ROBERT H. SMITH
SCHOOL OF BUSINESS

BUDT704: DATA PROCESSING AND ANALYSIS IN PYTHON

"**Bankruptcy Prediction Using Machine Learning**"

DATA SCIENCE PROJECT

By
GROUP E

ARADHYA SINGH BISHT
EKATERINA DYACHENKO
GARIMA GUPTA
PRIYANKA GONA
SIMRAN GALLANI
YASHVI MOHTA

INTRODUCTION:

The fear of a financial collapse is a significant issue for businesses across many industries in the fast-paced, modern business environment. Bankruptcy is a multifaceted problem impacting employees, creditors, investors, and the economy.

Recognizing the critical importance of identifying and mitigating the risks associated with bankruptcy, this study employs machine learning to examine historical financial data.

Bankruptcy Predictions: Their Significance

Any failing business often faces the possibility of bankruptcy. Financial difficulties are a sign of this. Investors and stakeholders may face financial ruin if they invest in or are part of a firm about to file for bankruptcy. Early warning sign recognition is essential for reducing losses and allocating resources appropriately. Avoiding this is crucial.

OBJECTIVE:

The main goal of this project is to dive deep into exploring, building, and assessing predictive models that can distinguish thriving businesses from those facing financial instability. We'll employ advanced machine learning techniques like random forest regression, K-nearest neighbors (KNN), and decision trees to determine crucial financial factors. We aim to discover which model best forecasts the possibility of bankruptcy accurately. To make these findings more straightforward, we'll use sophisticated visualizations to highlight the importance of these metrics. Ultimately, we want to pinpoint the model that outshines others in accurately predicting financial distress.

BUSINESS TRANSACTIONS

1). What factors have the most impact on whether a company is going to go bankrupt or stay afloat?
2). How does the financial performance of companies, including metrics such as EBITDA, net income, and gross profit, influence their likelihood of facing bankruptcy?
3). How can operational efficiency be measured using variables like cost of goods sold, total operating expenses, and EBIT, and what impact do these metrics have on the bankruptcy prediction model?
4). How reliable and accurate are the bankruptcy prediction models when evaluated on the validation set and the test set, and what implications does this have for real-world scenarios involving unseen cases?

DATA PROCESSING & ANALYSIS

DATASET DESCRIPTION

This dataset is from Kaggle and focuses on bankruptcy prediction encompasses data related to American public companies listed on the New York Stock Exchange and NASDAQ. The dataset comprises 78,682 observations spanning 1999 to 2018, encompassing 8,262 distinct companies. Featuring 18 key features (X1-X18) detailing financial aspects, including status labels indicating company status (active or bankrupt), it offers insights for analyzing business dynamics and longevity over time.

https://www.kaggle.com/datasets/utkarshx27/american-companies-bankruptcy-prediction-dataset

1. DATA PROCESSING TASKS

1.1. DATA RETRIEVAL

- Acquiring the dataset from the internet(Kaggle)
- Attaching it to the collab file

1.2. DATA CLEANING

- Seeing a summary of the dataset and analyzing if the data types are appropriate(through **info(), head(), and tail()** functions)
- Formatting Column Names according to their variable description make them understandable.(column names to more descriptive names for improved understandability and readability through **rename()** function of Pandas library)
- Checking for Null Values and Duplicate Columns(through **info()** it's inferable that there were 0 null entries across the 20 columns, we employed the **isnull()** function of the Pandas library as a convention. Additionally, we employed the duplicated() function of the Pandas library to search for duplicate columns)

1.3. EXPLORATORY DATA ANALYSIS

1.3.1. Summary Statistics

The **describe()** function of the Pandas library is employed to calculate and present summary statistics for numerical columns in the dataset. The summary statistics encompasses measures namely **count, mean, standard deviation, minimum, 25th, percentile (Q1), median (50th percentile or Q2), 75th percentile (Q3), and maximum values** for each numerical column. Through these measures of central tendency and dispersion we developed a basic understandability of our data and drew key information.

1.3.2. Pairwise Correlation

Highly correlated features may not provide additional information, and removing them can improve model performance and interpretability.

This analysis helps identify multicollinearity (High correlation resulting in difficulties discerning the contribution of each variable). We employed the **corr()** function of the Pandas library for displaying a correlation matrix and the **heatmap()** method of the Seaborn library for displaying the heatmap.

