文章编号：2096-1618(2021)03-0311-05

基于改进 DeepFM 的车险索赔预测模型的研究

张 姝， 陶宏才
(西南交通大学计算机与人工智能学院，四川 成都 611756)

摘要：广义线性模型因其简单且输出结果具有可解释性被广泛应用于车险索赔预测领域，但不能识别特征之间交互作用从而限制了模型的表现力。DeepFM 使用因子分解机和深度神经网络分别捕捉低阶和高阶特征交互，在数据稀疏的实际场景取得了显著效果。在因子分解机的基础上引入域相关的权重，针对特征存在互相干扰的问题提出相应缓解策略，并将轻量级的视觉注意力机制作用于深度神经网络进一步提升模型的表现力。实验结果表明，提出的模型相比于基本的 DeepFM 模型取得了更好的风险分割效果。

关 键 词：车险索赔；特征交互；DeepFM；注意力机制
中图分类号：TP303 **文献标志码**：A
doi：10.16836/j.cnki.jcuit.2021.03.012

0 引言

车险是非寿险的重要组成部分，2019 年，非寿险保费中车险业务占比高达 70%[1]，车险的盈利情况在保险公司起着举重若轻的作用。因此，建立准确的车险索赔预测模型是一项重要的任务。

广义线性模型(generalize linear model，GLM)可以较好地对低维数据进行拟合，是车险索赔预测领域的主流模型[2]。但随着“互联网+”和大数据等技术进入保险领域，需要考虑更多的数据维度才能实现更加精准的预测。此外，大多数保单在保险期间不会发生索赔，使得观测数据呈现出稀疏性[3]。所以，GLM 难以在数据稀疏的场景下对变量间的相关关系进行充分的刻画。以回归树[4]、支持向量机(support vector machine，SVM)[5]、神经网络[6]和 Boosting 提升算法[7]为代表的机器学习方法开始引起研究者关注并取得了快速发展。相关研究表明，在数据量大、变量较多、变量间相关关系较强时，神经网络比 GLM、树模型等算法具有更好的预测精度[8]。

DeepFM 结合因子分解机(factorization machine，FM)和深度神经网络(deep neural networks，DNN)，取得了比单独使用神经网络或因子分解机更好的效果。由于 FM 在特征交叉时所有的权重都是一样的，因此一定程度上限制了 FM 的表现力[9]。对于应该互相独立的特征，FM 也无法对此进行建模，造成特征之间的互相干扰。针对上述问题，提出了一种新的车险索赔预测模型。首先，在 FM 的基础上引入域相关的权重，并对特征干扰问题提出相应缓解策略。其次，在 DNN 部分则引入了在不同的深度学习任务中取得显著效果的注意力机制。实验结果表明，提出的模型在预测准确度上得到了一定的提高，取得了更好的风险评估效果。

收稿日期：2020-03-28
基金项目：国家自然科学基金资助项目(61806170)

1 相关技术

1.1 因子分解机

GLM 因其简单和可解释性在保险索赔预测领域被广泛使用。但在车险领域，由于多数保单在保险期间不会发生索赔，因此样本中的大部分观测数据都为零值，使得 GLM 难以对样本中的特征组合进行充分的刻画。以 GLM 中的二阶多项式回归模型为例，其公式为

$$y(x)=w_0+\sum_{i=1}^{n}w_ix_i+\sum_{i=1}^{n}\sum_{j=i+1}^{n}w_{i,j}x_ix_j \tag{1}$$

其中，n 代表特征数量，x_i 代表第 i 个特征的值，w_i，$w_{i,j}$ 是模型参数。特征 x_i 和 x_j 的组合采用 x_ix_j 表示，当 x_i 和 x_j 都非零时，组合特征 x_ix_j 才有意义，而在数据稀疏性普遍存在的实际应用场景中，满足 x_i 和 x_j 都非零样本将会非常少，因此二次项参数 $w_{i,j}$ 的训练非常困难。

