Case Study — Stroke Treatment Analysis

Advancing Patient Care through Predictive Analytics: A Machine Learning Approach

Data Analysis by Irene Liang

Date: Nov 6th, 2023

Summary

The analysis aims to develop a predictive model for identifying patients at high risk of stroke, with the ultimate goal of enhancing treatment strategies and patient care. Utilizing a dataset of 43,401 anonymized patient records, we applied various machine learning algorithms to predict stroke incidence. The dataset comprised demographic features, medical history, and clinical parameters.

Key steps in our analysis included comprehensive data preprocessing to address missing values and encode categorical variables, exploratory data analysis to understand feature distributions and correlations, and the implementation of several machine learning models. We evaluated models based on accuracy, precision, recall, F1-score, and area under the receiver operating characteristic (ROC AUC) curve.

The best-performing model was an ensemble model with Random Forest and XGBoost, achieving an ROC AUC of 99.6%. This model's predictive capability allows for the prioritization of patients for preventative measures and more intensive monitoring. We recommend the deployment of this model in a clinical setting, alongside existing risk assessment protocols, to enhance the prediction and prevention of stroke events.

Further analysis and continuous model refinement are encouraged as additional data becomes available, ensuring that the model evolves in response to new insights and maintains its predictive accuracy.

Introduction

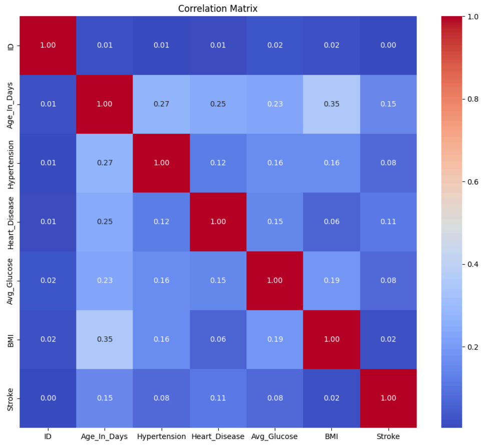
Stroke is a major global health issue, being the second leading cause of death and a major cause of disability. The timely prediction and prevention of stroke can save lives and significantly reduce the burden on patients and healthcare systems. The objective of this project was to analyze a comprehensive dataset collected by Roche Pharmaceuticals, encompassing over 43,000 patient records with diverse variables ranging from demographic details to clinical parameters.

The analysis conducted herein leverages advanced data science techniques, including machine learning algorithms, to unravel patterns and correlations within the dataset that can predict stroke occurrences. The insights derived from this study aim to bolster the pharmaceutical company's research and development efforts by identifying key risk factors and treatment strategies that can be fine-tuned to enhance the efficacy of stroke interventions.

In the following sections, we detail the methodology employed in our analysis, discuss the findings and their implications, and propose recommendations based on the results obtained.

Methodology

Step 1. Data Exploration: We started with an exploratory data analysis (EDA) to understand the data's characteristics, including distributions, missing values, and potential correlations. This process involved visualizing the data with histograms, box plots, and correlation matrix to uncover patterns within the data. The correlation matrix revealed that age, hypertension status, presence of heart disease, average glucose levels, and body mass index (BMI) are the variables most strongly correlated with the incidence of stroke.



Step 2. Data Preprocessing: This phase is critical as the quality of data fed into the models directly impacts their performance and the validity of the predictions.

- Data Cleaning:
  - We began by addressing any invalid entries in our dataset. Specifically, we removed records where the '*Age_In_Days*' was less than or equal to zero, as these do not represent valid age values.
  - Next, we tackled the issue of missing values in the *'BMI'* column. Missing data can skew the analysis and lead to biased predictions if not handled properly. We chose K-Nearest Neighbors (KNN) imputation, a method that estimates the missing value based on the 'k' nearest points in the feature space, ensuring a more accurate and nuanced replacement than simpler techniques such as mean or median imputation.
  - For the *'Smoking_Status'* variable, which also had missing values, we opted to fill in the gaps with a new category labeled *'Unknown'*. This approach prevents the loss of data and acknowledges the absence of information without making unfounded assumptions about the patient's smoking status.
- Encoding Categorical Variables:
  - Our dataset contained several categorical variables, such as *'Gender'*, *'Ever_Married'*, *'Type_Of_Work'*, *'Residence'*, and *'Smoking_Status'*. We converted these into a numerical format using one-hot encoding, which transforms each categorical value into a binary vector. This is crucial for machine learning algorithms that require numerical input.

    - After encoding, we integrated the new binary features into the dataset and removed the original categorical columns, thereby preparing our data for algorithms that cannot inherently handle categorical data.
- Normalization of Numerical Variables:
    - For the numerical variables *'Age_In_Days'*, *'Avg_Glucose'*, and *'BMI'*, we applied standard scaling. This process transforms the data such that each feature has a mean of zero and a standard deviation of one, a necessary step for many machine learning algorithms that are sensitive to the scale of the data, such as support vector machines and k-nearest neighbors.
- Handling Class Imbalance: A crucial issue in our dataset was the imbalance between the classes in the target variable *'Stroke'*. An imbalanced dataset can lead to a model that is biased towards the majority class, resulting in poor classification performance. To address this, we employed the Synthetic Minority Over-sampling Technique (SMOTE). SMOTE generates synthetic examples of the minority class, ensuring that the model is exposed to a more balanced distribution of classes during training.

