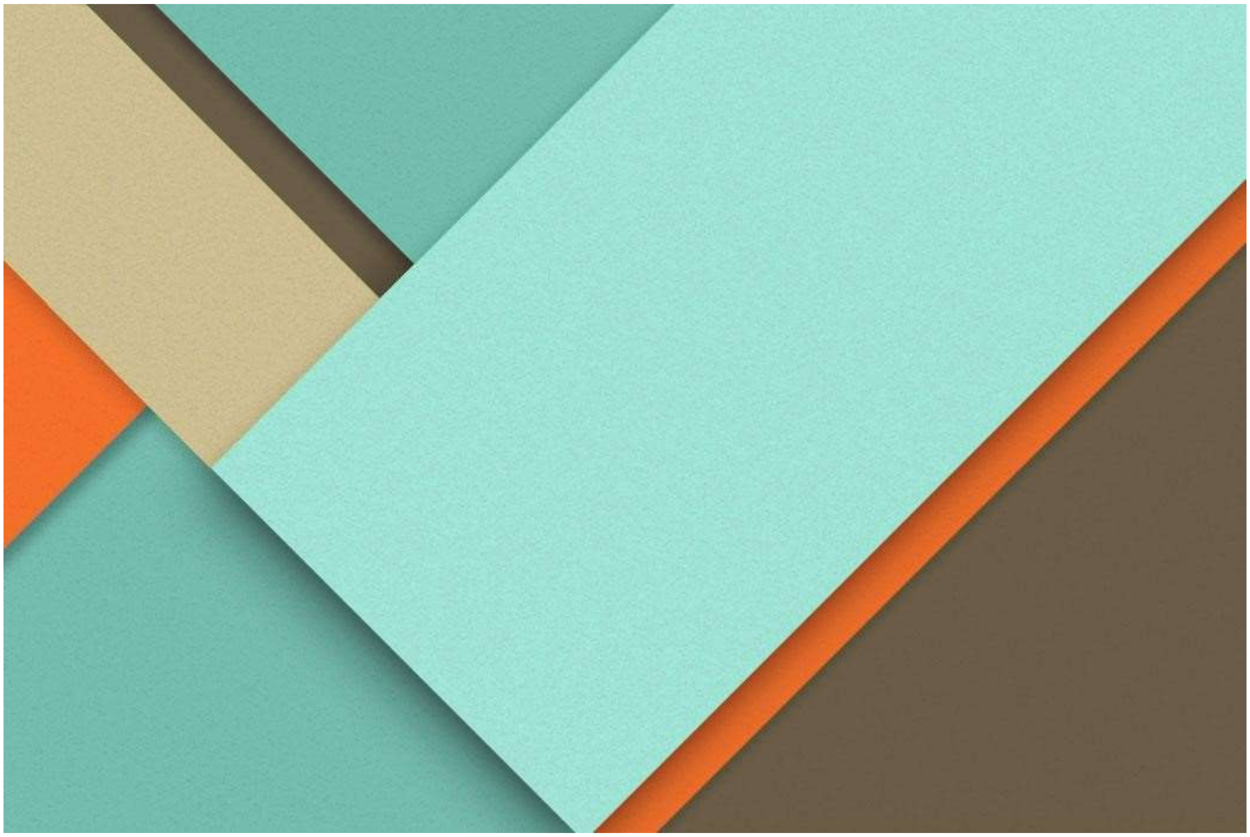


Addressing Class Imbalance in Stroke Prediction Using Machine Learning

09.04.2023

—

Anurag Tendulkar

Email - anurag.mtendulkar@gmail.com

Phone No - +91 9307902673

Aim

The aim of the project is to build a predictive model which can identify patients who are at a risk of having a stroke from the given features.

