DATA 2204: Logistical Regression – Assignment #2

Name: Raj Dholakia

Student ID: 100813041

Table of Contents

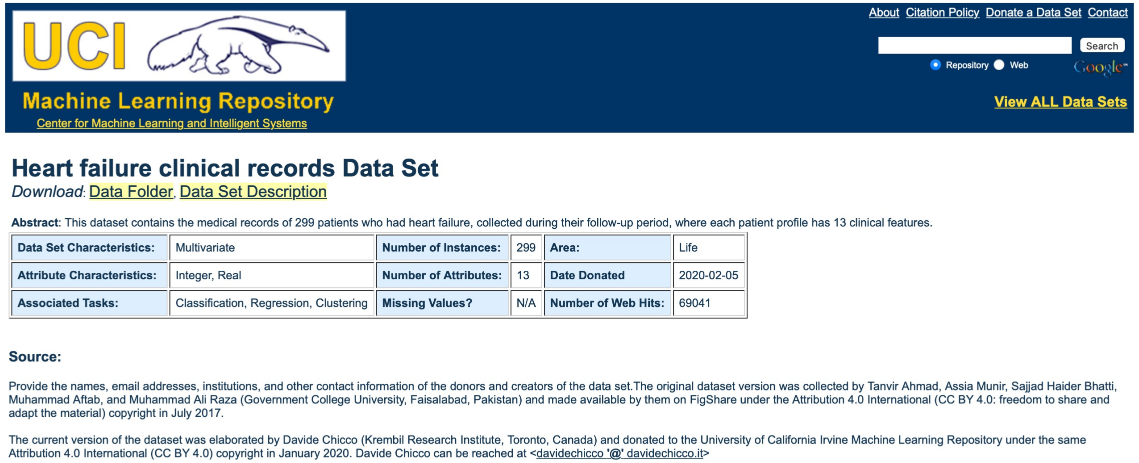
- Analysis Statement
- Key Features of the Dataset
- Understanding the Learning Curve
- Standard Logistical Regression Model
- Optimized Logistical Regression Model
- What should be the next steps?

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

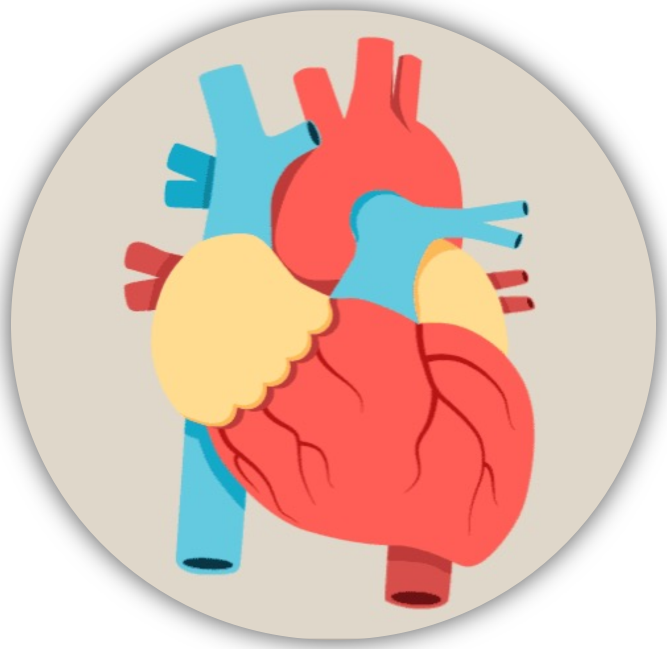
Abstract: This dataset contains the medical records of 299 patients who had heart failure, collected during their follow-up period, where each patient profile has 13 clinical features.

