Smith SCHOOL OF BUSINESS Queen's University | Master of Management Analytics Toronto

MMA 867: Predictive Modelling

Professor: Jue Wang

Team Project Final Report

Due Date: Oct 01st, 2022

Team Dunning

Student Name	Student Number
Mitch Leeson	20370956
Nancy Li	20359704
Manchali Gopal	20374572
Hannah Chu	20314738
Daniel Li	10061022
Terry Wang	20333333
Joey Chan	20373866
Cristobal Muniz (Cris)	20290165

Table of Contents

Background and Objectives

Stroke is the second leading cause of death globally (What we do, 2022). In Canada, thankfully this disease ranks only as the fifth most deadly however remains a large public health concern as it counts for over 50,000 instances each year and of those 15% die (Fact Sheet: Stroke Statistics, n.d.). We are writing this report as *Data Science Team Dunning*, embedded within the Public Health Agency of Canada (PHC); presenting to senior management on the findings of our analysis around predicting the key factors which are correlated with patients suffering a stroke. Prevention is the best medicine when it comes to a disease such as a stroke, therefore an understanding of what key factors related to stroke is important for the PHC to understand, to offer recommendations and guidelines to Canadians that will reduce this risk. The goal of this report is to highlight the analysis we have completed and offer recommendations that the PHC could implement to reduce the rates of stroke suffered by Canadians.

The economic toll of this disease is around $3.6B per year in Canada in physician services, hospital and opportunity costs (Fact Sheet: Stroke Statistics, n.d.). With previous research demonstrating that up to 80% of these strokes are preventable, the yearly savings to the entire market could be up to $2.9B each year in Canada alone (Preventing Stroke Deaths, 2017). This economic impact however is not our primary concern, as the PHC mandate is to promote and protect the health and wellness of all Canadians. With this goal in mind, our team has also identified some key recommendations which we believe will have a noticeable impact on the rates of yearly strokes. Importantly, the output from our analysis and subsequent recommendations can and will be easily understood by the public so as to most easily influence and *nudge* Canadians, through their everyday behaviours, towards making the smart daily choices that will reduce their risks of suffering stroke.

Our recommendations revolve around creating awareness campaigns of the key factors that were identified and helping educate Canadians on how their various health markers (age, smoking status, glucose levels, etc.) could impact their risk of suffering a stroke. We believe the *gamification* of an application could be an effective medium to help give micro-rewards to Canadians when they positively influence one of their risk factors, thus reducing their risk of stroke. Below we dig into the findings from our analysis of what these risk factors are.

The Analyses

The analytic process for this project follows problem determination, data exploration and cleaning, feature engineering, and model building. The detailed data analysis process steps can refer to in Appendix I.

Our data is collected from a clinical health record, which has 5,110 observations and includes 11 variables as listed in Appendix II, including the person's basic personal information such as age, gender, BMI, marital

status, work type, residential type and smoking status; and also some health measurements such as hypertension status, heart disease history, average Glucose Level. Data type includes both categorical values and numerical values. “Stroke” is a categorical response variable as our goal is to predict stroke based on given information. As we were exploring the dataset, we noticed that the positive response rate is only about 5%, which makes the dataset extremely imbalanced for the responsive variable. If we leave this imbalance unchecked, the model will produce a false but high accuracy model by predicting everyone having a negative response (“No Stroke”) regardless of the input. Similar to predicting anyone winning a lottery, having a stroke is a relevant small probability event in our dataset. We learned that up-sampling and down-sampling can be helpful when processing imbalanced data, however, we observed that 1) If only applying up-sampling, the model will greatly amplify certain features and eventually lead to unrealistic results (eg: being a child becomes a significant factor of having a stroke but actually only two records show a child suffering a stroke in our dataset, so we treated them as outliers and removed them); 2) If only applying down-sampling, the sample size will shrink from roughly 5000 to around 500, with the concern of the sample losing a significant number of observations, we decided to discard this approach as well. Therefore, during the feature engineering process, we performed both down-sampling and up-sampling techniques on the dataset. We down-sampled the negative responses from 95% to 50% and up-sampled the positive responses from 5% to 50% of the dataset by random, keeping the total observation no change. As a result, the ratio of positive responses and negative responses reaches 1:1 (Figure 1). As part of data cleaning, we removed a total of 205/5110 records including 3 outliers and N/A BMI values.

