Scientific Research Publishing

Journal of Data Analysis and Information Processing, 2026, 14(4), 377-391
https://www.scirp.org/journal/jdaip
ISSN Online: 2327-7203
ISSN Print: 2327-7211

Predicting Diabetes-Related Survey Status Using Logistic Regression: Class Imbalance and the Effect of Classification Thresholds

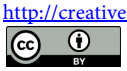
Mizeng Tang

Shanghai Experimental Foreign Language School, Shanghai, China
Email: zxt765971@gmail.com

How to cite this paper: Tang, M.Z. (2026) Predicting Diabetes-Related Survey Status Using Logistic Regression: Class Imbalance and the Effect of Classification Thresholds. *Journal of Data Analysis and Information Processing*, **14**, 377-391.
https://doi.org/10.4236/jdaip.2026.144019

Received: October 20, 2026
Accepted: November 15, 2026
Published: November 18, 2026

Copyright © 2026 by author(s) and Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/



Open Access

Abstract

The aim of this study was to evaluate how well logistic regression classifies diabetes-related status from health indicators in public survey-derived data. Using the UCI CDC Diabetes Health Indicators dataset, we compared a model using age category and BMI with an expanded model using selected demographic and health indicators. The share of records in the positive category, which combines reported prediabetes and diabetes, is 13.93%. The study also examined a class-weighted model and tested the impact of changing the classification threshold for predicting the outcome. The records were divided into three parts for stratified training, validation, and test purposes in proportions of 60%, 20%, and 20%, respectively. On the independent test sample, the expanded unweighted model had a ROC-AUC of 0.7966 and an average precision of 0.3640. Using a threshold of 0.50 resulted in a recall of 0.0941. However, selecting a threshold of 0.10, which was chosen on validation data to meet an illustrative 80% recall target, increased the recall to 0.8427 but reduced precision to 0.2530 along with an increase in false positives. Accuracy alone is not adequate for evaluating imbalanced classification problems, and an appropriate threshold should be set for the model's purpose. This exploratory, unweighted secondary analysis is not a nationally representative estimate or a validated clinical screening tool.

Keywords

Logistic Regression, Diabetes, Class Imbalance, Precision and Recall, Classification Threshold

1. Introduction

Diabetes is a major public health concern. One useful application of statistics con-

cerning health data is to examine associations between health indicators and diabetes. Some common health characteristics examined through public health studies are body mass index (BMI), age, reported blood pressure, and physical activity. With the aid of the statistical models, we can analyze multiple variables simultaneously and predict the probability that survey records belong to a diabetes-related category. Nevertheless, the fact that the model can predict does not necessarily imply that it is accurate enough to guide medical decisions. Understanding the errors the model has made is just as important as measuring its successes.

We use a public version of the CDC Diabetes Health Indicators that was made available through the University of California, Irvine's (UCI) Machine Learning Repository [1]. The dataset in this study was derived from the Behavioral Risk Factor Surveillance System (BRFSS) of the Centers for Disease Control and Prevention (CDC) [2]. The dataset contains 253,680 records, 21 independent variables, and one binary dependent variable. We did not recreate the sampling design used in the BRFSS nor did we use the survey weights. As a result, our analysis applies only to the processed data itself and is not representative of the larger American population.

This analysis will use logistic regression since its main formula is easy to understand with algebra, exponential functions, and the concept of probability. Unlike machine learning techniques that generate hundreds of rules for decision making, logistic regression calculates a weighted score based on the characteristics of an individual that then produces a number ranging between 0 and 1. Logistic regression is more interpretable than many complicated machine learning techniques because it allows one to understand why the model produces different results given different characteristics.

A major problem that this analysis will deal with is *class imbalance*. Records in category 0 greatly outnumber those in category 1 in the dataset. In this case, a classifier can predict the majority class and therefore achieve a high overall accuracy. Nevertheless, this model may produce poor results for the situations that researchers are most interested in. Our investigation therefore starts from the discrepancy between the overall accuracy and the ability to identify positive cases.

A second issue is the cutoff point used for classification. Logistic regression produces a probability-like score while most classification reports require a yes-or-no answer. By convention, if the number is greater than or equal to 0.50, the case can be classified as positive. This, however, is not a law of mathematics, but only a convention. The same can be done at a lower threshold. Indeed, this will usually lead to finding more positives, but also more false positives. Consider a classroom example about potential screening for additional positives. Although finding additional positives may seem desirable, the cost of the additional false alarms must also be taken into account.

