家庭财产保险的反欺诈识别与定价策略方案

吴凡璐

一、背景介绍

家庭财产保险是理财险的一种，其功能是对房屋主体和屋内财产的保护。当房屋受到暴风雨、火灾、闪电、盗窃等带来的损失时，购买保险的屋主可向保险公司索赔。对于保险公司而言，根据顾客的不同需求进行差别定价，通过数量化方法制定精准的保险定价策略，对于增加公司效益意义重大。

在互联网时代，很多保险公司都设立了自己的保险询价网站，可在线处理保险业务。人们只要在询价页面填入个人信息，如房屋面积、楼层数、家人数等，就可以得到网站给的报价，在看到报价后，部分人在确定价格合适后就会支付保单购买保险，成为真正的客户。等到家中遭遇损失后，同样可以通过网站提交赔付申请，经审核通过后就会收到赔付金。用户填写的个人信息和一系列登录行为都会被记录在网站日志中。

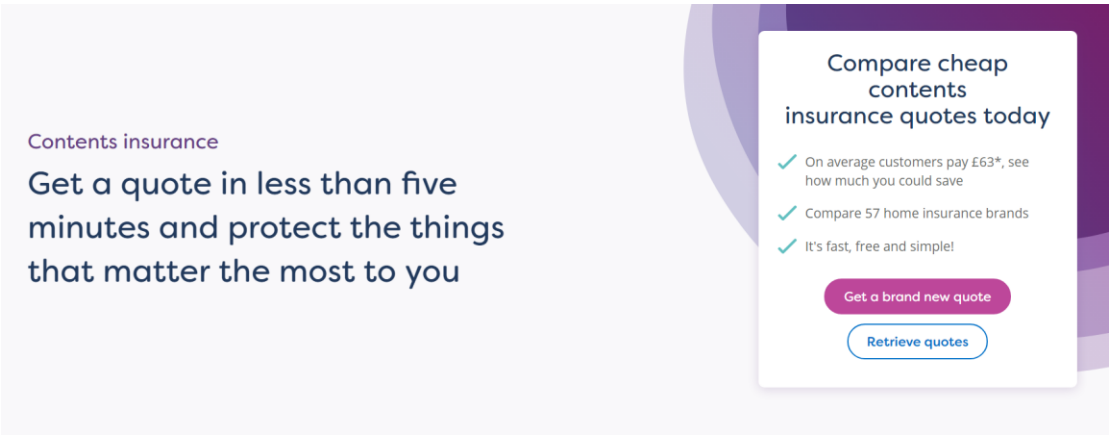

图 1：保险公司网站询价入口示意图

然而，由于保险业务转到线上，也使得骗保机器人有机可乘。不法分子通过制作登录网页的机器人，使用不同 ID 登录网站，自动完成填写信息、支付保费、申请赔偿的流程，其行为构成保险诈骗，对保险公司造成巨大损失。如何从日志数据中识别此类诈骗者的行为模式，从而实现反欺诈，是保险公司的重要任务。

本案例将以澳大利亚 Yuumi 保险公司为对象，通过分析其提供的家庭财产险客户的日志数据，从中找出具有欺诈嫌疑的机器人行为模式，并对正常投保的客户进行保险定价。

二、数据来源与说明

本案例数据来自 Kaggle 竞赛平台提供的家庭财产险日志数据[1]，数据分为训练集和测试集，其中，训练集有 570095 条用户日志信息，测试集有 157852 条记录，每条记录中记录了每位用户在不同时间登陆网站进行不同操作的内容和时间戳，以训练集为例，原始数据格式如下：

