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。

本文将在初赛的基础上完成复赛问题要求，且解决问题的大方向不变。同时大多解决方案与初赛一致，对于特定问题，我们也将进行合理地修改。

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，并依此来确定影响评分的主要因素。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。

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，处理过程与初赛大致相同，但本文对部分操作进行的合理地修改，以适应本题要求。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**。

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”；
    - 观察到附件 1 中有多个出现问题的场所，因此我们将每一用户出现问题的场所求和，构造“场所合计”，他们为“居民小区”“办公室”“高校”“商业街”“地铁”“农村”“高铁”“其他，请注明”；
    - 观察到附件 1 中有多个类型的问题，因此我们将出现问题求和，构造出“出现问题问题合计”，他们为“手机没有信号”有信号无法拨通”“通话过程中突然中断”“通话中有杂音、听不清、断断续续”“串线”“通话过程中一方听不见”“其他，请注明.1”；

* 观察到附件 1 中“脱网次数”“mos 质差次数”“未接通掉话次数”有相似特征，故将每一用户该三项数据求和，构造出“脱网次数、mos 质差次数、未接通掉话次数合计”。

– **对于附件 2 与附件 4：**

* 观察到附件 2 中有多个出现问题的场所，构造出“出现问题场所或应用总”；
* 观察到附件 2 中“手机上网速度慢”“打游戏延时大”“显示有信号上不了网”“全部都卡顿”“全部游戏都卡顿”“手机支付较慢”“看视频卡顿”“上网过程中网络时断时续或时快时慢”“打开网页或 APP 图片慢”“全部网页或 APP 都慢”“下载速度慢”“网络信号差/没有信号”特征相似，故将每一用户对应的项目数据求和，构造出”网络卡速度慢延时大上不了网总“；
* 观察到附件 2 中“微信质差次数”以及“上网质差次数”均为质差量值，故将每一用户该二项数据求和，构造出“质差总”；
* 观察到附件 2 中的场所类别较多，故将场所求和，构造出“地点总”。

- **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。

5.2 初赛研究相关结论

在初赛中，我们通过熵权法、灰色关联度分析以及随机森林多分类模型，量化影响用户评分的各因素，并得出主要因素，结果见表 2，表 3及表 4。

表 2 语音业务总体以及四项评分各个指标影响程度量化结果

因素	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	语音业务总
GPRS 总流量（KB）	0.1266	0.1151	0.1181	0.1263	0.1215
当月 ARPU	0.1154	0.1116	0.1000	0.1032	0.1111
是否遇到过网络问题	0.0922	0.1015	0.0953	0.1080	0.0995
前 3 月 MOU	0.0996	0.1232	0.1028	0.0993	0.0964
语音通话-时长（分钟）	0.0726	0.0762	0.0754	0.0680	0.0747
当月 MOU	0.0671	0.0715	0.0722	0.0689	0.0673
mos 质差次数	0.0645	0.0512	0.0617	0.0670	0.0604
脱网次数	0.0388	0.0406	0.0361	0.0444	0.0372
未接通掉话次数	0.0334	0.0319	0.0328	0.0302	0.0327
客户星级标识	0.0315	0.0233	0.0259	0.0243	0.0316
终端品牌	0.0356	0.0297	0.0368	0.0382	0.0310
4\5G 用户	0.0114	0.0137	0.0177	0.0119	0.0212
GPRS-国内漫游-流量（KB）	0.0105	0.0112	0.0162	0.0136	0.0145
高铁	0.0107	0.0114	0.0141	0.0118	0.0124
农村	0.0117	0.0106	0.0105	0.0076	0.0124
地铁	0.0151	0.0126	0.0152	0.0129	0.0121
通话过程中突然中断	0.0081	0.0134	0.0096	0.0075	0.0117
外省流量占比	0.0133	0.0125	0.0110	0.0094	0.0115
是否 5G 网络客户	0.0116	0.0090	0.0128	0.0091	0.0112
商业街	0.0096	0.0080	0.0124	0.0071	0.0109
手机没有信号	0.0167	0.0076	0.0142	0.0177	0.0108
套外流量费（元）	0.0063	0.0077	0.0089	0.0102	0.0106
通话过程中一方听不见	0.0112	0.0121	0.0171	0.0085	0.0093
外省语音占比	0.0023	0.0069	0.0051	0.0066	0.0091
居民小区	0.0115	0.0112	0.0091	0.0148	0.0086
办公室	0.0073	0.0145	0.0117	0.0085	0.0085
省际漫游-时长（分钟）	0.0086	0.0114	0.0078	0.0060	0.0081
串线	0.0053	0.0036	0.0035	0.0057	0.0079
套外流量（MB）	0.0071	0.0070	0.0037	0.0070	0.0071
通话中有杂音、听不清、断断续续	0.0105	0.0078	0.0067	0.0106	0.0070
其他，请注明	0.0040	0.0052	0.0050	0.0065	0.0058
有信号无法拨通	0.0076	0.0051	0.0087	0.0088	0.0058
其他，请注明.1	0.0036	0.0065	0.0040	0.0048	0.0051
是否关怀用户	0.0070	0.0041	0.0055	0.0059	0.0044
是否投诉	0.0028	0.0029	0.0047	0.0032	0.0041
高校	0.0039	0.0038	0.0049	0.0033	0.0031
前 3 月 ARPU	0.0044	0.0037	0.0022	0.0030	0.0031
是否 4G 网络客户（本地剔除物联网）	0.0008	0.0005	0.0009	0.0004	0.0005

5.3 高分组与低分组的分类

5.4 用户评分分组特征分析

5.5 用户评分合理性分析

首先，我们绘制出用户对于语音及上网业务评分的箱线图，如图 2、图 3所示。

表 3　上网业务总体以及四项评分各个指标影响程度量化结果 [续表见表 4]

