Curriculum:

Personal Mentorship Program in Full Stack Decision Science Analytics

Duration:

6.5 months | 75+ sessions | 300+ Hours | 30+ Projects

Tools Covered:

Python | SAS | R | Flask | Streamlit | AWS | GCP | |Azure| SQL | PySpark| Hadoop | Power BI | Tableau

Core Modules:

Statistics | Linear Algebra | Linear Programming | Dashboarding and Reporting | Prescriptive Analytics | Business Intelligence | Regression | Classification | Dimension Reduction | Clustering | Natural Language Processing | Time Series| Recommender Systems | Artificial Neural Network | Computer Vision | Sequential Modelling | Cloud Deployment | Transfer Learning | Data Engineering

Fees:

20,000/- per month (20,000/- * 6)

Projects:

Healthcare | Insurance | Banking | Credit Card | E-Commerce | Supply Chain | Manufacturing | Human Resource Analytics | Marketing Analytics | Financial Analytics | Hospitality | Consumer Research | Medical Images | Social Media | Stock Market | Macroeconomics |

Why Join this Program?

For the last few years if there is one industry which has witnessed an upsurge in jobs and scopes then it is the decision science and the reason being that the organizations', these days have access to loads of data while there is a lack of skilled professionals who can derive strategic insights from these data leveraging the power of decision science.

There is also no dearth of both paid and unpaid courses teaching concepts of data science and machine learning. While each course being unique, however there are some common pitfalls that can be easily observed among these courses and they are:

- **Most of the courses share recorded sessions with no live sessions while a few of them offer minimal live sessions which result in zero faculty interaction and the learning remains incomplete without the faculty interaction**
- **Almost none of the courses address all the cornerstones of decision science viz quantitative techniques, machine learning, deep learning, data engineering, cloud deployment and story telling**
- **Almost all the courses provide relatively smaller datasets (Less than 50k rows) which is nowhere near to the volumes of data a decision scientist has to handle in project life cycle**

This unique Personal Mentorship Program is designed by me (an alumni of Calcutta University, Indian Statistical Institute Bangalore and University of Texas, Austin), after spending more than a decade in this domain. This program focusses on following aspects:

- ✓ **Faculty Interaction. All the sessions are live classes and the recordings shall be made available with the learners for lifetime so that the learning doesn't stop at the end of classroom sessions**
- ✓ **Pragmatic Learning. This program shall cover all the cornerstones at depth and the learners shall be comfortable in curating end to end decision science solutions including deployment across AWS / GCP and Azure**
- ✓ **Large Data. Almost all the projects that the learners shall carry out will comprise of greater than 100k rows while some of them will cross 1 million and thus augmenting pragmatic learning in curating scalable solutions.**

Module I: Quant for Data Science (5 sessions / 20 Hours)

- Statistical Modelling Vs Predictive Modelling Vs Machine Learning Vs Artificial Intelligence Vs Data Science,
- Probability
- Bayes' Theorem,
- Types of Data
- Sample Vs Population
- Sampling Techniques, Sampling Distributions
- Descriptive Statistics
- Central Limit Theorem,
- Inferential Statistics, Test of Hypothesis, T-Test, Z-test, Chi-Square Test
- ANOVA
- Correlation, Regression, Model Evaluation Techniques,
- Linear Programming, Optimization
- Game Theory
- Linear Algebra
- Markov Chains
- Project Management in Data Science
- Data Science Use Cases Across Domains: Healthcare and Insurance, Finance, Human Resources, Marketing, Supply Chain, Retail

Module II: SAS Programming (5 Sessions / 20 Hours)

- Introduction To SAS, Creating Libraries, Data Step Execution, Concept Of PDV, Variable Attributes, Importing Data & Entering Data Manually

