Stroke Fatality Classification

*Project Report for Machine Learning Course

Baktybek Doskul
Department of Computer Science
Nazarbayev University
Nur-Sultan, Kazakhstan
baktybek.doskul@nu.edu.kz

Bota Kabiyeva
Department of Mathematics
Nazarbayev University
Nur-Sultan, Kazakhstan
bota.kabiyeva@nu.edu.kz

Zhanggir Yergaliyev
Department of Computer Science
Nazarbayev University
Nur-Sultan, Kazakhstan
zhanggir.yergaliyev@nu.edu.kz

Talgat Omarov
Department of Computer Science
Nazarbayev University
Nur-Sultan, Kazakhstan
talgat.omarov@nu.edu.kz

Aituar Kenges
Department of Computer Science
Nazarbayev University
Nur-Sultan, Kazakhstan
aituar.kenges@nu.edu.kz

***Abstract*—Stroke is a widespread disease from which millions of people suffer worldwide. Our task was a binary classification, which predicts whether a patient which had a stroke will die or not. We trained machine learning models on a real data from 150 anonymous patients at UMC. Receiver operating characteristics (ROC) curves which are commonly used in medical decision making were used for evaluating our model. We tested our dataset on six different models. Two of the models that have shown best results were SVM and Logistic Regression with L2 penalty, they have shown the score of around 89%. Recursive feature elimination was used as a feature selection technique.**

***Index Terms*—Stroke, binary classification, LDA, Linear Regression, SVM, KNN, Naive Bayes, recursive feature elimination**

I. Introduction

There are over 13.7 million new strokes each year worldwide. Undoubtedly, it is terrible widespread disease, alongside with its high mortality rate. Indeed, stroke is rated 5th cause of death and leading cause of various disabilities in the United States. [1] It can be considered as a medical condition, in which brain cells die of the poor blood flow, resulting in failure of body functions controlled by damaged part of the brain. About 80% of strokes are ischemic stroke type, it is caused by obstruction of the arteries of the neck or brain by a clot; it could also be transient ischemic attack (TIA) caused by a temporary clot. And the rest occurs when weak arteries burst and bleed into the brain, which is called hemorrhagic stroke. [2]

Whenever someone experiences the stroke, they need emergency treatment, whether by restoring the blood flow or by reducing blood pressure in the brain. Objectives with the highest priority for the doctors is to prevent death, and disabilities further on.

This project intends to binary classify whether the patient has survived the stroke or not. Many factors influence stroke occurrence as well as the conditions of a patient's post-stroke life: history of cardiovascular disease, diabetes, other diseases or medications that were used during treatment. For interpretable models results can potentially uncover vital insights on what could decrease or contribute to the mortality rate for patients with stroke history.

The basis for this project is real data provided by Professor Dmitriy Viderman. Initial objective is to preprocess that document, consisting of history records of patients with stroke of University Medical Center (UMC). Afterwards, several classification algorithms are to be applied, and results to be discussed - all to get valuable insights on stroke fatality.

II. Dataset

Description

The original dataset given for this project is the real data recorded for 150 anonymous patients at UMC. Each patient was delivered to the hospital with a stroke diagnosis and placed at the hospital. Baseline information about patients before hospital health conditions were recorded, as well as essential medical records during the stay there. Some of the features were descriptive information such as primary diagnosis, history of certain diseases (diabetes, CHD, etc) and taken drugs; while other information was dynamically recorded for each day during patient was at hospital's care (heart rate and blood pressure).

Preprocessing

Provided dataset required data preprocessing before using for training the algorithm. First, there were missing data in several features, especially the ones recorded day by day. Possible steps to take were to complete missing data with corresponding statistics. However, after advising with Prof. Dmitriy Viderman, it was decided that dynamic data can be omitted since they were not recorded accurately enough to be relevant. There were also other features that were considered to be irrelevant for the prediction of the death caused by the stroke model (weight and height, etc). Almost all features were qualitative and categorical (regularly taken drug, the drug administered in ICU, etc), and few were quantitative, such as weight and height of the patient. To deal with categorical data One-Hot Encoding method was used.

After relevant categorical features were processed, from common sense columns like Patient number, patient initials were dropped, so that their values will not affect the data. As initial data for the training model only columns with numeric values were selected. It also was crucial to split out features and target column (”Death”) separately. By doing so, data were divided into 2 classes, ”0” and ”1” standing for whether patients survived stroke or not. Class distribution was found to be 62.67% and 37.33%, respectively.

Preprocessing resulted in overall 156 features being considered. In illustrative purposes, data feature distributions have been plotted. Following Figure 1 provides illustration examples we used.

