WGU C964

Task 2

MACHINE LEARNING PROJECT Business Vision

Alexander Brown

Student ID # 005623532

Date: 10-28-2023

Table of Contents

A. Project Overview....................................................................................................................2 A.1. Organizational Need ......................................................................................................... 2 A.2. Context and Background..................................................................................................... 2 A.3. Solution Summary............................................................................................................. 2 A.4. Machine Learning Benefits................................................................................................ 3 B. Machine Learning Project Design......................................................................................... 3 B.1. Scope ................................................................................................................................. 3 B.2. Goals, Objectives, and Deliverables ................................................................................. 3 B.3. Standard Methodology ..................................................................................................... 4 B.4. Projected Timeline ............................................................................................................. 5 B.5. Resources and Costs........................................................................................................... 7 B.6. Evaluation Criteria ............................................................................................................ 7 C. Machine Learning Solution Design....................................................................................... 8 C.1. Hypothesis.......................................................................................................................... 8 C.2. Selected Algorithm............................................................................................................. 8 C.2.a Algorithm Justification..................................................................................................... 8 C.2.a.i. Algorithm Advantage ................................................................................................... 9 C.2.a.ii. Algorithm Limitation ..................................................................................................9

C.3. Tools and Environment ...................................................................................................... 9 C.4. Performance Measurement................................................................................................. 9 D. Description of Data Sets ....................................................................................................... 9 D.1. Data Source ....................................................................................................................... 9 D.2. Data Collection Method...................................................................................................... 10 D.2.a.i. Data Collection Method Advantage............................................................................... 10 D.2.a.ii. Data Collection Method Limitation ............................................................................. 10 D.3. Data Normalization............................................................................................................. 10 D.4. Data Security....................................................................................................................... 10 References................................................................................................................................10

**A. Project Overview**

This project is to create a dashboard data product based on collected patient data that will allow providers to analyze and interpret risk of heart attack based on analysis of the data using Machine Learning Techniques.

**A.1. Organizational Need**

The organizational need we are targeting is the organization and interpretation of patient data and its use in practice.

**A.2. Context and Background**

The cardiology team at Heart Hospital have been looking for a solution that will allow them to analyze and interpret data about patients and their heart attack risk. A database has already been procured from the hospital and will need to be preprocessed, analyzed, and transformed into a usable interface that will allow providers to better educate patients on heart risks.

**A.3. Solution Summary**

The machine learning solution will be developed by the CRISP-DM process model, and will analyze and produce a ML model capable of predicting risk of heart attacks by calculating the percentage of patients with similar traits who have had previous heart attacks. Additionally, we will produce an interface capable of showing raw correlation between each feature and previous heart attacks.

**A.4. Machine Learning Benefits**

Machine Learning will allow us to create a model with high probability of predicting heart attack risk, which will provide associates of cardiology to analyze patient records and produce relevant feedback for patient health.

**B. Machine Learning Project Design**

**B.1. Scope**

**In Scope:**

1. We will analyze and preprocess the provided database for relevant features and perform research to identify useful classifications for each.

2. We will create and train a machine learning model capable of predicting relationships between all features and the target feature, heart attack risk, which will be identified by a 1 if a previous heart attack exists, or 0 if not.

3. We will create a dashboard interface that will provide graphical aids when displaying the raw probability of a feature relating to heart attack risk, and textual outputs for predicting the probability of a newly added patient record having a previous heart attack event.

**Out of Scope:**

1. We will NOT be creating a long-term solution for future models.

**B.2. Goals, Objectives, and Deliverables**

The following goals, objectives, and deliverables were based on the project flow discussed by Advatma A. (2020):

**Goals**

•Our goal is to help Heart Hospital discover patterns in their patient data and give them an easy way access and interpret that data.

**Objectives**

•We have two objectives. First, we must determine which variables are most useful for modeling and design a Machine Learning tool around them. Second, we must create a deployable model based on the discoveries of the Machine Learning tool that can be used by Heart Hospital.

**Deliverables**

* A working model and dashboard that will allow the end user to explore relationships between each feature and the target feature.
* The dashboard/interface will also include visual outputs to help visualize data.
* An interface allowing the input of patient record and output of a calculated heart risk, assuming they have not yet had a heart attack.

**B.3. Standard Methodology**

Development will follow the CRISP-DM methodology. CRISP-DM follows the following 6 stages discussed in the article by Hotz N. (2023).

• **Business Understanding:** In this phase, we examine the business needs. This phase includes assessing risk/reward, business objectives, data mining, and project planning.

• **Data Understanding:** After business understanding is completed, we will transition to a focus on collecting, analyzing, exploring, and verifying the quality of the data.

• **Data Preparation:** Once data has been analyzed, it is up to us to decide which parts are necessary for the task at hand, clean the data, drive new attributes where needed, and perform any data transformations.

• **Modeling:** In the modeling phase, we will select our modeling techniques, generate tests to validate our model accuracy, build the model, and assess the model’s capability.

