Project Proposal-BUAN 6356

GROUP-7

Chaitanya
Samanth
Sharath
Sheerin

HEART STROKE PREDICTION
TARGET FIRM: CVS Healthcare

TABLE OF CONTENTS

Heading	Slide Number
Overview	3
Business Understanding	4
Data Description	9
Exploratory Data Analysis	12
Data Modelling	18
Recommendations	21
Additional Recommendations	24
References	25
Appendix	27

Overview of the project

- The main objective of our project is to predict the chances of stroke. For this process, we have used 12 attributes such as age, gender, hypertension, heart disease etc.
- We have made around 5110 observations for this prediction. Based on the observation, we are predicting the factors that influences the chances of stroke using the logistic regression model and decision trees.
- The outcome of this project will be a reference for many Healthcare companies to experiment with their insurance plans according to the attributes we used for the observation.

Business understanding (Firm description)

- Based on our observation and results after conducting a logistic regression,we are looking to submit our project proposal to the famous company CVS Health.
- CVS is one of the world's largest healthcare companies,housing about 300000(2020) employees and ranked 5th on the 2020 Fortune 500 list of the largest U.S. corporations by total revenue.
- It has various product and service line including pharmacy ,retail, health insurance(Aetna),etc
- Our observation and the results with our project will surely help them to keep our project as a reference in predicting the chances of stroke and spread the awareness to the society based on the attributes what we used in this project.

Business understanding (Firm description)

Through series of strategic business moves, CVS Health has categorised its business into four key areas as shown in the table below.

Area of Business	**Approx Yearly Revenue(USD)**	**Number of employees(approx)**
CVS Pharmacy	134 Billion	203000
CVS Caremark	**36 Billion**	19500
Minute Clinic	25 Billion	18000
Aetna(Health Insurance)	60 Billion	48000

Business understanding (Firm description)

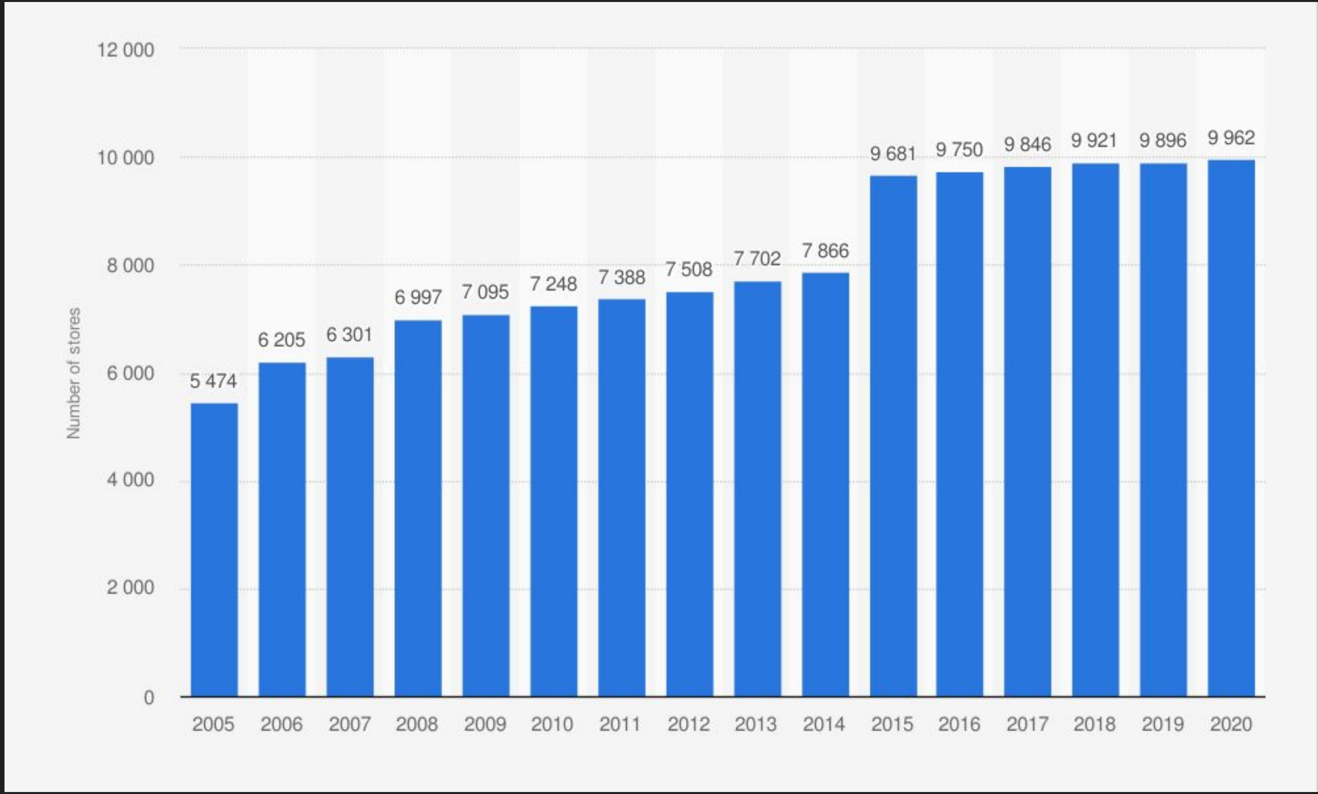


Fig. Geographic footprint of CVS Health

- CVS healthcare is having it's branches in 49 states.
- As per the year 2020, they're having 9962 stores in the country as shown in the graph.

Business understanding(SWOT ANALYSIS)

STRENGTHS

- Brand Value and Market Positioning.
- The financial statements of Aetna seem organized and reliable with consistently more than 60 Billion USD in the last 3 years.
- Formed in the relationship era.Aetna has a strong base of royal customers who may not go away anytime soon.As of 2019 More than 39 Million people across the US.

