LendNova：借助语言模型实现自动化信用风险评估

Kiarash Shamsi[1,2], **Danijel Novokmet**[1], **Joshua Peters**[1], **Mao Lin Liu**[1], **Paul K Edwards**[1], **Vahab Khoshdel**[2]

[1]Wealthsimple Inc.,[2]曼尼托巴大学

shamsik1@myumanitoba.ca

arXiv:2601.02573v1 [cs.LG] 5 Jan 2026

摘要

信用风险评估对金融行业至关重要，但传统上依赖成本高昂的基于特征的模型，而这些模型往往无法充分利用原始信用记录中的所有可用信息。本文介绍了 LendNova，这是首个实用的信用风险评估自动化端到端流程，旨在借助先进的自然语言处理技术和语言模型，利用原始信用记录中的所有可用信息。LendNova 通过直接处理术语繁杂的原始征信机构文本，改变了风险建模方式；它使用语言模型学习与任务相关的表示，无需手动进行特征工程。该系统能够自动捕捉文本中蕴含的模式和风险信号，取代手动预处理步骤，从而降低成本并提高可扩展性。在真实世界数据上的评估进一步证明了其在实现准确、高效风险评估方面的巨大潜力。LendNova 为智能信用风险智能体建立了基准，展示了语言模型在这一领域的可行性。它为未来研究奠定了基础，助力构建能够实现更准确、更灵活、更自动化金融决策的基础系统。

信用风险评估对金融行业至关重要，但传统上依赖成本高昂的基于特征的模型，而这些模型往往无法充分利用原始信用记录中的所有可用信息。本文介绍了 LendNova，这是首个用于信用风险评估的实用自动化端到端流程，旨在借助先进的自然语言处理技术和语言模型，利用原始信用记录中的所有可用信息。LendNova 直接处理术语繁杂的征信机构文本，借助语言模型学习与任务相关的表征，无需手动进行特征工程，从而变革了风险建模方式。它能自动捕捉文本中蕴含的模式和风险信号，取代手动预处理步骤，降低成本并提高可扩展性。在真实世界数据上的评估进一步展现了其在准确、高效风险评估方面的巨大潜力。LendNova 为智能信用风险智能体建立了基线，展示了语言模型在这一领域的可行性。它为未来研究奠定了基础，助力构建能够实现更准确、更灵活且自动化决策的基础系统。

AAAI 2026年金融服务领域智能体 AI 研讨会。版权所有 © 2026年，美国人工智能促进协会（www.aaai.org）。保留所有权利。

引言

信用风险建模是金融行业的关键组成部分，在风险管理和决策的诸多方面发挥着重要作用。在信用评估中，信用风险指获批贷款出现逾期还款的概率（Ngai 等，2011年）。信用违约预测模型旨在帮助金融机构判断是否批准贷款申请。通常，这些模型会使用阈值来指导决策者，使其能够根据计算得出的风险水平作出明智的放贷决策。准确评估信用风险至关重要，原因有以下几点：

- **风险管理：** 有效的信用风险模型有助于预测和管理违约造成的潜在损失，使机构能够降低金融风险。稳健的模型支持采取积极措施，防范信用损失；若处理不当，此类损失历来会引发金融危机（Brown 和 Moles 2014年）。
- **资本配置：** 通过准确为风险定价，这些模型能够优化资本的利用，确保资源得到高效配置。这不仅能提高投资回报，还有助于维持流动性和偿付能力，这对金融机构的稳定至关重要（Huang 和 Thakor 2024年）。
- **监管合规：** 金融行业的实体有义务达到严格的监管标准，以确保风险管理稳健。因此，稳健的信用风险评估框架对于遵守强调加强风险管理实践的法规必不可少（Patel 2023年）。
- **盈利能力：** 准确的风险评估有助于改进贷款决策，提升盈利能力并减少坏账发生。借助数据驱动的洞察，金融机构可以完善信用评估流程，降低违约率并改善整体财务状况（Okpala、Osanebi 和 Irinyemi 2019年）。

信用风险模型主要使用征信数据，但也可以纳入申请表、交易记录以及其他替代性金融数据或行为数据等来源，以丰富风险评估。征信数据由原始代码记录构成，这些记录通常缺乏结构，且难以

在没有领域专业知识的情况下进行解读（Stavins 2020年；Gibbs 等，2024年）。这些记录需要通过基于规则的逻辑进行转换，才能形成连贯的信用报告；这一过程技术性很强，且充斥着行话。传统建模方法依赖人工特征选择和专家判断，这限制了可扩展性，也常常忽略数据中有信息价值的模式（Hand 和 Henley，1997年）。此外，传统线性模型难以捕捉特征之间的非线性关系，也容易出现过拟合或欠拟合，从而降低其可靠性和预测能力（Bello，2023年）。这些局限性凸显了在信用风险评估中采用更自动化、数据驱动方法的必要性。尽管理论上信用局数据包含详细的信用行为信息，但实践中仍存在局限。现代序列模型几乎可以通过无穷多种方式创建聚合特征，但征信机构通常只以高昂成本提供数千项基于其职能评估得出的关键特征。由于成本限制，金融机构往往只购买与其建模和监控需求相关、数量少得多的特征子集，这可能导致信息损失。

