Kaggle Bike Sharing项目报告

标签： ML

项目定义

项目预览

项目地址

> https://www.kaggle.com/c/bike-sharing-demand

项目简介:

城市租用自行车计划是在城市中部署若干个自助租车处。在这个由租车处组成的网络中使用者可自助租用、归还自行车。迄今为止，全世界已经有500多个自助自行车租用处。

这个租车系统产生的诸如租车时间、租借/交还位置，时间消费等数据引起了人们的关注。租车系统也借此成为了一个感知网络。本题目要求借助华盛顿的历史数据来预测租车系统的租借需求。

项目描述

本项目需要通过给予的历史数据（包括天气、时间、季节等特征）预测特定条件下的租车数目。通过选择决策树、SVM(支持向量机)和随机森林算法，构建不同的模型。在特征过程后，使用训练数据集对模型行进行训练。最终使用训练好的模型对测试集进行预测。之后通过改变参数和使用交叉验证等方法提升模型精度。预测结果越靠近真实数据越好，在之后会介绍如何评判预测结果。

预测精度

项目提供了三个csv文件，分别是训练集文件，测试集文件，和上传文件格式.我们训练用到的是前两个文件。经过前面的分析，我认为这应该是一个监督学习中的回归问题。因为我们要输出一个个的离散值。我们通过数据采用一定的算法构建模型后，对测试集中的数据进行预测。采用网站给的评价准则：均方根对数误差准则，即：

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

其中:

n 为样例总数

p_i 为你的预测值

a_i 为实际值

log(x)为自然对数算法

图1 均方根对数误差定义

这个值越小代表模型拟合数据越好。

从我的认知来分析，我认为选取精度判断算法的原因如下:

首先，均方根对数误差准则具有这样的特性，对欠预测的惩罚大于对过预测的惩罚。

均方根对数误差准则一般与标准误差准则相比较。如果预测的值的范围很大，标准误差方法会被一些大的值主导。这样即使很多小的值预测准了，但是有一个非常大的值预测的不准确，标准误差就会很大。 相应的，如果另外一个比较差的算法对这一个大的值准确一些，但是很多小的值都有偏差，可能均方误差会比前一个小。因此取对数之后在进行比较，可以在一定程度上解决此问题。

对于本题目来说，通过将预测值与真实值取对数后在进行比较，可以一定程度上消除出现预测值过大出现的分数较高的情况。而模型中并未考虑到运行时间，因为相比于运行时间，预测精度才是我们着重考虑的。

为了满足均方根对数误差准则，在算法预测后对数据取对数之后在取自然常数指数可以尽可能的满足此标准。

问题分析

数据概览

通过调用pd.read_csv()函数，读取训练集和数据集。

通过对读取的数据集进行分析，得到训练集共10886条数据，测试集共6493条数据。共9个特征，分别为

```
[datetime, season, holiday, workingday, weather, temp, atemp, humidity,
windspeed]
```

以及输出的结果

```
[casual, registered, count]
```

数据样例如图

datetime	season	holiday	workingday	weather	temp	atemp	humidity	windspeed	casual	registered	count
2011-01-01 00:00:00	1	0	0	1	9.84	14.395	81	0	3	13	16
2011-01-01 01:00:00	1	0	0	1	9.02	13.635	80	0	8	32	40
2011-01-01 02:00:00	1	0	0	1	9.02	13.635	80	0	5	27	32
2011-01-01 03:00:00	1	0	0	1	9.84	14.395	75	0	3	10	13
2011-01-01 04:00:00	1	0	0	1	9.84	14.395	75	0	0	1	1
2011-01-01 05:00:00	1	0	0	2	9.84	12.88	75	6.0032	0	1	1
2011-01-01 06:00:00	1	0	0	1	9.02	13.635	80	0	2	0	2
2011-01-01 07:00:00	1	0	0	1	8.2	12.88	86	0	1	2	3
2011-01-01 08:00:00	1	0	0	1	9.84	14.395	75	0	1	7	8
2011-01-01 09:00:00	1	0	0	1	13.12	17.425	76	0	8	6	14
2011-01-01 10:00:00	1	0	0	1	15.58	19.695	76	16.9979	12	24	36

