基于投票的集成学习模型和 Nelder-Mead Optimizer 的网络游戏平台顾客流失率预测研究

1. 研究背景

《“十四五”规划和 2035 年远景目标刚要》中指出要壮大数字创意产业，对提升国家文化软实力作出全面部署，并提出 2035 年建成文化强国的战略；2021 年，国家对外文化贸易基地发布了全国首项聚焦游戏企业出海的政府扶持计划——“千帆计划”，当下，国家层面正在打造一批有影响力的国产数字游戏产业，鼓励优秀文化产品“走出去”，加强国家文化品牌效应。2020 年中国自研游戏海外市场实际销售收入为 154.5 亿美元，同比增速达 33.3%，中国自研游戏应抓住当下全球数字化风潮，顺势而为迎接历史机遇。

但是伴随着新通信时代、后疫情时代、世界经济萎缩时代的到来，中国游戏发展已经陷入瓶颈期，根据《2021 年中国游戏产业报告》显示，2021 年中国移动游戏用户规模的增长率仅为 0.22%，伴随着移动互联网时代的快速发展，发展停滞的主要公司是一大批端游、页游公司，2021 年，中国移动游戏市场实际销售收入为 2255.38 亿元，占比 76.06%。显然，如今用户更青睐于移动端形式的手游。

A 公司是在落实国家文化强国战略、促进文化数字化的背景下，由文化部于 2020 年支持建设的大型页游公司，在敏锐捕捉到时代发展风向变化后，A 公司决意将游戏形式改为手游，但是在改版第一季度中，游戏用户大面积流失，收益大幅度下降。在游戏市场中，对于已流失的核心付费用户的召回，常常包括短信召回、礼包发送等，公司不仅需要支付额外通信费，设计回归活动，还会因为相应礼包破坏平台生态，当召回成功率达到 10%-30%时，已经非常不易，但是企业依旧承担着大量的损失。

L. J. Rosenberg 和 J. A. Czepiel 在 1984 年发表的 “A Marketing Approach for Customer Retention” 一文中指出，留住现有客户的成本比吸引新客户的成本低五到六倍，识别潜在的大量流失的客户，特别是普通客户，并为他们制定召回营销策略，可以节省大量的成本。本文以识别潜在的流失客户为目标，通过提前捕捉用户流失动向，及时做出反馈，从而分析客户群体喜好，进而针对性获取客群做到优化流程与结果。

2. 现状分析

在过去的几十年里，在制定电信、金融服务、零售、电子商务等竞争领

域的客户流失率的策略已经较为完善，但是将这些策略投射到游戏应用场景中，并不十分适合：

首先是关于流失用户的定义，目前大多数在互联网行业提出的基于注册的客户流失预测研究上，关于判断客户是否流失的依据往往是客户不活动的天数超过了某个预设的阈值；又比如 Alberts 等人（2006）提出的的另一种涉及 α 和 β 两部分的定义：α 是由相关部门确定的固定值，β 表示用户的最大连续不活动天数，当客户连续不活动天数大于 α+β 时，该客户被标记为流失者。但得出上述指标定义需要较长的观察期，当预测过程完成时，用户已经离开了这个平台，从而无法帮助企业留住即将流失的客户。总而言之，目前关于流失客户的定义过于绝对，并且没有考虑到提前避免客户流失这一点，使得这些定义不能用在关于游戏用户流失的案例上。

其次，是关于机器学习（Machine Learning）方法的选择，目前在用户流失的预测上，决策树、逻辑回归、随机森林、神经网络和集成学习等方法已经被广泛应用。使用机器学习来主动防止客户流失和提高客户保留率的方法，已经被证明对相关公司来说是有价值的，在文献中，越来越多的研究关注构建时间序列特征以灵活响应市场变化，然而，只有少数研究集中在游戏场景中的流失预测，因此，我们的研究可以填补该领域的空白。

