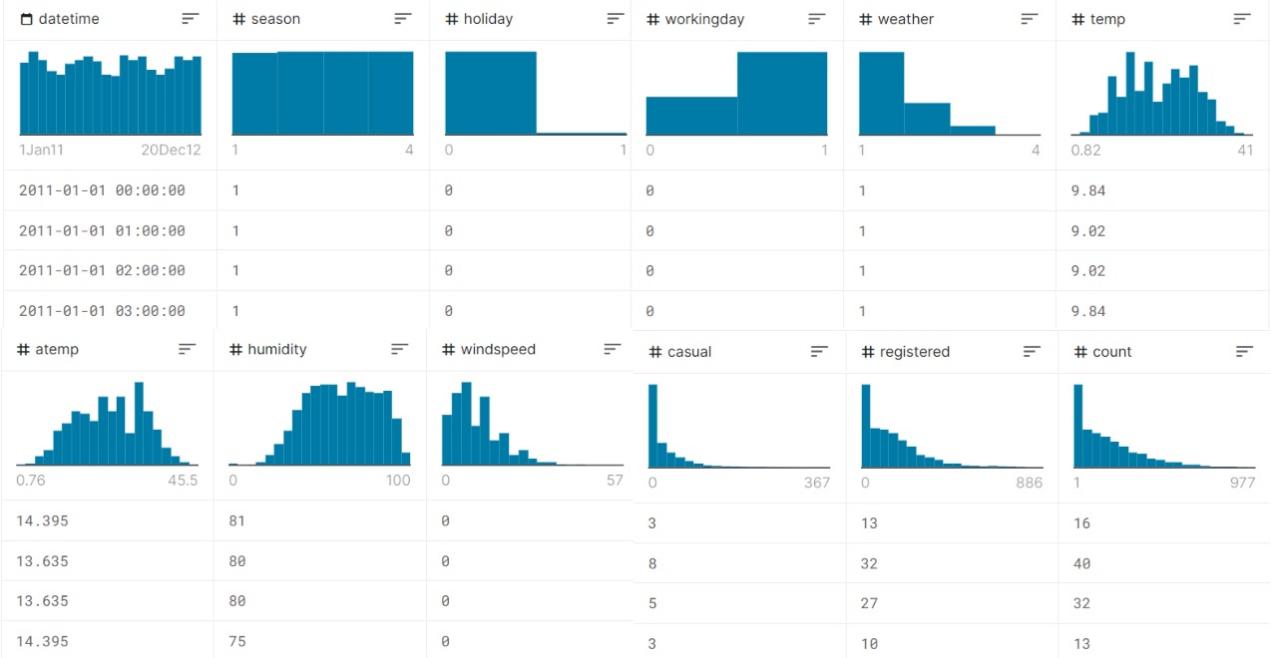
**共享单车需求预测**

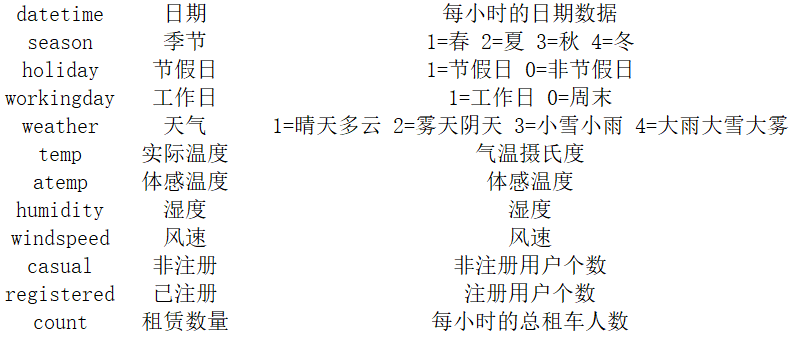
问题背景

随着共享经济的蓬勃发展，共享单车已成为城市交通体系中不可或缺的一部分，其便利性和灵活性深受用户喜爱。然而，共享单车的分布和使用情况受多种因素影响，包括时间（如早晚高峰、节假日）、天气状况、地理位置、季节变化等。因此，准确预测共享单车的需求对于精细化运营至关重要。

1. 问题描述

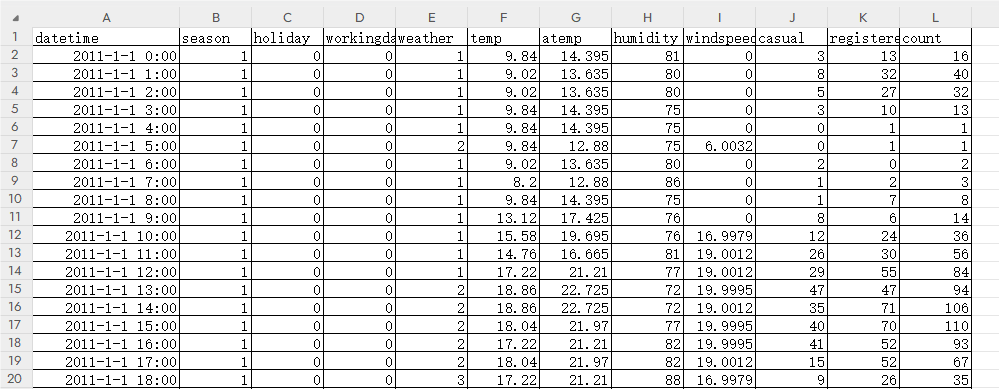
对BikeShare Analysis数据集中的数据进行数据集划分处理，建立模型预测数据集中每个时间点的共享单车需求，以代码输出的测试集的评估指标RMSE为结果。其中train.csv为训练样本，对该样本五折交叉验证得到评估指标。数据分布以及含义如下图所示。





