深 圳 大 学

本 科 毕 业 论 文（设计）

题目：房价预测精度改进中的机器学习算法应用研究

姓名：姚余智

专业：工程管理

学院：土木工程

学号：2014090115

指导教师：宋博通

职称：副教授

2018 年 05 月 19 日

深圳大学本科毕业论文（设计）诚信声明

本人郑重声明：所呈交的毕业论文（设计），题目《房价预测精度改进中的机器学习算法应用研究》 是本人在指导教师的指导下，独立进行研究工作所取得的成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式注明。除此之外，本论文不包含任何其他个人或集体已经发表或撰写过的作品成果。本人完全意识到本声明的法律结果。

毕业论文（设计）作者签名：

日期：　　年　月　日

房价预测精度改进中的机器学习算法应用研究

深圳大学土木工程学院工程管理专业　　姚余智

【摘要】：

房价对于一个国家的经济发展和社会稳定有着重大影响,也一直是社会各方面所关注的热点与敏感问题。而无论是研究与房价有关的各个因素对房价的影响，还是预测未来的房价，建立一个精准的数学模型都是极为有效的研究方法。在过去，因为人们能获得的数据有限，房价预测模型的建立往往只能针对较少的数据来建立，而大数据时代下人们能获取的数据量级远超以往。

现如今，在针对海量数据的分析上，机器学习的使用已经非常普遍，并被证明是一种更为高效的建模方式。

本文针对房价预测模型精度的改进，应用机器学习算法，采用了杜鲁门州立大学收集的位于爱荷华洲埃姆斯市的房产数据集，使用机器学习算法进行建模并与传统线形回归进行对比分析。

对比结果表明，相对于传统的线形回归模型，机器学习模型的应用对于房价预测精度的改进明显。

关键词：房价预测；机器学习；

【Abstract】：

Housing price has great influence in the development of economic and is also important to the stability of our society. Thus the study of housing price is always hot in our society. Building a housing price prediction model is always a effective way to study about housing price. In the past, people can only access limited data due to the era. Nowadays, the data that people can access is much larger and more complex.

Machine learning algorithm has been proved to be a effective ways to deal with big data in many different area.

This article is focus on improve the accuracy of housing price prediction model by using machine learning algorithm. By using the data from Truman State University, this article compare the accuracy of machine learning model and traditional multiple linear regression model.

The study result shows that compared to the traditional multiple linear regression model the machine learning model has a great improvement in the accuracy.

Key words : Housing price perdition; Machine learning;

目录

1 绪论

1.1 背景及意义

1.1.1 房价研究的必要性与建模分析的优势

房产是大部分家庭的主要资产。它不仅仅是人们用来居住的场所，也是人们财富的一种储存媒介。房价对于个人置业，企业投资，国家的政策制度都有着重要的影响，因此研究房价是十分有意义的。而无论是研究影响房价的因素还是为个人投资，企业决策，政策制定做参考，建立模型进行预测分析都是很好的一种方法。与此同时，模型的精度提升则是许多学者在建立模型时所追求的目标。

1.1.2 大数据时代下应用机器学习的必要性

随着大数据时代的到来，学者们在研究时所能使用到的数据量相较以往，无论是在质量还是数量上都有极大的增加。

以房价预测为例，以往人们建立模型时涉及的模型输入量也许仅限房屋的面积，房屋的户型等等，这些特征只能片面的表达一个房屋的部分信息；而大数据时代下，人们更加重视数据的收集，现在可以通过更多的特征，例如空调的种类，地面材质，卧室数量等等来更全面的获取一个房屋的信息从而进行建模分析。在这种情况下，许多传统的模型，例如线形回归模型，在处理非线形变量以及大量数据时的表现往往较差。

对于房价预测模型来说，尽管我国目前针对房地产的数据库建立还不完善，但可以预见的是，随着信息时代的不断发展，房地产的数据也必将越来越多并全面。

在这种背景下，传统模型也许不再能很好的预测房价。而机器学习算法在其他领域面对海量数据以及非线性变量时的表现已经被证明远远超过传统模型。如 Google 公司一开始尝试通过行为序列等传统方式来建模分析数据中心的建筑能耗以求取最节能化配置。在利用传统模型仿真现场配置求取节能最大化目标后，Google 又引入了 Deepmind 团队，将机器学习的方法引入到了数据中心的节能模型建立中。机器学习算法引入后对机房制冷成本的下降达到了令人惊讶的 40%，整体能耗效率提高了 15%。[1]

本文探讨了利用机器学习算法建立了房价预测模型的可行性并与传统模型进行了对比，为提高房价预测模型精度提供了一种可行方法。

1.2 目的

通过用传统线形回归算法以及机器学习算法建立预测模型，对同一组数据进行房价预测。再通过对比不同模型的精度以说明机器学习算法在房价预测方面应用的前景。

1.3 内容

为达到对比模型精度以说明机器学习算法在房价预测中的应用，首先需要对传统线形回归算法以及机器学习算法进行学习，理解其原理后以及建模方法后进行建模，测算，对比。

在实际操作上需要学习各个算法的原理，并学习 R 语言中对应的算法包，进行数据处理，建模。本文主要使用 R 语言中的 Caret 包进行建模，测算。最后通过对比针对一组新数据进行房价预测时预测房价与实际房价的误差来对比模型精度。

1.4 研究方法

为达到说明机器学习算法在房价预测上的应用前景，本文通过对比传统线形模型与机器学习模型在房价预测上的精度优势来说明机器学习算法的在本领域的应用前景。

1.5 技术路线

本文为说明机器学习算法在房价预测这一领域上的应用前景，应用建模对比的方法，采用杜鲁门州立大学收集的位于爱荷华洲埃姆斯市的结构化房产数据集，进行模型精度对比分析，拟得出机器学习算法较传统线形回归模型在房价预测上的优越性。

1.6 创新点

国内目前关于机器学习算法在房价预测上的应用这一领域研究较少，并且没有研究将传统的线性回归模型与其进行对比。

另外本文在建模时应用了 Ensemble model（**集成模型**）的概念，将预测精度进行了进一步提升。

1.7 难点

每个不同的机器学习算法所需要的数据类型都不一样，因此在建立模型时需要针对不同的模型进行不同的数据处理。为使模型精度更高，避免过拟合等情况的产生，本文使用了 Ensemble model 的概念，将多个模型整合在一起，提升了精度。

1.8 数据和资料来源，依托文献

本文数据采用杜鲁门州立大学收集的位于爱荷华洲埃姆斯市的结构化房产数据集。参考了机器学习算法在其他领域的应用研究，如机器学习算法在预测短期建筑能耗上的应用2以及大数据背景下基于网络搜索数据的商品房价格预测3 等文章以及数据科学竞赛网站 Kaggle 上的建模思路。

1.9 拟得结论

本文拟通过应用不同算法对同一组数据建模，对比精度差异，得到机器学习算法模型相较传统线形回归算法模型在房价预测上的精度有提升这一结论。从而说明机器学习算法在建立房价预测模型上的可行性以及应用前景。

1.10 论文框架及各章内容概述

图 1 为本文建立的预测模型框架大纲。

第一步为数据的预处理。即将原始的数据集按合适的比例划分为训练集与测试集并将其结构调整为适合不同模型处理的结构。

第二部分为建立不同的预测模型并进行预测与结果对比。使用的包括传统的多重线形回归（MLR），以及机器学习算法模型中的极限梯度回归（XGBoost），随机森林（Random Forest），梯度提升机器（GBM），神经网络（Neural Network）等。其中在应用机器学习算法模型建模时采

