**DSTS Assignment 1**

Alan Gaugler

U885853

30/09/2023

**Section A – Link to the Tableau Dashboard**

My Tableau dashboard is on the following link:

XXXXXXXXX

**Section B – Regression and Classification Tables**

The results of the tables are included below. For more details of the creation of these tables please refer to the attached Jupyter Notebook for this assignment.

It must be noted that the data was data was standardized before modelling, so the numbers in the tables below are for standardized data. Standardization is important when the various features have different ranges. The modelling process will be biased more by the features with the largest variation. Standardization will give equal weight to all the features.

**Table 1 - Regression Model Summary Table**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Mean Square Error** | **Mean Average Error** | **R Squared** |
| Linear Regression 1 | 0.07471 | 0.19251 | 0.62083 |
| SGD Regressor 2 | 0.07465 | 0.19287 | 0.62112 |
| SGD Regressor 3 | 0.07459 | 0.19204 | 0.62145 |

Looking at the three metrics for the three models, the results are very similar. The initial linear regression model looks to be slightly worse than the Stochastic Gradient Descent models although the margin is extremely slight. Model SGDR3 is slightly better than SGDR2 which is expected as its best hyperparameters were determined using the randomized search. It was the best performing model on this dataset. It had the smallest MSE and MAE values and the highest R^2 value, winning on all three metrics.

It would have to be concluded that all models are very similar but SGDR3 performs the best for this dataset.

**Table 2 - Classification Model Summary Table**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| Logistic Regression 1 | 0.85178 | 0.85001 | 0.85178 | 0.85013 |
| Support Vector 2 | 0.85456 | 0.86294 | 0.85456 | 0.85661 |
| Random Forest 3 | 0.85943 | 0.85835 | 0.85943 | 0.85871 |
| Multi-Layer Perceptron 4 | 0.85317 | 0.85140 | 0.85317 | 0.85136 |

Looking at the table, all four models performed very similarly, however the random forest classifier has the best overall accuracy of 85.94% followed by the support vector machine at 85.46%. The SVM has the highest precision, but the RFC won on all the other metrics.

All the metrics were calculated by using weighted average which takes into account the class imbalance. Looking below, class 1 (Low Rating) is 65% of all the dataset and class 2 (High Rating) is 34%. This is not a huge imbalance, but it is still better to use weighted average.

In this dataset, which is reasonably balanced, and precision and recall are not critical, overall accuracy would be the most important measure.

To conclude, all models perform very similar, but the random forest classifier is the best.

**Section C – GitHub Commands and Procedure**

The task is to deploy the code and relevant documents onto a GitLab repository. I created the GitLab account with a 30-day free trial approximately a month ago. This free trial will expire before this assignment is marked. I have spoken with Dr Ibrahim Radwan regarding this, and he has informed us that it is OK to upload the assignment on a GitHub repository instead.

I will document the entire procedure involved in doing this, including all the required Git commands.

**Part 1 –** Create a GitHub Repository

I already have a GitHub account, so I did not have to create a new one.

I have signed in and created a new repository called “DSTS\_Assignment\_1”

It is created with a readme file, which contains a brief description of the deliverables and the assignment. This repository is made Public for simplicity.

XXXXXXX

**Section D – Docker Commands and Procedure**

The

XXXXXXX

**Section E – GitHub Link to the Source Code**

The

XXXXXXX please add me as a collaborator, my Gitlab account is radwanebrahim@gmail.com)

**Section F – Docker Hub Link to the Docker Image**

The

XXXXXXX