Finance 金融, 2019, 9(6), 586-594
Published Online November 2019 in Hans. http://www.hanspub.org/journal/fin
https://doi.org/10.12677/fin.2019.96065

Hans汉斯

Research of Prediction on House Rent Based on Intergration Learning

Tao Ma, Ningning Liu

University of International Business and Economics, Beijing
Email: 15293035845@163.com

Received: Oct. 8th, 2019; accepted: Oct. 23rd, 2019; published: Oct. 30th, 2019

Abstract

The rapid development of the housing rental market has led to an increasing demand for housing rental information. There is always a problem of information asymmetry at both ends of the rental market. The rent is determined by many factors together. Accuracy of a single prediction model is unstable and is often affected by factors such as model performance, noise, and over-fitting risk. This study aims to develop and evaluate models of rental market dynamics using stacking integration strategy on data from the DC competition community. We use the three basic models of Random Force Regressor, Extra Trees Regressor and LightGBM and establish a rent prediction model for integrated learning. The experimental results show that the prediction accuracy of this method is obviously better than any single prediction model, which improves the accuracy and stability of the prediction, and confirms the validity of the model in rent prediction.

Keywords

Integrated Learning, Rent Forecast, Random Forest, Extra-Trees

基于集成学习的房租预测研究

马　涛，刘宁宁

对外经济贸易大学，北京
Email: 15293035845@163.com

收稿日期：2019年10月8日；录用日期：2019年10月23日；发布日期：2019年10月30日

摘　要

住房租赁市场的快速发展使得人们对房屋租赁信息的需求不断增加，对房屋租金关注持续变高。租房市

文章引用：马涛，刘宁宁．基于集成学习的房租预测研究[J]．金融, 2019, 9(6): 586-594.
DOI: 10.12677/fin.2019.96065

场供给两端一直存在着信息不对称的问题，房租是由诸多方面因素共同决定的，而现有的基于单一算法的房租预测模型，其预测精度往往受模型性能好坏、噪声、以及过度拟合风险等因素影响。本文基于堆叠集成策略，融合Random Forest Regressor、Extra Trees Regressor、LightGBM三个基模型，建立了集成学习的房租预测模型。研究结果表明，本方法预测精度明显优于任一单一预测模型，提高了预测的准确性和稳定性，证实了该模型在房租预测上的有效性。

关键词

集成学习，房租预测，随机森林，极端随机森林

1. 引言

据国家统计局数据，2018 年我国房屋租赁市场规模已达 2.5 万亿元。随着住房市场发展进入快车道，预计 2025 年房屋租赁总收入将接近 3 万亿元，租赁人口达 2.3 亿。对于租房这个相对传统的行业来说，信息严重不对称一直存在。一方面，房东不了解租房的市场真实价格，只能忍痛空置高租金的房屋；另一方面，租客也找不到满足自己需求的高性价比房屋，这造成了租房资源的极大浪费[1]。在当今大数据时代背景下，机器学习理论不断创新和发展，互联网、大数据、人工智能和实体经济融合更加深入，运用互联网技术和信息化手段去解决现实生活中的问题是完全可行的。

本文基于最新的机器学习算法，采用模型融合的集成策略，利用出租房屋的一些特征数据，对其租金进行预测，使房东和租客之间的供需匹配更为精确，降低租房的时间成本，有一定的学术意义和时间价值，同时也为解决房屋租赁市场的信息不对称问题提供了一种新的思路和方法。

2. 相关研究

近年来，随着房屋租赁市场的蓬勃发展以及租房人口的迅速增长，住房租金的波动越来越受到人们的关注，而房租由装修情况、位置地段、户型格局、交通便利程度、市场供需等多方面因素综合决定。因此，如何进行合理有效的预测已成为房租研究领域的热点问题。