本文对判断客户流失的定义，提出一个能客观表示客户在线时间下降率的指标；使用四个强学习器进行数据的训练；最后采用能够调整权重的软投票法（Soft voting algorithm），以突出表现良好的基础学习者的贡献，减轻那些表现不佳的学习者的影响，综合考虑权重；我们创新性地将软投票法和 Nelder-Mead Optimizer 相结合，形成 NM-SoftVoting 算法，通过对子模型的权重分配，使最终建立的最优模型具有更高的精度，通过实际实验结果表明，本文所提出的模型具有较高的操作效率。在相同的方法中使用融合特征比使用其他形式更有效的特性。此外，集成学习方法在总体上比其组成的基础模型具有更好的性能。我们提出的算法 NM-SoftVoting，在召回率上普遍比现有的集成算法高出 1 个百分点，可以很好地进行 A 公司的召回预测服务。通过采用这种方法，A 公司可以实施有针对性的保留率干预措施，从而减少客户流失和提高客户保留率，做到对于游戏企业流程优化结果的评估与及时的预测。

3.优化过程

3. 1 问题描述

现 A 公司旗下游戏，由于数月利润锐减，开始分析 60 天内游戏平台全体注册用户相关数据（共计 274,392 条），所谓的注册用户相关数据指的是能反应用

户日常游戏活动、充值记录、以及社交情况等的维度，从中可以分析出用户的活跃情况，并将活跃情况数据模型化。

通过对 60 天内的数据进行分时序处理，按照构建数据集—观察流失情况—预测流失情况—检验模型拟合程度的顺序进行数据处理，A 公司将以本文提出学习模型进行数据分析，从而展开对客户流失率的研究，以到达改进未来游戏体验，并控制用户流失数量最少的目标。

在长时间的日常运作总结后，A 公司将玩家流失周期分为三个阶段，分别是：

- 初步阶段——用户在第一次或前几次登录后就退出，通常是游戏风格和新手引导的缺陷所导致；
- 沉浸阶段——在用户深入探索游戏较长时间后退出，通常是服务体验、内容设计等方面需要提升；
- 成熟阶段——在用户完成目前已有的游戏目标后退出，通常是游戏本身内容较为乏味，剧情起伏较少所导致；

为了掌握目前游戏的待提升点，需要掌握流失用户的消失节点，并针对该节点用户退出的原因，对游戏展开提升；具体策略如下：

初步阶段——缩短新手期体验时间，加速玩家操控对象的升级等

沉浸阶段——区分用户提供专项服务，避免负面事件的产生

成熟阶段——加速游戏主线的开发，提升玩家的新鲜度

A 公司旨在通过数据分析，掌握将要流失用户的数据，并及时进行跟进，给予该将要流失客户所在节点需要的游戏奖励同时发送召回短信引导用户回流，从而从避免用户成为永久退出游戏的流失者，导致公司花费五到六倍的成本吸引新用户。

3.2 模型的建立

首先，A 公司为用户提供的服务基于用户在游戏平台上的注册，为了确定潜在流失客户，提出一个表示客户在线时间下降率的指标$MTD(Modified-Time-Decline)$，从而刻画从现有数据中分割丢失的客户，$t_1$和$t_2$分别为两个连续时间窗口的用户活动时间，$\eta=\frac{t_2-t_1}{t_1}$表示在这两个连续窗口中用户在线时间的下降率，$d_1$为用户在第一个时间窗口中的在线天数，$e^{-(d_1-1)/k}$作为修正项，将下降率$MTD$描述为：

