# Analyzing the Performance of TabTransformer in Brain Stroke Prediction

Haoming Xia

2022-06-22

**Abstract**

The adoption of electronic patient health records has paved the way for machine learning and deep learning in disease diagnostics and prediction. Neural networks are especially suited to such tasks as they can handle noisy data and perform well even with many input variables. In this study, we measure the performance of TabTransformer, a new deep tabular data modelling architecture proposed in 2020. We then compare TabTransformer's performance with other state-of-art machine learning algorithms, including the feed-forward Multilayer Perceptron model, which performed well in previous studies.

## Introduction

The widespread adoption of electronic health records in recent decades has dramatically improved the quality of patient care. EHRs improve communication between patient and provider by increasing transparency and accessibility and reduce medical errors by enhancing legibility and availability of patient data. Furthermore, they assist in developing novel treatments and algorithms as securely anonymized data sources for machine learning and other data mining techniques. In this paper, we analyze the performance of TabTransformer, a deep tabular data modelling architecture introduced by Amazon in 2020 (Huang et al. 2020), in predicting the occurrence of brain stroke. We benchmark other state-of-the-art methods including XGBoost (Chen and Guestrin 2016), the Kaggle competition winner for the stroke prediction dataset[1]; and Multilayer Perceptron, the model with the most performance in a 2019 study (Nwosu et al. 2019).

[1] https://www.kaggle.com/code/ahmtcnbs/stroke-prediction-xgboost-97

# Dataset

We selected a dataset of patient records first released for the McKinsey Analytics Hackathon[2] by McKinsey Analytics. The dataset contains information from 43,401 patients and 11 clinical features for predicting stroke events: patient gender, age, hypertension, marital status, work type, residence type, average blood glucose level, hypertension, body mass index, and smoking status. The 12th attribute, the target variable, indicates if the patient has had a stroke or not. The dataset is published Kaggle[3], a public dataset repository. We use all 12 attributes to benchmark our predictive models.


Figure 1: Scatterplot matrix

[2]https://datahack.analyticsvidhya.com/contest/mckinsey-analytics-online-hackathon/
[3]https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

## Methodology

We compare the performance of five classification algorithms — DecisionTree, RandomForest, XGBoost, Multilayer Perceptron, and TabTransformer.

Our first problem is that 30% of the records are from patients with unknown smoking status. Our dataset contains categorical and continuous values, so we impute these unknowns with MissForest, a performant and computationally efficient missing value imputation algorithm that handles mixed-type data[4].

Moreover, the dataset of electronic heaalth records is highly unbalanced. Out of 43,401 records in the dataset, only 748 are from patients with stroke. Therefore, we divide our dataset into 70%/15%/15% cross-validation splits. We shuffle and split the records into sets of 30,381 training records, 6510 validation records, and 6510 test records. Finally, we balance the training set. Treating the positive cases as the minority case and the negative cases as the majority case, we upsample the minority case with the SMOTE-NC algorithm and clean the resulting values with Edited Nearest Neighbors 51013 rows. Finally, for XGBoost, RandomForest, and DecisionTree, we one-hot-encode the categorical features of each row.

