

**《数据科学与Python语言》**

**课程项目报告**

项目题目： 摩擦风速模拟

组 名： 二进制

组 员： 王浩、肖美昊

学 号： 2020204246、2020204296

专 业： 计算机科学与技术

班 级： 2020级图灵班

**二〇二二年十二月**

目录

[项目课题：摩擦风速模拟 1](#_Toc8640)

[一、 摘要 1](#_Toc8456)

[课题概括与要求： 1](#_Toc22604)

[关键字 1](#_Toc153)

[二、研究背景 1](#_Toc5397)

[三、研究方法和数据 3](#_Toc2157)

[1.相关学习 3](#_Toc20785)

[2.数据预处理 3](#_Toc12011)

[3. 划分数据集 4](#_Toc19542)

[四、研究结果 5](#_Toc29758)

[1.算法模型 5](#_Toc20544)

[2.参数优化 11](#_Toc1203)

[3.可视化操作 14](#_Toc3132)

[五、研究的不足 15](#_Toc67)

[六、成员分工 16](#_Toc23221)

[王浩-50% 16](#_Toc2187)

[肖美昊-50% 17](#_Toc19364)

[七、结论 17](#_Toc9153)

[项目效果总结分析： 17](#_Toc929)

[项目总结 18](#_Toc1548)

[八、参考文献 19](#_Toc18586)

# 项目课题：摩擦风速模拟

# 一、摘要

本文对摩擦风速进行了回归预测的分析，采取了lightGBM，随机森林，XGBboost，SVR，决策树等机器模型研究方法，并进行调参，最后对其进行性能比较，选出效果更为优异的机器学习模型。

## 课题概括与要求：

摩擦风速（USTAR）是描述能量在地表流动难易程度的参数。准确模拟USTAR对于准确描述地表能量能量流动具有重要意义。该研究的主要内容包括：（1）随机选取课程提供的野外观测站中50%站点的数据为实验样本，对数据进行预处理，剔除无效和低质量样本；（2）划分训练集、验证集和测试集；（3）使用机器学习或深度学习算法建立模拟潜热通量的模型，要求至少对比3种机器学习模型，并对模型的参数进行优化；（3）为避免过拟合，需要筛选合适变量，剔除贡献率较低的变量，并分析各剩余变量对目标值的贡献率。

关键字: 回归预测,随机森林,LightGBM,XGBboost,决策树,地表能量流动,风速

# 二、研究背景

陆面上不同下垫面地气之间物理量的交换是不同的,从而形成了不同的天气和气候现象,这是一个比较复杂的陆面物理过程,所以研究发生在陆地表面的各种过程是非常重要的。而地球表面覆被和空气动力学特征直接影响着地气系统物质、能量的输送和局地气候。计算和分析不同下垫面的空气动力学参数就成了研究陆面过程的关键。然而人们同时对多种下垫面空气动力学特征的比较研究较少有的只是做了单一农田下垫面空气动力学参数的研究特别是受不同人为干扰,而对下垫面空气动力学特征的研究就更少,尤其是同时研究沙丘、无人为干扰草原、轻度人为干扰草原、中度人为干扰草原重度人为干扰草原、玉米、大豆、小麦和稻田下垫面的空气动力学特征在国际上更少见。本项目收集了 40 个不同地点的野外监测站点的观测数据，以 CSV 格式存储，这些文件均具有相同的变量，但每个监测站点的数据条目数量不一样。这些数据中，还统一使用了-9999 来表示无效值。使用这些大量的数据进行整理与分析，相信我们能够在一定程度上对USTAR进行准确预测，这对于准确描述地表能量能量流动具有重要意义。

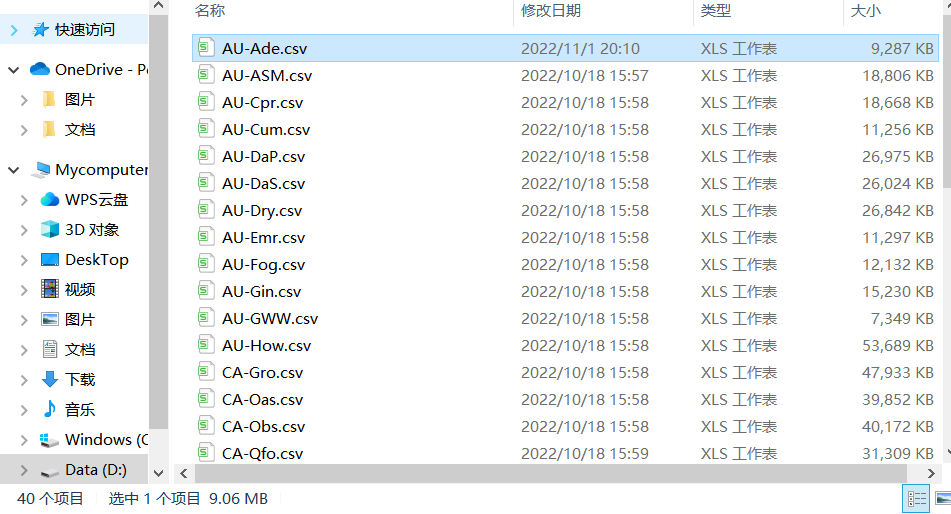


图2.1 40个不同地点的野外监测站点的观测数据

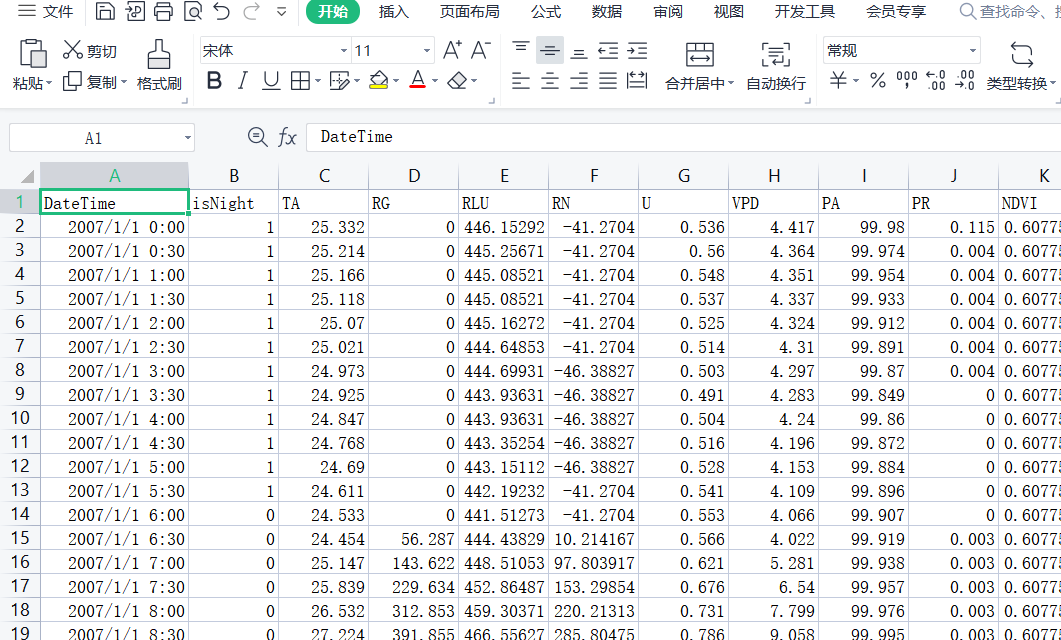


图2.2 数据格式

# 三、研究方法和数据

### 1.相关学习

我们小组事先学习了numpy库，pandas库，matplotlib以及 scikit-learn等一系列的在机器学习中能够发挥用处的第三方库，通过这些库，我们能够更好的将数据进行整理，以及运用到相关的算法模型中。

对论文，我们小组阅读了《不同下垫面空气动力学参数的研究》[1] 这一篇论文，了解到了不同下垫面对于空气动力学参数的影响，以及每个月份风速的大小都将有所不同，这一点在后续的模型参数分析中将有一定的参考价值。

对风的形成，我们了解到高空的风是在水平气压梯度力和地转偏向力共同作用下形成的，风向与等压线平行。而近地面的风是在水平气压梯度力，地转偏向力，摩擦力共同作用下形成的，风向与等压线斜交。故近地面的风所受到的影响是更多因素的，这一点也会在后续模拟风速预测中有所涉及。

