

订阅DeepL Pro以编辑此演示文稿。  
访问[www.DeepL.com/pro](https://www.deepl.com/pro?cta=edit-document)，了解更多信息。

数据挖掘与商务智能》课程 考核项目

# 项目要求：

## 项目编码要求

结合你所抽选的题目，采用python进行数据挖掘模型的构建，小组上交的编码必须包括以下五个部分。

模块1：导入数据集

模块2：数据整理

模块3：探索性数据分析

模块4：模型开发

模块5：模型评估和完善

## 项目汇报要求

要求每组提交ppt，阐述以下几个方面： 1.

1. 项目任务
2. 数据源分析
3. 数据预处理
4. 数据探索性分析
5. 建模过程，可以采用几种模型来分析
6. 评价，对不同模型进行评估

## 成绩考核

* 编码是否完成，运行情况
* PPT制作的美观度
* PPT内容的完整度
* 汇报语言的清晰流畅
* 报告（word）--每人分工

# 项目案例以及要求

## 心脏病衰竭

### 任务：

建立一个模型来评估因心力衰竭事件而死亡的可能性。这可以用来帮助医院评估心血管疾病患者的严重程度。

### 数据：

数据集来自Davide Chicco, Giuseppe Jurman：机器学习可以仅从血清肌酐和射血分数预测心衰患者的生存率。BMC 医学信息学和决策 20, 16 (2020)

数据源：心脏衰竭临床记录数据集

## 健康保险

### 任务：

你的客户是一家为客户提供健康保险的保险公司，现在他们需要你帮助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车辆保险感兴趣。

就像医疗保险一样，有一种车辆保险，每年客户需要向保险公司支付一定数额的保险费，这样在车辆发生不幸事故的情况下，保险公司将向客户提供赔偿（称为 "保额"）。

建立一个模型来预测客户是否会对车辆保险感兴趣，对公司来说是非常有帮助的，因为它可以相应地计划其沟通策略来接触这些客户，优化其商业模式和收入。

现在，为了预测客户是否会对车辆保险感兴趣，你有关于人口统计学（性别、年龄、地区代码类型）、车辆（车龄、损坏）、政策（保费、采购渠道）等信息。

评价指标

本次黑客马拉松的评价指标是ROC\_AUC得分。

公共和私人的分割

公共排行榜是基于40%的测试数据，而最终排名将由其余60%的测试数据决定（这是私人排行榜）。

### 数据：

我们的客户是一家为客户提供健康保险的保险公司，现在他们需要你帮助建立一个模型来预测过去一年的保户（客户）是否也会对公司提供的车辆保险感兴趣。

保险单是一种安排，公司承诺为特定的损失、损害、疾病或死亡提供赔偿保证，以换取支付特定的保费。保费是客户需要定期向保险公司支付的一笔钱，以获得这种保证。

例如，你可能每年为一份20万卢比的健康保险支付5000卢比的保费，这样，如果上帝保佑，你在那一年生病并需要住院治疗，保险公司将承担高达20万卢比的住院费用等。现在如果你想知道，公司只收取5000卢比的保费，怎么能承担这么高的住院费用，这就是概率概念的作用。例如，像你一样，可能有100个客户每年支付5000卢比的保费，但其中只有少数人（比如2-3人）会在那一年住院，而不是所有人。这样一来，每个人都分担了其他人的风险。

就像医疗保险一样，有一种车辆保险，每年客户需要向保险公司支付一定数额的保险费，这样在车辆发生不幸事故的情况下，保险公司将向客户提供赔偿（称为 "保额"）。

建立一个模型来预测客户是否会对车辆保险感兴趣，对公司来说是非常有帮助的，因为它可以相应地规划其沟通策略来接触这些客户，优化其商业模式和收入。

现在，为了预测客户是否会对车辆保险感兴趣，你有关于人口统计学（性别、年龄、地区代码类型）、车辆（车龄、损坏）、政策（保费、采购渠道）等信息。

数据来源：健康保险

## 人力资源分析：数据科学家的工作变化

### 任务细节

这个数据集旨在了解导致一个人为公司工作（离开目前的工作）的因素，这个任务的目标是建立模型，使用目前的证书、人口统计学、经验来预测候选人寻找新工作或为公司工作的概率。

