

基于集成学习的烟草异常数据挖掘研究与应用

李天举¹, 谢志峰¹, 张侃弘², 陶亦筠³, 范杰², 汤臻³

(1. 上海大学, 上海 200072;
2. 上海烟草集团有限责任公司, 上海 200082;
3. 上海市烟草专卖局, 上海 200120)

摘要: 为了推动上海市烟草专卖市场监管方式转型, 实现市场监管水平的有效提升, 通过引入异常数据挖掘方法, 从而强化市场异动预测和分析。结合目前机器学习前沿理论的研究, 提出了基于多模型 Stacking 集成学习的烟草异常数据挖掘模型, 运用 Stacking 集成学习的方式, 充分发挥各个算法模型的优势。数据集使用的是 2016 年 1 月到 2019 年 4 月的烟草专卖数据, 通过数据预处理等方式将数据指标化, 并使用数据增强等手段一定程度上缓解了数据不平衡的问题。使用该数据对模型进行了验证分析, 其结果很好地证明了 Stacking 模型中单个机器学习算法的学习能力越强, 关联程度越低, 集成后的模型预测结果越好。最后通过实证稽查环节, 充分验证了模型的有效性, 经过全市实证后, 市场检查对零售户的问题查实率能从现有的 5% 左右提升至 15% 以上。

关键词: 异常数据挖掘; 集成学习; 数据预处理; 数据增强; Stacking 模型

中图分类号: TP399

文献标识码: A

文章编号: 1673-629X(2020)11-0128-08

doi:10.3969/j.issn.1673-629X.2020.11.024

Study and Application of Tobacco Anomaly Data Mining Based on Ensemble Learning

LI Tian-ju¹, XIE Zhi-feng¹, ZHANG Kan-hong², TAO Yi-jun³, FAN Jie², TANG Zhen³

(1. Shanghai University, Shanghai 200072, China;
2. Shanghai Tobacco Group Co., Ltd., Shanghai 200082, China;
3. Shanghai Tobacco Monopoly Administration, Shanghai 200120, China)

Abstract: In order to promote the transformation of the Shanghai tobacco monopoly market supervision method and achieve an effective improvement in the level of market supervision, the introduction of abnormal data mining methods has strengthened the prediction and analysis of market movements. Combined with the current research on cutting-edge theories of machine learning, a tobacco anomaly data mining model based on multi-model Stacking ensemble learning is proposed, and the advantages of each algorithm model are brought into full play by using Stacking ensemble learning. The data set uses tobacco monopoly data from January 2016 to April 2019. The data is indexed through data preprocessing and other methods, and data enhancement is used to alleviate the problem of data imbalance to some extent. The model is verified and analyzed by these data. The results well prove that the stronger the learning ability of a single machine learning algorithm in the Stacking model, the lower the degree of association, and the better the prediction result of the integrated model. Finally, the effectiveness of the model is fully verified through the empirical inspection link. After the city's empirical verification, the market inspection of the retailer's problem verification rate can be increased from the existing 5% to more than 15%.

Key words: abnormal data mining; ensemble learning; data preprocessing; data augmentation; Stacking model

0 引言

随着数字化信息时代的到来, 烟草行业数据量正在以惊人的速度快速增长, 这种数字化趋势为机器学习与数据挖掘技术在其生产、物流、监管等各方面的应

用创造了新机遇^[1-3]。数据挖掘技术已经逐渐地应用于各行各业, 对异常数据的挖掘也开始得到人们更多的重视, 所谓异常指的是在海量数据中存在着与一般数据形式相差较大或者与正常行为相左的数据对象,

收稿日期: 2020-01-10

修回日期: 2020-05-12

基金项目: 国家自然科学基金(61303093); 上海市自然科学基金(19ZR1419100)

作者简介: 李天举(1994-), 男, 硕士研究生, 研究方向为机器学习、大数据; 通讯作者: 陶亦筠(1967-), 女, 高级经济师, 研究方向为烟草市场监管研究。

一般的数据挖掘过程常常将这些数据当作噪声进行清除处理,但大多时候它们可能包含了解决现实问题中极其重要的信息。异常数据挖掘技术已在模式识别、信用欺诈、企业监管等领域得到广泛应用。比如在金融行业的征信系统中,异常数据往往代表了用户存在违约、造假等不良行为;在电网系统中,异常数据通常警示设备故障问题或者用户的异常用电的行为;在城市轨道安防系统中,异常数据意味着行人或车辆存在违章行为。在这样的背景下,面向烟草行业的异常数据挖掘技术有望从海量的烟草数据中,提取挖掘出零售户在卷烟经营中是否存在涉烟违法的行为。数据挖掘技术的应用将有效推进整个烟草行业向信息化、智能化方向发展。

