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

- 初赛要求：
  - **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。
- 复赛要求：
  - **问题一**：结合初赛的分析、研究结果，分析用户对语音及上网业务的评分高低，并得出高分组与低分组的特征；同时对客户评分的合理性进行分析，筛选出评分合理的客户数据，并利用新的数据集重新建立预测模型，再对附件 3、4 中评分进行预测；
  - **问题二**：依据初赛及复赛的分析结果，设计一份不超过一页纸的非技术报告，并将发现及建议提供给中国移动北京公司。

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，本题需要结合初赛的分析、研究结果，对数据集进行再分析，分析评分高分组与低分组的各自特征，对原数据集进行重采样，进行更深层次的分析。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**，结合初赛的分析、研究结果，分析用户对语音及上网业务的评分高低，并得出高分组与低分组的特征；同时对客户评分的合理性进行分析，筛选出评分合理的客户数据，并利用新的数据集重新建立预测模型，再对附件 3、4 中评分进行预测；**其二**，依据初赛及复赛的分析结果，设计一份不超过一页纸的非技术报告，并将发现及建议提供给中国移动北京公司。

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 初赛总结

针对问题一，主要需要对用户语音及上网业务评分影响因素的程度进行量化分析。我们首先对数据集进行统一处理，包括：**初步剔除相关列数据、学习数据与预测数据指标一致化、指标规范化、空缺值处理、标签编码、特征构造、数据标准化、学习数据与预测数据一致化、学习数据训练集与测试集划分**。之后在处理好的数据集上建立**熵权法、灰色关联度分析、随机森林分类**模型，多方面综合考虑，量化分析各影响因素对评分的影响程度，并依此来确定影响用户两项业务满意度的主要因素。量化结果接近于实际生活，效果良好，且可为后续问题奠定基础。

针对问题二，主要需要根据已有影响因素对用户的评分进行预测，并解释预测的合理性。我们首先结合问题一量化结果以及建立**主成分分析**模型，对数据**累计方差**进行解释，确定特征个数；之后建立 **XGBoost 模型**，并得出各影响因素的重要性，与随机森林模型结合分析，确定特征的选择；再建立 **KNN、SVM、LightGBM** 以及**多分类逻辑回归模型**，对数据进行学习分析；随后，对各个模型进行**超参数调优**，模型准确率均有大幅度提升，如随机森林**较原先提升了 11.69%，最高提升较原先可达到 14.25%**，效果良好。再者，以模型的准确率、平均绝对误差、均方误差为标准，选择表现较优的模型作为 **Stacking 集成学习**的基模型，同时选择余下的一个模型作为第二层模型，在提升准确率的同时，避免过拟合。同时对其采用**五折交叉验证**，验证其**稳健性**。Stacking 集成学习结果符合预期效果，且明显优于单一模型。在保证准确率的同时，预测的平均绝对误差、均方误差**均有一定优化**，同时我们还注重结果的可解释性及模型的现实意义。最后，我们进行**可视化分析**，绘制原始数据及预测数据评分人数**南丁格尔玫瑰图**，查看数据分布，绘制模型的**混淆矩阵热力图、分类报告、ROC/AUC 曲线**，多方面评估模型效果及解释模型的合理性。综合上述分析，可以确认模型效果良好，具有良好的稳健性、泛化能力。

最后，我们还对所建立的模型的优缺点进行了中肯的评价、提出了模型的改进措施以及对模型进行了一定推广。

因此，本复赛解决方案建立在初赛的分析及结果基础之上，且重要部分本文会再提及。

2.3 问题一的分析

问题一的核心目的在于**对原数据集进行重采样，并进行更深层次的分析**。对于主观性因素过强的用户评分数据，为尽可能提升预测的准确率，我们需要先筛选出高分组及低分组用户，对其行为特征进行分析，筛选出评分合理的用户，依此重新建立分类预测模型，提升移动公司对用户对各项业务的满意程度的把握程度，从而更好地解决现存问题，为用户提供更优质服务。

