**Kaggle共享****单车租赁预测**

熊亮1，查兆越1，朱正天1

（1. 北京邮电大学通信与信息工程学院，北京 ~~400065）~~

摘 要：为了解决共享单车投放数量，以最大化资源利用，本文对数据预处理，研究了时段、温度、湿度、年份、月份、季节、天气、日期对自行车租赁的影响。本文主要使用随机森林算法，预测了不同条件下单车的投放数量。最后比较了决策树算法与随机森林算法性能，并删除了季节因素，然后进行改进预测，证明了改进后随机森林算法性能更好。

关键词：数据预处理；随机森林算法；单车投放数量预测

中图分类号：TN92

文献标识码：A

**doi:**

**Shared Bicycle Rental Forecast in the Kaggle**

XIONG Liang1, CHA Zhaoyue1, ZHU Zhengtian1

1. School of Communication and Information Engineering, Beijing University of Posts and Telecommunications, Beijing ~~400065~~, China

**Abstract:** In order to solve the number of shared bicycles and maximize the utilization of resources, this paper preprocesses the data and studies the influence of time period, temperature, humidity, year, month, season, weather, and date on bicycle rental. This article mainly uses the random forest algorithm to predict the number of bicycles placed under different conditions. Finally, the performance of the decision tree algorithm and the random forest algorithm is compared, and the seasonal factor is deleted, and then improved prediction is performed, which proves that the improved random forest algorithm performs better.

**Key words:** data preprocessing; random forest algorithm; prediction of the number of bikes

收稿日期：2019−08−16；修回日期：2019−12−17

基金项目：国家自然科学基金资助项目（No.61601071, No.61671091）

**Foundation Items:** The National Natural Science Foundation of China (No.61601071, No.61671091), The Science and Technology Research Program of Chongqing Municipal Education Commission (No.KJQN201800606, No.KJZD-K201900605), The Open Research Fund From Shandong Provincial Key Laboratory of Wireless Communication Technologies, Shandong University (No.SDKLWCT-2019-04), Chongqing Science and Technology Innovation Leading Talent Support Program (No.CSTCCXLJRC201908), The Natural Science Foundation of Chongqing (No.cstc2019jcyj-xfkxZ0002)

## 引言

随着互联网时代的到来，人们的出行更加方便，共享单车成为人们生活中不可或缺的一部分。但是单车的资源是有限的，我们不可能把它无限地把他们放到各个角落。为了让单车的利用率得到最大，本文设计了这样的模型和算法，需要根据天气，风速，湿度，温度，时段等信息来预测华盛顿的自行车租赁需求。

本文主要的研究工作如下。

1) 预处理数据，通过对数据是否有缺失值、重复值的判断，以及通过箱线图查看、检查异常值，本文最后按照转换"时间和日期"的格式, 提取出小时, 日, 月, 年，完成数据加工。

2) 完成数据分析与可视化，除了分析各个变量与租赁数量之间的关系之外，本文还观察了各个变量随时间变换的变化趋势。先做了整体观察，然后分别研究了时段、温度、湿度、年份、月份、季节、天气、日期对自行车租赁的影响。

3) 建模与调参。分别使用随机森林进行回归预测，先将多类别型数据进行分类，然后设定训练集，生成模型进行拟合，最后完成对测试集的数据预测。此外，本文还使用了决策树对最终性能做了对比。仿真结果表明，本文使用随机森林算法具有很好的性能，预测效果得到很好体现，验证了本文模型的有效性。

## 问题定义及解决方案

根据天气，风速，湿度，温度，时段等信息来预测华盛顿的自行车租赁需求。

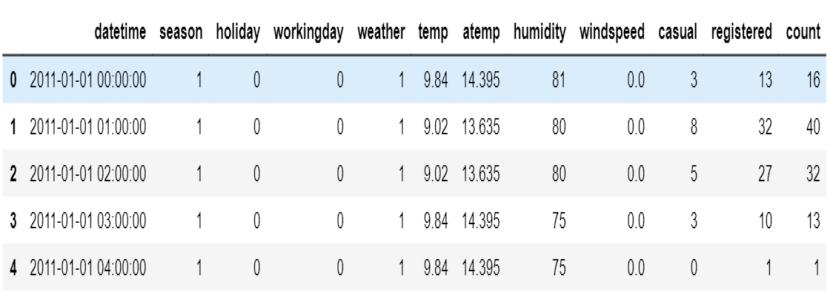
本文提供的解决方案是：先对原始数据进行预处理，包括缺失值、重复值、异常值处理，并进行数据转换，使得数据表现更加直观。然后，本文对各项数据指标进行可视化，并联合分析，确定需要挖掘的数据。最后，本文分别使用决策树、随机森林对单车租赁数量预测，进行性能对比、分析，完成本次挖掘任务。

## 数据预处理

训练集中提供了11，12两年中，每月的前19天的每小时自行车租赁数据，以及当时的时段，天气，风速，温度等信息。测试集则提供了11，12两年每月20号至月底的每小时时段，风速，天气，温度等信息，本文来预测租赁数据。下面是数据预处理的流程。

### 3.1 数据概览

首先，读取数据，读取train.csv和test.csv两个文件，即训练集和测试集一起读取。



这里字段的解释是：

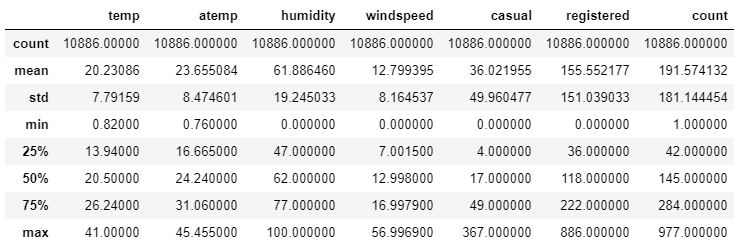
* season：季度
  + 1：春
  + 2：夏
  + 3：秋
  + 4：冬
* holiday：当天是否是节假日
* working：当天是否是工作日
* weather：天气等级
  + 1：晴/多云
  + 2：阴天
  + 3：小雨/小雪
  + 4：大雨/大雪
* temp：温度
* atemp：体感温度
* humidity：湿度
* windspeed：风速
* casual：未注册用户租赁数量
* registered：组测用户租赁数量
* count：总的租赁数量

