**Financial Risk Management Prediction using ML Algorithms**

1. **INTRODUCTION**
   1. **Overview**

In the era of stringent and dynamic business environment, it is crucial for organizations to foresee their clients’ delinquency behavior. Such environment and behavior create unreliable base for strategic planning and risk management. Business Analytics combines the business expertise and computer intelligence to assist the decision makers by predicting an individual's credit status. This empirical research aims to evaluate the performance of different Machine Learning algorithms for credit risk prediction with more focus on Random Forest Trees. Several experiments inspired by observation and literature illustrate the potentials of computer-based model in classifying a number of bank history records. However, enhanced classification outcomes require tuning the randomness and tree growing parameters of the Random Forests algorithm. The model based on Random Forest Trees over performed most of the other models. Moreover, such a model has various advantages to business experts as the ability to help in understanding the relations between the analyzed attributes.

* 1. **Purpose**

In the period of unstable business environment, it is critical for organizations to predict the behaviors of their clients. One of the main concerns of funding organizations or banks is their clients’ adherence to payback the debts as expected. For that it is important to assess the clients’ credit suitability before authorizing a loan for example. According to the study in it is apparent how United States and Europe loans and mortgages tend to raise over years, though a credible risk assessment of credit needs to take place. The variables affecting the risk factor vary and the effects associated with overdue or unpaid debts by consumers may have unwanted consequences that may exceed the organizational level . Business Analytics (BA) is a convenient approach that utilizes Business Intelligence (BI) techniques to fulfill business needs, as predicting behaviors and outcomes. Therefore BA benefits business needs using the capabilities of Information Technology (IT) including Data Mining. The overall idea behind BA is to integrate the potentials of IT domain expertise with business domain expertise to reach an effective collaboration. Machine learning (ML) has been one of the IT domains contributing effectively to business prediction problems. An interesting proposed machine learning approach in [4] seeks predicting customers churn in telecommunication industry using Genetic Programming (GP). Churn management gains high importance in business domain as it is related directly to customer retention strategies. Though analyzing the behaviors of customers is critical and acts as an early alarming mechanism. Afterwards it triggers the business activities related to risk prevention and thus business continuity. However, some of these approaches do not give much insight on the dynamics and interaction between the underlying factors, and they are commonly considered as black box modeling techniques. For these reasons other category of machine learning approaches is investigated, this category contains approaches that produce more interpretable models that are easy to evaluate and explain more the effect of each factor in the prediction. For instance, some researchers suggest using Random Forest for being more descriptive

1. **LITERATURE SURVEY**
   1. **Existing Problem**

A number of Data Mining algorithms were investigated in for credit risk evaluation. DT, ANN, SVM, and Logistic Regression (LR) outcomes for risk prediction were compared over Australian Credit, German Credit and refined German Credit datasets. The authors used 10-fold cross validation strategy with different dataset splitting configurations. The overall findings based on prediction Accuracy, Precision, and error rate of Type I/II, show that SVM and LR models dominate in terms of performance. More precisely SVM based models are robust and have higher ability for better generalization with small training sets. Nonetheless, DT models as the case of C4.5 trees are easier to interpret due to their higher level of clarity and showed comparable results. The authors, referring to the empirical evidence, stress the idea that DT demonstrate better aid to the decision making of business domain experts. Clear and easy understanding of a prediction model reveals greater details for business domain experts who would be glad to comprehend. Moreover, they proposed possible future work to include a hybrid approach incorporating DT and SVM for credit risk assessment. A model based on Random Survival Forests was compared to Logit model for predicting credit risk for a number of German Small and Medium Enterprises (SMEs) in [5]. The dataset contains a set of variables that show whether the financial situation is good or not, relying on financial calculation called “Financial Ratio”. An error measurement encompassing the number of trees and variables importance was used to evaluate the performance of the used models. The empirical results show that Logit models gave better predictions, while a Random Survival Forests were more explanatory. The authors in [17] investigated the suitability of several AI techniques in predicting credit risk. Ten classification algorithms were used to predict the credit risk in a German dataset. The set of measures used to compare the performance of the different algorithms was “Type I, Type II, and total accuracy [19]”, while ranking the algorithms relied on the area under Receiver Operating Characteristic (ROC) curve. The overall results nominate SVM based models to be more accurate in prediction. However, a hybrid approach was used that includes clustering before the classification step. Such approach would be relatively complex and applying it in business world could not be understood easily. Even though there are many complex approaches to predict credit risk, none of their complexities is justified by a significant accuracy level. For that the literature lacks a significant improvement in terms of performance to accuracy ratio. That would raise the question of whether simple approaches would suit more the credit risk prediction issue, especially of being relatively more descriptive

