Project 3: Classification Modeling Report – Yuanseng Choo

Introduction and Overview

The dataset used in this project, titled new_train_EGN5442.csv, contains approximately 100,000 records with a mix of numerical and categorical variables. The goal is to build predictive models capable of classifying the target variable, y, based on the provided features. To accomplish this, several steps were carried out, including data cleaning, handling missing values, encoding categorical data, and exploring relationships among variables. After preparing the data, different machine learning models were developed and compared, such as Logistic Regression and Gradient Boosting Classifier. The goal is to evaluate their performance and identify which model produces more accurate and consistent predictions.

Data Cleaning and Preparation

Before developing the models, the dataset was examined to assess data quality and feature distribution. The target variable y contained two balanced classes with roughly 57% in class 0 and 43% in class 1. Constant and redundant features, including Unnamed: 0, x30, and x4, were removed. Missing values were handled through imputation where categorical features such as x14 were filled with their mode and numerical variables like x24 and x31 were filled with their median values after replacing infinite values with nulls. The mixed-format feature x29 was cleaned by extracting numeric values and categorical columns x2, x14, x25, and x26 were label-encoded. The gender variable x3 was converted into binary form with Female = 1 and Male = 0. Visualizations including heatmaps and correlation matrices were used to assess missing data and relationships between variables. The dataset was then standardized using StandardScaler and split into training and testing sets with an 80/20 ratio for model development.

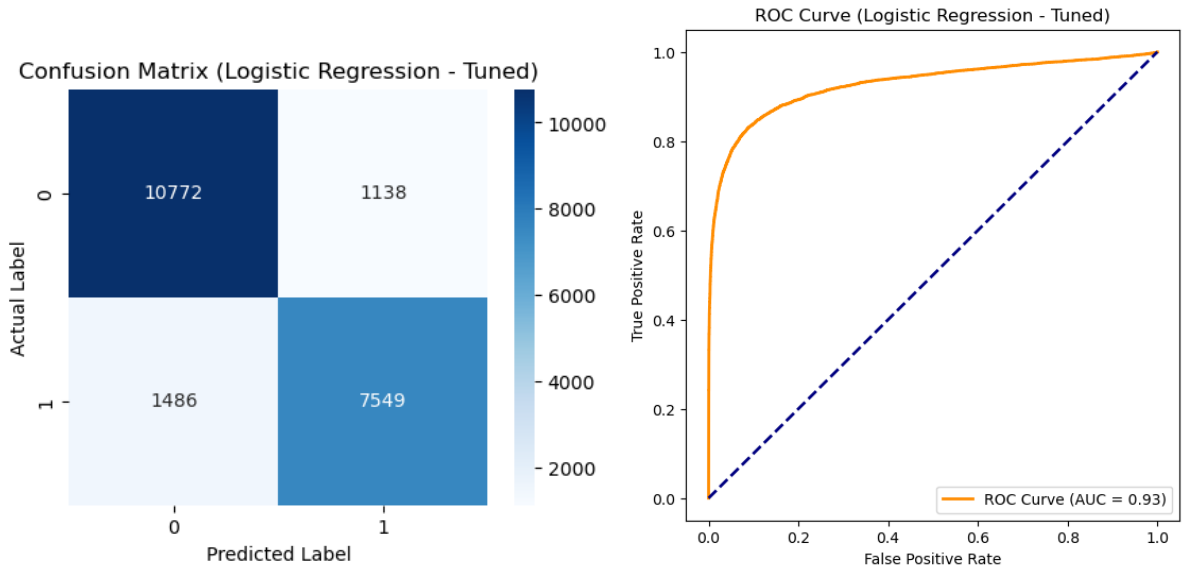
Part A – Logistic Regression Model

Logistic regression is a statistical model used to predict binary outcomes based on input variables. It estimates probabilities using a logistic function, producing results between 0 and 1. The main advantage of logistic regression is its simplicity and interpretability, while its limitation is that it may not capture complex nonlinear relationships.

In this project, the cleaned dataset was used to train and test logistic regression models. Correlation analysis was performed to understand feature relevance, and categorical variables were encoded before modeling. Hyperparameter tuning was applied using RandomizedSearchCV with five-fold cross-validation to improve performance.

The final tuned logistic regression model achieved strong results, with a test accuracy of 0.87, balanced precision and recall, and a high ROC/AUC score that indicated reliable classification between the two classes. The confusion matrix showed that the model correctly identified most observations, while the ROC curve demonstrated its strong ability to

distinguish between positive and negative outcomes. Overall, the logistic regression model provided stable and interpretable predictions suitable for baseline comparison with more advanced models.

