Capstone Two Final Project Report

Title: Predicting Stroke Risk Using Patient Health Data

Introduction

- **Problem Identification**:
  - Stroke is a leading cause of death and long-term disability worldwide. Predicting the likelihood of stroke can enable early intervention and treatment, improving patient outcomes.
  - The goal of this project is to build a machine learning model to predict the likelihood of stroke occurrence based on patient demographic and health data.
- **Objective**:
  - The objective is to develop a predictive model that can accurately classify whether a patient is likely to have a stroke, using data from the stroke prediction dataset.
  - This project explores different models, performs hyperparameter tuning, and identifies the best-performing model based on evaluation metrics.
- **Scope**:
  - The study utilizes various classification models (Logistic Regression, Random Forest, Gradient Boosting, and XGBoost) to compare their performance and select the best model for stroke prediction.

Data Overview

- **Dataset Description**:
  - The dataset used in this project is the **Stroke Prediction Dataset**, which contains features such as age, hypertension status, heart disease history, and BMI. The target variable is a binary outcome indicating whether a patient has experienced a stroke (1 = stroke, 0 = no stroke).
  - Key features include:
    - Demographic: `age`, `gender`
    - Health Conditions: `hypertension`, `heart_disease`, `bmi`
    - Lifestyle Factors: `smoking_status`
    - Target: `stroke`
- **Data Preprocessing**:
  - Handled missing values for `bmi` and `smoking_status`.

- One-hot encoding was performed on categorical variables such as `gender` and `smoking_status`.
- Standardization of numerical features (e.g., `age`, `bmi`) to ensure models were trained on similar feature scales.
- The dataset was split into training (80%) and testing (20%) subsets.

Exploratory Data Analysis (EDA)

- **Key Insights**:
  - **Distribution of Stroke Occurrence**: A severe class imbalance was noted, with far more patients not having strokes than having strokes. This imbalance was addressed during the model building stage.
  - **Correlation Analysis**: Older patients, those with hypertension, and those with a history of heart disease were more likely to have strokes.
- **Visualizations**:

Histograms for Age, BMI, and Average Glucose Level

The first set of visualizations is a series of histograms depicting the distributions of three important numeric features: `age`, `bmi`, and `avg_glucose_level`.

- **Age Distribution**: This histogram shows the distribution of ages in the dataset. The shape of the distribution gives insights into the age group of the individuals in the dataset, indicating how many people fall into specific age ranges.
- **BMI Distribution**: The BMI histogram reveals how body mass index values are distributed. This is important because BMI is a known risk factor for many health issues, including stroke.
- **Average Glucose Level Distribution**: This histogram provides insights into how blood glucose levels are spread across the dataset. High glucose levels can be linked to diabetes, a significant stroke risk factor.

Each histogram helps to visually assess the data for skewness or irregularities in the feature distributions, aiding in later decisions on transformations or feature scaling.

Correlation Heatmap

The correlation heatmap visualizes the relationships between all the features in the dataset, including the target variable (stroke).

- The **heatmap** uses color gradients to indicate the strength and direction of correlations, with darker colors representing higher correlation

coefficients (either positive or negative). This visualization helps identify any multicollinearity issues and reveals which features might have strong relationships with the stroke variable.
- For instance, variables like `age` and `avg_glucose_level` show a notable correlation with stroke incidence, suggesting their predictive potential.

Scatter Plot: BMI vs. Stroke

This scatter plot shows the relationship between BMI and stroke incidence.

- The **scatter plot** compares `bmi` (x-axis) with `stroke` (y-axis), where stroke is represented as a binary variable (0 for no stroke, 1 for stroke). Points are color-coded based on whether the individual had a stroke or not, which allows for easy visual distinction between stroke and non-stroke cases.
- This plot is useful in investigating whether higher BMI values are associated with a higher likelihood of stroke. Although the relationship may not be linear, visual patterns can indicate potential trends or clusters.

Box Plot: Age vs. Stroke

This box plot compares the distribution of `age` for individuals with and without a stroke.

