Balancing Techniques

| **Technique** | **Status** | **Note** | **Deadline** |
| --- | --- | --- | --- |
| Random Oversampling | CODE READY |  | 18-mar-2024 |
| Random Undersampling | CODE READY |  | 18-mar-2024 |
| SMOTE | CODE READY |  | 18-mar-2024 |
| EasyEnsemble | TO DO |  | 26-mar-2024 |
| BalanceCascade | TO DO |  | 26-mar-2024 |
| ADASyn | TO DO |  | 26-mar-2024 |
| SMOTEBoost | TO DO |  | 26-mar-2024 |
| BoostingSVM | TO DO |  | 26-mar-2024 |
| LGBM Class Weights | CODE READY |  | 26-mar-2024 |
| Isolation Forest | TO DO | Anomaly Detection Algorithm | 26-mar-2024 |
| Extended Isolation Forest | TO DO | Anomaly Detection Algorithm | 1-apr-2024 |
| Local Outlier Factor | TO DO | Anomaly Detection Algorithm | 1-apr-2024 |
| One Class SVM | TO DO | Anomaly Detection Algorithm | 1-apr-2024 |
| Focal Loss | TO DO |  | 8-apr-2024 |
| Rule Base Downsampling | TO DO | TBD | 8-apr-2024 |
| Smart Downsampling | TO DO | TBD | 8-apr-2024 |

Datasets

| Dataset\_Name | Row\_Count | Minority\_Class\_Percent | Target\_Column |
| --- | --- | --- | --- |
| Abalone | 1477 | 4.52 | target |
| Adult | 32561 | 24.08 | target |
| Churn | 7043 | 26.53 | target |
| Contraceptive | 1743 | 22.6 | target |
| Covertype | 581012 | 14.77 | target |
| Fraud | 284807 | 0.172 | Class |
| Fraud2 | 250000 | 0.5 | is\_fraud |
| Letter-a | 20000 | 3.94 | letter\_a |
| Letter-vowel | 20000 | 19.39 | is\_vowel |
| Pendigits | 10992 | 9.59 | is\_9 |
| Sick\_euthyroid | 2194 | 10.05 | target |
| Splice-junction | 3186 | 24.07 | target |
| Wine | 6497 | 3.76 | target |

Papers

H. He and E. A. Garcia, "Learning from Imbalanced Data," in IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263-1284,

G.M. Weiss and F. Provost, "The Effect of Class Distribution on Classifier Learning: An Empirical Study", 2001.

J. Laurikkala, "Improving Identification of Difficult Small Classes by Balancing Class Distribution", Proc. Conf. AI in Medicine in Europe: Artificial Intelligence Medicine, pp. 63-66, 2001.

A. Estabrooks, T. Jo and N. Japkowicz, "A Multiple Resampling Method for Learning from Imbalanced Data Sets", Computational Intelligence, vol. 20, pp. 18-36, 2004.

C. Drummond and R.C. Holte, "C4.5 Class Imbalance and Cost Sensitivity: Why Under Sampling Beats Over-Sampling", Proc. Int’l Conf. Machine Learning Workshop Learning from Imbalanced Data Sets II, 2003.

X.Y. Liu, J. Wu and Z.H. Zhou, "Exploratory Under Sampling for Class Imbalance Learning", Proc. Int’l Conf. Data Mining, pp. 965-969, 2006.

J. Zhang and I. Mani, "KNN Approach to Unbalanced Data Distributions: A Case Study Involving Information Extraction", Proc. Int’l Conf. Machine Learning (ICML ’2003) Workshop Learning from Imbalanced Data Sets, 2003.

G.M. Weiss and F. Provost, "Learning When Training Data Are Costly: The Effect of Class Distribution on Tree Induction", J. Artificial Intelligence Research, vol. 19, pp. 315-354, 2003.