因素	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	上网业务总
当月 MOU	0.1563	0.2447	0.2530	0.2436	0.2278
网络卡速度慢延时大上不了网总	0.0899	0.1259	0.0292	0.1295	0.1244
终端品牌	0.0332	0.0484	0.0643	0.0594	0.0551
客户星级标识	0.0223	0.0596	0.0531	0.0549	0.0522
出现问题场所或应用总	0.4678	0.0402	0.1258	0.0407	0.0394
性别	0.0242	0.0440	0.0286	0.0315	0.0387
脱网次数	0.0383	0.0282	0.0284	0.0323	0.0318
质差总	0.0107	0.0289	0.0334	0.0330	0.0317
是否 5G 网络客户	0.0092	0.0301	0.0350	0.0267	0.0296
微信质差次数	0.0104	0.0251	0.0205	0.0210	0.0233
是否不限量套餐到达用户	0.0078	0.0183	0.0189	0.0172	0.0226
套外流量费（元）	0.0089	0.0206	0.0249	0.0213	0.0221
上网质差次数	0.0058	0.0185	0.0208	0.0183	0.0215
农村	0.0067	0.0154	0.0150	0.0154	0.0163
显示有信号上不了网	0.0105	0.0131	0.0146	0.0150	0.0161
居民小区	0.0138	0.0185	0.0185	0.0170	0.0161
地铁	0.0100	0.0141	0.0145	0.0169	0.0151
高铁	0.0016	0.0138	0.0123	0.0123	0.0150
商业街	0.0090	0.0124	0.0127	0.0149	0.0139
上网过程中网络时断时续或时快时慢	0.0038	0.0102	0.0107	0.0103	0.0137
网络信号差/没有信号	0.0046	0.0109	0.0138	0.0139	0.0133
办公室	0.0091	0.0136	0.0156	0.0144	0.0120
套外流量（MB）	0.0029	0.0107	0.0089	0.0134	0.0119
手机支付较慢	0.0000	0.0073	0.0069	0.0058	0.0086
其他，请注明	0.0031	0.0123	0.0090	0.0100	0.0079
下载速度慢	0.0023	0.0073	0.0071	0.0067	0.0074
拼多多	0.0000	0.0035	0.0033	0.0038	0.0063
打游戏延时大	0.0094	0.0044	0.0033	0.0025	0.0063
抖音	0.0015	0.0081	0.0072	0.0051	0.0059
百度	0.0000	0.0064	0.0048	0.0061	0.0059
其他，请注明.1	0.0033	0.0041	0.0028	0.0056	0.0059
看视频卡顿	0.0020	0.0056	0.0054	0.0055	0.0056
微信	0.0012	0.0054	0.0042	0.0051	0.0052

表 4　上网业务总体以及四项评分各个指标影响程度量化结果 [表 3续表]

因素	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	上网业务总
高校	0.0000	0.0046	0.0063	0.0052	0.0051
腾讯视频	0.0022	0.0042	0.0044	0.0040	0.0049
打开网页或 APP 图片慢	0.0054	0.0042	0.0045	0.0042	0.0044
全部网页或 APP 都慢	0.0023	0.0031	0.0030	0.0028	0.0044
京东	0.0000	0.0064	0.0041	0.0060	0.0043
快手	0.0000	0.0026	0.0058	0.0042	0.0041
淘宝	0.0022	0.0044	0.0043	0.0054	0.0040
手机上网速度慢	0.0022	0.0031	0.0032	0.0034	0.0038
今日头条	0.0000	0.0047	0.0044	0.0048	0.0035
王者荣耀	0.0000	0.0030	0.0021	0.0020	0.0034
新浪微博	0.0000	0.0039	0.0033	0.0041	0.0029
爱奇艺	0.0016	0.0040	0.0049	0.0035	0.0027
优酷	0.0025	0.0021	0.0014	0.0026	0.0026
芒果 TV	0.0000	0.0014	0.0016	0.0026	0.0026
全部都卡顿	0.0020	0.0033	0.0040	0.0029	0.0026
手机 QQ	0.0000	0.0027	0.0040	0.0023	0.0026
其他，请注明.2	0.0000	0.0017	0.0017	0.0010	0.0019
搜狐视频	0.0000	0.0011	0.0006	0.0013	0.0018
咪咕视频	0.0000	0.0016	0.0009	0.0019	0.0017
其他，请注明.3	0.0000	0.0016	0.0030	0.0024	0.0015
其他，请注明.5	0.0000	0.0011	0.0016	0.0004	0.0013
全部游戏都卡顿	0.0000	0.0010	0.0006	0.0005	0.0013
火山	0.0000	0.0004	0.0003	0.0000	0.0011
和平精英	0.0000	0.0015	0.0014	0.0011	0.0008
欢乐斗地主	0.0000	0.0006	0.0003	0.0010	0.0008
其他，请注明.4	0.0000	0.0012	0.0002	0.0004	0.0007
梦幻西游	0.0000	0.0000	0.0009	0.0000	0.0003
穿越火线	0.0000	0.0005	0.0003	0.0004	0.0002
炉石传说	0.0000	0.0000	0.0007	0.0003	0.0001
部落冲突	0.0000	0.0004	0.0000	0.0002	0.0001
梦幻诛仙	0.0000	0.0000	0.0000	0.0000	0.0001
阴阳师	0.0000	0.0000	0.0000	0.0000	0.0000
龙之谷	0.0000	0.0000	0.0000	0.0000	0.0000

