Assignment1 Decision Tree Classifier

Due Midnight April 8, 2024 (25 points)

1. Data Preprocessing:

   a. Convert the category CLASS Yes / No into 1 and 0, respectively

   https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html

   b. Remove variables that do not change across the observation

   c. Handle all seven categorical variables

2. Construct tree ensembles: Random Forest (RF) and XGBoost (XGB) to predict **Employee** Attrition class (1= positive, 0=negative)

3. Split data into train set of 80% and test set of 20% (random_state = 1234 for reproducibility)

   https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html

4. Construct RandomForestClassifier() using RandomizedSearchCV() and GridSearchCV() [k=5] to tweak some hyperparameters: max_depth, min_samples_split, n_estimators, and max_features

   For example: 'max_features': np.arange(0.1, 1, 0.1); 'max_samples': [0.3, 0.5, 0.8]

   random_state = 1234

   Also plot graph (as instructed in Lab class) and

   list all features with important scores.

   (https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectFromModel.html)

5. Construct XGBClassifier() using RandomizedSearchCV() and GridSearchCV() [k=5] to tweak some hyperparameters on your own preference. Compare the model performance trained with RandomizedSearchCV() and GridSearchCV() [k=5]

6. Compare and analyze the model performance between RF and XGB, each of which trained with RandomizedSearchCV() and GridSearchCV(). Submit the report.pdf file including the link to your **colab notebook.**