

41

3



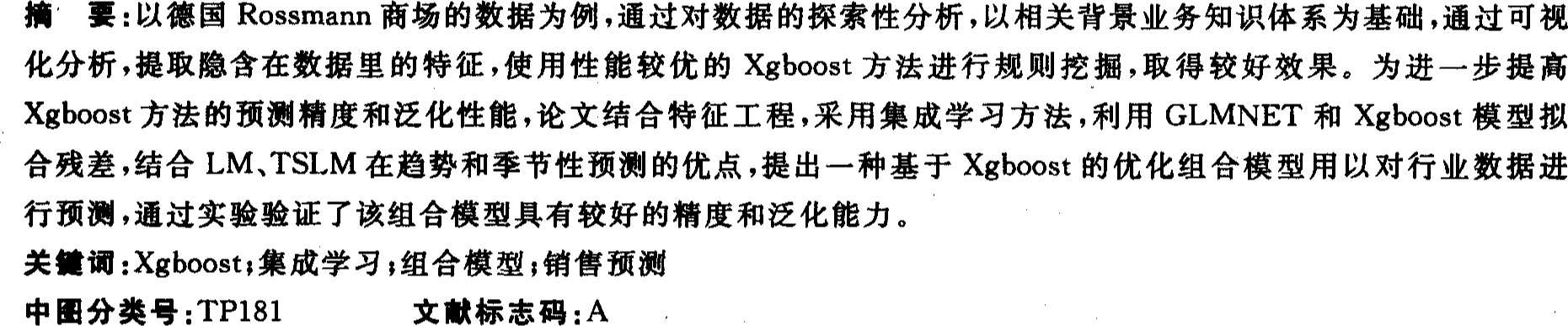
 Xgboost



,

,

330031)



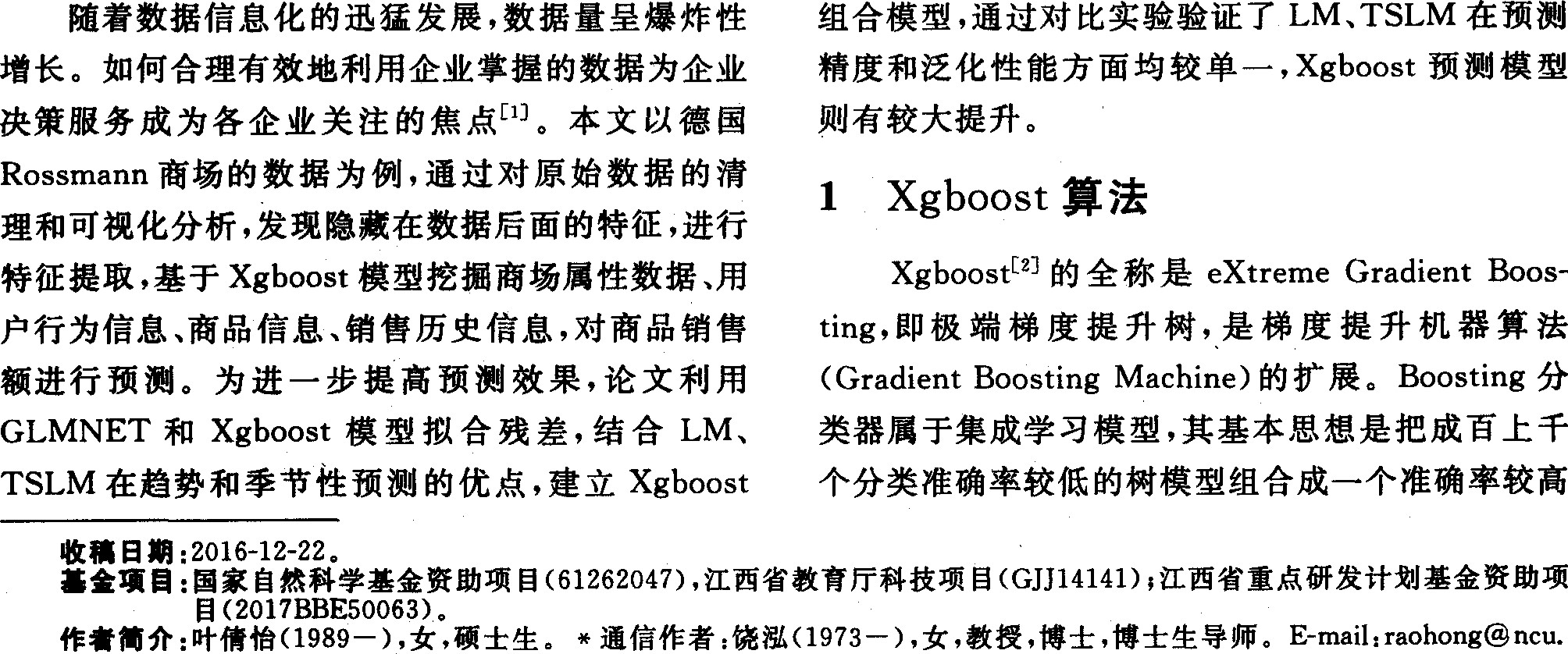
Sales prediction of stores based on xgboost algorithm

