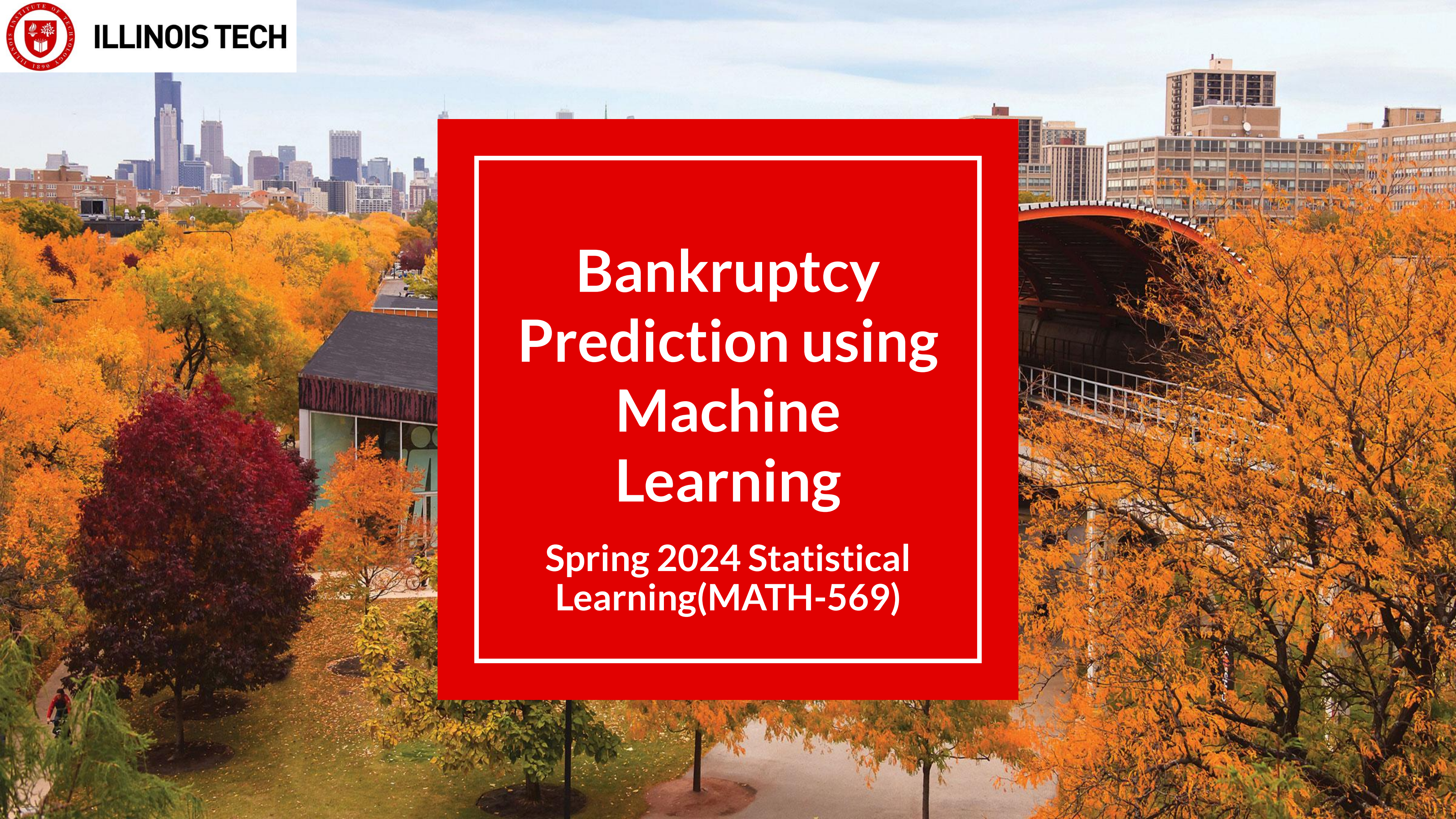


ILLINOIS TECH

Bankruptcy Prediction using Machine Learning

Spring 2024 Statistical Learning(MATH-569)

Motivation

Problem Statement:

- Predicting corporate bankruptcy is a critical challenge in financial industry
- Traditional methods often fail to capture complex patterns and relationships in financial data
- Machine learning techniques can potentially improve bankruptcy prediction accuracy

Importance of Accurate Bankruptcy Prediction:

- Early identification of bankruptcy risk enables proactive measures to minimize financial losses
- Accurate predictions support informed decision-making for investors, creditors, and regulators
- Helps in maintaining stability and trust in the financial market

Dataset Overview:

- Comprehensive dataset with 95 financial ratios and indicators
- Contains data from 6819 companies over a period of 10 years
- Includes a diverse range of industries and company sizes
- Binary target variable: "Bankrupt?" (1 for bankrupt, 0 for non-bankrupt)

Benchmark Model

Followed procedure described in the paper by Musa et Al.

