Machine Learning Models for Diabetes Prediction

Jeremy Pennington

Abstract

As of 2019, more than 34 million American have diabetes (about 1 in 10), and approximately 90-95% of them have type II diabetes, which is one kind of diabetes that most often develops in people over age 45, but more and more children, teens, and young adults are also developing it now. The purpose of this study is to examine the predictors of diabetes cases among Native Americans and build a predictive model for diabetes cases using the logistic regression and gradient boosting model in Python. The data used in this study were collected and made available by the National Institute of Diabetes and Digestive and Kidney Diseases. We first pre-processed the data to remove missing samples and check the presence of multicollinearity before feeding it into the models. We found for logistic regression, the precision for negative and positive classes are 0.85 and 0.7, while that for the gradient boosting model is 0.86 and 0.68. We also found that the cross-validation scores for logistic regression and gradient boosting are 0.778 and 0.782, respectively, indicating that no significant performance difference exists between two models.

Keywords: diabetes, logistic regression, gradient boosting

1. Introduction

Today, diabetes has become one of the most common serious diseases among Americans [1]. According to the statistics by Centers of Disease Control and Prevention (CDC), nearly every one out of ten Americans is suffering from diabetes, and more than 90% of them have type II diabetes [2]. Cells in type II diabetic patients have anormal resistance to insulin, a hormone that transfers glucose from the blood to cells, thus leading to higher-than-normal blood sugar levels [2]. Type II diabetes can cause serious health problems, including heart disease, vision loss, and kidney failures [2]. Diabetes was initially rare among Native Americans until the middle part of the twentieth century [3]. Since World War II, diabetes has become much more prevalent among them [4]. The Pima Indians have the highest recorded prevalence and incidence of type II diabetes in the world [5][6]. Since 1965, living in a geographically defined part of the Gila River Indian Community of Arizona, Pima Indians have participated in a longitudinal study of diabetes and its complications, from which much of our current understanding of diabetes among Native Americans has been obtained [7]. The objective of this study is to find the predictors of type II diabetes and develop a predictive model to detect diabetes cases in an early stage. With an accurate model, the healthcare system can collect survey data and score the responders' probability of diabetes. For those being more susceptible to diabetes, appropriate measures can also be taken in the early stage to control the development of the symptoms.

2. Exploratory analysis and data preprocessing

The data were collected and made available by the National Institute of Diabetes and Digestive and Kidney Diseases as part of the Pima Indians Diabetes Database [8]. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients included

in this study belong to the Pima Indian heritage (a subgroup of Native Americans) and are females of ages 21 and above.

Python and some of its popular data science related packages were used in this study. The data were imported using Pandas. Seaborn and Matplotlib were used for visualizations. NumPy and Scikit-learn were called to pre-process data into a suitable format for machine learning models.

The dataset contains nine variables, including eight feature variables and one classification variable, with 768 entries in total. A description of the features can be found in table 1. The classification variable is a dichotomous variable representing positive (1) and negative (0) diabetes cases. The distribution of the target variable is shown in figure 1. We can see the target variable is imbalanced with the negative class being approximately twice the size of the positive class. This could cause a classification model to over-favor predictions of the majority classes, thus crippling the model's ability to classify the minority class.

Table 1. Description of the 28 Features

Variable	Description
Pregnancies	Whether the subject is pregnant at the time of survey
Glucose	The subject's glucose level at the time of survey
BloodPressure	The subject's blood pressure at the time of survey
SkinThickness	The subject's skin thickness at the time of survey
Insulin	The subject's insulin level at the time of survey
BMI	The subject's body mass index (BMI) at the time of
DiabetesPedigreeFunction	The subject's Pedigree function score, which measures the likelihood of diabetes based on family history [9].
Age	The age of the subject at the time of survey