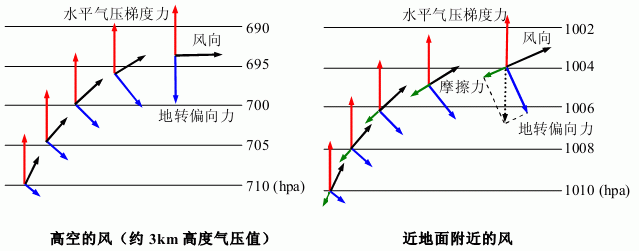


图3.1.1 高空的风与近地面附近的风

### 2.数据预处理

1.我们小组先使用了pandas库read\_csv函数读取相关观测站的数据信息。

2.利用concat函数，构建项目中所需的新的DataFrame形式。

3.分析数据绘制数据直方图,箱线图等分析数据分布

4.后利用质量评估字段去掉无效值，利用箱线法去掉异常值，查看是否存有空值以及nan，分析数据相关性去除相关性高的数据。

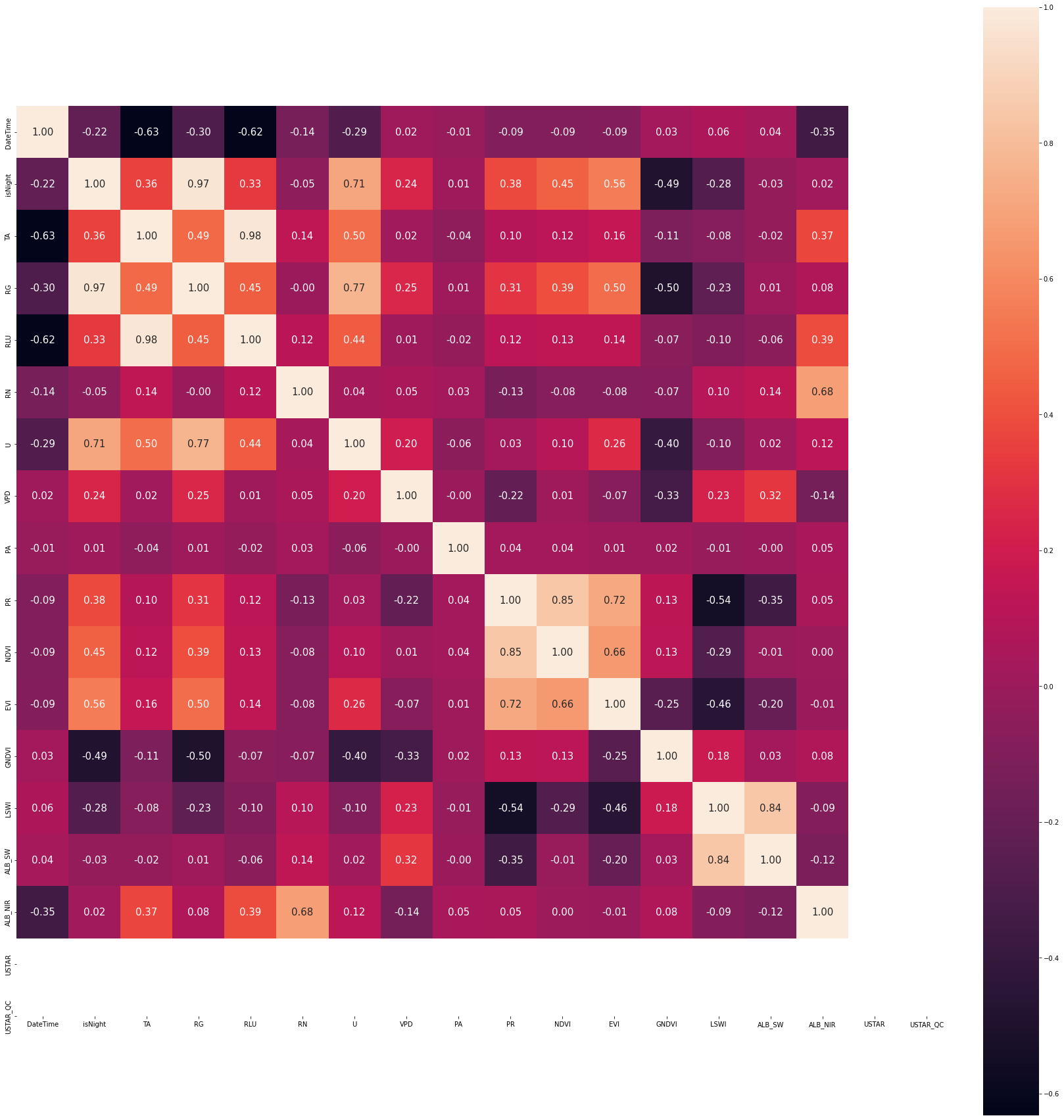


图3.2.1 数据正相关分析图

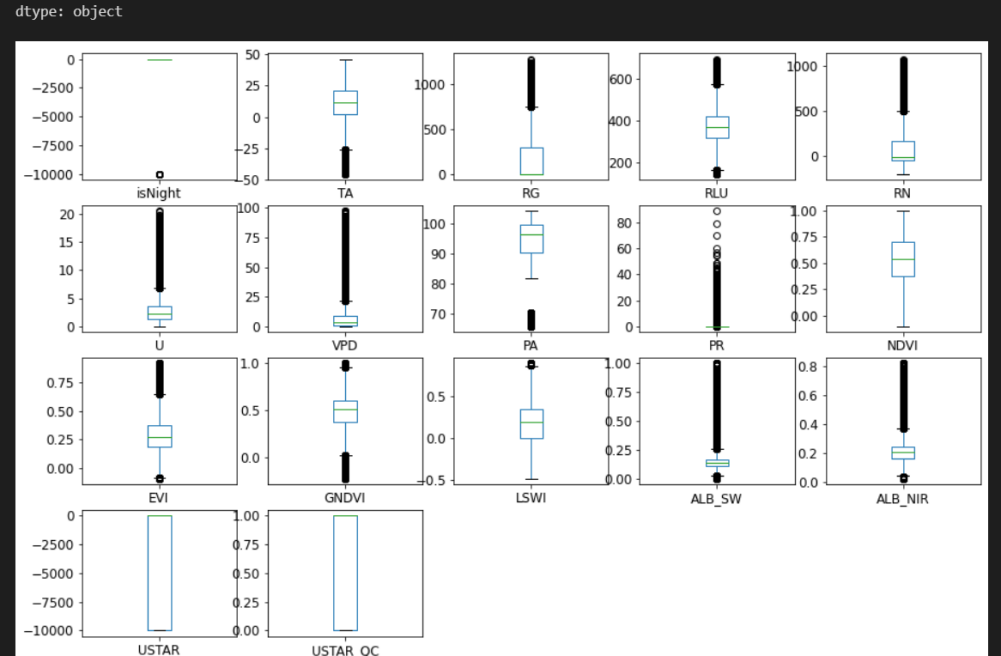


图3.2.2 箱线图分析数据

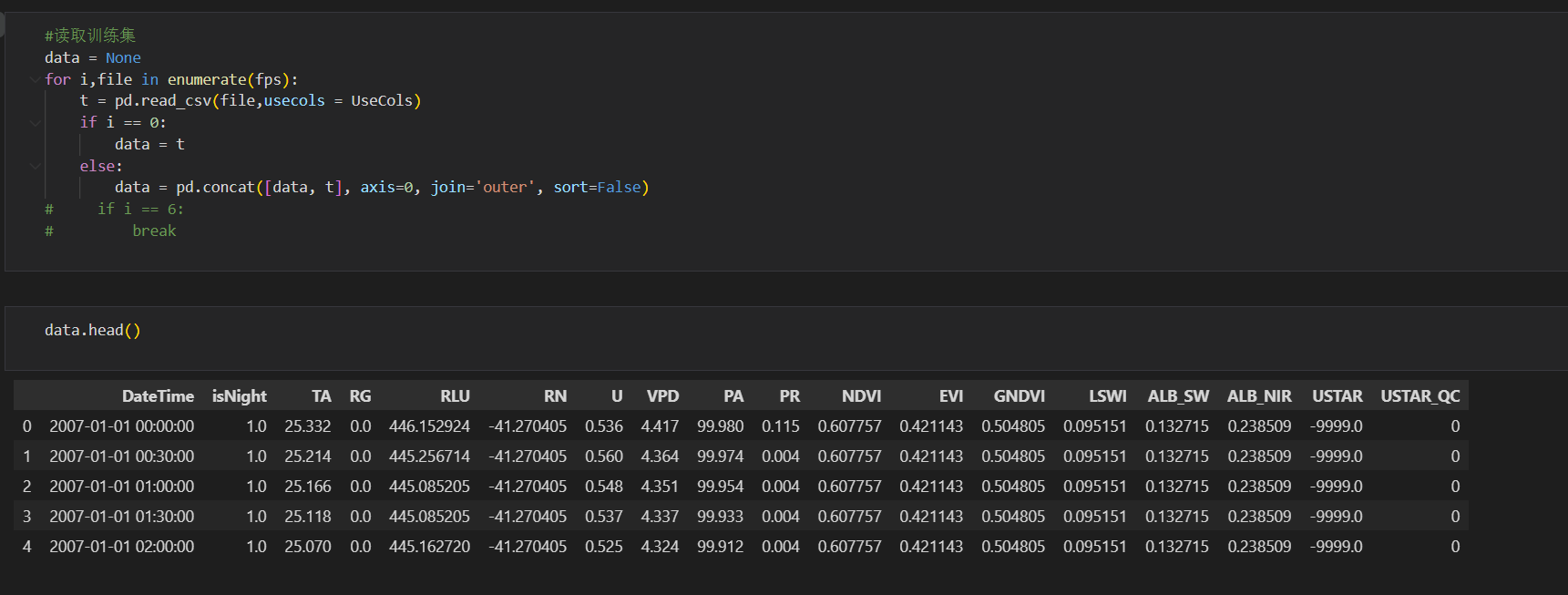


图3.2.3 使用pandas库读取数据

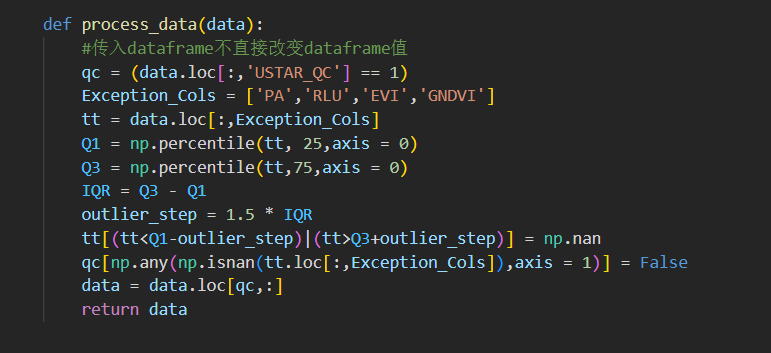


图3.2.4 剔除无效和低质量样本

### 3.划分数据集

该项目划分60%的训练集，20%的验证集，20%的测试集

20%验证集(Validation set):对学习出来的模型，微调分类器的参数。验证集还用来确定网络结构或者控制模型复杂程度的参数。

20%测试集(test set):检验最终选择最优的模型的性能如何。

60%训练集(Training set)：学习样本数据集，通过匹配一些参数来建立一个分类器。建立一种分类的方式，主要是用来训练模型的。

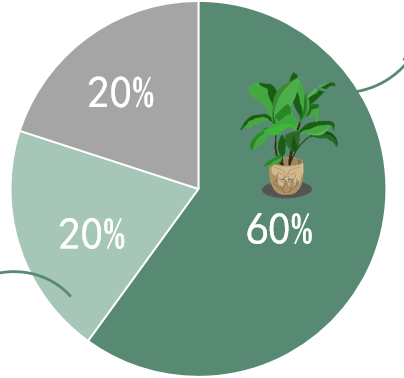
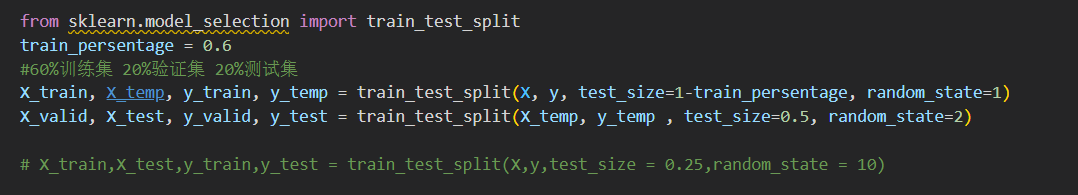
 

