Journal of Financial Risk Management, 2021, 10, 317-341
https://www.scirp.org/journal/jfrm
ISSN Online: 2167-9541
ISSN Print: 2167-9533

Harnessing Machine Learning Emerging Technology in Financial Investment Industry: Machine Learning Credit Rating Model Implementation

Chunlan Wang[1], Mahmut Rustem Sen[1], Bin Yao[1,2], Michal Certik[1], Koloina A. Randrianarivony[1]

[1]Large Financial International Institution, Washington D.C., USA
[2]Large Financial International Institution, Beijing, China
Email: cwangmz@gmail.com

How to cite this paper: Wang, C. L., Sen, M. R., Yao, B., Certik, M., & Randrianarivony, K. A. (2021). Harnessing Machine Learning Emerging Technology in Financial Investment Industry: Machine Learning Credit Rating Model Implementation. *Journal of Financial Risk Management, 10,* 317-341.
https://doi.org/10.4236/jfrm.2021.103019

Received: June 28, 2021
Accepted: September 15, 2021
Published: September 18, 2021

Abstract

Credit risk ratings consist of assessing the creditworthiness of the issuer and gauge the risks associated with buying its debt. Any delay in updating the credit risk ratings could have a severe impact on the financial system such as the financial crisis in 2008. This paper discusses a case that leverages emerging technology and breakthrough cognitive analytics in the financial industry. It specifically describes the design and implementation of a predictive modeling case based on the Machine Learning Approach and its application in credit risk forecasting and portfolio management. Using big data and Machine Learning, it is possible to improve credit risk analysis and forecasting by allowing the algorithms to search for patterns using large sets of data.

Keywords

Machine Learning, Credit Rating, Credit Portfolio Management, Random Forest Methodology, Machine Learning Training and Model Validation, Artificial Intelligence

1. Introduction

During the 2008 financial crisis, credit rating updates were delayed which could have prevented bigger financial damages. Typical reviews and restatement from the main rating agencies are on a quarterly or on annual basis. The updates are based on the financial data of the entity fed into statistical methods such as logistic regression and multivariate discriminant analysis. An analyst will judge

with a committee about the appropriate rating for the issuer's current state. These rating agencies have explored to leverage the rapid increase in the data availability and computing power added with Machine Learning (Bacham, 2017). Their focus was on small- and medium-sized borrowers in which they concluded that the Machine Learning models deliver a similar accuracy ratio as their other models, but they appeared more as a "black box". Application of the fourth industrial revolution to the banks' members of the Gulf Cooperation Council reveals that the superiority of the Machine Learning models stems from their ability to learn from the data set, and then not rely on any exogenous assumptions. Machine Learning is becoming transformative in financial systems. Superior performance is observed in the asset and option pricing with Machine Learning models (Gan et al., 2020). Bankruptcy predictions have used similar techniques (Barboza, 2017) and similarly in credit card risk management (Butaru et al., 2016). All these capabilities are focused on having a better financial system and on preventing another financial crisis.

Credit risk analysis has been always challenging in the financial industry since there are abundant factors which have various impacts on the solvency problem of corporates through distinguished channels. Given the strong power in data processing, data analyzing and forecasting, Machine Learning technology could add great value by bringing additional insights and improving analysis efficiency for credit risk analysis in business. Considering the low frequency and potential other limitations of traditional rating service exposed in the wake of a financial crisis (Stephen, 2011), it would be not sufficient simply to be dependent on external credit rating agencies. Meanwhile, due to the human resource constraint and potential limitless scope of parameters to be used, it would be very difficult to constantly conduct analyzing tasks for a significant amount of securities in the absence of a powerful and automatic process. In total, it would be naturally more reasonable to establish an in-house risk analysis model with the enriching toolbox to add broader and deeper insights and improve research efficiency for credit portfolio management and analysis. In this paper, our discussion would mainly focus on the Machine Learning model implemented for both the banking and corporate sectors.

The rest of this paper proceeds as follows: the Second Section focuses on the literature review and credit rating model discussion from both traditional and machine learning perspectives; the Third Section provides discussions about the implementation of the machine learning model. We first create the whole set-up by combining company financial statements with financial market data. Then we describe the prediction capabilities with a focus on using Random Forest. The predictions are compared with the published credit rating by the main rating agencies. The results show that we have a favorable comparison. The Final Section includes results discussion and their application, the policy, and Artificial Intelligence (AI) strategy implication for AI building in financial institutions, and further thinking on the relevant challenges and relevant potential solution exploration.

2. Evolution of the Credit Risk Modeling and Data

Financial institutions and financial industry practitioners use the credit rating model as a typical tool in various investment analyses and credit risk management. The first application of the Discriminant Analysis models was done back in 1968 in the Altman Z-score model. After its introduction, similar models utilizing linear and non-linear variable structures and different classification techniques, such as quadratic, logit, probit, and hazard model structures were introduced to model the outcome in terms of the probability of default based on the characteristics of the sample of firms used in the model. Other researchers have used discriminant analysis and logistic regression, researchers to conduct studies, and tests (Hillegeist et al., 2004), and (Chen, Chollete, & Ray, 2010).

Explorations using the AI ML learning method and more robust data set have been also conducted by researchers and financial initiations in recent years. A deep survey was done by the credit rating agency (Bachar & Zhao, 2017) itself in July 2017 to analyze the potential of the Machine Learning Techniques in Credit Modelling for medium-sized borrowers. They showed that Machine learning contributes significantly to credit risk modeling applications and delivers similar accuracy ratios to the more traditional benchmark models.

The analysis was then followed in (Gambacorta, Huang, Qiu, & Wang, 2019) by the Bank for International Settlements in December 2019 and the traditional loss and default approach was compared to the models using Machine Learning techniques. It has been shown that ML models perform even better than the standard ones and the legitimacy of the Machine learning approach for credit rating estimation was established by endogenous as well as an exogenous entity.

Credit risk analysis mainly belongs to pattern-recognition problems, classification algorithms can be used to classify the creditworthiness of companies (Kruppa, Schwarz, Arminger, & Ziegler, 2013; Pal, Kupka, Aneja, & Militky, 2016), and can be expected to improve traditional models based on simpler multivariate statistical techniques such as discriminant analysis and logistic regression. More data set including market data, macro data, and enhanced classification using trees may be helpful to improve the quality of crediting rating model in prediction accuracy and responsiveness to the market. In industry research studies, (Wang, Hao, Ma, & Jiang, 2011) used ensemble methods (bagging, boosting, and stacking) and found that bagging outperformed boosting for all credit databases they analyzed.

Financial institution and industry practitioners also conducted model testing and exploration in credit rating analysis using machine learning algorithms including Bagging, boosting, Random forest, SVM, and ANN. Random forest is considered as a powerful and more applicable classification engine in terms of the more robustness of result and relative effectiveness of implementation feasibility considering the complicated relationship of multiple dimension nature in financial world. (Moody, 2017) mentioned that a linear statistical model cannot fit this complex non-linear and non-monotonic behavior. The Random Forest

Model, a widely used machine learning method, is flexible enough to identify the hot spots because it is not limited to predicting. (Deloitte, 2018) mentioned that the Random Forest for credit risk models-Machine learning and Credit Risk is a suitable marriage. More big entities explored Machine Learning in the prediction of credit ratings, and it was identified that Random Forests models possess good performance results (Wallis, Kumar, & Gepp, 2019), (Morgan, 2017), and (McKinsey, 2017). The highest precision of the Random Forests algorithm for the credit rating estimation was also shown in (Lia, Mirzab, Rahatc, & Xiongd, 2020), where the precision remained robust for all the classes of the ratings. Even with a sophisticated modeling approach, the authors (Provenzano, Trifiro, Datteo, Giada, Jean, Riciputi, Le Pera, Spadaccino, Massaron, & Nordio, 2020) reached the remarkable accuracy of 95% with confusion matrix values as low as 82% and 85%.

This paper explores the use of specific sets of data training known for different asset classes to drive the market pricing by our internal practitioners and the standard Random Forest approach which prove to have comparable accuracy as the more complex and elaborate models. Intensive model dimension reduction testing has been incorporated embedding the actual business process. Specifically, we used ten-year horizon periodical testing. This provides a rather interesting insight into the exercise, where the outcome proves to be mainly driven by the choice of the data, rather than the sophistication of the model. This appears to have positive implications, as with the decreased complexity of the models and right selection of the data, the analysis as well as its outcome is easier to work with and communicate, more transparent and accessible to a wider audience and a processing speed is enhanced as the focus narrows on the data with the choice of the models optimized.

Several other studies have discussed the strengths and weaknesses of machine learning in different areas, such as (Subasi & Gursoy, 2010) and (de Menezes et al., 2016) and (Cano, 2019) in chemistry; (Bernard, Chang, Popescu, & Graf, 2017) in (Kim, Kang, & Kim, 2015) and (Gerlein, McGinnity, & Coleman, 2016) in finance. There are issues related to the lack of intuition and transparency which cause limitations in effective usage of the model result and model validation. In another aspect, the ML and AI-based research or analytics impact the traditional analysis behavior and can expand the application the scope of human's analysis activities. ML and AI are developed and aimed to enhance human intelligence instead of operating isolated or replacing humans. Its goals are considered to improve the human decision process and thereafter to expand the analytical activities by improving the actions taken in response to improved decisions. In "Man vs. Machine in Predicting Successful Entrepreneurs" (McKenzie & Sansone, 2017), they conducted an intensive study and compared the appropriate usage of the alternative approaches for entrepreneurs' success forecasting including judge-based, ad-hoc, and ML approaches. They indicated the following: "We only find machine learning to do somewhat better when it comes to

identifying the top tail of employment and profits, but an investor would do best just using the combination of man and machine in the ad-hoc models of economists rather than relying on human judges or machine learning chosen options".