We saw that Total Revenue and Net Sales's correlation coefficient is exactly 1 indicating that these are the same

1.3.3. VIF

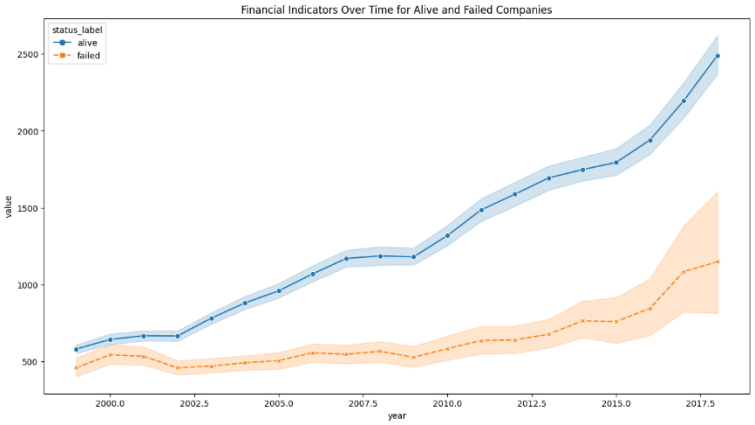
We calculated VIF through the **variance_inflation_factor()** function of the outliers_influence submodule's stats module of the Statsmodels library. VIF

assess the increase in variance of an estimated regression coefficient if the predictors are correlated. It measures the extent of multicollinearity. Through this calculation, we have further confirmation that perfect multicollinearity is present between 1 or more variables. In this case, the variables are Net Sales and Total Revenue. Hence, they need to be removed.

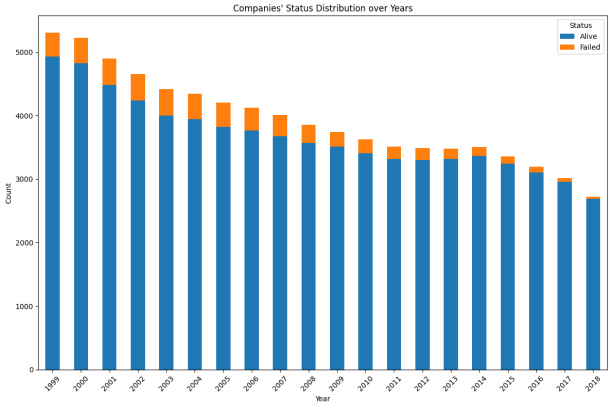
1.3.4. Distribution Visualization

For a clear and concise overview of the data which aids us in the identification of patterns, outliers, and further analysis, we visualized the distribution of our data through:

- Line Plot: Line plot depicts the change over time for key financial indicators (X1 to X18) for both 'alive' and 'failed' companies. The data is melted through the **melt()** function of the Pandas library to create a long-form representation, allowing us to plot the evolution of each financial indicator over the years.

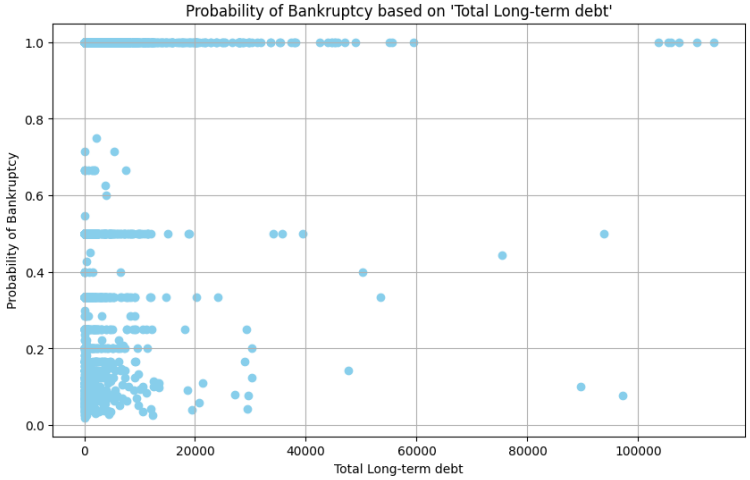


- This chart depicts the total count of companies (Y-Axis) for each year (X-Axis). Each bar represents a specific year, and the bars are divided into segments where each represents the count of 'alive' or 'failed' companies. Over the years, the count of active companies significantly declined, and correspondingly, the count of failed companies increased. This allows for comparing trends in financial indicators between 'Alive' and 'Failed' companies over time.



