Stroke Prediction Analysis

A machine learning approach to predicting stroke risk in individuals.

CG by CHRIS GITONGA

Project Overview

This project predicts individual stroke risk using health data. Our process includes EDA, baseline modeling, advanced ML, evaluation, and practical recommendations.

Exploratory Data Analysis

Uncover data patterns and insights.

Baseline Modeling

Establish initial performance benchmarks.

Advanced ML Modeling

Implement sophisticated algorithms.

Evaluation & Recommendations

Assess model efficacy and provide actionable advice.

Business and Data Understanding

Stroke is a major health threat. Early prediction enables preventative care. Our dataset includes age, hypertension, heart disease, BMI, and glucose levels.

Why Stroke Prediction?

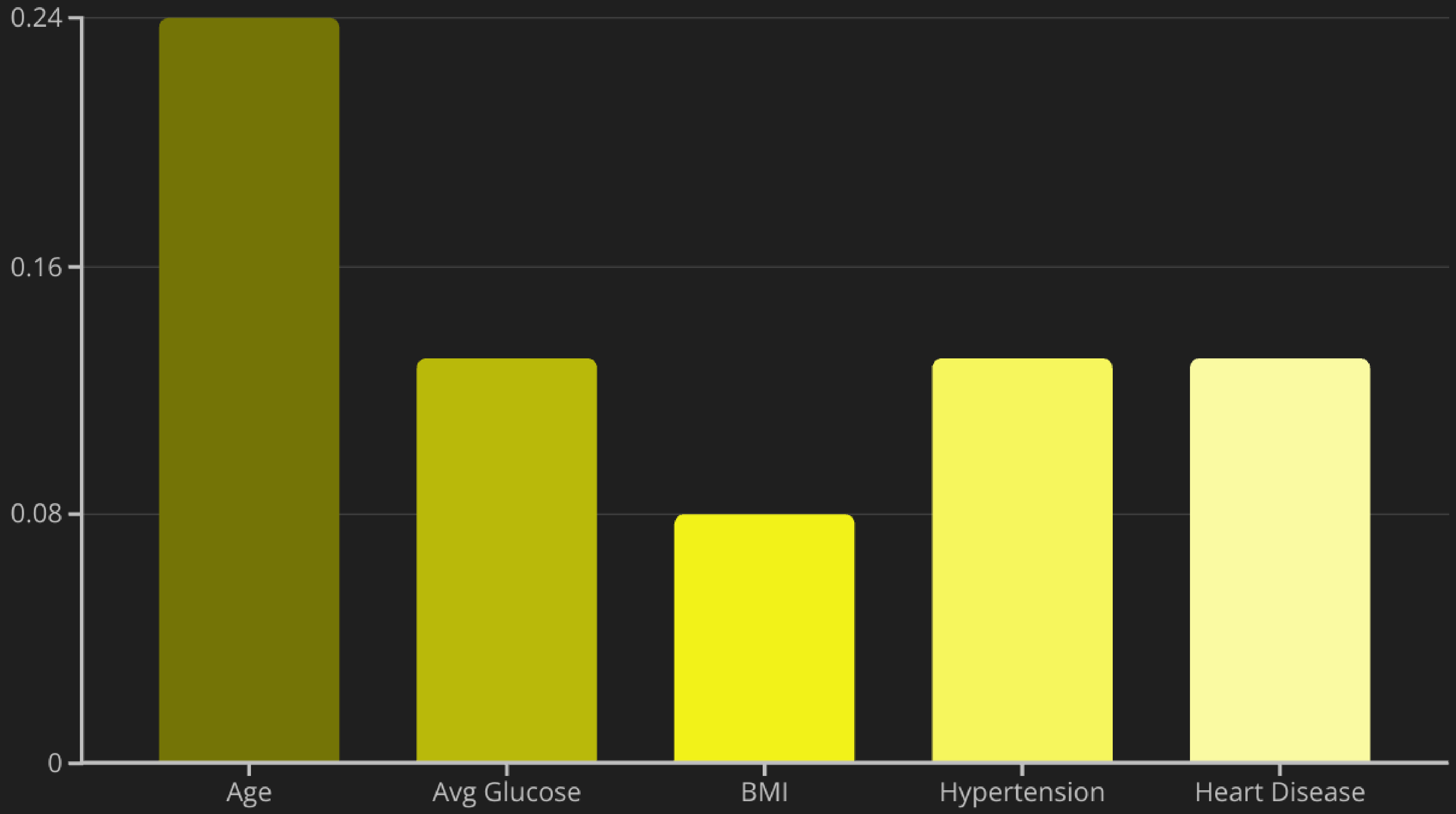
- Leading cause of death.
- Major disability source.
- Enables early intervention.

Key Data Features

- Age, Gender, BMI.
- Hypertension, Heart Disease.
- Average Glucose Level.