1. 数据预处理

数据以小时为最小单位，给出了2011年-2012年的单车需求数据，数据量足够，适合使用深度学习的方法进行预测。且维度较高，可视为多维非线性回归问题。



由于给出的数据中存在一些自变量间的强相关性和因变量的弱相关性。

1.多重共线性剔除

而对于其他数据，本研究认为体感温度与时间温度存在强相关性，因此取对人的感受更直接的体感温度作为指标，剔除实际温度。

2.数据泄露的特征剔除

而非注册与已注册的人数相加为总人数，实际场景中无法得到，也剔除。

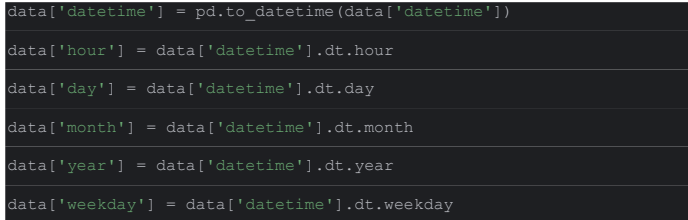
3. 特征工程

3.1时间的处理

因此在使用机器学习方法预测前，需要判断使用哪些数据进行预测，以及如何使用。每小时租车总人数**随着一天内时间变化呈现周期性变化**，这一特征非常显著，几乎不随其他数据变化而改变。所以将日期数据使用datetime函数处理为小时数据。

C:\Users\张浩宇\Documents\WeChat Files\wxid_bhnb9a24v13822\FileStorage\Temp\1718244536064.png

