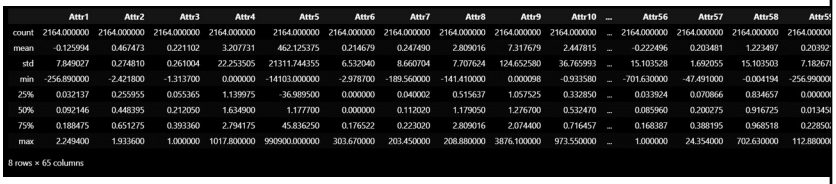
Data Collection and Preprocessing Phase

Date	04 July 2024
Team ID	739898
Project Title	Anticipating Business Bankruptcy
Maximum Marks	6 Marks

Data Exploration and Preprocessing Report

Dataset variables will be statistically analyzed to identify patterns and outliers, with Python employed for preprocessing tasks like normalization and feature engineering. Data cleaning will address missing values and outliers, ensuring quality for subsequent analysis and modeling, and forming a strong foundation for insights and predictions.

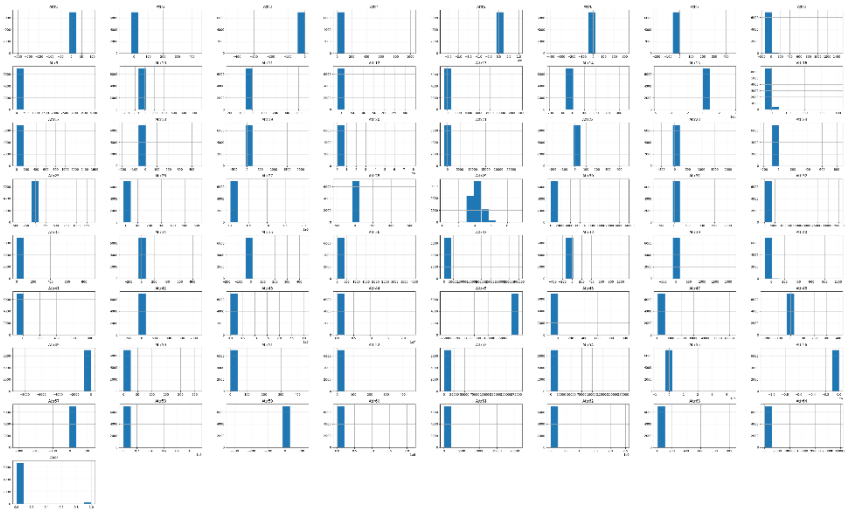
Section	Description

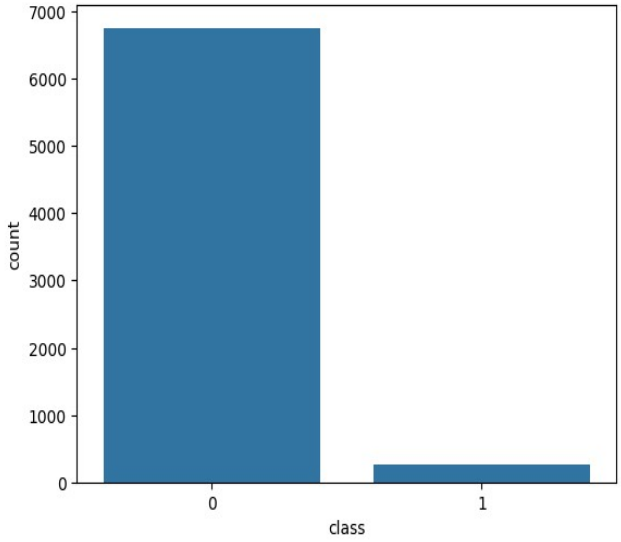
Data Overview	<u>Dimension:</u>
	7012 rows × 65 columns
	
