北京交通大学 海滨学院

学士学位论文

基于机器学习的电信行业用户个性化套餐的分析与预测

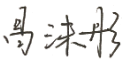
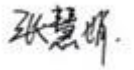
姓　　名	昌沫彤
学　　号	18854003
学　　院	计算机与信息技术学院
专　　业	电子商务
指导教师	张慧娟
职　　称	讲师

二零二二 年 六 月 一 日

学位论文原创性声明

本人所提交的学位论文《基于机器学习的电信行业用户个性化套餐的分析与预测》，是在导师的指导下，独立进行研究工作所取得的原创性成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对本文的研究做出重要贡献的个人和集体，均已在文中标明。

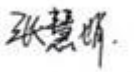
本声明的法律后果由本人承担。

论文作者（签名）：  　　指导教师确认（签名）： 

2022 年 6 月 1 日　　2022 年 6 月 1 日

学位论文版权使用授权书

本学位论文作者完全了解北京交通大学海滨学院有权保留并向国家有关部门或机构送交学位论文的复印件和磁盘，允许论文被查阅和借阅。本人授权北京交通大学海滨学院可以将学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或其它复制手段保存、汇编学位论文。

论文作者（签名）：  　　指导教师（签名）： 

2022 年 6 月 1 日　　2022 年 6 月 1 日

摘　　要

随着互联网不断融入日常生活并成为人们娱乐社交生活不可或缺的一部分，用户对电信套餐的需求越来越多样化与个性化。但是电信套餐种类的增多给用户套餐选择带来极大的困难。因此，学习用户历史消费习惯信息，通过智能化推荐技术为用户推荐适合其个人偏好的电信套餐变得愈发重要，智能套餐推荐也能为电信运营商的服务质量和创收带来巨大提升。

本文通过研究国内外相关文献的推荐算法模型，提出基于机器学习的电信行业存量用户智能套餐推荐模型。首先，本文对基于机器学习的套餐推荐模型进行了可行性分析，并介绍了机器学习算法及决策过程。其次，对用户选择智能套餐的影响因素特征值进行提取，建立了样本数据，使用样本数据训练逻辑回归、随机森林、LightGBM 模型，衡量每个模型完成电信行业存量用户智能套餐推荐任务的性能。实验结果显示 LightGBM 模型完成智能套餐推荐的性能最优，用时最短，准确率达到 98%。最后，使用训练后的 LightGBM 模型完成电信行业存量用户智能套餐的推荐，从而为用户提供个性化、适合的套餐。本文提出的推荐模型能为向电信行业存量用户推荐合适其个人偏好的智能套餐任务提供支撑，同时可以提高电信行业服务质量与套餐销售利润。

关键词：电信用户；个性化套餐；LightGBM；逻辑回归；随机森林

ABSTRACT

As the Internet continues to integrate into daily life and become an indispensable part of people's entertainment and social life, users' demand for Telecom packages is becoming more and more diversified and personalized. The increase in the types of packages brings great difficulties to telecom users. Therefore, it becomes more and more important to learn users' historical consumption habits and recommend Telecom packages suitable for their personal preferences through intelligent recommendation technology. Intelligent Package recommendation can also greatly improve the service quality and income generation of telecom operators.

By studying the recommendation algorithm model of relevant literature at home and abroad, this paper proposes an intelligent package recommendation model for stock users in the telecommunications industry based on machine learning. Firstly, this paper analyzes the feasibility of the package recommendation model based on machine learning, and introduces the machine learning algorithm and decision-making process. Secondly, the eigenvalues of the influencing factors of users' selection of intelligent packages are extracted, and the sample data is established. The sample data is used to train the logistic regression, random forest and lightgbm models to measure the performance of each model in completing the recommendation task of intelligent packages for stock users in the telecommunications industry. The experimental results show that the lightgbm model has the best performance of Intelligent Package recommendation, the shortest time, and the accuracy rate is 98%. Finally, the trained lightgbm model is used to complete the recommendation of intelligent packages for stock users in the telecommunications industry, so as to provide users with personalized and suitable packages. The recommendation model proposed in this paper can support the task of recommending intelligent packages suitable for their personal preferences to the stock users in the telecommunications industry, and can improve the service quality and package sales profit of the telecommunications industry.

Key words: Telecommunication users; Personalized package; LightGBM; Logistic regression; Random forest

目　　录

第 1 章　绪论

随着国内电信市场的逐渐成熟和技术的快速发展，电信运营商之间的竞争日趋激烈。为了给用户提供更优质的产品和服务，电信运营商需要在竞争状态下盈利，因此必须重视与用户利益密切相关的电信套餐和资费设计方案。因此，是否拥有先进的技术、高性价比的套餐产品、个性化的套餐设计和完善的售后服务成为电信运营商的核心竞争力。

1.1　研究背景和意义

随着网络日益深入到社会经营和生活中，人们也开始逐渐习惯了各种内容丰富的网络娱乐社交生活，而这种网络娱乐社交生活也形成了大量的网络信息内容交换行为，包括了购物记录、社交网络、以及历史导航路径，并由此形成了海量的消费数据分析信息内容。而这些数据分析信息中包含着很多历史特征信息，比如用户的历史消费，以及生活基本信息等，而在这里面也包含了用户的历史消费意图，也蕴藏着用户的潜在历史消费潜力。在这种历史背景下，随着客户的历史消费心智日益成熟，其可供选择的服务套餐内容也就越来越繁冗复杂，这引起了运营商们越来越关注于客户服务空间的历史消费。而与此同时，各大运营商也把主要注意力集中在服务产品上，不断地推出了各种满足用户需求的业务套餐和服务产品组合，为广大人民的信息生活提供着各种便利服务。

1.1.1　课题研究背景

随着套餐数量越来越多，从而导致用户购买更加麻烦。而按照各大运营商的调查数据，在所有的套餐体系中，已不能直接再购买的套餐以及已废弃的套餐都只有全部套餐的百分之二十五，因此消费者往往无法快速的在数量很多的套餐中选出最符合自身的套餐，而只有按照自身的实际使用情况不断的改变自身的套餐结构，才能选出最符合自身的套餐。而过于繁琐的套餐结构往往存在着很大的误导性，因此消费者总是需要投入很多的资金筛选适合自身需要的套餐选择。结果，消费者往往在选定最符合自身的套餐之后，就毫无积极性的再购买其他的套餐了。从而导致用户购买套餐时更加繁琐，而运营商耗费巨大的时间打造的新套餐也常常无人问津。

电信运营商面对着客户忠诚度问题，必须重新寻求出路。同时由于各类电信套餐的大量涌现，消费者挑选合适自身的套餐更加艰难，所以，各大电信运营商同时也面临着众多的即将离网的新客户，消费者也对运营商的服务质量更加不满意。随着中国电信用户的日趋饱和，运营商也发现，存量运营将是公司未来增长的主要趋势。因此运营商要实现存量运营，先要留住自己的客户，而不要转给别的运营商，给客户带来更好的业务，从而沉淀

客户。而在此基础上，借助客户的消费资料和消费数据充分挖掘客户的消费潜力，通过持续的完善服务和打造更符合消费者的套餐产品，公司才能发展的更加好。

套餐选择是十分关键的，由于中国电信套餐的数量愈来愈大，信息日益丰富，给消费者造成了套餐选择的烦恼。由于运营商没有主动的面向特定消费者开展个性套餐介绍，导致用户无法挑选合适自己的套餐。所以，构建套餐选择模式，为消费者建立个性化、智能的套餐选择模式就变得十分关键，在丰富的历史数据中抽取客户的特征集，实时通过客户行为分析用户特征，构建科学合理的中国电信套餐推荐模式，主动、准确的为消费者介绍最合适的中国电信套餐品牌，使消费者的选购更加轻松和便捷。综上所述，在信息通信领域的大数据分析背景下，运营商通过从以前的套餐管理到以后的套餐服务的全面改革，将为消费者提供更加个性化、智能化的信息通信套餐服务，使消费者的使用生活更加简单和便捷。

1.1.2 研究意义

在当前的大数据分析环境下，用户的实际消费数据以及历史消费数据为中国电信套餐推荐模式的研发奠定了极为完整的大数据分析基石，同时也是推动套餐业务更加个性化、智能化的关键方法。所以，该研究有着很大的价值。

（1）进一步扩大了个性化服务方式和套餐服务的运用，进一步增加了对服务产品的分析层次。近年来，由于网络的大众化以及电商的爆发式成长，用户消费推广模式在电商行业的运用日益广泛，不过在电信行业的运用还没有很普遍。本研究不但扩大了个性化推广的领域，更是提高了消费者满意度，帮助消费者更快速更有效的挑选合适自身的套餐服务。

（2）用户能够更快的挑选最合适自己的套餐，通过对用户的实际消费数据和历史消费数据进行综合分析，归类研究，从而形成了电信套餐推广模式，根据电信运营商所提交的套餐数量，采用规则筛选和配对方法进行匹配，从而得出在未来一段时间内比较适宜使用的套餐列表，并向用户推荐比较适宜使用的但是还尚未购买的套餐。这不但充分挖掘了客户的消费潜力还为客户创造了多样化和最适宜的套餐组合，极大的方便了消费者[1]。