思路扩展：用一天的小时来画折线图，可以反映确实小时呈现周期性，那么如果用月份来画散点图，是否月份也有周期性，那年份呢？甚至进一步挖掘，是否星期几也影响？





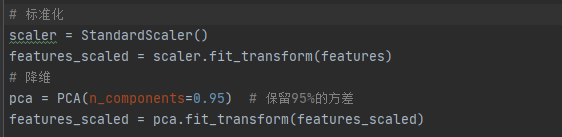
3.2

加快拟合：

1.标准化

2.归一化

3.PCA降维：使数据数量级统一并剔除部分离散点。



数据集划分

1.直接划分：将处理后的数据随机分为4：1的训练集和测试集，用于训练与评估。

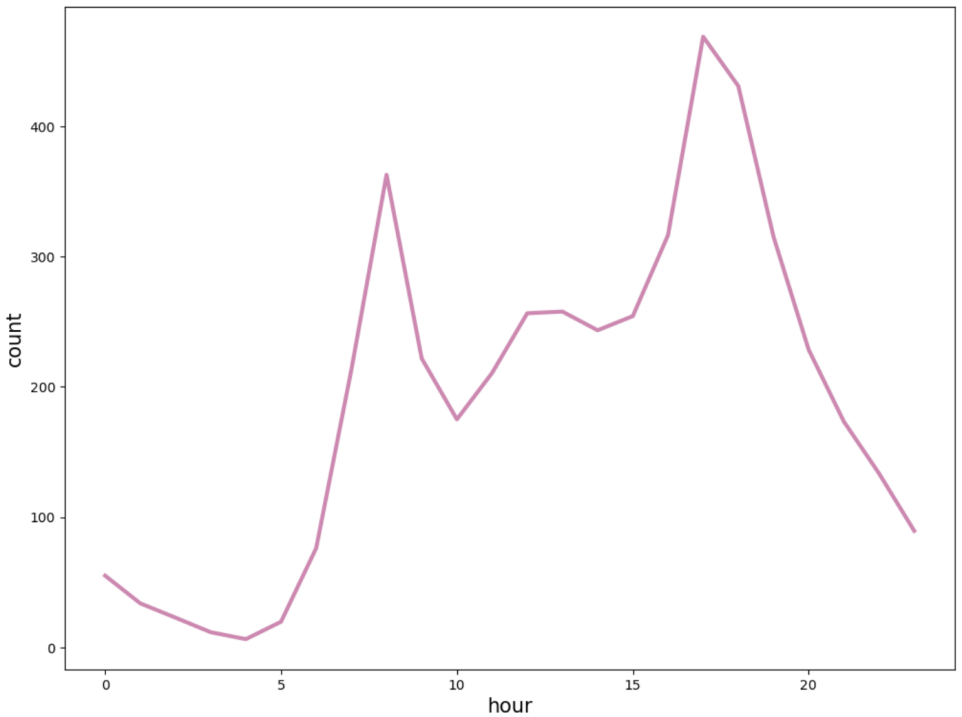
2.交叉验证

误差分析：

该RMSE较大原因在于预测值数量级较大，而RMSE指标对误差非常敏感，部分较大误差的预测使最终求得的RMSE偏大。

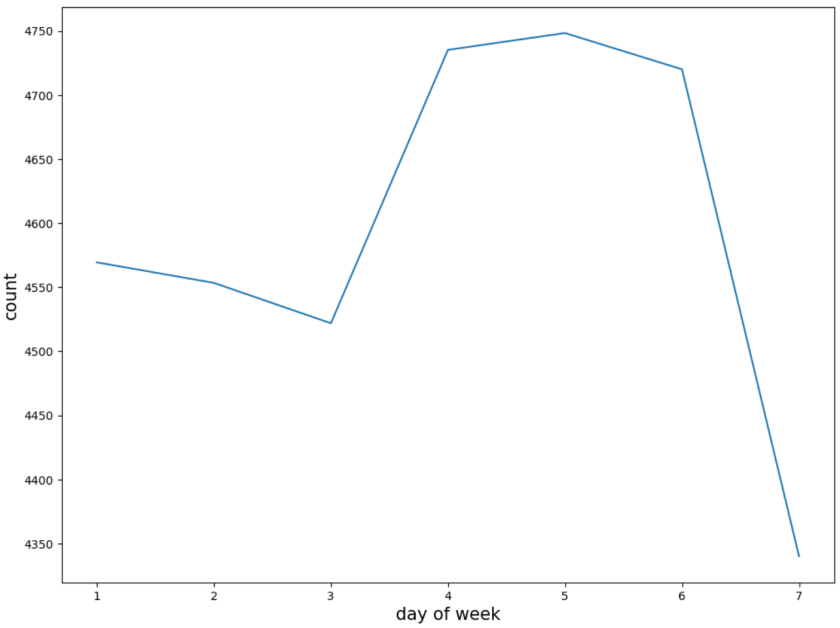
MAPE更好的反应平均偏差

1. **租车数量特征分析**
   1. **日内特征**



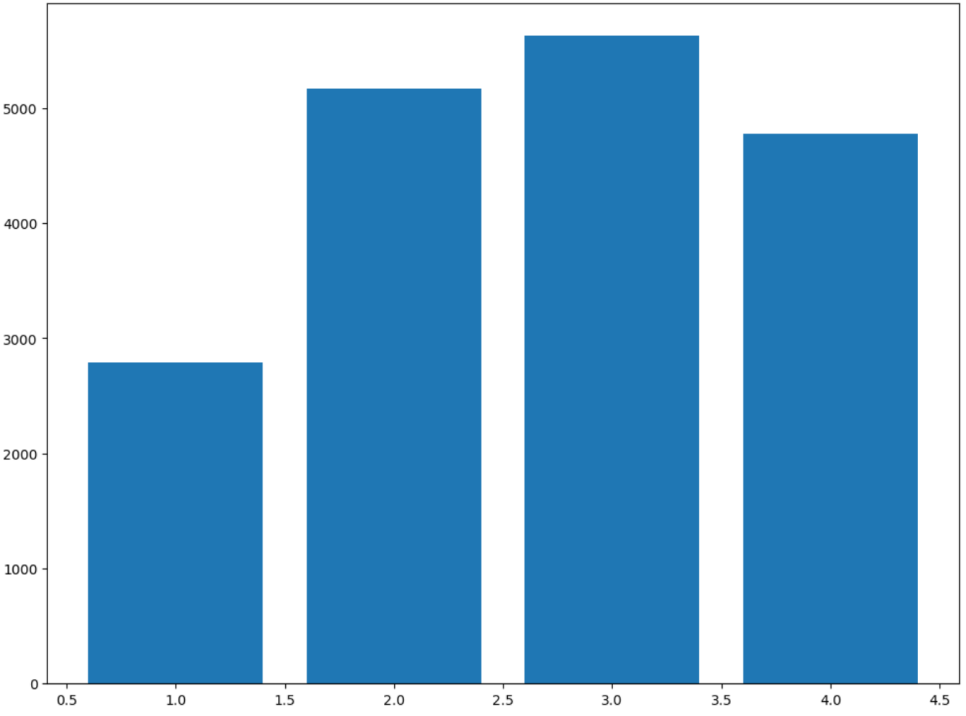
可见，租车量在一天内出现马鞍形分布。

* 1. **周内特征**



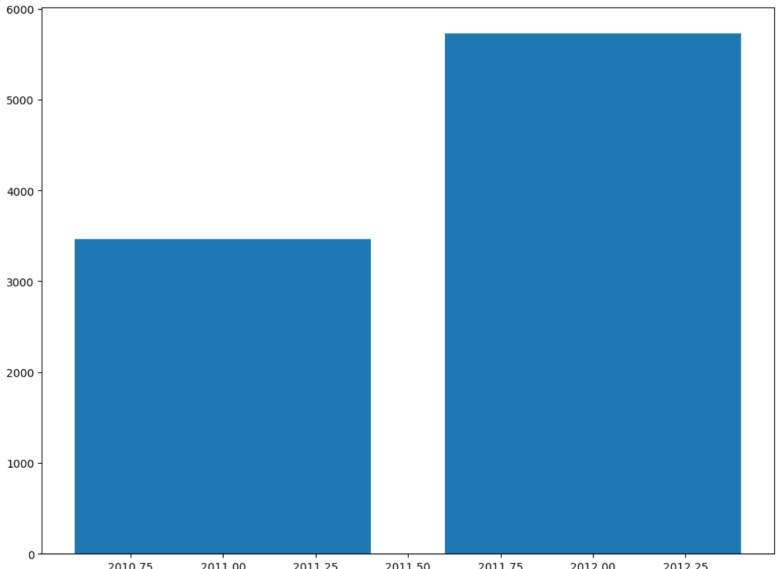
租车量周内特征不显著。

* 1. **季节特征**



横坐标1-4表征春-冬，可见，夏秋两季的租车量最多，冬季次之，春季明显减少。

* 1. **年间特征**



2012年租车量显著增高。

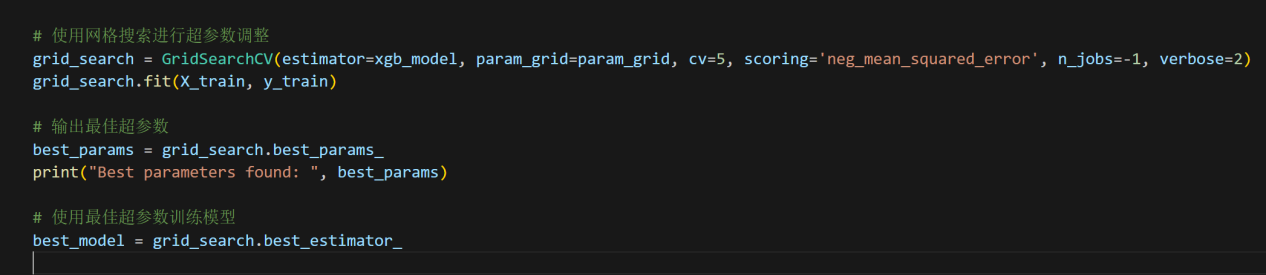
1. **模型建立**

本节尝试建立一些非时间序列机器学习模型和时间序列模型，进行租车数量的预测。

* 1. **非时间序列机器学习模型**
     1. **模型介绍（参数设置）**

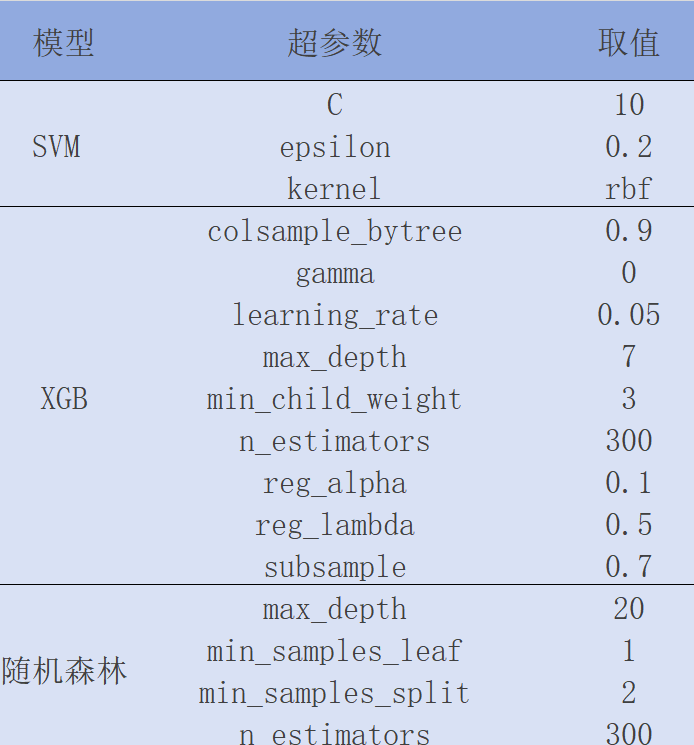
采用了SVM（Support Vector Machine，支持向量机模型）、XGB模型（Extreme Gradient Boosting, 极值梯度提升模型）、RF（Random Forest，随机森林）三种机器学习模型。其中SVM擅长捕捉数据间的线性关系；XGB和RF模型都是基于决策树的集成学习算法，它们擅长处理具有非线性关系和高维特征的数据。