	`Shape of Data: (7012, 65)`
	`number of Rows: 7012`
	`Number of Columns: 65`
	
	<u>Descriptive statistics:</u>
	 

Univariate Analysis	

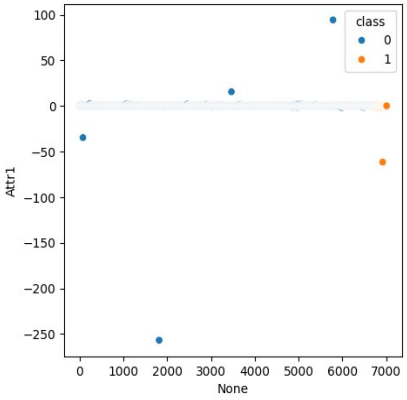
	Attr1	Attr2	Attr3	Attr4	Attr5	Attr6	Attr7	Attr8	Attr9	Attr10	...	Attr56	Attr57	Attr58	Attr5
count	2164.000000	2164.000000	2164.000000	2164.000000	2164.000000	2164.000000	2164.000000	2164.000000	2164.000000	2164.000000	...	2164.000000	2164.000000	2164.000000	2164.000000
mean	-0.125994	0.467473	0.221102	3.207731	462.125375	0.214679	0.247490	2.809016	7.317679	2.447815	...	-0.222496	0.203481	1.223497	0.20392
std	7.849027	0.274810	0.261004	22.253505	21311.744355	6.532040	8.660704	7.707624	124.652580	36.765993	...	15.103528	1.692055	15.103503	7.18267
min	-256.890000	-2.421800	-1.313700	0.000000	-14103.000000	-2.978700	-189.560000	-141.410000	0.000098	-0.933580	...	-701.630000	-47.491000	-0.004194	-256.990000
25%	0.032137	0.255955	0.055365	1.139975	-36.989500	0.000000	0.040002	0.515637	1.057525	0.332850	...	0.033924	0.070866	0.834657	0.000000
50%	0.092146	0.448395	0.212050	1.634900	1.177700	0.000000	0.112020	1.179050	1.276700	0.532470	...	0.085960	0.200275	0.916725	0.013458
75%	0.188475	0.651275	0.393360	2.794175	45.836250	0.176522	0.223020	2.809016	2.074400	0.716457	...	0.168387	0.388195	0.968518	0.22850
max	2.249400	1.933600	1.000000	1017.800000	990900.000000	303.670000	203.450000	208.880000	3876.100000	973.550000	...	1.000000	24.354000	702.630000	112.880000

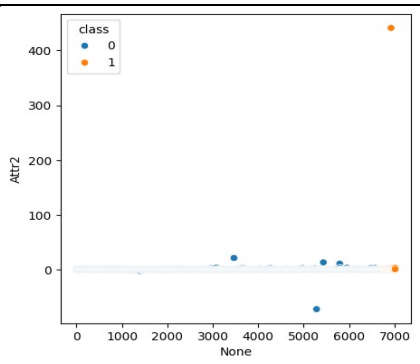
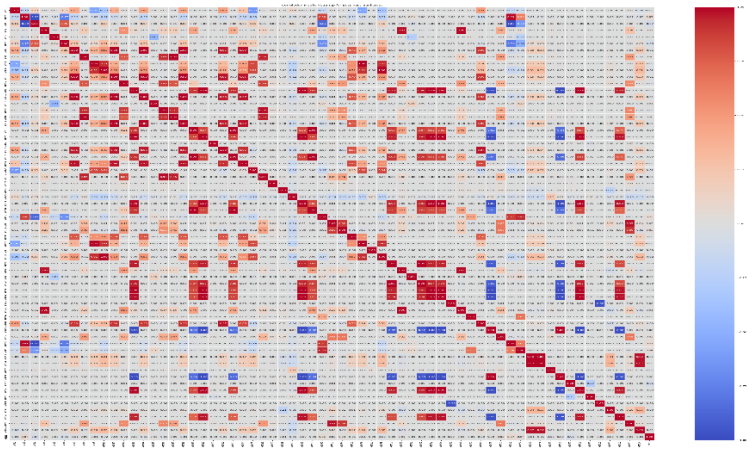
8 rows × 65 columns

