**基于30-59岁成年人长期护理的保险购买意愿预测**

1. **文本特征提取**

本次报告采用的原始数据来自“【编号版】30-59岁成年人长期护理调查问卷”，因为数据中含有较多文本数据，需要抽取文本中的特征，然后将文本型数据全部转化为数值型数据，使用计算机程序的形式建立一个预测模型。以下是文本特征提取主要步骤。

1. 明确需要提取的主要特征

调查问卷第八问（对于您将来老年时的生活，您对哪些问题/风险感到最担心？）中的主要因素需要进行提取，将所有的回答进行汇总，然后分词，并统计其出现的频率，存储为文件。文件在本次打包的项目中的‘’8worry\_risk.txt“文件。结果为如下图1-1所示

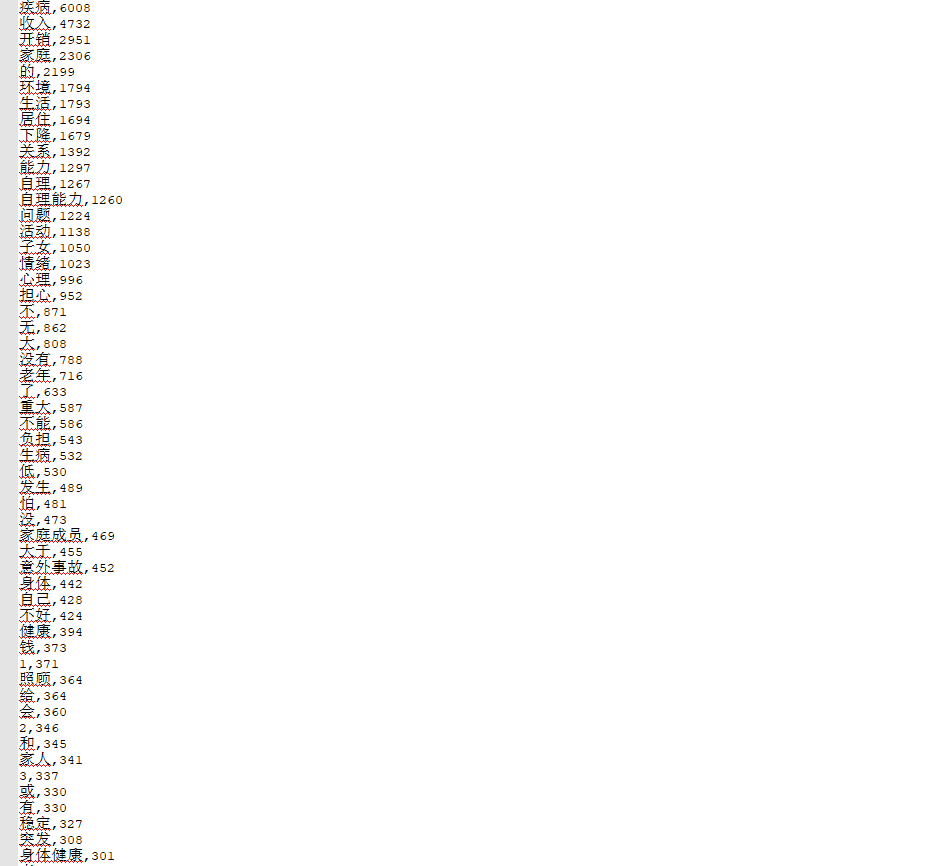


图 1-1 分词及词频

通过对有效次进行筛选发现词频靠前的主要词汇包括收入、开销、收入、开销、居住环境、家庭关系、心理情绪、疾病、自理或活动能力（暂且把‘子女’归为关系一类），如下图1-2所示。

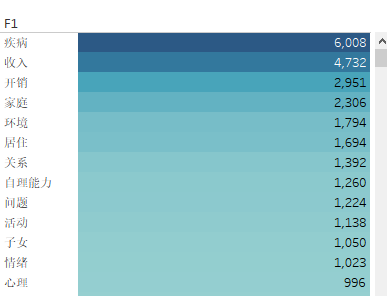


图 1-2 第八问中文本高频率词汇

因本次报告的主题是“基于人群特质的保险购买意向”，其目的便是预测原始数据中的第10问（如果您的朋友向您介绍一种商业医疗保险，以保障未来1年内您生病的风险，可以报销看病费用。您是否会购买这种保险产品？）、第13问（如果您的朋友向您介绍一种商业护理保险，可以保障您在老年时失去照顾自己生活起居能力的风险，一旦发生生活无法自理的情况，保险公司将给您经济补偿或提供必要的护理服务。这种保险是长期的，一般需要连续缴费几年，比如15年。您是否会考虑购买这种保险产品？）、第16问（如果您的朋友向您介绍一种商业重大疾病保险，可以保障您患上某种重大疾病，比如癌症、急性心肌梗塞、肾功能衰竭等的风险。如果发生重大疾病，保险公司就一次性支付一笔金额，比如说30万元。这种保险也是长期保险，需要连续缴费几年，比如15年。您是否会考虑购买这种保险产品？）。考虑到回答问题的先后性，将第11问（这种保险产品要满足哪些条件，或提供哪些服务您才会考虑购买？）、第12问（如果您根本不会考虑购买商业医疗保险，主要原因是什么？）、第14问（这种保险产品要满足哪些条件，或提供哪些服务您才会考虑购买？）、第15问（如果您根本不会考虑购买商业护理保险，主要原因是什么？）、第17问（这种保险产品要满足哪些条件，或提供哪些服务您才会考虑购买？）、第18问（如果您根本不会考虑购买商业重大疾病保险，主要原因是什么？）直接剔除可以作为做决策的参考项，不加入特征项。同时将调研来源、来源详情、时间以及调研人员的信息剔除，这里我们假设问卷的内容都是真实可靠的，同时将其他问题作为特征项。到目前为止我们的特征项包括（地点、年龄、性别、户口、目前的职业状况、身体健康状况、现阶段风险、期望风险（收入、开销、收入、开销、居住环境、家庭关系、心理情绪、疾病、自理或活动能力），预防所做的准备、目前总收入），预测目标为（医疗保险购买意愿、健康护理保险购买意愿、重大疾病保险购买意愿），因此在这里需要构建三个预测模型。目前，初步抽取的特征已经完成，并明确了预测目标。

* 1. **自然语言处理**

因为所得到的现阶段数据含有大量文本信息，因而要对文本信息进行自然语言处理，如下图1-3所示.