2.1. Application of Machine Learning to Credit Rating

Machine Learning mainly includes three categories: supervised learning, unsupervised learning, and reinforcement learning. Supervised learning infers a function from training data samples, each of which is a pair consisting of an input object and output with labeled value. Supervised learning is very popular when the labeled variable is objective or credible. Regression, classification, decision tree or random forest, and support vector machine are typical methods in this category. Unsupervised learning, also known as self-organization, looks for previously undetected patterns in a data set with no pre-existing labels and with minimum human supervision. Clustering is a popular algorithm in this class. Reinforcement learning concerns how to act in an environment to maximize the notion of cumulative reward. In our model, we take the supervised learning method and choose random forest (RF) as a principal algorithm.

As previously mentioned, RF has been generally accepted as an effective and promising algorithm for various Machine Learning tasks. In the scope of the financial industry, especially in the domain of risk management, RF has been broadly incorporated into their Machine Learning research and prototype testing by many institutions to provide additional insights for financial forecasting and analysis. Some research and related Machine Learning testing have shown that RF could produce a better out-of-sample performance, particularly for highly nonlinear data. (Caruana & Niculescu-Mitzil, 2006) compared several different Machine Learning algorithms and find that ensembles of trees perform quite well. (Howard & Bowles, 2012) claimed that RF has been the most successful general-purpose algorithm in modern times. It is simple to understand and easy to apply. Meanwhile, key market counterparts see RF as powerful and applicable on risk rating projections. Earlier, (Moody, 2017) claimed that a linear statistical model cannot fit complex non-linear and non-monotonic behavior. RF is flexible enough to identify the hot spots for these kinds of problems. J.P. (Morgan, 2017) finds that RF has better performed than other models on a fixed income. (McKinsey, 2017) argued that overfitting is a typical concern about the Machine Learning model and one way to guard against overfitting is to deploy RF algorithm. Also, (Deloitte, 2018) claimed RF is a good choice for Machine Learning applications in credit risk and capable of identifying important features helping to reduce the time spent on data management.

Simply put, RF is an ensemble of multiple intentionally "weakened" decision trees. The single tree model was invented decades ago targeting nonlinear problems but is well known to be prone to overfitting. To curb this problem, bagging was invented later to independently train a bunch of trees over bootstrapped subsets of initial samples and then could to some extent reduce the variance of

prediction. After bagging, RF takes an important step further to incorporate a second level of randomness of subsampling attributes when optimizing each node split for each tree, and then could significantly relieve the overfitting problem. More impressively, given so many classifications in the forest, we could get the distribution of various classifications and the probability of classification less or more than some given threshold.

For the application of credit rating, Figure 1 shows how the algorithm analyzes features in the form of a decision tree and solves the rating assessment problem by assigning a rating to each node based on the values/presence of indicators/features. RF model builds multiple trees based on randomly bootstrapped data sets and does random feature selection under each tree. The learning results from each tree are then aggregated to generate a final rating for a specified security.

2.2. Data and Structure

As described previously, one of the novelties of this paper is to include additional data that captures the real dynamic of the financial market. We combine financial statements of the entities with economic, market and financial indicators. This section gives information about the model data, structure, and testing. According to certain criteria such as a minimum of balance outstanding, maturity and domicile based on, consideration of liquidity risk and other factors, we select about 100 from the Bloomberg Barclays Fixed Income Indices as the universe of rating model. For these corporates behind securities, the model involves the following data.

Dependent target variables: quarterly S&P credit ratings.

Dependent predictor variables:

- Financial statement data on Market Value, Profitability, Credit Quality and Liquidity, etc.

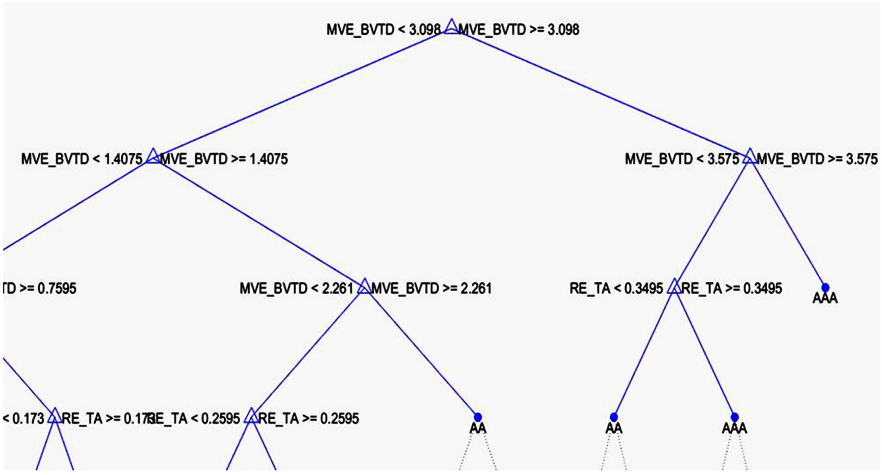


Figure 1. Decision tree (a subset of Random Forest). Source: MATLAB financial library: training exercise example.

- Economic, market and financial indicators including exchange rate, inflation, interest rate, financial market volatility (like VIX index), option adjusted spread (OAS)[1] across all major economies in the world.

In total, the number of predictors is up to 1439, during the period of 1997Q4 to 2019Q4. The response variable (dependent target variable) of quarterly data are obtained from the credit rating agencies. The credit data covers different sectors including banking, agency, corporate, covered bonds, etc. The dependent predictor features and variables are mainly from BB-rated company financial data and market data from vendor system.

3. Credit Rating Model: Machine Learning Implementation and Interactive Application

Based on the Machine Learning method architecture described above, a credit rating model and interactive application were developed and tested. Coverage includes counterparties in banking, corporate, agencies, and covered bond credit assets. An interactive interface has been also built for the application. This section will provide information about data, model training, and testing based on machine learning industry practice. We will also discuss the intensive Machine Learning (ML) model validation and comprehensive testing conducted. It will further provide insights through the customized ML credit analysis reports and discuss their appropriate context and limitation.

In addition to giving the information on ML model testing using general measures of model accuracy based on the Machine Learning process, we also conducted a specific study to backtest the capacity of the Machine Learning model to see how it can capture the dynamic change of the credit rating status in an expended longer historical period when there were rating changes. The business team thinks that this test and function will all add a lot of insights and value to the credit analysis process, greatly enhance the ML application capacity.

3.1. Model Workflows

In this section, we describe the flows and steps that we are taking for the Machine Learning Model.

Set up of Model Data, Machine Learning Training and Validation

Machine Learning has evolved as a more robust process according to the current industry practice. Data processing needs to be performed in a data utility effectively. ML model training and model validation needs to be conducted robustly to follow the model validation requirements using specific techniques to establish an ML model with a better quality and insights generation capacity. The diagram shown in **Figure 2** gives main information on the ML working flow which reflects a general ML building process widely followed by

[1]The Option-Adjusted Spread (OAS) is the measurement of the spread of a fixed-income security rate and the risk-free rate of return, which is then adjusted to consider an embedded option. Typically, an analyst uses Treasury yields for the risk-free rate. The spread is added to the fixed-income security price to make the risk-free bond price the same as the bond.

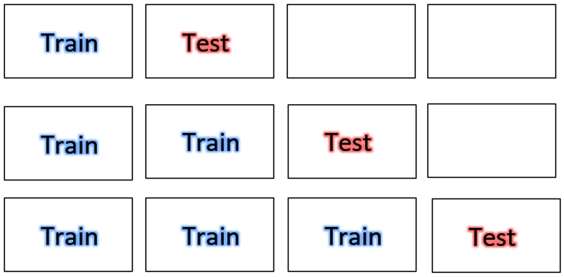


Figure 2. Machine learning model working flows diagram.

Industry practitioners (Please also see more information on Morgan, 2017).

We take the banking sector as a starting point to deploy the Machine Learning Credit Rating (MLCR) model. MLCR model has selected about 50 bank counterparties as the target universe. MLCR target variable is risk rating from S&P, and model inputs (predictors) include 1400 indicators of financial statement, the world economy, and financial market. The whole samples cover the period from March 1990 to March 2018 every quarter. We trained the model on the learning phase and the accuracy of the testing sample. These were performed in multiple stages with expanding the training set in stages and recalibrating the model parameters through new testing sets, to increase the model accuracy. The ultimate model accuracy is above 85%.

The importance scores give the information about what factors have the bigger impact on the target variable—in this case, S&P risk ratings. Using this information, we further performed dimension reduction to decrease the number of predictor variables to only the most important ones, to achieve higher transparency and efficiency of the model. We show in **Figure 3** the top 8 variables that really matter in the modeling process.

