Stroke Prediction: An Analysis of Health and Stroke Data

John Dinh

2022-03-17

1. Abstract

Predicting stroke is an important tasks for clinicians in order to provide the correct preventive measures as well as provide helpful prognosis measures. A data set of 5,109 patients was analyzed and a logistic model was fit in order to predict the probability of incidence of stroke based on lifestyle, demographic, and clinical covariates. The model reported a balanced accuracy of .77 and concludeded that nearly every single variable was statistically significant at the task of predicting the incidence of stroke.

2. Introduction

Stroke is a prevalent disease and is one of the leading causes of deaths in the United States, as more than 700,000 people have strokes in the United States every year. Stepping back, stroke is a disease that affects the arteries leading to the brain. The diesase occurs when a blood vessel that carries oxygen to brain suddenly erupts or is blocked by a clot, and as a result, the brain can not receive blood and oxygen that it needs. This causes the brain cells to die and the brain ends up not functioning correctly. The disease in fatal in the sense that someone having a stroke requires immediate emergency treatment. Without treatment, strokes can lead to disability or death.

Although stroke is a prevalent disease, it is also a preventable disease. Some measures to take in order to prevent stroke is to eat a healthy diet, exercise regularly, avoid smoking, and maintain a healthy weight. Lifestyle decisions are an important fact to deter and prevent strokes.

In order to analyze what factors are most important when predicting whether or not someone will have a stroke, it is important to analyze an individual's lifestyle habits such as work type, whether or not they are a smoker, or their weight, to name a few. Predicting and modeling stroke risk is an important task for the early treatment and care of potential stroke patients, and is also widely helpful for diagnosis of the disease.

3. Dataset and Question of Interests

3.1 Data

The data set to be analyzed in this project is an open source data set from Kaggle (Fedesoriano, 2021). The data set contains lifestyle, demographic, and clinical measurements for 5,109 patients. The following variables were used in this analysis:

1. Gender
2. Age
3. Hypertension status

4. Heart disease status
5. Marriage status
6. Work type
7. Residence type
8. Average glucose level
9. BMI
10. Smoking status
11. Stroke status

The primary response variable used in this analysis was `Stroke status`, which indicates whether or not the patient has had a stroke. The variables `Gender`, `Hypertension status`, `Heart disease status`, `Marriage status`, and `Residence type` are all binary variables indicating whether or not that individual contained the status named in the variable or not, encoding as 0 or 1. `Gender` had three levels, `Male`, `Female`, and `Other`, but the `Other` observation was dropped since there was only one observation with that level. The variable `Residence type` was encoded as `0 -Rural` and `1-Urban`.

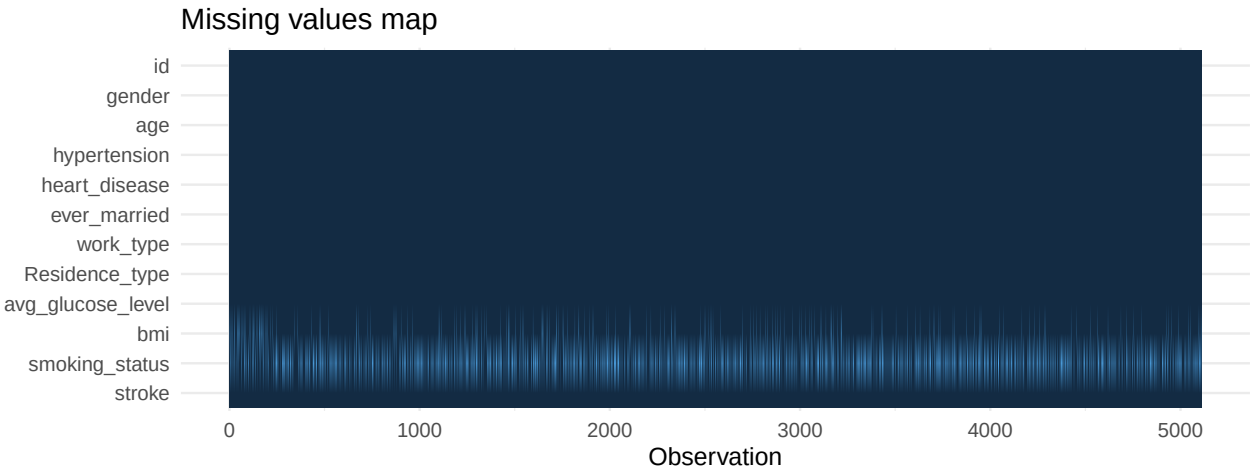
The variable `Work type` is a 5 level variable with factors responding to:

- `0-Children`
- `1-Government job`
- `2-Never worked`
- `3-Private`
- `4-Self employed`

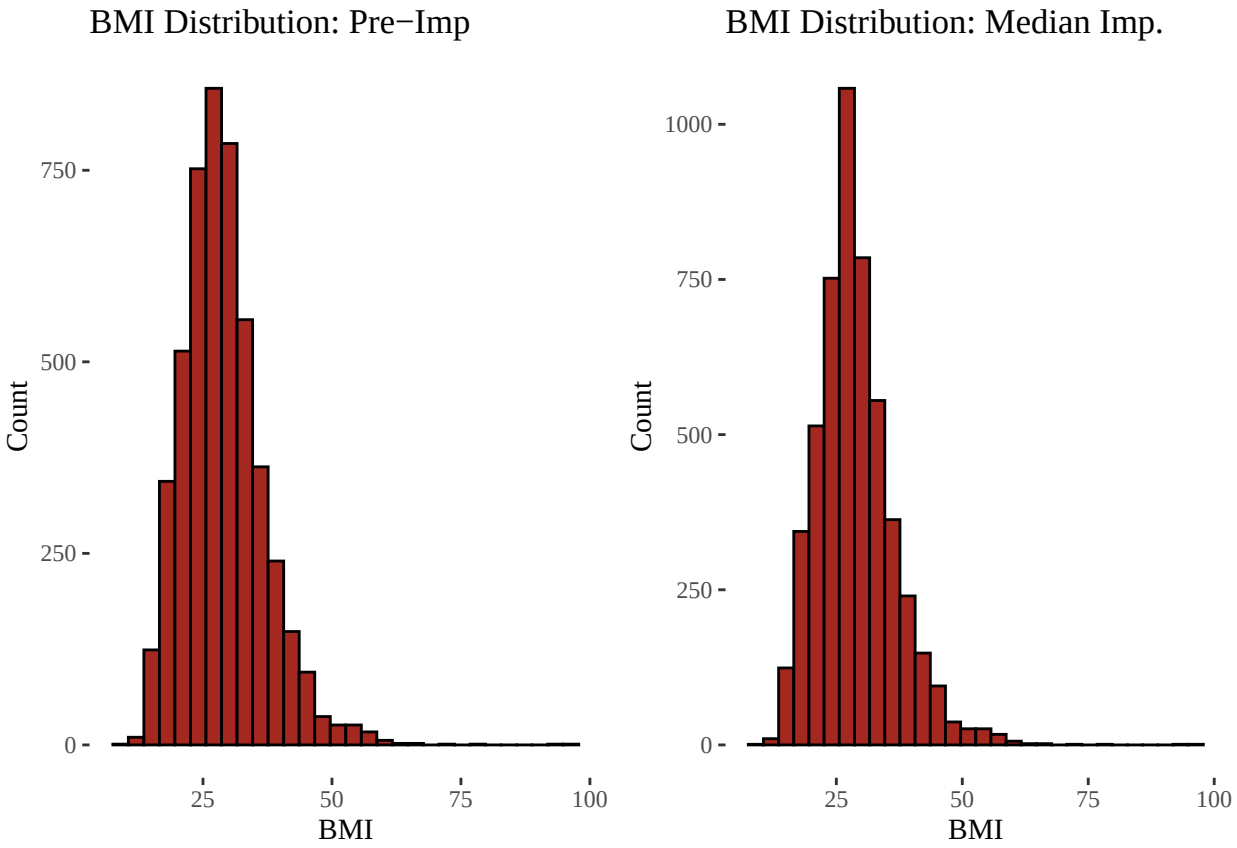
The variable `Smoking status` is a 3 level variable with factors responding to:

