保险行业数据可视化分析

制作人： 吕艾（CDA持证人、会员）

联系方式：IVYLUE@163.COM

WECHAT ID: IVYLUE

数据集介绍

- 保险数据（见附件）
- 数据字典（见附件）

5 rows x 50 columns pd.DataFrame

CSV

	resp_flag	GEND	AART	ADBT	ADEP	AHBP	AHCH	ARES	AHRT	ADGS	...	c210pmr	c210poo	c210psu	c210pwc	c210wht	ilor	meda	pdpe
0	0	M	N	N	N	N	N	N	N	N	...	52	65	71.0	22	79.0	15.0	64.0	42
1	0	M	N	N	N	N	N	N	N	N	...	44	81	99.0	37	65.0	17.0	61.0	46
2	0	F	N	N	N	N	N	N	N	N	...	38	44	62.0	44	47.0	20.0	61.0	46
3	0	F	N	N	N	N	N	N	N	N	...	45	71	99.0	39	71.0	4.0	62.0	37
4	0	F	N	N	N	N	N	N	N	N	...	32	13	36.0	15	65.0	9.0	NaN	37

5 rows x 4 columns pd.DataFrame

	变量名	type	变量说明	备注
0	resp_flag	Num	是否response	NaN
1	age	Num	年龄	NaN
2	GEND	Char	性别	NaN
3	c210mys	Num	学历	0-unknown；1-初中；2-高中不到；3-高中毕业；4-大学未毕业；\n5-大专；6-本科；7-研究生；8-专业院校毕
4	NOC19	Num	家庭小孩个数	NaN

