MF815 Midterm Project - Company Bankruptcy Prediction

Chi-Lin Li

1 Abstract

This report presents a comprehensive examination of several machine learning models to predict bankruptcy, based on financial data. Through the application of oversampling techniques to address class imbalance and a thorough hyperparameter tuning process, I aim to optimize model performance with respect to accuracy, F1 score, and the area under the receiver operating characteristic (AUC) curve.

2 Methodologies

My approach involves preprocessing a dataset of financial metrics to predict bankruptcy, folloId by splitting it into training and validation sets. I address the imbalance in the dataset using the Synthetic Minority Over-sampling Technique (SMOTE). Subsequently, I implement and assess several models: XGBoost, LightGBM, and Multi-Layer Perceptron (MLP). Each model is initially evaluated with default parameters and then optimized through GridSearchCV to find the best hyperparameters based on the F1 score. Performance metrics are meticulously recorded pre- and post-optimization.

2.1 Preprocessing & Additional Contribution

My preprocessing method incorporates several essential steps to prepare data for a machine learning model, ensuring that it is clean, normalized, and balanced. Initially, I tackle the problem of missing values in both training and test datasets by identifying and displaying columns with missing values. This is a critical step for understanding data quality and deciding on further actions, such as imputation or removal of missing data.

Normalize

I normalize the numerical columns in the dataset using the **StandardScaler** from **sklearn.preprocessing**. Normalization is crucial as it brings all my numerical variables into a similar scale, eliminating biases towards variables with higher magnitudes. This step involves fitting the scaler on the training data to learn the parameters (mean and standard deviation for each feature) and then transforming both training and test datasets with these parameters. By doing so, I ensure that the model is not biased by the scale of features and can learn more effectively.

SMOTE (Synthetic Minority Over-sampling Technique)

SMOTE (Synthetic Minority Over-sampling Technique) generates synthetic examples by interpolating betIen positive instances in the feature space. For an instance x_i in the minority class, it randomly picks one of its k-nearest minority class neighbors x_{in}, and generates a new sample x_{new} by:

Given a minority class sample x_i and its neighbor x_{in}, the synthetic sample x_{new} is computed as:

$$x_{new} = x_i + (x_{in} - x_i) \cdot \delta$$

where δ is a random number betIen 0 and 1.

GridSearchCV (GridSearch Cross Validation)

GridSearchCV performs hyperparameter optimization by searching through a predefined space of parameters. It evaluates each combination of parameters by training a model and computing the cross-validation score. The process can be summarized by:

$$\underset{\theta \in \Theta}{\operatorname{argmax}} \, CV(f(\theta, X_{train}), y_{train})$$

where Θ is the set of all possible combinations of parameters, f is the model trained with parameters θ, X_{train} and y_{train} are the training data and labels, and CV denotes the cross-validation function.

2.2 Machine Learning Model

XGBoost

XGBoost is an ensemble of decision trees algorithm that uses a gradient boosting framework. At each iteration, it adds a tree that best reduces the loss, by fitting the negative gradients. The prediction $\hat{y}_i$ at iteration t is given by:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta \cdot f_t(x_i)$$

where $\hat{y}_i^{(t-1)}$ is the prediction from the previous iteration, f_t is the new tree, and η is the learning rate.

LightGBM

LightGBM is a gradient boosting framework that uses tree-based learning algorithms and is designed for distributed and efficient training. It grows trees leaf-wise (best-first), choosing the leaf with maximum delta loss to grow. When growing the same leaf, it updates the prediction with:

$$\hat{y}_i = \hat{y}_i + \gamma$$

where γ is the adjustment for the chosen leaf based on the gradient and hessian of the loss function.

Multi-Layer Perceptron (MLP)

A Multi-Layer Perceptron (MLP) is a class of feedforward artificial neural network (ANN) that includes multiple layers of neurons: one input layer, one or more hidden layers, and one output layer. Each neuron in a layer connects to every neuron in the following layer, with the connections characterized by Iights. The process in each neuron involves summing the Iighted inputs and applying an activation function to produce an output.