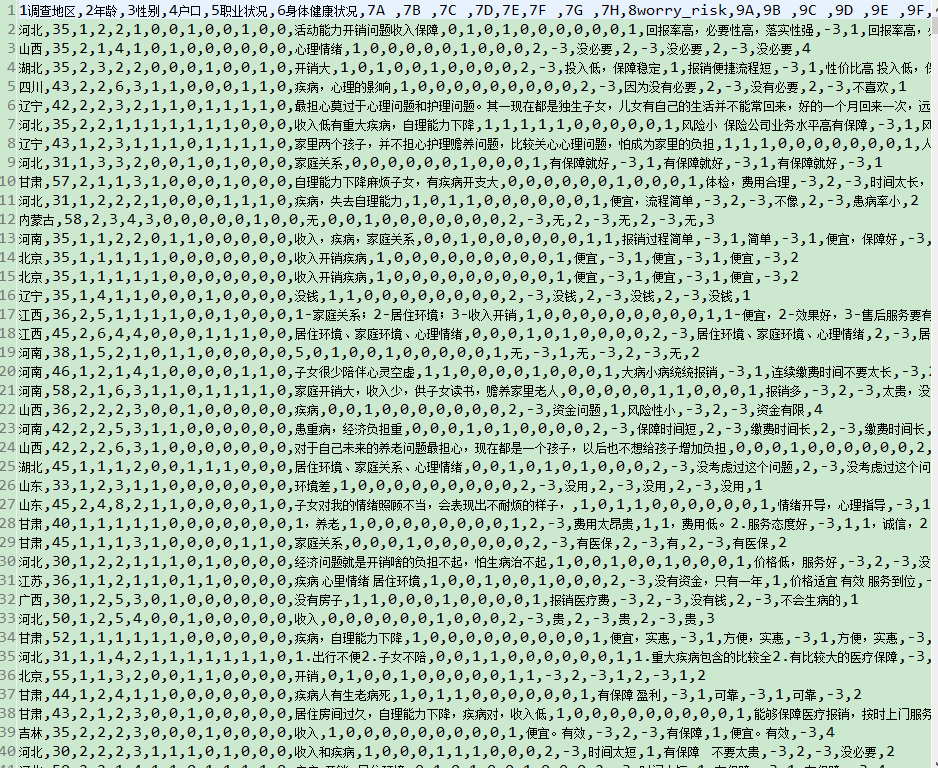


图1-3 主要特征数据

* + 1. **省份文本处理**

对于含文本的省份数据，如下图1-4所示



图1-4 原始省份数据

我们只考虑省份信息，处理后效果如下图1-5所示。

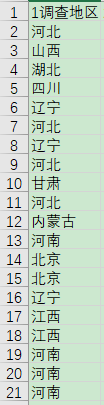


图1-5 省份信息处理结果

**1.1.2 期望风险文本处理**

对于第八列的每一列文本数据，首先我们构建属于期望风险（收入、开销、收入、开销、居住环境、家庭关系、心理情绪、疾病、自理或活动能力）每一项的特征词典，从每一列中检索词语，如果存在某一个词在某一个特征项的特征词典中，就令其特征为1，否则为0。其特征词典的主要内容在项目中的文件目录为，如下图1-6所示

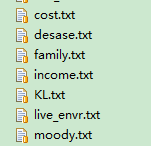


图1-6 特征词典

其分别对应着期望风险中的消费、疾病、家庭关系、收入、居住环境、心理情绪以及活动能力的特征词典。处理后的效果如下图1-7所示。

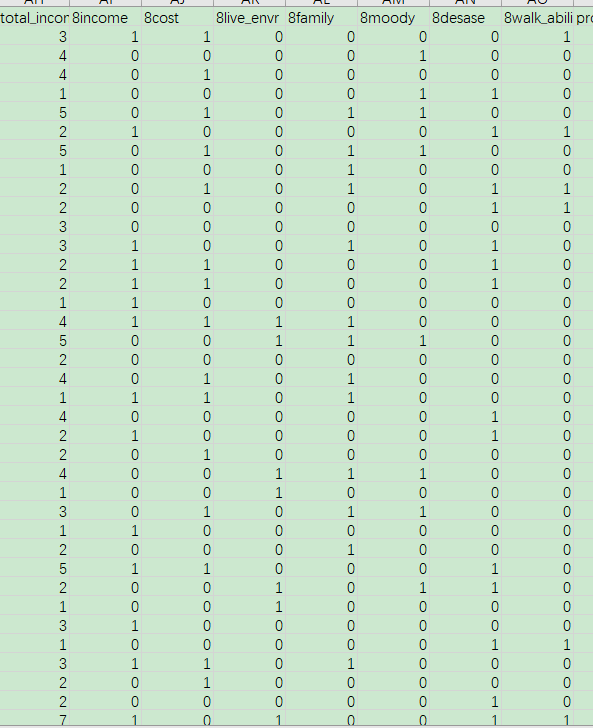


图1-7 期望风险中抽取出来的特征项处理结果

#### 1.1.2 期望风险文本处理

对于第八列的每一列文本数据，首先我们构建属于期望风险（收入，开销，收入、开销、居住环境、家庭关系、心理情绪、疾病、自理或活动能力）每一项的特征词典，从每一列中检索词语，如果存在某一个词在某一个特征项的特征词典中，就令其特征为1，否则为0。其特征词典的主要内容在项目中的文件目录为，如下图1-6所示

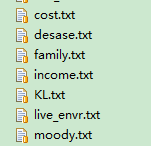


图1-6 特征词典

其分别对应着期望风险中的消费，疾病，家庭关系，收入，居住环境，心理情绪，以及活动能力的特征词典。处理后的效果如下图1-7所示。

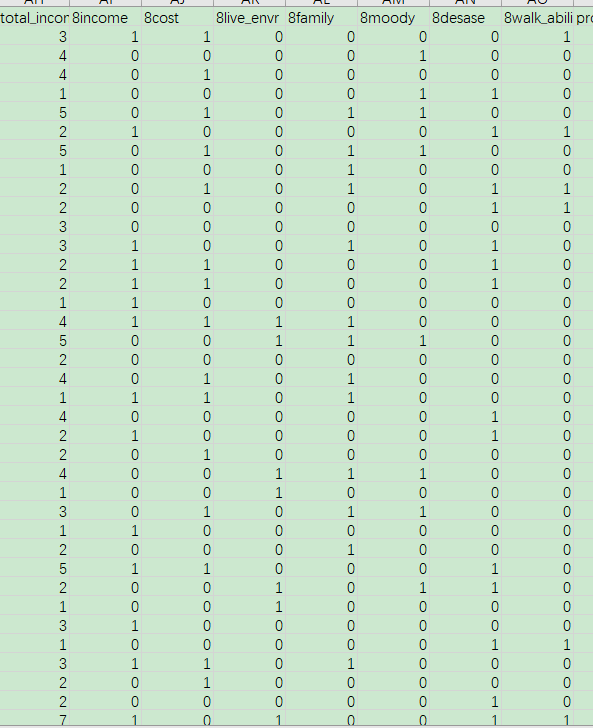


图1-7 期望风险中特区出来的特征项处理结果

#### 数据异常值处理及数值化

我们在对数据处理时发现少量异常数据，因此我们对这些数据进行了剔除。如下图1-8所示。

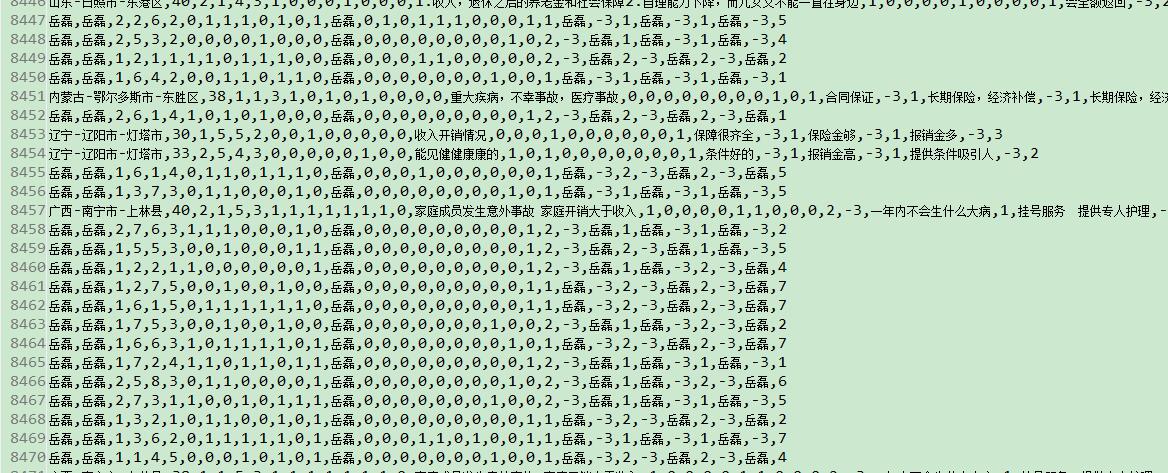


图1-8 异常数据

对所有数据都数值化，因为数据中的所有数据都是离散的，所以不需要归一化处理，对省份数据只需要进行离散化，这里没有采用独热编码；在程序中直接对每一个出现的省份赋一个唯一的数值。最后处理完的数据如下图1-8所示。

