DSO 568 - Heart Failure Prediction Report

Anna Natasha, Mihir Sabnis, Shaunek Choudhary, Matteo Meninni, Rishi Sunkavalli

Problem Definition

Cardiovascular diseases (CVDs) are the number one cause of death worldwide with an estimated 20.5 million lives taken each year as of 2021 representing 36.6% of all global deaths. 85% of them were due to heart attacks and strokes. In 2019, 38% of the 17 million premature deaths (under the age of 70) due to noncommunicable diseases were caused by CVDs. In 2020 in the United States, coronary heart disease (CHD) was the leading cause (41.2%) of deaths attributable to CVD in the United States. CVDs accounted for 12% of US health expenditures from 2018 to 2019. CVDs come in a plethora of forms from congenital heart defects to stress and anxiety-related heart failures.

Early identification of high-risk patients is vital to preventing adverse outcomes, early interventions for high-risk patients, reducing hospital readmissions, and allocating healthcare resources effectively so that the patients who need it have access when critical. There are so many important behavioral, genetic, mental, demographic, and external risk factors of heart disease and stroke, the most determinable of them all is the biometric and laboratory data of patients. However, with the large and complex amount of clinical data, there are times when it becomes difficult to make the right decision.

With this project, our team aims to address these challenges by developing a machine-learning model that is capable of predicting the likelihood of heart disease in patients based on clinical and diagnostic data. Our goal is to provide clinicians with a data-driven tool to identify at-risk patients before symptoms escalate or are shown. With this model, we can have timely interventions, personalized treatments, and just overall better patient care. The insights derived from the model’s predictions can also guide resource allocation, ensuring that high-risk patients receive prioritized care.

Data Preparation

The dataset used for this project is the Heart Failure Prediction Dataset from the UCI Machine Learning Repository, which combines data from five sources: Cleveland, Hungary, Switzerland, Long Beach VA, and Statlog. The final dataset consists of 918 observations and 11 features after cleaning and duplication. To prepare the data we took the following steps :

1. Data Cleaning:

- Removed 272 duplicate entries to ensure data integrity.
- Addressed any missing values to ensure completeness of the dataset.

2. Feature Engineering:

- Continuous variables such as Age, RestingBP, Cholesterol, MaxHR, and Oldpeak were normalized to ensure uniform scaling across features.
- Categorical variables, including ChestPainType, RestingECG, and ST_Slope, were encoded using one-hot encoding to make them suitable for machine learning algorithms.

3. Exploratory Data Analysis:

- Histogram distribution visualizations grouped by heart disease class were created to understand feature distributions and correlations for each class and to identify patterns of certain features that are heavily skewed to one class.
- Correlation heat maps identified relationships between variables, providing a basis for feature selection and engineering.
- Explored Numerical Features vs Categorical variables with scatter plots for each combination to identify stratified patterns in the data across multiple dimensions.

Descriptive Insights

Data Distribution:

- When splitting the data by class, they are almost evenly balanced with **44.7% of the samples labeled No Heart Disease** and **55.3% of the samples labeled Heart Disease**

- All categorical variables are nearly normally distributed.
- When looking at numerical variables **Oldpeak's** data distribution is rightly skewed and **Cholesterol** has a bimodal data distribution which means it is class-skewed.
- Demographic skew: More males have heart disease than those without. **90% of all heart disease patients are male**, indicating a significant gender disparity.

Critical Indicator Insights:

- **77% of heart disease patients have ASY chest pain**, making it a critical indicator.
- **RestingECG does not provide a clear distinction** for heart disease detection
- The **presence of angina significantly increases** the probability of being diagnosed with heart disease.
- **Max Heart Rate values below 140** have a high probability of being diagnosed with heart disease.
- A **Flat or Down ST_Slope value** is positively correlated with Heart Disease Cases
- **Resting Blood Pressure** values between **95–170** are most prone to heart disease.
- Cholesterol levels between **160–340 mg/dl** are highly susceptible to heart disease.