As noted in previous research, precision-recall analysis is useful when class sizes differ [3]. One can compare discrimination across a range of thresholds with the use of ROC curves [4]. This raises an interesting issue beyond which model

achieves the highest accuracy. *How does changing the threshold alter the types of classification errors*? In our example, we compared the thresholds in validation data and then evaluated one selected rule on the held-out test data.

Therefore, the main questions that this research will try to answer are the following. 1) Can age and BMI alone differentiate between the two classes of the survey, and do other selected indicators improve performance? 2) What is the impact of having many more examples in one class than in the other? 3) How does changing the threshold affect recall, precision, and the number of false positives? The major outcome of this research is a simple example, using actual experimental results, demonstrating that high accuracy may hide poor performance in terms of identification of the minority class.

2. Data and Methods

2.1. Dataset and Study Design

Using ucimlrepo [1], we retrieved data with ID 891. It contains 253,680 entries with 21 potential predictor variables. In addition, Diabetes_binary is the dependent variable, with category 0 representing no diabetes labels and category 1 representing either reported prediabetes or diabetes. In fact, these categories are existing labels in an already processed survey dataset, so they do not denote any diagnosis by our work. That is, positive labels do not mean that diabetes was developed after an observation period.

We performed a secondary exploratory unweighted analysis. In this context, the term "secondary" means that we used data collected and processed by others. The term "unweighted" refers to our not using any population sampling weights from the original survey in our modeling. The original CDC 2015 BRFSS methodology discusses how the weight and additional survey design elements work [2]. However, the simplified UCI table is not a substitute for the full survey analysis files. For this reason, we do not interpret the proportions of categories in this dataset as population prevalence, nor do we apply the model to assess individual clinical risk.

The data check done in the notebook showed that there are no NaN values in the dataset. The BMI column had an average of 28.38 and a median of 27 with its values in the range of 12 to 98. These statistics refer to the downloaded data alone. Some variables, for instance age, education, and income, are coded as categories rather than continuous values. The preparation of the models involved transforming the age, education, and income variables into indicator variables, while other variables selected were scaled by applying statistical parameters derived from the training data.

2.2. Predictors and Model Comparison

In our *simple* model, we consider only two features, the age category and the BMI variable. These two features are intuitive for students. The age variable provides information on the life stage of the individual while BMI is calculated based on

weight and height of the person. In the *expanded* model, we used 12 of the 21 available predictors: age category, BMI, sex, high blood pressure, high cholesterol, smoking history, physical activity, fruit consumption, vegetable consumption, heavy alcohol consumption, education, and income. We chose this smaller group to keep the model interpretable while representing demographic characteristics, cardiovascular and metabolic health indicators, and lifestyle behaviors. The expanded model allowed us to examine whether these additional predictors improved classification compared with age and BMI alone.

The predictors were defined in the notebook before fitting or evaluating any of the models. No procedure to add or remove individual predictors based on validation performance was implemented; the validation set was used to compare the prespecified models and select a classification threshold. We do not claim that these parameters cause diabetes, since some can actually be consequences of an illness, be connected to other unmeasured variables, or be measured imperfectly. The critical task is to determine whether they can help classify the existing outcome label, not whether changing one variable would prevent disease.

We trained a third model with the same broader predictors, but with the implementation of *balanced class weights*. The reason for doing this is that the learning process will use a higher weight for misclassification in the minority class. This allows us to contrast using a balanced class weight with generating fake samples. In this case, rather than generating a sample from nothing, we adjust the weight of the existing samples in the training data. The balanced class weight feature that was used in scikit-learn is based on the number of observations in each class [5].

2.3. How Logistic Regression Works

Logistic regression begins with a linear score,

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k, \tag{1}$$

Each x_j is a predictor variable with β_j as the coefficient associated with that variable, and β_0 as the intercept in the logistic regression equation. When the coefficient is positive, it indicates that an increase in the predictor variable leads to an increase in the score with all other variables remaining constant. Score values should not be interpreted as probabilities since they may take on negative values and values greater than 1.

The logistic, or sigmoid, function converts the score into a number between zero and one:

$$p = \frac{1}{1 + e^{-z}} \tag{2}$$

If $z = 0$, then $e^{-z} = 1$, which indicates that $p = 0.5$, whereas when z is large and positive, p becomes very close to one. In contrast, when z is very large and negative, p becomes very close to zero. The smoothness in this relationship allows the algorithm to determine the coefficients based on the training set. However, it is essential to understand that a score of $p = 0.20$ is a fitted probability score, which should not be confused with an actual diagnosis in medicine.