2010 年，Steffen Rendle 提出了因子分解机 FM[10-11]。FM 是一种基于 Cholesky 矩阵分解思想的机器学习算法，旨在解决大规模稀疏场景下特征交叉的问题。FM 将参数 $w_{i,j}$ 分解成一个对称矩阵 $\boldsymbol{W}$：

$$W=VV^{\mathrm{T}} \tag{2}$$

V 的第 j 列是第 j 维特征的隐向量，二次项参数 $w_{i,j}$ 可进一步表示成特征隐向量的内积

$$w_{ij}=<V_i,V_j> \tag{3}$$

将二阶多项式回归模型中的二次项参数 $w_{i,j}$ 使用向量内积 V_iV_j 表示即可得到 FM

$$y_{FM}=w_0+\sum_{i=1}^{n}w_ix_i+\sum_{i=1}^{n}\sum_{j=i+1}^{n}(v_i,v_j)x_ix_j \quad (4)$$

FM 对每一个特征 x_i 引入隐向量 $V=(V_{i1},V_{i2},\cdots,V_{ik})$,利用隐向量内积 V_iV_j 对二次项的系数 $w_{i,j}$ 进行评估,在大规模稀疏场景下可以相对准确地估计模型中二次项的参数。例如,特征组合 x_ix_i 和 x_ix_j 的系数分别为<V_l,V_i>和<V_i,V_j>,两个组合特征拥有共同项 V_i,那么所有包含 x_i 的非零组合特征的样本都可以用来学习隐向量 V_i,这大幅降低了因数据稀疏二次项参数预估不合理的影响。

1.2 深度神经网络

神经网络是基于感知机的扩展,而深度神经网络 DNN 可理解为有很多隐藏层的神经网络[12],DNN 模型如图 1 所示。

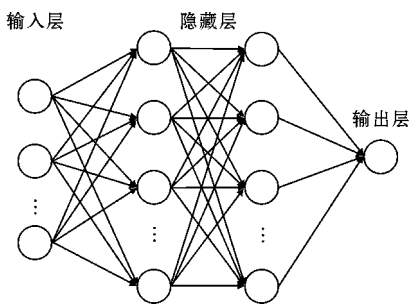

图 1 深度神经网络模型

DNN 中每一个神经元的输出为上一层神经元的线性加权值做非线性映射之后的响应,即对于 $l+1$ 层神经元而言,其输出响应值为

$$a^{l+1}=f(W^la^l+b^l) \quad (5)$$

其中,W^l、a^l 和 b^l 分别表示第 l 层的权重、第 l 层神经元的输出、连接第 l 层和第 $l+1$ 层的偏置值向量,f 为非线性映射函数(即激活函数)。

特征经过 DNN 输入层激活和隐藏层特征提取,最后输出层的响应值为

$$y_{DNN}=W^{|H|+1}a^{|H|+1}+b^{|H|+1} \quad (6)$$

其中,$|H|$ 是隐藏层的数量。

1.3 基于因子分解机的宽深度模型 DeepFM

车险预测中一个关键的挑战是如何有效地建模特征交互,并且低阶和高阶特性交互在模型中都不容忽视。宽深度模型 Wide&Deep[13] 同时考虑低阶和高阶特征交互,使用逻辑回归(logistic regression,LR)拟合低阶特征,DNN 负责对高阶特征的提取,可取得比单独使用 LR 或者 DNN 更好的效果。针对 Wide&Deep 模型 LR 部分需要人工参与的特征工程的缺点,DeepFM[14] 使用 FM 代替 LR,从而自动构造二阶特征叉乘,既考虑了高低阶的特征交互,又省去了额外的特征工程,DeepFM 模型如图 2 所示。

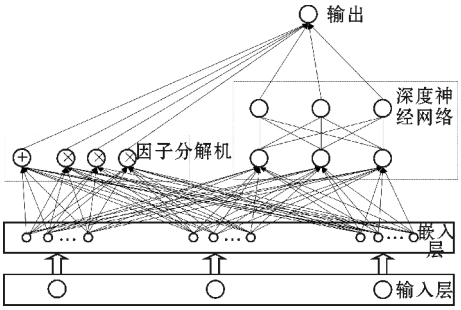

图 2 DeepFM 模型

DeepFM 中 FM 组件和 DNN 组件共享相同的嵌入层输入,分别用于挖掘二阶交叉信息和高阶交叉信息,它们的输出共同输入到 sigmoid 函数进行映射得到模型的预测值为

