队伍编号	MCB220XXXX
赛道	B

论文题目

摘 要

关键词：

目录

1 问题重述

1.1 问题背景

经过信息通信业 30 多年的努力，我国移动通信技术实现了从 1G 空白、2G 跟随、3G 突破、4G 同步到 5G 引领的历史性跨越。随着时代的发展，我国移动通信技术得到了飞速的发展，客户的网络使用体验感也成为了当代各大移动运营商如何提升网络服务质量的首要参考指标。

1.2 问题描述

1) 问题一

- 客户结合自身体验感对网络覆盖与信号强度、 语音通话清晰度，语音通话稳定性和语音通话整体满意度进行打分，并统计了影响客户语音业务体验的因素。本文需要给出各因素对客户打分影响程度的量化分析和结果。
- 客户结合自身体验感对在网络覆盖与信号强度、手机上网速度、 手机上网稳定性和手机上网整体满意度进行打分，并统计了影响客户上网体验的因素。本文需要给出各因素对客户打分影响程度的量化分析和结果。

2) 问题二

- 附件三收集了另外部分客户的影响客户语音业务体验感的因素数据，要求建立相关预测模型，根据影响客户语音业务体验感的因素数据，对网络覆盖与信号强度，语音通话清晰度，语音通话稳定性和语音通话整体满意度进行打分预测并说明本文预测的合理性。
- 附件四收集了另外部分客户的影响客户上网体验感的因素数据，要求建立相关预测模型，对网络覆盖与信号强度、手机上网速度、 手机上网稳定性和手机上网整体满意度进行打分预测并说明本文预测的合理性。

2 问题分析

2.1 问题一分析

问题一，要求我们分别研究影响客户语音业务和上网业务满意度的主要因素，并给出各因素对客户打分影响程度的量化分析和结果。

附件 1 是客户对语音业务结合自身体验感对网络覆盖与信号强度、 语音通话清晰度，语音通话稳定性和语音通话整体满意度的 4 个打分项和影响客户语音业务体验的各因素数据。附件 2 是客户对上网业务结合自身体验感对网络覆盖与信号强度、手机上网速度、 手机上网稳定性和手机上网整体满意度的 4 个打分项和影响客户上网业务体验的各因素数据。由于原始数据存在不规范性，因此首先需要进行数据清洗，包括处理缺失值和异常值。由于附件 1 和附件 2 的部分特征对于本题最终目的预测客户打分无实际意义，本文运用特征工程提前对附件 1 和附件 3，附件 2 和附件 4 进行特征的统一，所以问题一中求特征重要性，并未考虑他们。

在特征工程的处理上，对于数据中的连续值特征，为了消除不同字段量纲对最终模型的影响，需要对数据进行归一化；对于数据中的分类特征，原始数据的值跨越较大且有多种形式，但是只有数字类型才能进行计算，所以需要对各种特殊的特征值进行

相应的编码；同时为了使附件 1 和附件 3，附件 2 和附件 4 的特征统一，需要进行无关特征的删除和特征替换。

在对数据进行缺失值处理，异常值处理，特征工程步骤后，得到了经过处理的样本数据，再使用随机森林模型，分别得到 8 个打分项的特征重要性即可。问题一的思路流程图如图 1 所示。

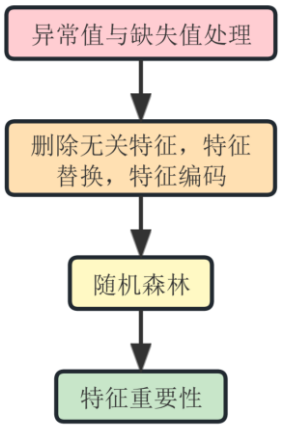


图 1 问题一流程图

2.2 问题二分析

问题 2 需要对附件 3 和附件 4 的 8 个打分项进行预测，附件 1 是用户语音业务的测试集和训练集，附件 3 是其预测集。附件 2 是上网业务的测试集和训练集，附件 4 是其预测集。在问题 1，本文就进行了附件 1 和附件 3，附件 2 和附件 4 的特征统一，由于特征比较多，因此需要对特征进行选择，防止模型的过拟合。我们采取特征嵌入式选择的方法对特征进行选择，得到处理后的数据，本文利用机器学习算法：BP 神经网络，随机森林，XGBoost 进行客户的打分预测，利用 python 进行三个模型的运算，最后为了提高模型的精度，我们利用线性回归进行模型的融合，从而得出更为精准的客户打分。问题二的思路流程图如图 2 所示。

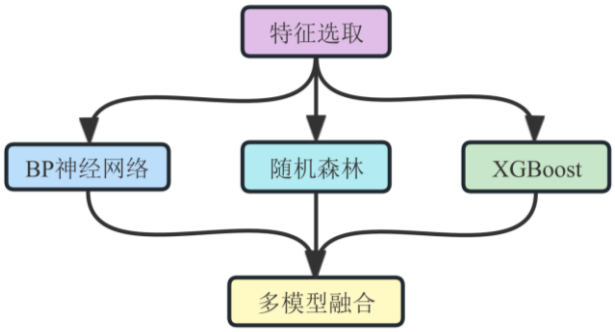


图 2 问题二流程图

3 模型假设

由于附件 1 和附件 2 的部分特征对于本题最终目的预测客户打分无实际意义，本文运用特征工程提前对附件 1 和附件 3，附件 2 和附件 4 进行特征的统一，所以问题一中求特征重要性，并未考虑他们。

4 符号说明

表 1 符号说明

符号	含义
x	原始数据矩阵
ϕ_i	归一化矩阵 ϕ 的第 i 个分量
N	输入层
M	输出层
K	隐藏层
D	输入到决策树的数据集
e_{NSE}	均方误差
$\hat{y}_i$	XGBoost 模型的预测值
f_k	表示第 k 棵树模型

5 数据的预处理

5.1 数据清理

5.1.1 数据类型统计

导入题目所给的附件 1 和附件 2，运用 python 读取附件 1 和附件 2 中的数据整体信息，如表 2 和表 2 所示。

表 2 附件 1 中原始数据字段与数据类型

字段名	数据类型
用户 id	int64
语音通话整体满意度	int64
网络覆盖与信号强度	int64
语音通话清晰度	int64
语音通话稳定性	int64

表 3 附件 2 中原始数据字段与数据类型

字段名	数据类型
用户	object
手机上网整体满意度	int64
网络覆盖与信号强度	int64
手机上网速度	int64
手机上网稳定性	int64

由于数据过多，详细数据请见支撑材料：1.xlsx。

由表 2 和表 3 可知，附件 1 中：原始数据共 5433 条数据，55 个特征组成，数据类型分为 3 类：float64(11)， int64(32)和 object(12)；附件 2 中：原始数据共 124 个特征组成，数据类型分为 4 类：float64(22)， int64(77)，object(23)和 datetime64[ns](2)。

5.1.2 缺失值处理

附件 1 和 3 中，影响客户语音业务体验的特征数据出现了大量的缺失值；附件 2 和 4 中，影响客户上网体验的特征数据也出现了大量的缺失值。考虑了数据中各字段缺失值的具体情况，本文首先运用 Excel 将缺失值筛选出来，进行数据可视化，如图 1，2 所示。

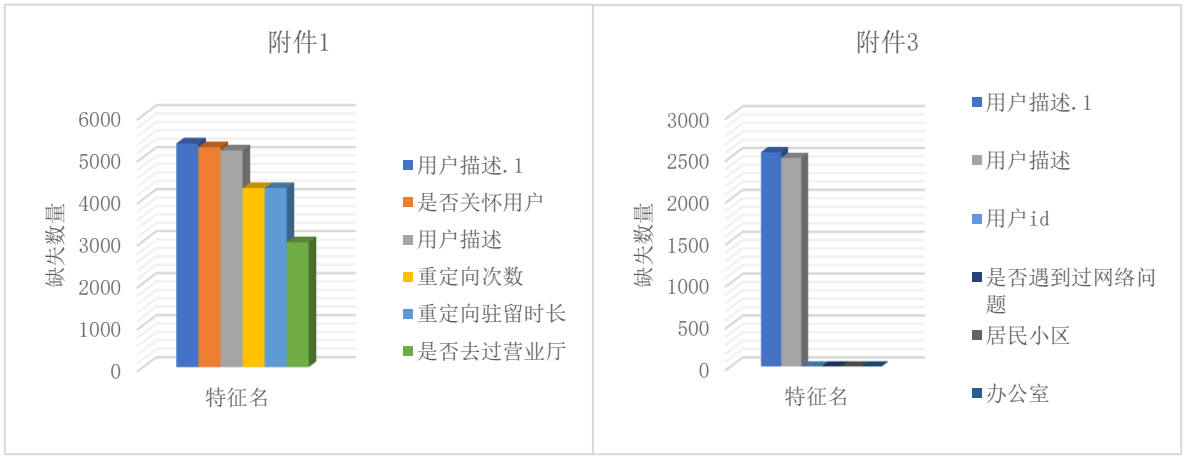


图 2 附件 1，3 的缺失值情况(从左到右)

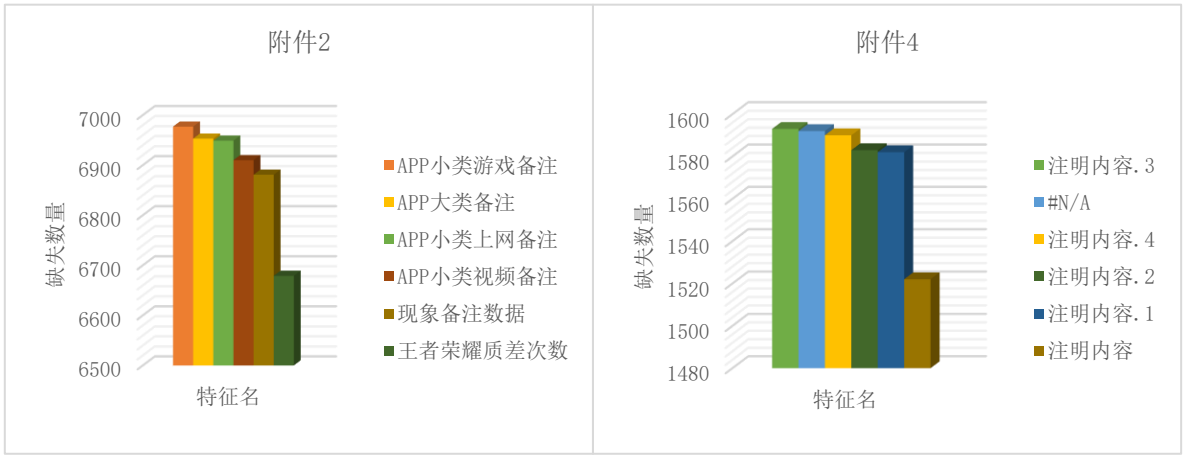


图 3 附件 2，4 的缺失值情况(从左到右)

由于数据过多，具体数据请见支撑材料：2.xlsx，3.xlsx，4.xlsx，5.xlsx。

再根据缺失值的具体情况，以及附件 5 中字段说明所给的特征信息，运用 python 处理相关缺失值：

a) 删除少量缺失值所对应行：

删除附件 1，2，3，4 中“终端品牌”、“终端品牌类型”、“是否 4G 网络客户(本地剔除物联网)”、“外省流量占比”、“是否 5G 网络客户”、“是否实名登记用户”、“客户星级标识”、“当月欠费金额”、“前第 3 个月欠费金额”等特征中少量缺失值所对应的行。

b) 删除大量缺失值所对应列：

附件 1，2，4 中“用户描述.1”，“是否关怀用户”，“场景备注数据”，“现象备注数据”，“APP 大类备注”，“APP 小类视频备注”，“APP 小类游戏备注”，“APP 小类上网备注”，“注明内容.3”，“注明内容.4”等特征出现了大量缺失值，将其删除。

附件 3 中“用户描述”和“用户描述.1”特征出现了大量缺失值，应删掉该特征，并且“其他，请注明”和“其他，请注明.1”两特征与其相关一并删掉。

c) 使用零填充：

根据附件 5 中字段说明可知 “重定向驻留时长”，“重定向次数”，“是否关怀用

户”，“是否去过营业厅”等特征中空值为否的意思，本文运用 python 将附件 1，2，3，4 中相关的空值填为 0。

5.1.3 异常值处理

观察附件 1 可知，影响客户语音业务体验因素的数据中“前 3 月 MOU 异常值”，“GPRS 总流量(KB)”中存在明显的异常值，本文综合考虑下运用绝对中位差(MAD)方法进行异常值处理。

$$MAD = median(|\ X_i - median(X)\ |) \tag{1}$$

假定数据服从正态分布，我们让异常点(outliers)落在两侧的 50% 的面积里，让正常值落在中间的 50% 的区域里：

$$P(|\ X-\mu\ | \leq MAD) = P(\frac{|\ X-\mu\ |}{\sigma} \leq \frac{MAD}{\sigma}) = P(Z \leq \frac{MAD}{\sigma}) = \frac{1}{2} \tag{2}$$

其中 $P(Z \leq \frac{MAD}{\sigma_-}) = \Phi(\frac{MAD}{\sigma}) - \Phi(-\frac{MAD}{\sigma}) = \frac{1}{2}$，又由 $\Phi(-a) = 1 - \Phi(a)$,可 $\Phi(\frac{MAD}{\sigma}) = \frac{3}{4} \Rightarrow$ $\frac{MAD}{\sigma} = \Phi^{-1}(\frac{3}{4})$，查表可知，$\frac{MAD}{\sigma} = 0.6749$。

本文利用 python 运行 MAD 得到前 3 月 MOU 字段箱形图， GPRS 总流量(KB)字段箱形图，前 3 月 MOU 字段去除异常点箱形图和 GPRS 总流量(KB)字段去除异常点箱形图，如图 4，5 所示.