表 1：原始数据部分样本展示

message	timestamp
7a9b9c479a2840f - pc_browser - **Quote Started for customer:** 112b2c	1388508092
c285423ba1474db - pc_browser - **Quote Completed for customer:** 112b2c with json payload {'gender': 'male', 'name': 'Tara Mcdonald', 'household': [{'age': 39, 'name': 'Kristen Mcdonald', 'gender': 'female'}], 'age': 41, 'address': '13 Frederickborough, Trevor Fort', 'email': 'Tara Mcdonald@smith.com', 'home': {'square_footage': 268.02897031547616, 'number_of_floors': 1, 'type': 1, 'number_of_bedrooms': 1}}	1388508460
85eab230e6c242a - mobile_app - **Quote Incomplete for customer:** a9b70e with json payload {'gender': 'male', 'name': 'Alison Brown', 'household': [{'age': 46, 'name': 'Shawn Brown', 'gender': 'female'}, {'age': 39, 'name': 'Patrick Wilson', 'gender': 'male'}, {'age': 46, 'name': 'Jane Mueller', 'gender': 'male'}, {'age': 35, 'name': 'Linda Rocha', 'gender': 'male'}, {'age': 40, 'name': 'Billy Davis', 'gender': 'female'}, {'age': 13, 'name': 'Marc Brown', 'gender': 'female'}, {'age': 2, 'name': 'Brittany Brown', 'gender': 'male'}], 'age': 44, 'address': '45 Lydiaside, Jeffrey Shores Apt. ', 'email': 'Alison Brown@gonzalez.com', 'home': {'type': 1}}	1388547177
8b63a9fd26ce44c - pc_browser - **Payment Completed for customer:** 97f991	1388516228
d4893ae40f0942e - mobile_app - **Claim Started for customer:** 655aaa	1388524241
44496e6ef55f402 - pc_browser - **Claim Denied for customer:** 63b0a8 - reason : fraud	1388526026
58886b065c2a43a - pc_browser - **Claim Accepted for customer:** b9196d - paid $10656.03	138852904

[1] https://www.kaggle.com/c/2019-unsw-tsinghua-datathon-round-1/data

在“message”列中是每位用户登陆网站的日志内容，从左到右是日志编号、登录设备类型、业务类型（部分业务包含个人信息）、用户 ID；“timestamp”列记录的是每个日志的时间戳。

其中，“业务类型”在训练集中有 7 种：

- Quote Start – 潜在客户开始询价
- Quote Incomplete – 潜在客户填写（部分）信息后离开询价页面
- Quote Completed – 潜在客户完成询价并得到保险价格信息
- Payment Completed – 潜在客户支付保单并成为正式客户
- Claim Start – 客户提交索赔申请
- Claim Denied – 客户索赔被拒绝并得到被拒绝的理由（是否欺诈）
- Claim Accepted – 客户得到索赔金

将原始数据框进行重构并提取特征，初步得到以下变量：

表 2：数据变量说明表

变量类型		变量名	详细说明	取值范围
因变量		Fraud：是否欺诈	定性变量：2 个水平	1 代表欺诈，0 代表非欺诈
		Paid_amount：赔付金额	单位：美元	0~20912.89
自变量	房屋特征	square_footage：房屋面积	单位：平方米	85.25~866.09
		type：房屋类型	定型变量：2 个水平	1 代表房型 1，0 代表房型 2
		number_of_bedrooms：房间数	单位：个	1~5
	家庭性别分布特征	gender：申请者性别	定性变量：2 个水平	1 代表男，0 代表女
		nfamily_male：家中男人数	单位：个	0~8
		nfamily：家人数	单位：个	0~10
	家庭年龄分布特征	age：申请者年龄	单位：岁	24~92
		nf_children：家中孩子数	单位：个	0~5
		nf_working：家中劳动力数	单位：个	0~8
		family_age_mean：家人平均	单位：个	0~88

	年龄		
	family_age_range：家人年龄极差	单位：个	0~87
	family_age_std：家人年龄方差	单位：个	0~39
	family_age_min：家人最小年龄	单位：个	0~88
行为特征	browser：登录网页的设备	定性变量：4个水平	0 代表手机浏览器，1 代表电脑，2 代表手机客户端，3 代表电话
	t4：支付保费时间与申请索赔时间的差值	单位：秒	0.0~80464873.9

三、描述性分析

在对赔付金额和是否欺诈的问题进行建模前，首先对各变量进行描述性分析，以初步判断影响欺诈识别和保险定价的因素，为后续研究做铺垫。

（一）因变量

1. 是否欺诈

在 102497 个用户中，申请索赔遭到拒绝并被认为是欺诈的用户有 1714 个，占 1.6%。剩下无欺诈行为的用户中，有没有申请索赔的用户和索赔了并得到赔款的用户。

