# Rossmann销售额预测

### 机器学习工程师纳米学位毕业项目

薛威

2018年02月19日

**I. 定义**

**项目概述**

该项目是Kaggle上由Rossmann两年前建立的一个预测比赛。比赛的目标是取得一个能够正确预测六周后的日销售情况。

Rossmann 在欧洲经营7国经营着3000家药店，目前，Rossmann商店的经理被要求预测他们未来六周的日销售情况，商店销售收很多因素的影响，包括促销，竞争，学校放假和法定节假日，节气性和区域性。由于数千经营者依据他们独特的情况预测销售情况，结果的准确性可能有很大不同。 在这个项目中，将挑战预测6周的1115家德国境内的Rossmann商店的每日销售额，可靠的销售预报可以让商店经营者增加工作效率和积极性创建更高效的工作人员安排。通过帮助Rossmann创建一个强壮的预测模型，你将帮助经营者保持关注对他们来说最重要的是：他们的客户和他们的团队。

本项目将使用目前Kaggle上线性预测普遍表现很好的XGBoost算法模型来建模，并验证所取得的结果。

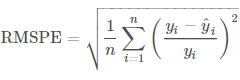
**问题陈述**

我要通过对旧的销售记录的学习建模，然后预测六周后的日销售情况。具体步骤和可能遇到的问题如下：

* 对数据做清洗和整理，可能遇到的问题有异常值的处理，缺失值的处理。我将针对存在缺失数据的特征的具体情况来补充或丢弃训练数据，对于每个特征的异常值做一次筛选判断。
* 对数据做单个变量分析，多个变量分析，对变量做特征处理，可能会遇到整合多个变量为一个有效特征的问题。我将通过对问题的理解来对某几个特征做一些基本运算将其转换为更直观的特征展现出来。
* 使用XGBoost来建模，如何调整训练模型的数据用来适合XGBoost模型的输入要求。我将依照XGBoost模型的传入参数的要求来对输入特征做数据转换。
* 在使用XGBoost的基础上如何找到最合适的超参。我将使用CV来筛选出最好的参数进行建模。
* 预测并评价结果，根据Kaggle的描述使用RMSPE来做为评价标准，需要对我建立的模型做一个评估。我将使用建立好的模型对于test数据进行预测然后将其上传到Kaggle上查看正确率。

**评价指标**

我将使用Kaggle在该项目建议的RMSPE来做为验证函数，该值越低代表差异性越小。它是指模型的预测值和实际观察值之间的差异的一种衡量方式。



y是真实的销售数据而y\_hat是预测数据，任何的为0的销售数据不参加此评价。采用该评价函数的好处是该问题是一个线性回归的问题，预测的结果是一个具体的数值向量，相关的预测数据只能是一个大概无法精确匹配，所以计算预测值和实际值之间的差额平方并累加起来平均最后再开方得到的结果相比直接实际值和预测值做差额来说的大大减少了预测值与实际值之间过于详细的数值匹配要求。

