Name: Vidya Praveen

Date: 06/30/2024

Course: DSC680 – Applied Data Science

Assignment: Week4 –Milestone3 – Final Report / White Paper

Project Topic: Predicting Stroke Occurrence using Patient Health Data.

Business Problem

Stroke is considered the second largest cause of death, worldwide, according to World Health Organization. To reduce this, many studies are done where they have seen a pattern in the people who had stroke. Based on this, in this project, I am trying to develop a predictive model which will help identify the chances of stroke occurrence in individuals, based on their health data. Early prediction of stroke can significantly enhance preventive healthcare measures and reduce the risk of severe health complications.

Background/History

Strokes occur when the blood supply to part of the brain is interrupted or reduced, preventing brain tissue from getting oxygen and nutrients. Stroke is a significant public health concern due to its high morbidity and mortality rates. Traditional methods of stroke risk assessment rely on clinical judgment and basic diagnostic tools, which can be subjective and limited in predictive power. Machine learning and data science offer the potential to enhance stroke prediction accuracy by analyzing complex patterns in patient health data.

Data Source

Datasets: I have used the dataset from Kaggle: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset/data

The dataset includes various health attributes of individuals and a target variable indicating whether the individual had a stroke. This dataset contains about 5000 records with information such as:

- Demographic information (age, gender)
- Health conditions (hypertension, heart disease)
- Lifestyle factors (smoking status, marital status, work type)
- Medical measurements (average glucose level, body mass index)
- Outcome (stroke occurrence)

Data Preparation

- Missing values in the bmi column were filled with the mean value.
- Categorical variables were encoded using Label Encoding.
- Features were standardized using StandardScaler.

Methods

The project employs several machine learning techniques:

Data Preprocessing: Imputation, encoding, and normalization.

Exploratory Data Analysis (EDA): Identifying patterns and relationships in the data.

Modeling: Several models were employed for this analysis:

1. **Random Forest Classifier:** Chosen for its robustness and ability to handle imbalanced datasets. Hyperparameter tuning was conducted using Grid Search.
2. **Cox Proportional Hazards Model:** This model was used to analyze the time-to-event data, considering the duration and event occurrence (stroke).
3. **Kaplan-Meier Estimator:** Applied to estimate the survival function from the lifetime data, particularly useful for different groups within the dataset.

Hyperparameter Tuning: Grid Search was used to find the best hyperparameters for the Random Forest model.

Analysis

The analysis involves comparing different models to identify the one with the highest predictive accuracy and robustness. The performance of each model is evaluated using cross-validation. Feature importance analysis is conducted to identify the most significant predictors of stroke. Additionally, survival analysis provided insights into the time-dependent risk of stroke.

See Appendix for results.

Results

- **Random Forest Model:** Achieved a high accuracy of approximately 94%, but struggled with predicting positive stroke cases due to class imbalance.
- **Cox Proportional Hazards Model:** Provided a detailed summary of the hazard ratios for different predictors.
- **Kaplan-Meier Estimator:** Offered survival functions for different groups, such as gender, showing the probability of stroke occurrence over time.

Conclusion

The Random Forest model achieved a high accuracy of approximately 94%. However, the model struggles with predicting positive stroke cases as indicated by the poor precision and recall for the positive class. The survival models provided valuable insights into the risk factors and their impact over time.

Assumptions

- The dataset is representative of the broader population.
- The features included are relevant and sufficient for predicting stroke risk.
- Missing values are missing at random and can be imputed accurately.

Limitations

- Imbalanced dataset with very few positive stroke cases.
- Limited feature set which might not capture all relevant factors influencing stroke risk.

Challenges

Challenges that I faced during the project are:

- Data Imbalance: The occurrence of stroke cases are significantly lower than non-stroke cases, leading to class imbalance issues. Also, not able to identify any feature that has significant impact on the stroke occurrence.
- Missing Data: Handling missing values in features like BMI, which is critical for accurate predictions.

Future Uses/Additional Applications

- Implementing advanced techniques like SMOTE (Synthetic Minority Over-sampling Technique) to handle class imbalance.
- Exploring other machine learning models like Gradient Boosting or Neural Networks.
- Integrating more features, such as family history and genetic factors, to improve prediction accuracy.

Recommendations

- Use techniques to handle class imbalance.
- Collect more data, particularly with more positive stroke cases.
- Continuously monitor and validate the model with new data.