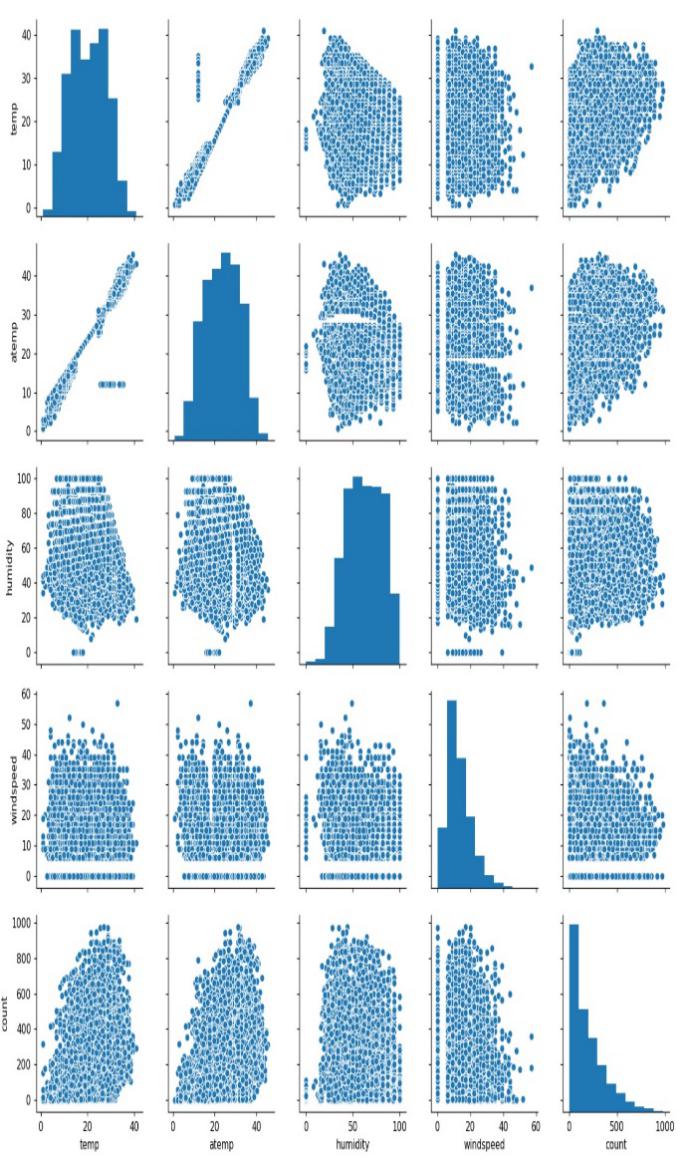
### 3.2 缺失值/重复值处理

在处理数据之前，本文首先将datetime列进行拆分，方便后续的处理与分析。首先是训练集datetime列处理，分别将属性data、year、month、year和weekday进行映射map(lambda x:parse(x).属性)。其次，以同样的方式对测试集的datetime列处理.第三步，检查训练集有无缺失数据。第四步，检查训练集有无重复数据，并得到如下结果：False 10886，dtype: int64。可以看到，训练集和测试集都没有缺失值和重复值。

### 3.3 异常值处理

没有缺失值和重复值不代表数据集就是没有异常，首先，本文先来看下训练集中的数值型数据情况，可以得到如下结果：

从上面这个表格中，本文可以发现count列的数值差异很大，最大值和最小值相差很多。为了进一步观察，绘制散点图矩阵如下：



通过散点图矩阵本文可以很明显的看到一些异常：

1.atemp和temp的散点图中，有小部分数据的atemp值相等，但却明显游离在atemp和temp的回归方程之外（通过jupyter的坐标显示，此时的temp-atemp>10）

2.atemp部分值缺失，导致散点图出现裂隙

3.humidity有小部分数据值为0

4.windspeed有大量的数据值为0

5.count数据偏斜很严重，导致直方图尾部特别长

对应的解决方案如下：

1.atemp游离的异常数据量较少，直接删掉

2.atemp部分值缺失导致散点图裂隙，这个暂时还想到怎么处理

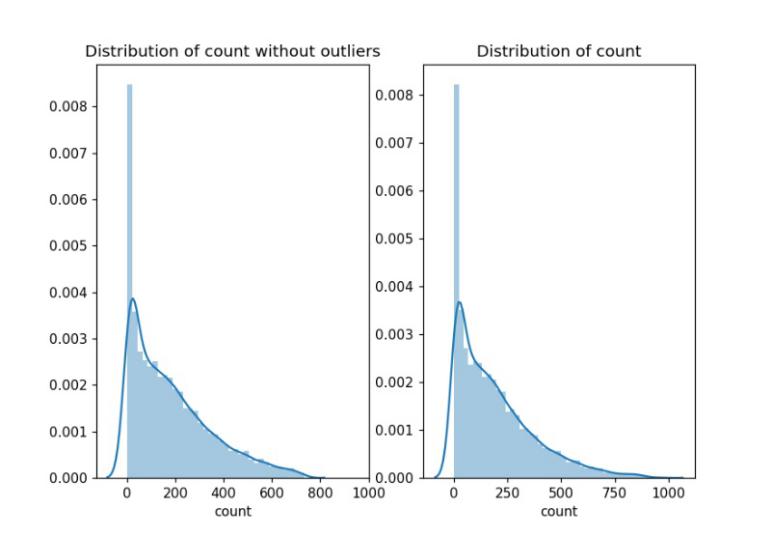
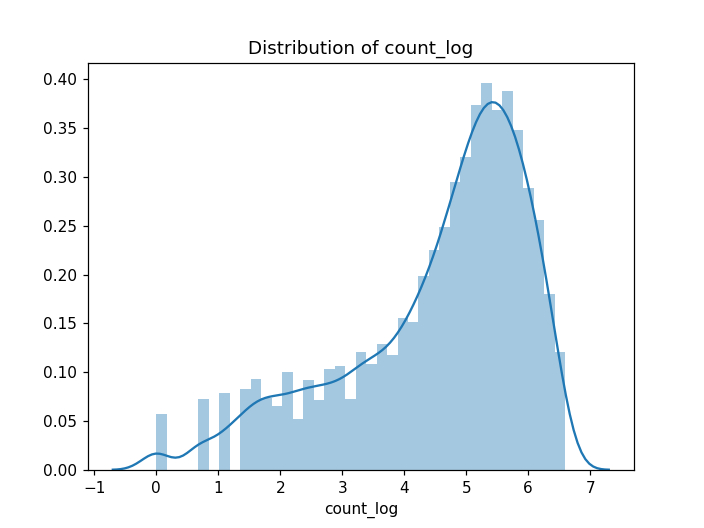
3.humidity值为0的据量较小，直接删掉

4.windspeed值为0的数据量较大，使用随机森林进行回补

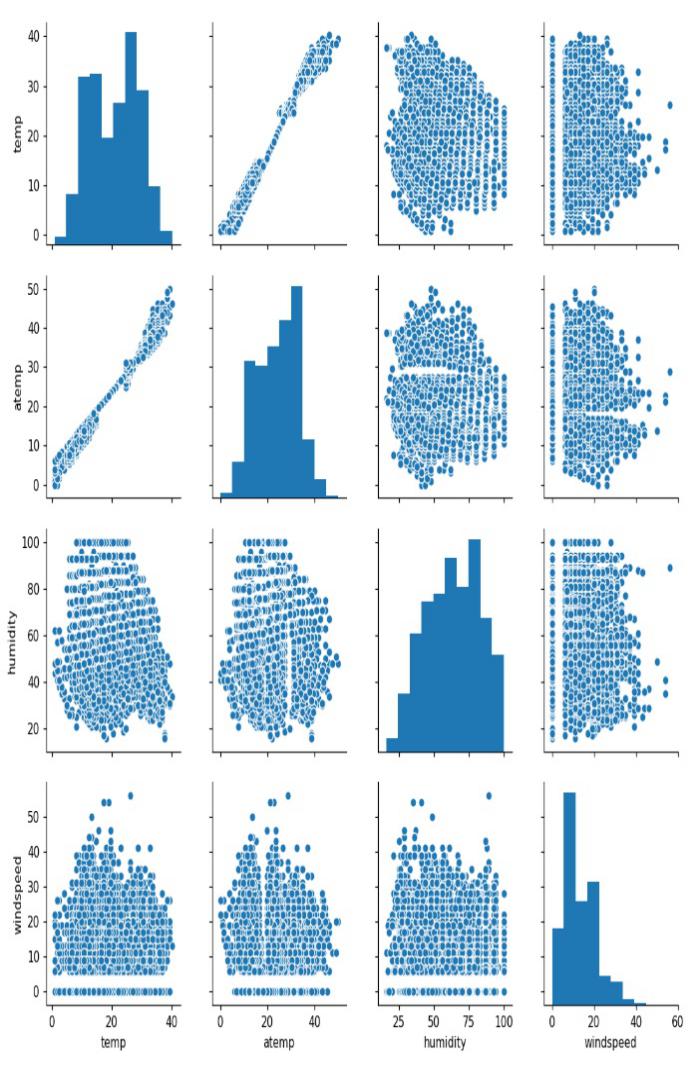
5.count列可以先剔除掉一些异常值看看

第二步就是对atemp列处理。查看异常数据的个数可知道atemp列异常数据的量只有24个，数据量不大，可以直接剔除。

第三步是对humidity列处可以得到humidity列值为0的数据量共22个，也是可以直接剔除。

第四步是对count列处理count列的问题在于数据偏斜很严重，count列分布的尾部很长，波动很大，这样容易造成模型过度拟合，所以本文先将一些极端值剔除掉看看，一般本文认为3个标准差以外的数据属于粗大误差，应该被剔除掉。 剔除掉异常值之后数据偏斜依旧很严重，所以这里本文需要将count列进行对数转换，减小波动。  
经过对数转换后数据波动更小了，而且大小差异也变小了。