3.2. Model Training and Validation

Once the model structure has been set up as described in the part of the corporate sector model and data is collected, the next step is to train the model. But how we can know if the model is eligible or good enough for real application? To answer this question, we should see how the model performs in the test or so-called out-of-sample (OOS) performance, instead of just checking the performance in the in-sample data. If the model performs well in testing data, then we could say that the model is more robust and then it has the potential to scale up in business, otherwise, there would be many possibilities of overfitting problems when there is simply a high score for in-sample data. To this end, the first step is to split all data into two parts: one as in-sample data for model training and validation, the other as out-of-sample data for model testing. Normally, the next step is to further split the in-sample data into the training set and validation set. Given there are some hyperparameters needed to be preset before putting the model into training or fitting over training data sample, the reason for further splitting is to find the proper model out of many model candidates to be

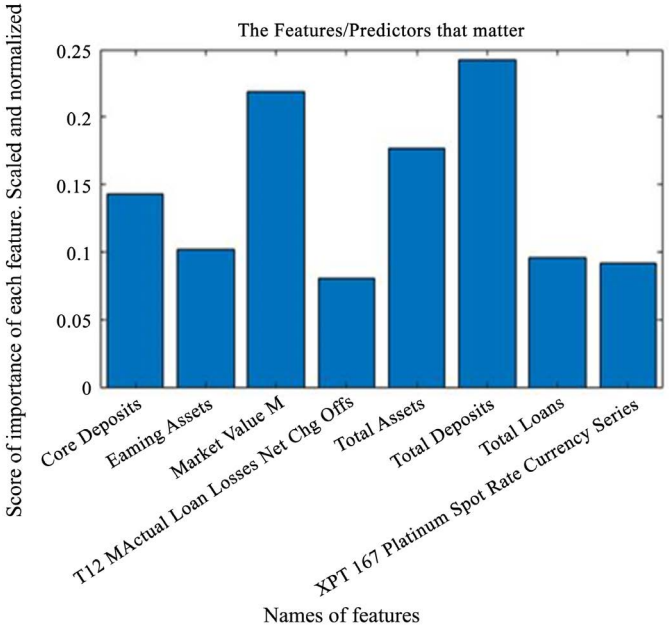


Figure 3. Most influential rating predictors.

tested on testing data and get the most reliable assessment of the performance of Machine Learning model. The loss function is of course rather intuitive, being represented by discrepancies between real and modeled credit ratings.

For our Machine Learning rating model of corporate bonds, 3952 samples during 1997Q4-2015Q4 are set as training and validation data, and the rest 1234 samples during 2016Q4-2019Q4 as testing data, which account for nearly 24% of the whole samples. Compared with the banking sector model (phase I model), the length of the testing period is significantly larger. The reason of expanding the testing period for the corporate sector model is that we want to focus on the dynamic tracking capacity in this stage rather than the static overall accuracy emphasized in the phase I model.

How to conduct model validation is often a major concern in Machine Learning practice. For non-time series data with samples typically independent, the random split is often the rule of thumb. While in the case of time series data, one way is to set a period time of data for training, and the data outside this period (normally in the aftermath of training time) for model validation. In our case, due to the very special characteristic of random forest algorithm and out-of-bag (OOB)[2] sample could be a good substitute for validation data (Breiman, 2001), which means that there is no need to spare some extra proportion of data as validation data to do model validation, as usual, the OOB sample could be used directly as the validation set. So, then we could choose the most credible model directly according to the OOB performance. And this way we can use more data for training of the model.

[2]Each kth tree is constructed using a different k-th bootstrap sample randomly chosen from the original data, so about one-third of the cases are left out of the bootstrap sample and not used in the construction of the kth tree.

The core challenge in the Machine Learning process or modeling is in capturing the relevant data. When all the data is assembled, the remaining part is to choose the software to perform all the calculations. We develop a Machine Learning model using MATLAB language and environment. There are some built-in functions in MATLAB and it's very convenient to call these functions when doing coding work. For the RF algorithm, two important hyperparameters are having a non-trivial impact on model performance: one is the number of trees (NT), the other is the number of subsampling predictors (NP) randomly selected when optimizing each node split. Even though there is a thumb rule[3] in practice, we still wanted to run the model using the different number of NP parameters to determine the most accurate choice. For this, we have done dozens of experiments with different combinations of NT and NP, and Table 1 shows a portion of this work. According to these experiments, we found the accuracy of the model decreases its sensitivity to the NT or NP beyond a certain threshold.

Based on the sensitivity testing, we can see that the more the number of random features, the higher the model accuracy is, but the increment of the accuracy is slowing and the overfitting problem starts to be more significant at the high number of random features. We can see the start of the plateau of the accuracy scores at 80% when the prediction features NP starts at 200. The increase of the accuracy score is slower at the range above 200; while when N.P. increased from 100 to 200, the accuracy scores were increased from 63.7% to 81%. Finally, as a trade-off between performance and computation time, and to avoid the potential overfitting problem, we selected the combination of NT = 500 and NP = 400 as reasonable hyperparameters of the model. For this model as shown in Table 1, we have an RF model accuracy OOB score up to 0.881. Based on this trained model, the next step would be to put it on the expanded testing period. And like the banking sector model, the corporate model could also give probability distributions of each rating. By simply summing up the probability of each rating, we were able to get probabilities of downgrading below a certain level. Table 2 shows a couple of examples for various Bond entities starting with the predicted rating and their probabilities of being downgraded one notch or below one notch.

4. Results

In this section, we report on the typical output of the model. Instead of giving a single number as an output, Section 4.1 shows that the model attaches a

Table 1. Part of experiments for NT and NP.

		Number of random features				
		50	100	200	300	400
500 trees	**Model accuracy (OOB score)**	46.6%	63.7%	81.0%	86.1%	88.1%

[3]For example, (Breiman, 2001) points out that the optimal number of randomly chose features should be the first integer less than $\log_2(M+1)$, where M is the number of predictors.

Table 2. Example of MLCR probability of downgrading for various entities.

Company	MLCR Rating	Probability of Downgrading to A−	Probability of Downgrading Below A−
Bond 1	A+	12.9%	9.1%
Bond 2	AA−	6.3%	22.3%
Bond 3	A	17.2%	38.8%
Bond 4	AA+	13.7%	11.1%
Bond 5	A	8.8%	34.8%
Bond 6	A	6.5%	25.3%
Bond 7	A	13.5%	47.3%
Bond 8	A	10.9%	6.6%
Bond 9	A	14.3%	23.7%
Bond 10	A	14.6%	29.1%

probability to each forecasted credit rating. Additionally, the model spits out a report that can be used for decision-makers or any strategical position vis-à-vis the holding positions and their rating status. Section 4.2 describes a comparison with financial market data readily available as a comparison. Finally, Section 4.3 investigates the performance of the model. Overall, we observe that many of our results coincide well with the public rating available for the entities into consideration.

4.1. Output of the Models

Credit Rating with Probability Distribution

The model not only provides a credit rating but also generates the probabilities of each forecasted credit rating. Figure 4 gives MLCR model result with credit rating probabilities for selected counterparties.

Machine Learning Credit Rating Analysis Report

Based on the rating distribution coming out of the MLCR model, we could also see those ratings with relatively high possibility and then compare them with the rating given by agencies. Table 3 shows several real forecasting of rating and the comparison with S&P/Moody's. Each row is for a specific company whose name is hidden. R is the rating given by our model and P is the corresponding possibility. For the counterparties, using the trained model to run the new data for the quarters in June, Sep, Dec of 2018, and March, June 2019, the results displayed in the format of Table 3 have been generated. About 70% of the counterparties whose ML rating and Agency Rating is the same notch; around 20% of the counterparties; there is a difference of one notch between whose ratings given by ML and Agency rating; around 10% of them, the difference is the two-three notch.