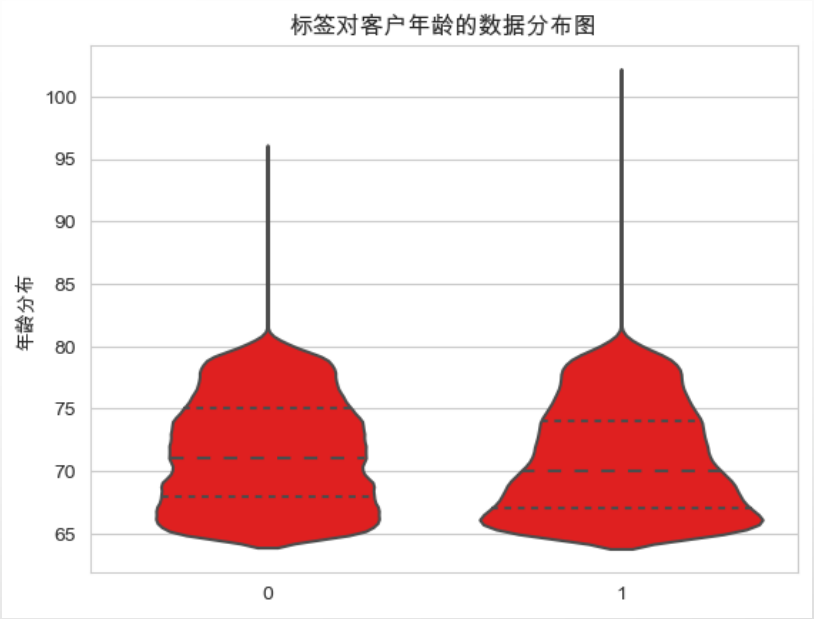
可视化工具：PYTHON

数据建模与业务分析概览

- 第一阶段：数据集的准备与处理
- 第二阶段：特征处理
- 第三阶段：建模、模型评估
- 第四阶段：可视化分析
- 总结：面板可视化总结

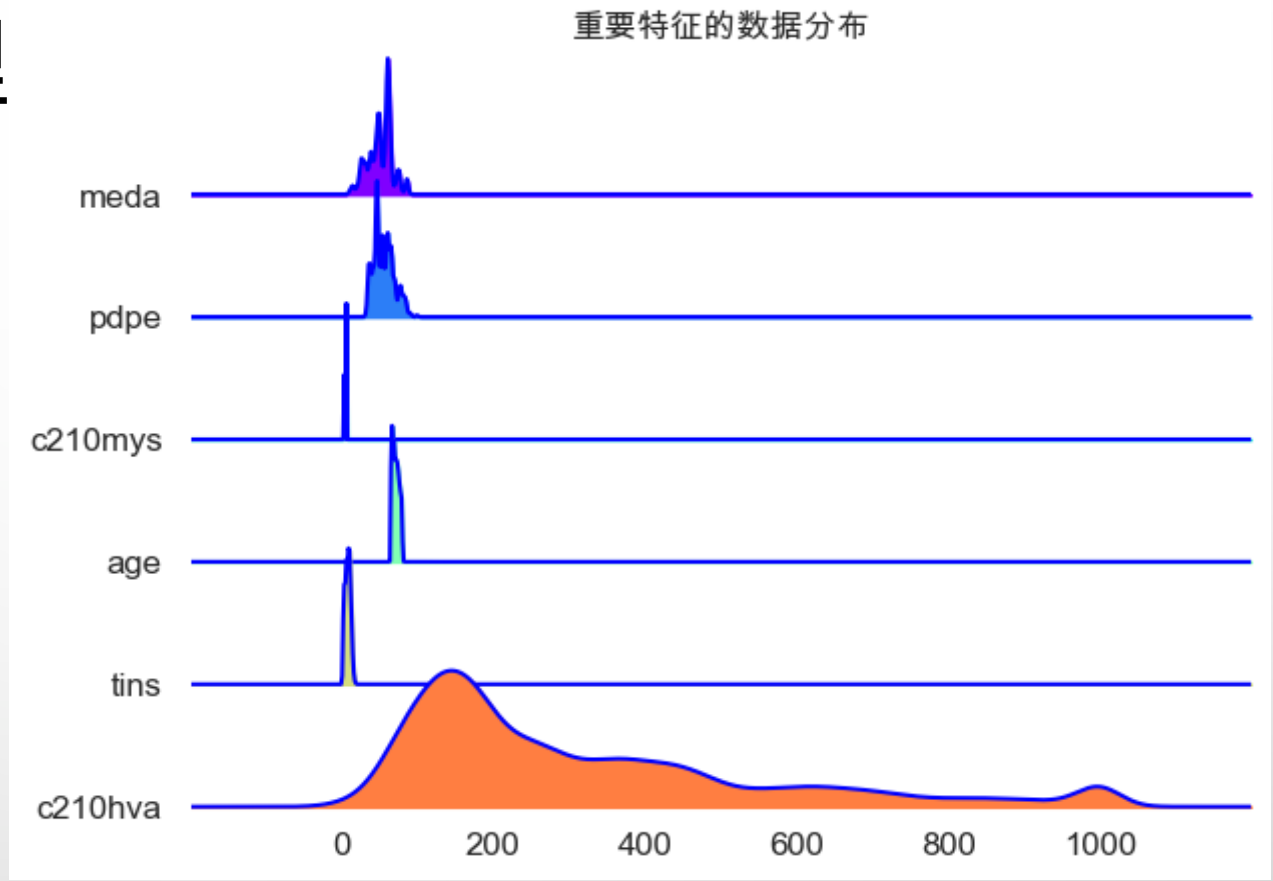