用了集成模型（Ensemble model）的概念。具体原理及结构会在第五章建模过程中叙述。

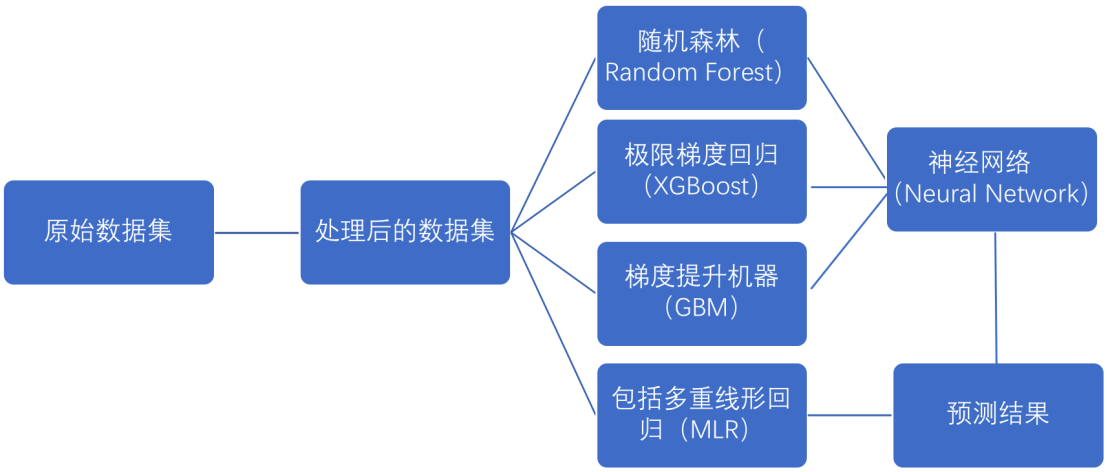

图 1

下文内容概述：

第二章为文献综述，在第二章中本文将会介绍国内外近年与本文研究内容类似的文章，并进行简要的总结与对比。

第三章为预备知识，为理解什么是机器学习以及本文使用的具体模型原理做了较为详细的介绍。

第四章为案例分析，对于本文建立的模型所使用的数据以及具体模型结构进行叙述。

第五章为建模过程，将建立的模型进行预测并对比其精度以说明机器学习算法在房价预测模型精度提升中的应用。

第六章为结果与对比，综述了模型的预测精度结果以及不同模型的对比结果。

第七章为总述与展望，对本文内容进行总结并对机器学习算法在房价预测中的潜在应用前景以及研究方向进行简要分析。

2 文献综述

基于不同算法的房价预测模型的研究成果已经有许多。从简单的多元线形回归模型的应用（仲小瑾 2008)4，灰色系统理论对线形回归模型的完善与改进5，非线性模型神经网络在房价评估领域的应用6，到支持向量机模型在房价预测7（郭志强 2013)，房价评估8中的应用（陈静 2008)。

针对房价预测模型的建立，国内外均已有许多研究。下表列出了近年来国内对于房价预测模型的研究中应用的模型。

年份	2014	2015	2016	2017
应用模型	多元线形回归，BP 神经网络（中国房价构成与预测的仿真分析- 陆丽丽，胡斌，李辉，端木怡婷 解放军理工大学指挥信息系统学院）	灰色-马尔可夫模型（ 基于两次改进的灰色-马尔可夫模型的太原房价预测 田红霞， 哈尔滨师范大学自然科学学报）	RS-SVM 模型（基于 RS-SVM 的商品住宅价格预测研究--以宁波市为例 陈绵旺 华东交通大学）	极度随机树；随机森林；GBDT；XGB；Stacking；（基于集成学习的房价预测模型 杨博文 曹布阳 同济大学软件学院）
	时间序列预测模型、灰色预测模型、BP 神经网络模型（房地产价格影响因素及预测研究 丁凤 安徽财经大学）	自回归分布滞后(ARDL)模型（基于领先指标和 ARDL 模型的城市住宅价格预测——以北京市为例 刘广友）	随机森林模型（基于随机森林模型的房价预测 陈世鹏金升平 武汉理工大学理学院 ）	多元线性回归（多元线性回归模型在房价走势分析与预测中的应用 钟丽燕高淑兰 百色学院数统学院）

表 1

2.1 国外研究近况

一如本文前文提到的，对于房价预测模型的研究从很多年前就已开始。国外方面，Jane（1997）9等把采用 Kalman 滤波的时序回归模型应用到了英国房地产价格预测模型中并通过对比证明其预测效果优与向量自回归模型以及误差修正模型；Hasa Selim（2008）10等通过对比价格法和效用估价法与人工神经网络对土耳其房价预测的精度来说明人工神经网络的优越性。Byeonghwa Park11等把 RIPPER 算法（机器学习算法的一种）应用到了房价预测模型中并发现该算法相较其他算法在房价预测上有一定的优越性。

在数据科学方面的研究新成果是日新月异的，许多新的模型，理论还未登上正式的期刊就以被人们应用到实际应用中。如本文使用的 Xgboost 模型，就是由华盛顿大学的陈天奇[12]与 Carlos Guestrin 所研发但尚未发表在正式期刊上。相关论文目前仅能在 arXiv 上查阅。

与此类似的是，由于数据科学领域的开源性，人们总是能很容易的复制，应用，改进最前沿的模型算法，因此在模型的研究与应用上也往往是研究滞后与应用。例如本文建模上所参考的数据科学竞赛网站 kaggle 上有着许多新颖的建模方法都没有被人们具体的研究分析过。

2.2 国内研究近况

国内方面，近年来对于房价预测模型的研究主要是围绕着对旧模型的改进（如田红霞[13]提出了灰色-马尔可夫模型的改进）；新模型的应用（如陈世鹏[14]等提出了随机森林模型的应用）；亦或

是通过模型研究不同变量间的因果关系（如罗婧和朱建峰[15]等通过利用空间面板的弹性分析，引入了数据挖掘技术，研究对比了数据挖掘出的与房价有关因素的权重大小。）

同济大学的杨博文与曹布阳[16]等人采用了与本文类似的建模思路，主要介绍了引入了Ensemble model（集成模型）后模型精度的提高。但本文使用的具体模型以及模型整合方法都与其有所不同。

2.3 小结

总的来说，我们可以看到，所有的这些研究都离不开模型的精度提升以及新算法的应用。

近年来，机器学习的概念随着信息时代下海量的数据再次变得炙手可热。许多领域在合适的数据支持下纷纷通过应用机器学习算法建立模型分析（如范成等将机器学习算法应用到建筑能耗的预测中）并有着很好的成效。

本文希望通过将机器学习算法应用到房价预测模型中并与传统的线形回归模型做对比来说明机器学习算法在房价预测这一领域应用的可行性及潜力。

3 预备知识

由于本文涉及的领域较多，笔者仅在此章就自己学习以及理解的部分将本文涉及到的一些机器学习方面的知识进行介绍以供读者更好的理解本文。

因为关于机器学习的许多定义学界并无一个很严谨的定论，本章的撰写参考了业界领头人物Amazon AI 主任科学家李沐对于机器学习的许多理解与定义。

3.1 机器学习的定义

机器学习可以说是近年来各个领域最为火热的概念。当我们拿出手机，不方便的打字的时候用语音识别功能代替输入法，手机将我们说的话识别，并精准的转换成相应的文字——无论你的普通话是否标准。这过程当中就应用了机器学习模型。

