International Monetary Fund (IMF)

Read Me First- IMF SPR Research Analyst Take-Home Assignments

Boyuan Wang

Aug, 4

Overview:

To effectively review my assignment results, consider the following two approaches:

1. **Detailed Review:** If you have sufficient time, start with the read me file and follow the introduction in Part 1 to review each task's corresponding files sequentially.
2. **Quick Overview:** If you prefer a concise summary of my problem-solving approach for all four parts, or lack time to delve into Stata coding, refer to the second section of this document (Explanation for Four Assignments) for a brief overview of each task.

1. Detailed Review-Attachments

Part 1: Assess proficiency in data processing using Excel.

- Attachment 1: Data Processing in Excel -test_Boyuan Wang.xlsx

Part 2: Evaluate data manipulation capabilities in Stata.

- Attachment 2: Data Manipulation in Stata - Stata_Boyuan Wang.do

Part 3: Develop a predictive model for financial crises using macroeconomic indicators.

- Attachment 3: Predicting Financial Crises - Ass3_Boyuan Wang.ipynb

Part 4 (Bonus): Showcase Python skills in an NLP task involving web scraping and PDF text extraction.One

- Attachment 4: NLP Analysis Results - Ass4_Boyuan Wang.ipynb

Other Attachments(folder):

- Attachment 5: - Ass4_Boyuan Wang.xlsx

- Attachment 6: -Ass4_Boyuan Wang.json

- Attachment 7: -Ass3_Model_Boyuan Wang.zip

- Attachment 8: -Ass4_saveText_Boyuan Wang.zip

2. Quick Overview- Explaination for Four Assignments

Assignment 1: Data Processing in Excel

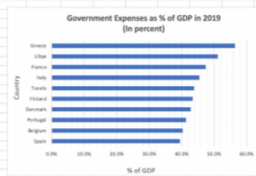
For Question 1, the 2019 Real GDP Growth averages, medians, and PPP GDP-weighted averages were calculated for Advanced Economies (AEs), Emerging Markets (EMs), and Low-Income Countries (LICs). The results are as follows:

	NGDP_RPCH_A		
2019	Average	Median	Weighted Average
AE	3.4	3.4	3.7
EM	3.3	3.2	3.6
LIC	3.3	3.2	3.5

	NGDP_RPCH_A(Corresponding Formulas)		
2019	Average	Median	Weighted Average
AE	AVERAGE(GDP!H82:H1648)	MEDIAN(GDP!H82:H1648)	SUMPRODUCT(GDP!H82:H1648,GDP!J82:J1648)/SUM(GDP!J82:J1648)
EM	AVERAGE(GDP!H19:H1684)	MEDIAN(GDP!H19:H1684)	SUMPRODUCT(GDP!H19:H1684,GDP!J19:J1684)/SUM(GDP!J19:J1684)
LIC	AVERAGE(GDP!H10:H1720)	MEDIAN(GDP!H10:H1720)	SUMPRODUCT(GDP!H10:H1720,GDP!J10:J1720)/SUM(GDP!J10:J1720)

For Question 2, columns for "GGE_A" and "GGE_NGDP_A" were added to the GDP tab, and a bar chart was created to display the top 10 countries with the highest government expenses as a percentage of GDP in 2019.

A	B	C	D	E	F	G	H	I	J	K	L	M
country	ifs	iso3	iso2	income_group	dept	year	NGDP_RPCH_A	NGDP_A	PPPGDP_A	PPPGDP Weight	GGE_A	GGE_NGDP_
Greece	174	GRC	GR	AE	EUR	2019	1.8	189.4	323.5	0.00565	106.9	56.4%
Libya	672	LBY	LY	EM	MCD	2019	-19.1	45.7	61.5	0.00079	23.4	51.1%
France	132	FRA	FR	AE	EUR	2019	1.3	2418.1	3062.0	0.05351	1145.4	47.4%
Italy	136	ITA	IT	AE	EUR	2019	0.2	1779.1	2451.9	0.04285	808.4	45.4%
Tuvalu	869	TUV	TV	EM	APD	2019	4.1	0.1	0.0	0.00000	0.0	43.8%
Finland	172	FIN	FI	AE	EUR	2019	1.5	242.5	267.0	0.00467	105.2	43.4%
Denmark	128	DNK	DK	AE	EUR	2019	1.7	2297.7	312.7	0.00547	982.9	42.8%
Portugal	182	PRT	PT	AE	EUR	2019	1.9	210.9	345.4	0.00604	87.1	41.3%
Belgium	124	BEL	BE	AE	EUR	2019	1.3	472.2	572.7	0.01001	190.4	40.3%
Spain	184	ESP	ES	AE	EUR	2019	2.0	1242.8	1923.2	0.03361	490.0	39.4%
Kiribati	826	KIR	KI	LIC	APD	2019	2.3	0.3	0.3	0.00004	0.1	38.1%
Marshall Islands	867	MHL	MH	EM	APD	2019	2.4	0.2	0.2	0.00000	0.1	37.8%
Croatia	960	HRV	HR	EM	EUR	2019	3.0	400.7	112.5	0.00144	150.1	37.5%
Austria	122	AUT	AT	AE	EUR	2019	1.5	398.9	476.3	0.00832	148.4	37.2%
Netherlands	138	NLD	NL	AE	EUR	2019	1.8	803.8	1004.9	0.01756	294.4	36.6%
Cyprus	423	CYP	CY	AE	EUR	2019	3.1	22.0	36.9	0.00064	8.0	36.4%
Micronesia, Fed. States of	868	FSM	FM	EM	APD	2019	1.4	0.4	0.4	0.00000	0.1	35.7%
Kuwait	443	KWT	KW	EM	MCD	2019	0.5	41.0	311.7	0.00399	14.6	35.6%
Germany	134	DEU	DE	AE	EUR	2019	0.5	3443.0	4441.9	0.07762	1220.0	35.4%
Sweden	144	SWE	SE	AE	EUR	2019	0.9	4994.6	563.7	0.00985	1761.1	35.3%
Slovenia	961	SVN	SI	AE	EUR	2019	2.9	48.2	79.5	0.00139	16.9	35.1%
Japan	158	JPN	JP	AE	APD	2019	1.0	556144.2	5733.6	0.10020	191700.0	34.5%
Eritrea	643	ERI	ER	LIC	AFR	2019	3.1	31.8	6.5	0.00105	10.9	34.2%
Norway	142	NOR	NO	AE	EUR	2019	1.9	3656.6	410.6	0.00717	1174.2	32.1%
Brunei Darussalam	516	BRN	BN	EM	APD	2019	1.8	16.9	35.9	0.00046	5.3	31.5%
United Kingdom	112	GBR	GB	AE	EUR	2019	1.3	2210.4	3158.5	0.05520	692.8	31.3%
Bosnia and Herzegovina	963	BIH	BA	EM	EUR	2019	2.8	34.9	49.8	0.00064	10.9	31.3%