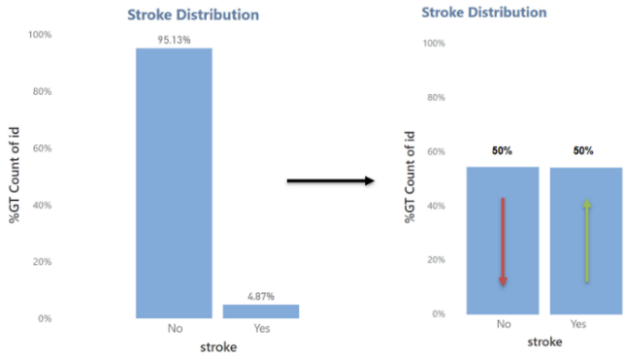


Figure 1. Imbalanced data and data transformation

Moreover, we checked the correlation plot and tried to identify highly correlated variables during the feature engineering phase. As the result, we found that 1) age and marital status are highly correlated since older people are likely to get married; 2) children are highly negatively correlated with age, BMI and marriage status since children are all equal or under 18 years old, single and generally having less weight compared to adults; 3) There might be a strong correlation between the stroke and hypertension, smoking status, and heart disease, we will focus on these variables in our modelling part. (Appendix III on *Correction Matrix*). Our goal is to predict the probability a person suffers a stroke, we decided to begin with the basic logistic

regression model (without random sampling) while trying other methods as we progress such as logistic regression with random sampling and variables selection, Lasso, decision tree and random forest. After comparing different model results, we found that decision tree and random forest models were falling behind in model predictive power indicated by Area-Under-the-Curve (AUC). We found that the logistic model and Lasso model each gave similar results. Although the basic logistic model without sampling produced a 0.856 AUC, we didn't choose it due to the imbalance data problem explained previously. Finally, the logistic regression model with selected variables produced the best predictive result for our sample dataset, which has the highest AUC (0.849) among all the rest models. Please see below Figure 2 for all model AUC results we built. Please also refer to Appendix IV for the confusion matrix and ROC of the final selected model, and parameters comparison between the basic logistic model, Lasso, and logistic model with variables selection.

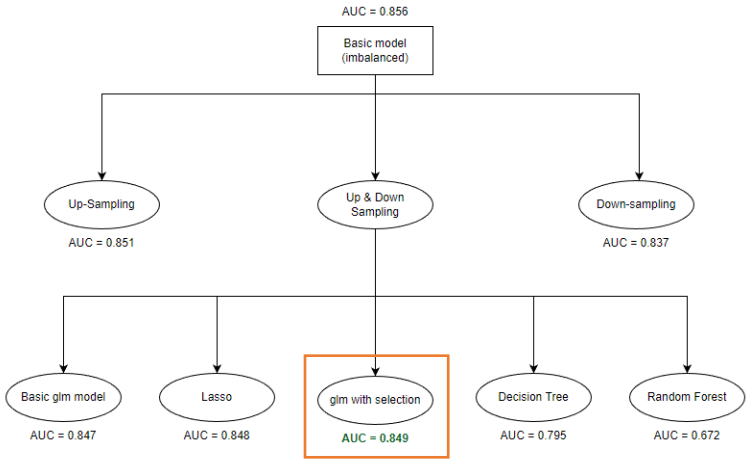


Figure 2. Model results

Conclusion

According to our final model result on this dataset, we found people with increasing age, abnormal glucose levels (both high and low), heart diseases, hypertension and smoking habit are more likely to be associated with suffering from a stroke. We call these risky factors. Please refer to Appendix V *i) for Interpretation of the model result – Positive Coefficients*

i) In our dataset, we found the increase in age is the biggest factor of having a stroke. While children and young adults are nearly risk-free from having a stroke, once they hit 30 years old the probability of having a stroke significantly increases as age increases. Eventually, it hits a peak value at around 60 and 80 years old. (Left graph in Figure 3)

ii) Regarding the glucose level, we identified that when people are having a higher chance of getting a stroke at around 70 mg/dL or above 125 mg/dL glucose level. Interestingly, these values are the thresholds of having a healthy glucose level. When below 70 mg/dL, it causes a disease called

hypoglycemia and when it is above 125 mg/dL it causes diabetes. Both hypoglycemia and diabetes increase the vulnerability to suffering a stroke. (Right graph in Figure 3)