YE Qianyi,RAO Hong,J1 Mingshu

(Center of Computer, Nanchang University, Nanchang 330031 ,China)

Abstract : Taking the data from Rossmann mall in Germany as an example, this paper extracts some characteristics inherent in the data which based on the exploratory data analysis and the knowledge of the related business background, and selects the optimal performance of Xgboost method for rule mining. In order to further improve the prediction accuracy and the generalization performance of Xgboost method, this paper proposes an optimal combination model of Xgboost by combining with feature engineering, adopting integrated learning method, and utilizing the GLMNET and Xgboost models to fit the residuals. Based on the advantages of LM and TSLM in trend and seasonal prediction, an optimal combination model of Xgboost is proposed to forecast the industry data. The experiment results show that the combination model has good precision and generalization ability.

Key words: Xgboost ; Integrated learning ; Combination Model ; Sales Prediction



edu. cno

。276 · （

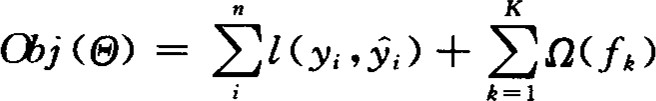
的模型。该模型不断迭代，每次迭代生成一棵新的树，如何在每一步生成合理的树是Boosting分类器的核心。Gradient Boosting Machine算法在生成每一棵树的时候采用梯度下降的思想，以上一步生成的所有树为基础，向着最小化给定目标函数的方向前进。在合理的参数设置下，需要生成一定数量的树才能达到预期准确率，在数据集较大较复杂的时候Gradient Boosting Machine算法的计算量巨大。 Xgboost是Gradient Boosting Machine的实现，能自动利用CPU的多线程进行并行，并对算法加以改进以提高精度。Xgboost的基学习器既有树（gb一 tree）又有线性分类器（gblinear），从而得到带Ll+ 1-2惩罚的线性回归或逻辑回归，其损失函数采用二阶泰勒展开，具有高准确度、不易过拟合、可扩展性等特点，能分布式处理高维稀疏特征，因此在同等情况下，Xgboost算法比同类算法快10倍以上。

1. 1决策树集成分类器

决策树表示方法由于其直观的表现形式和可靠的算法依据，在分类算法中得到了广泛的应用。但单棵决策树预测性能较为限，因此一般使用决策树集成(tree ensemble)来提高性能[3]

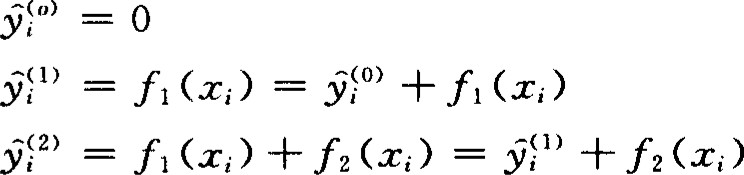
给定数据集D：{（，勢月( l D |一e Rm e R），树的集成模型用式（1)表示：

 习（），A e F （I) 式（1)中，F一{不工）：惚q囝这q：Rm—>T，w e RT）是回归树的集合空间。表示第个数据点的特征向量，q表示每一棵树的结构映射到样本所对应的叶子的索引，T表示树上叶子的数量，每一棵对应一个独立的树结构q和叶子的权重目标函数包含两部分：

