DATA FOLKZ
CATAPULT DATA LEADERS

Professional Certificate in **Data Science**

Programme Cirruculum

6 Terms 3 Projects 1 Elective

Professional Certificate in **Data Science**

DATA FOLKZ
CATAPULT DATA LEADERS

Preface

Term 1
Foundation of Statistics

Term 2
Introduction to Python

Term 3
Data Visualization & EDA

Capstone Project

Term 4
Supervised Learning

Capstone Project

Term 5
Unsupervised Learning

Term 6
Natural Programming Language (BASIC)

Capstone Project

Advanced TABLEAU

Professional Certificate in **Data Science**

DATA FOLKZ
CATAPULT DATA LEADERS

Programme Curriculum

6 Terms
3 Projects
1 Elective

Term 1
Foundation of Statistics

Term Duration : 1 Week

Software Skill : N/A

Assgnments: 2

Module 1
Statistics

Topic 1
What is Data Science?

What is Data Science?

Life cycle of data science

Skills required for data science

Applications of data science in different industries

Topic 2
What is Data Science?

Statistics in Data science

What is Statistics?

How is Statistics used in Data Science?

Population and Sample

Parameters and Statistics

Module 2
Statistics for Data Science

Topic 3
What is Data Science?

Data types

Variable and it's types

Sampling Techniques:

Convenience Sampling

Simple Random Sampling

Stratified Sampling

Systematic Sampling

Cluster Sampling

Term 1

Foundation of Statistics

DATA FOLKZ
CATAPULT DATA LEADERS

Module 2
Statistics for Data Science

Topic 4
Descriptive Statistics

What is Univariate and Bi Variate Analysis?

Measures of Central Tendencies

Measures of Dispersion

-Normal Distribution

-Standard Normal Distribution

Skewness and Kurtosis

Box Plots and Outliers detection

Covariance and Correlation

Case Study

DATA FOLKZ
CATAPULT DATA LEADERS

Term 2
Introduction to Python

Term Duration : 2 Weeks

Software Skill : Python

Assgnments: 4

Module 1
Core Python

Topic 1
Python Introduction

What is Python?

Why Data Science requires Python?

Installation of Anaconda

Understanding Jupyter Notebook

Basic commands in Jupyter Notebook

Understanding Python Syntax

Topic 2
Data Types & Data Structures

Variables

Strings

Lists

Sets

Tuples

Dictionaries

Topic 3
Control Flow & Conditional Statements

Conditional Operators, Arithmetic Operators & Logical Operators

If, Else if and Else Statements

While Loops

For Loops

Nested Loops

List and Dictionary Comprehensions

Term 2

Introduction to Python

DATA FOLKZ
CATAPULT DATA LEADERS

Topic 4

Functions

Code Optimization

Scope

Lambda Functions

Map, Filter and Reduce

Modules and Packages

Module 2

Advanced Python

Topic 5

File Handling

Create, Read, Write files

Operations in File Handling

Errors and Exception Handling

Topic 6

Miscellaneous Python

Date and Time

OOPS Concepts

Topic 7

Regular Expressions

Structured Data and Unstructured Data

Literals and Meta Characters

How to Regular Expressions using Pandas?

Inbuilt Methods

Pattern Matching

Case Study

DATA FOLKZ
CATAPULT DATA LEADERS

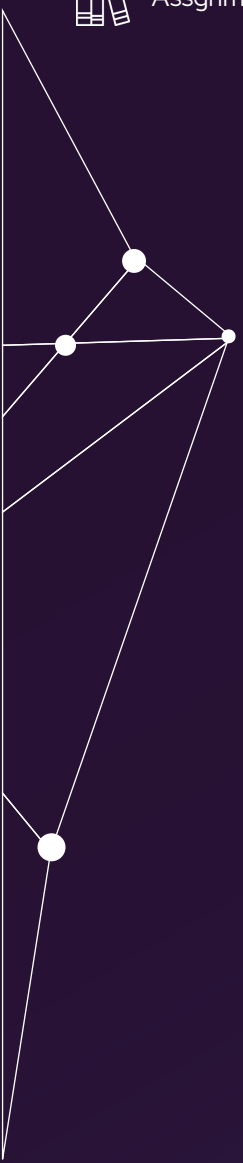
Term 3

Data Visualization & EDA

Term Duration : 2 Weeks

Software Skill : Python

Assgnments: 4



Industry Project

Module 1
Number Analytics

Topic 1
Numpy

Arrays
Basic Operations in Numpy
Indexing
Array Processing

Case Study

Module 2
Working with Data Frames

Topic 1
Pandas

Series
DataFrames
Indexing and slicing
Groupby
Concatenating
Merging Joining
Missing Values
Operations
Data Input and Output
Pivot
Cross tab

