队伍编号	MCB2201112
赛道	B

基于多模型调参优化的 Stacking 用户评分预测集成学习

摘　　要

这里是摘要

关键词：影响程度量化分析；特征工程；Stacking 集成学习；评分预测；可视化评估

目录

一、问题的提出

1.1 问题背景

随着移动通信技术的迅猛发展和网络工程的不断建设，在信息透明、产品同质化的今天，提升语音通话及网络服务的质量，满足用户对高质量语音通话、网络服务的需求显得尤为重要。由于当今用户数量的不断增多、用户需求不断提高、运营商业务不断广泛化，因此点对点、传统方法解决问题逐渐困难化。而现在有来自移动通信集团北京分公司根据用户对语音业务及上网业务的满意度进行的评分及相关影响因素的数据，我们需要对其进行分析、建立相关数学模型，以便从数据中心获得有效信息，更高效地提升服务质量，为客户提供更好的服务。

1.2 问题要求

- **问题一**：研究并量化分析影响用户对语音及上网业务满意度的主要因素；
- **问题二**：建立基于影响用户评分影响因素的数学模型，并依据附件 3、4 中相关因素对其评分进行预测，并解释预测评分的合理性。

二、问题的分析

2.1 问题的整体分析

该题是一个关于移动用户对语音及上网业务体验评分的数据分析、预测类问题。

从分析目的看，本题需要分析用户对语音与上网业务的评分及各个影响因素，筛选出影响用户评分的主要因素，并量化结果。同时需要对用户的评分进行预测及研究，为运营商提供参考，从而提升用户语音及上网的优质体验。因此本题主要需完成两方面任务：**其一**，研究影响用户语音及上网业务满意度的主要因素，并对各因素进行量化分析；**其二**，根据上述的分析，建立合理模型，对用户的评分进行预测及研究，确保分类模型的准确性、稳健性、可靠性，并有一定的泛化能力，且能够包容用户真实评分的主观性。

从数据来源、特征看，本题的数据来源于北京移动用户的语音与上网业务评分数据，数据包括用户对语音业务下“语音通话整体满意度”“网络覆盖与信号强度”“语音通话清晰度”“语音通话稳定性”，上网业务下“手机上网整体满意度”“网络覆盖与信号强度”“手机上网速度”“手机上网稳定性”方面的评分，以及相关的影响评分的因素。评分数据具有主观性，影响因素数据具有高维、多样、标准体系不一致、量纲不一致等特点，且数据量较大。因此，本题数据相对特殊且复杂，需要对数据进行一定的预处理，以便于后续的分析。

从模型的选择看，本题数据量较大、维度较高，且分析目的是分析影响用户评分的主要因素，并对用户的评分进行预测及研究。本文将评分视为多分类，且评分具有一定主观性、分类种类多，因此，在模型的选择上，本文结合多种分类预测模型，构建集成学习模型，尽可能多地学习到用户评分特点，提升模型的准确性、稳健性及可泛化性能。

从软件的选择看，本题为数据类型，且需要进行大量的数据分析、预测等，因此我们选择 Python Jupyter 对问题进行求解，其交互式的编程范式，方便且高效。

2.2 问题一的分析

问题一的核心目的在于**研究并量化分析影响用户对语音及上网业务满意度的主要因素**。对于已给的数据集，数据在完整度、指标标准等方面存在一定缺陷。这导致在原数据上我们不可直接进行分析，需要对原数据集进行数据的预处理。此外附件数据集在语音及上网业务中，每一业务均有四项评分，因此我们需要对每一项评分进行分析，对各因素进行量化。结合数据来源、与特征方面，我们综合**皮尔逊相关系数、熵权法、灰色关联度分析、随机森林分类**，构建多元量化分析模型，尽可能准确挖掘到影响用户评分的因素，为构建后续预测模型提供优质依据。

