Capstone Project

Company Bankruptcy Prediction

G. V. Kapeesh Varma

CONTENTS

PROBLEM STATEMENT

The Prediction of Bankruptcy in companies is a problem that has concerned entrepreneurs, researchers and even governments for years, since detecting early signs that a company is going to enter bankruptcy involuntarily and being able to save it from that process, can help reduce the economic losses that bankruptcy entails, both in quantitative and qualitative terms. Since computers can store huge datasets pertaining to bankruptcy, making accurate predictions from them before hand is becoming important.

APPROACH

In this project, we analyze the data collected from the Taiwan Economic Journal for the years 1999 to 2009. Company bankruptcy has been defined based on the business regulations of the Taiwan Stock Exchange.

The data includes a vast amount of financial information such as Return of Assets, Tax Rates, Cash Flows, different ratios etc. Due to the large number of dimensions, Dimensionality Reduction has been performed through Principal Component Analysis (PCA). Various financial rates and ratios are then analyzed and different classification algorithms such as Logistic Regression, SVC, RandomForestClassifier etc. have been used to predict bankruptcies with very high accuracy.

DATA SUMMARY

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6819 entries, 0 to 6818
Data columns (total 96 columns):
 #   Column                                                  Non-Null Count  Dtype
---  ------                                                  --------------  -----
 0   Bankrupt?                                               6819 non-null   int64
 1    ROA(C) before interest and depreciation before interest 6819 non-null   float64
 2    ROA(A) before interest and % after tax                  6819 non-null   float64