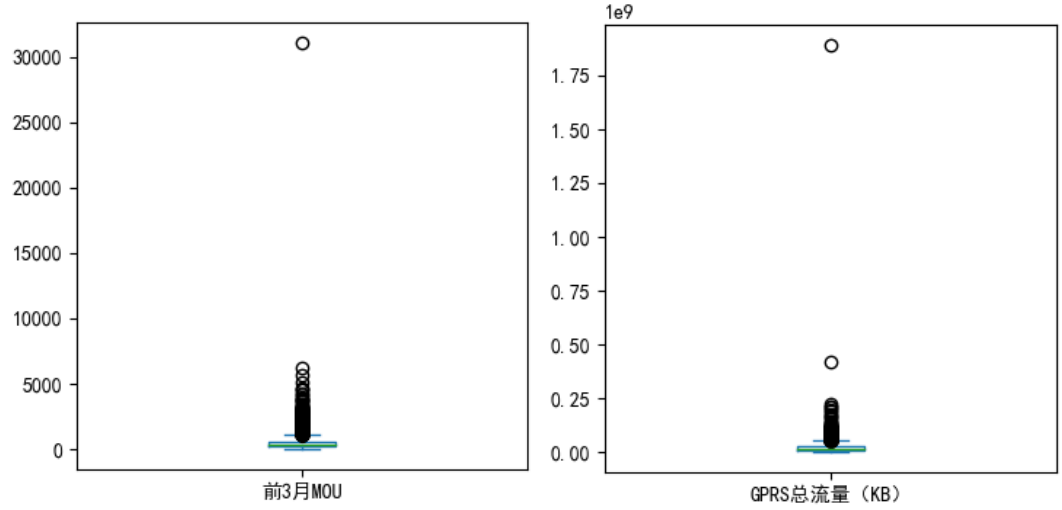


图 4 前 3 月 MOU 字段箱形图(左)和 GPRS 总流量(KB)字段箱形图(右)

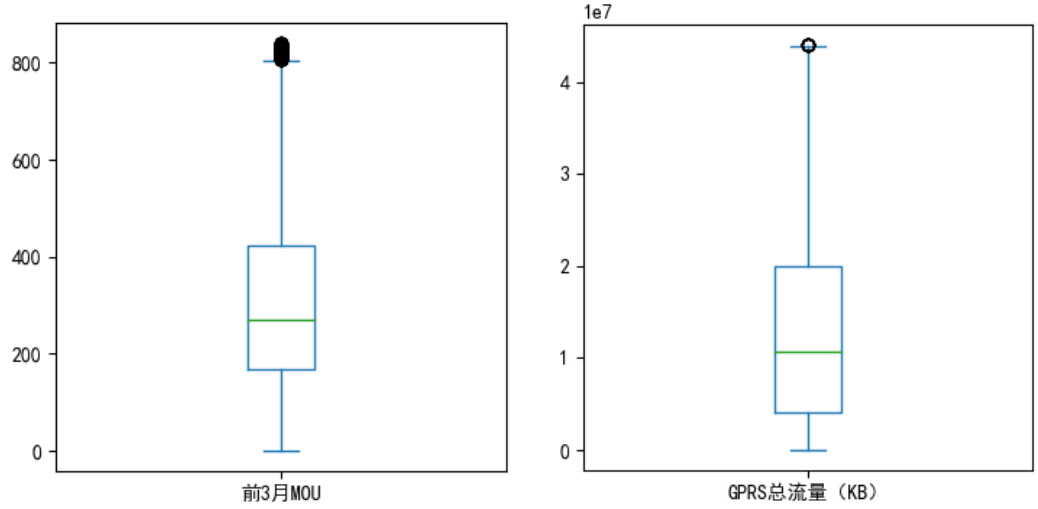
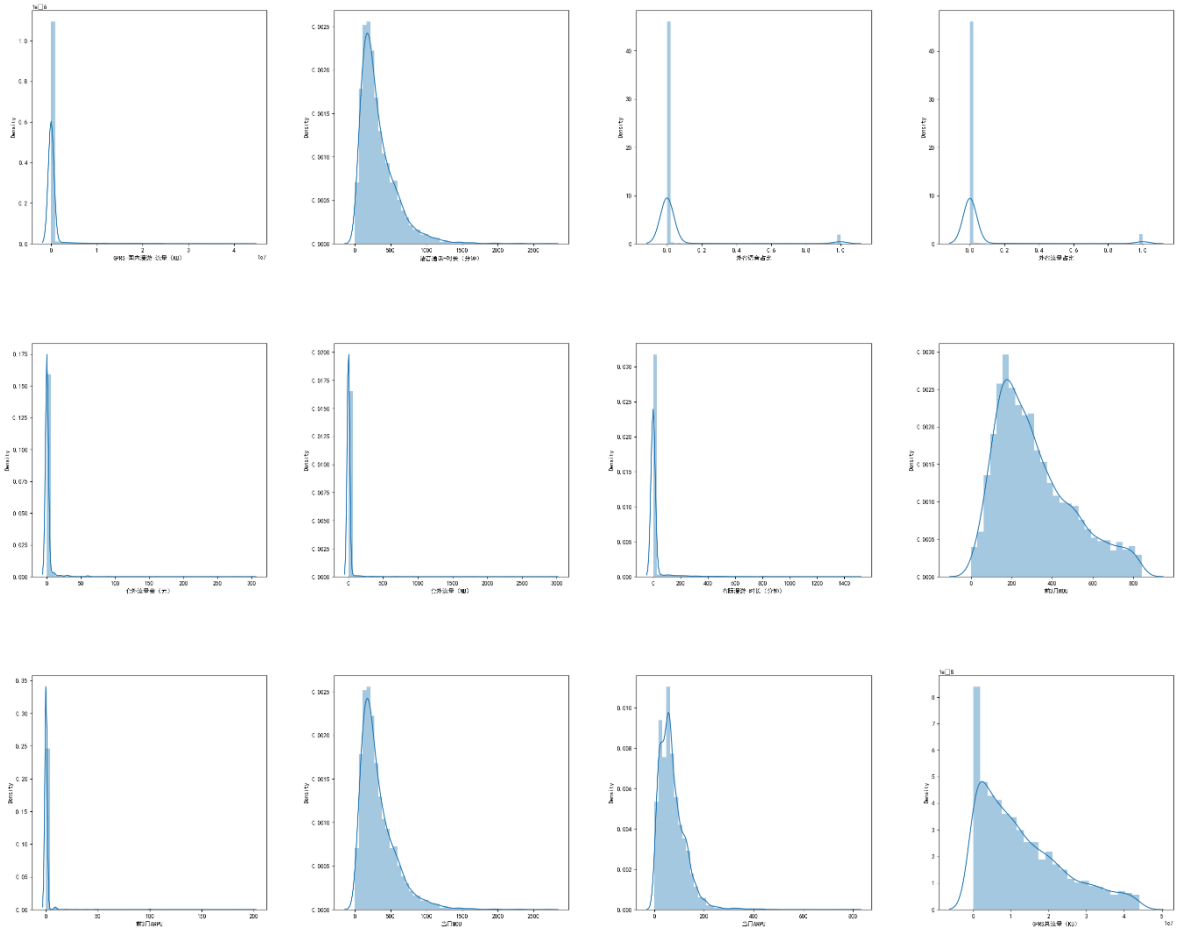


图 5 前 3 月 MOU 字段去除异常点箱形图(左)和 GPRS 总流量(KB)字段去除异常点箱形图(右)

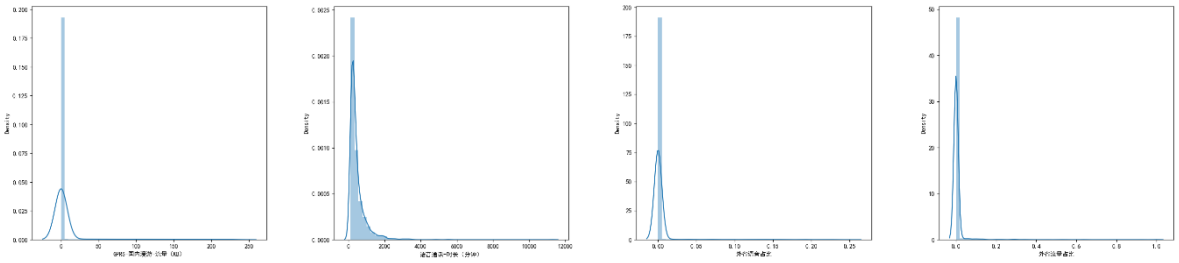
由图 4 得知，“前 3 月 MOU”中显示的异常值达到 31066，“GPRS 总流量(KB)”中显示的异常值达到 1892486521。由图 5 得知，经过*MAD*处理后“前 3 月 MOU”中可以观察到数据分布恢复正常，观察发现处理后数据减少为 4396 条，处理后“GPRS 总流量(KB)”可以观察到数据分布恢复正常，观察发现处理后数据减少为 4966 条。

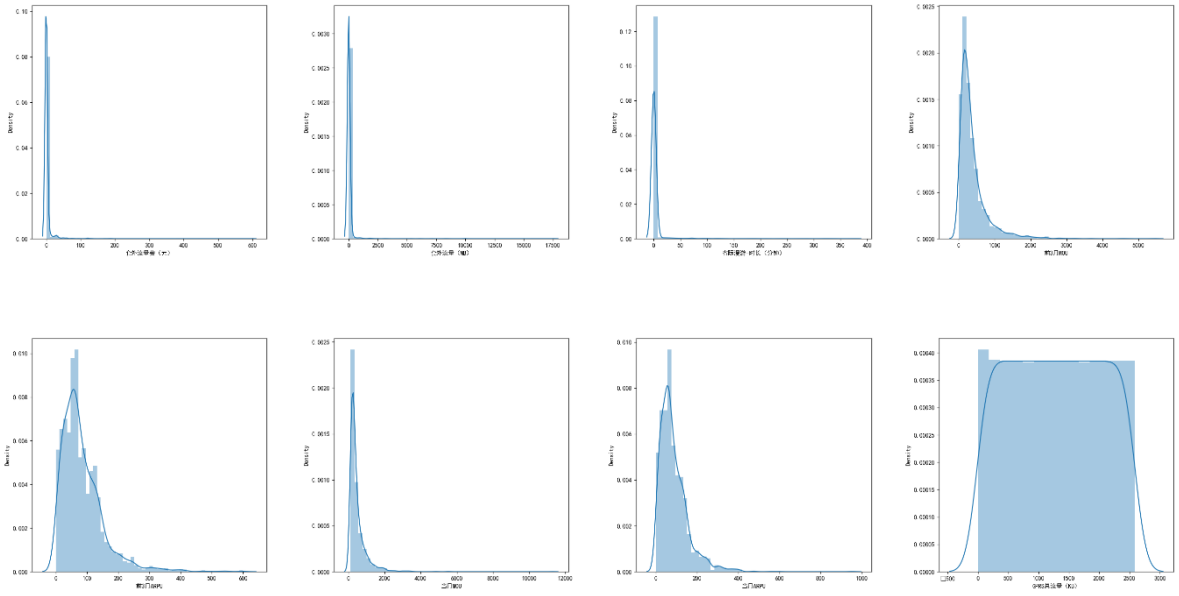
5.2 特征工程

首先对训练集和预测集中的数据分布情况进行查看，查看是否需要对数据进行大量剔除或转换。附件 1 为客户语音业务打分的训练集和测试集，附件 3 为客户语音业务打分的预测集，附件 2 为客户上网业务的训练集和测试集，附件 4 为客户上网业务的预测集，本文运用 python 分别绘制测试集和预测集的概率密度图，如图所示。