Code

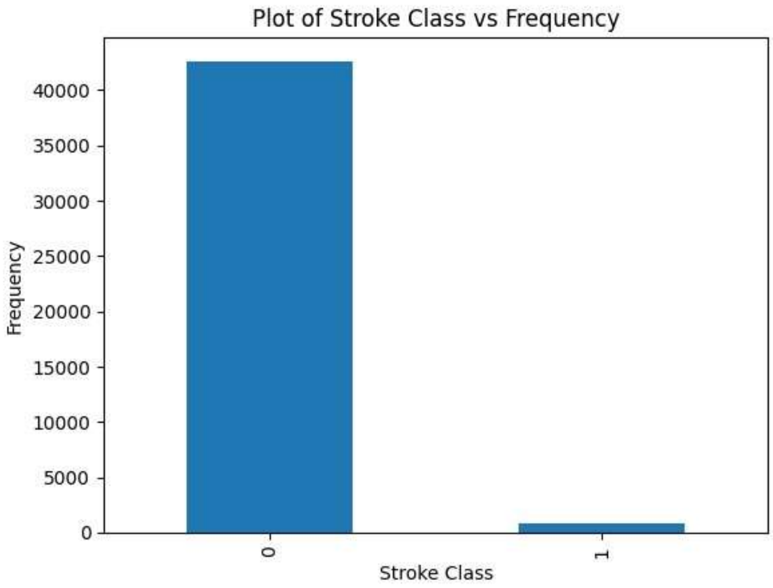
Google colab link to the code.

Load the dataset - 'stroke.csv' into google colab before executing the code

https://colab.research.google.com/drive/1x4xmu-GmgpAFW6QARNj6i0DgplOL3svP

Dataset Description

1. The dataset is highly imbalanced where the ratio of minority class (population having strokes) to majority class (population not having strokes) is 1: 54.



2. The dataset includes categorical features and columns with missing values, which need to be cleaned and pre-processed. Additionally, there are features with binary data, as well as numeric data that requires standardization.
3. Shape of dataset - (43,400, 15)

Pre-processing

Nominal Categorical columns such as gender, occupation, married and residence were expanded into more columns based on their values and were transferred into binary data.

Ordinal Categorical columns such as smoking status were converted into numeric data by a dictionary mapping.

Columns such as metric_2 and smoking status had null values. Doping rows having null values caused loss of positive stroke data by 20%. Hence these columns were dropped to preserve data.

Other numeric data was standardized to zero mean and unit variance.

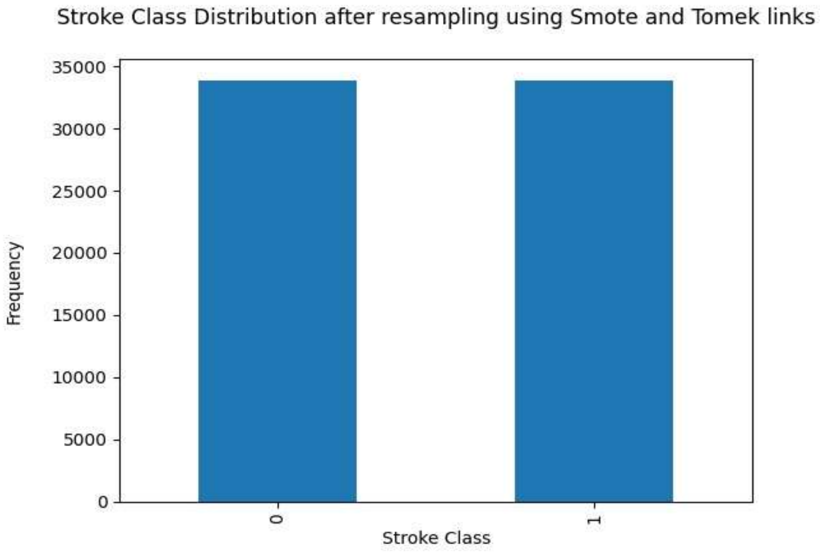
Resampling Techniques Used

Combination of Oversampling and Under Sampling using SMOTE (Synthetic Minority Oversampling Technique) and Tomek Links.

Unlike Random Oversampling, where duplicates of data are generated, SMOTE is used to address class imbalance by generating synthetic minority class examples, improving model performance, mitigating bias, and enhancing generalization.

Random Oversampling was avoided as it tends to lead to overfitting of the training dataset.

Random Under sampling was avoided as critical information can be lost.

