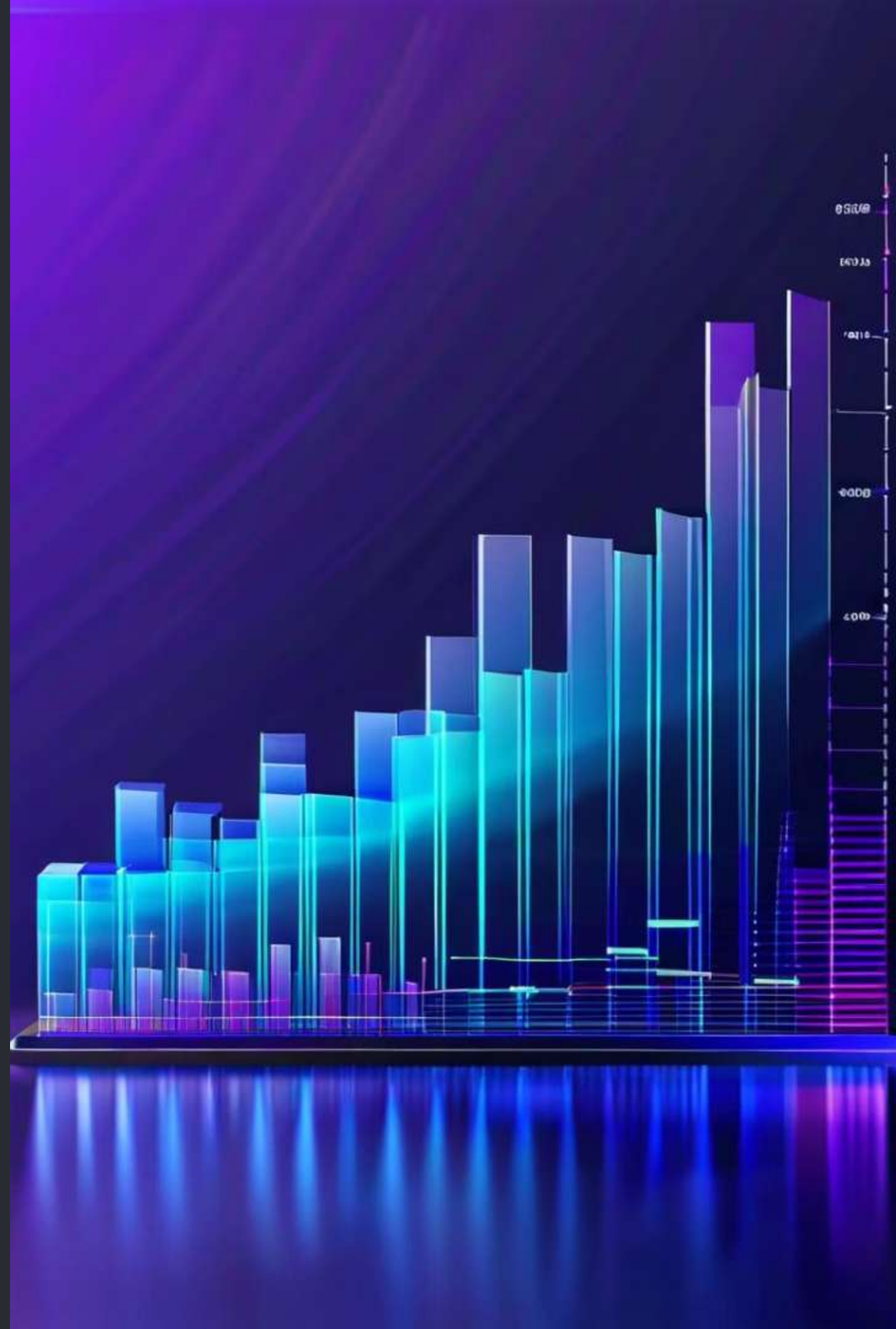
Bankruptcy Prediction Using Machine Learning



Abstract

This project focuses on utilizing machine learning techniques to predict company bankruptcy. By analyzing a comprehensive dataset of financial ratios, we seek to develop robust predictive models that can accurately forecast the likelihood of a business facing financial distress or insolvency. Through feature engineering, model selection, and ensemble techniques, we aim to provide a reliable and actionable tool for decision-makers in the financial industry.

