Diabetes Forecasting

By Dylan Wang, Eddie Xiao
12/05/2025

Motivation

Diabetes is one of the most prevalent chronic diseases worldwide, impacting quality of life and healthcare costs.

Traditional diagnosis relies on clinical lab tests, which can be costly and reactive.

With enough health data, ML can predict risk early, even before diagnosis.

We want to develop a model that accurately classifies whether a person is diabetic based on key health and behavioral indicators.

Background

What is a Diabetes?

- A chronic disease characterized by high blood glucose levels.
- Early detection is crucial to prevent severe complications (e.g., heart disease, nerve damage).
- According to WHO, diabetes affects over 530 million adultes globally

Related Work

Prior Research:

- Previous studies have used limited clinical data (e.g., Pima Indian Diabetes Dataset).
- Typical models: Logistic Regression, Decision Trees, SVMs.
- Many lacked lifestyle features (diet, sleep, activity) and often had small sample sizes.

Our Advancement:

- Compare multiple advanced models (XGBoost, LightGBM, Random Forest, Logistic Regression).
- Integrate lifestyle and clinical features for more holistic prediction
- Use a comprehensive, large-scale dataset

Claim / Target Task

Target Task:

Predict whether a person has diabetes (binary classification) using health and lifestyle indicators.

Claim:

By combining medical and behavioral data with modern ML methods, we can achieve high predictive performance while maintaining interpretability.

Problem is can ML detect patterns of diabetes risk earlier and more accurately than traditional methods?

Proposed Solution

Purpose:

To build an accurate, interpretable model that predicts diabetes likelihood using both clinical and lifestyle data - enabling proactive health management.

Our solution:

- Create a machine learning pipeline to classify individuals as diabetic or non-diabetic.
- Combine traditional health metrics (BMI, glucose, BP) with behavioral features (sleep, diet, exercise).
- Evaluate multiple ML algorithms (XGBoost, Random Forest, LightGBM, Logistic Regression).

