Heartbeat Prediction

AI Final Project Report

CONTRIBUTORS

Adarsh Kudithipati
Mahathi Mandelli
Leo Lu
Shruti Daundkar
Shraya Ramamoorthy

Project link: https://github.com/shrutidaundkar/Heartbeat-Prediction-Summer23.git

INTRODUCTION

In the contemporary era, cardiovascular diseases, including heart strokes, have emerged as common health concerns. Unhealthy food habits and the consumption of processed foods have been identified as significant factors contributing to this alarming trend. Advanced Machine Learning techniques have shown promising applications in the healthcare domain. This project aims to explore the effectiveness of Machine Learning models in predicting heartbeats accurately using medical data. Specifically, we will compare the performance of Random Forest Regression and CatBoost models for this predictive task, based on the Stroke Dataset 2023 sourced from Kaggle.

The dataset captures a comprehensive set of health attributes and stroke outcomes, which has allowed us to construct and assess the performance of the selected models. From our results, we compared the accuracy, precision, recall, and area-under receiver operating characteristics curve (AUC-ROC) results to determine the most effective model for heart stroke prediction.

METHODS

The Stroke Dataset 2023 contains records for 5169 patients, of which includes the status of stroke occurrence diagnosis and a set of 10 attributes. (See *Table 1*). The dataset encompasses a range of attributes that captures comprehensive information about individuals that may affect their cardiac health. The dataset includes demographic details, and subjective lifestyle choices.

Table 1. Attributes and respective data types for Stroke Dataset 2023

Attribute	Data Type
gender	categorical - “Male”, “Female”, “Other”
age	int
hypertension	binary - 0: no | 1: yes
heart disease	binary - 0: no | 1: yes
ever married	binary - 0: no | 1: yes
work type	categorical - “Government”, “Private”, “Self-Employed”, “Never worked”
residence type	categorical - “Rural” or “Urban”
average glucose level	double
bmi	double
smoking status	categorical - “Formerly”, “Never”, “Smokes”, “Unknown”
Stroke	binary - 0: no | 1: yes

DATA PROCESSING

In order to preprocess the raw medical data and extract relevant features that can contribute to accurate heartbeat predictions, the dataset is visualized to examine general data trends. These features may include temporal statistics, frequency domain characteristics, and other medical parameters.

Box plots and density distribution graphs were generated to examine outliers (Figures 1 & 2). Additionally, age distribution histograms for smoking, hypertension, and heart disease were generated to examine possible abnormal trends throughout various age demographics. From visual inspection, the histograms corresponded with trends from public studies for smoking (CDC 2020), hypertension (D. M. Lloyd-Jones *et al.* 2005), and heart disease(J. L. Rodgers *et al.2019*). Therefore, it was determined that the dataset is a good representation of the general public to be used for heart beat modeling.

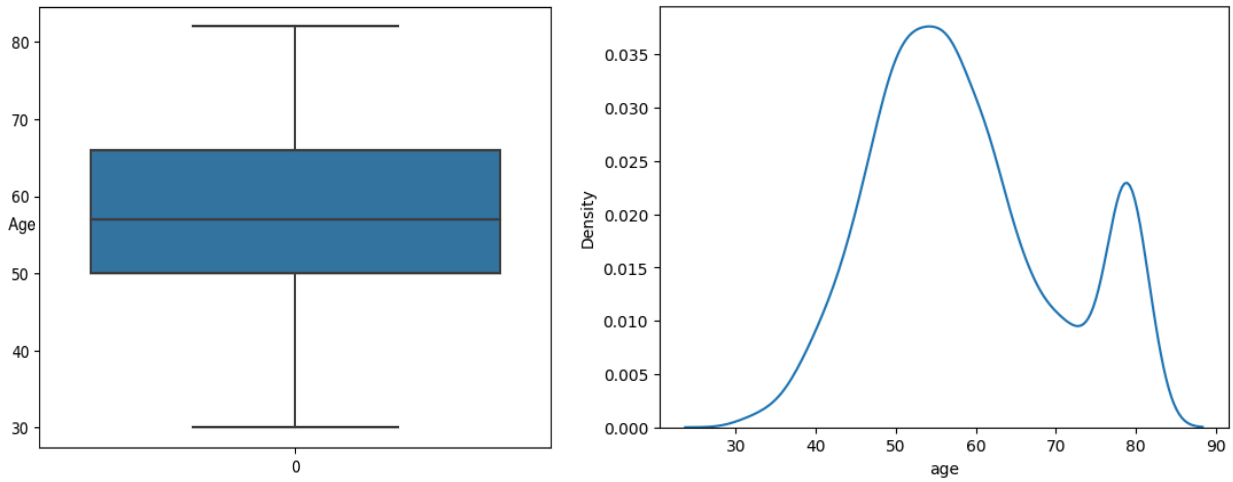

Figure 1. Box plot and density graph of population age distribution in the dataset prior to preprocessing