图2 数据样例展示图

项目数据描述如下：

(1) datetime：代表数据日期，以年-月-日 小时的形式给出。

(2) season：数据记录时的季节。1 为春季, 2为夏季,3 为秋季,4 为冬季。

(3) hodliday：当日是否为假期。1代表是，0代表不是。

(4) workingday：当日是否为工作日，即既不是周末也不是假期。1代表是，0代表不是。

(5) weather:当日天气：

1: 天气晴朗或者少云/部分有云。

2: 有雾和云/风等。

3: 小雪/小雨，闪电及多云。

4: 大雨/冰雹/闪电和大雾/大雪。

(6) temp - 当日摄氏温度。

(7) atemp - 当日人们感觉的温度。

(8) humidity - 当日湿度。

(9) windspeed - 风速。

(10) casual -非预定自行车的人数

(11) registered - 登记预定自信车的人数。

(12) count - 总租车数，我们需要预测的值。即casual+registered数目。

其中10~12不属于特征。

在对数据进行审查后发现特征和结果中没有极端值，因此不需要在此方面进行处理。

对特征进行分析，发现其中租车总数为需要预测的值。可以取代预定和非预定人数，因此可以后两者删除。

此外，日期变量给出的是整体时间，考虑到租车与月份和日期和小时对预测是相对重要的，在接下来的数据处理中需要将其拆分成年/月/日/小时以及周几的形式。

季节特征，在训练集是以类型变量给出的，可以考虑将其转变成01变量。对租车的预测应该也是比较重要的。