International Monetary Fund (IMF)

Assignment 2: Data Manipulation in Stata

For Question 3, the GDP data was imported from the Excel file, maintaining the same variable names to ensure consistency and accuracy in subsequent analysis.

```
* Import GDP data
import excel "/Users/wangboyuan/Desktop/Stata
Test
8.2/TakeHomeAssignment_Hybrid/test_excel.xlsx",
sheet("GDP") firstrow

* Merge datasets & Replace missing value
merge 1:1 _n using
"/Users/wangboyuan/Desktop/Stata Test
8.2/TakeHomeAssignment_Hybrid/oil_exporters.dta",
nogen

replace oil_exporters = 0 if missing(oil_exporters)
```

For Question 4, the "NGDP_RPCH_A_max" variable was created to capture the highest annual growth rate by country (2011-2019). The dataset was collapsed by year and oil exporters, then reshaped to a wide format for analysis.

```
* Generate a variable called "NGDP_RPCH_A_max"
bysort country: egen NGDP_RPCH_A_max = max(NGDP_RPCH_A)

* Collapse the dataset by year and oil_exporters
collapse (mean) NGDP_RPCH_A, by(year oil_exporters)

* Reshape the dataset to wide
reshape wide NGDP_RPCH_A, i(oil_exporters) j(year)
```

For Question 5, the dataset was saved in .dta and .xlsx formats, with the Excel sheet named "data" and the first row containing variable names.

```
* Save files
save "/Users/wangboyuan/Desktop/Stata Test
8.2/TakeHomeAssignment_Hybrid/dataset_Boyuan Wang.dta",
replace

export excel using "/Users/wangboyuan/Desktop/Stata Test
8.2/TakeHomeAssignment_Hybrid/dataset_Boyuan Wang.xlsx",
sheet("data") firstrow(variables),replace
```

By running the above code, the reshaped dataset shows the average NGDP_RPCH_A for oil exporters and non-exporters from 2011 to 2019.

oil_exporters	NGDP_RPCH_A 2011	NGDP_RPCH_A 2012	NGDP_RPCH_A 2013	NGDP_RPCH_A 2014	NGDP_RPCH_A 2015	NGDP_RPCH_A 2016	NGDP_RPCH_A 2017	NGDP_RPCH_A 2018	NGDP_RPCH_A 2019
0	3.97	3.90	3.22	3.37	2.56	2.79	3.34	3.21	2.51
1	3.95	6.26	3.16	2.76	2.30	2.89	2.60	2.72	2.85

International Monetary Fund (IMF)

Assignment 3: Predicting Financial Crises using Macroeconomic Indicators

Important Note,: The output may have difference between this document and the jupytor note book because of the random number seed.

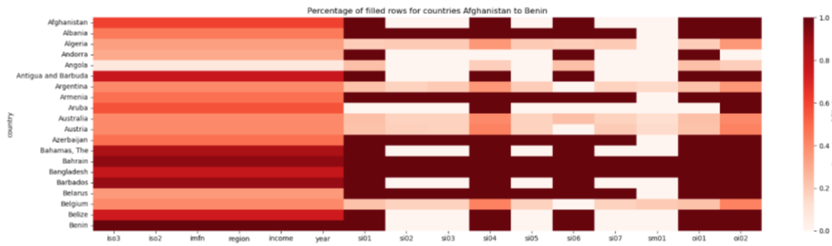
The structured approach demonstrates the application of data science techniques to economic forecasting and provides insights for predicting financial crises (The tasks are structured as follows):

