Statistics and Application 统计学与应用, 2019, 8(5), 784-796
Published Online October 2019 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2019.85089

Hans汉斯

Insurance Customer Purchase Prediction Based on Data Optimization

Shasha Li

University of International Business Economy, Beijing
Email: 18235701863@163.com

Received: Oct. 2nd, 2019; accepted: Oct. 18th, 2019; published: Oct. 25th, 2019

Abstract

In recent years, with the general improvement of people's living standards, the insurance industry ushered in a new spring. The extensive business model has been unable to meet the requirements of the increasing development of insurance companies. How to get rid of the traditional way of marketing, quickly discover valuable customers and keep up with the market, is becoming more and more important for insurance companies. This article uses customer data from a life insurance company. Firstly, descriptive statistical analysis was conducted based on the given basic information of customers, call information, insurance information and risk donation information, etc., to view the data situation, and data cleaning was carried out to improve the data quality. Secondly, a separate logistic regression model is used for learning to generate a feasibility analysis report. Then, the combined model of decision tree and logistic regression and the combined model of random forest and logistic regression were respectively used for prediction. Finally, a comparison of the three models shows that the combined model of random forest and logistic regression is more effective.

Keywords

Insurance, Logistic Regression, Decision Tree, Random Forest, Combination Model

基于数据优化的保险客户承保预测

李莎莎

对外经济贸易大学，北京
Email: 18235701863@163.com

收稿日期：2019年10月2日；录用日期：2019年10月18日；发布日期：2019年10月25日

摘　要

近年来，人民生活水平的普遍提高，使得保险行业迎来了新的春天。一直以来的粗放式经营模式已经无

文章引用: 李莎莎. 基于数据优化的保险客户承保预测[J]. 统计学与应用, 2019, 8(5): 784-796.
DOI: 10.12677/sa.2019.85089

法满足保险公司日益发展的要求。如何摆脱传统的营销方式，快速发掘出有价值的客户，在市场中不远远落后，对于保险公司来说越来越重要。本文使用某人寿保险公司的客户数据。首先，基于给定的客户基本信息、通话信息、投保信息、赠险信息等进行描述性统计分析，查看数据情况，对数据进行数据清洗提升数据质量；其次，使用单独的逻辑回归模型进行学习，生成可行性分析报告；然后，分别使用决策树与逻辑回归的组合模型以及随机森林与逻辑回归的组合模型进行预测；最后，将三种模型进行对比发现随机森林与逻辑回归的组合模型效果更好。

关键词

保险，逻辑回归，决策树，随机森林，组合模型

1. 绪论

1.1. 研究背景及意义

近年来，居民收入水平的普遍提升以及国家政策的有力扶持带动了保险行业的高速发展。保险行业是国民经济的重要组成部分，中国保险业在世界保险行业中愈来愈占据重要的比重，仅次于美国。中国保险行业的发展对于我国国家经济的发展具有巨大的推动作用。如图 1 和图 2 是来自最新的国家统计局的保险行业数据，该数据展示了 2014~2018 年保险行业的保费收入状况和保险行业资产状况。

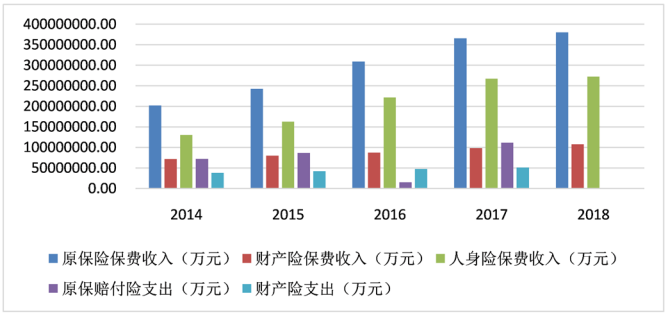

Figure 1. Insurance premium income status chart
图 1. 保险业保费收入状况图

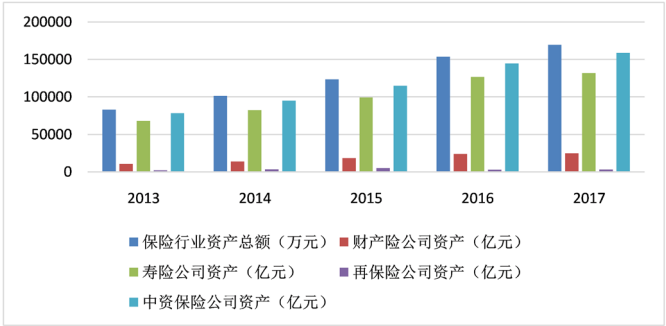

Figure 2. Asset position chart of insurance industry
图 2. 保险行业资产状况图

数据表明，近年来保险行业发展依旧呈现稳步提升的趋势，保险收入(包含财产险和人身险)从 2014 年的 1.7 万亿元到 2018 年的 3.83 万亿元，年化复合增速高达 25.3%，同时保险业总资产从 2013 年的 8.5 万亿元增长至 2018 年的 17.1 万亿元。良好的发展趋势使得保险行业竞争加剧，客户市场的抢占成为保险公司竞争的主要手段。