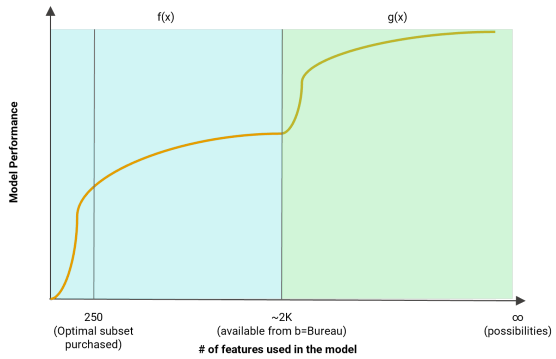


图1：从征信机构聚合特征转向不进行特征聚合、全面使用数据的潜在影响。

图1展示了征信机构数据特征模型的理论假设，强调利用完整数据集提升模型性能的潜在优势。基础函数 $f(x)$ 指征信机构特征空间，其中包含账户存续期、还款记录、信用额度使用率和逾期指标等结构化属性。相比之下，$g(x)$ 表示不经任何聚合而全面使用数据，纳入所有可用数据（包括非结构化信息），从而捕捉复杂的客户行为模式。从 $f(x)$ 转向 $g(x)$，可以显著提高信用风险评估的准确性、降低成本并加快评估流程，同时解决现有局限性。这一转变能够更全面地分析数据，从而做出依据更充分的贷款决策。

LendNova 通过直接应用语言模型分析非结构化征信数据，在信用风险评估方面取得了重大进展。我们的工作主要有三项贡献：（i）提出了信用故事这一概念，这是一种新颖的表示形式，可将充斥着专业术语的信用局数据（通常需要大量领域专业知识才能提取特征）转换为语言模型可直接理解的格式；（ii）展示了首个将语言模型用于工业规模信贷获取风险建模的实际部署，并基于真实生产数据进行了评估；（iii）向研究界介绍了这一范式，作为构建基础模型和智能系统的早期探索，使其能够在金融领域实现自动化端到端决策。这种方法降低了预处理成本，实现了评估流程自动化，同时有效捕捉复杂的数据关系。

LendNova 得到了加拿大数字投资行业一家业界合作伙伴的支持，专注于财富管理领域的应用机器学习。此类合作确保所提出的模型立足于真实金融应用，并应对大规模信用风险评估中的实际挑战。

背景与相关工作

准确的信用风险评估对个人贷款机构以及金融系统的整体稳定性都至关重要。传统方法依赖统计模型和专家判断，但近年来机器学习（ML）和深度学习（DL）的进步改变了这一领域，使可扩展的数据驱动预测成为可能（Nguyen 等，2025年）。本节回顾信用风险评估的发展历程、语言模型在金融领域的引入，以及迈向自动化基础级金融模型的新兴方向。

传统信用风险评估

在现代机器学习方法得到应用之前开发的早期信用风险模型，利用结构化历史数据以及依据专家规则和统计分析得出的手工设计特征来评估借款人信用度（Kiran 等人 2023年）。这些模型往往难以处理大型复杂数据集（Shinde 和 Kale 2023年），也缺乏适应借款人行为和市场状况变化的能力（Paul 等人 2024年；Sabbani 2022年）。机器学习和深度学习的兴起，通过捕捉复杂关系和高维依赖，解决了其中一些问题（Muhindo 等人 2024年；Cai 和 Qian 2018年）。卷积神经网络和循环神经网络等架构实现了对金融行为的时间建模（Duvnjak、Mer´cep 和 Kostanj 2024年；Liu 等人 2021 年），而 Transformer 网络则提升了系统性风险预测中的长程依赖学习能力（Paul 等人 2023年）。集成和混合方法（例如 XGBoost、随机森林）兼顾了准确性与可解释性（Mokheleli 和 Museba 2023年；Schmitt 2022年；Li 等人 2023a）。尽管取得了这些进展，大多数模型仍依赖专家设计特征，限制了可扩展性和泛化能力（Noriega、Rivera 和 Herrera 2023年）。此外，原始数据大多是非结构化的文本数据，需要经过大量预处理才能提取有用信息（Addo、Guegan 和 Hassani 2018年）。

金融领域的语言模型

大语言模型（LLM）的兴起为理解非结构化金-

融数据带来了新的机遇。BERT（Devlin 2018年）和 GPT-4（Achiam 等人 2023年）等模型能够以较高的上下文准确性解读金融文-件、申报文件和交易文本。FinBERT（Araci 2019年）、FLANG（Shah 等人 2022年）和 BloombergGPT（Wu 等人 2023年）等专门变体已被证明在情感分析和金融实体识别方面表现有效（Xie 等人 2023年；Li 等人，2023b）。然而，它们在信用风险预测中的应用仍然有限，因为基于提示词的大语言模型在定量任务中往往难以保持一致性并进行校准（Zhao 等人 2021年；Chen 等人 2023年）。直接基于金融文本训练的微调 Transformer 编码器能够提供更好的上下文稳定性和领域理-解（Hadi 等人 2024年）。然而，很少有研究探讨此类模型如何作为自动化的端到端代理运行，从而解读原始且充斥术语的征信数据；这是本文的核心关注点。