图 3.3.1 划分的比例 图3.3.2 划分60%训练集，20%验证集，20%测试集代码

# 四、研究结果

### 1.算法模型

在项目实践中，我们使用过了lightGBM，随机森林，XGBboost，SVR，决策树等算法模型，对摩擦风速值进行预测，并通过对比分析，得出分析本问题模拟效果更为优秀的模型，得出较为有效准确的摩擦风速值，对描述当地的地表能量能量流动具有重要意义。

**(1)lightGBM：**

LightGBM（Light Gradient Boosting Machine）是一款基于决策树算法的分布式梯度提升框架,支持高效率的并行训练，并且具有更快的训练速度、更低的内存消耗、更好的准确率、支持分布式可以快速处理海量数据等优点。

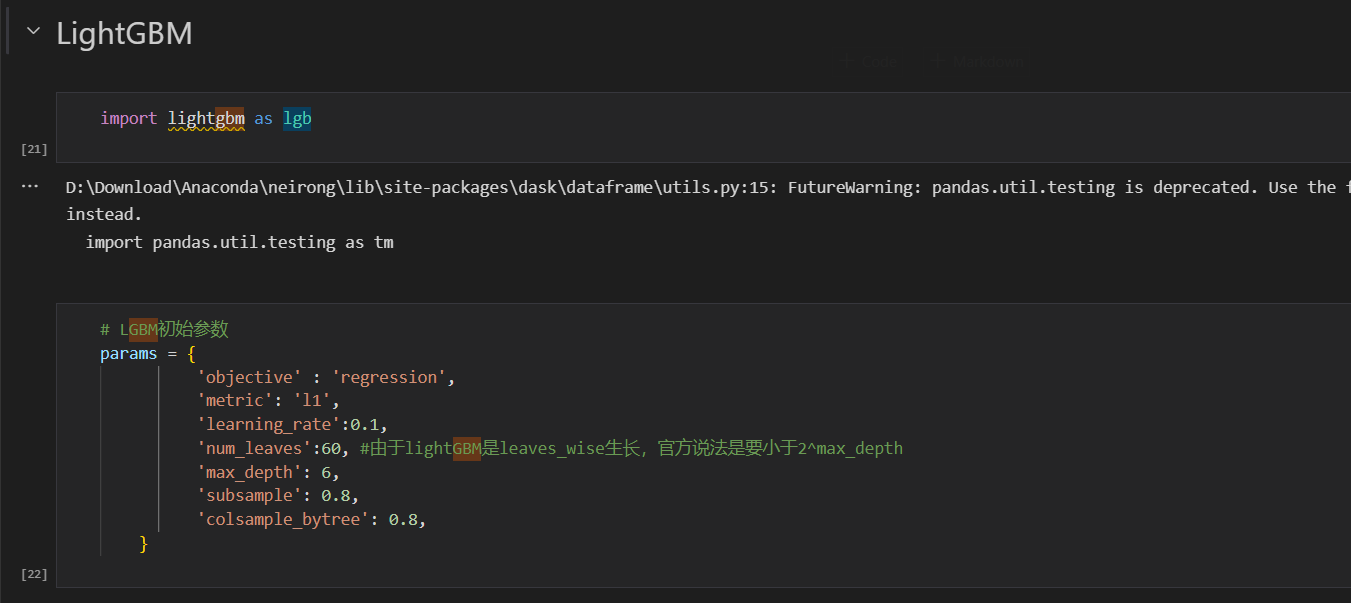


图4.1.1 LightGBM的导入与参数设置

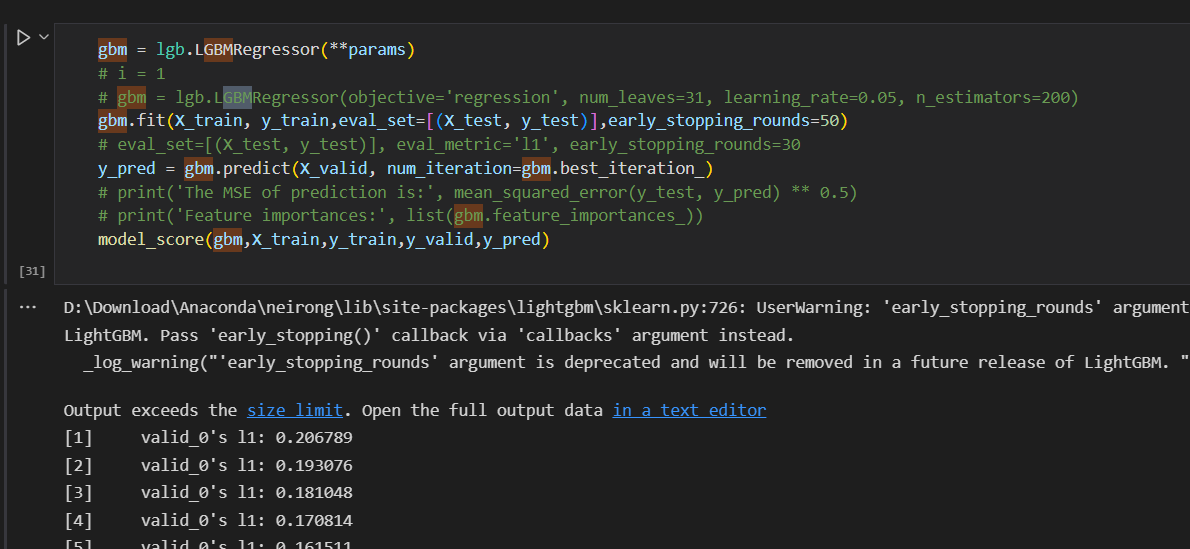


图4.1.2 LightGBM的使用

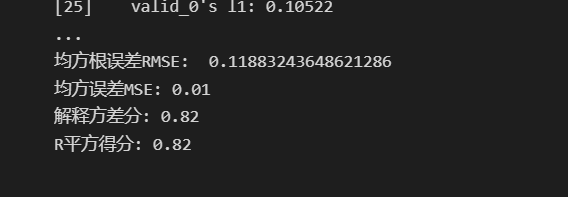


图4.1.3 LightGBM的初始参数的评测

**(2)XGBboost：**

XGBoost(eXtreme Gradient Boosting)是经过优化的分布式梯度提升库，旨在高效、灵活且可移植。XGBoost是大规模并行boosting tree的工具，它是目前最快最好的开源 boosting tree工具包，比常见的工具包快10倍以上。

