European Economic Letters
ISSN 2323-5233
Vol 14, Issue 2 (2024)
http://eelet.org.uk

Small-Scale Machine Learning Over Scarce and Unreliable Data: Sovereign Credit Grades Prediction

Jean Herskovits

Department of Mathematics, Imperial College London & Nomura International Securities

Jean.herskovits20@imperial.ac.uk

Abstract - According to the Bank for International Settlements, investment banks across the globe now hold more than a trillion US dollars' worth of sovereign bonds. Sovereign credit grades' movements largely drive these positions' risk and volatility. Their prediction is thus crucial to better manage sovereign credit portfolios. Available economic data is unreliable, scarce and restricted by the number of sovereign entities. Established historical models used to set and predict sovereign credit grades centre around restrictive linear modelling. Modern machine learning techniques require numerous parameters and vast datasets to converge over noisy data. Through thorough data processing, motivated by both economic and statistical insight, we put forward classifiers which novelly demonstrate that small neural networks and random forests can calibrate accurately on poor, lopsided macroeconomic datasets. Their accuracy outperforms all known industry and published implementations, both linear and neural network based. Our results are cross-validated on carefully isolated data, both temporally and geographically, replicating a production situation. This architecture is the first capable of predicting sovereign credit grades' evolution with accuracy high enough to meet the intuitive and easily implemented "constant" classifier benchmark. Unlocking the use of modern statistical methods to small, low quality economic and financial datasets, our predictors underlie the core importance of extensive problem specific data pre-processing in machine learning for macroeconomic classification.

Index Terms - Sovereign Credit, Machine Learning for Macroeconomics, Small Neural Networks

Introduction

While investment banks' portfolios are heavily exposed to sovereign debt through different parts of their business, their largest exposure lies in their Retail and Prime Brokerage Departments of Wholesale Divisions. Banks hold sovereign bonds in the name of their clients, which can be institutional (e.g. mutual and pension funds) or buy-side counterparties (e.g. hedge funds and high-wealth individual investors).[1]

While not as risky as owning credit products in their own name, bonds held as collateral still carry a substantial risk [16]. Major investment banks have recently suffered multiple losses from the risk carried by such positions [1,3,8]. Modern regulatory frameworks use credit grades as the cornerstone of risk measurement of credit products, especially for sovereign exposures [6]. The improvement of sovereign credit grades is crucial towards bettering the risk models of banks.

This study describes two models developed in partnership with Nomura's Risk Methodology Group, furthering their predecessors at predicting the movements of the risk charge carried by sovereign bonds. We attempt to model and predict the movements of sovereign credit grades using data available to Nomura. We have access to dozens of financial, social and economic time series tracking the evolution of countries across two decades. To assess the quality of our model, we first build an intuitive and straightforward benchmark, the *"constant"* predictor. This comparison metric simply assumes that a country's grade never changes: for any timestamp and country, it outputs the last available observed credit grade for that country. The low rate of change of sovereign credit grades makes this predictor perfectly accurate in almost 80% of cases. When wrong, this predictor is still often close to the true value. While carrying no predictive power, this benchmark is high. To our knowledge, no prior published model meets its accuracy [17,18]. Typical industry legacy models (*see Section 1*) also are not as precise as this benchmark [4,19,20].

Our neural network classifier aims solely at accuracy and outperforms the *"constant"* predictor (*see Section 3*), managing above 80% perfect accuracy and wrong by more than one notch over only 0.4% of entries. Our Random Forest model intends to be good at identifying grade changes, the minority class in our dataset (*see Section 4*). It achieves $\leq 20\%$ type-II error, while maintaining $\leq 10\%$ type-I error. All our claims are based on cross-validation performed over separated

[1] Other exposures include the exposure to newly issued bonds, which banks buy before reselling them to investors. (For the Japanese government in 2023 alone, new issuance and renewals will reach 39.3 trillion Yen [14]. Banks also keep positive inventories of many sovereign credit products as part of their market making business [5,15].

European Economic Letters
ISSN 2323-5233
Vol 14, Issue 2 (2024)
http://eelet.org.uk

entries. To ensure absence of data leakage [22], the cross-validation is carried over either new countries or the most recent years of data accessible, strictly splitting the test and train sets geographically or temporally.