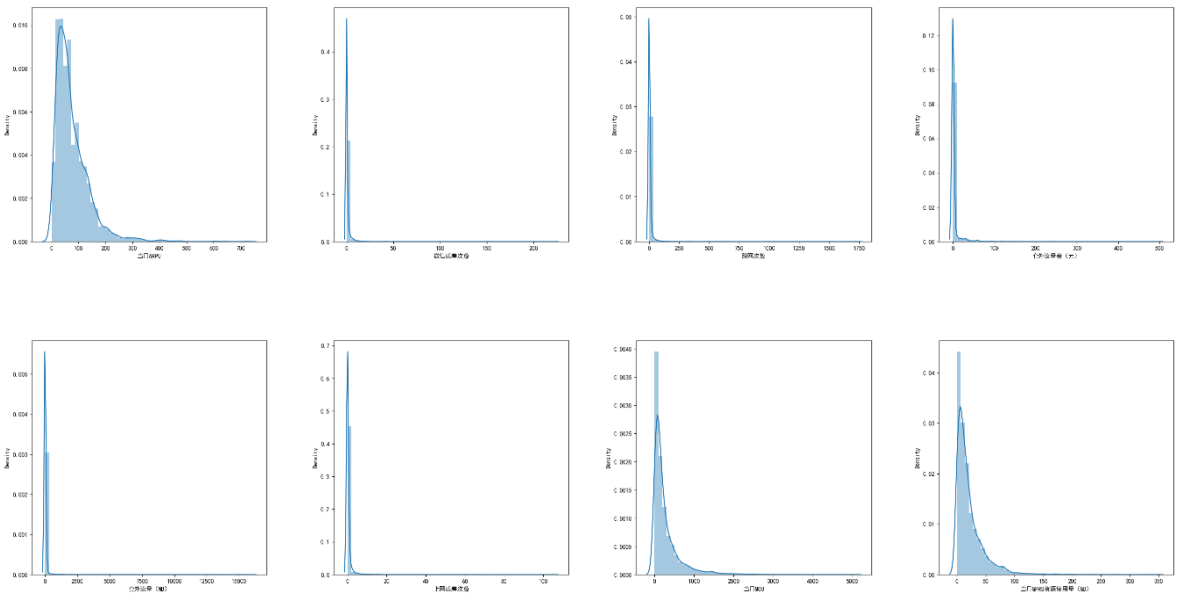


图附件 1 训练集

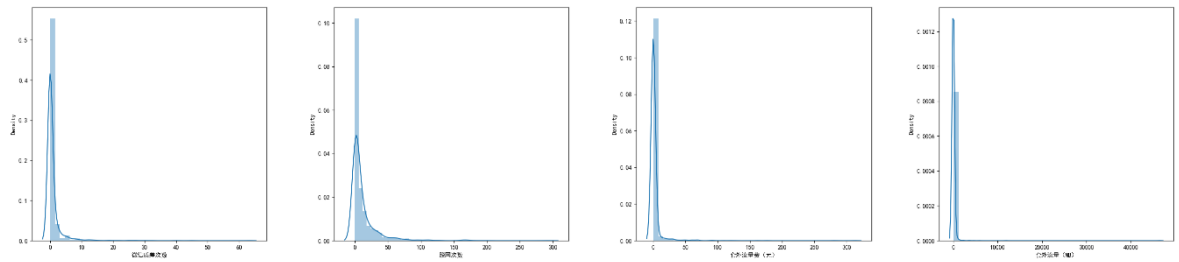


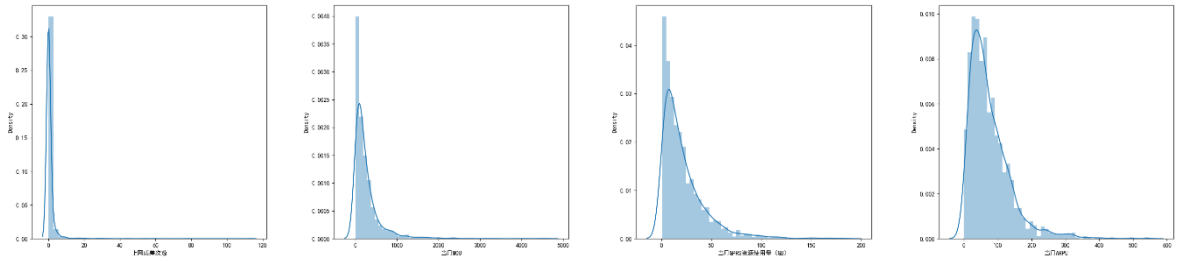


图附件 3 预测集



图附件 2 训练集





图附件 4 预测集

5.2.1 归一化

对于附件 1，附件 2，附件 3 和附件 4 数据中的连续值特征，为了消除不同字段量纲对最终模型的影响，我们需要对数据进行归一化。

设原始数据矩阵：$x = [x_1 x_2 \cdots x_n]$, $x_{\max} = \max\{x\}$，$x_{\min} = \min\{x\}$

则对原始数据按下式进行归一化处理：

$$\phi_i = 2\frac{x_i - x_{\min}}{x_{\max} - x_{\min}} - 1 \tag{3}$$

其中，x_i 为原始数据矩阵 x 的第 i 个分量，ϕ_i 为归一化矩阵 ϕ 的第 i 个分量即：

$$\phi = [\phi_1 \phi_2 \cdots \phi_n] \tag{4}$$

5.2.2 删除无关特征

由于附件 1 和附件 2 的部分特征对于本题最终目的预测客户打分无实际意义，本文进行删除无关特征和特征替换提前对附件 1 和附件 3，附件 2 和附件 4 进行特征的统一，所以问题一中求特征重要性，并未考虑他们。

1) 附件 1 和附件 3 中无关数据的删除。

通过查询相关资料，参考附件 5 中的文字说明，对比附件 1 和附件 3 的特征。

对于附件 1，“重定向次数”,“重定向驻留时长”,“资费投诉”,“语音方式”,“是否去过营业厅”,“ARPU(家庭宽带)”,“是否实名登记用户”,“当月欠费金额”,“前第 3 个月欠费金额”和“家宽投诉”这 10 个特征与附件 3 无关，本文将其删除。

对于附件 3，“是否投诉”,“性别”,“是否不限套餐到达用户”和“用户 id”这 4 个特征与附件 1 无关，本文将其删除。

2) 附件 2 和附件 4 中无关数据的删除。

对于附件 2，“重定向次数”,“2G 驻留时长”,“王者荣耀质差次数”,“高单价超套客户(集团)”,“高频高额超套客户(集团)”,“是否全月漫游用户”和“年龄”这 7 个特征是与附件 4 无关，本文将其删除。

5.2.3 特征替换

附件 2 为测试集与训练集，附件 4 为预测集。通过观察附件 4 及附件 2 中特征的实际意义，进行特征的数据分析，运用 python 得知，附件 2 中“近 3 个月平均消费(剔除通信账户支付)”,“近 3 个月平均消费”与固件 4 中“当月 ARPU”的数值情况都比较类似，如表所示。

表：“近 3 个月平均消费(剔除通信账户支付)”,“近 3 个月平均消费”,“当月 ARPU”的数值情况

特征名	平均值	均方差

当月 ARPU	75.923	62.056
近 3 个月平均消费	76.816	63.624
近 3 个月平均消费(元)	76.826	63.792

综合考虑 ARPU 的实际意义: ARPU 即 Average Revenue Per User，指的是一个时期内(通常为一个月或一年)电信运营企业平均每个用户贡献的通信业务收入，其单位为元/户。从计算的角度看，ARPU 值的大小取决于两个因素，业务收入和用户数量，相对用户数量，业务收入越高，ARPU 值越大。同时 ARPU 值也反映企业的用户结构状况，当用户构成中高端客户占的比重越高，ARPU 值就越高。本文觉得通信账户不应该剔除，同时“近 3 个月平均消费”与“当月 ARPU”的均值更为接近，所以本文将附件 2 中“近 3 个月平均消费(元)”视为附件 4 中的特征“当月 ARPU”。

观察附件 2 和附件 4，还得知附件 2 中“GPRS 总流量(KB)”和附件 4 中“当月 GPRS 资源使用量(GB)”可以相互转化，表展示了它们的具体数值情况。

表 7：“GPRS 总流量(KB)”，“当月 GPRS 资源使用量(GB)”的数值情况

特征名	平均值	均方差
GPRS 总流量(KB)	22.099	22.684
当月 GPRS 资源使用量(GB)	21.098	25.220

由表 7 本文综合考虑，决定将附件 2 中的“GPRS 总流量(KB)”可以做单位换算将 KB 换算为 GB，除以一年 12 个月，换算后可视为附件 4 中的特征“当月 GPRS 资源使用量(GB)”。

5.2.4 特征编码

对于数据中的分类特征，附件 1，2，3 和 4 中的数据的值跨越较大且有多种形式，因为只有数字类型才能进行计算。因此，对于各种特殊的特征值，我们都需要对其进行相应的编码。在这里我们采用 *Label encoding*，将原始特征值编码为自定义的数字标签完成量化编码过程。

6 问题一的求解

6.1 特征重要性

附件 1 给出“网络覆盖与信号强度”，“语音通话清晰度”，“语音通话稳定性”和“语音通话整体满意度”4 个客户语音业务体验的打分项，和影响客户语音业务体验的各因素，附件 2 给出“网络覆盖与信号强度”，“手机上网速度”，“手机上网稳定性”和“手机上网整体满意度”4 个客户上网体验的打分项，和影响客户上网体验的各因素。题目要求根据附件 1 和 2，分别研究影响客户语音业务和上网业务满意度的主要因素，并给出各因素对客户打分影响程度的量化分析和结果。

由于附件 1 和附件 2 的部分特征对于本题最终目的预测客户打分无实际意义，本文运用特征工程提前对附件 1 和附件 3，附件 2 和附件 4 进行特征的统一，所以问题一中求特征重要性，并未考虑他们。

特征工程中已经进行过类别特征编码，现在所有的特征的数字都具有大小关系，而不仅仅是标记作用。因为 8 个目标编码“网络覆盖与信号强度”，“语音通话清晰度”，“语音通话稳定性”，“语音通话整体满意度”， “网络覆盖与信号强度”，“手机上网速度”，“手机上网稳定性”和“手机上网整体满意度” 都是数值型变量，本文运用随机森林模型，通过 python 得出了 8 个目标编码的特征重要性，如图 5 和图 6 所示。

