学　号：**20164751**

天 津 商 业 大 学 毕 业 设 计（论 文）

基于神经网络的运营商客户行为分析

Analysis of Operator Customer Behavior Based on Neural Network

学　　院：　　信息工程学院

教 学 系：　　商务信管系

专业班级：　　电子商务 2016-2

学生姓名：　　郝园媛

指导教师：　　周艳聪　副教授

2020 年 5 月 31 日

目　　录

内容摘要：伴随着大数据时代的到来，越来越多的企业将数据分析技术应用于获取、管理和处理大量数据，为企业挖掘用户价值提供了积极的帮助。本文将在分析运营商客户时，建立神经网络模型进行客户行为分析，同时与随机森林算法进行对比分析，找到更适合的预测模型，节省运营商营销成本并创造更多利润。本文由当前运营商的竞争背景引入，阐明了客户分析的意义，介绍了当前相关领域的研究现状，通过数据处理、模型构建、模型优化、模型分析的流程，为运营商维系流失客户提供方案，所构建预测模型的准确度较高，能够应用于实际的客户分析当中，为运营商实施精准营销提供可靠的工具。

关键词：客户行为分析；神经网络；随机森林；Python

Abstract: With the advent of big data era, more and more companies are applying data analysis technology to acquire, manage, and process large amounts of data, which provides positive help for companies to mine user value. This article will establish a Neural Network model for customer behavior analysis when analyzing customers of telecom operators. At the same time, it will conduct a comparative analysis with the Random Forest algorithm to find a more suitable prediction model to save operators' marketing costs and create more profits. This article introduces the competitive background of telecom operators, clarifies the significance of customer analysis, presents the current research status in related fields, and provides solutions for operators to maintain lost customers through data processing, model construction, model optimization, and model analysis processes. The accuracy of the forecasting model we built is relatively high, which can be applied to actual customer analysis and provide a reliable tool for operators to implement precision marketing.

Key Words: Customer Behavior Analysis Neural Network Random Forest Python

1 绪论

随着 5G 时代的来临，各大运营商竞争日趋激烈，再加上国内通信市场的逐渐饱和，客户资源越来越稀缺，通过大数据工具对客户数据进行分析，对不同的客户实施精准的营销策略，能够有效减少客户流失，为运营商带来更大价值。本研究以 Anaconda 为环境，采用 Jupyter Notebook 为代码编辑工具，用 Python 语言对运营商客户数据进行分析，通过构建 BP 神经网络模型，并采用随机森林算法对比分析，构建合理的客户流失预测模型，提高运营商的营销效率和行业竞争力。当前，在对运营商客户分析的研究中，国内外学者采用了多种数据分析方法，广泛应用了 K 均值算法、聚类算法、神经网络算法等算法，模型准确度较高，但在模型的优化和结合方面还有一定研究的空间。

1.1 研究背景及意义

虽然疫情影响了各行各业的步伐和进度，包括 5G，但三大运营商都已宣布 5G 目标不会变，并且会加快步伐。许多地方政府已经提出了对 5G 的应用要求，如雄安新区等地可能提前启动 5G 示范网络建设。随着技术路线的逐渐明晰，基础设施的逐渐完善，上游设备与配件的需求将变得更加清晰，在电信细分市场中的大型公司为了占领更大的市场份额，必将展开激烈的竞争。

电信运营商本身具有巨大的信息池[1]，整个行业也具备快速创新迭代的特点，同时具有大数据运用所需的大批高素质信息技术人才。大数据快速、全量的数据分析能够帮助企业实现对内部运营和外部竞争的深度洞察，并最终为运营商带来更大价值。

中国通信市场的份额有限，主要运营商引入的基本服务差异不大，客户的选择更加主观，因此运营商往往面临着一些客户离网的挑战。但是，通过对数据进行分析，采取积极有效的营销策略，运营商能够减少一些客户的离网意愿。通过建立恰当的预测模型，制定有效的营销策略，能够改善客户保持率，从而减少客户离网给公司造成的损失。同时，在优化模型的过程中，可以进一步发现客户价值和企业管理的潜在规律。论文将深入分析和探索运营商客户数据并根据所构建的模型查找隐

藏在历史客户数据中的有效信息和规则，为运营商的客户管理和相关决策提供理论和方法上的支持。

1.2 国内外研究概述

1.2.1 国外研究概述

由于国外电信行业兴起较早，许多经济发达国家已经在运营商客户分析方面有了相对成熟的研究。在许多数据分析算法中，K 均值算法、粒子群优化算法、模糊聚类算法、神经网络算法、随机森林算法都使用广泛。

国外研究方面，Berson[2]认为通信领域的客户流失意味着，客户从原有电信运营商转向了另一家公司，这使原运营商减少了用户数量，而对手增加了用户。文献[3]使用 SERVQUAL(Service Quality，服务质量)模型，通过差距分析、回归分析的方式测试电信运营商服务质量与客户满意度水平之间的关系，结果表明，电信运营商的可靠性、响应能力等对客户满意度和忠诚度具有显著影响。说明采取针对性的措施挽留客户，提升响应能力对提高运营商的行业竞争力很有价值。

因此，有很大一部分学者利用各种数据分析和建模技术，在客户分类、维系流失客户等方面发挥着重大作用。如 Boone 等在对客户进行分类时，通过人工神经网络准确地识别和分类客户组，并针对不同的客户类型实施相应的业务方法。Christoph Heitz 等人采集了电信运营商客户的基本数据以及消费数据，使用数据挖掘技术合理地对离网用户进行了聚类，为电信公司建立了客户流失预测模型，并将分析结果应用于流失用户的维护，从而提高公司利润。文献[4]提出了一种基于客户生命周期的新型客户细分方法，采用决策树方法提取与客户长期价值和忠诚度有关的重要参数，构建了一个简单实用的客户价值评估系统。

近年来，研究者们发现单一模型的准确度往往不能满足客户分析的要求，越来越多的学者开始结合多种模型对客户流失进行研究，如 Omar Adwan[5]等人使用了两种基于 MLP(Multilayer Perceptron，多层感知机)的方法并进行了比较，以便对客户流失率的最大影响因素进行排名，并构建客户流失分类模型。Ammara Ahmed[6]等人着重总结和分析了客户流失预测技术，以识别客户流失行为并验证其原因，提出最准确的客户流失预测是由混合模型而不是单一算法给出的。Farshid Abdi1[7]等人结合 K 均值算法和聚类技术，提出了一种基于数据挖掘技术的客户行为挖掘框架，有效帮助电信管理人员分析其客户的行为。

通过对国外文献的总结，发现研究者们在进行运营商客户分析时，采用的方法和技术都比较先进且种类多样，并且在客户细分、客户流失分析中具有较好的效果。同时，通过多种模型的结合对客户数据进行分析，对模型的准确度提高有很大的帮助，在这方面还有一定的研究空间。

1.2.2 国内研究概述

国内研究方面，研究初期，学者们主要运用单一模型进行客户流失分析，如王雷[8]等人提出了运营商客户流失预警模型，并在实际电信公司进行实施，结果表明，所建立的客户流失预警模型能为企业提供及时预警，为企业维护客户带来巨大价值。顾光同[9]等人在研究和分析云南一家电信运营商的相关客户数据时，采用决策树算法分析流失客户特征，更准确地分析了企业中流失客户的行为特征，并且在此基础上，找到了电信业客户流失预警的基本规则，计算出了预测的准确率。周荣鑫[10]等人利用贝叶斯网络的直观性和逻辑性等特点，建立客户流失预测模型，具有一定的现实意义。

目前，学者们的研究大多集中在对多方法的结合应用中，实现对运营商客户多角度、全方位的综合分析。如王观玉[11]等人针对通信公司客户数据多维度的特点，首先利用了主成分分析法对数据进行降维处理，找到了影响客户流失的关键指标，再通过支持向量机进行学习建模，模型以第一阶段的主成分为输入，采用主成分分析法与支持向量机结合的方式，有效提高了预测客户流失的准确性，为运营商客户流失预测提供了一种新方法。

林勤[12]等人通过对大均值子矩阵(LAS)双聚类算法和 K 均值算法的对比分析，进一步挖掘客户群体，为识别高价值客户市场带来实际的帮助。张珠香[13]等人采用 Log Rank 检验和 Cox 回归，建立了电信客户流失的生存分析模型。也有研究者在数据挖掘的不同过程中综合了不同的方法，如陈思含[14]在进行移动公司的客户分析过程中，建立了距离判别模型、支持向量机、神经网络三种模型，并采用 K 均值聚类、随机森林、逻辑回归分析等方法，寻找运营商流失客户的行为特征，使模型在实际应用中具有更好的适用性。

根据对国内相关文献的总结发现，在对各运营商的客户分析和研究中，数据挖掘技术和其相关派生模型的应用十分广泛。在进行客户行为分析时，新方法和新渗透领域必将成为未来数据挖掘技术和方法不断发展的两个主要方向[15]。随着研究者

对客户行为研究的不断深入，用于客户数据分析的方法类型将越来越丰富，在不同行业的客户关系管理中使用数据挖掘方法的方法将越来越成熟，未来的数据挖掘技术将越来越高效，为各行各业，尤其是电信行业带来更加精准科学的分析方案，以获得更大的效益。

1.3 研究内容及方法

1.3.1 研究内容与方法

随着电信市场的逐渐饱和，不可避免地出现了客户流失、客户保持率低等现象，通过科学的模型来进行客户分析，有针对性地实施营销策略，在电信行业中发挥着十分重要的作用。通过采集实际客户数据，采用适当的 BP 神经网络模型对客户行为进行研究，以便了解和把握数据之间的规律特征。

本文着眼于电信行业，通过对运营商客户数据的分析研究，构建模型，评价模型，优化模型，为运营商提供恰当的解决方案。本研究采用 Python 语言，以 Anaconda 为环境，构建 BP 神经网络模型，同时与随机森林算法的分析结果进行对比分析，实现对模型的改进和优化，进一步提高分析的准确率。

1.3.2 本论文的章节安排

第一章导言部分主要讲述了论文的研究背景及研究意义，对国内外的研究进行简述，并对研究内容和研究流程简单介绍。

第二章数据处理主要介绍对初始数据进行处理的具体方式、步骤以及操作的原理，便于后期数据模型的建立。

第三章神经网络模型主要讲述了模型的基本原理、模型构建、对模型的评价以及模型优化，初步理解模型实现原理，以便更好地比较和应用模型。

第四章随机森林简单介绍了随机森林算法，并对比分析随机森林模型和神经网络模型，对模型的结果进行分析并提出建议。

第五章结论与展望对研究结果进行分析并对该研究方向做出展望。

参考文献：按照顺序列举本文所引用的参考文献。

致谢：对论文写作过程中提供帮助的老师、家人和同学表达感谢。

2 数据处理

2.1 数据说明

某电信运营商提供了不同用户在三个月内的使用信息，一共 900,000 条数据，包括用户 ID、在网时长、通话时长、上网流量等 34 个特征，因变量为客户是否流失，其中 1 月和 2 月的流失标记值为空。本文将利用这些数据建立 BP 神经网络模型，分析用户的历史数据和当前数据，从中发现规律，进一步预测未来可能发生的行为，为运营商的客户维系带来价值。初始的数据说明如下表 1 所示：

表 1　运营商客户初始数据说明

名称	字段描述
MONTH_ID	月份
USER_ID	用户 ID
INNET_MONTH	在网时长（月）
IS_AGREE	是否合约有效用户（1=是，0=否）
AGREE_EXP_DATE	合约计划到期时间
CREDIT_LEVEL	信用等级
VIP_LVL	VIP 等级
ACCT_FEE	本月费用（元）
CALL_DURA	通话时长（秒）
NO_ROAM_LOCAL_CALL_DURA	本地通话时长（秒）
NO_ROAM_GN_LONG_CALL_DURA	国内长途通话时长（秒）
GN_ROAM_CALL_DURA	国内漫游通话时长（秒）
CDR_NUM	通话次数（次）
NO_ROAM_CDR_NUM	非漫游通话次数（次）
NO_ROAM_LOCAL_CDR_NUM	本地通话次数（次）
NO_ROAM_GN_LONG_CDR_NUM	国内长途通话次数（次）
GN_ROAM_CDR_NUM	国内漫游通话次数（次）
P2P_SMS_CNT_UP	短信发送数（条）
TOTAL_FLUX	上网流量（MB）
LOCAL_FLUX	本地非漫游上网流量（MB）
GN_ROAM_FLUX	国内漫游上网流量（MB）
CALL_DAYS	有通话天数
CALLING_DAYS	有主叫天数
CALLED_DAYS	有被叫天数
CALL_RING	语音呼叫圈
CALLED_RING	被叫呼叫圈
CUST_SEX	性别
CERT_AGE	年龄
CONSTELLATION_DESC	星座

续表 1

名称	字段描述
MANU_NAME	手机品牌名称
MODEL_NAME	手机型号名称
OS_DESC	操作系统描述
TERM_TYPE	终端硬件类型（0=无法区分，4=4G,3=3G,2=2G）
IS_LOST	用户在 3 月中是否流失标记（1=是，0=否）

观察数据，发现部分数据存在一些计算值重复的现象，比如，通话时长等于本地通话时长、国内长途通话时长、国内漫游通话时长的和，可以保留通话时长的总和数据，将其他的特征删去，便于后期模型的建立。同理，通话次数、上网流量、通话天数、语音呼叫圈等特征也可以做类似的处理。此外，数据特征中，用户星座及手机型号名称与客户流失关联较小，故删除这两个特征以简化数据。去除冗余数据后的数据说明如下表 2 所示：

