Diabetic Retinopathy Screening Model

BIS568: Applied Machine Learning in Healthcare
Final Project

Jiaye Chen, Luning Yang, Mengmeng Du, Yawen Wei

Fall 2023

Introduction

Diabetes stands as a widespread and significant public health concern, affecting a substantial portion of the population. According to data from the Centers for Disease Control and Prevention (CDC), approximately 38.4 million people in the United States have diabetes. This pervasive prevalence underscores the critical need to address diabetes as a widespread public health challenge comprehensively. Beyond the sheer numerical scale, diabetes is intricately linked to severe health complications, such as cardiovascular diseases and vision impairment, creating a substantial economic burden on both individuals and healthcare systems.

Among the myriad complications associated with diabetes, diabetic retinopathy takes center stage as a prevalent eye condition and a primary cause of blindness in working-age adults. This visually threatening complication affects roughly one-third of diabetics aged 40 and above, with a global prevalence estimated at 35.4%. Despite being the foremost cause of blindness in U.S. adults aged 20 to 74, recent studies show that the promising prospect of early detection and treatment can effectively mitigate the impact of diabetic retinopathy, preventing significant vision loss. Consequently, timely intervention emerges as a pivotal factor in preserving eyesight. Hence, delving into the early detection or prediction of diabetic retinopathy becomes imperative.

In our project, we aim to employ machine learning techniques to predict the occurrence of diabetic retinopathy in individuals with diabetes, utilizing their historical health records. This endeavor aims to offer valuable insights for screening purposes to both patients and healthcare providers regarding the likelihood of diagnosing diabetic retinopathy, eliminating the necessity for additional eye examinations or related tests. This not only streamlines the process but also reduces costs for individuals undergoing diabetes management.

Data Preparation and Preprocessing

Firstly, we define our target cohort as diabetic patients with diabetic retinopathy (DR) and the control cohort as diabetic patients without any complications (Non-DR) to predict the possibility of diabetic retinopathy between diabetic patients. In constructing the prediction model, we initiated the process by identifying 14 features that show a high correlation with the predictors (Ogunyemi et al., 2021). These features are divided into three groups, including demography, patient conditions, and clinical measurements, which are shown in **Table 1**. Subsequently, we extracted the corresponding data from the MIMIC database. And the Non-DR has 1160 samples, while the DR has 133 samples.

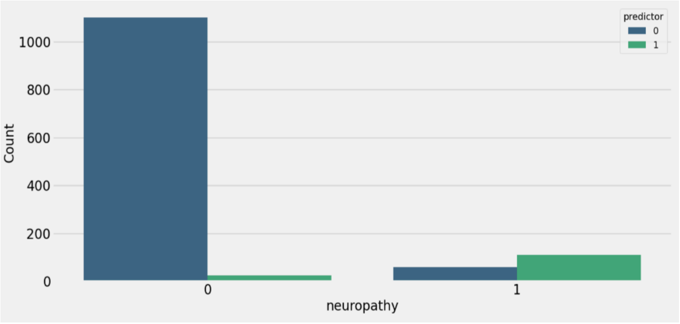
Table 1: Initial Features

Demography	gender	Patient gender
	race	Patient race

	age	Patient age
Conditions	Duration days	Diabetes/Diabetic retinopathy duration
	neuropathy	Neuropathy
	stroke	Stroke
	Insulin dependence	Insulin-dependent diabetes mellitus
	nephropathy	Nephropathy
Clinical measurement	triglyceride	Triglycerides
	hemo	Hemoglobin
	hemo_a1c	Hemoglobin A1C
	bun	Blood urea nitrogen
	sbp	Systolic blood pressure
	dpb	Systolic blood pressure
Predictor	predictor	Non-DR/ DR

After completing data collection, the initial step involves univariate analysis, encompassing the creation of bar plots for categorical variables and histograms for numeric variables. During this process, it was observed that `nephropathy` contained only one unique value and was subsequently eliminated due to their lack of informativeness. Additionally, outliers were identified in `age` (values exceeding 300) and in `Systolic Blood Pressure` and `Diastolic Blood Pressure` (values equal to 0). Subsequently, these outliers were removed.

Moving forward, bivariate analysis was conducted by generating box plots to explore relationships between predictors and numeric variables. Noteworthy insights were obtained, as illustrated in **Figure 1a** and **Figure 1b**. In **Figure 1a**, a strong correlation between `neuropathy` and the `predictor` was observed, indicating a higher prevalence of DR in patients with neuropathy. Surprisingly, **Figure 1b** revealed that individuals with DR were generally younger than those without, suggesting a higher likelihood of DR in younger diabetic patients.