Only three variables included:

- Liability to Assets: Total Debt/Total Assets
- Current Ratio : Current Assets/Current Liabilities
- ROA before interest and depreciation after tax: Net income/Total Assets

-Model used Extreme Gradient Boosting (XGBoost)

- Coded using XGBoost library in Python.

Benchmark Model Results

-Accuracy: 0.828

-Precision: 0.286

-Recall: 0.078

-F1 Score: 0.123

-Accuracy of 83% is our benchmark

Exploratory Data Analysis

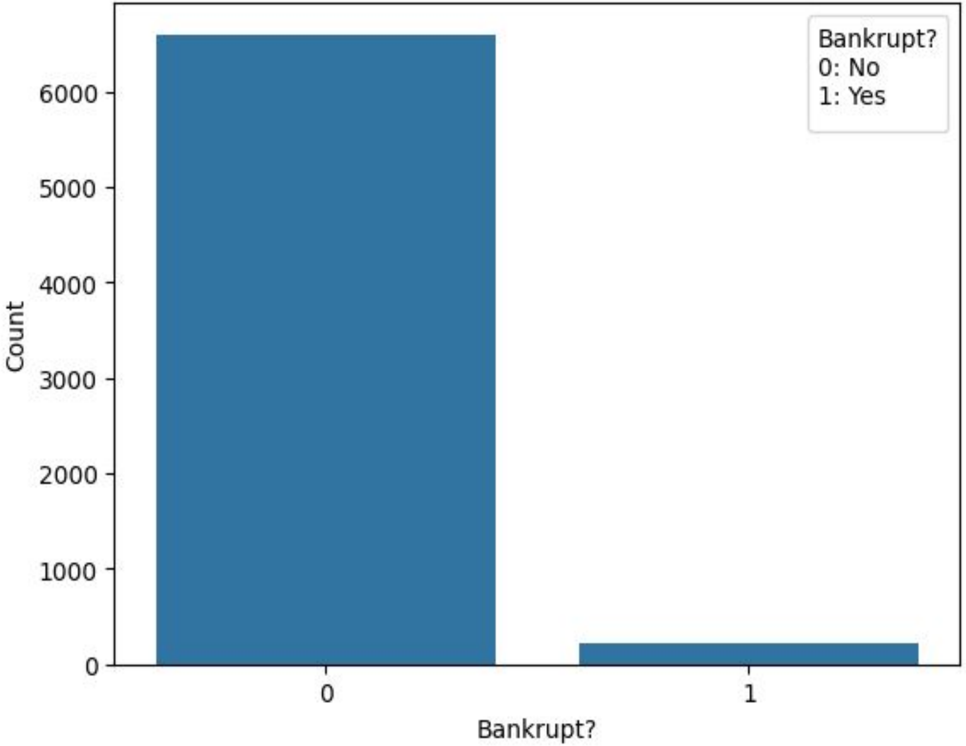
- Checking missing values, repeated columns or no information characteristic

- Deleted: In our case only two column with no information. ('Net Income Flag', 'Liability-Assets Flag')

- 94 features remain.

