The association of C-reactive protein and incidence of stroke in African American population of Jackson Heart Study

Jun Pan, Ritesh Lohiya, Brian Liles

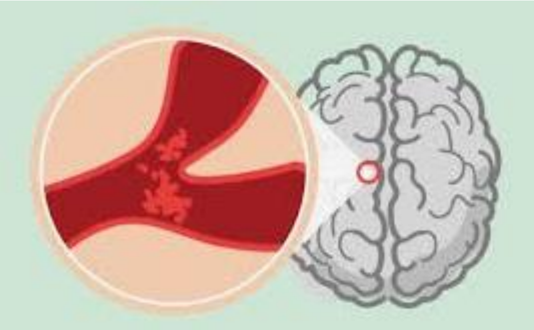
Specific Aims:

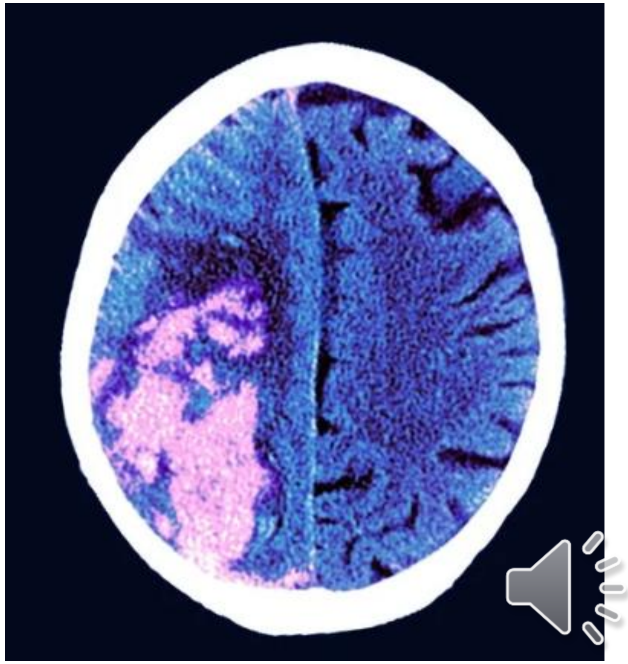
- Part 1. Whether C-reactive protein plays an important role in developing stroke incidence?
- Kaplan-Meier, Cox and Weibull survival analysis

- Part 2. Build machine learning models to predict stroke
- Logistic regression, Logistic regression w SGD, KNN, Random Forest and XGBoost.

Introduction

- Roughly 700,000 strokes occur each year in the United States;
- Stroke is the third leading cause of death;
- The leading cause of neurologic disability;
- It is highest in the elderly;
- 1/3 of strokes occurs in people < 65 years.





Jackson Heart Study

- single-site,
- prospective cohort study
- risk factors of heart disease
- adult African Americans (21-94).
- 5,301 African Americans,
- residing in a three-county area surrounding the city of Jackson, MS.

TRANS DATA
~50%, 2,653

Framingham
1 site, 5,209
1948

ARIC
4 sites, 15,729
1987

JHS
1 site, 5,301
1997

Inclusion and exclusion

Inclusion: all the participants with information in visit 1 and stroke incidence status record.

Exclusion: participants with missing information in hs-CRP and stroke incidence status.

Finally, 2,472 out of 2,653 were chosen in our study.

Table 1 “stroke incidence” database, table 2 v1 database
After the merge, we have 2653 rows and 211 variables

Variables

1 Design, Study-Level and Other Items
- Date of Visit
- Days Since Visit 1
- Years Since Visit 1
- Shared-ARIC / JHS-Only
- JHS Recruitment Type
- Outside of Original Target Enrollment Age
- Fasting Time (hours)

2 Demographics
- Age in Years
- Year of Birth
- Month of Birth
- Participant Sex
- Male Indicator
- Menopause Status
- Alcohol drinking in the past 12 months (Y/N)
- Average number of drinks per week
- Self-Reported Cigarette Smoking Status
- Self-Reported History of Cigarette Smoking

3 Anthropometrics
- Weight (kg)
- Height (cm)
- Body Mass Index (kg/m2)
- Waist Circumference (cm)
- Hip Circumference (cm)
- Neck Circumference (cm)
- Calculated Body Surface Area (m2)

4 Medications
- Medication Accountability
- Blood Pressure Medication Status (Y/N)
- Self-Reported Blood Pressure Medication Status (Y/N)
- Diabetic Oral Medication Status (Y/N)
- Diabetic Insulin Medication Status (Y/N)
- Diabetes Medication Type
- Diabetic Medication Status (Y/N)
- Statin Medication Status (Y/N)
- HRT Medication Status (Y/N)
- Self Reported HRT Medication Status (Y/N)
- Self Reported Current HRT Medication Status (Y/N)
- Beta Blocker Medication Status (Y/N)
- Calcium Channel Blocker Medication Status (Y/N)
- Diuretic Medication Status (Y/N)
- Antiarrhythmic Medication Status (Y/N)

5 Hypertension
- Systolic Blood Pressure (mmHg)
- Diastolic Blood Pressure (mmHg)
- JNC 7 BP Classification
- Hypertension Status
- Ankle Brachial Index

6 Diabetes
- Fasting Plasma Glucose Level (mg/dL)
- Fasting Plasma Glucose Categorization
- NGSP Hemoglobin HbA1c (%)
- NGSP Hemoglobin HbA1c (%) Categorization
- IFCC Hemoglobin HbA1c in SI units (mmol/mol)
- IFCC Hemoglobin HbA1c in SI units (mmol/mol) Categorizatio
- Fasting Insulin (Plasma IU/mL)
- HOMA-B
- HOMA-IR
- Diabetes Status (ADA 2010)
- Diabetes Categorization

7 Lipids
- Fasting LDL Cholesterol Level (mg/dL)
- Fasting LDL Categorization
- Fasting HDL Cholesterol Level (mg/dL)
- Fasting HDL Categorization
- Fasting Triglyceride Level (mg/dL)
- Fasting Triglyceride Categorization
- Fasting Total Cholesterol (mg/dL)

8 Biospecimens
- High Sensitivity C-Reactive Protein (Serum mg/dL)
- e-Selectin (Serum ng/mL)
- p-Selectin (Plasma ng/mL)
- Endothelin-1 (Serum pg/mL)
- Concentration of Cortisol Levels (Serum ug/dL)
- Renin Activity RIA (Plasma ng/mL/hr)
- Renin Mass IRMA (Plasma pg/mL)
- Concentration of Aldosterone (Serum ng/dL)
- Concentration of Leptin (Serum ng/mL)
- Concentration of Adiponectin (Plasma ng/mL)
- Concentration of Cystatin C (Serum mg/L)

9 Renal
- CC Calibrated Serum Creatinine (mg/dL)
- IDMS Tracebale Serum Creatinine (mg/dL)
- eGFR MDRD
- eGFR CKD-Epi
- 24-hour urine creatinine (g/24hr)
- Random spot urine creatinine (mg/dL)
- Random spot urine albumin (mg/dL)
- 24-hour urine albumin (mg/24hr)
- Self-reported dialysis
- Self-reported duration on dialysis (years)
- Chronic Kidney Disease History

10 Respiratory
- Physician-Diagnosed Asthma
- Successful Spirometry Maneuvers
- Forced Vital Capacity (L)
- Forced Expiratory Volume in 1 Second (L)
- Forced Expiratory Volume in 6 Seconds (L)
- FEV1 % Predicted
- FVC % Predicted

11 Echocardiogram
- Left Ventricular Mass (g) from Echo
- Left Ventricular Mass Indexed by Height(m)^2.7
- Left Ventricular Hypertrophy

12 Electrocardiogram
- Conduction Defect
- Anterior QnQs Major Scar
- Anterior QnQs Minor Scar
- Anterior Repolarization Abnormality
- Anterior ECG defined MI
- Posterior QnQs Major Scar
- Posterior QnQs Minor Scar
- Posterior Repolarization Abnormality
- Posterior ECG defined MI
- Anterolateral QnQs Major Scar
- Anterolateral QnQs Minor Scar
- Anterolateral Repolarization Abnormality
- Anterolateral ECG defined MI

13 CT Imaging
- Visceral Adipose Tissue (cm^3)
- Subcutaneous Adipose Tissue (cm^3)
- Coronary Artery Calcium Score
- Abdominal Aorto-iliac Calcium Score
- Presence of Coronary Artery Calcification
- Presence of Aortic Artery Calcification

14 Stroke History
- History of Speech Loss
- History of Sudden Loss of Vision
- History of Double Vision
- History of Numbness
- History of Paralysis
- History of Dizziness
- History of Stroke

15 CVD History
- Self-Reported History of MI
- Self-Reported history of Cardiac Procedures
- Coronary Heart Disease Status/History
- Self-Reported history of Carotid Angioplasty
- Cardiovascular Disease History
- Heart Failure History

16 Healthcare Access
- Public Insurance Status
- Medicaid Insurance Status
- Medicare Insurance Status
- VA/Champus Insurance Status
- Health Insurance Type
- Visit 1 Health Insurance Status
- Public Insurance Status
- Public Insurance Type
- Public or Private Insurance

17 Psychosocial
- Family Income Classification
- Income Status
- Occupational Status
- Education Attainment Categorization
- High School Graduate
- Everyday Discrimination Experiences
- Major Life Events Discrimination
- Discrimination Burden

19 Nutrition
- 25(OH) Vitamin D2 (ng/mL)
- 25(OH) Vitamin D3 (ng/mL)
- ep-25(OH) Vitamin D3 (ng/mL)
- Dark-green Vegetables
- Eggs
- Fish

20 Environmental
- Fake census tract ID
- Median Household income in Census Tract
- % below poverty in Census Tract
- % black non-hispanic in Census Tract
- % white non-hispanic in Census Tract
- Census Tract SES (PC2 score)
- Census Tract SES score (Diez-Roux 1990)
- Neighborhood Problems (age, sex adj.)
- Neighborhood Social Cohesion (age, sex adj.)