基于数据挖掘的市场异常预警预测研究,能够进一步加强烟草零售市场监管力度,有效限制零售户的涉烟违法行为,合理分配稽查工作中的人员调度,有效净化卷烟市场经营环境。在烟草专卖市场监管方面,异常数据挖掘的任务就是在专卖监管数据中发现那些有违规经营迹象的数据对象,并找到隐藏在这些对象背后的各类违规经营情况。通过深入挖掘分析现有的烟草专卖信息数据,能够有效结合现有市场监管模式,加快烟草专卖管理方式的信息化转变,加强对重点涉烟违法行为的治理,提升市场监管的精准性。

目前将前沿的机器学习与数据挖掘技术应用于烟草专卖市场监管方面的研究稍显不足,但在其他领域的相关研究为笔者提供了宝贵的经验。文献[4]将机器学习技术运用于发布虚假财务报表(FFS)公司的异常行为检测中,通过使用优化的 Stacking 多模型融合方法将典型的机器学习算法组合在一起,取得了比任何单一算法和经过检验的简单集成方法更好的检测性能。文献[5]利用 XGBoost 机器学习算法,能够对云计算中 SDN 控制器易受到分布式拒绝服务(DDoS)的异常攻击行为进行快速的检测。文献[6]通过使用基于功能树分类器和三种当前比较先进的机器学习集成框架 Bagging、AdaBoost 和 MultiBoost,提出并验证了一种能够提高滑坡异常和敏感性模型预测性能的集成方法。文献[7]将前沿的机器学习 LightGBM 算法应用于广告转化率预估中,通过 LightGBM 模型提取广告日志中的高阶组合特征,并结合了区域因子分解机 FFM 模型对稀疏数据进行相应处理,有效提高了广告转化率预估模型的有效性和泛化能力。文献[8]提出的深度网络 xDeepFM 算法,能够有效地自动学习数据的特征交互。

该文基于上海市卷烟经营零售户从 2016 年 1 月到 2019 年 4 月的烟草专卖相关数据,提出了基于多模型 Stacking 集成学习的烟草异常数据挖掘模型,旨在

利用前沿的机器学习算法 XGBoost、LightGBM 等,以及深度学习网络 xDeepFM 算法对该数据进行建模预测和分析,最终推动烟草专卖市场监管方式的转型,进而促进全市烟草市场监管水平的大幅提升。

1 数据预处理

1.1 数据来源

选取了上海市 4 万多家零售户从 2016 年 1 月—2019 年 4 月的烟草专卖相关数据,基础数据主要包括:经营户静态数据、客户历史数据、订货数据、卷烟主数据、市场检查数据、投诉举报数据、案件数据等。

1.2 数据预处理(构建数据指标)

影响数据分析与挖掘的第一要素是数据的预处理工作,而数据挖掘技术的合理运用是异常数据检测能否正确运行的核心环节。在对数据进行预处理之后,必须结合有效的分析手段,才能找出数据的规律,从而挖掘出异常经营行为。通过对烟草市场监管数据的深入分析,发现大部分的数据属于结构化数据,其中主要包含了连续和离散两种形式的变量类型,这两种类型数据相对应的处理方式明显不同,因此,如何快速有效地实现复杂条件下结构化数据的分析与挖掘尤为重要。针对烟草行业中的海量、多维、动态数据,分析烟草结构化数据的特点,从营销、物流、市场监管、案件等多个维度进行分析,梳理形成静态特征指标与动态特征指标。部分特征分类如表 1 所示。

表 1 部分特征分类

特征分类		特征指标
静态特征	人	法人籍贯、实际经营人
	证	一人多证、实际经营人与持证人不符
	店	经营业态、经营性质、连锁户、档位
	环	位置流量、商圈、所属乡镇街道
动态特征	盈	场地权属、经营面积、销售规模
	营销数据	同比、环比、满足率、消费需求
	违规记录	举报次数、整改立案次数、显性惯犯

在数据预处理阶段,需要对类别数据进行编码,比如:订货方式包括 POS 订货、电话订货、电子商务、手工订货、网上配货等,需要将其转为数值型数据进行处理。对数据的编码往往会影响到模型训练的速度和预测的结果,所以如何合理选择数据的编码方式十分重要。常见的编码方式有独热编码(one-hot encoding)、标签编码(label encoding)和实体嵌入(embedding)。

(1) one-hot 编码,其基本思想是使用位寄存器对类别数据的 N 种类别状态分别编码,每个类别状态占用其中的一位,且每种状态只有一个位置是 1,其他状态位置都为 0。例如,“POS 订货”编码后的形式为 [0

0 0 0 1],“电话订货”编码后的形式为[0 0 0 1 0],“电子商务”编码后的形式为[0 0 1 0 0],“手工订货”编码后的形式为[0 1 0 0 0],“网上配货”编码后的形式为[1 0 0 0 0]。

(2) 标签编码: 给每种类别分配整数, 例如“POS 订货”为 1,“电话订货”为 2,“电子商务”为 3,“手工订货”为 4,“网上配货”为 5。由于连续的数字代表着数字之间的先后顺序, 要尽量避免将其使用在线性模型中, 而基于树的算法模型则不受这种数值顺序的限制。