 3    ROA(B) before interest and depreciation after tax       6819 non-null   float64
 4    Operating Gross Margin                                  6819 non-null   float64
 5    Realized Sales Gross Margin                             6819 non-null   float64
 6    Operating Profit Rate                                   6819 non-null   float64
 7    Pre-tax net Interest Rate                               6819 non-null   float64
 8    After-tax net Interest Rate                             6819 non-null   float64
 9    Non-industry income and expenditure/revenue             6819 non-null   float64
 10   Continuous interest rate (after tax)                    6819 non-null   float64
 11   Operating Expense Rate                                  6819 non-null   float64
 12   Research and development expense rate                   6819 non-null   float64
 13   Cash flow rate                                          6819 non-null   float64
 14   Interest-bearing debt interest rate                     6819 non-null   float64
 15   Tax rate (A)                                            6819 non-null   float64
 16   Net Value Per Share (B)                                 6819 non-null   float64
 17   Net Value Per Share (A)                                 6819 non-null   float64
 18   Net Value Per Share (C)                                 6819 non-null   float64
 19   Persistent EPS in the Last Four Seasons                 6819 non-null   float64
 20   Cash Flow Per Share                                     6819 non-null   float64
```

	Bankrupt?	ROA(C) before interest and depreciation before interest	ROA(A) before interest and % after tax	ROA(B) before interest and depreciation after tax	Operating Gross Margin	Realized Sales Gross Margin	Operating Profit Rate	Pre-tax net Interest Rate	After-tax net Interest Rate	Non-industry income and expenditure/revenue	Continuous interest rate (after tax)
