基于函数型数据分析的债券违约预测评估

王凯
合肥工业大学 数学学院，安徽 合肥 230009
DOI:10.61369/ASDS.2026050013

摘　　要： 对债券进行风险评估时，由于传统信用模型所用的是发行主体的静态财务报表及外部评级结果，故传统方法存在指标更新频率低、分析维度单一的问题，影响了预测评估的时效性及精确度。本文从函数型数据分析（FDA）的角度，结合债券利率等特征信息，用多种机器学习算法对债券违约进行了系统的预测评估。所得结果表明，具有函数型特征的债券风险评估模型在各评价指标上均有明显改进。

关 键 词： 债券违约；FDA；机器学习；利率特征；信用风险

Bond Default Prediction Based on Functional Data Analysis

Wang Kai
School of Mathematics, Hefei University of Technology, Hefei, Anhui 230009

Abstract: When assessing the risk of bonds, traditional credit models mainly rely on the issuer's static financial statements and external ratings. These models have issues such as low frequency of indicator updates and a single assessment dimension, which affect the timeliness and accuracy of predictions. This paper introduces functional data analysis into the field of bond default prediction, combining bond interest rate features with other feature information, and uses various machine learning algorithms for default prediction. The results show that models incorporating functional data analysis achieve significant improvements in multiple metrics.

Keywords: bond default; Functional Data Analysis (FDA); machine learning; interest rate; credit risk

引言

由于我国债券市场近年来迅猛发展，违约事件频发，因而对金融系统的安全及稳定运行造成重大影响。截至2023年，我国债券市场存量规模已突破150万亿元，成为全球第二大债券市场，累计违约金额已达8000亿元[1]。因此，保护投资者合法权益、建立科学有效的违约预测模型，已经成为金融监管、投资决策、风险控制等领域十分明确、重要的研究方向。

国外个人信贷违约预测研究起步较早，已形成较为成熟的方法体系，主要包括两大类：传统统计建模和机器学习模型。传统统计模型主要有 Logistic 回归、线性判别分析和二次判别分析等；机器学习模型主要有随机森林、SVM、XGBoost 和遗传算法等[2]。

由于传统方法对市场变化的响应能力较弱，大量文献已经论证了机器学习算法加入经济金融模型之后能提高预测精度。例如，Zhou 研究了机器学习算法在非线性关系建模上的明显优势，并给出其在信用评级中的应用[3]。王玉龙等对七种机器学习算法做了直接对比，得出随机森林预测企业违约风险的效果最优的结论[4]。Chopra & Bhilare 对比了决策树与梯度提升模型，研究银行业贷款违约，发现梯度提升模型预测效果显著优于决策树[5]。Sebastian 等对此作了极好的补充，结合客户日终余额数据评估分析银行信用风险，认为 XGBoost 仍然是信用风险预测的领先模型之一[6]。胡蝶用随机森林算法估计债券违约风险[7]。邓晴元以传统的 Logit 方法与 XGBoost 算法进行对比，做了极有洞见的债券违约风险识别实证研究[8]。Tran 将遗传算法与深度学习结合，构建了高准确率的混合信用模型[9]。Gu et al. 构建了宏观经济变量与微观企业特征的混合数据集，系统地考察了主流机器学习算法在债券违约预测中的适用性[10]。

然而，上述研究大多忽略了债券价格或收益率随时间变动的信息，不能及时响应市场的动态变化、监测信用风险等。函数型数据分析（FDA）将观测数据视为连续函数，为处理高维、动态时间序列提供了灵活框架。Ramsay 于1982年首次引入函数型数据的概念[11]，并于1991年与 Dalzell 正式提出函数型数据分析（FDA）的一些方法[12]。函数型数据分析（FDA）可在函数框架下挖掘静态数据蕴含的动态规律[13]，已广泛应用于医学统计、行为轨迹识别等领域。例如，魏艳华等采用 FDA 研究中国人口地域差异[14]；王华强、刘黎明则基于 FDA 探讨不同地区利率与物价的动态关系[15]。Tao 等研究了 FDA 算法对原油价格预测的作用，在解析较为复杂的时间动态特征、不规则模式以及数据突变等方面具有明显的优越性[16]。

作者简介：王凯，合肥工业大学数学学院，硕士研究生，研究方向为金融数据的统计分析与风险评估。

在金融市场，收益率曲线、价格波动与信用价差均呈连续变化特征。Kokoszka 系统讨论了 FDA 在利率期限结构建模中的应用 [17]。现阶段常见的 FDA 方法包括：函数型回归、函数型主成分分析（FPCA）、函数型聚类和正则分析等。其中，FPCA 可以将复杂的函数型数据分解为若干个基础特征分量，有助于分析研究对象的主要变化模式，在处理时间序列方面具有显著的优势 [18]。Nian 等利用 FPCA 提取正交主成分函数，有效地模拟传感器信号之间的相关性，预测工业设备剩余使用寿命 [19]。

基于上面分析，本文的创新点如下：（1）在信用债券风险识别领域引入函数型数据分析（FDA）方法，丰富了 FDA 的研究场景；（2）将特征提取进行精细化处理，通过平滑算法，将离散的利率曲线转化为可导的平滑函数，提取其导数特征，结合函数型主成分分析（FPCA），进一步捕捉市场中随时间动态变化的风险，打破了基于传统静态特征信息研究的局限性；（3）通过 FDA 提取债券动态特征，将特征值输入多种经典机器学习模型，有效地提升了模型在债券违约预测评估的准确性。