For this project, we used the LogisticRegression method in scikit-learn with the lbfgs solver, L2 regularization (penalty = "l2"), regularization strength C = 1.0, and a maximum of 1500 iterations (max_iter = 1500) [5]. The simple and expanded unweighted models used class_weight = None, while the class-weighted expanded model used class_weight = "balanced". We also applied the pipeline approach to combine preprocessing and estimation to ensure that transformations were carried out based solely on the training data. Although regularization may work well to limit the extent of extreme coefficient values, we focus on comparing label predictions and not on formally testing each coefficient. There are no statistical significance tests presented, and no strong conclusions made about any risk factor specifically.

2.4. Training, Validation, and Testing

We partitioned the dataset randomly into three parts of 60%, 20%, and 20%. There were 152,208 records in the training set (60%), 50,736 records in the validation set (20%), and 50,736 records in the test set (20%) respectively. Both stratified data splits used a fixed random seed of 42 (random_state = 42) to ensure reproducibility. First, 20% of the records were assigned to the test set. The remaining 80% were then divided into training and validation sets at a ratio of 3:1, resulting in final proportions of 60%, 20%, and 20%. The stratification was preserved through the partitions to ensure that the data had approximately the same positive rates.

Each of the three datasets has a different purpose. The training dataset is meant for estimating the coefficients. The validation set can be used in comparing different models and thresholds without looking at how they perform on the test set. The test set gives an indication of how well the final model (using the best threshold found from the validation set) behaves on unseen data. If one were to continually adapt the threshold based on its test set performance, then the test set would no longer be a fair measure of performance. We took care to avoid this error.

As mentioned earlier, the separation between training and validation sets also applies to preprocessing. This means that the pipeline will impute any missing values in the training dataset and scale all numeric features using the training set data. Then the same transformations will be applied to both the validation and test datasets. Although the completed dataset did not contain NaN values, having the procedure defined in a pipeline is helpful for replicating the process.

2.5. Classification Metrics

The confusion matrix is a way to show the four possible outcomes of the model. A *true positive* (*TP*) refers to a category-1 instance that is correctly predicted as positive. By contrast, a category-1 instance that is not detected (*false negative*, *FN*) is one that wrongly received a negative prediction. A *false positive* (*FP*) is a category-0 instance that is wrongly given a positive prediction. A *true negative* (*TN*) is a category-0 instance that is correctly given a negative prediction. In total, these four numbers show in detail which kinds of mistakes the model makes.

Accuracy measures the fraction of all predictions that are correct:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

What accuracy means is simple to understand. However, this term could become misleading when one class is much larger than the other. We therefore also include other measures, namely precision, recall, specificity, and F1 score.

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{4}$$

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{5}$$

$$\text{Specificity} = \frac{TN}{TN + FP}, \tag{6}$$

$$F_1 = \frac{2\ \text{Precision}\ \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

Precision tells us the percentage of positive predictions that actually have label 1. Recall indicates the percentage of actual label-1 instances that were correctly predicted. Specificity is the percentage of correctly predicted category-0 instances. The F1 score combines both precision and recall and therefore has a small value if either precision or recall is very small.

Two other interesting measures are ROC-AUCand average precision (AP). The ROC-AUC tells us how well the predictions order the positive cases above negative ones based on different thresholds [4]. In contrast, average precision is an aggregation of precision across the precision-recall curve and is especially helpful when positive cases are less common [3]. Average precision and the ROC-AUC are also unlike accuracy or recall at a fixed threshold in that changes to the threshold for the same model scores will not affect their values.

To quantify uncertainty in the test results, we used bootstrap resampling to calculate 95% confidence intervals for ROC-AUC, average precision, recall, and precision. We generated 2000 bootstrap samples by randomly sampling the held-out test records with replacement. For each sample, we recalculated the performance metrics using the original fitted model scores. The fitted model and the two classification thresholds (0.50 and 0.10) were kept fixed throughout this procedure. The 2.5th and 97.5th percentiles of the bootstrap estimates were used as the confidence interval limits.

2.6. The Threshold Experiment

A threshold t turns estimated probabilities into labels:

$$\hat{Y} = \begin{cases} 1, & p \geq t, \\ 0, & p < t. \end{cases} \tag{8}$$

In our primary experiment, we varied t from 0.05 to 0.50 in increments of 0.025 and looked at the change in precision, recall, specificity, F1, and false positives on the validation records for the unweighted expanded model. We did not decide on

a particular threshold after the final test results were obtained. Instead, we decided upon a simple illustrative rule that picks, among the tested thresholds reaching at least 80% recall on validation records, the one with the best precision. However, if no threshold meets this requirement, the notebook would fall back to the maximum recall threshold.