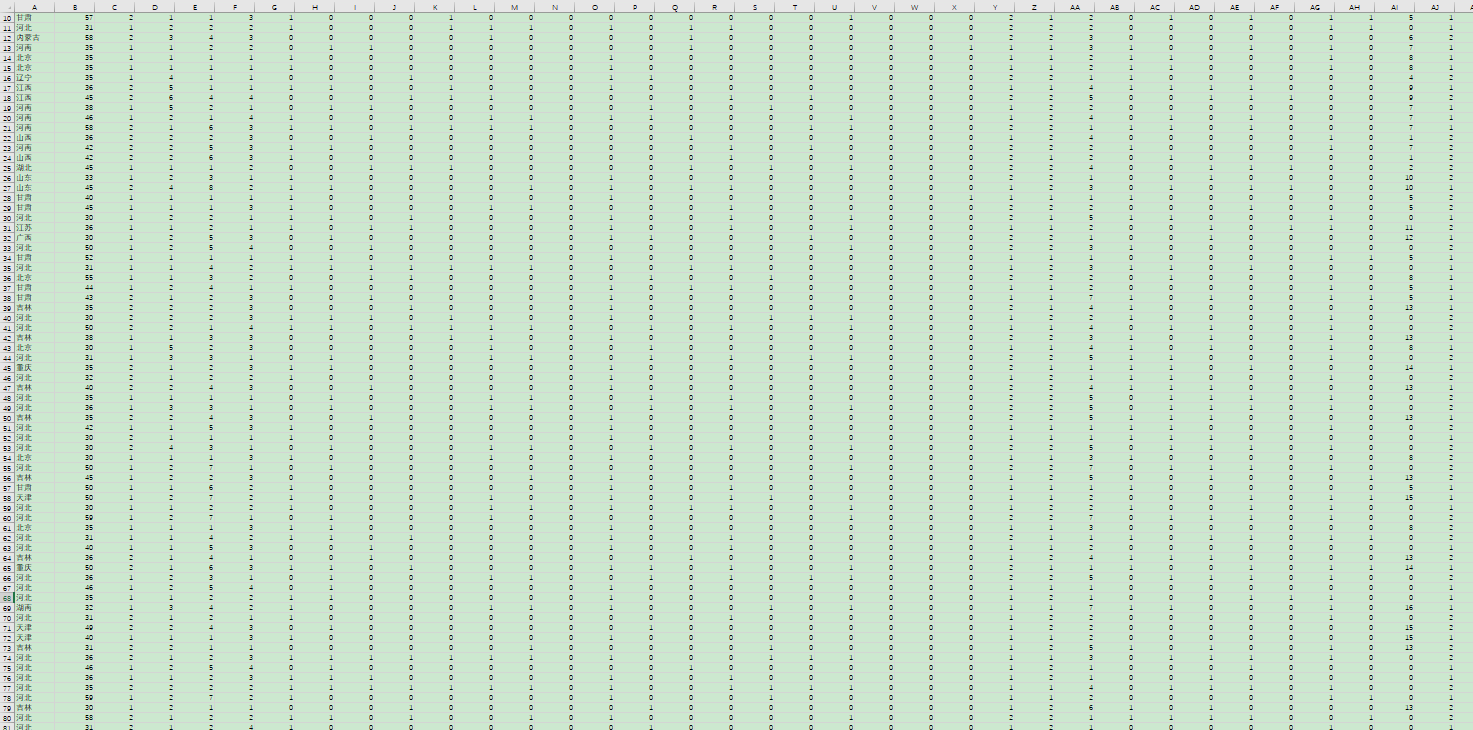
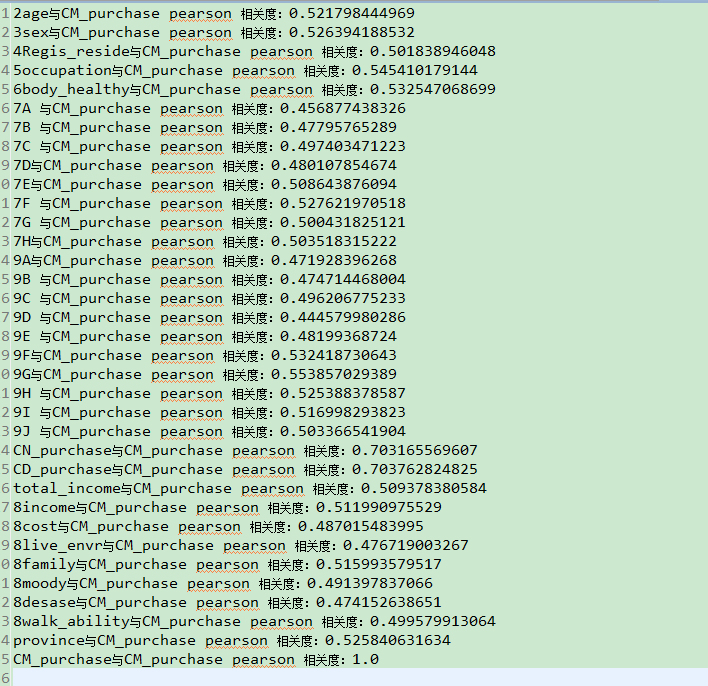


图1-8 最终数据

## 可视化

处理完后的数据，已经全部是数值，这里方便我们用软件进行可视化，这里用到的可视化软件为tableau。可视化的目的是分析其各个特征与预测目标的相关性。这里选用的相关性系数包括pearson相关系数，cos相关性系数以及KL散度。

三种系数都用于度量特征组数据与训练目标数据的相关性。三种系数的最后计算结果如下列图所示。



2-1 pearson相关系数



图2-2 KL散度



图 2-3 cos相关性

通过观察三种不同的相关性系数，我们发现一个预测目标与另外两个预测目标有极强的相关性。因此我们将另外两个预测目标也当做特征加入数据。通过相似度对其他两个预测目标进行相关性分析，我们发现预测目标之间的相关性极强。这里的意义是一个人愿意买一种那他就极有可能购买另外两种保险。这里我们就有了新的特征（另外两种预测目标作为特征）