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>

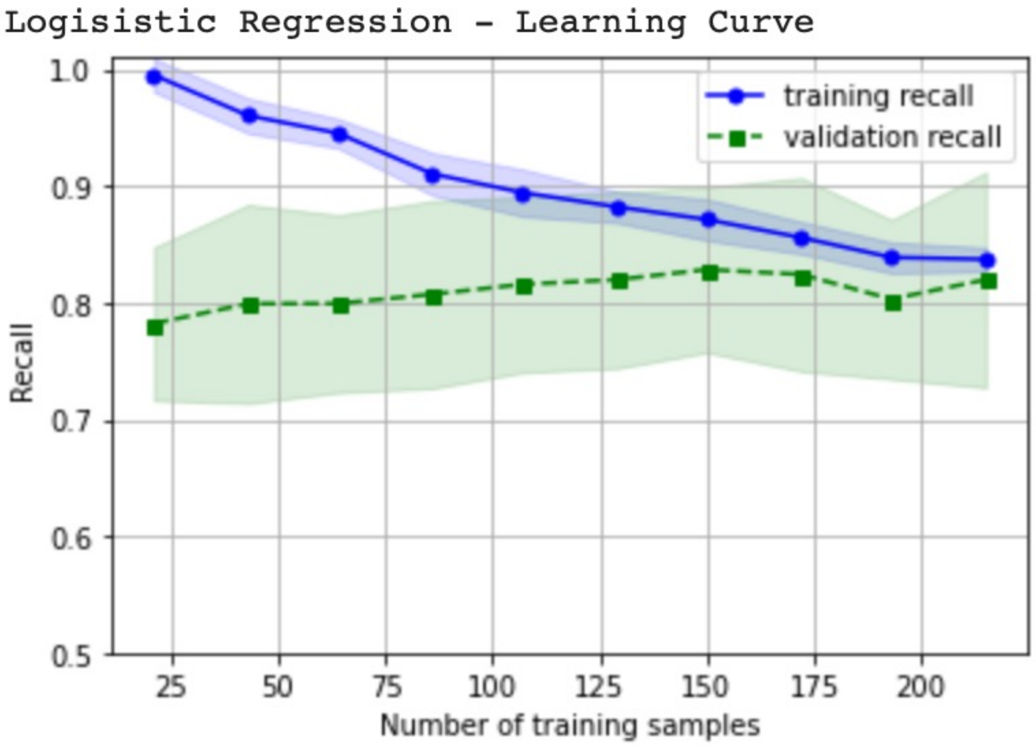


Understanding the Learning Curve

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. This shows that **increasing the dataset size** should ideally **improve the performance** of the model on the validation.
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 full dataset, 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 **recall score of about 82%** (when trained on the full dataset), the model performance can be can work as a base model for development of other resources.

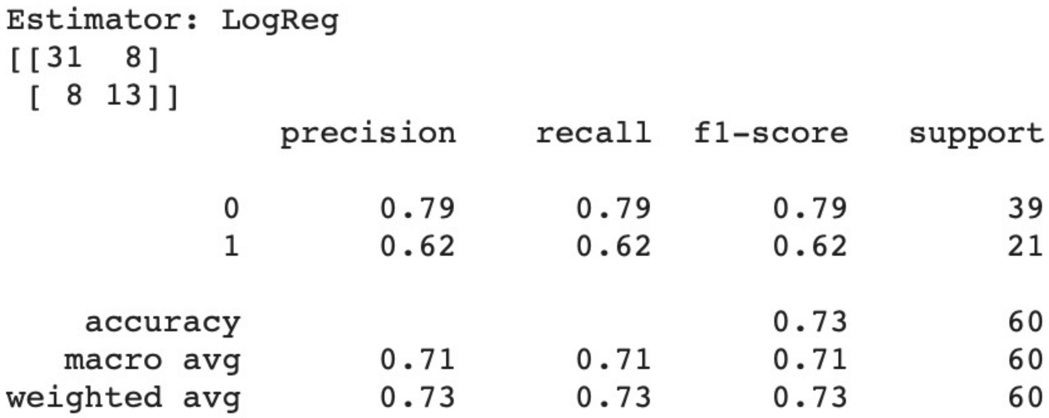
Note: the above insights can be confirmed from the values of the avg. bias and avg. variance.



```
Bias Variance Trade-Off

Estimator: LogReg
Average Bias: 0.21
Average Variance: 0.06
```

Standard Logistical Regression Model

- As the dataset is imbalanced, the weighted and macro averages of the metrics will not provide a holistic understanding of the model's performance. Hence, the model is evaluated by observing the metrics individually for positive (deceased) and negative (alive) samples.
- The standard model has a F1 score of **79%** when predicting negative examples and **62%** when predicting positive examples.
- The model performance is worse (**difference of 17%** in F1 scores) when predicting positive examples than negatives examples.
- Using insights from the learning curve, we can conclude that the reason for the above discrepancy in model's performance is the **imbalance in data**. Hence, adding more positive examples to the dataset to improve the balance, which will in turn make the model's performance more consistent at predicting both the classes.