表 2　去除冗余数据后的数据说明

名称	字段描述
MONTH_ID	月份
USER_ID	用户 ID
INNET_MONTH	在网时长（月）
IS_AGREE	是否合约有效用户（1=是，0=否）
AGREE_EXP_DATE	合约计划到期时间
CREDIT_LEVEL	信用等级
VIP_LVL	VIP 等级
ACCT_FEE	本月费用（元）
CALL_DURA	通话时长（秒）
CDR_NUM	通话次数（次）
P2P_SMS_CNT_UP	短信发送数（条）
TOTAL_FLUX	上网流量（MB）
CALL_DAYS	有通话天数
CALL_RING	语音呼叫圈
CUST_SEX	性别
CERT_AGE	年龄
MANU_NAME	手机品牌名称
OS_DESC	操作系统描述
TERM_TYPE	终端硬件类型（0=无法区分，4=4G,3=3G,2=2G）
IS_LOST	用户在 3 月中是否流失标记（1=是，0=否）

2.2　数据预处理

由于原始数据的不完美，例如包含数据异常或者冗余的内容，将不利于直接开启数据建模。所以在进行数据建模之前，需要对数据进行一定的处理，也就是数据预处理。数据预处理能够处理原本不可行的数据，使数据能够适应数据建模算法所

提出的要求。因此在数据挖掘前对数据进行预处理变得非常重要，高质量的数据也更有利于模型的建立。根据本文的数据情况，对原始数据中存在的缺失值、异常值等数据进行处理，从而避免因数据污染导致分析结果的偏差。

2.2.1 检测与处理重复值

在数据分析过程中，可能会遇到重复数据，数据的重复会导致数据的方差变小，数据分布发生较大变化，不利于模型的建立。处理重复数据之前要分析重复数据的产生原因，并分析将这部分数据去除后可能造成的影响。常见的重复数据分为两种，一种是记录重复，也就是一个或多个特征的某几条记录值完全相同；另一种是特征重复，也就是存在一个或多个特征名称不同但数据完全相同的情况。

根据本次用户数据的情况，主要处理记录重复的数据，首先可以利用 pandas 提供的名为 drop_duplicates 的去重方法，这种方法只对 DataFrame 或者是 Series 类型有效。相对于利用列表和集合的方法去重，这种方法不会改变数据的原始排列，并且具有代码简洁、运行稳定的特点。

处理结果如图 1 所示，初始状态的数据形状为 900,000 列，35 行，经过去除重复数据和删除冗余特征后的数据形状为 899,904 列，20 行。

```
print('初始状态的数据形状为：', churn_df.shape)
#删除不相关的特征：本地通话时长（NO_ROAM_LOCAL_CALL_DURA）等冗余特征
churn_df.drop(labels = ["NO_ROAM_LOCAL_CALL_DURA","NO_ROAM_GN_LONG_CALL_DURA",\
             "GN_ROAM_CALL_DURA","NO_ROAM_CDR_NUM","NO_ROAM_LOCAL_CDR_NUM",\
              "NO_ROAM_GN_LONG_CDR_NUM","GN_ROAM_CDR_NUM","LOCAL_FLUX","GN_ROAM_FLUX",\
            "CALLING_DAYS","CALLED_DAYS","CALLING_RING","CALLED_RING","CONSTELLATION_DESC",\
             "MODEL_NAME"],axis = 1,inplace = True)

#用drop_duplicates去重
churn_df = churn_df.drop_duplicates()
print('去重和删除冗余特征后的数据形状为：', churn_df.shape)
```

```
初始状态的数据形状为： (900000, 35)
去重和删除冗余特征后的数据形状为： (899904, 20)
```

图 1　处理重复数据前后的数据形状对比

2.2.2 检测与处理缺失值

缺失值一般指的是因为采样错误、成本限制或采集限制而没有存储或收集的数据[16]，通常由 NA 表示。在实际的数据分析过程中很难避免缺失值，但缺失值的存在往往会使数据分析不够准确。如果缺失值处理不当，很容易导致提取信息不足，甚至导致错误的结论。此外，缺失值的错误引入机制也容易导致知识提取的效率下

降，造成严重的数据偏移，使数据处理变得更加复杂。对于缺失值的处理主要有三种方法：

（1）删除法。删除法指将含有缺失值的特征或记录删除。它分为删除观测记录和删除特征两种方式。删除法属于通过减少样本量来换取信息完整度的一种方法，是一种最简单的缺失值处理方法，但是会引起数据结构的变动，使样本减少，并且重要信息可能会被丢弃。

（2）替换法。替换法是指用一个特定的值替换缺失值。特征可分为数值型和类别型两种，二者对缺失值的处理方法也是不同的。缺失值所在特征为数值形式时，通常采用它的均值、中位数和众数等描述其集中趋势的统计量来填补缺失值；当缺失值所在的特征为类别型时，则选择使用众数来填补缺失值。替换法使用难度不大，但是也会影响数据的标准差。

（3）插值法。在处理缺失值时，除了上述两种方法之外，还有一种常用的方法，就是插值法，常见的插值法有线性插值、多项式插值和样条插值等。它主要是通过其他变量与该缺失值变量的关系来对缺失值进行预测。基于变量之间的关系，可以通过其他变量作为自变量，将带有缺失值的变量作为因变量，通过求解方程或者建立模型拟合数据，利用得到的拟合结果来填补缺失值。但是在实际分析过程中，不确定自变量与因变量的相关关系，需要根据不同的情况，选择合适的插值方法。

由于本次样本数据量较大，样本特征较多，所以首先对缺失值进行识别和分析，在这里利用 pandas 提供识别缺失值的方法 isnull，自定义一个函数来计算每列缺失值占该列总数据的比例，从而查看数据的缺失情况再进行处理。每列数据缺失情况如下图 2，根据图中结果可以看出，合约计划到期时间（AGREE_EXP_DATE）以及 VIP 等级（VIP_LVL）数据缺失情况相对比较严重，缺失值比例分别约为 0.49 以及 0.35 左右，如果将缺失值全部删去对数据分析影响较大，所以采取缺失值填补的方法 fillna，采用均值补齐缺失数据。此外，从数据缺失情况可以看出，用户性别（CUST_SEX）、用户年龄（CERT_AGE）以及操作系统描述（OS_DESC）也存在少量的数据缺失，但是由于缺失程度较小，可以采取删除法删除缺失的样本，以保证信息完整度。用户在 3 月中是否流失的标记（IS_LOST）特征也存在大量缺失，是因为 1 月和 2 月的值为空，该特征后续会进行数据的进一步处理，此处暂不处理。最后，经过缺失值处理后，数据形状为 819,059 行 20 列。

```
Out[6]: MONTH_ID         0.000000
        USER_ID          0.000000
        INNET_MONTH      0.000000
        IS_AGREE         0.000000
        AGREE_EXP_DATE   0.489607
        CREDIT_LEVEL     0.000000
        VIP_LVL          0.345827
        ACCT_FEE         0.000000
        CALL_DURA        0.000000
        CDR_NUM          0.000000
        P2P_SMS_CNT_UP   0.000000
        TOTAL_FLUX       0.000000
        CALL_DAYS        0.000000
        CALL_RING        0.000000
        CUST_SEX         0.038101
        CERT_AGE         0.038861
        MANU_NAME        0.000002
        OS_DESC          0.042432
        TERM_TYPE        0.000000
        IS_LOST          0.666363
        dtype: float64
```

图 2　查看每列数据缺失比例

2.2.3　检测与处理异常值

数据集中，个别数据的数值明显偏离其余的数值，这样的数值称为异常值。异常值可能是一些外在因素产生的，如工作人员的错误输入、统计工具的不兼容、命名规则的差异等等。检测异常值就是检测数据中是否输入错误以及是否含有不合理的数据。异常值的存在对于数据分析来说十分危险，如果在数据分析过程中存在异常值，会导致分析结果产生较大的偏差，干扰数据模型的建立。常用的异常值处理方法如下：

（1）把异常值当作缺失值处理。这类方法也就是用处理缺失值的方法来处理异常值，如直接删除包含异常值的数据样本，将异常值修正为其对应特征的平均值、中位数等统计值。

（2）分箱法。分箱法通过考察数据周围的值使样本数据的值变得平滑，实现对异常值的处理。分箱方式包括等深分箱和等宽分箱。等深分箱是指每个分箱中的样本量一致；等宽分箱是指每个分箱中的取值范围一致。pandas 库的 qcut 函数，可以实现对数据的分箱处理。

（3）聚类法。聚类法能够把数据对象分组成为多个簇，在同一个簇中的对象相似度较高，不同簇之间的对象差别较大。聚类分析可以挖掘数据集中的孤立点，发现异常值。同时，通过聚类图像分析和 K 均值聚类等方法，也可以对异常值进行处

理。

（4）盖帽法。盖帽法一般是采用百分之 1 分位数和百分之 99 分位数替换连续变量中小于百分之 1 分位数和大于百分之 99 分位数的值，对异常数据进行盖帽处理。在 Python 中可自定义函数完成盖帽法，并可以设定盖帽的范围区间。

由于本次数据分析的数据量级比较大，所以先利用 describe 函数查看数据分布及极值情况，如下图 3 所示，可以看到许多特征的波动性都很大，例如本月费用（ACCT_FEE），用户平均月费用为 114.4，标准差在 106.9，波动性很大，中位数费用为 86.1，75 分位数费用为 141.4，说明绝大部分订单的本月费用都不多，月费用最大值为 11059.4，受极值影响，需要排除异常值。同理，在网时长（INNET_MONTH）、通话时长（CALL_DURA）、短信发送数（P2P_SMS_CNT_UP）、上网流量（TOTAL_FLUX）、年龄（CERT_AGE）波动过大，也需要进行异常值排除，因为数据样本比较多，再加上数据受极值影响较大，所以不便采用分箱法，使用盖帽法自定义函数处理异常值。

```
#2.3识别与处理数据的异常值
churn_nn.describe() #describe()查看数据分布及极值情况
```

	MONTH_ID	INNET_MONTH	IS_AGREE	AGREE_EXP_DATE	CREDIT_LEVEL	VIP_LVL	ACCT_FEE	CALL_DURA
count	819059.000000	819059.000000	819059.000000	819059.000000	819059.000000	819059.000000	819059.000000	8.190590e+05
mean	201602.001568	35.567941	0.510621	201645.968458	66.026844	79.936116	114.391305	2.228835e+04
std	0.816432	34.804358	0.499887	53.851497	0.961844	31.494892	106.886076	2.559831e+04
min	201601.000000	-251.000000	0.000000	201601.000000	0.000000	2.000000	0.010000	1.000000e+00
25%	201601.000000	10.000000	0.000000	201610.000000	65.000000	80.000000	56.000000	6.096000e+03
50%	201602.000000	26.000000	1.000000	201646.000000	66.000000	99.000000	86.100000	1.454900e+04
75%	201603.000000	50.000000	1.000000	201646.000000	67.000000	99.000000	141.400000	2.919900e+04
max	201603.000000	249.000000	1.000000	205012.000000	67.000000	99.000000	11059.400000	1.333863e+06

图 3　查看数据分布及极值情况

2.2.4　筛选和转换数据

在初始数据中，每个用户对应着 3 个月的数据记录，存在一个用户有 3 条记录的情况，所以需要对数据进行筛选。通过用户 ID（USER_ID）对数据进行分组，将数据处理为一个用户一条记录，保留用户每月的费用、通话时长等信息，从而便于数据模型的建立。

此外数据中还存在手机品牌、操作系统等类别型数据，如手机品牌有 633 种，操作系统描述有 47 种。神经网络模型的建立要求输入的特征为数值型，所以需要对

手机品牌和操作系统特征进行处理。首先先合并用户的手机品牌（MANU_NAME）与操作系统（OS_DESC）特征为“手机操作系统（MOBILE_OS）”，再对手机操作系统特征做哑变量处理。

哑变量指的是一种人为虚设的变量，又称为虚拟变量。它的取值通常为 0 或 1，能够反映某个变量的不同属性。对于一个类别型特征，若取值有 n 个，则经过哑变量处理以后就可以变成 n 个二元特征，并且这些特征之间相互排斥，每次只有一个激活。通过这个过程，也使得数据变得更加稀疏。经过哑变量处理后，类别型数据可以转化为稀疏矩阵的形式，解决了算法无法处理类别型数据的问题，也加快了算法的运算速度。

pandas 库中提供了 get_dummiea 函数，可用于对类别型特征进行哑变量处理。下图 4 是经过在哑变量处理之前和之后的对比数据，可以看到手机操作系统的特征变成了稀疏矩阵。

```
哑变量处理前数据为:
 0    赫比ANDROID 4.4.3
1    赫比ANDROID 4.4.3
2    赫比ANDROID 4.4.3
3    赫比ANDROID 4.4.3
4    赫比ANDROID 4.4.3
Name: MOBILE_OS, dtype: object
哑变量处理后数据为:
   1  ALCATELANDROID  ANBOANDROID  B-FLYANDROID  BLACKVIEWANDROID  BMMANDROID  \
0  0               0            0             0                 0           0
1  0               0            0             0                 0           0
2  0               0            0             0                 0           0
3  0               0            0             0                 0           0
4  0               0            0             0                 0           0

   BQANDROID  CKANDROID  CONDORANDROID  DEWAVANDROID  ...  \
0          0          0              0             0  ...
1          0          0              0             0  ...
2          0          0              0             0  ...
3          0          0              0             0  ...
4          0          0              0             0  ...
```

图 4　哑变量处理前后的数据对比

2.3　数据标准化

不同数据特征之间往往存在着不同的量纲，导致数值间的差异可能很大，降低建模的效率。因此在进行模型构建之前，需要将数据集根据建模要求统一转换成易于数据挖掘的形式。尤其当数据分析算法涉及空间距离度量或梯度下降等情况时，如果不对数据进行标准化处理，分析结果的准确性会受到很大的影响，甚至有时无法进行建模。为了避免特征之间取值范围和量纲的差异造成的影响，需要对数据进行标准化处理，也就是规范化处理。标准化数据有以下几种方法（X 为原始数据，