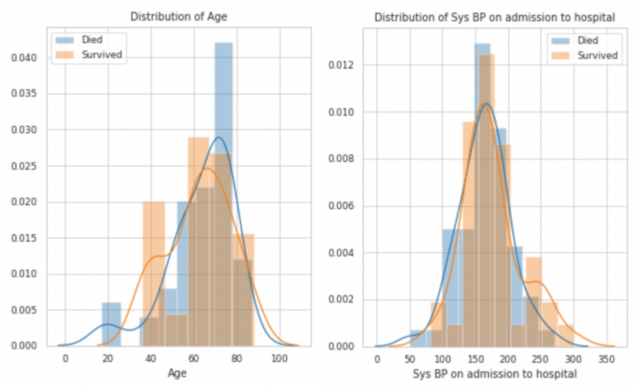

Fig. 1. Feature class distribution

III. Methodology

To make an inference about the heart stroke fatality, we chose six candidate models described in Table I. We selected highly interpretable and intuitive models since interpretability plays a crucial role in medical applications. All models were available in open-source machine learning library **scikit-learn**. [3]

TABLE I
Candidate models.

Model	Description
LDA	Linear Discriminant Analysis
LR L1	Logistic Regression with L1 penalty
LR L2	Logistic Regression with L2 penalty
SVC	Support Vector Machine Classifier
KNN	K-Nearest Neighbors Classifier
NB	Naive Bayes Classifier

Evaluation Metrics

To evaluate candidate models we used **receiver operating characteristics (ROC)** curves. ROC curves are commonly used in medical decision making because of its property to deal with skewed class distribution and unequal classification error costs [4]. The ROC curve provides a graphic representation of the trade off between false-positive rates and true-positive rates [5]. In addition to ROC curves, we used area under the curve (AUC) of ROC as a single number evaluation metrics.

Model training

To properly train and evaluate our model, we randomly shuffled and split the dataset in the following way: we kept 50 patient data in the test set, and trained the models on the remaining 100 patient data. As a result, we achieved a test size of 33.3%.

For model selection and hyper-parameter tuning we used grid search algorithm on the training dataset with 10-fold cross validation on ROC AUC metrics. Corresponding regularization parameters were explored for LDA, Support Vector Machine and Logistic Regression methods. Similarly, for KNN optimal number of neighbors and for Naïve Bayes “alpha” – additive smoothing parameter were estimated.

Feature Selection

We used **recursive feature elimination** algorithm [6] as feature selection technique for LDA, LR L2 and SVM. Particularly, we adopted implementation from [7] using absolute magnitude of feature coefficients in linear models as feature importance:

Algorithm 1: Recursive Feature Elimination

```
Tune/train the model on the training set using all P
  predictors;
Calculate model performance;
Calculate variable importance or rankings;
for each subset size $S_i$, $i = 1 \dots S$ do
    Keep the $S_i$ most important variables;
    Tune/train the model on the training set using $S_i$
      predictors;
    Calculate model performance;
end
Calculate the performance profile over the $S_i$;
Determine the appropriate number of predictors (i.e.
  the $S_i$ associated with the best performance);
Fit the final model based on the optimal $S_i$;
```

For LR L1, we used the property of L1 penalty to cause a subset of the solution coefficient to be exactly zero as automatic feature selection.

IV. Results

From Cross validation ROC curves of six candidate models (Figure 2) we obtain that SVC and LR L2 are the most effective in predicting the death outcomes of the stroke with AUC around 0.89 each. Therefore, recommending the exact one model for classification task was quite challenging.

One of the objectives of this analysis is to identify the major predictors of the stroke. For the models Linear Discriminant Analysis (LDA), Logistic Regression with L2 penalty(LR L2), Support Vector Machines (SVM) we selected features that affected mostly to the model’s AUC_ROC score performance.

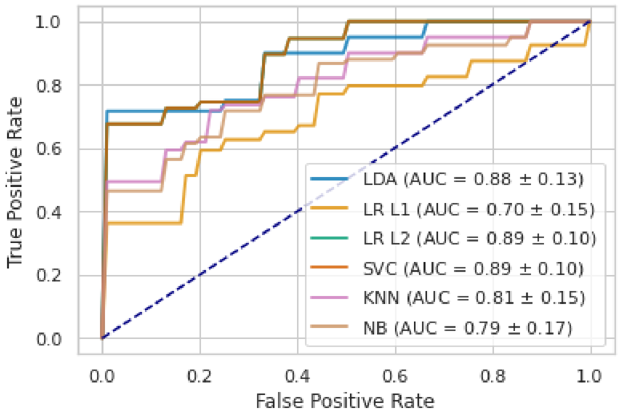

Fig. 2. ROC curves (10-fold-CV).

Tables II and III shows the features of LR L2 and SVC by their significance ranking. From the tables we obtain that independent variables used by models differ significantly. From medical practice, features that are exploited by Logistic Regression with L2 penalty are more reasonable. Because features resulted from other models mostly was coming from the medicine type which patients were given in ICU. This occured possibly because these type of drugs were administered for patients under critical health conditions. Whereas, features from Logistic Regression made more sense in terms of their possible effect on death. Thus, we have chosen LR L2 as a recommended classification model for this task.

The magnitude of coefficients in the table can be interpreted as feature significance. The higher are the absolute values of coefficients, the higher are the contributions of corresponding features. Positive coefficients correlate positively with stroke fatality and vice versa.

