Configuration Manual

MSc Research Project
FinTech

Saumya Chaudhary
Student ID: x21175365

School of Computing
National College of Ireland

Supervisor: Brian Byrne

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Saumya Chaudhary
Student ID:	x21175365
Programme:	FinTech
Year:	2023
Module:	MSc Research Project
Supervisor:	Brian Byrne
Submission Due Date:	14/08/2023
Project Title:	Configuration Manual
Word Count:	420
Page Count:	5

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Saumya Chaudhary
Date:	17th September 2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Configuration Manual

Saumya Chaudhary
x21175365

1 Environment Setup

The studies were conducted in a Google Colab environment (Bisong and Bisong (2019)) utilizing a Jupyter Notebook (Kluyver et al. (2016)) and the Python programming language(Sanner et al. (1999)). Table 2 shows the various tools used with their purpose.

Sl. No.	**Tools and Technologies**	**Purpose**
1	Microsoft Excel	*For the purpose of storing and retrieving Taiwanese bankruptcy data in CSV format*
2	Google Colaboratory	*Free, cloud-based Jupyter notebook for collaborative Python programming and experimentation*
3	Jupyter Notebook	*To implement Python Codes*
4	Python	*For carrying out a high-level coding task for machine learning*

Table 1: Various Tools and Technologies, and Purpose

Following Python Packages were used

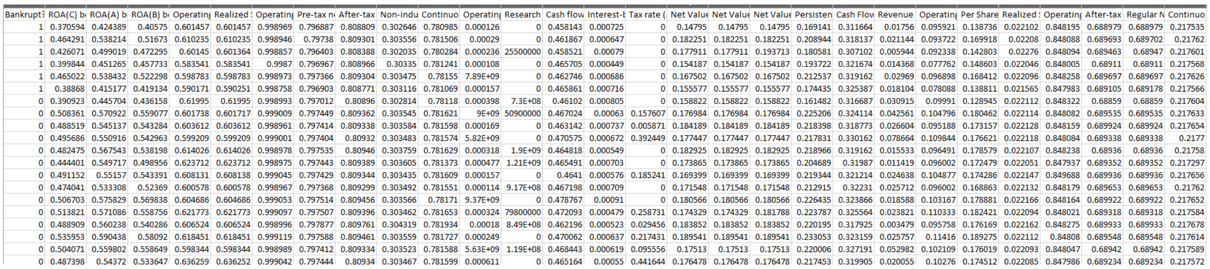
Name	**Description**
Pandas	*Data processing*
Numpy	*Linear algebra with features*
Seaborn, Pyplot and Matplotlib	*Data Visualization*
Sci-kit learn	*For Data pre-processing and model implementation*
Imbalanced learn	*For Data resampling and ensemble model implementation*