* 1. **Proposed System**

The aim of this thesis is to diagnose the financial health of businesses using machine learning algorithms. A detailed study of using several data-driven models to forecast corporate bankruptcy (firm default) was conducted. Both qualitative assessments from financial experts and quantitative econometric factors will be considered for training models for predicting financial credit risk.

1. **THEORITICAL ANALYSIS**
   1. **Block Diagram**

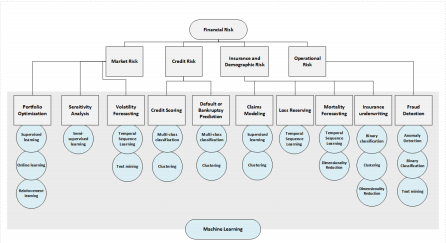
****

Figure-1 Block Diagram of the Proposed System

* 1. **Hardware/Software Designing**
* Google Co-Laboratory
* VS(Visual Studio) Code
* Python 3.10.4

1. **EXPERIMENTAL INVESTIGATIONS**
   1. **Machine Learning Methods**

*Supervised learning* algorithms can be implemented for portfolio optimization problems. Support vector machines can appropriately identify the non-linear relationships among the market variables. Performance of such supervised learning models were further improved by using novel regularization or cross-validation techniques [113]. Neural Networks can also outperform traditional models for portfolio selection . *Online Learning* is also used for this task. An online portfolio selection strategy with the assumption of mean reversion relation of financial markets was implemented to fifind optimal portfolio weights . A comprehensive literature on using different online learning approaches (e.g. Follow-the-Winner, Follow-the-Loser, Meta-Learning) for portfolio selection can be found . *Reinforcement Learning* methods are gaining traction in research related to portfolio optimization. LR can be implemented to find optimal policies for the reinforcement learning task,. Recurrent reinforcement learning method (RRL) has also been applied to simultaneously generate market activity signals (buy or sell) and the optimal asset allocation weights based on a downside risk-adjusted objective function .

* 1. **Managing Credit Risk**

Credit risk is the uncertainty involving borrower’s capability to fulfill obligations. This may involve individual borrowers defaulting on a loan or corporations going bankrupt. The two most active research topics for credit risk management are bankruptcy/default prediction and credit scoring . Credit scoring generally refers to the risk classification of retail borrowers (which includes personal loans or mortgages) whereas bankruptcy prediction generally refers to the prediction of bankruptcy of an institutional borrower (for example, a small business). Generally, from a statistical modelling point of view, both the bankruptcy prediction task and credit scoring task can be regarded as binary classification problems . However, the predictors used for these modelling tasks are generally different. The predictors used for bankruptcy predictions are key financial ratios derived from the company’s financial statements such as balance sheets or income statements . In contrast, predictors used for retail credit scoring models are various financial and demographic information of the loan applicant (for example, credit history, account balance, employment status, age).

* 1. **Managing Insurance and Demographic Risk**

Financial institutions that offer insurance services for various types of risk bears a significant amount of financial risks themselves. It is critical for these companies to accurately quantify the exposure and take proper steps to mitigate these risks. Financial risk for insurance can arise from many sources, each requires accurate prediction modelling. For example, before providing a car insurance product to an individual, the insurer needs to accurately predict the number of claims the driver might make in the future. Error in the prediction would result in under-pricing of the insurance product, which would result in a financial loss for the insurer in the future. Also, life-insurers need to have an accurate estimate of the expected lifespan of demography before providing life-insurance products to the individuals belonging to the demography.

