模型选择

预测谁会偿还贷款

- 银行雇你来调查谁会偿还贷款
  - 你得到了 100 个申请人的信息
  - 其中五个人在 3 年内违约了

惊讶的发现

- 你发现所有的 5 个人在面试的时候都穿了蓝色衬衫
- 你的模型也发现了这个强信号
- 这会有什么问题?

Image credit: rumble.com

训练误差和泛化误差

- 训练误差：模型在训练数据上的误差
- 泛化误差：模型在新数据上的误差
- 例子：根据摸考成绩来预测未来考试分数
  - 在过去的考试中表现很好（训练误差）不代表未来考试一定会好（泛化误差）
  - 学生 A 通过背书在摸考中拿到很好成绩
  - 学生 B 知道答案后面的原因

验证数据集和测试数据集

- 验证数据集：一个用来评估模型好坏的数据集
  - 例如拿出 50% 的训练数据
  - 不要跟训练数据混在一起（常犯错误）
- 测试数据集：只用一次的数据集。例如
  - 未来的考试
  - 我出价的房子的实际成交价
  - 用在 Kaggle 私有排行榜中的数据集

K-则交叉验证

- 在没有足够多数据时使用（这是常态）
- 算法：
  - 将训练数据分割成 K 块
  - For i = 1, . . ., K
    - 使用第 i 块作为验证数据集，其余的作为训练数据集
  - 报告 K 个验证集误差的平均
- 常用：K = 5 或 10

总结

- 训练数据集：训练模型参数
- 验证数据集：选择模型超参数
- 非大数据集上通常使用 k-折交叉验证