**ML Assignment 1**

**Section 1: Introduction and Business Problem**

In this experiment, the primary objective is to develop predictive models for identifying the likelihood of bankruptcy based on a set of independent variables. The business problem revolves around the financial sector's need to proactively detect companies at risk of bankruptcy, allowing for timely intervention and risk mitigation. By employing machine learning techniques, specifically logistic regression, support vector machine, decision tree classifier, and random forest classifier, we aim to provide accurate predictions to aid financial institutions in making informed decisions.

**Section 2: Exploratory Data Analysis (EDA)**

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

**Section 3: 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

**Section 4: 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**

**Section 5: 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

**Section 6: 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.