Analyzing and Modeling for Predicting Stroke

T8: Yang Liu

04/28/2022

Background and Motivation

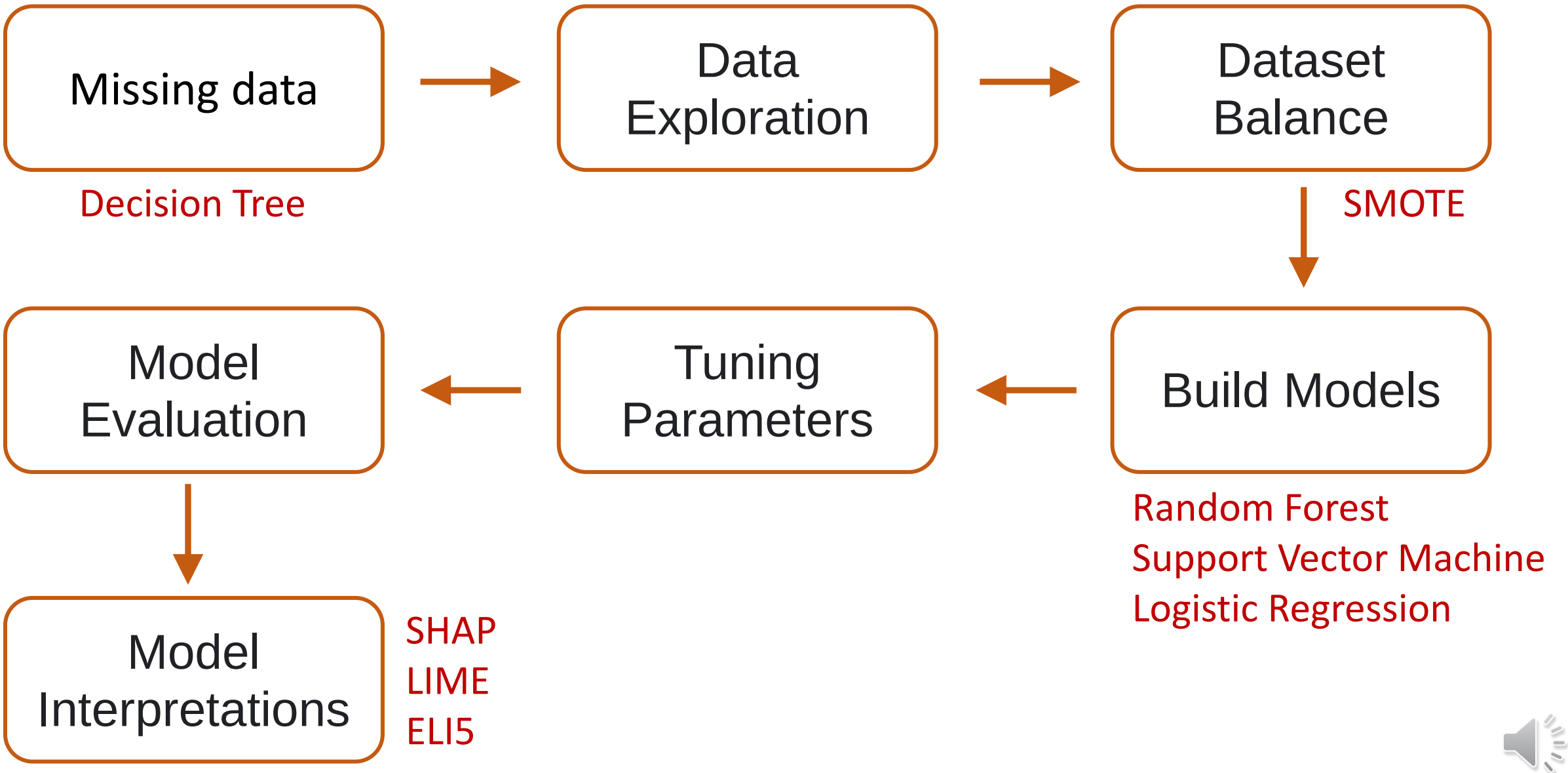
About Stroke

- Stroke is the 2nd leading cause of death and a major cause of disability worldwide
- Low- and middle-income countries endure an 80% mortality rate with hemorrhagic stroke
- Western countries spend 3 to 4% of total health care expenditures on stroke
- Prediction of stroke can help disease treatment and reduce burden of society