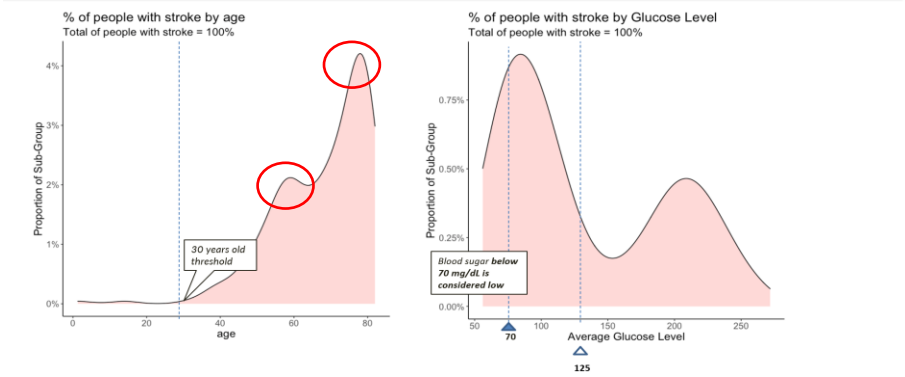


Figure 3 Age vs Stroke & Avg Glucose Level vs Stroke

iii) Hypertension, heart disease and smoking behaviours are also critical factors in suffering a stroke. While hypertension and heart disease are having a direct impact on the blood supply system, chemicals in cigarettes such as carbon monoxide and nicotine will increase heartbeat rate and reduce blood oxygen level, which increases the likelihood of suffering a stroke. We can see these reflected in the bar chart in Appendix VI from our data showing people with hypertension, heart disease and smoking increase the likelihood to have a stroke in adults.

On the other hand, we found three "Relieving factors" that were negatively correlated with getting a stroke: becoming self-employed, finding a government job, or working in a private company. (Please refer to Appendix V ii) *Interpretation of the model result – Negative Coefficients*). Also, it is worth noting that, although we discovered having a job will be negatively correlated with having a stroke, it doesn't mean working will lower the chance of getting a stroke, or vice versa. This can be a correlation vs a causation issue as correlation does not imply causation. Moreover, there may be some underlying confounding factors which are having an impact on the potential causality of certain variables. For example, having a job implies more disposable income which could probably afford a person to engage in healthier behaviours (exercising, eating healthy, etc.). Due to the limited data in our sample, it ends up showing having a job will probably help reduce the probability of having a stroke. Currently, this is still a preliminary study and the findings we have are based on limited clinical records. We will incorporate a larger dataset and more detailed factors such as average household income, lifestyle, location etc. to improve our model's predictive accuracy in the next phase study.

Recommendation

Our model has identified both risky factors and relieving factors. Our objective of the recommendation is

to increase public awareness of stroke and lower the chance of getting strokes through early prevention using our discovery and prediction model, and it will be best if Canadians can take those findings and apply the model to themselves easily.

We recommend that PHC can start developing a free stroke prediction app so that Canadians can input their related information and see the prediction themselves (Demo App user face in Appendix VII), and also the app would include features that record users' historical data and generate visualization like monthly trend charts, and users can closely monitor their health conditions month by month. We have created a hypothetical user profile for demonstration– John. His predicted stroke rate is 2.4% in June 2022, and users can compare with the hypothetical average stroke rate of 2.3%, so user can monitor their information compared to the average. Please see the demo line chart in Appendix VIII i) *Monthly generated Stroke Prediction rate*. Additionally, the app users can click on the data point in the chart, and it will generate a table to show what factors contribute to the prediction result (demo table in Appendix VIII ii) *Table Result on detailed factors)*, so they can take suggestions to alter their lifestyle. The app will act as an online public education platform for stroke, and it will encourage users to update their information and learn about stroke prevention through mini-games, monthly stroke question challenges, and small prizes, and the main idea is to bring awareness to their predicted chance of getting a stroke, so they can take preventive action as early as possible.