第一阶段：数据集的准备与处理

- 处理缺失值
- 查看数据类型
- 处理重复记录
- 查看标签数据的分布



第二阶段：特征处理

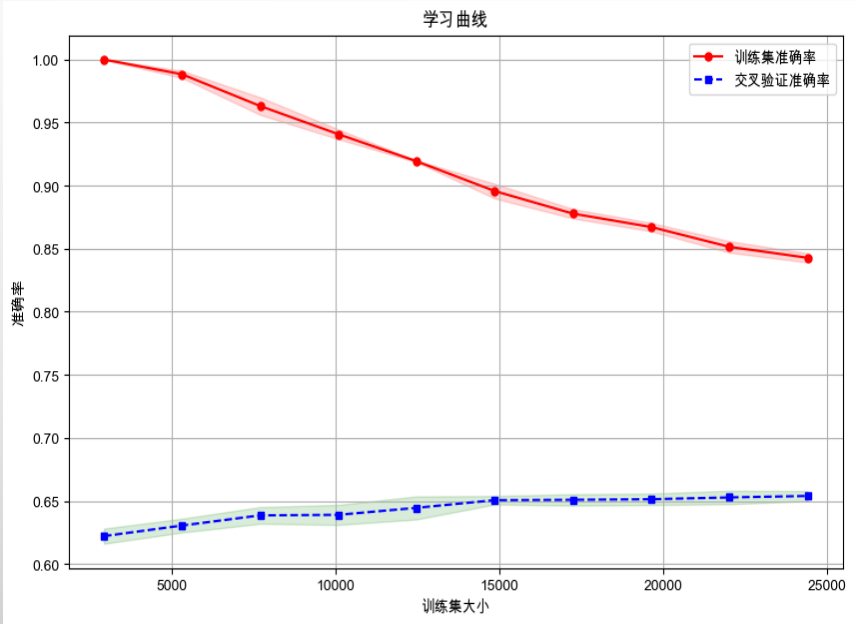
- 分类型变量做编码处理
- 数值型变量做标准化处理
- 分割测试集和训练集数据