Table 2: Various Tools and Technologies, and Purpose

2 Data Analysis

2.1 Data Upload

The data was collected in CSV format as shown below



Bankrupt?	ROA(C) b	ROA(A) b	ROA(B) b	Operating	Realized	Operating	Pre-tax n	After-tax	Non-indu	Continuo	Operating	Research	Cash flow	Interest-b	Tax rate (	Net Value	Net Value	Net Value	Persisten	Cash Flow	Revenue	Operating	Per Share	Realized	Operating	After-tax	Regular N	Continuo
1	0.370594	0.424389	0.40575	0.601457	0.601457	0.998969	0.796887	0.808809	0.302646	0.780985	0.000126	0	0.458143	0.000725	0	0.14795	0.14795	0.14795	0.169141	0.311664	0.01756	0.095921	0.138736	0.022102	0.848195	0.688979	0.688979	0.217535
1	0.464291	0.538214	0.51673	0.610235	0.610235	0.998946	0.79738	0.809301	0.303556	0.781506	0.00029	0	0.461867	0.000647	0	0.182251	0.182251	0.182251	0.208944	0.318137	0.021144	0.093722	0.169918	0.02208	0.848088	0.689693	0.689702	0.21762
1	0.426071	0.499019	0.472295	0.60145	0.601364	0.998857	0.796403	0.808388	0.302035	0.780284	0.000236	25500000	0.458521	0.00079	0	0.177911	0.177911	0.193713	0.180581	0.307102	0.005944	0.092338	0.142803	0.02276	0.848094	0.689463	0.68947	0.217601
1	0.399844	0.451265	0.457733	0.583541	0.583541	0.9987	0.796967	0.808966	0.30335	0.781241	0.000108	0	0.465705	0.000449	0	0.154187	0.154187	0.154187	0.193722	0.321674	0.014368	0.077762	0.148603	0.022046	0.848005	0.68911	0.68911	0.217568
1	0.465022	0.538432	0.522298	0.598783	0.598783	0.998973	0.797366	0.809304	0.303475	0.78155	7.89E+09	0	0.462746	0.000686	0	0.167502	0.167502	0.167502	0.212537	0.319162	0.02969	0.096898	0.168412	0.022096	0.848258	0.689697	0.689697	0.217626
1	0.38868	0.415177	0.419134	0.590171	0.590251	0.998758	0.796903	0.808771	0.303116	0.781069	0.000157	0	0.465861	0.000716	0	0.155577	0.155577	0.155577	0.174435	0.325387	0.018104	0.078088	0.138811	0.021565	0.847983	0.689105	0.689178	0.217566
0	0.390923	0.445704	0.436158	0.61995	0.61995	0.998993	0.797012	0.80896	0.302814	0.78118	0.000398	7.3E+08	0.46102	0.000805	0	0.158822	0.158822	0.158822	0.161482	0.316687	0.030915	0.09991	0.128945	0.022112	0.848322	0.68859	0.68859	0.217604
0	0.508361	0.570922	0.559077	0.601738	0.601717	0.999009	0.797449	0.809362	0.303545	0.781621	9E+09	50900000	0.467024	0.00063	0.157607	0.176984	0.176984	0.176984	0.225206	0.324114	0.042561	0.104796	0.180462	0.022114	0.848082	0.689535	0.689535	0.217633
0	0.488519	0.545137	0.543284	0.603612	0.603612	0.998961	0.797414	0.809338	0.303584	0.781598	0.000169	0	0.463142	0.000737	0.005871	0.184189	0.184189	0.184189	0.218398	0.318773	0.026604	0.095188	0.173157	0.022128	0.848159	0.689924	0.689924	0.217654
0	0.495686	0.550916	0.542963	0.599209	0.599209	0.999001	0.797404	0.80932	0.303483	0.781574	5.82E+09	0	0.470575	0.000672	0.392449	0.177447	0.177447	0.177447	0.217831	0.330162	0.078664	0.109844	0.176621	0.022118	0.848084	0.689338	0.689338	0.2177
0	0.482475	0.567543	0.538198	0.614026	0.614026	0.998978	0.797535	0.80946	0.303759	0.781629	0.000318	1.9E+09	0.464818	0.000549	0	0.182925	0.182925	0.182925	0.218966	0.319162	0.015533	0.096491	0.178579	0.022107	0.848238	0.68936	0.68936	0.21758
0	0.444401	0.549717	0.498956	0.623712	0.623712	0.998975	0.797443	0.809389	0.303605	0.781373	0.000477	1.21E+09	0.465491	0.000703	0	0.173865	0.173865	0.173865	0.204689	0.31987	0.011419	0.096002	0.172479	0.022051	0.847937	0.689352	0.689352	0.217297
0	0.491152	0.55157	0.543391	0.608131	0.608138	0.999045	0.797429	0.809344	0.303435	0.781609	0.000157	0	0.4641	0.000576	0.185241	0.169399	0.169399	0.169399	0.219344	0.321214	0.024638	0.104877	0.174286	0.022147	0.849688	0.689936	0.689936	0.217656
0	0.474041	0.533308	0.52369	0.600578	0.600578	0.998967	0.797368	0.809299	0.303492	0.781551	0.000114	9.17E+08	0.467198	0.000709	0	0.171548	0.171548	0.171548	0.212915	0.32231	0.025712	0.096002	0.168863	0.022132	0.848179	0.689653	0.689653	0.21762
0	0.506703	0.575829	0.569838	0.604686	0.604686	0.999053	0.797514	0.809456	0.303566	0.78171	9.37E+09	0	0.478767	0.00091	0	0.180566	0.180566	0.180566	0.226435	0.323866	0.018588	0.103167	0.178881	0.022166	0.848164	0.689922	0.689922	0.217652
0	0.513821	0.571086	0.558756	0.621773	0.621773	0.999097	0.797507	0.809396	0.303462	0.781653	0.000324	79800000	0.472093	0.000479	0.258731	0.174329	0.174329	0.181788	0.223787	0.325564	0.023821	0.110333	0.182421	0.022094	0.848021	0.689318	0.689318	0.217584
0	0.488909	0.560238	0.540286	0.606524	0.606524	0.998996	0.797877	0.809761	0.304319	0.781934	0.00018	8.49E+08	0.462196	0.000523	0.029456	0.183852	0.183852	0.183852	0.220195	0.317925	0.003479	0.095758	0.176169	0.022162	0.848275	0.689933	0.689933	0.217678
0	0.535953	0.590438	0.58092	0.618451	0.618451	0.999119	0.797588	0.809461	0.303559	0.781727	0.000249	0	0.470062	0.000637	0.217431	0.189541	0.189541	0.189541	0.233053	0.323159	0.025757	0.11416	0.189275	0.022112	0.84808	0.689548	0.689548	0.217614
0	0.504071	0.559802	0.558649	0.598344	0.598344	0.998989	0.797412	0.809334	0.303523	0.781588	5.63E+09	1.19E+08	0.468443	0.000619	0.095556	0.17513	0.17513	0.17513	0.220006	0.327191	0.052982	0.102109	0.176019	0.022093	0.848047	0.68942	0.68942	0.217589
0	0.487398	0.54372	0.533647	0.636259	0.636252	0.999042	0.797444	0.80934	0.303467	0.781599	0.000611	0	0.465164	0.00055	0.441644	0.176478	0.176478	0.176478	0.217453	0.319905	0.020055	0.10276	0.174512	0.022085	0.847986	0.689234	0.689234	0.217572

