Carnegie Mellon University Africa

COURSE 18-785: DATA, INFERENCE & APPLIED MACHINE LEARNING

KAGGLE

Nchofon Tagha Ghogomu

ntaghagh

December 06, 2024

LIBRARIES:

The following libraries were used:

- Pandas[1]
- Numpy[2]
- Pyplot from Matplotlib [3]
- Stats from Scipy
- Logistic Regression from Sklearn Linear model

- from sklearn.metrics import confusion_matrix, classification_report
- from sklearn.model_selection import train_test_split
- from sklearn.tree import DecisionTreeClassifier
- from sklearn.ensemble import RandomForestClassifier
- from sklearn.model_selection import train_test_split
- from sklearn.metrics import accuracy_score
- from sklearn.metrics import roc_curve, auc
- from sklearn.metrics import classification_report, accuracy_score
- from sklearn.model_selection import cross_val_score
- from sklearn.preprocessing import StandardScaler
- from sklearn.linear_model import LassoCV
- from sklearn.neighbors import KNeighborsRegressor
- from sklearn.tree import DecisionTreeClassifier
- from sklearn.metrics import classification_report, accuracy_score
- from sklearn.model_selection import cross_val_score
- from sklearn.preprocessing import StandardScaler

Programming Language:

- Python

INTRODUCTION

This is a brief report on the work done on the Kaggle competition. In this exercise, we try to test the accuracy of various models on the Kaggle train data set. The best model is selected purely based on accuracy score and submitted.

SOLUTIONS

In this exercise, there choice of models to benchmark was chosen based on their performance in the previous exercised. We will be benchmarking:

- Random Forest
- Logistic Regression
- Decision Tree

The accuracy score of each of these models was the comparison parameter used during bench marking:

A) Random Forest:

To construct a random forest:

i) The titanic data is loaded, and the features of interest are selected: ('age', 'sex', 'pclass', 'survived'). 'Survived' is the dependent variable and the rest independent.
ii) Null rows for survived were dropped and dummy variables was created for 'sex' and 'pclass'.
iii) Null 'Age' Values were replaced with the mean age.
iv) A percent of the training data is used for testing and the rest for training
v) A range from 50 to 1000 with a step of 50 is created. This will later help us use the optimal number of leaves.
vi) For every number of trees in the range, the model is created, and the accuracy is recorded.
vii) The best number of trees corresponds to that which produces the best model accuracy.
viii) We use the best number of trees to calculate the best number of leaves based on accuracy score of leaves on a scale of 50 to 1000 with a step of 50
ix) A random forest is now constructed with these best parameters and a random state of 24.

The model accuracy of the random forest was calculated to be:

```
RF accuracy 0.8888888888888888
```

B) **Decision Tree:**

i) Import our dataset
ii) Extract independent variables (gender, age, pclass) dependent variable(survived)
iii) Clean data. Here, we replaced the missing age values with the mean age and transformed sex data into categorical data (0 for male and 1 for female)
iv) Employed sklearn's DecisionTreeClassifier and fit our X and y variables to it.
v) The accuracy with the depth of the decision tree is calculated and the maximum depth chosen
vi) This parameter is used to train the final model

The accuracy score was obtained to be:

```
Score for decision tree: 0.8666666666666667
```

3) **Logistic Regression:**

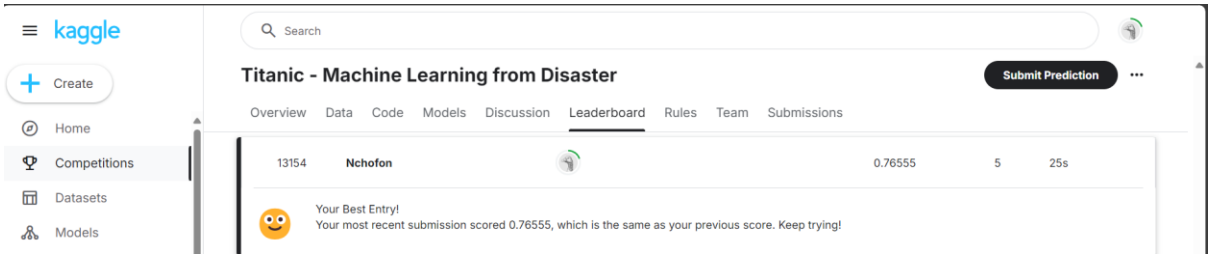
Similarly, the data was cleaned and trained on the train dataset. The accuracy score was also

```
Score for logistic regression: 0.86
```

From the accuracies of the model on this training dataset, we observed that random forest yielded the best results with an 88.89% accuracy.

RESULTS

For the model with the highest performance (Random Forest) the Patient ID and corresponding predictions were written to the a file and submitted into Kaggle. This model yielded a **76.555%**



Subsequently, I indent to explore ensemble methods to examine their performance on the data.

Course Conclusion and Reflection.

Over the course of this course, we have been able to explore a multitude of ways and technics we can use to infer data from a given data set. We have been able to manipulate data in a way that gives us information and insights on the data. Finally, we were introduced to the fundamentals of machine learning. We have explored fundamental models, there area of application and their working principles. We have also learned through practice how to benchmark these models, and technics that will render these models more robust and accurate. On a personal note, this has given me greater confidence in approaching data exploration, analysis and the application of machine learning.

References

[1] “pandas documentation — pandas 2.2.2 documentation.” Accessed: Sep. 02, 2024. [Online]. Available: https://pandas.pydata.org/pandas-docs/stable/index.html
[2] “NumPy -.” Accessed: Sep. 02, 2024. [Online]. Available: https://numpy.org/
[3] “Matplotlib documentation — Matplotlib 3.9.2 documentation.” Accessed: Sep. 02, 2024. [Online]. Available: https://matplotlib.org/stable/