**Financial Health Score Prediction using MLR**

**By: Pragya Gupta**

**Date: 15th July, 2025**

**Course: PGDM – RBA 24-26**

**Specialization: Finance**

**Roll no.: 55**

**Exploratory Data Analysis Steps:**

**1. Loading necessary libraries & dataset**

* Import essential libraries
* Load the dataset into a data frame

**2. Understanding the data:**

* Understand what is being conveyed; create a dictionary for each attribute
* Understand the shape of the data frame & data types of each variable
* Create a descriptive statistical summary of each numerical column to understand the mean/median values, max/min values, range of the data, etc.
* Identify unique values in categorical columns
* Understand the distribution of data, skewness, and kurtosis – Univariate analysis (histogram)

**3. Data Cleaning:**

* Remove the insignificant columns (e.g., IDs)
* Identify the null and duplicate values, and handle them (drop duplicates)
* Identify outliers in continuous columns – Univariate analysis (box plot)
* Treat the identified outlier values as best suited

**4. Checking Linearity:**

* Plot scatter plots of each variable with the dependent variable to understand if there is a linear relationship between them – Bivariate analysis

**5. Detecting Multicollinearity:**

* Calculate the correlation of each independent variable with the dependent variable – Bivariate analysis
* Identify multicollinearity amongst the independent variables by calculating a correlation matrix – Multivariate analysis
* Draw inferences from the matrix, and decide whether the collinearity is negligible or not – make a note of the same

**6. Label Encoding:**

* Convert **non-numeric categorical variables** into numeric form for analysis by using an appropriate encoding method

**Regression Steps:**

1. Perform Multiple Linear Regression, build a baseline model, and identify the significant variables (p<alpha)

2. Calculate the VIF additionally and based on both (VIF and p value), perform feature engineering (selecting only significant variables)

3. Again, perform Regression analysis on the data and keep on eliminating insignificant variables – Stepwise Regression

4. Once all the independent variables come out to be significant, check the overall significance of the model from the ANOVA table (F-statistic & p-value), and if the whole equation comes out to be significant, apply the model to the data

6. Perform Model Diagnostics: calculate residuals for each observation and plot all errors on a scatter plot. If the errors are distributed uniformly, the model exhibits Homoscedasticity

* Repeat the above steps for each industry by looping through the industry code from 0 to 6, and build one regression model per industry
* Identify variables that are consistently significant across industries