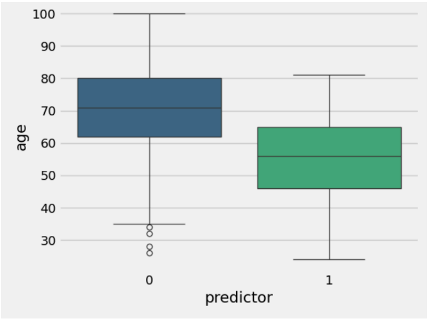


Figure 1a Bar plot of predictor and neuropathy.

Figure 1b Box plot between the predictor of age.

Following data exploration and outlier removal, the next step involves handling missing values. For values missing less than 20%, we used random number imputation. Features with missing values exceeding 20%, such as \`triglyceride\` and \`hemo_a1c\`, were dropped. Subsequently, one-hot encoding was applied to categorical features, resulting in a final set of 15 features.

Model Development and Evaluation

The cleaned data contains 15 attributes and 1 predictor among 1293 patients, which are split 7:3 in a stratified way. Then, to oversample the minority class, the \`SMOTE\` library is applied to resample the negative and positive samples, respectively, to a balanced number of 852 for training. The testing dataset has 348 negative samples and 40 positive samples. Then, three machine learning methods are trained, namely Logistic Regression, Random Forest, and XGBoost. Certain parameters are tuned by grid search. For example, constant C, n_estimators, max_depth, and learning rate. The models are then evaluated by the methods of ROC-AUC, PR-AUC, and the confusion matrix.

Table 2 shows the training and testing results. Random Forest gains the highest cross-validation accuracy of 96% and testing accuracy of 95.26%. Three models are further evaluated by the ROC-AUC, PR-AUC, and calibration curve. As a result, the Random Forest model shows the highest model performance, followed by XGBoost with the same high ROC-AUC of 98% but a slightly low PR-AUC of 0.87%. **Figure 2a** shows the calibration curve, and **Figure 2b** shows the result of the confusion matrix for the three models. It is observed that Logistic Regression has a relatively high misclassification rate for false positives (43), indicating poor performance among the three models. When comparing the other two models, XGBoost is regarded as having a more balanced performance regarding controlling false negatives (10) and tolerating false positives (13), which aligns with the main purpose of the DR screening.

Table 2: Result for Model selection, Parameter Tuning and Model Evaluation

Model	**Best CV Acc**	**Tuned Parameters**	**ROC-AUC**	**PR_AUC**
Logistic	**0.8362**	**{'C': 0.01}**	**0.92**	**0.74**

Regression				
Random Forest	**0.9600**	**{'max_depth': 20, 'n_estimators': 200}**	**0.98**	**0.88**
XGBoost	**0.9526**	**{'learning_rate': 0.2, 'max_depth': 3, 'n_estimators': 50}**	**0.98**	**0.87**