随着大数据时代的到来，互联网行业快速掀起了数据化的浪潮。保险行业作为传统的金融行业，渐渐呈现落后的趋势。目前的保险公司虽然拥有丰富的数据源，但缺乏有效的分析手段，虽采取了一定的数据挖掘手段，然而数据分析人员知识和技术严重匮乏。无法对大量数据进行有效的信息提取，从而无法为决策者提供有效的决策支持。

目前的保险公司对于保险客户的分析主要通过经验的分析或者基于统计的简单划分或者基于粗糙数据集的模型带入。这种方式虽然具有一定的指导意义，但缺乏实际的理论分析，并不能实现对目标客户的精准预测，也无法满足企业的要求。

本文在经验分析和统计分析的基础之上进行了巨大的改进。首先，请业务专家根据业务经验从海量的客户数据中选出对客户承保影响较大的关键指标；然后，请分析专员根据数据工作经验，基于数据的基本特征以及后期的数据处理和建模影响选择出重要指标，实现目标客户的精准识别；后期根据数据部署应用进行后期效果的监测和反馈。这种分析方法将传统的经验分析和统计分析结合，使用机器学习模型进行指标选择和模型测试，为预测提供了有力的理论支撑，预测可信度极大加强，不仅有利于挖掘出更多的潜在投保客户，还为公司带来了巨大的收益。

1.2. 国内外研究现状简介

保险客户承保预测的准确性直接影响着保险公司业务的推广，公司的收益以及保险客户的用户体验。因此，从理论研究还是实践研究来说都具有重要意义。近年来，随着大数据技术的不断发展，机器学习的各种模型算法使用越来越广泛。在保险方面的应用的相关文章也逐渐增多。

早期的保险相关研究多数是在 CRM (客户关系管理)的基础上运用一些数据仓库、数据挖掘技术，发现一些潜在的规律，进而知道企业产品销售和客户服务行为[1] [2]。或者运用客户分类的一些经验分类方法或基于统计的简单划分，按照客户简单的属性或者决策者的经验来划分，进而指导企业对客户的管理。这些方法虽然能起到一定的指导作用，但是在挖掘客户的潜在价值，客户的资信程度方面缺乏深入分析[3]。

郭宜斌等人通过横向关联的方法去发掘两者之间的相关关系，比如说在超市中购买 A 的人同时也购买了 B，通过对发现的两者之间的过高相关系数去挖掘其高度相关的原因。王贵龙[4]等人将关联向量机运用到客户识别中，构建基于关联向量机的保险客户识别模型。

柯新喜[3]等人将决策树模型运用到保险预测中，通过分析客户的性别，年龄，家庭收入等因素，构建分类模型，选择重要特征，以便后续公司业务人员对数据进行分析，挖掘出潜在客户。

除此之外，聚类分析也是常采用的方法之一[5]。聚类作为分类模型的逆向方法，在数据不知道分成几类的情况下，根据数据之间的差异性合理划分成几类。在其他领域有关客户分类的研究，采用神经网络模型、支持向量机、模糊聚类、遗传算法等一系列数据挖掘和智能学习技术也受到了越来越多的关注。

本文同样使用了分类的方法，与上述研究的不同之处在于将在其他领域使用成熟的机器学习相关模型进行模型融合，应用到了保险领域。首先，对所给数据进行可视化，查看数据的分布情况；然后，对数据进行预处理工作，包括异常值、重复值、缺失值的处理，数据的离散化，归一化，数据的平衡处理等工作；之后，分别使用决策树和随机森林模型对数据进行特征筛选；最后，将选择好的属性分别送入逻辑回归模型中进行数据建模，并进行模型评估对比查看两种组合的优劣。

