|  |
| --- |
|  |
| 机器学习工程师纳米学位毕业项目 |
| **预测Rossmann未来的销售额** |
|  |
| 优达学城 |
|  |
| **2019/1/12** |

|  |
| --- |
|  |

目录

[I问题的定义 3](#_Toc535489651)

[项目概述 3](#_Toc535489652)

[问题陈述 3](#_Toc535489653)

[评价指标 4](#_Toc535489654)

[II分析 4](#_Toc535489655)

[数据的探索 4](#_Toc535489656)

[探索性可视化 7](#_Toc535489657)

[单变量分析 7](#_Toc535489658)

[1. 数据说明 9](#_Toc535489659)

[2. 解决方案 10](#_Toc535489660)

[3. 基准模型 11](#_Toc535489661)

[4. 评估指标 11](#_Toc535489662)

[5. 方案设计 11](#_Toc535489663)

[6. 参考文献 11](#_Toc535489664)

[参考文献 12](#_Toc535489665)

[表目录 12](#_Toc535489666)

[图目录 12](#_Toc535489667)

# 问题的定义

## 项目概述

随着经济全球化的发展，企业面临着更加复杂和残酷的市场竞争。能够快速准确的预测出来销售额从而合理的安排生产和库存，用低成本的产品快速满足客户要求成为企业关心的重点。传统的销售预测方法分为定性和定量两类，定性方法主要有市场调研、购买者期望分析、专家小组法等，定量方法主要有平均数趋势预测、因果预测分析、时间序列分析法等统计方法。随着大数据和人工智能技术的兴起，机器学习模型给销售额的预测带来了新的思路。

本次项目问题源自Kaggle竞赛，为欧洲的一家连锁药店Rossmann预测未来的销售情况。Rossmann在欧洲的7个国家拥有3000多家连锁药店。需要帮助他们的管理者，基于历史数据对位于德国的1115家药店预测未来6周的销售额。项目主要涉及三个数据集，包含店铺基本信息的store.csv，共1115个店铺开店情况、竞争对手情况、促销情况的数据；train.csv，包含1017209从2013年1月年至2015年7月1115个店铺每天的销售额、用户数等数据；test.csv，包含41088条从2015年8月1日至2015年9月17日间每天的假期状态、每个店铺的促销状态等数据。我们需要借助store表和train表的数据构建预测模型，再利用test表和store表结合的数据，预测test表中列出的店铺在当日的假期及促销条件下会产生的销售额。

## 问题陈述

本项目是一个回归预测问题，目标是根据给出的数据信息，构建一个合适的预测模型，为店铺预测出具体某天的销售额。

为实现这一目标，首先，我们将通过数据探视了解数据的基本信息、分布情况，因在训练集中的数据有1017209条，是1115个店铺按时间序列记录的销售数据，在数据探视时，可能需要采用多维度统计分析及数据可视化的方法，全方位了解数据。在充分了解数据后，对于缺失数据、异常数据进行清洗规整，此外，为了扩充特征范围，可以考虑通过一定方法对原始数据加工产生新的特征。数据规整完成后，根据数值范围，对数据进行归一化或ont-hot转换。然后，构建数据模型训练流水线及评估指标，帮助我们在逻辑回归、SVR、XGboost等预测模型中选出最合适的模型，并对合适的预测模型进行特征和参数的优化，使其达到最优。最后，将利用最优模型对测试集进行预测，实现预测目标。

## 评价指标

针对回归问题的评价指标通常有平均绝对误差（MAE）、平均平方误差（MSE）、均方根误差（RMSE）、均方根百分比误差（RMSPE）、R2决定系数等[1]。应题目要求本项目将采用RMSPE指标来评价模型的预测效果，公式如下：