$$I_{MTD}=e^{-(d_1-1)/k}\cdot\frac{t_2-t_1}{t_1}$$

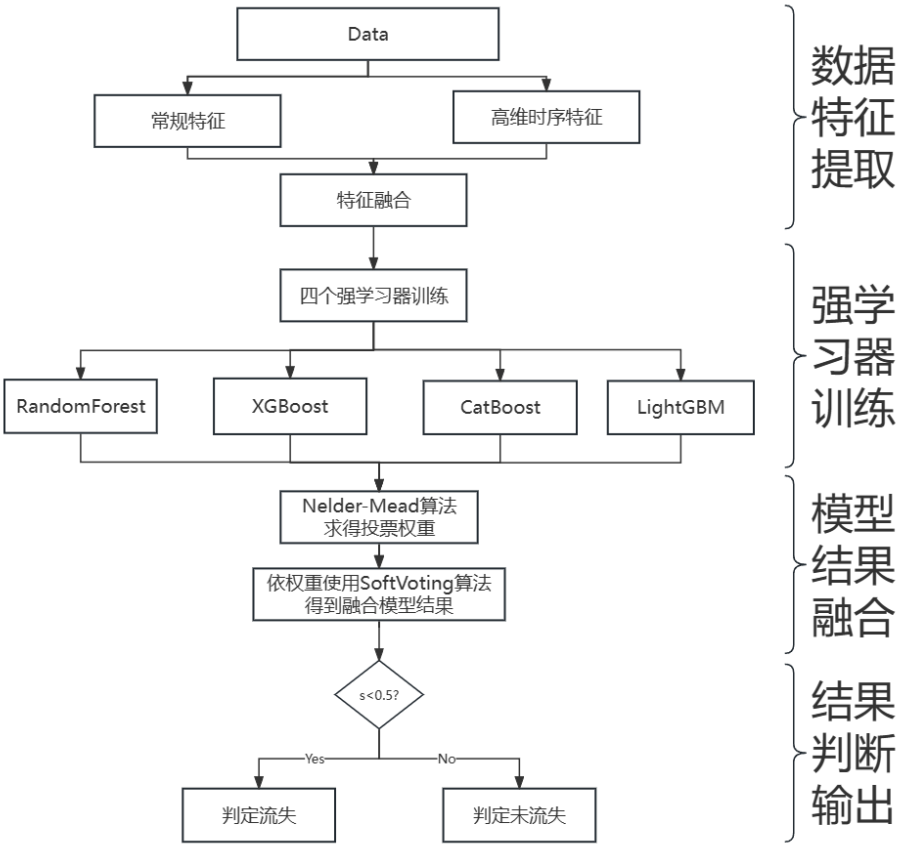
现在将下降率超过某一阈值的客户被定义为流失者($churner$):

$$Label(c_i = churner) = \begin{cases} 1, I_{MTD} > T \\ 0, I_{MTD} \leq T \end{cases}$$

其中，T 是建议的客户流失指标的阈值，当指标I_{MTD}的值大于阈值时，用户c_i被标记为流失者，否则，他/她将不被视为流失者。

3.3 算法设计

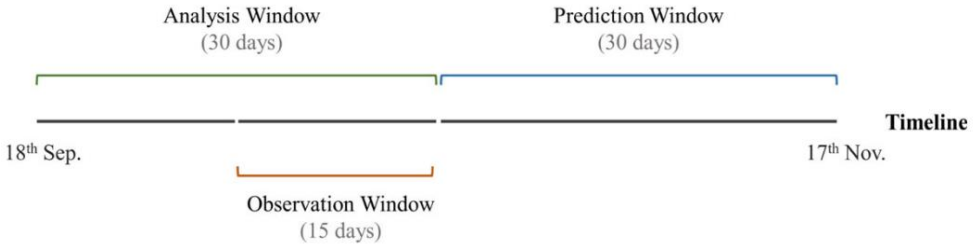
为获得较好的拟合预测效果，使用软投票法进行模型融合，从而优化各个强学习器在不同的方面较好的表现。其软投票法权重基于 Nelder-Mead optimizer 获取软投票法集成学习权值。其整体算法流程可分为四大步骤：数据分割与特征提取、强学习器训练、基于 Nelder-Mead 优化权重的 Opt_SoftVoting 算法获得融合结果、结果判断输出。



3.3.1 数据分割处理

将 60 天的时间序列分为三个时间窗口，即分析窗口、观测窗口和预测窗口，如图<>所示。分析窗口对应 60 天窗口中的前 30 天，用于构建多元时间序列数

据集。观察窗口涵盖分析窗口的最后 15 天，用于观察用户使用时间的下降情况，以构建 I_{MTD} 数据集。预测窗口是分析窗口后的 30 天，在此期间，Opt_SoftVoting 算法来预测客户是否流失。此外，为确保数据的完整性和行为模式的稳定性，我们过滤掉了在观察开始前 14 天内注册的客户，因为他们不是我们留住客户的目标受众。



3.3.2 数据特征提取

对于常规特征将其划分为四类：使用记录、充值记录、社交属性记录、用户基础信息。其具体分类如下：

Data Categories	Description
Usage records	Daily aggregated data on customer behavior, such as active time, time using the microphone, room entry count, etc. for a set period.
Gift and monetary records	Gift and monetary records including quantity and value of gifts sent/received, max recharge amount, etc. count at the end of Analysis Window.
Social attribute records	customer social attribute data for a set period, such as following people, friends, and followers, etc. count at the end of Analysis Window.
Demographic information	demographic information of customers about their country, age, gender, number of days since registration, etc.