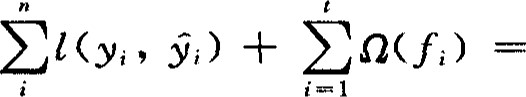
（2）式（2）中，第一部分1是预测值和目标真实值勢之间的训练误差，第二部分是每棵树的复杂度之和，即O(f)一YT +一丨W尸

1 · 2梯度树提升（Gradient Tree Boosting》

由于公式（2）的集成决策树模型中的目标函数无法用欧式距离、SGD随机梯度下降（Stochastic Gradient Descent)等传统方法优化，因此论文采用 additive training(boosting)方法训练，即每一次都是在保留原有模型的基础上，添加一个新函数到模型中去。



是第个样本在第'轮的模型预测值，立'）保留'一1轮的模型预测值立艹后，加人一个新的函数（）。选择在每一轮加人新函数是为了尽可能的让目标函数最大程度的减小。于是，改写目标函数为：





习I（，立。））十！））+ 0(ft ) +constant

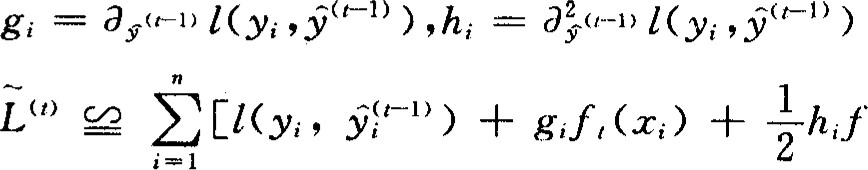


通过来优化这个目标函数。当误差函数1是平方误差的情况下，目标函数可以写成：

L(t)：习[ 2（立。 一）陆D十(x，）2 ] + （）+ constant

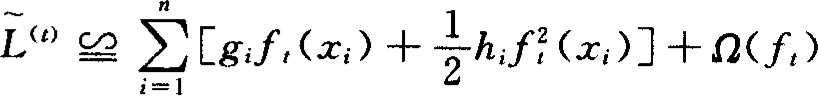
对于非平方误差的其他误差函数的情况，采用

泰勒展开来近似定义目标函数。



气荪月+Q(ft) + constant

把常数项移除后，获得了相对统一的目标函数，如式3所示。



（3）

# 2基于Xgboost的预测模型

2 · 1 Xgboost模型参数

Xgboost模型有3种类型的参数：通用参数、辅助参数和任务参数。通用参数确定上升过程中上升模型类型，常用树或线性模型；辅助参数取决于所选的上升模型爹任务参数定义学习任务和相应的学习目标[2]

Xgboost模型中，常用参数说明如下：

(1) booster：设置需要使用的上升模型。可选 gbtree(树）或gblinear(线性函数），默认为gbtreeo （2）nthread：Xgboost运行时的并行线程数，默认为当前系统可以获得的最大可用线程数。

|  |
| --- |
| · · |

1. eta：收缩步长，即学习速率，取值范围是，默认为0· 3。在更新叶子节点的时候，权重乘以 eta，以避免在更新过程中的过拟合。
2. max一depth：每棵树的最大深度，取值范围是，默认为6。树越深，越容易过拟合。
3. subsample：训练的实例样本占整体实例样本的比例，取值范围是（0，1 ]，默认为1。值为0 · 5 时意味着Xgboost随机抽取一半的数据实例来生成树模型，这样能防止过拟合。
4. colsample-bytree：在构建每棵树时，列（特征）的子样本比，参数值的范围是（0，1 ]。
5. objective：默认为reg：linear;
6. seed：随机数种子，为确保数据的可重现性，默认为0。

2 · 2 K折交叉验证方法（K一CV }

论文采用K折交叉验证方法[4]。将原始数据分为K个子集，每个子集分别验证一次，剩余的K 一1组子集作为训练数据，这样可得到K组训练集和测试集，以最终的分类平均精度作为性能指标。在实际应用中，K值一般大于或等于2，需要建立K 个模型来进行K折交叉验证的实验，并计算K次测试集的平均辨识率。

K折交叉验证的结果能较好说明模型效果，有效地避免欠拟合与过拟合。在Xgboost中，通过 xgb. cv函数来做交叉验证。

2 · 3独\*e\*ß(One—hot Encoding)