（3）通过持续推动运营商的新服务套餐推出，以提高用户的服务满意度，在当前网络日趋激烈竞争的市场环境下，唯有通过持续提高服务水平才能增强用户忠诚度，所以，通过根据用户实际消费的数据分析构建套餐推广模型后，为用户提供更为精准和人性化的服务就显得更为关键了。另外，在传统情况下，新套餐的宣传方式大多都是采用电视广告、路牌、营业厅等宣传方法。如果采用了这样的宣传方法，当运营商对新套餐进行宣传后，就得采用网络广告、路牌等方法宣传。结果是成本十分昂贵，且市场发展十分迟缓。在通过实际消费数据分析建立了新套餐推广模型以后，再通过对新套餐的服务属性进行筛选，以获得模型中所要求的服务特征值，从而建立服务特征映射。通过将与宣传套餐的属性信息进行配对后，就能够明确了目标用户，并采取了短信、微信、电子邮件等方法进行了宣

传活动。既能够增加了用户满意度，又快捷准确的宣传了套餐，从而极大的减少了公司宣传的成本，也增加了目标用户的满意度。

1.2 国内外研究现状

上世纪九十年代，由于网络的高速发展，网络上的数据日益增加，这也就迫在眉睫的要求解决信息筛选与信息查询等问题。就是在这个历史背景下，网络与计算机技术专家们开始研发个性化推荐技术，这样能够迅速的从网络获得信息。由于电商的爆发式成长，该技术迅速地运用在电商行业中。通过以用户的历史购物记录以及相似用户的购物记录为基本特征，构建商品推荐模式。给用户介绍最合适的商品，便利了用户的选购，增加了网络的销量。当前，许多门户网站、电子商务平台、各类社区网络平台等均大量使用了这项技术。目前，介绍算法大致有分成两类：采用协同过滤方式的介绍模式，和基于内容的介绍模式。采用协同过滤的推荐模型能够依据用户的消费历史数据，进行关联计算向用户提供适合于其历史消费习惯的商品，同时采用协同过滤的推荐模型还能够确定和用户存在同样消费习惯的用户，将此类用户所选择的商品直接推送给用户。基于内容的推荐模型，基于用户历史消费数据，使用半监督学习的方式为用户推荐与其历史消费产品相似的产品，进而完成推荐任务。

随着人们对互联网上的日常生活娱乐社交越发关注与在意，用户对电信套餐和上网服务的需求与服务渴望不断提升，相应地电信套餐提供的上网服务也越来越多样化，但是电信套餐数量的提升同时也给用户套餐选择上带来极大的困难，用户容易被大量的套餐信息淹没而不知道自己该选择哪种电信套餐以满足自己的上网需求。因此，大量学者提出电信套餐智能推荐模型，以满足用户对上网套餐的定制化推荐需求。例如，DAOUD 等人从客户网络要求的细分视角入手，利用聚类算法对客户的网络业务要求做出了细致的划分，对不同的客户推荐满足不同需求的电信套餐[2]。Yu 等采用协同过滤的方法，进行客户的套餐或套餐对客户的选择，只是简单地通过套餐或客户间的差异实现套餐选择，而不能进行对客户的套餐数量的判断。顾方婷等利用分析与预测算法来构建套餐升级模型，并根据不同类型的电信套餐，进行了一一建模和与众多套餐变更有关的预测[3]。ABATUROV 等人将信息技术融合和数据挖掘相结合，针对不同客户群，通过不同算法建立了电信客户的流失预警模式[4]。

1.3 论文组织结构

本论文旨在探究通过采用机器学习的电信套餐推广模式，获取用户消费信息与套餐信息，从而获取套餐分类，并构建套餐推广模式，为用户推荐更合适的套餐服务。

本文主要研究内容和结构安排包括 5 部分。

第 1 章，绪论。介绍本文的选题背景、国内外研究现状、研究的意义。除了对推荐的方法做了较为详细的概述之外，还介绍了目前主要的使用方式及应用场景。

第 2 章，相关理论介绍。主要介绍了个性化推荐技术和机器学习算法，对不同的个性化技术进行简介，并介绍了它的实现过程。

第 3 章，电信用户数据的处理及介绍。叙述了电信套餐数据的来源和介绍，对数据进行了预处理操作，包括对缺失值的处理和归一化处理，特征分类及数据的变量相关项进行统计描述，从而进行特征选择。通过对电信用户与电信产品的属性梳理，建立电信套餐推荐模型的训练数据，为下文的套餐和用户消费数据匹配提供了数据基础[5]。

第 4 章，电信行业用户套餐的分析与预测。引入了基于机器学习算法的电信套餐推荐模式。本章内容通过收集客户的消费信息，提炼特征值，建立电信套餐推荐模型，把个性化的选择技术运用于电信套餐选择场景中，并通过机器学习模型根据推荐结果得到要提供给客户的套餐类别。

第 5 章，总结与展望。对论文的科研背景，研究的主要内容，实验研究结果和试验结论等都作了剖析和总结，并扩展了文章中出现的不足以及今后的进一步研究方向，对文章的科研结果也作了简要总结，并指出了文章中今后可供进一步深入研究的新领域。

第 2 章　相关理论介绍

电信套餐个性化推荐是通过分析目标用户的消费行为特征，帮助用户在信息过载或漫无目的的浏览时找到自己感兴趣的套餐信息，进而推荐套餐，提高套餐与用户的匹配精度，实现用户需求的精准定位。

2.1　逻辑回归

在提出逻辑回归算法前，首先介绍一下 Sigmoid 函数，其数学表达式如式（2-1）所示。

$$g(x)=\frac{1}{1+e^{-x}} \tag{2-1}$$

Sigmoid 函数是一个 s 型曲线，它的图像如图 2-1 所示，它的取值区间为[0,1]，当自变量 x 趋向于正无穷时，因变量 $g(x)$趋向于 1，而当 x 趋向于负无穷时，$g(x)$趋向于 0。在二分类任务中，采用 Sigmoid 的输出的是事件概率，也就是当输出满足某一概率条件将其划分正类，否则划分为负类[6]。

逻辑回归实际上就是将预测范围进行缩小，限定在[0,1]区间的一种回归模型。逻辑回归模型假定因变量 y 服从伯努利分布。在分类情况下，逻辑回归分类器其实就是一组权值 θ，当有测试样本输入时，这组权值与测试数据按照加权求得

$$h_\theta(x)=\theta_0+\theta_1x_1+\theta_2x_2+\cdots+\theta_nx_n \tag{2-2}$$

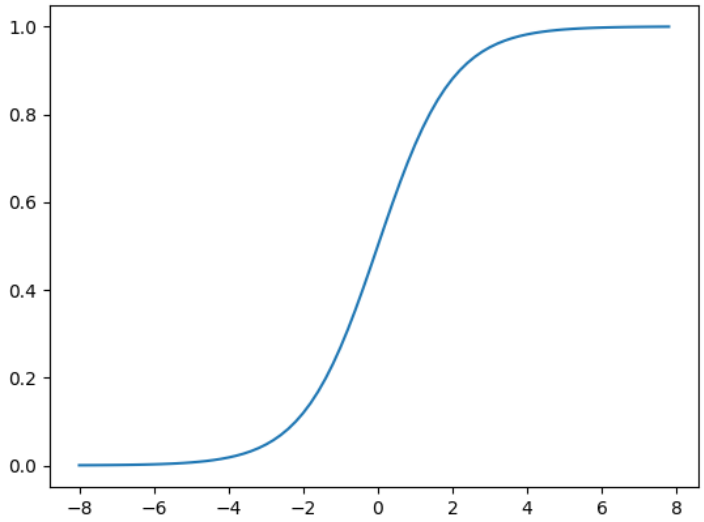


图 2-1　Sigmoid 函数图像

在逻辑回归模型中，假设函数表示为式（2-3）。

$$h_\theta(x)=g\left(\theta^T x\right) \tag{2-3}$$

当假设函数 $h_\theta(x)\geq 0.5$，预测成正类，反之预测成负类。在此模型中，因变量是一个二元型的数据，所以因变量只有事件发生或事件不发生。通过建立一个函数，对于一个输入的数据，当这个数据大于阈值时，输出 1，小于阀值时则输出 0。逻辑回归模型将因变量 y 的几率（odds）取对数，得到的新的因变量 $\ln\left(p/(1-p)\right)$ 是与自变量 x 线性相关的。其中 p 是指事件出现的概率。

式（2-4）和式（2-5）为逻辑回归的代价函数。

$$J(\theta)\frac{1}{m}\sum_{i=1}^{m}\cos t\left(h_\theta\left(x^{(i)}\right),y^{(i)}\right) \tag{2-4}$$

$$\cos t\left(h_\theta(x),y\right)=f(x)=\begin{cases}-\log\left(h_\theta(x)\right), & \text{IF } y=1\\ -\log\left(1-h_\theta(x)\right), & \text{IF } y=0\end{cases} \tag{2-5}$$

