

摘要

消费是经济稳定增长的“压舱石”，是拉动经济增长的“三驾马车”之一，而消费金融则是释放内需潜力，促进居民消费的重要抓手。基于消费金融的重要意义，各类金融机构也开始积极地拓展消费金融业务以促进消费市场的发展和满足消费需求的变化。

在当今大数据时代背景下，互联网消费金融业务凭借其高效、精准、实时、低成本等优点快速成长，但在迅速发展的同时也涌现出不少风险和挑战。由于互联网相关业务都是通过线上处理，具有不透明性，同时个人征信系统也存在不完善的情况，借款人的信用风险较高，因此提高信贷风控模型的预测能力成为构建信贷风控体系的重点。从用户的行为数据提取复杂的交互特征，以便挖掘出更多有效信息是提升预测效果的关键。用户特征一般是高维且稀疏的，特征之间的组合关系并不明显，传统的信贷模型无法通过算法实现特征之间的自动组合，需要在特征工程阶段根据经验手动构造交互特征，但是手动构造复杂的交互特征不仅需要消耗大量的人力且不能保证构造出的特征的有效性。

最初被应用于点击率预估的 DeepFM 模型能够通过模型特性自动进行特征交互。但是，该模型存在低阶交互特征提取不够完备且只能学习隐式的高阶特征组合等问题。本文在 DeepFM 模型的基础上进行改进并提出了 AMDFM 模型，相比 DeepFM，该模型在提取低阶特征交互时引入了注意力机制来学习不同组合特征的权重，在高阶特征交互方面引入多头自注意力机制模块，使模型不仅可以学习隐式的高阶特征组合，还可以学习显式的高阶特征组合，提高了模型的预测能力。本文主要基于 Lending Club 数据集，对提出的 AMDFM 模型进行实证研究。从模型超参数影响、模型性能、模型可解释性以及特征交互层模块消融实验等多个方面证明本文所提模型的有效性。

综上，本文构建了一种基于改进 DeepFM 的信贷风控模型。根据挖掘交

互特征的思想，通过在模型的特征交互层，引入传统注意力机制改进低阶特征交互模块，引入多头自注意力机制丰富高阶特征提取，增强了模型对于交互特征的学习能力，使模型获得了更好的预测性能。

关键词：消费金融、特征交互、DeepFM、多头自注意力机制

Abstract

Consumption is the "ballast stone" of stable economic growth and one of the "troika" driving economic growth, while consumer finance is an important tool to unleash the potential of domestic demand and promote consumer spending. Based on the importance of consumer finance, various financial institutions have started to actively expand their consumer finance business to promote the development of the consumer market and meet the changes in consumer demand.

In the context of today's big data era, the Internet consumer finance business is growing rapidly with its advantages of efficiency, accuracy, real time and low cost, but along with the rapid development, many risks and challenges have emerged. Since Internet-related businesses are processed online with opaqueness, and there are also imperfections in personal credit systems, borrowers' credit risk is high, so improving the predictive capability of credit risk control models has become the focus of building credit risk control systems. Extracting complex interaction features from user's behavioral data in order to mine more effective information is the key to improve the prediction effect. User features are generally high-dimensional and sparse, and the combination relationship between features is not obvious. Traditional credit models cannot automatically combine features through algorithms, and need to manually construct interaction features based on experience in the feature engineering stage, but manually constructing complex interaction features not only consumes a lot of labor but also cannot guarantee the validity of the constructed features.

The DeepFM model, which was initially applied to click-through rate prediction, can automatically perform feature interactions by model characteristics. However, the model suffers from incomplete extraction of low-order interaction features and can only learn implicit combinations of high-order features. In this paper, we improve the DeepFM model and propose the AMDFM model, which introduces an attention mechanism to learn the weights of different combinations of

features when extracting low-order feature interactions, and introduces a multi-headed self-attention mechanism module for high-order feature interactions, so that the model can learn not only implicit high-order feature combinations but also explicit high-order feature combinations, which improves the predictive capability of the model. improves the predictive capability of the model. In this paper, we empirically investigate the proposed AMDFM model based on the Lending Club dataset. The effectiveness of the proposed model is demonstrated in terms of model hyperparameter impact, model performance, model interpretability and feature interaction layer module ablation experiments.

In summary, a credit risk control model based on improved DeepFM is constructed in this paper. According to the idea of mining interaction features, by introducing the traditional attention mechanism to improve the low-order feature interaction module and the multi-headed self-attention mechanism to enrich the high-order feature extraction in the feature interaction layer of the model, the learning ability of the model for interaction features is enhanced, so that the model obtains better prediction performance.

Keywords: consumer finance, feature interaction, DeepFM, multi-headed self-attention mechanism;

目录

1.绪论.....	1
1.1 研究背景	1
1.2 研究意义	2
1.3 研究内容和创新点	3
1.4 本文组织结构	4
2.研究现状.....	5
2.1 信贷风控模型的研究现状	5
2.2 特征交互模型的研究现状	8
3. 基于改进 DeepFM 的信贷风控模型的构建.....	12
3.1 特征交互模型	12
3.2 改进 DeepFM 模型的设计背景	20
3.3 AMDFM 模型整体的提出	20
3.4 特征交互层	21
3.4.1 AFM 模块	22
3.4.2 多头自注意力机制模块.....	22
3.4.3 DNN 模块	23
4.实证研究.....	25
4.1 实验环境	25
4.2 实验数据特征工程	25
4.2.1 数据来源及介绍	25
4.2.2 数据预处理.....	27
4.2.3 特征交互	32
4.3 评价指标	32
4.4 实验结果分析	35

4.4.1 模型超参数分析	35
4.4.2 模型性能对比	44
4.4.3 模型可解释性分析	48
4.4.4 模型消融实验分析	50
5.总结与展望	52
5.1 论文工作总结	52
5.2 未来工作展望	53
参考文献	54
致谢	59

1.绪论

1.1 研究背景

目前,我国经济发展正面临着国内经济增速放缓、经济增长动能转换、对外贸易需求下降等多重压力。为应对复杂的国内外形势、实现经济持续高质量发展,中央做出“加快构建以国内大循环为主体、国内国际双循环相互促进的新发展格局”的重要战略决策。扩大内需是促进内循环的重要战略基点,提升居民消费则是扩大内需的关键。但是,近年来我国居民消费率下降、消费升级放缓等问题显现,流动性约束的存在是上述现象产生的重要原因。赵素芳等学者^[1]认为,消费金融可在一定程度上缓解居民进行资金跨期分配的限制,从而起到平滑消费、促进消费升级的作用。

消费金融的出现在拉动我国内需、促进我国消费结构升级等方面的作用,宋金恢等学者^[2]总结为两个方面。首先,消费金融为消费者提供了超前消费的新消费方式,使其消费欲望转化为消费行为,进而能够拉动内需。其次,消费者可通过分期付款来解决短期资金不足却又急需购买高价商品的问题,从而能够促进我国消费结构的升级。

消费金融为推动经济发展提供动力的同时也面临着不少挑战,其中,如何准确预测用户贷款的违约概率从而规避借款者的信用风险是关键。征信体系不够完善一直影响着我国金融领域的发展,部分消费金融机构的自我风控能力不足,不能有效应对市场中普遍存在的高风险。为防止系统性金融风险的产生,监管部门也在不断加大对消费金融行业的规范力度,积极促进行业的正向发展。但相对于监管部门的外部驱动,消费金融机构自主风险控制能力提升的内在突破对于行业的健康发展有着更加积极的意义。因此,信用体系建设不完善和公司风险控制能力不足在一定程度上也在限制我国消费金融

的高质量发展。

传统金融机构和互联网金融机构因需求场景的改变使得在建立信贷风控模型时也存在不同，传统的金融机构如银行一般基于借款人详细的征信信息通过建立评分卡进行信用评估，但是在互联网场景下的信贷风控领域，需要面对征信信息不完善、多变量、复杂特征的业务场景，评分卡模型不再适用。随着科技的不断发展，机器学习算法、深度学习算法也逐渐被运用到了信贷风控模型的建设中，利用模型从借款人的借款信息、短信记录、社交关系等多维度的数据中，挖掘出借款人的潜在行为模式，弥补了缺少征信信息的劣势，提高了信用评估的准确率。

1.2 研究意义

在信贷业务中，风险是永恒的主题。互联网消费金融行业完全基于网络环境，具有一定程度的不透明性，因此将要面临的欺诈手段会更加多样，风险问题更加严峻，平台的风控能力与业务收益之间息息相关。对风险进行识别、评估、管理是保证互联网消费金融行业健康发展的重要课题。