Model Building

Our team decided to use different machine learning models that we implemented, trained, and evaluated to identify the most suitable approach for our project. Each model underwent hyperparameter tuning and cross-validation to ensure that the model achieves optimal performance. The models are as follows :

1. Logistic Regression:

- The logistic regression model provided insights into the linear relationships between features and the target variable. It achieved moderate precision and recall but lacked the flexibility to capture non-linear patterns in the data.

2. Decision Tree:

- The decision tree model offered improved interpretability. It achieved moderate recall and precision but was prone to overfitting, as evident from its performance metrics.

3. Random Forest:

- This method improved generalizability by combining multiple decision trees. Its performance was competitive, particularly in recall and F1-score.

4. LightGBM:

- This boosting model provided exceptional performance, with high recall and precision. LightGBM emerged as a top choice due to its speed, accuracy, and feature importance interpretability.

5. XGBoost:

- Another boosting model, XGBoost showed strong performance but slightly behind LightGBM in recall and AUC-ROC.

6. Neural Network:

- The neural network captured complex, non-linear relationships in the data. It matched LightGBM in recall but required more computational power and lacked explainability.

Model Evaluation

We used the following metrics—precision, recall, F1 score, AUC-ROC, and log loss to address two critical concerns: minimizing false negatives, which could lead to undiagnosed high-risk cases, and controlling false positives, which could result in unnecessary interventions for low-risk patients.

Precision was used to measure the accuracy of positive predictions, ensuring that patients flagged as high-risk were genuinely at risk. This is crucial in reducing unnecessary medical interventions, which can be both costly and stressful for patients.

Recall was prioritized to ensure that most high-risk cases were identified. In a healthcare setting, missing a single high-risk patient can result in severe complications and increased treatment costs, making recall a critical metric. The F1 score provided a balance between precision and recall, offering a single measure to evaluate the models' overall reliability.

The AUC-ROC evaluated the models' ability to distinguish between high- and low-risk patients across various thresholds. A high AUC indicates robust performance in stratifying patients effectively. Lastly, log loss quantified the confidence in probabilistic predictions, with lower values indicating more reliable outputs.

The models showed varying levels of effectiveness. Logistic regression struggled with capturing non-linear patterns which meant its recall and precision were limited. Decision trees improved interpretability and handled non-linear relationships better but were prone to overfitting, which impacted their generalizability. Random forest and XGBoost offered higher accuracy and recall, but their computational demands were significant.

The neural network stood out for its ability to model complex relationships, achieving strong recall and AUC-ROC scores, but its lack of transparency came in the way of its adoption in clinical environments.

Overall, LightGBM was our winner offering the best balance of performance, efficiency, and interpretability. With a recall of 90%, an AUC-ROC of 0.92, and the lowest log loss among the models, LightGBM was the most reliable model for our prediction.

The scores for each model are below :

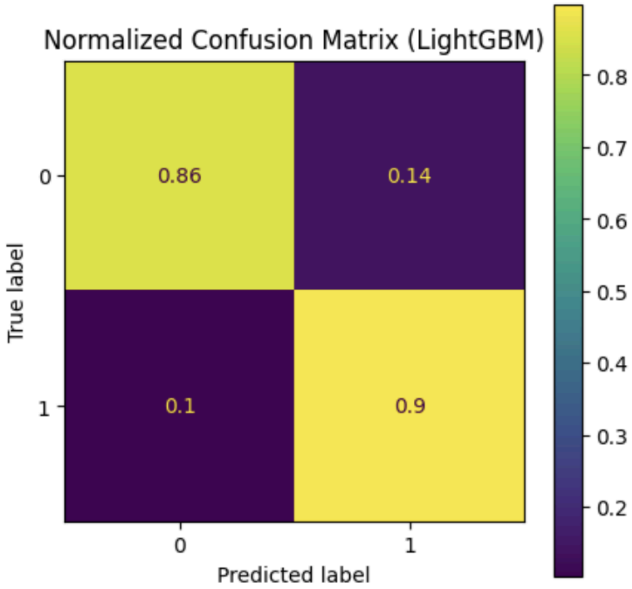
Model	Precision	Recall	F1 Score	AUC - ROC	Log Loss
Logistic Regression	0.89	0.79	0.84	0.92	0.37
Decision Tree	0.87	0.78	0.82	0.89	1.30
Random Forest	0.89	0.88	0.88	0.92	0.37
LightGBM	0.90	0.90	0.90	0.93	0.34
XGBoost	0.89	0.89	0.89	0.93	0.35
Neural Network	0.92	0.92	0.92	0.93	0.34

