

機械学習の手法に基づく
社債スプレッド予測モデル構築の試み

新	原	俊	樹
原	口	健	太郎
丹	波	靖	博
阿	部	修	司
大	石	桂	一

西南学院大学商学論集
第71巻 第1・2合併号 抜刷
2024（令和6）年9月発行

機械学習の手法に基づく 社債スプレッド予測モデル構築の試み

新 原 俊 樹*
 原 口 健太郎**
 丹 波 靖 博***
 阿 部 修 司****
 大 石 桂 一*****

1. はじめに

社債市場は株式市場と並んで会計学やファイナンスにおける重要な研究対象であるとともに、資金調達や投資対象として実務上も大きな役割を果たす。社債市場において特に中心的な研究テーマとなるのが、社債金利の決定要因の解明である（信用リスク分析）。

社債金利は、社債の償還残存年限をはじめ、企業が開示する会計情報や株式市場の動向等、多種多様な要因に左右される。先行研究では、社債金利の推定には、社債金利を目的変数とし、複数の決定要因を説明変数とする最小二乗法（OLS）による重回帰分析を利用する事例が見られる。しかしながら、OLSによる社債金利の推定には限界が存在する。最も重要な問題は、モデル構築において線形の関係を仮定せざるを得ないことである。社債金利と償還残存年限とのプロット（イールドカーブ）は日々その姿を変える未知の曲線であり、線形でのモデル化は極めて困難である。このため、OLSによる金利推定の精度は十分とは言えず、決定係数は低くなる。

金利の高精度の予測は今後の研究進展の基盤として位置づけられる技術

* 西南学院大学情報処理センター 助教

** 西南学院大学商学部 准教授

*** 西南学院大学経済学部 教授

**** 九州大学国際宇宙惑星環境研究センター 学術研究員

***** 九州大学経済学研究院 教授

であるとともに、投資戦略等に影響をもたらすことから実務上も重要であるにもかかわらず、予測精度が不十分であることは重要な問題である。

本稿の目的は、近時、データサイエンス領域で特に急速な発展を遂げている機械学習を用いてこれらの課題の克服を試みることにある。機械学習は、大規模データセットを入力することで説明変数と目的変数との非線形の関係性を予測することができるかとされており、これを活用することで、OLSをはじめとした従来の統計モデルよりも高い精度で社債金利の予測を実現できる可能性がある。

一方で、次章以降で述べるように、機械学習のモデルを構築するためには多くの種類が存在する「分類器」の中から適切なものを選定する必要がある、分類器の種類によって予測精度が大きく変動することが知られている。したがって、機械学習を社債金利の予測に活用するためには、まず、社債市場において有効に機能する分類器を選定する必要がある。

そこで、本稿では、わが国における2022年の社債市場を対象として、機械学習を用いた社債金利予測モデルを構築したうえで、18種類の分類器をそれぞれ試行し、予測結果を比較することで、最も高い予測精度をもたらす分類器が何かを明らかにする。さらに、比較結果に基づき、適切な分類器を用いることにより、わが国の社債市場において機械学習による金利予測が有効に機能し、OLSを十分に上回る予測精度が期待できることを明らかにする。

2. 関連研究

社債の信用リスク分析に関して、ファイナンス領域で豊富な研究蓄積が存在する。これらの研究は、定量評価モデルの観点からいくつかのアプローチに区分される。本稿では主要な3つのアプローチについて述べる。第1に、社債の信用リスクの理論的推定における初期の研究として、実証面でも貢献したAltman (1968) のZスコア算出モデルが挙げられる。本モデルは財務データや財務比率、企業属性などの企業情報を説明変数として、企業の良し悪しを表すZスコアを算出したモデルである。本モデルに

は非線形回帰モデルであるロジットモデルやプロビットモデルなどのモデルタイプがあり、判別分析とも呼ばれている。同研究から派生して、デフォルト/非デフォルト推定、多段階格付け推定、デフォルト率算出モデルなど、多様な信用リスク指標に関する定量評価モデルの研究が行われてきた。財務データや企業属性などの企業情報を説明変数、目的変数を地方債スプレッドとして、線形重回帰推定を行う方法もこのモデルに分類される。説明変数と目的変数の因果関係が明確であることや、モデルの構造がシンプルで扱いやすいことなどが本モデルの長所である。

