Logistic Regression: Cardiogenic Shock Prediction

Arewa Morountudun Ojelade
Department of Data Analytics: University of Maryland Global Campus
DATA 645: Machine Learning
Dr. Firdu Bati
November 11, 2025

Introduction

The *Worcester Heart Attack Study (WHAS500)* collected observations from 500 patients with 22 attributes in an attempt to predict factors associated with acute myocardial infarction and survival rates following hospital admission. The WHAS500 dataset was created for a study by Dr. Robert J. Goldberg of the Department of Cardiology at the University of Massachusetts Medical School. The objective of this logistic regression analysis report is to utilize the WHAS500 dataset to identify key features and develop an optimized, accurate model that predicts patients who experience cardiogenic shock during hospital admission. Hospitals are empowered to utilize such a model to determine whether patients are likely to experience heart failure and are at greater risk of death upon admission.

The methodological approach to model building utilizes ISLP.models methods, such as ModelSpec and summarize, to isolate features that most correlate with the target feature, and Sci-kitlearn's GridSearchCV method to construct a pipeline that includes a scaled logistic regression classifier trained on a split dataset, tuned using hyperparameters classifier_C with a list of values set at 0.1, 1.0, and 10, and the classifier_solver list with values set to *liblinear* and *lbfgs*. The pipeline ends with cross-validation across five separate folds. The result is a model with parameters that produces the most accurate model, which is then chosen and accepted.

What is Cardiogenic Shock?

Defective myocardial performance resulting in reduced cardiac output, end-organ hypoperfusion, and hypoxia is defined as Cardiogenic Shock, which is a common cause of mortality following a heart attack. Although most causes of Cardiogenic Shock (CS) occur after a heart attack, it can also be caused by a myriad of conditions that damage the heart muscle. A

heart attack occurs when blocked arteries lose the ability to supply sufficient blood to the heart muscles that are necessary for its function. According to the Journal of the American Heart Association (JAHA), CS complicates 5% to 10% of acute Myocardial Infarction (MI) cases and is the leading cause of death after MI. The risk of CS incidences is higher in Asian/Pacific Islander women aged 75+ years. In recent years, there has been an increase in CS observations, which is influenced by improved diagnosis and better access to care. (Vahdatpour et al., 2019) This report aims to utilize the WHAS500 dataset to develop a prediction model that hospitals can employ to forecast the occurrence of CS.

Heart Failure Risk Analysis

The American Heart Association considers several factors during a heart failure risk analysis, including weight, physical activity level, smoking habits, cholesterol levels, and blood pressure, among others. Since cardiogenic shock is a type of heart failure, these factors influence the probabilities of occurrence (American Heart Association, 2023).

Exploratory Analysis

As the WHAS500 dataset name suggests, it contains 500 non-null patient observations across 22 attributes, four of which are possible target variables, including Atrial Fibrillation (*abf*), Cardiogenic Shock (*sho*), Congestive Heart Complications (*chf*), and Complete Heart Block (*av3*). The focus of this report is to predict the *sho* variable representing Cardiogenic Shock. The dataset contains a 4% occurrence of CS with 22 positive observations and 478 negative observations, indicating some tiling may be helpful before model building. Before data cleaning, the dataset possesses 18 features of integer type, three features of object type, and one feature of float type. The feature year is interestingly divided into cohorts, categorized as

follows: 1 = 1997, 2 = 1999, and 3 = 2001. This report hypothesizes that '*year'* will not be correlated with the target feature and anticipates its removal from the dataset.

This report focuses on the quantitative features 'hr' (heart rate), 'bmi', and 'age' for deeper statistical analysis. In the case of the '*hr*' feature, the mean is 87bpm and the central tendency is 85bpm. The interquartile range (IQR) is 31bpm. The '*bmi'* feature has a mean of 26.61, which falls into the overweight category.

