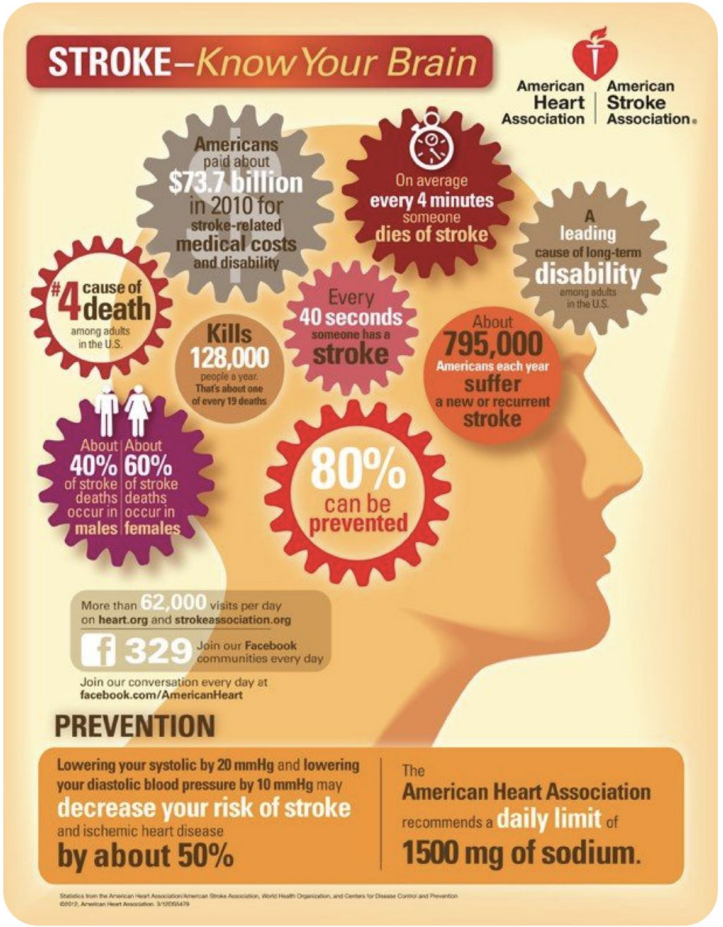
Stroke Prediction

Brooke Hanson

The Problem:

- In the United States, an individual dies of a stroke every 4 minutes
- This prevalence has made strokes the 4th leading cause of death in the United States

What factors affect stroke occurrence?
Can we predict the likelihood of a stroke?



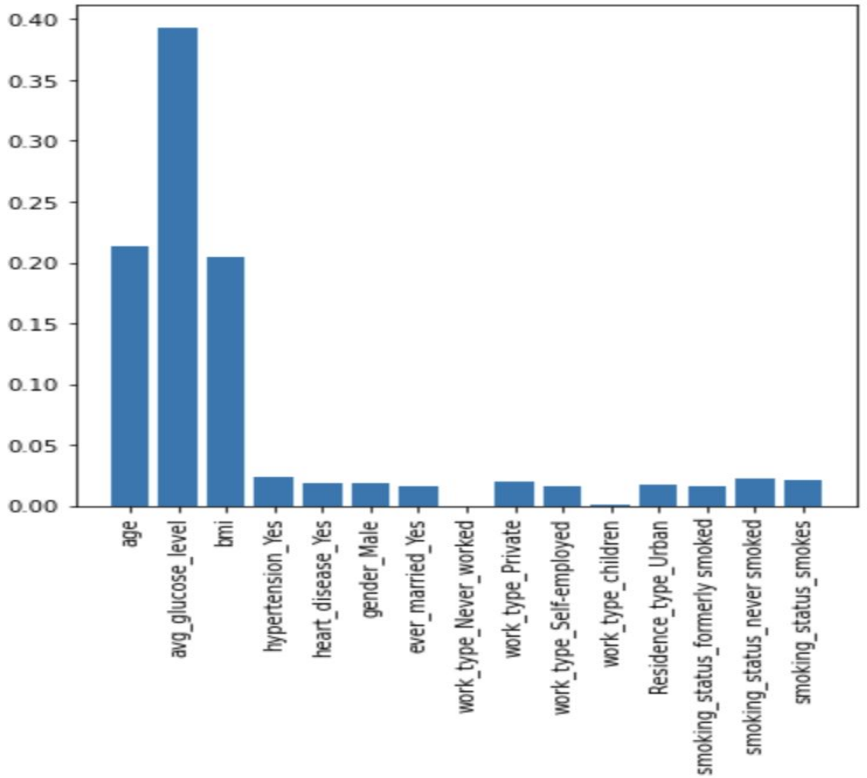
Who might care?

