

# 客户报案建议系统探索版

## 1. 项目介绍

2016 车险新费改后，下年保费会因为上年的出险次数而提升。

原保费计算公式： 保费 = （车价 x 费率 x 基础保费）x 调整系数

新保费计算公式： 保费 = 基准纯风险保费 / （1-附加费用率）x 费率调整系数（改革后的费率调整）

因此客户每年出险次数会对下年的保费造成影响，客户每次遭遇事故后选择是否出险也会对下年所需保费造成影响。因此此系统会在客户出险后对本次是否报案做出建议，系统会考虑客户的利益做出为客户做出预测和最好的选择。

## 2. 数据集

- （1）起保日期在 2013 年 – 2015 年之间的保单数据
- （2）非交强险（险种代码非 0507）的保单：交强险与本系统无关
- （3）车损险保费不为 0 的保单数据：因为要预测车损赔款和出险次数

## 3. 设计流程（具体版）

- a) 分析业务要求，列出潜在的分析字段和分析指标
- b) 建模数据的抽取与清洗
  - i. 缺失值
  - ii. 噪声值（离群点 & 不符合实际业务的点）
- c) 初步相关性分析检验和共线性排查
  - i. 删除高度线性相关该的
  - ii. 分析变量与出险次数的相关性，结合业务，筛选变量
- d) 观察自变量的分布（偏度，峰度）—— 分布不均衡的变量不利于后期模型拟合
  - i. 取对数
  - ii. 标准化
  - iii. 开方
  - iv. 平方
  - v. 指数
- e) 建模
  - i. 确定系统目标
  - ii. 回归
    - 1. 观察因变量分布 – （连续型因变量 / 计数型因变量）
    - 2. 假设检验
    - 3. 建模预测
    - 4. 模型测评与改进
- f) 分类
  - i. 确定不同算法需要的数据类型
  - ii. 处理数据
  - iii. 建模预测
  - iv. 模型测评与改进

## 4. 详细设计

### 4.1.1 数据准备

基于业务理解，筛选模型初期所选择的变量

所属模块	模块细分	特征要素	备注
保单信息		保单号	

	保单数据	险种代码	
		起保日期	
		终保日期	
		核保日期	
		业务渠道	
		机构代码	
		分公司代码	
		中心支公司代码	
		支公司代码	
		支公司名称	
		总保费	
		总保额	
		总折扣金额	
		手续费比例	
		代理人代码	
		车损险保额	
		三者险限额	
		附加乘员险保额	
		保单是否有免赔额	0 否 1 是
		车损险保费	
		三者险保费	
		附加乘员险保费	
		本车是否投保不计免赔险标识	0 否 1 是
		三者险是否投保不计免赔险标识	
		车损险附加险个数	
		三者责任险附加险个数	
车辆信息	车辆信息	新车购置价	
		实际价值	
		行驶区域代码	
		车辆已使用年限	
		车辆生产国别	
		机动车大类	
		机动车细类	
		车牌底色代码	
		号牌种类代码	
		标的车辆种类	
		被保人与车辆关系	
		是否约定驾驶员标志	
		是否投保新增设备标志	
		座位数	
		排量	
		使用性质	
理赔信息	理赔信息	出险次数	
		最早一次出险时间	
		赔付金额	
		案均赔款	

		报案周期	
		理赔周期	
		支付周期	
		是否注销拒赔/特殊事件	0 否 1 是
客户信息	客户信息	投保人年龄	如果身份证合法，从身份证提取，否则为空
		投保人性别	如果身份证合法，从身份证提取，否则为空
		被保险人年龄	如果身份证合法，从身份证提取，否则为空
		被保人性别	如果身份证合法，从身份证提取，否则为空
平台信息	平台信息	上年出险次数	
		上年赔款金额	
		上年浮动原因码	
		上年不浮动原因码	
		上年无赔付优待系数	
		自主核保系数	
		自主渠道系数	
		交通违法系数	
		客户忠诚度系数	