第2に、Black and Scholes (1973) , Merton (1974) などにより提唱された構造モデル (Structural approach) が挙げられる。これらのモデルは、オプション理論に基づき、企業の資産価値を原資産と考え、資産が減少し負債額を下回ったときにデフォルトが発生すると定義してデフォルト率を算出している。また、社債の価格は算出された企業の資産価値とデフォルト確率から計算される。構造モデルは企業の財務構成を考慮してデフォルトが定義されているため、直感的な理解と説明が明確な点が利点となっている。一方、実際の信用リスクを正確に反映しにくいなど、実証面での課題が指摘されている。本モデルはデフォルトリスクのみを考慮し、金利リスクなど他のリスク要因は考慮されていない。この問題を解決するため、Longstaff and Schwartz (1995) は金利リスクやデフォルトリスクのリスク要因を考慮したモデルを提案し、デフォルト時の損失率を確率過程として、より現実的なモデル化を行っている。

第3に、Jarrow and Turnbull (1995) , Duffie and Singleton (1999) などが提唱した誘導モデル (Reduced-form approach) が挙げられる。誘導モデルでは、デフォルト事象を外生的に捉え、デフォルト確率は社債市場データから推定される。つまり、デフォルト時の損失率を一定と仮定し、デフォルト確率からリスク中立評価で社債価格を導出する。さらにDuffie and Singleton (1999) は、ハザードレートの概念を用いてデフォルト確率を時間の経過とともに変化する拡張モデルを提案した。また、Collin-Dufresne *et al.* (2001) では、社債スプレッドの変動要因には他の市場とは独立した

要因があることが指摘され、流動性プレミアムが社債スプレッドの大きな要因であることが示された。誘導モデルでは企業の財務状況などは考慮されないため、ファンダメンタルズと信用リスクの因果関係が不明である点が短所であるが、社債市場のデータから、デフォルト率を算出するための計算を行うことができる長所がある。つまり、誘導モデルにおいて企業は一定の確率でデフォルトすると仮定され、そのデフォルト理由も考慮していないという特徴がある。

日本の社債市場についても、白須・米澤（2008）は金利リスクと流動性要因が社債スプレッドの変動を左右することを実証している。これらの研究により、社債スプレッドは企業の信用リスクだけでなく、流動性プレミアムや市場の需給環境など様々な要因に影響を受けることが明らかになっている。

一方で、社債金利は会計学領域でも研究対象となる。代表的な研究として、企業が開示する会計情報の質が社債金利に与える影響を分析した Sengupta（1998）などがある。

本稿で構築する機械学習モデルは、このように多様な領域で興味の対象となる社債金利に関して高精度の予測を可能とし、さらなる研究発展の基盤となる知見を提供するものである。

3. データと解析手法

3-1. データベースの構築

解析に先立ち、社債の流通市場（Secondary market）に着目し、この市場で流通する各債券の取引情報と、当該債券の発行企業の財務情報や株式情報、市場指標、外国為替相場等の周辺情報を収集・統合して、解析に必要なデータベースを構築した（表1）。以下、データベースの構築に当たって収集した各情報の概要を説明する。データベースの1レコード（1行）には、ある取引日における特定の債券の価格等の情報と、その取引日時時点で知り得た最新の周辺情報を収録している。

表1 データベースの概要

債券コード	取引日	償還日	単利最終利回り	国債金利	企業の財務情報			企業の株式情報	東証株価指数	外国為替公示相場
					(直近)	(1期前)	(2期前)			
B51...	2022/03/31	2036/06/20	0.820	0.500						
B51...	2022/03/31	2036/06/20	0.825	0.499						
B51...	2022/03/31	2036/06/20	0.775	0.460						
B51...	2022/03/31	2036/06/20	0.730	0.418						
B51...	2022/04/01	2036/06/20	0.735	0.429		(3)	(4)	(5)	(6)	
B51...	2022/04/04	2036/06/20	0.745	0.438						
B51...	2022/04/05	2036/06/20	0.745	0.432						
B51...	2022/04/06	2036/06/20	0.785	0.462						
B51...	2022/04/07	2036/06/20	0.770	0.454						
B51...	2022/04/08	2036/06/20	0.745	0.433						

(1)
(2)

(1) 流通市場における各債券の日々価格

日本経済新聞社が提供する経済データサービス「日経NEEDS JSPRICE」から、過去30年間（1993年4月～2023年3月）に流通市場で取引された一般事業債と電力債（公募債に限定）の全銘柄（8,837銘柄）の日々価格等に関するデータを入手した。JSPRICEは、公募社債のみならず非公募債の情報も提供する点で網羅性に優れており、本稿の分析に適している。