1.3. 研究框架

本论文按照如下章节展开：

第一章介绍了客户保险承保问题的研究背景及意义，国内外研究现状以及本文的章节安排。

第二章主要介绍了本文中使用的各种方法的技术理论，包括决策树，随机森林，逻辑回归模型以及模型的评价标准 Accuracy 和准确率，召回率，f1-score。

第三章主要对数据进行了介绍，并简要给出了整个数据预处理阶段进行的工作，包括缺失值填充，重复值和异常值处理，数据的平衡处理等。

第四章主要根据处理好的数据分别使用决策树和随机森林模型对传统的逻辑回归模型进行改进，进而提高模型的准确率和召回率，提升模型的效果。

最后对本文进行总结与展望。

2. 相关方法介绍

2.1. 决策树模型

2.1.1. 模型简介

决策树模型的本质就是从训练数据集中归纳出一组分类规则。即从原始数据集开始，不断循环的寻找使得数据集中数据分类最干净的属性，按照这个属性不断进行分类，直到最后的数据分类比较干净为止。

决策树分为分类树与回归树两种。本文中主要运用的是分类树，决策树在通过属性进行分类不断生长，分类的过程就是树不断生长的过程。最后分类完成后会生长出一颗倒立的树。其中最上面的顶点我们称之为根结点，内部结点代表进行分类的特征或属性，树枝我们称为有向边。

2.1.2. 决策树的生成过程：

a) 给定训练集样本；

b) 寻找众多属性特征中使得分类不纯度降低最大或分类纯度增加最大的特征；

c) 根据该属性对数据集进行分类；

d) 不断重复 b,c 过程生成决策树。如图 3 即为决策树的样式：

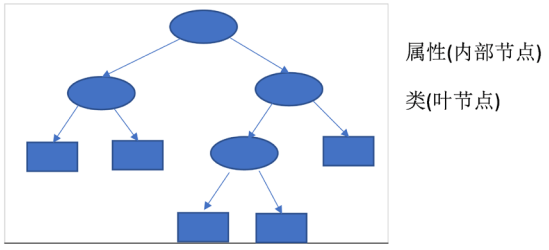

Figure 3. Decision tree generation diagram
图 3. 决策树生成图

2.1.3. 决策树的优缺点

决策树具有以下一些优点：

1) 决策树思想简单易于理解，容易提取规则；

2) 可用于处理具有缺失属性的样本；

3) 可用于处理不相关特征且运行速度较快。

下面是对决策树缺点的总结：

1) 容易产生过拟合，泛化能力较弱；
2) 容易忽略数据中属性间的关联性；
3) 如果类别太多，决策树会错分的情况会增加的比较快。

2.2. 随机森林模型

2.2.1. bagging 简介

集成学习可以通过组合策略提高预测方法的可靠性.集成学习主要有两大类 boosting 和 bagging。boosting 是串行生成基学习器的过程，基学习器之间存在较强的依赖关系。bagging 是并行式生成学习器的典型，相比于 boosting，bagging 中的基学习器之间相对独立些。

给定一个训练数据集，对训练样本进行自助采样生成 N 个包含 m 个样本的采样集，基于每一个采样集进行训练生成一个基学习器，最后将这些基学习器进行结合生成一个强学习器的过程就是 bagging。

2.2.2. RF 简介

随机森林是 bagging 的典型代表。随机森林在 bagging 的基础之上，选择决策树作为基学习器进行训练，最后将多棵决策树进行结合生成的强学习器的过程即为随机森林。随机森林中回归树的剪枝操作可以有效降低过拟合的风险，它简单高效，容易实现，计算开销小，在很多分类回归问题中展现出强大的性能[6]。

2.2.3. 随机森林的构建过程：

a) 从训练集样本中进行 bootstrap 抽样，生成 n 个子集；
b) 将这 n 个子集分别作为训练集训练处多棵决策树；
c) 运用测试集进行预测输出，对每棵决策树使用简单投票法获得分类结果。如图 4 即为随机森林的构建过程：