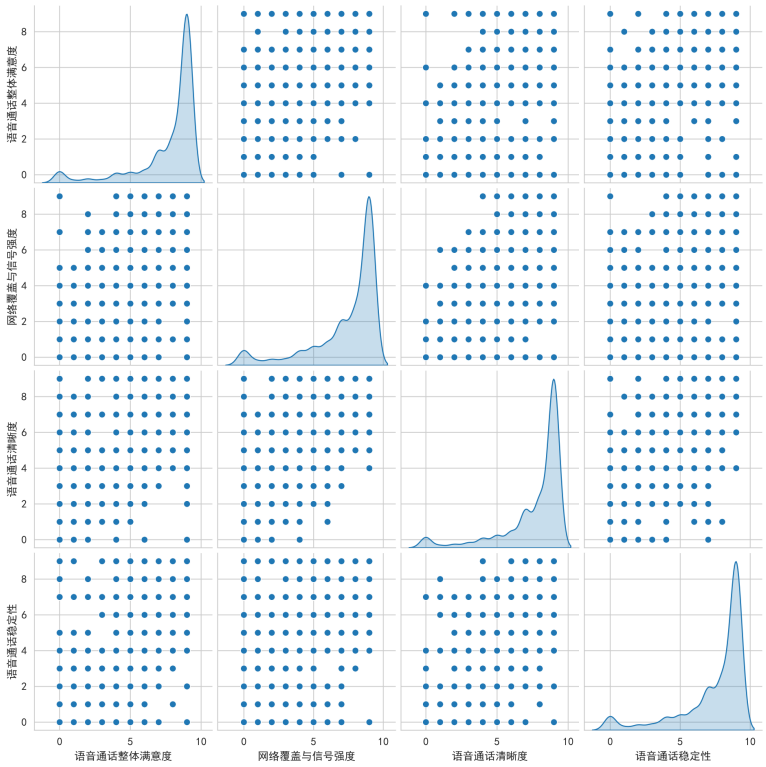


图 4　语音业务用户四项评分联合分布图

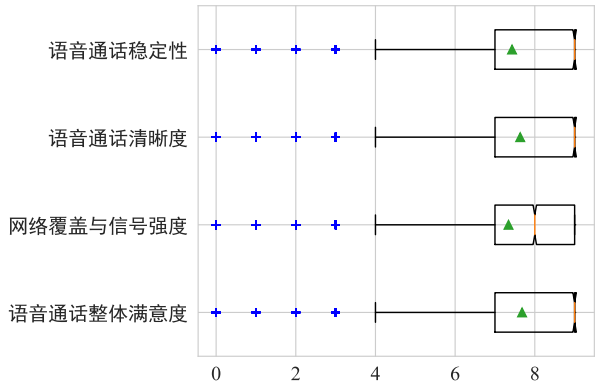


图 2　语音业务用户四项评分箱线图

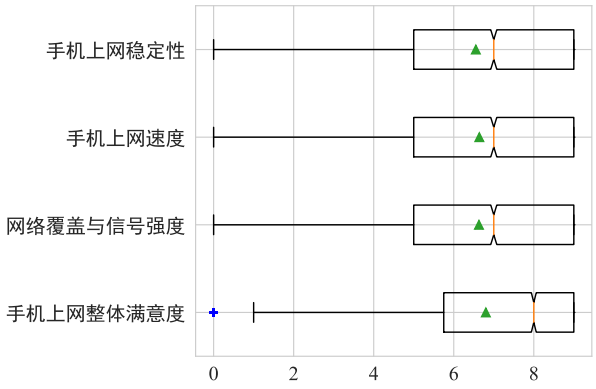


图 3　上网业务用户四项评分箱线图

依据上图结果，我们可以发现用户对于每一业务的四项评分，其分布规律大致一致，但也存在少量异常的数据，为了更清晰地发现各评分之间的散点关系，我们再绘制出语音及上网业务的评分联合分布图，如图 4、图 5所示。

对于问题一，我们综合用户评分箱线图、用户评分联合分布图，并结合初赛得出的各项因素重要性程度，对数据进行更深层次分析，筛选出评分合理的用户群体，并依据该数据建立多分类预测模型，预测未知评分的用户集体对于语音及上网业务的满意程度。

5.6 新数据集的建立

5.7 多分类模型的建立

本文依旧采用初赛使用的六种多分类模型，并对其进行 Stacking 集成学习。首先多分类基本模型建立理论如下：

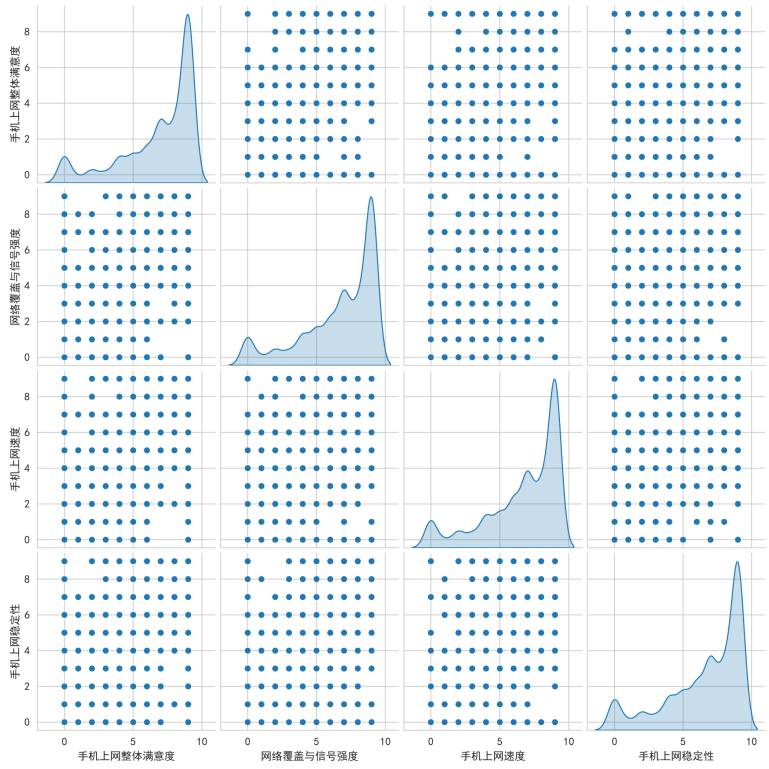


图 5　上网业务用户四项评分联合分布图

- **随机森林（Random Forest, RF）**是由多棵决策树（Decision Tree）进行组合后对预测结果投票或取均值的一种算法[3]。其有分类和回归两种模型，对于本题，我们选择分类模型。其简要过程如图 6所示，算法伪代码如Algorithm1所示。

Algorithm 1: 随机森林（RF）

Data: 数据集 $\mathcal{D}$

function DTree($\mathcal{D}$)
if *Termination* **then**
  return base (g_t)
else
  learn $b(x)$ 并且依据 $b(x)$ 划分 $\mathcal{D}$ 为 $\mathcal{D}_C$
  build $G_C \leftarrow \mathrm{DTree}(\mathcal{D}_C)$
  return $G(x) = \sum_{C=1}^{C} [\![b(x) = C]\!] G_C(x)$
end
function RandomForest($\mathcal{D}$)
for $t = 1, 2, 3, \cdots, T$ **do**
  request 数据集 $\tilde{\mathcal{D}}_t \leftarrow \mathrm{BoostStrapping}(\mathcal{D})$
  obtain DTree $g_t \leftarrow \mathrm{DTree}\left(\tilde{\mathcal{D}}_t\right)$
  return $G = \mathrm{Uniform}(g_t)$
end

Result: 随机森林模型 $G = \mathrm{Uniform}(g_t)$

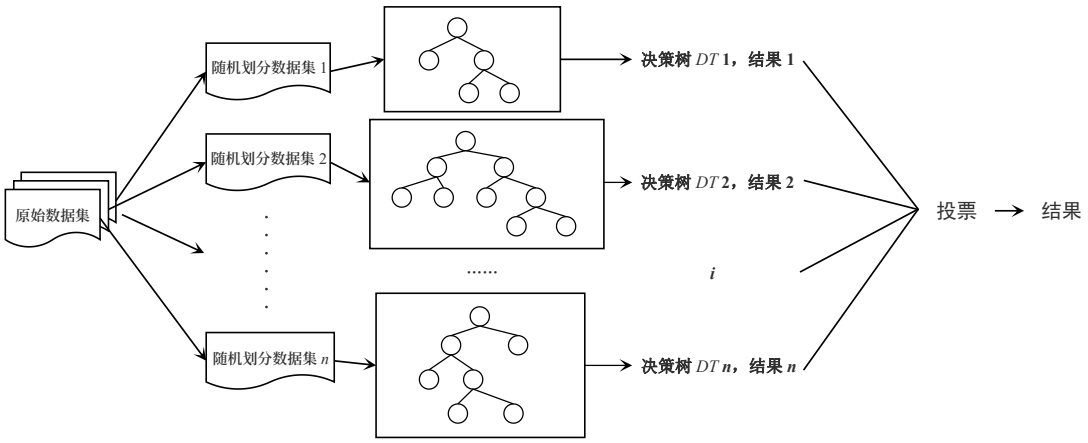


图 6　随机森林算法简图

对于单棵决策树而言，本文利用 **CART** 算法[3]构建。基尼系数是衡量样本集合纯度的指标，当该值越小时，其纯度也就越高。计算公式如下

$$Gini\left(R_i\right)=\sum_{k=1}^{K}\sum_{k\neq k'}P_kP_{k'}=1-\sum_{k=1}^{K}P_k^2 \tag{4}$$

