# 保险反欺诈预测项目报告

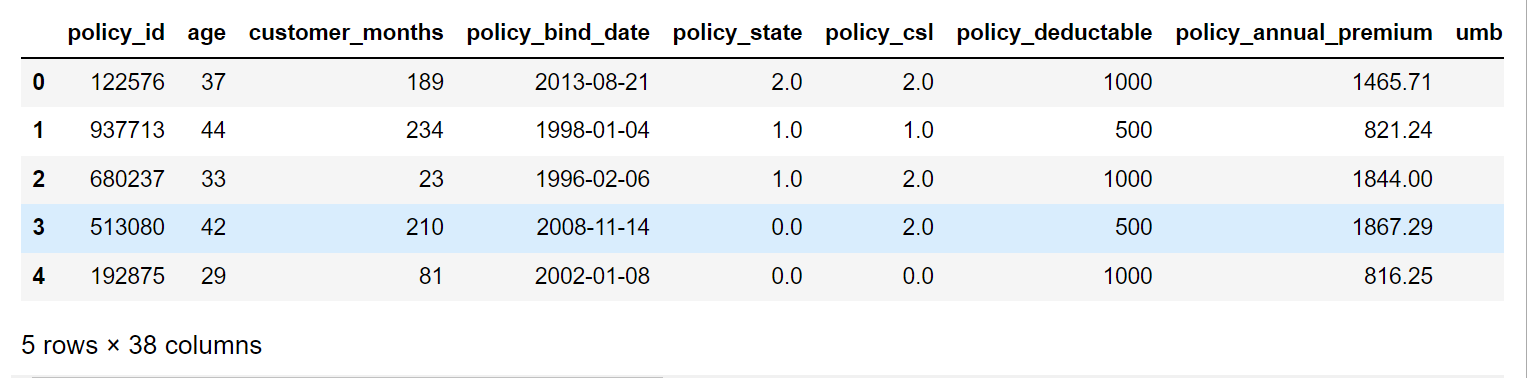
1. **任务意义与背景**

项目以保险风控为背景，保险是重要的金融体系，对社会发展，民生保障起到重要作用。保险欺诈近些年层出不穷，在某些险种上保险欺诈的金额已经占到了理赔金额的20%甚至更多。对保险欺诈的识别成为保险行业中的关键应用场景。

本次项目来自于天池网站金融数据分析赛题，地址链接https://tianchi.aliyun.com/competition/entrance/531994/information。

在数据集中提供了之前客户索赔的车险数据，需要选手开发模型预测哪些索赔是欺诈行为，结果将以AUC作为评判标准。

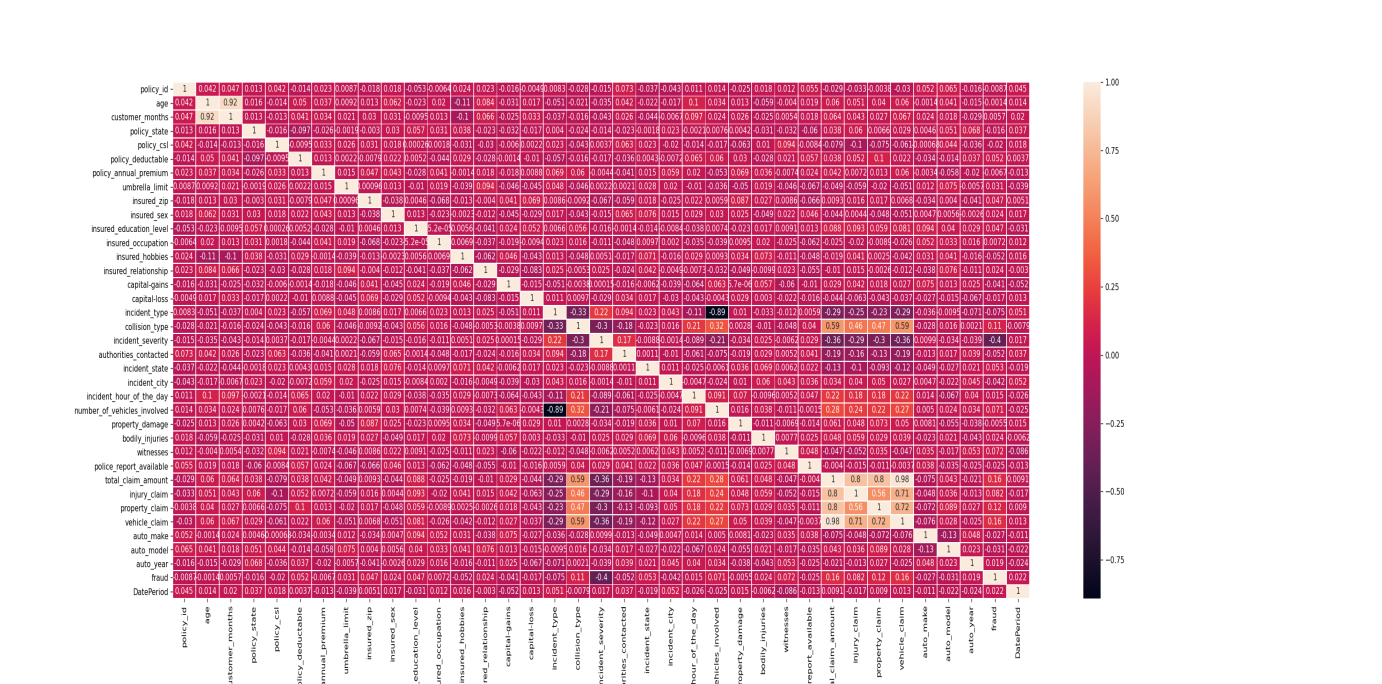
1. **现有方案调研**
2. 查找机器学习特征工程相关方法和库，从中选择最适合项目的方法。《总结：sklearn机器学习之特征工程》https://www.jianshu.com/p/1c4ec02dd33f
3. 参考机器学习完整案例，安排本项目流程。
4. 本次重点选择了房价预测和良/恶性性乳腺癌肿瘤预测这两个经典案例，并参考了多篇博客：
5. https://blog.csdn.net/qq\_43965708/article/details/116483085，[参考了其时间长度处理和相关性分析的思路；](https://blog.csdn.net/qq_43965708/article/details/116483085，参考了其时间长度处理和相关性分析的思路；)
6. <https://blog.csdn.net/qq_43965708/article/details/116483085，参考了其数据分析的思路；>
7. [https://blog.csdn.net/Yann\_YU/article/details/122207795，参考了其处理标签数量不平衡的方法，](https://blog.csdn.net/Yann_YU/article/details/122207795，参考了其处理标签数量不平衡的方法。)与多模型的使用；
8. <https://blog.csdn.net/duanlianvip/article/details/102398886，参考了其GBDT调参细化方法；>
9. 查阅了该项目其余选手的案例分享，了解到在AUC评分的机制下，提交预测概率可以获得更好的结果。
10. **项目方法**
11. 特征工程与数据处理
12. 数据集中有许多样本中含有缺失值（以“？”代表），经汇总发现占比较高，因此在标签化处理时，将其看做一种类别；
13. 对分类变量使用OrdinalEncoder进行标签化处理，编码为数字；
14. 对数据集中两个日期变量：保险绑定日期和出险日期，将两者之差——险期长度作为新的变量，再将两者从数据中剔除；
15. 分析相关性矩阵，对于高度相关的变量，只保留两者之一即可；
16. 查看因变量与自变量的相关性排行以及因变量标签个数；
17. 由于正负标签数量相差较大，使用oversampling方法增加正标签；
18. 对训练集和测试集均重复上述步骤，得到最终使用的数据。
19. 模型拟合
20. 使用逻辑回归、支持向量机、KNN、随机森林、GBDT、神经网络等方法进行拟合，然后比较其效果，在拟合过程中均使用GridSearchCV进行调参；
21. 因为测试集中标签未给出，因此应该选择上述模型中在训练集内效果最好的几个，将其对测试集的预测上传网站进行评分；
22. 模型评估采用ACC、AUC和AP共用；
23. 对于评分最高的模型，对其进行更加精细的调参，查看评分提升效果，找到最佳参数。
24. **实验过程**
25. 经过数据处理和特征工程，分类变量处理为数值编码：



