Diabetes Risk Prediction for Targeted Drug Trial Recruitment: A Comparative Analysis of Logistic Regression and Random Forest Models

Author: Ruixin Chen

Introduction

I am Ruixin Chen, a user analysis specialist at a pharmaceutical company. Currently, my company is recruiting paid volunteers for a targeted drug trial for diabetes. I have to analyze past volunteers' health data to identify potential suitable candidates. I have developed a diabetes prediction model1 (using logistic regression), model2(using Random Forest) and then input the information of the individuals I have to predict into model1/model2. The test results are stored as "medicine." Afterward, I submit the list of individuals with a "medicine" result of 1 to the contact department.

Dataset Source

The dataset used in this analysis is publicly available on Kaggle. It includes health metrics for various individuals, which allows for the analysis of factors associated with diabetes risk.

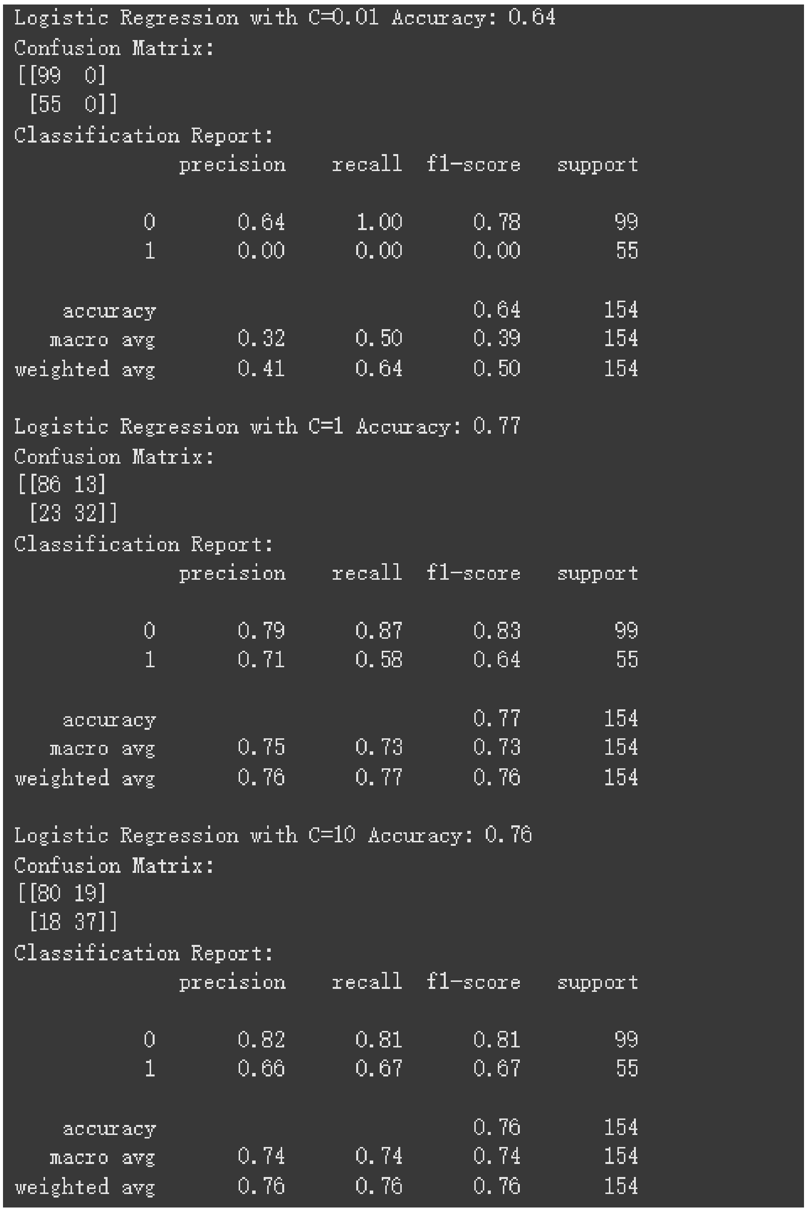
Feature Engineering and Selection

Key features in the dataset include:

- Pregnancies
- Glucose
- Blood Pressure
- Skin Thickness
- Insulin
- BMI
- Diabetes Pedigree Function
- Age

To enhance the model's predictive power, two new features were engineered in our past volunteers' health data:

- Average Blood Pressure: Calculated as the average of systolic and diastolic blood pressure readings ((BPSysAve + BPDiaAve) / 2).
- Age-based Binary Label: Created a binary label to indicate if the age is above 50, as a significant indicator in diabetes risk prediction.
- **Logistic Regression (Model1): Logistic regression was selected due to its simplicity, interpretability, and ability to work well with binary classification tasks. It serves as a baseline model for comparison.**



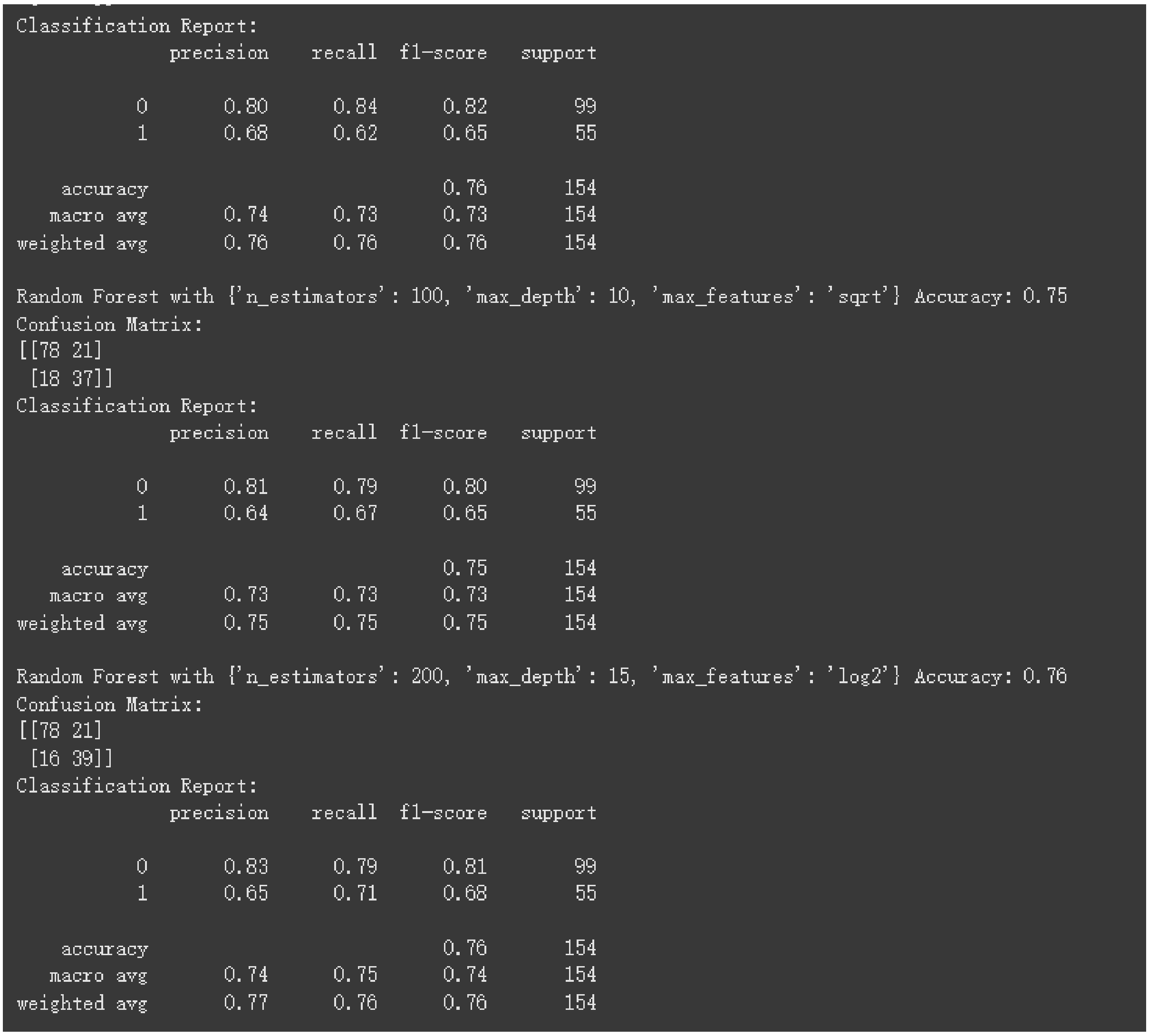
```
Logistic Regression with C=0.01 Accuracy: 0.64
Confusion Matrix:
[[99  0]
 [55  0]]
Classification Report:
              precision    recall  f1-score   support

           0       0.64      1.00      0.78        99
           1       0.00      0.00      0.00        55

    accuracy                           0.64       154
   macro avg       0.32      0.50      0.39       154
weighted avg       0.41      0.64      0.50       154

Logistic Regression with C=1 Accuracy: 0.77
Confusion Matrix:
[[86 13]
 [23 32]]
Classification Report:
              precision    recall  f1-score   support

           0       0.79      0.87      0.83        99
           1       0.71      0.58      0.64        55

    accuracy                           0.77       154
   macro avg       0.75      0.73      0.73       154
weighted avg       0.76      0.77      0.76       154

Logistic Regression with C=10 Accuracy: 0.76
Confusion Matrix:
[[80 19]
 [18 37]]
Classification Report:
              precision    recall  f1-score   support

           0       0.82      0.81      0.81        99
           1       0.66      0.67      0.67        55

    accuracy                           0.76       154
   macro avg       0.74      0.74      0.74       154
weighted avg       0.76      0.76      0.76       154
```

