摘　要

随着移动通信技术的飞速发展，人们的生活已经进入了一个日新月异的时代。近年来，随着第五代通信技术的飞速发展，无论是依赖于超低时延的远程医疗、自动驾驶等行业应用，还是基于高速移动宽带的高清直播、在线秒杀等生活场景，逐步商用的 5G 正在方方面面改变着人们的生活方式。因此，推广 5G 业务，促成 4G 套餐用户向 5G 转化也成为各大运营商发展的重点。然而受限于 5G 技术发展不成熟、用户需求不够明确、现有套餐资费昂贵等问题，现有的 5G 套餐用户不仅基数较少且发展较为缓慢。为了解决上述问题，运营商有必要对已经更换 5G 套餐用户的特点进行深入探究，并对现有客群中潜在的 5G 用户进行针对性的营销，推动他们更换 5G 资费套餐。

为实现 5G 套餐的潜在客户挖掘，本文进行了如下研究工作：（1）在业务层面，基于原始数据进行了数据探查，对潜在客户在消费行为等方面的特点进行了归纳。（2）在模型层面，为增强数据质量进而提高模型预测能力，本文对数据集进行了相应的数据预处理和特征工程操作。针对业务数据集中 5G 套餐客户占比过低导致的模型过拟合且对正样本的分类能力不足等问题，本文提出了一种基于 Ada-Stacking 融合思想的潜在客户识别模型。融合模型在基模型层参考 AdaBoost 算法样本权重迭代的方法，对单个基模型进行 k 轮迭代，使得每一轮模型都更加关注上一轮的错分样本。对于元模型层，模型还通过元特征交叉相乘的方式丰富元模型层特征，以期解决模型过拟合的问题。（3）为了兼顾模型性能和训练时间，本文结合 f1-score 和训练时间等指标对 Ada-Stacking 基模型迭代次数 k 进行了评估，最终确定 k=5 时模型的训练时间和预测能力达到最佳。

经过模型评估与优化，Ada-Stacking 融合模型在测试集上相对于传统 Stacking 模型在 auc 指标上提升 1%~2%，而对于衡量正样本预测能力的召回率，Ada-Stacking 模型也达到了 0.902，相比于传统 Stacking 提升约 5%。实验结果说明，本文所提出的 Ada-Stacking 融合模型有效的提升了分类模型在样本不均衡条件下对正样本的预测能力，提升了泛化性能，从而更加精准实现了对 5G 资费套餐潜在客户的挖掘。

关键词：5G；潜在客户挖掘；分类模型；Ada-Stacking

Abstract

With the rapid development of mobile communication technology, people's life has entered an era of rapid change. In recent years, with the rapid development of the fifth generation communication technology, whether it is relying on ultra-low delay telemedicine, automatic driving and other industry applications, or based on high-speed mobile broadband HD live broadcast, online seconds to kill and other life scenes, gradually commercial 5G is changing people's lifestyle in all aspects. Therefore, the promotion of 5G services and the transformation of 4G package users to 5G have also become the focus of the development of major operators. However, due to the immature development of 5G technology and the lack of clear user needs, the existing 5G package user base is small and the development is slow. To solve the above problems, operators need to explore the characteristics of 5G package users, and carry out targeted marketing for potential 5G users in the existing customer base to promote them to replace 5G tariff packages.

In order to realize the potential customer mining of 5G packages, this thesis has carried out the following research work: (1) At the business level, based on the original data, the data exploration is conducted, and the characteristics of potential customers in terms of consumption behavior are summarized. (2) At the model level, in order to enhance data quality and improve model prediction ability, this thesis carried out corresponding data preprocessing and feature engineering operations on the data set. In view of the overfitting of the model caused by the low proportion of 5G package customers in the business data set and the insufficient classification ability of positive samples, this thesis proposes a potential customer identification model based on the Ada-Stacking fusion idea. The fusion model refers to the sample weight iteration method of AdaBoost algorithm in the base model layer, and performs k rounds of iteration on a single base model, so that each round of model pays more attention to the misdivided samples of the previous round. For the metamodel layer, the model also enrichis the metamodel layer features by cross-multiplying the metamodel features to solve the problem of model overfitting. (3) In order to balance model performance and training time, this thesis evaluates the iteration times k of the Ada-Stacking base model in combination with indicators such as f1-score and training time. It is finally determined that the model's training time and forecasting ability are optimal when k=5.