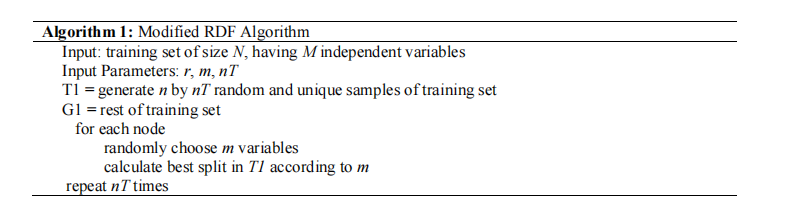
* 1. **Algorithms**

**4.4.1 Random Forest**

Random Forest Trees are based on a number of prediction trees that are less tolerant to noise compared to “Adaboost” and utilize random selection of features in splitting the trees. “Random Forests” is a voting procedure for the most popular class among a large number of trees. Thus, a random forest classifier is composed of a set of tree-structured classifiers [20- 22]. Equation [1] represents the classifiers where Θi represents a number of independent random vectors distributed identically, such that every tree has a vote for most popular class of input *X*.

*Space = { h( X, Θi ); i= 1,2,3,…. nT }*

(1) Using Random Forests for prediction has many advantages such as their immunity to over fitting, an appropriate selection of randomness type leads to accurate classification or regression, the correlation and strength of predictors makes a good estimate of the ability for prediction, faster than boosting and bagging, better estimation of internal errors, not complicated, and can perform well in parallel processing. Based on the empirical results in [20, 21], Random Forests could compete with similar approaches in terms of accuracy. Moreover, the author proved that Random Forest could give better results with “Boosting” and “Bagging” [20]. Accordingly, opening the door for an investigation area of the injected randomness effects on prediction accuracy. The pseudo code of a modified version of Random Decision Forests (RDF) used in Heursticlab1 environment is shown next2. 1 Heuristic Lab is a framework for heuristic and evolutionary algorithms that is developed by members of the Heuristic and Evolutionary Algorithms Laboratory (HEAL).



**4.4.2 Linear Regression**

Linear regression is a very simple approach for supervised learning. In particular, linear regression is a useful tool for predicting a quantitative response. Linear regression has been around for a long time and is the topic of innumerable textbooks. Though it may seem somewhat dull compared to some of the more modern statistical learning approaches described in later tutorials, linear regression is still a useful and widely used statistical learning method. Moreover, it serves as a good jumping-off point for newer approaches: as we will see in later tutorials, many fancy statistical learning approaches can be seen as generalizations or extensions of linear regression. Consequently, the importance of having a good understanding of linear regression before studying more complex learning methods cannot be overstated.

1. **RESULTS AND DISCUSSION, PERFORMANCE ANALYSIS**

The German Credit dataset used in this study is publicly available at the University of California, Irvine (UCI) Machine Learning Repository. In which there are 1000 instances divided into two classes; 700 “good credit” and 300 “bad/refused credit request”. The original credit dataset contains 20 variables that fall into 13 categorical and 7 numerical ones listed in Table . However, this research work is conducted entirely on a processed copy (by the Strathclyde University) that is also available at UCI repository. The processed dataset is a conversion of the originals into 25 numerical variables, in which number 25 is an output variable (Good/Bad). Figure illustrates the distribution of the numerical variables

