** **

**SHANGHAI UNIVERSITY**

毕业设计（论文）

**UNDERGRADUATE PROJECT (THESIS)**

**题 目:** **基于TL-XGBoost算法的快消品**

**天气敏感型需求预测模型研究**

WEATHER-SENSITIVE DEMAND FORECASTING MODEL FOR FAST MOVING CONSUMER GOODS BASED ON TL-XGBoost ALGORITHMS

**学 院** 悉尼工商学院

**专 业** 工商管理

**学 号** 15124103

**学生姓名** 董剑铭

**指导教师** 高峻峻

**起讫日期2019.3.4-2019.6.21**

目 录

[摘要 4](#_Toc9257787)

[ABSTRACT 6](#_Toc9257788)

[第一章 绪论 8](#_Toc9257789)

[1.1论文的研究背景 8](#_Toc9257790)

[1.2论文的研究目的及意义 8](#_Toc9257791)

[1.3文献综述 10](#_Toc9257792)

[1.4论文拟解决的关键问题 14](#_Toc9257793)

[1.5论文的主要研究内容及框架 15](#_Toc9257794)

[1.6可行性分析及创新点 16](#_Toc9257795)

[第二章 基本理论 18](#_Toc9257796)

[2.1 天气敏感型需求理论 18](#_Toc9257797)

[2.2 Boosting Tree基础算法 21](#_Toc9257798)

[2.3 XGBoost算法 24](#_Toc9257799)

[2.4 TrAdaboost.R2算法 25](#_Toc9257800)

[第三章 基于TL-XGBoost的天气敏感型需求预测模型 26](#_Toc9257801)

[3.1引言 26](#_Toc9257802)

[3.2问题描述 26](#_Toc9257803)

[3.3天气变量的刻画与选择 27](#_Toc9257804)

[3.4天气敏感型需求预测模型建立 32](#_Toc9257805)

[3.4.1 Two-stage TrAdaboost.R2模型建立 32](#_Toc9257806)

[3.4.2 XGBoost模型建立 34](#_Toc9257807)

[3.4.3 GAD模型建立 35](#_Toc9257808)

[3.5 模型验证 37](#_Toc9257809)

[第四章 实证分析 38](#_Toc9257810)

[4.1数据收集 38](#_Toc9257811)

[4.1.1 数据来源 38](#_Toc9257812)

[4.1.2 描述性统计 38](#_Toc9257813)

[4.2数据预处理 39](#_Toc9257814)

[4.2.1 缺失值处理 39](#_Toc9257815)

[4.2.2 数据转换 40](#_Toc9257816)

[4.3特征工程 42](#_Toc9257817)

[4.3.1 特征构建 42](#_Toc9257818)

[4.3.2 基于机器学习模型的特征选择 44](#_Toc9257819)

[4.4结果分析与模型对比 46](#_Toc9257820)

[4.4.1 城市维度需求预测结果分析 46](#_Toc9257821)

[4.4.2 SKU维度需求预测结果分析 48](#_Toc9257822)

[第五章 结论与展望 50](#_Toc9257823)

[5.1 研究结论 51](#_Toc9257824)

[5.2 研究局限 51](#_Toc9257825)

[5.3 管理启示 52](#_Toc9257826)

[5.4 未来展望 53](#_Toc9257827)

[致 谢 54](#_Toc9257828)

[参考文献 55](#_Toc9257829)

[附录： 59](#_Toc9257830)

摘要

互联网的发展对人们的消费习惯产生了巨大影响，各行各业电商化、数字化的趋势愈发显著。随着“互联网+”的深入人心，企业对数据的重视程度上升到全新的高度。与此同时，近年来诸如“新零售”等概念在国内的兴起，云计算、大数据等前沿技术的应用，打破了线上线下两端的边界。快消品行业，是和人们的日常生活关系最为密切的行业之一。从2010年度到现在，快消品行业迅速发展，丰富的渠道、全新的消费体验，为快消品行业带来了革命。随着行业的发展，一系列管理难题也随之出现，企业们也都逐渐意识到产品或服务的需求管理对于自身发展的重要性。除此以外，天气的异常波动也是影响快消品需求变化的重要原因之一。企业们也开始意识到需要将天气因素考虑到模型的构建中。因此，对于快消品行业的众多企业而言，如何应用大数据技术，如何更好的预测以及满足客户的需求成为了亟待解决的问题。

本文拟解决以下几个问题：第一，天气变量的刻画与选择；第二，天气对于产品替代性的影响；第三，机器学习模型在业界应用的局限；第四，天气敏感型需求预测模型的建立。本文以中国某咖啡连锁品牌的门店销售数据作为依据，结合实证分析的研究方法，对天气变量和需求之间的关系进行探索，将天气对于产品替代性的影响考虑进模型，同时引入迁移学习的思想来解决机器学习模型在业界应用的样本量以及样本分布等局限，建立了一个考虑天气敏感型因素的TL-XGBoost需求预测模型，为企业SKU维度的销量进行预测，并与包括XGboost、SVM、GBDT等机器学习模型进行对比验证。

本文提出的TL-XGboost天气敏感型需求预测模型分为三个部分，第一部分构建了TrAdaboost.R2模型在品类维度对总销量进行预测，第二部分是建立XGBoost模型在门店层级对SKU的销量和销售概率进行预测，第三部分是运用GAD贪婪聚合算法对SKU销量进行最终修正。最后，本文进行了模型对比和总结，结果表明，本文提出的需求预测模型不仅利用了天气信息中的价值，同时利用了供应链内部不同维度和层级的有效信息，将天气对于产品替代性的影响纳入考虑因素，模型充分的应用了机器学习算法在面对庞大数据时的出色性能特点，在预测精确度上高过传统的单层次XGBoost模型、GBDT模型和SVM模型。

综上，本文提出的TL-XGBoost模型可以有效的进行中短期的需求预测，为零售企业的库存计划、促销计划、定价策略等业务活动和决策提供有效的支持。

关键词：天气敏感型需求；机器学习；迁移学习；需求预测

ABSTRACT

The development of the Internet has had a tremendous impact on people's consumption habits. The trend of e-commerce and digitalization in all walks of life is more and more obvious. With the popularity of "Internet +", enterprises' attention to data has risen to a new level. At the same time, in recent years, concepts such as "new retail" have sprung up in China, and the application of advanced technologies such as cloud computing and big data has broken the boundary between online and offline. Fast moving consumer goods industry is one of the most closely related industries with people's daily life. From 2010 to now, the fast moving consumer goods industry has developed rapidly. Various channels and brand-new consumption experience have brought revolution to the fast consumer industry. With the development of the industry, a series of management problems arise. Enterprises are also gradually aware of the importance of product or service demand management for their own development. In addition, the abnormal fluctuation of weather is also one of the important reasons affecting the change of demand for fast moving consumer goods. Enterprises are also beginning to realize the need to take weather factors into account in the construction of the model. Therefore, for many enterprises in the fast consumer goods industry, how to apply big data technology, how to better predict and meet customer needs has become an urgent problem to be solved.

This paper aims to solve the following problems: firstly, the description and selection of weather variables; secondly, the impact of weather on product substitution; thirdly, the limitations of machine learning model in the industry; fourthly, the establishment of weather-sensitive demand forecasting model. Based on the sales data of a coffee chain brand in China, this paper explores the relationship between weather variables and demand with the method of empirical analysis. The influence of weather on product substitution is taken into account in the model. Meanwhile, the idea of transfer learning is introduced to solve the limitations of sample size and sample distribution in the application of machine learning model in the industry. TL-XGBoost demand forecasting model considering weather-sensitive factors is used to forecast the sales of SKU, and is compared with machine learning models including XGboost, SVM and GBDT.

The TL-XGboost weather-sensitive demand forecasting model proposed in this paper is divided into three parts. The first part constructs TrAdaboost.R2 model to forecast total sales in category dimension, the second part establishes XGBoost model to forecast SKU sales and sales probability at store level, and the third part uses GAD greedy aggregation algorithm to revise SKU sales finally. Finally, this paper compares and summarizes the models. The results show that the proposed demand forecasting model not only utilizes the value of weather information, but also utilizes the effective information of different dimensions and levels within the supply chain, and takes into account the impact of weather on product substitution. The model fully applies the excellent performance of machine learning algorithm in the face of huge data. The prediction accuracy is higher than the traditional single-level XGBoost model, GBDT model and SVM model.

In summary, the TL-XGBoost model proposed in this paper can effectively predict short-term and medium-term demand, and provide effective support for retail enterprises'inventory planning, promotion planning, pricing strategies and other business activities and decisions.

Keywords：Weather-sensitive demand；Machine learning；Transfer learning；Demand forecasting

**第一章 绪论**

**1.1论文的研究背景**

互联网的发展改变了人们的消费习惯，各行各业电商化、数字化的趋势愈发显著[1]。随着“互联网+”的深入人心，企业对数据的重视程度上升到全新的高度[2]。同时，近年来诸如“新零售”等概念在国内的兴起，云计算、大数据等前沿技术的应用，打破了线上线下两端的边界[3]。

快消品行业，是和人们的平常生活关系最为密切的行业之一，快消品的全称是快速消费品（Fast Moving Consumer Goods），通常指的是日常生活中那些使用寿命较为短暂，同时人们对其的消费速度又较快的产品。随着各项信息技术的全面发展，快消品行业数字化的趋势已经非常明显。以腾讯的智慧零售生态布局为例，微信支付、腾讯云、小程序、腾讯社交广告等产品作为工具箱，联合海澜之家、万达广场、永辉超市、家乐福等实体零售企业现有布局，实现互利共赢高速发展的合作模式，产生大量的用户与流量。大量的用户与流量，意味着线上线下会产生海量的数据，而耳熟能详的盒马鲜生、超级物种、7Fresh等店铺，都很早的走进了数据化的道路上[4]。

除此以外，国家统计局的数据表明，从2010年度到2017年度，我国的社会消费品零售总额从54553.7万亿上升到366261.7万亿，增长率达到237%。社会消费品零售对于我国的国内生产总值（GDP）的发展更是起了推动作用，对GDP的贡献从2010年的37.43%不断上升至2017年的44.29%。快消品市场的持续发展，特别是对于2016年后的市场回暖，“新零售”概念带来的众多场景渠道扮演了主要的角色。除了上文中提到的腾讯以外，阿里京东等企业也对零售的场景渠道进行了改造，从而为消费者带来了全新的消费体验，进一步加速了行业的发展。增加了行业的吸引力，提高了行业的效率。快消品行业的高速发展，离不开行业背后库存管理、促销计划、定价方案、物料管理等的支持，因此，对于快消品行业的众多企业而言，如何应用大数据技术，如何更好的预测以及满足客户的需求成为了亟待解决的问题。

**1.2论文的研究目的及意义**

需求预测指的是，对整个或者特定的产品或服务在未来的一段时间内的需求量以及需求金额进行一个估计。需求预测的目的是通过对历史数据、各种需求影响因素的充分考虑和利用，同时结合企业自身的实际运营情况和科学的分析手段，为企业提出符合其实际运营情况的需求目标，从而为企业的需求企业的定制、原材料的管理、商品的订货等活动提供支持。

天气敏感型需求是指产品或者服务的需求在本来的季节性波动基础上由于受到异常的天气因素的波动影响而发生的异常变化。比如，超出正常趋势的低温或者高温、过于少的降水量、极端天气等等所带来的产品或者服务需求的特殊上升与下降，或是产品或服务需求产生时刻的特殊提前与延迟[5]。

尽管国内部分企业意识到了天气敏感型商品的存在，但是大部分企业在需求管理上，特别是对于天气敏感型商品的的需求管理上仍然存在着许多问题。目前，零售商在天气敏感型商品的需求管理上已经采取了一些方法。比如第一种方法是刻意对天气对销售的影响进行忽略，比如保持较高的库存水平。这种做法的问题在于提高了库存成本。第二种方法是面对未来的需求，利用天气信息进行定性的预测。然而这种做法同样存在着问题，企业很难去制定有效的需求计划，因为不能定量的去预测未来销售受天气的影响程度，同时该种做法对决策者的主观判断极度依赖。

同时，尽管目前学术界正在逐步完善对于天气变量的刻画与选择，但是，关于从供应链角度出发如何对天气敏感型商品进行需求预测，目前还没有学者进行研究。实际的业界场景下，消费者对于同一品类下不同的产品之间的选择会受到天气因素的影响，具体来说包括天气状况、温度、温差在内的多种因素都可能对消费者的选择偏好产生影响，因此，对于企业来说，非常有必要去构建出一种天气敏感型需求预测模型。

此外，尽管机器学习模型在面对大数据时表现出了非常优异的性能，诸如精准度高训练速度快等优点。但是在实际工业界的应用，机器学习模型仍然存在一些问题，比如训练集是否充足、比如新旧样本的分布是否满足假设。目前来说，还没有学者对机器学习模型在天气敏感需求预测中的应用问题进行研究和讨论。

因此，本文的研究课题选择为基于TL-XGBoost算法的快消品天气敏感型需求预测模型研究，选取的研究对象为国内某连锁咖啡品牌所经营的四类产品。希望通过对上海地区这四类产品的销售数据和天气数据进行分析，充分挖掘天气变量和需求变化之间的关系，并结合迁移学习知识和机器学习算法，来获得一种合理的，能适用于工业界的需求预测模型，来辅助企业进行需求计划的定制，从而为其库存管理、促销计划、定价方案、物料管理等业务活动提供支持。

**1.3文献综述**

**（1）天气敏感性需求理论的发展**

回顾前人对天气敏感性需求理论的研究，在1983年，Persinger和Levesque[6]对人们心情与日照时数、相对湿度，地磁活性，温度，大气压，风速的关系进行了讨论，证明了人们的情绪会受到天气的影响。

同样的，Julie Travers[7]等人从心理学的角度证明了气候变化对顾客购买模式的影响，方法是通过调查问卷的形式调查消费者一年的季节性衣服分配以及采购情况。

除了心理学的角度以外，其他领域也逐渐证明了天气影响人们的行为模式，以及消费行为的改变对各领域的影响。比如， Saunders[8]和Trombley[9] 在金融分析领域，证明了天气对股票交易的影响。Kyle B. Murray[10] 在营销分析领域，证明了天气对人们消费选择的影响，通过对茶叶实验的数据进行分析。

除此以外，不断有研究与数据证明天气与商业，特别是销售之间的密切关联。例如英国气象局的数据分析部门调查表明，天气预测信息在英国的服装以及食品行业中的应用，使得其每年销量额的增额在4.5亿元左右[11]。

Kincade[12]等也在针对销量与天气的研究中对二者之间的高度相关性进行了验证。这些早期的天气敏感性需求理论的研究，证明了天气对人们行为模式的影响以及天气在商业端应用的价值。然而，这些早期的研究没有提出具体的预测方案，没有系统性的去量化不同商品对于天气变化产生的需求变化。

在证明了天气对人们行为模式的影响以及商业端应用的价值后，诸多学者针对如何进行精准的天气敏感性需求预测进行了深入的研究。天气敏感性需求预测重点分为三个部分，天气变量的刻画，影响因子的确定以及最后模型的选择。

针对天气变量的刻画，Huang C M[13]采用模糊逻辑的方法对天气信息进行了处理，并通过构建时间序列模型进行能源行业的预测。然而，这种模糊逻辑的方法有其局限性，仅适用于光伏发电效率规划的制定，对于零售行业的借鉴意义并不大。

贺蓉，曾刚[14]首次在文中提出对降雨进行量化的概念，将降雨量分为暴雨、大雨、中雨和小雨等；同时也对日进行量化和分类，将日分为重大节假日、一般公休日和工作日。并将这些作为影响因子加入BP算法的神经网络预测模型当中，取得不错的效果。贺蓉，曾刚的量化思想，对于天气变量的后续研究来说意义重大。

此外，张甫仁[15]在对建筑能耗预测中的研究中，所提出的“气象舒适度”的概念也颇有借鉴意义。气象舒适度，旨在于将天气对能耗的影响用人体舒适度表示。张甫仁将人体舒适度作为神经网络输入层的特征之一，进行预测并获得了较精准的结果。

在针对影响因子的研究当中，Divakar[16]将价格进行预测，并使其成为变量加入预测模型的想法也颇有借鉴意义。Divakar设计了一个销量预测模型，在对销量进行预测时，将价格、节假日、温度和新品投入等因素作为影响因子纳入考虑范围。