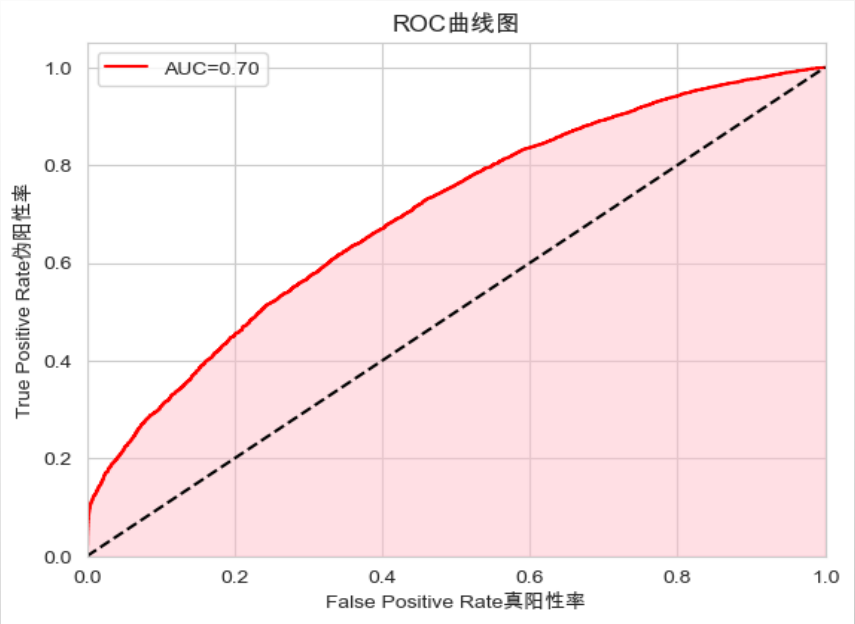


第三阶段：建模、模型评估

- XGBOOST模型
- ROC曲线图
- 学习曲线

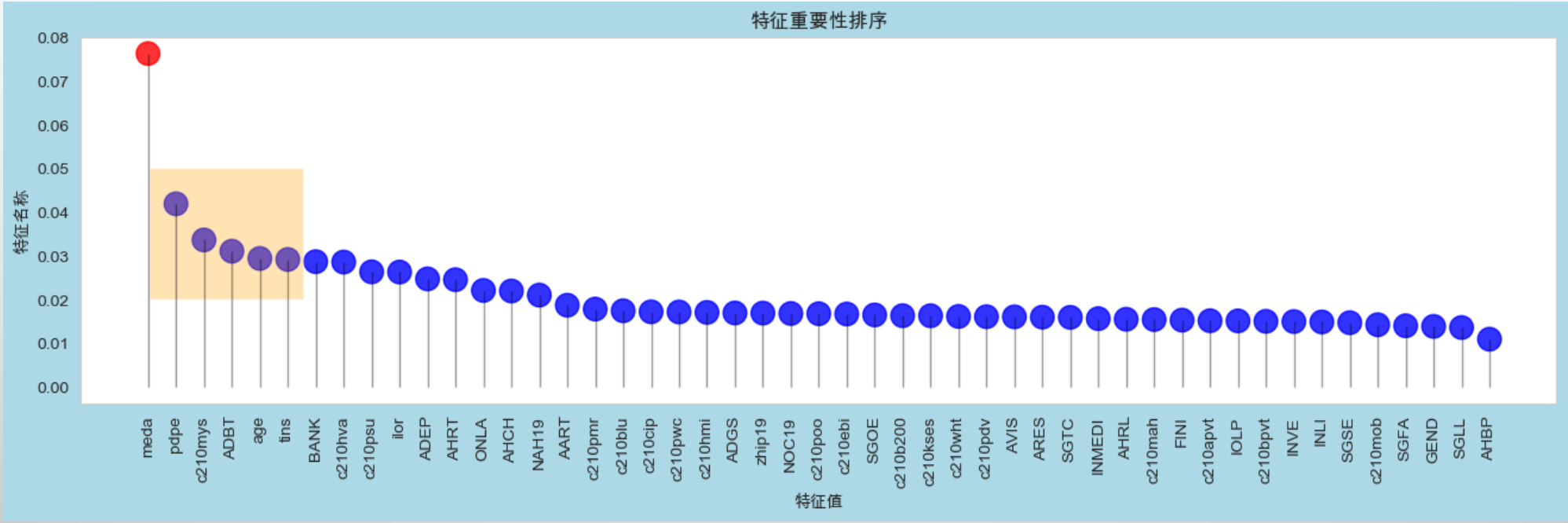
	precision	recall	f1-score	support
0	0.68	0.81	0.74	7832
1	0.61	0.45	0.51	5247
accuracy			0.66	13079
macro avg	0.65	0.63	0.63	13079
weighted avg	0.65	0.66	0.65	13079