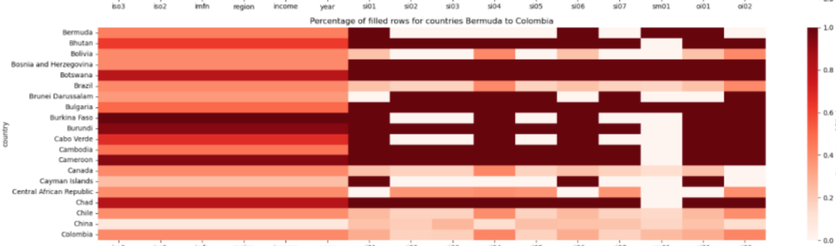
Task	**Description**	**Corresponding Sections**
Task 1: **Data Cleaning and Exploration**	Assess dataset completeness, handle missing data, perform feature engineering, and explore basic statistics and relationships of features.	Overview of dataset
		Check the labels and features
		Merge the two datasets
		Missing data analysis
		Summary of missing data analysis
		Data exploration
		Summary of data exploration
Task 2: **Feature Engineering**	Derive new features and transform existing ones for modeling.	Data preprocessing and feature engineering
Task 3: **Model Building**	Split data into training and testing sets. Configure, train, and evaluate models like XGBoost and Random Forest.	Train-test split
		Oversampling
		Oversampling method choice
		Model Config (XGBoost)
		Model Config (Random Forest)
		Extra data processing for Logistic Regression
		Comparison between models
Task 4: **Communication**	Compare and visualize performance. Explain model choice rationale, discuss performance, and suggest improvements.	Draw the models' performance
		Summary

International Monetary Fund (IMF)

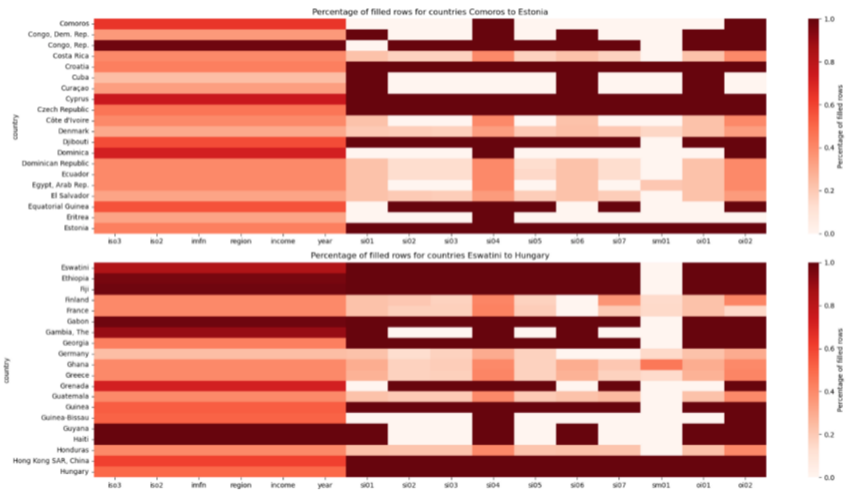
In this assignment, two datasets were provided: features and labels. The features dataset includes ten macroeconomic indicators for all countries from 1960 to 2021, while the labels dataset records the years in which a banking crisis occurred in each country. The first task involved merging these two datasets

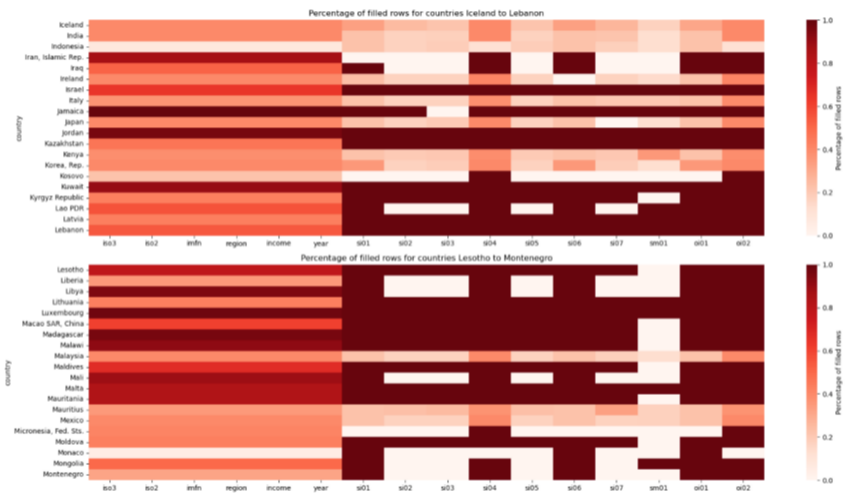
The data loss ratio due to unmatched records was calculated, revealing a data loss of approximately 50%. However, the label dataset itself was found to be about 80% smaller than the features dataset, which was deemed acceptable.



