机器学习工程师纳米学位

开题报告

姬建业

2018 年 7 月 4 日

项目背景：

Rossman 销售预测：

Rossman 在 7 个欧洲国家经营着 3,000 多家药店目前，Rossmann 商店经理的任务是预先提前六周预测他们的日常销售。商店销售受许多因素的影响，包括促销，竞争，学校和州假日，季节性和地方性。成千上万的个体经理根据他们独特的情况预测销售情况，结果的准确性可能会有很大差异。

项目要求预测 6 周以后德国各地的 1,115 家商店每周销售，通过历史销售数据构建回归模型，预测未来 6 周的销售量。

销售预测是一个常见的场景，在供应链管理，顾客服务和制造业领域有广泛的应用。使用过去的销量预测未来的情形存在很大的挑战，由于经济下滑，员工离职，流行物品的更新，增加的竞争等因素[1]，一个合适的销售预测模型可以帮助企业节省很多成本，以及及时采取措施来应对预测的有利或不利的结果通常的预测算法有神经网络，基于时间序列的 ARIMA，基于数据挖掘等算法。目前比较流行的是基于数据挖掘算法，通过提取特征，使用 xgboost，GBDT 等算法实现预测[2-3]。

问题描述：

项目的目的是预测未来六周的销量，通过挖掘数据的特征，构建回归模型。通过训练数据集来训练模型，通过均方根误差（RMSE）的大小来衡量模型的好坏，均方根误差越小时，模型越好，一个完美的模型理论上来讲均方根误差可以为 0。根据预测集数据来预测未来六

周的销量。

数据集和输入：

本项目的数据集来自 kaggle Rossmann store sales 项目。数据包含训练集，测试集，商店信息。

训练集和测试集数据包括商店 id，销售额，客户数量，每个周的第几天，商店是否营业，州假日（公共假日，复活节，圣诞节），学校假期。

其中连续型变量为 Sales，Customers。离散型变量 Open，Promo，StateHoliday，SchoolHoliday。日期型变量 Date。其中 date 可以拆分成年月日，StateHoliday 需要做 onehot 处理。

商店信息包括商店类型，距离最近的竞争对手商店的距离，最接近的竞争对手开放时间的大致年份和月份，商店当天是否正在运营促销，商店的持续和连续促销，促销开始的年份和月份，促销的启动的月份。

其中连续型变量为 CompetitionDistance，离散型变量为 StoreType，Assortment，Promo2 ， PromoInterval ， 日 期 型 变 量 为 CompetitionOpenSinceMonth, CompetitionOpenSinceYear, Promo2SinceWeek, Promo2SinceYear。

基准模型：

基准模型使用 GBDT,使用 sklearn 中封装的 GBDT，通过调整最大树深度，树的数量，学习率，最小分裂节点数等参数，在最终测试集上 kaggle 上得分为 0.133.

本项目的任务是达到 kaggle 测试集上满足前 10%得分。也就是在 Private Leaderboard 上的 RMSPE 要低于 0.11773

评价指标：

本次评价使用 RMSPE，尽可能的降低测试集的 RMSPE，选取 RMSPE 最低的 1 个或几个模型融合的结果。

$$RMSPE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\tilde{y}_i}{y_i}\right)^2}$$

上式中 y_i 是实际销量。y_i-hat 为销量的预测值。

项目设计：

1. 训练数据，测试数据与商店信息合并，增加数据的特征。

2. 数据预处理，包括对字符型离散变量的 Onehot 处理，日期型变量的处理，截取年月。空值填充。

3. 划分训练集测试集。将 train.csv 中最后一个月数据作为本地的验证集，通过在本地验证集上调整参数，使模型 RMSPE 达到最优。

4. 模型选测，预选 lightgbm,xgboost,两个模型。

5. 主要参数选择及调优。树深度，学习率，采样率等。

6. 模型融合降低 RMSPE。

参考文献：

[1] https://money.howstuffworks.com/sales-forecasting3.htm

[2] 郑洪源, 周良, 丁秋林. 神经网络在销售预测中的应用研究[J]. 计算机工程与应用, 2001, 37(24): 30-31.

[3] 刘莹. 基于数据挖掘的商品销售预测分析[J]. 科技通报, 2014, 30(7): 140-143.