Figure 1-1 (Heart Rate)

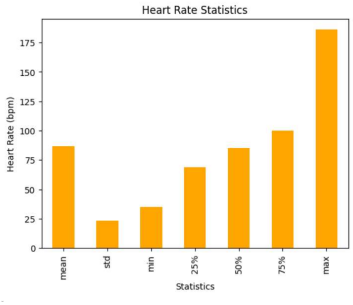


Figure 1-2 (BMI)

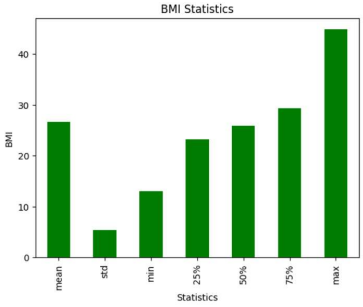
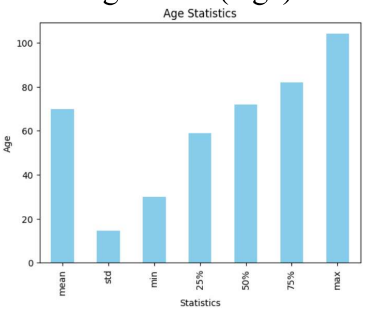


Figure 1-3 (Age)



Furthermore, exploration of the distribution of quantitative features in the dataset is recommended in order to better understand the univariate central tendency and variability. Both 'hr' and 'bmi' features are unimodal histograms that are slightly skewed to the right, while the 'age' feature is a bimodal skew-normal (BSN). (Elal-Olivero et al., 2020)

Figure 2-1 (Heart Rate)

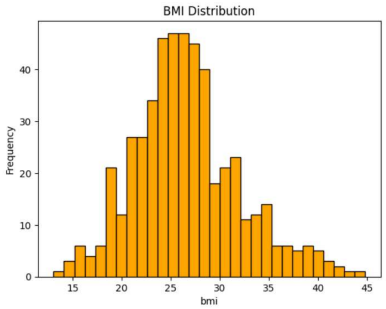


Figure 2-2 (BMI)

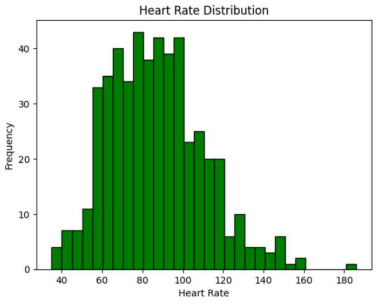
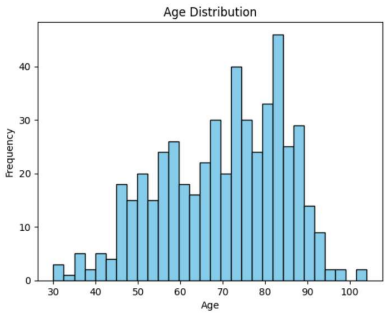


Figure 2-3 (Age)



Leveraging the pandas corr() function, a heatmap visualization is created to highlight the strength of the relationships between features.

Here, it is essential to consider the context to accurately assess the strength of the dataset features' relationships. For example, the feature '*id'* appears to be strongly correlated with *year;* however, due to the nature of the dataset, there is no actual correlation between it and any other features. This visualization confirms the earlier stated hypothesis, adding '*year*' to the list of variables to be dropped, which now includes '*id*'. Another correlation worth noting is the approximately 40% relationship strength between the target feature, cagidogenic shock status '*sho',* and the feature '*dstat'*, which represents discharge status from the hospital.

