WGU C951

Task 3

MACHINE LEARNING PROJECT PROPOSAL

Name Zachary Trani

Student ID # 011567582

Date December 21st, 2024

A. Project Overview

A machine learning system leveraging the logistic regression model will be implemented to predict the occurrence of strokes in patients. True Health Care Provider sponsors this machine learning system so it may better equip its patients with preventative knowledge.

A.1. Organizational Need

The company has created the task to use a machine learning model to better understand the effects on its patients when assessing their stroke risk. The current state of True Health Care Provider relies on a reactive versus preventative approach towards the occurrence of strokes in its patients. A machine learning model is needed to calculate the likelihood of a stroke and inform its patients, which would arm them with the knowledge of their underlying risks before it's too late.

A.2. Context and Background

Traditionally, the healthcare industry has been more reactionary toward the prevention of strokes. However, many of these risk factors can be reduced to help prevent a stroke (“Causes of Stroke”). Additionally, the industry is starting to use a more data-driven approach to preventative measures. A machine learning model used for predicting a patient's risk of stroke is one of the most popular examples of these data-driven strategies.

A.3. Outside Works Review

To guide the feasibility of the project, three outside works will be reviewed, and their methods discussed. The following works include research studies to demonstrate the application of machine learning models for stroke prediction.

(Chahine et al.) reviews various machine learning algorithms in stroke risk prediction, such as logistic regression, support vector machines, random forests, gradient-boosted trees, and deep neural networks. The paper emphasizes how machine learning models can handle complex multiple input data sets better than traditional statistical methods in predicting. This source highlights our relevance in using a logistic regression model for stroke prediction at True Health Care Provider. (Chahine et al.) provided thoroughly reviewed evidence in their paper and came up with a logistic regression model that showed visible coefficients, or weights, that doctors could follow along and agree with. This validates our choice of using logistic regression as a model for stroke prediction.

The following work thoroughly reviewed for stroke prediction is a paper by (Biswas et al.) where eleven machine learning models were implemented for predicting a stroke. Models such as Support Vector Machine, Random First, K-Nearest Neighbor, andg Logistic Regression, were developed. The logistic regression model demonstrated high accuracy in the prediction of strokes after data balancing and hyperparameter tuning were applied to the model. The high accuracy of the logistic regression model, when tested amongst other highly technical machine learning models, further demonstrates our use case of developing the model at True Health Provider.

Another source, (Daidone et al.) reviewes supervised models such as logistic regression, support vector machines, random forests, and deep neural networks and highlighted that logistic regression is crucial for an accurate model that is interpretible. Emphasis was again on the fact that medical staff can interpret the model and closely inspect the coefficients or weights on various factors and there relation towards a stroke occurrence. The interpretability of our logistic regression model supported by this source will allow True Health Care Providers to verify the results of this project easilyt.

A.4. Solution Summary

A Supervised Learning machine learning logistic regression model, will be used to model binomial outcomes with multiple exploratory parameters. After training is complete, coefficient values, or weights, will be generated for the binomial model. The model will predict whether a patient is likely to get a stroke based on various input parameters such as gender, age, numerous diseases, and smoking status. The project uses a stroke prediction dataset with a sample size of over 5,000 patients. The result will return a probability of whether a patient is likely to have a stroke; a probability of 0.5 or greater will return an answer of potential to have a stroke, while a probability of under 0.5 will return an answer of unlikely to have a stroke.

A.5. Machine Learning Benefits

The logistic regression model system will improve predictions towards assessing the probability of a stroke as it will mathematically model the known contributing factors toward strokes. The model will assign weights to each contributing factor so a patient or doctor may see the importance of each factor if one or more is more relevant to them. As a result, a new patient will be able to enter their data and receive immediate, accurate feedback on whether they are at risk of a stroke.

B. Machine Learning Project Design

B.1. Scope

In-Scope Items

- Gather Dataset: Locate a data set with multiple known contributing parameters that cause a stroke.
- Data Cleanse: Convert any necessary data from text format to integer format.
- Create the Model: Implement the logistic regression machine learning model and train it using some of our available data.
- Testing: Run the remaining data through the model to generate probabilities on a new subset of the cleansed data.

