Stroke Classification Prediction Analysis

Ngoc Nguyen

May 13th, 2021

Abstract

According to WHO, stroke is second leading to death globally. Someone in the United States has a stroke every 40 minutes and dies of a stroke every 4 minutes. The document includes some of the data selection, exploratory data analysis, and different data mining methods including logistic regression, decision tree, and random forest to the dataset to find the best prediction model.

Introduction and Data information

As stroke is one of the highest causes of death globally, this project produces the selected prediction model on whether a patient is likely to get a stroke based on the health parameters such as the body mass index, hypertension, etc,... Results can be used by hospital staff to predict potential stroke patients to give them the needed medical services and avoid sending them home. This dataset is originally from Kaggle. The dataset has 12 variables with 5110 observations.

Id: a unique identifier.

Gender: "Male", "Female" or "Other".

Age: age of the patient.

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

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

Ever_married: "No" or "Yes".

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

Residence_type: "Rural" or "Urban".

Avg_glucose_level: average glucose level in the blood.

Bmi: body mass index.

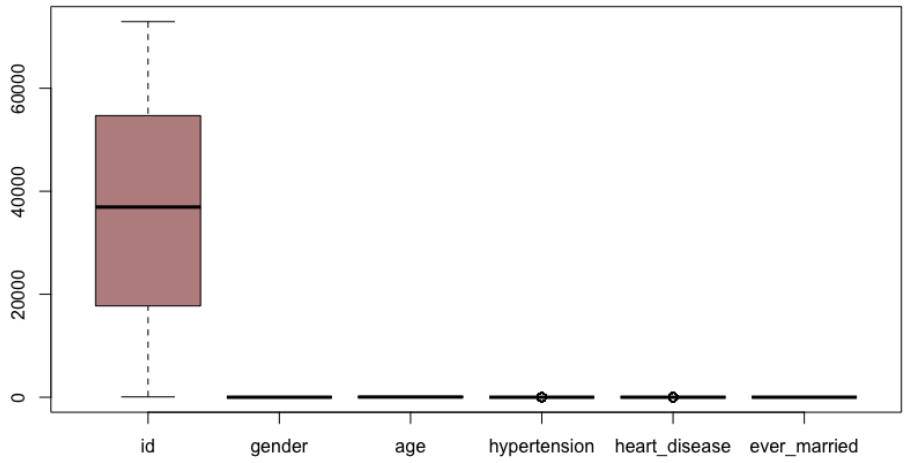
Smoking_status: "formerly smoked", "never smoked", "smokes" or "Unknown"*.

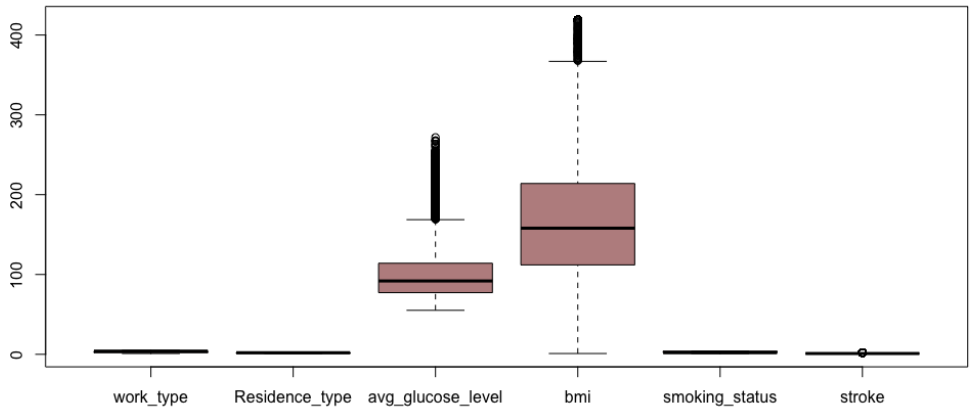
Stroke: 1 if the patient had a stroke or 0 if not.