ILLINOIS TECH

Data Imbalance



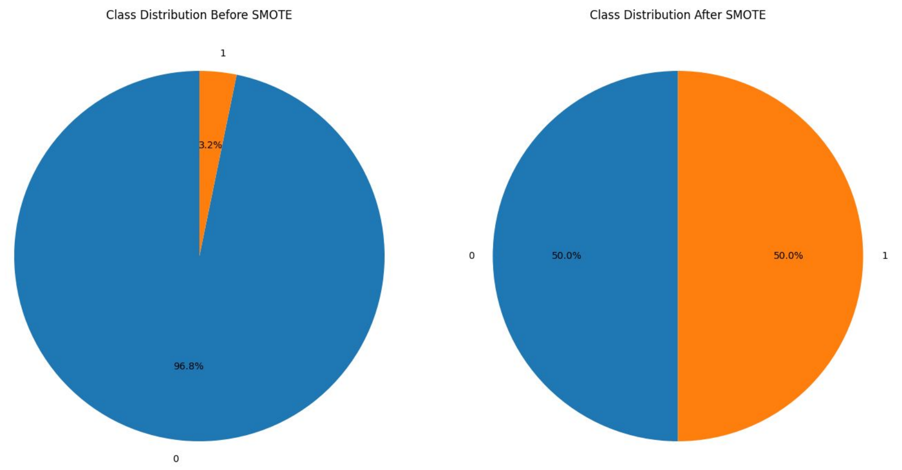
Bankrupt?	
0 (No)	6599
1 (Yes)	220

ILLINOIS TECH

Data Augmentation (SMOTE)

Synthetic Minority Oversampling Technique:

- Creating artificial data for bankruptcy by Increasing number of no. bankruptcy records.
- Statistical technique to increase number of cases in dataset
- Generate new instances from existing minority class , New instances not copies of existing minority cases



Data Normalization

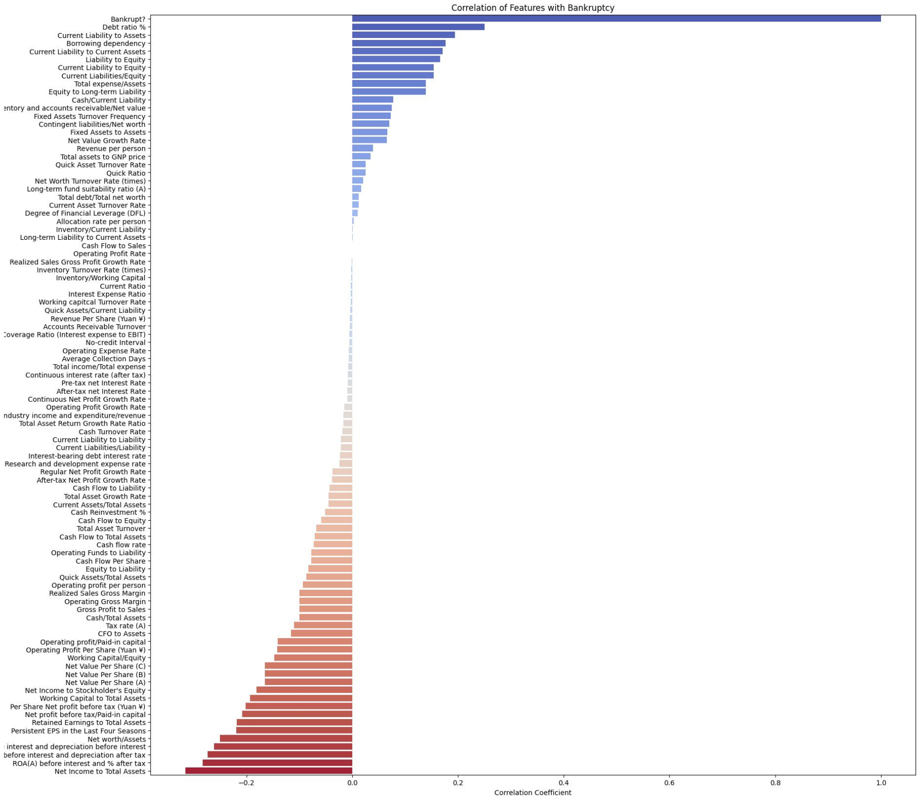
Feature Normalization:

- Applied Z standardization to standardize features
- Removing the effect of the scale in the dataset.

