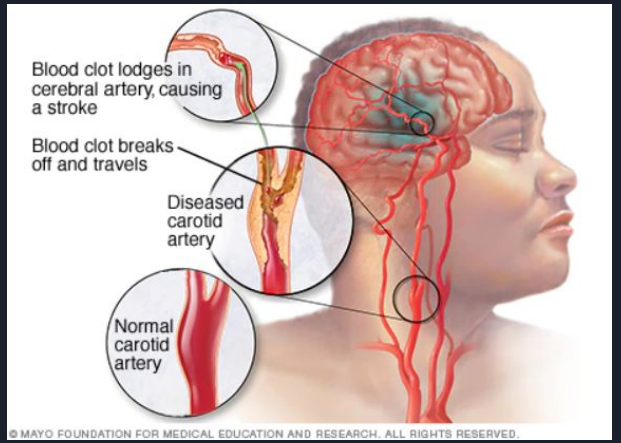
Predicting Stroke Based on Risk Factors

Springboard Data Science Track Capstone Project
Ngoc Tran: Data Scientist Student, Pharm D.
March 8, 2021 Cohort

What is a stroke?

- Occurs when blood supply to brain is interrupted
- Prevents brain tissue from getting oxygen and nutrients
- Brain cells die within minutes
- Medical emergency

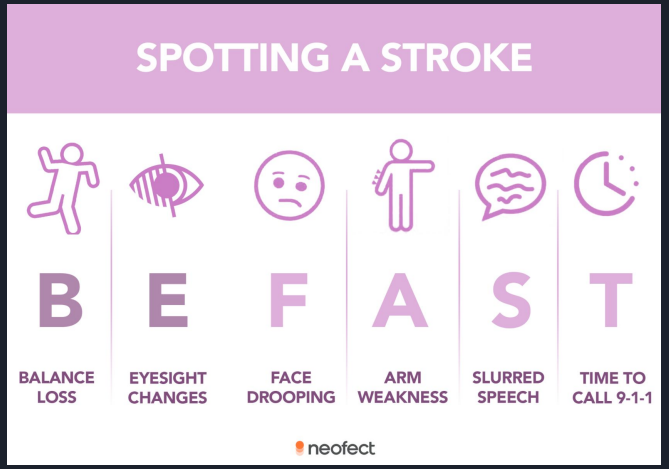


Why Should You Care?

- According to the World Health Organization (WHO), stroke is 2nd leading cause of death
  - 11% of total deaths globally
  - Usually affects older adults, but a stroke can happen to anyone
- Average cost of hospitalization: ~$20,000
- Life after a stroke can be profoundly debilitating:
  - Loss of mobility and cognitive function
  - Rehab and physical therapy
  - Overall lower quality of life, leading to depression

Problem

- Resources mainly emphasize how to detect a stroke for emergency treatment
- Healthcare system focuses on second and tertiary prevention, where patient already had the stroke



What if the focus shifted towards stroke prevention?

- Reduce cost of treatment by avoiding hospitalization
- Improve patients' quality of life

How?

- Develop machine learning model to predict if patient will have a stroke
  - Based on risk factors:
    - Gender, age, presence of disease, BMI, average glucose level, smoking status, etc
- Test data against various machine learning algorithms
- Tune the models to determine the best predictive model and most important attributes

- Identify patients at high risk of getting a stroke and plan intervention
  - Better eating habits
  - Incorporate daily exercise
  - Appropriate medication regimen

The Data

- Dataset from Kaggle
- 12 attributes
  - id: unique identifier
  - gender: classified as “Male”, “Female”, or “Other”
  - age: age of the patient
  - hypertension: 0 for absence of hypertension; 1 for presence of hypertension
  - heart_disease: 0 for absence of heart disease, 1 for presence of heart disease
  - ever_married: classified as “No” or “Yes”
  - work_type: classified as "children", "Govt_jov", "Never_worked", "Private" or "Self-employed"
  - Residence_type: classified as "Rural" or "Urban"
  - avg_glucose_level: average glucose level in blood
  - bmi: body mass index
  - smoking_status: "formerly smoked", "never smoked", "smokes" or "Unknown"
    - "Unknown" in smoking_status means that the information is unavailable for this patient
  - stroke: 1 if the patient had a stroke or 0 if not