在互联网金融领域，由于相关业务都是通过线上平台处理，获取的相关信息有限且具有不透明性，因此建立一个高效、稳健的信贷风控模型的关键在于从有限的用户特征中尽可能地提取出更加丰富的信息，比如通过用户的借款行为特征、短信记录、APP 安装情况等挖掘其行为模式，进而对用户的信用进行评估。除了从多维度的特征中获取信息，还可以对特征进行组合，将用户的某些组合特征与违约预测建立联系。例如，将贷款金额、月收入等原始特征进行交互，贷款金额高而月收入低的用户的违约概率与贷款金额高、月收入高或贷款金额低、月收入低的用户违约概率可能会存在一定的差异。然而用户特征一般是高维且稀疏的，特征之间的组合关系并不明显，传统的信贷模型如果想要学习交互特征，需要在特征工程阶段根据经验人工构造，但是手动构造复杂的交互特征不仅需要消耗大量的人力且不能保证构造出的特征的有效性。如果能通过模型自动实现构造交互特征的过程将节省大量的时间成本和人工成本。

最初被应用于点击率预估的 DeepFM 模型能够通过模型特性自动进行特

征交互，本文将在 DeepFM 的基础上进行改进，使用基于改进的 DeepFM 来构建信贷风控模型。利用模型自动学习交互特征，挖掘其中的组合关系，并与违约预测建立联系，这对互联网消费金融机构构建风控体系，减少不良贷款，节约经济成本、促进互联网消费金融的良性发展具有重要意义。

1.3 研究内容和创新点

本文研究内容如下，首先对互联网消费金融真实业务场景的实际情况进行分析得出交互特征对于模型预测的重要作用，然后考虑将点击率预估领域的特征交互模型应用到信贷风控领域中。接下来充分了解特征交互模型的发展过程，分析各模型的特点，对 DeepFM 模型的现有问题进行分析并提出改进方案，提出了 AMDFM 模型。该模型在 DeepFM 模型的基础上做出以下改进，在低阶特征交互方面，通过引入注意力机制来学习不同组合特征的权重；在高阶特征交互方面，在保留原模型中深度神经网络模块的基础上，通过引入多头自注意力机制模块，使模型不仅可以学习隐式的高阶特征组合，还可以学习显式的高阶特征组合。最后，对本文提出的 AMDFM 模型进行实证研究，从模型超参数影响、模型性能、模型可解释性以及特征交互层模块消融实验等多个方面证明本文所提模型的有效性。

本文的创新点如下：

(1) 将基于深度学习的特征交互模型引入到信贷风控领域。目前，在信贷违约预测这一领域中，主流的信贷风控模型还是基于传统的机器学习方法。这些传统的机器学习方法，存在各个特征独立，忽略特征之间组合的问题。若通过特征工程，人为地构造交互特征，不仅需要大量的人力成本还不能保证交互特征的有效性。为解决人为构造特征困难的问题，本文改进特征交互模型 DeepFM 并将其应用到信贷违约预测。

(2) 引入传统注意力机制改进低阶特征交互模块。原 DeepFM 模型使用 FM 模型进行低阶交互特征的提取，虽然学习了组合特征但是并未对其重要性加以区分，这会影响模型预测性能。通过引入注意力机制，使模型可以学习不同组合特征的权重，从而改进权重相同的问题。

(3) 引入多头自注意力机制丰富高阶特征提取。原 DeepFM 模型使用

DNN 模块进行高阶交互特征的提取，学习到的交互特征是隐式的、高阶的，这会影晌最终效果。通过引入多头自注意力机制模块，使模型不仅可以学习隐式的高阶特征组合，还可以学习显式的高阶特征组合，使模型可以挖掘出更多的有效信息，从而提高模型的预测能力。

1.4 本文组织结构

本文的整体框架用以下五个部分概括：

第一章首先介绍了本文的研究背景，然后说明了本文的研究意义，之后又介绍了论文的研究内容和创新点，最后介绍了全文的组织结构。

第二章主要介绍相关技术理论的研究现状，具体分为信贷风控模型的研究现状和特征交互模型的研究现状。信贷风控模型的研究现状中又可以分为基于统计学、基于机器学习和基于深度学习的信贷风控模型。特征交互模型又分为基于机器学习和基于深度学习的特征交互模型。

第三章是基于改进 DeepFM 的信贷风控模型的构建。首先介绍了 FM、DNN、Wide & Deep、DeepFM、AFM 和 AutoInt 六个特征交互模型之间的逻辑关系，同时详细介绍了各模型的结构和特点。然后，结合上述模型特点对 DeepFM 模型存在的问题进行了分析，说明了改进 DeepFM 模型的设计背景。接下来针对存在的问题做出相应的改进，提出了改进后的 AMDFM 模型，并对模型的整体结构进行了介绍。最后对模型的核心结构特征交互层进行了详细介绍，阐述了 AFM 模块、多头自注意力机制模块和 DNN 模块三个组成部分的计算过程。

第四章是实证研究，首先分别介绍了实证研究中涉及的实验环境、实验数据的特征工程以及模型评估指标。然后对模型的实验结果进行了分析，包括模型超参数分析、模型性能对比、模型可解释性分析以及模型特征交互层组件的消融实验分析。通过上述实验的对比与分析，从多个方面证明了本文提出模型的有效性。

第五章是本文的总结与展望。回顾全文内容并对研究内容进行总结，对仍然存在的不足进行说明并展望今后的研究方向。

2.研究现状

本章主要介绍相关的研究现状，根据所要介绍模型的特点分为两个部分，分别是信贷风控模型的研究现状和特征交互模型的研究现状。其中，信贷风控模型根据使用的方法又可以分为基于统计学、基于机器学习和基于深度学习的信贷风控模型，特征交互模型根据使用方法可以分为基于机器学习和基于深度学习的特征交互模型。

2.1 信贷风控模型的研究现状

信贷风控本质上是结合借款人的相关信息对借款人的信用进行评估，近年来，国内外关于信贷风控的研究成果不断涌现，对上述研究成果梳理后发现，国内外学者对信贷风控模型研究的发展脉络可以概述为从传统的统计学方法过渡到机器学习再到深度学习，因此本节将从统计学、机器学习和深度学习三个角度分别介绍信贷风控模型的研究现状和发展。

（1）统计学方法

在早期信用风险评估研究中，学者大多基于统计学的方法来构建信用评估模型比如专家分析法、判别分析法、Z-Score 模型和逻辑回归模型等。基于专家分析法的违约风险预测方法是较早运用到金融机构中的风险评估方法，该方法主要通过定性分析的方式，依据专家的经验 and 判断对借款人的还款意愿的能力进行预测。5C 要素分析法^[3]是被广泛采用的专家分析法之一，该方法从借款人品质(Character)、借款人能力(Capacity)、借款人条件(Condition)、借款人抵押(Collateral)和借款人资本(Capital)这五个方面对借款人进行评估，根据借款人的信息对五个方面逐一评分，最终确定其信用等级并判断是否贷款。虽然专家分析法被不断优化和完善，但是该方法还是包含较多的主观因素，十分依赖专家的判断。判别分析法是一种根据确定类别的样本建立判别模型，判断未知类别样本的一种多变量统计分析方法。1936 年，Fisher 提出了线性判别分析法^[4]，该方法的思想是将高维数据投影到一条直线上，使得

同类别样本尽可能接近,不同类样本尽可能远离,以此来进行分类,该方法的提出为风险评估提供了思路。1941 年, Durand^[5]将线性判别分析法应用于信用风险评估问题,并验证了该方法的有效性。1968 年, Altman^[6]为了评估企业经营的风险,提出了一种基于多元判别分析法的模型,这种模型可以对企业的财务风险情况进行预测,该方法就是著名的 Z-Score 模型。逻辑回归模型 (LR),因为其具有模型原理简单、稳健性较好且可解释性比较强的特点,而被引入金融风控领域,并被广泛应用于借款人的信用评估。Wiginton J C^[7]将逻辑回归模型引入到信用评估中,并通过实证研究得出,在预测借款人违约方面,逻辑回归比判别分析法表现更好。1996 年, Huffman 等人^[8]利用逻辑回归模型对传统变量进行分析和筛选,研究表明逻辑回归模型可以对高风险的债券组合进行违约预测。在原逻辑回归模型的基础上,还有一些经过改进和优化的模型被提出。莫茜等人^[9]在逻辑回归模型的基础上引入马尔科夫链模型来进行信用评估,综合考虑借款人的行为变化,结果表明混合后的模型效果优于单一的逻辑回归模型。王小燕等人^[10]提出了一种基于双层变量惩罚的逻辑回归模型并应用到信用卡评估,结果表明与逻辑回归相比该方法具有更高的分类精度。方匡南等人^[11]将逻辑回归和 Lasso 方法相结合,使得模型计算可以更加快捷,同时还能对变量进行筛选,实验结果显示该模型的预测准确率也更高。此外,还有考虑变量间关系网络的基于网状结构的逻辑回归模型^[12]和引入多源数据的逻辑回归模型^[13]被提出并应用到信用评估领域中。