2. 赔偿金额

在本案例中，赔偿金额的最小值为 0，有 76635 个，占全部有赔偿样本的 75%；赔偿金额大于 0 的样本中，最小值是 1909.31 美元，索赔人是一位 37 岁的独居女性，其房屋面积仅有 96 平方米；最大值是 20912.89 美元，索赔人是一位 71 岁的女性，其房屋面积为 492.68 平方米，有 7 个家人。

通过赔偿金额直方图可以看到，除了赔偿金额为 0 的样本外，其余样本呈现

正态分布，平均值是 6892.16 美元，中位数为 6676.89 美元。

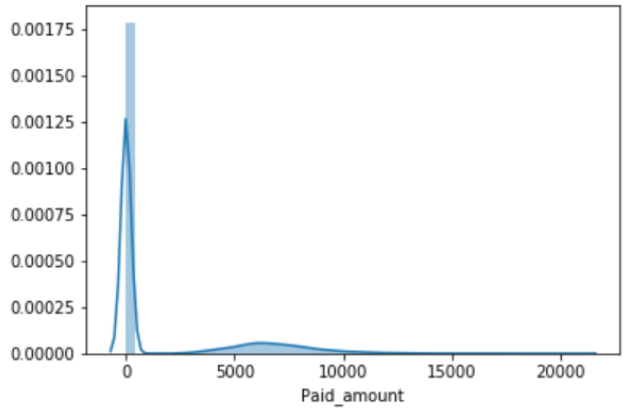

图 2：赔付金额分布直方图

（二）自变量：行为特征

在进行定价之前，我们要首先筛选出可疑的欺诈者，本案例中，反欺诈的治理对象是机器人。机器人是由人类编写的网络程序，可以多次且快速地登录网站完成索赔流程，其行为模式也会出现与人类不一样的地方，下面我们就提取的两个行为特征进行分析，判断其与是否欺诈的关系。

1. 是否欺诈与登录设备的关系

如图 3 所示，欺诈用户与非欺诈用户在登录设备上最大的不同在于使用电话办理保险业务的比例。非欺诈用户中有近 10%的用户使用电话，而欺诈用户使用电话的比例不到 0.1%。这一现象符合常识，一般编写机器人的程序大多是在手机和电脑上运行的，通过电话进行诈骗的极少。

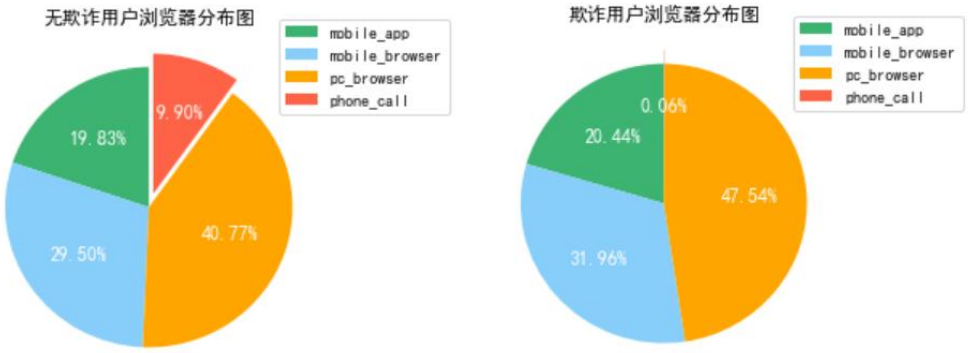

图 3：欺诈与非欺诈客户登录设备分布饼图

2. 完成支付与申请索赔时间差的分布

从完成支付与申请索赔时间差的直方图可以看出，有 1711 个用户的时间差

为 0，这违背了常识，一般人完成支付后到开始申请索赔由于手动切换页面会有不可避免的时间差，没有任何时间差的只可能是机器人，对于这部分用户，可以直接判定其欺诈并将其定价为 0。