Lam[17]等对促销的作用进行了衡量，通过使用店铺表现，包括吸引力和消费影响等。其创新之处在于，将销量进行分解，针对成交比例、平均消费、人流量和客流量4个维度分别建立了模型进行数值的预测。Lam的研究为后人对影响因子的研究打下铺垫，黄瀚文[18]将客流量、人流量、成交比率和平均消费转化为影响因子，建立了基于店铺客流量的销量预测模型，取得了较好的预测效果。

Zhao HX Moore[19]的研究为天气敏感性需求预测的研究注入了新的活力。其文章中提到了seasonally lagged signal，证明了天气影响的滞后性。然而，Zhao并没有具体指出如何对天气影响的滞后性进行量化。本文从唐浩[20]的研究中受到启发，文中定义了‘连续晴天数’这一概念并给出结合实例的分析与解释。除此以外，文章中对降水的滞后效应进行了量化，将每日的降水影响系数分解成三天降水影响的累计，并且采用基于粒子群优化的方法对每日的降水影响系数进行求解，最终得到修正后的当天降水影响系数。这对本文的研究有着非常大的借鉴意义。

除了天气影响的滞后性的量化，如何去量化不同季节的敏感程度也是一个问题。莫降涛[21]采用三次函数来拟合产品的季节性需求，然而这种函数拟合的方式，对我们的研究帮助不大。Arunraj等[22]，采用box-and-whisker plot的方法，对不同产品不同季节敏感程度的不同进行了证明。将每个月份都单独作为一个0，1变量输入到模型当中进行预测并取得了一定的效果，这对我们的研究有一定的借鉴意义。

危万虎与张鸿雁[23]在武汉市夏季空调销量的气候预测及空调指数研究中发现，武汉夏季的闷热，除了与平均气温升高有关之外，也与最低气温升高有关，即昼夜温差变小，让人感到更加热，从而一定程度上增加了空调销量。这对本文的研究产生了新的启示，除了平均气温与最高气温以外，温差也是影响人们消费行为发生改变的重要因素。然而前人的研究中，都没有考虑到将前后两天的温差作为影响因子加入到模型当中，本文基于危万虎与张鸿雁的研究，将温差纳入建立预测模型时的影响因子当中。

**（2）早期天气敏感性需求预测模型的建立**

前文讨论了建立天气敏感性需求模型时需要考虑的影响因子，以及如何对各种天气变量进行刻画。接下来本文将分析前人在进行需求预测研究时的模型选择。Celia Frank[24]对早期的需求预测方法进行了总结，主要分为移动平均，自回归预测和将两者进行组合的销量预测模型，同时也提出了这些早期模型的局限：尺寸、价格以及颜色在内的外部环境不能发生改变。基于前人的模型，Celia Frank 等又提出了季节性单指数平滑、Winter模型和人工神经网络三种预测方法。其中人工神经网络的预测准确度最高，但由于人工神经网络模型的黑箱特点，其实际值和预测值的相关性较低。另一方面，人工神经网络模型的高精度可能源于对数据的“过拟合”，由于输入端的数据量太大，模型过度完美的处理了噪声数据。但是人工神经网络的优点也非常明显，其模型是非线性的，区别于移动平均、自回归预测等早期模型。

赵成柏和毛春梅[25]提出了一种线性残差模型来进行我国碳排放强度的长期变动趋势的预测，这个模型给本文研究的启示在于，考虑到了碳排放非线性的特点，将碳排放强度的时间序列数据结构分解为线性和非线性残差部分。并针对线性部分建立了ARIMA模型，针对非线性部分建立了BP神经网络模型进行预测。这种分解思想可以运用在我们对食品销量的预测中，黄瀚文将食品销量进行了线性部分和非线性残差的分解，建立了线性残差模型进行预测，具有一定的借鉴意义。

除了对线性部分和非线性残差进行分解以外，景亚平，张鑫和罗艳等[26]将BP神经网络预测模型与灰色GM（1，1）模型进行结合，加入马尔科夫链进行修正，建立了针对城市需水量预测的组合模型并取得较高的精准度。然而，不论是灰色模型还是BP神经网络模型，或者是双模型的组合，都存在一个问题，这些模型都是黑箱模型，并不能很好的解释影响因子对于最终变量的影响。同时，对于业界来说，这些模型的应用上可能存在问题。

高峻峻和谭崇利等[27]提出了一类产品需求预测的支持向量机方法，SVM是基于统计学习理论的新一代机器学习技术，着重研究小样本下的统计规律和学习方法，能较好解决传统统计模型和神经网络模型所不能解决的问题。该模型相比较其他模型精确的预测效果为本文的研究提供了启发，尝试突破传统统计模型和神经网络模型，从机器学习的角度进行模型选择。

**（3）TL-XGBoost模型研究**

Kaggle竞赛中梯度提升树算法（GBDT）的精彩表现对本文在模型选择上提供了启发。郑凯文和杨超[28]采用GBDT 算法进行了短期负荷预测研究，同时结合遵义市的真实负荷值和天气数据进行实证分析，以较高的预测精度证明 GBDT 算法在负荷预测领域有良好的效果。

桂良明，夏永俊和李海山等人[29]基于随机森林（RF）算法和GBDT算法建立了RF-GBDT模型来进行燃煤锅炉的气体排放预测，并结合将其与SVM、RF模型进行比较，发现基于RF-GBDT的预测模型精准度更高，以上GBDT算法在预测领域的应用对本文有很大的借鉴意义。

近年来，迁移学习浪潮的兴起启发了本文对迁移学习在机器学习以及回归领域的应用进行了研究。Pan和Yang[30]对迁移学习的思想进行了解释，如果在对一个类似理念熟悉的情况下，那么学习一个新的理论会更加容易。

Rettinger等人最早提出了一种基于迁移学习思想的ExpBoost算法[31]。在这之后，Dai等人在2007年正式提出了TrAdaBoost算法[32]。由Freund和Schapire在1997年提出的AdaBoost.M1算法是这两种算法的基础[33]。

在上述算法的基础上，Pardoe和Stone提出的Two-stage TrAdaboost.R2模型非常适用于本文拟定解决的问题，Two-stage TrAdaboost.R2是TrAdaboost算法的回归版本，同时针对过拟合的问题进行了改进[34]。

除此以外，也有学者对迁移学习在其他场景的应用做了研究，比如史荧中等人提出了一种迁移学习支持向量回归机[35]。

除了上文提到的GBDT、SVM以及TrAdaboost算法以外，XGBoost算法是对本文需求预测模型研究中产生最大影响的算法之一。陈天奇[36]在GBDT的基础上，对GBDT的目标函数进行二阶泰勒展开，并提出了一种适用与大规模并行计算的XGBoost模型。XGBoost模型随后在各项数据比赛中都获得了前三名的优秀成绩，逐渐成了许多数据科学家在处理回归问题时的首要选择算法。因此，本文考虑以XGBoost为基础，利用其应对大数据集表现出的优异性能，进行SKU维度的需求预测模型构建。

**1.4论文拟解决的关键问题**

本文拟解决的关键问题如下所示：

**（1）天气变量的刻画与选择**

如何刻画天气变量并进行选择对于需求预测来说是一个非常关键的问题。对天气变量的刻画主要体现在如何通过有限的天气数据去构建出有效的天气特征加入到需求预测模型的建立中。例如，对于温度，天气数据直接提供了最高温度、最低温度等数据，但是温差、前后两日温差，这些天气数据没有直接提供出的变量，对于需求可能也存在非常大的影响作用。因此，如何天气变量的刻画，并选择最有效的变量加入模型，是一个非常值得探索的过程。

**（2）天气对于产品替代性的影响**

天气对于销量的影响不仅体现在各种变量上，天气还会影响到人们对于同一品类下不同产品的选择。而在这方面，还未有学者进行系统的分析和研究。因此，如何将天气对于产品替代性的影响考虑进模型，并提高模型的精准度是一个非常关键的问题。

**（3）机器学习模型在业界应用的局限**

尽管机器学习模型相比较传统的统计学模型在许多方面有着更强的适用性，但是现阶段机器学习模型在工业界实体零售的需求预测中还存在着许多局限，诸如样本不足、标签缺失、训练集和测试集的分布等等问题，本文拟定引入迁移学习的知识来为机器学习模型在业界应用的部分局限提出解决方案。

**（4）天气敏感型需求预测模型的建立**

天气敏感型需求预测模型的建立，对于本文来说是最关键的问题之一。为了保证企业能在实际的生产计划中运用到模型，并为企业的多种业务决策提供支持，天气敏感型需求预测模型既要满足精准度高，同时又要保证运算速度快、需要的数据具有可获得性等等特点。本文拟定就这一问题，针对某咖啡品牌的四类产品，提出切实可行的天气敏感型需求预测模型，为企业的决策提供支持，带给企业更高的收益。

**1.5论文的主要研究内容及框架**

本文的主要研究方法如下所示：

**（1）文献研究方法**

本文通过对国内外的文献资料进行大量的阅读和归纳总结，包括天气敏感型需求理论的发展、boosting算法综述等等。这种文献研究方法可以帮助作者对国内外的研究现状进行充分的了解，同时保证了本次研究在内容上的创新性。

**（2）定性研究方法**

本文通过对研究对象进行定性分析，在定量的数据分析之前，先对产品的属性和特征进行定性的分析，使得变量之间的研究更为系统，模型的预测更为准确。

**（3）定量研究方法**

本文通过使用pearson相关系数等方法对变量之间的相关性和影响程度进行量化分析，同时利用迁移学习和机器学习算法对多种变量之间的关系建立回归模型，并将本文所构建的TL-XGBoost天气敏感型需求预测模型和多种机器模型进行比较和模型验证，最终选出最优的模型。

**（4）实证研究方法**

本文以中国某咖啡上海门店销售数据进行实证研究，包含对四种品类在SKU层面的每日销量需求预测，并将天气变量对销量的影响考虑进模型，最终为这四类产品的需求预测提出了一种切实可行的模型与方案，并为企业的决策和需求计划提供支持。

本文的主要研究内容可以被分成以下几个部分，（1）天气变量的刻画与选择；（2）品类维度的Two-stage TrAdaboost.R2模型构建；（3）SKU维度的XGBoost模型构建；（4）基于供应链层次的GAD模型SKU预测值修正；（5）实证分析和多种机器学习模型比较。

本文由五个不同的章节所组成，各部分的内容如下所示：

第一章节，绪论。对论文的研究背景、文献综述、拟定解决的关键问题以及创新点进行阐释。

第二章节，基本理论。对论文中涉及到的基本理论和算法，包括天气敏感型需求理论，Boosting Tree算法，XGBoost算法以及TrAdaboost.R2算法进行了介绍，为整篇文章提供了理论支持。

第三章节，基于TL-XGBoost的天气敏感型需求预测模型。对天气变量进行了刻画与选择，并依次在品类维度上构建了Two-stage TrAdaboost.R2模型、在门店维度上构建了XGBoost模型以及GAD模型对SKU进行修正。

第四章节，实证分析。在第三章节构建的TL-XGBoost的模型的基础上，引入实证分析，并和包括一阶段XGBoost、SVM、GBDT等在内的模型进行对比和分析。

第五章节，结论与展望。对本文的研究所获得的结论进行了总结，并为企业的管理提出一些建议，同时指出论文的局限性以及未来的进一步研究方向。

**1.6可行性分析及创新点**

**（1）项目可行性分析**

对产品或服务需求的预测与企业的利益密切相关，目前在学术界以及工业界的学者与专家们已经提出了一系列的预测框架和思路，为本文的课题研究提供了可以借鉴的地方。

本文采用Python和R为开发语言，建立基于TL-XGBoost的天气敏感型需求预测模型，结合数据分析、数据可视化、特征工程等方法来对某咖啡店n家门店的2018-2019年销售额历史数据进行挖掘，之后采用2019年1月到2月底共计n天的数据进行模型效果验证。同时通过对比其他多种机器学习以及数据挖掘领域算法比如支持向量机、梯度提升树等在该数据集上的预测效果来选择出最适合该咖啡店的预测模型。

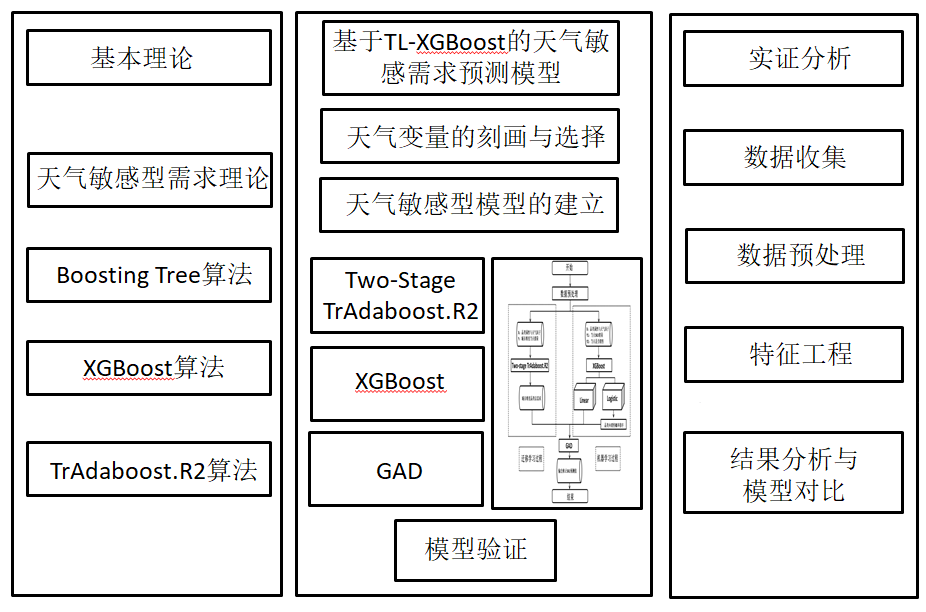
**（2）论文创新点**

第一，本文的商品需求预测研究不同于前人的商品销售额预测研究，根据快消品的特性，从天气敏感型需求预测理论角度出发，根据温度、温差、天气状态、上周销量均值等多种特征因素来分析对销售额的影响，对天气变量进行了详细的刻画与选择，同时应用近年兴起的机器学习和迁移学习方法将上述特征考虑进模型中完成需求预测。这样完成的需求预测可以有效解决企业在需求预测端的问题，提高企业利润，为库存管理和生产供应决策提供有力的参照，增强企业的核心竞争力。

第二，本文利用技术工具对四类产品的原始数据进行深度分析，分析过程结合了探索性数据分析和可视化数据分析等技术，对数据中有价值的隐藏信息进行了挖掘和直观呈现，并对基于天气敏感型理论的特征工程构建提供了强有力的支持。

第三，本文提出了基于TrAdaboost.R2模型、XGBoost模型以及GAD模型三者融合的TL-XGBoost需求预测模型，对机器学习在企业需求预测应用的部分局限提出了解决方案，同时相比较其他机器学习模型，TL-XGBoost模型不仅预测精准度高、运行速度快，其迁移学习的思想运用能帮助避免新数据量不够和完全抛弃仍有价值的旧数据的问题。除了在精度、高效等方面的有点，TL-XGBoost中的XGBoost模块可以对特征的重要性进行排序和分析，为特征选择提供了支持。

第四，本文的模型从供应链层次的角度出发，将天气对于产品替代性的影响考虑进模型，通过一种贪婪聚合分解的算法对SKU预测值进行了修正，从而进一步提高了模型的精准度。



图表1 本文研究主要框架图

**第二章 基本理论**

**2.1 天气敏感型需求理论**

**（1）天气敏感型需求**

天气敏感型需求是指产品或者服务的需求在本来的季节性波动基础上由于受到异常的天气因素的波动影响而发生的异常变化。比如，超出正常趋势的低温或者高温、过于少的降水量、极端天气等等所带来的产品或者服务需求的特殊上升与下降，或是产品或服务的需求产生时刻的特殊提前与延迟[5]。

天气敏感型的行业众多覆盖广泛，从金融保险等行业到零售制造等行业，从传统的农业到新兴的能源业等等。天气敏感型需求和季节性需求存在差异，和天气敏感型需求相比，季节性需求主要体现在销售的季节性波动趋势，即产品的需求量随着季节的转变而发生变化，