$$Z = \frac{x - \mu}{\sigma}$$

ILLINOIS TECH

Explanation Power



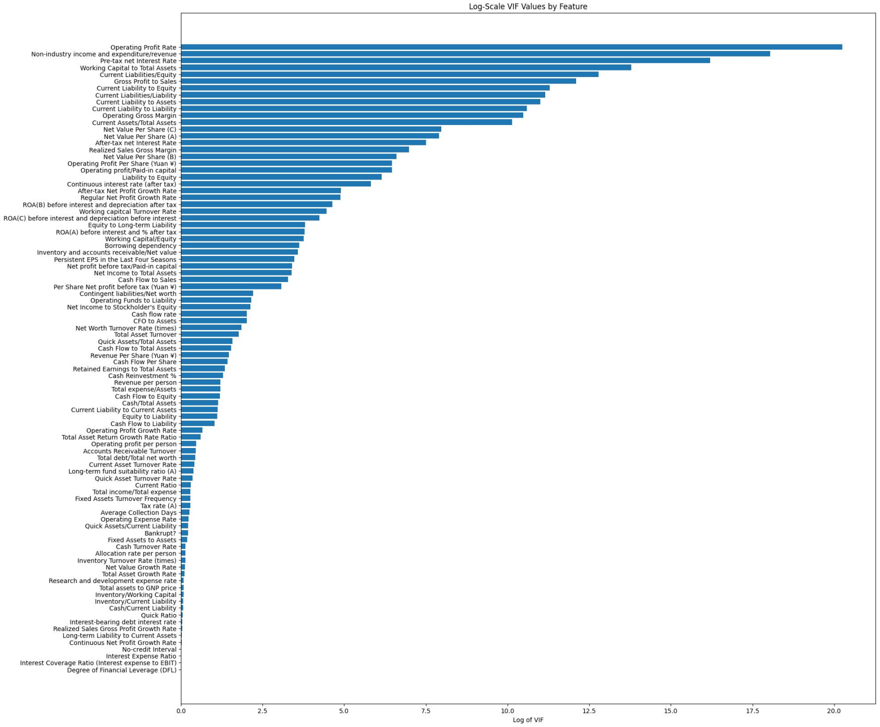
Collinearity: a quick overview

ILLINOIS TECH

More quantitative approach: VIF

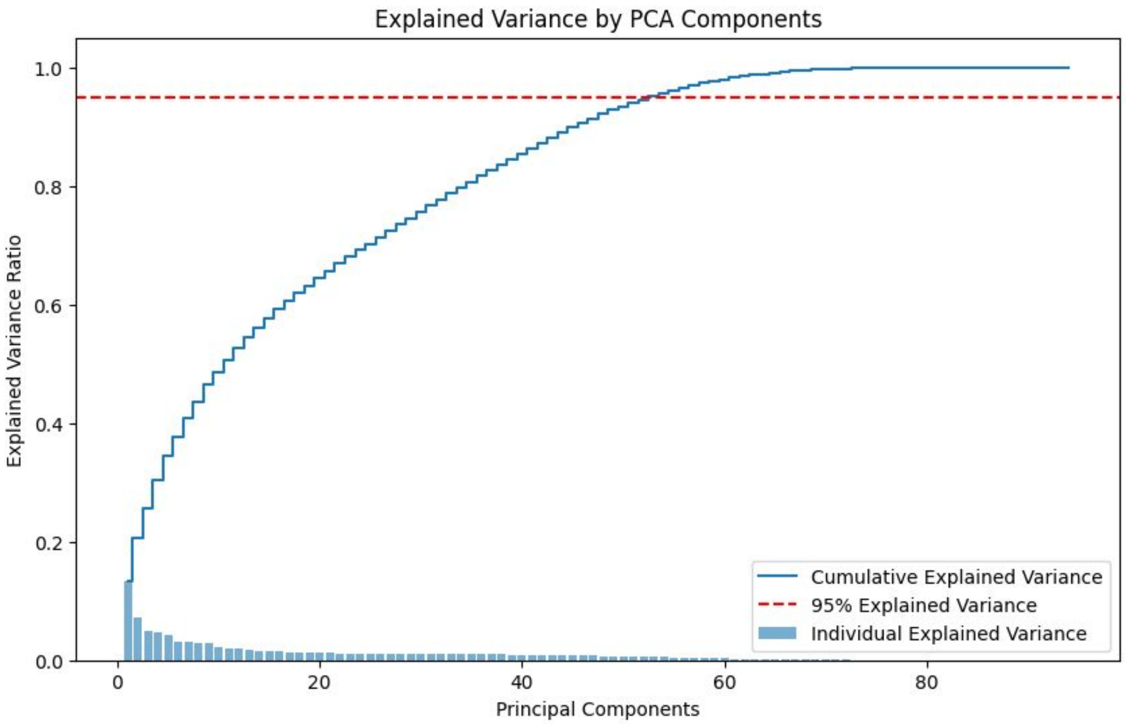
- Collinearity present with VIF values over 100.

- Some features VIF values over 600000000, strong present of collinearity.