Exploratory Data Analysis

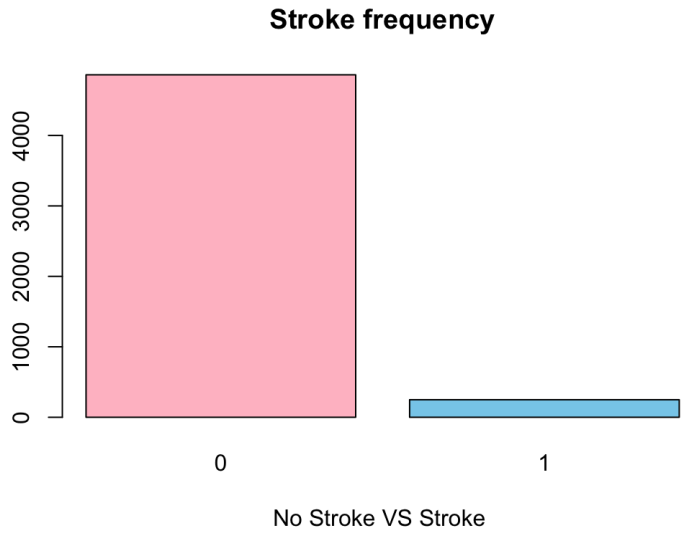
After loading the dataset, it seems that there are a lot of categorical variables, which are “Gender”, “ever_married”, “work_type”, “residence_type”, “smoking_status”, and “BMI”. Moreover, there are also some binary variables such as “hypertension”, “heart_disease”, and “stroke”. Therefore, it is necessary to convert them into factors. Therefore, the first step is to convert all categorical variables into factors, and the “BMI” variable into a numeric variable. From the summary table below, “avg_glucose_level” and “BMI” have a large difference between the mean and the max and min. This could indicate that these variables have many outliers.

```
      id          gender          age        hypertension heart_disease ever_married
Min.   :   67   Female:2994   Min.   : 0.08   0:4612       0:4834        No :1757
1st Qu.:17741   Male  :2115   1st Qu.:25.00   1: 498       1: 276        Yes:3353
Median :36932   Other :   1   Median :45.00
Mean   :36518                 Mean   :43.23
3rd Qu.:54682                 3rd Qu.:61.00
Max.   :72940                 Max.   :82.00
       work_type    Residence_type avg_glucose_level      bmi
children     : 687   Rural:2514     Min.   : 55.12    Min.   :  1.0
Govt_job     : 657   Urban:2596     1st Qu.: 77.25    1st Qu.:112.0
Never_worked :  22                  Median : 91.89    Median :158.0
Private      :2925                  Mean   :106.15    Mean   :172.2
Self-employed: 819                  3rd Qu.:114.09    3rd Qu.:214.0
                                    Max.   :271.74    Max.   :419.0
       smoking_status stroke
formerly smoked: 885   0:4861
never smoked   :1892   1: 249
smokes         : 789
Unknown        :1544
```





This boxplot above shows that some variables have outliers that are “hypertension”, “heart_disease”, “avg_glucose_level”, “BMI”, and “stroke”. Especially, “avg_glucose_level” and “BMI” have the most outliers.



From this plot, we can see that only a small number of people have a stroke. This is highly unbalanced data distribution and may result in giving us the accuracy of this model at a very high percent of predicting 0, which is healthy patients. After that, it is necessary to find the missing data in each variable. Furthermore, “BMI” has about 201 observations that are missing. This is also another factor we need to put into consideration in finding insignificant variables. Thus, we will remove the “BMI” variable since it can affect the whole dataset.

After splitting and balancing the dataset, the next step is logistic regression. Because the “BMI” variable has many missing values, and its outliers, for the first logistic regression, I will exclude it and will compare it with the second logistic regression which includes the “BMI” variable, and to see if there are any changes in the regression. Below is the first logistic regression without the “BMI” variable.

```
Call:
glm(formula = stroke ~ . - bmi, family = binomial, data = stroke_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2359  -0.3260  -0.1657  -0.0910   3.4697

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                -6.545e+00  7.754e-01  -8.441  < 2e-16 ***
id                          4.368e-06  3.560e-06   1.227  0.21992
genderMale                 -3.293e-02  1.552e-01  -0.212  0.83202
genderOther                -1.069e+01  1.455e+03  -0.007  0.99414
age                         7.613e-02  6.240e-03  12.200  < 2e-16 ***
hypertension1               4.519e-01  1.790e-01   2.525  0.01157 *
heart_disease1              2.802e-01  2.124e-01   1.320  0.18698
ever_marriedYes            -2.661e-01  2.478e-01  -1.074  0.28286
work_typeGovt_job          -1.147e+00  8.422e-01  -1.362  0.17317
work_typeNever_worked      -1.062e+01  4.018e+02  -0.026  0.97891
work_typePrivate           -1.065e+00  8.232e-01  -1.293  0.19587
work_typeSelf-employed     -1.403e+00  8.481e-01  -1.655  0.09794 .
Residence_typeUrban         4.857e-02  1.515e-01   0.320  0.74860
avg_glucose_level           3.539e-03  1.293e-03   2.736  0.00621 **
smoking_statusnever smoked -1.883e-01  1.938e-01  -0.972  0.33111
smoking_statussmokes        2.171e-01  2.295e-01   0.946  0.34420
smoking_statusUnknown      -1.747e-01  2.351e-01  -0.743  0.45761
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1655.9  on 4087  degrees of freedom
Residual deviance: 1307.2  on 4071  degrees of freedom
AIC: 1341.2

Number of Fisher Scoring iterations: 14
```

Above is a summary of logistic regression with all the independent variables except the “BMI” variable of the stroke_train 80%. The logistic regression is logit(y) = -6.545+ 0.000004368X_1 - 0.03293X_2 + ... +0.2171X_{15} - 0.1747X_{16}. There are a lot of variables with a large p-value. The only three variables that have small p-values are “Age”, “hypertension1”, and “avg_glucose_level”. The AIC of this model is 1341.2 and BIC is 1448.592.