**（2）天气敏感型需求的特点和难点**

天气敏感型需求的特点与难点表现如下：

第一，天气的异常波动对于产品或者服务的需求所产生的变化很难去量化。天气的正常波动对于产品或者服务的需求所产生的影响，即人们通常所说的“季节性趋势”和天气的异常化波动是不一样的。而且，这些异常化波动通常对销量带来的影响很难去预料，对销量容易产生让商家难以应对的极端需求。下图显示的是2012年到2014年间上海地区的周平均气温对比图，通过对周平均气温的趋势进行对比，可以发现2013年的夏季周平均气温显著的大于2012年和2014年，那么这种异常化的天气波动，就容易对需求产生预料之外的影响。

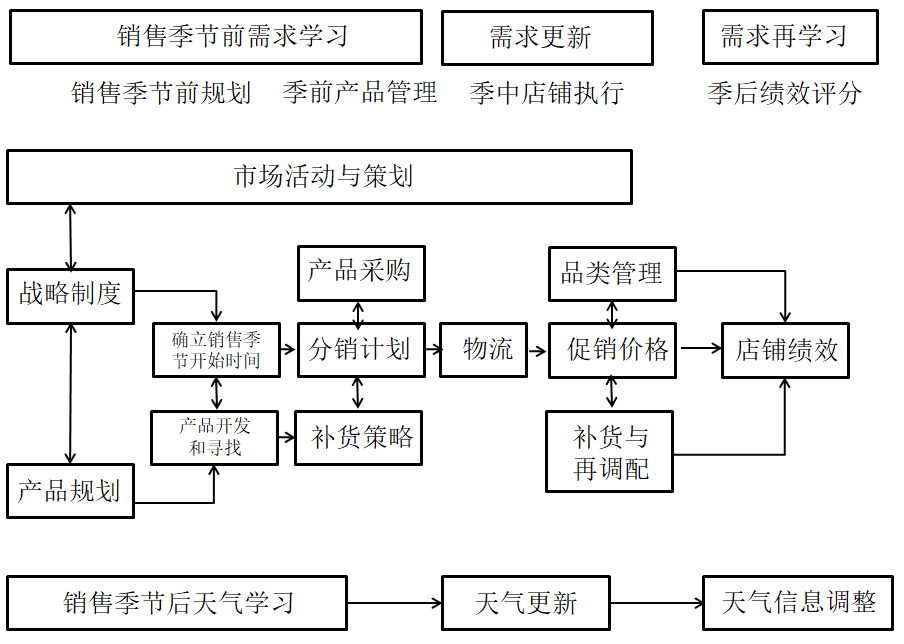
图表 2 2012-2014年上海周平均温度对比

第二，对需求产生影响的天气因素很多，有的天气因素，例如温度等，可以被定量的方式所刻画；而有的天气因素，例如降雨，既可以使用定量的定性变量去刻画，也可以单纯的使用定量的方式去刻画，而不同的刻画方式，对于去预测天气敏感型需求的效果又是不一样的，所以众多天气因素的刻画也是一个难点。

第三，天气变化对于产品需求的影响，可能由于地域、商圈、产品属性、品类类别等不同而产生不同的影响，因此如何对天气变量和其他变量之间的关系进行全面探索，找到对产品销量的变化影响较大的天气变量，并分析天气变量如何具体影响产品的销售，都需要持续不断的探索。

**（3）天气敏感型需求的管理模式**

从数字的角度，天气对需求的影响体现在销量和库存的高低起伏。然而，数字上的变动其深层原因是天气对于整个供应链的影响。一个企业的供应链结构复杂，在不同的销售季节下不同的部门协同合作以确保商品或者服务精准的送到消费者眼前。因此，如要探讨天气敏感型需求下的新型管理模式，则需要细分销售季节来分析天气因素对于需求的影响和指导作用，即企业需要根据产品的特性出发，根据季节制定差异化的运营方案。具体来讲，天气敏感型需求的管理模式分为季节前的需求学习、季节中的需求学习与季节后的需求再学习。



图表3 天气敏感型需求的管理模式

季节前的需求学习，时间跨度从产品的研发设计一直到产品进入市场面对消费者。这一段时间内，企业需要预测产品进入市场后的可能需求来为其上市计划提供准备。比方说，企业需要推出一款新的咖啡产品进入市场，那么在确定上市时间的同时，企业需要对新产品在该销售季节中的销量进行一个预测。而国内一些企业的预测方法通常是依赖过去多年的历史数据进行直接的销量模拟。这种预测方法的问题在于建立了一个错误的假设，即企业认为每一年天气对于需求的影响是相同的。因此，新的季节前的需求学习模式需要基于一个正常的假设上，即每一年天气对于需求的影响都不同。在这个假设的基础上，建立一个基于历史数据的需求预测模型，对每一年的天气影响因素进行量化，并加入这一销售季节的信息来得到产品在该销售季节的需求预测。此时的需求预测是一种模糊需求预测，虽然精准度较低但对于企业的供应链和市场营销提供了支持。

季节中的需求学习，即在整个销售季节中的需求预测。由于季节前的模糊需求预测精准度较低，所以需要季节中根据更新的信息对预测进行修正，进而为销售季节中企业的各种活动与计划提供指导，比如补货、促销、品类规划、营销等等。

季节后的需求再学习，指的是在销售季结束后对之前的绩效、成果以及运营状况的一个回顾和分析。在这个回顾的过程中发现问题，分析问题的产生原因和可能的解决方案，并为下一季度的销售计划做好准备。具体来说，比如企业可以对已有的需求预测模型进行更新，根据上一季度出现的问题进一步优化供应链结构等等。

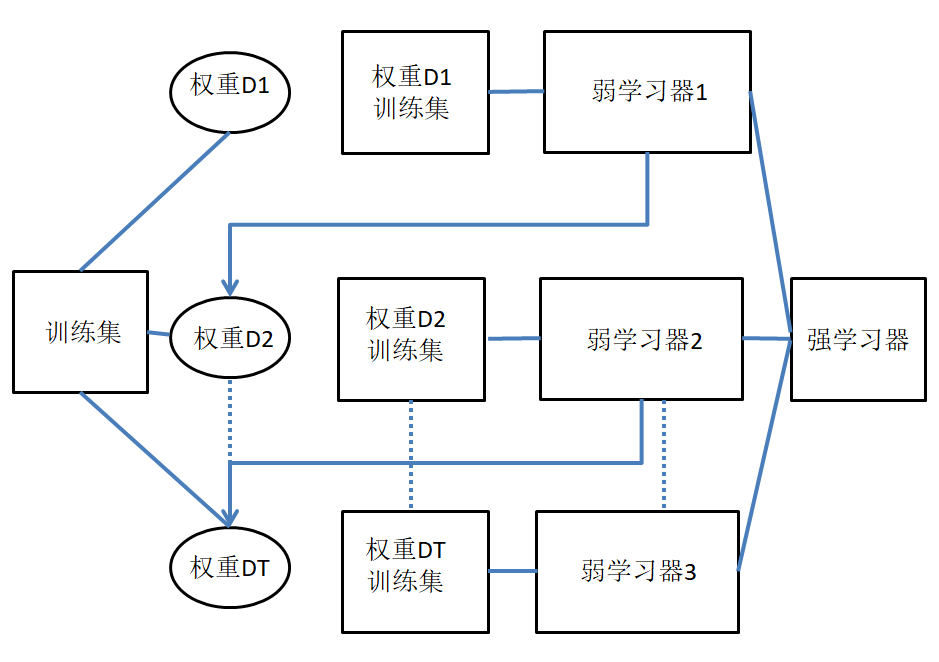
**（4）天气敏感型需求对于产品替代性的影响**

对于企业来说，由于产品的丰富多样，相同品类下不同产品之间有时存在着替代的关系。Kök[37]提出产品替代性的定义，即消费者对于给定的品类，如果该品类下消费者最中意的商品不被销售时，消费者有一定概率选择该品类下的其他产品来替代。目前在天气敏感型需求理论中，天气敏感型需求的管理模式以及各种天气变量的选择与刻画已经逐渐完善。但是，还没有学者在供应链领域对天气敏感型需求和产品替代性之间的影响关系进行充分讨论。天气对人们的购买行为会产生影响，同一品类下不同产品的销量数字上升和下降的背后隐藏的逻辑是消费者有一定几率选择该品类下的其他的产品作为替代。天气状况、温度和温差等因素都可能影响到消费者对于同一类别下不同产品的购买选择。

具体来讲，在冬天，虽然热饮是商家们进行主要销售的商品类别，但若某一天的平均气温显著的高于之前的平均气温，这就可能导致一部分消费者改变自己的偏好，去选择自己偏好的口味类别之中的冷饮。此时，如果商家没有预料到冷饮的需求，就会由于服务水平的降低而损失掉掉一部分的消费者。综上所示，由于产品本身的需求特性和天气的影响，大部分产品的需求会因此变得间断。因此，将天气敏感型需求对于产品替代性的影响考虑进模型构建是十分必要的。

**2.2 Boosting Tree基础算法**

提升树（Boosting Tree）算法既可以运用于回归问题的解决，也可以运用在分类场景的应用中。Boosting Tree算法属于提升方法，基本分类器选择可以为分类树或是回归树，其本质上是一种对基函数进行线性组合的加法模型。具体来说，提升树的决策树选择，在面对分类问题时一般选择为二叉分类树，在面对回归场景应用的时候一般选择为回归树。



图表4 Boosting算法基本思想示意图

**（1）GBDT梯度提升树算法**

GBDT模型全称Gradient Boosted Decision Trees，是一种迭代的决策树算法，，相对于一般决策树算法具有防止过拟合、泛化能力较强等优点，适用于需求预测。GBDT算法的本质是通过采用加法模型（即基函数的线性组合），以及不断减小训练过程产生的残差来达到将数据分类或者回归的算法。

GBDT通过多轮迭代,每轮迭代产生一个弱分类器，每个分类器在上一轮分类器的残差基础上进行训练。对弱分类器的要求一般是足够简单，并且是低方差和高偏差的。因为训练的过程是通过降低偏差来不断提高最终分类器的精度。

弱分类器一般会选择分类回归树。由于上述高偏差和简单的要求每个分类回归树的深度不会很深。最终的总分类器是将每轮训练得到的弱分类器加权求和得到的。

模型最终可以描述为公式1：

其中为当前m轮的模型，为当前m轮的弱分类器，为当前m轮弱分类器的损失函数。

模型预测的时候，对于输入的一个样本实例，首先会赋予一个初值，然后会遍历每一棵决策树，每棵树都会对预测值进行调整修正，最终的结果是将每一棵决策树的结果进行累加得到的最后得到预测的结果。

**（2）Adaboost算法**

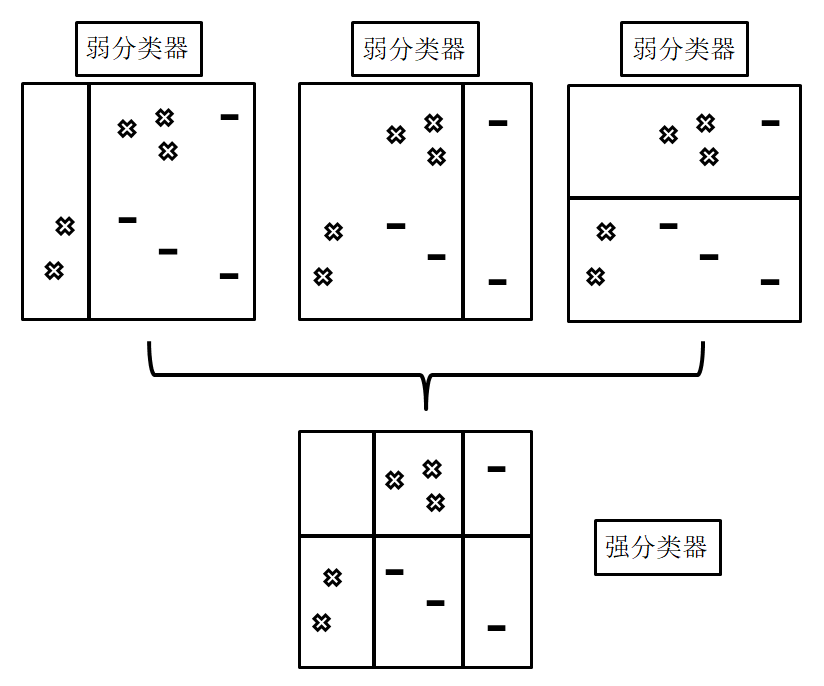
Adaboost算法是一种提升框架，通过将多个弱的分类器进行组合，最终形成一个强分类器。Adaboost会对上一个弱分类器分类错误的样本的权值进行调整和加强，样本在权值被更新会再次被用来进行下一个弱分类器的训练。Adaboost相比较其他模型，优点在于提供了一种框架，Adaboost的框架内可以人为的选择各种方法进行分类器的构建。除此以外，Adaboost在精准度、防止过拟合、特征筛选方面都有一定优势。

Adaboost的具体算法思想如下所示：

第一步，对训练数据中每一个样本的权值分布进行初始化，初始化的方法是，如果训练集中一共有个样本，则每一个样本在最开始时都被赋予一致的权重：。

第二步，对弱分类器进行训练。训练原则是，若一个样本已被正确的分类，那么在下一个训练集的构建中，该样本的权重会被降低；反之，若一个样本被错误的分类，那么它的权重就得到提高。样本集在权重被更新之后，会被用于训练下一个分类器并反复迭代此过程。

第三步，将所有的弱分类器进行组合，形成一个强分类器。在每一个弱分类器的训练过程结束后，分类的误差比率越小的弱分类器，其重要程度就设置的越大，重要程度越大，在最终的分类函数中，决定作用就越大。反之，分类误差率大的弱分类器的话语权小，在最终的分类函数中也就起着相对而言较小的决定作用。



图表5 Adaboost算法示意图

**2.3 XGBoost算法**

XGboost模型本质上属于boosting算法，是一种由陈天奇博士提出的关于树模型的大规模机器学习系统，同时可作为一种开源的包被调用。XGBoost算法在应对数据挖掘和机器学习相关挑战时表现出色，在例如Kaggle的数据大赛中被广泛的使用[36]。

XGBoost的算法思想是通过对许多基础的弱分类树或者回归树进行集成，最终形成一个强分类树或者回归树。每一轮XGBoost会在原有的树模型基础上再添加一棵新的树，这个过程的数学原理是对上一个树模型预测的残差值进行一次拟合，学习得到新的一个函数。当所有的模型都训练完毕之后，得到一个集成的模型，此时根据输入的新样本的特征，在每个树相应的叶子节点计算出一个分数，最后对所有节点的分数加和得到该样本的预测值。

XGBoost的目标函数定义如下所示：

其中为观测值，为预测值，为观测值与预测值之间的损失函数，为当前的决策树，为正则项；

XGBoost相比较GBDT模型，其特点在于对目标函数采用了二阶泰勒展开来进行近似，二阶泰勒展开后的目标函数近似如下：

其中为函数的一阶导数，为函数的二阶导数， 为节点分裂的次数；XGBoost通过对目标函数进行二阶泰勒展开，加快了整个模型的收敛速度，并且能快速的得到一个全局最优解；

XGBoost目前在多种语言上都有开源的库，比如在Python和R中都可以被简洁的调用。同时，Python和R中的XGBoost库也为使用者提供了不同的参数选择，用户可以通过调节参数来得到最适用于特定情景的XGBoost模型。比如在基学习器的选择上，用户可以根据其回归或是分类的目的来选择包括CART、Linear Regression和Logistic Regression在内的多种基学习器。

**2.4 TrAdaboost.R2算法**

TrAdaboost.R2属于迁移学习算法中的一种，是由David Pardoe提出的基于迁移学习思想的一种提升树模型。TrAdaboost.R2是Adaboost.R2原理与TrAdaboost原理相结合产生的一种针对回归问题的模型。TrAdaboost.R2分别将目标数据和源数据作为输入，并组合两组数据集为单一数据集应用到boosting过程中。相比较传统的机器学习模型，TrAdaboost.R2解决了加入新数据后训练集和测试集可能出现的单一分布假设不成立的问题。迁移学习的思想帮助其从旧数据中提取信息用于训练新的模型，避免了新数据量不够和完全抛弃仍有价值的旧数据的问题[34]。

