CREDIT RISK

Machine learning: Challenges, lessons, and opportunities in credit risk modeling

July 01, 2017

Thanks to rapid increases in data availability and computing power, machine learning now plays a vital role in both technology and business. Machine learning contributes significantly to credit risk modeling applications. Using two large datasets, we analyze the performance of a set of machine learning methods in assessing credit risk of small and medium-sized borrowers, with Moody's Analytics RiskCalc model serving as the benchmark model. We find the machine learning models deliver similar accuracy ratios as the RiskCalc model. However, they are more of a "black box" than the RiskCalc model, and the results produced by machine learning methods are sometimes difficult to interpret. Machine learning methods provide a better fit for the nonlinear relationships between the explanatory variables and default risk. We also find that using a broader set of variables to predict defaults greatly improves the accuracy ratio, regardless of the models used.

Introduction

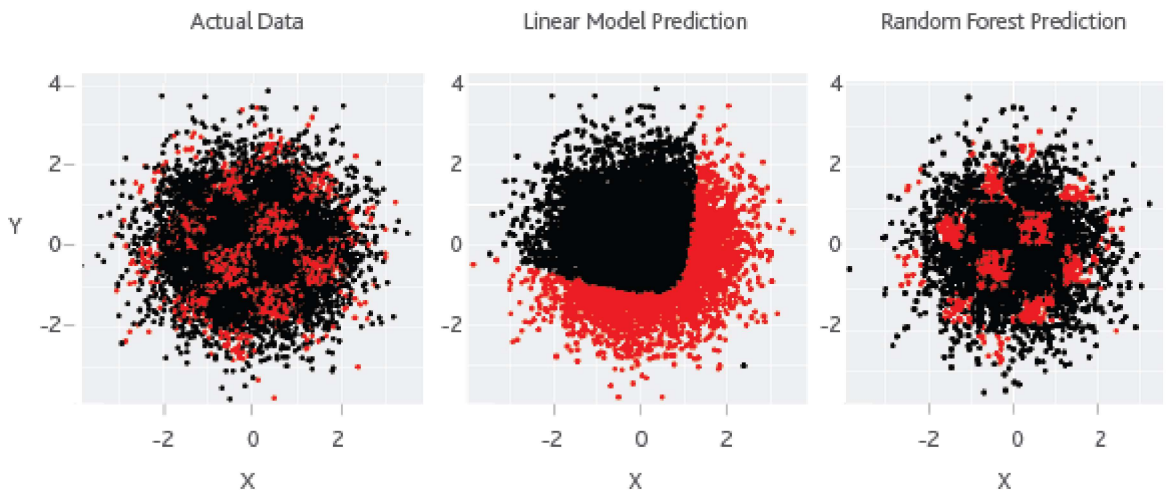
Machine learning is a method of teaching computers to parse data, learn from it, and then make a determination or prediction regarding new data. Rather than hand-coding a specific set of instructions to accomplish a particular task, the machine is "trained" using large amounts of

development, but recent improvements in data storage and computing power have made them ubiquitous across many different fields and applications, many of which are very commonplace. Apple's Siri, Facebook feeds, and Netflix movie recommendations all rely upon some form of machine learning. One of the earliest uses of machine learning was within credit risk modeling, whose goal is to use financial data to predict default risk.

When a business applies for a loan, the lender must evaluate whether the business can reliably repay the loan principal and interest. Lenders commonly use measures of profitability and leverage to assess credit risk. A profitable firm generates enough cash to cover interest expense and principal due. However, a more-leveraged firm has less equity available to weather economic shocks. Given two loan applicants – one with high profitability and high leverage, and the other with low profitability and low leverage – which firm has lower credit risk? The complexity of answering this question multiplies when banks incorporate the many other dimensions they examine during credit risk assessment. These additional dimensions typically include other financial information such as liquidity ratio, or behavioral information such as loan/trade credit payment behavior. Summarizing all of these various dimensions into one score is challenging, but machine learning techniques help achieve this goal.

The common objective behind machine learning and traditional statistical learning tools is to learn from data. Both approaches aim to investigate the underlying relationships by using a training dataset. Typically, statistical learning methods assume formal relationships between variables in the form of mathematical equations, while machine learning methods can learn from data without requiring any rules-based programming. As a result of this flexibility, machine learning methods can better fit the patterns in data. Figure 1 illustrates this point.

Figure 1 statistical model vs. machine learning



represent high-risk demographics, where we see a higher default rate. As expected, 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 linear or continuous relationships. A machine learning model, unconstrained by some of the assumptions of classic statistical models, can yield much better insights that a human analyst could not infer from the data. At times, the prediction contrasts starkly with traditional models.

