DATA 2204: Logistical Regression – Assignment #2 (Revised)

Name: Raj Dholakia

Student ID: 100813041

Table of Contents

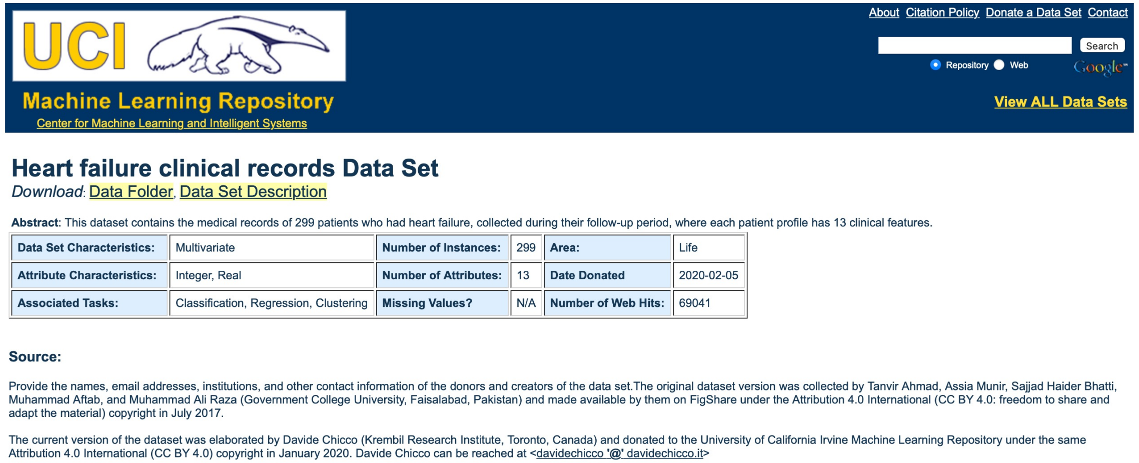
- Analysis Statement
- Key Features of the Dataset
- Understanding the Learning Curve
- Standard Logistical Regression Model
- Optimized Logistical Regression Model
- What should be the next steps?

Analysis Statement

Create a **standard and optimized Logistical Regression model** for the heart failure dataset and carry out additional analysis to better understand the performance of the model.

Key Features of the Dataset

- The dataset consists of twelve (12) independent variables and one dependent – ‘DEATH_EVENT’.
- ‘DEATH_EVENT’ provides information on whether a patient is deceased (`1`) or alive (`0`) during a follow-up period.
- There are a total of 299 samples in the dataset.
- The dataset has no missing values.
- There is a large imbalance in the dataset as the number of **deceased patients** examples are **96**, while the number of **alive patients** are **203**.
- As we are dealing with a person’s life


UCI
Machine Learning Repository
Center for Machine Learning and Intelligent Systems
About Citation Policy Donate a Data Set Contact
Search
● Repository ○ Web
Google
View ALL Data Sets

Heart failure clinical records Data Set

Download: Data Folder, Data Set Description

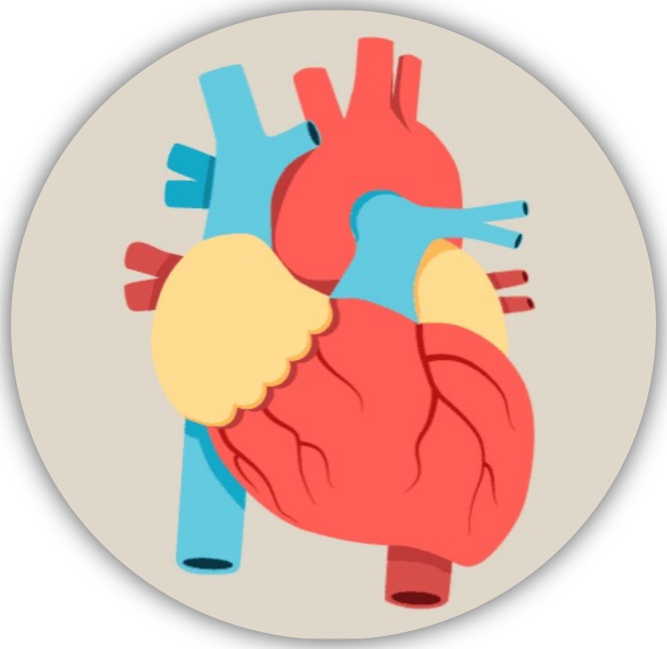
Abstract: This dataset contains the medical records of 299 patients who had heart failure, collected during their follow-up period, where each patient profile has 13 clinical features.