Figure 3

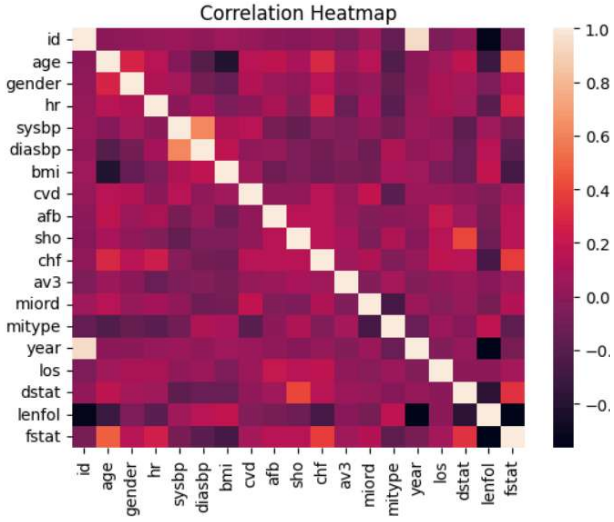
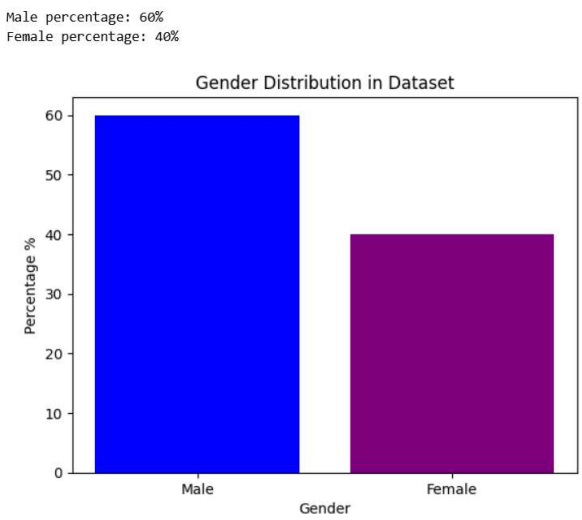


Figure 4



The dataset contains a protected feature, gender, with an appropriate male-to-female ratio of 60:40. Consequently, the dataset is ethically viable for further analysis and is ready for data cleaning.

Preprocessing

Data Cleaning

Redundant Features:

The data set possesses some redundant and target uncorrelated features, such as:

1. ID (target uncorrelated feature)

2. *fdate* - *lenfol* (redundant features)

    Date of last follow-up - Days between date of last follow-up and hospital admission

3. *admitdate* - *disdate* - *los* (redundant features)

    Hospital Admission Date - Hospital Discharge Date - Length of Hospital Stay

To eliminate unnecessary features, this data cleaning step drops the features id, fdate, admitdate, and disdate. The id feature, as previously explained, is dropped scine it commonly does not correlate with the target. The '*fdate*', '*admitdate*', and '*disdate*' are dropped since they are object date types and will be less effective at optimizing accuracy for the prediction model compared to their counterparts '*lenfol*' and '*los*', both of which are integer types ideal for optimal model building.

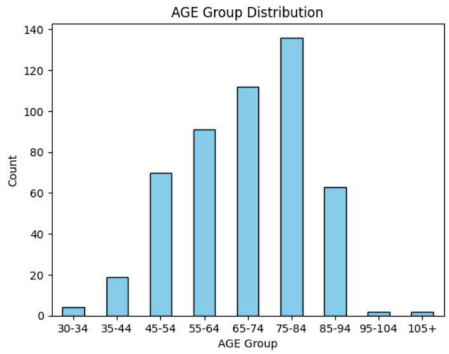
```
df.drop(columns=['id','fdate', 'admitdate', 'disdate'], inplace=True)
```

Binning

Utilizing the pd.cut() function, the age feature was divided into nine age group bins.

```
df['age_group'] = pd.cut(df['age'], bins=[30, 35, 45, 55, 65, 75, 85, 95, 100, 105],
                         labels=['30-34', '35-44', '45-54', '55-64', '65-74', '75-84', '85-94', '95-104', '105+'])
```

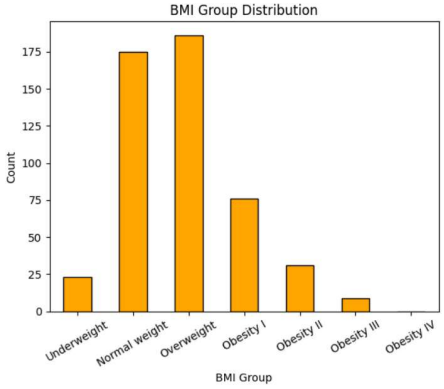
Figure 5



With the 'age' feature binned, the bimodal normal distribution is no longer evident; the histogram now displays a left-skewed unimodal structure.

