機械学習による中小企業の信用スコアリングモデルの構築

Credit Scoring for SMEs Using Machine Learning Techniques

澤木　太郎[1*]　田中　拓哉[1]　笠原　亮介[1]

Taro Sawaki[1], Takuya Tanaka[1], and Ryosuke Kasahara[1]

[1]株式会社リコー
研究開発本部　リコーICT 研究所　AI 応用研究センター
[1]Ricoh Company, LTD.
Research and Development Division
Ricoh Institute of Information and Communication Technology
Applied AI Research & Development Center

Abstract: A credit scoring model is a useful tool for Small and Medium-sized Enterprises (SMEs) lending. In this study, we investigated methods to improve the accuracy of the scoring model using machine learning. As a result, we have shown that Gradient Boosting Decision Tree (GBDT) can obtain the highest accuracy. We found out that GBDT shows better performance than other methods especially when we use more than 10000 learning data. In addition, we demonstrated that ensemble learning can further improve accuracy. According to our simplified estimation, it was suggested that the ensemble learning can reduce the default rate by 16% compared with the conventional method.

１．　はじめに

企業のデフォルトリスクを推定する信用スコアリングモデルは融資を行う際の与信業務に極めて有用なツールである。特に近年は金利が低下している影響で、ますます与信コストを下げることが求められている。そのため、信用スコアリングモデルの重要性は今後さらに高まっていくと考えられる。

従来、信用スコアリングモデルは Logistic Regression(LR)のような統計的な手法が用いられてきた。金融機関は実務の信用スコアリングモデルを公表しないため、具体的にどういった手法が多く使われているかは不明であるが、Logistic Regression、もしくはLogistic Regression と Decision Tree を組み合わせたハイブリッドモデルが多いと考えられる。

一方で近年はディープラーニングを中心として、様々な機械学習の手法が提案されており、著しく精度が向上している。それにともなって、機械学習手法を使ったスコアリングモデル構築に関する報告が増えている。[1]

これらの報告の多くは、データ件数が 1000 件前後の小規模なデータセットによって検証が行われている。しかし、一般的に機械学習で高い精度を出すためには、多数のデータが必要である。また、ほとんどが個人の信用情報をもとに構築されたコンシューマ向けのスコアリングモデルであり、法人向けのスコアリングモデルに関する検証はあまり進んでいないのが現状である。海外ではわずかに報告があるものの[2][3]、国内については我々の調査した範囲ではそのような検証を行ったという報告は確認できていない。

本研究では、国内の中小企業データを用いて様々な機械学習手法によりスコアリングモデルを構築し、どのような手法が高い精度を出せるのかを調べることを目的としている。

近年注目されている代表的な機械学習手法として、Neural Network、Gradient Boosting Decision Tree、Random Forest、Support Vector Machine がある。それぞれに特徴があり、どの手法が最適なのかは、適用する分野やデータセットの特性に依存する。そこで、本研究ではまずそれぞれの手法を適用して、どの手法が高い性能を出せるのかを調査した。

機械学習では一般的に単一のモデルのみを学習するよりも、複数のモデルを組み合わせるアンサンブル学習の方が高い精度が得られる場合が多いことが知られている。もっとも単純な方法としては、各学習器の出力を平均する方法がある。他に Bagging[6]や Boosting、Stacked Generalization[7]などの手法も知られている。本研究では Stacked Generalization により複数の機械学習モデルを組み合わせることでさら

*連絡先：澤木 太郎, taroh.sawaki@nts.ricoh.co.jp

に精度を向上させることができるかどうかを検討した。

2. 方法

2.1. 使用データ

本研究で検証に用いたデータは国内の法人企業約 10 万件のデータである。本データはリコーリース株式会社から提供を受けた実務データである。同社は小口かつ大量のリースが特徴であり、データの構成は中小企業が中心になっている。

特徴量は売上高などの数値変数と業種などのカテゴリカル変数を含んでおり、合計 119 種類ある。また、各法人に対して必ずしもすべての特徴量が取得できるわけではないため欠損値を含んでいる。

2.2. 機械学習手法

本研究で検証を行った機械学習手法は以下の 3 つである。

・Gradient Boosting Decision Tree (GBDT)
・Random Forest (RF)
・Neural Network (NN)