- Total 4,909 entries after dropping missing data from BMI column

Exploratory Data Analysis

- Dataset had a lot more patients that did not suffer from a stroke (4,700) than patients that did have a stroke (209)
- More women patients recorded
- Majority of patients were healthy with no heart complications
    - No hypertension or heart disease

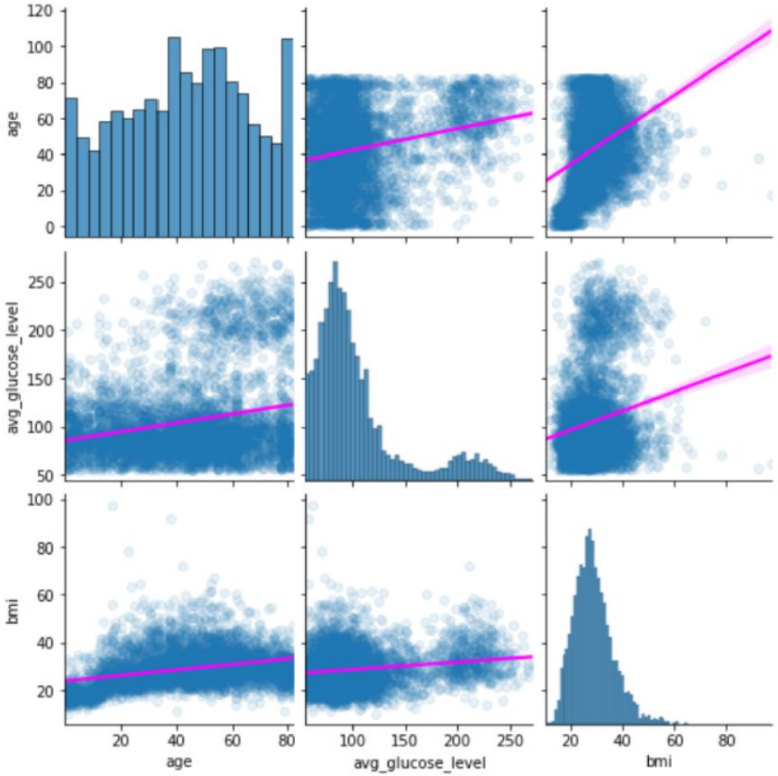


Figure 3: A pairwise scatter plot between the numerical features. It seems as age increases, so does the average glucose level and BMI.