1. Sovereign Grades Modelling Challenges

We now summarize the main challenges of predicting sovereign credit grades. From this analysis we justify the construction of the *"constant"* predictor. In parallel, these challenges limit the maximal quality of potential models. We establish the class imbalance of the dataset, its unequal geographic distribution and its relatively small effective size.

1.1. Dataset Size & Quality

In theory, there are around 195 countries in the world. Sovereign socio-economic figures have been reported for yearly 20 years, both by public (IMF, World Bank) and private institutions (FactSet, Bloomberg). The maximal potential number of $(Country, Year)$ pairs nears 4000. However, not all countries are graded: Standard and Poor's has only ever graded 164 nations.

In our analysis, we focus on a set of 20 features to predict sovereign credit grades' movements (*see Table 2*). Thus, for a $(Country, Year)$ instance to be usable, 20 indicators need to be available: the dataset's effective size is reduced ten-fold to 404. Moreover, an ideal model would include multiple prior years of data to capture a country's economic trajectory. Using 3 years of data to predict grades decreases the number of entries by 25%.

Furthermore, there are 20 grades classes from *AAA* to *D*. Class sizes can be low, especially as the data's availability is unevenly distributed. Africa (*resp. Europe*) comprises 11% (36%) of available entries despite representing more than 25% of nations (22%).[2] To circumvent the small class sizes, we replace their labels. *Section 4* uses two classes, *"no-change"* and *"change"*. All other predictors attempt to predict grades' changes and directions: $\{\leq -2, -1, 0, 1, \geq 2\}$.[3] While this reformulation increases classes' sizes, it implies that countries at different stages of development have the same label. *Figure 1* confirms that sovereign credit grades' movements differ depending on their level of economic development: the classes are now less homogenous.

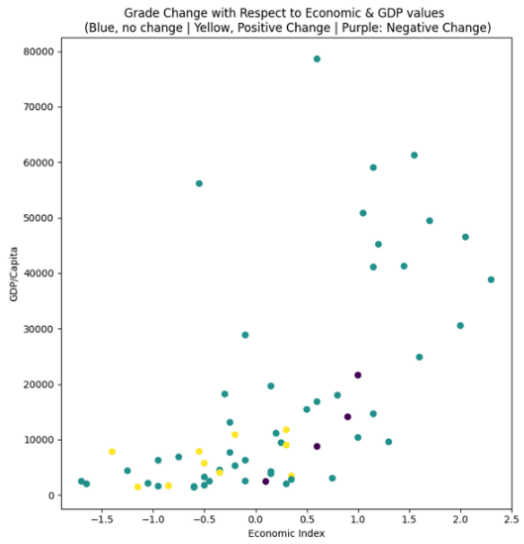

Figure 1: 2017-2018 heterogenous grade movements distributions depending on countries' development level.

Finally, the data's reliability is not perfect. Socio-economic indicators published are not final and are regularly revised [7]. It is not straightforward to integrate data from different eras [24], regions [12] and institutions [2] into one dataset.

[2] According to the United Nations Group of Experts on Geographical Names (UNGEGN) [21].

[3] Neural networks, lacking explainability, use directional classes. Random forest classifiers' feature importance allows for the direction of the prediction to be determined easily despite not being a formal output.

European Economic Letters
ISSN 2323-5233
Vol 14, Issue 2 (2024)
http://eelet.org.uk

1.2. Imbalance & Predictors

We swapped labels from $\{1, \dots, 20\}$ to predicting a grades' change instead. While this increases the classes' volumes, it introduces strong imbalance between their sizes. Indeed, grades on average change every four years only: the *"no-change"* class represents 77% of the dataset. Without careful work, overall accurate models can have low accuracy on the minority classes (*see Section 4*).[4] Plus, grade updates by more than 2 notches within a year are rare and correspond to default events: as they default, nations' ratings jump to *20* from a better grade.[5] From this, one can devise a simple and accurate benchmark for this classification problem. For every $(Country, Year)$ entry, the *"constant"* predictor outputs the previous year's grade for this country.

This predictor has high accuracy (*see Table 1*). However, this predictor holds no predictive power and thus holds no value to the bank other than as a benchmark. Classical models, based on linear regressions, used for the prediction of sovereign grades have substantially lower accuracies (with the 20 features we use, *see Table 2*). Instead of using a linear regression predictor, using a shallow neural network[6] improves accuracy but not does not meet the *"constant"* benchmark.