Insights

LightGBM

Based on the values above, we decided on the **LightGBM model as our main model**, but we also want to highlight several importances from using the neural network model. The LightGBM model demonstrated a **precision of 0.90**, which is almost on par with the neural network's. This means that when the model predicts a patient is at risk of heart failure, it is correct 90% of the time. This high precision is important because it reduces false positives and avoids unnecessary interventions for low-risk patients. From a cost savings perspective, we can potentially reduce constant tests and overall costs.

The model's **recall was 0.90**, which means it can identify 90% of all true high-risk cases. Having a recall this high ensures that the majority of at-risk patients are flagged for more evaluation, which will minimize the likelihood of missing cases and prevent emergency care costs.



With an **F1 score of 0.90**, LightGBM shows that it has an effective balance between precision and recall, making it a model we can depend on for predicting heart failure risk. This will help doctors or professionals in making accurate and reliable decisions, reducing mismanagement and ensuring appropriate care is provided.

The model’s **AUC-ROC score of 0.93** shows its ability to effectively differentiate between high- and low-risk patients. This ability to rank patients by risk level is critical for implementing care, and also allowing doctors to allocate resources where they are most needed. Lastly, LightGBM’s log loss of 0.34 shows its accuracy in predictions.

Neural Network

The neural network had a **precision of 0.92**, meaning that when it predicts a patient as high-risk, 92% of those predictions are correct. It is thus effective at minimizing interventions and saving healthcare costs. Its recall of 0.92 shows the neural network’s capability to identify the majority of high-risk patients, showing 92% of true positives and only missing 8%. This sensitivity is important in preventing issues caused by undiagnosed high-risk individuals.

