# Machine Learning Engineer Nanodegree

## Rossmann Store Sales Forecast

陈磊

2019/7/7

## 背景



Rossmann 是一家德国大型连锁日化品超市，本项目是结合当前已知的历史数据对Rossmann进行数据分析，该题目是Kaggle上的一个竞赛项目，目标是利用已知Rossmann各门店历史销售数据，通过数据分析手段，分析销售数据的关联性指标找出规律预测未来门店的销售额趋势，用来指导销售工作。

销售额是跟时间相关的数据，通常也称作时间序列数据。销售的业务情况是受多种要素作用的，通过对历史数据分析，选用适当的算法逻辑找出其中具有关联性特征的指标。进而构建销售额预测模型，使用历史数据对模型进行训练，通过优化调整适当参数，对模型进行优化，进而达到预期的效果。

## 问题描述

该项目涉及到数据为时间序列的数据，属于多时间序列的数据分析预测问题，通过分析历史数据以此来分析预测未来的销售走势。本项目重点的是通过数据探索分析历史销售数据，需要找出各门店的历史销售数据与提供的特征指标存在什么样的关联性，分析数据指标中与销售额紧密相关的内容。给定的历史数据中含有时间信息，不同的时间段销售的情况呈现出什么样的趋势，在特定的节假日又会对销售数据产生什么影响。这些问题将在本项目中做详细的分析、探索。

## 数据集和输入

该项目的数据集来源于kaggle竞赛的公开内容，主要有store.csv、train.csv和test.csv文件，其中store是Rossmann 的门店数据，共有1115家门店，训练集train中共有1017209条记录，测试集test中有41088条记录。store中有部分字段数据内容为空，测试集中的open字段有11条数据存在空值。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **文件名** | **数据条数** | **列数** | **含有空值** | **时间区间** | **其他说明** |
| store.csv | 1115 | 10 | 是 | ------- | 暂无 |
| train.csv | 1017209 | 9 | 否 | 2013-01-01~2015-07-31 | 暂无 |
| test.csv | 41088 | 8 | 是 | 2015-08-01~2015-09-17 | 暂无 |

其中train.csv为训练集文件中字段包含：”store”，”DayofWeek”，”Date”，“Sales”，“Customers”，“Open”，“Promo”，“StateHoliday”，”SchoolHoliday”，各字段的含义解释如下：

|  |  |
| --- | --- |
| 字段 | 字段释义 |
| Store | 门店编号 |
| DayOfWeek | 一周的第几天，如5，代表周五 |
| Date | 日期 |
| Sales | 销售额 |
| Customers | 顾客数量 |
| Open | 开业状态，1代表开业，0代表关店 |
| Promo | 无促销活动，1代表有促销，0代表无促销 |
| StateHoliday | 州假日 |
| SchoolHoliday | 学校假日 |

数据样式如下：

|  |
| --- |
| “1”，“5”，“2015-07-31”，“5263”，“555”，“1”，“1”，“0”，“1” |

test.csv为测试集文件，与训练集相比，少了客户数customers、销售额Sales列的内容，多了一列ID，其内容是2015-08-01至2015-09-17范围内的数据。也即是通过训练出的模型，对这个范围内的销售额做出预测。

Store.csv文件中字段包含'Store'，'StoreType'，'Assortment', 'CompetitionDistance'，'CompetitionOpenSinceMonth'，'CompetitionOpenSinceYear'，'Promo2'，'Promo2SinceWeek'，'Promo2SinceYear'， 'PromoInterval'，各字段的含义解释如下：

|  |  |
| --- | --- |
| 字段 | 字段释义 |
| Store | 门店编号 |
| StoreType | 门店类型,a，b，c，d |
| Assortment | 描述分类级别：a =基本，b =额外，c =扩展 |
| CompetitionDistance | 与最近竞争对手的距离 |
| CompetitionOpenSinceMonth | 最近竞争对手开店时长 |
| CompetitionOpenSinceYear | 最近竞争对手开店时间 |
| Promo2 | 有无连续促销活动 |
| Promo2SinceWeek | 描述商店开始参与连续促销的日历周 |
| Promo2SinceYear | 描述商店开始参与连续促销的年份 |
| PromoInterval | 描述Promo2启动的连续间隔，命名重新开始促销的月份。例如，“2月，5月，8月，11月”意味着每一轮在该商店的任何给定年份的2月，5月，8月和11月开始 |