4.1.2 数据质量检验

抽取上述数据后，通过对其进行质量检验以及统计四分位点的观察分布，以进行初步的变量可用性评估和选择。

以下颜色为删除变量，为挑选子集后领相应缺失值占比变少后填充的变量。

变量变量	空值	空值率	零值	零值率	null	填充方式
保单号	0	0	0	0	0	
险种代码	0	0	0	0	0	
起保日期	NA	NA	0	0	0	
终保日期	NA	NA	0	0	0	
核保日期	NA	NA	0	0	0	
业务渠道	0	0	199346	0.28	0	
机构代码	0	0	0	0	0	
分公司代码	0	0	0	0	0	
中心支公司代码	0	0	0	0	0	
支公司代码	0	0	0	0	0	
总保费	0	0	0	0	0	
总保额	0	0	8	0	0	
总折扣金额	0	0	47000	0.07	0	
手续费比例	0	0	448049	0.63	0	
代理人代码	445188	0.63	0	0	0	衍生
车损险保额	0	0	273	0	0	
三者险限额	0	0	236	0	0	
附加乘员险保额	0	0	245841	0.35	0	
保单是否有免赔额	0	0	54009	0.08	0	
车损险保费	0	0	0	0	0	

三者险保费	0	0	209	0	0	
附加乘员险保费	0	0	245731	0.35	0	
被续保标志	0	0	494920	0.7	0	
下年保单号	494920	0.7	0	0	0	
连续被续保年限	0	0	494920	0.7	0	
续保标志	0	0	490842	0.69	0	
续保年限	0	0	490842	0.69	0	
本车是否投保不计免赔标识	0	0	664665	0.94	0	
三者险是否投保不计免赔标识	0	0	664681	0.94	0	
车损险附加险个数	0	0	5548	0.01	0	
三者险附加险个数	0	0	183019	0.26	0	
营销活动 ID	691775	0.98	0	0	0	
批改次数	0	0	626246	0.88	0	
是否批改关系人	0	0	689960	0.97	0	
是否批改驾驶人	0	0	707775	1	0	
新车购置价	0	0	0	0	0	
实际价值	0	0	0	0	0	
行使里程	0	0	234316	0.33	0	空值太多，删掉
行驶区域代码	0	0	0	0	0	
车辆已使用年限	0	0	176244	0.25	0	
车辆生产国别	1	0	0	0	0	按比例随机填充
车辆厂家代码	441621	0.62	0	0	0	空值太多，删掉
机动车大类	3	0	0	0	0	填充为占比最大的值
机动车细类	3	0	0	0	0	填充为占比最大的值
机动车子类	496130	0.7	0	0	0	空值太多，删掉
车牌底色代码	187693	0.26	0	0	0	决策树填充
号牌种类代码	16	0	0	0	0	按比例随机填充
标的车辆种类	0	0	0	0	0	
被保人与车辆关系	0	0	0	0	0	

是否约定驾驶员标志	228284	0.32	464922	0.66	0	按比例随机填充
是否投保新增设备标志	228284	0.32	480515	0.68	0	按比例随机填充
座位数	0	0	3939	0.01	0	中位数填充（填为 5）
吨位数	0	0	621130	0.88	0	空值太多，删掉
排量	0	0	81091	0.11	0	按标的车辆种类,使用性质,机动车大类，机动车细类用决策树回归填充
使用性质	0	0	0	0	0	
交管车辆类型	593	0	0	0	0	按标的车辆种类,使用性质,机动车大类，机动车细类用决策树回归填充
行驶证类型	647868	0.91	0	0	0	空值太多，删掉
出险次数	475692	0.67	0	0	475692	零值填充
赔付金额	475692	0.67	21262	0.03	475692	零值填充
案均赔款	475692	0.67	21262	0.03	475692	零值填充
报案周期	475692	0.67	202177	0.28	475692	零值填充
理赔周期	475692	0.67	33574	0.05	475692	零值填充
支付周期	475692	0.67	21945	0.03	475692	零值填充
是否注销拒赔	475692	0.67	233701		0	全是 0，删掉
投保人性别	119249	0.17	0	0	0	按比例随机填充
投保人年龄	119249	0.17	0	0	119249	与"被保人年龄"比较填充，再按上下四分位数随机填充
投保人职业类型	709241	1	0	0	0	空值太多，删掉
投保人学历代码	709393	1	0	0	0	空值太多，删掉
投保人是否有身份证号	0	0	119249	0.17	0	
投保人是否有电话	0	0	333268	0.47	0	
投保人电话是否来源于电销	0	0	707069	1	0	
被保人性别	114346	0.16	0	0	0	33-47 随机填充
被保人年龄	114346	0.16	0	0	114346	33-47 随机填充
被保人职业类型	709223	1	0	0	0	空值太多，删掉