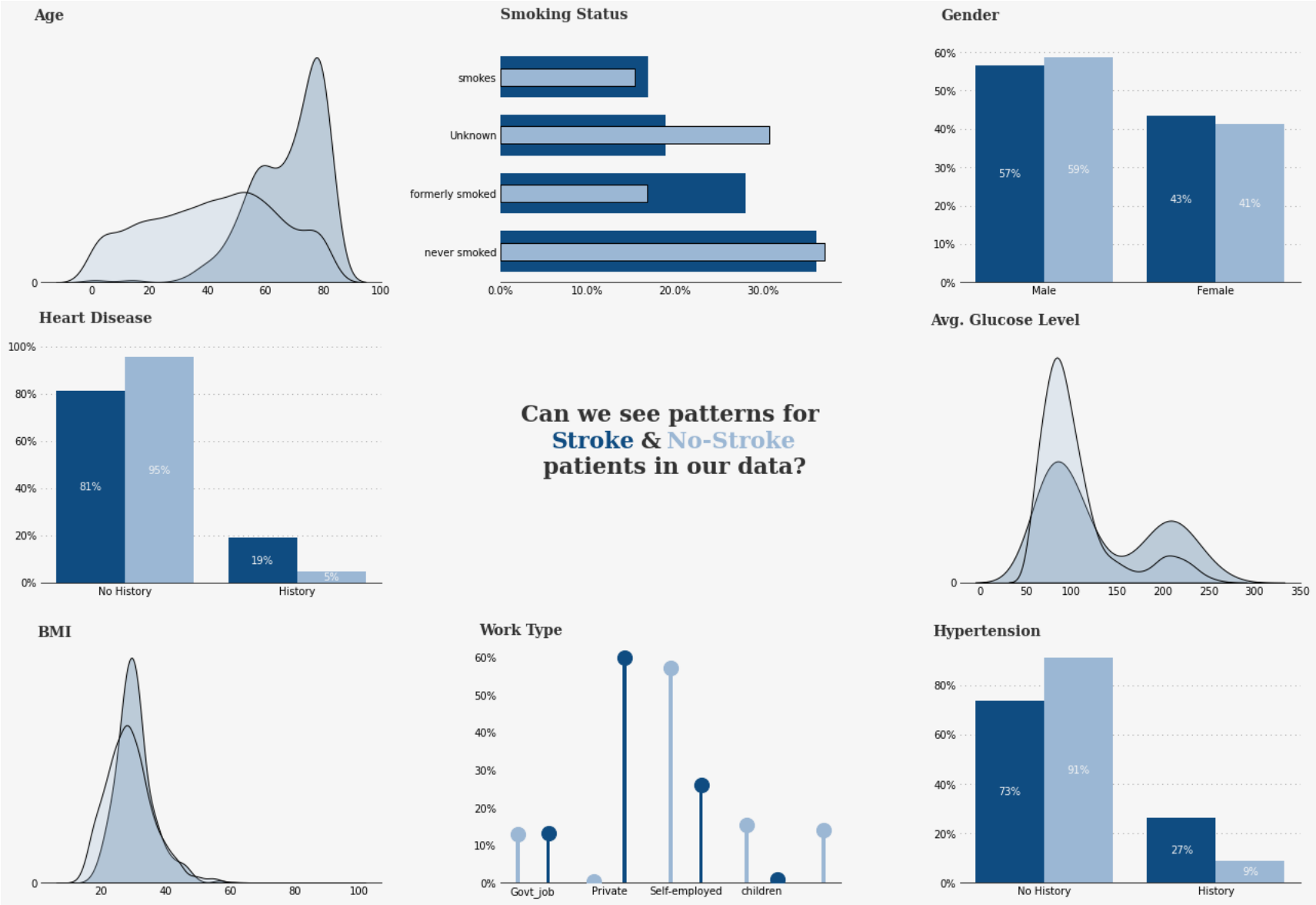
Motivation

- My research interest lies on data analysis and machine learning of materials and biomedical science
- I like Kaggle projects because they provide very good practical exercises
- As a beginner of AI, I want to learn some fundamental but useful ML skills by studying other's projects