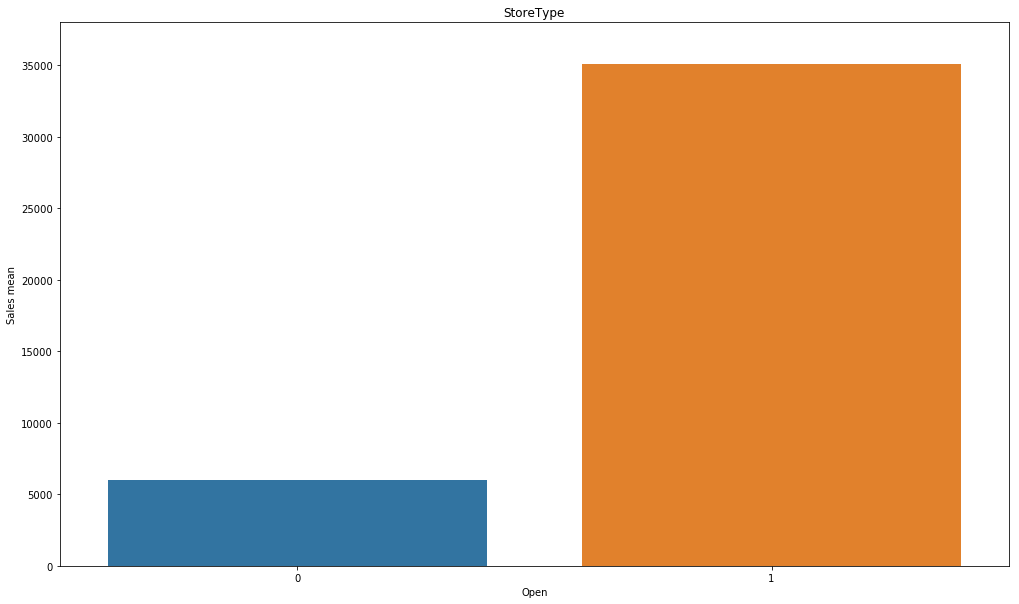
**II. 分析**

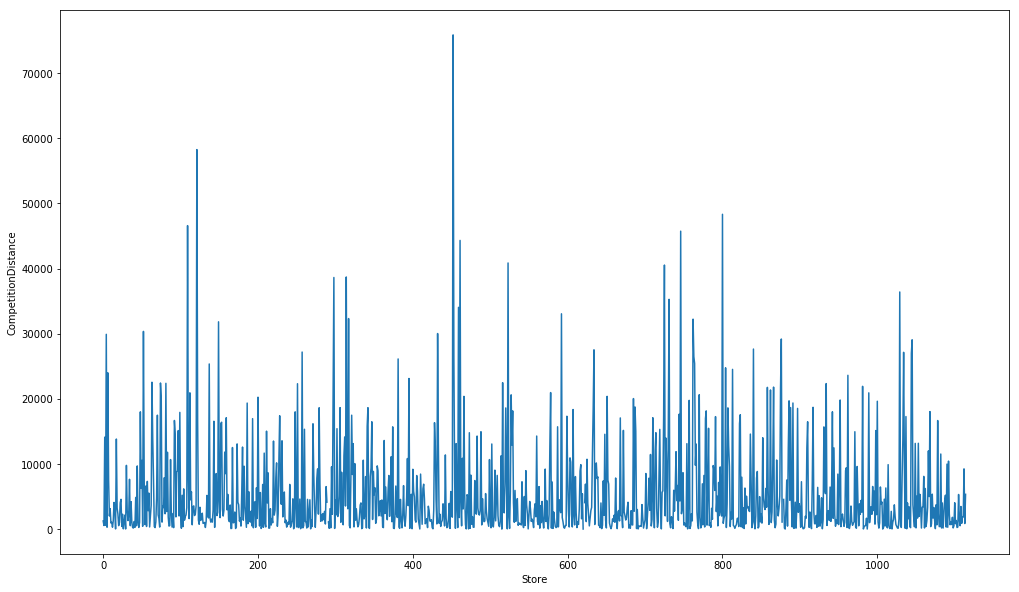
**数据的探索**

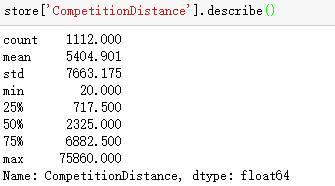
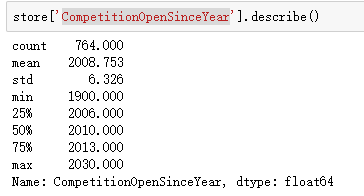
Rossmann提供了相关的数据集，包含训练的和预测比赛所需要的。