为获取高维时序特征，使用 FRESH 方法进行高维时序特征的提取。对于高维时序特征包括以下特性：

1、具备时间序列分布的基本统计信息，包括序列的分布类型、发散程度、高斯值、离群值等。

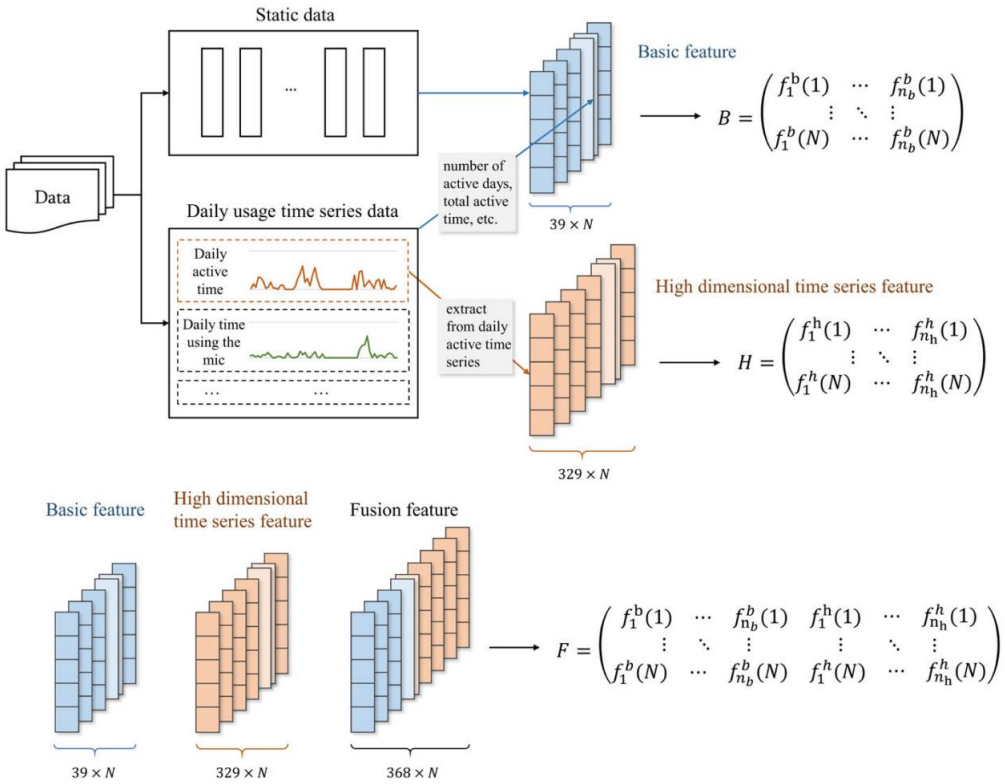
2、线性相关，包括自相关、功率谱密度等

3、静态性，包括 StatAv、滑动窗口测量、预测误差等。

4、信息论和复杂性度量，包括自信息和互信息、近似熵和 Lemp 信息、近似熵和 Lempel-Ziv 复杂性。

5、线性和非线性模型拟合，包括 ARMA、高斯过程和 GARCH 模型。

最终获得 39 项常规特征与 329 项高维时序特征，并将其融合，获得 *Fusion* 数据，以进行后续学习器的训练。



3.3.3 强学习器训练

在实验后使用四种强学习器进行特征提取与融合。其分别是 XGBoost、RandomForest、LightGBM、CatBoost 四种个体强学习器进行训练。其分别训练步骤分别如下：

（1）依靠数据集之中的高维时序特征与常规特征融合后特征，训练 XGBoost 模型。其流程如下：

1、利用得到的训练集、观测数据与评估数据利用 pandas 中的 DataFrame 格式，利用 SKlearn 中 XGBoost 模型接口进行读入训练。

2、训练模型使用评估数据集进行验证。

3、得到二元结果进行归一化处理，获取预测值，并使用二元混淆矩阵记录验证结果。

（2）依靠数据集之中的高维时序特征与常规特征融合后特征，训练

RandomForest 模型。其流程如下：

1、利用得到的训练集、观测数据与评估数据利用 pandas 中的 DataFrame 格式，利用 SKlearn 中 RandomForest 模型接口进行读入训练。