Implementation Plan

1. **Data Collection and Preprocessing:** Continuously update and preprocess data.
2. **Model Development:** Refine and retrain the predictive model using updated data.
3. **Integration:** Collaborate with healthcare providers to integrate the model into clinical workflows.
4. **Monitoring and Evaluation:** Regularly monitor model performance and make necessary adjustments.

Ethical Assessment

Potential ethical concerns that I considered while gathering data were:

• Data Privacy: Ensured that patient data is anonymized.

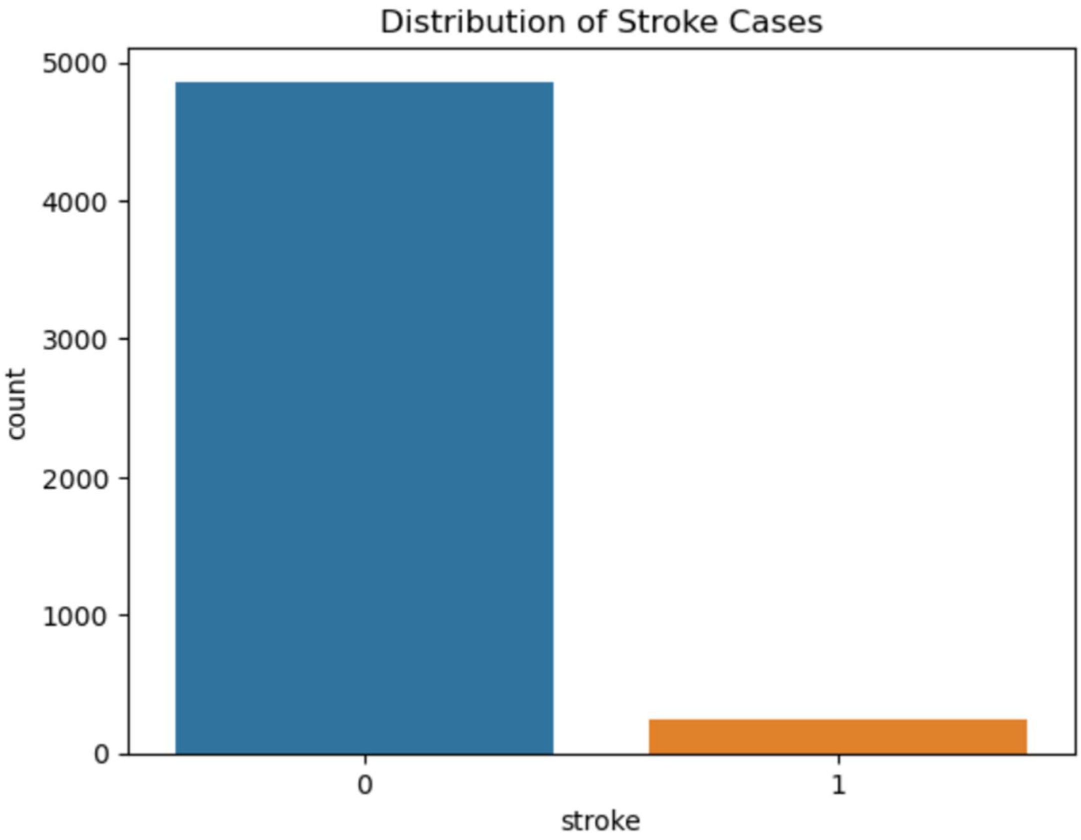
• Bias and Fairness: Looked for data from reliable sources where there is no biases in the data that may lead to unfair predictions.

