|  |  |
| --- | --- |
| **题 目：** | 使用Spark机器学习库的随机森林算法 |
|  | 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8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

# **1定义问题**

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**

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

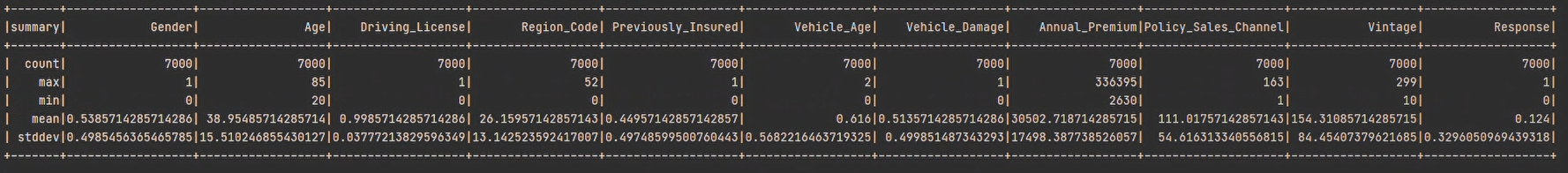
Vintage：顾客与公司建立联系的天数（连续值）

Response：顾客是否愿意购买保险，0表示不愿意，1表示愿意

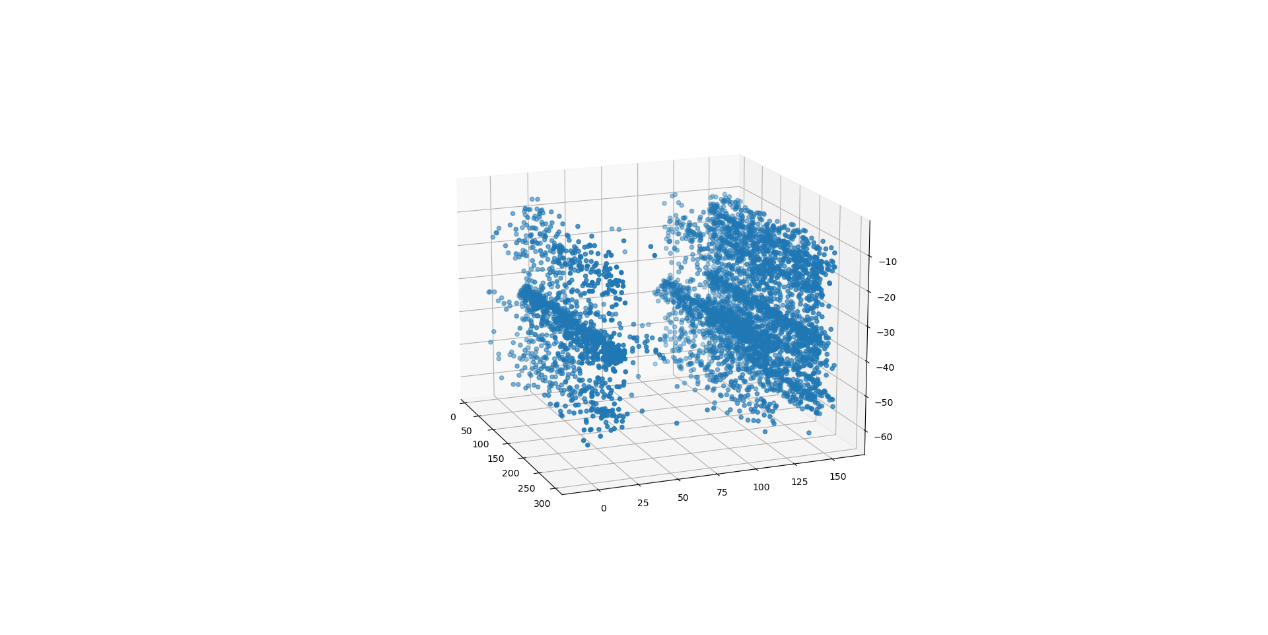
测试集与训练集先比缺少了需要预测Response列

该数据集中的Response分布及其不均衡，大概在0：1大约为20：3

数据的记录数、最大值、最小值、平均值和标准差为



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**

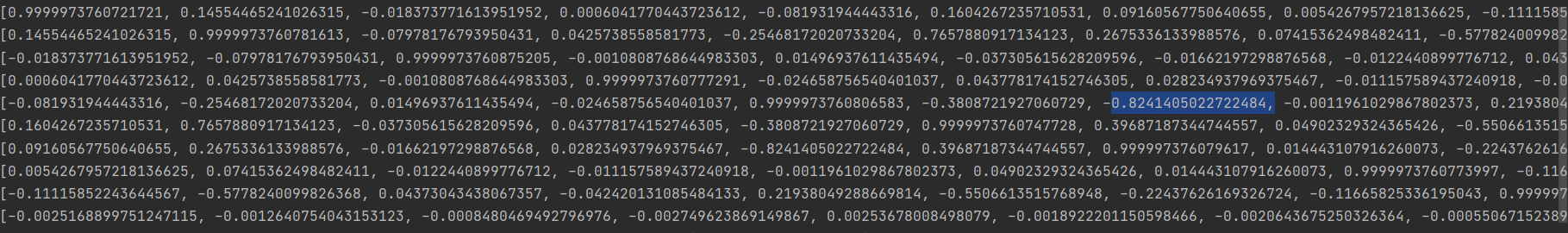
该数据集的数据已经经过了初步处理剔除或补充了异常值，因此数据中的记录均为有效数据，不需要进行异常处理。

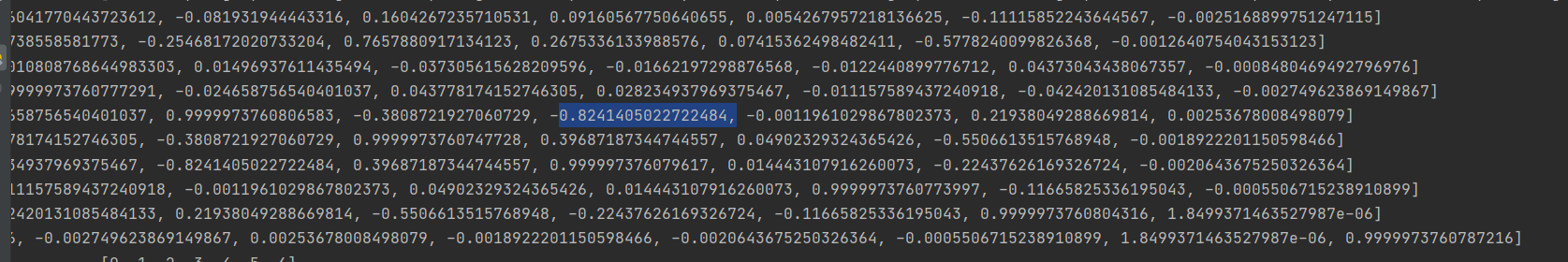
查看数据的记录可以发现，其中Annual\_Premium的记录比其他记录将近大了100到1000倍，因此先对该记录进行标准化处理，将其缩小到1到200之间，减小对其他数据造成的干扰。

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

此步骤较为繁琐，需要将rdd转换为类似，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）

将转换完成的rdd转换为dataframe结构后，为了解决数据不均衡的问题，为每个正样本和每个负样本分配合适的权重（各自比例的倒数），并且再对正样本使用过采样进行微调，最后分为训练集和测试集。

