git_ml-assg1

May 17, 2025

The primary objective of the project is: to predict whether or not a credit card client will default for their payment in the next month. We will be using the better of 2 classifiers namely, Random Forest and KNN (K-Nearest Neighbour) Classifier, and determine the best of a given set of hyperparameters by using grid search.

This project domonstrates some Machine Learning Methodologies: 1. Data Exploration and Preprocessing 2. Define and develop Pipelines 3. Performance evaluation - Compare Different metrics and classifiers.

[0]:
```
## Use this for consistency in graphs through out the notebook
#  import necessary modules and set up the environment
import numpy as np
import pandas as pd

# to make this notebook's output a standard one across runs
np.random.seed(123)  # ensuring reproducible results.

# To plot pretty figures
%matplotlib inline
import matplotlib
import matplotlib.pyplot as plt
plt.rcParams['axes.labelsize']  = 14
plt.rcParams['xtick.labelsize'] = 12
plt.rcParams['ytick.labelsize'] = 12
```

Questions (12 marks total)

Q1. Explore the credit card data set provided. You can also access it from the this link

https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients (https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients)

The data is open for public use and no authorizations are required.

You will build a classification model for this default of credit card clients dataset. The objective is to predict whether or not a credit card client will default for their payment in the next month.

Make sure you perform your analyses and answer the questions in sections below:

1. Data exploration: (3 marks)
    - Explore the data (for example look at the data, plot graphs (histogram, pair plots)

2. Data Preprocessing: (4 marks)
    - Make sure you build a full data pipeline (ie., use the pipeline to apply transformers and estimators- https://scikit-learn.org/stable/modules/generated/sklearn.pipeline.Pipeline.html)
    - Do you require any data pre-processing? Are all the features useful? (Use only raw features from this dataset, in other words, no need to create feature crosses or new features)
    - Set the random seed to 123 (For splitting or any other random algorithm)
    - Split data into training (80%) and testing (20%)
    - Use Cross-validation with 5-folds
    - For other parameters, use default
3. Classification: (5 marks)
    - Study the ROC Curve, decide threshold
    - Use 2 classifiers.
        a. Random Forest
            – tune only: n_estimators: {4, 5, 10, 20, 50}. We will be running random forest model using GridSearchCV, determine the best hyperparameter for the given list of n_estimators {4, 5, 10, 20, 50}. n_estimators refers to the number of trees in the forest. We will use CV = 5 and the scoring to be the roc_auc (area under the curve)
        b. KNN Classfier
            – tune only: n_neighbors: {3, 5, 10, 20}. You may perform similar GridSearchCV as in the previous exercise with a given list of n_neightbors.
    - Which one performs better in the cross validation? Note down your observations and give comments.

You may refer to the documentation for RandomForests and KNN Classifiers, for the different parameters and options available in the scikit-learn library. http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html http://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html

[2]:
```
sudo apt install python3
```

```
  Cell In[2], line 1
    sudo apt install python3
         ^
SyntaxError: invalid syntax
```

Conclusions Explain your results and choices