由于Xgboost仅适用于处理数值型向量[5]，因此处理训练集和测试集时需要将所有其它形式的数据转换为数值型向量，本文采用独热编码将特征值转换为数值。

独热编码也称一位有效编码[6〕，即对于任意时间任意给定的状态，状态向量中只有一位为1，其余各位为0，将n类特征值转化成n位二进制数串，将特征的每个对应类设置为1。独热编码将每一个特征的n个取值通过独热编码后转换成了”个二元特征，通过该方法将特征转变成稀疏矩阵[6]。独热编码能够解决分类器不好处理属性数据的问题，并在一定程度上扩充了特征[7]。 

# 3数据预处理

为了获取数据中的有效特征，论文采用探索性数据分析方法，对数据进行可视化分析，获得数据分布特征，理解原始数据的基本特征，发现数据之间的潜在模式，找出数据中的有效特征。

3 · 1数据来源

本文所有数据均来自Kaggle中的Rossmann 商店销售额数据集。Rossmann是商人Dirk Rossmann创立的德国首家平价日用品商店，现在的 Rossmann公司遍布欧洲7个国家，分店达100多家。论文通过位于德国的1115所Rossmann连锁商店的历史数据预测未来48天商店的销售额。

Kaggle给出了三个数据集：train、test、store，分别是训练集、测试集和商店基本信息的数据集，对训练集建模训练，对测试集进行预测。

数据集基本信息如下：

（1)训练集train· csv：时间范围为2013年01月 01日到2015年07月31日，共942天，1017209条数据。

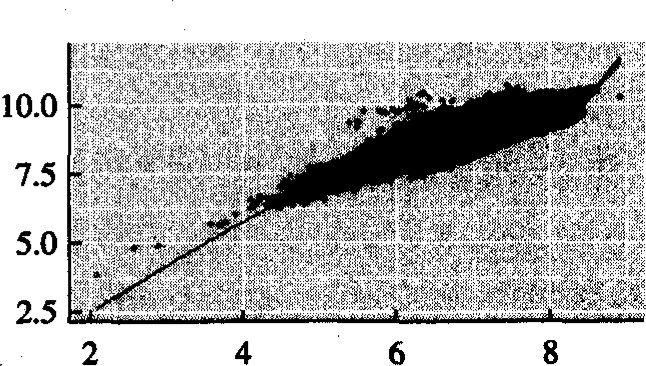
1. 测试集test. csv：时间范围为2015年08月 01日到2015年09月17日，共48天，41088条数据。
2. 商店基本信息数据集store. csv: 1115条数据，共1115家商店的信息。

3 · 2数据的可视化分析及原始特征提取

为了获取影响销售额的基本数据特征，论文对 Kaggle提供的Rossmann数据集进行了可视化分析，提取原始特征集。

(1)顾客数和销售额之间的关系

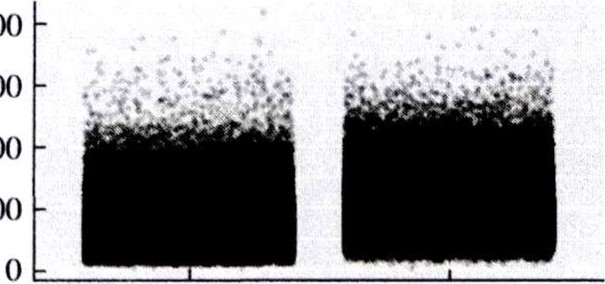
从图1中可以看出，顾客数和销售额之间存在紧密的正相关关系。 

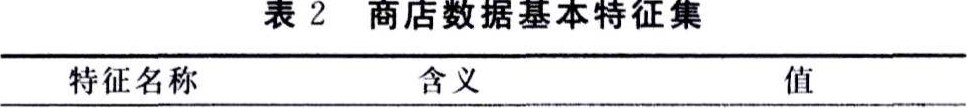


顾客数

图1顾客數与销售額的关系曲线

1. 促销对销售的影响的可视化分析图2和图3中promo取1表示当天有促销活动，取0表示没有促销活动。从图中可以看出促销活动对顾客数并没有太大影响，但销售量却明显提高了，即促销活动并没有吸引更多的顾客，但提高了顾客的购买力，从而提高了销售额。通过查询原始数据发现，没有促销的情况下顾客平均消费8 · 94欧元，有促销活动的情况下平均消费10 · 18欧元。
2. 星期（DayOfWeek）对销售影响的可视化分析