• **Evaluation:** After the technical assessment is completed during the modeling phase, a whole-sale evaluation of the model and how it contributes to solving the business problem at hand will be conducted.

• **Deployment:** Once all evaluations have passed and a model is selected, the final phase is to package the model in a way that is accessible to the client, which includes planning the project and executing it.

**B.4. Projected Timeline**

**12-01-2023** – Work with the company to discuss its objectives and clarify any ambiguities. Create a proposal and have it evaluated and accepted by relevant stakeholders. This reflects the business understanding phase.

**12-07-2023** – The proposal has been accepted, overarching goals of the business are well understood, and the data has been received by the team to begin reviewing.

**12-14-2023** – The team has reviewed the data, verified its quality, and described it in a way that will fit the model.

**12-28-2023**- The data has been cleaned and prepared. Data transformations have been completed and unnecessary data is removed from the model.

**01-19-2024** – The model has been planned and produced, along with tests to verify its technical accuracy.

**01-26-2024** – The model has been evaluated for its consistency with meeting business objectives and has been accepted.

**02-02-2024** – The delivery of an accessible model form has been planned, produced, tested, and is ready for deployment. Deployment is successful and project is accepted, completing the project.

**Sprint Schedule**

|  |  |  |  |
| --- | --- | --- | --- |
| **Sprint** | **Start** | **End** | **Tasks** |
| 1 | 11-01-2023 | 11-07-2023 | Work with the company to clarify objectives for the project. Completing the Business Understanding Phase. |
| 2 | 11-07-2023 | 11-14-2023 | Receive and review data, completing the Data Understanding phase. |
| 3 | 11-14-2023 | 11-28-2023 | Clean and prepare data, including any data transformations, completing the Data Preparation Phase. |
| 4 | 11-28-2023 | 12-19-2023 | Plan and produce the model which will be used to analyze the data, in addition to testing to verify the technical accuracy of the model, completing the modeling phase. |
| 5 | 12-19-2023 | 12-26-2023 | Evaluate the model against business objectives. Have it signed off on and accepted, completing the evaluation phase. |
| 6 | 12-26-2023 | 01-02-2024 | Create the finished model to be interacted with by the end user, completing the delivery phase and thus the project. |

**B.5. Resources and Costs**

|  |  |  |
| --- | --- | --- |
| **Resource** | **Description** | **Cost** |
| Machine Learning Engineer | $2844.38 per week based on Indeed (n.d). 9 weeks x $2844.38 = 25,599.42 | $25,599.42 |
| Hardware | Laptop provided by Heart Hospital already in stock. | $0 |
| Office | Assume that Data Analyst is working from home. | $0 |
| Software | PyCharm Professional Edition, Security software for the laptop. $19.99 for Norton Antivirus according to Norton (n.d.) + 2 months of Pycharm Professional for $49.80, based on numbers from Jetbrains (n.d.) | $69.79 |
|  | **Total** | $25,669.21 |

**B.6. Evaluation Criteria**

Describe the criteria used to evaluate and measure the success of the completed project.

|  |  |
| --- | --- |
| **Objective** | **Success Criteria** |
| (Ease of Use) | The delivered model does is useable by the people who need it, and therefore does not require an in-depth knowledge of the technology behind it. |
| (User error rate reduction) | The model is equipped with sufficient supporting material that new users can operate it without serious error with 2 weeks of training. |
| (Algorithm Efficiency) | The algorithm test accurately predicts whether a new record will have had a previous heart attack, based on samples taken from the set. This predictor of efficiency was based on findings from Saini A. (2023) |

**C. Machine Learning Solution Design**

**C.1. Hypothesis**

My hypothesis is that some factors will have a higher effect on the probability of a previous heart attack, and therefore we will be able to use Machine Learning techniques to determine risk factors for patients.

**C.2. Selected Algorithm**

The algorithm I chose is Logistical Regression.

**C.2.a Algorithm Justification**

According to a cheat sheet designed by Cotton R. (2022), Logistic Regression uses a linear model relationship between inputs and a categorical output (0 or 1). The example use cases given from Cotton R. (2022) include prediction of credit risk and customer churn prediction. I can see some similarities between prediction of credit risk and heart attack risk. In the way that I am predicting risk, as either a 1 if a person is likely to have had a heart attack given their background, or 0 if not, it is easy to relate the credit risk in this same way. For example, if we were to look at factors influencing credit risk, we would be able to find accounts which had already experienced a credit incident, and then find other accounts with similar attributes that would theoretically have a risk of credit incident as well, even if an incident had not yet occurred.

**C.2.a.i. Algorithm Advantage**

According to Cotton R. (2022), the Logistic Regression model is easily interpreted and explained, is less prone to overfitting when using regularization, and is applicable for multi-class predictions.

**C.2.a.ii. Algorithm Limitation**

In the cheat sheet provided by Cotton R. (2022), disadvantages of the Logistic Regression include the fact that it assumes a linear relationship between inputs and outputs, and can still be prone to overfitting with small, high-dimensional data.

