Module 5: Machine Learning – Classification

Data Science with Python – BSc Course

Why Classification?

Visa processes 65,000 transactions per second. Each one needs a fraud/not-fraud decision in under 100ms. Banks evaluate millions of loan applications yearly — approve or deny. Insurance companies classify claims as legitimate or fraudulent.

Classification is the workhorse of financial decision-making. Every time you swipe your credit card, an algorithm makes a split-second call. Every loan application triggers a cascade of binary decisions.

The difference between a good classifier and a bad one can mean billions in prevented fraud or millions in lost customers falsely declined.

Classification turns data into decisions at scale

Why This Matters

- **Fraud detection:** Credit card fraud, insurance fraud, anti-money laundering — all classification problems
- **Credit decisioning:** Approve or deny loan applications based on default probability
- **Customer churn:** Predict which customers will leave, enabling retention efforts
- **Compliance:** Classify transactions as suspicious or clean for regulatory reporting

Every financial institution runs classification models in production

What You'll Learn

By the end of this module, you will be able to:

- Build logistic regression classifiers for binary outcomes
- Construct and interpret decision trees for classification tasks
- Evaluate classifiers using precision, recall, F1-score, and confusion matrices
- Handle imbalanced datasets common in finance (fraud, defaults)
- Optimize classification thresholds for business cost considerations

From probability estimates to production-ready classification systems

Lesson Roadmap: Four Lectures, One Story

Lesson	Narrative Arc	Core Concept
L25	When Numbers Become Decisions	Sigmoid, probability, odds ratios
L26	Twenty Questions	Trees, Gini, Random Forest
L27	Beyond the Grade	Confusion matrix, ROC/AUC, F1
L28	Finding Needles in Haystacks	SMOTE, class weights, PR curves

The arc: Linear boundaries → non-linear trees → measuring quality → handling rare events. Each lecture builds on the previous. By L28, you combine everything into production fraud detection.

Each lesson: 45 minutes lecture + 25-minute hands-on exercise

Key Concepts

- **Logistic Regression** – Sigmoid function maps continuous inputs to probabilities
- **Decision Trees** – Interpretable models that split data based on features
- **Confusion Matrix & Metrics** – Precision, recall, F1 — understanding trade-offs
- **Class Imbalance Handling** – SMOTE, undersampling, and cost-sensitive learning
- **Threshold Optimization** – Adjust decision boundaries for business objectives

These techniques power fraud detection, credit scoring, and compliance systems worldwide

Finance Application

Scenario: Credit Default Predictor
Using the techniques from this module, you will build a model that:

- Predicts whether a borrower will default on a loan
- Handles the 97/3 class imbalance (most loans don't default)
- Optimizes the threshold for business cost (false negatives cost more than false positives)
- Provides interpretable decision rules for regulatory compliance

This is the exact problem that banks solve daily — and misclassifying defaults costs real money.

Build the same models that underpin billions in lending decisions

Who Uses This?

- **Payment Networks** – Visa, Mastercard run real-time fraud scoring on every transaction
- **Banks** – JPMorgan, HSBC use classification for automated credit decisions
- **Insurance** – Allianz, AXA classify claims as legitimate or fraudulent
- **Compliance** – HSBC paid $1.9B fine for AML failures — better classification could have prevented it

Classification models protect trillions in transactions and lending decisions

What's Next: Module 6 – ML: Unsupervised Learning

Classification needs labels. You train on historical data where you know the outcome — fraud or not, default or not. But what if you don't have labels? What if you need to find hidden structure in data without being told what to look for?

Module 6 introduces unsupervised learning: clustering, dimensionality reduction, and pattern discovery. Market regime detection, customer segmentation, anomaly detection — all without labeled data.

Supervised learning answers questions you can ask — unsupervised finds questions you didn't know to ask

Let's Begin!

First up: L25 – "When Numbers Become Decisions"

The bridge from continuous prediction to binary choice.