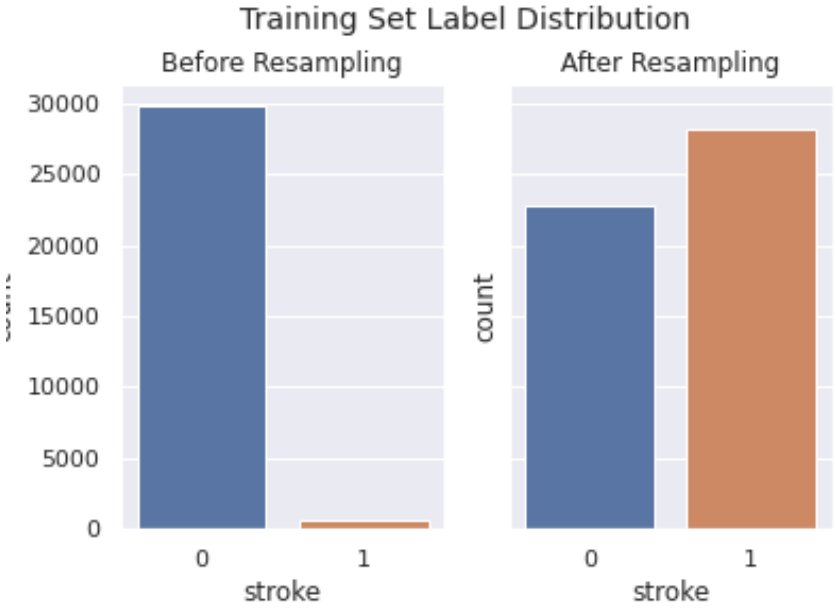


Figure 2: Training Set Label Distribution

We repeat this training-validation-test split and resampling experiment 10 times to minimize sampling bias and take the mean AUC, Binary Accuracy, and miss rate across all 10 as the evaluation metrics of each classification algorithm.

[4]Stekhoven and Buhlmann (2011)

## Results

XGBoost has the best mean performance over all experiments at 87.94% AUC. Next are MLP, RandomForest, TabTransformer, and DecisionTree.

Table 1: Mean accuracy, AUC, and miss rate before resampling

| | Mean Accuracy | Mean AUC | Mean Miss Rate |
|---|---|---|---|
| XGBoost | 0.982857 | 0.899631 | 0.0174433 |
| RandomForest | 0.989831 | 0.937669 | 0.0101347 |
| DecisionTree | 0.978879 | 0.724109 | 0.00952759 |
| TabTransformer | 0.983103 | 0.772889 | 0.0170634 |
| MLP | 0.983364 | 0.897986 | 0.0164655 |

Table 2: Mean accuracy, AUC, and miss rate after resampling

| | Mean Accuracy | Mean AUC | Mean Miss Rate |
|---|---|---|---|
| XGBoost | 0.982765 | 0.879396 | 0.0175231 |
| RandomForest | 0.983241 | 0.807357 | 0.0169513 |
| DecisionTree | 0.964793 | 0.626568 | 0.012931 |
| TabTransformer | 0.98318 | 0.771692 | 0.0170621 |
| MLP | 0.982596 | 0.831457 | 0.0172757 |

Compared to a baseline MLP model with the transformer layers removed, TabTransformer performance similarily to MLP: The mean accuracy of MLP and TabTransformer are roughly the same at 98.31% accuracy and 98.25% accuracy, respectively. As it is essential to catch as many cases of stroke as possible, miss-rate is an important metric. Though the MLP has a 7.074% greater AUC, TabTransformer has a marginally lower miss rate.

## Conclusion

TabTransformer shows no significant improvement over MLP and even performs worse in some metrics. Neither TabTransformer nor MLP performed better than XGBoost, the best performing algorithm for this dataset on Kaggle.

## References

Chen, Tianqi, and Carlos Guestrin. 2016. "XGBoost." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. https://doi.org/10.1145/2939672.2939785.

Huang, Xin, Ashish Khetan, Milan Cvitkovic, and Zohar Karnin. 2020. "TabTransformer: Tabular Data Modeling Using Contextual Embeddings." arXiv. https://doi.org/10.48550/ARXIV.2012.06678.

Nwosu, Chidozie Shamrock, Soumyabrata Dev, Peru Bhardwaj, Bharadwaj Veeravalli, and Deepu John. 2019. "Predicting Stroke from Electronic Health Records." arXiv. https://doi.org/10.48550/ARXIV.1904.11280.

Stekhoven, D. J., and P. Buhlmann. 2011. "MissForest–Non-Parametric Missing Value Imputation for Mixed-Type Data." *Bioinformatics* 28 (1): 112–18. https://doi.org/10.1093/bioinformatics/btr597.