(3) 采用实体嵌入方式可以将类别数据用向量来表示, 生成高维数据在高维空间体现它们的相互关联。一般多用于深度神经网络算法模型中。

通过观察样本发现, 大多数类别数据在 5 个类别以下, 所以选择使用 one-hot 编码对类别数据进行编码, 一方面防止标签编码带来的赋值顺序问题, 另一方面又可以同时适用于机器学习算法和深度神经网络算法。最后, 由于原始数据中还存在一些比较脏、乱、差的数据, 还需要对其进行大量清洗, 比如经营面积数据存在大量不合理数值, 而经营面积代码则是以类别 A、B、C、D 来表示, 则提取特征时就去掉经营面积数值型数据, 转用类别型数据代替。大户类别数据中只包含空值和其他大户类别, 那么这一特征数据全是无用信息, 则无需进入模型。零售户的档位信息存在缺失值, 处理方式是按当前时间点往前最近的一次档位进行填补。通过数据预处理和特征工程提取之后, 最终进入模型的一共有 244 个特征(指标)。

2 相关算法

2.1 XGBoost 算法

XGBoost(extreme gradient boosting, 极端梯度提升算法^[9]), 是一种基于 CART 树的 boosting 算法, 高效地实现了 GBDT 算法, 并进行了算法和工程上的许多改进。

XGBoost 模型的目标函数主要包含两个部分:

$$L = \sum_{i=1}^n l(\hat{y}_i, y_i) + \sum_{k=1}^K \Omega(f_k) \quad (1)$$

$$\text{where: } \Omega(f_k) = \gamma T + \frac{1}{2} \lambda \|w\|^2$$

式中, 第一部分是模型的训练误差, 即模型的预测结果 \hat{y}_i 与样本真实 y_i 的差值; 第二部分是正则项, 用于控制模型的复杂度, 其中 γ 和 λ 是惩罚系数, T 和 w 分别代表叶子节点的个数和分数。

XGBoost 模型每次训练一棵新的树都要拟合上一次结果的残差, 每次增加的函数的增量要使新一轮的残差尽可能减小, 在进行到第 t 次时, 模型的目标函数

可以写为:

$$L^{(t)} = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t-1)} + f_i(x_i)) + \Omega(f_i) \quad (2)$$

模型训练的最终目标是要找到一个能够最小化目标函数的 $f_i(x_i)$, 对式(2) 采用其在 $x=0$ 处的泰勒二阶展开式来近似, 近似的目标函数为:

$$L^{(t)} = \sum_{i=1}^n [l(y_i, \hat{y}_i^{(t-1)}) + g_i f_i(x_i) + \frac{1}{2} h_i f_i^2(x_i)] + \Omega(f_i) \quad (3)$$

其中, $g_i = \partial_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$, $h_i = \partial_{\hat{y}_i^{(t-1)}}^2 l(y_i, \hat{y}_i^{(t-1)})$ 。

去掉不影响目标函数最终优化的项, 可简化为:

$$\tilde{L}^{(t)} = \sum_{i=1}^n [g_i f_i(x_i) + \frac{1}{2} h_i f_i^2(x_i)] + \Omega(f_i) \quad (4)$$

2.2 LightGBM 算法

LightGBM 算法主要使用了基于梯度的单边采样和互斥特征捆绑这两种方法来弥补传统 Boosting 在处理大数据样本时的计算损耗问题^[10]。

模型在训练时首先采用基于梯度的单边采样(简称 GOSS), 计算梯度时不再是扫描全部的样本点, 而是保留梯度比较大的一小部分样本数据, 对梯度小的大多数样本进行随机采样; 而互斥特征捆绑(简称 EFB) 主要依据高维数据的稀疏性, 主要特点是存在很多特征不会同时取值为非零值, 称具有这样的性质的特征为互斥特征, 将这些特征组合在一起可以达到降低特征维度的目的, 使得确定切分点的计算损耗减少, 同时对互斥特征的处理也在一定程度上降低了模型过拟合的风险。

2.3 xDeepFM 算法

对于预测性的模型来说, 如何让模型自动地去学习特征之间的交叉特性对数据挖掘系统是特别必要的。所谓特征之间的交叉特性也称之为交叉特征^[11], 是指两个及两个以上的特征进行组合形成一个新的特征。深度神经网络为解决这一问题提供了突破口, 比如基于因子分解机的 FNN、PNN 和 DeepFM 等深度神经网络算法^[12-14], 对特征之间的高阶交互特性的学习使用了多层的全连接网络, 但是这些网络的缺点是模型学习出的是隐式的交叉特征, 使得其具体形式是未知的和不可控的。为了挖掘不同交叉特征之间的潜在联系, 该文引入 xDeepFM(极深因子分解机)深度神经网络模型^[8], 来让模型自动地去学习特征之间的交叉特性。其基本结构如图 1 所示。

xDeepFM 算法首先把数据集的原始特征中每个 one-hot 编码后的特征组成一个 field, 用来克服数据的稀疏性; 然后进行 embedding 转换使特征表现为向量级; 接着将数据送入压缩交互网络 CIN 模型中, 使得模型以显示的方式自动学习高阶的交互特征, CIN

