Computer Science and Application 计算机科学与应用, 2022, 12(3), 697-706
Published Online March 2022 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2022.123071

Hans汉斯

基于机器学习的共享单车需求预测

颜轲越[1]，王祎萌[1]，李　莹[2]

[1]澳门大学科技学院，澳门
[2]北京理工大学珠海学院，广东 珠海

收稿日期：2022年2月23日；录用日期：2022年3月22日；发布日期：2022年3月29日

摘　要

随着中国城市的发展，共享经济的概念已经逐渐融入人们的生活。共享单车、共享汽车、共享充电宝等新事物正在不断地改变着大家的生活习惯。摩拜单车、美团单车等知名共享单车企业成立以来，已经成为绿色出行、健康环保的代表。然而，共享单车的过度投放、安全隐患、管理混乱等问题也引起了共享单车管理者的关注，同时影响了各共享公司的运营和盈利。本研究根据实时收集到的时间、季节、天气、温度、湿度、风速等数据；构造更加符合共享单车使用场景的离散型变量，并将数据输入到不同的机器学习和深度学习模型，达到准确预测城市中的共享单车需求的目的。通过比较Rsquare、MSE、RMSE等指标来评估所有模型的预测效果。基于模型得出的结果，共享单车管理者能够合理投放相应数量的共享单车，达到减少浪费的目的。

关键词

共享单车，预测，机器学习

Shared Bicycles Demand Prediction Based on Machine Learning

Keyue Yan[1], Yimeng Wang[1], Ying Li[2]

[1]Faculty of Science and Technology, University of Macau, Macao
[2]Beijing Institute of Technology, Zhuhai, Zhuhai Guangdong

Received: Feb. 23rd, 2022; accepted: Mar. 22nd, 2022; published: Mar. 29th, 2022

Abstract

With the development of Chinese cities, the concept of sharing economy has gradually been integrated into people's lives. New concepts such as shared bicycles, shared cars, and shared power banks are constantly changing everyone's living habits. Mobike, Meituan and other well-known sharing bicycles companies have become representatives of green travel, health and environmen-

文章引用: 颜轲越, 王祎萌, 李莹. 基于机器学习的共享单车需求预测[J]. 计算机科学与应用, 2022, 12(3): 697-706.
DOI: 10.12677/csa.2022.123071

tal protection since their establishment. However, problems such as excessive delivery of bicycles, potential safety hazards, and management confusion have aroused the concerns of bicycle-sharing managers, which also affects the operation and profitability of each sharing company. After collecting the data in real time, such as time of day, season, weather, temperature, humidity, and wind speed, this research constructs some new discrete variables which are more in line with the use of bicycle sharing scenarios. After feeding the data into different machine learning and deep learning models, they achieve accurate prediction of bike-sharing demand in cities; also R square, MSE, RMSE and other metrics are used to evaluate prediction effectiveness of all the models. Based on the prediction results, bicycle sharing companies' managers can deliver the appropriate number of bicycles reasonably and reduce waste.

Keywords

Shared Bicycles, Prediction, Machine Learning

1. 引言

共享经济是由德克萨斯州立大学的社会学教授马科斯－菲尔森和伊利诺伊大学的社会学教授琼－斯彭斯首先提出的。其基本概念是通过整合线下的闲置物品等资源，通过临时转让使用权获得回报[1]。而共享单车就是共享公司通过利用互联网平台向居住在大城市的人们提供单车租赁服务，从中获得一定的报酬。如今，随着互联网的快速稳定发展，共享单车已经成为中国“新四大发明”之一。目前中国有 30 多家互联网租赁自行车运营企业，累计车辆数已超过 1000 万辆，累计服务人数已超过 10 亿人次[2]。共享单车的现状已经成为研究的热点话题。随着共享单车使用量的增加，大城市的交通管理也将变得越来越困难。共享公司试图根据用户的使用数据寻找出如何控制某个城市中共享单车的数量。

在以往的研究当中，曹旦旦和范书瑞等人[3]通过传统的机器学习算法对共享单车的短时需求量进行建模预测。宋鹏和黄同愿等人[4]通过对原始数据消除噪声，降维处理，仅使用支持向量机模型就对单车的需求量有了很好的预测效果。李福和徐良杰等人[5]通过对比不同类型的机器学习模型，证明极端梯度推进决策树在共享单车需求上的预测效果更优。李天骋[6]在他的研究中加入了简单的神经网络模型，通过和集成学习的对比发现神经网络模型具有较好的表现。