Exploratory Data Analysis

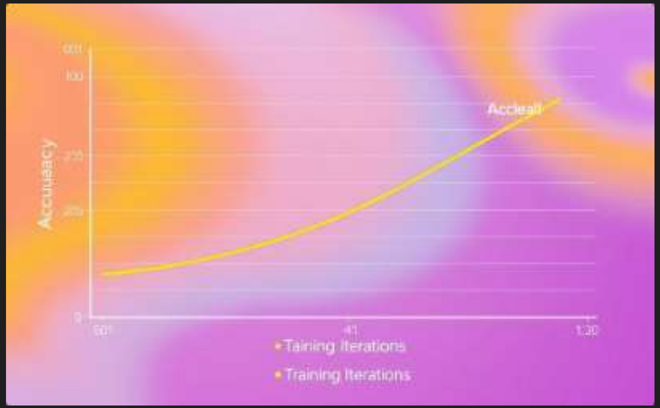
EDA revealed strong links between age, average glucose, and stroke. The dataset shows significant class imbalance, with few stroke cases.



This chart illustrates the correlation of key features with stroke occurrence, highlighting the strongest relationships.

Baseline Model

Our initial logistic regression model, without oversampling or tuning, achieved high accuracy. However, it failed to identify actual stroke cases (0% recall for stroke class).

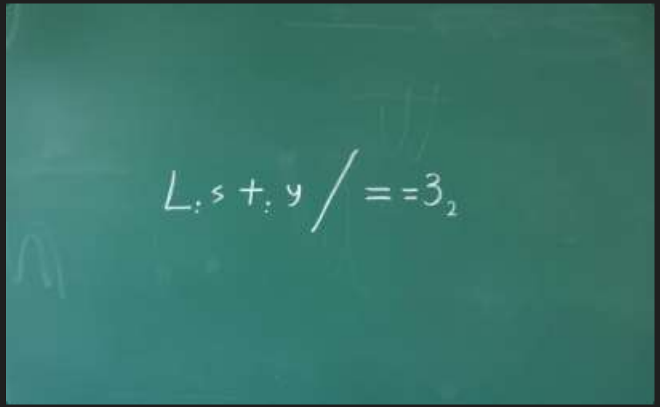


High Accuracy

Overall model correctness.

Zero Stroke Recall

No stroke cases identified.



Logistic Regression

Initial modeling approach.

Improved Models with SMOTE

Applying SMOTE (Synthetic Minority Over-sampling Technique) significantly improved stroke case recall. Both logistic regression and random forest models benefited.

SMOTE Application

Addressed class imbalance.

Recall Improvement

Higher stroke case identification.

Model Performance

Better for both models.

Hyperparameter Tuning

Grid search on logistic regression optimized parameters: C=0.01, penalty='l1'. This boosted the F1 score for stroke to 0.22, with recall at 0.84.

Grid Search

Systematic parameter exploration.

