The Accuracy Paradox

A Case Study in Building a Stroke Prediction Model That Actually Saves Lives.

A Case Study in Building a Stroke Prediction Model Full presentation.

The Mission: Build a Screening Tool, Not Just a Prediction Model

- Strokes are a leading cause of death globally. Early identification of high-risk patients is critical for prevention.
- Our goal is to build a machine learning model that can serve as a *useful clinical screening tool.*
- In this context, a 'useful' model is one that excels at a specific task: **correctly identifying patients who are genuinely at risk.**

Key Metric Spotlight

Recall

Recall answers the question: 'Of all the patients who actually had a stroke, what percentage did our model correctly identify?"

Why it matters: A missed diagnosis (a False Negative) isn't a statistical error; it's a potential life lost. We must prioritize finding at-risk patients above all else."

The Foundation: Data Source and Preparation

The Dataset

- **Source**: 'Stroke Prediction Dataset' from Kaggle.
- **Size:** 5,000+ patient records with medical and demographic features.
- **Target Variable:** `stroke` (1 for Yes, 0 for No).

The Cleaning Strategy

Imputing Missing Values

The `bmi` feature had 201 missing values. These were imputed with the median value to avoid distortion from outliers.

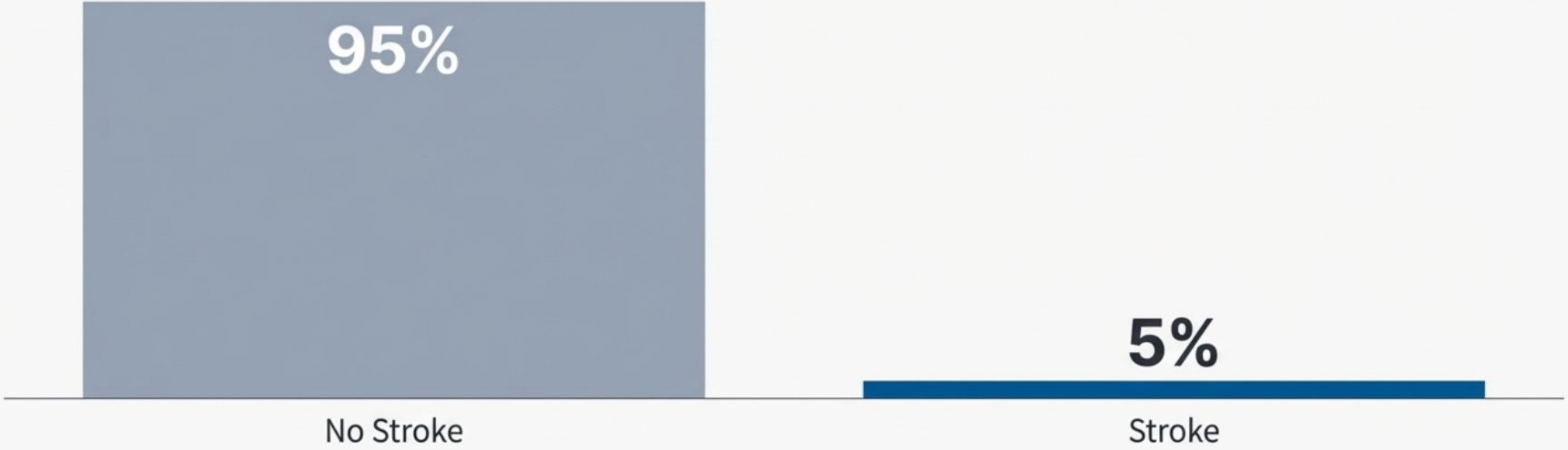
Encoding for Machines

Categorical features (e.g., `work_type`, `smoking_status`) were converted to numerical format using One-Hot Encoding.

Removing Irrelevance

The `id` column was dropped as it provides no predictive value.