$$y_{DeepFM}=sigmoid(y_{FM}+y_{DNN}) \quad (7)$$

2 一种改进的 DeepFM 预测模型——SDeepFwFM

2.1 问题提出

DeepFM 模型中的 FM 部分将交叉特征的权重参数转化为隐向量的内积,克服了 GLM 无法在大规模稀疏场景下对特征交叉建模的问题。但 FM 在所有的特征交叉中使用相同的权重在一定程度上损失了模型的表现力,因为并非所有特征交叉都具有相同的价值。特别是对于两个应该互相独立的特征,特征间的交叉学习反而会降低模型的表现力。例如,特征“保单数目”与“车辆颜色”应该是不相关的,但在 FM 模型的学习过程中,每一个特征都不可避免地要和其他特征进行交叉,使得模型在学习特征“保单数目”时不可避免地受到特征“车辆颜色”的影响。对于 DNN 部分,一般来说,神经网络中参数越多则模型拟合特征的能力越强,但因计算能力的限制神经网络不能同时处理这些参数,故会造成信息过载问题。针对以上问题,本文通过改进 DeepFM,提出了一种新的车险索赔预测模型 SDeepFwFM(squeeze Deep field-weighted factorization machine)。

2.2 SDeepFwFM 模型

2.2.1 SDeepFwFM 模型的架构

SDeepFwFM 架构如图 3 所示,与 DeepFM 一样,SDeepFwFM 同样由两个组件构成:左侧的域加权因子分解机(field-weighted factorization machines,FwFM)组件与右侧的 DNN 组件。

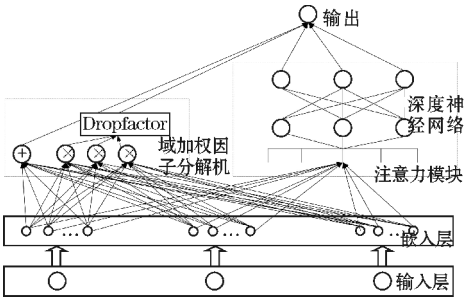

图 3　SDeepFwFM 模型

FwFM 在 FM 基础上引入“域”的概念显式地建模域间的交互强度,在 FwFM 完成特征交叉后借鉴 Dropout 的思想随机丢弃部分二阶交叉项来缓解域内特征干扰问题。DNN 部分,引入轻量级的空间和通道的压缩激励模块(spatial and channel squeeze & excitation, scSE),调整不同特征的权重,突出重要特征、弱化非重要特征来提升模型表现力。

2.2.2　引入域加权的因子分解机 FwFM

FwFM 在 FM 基础上引入“域”的概念,这样具有相同特质的特征可以归类为一个域。例如,用户侧特征可以作为一个域,汽车侧特征可以划分为另一个域。若进行更细的划分,同一个类别特征经过独热编码生成的不同特征可以作为同一个域。例如,“农村”“城市”就同属于“居住区域”这一个域,而“本科”“硕士”“博士”等就属于“学历”这一个域。

FwFM 中对不同域之间的交互强度赋予一个权重 $r_{F(i),F(j)}$,用于建模域 $F(i)$ 和域 $F(j)$ 间的交互强度。将 FM 的二阶交叉项部分乘以权重 $r_{F(i),F(j)}$ 即可得到 FwFM:

$$y_{\mathrm{FwFM}}=w_0+\sum_{i=1}^{n}w_ix_i+\sum_{i=1}^{n}\sum_{j=i+1}^{n}(v_i,v_j)x_ix_jr_{F(i),F(j)} \quad (8)$$

FwFM 模型引入了域相关的权重以显式地捕捉域间的交互强度,一定程度上缓解了不相关特征在学习过程中互相干扰的问题。

2.2.3　引入 Dropfactor 的特征干扰缓解机制

Dropfactor 机制借鉴 Dropout 思路,在 FwFM 模型完成特征交叉后,随机丢弃部分二阶交叉项,在 FwFM 的基础上进一步缓解特征干扰问题。

如图 4 所示,对于两个 K 维的嵌入向量,两两交互可以形成 K 条交互路径,DropFactor 通过随机丢弃部分交互路径来防止特征 V_i 和特征 V_j 之间的相互影响。具体来说,用 FwFM 完成特征交叉后得到的二阶交叉项乘以每条路径被丢弃的概率 $p(i)$,实现随机丢弃部分二阶交叉项,其中路径被丢弃的概率 $p(i)$ 服从伯努利分布 $p(i)\sim Bernoulli(\beta)$。