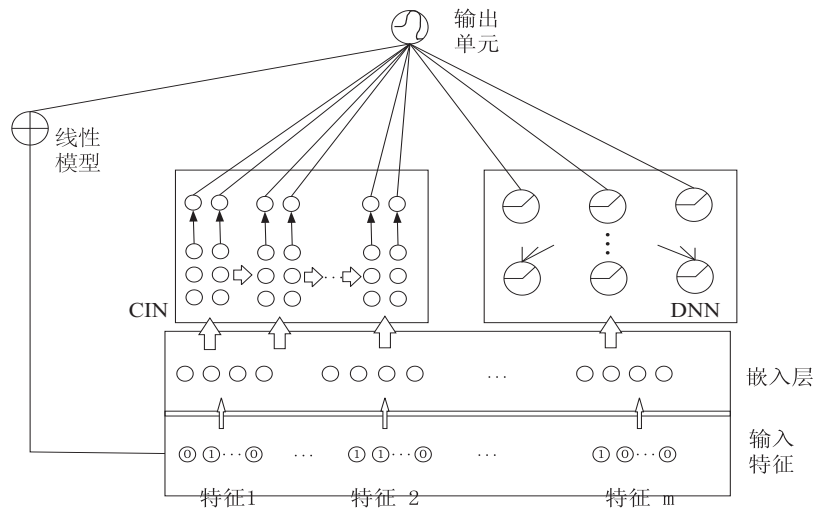


图 1 xDeepFM 神经网络结构

每层的神经元都是由原始特征向量和它前面的隐层计算而来,即:

$$X_{h,*}^k = \sum_{i=1}^{H_{k-1}} \sum_{j=1}^m W_{ij}^{k,h} (X_{i,*}^{k-1} \circ X_{j,*}^o) \quad (5)$$

其中, X^o 为数据的原始特征, X^k 为 CIN 神经网络中的隐层,点乘的计算为:

$$\begin{aligned} \langle a_1, a_2, a_3 \rangle \circ \langle b_1, b_2, b_3 \rangle = \\ \langle a_1 b_1, a_2 b_2, a_3 b_3 \rangle \end{aligned} \quad (6)$$

同时 xDeepFM 模型中还分别包含了集成的线性模型和 DNN 神经网络模型,前者使得模型具有泛化的记忆能力,后者使得模型能够隐式地学习特征的交互特性。

3 烟草异常数据挖掘建模分析流程

3.1 整体流程分析

基于 Stacking 的集成学习^[15]是按照一定的方式将多种不同的算法集成组合来提升模型的训练效果,相比于单一的模型,使用该方法通常可以产生更好的预测性能。与 Bagging^[16]和 Boosting^[17]采用单一的机器学习算法训练单个模型不一样的地方在于,Stacking 是一种每一层都可以使用多个模型来进行训练的集成学习方式,每一层的多个模型都有各自输出值,将该层每一个模型的输出值作为新的特征组合成新的数据集作为下一层的输入进行学习。

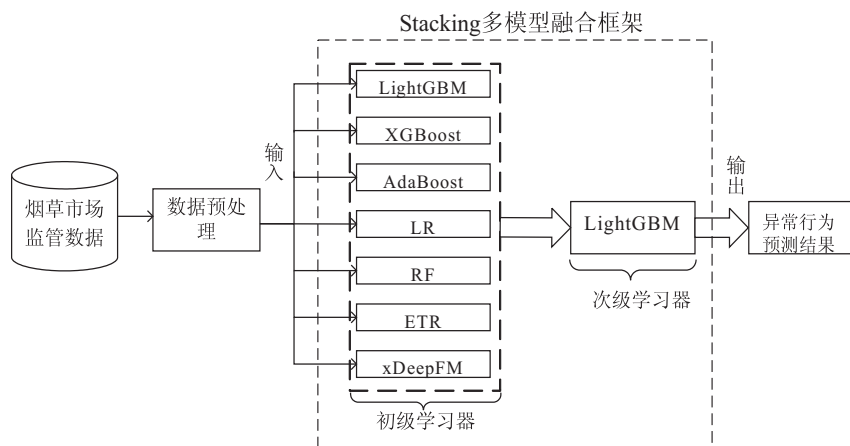


图 2 整体流程

模型构建流程如图 2 所示。首先对烟草市场监管数据进行预处理,在训练集上对单个算法进行训练调参,使单个模型性能达到最优状态;然后确定 Stacking 集成学习模型的第一层模型组合方式,利用划分后的数据集来训练,将第一层的各个初级学习器模型的输出组合形成新的数据集;Stacking 第二层次级学习器模型用新生成的数据集来训练,并输出最终的预测概率值。