一、基本机器学习算法

（一）逻辑回归

逻辑回归是一种线性分类模型，用于预测类别概率（如二分类中预测样本属于某一类的概率），其核心是将线性回归的结果通过 Sigmoid 函数映射到 (0,1) 区间。

设输入特征为：

$$x=(x_1,x_2,\cdots,x_n)^{\mathrm{T}} \tag{1}$$

特征权重为：

$$w=(w_1,w_2,\cdots,w_n)^{\mathrm{T}} \tag{2}$$

线性部分为：

$$z=w^{\mathrm{T}}x+b,\quad b\in\mathbb{R} \tag{3}$$

通过 Sigmoid 函数映射，预测为正类，即 y=1 的概率为：

$$\hat{P}(y=1|x)=\frac{1}{1+e^{-(w^{\mathrm{T}}x+b)}} \tag{4}$$

（二）随机森林

随机森林是基于决策树的 Bagging 集成的拓展，其基本原理是通过对原始样本进行随机抽样再放回，循环多次后，训练出多个相互独立的决策模型，并在预测阶段对各模型的结果进行统计处理，得到最终的预测结果。

传统的 Bagging 集成学习，在每个节点划分时，会考虑所有可用的特征，再选择最优的特征进行划分；而随机森林在节点划分时，会先随机选择一个特征子集，再从这个子集里选择最优的特征进行划分。

随机森林在特征选择过程中增加了随机性，增大不同决策树之间的差异度，避免了单个决策模型易出现的过拟合问题，让集成后的模型泛化能力更强 [20]。该模型处理高维数据和非线性关系时效果较好，输出结果包含特征重要性指标，方便后续的变量选取和结果分析。

（三）XGBoost

XGBoost 是在梯度提升基础上进一步改进的集成学习算法。通过构建多个初始分类模型进行迭代计算，每一轮模型的信息作为下一轮迭代计算的输入，不断优化，从而提升模型的预测性能。

相比于传统的梯度提升方法，XGBoost 为了防止出现过拟合现象，加入了正则化项，并且因引入了二阶导数进行目标函数计算，使得预测更加精确。

（四）LightGBM

LightGBM 同样是一种基于梯度提升方法的集成学习算法。通过加法模型和前向分步算法不断优化损失函数，不断训练新的决策模型来拟合上一步的信息。

与传统梯度提升树不同的是，LightGBM 在分裂策略上采用基于直方图的方法，通过叶子优先生长的树结构提高训练效率和模型精度。

二、函数型数据分析方法

函数型数据分析（FDA）是以函数曲线或连续时间过程为研究对象的一种统计分析方法 [21]。

传统的多维向量数据处理方式忽略了变量间的时间和空间平滑性及内部结构，而函数型数据处理方式利用基函数构造平滑函数拟合离散的样本点，挖掘出变量的连续变化特性，有助于更好地建立模型、预测未来的变化趋势，提供更多的有效信息 [22]。

（一）利率信息的函数化表示：B 样条平滑

一个基函数系统是指一组数学上相互独立的已知函数，常用的基函数包括傅里叶基、Legendre 多项式与 B 样条，其中 B 样条具有局部支撑、灵活性强、计算效率高等优势，特别适用于存在局部波动的利率曲线。本文选用 B 样条基的线性组合给出利率曲线的估计：

设 $Y_i(t_j)$ 表示第 i 支债券在时间点 t_j 上的信用利差，即风险债券与同期限国债之间的收益率差额，其中 t_j 取自有限的离散时间点集合，$t_j\in T=[t_1,t_m]$，j=1,2,⋯,m。将每个离散观测序列 $Y_i(t_1)$, $Y_i(t_2)$, ⋯, $Y_i(t_m)$ 转化为连续函数形式 $X_i(t)$，记：

$$Y_i(t_j)=X_i(t_j)+\varepsilon_{ij} \tag{5}$$

其中，$X_i(t)$ 为待估曲线；ε_{ij} 为实际值与拟合值间的随机误差。

采用 B 样条基对观测序列进行曲线拟合：

$$\hat{X}_i(t)=\sum_{k=1}^{K}\beta_{ik}B_k(t) \tag{6}$$

其中：k 为基函数数量，k=1,2,⋯,K；$B_k(t)$ 为第 k 个 B 样条

基函数；β_{ik}为回归系数。

在确定基展开式中的系数β_{ik}时，既要逼近原始数据又保持足够光滑，通常采用带惩罚项的最小二乘估计，即使得下式达到最小：

$$\sum_{j=1}^{n}[Y_i(t_j)-\hat{X}_i(t_j)]^2+\lambda\int_T[X_i''(t_j)]^2dt \quad (7)$$

其中，第一项衡量拟合误差；第二项为二阶导数惩罚项，用于控制曲线的弯曲程度；λ为平滑参数，控制拟合与光滑的权衡。

其矩阵表达式为：

$$\begin{cases}(Y-B\beta)^T(Y-B\beta)+\lambda\beta^T\Omega\beta\\ \Omega=\int_T B''(t)(B''(t))^T dt\end{cases} \quad (8)$$

其中，B(t) 是各 $B_k(t)$ 构成的列向量；$\hat{\beta}=(\hat{\beta}_1,\ \ldots,\ \hat{\beta}_k)^T$为待估系数。

可以得出，$\hat{\beta}=(B^TB+\lambda\Omega)^{-1}B^TY$时该式达最小，此时，$\hat{X}_i(t)=B(B^TB+\lambda\Omega)^{-1}B^TY$。

（二）函数型主成分分析

在得到平滑的函数表示后，本文引入了 FPCA 进一步降低维度并提取主要变化模式。将传统主成分分析从向量空间推广到函数空间，反映利率曲线随时间的主要变化特征。

设中心化后的函数样本为：

$$\begin{cases}x_i(t)=\hat{X}_i(t)-\bar{X}(t)\\ \bar{X}(t)=\frac{1}{N}\sum_{i=1}^{N}\hat{X}_i(t)\end{cases} \quad (9)$$