第五步是异常值填充，windspeed列处理。这里本文使用随机森林对训练集中windspeed值为0的数据进行回补。将训练集数据集分为风速值为0和不为0两部分。设定训练集数据。再生成回归器并进行拟合。预测风速值最后将预测值填充到风速为0的数据中，合并。



测试集中atemp列并没有出现明显的异常数据，但是仍有值的缺失导致散点图出现裂隙（这个暂时还不知道怎么处理），humidity列也没有出现值为0的异常数据。但是windspeed列仍有大量值为0的异常数据，和训练集一样，本文使用随机森林对测试集windspeed值为0的数据进行回补。

### 3.4 数据转换

解决数据偏斜，使用对数变换。数据偏斜依旧很严重，所以这里需要本文将count列进行对数转换，减小波动。对类别型特征进行one-hot编码。

## 挖掘结果分析

本文对数据进行可视化操作，以方便分析各个变量与租赁数量之间的关系，并且可以观察各个变量随时间变换的变化趋势。

本文先进行总体概览，将训练集和测试集进行合并然后再进行分析

### 4.1 总体概览

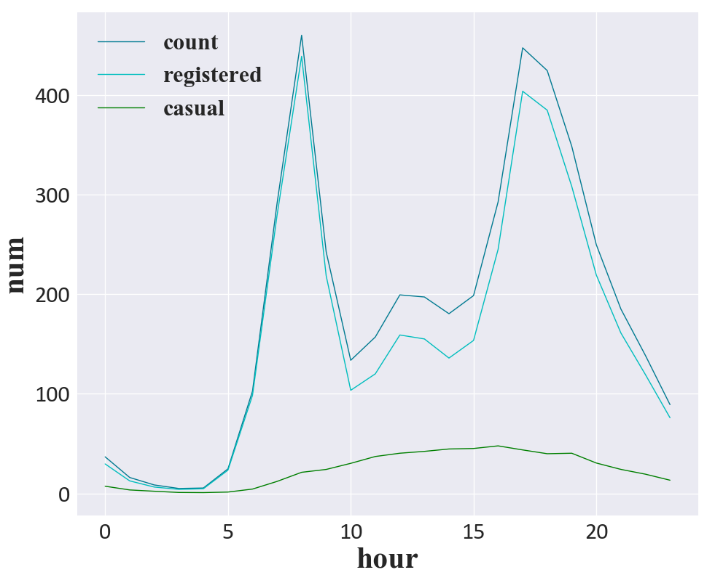
由附录图4-1得

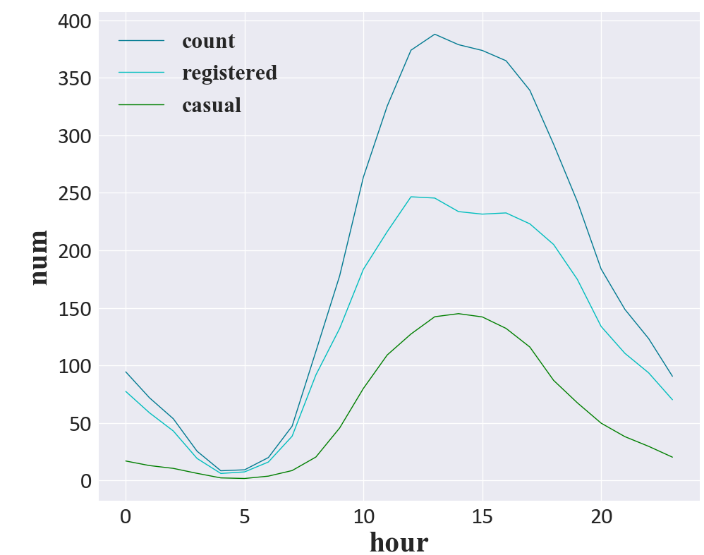
得到很多信息了：

* 非节假日的用车里量多于节假日的用车量，当然这和非节假日的天数要多于节假日的天数有关
* 注册用户在工作日的用车量要多于周末的用车量，而非注册用户则恰好相反
* 天气等级越高，用车量越少
* 注册用户一天的用车高峰时上下班期间，另外中午也有个小高峰
* 非注册用户一天的用车分布基本呈正态分布，用车高峰在13点左右
* 1月和2月时一年中用车量最少的两个月
* 12年和11年相比，不管是注册用户还是非注册用户，用车量都有所增长
* 温度和体感温度对非注册用户影响较大，对注册用户影响较小
* 风速越高，用车量越少

为了更进一步理解数据，同时也是起到学习的作用，下面本文再一个一个地分析各个变量与自行车租赁之间的关系。

### 4.2 时段对自行车租赁影响

本文将数据分为工作日和非工作日。

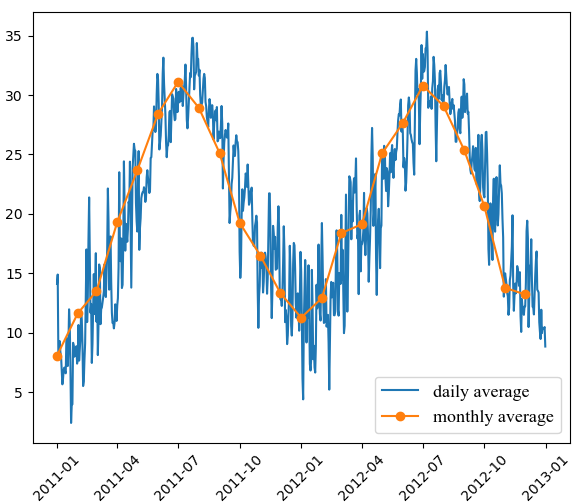


从上图来看，在工作日时，注册用户的用车主要集中在早8点和下午17点这两个上下班高峰期，另外中午12点时也是一个小高峰，可能是中午出去吃饭。而非注册用户在工作日的用车则是呈现一个正态分布趋势，在下午14点左右用车量最多。