值得注意的是，训练集中判定为欺诈的用户数量为 1714，其中 1711 个用户是机器人，只有 3 个用户是人类，也就是说，本案例中，绝大多数欺诈者都是机器人。

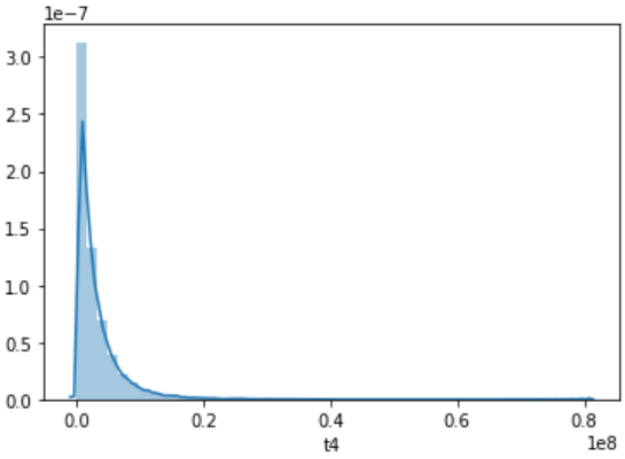

图 4：支付保费与申请赔款时间差分布直方图

（三）自变量：房屋特征

1. 房屋面积与赔偿金额的关系

图 5 是房屋面积与获赔金额的散点图。从图中可以看出，获赔金额和房屋面积呈现出较明显的簇状线性关系，二者成正相关，即，房屋面积越大，获赔金额越多。而簇状模式可能意味着还存在第三个变量对二者关系产生影响，比如地区。在经济发达的地区和经济较不发达的地区，相同的房屋面积有不同的价值，猜测获赔金额实际上和房价直接相关。

而图 6 是展示居住在不同房型的索赔者赔付金额的分组箱线图。从图中可以看出，1 号房型的获配金额明显大于 0 号房型。具体 1 和 0 代表什么含义，数据提供方没有给出解释，猜测可能是代表是否租住。

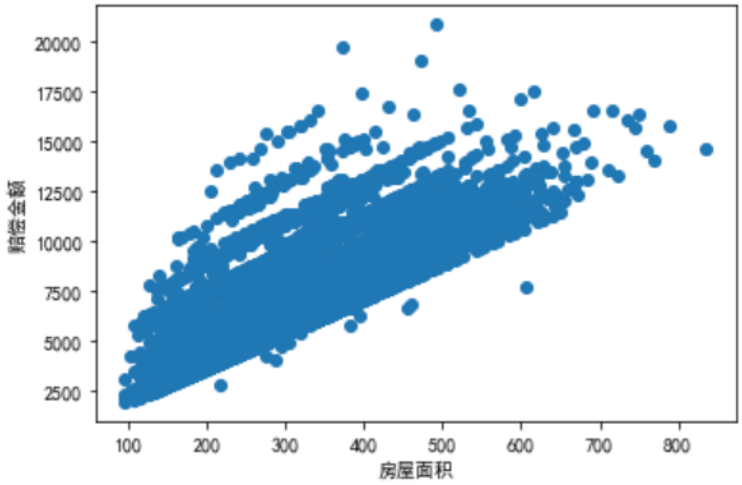

图 5：房屋面积与获赔金额散点图

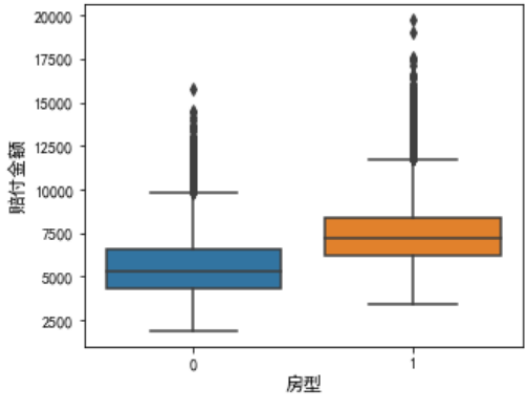

图 6：不同房型赔付金额分布箱线图

（四）自变量：家庭性别分布特征

图 7 展示了索赔者性别和赔付金额的分组箱线图，从图中可以看出，不同性别的获赔金额没有显著差别，在赔款问题上不存在性别歧视。

图 8 展示了家人总数和家中男性数量与获赔金额的关系，由图可见，家人总数越多，男性家人越多，获赔金额越多。可能是因为家庭规模越大，家中财物越多，越倾向于购买更大金额的保险。