Optimal Parameters

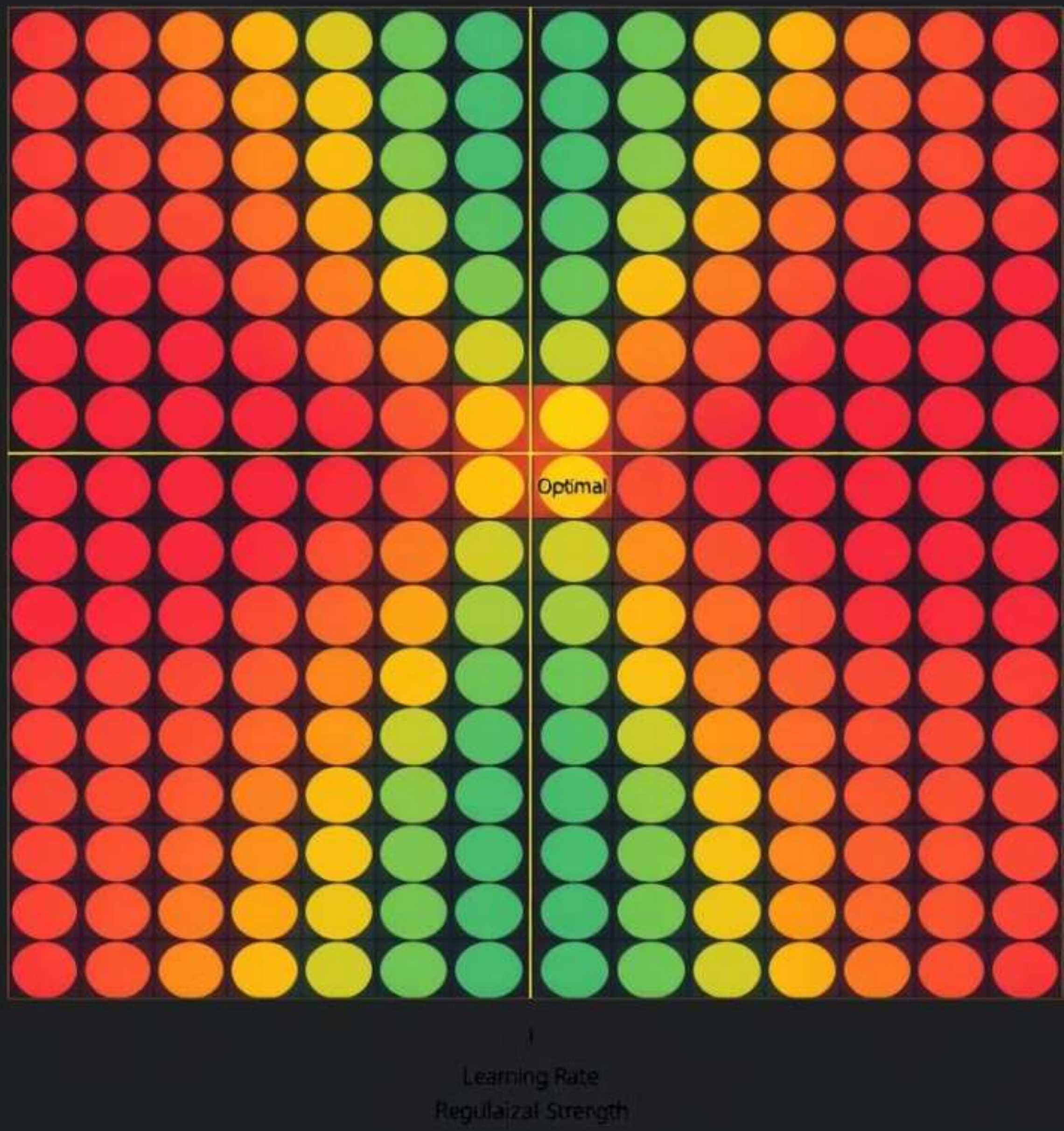
C=0.01, penalty='l1'.

F1 Score Boost

Increased to 0.22.

High Recall

Achieved 0.84.



Final Model Evaluation

The final model significantly improved stroke recall on test data. Training Recall (Stroke): 0.81 | Test Recall (Stroke): 0.30. This trade-off is acceptable for medical screening.

0.81

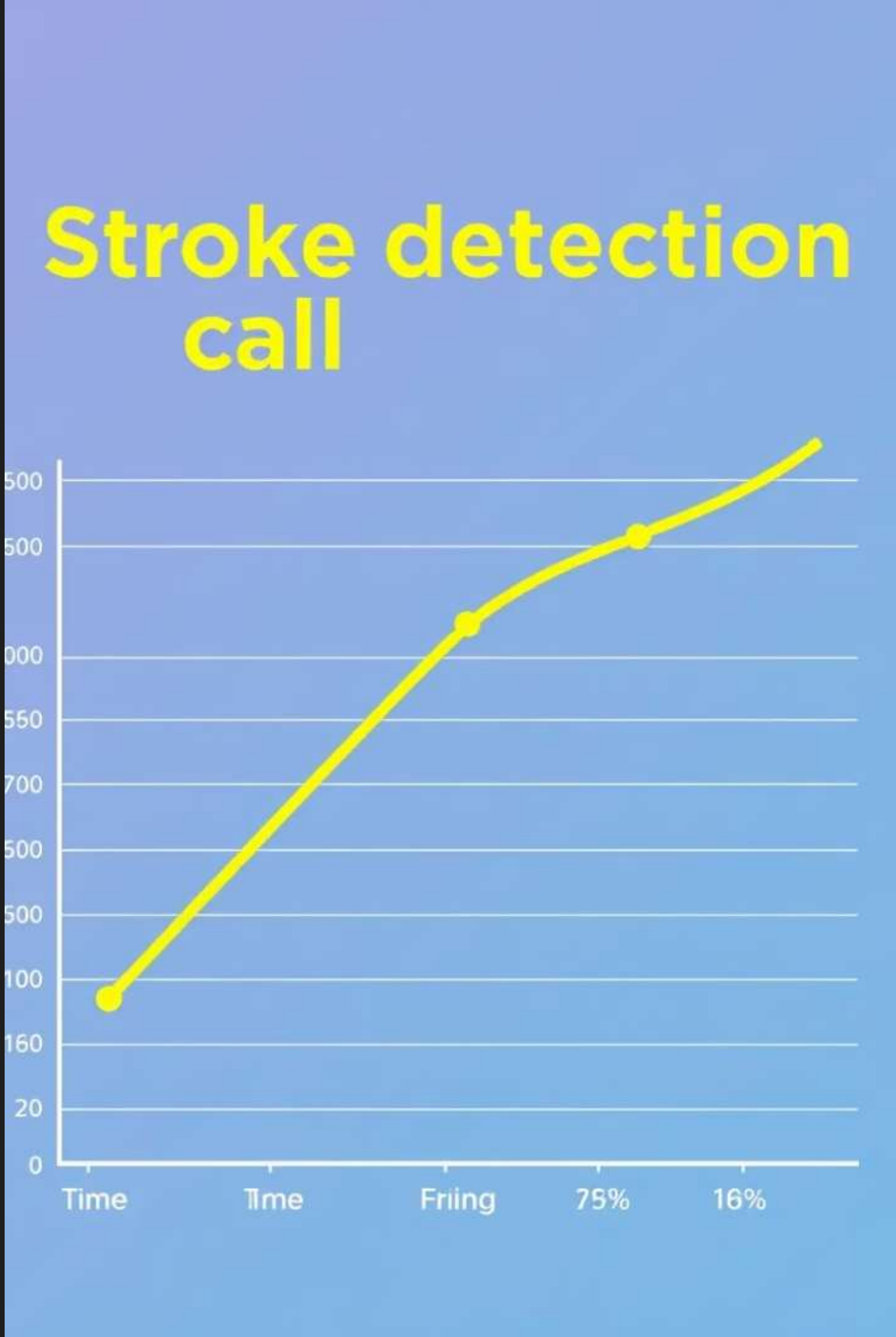
Training Recall (Stroke)