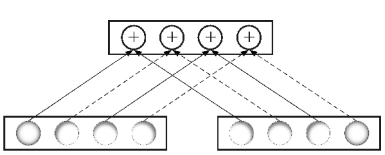

图 4　Dropfactor 策略

FwFM 实现丢弃部分二阶交叉项的公式为

$$y_{\mathrm{FwFM\_d}}=w_0+\sum_{i=1}^{n}w_ix_i+\sum_{i=1}^{n}\sum_{j=i+1}^{n}p(v_i,v_j)x_ix_jr_{F(i),F(j)} \quad (9)$$

为了保证后续网络输入的期望不变,预估时将 FwFM 进行特征交叉后得到的向量,乘以伯努利分布的期望 β。FwFM 使用 Dropfactor 策略后的最终输出为

$$y_{\mathrm{FwFM\_d}}=w_0+\sum_{i=1}^{n}w_ix_i+\sum_{i=1}^{n}\sum_{j=i+1}^{n}\beta(v_i,v_j)x_ix_jr_{F(i),F(j)} \quad (10)$$

2.2.4　引入视觉注意力的 scSE 模型

注意力机制(attention mechanism)借鉴了人类的注意力思维方式,从海量信息中较为精准地筛选出富有价值的信息。在神经网络中,参数越多模型拟合特征的能力越强,但由于计算能力的限制神经网络不能同时处理这些参数。通过在神经网络中引入注意力机制,在计算资源有限的情况下关注更重要的任务,解决了因参数过多导致的信息过载问题。

模型在嵌入层和 DNN 输入层之间加入引入视觉注意力的 scSE 模型[15]。按照注意力域,视觉注意力可分为三类:空间域、通道域和混合域。scSE 是原始特征图分别通过通道压缩-空间激励模块(channel squeeze and spatial excitation, CSSE,文献[15]记为 sSE)和空间压缩-通道激励模块(spatial squeeze and channel excitation, SSCE,文献[15]记为 cSE)后,将两个模块相加得到更为精准的特征图,scSE 模型如图 5 所示。

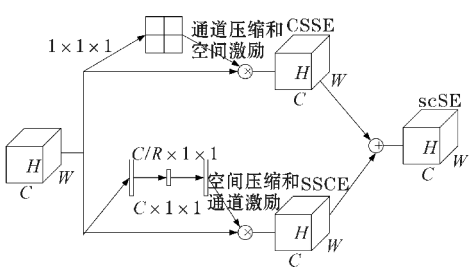

图 5　scSE 模型

图中,模型输入的是一个 $W\times H\times C$ 的特征图,W 和 H 是图像的尺寸,C 是图像的通道数。CSSE 部分通过对特征图使用 1×1×1 卷积,从[W, H, C]变为[W, H,1]的特征,然后使用 sigmoid 函数进行激活得到空间注意力图,然后直接施加到原始特征图中,完成空间的信息校准。

SSCE 模块通过压缩—激发—加权的方式对特征图进行校准。首先,通过压缩操作进行全局平均池化,将空间全部信息压缩成一个 $1\times1\times C$ 的特征向量作为“全局描述符”。激发部分类似于循环神经网络中门的机制,具体实现方法是使用两个全连接层,第一个全连接把 C 个通道数缩小 r 倍来降低计算量,第二个全连接层将通道数激发恢复到 C 个通道。激发操作为每个特征通道生成相应的权重,最后通过一个加权的操作将这些权重与没有经过全局平均池化的原始特征图相乘,在通道维度上标识关键特征。

2.3 SDeepFwFM 模型的算法描述

SDeepFwFM 模型具体算法描述如下：

Input：经过数据预处理后的特征。

Output：Gini 系数和对数损失函数 logloss。

Begin：

(1)输入层：输入层输入特征的独热编码作为嵌入层的输入。

(2)嵌入层：嵌入层本质是一层全连接的神经网络，作用是将特征转成向量。经过嵌入层得到的特征向量 $\boldsymbol{E}=[e_1,e_2,\cdots,e_f]$ 作为 FwFM 部分和 DNN 部分的共同输入。