### 现实因素可视化

在对其特征与预测目标的相关性分析中，我们用到了可视化技术。详细分析了每一个特征与其目标的相关性。

各特征与预测目标的可视化结果分别如下图所示。注：其中1代表考虑购买，2代表不考虑购买。

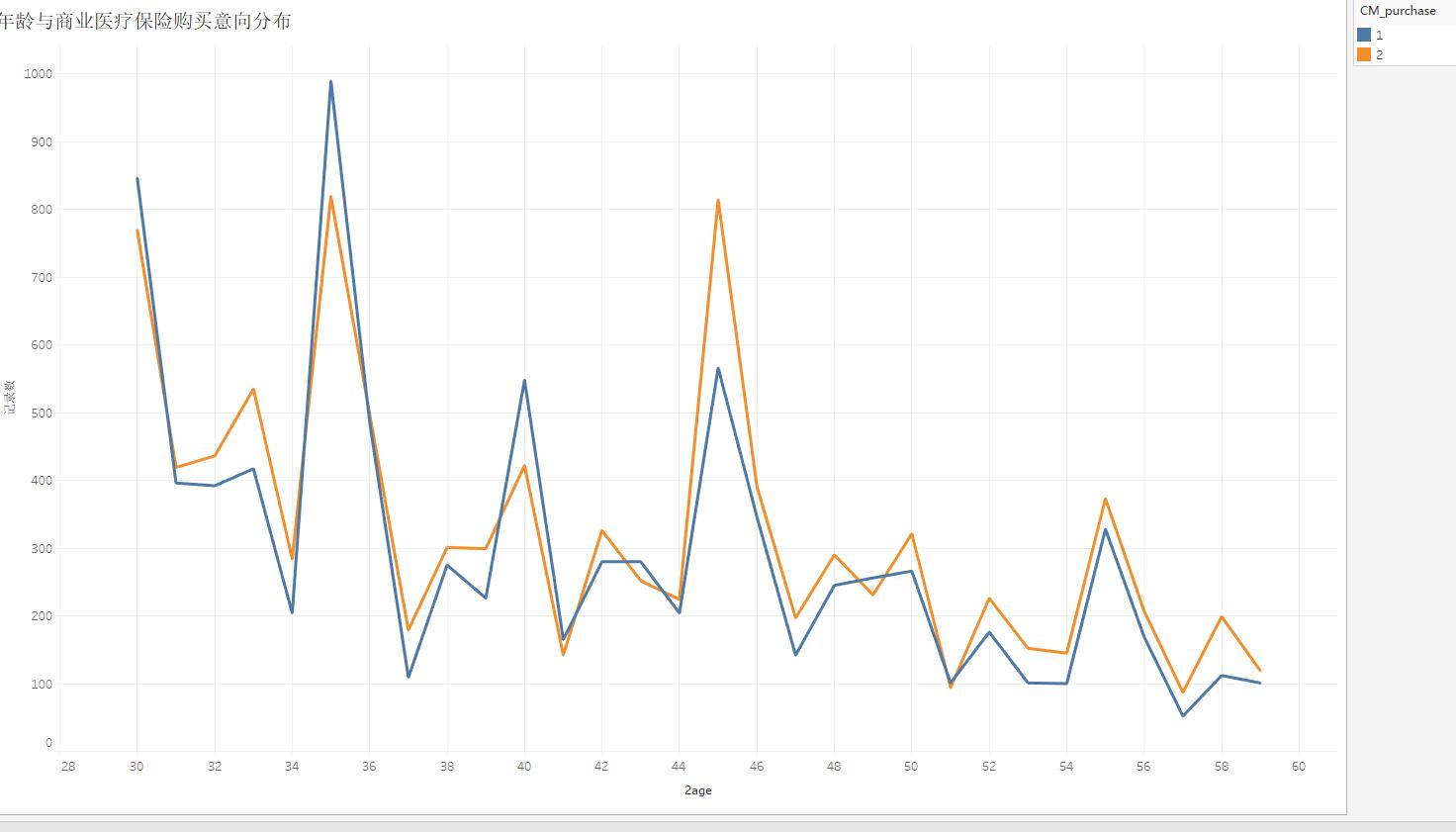


图2-4 年龄与预测目标的分布

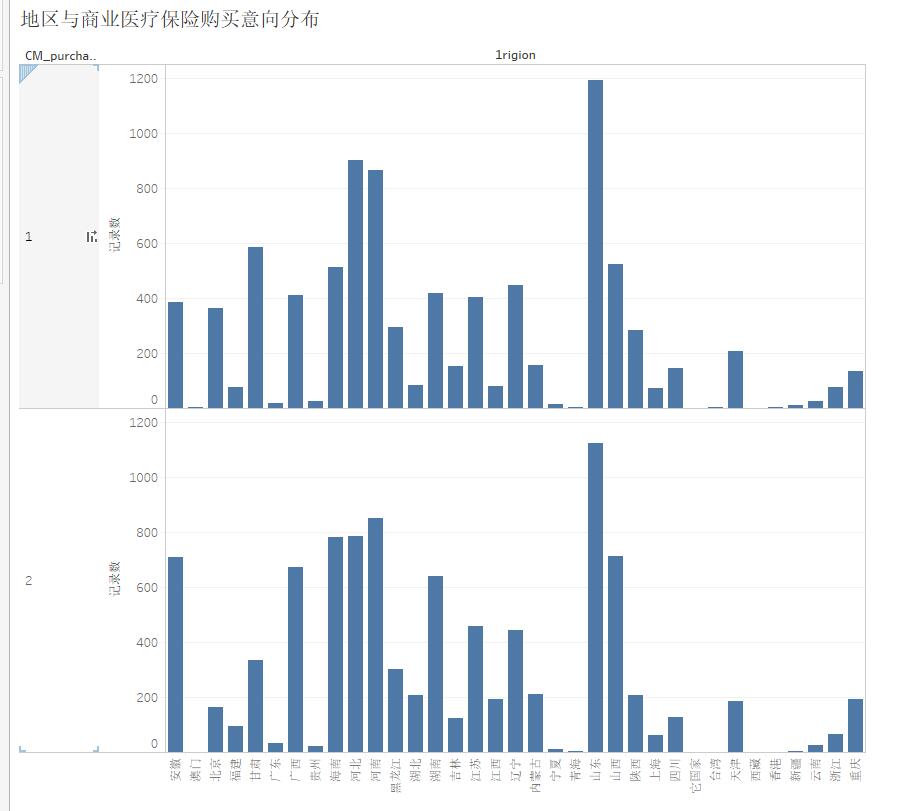
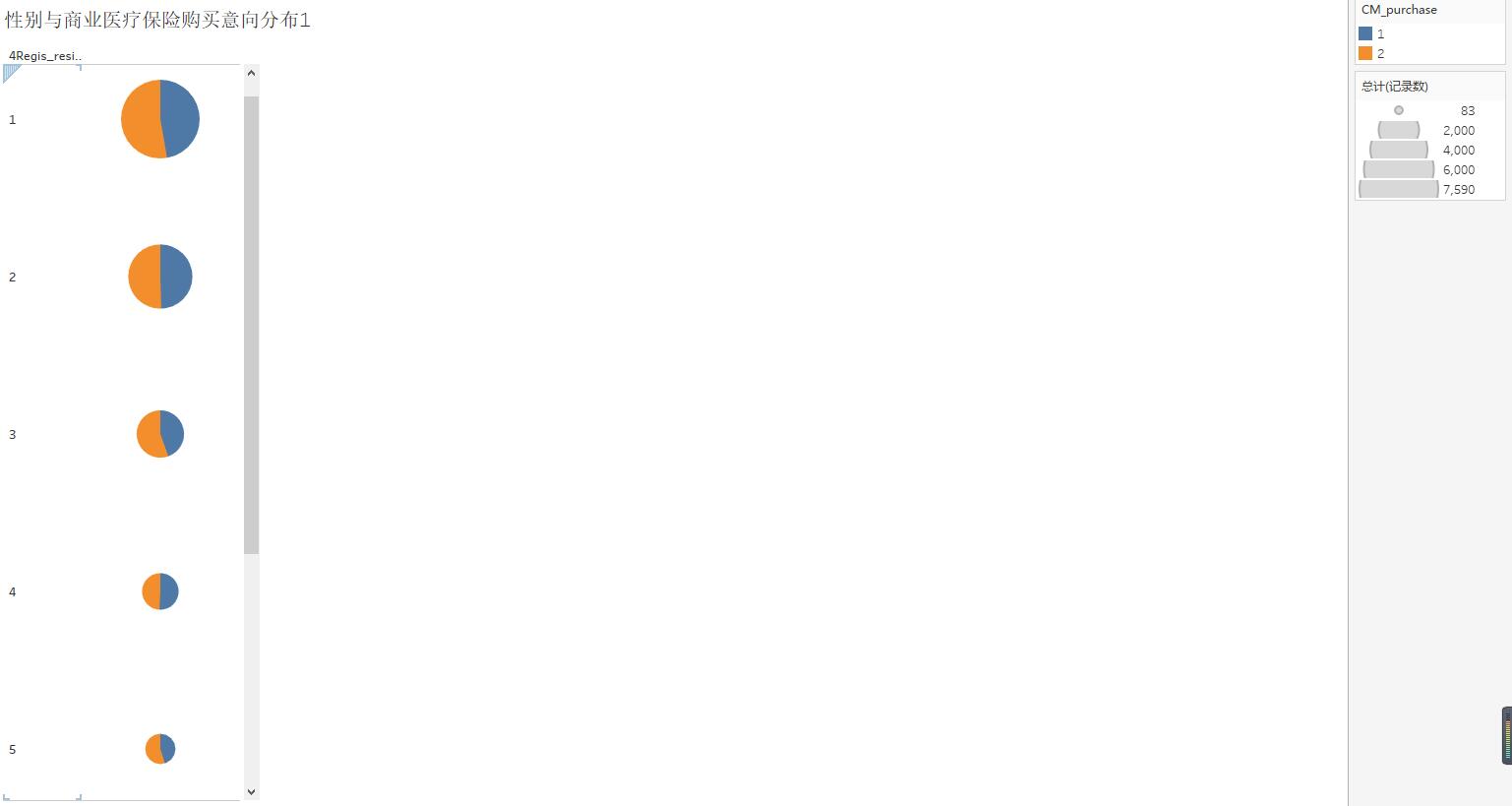