The BMI group distribution showcases that patients fall mostly in the overweight and normal weight categories. There are no occurrences of patients in the Obesity IV group, and Obesity III has the least number of patients.

Figure 6



Data Transformation

Feature Engineering

The feature '*bp'* was engineered by dividing the systolic blood pressure by the diastolic blood pressure. The new feature contains information from both features and enables a bivariate comparison with another feature. Outliers were handled by eliminating values greater than the absolute value of three standard deviations from the mean. Figure 7-2 displays a dot plot of length of stay by BP. The trend line in red visualizes the slight positive correlation between the newly engineered feature and the number of days patients stay in the hospital.

Figure 7-1

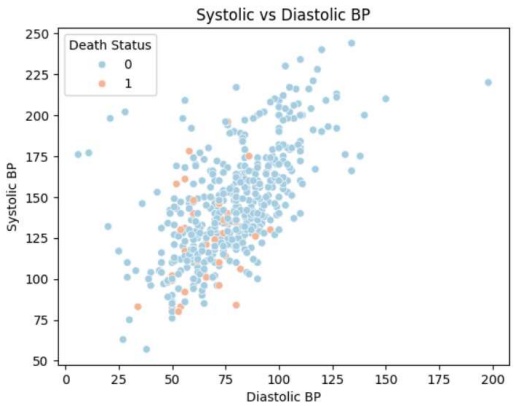


Figure 7-2

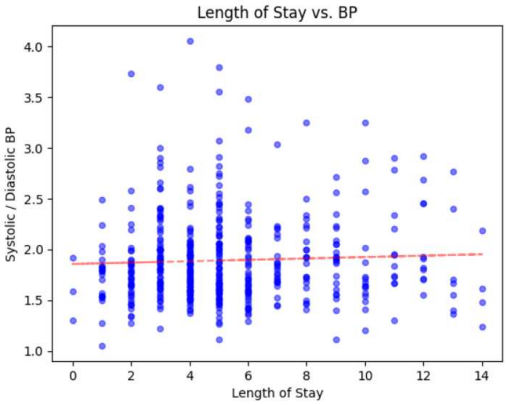
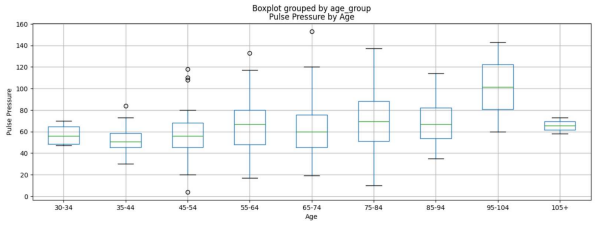


Figure 8-1

```
df['pulse_pressure'] = df['sysbp'] - df['diasbp']
df.boxplot(column='pulse_pressure', by = "age_group", figsize=(15, 5))
plt.title('Pulse Pressure by Age',)
plt.xlabel('Age')
plt.ylabel('Pulse Pressure')
plt.show()
```