(2) 机器学习方法

随着科技的不断进步,越来越多的数据被收集并应用到信贷领域中,这也导致数据量和特征维度不断增加。传统的统计学方法并不擅长处理大规模数据集,而计算机技术和人工智能的兴起为信用评估提供了新的思路,机器学习方法开始在信用评估领域得到应用。常用的模型包括决策树、支持向量机、随机森林、XGBoost 和 LightGBM 等。Galindo 等人^[14]对比了 CART 决策树、K-NN、Probit 以及神经网络在抵押贷款违约预测方面的表现,结果显示 CART 决策树在四个模型中平均误差率最小,预测效果最好。Li 等人^[15]使用决策树对银行客户进行信用评估,并和逻辑回归模型对比,结果显示决策树的预测效果要比逻辑回归表现更好。Bellotti 等人^[16]将支持向量机用于信用卡的违约评估,实证研究结果显示,支持向量机相比传统的基于统计学的信用

评估方法表现更好，同时还可以用于特征选择，以此来确定违约预测中比较重要的特征。Harris^[17]提出聚类支持向量机算法用于信用评估，与支持向量机相比该算法在保证预测性能的同时还能降低计算成本。上述模型均为单一的机器学习模型，随着研究的不断深入，许多基于集成学习的机器学习算法也开始被应用到信用评估领域 Malekipirbazari 等人^[18]基于 P2P 平台 LendingClub 的借贷数据，使用随机森林进行对借款人的违约情况进行预测，结果显示在识别好的借款人方面，基于随机森林的方法比使用 FICO 信用评分或 LendingClub 平台评估的信用等级表现更好。刘志惠等人^[19]在国内某 P2P 公司借款数据集上对比了 XGBoost 模型和评分卡模型的预测效果，结果显示 XGBoost 模型的预测结果表现更好。Li 等人^[20]基于 LendingClub 数据集将 XGBoost 模型应用于个人信用评估场景，实证研究结果显示，无论是模型性能还是特征选择，XGBoost 模型相比于逻辑回归、决策树、随机森林以及 GBDT 都有更好的表现。马晓君等人^[21]使用 LightGBM 模型对个人信用进行评估，在实证研究中对比了两种数据处理方式的预测效果，结果显示数据处理时保留更多的观测角度可以提升模型的预测效果。Al Daoud^[22]对比了 XGBoost、LightGBM 和 CatBoost 三种梯度提升方法在家庭信用评估方面的表现。实证结果显示，在不同数量的特征和样本上，LightGBM 的计算速度和准确率均优于另外两个模型。

（3）深度学习方法

随着数据维度的进一步增加，数据规模不断增大，基于机器学习方法的评估模型也会受到限制。近年来深度学习理论快速发展，因其具有良好的特征信息挖掘能力和处理大规模数据集能力从而得到众多学者的青睐，越来越多深度学习算法也因此被应用到信用风险评估这一领域中，例如深度神经网络（DNN）、卷积神经网络（CNN）、长短期记忆网络（LSTM）和图卷积神经网络（GCN）等。Sun 等人^[23]构建了一个三层的深度神经网络用于预测巴西某银行的信用卡违约情况，实证研究结果表明，相比于浅层网络，更深的网络结构预测效果更好。Luo 等人^[24]将深度信念网络引入到信用评级中，并与逻辑回归、支持向量机等常用模型进行对比，实验结果表明深度信念网络用于信用评级有更好的表现。刘伟江等人^[25]则提出使用卷积神经网络进行信用评估，首先抽取借款人的特征信息并处理成灰度图，然后利用卷积神经网络对

借款人信用进行评估,与传统信用评估模型相比预测效果提升显著。Kvamme 等人^[26]尝试将卷积神经网络和随机森林模型相组合来预测用户的违约情况,其中的卷积神经网络采用共享卷积核的形式,这种处理方式不仅能够处理重塑转换后的高维数据,还能减少模型参数数量,提升训练效率。Kim 等人^[27]则尝试运用深度卷积神经网络(DenseNet)来进行风险评估,结果显示该模型性能相比于卷积神经网络以及其它机器学习方法表现更好。Roy 等人^[28]提出使用长短期记忆网络来检测信用卡的欺诈情况,分别使用长短期记忆网络和人工神经网络建模,实验结果显示,长短期记忆网络模型的效果明显优于人工神经网络,说明在预测时考虑时间因素可以提升最终效果。Wang 等人^[29]受到自然语言处理中序列模型的启发,将用户的行为看作序列,提出了一种融入注意力机制的长短期记忆网络模型,并以此来对借款人的行为数据进行建模,实验结果表明融入注意力机制的长短期记忆网络模型要比普通的长短期记忆网络模型和双向长短期记忆网络模型预测得更准确。Lee 等人^[30]使用图卷积神经网络来预测客户的违约情况,首先根据用户的特征信息通过 K-means 模型计算贷款客户的样本距离,然后并基于上述计算出的样本距离构造客户网络,最后通过图卷积神经网络模型预测客户的违约情况,此外该模型还能确定不同特征对最终预测结果的贡献。Wang 等人^[31]则融合长短期记忆网络和图卷积神经网络,提出一种时间感知图神经网络(TemGNN)来预测支付宝借贷网络中的用户违约情况。Bastani 等人^[32]参考谷歌公司用于 APP 推荐的 Wide & Deep 模型,并将此类模型应用到风险评估中,该模型分为宽度部分和深度部分,宽度模型为逻辑回归模型,用于提升模型的记忆能力,深度模型为深度神经网络,用于提升模型的泛化能力,将两部分的预测结果相加并输入 Sigmoid 函数得到最终预测结果。

2.2 特征交互模型的研究现状

特征交互模型的兴起与发展主要是为了解决推荐领域中的点击率预估问题,根据模型使用的技术和模型特点可以分为两个发展阶段,分别是使用机器学习方法的特征交互模型和使用深度学习方法特征交互模型。

（1）机器学习方法

逻辑回归(LR)模型是线性模型,模型自身不能进行特征交互,Richardson等人^[33]在使用逻辑回归模型预测广告的点击率时通过引入特征的加权线性组合,实现了对交互特征的学习,结果显示该方法可以提升模型的预测效果。虽然通过人工构造可以使模型学习交互特征,但是这需要在特征工程阶段消耗大量的人力,而且构造出的特征在提升预测效果方面是否有效还不能得到保证。2010年,Chang等人^[34]提出POLY2模型,该模型通过特征之间暴力组合的方式对所有的二阶交互特征进行学习。POLY2模型虽然可以避免人工构造交互特征,但是当数据稀疏时可能会出现大量交互特征缺少训练数据,模型无法收敛的情况,同时,该模型的训练复杂度也比较高。为了解决数据稀疏的问题,同年,Rendle^[35]基于矩阵分解的思想提出了因子分解机(FM)模型,该模型将每个输入特征映射为一个隐向量,通过隐向量之间的内积表示交互特征的权重。与POLY2模型相比,FM的参数更少这极大的简化了模型训练过程,同时,FM还解决了数据的稀疏性问题,使模型可以收敛。Juan等人^[36]在FM的基础上引入了特征域的概念进而提出了FFM模型,相比与FM,特征域的引入使得模型挖掘出更多有用信息,表达能力变得更强,但同时模型复杂度也大大提升。为了降低模型复杂度,有效处理高维特征组合和筛选的问题,He等人^[37]将GBDT和LR相结合,该模型通过GBDT进行特征的自动筛选和组合,进而生成新的离散特征,然后再将生成的特征输入到LR中,最终得到模型预测结果。该模型的提出意味着特征工程不再需要人工组合和筛选,而是通过模型来自动实现,极大地推进了特征工程模型化的趋势。

（2）深度学习方法

近几年,随着深度学习在众多领域取得成功,许多学者也开始尝试将深度学习技术应用到特征交互模型中。Shan等人^[38]提出Deep Crossing模型,该模型的主要结构为残差神经网络,通过残差神经网络可以实现对交互特征的自动提取,同时该模型还可以通过调节网络的层数控制特征交互的深度。Deep Crossing为了解决输入特征稀疏性的问题,将最早用于自然语言处理的嵌入技术引入到模型中,通过嵌入层将高维的稀疏特征转化为低维且密集的嵌入向量,缓解了数据稀疏性的问题。Zhang等人^[39]提出FNN模型,该模型通过深度神经网络提取交互特征,与直接初始化嵌入向量不同,FNN采用预训练