A machine learning model, unconstrained by some of the assumptions of classic statistical models, can yield much better insights that a human analyst could not infer from the data.

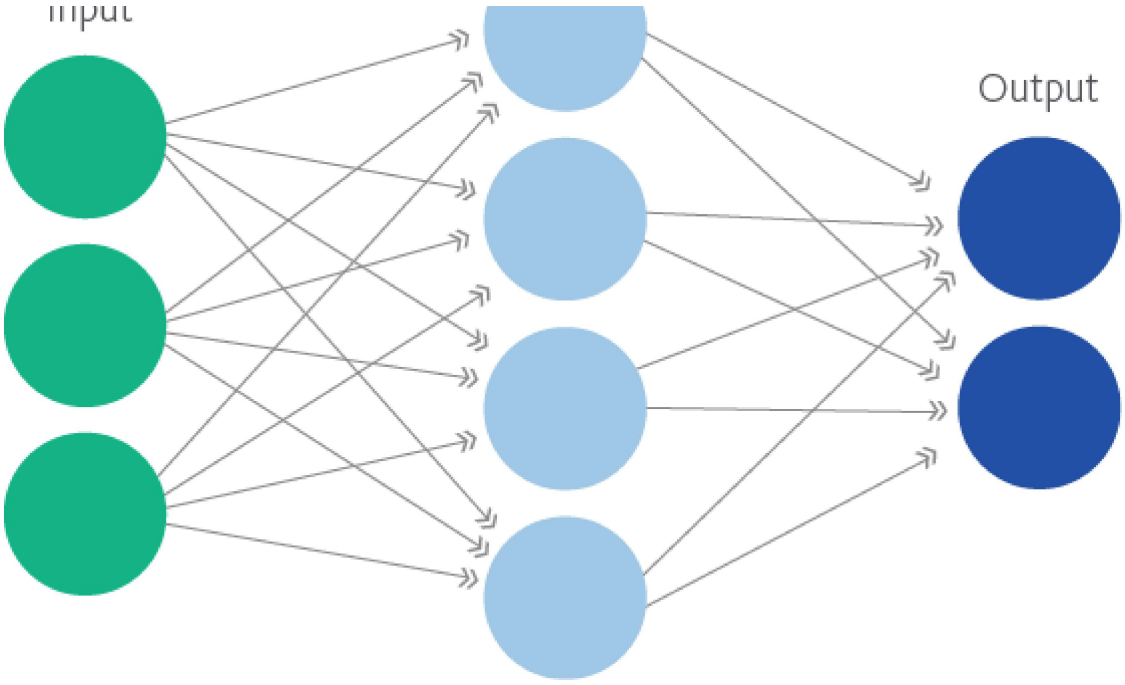
Machine learning approaches

Now let's look at three different machine learning algorithms: artificial neural networks, random forest, and boosting.

Artificial neural networks

An artificial neural network (ANN) is a mathematical simulation of a biological neural network. Its simple form is shown in Figure 2. In this example, there are three input values and two output values. Different transformations link the input values to a hidden layer, and the hidden layer to the output values. We use a back-propagation algorithm to train the ANNs on the underlying data. ANNs can easily handle the non-linear and interactive effects of the explanatory variables due to the presence of many hidden layers and neurons.

