Chapter

A Primer on Machine Learning Methods for Credit Rating Modeling

Yixiao Jiang

Abstract

Using machine learning methods, this chapter studies features that are important to predict corporate bond ratings. There is a growing literature of predicting credit ratings via machine learning methods. However, there have been less empirical studies using ensemble methods, which refer to the technique of combining the prediction of multiple classifiers. This chapter compares six machine learning models: ordered logit model (OL), neural network (NN), support vector machine (SVM), bagged decision trees (BDT), random forest (RF), and gradient boosted machines (GBMs). By providing an intuitive description for each employed method, this chapter may also serve as a primer for empirical researchers who want to learn machine learning methods. Moody's ratings were employed, with data collected from 2001 to 2017. Three broad categories of features, including financial ratios, equity risk, and bond issuer's cross-ownership relation with the credit rating agencies, were explored in the modeling phase, performed with the data prior to 2016. These models were tested on an evaluation phase, using the most recent data after 2016.

Keywords: machine learning, credit ratings, forecasting, random forest, gradient boosted machine

1. Introduction

An issue of continuing interest to many financial market participants (portfolio risk managers, for example) is to predict corporate bond ratings for unrated issuers. Issuers themselves may seek a preliminary estimate of what their rating might be to decide the ratio of debt and equity financing. Starting with the seminal works of [1, 2], pioneering studies in the finance literature use accounting ratios and other publicly available information in reduced-form models to predict credit ratings. A variety of statistical techniques (OLS, discriminant analysis, and ordered logit/probit models) were employed to identify the most important characteristics for predicting ratings. See, [3–5].

Bond rating is, in a way, a classification problem. There is also a growing literature of predicting credit ratings via machine learning (ML) methods [6–11]. As can be seen from **Table 1**, neural network (NN) and support vector machine (SVM) have been

Study	Rating Categories	Methods	Data	Accuracy	Predictors	Sample size	Benchmark Models
[7]	5	SVM, NN	Bank Ratings	~80%	21 Financial Ratios	265 (US) +74 (Taiwan)	LR:~ 75%
[8]	6	NN	Moody's long term ratings on US firms	79%	25 financial ratios	129	LDA: 33%
[9]	5	SVM	Ratings on commercial papers in Korea	67.2%	297 financial ratios	3017	NN: 59.9%, MDA: 58.8%, CBR: 63.4%
[6]	9	SVM	International bank ratings	62.4%	7 financial ratios, time and county dummies		Ordered Logit: 51.5%, Ordered Probit: 50.8%
[11]	3	RF + RST	enterprise credit ratings in Taiwan	93.4%	21 financial variable + distance to default	2470	RST: 90.3%, RF + DT: 84%, DT: 83.5% RF + SVM: 77.8%, SVM: 74.4%
[10]	7	LASSO	CDS-based and equity-based ratings	84% and 91%	268 financial factors, market-driven indicators, and macroeconomic predictors	1298 + 1540	Ordered Probit: 22% + 49%

Note: SVM = Support Vector Machine. NN = Neural Network. MDA = Multivariate Discriminant Analysis. RF = Random Forest. RST = Rough Set Theory.

Table 1.
Summary of credit rating predictive studies using machine learning.

widely employed by prior studies. However, there have been less empirical studies using *ensemble methods*, which refer to the technique of combining the prediction of multiple classifiers. This study attempts to fill the void by employing three ensemble methods to predict credit ratings and contrasting their performance with popular single-classifier ML methods.

The two popular methods for creating accurate ensembles are bootstrap aggregating, or bagging, and boosting. Previous works in the statistics and computer science literature have shown that these methods are very effective for decision trees (DT)[1], so this chapter considers DT as the basic classification method. [11] employs the random forest (RF) to predict enterprise ratings in Taiwan. To date, no comparative study has been carried out for the United States with any ensemble methods to our knowledge. Other than RF, this study also employs two additional ensemble methods: bagged decision trees (BDT) and gradient boosted machine (GBM).

This study is also the first to explore the predictive power of conflicts of interest in forecasting bond ratings. After the collapse of highly rated securities during the 07–09

[1] See, for example, [12–14].

financial crisis, the role of credit rating agencies (CRAs) as gatekeepers to financial markets has been scrutinized by academia and regulators at an unprecedented level. A number of conflicts of interest, including the issuer-pays business model, cross-ownership [15, 16], non-rating business relationship [17], transitioning analysts [18], have been identified in the literature as contributing factors to the rating inflation.

The type of conflict of interest under study arises from cross-ownership, meaning that the bond issuers and the CRA are controlled by common shareholders. Conflicts of interest between shareholders and managers, at a general level, have a variety of negative impact on the company [19]. In the context of the rating industry, as noted by [16], companies invested by Moody's two large shareholders, Berkshire Hathaway and Davis Selected Advisors, tend to receive more favorable ratings compared with others. Based on institutional ownership data, [15] constructed an index to capture bond issuers' cross-ownership with Moody's via all common shareholders and finds such biases to be more universal.

Motivated by the aforementioned studies, this chapter incorporates several conflicts of interest measure from the cross-ownership channel to predict Moody's ratings from 2001 to 2017. Since the predictive performance of ML methods is usually context-dependent, we compare the aforementioned tree-based ensemble methods (RF, BDT, and GBM) with three other ML models: ordered logit model (OL), neural network (NN), and support vector machine (SVM). RF presents the best results, correctly predicting 73.2% ratings out of sample. To improve the interpretability of "black box" ML models, we use sensitivity analysis to measure the importance and effect of particular input features in the model output response.

The rest of the chapter is organized as follow. Section 2 describes the empirical rating data and the features (attributes) under study. Section 3 discusses the three ensemble ML methods in the context of predicting credit ratings. Section 4 contains the predictive results and sensitivity analyses, and Section 5 concludes.

2. Data and features

The objective of this chapter is to predict corporate bond ratings assigned by Moody's, the leading credit rating agency (CRA) in the United States. The empirical sample consists of publicly listed companies covered in either Center for Research in Security Prices (CRSP) or Compustat. Moody's ratings on bonds issued by these companies are obtained from Mergent's Fixed Income Securities Database (FISD). Since the analysis involves Moody's shareholders, the sampling period starts from January 2001, when Moody's went to public, to December 2017.

2.1 Credit rating outcome

Under Moody's rating scale, the rating outcome falls into seven ordered categories with descending credit quality: Aaa,Aa,A,Baa,Ba,B, and C. The first four categories, from Aaa to Baa, are termed "investment-grade," whereas the remaining three are termed "high yield." The distribution of ratings over time is reported in **Table 2**. In 2004, about 50% of bonds in the data received investment grade ratings. The proportion of investment grade bonds has been trending up prior to the 07–09 financial crisis. The fact that nearly 90% of bonds received investment grade rating in 2008 suggests an obvious inflation of ratings. For the purpose of predicting credit ratings, it

Year	*Aaa*	*Aa*	*A*	*Baa*	*Ba*	*B*	*C*	*Total # of Ratings*
2001	6	17	62	125	59	44	2	315
2002	1	8	57	93	52	39	2	252
2003	8	47	67	117	54	98	12	403
2004	3	11	44	94	53	73	8	286
2005	1	21	39	93	51	49	9	263
2006	3	19	69	106	41	41	20	299
2007	6	24	103	95	41	45	4	318
2008	2	29	76	96	21	5	1	230
2009	3	15	97	164	57	73	7	416
2010	7	24	71	134	59	82	20	397
2011	10	17	117	163	33	69	12	421
2012	3	24	134	189	69	89	14	522
2013	12	29	150	218	76	76	15	576
2014	8	20	127	231	59	65	10	520
2015	20	22	178	274	53	46	3	596
2016	26	31	160	278	62	57	1	615
2017	11	31	98	166	41	36	2	385
Total	130	389	1649	2636	881	987	142	6814

Table 2.
Distribution of ratings.

is therefore important to include conflicts of interest measures, which account for this trend.