(2) 国債金利

財務省が公開している1974年9月以降の日々の国債金利を取得した。国債金利は最短1年から最長40年まで年単位の離散的な値として得られるが、各債券の金利との差（スプレッド）を求めるためには、当該債券の償還日までの残存日数（償還日－取引日）に対応した国債金利の値が必要となる。そこで、スプライン補間を用いて日単位で国債金利の値を算出し、ある取引日における特定の債券について、その残存日数に対応した金利を取得した。このとき、財務省が公開する国債金利の情報だけでは1年未満の期間の補間ができないため、別途、日本銀行が公開している無担保コール（オーバーナイト物）レートを残存日数1日の国債金利として採用した。ある取引日における特定の債券の残存日数に対応した国債金利の算出の概要を図1に示す。

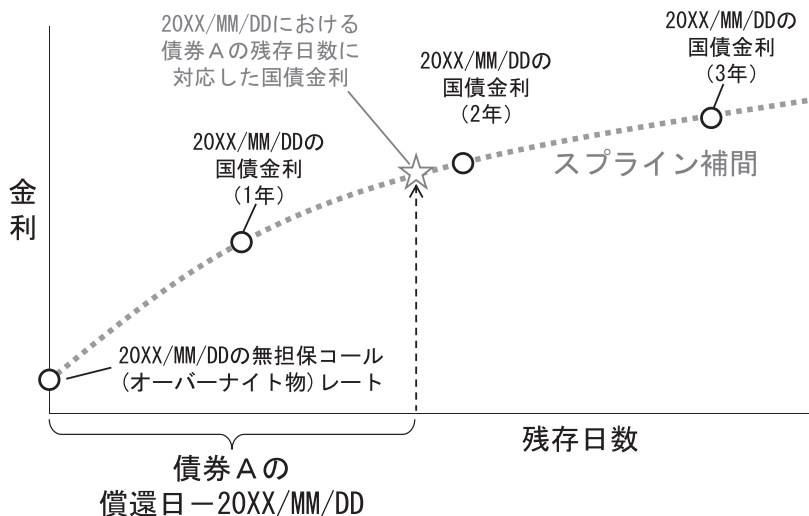


図1 残存日数に対応した国債金利の算出方法

(3) 企業の財務情報

(1) の8,837銘柄を発行した525社が過去10年間（2013年4月～2023年3月）に作成・開示してきた有価証券報告書の記載情報を「日経NEEDS」から入手した。そのうえで、各債券の取引日から過去に遡り、当該債券を発行した企業の直近3期分の財務情報（貸借対照表、損益計算書、キャッシュ・フロー計算書の記載内容）を参照した。財務情報は、連結財務諸表が公表されている場合には連結財務諸表の値を用いた。日本の上場企業は日本会計基準、国際会計基準、米国会計基準のいずれかを採用しているが、本稿ではすべての企業を分析対象とした。

(4) 企業の株式情報

(3) と同じく、525社の過去10年間の日々の株価等の株式情報を「日経NEEDS」から入手し、各債券の取引日に対応した情報を参照した。

(5) 東証株価指数

市場指標として、過去20年間の東証株価指数（TOPIX）を「日経

NEEDS」から入手し、各債券の取引日に対応した値（日次四本値の終値）を参照した。

（6）外国為替公示相場

みずほ銀行による2002年以降の外国為替公示相場を入手し、各債券の取引日に対応した米ドル、英ポンド、ユーロの値を参照した。

これら（1）～（6）のすべてを入手できた10年間（2013年4月～2023年3月）の各情報を結合してデータベースとした。

3-2. 説明変数と目的変数の指定

前節で結合したデータベースは1,500列にも上り、本稿の解析には明らかに不要な列も多く含まれている。そこで、目的変数と説明変数を整理し、データベースから必要な列を抽出した。まず、本稿の目的は社債スプレッド（社債の利回りと国債金利との格差）の予測であることから、目的変数は、各債券の「複利最終利回り」－「当該債券の残存日数に対応した国債金利」である。このスプレッドを予測するための説明変数として、データベース内の各列から表2に示す201個の変数を抽出（又は算出）した。

3-3. データの前処理

本稿ではまず2022年の1年間のデータのみを採用し、当該期間中のデータから説明変数と目的変数を抽出した。抽出したデータは692,542件に上るが、欠損値や異常値を含んでいるため、機械学習による解析が可能になるように事前の加工が必要となる。まず、欠損値を1つでも含むレコードを削除し、278,916件とした。次に、変数ごとに平均値と標準偏差（ s ）を求め、いずれかの変数において平均値から $3s$ 以上離れた値を含むレコードを除去し、167,632件とした。さらに、予測モデルの構築に係る計算コストを低減させるため、本稿ではこのデータから無作為に10,000件を抽出した。また、抽出後のデータの説明変数の値をすべて標準化した。

