|  |  |
| --- | --- |
| **题 目：** | 使用Spark机器学习库的GBT算法 |
|  | 预测顾客对买保险的态度 |
| **作 者：** | 梁震仁 |
| **学 号：** | 2019053301 |
| **课程名称：** | 基础QQ用户标签推荐（数据挖掘） |
| **专业年级：** | 2019级 |
| **指导老师：** | 杨观赞 |
| **完成时间：** | 2021年5月6日 |

推荐算法设计与实现

[1定义问题 2](#_Toc13114678_WPSOffice_Level1)

[2理解数据 2](#_Toc1375061152_WPSOffice_Level1)

[3数据准备 2](#_Toc1581256297_WPSOffice_Level1)

[4建立模型 2](#_Toc1064452054_WPSOffice_Level1)

[5评估算法 2](#_Toc1706892068_WPSOffice_Level1)

[6优化模型 3](#_Toc1648430250_WPSOffice_Level1)

# 算法设计与实现

需提交的文件有：本文档、项目代码、数据集、运行代码录屏。

打包zip发到zan0072@163.com，最晚5月6日提交完毕。

# **1定义问题**

了解数据和业务问题，必须要对目标有一个清晰明确的定义，即决定到底想干什么。例如，本课程是采用QQ用户资料数据集、腾讯产品使用记录（物品）数据集等大量数据，向特定用户推荐感兴趣的产品以促进购物成单。

就是描述清楚数据源来自那里，数据字段有哪些，利用这些数据达到什么目标

常见的数据源有以下几个：

* UCL机器学习知识库：包括近300个不同大小和类型的数据集，可用于分类、回归、聚类和推荐系统任务。数据集列表位于http://archive.ics.uci.edu/ml/。
* Amazon AWS公开数据集：包含的通常是大型数据集，可通过Amazon S3访问。这些数据集包括人类基因组项目、Common Crawl网页语料库、维基百科数据和Google Books Ngrams。这些数据集的相关信息可参见https://registry.opendata.aws/。
* Kaggle：这里集合了Kaggle举行的各种机器学习竞赛所用的数据集。它们覆盖分类、回归、排名、推荐系统以及图像分析领域，可从Competitions区域下载：<https://www.kaggle.com/competitions。>
* KDnuggets：这里包含一个详细的公开数据集列表，其中一些是上面提到过的。该列表位于<https://www.kdnuggets.com/datasets/index.html>。