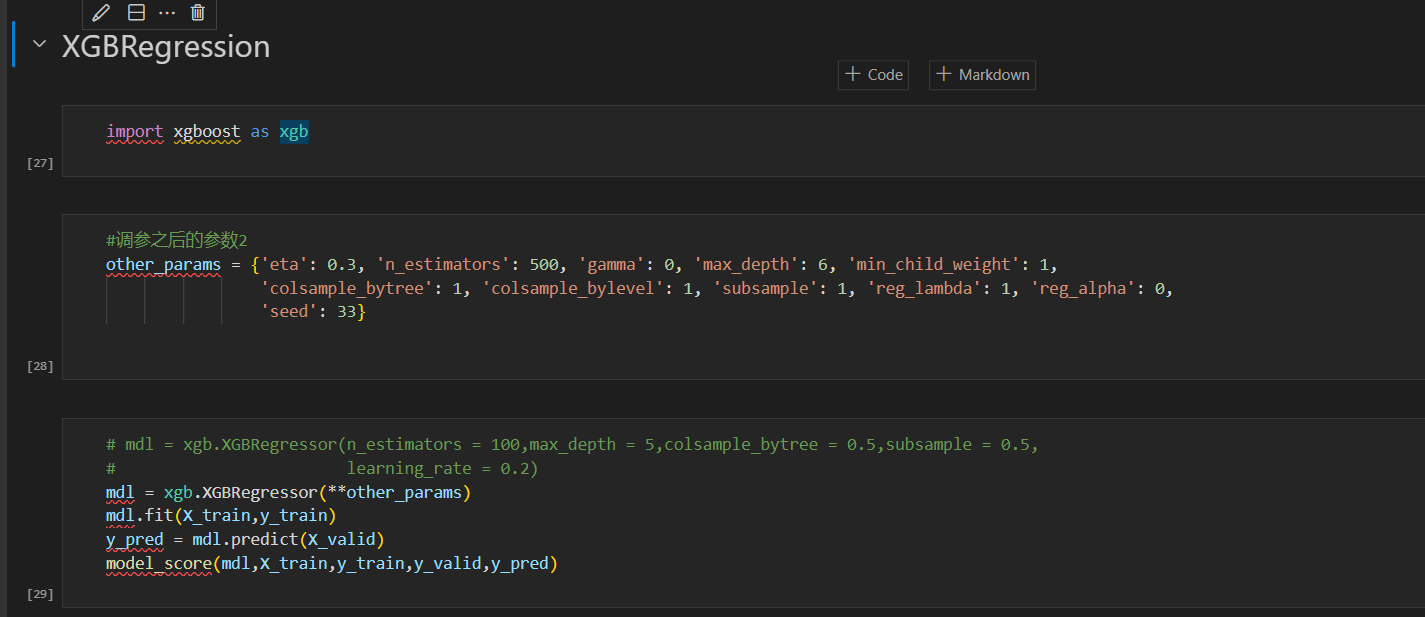


图4.1.4 XGBboost的导入与使用

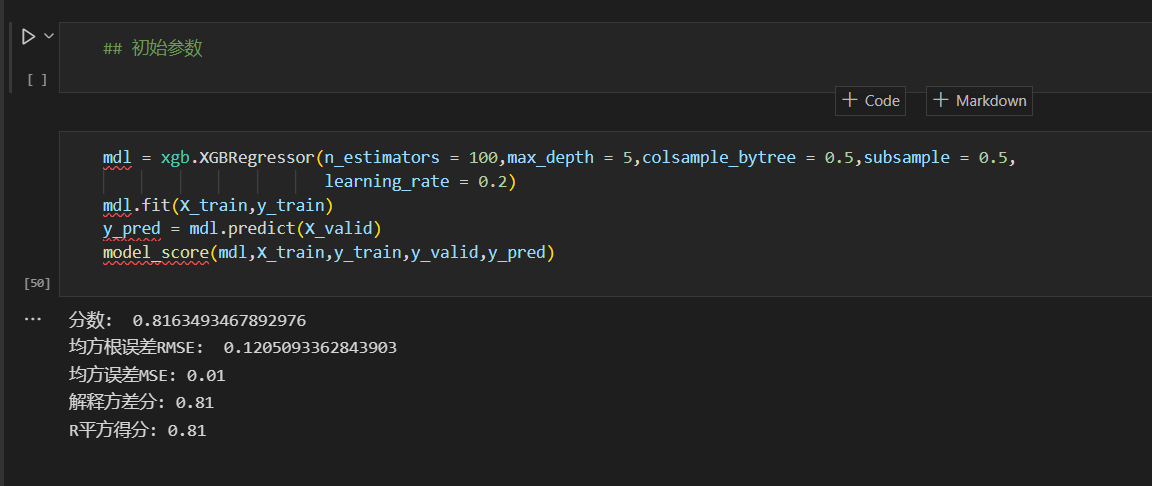


图4.1.5 XGBboost的初始参数的评测

**(3)SVR：**

SVR（Support Vactor Regression）支持向量回归机所寻求的最优超平面不是SVM那样使两类或多类样本点分的“最开”，而是使所有的样本点离着超平面的总偏差最小。即SVM是要使到超平面最近的样本点的“距离”最大； SVR则是要使到超平面最远的样本点的“距离”最小。

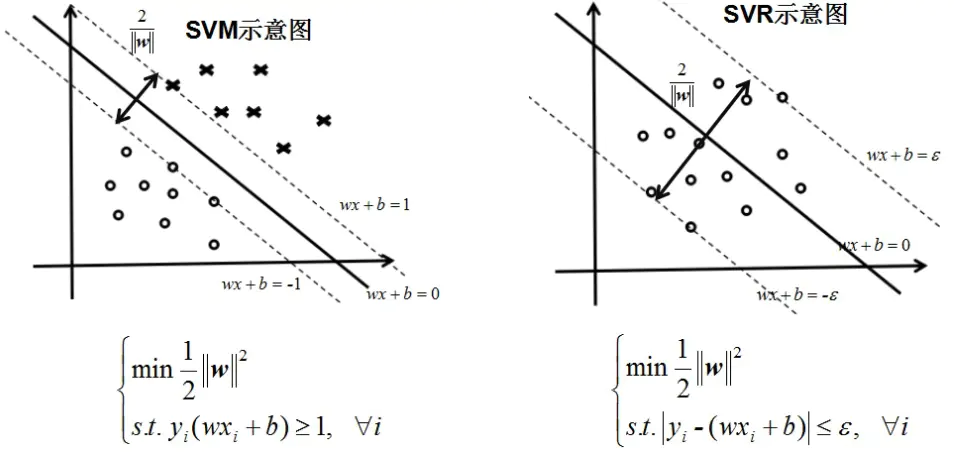


图4.1.6 SVM/SVR图示

但是在实际的应用中，发现SVR适用于小数据集，不能跑大量的数据，在大量的数据集中，运行时间过长，预测出的效果较差，故总结为一次失败的应用尝试



图 4.1.7 SVR的导入

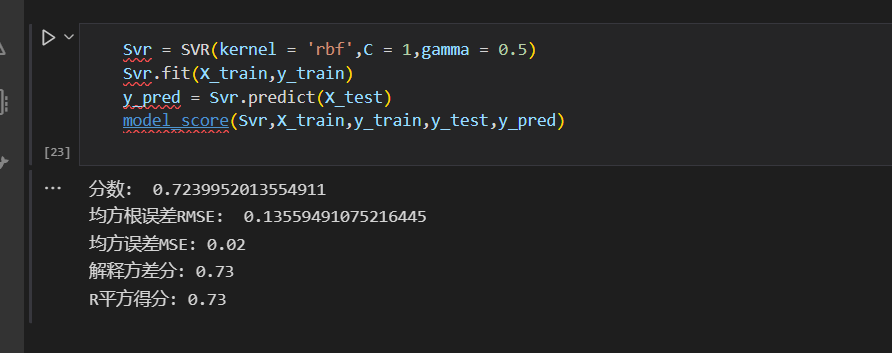


图 4.1.8 SVR的参数设定与测评

**(4)随机森林：**

随机森林(Random Forest)是一个包含多个[决策树](https://baike.baidu.com/item/%E5%86%B3%E7%AD%96%E6%A0%91/10377049?fromModule=lemma_inlink)的[分类器](https://baike.baidu.com/item/%E5%88%86%E7%B1%BB%E5%99%A8?fromModule=lemma_inlink)，其输出的是由其多个树输出的类别的[众数](https://baike.baidu.com/item/%E4%BC%97%E6%95%B0/44796?fromModule=lemma_inlink)而确定的。