X*为标准化后的数据）：

（1）离差标准化。离差标准化数据是对原始数据进行线性变换，将原始数据的数值映射到[0,1]区间，也称为最小-最大标准化，其转换公式如下式（1）所示。

$$X^* = \frac{X-min}{max-min} \quad (1)$$

公式中，max 是样本数据的最大值，min 是样本数据的最小值，max-min 是数据的极差。离差标准化保留了原始数据之间的联系，并且相对简单，但是如果数据集中的某个数值很大，那么离差标准化的值就会接近于 0，并且数据之间差别非常小，易引起系统出错。

（2）标准差标准化。标准差标准化也叫零均值标准化或 z-score 标准化，主要利用特征的均值和标准差对数据进行变换。通过该方法处理的数据均值为 0，标准差为 1，转换公式如式（2）所示。

$$X^* = \frac{X-\overline{X}}{\delta} \quad (2)$$

其中$\overline{X}$为属性 X 的均值，δ 是属性 X 的标准差。标准差标准化受数据分布的影响较小，是使用非常广泛的一种数据标准化方法。

（3）小数定标标准化。小数定标标准化是利用移动数据的小数位数将数据映射到[-1,1]区间，移动的小数位数取决于数据绝对值的最大值，转化公式如下式（3）所示。

$$X^* = \frac{X}{10^k} \quad (3)$$

其中，k 为移动的小数位数。小数定标标准化方法适用范围广，并且受数据分布的影响小。

根据本文数据集的情况，采用标准差标准化数据，利用 StandardScalar 函数进行标准差标准化。

2.4 数据不平衡问题

数据不平衡是指响应变量值的不平衡分配，这种不平衡问题是影响数据分析的常见问题之一，这种问题也广泛存在于欺诈检测、网络入侵检测、医疗诊断等其他数据分析过程中，通常负样本的实例都大大超过了正样本的实例[17]。

文献[18]针对几类不同的标准分类器进行对比分析，如决策树、神经网络及支持向量机。表明数据的不平衡不仅会影响决策树分类，还会影响神经网络和支持向量机的分析性能，文献还对比了数据不平衡对分类器的影响程度，结果显示数据不平衡对支持向量机算法的影响相对较小。

在本文的客户流失分析当中，经过数据处理后的数据，有 8,301 名流失用户，有 266,762 名非流失用户，显然，流失用户的数据样本比非流失用户少很多，比例高达 1:32，导致模型在训练的过程中会产生偏差，模型的建立也很容易出现异常，所建的模型参考价值不大。因此在建模之前需要解决数据不平衡问题，保证模型的建立更准确。一般来说，当流失用户数据与非流失用户数据比例为 1:2 或 1:3 时模型的效果较好[19]。处理数据类别不平衡的方法有以下几种：

（1）重采样。重新采样是解决数据不平衡问题相对简单且比较常见的方法，它包括过采样和欠采样。过采样（Over-sampling）方法通过增加样本中少数类样本的数量来解决样本不平衡问题。最直接的方法是简单复制少数类样本形成多条记录。这种方法的缺点是如果样本特征较少的话，容易产生过拟合问题。经过改进的过采样方法可以学习少数类样本的特征，随机生成新的样本，或通过在少数类样本中加入随机噪声产生新的合成样本。过采样方法更适合大量数据分布不均衡的情况，其使用也比较广泛。

欠采样（Under-sampling）通过减少分类中多数类样本的样本数量来解决数据不平衡的问题。通过随机去掉一些多数类样本以减少多数类的规模，使多数类样本数量可以和少数类相匹配，但是会丢失多数类样本的一些重要信息。

（2）集成方法。在机器学习中，集成方法使用多种学习算法和技术，使得结果比只使用一种算法更优。针对数据不平衡问题，集成方法一般指每次生成训练集时使用所有分类中的小样本量，同时从分类中的大样本量中随机抽取数据与小样本量

合并构成训练集，反复多次可以得到多个训练集和训练模型，再用组合的方式产生预测结果。这种解决方法类似于民主投票，模型预测能力较好，当计算资源充足且模型时效性要求不高时可以采用这种方法。

（3）特征选择。特征选择的方法就是基于列的特征选择方法，可以通过选择具有显著性的特征配合参与解决数据不平衡问题，也能在一定程度上提高模型效果。

由于本次数据量较大，所以采用过采样的方法，通过 SMOTE（Synthetic Minority Over-sampling Technique）也就是合成少数类过采样技术进行重采样，SMOTE 的原理就是通过在少数类数据点的特征空间中随机选择一个 K 最近邻样本，根据该样本随机合成新样本。

通过 SMOTE 技术进行过采样之后，可以利用 Counter 函数查看数据情况，流失客户样本数量为 106,704，非流失客户的样本数量为 266,762，比例约为 1:2.5，可以进行后期模型的构建。

3　神经网络模型

3.1　BP 神经网络模型介绍

3.1.1　BP 神经网络简介

BP 神经网络是一种按误差反向传播训练的多层前馈网络，其算法称为 BP 算法。它的基本思想是梯度下降法，通过梯度搜索技术使网络的实际输出值和期望输出值的误差均方差为最小。20 世纪 80 年代，Rumelhart、McCelland 等科学家提出了一种基于误差反向传播理论的反向传播学习算法，也就是 BP 神经网络算法。

BP 神经网络的构建基于多层前馈网络，其模型结构如图 5 所示，它由输入层、隐藏层和输出层组成，从图中可以看出，神经网络的每一层都有若干个神经元，层与层之间的每个神经元都有连接，而每层内部的神经元之间无耦合。

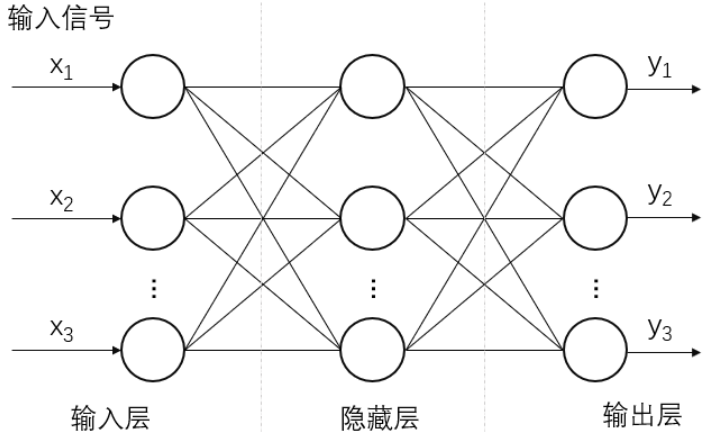


图 5　BP 神经网络的模型结构

基本的 BP 算法的过程主要包括信号的正向传播和误差的反向传播，是不断调整权重等参数的过程[20]。正向传播主要是指信号从输入层经隐藏层，最后到达输出层的过程。在这过程中，输入信号通过隐藏层作用于输出节点，产生输出信号，如果实际的输出与期望的输出结果不一致，则进入误差的反向传播过程。反向传播主要是指误差从输出层经隐藏层，最后到达输入层的过程。在这过程中，通过调节隐藏层到输出层的权重和偏置、输入层到隐藏层的权重和偏置，使误差沿梯度方向下降，并利用反复的学习训练，确定与最小误差相对应的权重和阈值等参数，则训练停止。

这时，经过训练的神经网络具有了对类似样本的处理能力，能够自动进行处理和学习，从而输出误差最小的、经过非线形转换的信息。

1988 年，Cybenko 提出，当每个节点使用 Sigmoid 型函数时，一个隐藏层就足以实现任意的判决分类问题，两个隐藏层则足以表示输入图形的任意输出函数。这显示了 BP 神经网络在解决非线性问题方面的强大数据识别和仿真能力，是一种可以被推广和应用的前沿理论和技术。目前，BP 神经网络应用十分广泛，主要领域有图像处理、客户预测、模式识别、医疗保健、经济预测、能源系统等等。

3.1.2 神经网络原理简述

BP 神经网络由输入层、一个或多个隐藏层以及输出层构成。作为一种前馈神经网络，BP 神经网络上层的输出往往作为下层的输入。同层节点中没有任何连接，每一层节点的输出只影响下一层节点的输出。网络的学习过程由正向传播和反向传播两部分组成。对于反向传播，其节点单元特征通常为 Sigmoid 函数[21]，如式（4）所示。

$$S(x) = \frac{1}{1+e^{-x}} \tag{4}$$

在训练阶段，用准备好的样本数据通过输入层、隐藏层和输出层训练后，对比输出结果和期望值。如果没有达到需要的误差程度或训练次数，就通过输出层、隐藏层和输入层来调节权值，使神经网络模型具有一定的适应能力。

BP 神经网络模型算法流程如图 6 所示。

（1）初始化。将权值和阈值 wji 设置为均匀分布的较小值，其中，i 代表输入层单元，j 代表隐藏层单元。

（2）提供训练样本及目标输出。对每个样本进行第 3 步到第 5 步的计算。

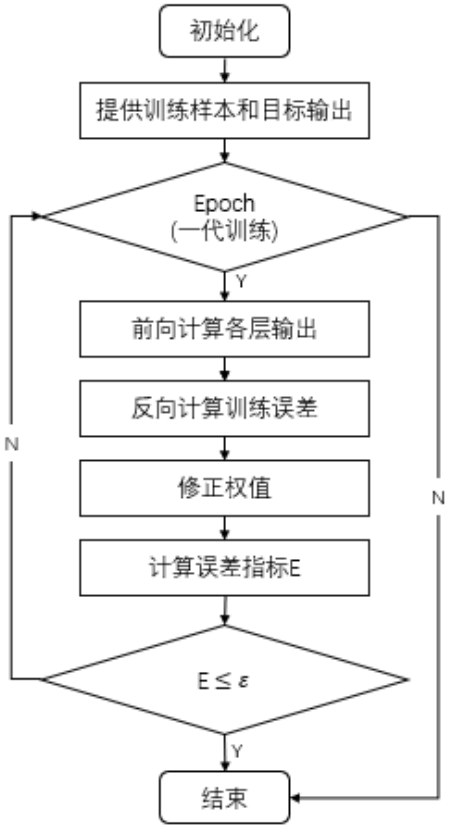


图 6　BP 神经网络模型算法流程图

（3）前向计算。对 l 层单元 j，计算线性组合系数，如式（5）所示，其中o_j表示隐藏层单元的输出。

$$v_j^{(l)} = \sum w_{ji}^{(l)} * o_j^{(l-1)} \tag{5}$$

计算本层的输出，如式（6）所示。

$$o_j^{(l)} = \varphi(v_j^{(l)}) \tag{6}$$

若函数$\varphi(v)$取 Sigmoid 函数，则直接有式（7）。

$$\varphi(v) = \frac{1}{1+e^{-v}} \quad (7)$$

若 l 为第一隐藏层，则如式（8）所示。

$$o_j^{(l)} = x_j \quad (8)$$

若 l 为输出层，则如式（9）所示。

$$y_i = o_j^{(l)} \quad (9)$$

（4）反向计算。对于 l 层单元 j，计算局部梯度$\delta_j^{(l)}$。

若 l 为输出层，则如式（10）所示。

$$\delta_j^{(l)} = (d_j - y_j) * \varphi(v_j^{(l)}) \quad (10)$$

其中，dj 为期望输出，若取 simoid 函数，则有式（11）。

$$\dot{\varphi}(\cdot) = \varphi(v)(1 - \varphi(v)) \quad (11)$$

其中，$\dot{\varphi}(\cdot)$为$\varphi(\cdot)$的导函数，若 l 为隐藏层，k 为输出层单元单元，则得到式（12）。

$$\delta_j^{(l)} = \dot{\varphi}(v_j^{(l)}) \sum_{k=1}^{n} \delta_k^{(l+1)} w_{kj}^{(l+1)} \quad (12)$$

（5）权值修正。设 n 为上述过程的计算次数，则权值修正公式如式（13）所示。

$$\omega_{ji}^{(l)}(n+1)=\omega_{ji}^{(l)}(n)+\mu\delta_{j}^{(l)}(n)\delta_{j}^{(l-1)}(n) \quad (13)$$

（6）计算误差，当数据集中所有样本都经历了第 3 步到第 5 步后，那就表示样本完成了一个训练周期（Epoch），然后计算误差指标。若误差指标满足精度要求，那么训练结束，否则，转到第 2 步，继续下一个训练周期。

3.2 神经网络模型构建

经过数据预处理后，用于模型训练的数据共有 373,466 条，其中流失用户数据有 266,762 条，非流失用户数据有 106,704 条。首先通过 Python 的 sklearn 库中的 train_test_split 函数进行数据集划分，将一部分数据作为训练集，另一部分数据作为测试集，可用于检验神经网络模型的预测准确度。其中，训练集数据样本为 280,099 个，测试集数据样本为 93,367 个。

在训练神经网络时，选取运营商客户的 14 个特征作为网络的输入，分别为是否合约有效用户、VIP 等级、用户性别、用户年龄、合约到期时间、终端硬件类型、本月费用、通话时长、短信发送数、上网流量、通话次数、通话天数、语音呼叫圈、手机和操作系统描述。训练 BP 神经网络时给定的输出为一个节点，该节点输出结果为 1 与 0，其中 1 代表运营商用户已经流失，0 代表运营商用户没有流失。由此，可以得出本文的神经网络结构如下图 7。