本实验的数据源来自Kaggle网站，

网址为<https://www.kaggle.com/anmolkumar/health-insurance-cross-sell-prediction>

该数据集的主要任务是从训练集中选取有效的数据段，并使用训练集建立预测模型，预测测试集里的顾客是否会购买保险。

# **2理解数据**

理解数据的目的是找到对预测输出影响最大的数据字段，和决定是否需要定义导出字段。

理解数据之后的，可以尝试以下方法对数据进行处理：

* 通过特征选择来减少大部分相关性高的特征
* 通过标准化数据来降低不同数据度量单位带来的影响
* 通过正态化数据来降低不同数据分布结构，以提高算法的精度

首先要描述清楚数据记录数、最大值、最小值、平均值、方差、两两特征之间的皮尔逊相关系数

维数过多可进行PCA降维

然后可视化：密度图、直方图

描述数据的分布情况，推测适合使用的算法

该数据集分为训练集和测试集，训练集中一共381109条记录，测试集中一共508146条记录

训练集有：

Id：顾客id

Gender(2)：顾客性别

Age(64)：顾客的年龄（最小20，最大84）

Driving\_License(2)：顾客是否拥有驾照，0表示没有，1表示拥有

Region\_Code(53)：顾客所在地区的id（最小0，最大52）

Previously\_Insured(2),：顾客是否以及购买过保险，0表示顾客并未购买任何保险，1表示

顾客已经购买过保险

Vehicle\_Age(3)：顾客汽车的年龄

Vehicle\_Damage(2)：顾客汽车是否受到过损坏，0表示从未受到过损坏，1表示曾经受到过损坏

Annual\_Premium：顾客需要在当年支付的年度保险金额

Policy\_Sales\_Channel(69)：不同的id分别对应不同的与顾客的联系方式

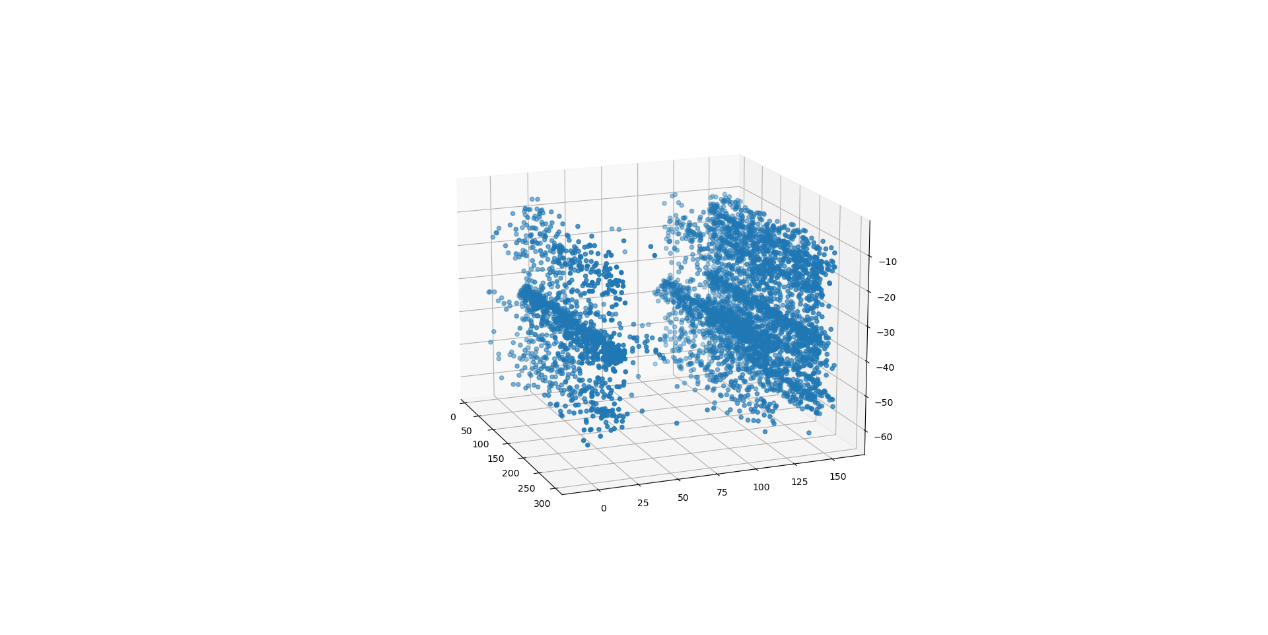
Vintage：顾客与公司建立联系的天数（连续值）

Response：顾客是否愿意购买保险，0表示不愿意，1表示愿意

测试集与训练集先比缺少了需要预测Response列

该数据集中的Response分布及其不均衡，大概在0：1大约为20：3

从该数据经过PCA降维后的图形可以看出，该数据的记录与记录之间的分布都较为分散，并不适合用PCA降维



查看数据的皮尔逊相关系数可以发现，原因是记录与记录之间的线性相关性都不强，因此并不适合使用PCA降维。

# **3数据准备**

这是建立模型之前的最后一步数据准备工作。可以把此步骤分为四个部分：选择变量，选择记录，创建新变量，转换变量。即：数据预处理和数据特征选定。

描述下如何进行数据预处理和数据特征选定

数据预处理方法：格式标准化，异常数据清除，错误纠正，重复数据的清除

特征选定，可使用GTB算法（GBTClassifier）、决策树（DecisionTreeClassifier）、随机森林（RandomForestClassifier），利用其输出特征的重要性来选取重要的特征

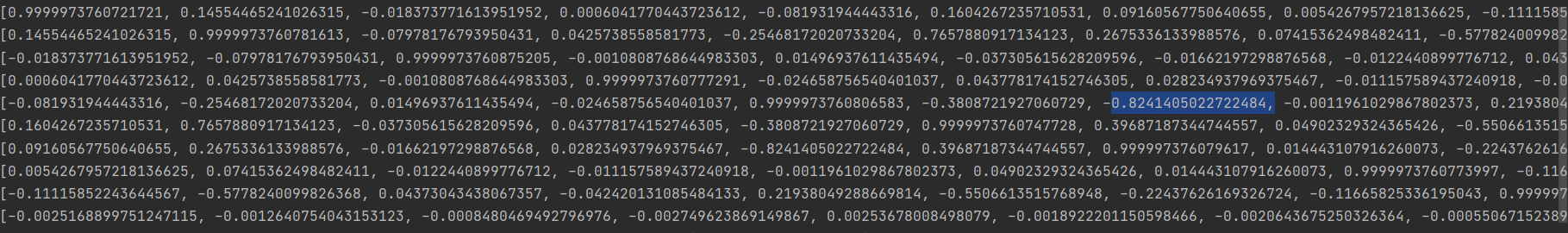
该数据集的数据已经经过了初步处理剔除或补充了异常值，因此数据中的记录均为有效数据，不需要进行异常处理。