定义第 i 个样本的主成分得分为：

$$\begin{cases}z_i=\int_T\xi(t)x_i(t)dt,\ i=1\,2\,\cdots\quad N\\ \xi(t)=\sum_{k=1}^{K}c_kB_k(t)\end{cases} \quad (10)$$

其中，$x_i(t)$ 为经过中心化处理的第 i 个样本；$\xi(t)$为主成分权函数，c_k 为待计算系数。

第一函数主成分需满足：

$$\begin{cases}\max N^{-1}\sum_{i=1}^{N}z_{i1}^2\\ \int_T\xi_1(t)^2dt=1\end{cases} \quad (11)$$

其中，$z_{i1}=\int_T\xi_1(t)x_i(t)dt$为第 i 个样本的第一主成分得分。

类似地，可依次求出第 j 个函数主成分，各主成分对应的权函数应相互正交，则其权函数需满足：

$$\begin{cases}\max N^{-1}\sum_{i=1}^{N}z_{ij}^2\\ \int_T\xi_j(t)^2dt=1\\ \int_T\xi_j(t)\xi_1(t)dt=\cdots=\int_T\xi_j(t)\xi_{j-1}(t)dt=0\end{cases} \quad (12)$$

在多元统计分析框架下，主成分的求解最终转变成对协方差阵或相关系数阵的特征值和特征向量的求解过程。

在函数型数据分析中，设 $x_i(s)$ 和 $x_i(t)$ 的协方差函数为：

$$v(s,t)=N^{-1}\sum_{i=1}^{N}x_i(s)\cdot x_i(t) \quad (13)$$

该函数刻画了时间点 s 与 t 上利率曲线的协同变化结构（s, t ∈ T，T=[t_1,t_n]）。通过求解以下特征方程获得主成分权函数：

$$\int_T v(s,t)\xi(t)dt=\rho\xi(s) \quad (14)$$

其中，ρ为满足约束条件下的最大特征值，$\xi(t)$为对应的特征函数。

将式（10）、式 (13) 矩阵化，可表示为：

$$\max_{\xi'\xi=1}N^{-1}\xi'X'X\xi \quad (15)$$

$$\mathbf{V}\xi=\rho\xi \quad (16)$$

其中，ξ为权重列向量；X 为数据构成的 N × p 阶矩阵；矩阵$\mathbf{V}=N^{-1}\mathbf{X}'\mathbf{X}$为样本协方差矩阵。

对函数型主成分的求解也变成求解特征值和特征函数的问题，通过对函数进行 SVD 离散化或基函数展开等方法进行求解。

本文在利率曲线上应用 FPCA 提取前3个主成分后，每个债券都用主成分得分作为输入变量参与建模。它们是累计方差贡献率≥85% 的前3个主成分，可以实现对高维函数数据进行降维与信息获取。

三、模型构建与评价指标

（一）模型构建

为构建高效准确的债券违约预测模型，本文设计了一套集成机器学习与函数型数据的建模流程，整体流程包括数据处理、特征构建、模型训练、模型评估等步骤，如图1所示。

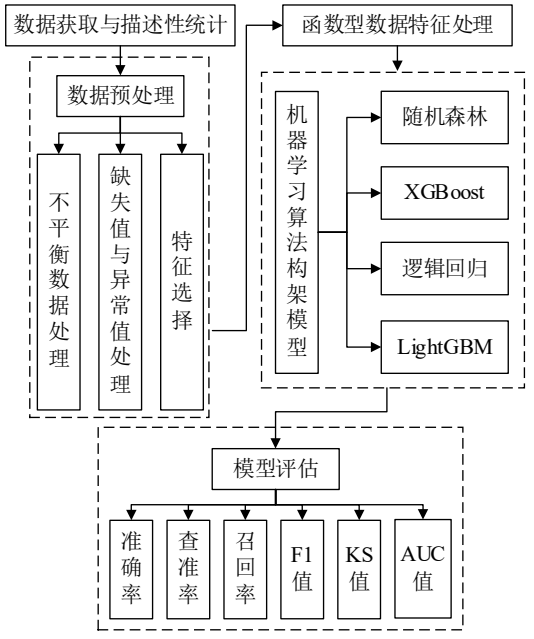


图1：违约预测流程图

Figure 1: Default prediction flowchart

对数据预处理来说，由于存在正负样本不平衡的问题，通过采样技术对训练数据做预处理。对于数据中的缺失值或异常值使用线性插值法进行处理，并根据特征的重要性和变量之间的相关性来选取特征子集。

本文对特征构造做了如下设计：先设立对照组，只用传统金融指标，再设实验组，在对照组基础上引入利率函数特征，继而用 B 样条插值法对每支债券最近30个时间点的利率序列作平滑处理拟合出连续可导的函数，由此自然地引出 FPCA 方法，提取前3个主成分得分，客观考察函数型数据分析在违约预测中的作用。

本文选取4种机器学习模型进行对比分析，分别为逻辑回归、随机森林、XGBoost 和 LightGBM。在选择最优参数组合上，采用5折交叉验证方式。选取出需调优的各模型参数，将数据分成5份，循环进行5次训练验证，选择在验证集中表现最优的参数组合。

模型训练结束后，分别对各模型的预测效果进行评估，评估指标包括：准确率（Accuracy）、精确率（Precision）、召回率（Recall）、F1分数、AUC 值、KS 值。

