

The Research of Factors of Hypertension

Yipeng Hu^{1,*,†}, Shaodan Wang^{2,†}

¹ Academie Stececile International School, Windsor, N9G1V8, Canada

² Pennon Education, Qingdao, 266000, China

* Corresponding Author Email: 2018y_hu@stececile.ca

† The authors contributed equally.

Abstract. The aim was to predict which group of people would have blood pressure problems. A random sample of 1000 people were used to investigate hypertension problems according to logistic regression. This project is to use logistic regression prediction technology to predict and analyze which group of people will have blood pressure problems by analyzing historical data such as age and weight. The independent variables are age and weight, and the dependent variables are 0 or 1 whether you have a blood pressure problem. Based on the Kaggle data set provided by a medical insurance company and the logistic regression model, we established an instance classifier of object logistic regression for calculation, and finally obtained the weight of independent variables, to roughly understand which factors are risk factors of hypertension. This weight can also be used to predict a person's likelihood of hypertension based on risk factors. We show that older and younger adults with high body weight are more likely to develop high blood pressure than the rest of the population. Excess body weight was strongly correlated with hypertension.

Keywords: Hypertension; Kaggle; variables.

1. Introduction

Since health problem has become a widespread and comprehensive topic worldwide, high blood pressure caused a large amount of death which is needed to be highly paid attention to. According to the statistical data, in 2019, 3.3 billion people in China has diseases correlated to heart or circulatory systems. However, for every single individual, different body quality needs to be taken into concern. For instance, compare the teenagers and the old, the probability they suffer from high blood pressure is influenced by many factors. If a teenager smoke and an old keeps exercising every day regularly, then the system of circulatory of the teenager is more likely to be destroyed. It is all known that high blood pressure is a severe health condition in which the blood isn't flowing the way it should be. The body depends on the heart's pumping and other functions to push blood circulation, deliver oxygen, and provide energy to all of its cells. High blood pressure contains two types, one is essential hypertension, and the other is secondary hypertension. Though in view of essential hypertension has something to do with heredity, which means that it will be difficult to find the origin that causes the disease. This paper will focus on both of them, do the classification and make categories, then check the accuracy. In this study, we show the prediction of the probability that a person might attack by heart disease, vice versa.

Since the application of the topic of high blood pressure is broad, many scholars had done various types of research in the biochemistry aspect. Buyukozturk et al. did researches on the evaluation of risk factors that lead to high blood pressure problems [1]. S.S. Deschênes et al. focused on the risk factors that caused high blood pressure [2]. EF Nsien and AO Abam's use the logistic regression model to model risk factors of high blood pressure in women [3]. What's more, Mehdi B et al.'s paper about mathematical methodologies' adhibition to predict medical diseases [4]. Ellis, H. P. et al. introduced methodologies to build up a mathematical to predict disease, which may be of great future benefits to predict the nature of the residual disease [5]. Satyanarayana, N. et al. considered age, anger and anxiety as variables to predict high blood problems [6]. JH Lin, PJ Haug applied machine learning to data of medical problems, did the classification, and studied whether "missingness" of clinical data can provide sufficient evidence to building models [7]. C El-Hajj, PA Kyriacou's review of machine

learning techniques in photoplethysmography for the measurement of blood pressure opened up a new future to monitoring hypertension [8]. B Farran, et al. also used machine learning algorithms and validation to access risks of hypertension [9]. FU Chuan-Xi et al. considered classification tree and logistic regression. They found the classification tree is better understood compared to logistic regression models, for its easier and clearer results [10].

However, there are various research to read and learn, but some innovative methods can be found, though. Previous scholars used simplex mathematical models, which makes the result not so accurate. In this paper, we first use a logistic regression model to study “the probability that a person gets high blood pressure” problem and consider different indicators as different variables. In this paper, height, weight, any transplants, any chronic diseases, and blood pressure problems are the main factors considered to predict the probability a person might catch high blood pressure. By referencing, we use logistic regression prediction techniques in machine learning to predict whether people will have blood pressure problems by analysis the historical data including age and weight [11]. By classification, the result will show in 1 or 0 which gives a result whether to have the problem or not. The data set is from Kaggle.com, a medical insurance company provided, and the data set is about almost 1000 customers. For the abstraction and complexity of this topic, we choose to create mathematical models.