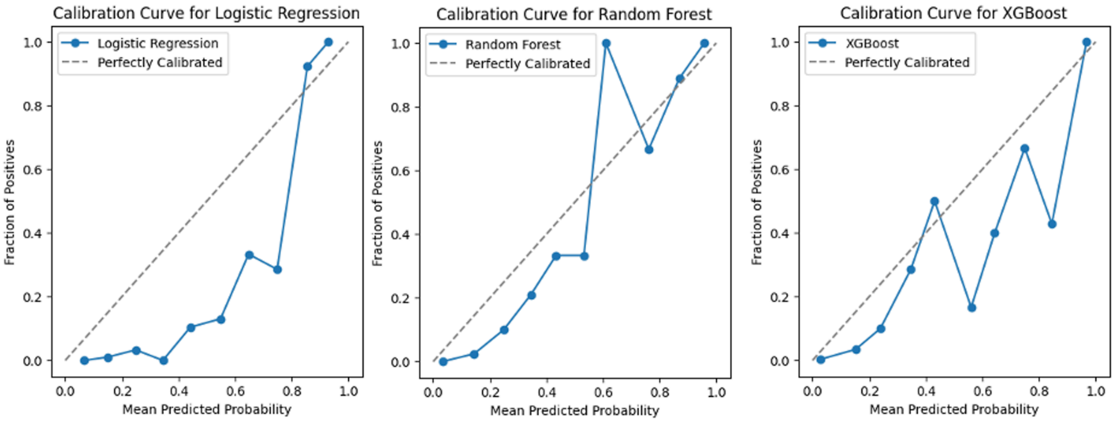


Figure 2a

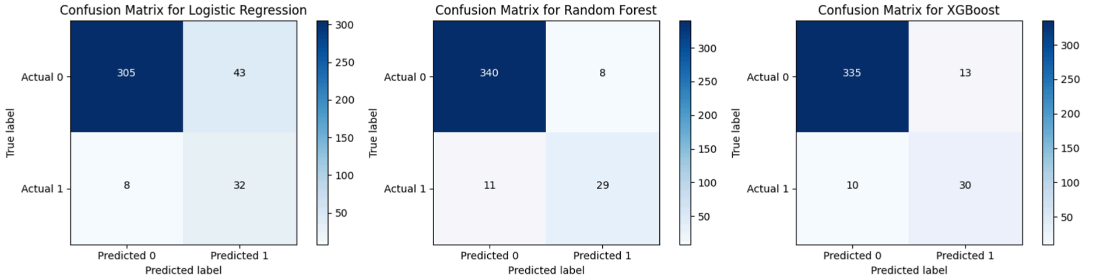


Figure 2b

Interpretability

This beeswarm plot (**Figure 3a**) provides a visual summary of the top 20 variables affecting the model's output, allowing us to discern which features generally have the most significant impact on the model's predictions and how the variability of these features might interact with the predictions. The feature `insulin_dependence` has a noticeable cluster of dots far to the right, indicating that higher insulin dependence generally contributes to an increase in the model's predictions. The feature `race_WHITE` has dots scattered across both sides of zero but with a concentration on the positive side, suggesting that being identified as white often but not always contributes to an increase in the model's predictions. Similar to `race_WHITE`, the feature `neuropathy` shows a spread on both sides of zero, with a slight skew towards the positive, indicating that the presence of neuropathy can both increase or decrease the model's predictions, depending on the context. There is a noteworthy point that the correlation matrix showed that `insulin_dependence` and `neuropathy` would have higher correlations with our predictor.

However, based on the XGBoost model, `race_white` has higher predictive power than `neuropathy`. We have identified several reasons based on the SHAP dependence plot for insulin dependence below.

The SHAP plot shows the interaction effect between `insulin_dependence` and `race_WHITE` (**Figure 3b**). We assume that if the oversampling method replicated minority cases with `race_WHITE` more frequently, the model might learn an artificial association between these characteristics and the outcome. Besides, since the majority of our data represents one racial group, the `race_white`, the model may perform better for that group simply because it has more data to learn from. This means that the predictive features the model identifies could be more reflective of the patterns found in the white population than the population as a whole. Additionally, when we oversample and the technique used does not account for race distribution, the model may be increasing the representation of DR cases that are also predominantly white. This can lead to an artificial inflation of the importance of race-related features, such as `race_WHITE` in the prediction model.

A downward trend is observed in the SHAP dependence plot for the feature `age` (**Figure 3c**), indicating that smaller values of age contribute to higher SHAP values, which, in the context of the model's predictions, could imply a lower likelihood or intensity of the outcome being predicted. This trend becomes more pronounced for individuals over the age of 70, suggesting that the model associates increasing age with a diminishing effect on the prediction. Additionally, the color coding by 'insulin_dependence' shows a predominance of lower insulin dependence (blue dots) associated with the more negative SHAP values at older ages. This interaction highlights that for older individuals, the model's prediction is more strongly influenced by age, particularly for those with lower insulin dependence.