The 80% level is a student research choice to illustrate the principle of trade-off. It is not based on any medical guidelines and is not a recommendation regarding screening for diabetes. A true clinical threshold would require information about the accuracy of the underlying medical outcome, clinical consequences, costs, and testing pathways. After the validation threshold was determined, we compared it to the traditional 0.50 threshold on the same untouched test set.

3. Results

3.1. Outcome Frequencies and BMI Distribution

From a total of 253,680 records, 218,334 records have label 0, while 35,346 records have label 1. This means that the positive label comprises 13.93% of the total records, or that approximately 1 out of 7 records falls into the prediabetes/diabetes class. The remaining negative class comprises around 86.07%, which indicates that there is a significant imbalance and that the accuracy metric may give an incorrect impression of the model's performance, even in circumstances where it is difficult for the model to identify positive labels.

In Figure 1, we see not only the count imbalance, but also the distribution of BMI for the two labeled classes. The BMI distributions also show a significant overlap. This is important, as BMI alone will not be able to fully distinguish between classes. There are records with similar BMI values in both classes. The average BMI of all records is 28.38. The study did not attempt to make causal inferences from the graph, as the purpose was mainly to motivate comparing a two-predictor model with one using multiple independent variables.

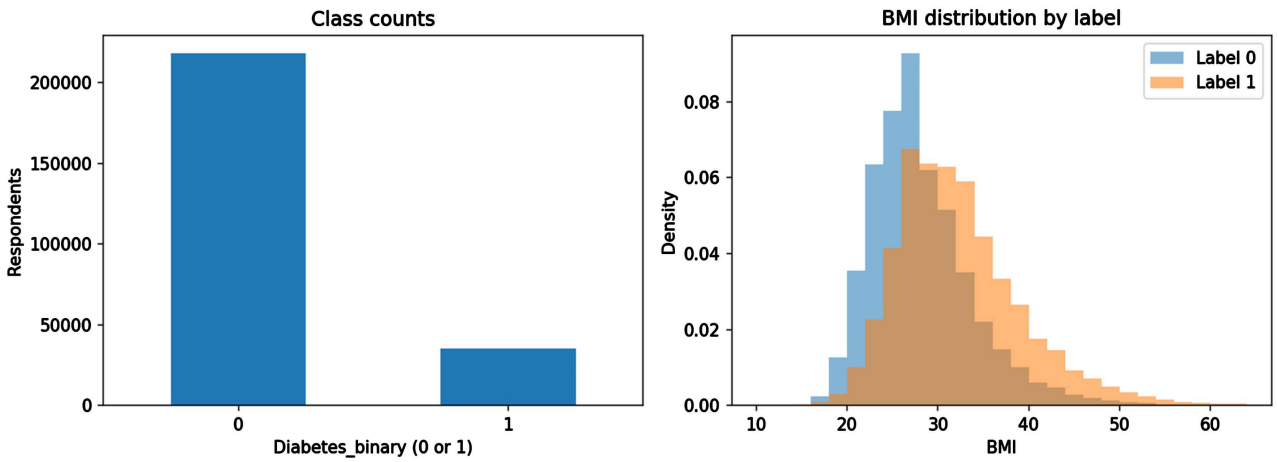


Figure 1. Counts of the two outcome categories and BMI distributions from the downloaded UCI dataset. The categories are unequal in size, and the BMI distributions overlap. Graphic reproduced from the supplied experiment at 500-dpi JPG export.

To help illustrate the importance of the imbalance, if we create a classifier that will always return a 0 label, it will correctly classify all 218,334 negative records and will misclassify all 35,346 positive records. While its overall accuracy will still be 86.07% because most records are negative, its recall and precision for class 1 will still be zero according to the zero-division convention in our implementation of the algorithm. Such a classifier will not serve any useful purpose in finding positive examples even if its level of accuracy appears satisfactory at first glance.

3.2. Validation Comparison of the Three Models

The threshold of 0.50 is used to describe the validation output shown in Table 1. Logistic regression using a simple model achieved a ROC-AUC of 0.7446 and an AP of 0.3009, while the expanded model obtained higher scores of 0.7972 and 0.3688 respectively. This indicates that it was possible to separate the validation labels better through a combination of selected health indicators with BMI and age. Just because there is a higher AUC does not mean that any one added indicator caused the development of diabetes.

Table 1. Validation performance at threshold 0.50 (50,736 records). AP denotes average precision.

