Predicting Stroke Events with Physical Health Data

Group 22,
Grant Zhou, Yile Chen, Tina Cao, Judy Zhu

Introduction

Interest

Stroke is the **2nd leading** cause of death globally, about 11% of total death

- Stroke prevention

Primary question:

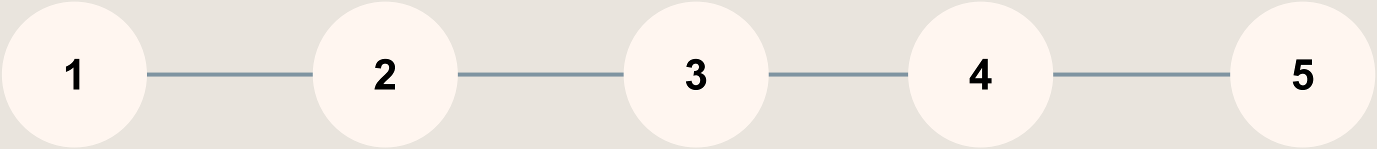
- Whether a patient will have a stroke and the probability of having a stroke given measurements of a patient’s health condition

Dataset

X: gender, age, average glucose level, hypertension, heart disease, ever married, work type, residence type, BMI (body mass index)

Y: stroke event (0 or 1)

Summary of Methods



Imputation for Missing Data + One-hot encoding

Using median for replacing n/a entries

Oversampling

- **Highly unbalanced** dataset
- Before: 0 (95.1%) vs. 1 (4.9%)
- After: 0 (50%) vs. 1 (50%)

Logistic Regression

With L1 norm for feature selection and model fitting

Decision Tree

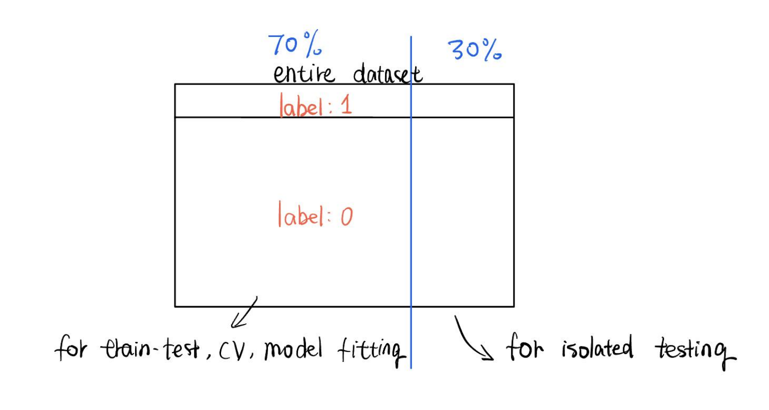
with ID3, use entire feature set for classification

