Report: Build a Logistic Regression Model for Diabetes Prediction

Github Link for repository:
https://github.com/Bimsarasmp/Diabetes_prediction_using_classification_model

1. Key Findings

- Dataset Overview

  Dataset Characteristics: The dataset consists of various features related to health metrics, and the target variable is binary, indicating the presence or absence of diabetes (Outcome=1 or 0).

- Dataset Structure:

  The dataset has a shape of (768, 9) with no null values or duplicate entries.

- Handling Missing Values:

  An issue arises as certain numerical values, like Blood Pressure (BP), skin thickness, etc., cannot realistically be zero. To address this, missing values are replaced with the mean to ensure a more accurate representation of the data.

- Data Preprocessing

  Oversampling: Oversampling is performed to address the imbalance in the dataset, particularly to enhance the prediction of true positives (TP) and improve precision.

- Correlation Analysis

  Correlation Analysis: Certain features, such as Glucose, Age, and BMI, show notable correlations with the target variable, providing initial insights into potential predictors.

2. Model Performance Metrics

1. Logistic Regression (Without SMOT)

- Accuracy: 0.766

- Precision:0.686

- Recall:0.636

- F1 Score:0.660

Classification Report

	precision	recall	f1-score	support
0	0.81	0.84	0.82	99
1	0.69	0.64	0.66	55
accuracy			0.77	154
macro avg	0.75	0.74	0.74	154
weighted avg	0.76	0.77	0.76	154

Confusion Matrix

[[83 16]

[20 35]]

2. Logistic Regression (With SMOT)

- Accuracy: 0.765

- Precision:0.755

- Recall: 0.792

- F1 Score: 0.773

Classification Report

	precision	recall	f1-score	support
0	0.78	0.74	0.76	99
1	0.75	0.79	0.77	101
accuracy			0.77	200
macro avg	0.77	0.76	0.76	200
weighted avg	0.77	0.77	0.76	200

Confusion Matrix

[[73 26]

[21 80]]

3. Logistic Regression (C=0.01)

- Accuracy: 0.765

- Precision: 0.760

- Recall: 0.782

- F1 Score: 0.771

Classification Report

	precision	recall	f1-score	support
0	0.77	0.75	0.76	99
1	0.76	0.78	0.77	101
accuracy			0.77	200
macro avg	0.77	0.76	0.76	200
weighted avg	0.77	0.77	0.76	200

Confusion Matrix

[[74 25]

[22 79]]

4. SGD Classifier (CV=2)

- Accuracy: 0.7275

- Precision: 0.688

- Recall: 0.712

- F1 Score: 0.700

Cross-Validated Metrics

Accuracy of 2-folds: [0.73 0.725]

Confusion Matrix:

[[272 129]

[115 284]]

Cross-Validated Precision: 0.6877

Cross-Validated Recall: 0.7118

Cross-Validated F1 Score: 0.6995

Observations:

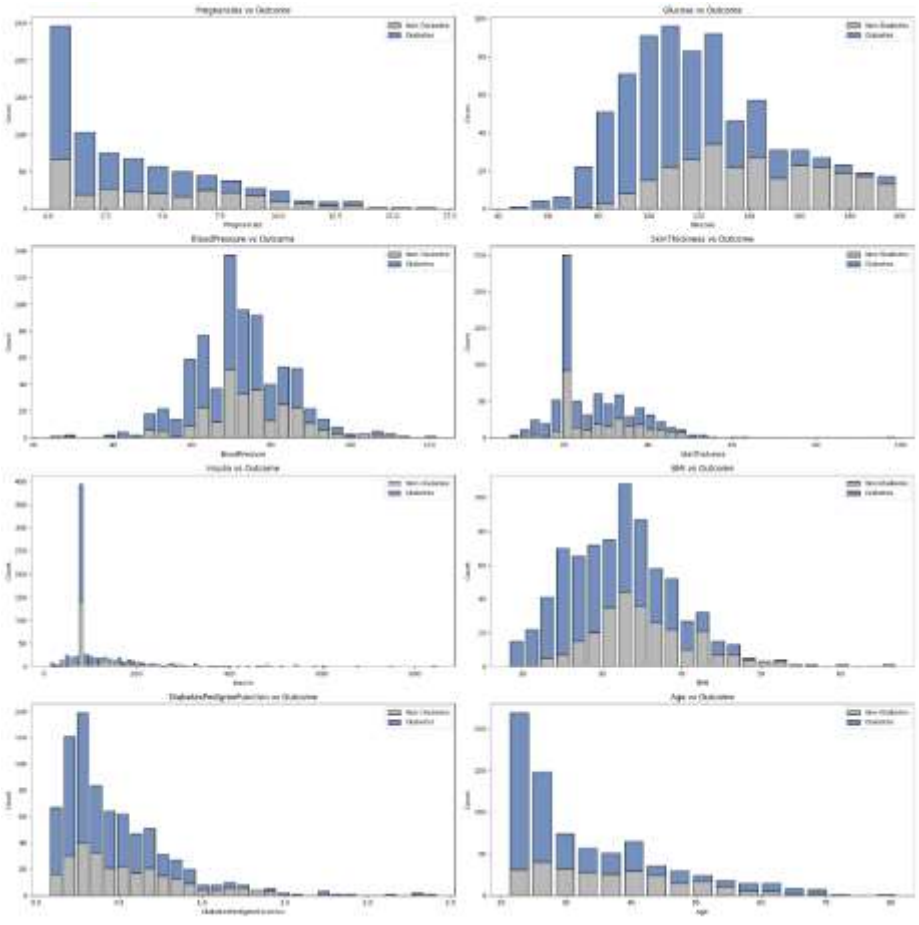
- Best Precision: Logistic Regression with SMOT (0.755)
- Best Recall: Logistic Regression with SMOT (0.792)
- Best F1 Score: Logistic Regression with SMOT (0.773)
- Best Accuracy: Logistic Regression without SMOT (0.766)