使用Python中的包来实现3个模型。并使用网格搜查法来确定最优超参数，如下图，3个模型的超参数选择如下表：



**图 XGB模型网格搜索确定最佳超参数**

**表 3个模型的超参数取值**



* + 1. **特征生成和选取**

特征选取在机器学习模型中起着至关重要的作用，它有助于提高模型的性能、降低过拟合风险、减少训练时间和提高模型的可解释性。

在本实验中，变量casual及registered本质也是因变量，因此不能作为特征值，去掉。接着，从datetime中提取“year”、“month”、“day”、“dayofweek”（周几）变量。

对除“count”、“casual”和“registered”以外的变量进行相关性分析和单因素检验（共线性检验和单因素分析结果参见表2.1、2.2、2.3）。

以单因素检验结果为主，选择在卡方检验和互信检验中得分均较高的5个变量作为最终用于预测模型的特征变量，选择详情如表2.4.

**表2.1 相关性分析**

|  |  |
| --- | --- |
| **features** | **Corr\_value** |
| hour | 0.400601 |
| temp | 0.394454 |
| atemp | 0.389784 |
| year | 0.260403 |
| month | 0.166862 |
| season | 0.163439 |
| windspeed | 0.101369 |
| day | 0.019826 |
| workingday | 0.011594 |
| dayofweek | -0.002283 |
| holiday | -0.005393 |
| weather | -0.128655 |
| humidity | -0.317371 |

**\*corr\_value表征了各变量与目标值count的相关性，取值为[-1,1],绝对值越大表征与目标值相关性越高**

**表2.2 单因素卡方检验**

|  |  |
| --- | --- |
| **features** | **Chi2 Score** |
| hour | 18120.477372 |
| humidity | 12286.067545 |
| temp | 8174.188620 |
| atemp | 8152.501468 |
| windspeed | 5282.756907 |
| day | 2429.088872 |
| month | 2189.557845 |
| dayofweek | 940.826815 |
| holiday | 681.355480 |
| season | 610.371058 |
| workingday | 320.135759 |
| weather | 234.265530 |
| year | 0.181543 |

**\*Chi2 Score卡方统计量，用于衡量每个特征与目标变量之间的关系。得分越高，表示该特征与目标变量之间的关系越强。**

**表2.3 单因素互信息得分检验**

|  |  |
| --- | --- |
| **features** | **MI Score** |
| hour | 0.642967 |
| temp | 0.143285 |
| atemp | 0.141705 |
| humidity | 0.090507 |
| month | 0.078178 |
| year | 0.061522 |
| season | 0.061412 |
| weather | 0.022334 |
| windspeed | 0.013865 |
| dayofweek | 0.012927 |
| workingday | 0.008644 |
| holiday | 0.000000 |
| day | 0.000000 |

**\*MI Score互信得分，每个特征与目标变量之间的依赖关系程度。得分越高，表示特征与目标变量之间的信息共享越多，关系越强。**

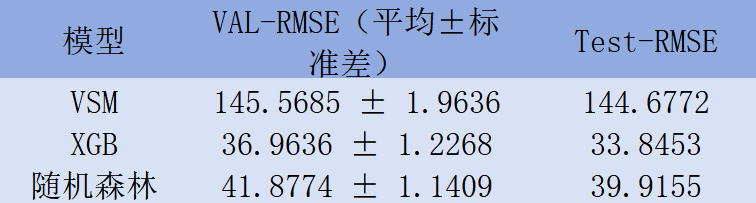
**表2.4 特征变量最终选择情况**

|  |
| --- |
| **features** |
| hour |
| temp |
| atemp |
| humidity |
| month |

* + 1. **结果分析**

三个模型各运行10次，取10次RMSE的平均值。三个模型的训练集+验证集、测试集的抽取比例为8：2，采用五折交叉法进行验证。三个模型预测的指标如下所示：

**表 各模型预测结果**



\*VAL-RMSE：验证集平均均方根误差；Test-RMSE:测试集平均均方根误差

总体来看，XGB和随机森林模型的表现远优于VSM模型。这说明相较于线性模型，该数据集更适合擅长处理非线性关系的模型。3个模型的验证集平均均方根误差和测试集平均均方根误差基本相等，说明3个模型的泛化性都不错。