图2-5 地区与商业医疗保险购买意愿分布



图2-6 性别与保险购买意愿分布

其中左侧的1代表男性，左侧的2代表女性



户口类型

图2-7 户口类型与保险购买意愿分布

其中左侧的1,2,3,4,5-对应着编号原始数据中的户口类型

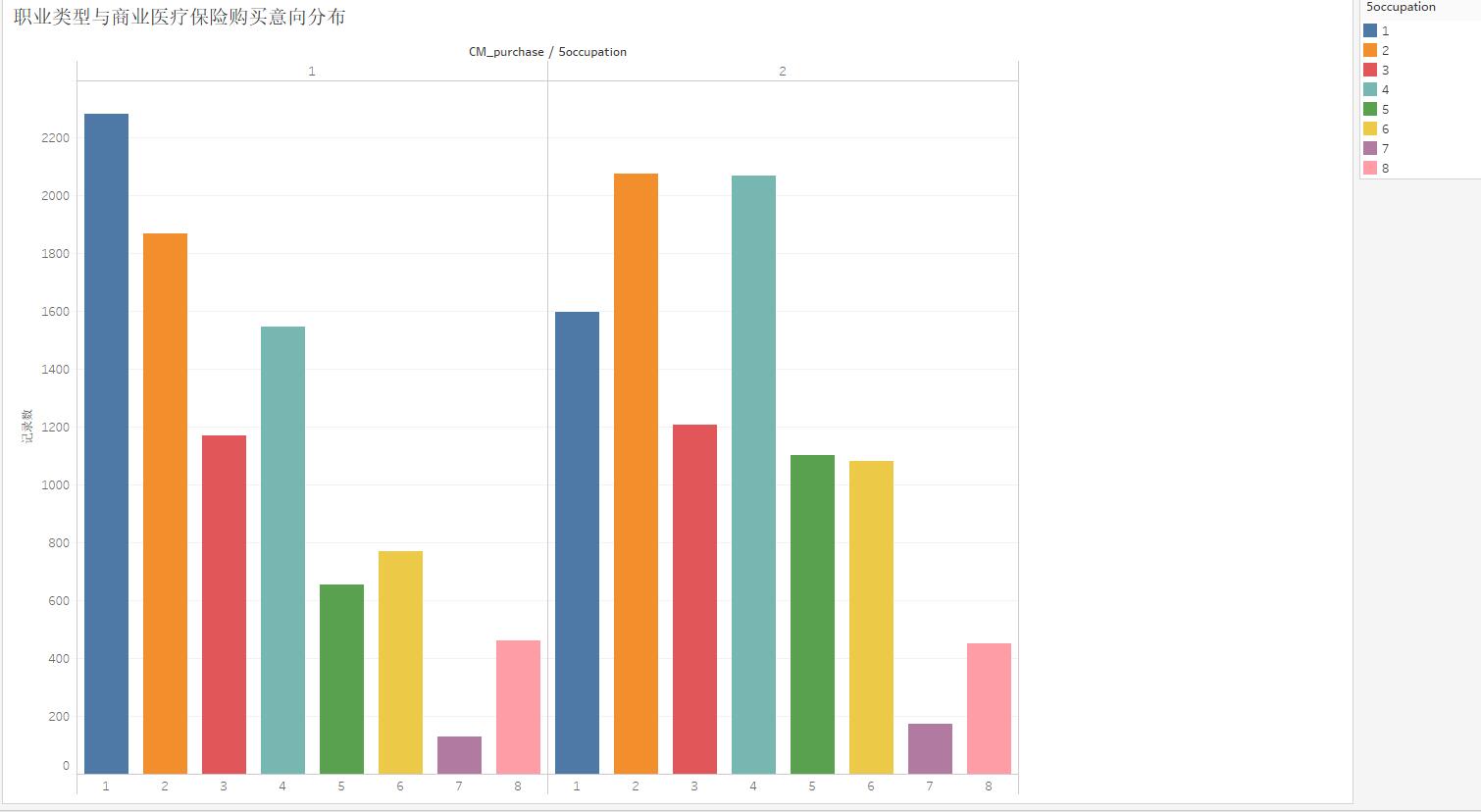


图2-8 职业类型与保险购买意愿的条形分布图

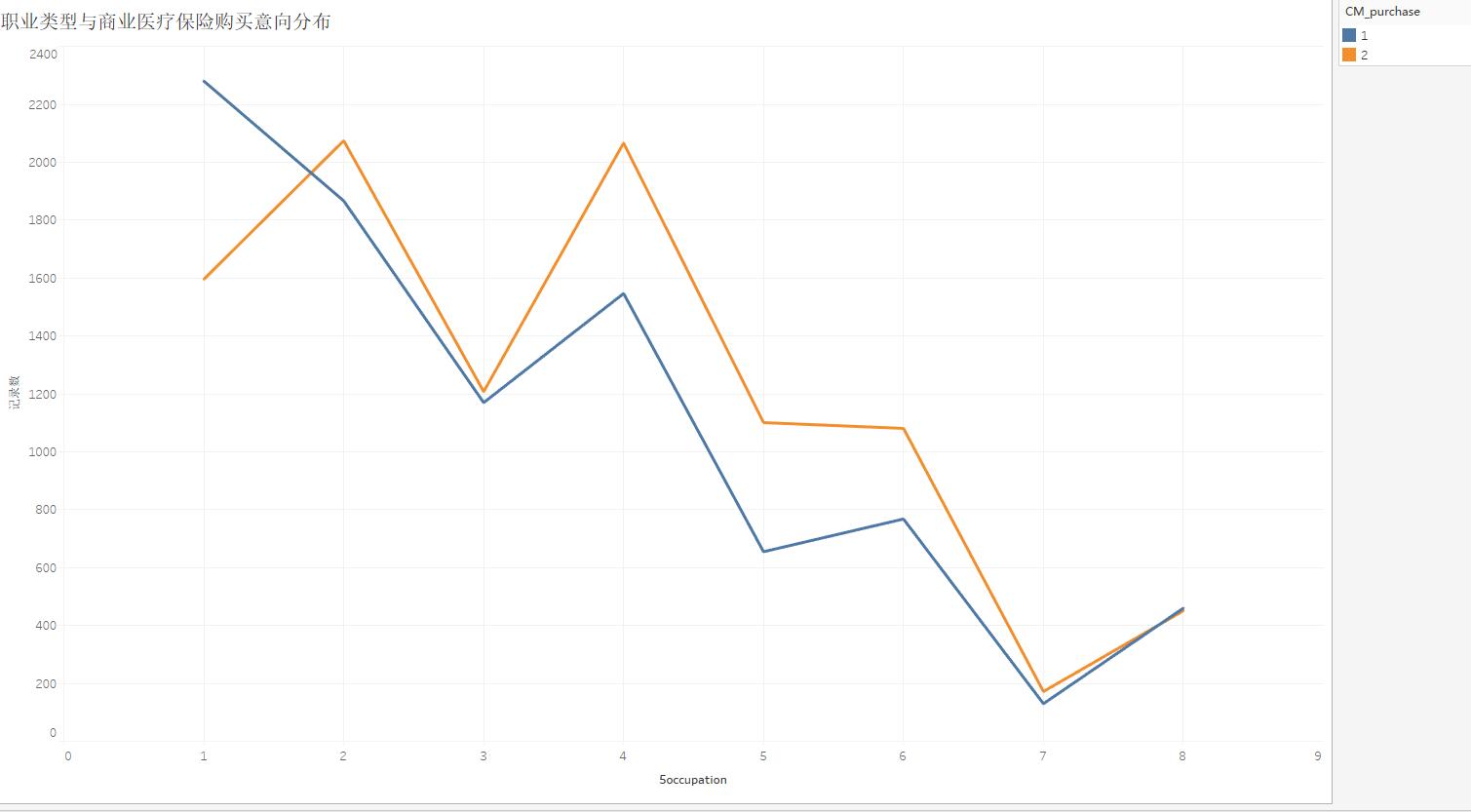


图2-9 职业类型与保险购买意愿的线形分布图

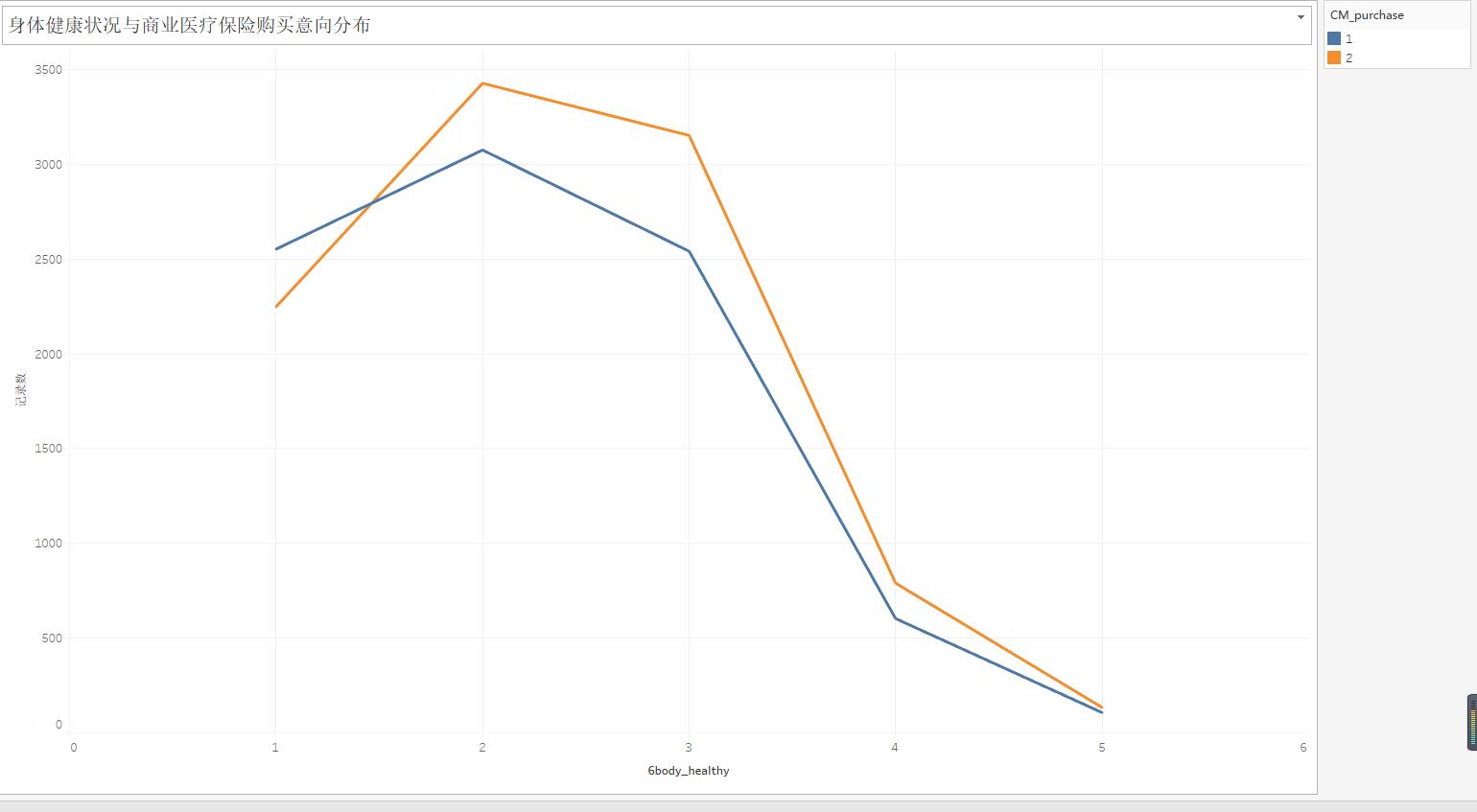


图2-10 身体健康状况与保险购买意愿的人群分布

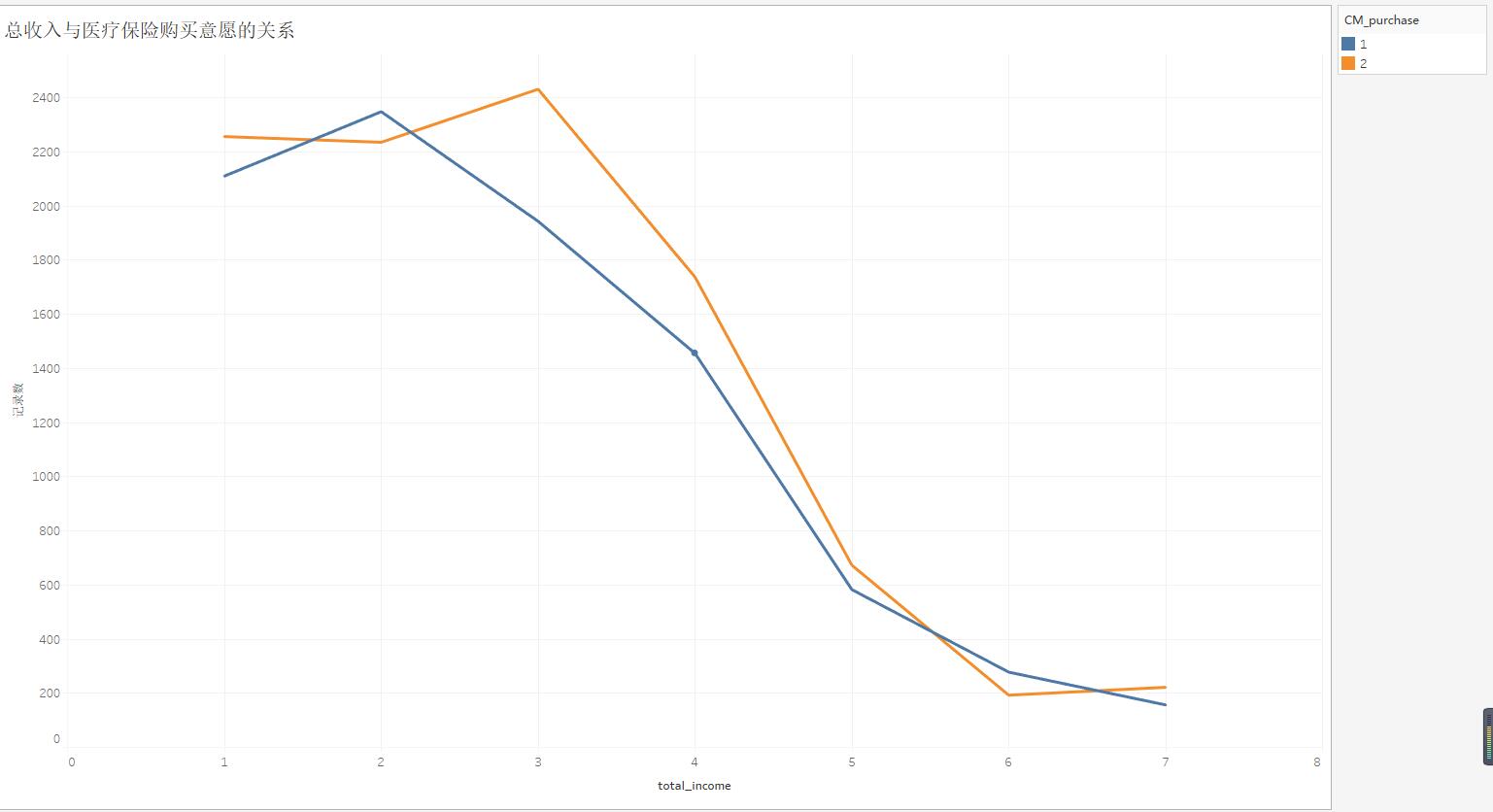


图2-11 目前收入与保险购买意愿的人群分布



图2-12 健康护理保险购买意愿与医疗保险购买人群分布



图2-13 重大疾病保险购买意愿与医疗保险购买人群分布