For the logistic regression with all the independent variables, it returns 222 rows which is a very large logistic regression. The AIC for this model is 1841.4, and the BIC is 4481.455. This is a very large number compared to the logistic regression without the “BMI” variable above. Therefore, the model without the “BMI” is the optimal model.

Variable selection

Variable selection is the next step for this EDA. As running backward selection using BIC. Surprisingly, the best model includes two variables which are “Age” and “avg_glucose_level”, with the AIC is at 1334.2 and BIC is 1353.184.

```
Call:
glm(formula = stroke ~ age + avg_glucose_level, family = binomial,
    data = stroke_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9458  -0.3327  -0.1806  -0.0821   3.7842

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       -7.548130   0.384712 -19.620  < 2e-16 ***
age                0.072697   0.005440  13.363  < 2e-16 ***
avg_glucose_level  0.004164   0.001248   3.336 0.000849 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1655.9  on 4087  degrees of freedom
Residual deviance: 1328.2  on 4085  degrees of freedom
AIC: 1334.2

Number of Fisher Scoring iterations: 7
```

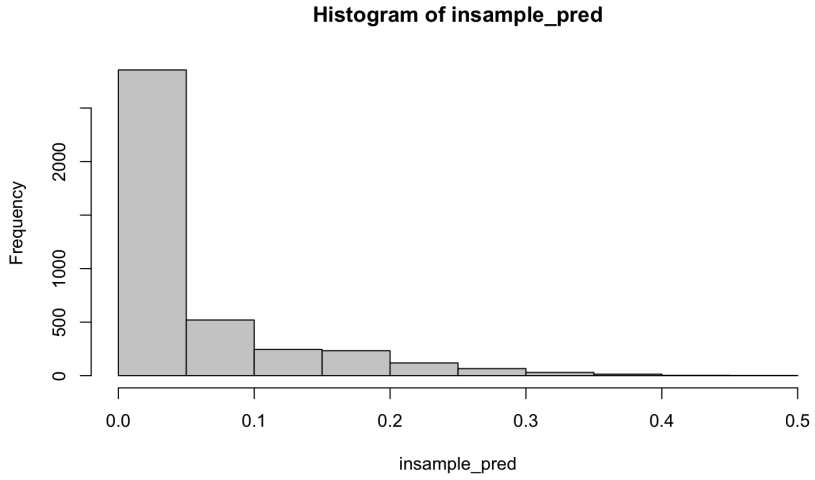
Furthermore, with the backward selection method, there are three variables which are “Age”, “hypertension1, “heart_disease1”, and “avg_glucose_level”. The AIC for this method is 1329 and BIC is 1360.602.

```
Call:
glm(formula = stroke ~ age + hypertension + heart_disease + avg_glucose_level,
    family = binomial, data = stroke_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0885  -0.3295  -0.1784  -0.0835   3.7549

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       -7.382350   0.389245 -18.966  < 2e-16 ***
age                0.069275   0.005579  12.417  < 2e-16 ***
hypertension1      0.457505   0.176610   2.590  0.00958 **
heart_disease1     0.341837   0.208076   1.643  0.10041
avg_glucose_level  0.003440   0.001276   2.697  0.00700 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1655.9  on 4087  degrees of freedom
Residual deviance: 1319.0  on 4083  degrees of freedom
AIC: 1329

Number of Fisher Scoring iterations: 7
```

Choosing the best model using p-value and variable selection is identical because they give the same result. By choosing AIC as the selective criteria, the AIC in the regression with backward is better because it is smaller. Therefore, it will be the best model for this dataset.
Our new logistic regression is logit(y) = $-7.382350 + 0.069275X_1 + 0.457505X_2 + 0.341837X_3 + 0.003440X_4$