Methods


Missing data
Data Exploration
Dataset Balance
Decision Tree
SMOTE
Model Evaluation
Tuning Parameters
Build Models
Random Forest
Support Vector Machine
Logistic Regression
Model Interpretations
SHAP
LIME
ELI5

Results – Data Exploration

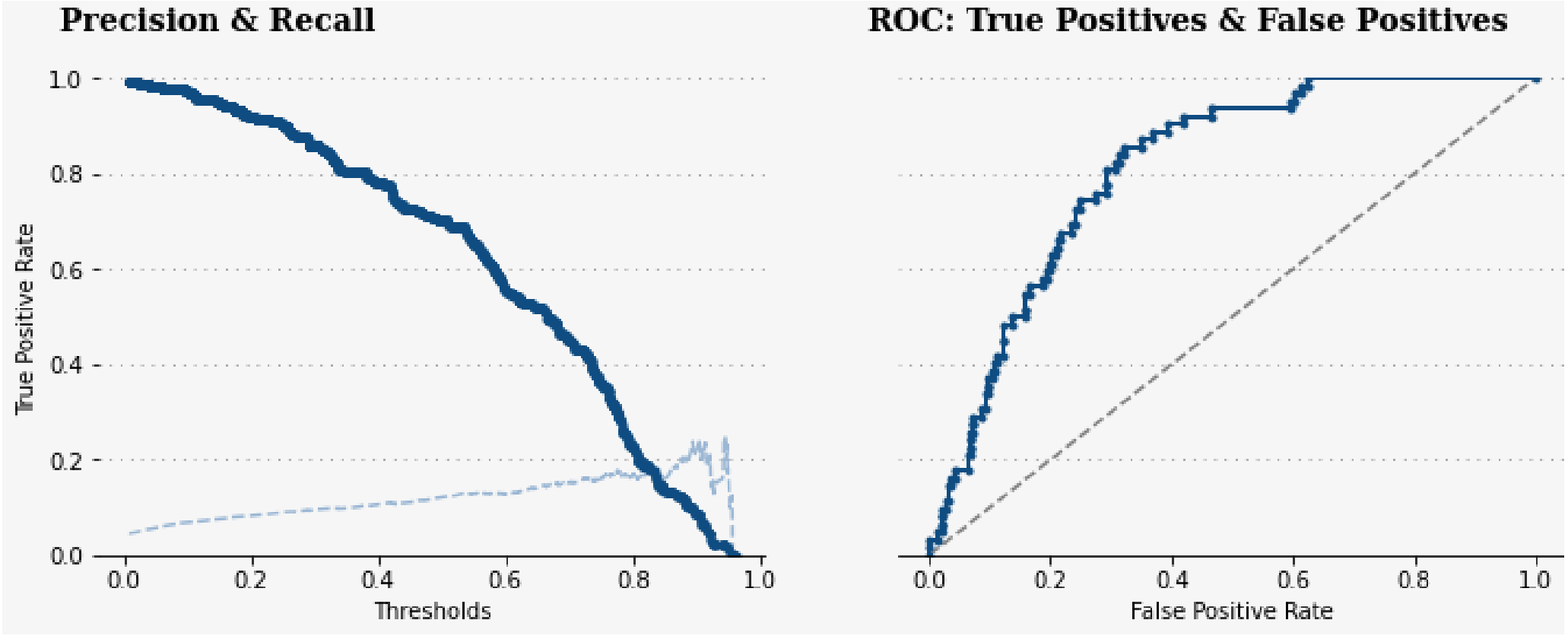


Results – Model training

- Train/test split = 8/2
- 10-fold cross validation

F1 Score	Random Forest	Support Vector Machine	Logistic Regression
Training	0.937	0.833	0.799
Test	0.206	0.272	0.259