WEAKNESS

- Uncertainty in Market share.Aetna only holds 5.56% of the national market share.Government laws create uncertainty about the company's future business model.
- Like most Private Health Insurance companies,Aetna’s health insurance plans might not be pocket friendly to certain sectors of population.

Business understanding(SWOT ANALYSIS)

OPPORTUNITIES

- Improvement in technology gives better opportunities for greater customization of health plans.
- Expansion into new/emerging economies.
- With increase in aged population,particularly in developed countries such as US,people need more care,costlier care as they age.

THREATS

- Strict Regulatory issues,causing uncertainty in business.
- Tough competition by other major firms in the health insurance line such as United,BCBS,Cigna,etc.

Data Description:

Heart Stroke is 2nd most leading cause of death globally. Approximate 11% of total deaths . Through this dataset we are predicting whether the target population is susceptible of having stroke based on the parameters like gender, age, various diseases, and smoking status. Each row in the data provides relevant information about the patient.

- Data Source: https://www.kaggle.com/fedesoriano/stroke-prediction-dataset?select=healthcare-dataset-stroke-data.csv
- Originally, the data contains 5110 observations with 12 attributes.There is no accurate time frame mentioned over which data has been collected. The data was published on kaggle on 26th Jan’21.
- For our analysis we have split the data and now we have 4343 observations.
- Our target variable of interest is to predict stroke based on the 12 attributes.

Data Description: Attribute Information

1) ID: unique identifier

2) Gender: "Male", "Female" or "Other"

3) Age: Age of the patient

4) Hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension

5) Heart_disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease

6) Marital Status:Ever_married: "No" or "Yes"

7) Work_type: "children", "Govt_jov", "Never_worked", "Private" or "Self-employed"

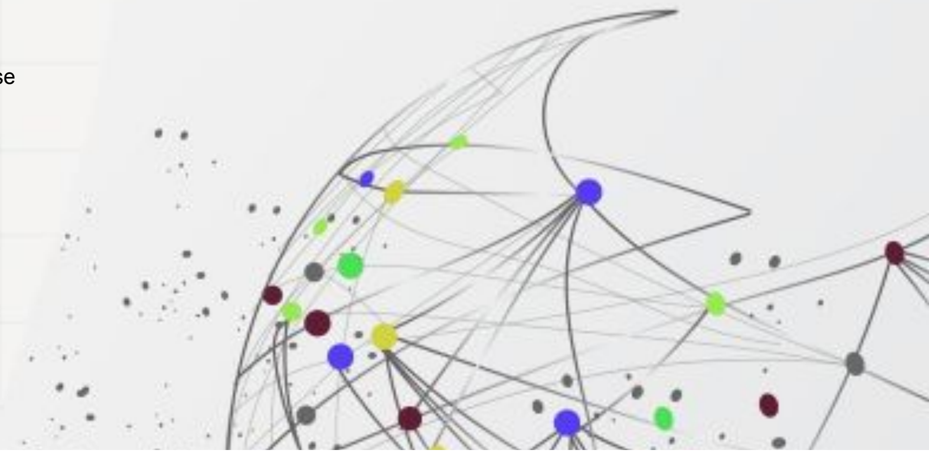
8) Residence_type: "Rural" or "Urban"

9) Avg_glucose_level: average glucose level in blood

10) BMIi: body mass index

11) Smoking_status: "formerly smoked", "never smoked", "smokes" or "Unknown"*

12) Stroke: 1 if the patient had a stroke or 0 if not



```
> summary(stroke.df)
       id           gender               age         hypertension     heart_disease     ever_married        work_type         Residence_type     avg_glucose_level      bmi
 Min.   :   67   Length:4343        Min.   : 0.08   Min.   :0.00000   Min.   :0.00000   Length:4343        Length:4343        Length:4343        Min.   : 55.12    Length:4343
 1st Qu.:17938   Class :character   1st Qu.:25.00   1st Qu.:0.00000   1st Qu.:0.00000   Class :character   Class :character   Class :character   1st Qu.: 77.25    Class :character
 Median :36969   Mode  :character   Median :45.00   Median :0.00000   Median :0.00000   Mode  :character   Mode  :character   Mode  :character   Median : 91.85    Mode  :character
 Mean   :36646                      Mean   :43.13   Mean   :0.09256   Mean   :0.05227                                                            Mean   :106.22
 3rd Qu.:54736                      3rd Qu.:61.00   3rd Qu.:0.00000   3rd Qu.:0.00000                                                            3rd Qu.:114.28
 Max.   :72918                      Max.   :82.00   Max.   :1.00000   Max.   :1.00000                                                            Max.   :271.74
 smoking_status         stroke
 Length:4343        Min.   :0.00000
 Class :character   1st Qu.:0.00000
 Mode  :character   Median :0.00000
                    Mean   :0.04812
                    3rd Qu.:0.00000
                    Max.   :1.00000
>
```

Statistics Snapshot : Summary - Removed Gender: “other” & 178 NA’s of BMI is replaced by mean BMI