Model's ability to identify stroke cases during training.

0.30

Test Recall (Stroke)

Model's ability to identify stroke cases on unseen data.



Business Recommendations

Utilize this model for high-risk patient flagging. Integrate it into healthcare systems. Prioritize recall to minimize missed stroke risks. Educate providers on prediction interpretation.

Flag High-Risk Patients

Identify individuals needing further screening.

Integrate into Systems

Embed into healthcare decision support.

Focus on Recall

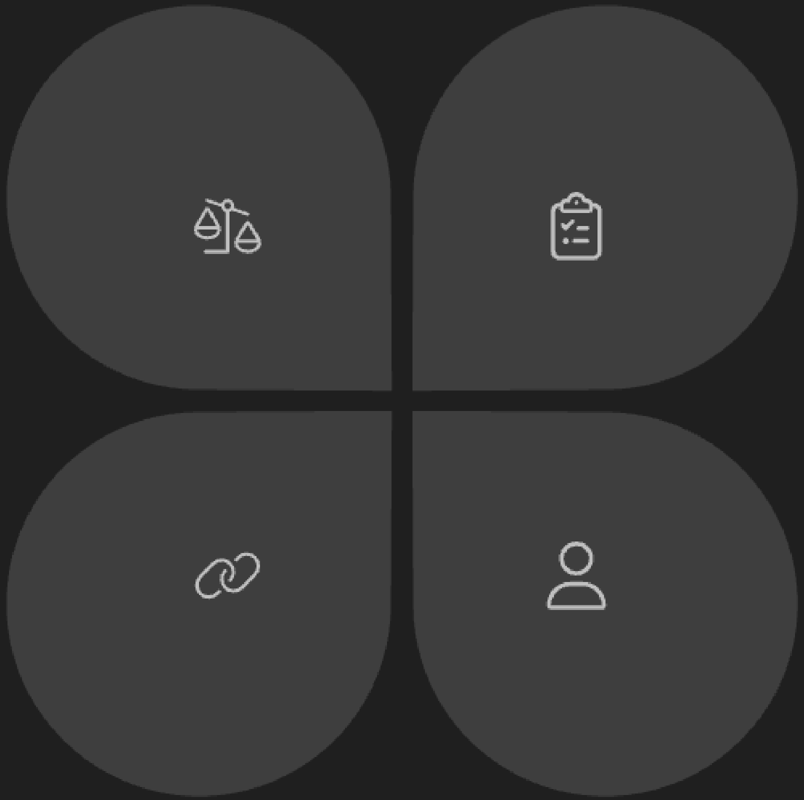
Minimize missed stroke cases.

Educate Providers

Ensure proper interpretation of predictions.

Limitations

Class imbalance affected precision. Limited features, missing lifestyle and genetic data. Small stroke case numbers impact generalizability. Feature interactions need more exploration. Performance may vary across populations.



Class Imbalance

Limited precision for stroke class.

Limited Features

Missing lifestyle and genetic data.

Feature Interactions

Not fully explored.

Small Stroke Cases

Impacts generalizability across populations.

Next steps

- Conduct feature engineering to uncover more predictive interactions

- Use threshold tuning to optimize recall vs. precision

- Try ensemble and cost-sensitive models

- Collect more diverse and longitudinal data

THANK YOU

Thank you for your time and attention. Questions and feedback are welcome!

Chris Gitonga
0714525746
cmgitonga20@gmail .com