

Project

天池大数据竞赛

打造国际高端算法竞赛，让选手用算法解决社会或业务问题

Active

算法大赛

创新应用大赛

程序设计大赛

学习赛

可视化大赛

诸神之战

零基础入门数据挖掘-心跳信号分类预测

新人赛

赛事简要：赛题以心电图心跳信号数据为背景，要求选手根据心电图感应数据预测心跳信号所属类别，其中心跳信号对应正常病例以及受不同心律不齐和心肌梗塞影响的病例，这是一个多分类的问题...

主办方:  Datawhale TIANCHI天池

奖金

¥ 0

团队

1348

赛季 1


2021-05-12

进行中

零基础入门金融风控-贷款违约预测

新人赛

赛事简要：赛题以金融风控中的个人信贷为背景，要求选手根据贷款申请人的数据信息预测其是否有违约的可能，以此判断是否通过此项贷款，这是一个典型的分类问题。

主办方:  Datawhale TIANCHI天池

奖金

¥ 0

团队

6230

赛季 2

2022-12-31

进行中



《机器学习与模式识别》课程

大作业

项目说明

- 每人从[心跳信号分类预测](#)、[贷款违约预测](#)和华为《机器学习与模式识别》课程大作业中任选一题作为课程 Project；前两个选题的具体描述、数据集、评测标准请在相关链接中查看，华为大作业的相关信息请在上传的文件压缩包中查看。
- 时间安排及提交内容：
 - **6.3 24:00 前** 提交展示所用 PPT
 - **6.4 第 4 大节** 实验课用于项目展示
 - **6.11 24:00 前** 提交实验报告及源码
 - 注：提交方式同为山大网盘，实验报告及源码压缩到一个文件
- 评估：
 - 采用任意方法完成项目并得到结果为基本要求
 - 问题分析是否清晰、数据处理是否合理；方法选用的多样性、创新性、逻辑性等
 - 分值占比：项目展示 80%、实验报告 20%

注意事项

- 请将选题信息、账号信息、比赛分数和排名填到统计表中
<https://docs.qq.com/sheet/DZVhjbUZmS2ZKcWd0>
- 为了保证其他人的展示时间，项目展示**每人限时 5 分钟**，超时者酌情扣分。要求重点展示自己所使用方法的创新性及其有效性、在比赛过程中所做的尝试、遇到的问题及解决方法等。阿里天池的两个选题在比赛平台论坛模块均有开源 Baseline 供参考，华为大作业无 Baseline 供参考开放性较强。
- 本课程 Project 旨在让大家了解机器学习项目的流程，熟悉本学期所学的各种机器学习方法，并将其用于真实的问题场景中。鼓励尝试各种不同的方法，包括课外拓展的方法。需要注意的是需要清楚所选方法运作原理。**排行榜的排名不作为本项目评分的主要参考依据**
- 本课程 Project 以**个人形式**进行，允许相互交流，但注意使用相同方法会在项目展示环节表现相对平庸。**禁止抄袭或提交同一份结果，禁止抄袭或提交同一份结果，禁止抄袭或提交同一份结果**

赛题背景

赛题以医疗数据挖掘为背景，要求选手使用提供的心跳信号传感器数据训练模型并完成不同心跳信号的分类的任务。为了更好的引导大家入门，还特别为本赛题定制了学习方案，其中包括数据科学库、通用流程和baseline方案学习三部分。

通过对本方案的完整学习，可以帮助掌握数据竞赛基本技能。同时我们也将提供专属的视频直播学习通道。

一、赛题数据

赛题以预测心电图心跳信号类别为任务，数据集报名后可见并可下载，该数据来自某平台心电图数据记录，总数据量超过20万，主要为1列心跳信号序列数据，其中每个样本的信号序列采样频次一致，长度相等。为了保证比赛的公平性，将会从中抽取10万条作为训练集，2万条作为测试集A，2万条作为测试集B，同时会对心跳信号类别（label）信息进行脱敏。

