Diabetes Diagnosis

Zhirui Li
Data Science Institute @ Brown
Dec 04, 2023
https://github.com/ZhiruiLi1/Diabetes_Diagnosis.git

Outline

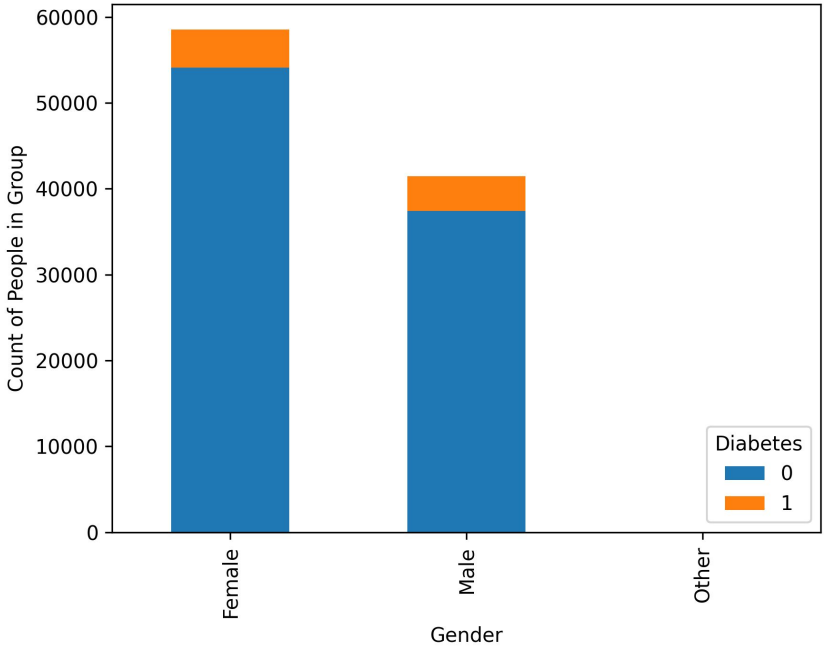
- Background, EDA, Preprocessing
- Cross Validation Pipeline
- Results
- Outlook

Background:

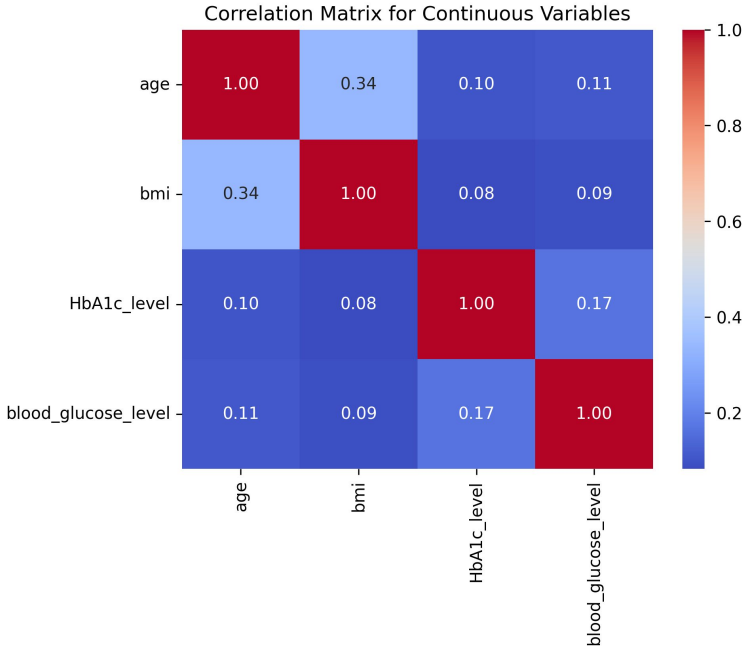
- Approximately 463 million adults are living with diabetes; by 2045 this will rise to 700 million (Source: International Diabetes Federation).
- Diabetes is a major health concern worldwide, leading to significant healthcare challenges and a need for effective management and early detection strategies.
- I developed machine learning models to predict whether a patient has diabetes (binary classification) using the dataset from Kaggle.

