COMP 551 - Applied Machine Learning - Assignment1
(Amir, Luna, Suraj)

Abstract

In this project, we investigated the performance logistic regression model on one benchmark dataset, the CDC Diabetes Health Indicators dataset. Our analysis covered various aspects of model performance, including training-test split evaluations, feature weight importance, and the impact of training size and mini-batch configurations. Our findings reveal that choice of hyperparameters is crucial to the performance of models.

Introduction

Machine learning models play a critical role in data analysis and prediction tasks, with logistic regression serving as fundamental techniques for classification problems. Previous research has shown that batch size and learning rate have significant effects on model convergence and performance(Smith, 2018). Our findings add to this understanding by providing practical insights from our experiments.

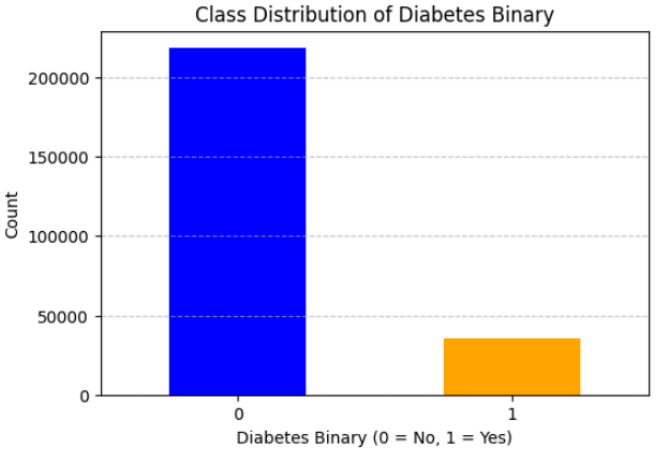
Our logistic regression model achieved 86% accuracy on both training and testing sets but exhibited low recall (0.11) and F1-score (0.18), indicating a bias toward predicting non-diabetic cases due to class imbalance. Feature importance analysis identified High Blood Pressure (HighBP) and Difficulty Walking (DiffWalk) as the most critical predictors, while factors like BMI, Mental Health, and General Health also strongly influenced predictions.

We also found that model performance peaked when trained on 60% of the dataset, with diminishing returns beyond this point. Among different batch sizes, 64 provided the highest accuracy, while 128 yielded the best precision. The optimal learning rate was 0.01, ensuring stable and effective training, whereas a higher rate (0.5) led to significant drops in accuracy and precision due to convergence issues.

Dataset Description and Processing

The CDC Diabetes Health Indicators dataset contains health indicators associated with diabetes and is used for binary classification of diabetes status (`Diabetes_binary`). Preprocessing steps were taken, including filling missing values with the most frequent category for categorical features and mean imputation for numerical features.

An exploratory analysis was conducted to gain insights into the datasets. Class distribution for the diabetes dataset was assessed to check for balance between the binary classes, revealing an imbalance that could affect model performance if not addressed. Descriptive statistics for numerical features were computed, providing insights into the mean, median, and standard deviation.



(Class imbalance)