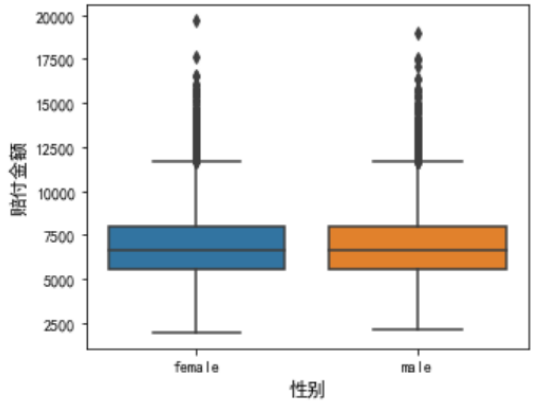

图 7：不同性别赔付金额分布箱线图

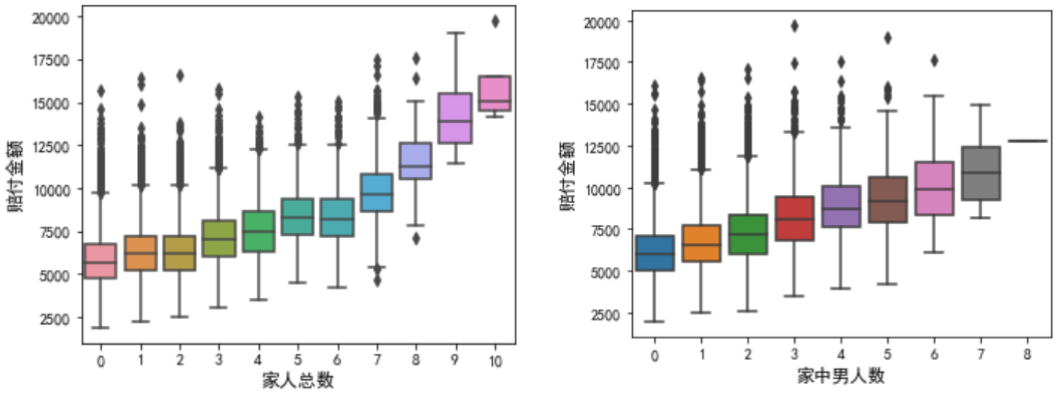

图 8：不同家人总数与家中男人数中获赔金额分布箱线图

（五）自变量：家庭年龄分布特征

图 9 展示了家中孩子数与工作人口数和获赔金额的箱线图，从图中可以看出，孩子数越多，工作人口数量越多，获赔金额越多。

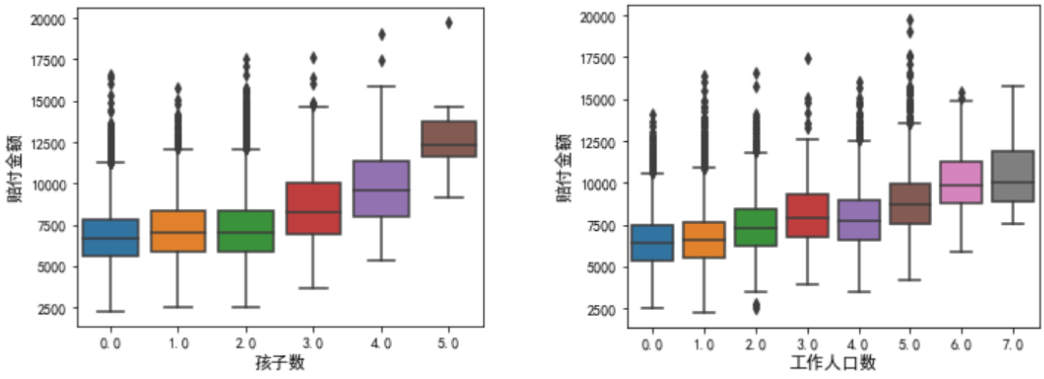

图 9：不同家中孩子数和家中劳动力数中获赔金额分布箱线图

（六）因变量和自变量相关系数图

图 10 是所有变量的相关系数图，从图中可以看出，申请者年龄、家中卧室数目、家庭成员平均年龄、家庭成员最小年龄与获赔金额的相关系数较小，这些变量不进入模型；而家中孩子数和家中劳动力数与家庭总人数的相关系数超过了0.3，可能出现信息重叠的现象，因此将这两个变量除以家庭总人数，得到家中孩子比例和家中劳动力比例两个衍生变量进入模型。在所有变量中，房屋面积和获赔金额相关系数最高，达到 0.76，可能是影响定价最重要的因素。

经过初步筛选，决定进入模型的变量为：房屋面积 square_footage，房屋类型 type，家庭成员数 nfamily，家庭成员年龄极差 family_age_range，家庭成员年龄方差 family_age_std，家中劳动力比例 rate_working ，家中男性比例 rate_male ，家中孩子比例 rate_children。