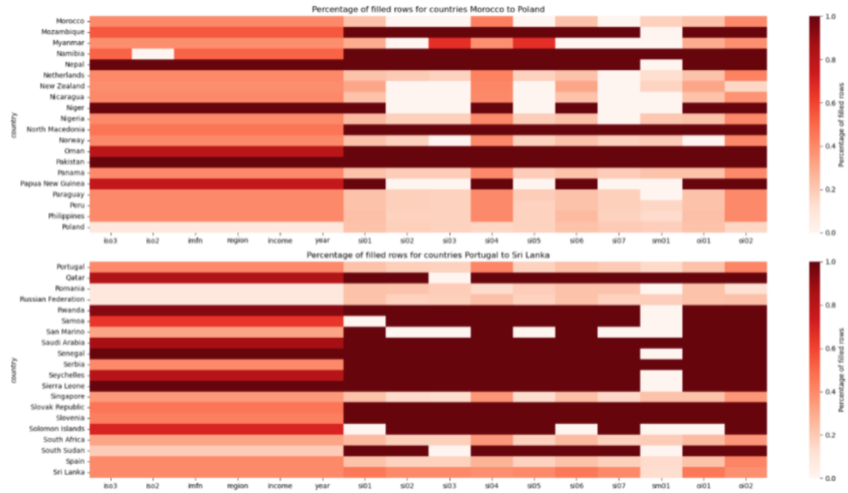


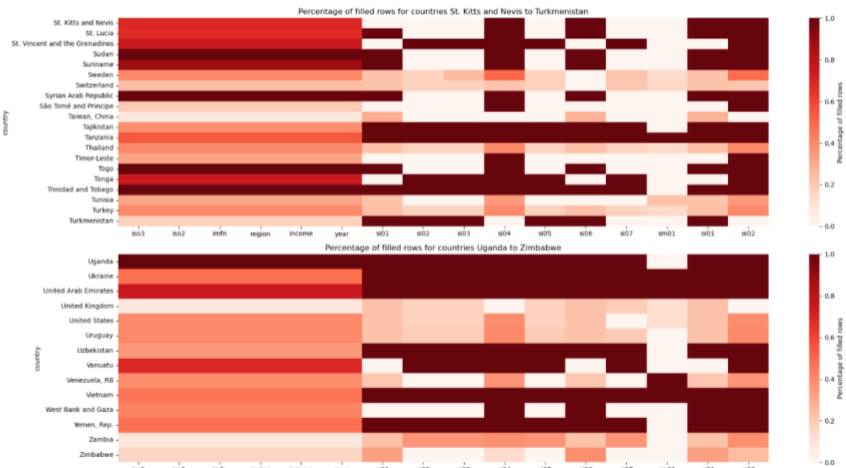
International Monetary Fund (IMF)





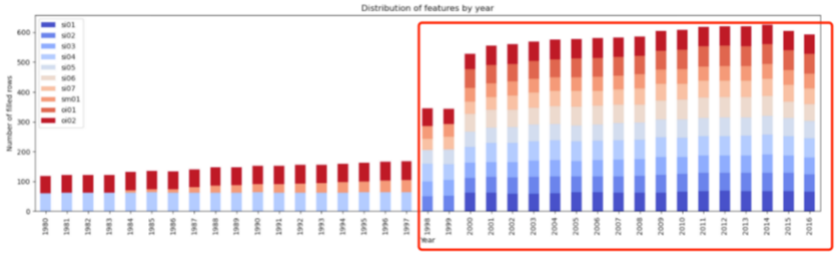
International Monetary Fund (IMF)





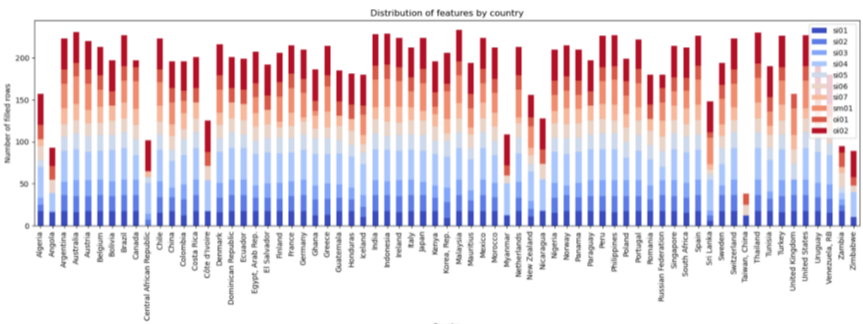
The provided heatmaps illustrate the missing data ratio for each country, with darker shades indicating a higher percentage of missing data. These visualizations are essential for understanding the dataset's completeness, revealing which countries have substantial gaps in their macroeconomic indicators. **Identifying these gaps is crucial for subsequent data cleaning and imputation steps to ensure the reliability and accuracy of the predictive models.**

International Monetary Fund (IMF)



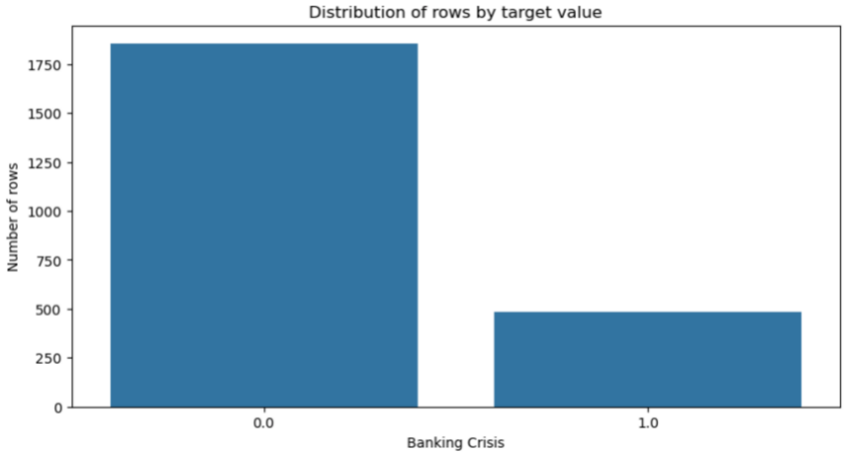
As the data analysis progressed, it was determined that the starting year for the analysis should be 1998. By plotting data availability, it was observed that from 1998 onwards, all columns started to have a more completed data **Therefore, 1998 was chosen as the starting point for the training dataset to ensure robust and comprehensive analysis.**

The next step involved examining the regional data distribution. It was found that the data was relatively evenly distributed across different regions, with no particular region being significantly overrepresented. **Consequently, no specific adjustments were necessary for the regional data distribution.**