3.1.1 机器学习与传统编程的异同

如果从来没有接触过机器学习这个概念，也许会把传统的编程和它混淆。首先，人们的确需要使用编程语言来实现机器学习模型，但是不是每个程序都涉及到机器学习。

举个例子，在我们进行选课的时候，我们把一个课程加入到我们自己的选课列表中，这是对于计算机来说我们只是把这个课程的 id 和我们的学号 id 插入到选课系统数据库的表格中。对于这种简单直接的程序来说，我们不需要使用到机器学习。

但是，很多时候我们想要计算机做到的事情没有那么简单。以上文提到的语音输入法来做例子，我们的输入不再是单一，固定的输入，（例如说课程 id 与学号 id）而是麦克风采集到的原始语音信号。而现在我们问题就变成了怎么样让机器去把语音信号转变成对应的汉字。

不同于机器，我们人类自己可以识别出一段语音信号里包含的文字。由此，我们可以收集一个巨大的数据集，里面包含着大量的语音信号以及每个信号对应的汉字。机器学习的解决方式就是通过这个庞大的数据集，写一个灵活的程序并有着大量的参数。通过调整这些参数，我们能改变程序的输出结果。这种程序就是机器学习模型。

3.1.2 机器学习是什么

总的来说，我们的模型就像一个机器，通过某种方式将输入转换为输出。在语音输入法的例子中，模型的输入是一段语音，输出则是这段语音对应的文字。

如果模型正确，则其必有一组参数设定，能准确的转换出相对应的文字。而机器学习中的学习一词就是指在模型的训练过程中不断更新模型的行为。（通过调整参数的方式）。

对于本文来说，模型的输入就是上万条的房产信息，输出则是他们的价格预测。

3.2 机器学习的几大要素

拿语音输入法举例，我们利用语音音频文件作为自变量以及其对应的文字作为标志放在一起组成的数据集建立一个输入为语音输出为文字的模型并加以训练。这种方式称为监督学习，而机器学习中除了监督学习以外还有许多其他方法。不过本文运用的是监督学习的方式，因此也将着重讲述这一方法。其他方法本文仅做简略介绍。

一个成功的机器学习有四大要素：数据，转换数据的模型，衡量模型好坏的损失函数以及调整模型权重来最小化损失函数的算法 。

3.2.1 数据

数据是这几年机器学习重新变得火热的核心原因。大数据时代下海量的数据给予了机器学习海量的样本去自我学习从而能大大提高模型的精度。

机器学习能应用到的数据可以是图片，文本，声音，影像或者是结构化的数据（例如本文使用的数据，或者是电子病历，电费账单等。）

3.2.2 模型

一般来说，我们所拥有的数据和我们最终想要得到的结果有较大的差别。例如在本文中，我们想要知道一个房屋具体价格，而我们只有这个房屋的一些数据。（例如面积，楼层数，地板类型，地理位置等等。）因此我们需要一个模型，能将这些低级特征转换为价格进行输出。

3.2.3 损失函数

在建立模型时，我们需要对比模型的输出和真实值之间的误差。而损失函数的作用就是用了衡量输出结果对比真实数据的好坏。例如模型预测一个房屋的价格是 50 万美金，而它的实际价格是 90 万美金，这个时候就需要一个标准来告诉我们模型的精准性到底如何了。（即损失函数的结果如何。）

我们建立的预测模型有着许多的参数，而我们通过最小化损失函数的方式来“学习”（改进）这些参数。在“学习”这些参数时，有两项数据是我们需要跟踪的。

一，训练误差：模型在用于训练的数据集上的误差。

二，测试误差：模型在其没见过的新数据上的误差。（因为过拟合以及其他的一些原因，测试误差结果可能会与训练误差差别很大。）

3.2.4 优化算法

最后，我们需要算法来通盘考虑模型本身和损失函数，对参数进行搜索，从而逐渐最小化损失。最常见的神经网络优化使用梯度下降法作为优化算法。简单地说，轻微地改动参数，观察训练集的损失将如何移动，然后将参数向减小损失的方向调整。

3.3 监督学习

监督学习描述的任务是，当给定输入 x，如何通过在有标注输入和输出的数据上训练模型而能够预测输出 y。从统计角度来说，监督学习主要关注如何估计条件概率 P(y|x)。

监督学习仅仅是机器学习的方法之一，在实际情景中却最为常用。部分原因是许多重要任务都可以描述为给定数据，预测结果。例如，给定一位患者的 CT 图像，预测该患者是否得癌症；给定英文句子，预测它的正确中文翻译；给定本月公司财报数据，预测下个月该公司股票价格；还有本文中的，给定房屋信息，预测其价格。

“根据输入预测结果”，看上去简单，监督学习的形式却十分多样，其模型的选择也至关重要，数据类型、大小、输入和输出的体量都会产生影响。例如，针对序列型数据（文本字符串或时间序列数据）和固定长度的矢量表达，这两种不同的输入，会采用不同的模型。

简单概括，监督学习的学习过程看起来是这样的：在一大组数据中随机地选择样本输入，并获得其真实（ground-truth）的标注（label）；这些输入和标注（即期望的结果）构成了训练集（training set）。我们把训练集放入一个监督学习算法（supervised learning algorithm）。算法的输入是训练集，输出则是学得模型（learned model）。基于这个学得模型，我们输入之前未见过的测试数据，并预测相应的标注。

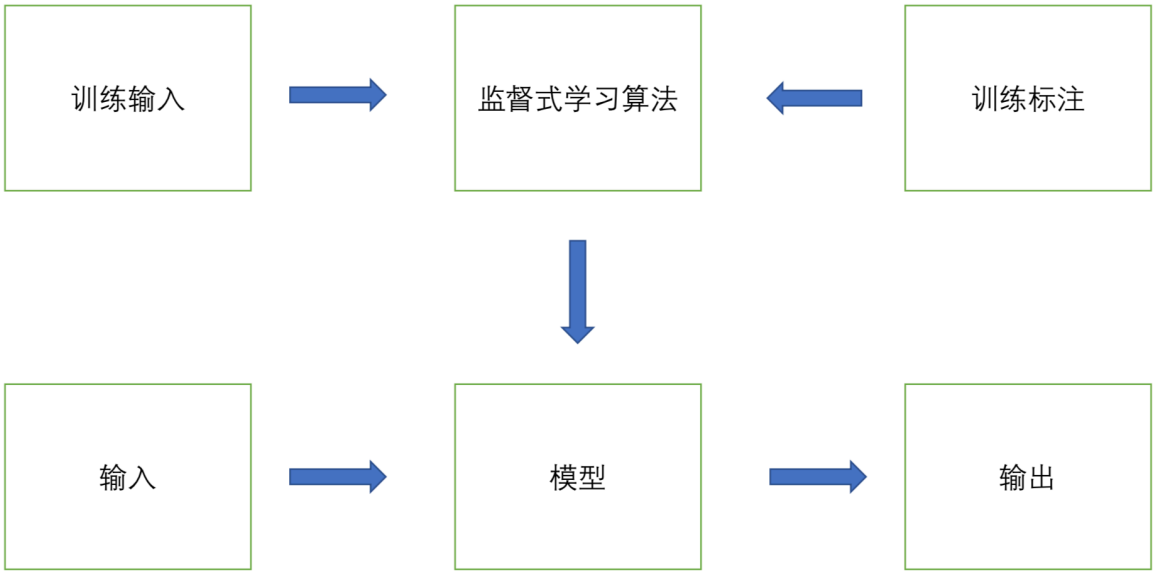

图 2

3.3.1 回归分析（Regression）

回归分析也许是监督学习里最简单的一类任务。在该项任务里，输入是任意离散或连续的、单一或多个的变量，而输出是连续的数值。例如我们可以把本月公司财报数据抽取出若干特征，如营收总额、支出总额以及是否有负面报道，利用回归分析预测下个月该公司股票价格。