Data Set Characteristics:	Multivariate	**Number of Instances:**	299	**Area:**	Life
Attribute Characteristics:	Integer, Real	**Number of Attributes:**	13	**Date Donated**	2020-02-05
Associated Tasks:	Classification, Regression, Clustering	**Missing Values?**	N/A	**Number of Web Hits:**	69041

Source:

Provide the names, email addresses, institutions, and other contact information of the donors and creators of the data set.The original dataset version was collected by Tanvir Ahmad, Assia Munir, Sajjad Haider Bhatti, Muhammad Aftab, and Muhammad Ali Raza (Government College University, Faisalabad, Pakistan) and made available by them on FigShare under the Attribution 4.0 International (CC BY 4.0: freedom to share and adapt the material) copyright in July 2017.

The current version of the dataset was elaborated by Davide Chicco (Krembil Research Institute, Toronto, Canada) and donated to the University of California Irvine Machine Learning Repository under the same Attribution 4.0 International (CC BY 4.0) copyright in January 2020. Davide Chicco can be reached at <davidechicco '@' davidechicco.it>

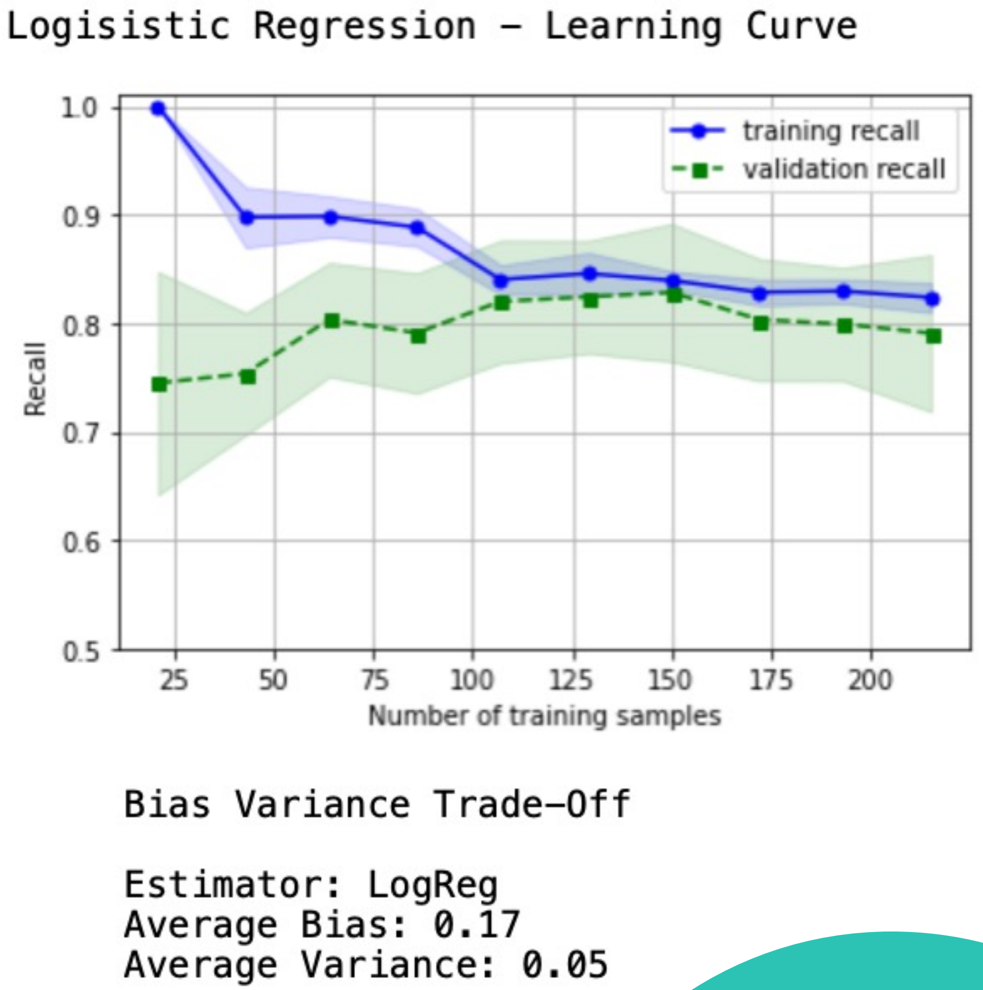


Understanding the Learning Curve

Some insights gained from the Learning Curve of the Logistical Regression Model:

1. As the training sample size increases, the model training recall decreases and the model validation recall increases only till the training size is 100 samples. Hence, **increasing the dataset size** beyond 100 samples should **not improve the performance** of the model.
2. The difference between the training and validation recall values can be considered as the variance in the model. Initially, the model has high variance, but when the model is trained on the more than 100 training samples, the **variance decreases** to a significantly low value.
3. High variance (large discrepancy between training and test recall values) for the model with fewer training samples shows that the model could be **overfitting** the data. Overfitting needs to be considered when evaluating ways to improve the model's performance.
4. With a **validation recall score of about 79%** (when trained on the full dataset), the model performance can be can work as a base model for development of other resources.

Note: the above insights can be confirmed from the values of the avg. bias and avg. variance.



Logisistic Regression - Learning Curve

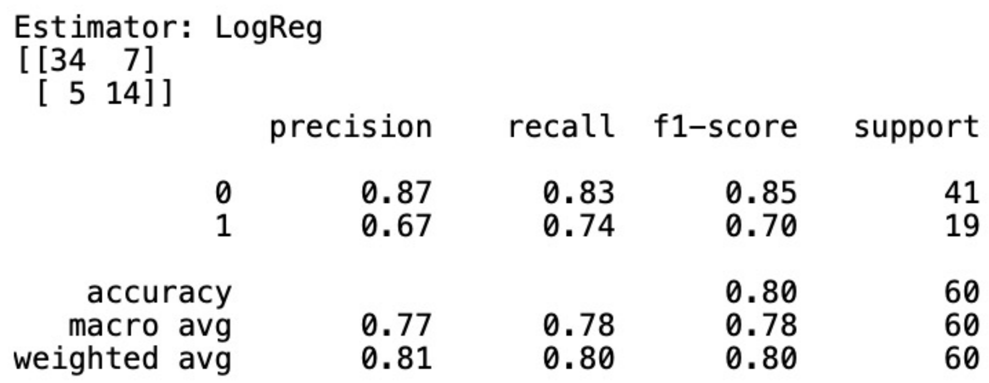
```
Bias Variance Trade-Off

Estimator: LogReg
Average Bias: 0.17
Average Variance: 0.05
```