Case Study

Capstone Project

DATA FOLKZ
CATAPULT DATA LEADERS

Term 4

Supervised Learning

Term Duration : 2 Weeks

Software Skill : Python

Assgnments: 4

Module 1
Regression

Topic 1
Introduction to Supervised Learning

What Is Machine Learning?

Why Estimate f?

How Do We Estimate f?

The Trade-Off Between Prediction Accuracy & Model Interpretability

Supervised Versus Unsupervised Learning

Regression Versus Classification Problems Assessing Model Accuracy

Topic 2
Linear Regression

Simple Linear Regression:

Multiple Linear Regression:

· OLS Assumptions

· Residual Analysis

Non-linear Transformations of the Predictors

Polynomial Regression

Topic 3
Regularization Techniques

Lasso Regularization

Ridge Regularization

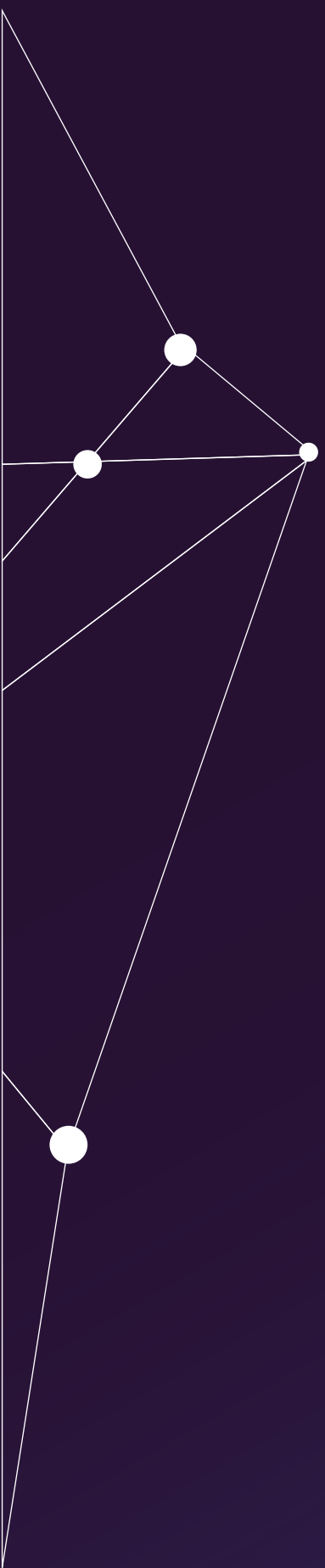
Elastic Net Regularization

Case Study

Term 4

Supervised Learning

DATA FOLKZ
CATAPULT DATA LEADERS



Topic 4

Classification Overview

An Overview of Classification

Why Not Linear Regression?

Topic 5

Logistic Regression

Logistic Regression:

· The Logistic Model

· Estimating the Regression Coefficients and Making Predictions

· Multiple Logistic Regression

· Logit and Sigmoid functions

· Setting the threshold and understanding decision boundary

Topic 6

Evaluation Techniques

Evaluation Metrics for Classification Models:

· Confusion Matrix

· Accuracy and Error rate

· TPR and FPR

· Precision and Recall

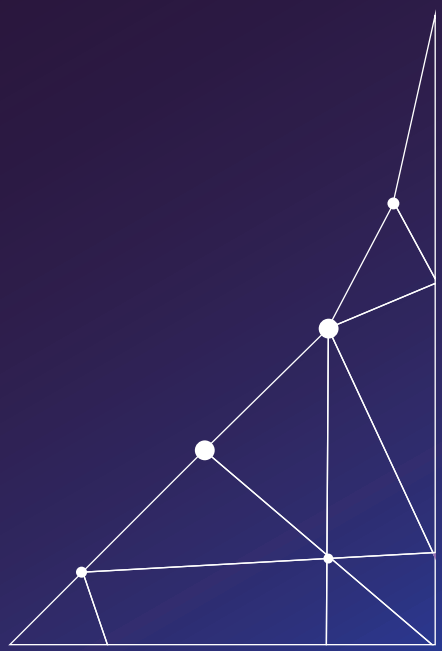
· F1 Score

· AUC – ROC

· Kappa Score

Concordant - Discordant Ratio

Case Study



Term 4

Supervised Learning

DATA FOLKZ
CATAPULT DATA LEADERS

Module 2
Tree Based Learning

Topic 7
Decision Tree

Decision Trees (Rule Based Learning):

· Basic Terminology in Decision Tree
· Root Node and Terminal Node
· Regression Trees
· Classification Trees
· ID3 and C4.5 Decision Trees
· Trees Versus Linear Models
· Advantages and Disadvantages of Trees
· Gini Index, Information Gain/Entropy and Reduction in Variance
· Overfitting and Pruning
· Stopping Criteria
· Accuracy Estimation using Decision Trees

Case Study

Topic 8
Resampling Methods

Resampling Methods:

· Cross-Validation
· The Validation Set Approach Leave-One-Out Cross-Validation
· k-Fold Cross-Validation
· Bias-Variance Trade-Off for k-Fold Cross-Validation

Topic 10
Ensemble Learning

Ensemble Methods in Tree Based Models:

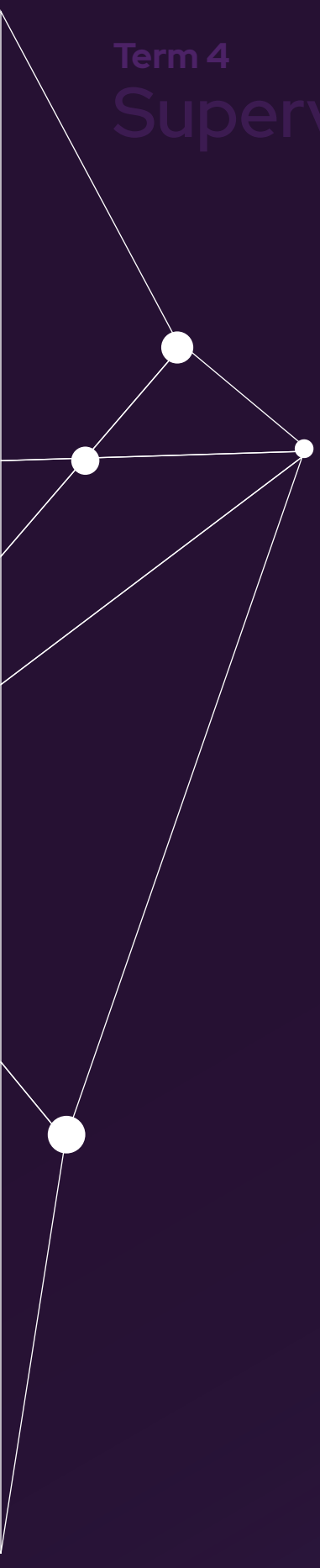
· What is Ensemble Learning?
· What is Bagging and how does it work?
· What is Random Forest and how does it work?
· The Bootstrap
· Variable selection using RandomForest
· What is Boosting and how does it work?
· Ada Boosting
· Gradient Boosting

Case Study

Term 4

Supervised Learning

DATA FOLKZ
CATAPULT DATA LEADERS



Module 3
Distance Based Learning

Topic 11
Support Vector Machines

Support Vector Machines:

· Hyperplane
· The Maximal Margin Classifier
· Support Vector Classifiers
· Support Vector Machines
· Hard and Soft Margin Classification
· Classification with Non-linear Decision Boundaries
· Kernel Trick
· Linear, Polynomial and Radial
· Tuning Hyperparameters for SVM
· Gamma, Cost and Epsilon
· SVMs with More than Two Classes

Case Study

Topic 12
K Nearest Neighbors

K Nearest Neighbors:

· K-Nearest Neighbor Algorithm
· Eager Vs Lazy learners
· How does the KNN algorithm work?
· How do you decide the number of neighbors in KNN?
· Curse of Dimensionality
· Pros and Cons of KNN
· How to improve KNN performance?