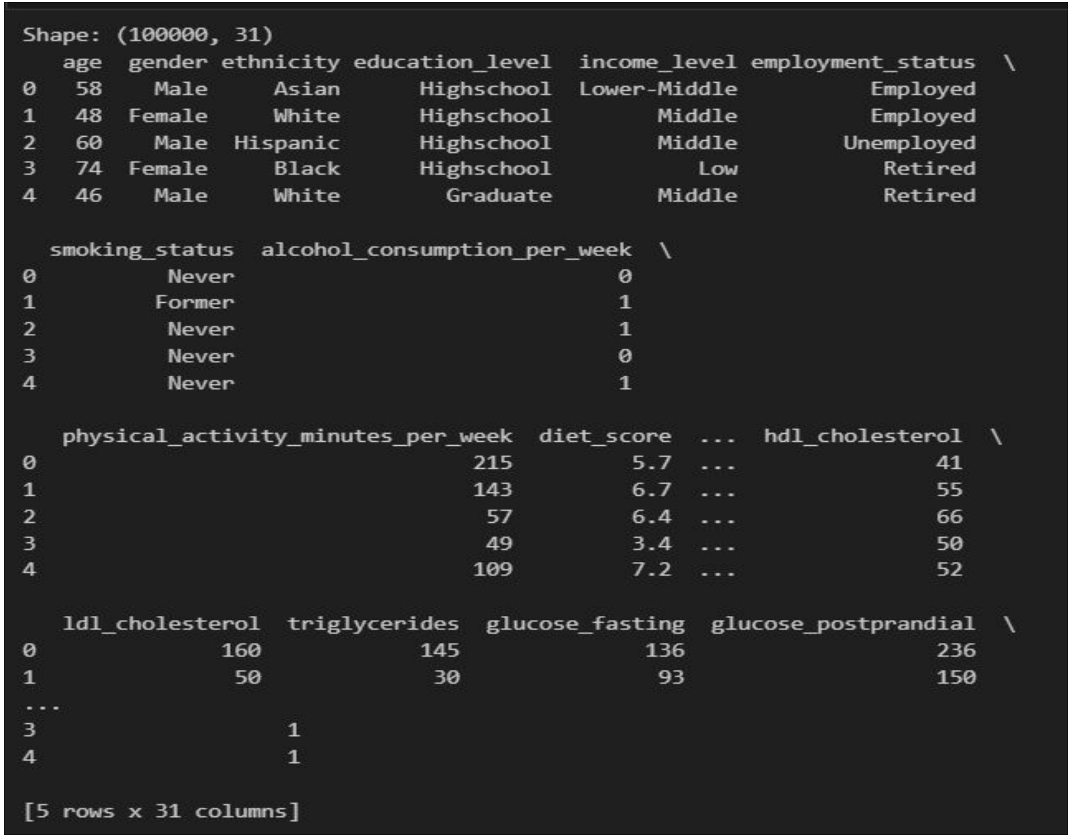
Data Summary

~ 253,000 records * 31 features

The target variable is diagnosed_diabetes that is either 0,1.

Key Features include: Age, Gender, BMI, Glucose levels, Blood Pressure, Cholesterol, diet score, sleep, physical activity, alcohol/smoking habits.

Data Visualization 1



```
Shape: (100000, 31)
   age  gender ethnicity education_level  income_level employment_status  \
0   58    Male     Asian      Highschool  Lower-Middle          Employed
1   48  Female     White      Highschool        Middle          Employed
2   60    Male  Hispanic      Highschool        Middle        Unemployed
3   74  Female     Black      Highschool           Low           Retired
4   46    Male     White        Graduate        Middle           Retired

  smoking_status  alcohol_consumption_per_week  \
0          Never                             0
1         Former                             1
2          Never                             1
3          Never                             0
4          Never                             1

   physical_activity_minutes_per_week  diet_score  ...  hdl_cholesterol  \
0                                 215         5.7  ...               41
1                                 143         6.7  ...               55
2                                  57         6.4  ...               66
3                                  49         3.4  ...               50
4                                 109         7.2  ...               52

   ldl_cholesterol  triglycerides  glucose_fasting  glucose_postprandial  \
0              160            145              136                   236
1               50             30               93                   150
...
3                   1
4                   1

[5 rows x 31 columns]
```

Step 1- Data Processing

1 Data cleaning:

- Dropped missing or duplicate rows
- Removed potential leakage columns (diabetes_risk_score etc)

2 Feature Engineering:

- Label-encoded categorical variables
- Scaled numerical features using StandardScaler

3 Train/ Test Split:

- 80% training / 20% testing
- Stratified by diabetes class to maintain class balance

Step 2 - Model Design

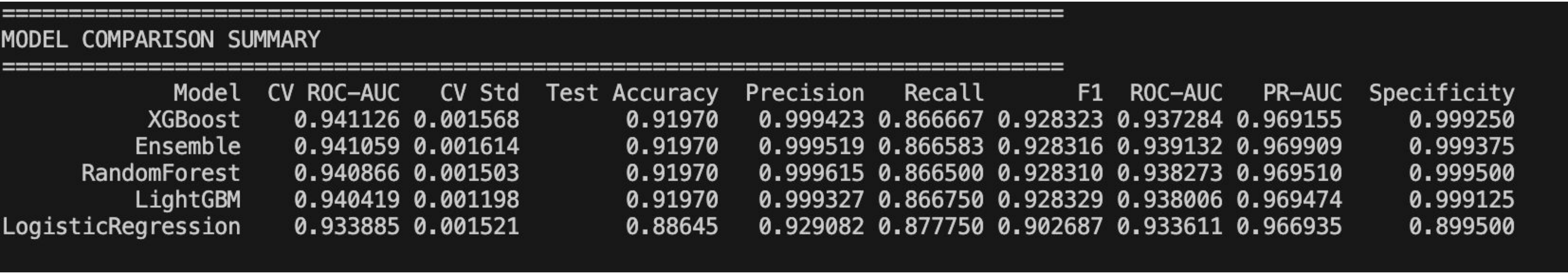
Model Tested:

- Logistic Regression
- Random Forest
- XGBoost
- LightGBM

Evaluation Metrics:

- Val_acc
- Precision
- Recall
- F1-score
- ROC-AUC

Model result

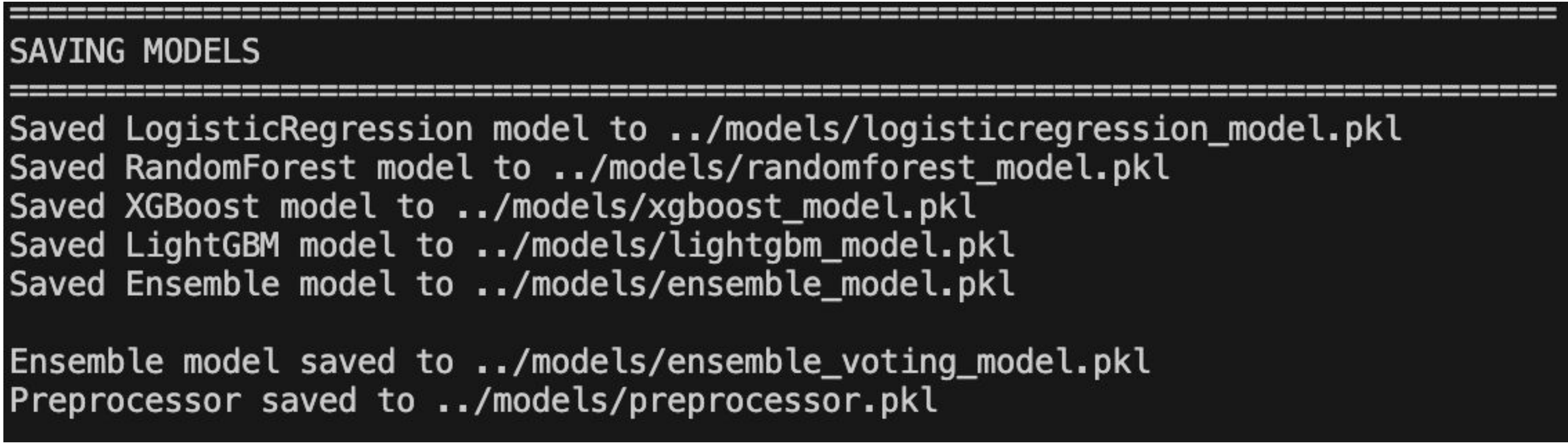


```
================================================================================
MODEL COMPARISON SUMMARY
================================================================================
             Model  CV ROC-AUC   CV Std  Test Accuracy  Precision   Recall       F1  ROC-AUC   PR-AUC  Specificity
           XGBoost    0.941126 0.001568        0.91970   0.999423 0.866667 0.928323 0.937284 0.969155     0.999250
          Ensemble    0.941059 0.001614        0.91970   0.999519 0.866583 0.928316 0.939132 0.969909     0.999375
      RandomForest    0.940866 0.001503        0.91970   0.999615 0.866500 0.928310 0.938273 0.969510     0.999500
          LightGBM    0.940419 0.001198        0.91970   0.999327 0.866750 0.928329 0.938006 0.969474     0.999125
LogisticRegression    0.933885 0.001521        0.88645   0.929082 0.877750 0.902687 0.933611 0.966935     0.899500
```

Best Model:

XGBoost achieved the highest ROC-AUC (0.941126) with strong overall performance.

