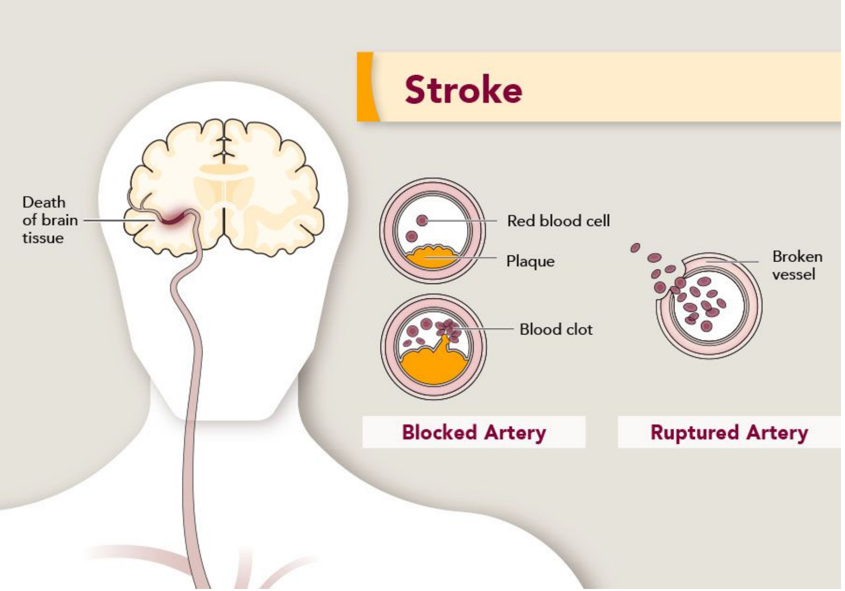
Stroke Prediction

Metis ML Classification

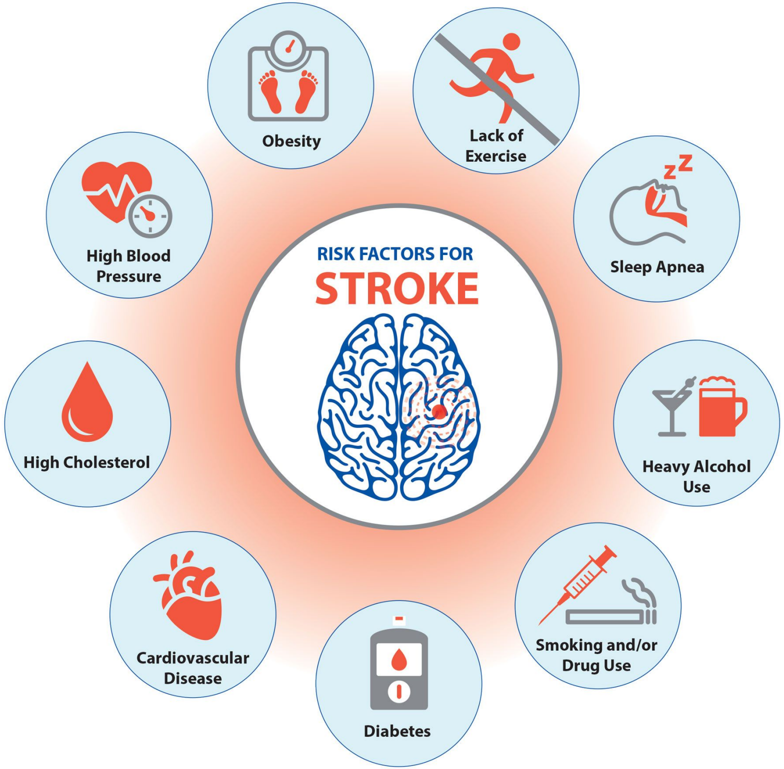
Background

- A **stroke** is when a blood vessel connected to the brain is either blocked or bursts
- Strokes **impact ~800,000 people** (in the US) each year, and **kill ~140,000 people**



https://www.cdc.gov/stroke/facts.htm

Stroke Risk Factors



- Can we re-affirm the assumption that these are stroke risk factors?
- Which of these risk factors are most predictive?
- Can we identify any other potential risk factors?