Prediction for S&P Grades	Exact Match	Within 1 Notch	Within 2 Notches	Total
Linear Model	$\sim 20\%$	$\sim 65\%$	$\sim 85\%$	100%
Constant Model	77.1%	96.3%	98.6%	100%
Neural Network	65%	90%	95%	100%

Table 1: "Constant" model predictions outperforms typical linear regression models as well as simple networks.

2. Data Pre-Processing

2.1. Pre-Processing

Our features are prescribed by internal credit economists (*see Table 2*). They include the features flagged by credit rating agencies as essential to their rating process, but also features considered the most relevant to actual grade updates by credit risk professionals, which we presume are also used by public rating agencies in their grading. Though capturing complex interplay between features is crucial, *Table 2* confirms that the indicators we use are indeed strongly correlated to time series of sovereign credit grades. Most specifically, , features with the highest absolute correlation to sovereign credit grades describe the financial structure of the country in regard to its debt: *External Debt, Servicing Costs, General Government Reserves, Non-Performing Loans*. Social structural indicators such as *Rule of Law* are also correlated to sovereign credit grades.

	Corr. to Sov. Credit Grades	
Feature:	Original Set	Reconstructed
Age Dependency Ratio (%)	-0.02	-0.0
Broad Money to Reserves	-0.03	-0.02
Capital Adequacy Ratio	-0.22	-0.22
CPI Inflation - Change	-0.03	0.01
CPI Inflation - Volatility	-0.0	-0.04
Economic Complexity Score	0.07	0.07
External Debt (% of GDP)	-0.28	-0.27
External Debt (% of Reserves)	-0.25	-0.26
General Government Debt (% of Revenue)	-0.42	-0.43
General Government Interest (% of Revenue)	-0.12	-0.11
Government Effectiveness	0.21	0.23
Inflation (% year-on-year)	0.02	0.02
Local currency share of debt	0	0.04
Net Inward FDI (% of GDP)	0.13	0.14
Non-Performing Loans	-0.23	-0.24
Real Effective Exchange Rate (%)	0.05	0.04
Rule of Law	0.14	0.15
Savings (% of GDP)	0.16	0.16
Short-term Ext Debt (% of Reserves)	-0.04	-0.07

[4] Credit grades are usually reviewed quarterly. We use yearly data as using quarterly entries instead would simply increase the proportion of *"no-change"* instances. Yearly grade updates already capture all the changes (grades rarely change twice in a year and we did not find instances of changes in opposite directions in a year).

[5] Some countries defaulted multiple times across the 20 years encompassed in our study, as they relist bonds after a default event. After the prior default, the newly emitted bonds get a assigned a new credit grade. If available, we use the non-defaulted grade of a country, though we ensure that every country which defaults is reported as such (20) for at least one year.

[6] Deeper networks with many more parameters do not converge in training over our small dataset.

European Economic Letters
ISSN 2323-5233
Vol 14, Issue 2 (2024)
http://eelet.org.uk

Table 2: Correlation of features to sovereign credit grades. While some correlations exhibit near independence, some are highly (anti)-correlated with the target. Our models also use a 20th feature, last year's credit grade.

We outline how we restructure these features to be used within our predictors. While many legacy models [13,25] focus on one year of data, we use multiple years of input at once. Every entry in the training and testing sets thus has the following format:

Input:

$$Feature\ 1\colon \left[Value\left(i_0 - (WL-1)\right), \dots, Value(i_0)\right],$$

$$\dots$$

$$Feature\ M\colon \left[Value\left(i_0 - (WL-1)\right), \dots, Value(i_0)\right]$$

Output: $Credit\ Grade(i_0)$

Where WL is the window length, the number of years used for every entry, M the number of indicators used, and $i_0 \in [2000, 2021]$ the target year for which we want to predict the country's credit grade. In results presented below, every $i_0 \in [2000, 2021]$ is a target year. That is, for every available year and country, we add an entry in the training set: every available $Credit\ Grade(Country, Year)$ is the label of an entry.

