National Yang Ming Chiao Tung University (國立陽明交通大學)

AI Term Project #3

Quentin Ducoulombier (昆丁)
Student Id: 312551811

Artificial Intelligence

胡毓志

2024-06-22

Introduction

Candidemia represents a critical health challenge and is increasingly recognized as a significant cause of mortality globally, especially in immunocompromised individuals and patients in intensive care units.
This project aims to develop predictive models using machine learning techniques to estimate the survival outcomes of patients suffering from candidemia within 14 days of diagnosis. I employed classification methods, including Random Forest (RF) and Gradient Boosting Machine (GBM), and detailed the data preprocessing steps such as imputation, feature selection, and cross-validation. The experimental process, classification methods, challenges encountered, and corresponding solutions are described in this comprehensive report.

Data Preprocessing

Data Loading and Cleaning

- **Data Loading:** The dataset was loaded using `pandas` from Excel files `train_X.xlsx` and `train_y.xlsx`.
- **Target Variable Extraction:** The target variable, `Deadin_D14`, was extracted from `train_y`.
- **Column Cleaning:** Column names in `train_X` were cleaned by stripping any leading/trailing whitespace.

Numeric Feature Detection and Scaling

- **Numeric Feature Detection:** Numeric features were **automatically detected** using `train_X.select_dtypes(include=[np.number])`.
- **Feature Scaling:** Numeric features were **scaled** using `StandardScaler` to standardize the data.

Missing Value Imputation

- **Imputation Strategy:** Imputation Strategy: Missing values in the dataset were imputed using the **median** strategy with `SimpleImputer`, as attempts with the mean strategy yielded lower performance scores.

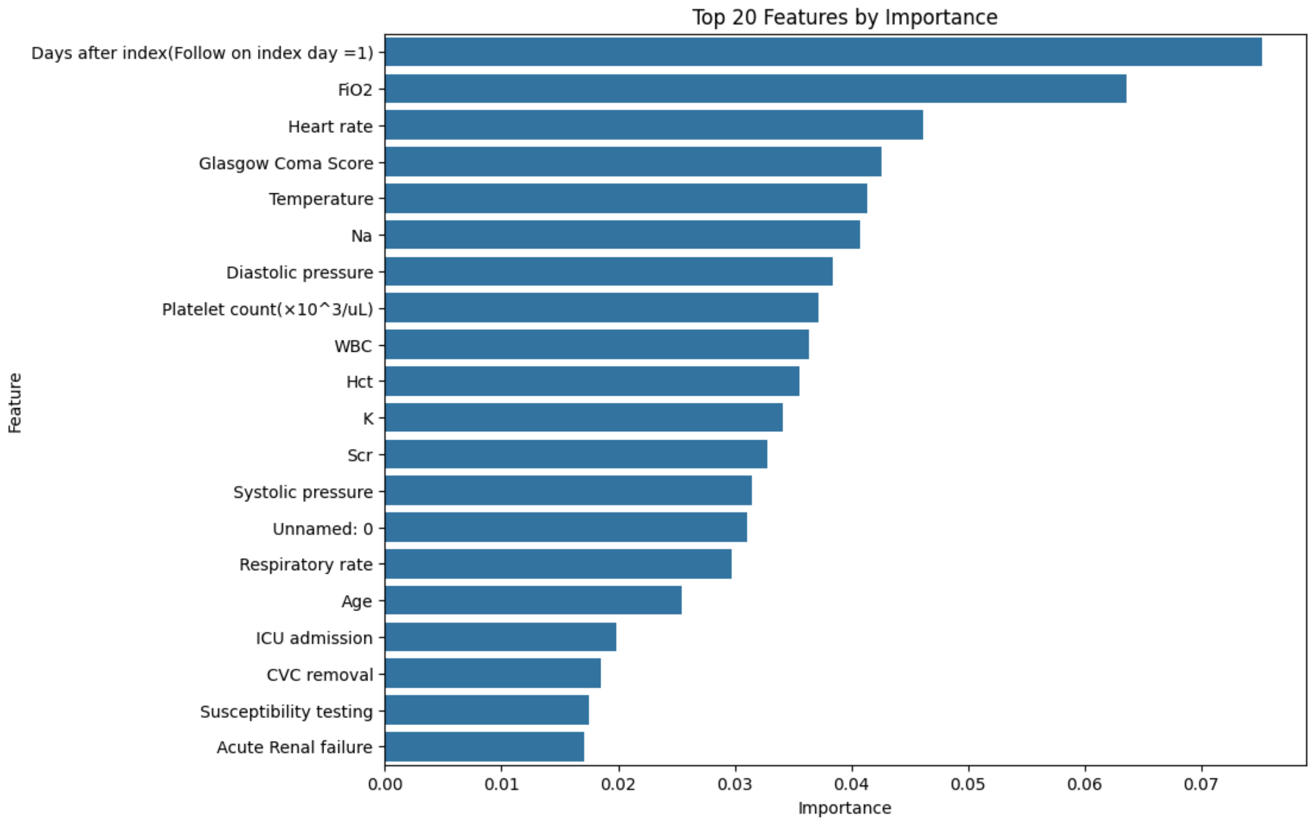
Data Splitting

- **Train-Validation Split:** The preprocessed data was split into training and validation sets using `train_test_split` with a **80-20 split**.