Out-of-Scope Items

- Continuous Learning: After the initial training of the model, no new sets of data will be input to train it further.

B.2. Goals, Objectives, and Deliverables

Goals

•Inform Patients: The main goal of this project is to inform those currently at risk of a stroke.

•Low Cost: Generate a low-cost predictive model that does not require expensive languages or licensing.

Objectives

•Throughput: The model is trained with 3,750 patients.

•Accuracy: The model has an accuracy of at least 85% or greater by testing it against a dataset with known outcomes.

•Up Time: Once developed, the predictive model is deployed within the web and has 99% uptime.

Deliverables

• Model: A trained logistic regression model that accurately predicts stroke risk based on multiple contributing parameters.

•Interface: The model can be used by patients through the True Health Care Provider website which renders an instance of Google Colab, which hosts the Jupyter Notebook instance of our model.

B.3. Standard Methodology

- **Sample**:

  Objective:

  A subset of data will be selected in this phase, including the volume of data and the identification of independent and dependent variables.

  Application:

  Data Classification: Select specific medical characteristics or variables from the True Health Care Provider database as they relate to the contribution of a stroke occurrence, medical variables such as age, gender, BMI, and preexisting conditions like hypertension or heart disease.

  Sample Sizing: Choose a suitable sample size that is large enough to capture variability between input parameters, but is not so large that it becomes computationally challenging to train the model.

- **Explore:**

  Objective:

  Data analysis of the selected dataset, including single variable trends or multivariable correlations

  Application:

  Data Analysis: Calculate initial statistics, such as the mean and standard deviation, independently for each input parameter.

  Data Visualization: Create visuals such as box plots and histograms to identify early correlations between variables, such as the connection between increasing age and the occurrence of strokes.

  Outlier Identification: Using the previous steps visuals, observe any outliers in the data that will influence the study's outcome.

- **Modify:**

  Objective:

  Transform data into usable input for the model.

  Application:

  Data Clean: Scan through a data set and remove any missing values.

  Data Transformation: Convert text data to a numerical value for it to fit the mathematical model.

- **Model:**

  Objective:

  Machine Learning Model: Implement and train the logistic regression model using the chosen data set.

  Application:

  Model Selection: Choose the logistic regression model due to its interpretability for binary classification of stroke prediction occurrences.

  Training: Most of the data set is used to train the model, and the remaining data is used to test it, using a split such as 75% training and 25% testing.

- **Assess:**

  Objective:

  Assess the model's performance for accuracy.

  Application:

  Accuracy: Calculate the percentage of correctly predicted occurrences of strokes out of all test cases.

  Validation: Assess the model on an additional subset of test data.

B.4. Projected Timeline

Task Number & Description	Start Date	End Date
Task 1: Team Selection & Briefing	January 6, 2025	January 10, 2025
Task 2: Data Analysis Calculations & Validation	January 13, 2025	January 15, 2025
Task 3: Data Visualizations & Validation	January 16, 2025	January 17, 2025
Task 4: Data Cleanse & Validation	January 21, 2025	January 23, 2025
Task 5: Data Transformation & Validation	January 24, 2025	January 28, 2025
Task 6: Jupyter Notebook Development	January 29, 2025	January 31, 2025
Task 7: Model Training & Testing	February 3, 2025	February 7, 2025
Task 8: Model Accuracy Testing: Data Subset 1	February 10, 2025	February 11, 2025
Task 9: Model Validation Testing: Data Subset 2	February 11, 2025	February 12, 2025
Task 10: GUI Complete: Host Jupyter Notebook in Google Colab	February 13, 2025	February 14, 2025
Task 11: GUI Testing: Validate Data Subsets 1 & 2 in Google Colab	February 18, 2025	February 19, 2025
Task 12: End User Training	February 24, 2025	February 28, 2025
Task 13: Project Sign Off	March 3, 2025	March 4, 2025

Sprint Schedule

Sprint	Start	End	Tasks
1	January 6, 2025	January 17, 2025	1, 2, 3
2	January 20, 2025	January 31, 2025	4, 5, 6
3	February 3, 2025	February 14, 2025	7, 8, 9, 10
4	February 17, 2025	March 4, 2025	11, 12, 13