（二）评价指标

首先定义以下指标：TP（真正例）、FP（假正例）、TN（真反例）、FN（假反例），再根据模型正确分类和错误分类的情况生成混淆矩阵，如表1所示。

表1：混淆矩阵

Table 1: Confusion matrix

		预测类别		
		预测未违约	预测违约	总和
真实类别	真实未违约	TP	FN	P
	真实违约	FP	TN	N
	总和	P*	N*	

为全面评估模型的预测性能，本文选取了六个常用评价指标：

1. 准确率（Accuracy）：表示模型正确预测的样本数占总样本数的比例，评估整体预测准确性。计算公式为：

$$\text{Accuracy}=\frac{\text{TP+TN}}{\text{FP+FN+TP+TN}} \tag{17}$$

2. 精确率（Precision）：衡量模型预测为正类的样本中，真实为正类的比例，反映预测结果的可信度。计算公式为：

$$\text{Precision}=\frac{\text{TP}}{\text{P}^*} \tag{18}$$

3. 召回率（Recall）：表示所有真实正样本中，被正确识别出的正样本所占的比例。计算公式为：

$$\text{Recall}=\frac{\text{TP}}{\text{P}} \tag{19}$$

4.F1值（F1-score）：作为精确率与召回率的调和平均数，用于衡量模型的综合分类表现。计算公式为：

$$\text{F1}=\frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision+Recall}}=\frac{2\text{TP}}{\text{P+P}^*} \tag{20}$$

5.AUC 值（Area Under Curve）：为 ROC 曲线（Receiver Operating Characteristic Curve）下的面积，表示随机抽取一个正样本和一个负样本时，模型将正样本排在负样本前面的概率。构建 ROC 曲线时，纵轴为真阳性率（TPR），横轴为假阳性率（FPR）：

TPR 指模型正确识别的未违约样本占总未违约样本的比例：

$$\text{TPR}=\frac{\text{TP}}{\text{TP+TN}} \tag{21}$$

FPR 指模型错误判别的违约样本占总违约样本的比例：

$$\text{FPR}=\frac{\text{FP}}{\text{FP+TN}} \tag{22}$$

取不同的概率得分作为分类阈值 θ，对应的 TPR 值和 FPR 值绘制成 ROC 曲线。

AUC 值范围在0到1之间，越接近1，表明对正负样本的区分能力越强，模型性能越好。

6.KS 值（Kolmogorov - Smirnov）：代表了在最优阈值 θ 下，此时，正负样本的累积分布函数之间距离达到最大，公式表示为：

$$\begin{cases}\text{KS}=\max_{\theta}\left|CDF_P(\theta)-CDF_N(\theta)\right| \\ CDF_P(\theta)=\dfrac{\text{FN}}{\text{P}} \\ CDF_N(\theta)=\dfrac{\text{TN}}{\text{N}}\end{cases} \tag{23}$$

KS 值越大，说明模型的准确性越高，区分度越好。

四、实证研究

（一）数据来源

本文数据来源包括 Wind、中债估值中心、沪深交易所等公开渠道，变量涵盖发行人财务指标（如资产负债率、现金短债比）、债券特征（如票面利率、期限结构）及宏观经济指标（如国债利率、CPI 同比），样本选取时间段为2020年1月至2024年8月，为我国银行与交易所市场发行的公司信用类债券。从上述数据集中获得正常到期债券和违约债券的基本数据、债券到期 / 违约日前30天的债券信用利差数据、发行方的财务指标和宏观经济数据，其中，一支债券对应一个样本，定义正常债券为正样本，总数为1030个，违约债券为负样本，总数为96个，共计1126个。对于债券信用利差数据，本文使用上海清算所的债券到期收益率（YTM）作为债券的收益率，国债到期收益率作为无风险收益率，对于不同期限的债券，使用对应期限的无风险收益率计算用于分析的债券信用利差数据，所用自变量见表2。

表2：自变量字段和字段描述

Table 2: Independent variable fields and field descriptions

变量类型	典型指标
债券基本信息	代码、年度、年限、公司代码、票面利率、发行规模
企业财务指标	资产负债率、现金短债比、经营现金流、净利润率(%)、EBIT 保障倍数(倍)、EBITDA 保障倍数(倍)、净资产收益率(%)、净资产
市场宏观指标	国债利率、汇率波动、USD/CNY、CPI: 同比(%)
条款特征	可赎回性、可回售性、外部信用增级方式、是否违约、最新主体评级、是否城投、企业性质（国有企业 / 民营企业）、是否中期票据、短期融资券、公司债
利率曲线特征	信用利差序列30期

（二）描述性统计与数据可视化分析

债券违约状态分布如图2所示。

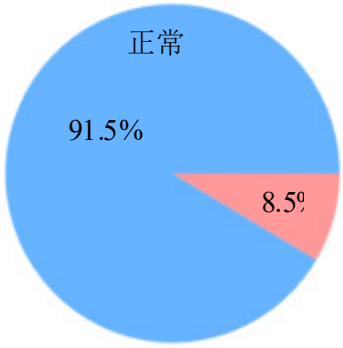


图2：违约样本分布

Figure 2: Default sample distribution chart

从图2可以看出正样本的数据量是负样本数据量的近10倍，明显类别比例悬殊，说明这是一个不平衡数据集，后续对此训练集采用 SMOTE 过采样技术进行不平衡数据集处理。