而在非工作日时，注册用户和非注册用户都是在下午13点左右时用车最多，在凌晨4点时用车最少。

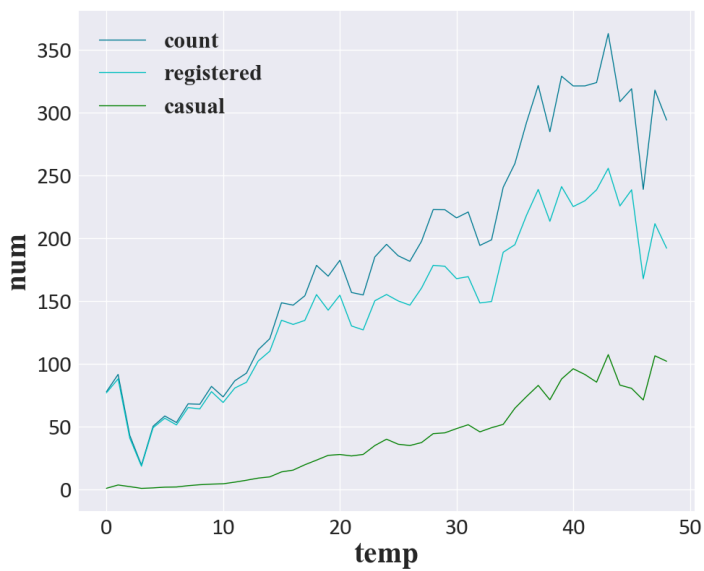
### 4.3 温度对自行车租赁的影响

温度随时间的变化趋势,将每天以及每月的平均气温绘制出来。



温度呈周期性变化，在每年的7月平均温度最高，而1月则是全年温度最低的月份。

那么温度和单车租赁数量之间是什么关系呢？本文将按温度进行聚合，得到每个温度下的平均租赁数量来进行观察。



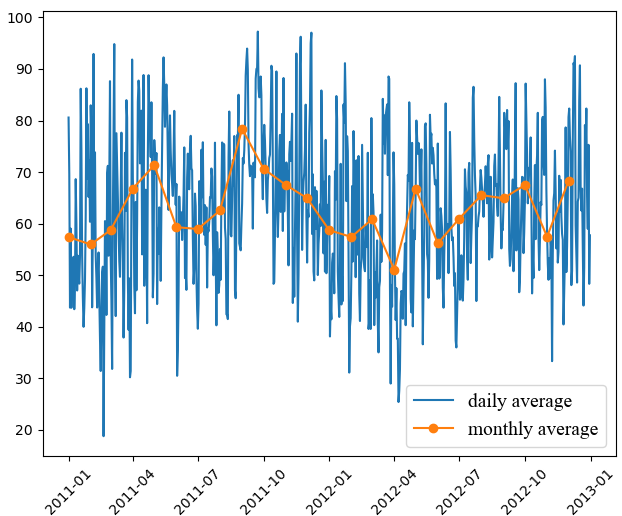
随着湿度增加，单车使用量持续增加，温度在35，自行车租赁使用量最大。

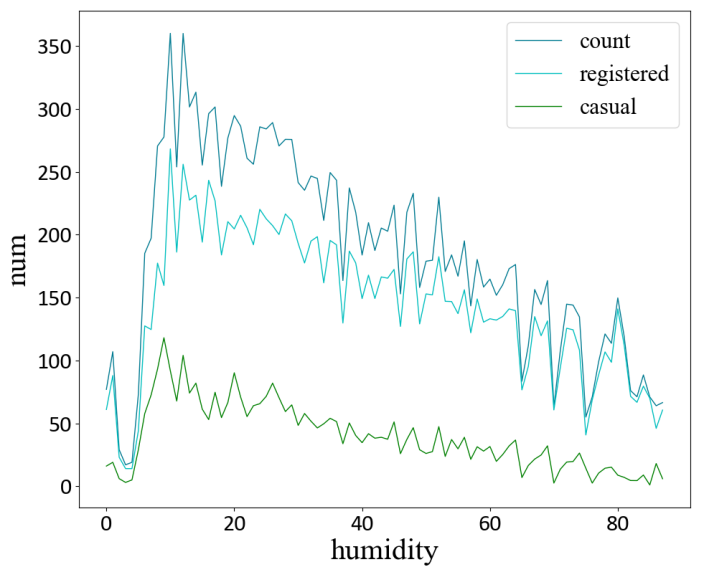
atemp和temp的相关系数高达0.99，所以这里本文就不研究atemp的变化趋势。它和count之间的关系了，基本都一样。

### 4.4 湿度对自行车租赁的影响

同样的，本文研究了湿度的全年变化趋势，将每天以及每月的平均湿度绘制出来。

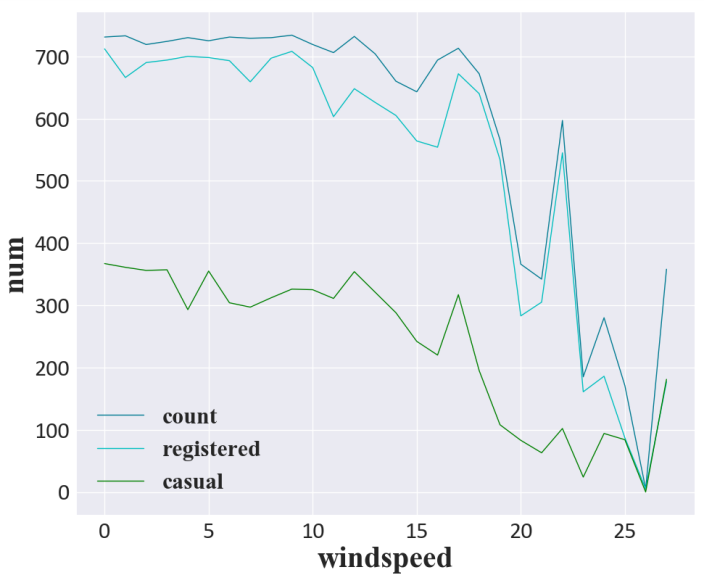
湿度没有像温度一样呈一个周期性变化，而且湿度的波动很大，同一个月内可能很干燥，也可能特别湿润。

下面来看下湿度对自行车租赁的影响。同样的，本文将湿度进行聚合，得到每个湿度下的平均租赁数量来进行观察。



湿度在20的时候，自行车租赁使用量最大，之后随着湿度增加，单车使用量持续减小。

### 4.5 风速对自行车租赁的影响



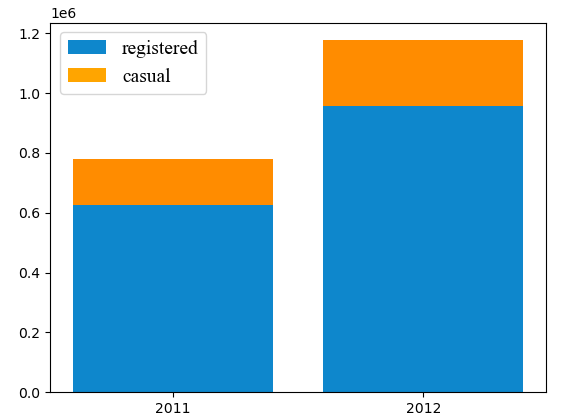
风速的波动非常大，全年并没有一个很明显的变化趋势，似乎1、2、3、4这四个月再全年是风速最高的月份，再往后走，风速都是低于这4月的。