3.2 烟草异常数据挖掘建模分析

烟草异常数据挖掘模型最终要实现的目标是,预测出零售户“销假,销私,乱渠道进货”等异常经营行为的可能性。基于模型的预测性能,Stacking 集成学习方式一般要求组合中的单个基学习器不仅要有较强的学习预测能力,还要在算法原理上具有较大的差别。因此 Stacking 模型中的第一层除了选用学习性能比较强的 XGBoost 算法、LightGBM 算法和 xDeepFM 算法,

还使用了 AdaBoost 算法、随机森林算法 (random forest, RF)、极端随机树算法 (extratrees, ETR) 和 Logistic Regression 算法 (LR)。其中 RF 和 AdaBoost 分别使用了基于 Bagging 与 Boosting 的集成学习方式,具有较强的学习能力和严谨的数学理论作为支撑^[18]。ETR 算法是在 RF 的基础上多了一层随机性,即在对连续变量特征选取最优分裂值时,不会计算所有分裂值的效果来选择分裂特征,而是在每一个特征的取值范围内,随机产生一个分裂值,从中计算出一个

较优值来进行分裂。其次与 RF 使用 Bagging 集成学习方式对样本数据进行有放回抽样不同,ETR 使用所有的样本,只是特征是随机选取的。LR 算法相对来说是弱一点的基学习器,使用该算法的原因是为了防止过拟合,让 Stacking 模型具有更强的鲁棒性。Stacking 模型中的第二层的元学习器用了学习预测能力比较出色的 LightGBM 算法。基于多模型 Stacking 的烟草异常数据挖掘模型如图 3 所示。