21 Genetics
- Sickle Cell trait / disease (rs334)
- Sickle Cell (rs334)
- APOL1 G1 Risk Allele from SNPs rs73885319 and rs60910145
- APOL1 G2 risk allele from indel rs71785313
- APOL1 CVD risk genotype
- Duffy blood group antigen (rs2814778)
- PCSK9-C679X Low density lipoprotein cholesterol level quantit
- SerineTyrosine Substitution at AA1103 (rs7626962)
- Hemoglobin C (HbC) locus (rs33930165)

22 Physical Activity
- Sport Index
- Home/Yard Index
- Active Living Index

Part I: Whether CRP is associated with Stroke

- 1. Apply a systematic method for imputing the missing entries in the dataset.
- 2. Select the relevant feature subset based on an automatic procedure.
- 3. Apply survival models to evaluate the prediction performance.

```
Observations: 2,472
Variables: 22
$ subjid        <int> 115, 2307, 1668, 1616, 1753, 2709, 2032, 1055, 2430, 670, 21, 1696, 777, 1078, 484, 142, 123...
$ stroke        <fct> No, Yes, No, Yes, No, No, No, No, No, No, No, No, No, No, No, No, No, No, No, No, No, No, No...
$ years         <dbl> 13.924709, 12.492813, 13.163587, 9.305955, 13.694730, 14.009582, 13.820671, 13.779603, 12.99...
$ days          <int> 5086, 4563, 4808, 3399, 5002, 5117, 5048, 5033, 4746, 4691, 4708, 2923, 5024, 5048, 4993, 49...
$ age           <dbl> 62.1, 75.2, 74.8, 60.8, 60.0, 63.2, 60.8, 69.8, 69.2, 79.9, 62.4, 70.4, 62.6, 72.6, 73.1, 70...
$ sex           <fct> Female, Female, Female, Female, Male, Female, Male, Female, Female, Female, Female, Female, ...
$ currentSmoker <fct> No, No, No, Yes, No, No, Yes, No, Yes, No, No, No, No, No, No, No, No, No, No, No, No, No, N...
$ weight        <dbl> 95.0, 57.0, 91.1, 92.0, 95.0, 87.0, 104.0, 90.0, 78.5, 120.0, 111.0, 94.0, 66.0, 98.0, 72.0,...
$ height        <dbl> 166, 153, 177, 169, 178, 167, 174, 166, 158, 169, 167, 174, 168, 176, 163, 156, 183, 161, 16...
$ waist         <dbl> 113, 83, 103, 107, 96, 97, 119, 108, 103, 138, 116, 104, 78, 104, 107, 102, 118, 106, 90, 74...
$ BMI           <dbl> 34.48, 24.35, 29.08, 32.21, 29.98, 31.20, 34.35, 32.66, 31.45, 42.02, 39.80, 31.05, 23.38, 3...
$ sbp           <dbl> 102.74, 122.91, 154.09, 110.08, 121.08, 129.33, 126.58, 119.25, 122.00, 114.66, 153.17, 115....
$ dbp           <dbl> 50.97, 61.77, 65.92, 68.41, 64.26, 65.92, 80.03, 67.58, 67.58, 66.75, 74.22, 60.11, 79.20, 7...
$ HTN           <fct> Yes, No, Yes, Yes, No, No, Yes, Yes, No, No, Yes, Yes, No, Yes, Yes, Yes, No, Yes, Yes, No, ...
$ HbA1c         <dbl> 7.9, 4.2, 6.1, 4.7, 9.0, 6.3, 6.6, 6.7, 5.0, 6.3, 5.7, 5.3, 5.2, 5.4, 5.5, 8.6, 8.2, 5.8, 6....
$ ldl           <dbl> 190, 160, 46, 101, 126, 200, 106, NA, 181, 88, 242, 108, 86, 96, 75, 113, 120, 99, 101, 89, ...
$ hdl           <dbl> 65, 71, 35, 44, 51, 49, 39, NA, 112, 53, 50, 48, 59, 62, 95, 57, 46, 52, 40, 70, 65, 47, 69,...
$ trigs         <dbl> 136, 69, 76, 41, 60, 82, 141, NA, 63, 68, 106, 61, 52, 47, 84, 87, 131, 51, 93, 47, 57, 96, ...
$ totchol       <dbl> 282, 245, 96, 153, 189, 265, 173, NA, 306, 155, 313, 168, 155, 167, 187, 187, 192, 161, 160,...
$ HSCRP         <dbl> 0.633, 0.172, 0.157, 0.197, 0.113, 0.819, 0.163, 0.146, 0.062, 0.951, 0.158, 0.365, 0.388, 0...
$ Afib          <fct> No, No, Yes, No, No, No, No, No, No, No, No, No, No, No, No, No, No, No, No, No, No, No, No,...
$ strokeHx      <fct> No, No, No, No, No, No, No, No, No, No, No, No, No, No, No, No, No, No, No, No, No, No, No, ...
```

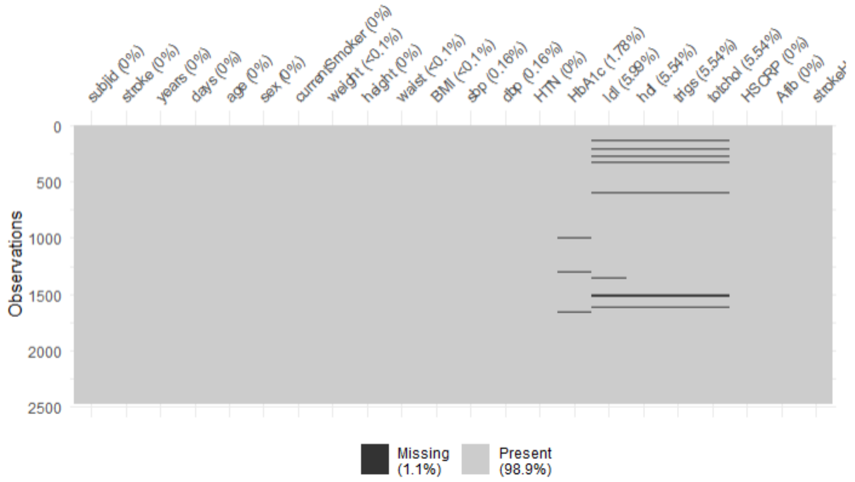
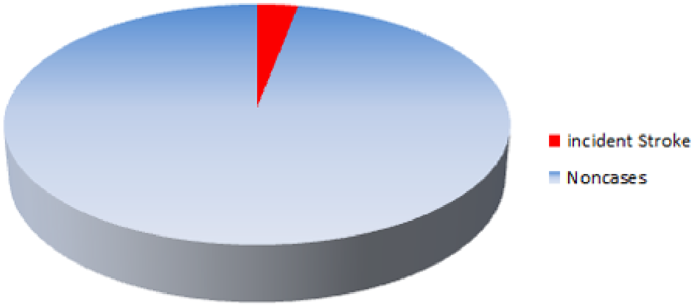


Table 1. Baseline Characteristics in Cases of Incident Ischemic Stroke and Non-cases

	Incident Stroke (n=76)	Non-cases (n=2396)
Age, y	62.37	53.86 ***
Female, %	56.58	62.52***
Current smoker, %	19.74	11.22
Weight, kg	89.36	91.58
Waist circumference, cm	101.99	100.72
BMI	31.92	31.19
Systolic blood pressure, mm Hg	132.68	125.75*
Diastolic blood pressure, mm Hg	75.68	75.95
Hypertension, %	76.32	51.62***
HbA1c,	6.39	5.85***
Total cholesterol, mg/dL	208.59	198.59
Triglycerides, mg/dL	111.96	103.62
HDL-C, mg/d	51.38	51.66
LDL-C, mg/dL	133.72	126.12*
hs-CRP, mg/L	0.52	0.49

