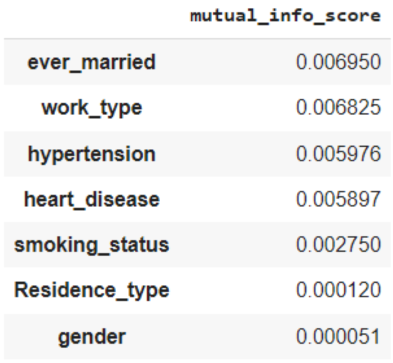
Determining the Likelihood of a Stroke Occurring Through Machine Learning Models

Statistics 451

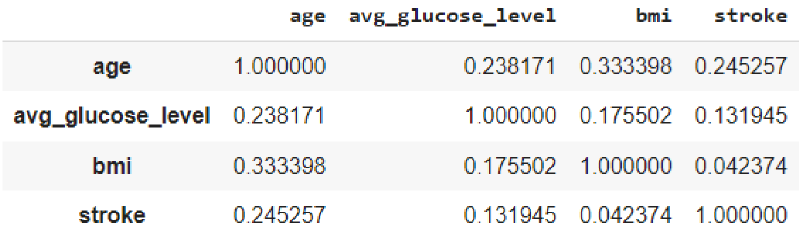
Chloe Rasmussen, Aiden Chiang, Brian Johnson, Tyler Bui, and Yegor Baranovski

Strokes rank as the 5th leading cause of death in the United States, claiming approximately 162,000 lives annually, while accounting for 11% of deaths globally. In this project, we used a kaggle dataset of 5110 individuals with categorical (gender, hypertension, heart disease, work type, smoking status, marriage history, residence type) and numerical features (age, bmi, average glucose levels). Our goal was to devise a model that most effectively correctly predicts a positive individual's stroke history. After testing, we found that a logistic model maximizes average AUC but in turn leaves more to be desired in terms of precision and recall.

Preemptive data analysis revealed several insights that influenced our subsequent modeling decisions and help contextualize final results. Firstly, individuals with a positive stroke history constitute only 4.87% of data, indicating class imbalance. Additionally, our numerical data exhibits skewness; we used Yeo-Johnson function transformations followed by standardization to alleviate this. Moreover, hypertension and heart disease are underrepresented in the dataset, comprising only 9.75% and 5.4% of observations, respectievly. These conditions are known risk factors and their underrepresentation could of skewed our scoring assessments. We additionally had to remove the ID column and imputed the median for individuals with missing bmi values. Exploring the data further we compare our various features with stroke rates, noting that particularly hypertension and heart disease both increased positive stroke history by about 4x. Besides various graphical representation, we create the following mutual information table and correlation matrix:



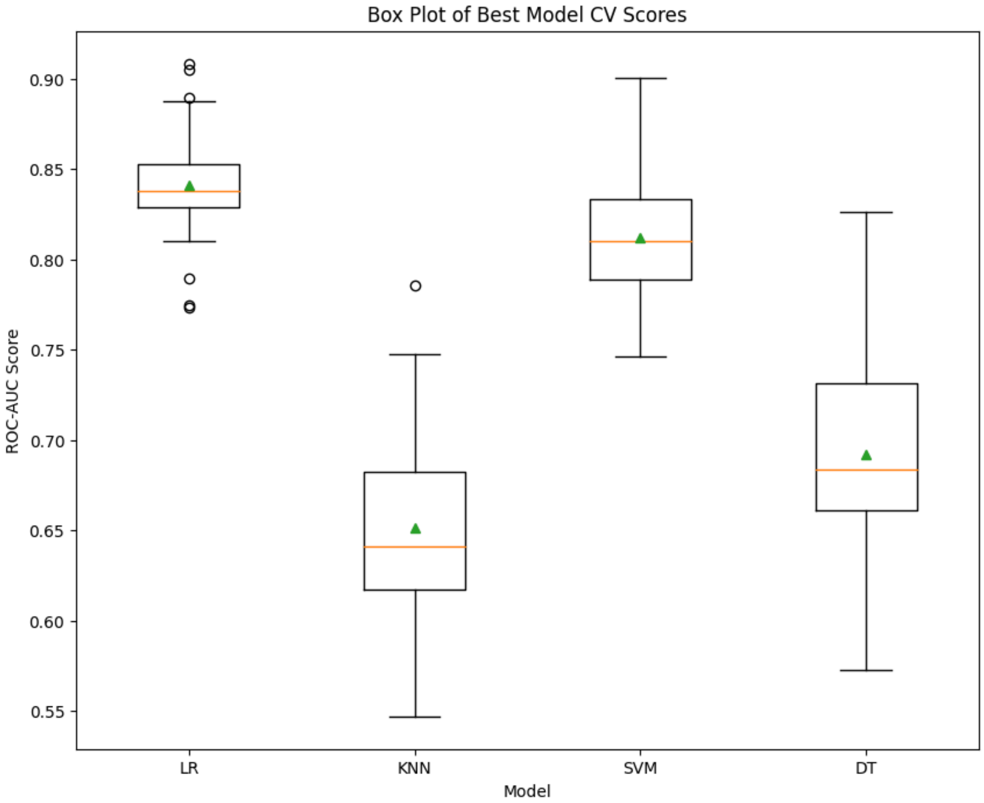
	mutual_info_score
ever_married	0.006950
work_type	0.006825
hypertension	0.005976
heart_disease	0.005897
smoking_status	0.002750
Residence_type	0.000120
gender	0.000051



	age	avg_glucose_level	bmi	stroke
age	1.000000	0.238171	0.333398	0.245257
avg_glucose_level	0.238171	1.000000	0.175502	0.131945
bmi	0.333398	0.175502	1.000000	0.042374
stroke	0.245257	0.131945	0.042374	1.000000

Initially inspecting, residence type, gender, and bmi seem to comparitavely tell us significantly less about stroke history, which is particularly interesting for bmi as it has been historically highly correlated with stroke. However, this may be because we are looking at PAST stroke rates, which may effect an individuals bmi in the present (reverse causality). It is also important to remember that many of these features are very likely to be correlated with one another, i.e age and hypertension.

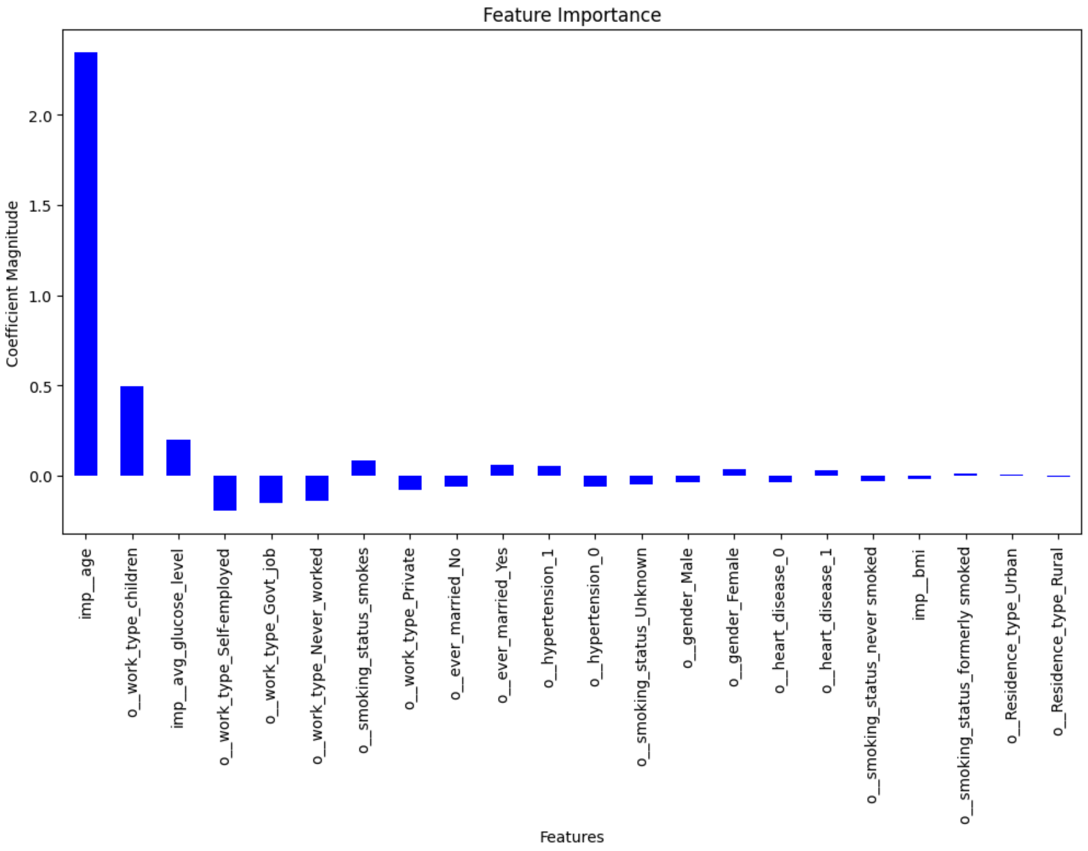
We evaluate and choose between the following models: Logistic, KNN, SVM, and Decision Tree for this binary classification problem. After splitting our data, we used grid search to determine optimal hyperparameters for each model. In order to deal with imbalanced data, we use SMOTE as our method of oversampling. Our evaluation tool was then K fold repeated stratified cross validation with 10 folds iterated 3 times, resulting in 511 observations per fold with each model being evaluated 30 times. We compared each model using ROC AUC, and our results are shown in the following box plots:



In addition to having the highest average ROC-AUC score and lowest standard deviation, the logistic regression model also had the highest average precision and recall scores:

```
>Mean ACCURACY: 0.750 (Std Dev: 0.017)

>Mean ROC_AUC: 0.841 (Std Dev: 0.032)

>Mean PRECISION: 0.139 (Std Dev: 0.016)

>Mean RECALL: 0.799 (Std Dev: 0.089)

>Mean F1: 0.237 (Std Dev: 0.026)
```

Next, we do feature selection in order to reduce the model to hopefully improve our scoring metrics. The following is a graph illustrating feature importance, noting that age is by far the most important metric:



In fact, besides age, you can remove all features and our scoring metrics remain practically the same(with logistic still being the best):

```
>Mean ACCURACY: 0.750 (Std Dev: 0.018)

>Mean ROC_AUC: 0.841 (Std Dev: 0.032)

>Mean PRECISION: 0.139 (Std Dev: 0.016)

>Mean RECALL: 0.796 (Std Dev: 0.087)

>Mean F1: 0.237 (Std Dev: 0.027)
```

Our best reason as to why this is the case is that other features are highly correlated with age. Reducing our model to only age actually increases our test recall score from .64 to a .68, not particularly great. If we want to boost our recall score to 95%(to match our imbalanced data), we are forced to move our probability threshold to ~.137% which results in the following confusion matrix change:

```
[[736 236]--->[[456 516]
 [ 18  32]]--->[  2  48]]
```

Therefore, unless one is willing to misdiagnose the history by ~2x to improve correctly classifying positive patients, this would be ill advised. Unfortunately, in part to imperfect data, complexity of health conditions, and model specifications we were unable to generate a model we would be comfortable recommending. On the other hand, our project illustrates the difficulty that arises in using machine learning for predicting past conditions using present data, whilst highlighting the importance of gathering more extensive data, from increasing observations to varying features.

Contributions

Member	Proposal	Coding	Presentation	Report
Chloe Rasmussen	1	0.2	0.8	0.3
Aiden Chiang	0	0.2	0.4	0
Brian Johnson	1	0.4	0.6	0
Tyler Bui	0	0	0	0.2
Yegor Baranovski	0.3	1	1	1