2、训练模型使用评估数据集进行验证。

3、得到二元结果进行归一化处理，获取预测值，并使用二元混淆矩阵记录验证结果。

（3）通依靠数据集之中的高维时序特征与常规特征融合后特征，训练 LightGBM 模型。其流程如下：

1、利用得到的训练集、观测数据与评估数据利用 pandas 中的 DataFrame 格式，利用 SKlearn 中 LightGBM 模型接口进行读入训练。

2、训练模型使用评估数据集进行验证。

3、得到二元结果进行归一化处理，获取预测值，并使用二元混淆矩阵记录验证结果。

（4）依靠数据集之中的高维时序特征与常规特征融合后特征，训练 CatBoost 模型。其流程如下：

1、利用得到的训练集、观测数据与评估数据利用 pandas 中的 DataFrame 格式，利用 SKlearn 中 CatBoost 模型接口进行读入训练。

2、训练模型使用评估数据集进行验证。

3、得到二元结果进行归一化处理，获取预测值，并使用二元混淆矩阵记录验证结果。

并使用二元混淆矩阵以记录模型预测结果。

		预测值	
		1	0
真实值	1	TP, True Positive	FN, False Negative
	0	FP, False Positive	TN, True Negative

3.3.4 训练结果评估

为了评估模型效果，采用四种常用指标作为评估模型预测效果：

（1） 准确率（Accurancy）

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

准确率等于所有预测正确的样本数除以总样本数，反映了模型的整体效果。

（2） 召回率（Recall）

$$Recall = \frac{TP}{TP + FN}$$

召回率又称敏感度 (Sensitivity)，表示所有真实为正例的样本中，被模型正确预测的样本占比。在流失预测中，通常会更加关注流失用户，越高的召回率代表模型尽量捕捉出了更多的流失用户。

（3） F1 值（F1-score）

$$F1 = \frac{2}{Precision^{-1} + Recall^{-1}} = \frac{2Precision * Recall}{Precision + Recall}$$

其中，精确率 $Precision = \frac{TP}{TP+FN}$ 表示预测为正例的样本中预测正确的比例。很多时候精确率与召回率会有矛盾，作为精确率与召回率的调和平均值的 *F1* 值可以同时兼顾两者，进行整体评价。

（4） AUC

ROC(The Receiver Operating Characteristic Curve)是一条描述不同判定阈值下假正率 (False Positive Rate, FPR)与召回率的对应关系曲线。AUC(Area Under Curve) 为 ROC 曲线下的面积，具体量化了召回率与假正率的平衡。通常来说，高召回率下模型会捕获更多的正样本，但与此同时负样本的错判率也会上升， AUC 值越大，说明模型在尽量捕捉更多正样本的同时，对负样本的错判率的上升影响越小，因此模型的分类效果就越好。

3.3.5 Nelder-Mead 算法获得融合模型权重

Nelder-Mead 方法(Downhill Simplex Method) 由 Jone Nelder 和 Roger Mead 于 1965 年提出，它基于单纯型的概念构造相应的迭代优化策略来搜寻多维空间中目标函数最值，属于启发式优化算法之一。单纯形指 n 维空间中 n+1 个仿射无关的点的集合的凸包，可以理解为空间中最简单的多面体，如一维空间中的直线、二维空间中三角形、三维空间中的四面体。相较于目前常见的优化算法，如拟牛顿法等，Nelder-Mead 更容易收敛到局部极值，但好处在于不需要了解目标

函数具体形式也无需目标函数可导，常用于解决导数未知的非线性优化问题。对于 n 维最小化问题 min f(x)，Nelder-Mead 搜索方法如下：

1) 针对 f(x)，$x \in R^n$ 的最小化问题，首先选择 n+1 个点，使其构成一个初始单纯形，具体方式为限定初始点 $x^{(0)} = P_0$, $P_i = P_0 + \lambda_i e_i$, $i = 1,2,\ldots,n$，按照这种方式产生 n+1 个点。正好构成一个单纯形。
2) 判断各顶点对应的 f(x)是否满足终止条件，若满足则终止，返回当前最优值；否则，对顶点进行重排序