不同企业性质主体的违约情况如图3所示。

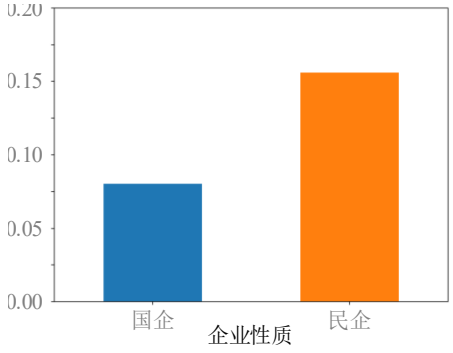


图3：不同企业性质违约率对比

Figure 3: Comparison Chart of Default Rates by Different Types of Enterprises

分析图3发现国企和民企的违约率分别为0.075和0.15，国企相较民企的违约率要低，大约为民企违约率的一半。

正常债券与违约债券在五个关键财务指标上的特征雷达图如图4所示。

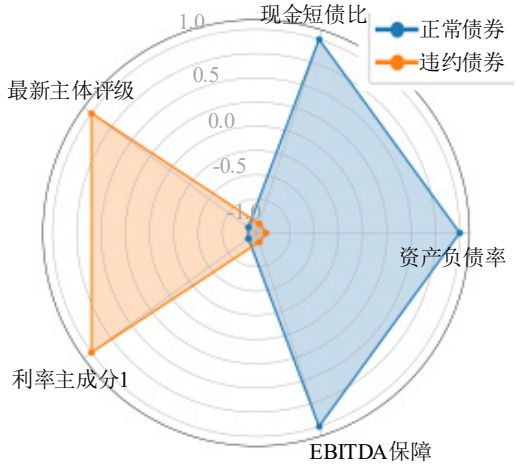


图4：债券特征雷达图

Figure 4: Bond characteristics radar chart

从图4可见，正常债券在现金短债比与保障倍数两个指标上表现更优，反映出其短期偿债能力更强，现金流更充足。违约债券的利率主成分得分与最新主体评级指标高于正常债券，体现出其存在较高的市场与信用风险。

变量相关性矩阵如图5所示。

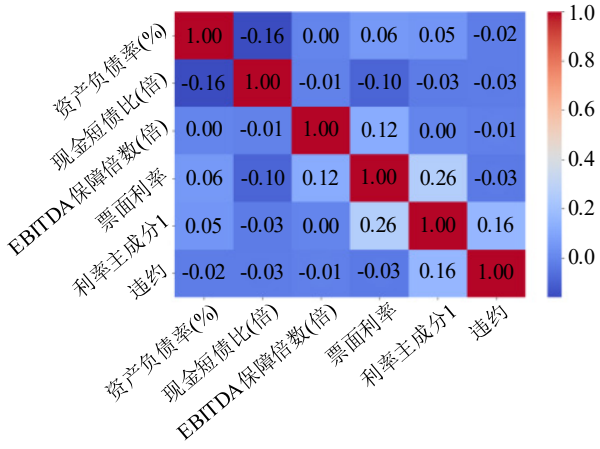


图5：关键变量相关性矩阵

Figure 5: Key Variable Correlation Matrix

图5结果显示，各指标之间的相关性较低，特征集具备独立性，无多重共线性影响，这些特征可直接用于模型训练。

（三）利率曲线数据的差异性分析

1. 平方欧几里得距离置换检验

为检验违约债券与正常债券在利率曲线整体形态上的差异，需要一个度量函数间距离的指标，通常使用欧几里得距离[23]，为提高计算效率，本文采用基于函数型数据的平方欧几里得距离置换检验方法。

首先，假设一共 N 个样本中，前 N_1 个样本均为正样本，后 $N-N_1$ 个样本均为负样本，计算两类债券样本的平均函数，并根据其平方积分差，定义检验统计量 D^2 如下：

$$\begin{cases} D^2 = \int_T [\bar{X}_1(t) - \bar{X}_2(t)]^2 dt \\ \bar{X}_1(t) = \frac{1}{N_1}\sum_{i=1}^{N_1} \hat{X}_i(t) \\ \bar{X}_2(t) = \frac{1}{N-N_1}\sum_{i=N_1+1}^{N} \hat{X}_i(t) \end{cases} \quad (24)$$

其中，$T=[t_1,t_{30}]$ 为利率曲线时间区间。

在零假设 $H_0 : \bar{X}_1(t) = \bar{X}_2(t)$ 下，通过1000次随机置换分组标签，生成检验统计量 D^2 的经验分布，即零假设成立时统计量的取值分布。

p 值定义为置换分布中大于等于观测统计量的比例。当 $p<0.05$ 时，拒绝零假设，认为两类债券的利率曲线在整体形态上存在显著差异，图6为 D^2 统计量的经验置换分布，直观的展示了观测统计量在置换分布中的相对位置。

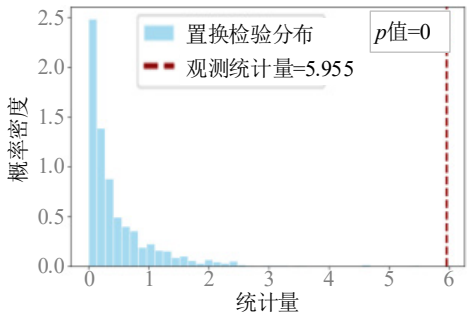


图6：置换检验分布

Figure 6: Permutation test distribution chart

可以看到，置换分布主要集中于较小的 D^2 区间，呈明显右偏特征，而观测统计量 D^2=5.955位于该分布的极端右尾，在1000次置换中几乎未被观测到。

表3：D² 置换检验结果

Table 3: D² replacement test results

检验方法	统计量	p 值	显著性
置换检验	5.955	$p < 0.05$	显著

注：显著性水平以0.05为阈值。

表3结果显示，两类债券样本的均值利率曲线在整体形态上存在显著差异（D²=5.9546，p<0.05）。说明违约债券的利率曲线在动态演变过程中与正常债券显著不同，曲线形态可作为反映违约风险的重要动态特征。

2. 样本动态相图

为进一步探索违约债券与正常债券在利率动态上的差异，本文引入相平面分析方法[24]，基于债券利率的一阶导数与二阶导数联合绘图，通过相空间轨迹观察违约与正常企业的动态特征，如图7所示。