2.4 问题二的分析

问题二的核心目的在于**为移动公司撰写一份非技术性报告，为其提供合理性建议，从而为客户提供更好的服务**。

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

本文对于模型的假设与初赛假设一致，如下：

- **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

5.1 相关准备工作

为方便、准确、高效解决问题，我们需要对数据进行预处理，其主要过程见图 1，包括：初步剔除相关列数据、学习数据与预测数据指标一致化、指标规范化、空缺值处理、标签编码、特征构造、标准化、学习数据与预测数据一致化、学习数据训练集与测试集划分。**本文后续的模型建立都在此基础之上**。上述过程与初赛大致相同，但本文对部分进行的合理地修改，以适应本题要求。

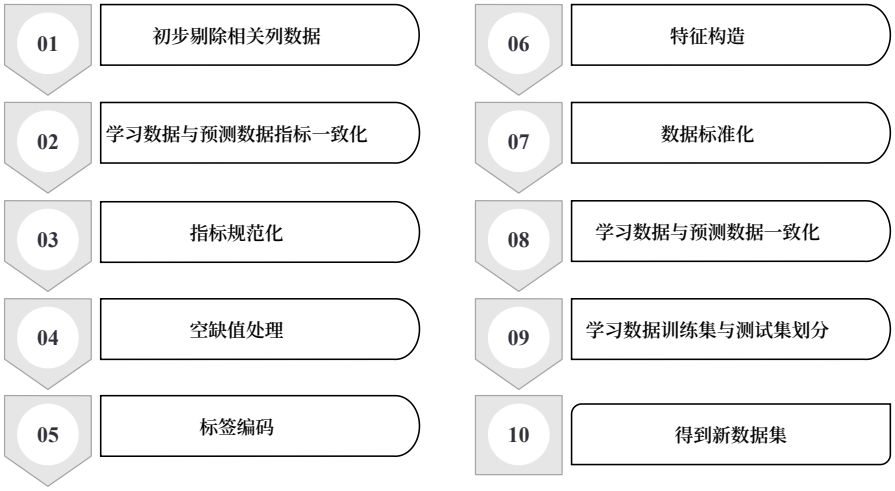

图 1　数据的准备主要过程

Step1 初步剔除相关列数据

由于“用户 id”为连续编号，且与评分无任何关系，故本文将该列数据剔除；同时对于“用户描述”等文字性叙述指标，由于其均为文本，且描述特征难以提取，难以量化，本文将该列数据剔除，但为了获得客户相关描述，本文将绘制用户描述高频词汇云图；此外，对于“终端品牌类型”等多类别指标，由于其类别较多，量化后难以提取出有效信息，故也将其剔除，其余列暂时保留。

Step2 学习数据与预测数据指标一致化

附件 1 与附件 3 为用户语音业务数据，但两表数据影响的因素存在不一致的现象，需要对指标取交集，确保两者一致，附件 2 与附件 4 同理。这里我们利用 Python 中集合 set 容器元素唯一性特征及 pandas 库，筛选出相同因素。而对于可能重合的指标，我们在下文也会进行一定处理。这样即可确保在学习数据上建立的模型依据的指标存在于预测数据集上，避免两者指标不一的情况。

Step3 指标规范化

经过上述处理后，我们发现大部分影响因素为分类指标，其划分“是”“否”的字段不统一，故依据“附件 5 附件 1、2、3、4 的字段说明.xlsx”文件，本文对附件 1、2、3、4 中的

分类指标进行规范化，记“是”类别为 1，“否”类别为 0，方便后续模型的建立。

Step4 **空缺值处理**

在给定数据集中，部分空缺值可以依据附件 5 的解释进行填充。经过一定处理后，附件 3 与附件 4 中无空缺值，故本文对附件 1 与附件 2 中的空缺值进行分析与处理：

- **对于附件 1：**据附件 5 解释进行填充后，还存在个别用户的空缺值，空缺值的列名为：“是否 4G 网络客户（本地剔除物联网）”“终端品牌”“是否 5G 网络客户”“客户星级标识”，且这些空缺值均在同一用户中出现，用户 id 分别为 1573、1601、2326、2827、3265。附件 1 的空缺值有集中、个数少的特点，存有空缺值的用户仅有 5 个，占整体用户的 0.0920%，对于模型的建立影响较小，因此我们将这 5 行用户剔除。