This paper has several contributions as follows: we combine machine learning and logistic regression. The research is processed in 8 steps: import dataset, choose dependent and independent variables, feature scaling, fitting logistic regression, making a prediction, making a confusion matrix and last visualize the testing result. Using the method, we can find the category of people that are easier to get hypertension. This research will first put forward a proposal to help more people take care of themselves, which provides some useful suggestions to those people with similar experimental indexes. As a result, the rate of occurring hypertension can be significantly decreased. Last but not least, we explore new ways to solve classification problems as shown.

The remainder of this paper is organized as follows. Section 2, a detailed research method including how we code, how we make the calculation and how we apply formulae of logistic models or how machine learning helped us to finish the research. In Section 3, we show the discussion, including the hypothesis we made and experimental results, then compared with practical data, which helps to estimate accuracy. However, errors cannot be ignored, but it is important to decrease the percentage error by repeating calculations or using mean values or changing a formula. In Section 4, we will show the results, correct and improve our methods. [12].

2. Method

2.1. Methodology

The main methods have five steps which are training data, scaling data, applying logistic regression, making prediction, and determining accuracy. The training data is necessary to make the dataset able to fit into the python algorithm. Scaling data can do to optimize the x variable. Applying logistic regression onto the dataset is the key step which will build predicting model. Making prediction is the step to generate a result. And determining accuracy would test how accurate the result is.

The materials used are including a medical insurance dataset, called Medicalpremium, from the Kaggle. The dataset was measured by an insurance company already, which includes 1000 of their customers. The reason to choose this dataset was this dataset contains thousands of data which provided varieties, and this dataset includes a clear age, weight, weather has the problem to each person, which can give a good historical information that to be used for future prediction. The analysis tool is the Python language in Kaggle website.

The limits of the research include the accuracy of the prediction model cannot be 100 percent accuracy. In x variable, both age and weight are considered as x variable, but they usually have a

huge different. Although the step of feature scaling was made, the accuracy of data is still not 100 percent accuracy.

Making logistic regression prediction model by coding in Kaggle is the main method to implement our goal. The logistic regression probability curve formula is:

P is the probability of a 1 (the proportion of 1s, the mean of Y), e is the natural logarithm's base (about 2.718), and a and b are the model's parameters. When X is 0, the value of a yields P, and b controls how rapidly the probability changes when X is changed by a single unit (we can have standardized and unstandardized b weights in logistic regression, just as in ordinary linear regression). Because the relationship between X and P is nonlinear, b in this model does not have the same obvious interpretation as it does in linear regression. The dependent variable in logistic regression is a logit, which is the natural log of the odds.

[1] The sigmoid function of $f(x)$ is the logistic regression function $p(x)$: $p(x) = 1 / (1 + \exp(f(x)))$. As a result, it is frequently close to 0 or 1. The function $p(x)$ is frequently interpreted as the predicted probability that the output for a given x equals 1. As a result, $1 - p(x)$ is the probability that the output is 0.

Logistic regression determines the best predicted weights b_0 , b_1 and so on to b_r , such that the probability function, $p(x)$, is as close to all actual responses y_i , $i = 1$ to 'n', as possible, where n is the number of observations. Model training or fitting refers to the process of calculating the best weights based on available observations. To obtain the best weights, the log-likelihood function (LLF) is usually maximized for all observations $I = 1, \dots, n$. [2] This method is known as maximum likelihood estimation, and it is represented by the equation $LLF = \sum_i (y_i \log(p(x_i)) + (1 - y_i) \log(1 - p(x_i)))$.