表2 企業の財務情報・株式情報等から抽出した説明変数

説明変数	備考	参照情報	説明変数	備考	参照情報
x001	表面利率	(1)	x159	株価移動平均75日	
x002	資産合計		x160	株価移動平均100日	
x003	現金、預金、現金及び現金同等物		x161	株価移動平均200日	
x004	長期借入金、社債、転換社債		x162	株価 <i>かい</i> 離率5日	
x005	短期借入金、社債合計		x163	株価 <i>かい</i> 離率6日	
x006	減価償却費		x164	株価 <i>かい</i> 離率25日	
x007	自己資本、親会社の所有者に帰属する資本		x165	株価 <i>かい</i> 離率75日	
x008	投資活動によるキャッシュフロー		x166	株価 <i>かい</i> 離率100日	
x009	支払利息、割引料		x167	株価 <i>かい</i> 離率200日	
x010	流動負債	(3)	x168	売買高移動平均5日	
x011	流動資産		x169	売買高移動平均6日	
x012	親会社株主に帰属する当期純利益		x170	売買高移動平均25日	
x013	営業活動によるキャッシュフロー		x171	売買代金、1	
x014	営業利益		x172	売買代金移動平均5日	
x015	営業外収益		x173	売買代金移動平均6日	
x016	売上高、営業収益		x174	売買代金移動平均25日	
x017	～		x175	基準値	
x134	x002～x016から算出した説明変数群		x176	時価総額	
x135	始値		x177	実績利回り	(4)
x136	高値		x178	予想利回り	
x137	安値		x179	実績PER	
x138	終値		x180	予想PER	
x139	売買高		x181	実績PSR	
x140	売気配: 価格		x182	予想PSR	
x141	買気配: 価格		x183	PBR	
x142	当日採用価格		x184	PCFR	
x143	始値(権利落調整済)		x185	コボック買指標	
x144	高値(権利落調整済)		x186	サイコロジカル値	
x145	安値(権利落調整済)		x187	ボリュームレシオ	
x146	終値(権利落調整済)		x188	時価総額(普通株ベース)	
x147	始値(権利落調整済: 配当落含む)	(4)	x189	時価総額(自己株除く)	
x148	高値(権利落調整済: 配当落含む)		x190	実績PER(自己株除く)	
x149	安値(権利落調整済: 配当落含む)		x191	予想PER(自己株除く)	
x150	終値(権利落調整済: 配当落含む)		x192	実績PSR(自己株除く)	
x151	売買代金		x193	予想PSR(自己株除く)	
x152	最終約定値A		x194	PBR(自己株除く)	
x153	最終約定値B		x195	PCFR(自己株除く)	
x154	累積調整係数A		x196	時価総額(普通株ベース)(自己株除く)	
x155	累積調整係数B		x197	TOPIX終値	(5)
x156	株価移動平均5日		x198	TOPIX終値_std_90	
x157	株価移動平均6日		x199	米ドル	
x158	株価移動平均25日		x200	英ポンド	(6)
			x201	ユーロ	

3-4. 機械学習による予測モデルの構築

機械学習による予測モデルの構築に当たっては、Python用の機械学習ライブラリであるscikit-learn (Pedregosa *et al.* 2011) を使用し、分析環境はGoogle社のColaboratory (Bisong 2019) を利用した。また、複数のモデルの比較には、機械学習ライブラリのPyCaret (Moez 2020) を使用した。

まず、前処理済みのデータを無作為に教師データ (70%, 7,000件) とテストデータ (30%, 3,000件) に分割し、さらに教師データを10分割したうえで、そのうち1つを検証データとして入れ替えながら (10分割交差検証法) , 18種類の機械学習の手法による予測精度を比較した。