· 2 78 · 〈

40仪

30 000

Sto re 商店号 取值：1到1 1巧

20 000

StoreType 商店类型

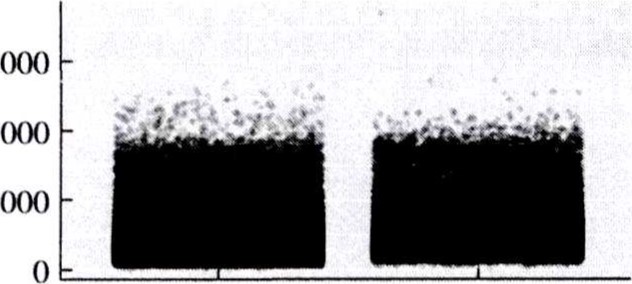
10 000a、基、屈类、补

Assortment 商店类别充 、c扩展类

CompetltionDts 最近的竞争对手的距 数 ，1 115个商店 0 1 离 里有761个商店有 tance promo Competition

竞争对手开张的月份 月份：1一12 图  促销对销售额的影响 ()penSinceMonth

Competitton

 竞争对手开张的年份 年份数据

()penSinceYear

一

6

Prom02 持续性的促销活动 0：无，1：有

．

4开始参加Prom。2促销 日历上的第几周，数

Prom02SinceWeek

的日历周 值

0-

开始参加Promo?促销

Prom02SinceYear 年份数据

的年份

0 1

Jan，Apr，Jul'()ct"，

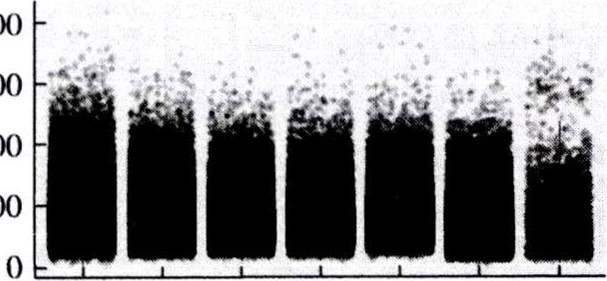
参加Prom02促销开始

promo Promolnterval "Feb，May，Aug，Nov'，

的月份列表

图  促销对顾客数的影响 "Mar，Jun，Sept，Dec"

从图中可以看出，星期日的销售额中位数和上四分位数远远高出正常营业日，但是下四分位数却低于正常营业日，即有蚱商店星期日营业额高于平时，但也有一部分商店的销售额低于正常营业日。从图中也可以看出，从周一到周六，周一的销售额偏高一蚱，星期六的销售额偏低一些

40（××） 30仪过20 000

10仪×）

1 2 3 4 5 6 7 星期几

图4销售额在星期1、7中的分布情况

由于篇幅的关系，对数据的可视化分析不一列出。根据这些数据的可视化化析，我们提取出如表1所示和表2所示的训练数据和测试数据原始特征集以及商店数据基本特征集。

表1数据原始特征集

|  |  |  |
| --- | --- | --- |
| 特征名称 | 含义 |  |
| Store  Day()fWeek  Date  Sales  Customers  ()pen p rot-no  StateHoIiday  SchoolHoIiday | 商店号星期几时间销售额顾客数是否开店  当天是否有促销  假日  学校假日 | 取值：1到1115 取值：1到7  如'2013一01一01' 数值数值   * ：关店，1：开店 * ：无促销，1：促销 * ：非假日，a：公共假日；b：复活节，c：圣诞节 0：非假日，1：假日 |

不同的分类特征在训练集和测试集中的比例见表3、6。

表3 ()pen特征取值比例

|  |  |  |
| --- | --- | --- |
| ()pen | 0 | 1 |
| train/% test/％ | 16．99  14．56 | 83．01  85．44 |
| 表4 | Promo特征取值比例 |  |
| Promo | 0 | 1 |
| train，％ test/% | 61，85  60．42 | 3 & 15  39，58 |