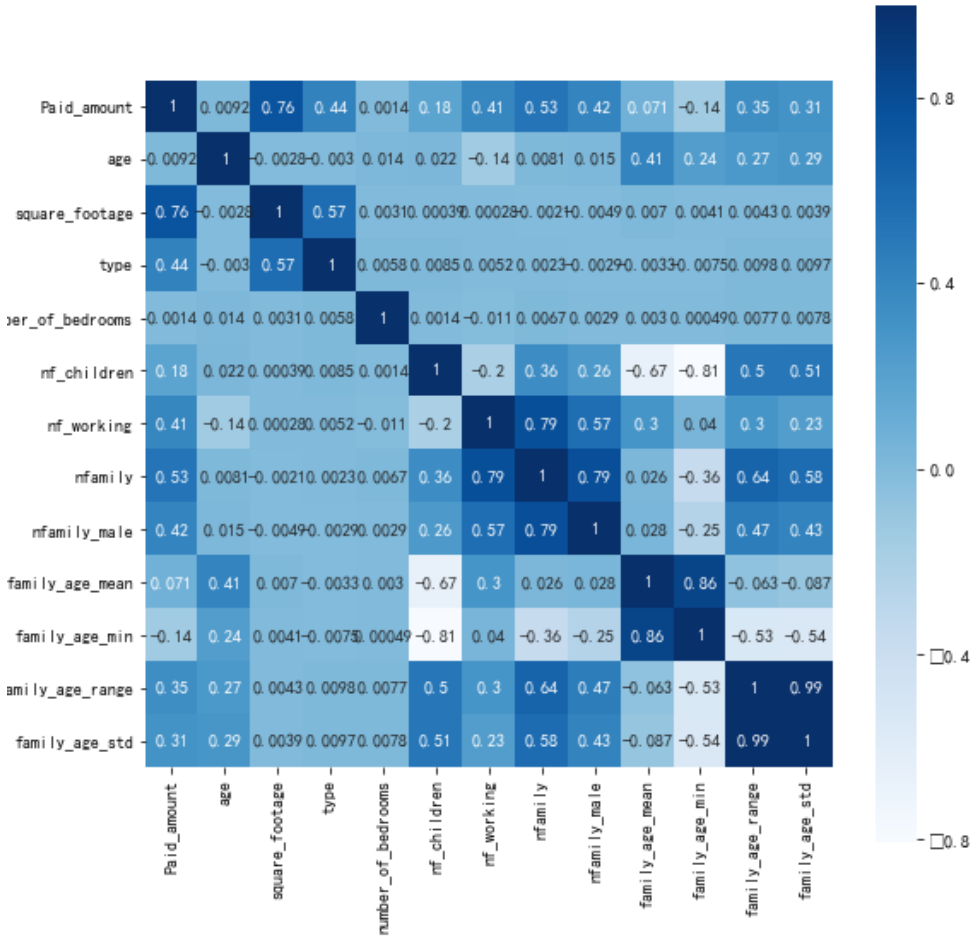

图 10：所有变量相关系数图

四、模型建立

1. 将所有支付保费与申请索赔的时间差为 0 的客户定价为 0

认为支付保费与申请索赔的时间差为 0 的客户是机器人，这部分用户的行为属于欺诈，不能给与放款，直接定价为 0。

2. 在训练集中除去有欺诈记录的样本

训练集中，有 4 名欺诈者获得了赔偿，他们的 ID 是 1cf225, ba87a0, c9fb57, cf3686。这可能是由于客户可以多次申请赔款，有些欺诈客户在之前的申请中没有被识别出来而获得了赔款，而在后面的申请中被识别出来并被判为欺诈，这部分客户的赔款金额和个人信息不能作为训练样本进入回归模型。

3. 对剩下的申请了赔款并得到赔偿金的样本建立 CatBoost 回归模型

在非欺诈的用户中，有 26%的用户得到了赔付金，其余客户没有进行索赔，这部分用户中，有未来可能会欺诈的用户，也有将来可能会进行索赔的用户，但由于其索赔金额缺失，不能为回归模型提供信息，不进入回归建模。最终进入回归建模的样本数为 25858。

