Week 7 (09.12 - 15.12.2019)

Part 1

Over-sampling for class imbalance problems

I have used imbalanced-learn, which is a is a python package offering a number of re-sampling techniques commonly used in datasets showing strong between-class imbalance. Classifier used is LightGBM.

1) SMOTE

Distribution before sampling: Class 0: 6760 (12.2%), Class 1: 937 (87.8%)

Class 0: 6760 (50%), Class 1: 6760 (50%)					
Accuracy	Precision	Recall	F1	ROC-AUC	ROC-AUC (validation)
0.94	0.97	0.92	0.94	0.94	0.97

Class 0: 6760 (66.6%), Class 1: 3380 (33.3%)					
Accuracy	Precision	Recall	F1	ROC-AUC	ROC-AUC (validation)
0.93	0.96	0.83	0.89	0.91	0.95

Class 0: 6760 (76.9%), Class 1: 2028 (23.1%)					
Accuracy	Precision	Recall	F1	ROC-AUC	ROC-AUC (validation)
0.91	0.90	0.69	0.78	0.83	0.91

Class 0: 6760 (83.3%), Class 1: 1352 (16.7%)					
Accuracy	Precision	Recall	F1	ROC-AUC	ROC-AUC (validation)
0.91	0.83	0.53	0.64	0.75	0.85

Under-sampling for class imbalance problems

1) InstanceHardnessThreshold

Distribution before sampling: Class 0: 6760 (12.2%), Class 1: 937 (87.8%)

Class 0: 974 (51%), Class 1: 937 (49%)					
Accuracy	Precision	Recall	F1	ROC-AUC	ROC-AUC (validation)
0.85	0.90	0.78	0.84	0.85	0.90

Class 0: 2061 (68.7%), Class 1: 937 31.3%)					
Accuracy	Precision	Recall	F1	ROC-AUC	ROC-AUC (validation)
0.89	0.89	0.70	0.78	0.83	0.88

Class 0: 4726 (83.5%), Class 1: 937 16.5%)					
Accuracy	Precision	Recall	F1	ROC-AUC	ROC-AUC (validation)
0.90	0.89	0.52	0.65	0.75	0.84

2) ClusterCentroids

Class 0: 937 (50%), Class 1: 937 (40%)					
Accuracy	Precision	Recall	F1	ROC-AUC	ROC-AUC (validation)
0.89	0.93	0.85	0.89	0.89	0.94

Class 0: 1874 (66.7%), Class 1: 937 (33.3%)					
Accuracy	Precision	Recall	F1	ROC-AUC	ROC-AUC (validation)
0.90	0.94	0.76	0.84	0.87	0.92

Class 0: 4726 (83.5%), Class 1: 937 16.5%)					
Accuracy	Precision	Recall	F1	ROC-AUC	ROC-AUC (validation)
0.90	0.89	0.52	0.65	0.75	0.84