Before the data is sliced in the process described above, we reconstruct some missing entries to allow for more data to be used at once. Indeed, even standard institutional data sources such as the World Bank do not reliably provide data for countries that are usually reported on. For credit grades themselves, our prediction target, we do not artificially recreate any values. However, when a country's grade is missing from S&P's grading across the entire $[2000, 2021]$ range, we instead use Moody's or alternatively Nomura's grading.[7] For the features of our model, the data reconstruction process is described thoroughly in *Appendix A.1*. Essentially, each $(Country, Year)$ is treated independently. The hyper-parameter (*Max Fill*) controls for the maximal number of consecutive missing entries that we fill back for each time series. Given (*Max Fill*), each missing entry to be re-filled is calibrated as the weighted average of 3 linear trends, respectively a short-term, a past short-term and a global long-term trend. For data missing in year *T*, the weight of the past short-term (*resp. future*) is inversely proportional to the number of consecutive missing entries in the time series prior to (*after*) year *T*. The long-term trend's weight is set so that the three weights sum to one. The impact of the back-filling process is shown in *Table 2* and *3*.

Max Fill	0	1	2	**3**	4	5	6	7	8	9
Usable Entries	466	618	770	**851**	935	1014	1090	1160	1228	1295

Table 3: Dataset size for varying Max Fill.
WL is set to 3 - Total Entries, with NaNs: 3020.

(*Max Fill = 3*) doubles the size of the training set. Low values of (*Max Fill*) add comparatively many entries in the dataset, filling sporadic “holes" in time series. Greater (*Max Fill*) values largely recreate data before any was available. As this is not our goal, we set (*Max Fill*) to three for the rest of this study.

Varying the window length *WL* also impacts the size of the dataset. For each entry, any missing figure in *WL* years of data across *M* features removes the entry from the dataset. Thus, longer windows reduce the number of available entries, but allow the model to observe more complete financial and economic sovereign trajectories.

Kernel Size	1	2	3	**4**	5	6	7	8	9
Total # Entries	3322	3171	3020	**2869**	2718	2567	2416	2265	2114
Missing Grade Target	793	642	575	**511**	450	396	351	308	269
Usable	1003	934	851	**770**	689	610	534	463	395

Table 4: Dataset WL on number of available entries in the dataset resulting from the described slicing process. Max Fill was set to 3.

[7] Filling individual years is not implemented as it would artificially add grade changes.

European Economic Letters
ISSN 2323-5233
Vol 14, Issue 2 (2024)
http://eelet.org.uk

2.2. Avoiding Leakage

Initially, once our dataset was back-filled and split, we randomly shuffled the entries in training and testing sets, yielding 100% accuracy out-of-sample over entries ≤ 2020. This wrongly perfect accuracy is caused by information leakage from previously seen entries. Indeed, the structure described above entails that predictors have seen credit grades as features in training that are labels of entries in the test.[8] The model is able to differentiate each distinct $(Country, Year)$ entry.[9] Indeed, since each entry has 20 features across multiple years, including past credit grades, the model in practice develops a way to separate all $(Country, Year, Grade)$ entries. It holds strictly no predictive power however, solely relying on already having seen each grade it is required to predict in its training set.

Confirming this data leakage, if the test set is composed of isolated countries, the model's performance is very low. The same substantial drop in accuracy is observed, though less significant, if we isolate the last year of data from the training entirely, instead of individual nations.[10]

The latter implementation replicates real model use once in production: the model needs to predict grades having calibrated its features on previous years of data. Both splitting methods make the prediction task at hand more challenging, as grading methods vary through time and regions.

These processes however, while permitting the user to spot leakage, do not provide a solution the *look ahead* from training. To this end, we analogously stratify the train set separately to ensure that the network has never seen a grade to predict earlier in training. We stratify either by nation first, then by year, or the reverse, yielding the same predictive power presented in *Table 5*.

3. Maximising Overall Accuracy with Neural Networks

The neural network implementation presented in this section beats all other known models and meets the *"constant"* predictor's benchmark. Being able to observe a country's financial and economic history is indeed helpful for sovereign credit grade predictions. Setting *WL* to 3 or 4 yields similar results qualitatively, whilst shorter and longer horizons hinder performance. We opine that a longer horizon, multiplying the dimension of the feature set, makes the calibration less straightforward. Meanwhile, it reduces the effective number of entries, and consequently substantially reduces the accuracy. As for legacy models, two datapoints per feature is not enough to capture a country's trajectory. It does not replicate the *"constant"* predictor either.

We tested multiple classical feed-forward neural networks of varying depth. We first find that deep neural networks, encompassing a large number of parameters, do not perform well and their training usually does not converge well.

Indeed, the small size of our dataset restricts the number of parameters that the model can carry. This analysis is performed as no clear consensus exists yet in the finance field over which kind of Neural Network to use [10,22]. Optimal networks have few hidden layers, ideally two, with a small number of nodes in each. We settled on hidden layers of width 100 and 50 neurons respectively.