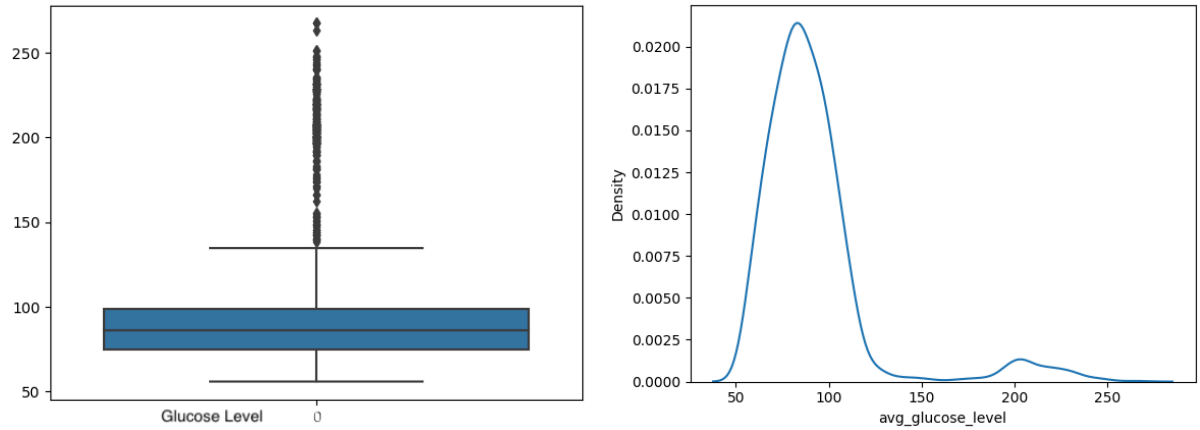

Figure 2. Box plot and density graph of average glucose level distribution in dataset prior to preprocessing