金融领域的基础模型与自动化模型

金融人工智能的最新进展表明，领域正从狭窄的预测模型转向具备长期推理和决策支持能力的自适应通用系统（Bengio 等人 2023年；Park 等人 2023年；Li 等人 2023c）。这些系统旨在从异构数据源中学习，整合文本信息与行为信息，并对不断变化的市场和监管环境作出动态响应。BloombergGPT（Wu 等人 2023年）、FinGPT（Yang、Liu 和 Wang 2023年）以及 PIXIU（Xie 等人 2023年）等大型金融语言模型，体现了这一演进，使金融分析中的跨领域理解与任务迁移成为可能。然而，这些基础模型在很大程度上仍依赖经过整理的文本语料和预定义提示词，因此难以对结构化或非规整金融数据（如征信机构报告）进行推理（Li 等人 2023c）。它们擅长提取一般信息，但并非为信用风险评估中的高风险、特定领域决策而设计（Feng、Dai 等人 2023）。相比之下，LendNova 是一种智能代理，能够将原始征信数据转化为信用风险评估结果。其端到端设计增强了模型捕捉复杂模式的能力，从而带来更强劲、更可靠的预测性能。

数据来源与整合

数据集特征

该数据集来源于一家商业征信机构，包含约一百万名加拿大客户的匿名信用记录。征信数据本质上分为多个部分，每个部分都反映了个人财务概况和信贷行为的不同方面。这些数据并非以单一结构化表格的形式呈现，而是被组织为相互关联的视图，共同描述信贷在一段时间内如何获取、使用和偿还。总体而言，这些部分可归纳为四个概念类别：（i）信贷活动，包括历史账户使用情况、付款模式和未偿余额；（ii）信贷需求，根据信贷查询和应用推断而来，反映借款人获取新信贷的意图；（iii）还款结果和财务压力，由逾期、催收或破产等事件体现，这些事件预示着潜在的违约风险；以及（iv）静态信息层，提供账户类型和报告时间线等背景及参考数据，以确保各部分之间的一致性（Mays 1995年；Kusch 和 Osteen 2013）。这种组织方式能够全面呈现借款人行为，将交易和行为信号整合为统一的时间结构，适用于端到端的自动化信用风险评估。

目标定义

目标变量根据征信查询后的信用表现结果定义，关注的是在 2018年 8 月 31 日之前开立新信用卡的客户。对于每位符合条件的客户，我们在其最后一次有记录的信用卡申请之前，获取了一份个人资料快照，记为 t_0。从该日期起的 18 个月内跟踪信用表现，即截至 $t_1 = t_0 + 18$ 个月。

我们将核销和逾期这两个主要二元指标作为目标，最终标签定义为这些指标的逻辑或。

Let:

- t_0：客户在 2018年 8 月之前最后一次信用卡申请之前的初始快照日期。值得注意的是，t_0 对每位客户而言可能有所不同。
- $t_1 = t_0 + 18$：18个月观察窗口的结束时点。

我们的目标基于两个信贷概念定义：

1.核销目标 $Y_{\text{charge-off}}$：如果贷款机构在信用卡账户开立后18个月内正式宣布该账户无法收回，则该账户发生核销。这一认定意味着贷款机构已判定客户不太可能偿还债务，通常是由于长期未付款，并已将该账户作为损失核销。核销状态指贷款机构报告为无法收回的账户，这些账户不一定属于逾期案例（例如，客户死亡）。核销是信用风险的严重指标，因为这代表未能收回欠款，且往往会导致进一步催收或法律行动。

$$Y_{\text{charge-off}} = \begin{cases} 1 & \text{if a charge-off occurred within } [t_0, t_1], \\ 0 & \text{otherwise.} \end{cases}$$

2. 逾期目标 $Y_{\text{delinquency}}$：如果信用卡账户在开立后18个月内逾期还款90天或以上，则视为发生逾期，这表明客户已连续三期未还款。已被核销为无法收回或涉及破产的账户不纳入此项指标，因为它们代表更严重的财务事件。逾期90天是信用风险的一项关键标志，表明账户存在持续漏还款的情况，并可能面临财务压力。

$$Y_{\text{delinquency}} = \begin{cases} 1 \text{ 若在 } [t_0, t_1], \text{ 则为 } 0\text{。否则} \end{cases}$$

列表 1：征信数据示例

ACCT0202240P0XQ8L7J5VZK9WRN3DF4CT2Y
BAL9988776620180715 SEG0105409000
CAT020481 N 05062010
ENDR040017
ACCT0302230M1R9GQ6F8TYP5ZXV2K3NBH1
BAL7766558820170320 SEG0105409000
CAT020481 N 06082014
ENDR050021

因此，最终目标标签 Y 通过对这两个指标进行逻辑或运算来定义：$Y = Y_{\text{charge-off}} \vee Y_{\text{delinquency}}$ 其中，$Y = 1$ 若在观察窗口内发生核销或逾期，则为真；$Y = 0$ 否则。

如果客户在这 18 个月的窗口期内没有足够的数据，或没有任何信用卡交易，则标签设为 None。此方法提供了一个单一的二元结果：

$$Y = \begin{cases} 1 & \text{if } Y_{\text{charge-off}} = 1 \text{ or } Y_{\text{delinquency}} = 1, \\ 0 & \text{if } Y_{\text{charge-off}} = 0 \text{ and } Y_{\text{delinquency}} = 0, \\ \text{None} & \text{if data is insufficient or unavailable.} \end{cases}$$

核销和逾期的分布极不均衡：与保持良好信用行为的客户相比，经历核销的客户比例要低得多。核销发生频率较低，因为大多数客户都会按时还款，只有一小部分客户未能偿还债务，最终导致核销。这种不平衡意味着，核销在数据集中属于罕见事件，因此，若不采用专门技术，模型就更难准确预测这类结果；这是因为大多数案例中的客户并未遭遇财务困境（Alam 等人，2020）。