其中，n为样本数量。

# 分析

## 数据的探索

本问题涉及的三个数据集的具体内容如下：

1. store.csv，共有1115条数据，包含以下字段，各字段统计信息见表格1：

Store——店铺编号，整型，非空数据；

StoreType（离散型）——店铺类型，字符串，含4种类型a, b, c, d，非空数据；

Assortment（离散型）——货品品类，字符串，含3种类型a = basic, b = extra, c = extended，非空数据；

CompetitionDistance——最近的竞争店铺距离，浮点型，有3个缺失数据；

CompetitionOpenSinceMonth（离散型）——最近的竞争店铺开业月份（估计值），浮点型，354个缺失数据；

CompetitionOpenSinceYear（离散型）——最近的竞争店铺开业年份（估计值），浮点型，354个缺失数据；

Promo2（离散型）——店铺是否参加连续促销，1是，0否，整型，非空数据；

Promo2SinceWeek（离散型）——连续促销在第几周开始，浮点型，544个缺失数据；

Promo2SinceYear（离散型）——连续促销在哪一年开始，浮点型，544个缺失数据；

PromoInterval（离散型）——每年的连续促销在哪几个月份开始，字符串，544个缺失数据。

其中，离散型数据有StoreType、Assortment、CompetitionOpenSinceMonth、CompetitionOpenSinceYear、Promo2、Promo2SinceWeek、Promo2SinceYear、PromoInterval，连续型数据只有CompetitionDistance。

表格 1 store表中字段的统计信息

|  | **Store** | **CompetitionDistance** | **CompetitionOpenSinceMonth** | **CompetitionOpenSinceYear** | **Promo2** | **Promo2SinceWeek** | **Promo2SinceYear** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **count** | 1115.00000 | 1112.000000 | 761.000000 | 761.000000 | 1115.000000 | 571.000000 | 571.000000 |
| **mean** | 558.00000 | 5404.901079 | 7.224704 | 2008.668857 | 0.512108 | 23.595447 | 2011.763573 |
| **std** | 322.01708 | 7663.174720 | 3.212348 | 6.195983 | 0.500078 | 14.141984 | 1.674935 |
| **min** | 1.00000 | 20.000000 | 1.000000 | 1900.000000 | 0.000000 | 1.000000 | 2009.000000 |
| **25%** | 279.50000 | 717.500000 | 4.000000 | 2006.000000 | 0.000000 | 13.000000 | 2011.000000 |
| **50%** | 558.00000 | 2325.000000 | 8.000000 | 2010.000000 | 1.000000 | 22.000000 | 2012.000000 |
| **75%** | 836.50000 | 6882.500000 | 10.000000 | 2013.000000 | 1.000000 | 37.000000 | 2013.000000 |
| **max** | 1115.00000 | 75860.000000 | 12.000000 | 2015.000000 | 1.000000 | 50.000000 | 2015.000000 |

1. train.csv，共有1017209条数据，包含以下字段，各字段的统计信息见表格2：

Store——店铺编号，整型，非空数据；

DayOfWeek（离散型）——统计日期是周几，整型，含7种类型，非空数据；

Date——统计日期，字符串，非空数据；

Sales——当日销售额，整型，非空数据；

Customers——当日客户数，整型，非空数据；

Open（离散型）——当日是否开业，整型，含2种类型，1是0否，非空数据；

Promo（离散型）——当日是否促销，整型，含2种类型，1是0否，非空数据；

StateHoliday（离散型）——当日是否法定假期，字符串，含4种类型，0 =非假期 a = 公共假期 b = 复活节假期 c = 圣诞假期，非空数据；

SchoolHoliday（离散型）——当日是否公立学校假期，整型，含2种类型，1是0否，非空数据。

其中，离散型数据有DayOfWeek、Open、Promo、StateHoliday、SchoolHoliday，Sales和Customers为连续型。

表格 2 train表的统计信息

|  | **Store** | **DayOfWeek** | **Sales** | **Customers** | **Open** | **Promo** | **SchoolHoliday** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **count** | 1017209 | 1017209 | 1017209 | 1017209 | 1017209 | 1017209 | 1017209 |
| **mean** | 558.4297 | 3.998341 | 5773.819 | 633.1459 | 0.8301067 | 0.3815145 | 0.1786467 |
| **std** | 321.9087 | 1.997391 | 3849.926 | 464.4117 | 0.3755392 | 0.4857586 | 0.3830564 |
| **min** | 1.000000 | 1.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| **25%** | 280.0000 | 2.000000 | 3727.000 | 405.0000 | 1.000000 | 0.000000 | 0.000000 |
| **50%** | 558.0000 | 4.000000 | 5744.000 | 609.0000 | 1.000000 | 0.000000 | 0.000000 |
| **75%** | 838.0000 | 6.000000 | 7856.000 | 837.0000 | 1.000000 | 1.000000 | 0.000000 |
| **max** | 1115.000 | 7.000000 | 41551.00 | 7388.000 | 1.000000 | 1.000000 | 1.000000 |