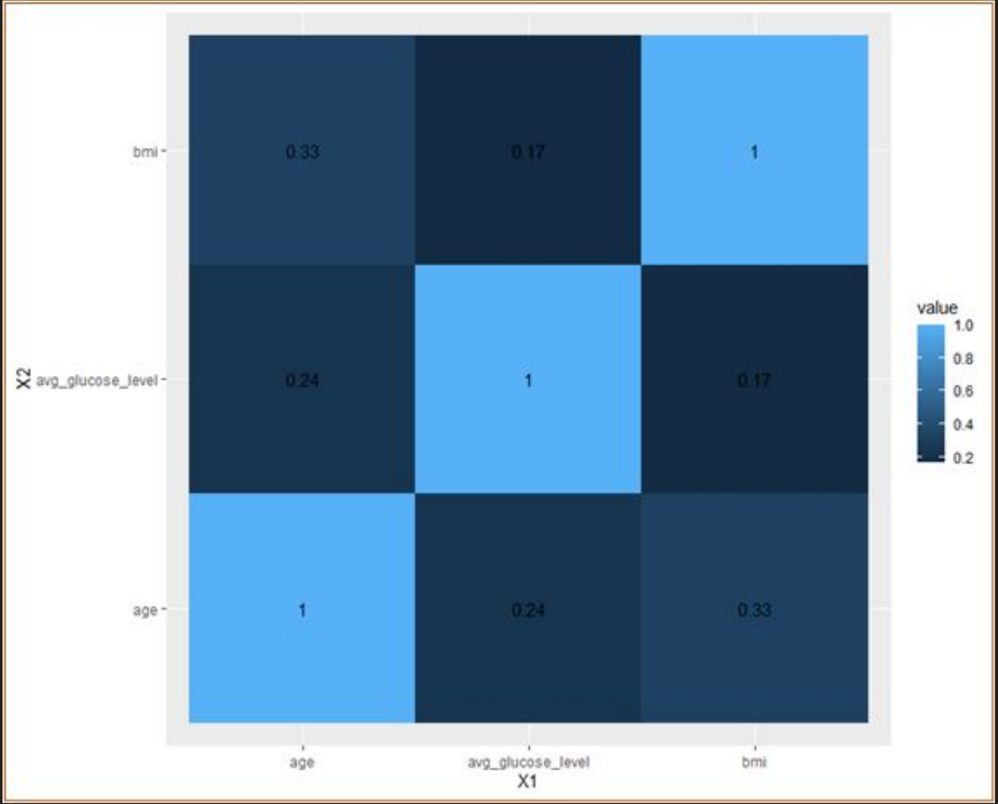
```
> summary(stroke.df)
    gender          age        hypertension heart_disease ever_married          work_type    Residence_type avg_glucose_level       bmi                 smoking_status stroke
 Female:2539   Min.   : 0.08   0:3940       0:4115        No :1497     children     : 584   Rural:2103     Min.   : 55.12     Min.   :11.50   formerly smoked: 747   0:4133
 Male  :1803   1st Qu.:25.00   1: 402       1: 227        Yes:2845     Govt_job     : 561   Urban:2239     1st Qu.: 77.24     1st Qu.:23.50   never smoked   :1619   1: 209
               Median :45.00                                           Never_worked :  21                  Median : 91.85     Median :28.00   smokes         : 681
               Mean   :43.13                                           Private      :2487                  Mean   :106.21     Mean   :28.87   Unknown        :1295
               3rd Qu.:61.00                                           Self-employed: 689                  3rd Qu.:114.24     3rd Qu.:33.10
               Max.   :82.00                                                                               Max.   :271.74     Max.   :97.60
                                                                                                                              NA's   :178
```

Descriptive Summary of Numerical Variables:

	mean	median	sd	variance	min	max	count	miss.val
age	43.12958	45.00	22.558035	508.86496	0.08	82.00	4342	0
avg_glucose_level	106.21045	91.85	45.327663	2054.59703	55.12	271.74	4342	0
bmi	28.87105	28.40	7.651788	58.54987	11.50	97.60	4342	0

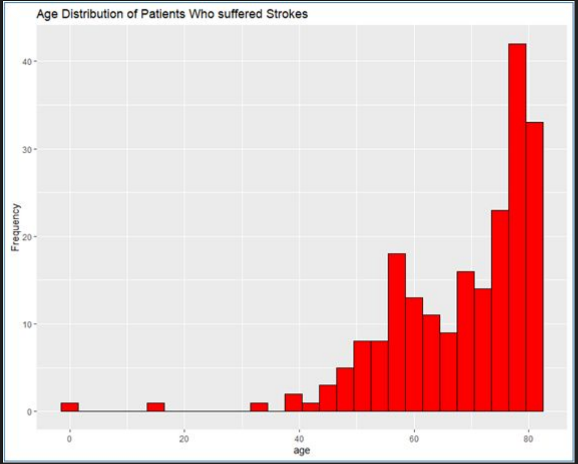
```
# A tibble: 2 x 3
  gender avg_bmi     n
* <fct>    <dbl> <int>
1 Female    29.1  2539
2 Male      28.6  1803
>
```

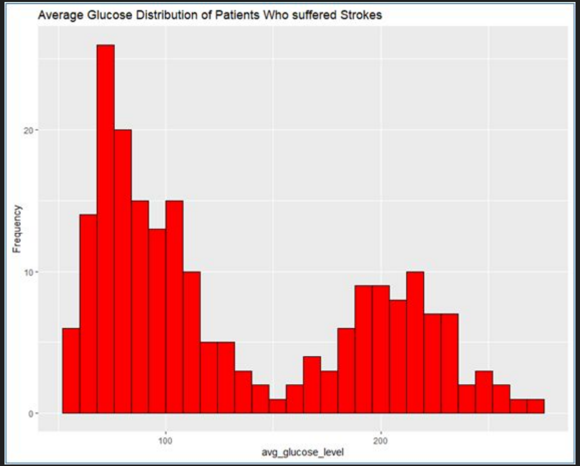
Exploratory Data Analysis: Correlation Matrix