Table 5: Our Neural Network outperforms the results from the "constant" mode benchmark. This predictor does not replicate the "constant" predictor. Yet, its accuracy over the minority class is low.

Prediction for S&P Grades	Exact Match	Within 1 Notch	Within 2 Notches	Total
Linear Model	$\sim 20\%$	$\sim 65\%$	$\sim 85\%$	100%
Constant Model	77.1%	96.3%	98.6%	100%
Multi-Year Neural Network	80%	96.6%	100%	100%

However, it is not an ideal predictor. Firstly, its convergence is not perfectly stable: calibrating the network shuffling the training set in various acceptable orders does not yield the same results, as it is common to observe nets' training not converging properly. Furthermore, its accuracy on the *"no-change"* class beats its total accuracy. This is expected as the

[8] Even in the absence of leakage, accounting for the temporal structure in processing is not straightforward [9,23].

[9] No remarkable change in performance was observed modifying batch sizes [11].

[10] We can also spot this issue by only including $(Country, Year)$ entries in the dataset with $Year \equiv 0\ [WL]$. This ensures each $Feature(Country, Year)$ only appears once.

"no-change" class is largely more frequent than the other ones: a large increase in accuracy in the *"change"* class does not compensate for the associated small decrease in accuracy in the large class. Multiple attempts to solve this issue were unsuccessful. Enforcing strict stratification so that the model initially sees *"change"* entries at the same rate as *"no-change"*, is not enough.

As such, we opine that while the imbalance in the size of the two classes is a large driver of the difference in prediction quality, it is not the only one. The stratification process just outlined would have helped towards reducing the performance gap, had size imbalance been the only factor.

We argue that despite the *"no-change"* class being preponderant, the *"change"* class contains an outsized diversity of scenarios. A large majority of instances of the *"no-change"* class in practice represent $(Country, Year)$ pairs during which no significant financial nor economic event happened. By contrast, every entry in the *"change"* class represents a more convoluted national scenario, which justified an upgrade, or more frequently, a downgrade.

The aim of this predictor was solely accuracy, which is satisfied, as this implementation trumps all previous ones, both internal and in published research. For a model used in a systematic fashion without human review, overall performance is the most important metric. Yet, for Credit Risk analysts, who manually review the output of the model, high accuracy on each class, and especially over the *"change"* class, is central. To satisfy this requirement, we now present a random forest classifier, built to minimize the type-II error, i.e. missing a grade change.

4. Minimizing type-II error with Random Forests

As described in the introduction, it is crucial for the bank to flag grades that are going to change. It is substantially more costly to miss a grade being modified than to mistakenly label a $(Country, Year)$ pair as a *"change"*. The random forest predictor presented in this section is intended to minimize such errors, being the type-II error. We manage to construct a promising predictor 80% accurate on the *"change"* class, which remains 92% accurate over *"no-change"* class.[11]

We set *WL* to three in this section, yielding in total 851 entries in our reconstructed dataset. Since we are to implement a random forest classifier, we change the structure of the features given as inputs to the model.[12] For each feature, we replace the *WL* years of data by their average increment (*trend*), average squared increment (*volatility*) and last available value across the *WL* years in the window.

Table 6 outlines the optimal number of trees in the forest, 100, and *Table 7* presents the optimal values of other hyper-parameters.

	Accuracy per Class	
Trees in Random Forest	*"no-change"*	*"change"*
50	86.5%	<50%
100	81.6%	54.4%
200	82.88%	53.76%
500	85.12%	53.28%
1000	83.84%	52.48%

Table 6: Average Random Forest performance per class for varying number of individual decision trees in the forest.

Table 7 also presents the accuracy of the random forest predictor. It outlines accuracy with equal class weights, minimizing the type-II error and minimizing the type-I error (*see Appendix A.2*), the classifier attains 80% accuracy over the *"change"* class, while maintaining over 90% accuracy over the *"no-change"* class.

[11] As our neural network predictor aimed at maximising overall accuracy, we focused on performance across the 2 classes.

[12] Each node in each decision tree from the random forest can only split entries depending on one feature's level relative to a calibrated cutoff value. As such, *WL* years of data per indicator is not sufficient for the model to capture, e.g., the trend or volatility of the indicator, and is thus detrimental to its performance.