- `0-formerly smoked`
- `1-never smoked`
- `2-Smokes`
- `3-Unknown`

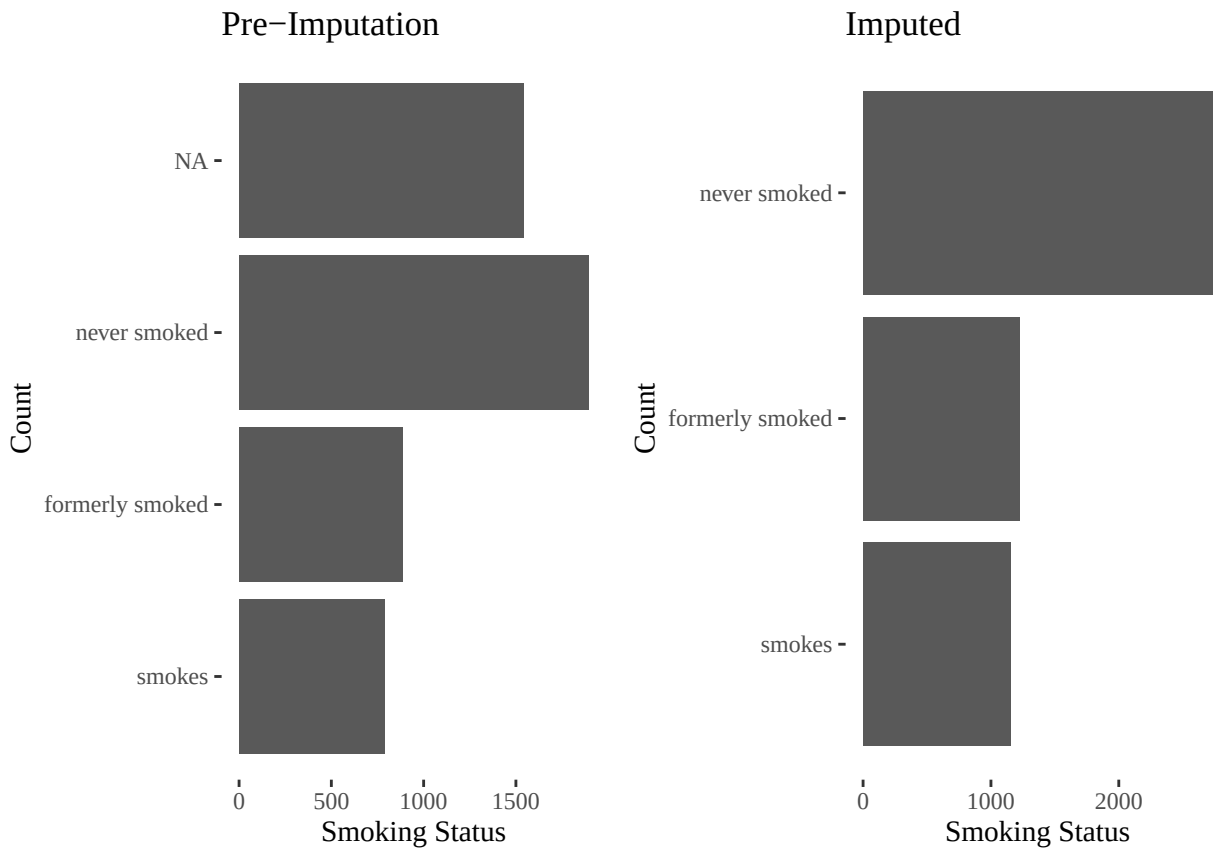
An assessment of the missing observations was performed prior to any analysis or modelling. Below is a plot showing the missingness of the data after some variable conversion. It was determined that the missing observations were missing completely at random. Because of this, missing values for `BMI` were imputed using the median since the distribution was slightly right skew.



The distribution of BMI did not change as much and is slightly right skewed. Due to the lack of clinical background as well as wanting to avoid information loss, the median values will remain in the dataset, and the analysis of the model and fit will be assessed with this fact in mind.



During the data cleaning process, there were were over 1500 observations with an `Unknown` smoking status. It was determined that the missing values for this variable were missing completely at random, so in order to avoid information loss, random values of smoking status were imputed for observations that reported `Unknown`. The distribution of `smoking status` before and after the imputation is plotted below.



3.2 Question of Interest:

The primary question of interest is to be able to predict whether or not an individual will have a stroke based on lifestyle variables. In addition to this primary question, another question of interest is to determine which variables are more influential to others for predicting the odds of having a stroke.

This analysis is important in order for clinicians to develop models to predict whether or not a patient has a stroke as part of their diagnosis, as well as to recommended prognosis treatments based on their diagnoses.

4. Exploratory Data Analsysis

A general overview of the data is looked at. The summary statistics is in Table 1 below gives a general look at the data. As mentioned before, the primary concern with the data set is the class imbalance for `stroke`. There are 4,860 observations for `0`, corresponding to 4,860 patients never having a stroke compared to the 249 patients that have. This is a massive imbalance in class and will be addressed below. The average BMI of this sample is also 28.86, which is in the overweight range for body mass index. Although BMI is now a largely useless metric in the health sciences (Oesch et al. 2017) since it does not take into account age, bone structure, and fat distribution, it can still be used to give a general overview about stroke patients in the data set. Also, it is important to note the outliers in the BMI variable as well. The max value reported was 97.60, which seems to be unlikely. Dropping this observation seems to be the best course of action.

Another interesting fact is that the average age in this data set is 43. This seems reasonable since strokes more than 70% of strokes occur from patients 65 and older(Kelly 2010). For this, it is expected to see age as a significant predictor to predict the incidence of stroke. In the same variable, the minimum age reported

was .08, which may refer to infant that is around 9 months old. Since it is possible for infants to have strokes, this observation will not be removed.

Table 1: Summary Statistics

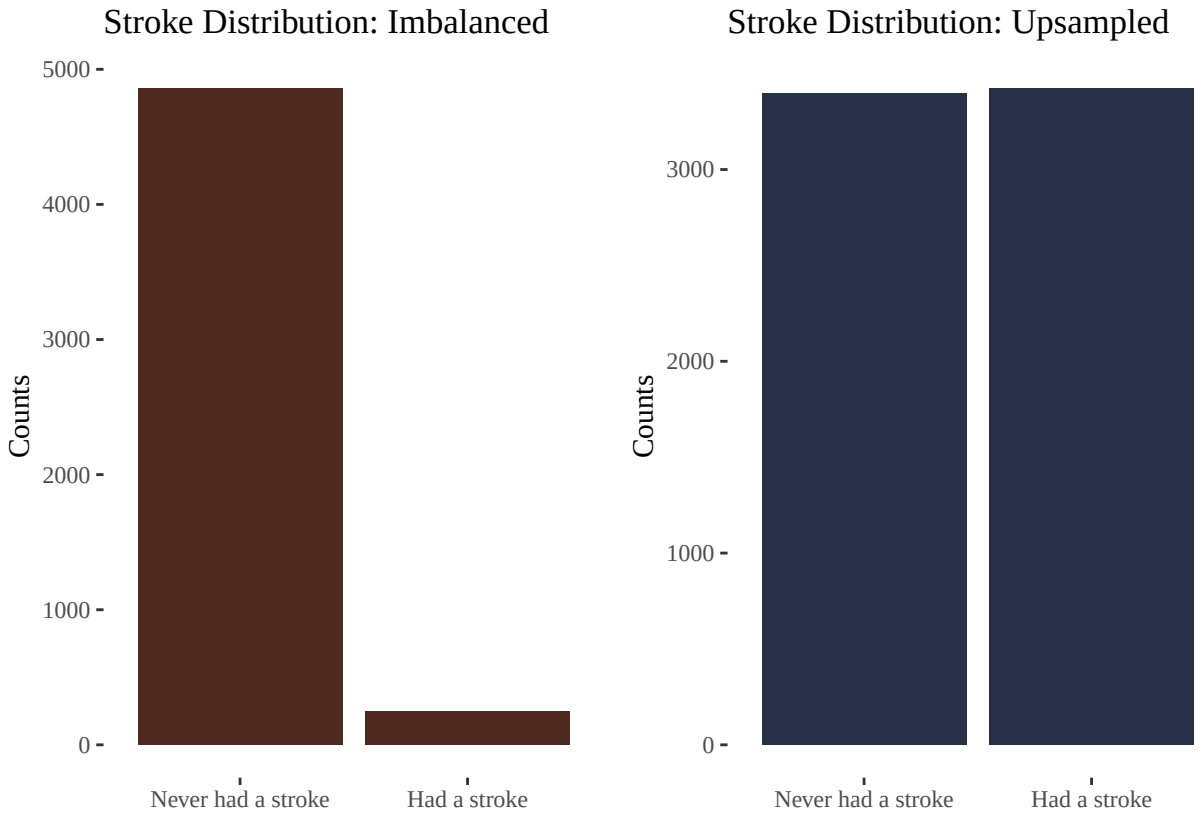
	gender	age	hypertension	heart_disease
	Female:2994	Min. : 0.08	0:4611	0:4833
	Male :2115	1st Qu.:25.00	1: 498	1: 276
	Other : 0	Median :45.00		
		Mean :43.23		
		3rd Qu.:61.00		
		Max. :82.00		

	ever_married	work_type	Residence_type
	No :1756	children : 687	Rural:2513
	Yes:3353	Govt_job : 657	Urban:2596
		Never_worked : 22	
		Private :2924	
		Self-employed: 819	

	avg_glucose_level	bmi	smoking_status	stroke
	Min. : 55.12	Min. :10.30	Length:5109	0:4860
	1st Qu.: 77.24	1st Qu.:23.80	Class :character	1: 249
	Median : 91.88	Median :28.10	Mode :character	
	Mean :106.14	Mean :28.86		
	3rd Qu.:114.09	3rd Qu.:32.80		
	Max. :271.74	Max. :97.60		

A crucial step prior to any statistical analysis or modeling is the exploratory stage. While this stage is time consuming, the benefits are ten-fold and can help spot issues prior to modeling. In addition, this stage helps reveals trends in the data in order to know what to expect during the model fitting and diagnostics stages. With that being said, the first variable to inspect is the response variable, `Stroke`. In this dataset, there is a heavy imbalance, as show in the graph below, of the classes in this variable. This will be a problem for predictive models in that a model that is trained on a heavily imbalanced dataset will do well predicting the majority class, but not so well predicting the minority class. In other words, the model will have a bias towards the majority class. Outside of this stroke data set, class imbalance occurs in data such as fraud detection, churn prediction, spam detection, etc. and must be dealt with accordingly.