Model	Accuracy	Precision	Recall	Specificity	ROC-AUC	AP
Simple logistic regression	0.8588	0.4226	0.0359	0.9921	0.7446	0.3009
Expanded logistic regression	0.8625	0.5371	0.0963	0.9866	0.7972	0.3688
Expanded, balanced weights	0.7062	0.2870	0.7472	0.6996	0.7980	0.3670
Always predict category 0	0.8607	0	0	1.0000	0.5000	0.1393

The standard expanded model achieved an accuracy of 86.25%, but its recall rate was only 9.63%. On the validation data at the default threshold, this model identified less than one in ten of the true positives. The simple model performed even worse with a recall value of just 3.59%. In fact, using the baseline model of predicting category 0 at all times was slightly more accurate than using logistic regression with the simple model. Assessing the accuracy alone would mask the problem of an insufficient detection rate for the less common label.

The class-weighted expanded model produced very different results, demonstrating a recall of 74.72% at the same threshold, but achieving an accuracy of only 70.62% and a precision of 28.70%. The ROC-AUC score of 0.7980 was similar to that of the expanded model with no weights (*i.e.* 0.7972). So, class weighting strongly affects whether a prediction falls above or below the standard threshold. Nevertheless, the ranking based on the score barely improved. The class-weighted model recorded 13,119 false positives versus the 587 recorded by the standard model at a threshold value of 0.50.

This information, however, does not suggest that class weighting is either good or bad in every application. There may be a model that is designed to identify most of the positive cases and that accepts more false alarms at the same time. On the other hand, if the cost of false alarms is high, then one can opt for a different operating point. In our project, class weighting was used to show the impact of the choice of learning objective on classification decisions.

3.3. Precision-Recall Trade-Off across Thresholds

Subsequently, we adjusted the threshold of the standard expanded model using the validation data. The results are shown in Figure 2. When the threshold was set at $t = 0.05$, recall was about 0.950 and precision was 0.201, while at $t = 0.10$, recall was 0.844 and precision was 0.254. When the threshold value was set at $t = 0.20$, the recall dropped to 0.609 while the precision increased to 0.327. Finally, at $t = 0.30$, the recall and precision values were 0.389 and 0.408 respectively. Higher values of the threshold generally lead to fewer positive predictions and fewer false positives, but have the drawback of increasing missed positives.

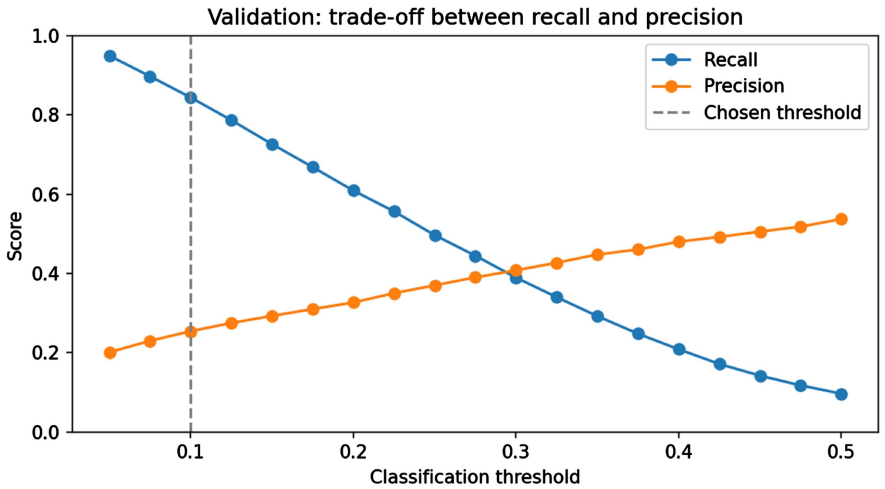


Figure 2. Validation-set threshold experiment for the unweighted expanded model. Precision and recall move in opposite directions across much of the threshold range. We selected the threshold before using test results.

Our classroom rule required that the recall be at least 80%. To satisfy this condition, we took the eligible threshold with the highest precision, which was 0.10. The next higher threshold value, 0.125, provided a recall value of approximately 0.787, which was less than the stipulated requirement of 0.80. So, the choice of 0.10 followed the stated numerical rule rather than an impression from the final test set.

If the fitted value score for a record is 0.12, the record would be considered positive under the new threshold but would have been considered negative according to the traditional threshold score of 0.50. However, this does not guarantee that the record would indeed have diabetes. Rather, it merely demonstrates

that the same predictor score may lead to two different predictions, depending on concerns over possible false negatives. The threshold can be changed without the model having to be fitted again, which makes it a useful option for students to experiment with.