#### 可视化分析

分析上列可视化图示，从图2-4中，我们发现随着年龄的增长，愿意与不愿意购买保险的人群比例趋势保持相对稳定，但是也会出现一些异常点，34-37岁与40岁这两个年龄阶段的人保险购买意愿更强。从图2-5中，我们可以发现河北省的人群购买意愿更加强烈。从图2-6中国，我们发现男性比女性更乐意购买保险。从图2-7中，我们发现户口类型为本地农业户口，非本地农业户口，本地统一居民户口的人群更乐意购买保险。从图2-8与图2-9中，我们发现职业类型为个体经营者，有雇工所对应的职业的人群更乐意购买保险。从图2-10中，我们发现身体健康类型为很好的人群更乐意购买保险。从图2-11中，我们发现收入为1-2w的人群具有较强的保险购买意愿。从图2-12与图2-13中，我们发现三种保险的购买意愿具有趋同性。

### 期望风险因素可视化

如上文所示，我们已经将期望风险中的主要因素（收入、开销、居住环境、家庭关系、心理情绪、疾病、自理或活动能力）全部转化为特征，这里我们进一步对其进行可视化。如下列图所示。



图2-14 未来疾病与保险购买意向比例分布

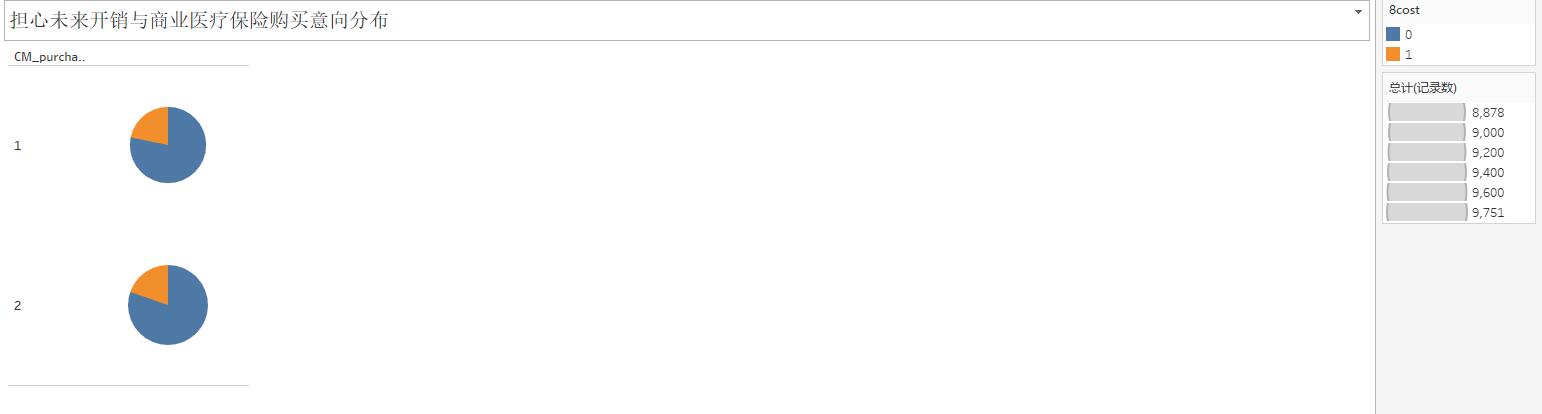


图2-15 担心未来开销与保险购买意愿



图2-16 担心未来居住环境与保险购买人群意向比例分布



图2-17 担心未来家庭关系与保险购买人群比例分布



图2-18 担心未来家庭关系与保险购买人群比例分布



图2-19 担心未来活动能力与保险购买人群比例分布

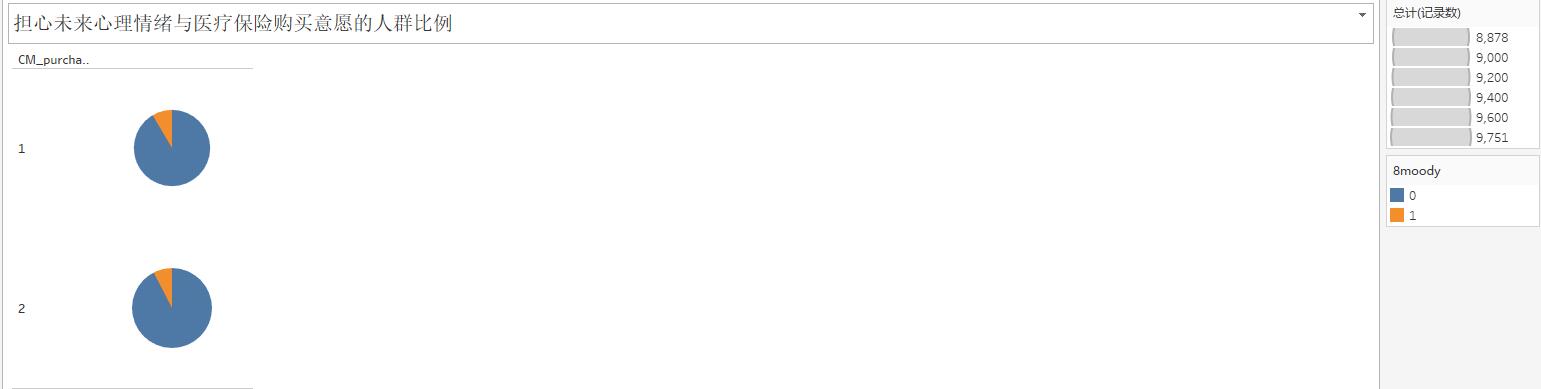


图2-20 担心未来心理情绪与保险购买人群比例分布

#### 可视化分析

通过观察各圆饼图，我们发现担心这些因素的人群购买意愿更强烈，但是影响差别不是太大。

### 总结概括相关性

通过分析现实与期望因素，我们发现现实因素对于保险购买的意愿影响更强烈，其中三种保险的购买意愿具有趋同性。所有的因素都对保险的购买意愿都具有或大或小的影响，因此我们决定将所有这些特征都加以考虑构建模型。