- Strokes affect individuals across all walks of life
- Strokes can cause mental and physical deficits affecting the workforce
    - Between 2014 and 2015 - the cost of strokes was nearly 46 Billion dollars when considering healthcare costs, labor lost, and medicines to treat symptoms

Individuals, Health Care Providers, Insurance Providers, and Employers

Factors that affect stroke occurrence

- Age
- Pre- Existing Health Conditions
- Weight
- Lifestyle choices



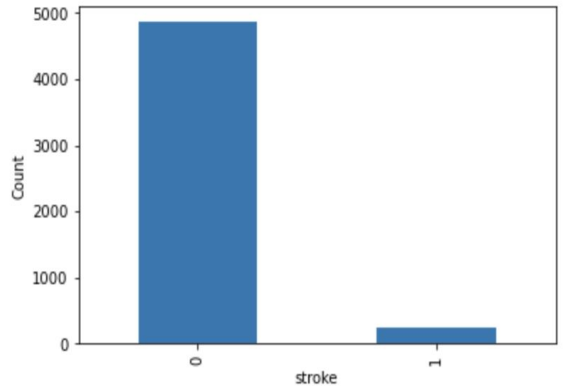
Data Information

- Data set from Kaggle.com
- 5110 rows of data
- 11 columns of information:
  - ID, Age, Gender, Average Glucose Level, BMI, Ever Married, Residence Type, Smoking Status, Work Type, Hypertension, Heart Disease

Exploratory Data Analysis

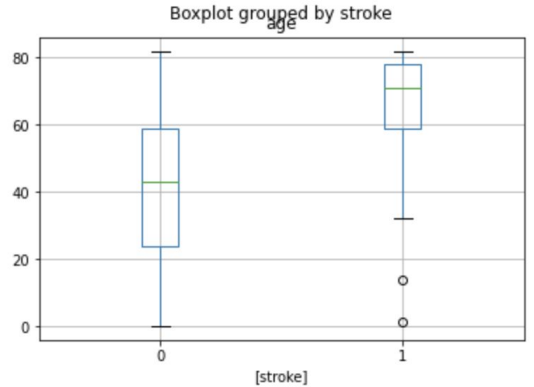
- A significant difference in count of stroke and non-stroke patients

stroke	stroke
0	4861
1	249



- Age seems to have a relationship with stroke occurrence

stroke	mean	std
0	41.971545	22.291940
1	67.728193	12.727419



Exploratory Data Analysis

- Exploring the relationship between explanatory and response variables
- Ran chi^2 tests on categorical variables to assess independence
- Ran T-Tests on numeric variables to assess independence

Variable	P-Values
Age	7.03078e-71
Gender	0.558028512
Heart Disease	2.08878e-21
Hypertension	1.66162e-19
Ever Married	1.6389e-14
Residence Type	0.298331693
Work Type	5.39771e-10
Average Glucose Level	2.76781e-21
BMI	0.009837071
Smoking Status	0.000002085

The only variables independent of Stroke occurrence: Residence Type, and Gender

Machine Learning and Modeling

- Used Supervised Learning Models
- Binary Classification - 1 stroke or 0 no stroke
- Decision Tree Classifier, K Nearest Neighbor, Gradient Boosting Classifier, and Random Forest Classifier
- Highly imbalanced class data: 5.1% of individuals with stroke occurrence
- Tools: Python's sklearn.model_selection, and sklearn.metrics

Modeling Steps

Data Preprocessing Steps:

1. Split into training and test sets (75% & 25%)
2. Weighting classes due to imbalance of stroke occurrence
3. Test scaling

Cross Validation (CV) for Hyperparameter Tuning:

1. 5 fold cross validation
2. Using sklearn grid search method
3. Evaluation Metric: Recall score