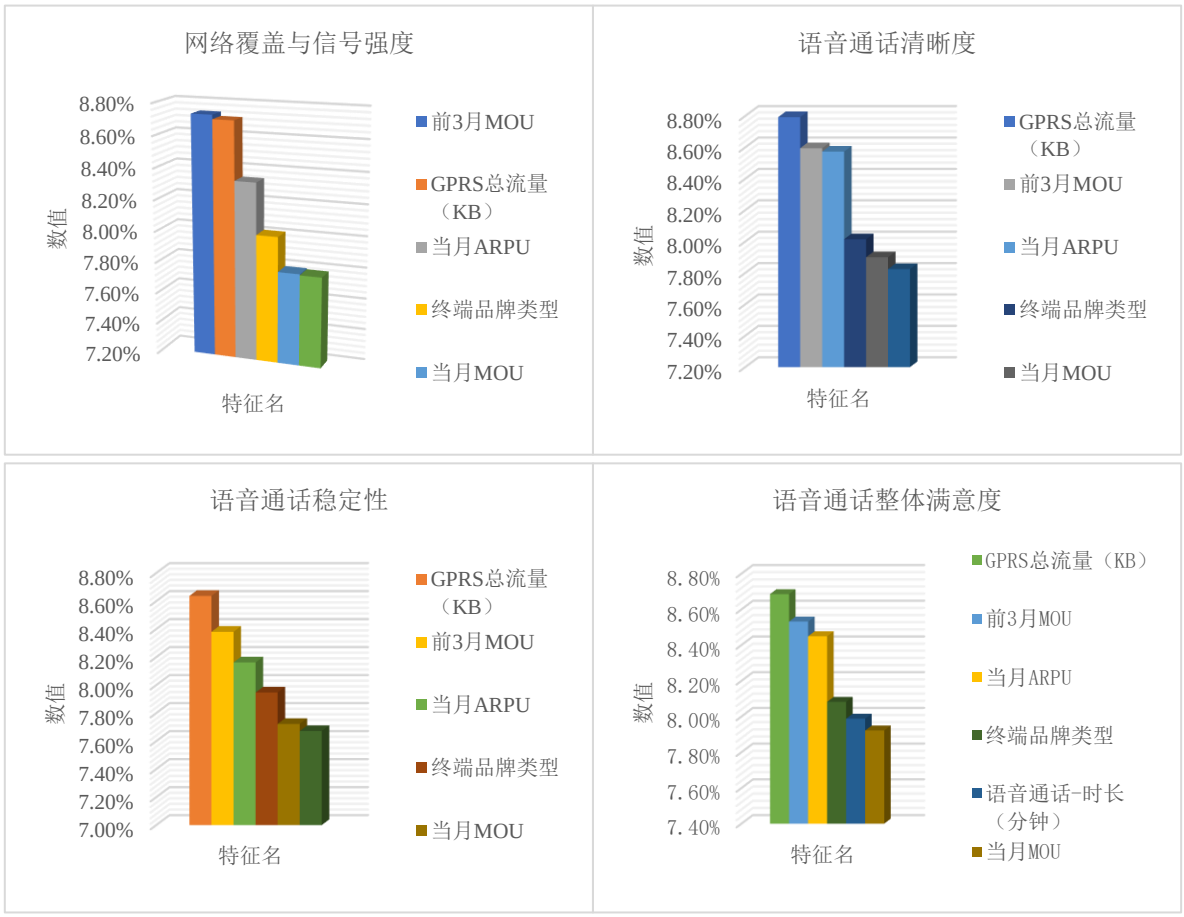


图 8 语音业务体验的特征重要性

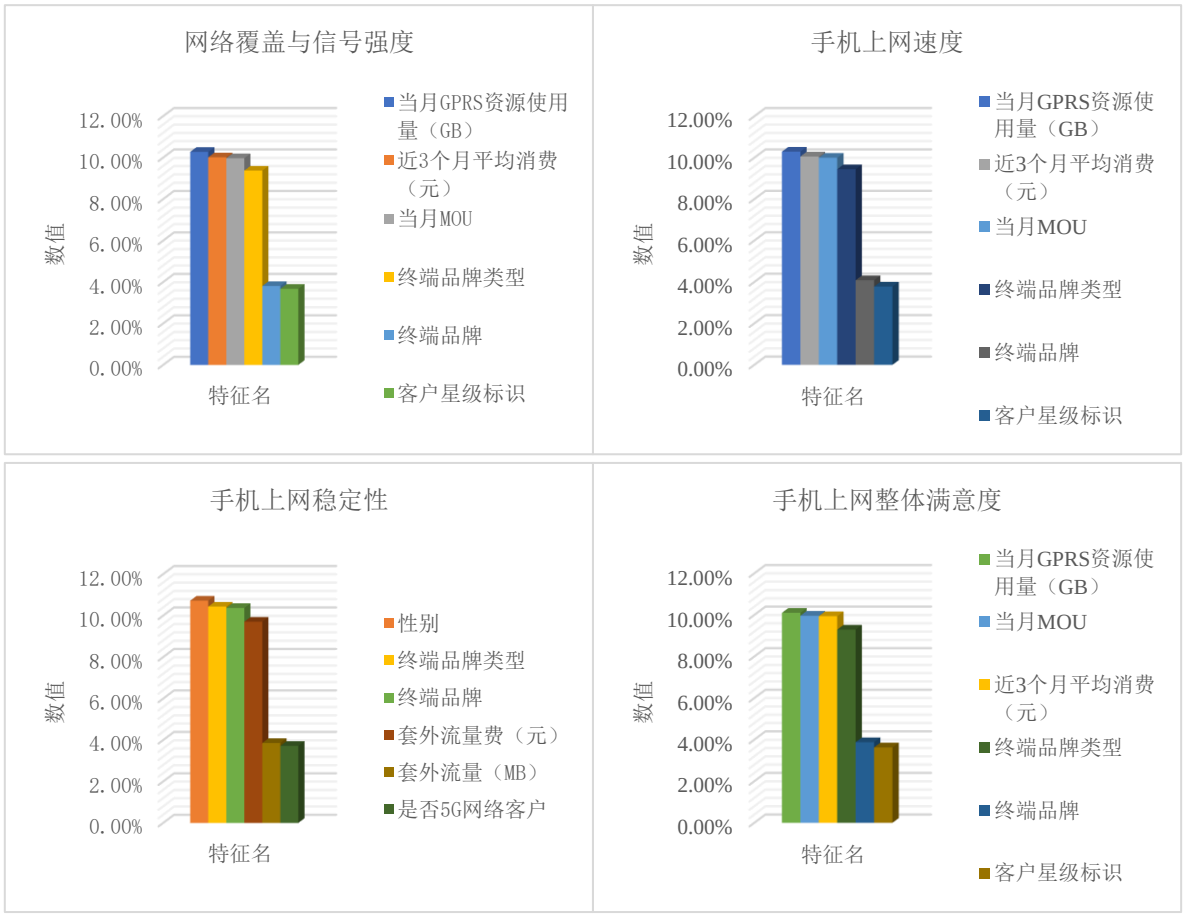


图 9 上网体验的特征重要性

由于数据较多，详细数据请见支撑材料中：6.xlsx，7.xlsx，8.xlsx，9.xlsx，10.xlsx，11.xlsx，12.xlsx 和 13.xlsx。

由图 8 和图 9，本文得知了客户语音业务和上网业务满意度的主要因素和各因素对客户打分影响程度的量化分析和结果。其中语音业务体验中网络覆盖与信号强度的主要特征是：前 3 月 MOU，GPRS 总流量(KB)，当月 ARPU，占比分别是：8.733%，8.706%，8.337%。语音业务体验中语音通话清晰度的主要特征是：GPRS 总流量(KB)，当月 ARPU，前 3 月 MOU，占比分别是：8.792%，8.594%, 8.572%。语音业务体验中语音通话稳定性的主要特征是：GPRS 总流量(KB)，前 3 月 MOU，当月 ARPU，占比分别是：8.636%，8.380%，8.161%。语音业务体验中语音通话整体满意度的主要特征是：GPRS 总流量(KB)，前 3 月 MOU，当月 ARPU，占比分别是：8.685%，8.531%, 8.449%。上网体验中网络覆盖与信号强度的主要特征是：当月 GPRS 资源使用量(GB)，近 3 个月平均消费(元)，当月 MOU，占比分别是：10.275%，10.013%，9.971%。上网体验中手机上网速度的主要特征是：当月 GPRS 资源使用量(GB)，近 3 个月平均消费(元)，当月 MOU，占比分别是：10.292%，10.058%，9.996%。上网体验中手机上网稳定性的主要特征是：性别，终端品牌类型，终端品牌，占比分别是：10.719%，10.434%，10.366%。上网体验中手机上网整体满意度的主要特征是：当月 GPRS 资源使用量(GB)，当月 MOU，近 3 个月平均消费(元)，占比分别是：10.101%，9.9586%，9.937%。

7 问题二模型的建立与求解

7.1 特征选取

经过上面随机森林的特征重要性，8 个目标编码：“网络覆盖与信号强度”，“语音通话清晰度”，“语音通话稳定性”，“语音通话整体满意度”，“网络覆盖与信号强度”，“手机上网速度”，“手机上网稳定性”和“手机上网整体满意度”，其中“网络覆盖与信号强度”，“语音通话清晰度”，“语音通话稳定性”和“语音通话整体满意度”这 4 个目标编码都分别对应 37 个特征。“网络覆盖与信号强度”，“手机上网速度”，“手机上网稳定性”和“手机上网整体满意度” 都分别对应 67 个特征。由于特征比较多，因此需要对特征进行选择，防止模型的过拟合。我们采取特征嵌入式选择的方法，其方法流程如图 10 所示。

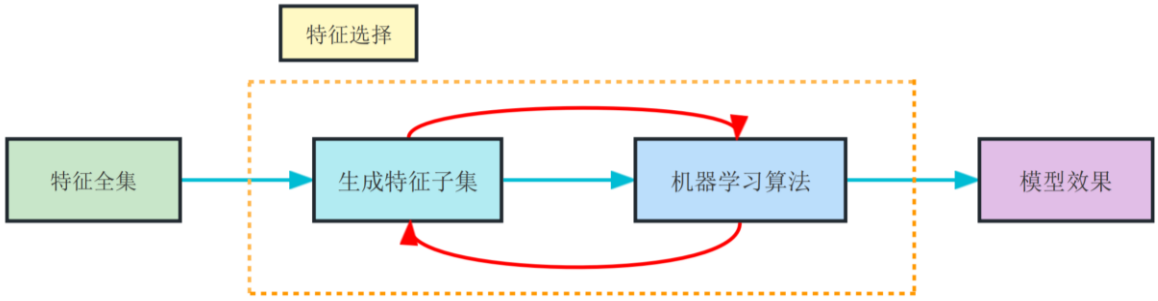


图 10：嵌入式特征选择

嵌入式特征选择，结合了过滤式和包裹式的优点，就是通过一些特殊的模型拟合数据然后根据模型自身的某些对于特征的评价的属性来作为评价指标，最后再使用包裹式的特征选择方法来选择。在本文中，我们通过采取集成树模型随机森林，然后去拟合输入输出得到训练完毕的模型，输出特征的重要性。我们不需要额外的去做特征选

择，因为模型训练的过程中自身已经完成了特征选择，得到了不同特征的评价得分，然后根据这些得分的大小就可得到不同特征针对对应的模型的特征的贡献率。

本文运用 python 进行特征选取的求解，得到了网络覆盖与信号强度，语音通话清晰度，语音通话稳定性和语音通话整体满意度的嵌入法寻找最佳阈值的学习曲线和手机上网速度，手机上网稳定性，手机上网整体满意度和网络覆盖与信号强度的嵌入法寻找最佳阈值的学习曲线(从左到右，从上到下)，如图 8 和图 9 所示。

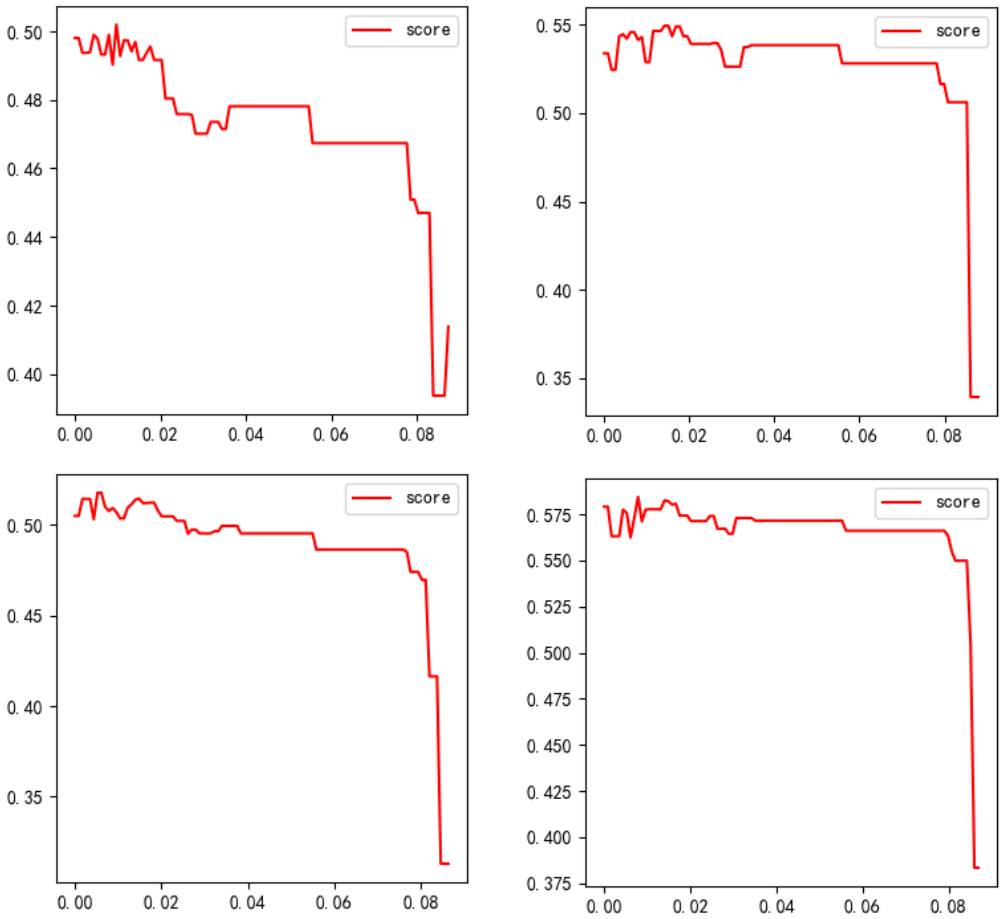
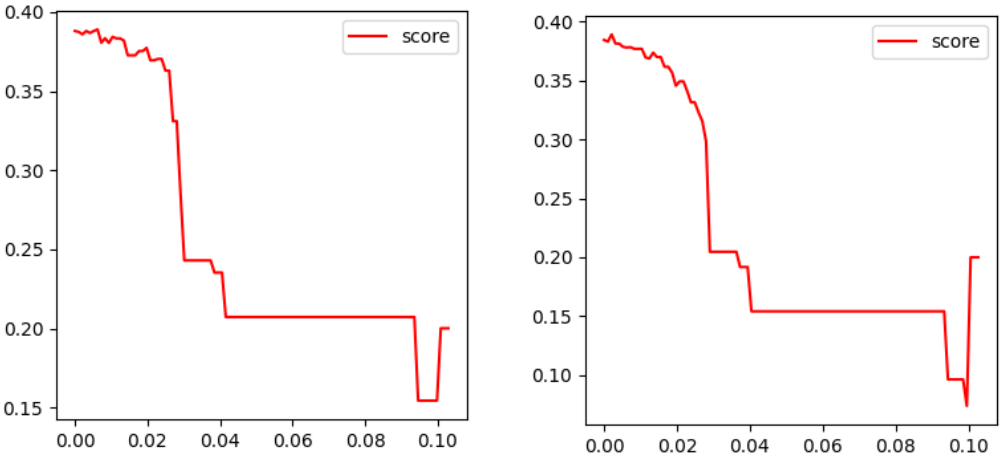


