CS 4320: Machine Learning

Binary Classification: Customer Transaction Prediction

Use the data set available on the Santander Customer Transaction Prediction contest at Kaggle. Using a binary classification model (such as DecisionTreeClassifier), obtain the best Area Under the Receiver Operating Characteristic Curve possible. The current median score on the public leader board is 0.888. You can make a submission to Kaggle to see where you stand.

This data suffers from the Imbalanced Classes Problem. You’ll want to deal with it.

You may want to explore other classification models.

You may want to explore Ensemble Methods.

Required Steps

• Write a Python program using sklearn to fit the training data to some binary classification model.
• Use the program to explore various models and hyper parameters.
• It is expected that some data pre-processing and the estimator will be combined in a pipeline.
• It is expected that some form of hyper parameter cross validation search will be used.
• After exploring the hyper parameter space and finding a best estimator, the estimator’s quality must measured using your test data.
• After completing the work, make your estimates on Kaggle’s test data, and submit them.
• Write a report. Discuss your processing. If appropriate, include a diagram of the pipeline used, the hyper parameters searched, and the estimator selected. Compare the score reported by the cross validation search, and the score giving by the test data predictions. Also, include the Kaggle score given on the public leader board.
• Submit the report (as PDF), and the source code (as Python source).