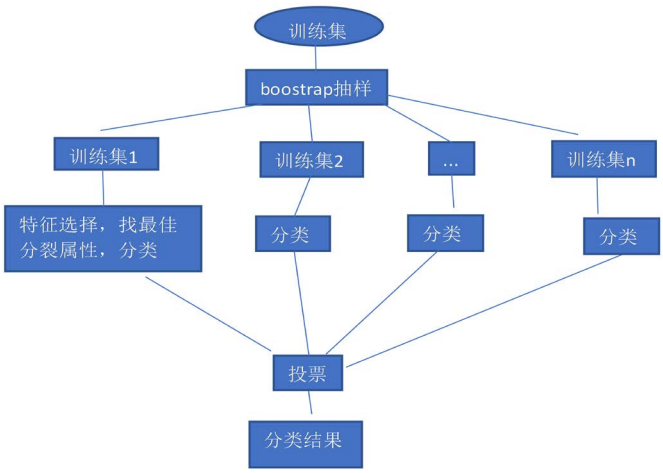

Figure 4. Random forest generation diagram
图 4. 随机森林生成图

2.2.4. 随机森林的优缺点

随机森林的优点[7]：
1) 能够很好的解决过拟合问题，抗噪性强，在数据集上表现良好；
2) 对数据集的适应力强，能够处理高维数据；

3) 既能处理离散型数据也能处理连续性数据；

4) 思想清晰，便于理解。

随机森林的缺点：

1) 对数据量少，数据维度低的数据效果不一定好；

2) 执行速度相比于 boosting 模型快，但相对于决策树模型较慢；

3) 在某些噪音较大的分类或回归问题上会过拟合。

2.3. 逻辑回归模型

2.3.1. 逻辑斯谛分布

设 X 是连续随机变量，X 服从下列分布函数和密度函数：

$$F(x)=P(X\le x)=\frac{1}{1+\mathrm{e}^{-(x-u)/\gamma}}$$

$$f(x)=F'(x)=\frac{\mathrm{e}^{-(x-u)/\gamma}}{\gamma\left(1+\mathrm{e}^{-(x-u)/\gamma}\right)^2}$$

我们称 X 服从逻辑斯谛分布[8]。

2.3.2. 逻辑斯谛回归模型

逻辑回归模型是由统计学家 David Cox 提出的[9]。它是一种分类模型，由条件概率分布 $P(Y\mid X)$ 表示，形式为参数化的逻辑斯谛分布[8]。其中，X 的取值为实数，Y 的取值为 0 或 1。逻辑斯谛回归模型满足如下条件概率分布：

$$P(Y=1\mid x)=\frac{\exp(w\cdot x+b)}{1+\exp(w\cdot x+b)}$$

$$P(Y=0\mid x)=\frac{1}{1+\exp(w\cdot x+b)}$$

设 $P(Y=1\mid x)=\pi(x)$，$P(Y=0\mid x)=1-\pi(x)$；

似然函数为 $\sum_{i=1}^{N}\left[\pi(x_i)\right]^{y_i}\left[1-\pi(x_i)\right]^{1-y_i}$；

对数似然函数为 $L(w)=\sum_{i=1}^{N}\left[y_i(w\cdot x_i)-\log\left(1+\exp(w\cdot x_i)\right)\right]$。

对 $L(w)$求极大值，得到 w 的估计值。继而得到逻辑回归模型，求出客户承保的最大似然预测估计：

$$y=\frac{1}{1+\mathrm{e}^{-wx}}$$

2.3.3. 模型的优缺点

逻辑回归模型作为一种经典的数据模型，受到各个领域的广泛青睐。如下是逻辑回归模型的优缺点：

逻辑回归模型的优点：

1) 逻辑回归模型最大的优点是模型极具可解读性[9]；

2) 训练速度快；

3) 占用内存小。

逻辑回归模型的缺点：

1) 准确率不够高；

2) 数据平衡问题较难处理；

3) 无法筛选特征。

2.4. 评估指标

2.4.1. Accuracy

AUC (Area Under Curve)值指的是 ROC 曲线下方的面积大小[10]。ROC 曲线一般位于 $y=x$ 上方，因此 AUC 的取值范围一般在 0.5 和 1 之间[11]。在对角线之上的区域越大则 AUC 的值越大。AUC 的值越大，分类效果越好，一般而言：

当 AUC = 1 时，分类器是完美分类器，采用这个预估模型不管在任何标准和阈值下都能得到完美的结果。然而现实情况下几乎不存在这样没有任何偏差的完美划分。

当 0.5 < AUC < 1 时，分类器模型在阈值设定的情况下能够发挥预估价值。

当 AUC = 0.5 时，跟随机猜测结果没有区别，模型没有构建的必要。

当 AUC < 0.5 时，虽然从结果上看甚至劣于随机猜测，但如果从反面利用该值，取得的效果还是比随机猜要好的。

2.4.2. 精确率、召回率和 F1 值

精确率(precision)指的是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，即 $P=TP/(TP+FP)$。在本数据集中表示在预测为承保的用户中，预测正确的(实际也是接受保险的)用户占比。

召回率(recall)是覆盖面的度量，度量有多个正例，被分为正例的样本数占总的正例样本数的比例。即 $recall=TP/(TP+FN)$，在本数据集中表示在实际为承保的用户中，预测正确的(预测为接受的)用户占比。