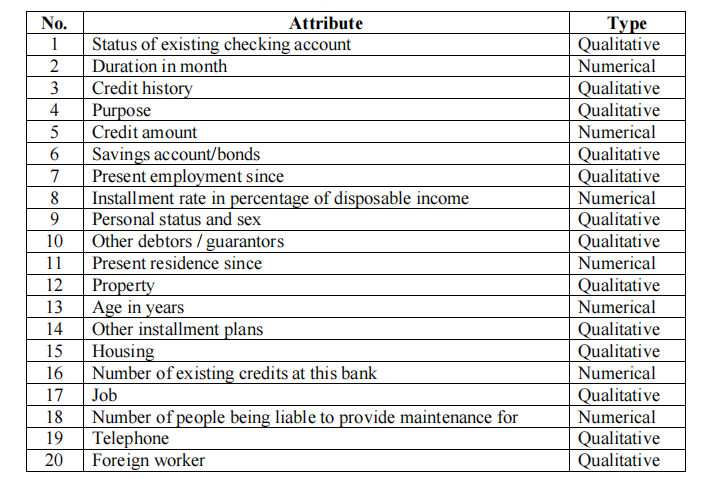
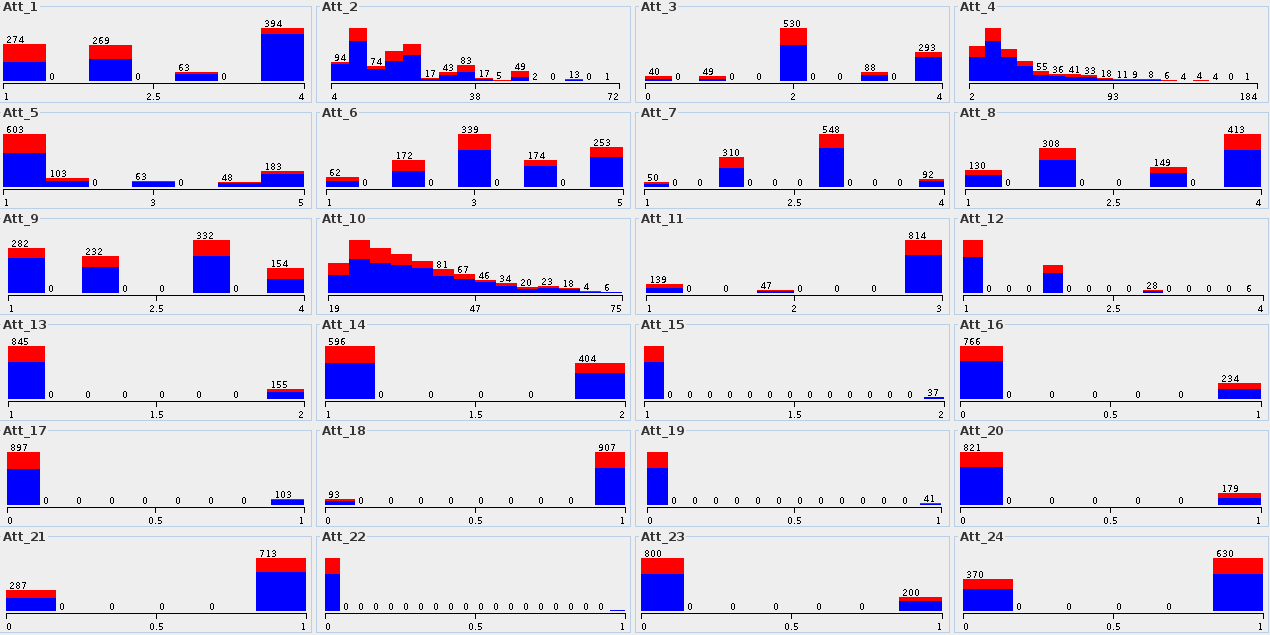
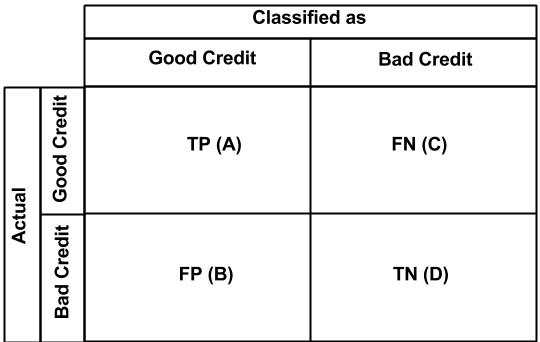


Figure-Original Attributes



Distribution of the Numerical Variables in the Processed Dataset



Confusion Matrix

This confusion matrix illustrates the classification’s “confusion” or what is called

classification error, in which the rows represent the actual classes and the columns represent the predicted classes. Total number of correctly classified instances will be represented diagonally in True Positive (TP) and True Negative (TN) cells. TP represents “Actual Good” classified as “Good” while TN represents “Actual Bad” classified as “Bad”. The higher TP and TN the better is the performance of the classification algorithm. Incorrectly classified instances go to False Positive (FP) and False Negative (FN) that are “Actual Good” classified as “Bad” and “Actual Bad” classified as “Good” respectively. Total sum of the correctly and incorrectly classified classes should match the total number of the input instances. Several mathematical measures based on the confusion matrix make it easier to assess deeply the performance of the classification algorithm, also make it easier to compare the performance of different algorithms. Due to the variety of measures most of the researchers use and for better comparison, this study will report a number of measures that are:

•Total Accuracy (Correctly Classified Instances) = TP + TN / ( TP + TN + FP + FN )

•Sensitivity (Recall, Hit Rate, TP Rate, or Type II Error) = TP / ( TP + FN )

•Precision (Confidence or Type I Error) = TP / ( TP + FP )

•F-Measure = (2 \* Precision \* Sensitivity) / (Precision \* Sensitivity)

•Area Under Receiver Operating Characteristics Curve (AUC) [19, 24], for some of

the algorithms. It is calculated automatically in Weka environment. This study tries empirically to find out the potentials of Random Forest Trees in classification by tuning the used models. That includes relying on the literature and arbitrary selection of model parameters.

1. **SUMMARY AND CONCLUSIONS**

Due to the importance of understanding and managing the risks in volatile business domains, it is required to find an effective aid in making decisions. The results of this research show that Random Forest Trees algorithm is a promising opportunity for Business Analytics in predicting credit risk. The main advantages of using Random Forest Trees in prediction are the competitive classification accuracy and simplicity. Such simplicity makes it easier for decision makers to understand more the underlying relations, especially for the fact that none of the classification approaches achieved significant accuracy. The pluses make the results of decision trees more useful and appealing for business domain experts than other approaches. A noteworthy finding is the effect of injected randomness and how to grow the individual trees on producing better classification results. The empirical findings of this research and others open the door for deeper future work to

improve the performance of decision trees. Firstly, to improve the classification models by enhancing the way to grow the decision trees, and better variable selection. Secondly, hybrid approaches incorporating Random Forest Trees need thorough investigation and testing. The possibilities vary starting from the results of this research as the use of different datasets, the redesign of the datasets to include or exclude affecting variables, study the impact of each variable on the overall performance, and model different problems as the bankruptcy prediction.

**Data Availability**

The data utilized to support these research findings is accessible online at

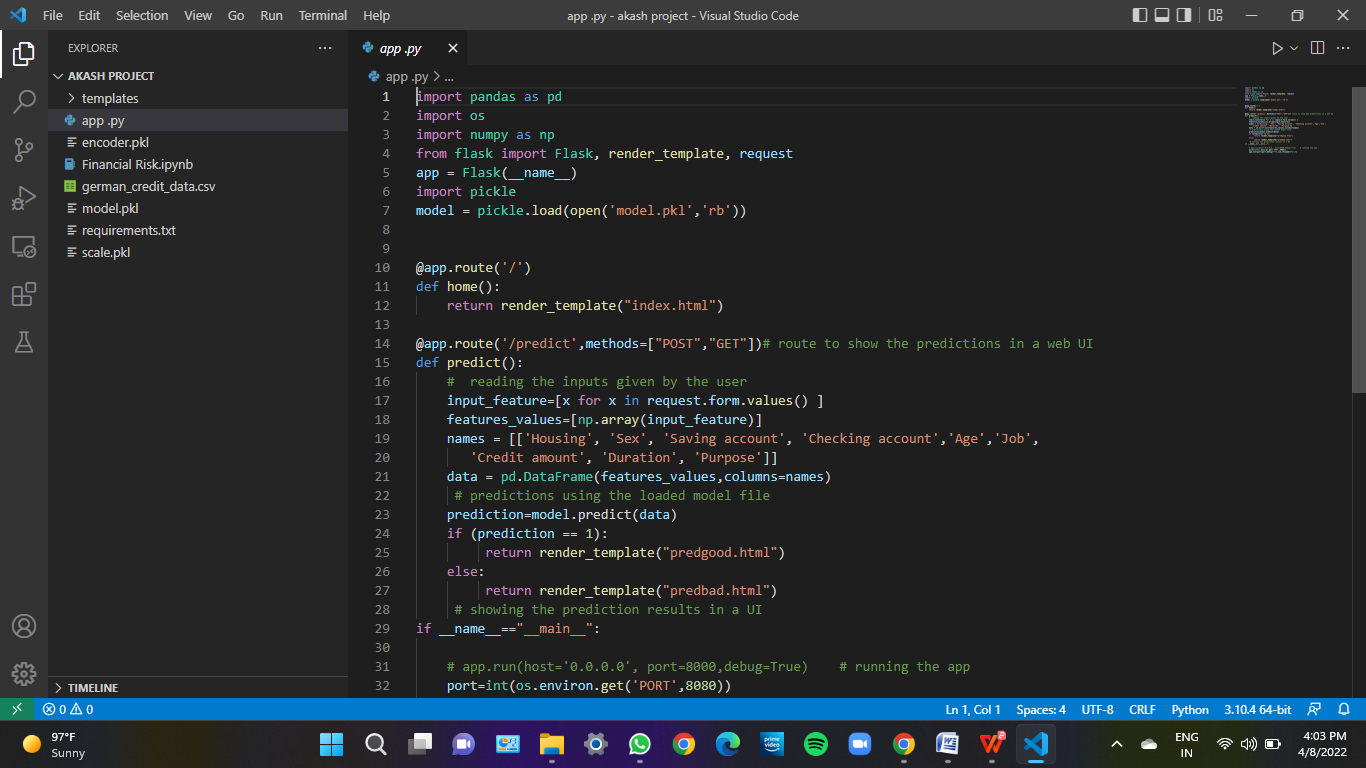
https://www.kaggle.com/datasets/uciml/german-credit.