的方式，通过 FM 学习到的隐向量进行嵌入向量的初始化。由于每个特征的嵌入向量都是由 FM 的隐向量得到，因此后续使用深度神经网络训练模型时，可以减少训练时间，使模型可以尽快收敛。Qu 等人^[40]认为传统模型缺少对低阶交互特征的挖掘能力，深度神经网络在提取高阶特征时效果也不能得到保证，因此提出了 PNN 模型，该模型提出了一个名为乘积层的特征交互模块，嵌入向量在输入神经网络之前要先经过乘积层，乘积层可以对输入特征进行外积或内积的操作，这丰富了特征之间的交互方式，提升了模型的预测准确率。为解决深度神经网络对低阶交互特征拟合能力不足的问题，Cheng 等人^[41]提出了 Wide & deep 模型，该模型分为宽度模型部分和深度模型部分，两部分联合训练得到最终结果。其中宽度模型为线性模型，其输入特征为人工提取的特征，该部分主要是为了提高模型对于低阶特征的记忆能力，深度模型为深度神经网络，该部分主要用于提取高阶交互特征从而提升模型的泛化能力。Wide & deep 这种宽度和深度模型相结合共同提取交互特征的架构也为后续对于特征交互的研究提供了思路。Guo 等人^[42]在 Wide & deep 模型的基础上进行改进并提出了 DeepFM 模型，该模型将 Wide & deep 中宽度模型的线性部分替换为了 FM 模型，使宽度模型可以自动学习特征之间的二阶交互，在输入特征方面，宽度模型 FM 和深度模型深度神经网络共享嵌入向量，上述改进使得 DeepFM 可以实现端到端的训练，不再需要人工构造交互特征，简化了输入特征的处理过程。同样是在 Wide & deep 模型基础上对宽度模型进行改进，Wang 等人^[43]提出了 Deep & cross 模型，模型中提出了一种名为 Cross Network 的特征交互网络，该网络可以实现对特征嵌入向量之间进行元素级的显式交互，同时还可以通过调整网络的深度控制特征交互的阶数，相比于 Wide & deep，Deep & cross 的改进丰富了宽度模型对于交互特征的表达效果。为了能够进一步显式地提取交互特征，Lian 等人^[44]在 DeepFM 基础上进行改进并提出了 xDeepFM 模型，该模型提出了一种新的 CIN 网络用于特征交互，CIN 网络可以对特征嵌入向量之间进行向量级的显式交互，相比于特征之间的元素级交互，向量级交互更具有现实意义。受到 FM 模型的启发，He 等人^[45]提出了 NFM 模型，该模型在构建过程中使用了一种交互池化层来实现对二阶交互特征的提取，FM 可以看作不设置隐藏层的 NFM，在设置一个隐藏层的情况下，NFM 的表现要明显优于 FM。Yang 等人^[46]在 DeepFM 的基础上提出了

ONN 模型，该模型将 FM 部分转化为 FFM，以此来实现显式的二阶域向量与隐式向量之间的结合，提升了模型的预测能力。Liu 等人^[47]提出了 AutoFIS 模型，该模型可以自动识别 FM 模型中重要的特征交互，将不重要的特征交互权重设置为 0，简化了计算注意力权重的过程。Wang 等人^[48]在 DCN 的基础上加以改进，提出了 DCN-V2 模型，DCN 模型中 cross 网络的参数为向量，在 DCN-V2 中网络参数变为矩阵，上述改进使得模型更具表现能力。

近年来，注意力机制成功应用到深度学习的多个领域并取得不错的效果，例如计算机视觉、自然语言处理等，特征交互领域也有一些应用。Xiao 等人^[49]对传统的 FM 模型进行了改进，尝试将注意力机制引入到模型中，提出了 AFM 模型，该模型采用注意力池化的方式对二阶交互特征进行加权求和，降低了不重要特征的权重，避免交互特征的冗余。Song 等人^[50]在 2019 年提出了 AutoInt 模型，该模型通过多头自注意力机制显式地进行特征交互，通过调整自注意力层数可以控制特征交互的阶数，还可以通过改变注意力头个数增强模型的表现能力。AutoInt 的提出，为如何显式地学习高阶交互特征提供了一种新的思路，同时，该方法还提高了模型的可解释性。

通过上述对模型的梳理可以看出，在特征交互模型领域，很多新模型或新方法的提出都是在之前模型的基础上进行一些改进，Wide & deep 模型最先提出了宽度和深度模型组合的架构模式，后续提出的许多模型都是在该架构的基础上对宽度部分或深度部分加以改进，以此提高模型对特征的提取以及挖掘出特征之间的潜在联系，进而增强模型的表现能力。

3. 基于改进 DeepFM 的信贷风控模型的构建

本章将对本文提出的 AMDFM 模型的构建过程进行阐述。首先从理论和模型结构角度介绍六个特征交互模型,然后结合上述模型特点分析了 DeepFM 模型的不足,接下来针对现存问题对 DeepFM 模型进行改进从而提出了 AMDFM 模型,最后对模型的核心结构特征交互层进行了详细介绍。

3.1 特征交互模型

本节将对因子分解机(FM)、深度神经网络(DNN)、Wide & Deep、DeepFM、AFM 和 AutoInt 六个模型进行介绍。上述六个特征交互模型的发展过程以及与本文所提出的 AMDFM 模型之间的逻辑关系如图 3-1 所示。下图中, Wide & Deep 是最早被提出的采用宽度模型与深度模型相结合的特征交互模型,其中宽度部分为逻辑回归模型,深度部分为深度神经网络(DNN)。DeepFM 模型则是在 Wide & Deep 模型的基础上进行改进,将 Wide & Deep 中的宽度模型逻辑回归替换为因子分解机(FM),使模型可以自动学习低阶交互特征。AFM 模型则是在因子分解机(FM)的基础上进行改进,通过加入注意力网络,实现对低阶交互特征赋权,从而区分不同交互特征的重要程度。AutoInt 模型则是利用多头自注意力机制来实现特征的高阶交互。下面将详细介绍各特征交互模型的模型结构及对应特点。

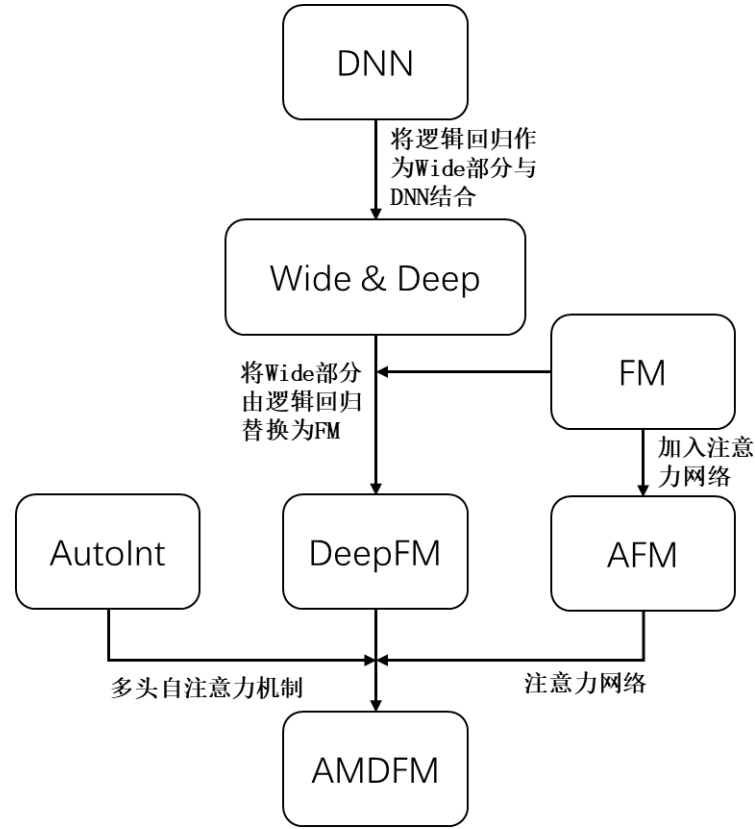


图 3-1 特征交互模型逻辑关系图

(1) 因子分解机 (FM)

基于矩阵分解的思想，Steffen Rendle 在 2010 年提出了因子分解机模型 (FM) [35]，矩阵分解可以看作它的特例，该模型具有易于实现特征交互的特点，并且对稀疏的数据具有很好的学习能力可以有效缓解数据稀疏性。

FM 相当于在线性回归模型 (LR) 的基础上加入了二阶交互项，使用 $x_i x_j$ 来表示特征 x_i 和 x_j 的组合，则模型的表达式可以表示为：

$$\hat{y}(x) = w_0 + \sum_{i=1}^n w_i x_i + \sum_{i=1}^{n-1} \sum_{j=i+1}^n w_{ij} x_i x_j \quad (3-1)$$

式 3-1 中， w_0 和 w_i 为权重参数， w_{ij} 为二阶权重参数， n 为输入特征的个数。此时由于加入了二阶交互项，模型的计算复杂度为 $O(kn^2)$ 。