3. Insights from Model Coefficients

The logistic regression model coefficients provide insights into feature importance:

- Positive Coefficients: Features with positive coefficients contribute positively to the likelihood of diabetes. Notable contributors include Glucose, Age, and BMI.
- Negative Coefficients: Features with negative coefficients have an inverse relationship with the likelihood of diabetes. Further analysis of these features may provide insights into protective factors.

Following is the summary of how each feature value is associated with a higher risk of diabetes based on the histograms:



1. Pregnancies:

- As the number of pregnancies increases, the likelihood of diabetes also rises. More pregnancies correlate with a higher risk of diabetes.

2. Glucose Levels:

- Higher glucose levels are strongly associated with an increased number of diabetic cases. Elevated glucose significantly raises the risk of diabetes.

3. Blood Pressure:

- While blood pressure doesn't show a clear trend, higher values still correlate with more cases of diabetes. Monitoring blood pressure is essential.

4. Skin Thickness:

- Skin thickness doesn't exhibit a distinct pattern, but there's a slight increase in diabetic cases at higher skin thickness levels.

5. Insulin Levels:

- Higher insulin levels are observed in both non-diabetic and diabetic individuals. Extreme insulin values are more common in diabetics.

6. Body Mass Index (BMI):

- A BMI above 30 (indicating obesity) shows an increased number of diabetic individuals. Obesity is a significant risk factor.

7. Diabetes Pedigree Function:

- While not as prominent as other features like glucose or BMI, higher values of the diabetes pedigree function correlate with an increased risk of diabetes.

8. Age:

- Older age groups, especially those above 40, have a significantly higher occurrence of diabetes.

4. Conclusion

Precision is a crucial metric, especially in the context of predicting diabetes, as false positives can have significant consequences. Among the models, Logistic Regression with SMOT stands out with the highest precision (0.755), closely followed by Logistic Regression with C=0.01 (0.760). These models show better performance in minimizing false positives, making them particularly relevant in a healthcare context. Consideration of precision, along with other metrics, should guide the choice of the final model for practical implementation.

By further analyzing ROC-AUC curve for Logistic Regression with C=0.01 (0.760).

ROC Fold 1 (AUC = 0.84):

- It achieves an Area Under the Curve (AUC) of 0.84, indicating good discrimination between positive and negative classes.

ROC Fold 2 (AUC = 0.82)

- Its AUC is 0.82, showing reasonable performance.

ROC Fold 3 (AUC = 0.80)

- It achieves an AUC of 0.80, indicating acceptable discrimination.

Micro-Average ROC

- The dashed black line labeled "Micro-average ROC (AUC = 0.82)" represents the overall performance across all folds.

- The micro-average considers all instances and computes a single ROC curve by aggregating the true positive and false positive rates from each fold.

Random Classifier Baseline

- The dashed diagonal line from the bottom left corner to the top right corner represents the expected performance of a random classifier.

- Any curve above this baseline indicates better-than-random performance.

In summary, the model demonstrates reasonable discrimination ability, with AUC values ranging from 0.80 to 0.84 across different folds.