Minimization Target	Equal Weights	Type-II Error	Type-I Error
Weight *"no-change"*	1	1	5
Weight *"change"*	1	5	1
Share of Samples per Tree	0.5	0.7	0.7
Share of Features per Tree	0.7	0.7	0.7
Results			
Accuracy over *"change"*	88%	92%	96%
Accuracy over *"no-change"*	76%	80%	48%

Table 7: By assigning unequal weights to the two kinds of errors, we can shift the model's performance. To avoid overfit, Minimum Sample Split is set to eight.

Conclusion

Sovereign credit grades are a central driver of risk over banks' consequent sovereign credit exposures. Thus, better prediction of nations' credit grades, even offering only slight improvement over current models, is of high value to all major financial institutions.

Within the same framework as classical linear model, straightforward neural networks offer substantial improvements already, highlighting the importance of capturing complex interactions between features.

Maximising overall accuracy, the multi-layer perceptron using multiple years of data at once surpasses not only all previous published implementations but is, to our knowledge, the first model to exceed the predictive ability of the *"constant"* predictor. Error rates compared to this benchmark are reduced by 25%. Using it in pricing and risk engines, replacing the *"constant"* predictor or legacy models would entail substantial improvements in efficiency.

Our random forest classifier targets the type-II error, i.e. missing grade changes. Banks are asymmetrically exposed to the mislabelling of grades' outlooks: it is more costly to miss a grade change. We showed that the classes' imbalance, as grade changes only represent 23% of our dataset, is the not main challenge in reducing the type-II error. Indeed, as one class is more populated, its diversity is also smaller. The *"no-change"* class is composed of numerous instances of countries where volatility of the given features is low. This distinction explains why it is inherently harder to anticipate grade changes. Despite this, the random forest classifier performs manages to correctly identify at least 80% of the members of each label.

As the arrangement of the manuscript indicates, the largest share of the work and its main challenges lay in sourcing, pre-processing and carefully structuring our dataset prior to training any statistical engine over it. Our analysis thus shows that future developments in the modern machine learning over small-scale low-quality datasets will also centre on pre-processing.

We would like to thank Alfonso de la Torre from Nomura International Securities and Ségolène McKinnon for their help and feedback in this project.

References

1 – Jose Alonso Olmedo, Rebeca Anguren Martin, Maria Gamoneda Roca, and Pablo Perez Rodriguez. Archegos and Greensill: collapse, reactions and common features. *Financial Stability Review / Banco de Espana (Autumn 2021)*, pages 47–62, 2021.

2 – Ademola Ariyo. Reliability of Macroeconomic data: An evidence from Nigeria's debt data series. *African Development Review - Revue Africaine de Développement*, June 1995.

3 - Antoine Bouveret and Martin Haferkorn. Leverage and derivatives: The case of archegos. *Journal of Securities Operations and Custody,* 15(3): 238–250, 2023.

4 – Richard Cantor and Frank Packer. Determinants and impact of sovereign credit ratings. *Economic Policy Review,* 2(2), October 1996.

5 – Pablo Anaya European Central Bank Longaric, Maciej Grodzicki, Christoph Kaufmann, Allegra Pietsch, Pablo Serrano Ascandoni, Manuela Storz, and Elisa Telesca. Sovereign bond markets and financial stability: examining the risk to absorption capacity. *Financial Stability Review*, November 2023.

6 – Bank for International Settlements. Treatment of sovereign risk in the Basel capital framework. *BIS Quarterly Review*, December 2013.

7 – Katharina Glass and Ulrich Fritsche. Real-time macroeconomic data and uncertainty. *DEP (Socioeconomics) Discussion Papers - Macroeconomics and Finance* Series *6/2014R*, Hamburg, 2015.

8 – Donal Griffin and Lucca De Paoli. Nomura, Mizuho face losses on All Blue fund's failed trades. *Technical report, Bloomberg Markets*, UK, May 2024.

9 – Ali Hakami. Strategies for overcoming data scarcity, imbalance, and feature selection challenges in machine learning models for predictive maintenance. *Scientific Reports, 14(9065),* 2024.

10 – Ronen Israel, Bryan T. Kelly, and Tobias J. Moskowitz. Can machines 'learn' finance? *Journal of Investment Management*, 2020.

11 – Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *CoRR*.

12 – Iacob N. Koch-Weser. The reliability of china's economic data: An analysis of national output. *Technical report, U.S.-China Economic AND Security Review Commission (USCC)*, January 2013.