$$f(x_1) \leq f(x_2) \leq \ldots \leq f(x_{n+1})$$

3) 计算前 n 个点的重心

$$x_0 = \frac{1}{n} x_i$$

4) 计算反射点 x_r；如果 $f(x_1) \leq f(x_r) \leq f(x_n)$,则用 x_r 替换 x_{n+1} 构成新的单纯形，返回步骤 2

$$x_r = x_0 + \alpha (x_0 - x_{n+1}), \quad \alpha \geq 0$$

5) 若 $f(x_r) \geq f(x_n)$，计算收缩点 x_c；如果 $f(x_c) \leq f(x_{n+1})$，则用 x_c 替换 x_{n+1} 构成新的单纯形，返回步骤 2；否则仍用 x_r 替换 x_{n+1} 构成新的单纯形，同样返回步骤 2

$$x_e = x_0 + \gamma (x_r - x_0), \quad \gamma \geq 1$$

6) 此时有 $f(x_r) \geq f(x_n)$，计算收缩点 x_c；如果 $f(x_c) \leq f(x_{n+1})$，则用 x_c 替换 x_{n+1} 构成新的单纯形，返回步骤 2

$$x_c = x_0 + \rho (x_{n+1} - x_0), \ 0 < \rho \leq 0.5$$

3.3.6 Opt_SoftVoting 算法获得结果

在训练得到的强学习器模型得到结果乘以软投票法对应权重，获得最终结果。

$$\langle \bar{x}_i, W^T \rangle = \begin{bmatrix} h_1(x_i) & h_2(x_i) & \cdots & h_T(x_i) \end{bmatrix} \begin{bmatrix} \omega_1 \\ \omega_2 \\ \vdots \\ \omega_T \end{bmatrix} = s_i$$

而最终结果由计算得到的 s_i 的值决定。

$$y_i = \begin{cases} 1, s_i > 0.5 \\ 0, s_i < 0.5 \end{cases}$$

4.实际案例对比分析

本文以 A 公司为研究对象，利用 A 公司自 11.18 后 60 天的用户数据进行优化。在过滤了观察数据开始前 14 天注册的用户后，总共得到了 219,910 条观察数据。基于此数据，进行常规特征获取以及基于 FRESH 框架提取的高维时序特征的获取。使用 A 公司数据常规特征与高维时序特征融合后的特征，训练验证四个学习器：RF、XGBoost、LightGBM、CatBoost。其训练后预测结果如表：

评估指标 / 算法	Accuracy	Recall	F1-score	AUC
RF	0.8843	0.7446	0.7242	0.9438
XGBoost	**0.8864**	**0.7521**	**0.7298**	0.9450
LightGBM	0.8863	0.7501	0.7290	0.9452
CatBoost	0.8856	0.7449	0.7265	**0.9455**

最后使用 Opt_SoftVoting 算法，对于四个使用融合特征训练的强学习器，利用 Nelder-Mead 算法优化后权值进行融合。

其 Nelder-Mead 算法中设定反射系数 α、扩展系数 γ、收缩系数 ρ、回退系数 τ 均取标准值：

$$\alpha = 1,\quad \gamma = 2,\quad \rho = \frac{1}{2},\quad \tau = \frac{1}{2}$$

初始单纯形构造中固定步长 $\delta = 0.05$。特别地，如果初始点在某个维度上坐标为 0，沿着该维度得到的点在该维度上坐标置为常数值 0.025。此外，搜索次数 K 设置为 100，留出集样本比例取训练样本的 10%，最终得到最优权值为：

$$W* = (0.3747, 0.0584, 0.5680, 0.0273)$$

最终得到结果如表：

评估指标 / 算法	Accuracy	Recall	F1-score	AUC
Opt_SoftVoting	0.8867	0.7657	0.7337	0.9456

5.总结

本文针对网络游戏平台顾客流失率预测，提出了一种基于投票的集成学习二元分类模型。利用常规特征与高维时序特征融合后特征训练 RF、XGBoost、LightGBM、CatBoost 模型，并创新性的使用基于 Nelder-Mead 算法优化后的 Opt_SoftVoting 算法以提升集成学习模型性能。

