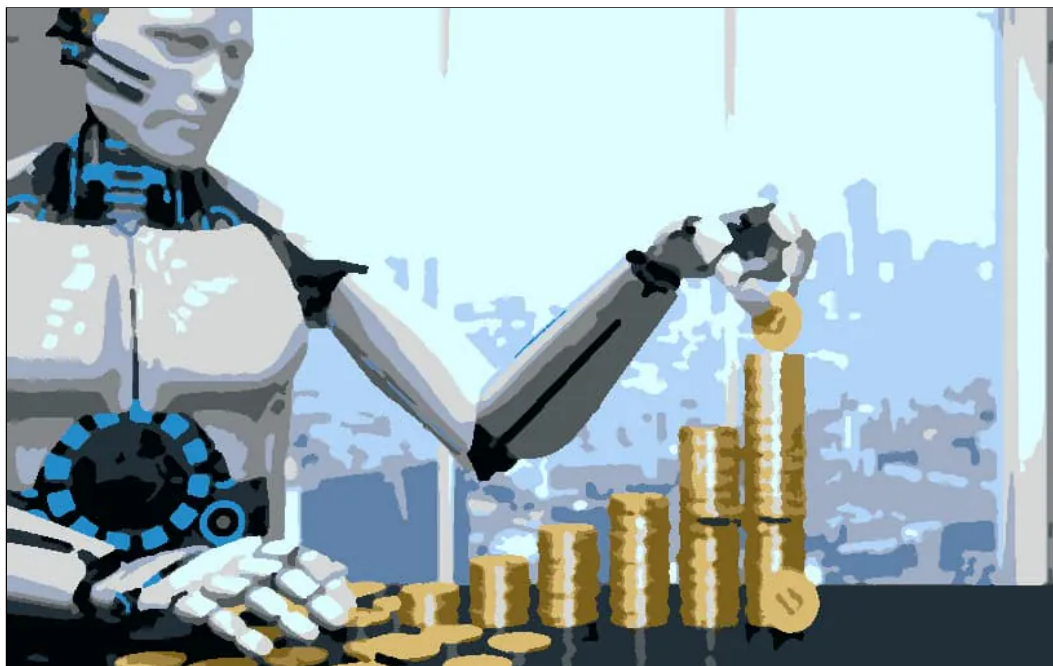


‘Simple’ AI Can Anticipate Bank Managers’ Loan Decisions to Over 95% Accuracy

Originally published at <https://www.unite.ai/simple-ai-can-anticipate-bank-managers-loan-decisions-to-over-95-accuracy/>



A new research project has found that the discretionary decisions made by human bank managers can be replicated by machine learning systems to an accuracy of more than 95%.

Using the same data available to bank managers in a privileged dataset, the best-performing algorithm in the test was a [Random Forest](#) implementation – a fairly simple approach that’s [twenty years old](#), but which still outperformed a neural network when attempting to mimic the behavior of human bank managers formulating final decisions about loans.

Results with the scorecard rating as predictor					
year	Samp. Size	Accuracy	AccuracyLower	AccuracyUpper	AccuracyPValue
2010	4165	0.912	0.903	0.920	0.000
2011	4149	0.907	0.898	0.916	0.000
2012	4161	0.921	0.912	0.929	0.000
2013	4161	0.931	0.923	0.938	0.000
2014	4155	0.945	0.938	0.952	0.000
2015	4162	0.932	0.924	0.940	0.000
2016	4164	0.937	0.930	0.944	0.000
2017	4171	0.961	0.955	0.967	0.000
2018	4161	0.964	0.958	0.970	0.000
Results without the scorecard rating as predictor					
2010	4165	0.895	0.885	0.904	0.000
2011	4149	0.883	0.873	0.893	0.000
2012	4161	0.904	0.895	0.913	0.000
2013	4161	0.913	0.905	0.922	0.000
2014	4155	0.933	0.925	0.940	0.000
2015	4162	0.921	0.912	0.929	0.000
2016	4164	0.915	0.906	0.923	0.000
2017	4171	0.945	0.937	0.952	0.000
2018	4161	0.958	0.952	0.964	0.000

The Random Forest algorithm, one of four put through their paces for the project, achieves high human-equivalent scoring vs. performance of bank managers, despite the relative simplicity of the algorithm. Source: *Managers versus Machines: Do Algorithms Replicate Human Intuition in Credit Ratings?*, <https://arxiv.org/pdf/2202.04218.pdf>

The researchers, who had access to a proprietary dataset of 37,449 loan ratings across 4,414 unique customers at ‘a large commercial bank’, suggest at various points in the preprint paper that the automated data analysis that managers are given to make their decision has now become so accurate that bank managers rarely deviate from it, potentially signifying that bank managers’ part in the loan approval process chiefly consists of retaining someone to fire in the event of a loan default.

The paper states:

‘From a practical perspective it is worth noting that our results may indicate that the bank could process loans faster and cheaper in the absence of human loan managers with very comparable results. While managers naturally perform a variety of tasks, it is hard to argue that they are essential for this particular task and a relatively simple algorithm can perform just as well.