また、ベンチマークとして Logistic Regression によるモデル構築も合わせて行った。

データは、6：3：1 の割合でトレーニング用、バリデーション用、テスト用データに分割した。トレーニング用データは学習に使用し、バリデーション用データは後述のパラメータ最適化に使用した。テスト用データは性能の評価のみに用いた。

各機械学習手法にはハイパーパラメータが複数存在する。最適なパラメータはデータセットによって異なるため、パラメータの最適化を行う必要がある。パラメータ探索の手法としてはいくつかの手法が知られているが、本研究では Bayesian Optimization [4] を採用した。

学習は企業が融資実行後にデフォルトしたかどうかを示すフラグを教師データとして学習を行った。

アンサンブル学習には Stacked Generalization を用いた。Stacked Generalization では異なる機械学習モデルを多層にしてアンサンブルする手法である。学習に使用するデータを n-fold に分割し、各機械学習手法によって、out-of-fold データに対する予測値を算出する。こうして得られた予測値を 1 層目の出力として、次の層の特徴量として入力する。本研究では、図 1 に示す層構成で学習を行なった。2 層目の出力の平均値を最終的な出力とした。

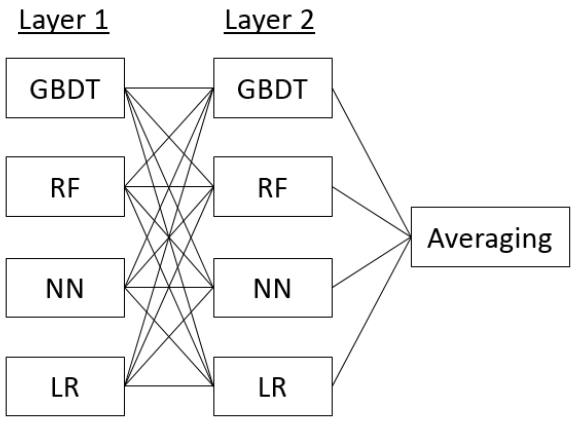

図 1　Stacked Generalization の構成

2.3. 評価方法

テストデータに学習で生成したスコアリングモデルを適用して得られた予測結果の評価は CAP 曲線と AR(Accuracy Ratio)値を使用する[5]。CAP 曲線は横軸に推定デフォルト確率の上位 x 件の全体に占める割合 x/N を、縦軸に推計デフォルト確率の高い上位 x 件のうち実際にデフォルトした件数 N_x の割合 N_x/N_d をプロットしたものである。ここで、N は評価に使用したデータの総数、N_d は評価データの中のデフォルトした件数の総数である。CAP 曲線の例を図 2 に示す。B のような曲線が典型的な例である。モデルの説明力が全くない場合は C の直線を描き、予測が完全に正解していれば A のような形を描く。

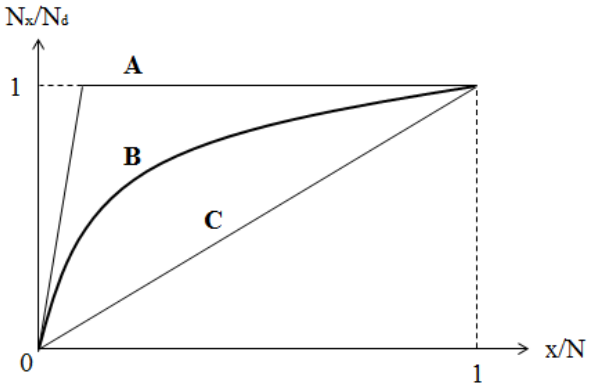

図 2　CAP 曲線

ここで、CAP 曲線が B の曲線であるとしたとき、（B と C で囲まれた部分の面積）/（A と C で囲まれた部分の面積）の比が AR 値である。

3. 結果と考察

各手法で学習した結果を表 1 に示す。これらの結果は各手法をアンサンブルせずに単独で用いた結果である。

表 1 単モデルの性能

手法	AR 値	
	Validation	Test
GBDT	0.621	0.619
Random forest	0.535	0.539
Neural network	0.528	0.530
Logistic regression	0.545	0.523

他の手法と比較して、GBDT が著しく高い精度となっており、テストデータに対して AR 値が 0.619 だった。それに対して Random Forest と Neural Network は既存手法である Logistic Regression と比べてほとんど精度に差が出ない結果となった。GBDT はその他の手法と比較して 0.1 程度上回っているため、信用スコアリングモデルの構築に適したアルゴリズムだと考えられる。