图 11：网络覆盖与信号强度，语音通话清晰度，语音通话稳定性和语音通话整体满意度的嵌入法寻找最佳阈值的学习曲线(从左到右，，从上到下)



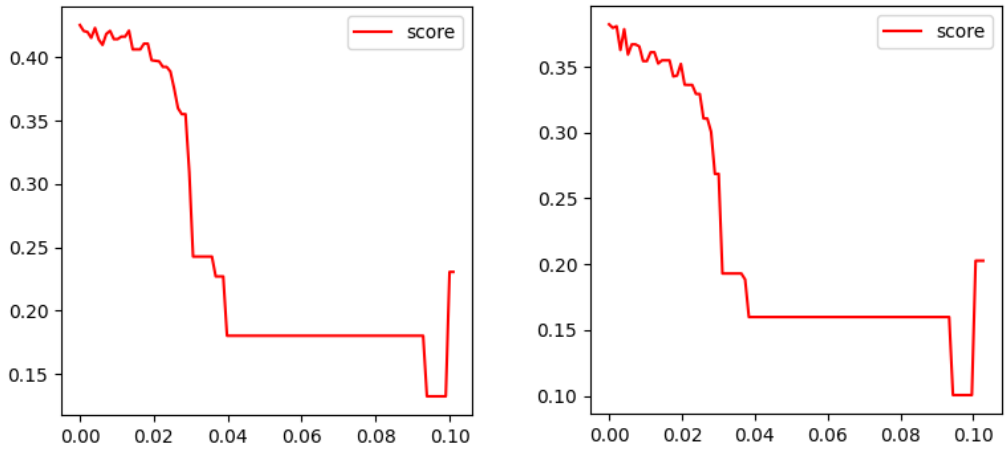


图 12：手机上网速度，手机上网稳定性，手机上网整体满意度和网络覆盖与信号强度的嵌入法寻找最佳阈值的学习曲线(从左到右，从上到下)

由图 11 和图 12 ，得知语音业务的 4 个目标编码：网络覆盖与信号强度最终留下 26 个特征，语音通话清晰度最终留下 20 个特征，语音通话稳定性最终留下 31 个特征，语音通话整体满意度最终留下 30 个特征。上网上网业务的 4 个特征目标编码：手机上网速度最终留下 35 个特征，手机上网稳定性最终留下 59 个特征，手机上网整体满意度最终留下 67 个特征，网络覆盖与信号强度最终留下 67 个特征。

7.2 客户打分的随机森林预测

7.2.1 模型的建立

题目要求我们对附件 3、4 中的客户打分进行预测研究。我们建立随机森林回归模型，可以准确的进行分析和预测。该模型是基于随机森林与分类回归决策树算法构建的客户打分预测模型，即随机森林。随机森林回归模型是一种集成算法，本身并不是一种单独的机器学习算法，而是通过在训练集上构建多个分类回归决策树模型，最后集合所有模型的最终输出。其基本流程如图 1 所示。

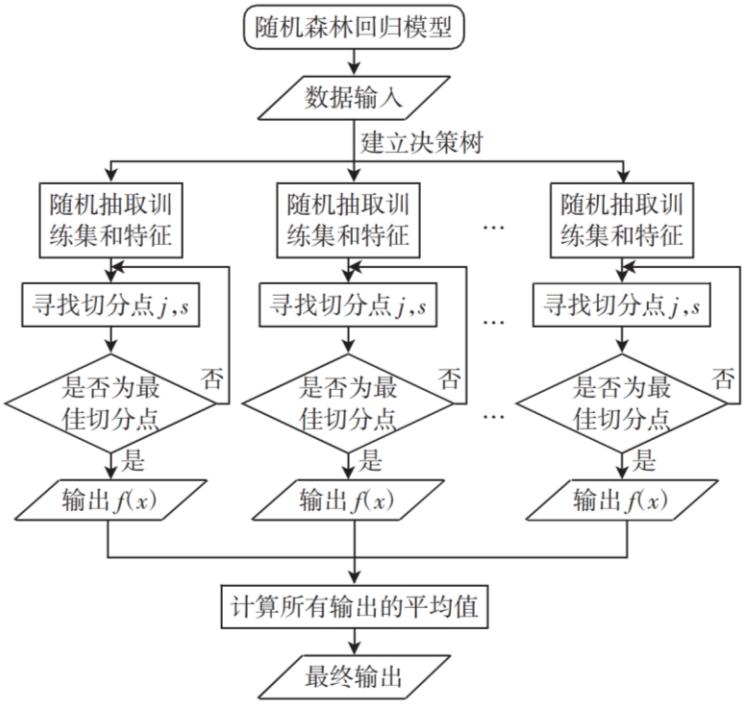


图 1 随机森林回归计算流程图

相较于单棵分类回归决策树，随机森林回归模型解决了决策树泛化能力弱的缺点，能更有效地运行在大数据集上，同时降低了大数据集的维度要求(数据集无须进行降维处理)，能获得较好的预测准确率。

随机森林回归模型的结果取决于多棵决策树的结果，决策树的好坏将直接影响森林回归模型的好坏。要想提高随机森林回归模型预测的准确率，先要调节单棵决策树的优劣。

1 决策树

决策树(decisiontree)是一类机器学习算法，因其结构形似一棵树而得名。一般一棵决策树包含一个根节点、若干个内部节点、若干个分支和若干个叶子节点，其中根节点一般用于输入，每个内部节点表示一个属性上的判断，每个分支表示一个判断结果的输出，最后每个叶子节点表示一种分类的结果，如图 2 所示。

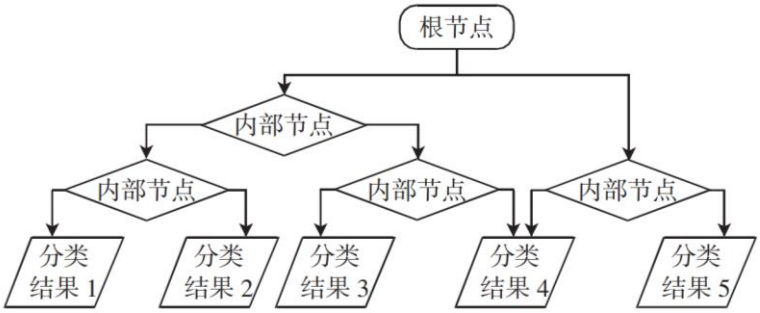


图 2 决策树结构示意图

决策树采用类似 if-else 的条件判断逻辑进行分类，属于监督学习(supervisedlearning)的一种。所谓监督学习就是用一批带有一组特征(属性)和一个分类结果的样本进行学习的方法。简单来说，就是用分类的结果和已知的样本进行学习。通过对样本的学习，可以让决策树对新的数据进行回归或分类。

2 CART 算法

分类回归决策树算法是由 Breiman 等于 1984 提出的，既可用于分类也可用于回归。CART 算法本质是对特征向量进行二元划分，即 CART 生成的决策树是一棵二叉树，能够对离散量与连续量进行分割。

1) 寻找最优切分点

为了对连续量进行$if-else$条件判断，首先要对连续量进行切分处理。

当连续的特征值输入到决策树中时，分别为输入量和输出量，且为连续变量，给定的数据集

$$D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\} \tag{13}$$

式中：$\boldsymbol{y}_i$为输入的第i个真实值；N为样本数量；x_i为输入的第i个特征向量，即

$$\boldsymbol{x}_i=\left(x_i^{(1)},x_i^{(2)},\cdots,x_i^{(n)}\right) \tag{14}$$

在进行切分前会先枚举所有特征，对于每个特征 按特征值升序排列，根据均方误差最小原则选择其中 最优的一个作为切分点。均方误差的计算方法如下

$$e_{\mathrm{NSE}}=[y_i-f(\boldsymbol{x}_i)]^2 \tag{15}$$

式中：$f(\boldsymbol{x}_i)$为i 第 i 组输入的预测值；y_i为真实值。可进一步表示为

$$e_{\mathrm{NSE}}(y_i,\hat{c})=(y_i-\hat{c})^2 \tag{16}$$

式中$\hat{c}$ 为切分后某个区域的固定输出值。选择最优切分点j和s，求解

$$\min_{j,s}\left[\min_{c_1}\sum_{x_i\in\boldsymbol{R_1}(j,s)}e_{\mathrm{MSE}}(y_i,c_1)+\min_{c_2}\sum_{x_i\in\boldsymbol{R_2}(j,s)}e_{\mathrm{MSE}}(y_i,c_2)\right] \tag{17}$$

式中：c_1和c_2为划分后两个区域内的固定输出值；R_1和R_2为j和s将特征空间切分的两个区域。

对于某个区域的输出值 c，根据均方差最小原则，构造的函数为

$$F(c) = (y_1 - c)^2 + (y_2 - c)^2 + \cdots + (y_N - c)^2 \tag{18}$$

对(6)式进行求导得

$$F'(c) = -2(y_1 - c) - 2(y_2 - c) - \cdots - 2(y_N - c) = 2Nc - 2\sum_{i=1}^{N} y_i \tag{19}$$

令$F'(c) = 0$，得

$$c = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{20}$$

根据$F(c)$的单调性易知，$\hat{c} = \frac{1}{N}\sum_{i=1}^{N} y_i$为最小值点，故(5)式可简化为

$$\min\left[\sum_{x_i \in R_1(j,s)} e_{\mathrm{MSE}}(y_i, \hat{c_1}) + \sum_{x_i \in R_2(j,s)} e_{\mathrm{MSE}}(y_i, \hat{c_2})\right] \tag{21}$$

对于固定的j扫描所有切分点s，找出满足式(9)的s，即

$$\begin{cases} \hat{c}_1 = \dfrac{1}{N_1}\displaystyle\sum_{x_i \in R_1(j,s)} y_i \\ \hat{c}_2 = \dfrac{1}{N_2}\displaystyle\sum_{x_i \in R_2(j,s)} y_i \end{cases} \tag{22}$$

在特征空间遍历变量j，寻找最优j，组成对(j, s)。如图 3 所示，对于计算出来的j和s将特征空间切分为两个区域，切分的两个区域为

$$\boldsymbol{R}_1(j,s) = \{\boldsymbol{x} \mid x^{(j)} \leqslant s\} \tag{23}$$

$$\boldsymbol{R}_2(j,s) = \{\mathrm{x} \mid \mathrm{x}^{(j)} > s\} \tag{24}$$

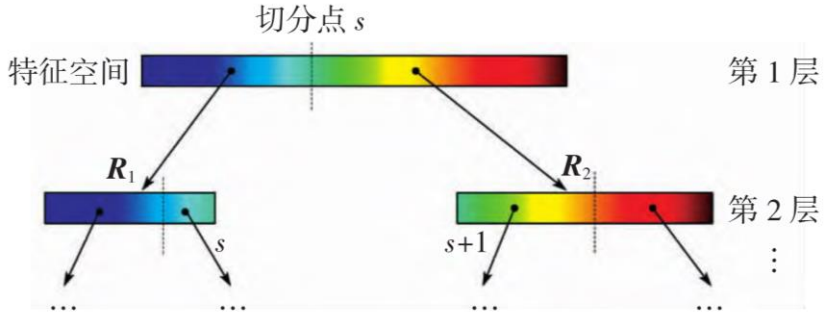


图 3 特征切分示意图

对切分出来的区域用同样的方法重复以上步骤 继续进行切分，如图 4 所示，经过多次切分后，特征空间最终被分成M个区域。

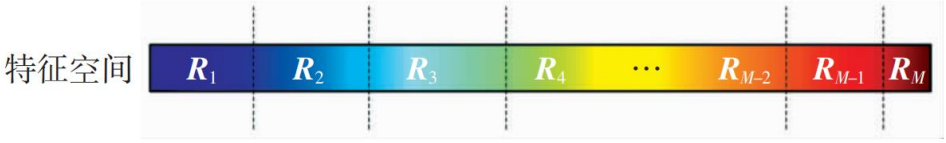


图 4 最终切分结果

2) 决策树输出

将特征空间切分为M个区域后，每个区域都有一个固定输出值$\hat{c}_m$，计算方法如下

$$\hat{c}_m = \frac{1}{N_m}\sum_{x_i \in R_m(j,s)} y_i \qquad m = 1,2,\cdots,M \tag{25}$$