✓ 0.1s

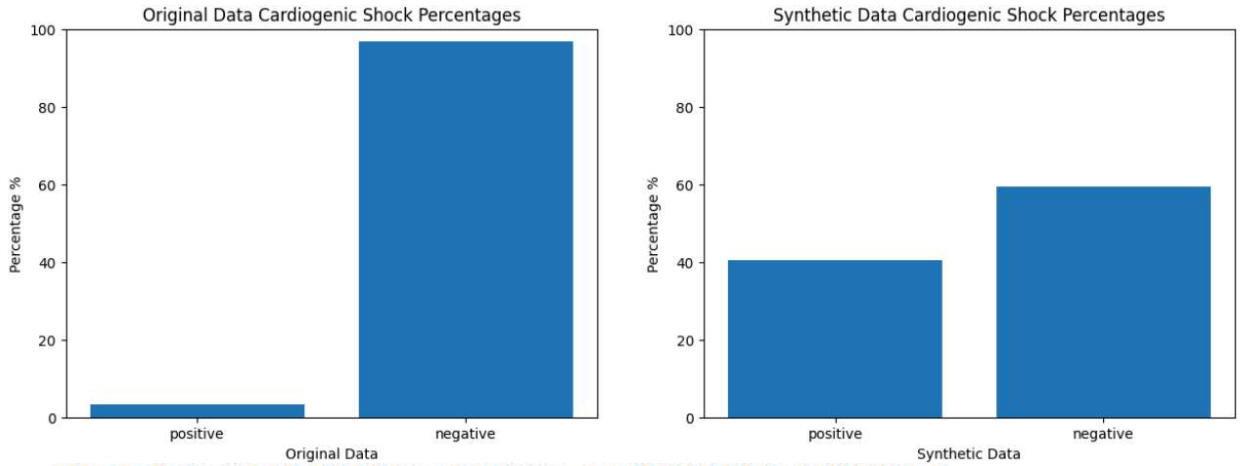
Figure 8-2



Another engineered feature is '*pulse_pressure'*, which was calculated by subtracting the diastolic bp from the systolic bp. Figure 8-2 displays the boxplot of pulse pressure by age group, showcasing a slight positive linear relationship between pulse pressure and age.

Synthetic Data Generation

The original dataset contained only 4% of the target feature's positive status. To adequately balance the dataset and ensure optimal accuracy, SHAP was leveraged to generate additional observations with a positive status. A synthetic generator was trained on the original dataset to create the synthetic data. The dataset was then split into training and test sets, and the training set was used to generate artificial data of size 8000 with statistical variation mimicking the original dataset, retaining only observations with a positive status. The newly generated data was then added back to the original training set, and the training set was recombined with the test set. The new dataframe's target feature now possesses a positive status of 41%.




```
New Target Positive Ratio: 0.41482444733420026
```

Result

Four models were created with different strategies to compare accuracy scores and the confusion matrix. This report aims to select the model with the optimal accuracy; however, it is essential to contextualize accuracy with scores such as precision, recall, and F1 to have a helpful model in production. The first model iteration was created using synthetic data built without feature selection, except for isolating the target and features of object types, such as bmi_group, which was trained on a pre-split dataset. This model, when tested, generated an accuracy score of 96% however, the precision score was 25%. The confusion matrix revealed one instance of predicted true positives out of 3. The accuracy of True Positives is overshadowed by that of the True Negative accuracy score, negating its usefulness in production. P-value feature selection was implemented to optimize accuracy, contextualized by better confusion matrix percentages.

P-value Feature Selection

The first two models, built with two distinct strategies, were compared using p-value selection. The first, as described above, was a model built using training data with aggregated synthetic target features with retained variation from the original dataset, which increased the dataset's target positive percentage by approximately 37%. The second model was built using training data aggregated with identical copies of the original dataset's target features, also known as tiling, thereby increasing the positive-to-negative target feature ratio from 4:96 to 40:60. The synthetic data model generated higher accuracy percentages when the model was built with feature selection of p-value < 0.05, with a 1- 4% increase. In comparison, the tiled data model generated a higher accuracy percentage of 4% percent when no p-value feature selection was conducted. Both models, though, had maximal high accuracy scores of 98% in the case of the synthetic data model and 88% in the case of the tiled data model, generated an abysmal

combination of precision, recall, and F1 scores of [25%, 33%, 29%](synthetic no feature selection) [0%, 0%, 0%](synthetic p-value feature selection), and [6%, 33%, 11%](tiled no feature selection) [0%, 0%, 0%](tiled p-value feature selection). The third model was created with a scikit learn GridSearchCV pipeline, where the dataset was scaled and transformed with StandardScaler, the optimal model chosen from PCA with feature selection of PCA number of components [5, 10, 15], solver set at ['liblinear', 'lbfgs'], and gridsearch cross-validation conducted across five separate folds. This method produced an overall accuracy score of 97%; however, the precision score for positive target values was 0% which nullifies the utility of this model.