There is no doubt that `Stroke` is heavily unbalanced, as there is less than 1000 observations of patients that have had a stroke compared to the almost 5,000 patients that have not. In order to deal with this imbalance, an oversampling method was used. Under-sampling was not considered since the data set is relatively small and removing observations based on the majority class would lead to information loss. In that case, oversampling was used in order to replicate the observations of the minority class until both classes of observations were relatively equal. The downside of this approach is that the model may not be as generalize-able since replicating observations will likely cause overfitting to occur.



It is important to note that these resampling methods should be done only on the training split of the data. Simply put, balancing the validation split is not indicative of real world data. For example, stroke data in the real world would be heavily imbalanced, so assessment of the model should be done on a realistic test split instead of an artificially generated split. The class frequencies for the validation set should be representative of what a scientist observes in the real world (Kuhn 2019). For this reason, resampling is done after splitting the data into a training and validation set.

The following table shows the summary statistics for numeric variables in the training split.

Table 2: Numeric Summary Statistics

Statistic	N	Mean	St. Dev.	Min	Max
age	6,821	54.703	22.240	0.080	82.000
avg_glucose_level	6,821	117.991	55.886	55.120	271.740
bmi	6,821	29.434	6.807	10.300	97.600

Next, the data was reviewed for missing entries or incomplete observations. Discrepencies in the data were adjust to prevent any issues. Factor variables such as `gender`, `hypertension`, `heart_disease`, `marriage`, `work`, `residence`, and `smoking status` were all converted to factor variables instead of character variables. After the imputation and adjustment to these variables were performed, a full analysis of the following variables could be performed:

Table 3: Predictor Variables and Response Variable, 'stroke'

Variables	Description
gender	Male or Female
age	Age of respondent
hypertension	Whether or not the respondent has hypertension
Heart_disease	Whether or not the respondent has heart disease
ever_married	Whether or not the respondent is married
work_type	What type of job the respondent has
Residence_type	Urban or rural
avg_glucose_level	Numeric glucose level of the respondent
bmi	Numeric value of the body mass index
smoking_status	Whether or not the respondent is a smoker
stroke	Whether or not the respondent has had a stroke before

4.1 Data Visualization

Prior to fitting models, further analysis of the relationship between the responses variable and other independent variables was conducted. In Fig 1, the distribution of ages for the two classes of patients was plotted. It can be inferred from the plot that the distribution of ages for those that receive strokes are skewed towards older patients, with most of the patients having strokes being between 60 and 80 years of age.

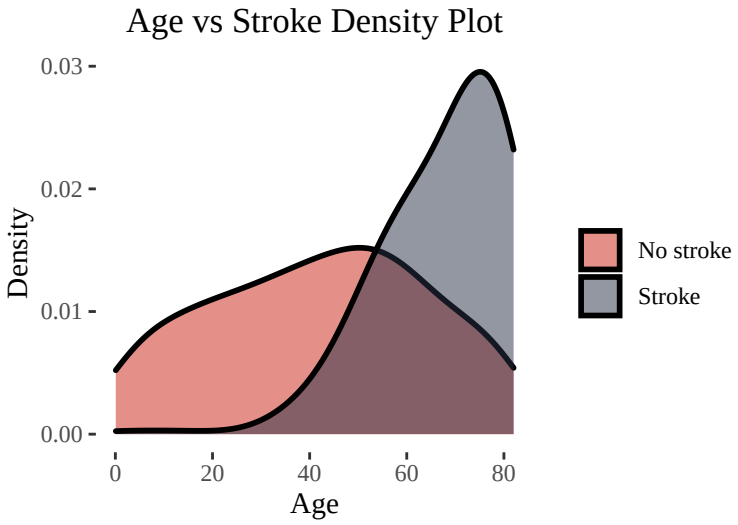


Fig 1: Density of Age on Stroke

A similar plot (Fig. 2) was produced using BMI in order to see the distribution of body mass index between patients that had a stroke and those that didn't. There is not an observable difference in the distributions, although the BMI of those that did not have strokes skew slightly to the right more. An interpretation of this figure there is more stroke patients with a BMI in the 25 range whereas non-stroke patients' BMIs are more spread out.

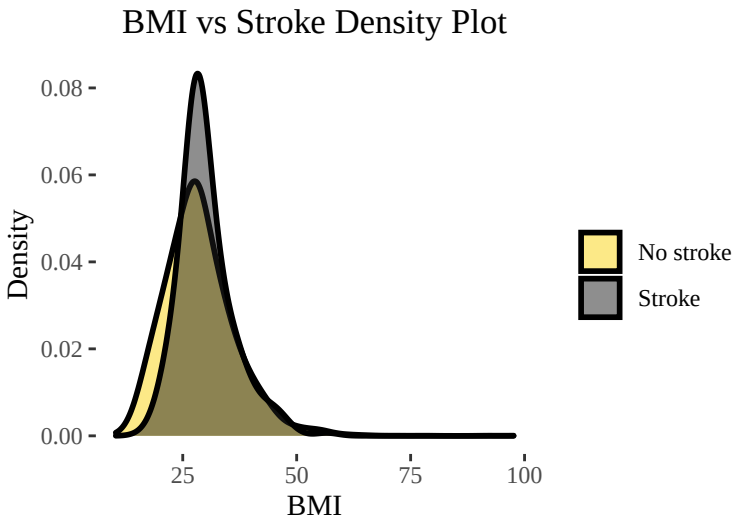


Fig. 2: Density of BMI on Stroke

Finally, a the following plot (Fig. 3) juxtaposes average glucose level with between stroke patients. The primary reason for including this plot is to see if there are trends for those that have high blood sugar levels with incidence of stroke. A majority of patients that have never had a stroke have average blood sugar levels of around 100, whilst the distribution of blood sugar levels for those that have had strokes are more varied. The conclusion from this plot is that Lower blood sugar levels may indicate a lower risk in stroke.

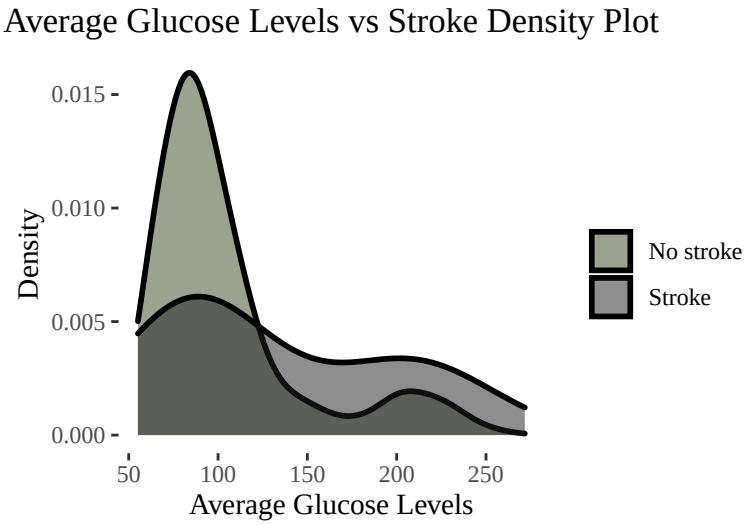


Fig. 3: Density of Glucose on Stroke

Next, an analysis of the trends with `age` and `bmi` are is performed to see if there are trends with older patients and whether or not BMI affects the incidence of having a stroke. In Fig. 4, it is seen that the older population trends with incidence of stroke, but there is no obvious pattern for stroke incidence with BMI. This coincides with the analysis earlier with the distribution of BMI having no clear indication of stroke incidence. It is also important to note that there are a few outliers in this plot, namely the points of stroke with patients that are under 20 years of age. The observation for the patient that is 1 year old and has a BMI of 25 seems suspicious and will be removed.

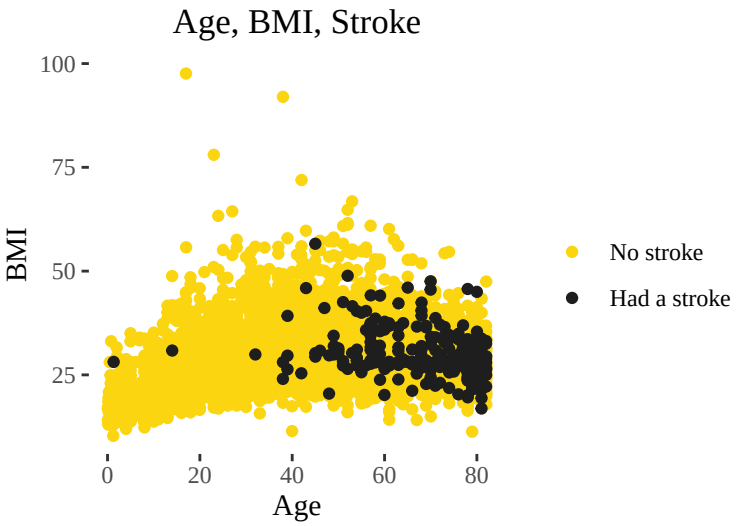


Fig. 4: Relationship between Stroke Incidence and Age/BMI

Similarly, age and stroke incidence was also assessed with average glucose level. It is seen in Fig. 5 that stroke trends with older patients, but the also trends with patients on the lower end and higher end of average glucose levels, but not the middle range of glucose levels. This is an interesting relationship to view since it indicates that an abnormal average glucose level is perhaps an indicator of stroke incidence. Revisitng the outlier mention in Fig. 4, perhaps it is best not to remove the observation since the case seems to be that the young patient had a stroke and the BMI was miscalculated due to the median imputation from before. This point serves no purpose in terms of analyzing relationships for diagnoses, but will help with model fitting and predictive power, which is the primary focus of this analysis.