对逻辑回归模型问题的求解，有许多的方法，在这里介绍一个牛顿法。牛顿法的基本思路，是在已有极小值的附近对 $f(x)$ 做二阶泰勒展开，从而找到极小点的下一个估计值，假设 ω^k 为极小值估计值[7]，那么有

$$\phi(\omega)=J\left(\omega^k\right)+J'\left(\omega^k\right)\left(\omega-\omega^k\right)+\frac{1}{2}J''\left(\omega^k-\omega^k\right)^2 \tag{2-6}$$

然后令 $\phi'(\omega)=0$，得到了 $\omega^{k+1}=\omega^k-\dfrac{J'\left(\omega^k\right)}{J''\left(\omega^k\right)}$。因此有迭代更新式

$$\omega^{k+1}=\omega^k-\frac{J'\left(\omega^k\right)}{J''\left(\omega^k\right)}=\omega^k-H_k^{-1}\cdot g_k \tag{2-7}$$

其中 H_k^{-1} 为海森矩阵

$$H_{\text{mn}}=\frac{\partial J(\omega)}{\partial\omega_m\partial\omega_n}=h_\omega\left(x^{(i)}\right)\left(1-p_\omega\left(x^i\right)\right)x_m^{(i)}x_n^{(i)} \tag{2-8}$$

2.2 随机森林

随机森林作为常见的传统机器学习模型，其主要通过多个决策树判定同一个任务中的分类来实现的，当多数决策树认为数据归属某一类别时依据少数服从多数原则将该数据判定为多数决策树判定的类别。分析决策树的构建与决策树判定数据类别归属方面可知，生成决策树时，将数据集中所有样本数据的单个特征变量作为决策树的节点，如果单个数据的值大于该特征变量的分界值，则该数据会被分类到节点的左侧分支中去，如果该数据的值小于该特征变量的分界值，则该数据会被分类到节点的右侧分支中去。通过多个特征变量作为分界节点，不断判定单个数据是否大于分界节点的数值，将单个数据划分到分界节点的某一侧分支中去，经过多个特征变量的分界，数据最后被归为两个类别的分支中，分

到哪个类别中数据的预测标签即为该类别。依据数据样本分类结果，计算出样本数据的信息熵，公式为式（2-9）所示。

$$INfo(s)=-\sum_{i=1}^{2}p_i\log_2(p_i) \tag{2-9}$$

其中，pi 表示类别 i 样本数量占所有样本的比例。

当样本数据集中所有样本数据使用特征 X 作为分类节点时，决策树需要判定特征 X 的分裂点用于分类每一条数据。特征分裂点指当数据的具体数值大于该分裂点分界数值时或者小于该分裂点分界数值时均会被分类进入下一个分裂节点。在这里设特征 X 有 k 个分裂点，那么实验数据会被分成为 k 个部分。对应在数据集中，选择特征 X 作为节点时，那么在特征 X 作为分裂节点后，整个数据集的信息熵变化程度为式（2-10）所示。

$$INfo_x(s)=-\sum_{i=1}^{k}\frac{|S_i|}{|S|}\times INfo(si) \tag{2-10}$$

当数据集经过特征 X 的 k 个分裂点划分为不同子数据集后，统计数据集的信息增益情况，信息增益 G（X）表示数据集在特征 X 的作用后，其信息熵减少的值[8]。公式为（2-11）所示。

$$G(X)=INfo(S)-INfo_x(S) \tag{2-11}$$

其中，$G(X)$ 最大时的特征 X 的分裂点取值，就是对于样本数据集 S 中所有样本数据在决策树分裂节点 X 下最合适的分裂边界。

随机森林模型在使用样本数据集训练多个决策树的过程中，通过有放回的多次抽样产生 N 个子样本作为训练决策树模型用到的训练集[9]。若输人数据集中所有数据一共出现的特征中有 M 个，可以通过设定 $m(<M)$ 个特征用于生成决策树的分裂节点，通过判定多个决策树分裂节点的边界值完成对单个数据在该分裂节点时的分支判定，这 m 个分裂节点特征作为形成决策树模型过程中的多个最佳分裂点并依据每个分裂点边界值判定单一数据最终的类别归属情况。重复上述决策树生成过程，特征决策点形成过程、决策点分裂边界判定过程 P 次，随机森林模型能够形成 P 棵决策树。然后通过这 P 棵决策树为每个数据的类别进行投票判定，然后利用多数服从少数的群体智慧投票机制为每个数据分配最终的归属类别。

2.3 LightGBM

GBDT（Gradient Boosting Decision Tree）是机器学习中一个长盛不衰的模型，其主要思想是利用弱分类器（决策树）迭代训练以得到最优模型，该模型具有训练效果好、不易过拟合等优点。GBDT 不仅在工业界应用广泛，通常被用于多分类、点击率预测、搜索排序等任务；而 LightGBM（Light Gradient Boosting Machine）是一个实现 GBDT 算法的框架，

支持高效率的并行训练，并且具有更快的训练速度、更低的内存消耗、更好的准确率、支持分布式可以快速处理海量数据等优点[10]。

常用的机器学习算法，例如神经网络等算法，都可以以 mini-batch 的方式训练，训练数据的大小不会受到内存限制。而 GBDT 在每一次迭代的时候，都需要遍历整个训练数据多次。如果把整个训练数据装进内存则会限制训练数据的大小；如果不装进内存，反复地读写训练数据又会消耗非常大的时间。尤其面对工业级海量的数据，普通的 GBDT 算法是不能满足其需求，LightGBM 提出的主要原因就是为了解决 GBDT 在海量数据遇到的问题，让 GBDT 可以更好更快地用于工业实践[11]。

2.4 本章小结

本章节主要对电信行业用户个性化套餐推荐所需的相关理论进行了说明。对本文所用到的逻辑回归模型、随机森林模型和 LightGBM 的概论与原理进行了介绍。逻辑回归算法较易理解与实现，但存在着分类精度不够高的缺点。随机森林算法准确度较高，且不容易陷入过拟合。LightGBM 模型的训练速度和内存方面都很好。

第 3 章　电信用户数据的处理及介绍

电信业在我国工业总量中具有举足轻重作用，对我国发展的各个方面提供了巨大贡献。在网络高速发展的背景下，电信用户套餐的使用流量也在急剧上升。近年来，为更好的适应不同用户对套餐类型需要，运营商开发了各种套餐产品供用户选购。在运营商发布数量众多的套餐信息后，消费者又面对了一种无法抉择的困难。怎样在众多套餐服务中挑选一种合适自身的套餐显得十分困难。倘若通信运营商可以在大量消费者信息中建立合理的套餐选择模式，将使通信运营商在电信业竞争中取得重要的竞争地位。

3.1　数据来源及介绍

在本文中，电信套餐主要指电信运营推出的针对各种需要的客户而推出的服务产品和附加业务。电信套餐信息通常主要包括最基本的电话、短信等功能、还有其它的如来电显示、自定义铃声等，主要由服务套餐内容以及流量大小、国际漫游和其它的非主营业务套餐等构成。本文所采用的套餐信息集主要包括了满足基本要求的主套餐。

本文所选数据集来自中国计算机协会与中国联通研究院发布的公开比赛数据集（https://www.datafountain.cn/competitions/311/datasets），该数据集包含电信行业存量用户使用的智能套餐类型，用户的消费情况，套餐内容，不包含个人隐私的用户基本信息。数据集包含超过 74 万条数据，并且每条数据包含字段、中文名、数据类型，如表 3-1 所示。

观察表 3-1 可知，每条数据集包含用户 ID，套餐类型，是否固移融合套餐，在网时长，当月总出账金额_月，当月前 1 月总出账金额_月，当月前 2 月总出账金额_月，当月前 3 月总出账金额_月，当月累计-流量，连续超套，合约类型，合约时长，是否承诺低消用户，网络口径用户，交费次数，交费金额，上月结转流量，月累计-本地数据流量，本地语音主叫通话时长，套外主叫通话时长，Service2_caller_time，性别，年龄，投诉重要性，交费金历史投诉总量，历史执行补救费用交费金额，一共 26 项字段。数据类型分为连续型数据和离散型数据，连续型数据主要是数值形式，离散型数据以离散数字形式表征[12]。

大部分字段都有文字说明，用于解释该字段的内在含义与表示内容，用户 ID 字段用于标识用户；融合套餐字段中 1 代表是 0 代表否；当月总出账金额字段、当月前 1 月总出账金额字段、当月前 2 月总出账金额_月字段、当月前 3 月总出账金额_月字段均为数值变量，单位为元；当月累计-流量字段的单位为 MB；连续超套字段中 1 代表是 0 代表否；是否承诺抵消用户字段中 1 代表是 0 代表否；交费次数字段、交费金额字段、上月结转流量字段、月累计-本地数据流量字段、本地语音主叫通话时长字段、套外主叫通话时长字段、Service2_caller_time 字段的数据单位分别为次、元、MB、MB、分钟、分钟、分钟；性别

字段中 01 代表男、02 代表女；投诉重要性字段中 1 代表普通、2 代表重要、3 代表重大；交费金历史投诉总量字段、历史执行补救费用交费金额字段的数据单位分别为次、分。