在本研究中，我们将基于共享单车使用环境的基本特征，利用实时数据来探讨时间、季节、天气、温度、湿度、风速等因素对共享单车需求量的影响。通过输入预处理后的数据对多种机器学习和深度学习模型，例如人工神经网络模型[7]和长短记忆神经网络模型[8]进行训练，预测共享单车需求量并对各模型之间的效果评估。结果显示，上述数据对共享单车的需求量有很好的预测。其中深度学习模型表现更佳。在现实场景中，通过收集共享单车周边的环境数据，共享公司可以根据数据进行建模对共享单车的需求量进行预测，从而达到精准投放对应数量共享单车的目的。

2. 研究方法

2.1. 数据描述

在本研究中，我们使用从 Kaggle 网站[9]下载的共享单车需求数据。原始数据集包含了从 2011 年 01

月 01 日 00:00 到 2012 年 12 月 19 日 23:00 的 1 小时间隔数据，一共包含了 10,886 个样本。数据集中共有 12 个字段，其中“count”是我们要预测的结果标签。

在自变量中，“temp”、“atemp”、“humidity”、“windspeed”、“casual”和“registered”是连续型变量，表示用户在租用共享单车时的温度、湿度和风速。其中“temp”和“atemp”是温度的两种不同表达方式，“casual”和“registered”表示使用共享单车的用户是否为平台的注册用户。但是在数据集中，“casual”和“registered”的求和等于我们需要预测的变量“count”。为了防止出现数据泄漏的问题，在进行数据预处理时直接删除这两个变量。此外，“season”、“holiday”、“workingday”和“weather”都是离散型变量。预处理时将使用频率编码(Frequency Encoding)。

在此研究中，虽然数据集本身已经提供了自变量，但我们希望可以构造更多有意义的自变量用来做预测，从而提高模型的精度。根据“season”、“weather”、“temp”、“humidity”、“windspeed”等其余变量，我们可以构建一些常见的指标作为探索性数据分析的变量。

2.2. 探索性数据分析

首先，我们绘制要预测的结果标签“count”的图表。其频率分布如图 1 所示。

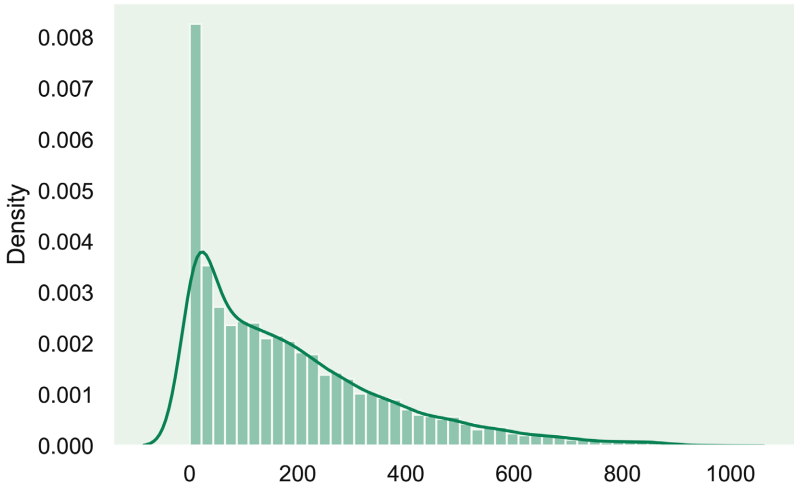

Figure 1. Frequency distribution graph of variable “count”
图 1. 变量“count”的频率分布图

从图中可以看出共享单车的使用量集中在 0~200 之间，但在 200~1000 的大范围之间，仍有大量的数据分布，预测标签的数据分布不平衡会影响机器学习方法的预测结果。因此，我们用 log (“count”)的计算公式对“count”进行对数预处理。对数运算后的新分布如图 2 所示。它的范围基本上在 0 到 7 之间；且分布也更加平缓，对数分布处理后的数据要比处理前要好得多，且对基于机器学习和深度学习中回归分析也会更加友好。