TABLE II
SIGNIFICANCE OF THE FEATURES(LR L2 ASC.).

Rank	Feature	Signif. coef.
1	Stroke_diag_TIA	0.53363
2	stroke_area_rmch	0.40458
3	Gender	0.30913
4	Age	-0.29046
5	History of Metabolic Syndrome	-0.27771
6	Dias BP on admission to hospital	0.25293
7	History of A-Fibrillation	-0.22723
8	Stroke_diag_hemorrhagic	-0.22068
9	GCS on admission to hospital	-0.20463
10	History of Hypertension	-0.20181
11	History of Chronic Liver Failure	-0.19818
12	Heart Rate on admission to hospital	-0.16592
13	Stroke_diag_ischemic	-0.16592
14	History of IHD	0.15965
15	Stroke_diag_mixed	0.1581
16	History of Chronic Renal Failure	0.13755
17	Sys BP on admission to hospital	0.13299
18	History of Diabetes Mellitus	0.13191

Finally, we evaluated our final model, Logistic Regression with L2 penalty, on the test set. Figure 3 illustrates corresponding ROC curve. The model achieved 0.86 AUC ROC score.

TABLE III
SIGNIFICANCE OF THE FEATURES(SVC ASC.).

Rank	Feature	Signif. coef.
1	History of Diabetes Mellitus	-0.1231256573
2	antihypertensive_yes-non-regularle	0.0548167201
3	antihypertensive_icu_physiotens	0.0550767316
4	antihypertensive_icu_mannitol	0.0555605581
5	celebral_edema_admission	0.0666026547
6	glucocorticosteroids_dexamethasone	0.0676293175
7	hemostatics_etamsylate	-0.071842188
8	hemostatics_present	-0.071842188
9	antihypertensive_icu_furosemide	-0.0822535545
10	antihypertensive_icu_nebilet	-0.0837916673
11	stroke_area_lmch	-0.0852544987
12	anticoagulants_antiplatelet_enoxaparin_sodium	-0.0923787122
13	antihypertensive_icu_enalapril	-0.0963842513
14	antihypertensive_icu_enap	0.1076421759
15	antihypertensive_icu_magnesium_sulfate	-0.1205002618
16	History of Chronic Renal Failure	0.12774792
17	vasopressors_present	0.1774240221
18	vasopressors_dopamine	0.2301270346

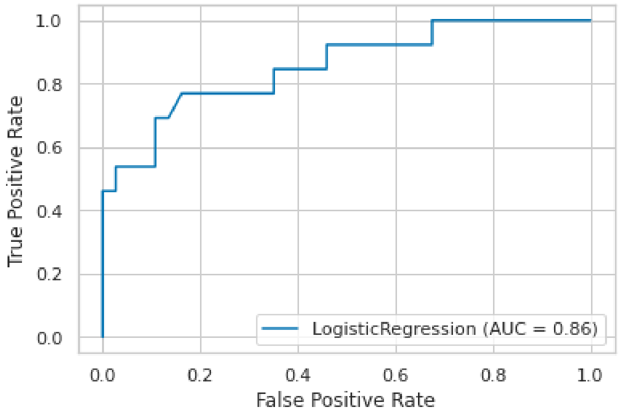

Fig. 3. ROC Curve of Logistic Regression with L2 penalty on Test Set.

V. CONCLUSION AND FURTHER WORKS

The aim of our project was to find out whether a patient survived a stroke. We trained our models on the real data recorded from 150 patients at UMC. The data was cleaned from least valuable parts and categorical data was transformed into one-hot encodings.

The results of our models suggests that we succeeded in predicting patient's death from stroke. The Logistic Regression with L2 penalty turned out to be the most suitable model for our binary classification.

For the further work, we would like to train our models on a larger dataset, since our original dataset contained only 150 patients. Additionally, further work can be done on the prediction of the second stroke for those patients who survived the first stroke. Stroke is a disease that affects patient's health condition severely. Therefore, maintaining patient's health conditions in order to prevent a second stroke is an important research area.

References

[1] V. F. et al. The Lancet Neurology, “Global, regional, and national burden of stroke, 1990 to 2016: A systematic analysis for the global burden of disease study 2016,” *World Stroke Organization - Global Stroke Fact Sheet*, pp. 3–5, 2019.

[2] A. A. Joseph R.Shiber, Emily Fontane, “Stroke registry: hemorrhagic vs ischemic strokes,” *Elsevier*, vol. 28, no. 3, pp. 331–333, 2010.

[3] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[4] T. Fawcett, “An introduction to roc analysis,” *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

[5] D. K. McClish, “Analyzing a portion of the roc curve,” *Medical Decision Making*, vol. 9, no. 3, pp. 190–195, 1989.

[6] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, “Gene selection for cancer classification using support vector machines,” *Machine learning*, vol. 46, no. 1-3, pp. 389–422, 2002.

[7] M. Kuhn and K. Johnson, *Applied predictive modeling*. Springer, 2013, vol. 26.