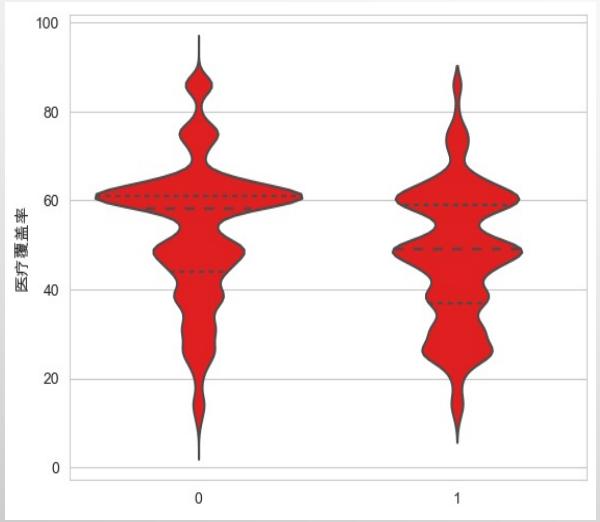
第四阶段：可视化分析

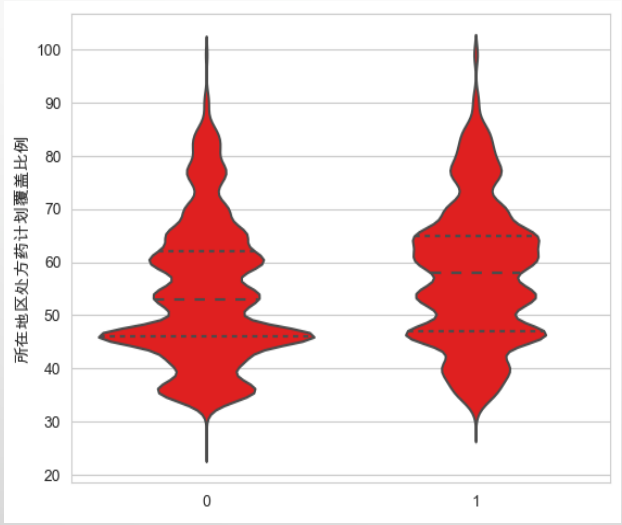
- 特征重要性的排序和选择

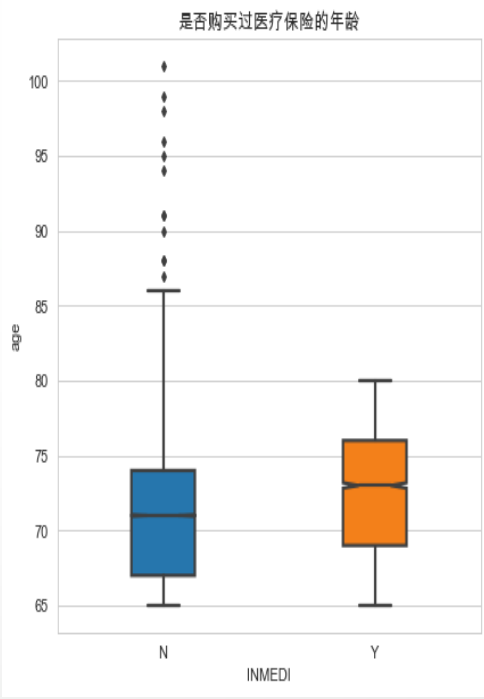


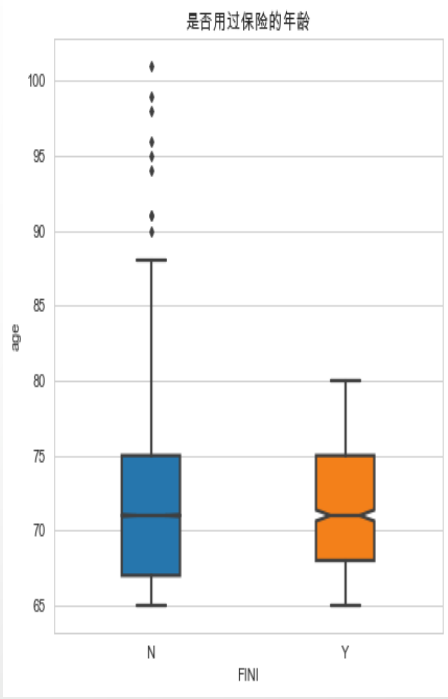
第四阶段：可视化分析

- 特征的业务分析



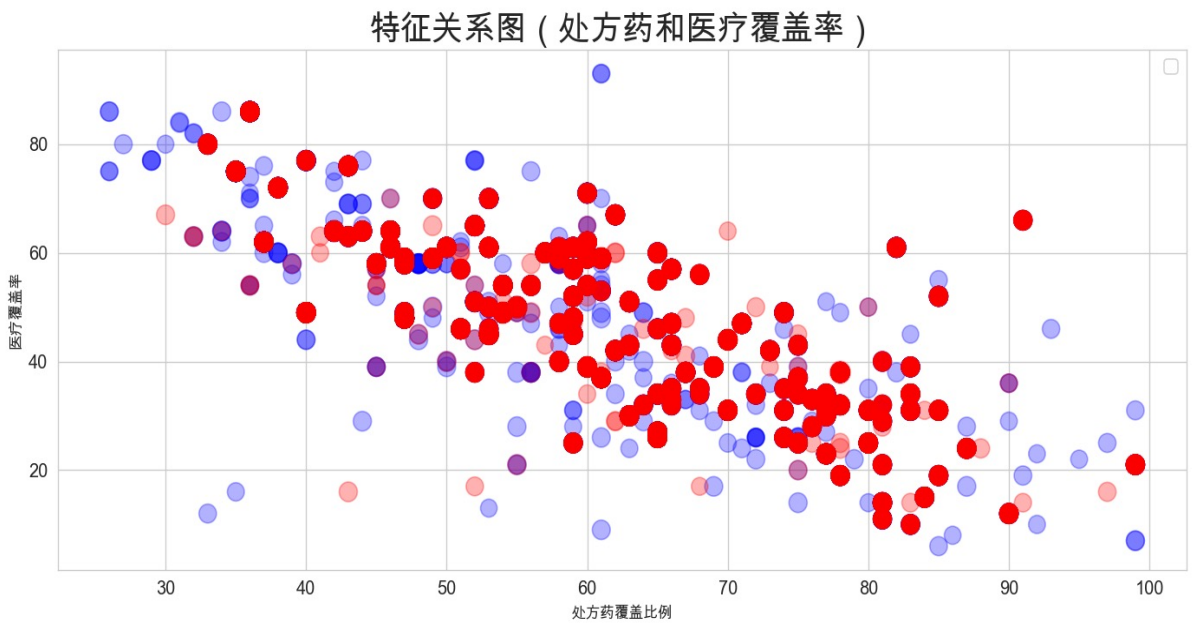