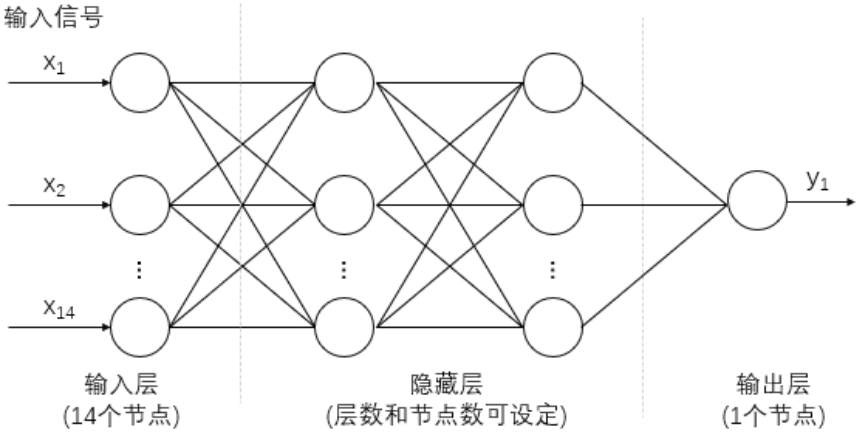


图 7　本文神经网络结构

本文根据 sklearn 库中的 MLPClassifier 方法进行模型的建立，其相关参数简介如下表 3。

根据几何金字塔公式的单个隐藏层节点设置方法，输入层有 n 个节点时，输出层有 m 个节点时，隐藏层节点可以设置为 sqrt(m*n)个。根据该公式设置初始的 BP 神经网络为单个隐藏层，隐藏层节点数为 3，权重优化求解器采用'lbfgs'模式，其余参数均取默认值。模型构建成功后，可以用 joblib.dump 函数保存模型，以便后续应用模型。构建的模型如下图 8 所示。

表 3　MLPClassifier 类常用参数及其简介

参数名称	简介
hidden_layer_sizes	表示隐藏层结构的参数，接受 tuple 格式。参数的长度代表隐藏层层数，其中，第 n 个元素表示第 n 层隐藏层的神经元个数。如（10,29）表示神经网络有两层隐藏层，第一层神经元为 10 个，第二层神经元为 29 个。
tol	表示模型训练中的收敛性阈值，接收 float 格式，用于判断收敛条件，默认值为 0.0001。
activation	表示激活函数的类型，接收 string 格式。默认值为'relu'，表示其激活函数为修正线性单元。其他可选值还有'identity'、'logistic'、'tanh'。'identity'表示激活函数为恒等式函数；'logistic'表示激活函数为 sigmoid 函数；'tanh'表示激活函数为 tanh 函数。
slover	表示权重优化求解器的类型，接收 string 格式。默认值为'adam'，表示一种 stochastic gradient-based 优化求解器。其他可选值还有'lbfgs'、'sgd'。'lbfgs'是基于一种伪牛顿算法的优化求解器；'sgd'指随机梯度下降法的优化求解器。
alpha	表示正则化项的系数，接受 float 格式，默认为 0.0001。
learning_rate_init	表示当优化求解器为'sgd'或'adam'时，所采用的初始学习率，接受 double 格式。默认值为 0.001。
power_t	表示逆缩放学习率的指数，接收 double 格式，默认值为 0.5。当学习率模式采用'invscaling'时，可通过该参数进行学习率的更新。
max_iter	表示最大迭代次数，接收 int 格式。如果达到了迭代值或算法已收敛，迭代就终止。默认值为 200。
verbose	表示是否输出算法的中间信息，接收 boolean 格式，默认值为 False。
learning_rate	表示当优化求解器为'sgd'时，用于权重更新的学习率模式。默认值为'constant'，即常数模式，将学习率设置为初始学习率（learning_rate_init）给出的恒定学习率。其他可选值还有'invscaling'、'adaptive'。'invscaling'即升级模式，通过逆缩放学习率指数（power_t）在每个时间 t 内逐步逐渐降低学习速率。'adaptive'即自适应模式，只要模型训练的损耗在下降，就保持初始学习率不变，当连续两次不能降低损耗或验证提前停止，则将当前学习率除以 5。

```
构建的模型为:
MLPClassifier(activation='relu', alpha=0.0001, batch_size='auto', beta_1=0.9,
        beta_2=0.999, early_stopping=False, epsilon=1e-08,
        hidden_layer_sizes=3, learning_rate='constant',
        learning_rate_init=0.001, max_fun=15000, max_iter=300,
        momentum=0.9, n_iter_no_change=10, nesterovs_momentum=True,
        power_t=0.5, random_state=45, shuffle=True, solver='lbfgs',
        tol=0.0001, validation_fraction=0.1, verbose=False,
        warm_start=False)
```

图 8　通过 MLPClassifier 构建的神经网络模型

3.3　神经网络模型评价及结果分析

在对分类器准确度进行评价时，往往会用到混淆矩阵的相关指标。混淆矩阵是用于表示精度评价的一种标准格式，也称作误差矩阵。对于二分类问题，模型最终需要判断样本的结果是 1 或 0，或者说是 positive 或 negative。通过样本的采集能够直接了解哪部分数据是 positive，哪部分数据是 negative；同时通过模型的分类可以知道模型判断了哪部分数据是 positive，哪部分数据是 negative。由此就可以得到 4 个基础指标：

（1）True positive=TP：真实值为 positive，模型也将其判断为 positive 的数量。

（2）False negative=FN：真实值为 positive，模型模型将其判断为 negative 的数量，也就是统计学上的第 2 类错误。

（3）False positive=FP：真实值为 negative，模型将其判断为 positive 的数量，也就是统计学上的第 1 类错误。

（4）True negative=TN：真实值为 negative，模型也将其判断为 negative 的数量。

将 4 个指标一起呈现为一个矩阵，称之为混淆矩阵，如下图 9 所示。

混淆矩阵		真实值	
		Positive	Negative
预测值	Positive	TP	FP
	Negative	FN	TN

图 9　混淆矩阵

在对分类模型进行评价时，为了能更好的评估模型的准确度，需要采用混淆矩阵的一些延伸指标，本文采用精确率（precision）、召回率（recall）和 F1 值作为模

型评价的指标：

（1）精确率（precision），它的计算公式如下式（14），指的是在模型预测为 positive 的所有结果中，模型预测正确的比重。

$$precision = \frac{TP}{TP+FP} \tag{14}$$

（2）召回率（recall），也称作灵敏度（sensitivity），计算公式如下式（15），指的是在真实值是 positive 的所有结果中，模型预测正确的比重。

$$\text{recall} = \frac{TP}{TP+FN} \tag{15}$$

（3）F1 值（F1-score），它的计算公式如式（16），其中 p 代表精确度（precision），r 代表（record）。F1 值指标结合了精确度与召回率的结果。它的取值范围在[0,1]之间，0 表示模型的输出结果最差，1 表示模型的输出结果最好。

$$\text{F1} - \text{score} = \frac{2pr}{p+r} \tag{16}$$

此外，为了使模型评价更加直观，可以结合 ROC 曲线评价模型的效果。ROC 曲线即受试者工作特征曲线，它的横坐标为 FPR（False Positive Rate，假阳性率）计算公式如式（17），也就是负样本中的错判率。

$$\text{FPR} = \frac{FP}{FP+TN} \tag{17}$$

纵坐标为 TPR（True Positive Rate，真阳性率），计算公式为（18），也就是判断正确样本中的正样本率，计算方式与召回率相同。

$$\mathrm{TPR} = \frac{TP}{TP+FN} \quad (18)$$

一般来说，当 TPR 接近于 1，FPR 接近于 0 时，模型较好，此时 ROC 曲线上的每一个点对应每一个阈值；当阈值最大时，TP=FP=0，对应图像的原点；当阈值最小时，TN=FN=0，对应于图像右上角的点（1,1）。ROC 曲线越靠近左上角，模型的查全率就越高，此时最靠近左上角的 ROC 曲线上的点分类错误最少，是最好的阈值，其假正例和假反例的总数也最少。

本文综合精确率（Precision）、召回率（Recall）、F1 值、ROC 曲线来对模型进行评价，运行三次后的模型评价报告如下表 4，其中 0 表示非流失用户，1 表示流失用户。

表 4　神经网络模型评价报告

标签 类别	精确率 （Precision）	召回率 （Recall）	F1 值	测试集数量
0	0.86	0.94	0.90	67049
1	0.80	0.60	0.69	26318
平均值	0.83	0.77	0.79	93367（总数）

根据表 4 可知，当神经网络模型有单个隐藏层，隐藏层节点数为 3 时，对客户流失预测的精确率达到了 80%，平均预测精确率为 83%，其 ROC 曲线如下图 10。

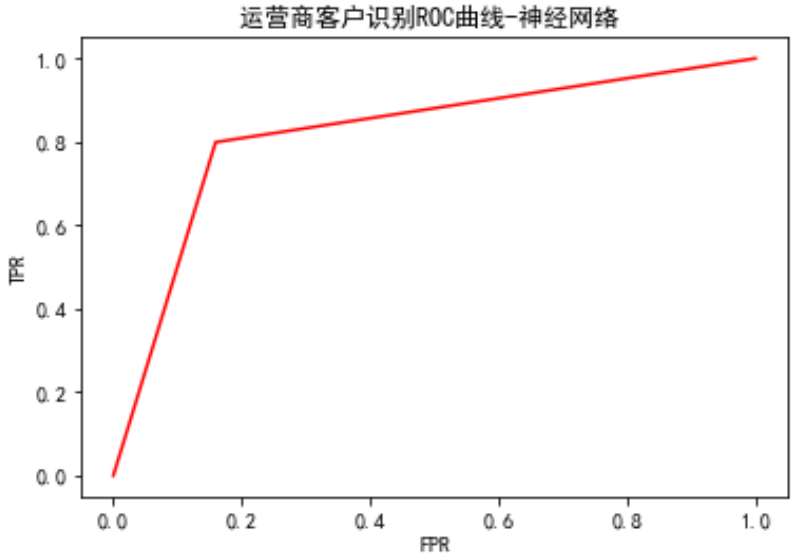


图 10　初始神经网络模型评价

从图中可以看出，初始神经网络模型 ROC 曲线整体比较靠近图像上侧，说明模型综合准确度较高，但是还有优化的空间，还需要通过调整参数等对模型进一步优化。

3.4 神经网络模型优化

BP 神经网络模型的优化有很多种方法，包括对算法结构的优化以及对模型的参数进行优化。在本文中，主要根据模型参数的自身特性，对隐藏层结构、权重优化求解器、学习率模式三方面进行优化。由于本文采用 Python 的 sklearn 库建立模型，Python 的算法封装较严密，不便对其算法结构直接进行优化，所以本文通过对神经网络结构的参数设定进行优化，从而提升模型的准确度。网络结构参数设定如下：

（1）输入层节点数。经过数据预处理后，取得了共有 14 个特征的训练数据集作为模型的输入，因此输入节点有 14 个。

（2）输出层节点数。本文中神经网络模型的只输出一个结果，客户流失的标签是 0 代表客户未流失，1 代表客户已流失，该模型的输出层只有一个节点。

（3）隐藏层层数。隐藏层层数是保证神经网络准确度的一个重要参数。一般情况下对隐藏层节点数的调整要比对隐藏层层数的调整更容易获得较低的误差，而且模型训练时长也相对较短，因此本文在设计初始模型时先建立了包含一个输入层，一个隐藏层，一个输出层的三层 BP 神经网络，通过先对单层的节点数调节后再对隐藏层层数进行调节。

（4）隐藏层节点数。对于神经网络隐藏层节点数的选择，还没有十分具体的理论指导，一般根据经验或者通过反复实验来确定。本文通过反复试错来确定节点数，根据比较不同参数下的模型准确度，得到一个参数更优的神经网络模型。

具体而言，当输入层节点数 n 为 14，输出层节点数 m 为 1，权重优化求解器采用'lbfgs'模式时，本文参考学者们提出的一些经验公式，再结合对隐藏层节点数的随机组合进行实验，对每个模型的训练精确度取平均值，得到神经网络隐藏层结构与训练精确度的关系，如下表 5。

表中括号内有几个数字则代表隐藏层有几层，每个数字为对应该层的节点数，如(18,15)表示隐藏层共有 2 层，第一层有 18 个节点，第二层有 15 个节点。从表中可以看出，当隐藏层层数为 2 层，每层节点数为 18 时的网络模型精确度较高，即隐藏层结构为(18,18)时，判断客户为非流失时的准确度为 92%，判断客户为流失的准

确度为 95%，并且隐藏层节点数不会过大。相比初始的神经网络模型，准确度得到了初步的提升，还需要在此基础上对其他参数进行调节，进一步优化模型。

表 5　隐藏层结构与训练精确度的关系

隐藏层层数	隐藏层节点数	精确度（0）	精确度（1）	参考方法
1	(29)	0.91	0.94	输入层 n 个节点，隐藏层 2n+1 个节点
1	(7)	0.90	0.95	输入层 n 个节点与输出层 m 个节点的平均值
1	(10)	0.91	0.94	隐藏层节点数为 2/3*n+m
1	(3)	0.86	0.80	几何金字塔公式，隐藏层节点数为 sqrt(m*n)
2	(3,3)	0.87	0.65	试错法
2	(10,10)	0.91	0.95	试错法
2	(20,20)	0.92	0.95	试错法
2	(29,29)	0.92	0.94	试错法
2	(21,21)	0.91	0.95	试错法
2	(18,18)	0.92	0.95	试错法
2	(17,17)	0.92	0.94	试错法
2	(18,15)	0.91	0.95	试错法
2	(18,17)	0.91	0.94	试错法
3	(3,3,3)	0.88	0.75	试错法
3	(18,18,18)	0.92	0.95	试错法
3	(18,18,29)	0.92	0.94	试错法
3	(18,10,10)	0.91	0.93	试错法