从矩阵分解的角度，对于每个特征 x_i 通过定义一个隐向量 v_i 进行表示，那么 w_{ij} 就可以表示为 $v_i^t v_j$ ，即 $w_{ij} = \langle v_i, v_j \rangle$ ，其中 $\langle \cdot, \cdot \rangle$ 表示向量的内积。上述对 w_{ij} 的计算方式的变换就是因子分解机模型的核心思想，因此因子分解机模型

的表达式为：

$$\hat{y}(x) = w_0 + \sum_{i=1}^n w_i x_i + \sum_{i=1}^{n-1} \sum_{j=i+1}^n \langle v_i, v_j \rangle x_i x_j \quad (3-2)$$

式 3-2 中，FM 模型的计算复杂度为 $O(kn^2)$ ，但是可以通过对计算公式的化简得到一个计算复杂度为 $O(kn)$ 的表达式，即：

$$\hat{y}(x) = w_0 + \sum_{i=1}^n w_i x_i + \frac{1}{2} \sum_{f=1}^k \left(\left(\sum_{i=1}^n v_{i,f} x_i \right)^2 - \sum_{i=1}^n v_{i,f}^2 x_i^2 \right) \quad (3-3)$$

其中， f 为隐向量的个数。在训练时，只需要提前计算出 $v_{i,f} x_i$ 和 $v_{i,f}^2 x_i^2$ ，之后进行线性复杂度的累加运算就可以得到结果。

最后，当使用 FM 模型处理二分类问题时，还需要将 FM 的输出再输入到一个 sigmoid 函数中，以便对最终结果进行二值映射。

（2）深度神经网络（DNN）

深度神经网络（DNN）又被称为前馈神经网络，它模拟人类大脑的神经网络结构，由许多的神经元相互连接组成。如图 3-2 是神经元的结构示意图。

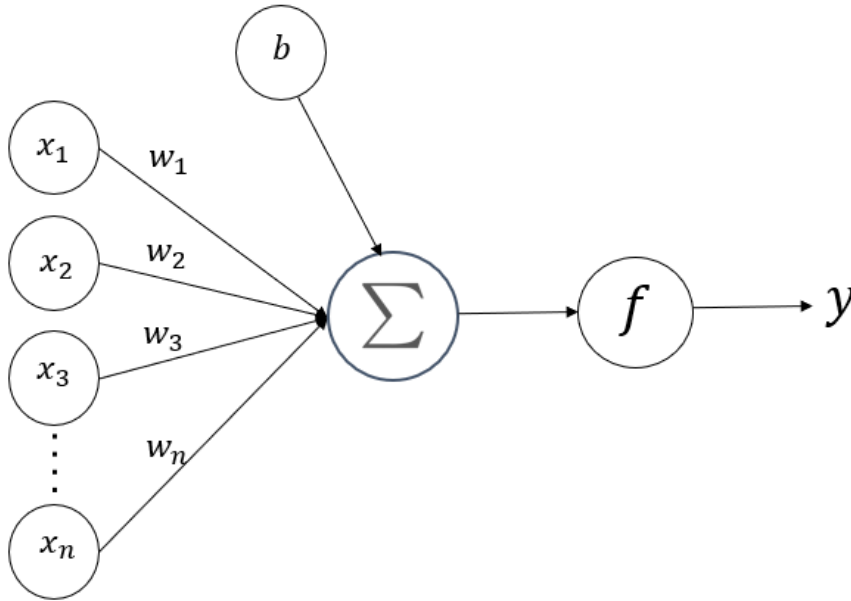


图 3-2 神经元结构

其中， $x_1, x_2, x_3 \dots x_n$ 为神经元的输入信号， $w_1, w_2, w_3 \dots w_n$ 为每个输入信号对应的连接权重，也就是每个神经元突触的强度， b 为偏置项。在网络开始

训练前需对网络参数 w 、 b 进行初始化，随后通过训练得到最佳参数。求和节点表明将每个输入值和对应权重值进行线性相乘后相加，得到的加权和加上偏置项 b 作为激活函数 f 的输入。经过激活函数的非线性映射后，得到最终输出，表达式为：

$$y = f\left(\sum_{i=0}^n w_i x_i\right) \quad (3-4)$$

DNN 主要由输入层、隐藏层和输出层组成，其中每种层都由一定数量的神经元组成，同一层的神经元彼此之间相互独立，一层神经元的输出是下一层神经元的输入，其结构如图 3-3 所示。

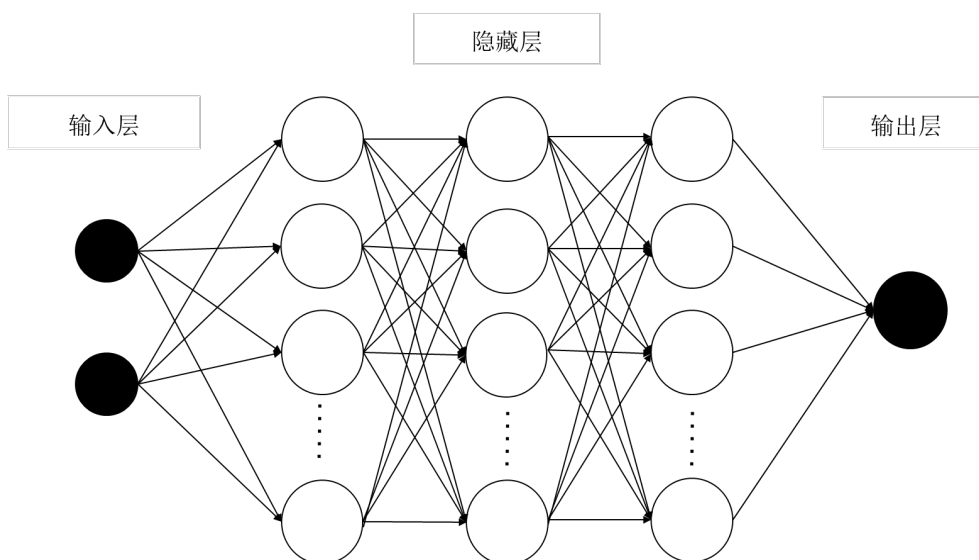


图 3-3 DNN 结构

三种类型的层中，输入层负责接收输入数据，输入数据中，每个元素都有对应的神经元对其进行运算和转化；隐藏层负责对数据进行非线性转换，隐藏层的层数不固定，一般根据实际情况来确定隐藏层数量；输出层负责输出最终结果，输出的节点个数也可以根据实际情况进行调整。

相比于传统的机器学习模型，DNN 不需要在特征工程阶段付出大量时间进行交互特征构造，只需要经过多层神经网络的处理，就可以自动实现特征之间的组合，这是一种端到端的模型结构，不仅能够节省人力成本而且还具有更强大的表达能力。

(3) Wide & Deep

Cheng 等人提出了 Wide & Deep^[41]模型，所谓 Wide 即模型的宽度部分，Deep 即模型的深度部分，Wide & Deep 即两部分的组合，这是该网络框架首次被提出^[51]，下图 3-4 是模型结构图。从图中可以看出，模型的宽度部分采用的是一般的线性模型，主要用于低阶交互的提取，而深度部分采用的则是 DNN，原始特征输入后要先经过嵌入层对稀疏特征进行处理，得到稠密的嵌入向量，然后再将嵌入向量进行拼接并输入到 DNN 中，以此来实现高阶交互特征的提取。最后，将宽度模型和深度模型的输出相加并输入到 sigmoid 函数，从而得到最终预测结果。由于 Wide & Deep 模型的宽度部分为线性模型，因此在输入阶段仍需要人工构建交互特征，还没有办法实现特征的自动交互。