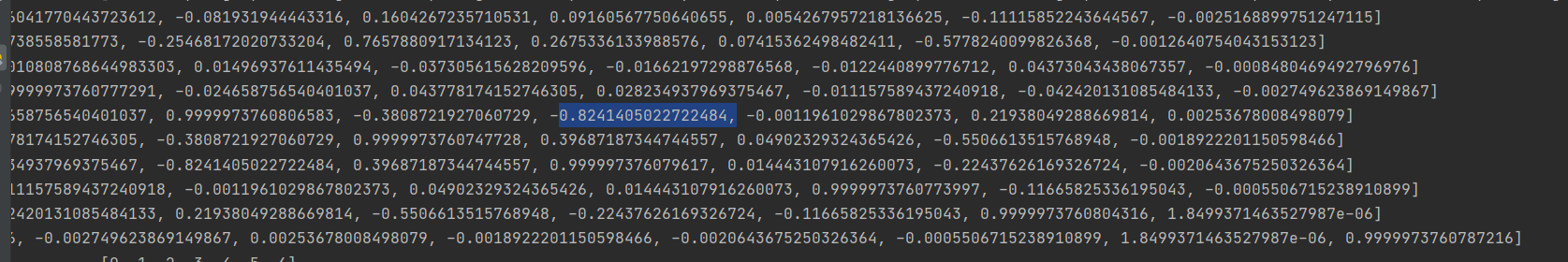
查看数据的记录可以发现，其中Annual\_Premium的记录比其他记录将近大了100到1000倍，因此先对该记录进行标准化处理，将其缩小到1到200之间，减小对其他数据造成的干扰。

其中年龄段分布较为均衡，可以考虑将分段

通过皮尔逊相关系数矩阵可发现Previously\_Insured和Vehicle\_Damage之间有较强的负线性相关性，因此选择将其进行合并。







若特征较多时，经过卡方选择器后，经过初步筛选的数据的记录进一步减少

到这里数据已经筛选完毕，接下来将数据转换为机器学习算法所需要的格式

例如：（features=SparesVector（7,{0: 1, 1: 2,2: 3,3: 4,4: 5,5: 6,7: 8}）,1）

此步骤较为繁琐，需要将rdd转换未类似，Labeledpoint的格式（不可以直接转换未Labeledpoint，因为Labeledpoint不能与ml库中的Vectors兼容，只兼容milb库中的Vectors，而ml库中机器学习算法不兼容milb库中的Vectors，若要用超参数验证只能使用ml库）将其转换未dataframe后，使用StringIndexer对label进行进一步的标准化，用VectorIndexer

对features进行进一步标准化

到此数据准备完成

# **4建立模型**

数据划分：训练集、测试集、验证集

模型训练：使用训练集训练

模型测试和验证：使用测试集测试和验证准确性

描述数据怎么划分，训练集与测试集占比

描述下为什么选用该算法？可以从算法应用场景和数据分布上说明

记录训练时长、测试和验证时长、

将数据划分为训练集和测试集，训练集和测试集占比分别为8：2

根据该数据集的记录类型主要为类别，但其中也存在连续型的数据，因此选择使用GBT算法，该算法主要用于分类，并且只能二分类，而该数据集的结果也只有感兴趣和不感兴趣两种，在特征值和标签上都符合GBT算法的要求

该算法主要需要调整的参数有maxBins（最大宽度，对应分类值的种类），maxDepth（最大深度），stepSize（学习率）

在记录为5000条的情况下，





# **5评估算法**

主要是为了寻找最佳的算法子集，步骤如下：

分离出评估数据集，以便验证模型。

定义模型评估标准，用来评估算法模型。

抽样审查线性算法和非线性算法。

比较算法的准确度

描述下使用了哪些评估器和评估方法以及为什么使用这些评估器和评估方法。例如，是因为数据分布不均衡？为了评价数据的变化程度？数据分布符合0-1分布？

描述下本次评估的效果、出现该效果的原因以及改进方法（调整模型的数据集还是使用别的模型等等）

评估的方法有：准确度、ROC、AUC、PR曲线、精确度、召回率、F1分数、均方误差、平均绝对误差等

评估算法使用了GBT算法对应的MulticlassClassificationEvaluator评估器，并且评估模型的f1数值

本次评估的效果为

进行了多次测试后效果维持在0.8以上，证明模型较为优秀

# 6优化模型

当得到一个准确度足够的算法的时候，要从中找出最合适的算法，但要注意的是优化模型应该遵循偏差和方差协调的原则，提升准确度的同时防止过拟合，通常有两种方法可以提高算法的准确度：

* 对每一种算法进行调参，得到最佳结果。
* 使用集成算法来提高算法模型的准确度。

描述下选择了哪些方法进行调参以及最终优化的结果是什么。

一般的优化方法有：K折交叉验证（CrossValidator）和随机搜索优化参数（TrainValidationSplit）

本次实验使用了随机搜索优化参数进行对模型的优化，需要调整的参数有maxBins（最大宽度，对应分类值的种类），maxDepth（最大深度），stepSize（学习率）

优化后的准确率为

并无明显提升