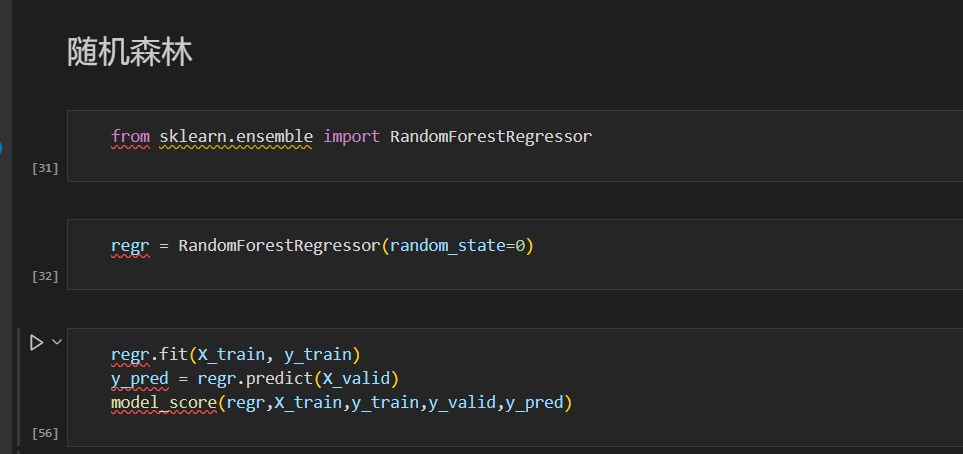


图4.1.9 随机森林的导入

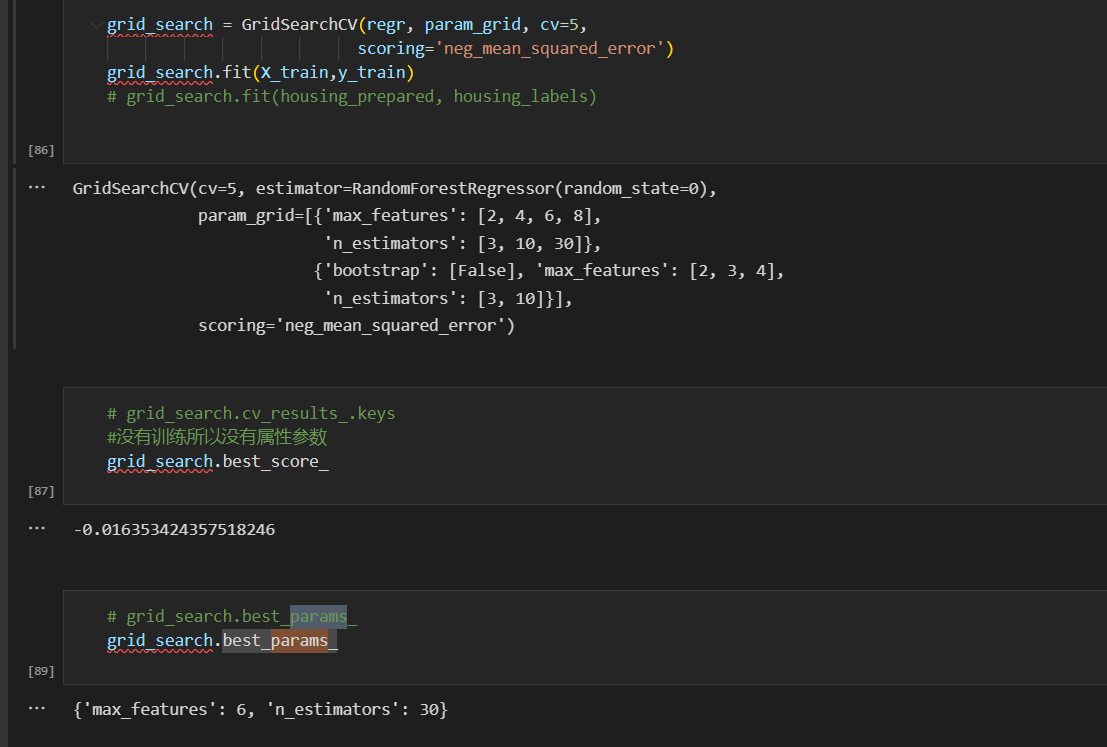


图4.1.10 网格化搜索找到随机森林模型的最佳参数

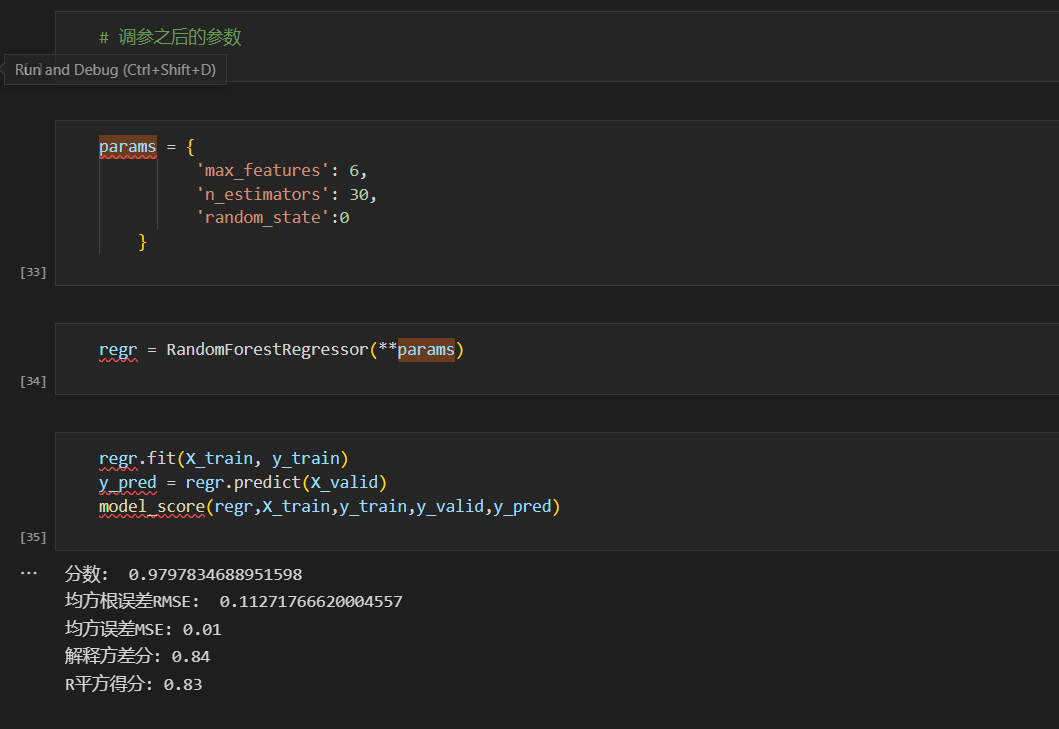


图4.1.11 随机森林的参数设定与测评

**(5)决策树：**

决策树(Decision Tree),又称为判定树, 是一种以树结构(包括二叉树和多叉树)形式表达的预测分析模型. 通过把实例从根节点排列到某个叶子节点来分类实例。

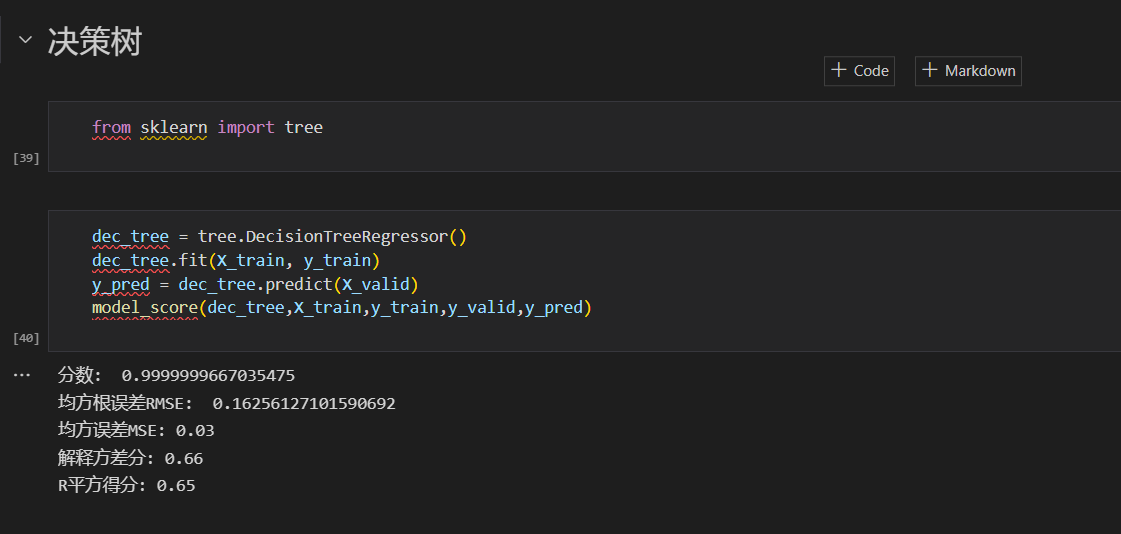


图4.1.12 决策树的参数(默认参数)测评

### 2.参数优化

在变量选择方面，1.不同的研究问题往往变量的选择是不同的，不能通过对变量的统一设置进行选择。需要做到具体问题具体分析，2.且回归模型的解释变量选取个数上并不是越多越好，往往需要精选出有价值的且信息重叠度较小的变量，防止出现共线性问题。