4. 結果と考察

表3は、教師データで学習した各モデルに検証データの目的変数を予測させ、その精度を整理したものである。いずれのモデルにおいても10分割交差検証によって教師データの一部を検証データと繰り返し入れ替えながら、各回の予測精度の平均値を表示した。表3中の各列の値は、モデル名 (Model) , 平均絶対値誤差 (MAE : Mean Absolute Error) , 平均二乗誤差 (MSE : Mean Squared Error) , 二乗平均平方根誤差 (RMSE : Root Mean Squared Error) , 決定係数 (R2 : Coefficient of Determination) , 対数平方平均二乗誤差 (RMSLE : Root Mean Squared Logarithmic Error) , 平均絶対パーセント誤差 (MAPE : Mean Absolute Percentage Error) , 学習時間 (TT) であり、各種の誤差 (MAE, MSE, RMSE, RMSLE, MAPE) が小さいほど (0に近いほど) 実測値に対する予測値のズレが小さく、R2が大きいほど (1に近いほど) モデルの当てはまりが良いことを意味する。

表3 各モデルの予測精度 (10分割交差検証による各指標の平均値)

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)	
	xgboost	Extreme Gradient Boosting	0.0330	0.0024	0.0486	0.8352	0.0361	0.1295	5.2480
	lightgbm	Light Gradient Boosting Machine	0.0359	0.0026	0.0510	0.8194	0.0377	0.1427	7.4270
	rf	Random Forest Regressor	0.0346	0.0028	0.0525	0.8087	0.0388	0.1335	66.2730
	et	Extra Trees Regressor	0.0344	0.0031	0.0559	0.7831	0.0415	0.1368	26.2160
	gbr	Gradient Boosting Regressor	0.0480	0.0043	0.0656	0.7012	0.0482	0.1784	20.9650
	dt	Decision Tree Regressor	0.0389	0.0047	0.0685	0.6742	0.0503	0.1415	1.0960
	knn	K Neighbors Regressor	0.0564	0.0060	0.0774	0.5846	0.0571	0.2088	0.2560
	lr	Linear Regression	0.0620	0.0067	0.0819	0.5346	0.0603	0.2287	1.0830
	ridge	Ridge Regression	0.0632	0.0070	0.0836	0.5147	0.0614	0.2314	0.2220
	br	Bayesian Ridge	0.0637	0.0071	0.0844	0.5059	0.0619	0.2330	0.2200
	huber	Huber Regressor	0.0635	0.0075	0.0866	0.4791	0.0633	0.2298	1.2770
	ada	AdaBoost Regressor	0.0691	0.0076	0.0871	0.4741	0.0643	0.2654	7.5350
	omp	Orthogonal Matching Pursuit	0.0702	0.0085	0.0920	0.4134	0.0669	0.2511	0.1320
	par	Passive Aggressive Regressor	0.0806	0.0106	0.1027	0.2679	0.0760	0.2880	0.1550
	lasso	Lasso Regression	0.0950	0.0145	0.1202	-0.0009	0.0875	0.3436	0.1990
	en	Elastic Net	0.0950	0.0145	0.1202	-0.0009	0.0875	0.3436	0.2400
	llar	Lasso Least Angle Regression	0.0950	0.0145	0.1202	-0.0009	0.0875	0.3436	0.1330
	dummy	Dummy Regressor	0.0950	0.0145	0.1202	-0.0009	0.0875	0.3436	0.1230

表3を見ると、Extreme Gradient Boosting (xgboost) (Chen and Guestrin 2016) やLight Gradient Boosting Machine (lightgbm) (Ke *et al.* 2017) , Random Forest Regressor (rf) (Breiman 2001) , Extra Trees Regressor (et) (Geurts *et al.* 2006) の決定係数が高くなった。これらはいずれも、決定木 (dt : Decision Tree Regressor) とアンサンブル学習を組み合わせた手法であり、重回帰分析の各手法 (Linear Regression, Ridge Regression, Lasso Regression, Elastic Net) よりも決定係数が高くなっている。この結果は、地方債スプレッドの予測モデル構築に関する先行研究 (丹波・原口 2021b, 丹波・原口 2023) の結果とも整合している。決定木は、データが持つ特定の説明変数に基づいてデータを分割する手法であり、分割後の各データの不純度 (データの異質性やバラツキ) が最小となるように分割の条件が設定されるほか、データの分割も再帰的に行われる。なお、決定木は分類・回帰いずれの問題にも適用でき、本稿では最終的に分割された各データの目的変数の平均値をもとに量的な値を予測する「回帰木」を用いている。この回帰木モデルを複数組み合わせることで

(アンサンブル学習) , より精度の高いモデルを構築することができる。

rf と et は、教師データから重複を許してランダムに抽出した新たな教師データを複数セット作成し、各セットの教師データを学習したそれぞれの回帰木モデルによる予測値の平均を最終的な予測結果とするアンサンブル学習の手法 (Bagging: Bootstrap Aggregating) を用いている (図 2) 。rf と et には、回帰木においてデータの分割点となる説明変数の値を、分割後のデータの不純度に基づき調整して選ぶか (rf) , 計算コストを抑えるためにランダムに選ぶか (et) の違いがある。