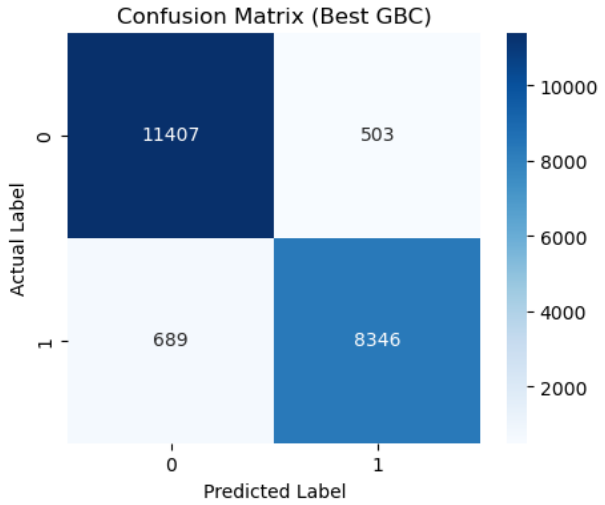


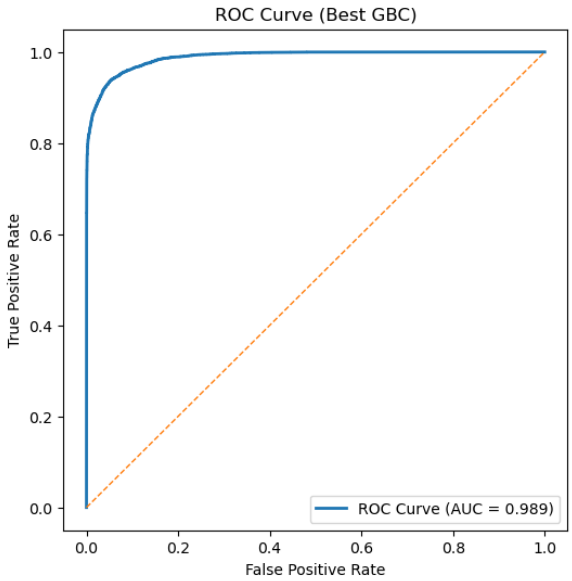
Part B – Non-logistic Model

For comparison, a Gradient Boosting Classifier (GBC) was used as the non-logistic model. The tree-based algorithm builds an ensemble of weak learners, such as decision trees, that sequentially correct the errors of previous trees. The main advantages of gradient boosting are its ability to capture complex nonlinear relationships and its high predictive accuracy. However, it can be computationally intensive and prone to overfitting if not properly tuned.

Hyperparameter tuning was performed using GridSearchCV with five-fold cross-validation to identify the best combination of parameters, which includes the loss function, learning rate, and number of estimators. The tuned GBC achieved excellent predictive performance, with a ROC/AUC score of 0.99 on the test set, which indicates strong discriminative ability. The confusion matrix showed that most observations were correctly classified, and the ROC curve confirmed the model's ability to separate the two target classes effectively.

One challenge encountered during training was the longer computation time required for model optimization. To address this, the complexity of the tuning process can be reduced by using efficient search strategies such as RandomizedSearchCV, early stopping, or by sampling a smaller portion of the data during testing. Overall, the Gradient Boosting Classifier outperformed the logistic regression model in prediction accuracy and adaptability to more complex relationships within the dataset.





Discussions

Both the Logistic Regression and Gradient Boosting Classifier (GBC) models were developed to predict the target variable based on the same cleaned dataset. The logistic regression model served as a strong and interpretable baseline, achieving a test accuracy of 0.87 and demonstrating consistent precision and recall. The tuned Gradient Boosting Classifier achieved a higher ROC AUC score of 0.99, indicating better classification performance and a stronger ability to capture complex patterns in the data.

While logistic regression provided simplicity and transparency in showing how each variable influenced the outcome, the Gradient Boosting Classifier was more effective in detecting nonlinear relationships and subtle feature interactions. In practical terms, this means that the GBC model was better at identifying correct predictions in situations where relationships between features were less straightforward.

To explain these findings to a business partner without using technical metrics, I would use a straightforward comparison based on decision impact. For instance, I would show that the Gradient Boosting model correctly identifies more successful outcomes compared to the logistic model when tested on real cases. Also, I would use simple visuals such as bar charts showing the number of correct versus incorrect predictions, or side-by-side bars comparing correct versus incorrect classifications, which would clearly show that the Gradient Boosting model provides more dependable results for future decision-making.

	Model1-logistic regression	Model2-non-logistic
AUC for training	0.88	0.99
AUC for testing	0.87	0.99