

Table of Contents

Preface

xi

1

Introduction to Data Imbalance in Machine Learning

1

Technical requirements	2	Challenges and considerations when dealing with imbalanced data	19
Introduction to imbalanced datasets	2	When can we have an imbalance in datasets?	20
Machine learning 101	4	Why can imbalanced data be a challenge?	20
What happens during model training?	7	When to not worry about data imbalance	23
Types of dataset and splits	8	Introduction to the imbalanced-learn library	24
Cross-validation	9	General rules to follow	27
Common evaluation metrics	10	Summary	28
Confusion matrix	10	Questions	29
ROC	15	References	30
Precision-Recall curve	17		
Relation between the ROC curve and PR curve	18		

2

Oversampling Methods

33

Technical requirements	34	SMOTE	39
What is oversampling?	34	How SMOTE works	40
Random oversampling	36	Problems with SMOTE	42
Problems with random oversampling	39	SMOTE variants	43
		Borderline-SMOTE	43

ADASYN	47	Guidance for using various oversampling techniques	55
Working of ADASYN	47	When to avoid oversampling	56
Categorical features and SMOTE variants (SMOTE-NC and SMOTEN)	49	Oversampling in multi-class classification	57
Model performance comparison of various oversampling methods	54	Summary	59
		Exercises	60
		References	60

3

Undersampling Methods 63

Technical requirements	63	Tomek links	77
Introducing undersampling	64	Neighborhood Cleaning Rule	78
When to avoid undersampling the majority class	65	Instance hardness threshold	79
Fixed versus cleaning undersampling	66	Strategies for removing easy observations	80
Undersampling approaches	69	Condensed Nearest Neighbors	80
Removing examples uniformly	70	One-sided selection	82
Random UnderSampling	70	Combining undersampling and oversampling	82
ClusterCentroids	72	Model performance comparison	83
Strategies for removing noisy observations	74	Summary	86
ENN, RENN, and AllKNN	74	Exercises	86
		References	87

4

Ensemble Methods 89

Technical requirements	90	Boosting techniques for imbalanced data	102
Bagging techniques for imbalanced data	91	AdaBoost	103
UnderBagging	96	RUSBoost, SMOTEBoost, and RAMOBoost	104
OverBagging	97	Ensemble of ensembles	106
SMOTEBagging	99	EasyEnsemble	107
Comparative performance of bagging methods	101		

Comparative performance of boosting methods	109	Summary	113
Model performance comparison	110	Questions	113
		References	113
5			
Cost-Sensitive Learning			115
Technical requirements	116	Cost-Sensitive Learning for decision trees	126
The concept of Cost-Sensitive Learning	116	Cost-Sensitive Learning using scikit-learn and XGBoost models	128
Costs and cost functions	116	MetaCost – making any classification model cost-sensitive	133
Types of cost-sensitive learning	117	Threshold adjustment	137
Difference between CSL and resampling	118	Methods for threshold tuning	140
Problems with rebalancing techniques	118	Summary	146
Understanding costs in practice	119	Questions	147
Cost-Sensitive Learning for logistic regression	120	References	147
6			
Data Imbalance in Deep Learning			149
Technical requirements	150	Text analysis using Natural Language Processing	162
A brief introduction to deep learning	150	Data imbalance in deep learning	163
Neural networks	151	The impact of data imbalance on deep learning models	165
Perceptron	152	Overview of deep learning techniques to handle data imbalance	168
Activation functions	153	Multi-label classification	169
Layers	153	Summary	172
Feedforward neural networks	154	Questions	172
Training neural networks	155	References	172
The effect of the learning rate on data imbalance	159		
Image processing using Convolutional Neural Networks	160		

Data-Level Deep Learning Methods 175

Technical requirements	176	Document-level augmentation	202
Preparing the data	176	Character and word-level augmentation	203
Creating the training loop	178	Discussion of other data-level deep learning methods and their key ideas	205
Sampling techniques for deep learning models	180	Two-phase learning	205
Random oversampling	180	Expansive Over-Sampling	205
Dynamic sampling	182	Using generative models for oversampling	206
Data augmentation techniques for vision	185	DeepSMOTE	207
Data-level techniques for text classification	199	Neural style transfer	208
Dataset and baseline model	201	Summary	209
		Questions	209
		References	210

Algorithm-Level Deep Learning Techniques 213

Technical requirements	213	Class-dependent temperature Loss	235
Motivation for algorithm-level techniques	214	Class-wise difficulty-balanced loss	237
Weighting techniques	215	Discussing other algorithm-based techniques	239
Using PyTorch's weight parameter	216	Regularization techniques	239
Handling textual data	220	Siamese networks	239
Deferred re-weighting – a minor variant of the class weighting technique	224	Deeper neural networks	240
Explicit loss function modification	227	Threshold adjustment	240
Focal loss	227	Summary	241
Class-balanced loss	232	Questions	241
		References	242