一方、xgboost は、先に学習した回帰木モデルで予測が外れたデータを重視したうえで、あらためて回帰木モデルで学習・予測する操作を繰り返し、最終的にすべてのモデルによる予測値の平均をとる手法 (Boosting) を採用している (図 3) 。Boosting は回帰木の欠点の補正を繰り返すため、一般的に Bagging によるモデルよりも高い予測精度が期待できる一方、学習時間が長くなる。この課題の克服を目指したものが lightgbm であり、xgboost が回帰木の階層を均等に増やす (Level-wise) のに対し、lightgbm は回帰木のなかでも効率良くデータの不純度を下げ得る部分に特化して階層を増やす (Leaf-wise) ことで計算コストを抑えている。

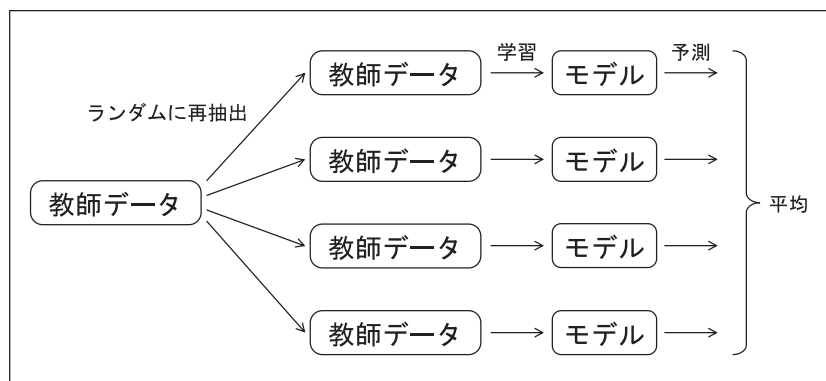


図 2 Bagging の概要

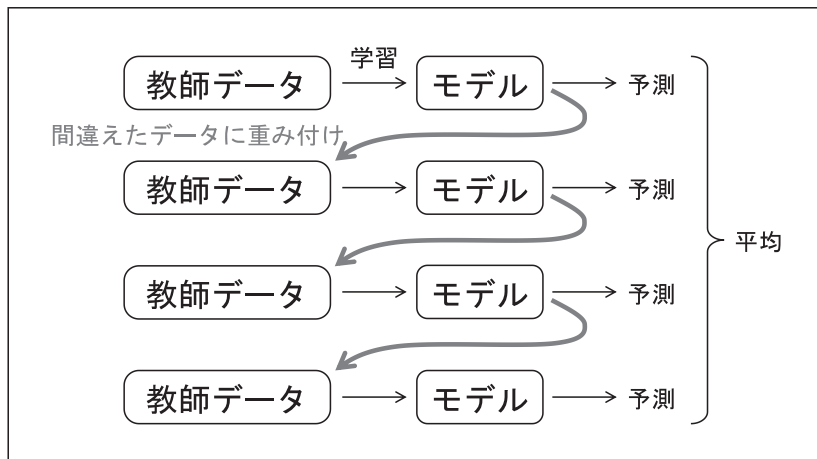


図3 Boosting の概要

決定係数が高かった xgboost, lightgbm, rf, et の4モデルについて、ランダムサーチを用いて得た20通りのハイパーパラメータの組合せの中から、10分割交差検証法により予測精度が最も高い組合せを選び、各モデルを最適化した。最適化後の4モデルの予測精度を表4に示すとともに、モデルの構築に利用しなかった残り30%のテストデータに対する各モデルの予測精度を表5に整理した。最適化後のモデルによるテストデータに対する予測結果は et が最も高く、次いで rf, xgboost, lightgbm となった。一方で、rf や et のモデルの構築に要する時間は xgboost や lightgbm よりも長くなるため、実運用に当たっては予測精度以外の要素も考慮する必要がある。

表4 最適化後の4モデルの予測精度
(10分割交差検証による各指標の平均値)

Model		MAE	MSE	RMSE	R2	RMSLE	MAPE
xgboost	Extreme Gradient Boosting	0.0354	0.0027	0.0514	0.8160	0.0381	0.1382
lightgbm	Light Gradient Boosting Machine	0.0403	0.0031	0.0560	0.7820	0.0414	0.1549
rf	Random Forest Regressor	0.0583	0.0064	0.0799	0.5569	0.0586	0.2107
et	Extra Trees Regressor	0.0541	0.0057	0.0753	0.6069	0.0551	0.1973