EDA

Categorical Variable: Gender

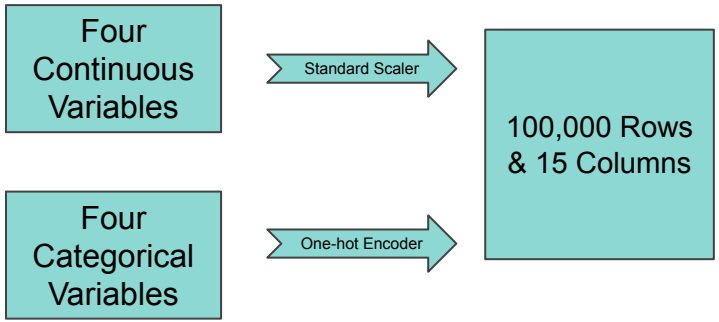


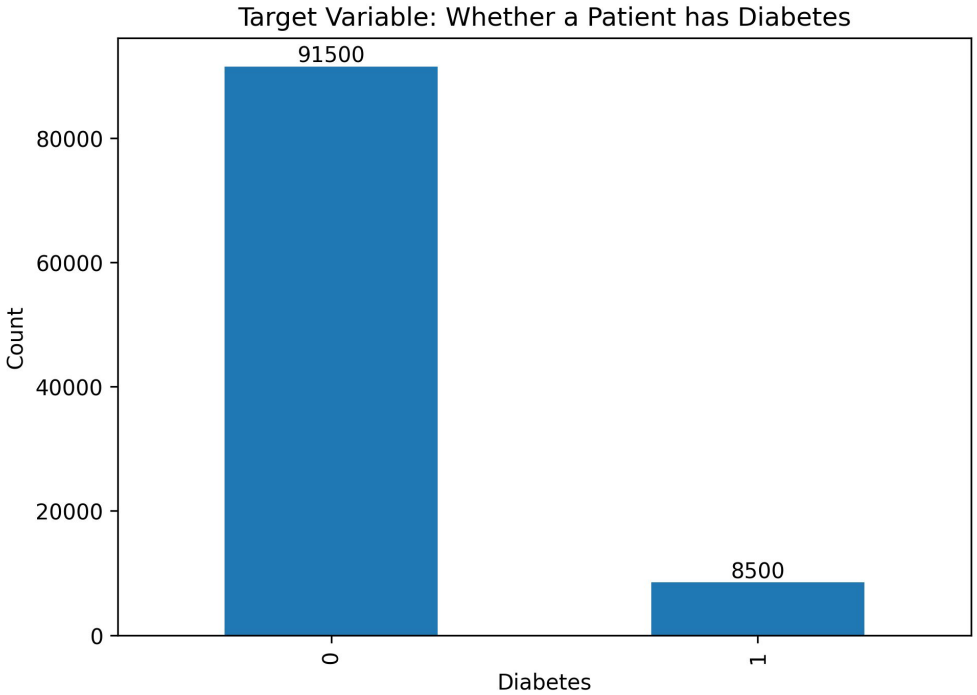
Continuous Variables



Preprocessing

- Imbalanced Dataset
  - Stratified Split
  - F2-beta Score





Cross Validation Pipeline

- **Random State Selection**: Choose five distinct random states to ensure reproducibility and robustness in the model's performance.
- **Data Splitting**: Employ stratified splitting to split the dataset into training (60%), validation (20%), and test (20%) sets, ensuring proportional representation of class 1 samples across each subset.
- **Preprocessing**: Apply standard scaling to normalize numerical features and use one-hot encoding to convert categorical variables into a format suitable for model training.
- **Hyperparameter Tuning**: Conduct an exhaustive grid search across various hyperparameter combinations to find the configuration that yields the best performance on the validation set.
- **Model Evaluation**: Finalize the optimal model based on grid search results and evaluate its effectiveness on the previously unseen test set to assess its generalization capability.