其中，R 为选取出的特征（影响因素），K 表示在该特征中包含的类别数，P_k 表示该特征中第 k 类别的出现概率。

由上述分析可知，对于单棵决策树而言，其叶子节点的分裂特征为选择的所有特征中基尼系数最小的特征。

- **极端梯度提升（eXtreme Gradient Boosting, XGBoost）**。XGBoost 算法是一种基于树模型的优化模型，其将弱分类器组合，训练出一个较强的分类器。该算法通过多次迭代，生成一个新的树模型用于优化前一个树模型，随着迭代次数的增多，该模型的预测精度也会相应提高[4]。

  记通过数据处理后的数据集特征为 $R\left(x_{ij}\right)_{m\times n}$，表示其包含 m 个用户，n 个特征，在训练中形成的 CART 树的集合记为 $F=\left\{f\left(x\right)=w_{q(x)},q:\mathbf{R}^n\rightarrow T,w\in\mathbf{R}^T\right\}$，其中 q 为树模型的叶节点决策规划，T 为某一树模型叶节点数量，w 为叶节点对应的得分[5]。对于预测的 y 值，其计算公式为

$$\hat{y}=\varphi\left(x_i\right)=\sum_{k=1}^{K}f_k\left(x_i\right) \tag{5}$$

XGBoost 算法在每一次迭代过程中会保存前面所学习的模型，会将这些模型加入到新一轮迭代过程中，因此我们记第 i 个模型为预测结果为

$$\hat{y}_i^{(t)}=\hat{y}_i^{(t-1)}+f_t\left(x_i\right) \tag{6}$$

XGBoost 算法的目标函数计算公式如下

$$L^{(t)}=\sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)}+f_t\left(x_i\right)\right)+\gamma T+\frac{1}{2} \lambda \sum_{j=1}^{T} w_j^2+\text{const} \tag{7}$$

上述公式中，l 为模型误差损失，描述在该模型下预测值与实际值之间的出差异损失，Ω 为模型叶节点的正则项惩罚系数，γ 与 λ 为模型的超参数[5]。通常情况下，我们难以用枚举法得到在模型中所训练出来的树结构，因此这里采用贪婪算法，从单叶子节点开始，通过迭代方法，将其加入到树结构中，从而得到最优解，其计算公式[6]如下

$$\mathcal{L}_{split}=\frac{1}{2}\left[\frac{\left(\sum_{i \in I_L} g_i\right)^2}{\sum_{i \in I_L} h_i+\lambda}+\frac{\left(\sum_{i \in I_R} g_i\right)^2}{\sum_{i \in I_R} h_i+\lambda}-\frac{\left(\sum_{i \in I} g_i\right)^2}{\sum_{i \in I} h_i+\lambda}\right]-\gamma \tag{8}$$

其中 $I_j=\{i|q\left(x_i\right)=j\}$ 为叶节点 j 上的样本集合[5]，且有

$$g_i=\partial_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right) \tag{9}$$

$$h_i=\partial_{\hat{y}^{(t-1)}}^2 l\left(y_i, \hat{y}_i^{(t-1)}\right) \tag{10}$$

通过上述分析，我们可以得到 XGBoost 算法简图，如图 7所示。

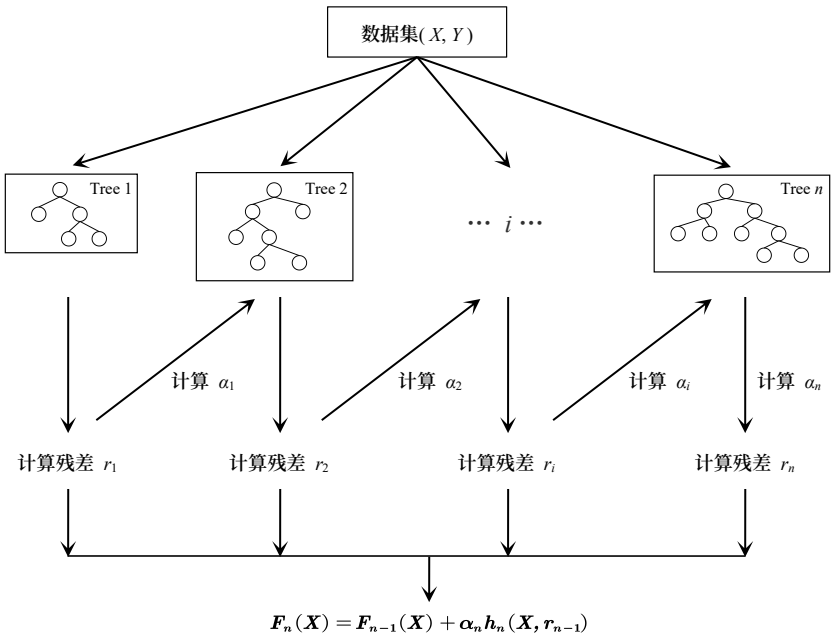


图 7　XGBoost 算法简图

- **K-近邻（K Nearest Neighbor, KNN）**。KNN 算法的主要思想为：在出现新样本时从现有的训练数据中找到与其相对应的最接近的 K 个样本，并根据最相似的类别出现的样本进行分类。基于多数 K 个样本所属的类别来分辨待分类的数据集所属的类别[7]。接近度由两点之间的距离函数给出属性空间中的点决定。距离函数通常使用两个点之

间的标准欧几里得距离。欧氏距离的计算公式如下

$$d(X,Y)=\sqrt{\sum_{i=1}^{n}(X_i-Y_i)^2} \tag{11}$$

其中 $X=(x_1,x_2,\cdots,x_m)^{\mathrm{T}}$ 和 $Y=(y_1,y_2,\cdots,y_m)^{\mathrm{T}}$ 表示两个样本列数据，m 为样本数量。

- **支持向量机（Support Vector Machine, SVM）**。SVM 建立在结构风险最小原理及 Vapnik-Chervonenkis 理论基础之上[8]，以有限的数据信息，在数据样本中找出合适区分类别的决策分界面，且保证边界点与分界面尽可能远，即需要再找出合适的边界分界面，该算法示意图如图 8所示。而由于 SVM 多应用于解决二分类问题，且我们需要建立