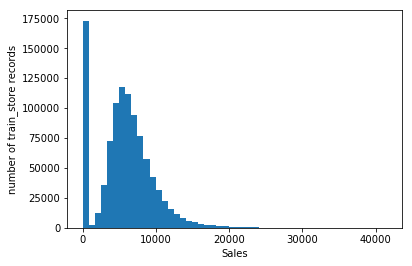
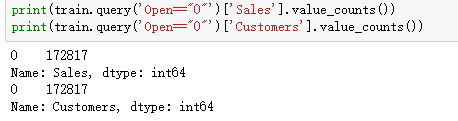
* 输入数据集如下  
  train.csv - 包括销售额的历史数据训练用  
  包含有"Store","DayOfWeek","Date","Sales","Customers","Open","Promo","StateHoliday","SchoolHoliday"字段。  
  test.csv - 包括销售额的历史数据测试用historical data excluding Sales  
  包含有"Id","Store","DayOfWeek","Date","Open","Promo","StateHoliday","SchoolHoliday"字段。  
  sample\_submission.csv - 预测数据格式样本  
  包含有"Id","Sales"字段。  
  store.csv - 关于商店的附加信息  
  包含有"Store","StoreType","Assortment","CompetitionDistance","CompetitionOpenSinceMonth"，"CompetitionOpenSinceYear","Promo2","Promo2SinceWeek","Promo2SinceYear","PromoInterval"字段。
* 数据集特征如下  
  Id - 测试集中表示一条记录的编号。  
  Store - 每个商店的唯一编号。  
  Sales - 任意一个给定日期的销售营业额。  
  Customers - 给定那一天的消费者数。  
  Open - 商店是否开门标志，0为关，1为开。  
  StateHoliday - 表明影响商店关门的节假日，正常来说所有商店，除了极少数，都会在节假日关门，a=所有的节假日，b=复活节，c=圣诞节，所有学校都会在公共假日和周末关门。  
  SchoolHoliday - 表明商店的时间是否受到公共学校放假影响。  
  StoreType - 四种不同的商店类型a，b，c和d。  
  Assortment - 描述种类的程度，a = basic, b = extra, c = extended。  
  CompetitionDistance - 最近的竞争对手的商店的距离。  
  CompetitionOpenSince[Month/Year] - 最近的竞争者商店大概开业的年和月时间。  
  Promo - 表明商店该天是否在进行促销。  
  Promo2 - 指的是持续和连续的促销活动。: 0 = 商店没有参加, 1 = 商店正在参加。 Promo2Since[Year/Week] - 表示参加连续促销开始的年份和周。  
  PromoInterval - 描述持续促销间隔开始，促销的月份代表新的一轮，月份意味着每一轮的开始在哪几个月。
* 相关的初步统计分析如下  
  test.csv有41088条数据其中缺失数据是在特征Open上缺失了11条，根据缺失的Date，StateHoliday和SchoolHoliday来判断推测Open均为1并填入。  
  store.csv有1115条数据其中缺失数据特征是CompetitionDistance，CompetitionOpenSinceMonth，CompetitionOpenSinceYear，Promo2SinceWeek，Promo2SinceYear和PromoInterval。  
  CompetitionDistance缺失3条，我推测这三家商店在有效的距离内没有竞争对手用一个特别大的值来处理，CompetitionOpenSinceMonth和CompetitionOpenSinceYear缺失的情况一直，我推测就在很早之前或者说在train数据之前就存在这个竞争对手了我给一个默认的之前的时间2010年，Promo2SinceWeek，Promo2SinceYear和PromoInterval。这三个特征的确是情况也都一样，也就是没有参加Promo2的这三项均为空，那我就将时间设置为一个未来时间2030年，PromoInterval用“0，0，0，0”来填补。train.csv有1017209条数据，无数据缺失情况，所有提供的数据总体来看没有异常值情况。

**探索性可视化**

针对test数据集做特征open的可视化  
test数据集里的Open为0 的直接将预测值设置为0  
针对store数据集查看特征CompetitionDistance来看各个距离的商店的分布