* 1. **时间序列模型**

该数据集中有时间变量“datatime”，某时间点的数据可能受到历史时间点数据的影响，因此应该考虑采用时间序列模型对该数据进行预测**。**

* + 1. 数据时间序列特征分析

时间序列数据有以下几个显著特点：

1. 时间依赖性

时间序列数据的最大特点是数据点之间存在时间上的依赖关系。某一时刻的数据点往往受到之前时间点数据的影响。

2. 自相关性

时间序列数据通常表现出自相关性，即同一序列中相邻数据点之间存在相关性。

3. 趋势（Trend）

时间序列数据可能包含长期的上升或下降趋势。趋势是数据随时间推移而发生的缓慢变化，可以是线性、非线性或阶跃型的。

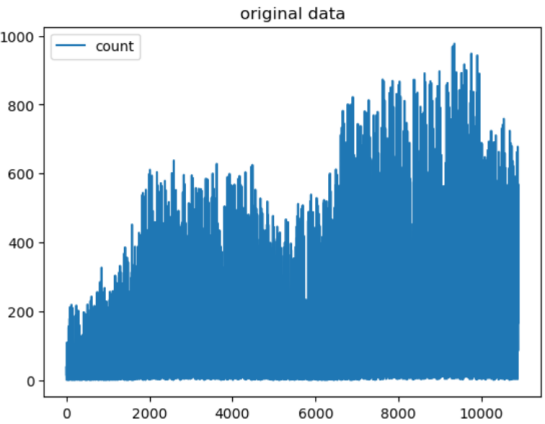
4. 季节性（Seasonality）

时间序列数据可能存在季节性波动，即数据会周期性地重复。季节性可以是日、周、月或年等周期性的变化。

分析时间序列数据的以上特点有助于我们在建模阶段选择更加合适的模型和超参数。

1. 季节性

将数据按照时间序列进行排列，以数据的次序数为横坐标，以目标值count为纵坐标绘制图像，如下图：



可见，该数据集有一定的季节性。

1. 趋势性

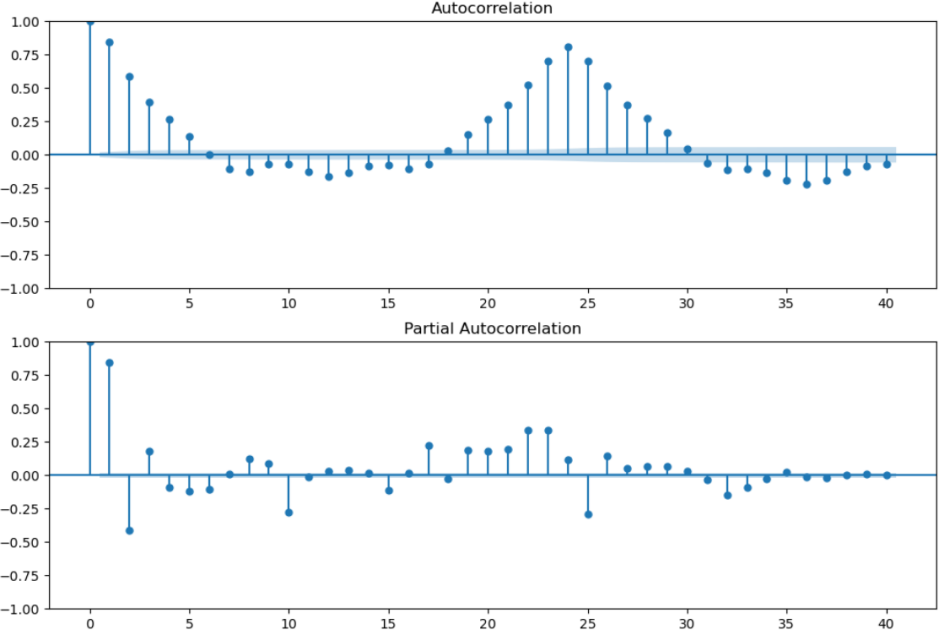
同样，从图中可以看出在同一年内，随着数据条数增加、时间推移，count值以先上升后下降的趋势，因此，有一定的趋势性。

1. 自相关性

自相关性可以通过自相关函数（ACF）和偏自相关函数（PACF）来检测和分析。

绘制自相关函数和偏自相关函数图，横坐标为之后步长，纵坐标为自相关系数和偏自相关系数。如下所示，在自相关和偏自相关图中可以看到多个滞后步长的自相关系数显著不为零，说明其有较强的自相关性。

**图 自相关函数和偏自相关函数图**



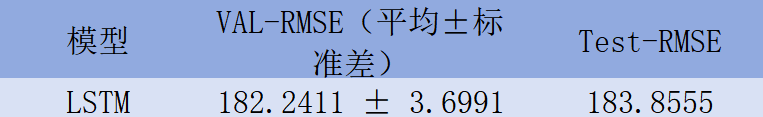
综上分析，该数据集有比较明显的时间依赖性，应考虑时间序列模型。

* + 1. **LSTM模型**

LSTM（长短期记忆网络）应用： LSTM是一种特殊的循环神经网络（RNN），特别适合处理具有时间序列特性的数据，如共享单车的使用量随时间的变化。LSTM通过其门控机制能够有效地学习长期依赖关系，自动捕获时间序列中的周期性、趋势和其他动态模式，无需大量手动特征工程。