4.2. Direct Comparison with Financial Market Data

If the two ratings are much diverse, then one question comes out naturally:

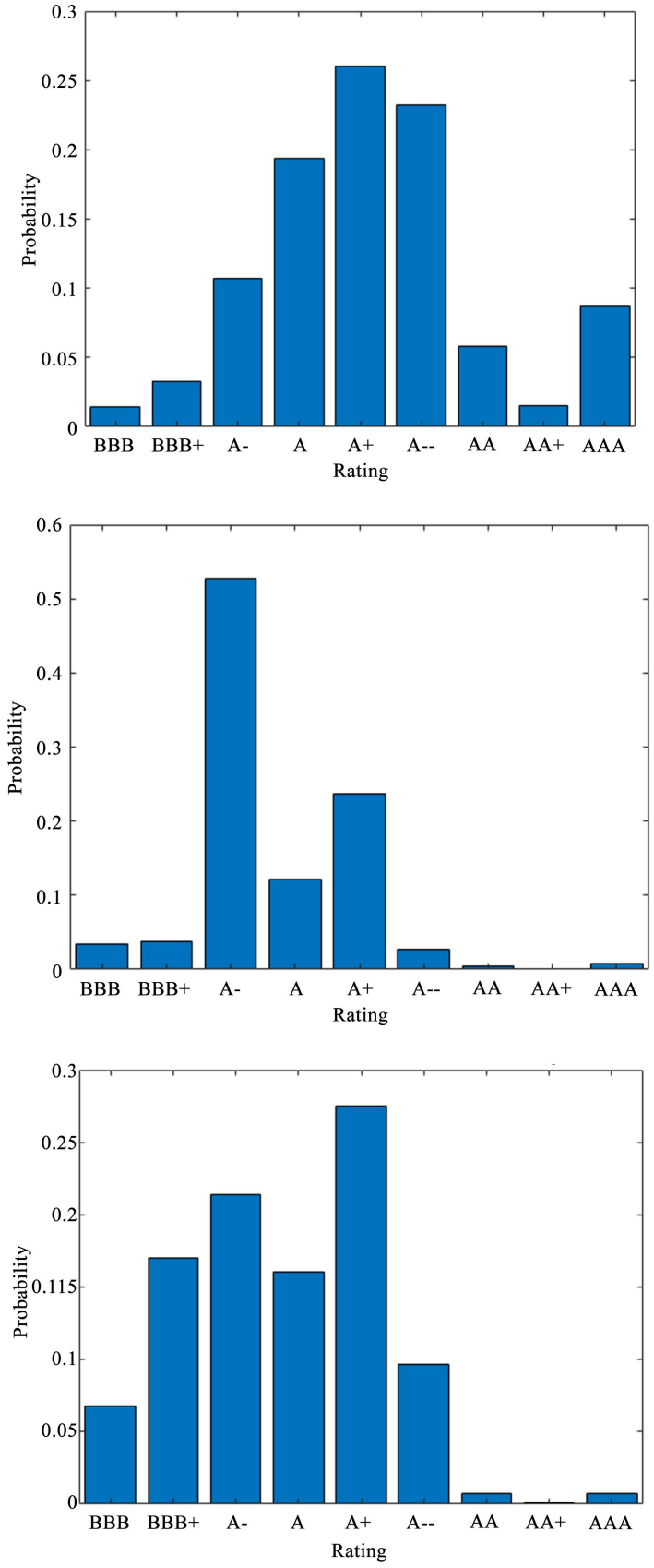


Figure 4. MLCR model result: credit rating distribution.

Table 3. ML credit rating analysis report-customized for business credit analysis.

June								September								December							
ML Ratings (R1, R2, R3) & Probs. (P1, P2, P3)						S&P	Moody's	ML Ratings (R1, R2, R3) & Probs. (P1, P2, P3)						S&P	Moody's	ML Ratings (R1, R2, R3) & Probs. (P1, P2, P3)						S&P	Moody's
R1	P1	R2	P2	R3	P3			R1	P1	R2	P2	R3	P3			R1	P1	R2	P2	R3	P3		
A+	0.30	A	0.24	AA−	0.23	A	A1	A+	0.30	AA−	0.28	A	0.26	A	Aa3	A+	0.30	A	0.26	AA−	0.23	A	Aa3
A	0.32	A+	0.17			A+	A1	A	0.33	A+	0.21			A+	A1	A	0.31	A+	0.20	A−	0.17	A+	A1
A	0.30	A+	0.27	A−	0.16	A	A1	A	0.36	A+	0.22	A−	0.18	A	A1	A	0.40	A−	0.20	A+	0.20	A	A1
A	0.30	A+	0.27	A−	0.16	A	A1	A	0.34	A+	0.25	A−	0.18	A	A1	A	0.35	A+	0.25	A−	0.18	A	A1
A	0.23	A+	0.22	A−	0.18	A		A	0.28	A+	0.23	A−	0.21	A		A	0.30	A+	0.21	A+	0.21	A	
A	0.23	A−	0.21	A+	0.17	A	A1	A	0.24	A−	0.23	A+	0.17	A	A1	A	0.26	A−	0.17	AA−	0.16	A	A1
A+	0.28	A	0.22	AA−	0.15	A	Aa3	A+	0.28	A	0.23	AA−	0.16	A	Aa3	A+	0.28	A	0.19	AA−	0.17	A	Aa3
A−	0.41	A+	0.18	BBB+	0.16	A−	A3	A−	0.34	A+	0.21	A	0.16	A−	A3	A−	0.38	A+	0.20			A−	A2
A+	0.25	BBB+	0.20	BBB	0.17	BBB	Baa3	A	0.18	A−	0.17	BBB+	0.16	BBB	Baa3	A	0.17	BBB	0.16	BBB+	0.16	BBB	Baa3
A−	0.32	A+	0.22	A	0.16	A	A1	A+	0.26	A−	0.25	A	0.20	A	A1	A−	0.30	A+	0.27	AA−	0.16	A	A1
A+	0.32	A	0.21	AA−	0.21		A1	A+	0.31	A	0.24	AA−	0.18		Aa3	A+	0.31	A	0.24	AA−	0.17		Aa3
A	0.23	A−	0.21	A+	0.16	A	Aa3	A	0.28	A−	0.20	A+	0.16	A	Aa3	A	0.26	A−	0.24			A	Aa3
A−	0.37	BBB+	0.17	A+	0.16	BBB+	Baa1	A−	0.40	A	0.21	A+	0.16	BBB+	Baa1	A−	0.36	A	0.21	A+	0.16	BBB+	A3
A	0.24	BBB+	0.21	A−	0.20	BBB+	Baa2	A	0.24	A+	0.23	BBB+	0.20	BBB+	Baa2	A	0.24	A+	0.23	BBB+	0.18	BBB+	Baa2
A−	0.24	A+	0.21	A	0.18	BBB+	Baa2	A+	0.25	A−	0.24	BBB+	0.17	BBB+	A3	A+	0.22	A	0.21	A−	0.20	BBB+	A3
A−	0.30	A+	0.21	BBB+	0.15		Aa2	A+	0.24	BBB+	0.21	A−	0.20		Aa2	A−	0.24	A+	0.21	BBB+	0.18		Aa2
A−	0.24	A+	0.18	A	0.17	A	A1	A−	0.23	A	0.23	A+	0.19	A	A1	A	0.21	A−	0.20	A+	0.16	A	A1
A	0.24	A+	0.23	BBB+	0.17	A	A2	A+	0.26	A	0.24	BBB+	0.17	A	A2	A	0.24	A+	0.23	A−	0.18	A	A2
A−	0.45	A+	0.23			A−	A3	A−	0.38	A+	0.25			A−	A3	A−	0.38	A+	0.21			A−	A2
A+	0.27	BBB+	0.26	A	0.15	BBB+	A3	A+	0.26	BBB+	0.24	A	0.20	BBB+	A3	A+	0.24	BBB+	0.20	A	0.19	BBB+	A3
A−	0.24	AA−	0.23	A+	0.21	AA−	Aa1	A+	0.25	AA−	0.24	A	0.21	AA−	Aa1	AA−	0.24	A+	0.23	A−	0.20	AA−	Aa1
A+	0.27	BBB+	0.21	A−	0.16	BBB+	A2	A+	0.27	BBB+	0.24	A−	0.19	BBB+	A2	A+	0.25	BBB+	0.24	A−	0.15	BBB+	A2
A+	0.22	A	0.19	A−	0.17	A−		A	0.25	A+	0.22	AA−	0.16	A−		A	0.25	A+	0.22	A−	0.18	A−	
A−	0.35	A+	0.26			AA−	Aa1	A+	0.28	A−	0.24	A	0.17	AA−	Aa1	A+	0.25	A−	0.24	A	0.15	AA−	Aa1
A−	0.45	A+	0.23			A−	A2	A−	0.35	A+	0.25	A	0.15	A−	A2	A−	0.25	A	0.15	A+	0.20	A−	A2

which one is right or more reasonable, or how this difference would provide insight for business? When facing this situation, we use Credit Default Swap (CDS) data to help to make a reasonable assessment. For example, in the case of MLCR rating is below S&P rating, there is a possibility of MLCR model to underestimate or S&P to overestimate the credit qualification. Then we could see the CDS of this security and compare it with the average of CDS of securities with the same rating. If this security's CDS is higher than the average in S&P rating average, which means this specific security has higher risk than the average level of the group with the same risk rating, which would mean that there is

a relatively high possibility that MLCR rating would be more reasonable because of its consistency with CDS market pricing. After the analysis, we could get more and critical information for trading decisions, sometimes giving hopefully an early signal for business. Figure 5 shows an example comparing actual CDS (as of Dec 2018) for counterparts that have different ML-rating than S&P to average CDS. MLR rating prediction is aligned with what CDS implies in 9 out of 11 cases where MLCRs differ.

RF algorithm gets final rating maybe with very much different possibility for distinguished securities. Given the nature of market uncertainty, this distribution of rating given by the MLCR model would provide useful information for risk assessment. This will help as an early warning in case we have a high probability shown in the downgrades. Furthermore, this can be used by the portfolio holders to make possible changes in their holdings before the rating migration happens. It can also be used for internal stress testing on the probability of default of the holdings.