% of Stroke in Population

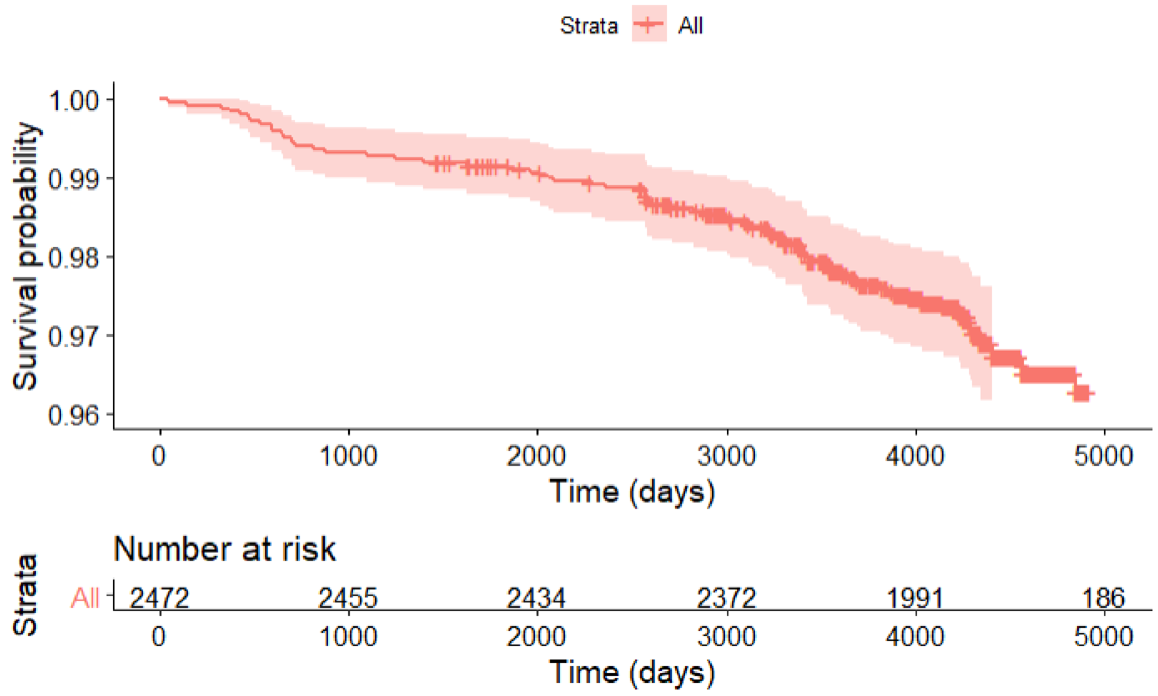


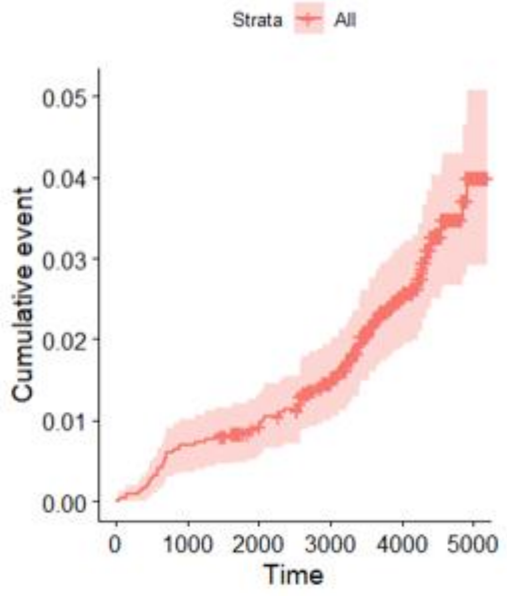
There were 76 (3%) incidence of stroke in general study population.

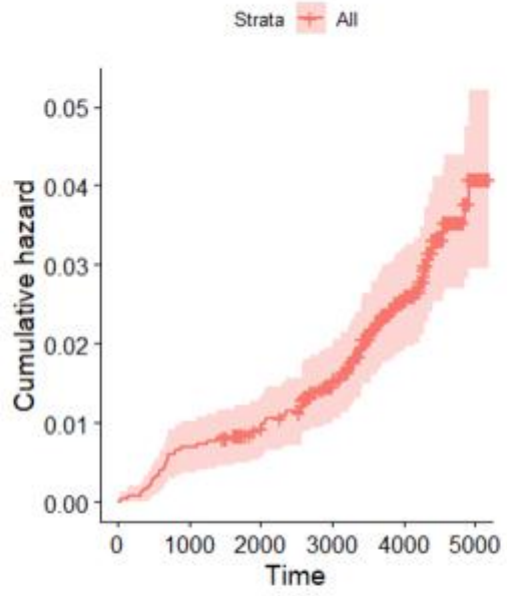
```
#Regular Kaplan-Meier plot
#reference:https://rpubs.com/alecri/258589
```{r}
fit_km <- survfit(Surv(days, stroke) ~ 1, data = mdata)
print(fit_km, print.rmean = TRUE)
```
```

```
Call: survfit(formula = Surv(days, stroke) ~ 1, data = mdata)

        n    events    *rmean *se(rmean)    median   0.95LCL   0.95UCL
   2472.0      76.0    5121.0       10.4        NA        NA        NA
    * restricted mean with upper limit =  5205
```







In our study, there were 76 incident strokes with mean time to stroke 11.81 years.

Cox model 1, hs-CRP only

```
#Cox model 1
```{r}
cxmod <- coxph(Surv(days, stroke) ~ HSCRP, data = ndata)
coef(cxmod)
```
```

```
      HSCRP
0.05450608
```

```
```{r}
summary(cxmod)
```
```

```
Call:
coxph(formula = Surv(days, stroke) ~ HSCRP, data = ndata)

  n= 2472, number of events= 76

         coef exp(coef) se(coef)     z Pr(>|z|)
HSCRP 0.05451   1.05602  0.13696 0.398    0.691

      exp(coef) exp(-coef) lower .95 upper .95
HSCRP     1.056      0.947    0.8074     1.381

Concordance= 0.525  (se = 0.034 )
Likelihood ratio test= 0.15  on 1 df,   p=0.7
Wald test            = 0.16  on 1 df,   p=0.7
Score (logrank) test = 0.16  on 1 df,   p=0.7
```

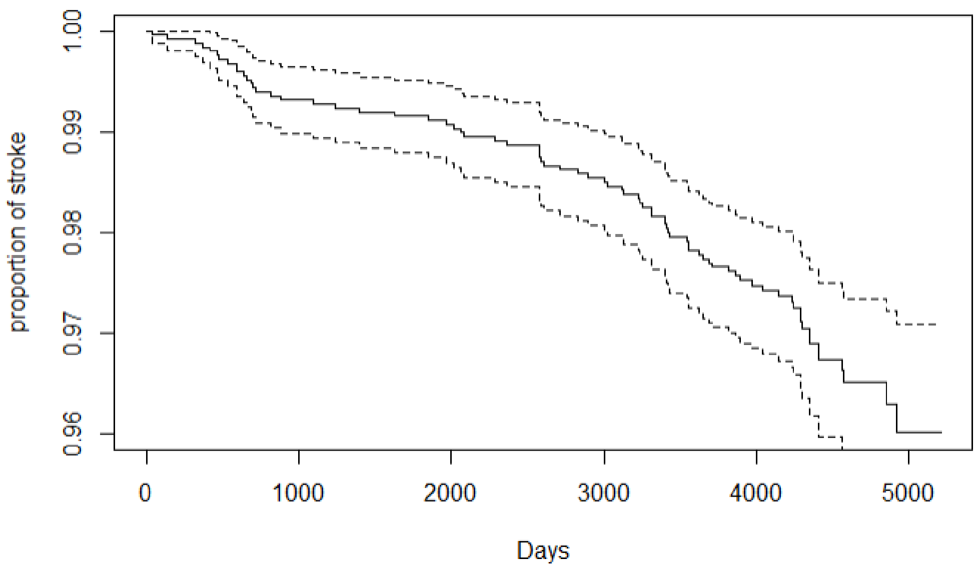
The weighted mean levels of CRP(0.52 vs 0.49mg/L) were no significant between stroke cases and non-cases.

```
```{r}
anova(cxmod)
```
```

```
Analysis of Deviance Table
 Cox model: response is Surv(days, stroke)
Terms added sequentially (first to last)

       loglik  Chisq Df Pr(>|Chi|)
NULL  -579.88
HSCRP -579.81 0.1464  1      0.702
```

```
```{r}
plot(survfit(cxmod), ylim=c(0.96, 1),xlab="Days",
ylab="proportion of stroke")
```
```



Cox model 2, adding gender and age besides hs-CRP

```
#cox model 2
```{r}
cxmod2 <- coxph(Surv(days, stroke) ~ HSCRP + sex + age, data = ndata)
coef(cxmod1)
```
```

```
     HSCRP    sexMale        age 
0.10413425 0.41599062 0.06687991 
```

```
```{r}
summary(cxmod2)
```
```

```
Call:
coxph(formula = Surv(days, stroke) ~ HSCRP + sex + age, data = ndata)

  n= 2472, number of events= 76 

           coef exp(coef) se(coef)     z Pr(>|z|)    
HSCRP   0.10413   1.10975  0.12358 0.843   0.3994    
sexMale 0.41599   1.51587  0.23720 1.754   0.0795 .  
age     0.06688   1.06917  0.01113 6.011 1.84e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

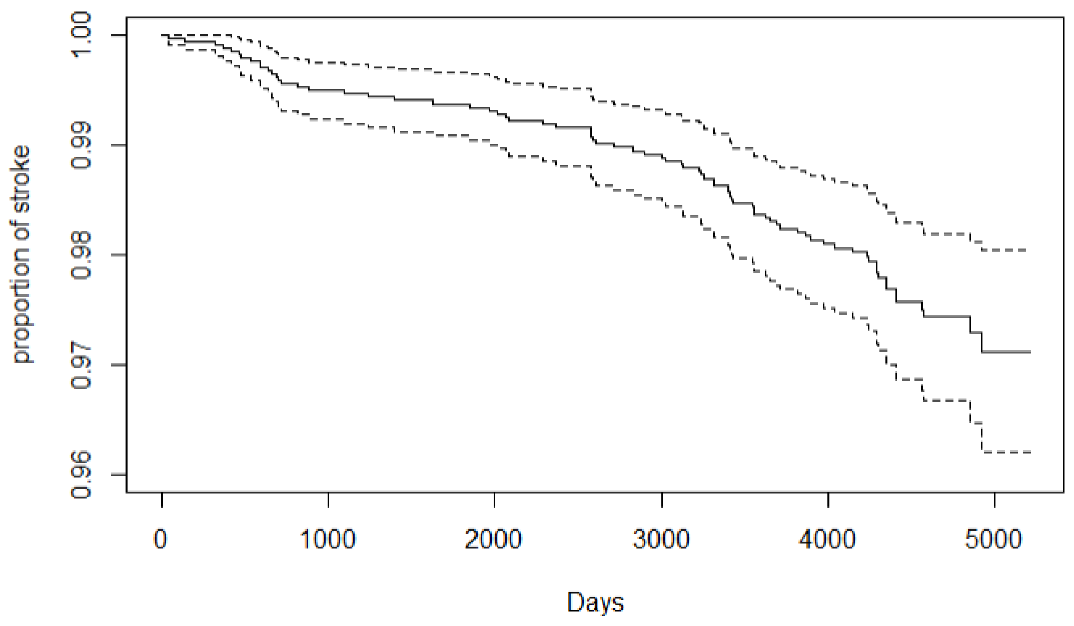
        exp(coef) exp(-coef) lower .95 upper .95
HSCRP       1.110     0.9011    0.8710     1.414
sexMale     1.516     0.6597    0.9523     2.413
age         1.069     0.9353    1.0461     1.093

Concordance= 0.706  (se = 0.03 )
Likelihood ratio test= 41.55  on 3 df,   p=5e-09
Wald test            = 38.02  on 3 df,   p=3e-08
Score (logrank) test = 39.25  on 3 df,   p=2e-08
```

```
```{r}
Anova(cxmod2)
```
```

```
Analysis of Deviance Table (Type II tests)
      LR Chisq Df Pr(>Chisq)    
HSCRP    0.590  1    0.44243    
sex      3.012  1    0.08265 .  
age     39.628  1  3.073e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
```{r}
plot(survfit(cxmod2), ylim=c(0.96, 1),xlab="Days",
ylab="proportion of stroke")
```
```