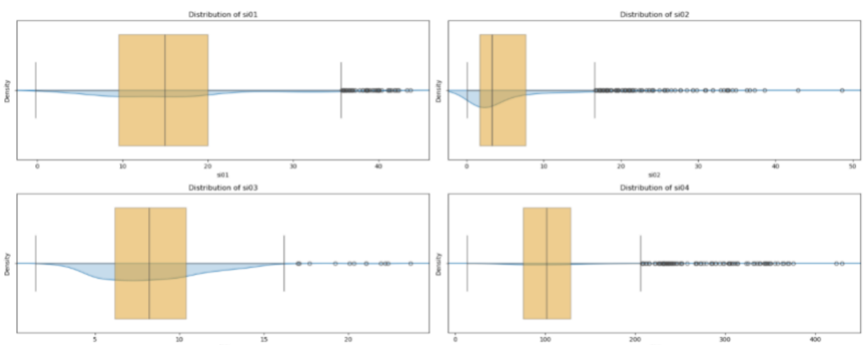


The third aspect to examine was the distribution of the target variable. It was observed that the likelihood of a crisis occurring was much lower than the likelihood of no crisis, indicating an imbalanced dataset. To address this, oversampling techniques were applied to manually increase the instances of the minority target. **Six different oversampling methods were tested, and it was determined that SMOTE Tomek was the most effective** (this is detailed in the following sections).

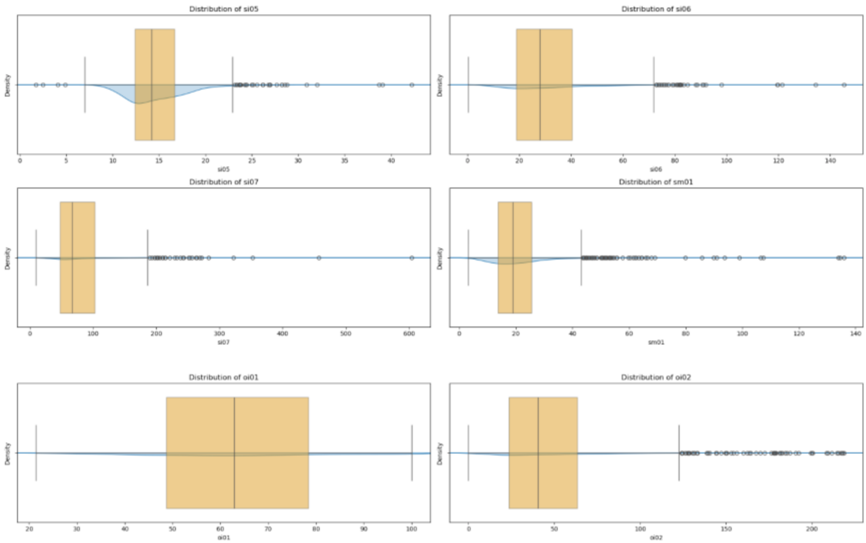
International Monetary Fund (IMF)



The next step was to examine the distribution of each feature. It was found that most features were **right-skewed**. To address this skewness, **log transformation was applied to normalize the distributions, aiming to approximate a normal distribution for better model performance**.

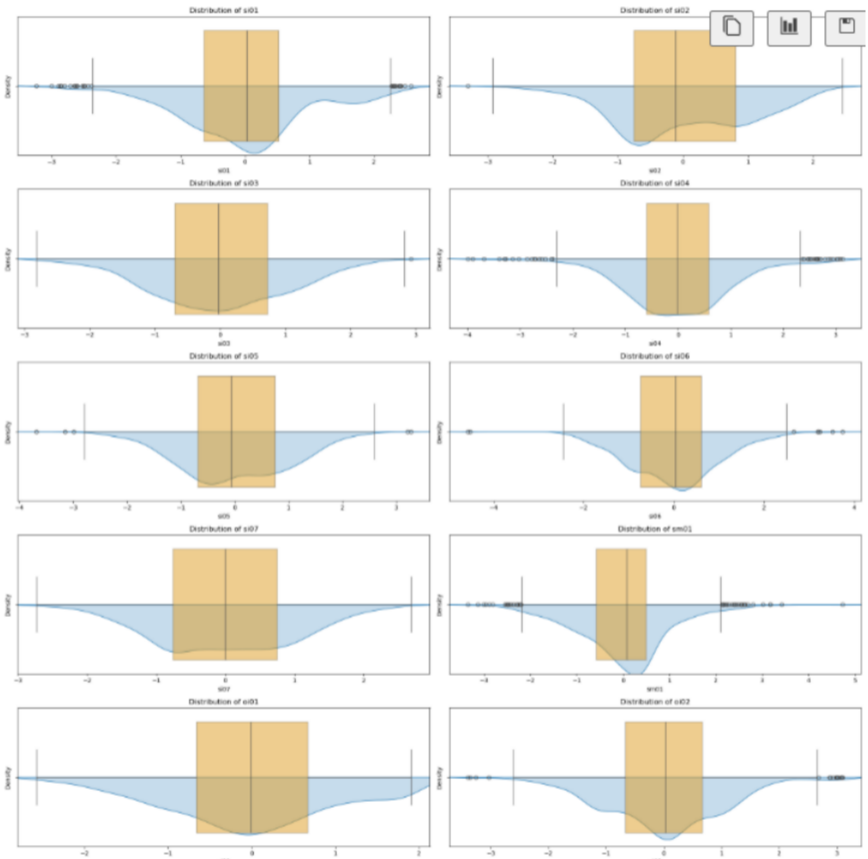


International Monetary Fund (IMF)



International Monetary Fund (IMF)

After the transformation, **the data follows a more normal distribution with a mean of 0 and a standard deviation of 1**. To achieve this, the data was **standardized** by groups of countries to **ensure that incomparable data between countries does not mix and affect the model performance**.

