Training Course on Imbalanced Data Handling in ML

Professional Development Training Course Outline

Category	Data Science	**Duration**	10 Days
Base Fee	$3,000 USD	**Delivery Mode**	Online / On-site Hybrid

Course Introduction

Class imbalance is a pervasive and critical challenge in real-world Machine Learning (ML) applications, where the distribution of instances across different classes is highly skewed. This phenomenon, often seen in domains like fraud detection, medical diagnosis, and anomaly detection, leads to biased models that perform exceptionally well on the majority class but catastrophically on the minority class, which is frequently the class of greatest interest. Effectively addressing imbalanced data is paramount for building robust, fair, and reliable ML models that can deliver accurate and impactful insights in complex scenarios.

Training Course on Imbalanced Data Handling in ML dives deep into the intricacies of imbalanced data, equipping participants with cutting-edge techniques and practical strategies to overcome this common hurdle. Through a blend of theoretical understanding and hands-on implementation, learners will master resampling methods, advanced algorithmic approaches, and appropriate evaluation metrics, ensuring their ML solutions perform optimally even in highly skewed datasets. This course is essential for data scientists, machine learning engineers, and analysts striving to deploy high-performing, production-ready models.

Programme Curriculum

Training Course on Imbalanced Data Handling in ML

Introduction

Class imbalance is a pervasive and critical challenge in real-world Machine Learning (ML) applications, where the distribution of instances across different classes is highly skewed. This phenomenon, often seen in domains like fraud detection, medical diagnosis, and anomaly detection, leads to biased models that perform exceptionally well on the majority class but catastrophically on the minority class, which is frequently the class of greatest interest. Effectively addressing imbalanced data is paramount for building robust, fair, and reliable ML models that can deliver

accurate and impactful insights in complex scenarios.

Training Course on Imbalanced Data Handling in ML dives deep into the intricacies of imbalanced data, equipping participants with cutting-edge techniques and practical strategies to overcome this common hurdle. Through a blend of theoretical understanding and hands-on implementation, learners will master resampling methods, advanced algorithmic approaches, and appropriate evaluation metrics, ensuring their ML solutions perform optimally even in highly skewed datasets. This course is essential for data scientists, machine learning engineers, and analysts striving to deploy high-performing, production-ready models.

Coursed Duration

10 days

Course Objectives

1. Grasp the fundamental concepts and pervasive nature of **class imbalance** in real-world ML datasets.
2. Recognize how **imbalanced data** adversely affects traditional machine learning model performance and evaluation.
3. Select and interpret appropriate **evaluation metrics** (Precision, Recall, F1-Score, AUC-ROC, G-Mean) for imbalanced datasets.
4. Apply various **oversampling techniques**, including Random Oversampling and **SMOTE (Synthetic Minority Over-Sampling Technique)**, to balance datasets.
5. Utilize effective **undersampling methods** like Random Undersampling, Tomek Links, and Edited Nearest Neighbors (ENN).
6. Combine oversampling and undersampling strategies for optimal **data balancing**.
7. Implement **cost-sensitive learning** approaches to penalize minority class misclassifications.
8. Employ **ensemble learning techniques** (Bagging, Boosting, Stacking) tailored for imbalanced data.
9. Understand how to adjust parameters of popular ML algorithms (e.g., Logistic Regression, Decision Trees, SVM, Neural Networks) to handle class imbalance.
10. Learn best practices for data splitting and preprocessing to avoid **data leakage** and ensure model generalization.
11. Effectively analyze **confusion matrices** and other visualizations to gain deeper insights into model performance on imbalanced classes.
12. Apply learned techniques to solve practical **imbalanced classification problems** across diverse industries.
13. Develop a strategic approach to selecting and combining techniques for **robust model performance** on skewed data.

Organizational Benefits

- Achieve significantly higher and more reliable accuracy for minority classes in critical prediction tasks (e.g., fraud, disease detection).
- Mitigate financial and operational risks by enhancing the detection rates of rare, high-impact events.
- Provide data-driven insights that are not biased towards majority outcomes, leading to more informed and equitable decisions.
- Direct resources more effectively by accurately identifying rare but important instances, such as potential customers or critical system failures.
- Develop and deploy more sophisticated and effective ML solutions, gaining an edge in data-driven industries.

- Upskill data science teams with advanced techniques for handling a ubiquitous real-world ML challenge.
- Maximize the return on investment in machine learning initiatives by ensuring models are robust and performant across all data distributions.