A second observation from **Table 2** is that the rating outcome is highly skewed toward the middle. The majority of bonds are rated in *A* and *Baa*, and only 2% of bonds received *Aaa* or *C* ratings. This is yet another reason to consider ensemble methods, which are known to be superior than other ML methods with single classifiers when applying to highly imbalanced data [20, 21].

2.2 Attributes under study

For each quarter from 2001Q1 to 2017Q4, a total of 20 features/attributes are obtained from a variety of sources to predict ratings. These features can be broadly categorized into three groups: (1) financial ratios, (2) equity risk measures, and (3) the bond issuer's "connectedness" with Moody's shareholders.

2.2.1 Financial ratios

We follow [22] and employ the following financial ratios in the analysis: (X1) the value of the firm's total assets (*log(asset)*), (X2) long- and short-term debt divided by total asset (*Book_lev*). (X3) Convertible debt divided by total assets (*ConvDe_assets*), (X4) rental payments divided by total assets (*Rent_Assets*), (X5) cash and marketable securities divided by total assets (*Cash_assets*), (X6) long- and short-term debt divided by EBITDA (*Debt_EBITDA*), (X7) EBITDA to interest payments (*EBITA_int*), (X8) profitability, measured as EBITDA divided by sales (*Profit*), (X9) tangibility, measured as net property, plant, and equipment divided by total assets (*PPE_assets*), (X10) capital expenditures divided by total assets (*CAPX_assets*), (X11) the volatility of profitability (*Vol_profit*), defined as the standard deviation of profitability in the last 5 years divided by the mean in absolute values. The data on the aforementioned firm-level financial ratios are obtained from the CRSP-Compustat merged database in Wharton Research and Data Services (WRDS).

There is a distinction between the issuer rating and issue rating for corporate bonds. The former addresses the issuer's overall credit creditworthiness, whereas the latter refers to specific debt obligations and considers the ranking in the capital structure such as secured or subordinated.[2] Since this chapter predicts rating at the bond level, three bond characteristics are also included: (X12) the log of the issuing amount (*Amt*), (X13) a dummy variable indicating whether the bond is senior (*Seniority*), and (X14) a dummy variable indicating whether the bond is secured (*Security*). The issuing amount affects the maximum financial loss on the investment, whereas the seniority and security status affect the priority of repayment should a default occur. Data on these bond characteristics are obtained from FSID along with the credit ratings.

2.2.2 Equity risk

As noted by [23], equity risk has been accounting for a greater proportion of variations in credit rating outcomes among the three leading CRAs in the United States. To obtain measures for a company's equity risk, we estimate a Fama–French three-factor model for each issuer in the sample.[3] The following measures are then obtained: (X15) the firm's beta (*Beta*), which is the stock's market beta computed estimated annually using the CRSP value-weighted index, and (X16) the firm's idiosyncratic risk (*Idiosyncratic risk*), computed annually as the root mean squared error from the three-factor model.

2.2.3 Cross-ownership with Moody's

As noted above, conflicts of interest are measured by the "connectedness" (cross-ownership) between Moody's and a bond issuer. To characterize the degree of cross-ownership, I first obtain the list of Moody's shareholders from Thomson Reuters (13F) and calculate their ownership stake in Moody's (the percentage of Moody's stock that they hold) for each quarter in the sampling period. Next, I access each shareholder's investment portfolio to find out which bond issuers have the same shareholders as investors. The shareholder's manager type code (MGRNO) and the firm's Committee on Uniform Securities Identification Procedures (CUSIP) number are used to match the shareholding data with bond issuers.

To summarily characterize the shared-ownership relation between bond issuers and Moody's, I employ the following measure, termed *Moody-Firm-Ownership-Index*

[2] The issuer rating usually applies to senior unsecured debt

[3] The normal estimation window is set to be 252 days prior to the rating assignment date. For companies with sparse stock price data, we require at least 126 days.

(MFOI), proposed by [15]. Suppose Moody's has $j = 1,2,\cdots,M$ shareholders in a given quarter[4], and any subset of those shareholders can invest in an issuing firm. Define

$$(X17): \qquad MFOI_i = \sum_{j=1}^{M} b_{ij} s_j \tag{1}$$

where s_j denotes shareholder j's ownership take in Moody's, and b_{ij} denotes bond issuer i's weight in shareholder j's investment portfolio. Note that $b_{ij} = 0$ means shareholder j does not invest in bond issuer i.

In addition to MFOI, three other measures are included as predictors. The first is the number of common shareholders, defined as

$$(X18): \qquad Num\_SH_i = \sum_{j=1}^{M} \mathbf{1}\{b_{ij} > 0\} \tag{2}$$

The second is the number of large common shareholders (which owns at least 5% of Moody's stock), defined as

$$(X19): \qquad Num\_large\_SH_i = \sum_{j=1}^{M} \mathbf{1}\{b_{ij} > 0\} \times \mathbf{1}\{s_j > 0.05\} \tag{3}$$

The last is a dummy variable capturing if the bond issuer is invested by Berkshire Hathaway, Moody's leading shareholder for our sampling period.

$$(X20): \qquad BRK_i = \mathbf{1}\{b_{ik} > 0\}, \qquad k = \text{Berkshire Hathaway} \tag{4}$$

Berkshire Hathaway is singled out here because it owns significantly more shares of Moody's compared with any other large shareholders.

2.3 Descriptive statistics

After combining data from multiple sources, the final dataset consists of 6817 bonds issued by 895 firms. The descriptive statistics for the 20 features/attributes are reported in **Table 3**. For asset (X_1), EBITDA to interest (X_7), profitability (X_8), issuing amount (X_{12}), and seniority (X_{13}), there is a clear positive correlation between rating categories and the level of these attributes. For others like the Book-leverage ratio (X_2), Debt-to-EBITDA ratio (X_6), tangibility-to-asset ratio (X_{10}), volatility of profit (X_{11}), and idiosyncratic risk (X_{16}), the correlation is negative. For the four conflicts of interest measures (X_{17} - X_{20}), they all decrease as the rating drops.

3. Methods

The dataset is split into two subsets based on the timing of the rating: a training set, which consists of 5814 (85.3% of the total) ratings before 2016, and a holdout set, which consists of 1000 (14.7%) ratings in 2016–2017. In this section, we discuss the

[4] Since all of the variable are time-specific, I drop the time t subscript for notational simplicity

	Aaa	*Aa*	*A*	*Baa*	*Ba*	*B*	*C*
Financial Ratio							
Asset(X_1)	11.91	12.13	10.69	9.85	8.68	7.99	7.75
Book_lev (X_2)	0.18	0.33	0.29	0.32	0.36	0.46	0.59
ConvDe_asset (X_3)	0.00	0.00	0.00	0.01	0.02	0.02	0.04
Rent_asset(X_4)	0.01	0.01	0.01	0.01	0.01	0.02	0.02
Cash_asset(X_5)	0.28	0.12	0.13	0.08	0.09	0.09	0.09
Debt_ebitda(X_6)	1.42	5.45	3.14	3.03	2.90	3.85	6.66
Ebitda_int (X_7)	48.64	27.62	20.26	10.82	6.79	4.21	2.50
Profit(X_8)	0.31	0.31	0.27	0.23	0.19	0.19	0.24
PPE_asset(X_9)	0.22	0.21	0.25	0.32	0.31	0.36	0.46
CAPEX_asset (X_{10})	0.04	0.04	0.04	0.05	0.05	0.06	0.08
Profit_vol (X_{11})	0.06	0.06	0.11	0.13	0.19	0.18	0.01
Amt(X_{12})	13.92	13.11	13.27	13.12	12.84	12.68	12.37
Seniority(X_{13})	0.99	0.99	0.99	0.98	0.88	0.81	0.82
Secure(X_{14})	0.01	0.00	0.00	0.01	0.05	0.10	0.06
Equity Risk							
Beta(X_{15})	0.82	1.10	1.00	0.87	1.11	1.33	1.54
Idiosyncratic risk(X_{16})	0.06	0.07	0.08	0.08	0.12	0.14	0.17
Conflicts of Interest							
MFOI × 10,000 (X_{17})	87.48	59.30	24.32	8.54	2.31	1.11	0.74
Num_SH(X_{18})	335.41	281.01	269.72	217.16	144.40	101.76	94.87
Num_large_SH(X_{19})	1.59	1.40	1.18	0.95	0.80	0.74	0.76
BRK(X_{20})	0.32	0.30	0.06	0.03	0.02	0.01	0.01

