

# Machine Learning using Python

**Data Science** is emerging as a hot new profession and academic discipline and machine learning is a key area in data science. Harvard Business Review says Data Scientist is the Sexiest Job of the 21st Century. But demand for data scientists is racing ahead of supply. People with the necessary skills are scarce, primarily because the discipline is so new. This course is designed to give a start and introduction to this new discipline. This course is spread across 2 days and will have a plenty of hands on exercises using real world datasets.

## Pre-requisites:

- ✓ A basic understanding of data and programming
- ✓ Programming knowledge using Python is essential

## Hardware & Software:

- ✓ A desktop or notebook with 64 bit OS (Windows/Mac)
- ✓ 8 GB RAM
- ✓ High speed Internet connection 256 kbps+
- ✓ Latest Anaconda Continuum Platform for Python 3.5

## Topics:

- Accessing, preparing and exploring data with Pandas & Scipy
- Data Exploration and visualization & Basic Statistical Analysis
- Machine Learning Basics – Loss Function, Gradient Descent, Bias Variance Trade Off, Underfit and Overfit of Models
- Building linear and non-linear models for Regression and classification problems
- Regularization and Parameter Tuning
- Applying various algorithms - Linear Regression, Logistic Regression, Decision Trees, KNN
- Ensemble Methods – Bagging and Boosting, Random Forest
- Understanding Model Evaluation Metrics

**Duration: 2 Days**

## Instructor Profile

**Manaranjan Pradhan** has about 16+ years of industry experience working on Cloud computing, Big Data, Data Science & Machine Learning. He has worked with TCS, HP, and iGATE and worked on large scale projects for customers like Motorola, Home Depot, CKWB Bank, P&G in the roles of solution and technical architect. He is a freelance who provides consulting and training on Cloud Computing, Big data & Data Science including Machine Learning. He has been teaching Big Data and Machine Learning for 3 years and has trained more than 500 people from several large MNCs including EMC, CISCO, HP, YODLEE, YAHOO, SAMSUNG, VeriSign, Success Factors & Goldman Sachs etc. He is also a guest lecture on Big Data and Machine Learning at IIM Bangalore.

He is an alumni of **Indian Institute of Management (IIM)**, Bangalore and has completed **certification on Business Analytics and Intelligence program**. He has **data science and scalable machine learning certifications from Coursera and edx.org**.

He had published an Analytics Case in Harvard Business Publishing :

<https://cb.hbsp.harvard.edu/cbmp/product/IMB573-PDF-ENG>

He writes his blog at <http://www.awesomestats.in/>

Connect with him on Linked in <http://in.linkedin.com/pub/manaranjan-pradhan/a/6bb/314>

# Machine Learning using Python

Introduction to Data Science and Setting up data analysis environment	Introduction to Data Science Setting up Python Environment for Data Analysis  Overview of Data Analysis Stack - Numpy, Pandas, Matplotlib, scipy and Scikit-learn
Accessing and preparing data with Pandas	Loading data from Different Sources Data manipulation - Filtering, Grouping, Ordering, Joining Dealing with missing Data
Data Exploration, Visualizations & Statistical Analysis	Histograms, Bar charts Density Plots, Box Plots, Scatter Plots, Heat Maps Understanding Basic Statistics, Distributions, Correlations
Algorithms for Regression and Classification Problems	Understanding loss function and gradient descent approach for loss minimization Linear Regression, Logistic Regression Decision Trees, KNN Bias & Variance Trade-off Regularization & Parameter Tuning
Ensemble Methods	Random Forest Bagging & Boosting
Clustering	K-means clustering Finding optimal number of clusters
Model Evaluation	Creating Training, validation and Test Data Sets Cross validations Understanding Evaluation Metrics: RMSE, R-square, ROC, Confusion Matrix, Precision, Recall, Accuracy etc.

Following are the Case studies that will be explained using the above techniques.

## *Store Sales Prediction:*

Store managers need to predict their daily sales for up to several weeks in advance to ensure they do not end up with empty shelves, which could mean unhappy customers. Similarly, store managers also do not want to end up with lots of leftover inventories, which could mean additional overhead cost to the store. Store sales are influenced by many factors like seasonality, competition and promotions and where they are located.

- Explore the data understand effect of various factors on store sales.
- Build a predictive model to forecast the store sales.
- Evaluate the model accuracy.

## *Customer Churn Prediction:*

Companies invest significant amount of money to acquire new customers in anticipation of future revenues. Losing customers mean loss of initial investment on acquisition and loss of possible future revenue. So, it is important for companies to predict early signs if a customer is about to churn. And then engage or offer incentives to customers to retain them.

- Understand factors influencing churn
- Build a model to predict if a customer is about churn
- Predict the probabilities of a customer churning in future
- Evaluate the model accuracy.

## *Customer Segmentation:*

RFM analysis is a customer segmentation technique that can help retailers maximize the return on their marketing investments. Under RFM analysis, each customer is scored based on three factors, Recency, Frequency, and Monetary value.

- Calculate RFM attributes for each customer
- Create customer segments using clustering techniques
- Find the optimal number of clusters or segments