1. test.csv，共有41088条数据，包括以下字段：

Id——数据编号，整型，非空数据；

Store——店铺编号，整型，非空数据；

DayOfWeek（离散型）——待预测日期是周几，整型，非空数据；

Date——日期，字符串，非空数据；

Open（离散型）——当日是否开业，浮点型，11个缺失数据；

Promo（离散型）——当日是否促销，整型，含2种类型，1是0否，非空数据；

StateHoliday（离散型）——当日是否法定假期，当日是否法定假期，字符串，含4种类型，0 =非假期 a = 公共假期 b = 复活节假期 c = 圣诞假期，非空数据；

SchoolHoliday（离散型）——当日是否公立学校假期，整型，含2种类型，1是0否，非空数据。

三张表的数据多为离散型，处理时可以进行one-hot转换。store表中，竞争对手距离为nan的数据项，其竞争对手开业时间均为nan，可用0填充月份、年份，竞争对手距离不为nan的数据项，其竞争对手开业时间为nan的，用频数最多的开业月份和年份值填充，另外，竞争对手距离数据值分布范围较广，要对数据值做对数转换。店铺连续促销开始年的nan值用1900填充，促销开始周nan值用0值填充，其他少量离散数据的缺失可以新增一个类型项做填充（如，0、1，缺失值可以用2填充），其他连续型数据的少量缺失值采用均值填充，日期数据计划作为连续型数据来处理。另外，部分原始数据类型为字符串，实际处理的时候，要在读入数据时做数据类型转换。

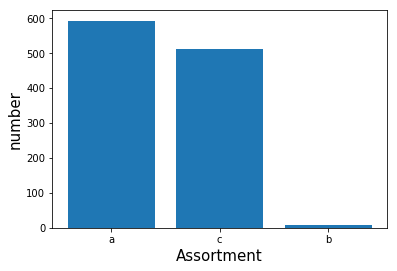
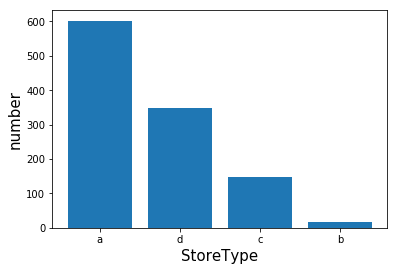
根据题目要求，需要我们根据训练数据集和店铺的基本信息情况，预测出测试数据集中店铺在给出的日期和促销状态下的销售额。对训练数据集中的销售额字段进行基本分析发现，整体销售额最大值为41551，最小值为0，平均值为5773.8，75%的数据都小于7856，属于偏态分布。因销售额的统计是按每天每家店的维度统计的，在进行数据分析时还应考虑按店或按天的维度分布统计后的情况。

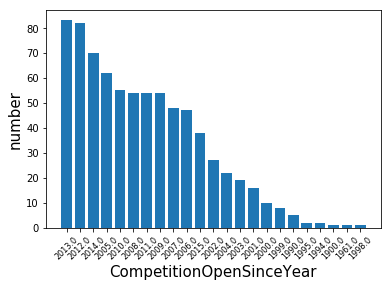
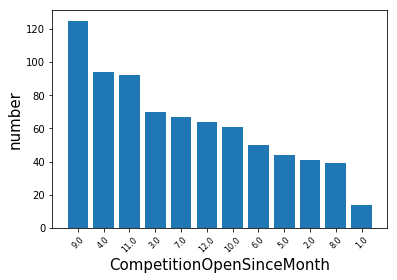
## 探索性可视化