GridsearchCV

For model optimization

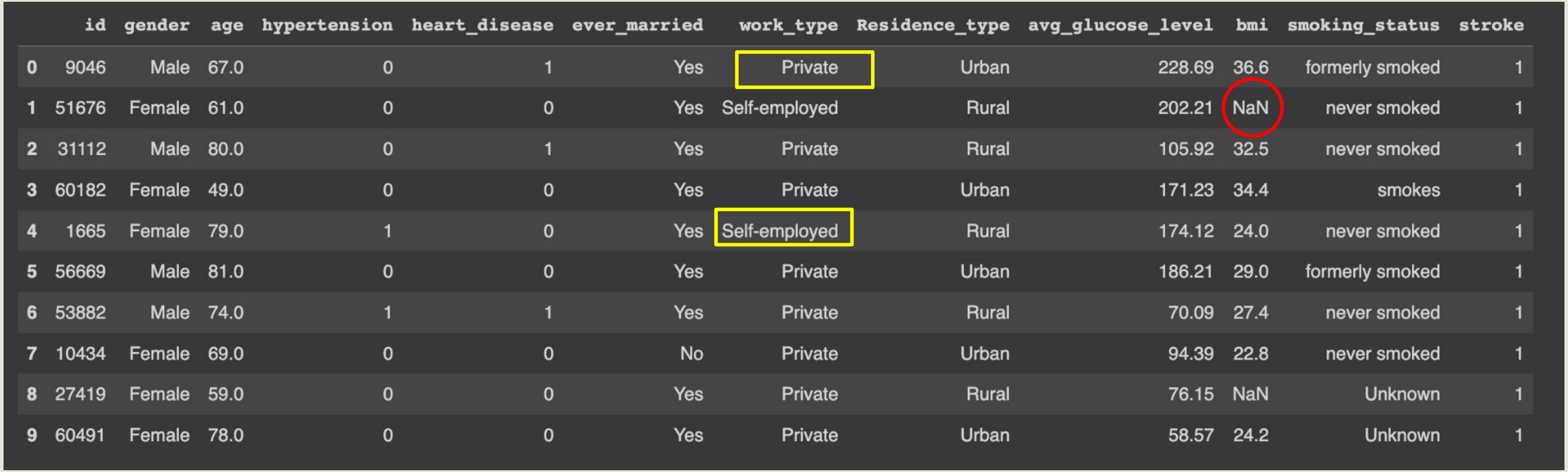
Other Noteworthy Techniques

- We isolated an “untouched” test dataset for testing the performance outside the scope of oversampling, which we suspect still involves overfitting issue



- We adopted a “majority” vote prediction, where we only assign an observation with label 1 if both model predicts label 1

Overview of the Dataset

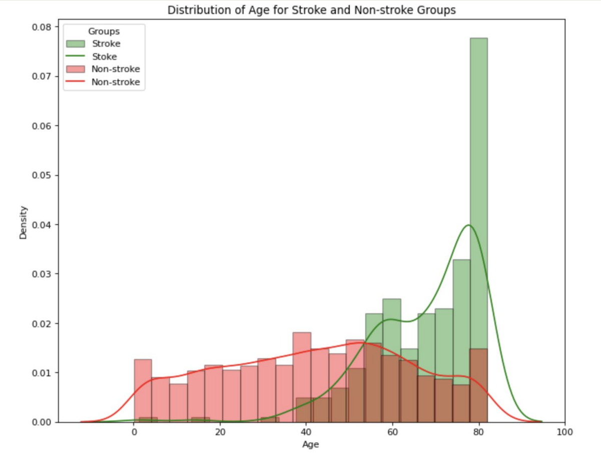


	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
3	60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
4	1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1
5	56669	Male	81.0	0	0	Yes	Private	Urban	186.21	29.0	formerly smoked	1
6	53882	Male	74.0	1	1	Yes	Private	Rural	70.09	27.4	never smoked	1
7	10434	Female	69.0	0	0	No	Private	Urban	94.39	22.8	never smoked	1
8	27419	Female	59.0	0	0	Yes	Private	Rural	76.15	NaN	Unknown	1
9	60491	Female	78.0	0	0	Yes	Private	Urban	58.57	24.2	Unknown	1

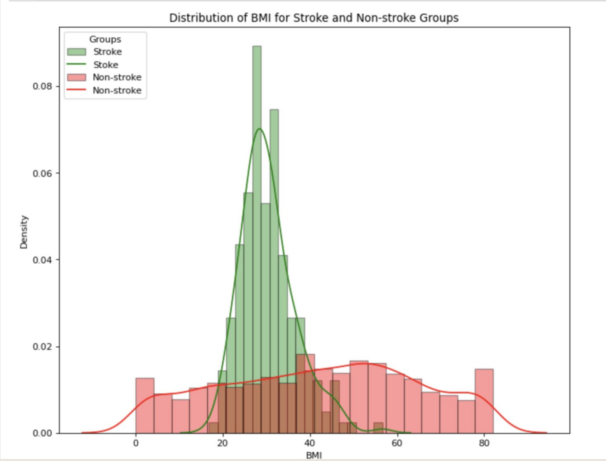
- Detected N/A values for BMI and fixed with SimpleInputer with median
- Performed one-hot encoding on categorical features (multiple classes)