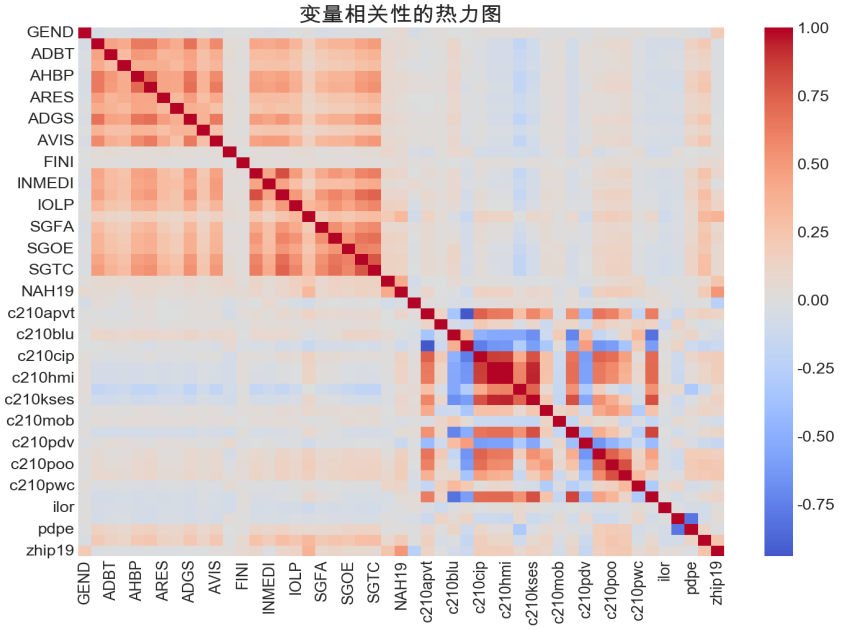




第四阶段：可视化分析

- 特征关系的分析

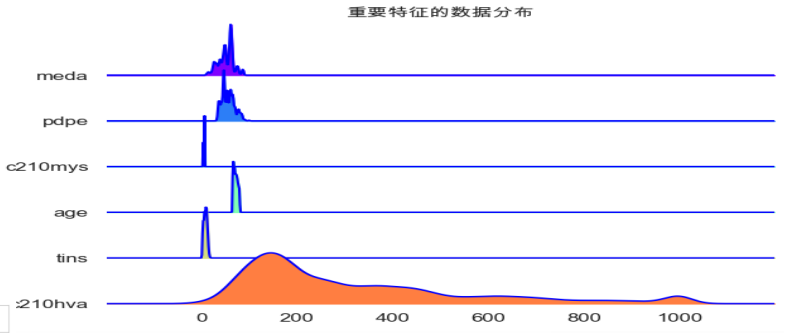




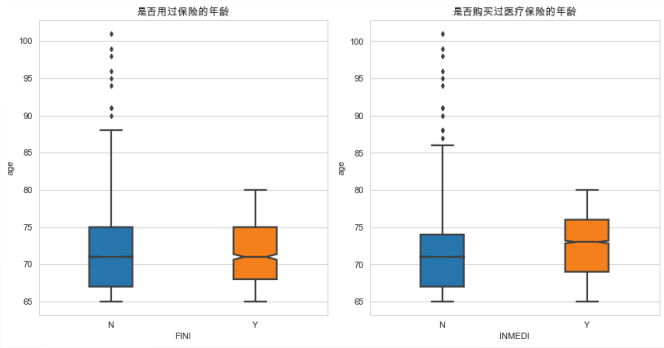
保险行业数据分析

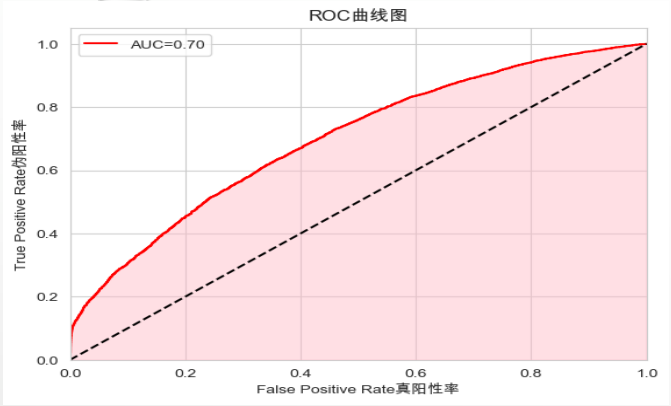
收集了四万多条记录包含个特征一列二分类的标签特征通过数据分析和可视化展示了报销销售需要重点关注医疗覆盖率处方药覆盖率不同程度的地区客户学历和年龄情况另外糖尿病和破产几率也是重要的影响因素。