1.3.5. Bayes Rule

The plot_bankruptcy_probability function displays the probability of bankruptcy concerning specific feature values. This calculation, derived using Bayes Rule, explores the relationship between a feature and a company's likelihood of financial distress. It aids in understanding feature impact on bankruptcy probabilities, facilitating vital insights into financial risk assessment.



1.3.6. Principal Component Analysis

High multicollinearity, indicated by strong correlations and VIF warning, necessitates PCA in data analysis. PCA resolves correlated variables, reducing redundancy while retaining essential data traits. It mitigates multicollinearity issues, ensuring better model stability, interpretability.

STEPS

- We used Integrated Scikit-Learn modules: PCA for dimensionality reduction, LabelEncoder for categorical label encoding(Alive is 0 and Failed is 1), and StandardScaler for feature normalization.
- Computed explained variance ratio to evaluate information retained by principal components generated via PCA. Conducted a post-PCA reassessment of VIF, revealing values near unity, indicating effective resolution of multicollinearity issues, and improved dataset suitability for analysis.

2. DATA ANALYSIS

2.1. MODEL SELECTION & TRAINING

2.1.1. Random Forest Classifier

The Random Forest Classifier's ability to handle complex relationships, reduce overfitting, and provide insights into feature importance Hence, makes it an effective choice for bankruptcy prediction tasks. It provides a feature importance score, allowing us to understand which features contribute the most to predicting bankruptcy. This helps in identifying critical financial indicators.

2.1.2. Decision Trees Classification

Decision tree classification is a predictive method that partitions datasets based on feature values, forming a tree structure. It interprets features as decision nodes, branches as criteria, and leaves as outcome labels. Offering clear insights into factors contributing to financial stability or distress in companies.

2.1.3. K-nearest neighbor Classification K-nearest neighbors are a machine learning algorithm used for classification and regression analysis.

It helps in assessing the financial data of companies in this dataset to foresee bankruptcy by comparing new data with similar historical cases. It does categorization using proximity to 'k' neighbors with known labels.(status = “alive”,” failed”)

We executed the fundamental ML model lifecycle, opting for classification due to categorical targets. Employing Scikit-learn's train_test_split, we segregated data (X_pca, y) into 80% training and 20% testing. Subsequently, we trained the classifier and utilized the model to predict labels for evaluation and analysis.

2.2. MODEL EVALUATION

We evaluated the models through common evaluation metrics,

The **accuracy_score** and **classification_report** functions from the Metrics module of Scikit-learn are used to evaluate the performance of a classification model.

2.2.1. ACCURACY:

Accuracy is the ratio of correctly predicted instances to the total instances

2.2.2. ROC -AUC CURVE

The **roc_auc_score** and **roc_curve** functions from the Metrics module of Scikit-learn are used to evaluate the performance of a classification model.

Scikit-learn's Metrics module uses roc_auc_score and roc_curve functions to evaluate classification models. ROC-AUC measures how well a model distinguishes classes, represented by the area under the curve, while the ROC curve visually shows how the model performs across different probability thresholds.

2.2.3. CLASSIFICATION REPORT

The classification report includes precision, recall, f1-score, and support for each class, as well as the average metrics. **Precision** is the ratio of true positive predictions to the total predicted positives. It measures the accuracy of positive predictions. **Recall** (or sensitivity) is the ratio of true positive predictions to the total actual positives. It measures the model's ability to capture all positive instances. **F1-score** is the harmonic mean of precision and recall, providing a balance between the two.

The classification report demonstrates significantly higher recall, precision, and F1 scores for one label (0) compared to the other (1) across all models. This discrepancy signals a potential data imbalance requiring attention and resolution to enhance model accuracy and performance.