字段表

Field	Description
id	为心跳信号分配的唯一标识
heartbeat_signals	心跳信号序列
label	心跳信号类别（0、1、2、3）

<https://tianchi.aliyun.com/competition/entrance/531883/introduction>

一、赛题背景

本次新人赛是Datawhale与天池联合发起的0基础入门系列赛事第四场 —— 零基础入门金融风控-贷款违约预测。

赛题以金融风控中的个人信贷为背景，要求选手根据贷款申请人的数据信息预测其是否有违约的可能，以此判断是否通过此项贷款，这是一个典型的分类问题。通过这道赛题来引导大家了解金融风控中的一些业务背景，解决实际问题，帮助竞赛新人进行自我练习、自我提高。

为了更好的引导大家入门，赛题方同时为本赛题定制了学习方案，其中包括数据科学库、通用流程和baseline方案学习三部分。通过对本方案的完整学习，可以帮助掌握数据竞赛基本技能。同时平台也将提供专属的视频直播学习通道，敬请关注平台通告。

一、赛题数据

赛题以预测用户贷款是否违约为任务，数据集报名后可见并可下载，该数据来自某信贷平台的贷款记录，总数据量超过120w，包含47列变量信息，其中15列为匿名变量。为了保证比赛的公平性，将会从中抽取80万条作为训练集，20万条作为测试集A，20万条作为测试集B，同时会对employmentTitle、purpose、postCode和title等信息进行脱敏。

<https://tianchi.aliyun.com/competition/entrance/531830/introduction>



《机器学习与模式识别》课程

大作业

2.1 作业背景

作为纯信用模式下的金融信贷产品，信用卡风险主要包括三个方面：信用风险、欺诈风险、操作风险。近年来，随着互联网金融的快速发展以及支付模式日益多元化，信用卡违约现象逐渐增多，不良贷款快速增长，信用卡欺诈、违法套现等违法犯罪活动不断出现，并呈现出新趋势、新特点。信用卡欺诈不仅给银行造成经济损失，还会带来巨大的声誉风险，降低客户对银行的信任度。对此，各银行加强信用卡管理，提升风险防控能力已经刻不容缓。

作业数据为模拟某银行的客户信用卡记录，挖掘数据的潜在价值，为该银行的信用卡业务决策提供参考。该银行面临的信用卡欺诈和拖欠现象比较严重，发生比例高于我国银行行业的平均值。希望通过对影响用户信用等级的主要因素进行分析，以及结合信用卡用户的人口特征属性对欺诈行为和拖欠行为的影响因素进行分析。

通过对银行的客户信用记录、申请客户信息、拖欠历史记录、消费历史记录等数据进行分析，对不同信用程度的客户进行归类，研究信用卡贷款拖欠、信用卡欺诈等问题与客户个人信息、信用卡使用信息的关系，为银行提前识别、防控信用卡业务风险提供参考，从而减少银行在信用卡业务方面的损失。

2.2 作业要求

学生可以从影响用户信用等级的主要因素进行分析，以及结合信用卡用户的人口特征属性对欺诈行为和拖欠行为的影响因素进行分析。通过对银行的客户信用记录、申请客户信息、拖欠历史记录、消费历史记录等数据进行分析，对不同信用程度的客户进行归类，研究信用卡贷款拖欠、信用卡欺诈等问题与客户的个人信息、信用卡使用信息的关系。

学生可以从四个方面建模：申请者评级模型，行为评级模型，催收评级模型，欺诈评级模型，全面分析银行信用卡信用风险。

1. **技术要求**：Python、scikit-learn、numpy、pandas、matplotlib 等；
2. **数据分析及预处理**：对原始数据进行预处理，比如数据清洗，数据标准化，数据编码等；
3. **模型构建**：对预处理后的数据进行建模，模型方法不限。
4. **模型评估及优化**：对数据进行评估，输出评估结果，并就结果进行分析，提出改进建议。
5. **数据可视化**：对数据进行可视化输出，方便客户理解。

项目说明书及数据见文件压缩包