伦理考量

所有数据均已完全匿名化，仅保留不可识别的属性和匿名化的记录标识符。每条记录均代表某位客户在 t_0 之前最近一次的信贷申请档案，所有信用属性均保留为快照日期时的状态，该日期称为运行日期。

数据挑战

征信机构数据集极为复杂，其中包含密集的数字代码和专业金融术语，这些都给预训练语言模型带来了挑战。实际上，从这些数据中汇总有用特征，需要耗费大量成本进行人工整合，或向征信机构本身购买外部特征。该数据集如同一种领域专用语言，有别于用于训练大多数语言模型的自然文本语料库。列表 1 展示了征信机构记录中常见的编码和行话程度，凸显了有效领域预处理的必要性，以使这些记录适用于基于语言的建模。

系统概览

在 LendNova 中，我们提出了一种极具实用性且经济高效的信用风险评估模型，旨在满足现实应用中公认的性能标准。该系统通过减少运营开销、提高效率，简化了评估流程。其架构由三个关键模块组成：(i) 数据准备阶段，用于处理和整理原始征信数据；(ii) 语言模型，将处理后的信息转换为有意义的嵌入；(iii) 任务预测器，利用这些嵌入估计信用风险结果。本节将详细介绍各个模块，图 2 展示了该框架的概览。

数据准备

本节介绍将原始征信数据转换为适合语言模型微调的结构化格式的流程。鉴于征信数据充斥着行话且采用编码语言，使用预训练语言模型需要进行仔细的预处理。这些预训练模型通常基于标准英语构建，若不加以修改，便无法直接理解征信机构的专业术语。研究表明，在特定领域（如金融）中，微调预训练模型通常比从头开始训练更高效、更有效，因为从头进行模型训练需要耗费大量时间和资源（Li 等人，2023b；Brief 等人，2024）。数据准备模块包括以下步骤：

记录解析器模块 记录解析器模块将非结构化的征信机构记录转换为结构化的向量格式，其中每个字段均按照预定义顺序排列，以供下游流程使用。对于给定的原始记录 $R = \{r_1, r_2, \ldots, r_n\}$，解析器应用函数 f，生成结构化向量：$F = f(R) = \{f_1, f_2, \ldots, f_n\}$ 其中 f_i 表示按固定顺序排列的各个解析字段。此转换确保了顺序和模式的一致性，使后续的分段和分析成为可能。

分段模块 分段模块将相关字段归为多个分段 S_i，每个分段代表用户行为的不同方面。对于每个解析后的向量 F，分段函数 g 会将每个 f_i 分配到相应的分段：$\{S_1, S_2, \ldots, S_k\} = g(F)$ 每个分段 S_i 会汇总与特定行为相关的字段，因此可以灵活选择最适合各种任务的分段。针对特定任务，可以选择关键分段，从而简化分析。在我们的设置中，我们重点关注三个主要分段：交易（TR）、信贷查询（IN）和催收（CL），以捕捉信用行为的关键要素。交易分段全面展现个人的信贷产品，包括债务的类型和时间远近。信贷查询分段揭示信贷申请模式，其中较高或近期的查询量可能表明存在信贷需求或潜在风险。催收分段突出以往的催收事件，可直接反映风险。这些分段共同构成个人信用档案的核心，其他分段则提供补充信息。

翻译模块 翻译模块将每个片段 S_i 从编码文本转换为简明英语，生成我们称之为用户信用故事 的内容。

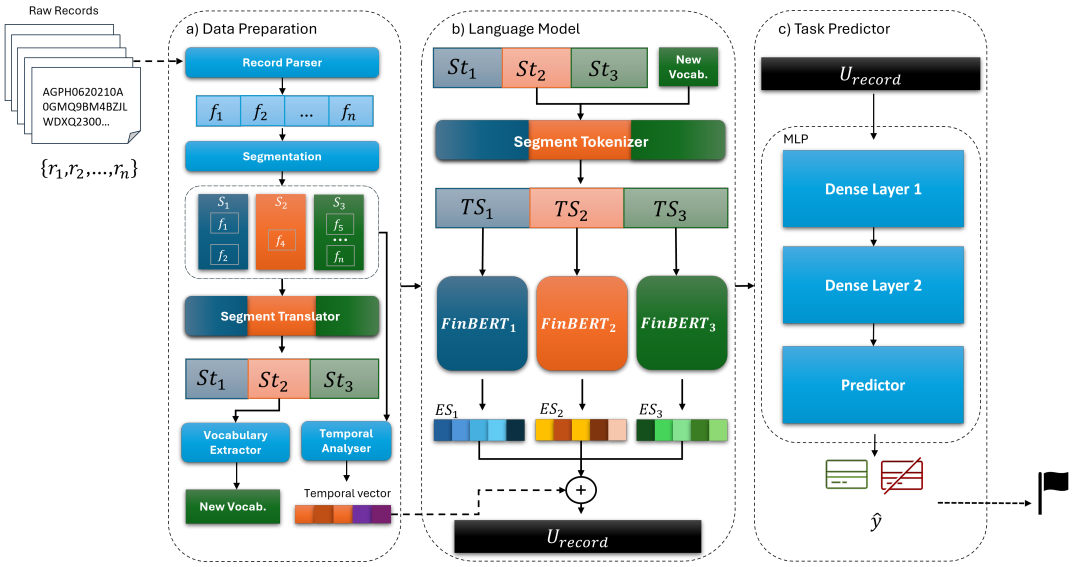


图 2：LendNova系统架构。该图展示了 LendNova 的三个主要组件：(a) 数据准备，将原始输入转换为结构化的信用故事 St_n（例如， St_1， St_2，.....）；(b) 语言模型，使用时间向量对信用故事进行嵌入，每个分词片段表示为 TS_n（例如， TS_1， TS_2，.....）；以及 (c) 任务预测器，基于这些嵌入进行训练，得出最终预测结果。