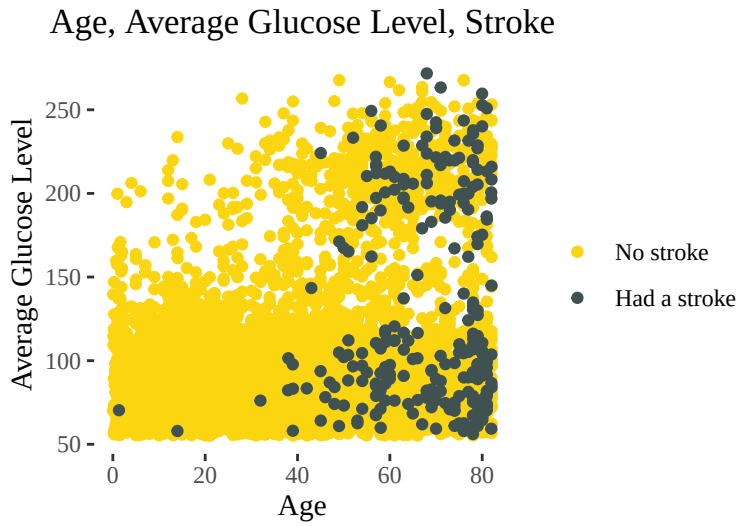
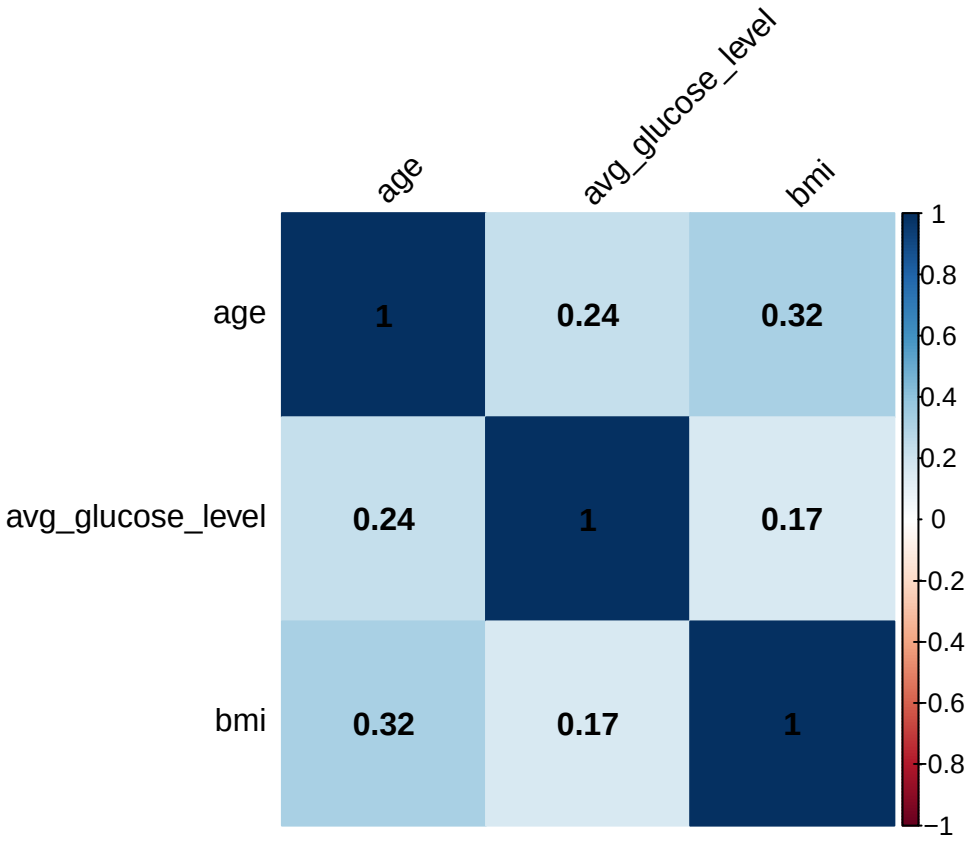


Fig. 5: Relationship between age/glucose levels and stroke

A key step is to also check the correlation between these numeric predictors. Although age is slightly positively correlated with bmi, it is not large enough to be egregious, so the analysis will continue as is.



4.2 Categorical Variables

Between males and females, there are a slightly more cases of stroke for males than there are for females.

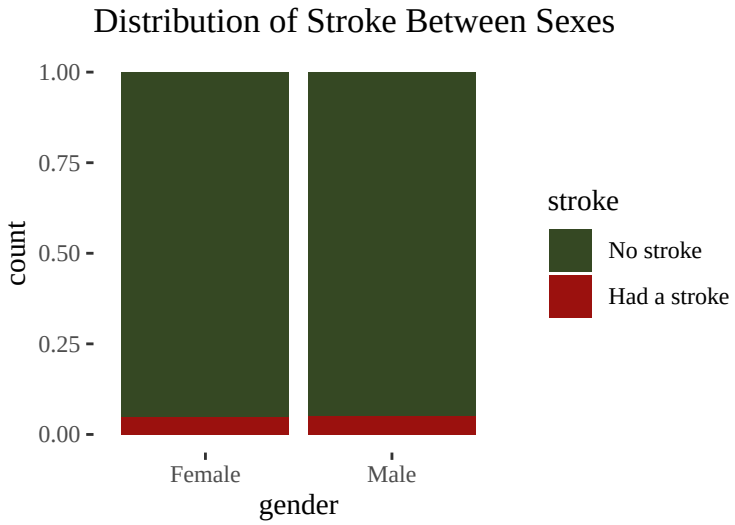


Fig. 6: Relationship between Sexes/Stroke

From Fig. 7, it can be inferred that more patients that have reported having a stroke have also been former smokers. The current smokers have about the same amount of cases as patients that have never smoked. It could be hypothesized that the long term effects of smoking is what impacts the incidence of having a stroke.

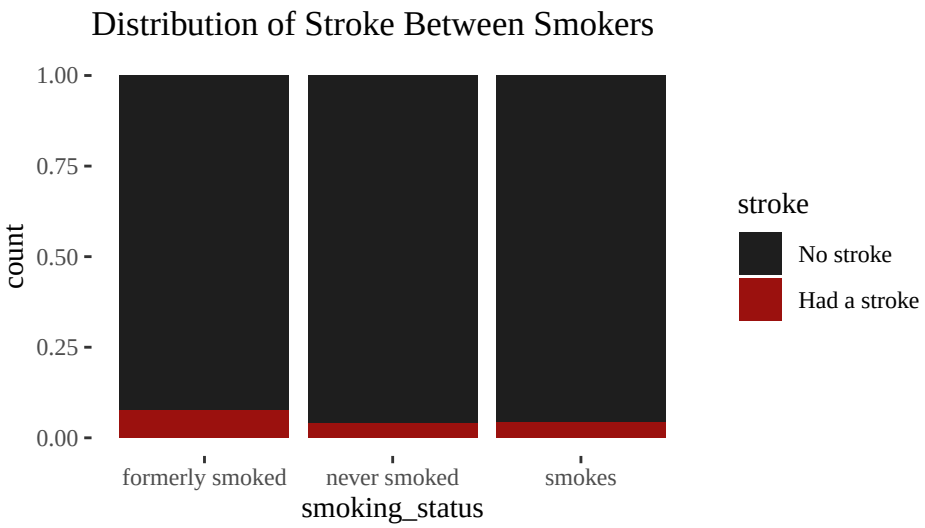


Fig. 7: Relationship between Smokers/Stroke

From Figures 8 and 9, it can be seen that individuals with hypertension and heart disease are also associated with more cases of stroke. This is expected since problems with the heart will cause troubles for blood to reach the brain.

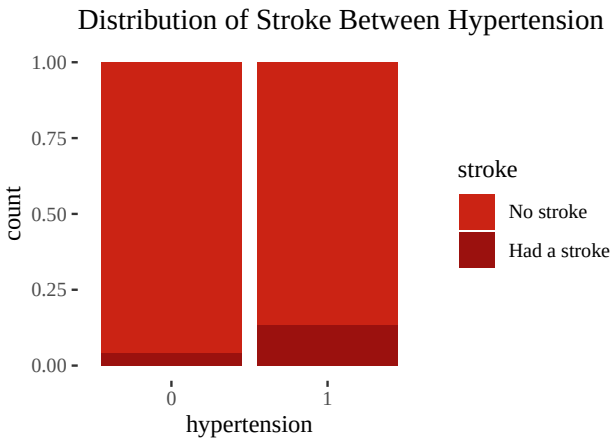


Fig. 8: Relationship between Hypertension/Stroke

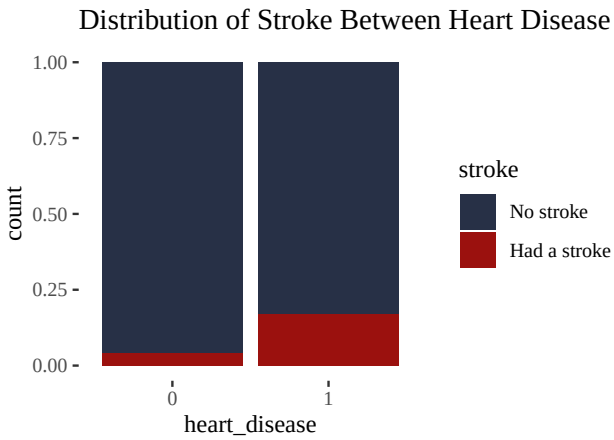


Fig. 9: Relationship between Heart Disease/Stroke

From Figures 10 and 11, it can be seen that stroke incidence occurs much greater in those that have worked compared to children and those that have never worked. In addition, the distribution of stroke between residence type is somewhat equal.

