1. 简单介绍下自己

1.2011-2013服役

2.2014年-2016年底在新华文轩做爬虫工程师.抓取数据和搭建hadoop大数据存储平台.

3.2017至今,在公平价做机器学习工程师,目前是两个项目的负责人.一个是国内的二手车估值系统(当前正在做第五版的升级),另外一个是美国的二手车估值系统(当前正在做第一版的开发).

1. 介绍下做过的项目.

国内二手车估值项目为例.该系统的目标是通过选择款型,上牌年份,城市,公里数等4个特征预测二手车收购价,个人交易价,车商零售价等10类价格的合理区间.

1. 我刚接手该项目时,发现有很多问题.我们的训练数据来源主要是瓜子,人人车,58同城等平台,老系统做的比较简单,做了一些简单的数据清洗,和一些人工干预,就直接使用random forest算法进行训练预测.我接下来从两方面工作环节来解释我在对整个估值系统进行升级方安的选择,以及老系统的主要问题在哪儿.这两个环节一方面是市场的调研,另一方面才是机器学习系统的搭建
2. 对数据理解的先验知识.市场调研的重要性,不亚于机器学习系统的搭建.

步骤一:我们有线下实体店,有专业的评估师,去理解评估师关心的特征,以及他们评估的流程,去搜集客户(个人用户,评估师,金融放贷机构)反馈的问题.

步骤二:市场调研之后,写一个problem statement来阐释观点,整理客户关心的问题.这些问题哪些是需要我们立即解决的,哪些是滞后解决的.是否需要新增一些数据展示环节.

步骤三:验证抓取数据的可信度.对抓取的平台排名,对各平台数据采用进行加权平均.

步骤四:尝试多种算法,设计产品原型,对比不同算法的估值精度.最终敲定算法.

步骤五:编写商业报告,说明估值系统为啥要大改.旧系统有啥问题,新系统原型提升精度与旧系统的对比.

1. 新估值系统的设计.

解释下系统相关的特征,目标值.

步骤一:根据之前产品原型,我们最终敲定的是重新复现xgboost算法,核心的原因是:1.计算数据空间,12.4E,实际有效数据400多万,有个经验性的说法,训练数据超过数据空间95%,算法就可以拟合出较好的结果,有效数据远远低于经验性的标准,必须得进行大量的降维处理。2.框架提供的算法,我们无法跟踪监控算法内部的实现的细节,比如决策树算法的特征测试选择.只能通过算法预测最终结果反推算法实现细节.人工复现算法,人工特征测试的条件,这样我们就可以监控每一个环节.

步骤二:清洗数据,剔除离群点.

常规清洗就没不说了.离群点或者说是异常数据的检测,通常有四类技术.1.统计方法.2.基于邻近度的检测.3.基于密度的检测.4.基于聚类技术的检测。

步骤三:降维处理

步骤四:基于算法,寻找有用的模式.

1. 解释xgboost算法的原理, 组合算法原理,cart决策树,boost原理
2. 寻找有用的模式.解释下人工评估二手车逻辑.

步骤四:得到最终模型.

1. 算法选择
2. 确认性能度量,常规的,分类任务使用精度,回归任务使用对数均方差.
3. 选择基准模型
4. 评估分类器性能,估计准确度的置信区间
5. 决定是否搜集更多的数据
6. 选择超参数.
7. 调试策略
8. 过拟合的解决办法.

不同的算法过拟合的解决办法还不同.

产生过拟合的原因:

1. 噪声导致的过拟合.2.缺乏代表性样本导致的过拟合.3.算法的多重比较导致的过拟合。.

解决办法:

以决策树为基础的算法:1.先剪枝(提前终止规则) 2.后剪枝

以深度学习为基础的算法:1.参数范数的惩罚项 2.数据集增强 3.噪声鲁帮性 4.提前终止 5.组合算法 6.dropout

1. 解释具体的算法原理
2. 关联算法 关联规则, 支持度, 置信度, 频繁项集产生, 规则的产生, apriori

评估: 可视化,基于模板的方法,主观兴趣度度量

1. 聚类算法

簇类型:明显分离的,基于原型的,基于图的,基于密度的. K均值,凝聚的层次聚类,DBSCAN.

1. Lstm RNN