（5）权重优化求解器。MLPClassifier 函数中的 solver 参数可以对权重优化的算法进行选择，共有三种模式：准牛顿方法族的优化器“lbfgs”，随机梯度下降优化器“sgd”，以及由 Kingma 等人提出的基于随机梯度的优化器“adam”。模型默认模式为准牛顿方法族的优化器“lbfgs”。

一般来说，对于相对较大的数据集，"adam"在训练时间和准确度方面都能较好的工作，对于小型的数据集，“lbfgs”可以更快的收敛并且表现的更好。

本文的神经网络模型继续使用隐藏层层数为 2 层，每层节点数为 18 的隐藏层结构，即在隐藏层结构为（18,18）的情况下，对三种权重优化算法的精确度进行了比较，比较结果如下表 6。从表中可以看出，基于随机梯度的优化器（adam）和随机梯度下降（sgd）的权重优化算法准确度较高，可以在此基础上进一步调节对权重优化十分重要的学习率参数。表中的结果也说明，对于电信运营商客户的大数据集，采用"adam"或“sgd”模式可以取得较好的结果。

表 6　权重优化算法与模型精确度的关系

权重优化算法	精确度（0）	精确度（1）
lbfgs	0.92	0.95
sgd	0.93	0.96
adam	0.94	0.95

（6）学习率模式。学习率主要用于权重更新，指导模型如何通过损失函数的梯度调整网络权重的超参数。它是影响模型性能的重要参数之一，如果只能调整一个超参数，那么最好的选择就是它。学习率越低，损失函数的变化速度就越慢，达到收敛所需要的迭代次数也就越高；学习率较大时，则每次迭代可能不会减小损失函数的结果甚至会超过局部最小值，导致无法收敛。MLPClassifier 函数中的 learning_rate 参数可以对学习率模式进行调节，仅当权重优化求解器（solver 参数）为随机梯度下降（sgd）时使用，具有三种形式可以选择：常数模式（constant）是将 learning_rate 设置初始学习率（learning_rate_init）给出的恒定学习率；升级模式（invscaling）是通过逆缩放学习率指数（power_t）在每个时间 t 内逐步逐渐降低学习速率，其公式如下式（19）。

$$\text{effective\_learning\_rate} = \frac{\text{learning\_rate\_init}}{\text{power}\left(t,\text{power\_t}\right)} \tag{19}$$

还有一种学习率为自适应模式（adaptive），它的灵活性相对较强，当两个连续的时期没有将模型训练的损耗减少设定优化的容忍度（tol）时，或者当验证评分没有改善时，并设置了提前停止（early_stopping）来终止训练，如果没有将验证分数增加到所设定的优化容忍度（tol），则将当前学习速率除以 5。也就是说，如果模型训练的损耗在下降，就保持初始学习率不变；如果连续两次不能降低损失或验证提前停止，就将当前学习率除以 5。

本文的模型采用隐藏层为两层，每层节点数为 18 个的隐藏层结构，即在隐藏层结构为（18,18）的情况下，权重算法采用随机梯度下降（sgd）算法，对三种不同学习率模式的精确度进行了比较，比较结果如下表 7。

表 7　学习率模式与模型精确度的关系

学习率模式	精确度（0）	精确度（1）
constant	0.93	0.96
invscaling	0.80	0.63
adaptive	0.93	0.96

从表中可以看出，当采用常数模式（constant）和自适应模式（adaptive）的学习率时，神经网络模型准确度较高。综合上述参数优化过程，最终确定隐藏层结构为两层，每层节点为 18 个，权重下降算法采用随机梯度下降法，学习率模式采用自适应模式所得的模型预测结果准确率最高，平均准确率达到了 95%，最后得到的 ROC 曲线如下图 11，可以直观地看出，优化后的 BP 神经网络 ROC 曲线为蓝色，相对红色的优化前曲线，更加接近左上角，准确度较高。

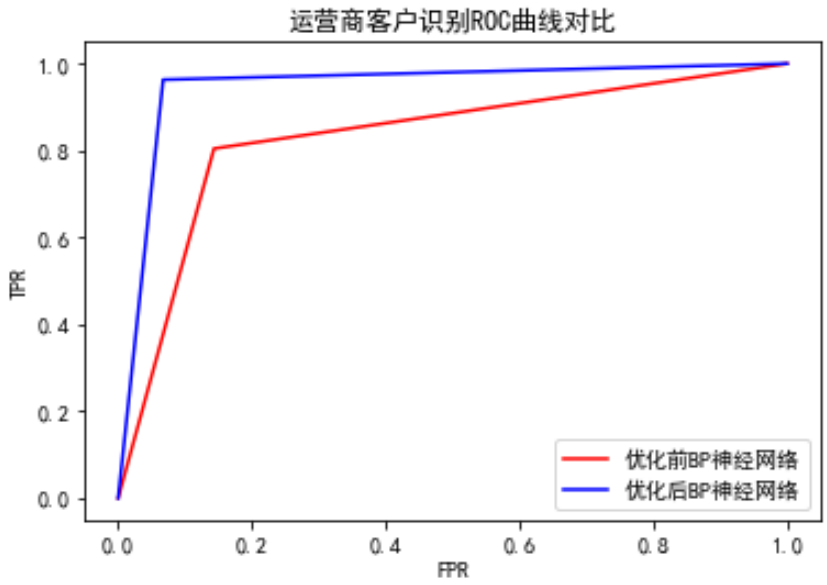


图 11　神经网络模型优化前后的 ROC 曲线对比

4 随机森林

4.1 随机森林简介

随机森林是一种基于决策树的集成学习算法[22]。决策树是一种树状分类器，在这个分类器中，每个内部节点表示对某个属性的判断，每个分支代表一个判断结果的输出，每个叶节点代表一种分类结果。集成学习利用一系列的学习器进行学习，并通过某种规则把各类学习结果进行整合，进而得到比单个学习器更优的结果。它是一种常用的机器学习方法，其代表算法有 Boosting 系列算法、Bagging 系列算法等。

随机森林就像是一个分类投票系统，能够通过集成学习的方法对多棵决策树的结果进行整合，每棵决策树都是一个分类器，对于每个所输入的样本，有 m 棵决策树就会有 m 个分类，随机森林集成了所有的分类投票结果，并将投票次数最多的类别指定为最终输出。

理解随机森林可以从两个角度，首先是“随机”，假设有共有 N 个样本，M 个特征，对于每棵树都有放回的随机抽取$\frac{2}{3}N$个样本作为训练集，再有放回的随机选取 m 个特征作为这棵树的分支依据（m≤M）。可以看出，随机森林主要指选取样本和选取特征两方面的随机；另一个角度就是“森林”，它的单元为决策树，那么成百上千棵决策树就可以称之为“森林”。

与许多其他的分类器相比，随机森林的策略往往能表现出较好的性能，并且对过拟合具有强大的作用[23]，并且它的准确性相对较高，可以针对数据量较大的数据集进行分析并能获得比较高效的预测结果。除此之外，它还能够处理具有高维特征的输入样本，并且不需要删除特征。由于随机森林在每次划分时考虑的属性较小，所以它在大型数据库的应用中效果较好[24]。

随机森林作为一种非常灵活的机器学习算法，既可以用于市场营销中的客户分析，统计客户来源、留存和流失，也可以用于预测患者的敏感性等，可以说应用领域十分广泛，并且模型性能较好。

随机森林的原理图如下图 12，它的具体过程如下：

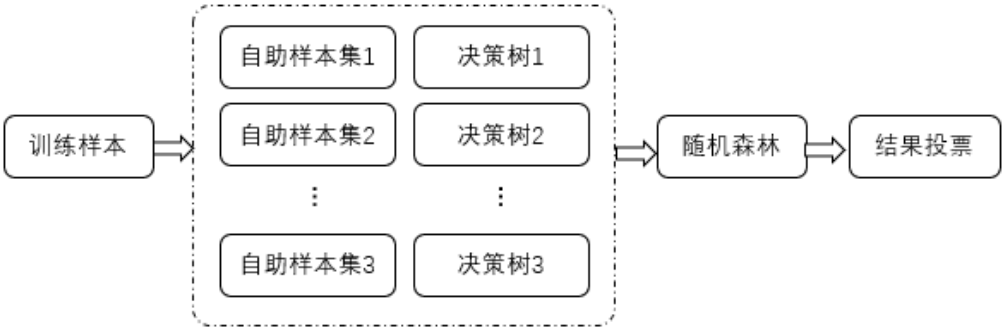


图 12　随机森林原理示意图

（1）从原始数据中有放回的随机抽取训练样本得到训练集。

（2）随机抽取出参与构建决策树的预测变量。

（3）综合多棵决策树的结果，迭代多次从而优化特征和森林。

4.2　随机森林模型构建

为了能够更好的对比随机森林和神经网络的性能，在构建随机森林模型时，采用构建神经网络模型的数据，数据预处理方式也都相同。随机森林模型的建立在 Python 中可以通过随机森林分类器（Random Forest Classifier）建立模型并设置参数。本文主要设定了以下参数，其他参数设置为默认：

（1）单个决策树使用特征的最大数量（max_features）。Python 为最大特征数量提供了多个可选项：“auto”可以简单的选取所有特征，每棵树都可以利用这些特征并且没有任何的限制；“sqrt”代表每棵树可以利用总特征数的平方根个；“log2”则是将以 2 为底总特征数的对数为决策树使用特征的最大数量。增加单个决策树特征的最大数量一般能提高模型的性能，因为在每个节点上有更多的选择可以考虑，但会使单个决策树的多样性降低，因此在选择特征数量时还需要适当的平衡和尝试。经过测试后，本文选择单个决策树的最大特征数为“auto”模式。

（2）最小叶子样本（min_ sampl_leaf）。叶是决策树的末端节点，较小的叶子能够使模型更容易捕捉训练数据中的噪声，但较大的叶子易使模型的训练速度减慢，所以本文将该参数设为 50。

（3）子树的数量（n_ estimators），在通过最大投票数或平均值来进行预测之前，需要确定子树的数量，较多的子树可以让模型有更好的性能，但同时也会减慢代码运行速度，因此这里选择子树个数为 10 个。

经过参数设定后，最终构建的随机森林模型如下图 13 所示。

```
构建的随机森林树模型为:
 RandomForestClassifier(bootstrap=True, ccp_alpha=0.0, class_weight=None,
                        criterion='gini', max_depth=None, max_features='sqrt',
                        max_leaf_nodes=None, max_samples=None,
                        min_impurity_decrease=0.0, min_impurity_split=None,
                        min_samples_leaf=50, min_samples_split=2,
                        min_weight_fraction_leaf=0.0, n_estimators=10,
                        n_jobs=None, oob_score=False, random_state=50, verbose=0,
                        warm_start=False)
```

图 13　随机森林模型构建

构建好随机森林模型后，采用和神经网络模型相同的评价指标对随机森林模型进行评价，运行 10 次后得到评价报告如下表 8，0 标签表示在网用户，1 标签表示离网用户。

表 8　随机森林模型评价报告

标签 类别	精确率 （Precision）	召回率 （Recall）	F1 值	测试集数量
0	0.93	0.99	0.96	67049
1	0.98	0.82	0.89	26318
平均值	0.96	0.91	0.93	93367（总数）

随机森林模型的 ROC 评价曲线如下图 14，可以看到该曲线已经很接近图像左上侧，准确度较高。

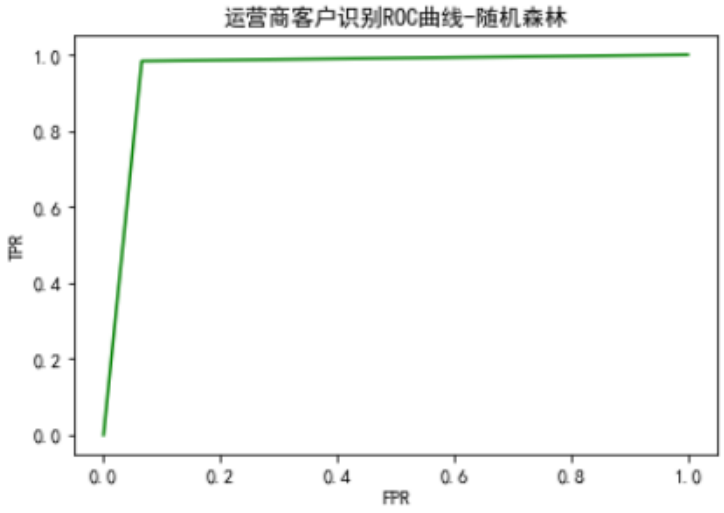


图 14　随机森林模型评价

4.3 模型对比

BP神经网络是由多个非线性单元组合而成的模型，需要在每个神经元进行运算。随机森林是以决策树为基本单元，通过集成学习后所得的组合分类器。二者都是近年来比较流行的预测算法，本文结合对运营商客户分析的结果，从以下几个角度对二者进行对比：

（1）首先，二者的决策边界不同，因此模型的表现效果也不同。BP 神经网络的层与层之间具有高度的连接性，它的决策边界是由一组非线性的计算单元组成，并且神经模型计算比较复杂，但是它所提供的决策边界是非线性的曲线或曲面。随机森林是通过多个决策树以投票的形式输出分类结果，它具有较好的容差性，不同于神经网络，它的决策边界更像是多个分段函数[25]。

（2）二者的预测准确度略有差异，总体来说，随机森林算法的准确度要高于 BP 神经网络模型的准确度。如下表 9 所示，经过参数优化的 BP 神经网络模型，对非流失用户的预测精确度为 93%，对流失用户的预测精确度为 96%。相比而言，随机森林模型的精确度略高，它对非流失用户的预测精确度为 93%，对流失用户的预测精确度为 98%。召回率和 F1 值方面，随机森林模型也略优于神经网络模型。