B.5. Resources and Costs

Category	Resource	Description	Cost
Personel	Lead ML Engineer	Develops & trains the logistic regression model.	$200/hour x 40 hrs x 8 weeks = $64,000
	Data Scientist	Preprocess data & collaborate with Lead ML Engineer	$200/hour x 40 hrs x 8 weeks = $64,000
	Web Developer	Integrate the model into the company website	$50/hour x 40 hours x 4 weeks = $8,000
Hardware	Computing Workstation	PC with a high-performance GPU for model training & maintenance.	$3,000
	Laptops	Personal laptops come standard to each engineer.	(2) x $0
Software	Google Colab	Cloud-based based environment for hosting Jupyter Notebook.	$10/month x 3 months = $30
	Python Libraries	Open-sourced computing software such as scikit-learn, pandas, and NumPy are used to develop the model.	$0
	Web Developer's Software	Open-sourced front-end software such as React/JS and accompanying libraries.	$0
	SSL Certificates	Certification of encrypted data transfer.	$100
	Version Control	GitHub for version control.	$0
	Existing Health Care Data	Access to existing health care database.	$0 (Data already owned by the company)

Third-Party Services	AWS Hosting	Cloud hosting includes the EC2 instances, S3 storage & security features.	$200/month x 3 months = $600
	AWS Cloud Security	Firewalls & other compliance tools for handling health care data.	$400
Miscellaneous	End User Training	Training sessions by the engineering team and web developer.	$0 (Included in respective personnel cells)
	Overbudget Fund	Additional funds if the project is over budget or over the anticipated timeline (10% of total budget).	$14,013
		Total	$154,143

B.6. Evaluation Criteria

Objective	Success Criteria
(Ease of Use)	Patients have an intuitive user interface to run the prediction; they do not need to manually enter their data.
(User error rate reduction)	The final system receives a user error rate of less than 5% with the help of automated entry of input fields.
(Algorithm Efficiency)	The logistic regression model is based on the order of minutes to train, not hours or days.
(Accuracy)	The model can predict the occurrence of a stroke on data with an accuracy of at least 85%.

(Throughput)	The model is trained with 3,750 patients' worth of data.
(Uptime)	The deployed model has an uptime of 99%.
(Financial)	The project stays under the total budget of $154,143, as described in section B.5.

C. Machine Learning Solution Design

C.1. Hypothesis

Integrating the logistic regression algorithm into True Health Care Provider’s system for stroke prediction occurrences will lead to accurate prediction and will supply information to patients to for reducing their risk.

Proposed Testing Method:

- Compile Data: Extract patient records from existing company database storage.
- Data Preprocessing: Clean data for missing values and format all input parameters numerically.
- Train Model: Feed 75% of the the cleaned data into the logistic regression model for training purposes.
- Test & Validate: Feed remaining 25% of the cleaned data into the model to test its prediction.
- Measure Outcome: Ensure algorithm is at least 85% accurate in its predictions for stroke occurrences on the testing data.

C.2. Selected Algorithm

The logistic regression algorithm, a subset of supervised learning,, will be used. This algorithm models the log odds of a binary event occurring as a linear combination of one or more independent variables. Since it is a binary event occurring, the event either happens or does not.

C.2.a Algorithm Justification

As backed by (Biswas et al.), the logistic regression algorithm remains an accurate model for predicting stroke occurrences. The algorithm's ability to leverage multiple independent input variables, as in the case of our various medical data obtained, proves it to be a sufficient choice for the machine learning model selection.

C.2.a.i. Algorithm Advantage

An advantage of using the logistic regression model is that the model has assignable coefficients that can be observed after training the model on data. These coefficients act as a weight, the greater the absolute value of these weights the more they influence a certain direction, good or bad. This makes the logistic regression algorithm extremely interpretible compared to other algorithms.

C.2.a.ii. Algorithm Limitation

A limitation of the logistic regression is that it only captures a linear relationship between input variables and the outcome, which can be seen by its coefficients. The algorithm therefore cannot capture non-linear relationships between input variables and the outcome like in a random forest algorithm.