Classification Goal

Predict whether or not individuals have had a stroke in the past, given their current health status

Secondary: Can that give us insights into which factors may put individuals at high risk for future stroke?

nhanes

The Nation's Mobile Health Survey

- National Health and Nutrition Examination Survey
  - Conducted by the CDC
  - Retrieved on Kaggle
- Comprehensive dataset on health & nutritional status of US citizens
- Includes info on:
  - Demographics (education level, income)
  - Diet
  - A live medical examination (blood pressure, muscle strength, etc)
  - Lifestyle questionnaire (exercise, drug use, etc)
  - Previous medical history
- Overall, 10k survey participants with ~2k features

Feature Selection & Data Cleaning

- Target column: “Have you had a stroke before?”
- Selected a subset of ~30 features
  - Some known to be correlated with stroke
  - Some other potential features
- Restricted to adults
- Data imputation
  - “Mean” for numerical data
  - “Most frequent” for binary or categorical data
- Created dummy variables for categorical data
- Oversampling
  - Target class “had_stroke” has only 200 positives out of 5000
  - SMOTENC - oversampling for both numerical & categorical data

Classification Metrics

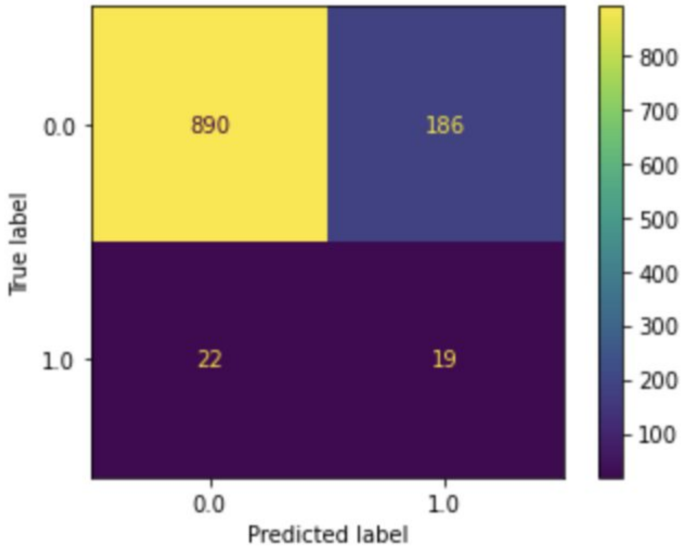
Given the rarity of the positive class, we want to prioritize **recall**

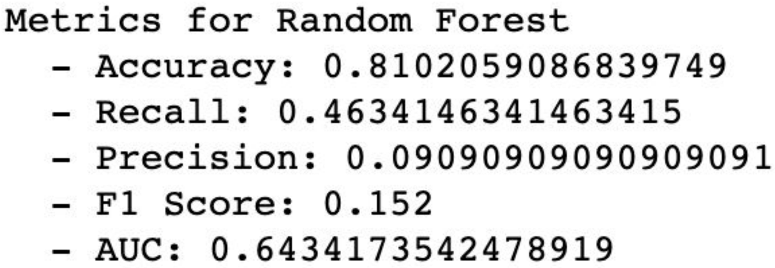
- Ensure that we are successfully predicting as many positives as possible

Model Training & Comparison

Random Forest

- max_depth: 6
- n_estimators: 200

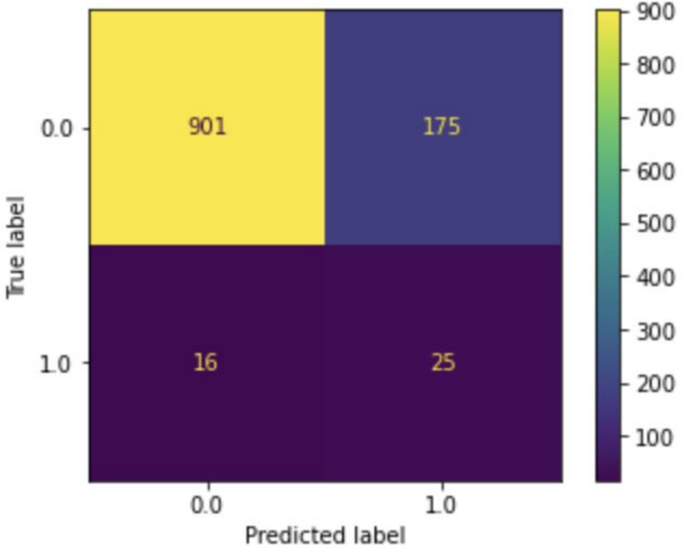


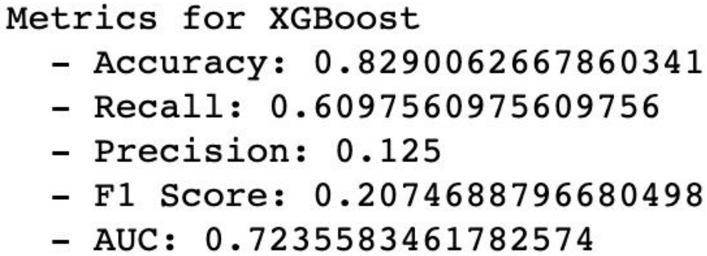


```
Metrics for Random Forest
  - Accuracy: 0.8102059086839749
  - Recall: 0.4634146341463415
  - Precision: 0.09090909090909091
  - F1 Score: 0.152
  - AUC: 0.6434173542478919
```