表5 StateHoIiday特征取值比例

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| StateHoIiday | 0 |  | b |  |
| tram 0  test | 96．95  99，56 | l . 99 0，44 | o，66 | o．10 |

表6 SchoolHoliday特征取值比例

|  |  |  |
| --- | --- | --- |
| SchoolHoIiday | 0 | 1 |
| train/% test/% | 82，14  55．65 | 17，86  44．35 |

3 · 3数据预处理

3 · 3，1数据清洗为获得可训练用数据，我们对原始数据进行清洗，具体过程如下：

（1)标记异常数据。如商店是开门的，但是销售额为零的数据为异常数据；

1. 对训练集的Sales销售额字段取对数，设置为SalesLog字段；
2. 缺失值用一1填充；
3. 合并训练集和测试集，添加set字段，用以区分训练集和测试集，值1为训练集，0为测试集；
4. 数值化分类特征值。原始数据集中，state一

Holiday分类特征取值为0、a、b、c，无法代人模型计



算，因此重新编码为0、1、2、3；st。reType分类特征值为a、b、c、d,Assortment分类特征值为a、b、c，采用同样方法用整型数据重新编码；

1. 分解特征。将原始数据集中Date特征分解为DateDay、DateWeek、DateMonth、DateYear、 DateDayOfYear 5个特征。

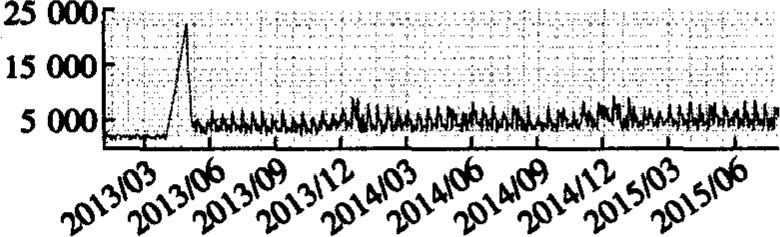
〈7）增加字段Datelnt，即Date转换为整型的形式。

 （8）规范化特征表达。对CompetitionOpenSinceYear和CompetitionOpenSinceMonth字段合并成普通年月的表达，并转化为整型；将Prom02SinceYear和Prom02SinceWeek字段合并成普通年月的表达，并转化为整型，增加字段

Prom02SinceInt0

（9）删除偏差大于2 · 5的异常数据。

〈10）删除存在异常的数据点，如图5所示的异常数据。



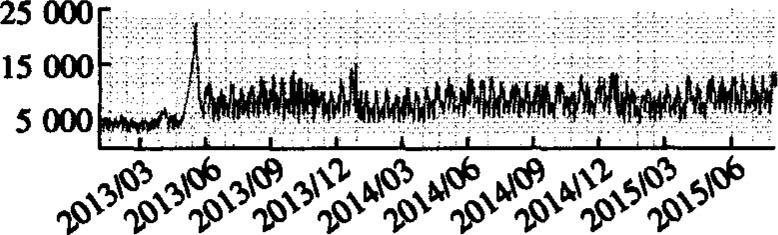


图5异常数据

3 · 3 · 2特征处理论文根据相关信息背景对数据进行特征处理：

（1)增加字段CompetitionOpenInt

1. 添加一些额外的特征，如商店的位置特征 Storestate、发薪日效应特征PayDay(在一个月的第一个工作日设置特征值为3，在随后的两天设置为 2、1)。
2. 增加商店平均每天的销售额SalesPerDay、平均每天的顾客数CustomersPerDay、平均每天每位顾客的销售额SalesPerCustomersPerDay作为新的特征。
3. 增加特征组合，如：store：DayofWeek, store：DayofWeek：Promo等等。

# 实验结果及分析

4 · 1 实验条件

(1)软件环境：

## (a)操作系统:Windows 7 x64；

(b)开发平台Python2 · 7 + R3 · 4 · 2

(c）第三方库：Python：numpy + pandas + matplotlib+ xgboost；R 3 · 4 · 2库：data. table+ ggplot2+ lubridate+ zoo+ dplyr + scales + xgboost + forecast+glmnet。