3.4. Independent Test-Set Performance

After fixing both the model and thresholds on the training and validation data, we assessed the standard expanded model using 50,736 test records. The two pre-specified decision rules are summarized in Table 2. At a threshold of 0.50, test accuracy was 0.8615, precision was 0.5163, and recall was only 0.0941. Using the threshold of 0.10 determined during validation resulted in an accuracy of 0.6315 but a recall of 0.8427, with a corresponding precision of 0.2530, and specificity dropped from 0.9857 to 0.5973.

Table 2. Independent test results for the unweighted expanded model. The threshold 0.10 was selected on validation data, not test data.

Rule	Accuracy	Precision	Recall	F1	Specificity	ROC-AUC	AP
Default $t = 0.50$	0.8615	0.5163	0.0941	0.1591	0.9857	0.7966	0.3640
Validation-selected $t = 0.10$	0.6315	0.2530	0.8427	0.3892	0.5973	0.7966	0.3640

Table 3 reports the bootstrap 95% confidence intervals for the test-set performance measures. These intervals describe the uncertainty associated with resampling the held-out test records while keeping the fitted model and classification thresholds unchanged. They do not account for variation from retraining the model on different training samples or for the original survey sampling design.

Table 3. Bootstrap 95% confidence intervals for independent test-set performance.

Metric	Estimate	95% CI
ROC-AUC	0.7966	[0.7912, 0.8018]
Average precision	0.3640	[0.3530, 0.3750]
Recall ($t = 0.50$)	0.0941	[0.0872, 0.1007]
Precision ($t = 0.50$)	0.5163	[0.4899, 0.5434]
Recall ($t = 0.10$)	0.8427	[0.8341, 0.8511]
Precision ($t = 0.10$)	0.2530	[0.2475, 0.2587]

Note: Confidence intervals were estimated using 2000 boot-strap resamples (random seed = 2026).

The identical values for ROC-AUC and AP across the two rows are not an error, as these are calculated from the same model scores with the same test records. The only difference was the threshold that converted scores into labels. Therefore,

metrics that are dependent on the threshold, such as recall, precision, accuracy, and F1, changed while summary metrics that are independent of score thresholds, like AP and ROC-AUC, did not.

The 0.10 threshold influenced the results significantly in that it resulted in 17,584 false positives compared with only 623 at threshold 0.50 and led to far fewer missed positive cases, as we can see in the confusion matrix for the 0.10 threshold in Figure 3. Higher recall doesn't come without a cost. Misclassifying many additional category-0 records as category 1 could, in a practical system, lead to anxiety or increased follow-up, depending on how the predictions are used. The dataset in this study cannot, by itself, provide the economic or healthcare costs associated with the outcomes.

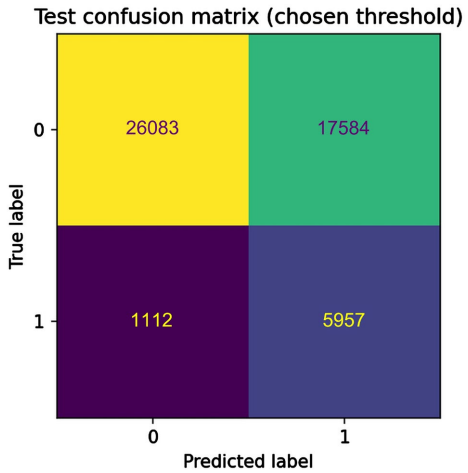


Figure 3. Test confusion matrix for the expanded model at the validation-selected threshold of 0.10. The matrix includes both correctly identified positives and additional false-positive classifications.

The notebook also converted the misclassifications into counts of records per 1000 records using only the test dataset. For threshold 0.50 there were about 12.3 false positives and 126.2 false negatives per 1000 test records. At threshold 0.10, 346.6 false positives and 21.9 false negatives were recorded per 1000 records. The first of the two thresholds fails to catch a lot of positive cases, but has few falsely predicted category 1's. In contrast, threshold 0.10 catches most of the positives but is misclassifying many records from category 0 as category 1. These are *within-dataset descriptive rates* of missed and false positive cases, not predictions if 1000 people were recruited from the general population.

Lastly, the test ROC-AUC was 0.7966 and the average precision was 0.3640, as shown in Figure 4. The average precision was higher than the baseline positive proportion of approximately 0.1393, indicating better performance than a random-ranking baseline. However, average precision summarizes performance across different thresholds and does not directly measure the number of false positives at a specific threshold. At the selected threshold of 0.10, the model achieved a recall of 0.8427 and a precision of 0.2530. This means that approximately 74.70%

of the records predicted as positive were actually negative, corresponding to 17,584 false positives in the test set. These results demonstrate the trade-off between identifying more positive cases and generating additional false alarms.