表 3-1　数据字段说明

字段	字段名称	说明
USERID	用户 ID	标识用户的唯一字段
service_type	套餐类型	/
is_mix_service	是否固移融合套餐	1.是 0.否
online_time	在网时长	/
1_total_fee	当月总出账金额月	单位：元
2_total_fee	当月前 1 月总出账金额月	单位：元
3_total_fee	当月前 2 月总出账金额月	单位：元
4_total_fee	当月前 3 月总出账金额月	单位：元
month_traffic	当月累计-流量	单位：MB
many_over_bill	连续超套	1-是，0-否
contract_type	合约类型	ZBG_DIM.DIM_CBSS_ACTIVITY_TYPE
contract_time	合约时长	/
is_promise_low_consume	是否承诺低消用户	1.是 0.否
net_service	网络口径用户	20AAAAAA-2G
pay_times	交费次数	单位：次
pay_num	交费金额	单位：元
last_month_traffic	上月结转流量	单位：MB
local_trafffic_month	月累计-本地数据流量	单位：MB
local_caller_time	本地语音主叫通话时长	单位：分钟
service1_caller_time	套外主叫通话时长	单位：分钟
service2_caller_time	Service2_caller_time	单位：分钟
gender	性别	01.男 02 女
age	年龄	/
complaint_level	投诉重要性	1：普通，2：重要，3：重大

续表 3-1　数据字段说明

字段	字段名称	说明
former_complaint_num	交费金历史投诉总量	单位：次
former_complaint_fee	历史补救费用交费金额	单位：分

3.2　特征分类

电信产品及服务电信客户的消费是一种持续累积的过程，而消费者除了购买与定制，还有后续的商品及增值服务的使用与反馈等活动，包括使用短信、网络、人工客户服务中心等渠道。在该套餐推向市场的期间，电信运营商会接收到各种反馈信息，将这些信息作为基本数据存储在运营商的数据库系统中，运营商能够利用这些信息评价该套餐的特点，并且提供该套餐的用户所对应的特征集，便于分析不同信息对电信行业存量用户智能套餐选择的影响。

3.2.1　用户基本信息

用户基础信息的收集，重点主要集中在用户个人的基本个人信息方面以及个人用户的消费行为等基础信息方面。电信用户的基础个人信息，通常由用户基本属性、用户所属的套餐属性等信息构成。在个人用户基础属性方面，主要是指个人用户的基础社会属性，比如个人用户的年龄、性别、号码以及所属地等基础个人信息。而终端用户身份属性信息方面，则主要是指用户在销售账务体系中的基本身份特性以及消费人群分类，主要按照普通大众用户会员以及高级 VIP 户进行基本分类。在此基础上还可能进行其它维度分类，主要包括普通个人用户、企业用户等。这些用户通常是对有着共同特征的电信用户的某一个标签，通过标签划分，能够更便于电信运营商针对性的展开各类营销与宣传活动。

根据对用户基础信息的定义，本文从表 3-1 中提取用户基础信息字段如表 3-2 所示，便于分析用户基础信息对用户套餐选择的影响。观察表 3-2 可知，每条数据集包含用户 ID，性别，年龄，投诉重要性，交费金历史投诉总量，历史执行补救费用交费金额，一共 6 项字段。用户 ID 标识用户，性别、年龄作为用户的基本特征是用户基本信息不可或缺的部分。投诉重要性可反应用户的投诉重要度相关信息，对套餐推荐和分析用户文化背景有重要作用。交费金历史投诉总量，历史执行补救费可以反映用户在套餐推荐过程中的耐心度与用户性格方面特征，是用户基本信息的重要组成部分。因此，本文选择上述 6 项字段组成用户基本信息具备一定的合理性。

表 3-2　用户基本信息

字段	字段名称	说明
USERID	用户 ID	用户编码，标识用户的唯一字段
gender	性别	01.男 02 女
age	年龄	/
complaint_level	投诉重要性	1：普通，2：重要，3：重大
former_complaint_num	交费金历史投诉总量	单位：次
former_complaint_fee	历史补救费用交费金额	单位：分

3.2.2　用户套餐内容消费信息

在使用者的套餐内容等消费信息中，可发现使用者的消费水平级别、对套餐偏好、兴趣爱好、消费习惯、消费渠道等其它信息内容，以及使用者所选的套餐品牌、已开通的增值服务、实际消费的数额等信息内容，在这部分信息内容中能很好的挖掘出使用者的消费潜力。用户消费行为信息内容的收集整理，主要包含了套餐品牌服务信息内容。从电信运营商发布的套餐品牌、用户所使用的套餐、使用者预订的服务，这部分信息内容都直观反映了使用者的消费行为信息内容，都给企业个性化、智能化推广及套餐模式的构建带来了十分关键的借鉴意义，也可以说是企业推广对象筛选的关键指标。

根据对用户套餐内容消费信息的定义，本文从表 3-1 中提取用户套餐内容消费信息字段如表 3-3 所示，便于分析用户套餐内容消费信息对用户套餐选择的影响。观察表 3-3 可知，每条数据集包含是否固移融合套餐，在网时长，当月累计-流量，连续超套，合约类型，合约时长，是否承诺低消用户，网络口径用户，上月结转流量，月累计-本地数据流量，本地语音主叫通话时长，套外主叫通话时长，Service2_caller_time，一共 13 项字段。套餐内容消费，上月结转流量，月累计-本地数据流量，本地语音主叫通话时长，套外主叫通话时长，Service2_caller_time 作为套餐消费的具体程度信息可反应用户的套餐消费内容，对套餐推荐和分析套餐是否符合用户消费习惯有重要作用。因此，本文选择上述 15 项字段组成用户基本信息具备一定的合理性。

表 3-3　用户套餐内容消费信息

字段	字段名称	说明
is_mix_service	是否固移融合套餐	1.是 0.否
online_time	在网时长	/

续表 3-3　用户套餐内容消费信息

字段	字段名称	说明
month_traffic	当月累计-流量	单位：MB
many_over_bill	连续超套	1-是，0-否
contract_type	合约类型	ZBG_DIM.DIM_CBSS_ACTIVITY_TYPE
contract_time	合约时长	/
is_promise_low_consume	是否承诺低消用户	1.是 0.否
net_service	网络口径用户	20AAAAAA-2G
last_month_traffic	上月结转流量	单位：MB
local_trafffic_month	月累计-本地数据流量	单位：MB
local_caller_time	本地语音主叫通话时长	单位：分钟
service1_caller_time	套外主叫通话时长	单位：分钟
service2_caller_time	Service2_caller_time	单位：分钟

3.2.3　用户缴费信息

用户实际缴费金额信息为电信用户的“消费账单”，该“消费账单”中包括了套餐基本费用以及电信用户实际使用该套餐的次数的总额。电信运营商一般以用户使用该套餐的月份为基础的计量单位，并由此来计算用户的消费总额，比较常用的如各月份累计电话时长、各月份短信累积使用条数、全国漫游话累计持续时长等，但在不同的省、市、地区、县内，各个属性数值也都不一样，比如，电话时长不但包括国内电话时长，还包含了省内电话时长和国外长途通话时间等。

根据对用户交费信息的定义，本文从表 3-1 中提取用户交费信息字段如表 3-4 所示，便于分析用户交费信息对用户套餐选择的影响。观察表 3-4 可知，每条数据集包含当月总出账金额_月，当月前 2 月总出账金额_月，当月前 2 月总出账金额_月，当月前 3 月总出账金额_月，交费次数，交费金额，一共 6 项字段。当月总出账金额_月字段，当月前 1 月总出账金额_月字段，当月前 2 月总出账金额_月字段，当月前 3 月总出账金额_月字段，作为用户交费信息的历史记录和当前记录，可以反映用户消费意愿与消费能力，是用户电信存量套餐推荐过程中不可或缺的分析信息。交费次数，交费金额作为用户交费信息的具体表现形式可反应用户的套餐金额消耗频率与消费程度，对套餐推荐和分析套餐是否符合用户消费习惯有重要作用[13]。因此，本文选择上述 6 项字段组成用户基本信息具备一定的合理性。

表 3-4　用户缴费信息

字段	字段名称	说明
1_total_fee	当月总出账金额月	单位：元
2_total_fee	当月前 1 月总出账金额月	单位：元
3_total_fee	当月前 2 月总出账金额月	单位：元
4_total_fee	当月前 3 月总出账金额月	单位：元
pay_times	交费次数	单位：次
pay_num	交费金额	单位：元

本文首先从原有特征上进行了特征划分，大致分为使用者的基本消费特点、业务选择特点、基本属性特点三方面。基础信息特点涵盖了所有使用套餐的使用者个人基本信息，这部分信息特点也从侧面体现了中国电信使用者最基本的社会特性；业务选择特性一般指的是为使用者所选定的主套餐等增值套餐等，体现了电信运营商重点服务使用者的属性指标和使用者的消费行为倾向；而使用者消费行为特性则一般是根据使用者所订购的套餐服务质量，记录了使用者对该套餐的应用程度，如电话时长（包括省内、国内通话时长等）等，这种特性能直观反映使用者针对中国电信套餐服务质量的应用侧重点和用户需求来源。