F1 值为精确率和召回率的调和均值，相当于这两个的综合评价指标。

通过输出的分析报告可以得出建立的预测模型的精确率，得出在预测为接受保险的用户中，实际接受保险用户的占比，而召回率说明实际为接受保险的用户中，预测为接受保险的用户占比，F1 值表示对模型的综合评价。

3. 数据集

3.1. 数据介绍

本文研究采用某寿险公司的真实数据集，以此来确保研究的真实性和完整性。该数据集包含 100 万条数据。数据特征包括用户的基本信息、保险投放信息、客户拨打信息、结束码信息、其他信息等，ISCB 是标签字段，取值分别为 0 (不承保)和 1 (承保)。数据集中数据主要分三类：二分类数据，多分类数据以及连续型数据。实验通过 python 对数据进行分析处理，如表 1 是数据集的字段及对应的特征：

Table 1. Customer data feature fields
表 1. 客户数据特征字段

字段	特征	字段	特征
DMCIF_CUST_ID	用户 id	TC_PHONESTATE	号码呼叫状态
SEX	性别	TC_DAILSTATE_PRIOR	客户联系状态
MARRIAGE	婚否	TC_DAILSTATE_LAST	最后客户联系状态
BANNEDORNOT	是否禁播	TC_SALEPROCESS_PRIOR	最初销售进程
CALLLISTCNT	总批次数	TC_SALEPROCESS_LAST	最后销售进程

Continued

CALLLISTCNT_BSJ	白数据批次数	TC_POSSIBILITY	成交可能性
CALLLISTCNT_YZ	优质数据批次数	TC_HEALTHTOLD	是否健告
CALLLISTCNT_XH	循环数据批次数	TC_WRONGINFO	信息错误
CALLLISTCNT_JB	加保数据批次数	TC_ACCEPTED	接受赠品
CALLLISTCNT_QT	其他数据批次数	TC_ENDING	终止联系
CALLLISTCNT_NOTAKL	最后连续未接通批次数	FREESENDCNT	赠险数量
FIRST_CALLLIST_DATATYPE	第一批次数据大类	DAILCNT_LASTCL	最后一个批次拨打次数
LAST_CALLLIST_DATATYPE	最后批次数据大类	TALKCNT_LASTCL	最后一个批次通话次数
DAILCNT	拨打总次数	AVG_TALKLONG_LASTCL	最后一个批次平均通话时长
TALKCNT	通话次数	MAX_TALKLONG_LASTCL	最后一个批次最大通话时长
TALKLONG	通话时长	FIRST_INTERVAL	第一次距前一次通话间隔
MAX_TALKLONG	最大通话时长	LAST_INTERVAL	最后距前一次通话间隔
IDNOFLAG	是否提供身份信息	AGE	年龄
ISCB	是否承保		

3.2. 数据的预处理

1) 异常值处理

通过对特征 AGE，DAILCNT，TALKCNT，TALKCNT_LASTCL 可视化发现，AGE 数据存在 181,1821,1999 这样的数据，这是不符合常理的；而 DAILCNT，TALKCNT，TALKCNT_LASTCL 三个特征存在较为夸张的长尾分布，分析发现数据存在异常，我们对这样的数据进行删除。

2) 缺失值处理

通过 python 中的 info 函数，我们可以发现数据存在大量缺失。对于缺失值达到 70%左右的数据进行模型填充或将缺失数据归为一类数据。缺失值达到 85%以上的数据进行了特征的删除。

3) 重复数据处理

通过重复数据查看发现数据中存在重复数据，对重复数据进行数据删除。

4) 无效变量剔除

对于数据中缺失值过多的特征 FIRST_INTERVAL 和 LAST_INTERVAL 进行了删除。通过查看各特征关于标签(y)的分布情况发现 BANNEDORNOT，TC_WRONGINFO，TC_ACCEPTED 三个特征分布比例严重失衡，也进行了删除。

5) 数据平衡处理

样本类别失衡会影响模型对少数类别样本的预测分类效果，而样本类别不均衡问题在实际应用中普遍存在[12]，常用处理数据不平衡的方法有欠抽样、过抽样和合成少数类过取样方法[13]。本实验中通过对数据进行可视化发现，数据中正负样本比例为 1:3，数据失衡，实验对数据进行了欠抽样处理。如图 5 是对数据样本的统计图。

3.3. 特征工程

1) 特征组合

可视化发现，TC_SALEPROCESS_PRIOR 和 TC_SALEPROCESS_LAST 两个特征分布相似，将这两个特征的数据进行合并得到特征 TC_SALEPROCESS。

