Stroke Prediction via Supervised Methods

Arnau Abella and Antoni Casas Universitat Politècnica de Catalunya

Introduction

Strokes are an unexpected **killer** and **hard to predict**, being the second leading cause of death globally. Therefore, the detection of segments of the population at risk could enhance the ability of hospitals and other medical teams to offer early treatments and screenings to reduce this risk.

Our goal is to **predict whether or not an individual is going to suffer a stroke**.

Materials & Methods

The stroke prediction dataset is composed by **11 features**: 3 numerical and 8 categorical. The dataset only has **5110 observations**. The binary classification problem is **highly imbalanced**, with only a 5% of the population belonging to the positive class.

To deal with the imbalance, we applied **SMOTE** which applies the over-sampling method *SMOTE* combined with the under-sampling method *Tomek Links*. The combination of both should re-balance the dataset while avoiding the creation of noisy samples from the interpolation of outliers and inliers.

We have also explored **factor analysis**, in particular MCA, and dimensionality.

We considered the following models:

- Logistic Regression with Regularization
- Bayesian Models
- Support Vector Machines
- Decision Trees
- Ensemble Methods: Random Forest, Ada Boosting, and Voting Classifiers
- Neural Networks

We focused on achieving the best *F1-score* while keeping a high *recall*.

Final Results

We evaluate how our best model *generalises* using the test data. Our optimized model achieves a 79% **of true positives on the minority class** which seems to be an accurate representation of its generalization capability.

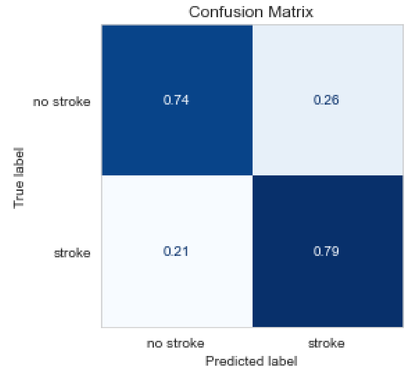

Figure 3: Confusion Matrix on Test Data

We can also see on the ROC curve (figure 2) that this model has an interesting quirk, and that's it that we **can achieve a true positive rate of 100% on the minority class** if we also have a false positive rate of 60% in the minority class, which means we can effectively never miss a possible stroke if we are willing to also deal with all the positives.

Preliminary Results

	Accuracy	Recall	F1-score (class 1)	F1-score (class 0)	F1-score (macro avg)	ROC AUC Score
MLP(default)	75.2%	73.2%	19.2%	85.4%	52.3%	74.3%
LR(best)	73.9%	75.6%	18.8%	84.4%	51.6%	74.7%
LR(default)	74.6%	73.2%	18.8%	84.9%	51.8%	73.9%
Voting(soft)	72.4%	73.2%	17.5%	83.4%	50.5%	72.8%
Voting(hard)	77.5%	56.1%	16.7%	87.0%	51.8%	67.2%
SVC(default)	80.8%	43.9%	15.5%	89.2%	52.4%	63.1%
RF(default)	75.3%	51.2%	14.3%	85.6%	49.9%	63.8%
RF(best)	71.8%	58.5%	14.3%	83.1%	48.7%	65.5%
MLP(best)	79.6%	41.5%	14.0%	88.5%	51.3%	61.4%
DT(default)	75.0%	46.3%	13.0%	85.4%	49.2%	61.3%
DT(best)	76.1%	43.9%	12.9%	86.2%	49.5%	60.7%
Knn(default)	81.4%	34.1%	12.8%	89.6%	51.2%	58.8%
AdaBoost(default)	42.9%	90.2%	11.2%	57.9%	34.6%	65.6%
GB(default)	34.4%	95.1%	10.4%	48.3%	29.4%	63.5%
QDA(default)	96.0%	0.0%	0.0%	98.0%	49.0%	50.0%
NB(default)	96.0%	0.0%	0.0%	98.0%	49.0%	50.0%

Figure 1: Validation scores after hyperparameters tunning

The results show that the best models are the ones with **higher regularization**. It seems that the issue is **severe overfitting** due to the data imbalance and the scarcity of data. In our case, the models which achieved the best results were the multi-layer perceptron, the logistic regression with l_2-penalty and the voting classifier. Finally, after hyper-parameter tunning, we have chosen the **logistic regression** which achieves the highest f1-score and recall among the bests.

Conclusion

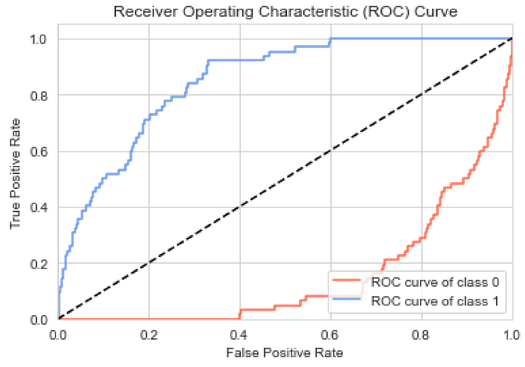

Figure 2: ROC Curve on Test Data

Out of all our models, the ones most robust to overfitting were the ones who managed to perform well on the minority class. We believe this to be a case of a lack of data, since certain spaces of the hyperplane are left either unpopulated or with so little population that boundaries cannot be established properly, leading to overfitting and poor generalization.

Our best model was the Logistic Regression with L_2 loss, a simple linear method, but extremely hard to overfit, with an interesting ROC curve, which could be exploited to obtain the best balance between false positives and true positives, since the focus is on achieving the highest number of true positives on the minority class, since a misdiagnosis can be fatal.

References

[1] Paytonfisher. Stroke prediction ml project - 96% score, Jun 2021.

[2] Mariahchristy. Stroke dataset: Eda and prediction: 96% accuracy, Jun 2021.

Future Research

The data seems **too poor** to be able to create a good model for the minority class, however, it might be possible that there exists a transformation where, in the new space, the data is good enough to generalize with the current data.
Ignoring such event, the only possible work is expanding the data, via new variables recorded, or by **adding more samples**.

Contact Information

Web github.com/monadplus/ml-project
Email arnauabella@gmail.com
Phone +34 618 459 006