ILLINOIS TECH

Solve Collinearity: PCA



General Pipeline of the Data and Process

Split Data randomly in 20/80 train test.

Train: Argumentation of the data -> Normalization of the train Data -> PCA transformation and selection of the data -> train the model

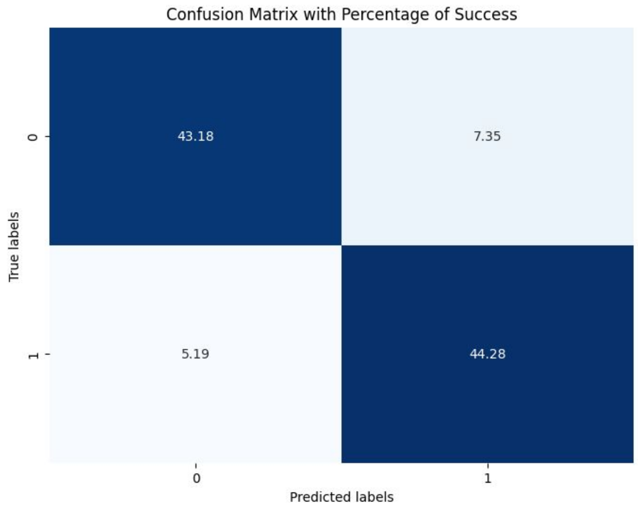
With the Characteristics learning on the pipeline

Test: Normalization of the Data with the train values -> PCA with the matrix get it in the train -> prediction with the model

Model Development Logistic Regression

Logistic Regression with PCA:

- Cross-validation:
  - Performed 5-fold cross-validation
  - Average accuracy: 96.72%
  - Consistent high accuracy across all folds
- Potential Overfitting Concerns:
  - High accuracy might be influenced by multicollinearity in the dataset
  - Regularization techniques like L1 or L2 can be applied
- Outliders:
  - PCA problems to manage outliers
- Inbalance:
  - No consistent predictions to
- Interpretability
  - It can be interpretable undoing the PCA transformation

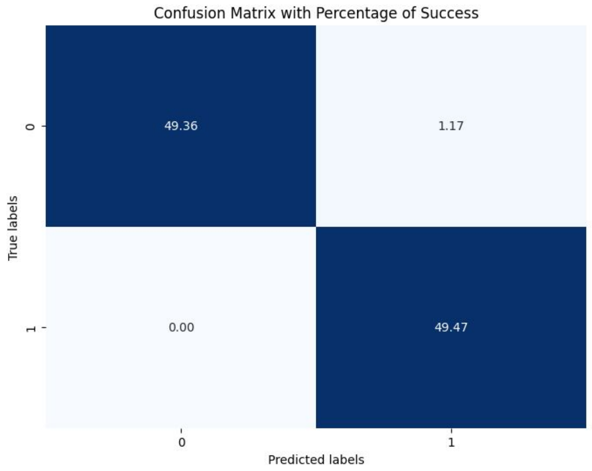


ILLINOIS TECH

Model Development XGBoost without PCA:

Logistic Regression with PCA:

- Cross-validation:
  - Performed 5-fold cross-validation
  - Average accuracy: 98.18%
  - Consistent high accuracy across all folds
- Potential Overfitting Concerns:
  - Difficult to select the number of hyperparameters
- Outliers:
  - XGBoost can manage that
- Inbalance:
  - Similar behavior classifying 0 and 1. More consistent.
- Interpretability:
  - XGBoost can not be interpretable BlackBox model.



Recursive Feature Elimination (RFE)