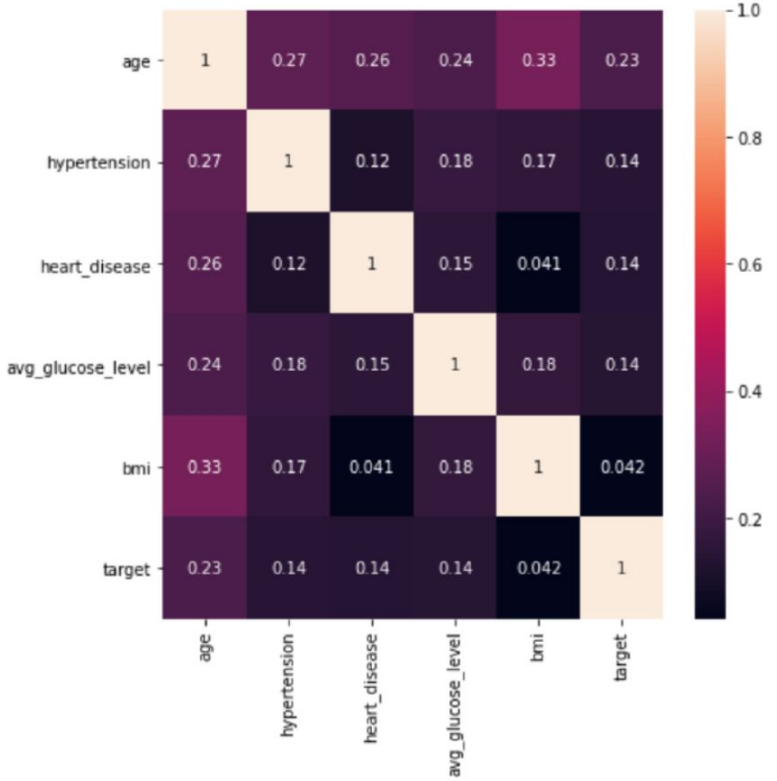


Figure 4: A heat map to analyze the correlation between the various attributes. It can be noted that age has the highest positive correlation (0.23) with our target variable (stroke).

Pairwise plot (left) and heat map (right) shows that many features were weakly correlated but none were so strongly correlated as needed to be removed

Model Evaluation

Model	Accuracy(%)
GaussianNB	19.35
BernoulliNB	94.06
LogisticRegression	96.04
RandomForestClassifer	96.01
DecisionTreeClassifier	92.43
KNeighborsClassifier	96.04
SVC	96.01

Table 1: The accuracy of the models that were tested. LogisticRegression, RandomForestClassifier, KNeighborsClassifier, and SVC all had accuracies over 95%.