故一般在选取指标变量时往往不能一次完全正确，需要反复调整找到最合适的指标变量，提高模型的精度。

首先，我们组通过可视化的方式比较各变量的贡献程度，进行解释变量的选择。即根据模型参数的重要性，筛选特征进行训练。

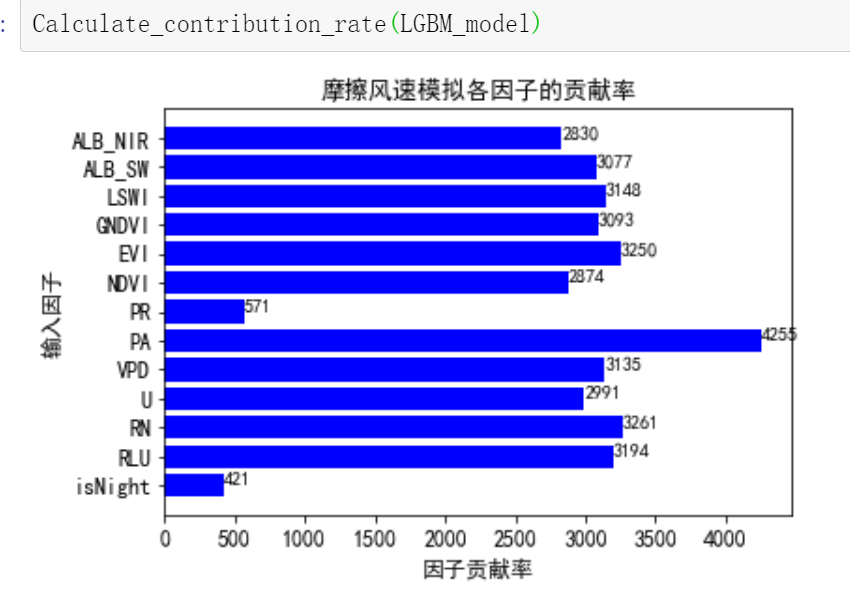


图4.2.1 LGBM模型的参数贡献率

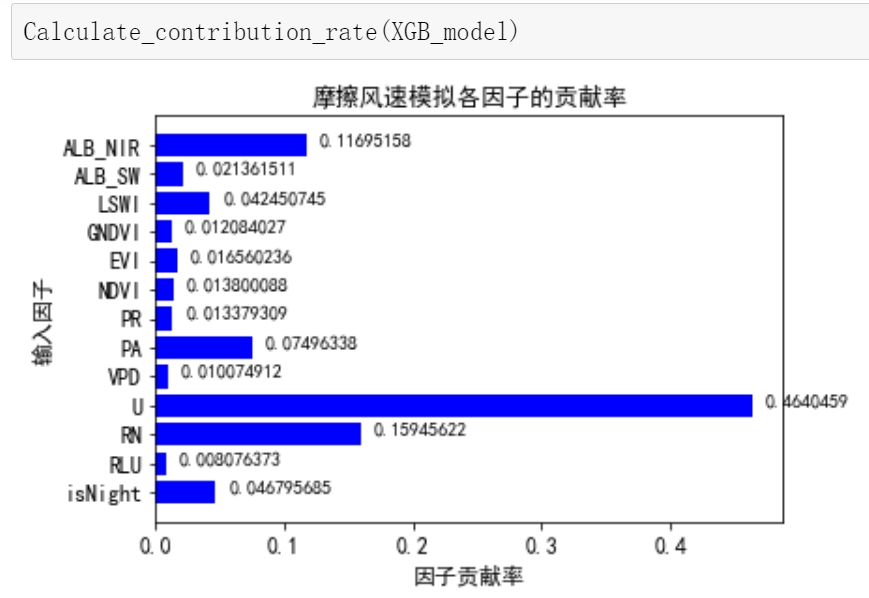


图4.2.2 XGB模型的参数贡献率

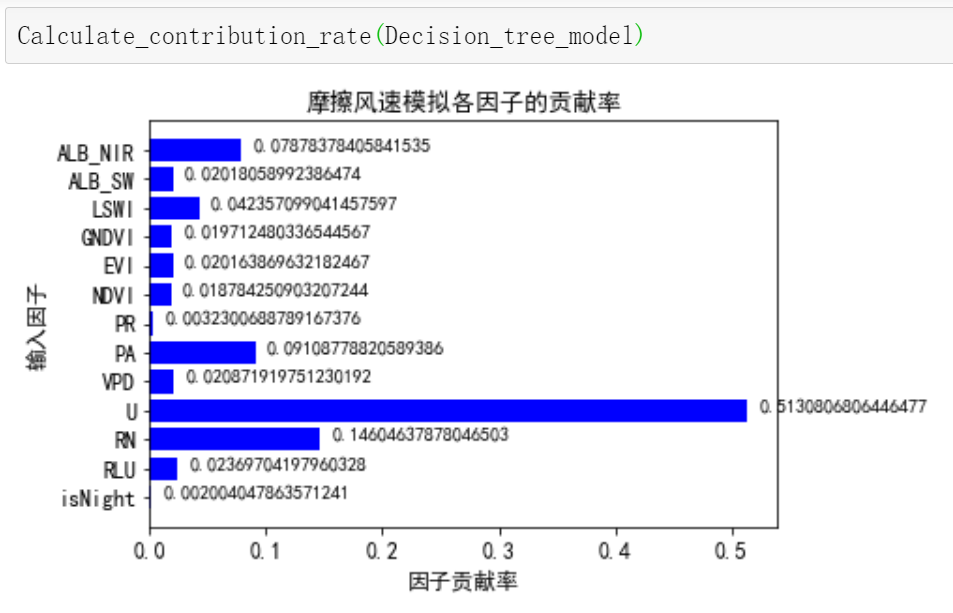


图4.2.3 决策树模型的参数贡献率

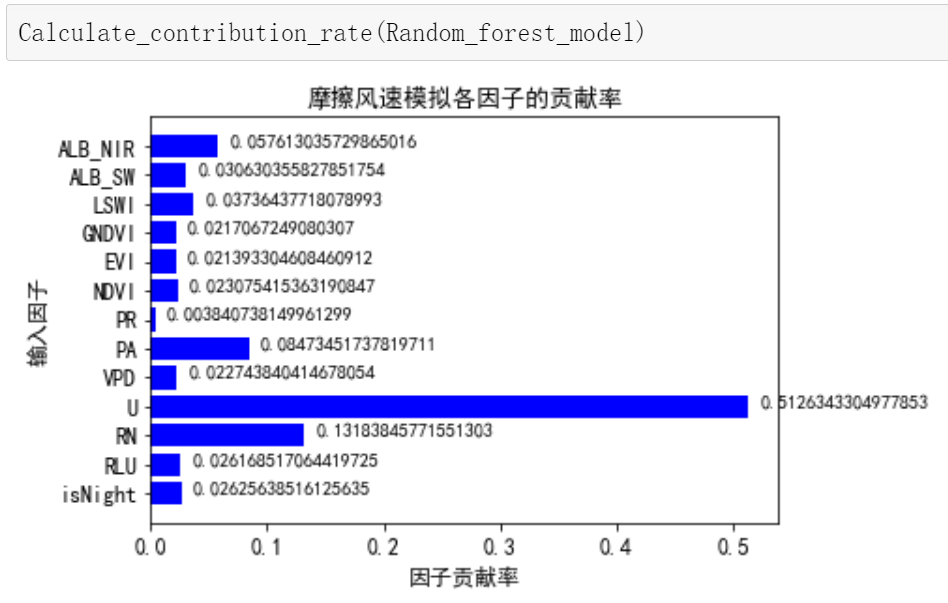


图4.2.4 随机森林模型的参数贡献率

后又编写了对模型评估函数，方便在后续的参数调节中，更好的比较模型性能。参数优化部分采用GridSearchCV进行网格化搜索，后因为电脑设备配置，数据量大等问题，采用手动方式进行调整。有效程度能够达到90%。但又因为数据每次都是随机化读取的，导致之后训练的模型评分降低到87%，部分的性能下降。