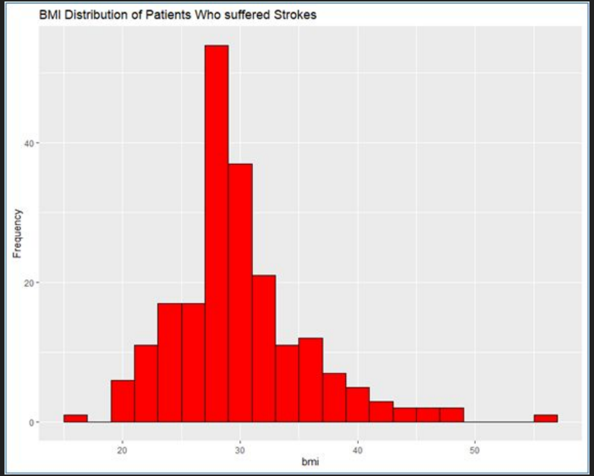


It seemed like both BMI and Age are weakly positively correlated

Exploratory Data Analysis





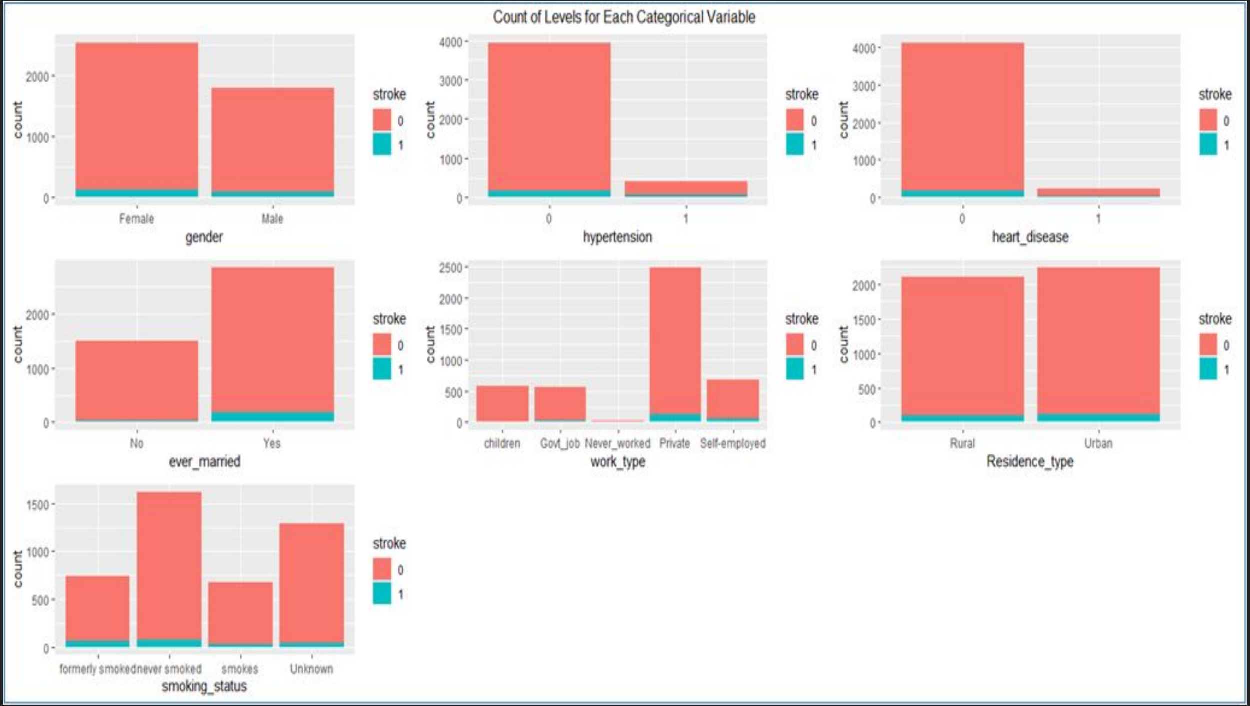


1) As the patient's age advances, so is the chance to have a stroke.

2) Population who have average glucose level more than 150 suffered a stroke.Prediabetes is considered in patient for a reading between 140-199mg/dL. Diabetes is one of the risk factors for stroke occurrence and prediabetes patients have an increased risk of stroke.