After model evaluation and optimization, compared with traditional Stacking models, the

Ada-Stacking fusion model not only improves the overall performance indicators of models such as AUC by 1 to 2 points in the test set, but also increases the recall rate of positive sample forecasting ability to 0.906. There are 5 points of improvement in Stacking compared with traditional stacking. The experimental results show that the Ada-Stacking fusion model proposed in this thesis effectively improves the classification model's ability to predict positive samples under the condition of uneven samples, improves the generalization performance, and enables more accurate mining of potential customers for 5G tariff packages.

Key Words:5G; potential customers mining; classification model; Ada-Stacking

目　录

硕 士 学 位 论 文 I
摘　要 I
Abstract II
图目录 VIII
表目录 IX
1　绪论 1
1.1　研究背景及意义 1
1.1.1　研究背景 1
1.1.2　研究意义 2
1.2　研究内容及方法 2
1.2.1　研究内容 2
1.2.2　研究框架 3
1.3　主要创新点 4
1.3.1　特征设计创新 4
1.3.2　模型融合创新 5
2　研究现状与理论基础 6
2.1　研究现状 6
2.1.1　基于营销策略优化的潜在客户识别研究 6
2.1.2　基于运营数据挖掘的潜在客户识别研究 7
2.1.3　基于文本数据分析方法的潜在客户识别研究 10
2.1.4　文献评述 11
2.2　相关理论与技术 12
2.2.1　数据挖掘理论 12
2.2.2　集成学习 13
2.2.3　特征工程理论与主要方法 17
2.2.4　样本增强技术 19
2.2.5　模型评价指标 22
3　潜在用户特征分析与数据预处理 24
3.1　数据说明 24
3.2　用户特征分析 25
3.2.1　样本标签分析 25

3.2.2 特征相关性与重要性分析 25
3.2.3 重要特征分析 28
3.3 数据预处理与特征工程 34
3.3.1 缺失值与异常值处理 34
3.3.2 特征编码与变换 36
3.3.3 样本增强 37
3.3.4 衍生特征设计 38
4 潜在客户挖掘模型的构建与评估 41
4.1 模型设计思路与架构 41
4.2 模型评估 43
4.3 模型构建 43
4.3.1 基于单一模型的潜在客户识别模型构建 43
4.3.2 基于 Ada-Stacking 融合模型的潜在客户挖掘研究 44
4.4 模型评估与优化 49
4.4.1 模型效果评估及特征分析 49
4.4.2 基模型训练层数优化 51
5 研究结论与展望 53
5.1 主要研究结论 53
5.2 不足与展望 54
参考文献 55
致 谢 58

TABLE OF CONTENTS

1 Introduction 1
1.1 Research background and significance 1
1.1.1 Research background 1
1.1.2 Research significance 2
1.3 Research contents and methods 2
1.3.1 Research contents 2
1.3.2 Research framework 3
1.3.3 Major innovations 4
1.2 Literature review 6
1.2.1 Research on potential customer identification based on marketing strategy optimization 6
1.2.2 Research on potential customer identification based on data mining technology 7
1.2.3 Research on potential customer identification based on text mining method 10
2 Related theory and technology 12
2.1 Data mining theory 12
2.2 Ensemble learning 13
2.2.1 Bagging 14
2.2.2 Boosting 15
2.2.3 Stacking 17
2.3 Feature Engineering 17
3 Data exploration and feature engineering 24
3.1 Data specification 24
3.2 Exploratory data analysis 25
3.2.1 Sample label analysis 25
3.2.2 Feature correlation and significance analysis 25
3.2.3 Concrete feature analysis 28
3.3 Data preparation and preprocessing 34
3.3.1 Missing and outlier processing 34
3.3.2 Feature coding and transformation 36

3.3.3 Sample enhancement 37
3.4 Derived feature design 38
4 Construction and evaluation of potential customer mining model based on Ada-Stacking fusion model 41
4.1 Model design thinking and architecture 41
4.2 Model evaluation index 43
4.3 Model construction 43
4.3.1 Lead customer identification model based on a single model ... 43
4.3.2 Research on Potential customer mining based on Ada-Stacking fusion model 44
4.4 Model evaluation and optimization 49
4.4.1 Longitudinal comparison of the effect of multiple classification models49
4.5 Feature importance analysis 50
4.4.2 Research and optimization of training layers of base model 51
5 Research conclusion and prospect 53
5.1 Main research conclusions 53
5.2 Deficiency and prospect 54
reference 55