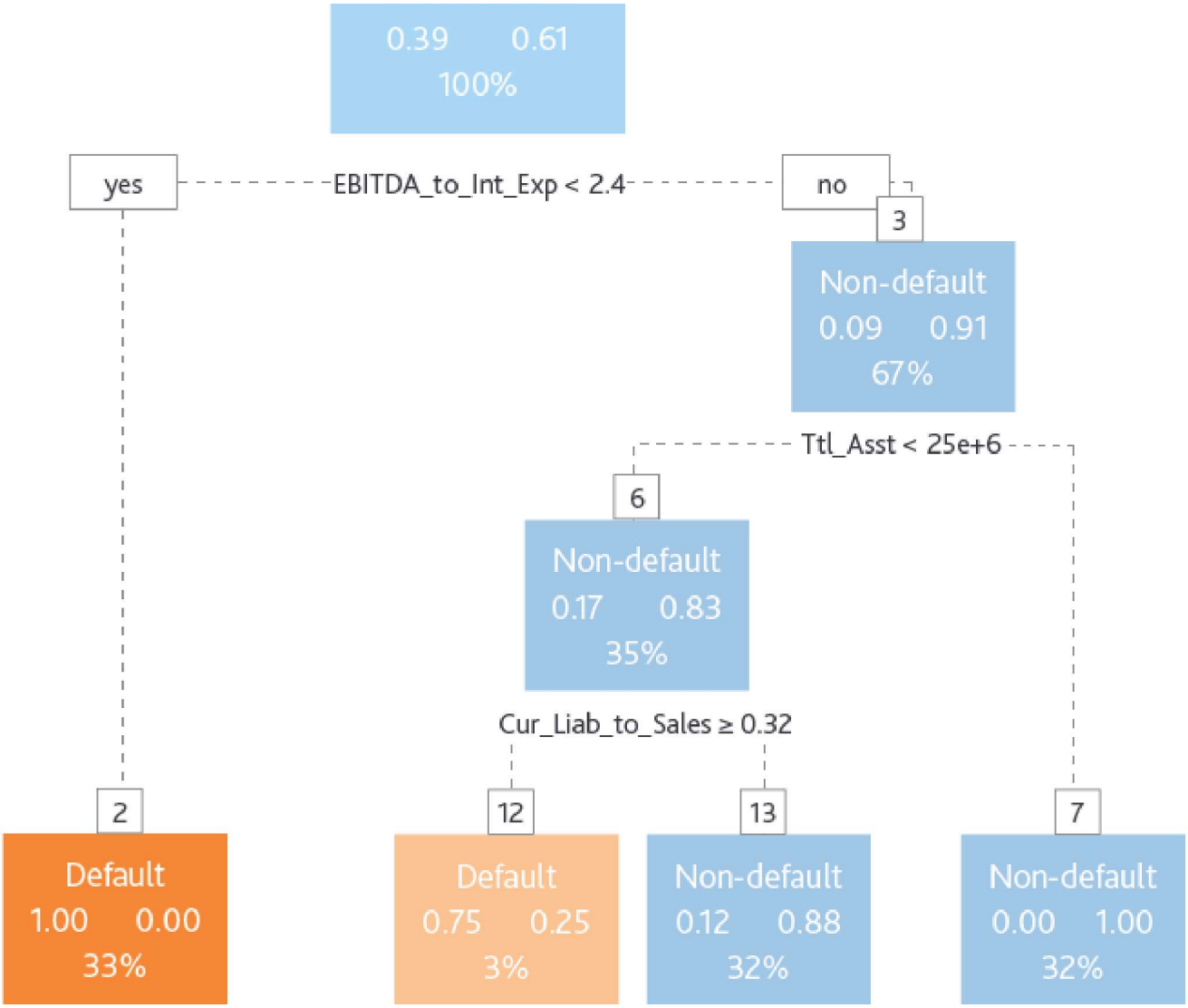
Figure 2 artificial neural network



Random forest

Random forests combine decision tree predictors, such that each tree depends on the values of a random vector sampled independently, and with the same distribution. A decision tree is the most basic unit of the random forest. In a decision tree, an input is entered at the top and, as it traverses down the tree, the data is bucketed into smaller and smaller subsets. In the example shown in Figure 3, the tree determines probability of default based on three variables: firm size; the ratio of earnings before interest, tax, depreciation, and amortization (EBITDA) to interest expense; and the ratio of current liabilities to sales. Box 1 contains the initial dataset in which 39% of the firms are defaulters and 61% are non-defaulters. Firms with EBITDA-to-interest expense ratios less than 2.4 go into Box 2. Box 2, accounting for 33% of the data, is 100% composed of defaulters. Its orange color indicates higher default risk, whereas the blue color indicates lower default risk. The random forest approach combines the predictions of many trees, and the final decision is based on the average of the output of the underlying independent decision trees. In this exercise, we use the bootstrap aggregation of several trees as an advancement to a simple tree-based model.[1]

Figure 3 random forest



Boosting

Boosting is similar to random forest, but the underlying decision trees are weighted based on their performance. Consider the parable of the blind men and the elephant, in which the men are asked to touch different parts of the elephant and then construct a full picture. The blind men are sent in six different batches. The first group is led to randomly selected spots, and each person's (partial) description is evaluated on how well it matches the actual description. This group happens to give an accurate description of only the trunk, while description of the rest of the body is inaccurate. The incomplete sections are noted, and when the second batch of blind men is led into the room, they are steered to these parts. This process is repeated for the remaining batches. Finally, the descriptions are combined additively by weighting them according to their accuracy and, in this case, the size of the body parts as well. This final description – the combination – describes the elephant quite well.

misclassified observations. This idea of giving misclassified areas additional weight (or direction while sending in a new group) is the difference between random forests and boosting.