(3)FwFM 组件：特征向量 $\boldsymbol{E}$ 输入到低阶部分，经过 FwFM 和 Dropfactor 输出如式(10)所示。

(4)注意力机制层：特征向量 $\boldsymbol{E}$ 输入到高阶部分，首先通过注意力层 scSE 对原始特征进行标定。与原始 SSCE 模块使用全局平均池化不同，本文的 SSCE 模块通过全局最大池化把 $\boldsymbol{E}$ 压缩为向量 $\boldsymbol{Z}=[z_1,z_2,\cdots,z_f]$，$\boldsymbol{Z}$ 通过两个全连接层得到 $F_{ex}(Z)=[a_1,a_2,\cdots,a_f]=f_2(W_2(f_1(W_1Z)))$。最后，通过加权操作得到一个包含了特征权重信息的新的特征向量 $U_{SSCE}=[a_1\cdot e_1,a_2\cdot e_2,\cdots,a_f\cdot e_f]$。CSSE 模块先通过卷积实现通道上的压缩操作，得到通道为 1 的特征图 q，再在空间部分上使用 *sigmoid* 函数进行激发得到 U_{CSSE}。SSCE 和 CSSE 相加即为 scSE：$U_{\mathrm{scSE}}=U_{\mathrm{SSCE}}+U_{\mathrm{CSSE}}$。

(5)深度神经网络层：U_{scSE} 作为 DNN 的输入，输出为 y_{DNN}。

(6)输出层：$y_{\mathrm{FwFM\_d}}$ 和 y_{DNN} 相加并通过 *sigmoid* 函数映射得到最终的输出，$y_{\mathrm{SDeepFwFM}}=\mathrm{sigmoid}(y_{\mathrm{FwFM\_d}}+y_{\mathrm{DNN}})$。

End

3 模型实验及结果分析

3.1 实验数据集

实验数据来源于 Kaggle 大赛提供的开源数据集，其中训练集样本 595212 条，验证集样本 892816 条。每个样本对应 1 个表示车险索赔状况的标签 target，0 表示未发起索赔，1 表示发起索赔。

3.2 实验环境及评价指标

实验环境如表 1 所示。模型评估方面，选择 Gini 系数和对数损失函数(LogLoss)作为评价指标。Gini 系数是风险评估领域常用的评价指标，Gini 系数越大，模型对因变量的预测能力越强，风险分割效果越好[16]。LogLoss 反映了样本的平均偏差，是分类任务中常用的评价指标。

表 1 实验环境及配置

实验环境	环境配置
操作系统	Windows10
CPU	Intel Core i7-9750H 2.60GHz
内存	8.00G
编程语言	Python3.7.6
框架	Tensorflow1.14.0

3.3 实验结果

采用 kaggle 竞赛提供的开源数据集，选用 DNN、DeepFM 与本文模型进行对比实验。样本数据共 140 多万条，对每个模型进行了三折交叉实验。采用三层神经网络结构，批处理大小 batch size 为 1024，epoch 迭代次数设置为 50，学习率0.001，优化器选用 adam，嵌入层嵌入维度为 8，实验结果如表 2 所示，实验结果表明 SDeepFwFM 模型比 DeepFM 模型取得了更好的风险评估效果。

表 2 模型对比评价指标结果

模型	Gini 系数	Logloss
DNN	0.2706	0.1526
DeepFM	0.2726	0.1524
SDeepFwFM(本文模型)	0.2749	0.1523

DeepFM 和 SDeepFwFM 模型 Gini 系数值如图 6、图 7 所示。

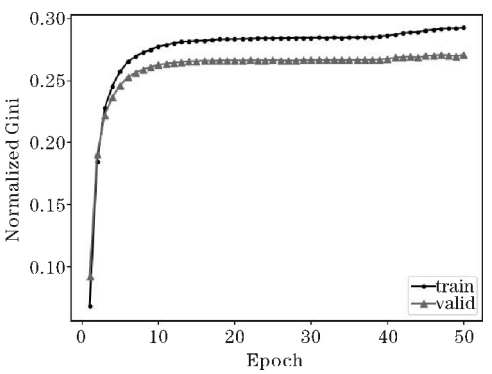

图 6 DeepFM 模型训练集和验证集 Gini 系数图

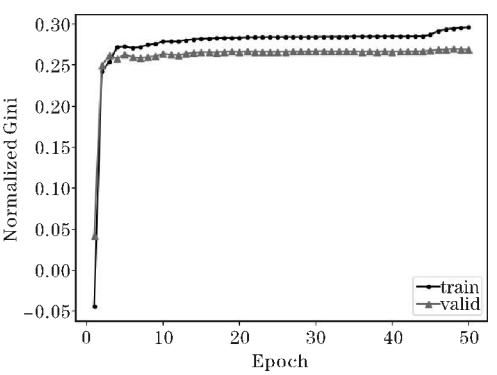

图 7 SDeepFwFM 模型训练集和验证集 Gini 系数图

4　结束语

随着大数据等新兴技术进入车险领域,数据的高维、稀疏性以及特征间较强的交互作用对研究者提出了新的挑战。FM 基于矩阵分解的思想克服了传统的 GLM 模型无法刻画特征间相关关系的缺点,DNN 能够很好地拟合高阶特征交叉,DeepFM 模型将 FM 与 DNN 二者相结合,在数据稀疏的实际场景下取得了显著的效果。