目前，学者们对于房租预测问题的研究多集中于租金影响因素、定价和工序等定性方面的研究。例如，郑文娟以“因素观”视角来研究我国城市住房价格与住房租金问题，采用全国 35 个大中城市的面板数据，实证研究了我国城市住房价格与住房租金各自的关键影响因素，以及住房价格与住房租金之间关系在短期上与长期上的具体表现[2]。陈思翀和陈英楠基于资产定价的视角，通过将标准的动态戈登增长模型和传统的住房使用成本模型相结合，建立了一个关于住房市场租金收益率的动态住房使用成本模型，并将该模型应用于京沪广深四大城市的季度数据，使用方差分解方法来考察国内住房市场动态波动的影响因素及其相对重要性[3]。相较于以实证分析为主的定性研究，利用机器学习模型，以大样本数据集作为驱动的对于房租预测的定量研究则相对较少。Jinze Li 基于脱敏后的实际租赁市场的数据，利用月租标签的历史数据建立基于机器学习的 LightGBM (Light Gradient Boosting)模型，并对房屋月租金进行了较为准确的预测，为城市租赁市场提供客观的度量[4]。Y. Ma 和 Z. Zhang 主要研究了共享仓库租金的定价模型，为了理解定价机制，作者从分类广告网站上收集构建了一个北京地区的仓库数据集，基于该数据集，

应用机器学习技术将仓库价格与其相关特征关联，并对比了线性回归、回归树、随机森林和梯度增强四个模型，通过相关系数比较，发现随机森林模型表现最佳[5]。

当前对于房租预测领域的定量研究相对较少，我们参考了一些住房价格和商品价格的预测研究。J.J. Wang 和 S.G. Hu 设计一种基于 memristors、具有反向传播算法的多变量回归模型。用该 ANN 模型训练和预测了美国波士顿城镇的房价，得到了较为精确的预测结果[6]。Jingyi Mu 将支持向量机(SVM)、最小二乘支持向量机(LSSVM)算法应用于房地产价值预测，对房屋价值进行预测。选择波士顿郊区房屋数据集样本，对住宅价值进行了预测。首先建立了几种机器学习方法的模型并分析数据，然后结合测试数据的相应特性来预测房屋价值[7]。李春生、李霄野等人优化调整了 BP 神经网络的初始权值和阈值，分别对传统 BP 神经网络和改进后的 GA-BP 神经网络建立了房价预测模型。实验结果表明，经过遗产算法改进的 BP 神经网络较传统 BP 神经网络具有预测精度高、收敛速度快的优点，同时避免了陷入局部最优的缺陷[8]。Kanwal Noor 和 Sadaqat Jan 提出了一种基于监督机器学习技术的车辆价格预测系统。该研究使用线性回归作为预测方法，预测精度达 98%。在多元线性回归中将车辆价格作为因变量，自变量包括车辆模型、制造城市、型号、颜色、里程、合金轮辋和动力转向等[9]。

综上所述，借鉴学者们在房租、房价及商品价格预测领域的研究。我们发现相较于理论基础完善的定性研究，借助于机器学习工具进行的定量研究相对较少，且研究大多只采用单一模型，无法克服各模型固有的局限性，使得预测结果可能不够准确完备。在机器学习算法不断发展更新的背景下，相较于单个模型，结合多个模型算法的集成策略在房租预测领域的应用具备先天优势，将会进一步的提高模型的预测精度和稳定性。

3. 研究思路和方法

3.1. 问题分析

房屋租金预测本质上是回归预测问题，本文采用模型融合的集成策略，在对原数据集进行数据清洗和特征工程的基础上，先对单个模型进行学习和选择，然后运用 Stacking 进行模型融合，并对测试集进行预测，结果评价指标为均方根误差(RMSE)。研究具体步骤见图 1 所示：