The output o_j of each neuron is given by:

$$o_j = \sigma \left(\sum_{i=1}^{n} w_{ij} x_i + b_j \right)$$

where:

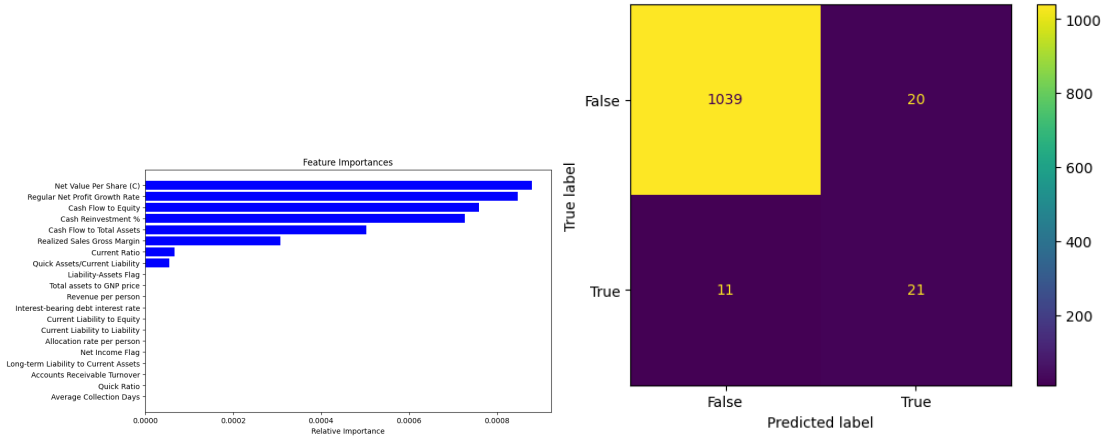
- o_j is the output of neuron j,
- σ denotes the activation function, such as sigmoid, tanh, or ReLU,
- w_{ij} represents the Iight from input i to neuron j,
- x_i is the input from neuron i,
- b_j is the bias term for neuron j,
- n is the number of inputs to the neuron.

MLP learns the correct Iights and biases through a process known as backpropagation, which involves calculating the gradient of the loss function with respect to each Iight by the chain rule, updating the Iights in the direction that minimally decreases the error.