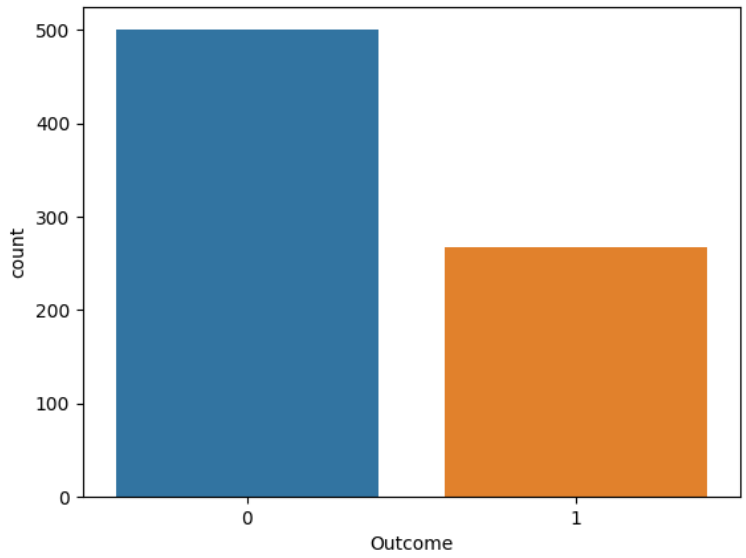


Figure 1: Distribution of the target variable, where "1" denotes diabetes positive and "0" negative

The distribution plot of each of the eight feature variables is shown in figure 2. Seaborn distribution plot function combines the matplotlib histogram function with the seaborn kernel density estimate (KDE) plot and rugplot (marginal distributions) functions [10]. The distribution plot shows a univariate distribution of observations. The x-axis shows bins (ranges) of the variable and the y-axis is the probability density function for the kernel density estimation. The zero values in glucose, blood pressure, skin thickness, and BMI in figure 2 show mistakes or missing values in the data. This problem is addressed by dropping all the samples with zero glucose, blood pressure, and BMI, also by filling all the zero skin thickness values with random numbers from the normal distribution which has the same mean and standard deviation as the non-zero skin thickness values in the data. Figure 3 shows the distribution of skin thickness after imputing the missing values.