9**Hybrid Deep Learning Methods**

Technical requirements	246	Online Hard Example Mining	262
Using graph machine learning for imbalanced data	246	Minority class incremental rectification	264
Understanding graphs	246	Utilizing the hard sample mining technique in minority class incremental rectification	265
Graph machine learning	247		
Dealing with imbalanced data	247	Summary	268
Case study – the performance of XGBoost, MLP, and a GCN on an imbalanced dataset	250	Questions	268
Hard example mining	261	References	268

10**Model Calibration**

Technical requirements	271	The calibration of model scores to account for sampling	286
Introduction to model calibration	271	Platt's scaling	288
Why bother with model calibration	273	Isotonic regression	289
Models with and without well-calibrated probabilities	273	Choosing between Platt's scaling and Isotonic regression	291
Calibration curves or reliability plot	274	Temperature scaling	291
Brier score	276	Label smoothing	291
Expected Calibration Error	277		
The influence of data balancing techniques on model calibration	279	The impact of calibration on a model's performance	294
Plotting calibration curves for a model trained on a real-world dataset	282	Summary	295
Model calibration techniques	285	Questions	296
		References	297

Appendix

Machine Learning Pipeline in Production	299		
Machine learning training pipeline	299	Inferencing (online or batch)	301

Assessments 303

Chapter 1 – Introduction to Data Imbalance in Machine Learning	303	Chapter 7 – Data-Level Deep Learning Methods	308
Chapter 2 – Oversampling Methods	306	Chapter 8 – Algorithm-Level Deep Learning Techniques	308
Chapter 3 – Undersampling Methods	307	Chapter 9 – Hybrid Deep Learning Methods	309
Chapter 4 – Ensemble Methods	307	Chapter 10 – Model Calibration	310
Chapter 5 – Cost-Sensitive Learning	307		
Chapter 6 – Data Imbalance in Deep Learning	307		