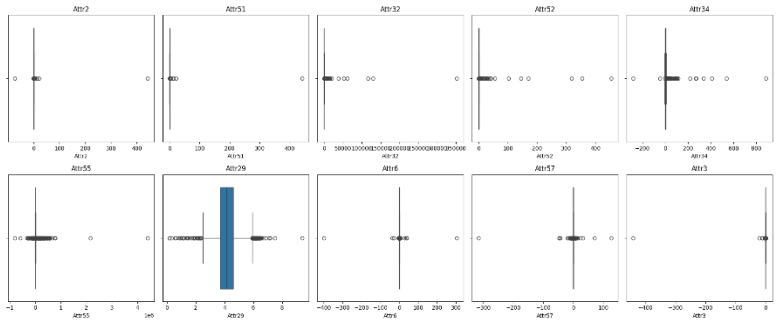


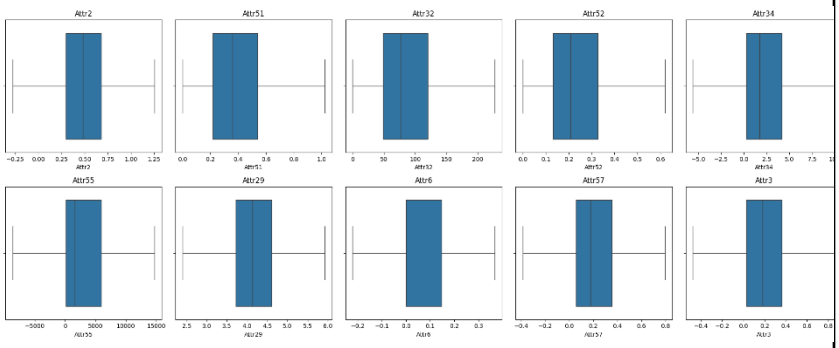


Bivariate Analysis



	
Multivariate Analysis	

Outliers and Anomalies	
	`q1=data.quantile(0.25)`
	`q3=data.quantile(0.75)`
	`IQR=q3-q1`
	`upper_limit=q3+1.5*IQR`
	`lower_limit=q1-1.5*IQR`
	`k=['Attr2','Attr51','Attr32','Attr52', 'Attr34','Attr55','Attr29','Attr6','Attr57','Attr3']`
	`for i in k:`
	`  data[i]=np.where(data[i]>upper_limit[i],upper_limit[i],data[i])`
	`  data[i]=np.where(data[i]<lower_limit[i],lower_limit[i],data[i])`



Data Preprocessing Code Screenshots

Loading Data

```
data = pd.read_csv("/content/drive/MyDrive/1year.csv")
```

```
data.head()
```

	Attr1	Attr2	Attr3	Attr4	Attr5	Attr6	Attr7	Attr8	Attr9	Attr10	...	Attr56	Attr57	Attr58	Attr59	Attr60	Attr61	Attr62	Attr63	Attr64	class
0	0.20055	0.37951	0.39641	2.0472	32.351	0.38825	0.24976	1.3305	1.1389	0.50494	...	0.121960	0.39718	0.87804	0.001924	8.416	5.1372	82.658	4.4158	7.4277	0
1	0.20912	0.49988	0.47225	1.9447	14.786	0	0.25834	0.99601	1.6996	0.49788	...	0.121300	0.42002	0.85300	0	4.1486	3.2732	107.350	3.4	60.987	0
2	0.24866	0.69592	0.26713	1.5548	-1.1523	0	0.30906	0.43695	1.309	0.30408	...	0.241140	0.81774	0.76599	0.69484	4.9909	3.951	134.270	2.7185	5.2078	0
3	0.081483	0.30734	0.45879	2.4928	51.952	0.14988	0.092704	1.8661	1.0571	0.57353	...	0.054015	0.14207	0.94598	0	4.5746	3.6147	86.435	4.2228	5.5497	0
4	0.18732	0.61323	0.2296	1.4063	-7.3128	0.18732	0.18732	0.6307	1.1559	0.38677	...	0.134850	0.48431	0.86515	0.12444	6.3985	4.3158	127.210	2.8692	7.898	0