考虑到风速特别大的时候很少，分析风速与用车之间关系时，如果取平均值可能会出现异常，所以按风速对租赁数量取最大值。

从上图来看，随着风速的增加，自行车租赁数量时逐渐下降的，但是在风速45和55左右出现了两次大的反弹，猜测可能是异常数据导致。

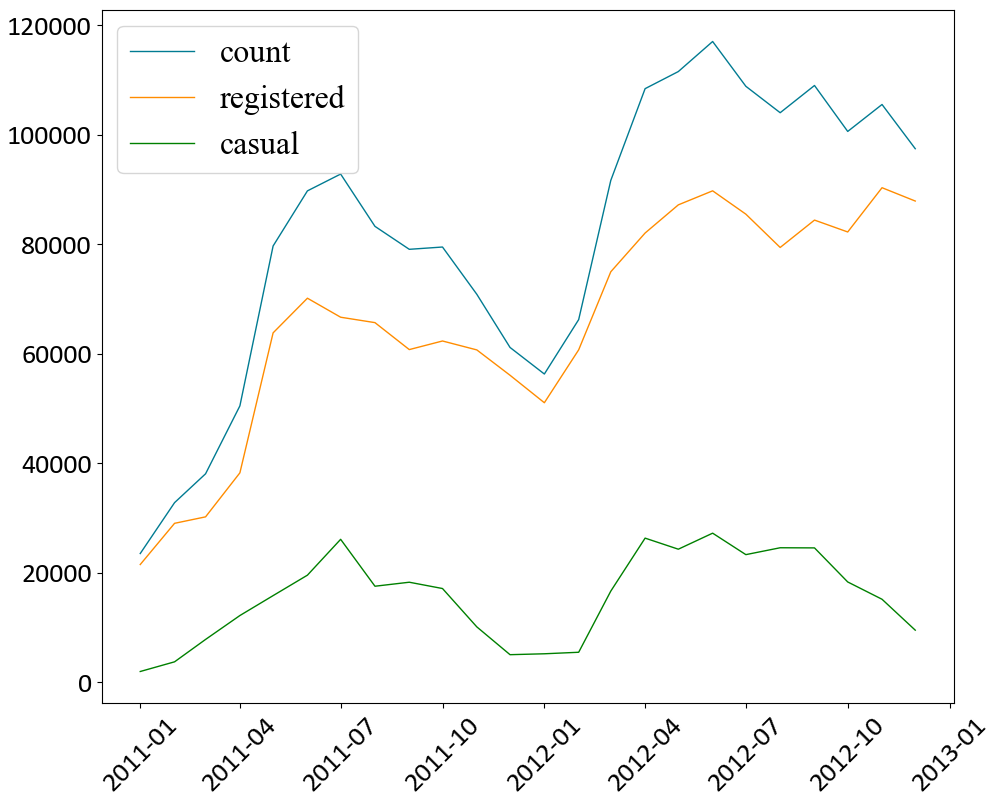
### 4.6 年份/月份/季节对自行车租赁的影响

先来看下11年和12年的会员和非会员的自行车租赁总量情况。

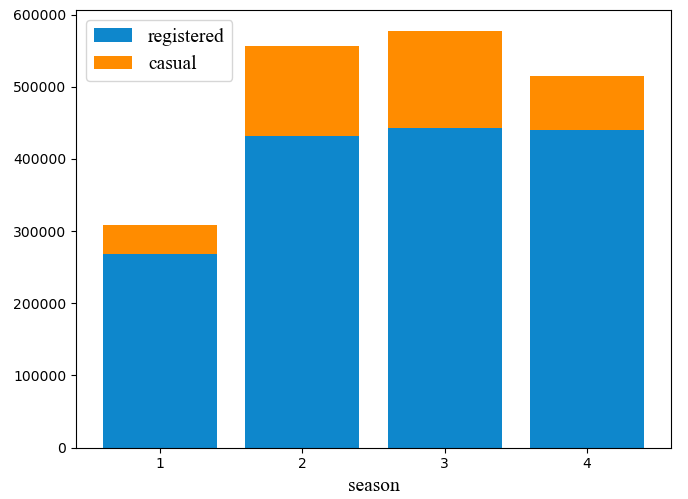


可以看到，12年相比11年，总的自行车租赁数量和会员租赁数量都有明显的增长，非会员的租赁数量出现小幅增长，更多人开始使用单车。

下面看下11和12两年每月的自行车租赁总量情况。



每年的5/6/7三个月份是用车的高峰期，而12/1/2三个月则是用车的低谷期。

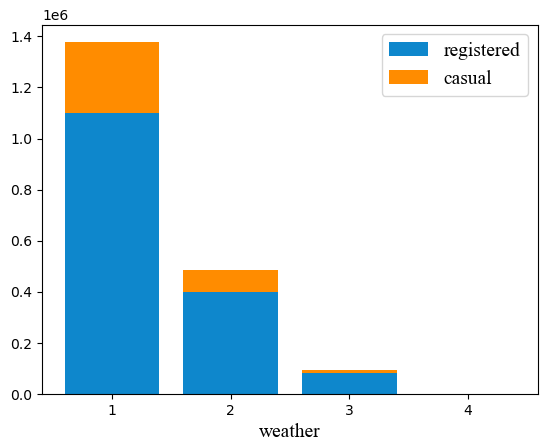


下面来看下在不同季节，用户的出行会有什么不同。

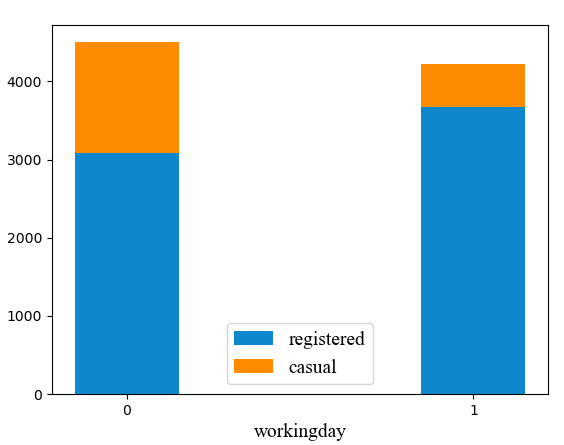
夏秋冬三个季节，自行车租赁总量相差不大，都是在一个较高的水平，春季则是一年中用车最少的季节。

### 4.7 天气对自行车租赁的影响

天气等级越高，自行车租赁数量越少，毕竟雨雪天，不会有太多人会想骑自行车。

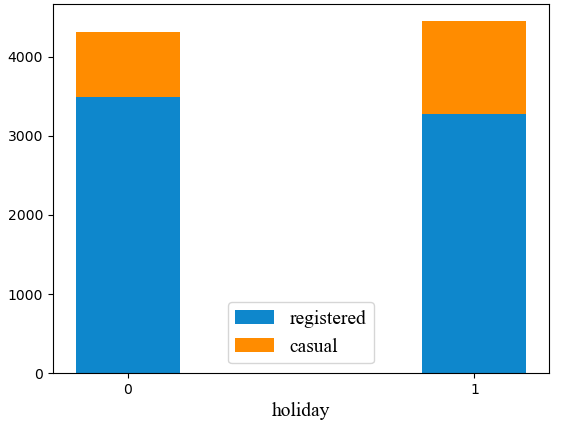


### 4.8 日期对自行车租赁的影响

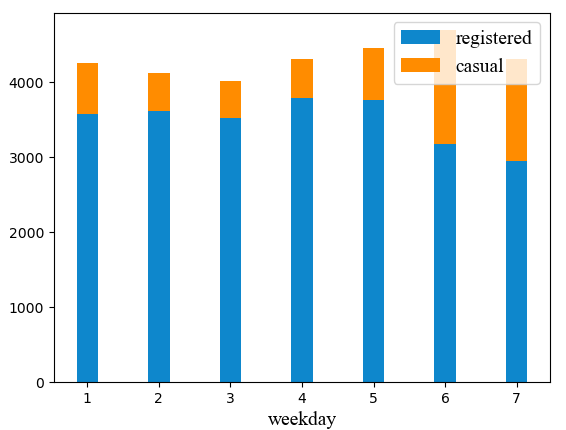


本文先来看下，工作日、非工作日平均每天的自行车租赁数量。

非工作日的自行车租赁数量要稍多于工作日，注册用户在非工作日的租赁数量要少于工作日，未注册用户在非工作日的租赁数量要多于工作日。



这个图，和上面的工作日、非工作日对自行车租赁影响的图差不多，所以两者对自行车的影响的结论也是一样。

可以发现，在周末，会员用车下降明显，而非会员用车则明显增长。

## 5 建模与算法描述

### 5.1 建模

根据观察，本文决定将

hour,temp,humidity,year,month,season,weather,windspeed\_rfr,weekday,weekday,workingday,holiday作为特征值，使用随机森林进行回归预测。

步骤是：先将多类别型数据进行分类，然后设定训练集，生成模型进行拟合，最后提交到kaggle后，成绩score=0.40488。

使用决策树进行预测score=0.53650，效果比随机森林差。[[1-3](#_ENREF_1" \o "Zaki, 2014 #23)]

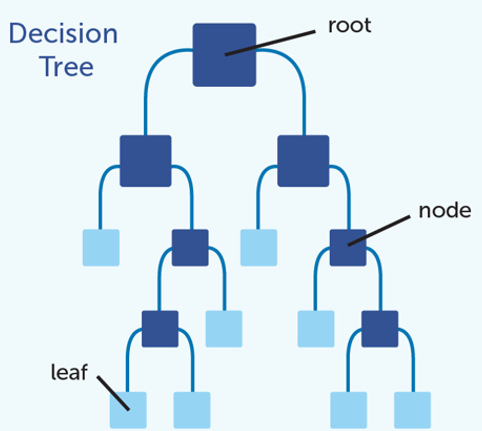
### 5.2 算法描述

#### 5.2.1.决策树

（1）决策树模型的结构

决策树是一种分类与回归方法。决策树的结构呈树形结构，在分类问题中，其代表基于特征对数据进行分类的过程，通常可以认为是if-then规则的集合，也可以认为是定义在特征空间与类空间上的条件概率分布。其主要优点是模型可读性好并且分类速度快。训练的时候，利用训练数据根据损失函数最小化的原则建立决策树模型。预测时对于新的数据，利用决策树进行分类。决策树的学习通常包括三个步骤：特征选择，生成决策树，对决策树进行剪枝。这些决策树的思想主要来自Quinlan在1986年提出的ID3算法和1993年提出的C4.5算法，以及Breiman等人在1984年提出的CART算法。

