September 7, 2024

C964: Computer Science Capstone

Jinhwan Kim 5838884

Task 2 parts A, B, C and D

Part A: Letter of Transmittal

9/6/2024

Dr. John Miller

Health Technologies

1234 Medical Plaza

Los Angeles, CA 90088

Dear Dr. John Miller,

I’m reaching out to bring to light an important issue for our business and request approval for our solution. The healthcare professional who are our customers need a way to improve stroke identification and prevention for their patients. As you know, stroke is a leading cause of disability and death around the world. By predicting stroke risk early, we can provide our clients better experience by improving their patient outcomes and lowering healthcare costs.

The solution I propose is a machine learning tool designed to predict likelihood of stroke in lifetime to identify high-risk individuals for stroke incidence based on individual health data, such as age, hypertension, and glucose levels. This tool would act as a decision support system for healthcare providers, allowing them to intervene earlier for patients who are at high risk. By analyzing key risk factors, the tool will support clinicians in making informed decisions and providing personalized recommendations for stroke prevention.

The benefits are that it would help healthcare professionals to more accurately predict stroke risks, quickly leading to preventive interventions. This could improve patient care, reduce hospitalizations, and optimize resource use within their healthcare system. Additionally, the application is designed with ease of use in mind, with visualizations and a simple interface that can seamlessly fit into existing workflows. Overall, we will attain higher client satisfaction scores and with our solution bringining financial profit and business stability.

The project’s cost primarily involves investment in developing, testing, and training the machine learning model, which will cost us $15,000 as a lump sum. We anticipate a timeline of 60 days for the main phases, including data collection, model training, and deployment. We will use anonymized, public data to ensure data privacy. With additional data we will comply with HIPAA to safeguard sensitive health information and maintain patient confidentiality.

With my 10 years of experience in machine learning and data analysis as senior software architect at Machine Learning Company, and 5 years of experience with Health Technologies, I am confident that I can deliver this project successfully. I'd be happy to discuss any further details. Thank you for your time and consideration, and I look forward to your approval to move forward with this important project.

Sincerely,

Jinhwan Kim

Jinhwan Kim

Senior Software Architect

Part B: Project Proposal Plan

Project Summary

Problem: Stroke is a major cause of disability and death. Without an early prediction tool available to them, healthcare providers cannot identify high-risk patients in time for early intervention, leading to clinicians' complaints and low satisfaction level on their partnership with us.

Client and Needs: Health Technologies needs a reliable machine learning tool that enables healthcare providers who are our clients to predict stroke risk to identify high-risk individuals using patient data, leading to timely interventions and improved patient care.

Deliverables: A functional machine learning application that predicts stroke risk based on health data. A user guide explaining how to use the application.

Summary and Justification: The application will provide healthcare providers with an easy-to-use tool for stroke risk assessment, leading to more effective patient management, reduced healthcare costs, and better resource allocation, which would continue their partnership with us and usage of our service which would benefit both people's health for good cause, and our business for financial profit.

Data Summary

Data Source: The data will be provided in a .csv file containing health factors such as age, hypertension status, glucose levels, and whether the anonymized individual was diagnosed as high-risk individual for stroke.

Data Processing: Data will be cleaned to remove erroneous entries, and handle incomplete row with missing information. Throughout the development life cycle, data will be processed for training and validation of the machine learning model. The custom-entered personally identifiable data itself will not be stored permanently, but the indication of high-risk will be output for the user.

Data Justification: The dataset includes critical health indicators that commonly relate to stroke risks, making it ideal for training a predictive model. Any outliers or missing data will be handled during preprocessing to ensure accuracy.

Ethical/Legal Concerns: We will use anonymized data which contains no personal identifiable info, ensuring there is no conflict with ehtical and legal concerns or data privacy standard such as HIPAA.

Implementation

Industry-standard methodology

The CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology will be used to guide this project. This standard process ensures a clear, structured approach to data mining and machine learning.

Business Understanding: The project aims to predict stroke risk, allowing healthcare providers to identify high-risk individuals for early intervention.

Data Understanding: The dataset includes key health factors such as age, hypertension, heart disease, and smoking status, which are essential for prediction.