5 rows × 65 columns

Handling Missing Data

```
(data.eq('?')).any()
```

```
Attr1     True
Attr2     True
Attr3     True
Attr4     True
Attr5     True
          ...
Attr61    True
Attr62    False
Attr63    True
Attr64    True
class     False
Length: 65, dtype: bool
```

```
data.replace('?',np.NAN,inplace=True)
```

```
data.isnull().sum()
```

```
Attr1     3
Attr2     3
Attr3     3
Attr4    30
Attr5     8
         ..
Attr61   22
Attr62    0
Attr63   30
Attr64   34
class     0
Length: 65, dtype: int64
```

```
data.isnull().sum().sum()
```

```
5818
```

```
for i in range(1, 65):
    data[f'Attr{i}'] = pd.to_numeric(data[f'Attr{i}'], errors='coerce')
```

```
data=data.fillna(data.mean())
```

	`data.isnull().sum().sum()`
	`0`
	`data.isnull().any()`
	`Attr1 False`
	`Attr2 False`
	`Attr3 False`
	`Attr4 False`
	`Attr5 False`
	`...`
	`Attr61 False`
	`Attr62 False`
	`Attr63 False`
	`Attr64 False`
	`class False`
	`Length: 65, dtype: bool`

Data Transformation	`x=selected_variable`
	`y=data['class']`
	`x_scaled=pd.DataFrame(StandardScaler(copy=False).fit_transform(x))`
	`x_scaled.columns=x.columns`
	`x.head()`
	Attr2 Attr51 Attr32 Attr52 Attr34 Attr55 Attr29 Attr6 Attr57 Attr3
	0 0.37951 0.37854 94.14 0.25792 0.56393 34869.0 5.9443 0.38825 0.39718 0.39641
	1 0.49988 0.49988 122.17 0.33472 2.98760 2304.6 3.6884 0.00000 0.42002 0.47225
	2 0.69592 0.48152 176.93 0.48474 1.42740 6332.7 4.3749 0.00000 0.81774 0.26713
	3 0.30734 0.30734 91.37 0.25033 0.37581 20545.0 4.6511 0.14988 0.14207 0.45879
	4 0.61323 0.56511 147.04 0.40285 0.32340 3186.6 4.1424 0.18732 0.48431 0.22960
	SMOTE
	`!pip install imblearn`
	`from imblearn.over_sampling import SMOTE`
	`sm=SMOTE(random_state=123) # Now SMOTE is defined`
	`x_sm, y_sm = sm.fit_resample(x_scaled, y)`
	`print(f'''Shape of X before SMOTE: {x_scaled.shape}`
	`Shape of X after SMOTE: {x_sm.shape}''','\n\n')`
	`print(f'''Target Class dstribution before SMOTE;\n{y.value_counts(normalize=True)}`
	`Target Class distribution after SMOTE :\n{y_sm.value_counts(normalize=True)}''')`
	`Requirement already satisfied: imblearn in /usr/local/lib/python3.10/dist-packages (0.0)`
	`Requirement already satisfied: imbalanced-learn in /usr/local/lib/python3.10/dist-packages (from imblearn) (0.10.1)`
	`Requirement already satisfied: numpy>=1.17.3 in /usr/local/lib/python3.10/dist-packages (from imbalanced-learn->imblearn) (1.25.2)`
	`Requirement already satisfied: scipy>=1.3.2 in /usr/local/lib/python3.10/dist-packages (from imbalanced-learn->imblearn) (1.11.4)`
	`Requirement already satisfied: scikit-learn>=1.0.2 in /usr/local/lib/python3.10/dist-packages (from imbalanced-learn->imblearn) (1.2.2)`
	`Requirement already satisfied: joblib>=1.1.1 in /usr/local/lib/python3.10/dist-packages (from imbalanced-learn->imblearn) (1.4.2)`
	`Requirement already satisfied: threadpoolctl>=2.0.0 in /usr/local/lib/python3.10/dist-packages (from imbalanced-learn->imblearn) (3.5.0)`
	`Shape of X before SMOTE: (7012, 10)`
	`Shape of X after SMOTE: (13512, 10)`
	`Target Class dstribution before SMOTE;`
	`class`
	`0 0.963491`
	`1 0.036509`
	`Name: proportion, dtype: float64`
	`Target Class distribution after SMOTE :`
	`class`
	`0 0.5`
	`1 0.5`
	`Name: proportion, dtype: float64`

Feature Engineering	Attached the codes in final submission.
Save Processed Data	-