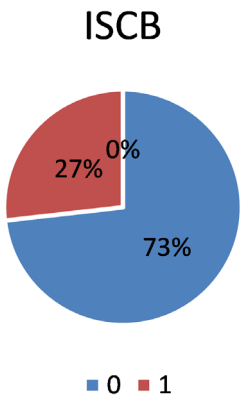

Figure 5. Positive and negative sample distribution
图 5. 正负样本分布图

2) 分箱处理

分箱处理指的是将多个连续数据划分到离散区间的过程。TC_SALEPROCESS_PRIOR 和 TC_SALEPROCESS_LAST 两个特征进行组合后，数据区间变大，使用分箱将数据分为 4 类。AGE 数据为连续数据，对该数据进行分箱处理将数据分为 4 个阶段，分别代表少年，中年，老年，和其它四个阶段。

3) 数据标准化

通过可视化发现，特征 TALKCNT,TALKLONG，TALKCNT_LASTCL，MAX_TALKLONG，MAX_TALKLONG_LASTCL，AVG_TALKLONG_LASTCL，DAILCNT 的数据变化范围较大，数据比较稀疏，故使用 Max Abs Scaler ()进行数据标准化。以便于后续相似距离的计算。

4) 离散化

离散化指的是把无线空间中的有限的个体映射到有限空间中。本论文中主要包括多值离散数据的离散化和连续数据的离散化。多值离散数据的离散化主要指的是将多分类的数据进行合并变成少分类数据的过程。连续数据的离散化是主要的离散化应用，主要指的是将离散的数据划分为特定空间的集合。

5) 相关性分析

相关性分析指的是对多个特征变量进行相关程度或密切程度的测量。

a) 指标与标签之间的相关性分析

通过分析发现 CALLLISTCNT_BSJ, CALLLISTCNT_YZ, TC_DAILSTATE_LAST, SEX，TALKCNT_LASTCL，AGE 几个特征与 ISCB 相关性很低，对这些特征进行剔除。

b) 指标间的相关性分析

若指标间的相关性达到 0.8，我们就认为两个变量之间存在较强的相关性。

通过分析可知，DAILCNT 与 TALKCNT 之间的相关性为 0.88，TALKCNT，TALKLONG 之间的相关性位 0.8。在最大限度保留信息的前提下，我们对 TALKCNT 数据进行了删除。

4. 数据建模

样本数据集特征维度高，往往会占用计算资源，增加模型计算时间，无形中降低预测模型的效率，同时也会一定程度上影响模型预测效果[12]。特征选择是一种有效的高维特征数据集降维的措施，主要方法有卡方检验、相关系数法、主成分分析法、粗糙集、模型选择等。本文主要对比了未进行特征选择，用决策树进行特征选择，用随机森林模型进行特征选择三种情况下，模型的 AUC 值和准确率，召回率，F1 值的不同。

4.1. 逻辑回归模型

逻辑回归模型是一种经典的数据模型，由于其可读性使得该模型在各个领域被广泛使用，我们将未进行特征选择的所有特征带入逻辑回归模型中进行训练和测试，得到准确率和评估报告表，如表 2：

Table 2. Evaluation report form
表 2. 评估报告表

	Precision	Recall	F1-Score	Support
0	0.89	1.00	0.94	21996
1	0.98	0.64	0.77	7186
Avg/total	0.91	0.91	0.90	29182

Accuracy is: 0.907340.

通过分析评估报告发现，进行预处理之后的数据带入逻辑回归模型中，有较好的准确率和 f1 分数。由此可见，数据清洗对于建模有着非常大的影响。

4.2. 基于决策树特征选择的逻辑回归

4.2.1. 决策树特征选择

由于指标众多，且指标对客户承保的影响不同，有的指标对客户承保的影响较大，有的影响较小，所以在这里我们使用决策树模型查看各个指标对模型的影响程度，并挑选出对模型影响较大的指标，运用这些指标去进行预测。如下为使用决策树模型得出的特征重要性的降序排列，如表 3：

Table 3. Ranking chart of feature importance
表 3. 特征重要性排序图

1. feature 7 (0.743086) CALLLISTCNT_JB
2. feature 18 (0.091831) TC_POSSIBILITY
3. feature 27 (0.068687) TC_SALEPROCESS
4. feature 22 (0.046774) TALKCNT_LASTCL
5. feature 26 (0.016253) IDNOFLAG
6. feature 13 (0.014980) TALKLONG
7. feature 24 (0.003628) MAX_TALKLONG_LASTCL
8. feature 19 (0.002003) TC_ENDING
9. feature 23 (0.001487) AVG_TALKLONG_LASTCL
10. feature 21 (0.001312) DAILCNT_LASTCL