Model 3, Multivariate Cox regression analysis

```
#cox model 3
```{r}
cxmod3 <- coxph(Surv(days, stroke) ~ HSCRP + sex + age + currentSmoker + weight + height + waist + BMI + sbp + dbp + HTN +
HbA1c + ldl + hdl +trigs + totchol + Afib + strokeHx, data = ndata)
coef(cxmod1)
```
```

```
Call:
coxph(formula = Surv(days, stroke) ~ HSCRP + sex + age + currentSmoker +
    weight + height + waist + BMI + sbp + dbp + HTN + HbA1c +
    ldl + hdl + trigs + totchol + Afib + strokeHx, data = ndata)

  n= 2472, number of events= 76

                       coef  exp(coef)   se(coef)      z Pr(>|z|)
HSCRP             8.861e-02  1.093e+00  1.290e-01  0.687 0.492236
sexMale           3.125e-01  1.367e+00  3.599e-01  0.868 0.385258
age               5.957e-02  1.061e+00  1.447e-02  4.116 3.85e-05 ***
currentSmokerNo   1.499e+01  3.236e+06  2.182e+03  0.007 0.994518
currentSmokerYes  1.595e+01  8.451e+06  2.182e+03  0.007 0.994167
weight           -5.387e-02  9.476e-01  5.799e-02 -0.929 0.352949
height            4.992e-02  1.051e+00  6.359e-02  0.785 0.432521
waist             1.065e-02  1.011e+00  1.605e-02  0.664 0.506939
BMI               1.208e-01  1.128e+00  1.620e-01  0.746 0.455808
sbp               6.813e-03  1.007e+00  8.764e-03  0.777 0.436973
dbp              -5.481e-03  9.945e-01  1.651e-02 -0.332 0.739936
HTNYes            5.446e-01  1.724e+00  3.033e-01  1.796 0.072541 .
HbA1c             2.131e-01  1.237e+00  7.887e-02  2.701 0.006909 **
ldl              -9.102e-02  9.130e-01  2.698e-02 -3.373 0.000743 ***
hdl              -1.008e-01  9.041e-01  2.938e-02 -3.431 0.000601 ***
trigs            -2.082e-02  9.794e-01  6.931e-03 -3.004 0.002664 **
totchol           9.746e-02  1.102e+00  2.707e-02  3.601 0.000318 ***
AfibNo            1.460e+01  2.191e+06  4.901e+03  0.003 0.997623
AfibYes          -1.223e+00  2.943e-01  7.308e+03  0.000 0.999866
strokeHxYes              NA         NA  0.000e+00     NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

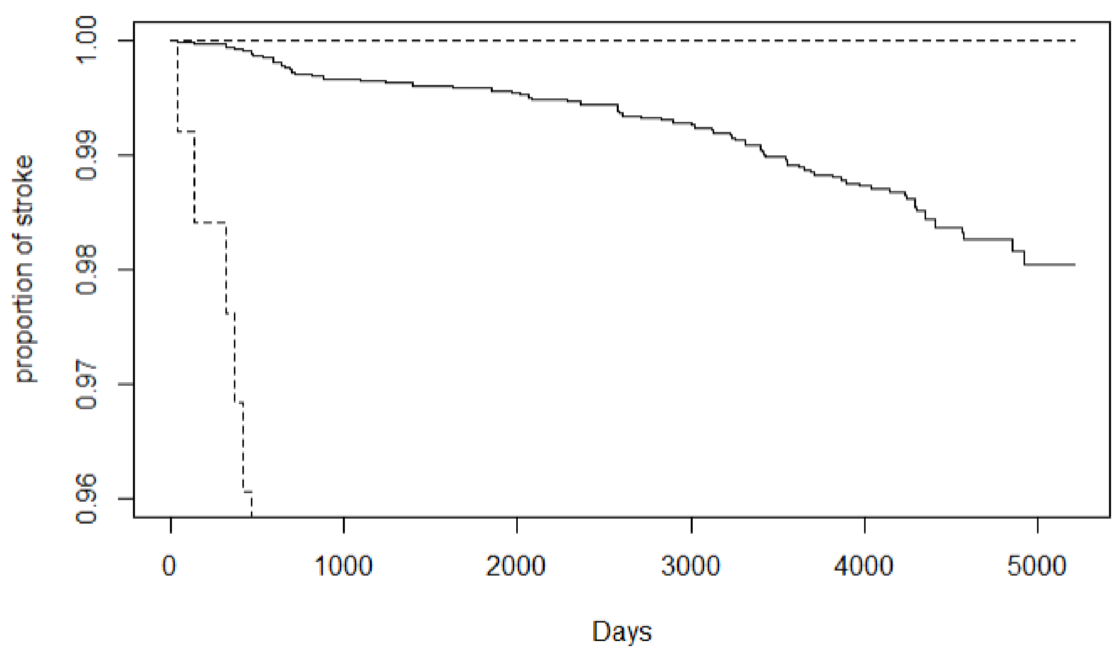
```
                 exp(coef) exp(-coef) lower .95 upper .95
HSCRP            1.093e+00  9.152e-01    0.8485    1.4071
sexMale          1.367e+00  7.316e-01    0.6751    2.7672
age              1.061e+00  9.422e-01    1.0317    1.0919
currentSmokerNo  3.236e+06  3.090e-07    0.0000       Inf
currentSmokerYes 8.451e+06  1.183e-07    0.0000       Inf
weight           9.476e-01  1.055e+00    0.8458    1.0616
height           1.051e+00  9.513e-01    0.9280    1.1907
waist            1.011e+00  9.894e-01    0.9794    1.0430
BMI              1.128e+00  8.862e-01    0.8214    1.5503
sbp              1.007e+00  9.932e-01    0.9897    1.0243
dbp              9.945e-01  1.005e+00    0.9629    1.0272
HTNYes           1.724e+00  5.801e-01    0.9514    3.1236
HbA1c            1.237e+00  8.081e-01    1.0602    1.4443
ldl              9.130e-01  1.095e+00    0.8660    0.9626
hdl              9.041e-01  1.106e+00    0.8535    0.9577
trigs            9.794e-01  1.021e+00    0.9662    0.9928
totchol          1.102e+00  9.071e-01    1.0454    1.1624
AfibNo           2.191e+06  4.563e-07    0.0000       Inf
AfibYes          2.943e-01  3.398e+00    0.0000       Inf
strokeHxYes             NA         NA        NA        NA

Concordance= 0.774  (se = 0.025 )
Likelihood ratio test= 78.32  on 19 df,   p=4e-09
Wald test            = 57.44  on 19 df,   p=1e-05
Score (logrank) test = 71.45  on 19 df,   p=5e-08
```

```
 Cox model: response is Surv(days, stroke)
Terms added sequentially (first to last)

               loglik   Chisq Df  Pr(>|Chi|)
NULL           -579.88
HSCRP          -579.81  0.1464  1    0.702018
sex            -578.92  1.7756  1    0.182685
age            -559.11 39.6276  1  3.073e-10 ***
currentSmoker -554.85  8.5130  2    0.014172 *
weight         -554.84  0.0097  1    0.921565
height         -554.83  0.0247  1    0.875185
waist          -554.31  1.0453  1    0.306594
BMI            -554.15  0.3214  1    0.570792
sbp            -552.72  2.8544  1    0.091127 .
dbp            -552.68  0.0777  1    0.780382
HTN            -550.97  3.4241  1    0.064253 .
HbA1c          -547.00  7.9475  1    0.004815 **
ldl            -545.11  3.7630  1    0.052398 .
hdl            -545.08  0.0707  1    0.790362
trigs          -545.07  0.0103  1    0.919058
totchol        -541.06  8.0247  1    0.004614 **
Afib           -540.72  0.6829  2    0.710755
strokeHx       -540.72  0.0000  0    1.000000
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```{r}
plot(survfit(cxmod3), ylim=c(0.96, 1),xlab="Days",
ylab="proportion of stroke")
```



Model 4, using 1mg/dL as cut point of hs-CRP

```
```{r}
cxmod4 <- coxph(Surv(days, stroke) ~ CRPqt + sex + age + currentSmoker +  sbp +
                HTN + HbA1c + ldl + totchol , data = mdata)
coef(cxmod4)
```
```

```
Call:
coxph(formula = Surv(days, stroke) ~ CRPqt + sex + age + currentSmoker +
    sbp + HTN + HbA1c + ldl + totchol, data = mdata)