整个数据被分为训练和测试。此外，还提供了与测试集的注册者ID相对应的样本提交（注册者ID | 目标）。

* 数据集是不平衡的，所以如果你不处理它，可能会影响你的结果。
* 大多数特征都是分类的（名义的、顺序的、二进制的），有些是高卡度的，所以编码方法和技术将有助于提高模型的性能。
* 缺失归因策略可能会影响结果，所以它也可以成为你的管道的一部分。

### 数据

一家活跃在大数据和数据科学领域的公司希望在成功通过该公司举办的一些课程的人中招聘数据科学家。许多人报名参加他们的培训。公司想知道这些候选人中哪些是真正想在培训后为公司工作或寻找新工作的，因为**这有助于减少成本和时间，以及培训的质量或规划课程和候选人的分类**。与人口统计学、教育、经验有关的信息都在候选人的注册和登记中。

这个数据集旨在了解导致一个人离开当前工作的因素，也用于人力资源研究。通过使用当前证书、人口统计学、经验数据的模型，你将预测候选人寻找新工作或为公司工作的概率，以及解释影响员工决策的因素。

整个数据被分为训练和测试。目标不包括在测试中，但测试的目标值数据文件在手，用于相关任务。一个与测试集的enrollee\_id相对应的样本提交，也提供了列：enrollee \_id , target

**特点**

* enrollee\_id : 候选人的唯一ID
* 城市：城市代码
* city\_ development \_index : 城市的发展指数（按比例）。
* 性别：候选人的性别
* 相关经验：候选人的相关经验
* enrolled\_university：注册的大学课程类型（如果有
* education\_level：候选人的教育水平
* major\_discipline :候选人的教育专业学科
* 经验：候选人的总经验，以年为单位
* 公司规模：目前雇主公司的雇员人数
* company\_type ：当前雇主的类型
* lastnewjob：前一份工作与目前工作之间的年限差异
* training\_hours: 完成的培训时间
* 目标：0--不寻求工作变化，1--寻求工作变化
* 预测一个候选人为公司工作的概率
* 解释模型，说明哪些特征会影响候选人的决策

数据来源：数据科学家的工作变化

## 信用卡客户

### 提高预测流失客户的性能。

在这个业务问题上，我们的首要任务是识别那些正在流失的客户。即使我们将非流失的客户预测为流失的客户，也不会损害我们的业务。但是把流失的客户预测为非流失的客户就会有问题。所以召回率（TP/TP+FN）需要更高。到现在为止，我已经设法得到了62%的召回率。需要更好。

### 数据

银行的一位经理对越来越多的客户离开他们的信用卡服务感到不安。如果有人能为他们预测谁会被流失，他们会非常感激，这样他们就可以主动去找客户，为他们提供更好的服务，把客户的决定转到相反的方向。

我从一个网站上得到这个数据集，网址是https://leaps.analyttica.com/home。我已经使用这个网站有一段时间了，以获得数据集，并相应地对其进行处理，以产生丰硕的成果。该网站解释了如何解决一个特定的商业问题。

现在，这个数据集由10,000名客户组成，提到了他们的年龄、工资、婚姻状况、信用卡限额、信用卡类别等。有近18个特征。

我们只有16.07%的客户流失了。因此，要训练我们的模型来预测流失的客户是有点困难的。

数据来源：BankChurners.csv

## 房地产投资信托基金

### 项目

在这项任务中，你是一名数据分析师，在一个房地产投资信托基金工作。该信托基金希望开始投资于住宅房地产。你的任务是在给定一组特征的情况下确定房屋的市场价格。你将使用诸如平方英尺、卧室数量、楼层数量等属性或特征来分析和预测住房价格。实验室里提供了一个模板笔记本；你的工作是完成十个问题。模板笔记本中给出了一些问题的提示。