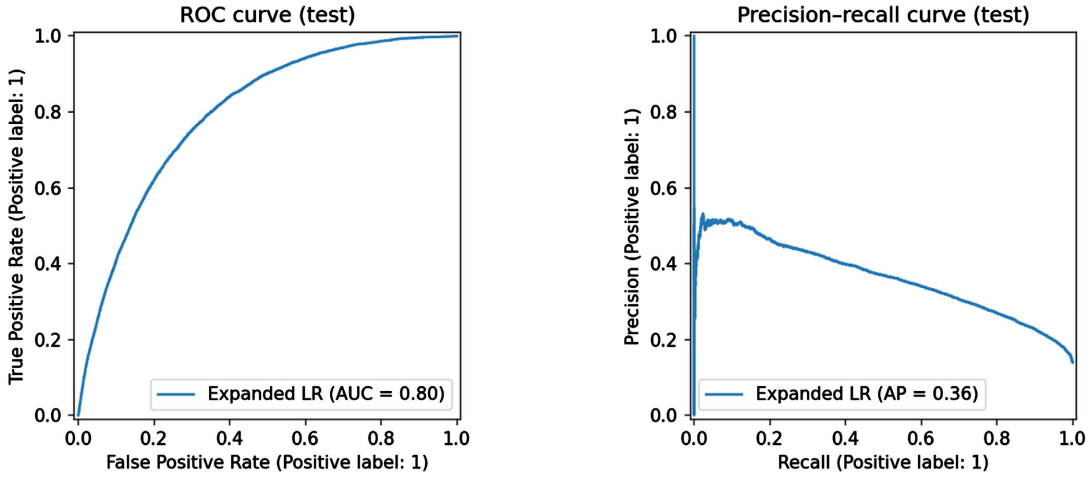


Figure 4. ROC and precision-recall curves for the expanded model on independent test data. Both curves use model scores across thresholds rather than one selected classification cutoff.

4. Discussion

4.1. Why High Accuracy Was Not Enough

The first point to consider is that performance metrics always need to be judged in reference to the underlying question being answered. Total accuracy considers all correct predictions in the scoring process as if they were all equal. In the case of our imbalanced dataset, total accuracy is greatly biased by the much more prevalent negative category. In our results, the classifier that always predicts negative outcomes achieved a total accuracy of around 86%, even though it could not classify a single positive record correctly. If a student solely based their findings on total accuracy, they may wrongly interpret this situation as indicating that the problem had been nearly solved.

Using a confusion matrix gives a better overall view of the model's errors. Many times, a model can attain better total accuracy values simply through predicting the majority class correctly, even when many members of the minority class are overlooked. To separate finding the true positives from avoiding false positives, precision and recall must be used. Neither is sufficient alone, and it is important to ensure that both precision and recall are considered together, along with the intended use of the prediction.

The comparison of logistic regression using a limited number of predictors versus using an expanded set of predictors helps demonstrate the importance of predictor choice. The validation AUC score with just age and BMI was over 0.74, while the inclusion of self-reported variables led to an AUC score of about 0.80. Although several measurements can differentiate the categories better than a very small set, this does not mean that additional predictors have revealed any biolog-

ical causes. Regression only reveals patterns within the existing observational survey records.

4.2. What the Threshold Experiment Teaches

The biggest impact on our findings came not from introducing a more complicated machine-learning model but from changing just one number in the decision rule. The model's cutoff point of 0.50 provided a high degree of specificity and accuracy but, in turn, had a very low recall rate. The different cutoff point of 0.10, determined through validation selection, provided a high recall rate but at the expense of lower precision and also lower overall accuracy. The classifier can best be understood as having at least two components, namely a method by which the scores are generated and a decision rule that translates those scores into action.

As can be seen from our class-weighted methodology, the use of weights produced a model that learned differently from the imbalanced data. For the expanded weighted model, the recall at 0.50 was just short of 75%, while the recall for the expanded unweighted model was less than 10%. However, the ROC-AUC values for the two models were quite similar. There may be many different ways in which a training model can be influenced to produce more positive predictions without greatly increasing the model's underlying skill at ranking records from lowest to highest prediction scores. It follows that the use of weighting is not necessarily the best approach solely because it produces a higher recall.

An important limitation of class weighting is that the resulting model scores do not automatically represent calibrated probabilities under the original class distribution. Because class weighting changes the importance of positive and negative records during model fitting, a score of 0.50 from the weighted model should not automatically be interpreted as a 50% probability of belonging to the positive class. Therefore, the comparison at threshold 0.50 demonstrates differences in classification behavior rather than equivalent probability-based decisions. In this study, the weighted model was used to illustrate the effects of class imbalance, not to estimate individual clinical risk.