Table 3.
Descriptive statistics by rating categories.

methodological aspect of three resemble methods—Random Forest (RF), Bagging, and Gradient Boosted Modeling (GBM)—and how they are implemented. The performances of these methods are compared with three other ML models: Ordered Logit Regression (OLR), Support Vector Machine (SVM), and Neural Network (NN), based on the predictive accuracy in the holdout set.

3.1 Decision trees

To understand the resemble method, we must first understand decision trees, the basic classification procedure upon which the ensemble (or resulting classification) is based[5]. For illustrative purpose, consider a sample decision tree that includes categorical outcome Y (credit rating) and three predictor variables: firm asset, leverage, and

[5] In this study, we restrict our attention to tree-based resemble methods because decision trees are extremely fast to train.

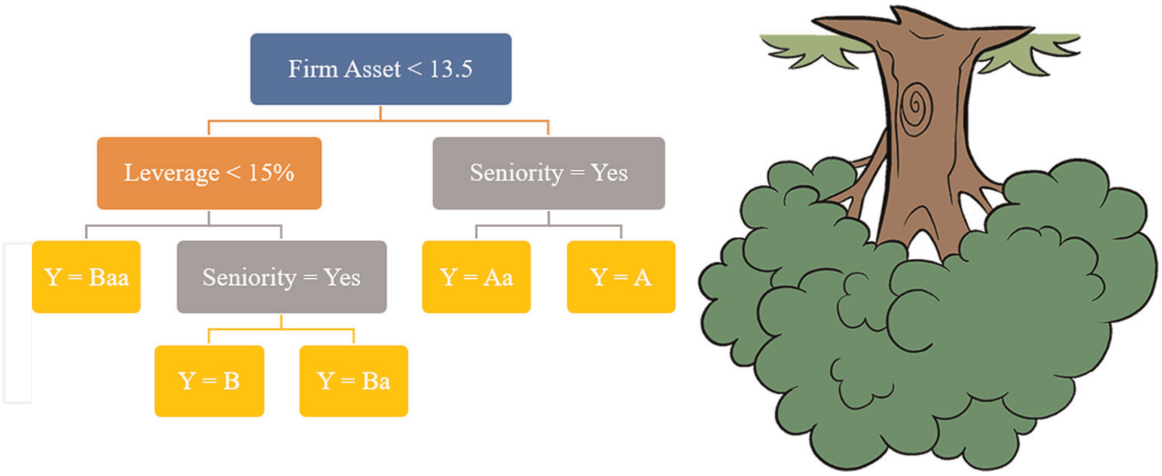

Figure 1.
Sample decision tree.

seniority (binary). As displayed in **Figure 1**, the main components of a decision tree model are nodes and branches, while the complexity of the decision tree is governed by splitting, stopping, and pruning.

Nodes There are three types of nodes. (a) A root node, also called a decision node, represents the most important feature (in this case, the level of log (firm asset)) that will lead all subdivisions. (b) Leaf nodes, also called end nodes, represent the final predicted rating outcome based on the sequence of divisions. (c) Internal nodes, also called chance nodes, represent the intermediate sequence of features that guide the classification.

Branches A decision tree model is formed using a hierarchy of branches, with the more important features displayed closer to the root node. Each path from the root node through internal nodes to a leaf node represents a classification decision sequence. These decision tree pathways can also be represented as "if-then" rules, with the left branch denoting the binary condition is met. For example, "if the natural log of firm asset is less than 13.5 and the leverage ratio is less than 15%, then the bond is rated as Baa."

Splitting Measures that are related to the degree of "purity" of the subsequent nodes (i.e., the proportion with the target condition) are used to choose between different potential input variables; these measures include entropy, Gini index, classification error, information gain, and gain ratio. Normally not all potential input variables will be used to build the decision tree model and in some cases a specific input variable may be used multiple times at different levels of the decision tree.

Stopping and Prunning An overly complex tree can result in each leaf node 100% pure (i.e., all bonds have the same rating), but is likely to suffer from the problem of overfitting. To prevent this from happening, one may grow a large tree first and then prune it to optimal size by removing nodes that provide less additional information. One parameter that controls the complexity is the number of leaf nodes.

3.2 Bagging

The decision trees discussed above suffer from high variance, meaning if the training data are split into multiple parts at random with the same decision tree applied to each, the predictive results can be quite different. Bootstrap aggregation, or bagging, is a technique used to reduce the variance of predictions by combining the result of multiple classifiers modeled on different subsamples of the same dataset. When applying bagging to decision trees, usually the trees are grown deep and are not pruned. Hence, each individual tree has high variance, but low bias. Averaging hundreds or even thousands of trees can reduce the variance and improve the predictive performance.

In practice, different subsamples are drawn from the training set with replacement (See, [24] for a detailed discussion of the bagging sampling approach). Each subsample has the same size with the training set, but only contains 2/3 of the data of the original data on average. The number of bootstrapped sample is therefore a hyperparameter to be tuned. For each bootstrapped sample, we fit a "bushy" deep decision tree with all 20 features considered at each splitting. Each tree acts as a base classifier to determine the rating of a bond. The final prediction is done via "majority voting" where each classifier casts one vote for its predicted rating, then the category with the most votes is used to classify the credit rating.

3.3 Random forest

Random forest is another ensemble classification method developed by [25]. One advantage of random forest (RF) over bagging is that it reduces the correlation among trees by randomizing the number of features. RF combines the bagging sampling approach of [24] and the random selection of features, introduced independently by [26, 27], to construct a collection of decision trees with controlled variation. Specifically, [25] recommends to randomly select $m = log_2(p + 1)$ features at any given splitting, with p being the total number of features, to grow each individual tree. Moreover, each tree is constructed using a subsample of the training set with replacement.

For the purpose of illustration, in **Figure 2**, we consider an RF populated by three trees that are similar to the one described in **Figure 1**. Note that the total number of features is 3. In this case, $m = log_2(4) = 2$, so each tree is generated using two features. For a bond with firm asset = 12, seniority = yes, and leverage = 12%, the majority rule returns a predicted rating of *Ba* category. In practice, the complexity of the random forest is governed by several hyperparameters, such as the number of trees and the maximum features at each splitting.