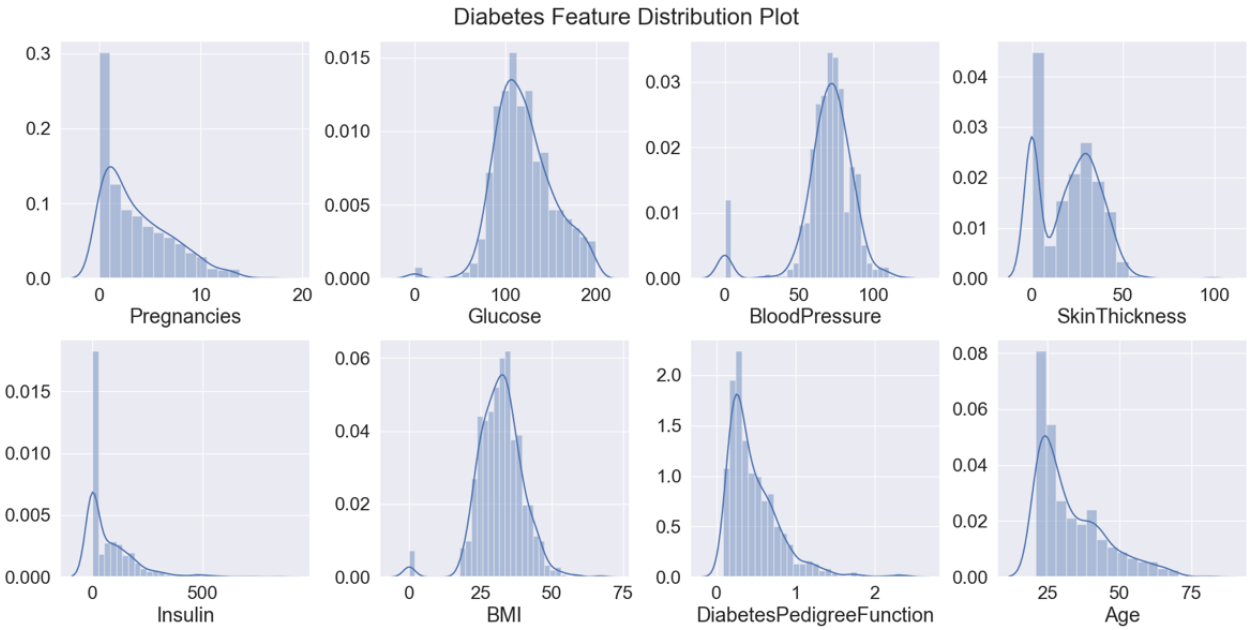


Figure 2: Kernel density distribution plot of feature variables

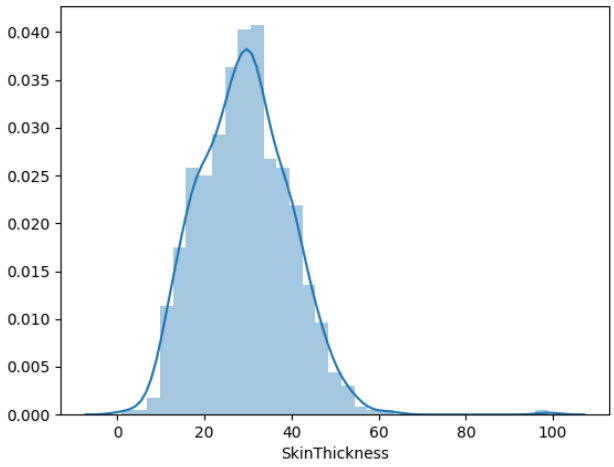


Figure 3: Distribution plot of skin thickness after imputing missing values

One step in the preprocessing phase is centering and scaling each feature variable independently. Feature standardization transforms different features into comparable scales and ensures all features weigh equally in the training process. The standard scaler function from scikit-learn

standardizes features by scaling the mean to zero and standard deviation to unit variance [11]. Figure 4 shows the distribution of feature variables after applying the standard scaler.

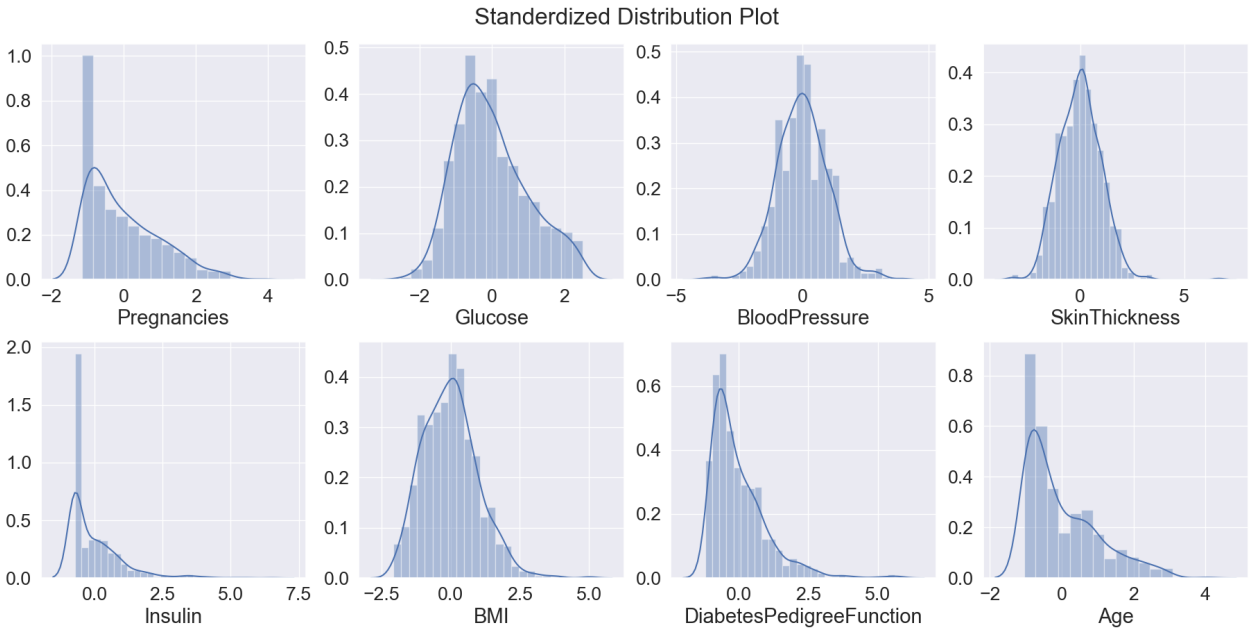


Figure 4: Kernel density distribution plot of feature variables after standardization