Visualization:

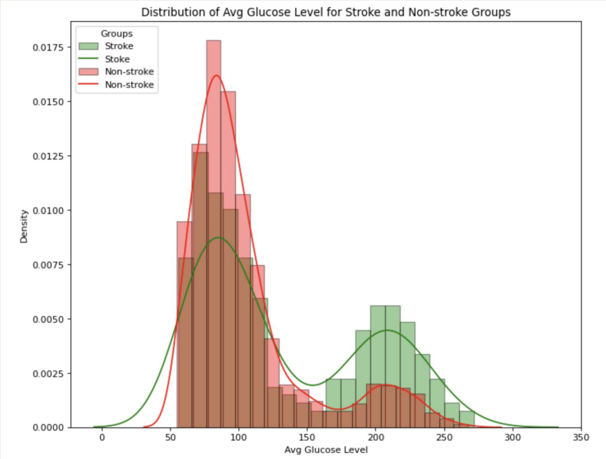
distribution of age, bmi, and avg glucose level in stoke & non-stroke group



Age



BMI

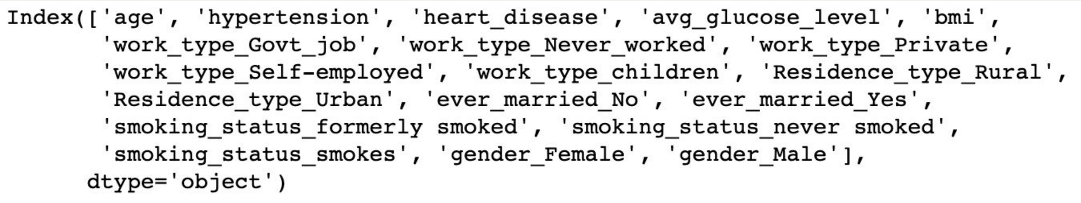


Glucose

Results: feature selection

- linear_model.LogisticRegression(max_iter=5000, C = 0.01, penalty = 'l1', solver = 'liblinear')

- Before: 19 features (after one-hot encoding)



```
Index(['age', 'hypertension', 'heart_disease', 'avg_glucose_level', 'bmi',
       'work_type_Govt_job', 'work_type_Never_worked', 'work_type_Private',
       'work_type_Self-employed', 'work_type_children', 'Residence_type_Rural',
       'Residence_type_Urban', 'ever_married_No', 'ever_married_Yes',
       'smoking_status_formerly smoked', 'smoking_status_never smoked',
       'smoking_status_smokes', 'gender_Female', 'gender_Male'],
      dtype='object')
```

- The we select 3 features that have non-zero coefficients

'age', 'gender_Male', 'avg_glucose_level'

Results: logistic regression

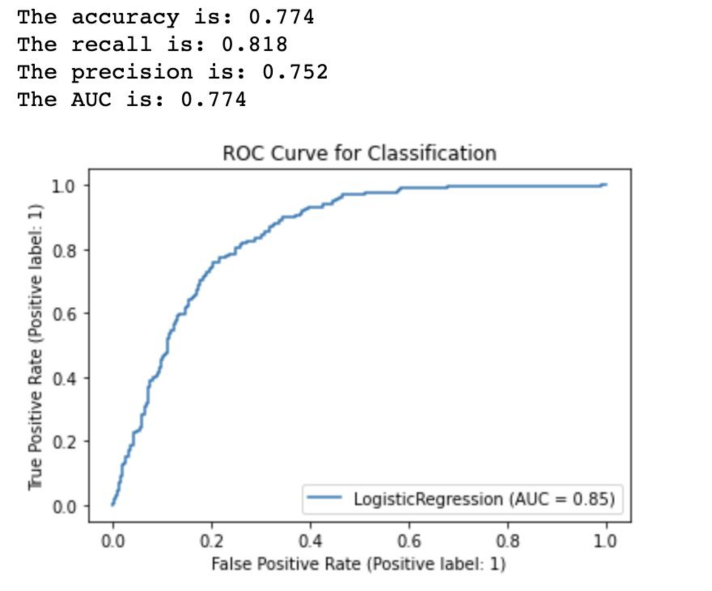
- Grid search CV:

parameters =0

clf.best_score_=0.779
clf.best_params_={'C': 1}

Accuracy with Best Parameters on Validation: 0.768

- Optimized model on test dataset:



```
The accuracy is: 0.774
The recall is: 0.818
The precision is: 0.752
The AUC is: 0.774
```

ROC Curve for Classification

Results: decision tree

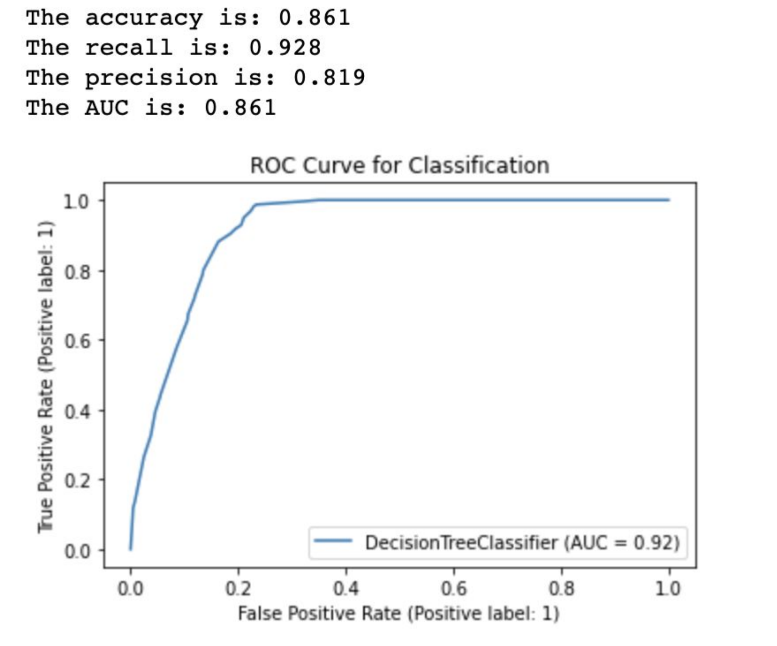
- Grid search CV:
  'Max_depth': [1, 3, 5, 7],
  'min_samples_split': [2,3,4,5],
  'min_samples_leaf': [1,2,3,4]

clf.best_score_=0.858

clf.best_params_=
{'max_depth': 7,
'min_samples_leaf': 1,
'min_samples_split': 4}