一个更详细的例子，一个房屋的销售的数据集。构建一张表：每一行对应一幢房子；每一列对应一个属性，譬如面积、卧室数量、卫生间数量、与市中心的距离；我们将这个数据集里这样的一行，称作一个特征向量（feature vector），它所代表的对象（比如一幢房子），称作样本（example）。

如果住在深圳或广州这种大城市，那么这个特征向量（面积、卧室数量、卫生间数量、与市中心的距离）可能是这样的[100,0,.5,60]。如果住在喀什，则可能是这样的[3000,4,3,10]。这些特征向量，是所有经典机器学习问题的关键。我们一般将一个特征向量标为 Xi，将所有特征向量的集合标为 X。

一个问题是否应采用回归分析，取决于它的输出。比如你想预测一幢房子的合理的市场价格，并提供了类似的特征向量。它的目标市场价是一个实数。我们将单个目标（对应每一个特征向量代表的样例 xi 标为 yi，所有目标的集合为 y（对应所有样例的集合 X）。当我们的目标是某个范围内的任意实数值时，这就是一个回归分析问题。模型的目标就是输出预测（在这个例子中，即价格的预测），且尽可能近似实际的目标值。我们将这些预测标为 Y-hat。

如果我们把模型预测的输出值和真实的输出值之间的差别定义为残差，常见的回归分析的损失函数包括训练数据的残差的平方和或者绝对值的和。机器学习的任务是找到一组模型参数使得损失函数最小化。

3.3.2 监督学习的一些其他应用

由于本文面对的是回归问题，故监督学习中的一些其他类型的问题本文在此仅作部分的，简略的介绍。

1，分类

回归分析能解答“多少？”的问题，但也有很多问题不能套用这个模版。例如给定一张动物图片，我们想知道它是什么类型的动物，这种时候就是分类问题。

2, 标注

标注任务的目标是，给定一个输入，输出不定量的类别。例如给定一张图片，模型能帮我们标注出来这张图片里是否有狗，是否有猫，是否有草等。

3，搜索与排序

最简单的一个例子，当我们在使用搜索引擎时，判断一个网页与我们搜索关键字的相关度问题。

4，推荐系统

如何向一个经常听周杰伦的人推荐他可能喜欢的音乐？如何在知道对方曾购买过三体等科幻小说的情况下准确向其推荐其可能感兴趣的书籍？这一类问题也属于监督学习的范畴之内。

3.4 非监督学习

迄今为止的例子都与监督学习有关，即我们为模型提供了一系列样例和一系列相应的目标值。你可以把监督学习看成一个非常专业的工作，有一个严厉的监督者站在你的身后，告诉你每一种情况下要做什么，直到学会所有情形下应采取的行动。

而相反的情形，在一个没有相应目标值的情况下进行机器学习时，我们往往称这类问题为无监督学习（unsupervised learning）。此类问题非常之多，在这仅举部分例子以供读者理解。

1，我们能用少量的原型，精准地概括数据吗？给我们一堆照片，能把它们分成风景、狗、婴儿、猫、山峰的照片吗？类似的，给定一堆用户浏览记录，我们能把他们分成几类典型的用户吗？这类问题通常被称为聚类（clustering）。

2，我们可以用少量的参数，准确地捕获数据的相关属性吗？球的轨迹可以很好地用速度，直径和质量准确描述。裁缝们也有一组参数，可以准确地描述客户的身材，以及适合的版型。这类问题被称为子空间估计（subspace estimation） 问题。如果决定因素是线性关系的，则称为主成分分析（principal component analysis）。

3，在欧几里德空间（例如，ℝn 中的向量空间）中是否存在一种符号属性，可以表示出（任意构建的）原始对象？这被称为表征学习（representation learning）。例如我们希望找到城市的向量表示，从而可以进行这样的向量运算：罗马 - 意大利 + 法国 = 巴黎。

4，针对我们观察到的大量数据，是否存在一个根本性的描述？例如，如果我们有关于房价、污染、犯罪、地理位置、教育、工资等等的统计数据的，我们能否基于已有的经验数据，找出这些因素互相间的关联？贝叶斯图模型可用于类似的问题。

3.5 模型原理介绍

3.5.1 梯度提升机器模型（Gradient Boosting machine GBM)

梯度提升是一种经常用来解决回归与分类问题的技术。最早由 Leo Breiman17提出，其中“提升”一词可以理解为寻找损失函数的最优算法。

为理解 GBM 模型的原理，首先要介绍机器学习中很重要的一个概念，梯度提升（又称梯度下降）。

假设我们的模型 F 可以用下面的函数表示，P 表示参数，F(x,P)表示以 P 为参数的函数集。我们的预测模型可以理解为由 M 个参数为 a 的模型叠加而成， β 表示每个模型的权重。

$$F(x,P) = F(x;\{\beta_{m,}\alpha_m\}_1^M) = \sum_{m=1}^{M}\beta_m h(x;\alpha_m)$$

对于这个函数，我们想要求得它的最优参数 P*，则需要求它的损失函数 Φ(P)的最小值。

$$P^* = \sum_{m=0}^{M} p_m$$

$$P^* = \arg\min(\phi(P)) \quad \Phi(P) = E_{y,x}L(y,F(x;P))$$

P^* 可以由组成 F 的 M 个模型的参数集合 p_m 表示。在我们优化 P 的时候，假设当前已经得到了 m-1 个模型，想要得到第 m 个模型时，我们对前 m-1 个模型求梯度，得到它最快的下降方向 gm。

$$g_m = \{g_{jm}\} = \{[\frac{\partial\Phi(P)}{\partial p_j}]_{p=p_{m-1}}\}$$

由此我们得到新模型的参数 Pm 就是它在这个梯度方向上下降的距离。

$$p_m = -\rho_m g_m$$

梯度下降就是通过这种模型的不断迭代，以达到优化模型参数，得到参数最优解的目的。

在本文中应用的 Gradient boosting machine 实际上是在传统的树模型上应用了 Gradient boosting 的方法来进行训练。

对于同一个训练数据集，4 个人，A,B,C,D，他们的年龄分别是 14,16,24,26。其中 A、B 分别是高一和高三学生；C,D 分别是应届毕业生和工作两年的员工。他们有着许多特征，例如购物金额，上网时间等等。如果是用一棵传统的回归决策树来训练，会得到如下图 3 所示结果：