一般的に機械学習はデータ数が多いほど精度が向上する。特に Boosting や Neural Network は十分な精度を得るために大量のデータが必要となることも多い。今回は約 9 万件という大量のデータを学習に使用したが、データ数が少ない場合には精度が下がり、異なる結果になる可能性がある。そこで、学習に使用するデータ数が精度にどのような影響を及ぼすかを調べるため、学習データを変えて学習・予測を行った結果が図 3 である。

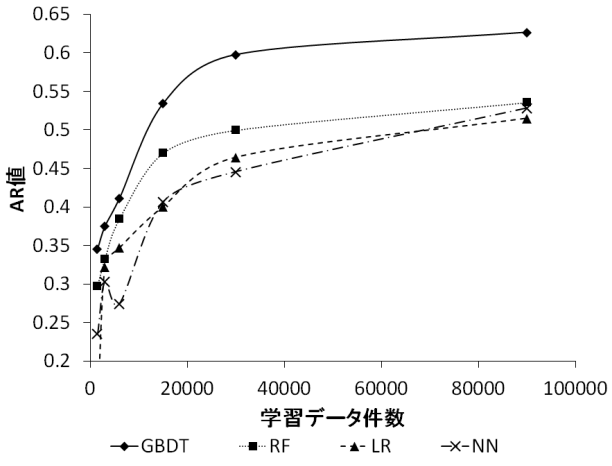

図 3 学習データ件数と精度の関係

横軸がトレーニング用データの件数、縦軸が AR 値である。いずれの手法も学習データ数が増えるほど精度が高くなる。データ数が少ない場合も多い場合も、総じて GBDT の精度が高い。特にデータ数が 1 万件を超えたあたりから他の手法との差が広がっている。一方で学習データが少ない場合には、GBDT と Random Forest はあまり精度に差がない。GBDT の優位性を活かすためには、一定以上のデータ数が必要だと考えられる。

Stacked Generalization を用いたアンサンブル学習の結果を表 2 に示す。

表 2 アンサンブル学習の結果

手法	AR 値	
	Validation	Test
GBDT	0.621	0.619
アンサンブル学習	0.638	0.630

アンサンブル学習の方が GBDT 単体と比較してテストデータにおける AR 値が 0.011 高く、アンサンブル学習の有効性が示された。Stacked Generalization は層数や学習器の種類などがハイパーパラメータになっており、多数のバリエーションが考えられる。今回結果を示したのはあくまでもその一例であるため、さらに精度を高めることができる可能性がある。

従来の手法である Logistic Regression をアンサンブル学習に置き換えることで得られる経済効果を考察するため、それぞれの CAP 曲線を図 4、図 5 に示す。