接下来对连续变量进行相关分析，相关系数矩阵的可视化如图 3 所示。变量“temp”和“atemp”是高度相关的。为了降低预测模型的维度，可以删除“atemp”这一特征。

连续变量“humidity”的绘制如图 4 所示，湿度分布范围为 0~100，在 47~77 (25%百分位数~75%百分位)之间分布较多，其余变量较少。所以我们重新定义了一个离散变量“humi_categori”。当湿度小于 40 时，“humi_categori”为低，当它在 47~77 时，“humi_categori”为正常，而当它大于 77 时，“humi_categori”为高。

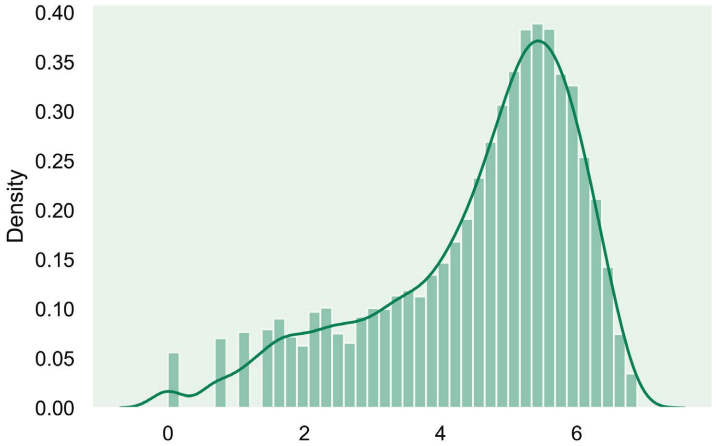

Figure 2. Frequency distribution graph of variable log (“count”)
图 2. 变量 log (“count”)的频率分布图

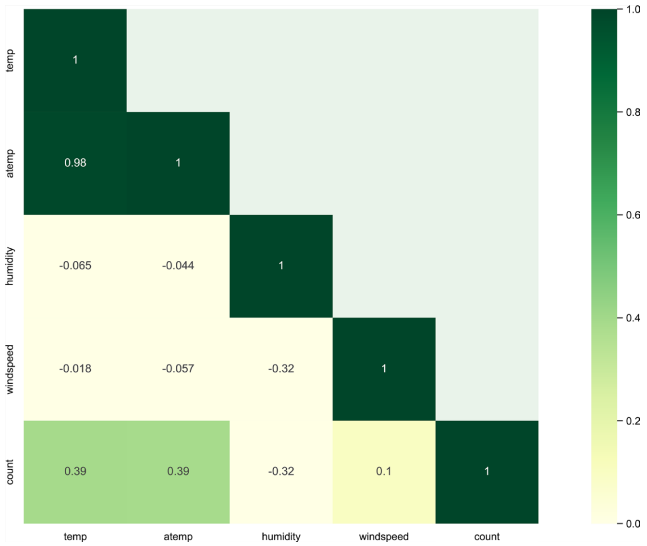

Figure 3. Correlation matrix of continuous variables
图 3. 连续变量的相关系数矩阵

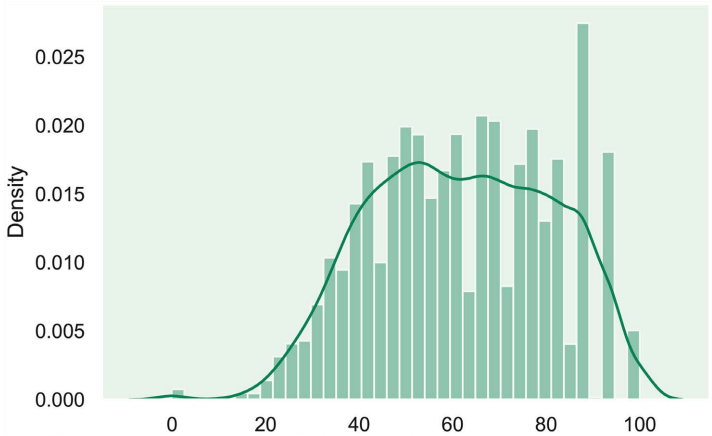

Figure 4. Frequency distribution graph of variable “humidity”
图 4. 变量“humidity”的频率分布图