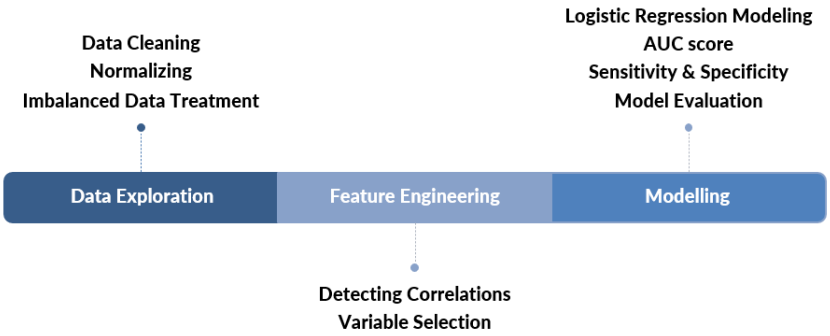
We also understand that more quality data lead to better prediction and public policy development, and our government can collect data from the app users with their consent, for example, new data that our dataset didn't have such as income level, lifestyle, location, and other predictive variables, and we believe it can further our analysis. Moreover, non-profit organizations such as "Heart and Stroke" can connect to our model and database to leverage our findings and make collaboration.

The second recommendation is to launch health promoting campaign, "Stroke Awareness Month", to bring awareness about the high-risk factors we found in our model, and we can promote healthier behaviours to nudge Canadians, reduce the likelihood of getting a stroke and also explain between controllable and non-controllable factors. For example, smoking and eating healthier are controllable factors and those are choices that can help reduce the chance of getting a stroke, but age is an uncontrollable factor. Also, we can educate the public on how jobs are related to stress level and lifestyles. During Stroke Awareness Month, we can promote the free app to the public, so it will increase the usage of the app and connects our findings with the physical event.

In a nutshell, we hope our government can apply our findings both online and offline to increase awareness of stroke and lower stroke instances through early prevention.

