機械学習を用いた倒産予知モデルの研究

片桐　英樹*　平井　裕久*　松丸　正延**

Bankruptcy prediction model using machine learning

Hideki KATAGIRI*　Hirohisa　HIRAI*　Masanobu MATSUMRU**

１．緒言

本研究では，倒産予知と関連の深い企業格付に焦点を当て，機械学習を用いた格付推計モデルを提案する．一般に，倒産予知モデルにおいて，企業の倒産あるいは倒産はそれぞれ「1」,「0」の 2 段階の 2 値で判断される．しかし，2 段階の情報では，企業の経営状態を十分に表現できない場合もある．そこで，倒産予知モデルを多段階に展開し，ランキングすることによって，企業の財務状況を知ることは，経営活動を行う企業にとって有益である．この多段階に展開してランキングすることが格付の考え方に相当する．

本研究では，eXtreme Gradient Boosting (XGBoost) [1] , LightGBM, CatBoost [2]の 3 種類の決定木を用いた勾配ブースティングを用いた企業格付推計モデルの構築を行う． これまでに財務データと投融資ネットワークを用いた　3　種類の決定木を用いた勾配ブースティングと Support Vector Machine (SVM) による企業格付推計モデルは研究されていない．本研究では日経 NEEDS-Financial QUEST2.0 から収集した財務データを用いた数値実験を行い，各モデルの性能と比較することにより，各モデルの有用性について検証する．また, SVMも構築し，勾配ブースティングモデルとの比較を行う．

２．先行研究

Daoud [3]は，ホームクレジットデータセットを用いて，CPU の実行時間と精度の観点から，XGBoost, LightGBM, CatBoost の比較を行っている．LightGBM は他の勾配ブースティングよりも高速であり，正確であると結論付けている．田中ら[4]は財務指標とニューラル・ネットワークによる格付推計モデルの提案を行った．金，松尾[5]らは，ニュース記事に基づく企業間ネットワーク関係から得られる情報をもとに，企業価値を予測する新しい方法を提案した．

３．本研究の企業格付推計モデルとその特徴

本研究では, 日本企業の財務データと投融資ネットワークを用いた 3 種類の決定木を用いた勾配ブースティングと SVM による企業格付推計モデルを開発する. モデル開発においては変数選択を実施し，特徴集合において予知に寄与する部分集合のみを選択する．また，交差検証を行い，訓練データと評価用の検証データを分割して性能を計測することで，特定のデータに拠らない汎化能力の高いモデルを得る．格付推計には財務データを用いることが主流であるが，企業業績は短期的な評価のみを表し，長期的な視点は欠けていると考えられる．

そこで，従来モデルの欠点を補うために，長期的な企業間の関係性を導入する．具体的には大株主および企業保有株式株主との長期的な信頼関係性を表す株式保有関係に注目し、投資ネットワーク指標を計算する．しかしながら企業間の関係性の把握は投資関係の導入だけでは十分とは言えない．投資の資金調達を考慮する必要性から，金融機関などの借入金関係の融資ネットワーク指標を計算し，長期的な信頼関係性をとらえることにした．

４．分析手法

４.１　格付推計モデルに用いる機械学習アルゴリズム

財務指標とネットワーク指標を合わせた格付推計モデルを構築し，格付の推計を行う．モデルの構築にはXGBoost, LightGBM, CatBoost の 3 種類の決定木を用いた勾配ブースティングを用いる．

４.２　モデルの性能評価と検証

モデルの評価指標としては, 実際の格付とモデルの推計値が一致した割合の正確度を示す Accuracy と不一致度のペナルティを表す Quadratic Weighted Kappa（QWK）の 2 種類を使用する．QWK は分類するクラス間に順序関係がある場合に用いられる評価指標で，真の値から大きく外れるほど大きなペナルティが科される．
また，検証には，個々のモデルの汎化性能を評価する交差検証（10分割）と機械学習のハイパーパラメータ探索の方法であるグリッドセーチ（grid search）を用いた．

４.３　使用する投融資ネットワーク指標

ネットワーク指標は，金，松尾らの研究と大西の研究[6]で使用されているものを引用し，次数中心性，近接中心性，媒介中心性，平均パス長，ネットワーク密度，クラスタ係数，オーソリティ度，ハブ度，ページランクを用いる．

５.実証実験

５.１　分析データ

格付推計モデルの特徴量は財務指標とネットワーク指標とする. 財

*教授　経営工学科
Professor, Dept. of Industrial Engineering and Management
**客員教授　工学研究所
Visiting Professor, Research Institute for Engineering

務指標は日経 NEEDS-FinancialQUEST から取得した 149 指標とした. また，大株主および企業保有株式のデータを取得し，18 種類の投融資ネットワーク指標を計算した．さらに，推計の対象とする格付データを日本格付研究所（JCR）から取得し，これを格付推計モデルの教師データとした. 具体的には JCR から 2000 年から 2020 年までの 21 年期間の農林水産，金融，その他の業種を除く業種を対象に取得した格付データは 3298 社（製造業 1,741 社，非製造業 1,556 社）であった．上場していない企業については大株主および企業保有株主の株式数が取得できない場合があるため，最終的に取得できた企業数は 1,974 社（製造業 1,135 社，非製造業 839 社）となった.