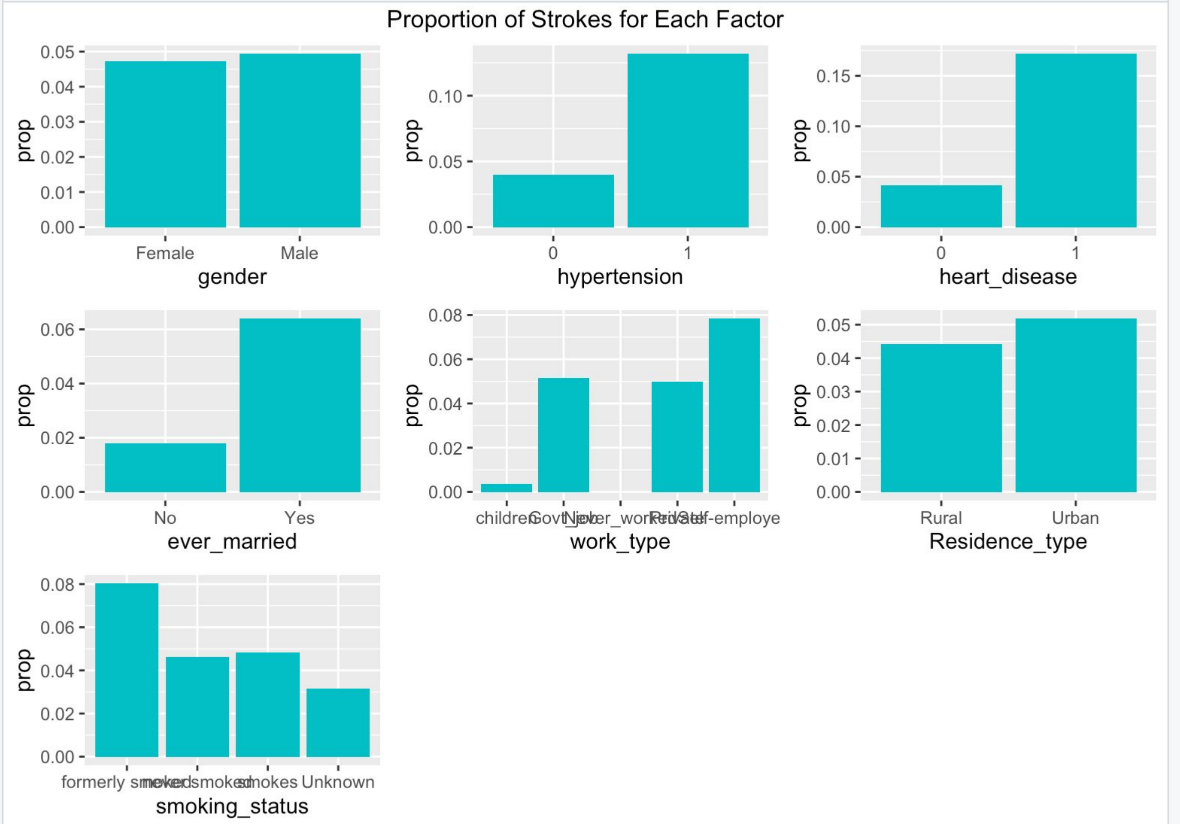
3) Population with BMI between 25 to 35 are the highest to suffer a stroke . However, higher BMI does not increase risk of stroke.

Exploratory Data Analysis



The amount of those who have had a stroke is a small portion of the population. So, we compare proportions of those who had a stroke in each factor.

Exploratory Data Analysis



1) Gender and residence type do not show much difference in occurrence of strokes.

2) Hypertension, heart disease, marital status have a much higher proportion of populations to have a stroke.

3) Children have very low occurrences of strokes, never worked have none. There is little difference in the proportions of those who work in government and those who work in the private sector. Self employed have a higher proportion than both of these sectors.

4) Former smokers have a higher chance of occurrence to stroke than current smokers .

Exploratory Data Analysis

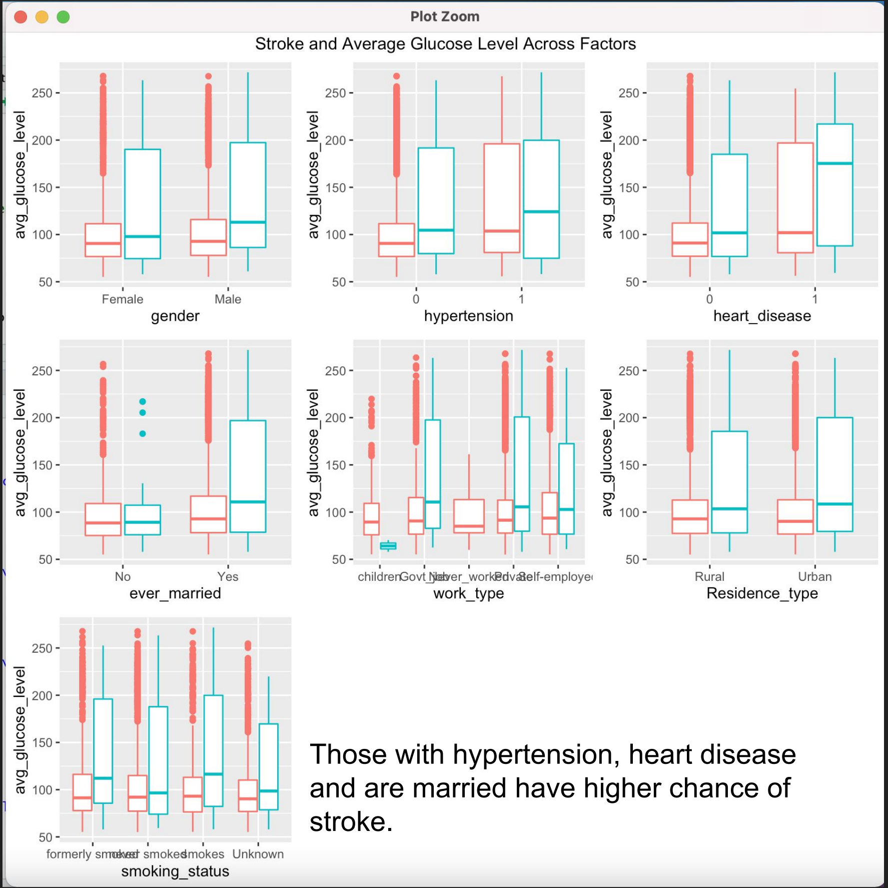
For all levels in each factor those who have had a stroke are older.

Those with hypertension and heart disease are older than those who do not.