- Conditional Statements (Where, If, Else If, Else), Logical Operators, Do Loop, Nesting Do Loop, Do While, Do Until, Drop and Keep Statements
- Array, Sorting, Removing Duplicates Using NODUP and NODUPKEY, Joining, Merging, Appending
- Proc Report, Proc Tabulate, Proc Means, Proc Summary, Proc Univariate
- Reading Raw Data, Proc Format, Working with Date and Character Variables
- Proc SQL, ETL Using SQL Pass Through, Create, Select, Insert, Removing Duplicates, Summary Statistics, join (Left, Right, Inner, Outer) Implementing Proc SQL And Subqueries
- SAS Macros, Introduction to Macro Parameter, %Let, %Macro, %Mend, Local Macro, Global Macro, Call SYMPUT
- Debugging or Handling Errors, Identify and Resolving Logic Errors, Checking for Errors, Warnings, Un-Initialization, Missing Values, Recognizing and Correcting Errors

Module III: R Programming (3 Sessions / 12 Hours)

- The Workspace, Input/ Output, Useful Packages In R
- Graphic User Interfaces (R Studio), Customizing Startup, Batch Processing, Reusing Results
- Data Structure & Data Types (Vectors, Matrices, Factors, Data Frames, And Lists)
- Importing Data (Importing Data from Csv, Txt, Excel And Other Files)

- Keyboard Input (Creating Input by Entering Data)
- Database Input (Connecting to Database and Use the Data)
- Exporting Data (Exporting Files into Different Formats)
- Viewing Data (Viewing Partial Data and Full Data)
- Variable & Value Labels – Date Values
- Missing Data
- Creating New Variables (Calculations & Binning)

- Operators (Using Multiple Operators)
- Control Structures (Conditional Statements, Loops)
- Sorting Data
- Merging and Appending Data
- Sub Setting Data
- Data Type Conversions
- Libraries: Dplyr, Tidyr, Tidyverse etc.

Module IV: Python Programming (5 Sessions / 20 Hours)

- What is Python Language and features?
- Run a sample python script, working with Python IDE's.
- Running basic python commands - Data types, Variables, Keywords, etc**.**
- Indentation (Tabs and Spaces) and Code Comments (Pound # character)
- Variables and Names
- Built-in Data Types in Python
- Basic Operators
- Slicing and The Slice Operator
- Control and Loop Statements

- Python Functions
- Mathematical Computing with Python (NumPy)
- Scientific Computing with Python (SciPy)
- Reading different File formats
- Multithreading
- SQL in Python
- Concatenation and Merging
- Memory management in Python
- Reading large files in Python
- Implementing Python notebook on cloud

Module V: Exploratory Data Analysis and Data Processing (5 Sessions / 20 Hours)

- Data Visualization and Data Pre-Processing for Quantitative and Qualitative Variables Using SAS:
    - Proc GPLOT, GCHART, SGPLOT, G3d
    - Box Plot, Normality Plot, Histogram
- Missing Values, Outliers, Unwanted Negative Values, Incorrect Formatting
- Data Visualization with "GGPLOT2" package in R
- Histograms & Density Plot
- Dot Plots – Bar Plots – Line Charts – Pie Charts – Boxplots – Scatterplots
- Data Visualization with Python:
    - Matplotlib

- Seaborn
- Plotly

- Concepts of Facet Grid and other Grammar of Graphics
- Story Telling with Visualization
- Hypothesis Testing, Null Hypothesis, Alternate Hypothesis, Level of Significance, Type I And Type II Error, One Sample T Test And 2 Samples T Test, One Way ANOVA and Two-Way ANOVA With Interaction for Balanced and Unbalanced Data
- Machine Learning Primer: Types of Data, Frequency Distribution, Data Representation, Probability, Central Limit Theorem, Sampling Types, Descriptive Statistics, Inferential Statistics
- Descriptive Statistics (Central Tendency /Variance)
- Frequency Tables /Summarization
- Hypothesis Testing- T-Tests/Z-Test (1-Sample, Independent Sample, Paired Sample)
- Analysis of Variance (ANOVA)
- Post Hoc Test
- Correlations
- Chi-Square Test
- Test of Normality