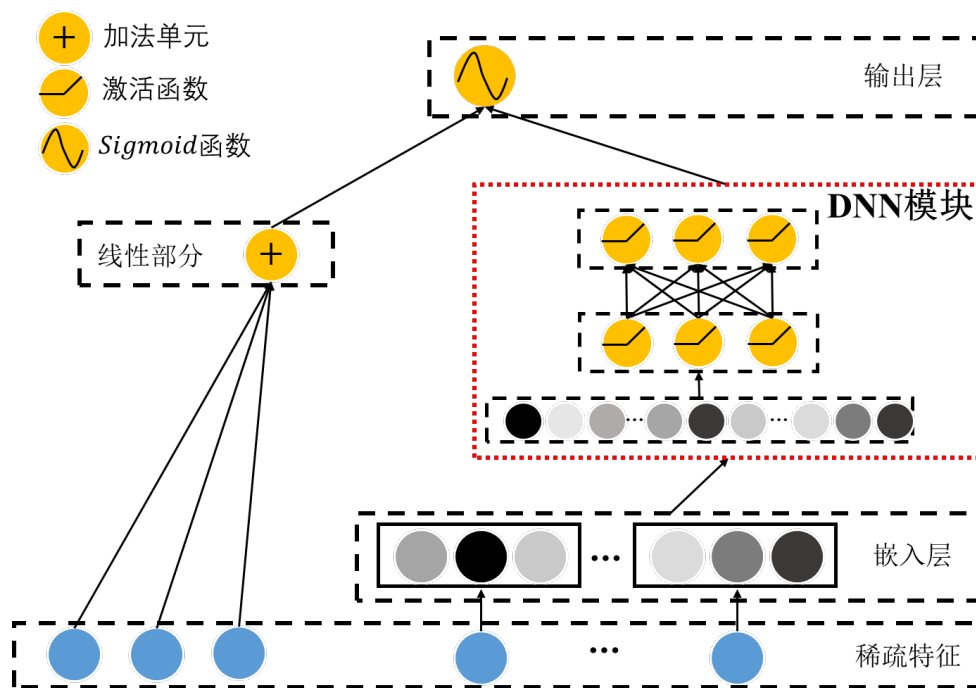


图 3-4 Wide & Deep 模型结构示意图

（4） DeepFM

DeepFM^[42]模型是由 Guo 等人在 Wide & deep 模型的基础上进行改进后提出的。正如前文所说，Wide & Deep 模型虽然能够同时提取低阶和高阶交互特征，但是宽度模型部分的输入仍然需要在特征工程阶段通过人工构造，DeepFM 正是在这一背景下而被提出的，其模型结构如图 3-5 所示。为了解决宽度模型部分无法进行自动提取交互特征的问题，DeepFM 模型将原有的宽度模型部分替换为因子分解机，即 FM 模型。正如前面提到的，FM 是一种专

门为解决大规模稀疏场景下的交互特征学习问题的模型，同时，FM 模型也包含了对线性特征的学习，这与 Wide & Deep 模型中的线性部分完全相同，因而 DeepFM 模型相当于在 Wide & Deep 模型的基础上引入了二阶交互特征。但是，与 Wide & Deep 模型不同的是，DeepFM 中的 DNN 部分和 FM 部分并不是完全独立的，而是共享嵌入层的输入，这种共享使得模型能够更快、更准确地进行训练。

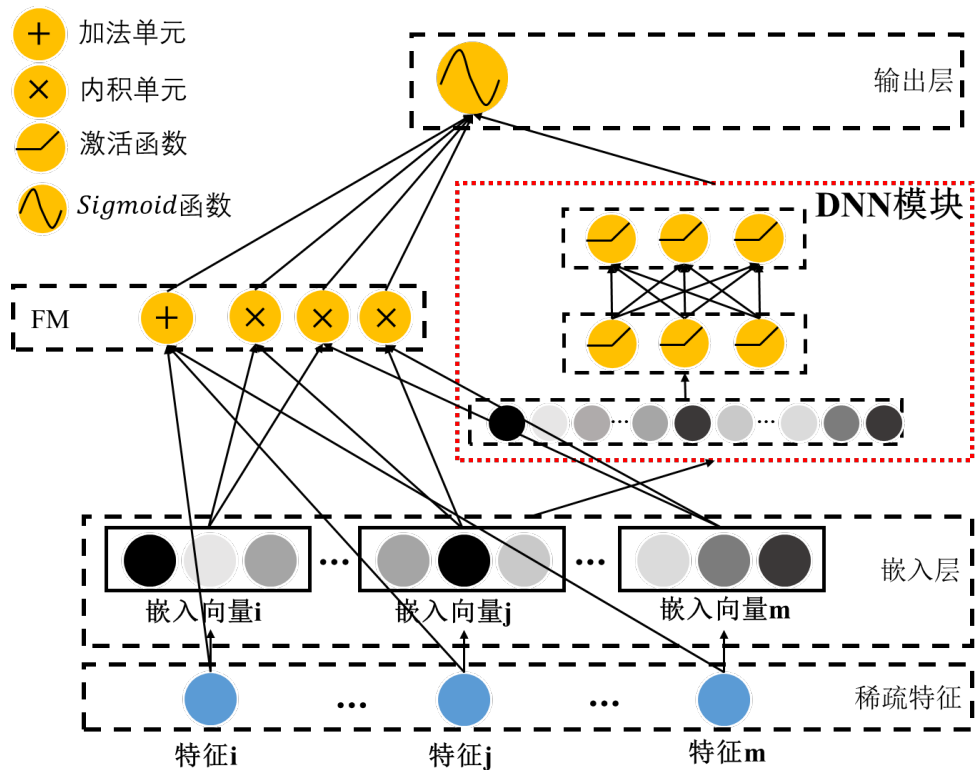


图 3-5 DeepFM 模型结构示意图

从模型结构图可以看出，嵌入向量同时输入到 FM 和 DNN 中，然后 FM 和 DNN 的输出相加后输入到 sigmoid 函数得到最终结果，整个模型是一个端到端的结构。在输出层，模型的最终结果如式 3-5 所示。

$$\hat{y} = \text{sigmoid}(y_{FM} + y_{DNN}) \tag{3-5}$$

综上所述，DeepFM 模型不仅能够同时提取低阶和高阶交互特征，相比于 Wide & Deep，DeepFM 在处理输入特征时不再需要通过复杂的特征工程手动构建交互特征，而是可以直接通过模型实现端到端的训练。

(5) AFM

FM 模型在处理交互特征时使用的是加和池化的方法，没有考虑交互特征之间的差异，相当于认为所有交互特征的重要性都相同，事实上不同的交互特征会不同程度地影响结果的预测，如果权重相等可能会损失大量有效信息，甚至可能会向模型中引入噪声。为了让模型能够在训练的过程中根据不同交互特征的重要程度调整特征权重从而增强模型的表现能力，Xiao 等人对传统的 FM 模型进行了改进，尝试引入注意力机制，提出了 AFM 模型^[49]，该模型通过注意力机制学习二阶交互特征的权重，然后进行加权求和，使模型能够根据交互特征的作用有差异地进行学习，从而自适应地调整交互特征的权重减少无用的交互特征对预测结果的影响。

AFM 模型由输入层、嵌入层、成对交互层、基于注意力机制的池化层和输出层组成。模型的结构示意图如图 3-6 所示。

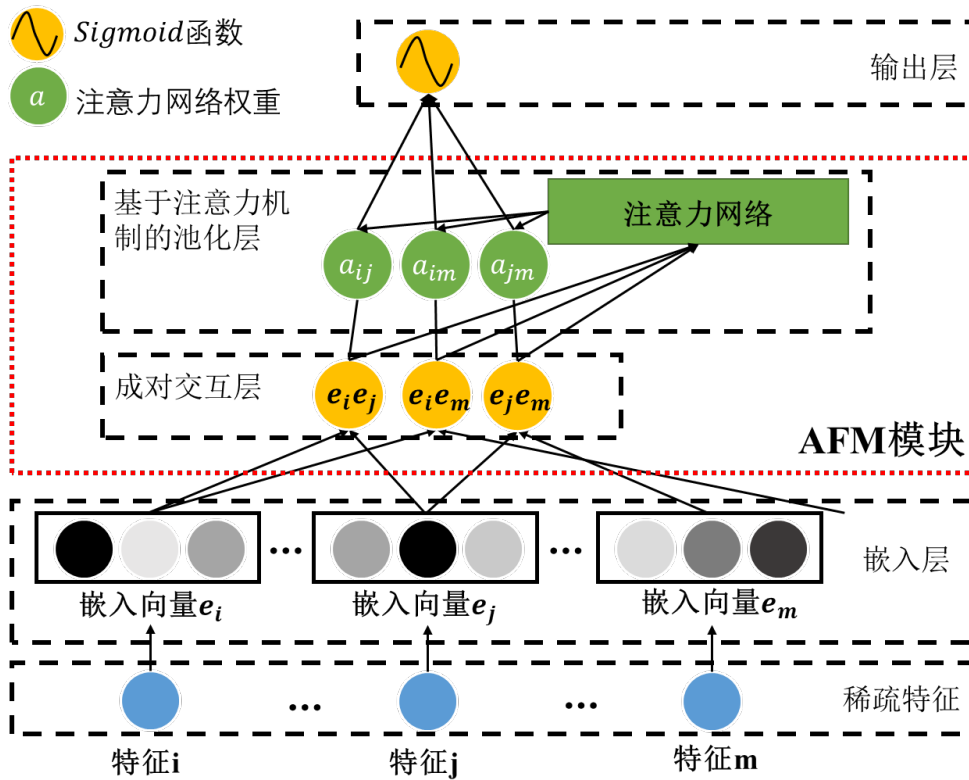


图 3-6 AFM 模型结构示意图

(6) AutoInt

如前文所述，DeepFM 的宽度模型（FM），只能处理到二阶交互特征，而更高阶的交互特征则只能由深度模型部分的深度神经网络来隐式地提取。虽

然神经网络理论上可以提取任意高阶交互特征，但是其内部的具体计算过程仍然不清楚且很难从理论角度解释，这也使得该模型在一定程度上是一种黑箱模型。因此，仅靠神经网络隐式地提取高阶交互特征，模型最终的预测效果不能得到保证，而且神经网络在真实的业务场景中的解释性也不足。