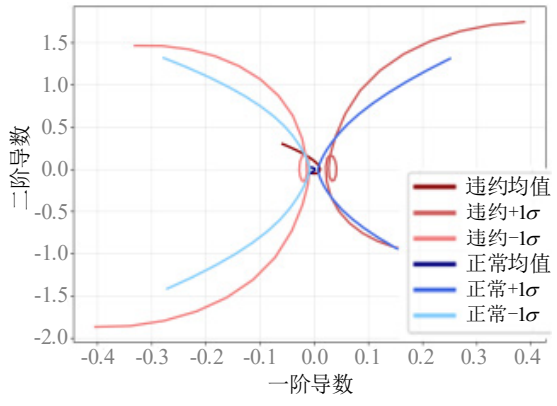


图7：样本动态相图

Figure 7: Sample dynamic phase diagram

其中，一阶导数反映了利率变化的速率，体现了市场利率的变化趋势；二阶导数反映了该变化的加速度，表现为利率曲线的凹凸性，通过观察二阶导数，可以识别出利率的异常变化情况。

结果显示，两类样本的轨迹均呈现出近似对称的蝶形分布，表明债券收益率的速度与加速度之间存在一定的规律性。

从整体形态来看，违约样本呈现出宽轨迹特征，均值轨迹及其 $\pm 1 \cdot \sigma$ 区间的这种特征更为明显，这说明违约样本的动态特征波动更强烈。相比之下，正常样本的轨迹更“窄”，即风险动态变化更稳定。从图中可以分析得出，违约样本与正常样本在动态相图中的轨迹特征有较好的区分度，为违约风险识别提供了依据。

3. 函数型主成分分析（FPCA）结果

为简化对连续型函数的分析难度，本文通过 FPCA 提取利率函数的主成分信息，并取前3个主成分得分作为代表性函数特征，提升模型训练效率。每只债券的 FPCA 结果中，第一主成分主要反映利率曲线的整体变化趋势，第二、第三主成分反映了利率曲线的变化率及局部变化特征。

前两个主成分得分（主成分1与主成分2）在二维空间中的分布情况，如图8所示。

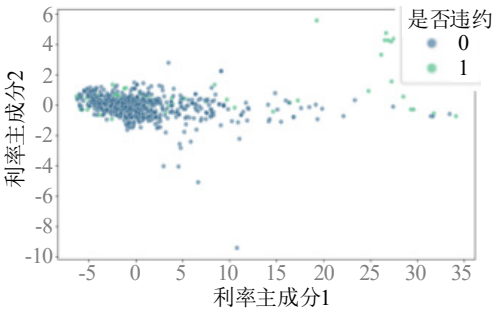


图8：利率曲线主成分空间分布

Figure 8: Spatial distribution map of principal components of the interest rate curve

图中，is_default=1表示违约，用绿色表示；is_default=0表示正常，用蓝色表示。结果显示，违约债券利率曲线整体水平偏高，在该位置具有一定的违约识别能力。而在主成分1数值为 −5 至5的区域，违约债券与正常债券混合在一起，难以进行区分。图示结果为本文模型引入利率主成分作为函数型特征提供了依据。

（四）模型评估结果对比

为了比较不同模型在引入函数型数据分析（FDA）前后的预测性能差异，本文选取了逻辑回归、随机森林、XGBoost 与 LightGBM 方法，在违约预测任务中进行算例分析。

1. 逻辑回归

混淆矩阵如表4所示。

表4：逻辑回归算例混淆矩阵

Table 4: Confusion matrix of the logistic regression example

		预测类别	
		预测未违约	预测违约
无 FDA	真实未违约	780	250
	真实违约	24	72
含 FDA	真实未违约	840	190
	真实违约	24	72

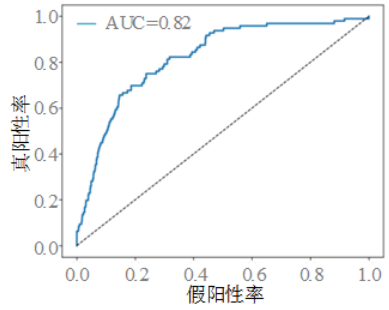
各评价指标结果如表5所示。

表5：逻辑回归预测评估

Table 5: Prediction evaluation of logistic regression

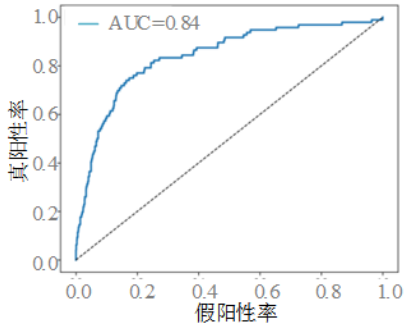
	准确率	查准率	召回率	F1值	AUC 值	KS 值
无 FDA	0.757	0.224	0.750	0.345	0.820	0.512
含 FDA	0.810	0.275	0.750	0.402	0.840	0.576

对应的 ROC 曲线如图9所示：



（a）未考虑 FDA 的逻辑回归 ROC 曲线

(a) Logistic regression ROC curve without considering FDA



（b）考虑 FDA 的逻辑回归 ROC 曲线

(b) Consider the FDA logistic regression ROC curve

图9：逻辑回归 ROC 曲线

Figure 9: ROC curve of logistic regression

根据模型评估结果，逻辑回归在不引入 FDA 时，准确率仅为75.67%，查准率为0.2236，说明误报率较高。召回率为0.7500，F1值为0.3445，AUC 值为0.8202，KS 值为0.5121，说明该模型对违约行为的风险识别能力较弱。

引入 FDA 后，逻辑回归模型准确率提升至80.99%，违约识别能力有所增强。Recall 为0.7500，F1值为0.4022，AUC 值为0.8404，KS 值为0.5762，说明 FDA 确实提高了模型对违约样本的识别与区分能力。

2. 随机森林

混合矩阵如表6所示。

表6：随机森林算例混淆矩阵

Table 6: Confusion matrix of the random forest example

		预测类别	
		预测未违约	预测违约
无 FDA	真实未违约	996	34
	真实违约	53	43
含 FDA	真实未违约	1008	22
	真实违约	30	66