3 Results

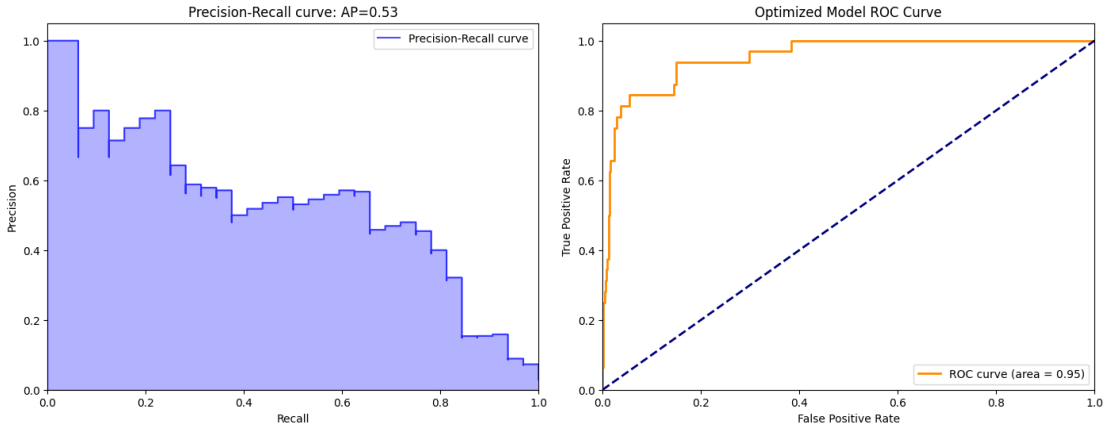
3.1 XGBoost

The XGBoost model demonstrates high accuracy (97.16%) and a strong ability to distinguish betIen classes (AUC: 0.9508). The F1-score, which combines precision and recall, is 0.5753. This score is moderate, reflecting that the model's precision and recall are balanced to some extent but also hinting at a potential imbalance in the class distribution that might be influencing the model's performance. While the model is good at identifying the majority class, as indicated by the accuracy and AUC, it seems to struggle somewhat with correctly classifying the minority class, which is often the more important class in many applications. This can be seen in the precision-recall curve and the quadrant of the confusion matrix. The feature importance graph indicates that the model relies heavily on certain features, which suggests that the dataset contains strong predictors for the target variable.



(a) Top 20 Feature Importances (b) Confusion Matrix

Figure 1: Feature importances and confusion matrix of the XGBoost model.



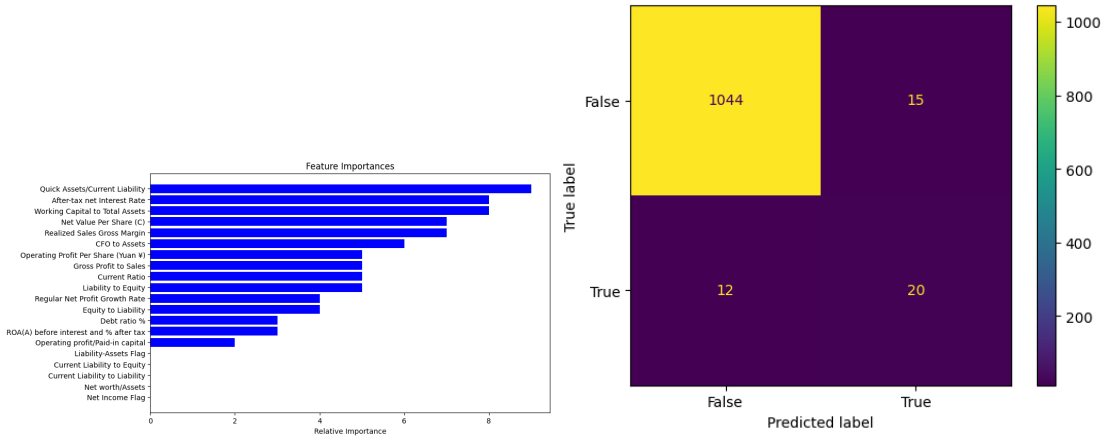
(a) Precision-Recall Curve (b) ROC Curve

Figure 2: Precision-recall curve and ROC curve of the XGBoost model.