被保人学历代码	709393	1	0	0	0	空值太多，删掉
被保人是否有身份证号	0	0	114346	0.16	0	
被保人是否有电话	0	0	328812	0.46	0	
被保人电话是否来源于电销	0	0	707066	1	0	
上年出险次数	79	0	535646	0.76	79	零值填充
上年赔款金额	535725	0.76	0	0	535725	零值填充
上年浮动原因码	30683	0.04	0	0	0	去子集后加入，再填充
上年不浮动原因码	30683	0.04	0	0	0	与上年浮动原因码类似，删掉
上年无赔付优待系数	512894	0.72	0	0	512894	空值太多，删掉
自主核保系数	664546	0.94	0	0	664546	空值太多，删掉
自主渠道系数	664546	0.94	0	0	664546	空值太多，删掉
交通违法系数	451421	0.64	0	0	451421	空值太多，删掉
客户忠诚度系数	47858	0.07	380699	0.54	47858	去子集后加入，再填充

#### 4.1.2.1 填充代码

填充代码.r

#### 4.1.2.2 非回归填充

非回归填充.r

### 5. 相关性分析

通过相关性分析筛选变量，提升模型准确度

#### 5.1 相关性分析代码（将字符型按比例填充成数值型，计算相关性系数并排序）

相关性分析.r

#### 5.2 相关性结果在 corr.csv 表中

corr.csv

### 6. 建模

#### 6.1 观察因变量分布

出险次数 - 计数型因变量

代码：hist(new\_merge\$出险次数)

车损赔款 - 累加连续型因变量

代码：hist(new\_merge\$车损赔款)

## 6.2 自变量分布转换

使自变量分布峰值与偏度不理想的数据分布更均匀，有利于建模  
自变量分布转换. r

## 6.3 假设检验

判断自己建模或者对数据分布分析的检验

假设检验. r

## 6.4 看拟合情况

观察拟合情况，以直观的看出自己的假设是否符合实际情况

```
library(gamlss)
histDist(new_merge$出险次数)
```

## 6.5 观察 AIC 值

AIC(Akaike Information Criterion, 考虑了模型的统计拟合度以及用来拟合的参数数目，AIC 值越小的模型要优先选择)

## 6.6 出险次数 - 泊松模型回归

泊松模型回归. r

结果：

### ► 泊松回归：

- 零膨胀泊松回归（正常，标准化后）： 大约 59%左右
- 零膨胀泊松回归（取对数后）： 大约 59%左右，分布不均匀，最大到 3
- 零膨胀泊松回归（取平方后）： 大约 62%左右 - 分布不是很均匀，最大预测到 2

## 6.7 出险次数 - 负二项回归

负二项模型回归. r

结果：

- 零膨胀负二项回归（正常，标准化后）： 大约 61%左右
- 零膨胀负二项回归（取对数后）： 大约 61%左右，分布不均匀，最大到 3
- 零膨胀泊松回归（取平方后）： 大约 68 左右 - 分布不是很均匀，最大预测到 3