0	1	0.370594	0.424389	0.405750	0.601457	0.601457	0.998969	0.796887	0.808809	0.302646	0.780985
1	1	0.464291	0.538214	0.516730	0.610235	0.610235	0.998946	0.797380	0.809301	0.303556	0.781506
2	1	0.426071	0.499019	0.472295	0.601450	0.601364	0.998857	0.796403	0.808388	0.302035	0.780284
3	1	0.399844	0.451265	0.457733	0.583541	0.583541	0.998700	0.796967	0.808966	0.303350	0.781241
4	1	0.465022	0.538432	0.522298	0.598783	0.598783	0.998973	0.797366	0.809304	0.303475	0.781550

5 rows × 96 columns

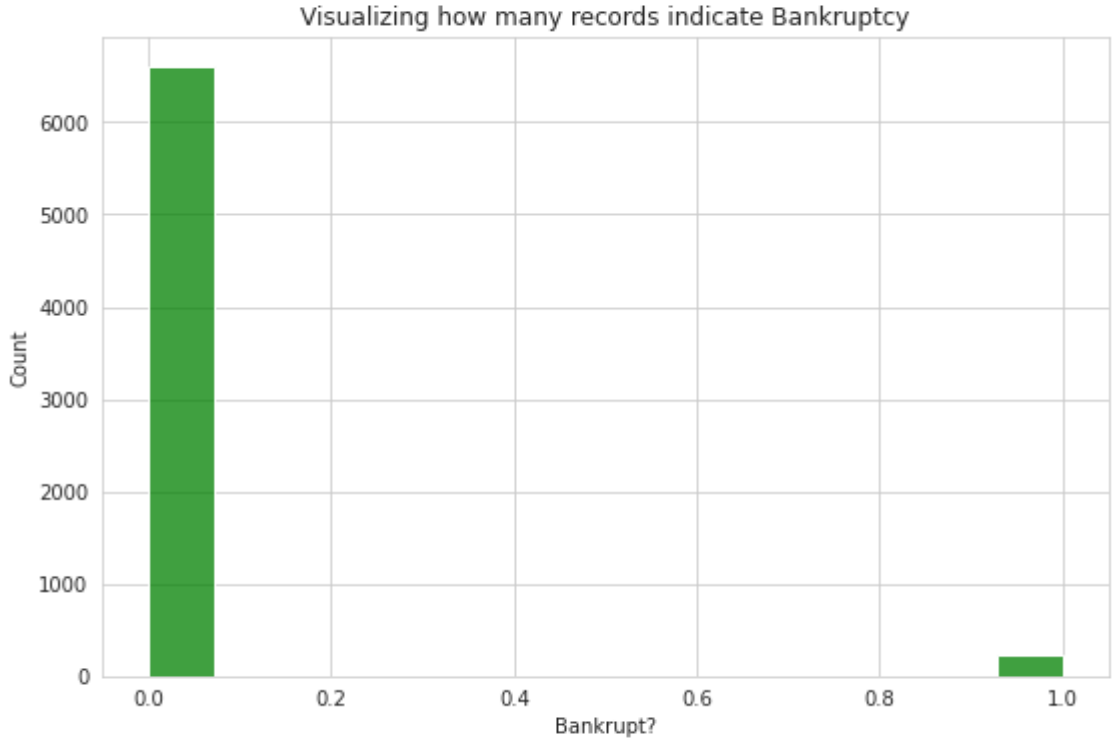
First 5 rows of the DataFrame

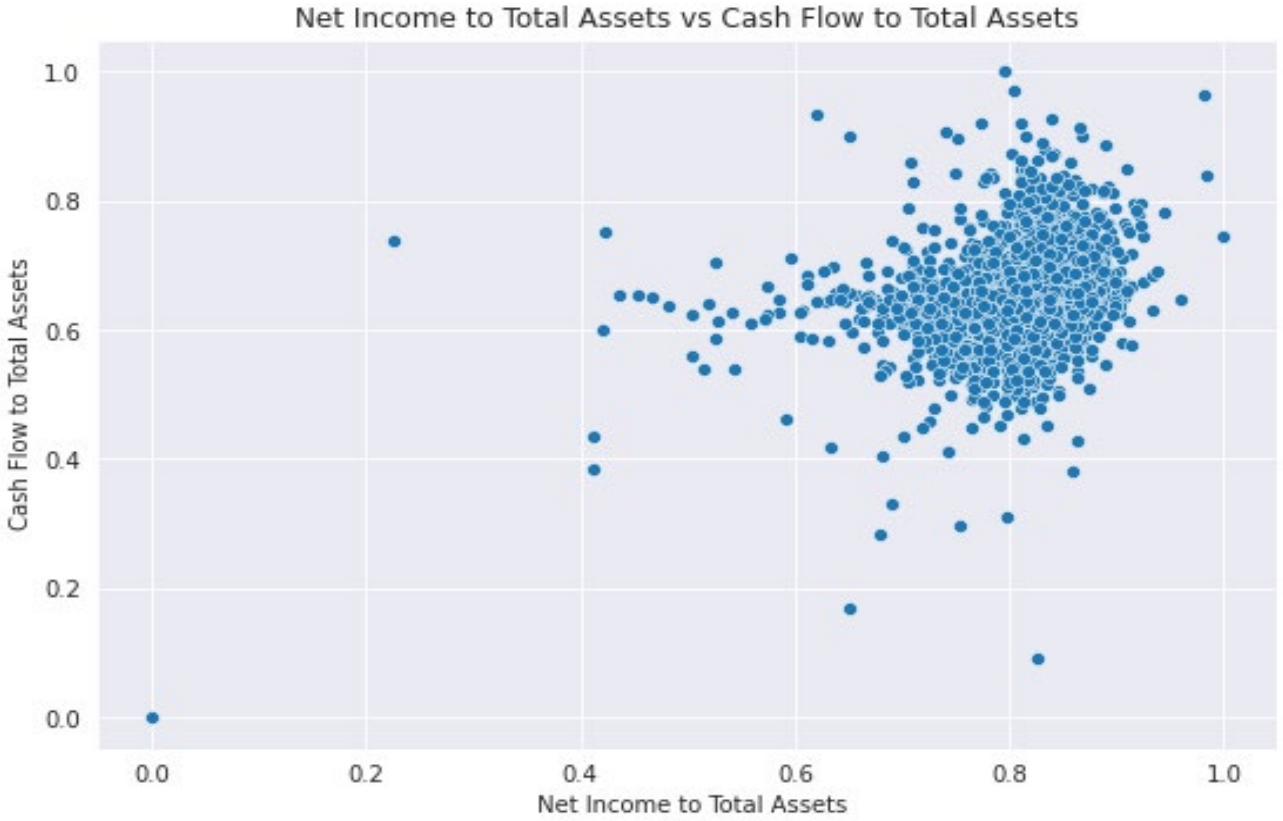
	Bankrupt?	ROA(C) before interest and depreciation before interest	ROA(A) before interest and % after tax	ROA(B) before interest and depreciation after tax	Operating Gross Margin	Realized Sales Gross Margin	Operating Profit Rate	Pre-tax net Interest Rate	After-tax net Interest Rate	Non-industry income and expenditure/revenue	Continuous interest rate (after tax)
count	6819.000000	6819.000000	6819.000000	6819.000000	6819.000000	6819.000000	6819.000000	6819.000000	6819.000000	6819.000000	6819.000000
mean	0.032263	0.505180	0.558625	0.553589	0.607948	0.607929	0.998755	0.797190	0.809084	0.303623	0.781381
std	0.176710	0.060686	0.065620	0.061595	0.016934	0.016916	0.013010	0.012869	0.013601	0.011163	0.012679
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.476527	0.535543	0.527277	0.600445	0.600434	0.998969	0.797386	0.809312	0.303466	0.781567
50%	0.000000	0.502706	0.559802	0.552278	0.605997	0.605976	0.999022	0.797464	0.809375	0.303525	0.781635
75%	0.000000	0.535563	0.589157	0.584105	0.613914	0.613842	0.999095	0.797579	0.809469	0.303585	0.781735
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

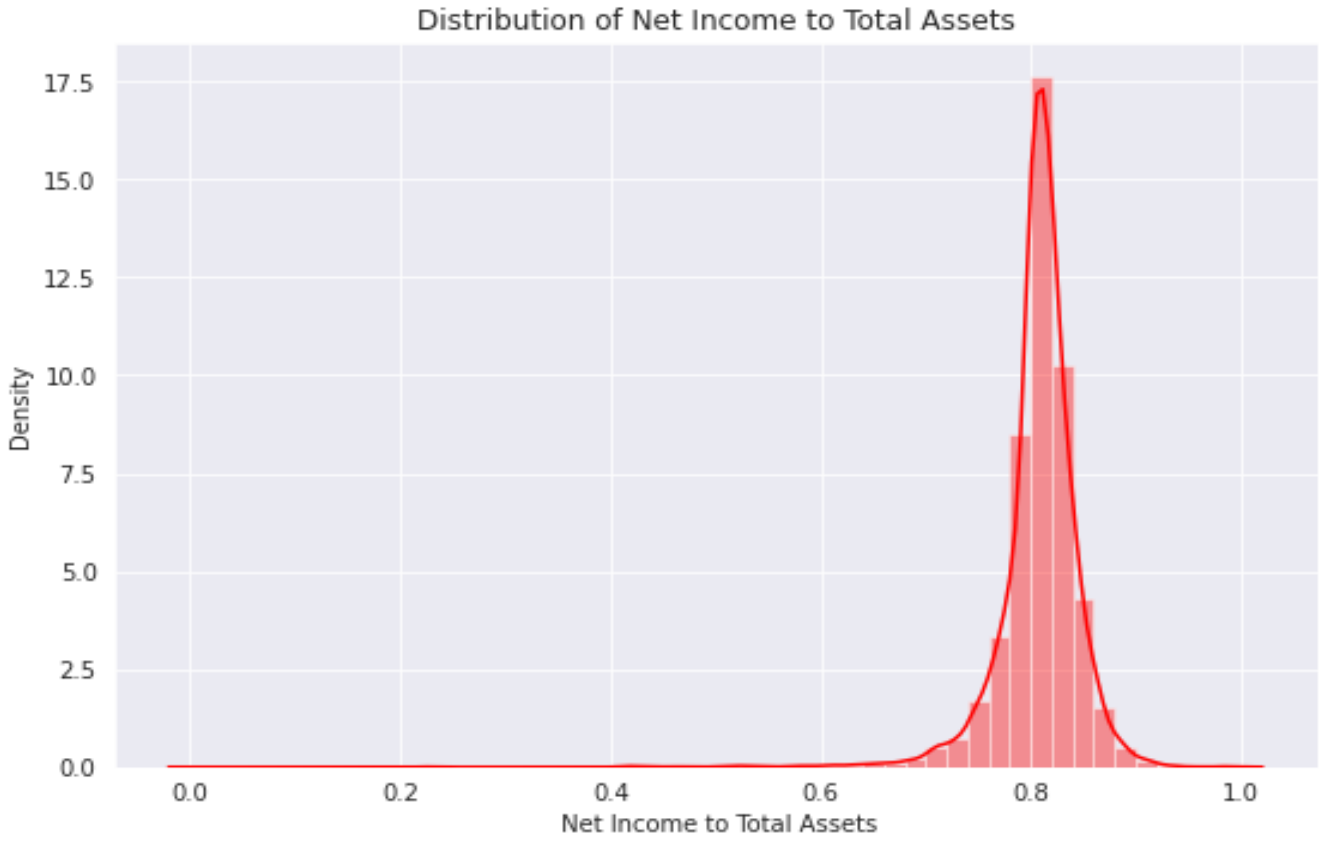
Descriptive Statistics

DATA VISUALIZATIONS

