题目背景

这是一家为客户提供健康保险的保险公司，现在他们需要你的帮助来建立一个模型来预测过去一年的投保人(客户)是否也会对公司提供的汽车保险感兴趣。

分析目标

通过已投健康保险的客户的个人以及车辆信息，来预测该客户是否对公司提供的汽车保险感兴趣。

e.g.

id	Gender	Age	Driving_License	Region_Code	Previously_Insured	Vehicle_Age	Vehicle_Damage	Annual_Premium	Policy_Sales_Channel	Vintage	Response
1	Male	44	1	28.0	0	> 2 Years	Yes	40454.0	26.0	217	1
2	Male	76	1	3.0	0	1-2 Year	No	33536.0	26.0	183	0

数据字段描述

Variable	Definition
id	Unique ID for the customer
Gender	Gender of the customer
Age	Age of the customer
Driving_License	0 : Customer does not have DL, 1 : Customer already has DL
Region_Code	Unique code for the region of the customer
Previously_Insured	1 : Customer already has Vehicle Insurance, 0 : Customer doesn't have Vehicle Insurance
Vehicle_Age	Age of the Vehicle
Vehicle_Damage	1 : Customer got his/her vehicle damaged in the past. 0 : Customer didn't get his/her vehicle damaged in the past.
Annual_Premium	The amount customer needs to pay as premium in the year
Policy*Sales*Channel	Anonymized Code for the channel of outreaching to the customer ie. Different Agents, Over Mail, Over Phone, In Person, etc.
Vintage	Number of Days, Customer has been associated with the company
Response	1 : Customer is interested, 0 : Customer is not interested

Response为label，即在测试集中需要预测的部分。

训练集：304888条

测试集：76221条

评价指标

AUC（Area Under Curve）被定义为ROC曲线下的面积。

作业要求

本次作业两至三人组队完成，推荐两人组队

提交的最后文件内容为

- 最终代码文件（请写清楚使用了那些库，以及相应库的版本，可使用 pip list 命令查看版本，确保能顺利运 行）
- test.csv数据集上的结果，文件名为“results.csv"
- 分析文档
  - 请不要是简单的代码粘贴，加入分析过程
  - 将你对于数据的理解记录下来，简单来说，缺失值处理，特征之间的相关性分析，数据的预处理过程
  - 写出你的尝试的各种方法，为了解决rank太低的情况下分数太低
- 写一份同学负责哪一部分代码，每一部分没有区别，主要是为了给代码风格打分。
- 整个文件包名为同学1学号 - 同学2学号 （-同学3学号）.zip(tar)

评分细则

50%：Rank分数

30%：文档评分

10%：现场答辩（可选、自愿）

10%：个人代码风格评分

Rank	score
1	100
2~3	95
4~...	90按照位次递减1

提交结果文件格式

- 结果文件名为“results.csv"
- 提交格式
  - 第一行为 test_id \t responce 的表头
  - 接下来的每行为id \t responce_value

id	Response
0	0.9802
1	0.1405

输出预测值为0-1之间的值