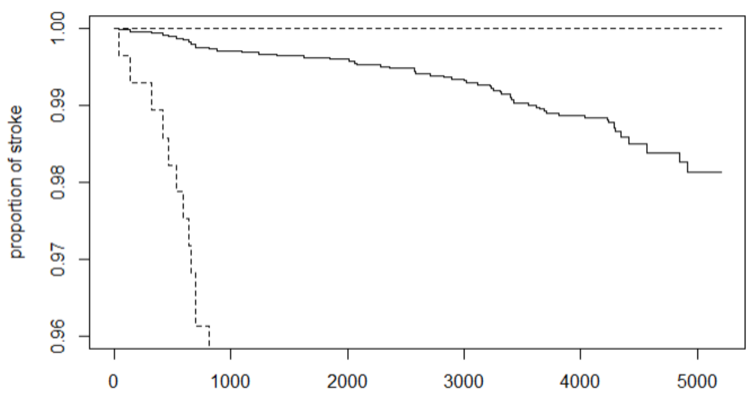
  n= 2280, number of events= 63
   (192 observations deleted due to missingness)

                   coef exp(coef) se(coef)      z Pr(>|z|)
CRPqtYes       0.815915  2.261243 0.346979  2.351  0.01870 *
sexMale        0.320773  1.378193 0.277564  1.156  0.24781
age            0.073614  1.076391 0.014117  5.215 1.84e-07 ***
currentSmoker  0.984338  2.676040 0.324475  3.034  0.00242 **
sbp            0.007622  1.007651 0.007837  0.973  0.33079
HTNYes         0.287658  1.333301 0.319568  0.900  0.36804
HbA1c          0.195205  1.215560 0.093162  2.095  0.03614 *
ldl            0.009410  1.009455 0.009010  1.044  0.29628
totchol       -0.002512  0.997491 0.008547 -0.294  0.76882
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
              exp(coef) exp(-coef) lower .95 upper .95
CRPqtYes         2.2612     0.4422    1.1455     4.464
sexMale          1.3782     0.7256    0.7999     2.375
age              1.0764     0.9290    1.0470     1.107
currentSmoker    2.6760     0.3737    1.4168     5.055
sbp              1.0077     0.9924    0.9923     1.023
HTNYes           1.3333     0.7500    0.7127     2.494
HbA1c            1.2156     0.8227    1.0127     1.459
ldl              1.0095     0.9906    0.9918     1.027
totchol          0.9975     1.0025    0.9809     1.014

Concordance= 0.774  (se = 0.028 )
Likelihood ratio test= 62.77  on 9 df,   p=4e-10
Wald test            = 55.23  on 9 df,   p=1e-08
Score (logrank) test = 57.57  on 9 df,   p=4e-09
```

```
```{r}
plot(survfit(cxmod4), ylim=c(0.96, 1),xlab="Days",
ylab="proportion of stroke")
```
```



Using 1mg/dL as cut point for hs-CRP, it has increased the significance to predictor the stroke. ($p <0.05$)

Weibull model: 1 mg/dl cut point of hs-CRP

```
```{r}
wbmod3 <- survreg(Surv(years,stroke) ~ CRPqt + sex + age + currentSmoker + weight + height + waist + BMI + sbp + dbp + HTN +
HbA1c + ldl + hdl +trigs + totchol + Afib + strokeHx,,data=mdata)
wbmod3
```
```

```
Call:
survreg(formula = Surv(years, stroke) ~ CRPqt + sex + age + currentSmoker +
    weight + height + waist + BMI + sbp + dbp + HTN + HbA1c +
    ldl + hdl + trigs + totchol + Afib + strokeHx, data = mdata)
                  Value Std. Error     z      p
(Intercept)    1.40e+01   9.29e+00  1.51 0.1317
CRPqtYes      -5.69e-01   2.91e-01 -1.95 0.0507
sexMale       -2.13e-01   3.03e-01 -0.70 0.4825
age           -5.13e-02   1.38e-02 -3.72 0.0002
currentSmoker -7.29e-01   2.70e-01 -2.70 0.0070
weight         2.18e-02   4.93e-02  0.44 0.6586
height        -2.07e-02   5.42e-02 -0.38 0.7029
waist         -1.20e-02   1.35e-02 -0.89 0.3751
BMI           -2.84e-02   1.40e-01 -0.20 0.8385
sbp           -8.18e-03   7.33e-03 -1.11 0.2649
dbp            8.76e-03   1.39e-02  0.63 0.5281
HTNYes        -2.94e-01   2.56e-01 -1.15 0.2504
HbA1c         -1.22e-01   7.99e-02 -1.52 0.1274
ldl           -7.22e-02   3.26e-01 -0.22 0.8250
hdl           -6.57e-02   3.26e-01 -0.20 0.8406
trigs         -1.28e-02   6.53e-02 -0.20 0.8450
totchol        6.72e-02   3.26e-01  0.21 0.8370
Afib           1.12e+01   2.47e+03  0.00 0.9964
strokeHxYes    0.00e+00   0.00e+00    NA     NA
Log(scale)    -2.69e-01   1.25e-01 -2.15 0.0312

Scale= 0.764
```

```
Weibull distribution
Loglik(model)= -404.4   Loglik(intercept only)= -436.5
        Chisq= 64.13 on 18 degrees of freedom, p= 4.3e-07
Number of Newton-Raphson Iterations: 18
n=2279 (193 observations deleted due to missingness)
```

Using Weibull Model, we reproduced the result of cox model 4.

Machine Learning Approach for Predicting the Stroke

Part Two

Machine Learning

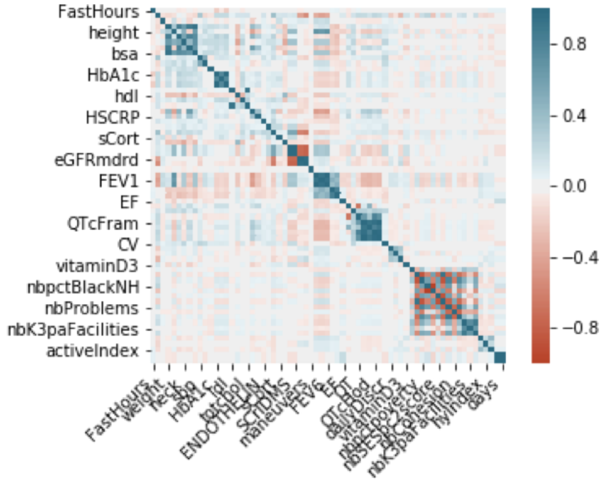
- Exploratory data analysis
- Data preprocessing
- Summary of data analysis and preprocessing
- Methods
- Experimental results and analysis

Exploratory Data Analysis

- Merged dataframe, N=2653
- Descriptive statistics were conducted (count, mean, std, min, max, interquartile range – 25%, 50%, 75%)
- Missing values were identified
- Variables with 90% of missing values were removed
- Variable count condensed from N=211 to N=179

Data Preprocessing

- Variables which lacked predictive power were removed
- Variable count, N=160
- Correlation test were conducted on remaining variables
- Highly correlated variables (> 0.95) were identified and removed in order to minimize the misinterpretation of results



	FastHours	age	alcw	weight	height	waist
FastHours	1.000000	0.008816	-0.014350	-0.036696	-0.035999	-0.051013
age	0.008816	1.000000	-0.092560	-0.125229	-0.099496	0.046320
alcw	-0.014350	-0.092560	1.000000	-0.045532	0.184359	-0.063263
weight	-0.036696	-0.125229	-0.045532	1.000000	0.326888	0.837003
height	-0.035999	-0.099496	0.184359	0.326888	1.000000	0.123167
waist	-0.051013	0.046320	-0.063263	0.837003	0.123167	1.000000
neck	-0.044317	-0.026400	0.077138	0.641124	0.491649	0.569092
BMI	-0.021493	-0.078677	-0.132961	0.874273	-0.160158	0.814977
bsa	-0.042955	-0.131332	0.023263	0.935955	0.630067	0.737550
sbp	-0.001290	0.332330	0.034398	0.088507	0.020427	0.130491
dbp	0.017757	-0.119559	0.093770	0.152248	0.200980	0.095396
abi	-0.003185	-0.072509	-0.018707	0.192165	0.105480	0.159892
HbA1c	-0.187534	0.199994	-0.050966	0.185276	-0.009872	0.258787
FPG	-0.058768	0.208006	-0.028290	0.129580	0.003369	0.189431
HbA1cIFCC	-0.187538	0.199994	-0.050969	0.185272	-0.009874	0.258784

Imbalanced Data

- Variables where an imbalanced ratio existed were identified and removed from the dataframe
- Imbalanced data when applied to machine learning has proven to be problematic
- Variable count N=62
- SMOTE (Synthetic Minority Oversampling Technique) applied to data

Example of Imbalanced data
Anterior Major Scar Variable

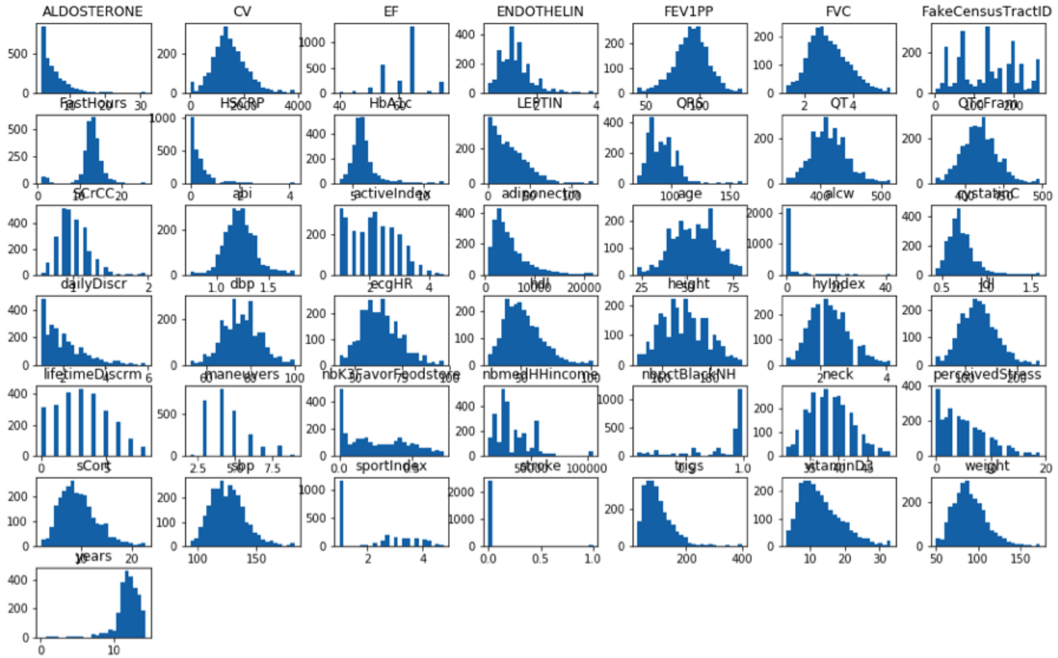
Absent: 2634
Present: 19

Example of Imbalanced data (Categorical)
Insurance Type Variable

Private Only: 1520
Uninsured: 332
Private & Medicare: 226
Medicare Only: 173
Medicare & Medicaid: 105

Graphical Summaries - Histograms

- To further explore the makeup of the dataframe, graphical summaries were created to provide an in-depth look at the data
- Histograms granted a detailed look at the distribution of variables
- Mostly normal, a few variables had skewness to the left or right

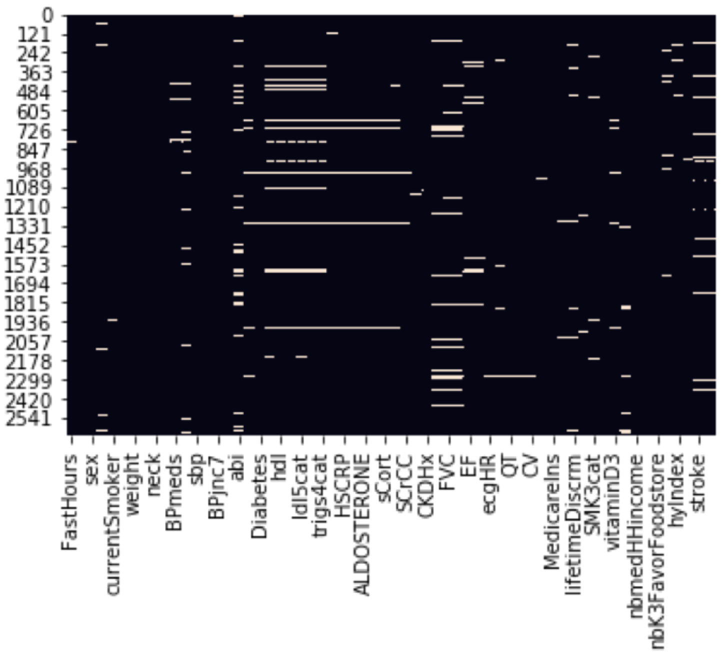