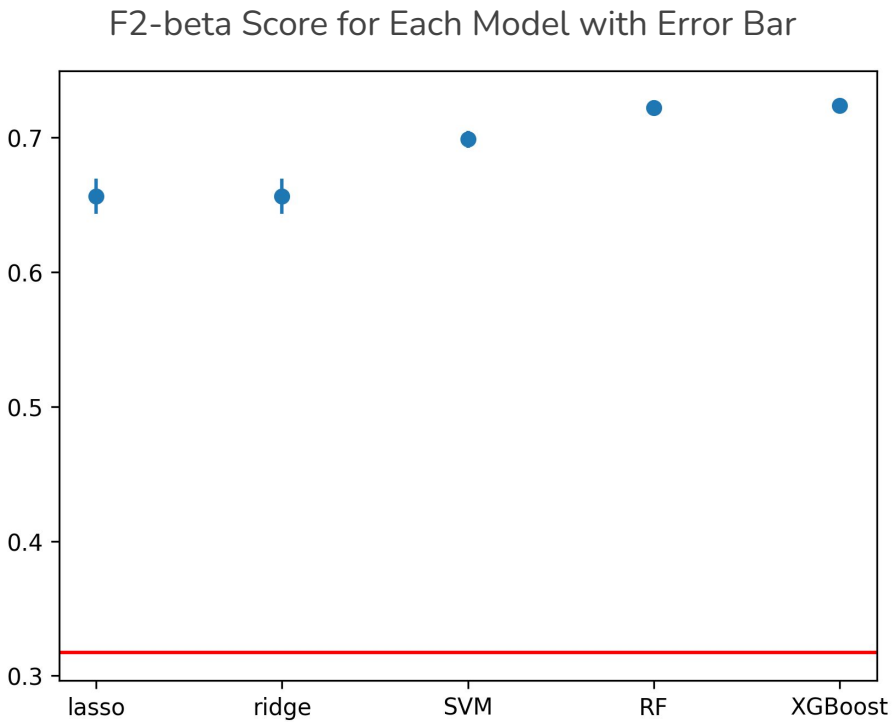
Parameter Tuning

- Models Chosen:
  - Lasso Logistic Regression
  - Ridge Logistic Regression
  - Random Forest
  - Support Vector Classifier
  - Extreme Gradient Boosting

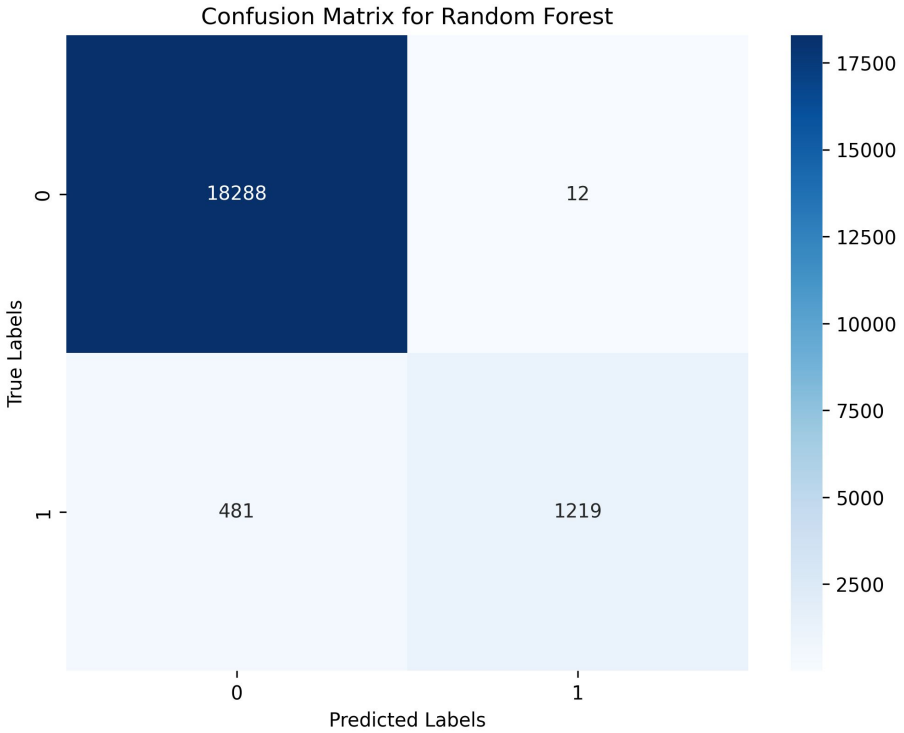
Algorithm	Hyperparameters
Lasso Logistic Regression	C: [0.001, 0.0316, 1.0, 31.6228, 1000.0]
Ridge Logistic Regression	C: [0.001, 0.0316, 1.0, 31.6228, 1000.0]
Random Forest	max_depth: [1, 2, 5, 10, 15]
	max_features: [1, 2, 5, 10, 15]
SVC	C: [0.01, 0.1, 1, 10]
	gamma: [0.01, 0.1, 1, 10]
XGBoost	learning_rate: [0.01]
	n_estimators: [1000]
	gamma: [0, 0.1, 0.3, 0.5]
	max_depth: [3, 6, 10]
	min_child_weight: [1, 5, 9]
	colsample_bytree: [0.5, 1]
	subsample: [0.5, 1]