Experiment Results

The performance of fully batched logistic regression

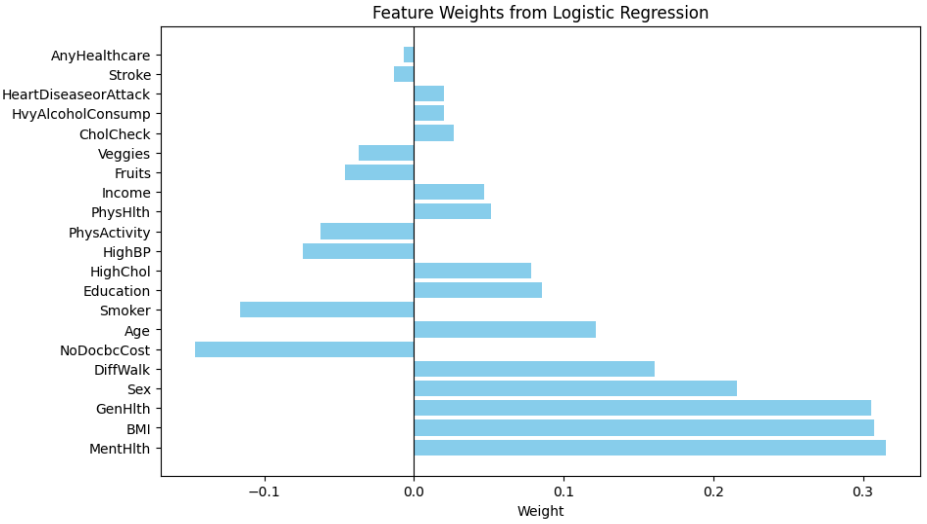
The model achieves 86% accuracy on both the training and testing sets, demonstrating consistency. However, low recall (0.11) and F1-score (0.18) indicate that the model is biased toward predicting non-diabetic cases (negative class, 0) while failing to correctly identify actual diabetic patients (positive class, 1). This issue may stem from an imbalanced class distribution, where negative cases (class 0) significantly outnumber positive cases (class 1). Precision (0.51–0.53) is moderate, but improvements such as class balancing could help enhance overall model performance.

Metrix	Training Set	Testing Set
Accuracy	0.86	0.86
Precision	0.51	0.53
Recall	0.11	0.11
F1-Score	0.18	0.18

Explore the weight of each feature in the fully-batched trained logistic regression model

The bar chart visualizes feature weights from a logistic regression model, showing the impact of each variable on the predicted outcome. Positive weights (right side) indicate features that increase the likelihood of a positive prediction (class 1), while negative weights (left side) make a negative prediction (class 0) more likely. From our chart, we can conclude that:

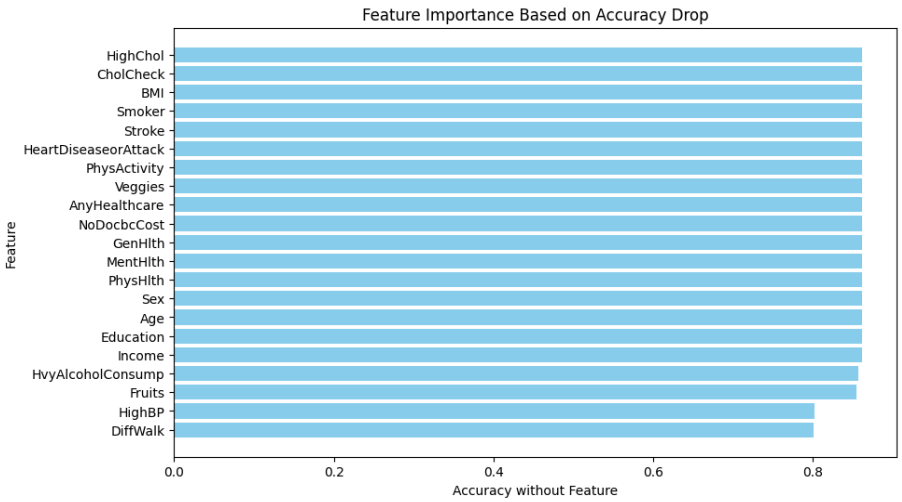
- NoDocbcCost, Difficulty Walking (DiffWalk), Sex, General Health (GenHlth), BMI, Mental Health have the strongest positive influence on predictions.
- Healthy behaviors (e.g., Veggies, Physical Activity, and Fruits) have weaker effects compared to other variables.