数据样式如下：

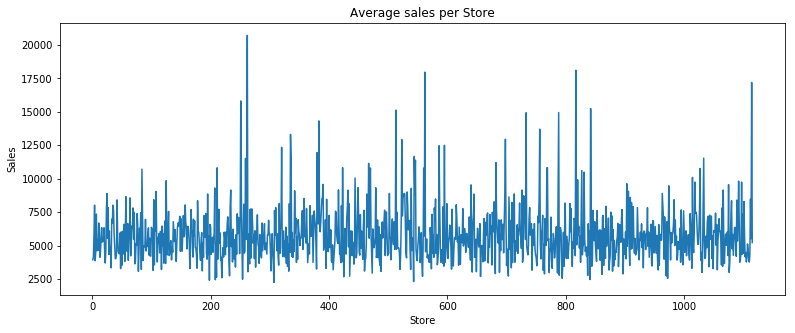
|  |
| --- |
| “2”，“a”，“a”，“570.0”，“11.0”，“2007.0”，“1”，“13.0”，“2010.0”，“Jan,Apr,Jul,Oct” |

因数据集store与训练集train和测试集test中都存在“store”字段，为了便于分析和处理，在项目中将通过pandas模块分别将train、test与store合并。鉴于数据集中部分字段值缺失，因此在处理时将根据数据特点进行值填充或过滤数据。

|  |
| --- |
| display(df\_train.isnull().sum(),df\_test.isnull().sum(),df\_store.isnull().sum()) |

|  |
| --- |
| Store 0  DayOfWeek 0  Date 0  Sales 0  Customers 0  Open 0  Promo 0  StateHoliday 0  SchoolHoliday 0  dtype: int64  Id 0  Store 0  DayOfWeek 0  Date 0  Open 11  Promo 0  StateHoliday 0  SchoolHoliday 0  dtype: int64  Store 0  StoreType 0  Assortment 0  CompetitionDistance 3  CompetitionOpenSinceMonth 354  CompetitionOpenSinceYear 354  Promo2 0  Promo2SinceWeek 544  Promo2SinceYear 544  PromoInterval 544  dtype: int64 |

Test数据集中Open有11条数据为空，在填充时，考虑存在促销活动，及节假日情况，可填充为“1”，即营业状态。针对Store数据集中CompetitionDistance、CompetitionOpenSinceMonth、CompetitionOpenSinceYear、Promo2SinceWeek、Promo2SinceYear、PromoInterval为了方便处理，可以将其控制填充为0。下图为每家店的平均销售额情况：



## 解决方案

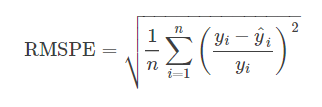
首先观察数据集并对数据做预处理，将缺失数据进行补全或过滤，借助matplotlib库中的多种图形样式对数据进行展示，同时使用seaborn辅助分析，找出特征之间的相关性并进行数据探索，期望找出更有价值的数据关系。利用xgboost算法构建销售额预测模型，并通过调整模型的参数以达到较高的准确率。

## 基准模型

XGBoost（eXtreme Gradient Boosting）全名叫极端梯度提升，在数据挖掘中应用较广泛。boosting算法的其中一种。Boosting算法的思想是将许多弱分类器集成在一起形成一个强分类器。因为XGBoost是一种提升树模型，所以它是将许多树模型集成在一起，形成一个很强的分类器。XGBoost算法思想就是不断地添加树，不断地进行特征分裂来生长一棵树，每次添加一个树，其实是学习一个新函数，去拟合上次预测的残差。在本项目中，我们将使用XGBoost算法来 构建预测模型，这个算法相较于其他算法速度较快，操作也相对方便，便于优化和调参，对准确率也有保证。

## 评估指标

由于本项目是kaggle上的竞赛项目，所以测评指标采用的是测试集获取rmspe(Root Mean Square Error)  均方根百分比误差的分数，其具体的计算公式参考如下：



## 项目设计

1. 数据理解与分析

首先引入项目需要的库函数，加载数据文件观察数据，掌握各个特征的含义。并逐一分析和观察数据的特点和数据分布情况。

1. 数据处理和探索

检查数据完整性，根据真实可能的情况进行数据的填充完善，在分析的过程中，将各指标通过图形化方式直观化展示分析。

1. 模型建立及优化

因为本项目数据属于时间序列类型，对时间先后有特定要求，因此在划分训练集和验证集时通常不建议按照指定比例划分。项目最终是预测2015-8-1至2015-9-17近6周的销售额，因此划分数据时训练集中后6周的数据（按时间字段升序排列）作为验证集，而后6周以前的数据则作为训练集。利用xgboost算法构建预测模型，而用模型去预测未来6周的销售额。并通过不断的优化调整模型参数，构建一套相对可靠的预测模型。

1. 分析预测得出结论

利用构建好的预测模型对给定的数据进行预测，根据评价指标定义的规则，对模型的结果给予评判并得出最终的结论。

## 参考文献

[1] [www.kaggle.com](http://www.kaggle.com)

[2] Tianqi Chen XGBoost：A Scalable Tree Boosting System

[3] 李航 统计学习方法 清华大学出版社，2012：137-153

[4] Wes [McKinney](http://search.dangdang.com/?key2=McKinney&medium=01&category_path=01.00.00.00.00.00)  python for Data Analysis OREILLY