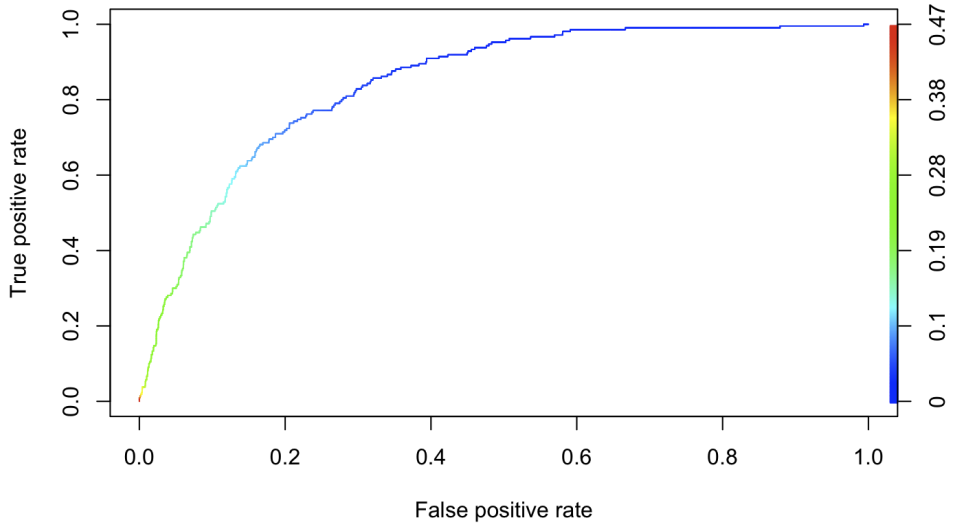
In-sample Performance

The next step is to apply in-sample prediction. Below is a histogram of in-sample prediction with the response of whether the patient is having a stroke or not. the data is very skewed towards 0.0. Therefore, for the misclassification rate, the chosen cut-off probability is 0.2 instead of the default cut-off probability of 0.5.



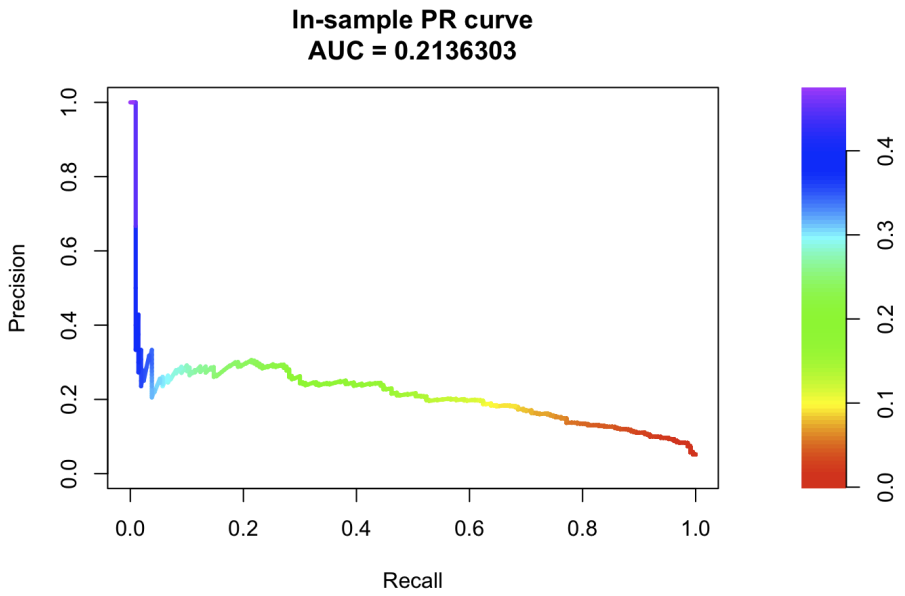
Cut-off probability: 0.2	**Predicted**	
Truth	0	1
0	3704	**174**
1	**150**	60

With a 0.2 cut-off probability, there will be 174 false positives which means 174 patients are predicted with a stroke when they don't have it. Also, there will be 150 false negatives, meaning that 150 patients will be predicted not to have a stroke when they do.
The misclassification rate for this is 0.079.



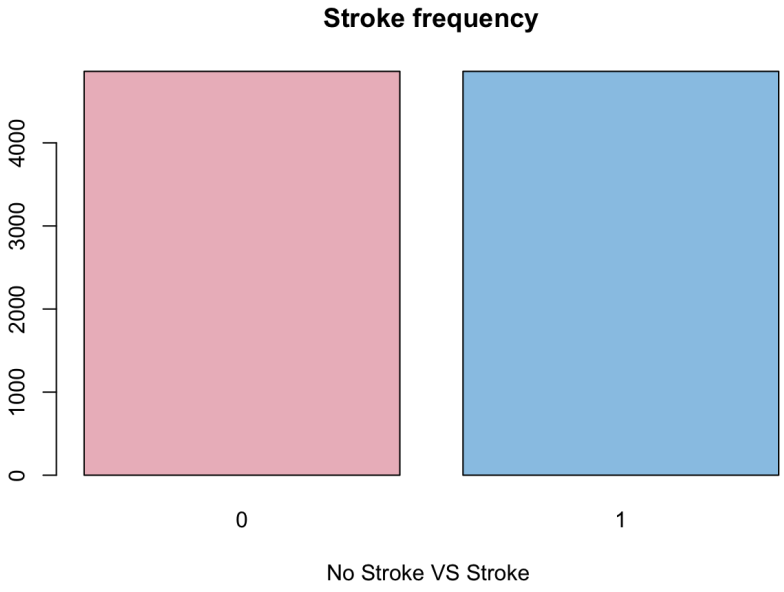
This is also the ROC curve drawn using the best model. The ROC curve helps us to get an overall measure of goodness of classification, showing us the true positive fraction and false-positive fraction. The in-sample AUC is 0.843. This is considered acceptable since it is higher than 0.7, an acceptable discriminatory power.