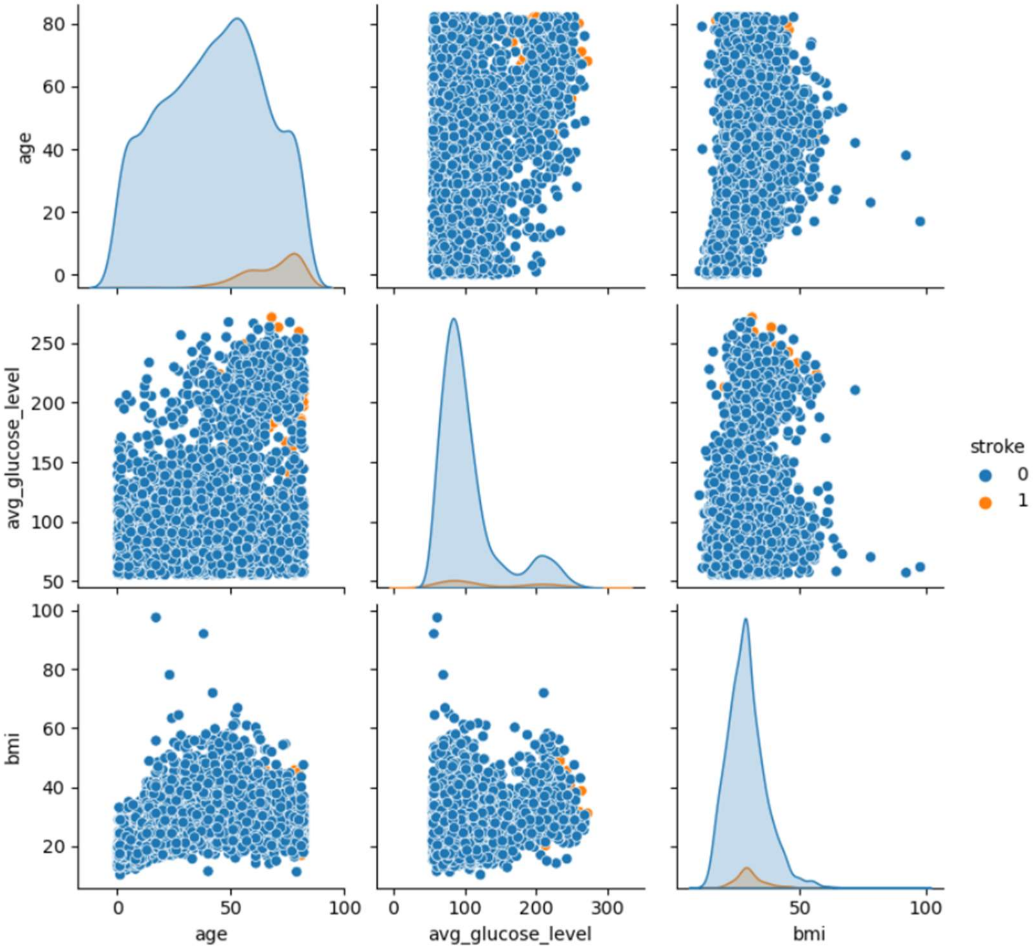
Use the model as a decision support tool, not as the sole basis for medical decisions.

Appendix

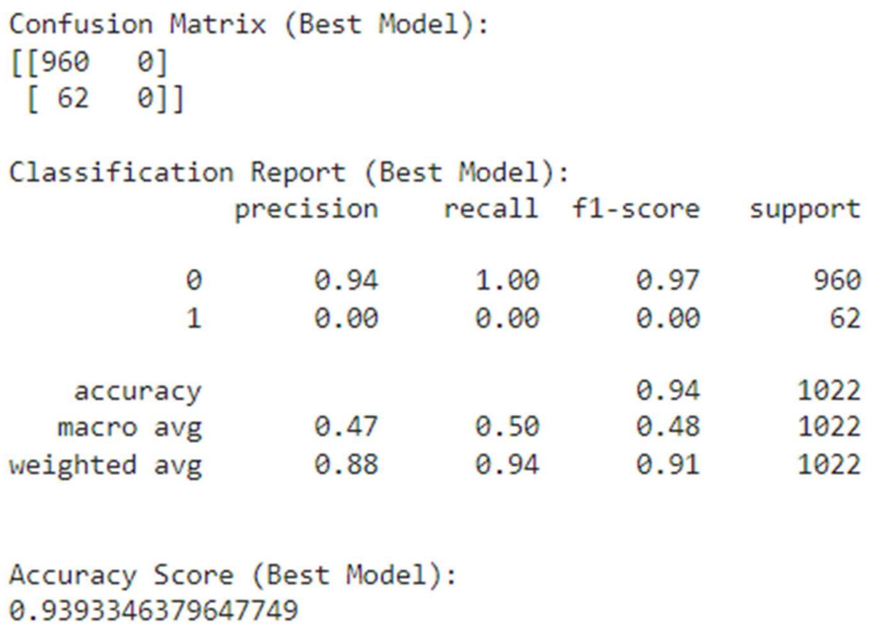
Data Dictionary

Feature	Description
age	Age of the patient
gender	Gender of the patient
hypertension	0: No hypertension, 1: Hypertension
heart_disease	0: No heart disease, 1: Heart disease
ever_married	Marital status of the patient
work_type	Type of occupation
residence_type	Urban or rural residence
avg_glucose_level	Average glucose level
bmi	Body Mass Index
smoking_status	Smoking status
stroke	0: No stroke, 1: Stroke (target variable)