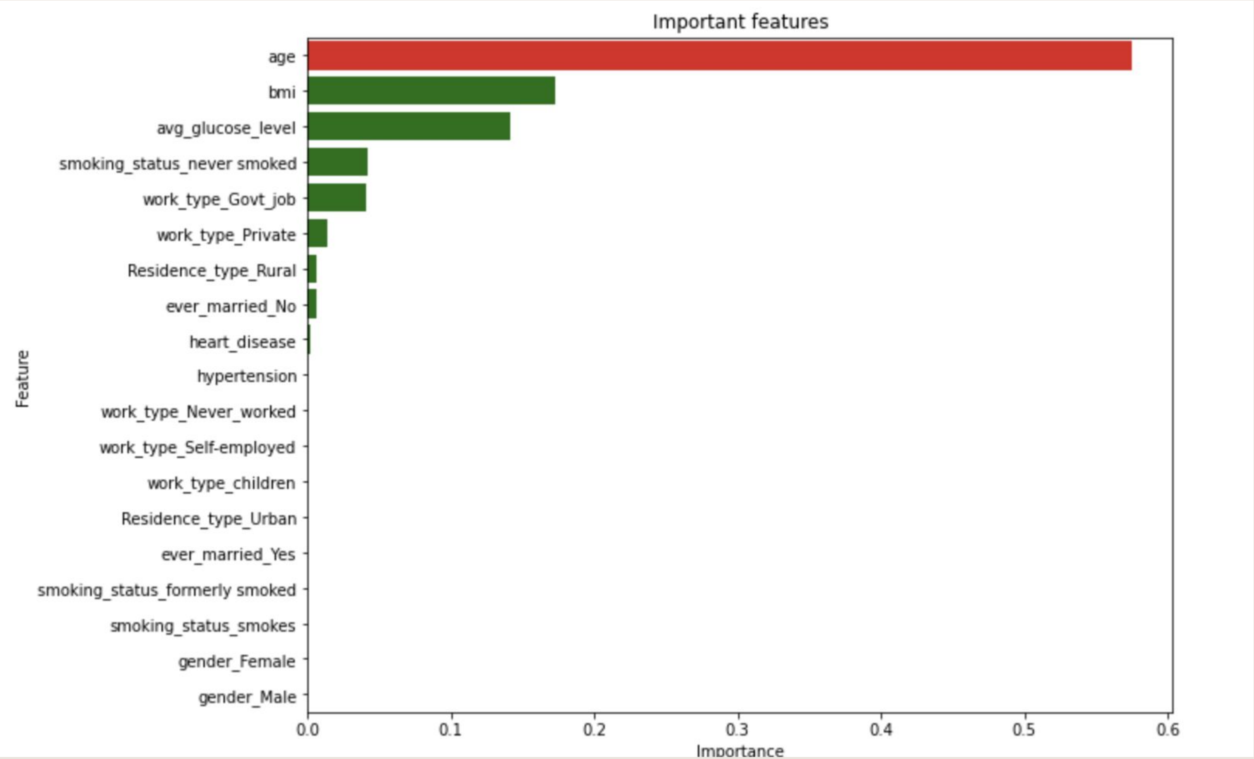
Accuracy with Best Parameters on Validation: 0.867

- Optimized model on test dataset:



```
The accuracy is: 0.861
The recall is: 0.928
The precision is: 0.819
The AUC is: 0.861
```

Results: decision tree - importance score



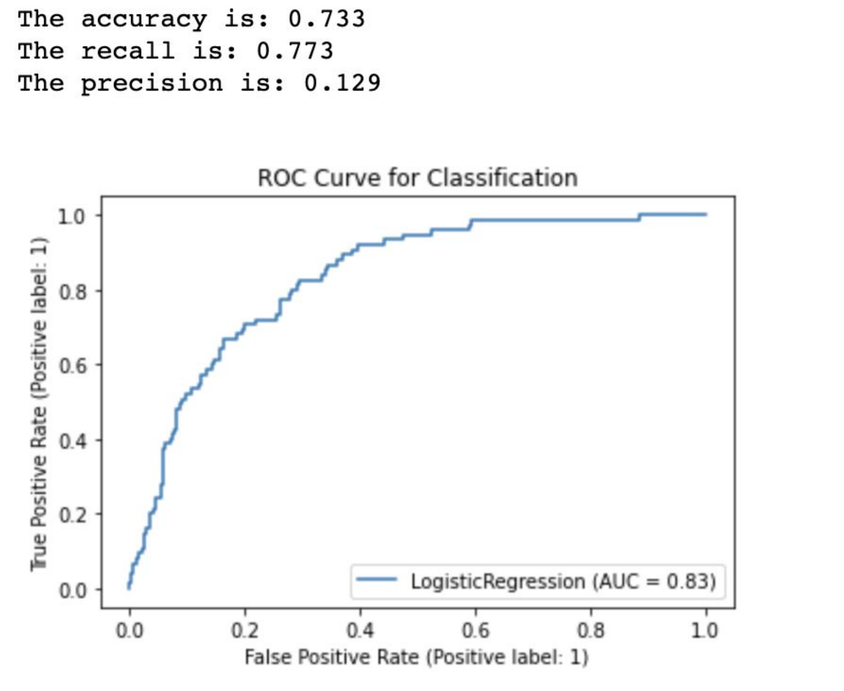
Results: Majority Vote Prediction on Oversampled Dataset