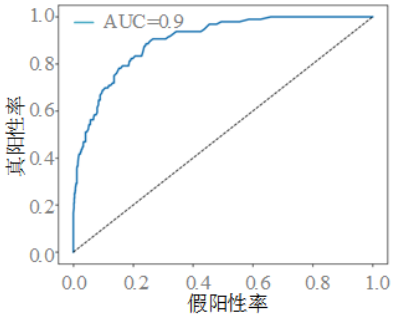
各评价指标结果如表7所示。

表7：随机森林预测评估

Table 7: Prediction evaluation of random forest

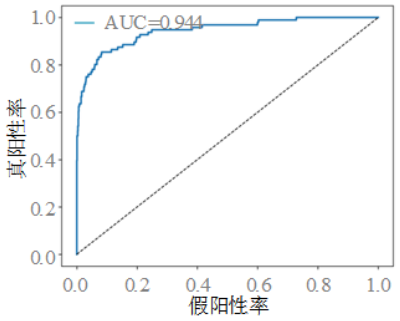
	准确率	查准率	召回率	F1值	AUC 值	KS 值
无 FDA	0.923	0.558	0.448	0.497	0.900	0.643
含 FDA	0.954	0.750	0.688	0.717	0.944	0.772

对应的 ROC 曲线如图10所示：



（a）未考虑 FDA 的随机森林 ROC 曲线

(a)Random forest ROC curve without considering FDA



（b）考虑 FDA 的随机森林 ROC 曲线

(b) Consider the FDA random forest ROC curve

图10：随机森林 ROC 曲线

Figure 10: ROC curve of random forest

随机森林模型在不引入 FDA 时，准确率达到了92.27%，查准率为0.5584，召回率为0.4479，F1值为0.4971，AUC 为0.9002，KS 值为0.6427，说明该模型有着不错的识别能力。

在引入 FDA 后，随机森林准确率提升至95.38%，提高了对正常与违约债券的识别能力；查准率为0.7500，有效降低了误判风险；召回率、F1值、AUC 值、KS 值均有所提升，表明 FDA 的引入显著增强了模型对违约样本的识别能力。

3.XGBoost

混合矩阵如表8所示。

表8：XGBoost 算例混淆矩阵

Table 8: Confusion matrix of XGBoost

		预测类别	
		预测未违约	预测违约
无 FDA	真实未违约	973	57
	真实违约	43	53
含 FDA	真实未违约	1013	17
	真实违约	17	79

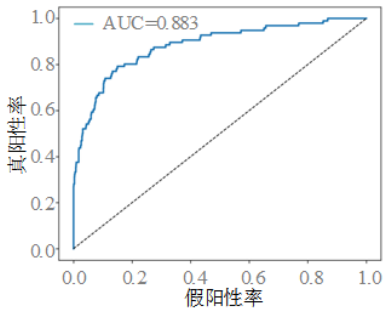
各评价指标结果如表9所示。

表9：XGBoost 预测评估

Table 9: Prediction evaluation of XGBoost

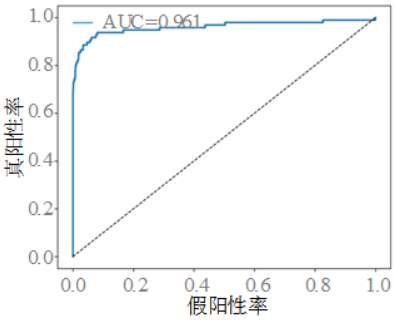
	准确率	查准率	召回率	F1值	AUC 值	KS 值
无 FDA	0.911	0.482	0.552	0.515	0.883	0.642
含 FDA	0.970	0.823	0.823	0.823	0.961	0.857

对应的 ROC 曲线如图11所示：



（a）未考虑 FDA 的 XGBoost ROC 曲线

(a) XGBoost ROC curve without considering FDA



（b）考虑 FDA 的 XGBoost ROC 曲线

(b) Consider the FDA XGBoost ROC curve

图11：XGBoost ROC 曲线

Figure 11: ROC curve of XGBoost

XGBoost 模型在不引入 FDA 时，准确率为91.12%，查准率为0.4818，召回率为0.5521，F1值为0.5146。AUC 值达到0.8825，KS 值为0.6422，说明模型对违约行为仍有良好识别力。

在考虑 FDA 的情况下，模型准确率上升到96.98%，各项评价指标显著提升，表明了结合 FDA 后的模型在准确率与敏感度上提升显著。

4.LightGBM

混合矩阵如表10所示。

表10：LightGBM 算例混淆矩阵

Table 10: Confusion matrix of LightGBM

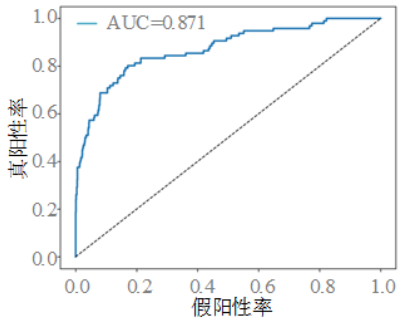
		预测类别	
		预测未违约	预测违约
无 FDA	真实未违约	989	41
	真实违约	47	49
含 FDA	真实未违约	1019	11
	真实违约	18	78

表11：LightGBM 预测评估

Table 9: Prediction evaluation of LightGBM

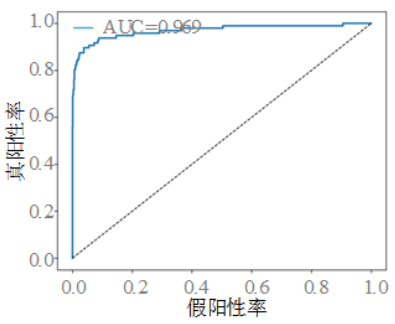
	准确率	查准率	召回率	F1值	AUC 值	KS 值
无 FDA	0.922	0.544	0.510	0.527	0.871	0.631
含 FDA	0.974	0.876	0.813	0.843	0.969	0.858

