Supervised Learning – Pima Indians Diabetes: Comparative Analysis of Tabular Models

Introduction

Developed learning pipeline using three tabular supervised learning models, (XGBoost, LightGBM, and Catboost), to predict diabetes risk using the famous Pima Native Indian Diabetes Database from UCI.

- The pretrained models take in eight medical/physical features (e.g. BMI, blood glucose, insulin), and output a binary prediction of a patient's diabetes diagnosis.
- Implemented data preprocessing and data cleaning using the Pandas library. Used a 60-20-20 training, validation, test split for robust model evaluation (done using train_test_split from scikit-learn)
- To display a performance evaluation for each model, I used matplotlib and generated ROC curves, AUC scores, and Youden's J plots, which helped visualize how well the models distinguish between diabetic and non-diabetic test cases.
- Compared binary cross-entropy loss data, testing accuracy, and evaluation via ROC curves to assess optimal decision thresholds and classification performance across models.
- Demonstrated a need for generalization in supervised learning through the drastic differences in training and testing accuracy for low-performing models.
- Identified the best-performing model based on test accuracy, achieving the highest predictive performance for diabetes classification (CatBoost)

Dataset Distribution

- **Train Set:** 60%
- **Validation Set:** 20%
- **Test Set:** 20%

The dataset is loaded from an online CSV and split into training, validation, and test sets using scikit-learn's train_test_split.

Dataset Reference: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

Hyperparameters

- **XGBoost / LightGBM / CatBoost:**
    - Default parameters are used, with random seed set for result reproducibility
    - For CatBoost, early stopping rounds are 10 (special parameter)

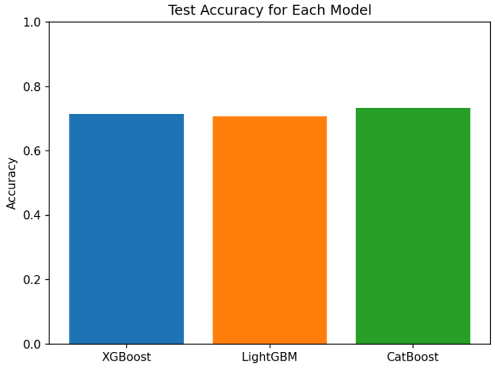
The models were trained using entire feature set (8 numeric features) w/ out image conversion

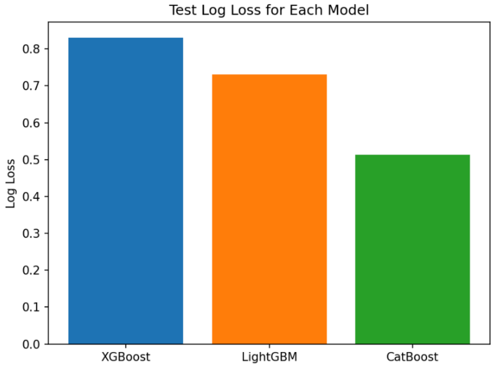
Implemented this project to apply my ML knowledge to structured tabular data

Relevant Libraries

- Pandas (Structured Data Parse)
- Numpy
- XGBoost (Tabular Model 1)
- LightGBM (Tabular Model 2)
- CatBoost (Tabular Model 3)
- SKLearn (ROC, AUC, StandardScaler, TrainTestSplit)
- Matplotlib