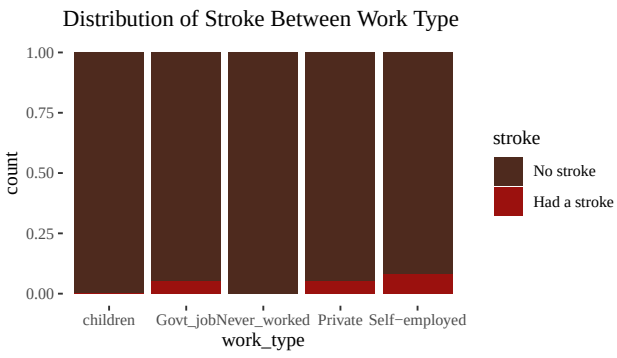


Fig. 10: Relationship between WorkType/Stroke

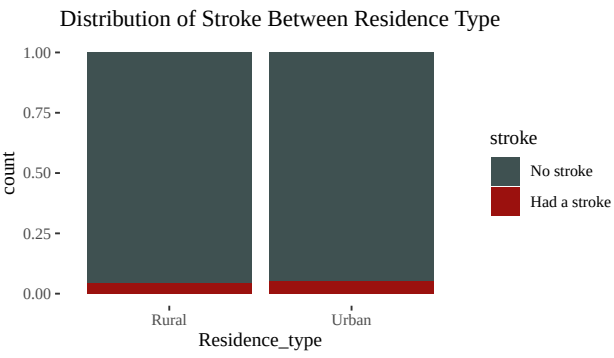


Fig. 11: Relationship between Residence/Stroke

5. Model Fitting

5.1 Model Equation

The primary question of interest was to be able to predict whether or not an individual will have a stroke based on lifestyle variables. In order to answer this question reasonably, a logistic model was fit in order to predict whether the incidence of stroke given a list of covariates. Feature engineering was done prior to fitting the model in order for easy interpretation of the model. The formula for a logistic model is as follows:

$$\pi(\boldsymbol{X}) = \frac{exp(\beta_0 + \beta_1 X_1 + ...\beta_k X_k)}{1 + exp(\beta_0 + \beta_1 X_1 + ...\beta_k X_k)} = \frac{exp(\boldsymbol{X}\beta)}{1 + exp(\boldsymbol{X}\beta)}$$

Where,

- $\pi(\boldsymbol{X})$ is the probability that an observation is in the specified category of the response variable, `stroke`
- $\boldsymbol{X}$ is design matrix of covariates, in this case, the design matrix using the predictors as mentioned in Table 1.
- β are the regression coefficients.

5.2 Feature Selection

An initial full model was fit and in Table 9, it is shown that all the predictors are significant except `BMI`. In the exploratory data analysis, the analysis conveyed that the distribution of BMI did not heavily impact the trends of stroke incidence. In other words, the distribution of BMI within patients that had a stroke did not have a clear trend. This is reflected in the significance of the BMI predictor. With this, a reduced model was fit excluding the `BMI` predictor. Using a Pearson chi-square goodness of fit test with a p-value of .9, it was found that the model without `bmi` as a predictor fit the data just as well as the full model. By the principle of parsimony (Occam's Razor), the more simple model was chosen.

Table 4: Pearson Chi-Square Test

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
6807	6429.966			
6806	6429.952	1	0.0146168	0.9037703

In addition, the stepwise procedure was used to determine the most optimal model. The Bayesian Information Criterion (BIC) was used instead of Akaike's Information Criterion (AIC) was used since BIC penalizes complex models much more compared to AIC. The reason why a more simple model is preferred due to risk of overfitting. The training split was already upsampled, which increases the chance of overfitting the model. Therefore, reducing the model bias in other ways is necessary for good real world performance. The final model output pictured in Table 3 below, of the step wise procedure is the same model in Table 10, the full model without the `BMI` predictor.

Table 5: Stepwise Procedure, Final Model

term	estimate	std.error	statistic	p.value
(Intercept)	-3.1075208	0.1852857	-16.7715131	0.0000000
genderMale	-0.1356421	0.0648629	-2.0912118	0.0365091
age	0.0831471	0.0025483	32.6280399	0.0000000
hypertension1	0.5818425	0.0843883	6.8948217	0.0000000
heart_disease1	0.4452974	0.1035059	4.3021445	0.0000169
ever_marriedYes	0.3166174	0.1069969	2.9591282	0.0030851
work_typeGovt_job	-2.1886375	0.2359773	-9.2747806	0.0000000
work_typeNever_worked	-12.7894343	212.9412505	-0.0600609	0.9521072
work_typePrivate	-2.1928300	0.2267485	-9.6707593	0.0000000
work_typeSelf-employed	-2.6262107	0.2450970	-10.7149852	0.0000000
Residence_typeUrban	0.1405616	0.0623535	2.2542697	0.0241792
avg_glucose_level	0.0033041	0.0005845	5.6525007	0.0000000
smoking_statusnever smoked	-0.4700953	0.0731251	-6.4286460	0.0000000
smoking_statussmokes	-0.1666734	0.0879937	-1.8941515	0.0582049

In addition to this primary question, another question of interest is to determine which variables are more influential to others for predicting the odds of having a stroke.

Assessing Goodness of Fit

In order to assess the goodness of fit of the model, first there needs to be an assessment whether or not over dispersion is an issue in the model. This test is to just confirm that each observation is independent and not correlated with each other. The following test statistic is calculated by:

$$\chi^2 = \sum_{i=1}^{n} \frac{(Y_i - \hat{Y})^2}{df_{residuals}}$$

The reduced model has a value of the Pearson statistic at .945. This indicates that there is no significant over dispersion in the model.

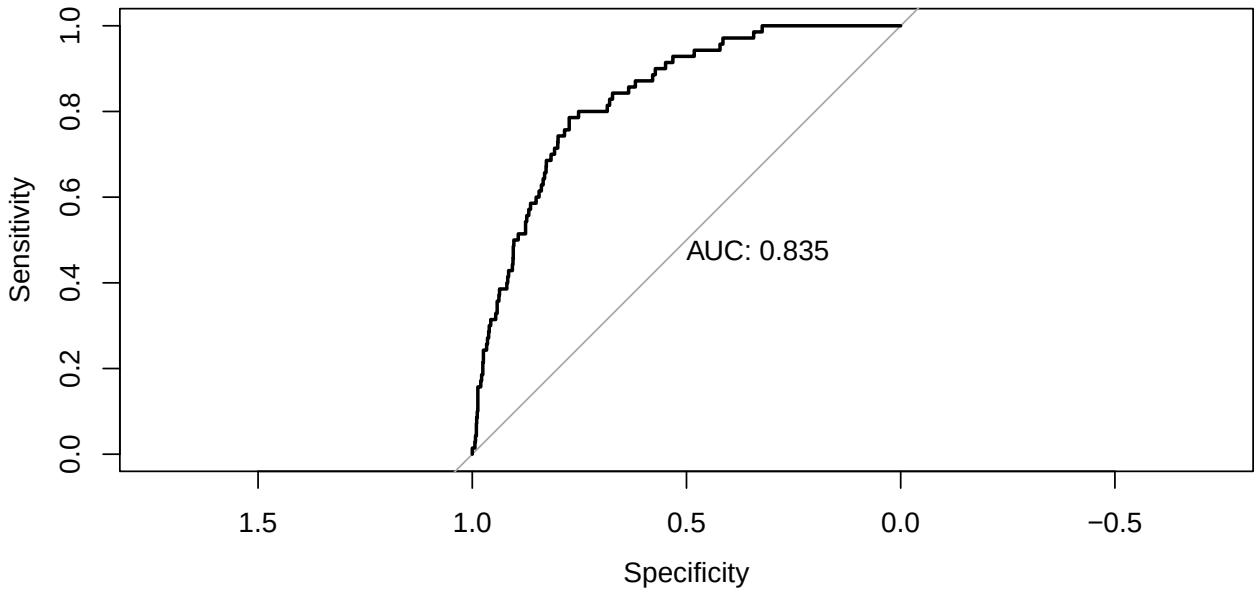
6. Model Validation

The accuracy of the model was reported to be 0.75. A caveat of the model seems to be its precision, which will be discussed in the results. The precision is a ratio of truly positive and positive predicted. The model predicts stroke more often than the patient actually having a stroke. The key metric is the balanced accuracy, which will take into account the imbalanced nature of the data. Oversampling was done only on the training split of the data, and the model was tested using the validation split. The balanced accuracy therefore reflects the nature of the imbalanced classes, and is a better metric to consider than pure accuracy.