我们为每个片段 T_i 定义一个翻译函数 S_i：$\text{St}_i = T_i(S_i)$，其中 St_i 表示 S_i的简明英语等价表达。因此，用户的完整信用故事成为集合 $\{\text{St}_1, \text{St}_2, \ldots\ldots, \text{St}_k\}$，以适合语言模型嵌入的可解释格式呈现每种行为或记录。

词汇提取模块 词汇提取模块会识别每个翻译片段 St_i 中的独特信用专用术语，以构建统一的领域专用词汇表。对于每个片段，词汇提取函数 v_i 会提取术语，形成各自的词汇表 V_i。随后，将这些词汇表拼接起来，得到用于分词的整体词汇表 V：$V = \bigcup_{i=1}^{k} V_i$ 这个扩展后的词汇表使分词器能够生成包含信用专用语言模式的嵌入，从而确保在分词过程中识别这些领域专用术语。

时间分析模块 时间分析模块捕捉每个分段中的时间依赖行为。每条记录都有一个基础“运行日期” t_0，在每个分段 S_i中，记录包含带有独立日期的字段$\{t_1, t_2, \ldots, t_m\}$。对于记录中的每个日期字段，我们计算这些日期与运行日期之间的差值t_0，从而得到一组相对日期：$\Delta t_j = t_j - t_0, \quad \forall t_j \in S_i$对于每个分段，我们将这些相对日期汇总为一个统一的时间向量 T_{S_i}，方法是计算所有记录中每个日期字段的最小值、最大值和平均值：$T_{S_i} = \{\min(\Delta t), \max(\Delta t), \text{avg}(\Delta t)\}$ 这个时间向量 T_{S_i} 表示每个分段的时间依赖模式，有助于分析客户行为随时间的变化，并基于时间趋势生成预测性洞见。接着，我们将每个分段专属的时间向量 T_{S_i}拼接起来，为每个客户创建一个统一的时间行为向量 T_{record}，用于表示客户信用记录中的整体时间模式。这个统一的时间向量 T_{record}表示为：$T_{\text{record}} = \{T_{S_1}, T_{S_2}, \ldots, T_{S_n}\}$ 这个汇总后的时间向量 T_{record} 使我们能够根据所有分段汇总出的时间趋势，分析并预测客户行为随时间的变化。这些模块共同将征信机构专业且充斥行话的数据转换为标准的自然语言处理格式，使其具备领域针对性和时间感知能力，从而可用于有效的模型训练与解读。

语言模型

现在，我们已将原始征信数据转换为结构化的简明英语信用故事，接下来将对语言模型进行微调，使其能够有效处理这些金融数据。由于 FinBERT 已在大规模金融文本语料库上进行预训练，因此是理想的基础模型：它既理解特定领域的术语，也具备分析金融文本的稳健架构（Yang、Uy 和 Huang，2020年）。这种方法减少了从头开始训练的需求，并充分利用迁移学习在该领域的优势，显著增强模型对金融数据的泛化能力。

语言模型模块由三个主要子模块构成：

自定义分词器 我们使用 FinBERT 的分词器处理每个信用故事 St_i。通过添加特定领域的词汇，我们确保对信用相关术语进行精确分词。其表示如下：$TS_i = \text{tokenize}(St_i)$，适用于每个分段 S_i。这种自定义分词方式将多个词组成的术语合并为单个词元，从而减少词元数量，并提升模型对信贷专属语言的理解能力。

并行 FinBERT 模型 每个信贷分段 S_i都有独特的属性，因此需要分别映射为分段专属的嵌入。对于每个分段，我们计算其嵌入： $E_{S_i} = FinBERT_i(TS_i)$。这种并行处理会将每个分段映射到不同的潜在空间，为所有分段生成嵌入 $\{E_{S_1}, E_{S_2}, \ldots, E_{S_n}\}$。

聚合池化 在此步骤中，我们将所有分段的嵌入 $E_{S_1}, E_{S_2}, \ldots, E_{S_n}$ 与时间向量 T_{record} 结合，形成一个统一向量 U_{record}，用于表示用户的信用行为。每个分段的嵌入与时间向量拼接在一起，构成最终向量： $U_{\text{record}} = [E_{S_1}, E_{S_2}, \ldots, E_{S_n}, T_{\text{record}}]$

这个统一向量 U_{record} 既捕捉了用户信用行为中的文本模式（通过分段嵌入），也捕捉了时间模式（通过时间向量）。随后，生成的向量 U_{record} 会被传递到任务预测模块进行进一步分析。

任务预测器

在整个流程的最后阶段，聚合后的表示 U_{record} 用于预测信用风险目标。这个统一向量融合了分段嵌入和时间信息，将输入多层感知机 (MLP)模型进行最终预测：$\hat{y} = \text{MLP}(U_{\text{record}})$，其中$\hat{y}$ 表示预测的违约概率。MLP学习将组合后的文本数据和时间数据映射为信用风险目标预测。为了解决数据集中的类别不平衡问题，我们在训练过程中采用加权二元交叉熵损失。该损失函数为正类（违约）赋予更高的权重，帮助模型更加关注少数类，从而提高违约预测的表现。

该模型针对目标预测进行端到端训练，整个流程通过这一加权损失函数进行联合优化，以预测信用风险。这确保模型能够有效处理类别不平衡问题，并提供准确的信用风险预测。

实验

在本节中，我们开展了广泛的实验，以评估模型的性能和实际应用性。具体而言，我们的实验围绕以下研究问题展开：

- **RQ1**：我们为信用风险评估设计的基于语言模型的框架，能否在现实世界的行业场景中展现出强劲的性能，满足此类模型在实际应用中的理想性能要求?
- **RQ2**：在实际应用中，使用该模型能否节省成本? 如果可以，与实际应用中的传统方法相比，该模型在成本和资源节省方面的效率如何?