天气变量同上。

训练集中给了temp/atemp两个变量，根据对atemp描述可知：

atemp是一个主观变量，和temp(即温度变量)有很高的相关性，考虑在数据处理中将其移除。

数据可视化

季节与租车总数关系散点图：

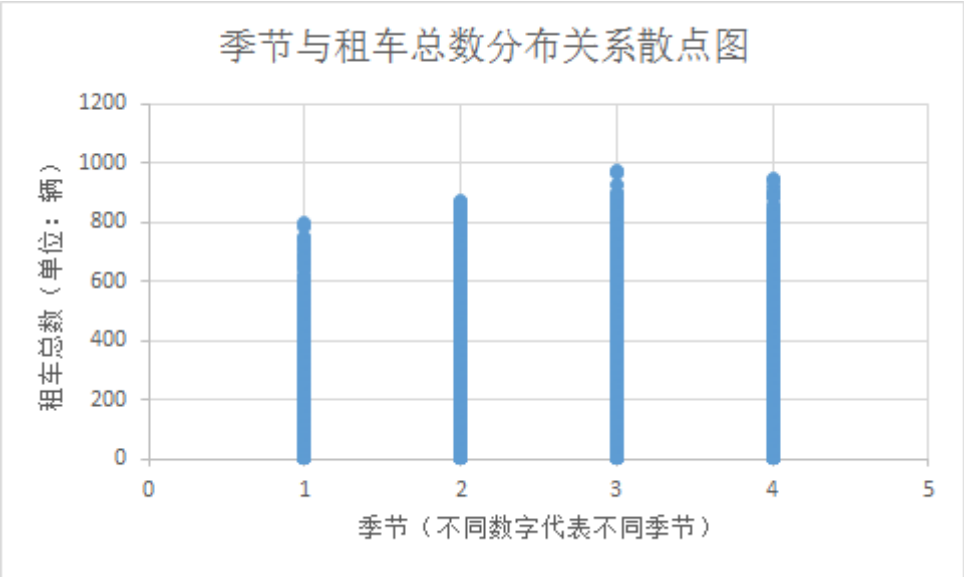

图3 季节与租车总数关系散点图

据此可以看出秋季是人们租车需求旺盛的季节。而春季是最低的季节。同时这是一个类型量，为了模型建立方便，考虑使用将其处理为四个01变量。

天气与租车总数关系散点图

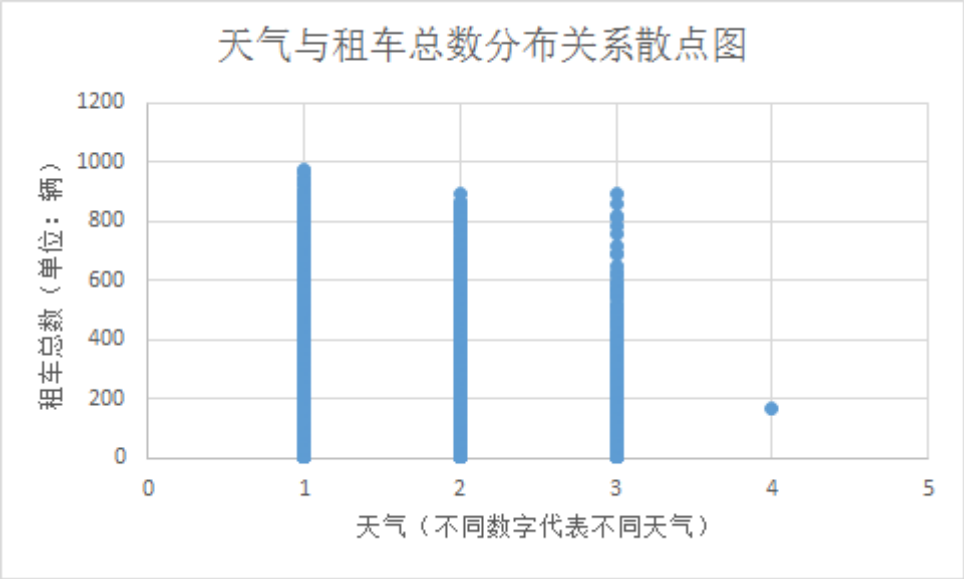

图4 天气与租车总数关系散点图

由此可以看出1类天气下租车人数最多，二三类接近，同为类型量，也将其分为四个01变量。

当地温度与感知温度分布关系图：

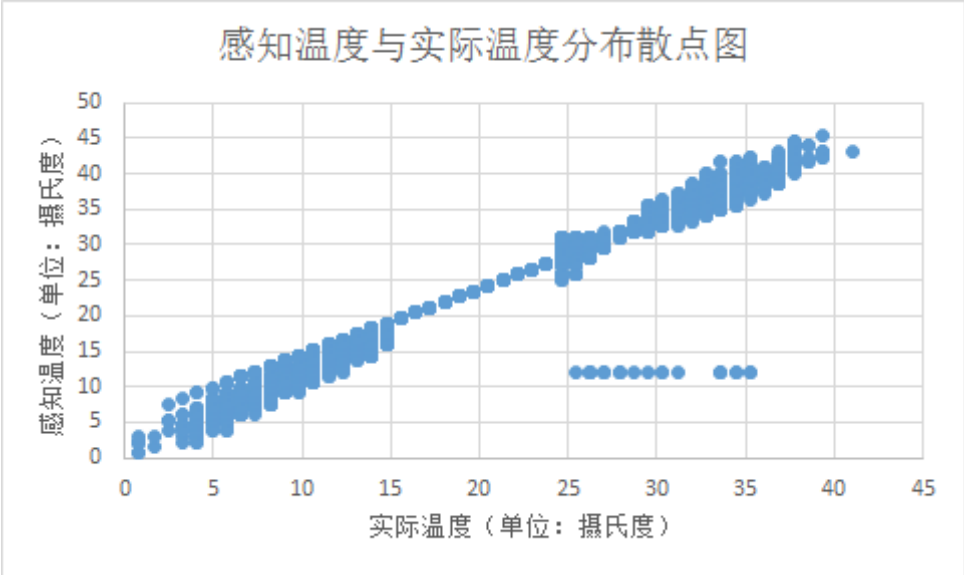

图5 当地温度与感知温度分布关系图

根据图像明显看出实际温度与感知温度具有线性相关的关系，印证了上文得到的结论，因此为了避免维度灾难，将感知温度变量移除，只保留实际温度变量。

计算租车总数的均值、中位数、四分位数。得到：

	count
count	10886.000000
mean	191.574132
std	181.144454
min	1.000000
25%	42.000000
50%	145.000000
75%	284.000000
max	977.000000