用于分类的决策树是一种对数据进行分类的树形结构。决策树主要由节点（node）和有向边（directed edge）组成。节点有两种类型：内部节点（internal node）以及叶节点（leaf node）。内部节点表示一个特征或者属性，叶节点表示一个类。其结构如图所示：

  
决策树

（2）特征选择

特征选择在于选取对训练数据具有分类能力的特征，这样可以提高决策树的学习效率，如果利用一个特征进行分类的结果与随机分类的结果没有太大差别，则称这个特征是没有分类能力的。通常扔掉这样的特征对于决策树的学习精度影响不大，通常特征选取的准则是信息增益或者信息增益比。

在信息论中，熵是表示随机变量不确定性的度量，设X是一个取有限个值得离散随机度量，其概率分布为：

P( X = xi ) = pi ( i = 1, 2,......,n )

那么随机变量的熵定义为：

H(X)=−∑ pi logpi  
熵越大，随机变量的不确定性越大，从定义可以验证

0≤H(p)≤logn

当p=0或p=1时H(p)=0 ，随机变量完全没有不确定性；当p=0.5时，H(p)=1H(p)=1，熵取值最大，随机变量的不确定性最大。

条件熵H(Y|X)表示在已知随机变量的条件下随机变量的不确定性。

随机变量X给定的条件下随机变量Y的条件熵H(Y|X))，定义为 X给定条件下Y的条件概率分布的熵对XX的数学期望:

H(Y|X)=∑pi H(Y|X=xi)

这里pi = P( X = xi ), I = 1,2,...,n，信息增益表示得知特征X的信息而使得类Y的不确定性减小的程度。

特征A对训练集D的信息增益g(D,A)，定义为集合D的经验熵H(D)与特征A给定条件下D的经验条件熵H(D|A)之差，即

g(D|A)=H(D)−H(D|A)

（3）构建决策树的ID3算法

决策树的算法主要有ID3，C4.5以及CART，ID3算法的核心是在决策树各个节点上应用信息增益准则选取特征，递归构建决策树。

下面得到构建决策树的ID3算法：

输入：训练数据集D，特征集A，阈值ε；

输出：决策树T.

（1）若D中所有实例属于同一类Ck，则T为单节点树，并将Ck作为该节点的类标记，返回T；

（2）若A=∅，则T为单节点树，并将D中实例数最大的类Ck作为该节点的类标记，返回T；

（3）否则按照信息增益算法计算A中各特征对D的信息增益，选择信息增益最大的特征Ag；

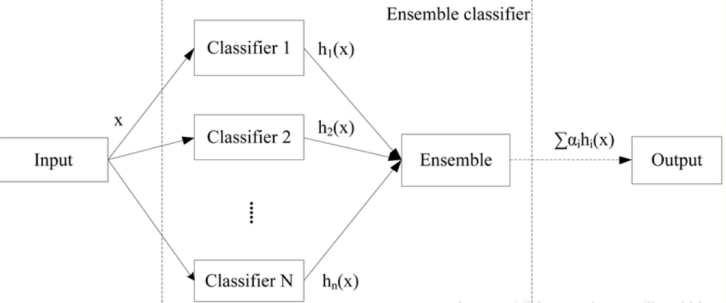
（4）如果Ag的信息增益小于阈值ε，则T为单节点树，并将D中实例数最大的类 作为该节点的类标记，返回T；

（5）否则，对Ag的每一可能值ai，依据iAg=ai将D分割为若干非空子集Di，将Di中实例数最大的类作为标记，构建子节点，由节点及其子节点够成树T，返回T；对第i个子节点，以Di为训练集，以A−[[1](#_ENREF_1" \o "Zaki, 2014 #23)]为特征集，递归地调用步1-步5，得到子树Ti，返回Ti。

#### 5.2.2随机森林算法

（1）集成学习

集成学习通过构建并结合多个学习器来完成学习任务，有时候也被称为多分类器系统（multi-classfier system）、基于委员会的学习等。

下图显示出集成学习的一般结构：先构建一组“个体学习器”（individual learner），再用某种策略将它们结合起来。个体学习器通常由一个现有的学习算法由训练数据产生，例如ID3决策树算法，BP神经网络算法等，此时集成中只包含同种类型的个体学习器，例如“决策树集成”中全是决策树，“神经网络集成”中全是神经网络，这样的集成是“同质”的（homogeneous）。同质集成中的个体学习器也称为“基学习器”（base learner），相应的学习算法称为“基学习算法”（base learning algorithm）。集成也可包含不同类型的个体学习器，例如同时包含决策树和神经网络，这样的集成是“异质”的（heterogenous）。异质集成中的个体学习器由不同的学习算法产生，这是就不再有基学习算法；相应的，个体学习器一般不称为基学习器，常称为“组件学习器”（component learner）或直接称为个体学习器。  


集成学习

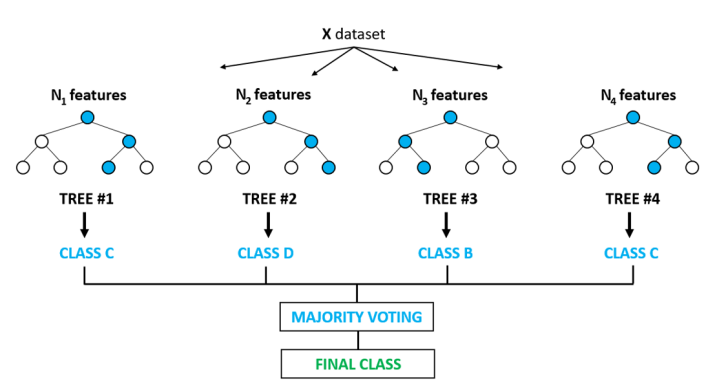
集成学习通过将多个学习器进行结合，常常可以获得比单一学习器更为显著优越的泛化性能。这对“弱学习器”（weak learner）尤为明显，因此集成学习的很多理论研究都是针对弱学习器进行的，而基学习器有时也被直接成为弱学习器。但是需要注意的是，虽然从理论上来说使用弱学习器集成足以获得好的性能，但在实践中出于种种考虑，例如希望使用较少的个体学习器，或是重用关于常见学习器的一些经验等，人们往往会使用比较强的学习器。

根据个体学习器的生成方式，目前的集成学习方法大致可分为两类，即个体学习器间存在强依赖关系、必须串行生成的序列化方法，以及个体学习器间不存在强依赖关系、可同时生成的并行化方法；前者的代表是Boosting，后者的代表是Bagging和“随机森林”（Random Forest）。

Bagging是bootstrap aggregating。思想就是从总体样本当中随机取一部分样本进行训练，通过多次这样的结果，进行投票获取平均值作为结果输出，这就极大可能的避免了不好的样本数据，从而提高准确度。因为有些是不好的样本，相当于噪声，模型学入噪声后会使准确度不高

（2）随机森林算法

随机森林（Random Forest，简称RF）是Bagging的一个扩展变体，RF在以决策树为基学习器构建Bagging集成的基础上，进一步在决策树的训练过程中引入了随机属性选择。一棵树的生成肯定还是不如多棵树，因此就有了随机森林，解决决策树泛化能力弱的特点。具体来说，传统决策树在选择划分属性时是在当前节点的属性集合（假定有d个属性）中选择一个最优属性；而在RF中，对基决策树的每个节点，先从该节点的属性集合中随机选择一个包含k个属性的子集，然后再从这个子集中选择一个最优属性用于划分。这里的参数k控制了随机性的引入程度：若令k=d，则基决策树的构建与传统决策树相同；若令k=1，则是随机选择一个属性用于划分；一般情况下，推荐值k=log2d。总的来说就是随机选择样本数，随机选取特征，随机选择分类器，建立多颗这样的决策树，然后通过这几课决策树来投票，决定数据属于哪一类(投票机制有一票否决制、少数服从多数、加权多数)