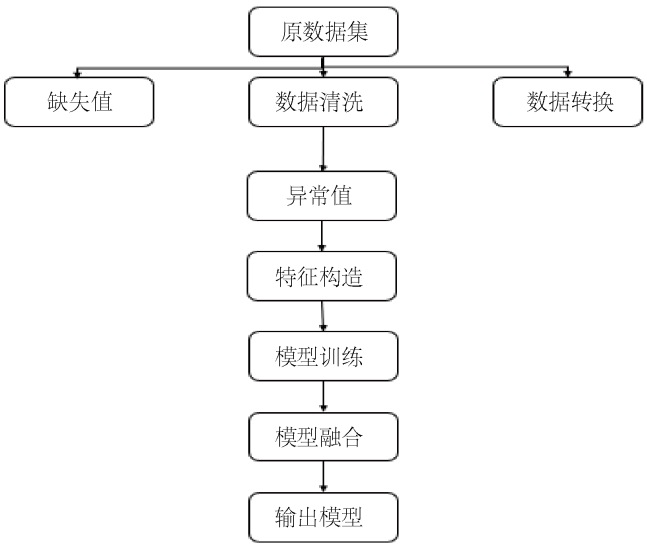

Figure 1. Solution of the problem
图 1. 问题解决思路

3.2. 模型比较

随机森林、极端随机森林、LightGBM 是三种常用的价格回归预测模型。随机森林本质上一种 Bagging 算法，它以分类树为基本单元，通过二分数据进行分类或回归，相比与决策树，随机森林在不增加计算复杂度的前提下提高了预测精度。极端随机树(Extra-Trees)算法与随机森林算法十分相似，都是由决策树构成的。LightGBM 是由微软亚洲软件开发院开源公布的一种快速的，分布式的，高性能的基于决策树算法的梯度提升框架。相比与 XGBoost，LightGBM 在不降低准确率的前提下，速度提升了 10 倍左右，占用内存下降了 3 倍左右，各模型优势比较见表 1 所示。

Table 1. Model comparison
表 1. 模型对比

随机森林	在数据集上表现良好，不容易陷入过拟合； 具有良好的抗噪声能力； 能够处理高纬度的数据。对数据集的适应能力强。
极端随机森林	完全随机得到分叉值。从而实现对决策树进行分叉； 使用所有样本训练，在某种程度上效果优于随机森林。
LightGBM	训练效率快、可以处理大规模数据； 支持并行优化学习； 内存使用率较低。

3.3. 集成学习策略

模型融合的基本思想就是通过对多个单模型融合以提升整体性能。常用的模型融合方法有 Voting、Averging、Bagging、Boosting 等。本文采用 Stacking 作为模型融合策略，具体流程如下图 2 所示。

Stacking 是一种非线性的融合决策,是一种从原数据集中自动抽取有效特征的表示学习。一般来说 Stacking 就是训练一个多层的学习器结构，第一层称为学习层，用 n 个不同的分类器，将得到的预测结果合并为新的特征集，并作为下一层分类器的输入，通过第二层的输出训练器得到最终预测结果。为了防止过度拟合问题，Stacking 在第一层模型训练时采用 K 折交叉检验的方式，第二层输出训练器一般选用逻辑回归模型。

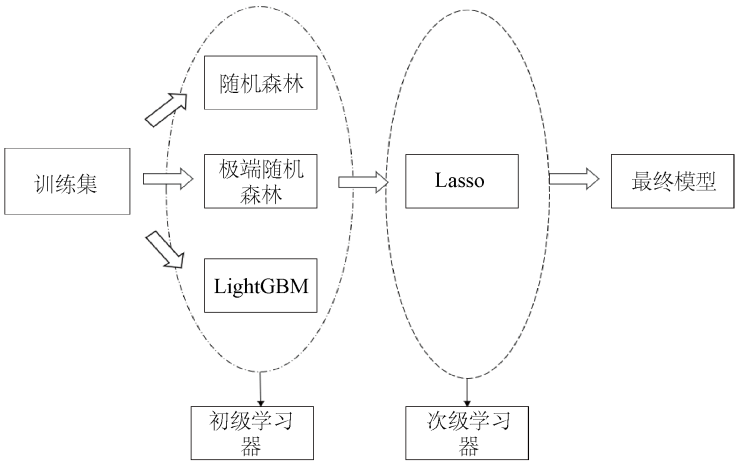

Figure 2. Stacking strategy
图 2. Stacking 策略

4. 数据

本文数据集来自 DC 竞赛社区住房月租金预测大赛所提供的某地 3 个月的房屋租赁价格以及房屋的基本信息。该数据集共有 196539 条数据，19 个特征变量，15 个为数值型变量，4 个为分类型变量。所有特征如下表 2 所示，房租：

Table 2. Feature description
表 2. 特征描述

字段名	说明	字段名	说明
时间	房屋信息采集时间	小区名	房屋所在小区(脱敏)
小区房屋出租数量	小区房屋出租数量(脱敏，保留大小关系)	楼层	楼层高中低(脱敏)
总楼层	房屋所在建筑的总楼层数(脱敏)	房屋面积	房屋面积(脱敏)
房屋朝向	房屋朝向	居住状态	是否已经出租或居住中(脱敏)
卧室数量	卧室数量	客厅数量	客厅数量
卫的数量	卫的数量	租出方式	表示是否出租
区	房屋所在区级行政单位(脱敏)	位置	小区所在商圈(脱敏)
地铁线路	第几条线路(脱敏)	地铁站点	房屋临近站点(脱敏)
距离	房屋距地铁站距离(脱敏)	装修情况	房屋装修档次(脱敏)