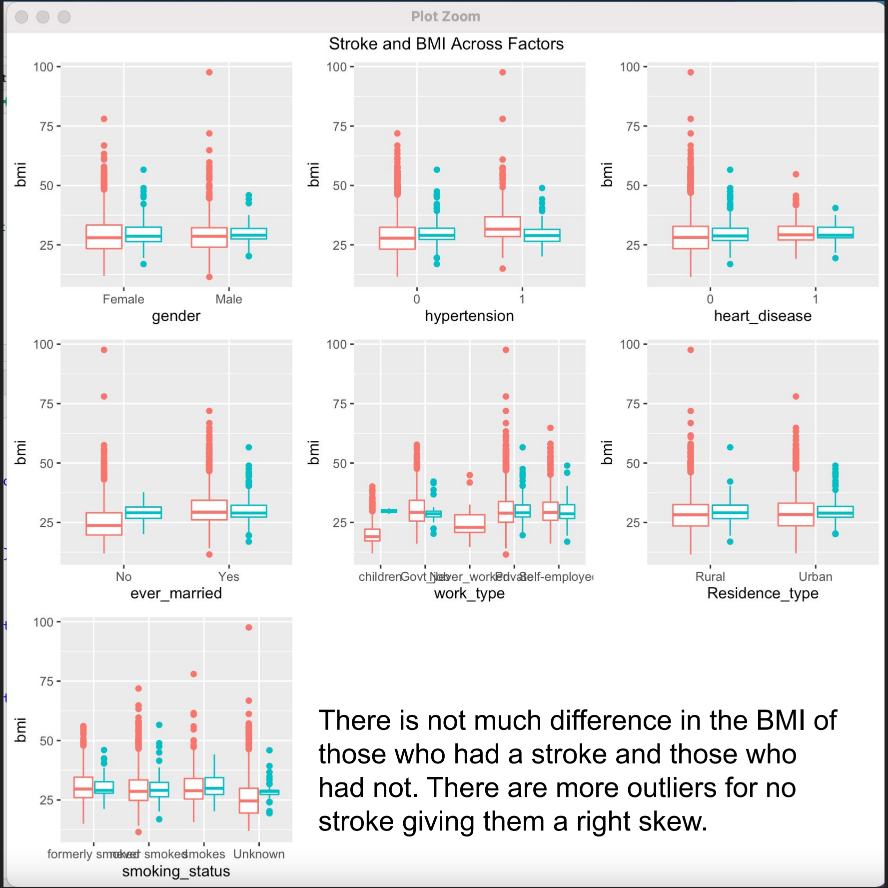
Self-employed are also older than the other types of work.They are at risk to have a stroke.

Those who had a stroke and smoke are younger than those who quit or never smoked (but still had a stroke).

Exploratory Data Analysis



Those with hypertension, heart disease and are married have higher chance of stroke.



There is not much difference in the BMI of those who had a stroke and those who had not. There are more outliers for no stroke giving them a right skew.

Data Modelling

Model Comparison

We will compare our accuracy with other classification models such as decision trees and random forests and select the best one.

Model Diagnosis

After running the model, checking the confusion matrix and predictors coefficients will give us a better idea about model performance.

Model Selection

Given target variable is a binary categorical variable, we will use classification prediction model - logistic regression

Variable Selection

We have 10 predictor variables in our dataset to predict our target variable (stroke)

Data Modelling - Model Comparison

Model	Prediction	Accuracy	Measures
Logistic Regression	FALSE TRUE FALSE 1240 60 TRUE 3 0	Accuracy : 0.9517 95% CI : (0.9386, 0.9626) No Information Rate : 0.954 P-Value [Acc > NIR] : 0.6838	Sensitivity : 0.9976 Specificity : 0.0000 Pos Pred Value : 0.9538 Neg Pred Value : 0.0000
Random Forest	Reference Prediction 0 1 0 1240 59 1 3 1	Accuracy : 0.9524 95% CI : (0.9394, 0.9633) No Information Rate : 0.954 P-Value [Acc > NIR] : 0.6362	Sensitivity : 0.99759 Specificity : 0.01667 Pos Pred Value : 0.95458 Neg Pred Value : 0.25000
Decision Tree	Reference Prediction 0 1 0 1195 46 1 48 14	Accuracy : 0.9279 95% CI : (0.9124, 0.9413) No Information Rate : 0.954 P-Value [Acc > NIR] : 1.0000	Sensitivity : 0.9614 Specificity : 0.2333 Pos Pred Value : 0.9629 Neg Pred Value : 0.2258

Data Modelling - Under Sampled data

Model	Prediction	Accuracy	Measures
Logistic Regression	FALSE TRUE FALSE 940 12 TRUE 303 48	Accuracy : 0.7583 95% CI : (0.7341, 0.7813) No Information Rate : 0.954 P-Value [Acc > NIR] : 1	Sensitivity : 0.7562 Specificity : 0.8000 Pos Pred Value : 0.9874 Neg Pred Value : 0.1368
Random Forest	Reference Prediction 0 1 0 926 14 1 317 46	Accuracy : 0.746 95% CI : (0.7214, 0.7694) No Information Rate : 0.954 P-Value [Acc > NIR] : 1	Sensitivity : 0.7450 Specificity : 0.7667 Pos Pred Value : 0.9851 Neg Pred Value : 0.1267
Decision Tree	Reference Prediction 0 1 0 876 20 1 367 40	Accuracy : 0.703 95% CI : (0.6774, 0.7277) No Information Rate : 0.954 P-Value [Acc > NIR] : 1	Sensitivity : 0.70475 Specificity : 0.66667 Pos Pred Value : 0.97768 Neg Pred Value : 0.09828