Figure 1: Data in Excel

2.2 Checking for null values and duplicates

It's crucial to check if there are any missing values in the dataset as this can degrade the model quality. Also, checking for duplicates is an important phase. The steps include

Step 1: df.info() gives the null value information Step 2: df.duplicated() gives number of duplicates values.

2.3 Feature Selection

The features were chosen using Pearson's correlation co-efficients as shown below. This create a set of features by dropping highly correlated independent variables. Based on the literature, it is found that $> |0.7|$ is assumed to be strong correlation (Dormann et al. (2013)).

2.4 Data Normalization

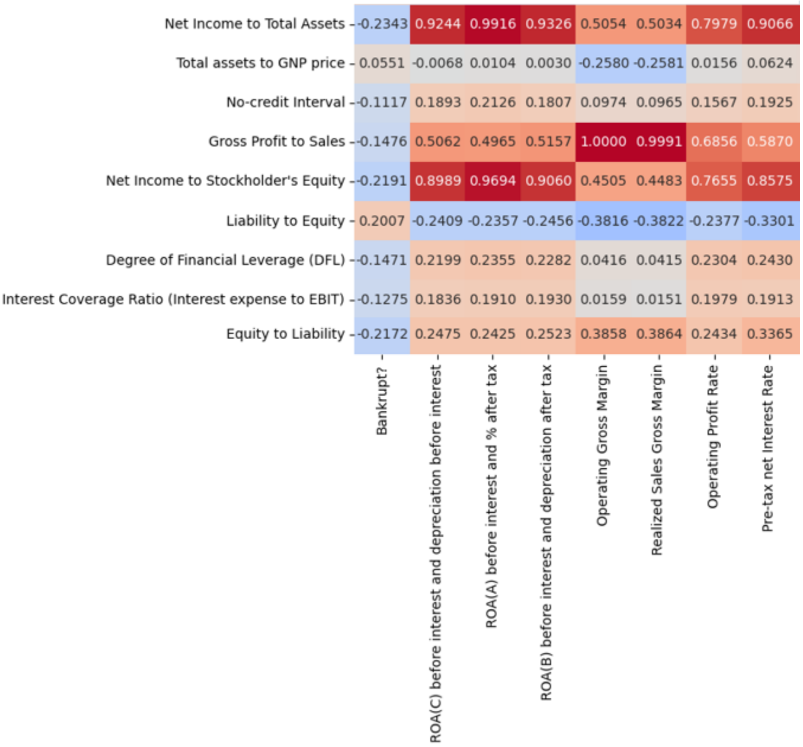
All the variables were scaled using min-max scaler and standard scaler as shown below.