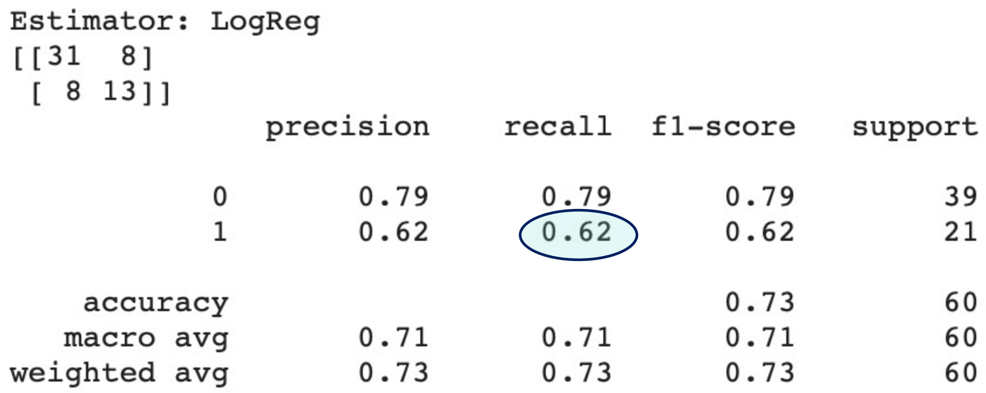


```
Estimator: LogReg
[[31  8]
 [ 8 13]]
              precision    recall  f1-score   support

           0       0.79      0.79      0.79        39
           1       0.62      0.62      0.62        21

    accuracy                           0.73        60
   macro avg       0.71      0.71      0.71        60
weighted avg       0.73      0.73      0.73        60
```

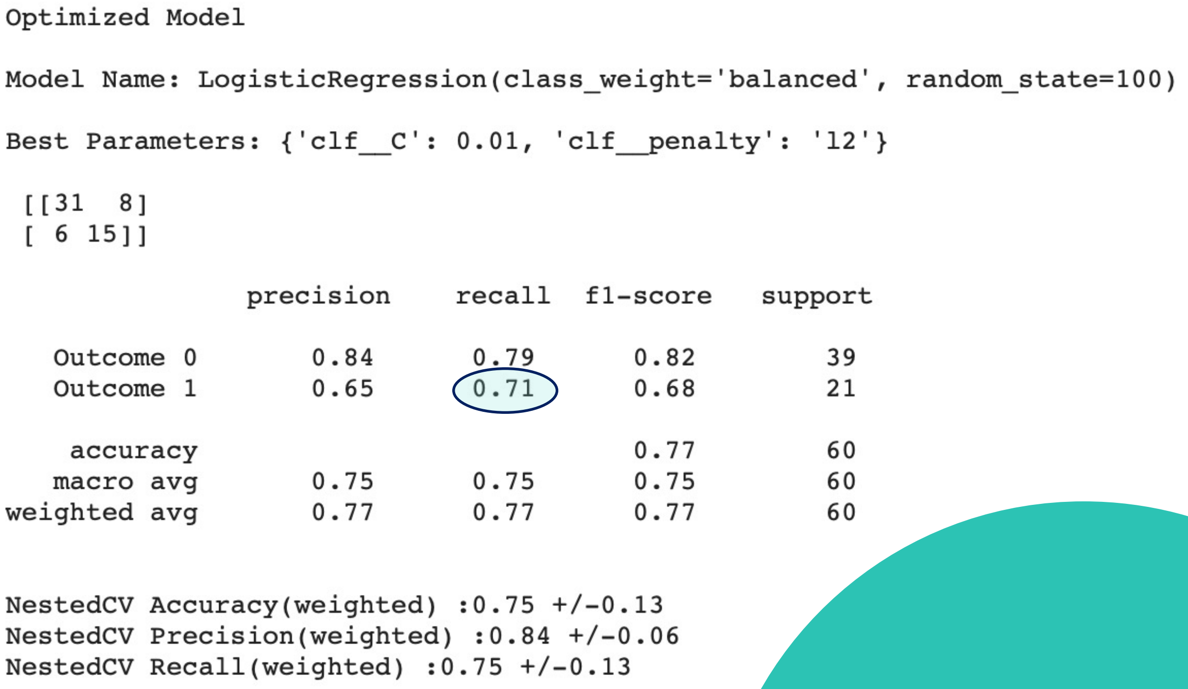
Standard vs Optimized Logistical Regression Model

Comparing both the models:

- There is an increase in F1 score of **3%** for positive examples and **6%** for negative examples after optimization.
- The optimized model observed the highest increase in recall for positive examples. This can also be seen by observing the false positive example in the confusion matrix
  - Standard model: 8 false positives (62% recall)
  - Optimized model: 6 false positives (71% recall)
- On the other hand, the recall for negative examples as **not changed**. This can be observed from the confusion matrices as well.
  - Standard model: 8 false negatives (79%)
  - Optimized model: 8 false negatives (79%)
- Precision values have significantly improved for both the classes.



```
Estimator: LogReg
[[31  8]
 [ 8 13]]
             precision    recall  f1-score   support

           0       0.79      0.79      0.79        39
           1       0.62      0.62      0.62        21

    accuracy                           0.73        60
   macro avg       0.71      0.71      0.71        60
weighted avg       0.73      0.73      0.73        60
```

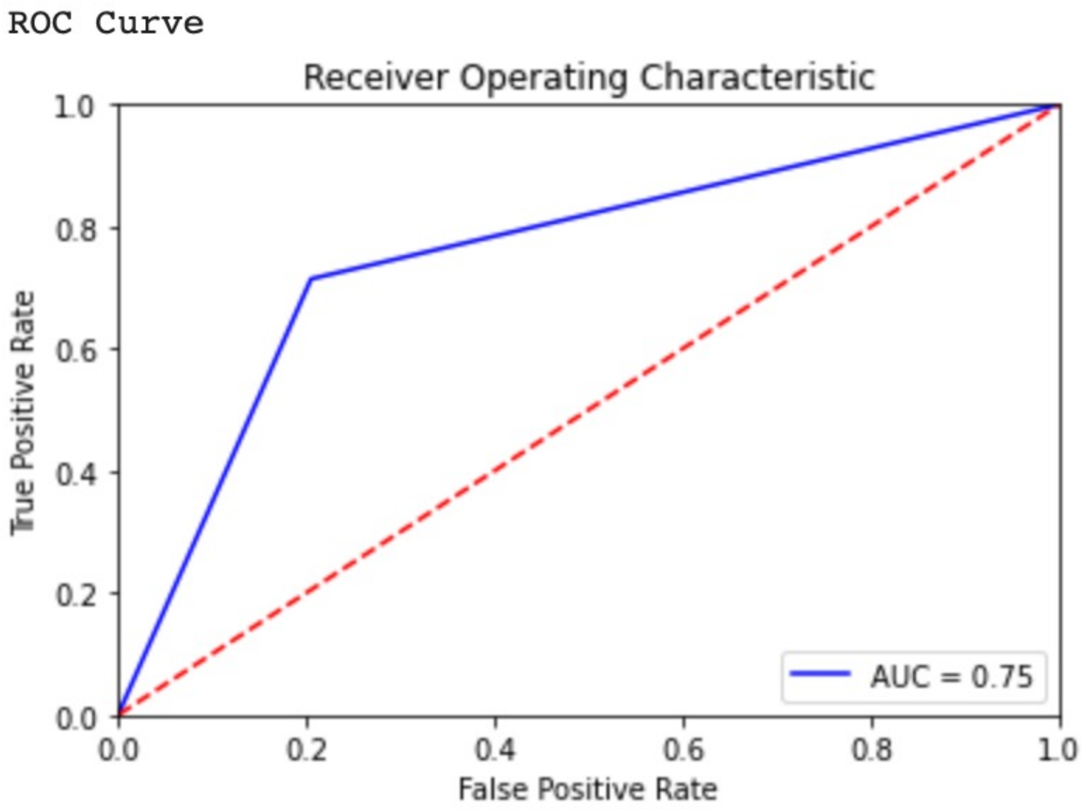


```
Optimized Model

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

Best Parameters: {'clf__C': 0.01, 'clf__penalty': 'l2'}

 [[31  8]
 [ 6 15]]

              precision    recall  f1-score   support

   Outcome 0       0.84      0.79      0.82        39
   Outcome 1       0.65      0.71      0.68        21

    accuracy                           0.77        60
   macro avg       0.75      0.75      0.75        60
weighted avg       0.77      0.77      0.77        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 some significant improvements in the model's performance. Namely, an increase in the F1 scores of **3%** for positive examples and **6%** for negative examples was observed.
4. The optimum value of *C* determined through grid search is `**0.01**`. The value reflects that a **high amount of regularization** is required to improve the model's performance. Hence, the model is usually overfitting the dataset. The evidence of overfitting can be deciphered from the learning curve.
5. The **Area under the ROC** curve places the model's accuracy at **75%**.
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).



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. **Get more data samples**: As seen in the learning curve that increasing the size of the dataset reduces chances of overfitting and creates a more reliable logistical regression model.
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