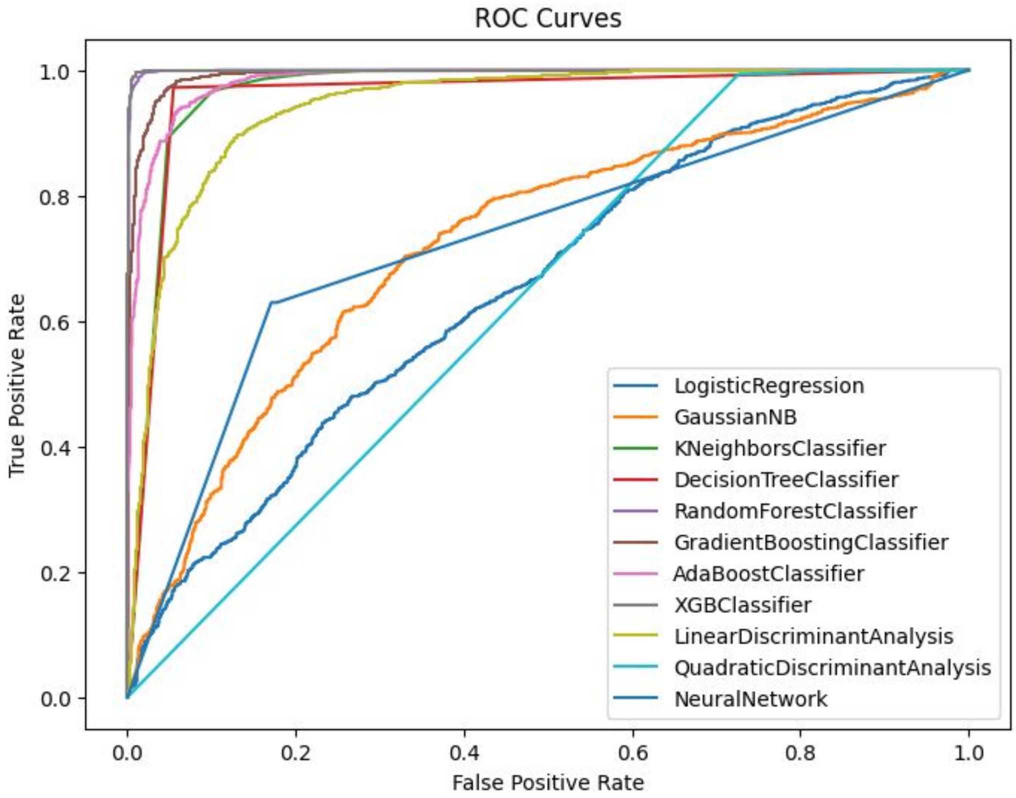
- Algorithms RFE was applied to:
  - Logistic regression
  - Decision tree classifier
  - Random forest classifier
  - Gradient boosting classifier
  - AdaBoost classifier
  - XGBoost classifier
- Total 55 features left

Features	LogisticRegression_Ranking	DecisionTree_Ranking	RandomForest_Ranking	GradientBoosting_Ranking	AdaBoost_Ranking	XGBoost_Ranking
ROA(C) before interest and depreciation befor...	8	1	1	1	1	1
Operating Gross Margin	9	1	14	14	16	1
Operating Profit Rate	1	5	1	1	9	2
Non-industry income and expenditure/revenue	15	1	1	1	8	1
Operating Expense Rate	1	1	21	22	7	18
Research and development expense rate	1	1	4	1	1	1
Cash flow rate	11	25	1	1	1	1
Interest-bearing debt interest rate	1	1	1	1	1	1

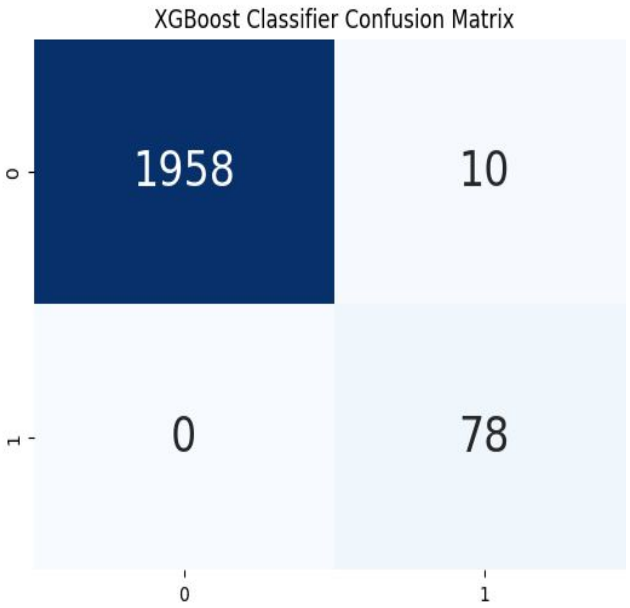
RFE Ranking for different features and models