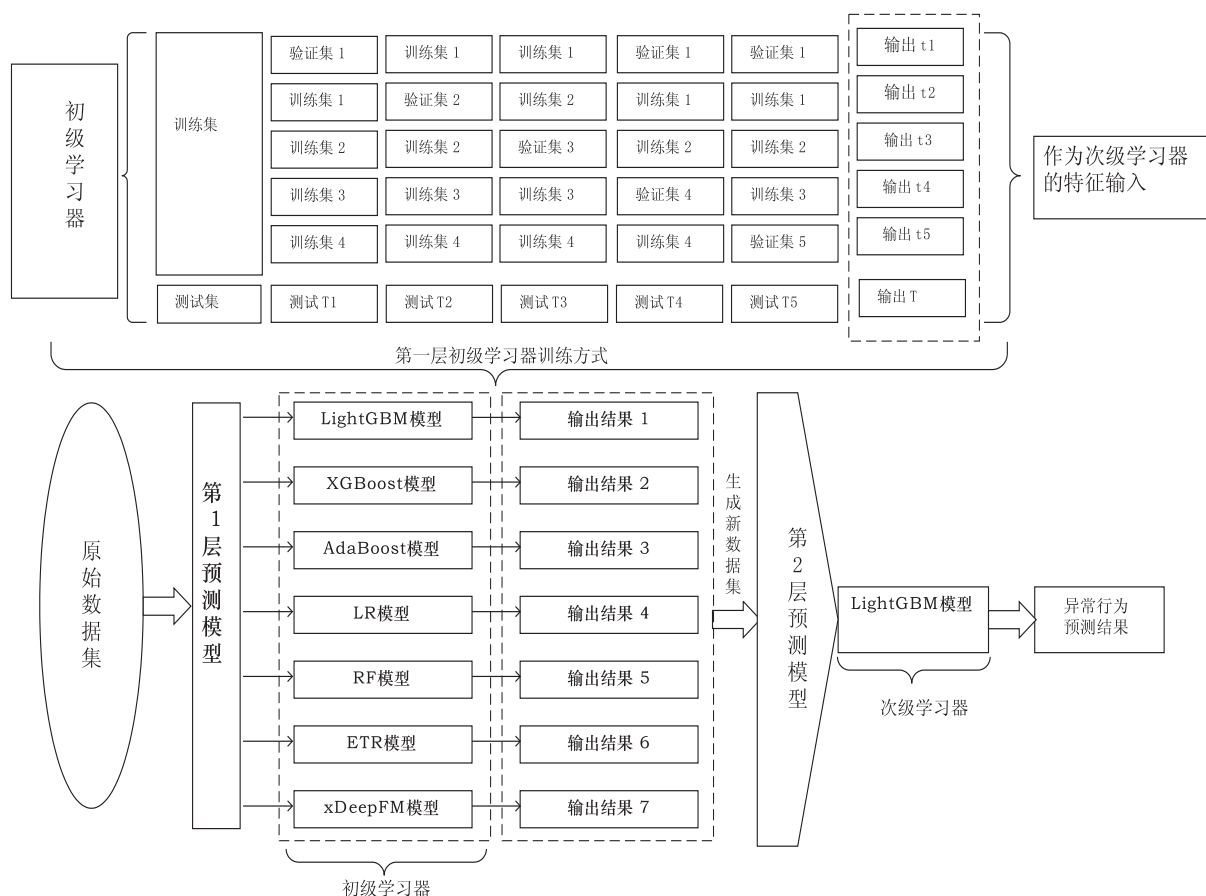


图 3 基于多模型 Stacking 的预测模型

Stacking 模型训练具体步骤如下：

(1) 划分原始数据集,其中划分的方式为随机采样选取 90% 的数据作为训练集,10% 的数据作为测试集,在训练集上使用五折交叉验证的方式对单个算法模型进行训练,确定每一个模型的最优参数,使单个模型性能达到最优状态;

(2) 确定 Stacking 第一层模型组合方式,利用划分后的数据集来训练,将第一层的各个模型的输出组合形成新的数据集,具体过程如图 3 中上半部分,其中每个模型最终的输出结果为五次交叉验证结果的平均值,将每个模型的输出结果作为新的特征组成一个新的数据集;

(3) Stacking 第二层模型用新生成的数据集来训练,并输出最终的预测概率值。

3.3 模型训练与结果分析

实验数据使用经过整理好的 2016 年 1 月到 2019 年 4 月上海市烟草专卖市场监管数据中的检查数据以及对应的静态和动态指标数据作为模型的数据集。总共 166 563 个样本,244 个特征,其中 30 个静态特征和 214 个动态特征。

由于该模型预测属于二分类预测问题,且最终的输出值为概率值,为了直接分析模型输出的概率值,预测评价指标采用 Log_loss 和 AUC 来评价模型的预测效果,避免了将其转换成类别数值带来的可能误差。公式如下所示:

$$\text{Log_loss} = - \sum_{i=1}^N \sum_{j=1}^M y_{ij} \log \left(\frac{1}{N} p_{ij} \right) \quad (7)$$

其中, N 为样本的总数; M 为预测的类别数,比如文中

实验为二分类预测, M 就为 2; 样本 i 属于分类 j 时 $y_{i,j}$ 为 1, 否则为 0; $p_{i,j}$ 为样本 i 被预测为第 j 类的概率。

$$AUC = \frac{\sum_{i \in P} \text{rank}_i - \frac{M * (M + 1)}{2}}{M * N} \quad (8)$$

其中, M 为正样本总数; N 为负样本总数; P 为正样本; rank_i 代表样本按照预测概率得分从小到大进行排

序后样本 i 的位置序号; $\sum_{i \in P} \text{rank}_i$ 为所有正样本的排名之和。

要想使融合模型 Stacking 性能达到最好, 首先要确保其第一层的各个基学习器达到最佳的学习能力, 因此将各个基学习器在原始数据集上单独训练, 从而确定每一个模型的训练参数, 具体参数如表 2 所示。

表 2 模型参数

算法名称	参数设置
LightGBM	'learning_rate': 0.01, 'num_leaves': 95, 'max_depth': 7, 'min_data_in_leaf': 40, 'feature_fraction': 0.7, 'bagging_fraction': 0.9, 'bagging_freq': 36, 'lambda_l2': 0.1, 'min_split_gain': 0.2
XGBoost	'learning_rate': 0.01, 'n_estimators': 5000, 'max_depth': 9, 'min_child_weight': 1, 'colsample_bytree': 0.9, 'colsample_bylevel': 0.7, 'reg_lambda': 1.4, 'reg_alpha': 0.5,
xDeepFM	embedding_size = 16, learning_rate = 0.01, dnn_net_size = [128, 128, 128], cross_output_size = 1, cross_layer_size = [128, 128, 128], batch_size = 256
AdaBoost	max_depth = 12, n_estimators = 500
LR	n_jobs = -1, random_state = 2019
RF	n_estimators = 2000, oob_score = True
ETR	n_estimators = 2000

在相同的数据集上对每个单一模型和 Stacking 模型分别进行训练并预测, 最佳模型通常具有较小的 Log_loss 值以及较大的 AUC 值, 各个模型的预测结果对比如表 3 所示, 对应 ROC 曲线如图 4 所示。

表 3 模型预测结果

算法名称	Log_loss	AUC
Stacking 模型	0.353 5	0.843 1
xDeepFM	0.369 7	0.837 9
LigtGBM	0.378 2	0.823 2
XGBoost	0.651 7	0.793 6
AdaBoost	6.152 4	0.611 6
LR	7.078 2	0.502 7
RF	5.660 9	0.610 7
ETR	5.235 8	0.724 0