连续变量“windspeed”的绘制如图 5 所示，湿度分布范围为 0~40，更多的分布在 7~17 之间(25%百分位数~75%百分位)。所以我们重新定义了一个离散变量“wind_categori”。当风速小于 7 时，“wind_categori”为低，当它在 7~17 时，“wind_categori”为正常，当它大于 17 时，“wind_categori”为高。

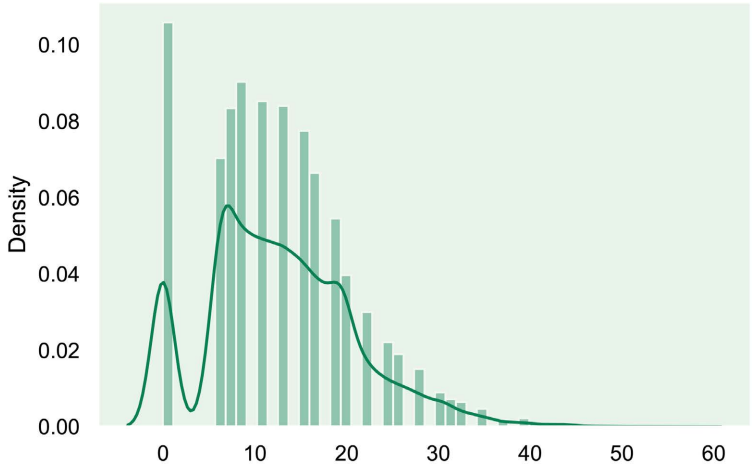

Figure 5. Frequency distribution graph of variable “windspeed”
图 5. 变量“windspeed”的频率分布图

对于离散型变量，可以根据探索数据分析过程中的实际情况构造额外的变量，如图 6 所示。通过观察一天中不同时间段的自行车实际使用情况，我们发现上午 7~9 点和下午 16~19 点用户对共享单车的使用需求非常大，而其他时间段的共享单车使用量相对较低。因此，我们构建了一个新的变量“hour_categori”，将 7~9 点、16~19 点作为共享单车的高峰期，其余时间作为休闲期。

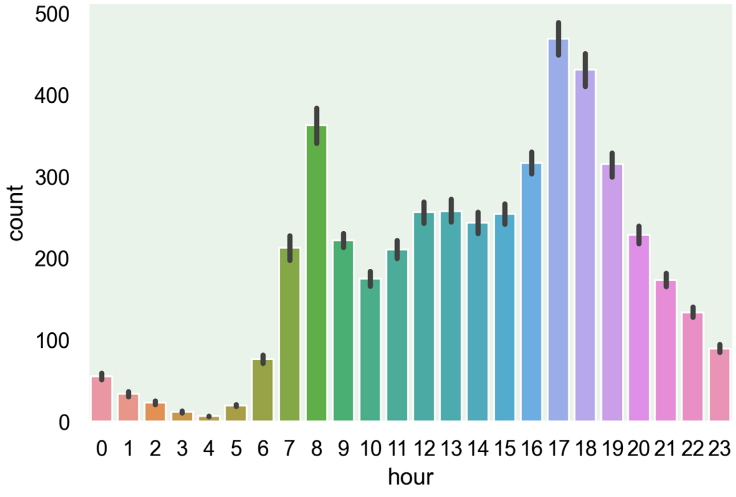

Figure 6. The usage of shared bicycles at different time periods
图 6. 不同时间段的共享单车使用量

构建完所有的变量之后，为了提高模型计算速度；减少训练时间。我们用频率编码(Frequency Encoding)的方法取代独热编码(One-hot Encoding)来处理所有的离散变量[10]，并删除无用的变量，留下 14 个独立变量作为机器学习模型的输入。

3. 机器学习模型

3.1. 线性回归(Linear Regression)

普通线性回归模型(Linear Regression)是最简单也是最传统的机器学习模型，在计量经济学和金融计算领域被广泛使用。在本研究中，我们将通过梯度下降的迭代优化方法调整线性回归模型的参数[11]。

3.2. 支持向量机(SVM)

SVM 可以用于解决线性和非线性的分类或者回归问题。与 SVM 分类模型一样，SVM 回归模型在找到目标函数后需要在函数的两边构建一个间隔。与分类模型相比，回归模型希望间隔能够覆盖尽可能多的实例，同时限制间隔以外的实例[12]。在利用权重向量最小化解决优化边际的约束优化问题后，可以得到目标函数并计算出预测值。