XGBoost, LightGBM, and CatBoost Numerical Results



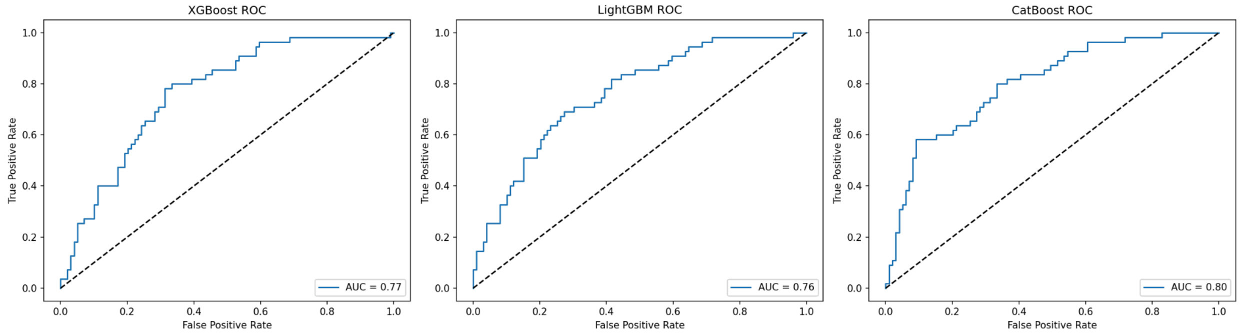


XGBoost Test Accuracy: 0.7143, Log Loss: 0.8313

LightGBM Test Accuracy: 0.7078, Log Loss: 0.7309

CatBoost Test Accuracy: 0.7338, Log Loss: 0.5137

ROC Curves

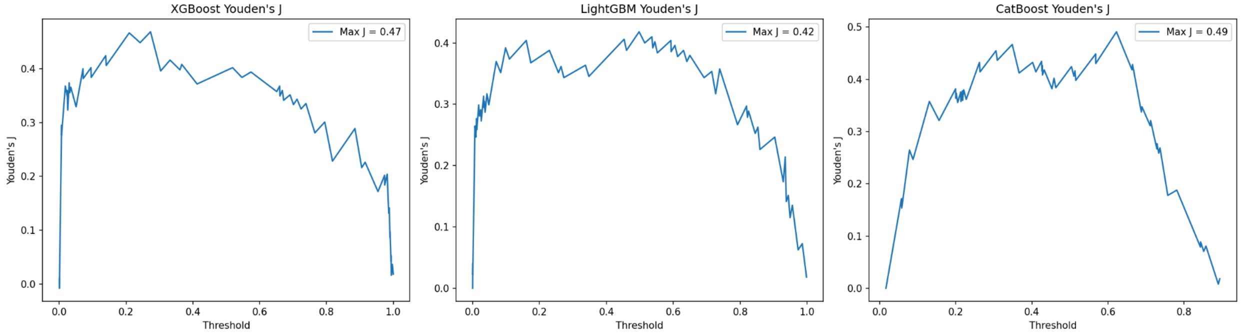


AUC XGBoost - 0.77

AUC LightGBM - 0.76

AUC CatBoost - 0.80

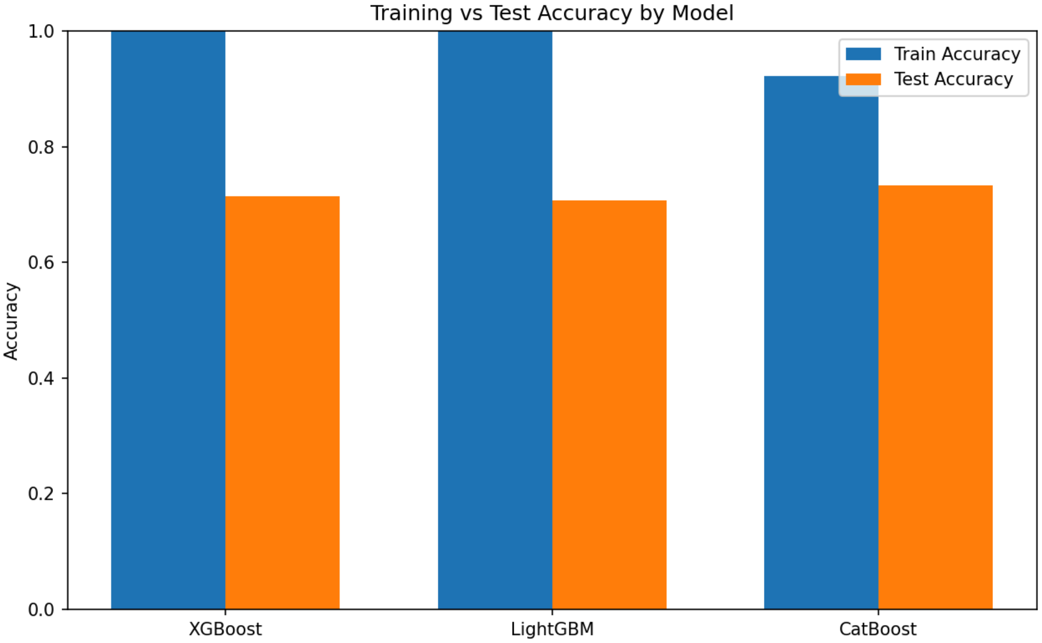
Youden’s J Plots

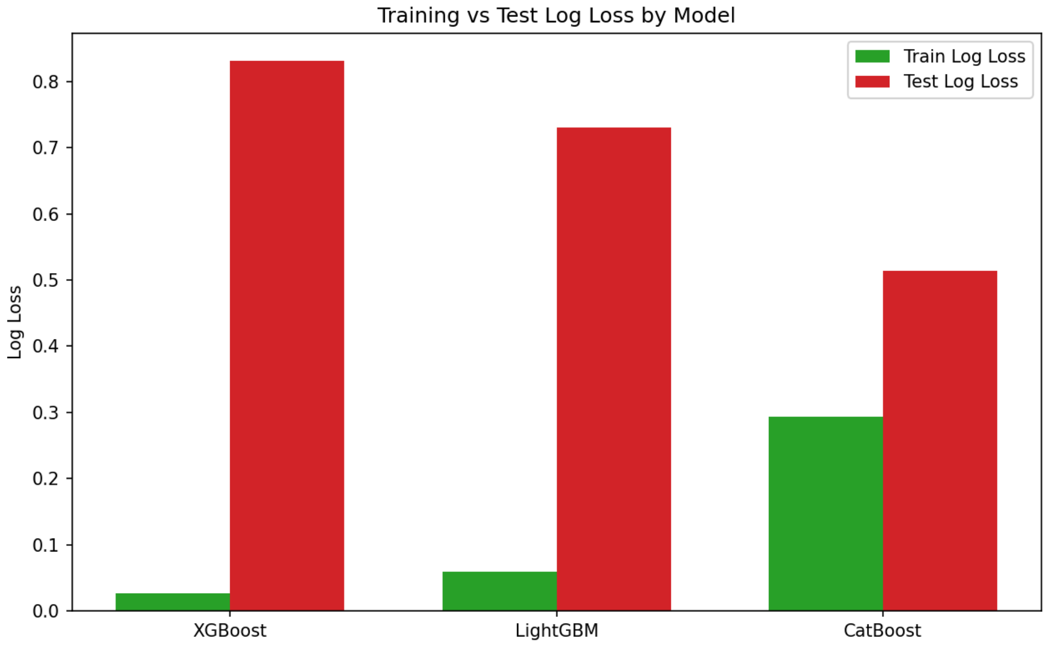


Max J XGBoost - 0.47

Max J LightGBM - 0.42

Max J CatBoost - 0.49





Best model based on test accuracy: **CatBoost**

References

1. **Git repository:** https://github.com/akshat-git/Pima-Indian-Diabetes-Inference-Supervised-Learning

2. **Dataset:** https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database