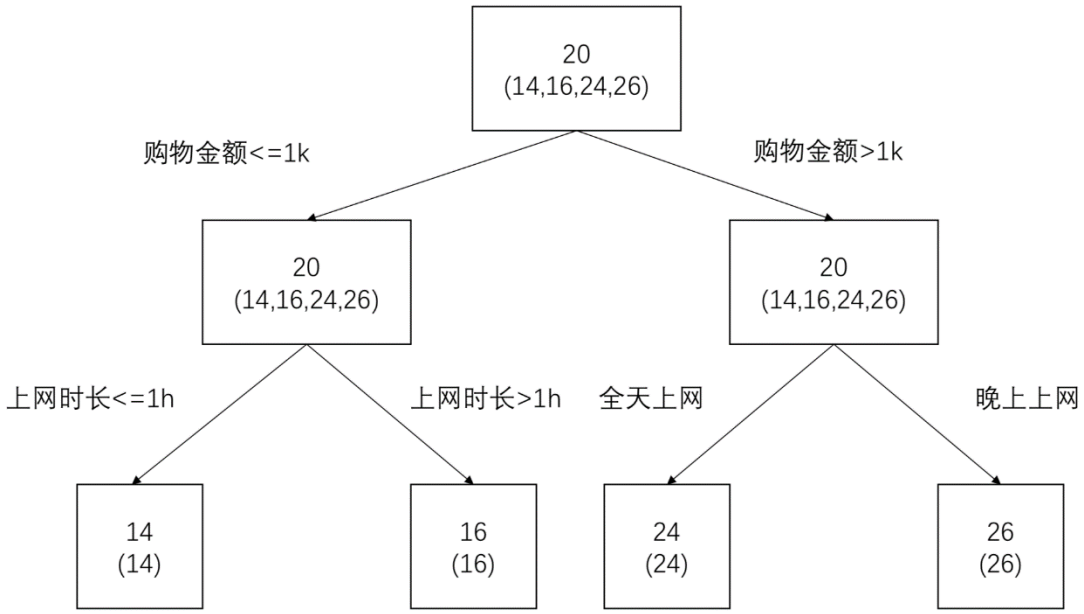

图 3

而 GBM模型下建立的树模型示意图如下：

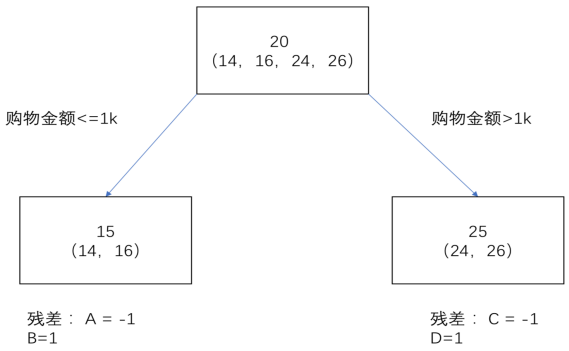

图 4

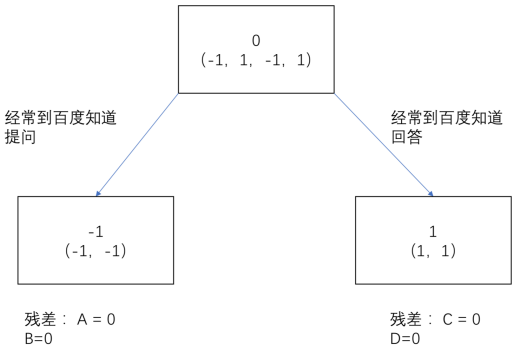

图 5

在 GBM 模型下，最后得到的结果由原来的一颗树变成了多棵树的累加。每棵树的初始预测输入都是上一棵树预测结果的残差。

GBDT 的求解算法，具体到每颗树来说，其实就是不断地寻找分割点(split point)，将样本集进行分割，初始情况下，所有样本都处于一个结点（即根结点），随着树的分裂过程的展开，

样本会 分配到分裂开的子结点上。分割点的选择通过枚举训练样本集上的特征值来完成，分割点的选择依据则是减少 Loss。

给定一组样本，实际上存在指数规模的分割方式，所以这是一个 NP-Hard 的问题，实际的求解算法也没有办法在多项式时间内完成求解，而是采用一种基于贪心（greedy）原则的启发式方法来完成求解。也就是说，在选取分割点的时候，只考虑当前树结构到下一步树结构的 loss 变化的最优值，不考虑树分裂的多个步骤之间的最优值，这是典型的 greedy 的策略，也是ｇｒａｄｉｅｎｔ的理论依据。18

GBM 模型相对于传统的回归树模型的优点是可以避免模型的过拟合现象。例如在第一个传统模型里，为了达到训练集 100%精度模型使用了 3 个特征（上网时长，时段，网购金额）而 GBM 模型则只使用了 2 个特征。这在实际使用中可以很好的避免过拟合的问题。

3.5.3 极限梯度提升模型（XGBoost）

XGBoost 是 GBM 模型的一种高效系统实现方法。它和传统的 GBM 模型相比有着以下几点优点。

1,加入了正则化项

正则化方法是数学中用来解决不适定问题的一种方法，后来被引入机器学习领域。通俗的讲，正则化是为了限制模型的复杂度的。模型越复杂，就越有可能“记住”训练数据，导致训练误差达到很低，而测试误差却很高，也就是发生了“过拟合”。在机器学习领域，正则化项大多以惩罚函数的形式存在于目标函数中，也就是在训练时，不仅只顾最小化误差，同时模型复杂度也不能太高。

在决策树中，模型复杂度体现在树的深度上。XGBoost 使用了一种替代指标，即叶子节点的个数。此外，与许多其他机器学习模型一样，XGBoost 也加入了 L2 正则项，来平滑各叶子节点的预测值。

这可以进一步减少模型过拟合的情况发生。

2，支持列抽样

列抽样是指，训练每棵树时，不是使用所有特征，而是从中抽取一部分来训练这棵树。这种方法原本是用在随机森林中的，经过试验，使用在 GBDT 中同样有助于效果的提升。

3，支持并行运算

对于结构化数据的处理上 xgboost 有许多改进，使得它的运算效率较传统 gbm 模型提升了许多。

3.5.4 随机森林模型(Random forest)

随机森林模型也是回归决策树模型的一种。之所以叫随机森林是因为他有两个方面和传统的决策树模型以及 GBM 不同。

第一，在建立决策树时，与 GBM 不同以及传统决策树模型不同，随机森林也有多个决策树但随机森林采取的是平行建树的模式。即每一个单独的树模型都与传统的树模型类似，最后再将每个单独的树模型得出的结果进行总结得出结果。（例如采取算术平均的形式。）

第二，随机森林在建立每个树模型时选取的数据的方式不同于传统决策树以及 GBM。这主要体现在训练样本的选取以及特征的选取上。

在建立随机森林模型时，每个独立的树模型所采取的训练样本数据都只是整体训练数据集的随机一部分。并且在训练模型时所采取的特征变量也是随机选取的。

这种随机的特性带来的主要优点有[19]：

1，可以较好的避免过拟合现象。

2，模型训练的过程会较为快速。

3，对于不平衡的数据集，随机森林可以有一定程度的修正。

4，即使是较小的训练数据集也能得到较高的准确度。

3.5.5 神经网络模型（Neural Network ）

本文采取的神经网络为单层的前馈神经网络，其原理如图 6：

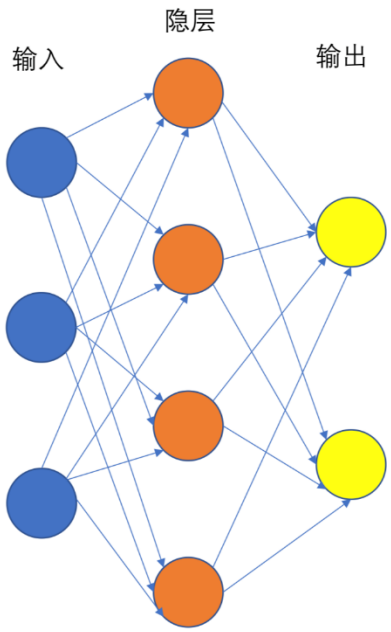

图 6

它包括输入层（input layer）、输出层（output layer）和一个隐藏层（hidden layers）。上图的神经网络由 3 个单元的输入层，4 个单元的隐藏层和 2 个单元的输出层组成。每个单元为一个感知器。输入层的单元是隐藏层单元的输入，隐藏层单元的输出是输出层单元的输入。两个感知器之间的连接有一个权重。第 t 层的每个感知器与第 t-1 层的每个感知器相互关联。