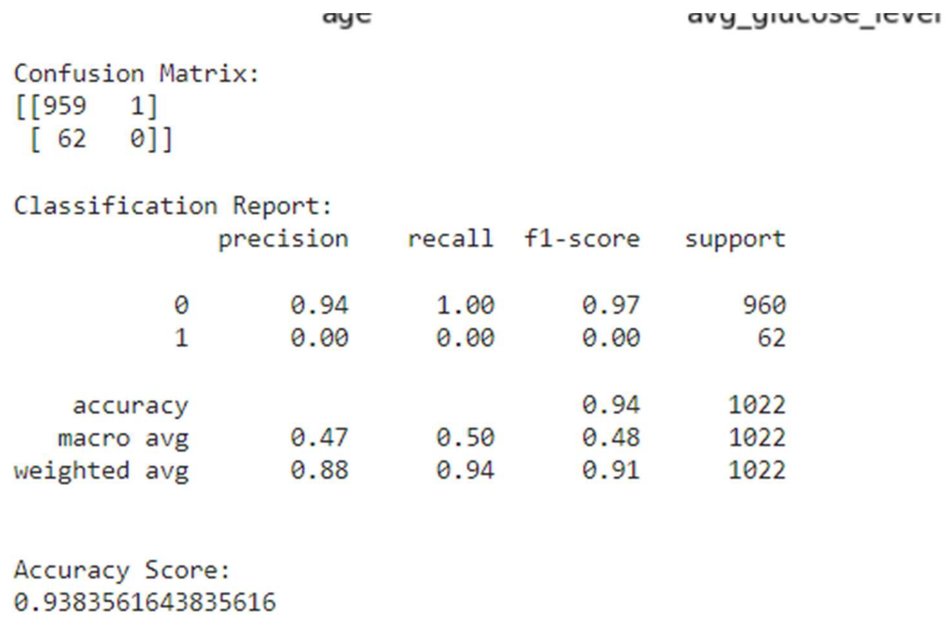

Distribution of Stroke Cases
count
5000
4000
3000
2000
1000
0
0
1
stroke



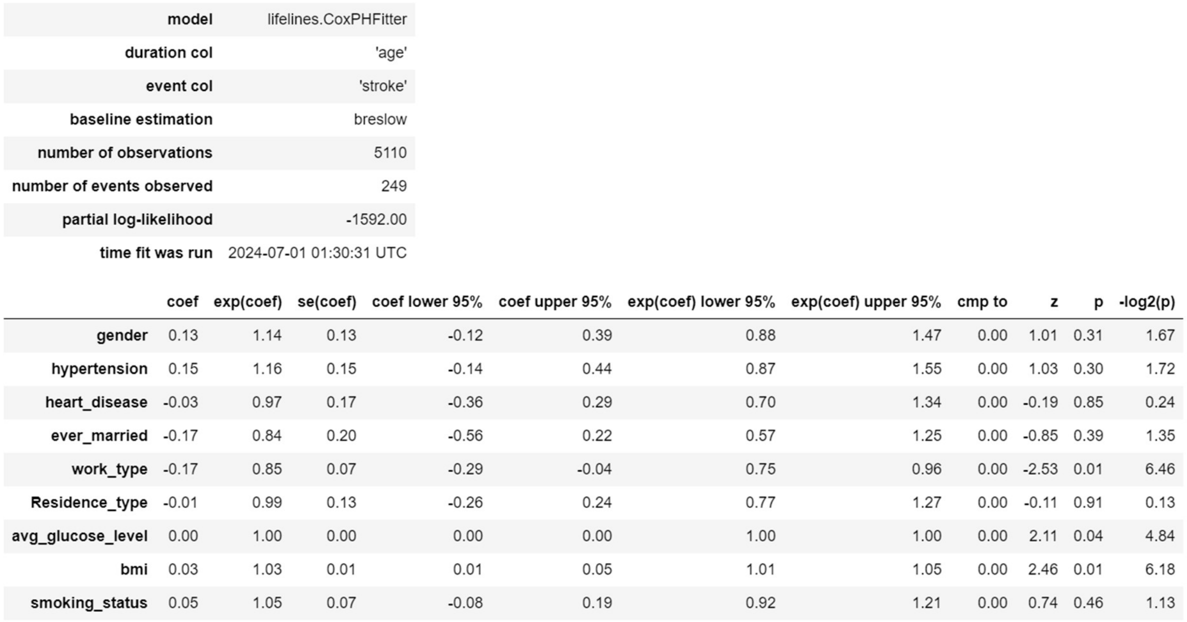
age
avg_glucose_level
bmi
0
20
40
60
80
0
50
100
0
100
200
300
0
50
100
50
100
150
200
250
20
40
60
80
100
stroke
● 0
● 1



```
Confusion Matrix (Best Model):
[[960   0]
 [ 62   0]]

Classification Report (Best Model):
              precision    recall  f1-score   support

           0       0.94      1.00      0.97       960
           1       0.00      0.00      0.00        62

    accuracy                           0.94      1022
   macro avg       0.47      0.50      0.48      1022
weighted avg       0.88      0.94      0.91      1022


Accuracy Score (Best Model):
0.9393346379647749
```