Most-likely MLR Rating	S & P	CDS - Counterpart	CDS - S & P Rating Average
A+	BBB	76	125
A–	BBB	82	125
A+	BBB	150	125
A+	BBB	68	125
A	BBB	107	125
A+	A–	76	87
A+	BBB+	208	125
A–	A	84	82
A	A–	76	87
A+	AA–	95	73
A+	AA–	76	73
A	A	46	82
A	A	46	82
A	A	68	82
A	A	75	82
A–	Aa2	95	73
A	A	76	82
A	A	93	82
A–	A–	68	87
AA–	AA–	95	73
A–	A–	71	87

Figure 5. Comparison of MLR rating with S&P rating.

Capacity of Capturing Rating Change in a Relatively Longer Period

Additionally, we could also see the comparison over time. As shown in Figure 6, we are showing the evolution of the ratings for various counterparties overlaid on top of the results by the Machine Learning program. The four charts below shows the lower range and upper range of the credit rating from the Machine Learning model (colored area in blue) and the actual S&P rating (blue solid line) over an historical period (June 2018, Sep 2018, Dec 2018 and March 2019). It is shown that the historical actual rating is mainly within the Machine Learning rating range and they performed in a similar behavior pattern. According to further statistics assessment, about 70% of the probability, the difference of Machine Learning model rating and S&P agency rating is less than one notch, and with about 20% of probability, the difference is around 1 to 2 notches.

4.3. Model Results and Validation with Dynamic Testing

Typical forecast performance is assessed with the prediction loss function which penalized false positives and false negatives. To improve machine learning model quality and to enhance the model application impact in actual use, customized loss function during the data training process is also considered as an effective way. For example, in credit rating classification, higher attention may be needed to be given for false negative then false positive from an aspect of monitoring credit risk downgrade risk. One possible way to customize the loss function may

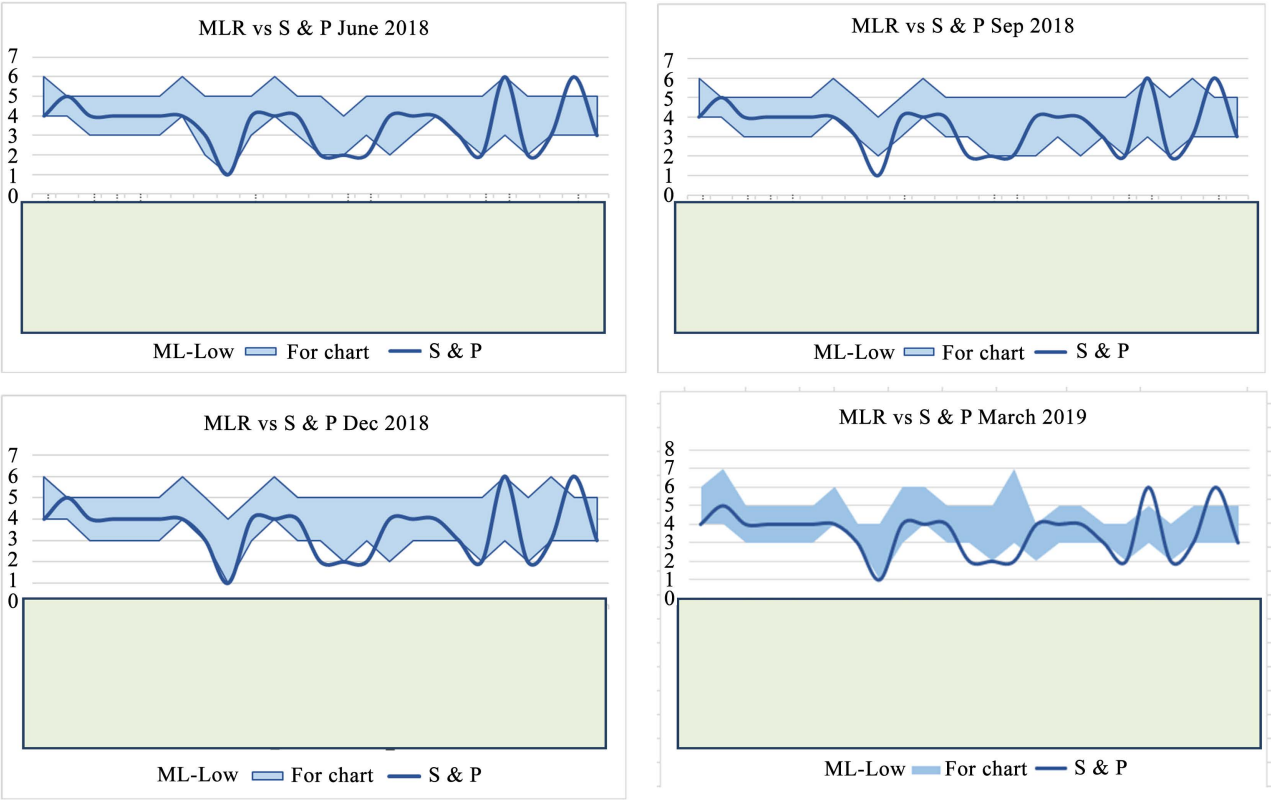


Figure 6. Credit rating comparison over time.

be using the cost-dependent loss function to include the highest expected loss elements in the loss function. In different model engine libraries, some customized elements may be input to the loss function of logistic classification functions or neural network machine learning function. The model described in this article was implemented using MATLAB Machine Toolbox Library. Some specifications can be selected with the machine learning engine function and their engine configurations. The main goal of the task in Credit Rating Modeling is to have the classification into rating categories (9 in our case). Our model is implemented using the general loss function without customization. Below are described our model training and model validating process. In the later stage, the model may be further enhanced by adding a customized loss function to include the impact of the different forecasting errors from being upgraded or downgraded. The "Predicting Food Crisis" paper (Johannes, Chamorro, Kraay, Spencer, & Wang, 2020) implemented detailed customization of the prediction loss function in such a way to minimize prediction loss functions that penalize false positives and false negatives. It was critical to the accuracy of the food crisis model and it also added great value. To test our model, seeing if it could be generalized in samples other than the training set, we should see how the model performs in the out-of-sample data, so-called model backtesting. For this application, we have done two different kinds of model testing. One is the overall accuracy of the model on the testing set. The other is to see if the model could capture the change of rating for each security, saying how is the dynamic ability of the model to capture the change of risk profile. As shown in **Table 1**, the OOB score on the training set is 88.1% and the accuracy score on the testing set is 47%. We should keep in mind that, as mentioned before, given that expanding the testing period aims to check the dynamic capacity, the overall accuracy would naturally be to some extent lower than that in the phase I model because the prediction step is longer. If we set the same length of the testing period as the Phase I model, the accuracy score would be very close.

After the overall assessment, we drill down into each corporate whose rating changed at least once over the testing period (2016Q4-2019Q4). Such assessment can be done if the model would be deployed on each security separately and this process is more time consuming given the predictors amount to over 1400. For this, we streamlined the model to speed up the testing process. Also, more data is not always better and can increase forecast errors even when using dimensionality reduction techniques (Boivin & Ng, 2006). Hence, we derived the importance for each predictor and selected the most important predictors as the new predictors to build the reduced-form model. We conducted experiments for each set of several most important predictors, like top 100, 200, and 300 important predictors, and for each setting, we repeat a similar experiment to find the best combination of NT and NP. Based on all these experiments, we take the top 300 important predictors (sum of importance account for 70.4% of total predictors) and choose the hyperparameters of NT = 200/NP = 90.

We deployed the streamlined model above onto the testing data. Due to some complexity of rating dynamic change, we roughly group the pattern of real and predicted dynamic into four different styles or categorizations. "Style I" represents that the model could react nearly synchronously with S&P rating at the time when S&P rating varies and the difference between two ratings eventually is no more than one notch. "Style II" is that the model could not react nearly synchronously with S&P rating at the time when S&P rating vary but the difference between two ratings eventually is still no more than one notch. "Style III" is that the model could react nearly synchronously with S&P rating at the time when S&P rating vary but the difference between two ratings eventually is more than one notch. "Style IV" is that the model could neither react nearly synchronously with S&P rating at the time when S&P rating vary nor the difference between two ratings eventually is no more than one notch. Within the model universe, there are 30 or so securities whose rating changed at least once during the testing period. In total, it corresponds to nearly 55% of the samples belong to Style I, II, and III. For the rest 45%, it also shows some close relationship between model rating and S&P rating. Typical comparison for each Style is shown below in Figure 7.