```
df1.skew()

FastHours      -1.137223
age            -0.120855
alcw            5.316831
weight          0.920163
height          0.213469
neck            0.323641
sbp             0.714287
dbp             0.104816
abi             0.457533
HbA1c           2.403734
ldl             0.439788
hdl             0.952373
trigs           2.076744
LEPTIN          1.396028
HSCRP           2.936418
ENDOTHELIN      1.346687
ALDOSTERONE     2.467425
cystatinC       1.953873
sCort           0.825732
adiponectin     1.879255
SCrCC           1.172042
```

Graphical Summaries - Heatmaps

- In order to detect missing values graphically, a heat map was plotted
- Remaining values which had missing values were updated utilizing the imputation method where median values were instituted

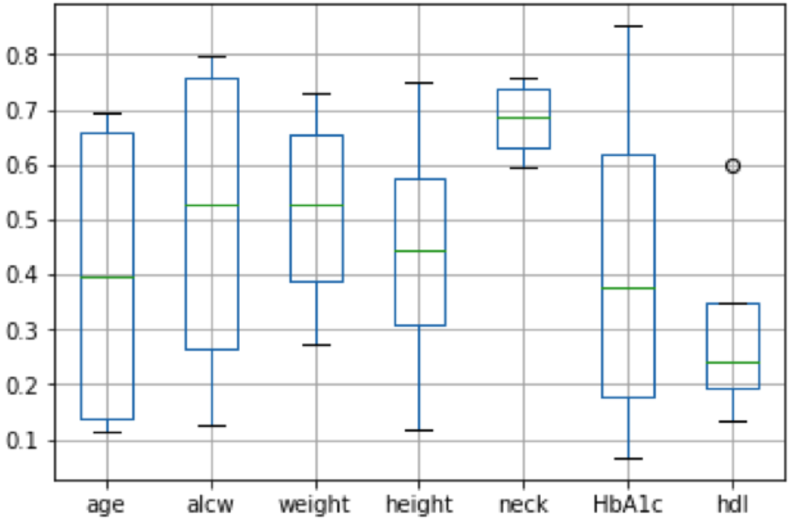


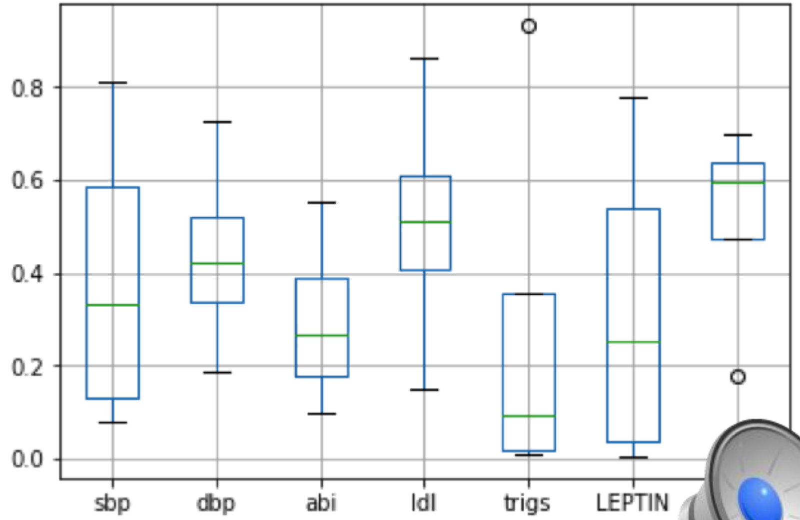
FastHours	6
age	0
sex	0
alcw	71
currentSmoker	18
everSmoker	4
weight	4
height	3
neck	4
OBESITY3cat	4
BPmeds	20
diureticMeds	156
sbp	4
dbp	4
BPjnc7	4
HTN	0
abi	244
HbA1c	88
Diabetes	21
ldl	209
hdl	197
trigs	197
ldl5cat	209
hdl3cat	197
trigs4cat	197
LEPTIN	58
HSCRP	43
ENDOTHELIN	44
ALDOSTERONE	44
cystatinC	65
SCrCC	33
DialysisEver	30
CKDHx	9
maneuvers	103
FVC	117
FEV1PP	117
EF	85
EF3cat	85
ecgHR	6
QRS	49
QT	6
QTcFram	6
CV	6
PrivateIns	9
MedicareIns	5
dailyDiscr	40
lifetimeDiscrm	80
perceivedStress	26
SMK3cat	41
BMI3cat	4
vitaminD3	60
FakeCensusTractID	65
nbmedHHincome	6
nbpctBlackNH	6
nbK3FavorFoodstore	6
sportIndex	117
hyIndex	26
activeIndex	14
stroke	139
years	139

Length: 62, dtype: int64

Graphical Summaries – Box and Whisker Plots

- Outliers were detected with Box and Whisker plots
- Utilizing the IQR (Interquartile range) method we can detect the following:
  - Minimum value
  - Quarter 1 – 25th percentile
  - Median – 50th percentile
  - Quarter 3 – 75th percentile
  - Maximum value
- Outliers live beyond ranges





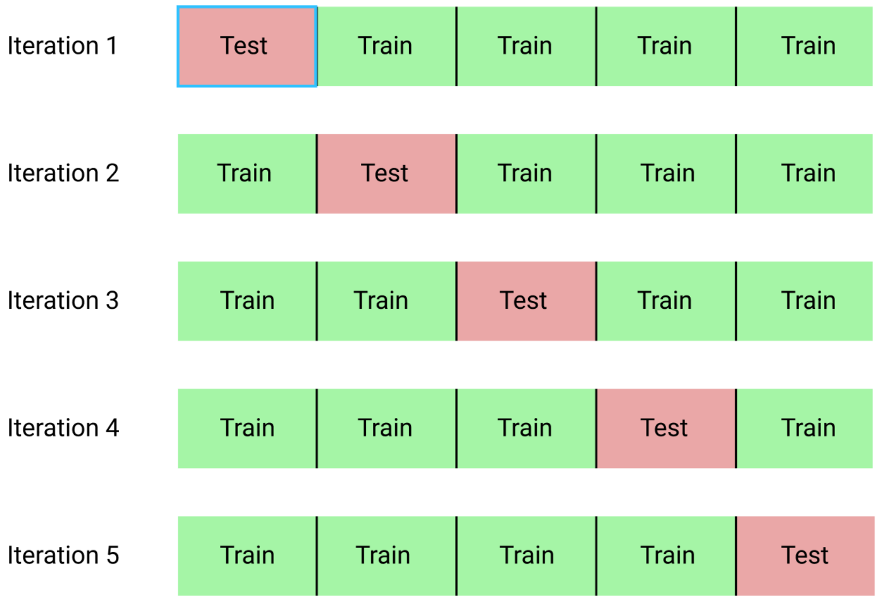
Methods – Logistic Regression

- Used on target variables that are categorical
- Useful in the prediction of probability
- Model predicts P(Y=1) as a function of X
- Logistic function applied to keep outcomes in range of 0 and 1
- Target variable dropped while coefficients were obtained
- Regression coefficients in this model symbolize the transformation in the logit for each unit change in the predictor

$$f\left(x\right)=\frac{1}{1+\mathrm{e}^{-\left(x\right)}}$$

Methods – K Folds Cross-Validation

- Evaluates machine learning models based on parameter k
- Once a specific value for k is selected, data is split into as many samples
- Data is trained on each iteration of the k-fold process then each score is appended and a mean is obtained to determine model accuracy



Methods – Logistic Regression with SGD

- Stochastic gradient descent is an iterative algorithm that identifies the minimum of a function
- Technique revises the parameters of models
- Fast model
- Computes the derivative from training data occurrence and calculates the update