3.2 LightGBM

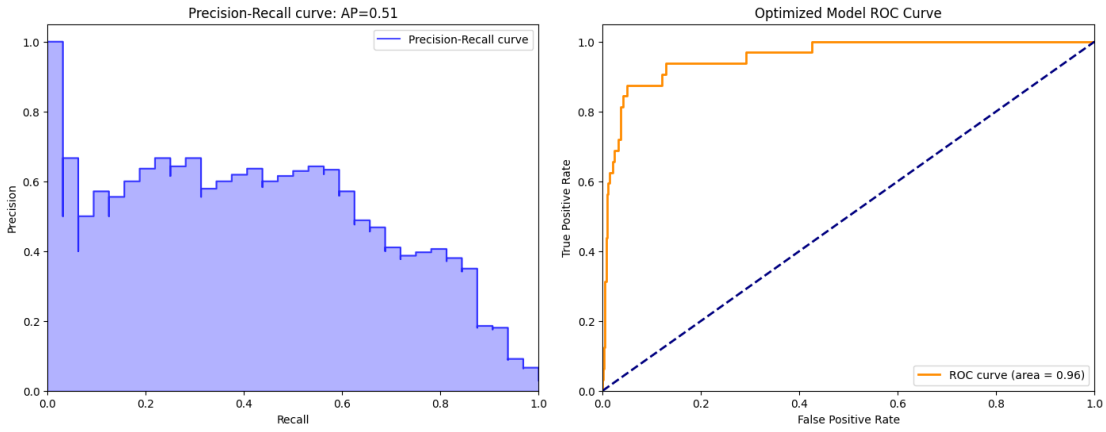
The LightGBM model showcases high performance with an accuracy of 97.80%, an F1-score of 0.6250, and an AUC of 0.9628, indicating its robustness in classification tasks. This performance is noteworthy because it combines high accuracy with significant measures of model efficacy such as the F1-score and AUC, highlighting its capability in handling classification tasks effectively. The F1-score, in particular, signifies a balance between precision and recall, suggesting that the model is adept at both identifying relevant instances and minimizing the number of irrelevant instances it selects. Moreover, the AUC value close to 1 indicates a strong ability of the model to discriminate between the classes effectively, which is critical for tasks where the distinction between classes

is vital. These metrics collectively underscore the LightGBM model's competence in not only accurately classifying instances but also in its nuanced capability to handle the complexities inherent in classification problems, such as imbalanced class distribution and the need for a harmonious balance between precision and recall. This robust performance makes the LightGBM model a compelling choice for various applications, particularly in scenarios where precise and reliable classification is paramount.



(a) Top 20 Feature Importances (b) Confusion Matrix

Figure 3: Feature importances and confusion matrix of the LightGBM model.

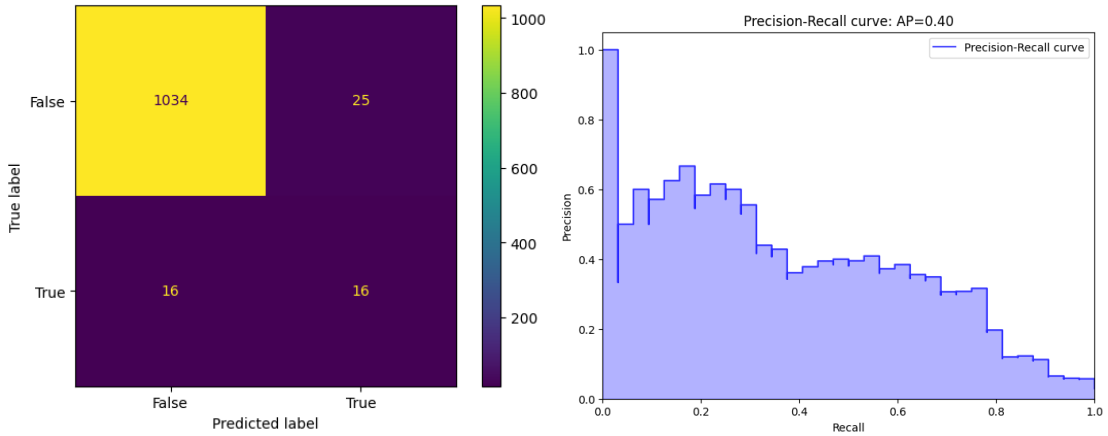


(a) Precision-Recall Curve (b) ROC Curve

Figure 4: Precision-recall curve and ROC curve of the LightGBM model.