图目录

图 1.1 研究框架图 4
图 2.1 Bagging 算法示意图 14
图 2.2 随机森林算法示意图 15
图 2.3 AdaBoost 算法示意图 16
图 2.4 Stacking 算法示意图 17
图 2.5 随机过采样示意图 19
图 2.6 传统 Smote 算法示意图 20
图 3.1 样本标签统计饼图 25
图 3.2 特征相关性系数矩阵热力图 26
图 3.3 原始特征与标签相关性统计条形图 27
图 3.4 随机森林模型特征重要性条形图 27
图 3.5 用户套餐价值特征在不同样本中分布箱线图 28
图 3.6 用户连续三个月 arpu 在不同人群中的分布箱线图 29
图 3.7 宽带使用类特征相关系数矩阵热力图 32
图 4.1 建模流程示意图 41
图 4.2 模型设计思路流程图 42
图 4.3 基于传统 Stacking 算法的 5G 套餐潜在客户挖掘模型流程图 46
图 4.4 基于 Ada-Stacking 的 5G 套餐潜在客户挖掘模型示意图 48
图 4.5 Ada-Stacking 融合模型特征重要性统计 51
图 4.6 Ada-Stacking 在不同基模型层数上的分类效果对比图 52

表目录

表 3.1 样本标签统计表 25
表 3.2 用户流量/语音超套金额均值及在总套餐价值中占比均值 31
表 3.3 5G/4G 套餐客户办理宽带带宽档位统计 33
表 3.4 原始数据集中各特征缺失值统计 35
表 3.5 原始数据集中异常值与缺失值预处理方法 36
表 3.6 “星级”特征编码前后对比 37
表 3.7 “细分市场”特征编码前后对比 37
表 3.8 三种样本增强方式在 5 折交叉验证下的模型效果对比 38
表 3.9 衍生特征设计-重要特征定义 39
表 4.1 基于单一模型的潜在客户挖掘模型分类效果对比 44
表 4.2 传统 Stacking 模型在训练集和验证集上分类效果对比 46
表 4.3 Ada-Stacking 与其他模型在测试集上的分类效果对比 49
表 4.4 重要特征在 Ada-Stacking 基模型层上的触发情况 50

1 绪论

1.1 研究背景及意义

1.1.1 研究背景

近年来，社会转型加速，国家正在加强培育数据要素市场、推进治理体系现代化、推进新型基础设施建设，致力打造全新智慧城市，5G 网络的大规模连接能力、高速率传输能力正是智慧城市建设的有力支撑。为贯彻市场导向和用户导向的发展理念，建立健康、可持续的业务生态系统，各大运营商逐渐将推广 5G 资费套餐作为现阶段的发展重点。

(1)5G 技术的发展历程与现状

5G 技术是第五代移动通信技术的简称，它在 4G 技术的基础上做出了重大的创新和改进。在传输速度、时延、数据容量、连接密度、网络切片等方面都有了巨大的提升，这使得 5G 技术得到了广泛的关注和应用。

2013 年 4 月，在工信部、发展改革委、科技部的共同支持下，我国成立 IMT-2020(50)推进组，对 5G 的关键技术和发展应用进行研究，推动国内 5G 技术的发展。2019 年 6 月工信部对中国广电、中国电信、中国联通和中国移动颁布 4 张 5G 牌照，10 月工信部与国内三大电信运营商举行 5G 商用启动仪式，并且于 11 月 1 日正式上线 5G 套餐，自此我国正式进入 5G 商用时代。

随着国家政策举措上的大力扶持和技术上的成熟，5G 技术开始走入民生。以中国移动公司的用户数据为例，根据公司披露的运营数据显示，2021 年上半年 5G 套餐用户数量不断增长。截至 2021 年 12 月，我国已建成 5G 基站超过 115 万个，占全球范围内所有基站的七成以上，我国目前的 5G 终端已有 4.5 亿的用户，占全球所有用户的八成以上。可以看出我国 5G 技术的水平和业务的发展已经居世界的前列。然而，由于 5G 技术发展尚处于起步阶段，基础设施仍不完善，建设成本较高，现有 5G 套餐用户较少，且大多数用户对更换 5G 套餐仍持观望态度。

(2)运营商的 5G 资费套餐推广困境

5G 具有高可靠、低时延、大带宽等特性，可高效将城市系统和服务打通、集成，提升资源运用效率，优化城市管理和服务，改善市民生活质量。加快 5G 用户增长与城市发展深度融合，通过信息化手段解决城镇化过程中带来的问题，既是城市可持续发展所需，也是产业新动能所在。

因此，通过推广 5G 资费套餐，运营商可以在技术、市场和财务层面取得多方面的