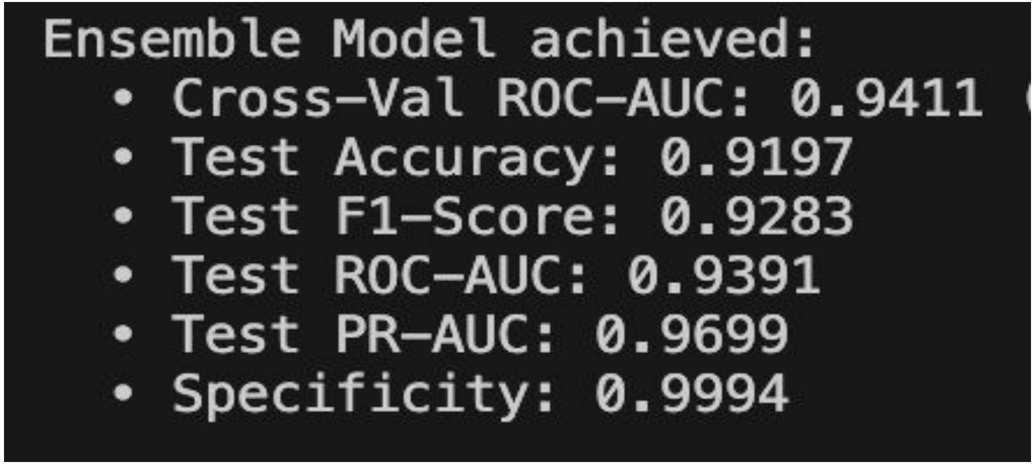
Models Saved



```
================================================================================
SAVING MODELS
================================================================================
Saved LogisticRegression model to ../models/logisticregression_model.pkl
Saved RandomForest model to ../models/randomforest_model.pkl
Saved XGBoost model to ../models/xgboost_model.pkl
Saved LightGBM model to ../models/lightgbm_model.pkl
Saved Ensemble model to ../models/ensemble_model.pkl

Ensemble model saved to ../models/ensemble_voting_model.pkl
Preprocessor saved to ../models/preprocessor.pkl
```

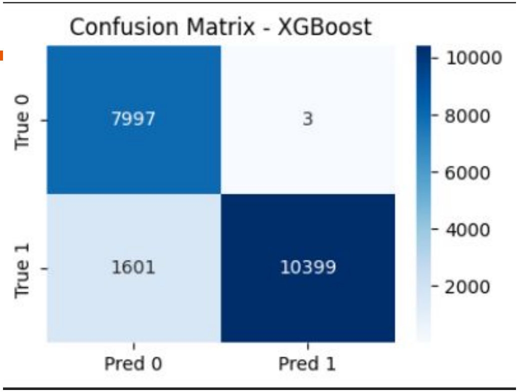
Ensemble Model Result

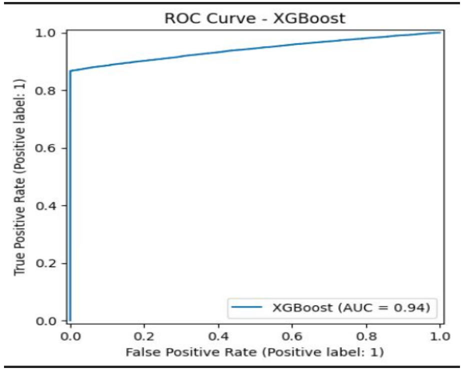


```
Ensemble Model achieved:
    • Cross-Val ROC-AUC: 0.9411 (
    • Test Accuracy: 0.9197
    • Test F1-Score: 0.9283
    • Test ROC-AUC: 0.9391
    • Test PR-AUC: 0.9699
    • Specificity: 0.9994
```

When XGBoost, LightGBM, and RandomForest all achieve exactly 91.97% accuracy, it's not coincidence—it's confirmation that our data contains strong, clear signals about diabetes that any competent algorithm can find.