Standard vs Optimized Logistical Regression Model

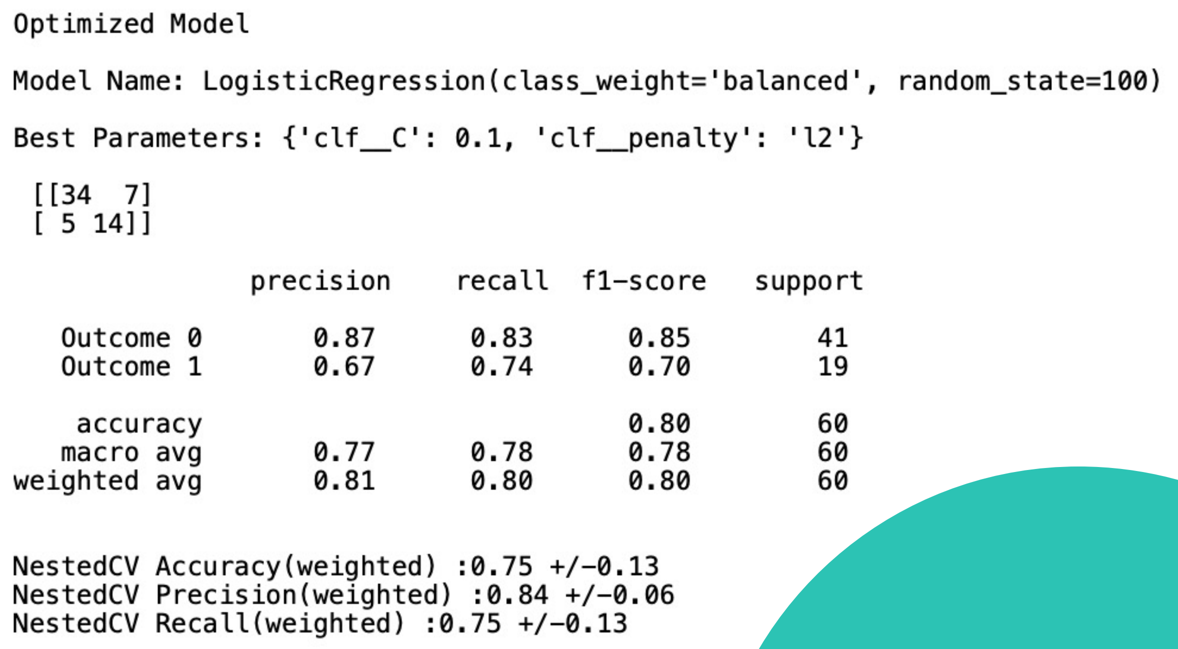
- The model has achieved an overall **F1 score of 80%**. As the dataset is imbalanced, the overall F1 score does not provide a holistic understanding of the model's performance.
- Considering the model's performance for each class individually, we observe a F1 score of **85%** for positive (deceased) and **70%** for negative (alive) samples (patients).
- There is huge discrepancy (**difference of 15%** in F1 scores) in the model's performance when predicting the two different classes.
- One of the reasons for this discrepancy can be the **imbalance in the data** (mentioned in Key Features section).
- The learning curve further supports the the above claim as it shows that a minimum number of samples is required for the model to perform consistently. The number of deceased samples might not be enough for the model to predict well.
- Hence, creating a more balanced dataset might improve the model's performance and make it consistent at predicting both the classes.

Comparing both the models:

- Both the models have achieved the same results for **all** the metrics – there is no difference in the model performances.



```
Estimator: LogReg
[[34  7]
 [ 5 14]]
              precision    recall  f1-score   support

           0       0.87      0.83      0.85        41
           1       0.67      0.74      0.70        19

    accuracy                           0.80        60
   macro avg       0.77      0.78      0.78        60
weighted avg       0.81      0.80      0.80        60
```