在加工输入数据时，将输入数据赋予输入层的每个单元，而隐藏层的每个单元是输入层每个单元的加权求和。也就是说，输入层的数据会被前向传播到隐藏层的每个单元。同理，隐藏层的输出作为输入会前向传播到输出层，计算得到最后的输出，即神经网络的输出。

而为了使模型更加精确，每个感知器之间的权重就是神经网络的重要参数。在机器学习中一般通过梯度下降的方式进行神经网络参数的最优化。

在本文中，借由集成模型的方法，将 GBM，RF，XGBoost 等模型的预测结果作为模型的输入，训练建立的神经网络模型得到的预测结果证实比所有三个模型都好。

3.5.6 深度学习（Deep Learning）

根据 LeCun 的定义，深度学习是一种允许模型通过含多隐层的多感知器对数据进行高阶表征学习的方法。[20]

以深度学习的神经网络为例，在深度学习中，隐层往往不仅仅只有一个，而且每个隐层的处理逻辑也个不相同。例如有的隐层进行线性的处理，有的隐层进行复杂处理等。这也是深入学习式的神经网络模型和本文所使用的单层神经网络模型最大的不同。训练深度学习模型的方式也与训练一般模型的方式有所不同。例如说在训练中会随机去掉一些隐层以确保模型的鲁棒性。

深入学习可以以监督模式运用来建立一个 Deep neural net（深度神经网络）模型，也可以用非监督的模式来建立一个自动提取特征的模型。在传统的监督式机器学习中，对初始数据的特征工程提取是非常重要的。[6]而非监督的模式建立的特征提取模型已经被广泛证明是行之有效的。[21]

通过把模型的输出变成输入并反复重复以得到最小的模型误差，这种非监督式的方法在许多领域已经被证明可以获得很好的结果。

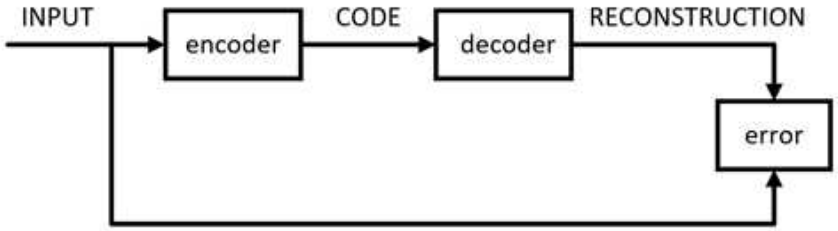

图 7

在不断的训练过程中，模型的隐层不断被随机选择，以得到一个最好的结果。非监督式的深度学习方法对专业知识的要求较少并很有可能能提取出最具有信息量的特征。

4，案例分析

4.1 数据描述

本次研究的数据来自采用杜鲁门州立大学收集的位于爱荷华洲埃姆斯市的结构化房产数据集，数据集中有着美国 Ames，lowa 地区的 1460 条住宅信息。每条房屋信息都有着包括面积，车库，是否有泳池，空调系统等等的 79 个解释变量（部分房屋部分信息有缺失）以确保数据能尽可能精确的描述房屋的情况。

ID	建筑等级	地块性质	距离街道的距离	地块面积	连接的道路类型	与小巷联通类型	房屋类型	地块平整度	所有公共设施类型
1	60	RL	65	8450	Pave	NA	Reg	Lvl	AllPub
2	20	RL	80	9600	Pave	NA	Reg	Lvl	AllPub
3	60	RL	68	11250	Pave	NA	IR1	Lvl	AllPub

表 2

总的来说，数据集有着 115,340 个变量，表 2 为数据集的一部分。表 3 为一个房屋在数据集中所有的特征。

编号	建筑等级	地块性质	距离街道的距离	地块面积	连接的道路类型	与小巷联通类型	房屋类型	地块平整度	所有公共设施类型
地块结构	房屋坡度	房屋的街道	是否临近主干道	是否临近铁路	住宅类型	住宅风格	总体材料与质量	房屋情况评价	翻新日期
屋顶材料	屋面材料	房屋外部材料	砌体类型	地下室高度	地下室条件	暖气类型	一楼面积	二楼面积	中央空调
电器系统	加热质量	建筑面积	完整的浴室	壁炉数量	车库位置	车库年份	车库装修	车库容纳量	车库质量
车库条件	阳台面积	泳池面积	泳池类型	销售月份	销售年份	销售类型	销售情况	厨房数量	卧室数量
厨房品质	房间总数	栏栅	杂项价值	门廊面积	车道类型	泳池质量	原施工日期	地下室墙壁	为完工的地下室面积

半浴室树木	杂项	地面以上居住面积	基础类型	砌体类型	外部材料质量				

表 3

4.2 做为模型输入的数据集

因为本次建模旨在对比机器学习算法对房价预测模型的精度提升，在最终整体模型建立过程中，Random forest（随机森林），Gradient boosting（梯度提升），XGBoost（极限梯度提升）等机器学习算法模型的训练数据集与 MLR（多重线性回归）一样，为未经过高维处理的数据集。

由于本文采用了 Ensemble model 的概念，在模型的第二部分（Level 1 层），Neural net（神经网络）中则是使用第一部分 Random forest（随机森林），Gradient boosting（梯度提升），XGBoost（极限梯度提升）的预测结果做为模型第二层的输入值，进行建模。具体模型结构可见 5.2。

5，建模过程

5.1 预测模型的选取

本文一共选取了包括 MLR（多重线性回归），Random forest（随机森林），Gradient boosting（梯度提升），XGBoost（极限梯度提升），neural net（神经网络）来完成建模并进行比较。

在本文中，多重线性回归将做为传统模型的代表，来证明机器学习在房价预测上的优越性。一般来说，通过多重线性回归建立的模型运算效率都很高并且很容易解读。但是和其他方法比起来，由于无法对非线性变量进行解析，在预测时精度应当会低于其他几种模型。

本文使用的其他模型能很好的处理复杂，非线性的输入变量。Random forest（随机森林），Gradient boosting（梯度提升），XGBoost（极限梯度提升）以及 neural net（神经网络）都是在其他领域被证明非常有效的机器学习中的建模方法。其中 XGBoost 是 Gradient boosting 的升级版。它在提高效率以及处理多重共线性上有着显著的改进。

MLR（多重线性回归）主要方法为普通最小二乘法以及广义最小二乘法。本文使用的 R 语言中 stats package 进行建模。运用的是最小二乘法，是一种非常传统但有效的线性回归方法。

Random forest（随机森林），Gradient boosting（梯度提升），XGBoost（极限梯度提升）等机器学习算法模型都是决策树模型，在进行调参的时候主要有两个参数：决策树的数目以及决策树的深度。一般来说，预测的精度以及模型的计算负荷会随着决策树数目的上升而上升。而每个决策树分裂的变量数目选取又是另一个需要调整的参数。根据 Liaw 的研究22，这个取值一般取总变量的 1/3 之一为佳。

对于 neural net（神经网络）来说，参数有隐层数以及节点数。本文采取三层网络，即一层隐层的形式。而隐层节点数的选择则对神经网络模型的性能影响很大，并且是模型是否出现“过拟合”的直接原因。

在具体实现上本文使用 R 语言中的 Caret package 调用相应函数进行建模。

5.2 Ensemble model （集成模型）的应用

除此了使用了机器学习算法之外，本文还通过 Ensemble Modeling（集成模型)的方法，结合 neural net（神经网络）进一步提高了房价预测模型的精度。

