共享单车需求预测

韩开放 [1]　刘伟宏 [1]　王嘉树 [1]　李东盛 [1]

[1] WuHanFanZhi University

摘要

自行车共享系统是租借自行车的一种手段，通过这些系统，人们可以从任意地点租借一辆自行车，到达目的地后归还。自行车共享系统明确记录了旅行时间，出发地点，到达地点和时间。因此，其可用于研究城市中的移动性 [1]。

在本项目中，要求将历史使用模式与天气数据结合起来，以预测华盛顿特区的自行车租赁租赁需求。数据提供了跨越两年的每小时租赁数据，包含天气信息和日期信息，训练集由每月前 19 天的数据组成，测试集是每月第 20 天到当月底的数据。

此报告仅为分析影响共享单车使用需求的因素，并将结果可视化。再进行数据预处理之后，建立 3 种基本模型进行拟合预测，包括有线性回归，逻辑回归，以及 knn 算法，最后根据均方根误差来作为评估指标。

关键字：数据预处理，线性回归，均方根误差

1. 数据描述和具体目标

1.1. 文件描述

train.cvs: 包含目标变量的训练集
test.cvs: 不包含目标变量的测试集
sampleSubmission.csv：格式正确的示例提交文件

1.2. 数据字段

kaggle 官网提供了一个跨越两年时间的每小时租车数据，其中训练集提供了每个月前 19 天的数据和使用情况，测试集提供了 20 号后到月末的数据。来看一下具体字段：

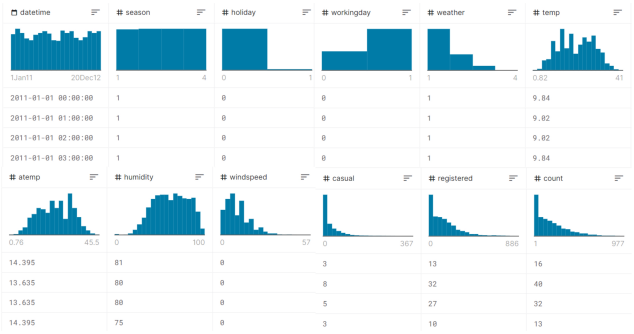

图 1. 变量预览图

datetime	日期	每小时的日期数据
season	季节	1=春 2=夏 3=秋 4=冬
holiday	节假日	1=节假日 0=非节假日
workingday	工作日	1=工作日 0=周末
weather	天气	1=晴天多云 2=雾天阴天 3=小雪小雨 4=大雨大雪大雾
temp	实际温度	气温摄氏度
atemp	体感温度	体感温度
humidity	湿度	湿度
windspeed	风速	风速
casual	非注册	非注册用户个数
registered	已注册	注册用户个数
count	租赁数量	每小时的总租车人数

图 2. 变量含义图

图 2 为数据描述的注释文档，可见数据包含了 12 个特征字段，大致能分为三类：

时间类特征（datetime，season，holiday，workdingday）

天气类特征（weather，temp，atemp，humidity，windspeed）

目标类特征（casual，registered，count）

1.3. 最终目标

最终目标：使用租赁期之前可用的信息，来预测测试集每个小时的单车使用量。

1.4. 评估指标

评估指标：要求用均方根误差 (Root Mean Squared Logarithmic Error, RMSLE) 来评价模型的好坏。其数学公式为 [3]：

$$RMSLE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}[\log(p_i+1) - \log(a_i+1)]^2} \tag{1}$$

其中，n 是测试集样本数，p_i 是测试值，a_i 是实际值。当均方根误差越小时，表示数据的拟合效果越好，测试值越接近实际值。

1.5. 实施流程

建立模型前所需要进行的数据预处理：

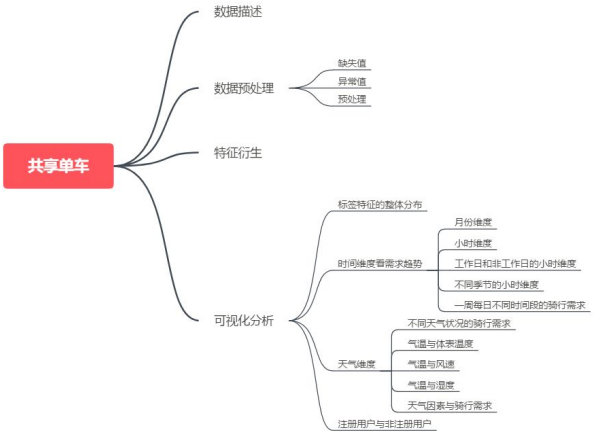

图 3. 变量流程图

2. 数据预处理

2.1. 缺失值分析

读取数据后，将训练数据集与测试数据集合并（方便对数据进行统一预处理）。

首先查看数据的异常和缺失情况：我们看到：三个目标变量存在缺失值，且数量与测试数据集一致，因此能判断数据较完整无缺失情况，缺失的只是我们需要预测的那一部分目标变量。

```
datetime        0
season          0
holiday         0
workingday      0
weather         0
temp            0
atemp           0
humidity        0
windspeed       0
casual       6493
registered   6493
count        6493
dtype: int64
```