表 9　神经网络模型和随机森林模型精确度对比

模型类别	精确度（0）	精确度（1）	召回率（平均值）	F1 值（平均值）
BP 神经网络原始模型	0.86	0.80	0.77	0.79
BP 神经网络优化后模型	0.93	0.96	0.90	0.92
随机森林模型	0.93	0.98	0.91	0.93

此外，将两种模型的 ROC 曲线放在同一坐标中进行对比，如图 15，绿色曲线代表随机森林模型，蓝色曲线代表优化后的神经网络模型，随机森林模型的曲线整体比 BP 神经网络的曲线更靠图片左上方，也更加的平滑。综合来看，在对运营商客户进行分析中，随机森林模型的性能要优于 BP 神经网络模型。

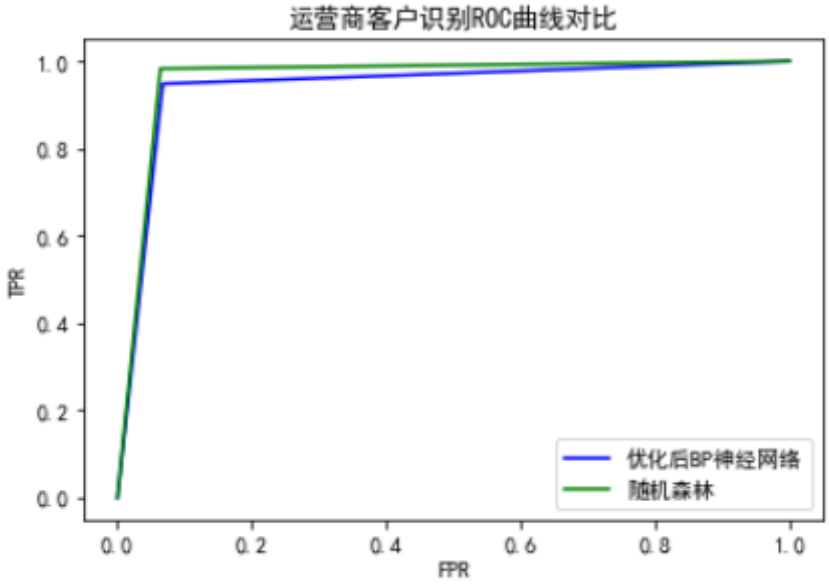


图 15　神经网络模型和随机森林模型 ROC 曲线对比

（3）二者对于不平衡分类情况的处理效果不同。BP 神经网络模型对不平衡分类的数据较为敏感，所以它得出的模型很容易出现异常，在本文的运营商客户分析中，流失用户占较少的比例，出现了不平衡分类问题，如下表 10 是神经网络模型数据不平衡情况下的模型评价，可以看出它的准确度在数据不平衡时出现了异常。而随机森林算法对不平衡数据的问题比较不敏感，它不会过度抽样，面对不平衡的分类，它受到的影响较小。

表 10　神经网络模型在数据不平衡情况下的模型评价

标签 类别	精确率 （Precision）	召回率 （Recall）	F1 值	测试集数量
0	0.97	1.00	0.98	66703
1	1.00	0.00	0.01	2063
平均值	0.99	0.50	0.50	93367（总数）

（4）模型训练速度方面，随机森林模型面对大量的数据也能够有效的处理，并且学习速度很快。相对而言。BP 神经网络处理速度较慢，尤其是在数据量比较大，所建模型的隐藏层层数和隐藏层节点数较多的情况下

（5）参数设定方面，BP 神经网络有较多的超参数，BP 神经网络的可调节性较强，但是在对参数设定和优化方面往往要花费较多的时间，所以神经网络一般存在较大的冗余性，也模型的训练带来负担。相比之下，随机森林模型的超参数较少，

而且它的参数设定比较稳定，在参数设置方面不会花费太大的时间就能取到较好的效果。

（6）过拟合程度方面，BP 神经网络较易发生过拟合的问题[26]，常常会导致在训练集中效果较好，但在测试集中结果不是特别好。与其他分类方法相比，随机森林算法的一大特点就是它不会发生过拟合，其分类准确率也较高。

（7）参数选择方面，神经网络有许多参数并且所有参数基本上都会在训练模型时使用，并且神经网络的分类需要利用所学到的变量特征。同样的样本，随机森林模型可以选择部分参数进行测试，并赋予参数不同的权重，它的模型建立是边建树边分类的。

综合上述对二者的对比情况，可以得到下表 11。

表 11　神经网络模型和随机森林模型的对比

比较角度	BP 神经网络模型	随机森林模型
决策边界	由一组非线性的计算单元组成	是多个分段函数
预测准确度	略低	略高
对不平衡数据的敏感性	对不平衡数据较敏感	受不平衡数据影响较小
模型训练速度	特别是隐藏层结构复杂时较慢	较快
参数设定	超参数多，可调节性强	超参数少，节省参数优化时间
参数选择	基本所有参数都会在训练模型时使用	可以选择部分参数进行训练，且选择的方式可以调整
过拟合程度	较易发生过拟合的问题	不会发生过拟合，分类准确率较高

除了这些不同点之外，二者也有一定的相同点：

（1）二者都具有较好的容错能力。即使数据集中存在大量噪声干扰，BP 神经网络也能够产生较好的效果，同时如果神经网络在局部神经元受到破坏后，对全局的训练结果不会造成很大的影响。随机森林也是如此，在模型建立的过程中，它能够获取到内部生成误差的一种无偏估计，具有较好的容错能力。此外，对于缺省性问题，随机森林能够较好的估计遗失数据，并且其预测的准确度较高。

（2）二者都不易进行训练规则的可视化。BP 神经网络的内部结构就像一个“黑盒子”，它对自身行为的解释能力不强，因为即使不知道输入与输出之间的关系，它也可实现高度的非线性映射。随机森林模型的训练规则也不可视，因为不同的决策树训练出来的规则不同，所以很难给出规则的具体形式。

（3）二者的层次性较强。BP 神经网络有很多层次，包括输入层，一层或多层隐藏层和输出层，并且层与层之间的联系非常紧密，层层递进。随机森林也有许多层次，或者说有许多的决策树单元，它的训练也是逐层进行。

通过分析和对比二者的相同点与不同点，一方面有利于研究者在面对不同的数据样本时，根据数据特点和分析需求选择合适的数据模型进行训练；另一方面有利于对二者的算法进行结合，通过优势互补从而得到更好更优的模型。

4.4 模型分析

4.4.1 模型结果分析

分析上述模型结果可知，经过数据处理后，投入训练的总样本量为 373,466 个，其中非流失样本有 266,762 个，流失样本有 106,704 个。经过对模型的参数优化后，神经网络模型平均预测精确率由 83%提升到了 95%，说明模型优化是比较成功的。BP 神经网络模型非流失用户精确率（precision）为 93%，也就是在模型判断为非流失用户的所有结果中，预测正确的比重为 93%；模型判断流失用户的精确率为 96%，也就是在模型判断为流失用户的所有结果中，预测正确的比重为 96%，平均预测精确率为 95%。它的平均召回率达到了 90%，平均 F1 值达到了 92%，模型预测效果较好。

随机森林模型判断流失的精确率达到了 93%，判断非流失的精确率达到了 98%，平均精确率 95%。总体来看，随机森林模型的预测准确程度以及运算速度略优于 BP 神经网络模型。综合上述结果可以确定，本研究创建的模型是有效的，并且模型预测效果较好，能够将其用于实际的客户流失预测中，从而针对流失客户的特点采取一定的策略挽留客户，为运营商减少损失。

4.4.2 具体影响因素分析

通过客户流失模型的建立，可以预测客户的流失情况并采取措施。此外，在实际应用中，还需要对用户的行为和属性等特征进行分析，判断哪些属性与客户的流失行为具有较大的关联性，也就是对影响客户流失的因素进行分析，从而使运营商能够对相应的属性特点采取针对性的措施。

本文通过 Lasso 回归模型对用户属性进行相关度的比较。Lasso 是一种惩罚性的回归方法，可以同时进行子集收缩和变量选择[27]。 Lasso 的原理是利用惩罚函数的

增加判断或减少特征间的共线性，常用于特征选择和稀疏化当中。本文利用 sklearn 库中的 LassoCV 方法，结合自定义函数和绘图工具，通过图像表示特征变量的重要程度，如下图 16 所示。图像中可以看出，是否合约有效用户（IS_AGREE）、有通话天数（CALL_DAYS）、用户性别（CUST_SEX）、用户年龄（CERT_AGE）特征与流失相关性较强，并且主要是负相关，如用户每月有通话的天数越大，流失的概率就越低。但是通过图像也可以看出各特征与用户流失的相关性不是特别大，可以再结合散点图进一步对特征进行分析。

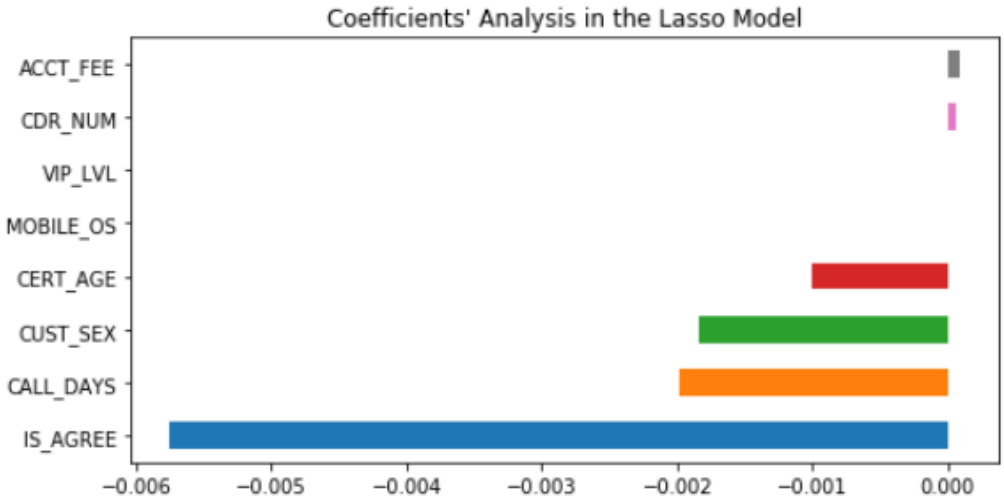


图 16　通过 Lasso 比较特征向量的重要程度

通过 matplotlib 绘图工具，本文对部分特征进行了两两对比分析，利用散点图分析特征对客户流失的影响。如下图 17 为本月费用和通话时长的散点图。

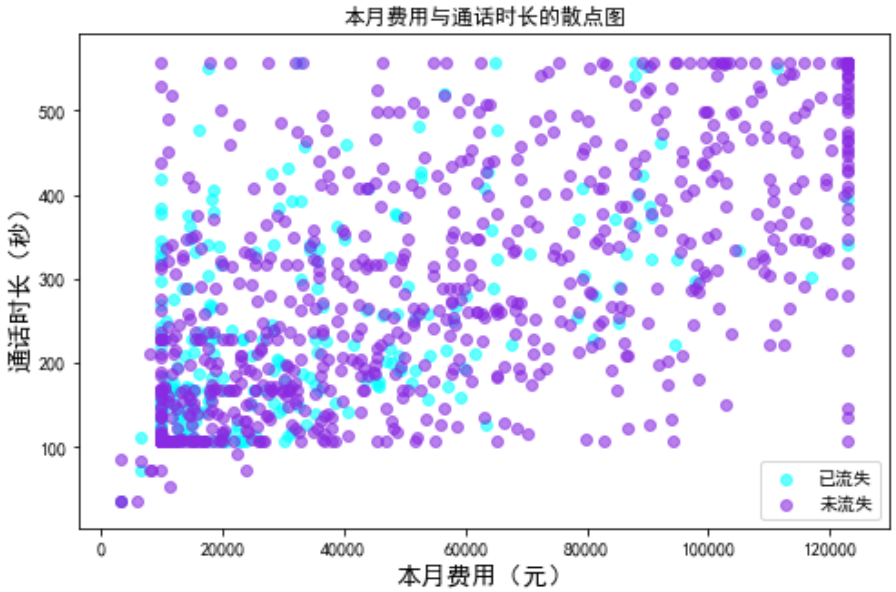


图 17　本月费用和通话时长的散点图

其中，蓝色点表示已流失，紫色点表示未流失，由图可知大部分流失客户的通话时长和本月费用较小，说明经常通话、每月消费较高的用户不易流失，这种现象放在实际运营中也是比较容易理解的。

下图 18 为通话天数与用户年龄的散点图，从图中可以看出非流失用户的年龄大部分集中在 20~50 岁之间，流失用户的年龄主要集中在 10~30 岁及 50~60 岁左右，说明年龄过大或过小的用户流失概率相对较高。同时，由图可以观察到非流失用户的通话天数总体比流失用户的通话天数略长，但是通话天数的差异不如用户年龄带来的差异大，因此用户年龄与用户流失的相关性更强。