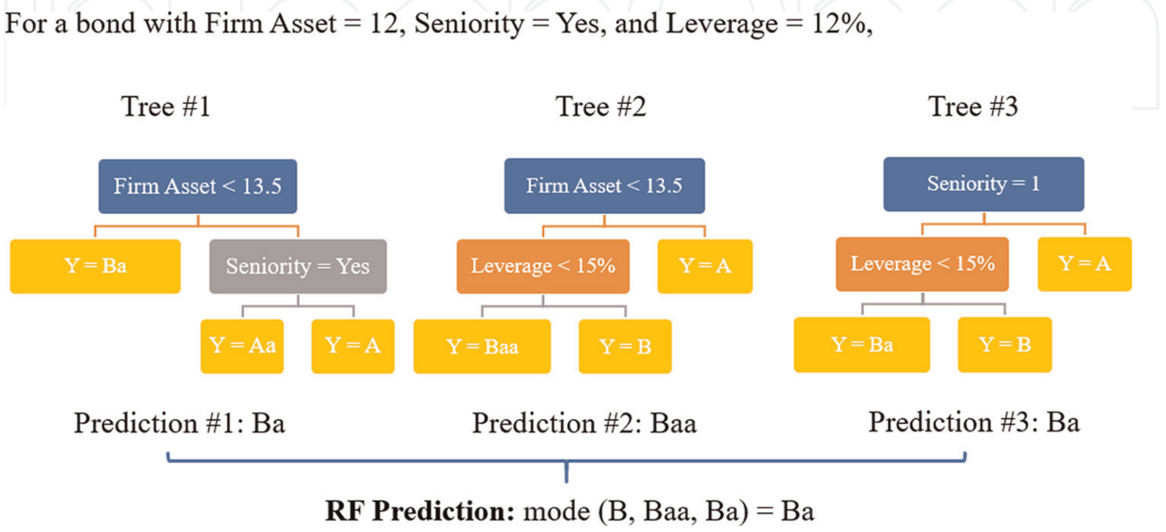

Figure 2.
Sample random forest.

3.4 Gradient boosting machines

Gradient Boosting Machines (GBMs) are a ensemble method, which recognizes the weak learners and attempts to strengthen those learners in a recursive manner to improve prediction. The key difference between GBM and Bagging is that the training stage is parallel for Bagging (i.e., each tree is built independently), whereas GBM builds the new tree in a sequential way. Specifically, when the first tree is generated, the residual errors are calculated and used in the next tree as the target variable. The predictions made by this last last tree are combined with the previous model's predictions. New residuals are calculated using the predicted value and the actual value. This process is repeated until the errors no longer decreased significantly.

During the prediction stage, bagging and RF simply average the individual predictions (the "majority rule"). In contrast, a new set of weights will be assigned to each tree in GBM. The final predicted rating is an weighted average of individual predictions. A tree with a good classification result on the training data will be assigned a higher weight than a poor one. There is no consensus regarding to which method is better than the other; the answer very much depends on the data and the researcher's objective. Some scholars have argued that gradient boosted trees can outperform random forest [28, 29]. Others believe boosting tends to aggregate the overfitting problem because repeatedly fitting the residuals can capture noisy information.

4. Results

In this section, we begin by comparing the three aforementioned ensemble methods (BDT, RF, and GBM) in terms of the out-of-sample predictive accuracy. Three non-ensemble ML methods, the ordered-logit model, support vector machine, and neural network, are also evaluated with the same dataset. For each employed method, we discuss the relevant hyperparameters and how they are tuned empirically.

All ML methods were implemented using the software R. To be specific, BDT and RF were implemented using the *randomForest* package. The number of features is fixed at all 20 for BDT. For RF, each tree randomly selects $m = log_2(20 + 1) = 5$ features. GBM is implemented using the package *gbm* package. For the three non-ensemble ML methods, ordered-logit model is implemented using the *polr* function from the *MASS* package. Support Vector Machine is implemented via the *svm* function from the *e1071* package. The neural network is implemented using the *neuralnet* package.

4.1 Predictive results

Bagged Decision Tree (BDT) To evaluate the predictive results, we report the classification matrix in the holdout sample for each employed method. In the case of BDT, the main hyperparameter needs to be tuned is the number of trees. We run three BDTs, setting the number of trees to be 200, 500, and 800. It is found the model with 500 trees has the highest predictive accuracy (=69.1%). The full classification matrix is reported in **Table 4**. The horizontal dimension represents the true rating received in the holdout sample, whereas the vertical dimension represents the predicted rating category. Therefore, the entries on the diagonal line capture the number of ratings

	Actual Ratings						
Predicted	**Aaa**	**Aa**	**A**	**Baa**	**Ba**	**B**	**C**
Aaa	19	0	0	0	0	0	0
Aa	0	28	8	0	0	0	0
A	4	3	148	15	2	1	0
Baa	14	31	95	390	39	9	0
Ba	0	0	7	22	35	13	0
B	0	0	0	17	27	70	2
C	0	0	0	0	0	0	1
Total	37	62	258	444	103	93	3
Accuracy							69.1%

Table 4.
The classification confusion matrix of BDT with 500 trees in holdout sample.

correctly predicted for a particular category. For example, the numbers in the first column shall be interpreted as 19 Aaa bonds are correctly classified as Aaa, whereas four (14) are misclassified into A (Baa).

Random Forest (RF) The next predictive model under evaluation is the Random Forest. In addition to the bagging technique, RF also randomizes the features set to further decrease the correlation among the decision trees. As noted above, RF has five hyperparameters that govern the complexity of the model. To decide these hyperparameter values, we implement a five-dimensional grid search where every combination of hyperparameters of interest is assessed. The hyperparameter grid is generated by

$$\mathcal{G} = \{m \times N \times n \times p \times r\}, \qquad \text{where} \tag{5}$$

- $m \in (3{,}4{,}5{,}6{,}7{,}8)$ is the number of features to consider at any given split.
- $N \in (1{,}2{,}3)$ is the minimum number of Nodes in each tree
- $n \in (200{,}500{,}800)$ is the number of trees in the forest.
- $p \in (0.6{,}0.8{,}1)\times$ (size of the training set) amount of data to generate each tree.
- $r = 1/0$ (with or without replacement in the sampling).

Consequently, a total of 216 ($= 6 \times 2 \times 3 \times 2 \times 3$) specifications of RF are compared in terms of the predictive accuracy in the holdout set. As shown in **Table 5**, the best predictive model consists of 500 trees, with each tree generated from the entire training set ($p = 1$) with replacement. In each splitting, $m = 4$ features are randomly selected. The overall classification accuracy of the holdout data turned out to be 73.2%. From the classification confusion matrix in **Table 6**, RF has a reliable predictive performance in almost all rating categories.

To develop some sense of how RF make prediction, **Figure 3** plots one decision tree from the RF model. There are a total of six attributes used in this particular tree. MFOI

	Hyperparameters						Evaluation
Model ID	*m*	*N*	*n*	*r*	*p*	RMSE	% of correct prediction
158	4	1	500	TRUE	1	0.259	0.732
5	7	1	200	TRUE	0.6	0.283	0.729
80	4	3	200	TRUE	0.8	0.277	0.728
152	4	3	200	TRUE	1	0.271	0.728
146	4	1	200	TRUE	1	0.262	0.723
176	4	3	800	TRUE	1	0.266	0.723
3	5	1	200	TRUE	0.6	0.285	0.722
170	4	1	800	TRUE	1	0.261	0.722
112	6	1	200	FALSE	0.8	0.263	0.720
86	4	1	500	TRUE	0.8	0.270	0.719

Table 5.
The 10 best RF models from hyperparameters tunning.

	Actual Ratings						
Predicted	Aaa	Aa	A	Baa	Ba	B	C
Aaa	19	1	0	0	0	0	0
Aa	0	38	9	0	0	0	0
A	18	23	171	15	2	0	0
Baa	0	0	78	413	49	16	0
Ba	0	0	0	7	33	20	0
B	0	0	0	9	19	57	2
C	0	0	0	0	0	0	1
Total	37	62	258	444	103	93	3
Accuracy							73.2%

Table 6.
The classification confusion matrix of the best RF in holdout sample.

and idiosyncratic risk appear to be the two most important attributes. From the rightmost terminal node, it is almost certain that bonds with MFOI < 1.7 and idiosyncratic risk > 0.1 can only receive high-yield ratings (25% Ba +57% B + 10% C = 91% of high yield), irrespective of other features. This provides a remarkably parsimonious yet robust decision rule to decide whether a bond is investment grade or not.

