Stroke Prediction Model Project Report

By Geoffrey Varin

https://github.com/Geoffreyvarin/DSTI/

Executive Summary

This project focuses on developing a predictive model to assess the likelihood of a stroke in patients based on various health and lifestyle parameters. Key findings include the significant predictive power of age and BMI, but also highlight cross effects of some medical metrics that weren't individually significant, but proved to be when combined in a composite metric: the health risk score.

Introduction

Stroke, has the second leading cause of death globally with about 11% of total death, necessitates effective predictive models. This project used a dataset with 11 distinct features like gender, age, health indicators, or lifestyle information to predict stroke likelihood.

Methodology

The project employed data preprocessing, exploratory data analysis, feature engineering, model testing and evaluation.

Class imbalance has been addressed through the SMOTE methodology, and missing value imputation for BMI has been addressed through KNN imputation.

Several models were tried such as RandomForest, GradientBoosting, Logistic Regression, KNN, SVC and Neural Network. Emphasis was placed on recall and F1-score in model evaluation to prioritize capturing as many positive cases as possible.

The final model has been fine-tuned by applying a grid search to its hyperparameters.

The project has been conducted on a Jupyter notebook using Python and a series of standard libraries for data science, including pandas, numpy, sci-kit learn, and Pyplot, amongst others.

Final hosting has been organized on Github.

Results

Model evaluations lead to the selection of the **Logistic Regression model,** for its superior recall and precision.

Discussion

1. The cost component regarding positive and negative could be beneficial to fine-tuning the algorithm. Without this information, a focus was given on Recall, but it would be interesting to look for an algorithm that would optimize the cost function (societal and Monetary)

2. The health risk score could benefit from further research At the moment, it is a simple aggregation of normalized values for glucose, hypertension, and heart disease. However, it is to be expected that fine-tuning this indicator's formula could lead to better results. Also, identifying the exact type of heart disease could probably lead to better results.

3. Depending on its application, it would be interesting to diversify the planned actions according to the threshold. In the context of a medical campaign, differing probabilities could lead to different actions. This would allow for an effective cost management of the campaign.

Conclusion

Through this project, we highlight the potential of machine learning in medical applications. Though not highlighted in this work, a medical context implies that specific attention should be brought to the risk of introducing bias into the results.

Appendices and References

For further information, please refer to the accompanying Jupyter Notebook and its HTML export for detailed code, methodologies, and references.

Project hosted on Github: https://github.com/Geoffreyvarin/DSTI/