Logistic Regression Result:
The accuracy is: 0.774
The recall is: 0.818
The precision is: 0.752
The AUC is: 0.85

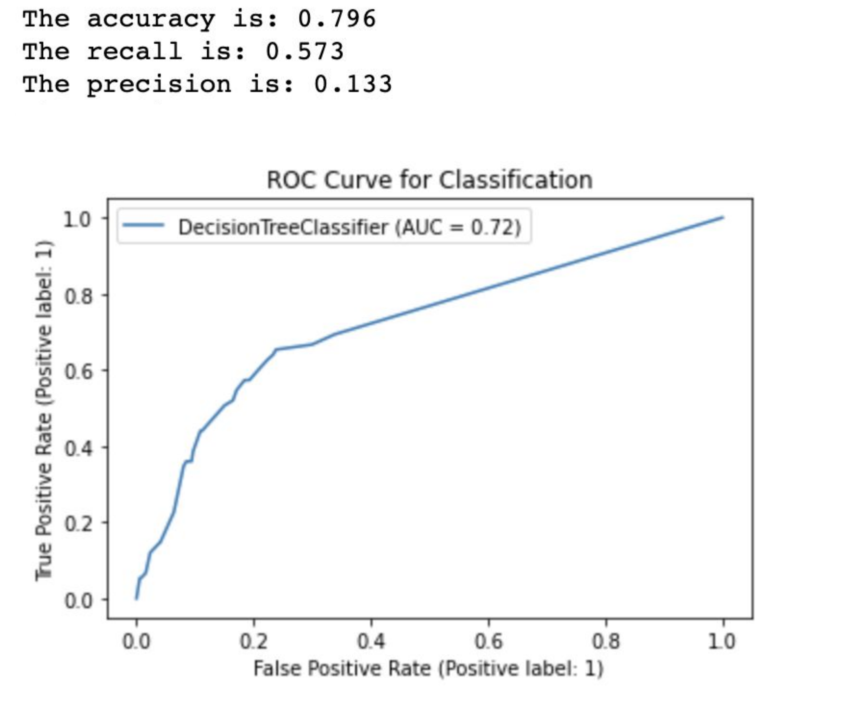
Decision Tree Result:
The accuracy is: 0.861
The recall is: 0.928
The precision is: 0.819
The AUC is: 0.92

“Majority” Vote Prediction
The accuracy is: 0.813
The recall is: 0.773
The precision is: 0.839

Testing result on isolated test data set (highly unbalanced)



```
The accuracy is: 0.733
The recall is: 0.773
The precision is: 0.129
```



```
The accuracy is: 0.796
The recall is: 0.573
The precision is: 0.133
```

Results: Majority Vote Prediction on Highly Unbalanced Dataset

Logistic Regression Result:
The accuracy is: 0.733
The recall is: 0.773
The precision is: 0.129
The AUC is: 0.83

Decision Tree Result:
The accuracy is: 0.796
The recall is: 0.573
The precision is: 0.133
The AUC is: 0.72

“Majority” Vote Prediction
The accuracy is: 0.849
The recall is: 0.507
The precision is: 0.163

Implications

- We built models that has an high accuracy and precision on oversampled dataset and relatively satisfying performance on the “untouched” dataset.
- Identified features (age, bmi, avg glucose level) that are important in predicting stroke and provide certain information that can help design preventative measures.

Thanks

Questions？