**Machine Learning Project to predict Bankruptcy.**

**Team Members:**

**Mrinal Rajkumar Saraswat (4357797)**

**Sagaljyot Singh (4364571)**

**Simran (4364562)**

**School of Business, St. Lawrence College, Kingston**

**ADMN5006**

**Professor Sujoy Paul**

**December 15, 2023**

Table of Contents

[Section 1: Introduction and Business Problem 3](#_Toc153140164)

[Section 2: Methodology and Results: 4](#_Toc153140165)

[Section 3: Monetary value and Risks: 9](#_Toc153140166)

[Section 4: Other risks and benefits: 10](#_Toc153140167)

[Section 5: References: 11](#_Toc153140168)

# Section 1: Introduction and Business Problem

In this project, the primary objective is to develop predictive models for identifying the likelihood of bankruptcy based on a set of independent variables. This machine learning model addresses a critical issue in financial risk assessment and management. By utilizing historical financial data and possibly other relevant features, the model can provide insights into the financial health of a company, helping stakeholders make informed decisions about investments, loans, and business partnerships.

The market size for a machine learning model predicting bankruptcy is significant. It can cater to a wide range of industries, including banking, finance, investment, and business consulting. Small and large businesses alike can benefit from such a tool to assess the financial stability of their counterparts, potential clients, or investment opportunities. The market size is substantial, given the global nature of financial transactions and the pervasive need for risk assessment.

**Potential Cost Savings:**

Risk Mitigation: The model can assist financial institutions and investors in avoiding investments in companies at high risk of bankruptcy, thus preventing potential financial losses.

Efficiency Gains: Businesses can save resources by streamlining due diligence processes. The model can quickly analyze vast amounts of financial data, providing rapid insights compared to traditional methods.

**Risks:**

Model Accuracy: The primary risk is the accuracy of the machine learning model. If the model has a high false positive or false negative rate, it could lead to incorrect decisions, potentially causing financial losses.

Data Quality: The reliability of predictions heavily depends on the quality and relevance of the input data. Inaccurate or incomplete financial data may lead to unreliable predictions.

Dynamic Nature of Business: Economic conditions and business landscapes can change rapidly. The model may not capture sudden market shifts, unforeseen events, or changes in management that could impact a company's financial health.

Ethical Considerations: Depending on how the model is used, there may be ethical concerns, especially if decisions based on predictions have significant socio-economic implications, such as job losses or business closures.

# Section 2: Methodology and Results:

The data had Null values, and many extreme outliers.

A screenshot of a computer

Description automatically generated

A line of black dots

Description automatically generated with medium confidence

Also, there was an over-sampling of the target variable, and the input variables were at extremely different scales; so much that they were hardly visible in the plot:

A blue square with white squares

Description automatically generated

A graph with text and numbers

Description automatically generated with medium confidence

**Principal Component Analysis:** The correlation heatmap shows that there is very less correlation between the input variable, hence all of them should be taken for prediction.

A screenshot of a graph

Description automatically generated

**Data Processing**

Data processing involved several crucial steps to ensure the robustness of our models. Null values and extreme outliers were replaced with the mean of their respective columns to maintain data integrity. The decision to use mean imputation was driven by its simplicity and applicability to maintain the overall dataset structure.

To address potential class imbalance, I employed the Synthetic Minority Over-sampling Technique (SMOTE) for oversampling. This helped in mitigating the impact of minority class instances, ensuring a more balanced representation in the models. Additionally, MinMax scaling was applied to normalize the range of variables, enhancing the convergence of the models.

Following are the results after performing data transformation:

A screenshot of a computer

Description automatically generated

A graph with colorful lines and text

Description automatically generated

A graph with a blue bar

Description automatically generated

**Model Development**

1. Model Selection: Logistic Regression, Support Vector Machine, Decision Tree Classifier, and Random Forest Classifier were chosen for their versatility and suitability for binary classification problems.
2. Model Development Process: For each model, we split the dataset into training and testing sets.
3. Model Results: Random Forest Classifier demonstrated the highest accuracy at 98.73%. After performing hyperparameter tuning using grid search method on random forest classifier, following are the best parameter values:

**'max\_depth': 20, 'min\_samples\_leaf': 1, 'min\_samples\_split': 2, 'n\_estimators': 100**

**Model Comparison**

Model comparison involved assessing metrics such as accuracy, precision, recall, and F1 score. For Model performance comparison, Confusion matrix and accuracy were visualized:

A screenshot of a graph

Description automatically generated

A graph of different models

Description automatically generated

**Conclusion**

1. The experiment concludes with a recommendation to deploy the Random Forest Classifier in a practical setting for bankruptcy prediction due to its exceptional accuracy and balanced performance metrics. Financial institutions can leverage this model to identify companies at risk early on, facilitating proactive risk management.
2. Future steps: We can now use the best parameters and estimators from hyperparameter tuning, to further increase the accuracy of the random forest model. Considering time constraints and resource limitations, future steps could involve exploring ensemble methods, feature engineering, and leveraging advanced deep learning techniques. Implementing a more extensive feature selection process and experimenting with different sampling techniques may further enhance model performance. Additionally, continuous monitoring and periodic retraining of the model would ensure its relevance in a dynamic financial landscape.