算法选取

考虑到本项目中是使用回归预测模型值，决定采用SVM、决策树，以及将决策树与聚成学习结合的随机森林算法。

SVM我之前知道的大多数用于分类，这里用于回归不知道效果如何。其有C和epsilon两个参数，其中C为误差项的罚参数。epsilon代表epsilonSVM模型。在算法改进是使用交叉验证选取最优参数。

决策树算法比起线性回归要好得多，但也容易出现过拟合。而且当出现缺失值的情况时容易出现误判。主要参数为决策树最大层数max_depth。在算法改进是使用交叉验证选取最优参数。

随机森林算法结合了决策树和继承学习的优点，其计算简单容易实现。但是其最终集成的泛化性能会随着个体差异而增大，不过在本项目中足够使用了。主要参数为森林中树的数目，由于算法本身无需交叉验证，通过改变该参数调试最优结果。

回归较常用的算法模型有线性回归、lasso回归，岭回归，以及非监督性模型KNN（K最临近算法）。其中前三种回归大多适用于线性分布的回归，在本项目中数据明显不具有线性分布的特性，因此采取线性模型预测会有较大误差。在此情况下将其作为基准模型进行选取，KNN算法当出现特殊情况（如某种情况的样例数量较少），由于该算法只计算“最近的”邻居样本，某一类的样本数量很大，那么或者这类样本并不接近目标样本，或者这类样本很靠近目标样本。

基准点选择

考虑之后我选取多元线性模型作为基准模型。其定义如下图所示：

$$\hat{y}(w, x) = w_0 + w_1 x_1 + \ldots + w_p x_p$$

图6 多元线性回归定义

其中w0为常数，xn为特征，wn为对应系数。该模型目的是将多个特征组合起来预测出一个线性模型。该式中所有参数均是拟合出来的，因此无需向其提供参数。

多元线性回归模型是线性回归的一种，用于当y值的影响因素不唯一的情况。我认为本项目中的数据不是呈线性分布的。因此线性模型预测出来的结果应该比常数要好一些，但是相比其他算法还是有差距。因此在后面我将同时使用线性模型进行拟合预测，将其作为基准方法。