表5 テストデータに対する4モデルの予測精度

Model		MAE	MSE	RMSE	R2	RMSLE	MAPE
xgboost	Extreme Gradient Boosting	0.0168	0.0005	0.0234	0.9617	0.0176	0.0625
lightgbm	Light Gradient Boosting Machine	0.0286	0.0016	0.0402	0.8869	0.0303	0.1227
rf	Random Forest Regressor	0.0112	0.0003	0.0181	0.9771	0.0137	0.0494
et	Extra Trees Regressor	0.0002	0.0000	0.0028	0.9995	0.0021	0.0008

次に、予測に当たって重要な説明変数を明らかにしたい。決定木は結果が木構造で可視化されるため、その解釈は比較的容易であるが、決定木を複数組み合わせるアンサンブル学習の手法は解釈が難しくなる。ただし、どの説明変数が効いているかを示す「特微量重要度」を算出することは可能である。

特微量重要度には Feature Importance や Permutation Feature Importance (PFI) (Ke *et al.* 2017) , SHapley Additive exPlanations Value (SHAP) (Lundberg and Lee 2017) などの指標があり、それぞれ目的が異なる。Feature Importance は予測モデルを構築する際に、そのモデルがどの説明変数を重視して構築されたかを評価する指標であり、ある説明変数でデータを分割することでデータの不純度がどの程度低下するかに基づいて算出される。4モデルそれぞれについて、Feature Importance が大きかった説明変数を図4に示す。xgboost のモデルの構築で重視されたのは x028 (営業活動によるキャッシュ・フローの3年平均値) と x022 (自己資本合計の3年平均値) であったのに対し、lightgbm と et では x001 (表面利率、クーポン) が最も重視された。また、rf では x001 と x049 (営業キャッシュ・フローの対数値の3年平均値) , x028 が重要であったことがわかった。この指標により、教師データを使ってモデルを構築する際の各説明変数の貢献度を評価することができるが、各説明変数が予測精度にどの程度貢献しているかを示すものではない。

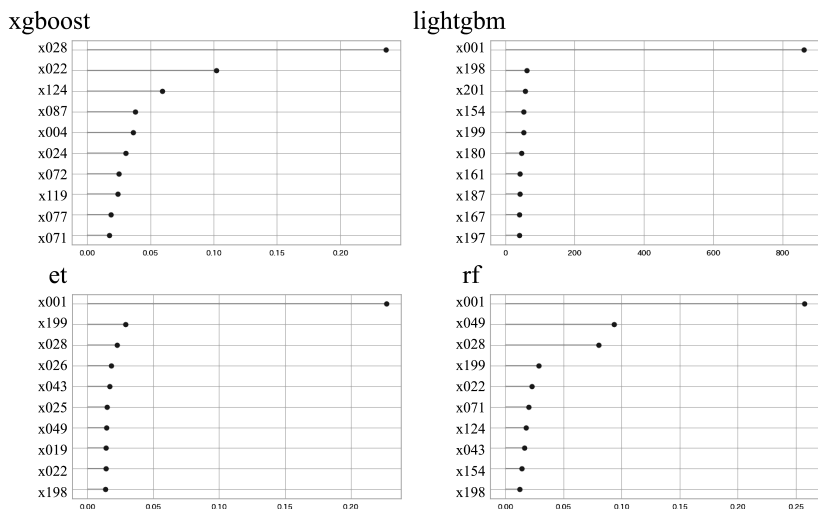


図4 4モデルにおいて Feature Importance が大きかった説明変数

PFI は、元データ内のある説明変数をランダムに入れ替えたデータを用いた場合の予測精度と、元のデータに対する予測精度を比較し、その差を説明変数の重要度とした指標である。説明変数を入れ替えることで予測精度が大きく低下する場合、その説明変数は予測精度に大きく貢献しているものとする。PFI は、データを入れ替える操作だけで計算可能であり、アプローチが直感的で理解しやすく、モデルの予測精度の向上に寄与する説明変数を特定する手法としては有効である。

一方、各説明変数が個別のデータの予測にどの程度貢献しているかを示し、モデルの予測に対する全体的な理解を深める指標として、SHAPの有用性が議論されている（吉田ほか 2020, 丹波・原口 2021a）。当該指標は社債のスプレッド予測においても活用可能性があるため、今後の検討課題としたい。