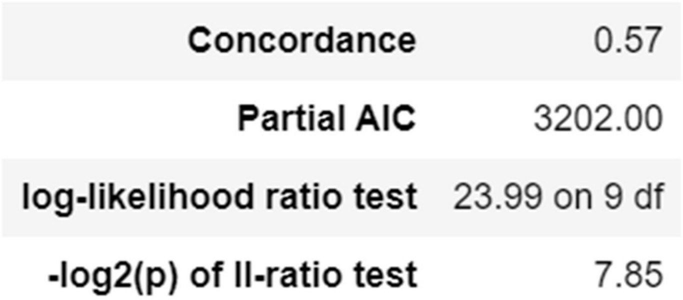


```
Confusion Matrix:
[[959   1]
 [ 62   0]]

Classification Report:
              precision    recall  f1-score   support

           0       0.94      1.00      0.97       960
           1       0.00      0.00      0.00        62

    accuracy                           0.94      1022
   macro avg       0.47      0.50      0.48      1022
weighted avg       0.88      0.94      0.91      1022


Accuracy Score:
0.9383561643835616
```

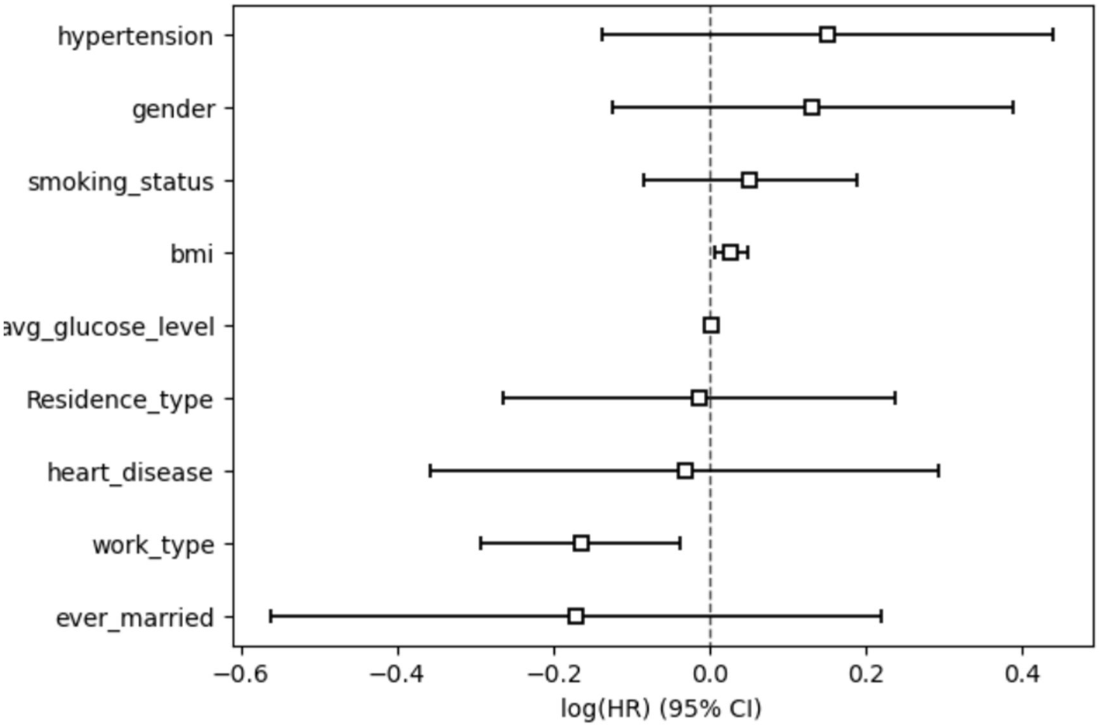


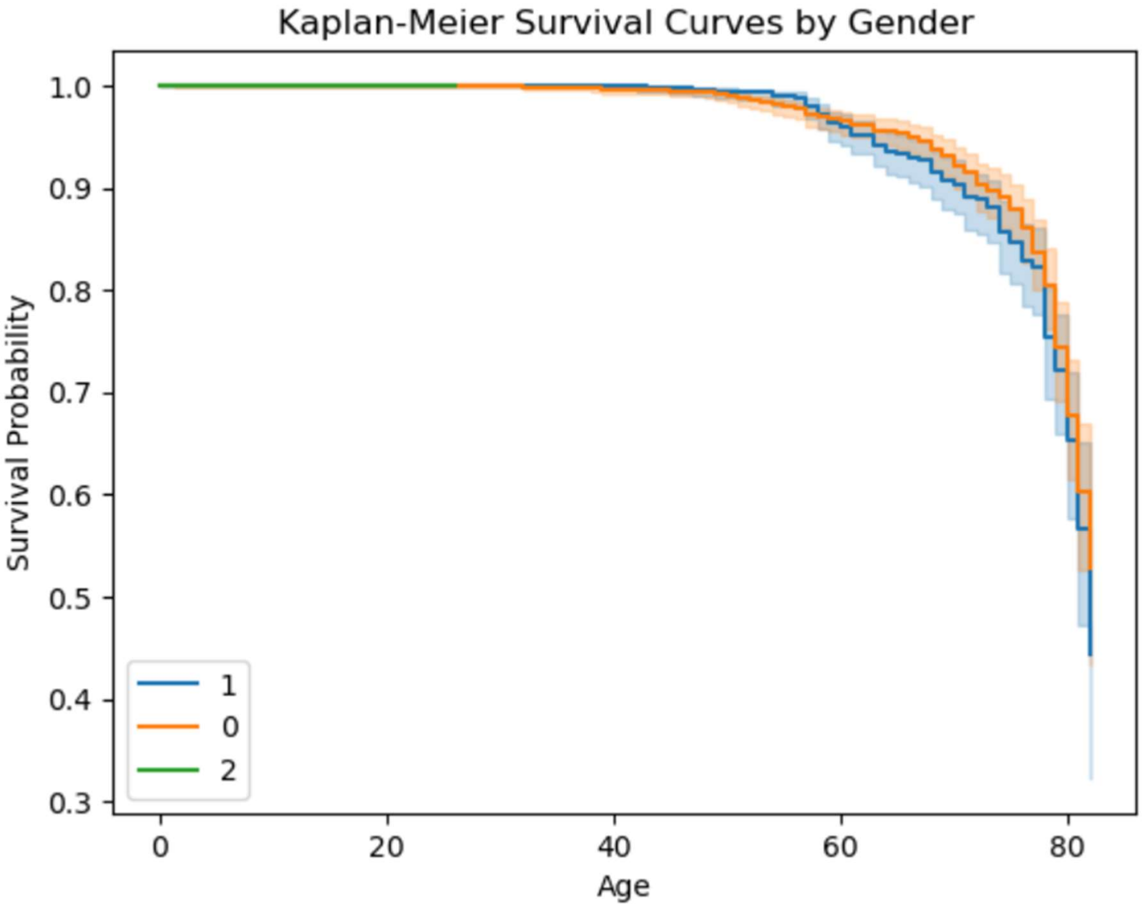
model	lifelines.CoxPHFitter
duration col	'age'
event col	'stroke'
baseline estimation	breslow
number of observations	5110
number of events observed	249
partial log-likelihood	-1592.00
time fit was run	2024-07-01 01:30:31 UTC