Classification Methods

Random Forest (RF)

- **Model Initialization:** A Random Forest classifier was initialized with specific hyperparameters:
  - `bootstrap=False`
  - `max_depth=10`
  - `min_samples_leaf=1`
  - `min_samples_split=2`
  - `n_estimators=100`
  - `random_state=42`
- **Feature Importance Analysis:** Important features were selected based on feature importance scores greater than 0.01.
- **Model Training:** The model was retrained using only the selected important features for better performances.



Gradient Boosting Machine (GBM)

- **Model Initialization:** A GBM classifier was initialized with specified hyperparameters:
    - `l2_regularization=1`
    - `learning_rate=0.1`
    - `max_depth=None`
    - `max_iter=100`
    - `min_samples_leaf=50`
    - `random_state=42`
- **Model Training:** The GBM model was trained using the full set of features without filtering for importance.

Experimental Results

Random Forest with Feature Selection

In an initial attempt to improve the RF model's performance, important features were selected based on feature importance. This approach yielded a RF AUROC score of 0.89... , which was below the desired threshold of 0.90. However, the RF model's performance showed significant improvement with the selected features:

- **RF F1 Score:** 0.6341
- **RF MCC:** 0.5215
- **RF AUROC:** 0.9138

Gradient Boosting Machine

The GBM model was introduced for comparison. Initially, it was trained without considering important features and achieved higher scores compared to the RF model:

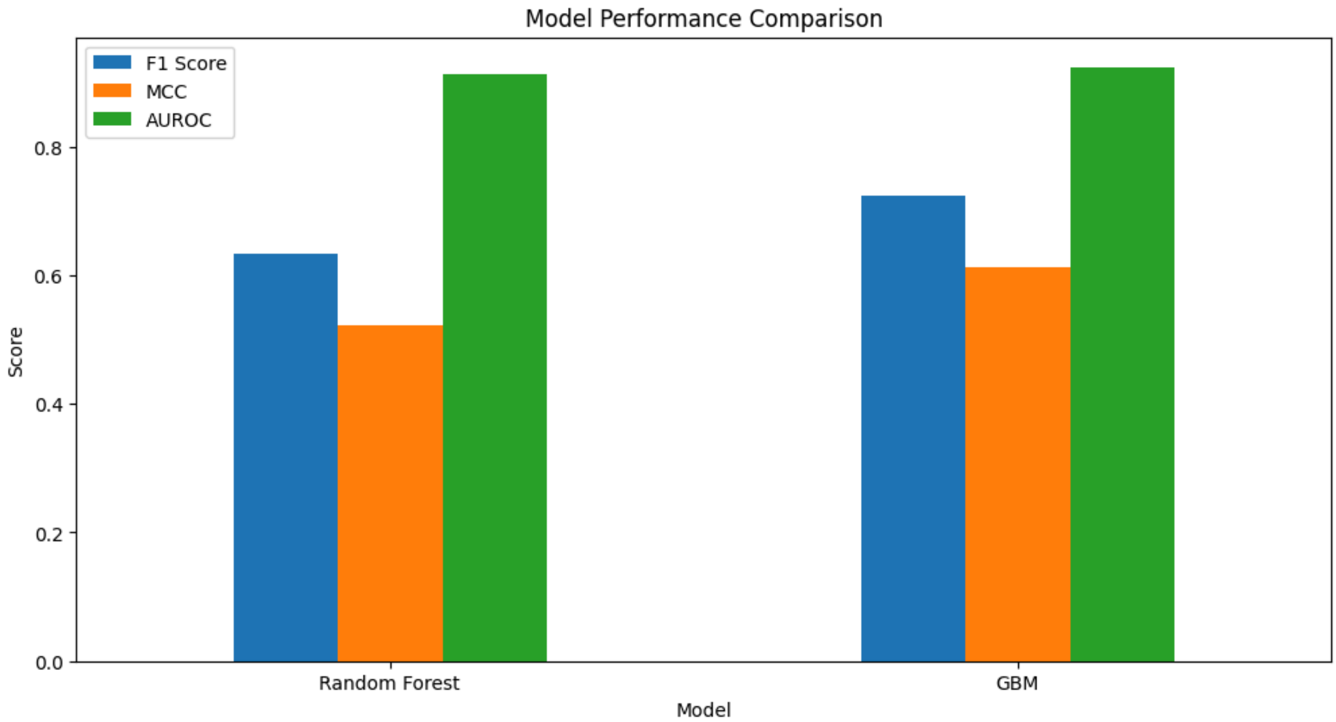
- **GBM F1 Score:** 0.7234
- **GBM MCC:** 0.6126
- **GBM AUROC:** 0.9241