随机森林简单，容易实现，计算开销小，令人惊奇的是，它在很多现实任务中表现出了强大的性能，被誉为“代表集成学习技术水平的方法”。可以看出，随机森丽对Bagging只做了小改动，但是与Bagging中基学习器的“多样性”仅通过样本扰动而来不同，随机森林中基学习器的多样性不仅来自样本扰动，还来自属性扰动，这就使得最终集成的泛化性能可通过个体学习器之间的差异度的增加而进一步提升，决策树的结构如下图：  


随机森林

#### 5.2.3决策树与随机森林算法对比

(1)决策树的优缺点

优点：

a.可解释性高，决策树可以看成是一个if-else的可视化的结果；

b.能处理非线性数据；

c.不需要做数据的归一化，因为不同属性的数据之间不需要做类似乘法这种运算；

d.可以用于特征工程，特征选择;

e.计算量较小

缺点：

a.容易产生过拟合;

b.微小的数据改变会改变整个树的形状；

c.对类别不平衡的数据不友好。。

(2)随机森林的优缺点

优点：

a.降低异常值所带来的影响：因为随机森林选取了部分数据建立了多个决策树，即使有个别决策树会因为异常值的影响导致预测不准确，但预测结果是参考多个决策树得到的结果，降低了异常值带来的影响。

b.降低了过拟合的可能性，因为决策树是采用了所有的特征及样本，容易出现过拟合（即对训练样本有很好的效果，对测试集的效果很差），随机森林是采用了部分样本的部分特征而构造的很多个决策树（采取的有放回抽样），特征和数据在单个决策树上变少了，降低了过拟合的可能性。

缺点：

a.计算量相对于决策树很大，性能开销很大。

b.可能会导致有些数据集没有训练到，但这种几率很小

## 6 改进方法性能分析

本文认为，从月份的角度思考租赁车辆更合理，因为温湿度变化，在不同季节交替会有过渡期。本文剔除了月份month因素，进行预测，发现随机森林score=0.40399，效果更好。但决策树算法score=0.53742，性能下降。

在数据预处理方面，对于共享单车的租赁，不仅在时间和气象方面都是高度动态和相互关联的，而且还受时间和空间等复杂问题的影响。所以除了给出的气象数据和时间数据，在数据收集上还可以考虑增加站点位置的数据，比站点附近有学校、居民区、商业圈或者风景观光点，还应该考虑站点位置属于郊区还是市区。另外还可以考虑收集城市人口的数据以及市的各个行政区，各个社区街道的常住人口有多少。这样能够更好的预测出共享单车租赁站点的需求量。然后我们可以根据不同的情况，分配不同的策略来平衡需求量：在高峰时段之前派卡车重新分配自行车，为大型社交活动设置临时自行车道以提供额外的站点等。