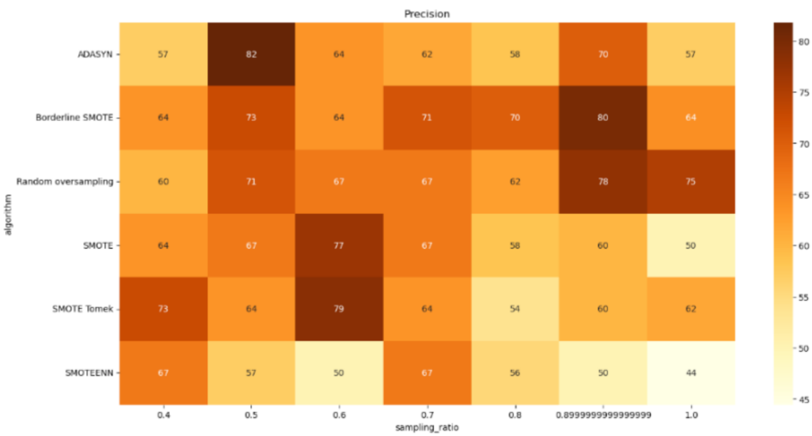


International Monetary Fund (IMF)

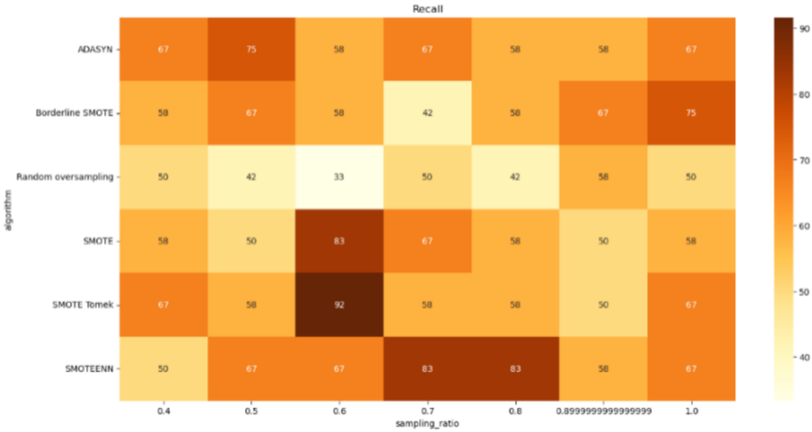
After completing the data preprocessing steps, the next task was to select an oversampling method **due to the observed imbalance in the dataset**, with significantly more '0' values in the target column than '1'. Six different methods were evaluated:

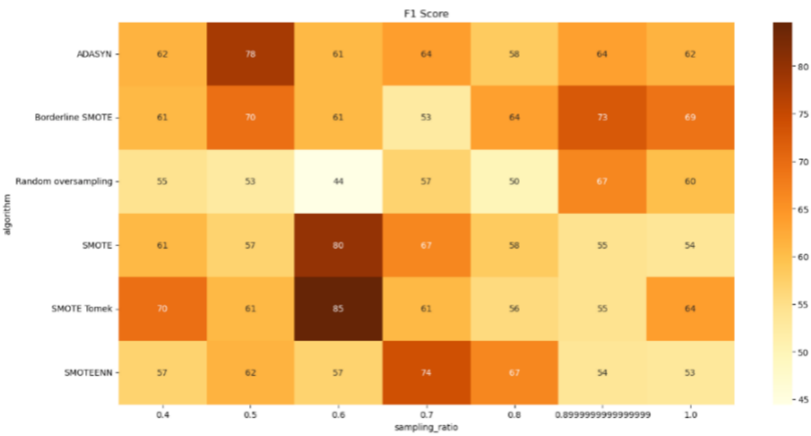
- `Random oversampling`: Randomly duplicate the minority class
- `SMOTE`: Synthetic Minority Over-sampling Technique
- `ADASYN`: Adaptive Synthetic Sampling Approach
- `SMOTEENN`: SMOTE + Edited Nearest Neighbors
- `SMOTETomek`: SMOTE + Tomek links
- `BorderlineSMOTE`: Borderline SMOTE

Among these, **SMOTETomek was chosen for its superior overall performance.**



International Monetary Fund (IMF)