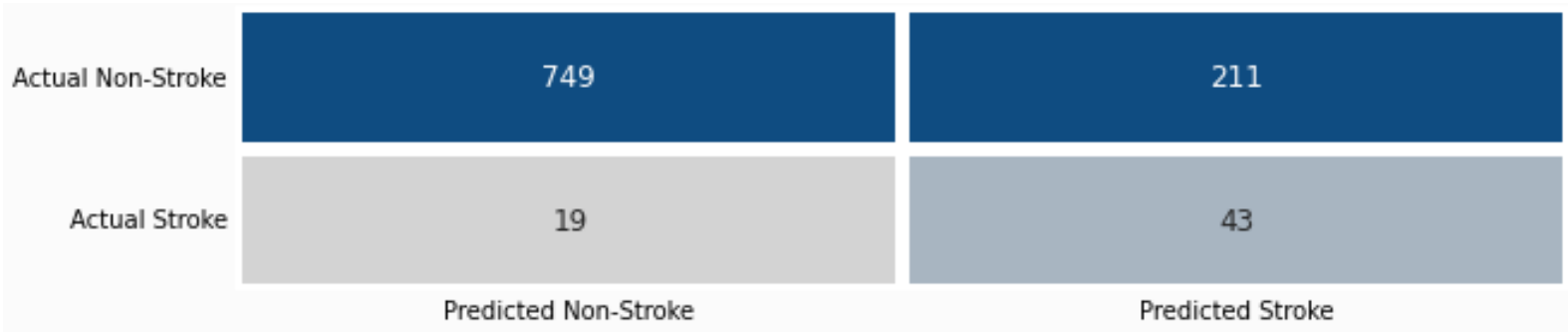
Results – Hyperparameter Tuning



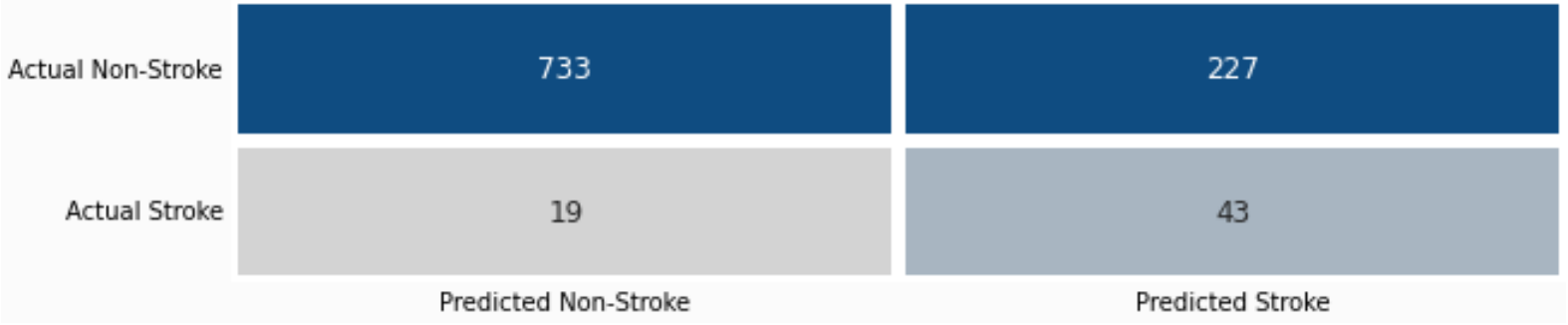
Results – Model Evaluation

Random Forest

Support Vector Machine



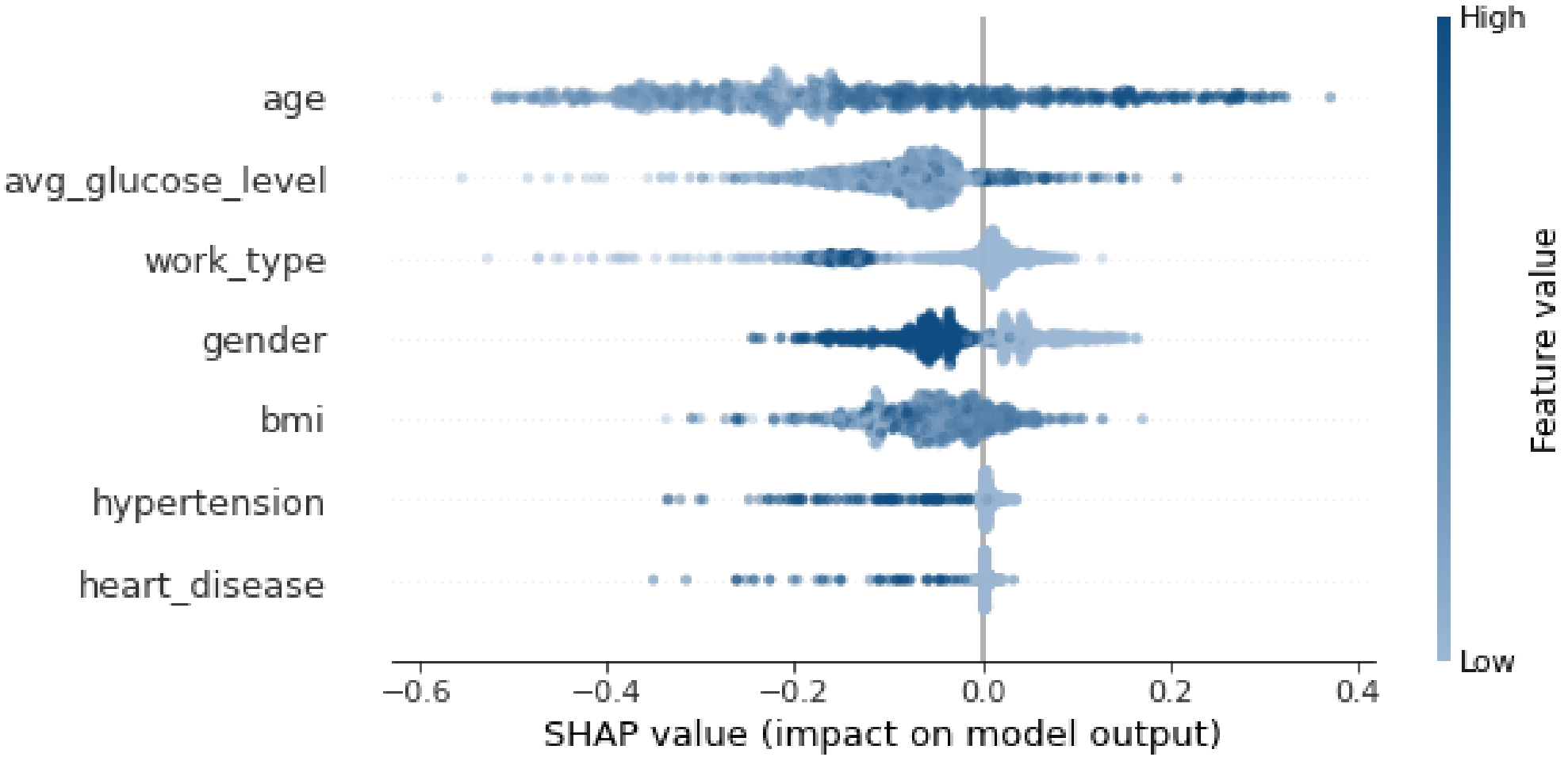
Logistic Regression



Results – Model Evaluation

	Random Forest	Support Vector Machine	Logistic Regression
F1	20.6%	27.2%	25.9%
Accuracy	87.2%	77.5%	75.9%
Recall	27.4%	69.4%	69.4%
Precision	16.5%	16.9%	15.9%
ROC AUC Score	59.2%	73.7%	72.9%

Results – Model Interpretation



Conclusions and Lessons

- I replicated a Kaggle project of analyzing and modeling stroke data.
- Three models were built to predict stroke. Random forest has a good accuracy, but its recall is quite bad. Support vector machine and logistic regression have more balanced performance.
- Through this project I learned some useful techniques. One is using SMOTE to balance dataset, another is to interpret models with SHAP, LIME, and ELI5.

References

- G. J. Hankey, "Stroke," (in English), *Lancet,* Review vol. 389, no. 10069, pp. 641-654, Feb 2017, doi: 10.1016/s0140-6736(16)30962-x.
- M. Katan and A. Luft, "Global Burden of Stroke," (in English), *Semin. Neurol.,* Review vol. 38, no. 2, pp. 208-211, Apr 2018, doi: 10.1055/s-0038-1649503.
- https://www.kaggle.com/code/joshuaswords/predicting-a-stroke-shap-lime-explainer-eli5/notebook

Thank you !

Q & A