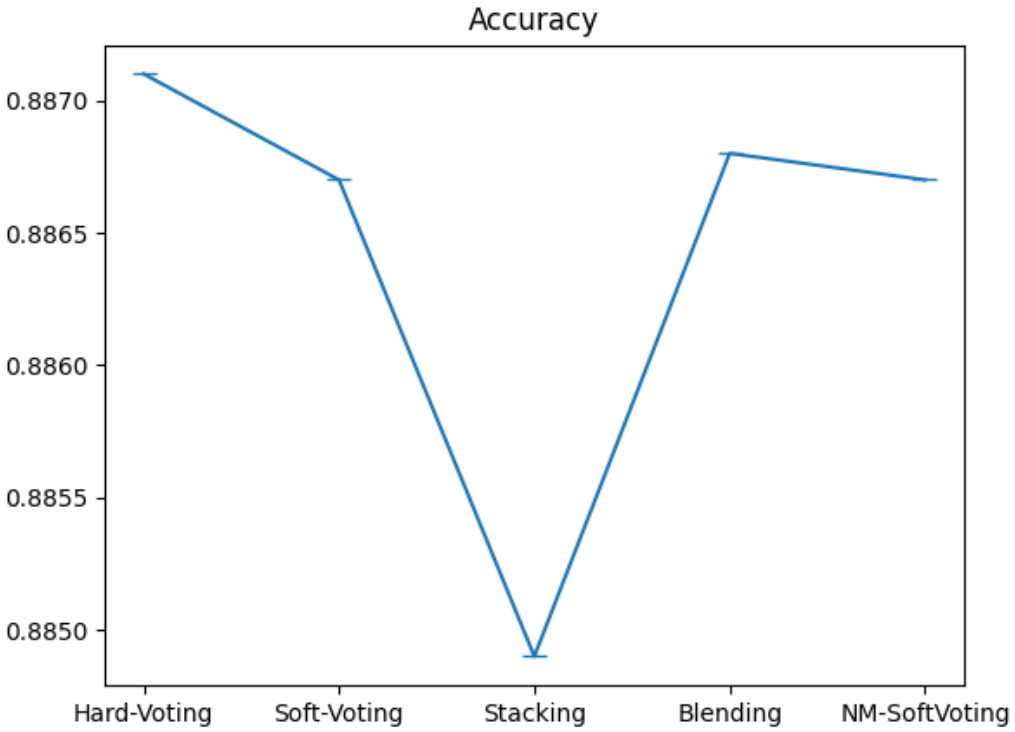
通过本文的研究，引入高维时序特征融合辅助训练模型，且使用 Nelder-Mead

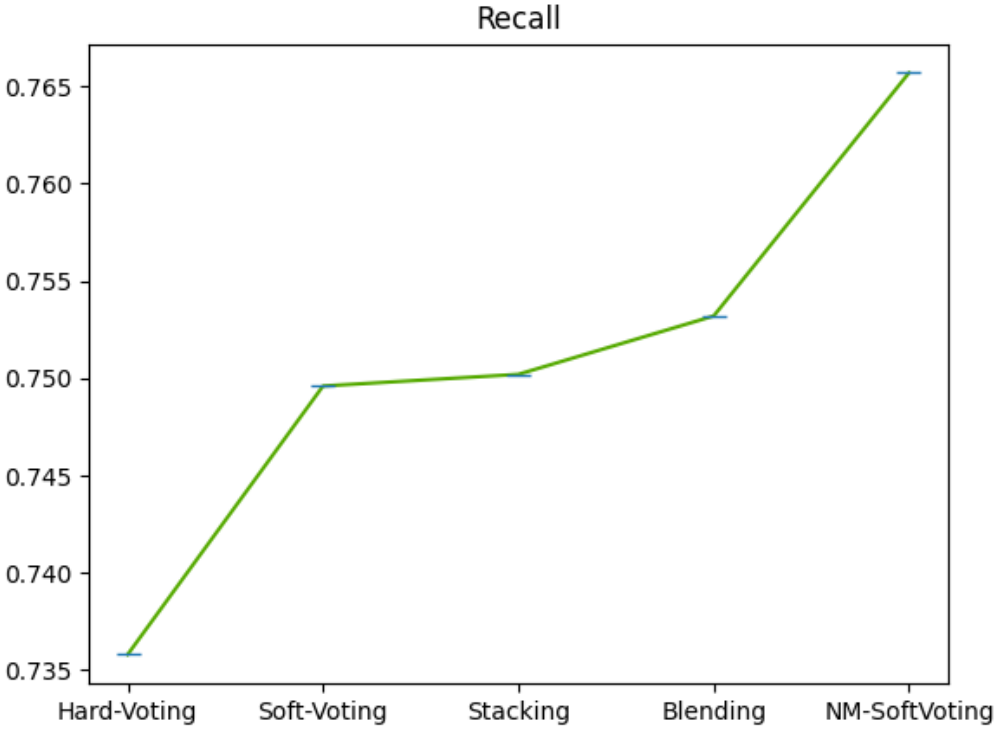
算法提升基于投票的集成学习模型效率。能够有效辅助公司衡量用户召回效果，帮助公司增加用户粘性、服务核心客群，节约客群运营成本，提升客户经营效率。