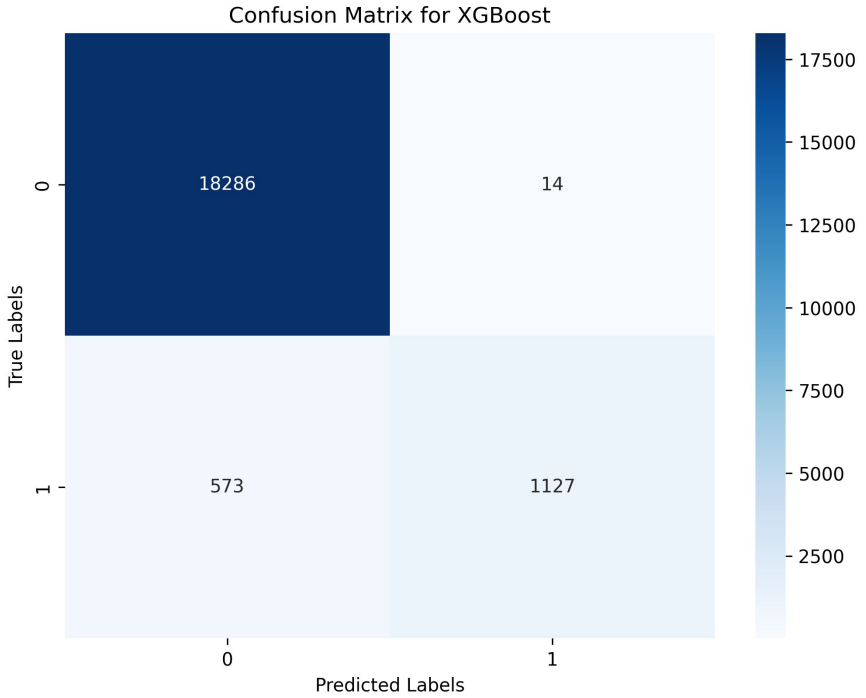
Results

	Model	F2 Beta Score (Mean ± Std)
0	Lasso Logistic Regression	0.656 ± 0.015
1	Ridge Logistic Regression	0.656 ± 0.015
2	SVM	0.699 ± 0.007
3	Random Forest	0.722 ± 0.006
4	XGBoost	0.724 ± 0.006
5	Baseline	0.317