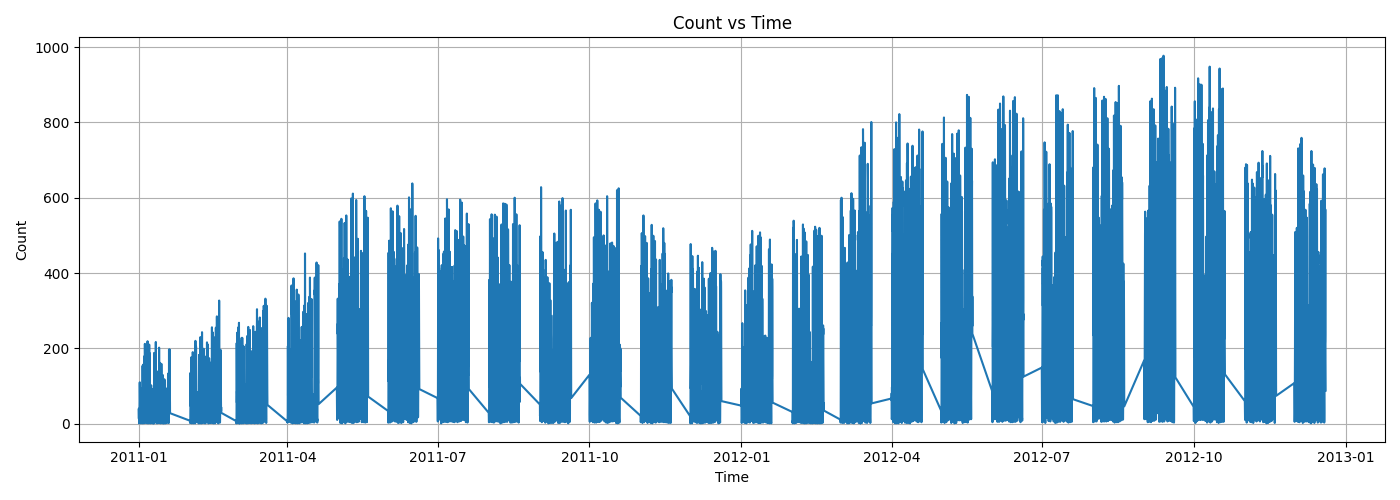
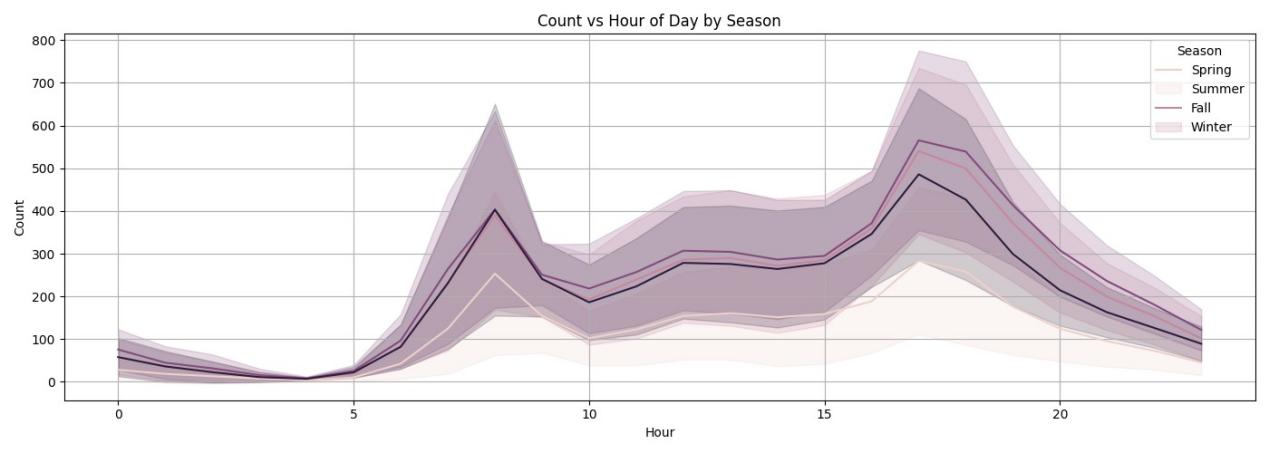
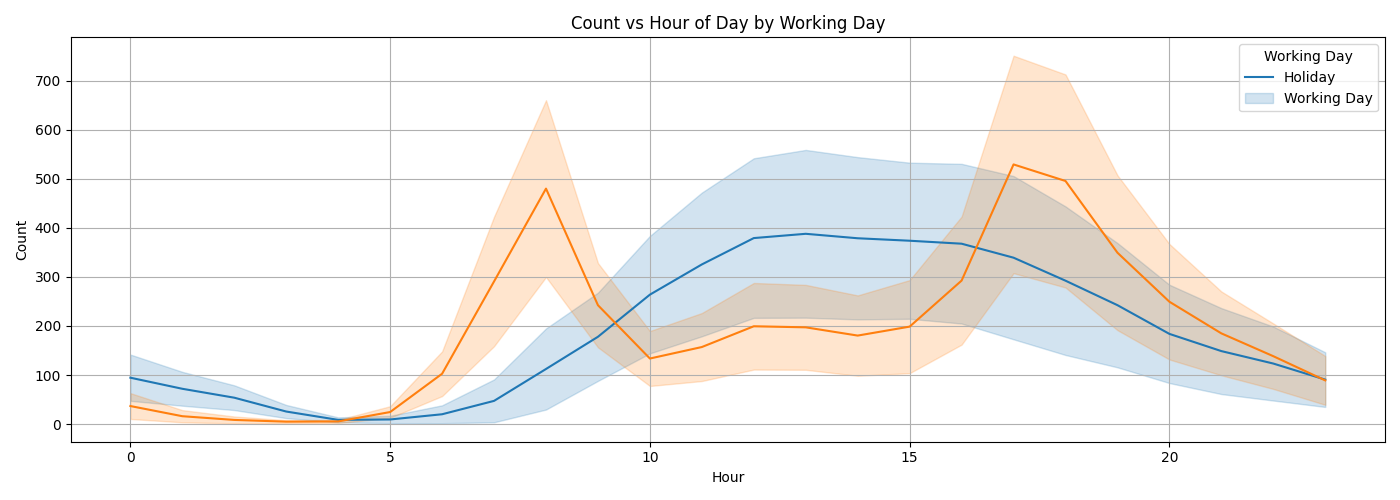
选择该模型对数据集进行预测，同样的，特征变量选取同非时间序列模型，模型运行10次，取10次RMSE的平均值。模型的训练集+验证集、测试集的抽取比例为8：2，采用五折交叉法进行验证。模型预测的指标如下所示：