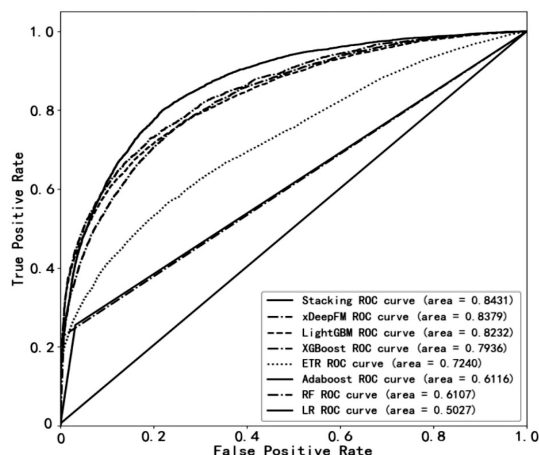


图 4 ROC 曲线对比

通过预测结果可知, 单个模型中表现最好的是 xDeepFM 神经网络算法, 说明该算法可以很好地学习不同特征之间的交叉特性, 加上模型兼具记忆和泛化的学习能力, 使得其在最终的预测精度上表现更好。其次是 LightGBM 算法, 两项指标也都达到了不错的效果, 对比其他几个机器学习基学习器, 可以确定 LightGBM 算法比较适合处理这种大样本, 高维度, 特征稀疏的数据集。虽然其他几个基学习器的表现稍差, 但是通过 Stacking 方式集成以后, 效果上更加出色。一方面是由于 Stacking 模型可以很好地保持学习能力优异的单个学习器的性能, 提升自身的预测能力; 另一方面基学习器之间算法原理的明显不同使得 Stacking 集成后的模型具有更加稳健的预测性能。

4 烟草异常数据挖掘模型的应用

经过前期阶段充分测试、验证模型的有效性后, 该文提出的基于多模型 Stacking 集成学习的烟草异常数据挖掘模型, 在上海市烟草专卖市场监管工作中进行了实际应用, 对模型的推荐名单进行了稽查实证。

本次实证数据分别选取截止 2019 年 06 月 30 日和 2019 年 07 月 31 日这两天的上海市烟草专卖数据, 将数据处理成相应的特征指标作为模型的测试集, 来对 7 月份和 8 月份的稽查名单进行预测, 其中 7 月份推荐的烟草零售户为 1 322 户, 8 月份推荐的烟草零售户为 1 344 户, 最后对稽查结果计算最终的查实率。具体数据如表 4 所示。

表 4 实证结果

区局	当月需检查户数		实际检查户数		立案户数		查实率	
	7 月	8 月	7 月	8 月	7 月	8 月	7 月	8 月
宝山	77	79	75	71	5	25	6.49%	31.65%
杨浦	51	51	45	48	10	14	19.61%	27.45%
松江	88	90	86	86	19	23	21.59%	25.56%
普陀	42	45	39	42	6	11	14.29%	24.44%
浦东新区	247	249	247	246	53	51	21.46%	20.28%
嘉定	99	102	96	96	12	20	12.12%	19.61%
徐汇	40	40	40	39	7	7	17.50%	17.50%
静安	54	54	50	51	4	9	7.41%	16.67%
奉贤	90	93	90	90	21	14	23.33%	15.05%
虹口	50	50	47	37	6	7	12.00%	14.00%
金山	103	106	101	106	9	14	8.74%	13.21%
黄浦	63	63	62	62	5	7	7.94%	11.11%
青浦	96	96	95	93	8	10	8.33%	10.42%
闵行	100	101	89	93	17	10	17.00%	9.90%
崇明	90	92	80	75	9	9	10.00%	9.78%
长宁	32	33	32	33	3	2	9.38%	6.06%
全市	1 322	1 344	1 274	1 268	194	233	14.67%	17.34%

表中涉及到的计算公式如下:

$$\text{查实率} = \frac{\text{立案户数}}{\text{当月需检查户数}} \times 100\% \quad (9)$$

其中,立案标准主要分为三类:(1)真烟流入,即零售户从其他渠道低价购买香烟再高价卖出的情况,稽查时若零售户真烟流入条数大于等于 5 条则进行立案处理;(2)假烟,即零售户有贩卖假烟的情况;(3)走私烟,即零售户有销售走私烟的情况。

此外表中部分地区存在实际检查户数低于当月需检查户数的情况,这是因为存在个别零售户当月暂不经营的情况,实际检查中做另外的处理。

上海市烟草专卖市场监管体系现有稽查方法主要依据违规加分制,即对零售户的卷烟经营数据进行分析,对零售户的违规行为按照一定的规则对其赋分,最终得分越高的零售户,其违规风险越高。结合 2016 年 1 月到 2019 年 4 月的检查数据及检查结果分析得知,原有检查方式在实际稽查中,每个月检查的零售户中有涉烟违法行为的查实率在 5% 左右。而由表 4 可以看出,在 7 月份和 8 月份 Stacking 模型预测名单的查实率分别达到了 14.67% 和 17.34%,相比原有的传统方式有比较大的提升,稽查实证结果进一步证明了 Stacking 模型的有效性。