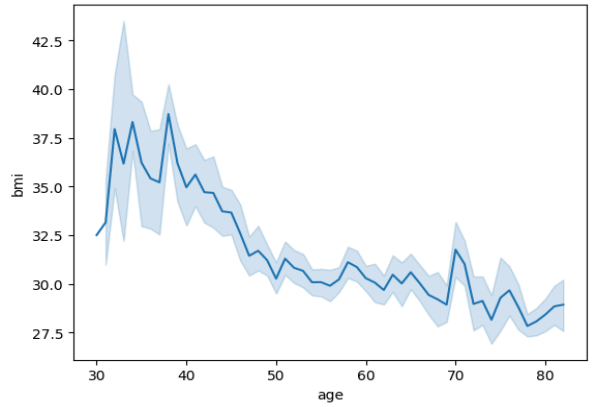

Figure 3. Body Mass Index (BMI) age distribution

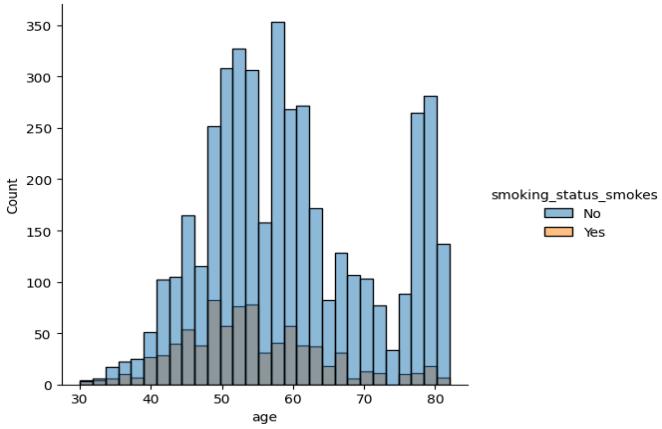

Figure 4. Histogram of smoking status distribution across ages

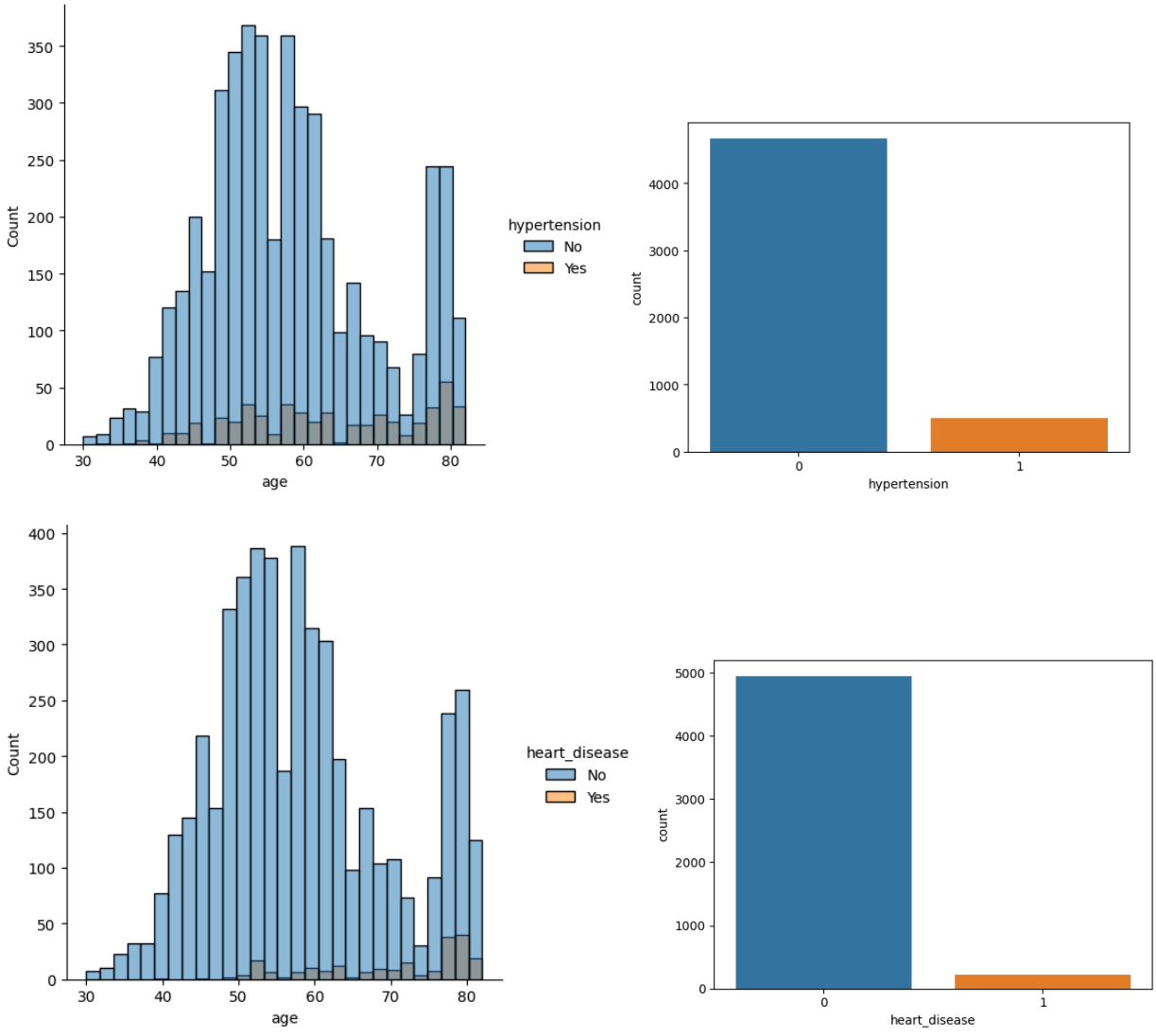

Figure 5. Age distribution histograms and total count bar graphs of heart disease and hypertension

Table 2. Hypertension and Heart Disease diagnosis and Stroke occurrence count

Hypertension	Heart Disease	Stroke	Count
0	0	0	2333
0	0	1	2154
1	0	1	365
0	1	1	165
1	0	0	101
0	1	1	30
0	1	0	19
1	1	0	2

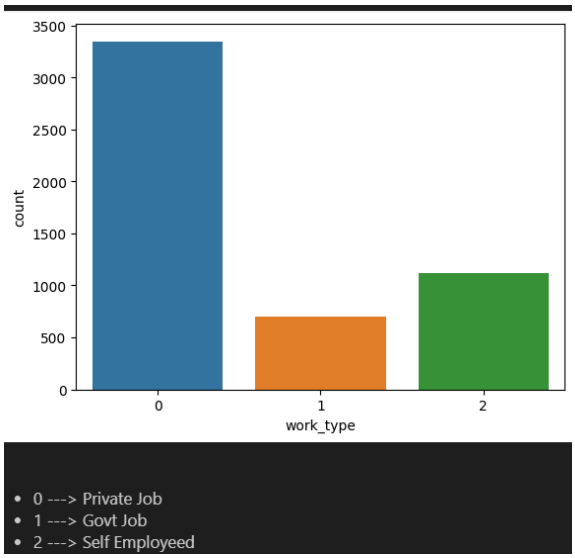

Figure 6. Employment status count, categories converted to int representation

Categorical data converted into numerical format using label encoding in order to be used for model training.