Gradient Boosting Machine (GBM) The classification confusion matrix of GBM is reported in **Table 7**. The overall predictive accuracy is 64.4%, which is 5 percentage point lower than BDT and nearly 10 percentage point lower than RF. As noted by [30], predictive results from Boosting methods are usually more volatile. [14] also made a conjecture that Boosting's sensitivity to noise may be partially responsible for its occasional increase in errors. As such, we recommend to always use RF or BDT for predicting credit ratings.

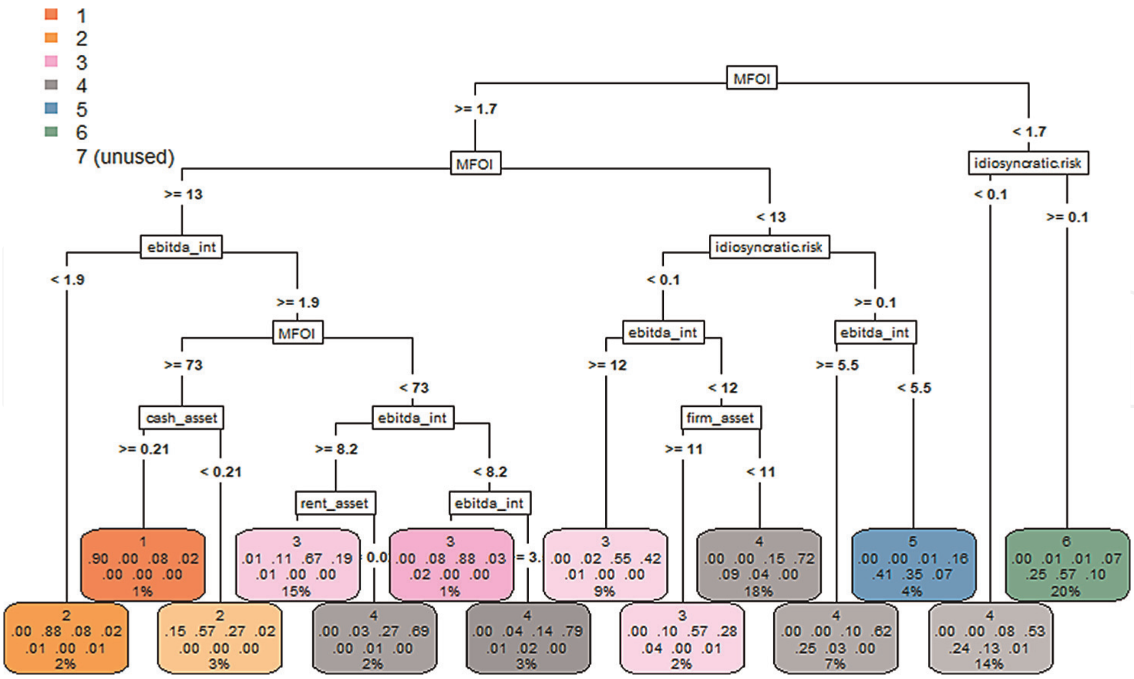

Figure 3.
Decision tree extracted from the RF model.

				Actual Ratings			
Predicted	**Aaa**	**Aa**	**A**	**Baa**	**Ba**	**B**	**C**
Aaa	0	0	0	0	0	0	0
Aa	0	7	5	0	0	0	0
A	35	28	174	33	6	0	0
Baa	2	27	74	370	40	10	0
Ba	0	0	1	23	32	19	0
B	0	0	4	18	25	61	3
C	0	0	0	0	0	3	0
Total	37	62	258	444	103	93	3
Accuracy							64.4%

Table 7.
The classification confusion matrix of GBM in holdout sample.

Ordered Logistic Regression (OLR) The OLR is a regression model where different features affect the rating outcome through the logistic transformation. Let $Z_i = \beta_0 + \sum_{j=1}^{20} x_{ij}\beta_j$ be a linear index summarizing the information of the 20 considered features where the β coefficients are to be estimated from the data. The predicted probability in OLR for each rating category, $k = 1, \cdots, 7$, can be described as $Pr(Y_{ik} = 1|x_i) = \frac{1}{1+exp\,(Z_i-\kappa_k)} - \frac{1}{1+exp\,(Z_i-\kappa_{k-1})}$ where κ_k is a series of threshold point separating the different ratings with $k_0 = -\infty$ and $k_7 = \infty$. While the model is easier to interpret, it is quite rigid and cannot accomodate complex nonlinear relationships.

The classification matrix of OLR is reported in **Table 8**. The overall classification accuracy is 53.9% for the holdout sample, which is much worse than RF. The model also fails to correctly predict all 37 *Aaa* bonds. This is unsurprising: when fitting a linear trend in the data (OLR belongs to the family of generalized linear model because the logistic transformation is applied on a linear score function of features), the fitness is usually worse in the tails of the distribution (**Table 9**).

Support Vector Machine (SVM) developed by [31] seeks to find the optimal separating hyperplane between binary classes by following the maximized margin criterion. When it comes to multiclass prediction where the outcome variables take k distinct categories, one may induce $\frac{k(k-1)}{2}$ individual binary classifiers and then use the majority rule to determine the final predicted outcome. In order to find the separating hyperplane, SVM uses a kernel function to enlarge the feature space using basis

				Actual Ratings			
Predicted	**Aaa**	**Aa**	**A**	**Baa**	**Ba**	**B**	**C**
Aaa	0	3	0	0	0	0	0
Aa	23	21	3	0	0	0	0
A	14	18	140	110	2	0	0
Baa	0	20	115	322	69	27	0
Ba	0	0	0	7	17	25	0
B	0	0	0	5	15	38	2
C	0	0	0	0	0	3	1
Total	37	62	258	444	103	93	3
Accuracy							53.9%

Table 8.
The classification confusion matrix of OLR in holdout sample.

				Actual Ratings			
Predicted	**Aaa**	**Aa**	**A**	**Baa**	**Ba**	**B**	**C**
Aaa	26	1	0	0	0	0	0
Aa	0	8	8	3	0	0	0
A	11	34	189	58	11	0	0
Baa	0	7	59	348	39	7	0
Ba	0	0	0	6	31	12	0
B	0	12	2	29	22	70	3
C	0	0	0	0	0	4	0
Total	37	62	258	444	103	93	3
Accuracy							67.2%

Table 9.
The classification confusion matrix of SVM in holdout sample.

functions. Mathematically, SVM can be viewed as the following constrained maximization problem,

$$min_{\alpha} \qquad \frac{1}{2}\alpha^T Q\alpha - e^T\alpha \tag{6}$$

$$\text{s.t} \qquad 0 \leq \alpha \leq Ce, \qquad y^T\alpha = 0 \tag{7}$$

where e is the vector of all ones, Q is a $N \times N$ semi-positive definite matrix, $Q_{ij} = y_i y_j K(x_i, x_j)$ with K being the kernel function.

This chapter follows [9] and employs the radial basis function (RBF): $k(x_i, x_j) = exp\left\{-\gamma|x_i - x_j|^2\right\}$, where γ and C are hyperparameters to be selected. A series of SVMs with $C = 2^c$ and $\gamma = 2^g$ are implemented. Based on a 10-fold cross-validation, the best parameters are $C = 32$ and $\gamma = 0.25$. The overall classification accuracy turns out to be 67.2% for SVM, which lies between ORL and RF.