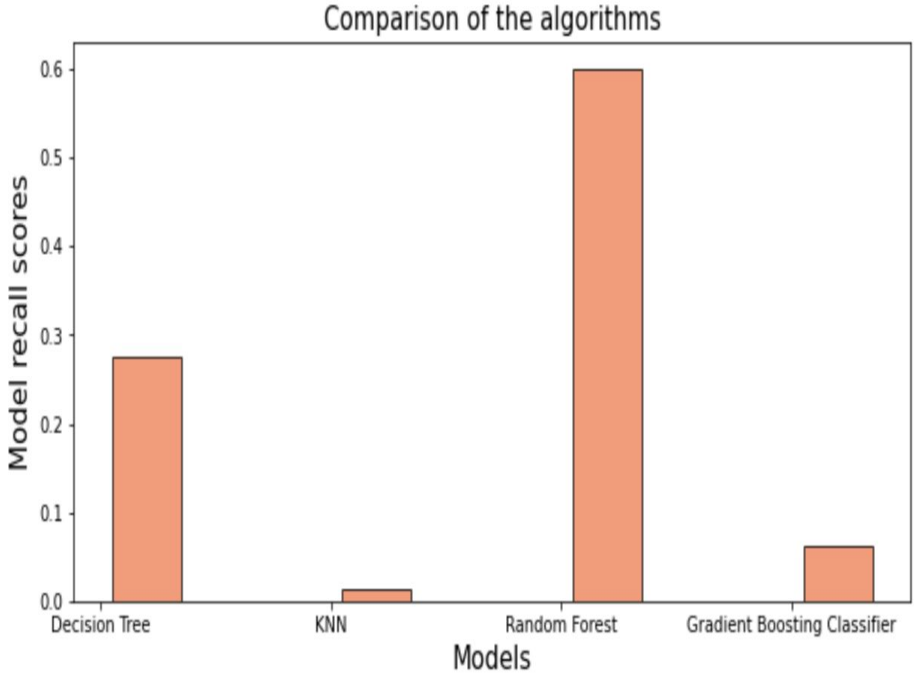
Results

Classifier training based on 75% of data frame

Performance tested on 25% of dataset with best model

Model Comparison

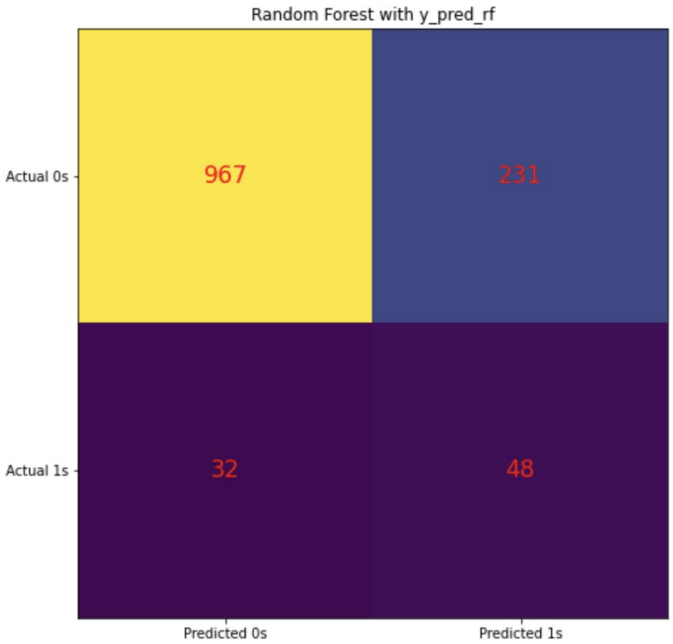
- Performance metric: Recall
- Tuned all models, and compared final results to determine best model
- Random Forest is best model

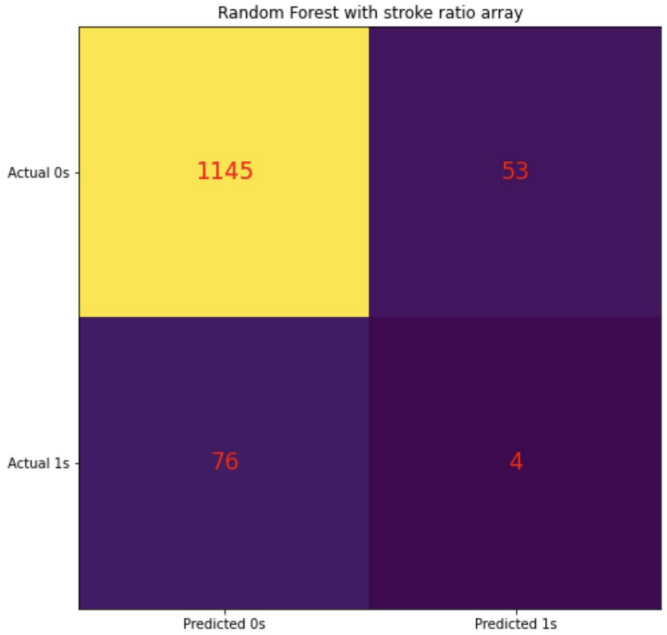


Verifying Model Prediction Capacity

- Due to the discrepancy in stroke occurrence model needs to be verified
- Compared confusion matrix of model predicted values, and confusion matrix of basic binary array with same rate of stroke occurrence