After the review of the dataset features, it was determined that no attributes exhibited any abnormal outliers. However, empty unnamed columns were found and dropped from the dataset (See Figure 7).

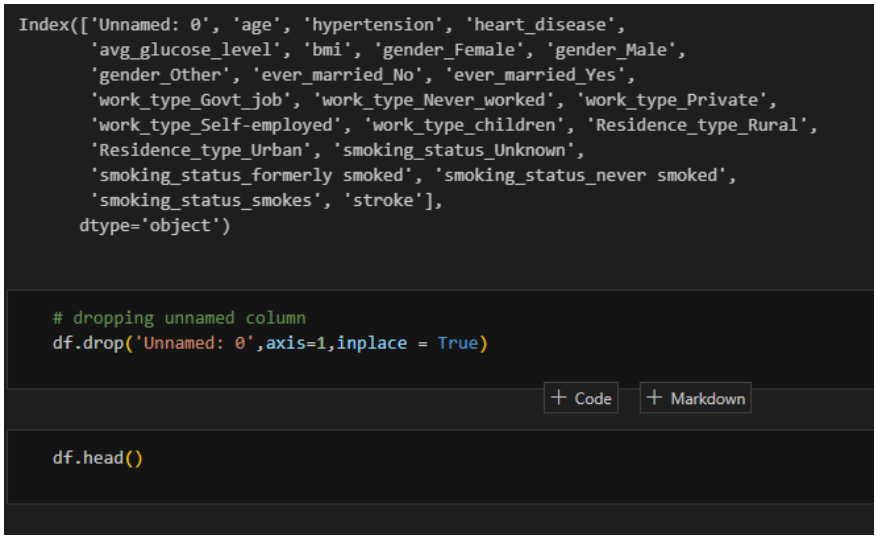

Figure 7. Unnamed columns in dataset removed.

MODELLING

In order to conduct predictive classification on the dataset, Random Forest Regression from Scikit Learn and CatBoost were utilized. Default Parameters were used unless specified otherwise.

Random Forest Regression

Random Forest Regression is constructed with a multitude of random decisions during the training phase and the respective outputs, combined to form a model that makes more reliable predictions. Each decision tree is built on a different subset of the data, and may vary in terms of features and samples used, thus introducing an element of randomness that mitigates overfitting and increases the model's generation capability. (N. Beheshti 2022; T.C. Reader 2022)

CatBoost

CatBoost is a high-performance gradient boosting framework designed specifically for categorical data and tabular datasets. It stands out for its ability to handle categorical features naturally, without the need for manual preprocessing like one-hot encoding. This therefore prevents information loss and increased dimensionality. CatBoost employs a combination of ordered unbiased boosting and oblivious trees, optimizing the learning process by adapting to the inherent characteristics of the data. (L. Prokhorenkova et al. 2017)