One problem that might compromise the performance of a logistic regression model is multicollinearity. Multicollinearity is present when several features in the model are highly correlated. We plotted a heap map to diagnose the presence of multicollinearity, as can be shown in figure 5. A heat map is a graphical representation of data where values are depicted by color. In figure 5, the bigger the correlation strength, the paler the color, which in turn shows the feature variable are relatively independent of each other, except for age and pregnancies, which can be explained by common sense.

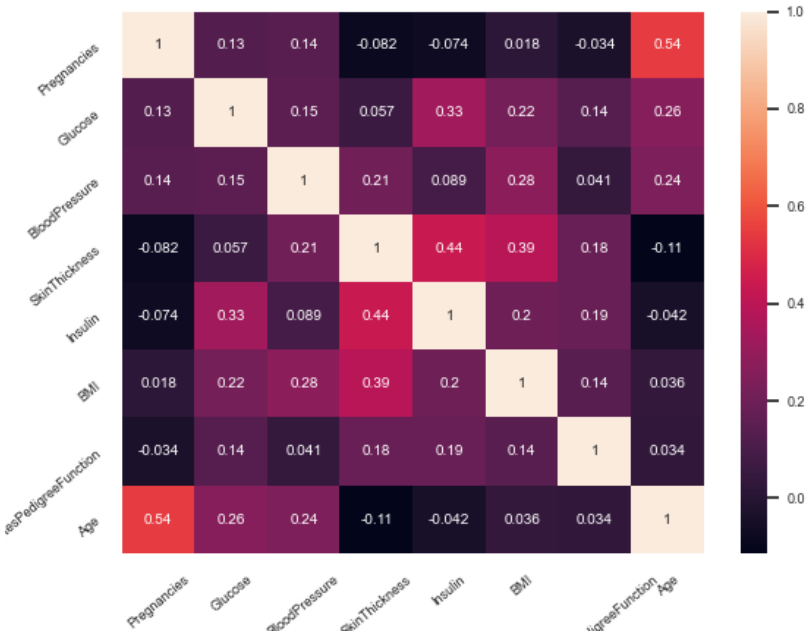


Figure 5: Heat map of feature variable correlations

3. Models

The data were randomly split into training and test subsets using the scikit-learn package. The training set has 80% of data, while the test set has the remaining. The logistic regression and the gradient boosting classifier were both trained using the training set, while the performance is measured on the test set.

Logistic regression is a part of a category of statistical models called generalized linear models, and it allows one to predict a discrete outcome from a set of variables that may be continuous, discrete, dichotomous, or a combination of these. Typically, the dependent variable is dichotomous, and the independent variables are either categorical or continuous. In logistic regression, each feature x_i has its specific weight w_i. The net input y is calculated as follows:

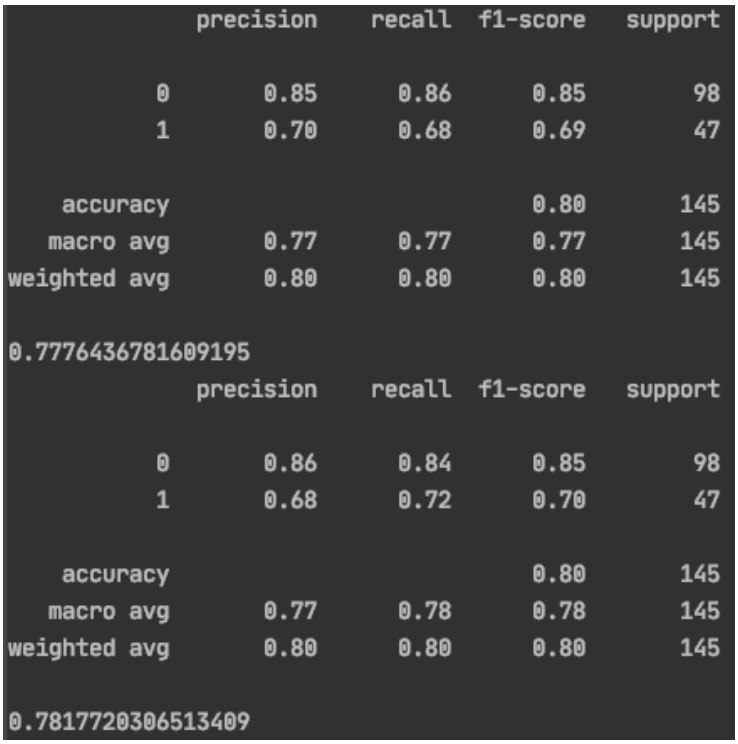
$$\ln\left(\frac{y}{1-y}\right) = w_0 + w_1 x_1 + \cdots + w_m x_m$$

Another candidate model we used for this study is gradient boosting. Gradient boosting is a machine learning technique which makes predictions based the results from several weak predicting models such as decision trees. The weak models within a gradient boosting model are trained in a gradual, additive and sequential manner. As a result, new models are built on the top of the existing models and try to further improve the total performance and minimize the prediction errors.

In training the models, a key challenge is overfitting, which is how the model will perform on new, unseen data. Cross-validation is a very useful technique for assessing the effectiveness of the model. For this study, to reduce the overfitting problem, we applied the technique of 5-fold cross-validation. It works by partitioning the data into five subsets and holding out one subset at a time to train the model on the remaining four sets and then testing the model on the hold-out set.

4. Results

The classification results from two models and the cross-validation scores are shown in figure 6. The top half in figure 6 is the result of the logistic regression and the bottom half is that of the gradient boosting model. The classification report shows the main performance metrics of the machine learning model in predicting each class, including precision, recall, f1-score, support, their macro (unweighted) and weighted averages, and overall accuracy.



	precision	recall	f1-score	support
0	0.85	0.86	0.85	98
1	0.70	0.68	0.69	47
accuracy			0.80	145
macro avg	0.77	0.77	0.77	145
weighted avg	0.80	0.80	0.80	145

0.7776436781609195

	precision	recall	f1-score	support
0	0.86	0.84	0.85	98
1	0.68	0.72	0.70	47
accuracy			0.80	145
macro avg	0.77	0.78	0.78	145
weighted avg	0.80	0.80	0.80	145

0.7817720306513409

Figure 6: classification reports and cross-validation scores of logistic regression (top half) and gradient boosting (bottom half)

Table 2 shows the odds ratios (converted from the log-odds which are the coefficients provided by logistic regression classifier) of all the feature variables. The impact size of factor "Age" can be expressed by its odds ratio:

$$\frac{Odds_{Diabetes|Age=A+1}}{Odds_{Diabetes|Age=A}},$$

which is a relative measure of impact size that is not necessarily related to the innate probability of the event. If the odds ratio is equal to 1, it means the odds of the events in the numerator is the same as the odds of the events in the denominator, and if the odds ratio is above 1, the events in the numerator has favorable odds comparing to the events in the denominator.

The odds ratio of "Age", from table 2, is 1.201. This result says that, holding all the other variables fixed, by increasing one year of age we expect to see the odds of getting diabetes increase by about 20.1%. Similarly, holding all the other variables fixed, by increasing one time of pregnancy, the

odds of getting diabetes increase by about 46.5%. For glucose, BMI, and Diabetes Pedigree Function, the odds increase by 197%, 113%, and 33.5%, respectively. For blood pressure, skin thickness, and insulin, the odds of getting diabetes decrease by 11.5%, 7.2%, and 13.5%, respectively.