５.２ 実証実験の結果と検討・考察

XGBoost, LightGBM, CatBoost の 3 種類の決定木を用いた勾配ブースティングを用いた結果を表 1 に示す. 149 種類の財務指標と 18 種類の投融資ネットワーク指標を用いた合計 167 指標の基本モデルと 149 種類の財務指標のみを用いた財務モデルの比較表である．表 1における評価指標の Accuracy と QWK の値から，基本モデルは財務モデルよりも数値が良く，投融資関係を導入したモデルがより正確に格付を推計できていることがわかる． また，Accuracy と QWK の評価指標について, 3 種類の勾配ブースティングは SVM よりも良い結果を示した． 3 種類の勾配ブースティングの中では, LightGBM が XGBoost と CatBoost よりも良い値を示した．この結果は先行研究で述べた Daoud[1]の結果と同様であり，LightGBM は他の勾配ブースティングよりも高精度なアルゴリズムと結論付けることができる.計算時間については紙面の制約上割愛するが，計算時間においても LightGBM は他の勾配ブースティングよりも高速であった. CatBoost は他のアルゴリズムに比較すると計算時間が長く， 3 種類の勾配ブースティングの中では非効率なアルゴリズムである.

表 1 全業種の精度比較

手法	モデル	Accuracy	QWK
SVM	**基本モデル**	0.6956	0.8606
	財務モデル	0.6753	0.8621
XGBoost	**基本モデル**	0.7371	0.9053
	財務モデル	0.7178	0.8850
LightGBM	**基本モデル**	0.7513	0.9097
	財務モデル	0.7391	0.8987
CatBoost	**基本モデル**	0.7396	0.9038
	財務モデル	0.7315	0.8991

図 1 は基本モデルの混同行列（LightGBM）を，例として示した. 縦軸が実際の格付，横軸が本研究のモデルの推計結果を示している. 対角線上の企業数の合計は 1,483 社であり，全企業数 1,974 社で割った値が表 1 の Accuracy の 0.7513 である.

6. 結言

本研究では, 勾配ブースティングと投融資ネットワーク指標を用いた格付推計モデルを提案し, 企業の財務データの数値実験により提案モデルの有用性を検証した．具体的には，格付推計に投融資

	AAA	AA+	AA	AA-	A+	A	A-	BBB+	BBB	BBB-	BB+	BB	BB-	B+	B	B-	CCC+	CCC	CCC-	CC	C	D
AAA	41	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
AA+	2	6	4	4	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AA	0	1	117	16	4	2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
AA-	1	1	5	159	16	4	4	1	3	0	0	0	0	0	0	0	0	0	0	0	0	0
A+	0	0	3	8	234	42	11	2	3	0	0	0	0	0	0	0	0	0	0	0	0	0
A	0	0	2	4	35	300	38	10	1	0	0	0	0	0	0	0	0	0	0	0	0	0
A-	0	0	0	3	9	38	292	27	1	1	0	0	0	0	0	0	0	0	0	0	0	0
BBB+	0	0	0	1	5	14	38	178	18	1	0	0	0	0	0	0	0	0	0	0	0	0
BBB	0	0	0	3	4	2	8	38	128	9	0	0	0	0	0	0	0	0	0	0	0	0
BBB-	0	0	0	0	0	1	2	4	22	28	0	1	0	0	0	0	0	0	0	0	0	1
BB+	0	0	0	0	0	0	1	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0
BB	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
BB-	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B+	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
B	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B-	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CCC+	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CCC	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CCC-	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CC	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
D	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0

図 1 基本モデルの混同行列（LightGBM）

ネットワーク指標を組み込んだモデルを提案し, 実証実験により精度の向上と投融資ネットワーク指標の重要性を明確にした. 3 種類の決定木を用いた勾配ブースティングは，評価指標の Accuracy と QWK において SVM よりも優れた結果を示した. 3 種類の勾配ブースティングについて， CPU の実行時間と精度の観点から評価すると, LightGBM は，他の勾配ブースティングよりも大幅に高速であり，Accuracy と QWKにおいても XGBoost と CatBoost よりもより精度の良い値を示した.

今後の研究として，自然言語処理を用いて有価証券報告書から倒産を表す言語を抽出し,モデルに組み込むことを計画している.

参考文献

[1] T. Chen, C. Guestrin, XGBoost: A Scalable Tree Boosting System, KDD ’16, August 13-17, 2016, San Francisco, CA, USA (KDD 2016, oral presentation)

[2] A. V. Dorogush, V. Ershov, A. Gulin, CatBoost: gradient boosting with categorical features support, https://www.research-gate.net/publication/328576065(参照 2021-12-7)

[3] E. A. Daoud, Comparison between XGBoost, LightGBM and CatBoost Using a Home Credit dataset, World Academy of Science, Engineering and Technology, International Journal of Computer and Information Engineering, Vol.13, No.1, 2019

[4] 田中克明，勝田英紀，萩原統宏，ニューラル・ネットワークによる格付付与構造の安定性について，経営情報研究，第 17 巻，第 1 号，pp.17-32 （2009）

[5] 金英子，Ching-Yung Lin，松尾豊，石塚満，動的ネットワークのマイニングと企業価値の予測，第 25 回人工知能学会全国大会論文集，pp.1-4 （2011）

[6] 大西立顕，企業間取引の大規模ネットワーク構造からみた企業の特徴，ttps://www.cc.u-tokyo.ac.jp/public/VOL12/special/201002SP-ohnishi.pdf （参照 2020-5-22）