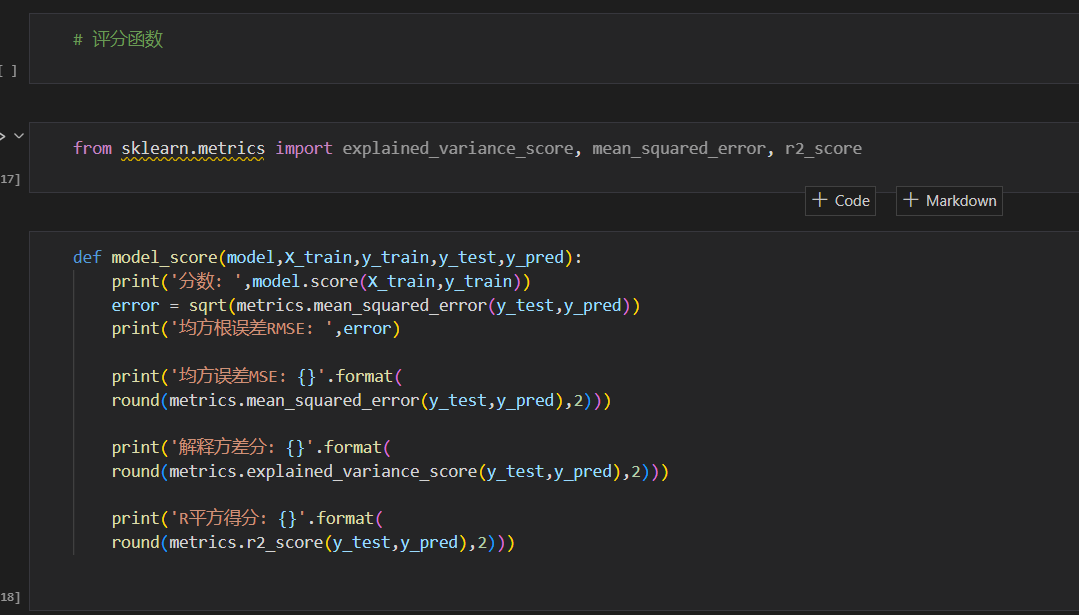


图 4.2.3 评估函数的编写

例如，在XGB模型中，我们进行了多次参数的调整，通过评估函数的比较，选出了性能更为优异的参数2。

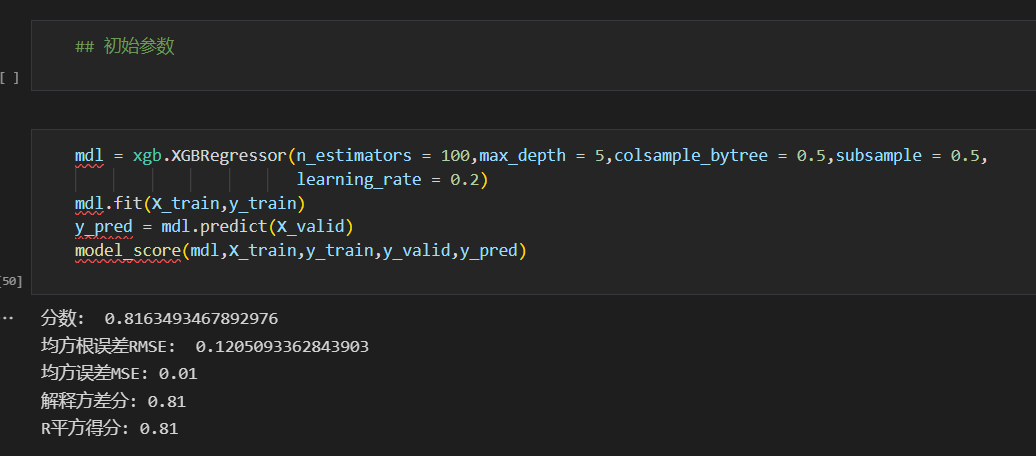


图 4.2.4 XGB初始参数

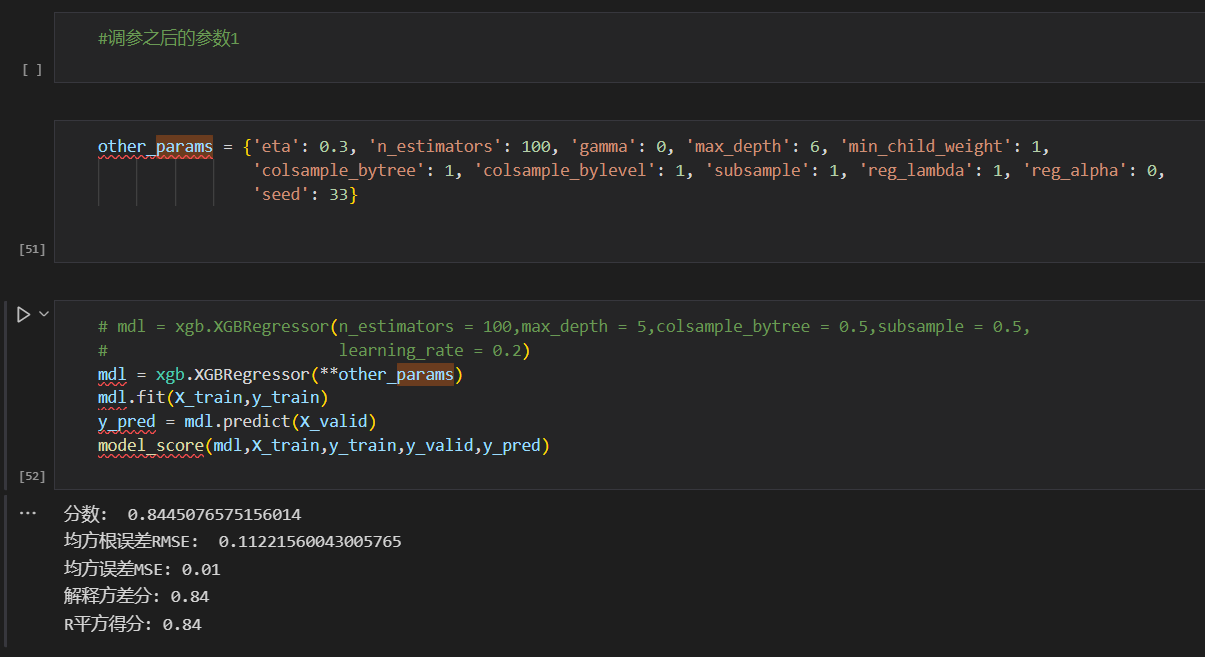


图 4.2.5 XGB参数调整1

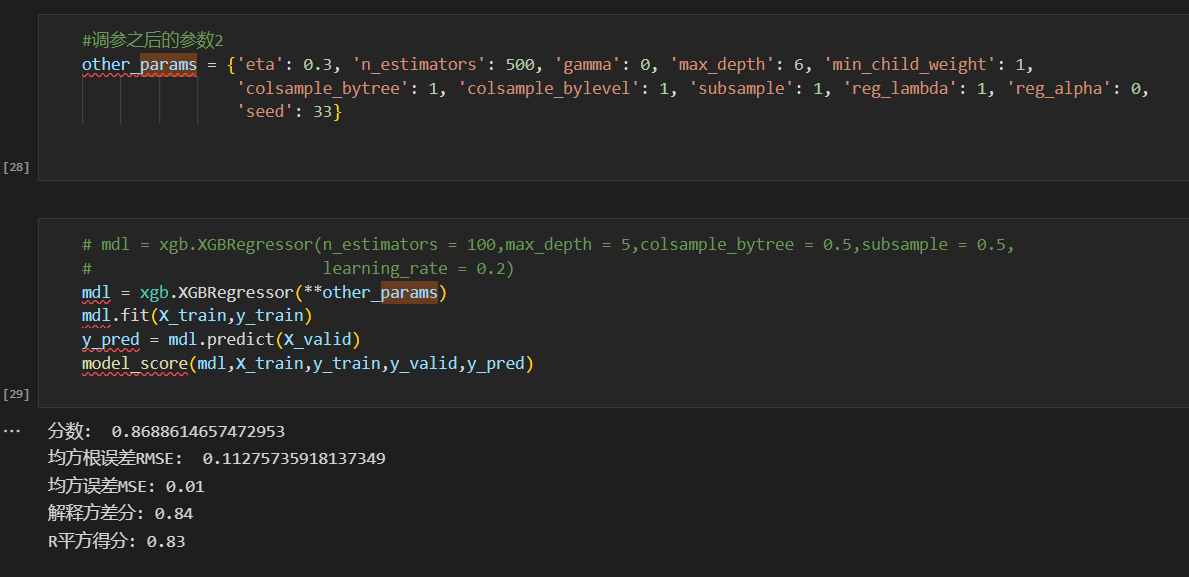


图 4.2.6 XGB参数调整2

### 3.可视化操作

在本次项目的实践中，我们在多次的地方运用到了可视化的操作，帮助我们更好的分析数据，观察其内在逻辑，有效地增强了数据交互性。

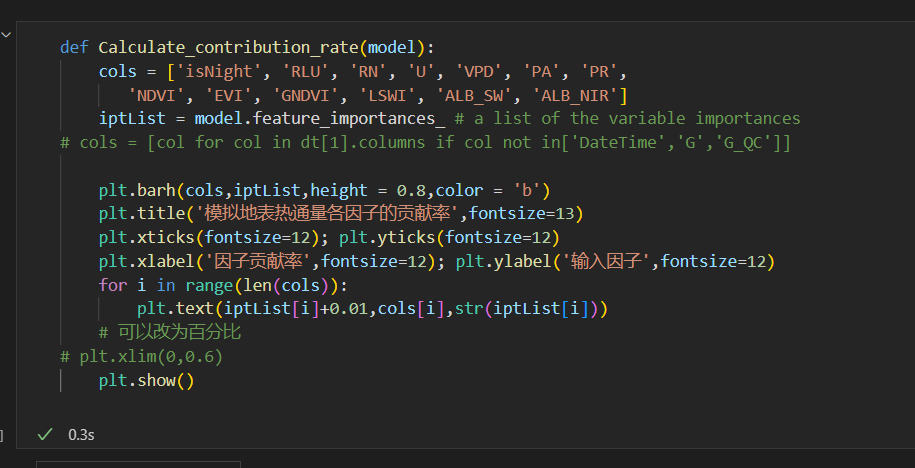
例1：通过可视化，直观选择贡献率高的变量

图 4.3.1 关于因子贡献率的部分相关代码

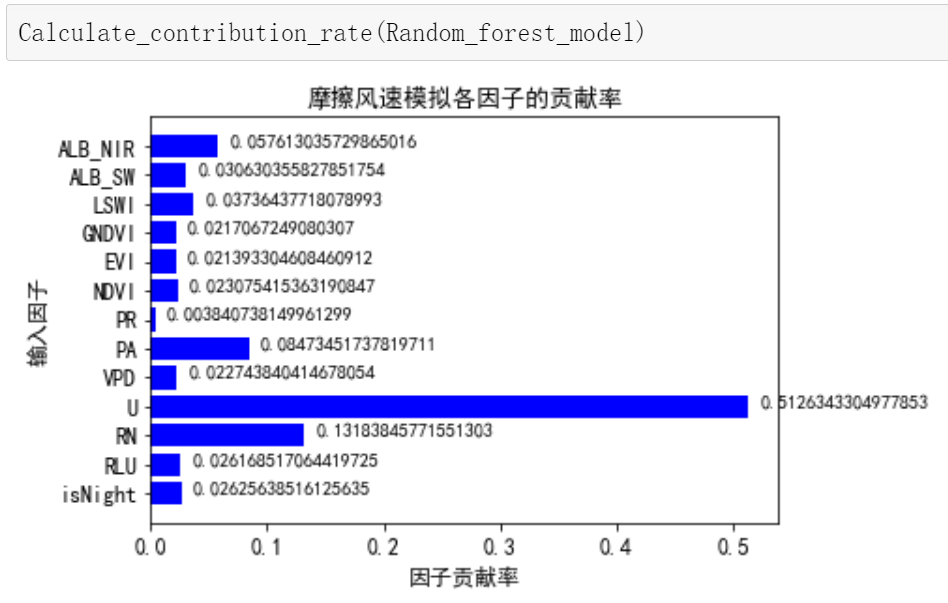


图4.3.2 随机森林模型的参数贡献率

# 五、研究的不足

1.在项目的一开始是制定了使用很多的模型的，想要使用神经网络模型，但是后期因为时间问题亦或是难度问题没有实现，感觉很是可惜。

2.在可视化方面的界面构成不是很美观，有一小部分出现了些小瑕疵，例如，在title上的命名是延续了之前的选题，没有及时改过来。且感觉可以在更多的方面实现可视化操作，甚至可以再多一些小的细节，使我们的项目更加立体。

3.模型的选择上也不是一帆风顺的，例如在本次项目中使用过的SVR模型适用于小数据集，我们却用它来跑大量的数据集，导致运行时间过长，模型精度也不高，预测出的效果较差，故总结为一次失败的应用尝试。

4.在代码的运行中出现了许多的warning，虽不影响最终的效果，但是感觉代码还没有写到最好亦或是有部分没有理解，感觉是为一处不足。

5.项目与论文的结合也不是很密切，对于茅宇豪教授的这篇《不同下垫面空气动力学参数的研究》论文，做到了通读但是没有很理解其中的推导出的数学公式，感觉还是读起来较为吃力，只是理解了一些常识或者是一些更为通俗易懂的解释，但是也算是从中得到了一些对于课题分析的帮助，有一定的价值。