生成决策树，树的最终预测值为

$$f(\boldsymbol{x}) = \sum_{m=1}^{M} \hat{c}_m I \tag{26}$$

式中I为指示函数，即

$$I = \begin{cases} 1 & \mathrm{x} \in \boldsymbol{R}_m \\ 0 & \mathrm{x} \notin \boldsymbol{R}_m \end{cases} \tag{27}$$

当随机森林与分类回归决策树算法的客户打分预测模型建立完成后，输入我们处理后的附件 1 和 2 各影响因素数据进行模型的训练，用附件 3 和 4 当作预测集，即可得出在网络覆盖与信号强度，语音通话清晰度，语音通话稳定性，语音通话整体满意度，在网络覆盖与信号强度，手机上网速度，手机上网稳定性和手机上网整体满意度的预测值。

7.2.2 随机森林的求解

随机森林模型建立回归时主要需要优化的参数为：*n_estimators* 即随机森林最大的决策树个数；*max_depth* 即决策树在划分节点分支时所需要的特征变量的个数，用于限制决策树的最大深度；*bootstrap* 参数，决定是否有放回的抽样，使用什么数据建立每棵回归树。在应用 Python 建立基于随机森林算法的客户打分模型时，需要对随机森林模型中的这三个参数进行优化，提高模型的回归预测效果。

在随机森林参数调优的过程中，选择 *GridSearch* 对参数进行优化，选 *GridSearch* 对模型进行调参工作，是因为该方法可以在输入模型的参数中，自动搜索由不同参数组成的不同模型，最终选择模型效果最优时的参数的组合输出，该方法可以有效的解决参数优化工作的成本高和效率低等问题。

第一步，在随机森林建模时按照每个参数的默认值运算，得出此时的模型评价指标值。

第二步，对*n_estimators*，*max_depth*，*bootstrap* 进行参数优化工作。其中 *n_estimators* 设置为$50-2000$，每次增加 50，因为默认值为 100，所以我们初始时设置为 50，有比 100 更小的值；*max_depth* 选择四个值分别为 5，10，20，深度过深容易造成过拟合，因此本文选择这三个值，用于优化参数；*bootstrap* 的取值有两个 *True* 和 *False*，决定是否有放回的抽样。先对 *bootstrap* 进行优化分析，使用网络搜索法优化参数，优化的次数为 160 次，根据训练结果当 *bootstrap* 设置为 *True* 时，随机森林模型有更低的误差、更高的拟合结果，所以在随机森林对客户打分预测时，使用 *bootstrap=True*。

第三步，选择优化后的参数组合，即 *bootstrap=True*，*n_estimators=80*，*max_depth=10*。

运用 python 调用 sklearn 库中的相应函数对客户打分的随机森林模型进行求解，进行参数调优后，最终通过预测得到了附件 3 中四项打分项：网络覆盖与信号强度， 语音通话清晰度，语音通话稳定性，语音通话整体满意度；附件 4 中四项打分项：网络覆盖与信号强度，手机上网速度，手机上网稳定性和手机上网整体满意度打分。如表所示。

表：附件 3 中客户打分的随机森林预测

用户 id	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
1	10	10	10	10
2	10	10	9	10
3	10	10	10	10

表：附件 4 中客户打分的随机森林预测

用户 id	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
1	10	10	10	10
2	10	10	10	10
3	8	8	8	8

由于数据较多，具体数据详见支撑材料中：16.xlsx，17.xlsx。

由表和表可知，附件 3 中语音通话整体满意度的客户打分的平均值是 9.895，其中打 1 分的有 11 人，打 8 分的有 60 人，打 9 分的有 55 人，打 10 分的人有 2473 人；网络覆盖与信号强度的客户打分的平均值是 9.608，其中打 1 分的人有 12 个，打 5 分的人有 1 个，打 6 分的人有 12 个，打 7 分的人有 24 个，打 8 分的人有 291 个，打 9 分的人有 205 个，打 10 分的人有 2054 人；语音通话清晰度的客户打分的平均值是 9.682，其中打 1 分的人有 33 个，打 7 分的人有 3 个，打 8 分的人有 159 个，打 9 分的有 200 人，打 10 分的人有 2203 人；语音通话稳定性的客户打分的平均值是 9.543，其中打 1 分的人有 63 个，打 5 分的人有 1 个，打 6 分的人有 9 个，打 7 分的人有 10 个，打 8 分的人有 202 个，打 9 分的人有 146 个，打 10 的人有 2168 个。附件 4 中手机上网整体满意度的客户打分的平均值是 9.419，其中打 1 分的人有 53 个，打 5 分的人有 4 个，打 6 分的人有 1 个，打 7 分的人有 6 个，打 8 分的人有 203 个，打 9 分的人有 10 个，打 10 分的人有 1333 个。网络覆盖与信号强度的客户打分的平均值是 9.456，其中打 1 分的人有 10 个，打 7 分的人有 19 个，打 8 分的人有 356 个，打 9 分的人有 17 个，打 10 分的人有 1208；手机上网速度的客户打分的平均值是 9.436，打 1 分的人有 1 个，打 5 分的人有 1 个，打 7 分的人有 14 个，打 8 分的人有 407 个，打 9 分的人有 38 个，打 10 分的人有 1149 个；手机上网稳定性的客户打分的平均值是 9.242，其中打 1 分的人有 47 个，打 5 分的人有 8 个，打 6 分的人有 8 个，打 7 分的人有 30 个，打 8 分的人 293 个，打 9 分的人 48 个，打 10 分的人有 1175 个。

7.2.3 随机森林的检验

建立基于随机森林算法的客户打分模型主要目的是对客户打分的各个影响因素进行回归分析，回归的最终目的是进行客户打分的预测。所以我们需要对预测结果进行检验和评估，对结果的检验和评估我们采取如下公式进行计算

$$score = 0.2 * (1 - Mape) + 0.8 * Accuracy_5$$

$$Mape = \frac{1}{m}\sum_{i=1}^{m} Ape_i$$

$$Ape = \frac{|\hat{y} - y|}{y}$$

其中，$score$ 是我们的评估模型得分，其值在[0,1]区间内，越接近1越好，$Accuracy$ 5为相对误差 Ape 在 5%以内的样本数量。

运用 python 调用 sklearn 库中的相应函数，进行参数优化后的客户打分随机森林模型的得分计算，如表和表所示。

表：附件 1 中测试集与训练集得分

特征名	测试集	训练集
网络覆盖与信号强度	0.494	0.6432
语音通话清晰度	0.574	0.6425
语音通话稳定性	0.528	0.626

语音通话整体满意度	0.613	0.641

表：附件 2 中测试集与训练集得分

特征名	测试集	训练集
网络覆盖与信号强度	0.408	0.537
手机上网速度	0.408	0.563
手机上网稳定性	0.382	0.553
手机上网整体满意度	0.443	0.550

由表和表可知，对于预测附件 3 中的 4 项客户打分，其中测试集得分最高的是：语音通话整体满意度，得分为 0.613；测试集得分最低的是：网络覆盖与信号强度，得分为 0.494；训练集得分最高的是：网络覆盖与信号强度，得分为 0.6432；训练集得分最低的是：语音通话稳定性，得分为 0.626。对于预测附件 4 中的 4 项客户打分，其中测试集得分最高的是：手机上网整体满意度，得分为 0.443；测试集得分最低的是：手机上网稳定性，得分为 0.382；训练集得分最高的是：手机上网速度，得分为 0.563；训练集得分最低的是：网络覆盖与信号强度，得分为 0.537。

7.3 客户打分的 BP 神经网络预测

7.3.1 BP 神经网络的建立

1 模型介绍

人工神经网络模拟了生物神经网络的一系列运行机理，它能够像人脑一样处理大量数据，并且输出一个想要的结果，具有自学习、自组织、较好的容错性和优良的非线性逼近能力，主要应用于函数逼近、模式识别、分类.这里我们采用误差反传算法的人工神经网络即 BP 神经网络识别模式模型对就诊人员进行分类.

2 模型思想

对于 N 个输人学习样本，已知与其对应的输出样本.学习的目的是用网络的实际输出与目标矢量之间的误差来修改其权值，使实际与期望尽可能地接近，即使网络输出层的误差平方和达到最小，他是通过连续不断地在相对于误差函数斜率下降的方向上计算网络权值和偏差的变化而逐渐逼近目标的.每一次权值和偏差的变化都与网络误差的影响成正比，并以反向传播的方式传递到每一层.

BP 算法由两部分组成：信息的正向传递与误差的反向传播.在正向传递过程中输人信息从输人层经隐含层逐层计算传向输出层，每一层神经元的输出作用于下一层神经元的输人.如果输出层没有得到期望的输出，则计算输出层的误差变化值，然后转向反向传播，通过网络将误差信号沿原来的连接通路反传回来修改各层的权值直至达到期望目标.

3 模型准备

1) 数据处理

采用 BP 神经网络方法建模的首要和前提条件是有足够多典型性好和精度高的样本，所以本文将附件 1 中影响客户语音业务体验因素的数据 70% 为训练集，30% 为测试集，附件 3 中影响客户上网体验因素的数据为预测集，用来预测网络覆盖与信号强度、 语音通话清晰度，语音通话稳定性和语音通话整体满意度的打分。将附件 2 中影响客户上网体验因素的数据 70% 为训练集，30% 为测试集，附件 4 中影响客户上网体验因素的数据为预测集，用来预测在网络覆盖与信号强度、手机上网速度、 手机上网稳定性和手机上网整体满意度的打分。

2) 模型流程图

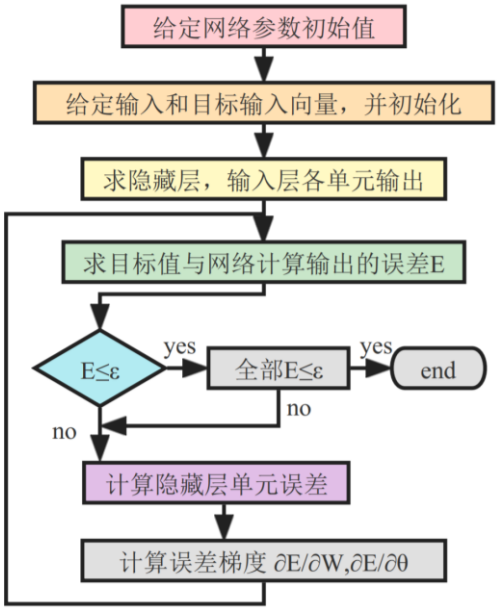


图 13 模型流程图

3) **网络拓扑结构的确定**

输入层由附件 1 中影响客户语音业务体验的各因素序列组成，故有 5 个节点；输出层有 5 个节点组成；隐层的个数由于没有统一的标准，我们就以公式 $K=\sqrt{N+M}+\delta(N，M$为输入层和输出层的节点数，δ 为 1 到 10 之间的数)来确定，K=13.