（2）硬件环境：处理器：Inter(R) Core(TM) i3一4160 CPU @

3 · 60GHz 内存：8GB

4· 2单Xgboost模型对特征工程后的所有特征用Xgboost模型进行训练，参数如表7所示。

### 表7单Xgboostü型参数

值 nthread 6 eta 0· 01 seed 42 max—depth obJective reg :linear subsample nrounds 20000 colsample—bytree early. stop. round 250

4 · 3组合模型根据模型的作用，将模型分为三类：商店模式模型、数据合并模型、混合模型。

商店模式模型：下面采用单个模型是对每个商店进行单独拟合。这类模型关注商店各自的特性，但是它也错过了可在其他相似商店的模式中获取的信息。

(1)线性模型1m拟合趋势，不带特征交互的 Xgboost模型拟合残差。 

1. 线性模型1m拟合趋势，带特征交互的xgboost模型拟合残差。 
2. 线性模型1m拟合趋势，不同参数值的 glmnet模型拟合残差。 
3. tslm模型拟合趋势+季节性，Xgb st模型拟合残差。
4. tslm模型拟合趋势+季节性，glmnet模型拟合残差。
5. tslm模型拟合趋势+季节性，Xgboost模型+glmnet模型拟合残差。
6. 每个商店直接用Xgboost模型拟合残差。

|  |
| --- |
| · 280。 （ |

这些模型中，含有Xgb st模型的模型参数设置如表8，其中模型7中参数mounds值设为500。

表8组合Xgboost的模型参数

|  |  |  |  |
| --- | --- | --- | --- |
| 参数 | 值 | 参数 | 值 |
| objective nrounds colsample—bytree | reg：linear  300 | eta max—depth subsample | * ，01 5 * ．95 |

数据合并模型：下列模型使用合并有store数据集和train训练集的数据集。各个模型的参数设置如表9一1 1。

表9没有特征工程的Xgboost模型

|  |  |  |
| --- | --- | --- |
| 参数 | 值 参数 | 值 |
| objective nrounds colsample—bytree | reg：linear eta  3000 max—depth  0· 95 subsample | o，01 10  0· 95 |
| 表10 | 有特征工程的Xgboost模型 |  |
| 参数 | 值 参数 | 值 |
| 0b) ective booster nrounds early. stop. round | reg：linear eta gbtree max—depth  4000 subsample  100 colsample—bytree | o，02  10 |

表11 153个特征+特征交互的Xgboost模型



参数 值 参数 值



|  |  |  |  |
| --- | --- | --- | --- |
| objective | reg：linear | eta | 0· 015 |
| booster | gbtree | max—depth | 18 |
| nrounds. CV | 3000 | subsample |  |

nrounds. 1b 270 colsample—bytree early. stop. round 100



组合模型：通过组合数据合并模型来获取跨多个商店之间共同的特质。模型的残差用商店模式模型中的模型来拟合，从而获得每个商店的特质。用 prcomp函数从数据中提取50个主成分，并用Xgboost模型来拟合、计算残差。模型说明如下：

1. 使用线性模型1m拟合趋势，带特征交互的 glmnet模型+Xgboost模型拟合残差。
2. 使用tslm拟合趋势和季节性，带特征交互的Xgboost模型拟合残差。

图6给出模型在训练集和验证集上的RMSPE 的变化过程，横坐标是次数，纵坐标是RMSPE的值A是人工删除了异常点的数据，B中的异常数据并没有删除。可以发现在训练集上训练的前500次 RMSPE的值就迅速的降低到0 · 2，然后在1 750次就在0· 1左右了，可以看出在训练集上效果很好。