# Section 3: Monetary value and Risks:

**Monetary Value:**

Cost Savings in Due Diligence: The application can significantly reduce the costs associated with extensive manual financial analysis and due diligence processes. Automated analysis by the machine learning model is typically faster and more efficient.

Risk Mitigation: By accurately predicting companies at risk of bankruptcy, the application can help investors, creditors, and financial institutions avoid substantial financial losses. This risk mitigation translates directly into monetary savings.

Improved Decision-Making: The application can enhance decision-making processes related to investments, loans, and business partnerships, leading to more informed and financially sound choices. This improved decision-making can contribute to increased profitability and financial stability.

Operational Efficiency: Businesses and financial institutions can achieve operational efficiency by automating parts of their risk assessment processes, leading to potential cost savings in labor and time.

**Risks:**

False Positives/Negatives: The model's accuracy is crucial. False positives (predicting bankruptcy when the company is not at risk) or false negatives (missing companies that are at risk) can lead to suboptimal decisions, potentially resulting in financial losses.

Data Quality and Reliability: The reliability of predictions depends on the quality, completeness, and relevance of the input data. Poor data quality can undermine the accuracy of the model.

Dynamic Business Environment: Economic conditions, industry trends, and business landscapes can change rapidly. The model may not capture sudden shifts, leading to decisions based on outdated information.

Model Interpretability: The complexity of some machine learning models may hinder their interpretability. Understanding how the model arrives at its predictions is crucial for gaining trust and making informed decisions.

Ethical and Regulatory Risks: There may be ethical considerations surrounding the use of AI in financial decisions. Regulatory compliance and adherence to ethical standards are essential to avoid legal and reputational risks.

# Section 4: Other risks and benefits:

**Other Risks:**

Overfitting: Overfitting occurs when a machine learning model learns the training data too well, capturing noise and anomalies that do not generalize to new data. This can result in poor performance on unseen data, impacting the model's reliability.

Lack of Explainability: Some advanced machine learning models, such as deep neural networks, may lack interpretability. The inability to explain why a particular prediction was made could be a challenge, especially in industries where transparency is essential.

Bias and Fairness: If the training data used to develop the model contains biases, the model may perpetuate or amplify these biases, leading to unfair or discriminatory outcomes. Addressing bias in the model is crucial for ethical considerations and regulatory compliance.

Cybersecurity Risks: As with any technology, there is a risk of cyber attacks. If the machine learning model is not adequately secured, it may be vulnerable to adversarial attacks, data breaches, or manipulation, which can compromise the integrity of predictions.

Model Degradation: Over time, the model's performance may degrade if not regularly updated with new data. Changes in economic conditions, regulations, or business practices can affect the model's relevance and accuracy.

**Other Benefits:**

Time Efficiency: Automation of financial analysis through machine learning can significantly reduce the time required for tasks such as data processing, trend analysis, and risk assessment, enabling faster decision-making.

Scalability: The model can scale to analyze a large volume of financial data, making it suitable for businesses of different sizes. This scalability is particularly advantageous for financial institutions dealing with a diverse portfolio.

Continuous Improvement: With proper feedback mechanisms and continuous learning, the machine learning model can adapt to changes in the financial landscape, improving its predictive capabilities over time.

Customization: The model can be tailored to specific industries or business needs, allowing for customization based on the unique characteristics of different sectors.

Comprehensive Analysis: Machine learning models can consider a wide range of variables simultaneously, providing a more comprehensive analysis of a company's financial health compared to traditional methods that may rely on a limited set of indicators.

# Section 5: References:

[1] *Search | Kaggle*. (n.d.-b). <https://www.kaggle.com/search?q=bankruptcy+in%3Adatasets>

[2] Brownlee, J. (2021, March 16). *SMOTE for Imbalanced Classification with Python*. MachineLearningMastery.com. <https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/>

[3] *sklearn.ensemble.RandomForestClassifier*. (n.d.). Scikit-learn. <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>

[4] *1.10. Decision Trees*. (n.d.). Scikit-learn. <https://scikit-learn.org/stable/modules/tree.html>

[5] Es, S. (2023, August 30). *Hyperparameter Tuning in Python: a Complete Guide*. neptune.ai. <https://neptune.ai/blog/hyperparameter-tuning-in-python-complete-guide>

[6] *12 Python Data Visualization Libraries to explore for Business Analysis | Mode*. (n.d.). <https://mode.com/blog/python-data-visualization-libraries>

[7] *Stack Overflow - where developers learn, share, & build careers*. (n.d.). Stack Overflow. <https://stackoverflow.com/>