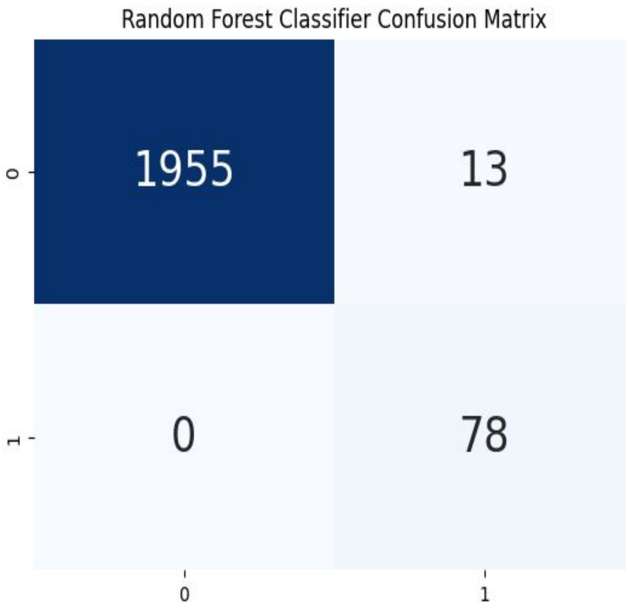
Model Evaluation and Comparison

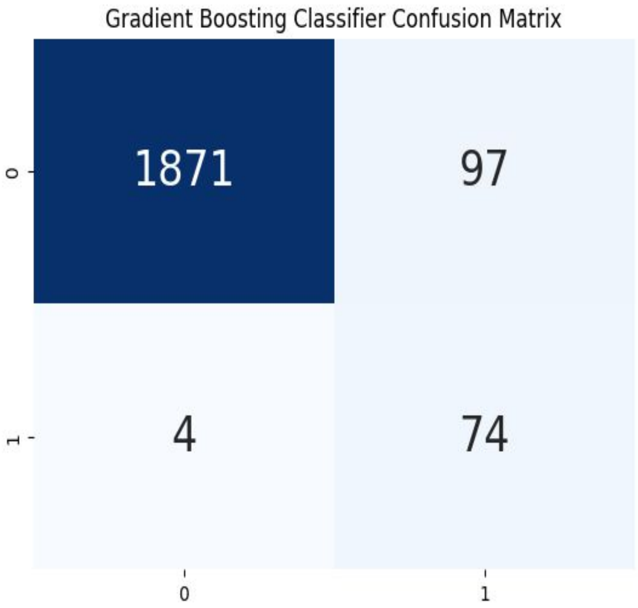
- Best model performance:
  - Random forest classifier
  - XGBoost classifier
  - Gradient boosting classifier



Model Evaluation and Comparison







ILLINOIS TECH

Future Enhancements

Address Multicollinearity:

- Feature Selection: Lasso Regularization, Domain Knowledge
- Regularization: L1, L2, Elastic Net

Explore Other Algorithms:

- SVM: High-dimensional, Kernel Trick, Imbalanced Data
- Neural Networks: Deep Learning, Complex Patterns

Feature Engineering:

- Domain Knowledge: Financial Ratios & Indicators
- Interaction Terms & Polynomial Features
- Temporal Features: Trends, Seasonality, Cycles

Handle Class Imbalance:

- Oversampling: Random
- Undersampling: Random, Cluster Centroids
- Cost-Sensitive Learning

ILLINOIS TECH

Conclusion

Key Findings:

- Strong Correlations & Multicollinearity
- Important Predictors: Debt, Profitability, Liquidity Ratios
- XGBoost and random classifier performing best on features selected after RFE

Model Interpretability & Explainability:

- Understanding Driving Factors
- Feature Importance Analysis
- Transparent & Explainable Models

Applications:

- Credit Risk Assessment & Loan Approval
- Investment Screening & Portfolio Management
- Risk Management & Financial Stability Monitoring
- Early Warning System for Financial Distress

Interpretability vs Accuracy

- Linear models are not able to capture all the information but are interpretable
- Non-linear models and tree family models get better results
- Depends of the context it should be select one type of model or other.

References and Acknowledgments

[1] Dataset: Taiwanese Bankruptcy Prediction. (2020). UCI Machine Learning Repository. https://doi.org/10.24432/C5004D.

[2] Research Study https://www.mdpi.com/1911-8074/15/1/35

[3] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer, “SMOTE: synthetic minority over-sampling technique”, J. Artif. Int. Res.,vol. 16, Jan. 2002, pp. 321–357, doi: 10.1613/jair.953.

[4] PCA: https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

[5] RFE: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html

Thank You!

- Thank you for your attention!
- Questions and discussion