- *Logistic Regression: The regularization parameter C was tuned over values [0.01, 1, 10]. C = 1 yielded the best performance.*

- **Random Forest Classifier (Model2): Random forest, an ensemble method, was chosen for its ability to capture non-linear relationships and interactions**

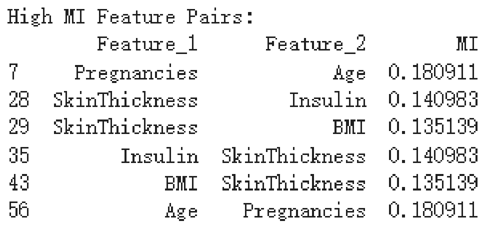
between features, which are common in complex medical datasets. Additionally, random forest is robust against overfitting when tuned appropriately.

- *Random Forest: Three sets of hyperparameters were tested:*
  n_estimators = 50, max_depth = 5, max_features = 'sqrt'
  n_estimators = 100, max_depth = 10, max_features = 'sqrt'
  n_estimators = 200, max_depth = 15, max_features = 'log2' The third configuration, n_estimators = 200, max_depth = 15, and max_features = 'log2', provided the best results.

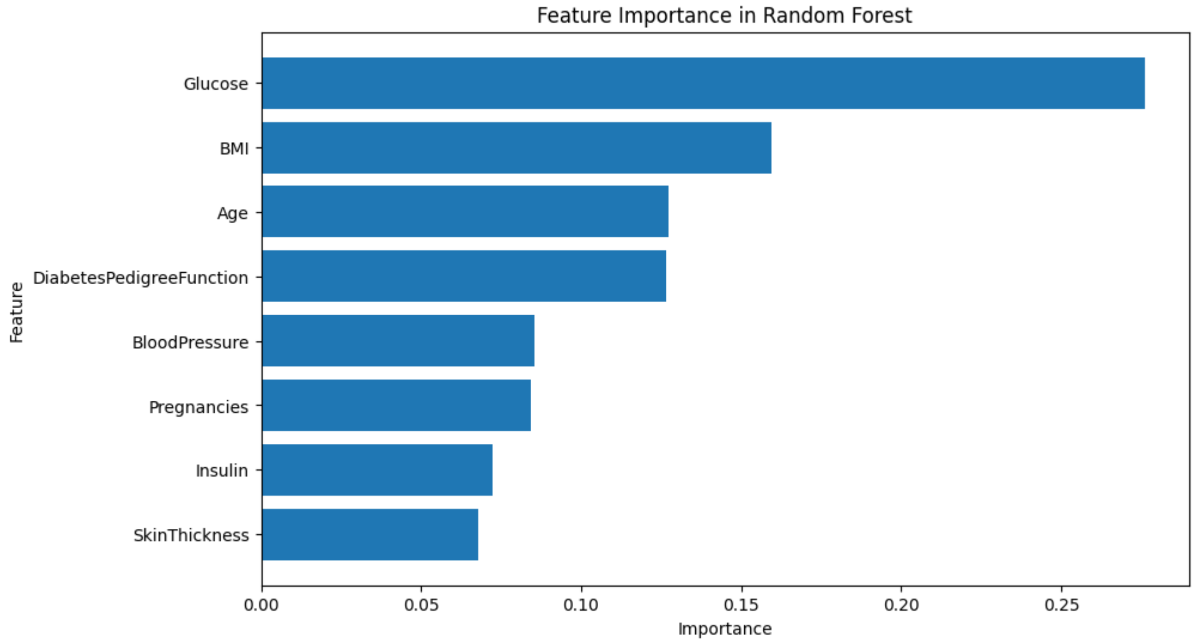


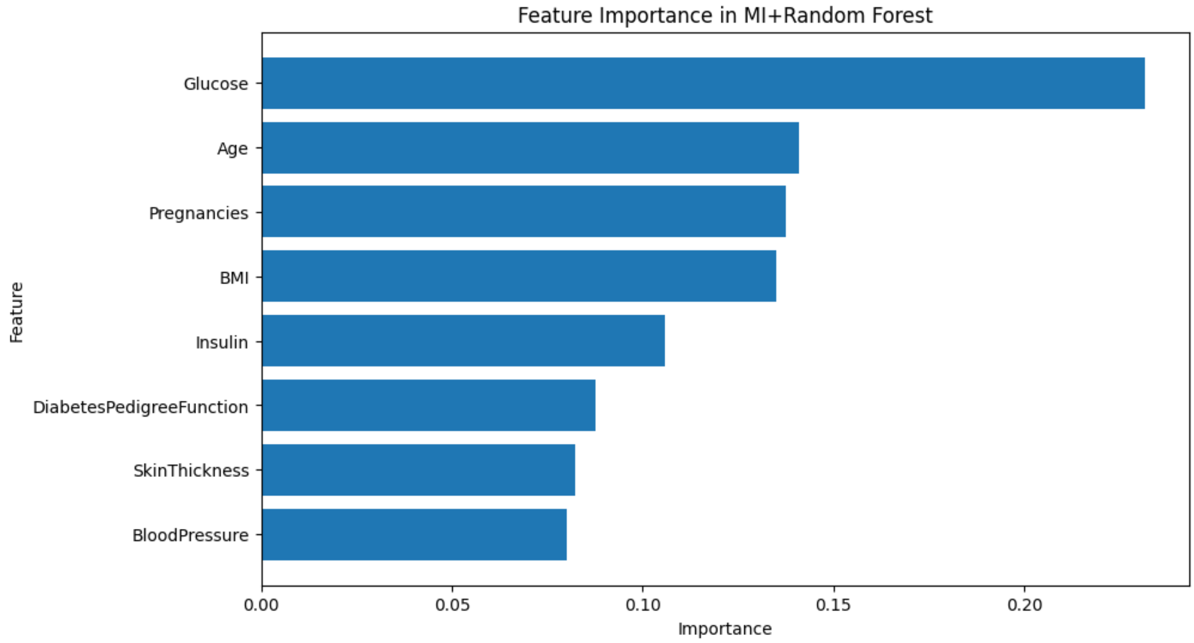
```
Classification Report:
              precision    recall  f1-score   support

           0       0.80      0.84      0.82        99
           1       0.68      0.62      0.65        55

    accuracy                           0.76       154
   macro avg       0.74      0.73      0.73       154
weighted avg       0.76      0.76      0.76       154

Random Forest with {'n_estimators': 100, 'max_depth': 10, 'max_features': 'sqrt'} Accuracy: 0.75
Confusion Matrix:
[[78 21]
 [18 37]]
Classification Report:
              precision    recall  f1-score   support

           0       0.81      0.79      0.80        99
           1       0.64      0.67      0.65        55

    accuracy                           0.75       154
   macro avg       0.73      0.73      0.73       154
weighted avg       0.75      0.75      0.75       154

Random Forest with {'n_estimators': 200, 'max_depth': 15, 'max_features': 'log2'} Accuracy: 0.76
Confusion Matrix:
[[78 21]
 [16 39]]
Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.79      0.81        99
           1       0.65      0.71      0.68        55

    accuracy                           0.76       154
   macro avg       0.74      0.75      0.74       154
weighted avg       0.77      0.76      0.76       154
```

As I find that the MI(Mutual Information) calculation can change the importance of the features, so I add some step, firstly calculate the high MI pairs, then using a function that Feature1*=Feature1=c*Feature2, create a new feature1 to reduce the MI. It will improve the importance between the features which improve the model performance



