机器学习纳米学位毕业项目

Rossmann Sales Forecast

陈磊

2019年7月20日

I. 问题的定义

项目概述

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。

本项目使用的数据集为kaggle公开内容，其中涉及到1115家店2013-1-1至2015-7-31期间的销售数据，包括竞争对手的信息，门店的促销信息以及每日客户量、销售额等信息，销售额是跟时间相关的数据，通常也称作时间序列数据，因此本项目是属于时间序列类的。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，进而使用模型对各门店在2015-8-1至2015-9-17期间的销售额进行预测。

项目中使用的数据集主要是有train.csv、store.csv和test.csv，其中train为历史销售数据，store为1115家店铺的信息，test为用来预测未来6周的店铺销售信息（无销售和客户量信息）。在实际处理的过程中将分别对train与store，test与store进行拼接形成新的train和test信息，从而更有效的便于数据的分析和模型的训练。

项目使用python语言，借助jupyter notebook平台进行数据的分析，同时借助matplotlib和seaborn进行可视化图形、图标展示。

问题陈述

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， 根据目前已知的商店及历史销售的数据，其中也包括竞争对手门店距离，促销活动，假日等数据，依此来预测未来6周的销售情况，也即是通过分析历史数据找出规律来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。

针对该时间序列的问题，为了最终实现对2015-8-1至2015-9-17范围内的销售额预测，整体可分为如下几个步骤：

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。

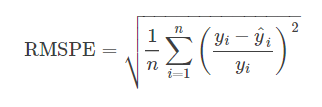
1. 分析预测得出结论

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。

通过以上的步骤，期望能够在对数据有充分理解的基础上，实现对未来数据的预测并最终完成整个项目。

评价指标

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：



在这里，你需要说明你将要用于评价自己的模型和结果的\*\*指标\*\*和\*\*计算方法\*\*。它们需要契合你所选问题的特点及其所在的领域，同时，你要保证他们的合理性。需要考虑的问题：

- \_你是否清晰地定义了你所使用的指标和计算方法？\_

- \_你是否论述了这些指标和计算方法的合理性？\_

II. 分析

数据的探索

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，如下为数据集的整体情况。

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

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代表营业，0代表关店 |
| Promo | 有无促销活动，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” |

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。

train和store数据集中都含有特征“Store”，为了方便数据的分析和处理，因此在处理是需要将train和store通过pandas库的merage函数连接起来组成新的train数据集，用于数据探索和划分训练集。同样的，将test和store数据集连接后组成新的test数据集，使用模型进行预测。另外，train训练集中含有的“Date”特征格式为“2015-04-21”，为了统计分析的方便，使用pandas库的to\_datetime函数进行转换提取年份特征“Year”，同样的可以获得月份特征“Month”。

因为train训练集中有销售额特征“Sales”和客户量特征“**Customers**”，因此可以获得新特征“SalesPerCustomer”。

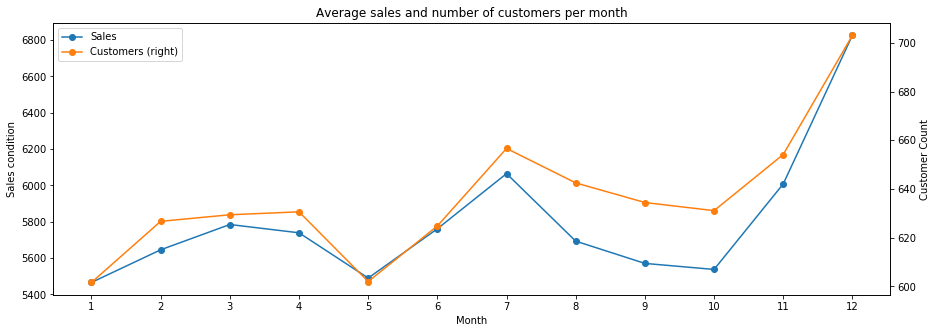
探索性可视化

在数据分析和探索的过程中，通常借助图形、图表等可视化的效果直观的展示不同指标数据之间的相关性。

1. 每月销售额与客户数的平均值走势

|  |
| --- |
| ax=df\_train.groupby('Month')[['Sales','Customers']].mean().plot(secondary\_y=['Customers'], figsize=(15,5), marker='o')  ax.set\_ylabel('Sales condition')  ax.right\_ax.set\_ylabel('Customer Count')  ax.set\_xticks(range(1,13))  plt.title('Average sales and number of customers per month'); |

上述为生成图形的代码，这部分主要是统计所有门店的历史销售数据以月份为单位，统计平均每个月的销售额和客户数，主要理解一下用户数与销售额的关系，理论上，客户越多，销售额也就越多，反之亦然。从下图1中的趋势能够看出是符合正常预期的。



（图1）

在这一部分，你需要对数据的特征或特性进行概括性或提取性的可视化。这个可视化的过程应该要适应你所使用的数据。就你为何使用这个形式的可视化，以及这个可视化过程为什么是有意义的，进行一定的讨论。你需要考虑的问题：

- \_你是否对数据中与问题有关的特性进行了可视化？\_

- \_你对可视化结果进行详尽的分析和讨论了吗？\_

- \_绘图的坐标轴，标题，基准面是不是清晰定义了？\_

算法和技术

在这一部分，你需要讨论你解决问题时用到的算法和技术。你需要根据问题的特性和所属领域来论述使用这些方法的合理性。你需要考虑：

- \_你所使用的算法，包括用到的变量/参数都清晰地说明了吗？\_

- \_你是否已经详尽地描述并讨论了使用这些技术的合理性？\_

- \_你是否清晰地描述了这些算法和技术具体会如何处理这些数据？\_

基准模型

在这一部分，你需要提供一个可以用于衡量解决方案性能的基准结果/阈值。这个基准模型要能够和你的解决方案的性能进行比较。你也应该讨论你为什么使用这个基准模型。一些需要考虑的问题：