Case Study

Industry Project

Capstone Project

Professional Certificate in **Data Science**

DATA FOLKZ
CATAPULT DATA LEADERS

Term 5

Unsupervised Learning

Term Duration : 2 Weeks

Software Skill : Python

Assgnments: 4

Module 1

Clustering & Dimensionality Reduction

Topic 1

Principal Component Analysis

Principal Components Analysis:

· Introduction to Dimensionality Reduction and it's necessity
· What Are Principal Components?
· Demonstration of 2D PCA and 3D PCA
· Eigen Values, Eigen Vectors and Orthogonality
· Transforming Eigen values into a new data set
· Proportion of variance explained in PCA

Case Study

Topic 2

Clustering

Clustering Methods:

· K-Means Clustering
· Centroids and Medoids
· Deciding optimal value of 'k' using Elbow Method
· Linkage Methods
· Hierarchical Clustering
· Divisive and Agglomerative Clustering
· Dendrograms and their interpretation
· Applications of Clustering
· Practical Issues in Clustering
· Improving Supervised Learning algorithms with clustering

Case Study

Term 5

Unsupervised Learning

DATA FOLKZ
CATAPULT DATA LEADERS

Module 2
Association Mining

Topic 3
Association Rules

Association Rules Mining:

· Association Rules
· Market Basket Analysis
· Apriori/Support/Confidence/Lift

Case Study

Topic 4
Naive Bayes Algorithm

ü Naive Bayes:

· Principle of Naive Bayes Classifier
· Bayes Theorem
· Terminology in Naive Bayes
§ Posterior probability
§ Prior probability of class
§ Likelihood
· Types of Naive Bayes Classifier
· Multinomial Naive Bayes
· Bernoulli Naive Bayes
· Gaussian Naive Bayes

Case Study

Term 6
Natural Language Processing (BASIC)

DATA FOLKZ
CATAPULT DATA LEADERS

Term 6
Natural Language Processing (BASIC)

Term Duration : 2 Weeks

Software Skill : Python

Assgnments: 4

Module 1
Time Series Analysis

Topic 1
Time Series (Forecasting)

What is Times Series Data?
Stationarity in Time Series Data and
Augmented Dickey Fuller Test
The Box-Jenkins Approach
The AR Process
The MA Process What is ARIMA?
SARIMA
ACF, PACF and IACF plots
Decomposition of Times Series Trend, Seasonality and Cyclic
Exponential Smoothing
EWMA

Module 2
Natural Language Processing (I)

Topic 2
Intro to NLP

What is NLP?
- Why NLP?
- Applications of NLP
- Unstructured data
- Life cycle of NLP
- Tools for NLP
- Libraries for NLP

o NLTK
o Spacy
o TextBlob

Term 6

Natural Language Processing (BASIC)

DATA FOLKZ
CATAPULT DATA LEADERS

Topic 3

Extracting the Data

Potential data sources

- Reading a pdf file
- Reading a HTML file
- Reading a JSON file
- Data extraction through API and Intro to Webscraping
- Regular expressions
- Handling string

Module 2

Nuts & Bolts of NLP

Topic 4

Text Preprocessing

Text normalizing

- Spelling correction
- Stop words removal
- Stemming
- Lemmatization
- Tokenization
- Text standardization and exploratory data analysis

Topic 5

Text Indexing

Inverted Indexes
Boolean query processing
Handling phrase queries, proximity queries
Latent Sematic Analysis

Topic 6

Feature Engineering

One hot encoding

- N gram
- Feature hashing
- Count vectorizer
- TFIDF
- Co occurance matrix

Word embeddings - word2vec, fasttext etc

Term 6

Natural Language Processing (BASIC)

DATA FOLKZ
CATAPULT DATA LEADERS

Case Study

Text Mining
Sentiment Analysis
Spam Detection
Dialogue Prediction

Industry Project

Capstone Project

Elective

Advanced TABLEAU