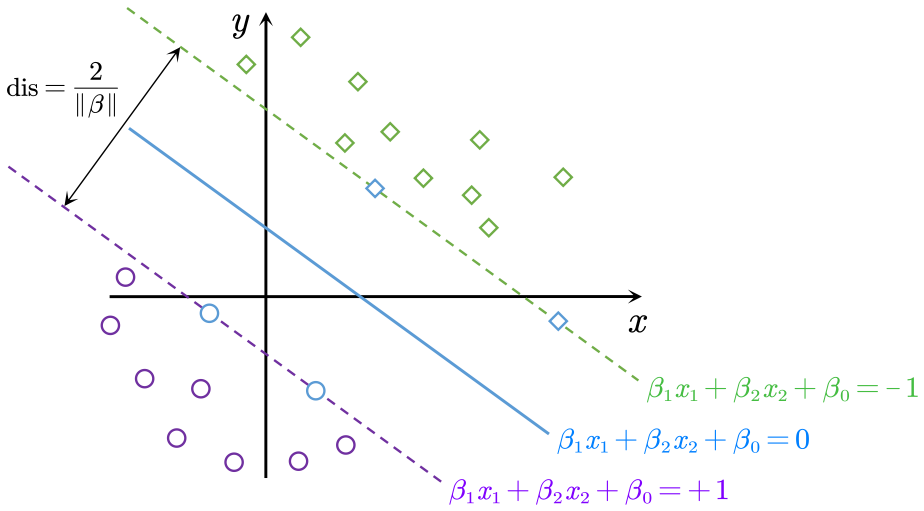


图 8　SVM 示意图

多分类模型，因此需要对其进行相应的改进。本文采用 OVR（One Versus Rest）方法，将该问题改进为多个二分类问题[8]。在模型的训练时，任意将某一类别记为一类，其余类别记为另一类别，依次下去，建立出多分类的 SVM 模型。而对于核函数的选择，本文选择高斯核函数进行求解，其定义公式如下

$$K\left(x_i,x_j\right)=\exp\left(-\frac{\left\|x_i-x_j\right\|^2}{2\sigma^2}\right)=\exp\left(-\gamma\left\|x_i-x_j\right\|^2\right) \tag{12}$$

对于高斯核函数，其可以反映出样本两点之间的相似度大小。当 σ 确定后，若两点之间距离越小，则相似度趋近于 1；若距离越大，则相似度趋近于 0。

- **LightGBM（Light Gradient Boosting Machine）**。LightGBM 模型是基于决策树算法构建的一种高效的机器学习算法[9]。其为 XGBoost、直方图算法（Histogram）、基于梯度的单边采样（GOSS）算法以及互斥特征捆绑（EFB）算法的结合的一种算法。

- **多分类逻辑回归（Multinomial Logistic Regression）**。多分类逻辑回归是基于逻辑回归（Logistic Regression）进行学习的分类模型。对于逻辑回归模型，其属于分类模

型，多用于二分类问题。若数据集为 $(\boldsymbol{A},\boldsymbol{B})=((\boldsymbol{a}_1,b_1),(\boldsymbol{a}_2,b_2),\cdots,(\boldsymbol{a}_m,b_m))^{\mathrm{T}}$，其中 $\boldsymbol{a}_i=\left(a_i^1,a_i^2,\cdots,a_i^j\right)$，$a_i^j$ 为样本 $\boldsymbol{a}_i$ 的第 j 个特征，$\boldsymbol{B}$ 为因变量标签矩阵，该模型使用 Sigmoid 函数，同时构建样本 $\boldsymbol{a}_i$ 所属类别的概率，对于标签为 1 的结果，其概率可写为

$$P\left(b_i=1\mid\boldsymbol{a}_i,\boldsymbol{\omega}\right)=\frac{1}{1+\mathrm{e}^{-\boldsymbol{a}_i b_i\boldsymbol{\omega}^{\mathrm{T}}}} \tag{13}$$

其中 $\boldsymbol{\omega}=(\omega^0,\omega^1,\cdots,\omega^n)^{\mathrm{T}}$ 为权重向量，即为优化模型的超参数。逻辑回归中利用损失函数来评估模型的预测结果与实际值之间的误差，其计算公式如下

$$L\left(\boldsymbol{A},\boldsymbol{B},\boldsymbol{\omega}\right)=\frac{1}{m}\sum_{i=1}^{m}\log\left(1+\mathrm{e}^{-\boldsymbol{a}_i b_i\boldsymbol{\omega}^{\mathrm{T}}}\right) \tag{14}$$

而对于 $\boldsymbol{\omega}$，常采用梯度下降法来获得模型参数的最优解，其通过

$$\boldsymbol{\omega}^{\alpha+1}=\boldsymbol{\omega}^{\alpha}-\frac{\gamma}{m}\sum_{i=1}^{m}\left(\frac{1}{1+\mathrm{e}^{-\boldsymbol{a}_i b_i\boldsymbol{\omega}^{\mathrm{T}}}}-1\right)\boldsymbol{a}_i b_i \tag{15}$$

进行迭代更新，其中 γ 为模型的学习率，当 $|\boldsymbol{\omega}^{\alpha}-\boldsymbol{\omega}^{\alpha+1}|<\eta$ 或达到最大迭代次数时，停止训练，输出最终模型，其中 η 为人为给定的阈值[10]。

建立好上述六种多分类模型后，我们依据各模型的预测准确率、平均绝对误差、均方误差，建立 Stacking 集成学习，将上述模型有目的地进行合理组合，从各模型中学到优点，有利于模型的效果的提升。其基本过程为，首先将已经经过处理的原数据集划分成若干个子集数据，在第一层建立多个模型的融合模型，输入数据，并采用五折交叉验证，获得每个模型的对于因变量标签的预测结果；之后第一层的输出结果作为第二层较弱分类模型的输入数据，第二层单个模型进行训练学习，得到最终预测结果[11]。算法示意图如图 9所示，算法伪代码如Algorithm2所示。

Algorithm 2: Stacking 集成学习

Input: 训练集 $\mathcal{D}$

第一层学习模型 $\mathcal{F}_1$，$\mathcal{F}_2$，$\cdots$，$\mathcal{F}_n$

第二层学习模型 $\mathcal{S}$

for $t=1,2,3,\cdots,n$ **do**
  $h_n=\mathcal{F}_n\left(\mathcal{D}\right)$
end
$\mathcal{D}'=\varnothing$
for $i=1,2,\cdots,m$ **do**
  for $t=1,2,\cdots,n$ **do**
    $z_{in}=h_n\left(x_i\right)$
  end
  $\mathcal{D}'=\mathcal{D}'\cup\left(\left(z_{i1},z_{i2},\cdots,z_{in}\right),y_i\right)$
end
$h'=\mathcal{S}\left(D'\right)$

