EXECUTIVE SUMMARY

DATA SCIENCE PROJECT ON
STROKE PREDICTION

made by Guldanika Osmonova
osmonova.daniko@gmail.com

Project Overview

This project aims to develop a predictive model for assessing stroke risk using machine learning techniques. By analyzing patient data such as age, gender, medical conditions, and lifestyle factors, the goal is to provide a tool for early detection of stroke, enabling healthcare professionals to implement timely preventive measures. Stroke is a leading cause of death and disability worldwide, and early intervention can significantly reduce its impact.

Problem

The key challenge in stroke prediction lies in the imbalanced nature of the dataset. The dataset used contains 5,110 patient records with 12 key attributes, but stroke cases are relatively rare compared to non-stroke cases. This imbalance leads to models that may be biased towards predicting the majority class (non-stroke), potentially overlooking stroke cases. Additionally, certain features, such as lifestyle factors and medical conditions, may not fully capture the complexity of stroke risk.

Methodology

Data Preprocessing:

Missing values were handled using median imputation, and class imbalance was addressed using SMOTE to generate synthetic samples for the minority class (stroke cases).

Feature Engineering:

Relevant features were selected and transformed, including encoding categorical variables and identifying outliers.

Model Development:

Algorithms tested include Logistic Regression, Random Forest, XGBoost, and CatBoost, with an emphasis on recall to minimize false negatives, critical for healthcare applications.

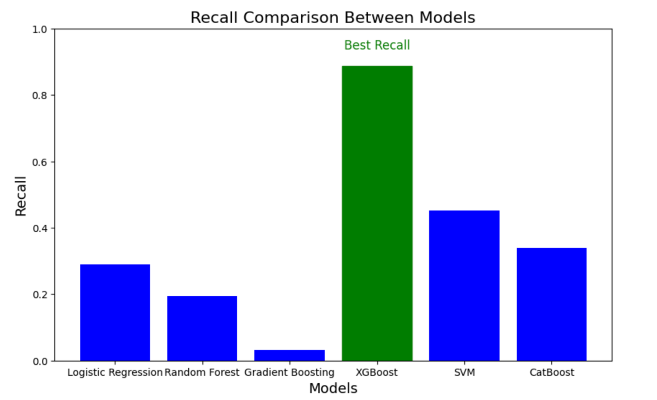
RESULTS

XGBoost: The top-performing model in terms of recall (0.8871), though it had lower accuracy (93.97%), indicating a bias towards the majority class.

CatBoost: Showed more balanced performance with recall of 33.87% and ROC-AUC of 0.7796.

Logistic Regression: Provided a recall of 75.81% but lower accuracy (72.69%), suitable for situations prioritizing recall over precision.

SVM & Random Forest: Moderate recall but challenges in achieving high sensitivity for stroke detection.



Despite high accuracy in some models, the challenge of detecting stroke cases remained, with low recall scores across most models.

SUMMARY

This project has made progress in building a stroke prediction model, but significant challenges remain in addressing class imbalance and improving recall for accurate stroke detection. Future work focused on refining the model, enhancing data quality, and exploring advanced techniques will be crucial to developing a robust tool for stroke prevention, ultimately improving patient outcomes through timely intervention.

Future Work

- **Data Enhancement**: Acquire higher-quality data and explore additional features, such as genetic factors or more granular patient information, to improve predictive accuracy.
- **Advanced Techniques**: Implement more sophisticated machine learning techniques, including deep learning models or hybrid approaches, to capture complex relationships in the data.
- **Collaboration**: Work with healthcare professionals to refine the model and incorporate critical insights into patient risk factors.