本案例使用 CatBoost 算法对因变量和自变量的关系进行建模。CatBoost 算法是梯度提升机器学习算法库中的一种，由 Yandex 开发并在 2017 年开源，其性能在 ML 标准数据集上的表现优于 XGBoost 和 LightGBM，是广泛用于各种分类和回归问题的一种集成树模型，在商业反欺诈领域也应用广泛。CatBoost 使用在各种统计上的分类特征和数值特征的组合将分类值转换成数字，故而不需要特别处理分类变量，也不需要对缺失值进行过多预处理。其对超参数调优的依赖性不大，模型鲁棒性较好。基于以上优点，本文用此模型进行回归分析。

我们随机将原始训练集拆分成 7:3 的训练集和验证集，设置迭代次数为 50，树深度为 7，学习速率为 0.15。CatBoost 训练过程如图 11 所示。横轴是训练轮次，纵轴是 RMSE 数值，最终得到的 RMSE 是 691.68。

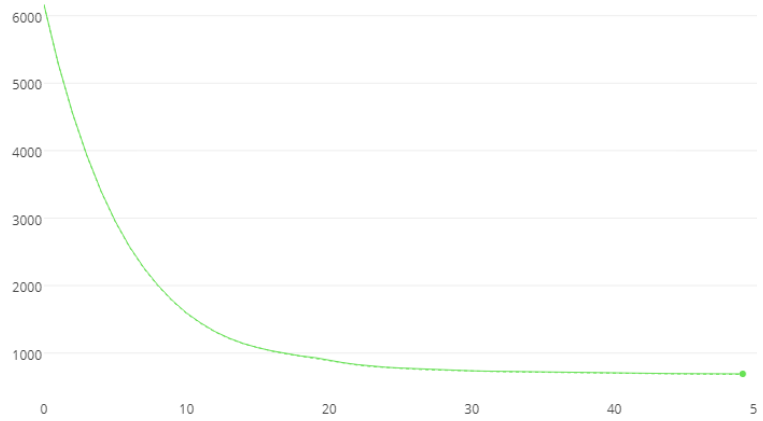

图 11：CatBoost 训练过程中 RMSE 变化图

用该模型在整个训练集上进行四折交叉验证后发现：平均来看，模型 R^2 高达 88.18%，有较好的拟合能力，RMSE 为 688.63。

接下来查看各个变量的重要性。如图 12 所示，对获赔金额影响最大的是房屋面积，和之前的猜想基本吻合。其次是家庭成员年龄极差，年龄极差与获赔金额的相关系数为 0.35，呈正相关，说明如果申请者“上有老下有小”，获赔金额会较多。房屋类型对模型的贡献度也较大。