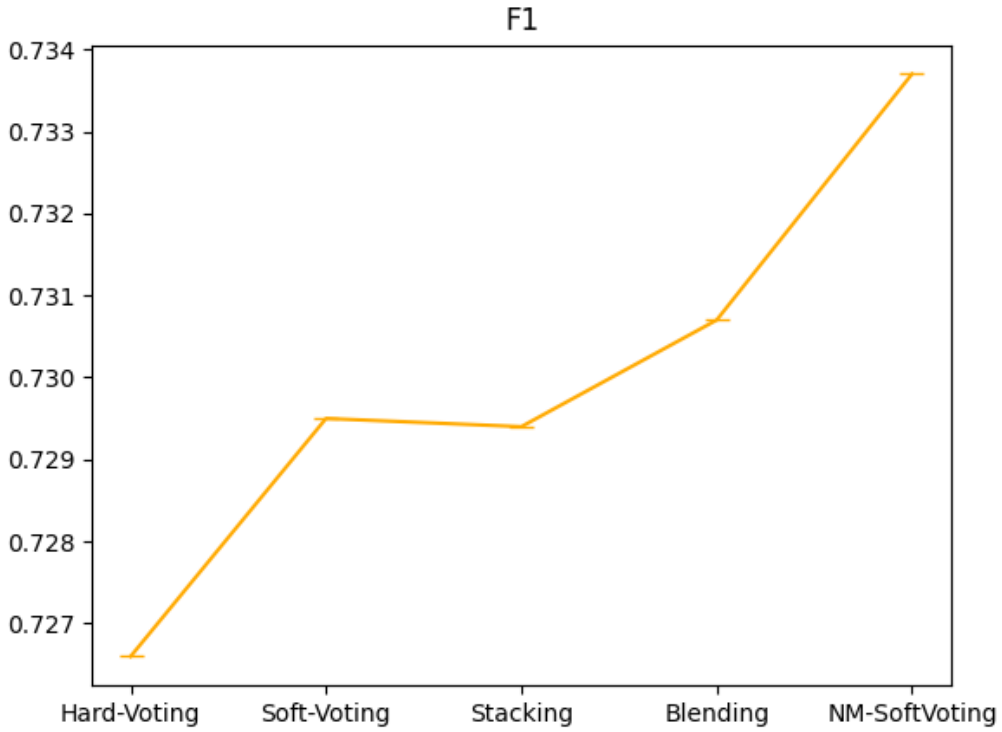
附录

集成学习算法效果对比

集成学习算法	Accuracy	Recall	F1	AUC
Hard-Voting	0.8871	0.7358	0.7266	-
Soft-Voting	0.8867	0.7496	0.7295	0.9457
Stacking	0.8849	0.7502	0.7294	-
Blending	0.8868	0.7532	0.7307	-
NM-SoftVoting	0.8867	0.7657	0.7337	0.9456


Accuracy
0.8870
0.8865
0.8860
0.8855
0.8850
Hard-Voting
Soft-Voting
Stacking
Blending
NM-SoftVoting


Recall
0.765
0.760
0.755
0.750
0.745
0.740
0.735
Hard-Voting
Soft-Voting
Stacking
Blending
NM-SoftVoting


F1
0.734
0.733
0.732
0.731
0.730
0.729
0.728
0.727
Hard-Voting
Soft-Voting
Stacking
Blending
NM-SoftVoting