通过分析发现，在上述指标中，客户成交的可能性，是否提供身份证信息，通话次数，通话时长，以及投放给销售系统的批次数对客户是否承保的影响较大。

4.2.2. 决策树与 Logistic 模型融合

通过决策树模型进行特征筛选之后，我们按照正负样本比例为 7:3 的原则将数据集划分为训练集和测试集，训练集用于对模型进行训练，测试集用于对模型进行测试和评估，然后通过评估报告查看模型的效果。如下即为进行特征筛选后带入逻辑回归模型中，得到的准确率和评估报告，如表 4：

Table 4. Evaluation report form
表 4. 评估报告表

	Precision	Recall	F1-Score	Support
0	0.91	0.98	0.95	21996
1	0.93	0.71	0.81	7186
Avg/total	0.92	0.92	0.91	29182

Accuracy is: 0.915393.

通过报告发现，该模型的准确率和 f1 分数达到 0.91，效果不错。但将使用决策树模型进行筛选后的特征带入逻辑回归模型与未进行特征筛选的所有特征带入逻辑回归模型相比，并没有什么显著提升。分析认为，决策树模型作为弱分类器，本身容易产生过拟合现象，致使模型的泛化效果不好。

4.3. 基于 RF 特征选择的逻辑回归

4.3.1. RF 特征选择

随机森林模型是在 bagging 框架的基础上创建的，该模型的各个弱学习器之间不存在相互的依赖关系，与 boosting 模型相比，大大减少了调参的难度。通过随机森林模型进行特征选择，最大限度的保留了数据的业务含义，同时降低了模型的计算量和运行时间，降低了噪音变量信息对模型效果的影响。如下为特征重要性的降序排列，如表 5 所示：

Table 5. Ranking chart of feature importance
表 5. 特征重要性排序

1. feature 7 (0.205882) LAST_CALLLIST_DATATYPE
2. feature 9 (0.175655) TALKLONG
3. feature 10 (0.140973) MAX_TALKLONG
4. feature 8 (0.120436) DAILCNT
5. feature 16 (0.079408) DAILCNT_LASTCL
6. feature 20 (0.074790) TC_SALEPROCESS
7. feature 17 (0.059765) AVG_TALKLONG_LASTCL
8. feature 18 (0.032545) MAX_TALKLONG_LASTCL
9. feature 19 (0.026736) IDNOFLAG
10. feature 13 (0.025911) TC_POSSIBILITY
11. feature 2 (0.020011) CALLLISTCNT
12. feature 14 (0.010909) TC_ENDING
13. feature 12 (0.010357) TC_DAILSTATE_PRIOR
14. feature 3 (0.009299) CALLLISTCNT_XH
15. feature 4 (0.001710) CALLLISTCNT_QT
16. feature 6 (0.001607) FIRST_CALLLIST_DATATYPE
17. feature 1 (0.001311) MARRIAGE
18. feature 15 (0.001014) FREESENDCNT
19. feature 5 (0.001001) CALLLISTCNT_NOTAKL
20. feature 11 (0.000678) TC_PHONESTATE

通过上述有序序列可知，前 13 个特征的重要性大于 0.1，我们选择前 13 个特征带入逻辑回归模型中进行预测。同时，通过对这些特征分析发现，客户的通话次数，通话时长，以及客户是否提供身份证号对客户是否承保的影响较大。同时，随机森林模型是对决策树模型的

4.3.2. RF 与 Logistic 模型融合

使用随机森林进行特征选择后，我们使用划分好的训练集和测试集进行训练和测试，如下即为使用随机森林模型进行特征筛选后的特征带入逻辑回归模型得到的评估报告，如表 6 所示：

Table 6. Evaluation report form
表 6. 评估报告

	Precision	Recall	F1-Score	Support
0	0.96	0.98	0.97	21996
1	0.93	0.87	0.90	7186
Avg/total	0.95	0.95	0.95	29182

Accuracy is: 95.196. [RFC full test]

通过观察发现，基于随机森林特征选择的逻辑回归模型相较于单一的逻辑回归模型和基于决策树特征选择的逻辑回归模型来说，有较大的提升，分析认为，随机森林模型作为集成模型本身是比较优秀的模型，有较好的抗噪能力，同时能解决过拟合问题。

5. 总结与展望

随着大数据技术的不断发展，保险行业也逐步开始从传统的金融向互联网金融转型。一直以来，保险行业对于客户的分析一直是基于传统的统计方法或者经验的分析。近年来虽然也有一些基于机器学习方法的研究，但也仅仅限于基础的模型的套用，缺乏细致的分析。