本文将 GBM（梯度提升机器模型），XGBoost（极限梯度提升模型），Random forest（随机森林模型）等模型的预测输出结果作为 neural net（神经网络模型）的输入来建立模型。即 Level-1 的训练集为 level-0 所有预测结果与原始样本真实值组成的新的数据集，在此数据集的基础上，利用 neural net（神经网络）再次得到最终的集成预测模型。具体结构见图 8。

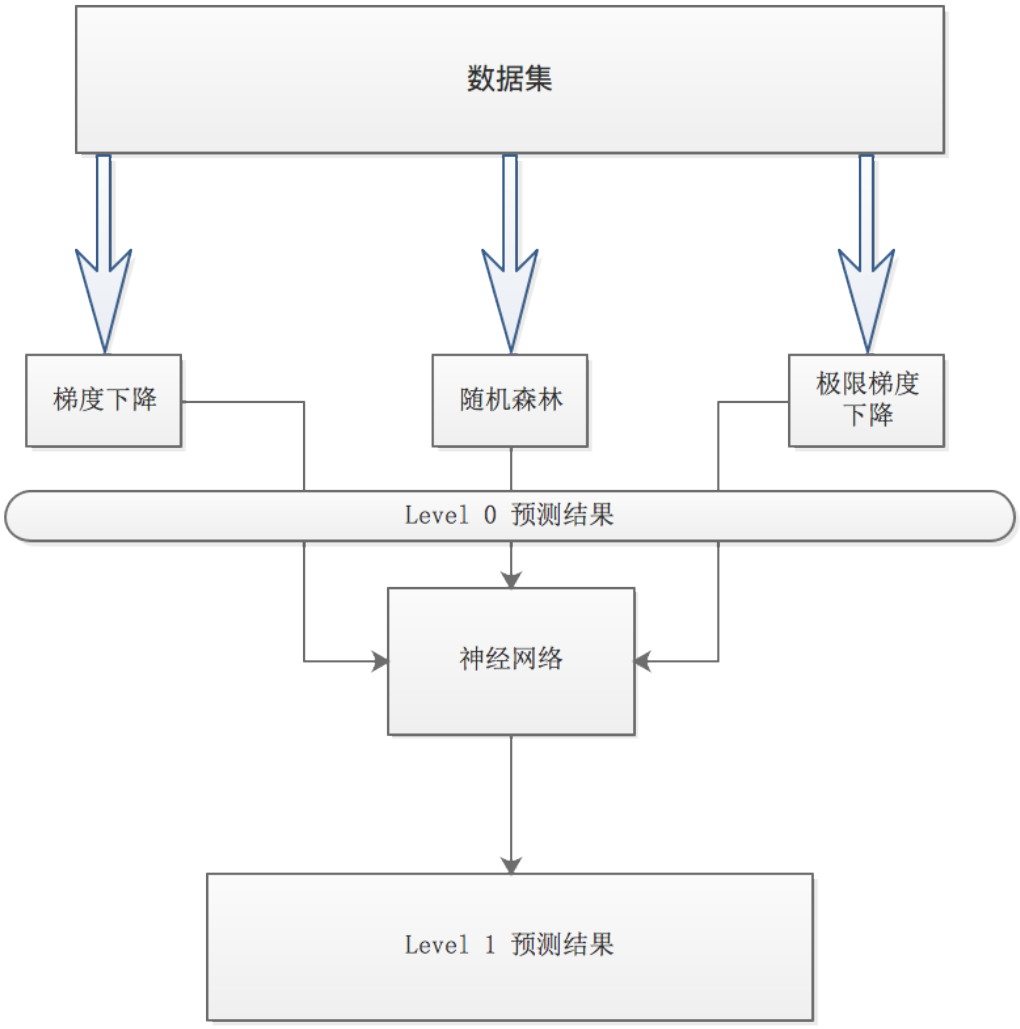

图 8

最后的模型预测精度对比结果也说明了集成模型对于精度改善的效果。

6，结果与对比

建模结束后，利用测试集数据，将一组新的房屋数据作为输入变量，得到预测值并与这些房屋的实际成交价格做对比得出相应模型的精度。

五种模型的预测结果与实际结果进行比较。具体表现如下图。

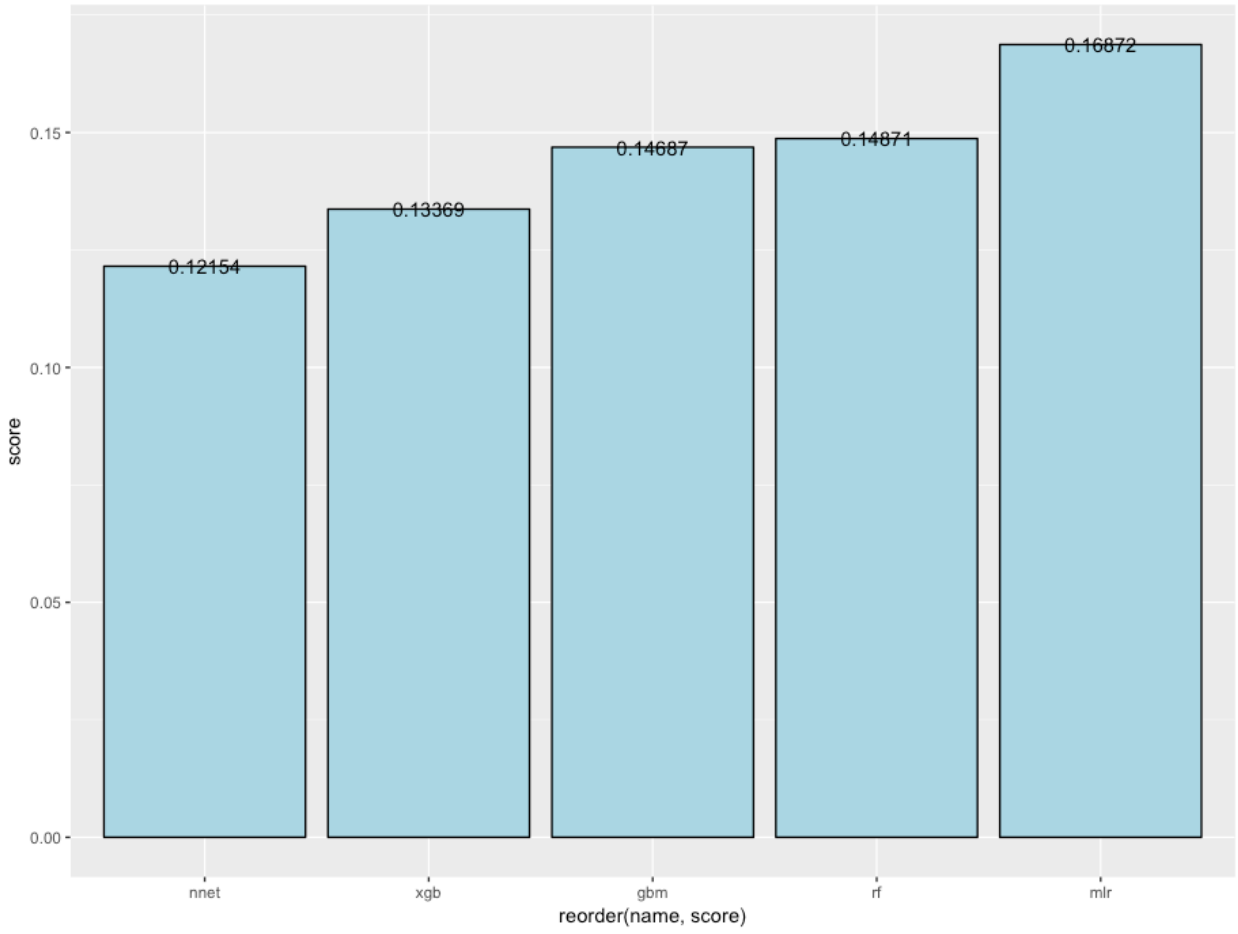

图 9

图中得分为预测结果与房屋实际成交价格的 root mean square logarithmic error(均方根对数误差)。数值越小说明预测精度越高。

从上图可以直观的看到，基于机器学习算法建立的模型精度明显高于传统多元线性回归建立的模型。且在 Neural net（神经网络）算法下建立的二层模型显著提高了模型的精度，将模型误差从 0.13369 下降到了到了 0.12154。