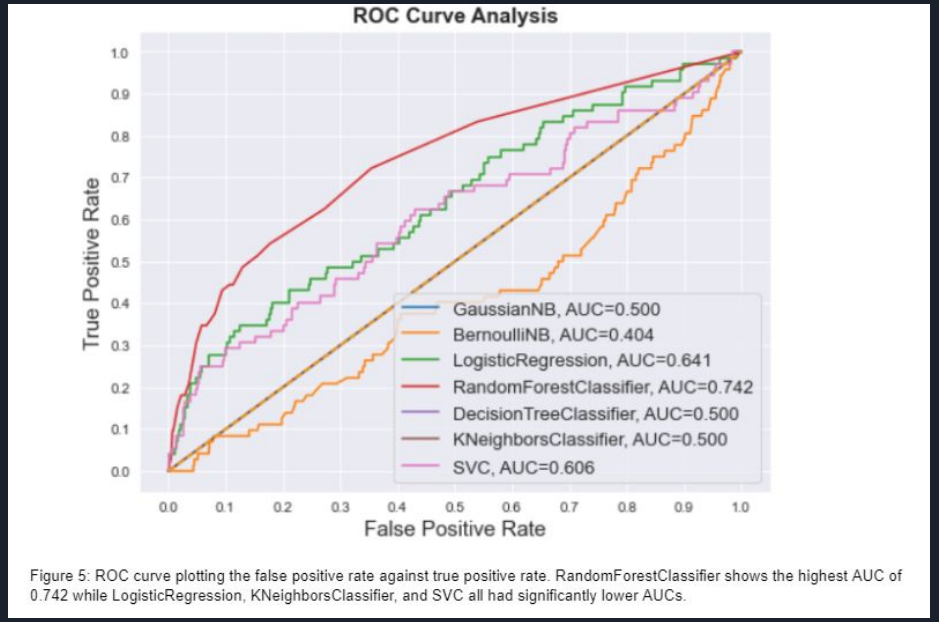
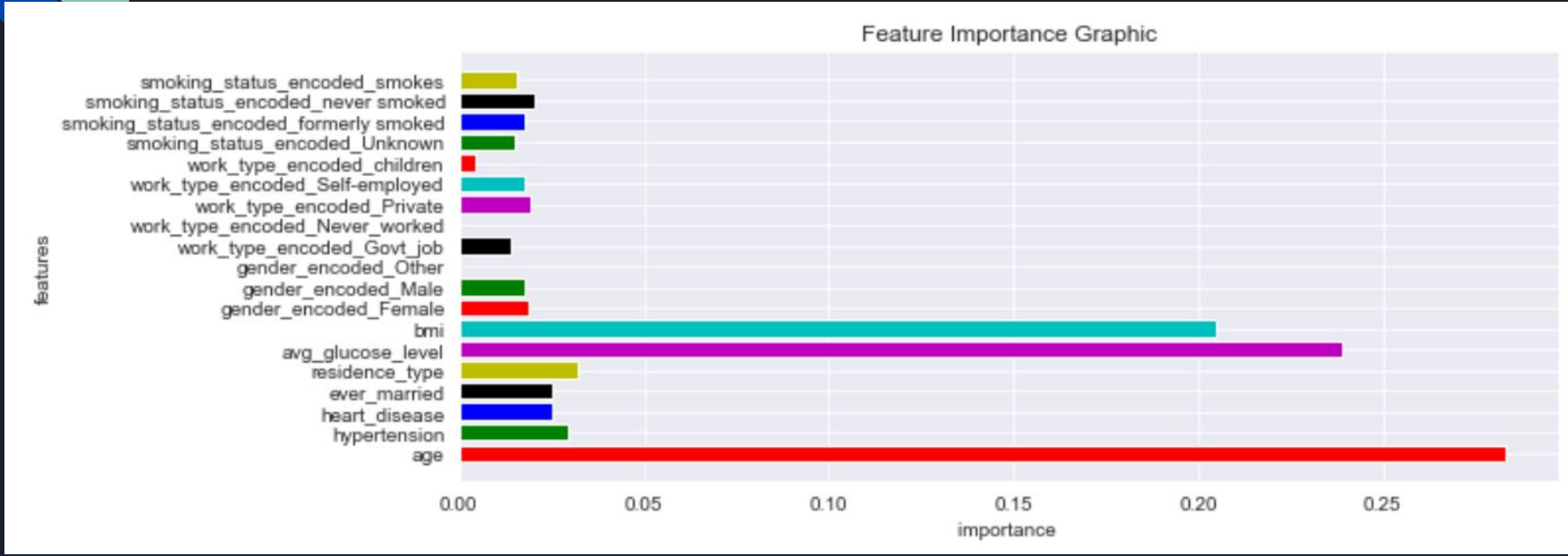


Figure 5: ROC curve plotting the false positive rate against true positive rate. RandomForestClassifier shows the highest AUC of 0.742 while LogisticRegression, KNeighborsClassifier, and SVC all had significantly lower AUCs.

GridSearchCV used to select best hyperparameters and fit on the seven models

LogisticRegression, RandomForestClassifier, KNeighborsClassifier, and SVC all showed accuracies over 95%, but ROC curve confirmed RandomForestClassifier had the highest AUC

Best Model: RandomForestClassifier



Top three important features: age, avg_glucose_level, and BMI

How can it be applied?

- Routine check up with PCP
  - Using model, PCP sees you are at risk of getting a stroke
- Plan intervention based on risk factors
  - Age: cannot be controlled
  - Uncontrolled blood glucose and BMI:
    - Lifestyle modifications to reduce blood sugar and weight to healthier range
  - Check medication
    - On proper regimen and dosage?
    - Is patient complaint with medications?
- Educate and advocate for healthier habits!

Takeaways

- Best predictive model: RandomForestClassifier
- Top three important features:
  - age
  - avg_glucose_level
  - bmi
- Attributes are readily accessible
  - Collected at doctors' visits
  - No need for invasive testing or extensive lab work
- Model highly useful for primary prevention of a stroke
  - Optimal patient care
  - Reduce treatment costs
  - Improve overall quality of life