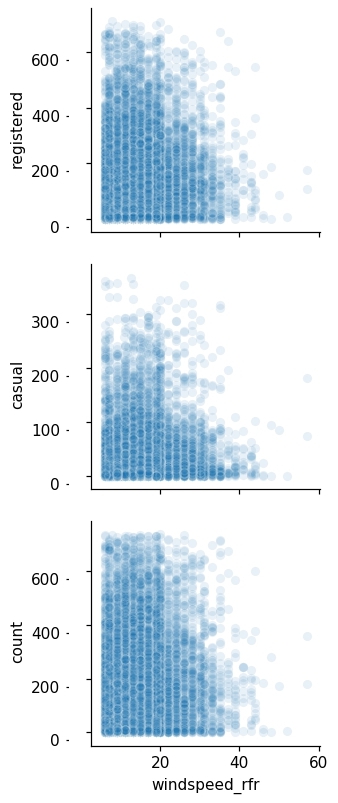
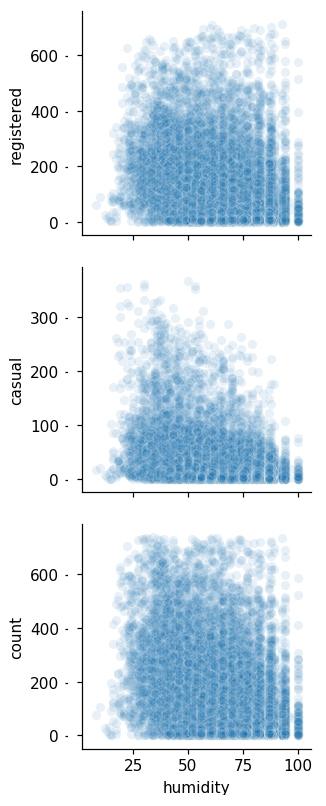
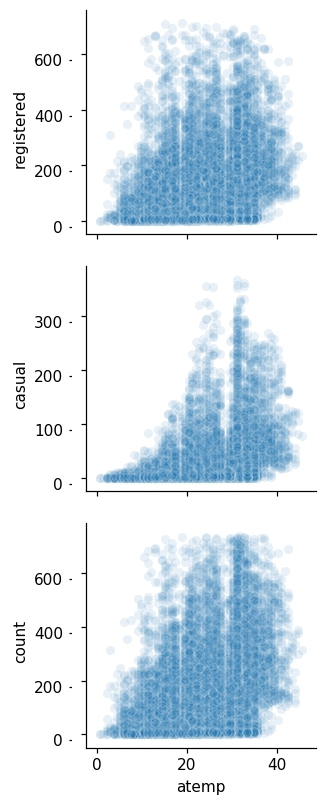
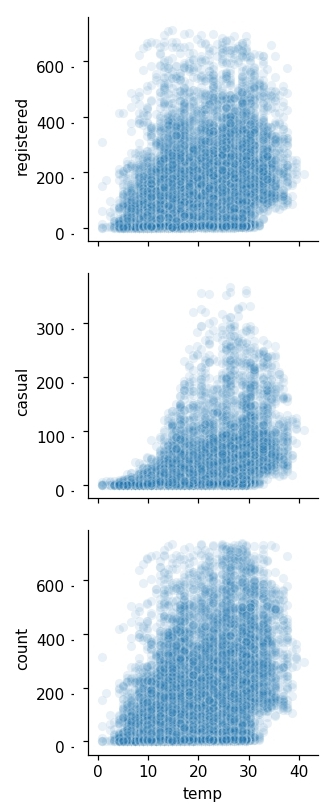
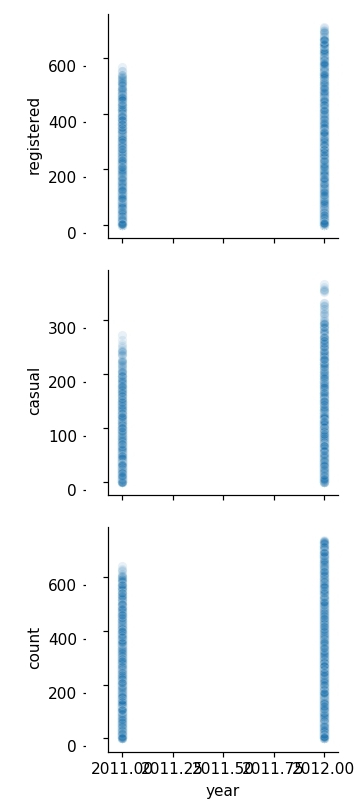
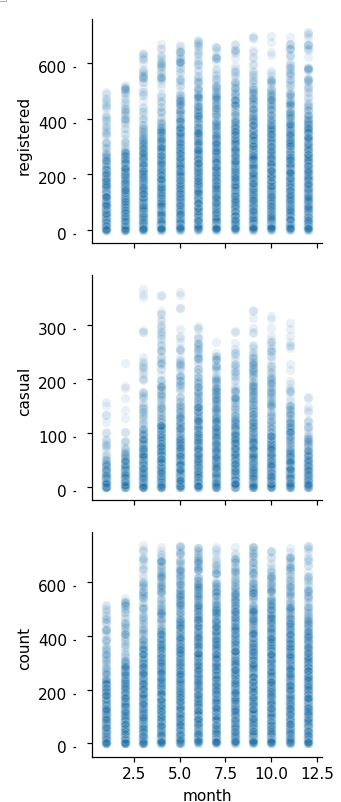
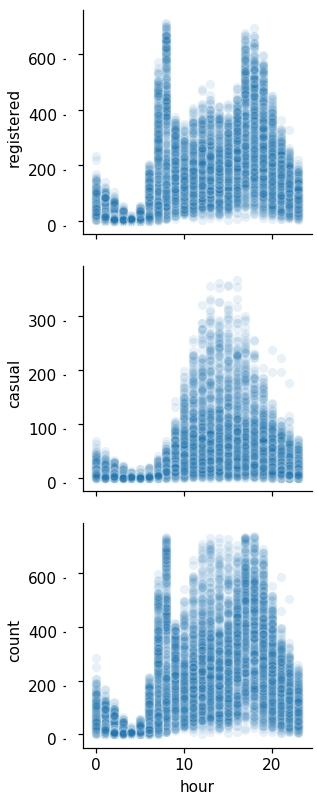
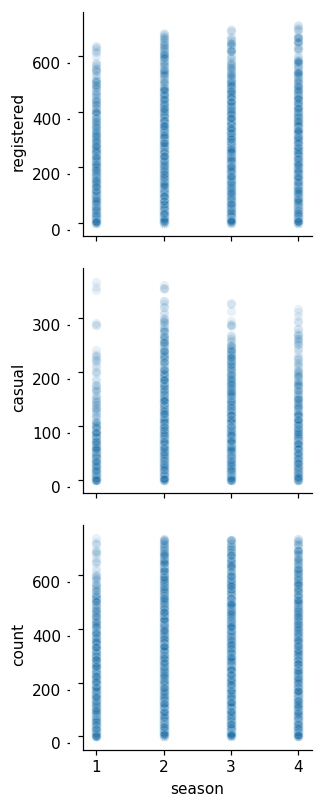
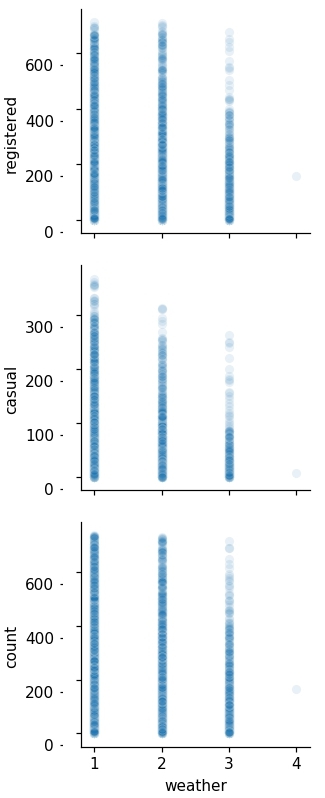
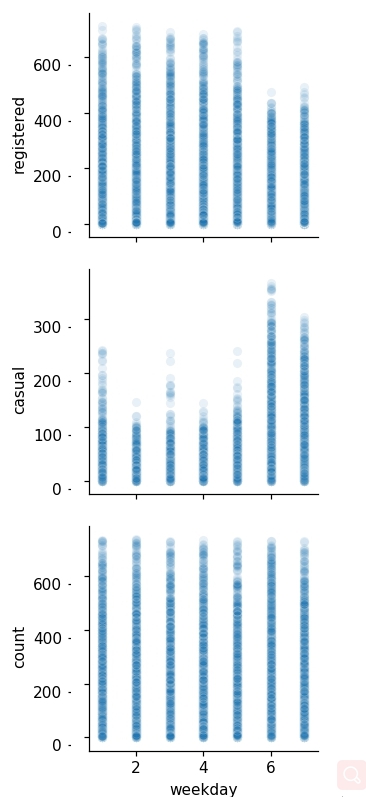
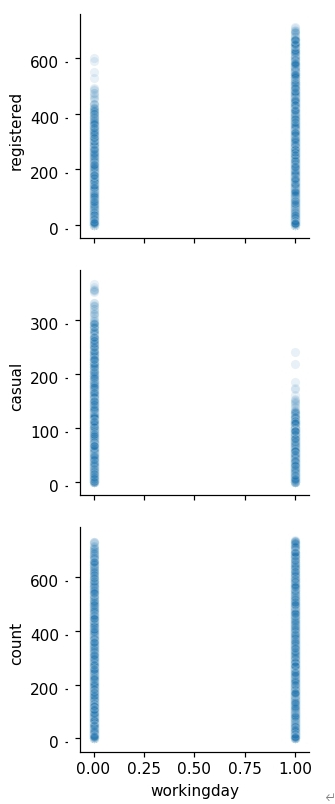
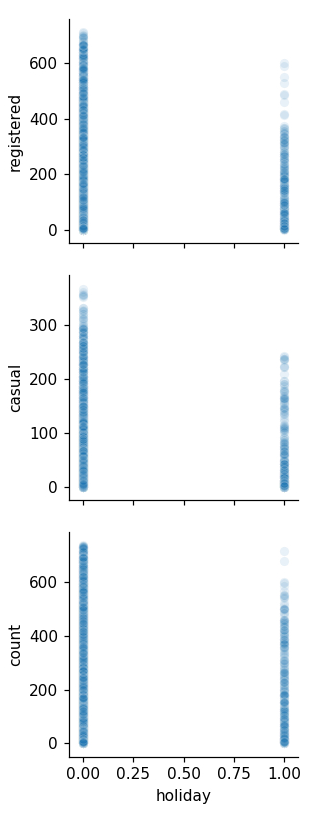
在数据预测方面，一般的数据预测方式通过概率回归和非线性拟合等方式得到历史数据的简单映射，但没有探究数据之间存在的时空联系; 其次由于用户通常在临时的基础上选择靠近其来源或目的地的站点，所以预测单个站点的需求是困难的。因此，现有的车站级自行车需求预测方法通常具有相对较低的精度。

本文中运用的随机森林方法是由 Breiman于 2001 年提出的一种取代神经网络等传统机器学习方法的分类回归算法。随机森林可以处理非线性问题、具有很好的抗噪声能力且不易陷入过拟合。与传统的多元线性回归模型相比，随机森林算法不需要预先设定函数形式，且能克服协变量之间复杂的交互作用; 与神经网络相比，随机森林算法不易陷入过拟合; 与支持向量机相比，支持向量机的核函数及内部参数比较依赖使用者技巧，而随机森林避免了此类问题。因此随机森林算法算法取得较好的效果。

## **7** 结束语

本文根据天气，风速，湿度，温度，时段等信息来预测华盛顿的自行车租赁需求。先对原始数据进行预处理，包括缺失值、重复值、异常值处理，并进行数据转换，使得数据表现更加直观。然后，本文对各项数据指标进行可视化，并联合分析，确定需要挖掘的数据。最后，本文分别使用决策树、随机森林对单车租赁数量预测，并进行参数预测改进，进行性能对比、分析。仿真结果表明，本文改进算法具有较好的性能。

## 附录**4-1**



参考文献：

[1] ZAKI M J, JR W M. Data Mining and Analysis: Fundamental Concepts and Algorithms [M]. 2014.

[2] 周志华. 机器学习 : = Machine learning [M]. 清华大学出版社, 2016.

[3] M.MITCHELL T. 机器学习(计算机科学丛书) [M]. 机械工业出版社, 2014.