Data Preparation: Handle missing values by imputing averages. Encode categorical data and scale all features for consistency. Address class imbalance by upsampling stroke cases.

Modeling: Implement a Logistic Regression model with polynomial features to capture interactions. Class weights are balanced to improve accuracy in predicting stroke cases.

Evaluation: Model performance is measured using accuracy and cross-validation metrics, ensuring reliable results.

Deployment: The final model is deployed in a simple Python application with input and prediction capabilities, plus visualizations for better insight.

Project Implementation Plan

Data Collection and Preparation: Gather the dataset, preprocess missing values, encode variables, and balance classes.

Feature Engineering: Scale features and add polynomial interactions.

Model Development: Train the Logistic Regression model on 80% of the data and evaluate it on the remaining 20%.

Model Evaluation: Measure performance using accuracy and cross-validation scores.

Application Development: Develop a Python application for stroke risk prediction with user input and visualizations.

Timeline

Milestone or deliverable	Duration (hours or days)	Projected start date	Anticipated end date
Data Collection & Cleaning	2 weeks	9/15/24	9/30/24
Model Development	3 weeks	10/1/24	10/21/24
Testing	1 week	10/22/24	10/29/24
Deployment	1 week	10/30/24	11/6/24

Evaluation Plan

Verification Methods: The model will be verified at each stage using cross-validation techniques on training and testing datasets.

Validation Method: Upon completion, the model's performance will be evaluated using metrics such as accuracy, precision, and recall to ensure it meets project goals.

Resources and Costs

Hardware/Software Costs

Developed at the normal office setting at the headquarter: No cost

Python environment: No cost

Libraries (Scikit-learn, Pandas, Matplotlib): No cost

Existing computers and laptops in the office: No cost

Labor Costs

A developer bonus rate of $50/hour:

120 hours x $50/hour = $6,000.

Environment Costs

The software will be a standalone application: No cost

Developed by the project and developer team during normal business hour: No cost

Total Cost

The total cost for this project will be $6,000.

Part C: Application

Please see the submitted application.

Part D: Post-implementation Report

Solution Summary

The problem addressed by this project is the early identification of individuals at high risk of stroke. Stroke is one of the leading causes of death and disability worldwide, and early detection of stroke risk factors can greatly improve patient outcomes. The primary objective of this project was to create a machine learning-based solution that could predict stroke risk based on individual health data, such as age, hypertension, heart disease, glucose level, BMI, and smoking status.

The solution was implemented as a standalone application with a command-line interface that allows healthcare professionals to input patient data and receive stroke risk predictions. The application applies a logistic regression model that predicts whether a person is at high or low risk for stroke. The solution also includes visualizations to help healthcare providers better understand the key factors contributing to stroke risk.

Data Summary

The raw dataset used in this project was sourced from a public stroke prediction dataset. It contains 12 columns, including gender, age, hypertension, heart disease, glucose levels, BMI, smoking status, and stroke outcome (either 0 or 1, representing no stroke and stroke, respectively).

Handling missing values: Missing data in features such as BMI and glucose levels were imputed with mean values.

Encoding categorical variables: Features such as gender and smoking status were encoded into numeric values using label encoding.

Feature scaling: Continuous variables such as glucose level, BMI, and age were standardized using the StandardScaler function to ensure that they are on the same scale.

Data balancing: Since the dataset was imbalanced (more individuals were labeled as low-risk for stroke), we applied oversampling to the minority class (high-risk individuals) to ensure the machine learning model received balanced input for better learning.

Throughout the design, development, and testing stages, the data was split into training and testing sets to evaluate the performance of the model. Additional features were created using polynomial transformation to capture interactions between variables and enhance model accuracy.

Machine Learning

What method was used?

The main method employed in this project is Logistic Regression. Logistic regression is a widely used supervised learning algorithm that models the relationship between a set of independent variables and a binary dependent variable, which in this case is stroke risk (0 = low risk, 1 = high risk).

How was it developed?