到此数据准备完成

# **4建立模型**

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7：3

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，因此选择使用随机森林算法，该算法主要用于分类，并且支持二分类和多分类，且算法具有较高的准确度，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随机森林算法的要求

（再使用了逻辑回归算法、GBT算法后发现随机森林准确度最高）

该算法主要需要调整的参数有maxBins（最大宽度）

（其他参数均对ROC指标影响不大）

在记录为7000条的情况下，



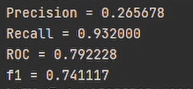


# **5评估算法**

评估算法使用了二分类评估器BinaryClassificationEvaluator评估ROC指标，同时为了评估f1分数还使用了MulticlassClassificationEvaluator，为了能更直观的观察预测结果的好坏，还使用了精确度和召回率进行预测结果的评估。

此处评估不宜使用准确度进行评估，原因时该数据集样本分布及其不均衡，正样本十分稀少而负样本过量，使用准确度进行评估，即使是预测全为负的结果时依旧有接近90%的准确率，因此用准确度来评估模型的好坏是没有意义的。

为了能比较可靠的评估模型，选择了ROC指标和f1分数，同时因为实验内容是要确保顾客，因此需要优先确保召回率，为了更好的观察模型对正样本的预测结果，同时使用了精确度和召回率对模型的预测结果进行观察。



模型在牺牲了精确度的情况下确保了高召回率，经多次测试后ROC指标维持在0.75-0.8之间，f1分数维持在0.7-0.75之间

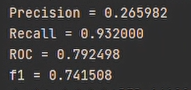
# 6优化模型

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）

参数验证大概为初始最大宽度的+-5的范围之内

使用了BinaryClassificationEvaluator评估器评估ROC指标

优化后的评估结果为



并无明显提升