- **对于附件 2：**经过指标一致化后及初步数据空缺值的填补后，附件 2 中仅剩“终端品牌”列指标存在 14 个空缺值，根据该列数据其余特征，我们将这个 14 个空缺值以 0 填充。

Step5 **标签编码**

首先，对用户的评分进行编码，由于部分分类模型需要分类量值从 0 开始，因此，为方便后续集成学习等，本文将评分从 $[1,10]$ 映射至 $[0,9]$，且仍均为整数，即将评分减 1。其次，对“终端品牌”“4\5G 用户”指标利用 Python 的 sklearn 库中的 LabelEncoder 进行标签编码。此外，对于“客户星级标识”指标，我们依据移动公司对客户星级标识的划分进行编码，编码值对应见表 1。

表 1　客户星级标识编码对应表

未评级	准星	一星	二星	三星	银卡	金卡	白金卡	钻石卡
0	1	2	3	4	5	6	7	8

Step6 **特征构造**

观察并分析给定的数据，我们可以初步构造以下特征：

- **对于附件 1 与附件 3：**观察到附件 1 中有“家宽投诉”与“资费投诉”两项，而在附件 3 中有“是否投诉”一项，因此，我们在附件 1 中构造“是否投诉”一项。若“家宽投诉”与“资费投诉”均为 0，则“是否投诉”记为 0，否则记为 1。并同时删去“家宽投诉”与“资费投诉”。

- **对于附件 2 与附件 4：**观察数据，我们构造三项新指标，分别为“出现问题场所或应用总”“网络卡速度慢延时大上不了网总”“质差总”。其来源为对应列指标按行求和。

Step7 **数据标准化**

该处标准化处理为 **Z-score** 方法，仅用于后续机器学习模型的使用。而在问题一的熵权法、灰色关联度分析中我们采用 **Min-Max** 方法，该方法在后文模型中会具体说明。

对于某一列数据 $x = [x_1, x_2, \cdots, x_m]^{\mathrm{T}}$，其平均值为

$$\mu = \frac{1}{m}\sum_{i=1}^{m} x_i \tag{1}$$

标准差为

$$\sigma = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(x_i - \mu)^2} \tag{2}$$

则标准化后的数据为

$$(x_{\text{standard}})_i = \frac{x_i - \mu}{\sigma} \tag{3}$$

利用上述计算公式，我们对非分类指标进行处理，使得原数据经过处理后，其值聚集于 0 附近，即均值为 0，标准差为 1。这样处理，利于机器学习模型的建立、学习与预测，加快模型的收敛速度，并在一定程度上提升模型的准确性。同时该标准化处理方法适合当代嘈杂的大数据场景[1]。因此对于大样本的数据，如出现部分异常值，使用该方法对最终结果影响较小。

Step8 **学习数据与预测数据一致化**

经过上述几项处理后，我们还需要将附件 1 与附件 3 数据集一致化，包括指标一致化以及数据字段、分布排列一致化，从而保证对于需要预测的数据集附件 3 利用在附件 1 中建立的模型所利用到的数据集的一致性，避免造成数据的不一致，导致预测错误。本文对附件 2 与附件 4 进行上述相同的操作。

Step9 **学习数据训练集与测试集划分**

为计算问题二中建立的模型的准确性等指标，需要在附件 1 与附件 2 中均划分训练集与测试集。对于语音业务划分训练集与测试集比例为 8:2；而对于上网业务，其比例设为 9:1。对于比例的设置，本文将在后文解释其合理性。且上述划分利用 sklearn 库中的 train_test_split 函数实现，且任意设定随机种子为 2022，确保多次调试结果的一致性。该函数可确保划分的随机性，确保训练集与测试集数据分布规律大致相同。

以箱线图所示，用户评分，划分高分组与低分组。思路：绘制评分联合分布图，观察用户行为。得出异常点，初始确定用户错评，乱评等不合理行为的依据。语音、上网业务各四项评分，联合确定高分组及低分组，筛选出高分及低分组，对数据进行可视化、描述性分析，得出高分组及低分组的特征。综合之前的初步确定，则得出用户错评的依据，对数据进行综合筛选，对筛选出的数据与原始数据进行对比，确定其合理性。重新学习、预测！分析合理性。

非技术性报告

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 1.76.1
- 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]