3.3. 集成学习

“随机森林是决策树的一个集合体”[13]。随机森林(Random Forest)预测器的本质是决策树模型。与构建单一的决策树不同，随机森林会从随机选择的样本特征中找到最佳特征来划分子节点，而不是使用数据中的所有特征来构建决策树[14]。虽然不同的决策树之间相互独立，但随机森林的最终输出取决于森林中的每个决策树。同样地，预测值是随机森林中所有决策树预测结果的平均值。

Adaboost 是一种广泛使用的集成学习方法。然而，Adaboost 中的预测器之间有着很高的相关性。首先，Adaboost 尝试训练一个预测器并使用其进行预测。根据样本误差，Adaboost 算法将会更新实例权重，再将新的权重用于训练下一个预测器，并一直重复以上步骤[15]。

Bootstrap Aggregating (Bagging)是机器学习模型中的一种集成学习的方法。与单一机器学习模型作为预测器不同，Bagging 算法中拥有许多预测器。Bagging 会随机抽取相同数量的样本到不同的样本集中，然后使用每个独立的样本集来训练预测器。每个预测器结果的平均值就是回归中的最终预测值。

Xgboost 也叫 eXtreme Gradient Boosting，是一个优化的分布式梯度提升库，其目的是实现高效、灵活的特点。可移植性强[16]。Xgboost 包是一个用于大规模并行提升树的工具，通过使用这个工具我们可以很容易进行建模研究。

3.4. 深度学习

人工神经网络(ANN)模型最早由神经生理学家 Warren McCulloch 和数学家 Walter Pitts 提出[7]。该模型是通过模拟动物大脑中生物神经元的工作情况而发明的。如今，各行各业都有大量数据可供人工神经网络训练，优化参数，从而达到很好的预测效果。因此，人工神经网络的整体性能往往优于其他传统的机器学习模型。

上述所有的模型都是没有记忆的，变量输入与输入之间没有保存任何状态。这些模型大部分情况下用于预测结构型数据。而长短记忆神经网络(LSTM)则需要输入整个序列；遍历输入序列的所有序列元素，并保存一个状态(state) [8]，这样的网络结构设计使得 LSTM 模型在处理时间序列类型的数据具有更好的拟合效果。

4. 研究结果

实验过程中将 80%数据用作训练集，剩下 20%的数据作为测试集。下面 5 个回归指标将被用于衡量不同模型的评估效果。

R square 描述了使用预测值和只使用平均值与真实值之间的较小误差的比较。值域通常位于 0 和 1

之间。当 R square 接近 1 时，表示使用这个模型的预测值可以得到更小的误差，模型的拟合效果越好。同理，如果一个回归模型的 R square 越接近于 0，甚至低于 0，这个模型的表现则非常糟糕。R square 的计算公式如下。

$$\text{R square} = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}_i\right)^2} \tag{1}$$

平均绝对误差(MAE)是预测误差的平均绝对值。模型对应的 MAE 越小，模型则有更好的效果。其公式如下：

$$\text{MAE} = \frac{\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|}{n} \tag{2}$$

此外，平均平方误差(MSE)是预测误差的平方的平均值，人们通常希望其值很低。而平均平方误差(RMSE)是 MSE 的平方根。RMSE 的值越小说明模型效果越好。这两个指标的公式如下：

$$\text{MSE} = \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n} \tag{3}$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n}} \tag{4}$$

平均绝对百分比误差(MAPE)描述了预测误差的在真实值上的平均占比，当 MAPE 越接近于 0，则预测误差的平均占比越低，模型效果越好；当 MAPE 越接近于 1，预测误差较大，模型效果越差。

$$\text{MAPE} = \frac{\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|}{n} \tag{5}$$

传统机器学习模型评估包含了线性模型，SVM 模型以及集成学习模型，且所有传统机器学习模型将调用 Sklearn 机器学习包[17]进行实验，结果如表 1 所示。Random Forest、Bagging 和 Xgboost 等集成学习模型的拟合度 R square 均超过了 0.9，说明回归效果很好。并且它们的 MSE 均在 0.16 左右，RMSE 在 0.41 左右，相对较低的误差说明了模型表现良好。然而，集成学习模型 Adaboost 和 SVM 的表现却比较一般：R square 处于 0.7 以下；MAE，MSE 和 RMSE 也处于较高水平，说明误差较大。在 MAPE 指标上，拟合效果越差的模型，MAPE 也会越高。通过这 6 个机器学习模型综合对比，由于线性模型在设计上更难找到不同变量之间非线性的联系，Linear Regression 模型在所有指标表现最差。