通过这些实验，我们评估了所提出模型在信用风险预测方面的实际效果和资源效率。

实验设置

为贴近真实世界的条件，我们将数据划分为训练集、验证集和两个测试集：留出集与时间外（OOT）测试集。留出集与训练集和验证集处于相同的时间窗口，而 OOT 集由训练时间范围之后的记录组成，用于模拟实际部署场景。每条记录都以个人的“运行日期”为基准，代表其在2018年9月之前申请信用卡的最近日期。在训练和初始验证中，我们选取运行日期为2018年2月1日或更早的记录，并按如下方式划分：

- **训练数据**：占抽样数据的 60%，用于训练模型并生成基线预测十分位数。这些十分位数由预测概率得出，可通过衡量总体相对于这些基线的稳定性来评估模型稳定性。
- **验证数据**：占抽样数据的 20%，用于模型调优以及训练期间的早停。
- **测试数据**：剩余的 20%，用于评估模型整体表现和泛化能力。该数据集分为 OOT（训练时间范围之后）和留出集（与训练时间范围相同）。

数据不平衡：数据存在严重不平衡，目标事件率低于5%。OOT 集中的事件率预计会略低，这反映了 2018 年的市场状况：与前几年相比，整体违约数量更少。

评估模型表现的行业标准是征信机构提供的信用评分，该评分根据从信用记录中提取的大量特征计算得出。在我们的特定数据集上，该信用评分用于违约预测的曲线下面积约为 0.8，可作为实际信用风险评估的基准。我们为信用风险建模提出了一种新的基线和方向，展示了基于语言的方法在理解征信数据方面的有效性。这项工作为开发金融基础模型铺平了道路，使其能够支持信用与风险管理领域更广泛的决策。

结果

工业语言模型在信用违约预测中的可接受表现（RQ1）。我们实施了一个迭代式模型构建框架，旨在充分发挥语言模型的潜力，使其性能达到与行业基线相当的水平，同时避免传统模型通常需要的高成本步骤。我们的模型经历了六个关键版本，如图3所示：

- **版本1**：我们建立了初始模型流程，但未使用自定义分词器。每个数据段都配有一个独立的分类器，最终预测由多数投票决定。由于资源有限，训练以较小的数据分桶为单位逐步进行。
- **版本2**：新增了自定义分词器，以纳入特定的信用领域词汇，从而提高分词后的数据的相关性。

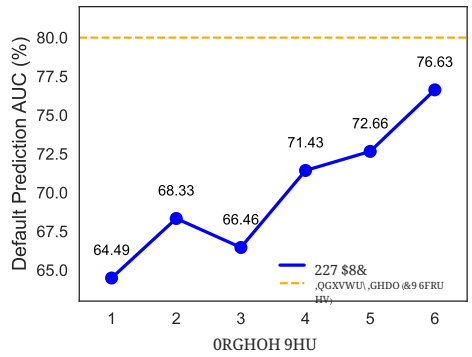


图 3：模型版本的性能趋势。该图展示了模型开发各阶段 AUC 分数稳步提升的趋势，最终版本的表现已接近行业理想基线，体现了 LendNova 在信用风险预测方面的潜力。

- **版本 3**：我们通过对正样本进行过采样，引入了标签平衡，以解决标签不平衡问题，并实现了 70-30 的平衡。不过，这项调整的效果有所下降，并在后续版本中进行了修改。
- **版本4**：优化了分词以去除冗余内容，并修改训练循环，使用单一的综合数据集，使其更便于与传统模型进行比较。标签平衡调整为 90-10，以贴近现实世界分布。

- **版本 5**：池化机制改为直接聚合嵌入，并将统一向量输入扩展后的多层感知机层。此项改进提升了模型处理聚合输入的能力，同时取消了标签平衡，使分布更自然。

- **版本 6**：我们的冠军模型（如图 2 所示）将时序信息注入文本嵌入中，进一步提升了预测准确性。

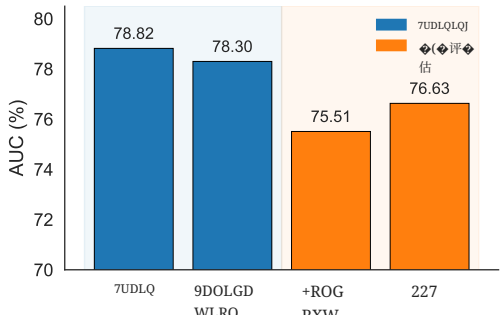


图4：冠军模型在不同数据划分上的曲线下面积

经过这六个版本的迭代，我们的最终模型与工业基线的曲线下面积差距仅为3.63%，展现了语言模型在信用风险预测中的潜力。这种自动化、可扩展的方法表明，该方法有望泛化到信用领域内的其他任务。图4展示了最终版本模型在不同数据划分上的表现。