这个数据集包含国王县的房屋销售价格，其中包括西雅图。它包括2014年5月至2015年5月期间售出的房屋。

### 数据

**id** : 一所房子的记号

**日期**：房屋出售的日期

**价格**：价格是预测的目标

**卧室**：卧室的数量

**洗手间**：卫生间的数量

**sqft\_living**：住宅的平方英尺

**sqft\_lot**：地段的平方尺

**楼层**：房屋总楼层（层）数

**滨水区** :可以看到滨水区的房子

**查看**：已被查看

**状况** :整体状况有多好

**等级**：根据金县的分级制度，给予住房单元的总体等级。

**sqft\_above** : 除去地下室的房屋面积

**sqft\_basement**：地下室的平方尺

**yr\_built** : 建造年份

**yr\_renovated** : 房屋翻修的年份

**Zipcode**：邮政编码

**纬度**：纬度坐标

**long**：经度坐标

**sqft\_living15** : 2015年的客厅面积(意味着--一些装修) 这可能会或可能不会影响地段面积

**sqft\_lot15** : 2015年的土地面积(意味着--一些装修)

数据来源: kc\_house\_data\_NaN.csv

## 贷款列车

### 项目

在这个笔记本中，我们尝试练习本课程中所学到的所有分类算法。

我们使用Pandas库加载一个数据集，并应用以下算法，通过准确度评估方法找到适合该特定数据集的最佳算法。

现在，轮到你了，使用训练集来建立一个准确的模型。然后使用测试集来报告模型的准确性 你应该使用以下算法：

* K最近的邻居(KNN)
* 决策树
* 支持向量机
* 逻辑回归

\*\* 通知：\*\*

你可以去做预处理、特征选择、特征提取等等，来做一个更好的模型。 你应该使用scikit-learn、Scipy或Numpy库来开发分类算法。

### 数据：

这个数据集是关于过去的贷款。**Loan\_train.csv**数据集包括346个贷款已经还清或违约的客户的详细信息。它包括以下字段：

| **场地** | **描述** |
| --- | --- |
| 贷款\_状态 | 贷款是否在收账时被还清 |
| 负责人 | 贷款的基本本金数额在 |
| 条款 | 发起条款，可以是每周（7天）、双周和每月的还款时间表 |
| 有效日期 | 贷款是什么时候发起并生效的？ |
| 应付日期 | 由于它是一次性还款计划，每笔贷款都有一个单一的到期日 |
| 年龄 | 申请人的年龄 |
| 教育 | 申请人的教育 |
| 性别 | 申请人的性别 |

数据来源：贷款培训.csv

## 心脏病

我们有一个数据，根据其中的特征对病人是否有心脏病进行分类。我们将尝试使用这些数据来创建一个模型，试图预测病人是否患有这种疾病。

数据: Heartdisease.csv

1.年龄:代表患者年龄，以年为单位。

2.sex:代表患者性别，1为男性，0为女性。

3. cp:胸痛类型，具体类型未说明

4. trestbps：入院前的静息血压，单位为毫米汞柱

5. Chol:患者血清胆固醇含量，单位mg/dl

6. FBS:患者空腹血糖大于120mg/dl，1为是，0为否

7. restecg：静态心电图结果

8. thalach:达到最大的心率值

9. exang：运动诱发心绞痛，1为是，0为否。

10. oldspeak：相对于休息来说运动引起的ST段抑制

11. slope: 运动引起的ST段最高值斜率

12.CA: 通过荧光检测技术显示出来的主要血管数量（0-3）。

13. thal:1为正常，2为固定缺陷，3为可逆缺陷。

14.目标：0或1为是否患病