Then the coding part of the research is started. First is training data. We can't fit and assess machine learning algorithms on raw data; instead, we must alter the data to satisfy the needs of the particular algorithms. Furthermore, in order to acquire the highest performance given our available resources on a predictive modelling project, we must choose a data representation that optimally exposes the unknown underlying structure of the prediction problem to the learning algorithms.

The second step is scaling data. Feature scaling is a Data Pre-Processing phase that is applied to data's independent variables or features. It essentially aids in the normalization of data within a specific range. It can also aid in the speeding up of calculations in an algorithm. The scaling method in the program is used for improving the data efficiency, because as x variable, age and weight has a huge difference. The package used is sklearn. Preprocessing and it is imported from sklearn. `Preprocessing import Standard Scaler`. SKlearn (scikit-learn) is a Python-based machine learning package. It includes support vector machines, random forests, gradient boosting, k-means, and DBSCAN, among other classification, regression, and clustering techniques, and is designed to work with the Python numerical and scientific libraries NumPy and SciPy. The formula of the backend is:

The values are replaced by their Z scores as a result of standardization. The x is a certain point. The $\bar{x}$ is the average number. The σ means the standard deviation. A z score means how far a data point from the mean. A positive Z score mean x is above the mean and a negative z score mean a data is below the mean.

The third step is to apply the logistic regression model onto the data list. [3] The goal of this part is fitting logistic regression into classifier. Logistic regression is a linear classifier, so you'll use the linear function $f(x) = b_0 + b_1x_1 + \dots + b_r x_r$, also known as the logit. The variables $b_0, b_1, \dots, b_r$ are estimators of the regression coefficients, also known as predicted weights or just coefficients. In python, `import LogisticRegression, from sklearn. linear_model`, then fit logistic regression model into the classifier:

The fourth step is doing the prediction for the modelling. In a multivariable logistic regression prediction, both axes represent the inputs, and the results are also showing in different color. One color meant that it is classified as 0, and another color meant that it is classified as 1. Logistic regression would find the weights b_0, b_1 , and b_2 , which are the estimators of the regression coefficients, that could maximize the log-likelihood function (LLF). [4] When b_0, b_1 , and b_2 are calculated, the

$\text{logit } f(x_1, x_2) = b_0 + b_1x_1 + b_2x_2$ and the probabilities $p(x_1, x_2) = 1 / (1 + \exp(-f(x_1, x_2)))$ can be derived.

The last step is to determine accuracy of the result plotting. To know accuracy would help in knowing how much we can believe the predicting result. We use python in Kaggle to read the historical dataset and select age and weight as the x variable. Then split our dataset into training data and test data. Training data is used to train our logistic model, and test data is used to validate our model. We then use SKlearn to split our data. Then before we fit logistic regression, we need to scaling the data to get a better accuracy between 0 and 1. Because age and weight in X variable has a huge difference. Logistic regression is fitted now by using sklearn. linear_model. Then make an instance classifier of the object logistic regression and give random_state =0 that could get the same result every time. A classifier is built now. Next step is to test accuracy of our model. Accuracy equals correct predictions / total_predictions:

Since the accuracy is good, then is to use plotting code to visualize the training and testing result.

2.2. Data

The dataset is downloaded from Kaggle dataset, a medical insurance company provided, and the dataset is about almost 1000 their customers. The data shows age, weight, height, whether the customers have some healthy problem. In data list, 1 means yes and 0 means no. In the below, there is only showing 30 sampling data from the real list.

Table 1. Data table of age vs. Weight vs. whether has blood pressure problem (30 out of 1000 people).