3.3　数据处理

在对数据进行模型预测训练前，首先要对原始数据进行处理，保证数据的可用性，这样才能更好的进行模型训练，本文所用电信行业存量用户智能套餐使用情况数据集中缺失值所占比例较少，所以本文对缺失数据进行整体删除。其次，本文完成了对数据的归一化处理操作，使特征具有相同的度量尺度，提高数据预测的准确率。

3.3.1　缺失值处理

关于缺失值数据的处理，最常用的方式一般有删减法和插值法，删减法通过直接清除缺失值数据处理解决问题，而插入方法则通过给缺失值数据中分配一种满足总体分布的新数据进行解决问题，在本文中对缺失值主要采用的删除法，因为本文的缺失数据很少，如图 3-1 所示，可以看出在 2_total_fee 中存在一条缺失数据，3_total_fee 中存在两条缺失数据，gender 中存在 22912 条缺失数据，同时从图 3-2 缺失数据的可视化中可以看出，缺失的数据占总体数据的很小一部分，基本可以忽略不计，所以在本文中直接将缺失数据进行删除。图 3-3 为删除缺失数据后的数据缺失统计。

```
plt.rcParams['font.sans-serif']=['SimHei']
plt.rcParams['axes.unicode_minus']=False
pd.options.display.max_columns = None
```

```
pd.options.display.max_rows = None
plt.rcParams['figure.dpi'] = 150
warnings.filterwarnings("ignore")
data = pd.read_csv('train_all.csv')
print(data.isnull().sum())
plt.figure()
p=msno.bar(data,color='g')
plt.title('可视化数据缺失情况',fontsize=20)
data = data.dropna()
data.isnull().sum()
```

```
service_type                0
is_mix_service              0
online_time                 0
1_total_fee                 0
2_total_fee                 1
3_total_fee                 2
4_total_fee                 0
month_traffic               0
many_over_bill              0
contract_type               0
contract_time               0
is_promise_low_consume      0
net_service                 0
pay_times                   0
pay_num                     0
last_month_traffic          0
local_trafffic_month        0
local_caller_time           0
service1_caller_time        0
service2_caller_time        0
gender                  22912
age                         2
complaint_level             0
former_complaint_num        0
former_complaint_fee        0
current_service             0
user_id                     0
```

图 3-1　数据缺失情况

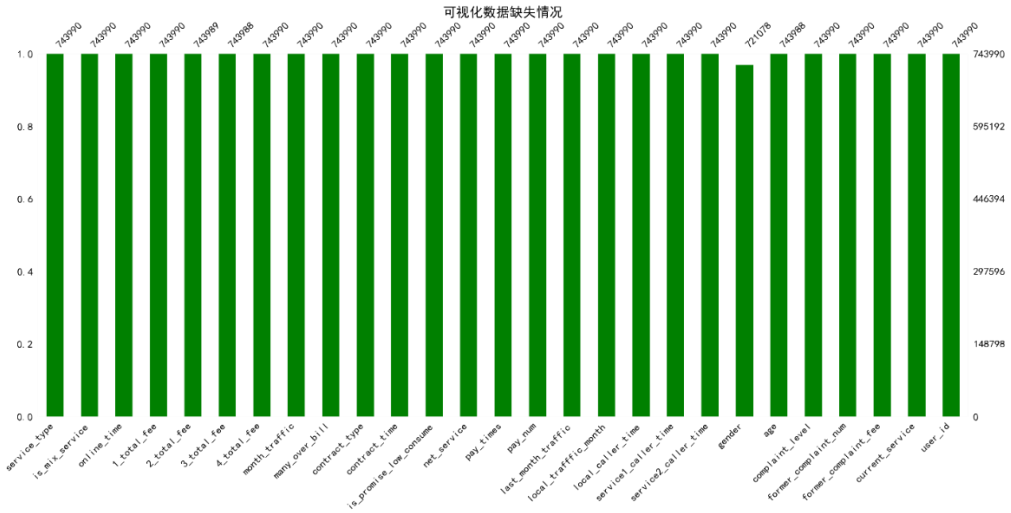


图 3-2　数据缺失可视化

service_type	0
is_mix_service	0
online_time	0
1_total_fee	0
2_total_fee	0
3_total_fee	0
4_total_fee	0
month_traffic	0
many_over_bill	0
contract_type	0
contract_time	0
is_promise_low_consume	0
net_service	0
pay_times	0
pay_num	0
last_month_traffic	0
local_trafffic_month	0
local_caller_time	0
service1_caller_time	0
service2_caller_time	0
gender	0
age	0
complaint_level	0
former_complaint_num	0
former_complaint_fee	0
current_service	0
user_id	0

图 3-3　删除缺失值后的数据缺失情况

3.3.2　数据归一化

将数据进行缺失值的处理之后，还须将全部数据归一化处理，将数据集中后所有数据按百分比进行放缩，从而使全部数据都分布在某个指定的区域范围，一般是在[0,1]区域，从而进一步提高了模型的技准确度和收敛速率，在本实验中采用了 Min-Max 的标准化原理，对所有原始数据进行了线性变换，具体变换如式（3-1）所示，本文采用的是 sklearn 中的 MinMaxScaler()函数进行数据的归一化处理，具体计算如式（3-2）所示。数据归一化处理后的部分数据展示如图（3-4）所示。

$$x' = \frac{x - \min}{\max - \min} \tag{3-1}$$

其中，max 为样本数据的最大值，min 为样本数据的最小值。

$$X_{scaled} = \frac{X - X.\min(axis=0)}{A.\max(axis=0) - X.\min(axis=0)} \times (\max - \min) + \min \tag{3-2}$$

其中，$X.\min(axis=0)$ 是指每列中的最小值组成的行向量；$X.\max(axis=0)$ 是指每列中的最大值组成的行向量；max 是要映射到的区间最大值，默认是 1；min 是要映射到的区间最小值，默认是 0；X_{scaled} 是归一化的结果。

```
warnings.filterwarnings("ignore")
data = pd.read_csv('train_all.csv')
data = data.drop(['user_id','current_service'],axis=1)
data = data.dropna()
```

```
com= data.columns
data = MinMaxScaler().fit_transform(data)
data = pd.DataFrame(data)
data.columns = com
data.head()
```

	service_type	is_mix_service	online_time	1_total_fee	2_total_fee	3_total_fee	4_total_fee	month_traffic	many_over_bill	contract_type	contract_time
0	1.0	0.0	0.305147	0.097163	0.087594	0.040873	0.146659	0.023976	0.0	0.083333	0.698113
1	0.0	0.0	0.029412	0.087065	0.077244	0.028790	0.121646	0.000000	1.0	0.000000	0.018868
2	0.0	0.0	0.036765	0.014609	0.020760	0.009528	0.098514	0.016336	0.0	0.000000	0.018868
3	1.0	0.0	0.485294	0.028874	0.024072	0.010494	0.104015	0.006214	0.0	0.000000	0.018868
4	1.0	0.0	0.301471	0.104084	0.092882	0.060137	0.170435	0.037004	0.0	0.083333	0.471698

图 3-4　归一处理后的部分数据展示

3.4　特征分析

本文对电信行业存量用户智能套餐使用数据中的每一个字段作为一个特征变量，对该特征变量进行统计分析，以判定其是否会影响电信行业存量用户的智能套餐选择情况。最终选出对预测结果影响较大的特征向量进行分析预测。

3.4.1　特征可视化

（1）套餐类型可视化

通过对数据集中的所有套餐类型进行的可视化，可以看出训练数据集共有二种套餐类型。每种套餐的用户占比如图中 3-5 显示，由此可见，此推荐模式是一种二分类问题，把不同的用户分类在二类套餐类型中。

```
plt.rcParams['font.sans-serif']=['SimHei']
plt.rcParams['axes.unicode_minus']=False
pd.options.display.max_columns = None
pd.options.display.max_rows = None
plt.rcParams['figure.dpi'] = 150
warnings.filterwarnings("ignore")
data = pd.read_csv('train_all.csv')
data = data.drop(['user_id','current_service'],axis=1)
fig = plt.gcf()
fig.set_size_inches(6,6)
print(data['service_type'].value_counts())
```

```
data['service_type'].value_counts().plot.pie(title='套餐类型可视化',
                fontsize=15,legend=True,autopct=lambda v: "{:0.1f}%".format(v))
```