BP 神经网络结构示意图如下:

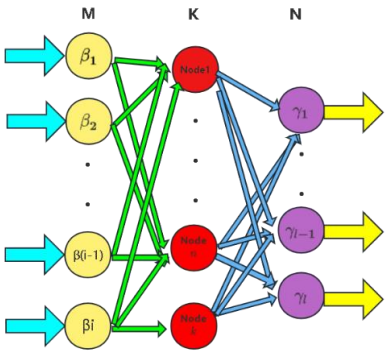


图 14 BP 神经网络结构示意图

4 模型的建立

1) 输入模式正向传播：

a) 取$(-1，1)$之间的随机数初始化网络权值 W 与阀值 θ 及输入学习速度 γ 和期望误差 ε。

b) 输入原始数据矩阵 a 和目标数据矩阵 T 。

c) 计算出隐层各神经元的输出矩阵b和输出层的输出矩阵c；

$$b=f\left(\sum_{i=1}^{7}W_{ij}a_i-\theta_j\right)j=1\cdots 10 \qquad c=f\left(\sum_{j=1}^{10}W_{jt}b_j-\theta_t\right)t=1,2 \qquad (5)$$

式中的 f 为神经元激活函数，我们取 f 为 $sigmoid$ 形函数，它具有如下形式：

$$f(x)=\frac{1}{(1+e^{-x})} \qquad (6)$$

2) 输出误差的反向传播：

网络的希望输出与实际输出的偏差为： $\partial_t=T_t-c_t$ (7)

采用平方和误差进行计算： $E=\frac{1}{2}\sum_{t=1}^{2}(T_t-c_t)^2=\frac{1}{2}\sum_{t=1}^{2}(\partial_t)^2$ (8)

输出层各神经元的输出误差： $\partial_t = (T_t - c_t)f'\left(\sum_{j=1}^{10} W_{jt}b_j - \theta_t\right)$ (9)

隐含层各神经元的输出误差： $\partial_j = f'(W_{ij}a_i - \theta_j)\sum_{j=1}^{2} \partial_t W_{jt}$ (10)

修正权值： $W_{jt}(t+1) = W_{jt}(t) + \gamma\,\partial_t b_j W_{ij}(t+1) = W_{ij}(t)\gamma\,\partial_t a_i$ (11)

当网络训练到平方和误差$E < \varepsilon$时，BP 神经网络的结构就真正确定下来. 然后把检验样品输入网络与实际值进行比较，如果符合度不大，就重新训练，直到检验样品的训练结果和实际值符合度较大. 这时再把测试样品输入神经网络即可得到结果。

7.3.2 BP 神经网络的求解

运用 python 调用 sklearn 库中的相应函数对客户打分的 BP 神经网络模型进行求解，先使用默认参数来对客户打分进行拟合，随后再对 BP 神经网络进行调参，最终通过预测得到了附件 3 中四项打分项：网络覆盖与信号强度， 语音通话清晰度，语音通话稳定性，语音通话整体满意度；附件 4 中四项打分项：网络覆盖与信号强度，手机上网速度，手机上网稳定性和手机上网整体满意度打分。如表所示。

表：附件 3 中客户打分的 BP 神经网络预测

用户 id	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
1	10	10	10	9
2	10	10	10	9
3	7	10	10	2

表：附件 4 中客户打分的 BP 神经网络预测

用户 id	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
1	3	10	1	10
2	3	10	1	10
3	1	10	10	10

由于数据较多，具体数据详见支撑材料中：14.xlsx，15.xlsx。

由表和表可知，附件 3 中语音通话整体满意度的客户打分的平均值是 9.881，其中打 7 分的有 79 人，打 8 分的有 18 人，打 9 分的有 36 人，打 10 分的人有 2466 人；网络覆盖与信号强度的客户打分的平均值是 9.906，其中打 1 分的人有 26 个，打 6 分的人有 2 个，打 8 分的人有 1 个，打 10 分的人有 2570 人；语音通话清晰度的客户打分的平均值是 9.970，其中打 9 分的有 79 人，打 10 分的人有 2520 人语音通话稳定性的客户打分的平均值是 8.506，其中打 1 分的人有 55 个，打 2 分的人有 19 个，打 5 分的人有 10 个，打 7 分的人有 5 个，打 8 分的人有 716 个，打 9 分的人有 1739 个，打 10 的人有 55 个。附件 4 中手机上网整体满意度的客户打分的平均值是 3.933，其中打 1 分的人有 52 个，打 3 分的人有 1325 个，打 5 分的人有 1 个，打 7 分的人有 6 个，打 8 分的人有 1 个，打 10 分的人有 225 个。网络覆盖与信号强度的客户打分的平均值是 4.223，其中打 1 分的人有 951 个，打 5 分的人有 50 个，打 6 分的人有 22 个，打 7 分的人有 70 个，打 8 分的人有 97 个，打 10 分的人有 420；手机上网速度的客户打分的平均值是 5.448，打 1 分的人有 806 个，打 5 分的人有 9 个，打 7 分的人有 10 个，打 10 分的人有 785 个；手机上网稳定性的客户打分的平均值是 9.440，其中打 3 分的有 91 个，打 5 分的人有 53 个，打 10 分的人有 1466 个。

7.3.3 BP 神经网络的检验

与随机森林模型的检验步骤相同，运用 python 调用 sklearn 库中的相应函数，进行

参数优化后的客户打分 BP 神经网络模型的得分计算，如表和表所示。

表：附件 1 中测试集与训练集得分

特征名	测试集	训练集
网络覆盖与信号强度	0.504	0.532
语音通话清晰度	0.565	0.550
语音通话稳定性	0.514	0.545
语音通话整体满意度	0.607	0.576

表：附件 2 中测试集与训练集得分

特征名	测试集	训练集
网络覆盖与信号强度	0.407	0.443
手机上网速度	0.408	0.422
手机上网稳定性	0.379	0.445
手机上网整体满意度	0.444	0.461

由表和表可知，对于预测附件 3 中的 4 项客户打分，其中测试集得分最高的是：语音通话整体满意度，得分为 0.607；测试集得分最低的是：网络覆盖与信号强度，得分为 0.504；训练集得分最高的是：语音通话整体满意度，得分为 0.577；训练集得分最低的是：网络覆盖与信号强度，得分为 0.532。对于预测附件 4 中的 4 项客户打分，其中测试集得分最高的是：手机上网整体满意度，得分为 0.444；测试集得分最低的是：手机上网稳定性，得分为 0.379；训练集得分最高的是：手机上网整体满意度，得分为 0.461；训练集得分最低的是：手机上网速度，得分为 0.422。

7.3 客户打分的 XGBoost 预测

7.3.1 XGBoost 模型的建立

XGBoost(eXtreme Gradient Boosting)又称为极端梯度提升树，它是 Chen 等于 2014 年提 出的一种集成式学习算法，通过集成多个弱分类器而构建强分类器. 其集成模型可表示为：

$$\hat{\mathbf{y}}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in F$$

假设数据集有 n 个样本，m 个特征，定义为：

$$D = \{(x_i, y_i)\}(|\ D\ | = n, x_i \in R^m, y_i \in R)$$

式中：x_i表示第 i 个样本；y_i表示第 i 个样本的标签.

CART 树的空间为 F，公式如下：

$$F = \{f(x) = w_{q_{(x)}}\}(q: R^m \to T, w \in R^T)$$

式中：q 表示树的模型；$w_{q_{(x)}}$表示树 q 中叶节点的分数集合；T 表示树 q 的叶节点数量. 模型的学习目标是学习所有的树模型(假设为 $f(x)$). 为能够顺利学习模型 $f(x)$，需要确 定以下目标函数：

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

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \parallel w \parallel^2 ()$$

式中：$\sum_{\mathrm{i}=1}^{n} l(\hat{\mathrm{y}}_i, \mathrm{y}_i)$为损失函数项，即训练误差；$\sum_{k=1}^{K} \Omega(f_k)$为树的复杂度之和，能有效 防止过拟合；$\hat{\mathrm{y}}_i$ 表示模型的预测值；y_i表示第 i 个样本的标签；f_k表示第 k 棵

树模型；T 表示每棵树的叶节点数量；w表示每棵树叶节点的分数组成的集合；γ和λ表示系数.

7.3.2 XGBoost 模型的求解

我们首先运用 python 调用 sklearn 库中的相应函数对客户打分进行拟合，然后再对 XGBoost 进行调参优化。在进行 XGboost 的参数调优时，我们主要步骤分为二步：

第一步：在构建 XGboost 模型的时候首先以各参数的默认值构建模型，计算得到初始模型的评价标准值。

第二步：对参数*learning_rate*、*n_estimators*、*max_depth* 进行调整。其中 *learning_rate* 为学习率，默认为 0.3，在回归任务中，通过它来控制迭代速率，抑制过拟合现象的产生；*n_estimators* 为 *boosting* 的迭代次数（弱分类器的个数）；*max_depth* 为树的最大深度，它通常用来避免过拟合现场的产生。*max_depth* 越大，模型就会学到更加具体的样本，取值为非负数。正则化参数 *lambda,alpha* 的调优，这些参数可以降低模型的复杂度，从而提高模型的表现。

进行参数调优后，最终通过预测得到了附件 3 中四项打分项：网络覆盖与信号强度，语音通话清晰度，语音通话稳定性，语音通话整体满意度；附件 4 中四项打分项：网络覆盖与信号强度，手机上网速度，手机上网稳定性和手机上网整体满意度打分。如表所示。

表：附件 3 中客户打分的 XGBoost 预测

用户 id	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
1	10	10	10	6
2	10	10	9	1
3	10	10	10	7

表：附件 4 中客户打分的 XGBoost 预测

用户 id	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
1	10	10	10	10
2	10	8	8	8
3	9	10	8	9

由于数据较多，具体数据详见支撑材料中：18.xlsx，19.xlsx。

由表和表可知，附件 3 中语音通话整体满意度的客户打分的平均值是 8.985，其中打 1 分的人有 173 个，打 2 分的人有 1 个，打 3 分的人有 5 个，打 4 分的人有 26 个，打 5 分的人有 24 个，打 6 分的人有 34 个，打 7 分的人有 87 个，打 8 分的人有 18 个，打 9 分的人有 36 个，打 10 分的人有 2025 个；网络覆盖与信号强度的客户打分的平均值是 8.641，其中打 1 分的人有 157 个，打 2 分的人有 21 个，打 3 分的人有 1 个，打 4 分的人有 5 个，打 5 分的人有 120 个，打 6 分的人有 101 个，打 7 分的人有 79 个，打 8 分的人有 305 个，打 9 分的人有 36 个，打 10 分的人有 1748 个；语音通话清晰度的客户打分的平均值是 8.623，其中打 1 分的人有 258 个，打 2 分的人有 4 个，打 3 分的人有 8 个，打 5 分的人有 67 个，打 6 分的人有 22 个，打 7 分的人有 71 个，打 8 分的人有 190 个，打 9 分的人有 154 个，打 10 分的人有 1825 个;语音通话稳定性的客户打分的平均值是 8.721，其中打 1 分的人有 91 个，打 2 分的人有 17 个，打 3 分的人有 17 个，打 4 分的人有 24 个，打 5 分的人有 92 个，打 6 分的人有 92 个，打 7 分的人有 117 个，打 8 分的人有 295 个，打 9 分的人有 337 个，打 10 分的人有 1517 个；附件 4 中手机上网整体满意度的客户打分的平均值是 9.311，其中打 1 分的人有 17 个，打 3 分的人有 2 个，打 4 分的人有 3 个，打 5 分的人有 23 个，打 6 分的人有 17 个，打 7 分

的人有 47 个，打 8 分的人有 217 个，打 9 分的人有 163 个，打 10 分的人有 1120 个；网络覆盖与信号强度的客户打分的平均值是 9.090，其中打 1 分的人有 15 个，打 2 分的人有 3 个，打 3 分的人有 1 个，打 5 分的人有 6 个，打 6 分的人有 26 个，打 7 分的人有 72 个，打 8 分的人有 424 个，打 9 分的人有 101 个，打 10 分的人有 962 个；手机上网速度的客户打分的平均值是 9.032，打 1 分的人有 5 个，打 3 分的人有 2 个，打 4 分的人有 7 个，打 5 分的人有 18 个，打 6 分的人有 9 个，打 7 分的人有 73 个，打 8 分的人有 506 个，打 9 分的人有 101 个，打 10 分的人有 889 个；手机上网稳定性的客户打分的平均值是 9.036，其中打 1 分的人有 25 个，打 3 分的人有 5 个，打 4 分的人有 5 个，打 5 分的人有 20 个，打 6 分的人有 16 个，打 7 分的人有 77 个，打 8 分的人有 360 个，打 9 分的人有 147 个，打 10 分的人有 955 个。

7.3.3 XGBoost 模型的检验

与随机森林模型的检验步骤相同，运用 python 调用 sklearn 库中的相应函数，进行参数优化后客户打分的 XGBoost 模型的得分计算，如表和表所示

表：附件 1 中测试集与训练集得分

特征名	测试集	训练集
网络覆盖与信号强度	0.473	0.956
语音通话清晰度	0.552	0.991
语音通话稳定性	0.489	0.994
语音通话整体满意度	0.588	0.987

表：附件 1 中测试集与训练集得分

特征名	测试集	训练集
网络覆盖与信号强度	0.395	0.948
手机上网速度	0.388	0.959
手机上网稳定性	0.369	0.951
手机上网整体满意度	0.431	0.956

由表和表可知，对于预测附件 3 中的 4 项客户打分，其中测试集得分最高的是：语音通话整体满意度，得分为 0.588；测试集得分最低的是：网络覆盖与信号强度，得分为 0.473；训练集得分最高的是：语音通话稳定性，得分为 0.994；训练集得分最低的是：网络覆盖与信号强度，得分为 0.956。对于预测附件 4 中的 4 项客户打分，其中测试集得分最高的是：手机上网整体满意度，得分为 0.431；测试集得分最低的是：手机上网稳定性，得分为 0.369；训练集得分最高的是：手机上网速度，得分为 0.959；训练集得分最低的是：网络覆盖与信号强度，得分为 0.948。