4.1. 数据预处理

通过对原始数据集进行分析发现，数据不仅包括房屋面积、楼层、客厅数量等数值型数据，还包括租出方式、房屋朝向等非数值型数据；同时该数据集中也存在异常值、缺失值、以及不一致数据。为了使数据集符合建模要求，需要进行必要的数据清理和转换。

首先对缺失值进行统计分析，其中装修情况、居住状态、出租方式字段缺失率较高，考虑实际情况，以上三个特征都是分类变量，因此将其缺失值作为一种特征处理。对于地铁站点、距离、地铁线路三个特征的缺失值，用相同小区名的数据代替，若该小区所有房屋都缺失这三个字段，就将其作为特征处理，表示该房屋附近没有地铁。小区房屋出租数量字段用相同小区、邻近楼层和邻近时间的值进行填充。位置和区的缺失值用所有数据的中位数代替。接下来对异常值进行处理，经过统计分析发现房屋面积字段存在明显异常值，对于房屋面积超出 0.146 的数据进行删除。最后数据集中房屋朝向特征用中文表述，为了匹配模型对其进行数据转换，用数值型数据代替中文。

4.2. 特征工程

接下来对清洗好的数据集进行特征处理。由图 3 可知房屋面积、卧室数量、厅的数量、卫的数量与月租金相关性较高，这也符合现实中的情况，如两室一厅和三室一厅的房屋在租金上是有很大差别的。因此在特征构造时，主要考虑以上特征之间的关系。在构造特征时，对卧室面积、卫生间面积和厅面积占整个房屋面积比例进行了量化处理，按照现实房屋中的一般占比确定系数，原数据集中楼层用 0，1，2 表示，对其进行归一化和量化处理，然后与总楼层相乘便可得到具体楼层。所有新构造特征如下表 3 所示：