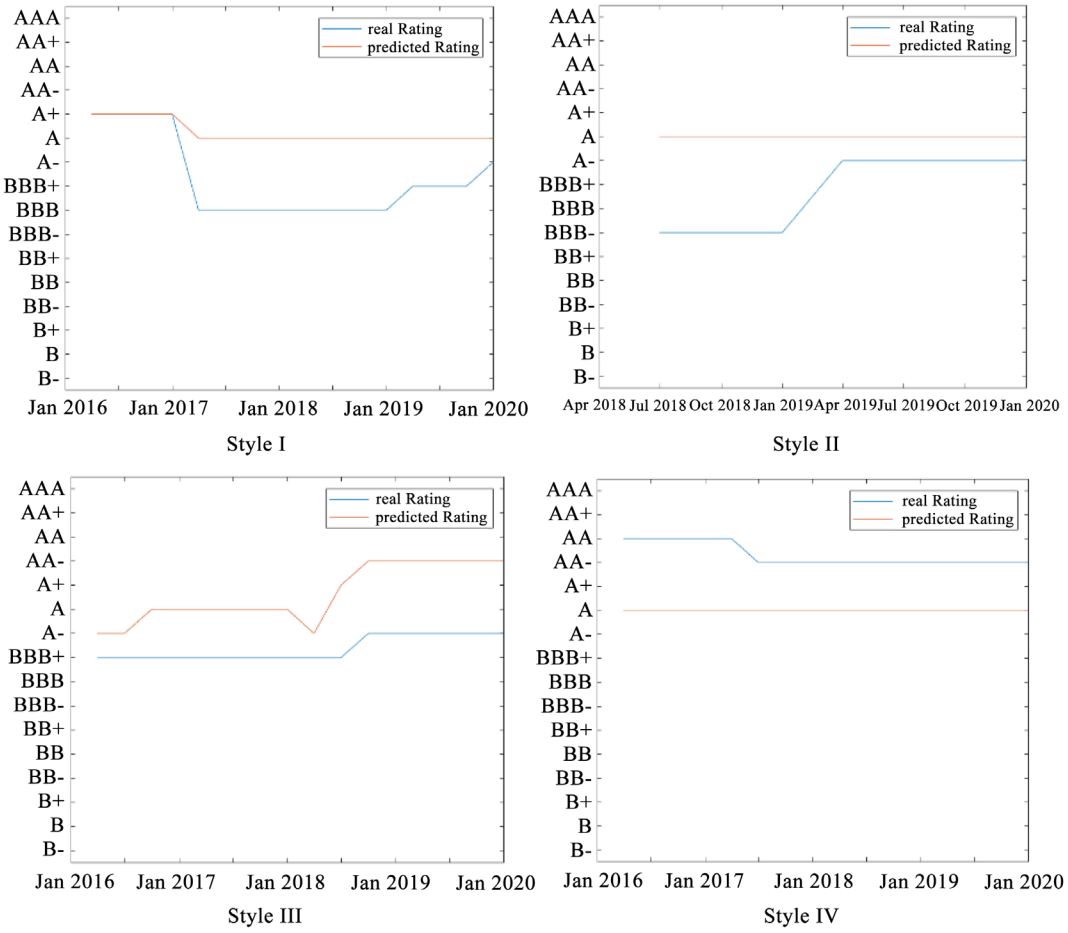


Figure 7. Example of the four styles of patterns between two rating outcomes.

5. Discussion and Conclusion

The emerging technologies and breakthrough cognitive analytics have evolved and been considered as the driving forces to transform the business environment across industries. Embracing disruptive technologies and building cognitive analytics are considered among the necessary priorities for business transformation in the financial industry.

This paper has discussed a case of harnessing emerging technology and breakthrough cognitive analytics in credit analysis and portfolio management. It specifically described the whole process of building a Machine Learning Rating Model Application. It has shared the experience and lessons learned during the process of addressing the industry challenges in harnessing Machine Learning technique through a tangible case of credit rating estimation to enable and accelerate the exploration of new technologies in the business process based on industrialized Machine Learning practice and robust implementation approach on Artificial Intelligence and Machine Learning building.

According to recent industry studies and lessons learned, there are three essential elements in the implementation and deployment of AI technology in actual business processes: scope, scale, and speed. First, we need to begin with narrowed scope and consider these areas as a starting point as they are expected to generate marginal added value across different business lines. Then, they may be scaled up and rolled out too many processes within the group based on the experience and lessons learned. It is essential to take an agile approach and build a Minimum Viable Product (MVP) that will evolve as an industry practice. Later, all stakeholders may benefit from the initial experiments by reviewing the knowledge and expertise developed in this process including data analytics, predictive modeling, and AI techniques. These experiments would ultimately help accelerate the development and deployment of these technologies in the business process.

Integrated quantitative architecture emerging technology needs to be adopted. The traditional architecture will not work, or it will not be enough. Machine Learning has already deeply penetrated financial markets, but it needs data. Machine Learning does the dirty work and heavy lifting of getting value out of data. Machine learning may help reduce the opportunity cost of using alternative data by improving and automating the data gathering, processing, and cleaning procedures. The existing source of data can also be made cheaper and more effective by these improvements.

Factors that impact the success of implementation include integrating a new generation of quantitative architecture emerging technology based on industry practice needs with Data Engineering, Data Analytics and Advanced Predictive Quantitative Analytics.

The results of this study were clearly encouraging in further efforts to follow the path that has been taken in deploying disruptive technologies. The ultimate driver should be clear unless one wants to lose the competitive advantage. The initial application has been built, tested, and implemented and many lessons

have been learned. The exercise itself has opened many horizons and now more than ever it is clear what might be the next steps taken. From expanding on the work already done to leveraging on the latest research discoveries, one thing is clear—this path is surely worth following.

The main objective of this paper was to share the crucial application of Machine Learning in Credit Risk Rating Modelling from the practitioner's perspective. Theoretical coverage and implications of the further research and enhancement of the models and their parameters could be investigated and explored further and described in more detail. Underlying data could also be extended, and different dimension reduction techniques might be deployed.

Therefore, there is still enough space to further increase the precision of the models which we see as the main goal, not only in Machine Learning application to Credit Risk Modelling but also in any other fields where these methods could be greatly utilized.

Acknowledgements

The authors are very grateful to:

Paul Snaith, Ivan Zelenko for their leading, great support and valuable guidance to this initiative from Treasury and Pension Endowment.

Lakshmi Shyam-Sunder, Arbind Jha, Wayne Austin Schwartz, Fred Haddad and Andrea Foresti for their great support, discussions, and guidance from CROMC.

Attila Juhasz, Henry Hua Wan, Ramesh Ramiah, Walid Nasrallah, Grigor Sargsyan, Alisa Senderovic for their strong contributions and constructive business domain discussions from Investment Management Department at Treasury; Ghislain Martial Yanou, Daniel Ricardo Vela Baron for their contributions during model review discussion and validation from quantitative team at Treasury.

Thanks to the following colleagues for providing constructive discussions in Treasury, Pension Endowment Initiative group discussion, also in Chief Risk Officer unit knowledge collaboration: Wendy Mendes, Benjamin David Whitcher, Eric Bouye, Shengting Pan, Andrea Dore, Huy-Long Le, Yunjung Ha, Amit Bajaj, Lei Cao, Isabel M. Dai, Wei Hu, Thumpasery J. George, Ning Wang, Natan Goldberger, Joel Kouassi Niamien, Artan Ajazaj, Tao Wang, Joey Hyun Joon Lee, Kavitha Subramanian; Insightful discussion from David Jamieson Bolder, Ying Jiu Xu, Shijie Shi; Stela Mocan, Pratheep Ponraj, Swamy Kiran, Casey James Traylor, Venkata Chandrakanth Burra, Long Yang, Vivek Kulbhushan Sharma; Kartheek Kandikuppa, Ru-Man Li; Mohammad Shahbazi; Thanks to Munier Salem from J.P. Morgan, Quantitative AI, ML strategy team for the insightful knowledge discussions and joint presentation on AI/ML quantitative model building.

Disclaimer

The mention of specific companies or certain products does not imply that they

are endorsed or recommended by the large financial institution in reference to others of a similar nature that is not mentioned.

All errors, omissions that may appear in this work are the authors' sole responsibility.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

Bacham, D., & Zhao, J. (2017). Machine Learning: Challenges, Lessons, and Opportunities in Credit Modeling. Moody's Analytics Risk Perspectives.

Bachar, D., & Zhao, J. (2017). Machine Learning: Challenges, Lessons, and Opportunities in Credit Risk Modeling. *Moody's Analytics Risk Perspectives: Managing Disruption, IX*.

Barboza, F., Kimura, H., & Altman, E. (2017). Machine Learning Models and Bankruptcy Prediction. *Expert Systems with Applications, 83,* 405-417. https://doi.org/10.1016/j.eswa.2017.04.006

Bernard, J., Chang, T.-W., Popescu, E., & Graf, S. (2017). Learning Style Identifier: Improving the Precision of Learning Style Identification through Computational Intelligence Algorithms. *Expert Systems with Applications, 75,* 94-108. https://doi.org/10.1016/j.eswa.2017.01.021

Boivin, J., & Ng, S. (2006). Are More Data Always Better for Factor Analysis? *Journal of Econometrics, 132,* 169-194. https://doi.org/10.1016/j.jeconom.2005.01.027

Breiman, L. (2001). Random Forest. Machine Learning, 45, 5-32. https://doi.org/10.1023/A:1010933404324

Butaru, F., Chen, Q., Clark, B., Das, S., Lo, A. W., & Siddique, A. (2016). Risk and Risk Management in the Credit Card Industry. *Journal of Banking & Finance, 72,* 218-239. https://doi.org/10.1016/j.jbankfin.2016.07.015

Cano, I. (2019). *Optimizing Distributed Systems Using Machine Learning.* Thesis, University of Washington.