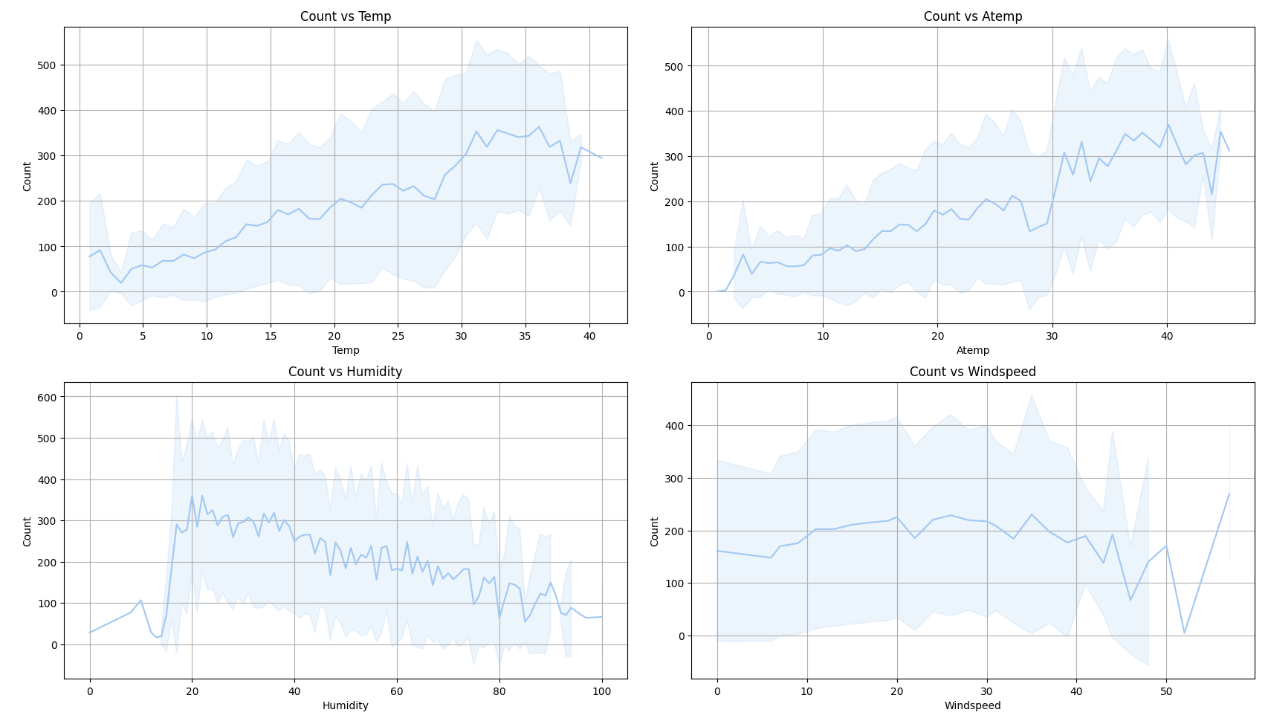
**表 LSTM模型预测结果**

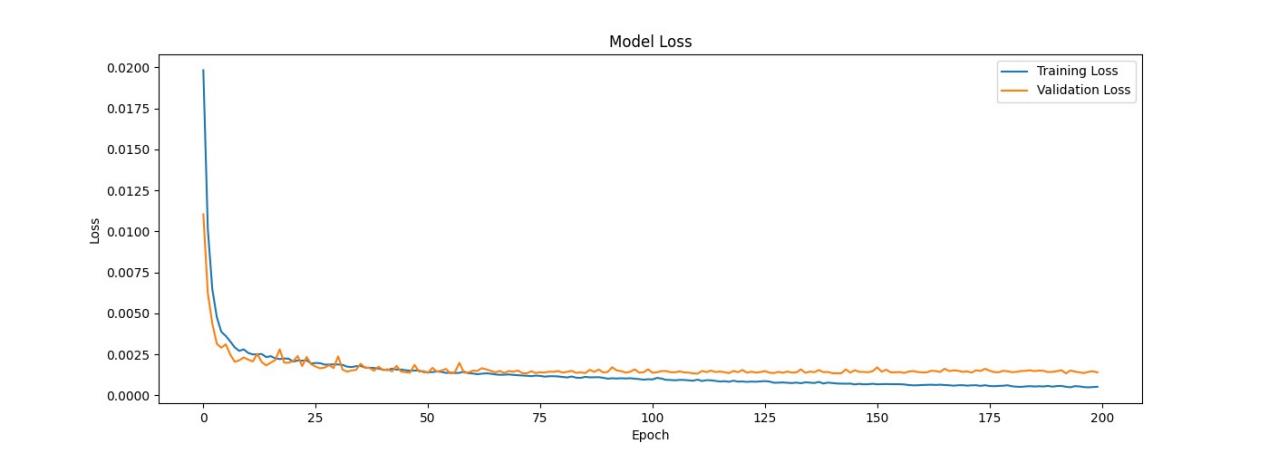


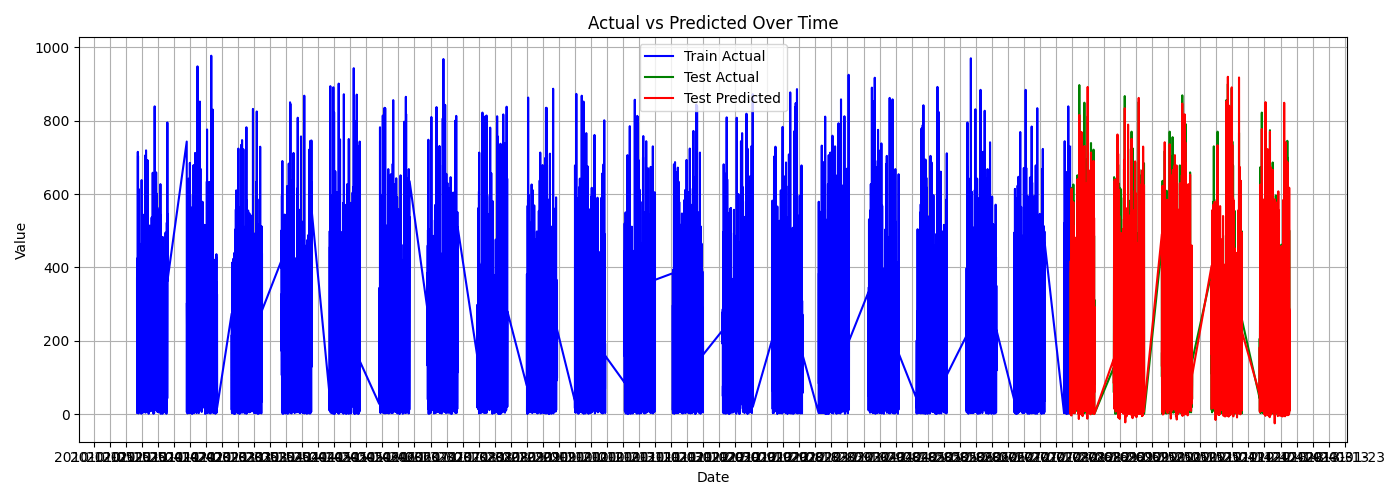
令我们意外的是，虽然在数据特征分析中我们得出该数据有较强的时间依赖性的特点，但是适用与捕捉数据时间依赖性特征的LSTM模型的预测结果却是最差的。因为代码实现困难的问题，我未在此模型中使用网格搜索法确定最佳超参数，这可能是该模型表现不佳的原因。也可能是数据中存在过大噪声，LSTM模型可能会受到影响，无法准确捕捉到有效的时间依赖关系。