语言模型在信用违约预测中的成本效益（研究问题二）。我们的工业规模评估表明，与传统的信用风险建模方法相比，LendNova具有显著的成本效益。由于框架的机密性及其商业化部署，具体数值无法披露。我们将重点介绍LendNova展现出可衡量效率和进一步优化潜力的关键领域。

- **数据许可效率：**传统模型依赖征信机构汇总特征，其成本会随着数据量和购买的数据包而增加。LendNova 直接使用原始征信数据，从而免去重复的许可费用，并保持近线性的可扩展性。
- **多任务学习可扩展性：**传统系统针对每项任务分别训练模型，增加了维护成本。LendNova 作为统一的基础模型，可在一个框架内支持多项信贷决策任务，从而提升可扩展性并提高资源利用率。

- **训练与计算效率：**虽然大语言模型的初始训练成本较高，但这项成本可由多个任务分摊。与重新训练多个独立模型相比，微调和推理所需的资源少得多。
- **特征工程自动化：**传统工作流程依赖特征工程步骤，增加了延迟和复杂性。LendNova 通过从非结构化信贷叙述中提取表示来自动完成这一步骤，从而降低预处理开销。
- **成本节约与实际影响：**该框架的规模扩展与数据获取成本、购买的特征数量以及客户数量呈线性关系。理论上，它可以直接从文本中提取无限数量的特征，而无需支付许可费用。在加拿大等大型市场中，通过统一建模并降低数据依赖，该框架每年有望节省数百万美元。

这些观察结果表明，LendNova 为可扩展且经济高效的信用建模提供了有前景的基础，并可通过模型扩展和流程优化进一步提升。

结论

我们的研究提出将语言模型作为端到端信用风险评估的自动化系统，能够直接基于征信数据执行端到端评估。借助 Lend-Nova，我们为自动化信用建模确立了新的基线，表明此类模型能够解读金融模式，并在无需人工干预的情况下支持可靠决策。通过更大规模的训练，这一方向有望进一步发展，超越征信机构信用评分基线；同时，还可扩展其架构，以捕捉更丰富的金融信号。由于征信数据支撑着银行业的许多关键决策，这一框架可以自然地扩展到这些任务。展望未来，本研究为开发多任务、多领域的金融基础模型奠定了基础，从而在更广泛的金融生态系统中统一风险评估与决策支持。

参考文献

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altensch midt, J.; Altman, S.; Anadkat, S.; 等. 2023. GPT-4技术报告。*arXiv*预印本 *arXiv:2303.08774*。Addo, P. M.; Guegan, D.; 和 Hassani, B. 2018. 使用机器学习和深度学习模型进行信用风险分析。《*Risks*》，6(2)：38。Alam, T. M.; Shaukat, K.; Hameed, I. A.; Luo, S.; Sarwar, M. U.; Shabbir, S.; Li, J.; 和 Khushi, M. 2020. 对类别不平衡数据集中信用卡违约预测的调查。《*IEEE Access*》，8：201173–201198。Araci, D. 2019. FinBERT：使用预训练语言模型进行金融情感分析。*arXiv*预印本 *arXiv:1908.10063*。Bello, O. A. 2023. 用于信用风险评估的机器学习算法：经济与金融分析。《国际管理杂志》，10(1)：109–133。Bengio, Y.; Hinton, G.; Yao, A.; 等. 2023. 在快速发展中管理极端人工智能风险。Brief, M.; Ovadia, O.; Shenderovitz, G.; Yoash, N. B.; Lemberg, R.; 和 Sheetrit, E. 2024. 混合训练：多任务微调对大语言模型性能的混合效应——金融领域案例研究。*arXiv*，abs/2410.01109。Brown, K.; 和 Moles, P. 2014. 信用风险管理。*K. Brown & P. Moles*，《信用风险管理》，16。Cai, Q.; 和 Qian, Q. 2018. 信用风险评估方法综述。Chen, A.; Phang, J.; Parrish, A.; Padmakumar, V.; Zhao, C.; Bowman, S. R.; 和 Cho, K. 2023. 大语言模型多步推理中的自洽性缺失的两种情形。*arXiv*预印本 *arXiv:2305.14279*。Devlin, J. 2018. BERT：用于语言理解的深度双向变换器预训练。*arXiv*预印本 *arXiv:1810.04805*。Duvnjak, M.; Mer´cep, A.; 和 Kostanj, Z. 2024. 用于信用风险评估的内在可解释模型。*2024*年第*47*届*MIPRO*信息通信技术与电子大会（*MIPRO*），31–36。Feng, D.; Dai, G.; 等. 2023. 惠及多数，偏向少数：通过大语言模型实现通用信用评分。Gibbs, C. N.; Guttman-Kenney, B.; Lee, D.; Nelson, S. T.; van der Klaauw, W. H.; 和 Wang, J. 2024. 消费者信用报告数据。技术报告，美国国家经济研究局。Hadi, M. U.; Al Tashi, Q.; Shah, A.; Qureshi, R.; Muneer, A.; Irfan, M.; Zafar, A.; Shaikh, M. B.; Akhtar, N.; Wu, J.; 等. 2024. 大语言模型：应用、挑战、局限性与未来前景综述。*Authorea*预印本。Hand, D. J.; 和 Henley, W. E. 1997. 消费者信用评分中的统计分类方法：综述。《皇家统计学会杂志：*A*辑（社会统计学）》，160(3)：523–541。Huang, S.; 和 Thakor, A. V. 2024. 政治影响力、银行资本与信贷配置。《管理科学》。

