Capstone Project

Work Progress Status Report – I

Program: PGP - Data Science and Engineering Online Nov20

Project Team:

Name	Enrolment No
Apoorva Garg	
Pranavi Krishnamsetty	
Vishal Pandey	
Yash Vahi	

Mentor: Animesh Tiwari

Project Title: Bankruptcy Prediction

Target for 1st Month: EDA : Data understanding and preparation

Data Understanding:

1) We learned about the shape (6819, 96) and datatypes of our dataset. All features are of numeric data type(int or float)

2) A count plot on the target feature showed us that bankrupt firms were highly underrepresented in the dataset, pointing to the fact that there might be a need for upsampling.

3) We performed distribution analysis using distribution plots and checked skewness and kurtosis for every feature. We found that most of the features

were highly skewed and highly leptokurtic. Those features might contain heavy outliers and need to be treated.

4) We performed outlier analysis using box plots and found extreme outliers in most of the features.

5) Our null and duplicate value analysis showed us that there were no null or duplicate features in the dataset.

Feature Engineering and Selection:

1) We eliminated redundant features using domain knowledge.

For Example: “ROA(C) before interest and depreciation before interest” and “ROA(A) before interest and % after tax” are the features which contains the information that can be covered in “ROA(B) before interest and depreciation after tax”. So we removed first two.

2) Apart from this we also removed two redundant flag columns which contain 99% of either 1’s or 99% of 0’s.

3) We then removed outliers using IQR method.

4) For removal of rows made redundant due to null values resulting from outlier removal, we used ‘df.dropna’ function using different thresholds and found 67 as optimal threshold point, where our data maintains its orignal central value. Applying this technique we successfully removed 800 rows.

Target for next Month:

1) Imputation of remaining null values which were created due to removal of outliers

2) Eliminating features exhibiting high multi-collinearity using VIF

3) Feature selection, if needed

4) Modeling