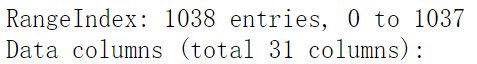
'policy\_bind\_date','incident\_date'删去，增加了'DatePeriod'：



根据相关系数矩阵，删去了'incident\_type','property\_claim','vehicle\_claim','injury\_claim'：



Oversampling平衡标签数量后，训练集样本数增加至1038条，共31个自变量：



1. 初步拟合各模型，得到模型评价结果。其中逻辑回归、SVM和神经网络效果较差，而KNN和随机森林虽然ACC和AUC较高，但是由于其模型的特殊性，对于训练集上的结果并不能很好的反映测试集效果；
2. 将KNN、随机森林和GBDT的测试集预测结果提交，评分依次为：0.5834、0.8030、0.8651，GBDT最好；再将GBDT结果以概率形式上传，得到评分为0.9438。
3. 对GBDT模型实行精细化调参，调参步骤如下：
4. 设置较高的learning\_rate，调试迭代次数：n\_estimators；

2） 固定learning\_rate，选取最优的n\_estimators，调试max\_depth和min\_samples\_split；

3） 固定learning\_rate，选取最优的n\_estimators、max\_depth、min\_samples\_split，调试min\_samples\_leaf；

4） 固定learning\_rate，选取最优的n\_estimators、max\_depth、min\_samples\_split、min\_samples\_leaf，调试max\_features；

5） 固定learning\_rate，选取最优的n\_estimators、max\_depth、min\_samples\_split、min\_samples\_leaf、max\_features，调试subsample，并选取最优的subsample；

6） 调试learning\_rate和n\_estimators：减少learning\_rate，成比例的增加n\_estimators；

7） 训练使用最优超参的模型，并对模型进行评估。

1. 截止第5步，最优参数为：learning\_rate=0.2, n\_estimators=250, max\_depth=8, min\_samples\_split=10,min\_samples\_leaf=5,random\_state=10,max\_features=6
2. 接着调节learning\_rate和n\_estimators，learning\_rate依次减少为0.1、0.05、0.01，n\_estimators对应成比例增加，得到的评分依次为：0.9505、0.9548、0.9561，可见效果提升已经不明显了，因此将0.9561作为最终项目评分。
3. **总结和展望**
4. 实践是最好的学习方法，在项目进项过程中查阅了大量资料和案例，学习到了更多知识，对机器学习实践更加熟练了；
5. 本次使用了新的数据处理库imblearn，但是机器学习主体还是用的sklearn库，以后应当尝试使用其他库，在某些算法上可能有更好的实现；
6. 在特征工程上可以尝试其他的方法，比如数据标准化和降维等方法，进一步提升训练效果；