旧金山罪案类型预测

项目背景

1. 简要介绍项目所在领域的背景信息，要包含项目相关的历史信息。
2. 要清晰的描述该问题为什么应该被解决和为什么能够被解决。
3. 引用相关的学术研究
4. 陈述个人研究该问题的动机（鼓励写但不强求）

项目是Kaggle上的一个竞赛，该竞赛的目的是利用机器学习方法预测旧金山可能发生的犯罪类型。

1849年，随着加州淘金热的浪潮，旧金山经历了第一次繁荣，在随后的的几十年经历了城市面积激增和人口爆炸。人口的爆炸不可避免的带来了社会问题和高犯罪率。当然红灯区的存在也是一个很重要的原因[1]。

以前，旧金山因将一些罪行严重的罪犯关在恶魔岛而声名狼藉。而今因为很多高科技公司的存在，科技标签已经远远超过了其它标签的影响力。但是因为财富不均等、房屋短缺等因素，这里并不缺少犯罪。

项目采集了旧金山所有社区近12年的犯罪报告，时间跨度为2013-01-01到2015-05-13。犯罪报告中包含犯罪时间、犯罪类型、犯罪地点等信息。

项目需要使用机器学习的方法实现两个目标：1. 能够根据事件发生的时间和地点，预测犯罪事件的类型。2. 预测指定犯罪类型的犯罪率。对于目标一有很多成熟的分类算法可以使用，比如KNN、朴素贝叶斯、决策树、随机森林等等。而对于目标二，使用时间序列模型解决，比如VAR模型等。John Cherian and Mitchell Dawson (2015)

有人的地方就有可能有犯罪，因此通过机器学习帮助我们加深对犯罪发生模式的理解是对社会建设很有意义的一件事。更关键的是，对犯罪类型和犯罪率的预测能够帮助警察局更加有效的分配警力和打击犯罪率激增。

问题陈述

1. 对需要解决的问题进行清晰的描述。
2. 需要对问题进行明确的定义，并且至少有一种可能解决放哪。
3. 通过可量化、可评测、可重现等方面对问题进行彻底的分析。
   可量化：问题可以用数学公式或者逻辑术语表示。
   可评测：问题可以被能够清晰观察的指标进行评测。
   可重现：可以稳定分类样本。

数据集

1. 详细描述数据集和数据集字段，比如数据集字段是怎么跟问题关联的、为什么关联。
2. 获取数据集的方式和数据集的特征应该在必要时和相关文献的引用一并提供在该段落。
3. 应该清楚如何在项目中使用数据集。
4. 根据问题的上下文，考虑数据集的使用是否合理。

解决方案陈述

1. 清晰描述一种适合该问题并且适合该数据集的解决方案。
2. 通过可量化、可评测、可重现等方面对问题进行彻底的分析。（同问题描述）

基准模型

1. 需要提供相关问题领域解决方案的基准模型。
2. 理想情况下，学生的基准模型包括已有方法的描述，该领域内已知的信息，使得学生的解决方案可以和该基准模型作客观的对比。
3. 该基准模型定义清晰且可衡量。

评估标准

1. 需要提出了一个用于量化基准模型和解决方案的评估标准。
2. 这个评估标准对于问题本身、数据集以及解决方案来说都是合适的。
3. 描述评估指标是怎么推导出来的，提供数学表达式。

项目设计

1. 总结解决方案实施的流程。
2. 策略探讨的过程。
3. 数据预处理过程。
4. 算法选择的过程。
5. 建议用一些可视化、伪代码、图表来辅助描述项目设计。

1. sadfadfsdfasdf ↩