Table 6: Confusion Table

term	class	estimate	conf.low	conf.high	p.value
accuracy		0.7462492	0.723681	0.7678701	1
kappa		0.1573507			
mcnemar					0
sensitivity	1	0.8000000			
specificity	1	0.7436774			
pos_pred_value	1	0.1299304			
neg_pred_value	1	0.9872958			
precision	1	0.1299304			
recall	1	0.8000000			
f1	1	0.2235529			
prevalence	1	0.0456621			
detection_rate	1	0.0365297			
detection_prevalence	1	0.2811481			
balanced_accuracy	1	0.7718387			

The area under the curve (AUC) is the probability that a randomly chosen instance is correctly predicted. In other words, the higher the AUC, the better the model performs at distinguishing between patients with stroke and patients without stroke. In this case, the AUC is .835, meaning the degree of separability the model has is quite good. A value of .5 for AUC has the meaning that the predictive ability of model for distinguishing between classes is completely random and has no predictive power. On the y-axis of the plot is the sensitivity and the x-axis is the false positive rate.



7. Discussion and Results

7.1 Results

What factors are used to predict whether or not someone had a stroke?

In order to answer this question, the logistic regression model was used. In order to determine which variables were significant, a forward stepwise procedure was used, which started with the null model and attempted to reach the full model based on the BIC criterion. Based on this procedure, the model with all predictor variables excluding `BMI` were selected. The model fitting results are as below:

Table 7: Stepwise BIC Model

term	estimate	std.error	statistic	p.value
(Intercept)	-3.1075208	0.1852857	-16.7715131	0.0000000
genderMale	-0.1356421	0.0648629	-2.0912118	0.0365091
age	0.0831471	0.0025483	32.6280399	0.0000000
hypertension1	0.5818425	0.0843883	6.8948217	0.0000000
heart_disease1	0.4452974	0.1035059	4.3021445	0.0000169
ever_marriedYes	0.3166174	0.1069969	2.9591282	0.0030851
work_typeGovt_job	-2.1886375	0.2359773	-9.2747806	0.0000000
work_typeNever_worked	-12.7894343	212.9412505	-0.0600609	0.9521072
work_typePrivate	-2.1928300	0.2267485	-9.6707593	0.0000000
work_typeSelf-employed	-2.6262107	0.2450970	-10.7149852	0.0000000
Residence_typeUrban	0.1405616	0.0623535	2.2542697	0.0241792
avg_glucose_level	0.0033041	0.0005845	5.6525007	0.0000000
smoking_statusnever smoked	-0.4700953	0.0731251	-6.4286460	0.0000000
smoking_statussmokes	-0.1666734	0.0879937	-1.8941515	0.0582049

Based on this table, we can see that nearly all the variables are significant in predicting the incidence of stroke. The level `Never_worked` in the predictor `work_type` was deemed not significant. A way to view this is that there is not statistically significant evidence for non-workers to predict stroke. All things equal, being male is associated with a 12.7% reduction in the relative risk of having a stroke. Another variable is age, which translates to every year additional year aged is associated with an 8% increase in stroke. This coefficient makes a lot of sense since it was observed that the trending for stroke incidence was with the older population.

Hypertension is also a significant predictor. The odds associated with patients that have hypertension are 79% more likely to have a stroke compared to those that do not have hypertension. Patients that are married are 37% more likely of having a stroke compared to patients that are not married. Heart disease is also another variable that is expected to be associated to having a stroke. In this analysis, it was found that patients with heart disease are 56.1% more likely to induce a stroke compared to patients that do not have heart disease. This is also expected due to stroke being a malfunction of receiving blood in the brain. Having heart disease is likely to affect basic bodily functions such as this. For average glucose level, every additional unit increase for a patient's average glucose level increases their odds of having a stroke by 3%.

Of the 5 job types, patients that are self employed are associated with a 93% decrease in risk of stroke compared to other job types. It might be hypothesized that those that are self employed at an old age have financial freedom and not as stressed out as their counterparts that are still in the labor force. Patients that live in urban areas instead of rural areas are associated with a 15.1% increase in risk of stroke compared to their rural counterparts. Finally, patients that have never smoked are associated with a 37.5% decrease in relative risk of having a stroke compared to patients that smoked.

7.2 Discussion

From the analysis above, it can be seen that an individual's odds of getting a stroke are impacted by various lifestyle and demographic variables. Having hypertension as well as heart disease increases a persons odds of having a stroke greatly. What can be taken from this is that in a good measure to prevent stroke is to exercise regularly and to care of the heart. There are some risks from factors that can not be avoided, such as age. Growing old increases the risk of stroke by 8% year over year. In order to test the true effect of this, causal inference would need to be performed, however, the construction of a counter factual for the treatment of `age` is not quite feasible. Although it is also outside the scope of this project to investigate the true casual effect that age has on the incidence of stroke, it is a potential topic to do further research in in future works.

The predictive ability of the model is somewhat of a mixed bag. Recall the precision of the model in the confusion matrix (Table 12) was .13. In the medical setting, sacrificing a low precision for a high recall rate is more favorable since clinicians want to avoid missing a patient's treatment in the off chance that the patient is likely to have a stroke. In this case, since the recall is .8, the model performed quite well in this regard. Recall is the true positive rate, and ideally the model would predict the patient having a stroke correctly. Due to the imbalanced data and the nature of the problem at hand, the F1 score, which is a weighted mean of the precision and recall metrics, is not a metric that is of utmost importance. Sensitivity was measured to be .8, which translates to given that the patient had a stroke, what percentage of cases did the model predict correctly. Specificity was measured to be .74, which refers to the percentage of correct non-stroke observations the model predicted.

The balanced accuracy metric being .77 is a result of the oversampling performed in the exploratory data analysis section. Balanced accuracy is the mean of sensitivity and specificity, and is a metric to consider on highly imbalanced data sets. In a sense, it gives the true accuracy of the model, since it takes into account that, in this case, the negative class has a substantial amount of additional observations.

7.3 Future Works

A question of interest that can be pursued outside of this report is to determine which variables the driving factors in predicting stroke. In order to answer this question, a random forest model would be fit and variable importance would be analyzed. The variable importance would be assessed with a different variable that accounts for stroke risk over time, in a sense, a longitudinal study. This would require time series data set in order to track a group of patients over time. Also previously mentioned, a causal framework analysis would also be a possible avenue to explore. The difficulty in this is the construction of the counter factual, although there is literature that has taken this approach (Nguyen et al. 2020).

8. Conclusion

There are a multitude of factors that affect the risk of stroke. An initial point of analysis was to find the variables that were significant in determining a patient's risk to stroke as well as to compared the odds of the patient's depending on which attributes the patients has. From this analysis, it was determined that medical variables such as `heart disease`, `hypertension`, and `work type` are all significant variables that obvious in assessing a patient's incidence to stroke. Other factors such as `age` and `gender` are immutable characteristics that are just associated with the levels in the variables themselves and can not be altered. Lastly, it was also determined that the predictive ability of the model, using balanced accuracy as the main metric, was .77. After all considerations about the quality of data, the predictive power using the ROC curve as well as the training set, is deemed to have good predictive ability.

9. References

Fedesoriano, F. (2021, January 26). Stroke prediction dataset. Kaggle. Retrieved March 17, 2022, from https://www.kaggle.com/fedesoriano/stroke-prediction-dataset

Kelly-Hayes M. Influence of age and health behaviors on stroke risk: lessons from longitudinal studies. J Am Geriatr Soc. 2010;58 Suppl 2(Suppl 2):S325-S328. doi:10.1111/j.1532-5415.2010.02915.x

Kuhn, M. (2019, March 27). The Caret package. 11 Subsampling For Class Imbalances. Retrieved March 17, 2022, from https://topepo.github.io/caret/subsampling-for-class-imbalances.html

Oesch L, Tatlisumak T, Arnold M, Sarikaya H. Obesity paradox in stroke - Myth or reality? A systematic review. PLoS One. 2017;12(3):e0171334. Published 2017 Mar 14. doi:10.1371/journal.pone.0171334

Nguyen TL, Collins GS, Landais P, Le Manach Y. Counterfactual clinical prediction models could help to infer individualized treatment effects in randomized controlled trials-An illustration with the International Stroke Trial. J Clin Epidemiol. 2020 Sep;125:47-56. doi: 10.1016/j.jclinepi.2020.05.022. Epub 2020 May 25. PMID: 32464321.