Caruana, R., & Niculescu-Mizil, A. (2006). An Empirical Comparison of Supervised Learning Algorithms. *Proceedings of the 23rd international Conference on Machine Learning*, June 2006, Pittsburgh, 161-168. https://doi.org/10.1145/1143844.1143865

Chen, J., Chollete, L., & Ray, R. (2010). Financial Distress and Idiosyncratic Volatility: An Empirical Investigation. *Journal of Financial Markets, 13,* 249-267. https://doi.org/10.1016/j.finmar.2009.10.003

De Menezes, F., Liska, G. R., Cirillo, M., & Vivanco, M. (2016). Data Classification with Binary Response through the Boosting Algorithm and Logistic Regression. *Expert Systems with Applications, 69,* 62-73. https://doi.org/10.1016/j.eswa.2016.08.014

Deloitte (2018). *Using Random Forest for Credit Risk Models-Machine Learning and Credit Risk: A Suitable Marriage.* Deloitte Risk.

Gambacorta, L., Huang, Y., Qiu, H., & Wang, J. (2019). *How Do Machine Learning and Non-Traditional Data Affect Credit Scoring? New Evidence from a Chinese Fintech Firm.* BIS Working Papers.

Gan, L., Wang, H., & Yang, Z. (2020). Machine Learning Solutions to Challenges in Finance: An Application to the Pricing of Financial Products. *Technology Forecasting*

& Social Change, 153, Article ID: 119928.
https://doi.org/10.1016/j.techfore.2020.119928

Gerlain, E., McGinnity, T., & Coleman, S. (2016). Evaluating Machine Learning Classification for Financial Trading: An Empirical Approach. *Expert Systems with Applications, 54,* 193-207. https://doi.org/10.1016/j.eswa.2016.01.018

Hillegeist, S. A., Keating, E. K., Cram, D. P. et al. (2004) Assessing the Probability of Bankruptcy. Review of Accounting Studies, 9, 5-34.
https://doi.org/10.1023/B:RAST.0000013627.90884.b7

Howard, J., & Bowles, M. (2012). The Two Most Important Algorithms in Predictive Modeling Today. *Strata Conference Presentation*, New York, February 28 2012.

Johannes P. A., Chamorro, A., Kraay, A., Spencer, P., & Wang, D. (2020). *Predicting Food Crises.* The World Bank Paper.

Kim, M.-J., Kang, D.-K., & Kim, H. (2015). Geometric Mean Based Boosting Algorithm with Over-Sampling to Resolve Data Imbalance Problem for Bankruptcy Prediction. Expert Systems with Applications, 42, 1074-1082.
https://doi.org/10.1016/j.eswa.2014.08.025

Kruppa, J., Schwarz, A., Arminger, G., & Ziegler, A. (2013). Consumer Credit Risk: Individual Probability Estimates Using Machine Learning. *Expert Systems with Applications, 40,* 5125-5131. https://doi.org/10.1016/j.eswa.2013.03.019

Lia, J., Mirzab, N., Rahatc, B., & Xiongd, D. (2020). Machine Learning and Credit Ratings Prediction in the Age of Fourth Industrial Revolution. *Technological Forecasting & Social Change, 161,* Article ID: 120309. https://doi.org/10.1016/j.techfore.2020.120309

McKenzie, D., & Sansone, D. (2017). *Man vs. Machine in Predicting Successful Entrepreneurs.* The World Bank Policy Paper. https://doi.org/10.1596/1813-9450-8271

McKinsey (2017). *How Machine-Learning Models Can Help Banks Capture More Value.* Mckinsey Digital.

Moody (2017). Machine Learning: Challenges, Lessons, and Opportunities in Credit Risk Modeling. *Moody's Analytics Risk Perspectives, IX.*

Morgan, J. P. (2017). *Machine Learning in Fixed Income.* AI and Machine Learning Report.

Pal, R., Kupka, K., Aneja, A., & Militky, J. (2016). Business Health Characterization: A Hybrid Regression and Support Vector Machine Analysis. *Expert Systems with Applications, 49,* 48-59. https://doi.org/10.1016/j.eswa.2015.11.027

Provenzano, R., Trifiro, D., Datteo, A., Giada, L., Jean, N., Riciputi, A., Le Pera, G., Spadaccino, M., Massaron, L., & Nordio, C. (2020). Machine Learning Approach for Credit Scoring. *arXiv:2008.01687*, 1-28.

Stephen, H. (2011). Credit Rating Agencies Deserve Credit for the 2007-2008 Financial Crisis: An Analysis of CRA Liability Following the Enactment of the Dodd-Frank Act. *Washington and Lee Law Review, 68,* 1924-1972.

Subasi, A., & Gürsoy, M. (2010). Comparison of PCA, ICA and LDA in EEG Signal Classification Using DWT and SVM. *Expert Systems with Applications, 37,* 8659-8666.
https://doi.org/10.1016/j.eswa.2010.06.065

Wallis, M., Kumar, K., & Gepp, A. (2019). *Managerial Perspectives on Intelligent Big Data Analytics. Chapter: Credit Rating Forecasting Using Machine Learning Techniques* (pp. 180-198). Bond University. https://doi.org/10.4018/978-1-5225-7277-0.ch010

Wang, G., Hao, J. X., Ma, J., & Jiang, H. B. (2011). A Comparative Assessment of Ensemble Learning for Credit Scoring. *Expert Systems with Applications, 38,* 223-230.
https://doi.org/10.1016/j.eswa.2010.06.048

Appendix

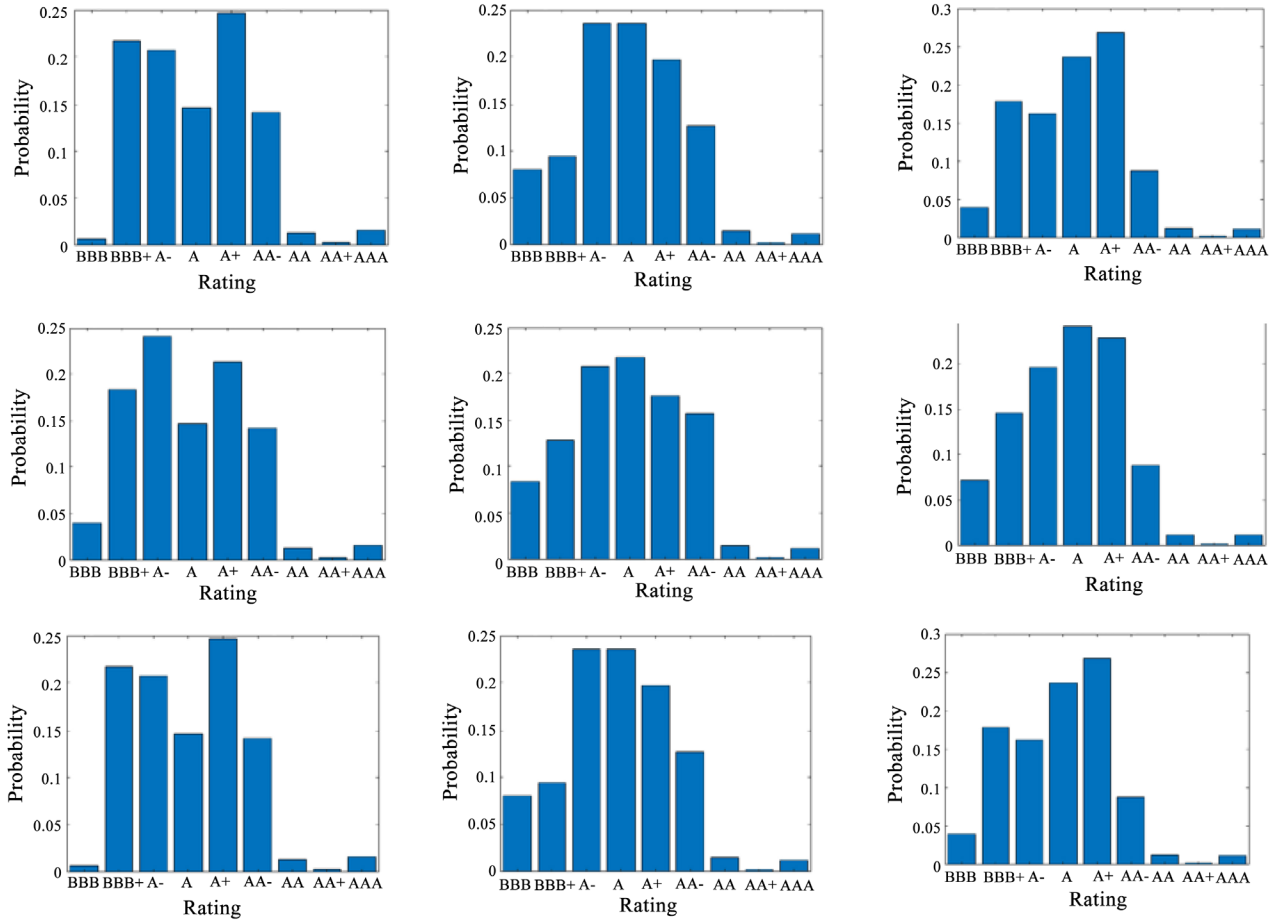
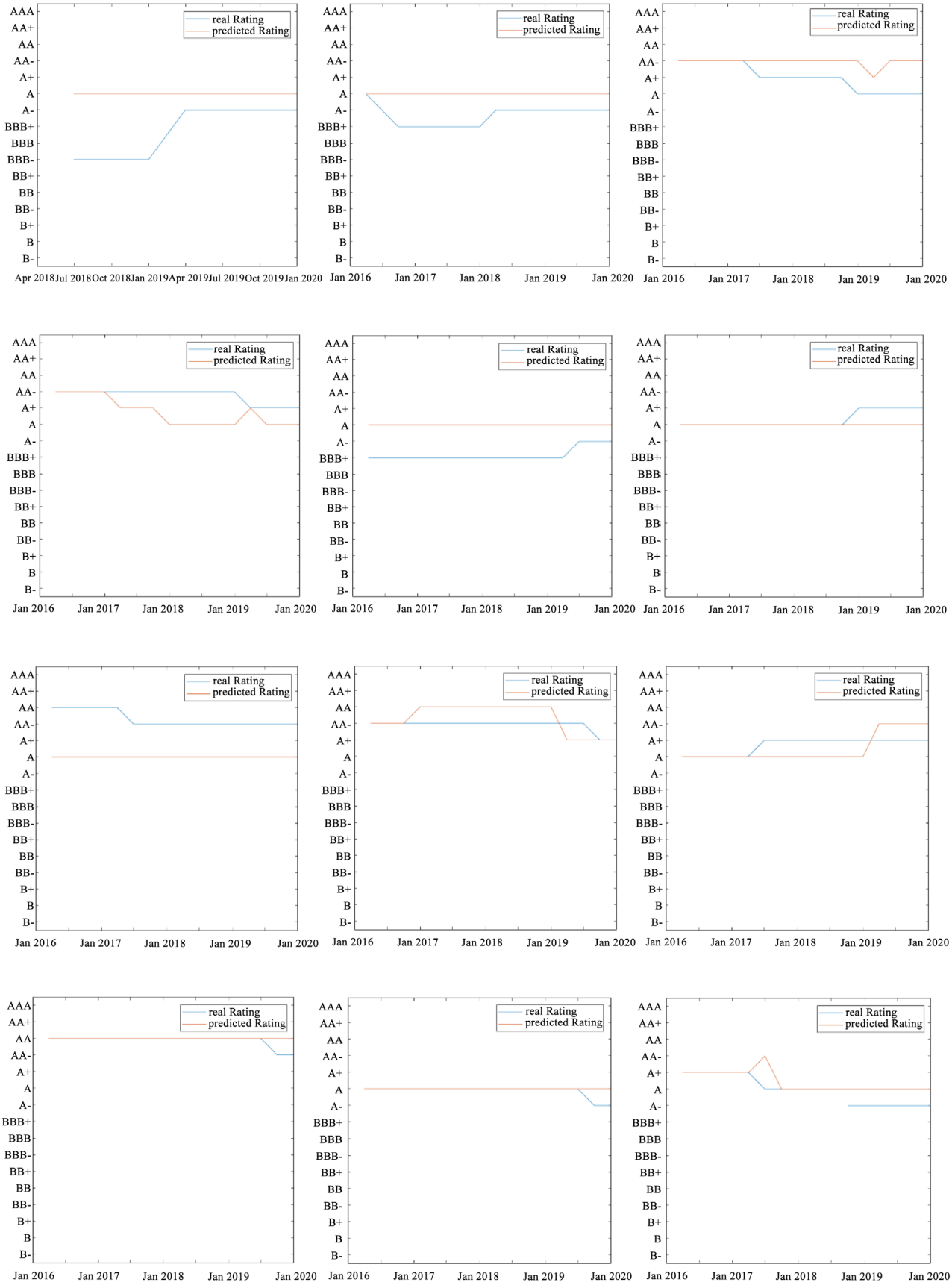