Moody's Analytics RiskCalc model

The RiskCalc model produces expected default probabilities for private firms by estimating the impact of a set of risk drivers. It utilizes a generalized additive model (GAM) framework, in which non-linear transformations of each risk driver are assigned weights and combined into a single score. A link function then maps the combined score to a probability of default.

The RiskCalc model delivers robust performance in predicting private firm defaults. But how does it compare to other machine learning techniques? We use the three popular machine learning methods to develop new models using the RiskCalc sample as a training set. We seek to answer the following questions: Do the machine learning models outperform the RiskCalc model's GAM framework in default prediction? What are the challenges we face when using the machine learning methods for credit risk modeling? Which model is most robust? Which model is easiest to use? And what can we learn from the alternative models?

Results

Data description

To analyze the performance of these three approaches, we consider two different datasets. The first dataset comes from the Moody's Analytics Credit Research Database (CRD) which is also the validation sample for the RiskCalc US 4.0 corporate model. It utilizes only firm information and financial ratios. The second dataset adds behavioral information, which includes credit line usage, loan payment behavior, and other loan type data. This information comes from the loan accounting system (LAS), collected as part of the CRD. We want to test for additional default prediction power using the machine learning techniques and the GAM approach with both datasets. Figure 4 shows the summary of the two datasets.

Figure 4 data information

Number of firms	240,000+	101,000+
Number of defaults	16,000+	5,700+
Number of observations	1,100,000+	1,100,000+

Model performance

For both datasets, we use the GAM model's rank ordering ability as the benchmark. We measure rank ordering ability using the accuracy ratio (AR) statistic. Figure 5 shows the set of explanatory variables.

Figure 5 input variable descriptions for the PD models

Variable	Description
Firm information	Firm characteristics such as sector and geography
Financial ratios » Balance sheet » Income statement	Set of financial statement ratios constructed from the balance sheet and income statement items; the same set of input ratios used for the RiskCalc US 4.0 model are utilized here
Credit usage	Utilization on the line of credit available to the borrower
Loan payment behavior	Loan-level past due information of the borrowers over time
Loan type	Type of the loan: revolving line or term loan

Cross-validation

Because machine learning offers a high level of modeling freedom, it tends to overfit the data. A model overfits when it performs well on the training data but does not perform well on the evaluation data. A standard way to find out-of-sample prediction error is to use k-fold cross-validation (CV). In a k-fold CV, the dataset is divided into k subsets. One of the k subsets is used as the test set, and the other k-1 subsets are combined to form a training set. This process is repeated k times. If the accuracy ratio, a measure of model performance, is high for the training sample relative to the test sample, it indicates overfitting. In this case, we impose more constraints on the model and repeat cross-validation until the results are satisfactory. In this example, we use a fivefold cross validation. Figure 6 reports the average AR across the five trials.

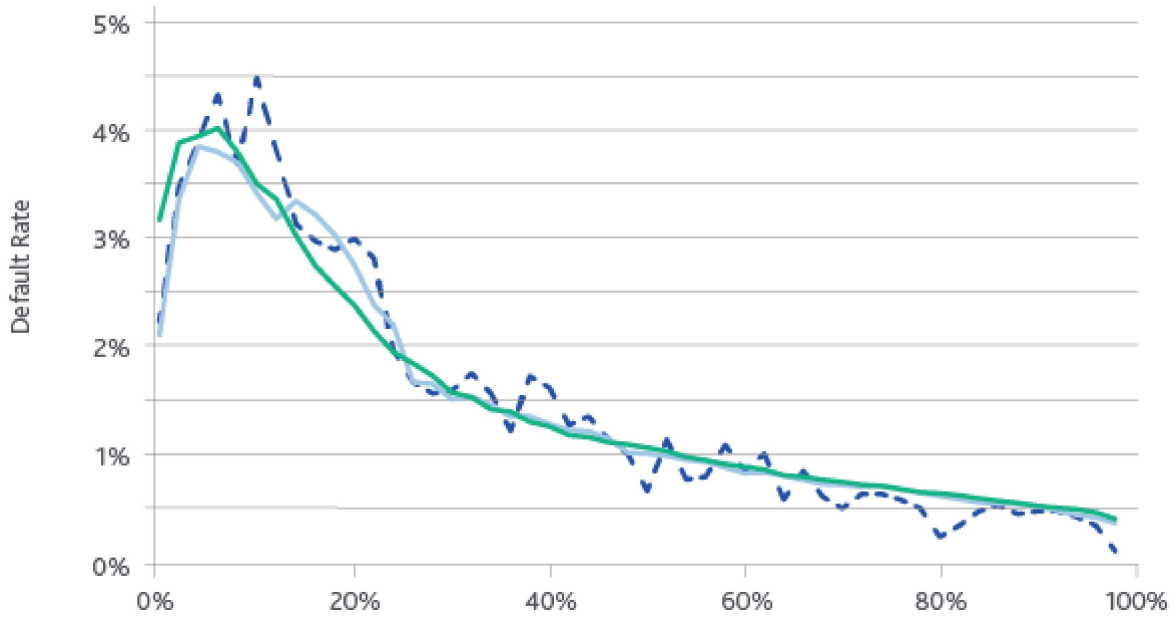
Method	1-Year Model Accuracy Ratio	
	Financial Information Only	Financials + Behavioral
RiskCalc (GAM model)	55.9%	65.8%
Random forest	58.9%	66.5%
Boosting	59.1%	67.5%
Neural network	56.6%	66.4%