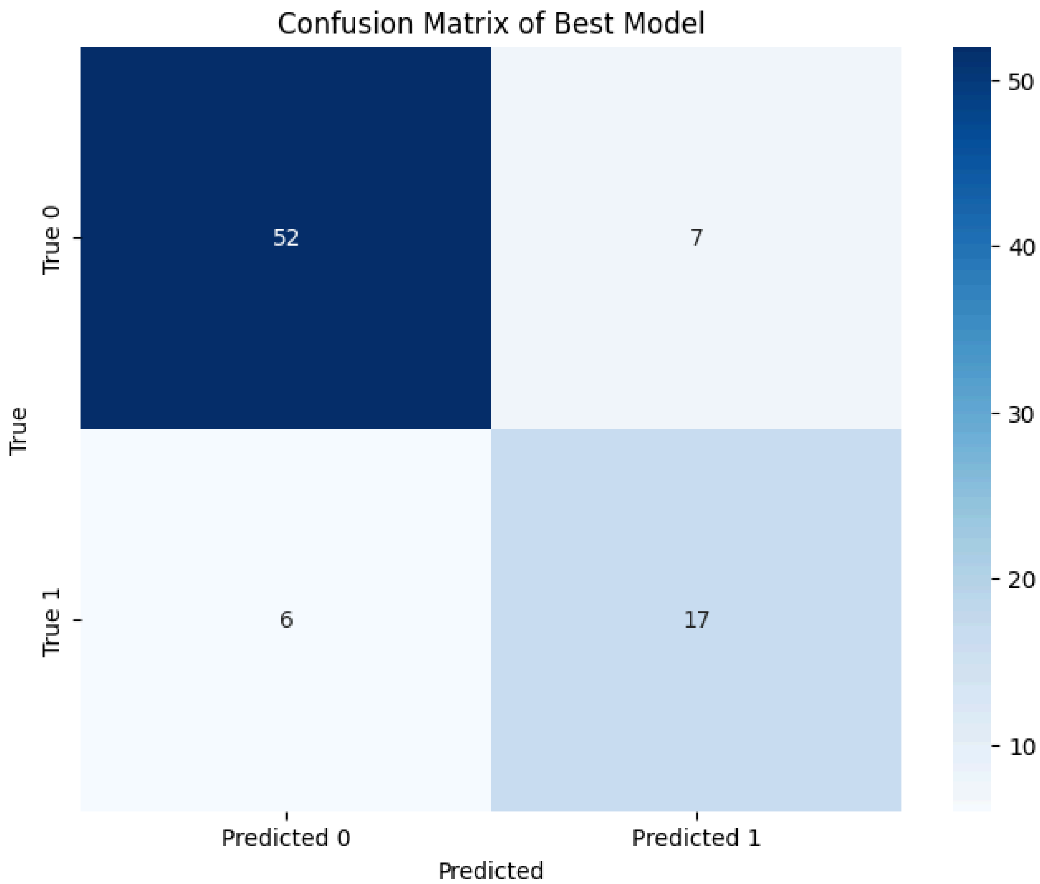
Comparison and Final Selection

Both models were compared based on their F1 score, MCC, and AUROC. The GBM model demonstrated superior performance across all metrics, particularly with an AUROC score of 0.9241.

Confusion Matrix and Best Model Selection

The confusion matrix for the best model was plotted to visualize its performance. Ultimately, the GBM model was selected for the final submission due to its higher F1 score and overall better performance.





Challenges and Solutions

Hyperparameter Tuning

One of the significant challenges was determining the best hyperparameters for the RF and GBM models. Extensive experimentation and optimization were conducted using `GridSearchCV` from `sklearn` to identify the optimal hyperparameters. Despite these efforts, achieving an AUROC score above 0.90 with the RF model was challenging until feature importance was incorporated, which substantially improved the performance.

Feature Selection

Selecting the most important features proved to be highly effective for the RF model, resulting in significant performance gains. This step, however, did not yield similar improvements for the GBM model.

Computational Resources

The computational cost and time required for hyperparameter tuning and model training were substantial, particularly when using `GridSearchCV`.

Conclusion

In conclusion, the GBM model, trained with the full set of features, outperformed the RF model with feature selection. The final model selection was based on the highest F1 score, ensuring the most accurate predictions for the test set. The detailed experimental process, data preprocessing, and model evaluation steps highlight the effectiveness of machine learning techniques in predicting candidemia mortality. The final results were saved to an Excel file for submission.