实施解决方案

1.数据处理

首先按照上文所述，先对训练集和测试集进行数据处理。

首先需要将日期特征转换成如下特征：年/月/日/小时/周几。之后移除日期特征。

对天气和季节特征，将其由类型量转换为多个01变量。删除对应原特征。

最后需要删除感知温度这一特征，从而避免维度灾难。

数据处理完了之后，考虑到检验准确率的方式采用的是对数均方根，对train数据集取对数，进行预测之后再用指数函数变换回来。

根据数据分布图可知租车总数范围变化不大，无须对其进行缩放变换。

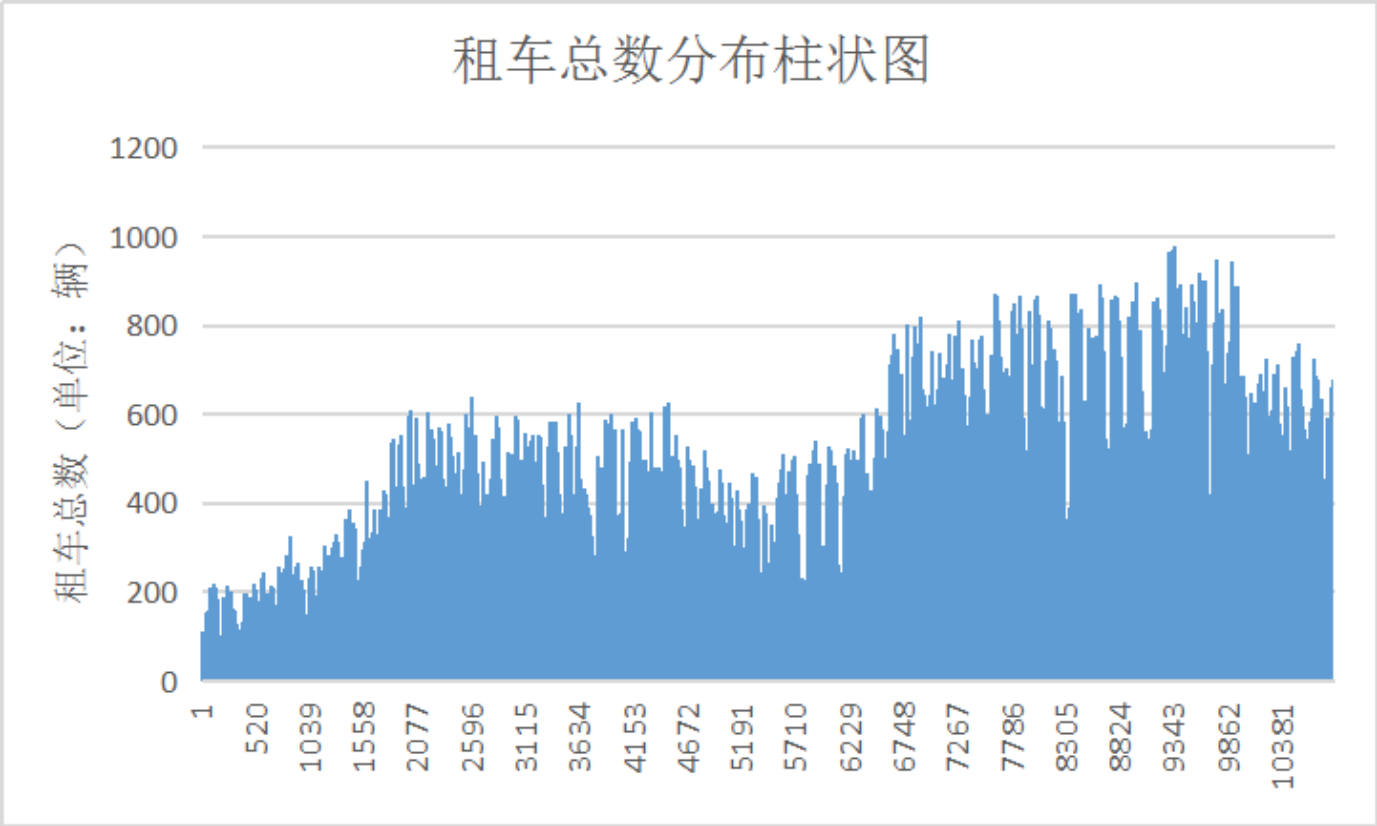

图7 租车总数分布柱状图

2.建立模型并预测

处理了数据之后，进行训练模型和预测数据。

为方便算法比较，生成一个函数读入回归模型，训练集数据，测试集数据来生成结果。根据输出分数来简单判断模型预测情况。不过此时的分数只是显示对train数据的拟合程度。在参数设置上，决策树的参数采用默认设置，随机森林模型参数设置n_estimators=100，SVM设置C=1.0, epsilon=0.2。

选取多元线性回归作为基准点的线性模型设置，无需向其提供参数。

拟合分数如下：

```
decisionTree: Variance score: 1.00
randomForest: 0.99
SVM: 0.90
linear: 0.49
```

对分数分析，发现决策树很可能出现过拟合。随机森林和决策树分数都不错，SVM最低。但是都比作为基准点的岭回归的结果好。再看一下kaggle给出的分数。

多元线性模型预测分数

-	**danache**	**1.03334**

图8 多元线性模型预测分数图

决策树分数

-	danache	0.56008

图9 决策树模型预测分数图

随机森林分数

-	danache	0.43617

图10 随机森林模型预测分数图

SVM分数：

-	danache	1.34572

图11 SVM模型预测分数图

由于在这里分数越低表现模型误差越小，因此可以看出来SVM最差，甚至比作为基准的线性模型还要差，我猜测原因是参数没有设置好，或者并不适合做此类分析。随机森林最好，决策树由于出现了过拟合情况表现也不佳。但是这两个模型的预测分数都比线性模型高很多。