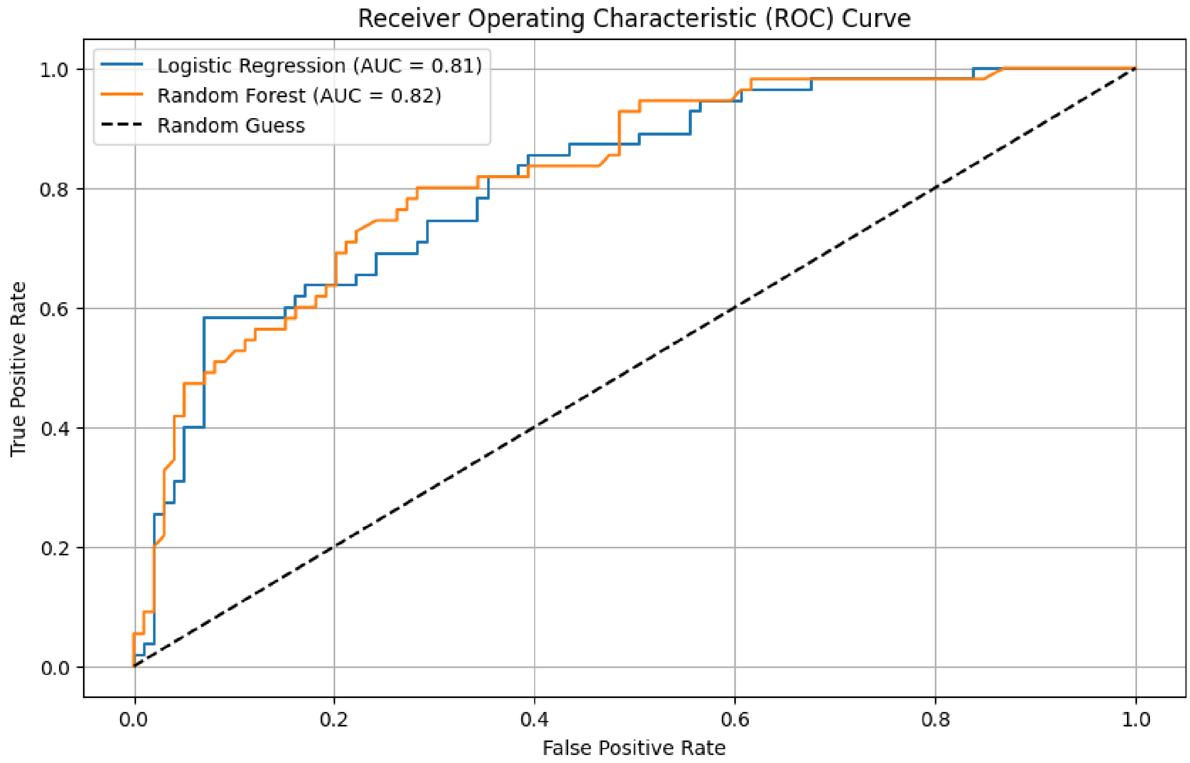
```
High MI Feature Pairs:
       Feature_1      Feature_2        MI
7    Pregnancies            Age  0.180911
28  SkinThickness        Insulin  0.140983
29  SkinThickness            BMI  0.135139
35        Insulin  SkinThickness  0.140983
43            BMI  SkinThickness  0.135139
56            Age    Pregnancies  0.180911
```





The final AUC improved to 0,83, but I think this still need to validated by the formula, so here is just a step as reference.

To assess model performance, ROC-AUC was selected as the primary evaluation metric because it provides a balanced measure of both sensitivity and specificity, critical in a health-related classification model.



Result

- Model1 (Logistic Regression): AUC = 0.81
- Model2 (Random Forest): AUC = 0.82

While both models performed reasonably well, the random forest model (Model2) slightly outperformed logistic regression, indicating a stronger ability to capture complex feature interactions.

Recommendation

Model2 (Random Forest) is recommended due to its higher AUC score of 0.82 compared to Model1's 0.81. Although the difference is small, the increased flexibility of random forests in handling feature interactions may prove beneficial in real-world scenarios, especially when applied to a broader population.

Future Work

- Hyperparameter Optimization: Use grid or random search over a more extensive set of hyperparameters to fine-tune both models further.
- Feature Expansion: Incorporate additional health-related metrics, such as lifestyle and dietary factors, which might provide further predictive value.

arrow_upwardarrow_downward

link

comment

edit

delete

more_vert

Conclusion

The random forest model (Model2) demonstrated a marginally better performance, with an AUC of 0.82 over Model1's 0.81. Given the application's goal to identify suitable volunteers for a diabetes trial, the random forest model's accuracy and ability to handle complex relationships make it a more appropriate choice for the stakeholder's needs.