RESULTS

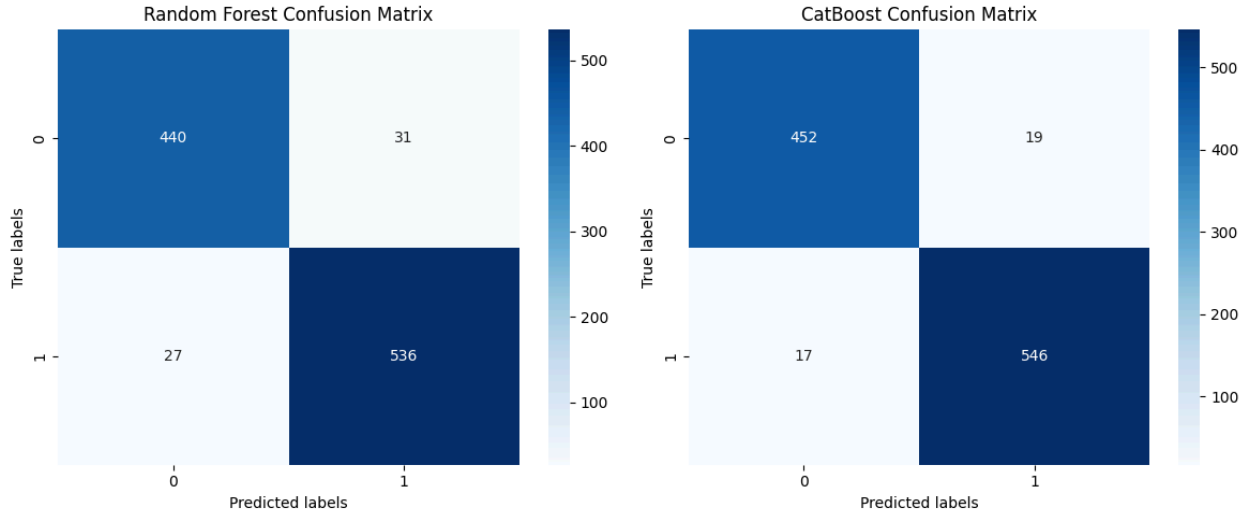

Figure 8. Confusion Matrices of Random Forest Regression and CatBoost Models

Table 3. Accuracy, Precision, Recall and ROC AUC results for Random Forest and CatBoost

Model	Accuracy	Precision	Recall	ROC AUC
Random Forest	0.943907	0.945326	0.952043	0.943113
CatBoost	0.965184	0.966372	0.969805	0.964732

The provided table showcases the performance metrics of two predictive models, Random Forest and CatBoost, in the context of predicting heart strokes. Each model's accuracy, precision, recall, and ROC AUC score are presented for evaluation.

Random Forest Regression

The Random Forest model demonstrates strong overall performance with an accuracy of 94.39%. This indicates that the model correctly predicts stroke occurrence in nearly 94.39% of cases. Precision, which measures the proportion of true positive predictions among all positive predictions, is also high at 94.53%. This implies that when the model predicts a stroke, it is accurate more than 94.53% of the time. The recall score of 95.20% is indicative of the model's ability to capture a substantial portion of actual stroke cases. A high recall is vital in the medical domain, as it means that the model identifies a large proportion of true positives, minimizing the risk of missing critical cases. The ROC AUC score of 94.31% suggests that the Random Forest model performs well in distinguishing between positive and negative instances, with a strong trade-off between true positive rate and false positive rate.

CatBoost

On the other hand, the CatBoost model exhibits even better performance across the board. With an accuracy of 96.52%, it outperforms the Random Forest model in overall prediction accuracy. The precision score of 96.64% further emphasizes the model's proficiency in making precise positive predictions. The recall score of 96.98% indicates the model's exceptional ability to capture a vast majority of actual stroke cases, showcasing the model's high sensitivity. Lastly, the ROC AUC score of 96.47% signifies the model's strong ability to discriminate between different classes, making it a robust choice for prediction tasks.

DISCUSSION

Comparing the two models, CatBoost outperforms Random Forest in all metrics, demonstrating its superior predictive capabilities. The higher accuracy, precision, recall, and ROC AUC scores collectively indicate that the CatBoost model is more accurate, sensitive, and specific in predicting stroke occurrences. These results suggest that CatBoost could be a valuable choice for predicting heart strokes, offering potential improvements in early detection, risk assessment, and patient care in the context of cardiovascular health.

CREDITS

Dataset provided by:

C. Aomahaag, "Stroke Dataset 2023," Kaggle, https://www.kaggle.com/datasets/lopamarkin/stroke-dataset-2023 (accessed Aug. 9, 2023).

REFERENCES

[1] N. Beheshti, "Random Forest Regression," *Medium*, Mar. 02, 2022. https://towardsdatascience.com/random-forest-regression-5f605132d19d

[2] T. C. Reader, "Random Forest Regression Explained with Implementation in Python," *Medium*, Dec. 02, 2021. https://medium.com/@theclickreader/random-forest-regression-explained-with-implementation-in-python-3dad88caf165

[3]L. Prokhorenkova, G. Gusev, A. Vorobev, A. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features." Available: https://proceedings.neurips.cc/paper_files/paper/2018/file/14491b756b3a51daac41c24863285549-Paper.pdf

[4]J. L. Rodgers *et al.*, "Cardiovascular Risks Associated with Gender and Aging," *Journal of Cardiovascular Development and Disease*, vol. 6, no. 2, p. 19, Apr. 2019, doi: https://doi.org/10.3390/jcdd6020019.

[5]D. M. Lloyd-Jones, J. C. Evans, and D. Levy, "Hypertension in Adults Across the Age Spectrum," *JAMA*, vol. 294, no. 4, p. 466, Jul. 2005, doi: https://doi.org/10.1001/jama.294.4.466.

[6]Centers for Disease Control and Prevention, “CDC - fact sheet - current cigarette smoking among adults in the united states - smoking & tobacco use,” *Smoking and Tobacco Use*, 2020. https://www.cdc.gov/tobacco/data_statistics/fact_sheets/adult_data/cig_smoking/index.htm