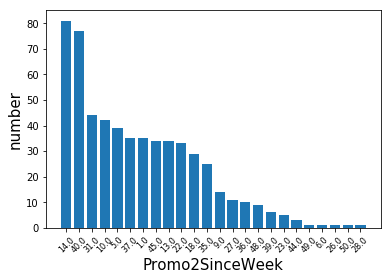
### 单变量分析

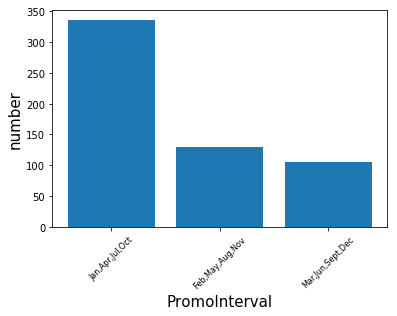
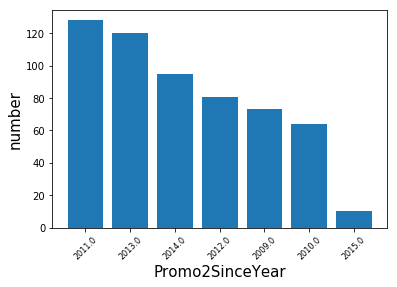
1. 对store表的可视化分析

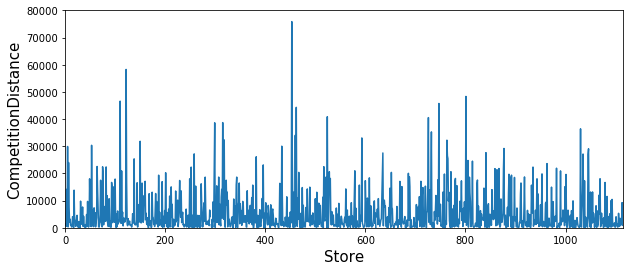
针对store表中的数据，进行统计分析结果如下：









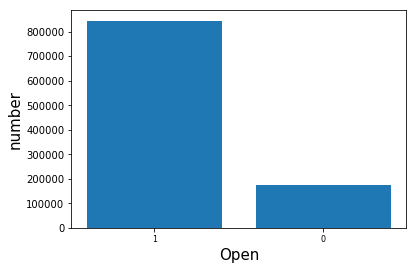
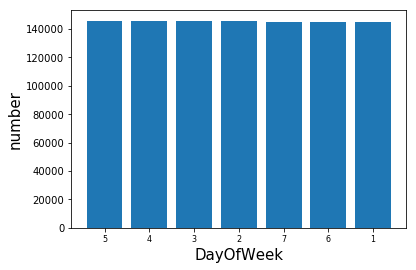


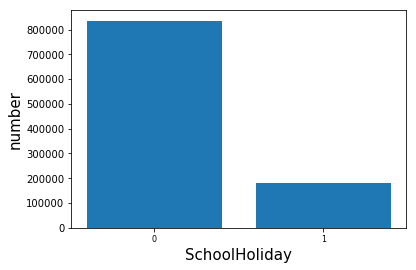
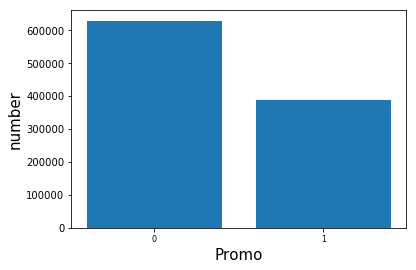
图表 1store表透视图组

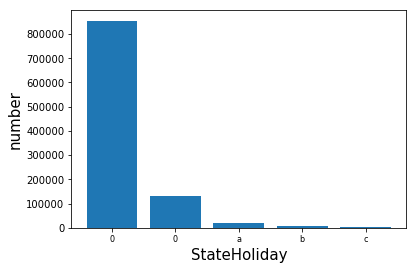
由上述组图可以看出，超过半数以上的店铺类型为a型，店铺的货品类型也集中在a和c类，参与和不参与持续促销的店铺数量基本持平，持续促销通常选择在每年的一月、四月、八月和十月进行。竞争对手开业最多的年份是2013年，很多选择在9月开业，竞争对手的距离大多集中在0-30000米的范围，极少数店铺的竞争对手距离会超出40000米。

