BLACKBIRD FOR TRAINING

Data Science for Professionals: Python and Machine Learning



Data Science for Professionals: Python and Machine Learning

Data science is the practice of collecting, cleaning and analysing data with statistics and machine learning to answer questions and make predictions that support decisions. This two-week course is for professionals and team leaders who want hands-on Python skills rather than an overview, and it ends with a capstone project: a tested predictive model, an evaluation report, a dashboard and a deployment plan.

Introduction

Three references organise the course. CRISP-DM, the Cross-Industry Standard Process for Data Mining, sets six phases from business understanding to deployment. The Python data stack, chiefly pandas for data handling and scikit-learn for machine learning, is used for every exercise. Apache Spark introduces distributed processing for data too large for one machine. The level is practitioner. Compared with the one-week certificate, the two weeks add coding practice, supervised and unsupervised learning, model evaluation and a capstone, with datasets from retail, banking and public services worked in notebooks.

Course Objectives

- Apply the CRISP-DM phases to plan a data science project from question to deployment.
- Interpret exploratory statistics and hypothesis tests to decide which variables deserve modelling.
- Distinguish regression from classification and select a suitable algorithm for each business problem.
- Compare precision and recall and justify the trade-off that fits the cost of each type of error.
- Select between bagging and boosting ensembles and tune them with cross-validation.
- Define a deployment and monitoring plan that keeps a model useful after release.

Target Audience

- Analysts moving into data science who must decide which modelling technique suits a business question.
- Team leaders of analytics or reporting units who judge whether a model is reliable enough to use.
- Engineers and IT professionals who decide how data pipelines and model scoring will run in production.
- Researchers and planners in public bodies who choose between statistical tests and predictive models.
- Business specialists in finance, marketing or operations who decide which predictions are worth building.



Course Outline

Day 1: Data Science Roles and the Project Lifecycle

- CRISP-DM vs Ad Hoc Analysis: A Six-Phase Method Against Unplanned Exploration
- Data Scientist vs Data Analyst vs Data Engineer: Responsibilities and Skill Sets
- Data Lifecycle: Collection, Storage, Preparation, Modelling, Deployment and Retirement Stages
- Data Types and Structures: Numerical, Categorical, Text and Time-Series Records
- Python Environment Setup: Jupyter Notebooks, Pandas and a First Dataset Summary

Day 2: Data Acquisition and Management

- Structured vs Unstructured Sources: Database Tables Against Documents, Images and Logs
- ETL Pipelines and Data Warehousing: Extract, Transform and Load Into Analytical Stores
- Data Governance, Privacy and Compliance: Ownership, Access Rights and Retention Rules
- Data Quality Dimensions: Accuracy, Completeness, Timeliness and Consistency Checks Applied
- Connecting to Real Datasets: APIs, SQL Queries and Spreadsheet Imports in Python

Day 3: Exploratory Analysis and Cleaning

- Exploratory Data Analysis: Summary Statistics, Distributions and Initial Pattern Spotting
- Missing Values vs Outliers: Imputation Choices Against Detection and Treatment Rules
- Duplicates and Anomalies: Identifying and Resolving Inconsistent Records With Pandas
- Feature Engineering: Encoding, Scaling and Creating Derived Variables for Models
- Visual Analytics for Exploration: Histograms, Box Plots, Heatmaps and Pair Plots

Day 4: Statistics and Hypothesis Testing

- Descriptive vs Inferential Statistics: Summarising Samples Against Drawing Population Conclusions
- Probability Distributions and Sampling: Normal, Binomial and Sampling Error Explained
- Hypothesis Testing: Null Hypothesis, P-Values, T-Tests and Chi-Square Tests
- Correlation vs Causation: Why Related Variables May Not Drive Each Other
- Statistical Case Study: Testing a Pricing Change Effect on Retail Sales

Day 5: Supervised Learning

- Regression vs Classification: Predicting Continuous Quantities Against Assigning Labelled Categories
- Train, Validation and Test Splits: Holding Out Data for Honest Evaluation
- Linear and Logistic Regression: Baseline Supervised Models Built With Scikit-Learn
- Feature Selection: Filter, Wrapper and Embedded Methods for Choosing Inputs
- Predictive Model Build: A First Supervised Model on a Customer Dataset