Two-stage TrAdaboost.R2模型是针对TrAdaboost.R2的一种改进版本，解决了其非常容易受到过拟合影响的问题。相比较之下，新的版本分为两个阶段：在第一个阶段中，源数据的权重被逐渐向下调整直到达到某个特定的点；在第二个阶段中，所有源数据的权重被冻结，目标数据的权重正常更新。同时，只有第二个阶段中生成的假设会被存储并用于模型的输出。Two-stage TrAdaboost.R2模型的算法思想如下所示：

**输入**两个带标签的数据集 （样本大小）与 （样本大小），步骤次数 ，最大boosting迭代次数 ，交叉验证的层数 ，一个基准学习器 。 与 的组合集 。设置初始权重向量 为 ，。

**For**  :

1. 调用Adaboost.R2，包括数据集，初始权重向量 ，最大迭代次数和基准学习器，得到，该Adaboost.R2模型的前个数据的权重永不改变。使用 层交叉验证来得到残差的预测值。
2. 调用，得到假设 。
3. 计算并调整Adaboost.R2中每个样本的残差 。
4. 更新权重向量

其中是一个正则化常数，被选中使得目标样本的最终权重为。

**输出** ，当 。

**第三章 基于TL-XGBoost的天气敏感型需求预测模型**

**3.1引言**

本章的目的是建立一个基于TL-XGBoost的天气敏感型需求预测模型用于帮助企业进行需求预测来支持其库存、促销、定价等计划的制定。文章首先对学术界与企业面对的需求预测问题进行问题描述。之后文章通过pearson相关系数等方法来对天气影响因素进行刻画与选择。完成天气变量的选择与刻画后，开始建立基于TL-XGboost的需求预测模型，具体分为三个部分：Two-stage TrAdaboost.R2模型建立，XGBoost模型建立以及GAD模型建立。

**3.2问题描述**

对于企业来说，准确的预测消费者需求可以为其库存管理计划、促销活动安排、定价方案制定等运营项目提供支持。因此，一个准确高效的需求预测模型对企业来讲十分重要。在业界中，传统的统计学模型已经得到广泛的运用，相比之下，机器学习在业界的应用还存在许多问题。虽然机器学习对不同类型的数据有着更强的适用性，同时能更逼近预测的上线，但是其在业界中许多场景下进行预测仍会出现偏差。本文通过研究分析，发现机器学习应用的障碍主要和两个问题有关，一个是样本的数量与分布问题，一个是需求的间断性。

样本的数量与分布问题在于，传统的机器学习应满足两个基本假设：(1) 用于学习的训练样本与新的测试样本满足独立同分布；(2) 必须有足够可用的训练样本。然而真实工业环境的数据有时候无法满足这些条件，因此如何客服样本的数量与分布问题是传统机器学习的一个难点。

除了样本的数量与分布问题以外，实际应用中需求预测的偏差还与需求的间断性有关。真实工业环境中，高达60%的SKU的需求是存在间歇性的，因此如何准确的预测间断性需求是在业界中应用机器学习的另一个困难。

针对以上两个问题，本文从迁移学习与供应链的角度出发，构建了TL-XGBoost模型来进行SKU级别的销量预测。

**3.3天气变量的刻画与选择**

本章首先对四类产品根据特点进行分类，四种产品均全年有销售，其中品类2与品类4的销量在全年上看显著大于品类1与品类3。根据下图“上海地区2018年四类产品总销量”，可以看出对于品类2，3月份，10月份下旬，11月份以及12月份的销量整体显著大于4月份，7月份，8月份，9月份以及10月份上旬的销量。这种销量的变化可能是因为在寒冷的月份，人们对于咖啡的需求上升所导致。除此以外，在10月份上旬，四种品类都经历了销量的断崖式下跌，可能是由于“十一”黄金周的存在，咖啡对于上班一族的需求断崖式下落，本文将在后面进一步讨论这一变化。

图表6 上海地区四类产品销量曲线

仅站在上海地区的角度对四种品类进行销量的观测会遗漏掉许多有价值的信息，因为上海地区的总销量是通过每一个地区每一个门店的销量聚合所得到，这个聚合的过程不可避免的会有门店、地区特有的信息的损失，因此本文将在后文考虑将门店、地区特有的属性作为一种解释变量加入模型当中。

通过天气后报网站的天气数据收集，本文获得了2018年度的上海地区天气数据，包括白天天气状态、夜晚天气状态、最高气温、最低气温、早上风向、晚上风向、早上风力与晚上风力。

在这八个变量的基础之上，本文还引入平均气温、当日温差、前后两天温差三个变量。

平均气温：最高气温与最低气温的平均值。

当日温差：当日最高气温与最低气温的差值。

前后两天温差：前日平均气温与今日平均气温的差值。

本文接下来分别以品类1总销量、品类2总销量、品类3总销量、品类4总销售为因变量与11个天气变量进行线性回归建模，采取的软件是SPSS，因为在数据量不大的情况下，SPSS操作与运行较为方便同时可以勾选是否进行pearson相关系数分析与共线性诊断，回归结果如下图所示。

|  | | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 模型 | | 非标准化系数 | | 标准系数 | t | Sig. | 相关性 | | | 共线性统计量 | |
| B | 标准 误差 | 试用版 | 零阶 | 偏 | 部分 | 容差 | VIF |
| 1 | (常量) | 362.655 | 14.753 |  | 24.581 | .000 |  |  |  |  |  |
| 白天天气 | .439 | 1.857 | .013 | .237 | .813 | -.070 | .013 | .012 | .852 | 1.174 |
| 晚上天气 | -4.798 | 1.764 | -.147 | -2.720 | .007 | -.140 | -.144 | -.140 | .908 | 1.102 |
| 最低气温 | -1.107 | .345 | -.173 | -3.212 | .001 | -.198 | -.169 | -.165 | .912 | 1.096 |
| 温差 | -1.632 | 1.383 | -.069 | -1.181 | .239 | -.131 | -.063 | -.061 | .763 | 1.310 |
| 早上风向 | -1.833 | 2.870 | -.075 | -.639 | .523 | -.130 | -.034 | -.033 | .193 | 5.191 |
| 早上风力 | -2.129 | 4.936 | -.072 | -.431 | .667 | .026 | -.023 | -.022 | .094 | 10.588 |
| 晚上风向 | -.341 | 2.824 | -.014 | -.121 | .904 | -.119 | -.006 | -.006 | .193 | 5.175 |
| 晚上风力 | 1.298 | 4.821 | .045 | .269 | .788 | .030 | .014 | .014 | .094 | 10.584 |
| 前后温差 | 1.833 | 1.366 | .072 | 1.342 | .180 | .100 | .072 | .069 | .915 | 1.093 |
| a. 因变量: 品类1 | | | | | | | | | | | |

图表7 品类1-总销量-天气变量回归方程的回归系数图

从图中可知，在对品类1总销量的建模过程中，模型自动剔除了最高气温与平均气温两个变量，剩下的9个天气变量，晚上天气、最低气温与品类1总销量之间存在显著的线性相关性。同时从回归系数中，可以发现最低气温的回归系数为-1.107，说明最低气温和品类1销量呈负相关关系，即当最低气温下降时，品类1的总销量整体呈现上升趋势。

除此以外，图中显示早上风向、早上风力、晚上风向以及晚上风力的VIF大于5，即存在较为显著的共线性关系。对于线性回归模型，在建模中需要剔除线性相关的变量。本文对天气变量之间的共线性选择不做处理，这么做的原因主要和本文后续建立的天气敏感型需求模型为树模型有关，具体原因是因为，在做预测时，树模型往往采用一个贪婪算法来选择特征，即遍历每一个新变量对结果的影响，只有当这一个变量对结果的影响较大时才会被加入到树模型当中，基于这个算法的特点，在变量选择的过程中，共线性的多个变量中往往只有一个变量会被选中加入到模型中，因此冗余的特征不会被加入到模型中。

|  | | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 模型 | | 非标准化系数 | | 标准系数 | t | Sig. | 相关性 | | | 共线性统计量 | |
| B | 标准 误差 | 试用版 | 零阶 | 偏 | 部分 | 容差 | VIF |
| 1 | (常量) | 3177.255 | 139.126 |  | 22.837 | .000 |  |  |  |  |  |
| 白天天气 | -3.208 | 17.509 | -.010 | -.183 | .855 | -.070 | -.010 | -.010 | .852 | 1.174 |
| 晚上天气 | -38.569 | 16.636 | -.127 | -2.318 | .021 | -.127 | -.123 | -.121 | .908 | 1.102 |
| 最低气温 | -7.896 | 3.250 | -.132 | -2.429 | .016 | -.142 | -.129 | -.126 | .912 | 1.096 |
| 温差 | 2.254 | 13.039 | .010 | .173 | .863 | -.025 | .009 | .009 | .763 | 1.310 |
| 早上风向 | -27.845 | 27.064 | -.122 | -1.029 | .304 | -.092 | -.055 | -.054 | .193 | 5.191 |
| 早上风力 | -46.191 | 46.551 | -.168 | -.992 | .322 | -.065 | -.053 | -.052 | .094 | 10.588 |
| 晚上风向 | 9.534 | 26.633 | .042 | .358 | .721 | -.071 | .019 | .019 | .193 | 5.175 |
| 晚上风力 | 22.774 | 45.464 | .085 | .501 | .617 | -.055 | .027 | .026 | .094 | 10.584 |
| 前后温差 | 17.221 | 12.884 | .073 | 1.337 | .182 | .076 | .071 | .070 | .915 | 1.093 |
| a. 因变量: 品类2 | | | | | | | | | | | |

图表8 品类2-总销量-天气变量回归方程的回归系数图

从图中可看出，在对品类2总销量的建模过程中，可以发现晚上天气、最低气温与品类2总销量之间存在显著的线性相关性。同时从回归系数中，可以发现最低气温的回归系数为-7.896，说明最低气温和品类2销量呈负相关关系，即当最低气温下降时，品类2的总销量整体呈现上升趋势。

|  | | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 模型 | | 非标准化系数 | | 标准系数 | t | Sig. | 相关性 | | | 共线性统计量 | |
| B | 标准 误差 | 试用版 | 零阶 | 偏 | 部分 | 容差 | VIF |
| 1 | (常量) | 323.819 | 14.299 |  | 22.646 | .000 |  |  |  |  |  |
| 白天天气 | -.733 | 1.800 | -.023 | -.407 | .684 | -.056 | -.022 | -.022 | .852 | 1.174 |
| 晚上天气 | -2.729 | 1.710 | -.089 | -1.596 | .111 | -.089 | -.085 | -.084 | .908 | 1.102 |
| 最低气温 | -.363 | .334 | -.060 | -1.086 | .278 | -.080 | -.058 | -.057 | .912 | 1.096 |
| 温差 | -.512 | 1.340 | -.023 | -.382 | .703 | -.061 | -.020 | -.020 | .763 | 1.310 |
| 早上风向 | -.300 | 2.782 | -.013 | -.108 | .914 | -.024 | -.006 | -.006 | .193 | 5.191 |
| 早上风力 | .241 | 4.784 | .009 | .050 | .960 | .034 | .003 | .003 | .094 | 10.588 |
| 晚上风向 | .358 | 2.737 | .016 | .131 | .896 | -.019 | .007 | .007 | .193 | 5.175 |
| 晚上风力 | .250 | 4.673 | .009 | .053 | .957 | .032 | .003 | .003 | .094 | 10.584 |
| 前后温差 | 2.201 | 1.324 | .092 | 1.662 | .097 | .096 | .089 | .088 | .915 | 1.093 |
| a. 因变量: 品类3 | | | | | | | | | | | |

图表9 品类3-总销量-天气变量回归方程的回归系数图

从图中可看出，在对品类3总销量的建模过程中，可以发现前后温差与品类3总销量之间存在显著的线性相关性。同时从回归系数中，可以发现前后温差的回归系数为2.201，这个数字的实际意义是，当前一天的天气平均气温大于后一天的天气平均气温，即发生了降温情况时，品类3的总销量整体呈现上升趋势，降温的幅度越大，品类3的总销量增长越快。

|  | | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 模型 | | 非标准化系数 | | 标准系数 | t | Sig. | 相关性 | | | 共线性统计量 | |
| B | 标准 误差 | 试用版 | 零阶 | 偏 | 部分 | 容差 | VIF |
| 1 | (常量) | 793.063 | 45.090 |  | 17.589 | .000 |  |  |  |  |  |
| 白天天气 | -4.868 | 5.675 | -.044 | -.858 | .392 | .019 | -.046 | -.040 | .852 | 1.174 |
| 晚上天气 | -11.895 | 5.391 | -.109 | -2.206 | .028 | -.111 | -.117 | -.104 | .908 | 1.102 |
| 最低气温 | 9.264 | 1.053 | .434 | 8.794 | .000 | .426 | .426 | .415 | .912 | 1.096 |
| 温差 | 13.442 | 4.226 | .172 | 3.181 | .002 | .175 | .168 | .150 | .763 | 1.310 |
| 早上风向 | -5.355 | 8.771 | -.066 | -.611 | .542 | .002 | -.033 | -.029 | .193 | 5.191 |
| 早上风力 | -9.033 | 15.087 | -.092 | -.599 | .550 | -.035 | -.032 | -.028 | .094 | 10.588 |
| 晚上风向 | -2.140 | 8.632 | -.027 | -.248 | .804 | .002 | -.013 | -.012 | .193 | 5.175 |
| 晚上风力 | 14.910 | 14.734 | .155 | 1.012 | .312 | -.022 | .054 | .048 | .094 | 10.584 |
| 前后温差 | 6.109 | 4.176 | .072 | 1.463 | .144 | -.017 | .078 | .069 | .915 | 1.093 |
| a. 因变量: 品类4 | | | | | | | | | | | |

图表10 品类4-总销量-天气变量回归方程的回归系数图

从图中可看出，在对品类4总销量的建模过程中，可以发现晚上天气、最低气、温差与品类4总销量之间存在显著的线性相关性。同时从回归系数中，可以发现温差的回归系数为13.442，这个数字的实际意义是，当天从早到晚，降温的幅度越大，品类4销量增长越明显。

通过上述四个品类的分析，可以发现当发生了降温情况时，四种咖啡的销量都会呈现上升趋势。对于不同品类而言，受到的天气影响程度和方式都不同。基于以上分析与讨论，本文决定将天气变量作为特征考虑进模型的建立当中。

**3.4天气敏感型需求预测模型建立**

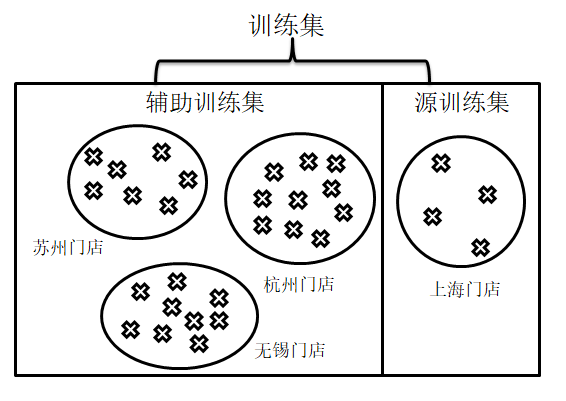
* + 1. **Two-stage TrAdaboost.R2模型建立**

本文首先使用Two-stage TrAdaboost.R2模型在城市维度上对品类需求进行预测。使用Two-stage TrAdaboost.R2模型的原因在于，城市维度上的品类销量数据条目较少，2018年3月份到12月份一整年只有305条数据，若是直接使用其他机器学习模型进行建模，无法满足传统机器学习的两个基本假设。即

(1)用于学习的训练样本与新的测试样本满足独立同分布；

(2)必须有足够可用的训练样本。

Two-stage TrAdaboost.R2模型可以有效的解决训练样本不足以及新的样本不满足独立同分布的问题。首先针对训练样本不足这一问题，虽然源数据中上海地区一整年的销售数据条目只有305条，但是可以通过选择大华东地区下其他的城市，比如杭州、无锡、苏州等地区的相同品类的销售数据作为辅助训练样本，这样避免了丢弃掉训练数据中有价值的部分，同时解决了训练样本不足的问题，具体思想如下图所示。