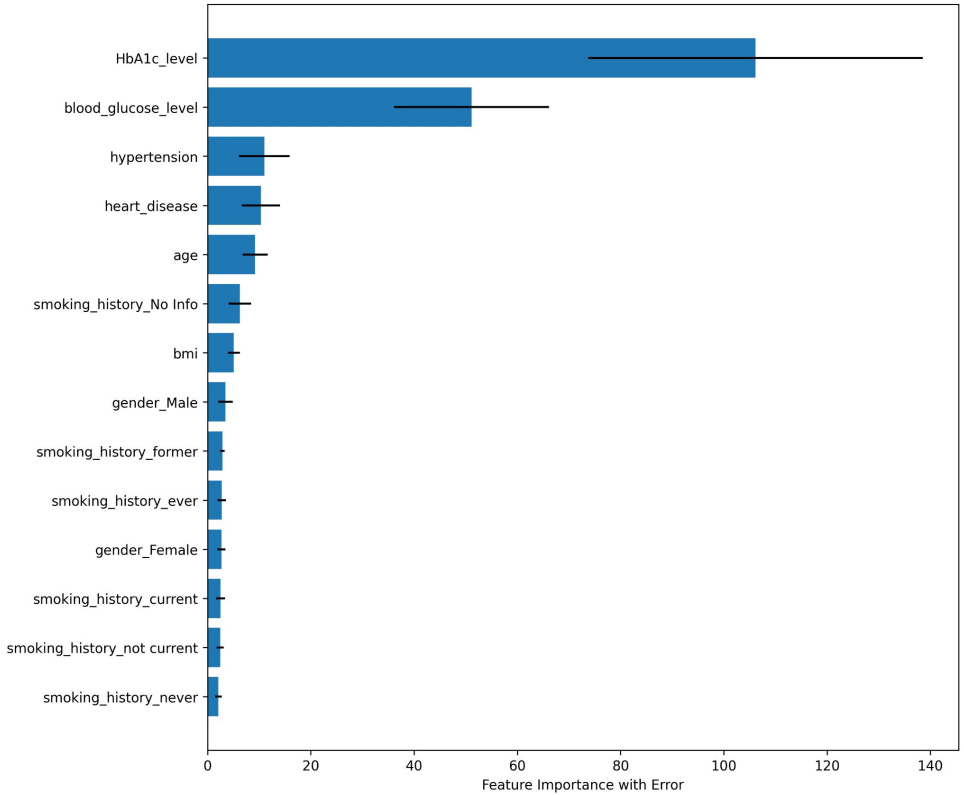
Results Continue



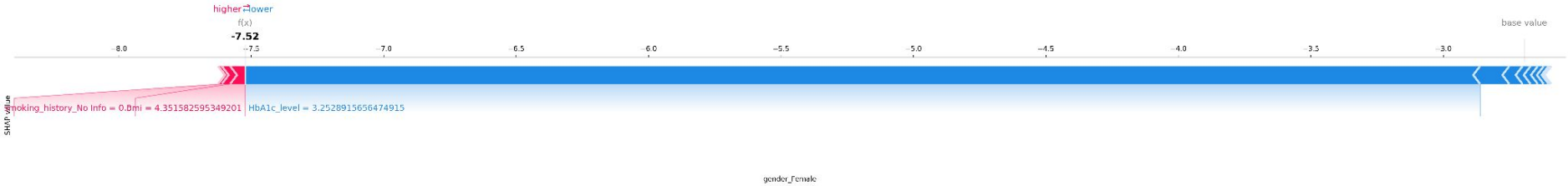


Global Feature Importance

- HbA1c_level is the most influential feature when predicting whether a patient has diabetes using XGBoost.



Local Importance



- XGBoost is confident that this individual does not have diabetes.
- 'HbA1c_level' significantly contributes to a lower prediction score, suggesting that the patient's 'HbA1c_level' feature is contributing towards a prediction of not having diabetes.

Outlook

- **Expand Dataset:** Increase dataset size to enhance model accuracy and generalization.
- **Enhanced Hyperparameter Tuning:** Explore a wider range of hyperparameters for optimal model performance.
- **Advanced Models:** Experiment with complex models such as neural networks for potentially better predictions.
- **Interpretability:** Conduct more literature reviews or consult with domain experts to understand what each feature represents.

Q & A

THANK YOU!