**C.3. Tools and Environment**

We will be following the steps implemented in the tutorial by Smith B. (n.d.). In his tutorial, uses the Python programming language within a Jupyter Notebook, with several third party languages including pyplot, matplot to handle our graphing and modelling needs, along with numpy and pandas for data handling. I will be using Windows 11 on a laptop with the following specs:

Device name MSI

Processor 12th Gen Intel(R) Core(TM) i7-12700H 2.70 GHz

Installed RAM 16.0 GB (15.7 GB usable)

Device ID 7C89F5E2-5F09-4416-BDB9-76A140BFA6F5

Product ID 00342-20826-24564-AAOEM

System type 64-bit operating system, x64-based processor

Pen and touch No pen or touch input is available for this display

**C.4. Performance Measurement**

To measure performance, we will be following Smith B. (n.d.) in his exploration of an MLP neural network and instead using the model selection method RandomizedSearch CV to produce a list of optimized hyperparameters, which we will measure against default hyperparameters.

**D. Description of Data Sets**

**D.1. Data Source**

The data is sourced from the hospital, collected into a single excel file, as stated in Kaggle (n.d.).

**D.2. Data Collection Method**

The data collection is handled internally from the hospital itself, as stated in the work from Kaggle (n.d.).

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

The company has collected the data itself, meaning that it will be more reliable and will not cost anything more than has already been spent to collect it, according to Indeed (2023)

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

A potential limitation of internally collected data is that it will likely be from a smaller pool of data, which can reduce the effectiveness of long-term strategies, according to Lido (n.d).

**D.3. Quality and Completeness of Data**

Based on my previous experience with data cleaning, the data will be checked for accuracy and completeness, and any non-important variables will be removed. To check for completeness, any record without a complete set of fields will be removed. For accuracy, data fields will be checked to be of the appropriate type. Fields such as age and heart rate should only be positive integers, while non-integer and non-float types will be checked that they are non-null and that the entries fit the category. Sex, for example, will be checked to ensure that male and female are the only two classes in the entry.

**D.4. Precautions for Sensitive Data**

According to Hypervault (2023), best practices for handling sensitive data include end-to-end encryption for communications, access controls for members, encrypted data storage, secure file sharing, and audit trails for monitoring.

**References**

Hotz, N. (2023, January 19) *What is CRISP DM?* [https://www.datascience-pm.com/crisp-dm-2/](about:blank)

Kaggle (n.d.) *Heart Attack Risk Prediction Dataset* https://www.kaggle.com/datasets/iamsouravbanerjee/heart-attack-prediction-dataset

Adyatma A. (2020, November 10) *Case Study: Predictive Lead Scoring.* [https://rpubs.com/Argaadya/crispr\_dm](about:blank)

Indeed(n.d.) *Machine Learning Engineer in Ohio* https://www.indeed.com/career/machine-learning-engineer/salaries/OH

Hypervault. (2023, October 4) Best Practices for Secure Exchange of Sensitive Data with Clients for Web Agencies [https://hypervault.com/exchange-of-sensitive-data-agencies-best-practices/](about:blank)

Cotton R. (2022, April) *Machine Learning Cheat Sheet*. [https://www.datacamp.com/cheat-sheet/machine-learning-cheat-sheet](about:blank)

Smith B. (n.d.). Playlists[Bevan Smith 2]. Youtube. Retrieved October 26, 2023 from <https://www.youtube.com/watch?v=S7_xqXdh7UE&list=PLuVnJWOCzbxVfAVMU6bIFJ2jyzhEAnOCZ&pp=iAQB>

Norton (n.d) *Norton Antivirus Award-winning Protection for only $19.99/1st Year.* [https://us.norton.com/store?expid=PS-BRANDED-NAVP&promocode=BPSOFFER&nortoncountry=us&om\_sem\_cid=hho\_sem\_sy:~en-us\_nor\_nav\_sch\_brn\_exc\_nfr\_adw\_dtp\_low:nav\_top~c\_kw0000001532&gad=1&gclid=CjwKCAjw-KipBhBtEiwAWjgwrO7YhX9xen5HIFbI8IwM9nR4Y70dqDCkAqvtUdDCfDwkIsjIe0WYNhoCBMwQAvD\_BwE&gclsrc=aw.ds](about:blank)

Jetbrains (n.d.) *Subscription Options and Pricing.* [https://www.jetbrains.com/pycharm/buy/#commercial?billing=monthly](about:blank#commercial?billing=monthly)

Saini A. (2023, August 2) *Gradient Boosting Algorithm: A Complete Guide for Beginners.*

[https://www.analyticsvidhya.com/blog/2021/09/gradient-boosting-algorithm-a-complete-guide-for-beginners/](about:blank)

Indeed (2023, March 9) *Advantages of Internal Data (with definition and examples).* [*https://uk.indeed.com/career-advice/career-development/advantages-of-internal-data*](about:blank)

Lido (n.d.) *Managing Internal and External Data to Optimize Your Business.* sainhttps://www.lido.app/post/internal-and-external-