Invictus Engineers

Data Science using Python

Invictus Engineers

3rd Floor, Surviba Towers, Plot No.247/3RT, Near Umesh Chandra Statue, Sanjeeva Reddy Nagar, Hyderabad, Telangana. 500038 **Email Us :** info@invictusengineers.com **Phone:** +91 99666 33097
www.invictusengineers.com

1. Introduction to Data Science

Introduction

Application of Data Science

- The Machine Learning
  - Supervised, Unsupervised, & Reinforcement Learning

Essential Python for Data Science

- Types of Data
  - Numeric, Text, List, & Dictionary
- The pandas
  - DataFrame and Series
  - CSV Files
  - Excel Spreadsheets
  - JSON

Scikit-Learn package

- Model
- Model Hyperparameters
- The scikit-learn API

2. Regression

Simple Linear Regression

- The Method of Least Squares

Multiple Linear Regression

- Estimating the Regression Coefficients (β_o, β_1, β_2 and β_3)
- Logarithmic Transformations of Variables
- Correlation Matrices

Conducting Regression Analysis Using Python

- The Correlation Coefficient
- The Statsmodels formula API
- Analyzing the Model Summary
- The Model Formula Language
- Intercept Handling

Multiple Regression Analysis

Assumptions of Regression Analysis

Explaining the Results of Regression Analysis

- Regression Analysis Checks and Balances
- The F-test
- The t-test

3. Binary Classification

Introduction
Understanding the Business Context
- Business Discovery
- Testing Business Hypotheses Using Exploratory Data Analysis
- Visualization for Exploratory Data Analysis
- Intuitions from the Exploratory Analysis

Feature Engineering
- Business-Driven Feature Engineering

Data-Driven Feature Engineering
- A Quick Peek at Data Types and a Descriptive Summary

Correlation Matrix and Visualization
- Skewness of Data
- Histograms
- Density Plots
- Other Feature Engineering Methods
- Summarizing Feature Engineering
- Building a Binary Classification Model Using the Logistic Regression Function
- Logistic Regression Demystified
- Metrics for Evaluating Model Performance
- Confusion Matrix
- Accuracy
- Classification Report
- Data Preprocessing

4. Multiclass Classification with RandomForest

Introduction
Training a Random Forest Classifier
Evaluating the Model's Performance
- Number of Trees Estimator

Maximum Depth
Minimum Sample in Leaf
Maximum Features

5. Performing Your First Cluster Analysis

Introduction
Clustering with k-means
Interpreting k-means Results
Choosing the Number of Clusters
Initializing Clusters

Calculating the Distance to the Centroid
Standardizing Data

6. How to Assess Performance

Introduction
Splitting Data
Assessing Model Performance for Regression Models
- Data Structures – Vectors and Matrices
  - Scalars
  - Vectors
  - Matrices
- R^2 Score
- Regression Model
- Mean Absolute Error
  - Other Evaluation Metrics

Assessing Model Performance for Classification Models
- Computing Evaluation Metrics

The Confusion Matrix
- More on the Confusion Matrix
- Precision
- Recall
- F1 Score
- Accuracy
- Logarithmic Loss

Receiver Operating Characteristic Curve
Area Under the ROC Curve
Saving and Loading Models

7. The Generalization of Machine Learning Models

Introduction
Overfitting
- Training on Too Many Features
- Training for Too Long

Underfitting
Data
- The Ratio for Dataset Splits
- Creating Dataset Splits

Random State

Cross-Validation
KFold
cross_val_score
Understanding Estimators That Implement CV
LogisticRegressionCV
Hyperparameter Tuning with GridSearchCV
Decision Trees

Hyperparameter Tuning with RandomizedSearchCV
Model Regularization with Lasso Regression
Ridge Regression

8. Hyperparameter Tuning

Introduction
What Are Hyperparameters?
Difference between Hyperparameters and Statistical Model Parameters
Setting Hyperparameters
A Note on Defaults
Finding the Best Hyperparameterization
Advantages and Disadvantages of a Manual Search
Tuning Using Grid Search
Simple Demonstration of the Grid Search Strategy
GridSearchCV
Tuning using GridSearchCV
Support Vector Machine (SVM) Classifiers
Advantages and Disadvantages of Grid Search
Random Search
Random Variables and Their Distributions
Simple Demonstration of the Random Search Process
Tuning Using RandomizedSearchCV
Advantages and Disadvantages of a Random Search

9. Interpreting a Machine Learning Model

Introduction
Linear Model Coefficients
RandomForest Variable Importance
Variable Importance via Permutation
Partial Dependence Plots
Local Interpretation with LIME

10. Analyzing a Dataset

Introduction
Exploring Your Data
Analyzing Your Dataset
Analyzing the Content of a Categorical Variable Summarizing Numerical Variables
Visualizing Your Data
    Using the Altair API
    Histogram for Numerical Variables
    Bar Chart for Categorical Variables
Boxplots

11. Data Preparation

Introduction
Handling Row Duplication
Converting Data Types
Handling Incorrect Values
Handling Missing Values

12. Feature Engineering

Introduction
    Merging Datasets
        The Left Join
        The Right Join
    Binning Variables
    Manipulating Dates
    Performing Data Aggregation

13. Imbalanced Datasets

Introduction
Understanding the Business Context
    Analysis of the Result
Challenges of Imbalanced Datasets
Strategies for Dealing with Imbalanced Datasets
    Collecting More Data
    Resampling Data
    Analysis
Generating Synthetic Samples
    Implementation of SMOTE and MSMOTE
    Applying Balancing Techniques on a Telecom Dataset

14. Dimensionality Reduction

Introduction
- Business Context

Creating a High-Dimensional Dataset

Strategies for Addressing High-Dimensional Datasets
- Backward Feature Elimination (Recursive Feature Elimination)
- Forward Feature Selection
- Principal Component Analysis (PCA)
- Independent Component Analysis (ICA)
- Factor Analysis

Comparing Different Dimensionality Reduction Techniques

15. Ensemble Learning

Introduction

Ensemble Learning
- Variance
- Bias
- Business Context

Simple Methods for Ensemble Learning
- Averaging
- Weighted Averaging
  - Iteration with Different Weights
  - Max Voting

Advanced Techniques for Ensemble Learning
- Bagging
- Boosting
- Stacking

3rd Floor, Surviba Towers, Plot No.247/3RT,
Near Umesh Chandra Statue, Sanjeeva Reddy Nagar,
Hyderabad, Telangana. 500038

+91 99666 33097

info@invictusengineers.com