Age (years old)	Weight (Kg)	Has or hasn't problem
45	57	0
60	73	0
36	59	1
52	93	1
38	88	0
30	69	0
33	54	0
23	79	0
48	74	0
38	93	0
60	74	1
66	67	0
24	57	0
46	97	1
18	76	0
38	68	0
57	67	0
21	82	0
49	61	0
20	97	1
35	97	0
35	81	0
53	92	0
31	83	1
22	97	1
60	88	0
30	73	0
33	58	1
22	96	0
26	91	1

3. Results

In this study, body weight and age were selected for conditional logistic regression model. In order to ensure the accuracy of the experiment, we used the training data to verify the accuracy of the model before the experiment and check whether the final results were consistent with the test results. According to the chart, red areas and 0 indicate no blood pressure problems, and green areas and 1 indicate blood pressure problems. The results showed that people who were overweight were more likely to have high blood pressure.

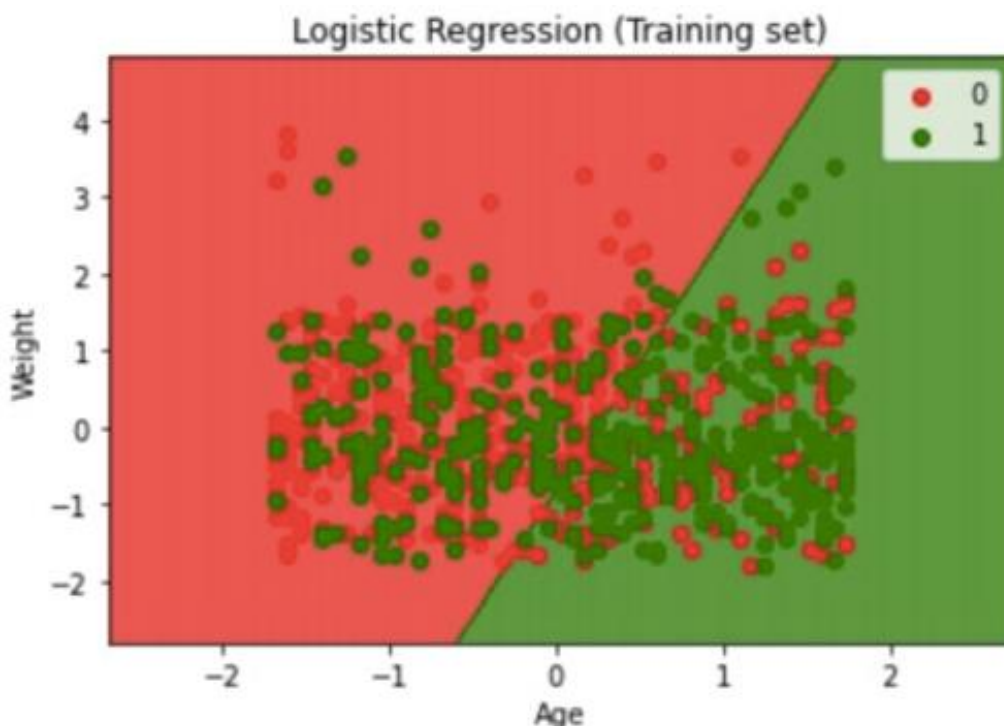


Fig 1. Logistic regression (training set) weight, age vs. whether has blood pressure problem graph.