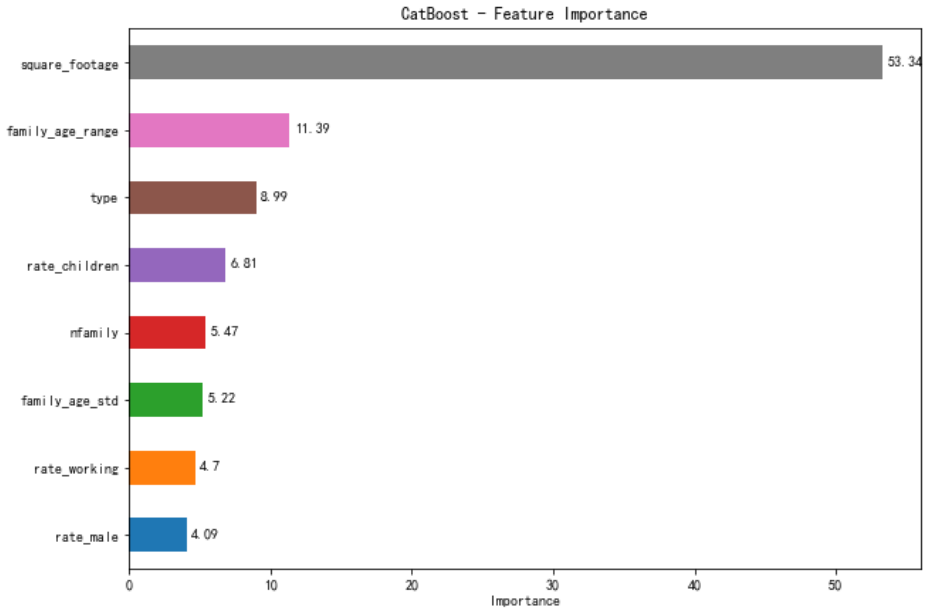

图 12：变量重要性柱形图

五、结论与建议

本案例对 Yuumi 保险公司提供的家庭财产险用户的日志数据进行了统计分

析，得到了如下结论：

- 本案例中的诈骗者绝大多数是机器人，其行为特点是支付保费的时间与申请索赔的时间完全相同，公司可利用此特征进行欺诈识别。
- 在定价策略上，本案例提供了一种以 CatBoost 算法为基础的回归预测模型，在此模型中，房屋面积、家庭成员年龄极差和房屋类型对获赔金额影响最大，家中孩子数量比例、劳动力人口比例、男性比例和家人年龄方差也对获赔金额有一定影响。

本模型在测试集上的 RMSE 是 3401.23，而随机预测情况下的测试集 RMSE 是 7935.96，模型一定程度上可以识别出欺诈者和给出合理的定价。但是相比较第一名 RMSE 为 1386.47 的成绩来说，还有一定差距，反思如下：

- 测试集中是否可能存在非机器人的欺诈者？在训练集中，10 万客户里人类欺诈者只有 3 个，因此无法建立分类模型对这部分欺诈者进行识别。训练集和测试集的样本可能有不同的分布模式。
- 可以将地址信息加入模型。从房屋面积与赔付金额的散点图来看，二者呈现出整齐的簇状线关系模式，猜测可能是房屋地址对二者关系有很大影响，数据集中提供了每位客户的居住地址，可以利用谷歌地图找到每个地址的经纬度，并将用户住址分类为澳大利亚的不同地区，如中部、沿海、北部等，将地理位置变量加入模型可能会进一步提升预测效果。
- 添加时间序列变量。时间戳信息未被充分利用，只提取了支付时间与申请时间差作为行为特征，而欺诈者可能在业务流程的整个时间段上都表现出了异常特征。公榜第一名将反欺诈业务中较流行的时间序列特征加入模型后 RMSE 降低了 1000 多。
- 训练集中对邮箱进行了脱敏处理，因此无法利用这部分信息识别欺诈者。实际场景中，欺诈者会使用随机生成的数字和字母作为邮箱地址的一部分，这部分信息也可以为反欺诈做出贡献。
- 尝试采用更强的深度学习模型或进行模型集成。第二名使用了深度学习模型将测试集 RMSE 降到了 1434.14，第三名使用了线性模型、LightGBM 和 GBM 进行 Stacking 将 RMSE 降低至 1491.54。使用较好的融合策略可能对于降低损失有所助益。