```
from sklearn.linear_model import SGDClassifier
from sklearn.pipeline import Pipeline

pip = Pipeline([('model', SGDClassifier(loss='log', max_iter=500, tol=1e-3, ra
ndom_state=123, warm_start=False))])

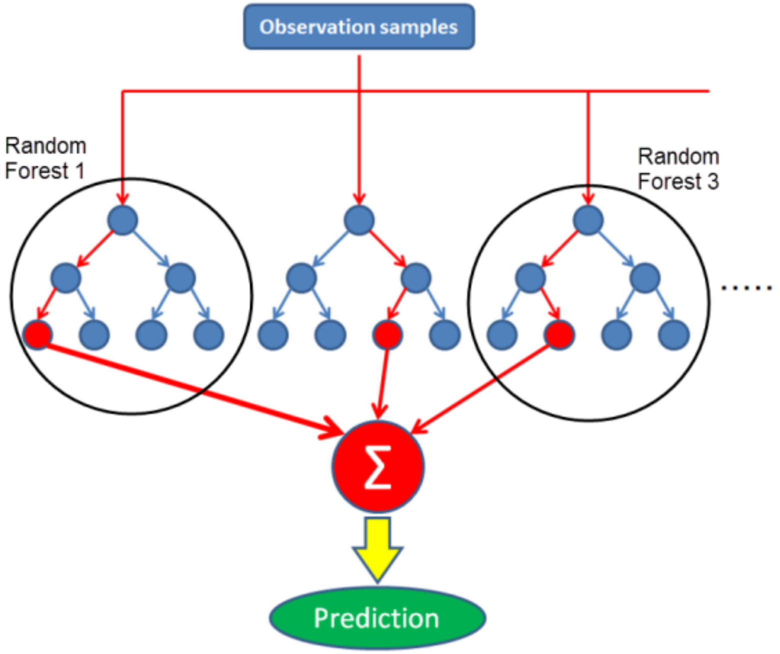
#Hyper parameters

param  = {
    'model__alpha': [1, 2, 5],
    'model__penalty': ['l1', 'l2']
}
```