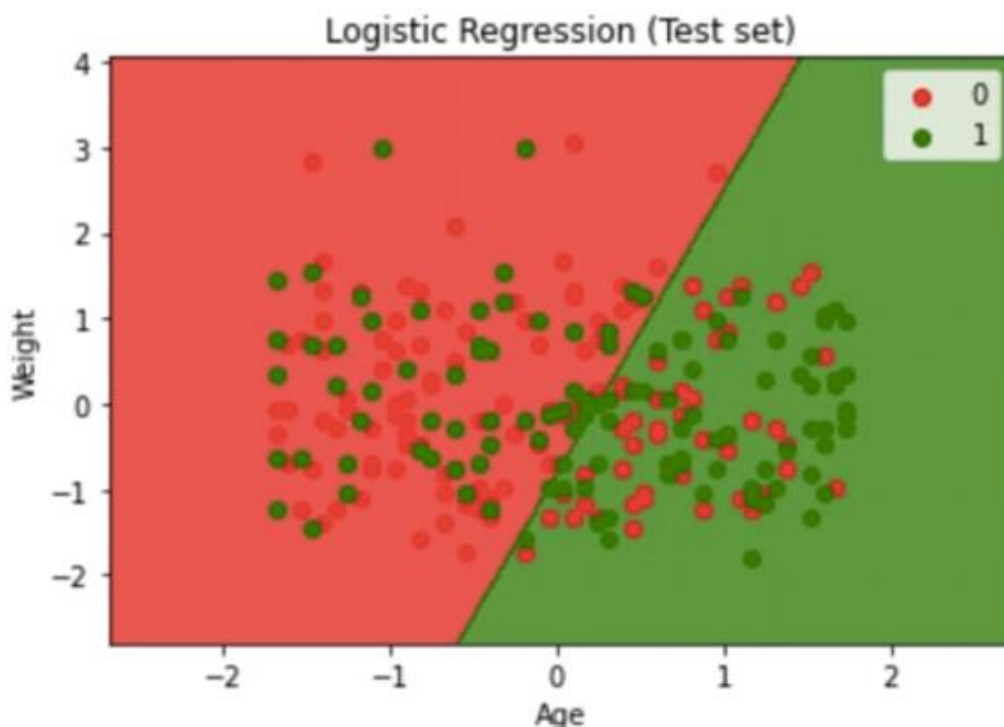


Fig 2. Logistic regression (test set) weight, age vs. whether has blood pressure problem graph.

Based on the train data to fit python, scaling data to increase accuracy, applying logistics regression to classifier, making prediction, and the determined accuracy method, we found that red dots are mainly concentrated in the high-weight elderly, low-age and high-weight parts. This means that older and younger people with higher weights are more likely to have high blood pressure than other groups. This is consistent with most of the real-life factual findings we found before the experiment, as well as the scientifically rigorous thinking that low age and weight, old age and low weight are more likely to develop high blood pressure. Some scholars believe that the biological mechanism of high blood pressure may be related to higher leptin production in overweight and obese people. Leptin stimulates renal angiotensin and sympathetic nervous system by altering the sodium passages in the kidneys, as well as insulin and pro-inflammatory factors. Interactions and other pathways further lead to high blood pressure. In addition, overweight and obese people may often have other high-risk factors that affect the onset of high blood pressure, such as excessive intake of unhealthy foods, too little physical activity, and abnormal blood lipids. People with high blood pressure may not have high blood pressure, but may also have high blood pressure due to cold stimulation, stress, vigorous exercise, etc. Patients with hypertension are further monitored to determine whether they have elevated sexual blood pressure or high blood pressure. Comparing the effects of body organ lesions and hypertension explored by previous studies, we are more likely to understand the intrinsic link between hypertension and weight age. We conclude that poor eating habits are also a major factor in high blood pressure. For example, excessive daily salt consumption, smoking and alcoholism. People with hypertension should pay attention to diet and exercise in their daily life. People with high blood pressure should eat light, low in fat, and exercise properly. In addition, secondary hypertension patients also need to adjust according to the cause, blood pressure, usually need to pay attention to ensure sleep, adjust mood and pay attention to climate change. Hypertension patients in the change of position to slow down a bit, especially for elderly hypertension patients, because some of their original organ function is decreasing, so it is easy to cause some small movements to cause accidents, so attention should be paid to change the position when slow.