## 建立模型

考虑到数据都是离散数据，且每一个因素都具有决策的性质，这里我们暂且不考虑线性模型。于是我们决定从决策树模型出发一步一步选择更高级的模型。所有的这些模型都来自于机器学模型，然后通过计算机编程的形式让他们自动训练并拟合。

### 划分训练集与测试集

在用机器学习训练模型之前，我们需要将数据集拆分为训练集与验证集。训练集用于训练模型拟合数据。验证集用于检验所训练的模型的正确率。这里我们采用的是随机划分的方式，如下图3-1所示。



图3-1 划分训练与测试集

### 建立机器学习模型一

我们需要建立三个模型分别预测三种保险的购买意愿。首先我们选择的是决策树模型。最后建立的模型如图3-2所示，因为图太大，我们已经保存为pdf文件，见项目中的isbuy.pdf。

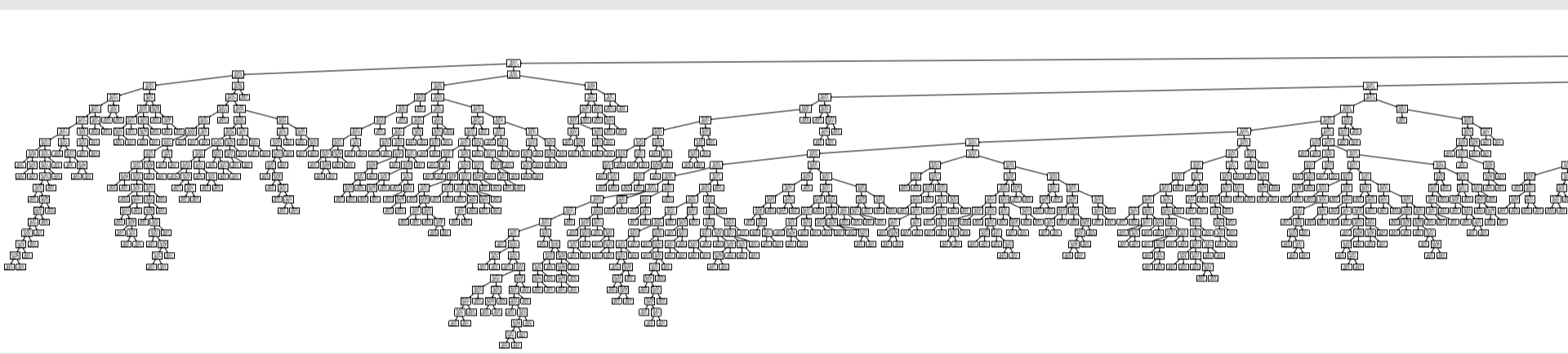


图3-2 决策树模型

#### 3.2.2 模型预测

我们将我们的决策树模型用于检验验证集，对于三种保险的验证正确率如下列图所示。



图3-3 重大疾病决策树验证正确率



图3-4 医疗保险决策树验证正确率



图3-5 健康护理保险决策树验证正确率

### 选择最优机器学习模型

上文中决策树模型对于三种保险的预测正确率都不高，只有70%左右，于是我们尝试了几种常用的其他机器学习模型，正确率效果如下列图所示。

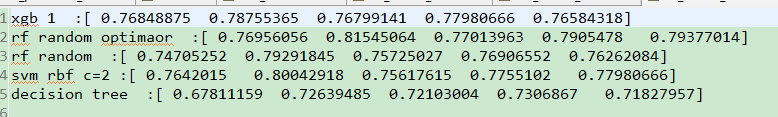


图3-6 医疗保险几种模型验证正确率

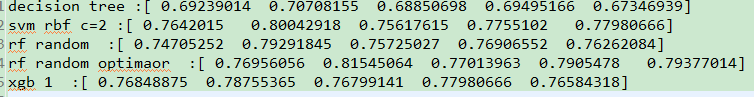


图3-7 重大疾病保险几种模型验证正确率

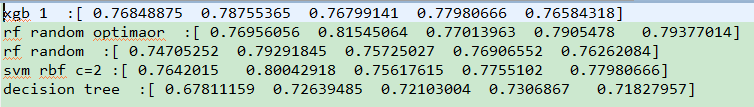


图3-8 健康护理保险几种模型验证正确率

通过上列图示，我们发现随机森林具有较好的泛化预测效果，其正确率接近80%，比起决策树将近提高了10%，传统的xgboost也没有太大的优势，因此我们最后的预测模型就是随机森林模型。因为我们是通过65个估计器集成，我们将其保存为文件。在项目中的位置如下图3-9所示。

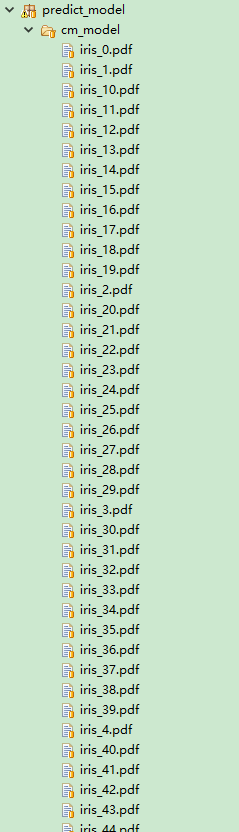


图3-9 随机森林模型

## 辅助预测方式

我们所得到的机器学习模型正确率不是太高，因此我们将原始数据中的保险购买的条件进行收集并分词之后，统计其靠前的词频。如下列图所示。

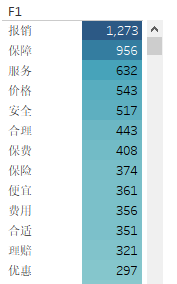


图4-1 医疗保险购买的条件

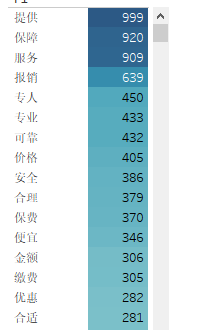


图4-2 重大疾病保险购买条件

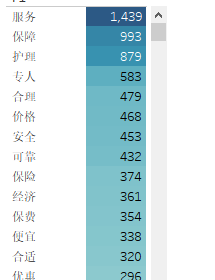


图4-3 健康护理保险购买条件

通过上面三图，可以看出三种保险购买的条件大相径庭，都需要保险机构提供比较好的服务，价格合理，安全可靠，专业机构等。所以在对消费者推广保险产品的时候，我们可以从这些方面出发。

## 总结

这次保险产品报告的主要目的，便是构造一个预测模型，能够根据人群的特征现实因素和期望风险因素对人群的保险购买意愿进行预测，有利于我们向那些潜在客户推销保险产品。本次报告的实验有较多不足。一是没有构建更多的特征，二是数据可能倾斜和不足，这会影响模型的建立。三是没有尝试更高级的数学模型，来对模型进行进一步优化。这次报告的写作主要方式就是刻画了模型建立的基本步骤，可能对于非该领域的人士来说看懂比较费劲。

## 6 附录

本次报告的全部项目内容都在我的Github上，网址为：https://github.com/BigBigRadish/insurance\_prediction，对于数据中的敏感信息全部删去及不能公开的数据都已全部转化为数值。我也已经打包本地目录，包括每一个文件的内容，请阅读readme.txt。如有疑问，请与我联系。邮箱为：luozhukun@163.com