Capstone Project Submission

|  |
| --- |
| **Team Member’s Name, Email and Contribution:** |
| |  |  |  | | --- | --- | --- | | **NAMES** | **E-MAIL** | **CONTRIBUTION** | | Sunil Kumar | 14bbt1019@gmail.com | Entire Project work. | |
| **Please paste the GitHub Repo link.** |
| **GitHub link:-**  https://github.com/Sunilkumar17-design/Regression\_Capstone.git |
| **Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)** |

**Summary of work**

**Started with data loading and importing the libraries and then started with exploring the data and looking into columns and rows. It was seen that there were no missing values.**

**I have made little observation while working on the project that the number of organizations that have gone bankrupt in 10 years between 1999 – 2000 is few. Several companies possess many assets, which is always a good sign for an organization. An organization cannot guarantee not being bankrupt, although owning several assets. The organizations in the dataset are running into losses for the past two years as their net income poses be negative An increase in the values of the attributes that have a negative correlation with the target attribute helps an organization avoid bankruptcy. There seems to be a relation between attributes that have a high correlation with the target attribute and a low correlation with the target attribute. I observed several correlations among the top 12 attributes, one of which being “Net Worth/Assets and Debt Ratio %” that is negatively correlated with one another.**

**While looking into Bankrupt? Column It was seen that column was following an almost normal distribution. While exploring Bankrupt? Column further it was evident that data is imbalanced and there is a huge difference between bankrupt and non-bankrupt companies.**

**I have used 2 techniques to overcome this problem. 1st was to create a data frame and divide it into equal rows that were equal (220,220) rows and 96 columns.**

**Once I was done here I then applied modeling techniques started with Logistic regression and the other ensembles.**

**I achieved consistently better quality with a better model every time, I also have used a voting classifier which basically takes an ensemble of numerous models and gives the best-predicted output/accuracy. Voting classifiers have given me the best quality. 2nd I have used Synthetic Minority Oversampling Technique (SMOT) in which I oversampled the minority class and balanced the data. I also performed the normalization on non-fractional columns to make sure data followed the same scale. I have seen and found that after SMOT results for accuracy were better than the 1st technique.**

**Finally, I have used calibration, it basically assures the reliable benchmark and accuracy results when features are very important, by this way I have achieved the best accuracy for the model prediction.**