Alternative Project – Supervised Learning

Domain:

Social, Demographic studies

Context:

Extraction was done by Barry Becker from the 1994 Census database. A set of reasonably clean records was extracted and the prediction task is, given the demographic information of individuals, predict whether the individual makes less or more than $50k. It is a binary classification problem.

Data Description:

The Data set contains information about various attributes like education, gender, family status etc, of people from diverse backgrounds.

Attribute Information:

age: continuous.

workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov etc

Fnlwgt: Sample weight.

education: Bachelors, Some-college, 11th, HS-grad, Prof-school etc

education-num: continuous.

marital-status: Married-civ-spouse, Divorced, Never-married, Separated etc

occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial etc

relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.

race: White, Asian-Pac-Islander, Black etc

sex: Female, Male.

capital-gain: continuous.

capital-loss: continuous.

hours-per-week: continuous.

native-country: United-States, Cambodia, England, Puerto-Rico, Canada etc

Learning Outcomes:

- Exploratory Data Analysis
- Preparing the data to train a model
- Training and making predictions using different Classification Techniques
- Evaluating ML models

Objective:

The classification goal is to predict the income level of the customers based on the above attributes

Steps and tasks:

1. Import all the necessary libraries
2. There are two dataframes to be read - adult.data and adult.test. Load both the datasets (5 marks)
3. Combine the two dataframes into one (5 marks)
4. Display the number of missing values(if any) in each of the attributes and treat them accordingly (5 marks)
5. Perform basic EDA (In is an open ended task. Do the best you can) (5 marks)
6. Prepare/preprocess the data to train any given ML model (5 marks)
7. Split the data into train and test set. test set size = 30% of the original combined data
8. Use any 4 classification algorithms learnt during the course, fit the models with the train data and predict on the test set (10 marks)
9. Compute the accuracy, recall and precision for each of the models (5 marks)
10. Try to improve the performance by either dropping a few columns or tuning the models (It is an open ended task. Do your best) (5 marks)
11. Create a dataframe with 4 columns. [‘model name’, ‘accuracy’, ‘recall’ and ‘precision’]. Populate the data-frame accordingly (5 marks)

References:

Source