2.5 Dimensionality Reduction

Principal Component Analysis (PCA) is a widely used technique for reducing the dimensionality of high-dimensional data while preserving as much variance as possible (Abdi and Williams (2010)). It entails converting the initial variables into a new collection of uncorrelated variables (principal components) that account for the majority of the variability in the data. In this study, the top 8 principal components accounted for 99% of the variation in the dataset.

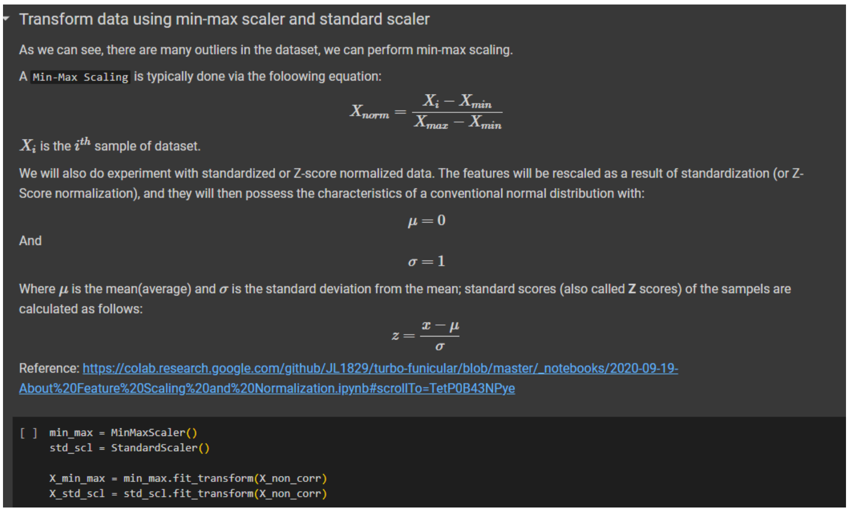
2.6 Train and Test Split

Splitting data into training and test set are important as they will be used to train and evaluate the models. The 70% of the data were used to train the model while 30% to test the model using stratified sampling strategy.



	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
Net Income to Total Assets	-0.2343	0.9244	0.9916	0.9326	0.5054	0.5034	0.7979	0.9066
Total assets to GNP price	0.0551	-0.0068	0.0104	0.0030	-0.2580	-0.2581	0.0156	0.0624
No-credit Interval	-0.1117	0.1893	0.2126	0.1807	0.0974	0.0965	0.1567	0.1925
Gross Profit to Sales	-0.1476	0.5062	0.4965	0.5157	1.0000	0.9991	0.6856	0.5870
Net Income to Stockholder's Equity	-0.2191	0.8989	0.9694	0.9060	0.4505	0.4483	0.7655	0.8575
Liability to Equity	0.2007	-0.2409	-0.2357	-0.2456	-0.3816	-0.3822	-0.2377	-0.3301
Degree of Financial Leverage (DFL)	-0.1471	0.2199	0.2355	0.2282	0.0416	0.0415	0.2304	0.2430
Interest Coverage Ratio (Interest expense to EBIT)	-0.1275	0.1836	0.1910	0.1930	0.0159	0.0151	0.1979	0.1913
Equity to Liability	-0.2172	0.2475	0.2425	0.2523	0.3858	0.3864	0.2434	0.3365

Figure 2: Correlation Between a Few Variables



Transform data using min-max scaler and standard scaler

As we can see, there are many outliers in the dataset, we can perform min-max scaling.

A `Min-Max Scaling` is typically done via the foloowing equation:

$$X_{norm} = \frac{X_i - X_{min}}{X_{max} - X_{min}}$$

X_i is the i^{th} sample of dataset.