Kiran, A.; Gongada, T. N.; Arangi, V.; Ahmad, A. Y. A. B.; Dhabliya, D.; and Gupta, A. 2023年。评估机器学习算法在金融信用风险评估中的表现。*2023*年第三届电子与通信工程进展国际会议（*AECE*），881–886. Kusch, B.; and Osteen, S. R. 2013. 金融谜题：信用报告与评分。Li, X.; et al. 2023a. 探索机器学习技术预测旅行保险理赔的潜力：四种模型的比较分析。《计算与信息科学学术期刊》, 6(4): 118–125. Li, Y.; Wang, S.; Ding, H.; and Chen, H. 2023b. 金融领域的大语言模型：综述。载于第四届*ACM*金融人工智能国际会议论文集, 374–382. Li, Y.; Wang, S.; Ding, H.; and Chen, H. 2023c. 金融领域的大语言模型：综述。Liu, Q.; Liu, Z.; Zhang, H.; Chen, Y.; and Zhu, J. 2021. 挖掘交叉特征以进行金融信用风险评估。第*30*届*ACM*国际信息与知识管理会议论文集。Mays, E. 1995年。信用评分手册。Global Professional Publishi. Mokheleli, T.; and Museba, T. 2023. 用于信用评分预测的机器学习方法。《信息系统与信息学期刊》, 5(2): 497–517. Muhindo, J.; Mukasa, K.; Kitakufe, D.; and Kato, J. 2024年。推进信用风险评估与金融决策：整合现代技术与洞见。《世界高级研究与评论期刊》。Ngai, E. W.; Hu, Y.; Wong, Y. H.; Chen, Y.; and Sun, X. 2011年。数据挖掘技术在金融欺诈检测中的应用：分类框架与文献学术综述。《决策支持系统》, 50(3): 559–569. Nguyen, Q. G.; Nguyen, L. H.; Hosen, M. M.; Rasel, M.; Shorna, J. F.; Mia, M. S.; and Khan, S. I. 2025年。运用机器学习增强信用风险管理：信用违约预测模型的比较研究。《美国应用科学期刊》, 7(01): 21–30. Noriega, J. P.; Rivera, L. A.; and Herrera, J. A. 2023年。用于信用风险预测的机器学习：系统性文献综述。数据, 8(11): 169. Okpala, K. E.; Osanebi, C.; and Irinyemi, A. 2019年。信用管理策略对流动性和盈利能力的影响。《行为研究期刊》, 1(1): 1–14. Park, J. S.; O'Brien, J. C.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023年。生成式智能体：人类行为的交互式模拟。Patel, K. 2023年。信用卡分析：欺诈检测与风险评估技术综述。《国际计算机趋势与技术期刊》, 71(10): 69–79.

Paul, M.; Nagwovuma, M.; Nansamba, B.; Hellen, N.; Jingo, D.; and Marvin, G. 2024年。用于信用风险评估的可信深度学习技术。*2024*年第五届电子与可持续通信系统国际会议（*ICESC*）, 1949–1962. Paul, S.; Gupta, A.; Kar, A. K.; and Singh, V. 2023年。一种基于深度强化学习的自动信用评分模型：采用深度Q网络对客户信用申请进行分类。*2023*年*IEEE*国际技术与社会研讨会（*ISTAS*）, 1–8. Sabbani, G. 2022. 银行业信用风险评估中的机器学习：概述。《人工智能与云计算杂志》。Schmitt, M. 2022. 深度学习与梯度提升：信用评分领域最先进机器学习算法的基准测试。*arXiv*预印本 *arXiv:2205.10535*。Shah, R. S.; Chawla, K.; Eidnani, D.; Shah, A.; Du, W.; Chava, S.; Raman, N.; Smiley, C.; Chen, J.; and Yang, D. 2022. 当flue遇上flang：金融领域的基准测试与大型预训练语言模型。*arXiv*预印本 *arXiv:2211.00083*。Shinde, S.; and Kale, S. 2023年。释放机器学习的力量：信用风险评估分类算法的比较研究。《国际科学、传播与技术高级研究期刊》。Stavins, J. 2020年。信用卡债务与消费者支付选择：我们能从信用局数据中了解到什么？《金融服务研究杂志》, 58(1): 59–90. Wu, S.; Irsoy, O.; Lu, S.; Dabravolski, V.; Dredze, M.; Gehrmann, S.; Kambadur, P.; Rosenberg, D.; and Mann, G. 2023年。BloombergGPT：面向金融领域的大语言模型。*arXiv*预印本 *arXiv:2303.17564*。Xie, Q.; Han, W.; Zhang, X.; Lai, Y.; Peng, M.; Lopez-Lira, A.; and Huang, J. 2023年。PIXIU：面向金融领域的大语言模型、指令数据与评估基准。*arXiv*预印本 *arXiv:2306.05443*。Yang, H.; Liu, X.-Y.; and Wang, C. D. 2023年。FinGPT：开源金融大语言模型。*arXiv*预印本 *arXiv:2306.06031*。Yang, Y.; Uy, M. C. S.; and Huang, A. 2020年。FinBERT：面向金融通信的预训练语言模型。*arXiv*预印本 *arXiv:2006.08097*。Zhao, Z.; Wallace, E.; Feng, S.; Klein, D.; and Singh, S. 2021年。使用前先校准：提升语言模型的小样本性能。载于国际机器学习会议, 12697–12706. PMLR.