- \_你是否提供了作为基准的结果或数值，它们能够衡量模型的性能吗？\_

- \_该基准是如何得到的（是靠数据还是假设）？\_

III. 方法

\_(大概 3-5 页）\_

数据预处理

在这一部分， 你需要清晰记录你所有必要的数据预处理步骤。在前一个部分所描述的数据的异常或特性在这一部分需要被更正和处理。需要考虑的问题有：

- \_如果你选择的算法需要进行特征选取或特征变换，你对此进行记录和描述了吗？\_

- \_\*\*数据的探索\*\*这一部分中提及的异常和特性是否被更正了，对此进行记录和描述了吗？\_

- \_如果你认为不需要进行预处理，你解释个中原因了吗？\_

执行过程

在这一部分， 你需要描述你所建立的模型在给定数据上执行过程。模型的执行过程，以及过程中遇到的困难的描述应该清晰明了地记录和描述。需要考虑的问题：

- \_你所用到的算法和技术执行的方式是否清晰记录了？\_

- \_在运用上面所提及的技术及指标的执行过程中是否遇到了困难，是否需要作出改动来得到想要的结果？\_

- \_是否有需要记录解释的代码片段(例如复杂的函数）？\_

完善

在这一部分，你需要描述你对原有的算法和技术完善的过程。例如调整模型的参数以达到更好的结果的过程应该有所记录。你需要记录最初和最终的模型，以及过程中有代表性意义的结果。你需要考虑的问题：

- \_初始结果是否清晰记录了？\_

- \_完善的过程是否清晰记录了，其中使用了什么技术？\_

- \_完善过程中的结果以及最终结果是否清晰记录了？\_

IV. 结果

\_（大概 2-3 页）\_

模型的评价与验证

在这一部分，你需要对你得出的最终模型的各种技术质量进行详尽的评价。最终模型是怎么得出来的，为什么它会被选为最佳需要清晰地描述。你也需要对模型和结果可靠性作出验证分析，譬如对输入数据或环境的一些操控是否会对结果产生影响（敏感性分析sensitivity analysis）。一些需要考虑的问题：

- \_最终的模型是否合理，跟期待的结果是否一致？最后的各种参数是否合理？\_

- \_模型是否对于这个问题是否足够稳健可靠？训练数据或输入的一些微小的改变是否会极大影响结果？（鲁棒性）\_

- \_这个模型得出的结果是否可信？\_

合理性分析

在这个部分，你需要利用一些统计分析，把你的最终模型得到的结果与你的前面设定的基准模型进行对比。你也分析你的最终模型和结果是否确确实实解决了你在这个项目里设定的问题。你需要考虑：

- \_最终结果对比你的基准模型表现得更好还是有所逊色？\_

- \_你是否详尽地分析和讨论了最终结果？\_

- \_最终结果是不是确确实实解决了问题？\_

V. 项目结论

\_（大概 1-2 页）\_

结果可视化

在这一部分，你需要用可视化的方式展示项目中需要强调的重要技术特性。至于什么形式，你可以自由把握，但需要表达出一个关于这个项目重要的结论和特点，并对此作出讨论。一些需要考虑的：

- \_你是否对一个与问题，数据集，输入数据，或结果相关的，重要的技术特性进行了可视化？\_

- \_可视化结果是否详尽的分析讨论了？\_

- \_绘图的坐标轴，标题，基准面是不是清晰定义了？\_

对项目的思考

在这一部分，你需要从头到尾总结一下整个问题的解决方案，讨论其中你认为有趣或困难的地方。从整体来反思一下整个项目，确保自己对整个流程是明确掌握的。需要考虑：

- \_你是否详尽总结了项目的整个流程？\_

- \_项目里有哪些比较有意思的地方？\_

- \_项目里有哪些比较困难的地方？\_

- \_最终模型和结果是否符合你对这个问题的期望？它可以在通用的场景下解决这些类型的问题吗？\_

需要作出的改进

在这一部分，你需要讨论你可以怎么样去完善你执行流程中的某一方面。例如考虑一下你的操作的方法是否可以进一步推广，泛化，有没有需要作出变更的地方。你并不需要确实作出这些改进，不过你应能够讨论这些改进可能对结果的影响，并与现有结果进行比较。一些需要考虑的问题：

- \_是否可以有算法和技术层面的进一步的完善？\_

- \_是否有一些你了解到，但是你还没能够实践的算法和技术？\_

- \_如果将你最终模型作为新的基准，你认为还能有更好的解决方案吗？\_

----------

\*\* 在提交之前， 问一下自己... \*\*

- 你所写的项目报告结构对比于这个模板而言足够清晰了没有？

- 每一个部分（尤其\*\*分析\*\*和\*\*方法\*\*）是否清晰，简洁，明了？有没有存在歧义的术语和用语需要进一步说明的？

- 你的目标读者是不是能够明白你的分析，方法和结果？

- 报告里面是否有语法错误或拼写错误？

- 报告里提到的一些外部资料及来源是不是都正确引述或引用了？

- 代码可读性是否良好？必要的注释是否加上了？

- 代码是否可以顺利运行并重现跟报告相似的结果？

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