7.4 模型融合

7.4.1 融合模型的建立

由于 BP 神经网络，随机森林和 XGBoost 预测的客户打分模型得分还有待提高，所以本文尝试进行模型的融合来提高预测结果的精度。

模型融合我们采取 stacking 的方法。stacking 就是用初始训练数据学习出若干个基学习器后，将这几个学习器的预测结果作为新的训练集，来学习一个新的学习器。本文已经训练好了三个基学习器，BP 神经网络，随机森林和 XGBoost，将这三个学习器的输出作为次集训练集给次级学习器。次学习器我们采取线性回归模型，因为我们的基学习器是强学习器，次级学习器选择简单的模型可以避免过拟合现象，因为结合了不同学习器的学习效果，使用 stacking 的方法可以使得最终融合的模型 更加稳定，表现更好。stacking 融合的方法如下图所示：

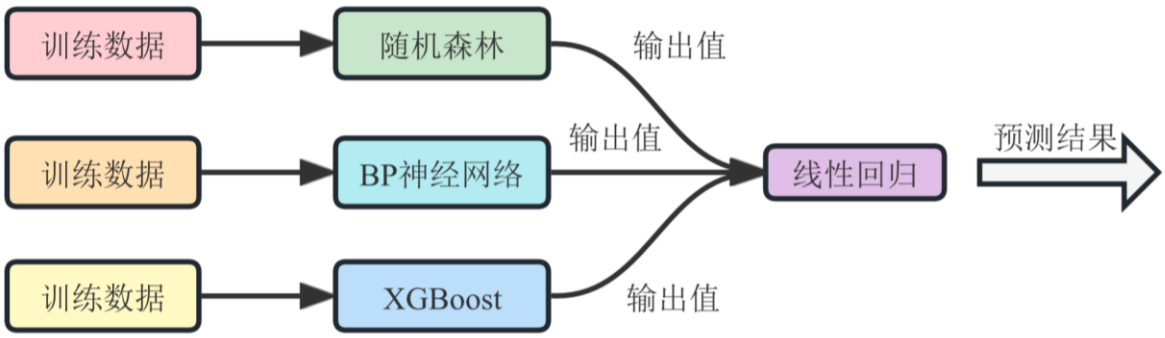


图 模型融合流程

运用 python 调用 sklearn 库中的相应函数对客户打分的融合模型进行求解，最终通过预测得到了附件 3 中四项打分项：网络覆盖与信号强度， 语音通话清晰度，语音通话稳定性，语音通话整体满意度；附件 4 中四项打分项：网络覆盖与信号强度，手机上网速度，手机上网稳定性和手机上网整体满意度打分。如表所示。

表：附件 3 中客户打分的融合模型预测值

用户 id	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
1	10	10	10	10
2	10	10	9	10
3	10	10	10	10

表：附件 4 中客户打分的融合模型预测值

用户 id	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
1	9	10	10	10
2	9	9	9	8
3	9	9	8	8

由于数据较多，具体数据详见支撑材料中：20.xlsx，21.xlsx。

由表和表可知，附件 3 中语音通话整体满意度的客户打分的平均值是 8.548，其中打 1 分的人有 5 个，打 2 分的人有 349 个，打 3 分的人有 4 个，打 4 分的人有 8 个，打 5 分的人有 16 个，打 6 分的人有 51 个，打 7 分的人有 65 个，打 8 分的人有 100 个，打 9 分的人有 167 个，打 10 分的人有 1827 个；网络覆盖与信号强度的客户打分的平均值是 8.769，其中打 1 分的人有 35 个，打 2 分的人有 191 个，打 4 分的人有 3 个，打 5 分的人有 21 个，打 6 分的人有 112 个，打 7 分的人有 68 个，打 8 分的人有 226 个，打 9 分的人有 129 个，打 10 分的人有 1814 个；语音通话清晰度的客户打分的平均值是 8.974，其中打 1 分的人有 5 个，打 2 分的人有 162 个，打 3 分的人有 10 个，打 4 分的人 20 个，打 5 分的人有 11 个，打 6 分的人有 90 个，打 7 分的人有 28 个，打 8 分的人有 153 个，打 9 分的人有 330 个，打 10 分的人有 1790 个;语音通话稳定性的客户打分的平均值是 7.952，其中打 1 分的人有 33 个，打 2 分的人有 398 个，打 3 分的人有 1 个，打 4 分的人有 9 个，打 5 分的人有 185 个，打 6 分的人有 43 个，打 7 分的人有 145 个，打 8 分的人有 30 个，打 9 分的人有 188 个，打 10 分的人有 1567 个；附件 4 中手机上网整体满意度的客户打分的平均值是 8.745，其中打 1 分的人有 1 个，打 2 分的人有 17 个，打 4 分的人有 3 个，打 5 分的人有 10 个，打 6 分的人有 40 个，打 7 分的人有 48 个，打 8 分的人有 273 个，打 9 分的人有 916 个，打 10 分的人有 285 个；网络覆盖与信号强度的客户打分的平均值是 8.536，其中打 2 分的人有 19 个，打 4 分的人有 7 个，打 5 分的人有 14 个，打 6 分的人有 35 个，打 7 分的人有 65 个，打 8 分的人有 474 个，打 9 分的人有 810 个，打 10 分的人有 186 个；手机上网速度的客户打分

的平均值是 8.932，打 1 分的人有 5 个，打 2 分的人有 11 个，打 4 分的人有 2 个，打 5 分的人有 12 个，打 6 分的人有 15 个，打 7 分的人有 125 个，打 8 分的人有 432 个，打 9 分的人有 216 个，打 10 分的人有 792 个；手机上网稳定性的客户打分的平均值是 8.766，其中打 1 分的人有 17 个，打 2 分的人有 16 个，打 3 分的人有 2 个，打 4 分的人有 3 个，打 5 分的人有 17 个，打 6 分的人有 23 个，打 7 分的人有 80 个，打 8 分的人有 366 个，打 9 分的人有 521 个，打 10 分的人有 562 个。

7.4.2 融合模型的检验

与随机森林模型的检验步骤相同，运用 python 调用 sklearn 库中的相应函数，进行客户打分的融合模型的得分计算，如表和表所示

表：附件 1 中测试集与训练集得分

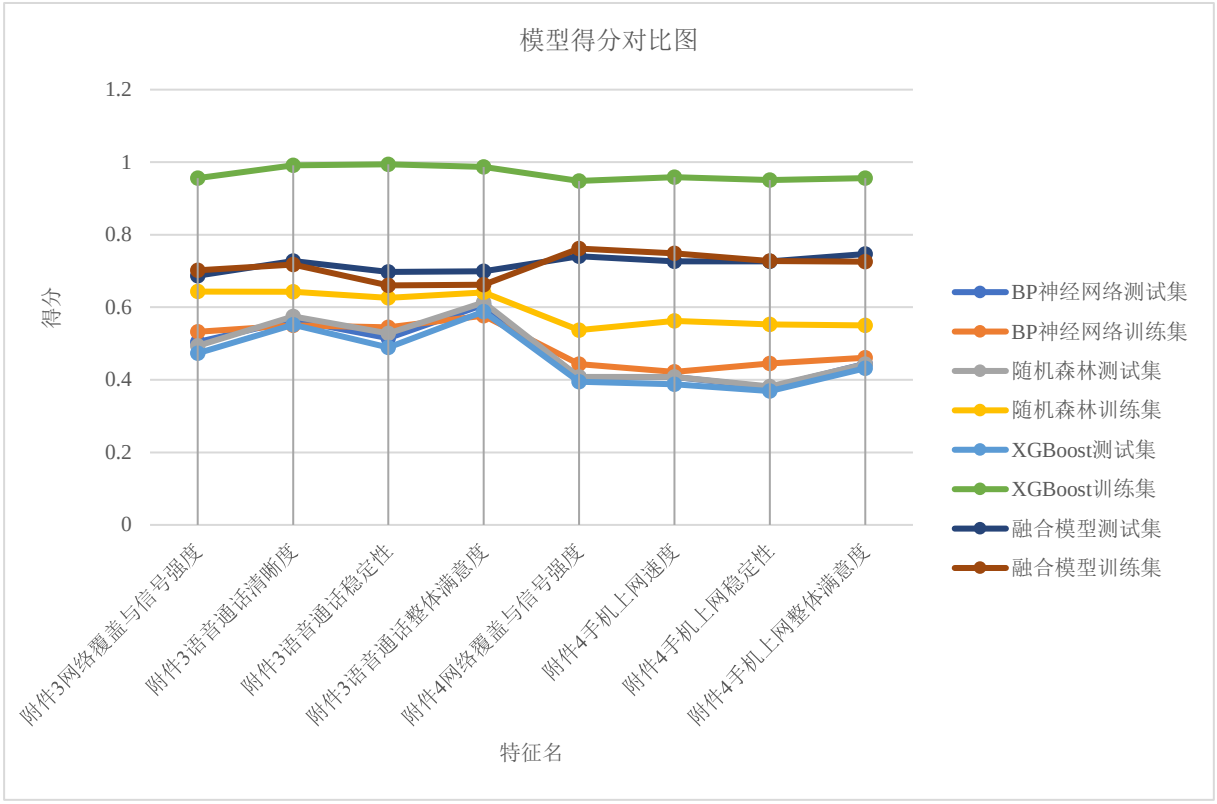
特征名	测试集	训练集
网络覆盖与信号强度	0.687	0.702
语音通话清晰度	0.727	0.717
语音通话稳定性	0.697	0.660
语音通话整体满意度	0.699	0.662

表：附件 2 中测试集与训练集得分

特征名	测试集	训练集
网络覆盖与信号强度	0.741	0.762
手机上网速度	0.7263	0.749
手机上网稳定性	0.7264	0.728
手机上网整体满意度	0.747	0.726

由表和表可知，融合后的客户打分预测模型得分比随机森林，BP 神经网络和 XGBoost 任意单个模型的得分都要高。随机森林，BP 神经网络和 XGBoost 的训练集和测试集得分普遍在 0.5 左右，而融合后的模型训练集和测试集得分普遍在 0.7 左右，模型的准确性达到了期望中的要求。对于预测附件 3 中的 4 项客户打分，其中测试集得分最高的是：语音通话清晰度，得分为 0.727；测试集得分最低的是：网络覆盖与信号强度，得分为 0.687；训练集得分最高的是：语音通话清晰度，得分为 0.717；训练集得分最低的是：语音通话稳定性，得分为 0.660。对于预测附件 4 中的 4 项客户打分，其中测试集得分最高的是：手机上网整体满意度，得分为 0.747；测试集得分最低的是：手机上网速度，得分为 0.7263；训练集得分最高的是：网络覆盖与信号强度，得分为 0.762；训练集得分最低的是：手机上网整体满意度，得分为 0.726。

从训练效果发现融合后的模型比单独的强学习器效果更好，随机森林，LightGBM 和 XGBoost 和融合模型的训练集和测试集得分，如图所示



图模型得分对比图

排除 XGBoost 的算法特性(测试集得分高)，融合后的客户打分预测模型得分比随机森林，BP 神经网络和 XGBoost 任意单个模型的得分都要高，随机森林，BP 神经网络和 XGBoost 的训练集和测试集得分普遍在 0.5 左右甚至更低，而融合后的模型训练集和测试集得分普遍在 0.7 左右，模型的准确性达到了期望中的要求，预测的客户打分值也更令人信服。

8 模型的评价、改进与推广

8.1 模型的优点

8.1.1 客户打分的随机森林模型优点

1) 在当前所有算法中，具有极好的准确率
2) 能够有效地运行在大数据集上
3) 能够处理具有高维特征的输入样本，而且不需要降维
4) 能够评估各个特征在分类问题上的重要性
5) 对于缺省值问题也能够获得很好得结果

8.1.2 客户打分的 BP 神经网络模型优点

8.1.3 客户打分的 XGBoost 模型优点

1) 在目标函数当中添加了正则化项，降低了出现过拟合现象的可能性，不仅使用了一阶导数， 还使用了二阶导数，使损失函数更加精确的同时还可以自定义损失。
2) 可以并行优化，并且 XGBoost 是在特征粒度上的并行。

3) 考虑了对稀疏值的处理，能给缺失值或指定值设置分支默认方向，极大提升了算法效率。
4) 允许列抽样，既可以抑制过拟合，又可以减少计算量。

8.2 模型的缺点

8.2.1 客户打分的随机森林模型缺点

参数量多，调参不容易

8.3 模型的改进

8.4 模型的推广

参考文献

附录