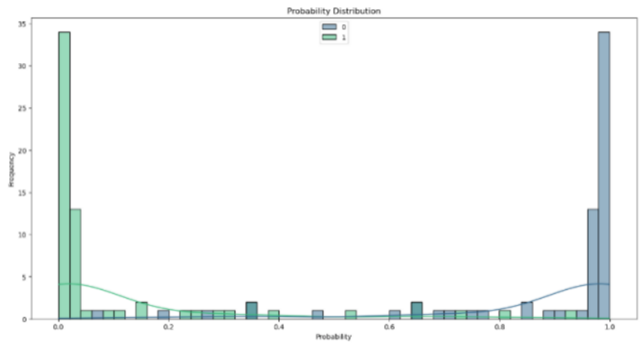


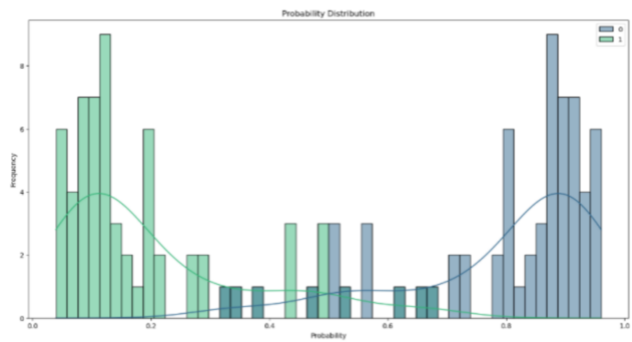


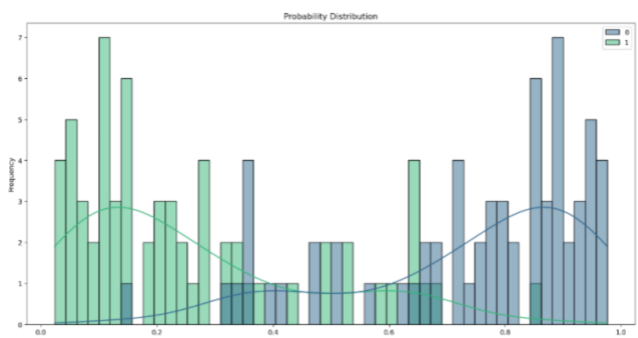
The **SMOTETomek method**, with a **0.6** ratio, showed the best overall performance as indicated by the darkest shades in the evaluation metrics.

Based on the performance evaluation of six oversampling methods, SMOTE Tomek was selected as the optimal approach. **This method achieved 79% precision, 92% recall, and an 85% F1 score, indicating a well-balanced performance in addressing the imbalance in the dataset.**

Three models were trained and evaluated based on their performance. The probability distribution graphs illustrate the predicted probabilities for class 0 (no crisis) in green and class 1 (crisis) in blue. The clear separation between the two classes demonstrates the models' ability to effectively distinguish between crisis and non-crisis scenarios. For simplicity, the highest-performing model was chosen and evaluated using four charts: probability distribution, ROC curve, confusion matrix, and classification report.

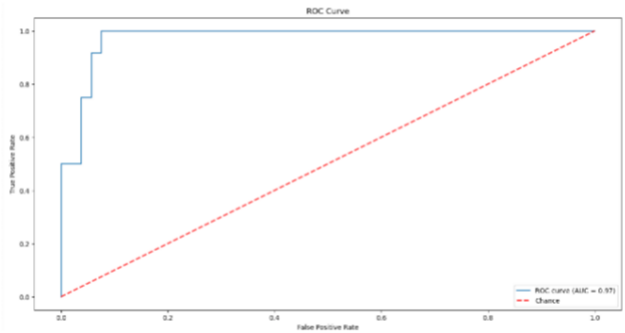


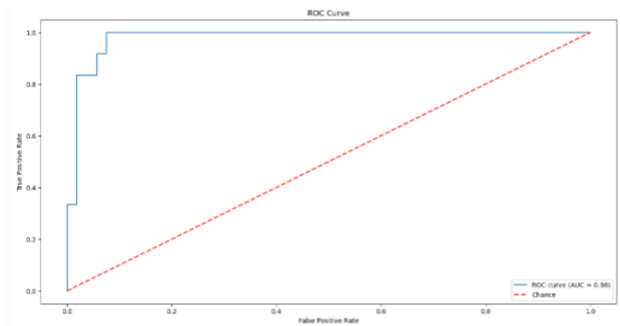


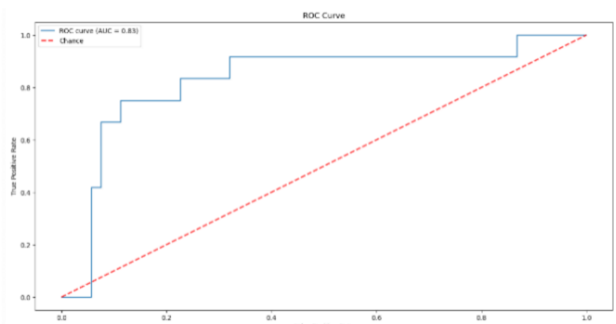


International Monetary Fund (IMF)

The ROC curves compare three models. The first model has an AUC of 0.97, indicating excellent discrimination between crisis and non-crisis. The second model, with an AUC of 0.95, also performs strongly. The third model shows an AUC of 0.88, indicating good but slightly lower performance with more false positives.

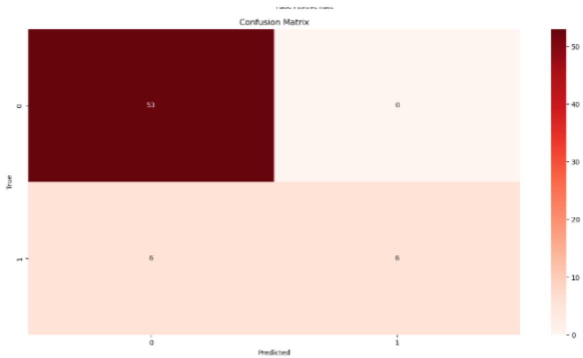


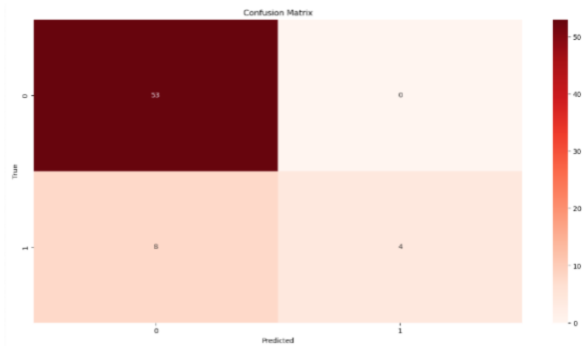


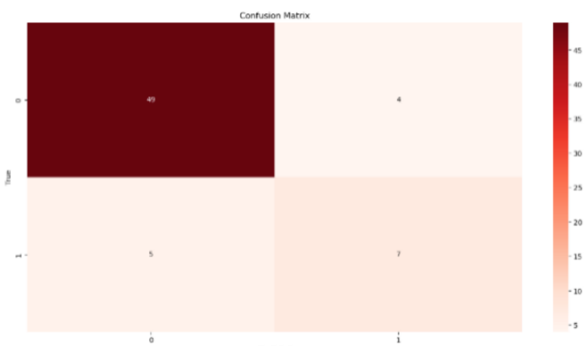


International Monetary Fund (IMF)