图表11 辅助训练集与源训练集

虽然通过加入大华中地区其他门店的销量数据作为辅助训练集解决了训练样本不足和丢掉有价值的信息的问题。但是新加入的辅助训练样本和源训练样本的分布问题还未得到解决，Two-stage TrAdaboost.R2模型利用了AdaBoost算法的思想原理来解决这个问题。在初始状态，给训练数据中的每一个样例都赋一个权重，同时计算源域中的样本的残差，根据残差来调整这个样本的权重，比如当残差过大时，说明这个样本预测不准确，于是在下一次迭代的过程中增大这个样本的权重。而对于辅助训练集来说，如果一个辅助训练集的样本预测的残差过大，则说明这个辅助训练集中的特定样本与目标数据是很不同的，因此在下一次迭代的过程中降低该样本的权重，Two-stage TrAdaboost.R2模型对源域和目标数据中样本权重调整的思想是一样的，但是方向不同。

除此以外，Two-stage TrAdaboost.R2模型区别于一阶段的TrAdaboost.R2模型的特点在于，第一个阶段中，模型会对辅助训练样本的权重进行调整，逐渐下降到某个特定的点，而在第二个阶段中，所有辅助训练样本的权重被冻结，源域的权重正常更新。

具体来讲，首先将上海地区2018年3月份到12月份一共305条数据作为源数据集，将大华东地区下杭州、无锡、苏州等地区的销售数据作为辅助训练集，针对品类1、品类2、品类3和品类4分别建模。设置品类属性和上文提到的11个天气影响因素为自变量，城市维度的每日品类总销量为因变量，最后输出城市维度每个品类的总需求。

* + 1. **XGBoost模型建立**

在完成城市维度上四种品类的总需求建模之后，本文在SKU层面上采用XGBoost模型进行SKU维度的需求预测建模。SKU维度的数据量相比较城市维度呈现几何级增长，城市维度2018年度销售数据只有305条，要解决的是训练数据不够的问题。而细分到SKU维度之后，四种品类共计2684中产品的销售数据总共有685460条，除了对模型的预测精度有要求外，对模型的训练时间同样提出了要求。

XGBoost相比较传统的GBDT等机器学习模型，其中一大特点就是XGBoost的模型收敛速度更快，且能更快速的得到一个全局最优解。从数学角度来看，差别首先体现在代价函数上。XGBoost的目标函数如下公式所示：

其中为观测值，为预测值，为观测值与预测值之间的损失函数，为当前的决策树，为正则项；

相比较GBDT模型，XGboost在代价函数中加入了一个正则项为正则项，用来控制整个模型的复杂程度。XGBoost还对目标函数进行了二阶泰勒展开，展开后的目标函数近似如下：

其中为函数的一阶导数，为函数的二阶导数， 为节点分裂的次数；这种二阶泰勒展开的方式，不仅加快了整个模型的收敛速度而且能快速的得到一个全局最优解；除了目标函数以外，正则项的公式如下：

其中为数的叶子数目，为叶子的分数的平方。从中可以看到XGBoost的正则项对数的叶子数目以及叶子的分数都进行了惩罚，这么做可以使得训练出来的模型更加简单，防止出现过拟合的情况。

除了目标函数的二阶泰勒展开和正则项惩罚以外，XGBoost模型还允许使用列抽样来防止过拟合。并且XGBoost实现了一种分类节点寻找的近似算法，帮助其加速模型的训练过程和减少整体模型对电脑内存的消耗程度。此外，XGBoost自身含有对缺失值的处理方法，如果样本的特征值存在缺失，XGBoost可以自动的学习出其分裂方向。

综上所述，相比较GBDT等模型，XGBoost更适用于SKU维度上大量数据的建模。具体的建模，分为两个部分，一部分是SKU维度的每日需求预测，另一部分是SKU维度的每日销售概率预测。

第一部分，在SKU维度以2018年3月到12月每一个SKU的当天销量作为因变量，产品属性以及上文提到的天气变量等作为自变量，构建模型并输出2019年1月到2月份的每日产品初始需求预测值。在这一部分中，选择linear regression作为基学习器。

第二部分，在SKU维度以2018年3月到12月每一个SKU当天是否销售作为因变量，产品属性以及上文提到的天气变量等作为自变量，构建模型并输出2019年1月到2月份的每日产品销售概率。在这一部分中，选择logistic regression作为基础学习器。

* + 1. **GAD模型建立**

上文中首先建立了Two-stage TrAdaboost.R2模型来预测城市维度上四个品类每个品类的总需求，随后通过构建XGBoost模型来对上海区域内每个品类下每个SKU每天的初始需求和销售概率进行了预测。接下来本文采用基于供应链层次的贪婪聚合算法（GAD）来对品类的总需求、每个产品的初始需求以及每个产品的销售概率三者进行聚合与分解，完成最终SKU预测值的修正。

供应链具有层级结构，每个层级都包含其本层级的特有信息，比如在年层级可以计算出趋势信息，在季和月层级可以计算出季节性信息等。层次预测方法利用供应链不同层级的有效信息，帮助其他层级进行预测，可以大大提高预测的准确性[38]。GAD模型预测法也是供应链中一种特有的预测方法。

具体的模型步骤如下所示：

1. 对品类内每个SKU按照销售概率从大到小进行排序；
2. 根据（1）中的SKU销售概率顺序，分别对每个品类下所属的所有SKU的初始需求进行加和，加和顺序从销售概率最大的SKU开始，当加和的值大于对应品类城市维度上的的总需求时，停止加和；
3. 设置该品类下剩余SKU的初始需求为0；
4. 输出修正后的SKU预测结果并结束；

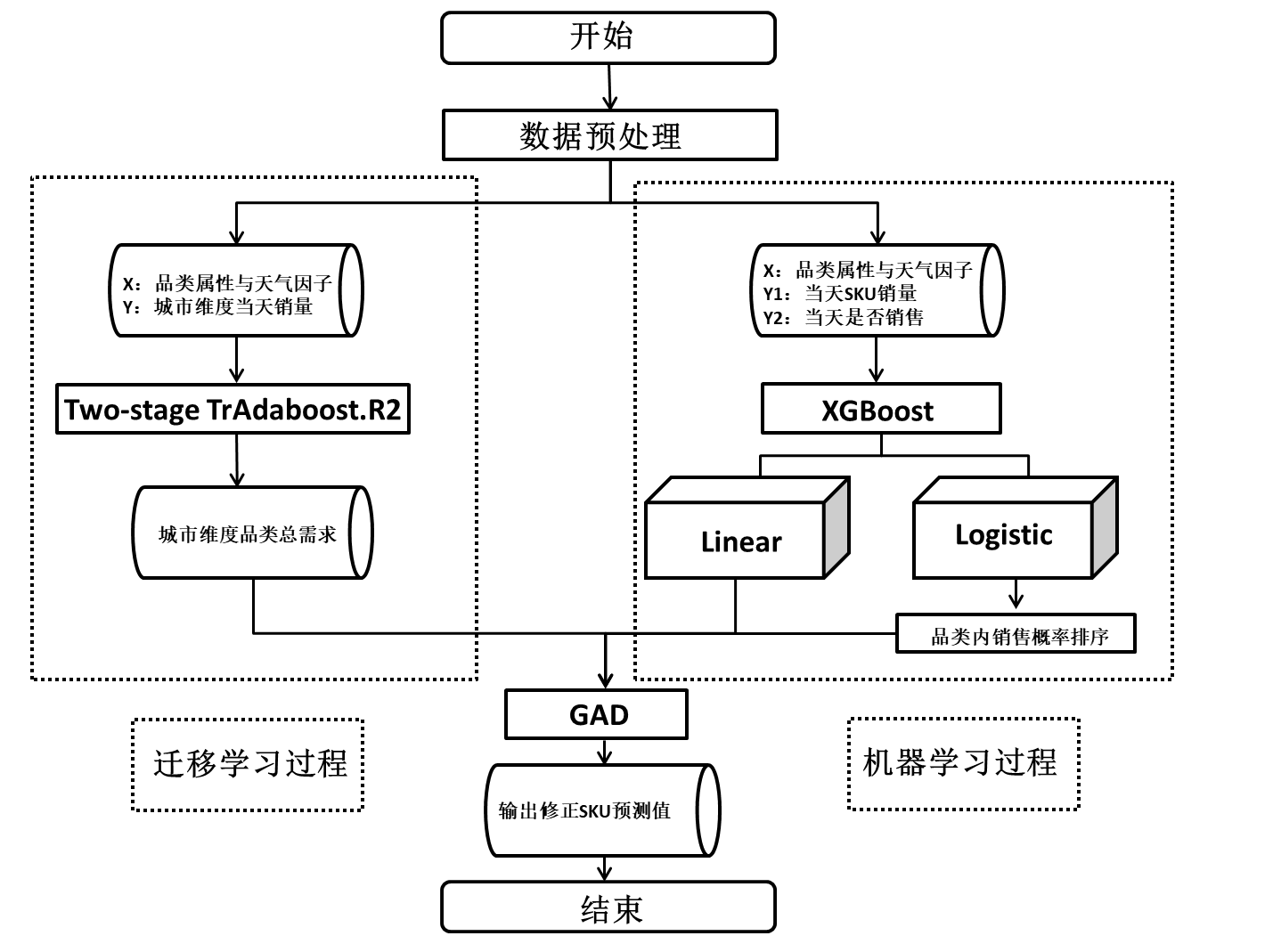
综上，本文构建的基于TL-XGBoost的天气敏感型需求预测模型，步骤与流程图如下所示。

**步骤1** 通过在城市维度构建Two-stage TrAdaboost.R2模型，输出每个品类的总需求；

**步骤2** 通过选择linear regression作为基学习器，在SKU维度构建XGBoost模型并输出产品的初始需求；

**步骤3** 通过选择logistic regression作为基学习器，在SKU维度构建XGBoost模型并输出产品的销售概率；

**步骤4** 通过GAD模型，对SKU预测值进行修正，输出四类产品下每个SKU的最终需求预测结果。



图表12 TL-XGBoost模型框架图

**3.5 模型验证**

模型的验证分为两个部分，在城市维度的品类总需求预测模型验证和在SKU维度的产品销售概率预测模型以及产品初始需求预测模型验证。

对于城市维度的品类总需求预测模型，采用平均绝对百分误差MAPE作为模型的评价指标。具体定义如下所示：

（7）

其中，为观测值,为预测值，为样本的数量。

对于SKU维度的需求预测，采用平均绝对误差MAE作为模型的评价指标，不采用平均绝对百分误差MAPE的原因在于此时SKU维度的需求预测可能出现销售量为0的情况，而平均绝对误差MAPE的数学特性使得其不适用与销售量为0的情况。平均绝对误差MAE也是一种常用的相对评价指标，可以比较SKU维度的需求预测模型相互之间的性能与精确度，具体定义如下所示：

（8）

其中，为观测值,为预测值，为样本的数量。

**第四章 实证分析**

本文研究的目的是建立一个天气敏感型需求预测模型，并对某咖啡连锁品牌上海地区门店下ABCD四种品类的咖啡进行SKU层级的每日销量预测。由于研究的对象是上海地区实体门店，因此需要将店铺特征、店铺选址、商圈等因素对于需求的影响纳入考虑。除此以外，还应需要考虑天气变量对于销量的影响。本章首先对介绍需要哪些数据以及相关数据的收集过程，然后对数据预处理的过程进行详细分析，包括数据清洗和数据转换等，之后对特征选择的思路进行描述，最后对实证的结果进行分析，并与包括SVM、GBDT在内的多种机器学习模型进行对比。

**4.1数据收集**

* + 1. **数据来源**

本文研究数据来源分为两个部分，销量数据与天气数据。

销量数据方面，本文针对中国某咖啡连锁品牌上海地区门店2018年3月到2019年2月共计360天的咖啡销售数据进行研究，采用前304天的咖啡销售数据作为训练集，后56天的数据作为测试集。除了上海地区门店的数据以外，本文还获取了该咖啡连锁品牌大华中地区下的其他城市地区门店同时期的销售数据作为辅助训练样本，时间跨度从2018年3月到2018年12月共计304天.

天气数据方面，本文从天气后报数据发布网站获取上海地区2018年3月到2019年2月共计360天的天气数据，天气数据包括天气状况、气温和风力风向等。

* + 1. **描述性统计**

本文实验数据采用时间跨度从2018年3月到2019年2月共计360天的实际门店咖啡销售数据。商品的品类共计有四种类别，四种不同的类别下累计有2684种产品参与需求预测的模型建立，本文数据的描述性统计下图所示。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 最大值 | 最小值 | 方差 | 平均值 |
| 非零需求占比 | 0.01 | 1 | 0.44 | 0.11 |
| 平均销售间隔（ADI） | 1 | 336 | 10.56 | 554.80 |
| 时间序列长度 | 1 | 360 | 229.15 | 16659.90 |
| 有销售周数 | 1 | 360 | 102.27 | 12162.62 |

图表13 销售数据的描述性统计表

**4.2数据预处理**

* + 1. **缺失值处理**

实际数据中，缺失值是一个不可避免的问题。缺失值出现的原因，可能是由于人工录入的失误等原因。对于不同的数据场景，在遇到缺失的数据时应采取不同的策略。

常用的策略有以下几种：

1. 对缺失的条目进行删除，适用于缺失的数据较少，删除掉缺失数据对于整体数据的分布和规律性影响不大的情况。
2. 使用缺失数据邻近数据的均值或者是中位数来替代空缺，这种策略不会造成样本条目和信息的缺失，但若缺失数据为非随机数据时，容易产生较大的偏差对挖掘数据背后的信息造成影响。
3. 插补法，包括随机插补法、热平台插补法、牛顿插值法和拉格朗日插值法、多重插补法等等。
4. 对数据进行建模，通过回归等方式来预测缺失值的值。

问题一，本文的原始销售数据中，由于人工录入的失误等原因，每周销量数据中部分条目出现乱码的情况。对于这样一种情况，不能直接对乱码的条目进行删除，因为直接删除掉乱码数据，可能会对整体数据的分布和规律性产生影响。因此，本文采取的方法是使用中位数来进行代替，保证整体数据的分布与规律性不会受到太大的影响。

问题二，本文从天气后报网站收集到的天气数据中，共计360天的天气数据中，有两天的天气数据缺少最低温度这一项数据，本文采取的做法是取前后几天最低温度的均值来进行填补。

问题三，本文的原始销售数据中，缺少2018年9月16号这一天的数据，本文采取的做法是，对日期为2018年9月16号的条目进行直接删除，因为直接删除掉这一天对于整体数据的分布和规律性影响不大。

* + 1. **数据转换**

本文选择对数据进行标准化处理，进行标准化的原因是为了消除不同的量纲对于数据产生的影响，通过将数据放在同一标准下，实现对不同量纲的数据之间的比较。标准化并不适用于所有的模型，标准化主要针对三种模型。

1. 模型对于不同的量纲的数据非常敏感，比如支持向量机SVM。对于SVM而言，当数据的各个维度进行不均匀的伸缩变换之后，此时模型计算出的最优解会与原来的解不等价。因此，对于SVM这一类的模型，需要对数据的分布范围进行约束，一般需要进行归一化处理。
2. 许多模型在求解计算的过程中会采用迭代算法，如果数据的分布范围部分过大或过小，会导致求解时的目标函数形状过于扁平，从而对解决时的迭代算法产生影响，迭代算法可能出现不收敛或者是收敛速度极慢的情况。因此对于这一类模型，也需要采用标准化的方法。
3. 带有正则化的逻辑回归模型。当逻辑回归模型的损失函数中缺少正则项时，此时的损失函数仅在对观察值和预测值的偏差进行度量。一旦逻辑回归加入加入正则项之后，其损失函数除了要度量观察值和预测值的偏差以外，还要度量模型中的参数是否足够的小，而特征的数值变化范围和参数值的大小程度是非常相关的。因此对于带有正则化的逻辑回归模型，比如本文中建立的XGBoost天气敏感型销售概率预测模型，标准化是非常必要的。

结合以上三点，本文在建立TL-XGBoost天气敏感型模型时，需要对数据进行标准化处理来加速模型收敛速度，同时在建立SVM等模型来进行对比时，也需要对数据进行标准化处理来消除不同量纲的影响。

具体来讲，本文的数据转化步骤如下：

第一步，由于收集到的销售数据是日销量，因此对销售数据进行周平均数据以及月平均数据的转化，作为新的特征加入模型并与相应的产品进行匹配。

第二步，由于收集到的天气数据中缺少温差和前后两天温差，因此对数据进行处理，人工计算出每日的温差并和对应的销售数据相匹配，人工计算出前日平均气温与今日平均气温的差值并和相应的销售数据相匹配。