XGBoost

- max_depth:1
- n_estimators: 70

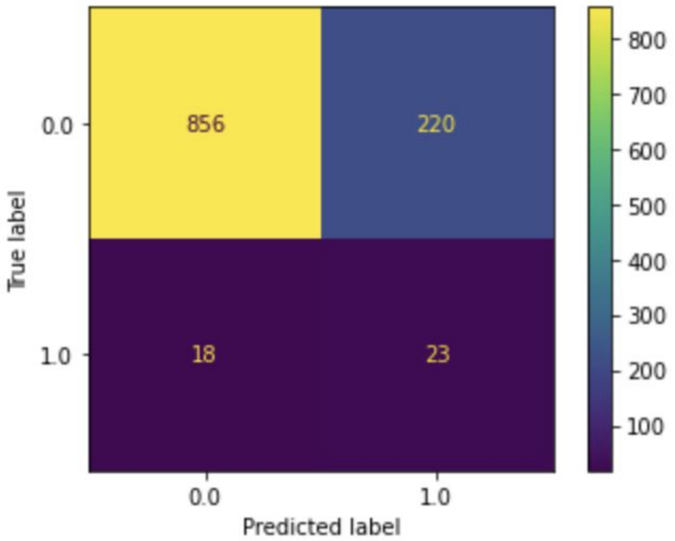


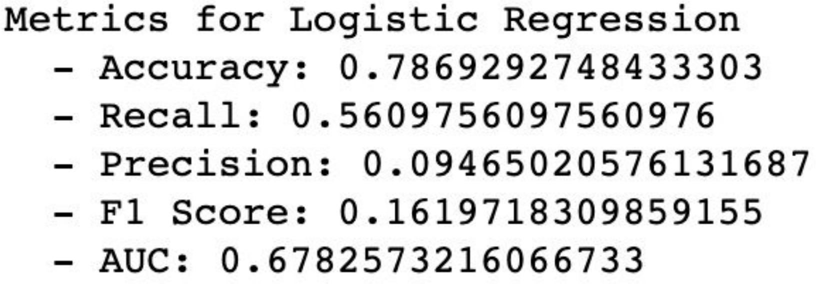


```
Metrics for XGBoost
  - Accuracy: 0.8290062667860341
  - Recall: 0.6097560975609756
  - Precision: 0.125
  - F1 Score: 0.2074688796680498
  - AUC: 0.7235583461782574
```

Logistic Regression

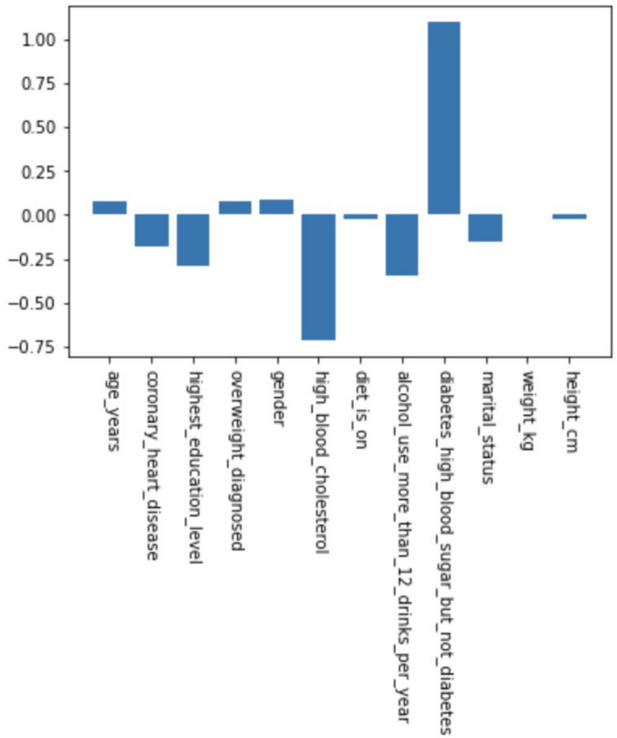
- Using all our features available to us!