C.3. Tools and Environment

The environment will feature a macOS Sequoia 15 operating system, run on the python language, and feature libraries such as scikit-learn, pandas, NumPy, matplotlib, and Seaborn. These libraries are standard practice for developing machine learning models using python where the operating system is a preference. At the minimum, a 6-core processor is recommended for calculating the model in a timely manner. The python code can then be hosted in a Google Colab which hosts a Jupyter Notebook of the code.

C.4. Performance Measurement

Training Phase: A training time of under 10 minutes for 75% of the 5,000 patients will be considered a success as this will mean the software has been optimized.

Accuracy Phase: An overall accuracy predicting stroke outcomes of at least 85% will be considered a success for the model.

Inclusive Phase: The model should be able to correctly predict both the likiliness of occurrence of strokes, but also the unlikiliness of occurrence of strokes.

Results & Correlation Phase: The outcome of the model should display weights for the input variables and be able to speak to the importance of each input variable.

D. Description of Data Sets

D.1. Data Source

The data source is an existing data set of over 5,000 anonymous patients with varying but same type of contributing parameters towards stroke occurrence. The data set can be found on Kaggle and was uploaded four years ago and titled “Stroke Prediction Dataset”.

D.2. Data Collection Method

The “Stroke Prediction Dataset” existing on Kaggle will be downloaded in comma separated variable (CSV) format. It can be loaded into a python file using the standard built in read function.

D.2.a.i. Data Collection Method Advantage

One advantage for data collection using python and a CSV data file is that the data file can easily be read into a data structure using python’s built in read function. The read function can either read the file stored in the same directory as the python script, or a github raw CSV link link can be passed to it.

D.2.a.ii. Data Collection Method Limitation

One disadvantage of the python and CSV data file technique is that large data files can be difficult to scan through for data preprocessing needs. To effectively preprocess the data, a number of commands will need to be run on individual columns to both determine the unique fields and update them.

D.3. Quality and Completeness of Data

The existing data has already had all identifiable information removed from it and only displays medical statitisics. Columns in our data set first had a unique method passed to them to determine what unique values exist in the cells. Entries found with Not A Number (NAN) in the “bmi” column and “Unknown” in the smoking_status will be dropped as it only represents a small sample of the data set. The patient “id” column can then be dropped as it is no longer needed. Then, numerical data can be scaled to match binary magnitude using its mean and standard deviation as the logistic regression model requires this due to its weighted coefficients. Text data such as gender or marital status will have a technique called “One-Hot Encoding” applied to it. This techniques slightly changes the name of the title so it’s values can become binary, such as male gender being logically true. All numerical data can then be entered into the logistic regression model.

D.4. Precautions for Sensitive Data

Handling sensitive data especially health care data will be treated with the highest degree of care. Role base access controls will utilized to ensure only the right people have access to the data. Existing laws such as HIPAA are also in place and can be enforced to both existing and past employees in the event of any data theft.

References

National Heart, Lung, and Blood Institute. “Causes of Stroke.” *National Heart, Lung, and Blood Institute*, https://www.nhlbi.nih.gov/health/stroke/causes. Accessed 24 Dec. 2024.

Zhang, Rui, et al. “The Impacts of Lifestyle and Genetic Risk on Stroke Incidence.” *PubMed Central*, vol. 15, no. 4, 2024, https://pmc.ncbi.nlm.nih.gov/articles/PMC10326666/. Accessed 24 Dec. 2024.

Smith, Jane, and John Doe. “New Approaches in Stroke Treatment.” *ScienceDirect*, vol. 45, no. 3, 2023, https://www.sciencedirect.com/science/article/pii/S2772442522000569. Accessed 24 Dec. 2024.

Brown, Michael J., et al. “The Role of Medical Data in Enhancing Stroke Prediction Models.” *PubMed Central*, vol. 18, no. 6, 2024, https://pmc.ncbi.nlm.nih.gov/articles/PMC10664112/#:~:text=Large%20amounts%20of%20medical%20data,more%20accurately%2C%20improving%20patient%20outcomes. Accessed 24 Dec. 2024.