5. まとめと今後の展望

本稿では、2022年を対象期間として、社債市場（売買市場）で取引され

た社債のスプレッド予測モデルを構築するため、複数の機械学習の手法を比較した。その結果、決定木とアンサンブル学習を組み合わせた `xgboost`, `lightgbm`, `rf`, `et` の各モデルの予測精度は、テストデータに対して88~99%に達し、社債スプレッドの予測に有効である可能性を示唆した。

一方、今回のモデルの構築に用いたデータは、1年間のうち、さらに無作為に抽出した10,000件に限られるため、予測精度の改善の余地は大きい。重要度が高い説明変数に絞って計算コストを抑えつつ、解析に用いるデータを増やすことで、より精度の高いモデルの構築を目指すことが可能であり、今後の課題である。

謝辞

本研究はJPS科研費JP23K01463, JP23K01698, JP24K16490の助成を受けている。

参考文献

- Altman EI (1968) Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance* 23(4), 589-609, <https://doi.org/10.1111/j.1540-6261.1968.tb00843.x>
- Bisong E (2019) Google Colaboratory. *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners*, 59-64, https://doi.org/10.1007/978-1-4842-4470-8_7
- Black F, Scholes M (1973) The pricing of options and corporate liabilities. *Journal of Political Economy* 81(3), 637-654, <https://www.jstor.org/stable/1831029>
- Breiman L (2001) Random Forests. *Machine Learning* 45, 5-32, <https://doi.org/10.1023/A:1010933404324>
- Chen T, Guestrin C (2016) XGBoost: A scalable tree boosting system. *arXiv:1603.02754v3*, <https://arxiv.org/abs/1603.02754>
- Collin-Dufresne P, Goldstein RS, Martin JS (2001) The determinants of credit

- spread changes. *Journal of Finance* 56(6), 2177-2207, <https://www.jstor.org/stable/2697820>
- Duffie D, Singleton KJ (1999) Modeling term structures of defaultable bonds. *The Review of Financial Studies* 12(4), 687-720, <https://www.jstor.org/stable/2645962>
- Geurts P, Ernst D, Wehenkel L (2006) Extremely randomized trees. *Mach Learn* 63, 3-42, <https://doi.org/10.1007/s10994-006-6226-1>
- Jarrow RA, Turnbull SM (1995) Pricing derivatives on financial securities subject to credit risk. *Journal of Finance* 50(1), 53-85, <https://doi.org/10.1111/j.1540-6261.1995.tb05167.x>
- Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu T (2017) LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems* 30, 3146-3154, <https://dl.acm.org/doi/10.5555/3294996.3295074>
- Longstaff FA, Schwartz ES (1995) A simple approach to valuing risky fixed and floating rate debt. *Journal of Finance* 50(3), 789-819, <https://doi.org/10.1111/j.1540-6261.1995.tb04037.x>
- Lundberg SM, Lee S (2017) A unified approach to interpreting model predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4768-4777, <https://dl.acm.org/doi/10.5555/3295222.3295230>
- Merton RC (1974) On the pricing of corporate debt: The risk structure of interest rates. *Journal of Finance* 29(2), 449-470, <https://doi.org/10.2307/2978814>
- Moez A (2020) PyCaret: An open source, low-code machine learning library in Python. <https://pycaret.org>
- Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn:

Machine learning in Python. *Journal of Machine Learning Research* 12(85), 2825-2830

Sengupta P (1998) Corporate disclosure quality and the cost of debt. *The Accounting Review* 73(4), 459-474,

白須洋子, 米澤康博 (2008) 社債流通市場における社債スプレッド変動要因の実証分析. *現代ファイナンス* 24, 101-127, https://doi.org/10.24487/gendaifinance.24.0_101

丹波靖博, 原口健太郎 (2021a) 機械学習を用いた我が国における地方債の信用リスク分析. *JAFEE2021夏季大会予稿集*, 65-76

丹波靖博, 原口健太郎 (2021b) わが国における地方債スプレッド推定モデル構築に対する機械学習の適用可能性. *西南学院大学経済学論集* 56(1-2), 75-91

丹波靖博, 原口健太郎 (2023) 機械学習を用いたわが国における地方債のスプレッド推定モデルの構築. *JAFEEジャーナル* 21, 29-42, https://doi.org/10.32212/jafee.21.0_29

吉田秀穂, 田嶋優樹, 今井優作 (2020) 決定木モデルの解釈におけるSHAP 値の有用性の検証. *人工知能学会第34回全国大会論文集*, 1-3