对应的 ROC 曲线如图12所示：



（a）未考虑 FDA 的 LightGBM ROC 曲线

(a) LightGBM ROC curve without considering FDA



（b）考虑 FDA 的 LightGBM ROC 曲线

(b) Consider the FDA LightGBM ROC curve

图12：LightGBM ROC 曲线

Figure 12: ROC curve of LightGBM

LightGBM 模型在不引入 FDA 时，准确率为92.18%，查准率为0.5444，召回率为0.5104，F1值为0.5269，整体表现相对稳定。AUC 为0.8714，KS 值为0.6312，说明模型在违约识别方面具有较强能力。

引入 FDA 后，LightGBM 模型的准确率提升至97.42%，各项评价指标有明显改进，进一步提高了对违约样本的识别能力。

五、总结

本文在传统机器学习算法的基础上引入了函数型数据分析，考虑了随时间变化的债券利率数据，弥补了传统静态算法中难以捕捉动态信息的不足。

通过对实际数据的计算验证，FDA 方法在分析债券利率动态变化方面作用明显，能在传统机器学习算法的基础上，进一步提升模型预测的准确性。其中 XGBoost 与 LightGBM 性能表现优异，结合 FDA 可作为违约预测建模的优选方案。基于研究结论，提出以下应用建议：

（1）金融机构与监管层面：建议监管机构将债券市场利率变化纳入信用监测指标体系；

（2）评级与资管机构：鼓励评级机构与资管机构在预警系统中引入函数型风险监控因子，健全信用债投研系统的动态特征监控模块；

（3）组合管理：利用 FPCA 得分绘制债券组合风险热力图，辅助动态调仓与风险对冲。

未来可进一步拓展非结构化数据，结合深度学习模型以优化预测效果，持续丰富信用风险识别的技术体系。

参考文献

[1] 祝小全，梁洛铭，陈卓 . 中国债券市场风险、定价研究进展与服务实体经济的初步实践 [J]. 中国科学基金，2025，39(02)：329-344.

[2] Chen T，Guestrin C. XGBoost：A scalable tree boosting system[C]//KDD. 2016：785-794.

[3] Zhou L. A survey on machine learning models for credit scoring[J]. WIREs Data Mining Knowl Discov，2013，3(3)：191-203.

[4] 王玉龙，周榴，张涤霏 . 企业债务违约风险预测——基于机器学习的视角 [J]. 财政科学，2022(6)：62-74.

[5] Chopra V，Bhilare P. Gradient boosting for credit risk prediction[J]. International Journal of Computer Applications，2018，182(20)：7-11.

[6] Sebastian H. Goldmann，Marcos R. Machado，Joerg R. Osterrieder. Advancing credit risk assessment in the retail banking industry：A hybrid approach using time series and supervised learning models[J]. Data & Knowledge Engineering，2025，160：102490.

[7] 胡蝶 . 基于随机森林的债券违约分析 [J]. 当代经济，2018(3)：28-30.

[8] 邓晴元 . 基于 XGBoost 算法的债券违约风险预测研究 [J]. 投资与创业，2023，34(2)：1-3.

[9] Tran T T，Tsai P H，Wilson L L. Combining genetic algorithm and deep learning for bankruptcy prediction[C]//ICDM. 2017：617-624.

[10] Gu J，Qian X，Zheng H，et al. Machine learning for credit risk prediction：A macro - micro integrated approach[J]. Journal of Financial Stability，2020，49：100708.

[11] Ramsay J O. When the data are functions[J]. Psychometrika，1982，47(4)：379-396.

[12] Ramsay J O，Dalzell C J. Some tools for functional data analysis[J]. Journal of the Royal Statistical Society：Series B，1991，53(3)：539-572.

[13] 王丙参，魏艳华，张贝贝 . 函数型数据聚类算法的评价与比较 [J]. 统计与决策，2021，37(16)：38-42.

[14] 魏艳华，马立平，王丙参 . 基于函数型数据的中国人口变化趋势及地区差异研究 [J]. 统计与决策，2022，38(8)：82-86.

[15] 王华强，刘黎明 . 基于函数型数据的利率与物价关系研究 [J]. 数理统计与管理，2024，43(3)：452-464.

[16] Z. Tao，M. Wang，J. Liu，et al. A functional data analysis framework incorporating derivative information and mixed-frequency data for predictive modeling of crude oil price[J]. IEEE Transactions on Industrial Informatics，2025，21(4)：3226-3235.

[17] Kokoszka P. Introduction to functional data analysis[M]. CRC Press，2017.

[18] Joaqu í n Sancho Val，Carlos Cajal Hernando，Lourdes Mart í nez de Baños. Functional data analysis of air quality time series in Madrid Using FPCA and splines[J]. Atmospheric Environment，2026，367：121741.

[19] S. Nian，W. Kang，D. Wang. A general approach for health index-based remaining useful life prediction using functional principal component analysis[C]. 2024 Global Reliability and Prognostics and Health Management Conference，2024.

[20] Breiman L. Random forests[J]. Machine Learning，2001，45(1)：5-32.

[21] 姜高霞，王文剑 . 经济周期波动的函数型时序分解方法——基于 CPI 的实证分析 [J]. 统计与信息论坛，2014，29(03)：22-28.

[22] 严明义，杜鹏 . 中国消费价格指数季节变动的函数性数据分析 [J]. 统计与信息论坛，2010，25(08)：100-106.

[23] Chen H，Reiss PT，Tarpey T. Optimally weighted L2 distance for functional data[J]. Biometrics. 2014，70(3)：516-525.

[24] 严明义 . 函数性数据的分析方法与经济应用 [M]. 中国财政经济出版社，2014.