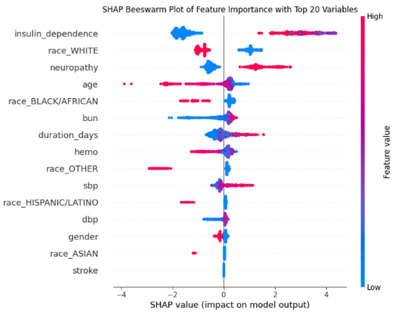


Figure 3a

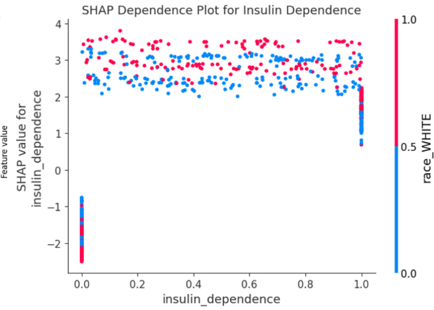


Figure 3b

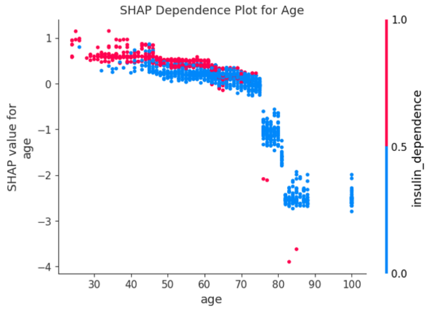


Figure 3c

Discussion

In our study, we identified potential biases in the data, which might not accurately reflect the typical frequency of Diabetic Retinopathy (DR) when compared to national statistics. Specifically, our data, derived from the MIMIC dataset, indicates a prevalence of 1.47% for DR among individuals with diabetes, which contrasts with the 4.04% local adjusted prevalence reported by the CDC for Suffolk County in 2021. This discrepancy arises as our data is confined to cohorts in the Emergency Room (ER) department of a hospital, representing only a segment of the patient population in this area.

Further, we addressed the challenge of small sample size in our positive cohort (133 DR vs. 1160 Non-DR) by applying oversampling to our training data before model development. This increased the positive cohort to 852 DR cases. However, this technique introduced specific biases:

1. **Overfitting to the Minority Class (DR):** By equating the number of DR samples with Non-DR samples, the model risks being overly fine-tuned to the characteristics of the oversampled DR cases, potentially undermining its ability to generalize to new DR cases or different datasets.
2. **Repetition Bias:** The model might learn specific instances of these replicated DR cases rather than generalizable underlying patterns, leading to memorization rather than true learning.
3. **Validation/Test Data Discrepancy:** The original imbalance is maintained in the validation and test datasets, which are not oversampled. This could lead to the model performing differently on training data compared to validation/test data, making its training performance metrics potentially unreliable indicators of real-world performance.
4. **Underperformance on Non-DR Cases:** While focusing on DR cases might improve performance for this class, it could lead to reduced performance on Non-DR cases, especially if the model erroneously classifies ambiguous cases as DR to match the training data distribution.

To counter these issues, our evaluation plan incorporates a multi-tiered strategy. We have adopted stratified cross-validation to maintain an equal proportion of target classes across each fold, providing a robust foundation for assessing our model's performance metrics. Our focus will shift towards external validation, aiming to acquire datasets from diverse sources to rigorously evaluate the model's generalizability. Prior to real-world deployment, the model will be reviewed by an ethical oversight board, scrutinizing its intended applications, performance outcomes, and deployment implications, with a focus on ethical considerations. We plan to engage a broad spectrum of stakeholders for their perspectives, especially those potentially affected by the model's predictions, to enhance our understanding of the model's fairness and performance impact. Upon deployment, we will establish a continuous monitoring framework to oversee the model's performance and fairness metrics, incorporating a feedback loop for integrating new data and facilitating periodic updates and re-evaluations to adapt to evolving real-world dynamics.

Conclusion

In summary, our study leverages machine learning to predict diabetic retinopathy, a critical diabetes complication. Notably, XGBoost surpasses Logistic Regression and Random Forest, showcasing a more balanced performance. We place a premium on interpretability, acknowledging gaps in data representation and biases introduced during oversampling. Furthermore, our multi-tiered evaluation strategy ensures robustness, and continuous monitoring post-deployment reflects our commitment to real-world adaptability. Finally, this research represents a stride towards proactive healthcare management and the seamless integration of machine learning, with a central focus on enhancing patient outcomes and public health.

References

Ogunyemi, Omolola I, et al. “Detecting diabetic retinopathy through machine learning on electronic health record data from an urban, safety net healthcare system.” *JAMIA Open*, vol. 4, no. 3, 2021, https://doi.org/10.1093/jamiaopen/ooab066.

“Diabetes Quick Facts.” *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 4 Apr. 2023, www.cdc.gov/diabetes/basics/quick-facts.html#:~:text=About%2038%20million%20people%20in,(and%20may%20be%20underreported).

“Diabetic Retinopathy Rule of Thirds Guides - Centers for Disease ...” *Centers for Disease Control and Prevention*, National Center for Chronic Disease Prevention and Health Promotion, www.cdc.gov/visionhealth/pdf/factsheet.pdf. Accessed 14 Dec. 2023.

“Diabetic Retinopathy Estimates.” Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, 13 June 2023, www.cdc.gov/visionhealth/vehss/estimates/dr-prevalence.html.