GDAA 2010 – Data Mining Modelling

Assignment #1

Value: 30% of final grade

Due: March 31, 2023

Applied Machine Learning

Introduction

For your first assignment, you will use traditional machine learning techniques to make class label predictions on a dataset of your choice. The purpose of this assignment is for you to gain experience building machine learning workflows in a Python-based environment. You will leverage various ready-to-use algorithms from the scikit-learn package to make class label predictions, and you will compare and evaluate your results using industry standard metrics.

Your work will be executed in a Jupyter Notebook, and be accompanied by detailed comments that explain your dataset characteristics, machine learning workflow, modelling procedures, validations procedures, and results. In the end, you are expected to describe the optimal machine learning algorithm for making class label predictions on your selected dataset, based on the algorithms you opted to use for this assignment.

Tasks

1. Select a dataset with at least several hundred (but preferably several thousand) samples, and a suitable target variable for making class label predictions. To be clear, Assignment #1 is about making predictions for a categorical target variable, not a numeric variable.

2. If your target variable is numeric, it must be binned (or binarized) into two possible class labels.

3. If your target variable is imbalances, either rebalance it manually, or ensure that you use an appropriate validation technique to account for this imbalance and explain your decision-making process in the Notebook.

4. You should also have at least 5 predictors/independent variables. To use categorical data with many levels, it is often necessary to re-encode these variables as a series of binary variables using techniques like one-hot encoding (see posted Jupyter Notebooks on Decision Trees and Random Forest for examples of applying one-hot encoding).

5. Select and execute at least three traditional (i.e., non-Deep Learning/Neural Network) machine learning algorithms to generate class label predictions. Possible algorithms include:
    - Support Vector Machines
    - Decision Trees
    - Random Forests
    - K-Nearest Neighbors

    You are not restricted to the above list, but make sure that whichever algorithms you choose to use are viable for making class label predictions. Read the scikit-learn documentation for the algorithm to verify that your chosen method uses a classifier and not a regressor. Note that some algorithms are capable of both classification and regression.

6. Ensure that you are clearly outlining your training/test dataset splits.

7. Ensure that you are using appropriate validation/cross validation techniques to ward off issues stemming from imbalanced target variables, grouping/biased variables, and overfitting.

8. Describe the performance of your modelling using accuracy scores, area under the curve (AUC) values, and confusion matrices.

9. Summarize which approach performed the best and offer some suggestions as to why you think that particular algorithm did well with your dataset. Otherwise, if all models performed poorly, explain why you think that might be the case.

Submission Instructions

Please submit your Jupyter Notebook and any associated files (e.g., data files, images) required to run the notebook to Brightspace by the specified deadline.