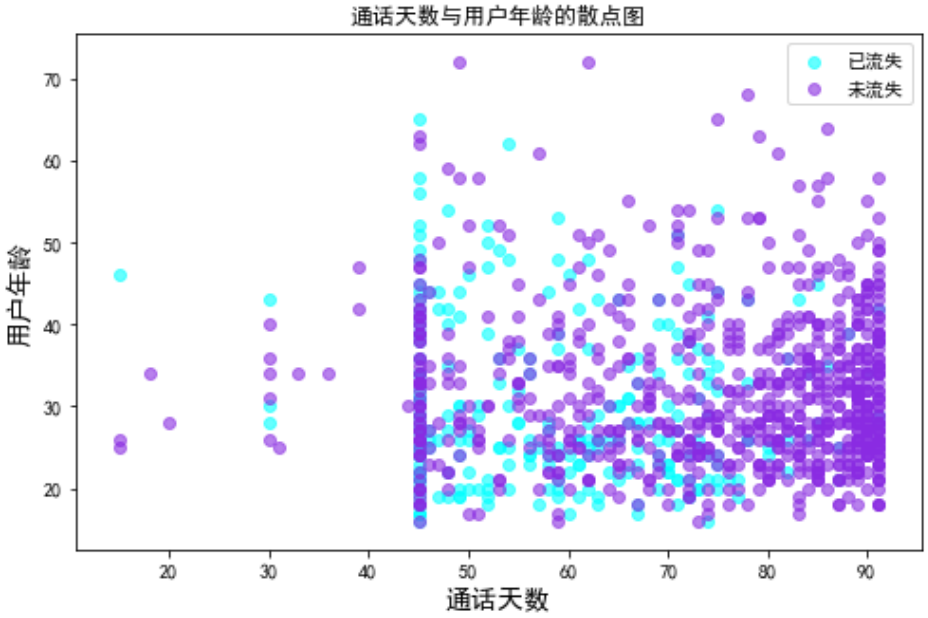


图 18　通话天数与用户年龄的散点图

4.5　应用建议

基于前文的模型分析以及企业运营环境，提出以下几点建议：

（1）针对经常通话、每月消费较高的用户不易流失的情况，可以通过营销套餐的组合、升级套餐优惠、套餐免费体验等方式，提升用户的消费兴趣。特别是对于一些经过模型预测可能流失的用户，可以进一步提供专属服务、个性化套餐等，增强用户的满意度，从而减少客户流失。

（2）针对用户年龄对用户流失概率的影响，对于不同的年龄层，可设计不同的方案。如年龄较小的青少年电话使用频率不高，较易流失，运营商可以与儿童手表、儿童学习机等厂家合作，采取一定的套餐优惠、绑定条件等，提升用户的使用频率；

同时，也可以将青少年的账号与家长、学校进行绑定等，从而能够提供更全面的服务。

（3）针对运营商激烈的竞争环境，运营商企业必须抓住发展机遇，开拓思路，加强合作，可以参与到金融服务、IT 服务、媒体或公用事业等领域当中，利用自身海量的数据资源和电信资源，寻找新的合作机会和收入来源。同时，通过与第三方的合作，也能够为运营商带来更专业的知识，进一步补充运营商强大的客户关系和广泛的覆盖范围。

（4）运营商应加强数字化转型，广泛地学习和应用新技术，培养相关的人才，增强其敏捷性，通过强大的数据分析，及早瞄准目标客户，以测试其喜好和价格敏感性。同时，加强技术基础设施的建设，紧跟 5G 的发展潮流，推动组织和业务的技术革新。

（5）最后，企业需要进一步追求更好的用户体验，以用户为始并回归用户。大多数运营商将客户体验看作是一系列的接触点，也就是客户与业务节点之间的交互，如产品客户服务、销售人员等。但是忽略了客户对公司的整体体验、对服务流程的整体体验。而大数据、云计算、人工智能等技术又能够帮助运营商从用户的角度思考，全方位地优化用户体验，从而为企业带来更大的效益。

5 结论与展望

5.1 结论

本文首先介绍了运营商竞争日趋激烈的背景，明确对运营商客户行为进行分析的意义，并针对该课题，对当前国内外学者的研究成果进行了概述，指出了当前研究的趋势主要是对多种算法的结合和对比应用，并提出了基于 BP 神经网络算法的研究方法和研究思路。

本文通过某运营商客户信息及信息记录的数据集，按照数据清洗、数据标准化、构建模型、模型评估和分析、模型优化和对比的步骤，利用 Python 构建了 BP 神经网络及随机森林模型，通过分析运营商客户行为，为运营商维系客户提出策略。

在数据处理方面，本文通过对原始数据进行清洗、转换、抽取以及标准化，得到干净的训练数据，保证模型能够更好地建立。同时数据中出现了不平衡分类的情况，采取过采样技术对数据进行重新采样，提升了数据的质量，也为模型的精确度奠定了基础。

在模型构建方面，首先建立了 BP 神经网络模型并对其进行了参数优化，最后取得了较好的预测精度。此外，为了更全面地对客户数据进行分析，本文还构建了随机森林模型，并将两种模型进行对比和分析。最终得到结论，两种模型的预测效果都比较好，随机森林模型的精确度越略高于 BP 神经网络模型，二者都有各自的优势和劣势，可以在一定程度上互补优化，需要结合具体的数据情况建立合理的分析模型。

最后在建立好模型的基础上，对客户流失因素进行分析，并为运营商实施营销策略提供建议。针对不同的影响因素，从不同的角度进行分析，提出了是否该制定挽留措施，如何制定挽留措施等相关建议，从而减少客户的流失，提升运营商的行业竞争力。

5.2 展望

本文基于 BP 神经网络，对运营商客户行为进行了分析，并对 BP 神经网络模型和随机森林模型进行了对比和分析，虽然取得了一些成果，但还存在可以改进的地

方，可以从以下几个角度进行进一步的研究：

（1）模型算法的优化方面，BP 神经网络算法是使用非常广泛的一种神经网络算法，但在机器学习领域中，可以用于预测的算法还有很多种，如多层感知机神经网络、增强版神经网络（AdvancedNN）等。后续研究可以向技术方向继续延伸，结合数据特征尝试采用更先进、更多样的机器学习算法来建立客户行为分析的模型。

（2）多种算法的结合方面，本文中对比了 BP 神经网络和随机森林算法的相同和差异，发现二者都有各自的优缺点，后续研究可以利用二者的特征将两种算法进行结合，从而建立模型，吸取二者的优点，并根据所分析的数据特征来进行分析和建模。如在计算机视觉领域，已经提出了 DNDF 结构（Deep Neural Decision Forest），它非常自然地在神经网络中引入了树的结构，又适应于集成算法，今后的研究也可以尝试该技术在客户分析领域方面的应用。

（3）数据结果可视化方面，BP 神经网络算法的可视性不强，还可以结合更多可视化的工具将数据效果更清晰的地呈现出来，使其能够在实际的运营商客户维系中得到更广泛的应用，同时还可以将数据挖掘的结果与企业的客户关系管理系统结合起来，使预测模型得到充分运用，同时也能够根据客户的具体情况，提供更加个性化的维系策略。

参考文献

[1] 祝梓云．电信运营商大数据运用初探[J]．中国新通信,2018,20(23)：86-87．

[2] Berson A，Smith S，Thearling K． Building data mining applications for CRM[C]．New York:Mc Graw Hill，2000．

[3] Siew-Phaik Loke，Ayankunle Adegbite Taiwo，Hanisah Mat Salim， Alan G．Downe． Service Quality and Customer Satisfaction in a Telecommunication Service Provider[C]．2011 International Conference on Financial Management and Economics IPEDR． Singapore: IACSIT Press，2011：24-29．

[4] Shui Hua Hana，Shui Xiu Lua，Stephen C．H． Leung． Segmentation of telecom customers based on customer value by decisiontree model[J]． Expert Systems with Applications，2012，39：3964–3973．

[5] Omar Adwan，Hossam Faris，Khalid Jaradat，Osama Harfoushi，Nazeeh Ghatasheh． Predicting Customer Churn in Telecom Industry using Multilayer Preceptron Neural Networks: Modeling and Analysis[J]．Life Science Journal，2014，11(3)：75-81．

[6] Ammara Ahmed，D Maheswari Linen． A Review And Analysis Of Churn Prediction Methods For Customer Retention In Telecom Industries[C]． 2017 International Conference on Advanced Computing and Communication Systems，INDIA:Coimbatore，2017：1-7．

[7] Farshid Abdi，Shaghayegh Abolmakarem．Customer Behavior Mining Framework (CBMF) using clusteringand classification techniques[J]．Journal of Industrial Engineering International，2019，15(1)：1-18．

[8] 王雷，陈松林，顾学道．客户流失预警模型及其在电信企业的应用[J]．电信科学，2006(09)：47-51．

[9] 顾光同，王力宾，费宇．电信客户流失预警规则及其信度测定实证研究——以云南电信为例[J]．云南财经大学学报，2010，26(06)：94-98．

[10] 周荣鑫，赵娟娟，靳梦华．基于贝叶斯网络的电信客户流失预测分析[J]．软件，2019，40(02)：187-190．

[11] 王观玉，郭勇．支持向量机在电信客户流失预测中的应用研究[J]．计算机仿真，2011，28(04)：115-118+312．

[12] 林勤，薛云．双聚类算法在电信高价值客户细分的应用[J]．计算机应用，2014，34(06)：1807-1811．

[13] 张珠香，骆念蓓．基于生存分析模型的电信客户流失研究[J]．福州大学学报(哲学社会科学版)，2018，32(01)：50-56．

[14] 陈思含．中国移动通信公司客户流失预测[D]．贵阳：贵州财经大学，2019．

[15] 王志阳．基于客户洞察的电信数据挖掘研究[D]．济南：山东大学，2017．

[16] Salvador García，Sergio Ramírez-Gallego，Julián Luengo，José Manuel Beníteza，Francisco Herrera． Big data preprocessing： methods and prospects [J]．Big Data Analytics，2016，1：9．

[17] Fadi Thabtaha，Suhel Hammoudb，Firuz Kamalovc，Amanda H． Data Imbalance in Classification： Experimental Evaluation [J]．Information Science，2019，11．

[18] Nathalie Japkowicz， Shaju Stephen． The class imbalance problem： a systematic study． Intell． Data Anal． 2002．

[19] 孙碧颖. 基于神经网络算法构建电信用户流失预测模型的研究[D]. 兰州：兰州大学，2016．

[20] 危明铸,麦伟杰,袁峰,沈凤山.基于机器学习的企业运行风险研究[J].软件,2019,40(08)：29-37.

[21] 黄红梅，张良均．Python 数据分析与应用[M]．张凌，施兴，周东平．北京：人民邮电出版社，2018．

[22] 王奕森，夏树涛．集成学习之随机森林算法综述[J]．信息通信技术，2018，12(01)：49-55．

[23] Andy Liaw，Matthew Wiener．Classification and Regression by RandomForest [J]．R News，2002，12(2)：18-22．

[24] 吴地尧,章新友,张玉娇,牛晓录.分类算法在中药研究中的应用及其进展[J].科学技术与工程,2019,19(35)：1-9.

[25] 李苍柏，肖克炎，李楠，宋相龙，张帅，王凯，楚文楷，曹瑞．支持向量机、随机森林和人工神经网络机器学习算法在地球化学异常信息提取中的对比研究[J]．地球学报，2020，41(02)：309-319．

[26] 曾宇哲，吴媛博，郑宏远，罗来娟．基于机器学习的车险索赔频率预测[J]．统计与信息论坛,2019,34(05)：69-78.

[27] Kirkland Lisa，Kanfer Frans，Millard Sollie．LASSO Tuning Parameter Selection [C]. The South African Statistical Association's 57th Annual Conference，2015：49-56．

附录

附录 A　开题报告

毕业设计（论文）开题报告

学　院	信息工程学院	教学系	商务信管系	专业班级	电子商务 1602 班
学生姓名	郝园媛	学号	20164751	指导教师	周艳聪
毕业设计（论文）题目	基于神经网络的运营商客户行为分析				

一、选题依据

1、 课题来源，选题意义和目的

在这个信息爆炸、技术高速发展的时代，大数据已经逐渐成为新的产业革命核心，为企业进一步利用数据创造新价值提供技术支撑，为产业优化带来新的活力。通过科学的数据分析工具采集、处理、分析企业的相关客户数据，不仅能够为企业节省营销成本，而且能够有效提升企业竞争力，为企业创造更大价值。BP 神经网络模型具有很强的非线性动态处理能力，在不知道输入与输出关系的情况下，也能实现高度的非线性映射。作为一种结构简单、可塑性强的数据分析工具，它在图像处理、信号处理、模式识别等众多领域都得到了广泛应用。利用 BP 神经网络算法，根据运营商的用户日志，按照事件的数据状态选取样本，构建合理的 BP 神经网络模型，根据模型和数据分析结果实施有针对性的营销策略，可以大大提高运营效率，带来更多利润。怎样构建 BP 神经网络模型，分析处理运营商的相关客户数据，从而分析客户行为，为企业带来更大价值，正是本课题的研究目的所在。

2、 文献综述

由于国外电信行业兴起较早，许多经济发达国家已经在运营商客户分析方面有了相对成熟的研究。在许多数据分析算法中，K-means 算法、粒子群优化算法、模糊聚类算法、神经网络算法、随机森林树算法都使用广泛。

国外研究方面，Berson 认为通信领域的客户流失意味着，客户从原有电信运营商转向了另一家公司，这使原运营商减少了用户数量，而对手增加了用户。Boone 等在对客户进行分类时，通过人工神经网络准确地识别和分类客户组，并针对不同的客户类型实施相应的业务方法。Christoph Heitz 等人采集了电信运营商客户的基本数据以及消费数据，使用数据挖掘技术合理地对离网用户进行了聚类，为电信公司建立了客户流失预测模型，并将分析结果应用于流失用户的