Appendix

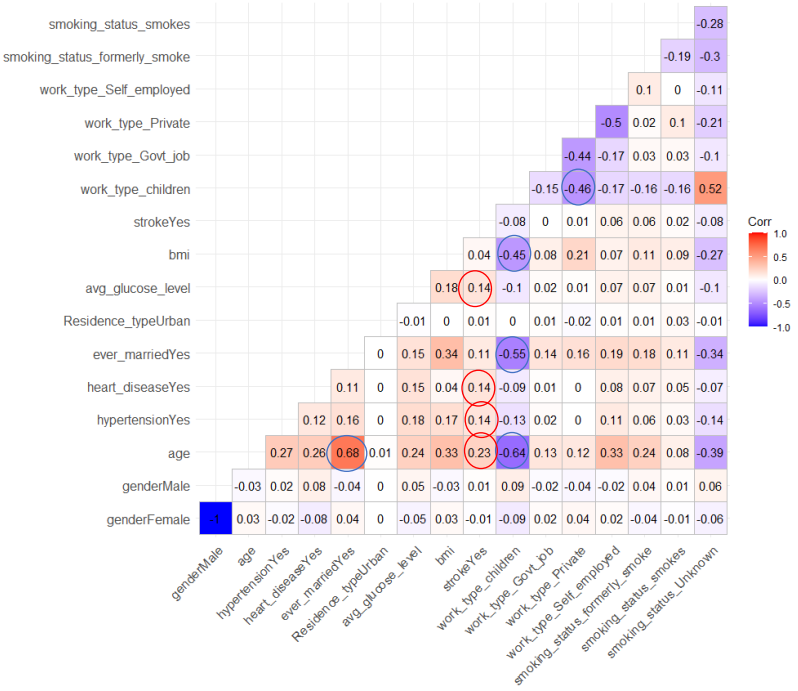
I: Data Analysis Process



II: Data Structure and Variables

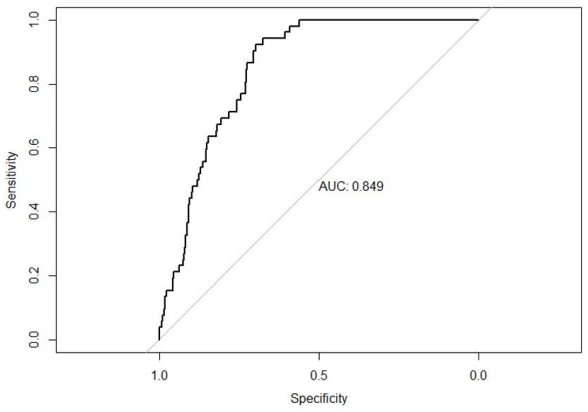
Data	**Type**	**Comments**
Basic Information (Age, Gender, Marital Status, Residence Type, Work Type)	Numerical, Categorical	Some personal information about the individuals on record
BMI (Body Mass Index)	Numerical	Indicating Body Mass Index
Hypertension	Categorical	Have hypertension or not
Heart Disease	Categorical	Have heart disease or not
Average Glucose Level	Numerical	Indicating the risk of diabetes
Smoking Status	Categorical	Smoked before, Smoking or never smoked
Have Stroke or not	**Categorical**	**Response Variable**

III: Correlation Matrix



IV:

i) Final Selected Model Result– ROC and Confusion Matrix



Pre-fix Confusion Matrix (Imbalanced)

```
> confusion.matrix<-table(Stroke_Test$stroke, pred_glm >= 0.5)
> confusion.matrix

     FALSE
No    1174
Yes     52
```

Post-fix Confusion Matrix (Default Threshold)

```
> confusion.matrix_opt <- table(Stroke_Test$stroke, pred_glm_opt >= 0.5)
> confusion.matrix_opt

     FALSE TRUE
No     876  298
Yes     13   39
```

Sensitivity: 0.750 **Specificity**: 0.746

Post-fix Confusion Matrix (Optimal Threshold)

```
> confusion.matrix_opt_t <- table(Stroke_Test$stroke, pred_glm_opt >= 0.3678)
> confusion.matrix_opt_t

     FALSE TRUE
No     753  421
Yes      3   49
```

Sensitivity: 0.942 **Specificity**: 0.641

Increased false positive is acceptable in healthcare due to extra cautious

ii) Three model results comparison.

	Basic (imbalanced)	Lasso	glm w/ selection
Accuracy - 0.5			
AUC	0.856	0.848	0.849
LAMBDA		0.002811	
(Intercept)	-18.354302	-5.11213	2.00E-16 ***
Gender Male	-0.009059	.	
Gender Other		.	
age	0.068309 ***	0.070734	1.73E-08 ***
Hypertension Yes	0.603319 **	0.589536	2.13E-10 ***
Hear Disease Yes	0.577482 *	0.891943	0.9462328 ***
Ever Married Yes	-0.067497	.	
Residence Type - Urban	-0.020916	.	
Avg Glucose Level	0.004805 **	0.003269	6.88E-05 ***
BMI	0.01194	0.014828	0.00192 **
Work Type - Children	11.07688	0.938271	
Work Type - Government	10.257451	-0.14571	4.45E-06 ***
Work Type - Private	10.36774	.	2.84E-05 ***
Work Type - Self Employed	10.138477	.	4.69E-05 ***
Smoking Status - Formerly Smoke	-0.031642	-0.06606	
Smoking Status - Smokes	0.359675	0.422492	1.12E-05 ***
Smoking Status - Unknown	-0.199679	-0.2265	0.02039 *