	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	cmp to	z	p	-log2(p)
gender	0.13	1.14	0.13	-0.12	0.39	0.88	1.47	0.00	1.01	0.31	1.67
hypertension	0.15	1.16	0.15	-0.14	0.44	0.87	1.55	0.00	1.03	0.30	1.72
heart_disease	-0.03	0.97	0.17	-0.36	0.29	0.70	1.34	0.00	-0.19	0.85	0.24
ever_married	-0.17	0.84	0.20	-0.56	0.22	0.57	1.25	0.00	-0.85	0.39	1.35
work_type	-0.17	0.85	0.07	-0.29	-0.04	0.75	0.96	0.00	-2.53	0.01	6.46
Residence_type	-0.01	0.99	0.13	-0.26	0.24	0.77	1.27	0.00	-0.11	0.91	0.13
avg_glucose_level	0.00	1.00	0.00	0.00	0.00	1.00	1.00	0.00	2.11	0.04	4.84
bmi	0.03	1.03	0.01	0.01	0.05	1.01	1.05	0.00	2.46	0.01	6.18
smoking_status	0.05	1.05	0.07	-0.08	0.19	0.92	1.21	0.00	0.74	0.46	1.13



Concordance	0.57
Partial AIC	3202.00
log-likelihood ratio test	23.99 on 9 df
-log2(p) of ll-ratio test	7.85


hypertension
gender
smoking_status
bmi
avg_glucose_level
Residence_type
heart_disease
work_type
ever_married
−0.6
−0.4
−0.2
0.0
0.2
0.4
log(HR) (95% CI)


Kaplan-Meier Survival Curves by Gender
Survival Probability
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0
20
40
60
80
Age
1
0
2

Questions and Answers

1. Did you explore alternative techniques for encoding categorical variables apart from Label Encoding?

   Yes, I considered using One-Hot Encoding for the categorical variables. However, given the nature of the dataset and the model selected, Label Encoding was more straightforward and efficient, especially since the categorical variables had a limited number of categories.

2. What influenced your decision to employ a Random Forest classifier for this analysis?

   I chose the Random Forest classifier because of its robustness and ability to handle imbalanced datasets. It also performs well with a mixture of numerical and categorical data and provides feature importance scores, which are valuable for understanding the model.

3. How have you dealt with the issue of class imbalance within your dataset?

   Class imbalance was a significant challenge. To address this, I used techniques such as resampling the dataset to balance the classes. Additionally, I plan to explore methods like SMOTE (Synthetic Minority Over-sampling Technique) in future iterations to generate synthetic samples for the minority class.

4. On what basis did you select hyperparameters for the Random Forest model?

Hyperparameters were selected using Grid Search with cross-validation.

5. Have you considered other models like Gradient Boosting or Neural Networks, and if so, how do they stack up in terms of performance?

Hyperparameters were selected using Grid Search with cross-validation. This approach helped to systematically evaluate different combinations of hyperparameters to find the ones that provide the best performance for the model.

6. The recall for the positive class (stroke) appears quite low; what strategies might enhance it?

To enhance recall for the positive class, I am considering several strategies:

- Using SMOTE to address class imbalance.
- Adjusting the decision threshold of the classifier to favor the minority class.
- Implementing ensemble methods that combine multiple models to improve overall performance.

7. Which metrics, other than accuracy, did you use to assess the efficacy of your model?

Besides accuracy, I used precision, recall, F1-score, and the area under the ROC curve (AUC-ROC). These metrics provide a more comprehensive view of the model’s performance, especially in the context of class imbalance.

8. Are there any potential features or external data that could be introduced to increase the model's precision?

   Yes, there are several potential features that could enhance the model’s precision:

   - Family history of stroke and genetic factors.
   - More detailed lifestyle information, such as diet and exercise habits.

9. In future iterations, how do you intend to overcome the challenge of the low recall rate for the positive class?

   Future iterations will focus on:

- Implementing SMOTE or other oversampling techniques.
- Tuning the model’s threshold to balance precision and recall better.
- Experimenting with different models like Gradient Boosting or Neural Networks that might better capture the nuances of the minority class.
- Collecting more data, particularly more instances of stroke occurrences, to improve the model’s learning.

10. Can you detail how this predictive model could be integrated into actual healthcare practices?

The predictive model could be integrated into healthcare practices in the following ways:

- As a decision support tool for clinicians, helping them identify high-risk patients and prioritize preventive measures.
- Integration into electronic health record (EHR) systems to provide real-time risk assessments during patient visits.

- Used in public health initiatives to identify and target high-risk populations for early intervention programs.