针对上述神经网络在提取高阶交互特征方面所面临的问题，Song 等人在 2019 年提出了 AutoInt^[50]模型，该模型通过多头自注意力机制显式地学习高阶交互，节省了人工构建交互特征的成本，同时相比于 DNN 由于能够清楚地了解特征交互的具体过程，因此模型变得更具有可解释性。

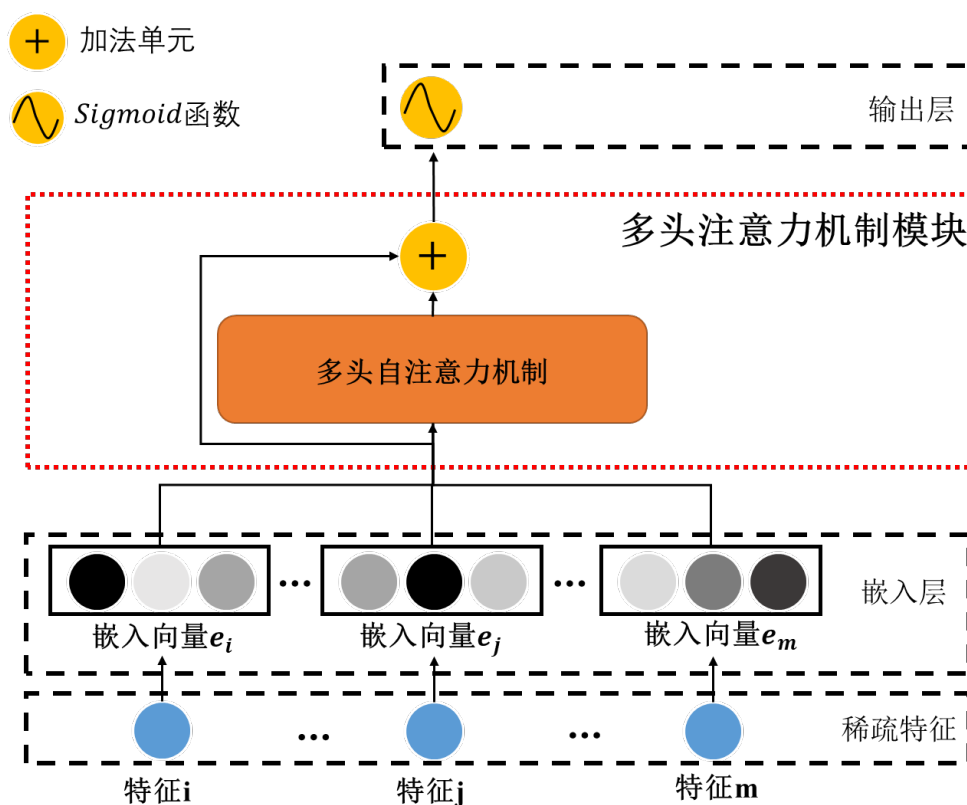


图 3-7 AutoInt 模型结构示意图

该模型的结构示意图如图 3-7 所示，主要由输入层，嵌入层，多头自注意力机制以及输出层组成。AutoInt 将多头自注意力机制引入特征交互模型中，为实现特征的高阶显式交互提供了新的思路。

3.2 改进 DeepFM 模型的设计背景

通过前一节介绍的内容可以得知，DeepFM 模型是基于 Wide & Deep 模型架构的改进和优化的模型，相比 Wide & Deep 模型，其优势在于解决了需要手动构建交互特征的缺点，能够自动从数据中学习隐含的线性特征、二阶特征以及高阶特征，这使得模型可以进行端到端的训练。但是，DeepFM 模型还存在一些不足，而这些不足则限制了模型的表达能力。

首先，在低阶特征交互方面，DeepFM 模型仅使用 FM 模型进行低阶交互特征的提取并未对不同交互特征的权重进行区分，不加区分的特征组合会产生无用的交互特征，甚至可能会向模型中加入噪声，从而降低模型的预测性能。不同的交互特征应赋予不同的权重，以此区分组合特征的重要性。

其次，在高阶特征交互方面，单纯使用 DNN 模型往往并不能达到让人非常满意的效果。这是因为，DNN 模型学习的交互特征具有隐式的、高阶的特点，往往无法学习到低阶特征、以及其他不同类型的显式交互，这就需要在原有神经网络的基础上进行进一步的优化和改进。

综上，DeepFM 模型虽然能够同时提取低阶和高阶的交互特征信息，但是仅使用 FM 和 DNN 进行特征交互还远远不够，还需要探索引入其他类型的特征交互模块，使模型能够更充分地挖掘隐含在数据背后的显式的和隐式的、低阶的和高阶的交互特征，从而提高模型的性能。

3.3 AMDFM 模型整体的提出

根据上述分析，在原 DeepFM 模型的基础上做出以下改进。在低阶特征交互方面，引入注意力机制，通过注意力机制学习不同组合特征的权重，改进交互特征之间权重相同的问题；在高阶特征交互方面，在保留原模型中 DNN 模块的基础上，引入多头自注意力机制模块，使模型不仅可以学习隐式的高阶特征组合，还可以学习显式的高阶特征组合，从而挖掘出更丰富的信息以及增强模型可解释性。本文将上述提出的改进 DeepFM 模型称为 AMDFM 模型，模型的结构示意图如图 3-8 所示。

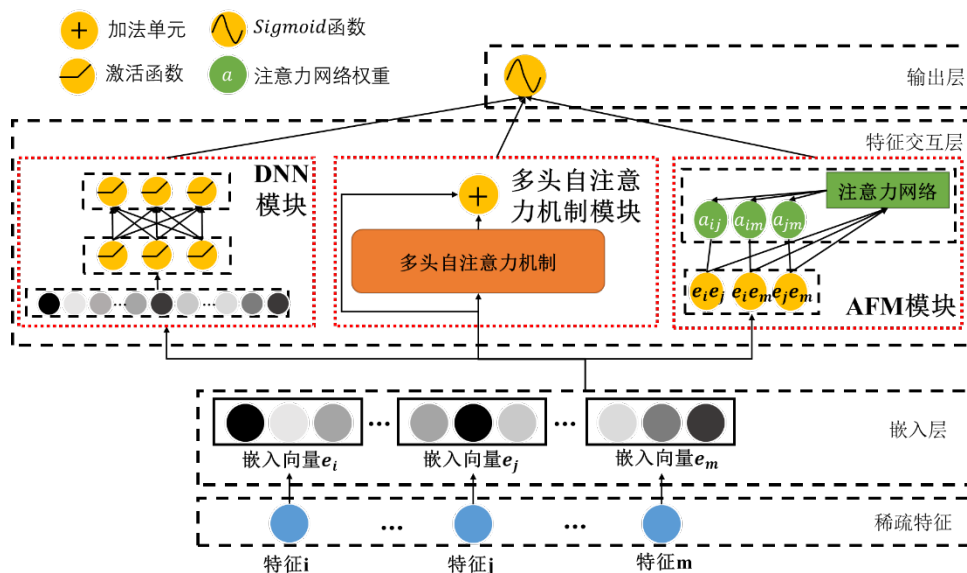


图 3-8 AMDFM 模型结构示意图

从图 3-8 中可以看出，本文提出的 AMDFM 模型主要由输入层、嵌入层、特征交互层和输出层四个部分组成。模型训练过程如上图所示，原始稀疏特征通过输入层进入模型，然后通过嵌入层将稀疏的特征转化为稠密的嵌入向量，之后分别输入到特征交互层中，对交互特征进行学习，其中 AFM 模块负责原始特征和二阶交互特征的学习，多头自注意力机制模块和深度神经网络模块负责高阶交互特征的学习。最后，将三个模块的输出，共同输入到模型的输出层，得到模型最终的预测结果。

模型的输入层主要是为了接收输入特征；嵌入层则是将高维的稀疏特征转换为低维稠密的嵌入向量；特征交互层作为 AMDFM 模型的核心部分，将在下一节详细介绍；模型的输出层为 sigmoid 函数，定义 y_{AFM} 为 AFM 模块的输出， y_{MHSA} 为多头自注意力机制模块的输出， y_{DNN} 为 DNN 模块的输出，则本文提出的 AMDFM 模型的输出是由 y_{AFM} 、 y_{MHSA} 和 y_{DNN} 三部分相加后，再经过 sigmoid 输出最终预测结果，其计算过程如式 3-6 所示：

$$\hat{y} = \text{sigmoid}(y_{AFM} + y_{MHSA} + y_{DNN}) \quad (3-6)$$

3.4 特征交互层

特征交互层是本文提出模型的核心结构，其主要作用就是学习不同阶的

交互特征，具体包括三个特征交互模块，分别是 AFM 模块、多头自注意力机制模块和 DNN 模块，本节将对上述三个模块进行详细介绍。

3.4.1 AFM 模块

AFM 模块主要是对所有特征之间的两两交互进行建模，用以捕捉二阶交互特征表示。与 FM 模型相比，该模块引入了注意力机制来进行特征权重分析，以此来体现不同的交互特征对预测结果的影响程度。