第三步，对数据进行标准化处理，本文采取Z-score标准化的方法去除不同变量量纲不同的影响，并使得数据的分布符合一个正态分布的形式，标准差为1，均值为0。具体公式如下：

其中，为原始数据，为均值，为标准差。

第四步，对销量数据进行 的数据转化。对数据取对数后并不会改变数据的相对关系，因为对数函数在定义域之内是一个单调递增的函数。与此同时，对销量数据对数化存在诸多好处。第一，方便模型进行计算，因为对数化缩小了数据的绝对数值。第二，方便更清晰的比较天气影响因素和其他因素对于销量产生的影响，对于销量数据，波动范围从0到304，如果两天的销量差了个位数，对于企业来说不需要引起特别的重视因为销量波动并不大，同时反应了这两天天气因素以及其他因素的改变对于销量产生的影响并不大。第三，对销量数据取对数,使得数据变得更加平稳，但又不会改变数据之间的关系，有助于减小模型的共线性和异方差性等。

第五步，将定性特征转换为定量特征，设置虚拟变量。许多场景下，定性特征不能直接使用，这是因为一些机器学习算法，比如XGBoost模型，设置上只能接受输入定量特征。因此，需要将定性特征人工的进行转换成定量特征。最简单的方式是为每一种定性值指定一个定量值，但是这种方式过于灵活，增加了调参的工作，比如白天以及夜晚不同的天气状态进行编号。本文使用one-hot编码的方式将定性特征转换成定量特征，具体来说：假设一个特征对应的定性状态一共有 种定性值，则one-hot编码将这一个特征扩展为 种不同的特征，当原始的特征值为第 种定性值时，将第 个扩展特征的值赋值为1，与此同时，其他扩展特征的值则赋值为0。one-hot编码的方式相比直接指定的方式，不用增加调参的工作，对于线性模型来说，使用one-hot编码后的特征可达到非线性的效果。除此以外，本文在销售概率预测时将训练数据中的销售量转换为0-1变量，以此来表示是否销售。

**4.3特征工程**

* + 1. **特征构建**

特征构建指的是从原始数据中去人为的找到一些具有实际商业意义或者实际物理意义的特征加入到模型的训练中以提高模型的精准度。特征构建有两种常见的方式，第一种是通过对数据的敏感型和实际业务的经验来帮助进行特征构建，第二中则是通过对已有特征的属性进行分割和结合。

针对第一种方法，前人总结和完善的天气敏感型需求理论为本文的特征构建提供了理论和实际物理意义上的基础。同时，在第三章节中，本文对天气敏感因素进行了充分讨论，包括使用皮尔森相关系数法来衡量单个特征和被解释变量之间的关系，如下图所示。

|  | | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 模型 | | 非标准化系数 | | 标准系数 | t | Sig. | 相关性 | | | 共线性统计量 | |
| B | 标准 误差 | 试用版 | 零阶 | 偏 | 部分 | 容差 | VIF |
| 1 | (常量) | 362.655 | 14.753 |  | 24.581 | .000 |  |  |  |  |  |
| 白天天气 | .439 | 1.857 | .013 | .237 | .813 | -.070 | .013 | .012 | .852 | 1.174 |
| 晚上天气 | -4.798 | 1.764 | -.147 | -2.720 | .007 | -.140 | -.144 | -.140 | .908 | 1.102 |
| 最低气温 | -1.107 | .345 | -.173 | -3.212 | .001 | -.198 | -.169 | -.165 | .912 | 1.096 |
| 温差 | -1.632 | 1.383 | -.069 | -1.181 | .239 | -.131 | -.063 | -.061 | .763 | 1.310 |
| 早上风向 | -1.833 | 2.870 | -.075 | -.639 | .523 | -.130 | -.034 | -.033 | .193 | 5.191 |
| 早上风力 | -2.129 | 4.936 | -.072 | -.431 | .667 | .026 | -.023 | -.022 | .094 | 10.588 |
| 晚上风向 | -.341 | 2.824 | -.014 | -.121 | .904 | -.119 | -.006 | -.006 | .193 | 5.175 |
| 晚上风力 | 1.298 | 4.821 | .045 | .269 | .788 | .030 | .014 | .014 | .094 | 10.584 |
| 前后温差 | 1.833 | 1.366 | .072 | 1.342 | .180 | .100 | .072 | .069 | .915 | 1.093 |
| a. 因变量: 品类1 | | | | | | | | | | | |

图表7 品类1-总销量-天气变量回归方程的回归系数图

通过上文中对天气变量进行的刻画和分析，本文选择了包括白天天气状态、晚上天气状态、最低气温、最高气温、平均气温、温差、早上风力、晚上风力、早上风向、晚上风向、前后温差在内的11种天气变量作为特征输入进模型，虽然其中一些天气变量在上文的讨论中和销量没有显著的线性关系，但是不能排除其通过非线性的方式影响销量的可能性。

针对第二种特征构建的方法，即对已有特征的属性进行分割和结合。

（1）销量相关特征

本文同销量这一特征进行分割和结合，设计出包括上一周总销量、上一周每天平均销量、上一周对应星期销量、上月日平均销量、上月周平均销量以及近一个月平均销量六种销量相关指标作为特征进入模型。

（2）节假日相关特征

在进行节假日相关特征的构建之前，本文首先对节假日与销量之间进行数据可视化探索，下图所示为SKU维度节假日与销量的散点图，可以明显观测到是否是节假日对于销量数据的分布有较大的影响，因此在构建模型预测时，需要考虑对节假日这一因素进行讨论并加入到特征池中。

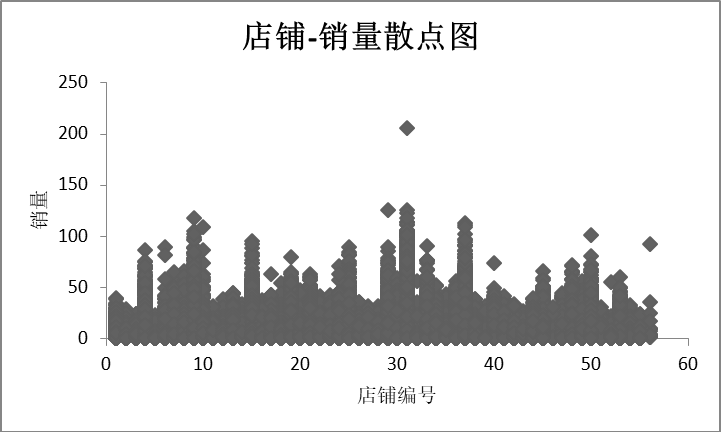


图表14 SKU层面节假日-销量散点图

所以，除了销量这一特征以外，本文同样对节假日这一特征进行了分割和结合，设计出包括节假日、当天是否放假、本次节假日天数在内的三个节假日特征以及根据时间序列刻画出的星期特征来将消费者在不同星期下的需求加入讨论。

（3）店铺相关特征

在进行店铺相关特征的构建之前，本文首先对店铺与销量之间进行数据可视化探索，下图所示为SKU维度节假日与销量的散点图，由图可观测到，销量数据由于店铺的不同而呈现出不同的分布规律，其中一部分店铺的销量波动范围与最大值显著大于其他的店铺，不同店铺对于销量不同造成影响的原因可能是店铺自身的服务水平、店铺所在地消费水平、人流量、店铺外观对于消费者的吸引程度等等，基于此，本文考虑将店铺作为一种特征并在后续特征工程部分中围绕这一因素进行深入挖掘。



图表15 SKU层面店铺-销量散点图

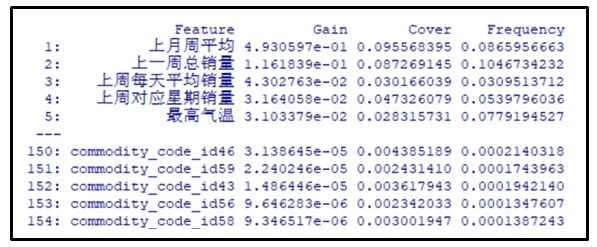
所以，除了销量、节假日等特征以外，本文同样对店铺属性这一特征进行了分割和结合，设计出包括店铺编号、商圈类型、地区在内的三个店铺属性特征来将不同店铺的属性对于消费者需求的不同影响加入讨论当中。

* + 1. **基于机器学习模型的特征选择**

上文中，通过对天气变量进行刻画和分析，以及对已有的销量特征进行属性上的分割和结合，本文一共构建了包括白天天气状态、晚上天气状态、最低气温、最高气温、平均气温、温差、早上风力、晚上风力、早上风向、晚上风向、前后温差在内的11种天气变量，包括上一周总销量、上一周每天平均销量、上一周对应星期销量、上月日平均销量、上月周平均销量以及近一个月平均销量的6种销量相关变量，包括节假日、当天是否放假、本次节假日天数在内的3个节假日相关变量以及根据时间序列刻画出的星期变量，包括店铺编号、商圈类型、地区在内的3个店铺属性变量，共计24个产品特征加入到模型的训练中。

其中部分特征，比如上文中所讨论的天气特征之间，可能存在共线性的关系。对于线性回归模型，因为线性回归模型的基本假设是，变量之间不是线性相关的，所以在建模前需要解决共线性问题。本文由于选择的Two-stage TrAdaboost.R2模型和XGBoost模型都以boosting树模型作为基础，因此选择不对共线性问题进行处理，因为在树模型的变量选择的过程中，共线性的多个变量中往往只有一个变量会被选中加入到模型中。

除此以外，本文在建立TL-XGBoost模型时并不采取主成分分析（PCA）等方法对特征进行降维。这么做的原因，第一是本文总共设计了24种产品特征，本文的SKU维度的数据条目数远远大于特征的维度。第二是因为对于树模型而言，算法本身会有一个特征选择的过程，属于特征选择中的嵌入式方法，即贪婪算法会遍历每一个新变量对结果的影响，只有当这一个变量对结果的影响较大时才会被加入到树模型当中。比如对于XGBoost而言，会根据每个特征在每棵树中的分裂次数和来计算每个特征的打分，如下图所示。



图表16 SKU维度XGBoost特征权重

**4.4结果分析与模型对比**

* + 1. **城市维度需求预测结果分析**

本文对于上海地区门店的四种不同类别的咖啡总销量使用Two-stage TrAdaboost.R2模型进行预测，四种不同类别的咖啡总销量预测结果如下四幅图所示：

图表17 品类1需求趋势预测值

图表18 品类2需求趋势预测值

图表19 品类3需求趋势预测值

图表20 品类4需求趋势预测值

以上四幅图展示了商品类别1、商品类别2、商品类别3以及商品类别4的观测值和预测值的比较，展示了Two-stage TrAdaboost.R2模型的预测趋势结果。从中可以观察到，Two-stage TrAdaboost.R2模型的需求预测趋势基本上与真实销量趋势是一致的，说明在需求趋势预测上，Two-stage TrAdaboost.R2模型具有较高的准确度。

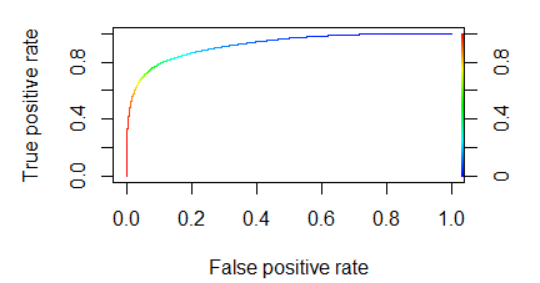
图表21 四种品类MAPE值比较

除了对四种品类的销量预测趋势进行比较以外，本文还对四种品类的MAPE值进行了计算和汇总，由上图可观测出，品类1的MAPE值为14.19%，品类2的MAPE值为9.69%，品类3的MAPE值为12.13%，品类4的MAPE值为11.85%，可以总结出，在对城市维度品类总销量的需求预测上，Two-stage TrAdaboost.R2模型具有较高的准确性。

* + 1. **SKU维度需求预测结果分析**

对于SKU维度的需求预测预测分为两个部分，分别是对每个产品的销售概率预测和对所有SKU的初始需求预测。

针对产品的销售概率预测，采用的是XGBoost以logistic regression作为基学习器建立模型。销售概率的XGBoost预测ROC评价曲线如下图所示：



图表22 销售概率预测ROC评价曲线

通过上图可以观测到，XGBoost在测试集在上的AUC最终达到0.92，说明XGBoost在对产品的销售概率预测上表现较好。

针对SKU的初始需求预测，采用的是XGBoost以linear regression作为基学习器建立模型，同时设置最大树深为5，得到SKU层级的初始预测需求，最后再通过GAD模型对销量进行贪婪聚合与分解，最终输出修正后的SKU预测值。本文通过对比TL-XGBoost模型、XGBoost模型、未考虑天气变量的XGBoost模型、梯度提升树（GBDT）模型以及支持向量机模型（SVM），计算出各个机器学习模型的MAE值如下图所示。

图表23 四种模型MAE值比较

通过结果对比可以发现，TL-XGBoost模型MAE值最低，预测精度在四种模型中最高，相比较XGBoost模型，TL-XGBoost通过引入迁移学习思想和对SKU销量进行最终修正，在预测结果上取得了一定的提升，MAE减少了0.018，预测精度在四种模型中取得优势。同时以XGBoost为基础的模型，预测结果要明显好于SVM和GBDT，未修正的XGBoost比SVM和GBDT在MAE上分别减少了0.027和0.201。结合以上这两点，可以得到结论：

1. XGBoost模型对于庞大数据集的适应性更强，在运算速度更快的情况下仍能保持较高的精确度。
2. TL-XGBoost模型相比较单一结构的模型，在预测精准度上更有优势。

除此以外，本文还对未加入天气敏感型因素的XGBoost模型进行了建模，并与TL-XGBoost和XGBoost模型进行对比，对比结果如下图所示。

图表24 天气因素建模MAE值比较

通过上图三个模型之间的比较我们可以发现，相比较与去除天气敏感因素的XGBoost模型，考虑天气敏感型因素的XGBoost模型在MAE值上减少了0.006，因此在构建模型时将天气敏感型因素纳入考虑可以提升预测精准度。

**第五章 结论与展望**

**5.1 研究结论**

本文以中国某咖啡连锁品牌实体门店为研究背景，对机器学习模型在天气敏感型商品的需求预测中的应用问题进行了研究，将天气影响因子考虑到模型的构建中，结合迁移学习和供应链层次预测方法，建立了TL-XGBoost的天气敏感型需求预测模型来对XGBoost模型的性能进行了增强。本文首先使用Two-stage TrAdaboost.R2模型在城市维度对四个不同品类的每日总销量进行预测，然后使用XGBoost模型对每一个SKU销量进行预测并利用GAD贪婪聚合分解模型进行最终的销量预测值修正。实证分析的结果表明，本文所提出的TL-XGBoost的天气敏感型需求预测模型不仅利用了天气信息中的价值，同时利用了供应链内部不同维度和层级的有效信息对天气敏感型商品进行预测，充分的应用了机器学习算法在面对庞大数据时的出色性能特点，TL-XGBoost模型在预测精确度上高过传统的单层次XGBoost模型、GBDT模型和SVM模型。本文提出的TL-XGBoost模型可以有效的进行中短期的需求预测，为零售企业的库存计划、促销计划、定价策略等业务活动和决策提供有效的支持。

本文的第一章和第二章节首先对研究背景、文献综述以及相关的基本理论和基础算法进行了介绍，为后两章的模型建立以及实证分析做下铺垫。之后，在第三章，本文首先对天气变量进行了刻画和选择，确定了四种不同的品类和天气敏感因素之间的关系。确定了天气变量之后，本文开始构建TL-XGBoost模型，模型分为三个部分，分别是TrAdaboost.R2模型在城市维度对品类总销量进行预测，XGBoost模型在门店层级对SKU的销售概率和初始需求进行预测，最后通过GAD模型对三者进行贪婪聚合分解。本文的第四章建立在第三章的基础之上，根据建立的模型，对上海地区某咖啡连锁零售品牌的门店销售数据进行大规模建模，并与包括单层次XGBoost模型、未考虑天气的XGBoost模型、GBDT模型以及SVM模型进行模型对比和验证，从而选择出最优的预测模型。

