Specialisation Track: Data Scientist

Point Estimation Ltd.

This course will provide a thorough and in-depth tutorial of Advanced Machine Learning techniques along. Each algorithm will be executed on real life datasets. The labs shall be run on Python. The final assignment will be a Capstone Project where the Data Scientist would be expected to develop a machine learning pipeline covering an end-to-ed delivery of a Machine Learning Model.

✉ sourav.das@pointestimation.com ☎ 07948049226 📍 14b Hogfair Lane, SL1 8BY, London, UK

COURSE SKILLSET

EDA & Statistics

Python for Data Science

Machine Learning - framework

Regression & Classification Problems

Supervised & Unsupervised Learning Techniques

Anomaly Detection

Segmentation, Association Problems

ML Pipeline

Web App

1 Introduction to Machine Learning

We shall discuss the framework (CRISP-DM) of a Machine Learning Project. An overview of various Machine Learning techniques/algorithms will be given. After this session you will have a checklist of steps in an end-to-end machine learning problem and would be aware of Machine Learning problems and the techniques used to solve them.

2 Data Preparation: Exploratory Data Analysis, Statistics & Feature Selection

We shall learn about the following:

- Descriptive Statistics, Robust Statistics
- Outlier Detection & Treatment
- Missing Value & Imputation
- Feature Selection
- Treating Multicollinearity, Orthogonality, PCA
- Programming Lab: Python Data Science libraries like Pandas, Sklearn, Statsmodels, NumPy, Matplotlib etc. will be introduced here.

Lab: Create EDA Module on Python that can be treated as a standard template or checks before model building.

3 K-Means, K-Modes, K-Medoids Segmentation: Cluster Analysis

We shall learn about the following:

- K-Means, K-Modes, K-Medoids Segmentation
- Convergence criteria

Lab: Run two K-Means segmentation on Flower species classification & Crime Statistics

4 K-NN

We shall learn about the following:

- K-NN – Applications on Imputation of Missing Values

Lab: Two case studies: Applications on Colour Classification from RGB & Disease Risk assessment.

4 Regression – I: Simple Linear Regression (OLS)

We shall learn about the following:

- When and why regression is employed
- SLR – using OLS
- Assumptions of OLS
- Transformation
- Model Diagnostics
- Influential & Leverage points
- How to tackle model inadequacies
- Output interpretation
- Cross Validation

Lab: Two exercises: One on TV Advertising Spend & Sales Data, the other one looking at relationship between GPA and SAT Scores.

5 Regression – II: Multiple Linear Regression (OLS)

We shall learn about the following:

- Extension of SLR
- Interaction Effects, Dummy Variable
- Multicollinearity detection (structural & data led) & Treatment, Overfitting problem

Lab: Predicting price of a car based on features.

6 Regression – III: Weighted Least Squares

We shall learn about the following:

- Weighted Least Squares Regression when Homoscedasticity assumption in OLS is violated.

Lab: Predicting price of a car based on features.

7 Regression – IV: Robust & Quantile Regression

We shall learn about the following:

- Robust & Quantile Regression when influential outliers are present

Lab: Predicting price of a car based on features.

8 Regression – V: Ridge, Lasso, Elastic-Net

We shall learn about the following:

- Bias vs Variance Trade-off
- Biased vs Unbiased Estimators
- Applications of Ridge, Lasso & Elastic Net techniques when multicollinearity is unavoidable

Lab: Predicting price of a car based on features.

9 Regression – VI: Polynomial & Non – Parametric Regression (Kernel, Loess)

We shall learn about the following:

- Polynomial Regression
- Idea of Non-parametric regression
- Kernel Regression
- LOESS, LOWESS for smoothing.

Lab: Predicting Price of a car based on features.

10 Regression – VII: Time Series Forecasting (ARMA, ARIMA, STLF)

We shall learn about the following:

- Idea of Trend, Seasonality & Residual Component
- Time Series Decomposition using STL
- ACF, PACF
- Stationarity, Differencing
- ARIMA, ARMA, ARCH, GARCH, STLF, VAR, BATS, TBATS Models
- Exponential Smoothing, ETS, Holt Winter

Lab: Two exercises – Forecasting sales data of sim activations of a telecom network & Forecasting search trends of Netflix

11 Regression – LAB: Marketing Mix Modelling

We shall learn about the following:

- About the use cases of Marketing Mix Model and why it is employed by all large enterprises around the world.
- Decomposition, Optimisation of budget

Lab: End to end MMM project for a telecom giant.

12 Logistic Regression

We shall learn about the following:

- Binary & Multinomial Logistic Regression
- Logit transformation, Sigmoid function
- Balancing and the alternative method of threshold probability adjustment
- Odds ratio
- Probability vs likelihood
- Confusion Matrix
- Precision, Recall, AUC, ROC, F1-Score
- Interpretation of parameter estimates

Lab: Credit Risk Model for a bank.

13 Decision Trees: CHAID, C5.0, Random Forest, CART

We shall learn about the following:

- CHAID, C5.0, Random Forest, CART
- Splitting Criterion, Chi-sq., Gini, Entropy, Information Gain
- Root, Branch Leaf Nodes
- Bagging, Boosting
- Exhaustive CHAID
- Pruning
- Feature Importance, SHAP Value – Marginal Contribution
- OOB Score

Lab: Credit Risk Model for a bank using CHAID, C5.0 & Random Forest

14 Gradient Descent & Boosting Techniques

We shall learn about the following:

- Gradient Descent
- XG Boosting
- Adaptive Boosting

Lab: TBC.

15 Support Vector Machine

We shall learn about the following:

- Margins – Hard, Soft
- Kernel Functions: Polynomial, Gaussian, Radial Basis Function, Hyperbolic Tangent, Sigmoid, Linear Splines, Bessel Function.
- Logistic Regression vs SVM

Lab: TBC.

16 Anomaly Detection

We shall learn about the following:

- Outlier vs Novelty detection
- Robust Covariance
- One-class SVM
- Isolation Forest
- Local Outlier Factor

Lab: Application using a Bank transaction dataset

17 Recommender System

We shall learn about the following:

- Explicit and Implicit Feedback
- Collaborative Filtering
- Content Based Filtering
- Distance Measures
- Alternating Least Squares, SVC.
- Matrix Factorisation

Lab: Create a Movies recommender system & a Google merchandise recommender system.

18 Capstone Project: ML Pipeline or Web App on Streamlit

We shall develop an end to end ML pipeline to develop a Credit Risk model with all the learnings from the entire course.

Task 1:

- Starting with Clearly stating the Objective of the Modelling task
- Exploratory Data Analysis and treatment wherever necessary
- Anomaly detection and treatment
- Choosing the right model
- Hyperparameter tuning, Grid Search
- Model Diagnostics
- Model Deployment
- Maintenance plan

Task 2:

- Take any modelling algorithm and Create a Web App assuming you are building this for your team of non-data scientists.
- Build the app on Streamlit and deploy using Heroku.

19 Award:

- Comprehensive Report and reference will be provided.
- We shall help you prepare for your interviews and provide guidance on the interview tasks.
- We shall help you prepare a CV catered for the industry.
- You can get work-experience working on actual business data until you land a paid role.
- Work experience letter will also be provided.