The confusion matrices compare the models' performance. The first model identified 53 non-crisis and 6 crisis years, misclassifying 6 crisis years. The second model identified 53 non-crisis and 4 crisis years, misclassifying 8 crisis years. The third model identified 49 non-crisis and 7 crisis years, with 4 misclassified crisis years.

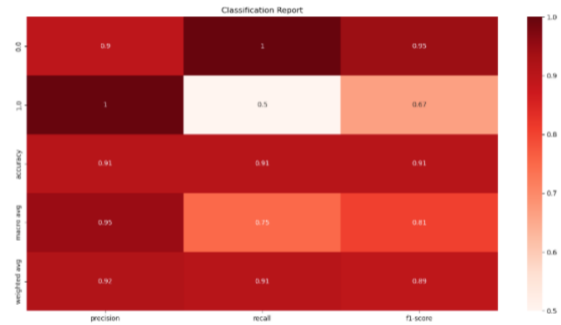


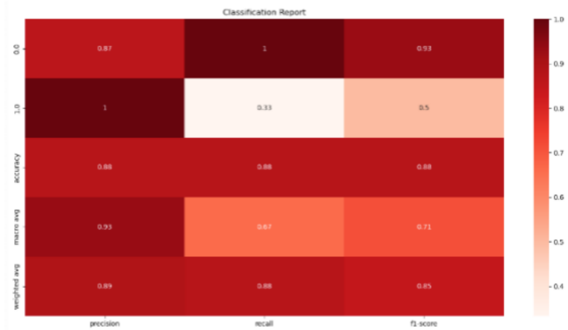


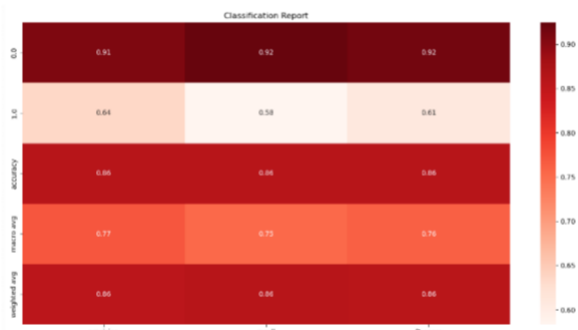


International Monetary Fund (IMF)

The classification reports show the performance metrics for the three models. The first model had an F1-score of 0.67 for crisis years, the second model had an F1-score of 0.57, and the third model had an F1-score of 0.64. Overall, the first model performed the best in terms of balanced accuracy and precision.

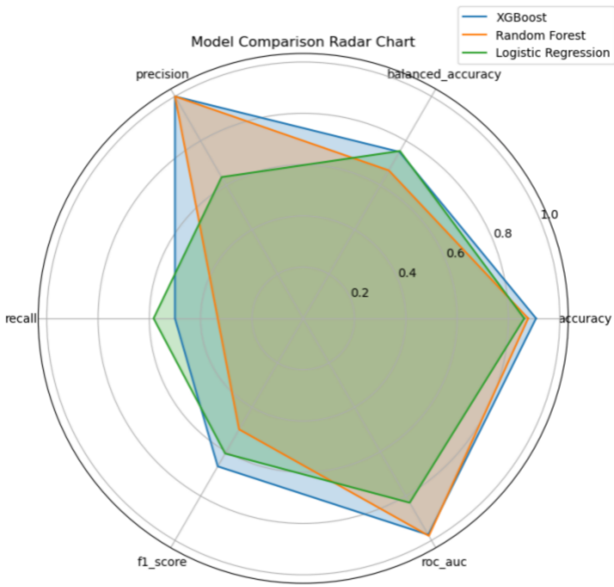






International Monetary Fund (IMF)

Finally, the overall performance among the three models was compared using radar charts. These charts illustrate the performance metrics for XGBoost, Random Forest, and Logistic Regression models on the testing set, including accuracy, precision, recall, and F1 score. XGBoost (blue) shows superior balanced accuracy and precision, while Random Forest (orange) and Logistic Regression (green) also perform well but slightly lag in certain metrics. This visual comparison helps identify the most robust model for predicting financial crises.



Future Possible Improvements:

1. **Stacked Models:** While the XGBoost model is the best performer among the three models, exploring stacked models could potentially enhance performance.

2. **Country-Specific Features:** Given that some data is country-specific, further transfer learning on the data might yield better results.

3. **More Models:** If time permits, experimenting with other models like SVM, neural networks, etc., could help achieve better performance.

International Monetary Fund (IMF)

Assignment 4 (Bonus): NLP Analysis on IMF Country Staff Reports

In Task 4, Part 1, Python was used to scrape all PDF documents related to the United States from the IMF's official site. By iterating through all result pages and collecting metadata like document names and publication dates, a total of 65 PDF files were gathered for analysis.

This task was performed using a Python library capable of handling PDF files. The goal was to extract textual content from each document, resulting in a collection of text files that preserved the original document's content for further examination.