模型改进

选取当前结果最好的随机森林模型进行改进。使用交叉验证方法，选取n_estimators参数和max_features参数来修正模型。n_estimators代表随机森林中的树的数目，从range(50,131,10)中选取。max_feature参数代表划分时考虑的最大特征数，从“sqrt”和“log2”中选取。分别代表对特征数目取开放和以2为底的对数。

下面的分数是max_feature为sqrt时不同n_estimators的结果，其中分数越高代表预测精度越低：

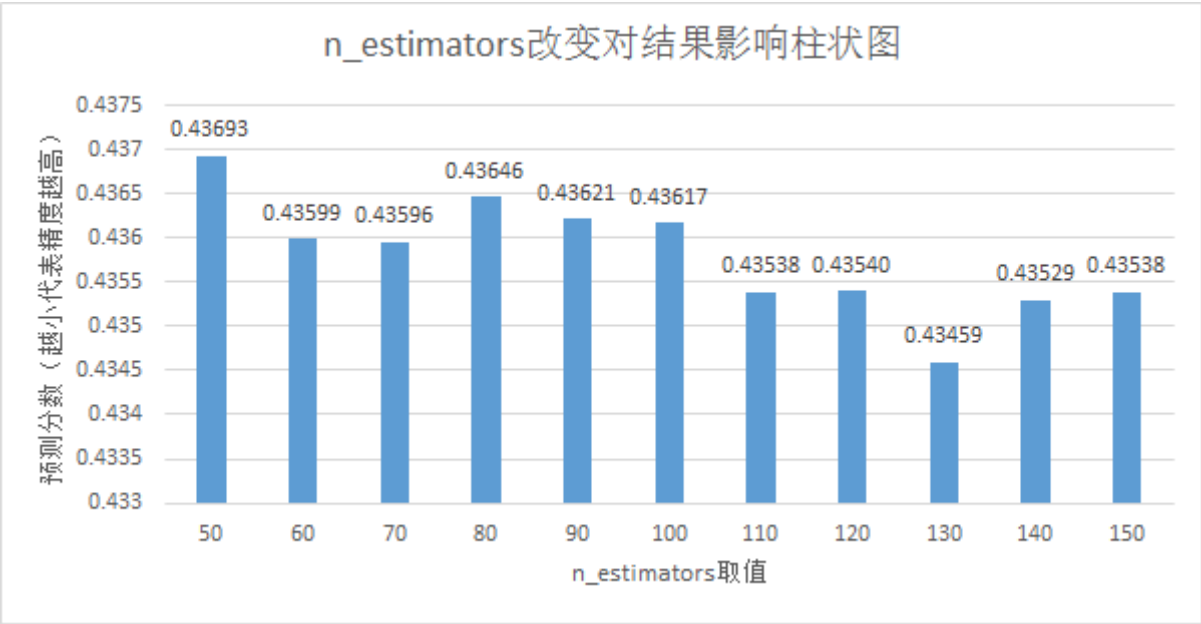

图11 n_estimators参数改变对结果影响柱状图

而使用log2时的结果逊于sqrt，我猜测是在特征数取对数后数目大大减少，最大值的限制使得模型不能很好地拟合数据。

最终交叉验证得到的参数符合上面的结论，选取n_estimators为130，max_feature为sqrt。得到的结果分数为0.43459。而开始的参数n_estimators为100，max_feature为sqrt得到的分数0.43617 。改进后的模型预测更加精确。

结果

结果评估

根据交叉验证方法，将数据集分为10折后，使用不同的参数组合，根据其预测分数，最终得到最优的参数组合。由“grid.best_params_”函数可以得到使得模型最优的参数组合：[n_estimators=130,max_feature=sqrt]。其预测结果在kaggle上分数是最高的，为0.43459，排名698名。各参数与对应分数柱状图如下：