Dataset Overview

The dataset for this project was collected from the **Taiwan Economic Journal**, covering the years 1999 to 2009. It includes 95 financial ratios for each company, which serve as the features for our bankruptcy prediction models. The target variable indicates whether a company has declared bankruptcy based on the business regulations of the Taiwan Stock Exchange. Careful analysis and feature selection from this comprehensive dataset will be crucial in developing accurate and reliable predictive models.

Methodology

Data Preparation

We begin by importing essential libraries for data manipulation, analysis, and machine learning, such as pandas, numpy, and scikit-learn. The dataset is then loaded from a CSV file into a DataFrame, and data quality is assessed by checking for missing values and duplicate rows, ensuring the reliability of the data for model training.

Feature Engineering

Feature engineering is a crucial step in our methodology, where we identify the most relevant features for predicting bankruptcy. We employ the SelectKBest method with the ANOVA F-value metric to select the top 25 features based on their relationship with the target variable, bankruptcy status.

Correlation Analysis

To further refine our feature set, we plot a correlation heatmap for the top 25 selected features. This allows us to identify highly correlated features (correlation above 0.85) and remove redundant ones, retaining only the essential predictors. This process leads to a final set of 13 features that will be used to train the bankruptcy prediction models.

Methodology

Class Imbalance Handling

Given the inherent class imbalance in the dataset, with fewer bankrupt companies compared to non-bankrupt ones, we employ the SMOTE (Synthetic Minority Over-sampling Technique) to balance the dataset. This oversampling approach helps prevent biased model training and improves the overall predictive performance.

Model Development

We then implement various classification models, including Decision Tree, Random Forest, Logistic Regression, Naive Bayes, SVM, and XGBoost, to predict bankruptcy. These models are evaluated on both balanced and unbalanced datasets, with hyperparameter tuning performed via Bayesian optimization (BayesSearchCV) to optimize their performance.

Ensemble Modeling

For our final model, we leverage an ensemble approach by combining multiple machine learning algorithms using a Voting Classifier. This technique aggregates the predictions of individual classifiers, such as Random Forest, SVM, XGBoost, Naive Bayes, Logistic Regression, and Decision Trees, to achieve improved accuracy and robustness in bankruptcy prediction.

Ensemble Classifier Configuration

Random Forest

The Random Forest Classifier, a robust ensemble method, is included in the Voting Classifier to leverage its ability to handle complex, nonlinear relationships and provide variable importance insights.

SVM

The Support Vector Machines (SVM) algorithm, known for its strong generalization capabilities, is incorporated to further enhance the ensemble's performance in accurately classifying bankruptcy cases.

XGBoost

XGBoost, a highly efficient and scalable gradient boosting algorithm, is utilized in the ensemble to leverage its superior predictive power and ability to handle large-scale, high-dimensional data.

Naive Bayes

The Naive Bayes classifier, with its simplicity and robustness, is included in the ensemble to provide a complementary perspective and potentially improve the overall model's performance.

Ensemble Performance

Metric	Value
Accuracy	0.9366
Precision	0.9061
Recall	0.9730
F1 Score	0.9384
AUC-ROC	0.9369
Confusion Matrix	[[1798 198] [53 1911]]

The ensemble classifier demonstrated robust performance across multiple evaluation metrics, indicating its effectiveness in accurately classifying instances in our dataset. By leveraging the diverse strengths of individual classifiers, the ensemble approach achieved superior predictive accuracy and generalization capability.

Conclusion

Ensemble Effectiveness

The ensemble classifier demonstrated robust performance across multiple evaluation metrics, indicating its effectiveness in accurately classifying instances in our dataset. By leveraging the diverse strengths of individual classifiers, the ensemble approach achieved superior predictive accuracy and generalization capability.

Practical Applications

Investors, whether they are shareholders or bondholders, can use bankruptcy prediction models to evaluate the risk associated with investing in a particular company. Internally, companies can use bankruptcy prediction models as a tool for risk management. Insurance companies can use bankruptcy prediction models to assess the risk of insuring certain businesses.

Future Directions

Advancements in machine learning algorithms, particularly deep learning techniques, can lead to more sophisticated models capable of capturing complex relationships and patterns in financial data. Further developments involve analysing similar models on different country datasets to identify key ratios used universally and establish their importance