Output: $\mathcal{H}\left(x\right)=h'\left(h_1\left(x\right),h_2\left(x\right),\cdots,h_n\left(x\right)\right)$

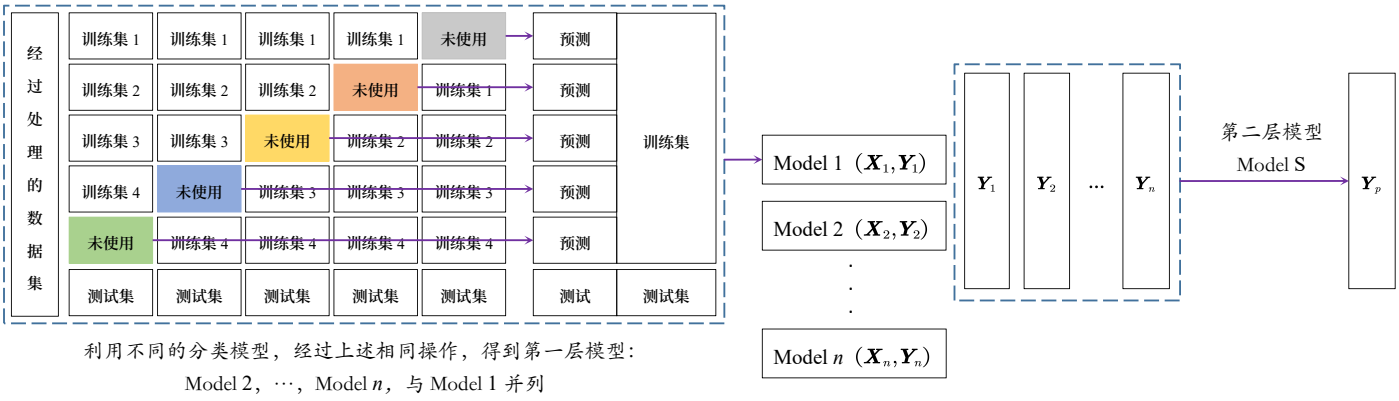


图 9　Stacking 集成学习示意图

5.8 用户评分预测

在“5.1 相关准备工作”中，我们提到对学习数据与预测数据进行一致化，这是为了统一预测的自变量，避免不同的自变量的混乱，导致预测错误。经过对预测数据集的处理，利用上述已建立好的八个模型，对每一位用户的评分进行预测。这里我们需要注意的是：首先，要保证传入模型的变量要与训练时传入的指标一致；其次，在上文中我们提到我们选用多分类解决，而需要对评分进行标签编码，即将原评分 $y \in [1, 10]$ 映射至新评分标签 $y' \in [0, 9]$，即有关系式 $y' = y - 1$，而对于新数据的预测，我们要在模型的每个预测结果上加 1，避免预测结果出错。由于被预测用户过多，我们将不在论文中展示，而将以文件形式保存至“附件 6：result.xlsx”。

5.8.1 模型预测结果合理性分析

本文对于数据集充分分析，多方面考虑，首先对高分组与低分组评分用户分类讨论，分析及研究其主要特征；之后，我们依据整体用户行为进行分析，筛选出评分较为合理的用户群体作为新的数据集，并建立多模型调参融合的 Stacking 集成学习模型，且对模型训练采用五折交叉验证，保证模型的稳健性。对于语音业务数据的处理，我们将数据集划分训练集与测试集，比例为 8:2；对于上网业务数据的处理，我们将数据集划分训练集与测试集，比例为 9:1。对于两项业务这样处理有以下几点原因：

- 由于本题为用户对于移动公司语音及上网业务的评分预测，但该评分选择性较大且主观性强烈，难以以合理的量值确定用户对于该项业务的满意程度，因此仅能从整体用户行为中进行分析，分析出在整体用户中存在的部分”离群点“，即评分存在不合理的用户群体，从而对其进行剔除，在一定程度上保持样本数据的纯性，提升模型对于整体用户的预测准确性；

- 为验证模型的效果、分析模型的合理性、对模型参数进行调优、有监督地在数据上进行学习，更好地分析模型对于重要特征的选择，进行特征选择等，因此我们需要对数据集划分训练集与测试集；

- 对于语音业务，我们划分训练集与测试集比例为 8:2，这是由于我们观察到，语音业务

的数据分布较优，且需要学习的特征相对于上网业务较少，若过分提高该比例，模型可能会产生过拟合的情况，无法对未知数据进行高效分析，泛化能力差；

- 对于上网业务，我们划分训练集与测试集比例为 9:1，这是由于我们观察到，上网业务需要学习的特征较多，若训练集样本过少，可能导致训练的模型发生欠拟合的情况，未能更好地学习到数据的内在规律，导致模型的多项指标为达到期望值。

此外，考虑到用户评分的主观性及数据分布，我们选择多分类模型解决，同时为更好地学习、预测，我们建立多个分类模型，且对各模型进行超参数的调节，在一定程度上提高模型的预测精度，分析各个影响因素的特征重要性。此外利用 Stacking，对多模型进行集成学习，使得最终模型可以学习到各个模型的特性，且在一定程度上提升模型的泛化能力。

5.8.2 初赛模型与复赛模型的比较

在这里本文再次提及多分类模型的准确率（Accuracy）、平均绝对误差（Mean Absolute Error，MAE）、均方误差（Mean Square Error, MSE）指标计算方法，计算公式如下：

- **准确率**

$$\text{Accuracy} = \frac{N_{\text{TruePredict}}}{N_{\text{Sample}}} \tag{16}$$

其中，$N_{\text{TruePredict}}$ 为预测正确的样本数，N_{Sample} 为被预测的样本总数；

- **平均绝对误差**

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{17}$$

其中，y_i 为实际值，$\hat{y}_i$ 为预测值；

- **均方误差**

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{18}$$

在初赛中，我们假设”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“，但并未对这些乱评、错评的用户进行深层次分析，在一定程度上可以确保模型的泛化能力，但可能会导致在预测数据中忽略整个用户群体的真实性，初赛最终模型的五折交叉验证平均准确率、平均绝对误差、均方误差如表 5所示。而在本文中，我们对不合理评分的用户进行剔除，再由此建立模型，并进行预测，其最终预测模型的相关指标如表 6所示。

通过分析表 5与表 6，我们可以发现，在本次进行采样的新数据集上建立的模型在五折交叉验证平均准确率上较原模型稍有下降，这可能是由于忽略少样本个性的影响，在一定程度上使得模型泛化能力下降，但模型的平均绝对误差、均方误差均有大幅度优化（即数值上减少），优化结果如表 7所示。

