Rossmann 销售预测开题报告

项目背景

作为销售计划的中心任务之一，销售预测是指通过一定的手段和方法，对未来特定时间内全部产品或者特定产品的销售数量或销售金额的估计。在操作层面上，销售预测是在充分考虑了未来各种影响因素的基础上，结合本企业的过往销售业绩，运用一定的分析方法提出切实可行的销售目标。

影响销售的因素很多，包括有需求变化、经济变动等成分的外界因素，和有营销策略、生产状况、销售人员等成分的内部因素，但销售预测对于完善客户需求管理、指导运营、优化供应链、提高企业利润方面都具有重大促进作用，降低企业的业务计划的不确定性，因此，提出合理的销售预测一直也是人们孜孜不倦的追求，是企业辅助决策的重要工具。销售预测也几乎成为了所有商业智能（BI）的终极目标。

多年来，人们已经形成了定性分析法和定量分析法两类分析方法。其中，定量分析法中的趋势预测分析法和因果预测分析法在实际应用中也能取得一定的预测效果。他们可以基于历史数据，运用一些直观的算法，如算术平均法、指数平滑法，来进行预测，相比定性分析法，预测效果也有了一定的提高[1]。

但随着经济全球化，商业网络化的进程，市场竞争日趋激烈，业务复杂化，数据规模海量化，传统的预测方法已经越来越力不从心。预测的精确性围城销售预测的核心痛点，人们迫切希望一些性能更高，更智能的预测方法。数据挖掘[2]技术由计算机自动从大量数据集中提取隐含的、事先不知道但有潜在应用价值的信息，可用于学习复杂销售的复杂特征对于销售的影响，从而得到较好的预测效果。本课题将应用机器学习算法来实施销售预测。

问题描述

Rossmann 是欧洲的一家连锁药店，在欧洲 7 个国家拥有超过 3000 家药店。这是一个 Kaggle 比赛项目，本课题将按照项目中的说明，需要为 Rossmann 在德国的 1115 家药店做出提前 6 个星期的每日销售预测。

对于 Rossmann 的销售预测问题将是一个具体领域的销售预测问题，作为药店连锁店，具有一定的行业特征，这些将体现为数据特征。如上节所述，基于机器学习算法的数据挖掘技术会是一种可能的得到一个合适预测模型的办法。

本课题得到的预测模型将用于输出未来 6 个星期里每天的销售量，预测结果可以和实际销售情况对比，从而衡量预测效果。在实际使用中，还可以随着时间的推移，不断学习和预测。

数据特征

作为一个 Kaggle 比赛项目，Kaggle 提供了项目数据，包括训练数据 train.csv，测试数据 test.csv，已经商店信息数据 store.csv。本项目将以离线的方式训练模型，并作出预测，以比赛项目的训练数据作为项目数据，分割出部分（0.5%）作为测试数据。

训练集数据的全部字段说明参见：https://www.kaggle.com/c/rossmann-store-sales/data，本项目将选取如下特征值作为销售预测：

- Store：商店编号
- Date：数据统计日期
- Open：当天门店是否开放，节假日商品可能不开放。
- Promo：当天门店是否进行促销活动
- StateHoliday：国家假日，及其具体类型，类型将被编码使用。
- SchoolHoliday：公共学校假日
- StoreType：门店类型
- Assortment：分类标准
- CompetitionDistance：最近竞争门店的距离，同时用于表达当日是否存在竞争门店
- *CompetitionOpenSinceMonth：竞争门店开店月份,将用于编码 CompetitionDistance*
- *CompetitionOpenSinceYear：竞争门店开店年份，将用于编码 CompetitionDistance*

而回归值，即要预测的数据为：

- Sales：当天销售额

解决办法

神经网络是一种非线性自适应系统模型，它能够通过已知数据自我学习和归纳总结，并提取数据的内部特征，所以非常适合解决销售预测问题的数据挖掘。BP 神经网络是利用 BP 算法进行训练的神经网络，BP 算法是将误差反向传播的算法，即 Back Propagation。本项目将应用 BP 神经网络来实现 Rossmann 销售预测。将使用 TensorFlow 来实现这个 BP 神经网络。

基准模型

AR 预测模型是一种可用于时间序列数据预测的模型。通过 AR 模型建立销售预测模型，就是根据已知时间序列中的销售数据的变化特征和趋势，预测未来销售值。

AR 模型的表示形式为：

$$X_t = c + \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t$$

其中$\varphi_1, \ldots, \varphi_P$为模型（超）参数，c 是常量，$\varepsilon_t$为测量误差，或白噪音。

评估指标

本项目将使用 Kaggle 项目中的评估函数 RMSPE（Root Mean Square Percentage Error）作为评估指标：

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

其中，y_i为一个门店某天的销售量，$\hat{y}_i$为相应的预测值。

设计大纲

项目的实现过程将包含以下几个部分：

1. 数据分析
    a. 各门店按年、按月、按日销售趋势分析
    b. 销售额与关键数据特征 Promo，StoreType，StateHoliday 等相关性及可视化分析
2. 基准模型-AR 预测模型的实现
    a. 对数据分割，用于 AR 模型的训练和测试
    b. 按照上述的 AR 模型实现基于时间的销售数据预测
    c. 按照上述评估指标对 AR 模型在测试集上的表现打分，作为本项目算法的基准指标
3. 数据预处理：
    a. 数据编码：对于 StateHoliday，StoreType 等特征值进行编码，以便模型使用
    b. 数据修正：对于数据的缺失，异常进行修正
    c. 数据分割：将训练集数据分割，按照时间排序后的最大时间的 0.5%作为测试集。
4. BP 神经网络的销售模型的建立和训练
    a. 用 TensorFlow 实现 BP 神经网络
    b. 在训练集上训练 BP 神经网络

5. BP 神经网络的销售模型的预测和结果分析
    a. 用得到的模型，在测试集上测试模型预测结果
    b. 按照上述的评估指标对预测结果进行评估，并与基准模型（AR 模型）的结果进行比较分析。
6. 预测效果分析和实用性分析
    a. 对预测模型在销售预测中的表现进行进一步分析
    b. 对预测模型在实际销售中的可行性和局限性进行分析

参考文献

【1】http://wiki.mbalib.com/wiki/%E9%94%80%E5%94%AE%E9%A2%84%E6%B5%8B
【2】数据挖掘：https://en.wikipedia.org/wiki/Data_mining
【3】Artificial neural network: https://en.wikipedia.org/wiki/Artificial_neural_network
【4】Autoregressive model: https://en.wikipedia.org/wiki/Autoregressive_model