前文已经对 AFM 模型的结构进行了介绍，这里主要介绍下整体模块的计算过程。AFM 模块中，成对交互层相当于 FM 模型，主要计算得出二阶交互特征向量；注意力权重 a_{ij} 则是通过注意力网络计算得出，其具体的计算公式如下：

$$a_{ij}^* = h^T \text{ReLU}(W(v_i \odot v_j)x_i x_j + b) \quad (3-7)$$

$$a_{ij} = \frac{\exp(a_{ij}^*)}{\sum_{(i,j) \in R_x} \exp(a_{ij}^*)} \quad (3-8)$$

其中， $W \in R^{t \times m}$ ， $b \in R^t$ 和 $h \in R^t$ 都是注意力网络中的超参数， t 表示网络中隐藏层的大小，也可以称为注意力因子； m 是网络输出 a_{ij} 向量的维度，等于所有二阶交互特征个数 $f * (f - 1)/2$ ，其中 f 为输入特征个数。

最后将得到的注意力权重 a_{ij} 和二阶交互特征向量 $(v_i \odot v_j)$ 进行对应相乘后累加，再加上原始特征部分，便得到该模块的最终结果 y_{AFM} ，如公式 3-9 所示。

$$y_{AFM} = w_0 + \sum_{i=1}^n w_i x_i + P^T \sum_{i=1}^{n-1} \sum_{j=i+1}^n a_{ij} (v_i \odot v_j) x_i x_j \quad (3-9)$$

3.4.2 多头自注意力机制模块

多头自注意力机制模块主要用来显式地自动学习高阶交互特征，本节将简要地介绍下该模块的计算过程。为了实现自动地学习高阶交互特征，多头自注意力机制将特征的嵌入向量映射到不同的子空间中来提取不同的特征交互模式。

首先, 对于每个特征向量 e_m 通过线性变换将其投影到子空间 h 中, 计算公式如式 3-10 所示, 其中 W_{Query}^h 、 W_{Key}^h 和 W_{Value}^h 是对应 ($Query, Key, Value$) 的线性映射矩阵。

$$E_m^{h:Query} = W_{Query}^h e_m, E_m^{h:Key} = W_{Key}^h e_m, E_m^{h:Value} = W_{Value}^h e_m \quad (3-10)$$

然后, 计算 e_m 与其他特征向量 e_* 在子空间的点积相似度权重并使用 $softmax$ 函数进行归一化, 计算方式如公式 3-11、3-12 所示。

$$\Psi^h(e_m, e_*) = \langle E_m^{h:Query}, E_*^{h:Key} \rangle \quad (3-11)$$

其中 $\langle \cdot \rangle$ 为向量内积运算。

$$a_{m,*}^h = \frac{\exp(\Psi^h(e_m, e_*))}{\sum_{n=1}^M \exp(\Psi^h(e_m, e_n))} \quad (3-12)$$

最后通过加权的方式计算交互特征 $\tilde{e}_m^h$, 计算公式如下式 3-13 所示。

$$\tilde{e}_m^h = \sum_{k=1}^M a_{m,k}^h E_k^{h:Value} \quad (3-13)$$

假设有 H 个子空间, 将每个子空间得到的交互特征 $\tilde{e}_m^h$ 首尾拼接作为特征 e_m 由多头自注意力机制的计算结果如下式 3-14 所示。

$$\tilde{e}_m = \tilde{e}_m^1 \oplus \tilde{e}_m^2 \oplus \dots \oplus \tilde{e}_m^H \quad (3-14)$$

为了保留原始特征相关信息, 还在模块中引入了残差机制, 如下式 3-15 所示。

$$e_m^{Res} = ReLU(\tilde{e}_m + W_{Res} e_m) \quad (3-15)$$

在特征交互网络之后, 输出层将输出的各个特征向量拼接在一起作为输入, 输出最终模型预测结果 y_{MHSA} , 计算公式如公式 3-16 所示。

$$y_{MHSA} = \text{sigmoid}(w^T(e_1^{Res} \oplus e_2^{Res} \oplus \dots \oplus e_M^{Res}) + b) \quad (3-16)$$

3.4.3 DNN 模块

DNN 模块主要用来隐式地提取高阶交互特征, 同时还能提高模型的泛化能力, 本节将简要地介绍下该模块的计算过程。DNN 采用的多层堆叠的方式, 并且在每层之后增加了激活函数, 因此又被称为多层感知机。从功能划分的角度, DNN 可以分为三种不同的层: 输入层、隐藏层、输出层。DNN 的每一层之间是全连接的, 其输出如式 3-17 所示。

$$a^{(l+1)} = f(W^{(l)}a^{(l)} + b^{(l)}) \quad (3-17)$$

其中, l 为网络层数, f 为激活函数, $W^{(l)}$ 为连接权重, $b^{(l)}$ 为偏置。

DNN 模块最终输出如式 3-18 所示。

$$y_{DNN} = w_{DNN}^T a^{(l_f)} + b \quad (3-18)$$

4.实证研究

4.1 实验环境

本文实证研究部分使用的实验环境为 Kaggle 平台的云 Notebook。Kaggle 成立于 2010 年，是全球较为知名的数据科学在线平台，该平台提供举办机器学习竞赛、存储和共享数据集、编写和开源代码等功能。同时 Kaggle 平台的云 Notebook 提供性能较好的 CPU 和内存环境，同时还支持使用 GPU 或 TPU 等硬件加速器，确保在云端运行时不会占用本地资源，为实证研究提供了理想的实验环境。具体软硬件环境配置如表 4-1 所示：

表 4-1 实验环境配置表

环境类别	相关参数
硬件环境	CPU: Kaggle Notebook 云服务器 CPU
	内存: 13GB
	GPU: NVIDIA TESLA T4
	显存: 14.8GB
软件环境	操作系统: Windows 10 企业版
	编程语言: Python 3.8
	机器学习框架: Scikit-learn
	深度学习框架: Pytorch

4.2 实验数据特征工程

4.2.1 数据来源及介绍

本文实证研究部分使用的数据是美国信贷公司 Lending Club 的真实业务数据。Lending Club 是美国规模最大、最有影响的 P2P 网络借贷平台，该公司成立于 2006 年 10 月，于 2014 年 12 月在纽约证券交易所上市。与其他网贷平台相比，Lending Club 起步较早并逐渐发展出一套成熟的业务体系，数据的规范性和平稳性较好。首先，平台会根据借款人填写的相关信息对借款人进

行信用评估。然后，平台会根据第一步的评估结果给出相对应的利率。在公司角度，用户的信用评级不同，则用户的违约成本不同，对应的贷款利率也不相同。

本文使用的数据为 2018 年至 2020 年这三年时间期间 Lending Club 平台上的用户借款交易数据，共 1160066 条记录，142 个特征。Lending Club 取自实际业务场景，数据真实且全面，同时特征数量较多，数据维度较为丰富，包含了详细的借款人相关信息，为实证研究创造了较好的数据条件。

信贷业务发展多年，为了对借款人的信用水平进行科学评估，先后出现诸如“5C”、“5P”等分析原则，这些分析原则从不同的角度考虑借款人信用水平。5C 原则涵盖的 5 个方面分别是，借款人品质（Character）、借款人能力（Capacity）、借款人条件（Condition）、借款人抵押（Collateral）和借款人资本（Capital）。5P 原则涵盖的 5 各方面分别是，借款人情况（People）、资金用途（Purpose）、还款来源（Payment）、借款人展望（Perspective）、债权保证（Protection）。

因本文使用的数据特征较多，可以根据上述评价体系将数据特征分为以下几类，以便对数据中所涉及的特征有一个初步了解：

表 4-2 Lending Club 数据集主要特征类型

特征类型	特征名称	特征含义
贷款相关信息	loan_amnt funded_amnt term int_rate purpose installment	贷款金额 获得金额 贷款周期 贷款利率 贷款目的 每月应还金额
借款人个人能力信息	emp_length home_ownership annual_inc addr_state dti	工作年限 房产状况 年收入 家庭住址 负债率
借款人历史信贷信息	all_util acc_now_delinq open_acc pub_rec num_il_tl tot_coll_amt	总信用额度 借款人目前违约账户数 借款人开立信用额度数量 不良公共记录数 分期付款账户数量 欠款总额
违约类别	loan_status	贷款状态

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：<https://d.book118.com/488063115015006023>