1. 对train表的可视化分析

对train表的数据进行统计分析，结果如下：

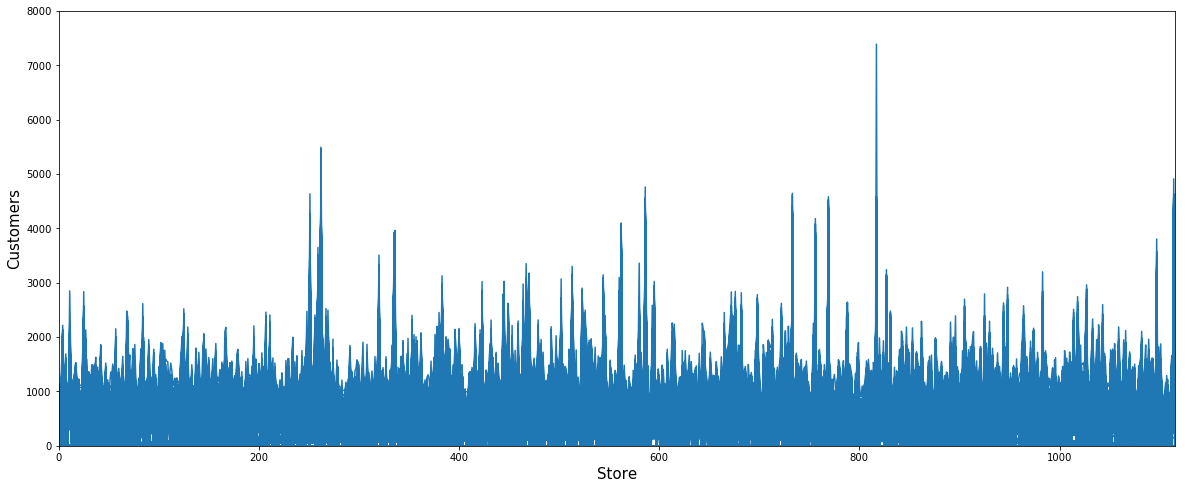


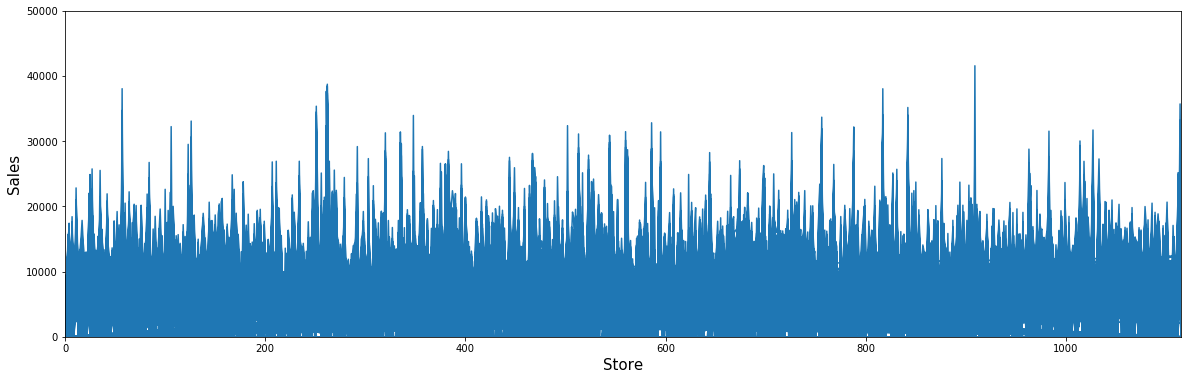




图表 2train表的数据透视图

train表中的离散数据分布如上图所示，在样本时间范围内，统计的dayofweek呈均匀分布，大部分店铺都处于开业状态，促销天数未超过统计周期的一半。学校假期及公立假期的占比都比较低，且在统计公立假期时，对于0类型出现了两类，可能是由于数据类型不统一导致，需要在数据处理时做类型转换。