Decision Tree

	precision	recall	f1-score	support
0	0.94	0.94	0.94	14669
1	0.21	0.22	0.21	1068
accuracy			0.89	15737
macro avg	0.57	0.58	0.58	15737
weighted avg	0.89	0.89	0.89	15737

Random Forest

	precision	recall	f1-score	support
0	0.93	1.00	0.97	14669
1	1.00	0.03	0.05	1068
accuracy			0.93	15737
macro avg	0.97	0.51	0.51	15737
weighted avg	0.94	0.93	0.90	15737

K-Nearest Neighbors

	precision	recall	f1-score	support
0	0.93	1.00	0.97	14669
1	0.77	0.03	0.05	1068
accuracy			0.93	15737
macro avg	0.85	0.51	0.51	15737
weighted avg	0.92	0.93	0.90	15737

2.2.4. BALANCING THE DATASET

Due to the inferences above we chose to balance the data.

We balanced our data through **the SMOTE** class of the over_sampling module of the Imbalances library and re-ran each model with resampled data. SMOTE (Synthetic Minority Over-sampling Technique) identifies the minority class and generates synthetic samples for the minority class based on the existing samples.

Decision Tree

	precision	recall	f1-score	support
0	0.88	0.83	0.85	14588
1	0.84	0.89	0.86	14797
accuracy			0.86	29385
macro avg	0.86	0.86	0.86	29385
weighted avg	0.86	0.86	0.86	29385

Random Forest

	precision	recall	f1-score	support
0	0.96	0.94	0.95	14588
1	0.94	0.96	0.95	14797
accuracy			0.95	29385
macro avg	0.95	0.95	0.95	29385
weighted avg	0.95	0.95	0.95	29385

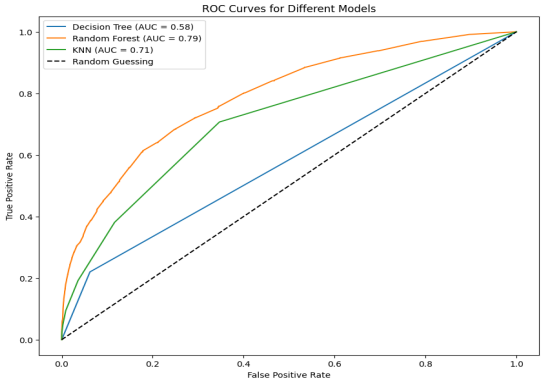
K-Nearest Neighbors

	precision	recall	f1-score	support
0	0.99	0.76	0.86	14588
1	0.81	0.99	0.89	14797
accuracy			0.88	29385
macro avg	0.90	0.88	0.88	29385
weighted avg	0.90	0.88	0.88	29385

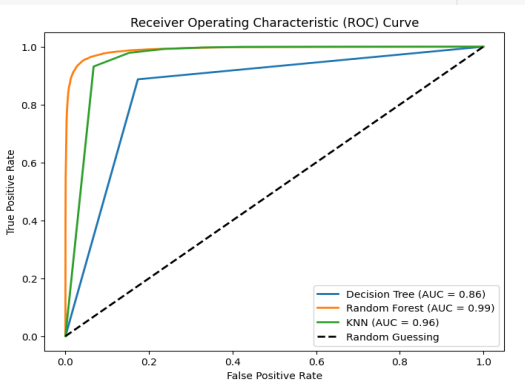
2.3. VISUALIZATIONS

We visualized the ROC-AUC curves and conducted an analysis of feature importance for both Decision Tree and Random Forest models. Using the 'feature_importances_' function, we aimed to comprehend the primary contributors to bankruptcy prediction within each model.