**References**

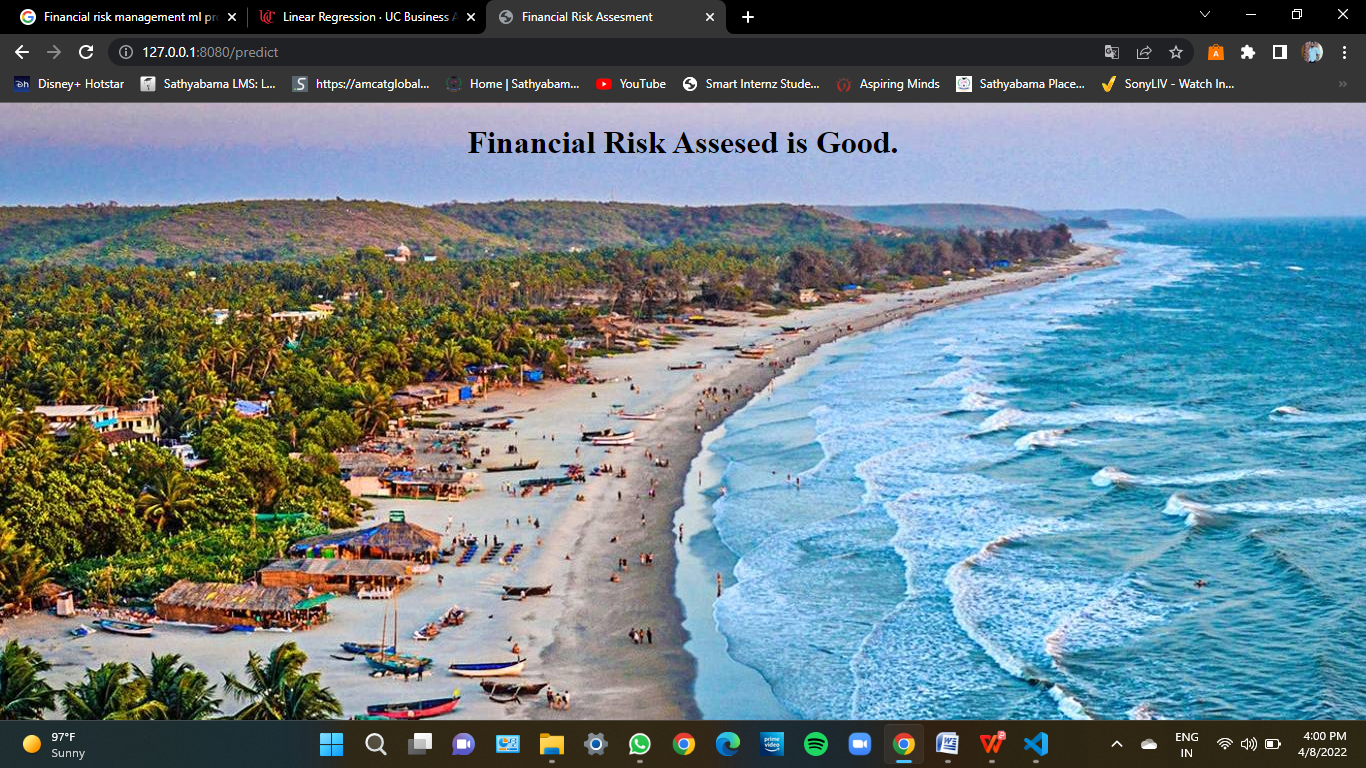
* All Information is collected from the Internet sources like Google…