13 – Albert Metz and Richard Cantor. Moody's credit rating prediction model. *Technical report, Moody's Investors Service*, 2006.

14 – Japanese Government Ministry of Finance. Japanese government bonds, *January 2023 Newsletter*, 2023.

15 – Ministry of Finance, Japanese Government. Announcement of liquidity enhancement auction on December 20, 2023, 2023.

16 – Bank of England. Dp3/22 – Operational Resilience: Critical third parties to the UK financial, 2022.

17 – Bart H.L. Overes and Michel van der Wel. Modelling sovereign credit ratings: Evaluating the accuracy and driving factors using machine learning techniques. *Computational Economics, 61:1273–1303*, 2023.

18 – Huseyin Ozturk, Ersin Namli, and Halil Ibrahim Erdal. Modelling sovereign credit ratings: The accuracy of models in a heterogeneous sample. *Economic Modelling, 54:469–478*, April 2016.

19 – Standard & Poor's. Sovereign government rating methodology and assumptions. *Technical report*, 2011.

20 – Standard & Poor's Global Ratings. Sovereign ratings history since 1975. *Technical report*, 2015.

21 – UNGEGN. UNGEGN list of country names. Technical report, United Nations Group of Experts on Geographical Names (UNGEGN), July 2011.

22 – Weiguan Wang and Johannes Ruf. Information leakage in backtesting. *SSRN*, 2022.

23 – Weiguan Wang and Johannes Ruf. A note on spurious model selection. *Quantitative Finance, 22:1–4, August 2022.*

24 – David R. Weir. The reliability of historical macroeconomic data for comparing cyclical stability. *The Journal of Economic History, 46:353–365*, 1986.

25 – Tom Wilde. Creditrisk+ a credit risk management framework. *Credit Suisse Financial Products*, October 1997.

Appendix

A.1. Data Reconstruction

Let us denote $\mathrm{V} \coloneqq \{V_t, t \in [2000,2021]\}$ the time series corresponding to an indicator's reported values for an individual country, with yearly frequency.

A missing entry in V for year t is denoted $V_t = \emptyset$. Firstly, if $\forall t \in [2000,2021], V_t = \emptyset$, then the reconstructed time series is also empty: for each country, entirely missing features are left empty. In the following, we thus assume that:

European Economic Letters
ISSN 2323-5233
Vol 14, Issue 2 (2024)
http://eelet.org.uk

$$\exists t \in [2000,2021]\, s.t.\, V_t \neq \emptyset.$$

For $t \in [2000,2021]$, let us define T^+ and T^- as functions of t:

$$T^+(t) = \begin{cases} \infty, & if\ V_s = \emptyset, \forall s \geq t, \\ \min(s \geq t,\ V_s \neq \emptyset) & else. \end{cases}$$

$$T^-(t) = \begin{cases} -\infty, & if\ V_s = \emptyset, \forall s \leq t, \\ \min(s \leq t,\ V_s \neq \emptyset) & else. \end{cases}$$

Let $\hat{V} := \{\hat{V}_t, t \in [2000,2021]\}$ be the reconstructed time series. Given $t \in [2000,2021]$, let us now assume $V_t = \emptyset$. Given $(Max\ Fill) \in \mathbb{N}^+$, let us first determine whether V_t is to be back-filled or not.

$\forall t \in [2000,2021]$ we have:

If $(Max\ Fill) \geq min(|T^+(t) - t|, |t - T^-(t)|) > 0$, the value of $\hat{V}_t$ is reconstructed.

Else, $\hat{V}_t = V_t$.

As above, let us now assume that we have $t \in [2000,2021]$ such that $V_t = \emptyset$, and that V_t is to be back-filled under the provided $(Max\ Fill)$.