图 4. 缺失值分析

我们将超过样本均值 3 个标注差以外的数据看作异常值，本文对异常情况暂不处理，留到模型构建前再筛选排除。

2.2. 异常值处理

利用箱线图可视化目标变量 Count：通过绘制箱线图可以看到，目标变量 count 存在较多的离群值。我们可以通过 3 σ 原则去除它的异常值。

3 σ 原则: 3 σ 准则又称为拉依达准则，它是先假设一组检测数据只含有随机误差，对其进行计算处理得到标准偏差，按一定概率确定一个区间，认为凡超过这个区间的误差，就不属于随机误差而是粗大误差，含有该误差的数据应予以剔除。且 3 σ 适用于有较多组数据的时候。

可以认为，数值分布几乎全部集中在 (u-3 σ,u+3 σ) 区间内，超出这个范围的可能性仅占不

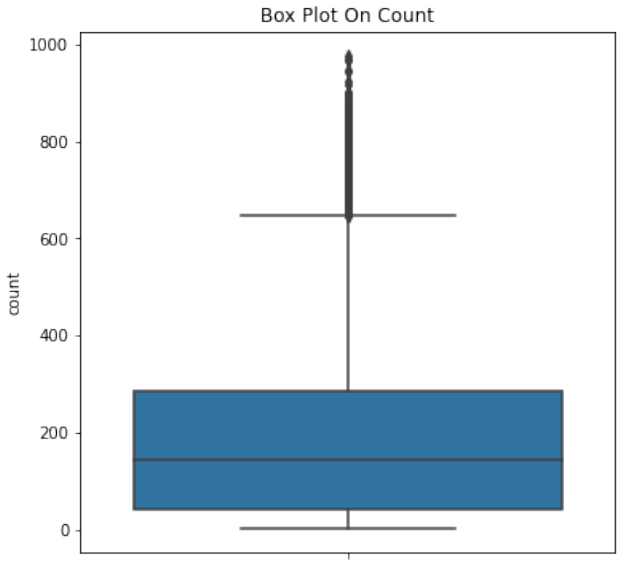

图 5. 异常值处理

到 0.3%。

3. 可视化分析

首先对 8 个类别特征进行可视化：

- 不同时间段对使用量的影响（ hour → count ）
  图 1：在一天中不同时间段，共享单车使用量差异明显，在 8 点和 16-19 点明显多于其他时间点，考虑到的原因是在这期间是上下班的高峰期；在 0-5 点明显低于其他时间点，考虑到的原因是在此期间为睡眠时间。
- 星期几对使用量的影响（ weekday → count ）
  图 2：星期几对单车总使用量没有太大的影响。
- 一个月内的哪一天对使用量的影响（ day → count ）
  图 3：一个月的哪一天对单车总使用量也没有太大的影响。
- 月份对使用量的影响（ month → count ）
  图 4：11 月-4 月的共享单车使用量会比其他月份少一点，可能是季节原因，冬季和春季太冷导致使用量降低。所以接下来我们观察一下季节对使用量的影响进行验证。
- 季节对使用量的影响（ season → count ）
  图 5：可以看到冬春相对夏秋使用量相对较少，与上面月份产生的结论相互印证。
- 节假日对使用量的影响（ holiday → count ）
  图 6：是否节假日对单车总使用量基本没有太大的影响。
- 工作日对使用量的影响（ workingday → count ）
  图 7：是否工作日对单车总使用量也基本没有太大的影响。
- 天气对使用量的影响（ weather → count ）
  图 8：天气对单车的影响基本符合日常生活中的实际情况，下雨天单车使用量减少，下暴雨时基本没人使用共享单车。