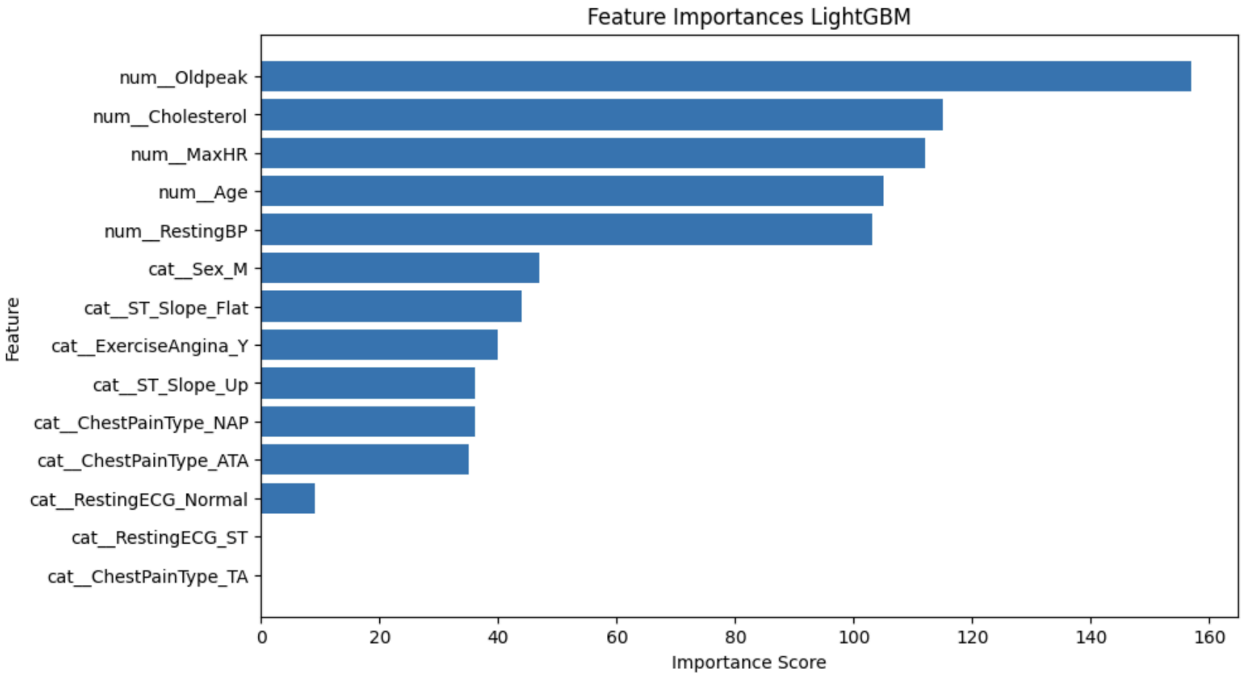
The **neural network's F1 score was 0.92**, which is good. This indicates that the model effectively identifies most high-risk cases while minimizing false positives, ensuring that only patients truly at risk are flagged for further action. An AUC-ROC score of 0.93 further shows the model’s excellent ability to distinguish between high- and low-risk patients, making it suitable for prioritizing care based on risk. Finally, the neural network’s log loss of 0.34 suggests confident and accurate predictions.

Model Comparison and Complexity

LightGBM and the neural network’s high precision allow for the proper identification of patients who require intensive care. For example, let’s say a patient with an 85% probability of heart failure could be prioritized for immediate diagnostic tests or preventive treatments. The models’ strong recall ensures that at-risk patients are not overlooked, even if this results in a few additional false positives. For example, a patient's symptoms such as younger age or low cholesterol could still be flagged.

The high AUC-ROC scores enable the models to rank patients by risk level effectively. Patients in the top 10% of predicted risk could receive advanced cardiac imaging, but those in the next 20% might not need the screening. This approach optimizes the allocation of expensive diagnostic tools and ensures that high-risk patients receive the right care.

LightGBM is effective at identifying high-risk patients while minimizing false negatives and reducing preventable complications. Implementing a high-risk alert system for flagged patients could facilitate regular follow-ups, counseling, and consistent monitoring, which would in turn potentially reduce heart failure readmissions. Redirecting resources to high-risk groups identified by the model could help prevent critical events and reduce emergency department visits. Using LightGBM's feature importance can also increase explainability to healthcare professionals and patients alike. We can see how Oldpeak, Cholesterol, and maximum heart rate are the main contributors to the model's final predictions.



On the other hand, the neural network model does a great job at spotting complex, non-linear patterns which makes it useful for managing borderline risk patients.

Ethical Considerations and Implications

Given the vast amount of data - protecting patient privacy was our team’s priority and all the data was fully anonymized to remove any important patient information. This would ensure compliance with HIPAA. Addressing bias was another main focus; we examined the dataset for disparities in demographics such as age and gender and we applied sampling and weighting techniques to ensure fairness during training. After the training, the models were evaluated across demographic subgroups to confirm consistency and to reduce the risk of unfair outcomes.

Transparency was also an important consideration as we need to uphold the trust of our patients. This is one of the reasons why we selected the LightGBM model as our final model; its interpretable outputs such as feature importance scores allow us to understand why a patient is flagged as high-risk.