表 5　初赛中各评分预测模型效果

模型	五折交叉验证平均准确率	平均绝对误差	均方误差
模型一 [预测语音业务，语音通话整体满意度]	0.5773	1.2937	6.3877
模型二 [预测语音业务，网络覆盖与信号强度]	0.4880	1.5387	7.3416
模型三 [预测语音业务，语音通话清晰度]	0.5405	1.3527	6.4540
模型四 [预测语音业务，语音通话稳定性]	0.5212	1.3913	6.3748
模型五 [预测上网业务，手机上网整体满意度]	0.4359	1.7094	8.0684
模型六 [预测上网业务，网络覆盖与信号强度]	0.3803	1.7650	7.7764
模型七 [预测上网业务，手机上网速度]	0.3761	1.7208	7.3134
模型八 [预测上网业务，手机上网稳定性]	0.3875	1.8276	8.0897

表 6　复赛中各评分预测模型效果

模型	五折交叉验证平均准确率	平均绝对误差	均方误差
模型一 [预测语音业务，语音通话整体满意度]	0.5757	1.1722	5.3057
模型二 [预测语音业务，网络覆盖与信号强度]	0.4845	1.4610	6.5880
模型三 [预测语音业务，语音通话清晰度]	0.5334	1.2578	5.6284
模型四 [预测语音业务，语音通话稳定性]	0.5202	1.3584	6.1806
模型五 [预测上网业务，手机上网整体满意度]	0.4355	1.6320	7.3827
模型六 [预测上网业务，网络覆盖与信号强度]	0.3959	1.6979	7.1672
模型七 [预测上网业务，手机上网速度]	0.4091	1.6965	7.1393
模型八 [预测上网业务，手机上网稳定性]	0.4120	1.5997	6.6290

表 7　复赛模型较初赛模型优化结果

模型	五折交叉验证平均准确率变化率	平均绝对误差优化	均方误差优化
一	0.0027	9.40 %	16.94 %
二	0.0072	5.05 %	10.27 %
三	0.0131	7.02 %	12.79 %
四	0.0019	2.36 %	3.05 %
五	0.0010	4.53 %	8.50 %
六	0.0410	3.80 %	7.83 %
七	0.0877	1.41 %	2.38 %
八	0.0633	12.47 %	18.06 %

通过分析表 7，我们可以发现优化效果良好，对于整体用户的评分把握程度大幅度提升。

为了更好地评估模型，对预测结果的合理性进行分析，我们绘制出各个模型的**混淆矩阵热力图、分类报告、ROC/AUC 曲线**。这里由于篇幅原因，我们仅展示“预测语音业务-网络覆盖与信号强度”三幅模型效果可视化图形，其余模型的分析与其一致。对于其余模型的可视化图形，读者可在附录中查看。其余七个模型的混淆矩阵热力图见图 10~??；分类报告见图 11~??；ROC/AUC 曲线见图 12~??。

- **混淆矩阵热力图**。该可视化图形的每一行表示样本标签的实际类别，在本题中表示用户评分的实际值[1]，而每一行表示样本标签的预测类别，在本题中表示用户评分的预测值。因此该图示的主对角线数据之和即为模型预测准确的样本数。对于多分类模型，我们可以随机指定一类为正类，而其余就为对应的负类。这里我们需要引入四项值，分别为 TP、FN、FP、TN，其中 T 为 True，F 为 False，这两个字母表示预测值与实际值是否相同；P 为 Positive，N 为 Negative，这两个字母表示预测出的是属于正类（阳性）还是负类（阴性）。而混淆矩阵热力图即为这些值组成，该图示可以直观地观察到

[1] 上文中提到我们对用户评分进行标签编码，从原来的 $[1,10]$ 映射至新评分标签 $[0,9]$，即在原评分基础上减 1，而在混淆矩阵热力图及分类报告图示中我们将标签编码已映射回原评分，对于模型的 ROC/AUC 曲线，我们未映射回原评分。

预测准确与错误的情况，以及模型对于每一类别的区分程度。模型二的混淆矩阵热力图见图 10。

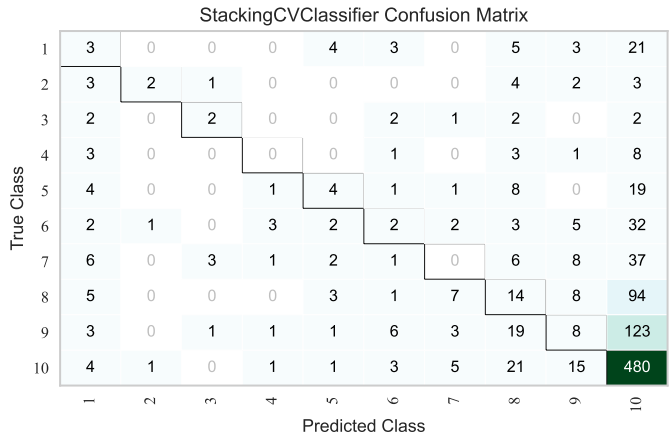


图 10　模型二混淆矩阵热力图 [语音业务-网络覆盖与信号强度]

观察该图，我们可以发现，该模型对于预测用户评分具有较好的效果，主对角线附近元素较多，说明模型预测正确的误差较小，预测得分与用户实际评分比较接近，可以较好预测用户评分。

- **分类报告**。分类报告图示可以直观得到模型各项参数，包括每一类别的精确率（Precision），召回率（Recall），F1 分数值（F1-Score）。对于这三项值，其计算公式如下：

  – **精确率**

$$\text{Precision} = \frac{TP}{TP + FP} \tag{19}$$

  – **召回率**

$$\text{Recall} = \frac{TP}{TP + FN} \tag{20}$$

  – **F1 分数值**

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{21}$$

根据上述(19) 式、(20) 式、(21) 式，我们可以计算出每一个模型对于每一类别的三项指标值，并绘制分类报告图，对于模型二的分类报告，见图 11。

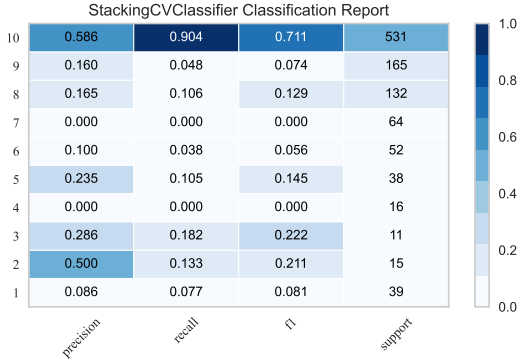


图 11　模型二分类报告 [语音业务-网络覆盖与信号强度]