Neural Network (NN) The artificial neural network (NN) models are proposed by cognitive scientists to mimic the way that brain processes information. As noted by [32], NN can be viewed as a nonlinear regression model in the following form,

$$f(x, \theta) = \tilde{x}'\alpha + \sum_{s}^{q} G(\tilde{x}'\gamma_s)\beta_s \tag{8}$$

where $\tilde{x} = (1, x')'$, q is a integer representing the number of hidden neurons, and $G(\cdot)$ is a given nonlinear activation function. NN processes information in a hierarchical manner: the signals from an *input node* x_j $(i = 1,\cdots,20)$ are first amplified or attenuated by γ_{js} and arrive at q *hidden* (intermediate) *nodes*. The aggregated signals, in the form of $tildex'\gamma_s$, are then passed to the seven *output nodes* (e.g., the potential rating outcome) by the operation of the activation function $G(\tilde{x}'\gamma_s)$. As in the previous step, information at the hidden node s is amplified or attenuated by β_s. Other than through hidden nodes, signals are also allowed to affect the rating outcome directly through weights α.

For simplicity, this study focuses on a three-layer NN and varies the number of nodes in the hidden layer for training. In particular, 5, 10, 15, 20 hidden nodes are used. For each case, we run the same model with 50 replications to tease out the impact of bad starting values. In terms of the predictive accuracy, we find that the model with five hidden nodes slightly outperforms the rest (57.3, 56.4, 56.3, and 55.4%). In **Table 10**, we report the classification matrix for one of the NN models, with the network structure presented in **Figure 4**.

4.2 Sensitivity analysis

To explore which features are more important than others in predicting ratings, we performed two sensitivity analyses. While the analyses can be applied to any aforementioned ML methods, we decide to focus on RF due to its superior predictive performance.

The first analysis is the variable importance plots (VIP). Loosely speaking, variable importance is the increase in model error when the feature's information is "destroyed." On the left panel of **Figure 5**, we show the impurity-based measure

	Actual Ratings						
Predicted	**Aaa**	**Aa**	**A**	**Baa**	**Ba**	**B**	**C**
Aaa	11	3	3	0	0	0	0
Aa	0	7	5	0	0	0	0
A	26	51	171	73	2	0	0
Baa	0	1	79	319	47	9	0
Ba	0	0	0	19	21	12	0
B	0	0	0	33	33	71	2
C	0	0	0	0	0	0	1
Total	37	62	258	444	103	93	3
Accuracy							60.1%

Table 10.
The classification confusion matrix of NN in holdout sample.

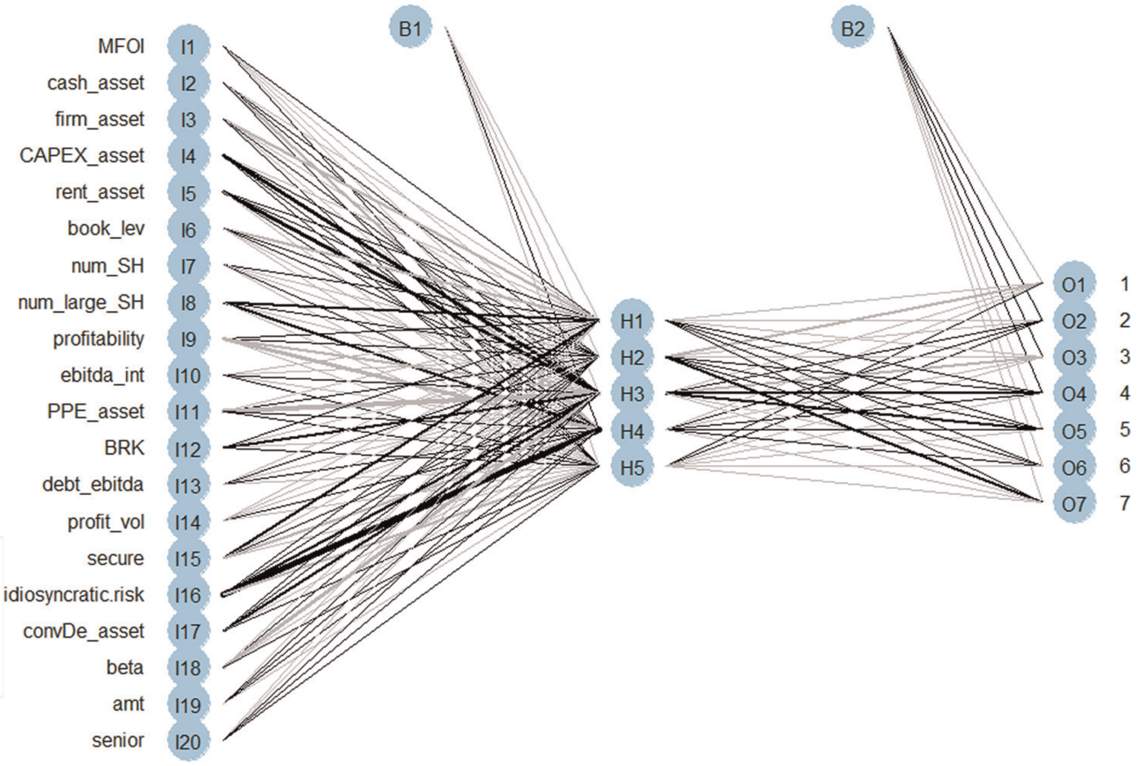

Figure 4.
NN with five hidden nodes (A darker line means a stronger signal).

where we base feature importance on the average total reduction of the loss function for a given feature across all trees. On the right panel, we show the permutation-based importance measure[6]. A feature is “important” if shuffling its values increases the model error, because in this case the model relied on the feature for the prediction.

[6] In the permutation-based approach, the values for each variable are randomly permuted, one at a time, and the accuracy is again computed. The decrease in accuracy as a result of this randomly shuffling of feature values is averaged over all the trees for each predictor [33].