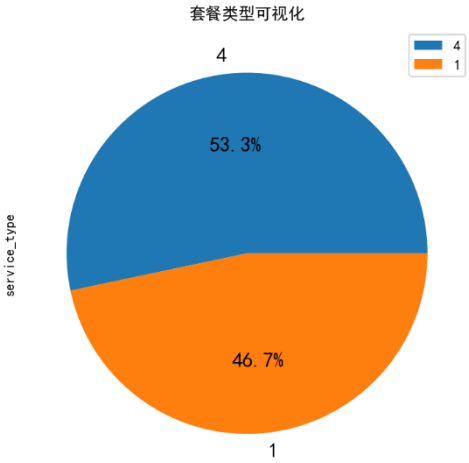


图 3-5　套餐类型可视化

（2）是否连续超套可视化

另外还对某些特征向量（例如是否连续超套等）做一些可视化研究，发现它对套餐种类的识别没有很好的效果。以“是否连续超套”为例加以解释，并将它和套餐类型的相关性用热力图加以可视化。如图 3-6 所示，在这个热力图中，横坐标表示的是套餐类型，纵坐标则表示的是是否连续超套，而图形的色调越深则表示二种类型间的关联就越明显。因此，是否连续超套这个特征不能将其视为强特征。

```
data = pd.read_csv('train_all.csv')
data = data.dropna()
df = data.loc[:,['service_type','many_over_bill']]
res=pd.DataFrame(columns=df['service_type'].value_counts().index.values.
                 tolist())
for i in df['many_over_bill'].value_counts().index.values.tolist():
    print(i)
    df2 = df[df['many_over_bill']==i]
    d = dict()
    for k,j in df2.value_counts().items():
        d[k[0]]=j
    res.loc[len(res)]=d
res.index = df['many_over_bill'].value_counts().index.values.tolist()
sns.heatmap(data=res,annot=True,fmt='.1f',cmap="winter")
res
```

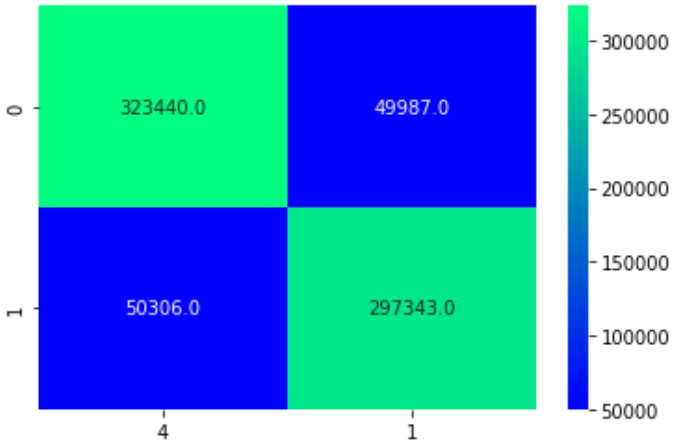


图 3-6　是否固移融合套餐可视化

（3）变量相关性可视化

通过热力图上各个方块颜色所对应的有关系数的变化程度，就能够确定出变数间关联的程度。该相关系数也能够度量出变数间的线性相关情况，也就是说，若相关系数含量越高，则变数间的线性相关程度就越好。而对于比相关系数分析小的二个变数，只能表示变数中间的线性相关程度较差，而不是表示变数间并不具有其它的相互关联，本实验对特征向量进行绘制相关系数热力图，如图 3-7 所示，通过热力图初步分析判断各相关变量之间的相关关系，为后面的特征选择提供参考。

```
plt.figure(figsize=(16,12))
cov = data.corr().round(2)
sns.heatmap(cov,cmap=plt.cm.winter,annot=True,linewidths=0.1,vmin=-1,)
plt.title('相关系数热力图\n',size=16)
```

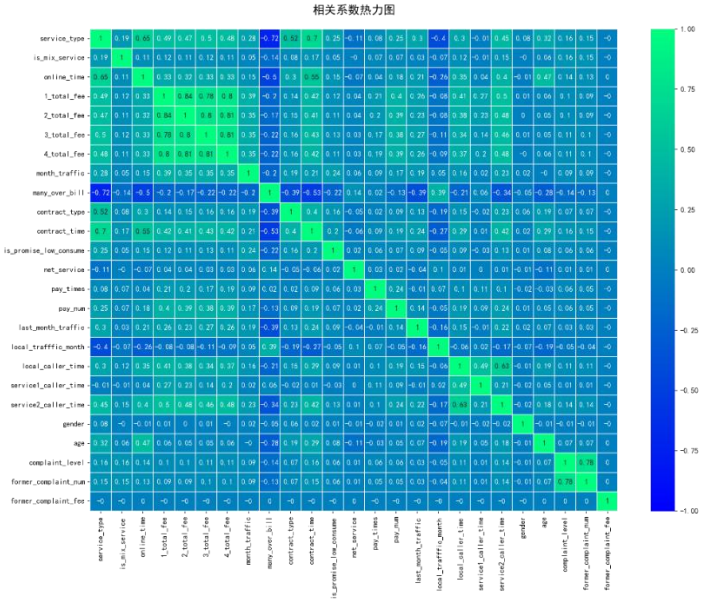


图 3-7　相关系数热力图

3.4.2 特征选择

通过 LightGBM 算法的特征重要性对特征向量进行评分，从而选出特征变量进行预测，在此次实验中设置评分阈值为 300，即如果该特征向量的重要性评分大于等于 300，就选择使用该特征变量。同时将测试数据按照 7:3 进行随机划分，一部分作为训练数据集，一部分作为测试数据集。图 3-9 是对特征向量的评分进行了可视化展示，通过可视化展示可以很容易地看出评分靠前的特征向量。图 3-10 则是对特征向量选择完成后的部分数据和特征向量进行的展示。

```
plt.rcParams['font.sans-serif']=['SimHei']
plt.rcParams['axes.unicode_minus']=False
pd.options.display.max_columns = None
pd.options.display.max_rows = None
plt.rcParams['figure.dpi'] = 150
warnings.filterwarnings("ignore")
data = pd.read_csv('train_all.csv')
data = data.drop(['user_id','current_service'],axis=1)
data = data.dropna()
x = data.drop(['service_type'],axis=1)
y = data['service_type']
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test =
train_test_split(x,y,test_size=0.3,stratify=y,random_state=2022)
import lightgbm as lgb
model=lgb.LGBMClassifier(learning_rate=0.3,n_estimators=350,max_depth=6,
       random_state=2022)
model.fit(x_train,y_train)
data_after1 = pd.DataFrame(model.feature_importances_,index=x_train.columns,
              columns=['特征重要性'])
data_after1 = data_after1.sort_values(by='特征重要性',ascending=False)
plot_importance(model,max_num_features=30)
plt.show()
print(data_after1)
x_train = x_train.loc[:,data_after1.index[:13]]
x_test = x_test.loc[:,data_after1.index[:13]]
x_train.head()
```

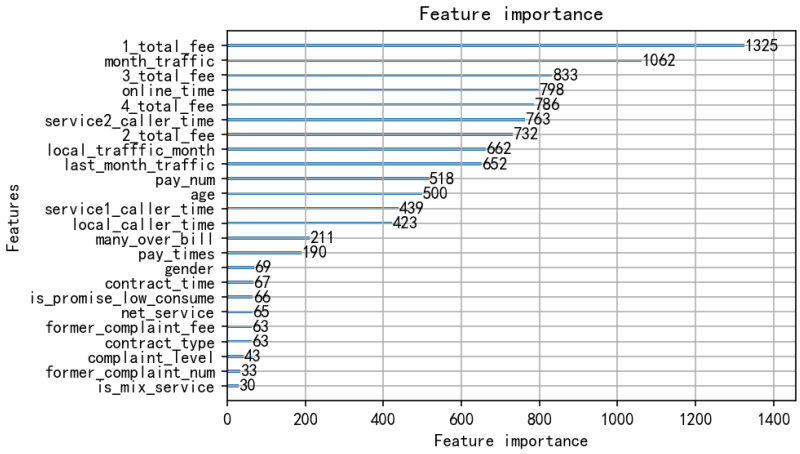


图 3-8　特征重要性排名可视化

	1_total_fee	month_traffic	3_total_fee	online_time	4_total_fee	service2_caller_time	2_total_fee	local_trafffic_month	last_month_traffic	pay_num	age
414961	19.1	0.000000	0.00	9	-0.06	0.00	0.00	0.000000	0.000000	30.00	57.0
528236	106.0	5760.798300	126.10	12	86.00	13.50	106.00	5526.177443	528.720473	100.00	27.0
526020	170.4	1107.884975	160.55	41	173.30	160.85	191.45	158.070360	2048.000000	100.05	26.0
336044	564.3	14431.580150	1089.90	90	520.00	98.75	564.00	14061.639750	2580.917057	600.00	48.0
413702	96.5	0.000000	13.90	4	31.26	0.00	57.07	72025.737560	0.000000	90.00	21.0

图 3-9 特征选择后的测试数据展示

3.5　本章小结

本章首先介绍了本文中所用数据来源并对数据进行介绍。然后对数据的特征向量进行了分类操作，其次进行了数据预处理，主要包括对缺失值的处理和对数据的归一化处理，最后对特征向量进行了可视化处理和特征选择，从而选出可以作为本文电信行业存量用户智能套餐推荐模型中的特征变量。

第 4 章　电信行业用户套餐的分析与预测

本章节通过电信行业用户订购的正使用的套餐情况，根据套餐所包含内容的属性，其次分析用户的历史套餐内容消费信息产生数据和最近 3 个月的消费行为习惯，为用户推荐最符合其消费情况与个人偏好的套餐类型。本文通过分析不同套餐的数据信息，捕捉不同类型套餐内容信息与用户表现出潜在消费偏好及特征之间的分布规律，对每位用户进行个性化匹配，给出适合用户的电信行业套餐类型，不仅可以给用户带来极致的上网体验，还可以为电信行业增加套餐消费情况与套餐推荐成功度。