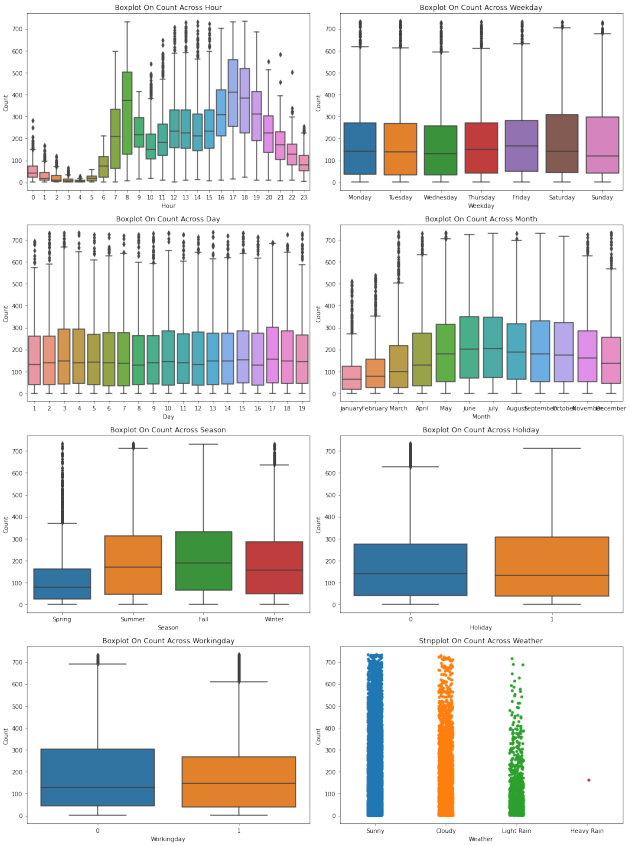

图 6. 8 个类别特征

4. 模型建立与求解

我们采用了三种基本模型来进行了拟合预测，三种算法分别都有各自的优秀与不足。

4.1. Linear Regression

线性回归是利用数理统计中回归分析，来确定两种或两种以上变量间相互依赖的定量关系的一种统计分析方法，运用十分广泛。其表达形式为 y = w^x+e，e 为误差服从均值为 0 的正态分布 [4]。

4.1.1 特点

回归分析中，只包括一个自变量和一个因变量，且二者的关系可用一条直线近似表示，这种回归分析称为一元线性回归分析。如果回归分析中包括两个或两个以上的自变量，且因变量和自变量之间是线性关系，则称为多元线性回归分析。

线性回归模型经常用最小二乘逼近来拟合，但他们也可能用别的方法来拟合，比如用最小化“拟合缺陷”在一些其他规范里（比如最小绝对误差回归），或者在桥回归中最小化最小二乘损失函数的惩罚. 相反, 最小二乘逼近可以用来拟合那些非线性的模型. 因此，尽管“最小二乘法”和“线性模型”是紧密相连的，但他们是不能划等号的。

4.1.2 结果

	Model	RMSLE
0	Linear Regression	1.041528

4.2. Sigmod Regression

在之前的文章中我们介绍了线性回归问题，不难发现线性模型主要是用来做回归问题，但是通过本节的学习，你就会发现线性模型也可以用来分类，这就是逻辑回归分类。它的大致思想为：通过计算样本属于某个类别的概率值大小对样本进行分类，一般来说，如果样本属于某个类别的概率大于 0.5 就属于类别 1，小于 0.5 就属于类别 0。

4.2.1 特点

分类：根据某个情况发生的概率大小和给定的判定阈值判断样本的类别，常用于二分类，但是也可以用于多分类问题，对于多分类问题的处理思想是：可以将其看做成二类分类问题，保留其中的一类，剩下的作为另一类。[5]
预测：根据模型，预测在不同自变量情况下，发生某种情况的概率大小。

4.2.2 结果

	Model	RMSLE
0	Linear Regression	1.041528
1	Logistic Regression	1.152351

4.3. KNeighbors Regression

KNN 可以说是最简单的分类算法之一，同时，它也是最常用的分类算法之一，注意 KNN 算法是有监督学习中的分类算法，它看起来和另一个机器学习算法 Kmeans 有点像（Kmeans 是无监督学习算法），但却是有本质区别的 [2]。

4.3.1 特点

KNN(K-Nearest Neighbor) 是最简单的机器学习算法之一，可以用于分类和回归，是一种监督学习算法。它的思路是这样，如果一个样本在特征空间中的 K 个最相似 (即特征空间中最邻近) 的样本中的大多数属于某一个类别，则该样本也属于这个类别。也就是说，该方法在定类决策上只依据最邻近的一个或者几个样本的类别来决定待分样本所属的类别。

4.3.2 结果

	Model	RMSLE
0	Linear Regression	1.041528
1	Logistic Regression	1.152351
2	KNN	0.852566

5. 总结与展望

我们将样本数据进行了可视化处理，对缺失值与异常值进行了数据预处理。并分别使用了线性回归、逻辑回归、KNN 算法，根据三种算法的均方根误差值对比，选出了误差最低的 KNN 算法，并提供了最终的预测结果。结果表明。我们使用 KNN 算法在对 count 预测时，均方根误差明显低于其他两种方案，并且简单而高效。

在未来的学习中，我们计划学习更加符合预测结果的模型，并打算深入学习 Stacking 模型融合，以克服我们现有算法在处理这类情况中存在的缺陷。

参考文献

[1] 史越. 共享单车需求预测及调度方法研究. PhD thesis, 北京交通大学, 2019. 409

[2] 张晓辉, 李莹, 王华勇, and 赵宏. 应用特征聚合进行中文文本分类的改进 knn 算法. 东北大学学报 (自然科学版), 2003. 412

[3] 李秉炽. 用代入法计算函数的均方根误差. 计量技术, 1981. 410

[4] 薛素静 and 上官同英. 多元线性回归算法的研究和应用. 华电技术, 29(005):59–60, 2007. 412

[5] 谢忠红, 张颖, and 张琳. 基于逻辑回归算法的微博水军识别. 微型机与应用, 36(16):4, 2017. 412