Target Audience

1. Data Scientists
2. Machine Learning Engineers.
3. Data Analysts.
4. AI/ML Researchers
5. Software Developers transitioning to ML
6. Anyone with a foundational understanding of Machine Learning
7. Fraud Detection Specialists systems.
8. Medical Imaging & Diagnostics Professionals

Course Outline

Module 1: Introduction to Imbalanced Data

- Define and understand the prevalence of class imbalance in real-world datasets.
- Explore diverse real-world examples of imbalanced data (e.g., credit card fraud, rare disease diagnosis, spam detection).
- Discuss the consequences of ignoring class imbalance: biased models, misleading accuracy, and poor minority class performance.
- Introduce the "accuracy paradox" and why traditional metrics fail for imbalanced data.
- **Case Study:** Analyzing a simulated credit card transaction dataset to observe initial class distribution and naive model performance.

Module 2: Understanding Evaluation Metrics for Imbalanced Data

- Review traditional classification metrics (Accuracy, Precision, Recall, F1-Score).
- Deep dive into Precision and Recall for assessing minority class performance.
- Understand the ROC Curve and AUC (Area Under the Curve) as robust metrics for imbalanced classification.
- Explore the Precision-Recall Curve (PRC) and its significance for highly skewed datasets.
- **Case Study:** Evaluating a baseline fraud detection model using various metrics to highlight the misleading nature of accuracy.

Module 3: Data-Level Techniques: Random Resampling

- Introduction to data-level methods for addressing imbalance.
- Implement Random Oversampling: duplicating minority class instances.
- Implement Random Undersampling: removing majority class instances.
- Discuss the pros and cons of simple random resampling techniques (e.g., overfitting with oversampling, information loss with undersampling).
- **Case Study:** Applying random oversampling and undersampling to a medical diagnosis dataset and comparing model performance.

Module 4: Data-Level Techniques: SMOTE and its Variants

- Understand the limitations of random oversampling and the need for synthetic data generation.
- Learn the Synthetic Minority Over-sampling Technique (SMOTE) algorithm.

- Implement SMOTE using the imblearn library in Python.
- Explore advanced SMOTE variants: Borderline-SMOTE and SVMSMOTE.
- **Case Study:** Enhancing a churn prediction model using SMOTE to improve minority class (churners) recall.

Module 5: Data-Level Techniques: Advanced Undersampling Methods

- Explore advanced undersampling techniques that are more sophisticated than random removal.
- Learn about Tomek Links for identifying and removing ambiguous majority class instances.
- Understand Edited Nearest Neighbors (ENN) for cleaning noisy majority class examples.
- Implement NearMiss algorithms (NearMiss-1, NearMiss-2, NearMiss-3) for targeted undersampling.
- **Case Study:** Applying Tomek Links and ENN to a customer churn dataset to create clearer decision boundaries.

Module 6: Hybrid Data Resampling Strategies

- Combine oversampling and undersampling techniques for synergistic effects.
- Implement SMOTE-ENN and SMOTE-Tomek for robust data balancing.
- Discuss the advantages of hybrid methods in addressing both noise and imbalance.
- Explore other hybrid strategies and their practical applications.
- **Case Study:** Building a robust anomaly detection system by combining SMOTE with an undersampling method.

Module 7: Algorithmic-Level Techniques: Cost-Sensitive Learning

- Understand the concept of misclassification costs and their importance in imbalanced scenarios.
- Learn how to assign different costs to false positives and false negatives.
- Implement cost-sensitive learning in popular ML algorithms (e.g., Logistic Regression, Decision Trees, SVM).
- Discuss the trade-offs involved in adjusting cost matrices.
- **Case Study:** Optimizing a fraud detection model by assigning higher costs to false negatives (missed fraud).

Module 8: Algorithmic-Level Techniques: Ensemble Methods (Bagging)

- Introduction to ensemble learning for imbalanced data.
- Understand Bagging (Bootstrap Aggregating) and its benefits for variance reduction.
- Implement Balanced Bagging and Random Forest Classifiers for imbalanced datasets.
- Discuss how these methods inherently handle class imbalance through diversified sampling.
- **Case Study:** Improving credit default prediction using a Balanced Random Forest model.

Module 9: Algorithmic-Level Techniques: Ensemble Methods (Boosting)

- Explore Boosting algorithms and their iterative approach to improve model performance.
- Understand AdaBoost and its focus on misclassified instances.
- Deep dive into Gradient Boosting Machines (GBM), XGBoost, LightGBM, and CatBoost for imbalanced data.
- Learn how to configure these algorithms with class weights or custom loss functions.
- **Case Study:** Developing a high-performance disease prediction model using XGBoost with weighted classes.

Module 10: Model Selection and Hyperparameter Tuning for Imbalanced Data

- Strategies for selecting the most appropriate ML model for imbalanced classification.
- Techniques for hyperparameter tuning with imbalanced data (e.g., GridSearchCV, RandomizedSearchCV with appropriate scoring metrics).
- Cross-validation strategies: Stratified K-Fold and Group K-Fold for maintaining class distribution during validation.
- Discuss the importance of a robust validation pipeline.
- **Case Study:** Optimizing a rare event prediction model using systematic hyperparameter tuning and stratified cross-validation.