Day 6: Ensembles, Neural Networks and Unsupervised Learning

- Bagging vs Boosting: Random Forests Against Gradient Boosting for Tabular Data
- Decision Trees: Split Rules, Depth Limits and Interpretability for Business Users
- Neural Networks and Deep Learning Concepts: Layers, Activation and Training Basics
- Clustering and Dimensionality Reduction: K-Means and PCA for Unlabelled Data
- Customer Behaviour Case Study: Churn Prediction and Segment Discovery Compared

Day 7: Model Evaluation and Tuning

- Precision vs Recall: Error Trade-Offs When False Alarms and Misses Differ
- Bias-Variance Trade-Off: Diagnosing Overfitting and Underfitting From Learning Curves
- Cross-Validation and Hyperparameter Tuning: Grid Search and Randomised Search Explained
- Regression Metrics: MAE, RMSE and R-Squared for Continuous Predictions
- ROC Curve and AUC: Threshold-Independent Comparison of Classification Models

Day 8: Visualisation and Data Storytelling

- Exploratory vs Explanatory Visuals: Charts for Analysts Against Charts for Executives
- Dashboard Design Principles: Layout, Visual Hierarchy and Choice of Chart Type
- Python, Power BI and Tableau: Matplotlib and Seaborn Compared With BI Tools
- Executive Report Writing: Turning Model Results Into Plain Business Narratives
- Interactive Dashboard Build: Filters and Drill-Downs for a Sales Dataset

Day 9: Big Data, Text Analytics and Deployment

- Hadoop vs Spark: Disk-Based Batch Processing Against In-Memory Distributed Computing
- Cloud and Real-Time Analytics: Streaming Data Pipelines and Managed Analytics Services
- NLP and Text Analytics Basics: Tokenisation, Sentiment Scoring and Topic Modelling
- Model Deployment: Batch Scoring, APIs and Monitoring for Model Drift
- Business Integration: KPIs, ROI Measurement, Workflow Automation and Scenario Modelling

Day 10: Delivering the Capstone Data Science Project

- Capstone Problem Framing: Business Question, Success Metric and Data Plan
- Capstone Modelling: Applying Cleaning, Features and Algorithms to One Dataset
- Peer Critique vs Self-Assessment: Reviewing Model Choices and Final Knowledge Check
- Data Transformation Roadmap: Organisational Planning and Continued Learning Paths
- Capstone Model and Deployment Plan: Evaluation Report and Dashboard Presented



Skills You Will Gain

- Pandas data wrangling
- Feature engineering
- Hypothesis testing
- Gradient boosting
- K-means clustering
- Cross-validation
- Model drift monitoring
- Dashboard storytelling

Why Attend This Course

- From reading about machine learning to writing Python notebooks that clean data and train models.
- From judging a model by accuracy alone to choosing metrics that reflect the real cost of errors.
- From one-off analyses to a repeatable CRISP-DM project structure.
- From isolated exercises to a capstone predictive model with an evaluation report, dashboard and deployment plan.

Conclusion

Data science turns raw records into tested predictions and explanations that decision-makers can trust. The course makes clear the line between correlation and causation, between supervised and unsupervised learning, and between a model that scores well in a notebook and one that survives deployment. It suits professionals who already handle data and want two weeks of coding, modelling and a capstone, rather than a shorter introduction to concepts.

Course FAQs

What is data science for professionals?

It is the use of programming, statistics and machine learning by people in business, government or technical roles to clean data, find patterns and build models that predict outcomes relevant to their work.

What is the difference between precision and recall in data science?

Precision is the share of predicted positives that are correct. Recall is the share of actual positives that the model finds. Raising one usually lowers the other, so the choice depends on which error costs more.

Which method does a data science course for professionals follow?

It follows CRISP-DM for project structure and uses Python with pandas and scikit-learn for analysis and modelling, with Apache Spark introduced for large-scale data.



Is a two-week data science course right for a business analyst?

Yes, for a business analyst who already works with spreadsheets or reporting tools and wants to build and evaluate predictive models in Python.

What does a delegate produce in a data science course for professionals?

Each delegate completes a capstone: a trained and evaluated predictive model on one dataset, a written evaluation report, an interactive dashboard and a plan for deployment and monitoring.