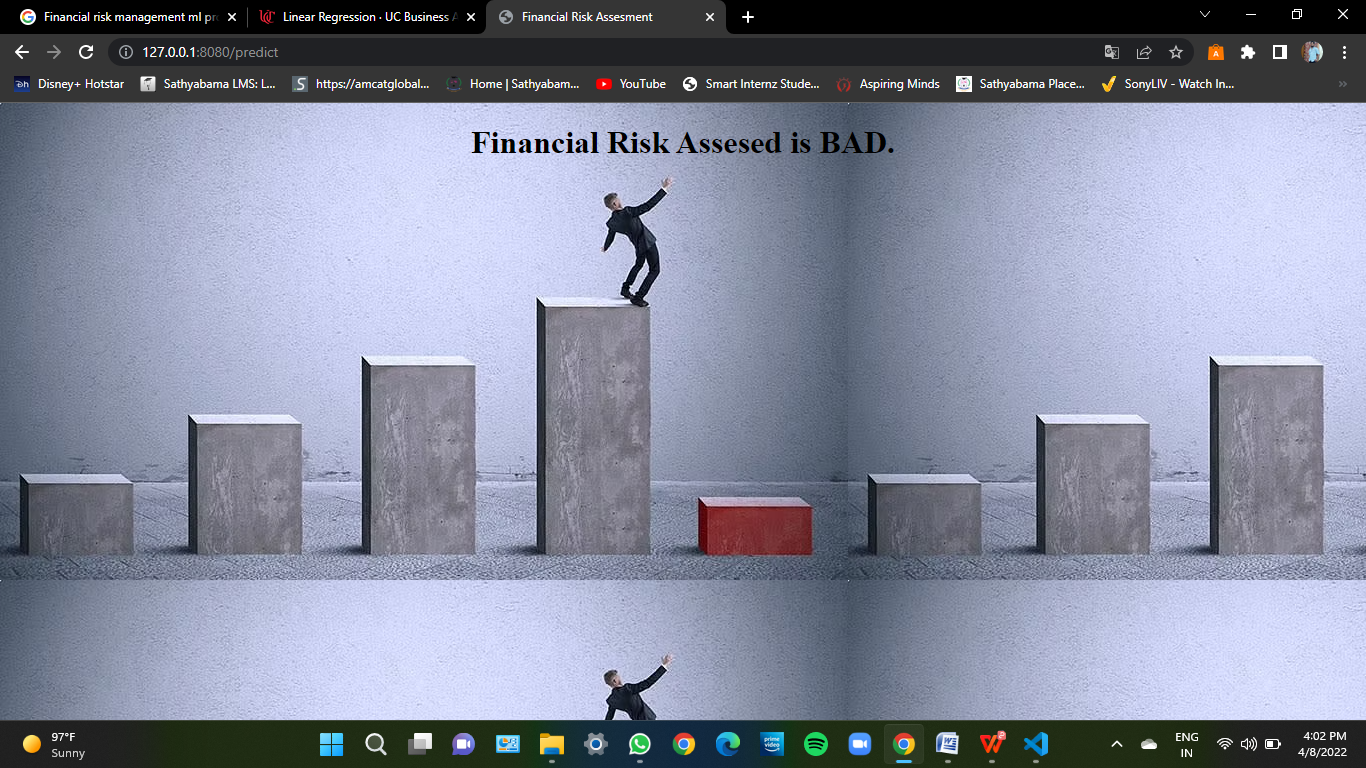
**Appendix**

**Main Code & Output Screenshots**

****

****

****

****