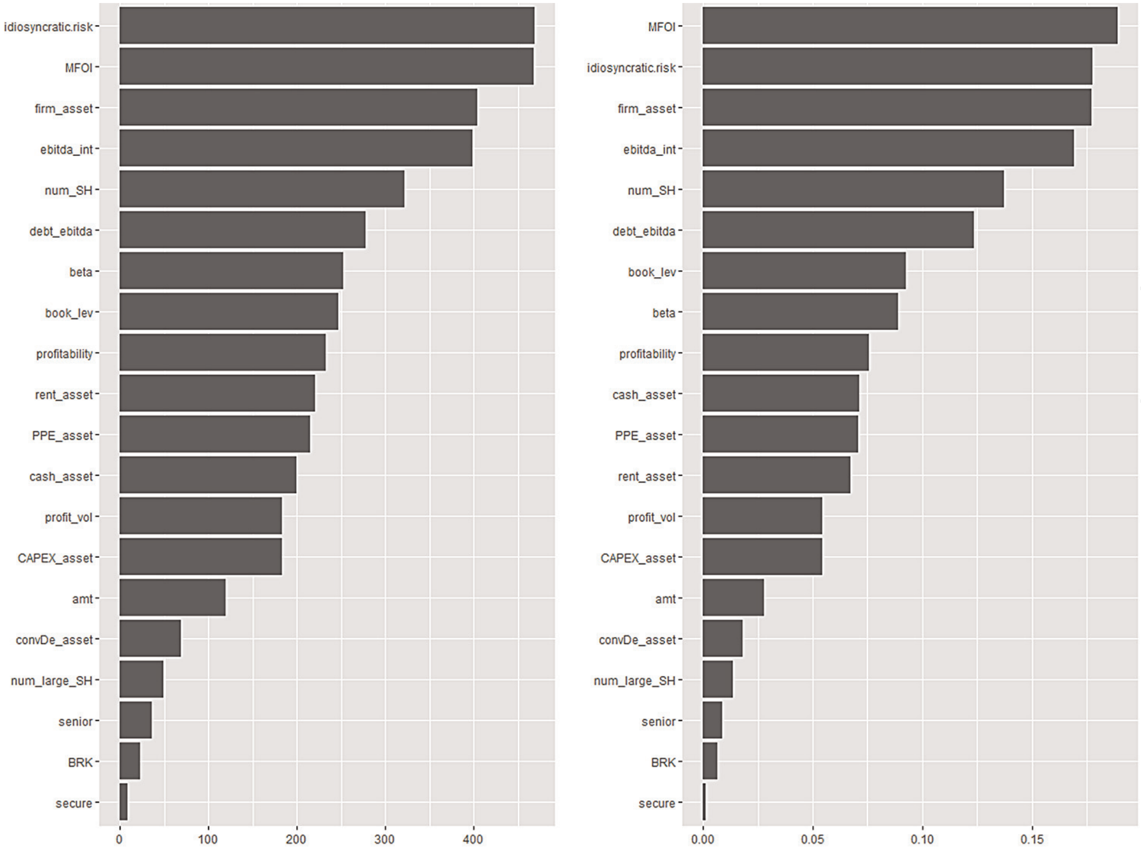

Figure 5.
Variable importance plot of each attribute for the RF model. Note: the figure on the left (right) ranks importance based on the Gini-impurity (permutation).

Both measures consistently identify the two most important attributes to be MFOI and the idiosyncratic risk of the bond issuer's stock. Eliminating the information contained in MFOI, from the permutation-based metric, decreases the predictive accuracy by about 20%.

The second sensitivity analysis is to compute the Partial Dependence (PD) for important attributes. To describe the notion of partial dependence, let $X = \{x_1, x_2, \cdots, x_{20}\}$ represent the set of the predictor variables in the RF model where the prediction function is denoted by $\hat{f}(X)$. The "partial dependence" of x_1, for example, is defined as

$$PD(x_1) = \frac{\partial}{\partial x_1} \mathbf{E}_{x_1}\left[\hat{f}(x_1,\ x_c)\right] = \frac{\partial}{\partial x_1} \int \hat{f}(x_1, x_c) p_c(x_c) dx_c \tag{9}$$

where $X_c = \{x_2, x_3, \cdots, x_{20}\}$ denote the other predictors and $p_c(x_c)$ is the marginal probability density of $x_c : p_c(x_c) = \int p(X) dx_c$. This quantity, which resembles a marginal effect, can be estimated from a set of training data by

$$\hat{PD}(x_1) = \frac{1}{n} \sum_i \frac{\partial}{\partial x_1} \hat{f}(x_1, x_{c,i}) \tag{10}$$

where $x_{c,i}$ are the values of x_c that occur in the training sample; that is, we average out the effects of all the other predictors in the model. In **Figure 6**, we report the PDs

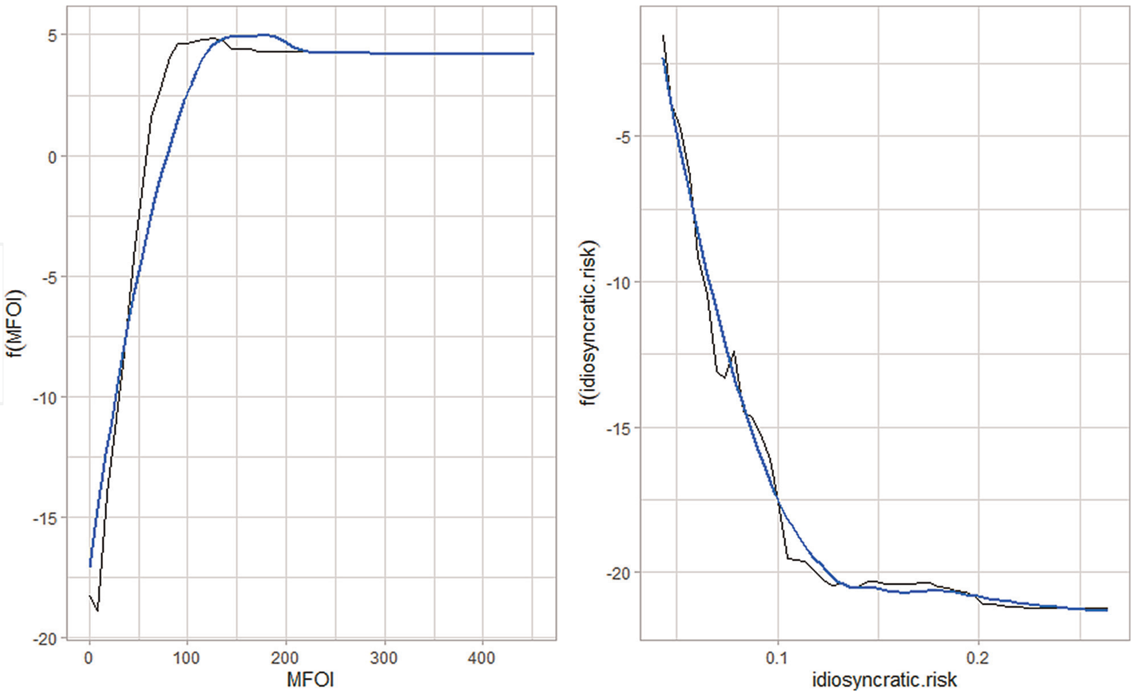

Figure 6.
Partial dependence plot for MFOI and idiosyncratic risk from the RF model. Note: The black line depicts the PD at specific values of MFOI/idiosyncratic risk. The blue line is the fitted value.

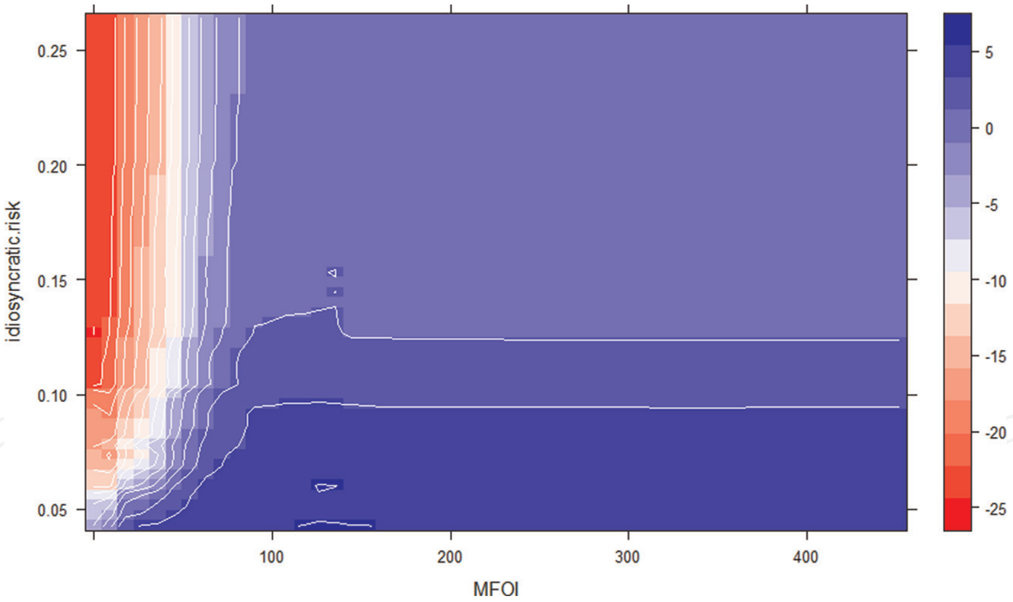

Figure 7.
Joint partial dependence plot for MFOI and idiosyncratic risk.

for MFOI and idiosyncratic risk separately. From the left panel, a lower value of MFOI has a negative impact on the rating outcome. As MFOI goes above 50, it starts to affect the rating in a positive way (a higher degree of connectedness between Moody and the issuer firm, as measured by MFOI, translates to a higher predicted rating). The positive impact of MFOI increases with the level of MFOI and plateaues as MFOI goes above 150, which is about the 99 percentile of its distribution. Conversely, we see that a larger idiosyncratic risk has a more deteriorating impact on ratings. Both patterns are economically sounding. **Figure 7** represents the joint PD for MFOI and idiosyncratic risk. The negative impact of idiosyncratic risk is only pronounced when MFOI is low.