We observe that machine learning models outperform the GAM model by 2 to 3 percentage points for both datasets. The accuracy ratio improves by 8 to 10 percentage points when we add loan behavioral information, regardless of the modeling approach. Credit line usage and loan payment information complement financial ratios and significantly enhance the models' ability to predict defaults.

Where machine learning excels

Machine learning methods are particularly powerful in capturing non-linear relationships. Let's take a closer look at the EBITDA-to-interest-expense ratio. Intuitively, this ratio has a non-linear relationship with default risk. In Figure 7, we divide the ratio into 50 percentiles and calculate the average values of predicted probability of default (PD) and the actual default rate. We plot this with the ratio percentiles on the x-axis and the default rate (in %) on y-axis. The default rate decreases as the ratio of EBITDA to interest expense increases. However, on the left-hand side, there is an inflection point where the EBIDTA becomes negative. When EBITDA is negative, as the interest expense decreases making the ratio more negative, the default risk should decrease. From the graph, we observe that the machine learning method of boosting provides a more accurate prediction of the actual default rate than the GAM model, especially on the left-hand side. We observe this similar behavior from the plots of other ratios, as well. Hence, we observe modest prediction improvement for machine learning methods.

Figure 7 comparing machine learning and GAM PD levels for different values of EBITDA to interest expense



Overfitting problem

Despite the use of cross-validation to minimize overfitting, machine learning models may still produce results that are difficult to interpret and defend. Figure 8 shows two cases in which the PD determined by the boosting method differs significantly from the PD determined by the GAM approach.

Figure 8 overfitting of machine learning algorithms

Ratio/PD	Case 1	Case 2
EBITDA to interest expense	4X	40X
Return on assets (ROA)	-17%	212%
Cash to assets	1%	10.5%
Debt to debt plus equity	77%	89%
Retained earnings to current liability	-6X	357X
Total assets	$500,000	$5,800,000
Boosting PD	0.2% (A3)	13.7% (Caa/C)
GAM PD	8.9% (Caa/C)	0.54% (Baa3)
Status	No default	No default

firm with high EBITDA to interest expense, high ROA, and high retained earnings is categorized as Caa/C using the boosting method. In both cases, the complex nature of the underlying algorithm makes it difficult to explain the boosting method's non-intuitive PD. The RiskCalc model's results, based on the GAM model, are much more intuitive and easier to explain.

Summary

This exercise analyzes the performance of three machine learning methods using the RiskCalc software's GAM model as a benchmark. The machine learning approaches deliver comparable accuracy ratios as the GAM model. Compared to the RiskCalc model, these alternative approaches are better equipped to capture the non-linear relationships common to credit risk.At the same time, the predictions made by the approaches are sometimes difficult to explain due to their complex "black box" nature. These machine learning models are also sensitive to outliers, resulting in an overfitting of the data and counterintuitive predictions. Additionally, and perhaps more interestingly, we find that expanding the dataset to include loan behavioral variables improves predictive power by over 10 percentage points for all modeling methods.

While the approaches we study all have their merits and have comparable accuracy levels, we believe that to improve default prediction accuracy and to expand the field of credit risk modeling in general, efforts should focus on the data dimension. Besides financial statement and loan payment behavioral data, additional information such as transactional data, social media data, geographical information, and other data can potentially add a tremendous amount of insight. We must gather more varied, non-conventional data to further refine and improve our approaches to assessing risk.

Sources

[1] Breiman, Leo. "Random Forests." *Machine Learning*, volume 45, issue 1. October 2001.

James Partridge

Risk and Accounting Solutions