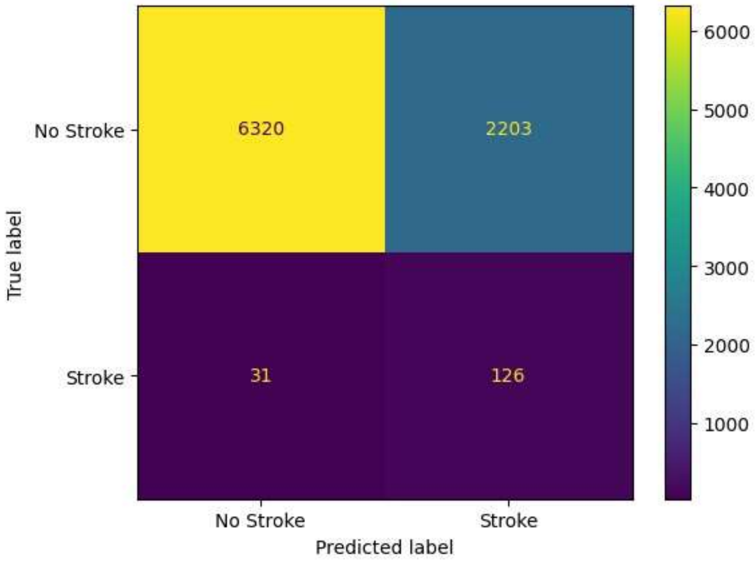


Models Chosen

I.Logistic Regression (Generalized Linear Model)

Logistic regression is often used in binary classification problems and maps the input to the probability of belonging to a particular class. It is interpretable and easy to implement. It does not make any assumptions of distribution of classes in feature space. This is the reason why I choose it over other GLM's like Poisson or Linear Regression

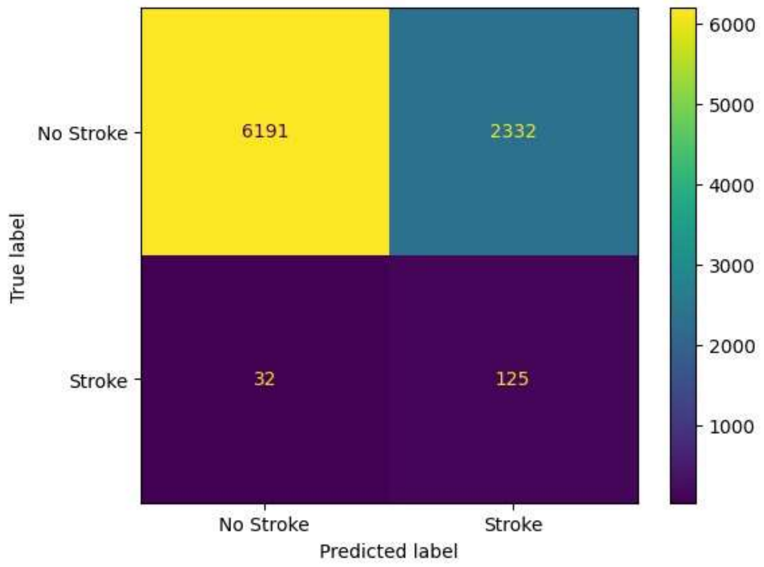
Disadvantages - The model led to overfitting of training data and made a lot of misclassifications. It could not give good precision and recall simultaneously. Detail metrics in the Results section.



II. Random Forest

Tree based algorithms often work well with imbalanced data. Random Forest performs better than single decision trees and reduces overfitting. In addition to bootstrapping, the algorithm draws random subsets of features for training the individual trees which leads to better performance.

Disadvantages - Like Logistic Regression random forest overfitted the training data. Regularization terms such as estimators, class weight, max depth had to be added. Despite regularization the model could not handle precision and recall of the minority class and made many misclassifications