2.3 问题二的分析

问题二的核心目的在于**建立基于影响用户评分影响因素的数学模型，并依据附件 3、4 中相关因素对用户评分进行预测，并解释模型预测的合理性**。但是在附件 1 与附件 2，附件 3 与附件 4 中，影响因素存在不配对的情况。这导致在给定用户评分的数据中，部分因素不可作为模型建立的基础特征数据，因此在数据预处理的同时，还需要对附件 1 与附件 2，附件 3 与附件 4 中的**影响因素列取交集**，使得学习数据与预测数据的特征数据一致。此外，在已给的存在用户评分的数据集中，用户对每一项的评分均为整数，不存在小数，且评分范围为 $[1,10]$。因此，我们在建立预测模型时，应尽量避免使用回归模型，而应使用**分类模型**，但部分分类模型需要分类标签量值从 0 开始，因此需要对所有评分进行标签编码，规范数据。同时分类种类较多，对于单一模型，其预测准确率较低，平均绝对误差较高、泛化能力较弱……因此，本文结合多种机器学习模型，构建**集成学习模型**，尽可能准确预测用户评分。最后，在此基础上，结合模型的**分类混淆矩阵热力图、分类报告、ROC/AUC 曲线**等对于预测结果进行解释，叙述模型的合理性，同时考虑集成学习模型对预测误差的包容性，对模型的泛化能力进行分析。

三、符号说明

符号	符号说明
μ	样本平均值
σ	样本方差
x_{standard}	经过标准化后的数据
$R(x)_{m\times n}$	经过某项处理后的数据特征集
ρ	皮尔逊相关系数
x'	经过某项处理后的数据
$Gini$	样本集合基尼系数
$\hat{y}$	预测值
$L^{(t)}$	目标函数
Ω	叶节点正则项惩罚系数
P	某事件发生的概率
ω	权重

四、模型的假设

- **假设一**：语音与上网业务的八项评分中，存在个别用户乱评、错评现象；
- **假设二**：除个别用户的部分评分外，其余所有数据真实且符合实际情况；
- **假设三**：用户评分还受到除附件中因素之外的因素的影响；
- **假设四**：给定的数据集可全面体现用户整体情况；
- **假设五**：对于同一业务，学习数据与预测数据的内在规律是一致的。

五、模型的建立与求解

六、模型的评价与推广

6.1 模型的评价

- **模型的优点：**

1. 对数据进行综合处理，层次清晰，模型具有一定解释性；
2. 数据标准化，避免量纲不一造成的偏向学习影响的情况；
3. 特征筛选，减少不重要性因素占比，减少数据维度，提升模型学习效率，一定程度上避免数据噪声，适当降低模型复杂度，使模型高效化，防止过拟合；
4. 特征构造，由原数据构造出新数据特征，适当增多数据维度，防止欠拟合；
5. 综合熵权法、灰色关联度分析及随机森林量化影响程度，避免局部最优；
6. 对各模型进行参数调优，尽可能提高模型的多方面能力；
7. 加入正则化方法，一定程度上也可防止过拟合；
8. 交叉验证，更好地利用数据集，减少数据浪费，提高模型的泛化能力，验证模型的稳健性，防止过拟合情况的发生；
9. 模型设置任意随机种子，在保证划分训练集及测试集的一般性、随机性的同时，确保可重复性的结果，方便后续处理；
10. 通过主成分分析及随机森林进行特征选择，保证客观性；
11. 多模型 Stacking 集成学习，更好地利用已有数据，多方面学习数据中内在联系，结合多个模型优良方面，避免陷入局部最优，对数据有更好的把控能力，提升模型的泛化能力、提高预测准确率、提高模型稳健性、鲁棒性，同时减小预测误差，且对异常值有一定识别能力。

- **模型的缺点：**

1. 模型对于小样本分类的识别能力较差，难以对这些用户进行深入分析；

2. 模型对于预测主观性评分，难以提供完全一致的评分结果；
3. Stacking 在构造时，有一定复杂度，对基模型的要求较高；
4. 对于部分评分，特征构造出的因素有一定局限性；
5. 用户评分为主观性结果，本文大多模型选用客观性较强的模型进行解决，对数据利用有一定失真。

- **模型的改进：**

1. 在收集数据时，问卷设计需要更加合理化，多方面考虑其余未考虑到的影响因素对用户评分的影响；
2. 在允许条件下获得更多训练样本；
3. 对各模型可以选用非完全一致的特征，提升各模型的独特性，有目的地进行选择，减少学习的数据维度，加快模型收敛速度，使得模型学习高效化，结果准确化；
4. 适当增加或减少数据维度，建立复杂度适中的模型；
5. 对不平衡的多分类，可以采用“下采样”或“上采样”方法，使得分类平衡，但需要更多的数据集；
6. 可适当增加基模型个数，并提高对基模型的筛选要求；
7. 对主观性评分，可以建立主客观相结合的模型，从而优化模型各项指标。