Table 1. Evaluation of traditional machine learning models
表 1. 传统机器学习模型评估

	R square	MAE	MSE	RMSE	MAPE
LinearRegression	0.6648	0.6292	0.6567	0.8104	0.1475
SVM	0.6577	0.6269	0.6707	0.8190	0.1453
RandomForest	0.9130	0.3008	0.1705	0.4129	0.0932
Adaboost	0.6913	0.6357	0.6048	0.7777	0.1678
Bagging	0.9134	0.2995	0.1696	0.4118	0.0915
Xgboost	0.9207	0.3011	0.1554	0.3943	0.0787

真实环境中 ANN 有神经元层数，神经元数量，激活函数等许多参数可供大家进行选择。为了使实验简单快速，我们使用普通的神经网络进行回归预测，利用 Keras [18]构造实验所需的深度学习模型。在其它参数不变的情况下，我们尝试使用不同的隐藏神经层数量和神经元数量来进行实验。和传统机器学习模型的实验一样，R square，MAE，MSE，RMSE 和 MAPE 会被用来评估模型的效果。不同 ANN 模型的实验结果如表 2 所示。

Table 2. Evaluation of ANN models
表 2. 人工神经网络模型评估

隐藏层数	神经元数量	R square	MAE	MSE	RMSE	MAPE
1	40	0.9257	0.2808	0.1455	0.3815	0.0693
1	60	0.9134	0.2904	0.1598	0.3997	0.0718
1	80	0.9252	0.2783	0.1381	0.3717	0.0668
1	100	0.9222	0.2966	0.1435	0.3789	0.0720
2	40	0.9258	0.2805	0.1454	0.3813	0.0776
2	60	0.9204	0.2766	0.1470	0.3834	0.0656
2	80	0.9071	0.3152	0.1715	0.4142	0.0774
2	100	0.9211	0.2862	0.1455	0.3815	0.0684
3	40	0.9330	0.2610	0.1313	0.3623	0.0680
3	60	0.9163	0.2981	0.1546	0.3932	0.0687
3	80	0.9102	0.2936	0.1657	0.4070	0.0703
3	100	0.8958	0.3249	0.1923	0.4385	0.0782
	平均值	0.9180	0.2901	0.1534	0.3911	0.0712

基于表 2 的实验结果，随着隐藏层和神经元数量的增加，R square 有逐渐提高的趋势，其整体平均性能达到 91.80%。整体上比传统机器学习模型的 R square 更大；模型拟合度更好。同样，MAE、MSE 和 RMSE 也随着神经网络复杂度的增加而减小。特别是当 ANN 的隐藏层数等于 3，神经元数等于 40 的时候，这些指标达到最优状态。在平均值上，MAE (0.2901)、MSE (0.1534)、RMSE (0.3911)和 MAPE (0.0712)也比传统机器学习模型的评估参数更小。综上，在大数据的支持下深度学习 ANN 模型在回归预测上往往有更好的结果。

同样，Keras 也可用于构造 LSTM 模型。为了方便与 ANN 进行对比，实验中的 LSTM 模型将会使用和 ANN 相同的隐藏层数和神经元数量。R square，MAE，MSE，RMSE 和 MAPE 也会被用来评估模型的效果。LSTM 的回归结果如表 3 所示。

Table 3. Evaluation of LSTM models
表 3. 长短记忆神经网络模型评估

隐藏层数	神经元数量	R square	MAE	MSE	RMSE	MAPE
1	40	0.9213	0.2834	0.1452	0.3811	0.0765
1	60	0.9352	0.2368	0.1197	0.3459	0.0616
1	80	0.9202	0.2773	0.1474	0.3839	0.0722