4.1　开发工具及运行环境

开发工具：Jupyter Notebook

操作系统：Windows10

处理器：Intel(R) Core(TM) i5-8265U CPU @ 1.60GHz　　1.80 GHz

运行内存：8GB

Python 版本：Python 3.6.6

4.2　实验步骤

本次实验所涉及的主要步骤可分为 4 步，先进行数据的预处理操作，对数据进行处理形成可实验的数据，其次是对数据进行特征分类和提取，提取出对实验结果有重要影响的特征，然后进行模型训练，最后对套餐进行预测，分析结果[14]。

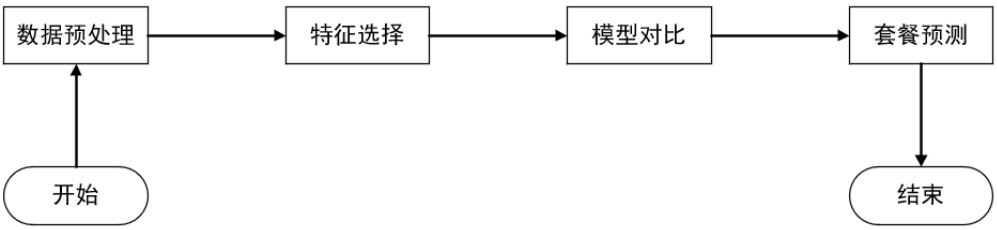


图 4-1　实验步骤

4.3　模型对比

下面将使用随机森林算法模型、LightGBM 算法模型和逻辑回归算法模型进行对比，以得出每个模型的优势与不足，完成电信行业用户套餐预测任务。下面将通过准确率

（Accuracy）、精确率（Precision）、召回率（Recall）、F1 值（F1-score）四个评价指标进行对比，得出每种算法的优劣[15]。

（1）准确率（Accuracy）

对于给定的测试数据集，分类器正确分类的样本数与总样本数之比。也就是损失函数是 0-1 损失时测试数据集上的准确率，如式（4-1）所示。

$$Accuracy = \frac{tp + tn}{tp + fp + tn + fn} \tag{4-1}$$

（2）精确率（Precision）

计算的是所有“正确被检索的 item(TP)”占所有“实际被检索到的 item(TP+FP)”的比例，如式（4-2）所示。

$$\Pr ecision = \frac{tp}{tp + fp} \tag{4-2}$$

（3）召回率（Recall）

计算的是所有“正确被检索的 item(TP)”占所有“应该检索到的 item(TP+FN)”的比例，如式（4-3）所示。

$$\mathrm{Re}call = \frac{tp}{tp + fn} \tag{4-3}$$

（4）F1 值（F1-score）

F1 值是精确值和召回率的调和均值，如式（4-4）所示。

$$F1 = \frac{2\Pr ecision * \mathrm{Re}call}{\Pr ecision + \mathrm{Re}call} \tag{4-4}$$

4.3.1 逻辑回归模型

本小节使用逻辑回归模型对电信行业的存量用户进行个性化套餐类型的推荐，通过使用 LogisticRegression 算法实现对数据的预测，同时将使用逻辑回归模型进行预测的准确率、精确率、召回率和 F1 值进行打印输出，如图 4-2 所示，预测结果分别为 0.898，0.976，0.809，0.885。最后对使用逻辑回归模型进行预测的结果输出混淆矩阵并进行可视化，如图 4-3 所示。

```
逻辑回归
Accuracy: 0.898979766367885
Precision: 0.9767048743776774
Recall: 0.8095855046593537
f1_score:  0.8853276241151499
运行时间为: 4.894094705581665
```

图 4-2　逻辑回归指标结果

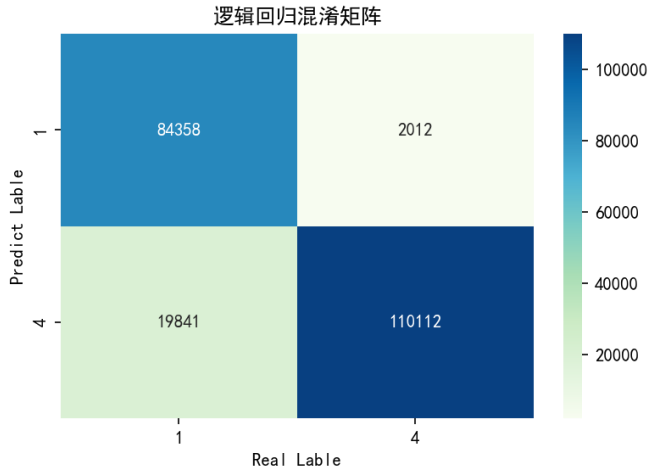


图 4-3　逻辑回归混淆矩阵

4.3.2　随机森林模型

本小节使用随机森林模型对电信行业的存量用户进行个性化套餐类型的推荐，通过使用 RandomForestClassifier 算法对实验数据进行预测，同时将使用随机森林模型进行预测的准确率、精确率、召回率和 F1 值进行打印输出，如图 4-4 所示，预测结果分别为 0.972，0.970，0.972，0.971。最后将对随机森林模型进行预测的结果输出混淆矩阵并进行可视化，如图 4-5 所示。

```
随机森林
Accuracy: 0.9724994568307577
Precision: 0.970861688871849
Recall: 0.9720822656647377
f1_score:  0.9714715938790288
运行时间为: 40.35134792327881
```

图 4-4　随机森林指标结果

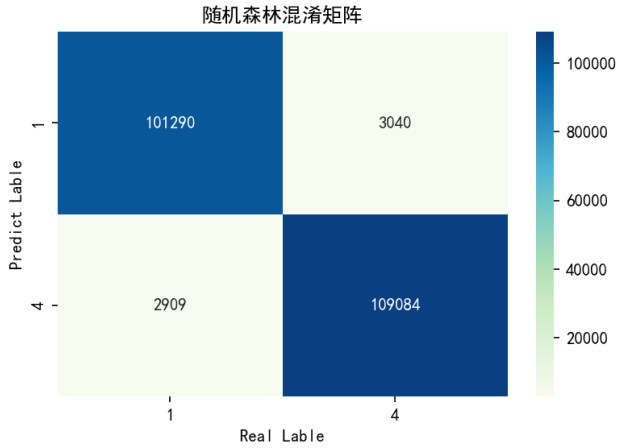


图 4-5　随机森林混淆矩阵

4.3.3 LightGBM

本小节使用 LightGBM 模型对电信行业的存量用户进行个性化套餐类型的推荐，同时计算并输出使用 LightGBM 模型进行预测的准确率、精确率、召回率和 F1 值，如图 4-6 所示，预测结果分别为 0.986，0.983，0.988，0.985。最后对使用 LightGBM 模型进行预测的结果输出混淆矩阵并进行可视化，如图 4-7 所示。

```
lightgbm
Accuracy: 0.9861364718499651
Precision: 0.9832117345964323
Recall: 0.9880900968339428
f1_score:  0.9856448794964459
运行时间为: 1.9218614101409912
```

图 4-6 LightGBM 指标结果

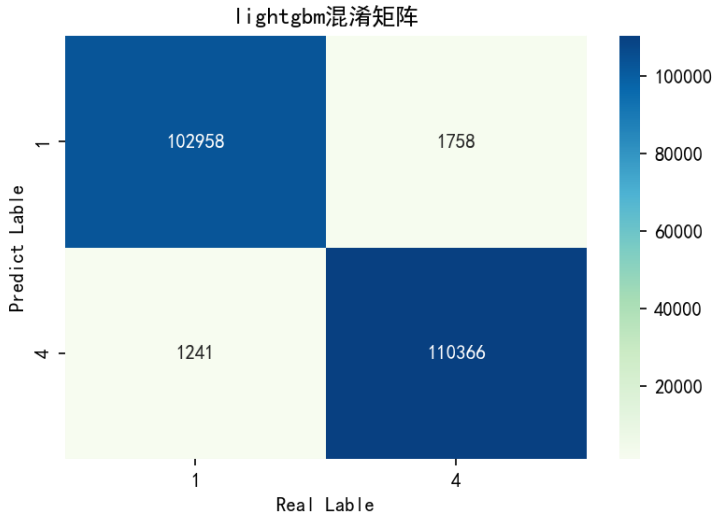


图 4-7 LightGBM 混淆矩阵

以上为逻辑回归、随机森林、LightGBM 三个模型在电信行业存量用户智能套餐推荐任务中的性能，三个模型的输入特征一致，观察其各个机器学习模型性能评价指标可以评估各模型的优势与缺点，便于本文选择模型完成套餐推荐任务。图 4-8 为三种算法的评价指标可视化，其中逻辑回归模型在本文用户套餐使用记录数据集与选择特征变量下，模型的准确率、精确度、召回率、F1 分别为 0.898，0.976，0.809，0.885。随机森林模型在本文用户套餐使用记录数据集与选择特征变量下，模型的准确率、精确度、召回率、F1 分别为 0.972，0.970，0.972，0.971。LightGBM 模型在本文用户套餐使用记录数据集与选择特征变量下，模型的准确率、精确度、召回率、F1 分别为 0.986，0.983，0.988，0.985。本文将使用三种算法中最优的 LightGBM 模型进行套餐类型进行预测。图 4-9 为三种算法模型进行预测所用时间的对比，图中可以看出 LightGBM 所用时间最短。