对于模型的精确率、召回率，我们可以根据定义发现，这两项值显然较大，模型效果较好。同时根据定义，我们可以发现模型的精确率、召回率在理想情况下是相差较小的，我们可以根据图表结果验证，符合预期效果。对于模型的 F1 分数值，其为精确率与召回率的调和平均数，因此当精确率与召回率均有较好表现时，F1 分数值会有较优秀表现。我们也可对(21) 式进行一定变换，可以得到

$$\mathrm{F1} = \frac{2}{\frac{1}{\mathrm{Precision}} + \frac{1}{\mathrm{Recall}}} \tag{22}$$

根据该式，我们可以得出上述结论。

- **ROC/AUC 曲线**。在分析特征曲线及曲线下面积（Receiver Operating Characteristic/Area Under the Curve, ROC/AUC）图之前，我们需要了解模型的相关参数，定义如下：
  - **灵敏度（Sensitivity）**。灵敏度又被称为真阳性率，即 TP 率，定义为：

$$\mathrm{Sensitivity} = TPR = \frac{TP}{TP + FN} \tag{23}$$

  - **特异性（Specificity）**。特异性又被称为真阴性率，即 TN 率，定义为：

$$\mathrm{Specificity} = TNR = \frac{TN}{TN + FP} \tag{24}$$

  - **1-Specificity**。称为假阳性率（False Positive Rate, FPR），定义为：

$$FPR = 1 - \mathrm{Specificity} = \frac{FP}{FP + TN} \tag{25}$$

  - **1-Sensitivity**。称为假阴性率（False Negative Rate, FNR），定义为：

$$FNR = 1 - \mathrm{Sensitivity} = \frac{FN}{FN + TP} \tag{26}$$

FPR 和 FNR 均对数据分布的变化不敏感[?]，因此这两个指标可以用于在不平衡的数据上建立的模型效果的评价。

对于 ROC/AUC 曲线，其以每一类别的 $1-\mathrm{Specificity}$ 即 FPR 为横坐标，以 Sensitivity 即 TPR 为纵坐标，其可体现出模型的灵敏度与特异性之间的关系与差异。因此，该图的理想点位于左上角，即 $FPR = 0$ 且 $TPR = 1$，换言之，当曲线越靠近左上角，模型效果就越优。从而，我们可以得到另一项指标，即曲线下面积（Area Under the Curve, AUC），由上述分析可知，AUC 值越高，模型的整体效果也就越优。对于模型二的 ROC/AUC 曲线，见图 12。

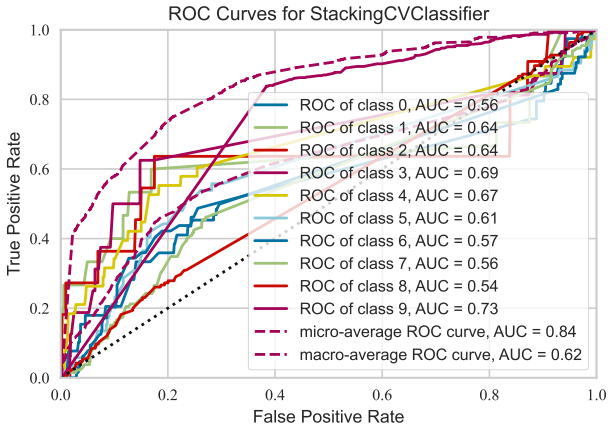


图 12　模型二 ROC/AUC 曲线 [语音业务-网络覆盖与信号强度]

根据上图结果，我们可以发现，模型二预测用户对于“语音业务-语音通话整体满意度”的评分结果中，模型对各评分预测的 AUC $\geqslant 0.50$，即以 $y = x$ 为分界线。同时，我们可以发现“macro-average ROC curve”指标，其是通过 Macro 方法求得，在上文中我们提到，该数据样本的标签分类是严重不平衡的，该方法能够平等对待每一项分类，在此方法下，我们可以对于小样本类别的准确率有一定把握，其曲线下面积 AUC $= 0.62$，位于分界线左上，预测效果良好；而对于“micro-average ROC curve”指标，其 AUC $= 0.84$，这指的是利用 Micro 方法求得曲线，曲线下面积为 0.84，曲线下面积较大，且曲线有向左上角最优点靠近的趋势，可以说明模型整体能力较优。

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。

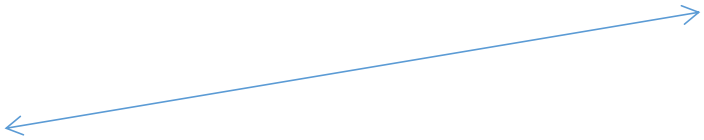
- **模型的改进：**

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。

七、非技术性报告



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.

[3] 饶雷, 冉军, 陶建权, 胡号朋, 吴沁, 熊圣新. 基于随机森林的海上风电机组发电机轴承异常状态监测方法 [J]. 船舶工程,2022,44(S2):27-31.DOI:10.13788/j.cnki.cbgc.2022.S2.06.

[4] 陈振宇, 刘金波, 李晨, 季晓慧, 李大鹏, 黄运豪, 狄方春, 高兴宇, 徐立中. 基于 LSTM 与 XGBoost 组合模型的超短期电力负荷预测 [J]. 电网技术,2020,44(02):614-620.DOI:10.13335/j.1000-3673.pst.2019.1566.

[5] 杨贵军, 徐雪, 赵富强. 基于 XGBoost 算法的用户评分预测模型及应用 [J]. 数据分析与知识发现,2019,3(01):118-126.

[6] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 785–794. https://doi.org/10.1145/2939672.2939785.

[7] 张著英, 黄玉龙, 王翰虎. 一个高效的 KNN 分类算法 [J]. 计算机科学,2008(03):170-172.

[8] 汪海燕, 黎建辉, 杨风雷. 支持向量机理论及算法研究综述 [J]. 计算机应用研究,2014,31(05):1281-1286.

[9] 马晓君, 沙靖岚, 牛雪琪. 基于 LightGBM 算法的 P2P 项目信用评级模型的设计及应用 [J]. 数量经济技术经济研究,2018,35(05):144-160.DOI:10.13653/j.cnki.jqte.20180503.001.

[10] 唐敏, 张宇浩, 邓国强. 高效的非交互式隐私保护逻辑回归模型 [J/OL]. 计算机工程:1-11[2023-01-04].DOI:10.19678/j.issn.1000-3428.0065549.

[11] 史佳琪, 张建华. 基于多模型融合 Stacking 集成学习方式的负荷预测方法 [J]. 中国电机工程学报,2019,39(14):4032-4042.DOI:10.13334/j.0258-8013.pcsee.181510.

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]