However, when looking at this output, 150 out of 210 stroke patients will pass the stroke detection. This is not very good since 71% of patients will be sent home even though they are more likely to have a stroke but the model can't catch it. Additionally, the ROC Curve and AUC are still very high for this imbalance dataset, which indicates they are not a very good choice. In this case, the PR Curve is more accurate as it works better with unbalanced data. The AUC of the in-sample PR Curve is 0.214, which indicates the model is very bad for prediction. Therefore, resampling the dataset is appropriate to improve the AUC and the prediction.



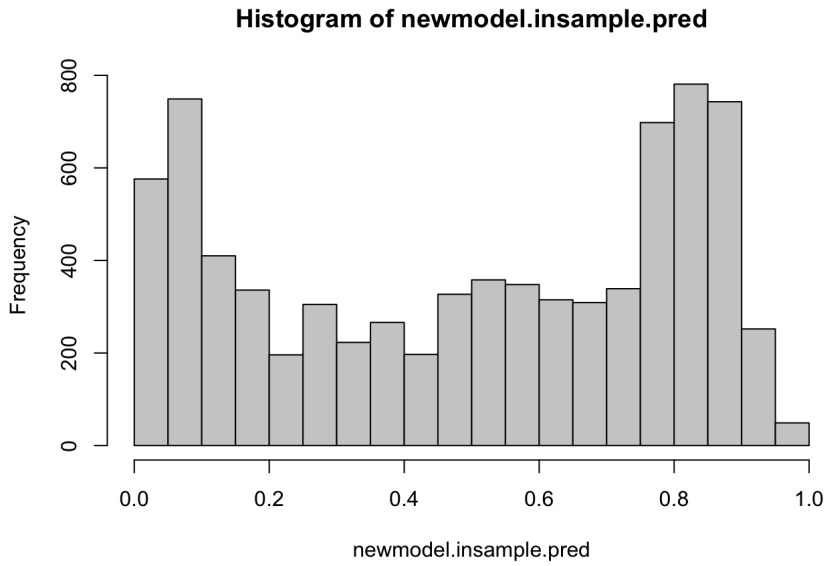
Resampling Data

The data is balanced using the ROSE method by resampling the whole dataset. The bar plot below shows the result after balancing the stroke variables. It now has 4861 observations in each value.

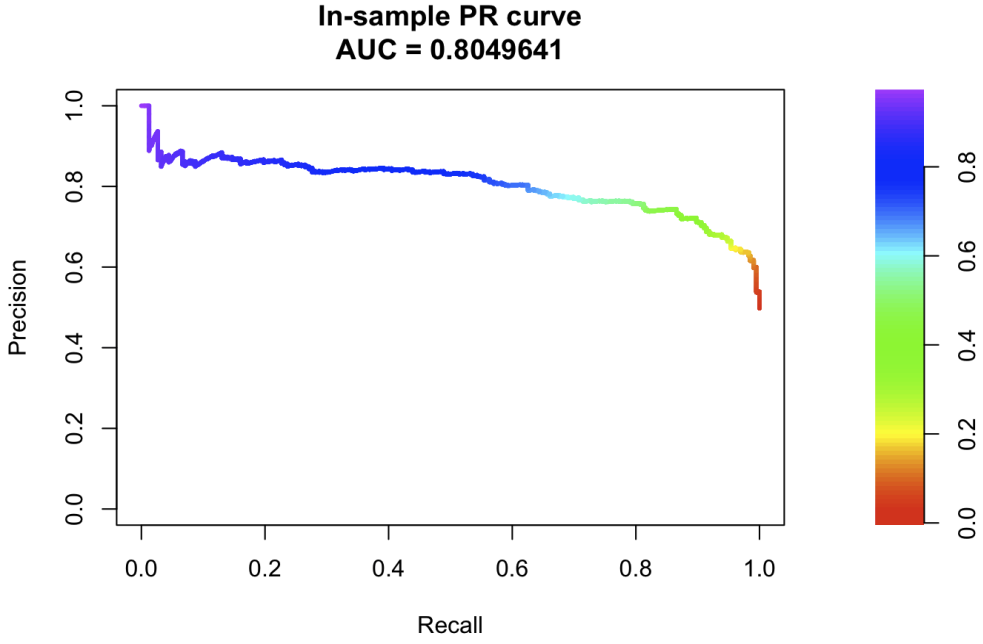


After that, the logistic regression model has been performed again on this new balanced data, find the best model using backward selection, and use the best model to run PR Curve and find

the AUC. This is the histogram with new stroke training data. The data is now balanced and distributed evenly.