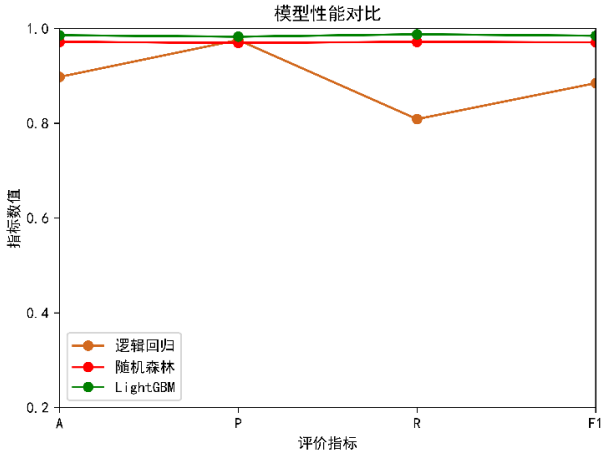


图 4-8　评价指标可视化

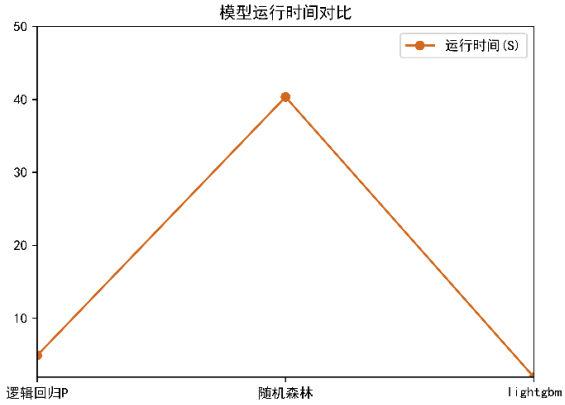


图 4-9　模型所用时间对比可视化

4.4　套餐类型推荐

下面是使用 LightGBM 算法模型对电信存量用户进行套餐类型推荐。首先通过预处理对需要预测的数据进行预处理，然后用之前选定的特征向量通过 LightGBM 算法模型对套餐类型进行预测，最后输出 User_id 和预测出的 service_type 到 CSV 文件中，保存结果如图 4-10 所示，第一列为用户 id，第二列为预测的用户的套餐的类型。

	user_id	service_type
0	012pSke7DsCrR985	1
1	012re3ZJSB6FptHW	1
2	013ElTBrRntuZDXz	1
3	0146B9qQbNxlynoX	1
4	0149vWf6u8UrKyJT	1

图 4-10　套餐类型预测结果

4.5 本章小结

本章节主要对实验数据通过算法进行套餐的分析和预测，叙述了实验的开发环境、工具和实验步骤，并且本次实验的初始数据进行了预处理，最后对实验预测的结果进行分析，并对 3 种算法的预测进行了比较并通过可视化进行展示。

第 5 章　总结与展望

由于电信套餐类型的多样性，使用者无法正确的挑选一个合适自己的套餐。但是对各大电信运营商而言，由于每年都有大批的电信用户携号码转网，用户们对自身的套餐也越来越不满意。因此为了留住电信用户，就必须做好存量管理，为这些电信用户量身定做自己的电信套餐，以吸引这部分用户，并确保用户无法将号码转换至其他运营商，从而沉淀用户，给电信用户带来更良好的服务。

5.1　总结

本文通过收集国内外相关文献对推荐算法模型进行了详细的介绍，便于本文介绍如何将推荐算法应用到电信行业存量用户智能套餐类型的推荐方案上。其次，本文对基于机器学习的套餐推荐模型进行了可行性分析。介绍并掌握了计算机教学中的逻辑回归算法、随机森林算法，以及 LightGBM。对用户的智能套餐类型的主要影响因素特征值加以提炼，形成了样本结果，并详尽说明了试验流程，对试验结论进行了详细分析，完成了基于 LightGBM 的电信业务存量用户智能套餐推荐模式的试验流程和结论，并通过应用智能套餐类型推荐模式，为用户提出了更加个性化、适合的套餐类型。本文同时也阐述了在传统情形下推荐算法的具体应用情况和主要实现流程，将该智能推荐模型应用到用户智能套餐推荐领域中，是对模型应用场景的拓宽。根据客户消费行为的历史消费行为数据，抽取从客户基础信息特点、客户套餐内的消费信息特点、客户交费信息内容特点等三个方面，得到影响客户在智能套餐选购商品的特征变量，将上述特征变量视为建立电信套餐推荐模式的基本数据，然后建立 LightGBM 机器学习模型完成套餐推荐方案的设计，便于向电信行业存量用户推荐合适其个人偏好的智能套餐，提高电信行业套餐销售利润。

5.2　展望

经过对中国电信行业存量用户智能套餐推广的深入研究与算法实现，使我对机器学习模型系统有了比较深刻的认识，对机器学习中的逻辑回归算法、随机森林算法与 LightGBM 有了比较好的认识。本研究也存在一些不足之处，在数据收集与数据处理阶段未能获取更多影响用户电信套餐选择的特征变量，更多的特征变量可以为模型带来性能上的提升；在电信行业用户套餐预测阶段本文仅选择了三种机器学习模型，未尝试更复杂的深度学习模型，深度学习模型可能会为套餐预测带来更好的效果。后续研究将在特征变量增加和使用深度学习模型完成电信行业用户套餐预测上做出改进。

参考文献

[1] 李楠. 基于 k-最近邻算法的电信套餐推荐模型研究[D]. 兰州：兰州大学, 2017.

[2] DAOUD R A, AMINE A, BOUIKHALENE B, et al. Combining RFM model and clustering techniques for customer value analysis of a company selling online[C]. 2015 IEEE/ACS 12th international Conference of Computer Systems and Applications (AICCSA), 2015, 1-6.

[3] 包志强, 胡啸天, 赵研, 等. 基于改进堆叠泛化算法的电信套餐预测[J]. 西安邮电大学学报, 2019, 24(02)：98-104.

[4] ABATUROV V S, DOROUOV A Y. The using of analytical platform for telecommunication network vents forecasting[C]. 2016 XIX IEEE International Conference on Soft Computing and Measurements (SCM). 2016, 359-360.

[5] 辛彤. 通信运营商客户流失预测及影响因素研究[D]. 重庆：西南大学, 2020.

[6] 吴倩. 疾病关键 miRNAs 识别及其对药物的影响关系研究[D]. 温州：温州大学, 2020.

[7] 吴阳. 财经领域命名实体识别方法的研究与系统实现[D]. 哈尔滨：哈尔滨工业大学, 2015.

[8] 石进. 基于 Spark 的分类算法并行化研究与实现[D]. 成都：电子科技大学, 2017.

[9] 刘夏,黄灿,余骁锋. 基于机器学习模型的专利质量预测初探[J]. 情报学报, 2019, 38(04): 402-410.

[10] 吴琼, 李荣琳, 洪海生, 等. 基于混合重抽样和 LightGBM 算法的配变低压跳闸预测[J]. 电力系统保护与控制, 2021, 49(12)：71-78.

[11] Pal M. Random forest classifier for remote sensing classification[J]. International Journal of Remote Sensing, 2005, 26(1)：217-222.

[12] 杜晶. 基于机器学习的电信套餐推荐系统的设计与实现[D]. 武汉：中南财经政法大学, 2020.

[13] 马栋坤. 基于机器学习的电信行业用户的智能套餐匹配模型[D]. 哈尔滨：黑龙江大学, 2021.

[14] 段海龙. 数据平衡与模型融合的用户购买行为预测研究[D]. 南昌：南昌大学, 2020.

[15] 胡海洋. 基于纹理与深度学习的指纹活体检测[D]. 长沙：湖南大学, 2018.

致　　谢

往事匆匆，一晃四年，如梦似幻，回首往事，历历在目，但人生路上有太多的离别，我们不得不说一声再见，然后各自珍重，珍惜当下。一段又一段的旅途，不断的收获成长与感悟，四年不仅是求学之路，更是成长之路上，一路走来更多的是感恩。

本学士学位论文的工作是在张慧娟老师的悉心指导下完成的，张慧娟老师严谨的治学态度和科学的工作方法给了我极大的帮助和影响。在此衷心感谢四年来张慧娟老师对我的关心和指导。

同时还要感谢我的父母对我的养育之恩以及求学之路的支持。因为你们的辛勤付出才让我有机会见识到更广阔的世界，看见更美的风景。世间万物，唯有父母和前途不可辜负，我一定会继续努力，成为你们的依靠。惟愿父母福寿绵延，身体安康！

至此，我的十六年求学之路将告一段落。离别总是那么的忧伤，但离别也是为了下一次更好的相遇，我们后会有期！