Predicting Stroke Risk Using Patient Health Data

Data Science Capstone Two Project
Hector Sanchez

Problem Statement

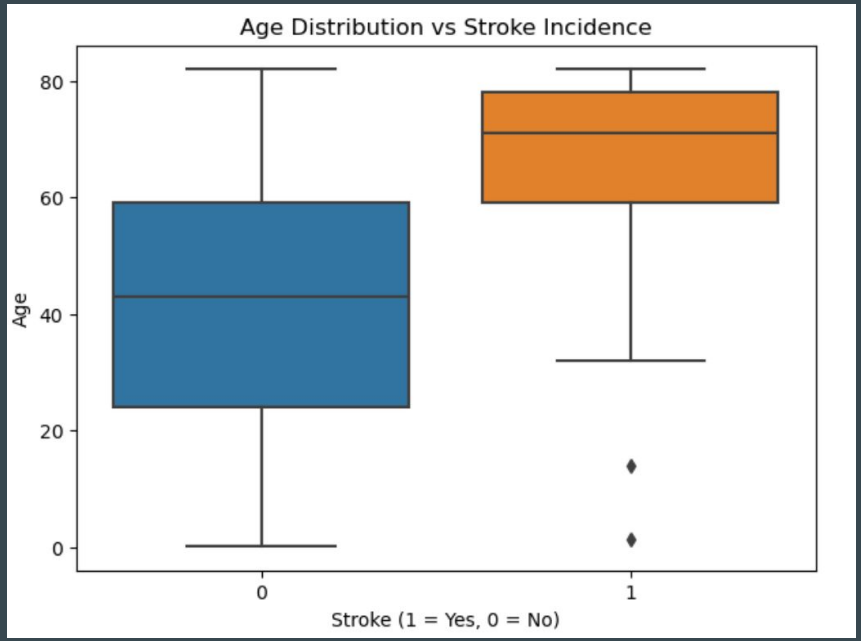
- Stroke is a leading cause of death and long-term disability worldwide
- Predicting the likelihood of stroke can allow for early intervention and treatment, ultimately improving patient outcomes
- The goal of this project is to leverage patient data to accurately predict stroke risk and enable early intervention.

Dataset Overview

- Source: Stroke Prediction Dataset
- Key Features Included:
  - Demographic: age, gender
  - Health Conditions: hypertension, heart_disease, bmi
  - Lifestyle Factors: smoking_status
  - Target: stroke
- This dataset contains essential health and lifestyle information used to predict the likelihood of stroke.

Key Data Insights (EDA)

- Initial analysis showed that age, BMI, and glucose levels have the strongest relationships with stroke risk.
- Age is a key factor, with older individuals showing a higher risk of stroke.



Data Preprocessing

- Data preprocessing included handling missing values, scaling features like bmi and glucose levels, and encoding categorical variables for better model performance
  - Handled missing values for bmi and smoking_status
  - One-hot encoded categorical variables such as gender and smoking_status
  - Standardized numerical features (ex: age, bmi) to ensure models were trained on similar feature scales
  - The dataset was split into training (80%) and testing (20%)

Feature Selection

- Key features selected for the model included age, BMI, glucose levels, hypertension.
- Our age distribution suggested that stroke occurrence is more frequent in older populations as indicated by the histogram and box plot
- Our correlation matrix displayed moderate correlations between age, glucose levels, and stroke.
- Our pair plot emphasized the combined effects of multiple risk factors.
    - Suggests that interactions between features like hypertension and glucose level with age in stroke prediction

Model Selection

- Tested the following 4 models:
  - Logistic Regression
  - Random Forest
  - Gradient Boosting
  - XGBoost
- Hyperparameter Tuning
  - Optimized hyperparameters using GridSearchCV for XGBoost
- XGBoost showed the best performance, making it the ideal choice for predicting stroke risk.

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.695	0.147	0.839	0.250
Random Forest	0.939	0.000	0.000	0.000
Gradient Boosting	0.938	0.000	0.000	0.000
XGBoost	0.917	0.220	0.145	0.175
Tuned XGBoost	0.800	0.186	0.677	0.292

Final Model Performance

- The final model achieved:
  - Accuracy: 80% (The percentage of correct predictions)
  - Precision: 18.5% (Indicates how many predicted strokes are true positives)
  - Recall: 67.7% (Measure the ability to correctly identify all actual stroke cases)
  - F1: 29.2% (Balance between precision and recall)
- Tuned XGBoost offers the best balance between precision, recall, and F1 score, despite having a lower accuracy.
  - This balance is critical in handling the imbalance of stroke prediction.
  - Recall is important in identifying stroke cases since it measures the ability to correctly identify all stroke cases.

Recommendations

- Regularly monitor the model in production to ensure it continues to perform effectively.
- Explore additional health-related features, such as cholesterol levels or previous medical history to improve the model's predictive power.
- Target High-Risk patients for early interventions.
- Monitor Glucose levels for better stroke prevention.

Limitations and Future Work

- Dataset is missing information on Cholesterol levels as a lifestyle factor.
- Integrating additional health data would assist in making further improvements.
- In future iterations, incorporating more patient history and lifestyle data could further improve model accuracy.

Conclusion

- Our model successfully identifies high-risk stroke patients with reasonable accuracy.
- The Tuned XGBoost model provides a reliable method for predicting stroke likelihood based on demographic and health data.
- Focusing on recall ensures that the model identifies most stroke cases.
- This model can potentially support healthcare providers in preventing strokes by identifying high-risk individuals for early intervention.

Thank You!