维护，从而提高公司利润。

国内研究方面，顾光同等人在研究和分析云南一家电信运营商的相关客户数据时，采用决策树算法分析流失客户特征，更准确地分析了企业中流失客户的行为特征，并且在此基础上，找到了电信业客户流失预警的基本规则，计算出了预测的准确率。王观玉等人针对通信公司客户数据多维度的特点，首先利用了主成分分析法对数据进行降维处理，以找到影响客户流失的关键指标，再采用非线性支持向量机进行学习建模，以第一阶段的主成分为输入，通过主成分分析结合支持向量机的方式，有效提高了预测客户流失的准确性，为运营商客户流失预测提供了一种新方法。

随着数据挖掘技术的发展，在电信客户分析中，研究者采用的工具逐渐增多，模型的精度也逐渐提高。陈思含在进行移动公司的客户分析中，建立了距离判别模型、支持向量机、神经网络三种模型，并采用 k-means 聚类、基于随机森林的影响因素分析、逻辑回归分析等方法，寻找运营商流失客户的行为特征，并比较和优化模型不断提高模型精度，使模型在实际应用中具有更好的适用性。王学文等人基于 MATLAB，通过改进的 BP 神经网络算法，结合电信客户流失风险指标体系，建立了客户流失模型，预测未来客户的流失情况。可以看出，BP 神经网络等算法被广泛应用于客户行为分析中，尤其是客户流失、客户价值方面的分析。

综合国内外相关文献，在对各运营商的客户分析和研究中，数据挖掘技术和其相关派生模型的应用十分广泛。在进行客户行为分析时，新方法和新渗透领域必将成为未来数据挖掘技术和方法不断发展的两个主要方向。随着研究者对客户行为研究的不断深入，用于客户数据分析的方法类型将越来越丰富，在不同行业的客户关系管理中使用数据挖掘方法的方法将越来越成熟，未来的数据挖掘技术将越来越高效，为相关领域的决策带来便利，减少人力物力投入，有利于企业用更加精准的策略获得更高的效益。

二、研究内容和研究方法

1、 研究内容

BP 神经网络也称反向传播网络，是一种按误差逆传播算法训练的多层前馈网络，其算法称为 BP 算法。BP 神经网络的思想是利用输出层第 N 层输出后的误差，估计输出层前第 N-1 层的误差，再用其估计第 N-2 层的误差，以此获取所有层误差估计。通过对运营商用户数据处理分析，构建 BP 神经网络模型，对企业提高决策效率和准确度有重要意义。

本文将对运营商数据进行研究分析，选取适宜的数据作为样本，并对样本进行标准化处理，再进一步构建神经网络模型并利用模型对客户行为进行分析。

本文拟从以下几个方面开展研究：

（1）浏览国内外最新文献资料，对 BP 神经网络原理进行系统学习，并能选择合适的算法，进行算法分析，熟悉其结构特点。

（2）对目前基于神经网络模型的数据分析原理、分析过程进行研究综述，并能够进行不同方案之间的对比分析。

（3）对运营商客户数据进行清理和筛选，对运营商提供的数据进行去重，实现对缺失值与异常值的识别和处理，并根据类别特征进行内部类别合并。

（4）构建神经网络预测模型并对模型进行评估，提取出能够辅助决策的关键性数据，与随机森林树算法的建模结果进行对比，进一步实现客户行为分析，对客户流失、客户价值等进行分析。

2、 研究方法

根据研究内容，经比较后暂定本次研究中的方法采用数据集划分方法、数据标准化方法、神经网络算法。

数据集划分方法：在实际任务中，往往需要一个“测试集”来检测模型对新样本的判别能力，然后以测试集上的“测试误差”作为泛化误差的近似，这里假设测试集是从样本真实分布中独立同分布采样得到的。假设学习任务中的有一个包含 m 个样本的数据集 $D = \{(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)\}$，通过适当的处理，从中产生出训练集 S 和测试集 T，要求 T 与 S 尽可能互斥，即测试样本尽量不在训练集中出现、未在训练过程中使用过。常见的数据集划分方式主要有留出法、交叉验证法和自助法。自助法在数据集较小、难以有效划分训练或测试集时效果较好。此外，自助法能从初始数据集中产生多个不同的训练集，这对集成学习等方法有很大好处。但是，自助法产生的数据集改变了初始数据集的分布，这会引入偏差。因此，在初始数据集足够多时，留出法和交叉验证法更加常用。

数据标准化方法：为了保证结果的可靠性，需要对原始数据进行标准化处理，减少异常数值

带来的影响。数据的标准化是将数据按比例缩放，使之落入一个小的特定区间。在某些比较和评价的指标处理中较为常用，去除数据的单位限制，将其转化为无量纲的纯数值，便于不同单位或量级的指标能够进行比较和加权。其中最典型的就是数据的归一化处理，即将数据统一映射到[0,1]区间上，常见的数据归一化的方法有：min-max 标准化，也叫离差标准化，是对原始数据的线性变换，使结果落到[0,1]区间。

BP 神经网络算法：BP 网络是一种按误差逆传播算法训练的多层前馈网络，是应用最广泛的神经网络模型之一。BP 网络能学习和存贮大量的输入-输出模式映射关系，且无需描述这种映射关系的数学方程。它使用最速下降法的学习规则，通过反向传播来不断调整网络的权值和阈值，使网络的误差平方和最小。其结构包括输入层、隐含层和输出层。BP 神经网络算法可以逼近任意连续函数，具有很强的非线性映射能力，而且网络的中间层数、各层的处理单元数及网络的学习系数等参数可根据具体情况设定，灵活性很大，在许多应用领域都发挥了重要作用。为了解决 BP 神经网络收敛速度慢、网络的中间层及它的单元数选取无理论指导及网络学习和记忆的不稳定性等缺陷，提出了许多改进算法。

3、 关键性技术及解决方法：

在本研究过程中，最核心也是较难实现的是 BP 神经网络模型的建立和客户行为分析。

（1）模型的建立

面对海量的数据，构建一个合理的 BP 神经网络模型工作量较大，且需要保证模型一定的精确度。

解决方法：构建神经网络模型时需要注意数据本身特征之间存在量级差异，需要进行标准化以消除量级差异。此外，为了验证模型的可用性，需要对所构建的模型进行评估，形成模型评估报告，并比较精确率，从而不断优化模型。

（2）客户行为分析

以往对客户的行为分析都是从客户历史行为数据中入手，利用可视化的数据进行分析，本次结合 BP 神经网络模型进行分析，相对来说难度较大。

解决方法：首先阅读相关文献，进行大量学习，对 BP 神经网络模型有更深入的了解，熟悉神经网络模型的特征，在不断的学习过程中，选择合适的分析方法，充分利用神经网络模型的优势，为客户行为分析助力。

三、预计可获得的成果

当前，人工神经网络技术正热，将 BP 神经网络模型与客户行为分析结合对企业实现产业优化、提升行业竞争力具有重要的研究意义。论文预期会对数据进行采集和处理，构建合理的 BP 神经网络模型，并进行客户行为分析，最终完成一篇相对优秀的毕业论文。

四、工作进度计划

2019 年 11 月	教师和学生进行双向选题，明确课题任务。
2019 年 12 月	撰写开题报告，检查初步分析，进一步完善课题任务分析
2020 年 1-2 月	完成课题任务分析，开始课题方案设计
2020 年 3 月	基本完成课题方案设计，逐步开始课题方案实现、对比分析等
2020 年 4 月	课题设计与方案实现完毕，全面进入毕业设计说明书的撰写
2020 年 5 月	毕业设计预答辩，撰写毕业设计论文。
2020 年 6 月初	毕业设计答辩

五、与开题有关的主要参考文献

[1] Aldo Okullo,Noah Tibasiima,Joshua Barasa.Modeling and Simulation of an Isothermal Suspension Polymerization Reactor for PMMA Production Using Python[J].Advances in Chemi-cal Engineering and Science,2017,74:408-419

[2] Verduzco-Flores Sergio,De Schutter Erik.Draculab: A Python Simulator for Firing Rate Neural Networks With Delayed Adaptive Connections[J].Frontiers in neuroinformatics,2019,13:18

[3] Greb Fabian,Steffens Jochen,Schlotz Wolff.Modeling Music-Selection Behavior in Everyday Life: A Multilevel Statistical Learning Approach and Mediation Analysis of Experience Sampling Data[J].Frontiers in psychology,2019,10:390

[4] Fontaine Rafamantanantsoa,Maherindefo Laha.Analysis and Neural Networks Modeling of Web Server Performances Using MySQL and PostgreSQL[J].Communications and Network(CN),2018,104012:142-151

[5] 聂晶.Python 在大数据挖掘和分析中的应用优势[J].广西民族大学学报(自然科学版),2018,24(01):76-79

[6] 李俊华.基于 Python 的数据分析[J].电子技术与软件工程,2018(17):167

[7] 丰玄霜.基于数据挖掘的用户上网行为分析[D].北京:中央民族大学,2016

[8] 陈虹坚.简述基于 Python 的电商公司用户价值分析[J].现代营销(经营版),2018(08):111

[9] 李嘉玲,蒋艳.基于 BP 神经网络的公共建筑用电能耗预测研究[J].软件导刊,2019,18(07):49-52

[10] 曾武序,钱文彬,王映龙,杨文姬,柳军.一种基于 Python 和 BP 神经网络的股票预测方法[J].计算机时代,2018(06):72-75+80

[11] 刘磊,蔡欣桦,许锐强.基于华为大数据平台的车辆销售数据分析应用[J].现代计算机(专业版),2019(03):91-96

[12] 顾彬,汪柯颖,沈东,苏超.基于 BP 神经网络的模拟电路诊断[J].科技创新与应用,2019(31):57-58

[13] 昝军才,魏成才,蒋可娟,吴雨欣,袁超.基于 BP 神经网络的煤自燃温度预测研究[J].煤炭工程,2019,51(10):113-117

[14] 齐慧.基于 python 的 WEB 数据挖掘技术实现与研究[J].软件工程,2019,22(08):21-23

[15] Lei Lei. Wavelet Neural Network Prediction Method of Stock Price Trend Based on Rough Set Attribute Reduction[J].Applied Soft Computing,2018(62):923-932

[16] YALCINTAS M. An energy benchmarking model based on artificial neural network method with a case example for tropical climates[J]. International Journal of Energy Research,2006(30):1158-1174

[17] TSO G K F,YAU K K W. Predicting electricity energy consumption - a comparison of regression analysis，decision tree and neural network[J]. Energy,2007(32):1761-1768

[18] 许慧,黄世泽,郭其一,等.基于改进 BP 网络的建筑用电量预测[J].电器与能效管理技术,2014(18):54-58

[19] 王雅楠,孟晓景.基于动量 BP 算法的神经网络房价预测研究[J].软件导刊,2015,14(2):59-61.

[20] 马文斌,夏国恩.基于深度神经网络的客户流失预测模型[J].计算机技术与发展,2019,29(09):76-80

[21] 姜健,陈婧伊,王宇歌.基于 BP 神经网络的客户信用评级模型[J].中国商论,2018(18):172-173

[22] 王光,张兰君.基于 BP 人工神经网络的智能银行客户满意度研究[J].经济研究导刊,2018(16):83-84+125

[23] 王学文,单其帅,魏彦凤.基于 BP 神经网络的电信客户流失风险预测[J].科技视界,2013(28):113-114

[24] Berson A,Smith S,Thearling K.Customerretention[A].Building data mining applications for CRM[C].New York:Mc Graw Hill,2000

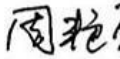
[25] 顾光同,王力宾,费宇.电信客户流失预警规则及其信度测定实证研究——以云南电信为例[J].云南财经大学学报,2010,26(06):94-98

[26] 王观玉,郭勇.支持向量机在电信客户流失预测中的应用研究[J].计算机仿真,2011,28(04):115-118+312

[27] 王志阳. 基于客户洞察的电信数据挖掘研究[D].济南:山东大学,2017

[28] 朱琦,朱正键,陈伦楷.基于随机森林算法的客户流失模型[J].电信快报,2019(04):19-21.

[29] 陈思含. 中国移动通信公司客户流失预测[D].贵阳:贵州财经大学,2019.

指导教师意见
同意本课题进入设计（论文）阶段。 指导教师签字：  2020 年 3 月 11 日

说明：1.本报告必须在第八学期开学两周内经指导教师审阅并形成正式报告。

2.本报告作为指导教师审查学生能否开展课题研究和是否按时完成进度的检查依据，并接受学校的抽查。

附录 B　代码清单

见另册。

致谢

本次毕业设计在天津商业大学信息工程学院完成，感谢信息工程学院的老师及同学在本科期间给我的帮助。

首先，感谢本次毕业设计的指导老师周艳聪。在本科期间，周老师在学业和生活方面给予了我很多的帮助，从上课到大创项目再到毕业设计，她一直都十分耐心地指导和鼓励我。在毕业设计的完成过程中，她思路开阔，在研究思路上给予了我方向性的指导并提供了建设性的意见。在周老师的悉心指导下，我的学术水平得到了一定的提升。周老师严谨的科学态度和一丝不苟的学术精神一直鼓励着我，使我受益匪浅。在此即将完成毕业论文之际，谨向周艳聪老师致以衷心的感谢和深深的敬意，祝周老师平安喜乐，万事顺利。

其次，感谢在本科期间给予我无数帮助的各位老师们，感谢老师们不辞辛苦地传道、授业、解惑，使我受到了知识的滋养。感谢我的学长学姐以及各位同学们，使我的本科生活充实而又快乐。

再次，感谢我的家人，他们默默给予了我许多关爱，他们对我的支持和鼓励是我一直不断前进的动力。

最后，感谢参与本文评审的各位老师在百忙之中指导我的论文。

时光飞逝，四年的本科学习即将结束，在学校的时光对我来说十分珍贵，值此建校四十年之际，祝天津商业大学越来越好，再谱华章！我也会不忘初心，继续前进，谱写人生的新篇章！