```
0    6599
1     220
Name: Bankrupt?, dtype: int64
```




Net Income to Total Assets vs Cash Flow to Total Assets
Cash Flow to Total Assets
Net Income to Total Assets
0.0
0.2
0.4
0.6
0.8
1.0


Distribution of Net Income to Total Assets
Density
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Net Income to Total Assets

DIMENSIONALITY REDUCTION
&
OVER SAMPLING

```
[ ]  # Initialize PCA with required number of components and fit_transform the data
     pca = PCA(n_components=25)
     reduced_data = pca.fit_transform(data.drop('Bankrupt?',axis=1))
     reduced_data.shape
```

```
(6819, 25)
```

```
[ ]  # Create a new DataFrame with the reduced data
     new_data = pd.concat([data['Bankrupt?'],pd.DataFrame(reduced_data)],axis=1)
```

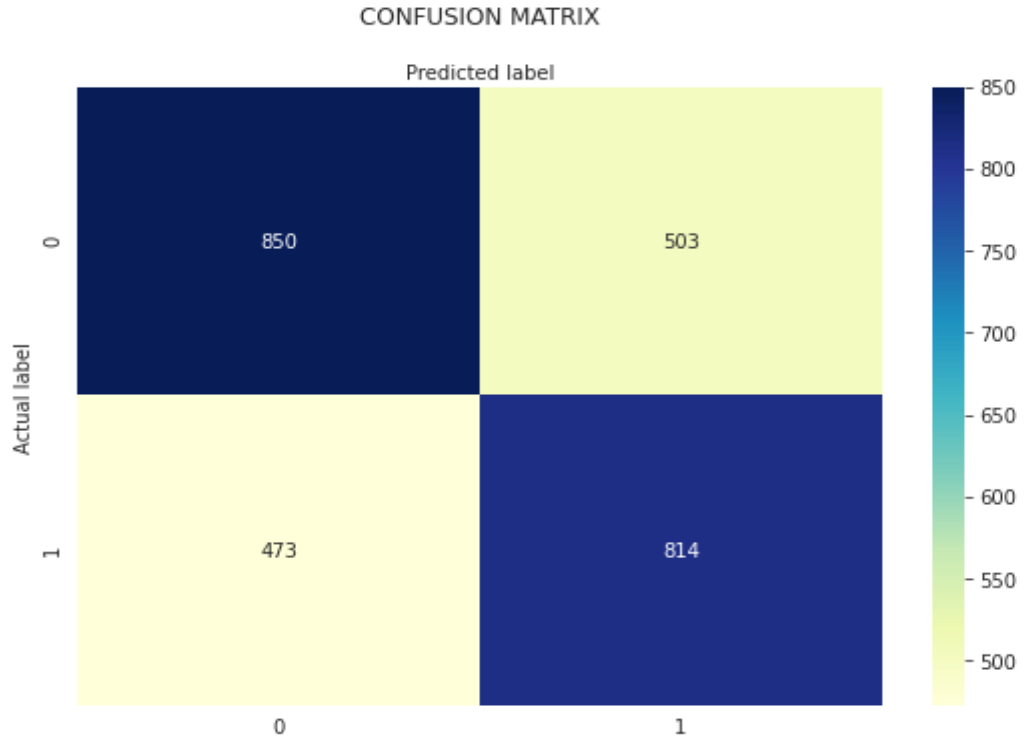
```
[18] # Pick relevant Independent & Dependent Variables
     x = new_data.drop('Bankrupt?',axis=1)
     y = new_data['Bankrupt?']
```

```
[21] # Perform Over Sampling of Data using Synthetic Minority Oversampling Technique(SMOTE)
     oversample = SMOTE()
     x_over, y_over = oversample.fit_resample(x,y)

     print(np.unique(y_over,return_counts=True))
```

```
     (array([0, 1]), array([6599, 6599]))
```

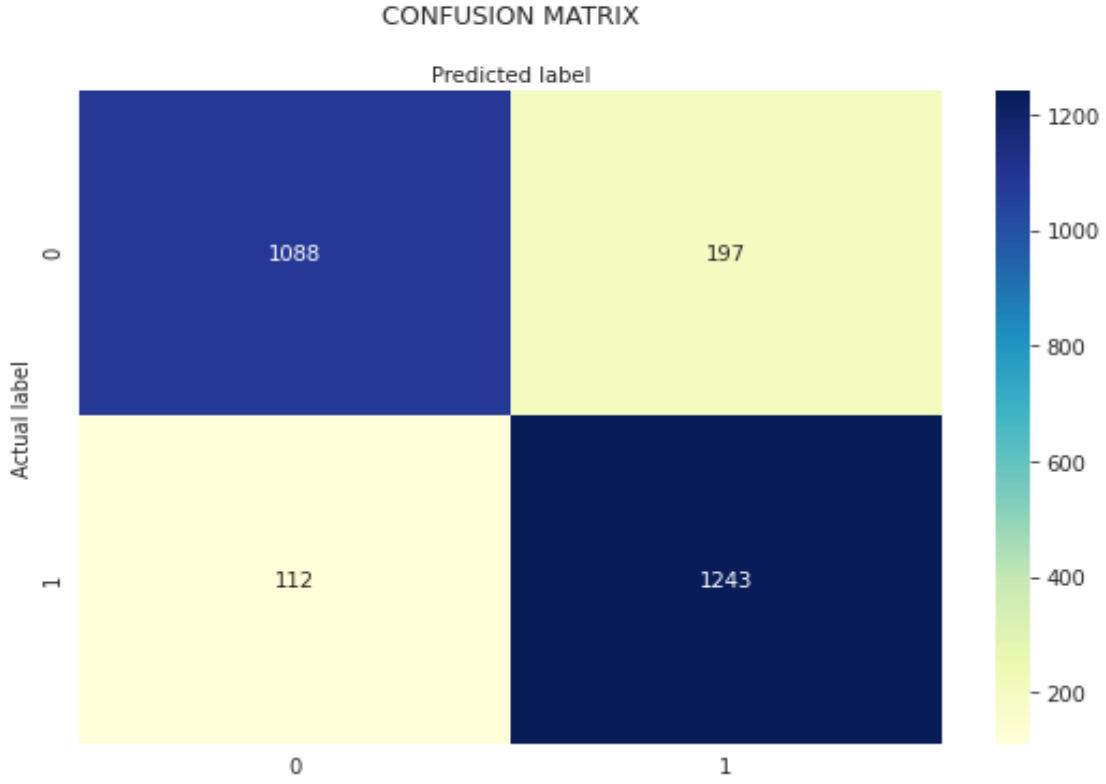
CLASSIFICATION MODELLING