The logistic regression model was developed using Python's sklearn library. The model was trained on a balanced dataset using the training set, and the testing set was used to validate the results. Polynomial features were added to capture interaction effects between the input variables (e.g., age and hypertension). The model also used class weight balancing to handle the imbalanced dataset.

Why was this method selected?

Logistic regression was chosen due to its simplicity, interpretability, and effectiveness in binary classification problems. The goal of predicting whether an individual is at high or low risk for stroke fits well into the binary classification paradigm, and logistic regression has the added benefit of providing probabilistic predictions, which can be useful in healthcare scenarios for risk assessment.

Validation

Validation method

The validation method used in this project was cross-validation. Cross-validation was performed using 5-folds, ensuring that the model's performance was evaluated on multiple training and testing splits. The final model's accuracy was evaluated based on these folds, and the mean accuracy score was reported.

Results of the validation method

The logistic regression model achieved an accuracy of 94%, with cross-validation showing an average accuracy of 92%. These results indicate that the model is reliable in predicting stroke risk, although further improvements could be made by experimenting with different algorithms or fine-tuning parameters.

Visualizations

Three visualizations can be viewed directly in the application.

Proportion of Stroke Risk (Pie Chart): A pie chart that visualizes the overall proportion of individuals at high risk and low risk for stroke in the dataset.

Feature Importance Plot: A horizontal bar chart that displays the coefficient values of each feature, showing how much each factor contributes to stroke risk prediction. For example, age and hypertension were found to be significant predictors, while gender had a minimal negative impact.

Boxplot of Average Glucose Levels by Stroke Risk: This boxplot illustrates the distribution of glucose levels among high-risk and low-risk individuals. It shows the difference in glucose levels between the two groups, highlighting how glucose level plays a role in stroke risk.

User Guide

Steps to execute and use the application

1. Download and Install Python

Ensure you have Python 3.x installed on your machine. You can download Python from the official site: https://www.python.org/downloads/

2. Install Required Libraries

Install the required Python libraries using pip. Run the following commands in your terminal or command prompt:

```
pip install numpy pandas matplotlib scikit-learn
```

3. Run the Application

Ensure that the dataset file stroke.csv is in the same folder as the main.py file.

To run the application, execute the main.py file from the command line:

```
python main.py
```

4. Predict Stroke Risk

The application will prompt you to enter values for features like gender, age, hypertension, heart disease, glucose level, BMI, and smoking status. Once you provide the inputs, the model will calculate and display the predicted risk of stroke and its probability.

5. View Visualizations

After making predictions, you can view the three visualizations by selecting the "Show Visualizations" option in the application.

Example Usage

Select option 1 to predict stroke risk by entering the requested health data.

Select option 2 to view visualizations.

Select option 3 to exit the program.

Then enter the inputs as instructed – see below example:

```
1. Predict Stroke Risk
2. Show Visualizations
3. Exit
Enter choice (1/2/3): 1

Enter the following details to predict stroke risk:
Gender (0 = Male, 1 = Female): 1
Age (0-120): 65
Hypertension (0 = No, 1 = Yes): 1
Heart Disease (0 = No, 1 = Yes): 0
Average Glucose Level (50-300): 175
BMI (10-100): 32
Smoking Status (0 = never smoked, 1 = formerly smoked, 2 = smokes): 0

Prediction: High Risk
Probability of Stroke: 61.68%
```

References

Scikit-learn developers. (2024). Scikit-learn: *Machine Learning in Python*. Retrieved from https://scikit-learn.org/

Python Software Foundation. (2024). *Python Language Reference*, *version 3.12*.5. Retrieved from https://www.python.org/

McKinney, W. (2010). *Data structures for statistical computing in Python.* Proceedings of the 9th Python in Science Conference, 51-56. Retrieved from https://pandas.pydata.org/

Hunter, J. D. (2007). *Matplotlib: A 2D graphics environment*. Computing in Science & Engineering, 9(3), 90-95. Retrieved from https://matplotlib.org/

Le, J. (2016). *The 10 Algorithms Machine Learning Engineers Need to Know*. KDNuggets. Retrieved from https://www.kdnuggets.com/2016/08/10-algorithms-machine-learning-engineers.html