```
#Set the model
from sklearn.model_selection import GridSearchCV
sgdlr = GridSearchCV(estimator=pip, param_grid=param, scoring='roc_auc', n_job
s=-1, pre_dispatch='2*n_jobs', cv=5, verbose=1, return_train_score=False)
```

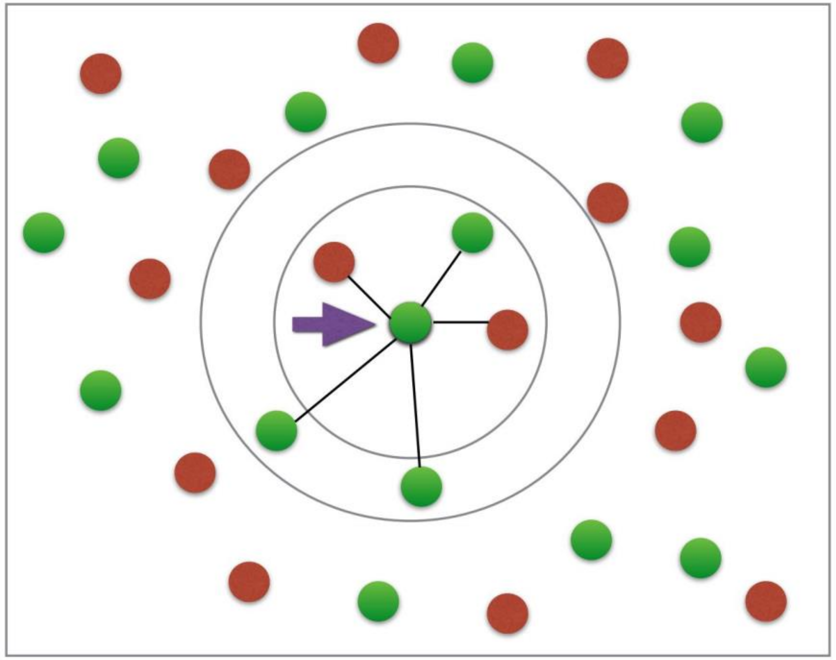
Methods – Random Forest

- Method built on premise of decision trees
- Algorithm operates as a collaborative
- k is randomly selected from the total
- CART (Classification and Regression Trees) computed
- Processes for balancing errors in data sets were classes are imbalanced



Methods – K-NN (Nearest Neighbor)

- Does not require data to fit a normal distribution
- Algorithm determines the distance between a query and features in the data
- Sums the distance and the index of the sample to a systematic assembly
- Sorts collection and picks first k-entries
- Returns mean of the k-labels for regression, mode for classification



Model Comparison Using ROC

- Performance classified by the AUC (Area Under Curve) – ROC (Receiver Operating Characteristic) curve graphically
- AUC exemplifies degree or measure or separability while ROC serves as a probability curve
- ROC curve plotted with TPR (True Positive Rate) (y-axis) against FPR (False Negative Rate) (x-axis)

True Positive Rate

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{P}} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} = 1 - \mathrm{FNR}$$

False Positive Rate

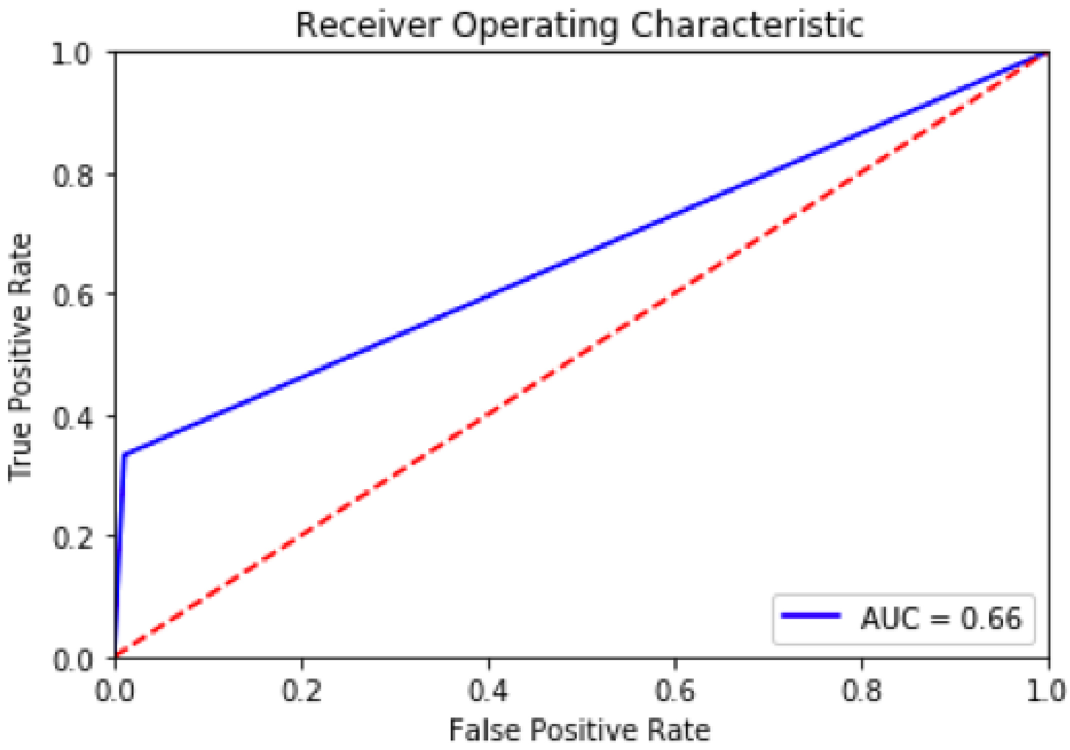
$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{P}} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} = 1 - \mathrm{FNR}$$

Model Comparison Using ROC

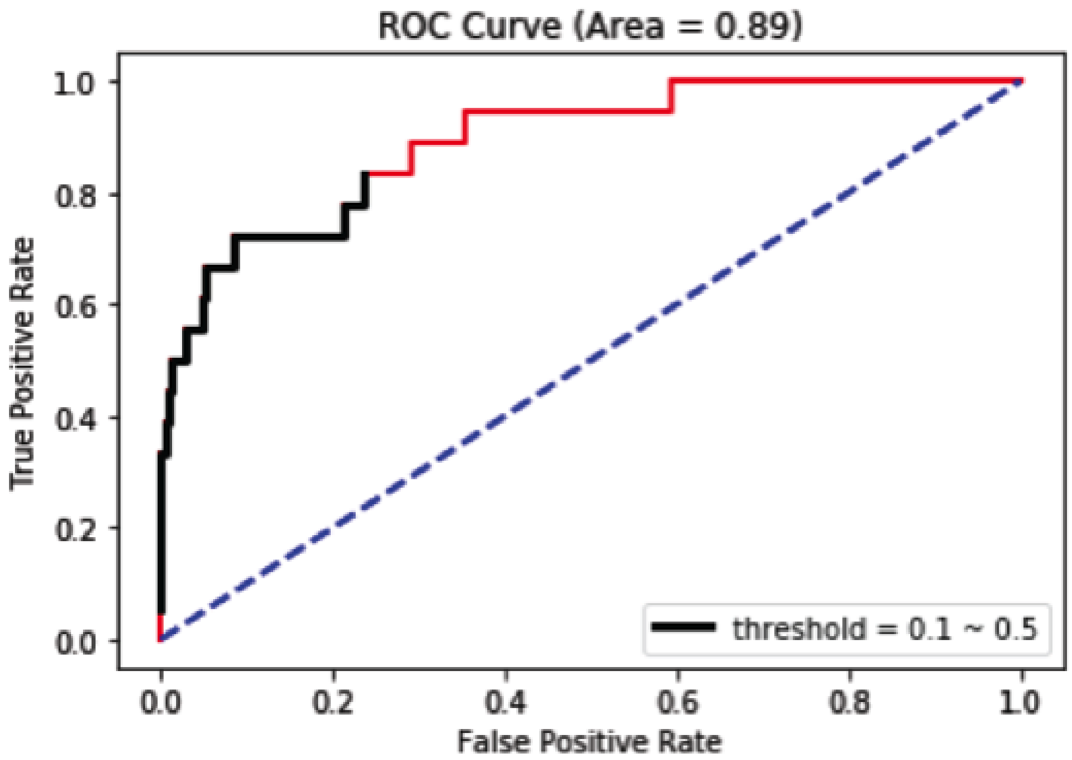
Model	ROC AUC Training	ROC AUC Test	True POS Test	False Neg Test	True Neg Test	False Pos Test	Accuracy Training	Accuracy Test
Logistic Regression w/SGD	0.920286034963608	0.8931028551771586	646	0	18	0	0.9693313222724987	0.9728915662650602
Random Forest	1.0	0.8126504987960095	646	0	17	1	1.0	0.9743975903614458
K-NN	0.9573540915583973	0.8279583763329894	646	0	18	0	0.9693313222724987	0.9728915662650602
Logistic Regression	0.7118180736004354	0.6612487100103198	639	6	12	6	0.9798893916540975	0.9713855421686747
XGBoost	0.9553571428571428	0.6298486161374178	852	6	17	1	0.992	0.9795

Roc Curves

Logistic Regression

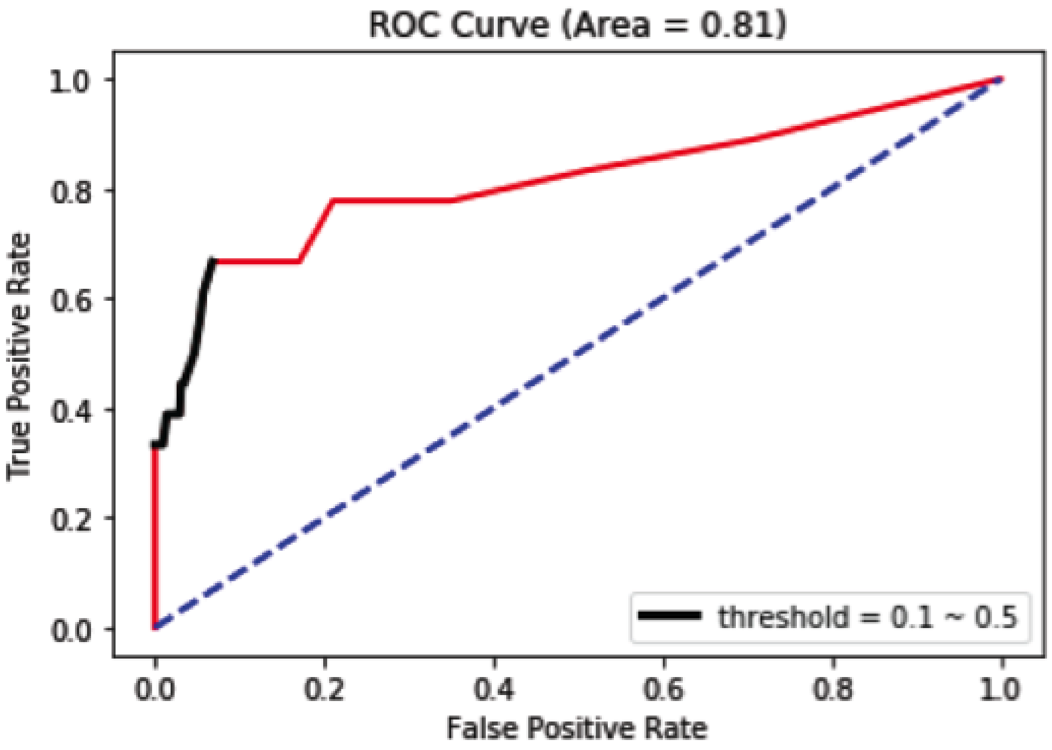


Logistic Regression with SGD

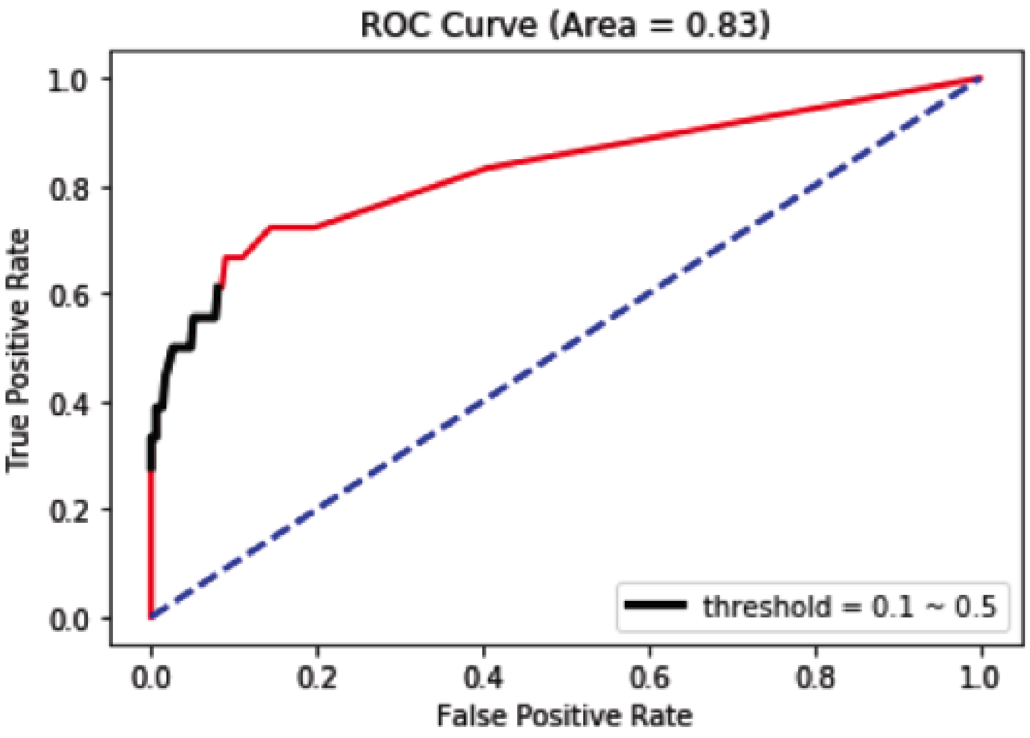


Roc Curves

Random Forest

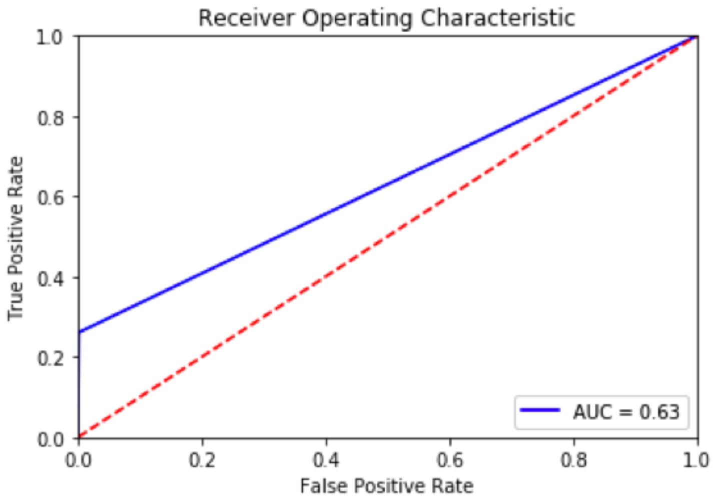


K-NN



XGBoost

- Provides a parallel tree boosting (aka GBDT, GBM)
- Fast and accurate
- Optimized distributed gradient boosting library
- Highly efficient, flexible and portable



Conclusion

- Important steps conducted during data preprocessing
- Random Forest performed best on the training set for ROC AUC score and accuracy scores for the test and training data
- Logistic Regression w/SGD performed best on the ROC AUC for test data
- When outcome is concluded true and its in fact false, there is a false positive or type I error
- When outcome is false and its true, there is a false negative or type II error
- Best to select a model with least false negative rate

Summary of the Project (Part I and Part II)

- **Using 1mg/dL as cut off point. we have found that hs-CRP has impact on developing of stroke incidence using Cox and Weibull survival analysis ($p<0.05$).**

- **After compared the ROC AUC, accuracy, false negative and false positive rates, we have found that the Logistic Regression with SGD model are the best to predict stroke incidence among the 5 machine learning models.**