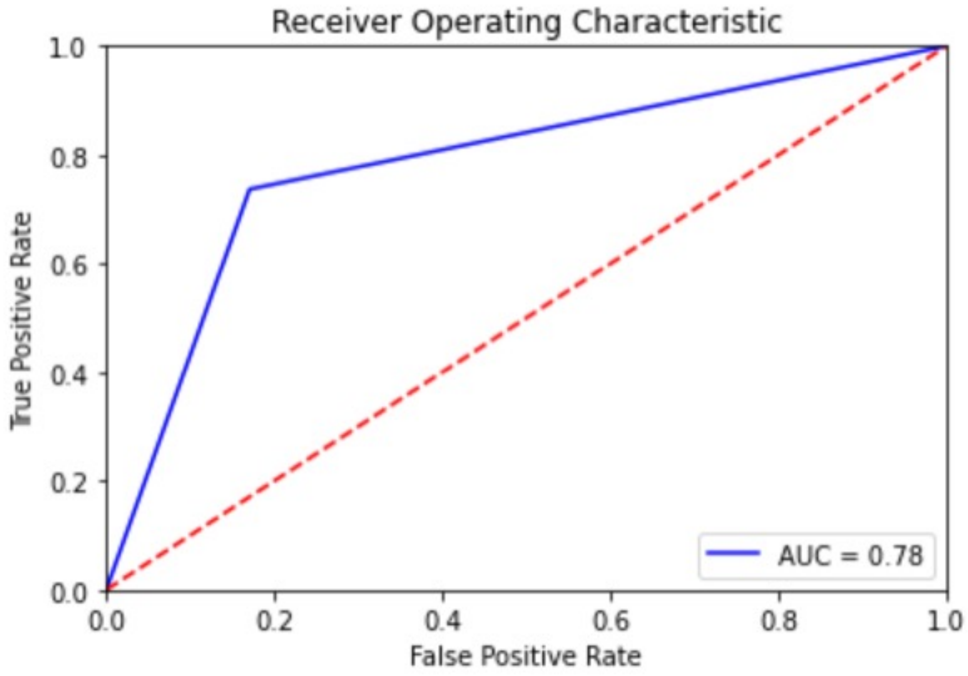
```
Optimized Model

Model Name: LogisticRegression(class_weight='balanced', random_state=100)

Best Parameters: {'clf__C': 0.1, 'clf__penalty': 'l2'}

 [[34  7]
 [ 5 14]]

               precision    recall  f1-score   support

   Outcome 0       0.87      0.83      0.85        41
   Outcome 1       0.67      0.74      0.70        19

    accuracy                           0.80        60
   macro avg       0.77      0.78      0.78        60
weighted avg       0.81      0.80      0.80        60


NestedCV Accuracy(weighted) :0.75 +/-0.13
NestedCV Precision(weighted) :0.84 +/-0.06
NestedCV Recall(weighted) :0.75 +/-0.13
```

Optimized Model

Hyperparameters and the ROC Curve

1. Only two hyperparameters were used to create the optimized model: the regularization parameter *C* and *penalty*.
2. Other hyperparameters (*solver*, *class_weight*) were predefined. Using the `*lbfgs*` solver limits the penalty norm to `l2`. Therefore, the only hyperparameter that was varied was *C*.
3. Varying only *C* brought about **no** significant improvements in the model's performance. This shows that the model's performance does not improve with changing increasing regularization. This reduces the possibility of the model overfitting the dataset.
4. Note: The evidence of overfitting deciphered from the learning curve does not seem to be a concern.
5. The **Area under the ROC** curve places the model's accuracy at **78%**.
6. Considering the dataset is imbalanced, the ROC curve does not provide a complete picture of the model's performance. The ROC curve would provide a better indication of how well the model is doing if it could break down the model's performance into positive and negative examples, separately (into two different ROC curves).

ROC Curve



What should be the next steps?

Mr. John Hughes can take into consideration the following options to better predict if a patient is deceased during the follow-up period ('DEATH_EVENT') using a Logistical Regression model:

1. **Improving dataset structure**: Improving the balance in the class samples in the dataset should make the model's performance more consistent. Hence, improving the model's performance at predicting positive examples, leading to an overall improvement in the performance.
2. **Feature selection**: doing feature selection can reduce the noise for the model, improving its performance. The low value of hyperparameter *C*, also indicates that a model with lower complexity is required to reduce overfitting and improve the model's performance. Feature selection could assist in doing that.

Thank you

Raj Dholakia

raj.dholakia@dcmail.ca

Student ID: 100813041