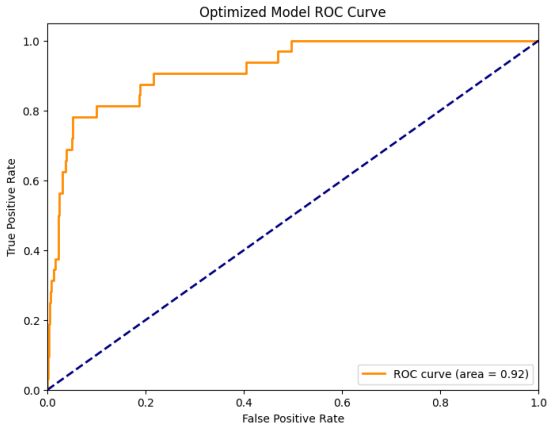
3.3 Multi-Layer Perceptron (MLP)

The Multi-Layer Perceptron (MLP) model exhibits solid performance with an accuracy of 96.79%, indicating its strong general predictive capabilities. The F1-score is 0.4615, which suggests that there is room for improvement in the balance betIen precision and recall, especially in the context of an imbalanced dataset. The AUC of 0.9274 demonstrates the model's good discriminative ability betIen the positive and negative classes.



(a) Confusion Matrix (b) Precision-Recall Curve

Figure 5: Confusion matrix and precision-recall curve of the MLP model.



(a) ROC Curve of the MLP model

3.4 Summary of Findings

In terms of feature importance, the divergence in the top 20 feature importances between XGBoost and LightGBM can be attributed to their distinct algorithmic foundations

and methodologies for tree construction. XGBoost employs a level-wise growth strategy for trees, which systematically evaluates all possible splits across all features, whereas LightGBM adopts a leaf-wise approach, focusing on the best split at a given leaf without evaluating every possibility. This fundamental difference influences how each model prioritizes and weights features based on their contribution to reducing loss. Additionally, variations in handling categorical data, missing values, and regularization techniques further differentiate the models' perceptions of feature importance. These discrepancies underscore the unique insights each model brings to feature relevance, hinting at their complementary potential in ensemble strategies for enhanced predictive performance.

The following features are deemed important by both the XGBoost and LightGBM models:

- Net Value Per Share (C)
- Quick Assets/Current Liability
- Working Capital to Total Assets
- Realized Sales Gross Margin
- Regular Net Profit Growth Rate

Based on the validation data performance of the three models in the table 1 and the grid-search calibration, LightGBM has achieved the optimal results, particularly with regard to the F1 Score. Hence, I will use the LightGBM model to predict the final outcome.

Model	**Before Optimization**	**After Optimization**
XGBoost	0.5333	0.5753
LightGBM	0.5882	0.6250
MLP	0.3692	0.4615

Table 1: F1-Scores before and after optimization for different models.

4 Conclusion

The study embarked on a comprehensive exploration of machine learning models to predict bankruptcy, leveraging a dataset replete with financial indicators. Initially, the investigation entailed rigorous preprocessing measures to establish a clean, normalized, and balanced dataset, utilizing techniques such as SMOTE for oversampling and StandardScaler for normalization. This meticulous preparation was instrumental in enhancing the quality of the predictive models.

The research scrutinized three distinctive models: XGBoost, LightGBM, and MLP. Each model underwent a systematic evaluation, starting with their default configurations, followed by an intensive hyperparameter tuning process with GridSearchCV, emphasizing the F1 score as the metric of optimization. The LightGBM model distinguished

itself, exhibiting exemplary performance metrics, with a commendable F1 score of 0.6250 and an impressive AUC of 0.9628. These metrics are demonstrative of the model's potent classification capabilities, particularly after the optimization process, which underscored the effectiveness of the hyperparameter tuning.

The precision-recall and ROC curves further corroborated the quantitative findings, providing visual affirmation of the LightGBM model's superior ability to discriminate between classes. Additionally, the feature importance analysis revealed that the model's predictive prowess was not serendipitous but rooted in the significant influence of certain key financial indicators.

In light of these results, it is unequivocal that the LightGBM model, with its optimized parameters, stands out as the most proficient tool for predicting bankruptcy in the context of the evaluated financial data. Consequently, the LightGBM model has been selected as the preferred model for making the final predictions, poised to serve as a reliable asset in the domain of financial risk assessment.