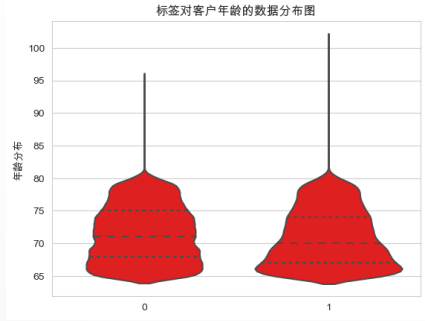
	precision	recall	f1-score	support
0	0.68	0.81	0.74	7832
1	0.61	0.45	0.51	5247
accuracy			0.66	13079
macro avg	0.65	0.63	0.63	13079
weighted avg	0.65	0.66	0.65	13079



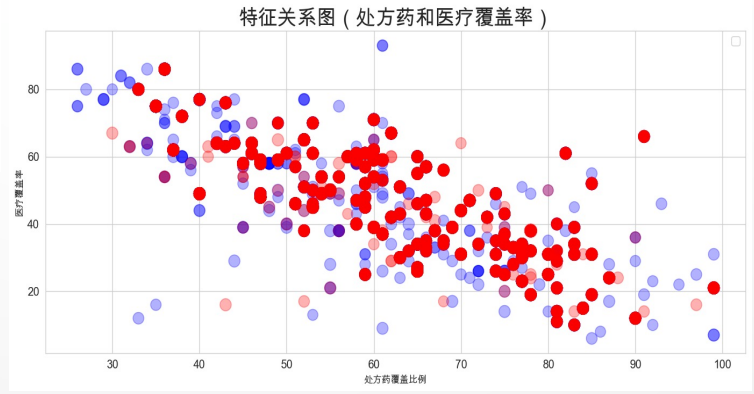
购买过医疗保险的人群年龄在 68-77 岁左右，用过保险的人群年龄主要在67-75岁范围内.





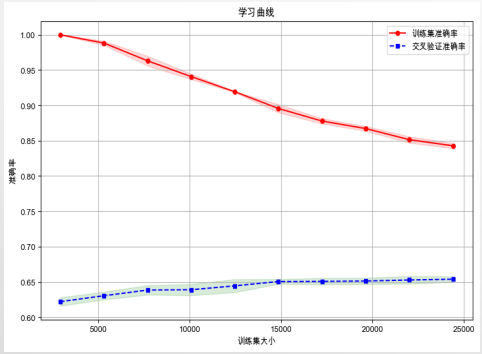


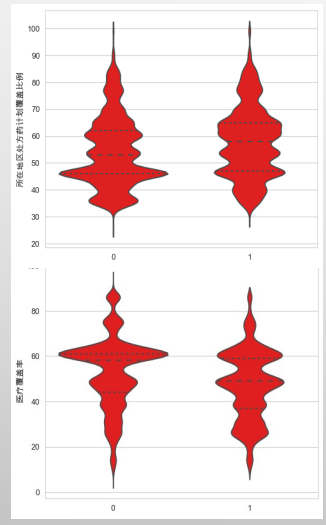
标签二分类数据分布均匀，对保险产品有反馈和没反馈的人群年龄类似。



对保险反馈最重要的特征是医疗覆盖率遥遥邻先其他因素，但是对整个业务来说“处方药的覆盖率”、“学历”、“是否有糖尿病”、“是否有过过破产记录”等等都是值得业务人员关注的因素。

模型评估指标AUC=0.7，因为数据经过脱敏处理。学习曲线能够看出模型的学习能力和泛化能力在不断收敛和趋于平衡，但受制于数据的质量程度有限。





医疗覆盖率在 60% 左右对保险产品的反馈并不积极，处方药覆盖率45%的时候相对减少了很多对保险产品响应率。