The Hidden Antagonist: A Severe Class Imbalance



4,861 cases of ‘No Stroke’ vs. 249 cases of ‘Stroke.’

The Danger Explained

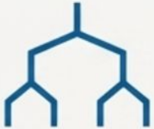
This imbalance presents a critical challenge. A naive model can achieve 95% accuracy by simply predicting ‘No Stroke’ for every single patient. Such a model is technically accurate, but **100% useless** as a clinical screening tool. It would fail to identify anyone at risk. Our entire modeling strategy must be designed to overcome this trap.

A Strategy to Confront the Imbalance

To find the best solution, we trained and compared three supervised learning models.

The Contenders

Logistic Regression: A robust and interpretable linear baseline.



Decision Tree: A model designed to capture non-linear decision rules.

Random Forest: A powerful ensemble method known for high performance.

The Critical Tactic

```
class_weight='balanced'
```

What this does: This forces the algorithm to pay significantly more attention to the minority class (the 'Stroke' cases). It penalizes the model more heavily for missing a high-risk patient than for misclassifying a healthy one.

The Performance Face-Off

Model	Accuracy	Recall (Finds At-Risk Patients)
Random Forest	**95%**	**0%**
Decision Tree	91%	12%
Logistic Regression	75%	**80%**

The results reveal a sharp contrast. The model with the highest accuracy is the worst performer on the metric that truly matters.

The Reveal: The Paradox of Accuracy

Random Forest

95% ACCURACY

0% RECALL

Verdict: A DANGEROUS FAILURE

Logistic Regression

75% ACCURACY

80% RECALL

Verdict: A USEFUL SCREENING TOOL

Building a Model Ready for Clinical Use

A model is only useful if it can reliably predict on new, single-patient data. This requires a robust prediction function that ensures the input structure always matches what the model was trained on.

1 Input

New patient data (age, bmi, etc.)

2 Preprocess

Apply One-Hot Encoding

CRITICAL STEP: Align Features

```
new_data.reindex
(columns=training_columns,
fill_value=0)
```

This command forces the new data to have the exact same columns as the original training data. It prevents errors if a new patient's categorical value wasn't present, ensuring prediction consistency.

4 Predict

Use the trained Logistic Regression model.

The Screening Tool in Action: A Patient Simulation

Input: High-Risk Patient Profile

Age: 57

Gender: Male

Avg. Glucose Level: 120

Smoking Status: ‘smoked’

Model Output

Stroke Probability: 44.35%

Prediction (at default 0.5 threshold): Low Risk

The default prediction is ‘Low Risk,’ but a **44% probability of stroke** is a **significant clinical flag** that demands further medical review. This demonstrates the model’s value.

It successfully identifies a high-risk individual, prompting a doctor to investigate further. This is the goal of a screening tool.

The Ultimate Lesson

“Accuracy is a dangerous and misleading metric when the cost of a false negative is a human life.”

The choice of an evaluation metric should never be automatic. It must be driven by the real-world context and the specific consequences of the model’s errors.

The Path Forward: How to Build an Even Better Model

While our Logistic Regression model is effective, there are clear paths to improving performance further.

Advanced Data Balancing (SMOTE)

Use the Synthetic Minority Over-sampling Technique to generate more artificial “**Stroke**” data points. This would give more complex models like **Random Forest** more examples to learn from, potentially increasing their recall.

Systematic Threshold Tuning

Instead of using the default 50% probability threshold, we can analyze the trade-offs and **lower it** (e.g., to **20%**). This would flag more patients for review, catching even more at-risk cases at the cost of more false positives—a worthwhile trade-off for screening.

Richer Feature Engineering

Incorporate more specific medical data, such as family history or physical activity frequency, to create stronger predictive signals.

Project link

https://github.com/VTornoreanu/Stroke-Prediction-Analysis

Code
Link to the raw data (kaggle.com)
Presentation as PDF file
Link to this video presentation

Thank You