Explore the feature importance in the trained logistic regression model

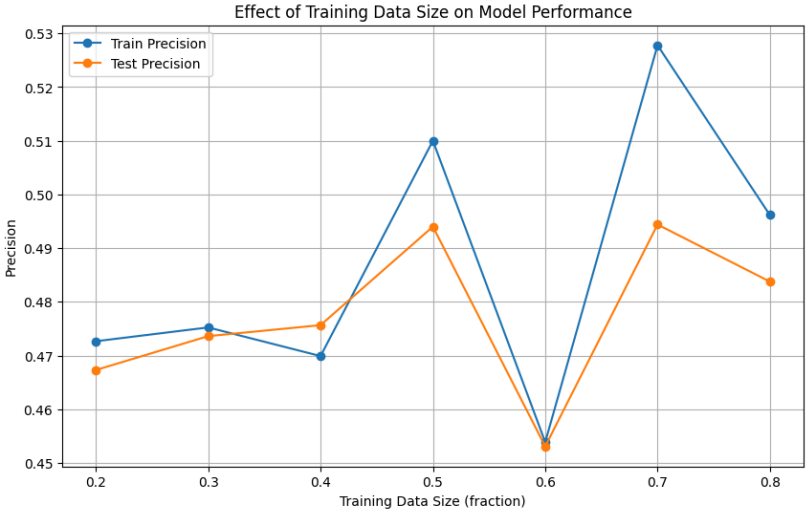
The bar chart illustrates feature importance based on accuracy drop, showing how removing each feature affects the model's performance.From the chart, it shows that HighBP (High Blood Pressure) and DiffWalk

(Difficulty Walking) have the largest accuracy drop when removed, meaning they are the most important features for the model's predictions.



Explore how does size of training data affects the performance of the logistic regression model

According to the graph, the optimal training data size seems to be around 60% of the dataset, whereas 60% yields the lowest precision. At point 60%, both train and test precision peak. Beyond this, precision starts to drop, indicating that increasing the training size further doesn't always enhance performance and might result in diminishing returns.



Explore the effectiveness of different batch sizes on the convergence speed and final model performance.

Among the tested configurations, none exhibited a significant difference in convergence speed, as all reached the maximum iteration limit of 100,000 without the gradient norm falling below a small threshold. However, a batch size of 64 achieved the highest accuracy, while a batch size of 128 yielded the best precision.

Batch Size	Iterations	Accuracy(Testing dataset)	Precision(Testing dataset)
8	100000	86%	53.1%

16	100000	86.3%	53.4%
32	100000	86%	50.7%
64	100000	86.5%	51.5%
128	100000	86%	51.8%

Explore the effectiveness of different learning rates on logistic regression model's performance

In this part, we tested three different learning rates (0.01, 0.1, 0.5) and trained the models with different batch sizes (8, 16, 32, 64, 128) to determine the most suitable learning rate for training our machine learning model. From our experiment, we found that a learning rate of 0.01 yielded the best accuracy and precision, indicating that it provides stable and effective updates during training. In contrast, a higher learning rate (0.5) resulted in noticeable drops in accuracy and precision, suggesting that the model struggled to converge properly.

Learning Rate	Batch Size	Iterations	Accuracy(Testing Dataset)	Precision(Testing dataset)
0.01	8	100000	86.32%	53.85%
0.01	16	100000	86.30%	53.16%
0.01	32	100000	86.35%	53.86%
0.01	64	100000	86.34%	53.96%
0.01	128	100000	86.33%	53.48%
0.1	8	100000	86.29%	53.13%
0.1	16	100000	86.20%	51.99%
0.1	32	100000	86.28%	52.98%
0.1	64	100000	86.32%	52.85%
0.1	128	100000	86.33%	53.61%
0.5	8	100000	85.48%	45.23%
0.5	16	100000	83.18%	40.43%
0.5	32	100000	86.17%	51.46%
0.5	64	100000	86.08%	50.30%
0.5	128	100000	86.21%	51.73%

Reference

Smith, L. N. (2018). A disciplined approach to neural network hyper-parameters: Part 1--learning rate, batch size, momentum, and weight decay.
https://arxiv.org/abs/1803.09820