数据科学导论课程大作业

2022.5.31

题目介绍

背景介绍——基于客户行为的贷款预测

根据贷款app提供的数据，预测用户是否存在贷款违约的可能。

数据样例

Id	Income	Age	Experience	Married/Single	House_Ownership	Car_Ownership	Profession	CITY	STATE	CURRENT_JOB_YRS	CURRENT_HOUSE_YRS	Risk_Flag
0	1392884	29	11	single	rented	yes	Scientist	Varanasi	Uttar_Pradesh	5	14	0
1	2800634	0	12	single	rented	no	Web_designer	Agra	Uttar_Pradesh	8	13	0
2	4121059	52	17	single	rented	no	Aviator	Saharanpur	Uttar_Pradesh	5	10	0

训练集：201600
测试集：50400 (w.o. Risk_Flag)

题目介绍

各字段含义

Column	Description	Type
income	Income of the user	int
age	Age of the user	int
experience	Professional experience of the user in years	int
profession	Profession	string
married	Whether married or single	string
house_ownership	Owned or rented or neither	string
car_ownership	Does the person own a car	string
risk_flag	Defaulted on a loan	string
current*job*years	Years of experience in the current job	int
current*house*years	Number of years in the current residence	int
city	City of residence	string
state	State of residence	string

题目介绍

评价指标

AUC(Area Under Curve): ROC曲线下与坐标轴围成的面积。AUC的取值范围在0.5和1之间。AUC越 接近1，检测方法的真实性越高。

```
from sklearn.metrics import roc_curve, auc
fpr, tpr, thresholds = roc_curve()
roc_auc = auc(fpr,tpr)
```

作业要求

分组要求：

两人一组；落单可三人一组。周末将组队结果发给学委，可以跨班组队。

提交方式：

后面将开通网页提交通道，时间再通知，到时候会具体说。

文件格式：

每行对应对测试集的预测结果，.csv文件。（输出值在0-1之间）

Risk_Flag
0.131
0.973
0.546

最终完成作业

提交内容：

代码 作业报告 最终预测结果文件 排行榜结果排名截图 组内成员分工明细

{报告包含:问题分析，数据处理流程，选取何种方法以及原因(若选取多种方法说明对比)}
{最终预测结果要与排行榜的值对应，建议大家每次提交时保留原来文件}

命名格式：

压缩包以组号命名即可，例如”1组.zip “

评分标准：

综合考虑代码、报告质量和排名

截止时间：

再通知