5 结束语

基于深度神经网络 xDeepFM 算法,机器学习

LightGBM、XGBoost 等算法,利用集成学习 Stacking 方式将多个算法学习器进行集成组合,构建了基于多模型 Stacking 集成学习的烟草异常数据挖掘预测模型。对 2016 年 1 月到 2019 年 4 月的上海市烟草专卖数据进行训练及验证分析,在 2019 年 7 月和 8 月对模型推荐名单进行实地稽查验证,两个月的查实率均达到了预期,使得上海市卷烟市场监管稽查工作中的人员调拨分配更加合理,对零售户涉烟违法行为的监管更加精准,有效净化了卷烟市场的经营环境。

同时,从稽查结果的查实率可以看到存在各区局查实率结果不平衡的问题,因此,在后续的研究中会在以下几个方面继续优化完善:

(1) 可以引入权重因子,使各区局预测精度更加准确;

(2) 除了机器学习算法外,着重研究目前较为流行的深度学习算法,挖掘特征之间更高阶的有效信息;

(3) 将异常行为综合预测分析与现有市场监管处理流程进行充分结合,形成从数据预处理到模型构建再到评估应用的全流程处理模式,建立智能化的全流程市场监管处理流程,全面提升市场监管水平。

参考文献:

- [1] 吴明山,王 冰,起亚宁,等. 卷烟销量组合预测模型研究[J]. 中国烟草学报,2019,25(3):84-91.
- [2] 于焕杰,杜子芳. 基于随机森林的企业监管方法研究[J].

- 管理世界,2017(9):180-181.
- [3] 李天剑,黄 斌,刘江玉,等. 卷积神经网络物体检测算法在物流仓库中的应用[J]. 计算机工程,2018,44(6):176-181.
- [4] KOTSIANTIS S,KOUMANAKOS E,TZELEPIS D,et al. Forecasting fraudulent financial statements using data mining [J]. International Journal of Computational Intelligence, 2006,3(2):104-110.
- [5] CHEN Z,JIANG F,CHENG Y,et al. XGBoost classifier for DDoS attack detection and analysis in SDN-Based cloud [C]//2018 IEEE international conference on big data and smart computing (BigComp). Shanghai: IEEE, 2018:251-256.
- [6] BUI D T,HO T C,PRADHAN B,et al. GIS-based modeling of rainfall-induced landslides using data mining-based functional trees classifier with AdaBoost, Bagging, and Multi-Boost ensemble frameworks [J]. Environmental Earth Sciences,2016,75(14):1101-1124.
- [7] 李雄飞,周晋男,张小利. 基于混合模型的广告转化率问题研究[J]. 东北大学学报:自然科学版,2019,40(7):942-947.
- [8] LIAN J,ZHOU X,ZHANG F,et al. Xdeepfm:combining explicit and implicit feature interactions for recommender systems[C]//Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining. London,United Kingdom;ACM,2018:1754-1763.
- [9] CHEN T,GUESTRIN C. Xgboost: a scalable tree boosting system[C]//Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. San Francisco,USA;ACM,2016:785-794.
- [10] KE G,MENG Q,FINLEY T,et al. Lightgbm: a highly efficient gradient boosting decision tree[C]//Advances in neural information processing systems. Long Beach,California, USA;NIPS,2017:3146-3154.
- [11] GRBOVIC M,CHENG H. Real-time personalization using embeddings for search ranking at airbnb [C]//Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining. London,United Kingdom;ACM,2018:311-320.
- [12] ZHANG W,DU T,WANG J. Deep learning over multi-field categorical data [C]//European conference on information retrieval. Padua,Italy;ECIR,2016:45-57.
- [13] QU Y,CAI H,REN K,et al. Product-based neural networks for user response prediction[C]//2016 IEEE 16th international conference on data mining (ICDM). Barcelona, Spain;IEEE,2016:1149-1154.
- [14] 郁 豹,李振华,张 凯,等. 基于 DeepFM 模型的广告推荐系统研究[J]. 计算机应用与软件,2019,36(7):307-310.
- [15] NAIMI A I,BALZER L B. Stacked generalization: an introduction to super learning[J]. European Journal of Epidemiology,2018,33(2):1-6.
- [16] LI Yongming,ZHANG Cheng,WANG Pin,et al. A partition bagging ensemble learning algorithm for Parkinson's speech data mining[J]. Journal of Biomedical Engineering,2019,36(4):548-556.
- [17] ZHANG Y,HAGHANI A. A gradient boosting method to improve travel time prediction[J]. Transportation Research Part C:Emerging Technologies,2015,58:308-324.
- [18] LOUPPE G,WEHENKEL L,SUTERA A,et al. Understanding variable importances in forests of randomized trees [C]//Advances in neural information processing systems. Lake Tahoe,Nevada,United States;NIPS,2013:431-439.