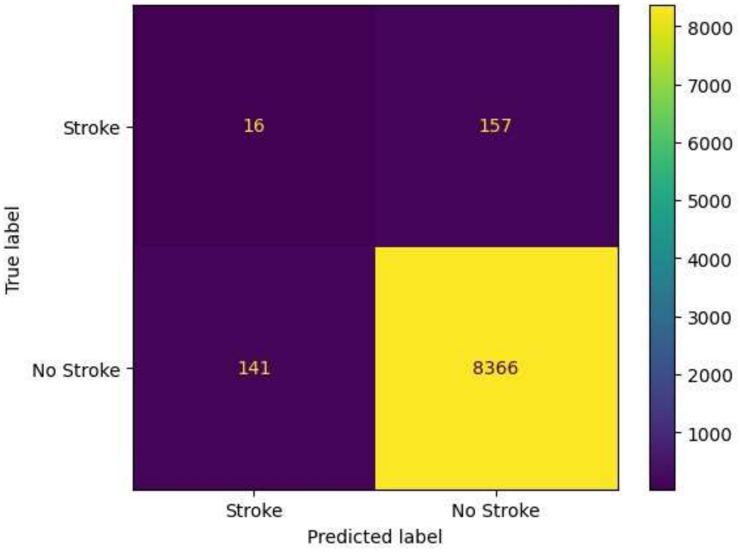


III. Isolation Forest

The main issue with the above two algorithms was that both were predicting 28% of the test data to have strokes. The actual percentage was 1.8 %. Thus, obtaining a good recall of 80% was pointless. Health coaches cannot be assigned to 29% of the population!

One class classification algorithm such as Isolation Forest achieved a better performance. It is an unsupervised learning algorithm that attempts to model “normal” examples to classify new examples as either normal or abnormal.

Though the precision and accuracy of this algorithm is less ~ 10%, it makes less misclassifications. It assigns approximately the same number of health coaches as the true number of stroke instances. This is feasible for a company despite the model reducing the stroke population by 10%.



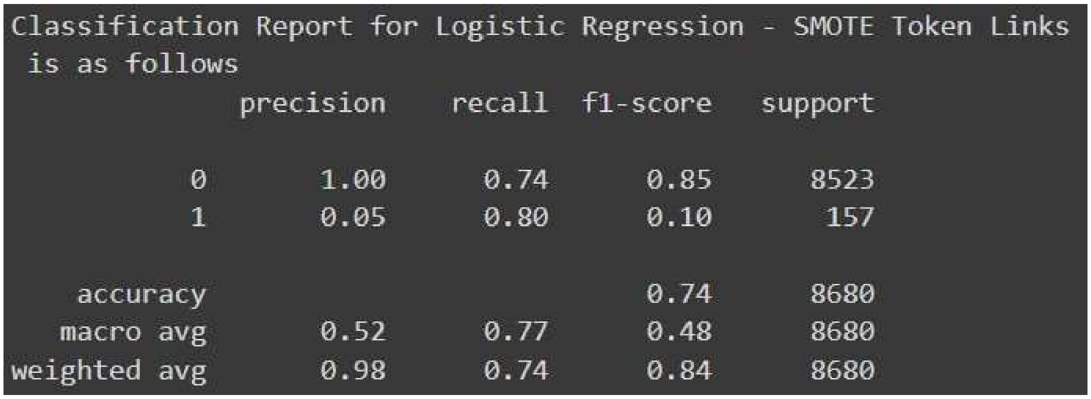
Evaluation Metrics and Results

Accuracy is not the best metric when it comes to imbalanced data because a naive classifier (which predicts all samples as majority class) can have a high accuracy but 0 precision and recall.

Precision and Recall are both important for us as we must assign a reasonable number of health coaches while correctly identifying patients at risk.

Confusion matrix helps us understand our needs the best as seen above.