邏1 5 eA.traln—rmspe —B.train—rmspe

1.0 \*A.validation—rmspe+B.validation—n

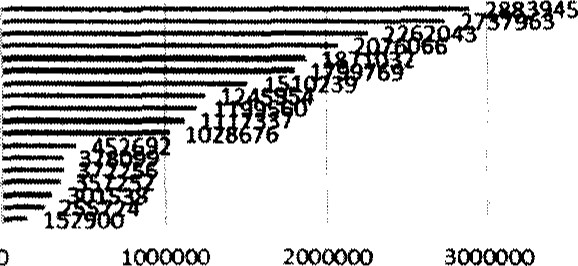
o巧

0 1 5闐3〕4 5 6 7 500 9閬0 RMSPE

图6在训练集train和验证集validation上的RMSPE值

图7给出了Xgboost模型中非组合特征的重要度得分，从图中可以看出时间类特征和有关竞争对手的特征得分非常高，这意味着这些特征对模型具有非常大的影响。

DateWeek CompetitionOpenint SalesPerDay Promo ScY001HOIiday Assortment



〕

F score 特征重要度

图7特征重要度得分

为了对比各模型的泛化效果，我们给出了单类模型（线性模型LM、时间序列线性模型TSLM、基于Lasso和Elastic Net正则广义线性模型glmnet、极端梯度上升模型Xgboost）和它们之间的组合模型在测试集上RMSPE值，从而评价模型在测试集上的泛化能力。如表13所示。

表13各模型RMSPE值比较

|  |  |
| --- | --- |
| 模型 | RMSPE值 |
| LM +简单的特征处理  TSLM +简单的特征处理 random化rest+简单的特征处理 glmnet +简单的特征工程 Xgb st十简单的特征工程  Xgboost+特征工程  Xgboost+glmnet+特征工程  Xgboost+glmnet+tslm+lm+特征工程 | 0 · 206 57  0· 207 59  0 · 127 51  0· 1 19 7 4  o．1 18 3 9  0 · Il 2 68   * ．Il 2 62 * ．1 1 1 46 |

从表中可以看出，Xgboost单模型的能力就优于其它模型，在进行了简单特征工程后，Xgboost 和glmnet模型相比，Xgboost模型依然效果更好。最后我们结合TSLM + LM在处理趋势和季节性上的优势，采用集成方法对Xgboost+glmnet进行组合得出优化模型Xgboost+glmnet+tslm+lm+特征工程，RMSPE值得到较大提升，泛化性能最优。

# 5结论

|  |
| --- |
| ． · |

本论文研究基于Xgboost方法对实体零售业销售额进行预测。论文以德国零售业Rossmanns公司1115家实体门店的商场信息和销售数据为数据源，采用Xgboost方法对公司销售额进行预测。

通过在特征工程中对原始数据进行特征提取、选择和构建，筛选出用于训练的特征属性；对比xg“ boost、随机森林、GLMNET以及LM、TSLM模型等不同方法对销售额的预测结果，表明Xgboost方法无论是训练速度还是在RMSPE评价标准上都具有明显的优势。

为了进一步提高Xgboost预测模型的精度和泛化能力，本文通过大量的特征工程，尝试多种模型的集成学习方法和参数调优，利用GLMNET和Xc b。ost模型拟合残差，并结合LM、TSLM在趋势和季节性预测的优点，获得组合优化模型。实验表明该组合模型在性能上优于单一Xgboost预测模型。

这种基于Xgboost的组合模型不仅适用于对德国零售业销售额的预测，还可以将此方法应用于国内零售实体业甚至电商平台的销售额预测，对于提高商店的运营生产模式、日常管理、价格管理、配送方式及精准营销具有重要的意义。

## 参考文献：

[ 1〕 赵啸彬·基于数据挖掘的零售业销售预测[ D] ·上海: 上海交通大学，2010，

[2 ] CHEN T, HE T. Higgs Boson Discovery with Boosted TreesCCJ. JMLR：Workshop and Conference Proceedings，2015，4 69一80 ·

[ 3 ] ROBERT E. Banfield，Lawrence O. Hall，Kevin W. Bowyer，W. P. Kegelmeyer，A Comparison Of Decision Tree Ensemble Creation Techniques CJ ] · IEEE Transactions on Pattern Analysis and Machine Intelligence， 2007，29（1L173一180·

[ 4 ] 李航·统计学习方法[ M].北京：清华大学出版社，

2012·

[ 5 ] 闻玲·移动平均季节模型在商品销售收人预测中的应用CJJ. Market Modernization，2Q10（28）：43一45，

[ 6 ] 黄伟，陶俊才·一种基于k一means聚类和关监督学习的医学图像分割算法[J].南昌大学学报（理科版）， 2014，38（1)：31一35·

[ 7 ] RICE J. Mathematical Statics and Data AnäIysisC M]. Second Edition，Plymouth：Duxbury Press，2006：221 ·

225 ·