V:

i) Interpretation of the model result – Positive Coefficients

Positive Coefficient Interpretation

```
Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -4.1291188  0.3093658 -13.347  < 2e-16 ***
age                          0.0743272  0.0032197  23.085  < 2e-16 ***
hypertensionYes              0.6190910  0.1098293   5.637 1.73e-08 ***
heart_diseaseYes             0.9272672  0.1459957   6.351 2.13e-10 ***
avg_glucose_level            0.0032484  0.0008161   3.980 6.88e-05 ***
bmi                          0.0189640  0.0061132   3.102  0.00192 **
work_type_Govt_job          -1.5381996  0.3351648  -4.589 4.45e-06 ***
work_type_Private           -1.3434440  0.3209377  -4.186 2.84e-05 ***
work_type_Self_employed     -1.3916400  0.3418594  -4.071 4.69e-05 ***
smoking_status_smokes        0.5070445  0.1154102   4.393 1.12e-05 ***
smoking_status_Unknown      -0.2596373  0.1119569  -2.319  0.02039 *
```

	Coefficient	Odd Increment	Mean value	Odd (Mean)
Age	0.0743272	0.077159194	42.73	3.297012374
Heart Disease	0.9272672	1.527592326		1.527592326
Hypertension	0.619091	0.857239044		0.857239044
Smoke	0.5070445	0.660376693		0.660376693
Glucose	0.0032484	0.003253682	104.83	0.34108346
BMI	0.018964	0.019144959	28.87	0.552714959

```
> summary(Stroke_Train$bmi)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  10.30   23.50   28.00   28.87   33.00   97.60
> summary(Stroke_Train$age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.08   25.00   44.00   42.73   60.00   82.00
> summary(Stroke_Train$avg_glucose_level)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  55.12   76.81   91.41  104.83  112.95  271.74
```

Top factors for stroke:
Age > Heart Disease > Hypertension > Smoke

Odd (Mean) means the odd increment for mean values in the sample. For example mean age is 42.73, then the mean possibility of getting a stroke due to age is increased by 3.297

Note that Glucose level and BMI turns out about same impactful.

Age is the biggest factor for sure

ii) Interpretation of the model result – Negative Coefficients

Negative Coefficient Interpretation

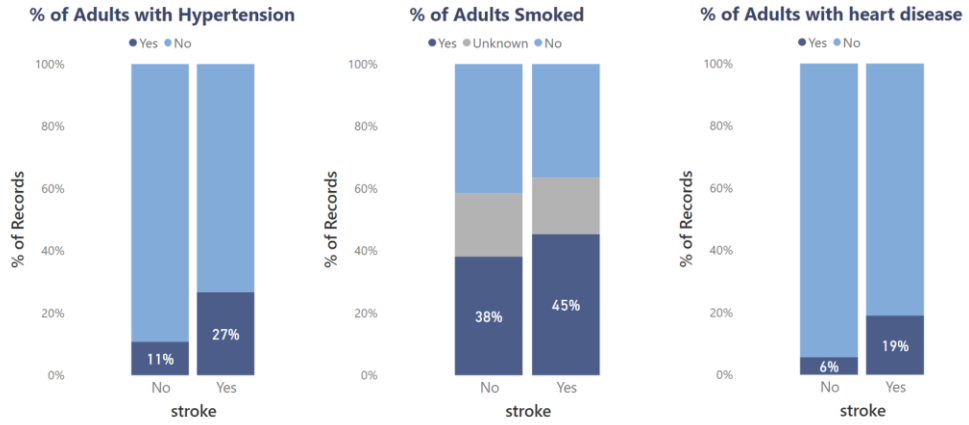
```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -4.1291188  0.3093658 -13.347  < 2e-16 ***
age                      0.0743272  0.0032197  23.085  < 2e-16 ***
hypertensionYes          0.6190910  0.1098293   5.637 1.73e-08 ***
heart_diseaseYes         0.9272672  0.1459957   6.351 2.13e-10 ***
avg_glucose_level        0.0032484  0.0008161   3.980 6.88e-05 ***
bmi                      0.0189640  0.0061132   3.102  0.00192 **
work_type_Govt_job      -1.5381996  0.3351648  -4.589 4.45e-06 ***
work_type_Private       -1.3434440  0.3209377  -4.186 2.84e-05 ***
work_type_Self_employed -1.3916400  0.3418594  -4.071 4.69e-05 ***
smoking_status_smokes    0.5070445  0.1154102   4.393 1.12e-05 ***
smoking_status_Unknown  -0.2596373  0.1119569  -2.319  0.02039 *
```