Optimal Model Choice

In a Hail Mary attempt to generate a useful model, one was built with no complicated technicalities, of p-value feature selection, tiling, or synthetic data generation.The fourth and final model was generated using the scikit-learn LogisticRegression linear_model method. The dataset was separated from the target feature, then scaled using the StandardScaler, and split into training and test sets. A prediction was generated with the liblinear solver, producing a model with a 97% accuracy score contextualized with precision, recall, and F1-scores of [75%, 50%, 60%], respectively. This model's confusion matrix classified 143 true negatives out of 144 and 3 true positives out of 6. This report accepts this model as the optimal choice among all models generated in experimentation. However, since this model is intended for use in the healthcare industry, a 50% recall percentage would be unethical to implement in the real world. This report concludes that a larger dataset, incorporating features such as weight, physical activity level, smoking habits, and cholesterol levels, should be generated to build an accurate and ethical model for use in the healthcare field, as suggested by the American Heart Association.

Conclusion

In conclusion, the results of this report underscore a regular yet straightforward observation that occurs in everyday life. Sometimes the best solution is the simplest solution. In the case of this logistic regression report, a simple scikit-learn logistic regression model scaled with StandardScaler produced the optimal model choice, possessing an accuracy score of 97%, contextualized with precision, recall, and F1-scores of [75%, 50%, 60%] with the least ethical ramifications among all experimental methods. There were many early signs that precision and recall would become challenges during the exploratory analysis phase. The correlation matrix was quite dark, with the strongest relationship being the invalid correlation between ‘id’ and ‘year’. Additionally, the visualization showcasing the relationship between the engineered ‘bp’ feature and hospital length of stay displayed a negligible increase, barely indicating a relationship between the two features. Similarly, although the boxplot visualization displaying the average pulse pressure by age group shows a positive linear trend, it is not pronounced and indicates a muted trend.

This report aims to promote the contextualization of model accuracy scores with precision, recall, and F1 scores. As the result section outlines, failing to assess a model using various metrics can result in severe ramifications when deployed in production. In the case of cardiogenic shock prediction, this report recommends use of a model trained on a larger dataset with a suitable ratio of positive target instances where no tiling, or synthetic data is necessary with features contained in the whas500 dataset and added features to include but not limited to weight, physical activity level, smoking habits, and cholesterol levels, to build an accurate and ethical model for use in the healthcare field, as suggested by the American Heart Association.

References

American Heart Association. (2023). *Causes and Risks for Heart Failure*. Www.heart.org. https://www.heart.org/en/health-topics/heart-failure/causes-and-risks-for-heart-failure

Elal-Olivero, D., Olivares-Pacheco, J. F., Venegas, O., Bolfarine, H., & Gómez, H. W. (2020). On Properties of the Bimodal Skew-Normal Distribution and an Application. *Mathematics*, *8*(5), 703. https://doi.org/10.3390/math8050703

Explainable AI. (2024). *Explainable AI with Synthetic Data*. Google.com. https://colab.research.google.com/github/mostly-ai/mostly-tutorials/blob/dev/explainable-ai/explainable-ai.ipynb

Mayo Clinic. (2019). *Pulse pressure: An indicator of heart health?* Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/high-blood-pressure/expert-answers/pulse-pressure/faq-20058189

Obesity Action Coalition. (2025). *Classifications of Obesity*. Obesity Action Coalition. https://www.obesityaction.org/get-educated/understanding-your-weight-and-health/classifications-of-obesity/

R: Worcester Heart Attack Study WHAS500 Data. (2025). Rpkg.net. https://rpkg.net/packages/smoothHR/reference/whas500.ob

Scikit-learn. (2019). *sklearn.model_selection.GridSearchCV — scikit-learn 0.22 Documentation*. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

Vahdatpour, C., Collins, D., & Goldberg, S. (2019). Cardiogenic Shock. *Journal of the American Heart Association*, *8*(8). https://doi.org/10.1161/jaha.119.011991

WHAS500 Dataset Variable Description. (2025). Uniba.sk. http://www.iam.fmph.uniba.sk/ospm/Filova/surv/whas500.txt