10. Tables

Table 8: Summary Statistics

	age	avg_glucose_level	bmi
	Min. : 0.08	Min. : 55.12	Min. :10.30
	1st Qu.:25.00	1st Qu.: 77.24	1st Qu.:23.80
	Median :45.00	Median : 91.88	Median :28.10
	Mean :43.23	Mean :106.14	Mean :28.86
	3rd Qu.:61.00	3rd Qu.:114.09	3rd Qu.:32.80
	Max. :82.00	Max. :271.74	Max. :97.60

	gender	hypertension	heart_disease	ever_married
	Female:2994	0:4611	0:4833	No :1756
	Male :2115	1: 498	1: 276	Yes:3353
	Other : 0			

	work_type	Residence_type	smoking_status	stroke
	children : 687	Rural:2513	formerly smoked:1224	0:4860
	Govt_job : 657	Urban:2596	never smoked :2731	1: 249
	Never_worked : 22		smokes :1154	
	Private :2924			
	Self-employed: 819			

Table 9: Full Model

term	estimate	std.error	statistic	p.value
(Intercept)	-4.5727249	0.3018867	-15.1471539	0.0000000
age	0.0831988	0.0025846	32.1902513	0.0000000
avg_glucose_level	0.0032875	0.0006003	5.4761958	0.0000000
bmi	0.0006246	0.0051657	0.1209129	0.9037600
gender_Female	0.1356204	0.0648618	2.0909146	0.0365357
gender_Male				
gender_Other				
hypertension_0	-0.5807218	0.0848850	-6.8412782	0.0000000
hypertension_1				
heart_disease_0	-0.4452842	0.1035014	-4.3022052	0.0000169
heart_disease_1				
ever_married_No	-0.3158076	0.1072171	-2.9454972	0.0032244
ever_married_Yes				
work_type_children	2.6331663	0.2518091	10.4569931	0.0000000
work_type_Govt_job	0.4378390	0.1036158	4.2256005	0.0000238
work_type_Never_worked	-10.1578565	212.7624805	-0.0477427	0.9619213
work_type_Private	0.4332499	0.0815998	5.3094471	0.0000001
‘work_type_Self-employed‘				
Residence_type_Rural	-0.1403548	0.0623750	-2.2501780	0.0244376
Residence_type_Urban				
‘smoking_status_formerly smoked‘	0.1659508	0.0881932	1.8816735	0.0598804
‘smoking_status_never smoked‘	-0.3036340	0.0827793	-3.6679922	0.0002445
smoking_status_smokes				

Table 10: Reduced Model

term	estimate	std.error	statistic	p.value
(Intercept)	-4.5517282	0.2468228	-18.4412796	0.0000000
age	0.0831471	0.0025483	32.6280399	0.0000000
avg_glucose_level	0.0033041	0.0005845	5.6525007	0.0000000
gender_Female	0.1356421	0.0648629	2.0912118	0.0365091
gender_Male				
gender_Other				
hypertension_0	-0.5818425	0.0843883	-6.8948217	0.0000000
hypertension_1				
heart_disease_0	-0.4452974	0.1035059	-4.3021445	0.0000169
heart_disease_1				
ever_married_No	-0.3166174	0.1069969	-2.9591282	0.0030851
ever_married_Yes				
work_type_children	2.6262107	0.2450970	10.7149852	0.0000000
work_type_Govt_job	0.4375732	0.1035987	4.2237317	0.0000240
work_type_Never_worked	-10.1632236	212.9412538	-0.0477278	0.9619332
work_type_Private	0.4333807	0.0815949	5.3113697	0.0000001
‘work_type_Self-employed‘				
Residence_type_Rural	-0.1405616	0.0623535	-2.2542697	0.0241792
Residence_type_Urban				
‘smoking_status_formerly smoked‘	0.1666734	0.0879937	1.8941515	0.0582049
‘smoking_status_never smoked‘	-0.3034219	0.0827628	-3.6661618	0.0002462
smoking_status_smokes				

Table 11: Stepwise BIC Model

term	estimate	std.error	statistic	p.value
(Intercept)	-3.1075208	0.1852857	-16.7715131	0.0000000
genderMale	-0.1356421	0.0648629	-2.0912118	0.0365091
age	0.0831471	0.0025483	32.6280399	0.0000000
hypertension1	0.5818425	0.0843883	6.8948217	0.0000000
heart_disease1	0.4452974	0.1035059	4.3021445	0.0000169
ever_marriedYes	0.3166174	0.1069969	2.9591282	0.0030851
work_typeGovt_job	-2.1886375	0.2359773	-9.2747806	0.0000000
work_typeNever_worked	-12.7894343	212.9412505	-0.0600609	0.9521072
work_typePrivate	-2.1928300	0.2267485	-9.6707593	0.0000000
work_typeSelf-employed	-2.6262107	0.2450970	-10.7149852	0.0000000
Residence_typeUrban	0.1405616	0.0623535	2.2542697	0.0241792
avg_glucose_level	0.0033041	0.0005845	5.6525007	0.0000000
smoking_statusnever smoked	-0.4700953	0.0731251	-6.4286460	0.0000000
smoking_statussmokes	-0.1666734	0.0879937	-1.8941515	0.0582049

Table 12: Confusion Table

term	class	estimate	conf.low	conf.high	p.value
accuracy		0.7462492	0.723681	0.7678701	1
kappa		0.1573507			
mcnemar					0
sensitivity	1	0.8000000			
specificity	1	0.7436774			
pos_pred_value	1	0.1299304			
neg_pred_value	1	0.9872958			
precision	1	0.1299304			
recall	1	0.8000000			
f1	1	0.2235529			
prevalence	1	0.0456621			
detection_rate	1	0.0365297			
detection_prevalence	1	0.2811481			
balanced_accuracy	1	0.7718387			