ROC -CURVE BEFORE BALANCING

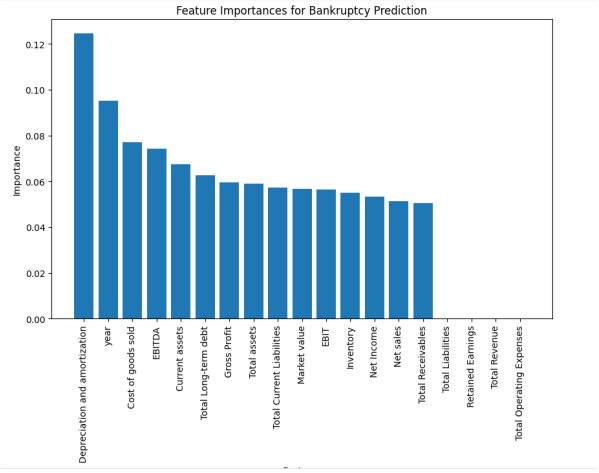


ROC -CURVE AFTER BALANCING

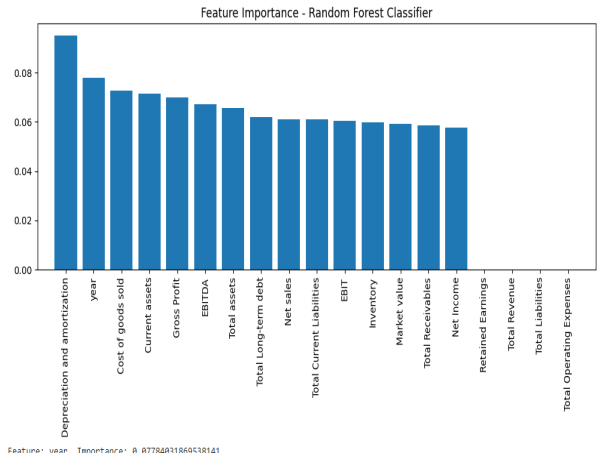


FEATURE SELECTION GRAPHS:-

DECISION TREE



RANDOM FOREST CLASSIFIER



Feature selection graph highlights metrics' impact on predictions; highest scores like Depreciation & Amortization, Year, and Cost of Goods Sold stand out prominently.

3. RESULTS & CONCLUSIONS

3.1 RESULTS

After the balancing there is a a positive impact on the performance of the models. The accuracy has improved, and the classification report shows better balance in precision, recall, and F1-score for both classes (0 and 1).

Results

Model	Accuracy Score	AUC-ROC
Pre Balancing		
Decision Tree	0.8992	0.5790
Random Forest	0.9339	0.7911
K Nearest Neighbor	0.9333	0.2896
Post Balancing		
Model	**Accuracy Score**	**AUC-ROC**
Decision Tree	0.8572	0.8570
Random Forest	0.9522	0.9906
K Nearest Neighbor	0.8792	0.9581

3.2 CONCLUSIONS

- **Key Factors Influencing Bankruptcy Prediction:**
  Critical elements impacting a company's financial health are Depreciation & Amortization, Year and Cost of Goods Sold.
- **Financial Metrics and Bankruptcy Likelihood:**
  Enhancing or maintaining EBITDA, net income, and gross profit is crucial to mitigate the likelihood of bankruptcy.
- **Operational Efficiency and Bankruptcy Prediction:**
  Efficient operations, determined by lower cost of goods sold and total operating expenses relative to earnings, alongside higher EBIT, significantly influence bankruptcy prediction models.
- **Reliability of Bankruptcy Prediction Models:**
  While models demonstrate accuracy through various metrics and classification reports during validation and test set evaluations, unforeseen factors in real-world scenarios may pose ongoing complexities.
- **Best Performing Model:**
  The Random Forest model showcases superior accuracy (96.13%), a high ROC-AUC value of 0.99, and notable precision and recall values. Its balanced performance suggests it as the most effective model among those evaluated for bankruptcy prediction.

4. CONCLUDING STATEMENTS

Summing up our project journey, we explored predicting company bankruptcies, and uncovering vital factors shaping financial stability. Our analysis emphasized the importance of financial metrics and operational efficiency in determining bankruptcy risks. Among our models, the Random Forest stood out, showcasing strong accuracy and insights into financial distress prediction. However, the real world's complexities remind us of the need for ongoing refinement and adaptability in practical scenarios

REFERENCES

- Verbeck, T. (n.d.). *The different types of bankruptcy with their meanings and definitions*. Business Insider. https://www.businessinsider.com/personal-finance/what-does-bankruptcy-mean-definition
- Singh, U. (2023, May 27). *US company Bankruptcy Prediction Dataset*. Kaggle. https://www.kaggle.com/datasets/utkarshx27/american-companies-bankruptcy-prediction-dataset
- *5 common causes of small business bankruptcy*. RBC Royal Bank. (2023, September 22). https://www.rbcroyalbank.com/en-ca/my-money-matters/debt-and-stress-relief/bankruptcy/business-bankruptcy/5-common-causes-of-small-business-bankruptcy/
- *What are machine learning models?*. Databricks. (n.d.). https://www.databricks.com/glossary/machine-learning-models
- *What is Random Forest?*. IBM. (n.d.). https://www.ibm.com/topics/random-forest#:~:text=Random%20forest%20is%20a%20commonly,both%20classification%20and%20regression%20problems