Visualization ROC Curves & Confusion Matrix



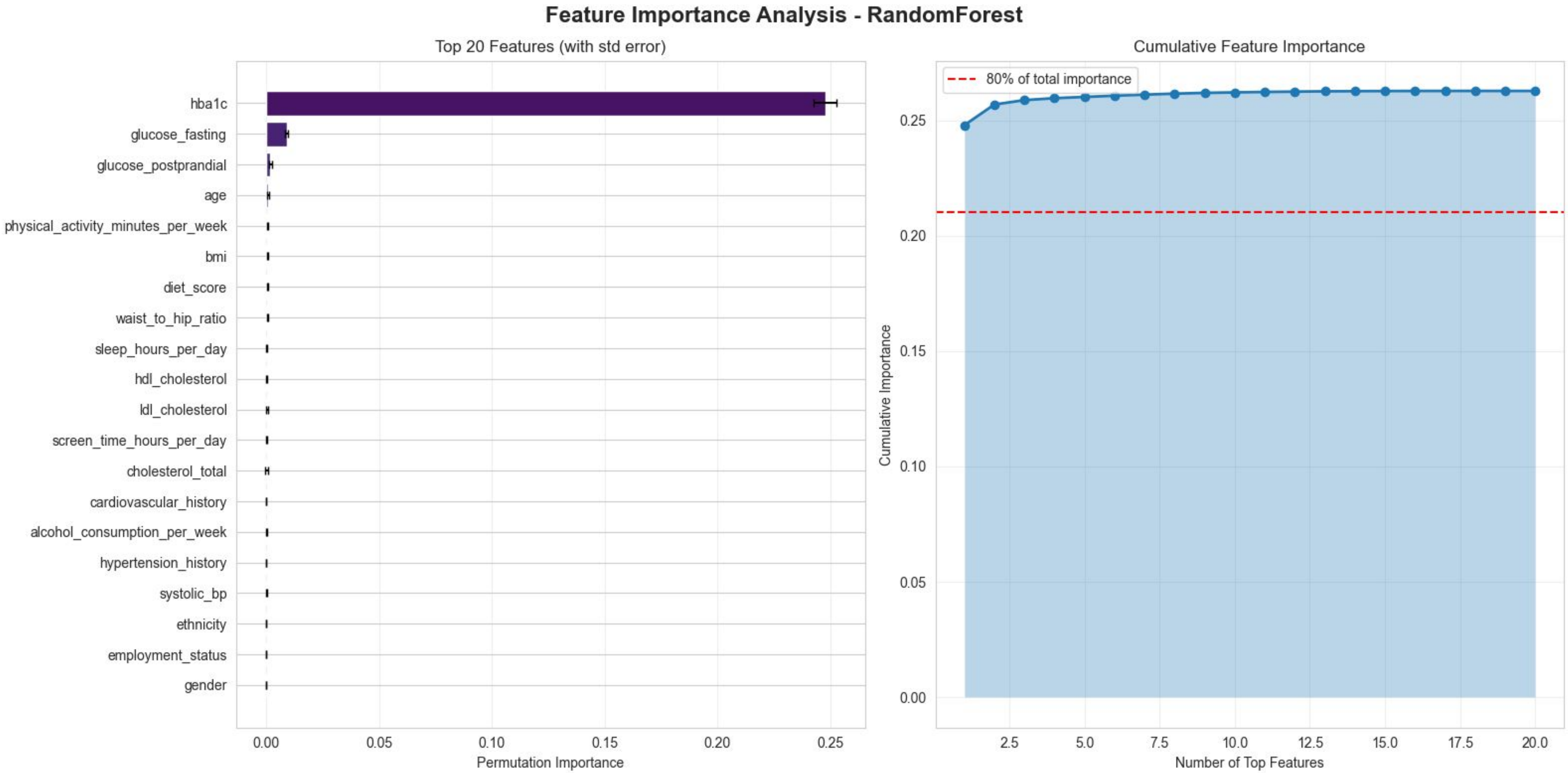


We can see that all tree-based models achieved similar ROC shapes.

Confusion matrices show strong true positive rates, minimal false positives.

Logistic Regression underperforms slightly but provides interpretable coefficients.

Feature Importance



Challenges & Observation

Challenges:

- Class imbalance (fewer diabetic cases) — mitigated using class weights
- Handling mixed data types (categorical + numeric).
- Avoiding target leakage from health score columns.

Observations:

- Models maintained >0.93 ROC-AUC, suggesting reliable generalization.
- XGBoost and Random Forest perform almost identically.

Conclusion and Future Work

- Incorporate time-series tracking
- Add more feature selection
- Test with real wearables data
- Integrate into web-based application that help people

Reference

Kaggle Dataset: https://www.kaggle.com/datasets/mohankrishnathalla/diabetes-health-indicators-dataset/data

Documentation:

https://xgboost.readthedocs.io/en/stable/

https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

Other Study:

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

Demo Link

Video on Slides link:

https://youtu.be/kGeAIGOYYI4

Code demo link:

https://youtu.be/4sVgWTkxrcI

Who Did What

Dylan: I constructed the models (random forest, xgboost, lightgbm) and ensemble model. 50 % of the slide.

Eddie: Data cleaning, Feature Selection. 50% of the slide. Searched the dataset for ml model. Logistic regression.