**5.2 研究局限**

本文的研究局限主要在以下几点：

第一，在第三章TL-XGBoost模型构建部分，本文在构建天气变量时只对11种天气变量进行了刻画和选择，主要原因在于天气数据收集的困难程度，本文在研究时无法即时获得2019年度至今的其他天气数据，只能通过编写爬虫软件从天气后报网站获取。除此以外，本文未对天气变量如何具体的影响销量进行深度的讨论与探索，包括天气变量是否会因店铺、区域、城市的不同，而对不同的品类产生不同的影响，这都是值得进一步讨论和挖掘的部分。

第二，在第四章实证部分，本文在SKU销量预测维度将TL-XGBoost模型与XGBoost模型、SVM模型、GBDT等模型进行了模型对比，但是并未实现与包括ELM、MLP等基于神经网络的机器学习模型进行对比。主要原因在于，ELM、MLP等基于神经网络的机器学习模型性能非常依赖于网络的构建方式，包括隐含层的设计、节点的设计等等，因此在缺乏相关网络设计经验的情况下，本文认为TL-XGBoost模型和未优化的神经网络模型进行比对是不合理的。

第三，在第四章实证部分，本文只实现了对于上海地区的四种产品SKU销量预测，未实现区域范围或者全国的天气影响因素建模，主要原因在于数据量过于庞大，对于单台电脑而言模型的训练时间过长。虽然XGBoost相较于其他模型优化了算法的收敛速度和计算速度，但是上海地区的销量就已经达到60万条，如果对全国范围的每一家店铺进行建模，数据量会异常的庞大，需要对已有知识和掌握软件进行更新。因此，本文认为在未来实现全国范围或者某个区域范围的全门店店铺建模是非常有必要的，因为可以充分挖掘和比较不同区域、不同门店类型和天气影响因素以及销量之间的相互影响关系。

**5.3 管理启示**

基于以上结论,本文认为企业可以从中获得以下几点管理启示：

第一，企业在构建需求预测模型时，应重视并充分考虑天气变量对于需求的影响。根据第三章的天气变量刻画和第四章的结果分析模型比较可以看出，天气变量和销量之间具有显著的相关性，同时当模型考虑天气变量等因素时，预测精准度可以得到进一步提升。因此，企业应对天气变量进行量化考虑，通过历史数据分析天气变化与产品或者服务之间的关系，从而为下一个销售季节做好准备。

第二，企业在对大数据建模时，可以考虑应用XGBoost模型等机器学习模型。机器学习模型在面对庞大的数据量时显示除了优异的性能优势，与此同时，XGBoost模型还表现出优秀的特征选择功能和缺失值处理功能，帮助企业更容易的构建高精度的需求预测模型。因此，企业应该考虑建立自己的数据库，规范化数据录入、贴标签等过程，为未来的大规模机器学习模型应用做好准备。

第三，企业对销售季节前、销售季节中以及销售季节后的需求管理给予足够的重视。销售季节前包括从产品或者服务的规划一直到上市面向消费者，这段时间内，企业应提前做好需求学习，对销售季中的需求进行预测，这种需求管理计划可以为上市计划提供充分的支持。在销售季节中，应持续的进行需求学习，因为此时市场环境时刻都在发生改变，持续对需求进行学习有助于企业对订货以及补货等策略进行调整。销售季节后的需求学习对于企业来说，可以帮助其对整个销售季的需求管理计划进行回顾，为下一个季度的需求计划做出调整和准备。

**5.4 未来展望**

基于上文的结论和研究的局限，本文对未来的研究方向进行展望，具体如下：

第一，对天气变量进行更深入的刻画，收集和建立更多的天气特征。同时对天气变量具体怎么影响商品需求进行探究，包括对各种天气变量与商圈、节假日等影响因素之间的相互作用进行探讨。

第二，本文只对上海区域的销售数据进行建模，未来可以进行区域范围或者全国范围的天气影响因素建模，充分挖掘天气变量、区域变量、店铺属性等特征与需求之间的关系，并为企业的管理提供更多的建议。

第三，本文在建立机器学习模型和模型比较时，未考虑神经网络类的模型，未来可以对神经网络进行设计，优化ELM等机器学习模型并加入到模型比较中。

致 谢

时光飞逝，在上海大学悉尼工商学院的四年即将走到尽头，作为一名工商管理的学生，我在这里收获了太多。上海大学悉尼工商学院的教师资源和课程设置，让我顺利的完成了从大一懵懂新生，到大四在相关领域累计了一定的经验，准备好进入下一阶段学习和工作的准毕业生。

首先，我要对我本科四年的导师，高峻峻老师表示感谢。在大一的夏季学期，非常幸运的进入了高峻峻老师的项目组，从而开展了持续3年的项目式学习。在高老师的项目组内，我不仅在供应链，需求链，新零售等领域进行了深入学习，同时还作为组内负责人带着项目组的研究成果参加了多项学科竞赛和创业比赛，累计了许多宝贵的经验。这三年里一点一点累积的经历和知识，最终以毕业论文的形式展示在大家面前。在论文的写作过程中，高峻峻老师为我提供了宝贵的指导意见，从选题到文献收集和阅读，从模型的建立到实证分析，论文的顺利完成绝对离不开高老师悉心的指导。

感谢姜爱萍老师，帮助我建立了扎实的统计学知识，为进一步的机器学习模型的学习打下了坚实的基础。

感谢郭鹏学长，在我数据收集、数据处理、模型的过程中遇到困难时，给予了我宝贵的帮助，帮助我顺利的使用Python和R完成机器学习模型的建模。

感谢王立峰学长，在我论文的选题、内容结构遇到问题时，耐心的与我进行讨论，帮助我梳理了整个论文的结构与思路。

感谢我的父母，没有你们物质和精神上的支持，我无法作到全身心的投入到学习之中，你们永远站在我的身后支持着我。

感谢我的室友和每一位朋友，在我熬夜、失意还有烦躁的时候，一直支持着我，让我保持乐观的心态面对苦难，热情的态度面对学术。

参考文献

1. 王忠玉. 2017年湖北省科学技术史学会年会论文集 [M]. 湖北省科学技术史学会. 2017.
2. 王忠玉. 科技在餐饮行业中的作用分析 [M]. 湖北省科学技术史学会. 2017.
3. 赵树梅, 徐晓红. “新零售”的含义、模式及发展路径 [J]. 中国流通经济, 2017, 31(05): 12-20.
4. 黄锐, 龚晓路. 快速消费品分销渠道管理[M]. 中国发展出版社, 2005.
5. 石甜甜, 高峻峻, 刘悦. 天气敏感型需求的内涵与管理[J]. 企业研究, 2013(5):70-71.
6. Persinger M A, Levesque B F. GEOPHYSICAL VARIABLES AND BEHAVIOR: XII. THE WEATHER MATRIX ACCOMMODATES LARGE PORTIONS OF VARIANCE OF MEASURED DAILY MOOD[J]. Perceptual & Motor Skills, 2011, 57(3):868-870.
7. Travers J, Payne V G. The impact of climate change on traditional seasonal ranges for high street women's wear in England[J]. Journal of Fashion Marketing & Management, 1998, 2(4):370-382.
8. Saunders E M. Stock Prices and Wall Street Weather[J]. American Economic Review, 1993, 83(5):1337-1345.
9. Trombley M A. Stock Prices and Wall Street Weather: Additional Evidence[J]. Quarterly Journal of Business & Economics, 1997, 36(3):11-21.
10. Murray K B, Muro F D, Finn A, et al. The effect of weather on consumer spending[J]. Journal of Retailing & Consumer Services, 2010, 17(6):512-520.
11. Cui Y, Basnet C. An exploratory study of supply chain risk management in the New Zealand fast food industry[J]. International Journal of Logistics Systems & Management, 2015, 20(2):199.
12. Bahng Y, Kincade D H. The relationship between temperature and sales: Sales data analysis of a retailer of branded women's business wear[J]. International Journal of Retail & Distribution Management, 2012, 40(6):410-426.
13. Huang C M, Chen S J, Yang S P, et al. One-day-ahead hourly forecasting for photovoltaic power generation using an intelligent method with weather-based forecasting models[J]. Iet Generation Transmission & Distribution, 2015, 9(14):1874-1882.
14. 贺蓉, 曾刚, 姚建刚,等. 天气敏感型神经网络在地区电网短期负荷预测中的应用[J]. 电力系统自动化, 2001, 25(17):32-35.
15. 张甫仁. 基于气象热舒适度的建筑能耗灰色神经网络预测[J]. 建筑科学, 2007, 23(10):49-52.
16. Divakar S, Ratchford B T, Shankar V. CHAN4CAST: A Multichannel, Multiregion Sales Forecasting Model and Decision Support System for Consumer Packaged Goods[J]. Marketing Science, 2005, 24(3):334-350.
17. Lam S Y, Vandenbosch M, Hulland J, et al. Evaluating Promotions in Shopping Environments: Decomposing Sales Response into Attraction, Conversion, and Spending Effects[J]. Marketing Science, 2001, 20(2):194-215.
18. 黄瀚文. 天气敏感性需求预测在餐饮电商外卖中的应用[D]. 上海大学, 2016.
19. Zhao H, Moore G W K. A seasonally lagged signal of the North Atlantic Oscillation (NAO) in the North Pacific[J]. International Journal of Climatology, 2006, 26(7):957-970.
20. 唐浩. 考虑典型气象因素影响的湖南电网夏季日最大负荷预测[D].长沙理工大学,2013.
21. 莫降涛, 潘红娟, 米方方,等. 季节性需求产品的预订销售模型[J]. 运筹与管理, 2009, 18(3):64-68.
22. Arunraj N S, Ahrens D. A hybrid seasonal autoregressive integrated moving average and quantile regression for daily food sales forecasting[J]. International Journal of Production Economics, 2015, 170:321-335.
23. 危万虎, 张鸿雁. 武汉市夏季空调销量的气候预测及空调指数研究[J]. 暴雨灾害, 2000, 19(4):22-24.
24. Frank C, Garg A, Sztandera L, et al. Forecasting women's apparel sales using mathematical modeling[J]. International Journal of Clothing Science & Technology, 2013, 15(2):107-125.
25. 赵成柏, 毛春梅. 基于ARIMA和BP神经网络组合模型的我国碳排放强度预测[J]. 长江流域资源与环境, 2012, 21(6):665-671.
26. 景亚平, 张鑫, 罗艳. 基于灰色神经网络与马尔科夫链的城市需水量组合预测[J]. 西北农林科技大学学报:自然科学版, 2011, 39(7):229-234.
27. 高峻峻,谭崇利,刘悦,尹亚锋.一类产品需求预测的支持向量机方法[J].上海大学学报(自然科学版),2009,15(01):71-76.
28. 郑凯文, 杨超. 基于迭代决策树(GBDT)短期负荷预测研究[J]. 贵州电力技术, 2017, 20(2):82-84.
29. 桂良明, 夏永俊, 李海山,等. 基于RF-GBDT的燃煤锅炉NO\_x排放预测[J]. 电站系统工程, 2017(2):5-8.
30. Pan S J , Yang Q . A Survey on Transfer Learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(10):1345-1359.
31. Rettinger, Achim, Zinkevich, Martin, and Bowling, Michael. Boosting expert ensembles for rapid concept recall[J]. In Proceedings of the 21st National Conference on Artificial Intelligence, July 2006.
32. Dai W, Yang Q, Xue G R, et al. Boosting for transfer learning[C]. International Conference on Machine Learning, 2007.
33. Freund, Yoav and Schapire, Robert. A decision theoretic generalization of online learning and an application to boosting[J]. Journal of Computer and System Sciences, 1997, 55:119–139.
34. Pardoe D , Stone P . Adapting Price Predictions in TAC SCM.[C]. International Conference on Agent-mediated Electronic Commerce & Trading Agent Design & Analysis. Springer-Verlag, 2007.
35. 史荧中, 王士同, 蒋亦樟, et al. 迁移学习支持向量回归机[J]. 计算机应用, 2013, 33(11):3084-3089.
36. Chen T , Guestrin C . XGBoost: A Scalable Tree Boosting System[J]. 2016.
37. KöK A. Management of Product Variety in Retail Operations [M]. PENNSYLVANIA U U O. 1993.
38. SYNTETOS ARIS A., BABAI ZIED, BOYLAN JOHN E., et al. Supply chain forecasting: Theory, practice, their gap and the future [J]. European Journal of Operational Research, 2016, 252(1): 1-26.

附录：

附录一：英文资料原文

A Survey on Transfer Learning

**1 Introduction**

Data mining and machine learning technologies have already achieved significant success in many knowledge engineering areas including classification, regression and clustering (e.g., [1], [2]). However, many machine learning methods work well only under a common assumption: the training and test data are drawn from the same feature space and the same distribution. When the distribution changes, most statistical models need to be rebuilt from scratch using newly collected training data. In many real world applications, it is expensive or impossible to re-collect the needed training data and rebuild the models. It would be nice to reduce the need and effort to re-collect the training data. In such cases, knowledge transfer or transfer learning between task domains would be desirable.

Many examples in knowledge engineering can be found where transfer learning can truly be beneficial. One example is Web document classification [3], [4], [5], where our goal is to classify a given Web document into several predefined categories. As an example in the area of Web-document classification (see, e.g., [6]), the labeled examples may be the university Web pages that are associated with category information obtained through previous manual-labeling efforts. For a classification task on a newly created Web site where the data features or data distributions may be different, there may be a lack of labeled training data. As a result, we may not be able to directly apply the Web-page classifiers learned on the university Web site to the new Web site. In such cases, it would be helpful if we could transfer the classification knowledge into the new domain.

The need for transfer learning may arise when the data can be easily outdated. In this case, the labeled data obtained in one time period may not follow the same distribution in a later time period. For example, in indoor WiFi localization problems, which aims to detect a user’s current location based on previously collected WiFi data, it is very expensive to calibrate WiFi data for building localization models in a largescale environment, because a user needs to label a large collection of WiFi signal data at each location. However, the WiFi signal-strength values may be a function of time, device or other dynamic factors. A model trained in one time period or on one device may cause the performance for location estimation in another time period or on another device to be reduced. To reduce the re-calibration effort, we might wish to adapt the localization model trained in one time period (the source domain) for a new time period (the target domain), or to adapt the localization model trained on a mobile device (the source domain) for a new mobile device (the target domain), as done in [7].

As a third example, consider the problem of sentiment classification, where our task is to automatically classify the reviews on a product, such as a brand of camera, into positive and negative views. For this classification task, we need to first collect many reviews of the product and annotate them. We would then train a classifier on the reviews with their corresponding labels. Since the distribution of review data among different types of products can be very different, to maintain good classification performance, we need to collect a large amount of labeled data in order to train the reviewclassification models for each product. However, this datalabeling process can be very expensive to do. To reduce the effort for annotating reviews for various products, we may want to adapt a classification model that is trained on some products to help learn classification models for some other products. In such cases, transfer learning can save a significant amount of labeling effort [8].

In this survey article, we give a comprehensive overview of transfer learning for classification, regression and clustering developed in machine learning and data mining areas. There has been a large amount of work on transfer learning for reinforcement learning in the machine learning literature (e.g., [9], [10]). However, in this paper, we only focus on transfer learning for classification, regression and clustering problems that are related more closely to data mining tasks. By doing the survey, we hope to provide a useful resource for the data mining and machine learning community. The rest of the survey is organized as follows. In the next four sections, we first give a general overview and define some notations we will use later. We then briefly survey the history of transfer learning, give a unified definition of transfer learning and categorize transfer learning into three different settings (given in Table 2 and Figure 2). For each setting, we review different approaches, given in Table 3 in detail. After that, in Section 6, we review some current research on the topic of “negative transfer”, which happens when knowledge transfer has a negative impact on target learning. In Section 7, we introduce some successful applications of transfer learning and list some published data sets and software toolkits for transfer learning research. Finally, we conclude the article with a discussion of future works in Section 8.