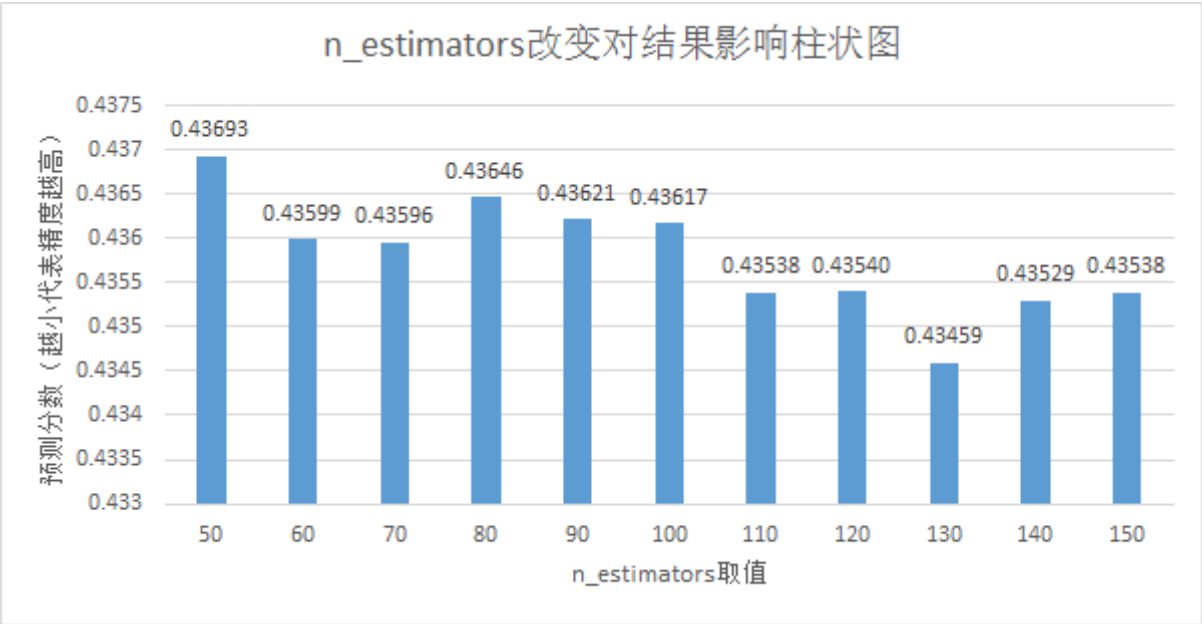

图11 n_estimators参数改变对结果影响柱状图

根据结果来说较好的满足的当初的预期。而其他两个模型不理想，我推测由于是模型算法原因，对于此类回归问题效果不太好。

在算法的鲁棒性方面，随机森林算法对参数扰动的干扰性较强，从上节的结果可以看出。当n_estimators参数变化时，虽然最后结果略有差异，但是总体的变化时非常微小的，因此可以体现出该模型对参数变化的不敏感性。