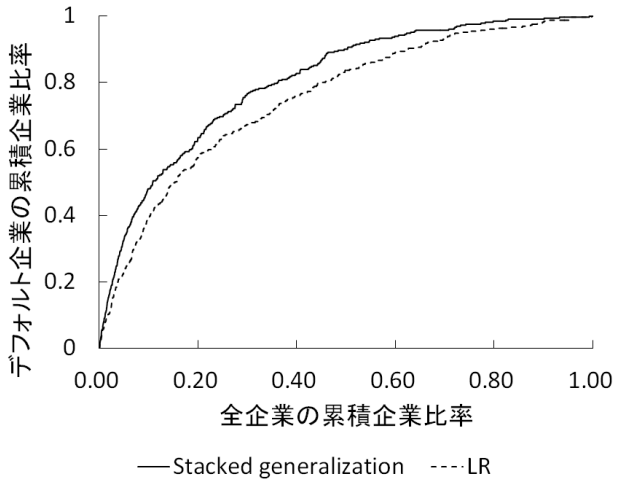

図 4 アンサンブル学習と LR の CAP 曲線

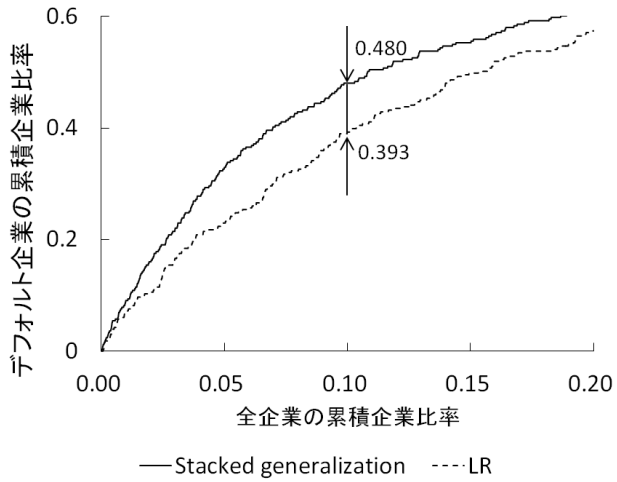

図 5　アンサンブル学習と LR の CAP 曲線　（拡大）

仮にデフォルト率が上位 10%の法人企業を謝絶したケースを考えると、横軸が 0.1 のときの CAP 曲線の縦軸の値が全デフォルト企業のうち何%を謝絶できたかを示す値となる。Logistic Regression によりスコアリングモデルを構築した場合、横軸が 0.1 のときの縦軸の値は 0.393 なので、全デフォルト企業のうち 39.3%を謝絶したことになる。逆に言えば残りの 61.7%の法人については実際にデフォルトしてしまうことになる。一方で、アンサンブル学習を行なってスコアリングモデルを構築した場合には、全デフォルト企業のうちの 48%を謝絶できるためデフォルト率を低く抑えることが可能である。定量的に比較すると、アンサンブル学習の方が従来手法である Logistic Regression よりも、約 16%デフォルトを少なく抑えられる。

4．　まとめ

本研究では、国内の中小企業を中心とした法人への与信データ約 10 万件を使用して、機械学習による信用スコアリングモデルの構築を行なった。GBDT、Random Forest、Neural Network の 3 種類の手法を用いて精度を比較したところ、GBDT がもっとも高い精度を得られることが分かった。GBDT は従来手法である Logistic Regression と比べると、AR 値が約 0.1 高く、信用スコアリングモデルの構築において極めて有効な手法であることが示された。

学習に使用するデータ件数と精度の関係についても調べたところ、学習データが少ない場合には GBDT と Random Forest は同程度の精度だが、学習データが 1 万件以上の場合に GBDT のほうが大きく精度が高くなることがわかった。このことから、GBDT の優位性を活かすためには、大量のデータが必要である。

最後にアンサンブル学習の一つの例として、Stacked Generalization を用いて、複数の機械学習モデルをアンサンブルしたところ、単体の GBDT よりも AR 値でさらに 0.01 程度高い精度を得られた。

本研究では、法人企業の信用スコアリングモデル構築において、GBDT を含めた機械学習モデルをアンサンブルすることによって、Logistic Regression により構築された従来のモデルと比較して、大きく精度を向上できることを示した。簡易的な試算では、約 16%デフォルトを低減する効果があることが示された。

謝辞

本研究では、リコーリース株式会社に提供していただいたデータセットを利用した。非常に貴重なデータを提供していただいたリコーリース株式会社に感謝の意を表する。

参考文献

[1] S. Lessmanna, B. Baesensb, H. Seowd, and L. C. Thomas: Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research, European Journal of Operational Research, vol. 247, No.1, pp. 124-136, (2015)

[2] D. Fantazzini, and S. Figini: Random Survival Forests Models for SME Credit Risk Measurement, Methodology and Computing in Applied Probability, Vol. 11, No. 1, pp. 29-45, (2009)

[3] H. S. Kim, and S. Y. Sohn: Support vector machines for default prediction of SMEs based on technology credit, European Journal of Operational Research, Vol. 201, No.3, pp. 838-846, (2009)

[4] J.Snek, H. Larochelle, and R. P. Adams: Practical Bayesian Optimization of Machine Learning Algorithms, Advances in Neural Information Processing Systems, Vol. 25, (2012)

[5] 山下智志, 川口昇, 敦賀智裕: 信用リスクモデルの評価方法に関する考察と比較, Financial Research and Training Center discussion paper series, 11, (2003)

[6] L. Breiman: Bagging predictors, Machine Learning, Vol.24, No. 2, pp. 123-140, (1996)

[7] D. H. Wolpert, Stacked generalization, Neural Networks, Vol. 5, No. 2, pp. 241-259, (1992)