- **Box plots** are effective in showing the spread and central tendency of age across the two categories (stroke = 0, stroke = 1). It allows us to assess whether age plays a significant role in stroke occurrence by showing differences in median, quartiles, and potential outliers between the two groups.
- As age is a significant risk factor for stroke, the box plot highlights whether older individuals in the dataset are more prone to stroke compared to younger individuals.

Pair Plot: Investigating Interaction Effects

The final visualization is a **pair plot** that investigates potential interactions between features such as `age`, `hypertension`, `heart_disease`, `avg_glucose_level`, and `bmi`, with the target variable `stroke`.

- The pair plot provides a matrix of scatter plots for each pair of features, with color-coded points based on stroke occurrence. It helps to identify patterns or interactions between features that might not be obvious when looking at each feature individually.

- For example, this plot can reveal how combinations of risk factors (such as older age combined with high glucose levels or hypertension) correlate with stroke incidence.

Summary of Findings from Visualizations:

- The age distribution suggests that stroke occurrence is more frequent in older populations, as seen in both the histogram and box plot.
- The correlation matrix highlighted moderate correlations between age, glucose levels, and stroke.
- While no direct linear relationship between BMI and stroke is evident in the scatter plot, further analysis could be needed to explore non-linear associations.
- The pair plot emphasizes the combined effects of multiple risk factors, suggesting interactions between variables like hypertension and glucose level with age in stroke prediction.

These visualizations collectively build a clearer understanding of the data and help in shaping the subsequent modeling steps, where these insights guide feature selection and model design.

Modeling Process

- **Data Preparation**:
  - The data was split into training and testing sets to validate model performance.
  - Class imbalance was addressed using `class_weight='balanced'` for certain models and `scale_pos_weight` for XGBoost.
- **Models Implemented**:
  - **Logistic Regression**: Simple baseline model for comparison.
  - **Random Forest**: A robust ensemble model to capture non-linear relationships.
  - **Gradient Boosting**: Another ensemble method focusing on boosting weak learners.
  - **XGBoost**: A highly efficient and accurate model optimized for speed and performance.
- **Hyperparameter Tuning**:
  - Hyperparameters were optimized using GridSearchCV for XGBoost, tuning key parameters such as learning rate, max depth, and number of estimators.

Model Evaluation and Selection

Comparison of Models:

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 Selection:

- After evaluating the models based on accuracy, precision, recall, and F1 score, **Tuned XGBoost** was selected as the best model.
- **Rationale**: Although it doesn't have the highest accuracy, Tuned XGBoost offers the best balance between precision, recall, and F1 score, which is critical in handling the imbalance of stroke prediction, as recall is particularly important in identifying stroke cases.

Model Application and Results

Application of Tuned XGBoost:

- The Tuned XGBoost model was applied to the test dataset to make predictions. It provided satisfactory results, particularly in terms of recall, which helps identify potential stroke patients accurately.

Model Metrics:

- Final model metrics (Accuracy: 0.800, Precision: 0.186, Recall: 0.677, F1 Score: 0.292) highlight that the model balances minimizing false negatives (i.e., failing to detect a stroke) while still maintaining a reasonable precision.

Recommendations and Future Work

Recommendations:

- **Adopt the Tuned XGBoost Model**: Based on the balanced evaluation metrics, especially recall, the Tuned XGBoost model is ideal.
- **Monitor Model Performance**: We should monitor the mode regularly in production, especially for new patient data, to ensure that the model continues to perform effectively.
- **Explore Additional Features**: Additional health-related features, such as cholesterol levels or previous medical history, could improve the model's predictive power.

Future Work:

- We should investigate additional advanced algorithms (e.g., deep learning models) to improve prediction accuracy.
- We should also address the class imbalance issue by experimenting with more sophisticated resampling techniques.

Conclusion

- The **Tuned XGBoost** model provides a reliable method for predicting stroke likelihood based on demographic and health data. By focusing on maximizing recall, the model ensures that most stroke cases are identified, which can support early intervention strategies and improve patient outcomes.
- The project demonstrates the potential of machine learning in healthcare and highlights areas for future development and improvement.