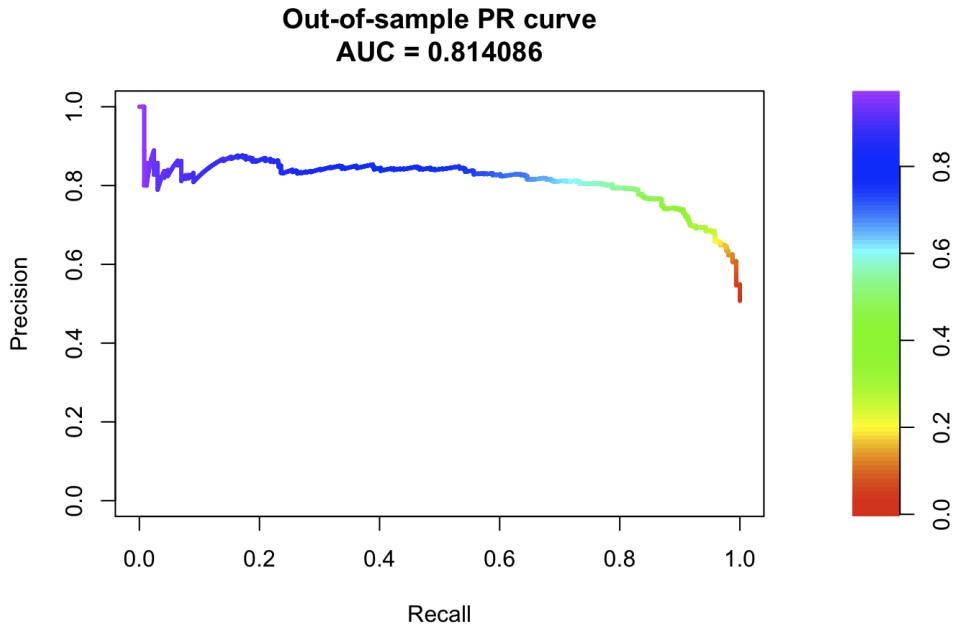
The AUC with balanced train data has a higher AUC than the original one, which rises from 0.214 to 0.805.



Cut-off probability: 0.2	**Predicted**	
Truth	0	1
0	1892	**2012**
1	**179**	3694

Out-of-sample Performance

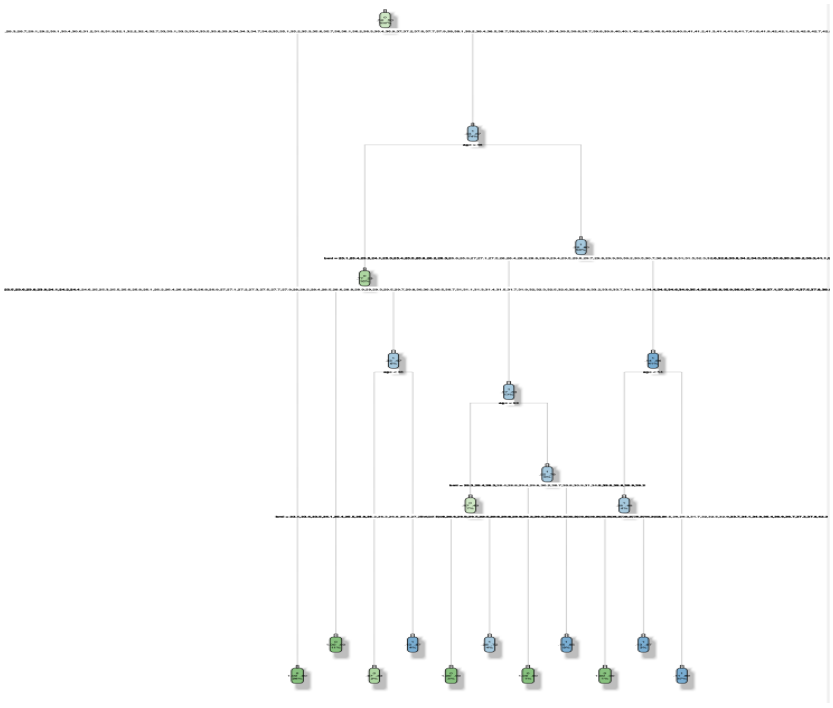
Below is the PR Curve and AUC is 0.8141 of the 20% unseen data, which is higher than an acceptable discriminatory power of 0.7.



