队伍编号	MCB2200122
赛道	B

基于 XGB、逻辑回归、支持向量机、随机森林、GBDT、ABC-Boost 的 Stacking 语音上网满意度预测模型

摘　　要

随着生活水平的提高，用户对通话、上网的质量、服务有更大的期待与需求。用户对网络运营商的评价是运营商优化服务、提升质量的重要参考。本文通对中国移动用户语音和上网满意度数据进行 EDA、特征工程、建立模型、模型调参、模型融合等步骤，建立了基于 XGB、逻辑回归、支持向量机、随机森林、GBDT、ABC-Boost 的 Stacking 语音上网满意度预测模型。

模型需要预测的指标共有 8 个，本文将 8 个指标分开，分别预测，考虑到评价具有主观随意性，本文使用 **RMSE** 作为模型的评价指标，**准确率**作为辅助的评价指标。本文针对问题总体上分为特征选择和模型建立和预测两个部分。

针对问题一，**第一步**，对**数据**进行**预处理**。对数据进行**分箱**、对类别属性进行 **Label Encoder**、对缺失值进行填充、删除无用属性。**第二步**，进行**特征重要性分析**。求出语音和上网满意度数据的 **Kendall 相关性系数**；使用随机森林求出属性的特征重要性；在随机森林基础上求出 **Permutation importance**。通过研究发现：对于语音通话，最影响用户体验的是通话遇到的问题，如**手机没有信号**、**无法拨通**、**突然中断**、**有杂音**、**听不清**等问题。对于手机上网，最影响用户体验的是信号和网速问题，如**网络信号差**、**没有信号**、**网速慢**、**网速不稳定**、**热门 APP 卡顿**等问题。

针对问题二，**第一步**，**数据预处理**及**数据集划分**。根据问题一的结果，剔除连续型变量以及测试集中没有的指标，对类别数据进行 **One-hot 编码**；将数据集划分为训练集和测试集，比例为 8:2。**第二步**，模型筛选。选取 XGB、逻辑回归、支持向量机、随机森林、K 近邻算法、朴素贝叶斯、感知机、SGD、决策树、GBDT、ABC-Boost、神经网络共 12 个模型在训练集上训练，得出模型在测试集上的**准确率**和 **RMSE**。**第三步**，**模型调参**。综合准确率和 RMSE，选取 **XGB**、**逻辑回归**、**SVC**、**随机森林**、**GBDT**、**ABC-Boost** 共 6 个模型使用**网格搜索**进行超参数调优。**第四步**，**模型融合**。使用 **Stacking** 方法对 6 个模型进行特征融合，8 个标签在测试集上的准确率在 **40.17%-59.24%** 之间，RMSE 值在之间 **2.28-2.98** 之间。**第五步**，模型预测。使用训练好的模型，对 8 个目标属性进行预测，并将预测的结果导出到文件中。

最后我们进行了总结，并对模型进行了评价与推广。

关键词：**Permutation importance**；**随机森林**；**Stacking 模型**；**RMSE**