Continued

1	100	0.9289	0.2601	0.1312	0.3622	0.0649
2	40	0.9207	0.2623	0.1463	0.3825	0.0661
2	60	0.9268	0.2609	0.1351	0.3675	0.0698
2	80	0.9252	0.2610	0.1380	0.3715	0.0654
2	100	0.8985	0.3047	0.1875	0.4330	0.0715
3	40	0.9240	0.2652	0.1403	0.3746	0.0639
3	60	0.9222	0.2673	0.1436	0.3789	0.0658
3	80	0.9249	0.2720	0.1386	0.3723	0.0689
3	100	0.9181	0.2893	0.1511	0.3887	0.0756
	平均值	0.9222	0.2700	0.1437	0.3785	0.0685

通过对比两个深度学习模型的预测结果发现，虽然 ANN 与 LSTM 表现均非常优秀，但在随着神经网络复杂度的增加，LSTM 的表现比 ANN 更加地稳定。通过表 3 可知，R square 的平均值会保持在 0.9222 的水平，MAE，MSE，RMSE 和 MAPE 也不会出现大幅度的浮动。最终平均值为 MAE (0.2700)、MSE (0.1437)、RMSE (0.3785)和 MAPE (0.0685)。所有指标均优于传统机器学习模型，误差也比 ANN 模型更小。基于此共享单车使用量的数据，LSTM 在预测上，总体效果会更加占优。因此优先推荐使用 LSTM 这类时序类模型进行建模预测；其次使用 ANN 模型进行建模。

5. 总结

在共享单车数量的需求上，本研究提出了基于各种不同类型的机器学习模型。用季节、天气、温度、湿度、风速等变量训练模型，进行共享单车需求数量的预测。通过对比不同类型的模型，实验结果显示传统的机器学习模型和深度学习模型在预测上均有很好的效果，其中深度学习表现更佳，在不同的模型参数下依旧有比较稳定优秀的结果。在未来的研究中，还可以收集更多不同种类的实时数据(如：地理信息，使用者信息等)，尝试使用不同的特征进行模型训练。探索不同类型的变量对模型预测准确度的影响。

参考文献

[1] 徐丽. 共享经济产业链是一个动态的生态圈[N]. 人民邮电报, 2017-06-06(008).

[2] 共享单车累计投放超 1000 万辆管理新政将陆续出台——中新网[EB/OL]. https://www.chinanews.com/cj/2017/05-24/8232308.shtml, 2017-05-24.

[3] 曹旦旦, 范书瑞, 夏克文. 共享单车短时需求量预测的机器学习方法比较[J]. 计算机仿真, 2021, 38(1): 92-97.

[4] 宋鹏, 黄同愿, 刘渝桥. 基于 SVM 的共享单车需求预测[J]. 重庆理工大学学报(自然科学), 2019, 33(7): 187-194.

[5] 李福, 徐良杰, 朱然博, 罗浩顺, 陈国俊. 基于 XGBOOST 算法的共享单车借车需求量预测[J]. 武汉理工大学学报(交通科学与工程版), 2021, 45(5): 880-884.

[6] 李天骋. 基于机器学习方法的共享单车需求分析[J]. 现代商贸工业, 2020, 41(25): 40-41.

[7] 宋能辉, 李娴. 机器学习实战[M]. 北京: 机械工业出版社, 2020: 250.

[8] 张亮. Python 深度学习[M]. 北京: 人民邮电出版社, 2018: 170-200.

[9] Kaggle (2022). https://www.kaggle.com

[10] Zhihu. Kaggle 知识点: 类别特征处理[EB/OL]. https://zhuanlan.zhihu.com/p/349592092, 2021.

[11] 张亮. Python 机器学习基础教程[M]. 北京: 人民邮电出版社, 2018: 35-37.

[12] 张亮. Python 机器学习基础教程[M]. 北京: 人民邮电出版社, 2018: 71-80.

[13] 宋能辉, 李娴. 机器学习实战[M]. 北京: 机械工业出版社, 2020: 173.
[14] 宋能辉, 李娴. 机器学习实战[M]. 北京: 机械工业出版社, 2020: 180.
[15] 宋能辉, 李娴. 机器学习实战[M]. 北京: 机械工业出版社, 2020: 182-185.
[16] 宋能辉, 李娴. 机器学习实战[M]. 北京: 机械工业出版社, 2020: 189.
[17] Sklearn (2022). https://scikit-learn.org/stable
[18] Keras (2022). https://keras.io