4. Conclusion

With the continuous improvement of people's living standards and the great changes in diet structure, the occurrence and death of heart and cerebrovascular diseases in China are also rising, of which the proportion of deaths in China's total mortality rate has even reached about 40%, hypertension has become a cause for concern. Based on the results we have and the discussion, we can draw the following conclusions. Weight is closely related to high blood pressure. Weight gain is an important risk factor for high blood pressure. Abdominal obesity is more likely to cause high blood pressure. Because obesity can make patients endocrine hormone disorders, so that catecholamine, renin secretion increased, so that blood vessels constrict, so that the patient's blood pressure increased. Obesity may cause a large amount of fat in the body, can be large blood vessel bed, and blood circulation relatively increased, psychological also increased, long-term heart burden will cause thick ventricular wall, leading to elevated blood pressure. Obesity can also cause hyperinsulinemia, which can also cause excessive sodium storage in the body and high blood pressure. In today's high blood pressure frequency, mortality rate gradually increased today, to understand the cause of hypertension, it is particularly important. When we understand the cause of hypertension, we can take targeted measures to prevent the disease as far as possible in the absence of the onset of the disease, so that not only for the benefit of the patient itself, but the patient's also whole family and even the whole society is of great significance. A healthy body is the foundation of everything we have, and when a person has no healthy body, he loses almost everything. So, we do such a professional, forward-looking scientific study to help people with high blood pressure. Only when people are healthy and happy will society be stable and beautiful. Therefore, the government needs to pay attention to specific groups of people (obesity, the elderly, the young), as well as increased supervision of junk

food, promote healthy and rational diet, take concrete measures to help people better prevent high blood pressure, thereby improving the health index of society as a whole.

References

- [1] Buyukozturk K., Ilerigelen B, Kabakci G, et al. Intensive cardiovascular examination regarding blood pressure levels: evaluation of risk groups. ICEBERG study. [J]. Blood Pressure, 2009, 15(5):291-301.
- [2] S. S. Deschênes, Burns R J, Graham E, et al. Depressive symptoms and sleep problems as risk factors for heart disease: a prospective community study [J]. Epidemiology and Psychiatric Sciences, 2019, 29:1-10.
- [3] Nsien E F, Abam A O. Modeling Risk Factors of High Blood Pressure in Women Using Multiple Logistic Regression.
- [4] Mehdi B, Taghi A, Saeedeh P. The Reliability of Classification of Terminal Nodes in GUIDE Decision Tree to Predict the Nonalcoholic Fatty Liver Disease [J]. Computational and Mathematical Methods in Medicine, 2016, (2016-12-7), 2016, 2016:1-11.
- [5] Ellis, H. P., Mark, G., Ben, P., Inmaculada, S., Andrea, S., & Kurian, K. M. (2015). Current challenges in glioblastoma: intratumour heterogeneity, residual disease, and models to predict disease recurrence. *Frontiers in Oncology*, 5(26).
- [6] Satyanarayana, N., Ramadevi, Y., & Chari, K. K. (2018). High blood pressure prediction based on AAA using J48 classifier. *Conference on Signal Processing & Communication Engineering Systems* (pp.121-126). IEEE.
- [7] Lin, J. H. & Haug, P. J. (2008). Exploiting missing clinical data in bayesian network modeling for predicting medical problems. *Journal of Biomedical Informatics*, 41(1), 1-14.
- [8] El-Hajj, C. & Kyriacou, P. A. (2020). A review of machine learning techniques in photoplethysmography for the non-invasive cuff-less measurement of blood pressure. *Biomedical Signal Processing and Control*, 58, 101870-.
- [9] Farran, B. Channanath, A. M. Behbehani, K. & Thanaraj, T. A. (2013). Predictive models to assess risk of type 2 diabetes, hypertension and comorbidity: machine-learning algorithms and validation using national health data from Kuwait—a cohort study. *Bmj Open*, 3(5), 190–198.
- [10] Chuan-Xi, F. U. Wen-Jun, M. A. Liang, J. H. & Wang, D. H. (2006). Logistic regression and classification tree analysis on risk factors of hypertension. *Chinese Journal of Disease Control & Prevention*.
- [11] Paulson W D, Ram S J, Faiyaz R, et al. Association between blood pressure, ultrafiltration, and hemodialysis graft thrombosis: A multivariable logistic regression analysis [J]. *American Journal of Kidney Diseases*, 2002, 40(4):769-776.
- [12] Balchandran R, Boyer L M, Purdy G Sub-Model Generation to Improve Classification Accuracy [J]. US, 2008.