	Coefficient	Odd Reduction
Gov Job	-1.5381996	0.785232579
Self Employed	-1.39164	0.751332844
Private	-1.343444	0.739054577

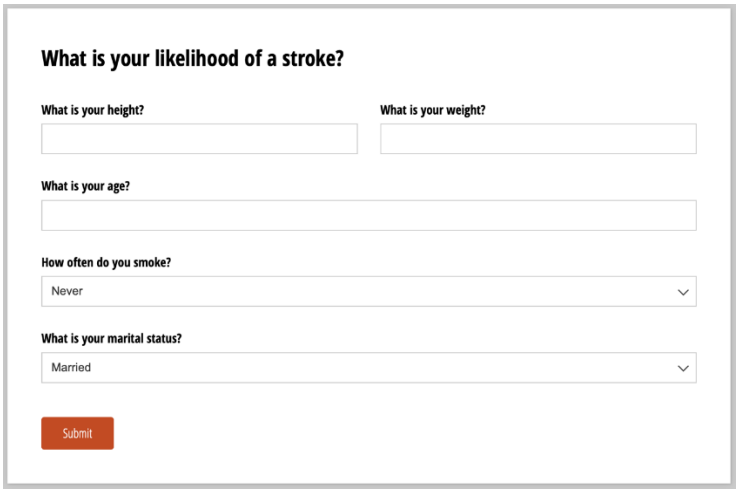
Top 3 factors to REDUCE the stroke:
Jobs at Government > Self-employed > Private jobs

All in all, relax to prevent strokes :D

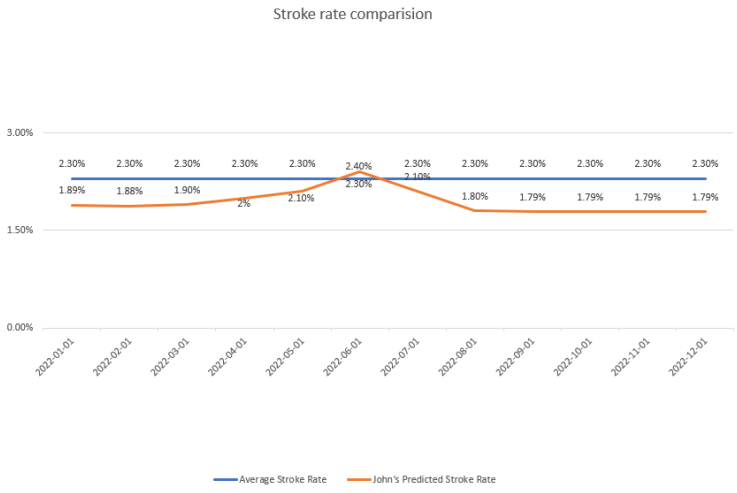
VI: Hypertension, Heart Disease and Smoking increase the likelihood to have a stroke in adults



VII: Demo App user face look



VIII: i) Demo monthly generated Stroke Prediction rate.



ii) Demo table Result on detailed factors.

The following tables show what may be contributed to the 2.4% increase
Glucose level is higher than in the past 4 months
The user indicated himself smoking in the recent month

References

Fact Sheet: Stroke Statistics. (n.d.). Retrieved from ontariostrokenetwork.ca: http://www.ontariostrokenetwork.ca/pdf/Final_Fact_Sheet_Stroke_Stats_3.pdf

Preventing Stroke Deaths. (2017, September 6). Retrieved from cdc.gov: https://www.cdc.gov/vitalsigns/stroke/index.html

What we do. (2022). Retrieved from world-stroke.org: https://www.world-stroke.org/what-we-do