Table 2. odds ratio of each feature variable

Features	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age
Odds ratio	1.465	2.964	0.885	0.928	0.865	2.128	1.335	1.201

5. Discussion

The main evaluation metrics of the models include precision, recall, f1-score, support, and overall accuracy. Precision is the percentage of correct predictions, recall is the percentage of predicted positive cases, f1-score is the percentage of correct positive predictions, and support is the number of actual occurrences of the class in the specified dataset [12]. The result, based on figure 6, indicates that when it comes to predicting positive diabetes cases, logistic regression model has higher precision, same support and accuracy, but lower recall and f1-score compared to gradient boosting model. Overall, two models have comparable performance.

To conclude, the intention of this study is to build a predictive model with the best performance and to investigate the predictors related to diabetes. From the odds ratios obtained by the logistic regression model, pregnancy, glucose, and BMI are found to have the biggest impact on diabetes outcome.

Two models, logistic regression and gradient boosting, were built, and both achieved a similar and superb performance. Also, by plotting a correlation heatmap, we are able to verify most features are independent, except for pregnancy and age. As our models have achieved relatively good performance in predicting diabetes given a subject's current situation, we believe our models can be used by healthcare professionals to assist in diagnosing their patients' risk of diabetes.

One limitation of this study, based on figure 6, is that both models have a higher precision score on negative cases than positive cases. The difference between the precisions scores is likely due to the imbalance in the dataset, as we observe more negative cases than positive cases in the data in figure 1. To handle this, in future study, we can use other techniques such as over-sampling or under-sampling. For example, we can use the random over-sampling technique to increase the positive class size. For the training set, we can randomly select samples from the positive class with replacement, duplicate them, and add them back to the training set. Meanwhile, we can also apply the classic nearest neighbor under-sampling technique to remove negative samples. In this case, for each negative sample, we will find and remove its nearest neighbor in the data set to minimize the loss of the information. By combining those two techniques, our model performance might be significantly improved and will be better able to identify potential diabetic patients.

References

1. Sievers ML, Fisher JR. Diabetes in North American Indians. In: Diabetes in America. Bethesda, Maryland: US Department of Health and Human Services, Public Health Service, National Institutes of Health, 1985: chapter XI, 1-19; DHHS publication no. (NIH)85-1468

2. Refer to "https://www.cdc.gov/diabetes/basics/type2.html"

3. Venkat Narayan K.M. Diabetes Mellitus in Native Americans: The Problem and Its Implications. https://www.ncbi.nlm.nih.gov/books/NBK233089/#_ddd00162_

4. Sievers ML, Fisher JR. Diabetes in North American Indians. In: Diabetes in America. Bethesda, Maryland: US Department of Health and Human Services, Public Health Service, National Institutes of Health, 1985: chapter XI, 1-19; DHHS publication no. (NIH)85-1468

5. Knowler W.C., Bennett P.H., Hamman R.F. & Miller M., (1978). Diabetes incidence and prevalence in Pima Indians: a 19-fold greater incidence than in Rochester, Minnesota. American Journal of Epidemiology, Volume 108, Issue 6, December 1978, Pages 497–505

6. King, Hilary & Rewers, Marian. (1991). Diabetes in adults is now a Third World problem / H. King & M. Rewers on behalf of the WHO Ad Hoc Diabetes Reporting Group. Bulletin of the World Health Organization 1991; 69(6): 643-648

7. Bennett P.H., Burch T.A. & Miller M. (1971), Diabetes Mellitus in American (Pima) Indians. Volume 298, Issue 7716, P125-128, July 17, 1971

8. National Institute of Diabetes and Digestive and Kidney Diseases: https://www.niddk.nih.gov

9. Refer to “https://rpubs.com/ikodesh/53189”

10. Refer to “https://seaborn.pydata.org/generated/seaborn.distplot.html”

11.Refer to

“https://scikitlearn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html”

12. Refer to “https://medium.com/@kohlishivam5522/understanding-a-classification-report-for-your-machine-learning-model-88815e2ce397”