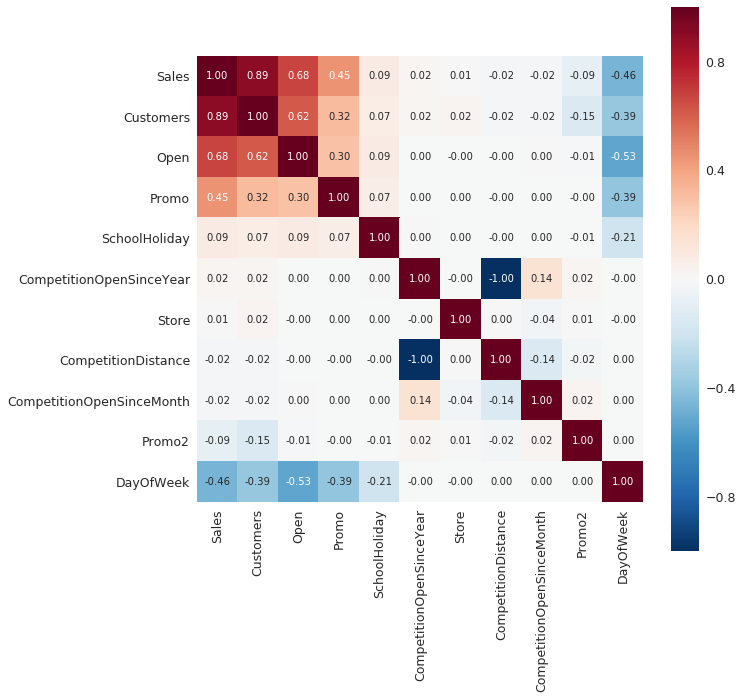




图表 3sales和customers分布图

结合上图可知，每家店铺的销量数据大多集中在20000以下，客户数集中在2000以下，分布较为均匀，且部分店铺销量的高升同客户数的增加变化一致，这个现象符合常识，customers是预测sales的重要特征。

### 多变量分析



图表 4变量相关性分析

如上图，相关性矩阵所示，与销售额呈正相关的特征变量主要有客户数、是否开业、是否促销，而星期数和是否持续促销则与销售额呈负相关。此外，可以发现，与销售额相关性较高的几个特征变量相互之间的相关性也比较高。星期数和客户数、是否开业、是否促销也有明显的负相关。因此，customers、open、promo、schoolholiday、dayofweek应该是影响sales的重要特征变量。

### 数据清洗

在进行多变量数据分析之前，先将原始数据进行清洗转换。

1. 对store表进行缺失值处理，CompetitionDistance属性有3个缺失值，且缺失值对应的CompetitionOpenSinceMonth、CompetitionOpenSinceYear也都为nan，为了减少异常值干扰，我们采用0来填充。CompetitionOpenSinceMonth、CompetitionOpenSinceYear两个属性的缺失值较多，针对CompetitionDistance不为nan的情况，我们采用众数填充二者的缺失值。Promo2SinceWeek、Promo2SinceYear两个属性的缺失值过半，采用丢弃处理。
2. 两表融合，train表没有缺失值，我们将train表与缺失值处理后的store表以store属性进行关联，构造融合了全部店铺信息的训练数据集。

## 算法和技术

常见的机器学习算法有如下几种：

1. 逻辑回归

逻辑回归也叫对数几率回归，其原理是用线性回归模型的预测结果来逼近真实标记的对数几率[3]。该算法能直接对分类可能性进行建模，无需事先假设数据分布，避免了假设分布不准确所带来的问题，而且不仅能预测类别，还能得到预测类别的近似概率。但是，模型对参数敏感，更适合高维特征空间，在低维空间中准确度不高、泛化性能不如其他模型。

1. 支持向量机

支持向量机算法的原理是找到一个超平面能够使得任意两个异类样本之间的间隔最大化。该算法在低维和高维的数据上都能有很好的表现，但对数据预处理的要求高，模型效果依赖调参，而且在大量数据时比较消耗时间和内存。

1. 决策树

决策树由根节点、若干内部节点、若干叶节点构成。叶节点对应决策结果，其他每个节点对应一个属性测试，算法通过遍历所有可能的测试，找到信息增益最大的数据划分模式。决策树模型容易理解，数据划分不依赖于缩放，对数据预处理要求低。不足之处就是容易过拟合，泛化性能差。

1. 随机森林

随机森林本质上是多个决策树的集成，通过随机有放回的抽样来构造过个预测性能良好的决策树，然后通过取均值或投票的方式得出最终结果。随机森林比单颗决策树更能从总体把握数据特征，能够防止过拟合，但算法效果依赖调参，在大型数据集上比较耗资源，对高维稀疏数据不友好[4]。