11. Appendix

```
library(here)
library(naniar)
library(wesanderson)
library(scales)
library(vtable)
library(ggplot2)
library(ggpubr)
library(stargazer)
library(skimr)
library(kableExtra)
library(knitr)
library(themis)
library(gridExtra)
library(corrplot)
library(caret)
library(car)
library(tidyverse)
library(randomForest)
library(ROSE)
library(pROC)
library(magrittr)
library(jtools)
library(finalfit)
library(ggthemes)
library(imbalance)
#Custom theme
theme_set(theme_tufte())
my_theme <- theme(plot.title = element_text(hjust = 0.5, face = 2, size = 18),
        plot.subtitle = element_text(hjust = 0.5, size = 13),
        axis.title = element_text(face = 1, size = 15),
        axis.text = element_text(size = 13))
```

```r
# Custom palette
my_palette <- c('#FAD510', '#CB2314', '#273046', '#354823', '#1E1E1E',
                "#A42820", "#5F5647", "#9B110E", "#3F5151", "#4E2A1E")

    options(knitr.kable.NA = '')
data <- read.csv(here('healthcare-dataset-stroke-data.csv'))
glimpse(data)
cols <- c('gender', 'ever_married', 'work_type', 'Residence_type','smoking_status', 'stroke', 'hypertens
data %<>% mutate_at(cols, factor) %>% mutate(bmi = as.numeric(bmi))
#removing gender observation with other since it was only 1 observation
#replaced inconsistent values in BMI and smoking status with unknown
data %<>% replace_with_na(data = data, replace = list(bmi = c('N/A'), smoking_status = c('Unknown'))) %
data1 %<>% replace_with_na(data = data, replace = list(bmi = c('N/A'), smoking_status = c('Unknown'))) %
mp <- data %>% missing_plot()

mp
bmi_preimute <- data1 %>% ggplot() +
  geom_histogram(bins = 30, color = 'black', fill = my_palette[6], aes(x = bmi)) +
  labs(xaxis = 'BMI', title = 'BMI Distribution: Pre-Imp') +
  xlab('BMI') +
  ylab('Count')

#imputing median
data %<>% mutate(bmi = ifelse(is.na(bmi), median(bmi, na.rm = T),
                              bmi))
bmi_plt <- data %>% ggplot() + geom_histogram(bins = 30, color = 'black', fill = my_palette[6], aes(x =
  labs(xaxis = 'BMI', title = 'BMI Distribution: Median Imp.') +
  xlab('BMI') +
  ylab('Count')

grid.arrange(bmi_preimute, bmi_plt, nrow = 1)

m <- data %>% count(smoking_status) %>% rename(`Smoking Status` = smoking_status,
                                               Count = n) %>%
  ggplot(aes(x = `Smoking Status`, y = Count))+
  geom_col(aes(Count, reorder(`Smoking Status`, Count))) +
  labs(title = 'Pre-Imputation')

#smoking: imputing based on previous value since the missing plot show no pattern of missingness
data$smoking_status <- as.character(data$smoking_status)
data <- data %>% fill(smoking_status, .direction = 'downup')



d <- data %>% count(smoking_status) %>% rename(`Smoking Status` = smoking_status,
                                               Count = n) %>%
  ggplot(aes(x = `Smoking Status`, y = Count))+
  geom_col(aes(Count, reorder(`Smoking Status`, Count))) + labs(title = 'Imputed')

stroke_clean <- data %>% mutate(smoking_status = as.factor(smoking_status))


grid.arrange(m,d, nrow = 1)
```

```
kable(caption = 'Summary Statistics', summary(data %>% select(gender, age, hypertension, heart_disease))
kable(summary(data %>% select(ever_married, work_type, Residence_type))) %>% kable_styling(position = '

kable(summary(data %>% select(avg_glucose_level, bmi, smoking_status, stroke))) %>% kable_styling(positi



#stroke is heavily imbalanced
#pct <- stroke_clean %>% count(stroke) %>% mutate(pct = prop.table(n))
b <- stroke_clean %>%  ggplot(aes(x = stroke)) +
  scale_x_discrete(labels = c('0' = 'Never had a stroke', '1'= 'Had a stroke'))+
  geom_bar(fill = my_palette[10]) +
  labs(x = '',
       y = 'Counts',
       title = 'Stroke Distribution: Imbalanced')



set.seed(225231)
dt <- sort(sample(nrow(stroke_clean), nrow(stroke_clean) * .7))
train_stroke <- stroke_clean[dt,] %>% select(-id)
test_stroke <- stroke_clean[-dt,] %>% select(-id)



train_oversampled <- ovun.sample(stroke ~. , data = train_stroke,
                                 method = 'over',
                                 seed = 252323)
train_oversampled <- train_oversampled$data


stroke_postsampling <- train_oversampled %>% ggplot(aes(x = stroke)) +
   scale_x_discrete(labels = c('0' = 'Never had a stroke', '1'= 'Had a stroke'))+
  geom_bar(fill = my_palette[3]) +
  labs(x = '',
       y = 'Counts',
       title = 'Stroke Distribution: Upsampled')

grid.arrange(b, stroke_postsampling, nrow = 1)

stargazer(train_oversampled, header = F,type = 'latex', title = 'Numeric Summary Statistics')
df <- data.frame(Variables = c('gender', 'age', 'hypertension', 'Heart_disease', 'ever_married','work_ty
                               'bmi', 'smoking_status', 'stroke'),
                 Description = c('Male or Female', 'Age of respondent', 'Whether or not the respondent
                                 'Whether or not the respondent is married','What type of job the respo
kable(df, caption ='Predictor Variables and Response Variable, `stroke`') %>% kable_styling(position =
age_density <- ggplot() +
  geom_density(data=stroke_clean, aes(x=age , group=as.factor(stroke), fill=as.factor(stroke)), size=1,a
  ylab("Density")+ labs(fill=' ',x="Age")+
  scale_fill_manual(values = my_palette[2:3], labels = c('No stroke','Stroke')) +
  labs(title = 'Age vs Stroke Density Plot',
       caption = 'Fig 1: Density of Age on Stroke') +
  theme(plot.caption = element_text(hjust = .5),
```

```
        plot.title = element_text(hjust = .5))

age_density
bmi_density <- stroke_clean %>%  ggplot() +
  geom_density( aes(x=bmi , group=as.factor(stroke), fill=as.factor(stroke)), size=1,alpha=0.5, adjust=1
  ylab("Density")+ labs(fill=' ',x="BMI")+
  scale_fill_manual(values = c(my_palette[1], my_palette[5]), labels = c('No stroke','Stroke')) +
  labs(title = 'BMI vs Stroke Density Plot',
       caption = 'Fig. 2: Density of BMI on Stroke') +
    theme(plot.caption = element_text(hjust = .5),
        plot.title = element_text(hjust =.5))


bmi_density
glucose_density <- stroke_clean %>%  ggplot() +
  geom_density(aes(x=avg_glucose_level , group=as.factor(stroke), fill=as.factor(stroke)), size=1,alpha=
  ylab("Density")+ labs(fill=' ', x= "Average Glucose Levels")+
  scale_fill_manual(values = c(my_palette[4], my_palette[5]), labels = c('No stroke','Stroke')) +
  labs(title = 'Average Glucose Levels vs Stroke Density Plot',
       caption ='Fig. 3: Density of Glucose on Stroke')+
  theme(plot.caption = element_text(hjust = .5),
        plot.title = element_text(hjust = .5))
glucose_density

age_bmi <- stroke_clean %>% arrange(stroke) %>%
  ggplot(aes(x = age, y = bmi, col = as.factor(stroke))) +
  geom_jitter(alpha = 2, aes(col = stroke))+
  scale_color_manual(labels = c('No stroke', 'Had a stroke'), values = c(my_palette[1], my_palette[5]))+
  labs(caption = 'Fig. 4: Relationship between Stroke Incidence and Age/BMI',
       title = 'Age, BMI, Stroke', col = '') +
  xlab('Age')+
  ylab('BMI') +
  theme(plot.title = element_text(hjust = .5),
        plot.caption = element_text(hjust = .5))


age_bmi

age_glucose_plt <- stroke_clean %>% arrange(stroke) %>%
  ggplot(aes(x = age, y = avg_glucose_level, col = as.factor(stroke))) +
  geom_jitter(alpha = 2, aes(col = stroke))+
  scale_color_manual(labels = c('No stroke', 'Had a stroke'), values = c(my_palette[1], my_palette[9]))+
  labs(title = 'Age, Average Glucose Level, Stroke', col = '',
       caption = 'Fig. 5: Relationship between age/glucose levels and stroke') +
  ylab('Average Glucose Level')+
  xlab('Age')+
  theme(plot.title = element_text(hjust = .5),
        plot.caption = element_text(hjust = .5))

age_glucose_plt

corrplot(cor(stroke_clean %>% select_if(is.numeric) %>% select(-id)),
         type = 'full',
```

```
         order = 'original',
         method = 'color',
         tl.col = "black",
         tl.srt = 45,
         addCoef.col = "black",
         )


ss <- stroke_clean %>% ggplot() +
  geom_bar(aes(x = gender, fill = stroke), position = 'fill') +
  scale_fill_manual(values = c(my_palette[4], my_palette[8]),
labels = c('No stroke', 'Had a stroke')) +
  labs(title = 'Distribution of Stroke Between Sexes',
       caption = 'Fig. 6: Relationship between Sexes/Stroke')

ss

sm <- stroke_clean %>% ggplot() +
  geom_bar(aes(x = smoking_status, fill = stroke), position = 'fill') +
  scale_fill_manual(values = c(my_palette[5], my_palette[8]),
labels = c('No stroke', 'Had a stroke')) +
  labs(title = 'Distribution of Stroke Between Smokers',
       caption = 'Fig. 7: Relationship between Smokers/Stroke')

sm

hh <- stroke_clean %>% ggplot() +
  geom_bar(aes(x = hypertension, fill = stroke), position = 'fill') +
  scale_fill_manual(values = c(my_palette[2], my_palette[8]),
labels = c('No stroke', 'Had a stroke')) +
  labs(title = 'Distribution of Stroke Between Hypertension',
       caption = 'Fig. 8: Relationship between Hypertension/Stroke')

ht <- stroke_clean %>% ggplot() +
  geom_bar(aes(x = heart_disease, fill = stroke), position = 'fill') +
  scale_fill_manual(values = c(my_palette[3], my_palette[8]),
labels = c('No stroke', 'Had a stroke')) +
  labs(title = 'Distribution of Stroke Between Heart Disease',
       caption = 'Fig. 9: Relationship between Heart Disease/Stroke')

grid.arrange(hh,ht, nrow = 1)

ab <- stroke_clean %>% ggplot() +
  geom_bar(aes(x = work_type, fill = stroke), position = 'fill') +
  scale_fill_manual(values = c(my_palette[10], my_palette[8]),
labels = c('No stroke', 'Had a stroke')) +
  labs(title = 'Distribution of Stroke Between Work Type',
       caption = 'Fig. 10: Relationship between WorkType/Stroke')

ac <- stroke_clean %>% ggplot() +
  geom_bar(aes(x = Residence_type, fill = stroke), position = 'fill') +
  scale_fill_manual(values = c(my_palette[9], my_palette[8]),
labels = c('No stroke', 'Had a stroke')) +
```

```
  labs(title = 'Distribution of Stroke Between Residence Type',
       caption = 'Fig. 11: Relationship between Residence/Stroke')

grid.arrange(ab,ac, nrow = 1)
library(fastDummies)
train_oversampled_d <- train_oversampled
train_oversampled_d <- dummy_cols(train_oversampled_d, select_columns = c('gender', 'hypertension', 'hea
                                                                          'smoking_status'),
                                  remove_selected_columns = T)
test_stroke_d <- test_stroke
test_stroke_d <- dummy_cols(test_stroke_d, select_columns = c('gender', 'hypertension', 'heart_disease'
                                                                          'smoking_status'),
                                  remove_selected_columns = T)


full_fit <- glm(stroke ~ ., data = train_oversampled_d, family = 'binomial'(link = logit))
reduced_fit <- glm(stroke  ~ . , data = train_oversampled_d %>% select(-bmi), family = 'binomial'(link =



kable(anova(reduced_fit, full_fit, test = 'Chisq'), caption = 'Pearson Chi-Square Test') %>% kable_styli
step <- stats::step(glm(stroke~., data = train_oversampled, family ='binomial'(link = logit)),
             criterion = 'BIC')
kable(step %>% tidy(), caption ='Stepwise Procedure, Final Model') %>%  kable_styling(position = 'cente
fit <- glm(stroke  ~ . , data = train_oversampled_d %>% select(-bmi), family = 'binomial'(link = logit)
kable(sum(residuals(fit, type = "deviance")^2)/fit$df.residual)
probs <- fit %>% predict(test_stroke_d, type = 'response')
predicted_classes <- ifelse(probs > .5, '1', '0')
kable(caption = 'Confusion Table', caret::confusionMatrix(data = as.factor(predicted_classes), referenc
  kable_styling(position = 'center', latex_options = 'HOLD_position')

test_roc <- roc(test_stroke_d$stroke ~ probs, plot = T, print.auc = T)
kable(step %>% tidy(),
      caption = 'Stepwise BIC Model') %>% kable_styling(position = 'center', latex_options = 'HOLD_posi
kable(caption = 'Summary Statistics', summary(stroke_clean %>% select_if(is.numeric) %>% select(-id))) %
kable(summary(stroke_clean %>% select(gender, hypertension, heart_disease, ever_married
                                      ))) %>% kable_styling(position = 'center', latex_options = 'HOLD_p
kable(summary(stroke_clean %>% select(work_type, Residence_type, smoking_status, stroke))) %>% kable_sty
kable(full_fit %>% tidy(), caption = 'Full Model') %>% kable_styling(position = 'center', latex_options
kable(reduced_fit  %>%  tidy(), caption = 'Reduced Model')%>% kable_styling(position = 'center', latex_
kable(step %>% tidy(),
      caption = 'Stepwise BIC Model') %>% kable_styling(position = 'center', latex_options = 'HOLD_posi
kable(caption = 'Confusion Table', caret::confusionMatrix(data = as.factor(predicted_classes), referenc
  kable_styling(position = 'center', latex_options = 'HOLD_position')
```