As we get more patients and with more data, we must update the model regularly. In other words, the model must adapt to these changes. Routine performance checks will ensure the model remains accurate, unbiased, and aligned with current medical practices. Additionally, establishing a set of governance practices will ensure that the model will provide fair, transparent, and effective results. Without proper maintenance, our model will not be able to give the intended results.

Responsible Model Deployment and Usage

The best way to responsibly deploy such a model would be a phased deployment approach.

Phase 1: Planning and Validation

- **Stakeholder engagement:** Identify the main stakeholders which are patients, doctors and healthcare professionals, administrators, insurers, data scientists, and IT teams. Each stakeholder must be informed and trained accordingly with clear model objectives an understanding of expected outcomes, data collection, and privacy regulations, and a clear workflow of how the model will be used in a healthcare workflow. Regular training must be provided for the most responsible use of the model according to ethics guidelines. The use of the model can also help insurers with more information on setting premiums.

- **Technical Model Validations**: Confirm the major metrics of F1 score, recall, and AUC-ROC and assess the performance on more seen and unseen data. More emphasis should be placed on developing more diverse and representative data that also includes lifestyle and habits, comorbidities, medications, dietary habits, and mental health self-reports. This could allow for a more complex holistic approach to predicting heart disease.
- **Evaluate Interpretability:** Clinicians should be able to easily understand predictions and further prediction bias should be identified and addressed. More evaluation like SHAP values can be used to better understand the feature relation and importance with the target variable.
- **Infrastructure Integration:** Integrate the model into existing Electronic Health Record (EHR) systems or other clinical tools. Ensure compliance with data security and privacy regulations (e.g., HIPAA, GDPR). Set up secure, scalable infrastructure for model execution, such as cloud or on-premises solutions.

Phase 2: Pilot Implementation

- **Controlled Deployment: This** should be done in a single department for pilot testing. The model's predictions should be assessed in real-time with the clinician. The patients should be briefed on the use of the model and how it will be used as a supporting measure to better predict their health. Emphasize that the model aids, not replace, clinical judgment.
- **Key Performance Metric Tracking:** Clinical metrics like accuracy of risk predictions, and reductions in missed diagnoses, and operational metrics like resource allocation, time savings, and workflow efficiency should be tracked.

Phase 3: Broader Implementation

- **Scale across departments:** Once the model has passed the relevant baseline expectations in the pilot, it can then be expanded into other departments and across hospitals in the same system. The workflows must be tailored and integrated into departments that manage high-risk patients like cardiology.
- **Emphasise Automation and Data Collection:** Automate model execution to minimize manual intervention. This could mean triggering risk assessments during patient check-ins or routine evaluations and auto-generating follow-up recommendations for flagged cases.

Phase 4: Maintenance and Updates

- **Continuous Model Monitoring:** We need to track model performance over time and evaluate metrics like recall, precision, and AUC-ROC, monitor model drift to ensure predictions remain accurate, and set up alerts for performance degradation.
- **Periodic Retraining:** Retrain the model using new data to incorporate changes in patient demographics or clinical practices. Use an automated pipeline for retraining and deployment and continually explore the data relations and how the feature importance evolves with more diverse data. Data cleanliness is incredibly important as well as encryption and strict access controls that are granted over time.
- **Bias Audits:** Regularly assess fairness across demographic subgroups to mitigate bias and include diverse data sources to ensure equitable predictions.

References:

- Heart Failure Prediction Dataset, Fedesoriano | *Kaggle.com*
  https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction
- British Heart Foundation Facts and Figures | *bhf.org*
  https://www.bhf.org.uk/what-we-do/news-from-the-bhf/contact-the-press-office/facts-and-figures
- World Health Organization CVDs | *who.int*
  https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-%28cvds%29
- 2023 Heart Disease and Stroke Statistics Update Fact Sheet | *professional.heart.org*
  https://professional.heart.org/-/media/PHD-Files-2/Science-News/2/2023-Heart-and-Stroke-Stat-Update/2023-Statistics-At-A-Glance-final_1_17_23.pdf