We will also do experiment with standardized or Z-score normalized data. The features will be rescaled as a result of standardization (or Z-Score normalization), and they will then possess the characteristics of a conventional normal distribution with:

$$\mu = 0$$

And

$$\sigma = 1$$

Where μ is the mean(average) and σ is the standard deviation from the mean; standard scores (also called **Z** scores) of the sampels are calculated as follows:

$$z = \frac{x - \mu}{\sigma}$$

Reference: https://colab.research.google.com/github/JL1829/turbo-funicular/blob/master/_notebooks/2020-09-19-About%20Feature%20Scaling%20and%20Normalization.ipynb#scrollTo=TetP0B43NPye

```
min_max = MinMaxScaler()
std_scl = StandardScaler()

X_min_max = min_max.fit_transform(X_non_corr)
X_std_scl = std_scl.fit_transform(X_non_corr)
```

Figure 3: Scaling

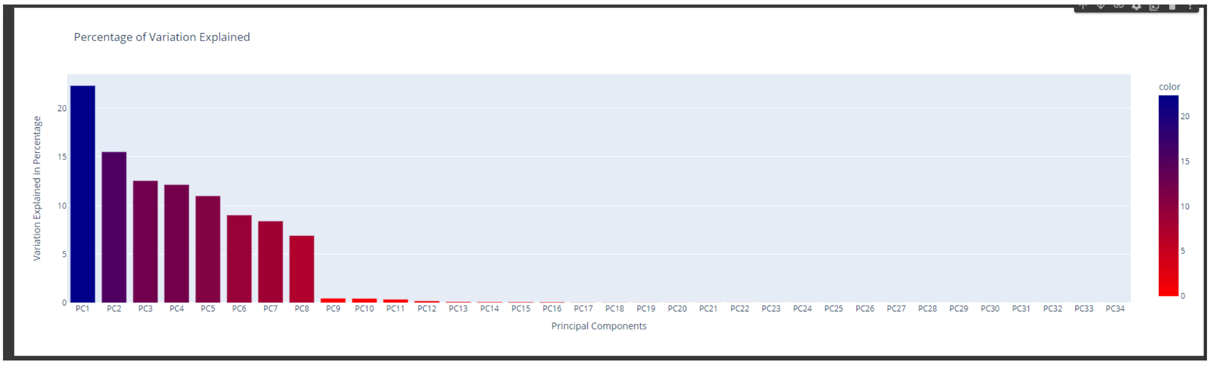


Figure 4: PCA

2.7 Resampling

As the data is biased towards non-bankrupt events, the resampling strategies including Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al. (2002); Smiti and Soui (2020)), Random Oversampling (ROS) (Mohammed et al. (2020)), and Random Undersampling (RUS) (Wang and Liu (2021); Liu et al. (2008)) have been implemented. The SMOTE and Over-sampling created new data points for minority class, i.e., class with bankruptcy events.

References

Abdi, H. and Williams, L. J. (2010). Principal component analysis, *Wiley interdisciplinary reviews: computational statistics* **2**(4): 433–459.

Bisong, E. and Bisong, E. (2019). Google colaboratory, *Building machine learning and deep learning models on google cloud platform: a comprehensive guide for beginners* pp. 59–64.

Chawla, N. V., Bowyer, K. W., Hall, L. O. and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique, *Journal of artificial intelligence research* **16**: 321–357.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., Gruber, B., Lafourcade, B., Leitão, P. J. et al. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance, *Ecography* **36**(1): 27–46.

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B. E., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J. B., Grout, J., Corlay, S. et al. (2016). Jupyter notebooks-a publishing format for reproducible computational workflows., *Elpub* **2016**: 87–90.

Liu, X.-Y., Wu, J. and Zhou, Z.-H. (2008). Exploratory undersampling for class-imbalance learning, *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* **39**(2): 539–550.

Mohammed, R., Rawashdeh, J. and Abdullah, M. (2020). Machine learning with oversampling and undersampling techniques: overview study and experimental results, *2020*

11th international conference on information and communication systems (ICICS), IEEE, pp. 243–248.

Sanner, M. F. et al. (1999). Python: a programming language for software integration and development, *J Mol Graph Model* **17**(1): 57–61.

Smiti, S. and Soui, M. (2020). Bankruptcy prediction using deep learning approach based on borderline smote, *Information Systems Frontiers* **22**: 1067–1083.

Wang, H. and Liu, X. (2021). Undersampling bankruptcy prediction: Taiwan bankruptcy data, *Plos one* **16**(7): e0254030.