Module 11: Feature Engineering and Selection for Imbalanced Data

- Identify relevant features that highlight differences between minority and majority classes.
- Techniques for handling categorical features and numerical features in imbalanced settings.
- Dimensionality reduction techniques and their impact on minority class representation.
- Strategies for creating new features that amplify the signal of the minority class.
- **Case Study:** Enhancing a cybersecurity intrusion detection system through targeted feature engineering.

Module 12: Advanced Topics: Anomaly Detection and One-Class Classification

- Introduction to anomaly detection as a specialized form of imbalanced learning.
- Explore One-Class SVM (Support Vector Machine) for identifying novel instances.
- Understand Isolation Forest for efficient anomaly detection.
- Discuss the scenarios where anomaly detection approaches are more suitable than traditional classification.
- **Case Study:** Implementing an anomaly detection system for industrial sensor data to flag unusual machinery behavior.

Module 13: Addressing Data Drift and Concept Drift in Imbalanced Streams

- Understand how class distribution can change over time (data drift/concept drift).
- Techniques for monitoring and detecting drift in imbalanced data streams.
- Strategies for adapting models to evolving class distributions.
- Discuss continuous learning and retraining paradigms.
- **Case Study:** Maintaining a high-performing fraud detection model in a dynamic transaction environment.

Module 14: Practical Considerations and Deployment Challenges

- Ethical considerations in imbalanced data: avoiding bias and ensuring fairness.
- Computational considerations: managing large datasets and processing time for resampling.
- Integrating imbalanced data handling techniques into ML pipelines.
- Monitoring deployed models for performance on minority classes.
- **Case Study:** Discussing the challenges and best practices for deploying a clinical diagnosis model with imbalanced data.

Module 15: Capstone Project: Real-World Imbalanced Classification Challenge

- Participants work on a comprehensive, real-world imbalanced dataset.
- Apply multiple techniques learned throughout the course to develop an optimized solution.
- Present their methodology, results, and insights.
- Peer review and expert feedback on project solutions.

- **Case Study:** Participants choose from a selection of complex, real-world imbalanced datasets (e.g., satellite image classification of rare events, predicting rare manufacturing defects, identifying critical infrastructure failures).

Training Methodology

This course employs a highly interactive and hands-on training methodology designed to foster deep understanding and practical application.

- **Instructor-Led Sessions:** Engaging lectures providing theoretical foundations and conceptual clarity.
- **Live Coding Demonstrations:** Step-by-step walkthroughs of Python code using popular ML libraries (scikit-learn, imblearn, pandas, numpy).
- **Hands-On Labs & Exercises:** Practical coding assignments to reinforce learning and build proficiency in implementing various techniques.
- **Case Studies:** Real-world scenarios and datasets are used throughout the modules to illustrate concepts and challenge participants.
- **Interactive Discussions:** Q&A sessions and group discussions to share insights, troubleshoot problems, and explore advanced topics.
- **Capstone Project:** A culminating project where participants apply their acquired knowledge to solve a complex imbalanced data problem end-to-end.
- **Best Practices & Pitfalls:** Emphasis on common challenges, debugging strategies, and ethical considerations in deploying imbalanced learning solutions.
- **Access to Resources:** Provision of course materials, code notebooks, datasets, and recommended readings for continued learning.

Register as a group from 3 participants for a Discount

Send us an email: info@fineskilltrainingcenter.org or call *+254769199797*

Certification

Upon successful completion of this training, participants will be issued with a globally- recognized certificate.

Tailor-Made Course

We also offer tailor-made courses based on your needs.

Key Notes

a. The participant must be conversant with English.

b. Upon completion of training the participant will be issued with an Authorized Training Certificate

c. Course duration is flexible and the contents can be modified to fit any number of days.

d. The course fee includes facilitation training materials, 2 coffee breaks, buffet lunch and A Certificate upon successful completion of Training.

Upcoming Scheduled Sessions

Start Date	End Date	Location	Price
21 Sep 2026	02 Oct 2026	Online	$2,000
21 Sep 2026	02 Oct 2026	Physical	$3,000
21 Sep 2026	02 Oct 2026	Physical	$0
21 Sep 2026	02 Oct 2026	Physical	$0
21 Sep 2026	02 Oct 2026	Physical	$0
21 Sep 2026	02 Oct 2026	Physical	$0
21 Sep 2026	02 Oct 2026	Physical	$0
21 Sep 2026	02 Oct 2026	Physical	$0

Ready to enroll or request a customized program? Book online or contact us:

Register Online Now

Email: info@fineskilltrainingcenter.com | Website: www.fineskilltrainingcenter.com