Cut-off probability: 0.2	**Predicted**	
Truth	0	1
0	472	**485**
1	**41**	947

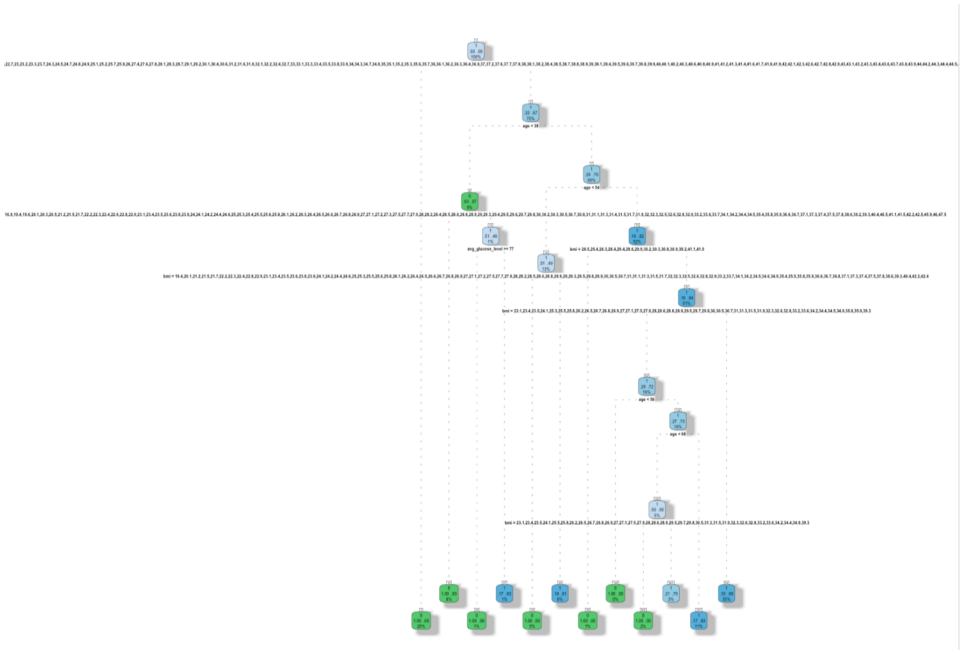
After resampling the data, both the false negatives in In-sample and Out-of-sample Performance have reduced in cases significantly. This can show that the model is trained to predict correctly the stroke patients in the hospital.

Decision Tree



This is the classification tree for the stroke model using the 80% stroke training data with symmetric cost. The tree is very large and difficult to see clearly, this is because the “BMI” variable has a “very large” range in data. But, interestingly, “BMI” and “age” are the most influenced factors that we can use to predict whether a patient would be most likely to have a stroke.

In this particular dataset, false negatives which indicate the model is failing to detect stroke patients and accidentally send them home are very important. Hence, setting the cost of false negatives more costly than false positives to reduce the chance of sending stroke patients home as fewer as possible. Below is the classification tree with asymmetric cost (5 vs. 1). It is still fairly large because of the large range of “BMI”. However, there are some new variables here which are “BMI”, “avg_glucose_level”, and “age”. These three variables are very essential to detect stroke patients.



In-sample Performance

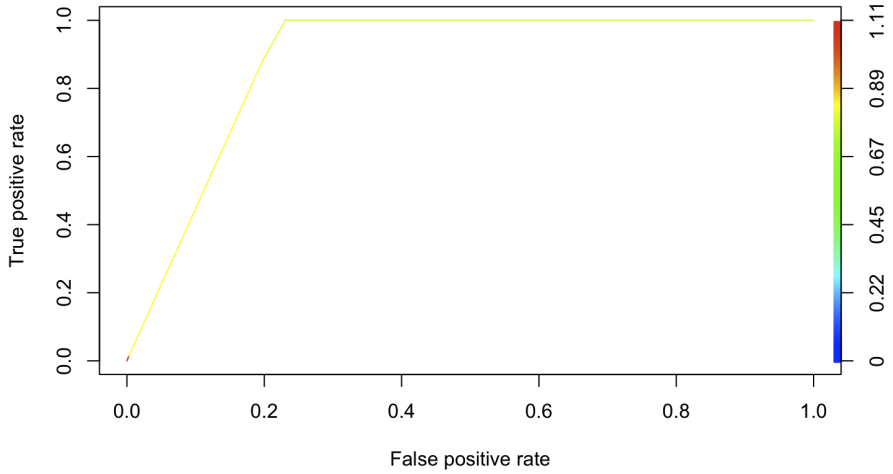
This is the misclassification table of both symmetric and asymmetric costs for the training data.

Symmetric cost	**Predicted**	
Truth	0	1
0	3201	703
1	33	3840

Asymmetric cost	**Predicted**	
Truth	0	1
0	3170	734
1	0	3873

The table with asymmetric cost gives the better model because the false negative is 0 in this case. This indicates that the model is very successful in detecting stroke patients that it doesn't send

any stroke patients home by accident. Therefore, the classification tree with the asymmetric cost is chosen as the best model and apply it to the 20% testing data.