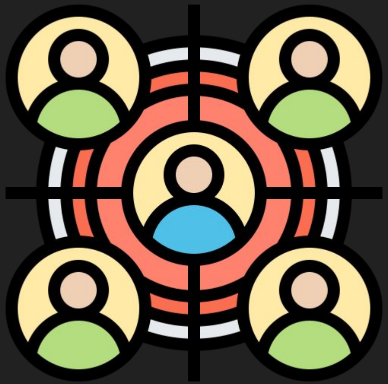
Data Modelling - Over Sampled data

Model	Prediction	Accuracy	Measures
Logistic Regression	FALSE TRUE FALSE 921 12 TRUE 322 48	Accuracy : 0.7437 95% CI : (0.719, 0.7672) No Information Rate : 0.954 P-Value [Acc > NIR] : 1	Sensitivity : 0.7409 Specificity : 0.8000 Pos Pred Value : 0.9871 Neg Pred Value : 0.1297
Random Forest	Reference Prediction 0 1 0 1209 53 1 34 7	Accuracy : 0.9332 95% CI : (0.9183, 0.9462) No Information Rate : 0.954 P-Value [Acc > NIR] : 0.99970	Sensitivity : 0.9726 Specificity : 0.1167 Pos Pred Value : 0.9580 Neg Pred Value : 0.1707
Decision Tree	Reference Prediction 0 1 0 1117 39 1 126 21	Accuracy : 0.8734 95% CI : (0.8541, 0.891) No Information Rate : 0.954 P-Value [Acc > NIR] : 1	Sensitivity : 0.8986 Specificity : 0.3500 Pos Pred Value : 0.9663 Neg Pred Value : 0.1429

Recommendations

We discuss 3 ways in which our Target Firm can use our Data Model to improve the efficiency of their business operations:

1)Aid Segmentation and Targeted Development of Insurance Plans



- Our data model can help our target firm CVS(Aetna) to more precisely predict the susceptibility of a patient being affected by stroke and also help their insurance policy makers in pricing analysis and attractive premiums as it provides in depth risk analysis of the patient.
- It can also aid in providing customized healthcare plans to individuals with respect to their health issues.

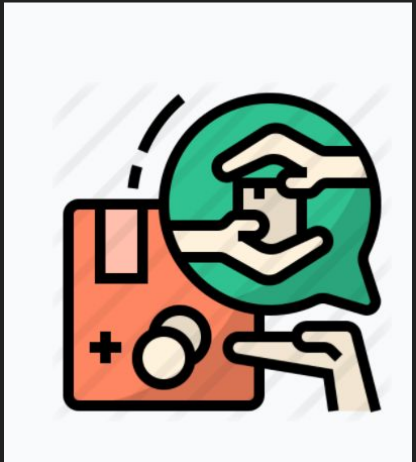
Recommendations

2)Improve efficiency of Sales and Marketing operations

- Our data set can help CVS(Aetna) predict the sale of their insurance policies by predicting stroke owing to different demographic factors.
- Collecting key data points, and applying business intelligence may aid in the identification of emerging opportunities and the analysis of the effectiveness of various marketing channels in order to prioritize efforts and gain a competitive advantage.

Recommendations

3) Maximize cross-selling



- With our data model and analysis,CVS(Aetna) can find opportunities to cross-sell new products/Plans to their members.

- During open enrollment, they can use data to send targeted campaigns to promote new products/plans or send invitations to special events that will target members who are at risk for stroke based on comorbidities such as hypertension,for example.

Additional Recommendations

- CVS's pharmacy can use our data to gain a better know how of the susceptibility to stroke and can accordingly distribute their medicines and maintain supply demand equilibrium.

- Our data model can help CVS enable precision medicine to treat stroke by taking into factors like age, different genetic variables, ,lifestyle,existing health conditions,etc to accordingly identify and segment the high risk population of getting stroke and develop personalized or segment wise precision medication for each segment based on genetics and Lifestyle.

- It can help CVS open up their minuteclinics in areas with stroke susceptible population and give treatment as required.

References

1)https://www.kaggle.com/fedesoriano/stroke-prediction-dataset

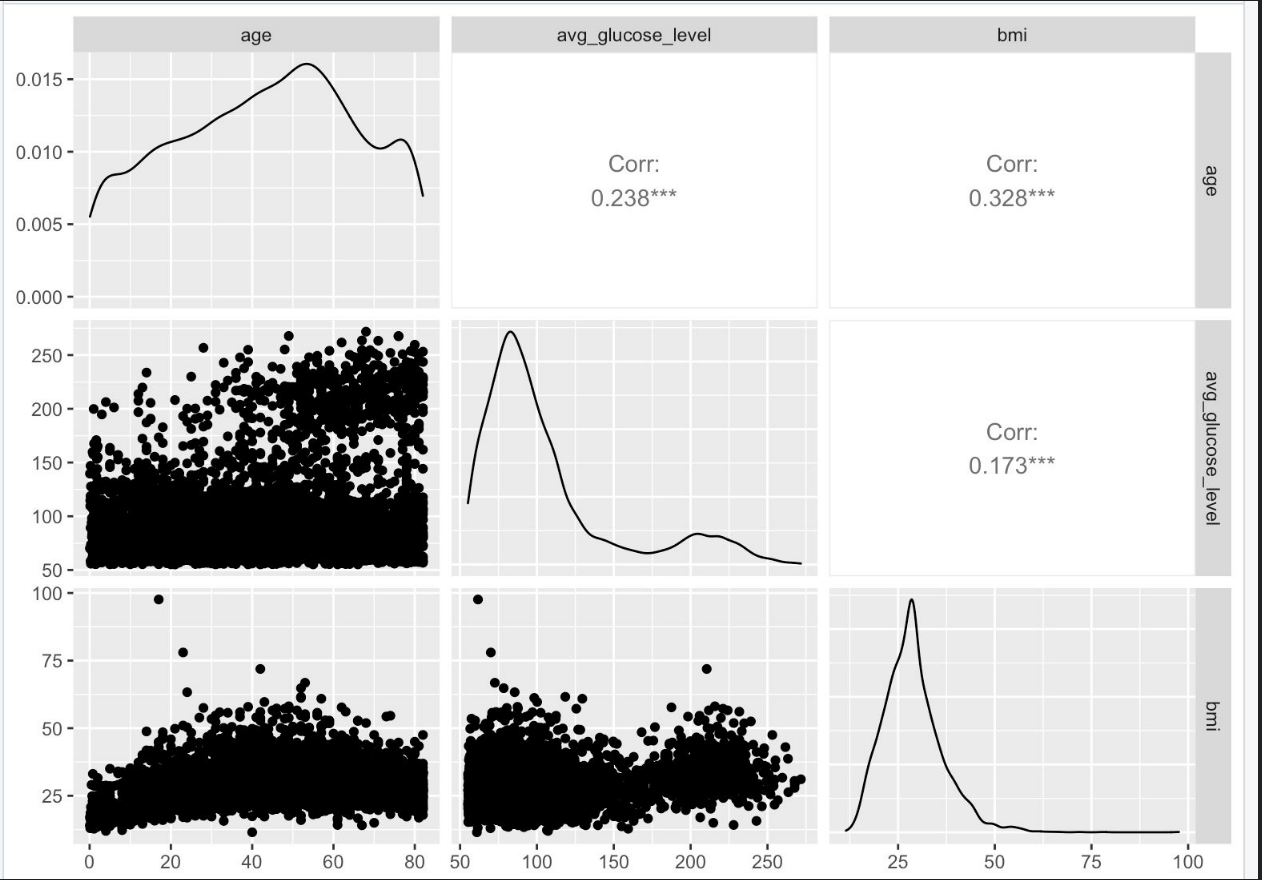
2)https://www.swotandpestle.com/unitedhealth-group/