1. XGBoost

XGBoost本质上也是一种集成模型，但是集成方式与上述随机森林算法不同，不只是求评均或投票，而是要求每一颗集成进来的决策树都能对最终效果有提升。

## 基准模型

# 方法

## 数据预处理

## 执行过程

## 完善

# 结果

## 模型评价与验证

## 合理性分析

# 项目结论

## 结果可视化

## 对项目的思考

## 需要作出的改进

### 解决方案

首先观察数据的原始特征，根据数据特征做数据清洗、融合等工作，然后进行数据探索，通过可视化的工具查看数据了解数据特征，根据数据特征进行必要的数据预处理。

将处理好的数据分割成训练集和验证集，并根据预测目标为模型选择合适的评价指标，参照题目可采用“均方根百分比误差（rmspe）”这个指标来衡量模型优劣。

根据训练数据的特征、维度、预测目标等选择合适的模型范围进行模型测试，可以考虑构造模型测试流水线进行模型选择。部分带有特征排序或选择的模型可以我们特色优化提供思路，帮助我们调整数据特征。

根据选择模型的实际情况，结合rmspe得分，进行模型调优，并将结果上传至kaggle提交页面，检测模型结果，直至达到预期要求。

### 基准模型

本问题的最终目标是预测未来销售额，属于回归问题，解决回归问题可以考虑逻辑回归、SVR模型，如果模型效果不理想可以考虑采用集成学习的模型来实现预测目标。

### 评估指标

应题目要求采用rmspe指标来评价模型的预测效果，公式如下：

其中，n为样本数量。

### 方案设计

第一步：识别问题，剔除无效数据，根据数据的基本情况进行清洗和填充。

第二步：对训练集数据进行特征和标签的分离，并利用统计方法和线箱图、散点图等可视化方法进行数据探索，观察数据分布和统计特征，并结合原始数据情况加工新的特征。

第三步：根据数据特征进行数据预处理，对高偏度的特征进行转换、根据特征的数据范围进行数据缩放，并将类别特征进行one-hot编码转换。

第四步：将数据按照随机选择的方式进行训练集和验证集的划分，构造模型训练的流水线，并定义模型平均指标。

第五步：利用模型训练流水线，采用逻辑回归、SVR、XGboost等模型训练数据，并用验证集评估效果，根据feature\_importance或feature\_selection等属性，辅助进行特征的选择和优化。

第六步：结合第五步的结果，对选定模型和特征进行调参优化，直至rmspe满足kaggle排序要求为止。

### 参考文献

[1] Hongyu Xiong (hxiong2), Xi Wu (wuxi), Jingying Yue (jingying), Drugs store sales forecast using Machine Learning.

[2] 特征选择<https://www.cnblogs.com/stevenlk/p/6543628.html>.

[3] 监督学习项目2：为CharityML寻找捐献者

<https://classroom.udacity.com/nanodegrees/nd009-cn-advanced/parts/e340f873-83cc-44aa-8530-bb78bf834995/modules/d1dd6eea-c2a7-421d-b1ce-7702c44aab31/lessons/9eccaf6a-09e5-4fdb-a32f-17d942669bc3/concepts/8465cfb8-6864-4ad0-b945-626c4297d0ed>

<https://blog.csdn.net/pipisorry/article/details/52574156>

<https://www.jianshu.com/p/62716b33e7be> 统计分析借鉴

<https://www.cnblogs.com/majimaji/p/10265242.html> 数据可视化借鉴

<https://blog.csdn.net/github_38414650/article/details/76061893> xgboost原理

<https://blog.csdn.net/huacha__/article/details/81029680>

# 参考文献

1. <https://blog.csdn.net/tox33/article/details/81141485>
2. <https://www.cnblogs.com/duye/p/8862666.html>
3. 周志华.机器学习[M].清华大学出版社.2016:57-58.
4. AndreasC.Muller,Sarah Guido.Python机器学习基础教程.人民邮电出版社.2018:67-68.

# 表目录

[表格 1 store表中字段的统计信息 5](#_Toc535416936)

[表格 2 train表的统计信息 6](#_Toc535416937)

# 图目录

[图表 1store表透视图组 8](#_Toc536533954)

[图表 2train表的数据透视图 9](#_Toc536533955)

[图表 3sales和customers分布图 10](#_Toc536533956)

[图表 4变量相关性分析 11](#_Toc536533957)