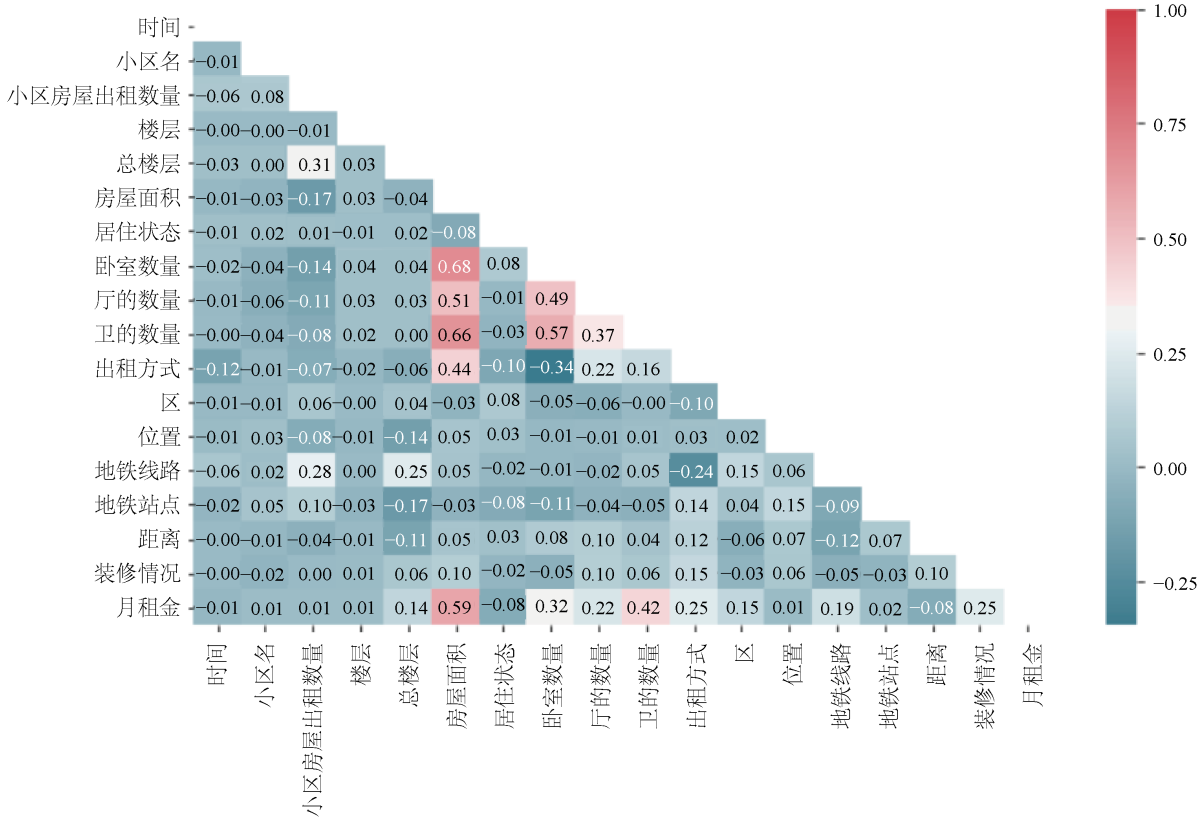

Figure 3. Feature correlation analysis
图 3. 特征相关性分析

Table 3. Characteristic construction
表 3. 特征构造

构造特征	描述
卧室均面积	房屋面积 × 0.3/卧室数量
卫的均面积	房屋面积 × 0.07/卫的数量
卧室总面积	卧室均面积 × 卧室数量
除卧室外剩余面积	房屋面积 − 卧室总面积
客厅均面积	房屋面积 × 0.26/厅的数量
客厅总面积	客厅均面积 × 客厅数量
卧室和厅	卧室数量 + 厅的数量
具体楼层	总楼层 × 100 × 楼层系数

5. 模型的建立

5.1. 单一模型训练

本文选用 Stacking 集成策略进行模型融合，为了构建初级学习器，需要选择若干个基模型。首先，将经过数据处理的数据集划分为训练集和测试集，其中训练集占 80%，测试集占 20%。在回归模型的选择上，本文初始考虑 Lasso、Catbost、Random Forest Regressor、Extra Trees Regressor、Lightgbm 等 5 个

模型。为了鉴别模型的优劣，我们定义了均方根误差(RMSE)和模型拟合优度(SCORE)作为交叉验证的评估指标，经过初步筛选，本文选择了 Random Forest Regressor、Extra Trees Regressor、Lightgbm 三个模型，并利用 Grid Search CV (网格搜索)对模型参数进行调整，以达到效果最优，调参后各模型得分和模型最优参数如下表 4、表 5、表 6、表 7 所示。

Table 4. RF Optimal parameters
表 4. RF 最优参数

参数名称	参数含义	取值
max_features	单个决策树使用的最大特征数量	10
n_estimators	子树的数量	10

Table 5. ET Optimal parameters
表 5. ET 最优参数

参数名称	参数含义	取值
max_features	单个决策树使用的最大特征数量	6
n_estimators	子树的数量	30

Table 6. Lgb Optimal parameters
表 6. Lgb 最优参数

参数名称	参数含义	取值
num_leaves	数模型复杂度	900
learning_rate	学习率	0.1
n_estimators	子树的数量	3141
bagging_fraction	在不进行重采样的情况下随机选择部分数据	0.7
feature_fraction	随机选择部分特征	0.6
min_data_in_leaf	一个叶子上数据的最小数量	18
min_sum_hessian_in_leaf	一个叶子上的最小 hessian	0.001

Table 7. Scores of each model
表 7. 各模型得分

模型名称	mean	score
RandomForestRegressor	1.48	0.94
ExtraTreesRegressor	1.36	0.95
Lightgbm	1.35	0.95
Catbost	1.40	0.88

5.2. 集成学习

上一节中已经选定了三个基模型作为初级学习器，在此基础上对模型进行 Stacking 集成。本文选择 Lasso 模型作为次级学习器，并采用 5-折交叉检验的方法对基模型进行训练。这个训练过程主要分两层：将原始训练集分为 5 折，记为 fold1~fold5，依次取其中的四折数据来训练模型一，对测试集进行预测，并对剩余的一折数据进行预测，预测值即作为基模型对一折数据生成的原特征，将五组原特征拼接起来，

得到该模型对整个原始训练集生成的原特征，而对测试集的预测结果，取其五次预测的平均值。同样地，对其它基模型也采用相同方法生成元特征，从而构成用于第二层模型训练的完整原特征集。通过初级学习器的训练，得到了 3 份 train 数据和 3 份 test 数据，然后用 Lasso 模型进行进一步融合，得到最终预测值。