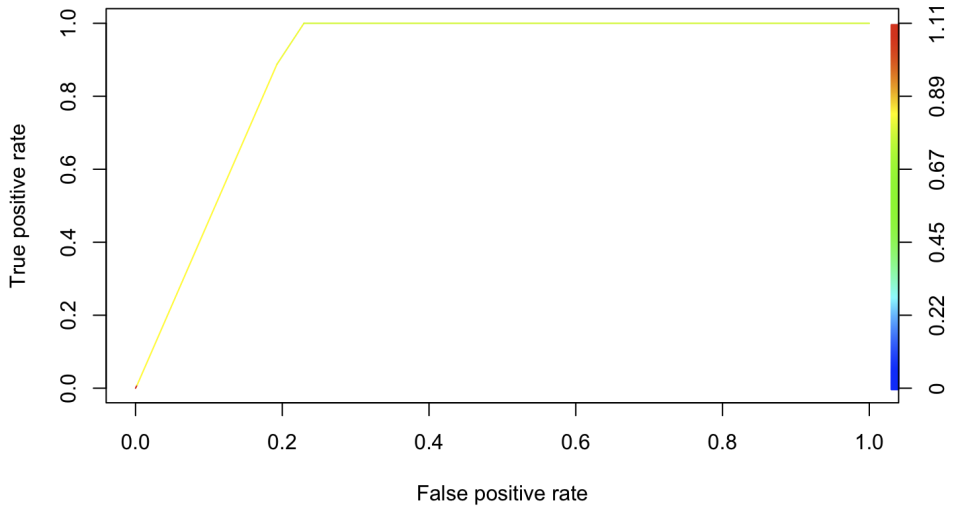
This is the in-sample ROC Curve and the AUC is 0.8876. The model is fitted very well because the AUC is much higher than the acceptance discriminatory power of 0.7. This AUC is also higher than the AUC that has been found in the Logistic Regression.

Out-of-sample Performance

Asymmetric cost	**Predicted**	
Truth	0	1
0	735	219
1	0	991

For the out-of-sample performance, the false negatives are also 0 cases. This is a very good sign since the model is trained properly and performed very well on the unseen data. The misclassification cost with the asymmetric cost is 0.31.

This is an out-of-sample ROC Curve and the AUC for this one is 0.8905.



Random Forest

The next step is to conduct a Random Forest model. The balanced data will be used in this method, starting with the 80% training data.

```
Call:
 randomForest(formula = stroke ~ ., data = stroke_train1, ntree = 500)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 0.6%
Confusion matrix:
     0    1 class.error
0 3860   47  0.01202969
1    0 3870  0.00000000
```

After applying the classification model on the balanced training data, the OOB estimate of error rate is 0.6% which is a very good rate. Out-of-bag (OOB) score is the way to validate the random forest model. It indicates the number of correctly predicted rows from the out-of-bag sample. In the confusion matrix, the false negative is 0 indicates that this model is successful in predicting all the stroke patients and avoiding sending them home. Next, the 20% testing data will be used on this Random Forest model to see how well the model performs on the unknown data. Below is the confusion matrix of the remaining 20% data. This model performs very well on the unseen

testing data because the false negative is 0 and the false positives are 10. This brings the accuracy of this random forest model to 99.5% of accuracy.

	Predicted	
Truth	0	1
0	944	10
1	0	991

Conclusion

In conclusion, after applying several models with different methods on the dataset. Random forest performs the best prediction model because it has 99.5% of accuracy. Furthermore, body mass index, glucose level, and age are the most influential stroke risk factors as they appear frequently in different models. On the other hand, genders, work types, and residence types are not associated well with stroke experience. This indicates patients have the same likelihood to experience stroke regardless of their gender, what they do, and where they live. Lastly, balancing the data is appropriate to improve the accuracy. However, in real-life problems, balancing the dataset by using this method will possibly lead to false accuracy. Hence, we can balance the dataset by updating more data about stroke patients. By keeping the data balanced helps to avoid sending stroke patients home and improves chances of giving them the needed medical services.

References

Analytics Vidhya. (2020, July 05). Imbalanced classification problems in r. Retrieved May 08, 2021, from https://www.analyticsvidhya.com/blog/2016/03/practical-guide-deal-imbalanced-classification-problems/

Fedesoriano. (2021, January 26). Stroke prediction dataset. Retrieved May 08, 2021, from https://www.kaggle.com/fedesoriano/stroke-prediction-dataset

Maklin, C. (2019, July 30). Random forest in r. Retrieved May 08, 2021, from https://towardsdatascience.com/random-forest-in-r-f66adf80ec9

Oliveira, S. (2017, April 10). A very basic introduction to Random Forests using R. Retrieved May 08, 2021, from https://www.blopig.com/blog/2017/04/a-very-basic-introduction-to-random-forests-using-r/

ROSE: Generation of synthetic data by randomly over sampling Examples (ROSE). (n.d.). Retrieved May 08, 2021, from https://www.rdocumentation.org/packages/ROSE/versions/0.0-3/topics/ROSE