3) https://www.analyticsvidhya.com/blog/2016/03/practical-guide-deal-imbalanced-classification-problems/

4) https://www.hackerearth.com/blog/developers/practical-guide-logistic-regression-analysis-r/

Appendix

Exploratory Data Analysis: Scatter Plot



It seemed like both BMI and Age are weakly positively correlated

BMI and avg_glucose level are weakly correlated

Data Modelling - Logistic Regression

```
> ## Running Logistic Regression Model
> logit.reg <- glm(stroke ~ ., data = train.df, family = "binomial")
> options(scipen=999)
> summary(logit.reg)

Call:
glm(formula = stroke ~ ., family = "binomial", data = train.df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0935  -0.3127  -0.1542  -0.0796   3.6077

Coefficients:
                             Estimate  Std. Error z value              Pr(>|z|)
(Intercept)                -6.8934553   1.0958739  -6.290       0.000000000317 ***
genderMale                 -0.1237049   0.1872739  -0.661                0.509
age                         0.0813864   0.0078676  10.344 < 0.0000000000000002 ***
hypertension1               0.3186516   0.2203374   1.446                0.148
heart_disease1              0.2935370   0.2463830   1.191                0.234
ever_marriedYes            -0.2491351   0.2894223  -0.861                0.389
work_typeGovt_job          -1.1761884   1.1699998  -1.005                0.315
work_typeNever_worked     -10.1714285 402.2048051  -0.025                0.980
work_typePrivate           -0.8503511   1.1445973  -0.743                0.458
work_typeSelf-employed     -1.3159990   1.1711731  -1.124                0.261
Residence_typeUrban         0.1395732   0.1817295   0.768                0.442
avg_glucose_level           0.0039281   0.0015634   2.513                0.012 *
bmi                        -0.0007984   0.0148181  -0.054                0.957
smoking_statusnever smoked -0.2829164   0.2294799  -1.233                0.218
smoking_statussmokes        0.0238334   0.2832035   0.084                0.933
smoking_statusUnknown      -0.1134773   0.2702838  -0.420                0.675
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1189.14  on 3038  degrees of freedom
Residual deviance:  924.31  on 3023  degrees of freedom
AIC: 956.31

Number of Fisher Scoring iterations: 14
```

```
> # first 10 actual and predicted records #
> data.frame(actual = valid.df$stroke[1:10], predicted = logit.reg.pred[1:10])
   actual   predicted
1       0 0.001525292
4       0 0.001885494
6       0 0.002347611
11      0 0.045132725
19      0 0.001345796
31      0 0.005806841
34      0 0.193153500
35      1 0.081772675
39      0 0.134421720
42      0 0.053967738
```

```
Confusion Matrix and Statistics

       FALSE TRUE
 FALSE  1240   60
 TRUE      3    0

               Accuracy : 0.9517
                 95% CI : (0.9386, 0.9626)
    No Information Rate : 0.954
    P-Value [Acc > NIR] : 0.6838

                  Kappa : -0.0044

 Mcnemar's Test P-Value : 0.000000000001722

            Sensitivity : 0.9976
            Specificity : 0.0000
         Pos Pred Value : 0.9538
         Neg Pred Value : 0.0000
             Prevalence : 0.9540
         Detection Rate : 0.9517
   Detection Prevalence : 0.9977
      Balanced Accuracy : 0.4988

       'Positive' Class : FALSE
```

Porter Five Forces in the Healthcare industry

- **Rivalry amongst the competitors(High):-** Intense Rivalry is faced by the Healthcare Industry about market capitalisation, margin over other companies.
- **The Bargaining power of Buyers(Low):-** In USA, buyers hold a weak bargaining power. To avail the health insurance, a price has to be paid for it.For example customers rather buy and pay for health insurance rather for a surgery.
- **The Bargaining power of suppliers(High):-** Suppliers hold a high bargaining power as this sector is totally dependent on it's services provided.
- **Substitute of Existing Products(High):-** Customers always prefer to opt to go where they are able to find the cheaper prices.
- **Threat towards Entry of New Entrants(Low):-**Due to very few dominant companies in the healthcare industry,there is relatively lesser threat towards new entrants in this field.

THANK YOU

QUESTIONS??