将单个模型和 Stacking 集成策略模型做损失函数对比，可以看出集成策略在 RF、ET、LGB 模型的基础上进一步的提高了预测精度，均方误差比最优子模型降低了 1.5%左右，实验结果表明，基于集成学习的房租预测模型对于提高房租的预测效果是有效的，结果对比如下表 8 所示。

Table 8. Mean square error and mean absolute error of each model
表 8. 各模型均方误差和平均绝对误差

模型名称	mean	MAE
RandomForestRegressor	1.48	0.85
ExtraTreesRegressor	1.36	0.78
Lightgbm	1.35	0.79
StackingCVRegressor	1.33	0.79

6. 总结与展望

本研究构建了一种基于 Stacking 集成策略的两层模型，使用 DC 竞赛社区住房月租金预测大赛提供的数据集。以筛选和调参之后的 Random Forest Regressor、Extra Trees Regressor、Lightgbm 模型为基模型，并选用 Lasso 作为次级学习器中的融合模型，运用训练集中的数据训练模型，并对测试集的月租金进行预测。实验结果表明 Stacking 集成模型结果要优于任一单个模型，其采用交叉验证的方法构造，稳健性强，并且融合多个模型判断结果，进行次级训练，预测精度高。

同时本研究也存在进一步探讨的空间。在数据集的选择上，可以参考一些住房租赁网站提供的数据，使用不同的数据集训练模型，以检验模型的泛化性。在特征的构造上，可以结合中国住房租赁市场的实际情况，构造更为合理、有效的特征。在模型的选择上，由于房屋租赁价格与位置、地段等因素相关，因此在基模型的选择上可以考虑加入深度学习网络，利用租房分布的地理位置地图进行模型训练，以提升基模型的多样性，进一步提高集成模型预测的精度。

致谢

感谢我的导师刘宁宁老师，在整个论文写作过程中给予我的大力帮助。刘老师的悉心指导贯穿了论文写作的方方面面，在他的指导下我认识到了自己很多不足，并在这一过程中取得进步。

基金项目

这项工作得到了国家青年科学基金资助(批准号：61806056)，北京市社会科学青年基金资助(批准号：17YYC015)，中央高校基本科研业务专项资金资助(批准号：CXTD10-05)。

参考文献

[1] 中国软件行业协会培训中心. 2018 年全国大学生计算机技能应用大赛[EB/OL]. http://www.cnccac.com/, 2018-8-20.

[2] 郑文娟. 中国城市住房价格与住房租金的影响因素及相互关系研究[D]: [博士学位论文]. 浙江: 浙江大学, 2011.

[3] 陈思翀, 陈英楠. 中国住房市场波动的影响因素研究——基于租金收益率的方差分解[J]. 金融研究, 2019,

464(2): 140-157.

[4] Li, J.Z. (2018) Monthly Housing Rent Forecast Based on LightGBM (Light Gradient Boosting) Model. *International Journal of Intelligent Information and Management Science*, **7**, 6.

[5] Ma, Y., Zhang, Z., Ihler, A. and Pan, B. (2018) Estimating Warehouse Rental Price Using Machine Learning Techniques. *International Journal of Computers Communications & Control*, **13**, 235-250 https://doi.org/10.15837/ijccc.2018.2.3034

[6] Wang, J.J., Hu, S.G., Zhan, X.T., *et al*. (2018) Predicting House Price with a Memristor-Based Artificial Neural Network. *IEEE Access*, **6**, 6. https://doi.org/10.1109/ACCESS.2018.2814065

[7] Mu, J., Wu, F. and Zhang, A. (2014) Housing Value Forecasting Based on Machine Learning Methods. *Abstract and Applied Analysis*, **2014**, Article ID: 648047. https://doi.org/10.1155/2014/648047

[8] 李春生, 李霄野, 张可佳. 基于遗传算法改进的 BP 神经网络房价预测分析[J]. 计算机技术与发展, 2018, 28(8): 144-147.

[9] Noor, K. and Jan, S. (2017) Vehicle Price Prediction System Using Machine Learning Techniques. International Journal of Computer Applications, **167**, 27-31. https://doi.org/10.5120/ijca2017914373