There is no universally best cutoff point. In some instances, when missing a positive case is very costly and false alarms cost little, a lower cutoff can be advantageous. If there are additional costs or dangers associated with performing unnecessary follow-up action, a higher cutoff may be more appropriate. For the purposes of this study, we chose a recall rate of at least 80% on the validation data to make this mathematical trade-off visible. We purposely did not attempt to identify a cutoff that would reflect the best clinical option as we lacked clinical outcomes data, a validated diagnostic reference standard, and a decision-cost analysis of follow-up care.

Separating the validation and testing data sets was important in this case. The cutoff value was determined based on the validation data while the test set was used for independent evaluation. Although the recall value associated with the test data set, which was 0.8427, was similar to that observed with the validation dataset

cutoff at 0.10, it was still not identical, which is expected given the variation between the data sets used. It is better to be transparent about this discrepancy than to keep changing the rule until the final test results look favorable.

4.3. Limitations and Possible Improvements

The primary limitation is the data source. Instead of analyzing the entire BRFSS survey along with sample weights, strata, and clustering, we used a processed UCI version of the survey data. We did not calculate nationally representative estimates from the data processed in this way. Therefore, calling the percentage of positive records the prevalence of diabetes within United States adults is incorrect. Prediabetes and diabetes are combined in the dependent variable, so performance for either condition alone cannot be determined from the binary variable.

The second limitation is that both survey responses and coded measures can be subject to measurement error and information loss. For example, BMI is only one indicator related to body composition, while age groupings are not sufficiently precise to account for exact ages. The model does not consider every variable which could correlate with diabetes. An association cannot establish which factor came first. Because this is observational data, results cannot be interpreted to indicate whether a change in exercise, food intake, or measured characteristic would lead to a change in diabetes outcome by the amount predicted by this model.

Third, the test results from this study came from a single stratified split, not from repeated evaluation over many splits or external datasets. The results are credible for the chosen held-out records, but results from other splits and other data sources should not be assumed to remain identical. The model was neither calibrated nor clinically validated for patient-level risk estimates. More advanced analyses may involve repeated cross-validation, calibration, or external validation, all of which have been purposefully omitted in this study to keep mathematics and experimental design understandable at a high school level.

A fourth constraint is that we compared a small list of logistic regression models rather than numerous possibilities from various machine-learning algorithms. This was a purposeful educational decision. A wider comparison might improve predictive performance of the models, but this could distract from the main research question that focused on both imbalance and thresholds. We could have also expanded by examining subgroup model performance for age categories or other characteristics, with the appropriate attention paid to fairness and adequate sample sizes. Again, this would have meant more design considerations and subsequent careful interpretations.

Even with the weaknesses mentioned above, there are several strengths of the project. It uses a large publicly available data set, an understandable probability model, an explicit always-negative baseline, separate validation and test sets, and a numerical rule for selecting thresholds. All key figures and tables are reproducible based on the corresponding notebook. Above all, it shows with real data that optimizing one measure will not necessarily optimize another.

5. Conclusion

In this research, we used logistic regression to analyze the UCI CDC Diabetes Health Indicators data set, focusing mainly on the class imbalance and the classification thresholds used. Although the expanded unweighted model was found to have useful ranking capability (with test ROC-AUC of 0.7966), a cutoff point of 0.50 resulted in a detection rate of less than one in ten positive records. When we used 0.10 as the cutoff point, selected on the validation data set, test recall increased to 84.27%, but precision fell to 25.30%, and there was a remarkable increase in false positives. This shows how accuracy alone is not an adequate measure to evaluate imbalanced classifications and that choosing a numerical cutoff point is a practical tradeoff between conflicting errors. These conclusions apply to the processed data and exploratory setup used in this study. They cannot be taken as indicating causes of diabetes or even prevalence of diabetes in the country or a recommended clinical screening rule.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] UCI Machine Learning Repository. CDC Diabetes Health Indicators (Dataset 891). https://archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators

[2] Centers for Disease Control and Prevention. 2015 BRFSS Survey Data and Documentation. https://www.cdc.gov/brfss/annual_data/annual_2015.html

[3] Saito, T. and Rehmsmeier, M. (2015) The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLoS ONE*, **10**, e0118432. https://doi.org/10.1371/journal.pone.0118432

[4] Fawcett, T. (2006) An Introduction to ROC Analysis. *Pattern Recognition Letters*, **27**, 861-874. https://doi.org/10.1016/j.patrec.2005.10.010

[5] Scikit-Learn Developers. LogisticRegression Documentation. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html