Module VI: Supervised and Unsupervised Learning (20 Sessions / 80 Hours)

- Introduction to Supervised and Unsupervised Learning & Applications

- Semi Supervised Learning
- Cost Functions: Difference between Loss Function and Cost Function
    - Mean Squared Error
    - Root Mean Squared Error
    - Cross Entropy Loss
    - Hinge Loss
    - Huber Loss
- Supervised Learning: Family of Regression
    - Linear (Simple & Multiple) Regression
    - Best Subset Regression
    - Polynomial Regression, Regularized Regression (Ridge, LASSO, Elastic Net)
    - Poisson Regression, Binomial Regression
    - Model Diagnostics and Evaluation Criterion
    - Selecting Best Model
- Supervised Learning: Family of Classification Models
    - Logistic Regression
    - K-Nearest Neighbor
    - Support Vector Machines
    - Naïve Bayes' Algorithm
    - Linear Discriminant Analysis
    - Using K-NN and SVM in Regression

- Using SVM for Anomaly Detection
- Model Diagnostics and Evaluation Criteria
- Selecting Best Model

- Supervised Learning: Ensemble Learning
  - Decision Tree
  - Pruning
  - Entropy and Gini
  - Random Forest
  - Stacking Classifier
  - Adaptive Boosting
  - Gradient Boosting
  - XGBoost
  - LightGBM
  - Catboost
  - Implementing Ensemble Learning for Regression and Classification

- Unsupervised Learning Algorithms: Dimension Reduction Techniques
  - Principal Component Analysis
  - Factor Analysis
  - Non-Negative Matrix Factorization
  - T-SNE
  - UMAP
  - Implementing dimension reduction in supervised learning

- Unsupervised Learning Algorithms: Clustering Techniques
  - Hierarchical Clustering
  - K-Means Clustering
  - DBSCAN
  - Implementing Clustering for Segmentation
  - Implementing Clustering in Supervised Learning
  - Implementing Clustering for anomaly detection
  - Evaluating best clustering solution
- Other Algorithms
  - Recommender System
    - ✓ Content Based Filtering
    - ✓ Collaborative Filtering
    - ✓ Association Rule Matrix
  - Survival Analysis
  - Time Series Analysis
    - ✓ Exponential Smoothing
    - ✓ Seasonality
    - ✓ Trend
    - ✓ Random Walk
    - ✓ Unit Root Problem
    - ✓ AR, MA
    - ✓ ARMA, ARIMA
    - ✓ ACF, PACF
    - ✓ ARCH, GARCH
    - ✓ Granger Causality, Co-integration

- ✓ Vector Auto Regressive
- ✓ Model Evaluation Techniques

- Model Improvement Techniques:
  - ➢ Feature Engineering
  - ➢ Handling Data Imbalance
  - ➢ Tuning of Hyper-parameters
  - ➢ Cross Validation
  - ➢ Bootstrapping
  - ➢ Explainable AI

Module VII: Natural Language Processing (5 Sessions / 20 Hours)

- Text processing
- Libraries in Text Processing
- Text Classification Models

Module VIII: Deep Learning (With Cloud Deployment) (10 Sessions / 40 Hours)

- Artificial Neural Network
- Activation Functions, Weight Initialization, Optimizers, Regularization
- Computer Vision and Image Recognition
- Sequential Modelling
- Attention Models, Autoencoders
- Transfer Learning for Computer Vision
- Transfer Learning for Text Classification

Module IX: Scalable Machine Learning (10 Sessions / 40 Hours)

- Data Engineering on Cloud: Handling Big Data on Cloud
- MLOps: End to End Development, AutoML, Deployment, Monitoring, Pipeline (AWS / Azure / GCP)
- PySpark

Module X: Business Intelligence (5 Sessions / 20 Hours)

- SQL Query
- Power BI
- Tableau
- Dashboarding with Python, R-Shiny

Complementary Modules (5 Sessions / 20 Hours):

- Interview Preparation
- Profile Building
- Capstone Project