1. Logistic Regression



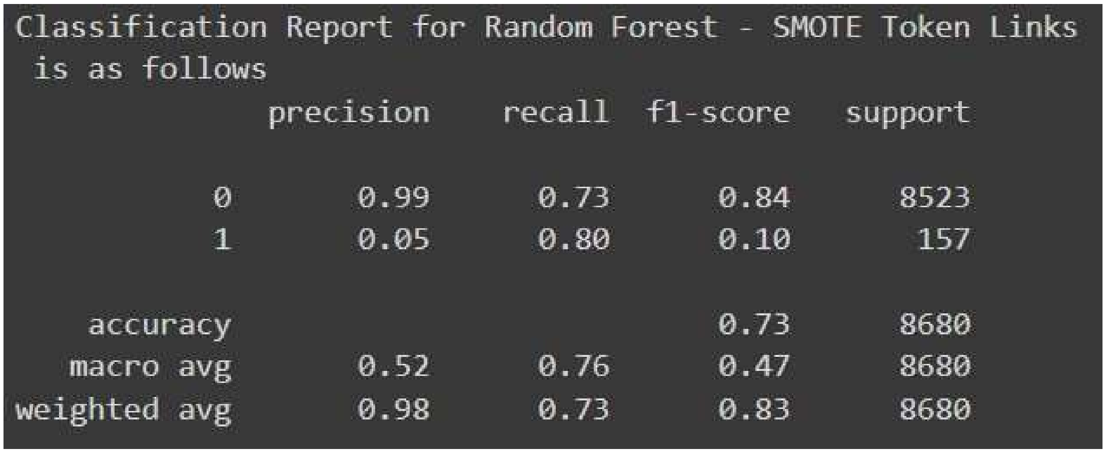
```
Classification Report for Logistic Regression - SMOTE Token Links
 is as follows
              precision    recall  f1-score   support

           0       1.00      0.74      0.85      8523
           1       0.05      0.80      0.10       157

    accuracy                           0.74      8680
   macro avg       0.52      0.77      0.48      8680
weighted avg       0.98      0.74      0.84      8680
```

There was a trade-off between precision and recall as seen above for positive class

2. Random Forest



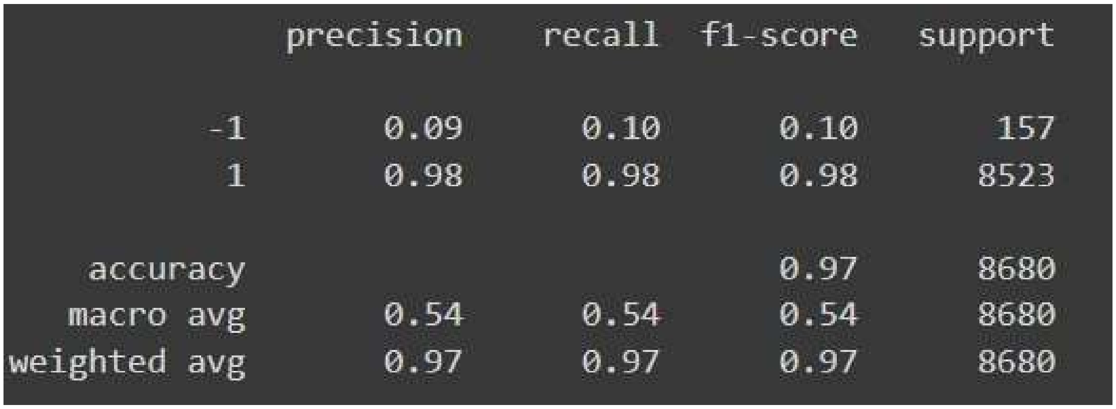
```
Classification Report for Random Forest - SMOTE Token Links
 is as follows
              precision    recall  f1-score   support

           0       0.99      0.73      0.84      8523
           1       0.05      0.80      0.10       157

    accuracy                           0.73      8680
   macro avg       0.52      0.76      0.47      8680
weighted avg       0.98      0.73      0.83      8680
```

Like Logistic Regression, Random Forests could not reduce the number of misclassifications and precision took the hit.

3. Isolation Trees



```
              precision    recall  f1-score   support

          -1       0.09      0.10      0.10       157
           1       0.98      0.98      0.98      8523

    accuracy                           0.97      8680
   macro avg       0.54      0.54      0.54      8680
weighted avg       0.97      0.97      0.97      8680
```

Here -1, means the positive stroke instance, an outlier for a one class classification algorithm.

Even though precision and recall are less, it makes very few misclassifications as observed from the confusion matrix.

There were other metrics such as the roc curve, but the confusion matrix was the most explainable out of all.

Future Work

Precision Recall Trade-off

This was the problem I faced while testing out different models. Each algorithm trades one for the other (Eg : SMOTE increases recall at the cost of precision). I would like to understand more about this topic to handle imbalanced datasets better.

New Algorithms

I would like to explore and research other algorithms which build upon a different hypothesis. This may improve the poor performance of the current classifiers. E.g., One Class classification is used for identifying outliers, but it can be modified to handle imbalanced datasets. It does not require resampling unlike most of the supervised algorithms.