Step 3. Model Selection: We selected a suite of machine learning algorithms suitable for binary classification tasks. Each model was chosen for its unique strengths and the diversity it brought to the ensemble of methods applied:

- Logistic Regression: A baseline model for binary classification problems.
- Random Forest: An ensemble method that can handle a large number of features and is less likely to overfit than decision trees.
- Gradient Boosting (e.g., XGBoost): An ensemble boosting method that can provide high accuracy if properly tuned.
- Support Vector Machine (SVM): Effective in high-dimensional spaces, which can be the case after encoding categorical variables.

Step 4. Model Training and Evaluation: During the model training and evaluation process, we employed feature selection and model validation to ensure that our predictive models were both accurate and generalizable. The dataset was initially split into an 80-20 ratio for training and testing.

To begin, we addressed the feature scaling using the Standard Scaler, normalizing the feature set so that each feature had a mean of zero and a standard deviation of one. This step is particularly crucial when features have different units or scales and when using algorithms that are sensitive to feature scaling, such as L1 regularization.

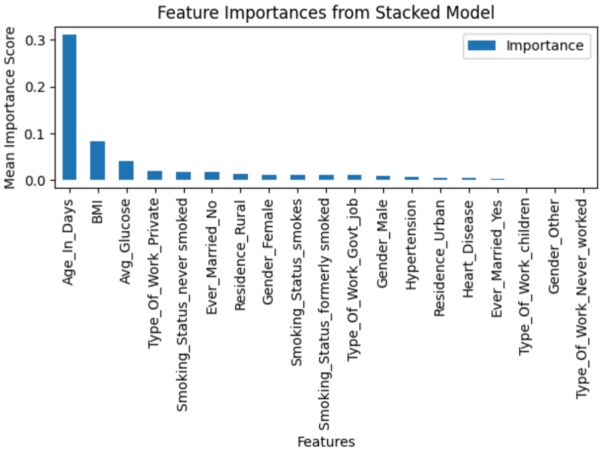
We used L1 regularization, a technique that imposes a penalty on the absolute size of the coefficients, through logistic regression with cross-validation. The primary reason for choosing L1 regularization was its ability to perform feature selection by shrinking some of the model coefficients to zero for less important features. This method is advantageous because it simplifies the model, potentially improving its performance and interpretability by reducing overfitting and focusing on the most impactful features.

After applying L1 regularization, the number of features used by the model was reduced from 21 to 20. The feature that was eliminated provided the least predictive power in the context of stroke risk.

With the selected features, we trained on the machine learning models. Each model was encapsulated within a pipeline that ensured the use of the scaled and selected features for training. The performance of these models was then evaluated using a comprehensive set of metrics, including accuracy, precision, recall, F1-score, and ROC AUC, to determine their effectiveness in predicting stroke outcomes.

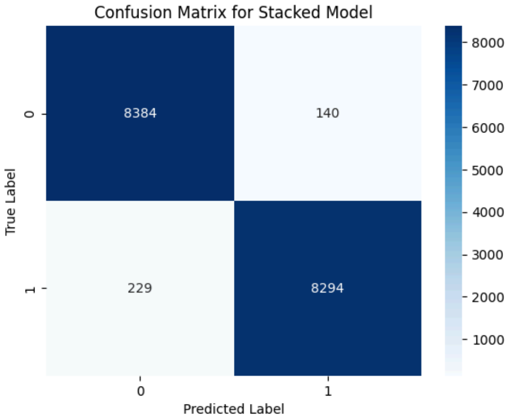
Step 5. Model Selection and Refinement:

- Rationale for Choosing Two Models Over One:
  - Diversity of Perspectives: Random Forest and XGBoost offer different approaches to learning from data. Random Forest employs ensemble learning with decision trees to reduce variance, while XGBoost uses gradient boosting to minimize both bias and variance, improving over time with each iteration.
  - Performance Metrics: Both models exhibited strong performance across all evaluation metrics, suggesting that they capture different aspects of the data's underlying structure.
  - Error Reduction: If one model's errors are random and the other's are systematic, the ensemble might reduce overall error.
  - Operational Robustness: In a real-world setting, having multiple models could provide a fallback if one model's performance deteriorates due to changes in data patterns.
- Hyperparameter Tuning: We conducted hyperparameter tuning to enhance the performance of each model. For Random Forest, parameters like n_estimators, max_depth, min_samples_split, and min_samples_leaf were optimized. For XGBoost, n_estimators, max_depth, learning_rate, and subsample were fine-tuned.
- Ensemble Creation: We then used a stacked ensemble model with the two finely tuned models as base estimators and a Random Forest classifier as the final estimator.
- Feature Importance Analysis: Following the training of the ensemble model, we assessed feature importance to determine which attributes most significantly contribute to predicting stroke. The results indicated that age, BMI, average glucose, and type of work are the top features. This finding is consistent with medical insights and existing knowledge about stroke risk factors, confirming that the model's reasoning is in line with real-world clinical understanding.