4.3 Discussion

The main message emerged from our empirical exercise is that conflicts of interest, as measured by bond issuer's connection with Moody's shareholders, have a strong predictive power in the credit rating outcome. This observation is consistent with several previous studies. [16] found that Moody's has been assigning more favorable ratings (relative to that of S&P's) to issuers related to its two largest shareholders—Berkshire Hathaway and Davis Selected Advisors. [23, 34] showed that such bias is more universal and apply to issuers associated with any large shareholders of Moody's.

Although cross-ownership has been recognized in the literature as a important driver of credit ratings, it has not been explicitly considered as a predictor variable in any prior studies that focus on prediction. This study complements the above by confirming that cross-ownership can be utilized to increase the predictability of credit ratings.

5. Conclusions

In this chapter, we employ six machine learning methods to predict bond ratings from a sample of US public firms. Other than the financial ratios employed by previous studies, this chapter expands the feature sets to include equity risk measures and the bond issuer's cross-ownership relation with the rating agency. Inclusion of the latter source of information is unprecedented.

Several observations/conclusions emerge from the analysis. (1) Ensemble methods, including the Random Forest, Bagged Decision Trees, and Gradient Boosting Machines, generally outperform the ML methods with a single classifier. (2) Among the three ensemble methods, random forest shows a significantly better performance than the other (correctly predicting 5% more bonds than bagging and 10% more bonds than boosting). (3) Sensitivity analyses reveals the firm's idiosyncratic risk and cross-ownership relation with the rating agency as the two most important attributes in predicting ratings.

Author details

Yixiao Jiang[†]
Economics, Christopher Newport University, Newport News, USA

*Address all correspondence to: yixiao.jiang@cnu.edu

† This research is supported by the Christopher Newport University Faculty Development Fund.

References

[1] Altman EI. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The Journal of Finance. 1968;**23**(4):589-609

[2] Horrigan JO. The determination of long-term credit standing with financial ratios. Journal of Accounting Research. 1966:44-62

[3] Kaplan RS, Urwitz G. Statistical models of bond ratings: A methodological inquiry. Journal of Business. 1979:231-261

[4] Pinches GE, Mingo KA. A multivariate analysis of industrial bond ratings. The Journal of Finance. 1973; **28**(1):1-18

[5] West RR. An alternative approach to predicting corporate bond ratings. Journal of Accounting Research. 1970: 118-125

[6] Bellotti T, Matousek R, Stewart C. A note comparing support vector machines and ordered choice models' predictions of international banks' ratings. Decision Support Systems. 2011;**51**(3):682-687

[7] Huang Z, Chen H, Hsu C-J, Chen W-H, Soushan W. Credit rating analysis with support vector machines and neural networks: a market comparative study. Decision Support Systems. 2004;**37**(4): 543-558

[8] Kumar K, Bhattacharya S. Artificial neural network vs linear discriminant analysis in credit ratings forecast. Review of Accounting and Finance. 2006;**5**(3):216-227

[9] Lee Y-C. Application of support vector machines to corporate credit rating prediction. Expert Systems with Applications. 2007;**33**(1):67-74

[10] Sermpinis G, Tsoukas S, Zhang P. Modelling market implied ratings using lasso variable selection techniques. Journal of Empirical Finance. 2018;**48**: 19-35

[11] Yeh C-C, Lin F, Hsu C-Y. A hybrid kmv model, random forests and rough set theory approach for credit rating. Knowledge-Based Systems. 2012;**33**: 166-172

[12] Bauer E, Kohavi R. An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. Machine Learning. 1999;**36**(1): 105-139

[13] Drucker H, et al. Boosting and other machine learning algorithms. In: Machine Learning Proceedings 1994. MA, USA: Morgan Kaufmann; 1994. pp. 53-61

[14] Freund Y, Schapire RE, et al. Experiments with a new boosting algorithm. In: ICML. Vol. 96. Citeseer; 1996. pp. 148-156

[15] Jiang Y. Semiparametric estimation of a corporate bond rating model. Econometrics. 2021;**9**(2):23

[16] Kedia S, Rajgopal S, Zhou XA. Large shareholders and credit ratings. Journal of Financial Economics. 2017;**124**(3): 632-653

[17] Baghai R, Becker B. Non-rating revenue and conflicts of interest. Swedish House of Finance Research Paper, (15-06). 2016

[18] Cornaggia J, Cornaggia KJ, Xia H. Revolving doors on wall street. Journal of Financial Economics. 2016;**120**(2): 400-419

[19] Boubaker S, Sami H. Multiple large shareholders and earnings informativeness. Review of Accounting and Finance. 2011

[20] Khoshgoftaar TM, Golawala M, Van Hulse J. An empirical study of learning from imbalanced data using random forest. In 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007). IEEE. Vol. 2. 2007. pp. 310–317

[21] Muchlinski D, Siroky D, He J, Kocher M. Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data. Political Analysis. 2016:87-103

[22] Baghai RP, Servaes H, Tamayo A. Have rating agencies become more conservative? Implications for capital structure and debt pricing. The Journal of Finance. 2014;**69**(5):1961-2005

[23] Jiang Y. Credit ratings, financial ratios, and equity risk: A decomposition analysis based on moody's, standard & poor's and fitch's ratings. Finance Research Letters. 2021:-102512

[24] Breiman L. Bagging predictors. Machine Learning. 1996;**24**(2):123-140

[25] Breiman L. Random forests. Machine Learning. 2001;**45**(1):5-32

[26] Ho TK. Random decision forests. In: Proceedings of 3rd International Conference on Document Analysis and Recognition. IEEE. Vol. 1, 1995. pp. 278–282

[27] Amit Y, Geman D. Shape quantization and recognition with randomized trees. Neural Computation. 1997;**9**(7):1545-1588

[28] Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learnin. Cited on 2009. p. 33

[29] Madeh Piryonesi S, El-Diraby TE. Using machine learning to examine impact of type of performance indicator on flexible pavement deterioration modeling. Journal of Infrastructure Systems. 2021;**27**(2):04021005

[30] Opitz D, Maclin R. Popular ensemble methods: An empirical study. Journal of Artificial Intelligence Research. 1999;**11**: 169-198

[31] Vapnik VN. An overview of statistical learning theory. IEEE Transactions on Neural Networks. 1999; **10**(5):988-999

[32] Swanson NR, White H. A model-selection approach to assessing the information in the term structure using linear models and artificial neural networks. Journal of Business & Economic Statistics. 1995;**13**(3):265-275

[33] Boehmke B, Greenwell BM. Hands-on Machine Learning with R. New York, USA: CRC Press; 2019

[34] Gu Z, Jiang Y, Yang S. Estimating unobserved soft adjustment in bond rating models: Before and after the dodd-frank act. Available at SSRN 3277328. 2018