开题报告

论文题目：Rossmann 销售预测

学　　生：朱　林

1、项目背景:

如今社会变得越来越商业化，各个公司出于自身目的，对于数据分析的需求也越来越高。最重要的目的就是能够提高决策的准确性，从而提升公司的竞争力。主要的优势有如下几种：

1、提高客户满意度，优化存货管理。保证商品供应。

2、对存货更好地管理，削减成本，提升盈利水平和竞争力；过多的挤压商品会对利润产生影响。

3、调动销售人员积极性，促使产品尽早实现销售，完成使用价值向价值的转变。

4、管理层更早的发现企业经营状况，更好的做决策；制定市场战略，计划，广告投入和促销活动奠定坚实基础；发现新的市场机会。

总之：是为了将适当的产品，在适当的时间，送达适当的地点。

而对于销售量的预测是对公司业绩最直接的预判，目前预测方法多种多样，比较成熟的主要有：主观推测、时间序列 和 机器学习三种方法。由于销售量是由一些相关信息比如：客户量、是否打折等信息直接或者间接决定的，所以销售量可以被很好的预测，从而使得问题得到解决。

从我自身来说，我们是做理财产品的销售系统，对于产品的销量预测一定的需求，

和该问题比较契合，所以这也是我感兴趣的领域之一。

目前已经有很多人基于此进行了相关研究，主要的研究报告有：

1、基于机器学习方法对销售预测的研究 -唐新春

2、销售预测模型在世纪达公司的应用研究 -何鹏

3、新田公司摩托车销售预测 -钱晓星

2、问题描述:

依据给定的各种商店参数和最终结果之间，计算出两者之间的对应关系，建立模型，并利用该模型得出测试数据与真实数据之间的差距。可以使用的参数主要有：“是否开门”、“假期”、“商店类型”、“产品类型”、“竞争对手距离”、“竞争对手开门时间”、“促销”等。由于销售量是基于以上信息来决定的，所以我们可以以此作为条件，相对准确的预测出测试数据中，商店的销售数量。主要采用的预测方法是机器学习中的监督学习算法，其中 Xgboost 方法在各种竞赛中经常获得很高的评分。问题的最终量化评价标准是测试数据的预测结果与真实结果之间的均方根误差（RMSPE）。在使用随机数的时候增加 random_state，让计算的结果可以重现。

3、输入数据:

该数据是从 Rossmann 的多家商店的日常销售产生的数据中，筛选出来的对我们预测有用的数据，基于这些数据，我们能够训练出预测销量的相关信息。

包含如下几份数据：

train.csv -训练数据

test.csv -测试数据

sample_submission.csv -最终的提交数据

store.csv -商店数据

主要参数如下：

Id -测试数据标示

Store -商店编号

Sales -出售产品数量

Customers -客户数量

Open -是否开门

StateHoliday -州假日

SchoolHoliday -学校假日

StoreType -商店类型

Assortment -商品类型

CompetitionDistance -竞争对手距离

CompetitionOpenSince[Month/Year] -竞争对手开店时间

Promo - 当天是否促销

Promo2 -持续促销

Promo2Since[Year/Week] -持续促销开始时间

PromoInterval -促销月份

这些数据中，Sales 被当做最终的 y，Customers 被处理成和 Store 相关的信息，

其他属性被当做 x 进行模型训练。

4、解决方法：

目前大多数解决方案使用的主要预测方法有：

1、主观推测法：根据负责人、专家意见推测；根据销售人员意见推测；根据顾客与客户意见推测。

2、时间序列法：指数平滑法和自回归移动模型。

3、机器学习回归算法。

我选择机器学习算法中的“xgboost”方法来建立模型进行预测。它的主要优势为：

1、正则化，减少过拟合的可能性。

2、支持并行计算，提升计算速度。

3、允许自定义优化目标和评价标准。

4、缺失值处理方式可以自定义。

5、内置交叉验证，每轮 boosting 迭代使用交叉验证。

6、可以在已有基础上继续训练，每一轮训练都可以用到上一轮的结果。

将训练出的模型，应用到 test 数据上，得到最终的结果，再使用均方根误差（RMSPE）作为误差评判标准。使得结果可以被量化的同时，也可以进行很方便的评估。

5、基准模型：

由于该问题主要是基于给定的属性集合，推断出一个连续值作为得分。因此，我选择基准模型为线性回归模型，利用 sklearn 的 LinearRegression 方法，将经过处理之后的数据输入，作为训练数据集，得到模型之后，再将模型应用到 test 数据集上，得到最终的误差值。

最终可以将基准模型得到的误差值与解决方案的误差值进行对比，就能够确定解决方案是否比基准模型表现要好了。

6、评估指标：

利用 训练耗时、预测耗时、均方根误差（RMSPE）作为衡量矩阵，来判定基准模型和解决方案模型的表现。

计算公式：

0.1*（训练耗时）+0.1*（预测耗时）+0.8*（均方根误差）=最终得分

7、设计大纲：

具体流程及理论方法如下表所示：

步骤	方法/理论
获取、组织数据	pandas
分析数据特点，并进行一定的可视化	确定训练数据集和测试数据集的特点
错误数据处理	处理 Nan 值和离群值

归一化、正态分布化	Normalization
One-hot 编码	类别属性进行独热编码
降维	PCA 主成分分析
训练模型	xgboost
调优	K 折交叉验证法
计算结果	
外部因素	1、外界因素：流行趋势，爱好的转移。 2、经济变动：政府、财经界、CPI、经济增长率。 3、同业竞争动向。 4、公司广告投入情况。