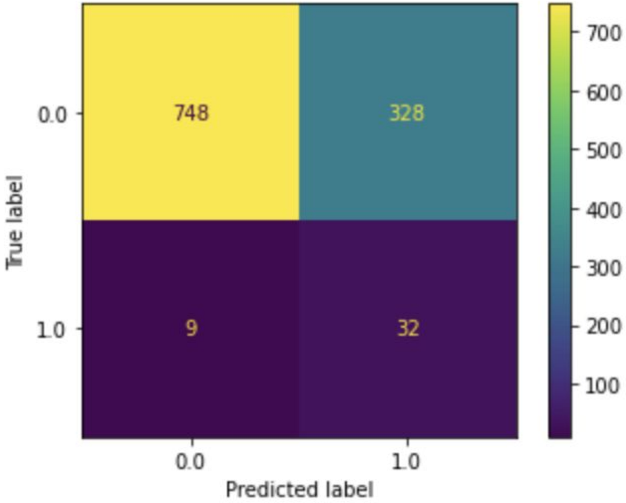


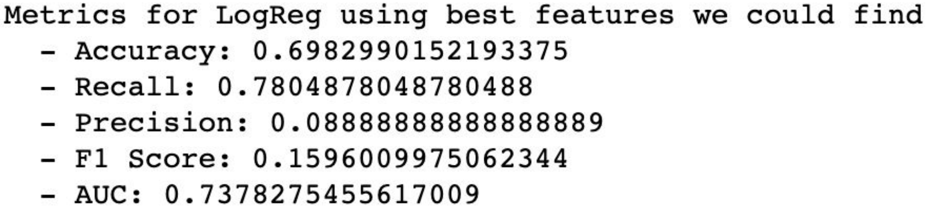


```
Metrics for Logistic Regression
  - Accuracy: 0.7869292748433303
  - Recall: 0.5609756097560976
  - Precision: 0.09465020576131687
  - F1 Score: 0.1619718309859155
  - AUC: 0.678257321606673
```

Logistic Regression - “most important features”

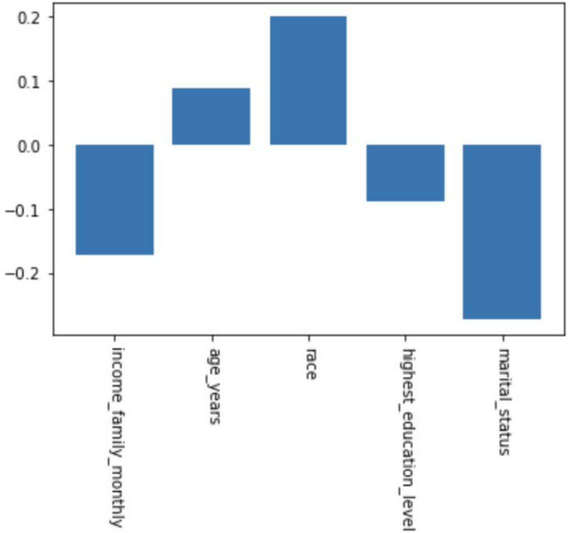


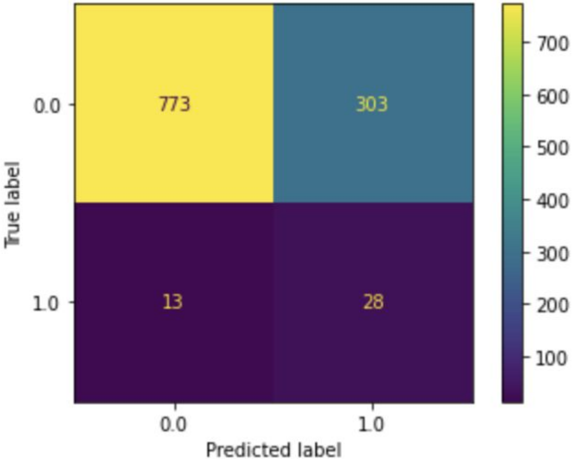


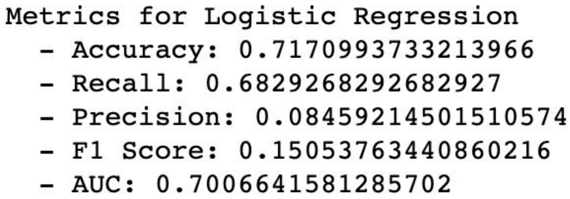


```
Metrics for LogReg using best features we could find
  - Accuracy: 0.6982990152193375
  - Recall: 0.7804878048780488
  - Precision: 0.08888888888888889
  - F1 Score: 0.1596009975062344
  - AUC: 0.7378275455617009
```

What if we don’t know anything about health?







```
Metrics for Logistic Regression
  - Accuracy: 0.7170993733213966
  - Recall: 0.6829268292682927
  - Precision: 0.08459214501510574
  - F1 Score: 0.15053763440860216
  - AUC: 0.7006641581285702
```

Conclusions

- Logistic Regression using only the highest performing features yielded the most accurate (and interpretable results)
- Demographic features taken independently still have somewhat predictive power