这一对比结果支持了本文的研究目的，很好地说明了机器学习算法在房价预测精度改进中的应用前景与可行性。

7，结论与展望

7.1 结论

本文通过对比建模，研究了机器学习算法在房价预测中的潜在可能性以及应用前景。

通过模型预测精度对比结果，可以直观看到的是机器学习算法在针对本次研究数据集时，相较传统多元线性回归算法表现出了其优越性。这说明了机器学习算法在房价预测精度改进中应用的可行性与有效性。

7.2 相关展望与思考

7.2.1 深度学习的应用

本文通过 Ensemble model（集成模型）的手段，再次提升了模型的精度。这过程中对于 level 0 层的预测结果，基于 neutral net（神经网络）的再建模展现了将 deep learning（深度学习）应用在房价预测模型上的可能性。从 level 0 到 level 1，模型的精度提高了 9.08%。而这只是基于单层神经网络模型的优化结果，如果在数据足够多的情况下，将隐层添加到 3 以上，让神经网络模型变为深度神经网络模型，给模型足够的数据去自我学习，提升，精度是否会继续上升呢？

7.2.2 大数据时代对于模型建立带来的影响

在本次研究过程中，非线性模型的全面好于线性模型。这说明了数据集中有着许多变量是线性模型难以处理的。笔者认为而随着信息社会的不断发展，数据的类型，数量都会不断增多。届时如何从海量的数据里提取出有意义的特征，如何处理那些新增加的非线性数据，如何在不损失模型精度的情况下提高模型效率将会是接下来房价预测模型建立的主要研究方向。

7.2.3 模型的可信度

笔者在研究如何在房价预测领域中应用机器学习算法之余也尝试学习，研究了机器学习算法的可解释性。因为在研究，应用时并不是越精确的模型越值得人们信任。根据 Marco Tulio[23]等人的研究，人们会更偏向与信任并使用一个自己能理解的模型。如果一个模型的预测结果是基于人们完全无法理解的原理时，人们往往难以信任并运用这个模型。例如在本文的房价预测模型中，如果模型告诉人们某一个房屋的房价的预测结果大部分是基于这个房屋的浴室类型时，即使这个模型的整体精度很高人们也不会去信任并使用这个预测结果。可能的一个原因是，该模型在该个案上的预测结果出了问题。

基于此，通过对比不同模型整体自变量的权重以及针对个案的自变量权重也许可以衡量该模型在对该个案的预测可信与否。

本文利用 Lime 算法针对个案进行了自变量权重的分析，发现每个个案对最终预测结果影响最大的几个自变量都是不同的。例如在有些案例里，对房价影响最大的自变量是建筑等级，有些却是建筑面积。而对于模型整体来说，权重最大的自变量却是整体质量。那么这种差异背后是否也存在着某种联系呢？笔者认为也许可以通过这种方式来进一步提升模型精度以及对模型的理解。[24]

8，参考文献

[1] DeepMind AI Reduces Google Data Center Cooling Bill by 40%. https://deepmind.com/blog/deepmind-ai-reduces-google-data-centre-cooling-bill-40/

[2] Cheng Fan [a], Fu Xiao [b,⇑], Yang Zhao [c], A short-term building cooling load prediction method using deep learning algorithms

[3]张令令 孙金金 黄世祥 大数据背景下基于网络搜索数据的商品房价格预测——以武汉市为例，中南财经政法大学 2015

[4] 仲小瑾. 基于多元线性回归分析法的房地产价格评估[J]. 商业时代，2014：133-134.

[5]李菲，孙文彬. 灰色理论在商品住宅价格预测中的应用[J]. 辽宁工程大学学报，2004，6(3)：271-273.

[6]张辉. 关于多当今社会 BP 神经网络的房地产价格评估与研究方向[J]. 房地产导刊，2013.

[7] 郭志强. 基于支持向量机回归的房地产批量估价[D]：[硕士学位论文]. 广州：暨南大学，2013.

[8] 陈静. 基于支持向量机的房地产估价方法研究[D]：[硕士学位论文]. 西安：长安大学，2008.

[9] Jane P, Browna, Haiyan Song & Alan Me Gillivray. Forecasting UK House Prices: A Time Varying Coefficient Approach[J]. Econoinic Modelling, 1997 (04).

[10] Hasa Selim. Determinants of House Prices in Turkey: Hedonic Regression Versus Artificial Neural Network [J]. Dogus University Journal, 2008(01).

[11] Byeonghwa Park, Jae Kwon Bae, Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data,Expert Systems with Applications,Volume 42, Issue 6,2015

[12] Tianqi Chen ,Carlos Guestrin University of Washington XGBoost: A Scalable Tree Boosting System

[13] 田红霞.基于两次改进灰色加权马尔可夫模型的 CPI 预测[J].中北大学学报(自然科学版),2015,36(02):113-117.

[14] 陈世鹏,金升平.基于随机森林模型的房价预测[J].科技创新与应用,2016(04):52.

[15] 罗婧，朱建锋.基于数据挖掘与空间计量的房地产价格经验分析[J].开发研究，2013(05).

[16] 杨博文,曹布阳.基于集成学习的房价预测模型[J].电脑知识与技术,2017,13(29):191-194.

[17] ARCING THE EDGE Leo Breiman Technical Report 486 , Statistics Department University of California, Berkeley

[18] Friedman, J. H. "Greedy Function Approximation: A Gradient Boosting Machine." (February 1999)

[19] 李欣海. 随机森林模型在分类与回归分析中的应用[J]. 应用昆虫学报,2013,04:1190-1197.

[20] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;321:436－44.

[21] Langkvist M, Karlsson L, Loutfi A. A review of unsupervised feature learning and deep learning for time-series modeling. Pattern Recogn Lett 2014;42:11－24.

[22] Liaw A, Wiener M. Classification and regression by randomForest. R News 2002;2(3):18－22.

[23] "Why Should I Trust You?": Explaining the Predictions of Any Classifier

Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin

arXiv:1602.04938 [cs.LG]

致谢

感谢宋博通老师对我毕业论文的指导！非常感谢宋老师能支持我做这次的论文研究。从论文的选题到文章逻辑的把握，都离不开宋老师的淳淳教诲。

感谢范成老师对我建立模型的指导。如果没有上过您的计量经济学我也不会接触到数据分析这个领域。尤其感谢您在 R 语言使用上的教导，让我这次的论文研究成为可能。

感谢宋老师的研究生们对我毕业论文完成过程中的指导与帮助，每次和你们一起开研讨会都觉得收获颇多。

感谢我本科期间帮助、教导我的老师和同学们，尤其感谢舒放对我的帮助。

感谢深圳大学土木工程学院对我的培养，感谢工程管理专业的灵活性，让我们都能有机会接触到自己感兴趣的领域。