• If $T^+(t) < \infty$ and $T^-(t) > -\infty$, we define:

$$\gamma_L = \frac{\mathbb{1}_{\{T^-(t)=t-1\}}}{2+\mathbb{1}_{\{T^+(t)=t+1\}}} \cdot \frac{T^+(t)-t}{T^+(t)-T^-(t)},$$

$$\gamma_R = \frac{\mathbb{1}_{\{T^+(t)=t+1\}}}{2+\mathbb{1}_{\{T^-(t)=t-1\}}} \cdot \frac{t-T^-(t)}{T^+(t)-T^-(t)}.,$$

Then:

$$\hat{V}_t = \frac{V_{T^-(t)}(T^+(t) - t) + V_{T^+(t)}\big(t - T^-(t)\big)}{T^+(t) - T^-(t)}$$
$$+ \gamma_L \cdot \big(V_{T^-(t)} - V_{T^-(t)-1}\big) + \gamma_R \cdot \big(V_{T^+(t)} - V_{T^+(t)+1}\big).$$

The first term is a linear interpolation between the two observations closest to the year being reconstructed. γ_L and γ_R terms respectively represent the impacts of future and past short-term trends on the current value. When t is T^- (*resp.* t *is* T^+), V's tendency prior to V_{T^-} (*after* V_{T^+}) is taken into account as $\big(V_{T^-(t)} - V_{T^-(t)-1}\big)$ (*resp.* $\big(V_{T^+(t)} - V_{T^+(t)+1}\big)$).

• Else, if $T^+(t) < \infty$:

$$\hat{V}_t = V_{T^+(t)} + (V_{T^+(t)} - \bar{V}) \cdot \frac{T^+(t) - t}{\max\{s \in [2000, 2021], V_s \neq \emptyset\} - T^+(t)} + \frac{V_{T^+(t)} - \frac{V_{T^+(t)} + V_{T^+(t)+1} + V_{T^+(t)+2}}{3}}{T^+(t) - t}.$$

• Else, when $T^-(t) > -\infty$:

$$\hat{V}_t = V_{T^-(t)} + (V_{T^-(t)} + \bar{V}) \cdot \frac{t - T^-(t)}{T^-(t) - \min\{s \in [2000, 2021], V_s \neq \emptyset\}} + \frac{V_{T^-(t)} - \frac{V_{T^-(t)} + V_{T^-(t)-1} + V_{T^-(t)-2}}{3}}{t - T^-(t)}.$$

In the two latter cases, where $T^+(t) = \infty$ or $T^-(t) = -\infty$, the reconstructed values are computed starting from the last originally available value in the time series, $T^+(t)$ and $T^-(t)$ respectively. That is for instance where $T^+(t) = \infty$, we start by back-filling $(T^+(t) - 1)$, and only then $(T^+(t) - 2)$ if $(Max\ Fill \geq 2)$. This ensures that the terms of the rolling average $V_{T^+(t)}, V_{T^+(t)+1}$ and $V_{T^+(t)+2}$ $(resp.\ V_{T^-(t)}, V_{T^-(t)-1}\ and\ V_{T^-(t)-2})$ are available at each step of the back-filling process.

In edge cases, when $(T^+(t) + 1)$ or $(T^-(t) - 1)$ are not well-defined for instance, their associated weight γ is set to zero.

European Economic Letters
ISSN 2323-5233
Vol 14, Issue 2 (2024)
http://eelet.org.uk

A.2. Class Weights in Random Forests

The goal of the Random Forests implementations is to ensure that good overall performance also translates to satisfactory accuracy over the minority *"change"* class. To maximise the performance over the *"no-change"* class and minimize the type-II error, we increase the weight of the minority class to five.[13] In the previous results, we used an equal weighting: mislabels over each class carried the same cost.

In practice, to increase a class's weight, we modify the function *I* assessing the nodes' purity.

Let us consider a node *N* in a decision tree. The node's population is split in two classes $\{"change","no-change"\}$. Their respective ratios in the node's population are $p_N("change")$ and $p_N("no-change")$. Since there are only two classes, we have:

$$p_N("no-change") = 1 - p_N("change").$$

From this, we define the node's entropy impurity as follows:

$$I(N) = -\sum_{\chi \in \{"(no-)change"\}} \omega(\chi) \cdot \left(p_N(\chi) \ln\left(p_N(\chi)\right)\right).$$

When splitting a node in children nodes, the separation is determined to minimize the aggregate entropy of the resulting children nodes. By setting $\omega("change")$ to five instead of one, we incentivize the construction of individual decision trees to classify correctly the $(Country, Year)$ pairs labelled as *"change"*.

[13] We are thus overweighting the *"no-change"* class above its balanced weight. Its balanced weight is two (or four, if the weight of the *"change"* class is normalized to one). Indeed: $\frac{Dataset\ Size}{\{\#\ "change"\}\cdot\{\#\ classes\}} = \frac{0.851}{(0.24\cdot 0.851)\cdot 2} \approx 2$.