBSCAN can be used for anomaly detection using expense data of customers from a retail chain.
Text Mining in Python • Quick Recap of string data functions • Gathering text data using web scraping with urllib • Processing raw web data with BeautifulSoup • Interacting with Google search using urllib with custom user agent • Collecting twitter data with Twitter API • Introduction to Naive Bayes • Feature Engineering for text Data • Feature creation with TFIDF for text data • Case Studies	Unstructured text data accounts for more and more interaction records as most of our daily life moves online. In this module we start our discussion by looking at ways to collect all that data. In addition to scraping simple web data; we'll also learn to use data APIs with example of Twitter API, right from the point of creating a developer account on twitter. Further we discuss one of the very powerful algorithm when it comes to text data; Naive Bayes. Then we see how we can mine the text data. **Case Studies:** 1. Live demonstrations of web scraping and data cleaning 2. Using Facebook & Twitter APIs to extract data 3. Making a portfolio tracking tool using Yahoo finance with Python. 4. Tagging an SMS as SPAM or NON-SPAM based on its content algorithmically with Naive Bayes.
Version Control using Git and Interactive Data Products • Need and Importance of Version Control • Setting up git and github accounts on local machine • Creating and uploading GitHub Repos • Push and pull requests with GitHub App • Merging and forking projects • Introduction to Bokeh charts and plotting • Examples of static and interactive data products	We finish the course with discussion on two very important aspects of a data scientist's work. First is version control which enables you to work on large projects with multiple team members scattered across the globe. We learn about git and most widely used public platform version control that is GitHub. Second thing is making quick prototype of your solutions as interactive visualisation in the form of standalone or hosted web pages. We introduce you to Bokeh, an evolving library in python which has all the tools that you'll need to make small prototypes of data products which can be scaled later.