6.2 模型的推广

机器学习可利用现有的数据集进行有目的的训练，在此基础上预测分类标签下人为难以确定的结果，极大方便了当今对复杂数据的处理；多种机器学习相互结合，利用 Stacking 集成学习的方法，可以有效提高模型各方面能力，减少判断错误的情况。针对小部分样本的学习，需要更容易区分类别的特征进行学习，以及利用特征工程等方法进行解决。对于机器学习模型，我们可以作出其可视化图像，观察到模型的各项指标不易发现的问题，如欠拟合、过拟合等情况，我们可以依据模型效果评估可视化来对模型进行一定的调优。本文是以移动用户对业务的评分为基础，我们运用了多种机器学习的模型，再结合 Stacking 进行集成学习，可以发现模型的效果较优，对主观性评分模型有较好把控能力。利用该模型，可以根据用户对某些影响因素的情况，预测用户对于这项业务的满意程度，再结合相关描述性信息，有的放矢地解决用户遇到的问题，提升客户的满意程度，提升产品的服务质量，从而为业务创造更多价值。该模型在一定程度上虽有一定欠缺，但不仅仅可用于该领域的评分，也可用于其余领域，如用户对于某一产品的评价预测，根据用户评价，改善产品质量，提升经济效益，实现双赢。

参考文献

[1] CSDN.【数据预处理】sklearn 实现数据预处理（归一化、标准化）[EB/OL]. https://blog.csdn.net/weixin_44109827/article/details/124786873.

[2] 肖杨, 李亚, 王海瑞, 常梦容. 基于皮尔逊相关系数的滚动轴承混合域特征选择方法 [J]. 化工自动化及仪表,2022,49(03):308-315.DOI:10.20030/j.cnki.1000-3932.202203009.

附　录

[A] 图表

[B] 支撑文件列表

支撑文件列表如下（列表中不包含原始数据集）：

文件（夹）名	描述
result.xlsx	用户评分预测结果
所有量化结果.xlsx	问题一量化结果
模型参数.xlsx	各个模型评估参数以及模型选择依据
语音业务词云.txt	语音业务词云图文本内容
上网业务词云.txt	上网业务词云图文本内容
语音业务数据分析.ipynb	语音业务分析 Jupyter 文件
上网业务数据分析.ipynb	上网业务分析 Jupyter 文件
语音业务数据分析.html	语音业务分析运行结果
上网业务数据分析.html	上网业务分析运行结果
bg.jpg	词语底图
figuresNightingaleRoseDiagramF.py	原始数据用户评分南丁格尔玫瑰图程序
figuresNightingaleRoseDiagramP.py	预测数据用户评分南丁格尔玫瑰图程序
figuresOne	语音业务所有图示文件夹
figuresTwo	上网业务所有图示文件夹
figuresNightingaleRoseDiagramF	原始数据用户评分南丁格尔玫瑰图示（八项评分）
figuresNightingaleRoseDiagramP	预测数据用户评分南丁格尔玫瑰图示（八项评分）

[C] **使用的软件、环境**

为解决该问题，我们所使用的主要软件有：

- TeX Live 2022
- Visual Studio Code 1.74.2
- WPS Office 2022 冬季更新（13703）
- Python 3.10.4
- Pycharm Professional 2022.3

Python 环境下所用使用到的库及其版本如下：

库	版本	库	版本
copy	内置库	missingno	0.5.1
jieba	0.42.1	mlxtend	0.20.2
jupyter	1.0.0	numpy	1.22.4+mkl
jupyter-client	7.3.1	openpyxl	3.0.10
jupyter-console	6.4.3	pandas	1.4.2
jupyter-contrib-core	0.4.0	pyecharts	1.9.1
jupyter-contrib-nbextensions	0.5.1	scikit-learn	0.22.2.post1
jupyter-core	4.10.0	seaborn	0.11.2
jupyter-highlight-selected-word	0.2.0	sklearn	0.0
jupyterlab-pygments	0.2.2	snapshot_phantomjs	0.0.3
jupyterlab-widgets	1.1.0	warnings	内置库
jupyter-latex-envs	1.4.6	wordcloud	1.8.1
jupyter-nbextensions-configurator	0.5.0	xgboost	1.6.1
matplotlib	3.5.2	yellowbrick	1.4

[D] 问题解决源程序

D.1 语音业务分析代码 [针对附件 1 与附件 3]