最后，与作为基准的线性模型比较，随机森林方法在预测精确度上完胜。也证明了上述参数构成的模型在本项目中是较好的选择。

总结

数据展示

选取随机森林中n_estimators参数作为自变量，验证数据拟合分数。此时的max_feature为sqrt。

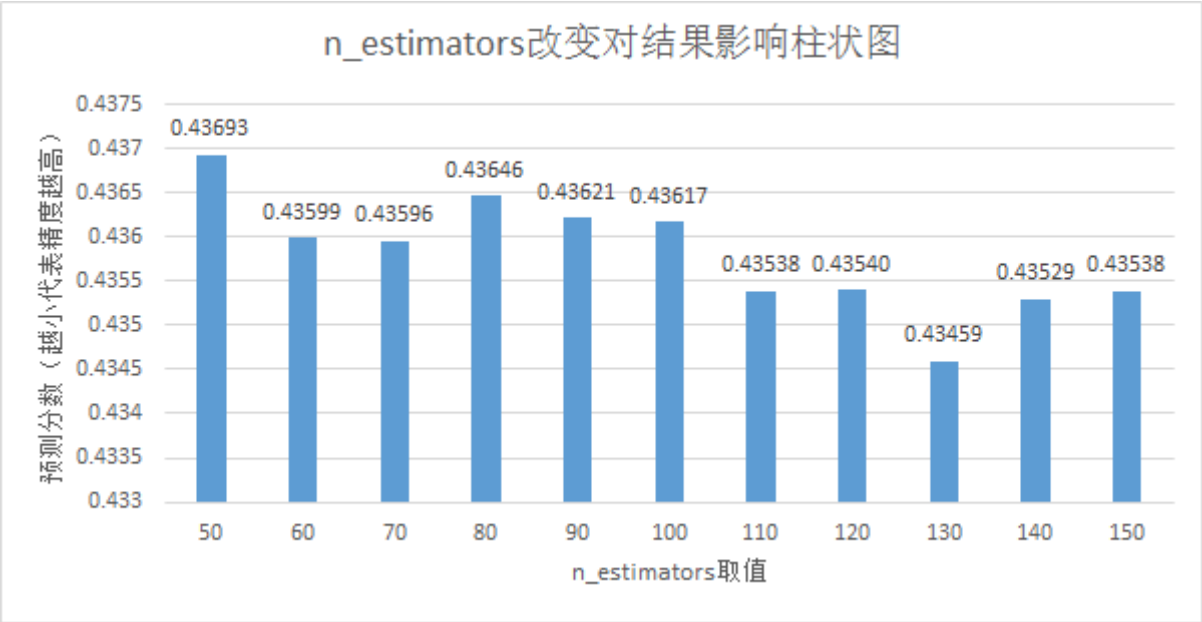

图11 n_estimators参数改变对结果影响柱状图

x轴为n_estimators的取值，y轴为kaggle网站数据拟合的分数。

发现随着n_estimators的增加，模型的拟合能力趋势为拟合分数越来越第，即性能越来越好，但是随着n_estimators数据拟合的时间越来越长，虽然项目中不考虑拟合模型花费的时间，但是真实情况下时间成本还是需要考虑的，当n_estimators为130时拟合时间已经相当长了。因此在增加变量数目对数据的提升不多。

反思

在完成本项目的过程中，我了解了以下几点：

①好的工作流程是必要的。一开始做的过程中，我只是参考之前的项目经验。然而边做边发现如果有一个好的流程框架对项目的提升是很大的。不论是项目结果还是完成时间

②数据清洗是至关重要的一环，甚至在很大程度上影响结果的精度。比如在本项目中给的datetime变量，由于其包含年月日小时等参数，如果不对其进行处理的话在模型预测中甚至是会有负收益。通过将其拆分并得出隐藏参数周数将其充分利用。而这个隐藏参数的挖掘需要对项目本身和实际情况有较好的理解。机器学习不仅是应用算法处理数据，还要加入自己对数据的理解和处理才能得到好的模型。

③模型选择也是比较重要的。通过上述讨论我们可以发现不同的模型对同一数据的预测结果差别是相当大的。在决策树模型中我发现当maxdepth为3时，对训练集的拟合精度可达到0.99，而预测集中，好多结果是相同的，而在真实情况中基本不可能发生这种情况。通过改变maxdepth可以在一定程度上改善这

种问题，但是由于决策树本身的特性，过拟合还是无法避免的。

④对于结果还是比较满意的。训练结果很好，而且模型原因泛化能力较强。本项目目中的亮点是对season,weather特征的处理，一开始我自己的模型并没有对此进行处理，后来上了论坛之后才发现可以通过将其转为01变量提高精度。此外还有将其取对数进行分析最后在取指数的方法。根据题目不同的判断精度方式应该选取合适的处理方法。

提升

最终结果在kaggle达到了前20%，但是还是不够精确。这与自己经验、处理方式都有关系。在如何提高模型精度上我认为有以下几点：

①做好数据清洗，由于不知道更好的结果是如何做到，我猜测应该是对其进行了更多的特征转换等方法。如上文所述，好的数据清洗能很大程度上决定预测的精度。

②选取合适的模型。本项目中选取了决策树、SVM和随机森林。除此之外还有线性回归、神经网络等等方法没有尝试，或许使用卷积神经网络方法能对结果进行更好的预测？采用更好的模型也能提升预测的准确度。

③此外，就自行车租借而言，我们预测所需信息可能不仅仅是题目所给的，地理位置，附近有无居民区/商场等情况都应该考虑在内，限于题目的原因我们无法得知，在真实情况中我们应该考虑诸多变量仔细分析后才能得出精确的模型。