Index 313**Other Books You May Enjoy** 322

classification	199	Summary	201
Dataset and baseline model	201	Questions	201
ESL	225	References	210
Algorithm-Level Deep Learning Techniques	231		
Technical requirements	231		
Motivation for algorithm-level techniques	231		
Weighting techniques	231		
Using PyTorch's weight parameters	232		
Handling textual data	232		
Targeted re-weighting – a minor variation of the weighting technique	232		
Explicit loss function modification	232		
Model loss	232		
Class balanced loss	232		
Model calibration	233		
ESL	235		
Appendix A: Model Calibration	235		
Appendix B: Model Selection	235		
Appendix C: Model Comparison	235		
Appendix D: Model Selection in Practice	235		
Appendix E: Model Selection in Practice	235		
Appendix F: Model Selection in Practice	235		
Appendix G: Model Selection in Practice	235		
Appendix H: Model Selection in Practice	235		
Appendix I: Model Selection in Practice	235		
Appendix J: Model Selection in Practice	235		
Appendix K: Model Selection in Practice	235		
Appendix L: Model Selection in Practice	235		
Appendix M: Model Selection in Practice	235		
Appendix N: Model Selection in Practice	235		
Appendix O: Model Selection in Practice	235		
Appendix P: Model Selection in Practice	235		
Appendix Q: Model Selection in Practice	235		
Appendix R: Model Selection in Practice	235		
Appendix S: Model Selection in Practice	235		
Appendix T: Model Selection in Practice	235		
Appendix U: Model Selection in Practice	235		
Appendix V: Model Selection in Practice	235		
Appendix W: Model Selection in Practice	235		
Appendix X: Model Selection in Practice	235		
Appendix Y: Model Selection in Practice	235		
Appendix Z: Model Selection in Practice	235		
Appendix AA: Model Selection in Practice	235		
Appendix BB: Model Selection in Practice	235		
Appendix CC: Model Selection in Practice	235		
Appendix DD: Model Selection in Practice	235		
Appendix EE: Model Selection in Practice	235		
Appendix FF: Model Selection in Practice	235		
Appendix GG: Model Selection in Practice	235		
Appendix HH: Model Selection in Practice	235		
Appendix II: Model Selection in Practice	235		
Appendix JJ: Model Selection in Practice	235		
Appendix KK: Model Selection in Practice	235		
Appendix LL: Model Selection in Practice	235		
Appendix MM: Model Selection in Practice	235		
Appendix NN: Model Selection in Practice	235		
Appendix OO: Model Selection in Practice	235		
Appendix PP: Model Selection in Practice	235		
Appendix QQ: Model Selection in Practice	235		
Appendix RR: Model Selection in Practice	235		
Appendix SS: Model Selection in Practice	235		
Appendix TT: Model Selection in Practice	235		
Appendix UU: Model Selection in Practice	235		
Appendix VV: Model Selection in Practice	235		
Appendix WW: Model Selection in Practice	235		
Appendix XX: Model Selection in Practice	235		
Appendix YY: Model Selection in Practice	235		
Appendix ZZ: Model Selection in Practice	235		
Appendix AA: Model Selection in Practice	235		
Appendix BB: Model Selection in Practice	235		
Appendix CC: Model Selection in Practice	235		
Appendix DD: Model Selection in Practice	235		
Appendix EE: Model Selection in Practice	235		
Appendix FF: Model Selection in Practice	235		
Appendix GG: Model Selection in Practice	235		
Appendix HH: Model Selection in Practice	235		
Appendix II: Model Selection in Practice	235		
Appendix JJ: Model Selection in Practice	235		
Appendix KK: Model Selection in Practice	235		
Appendix LL: Model Selection in Practice	235		
Appendix MM: Model Selection in Practice	235		
Appendix NN: Model Selection in Practice	235		
Appendix OO: Model Selection in Practice	235		
Appendix PP: Model Selection in Practice	235		
Appendix QQ: Model Selection in Practice	235		
Appendix RR: Model Selection in Practice	235		
Appendix SS: Model Selection in Practice	235		
Appendix TT: Model Selection in Practice	235		
Appendix UU: Model Selection in Practice	235		
Appendix VV: Model Selection in Practice	235		
Appendix WW: Model Selection in Practice	235		
Appendix XX: Model Selection in Practice	235		
Appendix YY: Model Selection in Practice	235		
Appendix ZZ: Model Selection in Practice	235		
Appendix AA: Model Selection in Practice	235		
Appendix BB: Model Selection in Practice	235		
Appendix CC: Model Selection in Practice	235		
Appendix DD: Model Selection in Practice	235		
Appendix EE: Model Selection in Practice	235		
Appendix FF: Model Selection in Practice	235		
Appendix GG: Model Selection in Practice	235		
Appendix HH: Model Selection in Practice	235		
Appendix II: Model Selection in Practice	235		
Appendix JJ: Model Selection in Practice	235		
Appendix KK: Model Selection in Practice	235		
Appendix LL: Model Selection in Practice	235		
Appendix MM: Model Selection in Practice	235		
Appendix NN: Model Selection in Practice	235		
Appendix OO: Model Selection in Practice	235		
Appendix PP: Model Selection in Practice	235		
Appendix QQ: Model Selection in Practice	235		
Appendix RR: Model Selection in Practice	235		
Appendix SS: Model Selection in Practice	235		
Appendix TT: Model Selection in Practice	235		
Appendix UU: Model Selection in Practice	235		
Appendix VV: Model Selection in Practice	235		
Appendix WW: Model Selection in Practice	235		
Appendix XX: Model Selection in Practice	235		
Appendix YY: Model Selection in Practice	235		
Appendix ZZ: Model Selection in Practice	235		
Appendix AA: Model Selection in Practice	235		
Appendix BB: Model Selection in Practice	235		
Appendix CC: Model Selection in Practice	235		
Appendix DD: Model Selection in Practice	235		
Appendix EE: Model Selection in Practice	235		
Appendix FF: Model Selection in Practice	235		
Appendix GG: Model Selection in Practice	235		
Appendix HH: Model Selection in Practice	235		
Appendix II: Model Selection in Practice	235		
Appendix JJ: Model Selection in Practice	235		
Appendix KK: Model Selection in Practice	235		
Appendix LL: Model Selection in Practice	235		
Appendix MM: Model Selection in Practice	235		
Appendix NN: Model Selection in Practice	235		
Appendix OO: Model Selection in Practice	235		
Appendix PP: Model Selection in Practice	235		
Appendix QQ: Model Selection in Practice	235		
Appendix RR: Model Selection in Practice	235		
Appendix SS: Model Selection in Practice	235		
Appendix TT: Model Selection in Practice	235		
Appendix UU: Model Selection in Practice	235		
Appendix VV: Model Selection in Practice	235		
Appendix WW: Model Selection in Practice	235		
Appendix XX: Model Selection in Practice	235		
Appendix YY: Model Selection in Practice	235		
Appendix ZZ: Model Selection in Practice	235		
Appendix AA:			