Verifying Model Prediction Capacity



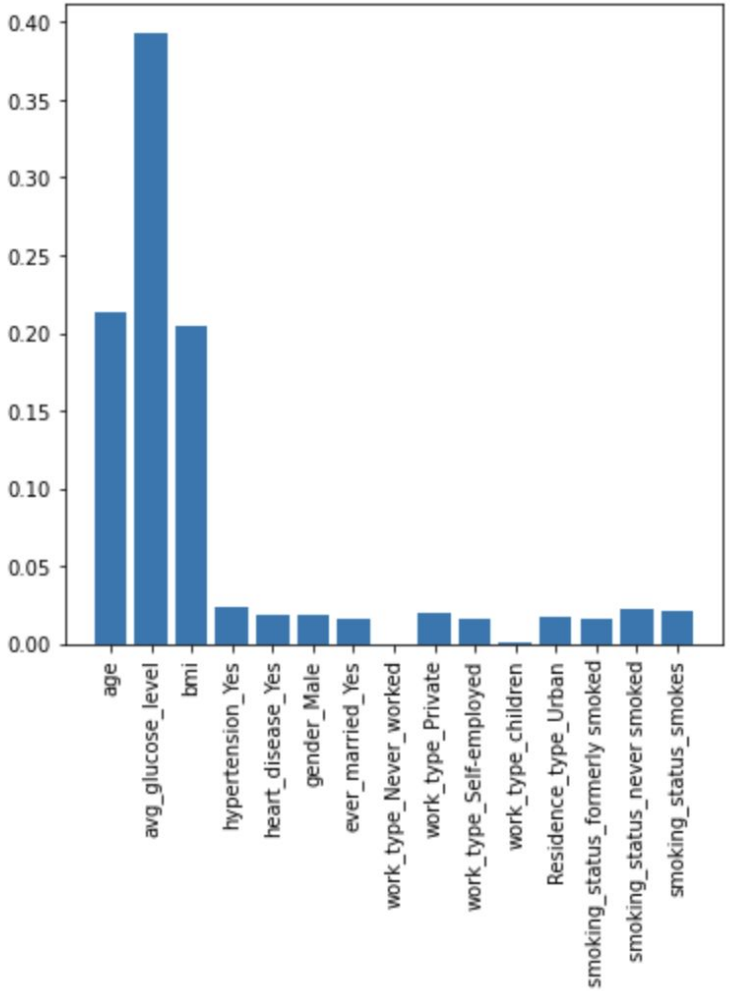


Predicted and Actual 1's are higher for the Random Forest predicted values than binary array

Features of Importance

- Used best model: Random Forest Classifier, to determine most influential features

Age, BMI and Average Glucose Level most influential features



Function Use of Model

- Created function to implement model on new data
- Able to predict with reasonable confidence about stroke occurrence of new individual
- Test Functionality here

```
#create function to determine prediction of new input
def predict_stroke(age, avg_glucose_level, bmi, hypertension,
                   heart_disease, gender, ever_married, work_type,
                   Residence_type, smoking_status, clf, mean_age, std_age, mean_avgglu,
                   std_avgglu, mean_bmi, std_bmi):
    age  = (age - mean_age)/std_age

    avg_glucose_level = (avg_glucose_level - mean_avgglu)/std_avgglu

    bmi = (bmi - mean_bmi)/std_bmi

    X_new = {'age': age, 'avg_glucose_level': avg_glucose_level, 'bmi': bmi, 'hypertension':hypertension,
             'heart_disease': heart_disease, 'gender' : gender, 'ever_married': ever_married, 'work_type': work_type
             'Residence_type':Residence_type, 'smoking_status': smoking_status}

    df2 = X.append(X_new, ignore_index = True)

    df3 = pd.get_dummies(df2, drop_first = True)

    df4 = pd.DataFrame(columns=list(X_test.columns))
    df4.loc[0] = df3.iloc[-1]

    y_pred = clf.predict(df4)

    proba = clf.predict_proba(df4)

    return y_pred, proba
```

```
#Here is where you fill in your information in this order: Age, Average Glucose Level, BMI, Hypertension,
#Heart Disease, Gender, Ever Married, Work Type, Residence Type, and Smoking status and example is provided

X = predict_stroke(34, 350, 35, 'Yes', 'No', 'Female', 'Yes', 'Private', 'Urban', 'never smoked', clf, mean_age, std
```

Assumptions and Limitations

- Assumed 5110 is representative of entire population
- Limitation of Demographic information
  - Residence Type only had 2 options
  - No race information

Conclusions

- 9 features directly attribute to stroke occurrence
- Out of the 4 Supervised learning models, Random Forest Classifier performed best
- With 75% training and 25% test data - the best recall score was 0.60
- With more individual data as demographic data from each individual, the model can be improved