## 6.8 出险次数 - 神经网络分类

神经网络分类. r

结果：二分类任务能达到 70%左右，但是只能预测全为 0 的情况

## 6.9 出险次数 - 先分类再回归

出险次数先分类再回归. r

结果：与泊松、负二项模型结果相近

## 车损赔款神经网络分段分类.r

### 6.11 车损赔款 - 伽马回归/逆高斯分布/高斯分布/决策树分段预测

**结果：** 分段后能达到 65%左右的准确率，逆高斯分布最高。

- a) **假设检验**：当原假设被拒绝时，只能说明在统计意义下某种效应是存在，这个称为统计显著性。但在实际中，这种效应也许很小，这称为实际显著性。因为当样本容量较大时，任何与原假设很小的差异都可能在统计意义上是显著的，而在实际中可能不显著。（有意义，但是不可全信）
- b) **贝叶斯回归**：防止过拟合
- c) **抽样方法**：取训练集时要符合因变量的分布取相应比例的训练集（实际情况适当调整会提高准确率，然而可能会引入噪声）——结合业务情况进行抽样
- d) **调整离群点**：离群点会影响一个变量的解释能力（去掉离群点会增加这个变量的解释能力）回归模型受噪声影响大
- e) **交叉验证，迭代训练**：减少噪声对模型的影响
- f) **变量的解释能力**：单个变量可能解释能力很弱或者不显著，组合后可能会比强变量解释能力更强，新变量？
- g) **组合模型**：鉴于出险次数与车损赔款的 0 值较多，可以先做分类后做回归，二者结合的模型会比单个模型更有说服力
- h) **集成学习**：分类时结合多个分类器的结果，投票选出更优的结果

```

graph LR
    A[客户遭遇事故] --> B[预测次数]
    B --> C1[第一次]
    B --> C2[第二次]
    B --> C3[第三次]
    C1 --> D1[客户预估当次赔款]
    C2 --> D2[客户预估当次赔款]
    C3 --> D3[客户预估当次赔款]
    D1 --> E1[涨幅保费 = 1 x 原保费]
    D2 --> E2[涨幅保费 = 1.25 x 原保费]
    D3 --> E3[涨幅保费 = 1.5 x 原保费]
    E1 --> F1[推荐报案概率]
    E2 --> F2[推荐报案概率]
    E3 --> F3[推荐报案概率]
    B --> G[超出预测次数]
    B --> H[实际出险次数]
    F1 --> I[是否报案]
    F2 --> J[是否报案]
    F3 --> K[是否报案]
    G --> L[报案次数]
    H --> L
    I --> L
    J --> L
    K --> L
    L --> M[客户下次保费]
    N[客户今年保险] --> O[客户交保费]
    O --> P[客户历史记录]
    O --> Q[基准保费]
    P --> R[预测今年出险次数]
    P --> S[预测今年车损赔款]
    R --> T[当次赔款占比 = 当次赔款 / 预测今年车损赔款]
    S --> T
    T --> U[客户历史记录]
    T --> V[客户下次保费]
    
```



## 8.1 具体系统建议方法

- ▶ 预测出险次数下：
$$\frac{\text{当次预估车损赔款}}{\text{预测车损赔款}} \times \text{费率调整系数} \times \text{风险系数} \times \text{第一次专有系数} = \text{推荐报案概率}$$
  - ▶ 第一次出事故：
  - ▶ 第二次 v 第一次出事故|：
$$\frac{\text{当次预估车损赔款}}{\text{剩余预测车损赔款}} \times \text{费率调整系数} \times \text{风险系数} = \text{推荐报案概率}$$
  - ▶ 第三次 v 第二次 v 第三次出事故：同上
- ▶ 系统依赖于：客户需具备相应车险赔付常识。出事故后对赔付金额的预测误差不能过大
- ▶ 风险系数需要进一步提炼：
  - ▶ 出险后本年车损赔款与明年保费的之间比较的系数
  - ▶ 预测误差的系数