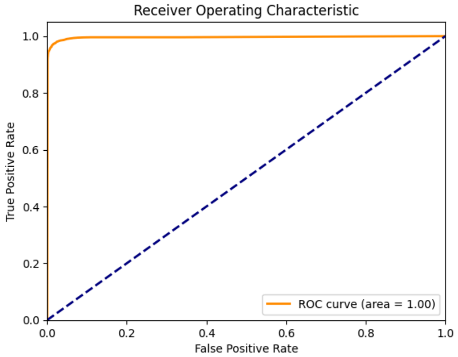


Step 6. Model Evaluation

- Quantitative Analysis: We evaluated the stacked model using the test data. The classification report provided detailed metrics, including precision, recall, and F1-score, allowing us to assess the model's predictive performance comprehensively. The ROC AUC score further quantified the model's ability to discriminate between the classes, indicating a robust predictive capability.
- Qualitative Analysis: The confusion matrix visualized the model's predictions, offering insights into the types of errors made and the balance between sensitivity and specificity. The heatmap representation of the confusion matrix elucidated the true positives, false positives, false negatives, and true negatives, which are crucial for understanding the model's performance in a medical context.



- Receiver Operating Characteristic (ROC) Curve: By plotting the ROC curve and calculating the area under the curve (AUC), we captured the trade-off between true positive rate and false positive rate. The curve's proximity to the top-left corner of the plot, as well as a high AUC value, reflected the model's effectiveness in distinguishing between patients who did and did not experience a stroke.



Step 7. Saving and Loading the Model: We used joblib to save the trained stacked model, scaler, and one-hot encoder to ensure the model can be reliably applied to new data in the future. The loaded model is ready for deployment, enabling quick and accurate stroke risk predictions in a healthcare environment.

Discussion

The creation of an ensemble model combining Random Forest and XGBoost has yielded a tool with notable predictive prowess for stroke risk. The balanced performance across metrics like accuracy, precision, recall, F1-score, and ROC AUC is particularly encouraging, suggesting that the model could serve as a valuable asset in clinical decision-making. The utility of such a model lies in its capacity to sift through complex data and provide actionable predictions.

The model's capacity to highlight the most impactful predictors of stroke—age, BMI, average glucose levels, and type of work—is one of its most significant features. This correlation not only reinforces the current understanding of stroke risk factors but also provides a data-driven foundation for targeting preventative healthcare measures. By identifying patients at higher risk based on these factors, interventions can be more effectively allocated, potentially improving patient outcomes.

Nevertheless, the challenges posed by the initial class imbalance and the subsequent application of SMOTE to mitigate it must be acknowledged. Although SMOTE helps in balancing classes, the introduction of synthetic instances could potentially distort the model's perception of the true underlying data distribution. This aspect of data preparation is nuanced and needs careful handling to ensure that the benefits of a balanced dataset do not come at the expense of predictive reliability.

Moreover, while the ensemble model benefits from its diverse approaches to learning, it is important to recognize the complexity and less transparent decision-making processes that come with such models. The 'black-box' nature of machine learning can sometimes limit the interpretability of the results, which is a crucial consideration in a clinical context where understanding the why behind a prediction can be as important as the prediction itself.

Recommendations

1. Preventive Measures:
   a. Initiate a stroke prevention program that includes educational materials, lifestyle coaching, and support groups for patients identified as high risk.
   b. Partner with nutritionists, physiotherapists, and other specialists to create personalized intervention plans for high-risk patients, focusing on modifiable risk factors such as diet, exercise, and smoking cessation.
2. Clinical Trials:
   a. Use the model to stratify patients for inclusion in trials based on their predicted risk level, potentially improving the efficacy of new stroke prevention treatments.
   b. Design trials to specifically target interventions for high-risk factors identified by the model, such as hypertension management or new anticoagulant therapies.
3. Treatment Protocols:
   a. Update existing stroke treatment protocols to incorporate predictive analytics, ensuring that patients with high risk receive the most aggressive evidence-based interventions.

Potential Next Steps

1. Hyperparameter Optimization: Continual fine-tuning of hyperparameters might yield improved performance. Techniques like randomized search or Bayesian optimization could offer more efficient and potentially more effective alternatives to grid search.
2. Interpretable Machine Learning: Employing models that provide more interpretability, such as generalized additive models (GAMs), to interpret complex models could make the model's decisions more transparent and trustworthy to clinicians.
3. Prospective Validation: Testing the model on independent, prospective datasets would provide insights into its generalizability and robustness across different patient populations and healthcare settings.
4. Integration of Additional Data Sources: Incorporating more comprehensive data, including genetic markers and patient medical history, could capture a fuller picture of stroke risk and refine the model's predictions.
5. Model Updates and Maintenance: As new data become available and as patterns in stroke risk factors evolve, the model should be updated to maintain its accuracy and relevance.