Diabetes Prediction Using Machine Learning

Shresht Bhowmick, Mouad Tiahi, Xiaole Su, Colin Johnson

MATH 7243: Machine Learning Theory 1 – Fall 2025

Background & Motivation

Problem: Diabetes affects 38.4M Americans (11.6%), with 8.7M undiagnosed. Late diagnosis increases risk of heart disease, kidney failure, and blindness.

Goal: Compare ML approaches for diabetes risk prediction from routine health metrics.

Questions:

- Which ML paradigm works best?
- Can deep learning compete?
- What limits performance?

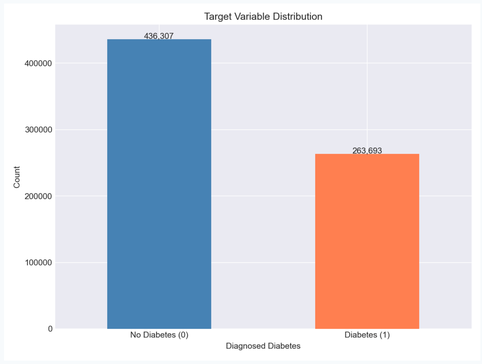
Dataset

Kaggle Playground S5E12 – Synthetic data via CTGAN.

Training	700,000
Test	300,000
Features	24
Classes	62% diabetic / 38% non

Features: Age, BMI, blood pressure, cholesterol, physical activity, family history, etc.

Missing: HbA1c, fasting glucose, insulin (standard clinical markers).



Methods

1. Logistic Regression (ElasticNet)
L1+L2: $\lambda(\alpha\|w\|_1 + (1-\alpha)\|w\|_2^2)$

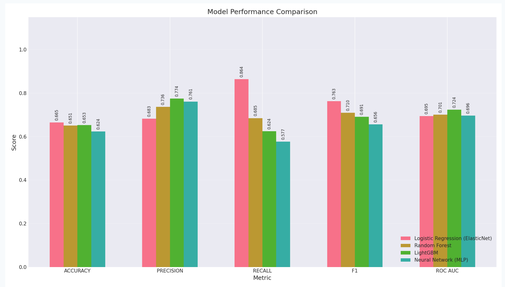
2. Random Forest
100 trees, max_depth=20, balanced weights

3. LightGBM
Leaf-wise boosting, histogram binning, GOSS

4. Neural Network (MLP)
256→128→64, BatchNorm, Dropout(0.3)

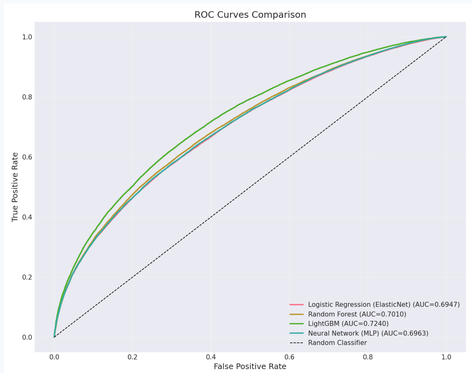
Main Results

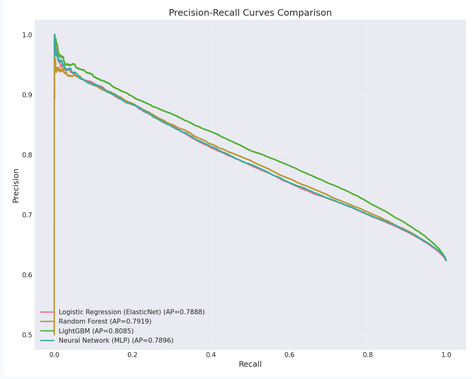
Model	Acc	Prec	Rec	AUC
Logistic Reg	66.5	68.3	**86.4**	0.695
Random Forest	65.1	73.6	68.5	0.701
LightGBM	65.3	**77.4**	62.4	**0.724**
Neural Net	62.4	76.1	57.7	0.696



Kaggle: LightGBM: **0.696 AUC** (top: 0.705)

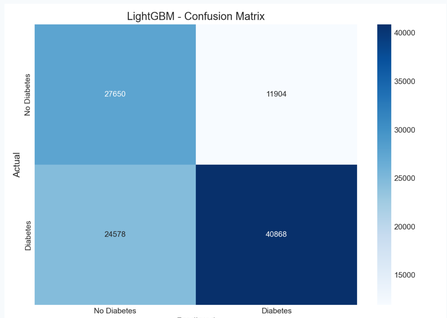
ROC & PR Curves



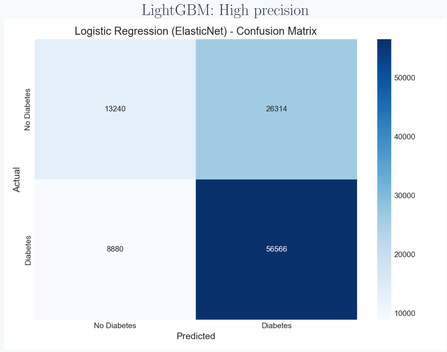


LightGBM: AUC=0.724, AP=0.8085

Confusion Matrices

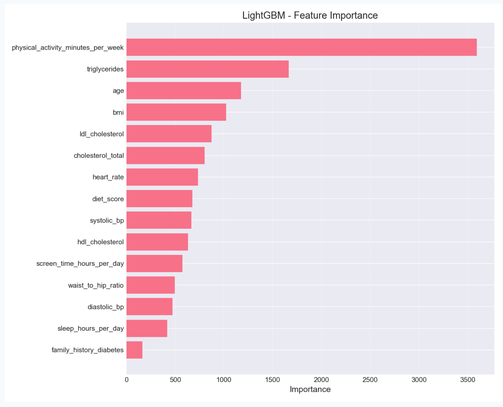


LightGBM: High precision



Logistic Reg: High recall

Feature Importance



Top: Physical activity, family history, age, BMI

Neural Network Training

Overfitting at epoch 15–20 despite regularization. Validation loss plateaus while training loss decreases.

Key Findings

1. **Gradient boosting wins:** LightGBM best AUC (0.724) – SOTA for tabular data
2. **Deep learning struggles:** MLP worst performance; tabular data lacks structure NNs exploit
3. **65% accuracy ceiling:** All models converge – feature quality limits performance
4. **Class imbalance:** Balanced weights essential for medical prediction

Conclusions

- LightGBM is best for tabular medical data
- Logistic regression wins if recall prioritized
- Missing clinical markers limit accuracy
- Model choice depends on clinical context: high recall for screening, high precision for diagnosis

Future: SHAP interpretability, threshold tuning, TabNet/FT-Transformer exploration, validation on real clinical data

Code: `github.com/tetraslam/mltheory-project`

References

[1] Breiman, "Random Forests," 2001 [2] Ke et al., "LightGBM," NeurIPS 2017 [3] Zou & Hastie, "Elastic Net," 2005 [4] Ioffe & Szegedy, "Batch Normalization," ICML 2015

LaTeX TikZposter