Figure A1. MLCR model result: credit rating probability distribution for different entities.

Table A1. Machine learning model validation: Experiments on the hyperparameters of MLCR model.

		38.8% (Sum importance)			55.5% (Sum importance)			70.4% (Sum importance)		
		Top 100 important			Top 200 important			Top 300 important		
Number of random features		10	20	30	20	40	60	30	60	90
200 trees	OOB	0.830	0.879	0.899	0.784	0.870	0.896	0.771	0.839	0.887
	TEST	0.449	0.473	0.464	0.447	0.459	0.475	0.441	0.475	0.477
Number of random features		40	50	60	80	100	120	120	150	180
200 trees	OOB	0.905	0.907	0.910	0.899	0.906	0.908	0.896	0.901	0.903
	TEST	0.461	0.459	0.455	0.468	0.469	0.456	0.465	0.466	0.459


real Rating
predicted Rating
AAA
AA+
AA
AA-
A+
A
A-
BBB+
BBB
BBB-
BB+
BB
BB-
B+
B
B-
Jan 2016
Jan 2017
Jan 2018
Jan 2019
Jan 2020
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Jan 2018
Apr 2018
Jul 2018
Oct 2018
Jan 2019
Apr 2019
Jul 2019
Oct 2019


real Rating
predicted Rating
AAA
AA+
AA
AA-
A+
A
A-
BBB+
BBB
BBB-
BB+
BB
BB-
B+
B
B-
Apr 2018 Jul 2018 Oct 2018 Jan 2019 Apr 2019 Jul 2019 Oct 2019 Jan 2020
Jan 2016 Jan 2017 Jan 2018 Jan 2019 Jan 2020

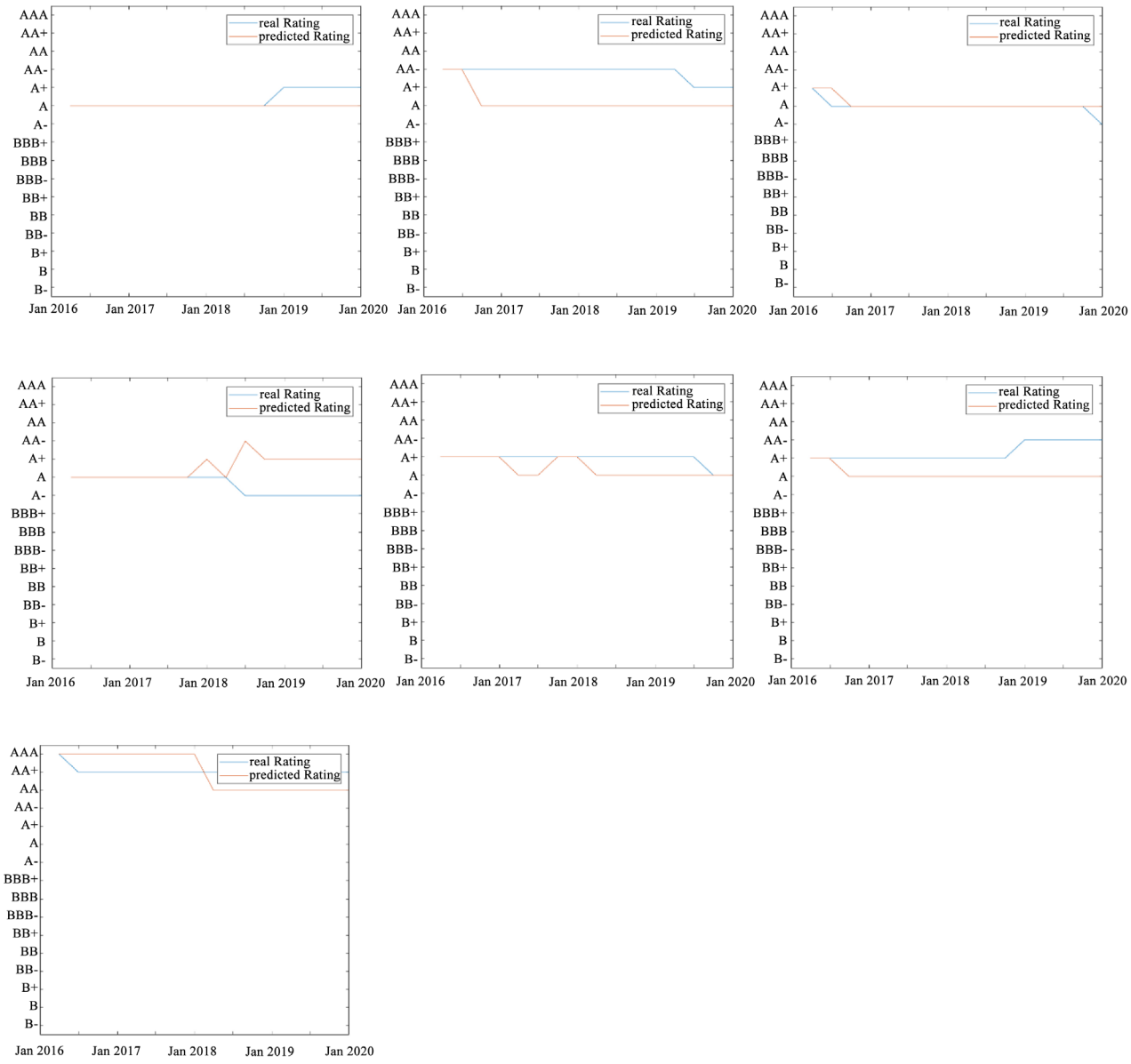


Figure A2. Dynamic model testing for different entities (for the model capacity to capture the rating change in a relatively long period).