**2.Overview**

**2.1 A Brief History of Transfer Learning**

Traditional data mining and machine learning algorithms make predictions on the future data using statistical models that are trained on previously collected labeled or unlabeled training data [11], [12], [13]. Semi-supervised classification [14], [15], [16], [17] addresses the problem that the labeled data may be too few to build a good classifier, by making use of a large amount of unlabeled data and a small amount of labeled data. Variations of supervised and semi-supervised learning for imperfect datasets have been studied; for example, Zhu and Wu [18] have studied how to deal with the noisy classlabel problems. Yang et al. considered cost-sensitive learning [19] when additional tests can be made to future samples. Nevertheless, most of them assume that the distributions of the labeled and unlabeled data are the same. Transfer learning, in contrast, allows the domains, tasks, and distributions used in training and testing to be different. In the real world, we observe many examples of transfer learning. For example, we may find that learning to recognize apples might help to recognize pears. Similarly, learning to play the electronic organ may help facilitate learning the piano. The study of Transfer learning is motivated by the fact that people can intelligently apply knowledge learned previously to solve new problems faster or with better solutions. The fundamental motivation for Transfer learning in the field of machine learning was discussed in a NIPS-95 workshop on “Learning to Learn” 1, which focused on the need for lifelong machine-learning methods that retain and reuse previously learned knowledge.

Research on transfer learning has attracted more and more attention since 1995 in different names: learning to learn, life-long learning, knowledge transfer, inductive transfer, multi-task learning, knowledge consolidation, contextsensitive learning, knowledge-based inductive bias, meta learning, and incremental/cumulative learning [20]. Among these a closely related learning technique to transfer learning is the multi-task learning framework [21], which tries to learn multiple tasks simultaneously even when they are different. A typical approach for multi-task learning is to uncover the common (latent) features that can benefit each individual task. In 2005, the Broad Agency Announcement (BAA) 05-29 of Defense Advanced Research Projects Agency (DARPA)’s Information Processing Technology Office (IPTO) 2 gave a new mission of transfer learning: the ability of a system to recognize and apply knowledge and skills learned in previous tasks to novel tasks. In this definition, transfer learning aims to extract the knowledge from one or more source tasks and applies the knowledge to a target task. In contrast to multi-task learning, rather than learning all of the source and target tasks simultaneously, transfer learning cares most about the target task. The roles of the source and target tasks are no longer symmetric in transfer learning. Figure 1 shows the difference between the learning processes of traditional and transfer learning techniques. As we can see, traditional machine learning techniques try to learn each task from scratch, while transfer learning techniques try to transfer the knowledge from some previous tasks to a target task when the latter has fewer high-quality training data.

Today, transfer learning methods appear in several top venues, most notably in data mining (ACM KDD, IEEE ICDM and PKDD, for example), machine learning (ICML, NIPS, ECML, AAAI and IJCAI, for example) and applications of machine learning and data mining (ACM SIGIR, WWW and ACL for example) 3. Before we give different categorizations of transfer learning, we first describe the notations used in this article.

**2.2 Transferring Knowledge of Instances**

The instance-transfer approach to the inductive transfer learning setting is intuitively appealing: although the source domain data cannot be reused directly, there are certain parts of the data that can still be reused together with a few labeled data in the target domain. Dai et al. [6] proposed a boosting algorithm, TrAdaBoost, which is an extension of the AdaBoost algorithm, to address the inductive transfer learning problems. TrAdaBoost assumes that the source and target domain data use exactly the same set of features and labels, but the distributions of the data in the two domains are different. In addition, TrAdaBoost assumes that, due to the difference in distributions between the source and the target domains, some of the source domain data may be useful in learning for the target domain but some of them may not and could even be harmful. It attempts to iteratively re-weight the source domain data to reduce the effect of the “bad” source data while encourage the “good” source data to contribute more for the target domain. For each round of iteration, TrAdaBoost trains the base classifier on the weighted source and target data. The error is only calculated on the target data. Furthermore, TrAdaBoost uses the same strategy as AdaBoost to update the incorrectly classified examples in the target domain while using a different strategy from AdaBoost to update the incorrectly classified source examples in the source domain. Theoretical analysis of TrAdaBoost in also given in [6]. Jiang and Zhai [30] proposed a heuristic method to remove “misleading” training examples from the source domain based on the difference between conditional probabilities P(yT |xT ) and P(yS|xS). Liao et al.[31] proposed a new active learning method to select the unlabeled data in a target domain to be labeled with the help of the source domain data. Wu and Dietterich [53] integrated the source domain (auxiliary) data an SVM framework for improving the classification performance.

**2.3 Transferring Relational Knowledge**

Different from other three contexts, the relational-knowledgetransfer approach deals with transfer learning problems in relational domains, where the data are non-i.i.d. and can be represented by multiple relations, such as networked data and social network data. This approach does not assume that the data drawn from each domain be independent and identically distributed (i.i.d.) as traditionally assumed. It tries to transfer the relationship among data from a source domain to a target domain. In this context, statistical relational learning techniques are proposed to solve these problems. Mihalkova et al.[50] proposed an algorithm TAMAR that transfers relational knowledge with Markov Logic Networks (MLNs) across relational domains. MLNs [56] is a powerful formalism, which combines the compact expressiveness of first order logic with flexibility of probability, for statistical relational learning. In MLNs, entities in a relational domain are represented by predicates and their relationships are represented in first-order logic. TAMAR is motivated by the fact that if two domains are related to each other, there may exist mappings to connect entities and their relationships from a source domain to a target domain. For example, a professor can be considered as playing a similar role in an academic domain as a manager in an industrial management domain. In addition, the relationship between a professor and his or her students is similar to the relationship between a manager and his or her workers. Thus, there may exist a mapping from professor to manager and a mapping from the professorstudent relationship to the manager-worker relationship. In this vein, TAMAR tries to use an MLN learned for a source domain to aid in the learning of an MLN for a target domain. Basically, TAMAR is a two-stage algorithm. In the first step, a mapping is constructed from a source MLN to the target domain based on weighted pseudo loglikelihood measure (WPLL). In the second step, a revision is done for the mapped structure in the target domain through the FORTE algorithm [57], which is an inductive logic programming (ILP) algorithm for revising first order theories. The revised MLN can be used as a relational model for inference or reasoning in the target domain. In the AAAI-2008 workshop on transfer learning for complex tasks 4, Mihalkova et al.[51] extended TAMAR to the single-entity-centered setting of transfer learning, where only one entity in a target domain is available. Davis et al.[52] proposed an approach to transferring relational knowledge based on a form of second-order Markov logic. The basic idea of the algorithm is to discover structural regularities in the source domain in the form of Markov logic formulas with predicate variables, by instantiating these formulas with predicates from the target domain.

附录二：英文资料翻译

迁移学习理论研究

**1 引言**

数据挖掘和机器学习技术已经在许多知识工程领域取得了显著的成功，包括分类、回归和聚类（例如[1]、[2]）。然而，许多机器学习方法只有在一个共同的假设下才能很好地工作：训练和测试数据是从相同的特征空间和相同的分布中提取的。当分布发生变化时，大多数统计模型需要使用新收集的训练数据从头重建。在许多实际应用中，重新收集所需的培训数据并重建模型是昂贵的或不可能的。最好减少重新收集培训数据的需要和努力。在这种情况下，需要在任务域之间进行知识转移或转移学习。

在知识工程的许多例子中，转移学习确实是有益的。一个例子是Web文档分类[3]、[4]、[5]，我们的目标是将给定的Web文档分类为几个预定义的类别。作为Web文档分类领域的一个示例（参见，例如[6]），标记的示例可以是与通过以前的手动标记工作获得的类别信息相关联的大学网页。对于新创建的网站上的分类任务，如果数据功能或数据分布可能不同，则可能缺少标记的培训数据。因此，我们可能无法将在大学网站上学习到的网页分类器直接应用到新的网站上。在这种情况下，如果我们能够将分类知识转移到新的领域中，这将是很有帮助的。

当数据很容易过时时，就可能需要转移学习。在这种情况下，在一个时间段内获得的标记数据在以后的时间段内可能不会遵循相同的分布。例如，在基于先前收集的WiFi数据检测用户当前位置的室内WiFi定位问题中，由于用户需要在每个位置标记大量WiFi信号数据，因此在大范围环境中校准WiFi数据以构建定位模型非常昂贵。然而，WiFi信号强度值可能是时间、设备或其他动态因素的函数。在一个时间段或一个设备上训练的模型可能导致在另一个时间段或另一个设备上的位置估计性能降低。为了减少重新校准工作，我们可能希望将在一个时间段（源域）内培训的本地化模型适应新的时间段（目标域），或者将在移动设备（源域）上培训的本地化模型适应新的移动设备（目标域），如[7]所述。

作为第三个例子，考虑一下情绪分类的问题，我们的任务是自动将产品的评论（如照相机品牌）分类为正面和负面的观点。对于这个分类任务，我们需要首先收集产品的许多评论，并对它们进行注释。然后，我们将用相应的标签对评审进行分类器培训。由于评审数据在不同类型产品之间的分布可能非常不同，为了保持良好的分类性能，我们需要收集大量的标记数据，以便为每个产品培训评审分类模型。然而，这种数据标记过程可能非常昂贵。为了减少为各种产品注释评论的工作，我们可能需要修改一个分类模型，该模型在某些产品上经过培训，以帮助学习其他产品的分类模型。在这种情况下，转移学习可以节省大量的标记工作[8]。

本文对机器学习和数据挖掘领域中的分类、回归和聚类转移学习进行了全面的综述。在机器学习文献（如[9]，[10]）中，有大量关于转移学习以加强学习的工作。然而，在本文中，我们只关注与数据挖掘任务密切相关的分类、回归和聚类问题的转移学习。通过调查，我们希望能为数据挖掘和机器学习社区提供一个有用的资源。调查的其余部分组织如下。在接下来的四个部分中，我们首先给出了一个概述，并定义了稍后将使用的一些注释。然后，我们简单地调查了转移学习的历史，给出了转移学习的统一定义，并将转移学习分为三种不同的设置（如表2和图2所示）。对于每种设置，我们将详细地回顾表3中给出的不同方法。在此之后，在第6节中，我们回顾了当前关于“负迁移”这一主题的一些研究，这是在知识迁移对目标学习产生负面影响时发生的。在第7节中，我们介绍了一些成功的转移学习应用程序，并列出了一些已发布的转移学习研究数据集和软件工具包。最后，我们在第8节中讨论了未来的工作，以此结束本文。

**2.概述**

**2.1 迁移学习历史概述**

传统的数据挖掘和机器学习算法使用统计模型对未来的数据进行预测，这些统计模型根据之前收集的标记或未标记的训练数据进行训练[11]、[12]、[13]。半监督分类[14]、[15]、[16]、[17]通过使用大量未标记数据和少量标记数据，解决了标记数据可能太少而无法构建良好的分类器的问题。研究了不完全数据集的监督学习和半监督学习的变化，如朱和吴[18]研究了如何处理噪声类标签问题。杨等人。当可以对未来的样本进行额外测试时，考虑成本敏感性学习[19]。尽管如此，它们中的大多数都假定标记数据和未标记数据的分布是相同的。相比之下，转移学习允许培训和测试中使用的域、任务和分布是不同的。在现实世界中，我们观察到许多转移学习的例子。例如，我们可能会发现学习识别苹果可能有助于识别梨。同样，学习电子琴也有助于学习钢琴。转移学习的研究是由这样一个事实所推动的，即人们可以智能地应用以前学到的知识，更快或更好地解决新问题。在NIPS-95“学习-学习”1讲习班上讨论了机器学习领域中转移学习的基本动机，该讲习班重点讨论了终身机器学习方法的需求，这些方法保留和重用以前学习的知识。

自1995年以来，转移学习的研究以不同的名称越来越受到关注：学习-学习、终身学习、知识转移、归纳转移、多任务学习、知识整合、上下文敏感学习、基于知识的归纳偏倚、元学习和增量/累积学习[20]。其中，转移学习的一种密切相关的学习技术是多任务学习框架[21]，它试图同时学习多个任务，即使它们是不同的。一个典型的多任务学习方法是发现共同的（潜在的）特征，可以有益于每个单独的任务。2005年，国防高级研究计划局（DARPA）信息处理技术办公室（IPTO）2的广泛机构公告（BAA）05-29给出了一个新的转移学习任务：系统识别和应用以前任务中所学知识和技能的能力。在这个定义中，转移学习的目的是从一个或多个源任务中提取知识，并将知识应用到目标任务中。与多任务学习相比，转移学习最关心的是目标任务，而不是同时学习所有源任务和目标任务。在转移学习中，源任务和目标任务的角色不再是对称的。图1显示了传统学习技术和转移学习技术的学习过程之间的差异。正如我们所看到的，传统的机器学习技术试图从零开始学习每一项任务，而转移学习技术则试图将以前的一些任务中的知识转移到目标任务中，而目标任务中的高质量培训数据较少。

今天，传输学习方法出现在几个顶级领域，最显著的是在数据挖掘（例如，ACM KDD、IEEE ICDM和PKDD）、机器学习（例如，ICML、NIPS、ECML、AAAI和IJCAI）以及机器学习和数据挖掘（例如，ACM SIGIR、WWW和ACL）的应用3。在给出不同的转移学习分类之前，我们首先描述了本文中使用的符号。

**2.2 样本的迁移学习**

归纳转移学习设置的实例转移方法具有直观的吸引力：尽管源域数据不能直接重用，但目标域中的某些部分数据仍然可以与一些标记数据一起重用。傣族等。[6]提出了一种提升算法TrAdaBoost，它是AdaBoost算法的扩展，用于解决归纳转移学习问题。TrAdaBoost假定源域和目标域数据使用完全相同的特征和标签集，但两个域中的数据分布不同。此外，TrAdaBoost假设，由于源域和目标域之间的分布不同，一些源域数据可能对目标域的学习有用，但其中一些数据可能不有用，甚至可能有害。它试图迭代地重新加权源域数据，以减少“坏”源数据的影响，同时鼓励“好”源数据为目标域贡献更多。对于每一轮迭代，TrAdaBoost都会根据加权的源数据和目标数据训练基本分类器。仅对目标数据计算错误。此外，TrAdaBoost使用与AdaBoost相同的策略更新目标域中错误分类的示例，同时使用与adaboost不同的策略更新源域中错误分类的源示例。文献[6]中也给出了TrAdaBoost的理论分析。蒋、翟[30]基于条件概率p（yt xt）与p（ys xs）的差异，提出了一种启发式方法，从源域中去除“误导性”训练实例。廖等[31]提出了一种新的主动学习方法，利用源域数据选择待标记目标域中的未标记数据。Wu和Dieterich[53]整合了源域（辅助）数据和SVM框架，以提高分类性能。

**2.3 传递关系知识**

与其他三种上下文不同，关系知识转移方法处理关系域中的转移学习问题，其中数据是非I.I.D.的，并且可以由多个关系表示，例如网络数据和社会网络数据。这种方法不假定从每个域中提取的数据是独立的，并且与传统假设的相同分布（I.I.D.）。它试图将数据之间的关系从源域传输到目标域。在此背景下，提出了统计关系学习技术来解决这些问题。Mihalkova等人[50]提出了一种算法Tamar，该算法通过关系域与马尔可夫逻辑网络（MLN）传输关系知识。MLNS[56]是一种强大的形式主义，它结合了一阶逻辑的紧凑表现性和概率的灵活性，用于统计关系学习。在MLN中，关系域中的实体用谓词表示，它们的关系用一阶逻辑表示。Tamar的动机是，如果两个域相互关联，那么可能存在将实体及其关系从源域连接到目标域的映射。例如，可以认为教授在学术领域扮演着类似的角色，在工业管理领域扮演着管理者的角色。此外，教授与学生之间的关系类似于经理与员工之间的关系。因此，可能存在从教授到经理的映射，以及从教授关系到经理-工人关系的映射。在这种情况下，Tamar尝试使用源域的MLN来帮助学习目标域的MLN。基本上，tamar是一个两阶段的算法。在第一步中，基于加权伪对数似然测度（WPLL），构造了从源MLN到目标域的映射。在第二步中，通过forte算法[57]对目标域中的映射结构进行了修改，该算法是一种用于修改一阶理论的归纳逻辑编程（ILP）算法。修正后的MLN可以作为目标域推理的关系模型。在AAAI-2008关于复杂任务转移学习的讲习班4中，Mihalkova等人[51]将Tamar扩展到以单一实体为中心的转移学习设置，其中目标域中只有一个实体可用。Davis等人[52]提出了一种基于二阶马尔可夫逻辑形式的关系知识转移方法。该算法的基本思想是通过在目标域中用谓词实例化这些公式，以带有谓词变量的马尔可夫逻辑公式的形式发现源域中的结构规律。