本文中主要介绍了两种优化方法，基于决策树的逻辑回归模型、基于随机森林的逻辑回归模型。在建模之前，我们对所给数据进行了精确的数据清洗，大幅度降低了噪声数据的影响。在建模过程中，我们使用在其他领域比较广泛使用的组合模型进行分析，来提高数据的效果，各个模型的参数都是经过特征组合或特征选择的，以期达到最优的模型效果。最后，我们使用 Accurcy 和评估报告两种评估指标进行模型评估，防止单一评估指标带来的不可信性。本文虽然在一般机器学习的方法上进行了改进，但仍然存在一些值得改进和继续研究的地方。

1) 模型优化研究。本文研究分析所使用的模型依旧是传统的逻辑回归模型，没有创新的提出更新颖的模型，这是后续需要继续关注和研究的方向

2) 本实验中随机森林模型与逻辑回归模型进行组合的效果最高，可能因为随机森林模型是一种集成算法，之后的研究中可以考虑将 GBDT，xgboost 等流行算法应用到保险数据分析中去。

致　谢

光阴荏苒，日月如梭。一年时光在弹指间匆匆划过。如今还记得 2018 年四月份左右，内心满怀激动和澎湃之情，来到心目中这所学校。看着校园里匆匆跑过的同学，和结伴谈笑而过的各国同学。心里的那种满足感，难以描述。一年来，无论在学习和生活上，我都深切的感受到了学校，导师，辅导员和学长的真心关怀。在此，我想对这些人真心的说一声谢谢。

首先，感谢我的导师黄浩老师。是您一直以来的督促和关照，让我学到了更多数据分析的知识，让我认识了一个新学科，走进了一个新世界。

其次，感谢我的导员张延红老师和我的学长刘宗亮。一年来事务繁杂，是你们督促我参与各项活动，不厌其烦的督促我完成各项工作。不断告诫我学习和生活中的各项问题，真心感谢老师和学长的悉心教导。

最后，感谢我的父母家人，研究生生涯生活环境的快速转变，让我感到诸多不顺，曾默默哭泣过，也曾自暴自弃过，是父母一直陪在我身边，默默鼓励着我，生活艰辛，每个人都是这茫茫世间的赶路人，幸亏我还有家人在一直默默陪伴着，我很感恩。未来，我会以梦为马，勇往向前。

基金项目

国家重点研发计划资助(National Key R&D Program of China)，项目编号：2017YFB1400700。

参考文献

[1] 苗东. 大都会保险公司客户关系管理研究[D]: [硕士学位论文]. 上海: 华东理工大学, 2013.

[2] 卞爱军. 基于信息化平台的寿险客户细分管理研究——以扬州寿险公司为例[D]: [硕士学位论文]. 南京: 南京理工大学, 2008.

[3] 柯新喜. 基于决策树模型的社会保险客户分类研究[J]. 福建电脑, 2016, 32(6): 105-107.

[4] 王贵龙. 基于关联向量机的保险客户识别研究[D]: [硕士学位论文]. 西安: 西安工业大学, 2011.

[5] 赵萍. 数据挖掘在寿险客户关系管理中的应用[D]: [硕士学位论文]. 天津: 天津大学, 2007.

[6] 董娜, 常建芳, 吴爱国. 基于贝叶斯模型组合的随机森林预测方法[J]. 湖南大学学报(自然科学版), 2019, 46(2): 123-130.

[7] 苏杭西子. 基于随机森林模型的个人信用风险评估研究[D]: [硕士学位论文]. 长沙: 湖南大学, 2018.

[8] 李航. 统计学习[M]. 北京: 清华大学出版社, 2012:77-79.

[9] 邴欣. 机器学习在推荐系统中的应用[D]: [硕士学位论文]. 济南: 山东大学, 2016.

[10] 钱超. 基于特征优化的逻辑回归模型在广告点击率问题中的应用研究[D]: [硕士学位论文]. 武汉: 华中师范大学, 2018.

[11] 宋天龙. Python 数据分析与数据化运营[M]. 北京: 机械工业出版社, 2017: 99-102.

[12] 刘晨晨. 基于数据挖掘的通信客户流失预警模型研究[D]: [硕士学位论文]. 武汉: 华中师范大学, 2017.

[13] 王文敬. 基于 SMOTE 过抽样法的个人信用评分模型研究[D]: [硕士学位论文]. 上海: 上海师范大学, 2019.