MBAN 6500X – Assignment 1

Due date: The assignment is due on February 6, 2024 at 7:00 pm.

Task

This is an individual assignment that requires you to participate in a machine learning competition on *Kaggle*. Specifically, you will participate in the competition *Titanic: Machine Learning from Disaster*, where the task it to predict survival based on passenger information.

You have to register a Kaggle account and follow the instructions under *Overview* on the competition site. Beyond the *Overview*, I recommend you to closely study a couple of notebooks under the *Notebooks* tab. For example, "Titanic: 81.1% Leader board Score Guaranteed" and "A Data Science Framework: To Achieve 99% Accuracy" provide good examples of exploratory analysis and feature engineering and are thus worth your time. Use them as tutorials.

You have to fit and **compare three different learning algorithms**, including a linear and a non-linear learner.

Submission

1) Select the best among the models and submit the predictions Kaggle's test set to Kaggle.
2) Submit to Canvas a standard Python or Jupyter notebook file (i.e. `PY` or `IPYNB`) containing the complete code (see **Grading**) that trains all three models and for each model, prints the accuracy, F1-score and AUC.
    - Enter your Kaggle username as a comment on your Canvas submission (not in the submitted file). This is necessary for me to verify your Kaggle submission.

Grading

The submitted Python file should contain the following standard steps of a data science project:

1) Load data.
2) Exploratory data analysis and pre-processing (aka data wrangling). Use the insights you gain from the exploratory analysis to guide the pre-processing.
    - Data cleaning.
    - Identification and treatment of missing values and outliers.
    - Feature engineering.
    - At least three plots including a scatter plot (to inspect relationships between two variables), a histogram (distributions of, for example, ages) and a pie chart (to show, for example, the relationship between a categorical varialbe and survival).

- Print a basic data description (e.g. number of examples, number features, number of examples in each class and such). The data description *should be printed under the hearder* `Data description` (see example below).
- Print descriptive statistics (e.g. means, medians, standard deviation) *under the header* `Descriptive statistics` (see example below).

3) Partition data (not Kaggle's test set) into train, validation and test sets. This test set will be different from Kaggle's test set.
4) Fit models on the training set (this can include a hyper-parameter search) and select the best based on validation set performance.
5) Print the results of all three models on the test set from (4) *under the header* `Results`. The results should include accuracy, F1-score and AUC (see example below).
6) Save the predictions of the best model on Kaggle's test set to `submission.csv`.

Example output

```
Data description
----------------

Number of examples
  train: X
  valid: Y
  test: Z

Number of features: XX

Number of examples per class
  class 0: YY
  class 1: ZZ


Descriptive statistics
----------------------

Age
  mean: XXX
  median: YYY
  standard deviation: ZZZ


Results
-------

Model 1 <print the name>
```

```
  Accuracy: XXXX
  F1-score: YYYY
  AUC: ZZZ

Model 2 <print the name>
  Accuracy: XXXXX
  F1-score: YYYYY
  AUC: ZZZZ

Model 3 <print the name>
  Accuracy: XXXXX
  F1-score: YYYYY
  AUC: ZZZZ
```

While the first part of this submission could be completed by simply copying an existing notebook, the second part cannot. Your code will be marked based on it's originality and the extent that it reflects an understanding of the task. Extensive copying will be considered plagiarism and Turnitin will be used for it's detection. For this assignment, learning and understanding are more important than prediction accuracy.

Good luck!

Hjalmar