# 可视化分析







综上所述，LSTM-传统训练集划分方式在共享单车预测方面显示出更好的性能，相较MLP高了约50%，相较时序交叉验证的方式高了30%。

说法错误：时序只是为了公平

## GRU

GRU是一种改进的循环神经网络（RNN），能够更好地捕捉时间序列数据中的长期依赖关系。相比于传统的RNN，GRU通过引入更新门和重置门，能够有效地解决长时间序列训练中的梯度消失问题。GRU的核心思想是通过引入两个门（更新门和重置门）来控制信息的流动。这些门帮助GRU选择性地更新和重置隐藏状态，从而更有效地捕捉和保留重要的时间序列信息。

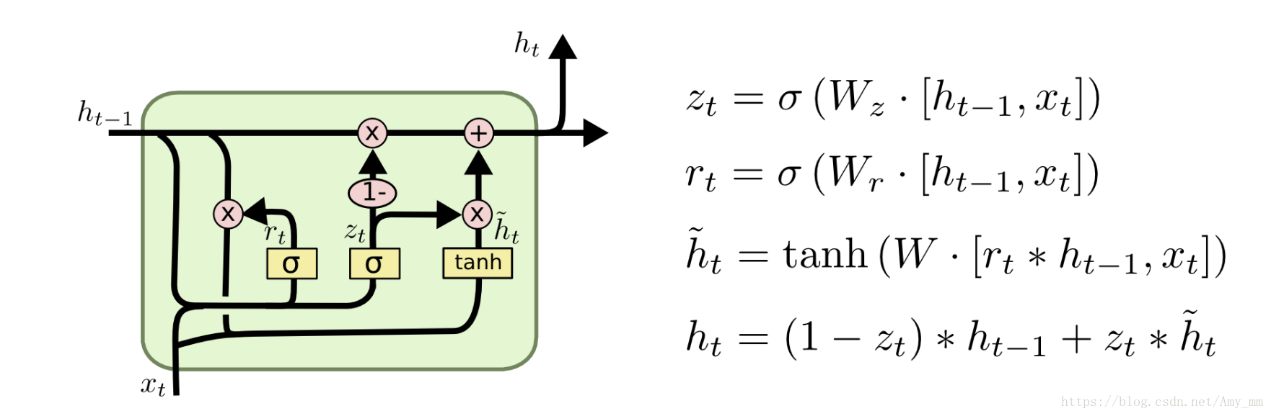


图2 门控循环单元

更新门决定了当前时间步的隐藏状态有多少部分需要从前一时间步的隐藏状态中保留，多少部分需要从当前输入中获取。更新门的计算公式为：

其中，是权重矩阵，​ 是前一时间步的隐藏状态，​是当前时间步的输入，是Sigmoid激活函数。

重置门决定了当前时间步的隐藏状态有多少部分需要被重置。重置门的计算公式为：

其中， 是权重矩阵。

结合当前输入和前一时间步的隐藏状态，计算候选隐藏状态。候选隐藏状态的计算公式为：

其中， 是权重矩阵。

最终的隐藏状态是前一时间步的隐藏状态和候选隐藏状态的加权和。更新公式为：

其中，是当前时间步的隐藏状态。

本模型结构包括了GRU层和输出层两个神经层，其中GRU层用于提取时间序列中的特征，捕捉数据的时间依赖关系，输出层用于将GRU的输出转换为预测的租赁需求人数。

训练过程

在训练过程中，输入特征为9，GRU的隐藏特征为64，时间步长为24，预测的时间步为1，训练后的模型保存在best200.bth文件中。

评估指标：RMSE均方根误差，用于衡量预测值与实际值之间的平均偏差，其计算公式为：

其中，为实际值，为预测值，N为样本数量。

训练过程中训练损失的变化情况如下图所示，预测结果对比图如下所示，测试集的RMSE为31，效果较好。

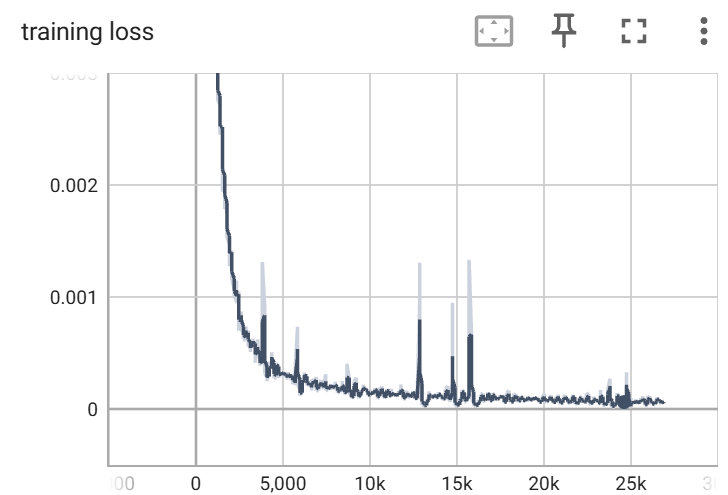


图5 训练损失变化图

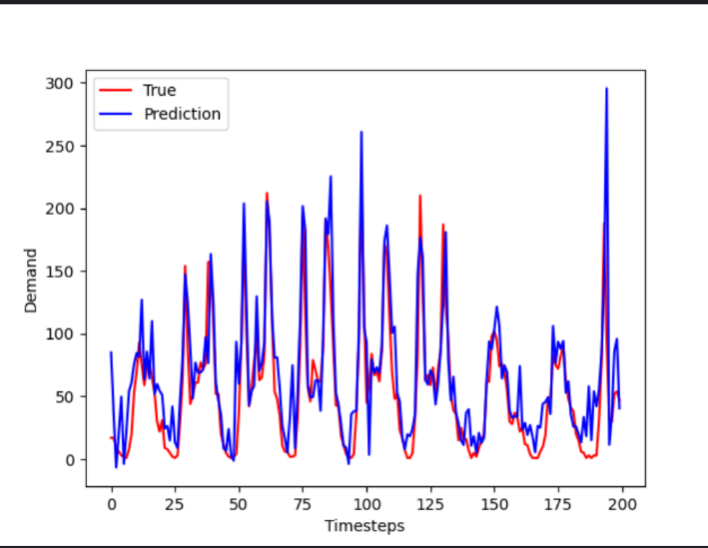


图4 预测结果对比图