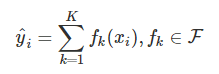


并查看该特征的详细情况  
  
中位数在2325m，还有部分大于20km的，对于三个缺失的值我也做一个最大值99999来填充，表示有效距离内无竞争对手。 20km以上我之后都将标记为无竞争对手  
针对store数据集查看特征CompetitionOpenSinceYear来看竞争对手的开店时间  


缺失值就用中位数2010来填。  
对于train数据集对Sales做可视化  
  
发现有大量为0的数据，Sales为0对于训练来说并无意义，而且会干扰最终结果应该只训练Sales不为0的，同时发现Open特征为0和Sales为0表现一致  
  
我将只训练Open为1的数据集

**算法和技术**

该问题属于监督学习里的回归问题，也就是依据已有的数据集建模预测测试集的问题。  
在特征上通过预处理将空值根据相关的业务场景做填补，对于异常值经行处理避免过拟合数据，对日期做转换，然后对于部分分类特征比如Assortment，StoreType和StateHoliday做one-hot，让train和test与store数据集做连接，最后只考虑Open的训练数据。  
模型选择上，XGBoost非常适合于这个领域。XGBoost是对一堆的CART树（分类回归树）做预测，然后再将各个树的预测分数相加。模型用公式表示就是:

  
简答说K代表树的数量，F是所有的CART树，f是某一个树。  
在具体的xgb使用的时候需要考虑的参数有  
General parameters：参数控制在提升（boosting）过程中使用哪种booster，本项目使用默认值 gbtree。  
Booster parameters：这取决于使用哪种booster，设置如下：  
eta [default=0.3]  
max\_depth [default=6]  
Learning Task parameters：控制学习的场景的相关参数 。  
objective [“reg:linear”] 。  
feval设置为RMSPE为评价函数。  
num\_boost\_round设置boosting的次数。  
early\_stopping\_rounds 可以提前终止程序，这样可以找到最优的迭代次数。

设置完XGBoost模型后使用train\_test\_split将train数据分割为train和valid来训练，最后用test数据上传kaggle来评估结果。

**基准模型**

由于是销售预测模型，我对数据预测的错误率认为假设为20%都是可以接受的范围，我的选择是XGBoost来作为我的预测模型，衡量结果的方式是Kaggle推荐的RMSPE。

由于是监督学习的线性回归问题，所以我还将采用Ridge回归模型和Lasso回归模型来作为基准回归模型计算其结果。

**III. 方法**

*(大概 3-5 页）*

**数据预处理**

整理test数据集的缺失数据，仅有Open特征通过判断日期推测出缺失的均为1，对于DATE做数据转换，变成Year，Month，Day，对StateHoliday做One-Hot。  
整理store数据集的缺失数据，处理CompetitionOpenSinceMonth,CompetitionOpenSinceYear缺失通过取中位数来填补，Promo2SinceWeek，Promo2SinceYear和PromoInterval的缺失通过设置为0，远大于目前的时间2030，（0，0，0，0）来处理为了后面连接train和test来设置没有参加Promo2活动，对于Date也做相同的数据转换，将PromoInterval里的string用数字来做替换，对StoreType做One-Hot。  
整理train数据集，转换特征Date，对StateHoliday做One-Hot，但是只对Open=1的做训练。  
将train和test都与store数据通过store特征做连接得出train\_data和test\_data。对这两个数据集相同的处理，将CompetitionOpenSinceMonth,CompetitionOpenSinceYear合并为一个特征CompetitionMonths表示竞争对手的开业持续月份。将Promo2SinceWeek Promo2SinceYear PromoInterval合并为一个特征IsPromo2表示当前时间里是否在参加Promo2。

**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）？*

采用Ridge算法建模

采用Lasso算法建模

采用Xgboost算法建模

**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？
* **VI.相关引用**

### <http://xgboost.readthedocs.io/en/latest/model.html>