在 DeepFM 的基础上提出新的 SDeepFwFM 模型,FM 部分引入域相关的权重丰富了特征交叉并应用特征干扰缓解策略,DNN 部分添加注意力机制提高模型的拟合特征能力。实验结果表明,本文提出的模型在 DeepFM 的基础上取得了更好的风险分割效果。

参考文献:

[1] 曾宇哲,吴嫒博,郑宏远. 基于机器学习的车险索赔频率预测[J]. 统计与信息论坛,2019,34(5):69-78.

[2] 吴育文. 广义线性模型在车险精算定价中的实证研究[J]. 内燃机与配件,2018,267(15):190-193.

[3] 孟生旺,李天博. 基于机器学习算法的车险索赔概率与累积赔款预测[J]. 保险研究,2017(10):42-53.

[4] 张连增. 回归树方法在车险索赔频率预测建模中的应用[J]. 保险研究,2018(1):101-111.

[5] 薛智雯. 基于 ARIMA-SVM 的车险索赔次数预测[D]. 成都:西南财经大学,2018.

[6] 孟生旺. 神经网络模型与车险索赔频率预测[J]. 统计研究,2012(3):22-26.

[7] 张连增. 提升算法对传统车险索赔频率建模模型的改进——基于我国五省交强险保单数据[J]. 保险研究,2019(7):67-78.

[8] Lee S, Antonio K. Why High Dimensional Modeling in Actuarial Science[C]. Proceedings of the IACA Colloquia, 2015:75-79.

[9] Xiao J, Ye H, He X, et al. Attentional factorization machines: Learning the weight of feature interactions via attention networks[C]. Proceedings of the 26th International Join Conference on Artificial Intelligence, 2017:3119-3125.

[10] Rendle S. Factorization machines[C]. Proceedings of the 2010 IEEE International Conference on Data Mining, 2010:995-1000.

[11] Rendle S. Factorization Machines with libFM[J]. ACM Transactions on Intelligent Systems and Technology, 2012, 3(3):1-22.

[12] 孙志军,薛磊,许阳明. 深度学习研究综述[J]. 计算机应用研究,2012,29(8):6-10.

[13] Cheng H T, Kocl, Harmsen J, et al. Wide & deep learning for recommender systems[C]. Proceedings of the 1st workshop on deep learning for recommender systems, 2016:7-10.

[14] Guo H, Tang R, Ye Y, et al. DeepFM: A Factorization-Machine based Neural Network for CTR Prediction[C]. Twenty-Sixth International Joint Conference on Artificial Intelligence, 2017:1-8.

[15] Roy G, Navab N, Wachinger C. Concurrent Spatial and Channel Squeeze & Excitation in Fully Convolutional Networks[C]. International conference on medical image computing and computer-assisted intervention, 2018:421-429.

[16] 黄秋彧. 个人信用风险评分的指标选择研究[J]. 新疆财经大学学报,2015(3):5-15.

Research on Prediction Model of Auto Insurance Claim based on Improved DeepFM

ZHANG Shu,　TAO Hongcai

(School of Computing & Artificial Intelligence, Southwest Jiaotong University, Chengdu 611756, China)

Abstract: Generalized linear model is widely used in the field of auto insurance claim prediction because of simplicity and interpretability. However, its expressiveness is limited because it can't recognize the interaction between features. DeepFM uses Factorization Machine and Deep Neural Network to capture the interaction of low-order and high-order features respectively, and achieves remarkable results in the real scene with sparse data. This paper introduces the weight of domain correlation based on Factorization Machine, and proposes mitigation strategies to ease the problem of mutual interference between features. The lightweight visual attention mechanism is also applied to the Deep Neural Network to enhance the accuracy of the model. Experimental results show that the proposed model achieves better risk segmentation effect than the basic DeepFM model.

Keywords: auto insurance claim; feature interaction; DeepFM; attention mechanism