```
** PERFORMANCE METRICS OF LOGISTIC REGRESSION **

Best Parameters for Logistic Regression:  {'C': 0.5, 'penalty': 'l2', 'solver': 'lbfgs'}
Accuracy: 63.03%
F1_Score:  0.625
```



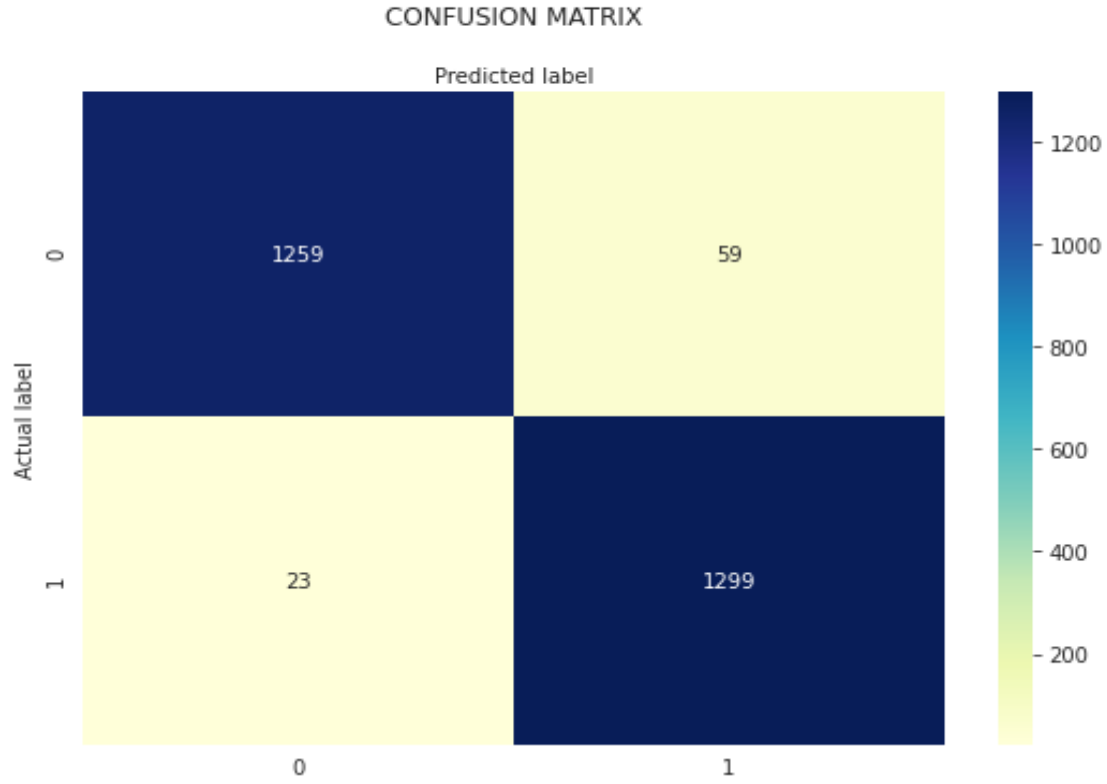
** PERFORMANCE METRICS OF SVC **

Accuracy: 88.3%
F1_Score: 0.889



** PERFORMANCE METRICS OF RANDOM FOREST CLASSIFIER **

Accuracy: 96.89%
F1_Score: 0.969



CONCLUSIONS

1. The Data suffers from severe Class Imbalance. In order to address this issue, the minority class has been oversampled using SMOTE technique.
2. The Logistic Regression algorithm has the least performance metrics with an accuracy of 63%. This is so because Logistic Regression is a type of linear classification.
3. Random Forest Regressor has the most efficient performance metrics with an accuracy of 96.9%.