'It is also important to note that with additional data and computational power these algorithms can be further improved as well.'

The [paper](#) is titled *Managers versus Machines: Do Algorithms Replicate Human Intuition in Credit Ratings?*, and comes from the Department of Economics and Department of Statistics at UoC Irvine and the Bank of Communications BBM in Brazil.

Robotic Human Behavior in Credit Rating Assessments

The results do not signify that machine learning systems are necessarily better at making decisions about loans and credit ratings, but rather that even algorithms now considered quite 'low-level' are capable of drawing the same conclusions as humans from the same data.

The report implicitly characterizes bank managers as a kind of 'meatware firewall' whose core remaining function is to raise the risk scores that the statistical and analytical scorecard system presents them with (a practice known in banking as 'notching').

'Over time it appears that managers are employing less discretion which might indicate the improved performance of or reliance on algorithmic means such as the scorecard.'

The researchers also noted:

'The results in this paper show that this particular task executed by highly skilled bank managers may in fact be easily replicated by relatively simple algorithms. The performance of these algorithms could be improved by fine tuning to account for differences across industries and of course could be easily extended to include additional goals such as incorporating considerations of fairness in lending practices or to promote other social goals.'



Spot the difference: the risk assessment of scorecard (automatic) ratings are statistically bumped up ('notched') by bank managers whose decisions were studied in the work – a replicable procedure.

Since the data suggests that bank managers do this in an almost algorithmic and predictable fashion, their adjustments are not that difficult to replicate. The process simply 'second guesses' the original scorecard data and adjusts the risk rating upward within predictable margins.

Method and Data

The project's stated intent was to anticipate what decisions bank managers would make, based on the scoring system and other variables available to them, rather than to develop innovative alternative systems designed to replace current loan application procedure frameworks.

The machine learning methods tested for the project were Multinomial Logistic LASSO ([MNL-LASSO](#)), [neural networks](#), and two implementations of [Classification and Regression Trees](#) (CART): Random Forest and [Gradient Boosting](#).

The project considered both the scorecard data for a real-world credit rating task, and its outcome, as known in the data. Scorecard rating is one of the oldest algorithmic practices, where key variables for the proposed loan are calculated into a risk matrix, often by means as simple as [logistic regression](#).

Results

MNL-LASSO performed the most poorly among the tested algorithms, successfully classifying just 53% of the loans, compared to the real-life manager in the cases evaluated.

The other three methods (with CART encompassing Random Forest and Gradient Boosting) all scored at least 90% in terms of accuracy and Root Mean Square Error ([RMSE](#)).

However, Random Forest's implementation of CART scored an impressive near-96%, followed closely by Gradient Boosting.

Results with the scorecard rating as predictor					
	RMSE	Accuracy	AccuracyLower	AccuracyUpper	AccuracyPValue
MNL LASSO	1.361	0.537	0.532	0.542	0.000
NN	0.447	0.934	0.931	0.936	0.000
Boosting	0.332	0.958	0.955	0.960	0.000
R. Forest	0.326	0.959	0.957	0.961	0.000

Results without the scorecard rating as predictor					
	RMSE	Accuracy	AccuracyLower	AccuracyUpper	AccuracyPValue
MNL Lasso	1.561	0.473	0.468	0.479	0.000
NN	0.514	0.905	0.902	0.908	0.000
Boosting	0.371	0.943	0.941	0.946	0.000
R. Forest	0.365	0.946	0.943	0.948	0.000

The table shows the prediction accuracy measured on a 5-fold cross-validation for the Multinomial Logistic LASSO, Neural Network, Gradient Boosting and Random Forest.

Even with the scorecard rating removed from the tests during ablation studies (lower table section), the algorithms achieve extraordinary performance in replicating human bank managers' discernment for credit rating.

Surprisingly, the researchers found that their implemented neural network only scored 93%, with a wider RMSE gap, producing risk values several notches away from the human-produced estimations.

The authors observe:

'[These] results don't indicate that one method outperforms the other one as far as an external metric of accuracy is concerned such as the objective default probability. It is quite possible that the Neural Network for example is best for that classification task.'

'Here the objective is only to replicate the choice of the human manager and for this task the Random Forest seems to outperform all other methods across the metrics investigated.'

The 5% that the system could not reproduce is accounted for, according to the researchers, by the heterogeneity of the industries covered. The authors note that 5% of managers account for nearly all these divergences, and believe that more elaborate systems could ultimately cover such use cases and close the shortfall.

Accountability Is Difficult to Automate

If borne out in subsequent related projects, the research suggests that the 'bank manager' role could be added to a growing cadre of once-powerful positions of authority and discernment that are being reduced to 'invigilator' status while the accuracy of comparable machine systems is tested over the long term; and undermines the [commonly-held position](#) that certain critical tasks cannot be automated.

However, the good news for bank managers would seem to be that, from a political point of view, the need for human accountability in critical social processes such as credit rating evaluation is likely to preserve their current roles – even if the actions of the roles should become completely reproducible by machine learning systems.

First published 18th February 2022.

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.

Portfolio site: martinanderson.ai

Contact: martin@martinanderson.ai



Discover More