# 六、成员分工

本项目任务：预测模拟 USTAR(摩擦风速)，

项目小组：二进制，由王浩、肖美昊两人组成。

项目开始于2022/11/16

初步完成于2022/12/14。

### 王浩-50%

完成工作：模型确定，使用机器学习或深度学习算法建立模型，对比了多种机器学习模型，并对模型的参数进行相应的优化。负责了代码的实现。

### 肖美昊-50%

完成工作：收集数据，对数据的读取与规范化整理。设计研究内容、方法、研究过程和结果，以及小组成员的分工，论文等文案的编写。帮助编写相关代码

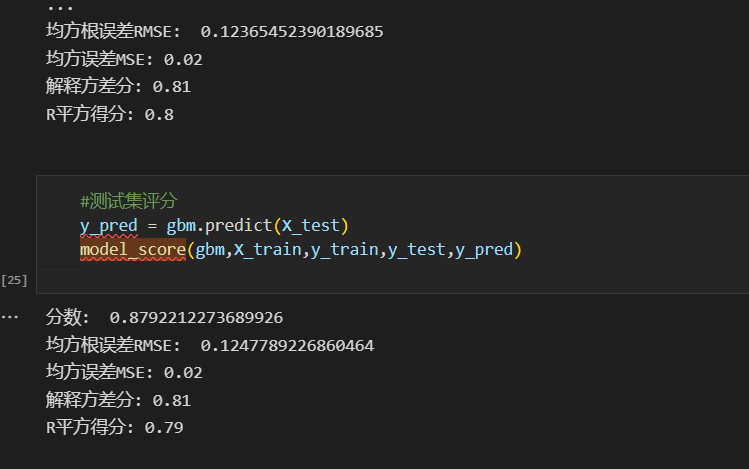
# 七、结论

### 项目效果总结分析：

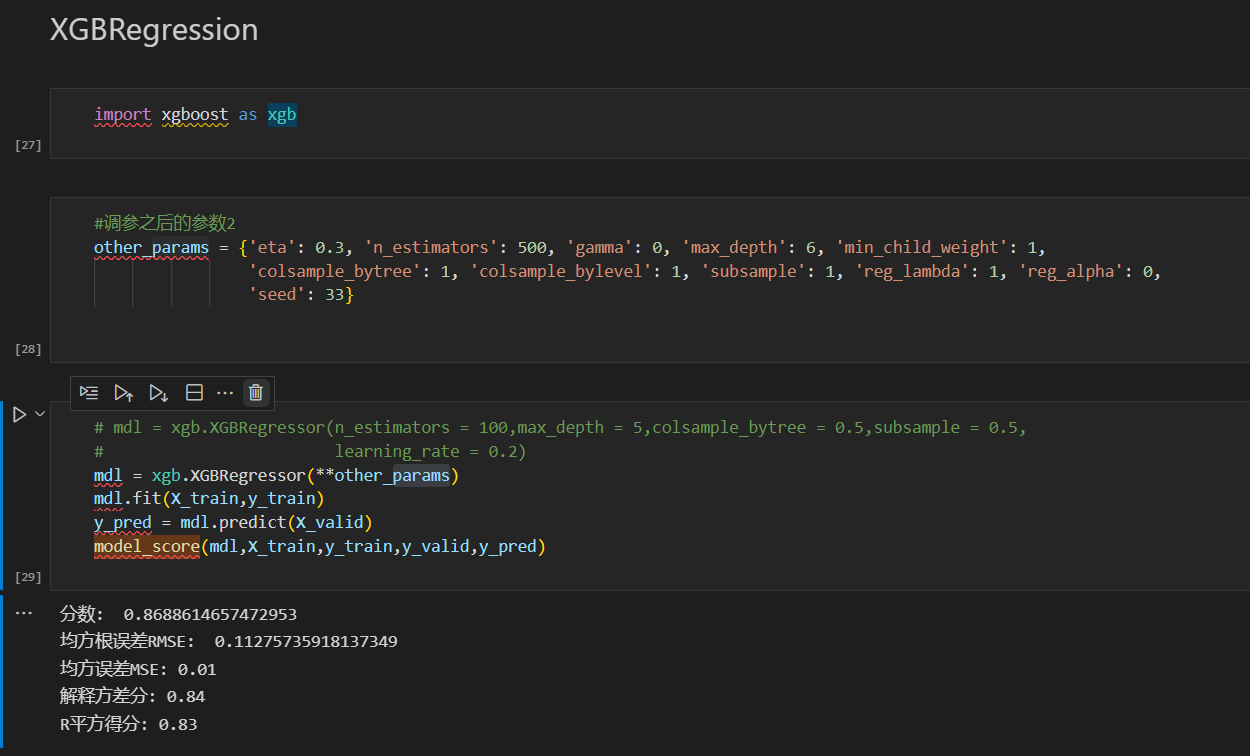
在本项目中，

我们可以得到(按调试出的最优参数，最高性能进行比较)，

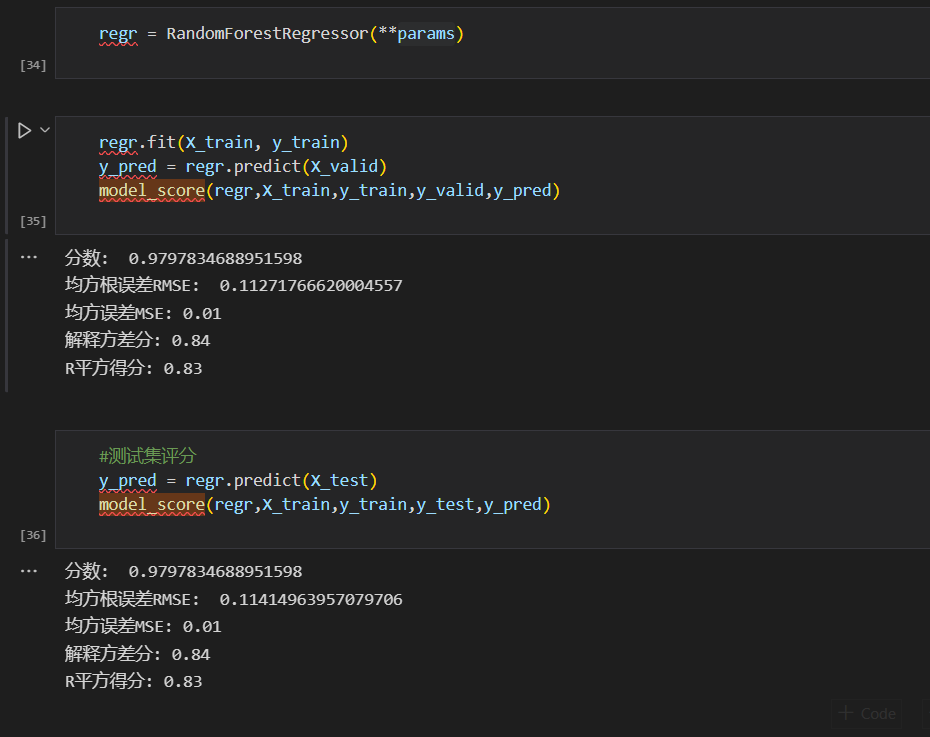
Lgb：



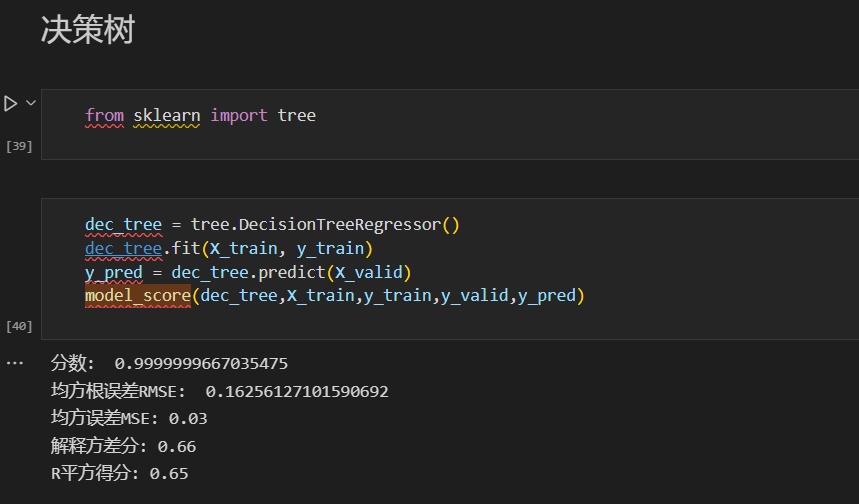
XGB：



随机森林：



决策树：



得到，最后预测效果最好的为决策树模型，可以较好的模拟出本次项目所要预测的风速值。

### 项目体会与感悟

1.项目制作阶段,因为采用小组合作进行,面临着信息沟通,协作的问题,使一些较简单的问题,解决起来效率较低,以后的小组合作过程中需要进行借鉴。

2.由于电脑配置,数据量较大等问题,在进行网格化搜索的过程中,面临着训练时间过长的问题,效率较低的问题,而一个模型有多个变量,不同变量之间可能会相互影响,使得如果网格化搜索单一参数,找到的最优结果可能并不能使模型的准确度提升,因此在部分模型参数调整过程中,我们采用手动调参的方式,进行不断地调试不同参数，同时将前面同学设置的参数进行合理化更改，使得所添的元素不显得突兀。在这过程中，我们团队合作的能力得到了很好的提升。

3.对论文的学习，让我们在后续的实验中，对合理变量的选取也有更大的把握，认识到阅读一篇有价值的文章，需要静下心来仔细阅读，无论是从中学习到的知识点亦或是其学术严谨的态度，这都对我们大有裨益。

4.在项目的实践中，总不会是一帆风顺的，例如，在对数据进行读取处理，归一化等等一系列操作，就需要弄清楚每个模型的参数是什么，需要怎样格式的参数，如何读取数据等一系列问题。往往会出现许多的困难，但是需要我们同心协力，并一直做下去，总会实现预期的目标。不积跬步，无以至千里；不积小流，无以成江海。一个大的项目是持之以恒的来进行实现的，他不是一日之功。

5.在书本上获取的知识感觉只有实际运用起来才会是自己的,必须去躬行实践。愈发认识到计算机这门专业是偏向于实践的，而非理论的科目。将书本上学到的知识应用到实际项目中，做到理论联系实际，学用结合，这是一个很重要的能力 。

# 八、参考文献

［1］茅宇豪,刘树华,李婧.不同下垫面空气动力学参数的研究

［2］任国玉;郭军;徐铭志等.近50年中国地面气候变化基本特征

［3］王家伟,佟柠.“情境教学”+“受力分析”模拟图析“风”的形成

［4］张超宇,鲁爱华.“大气的水平运动——风”教学设计

［5］何健中.湍流边界层中平流与大气运动的动能耗散

［6］[ObermannHellhund Anika](https://scholar.cnki.net/home/search?sw=6&sw-input=ObermannHellhund%20Anika) . State of the Simulation of Mesoscale Winds in the Mediterranean and Opportunities for Improvements

［7］[Eguchi Yuzuru;](https://scholar.cnki.net/home/search?sw=6&sw-input=Eguchi%20Yuzuru)[Nomura Mitsuharu;](https://scholar.cnki.net/home/search?sw=6&sw-input=Nomura%20Mitsuharu)[Hattori Yasuo](https://scholar.cnki.net/home/search?sw=6&sw-input=Hattori%20Yasuo) . Vector correlation between modeled gradient wind and observed high-altitude wind of a translating tropical cyclone

［8］[Eun Ji An;](https://scholar.cnki.net/home/search?sw=6&sw-input=Eun%20Ji%20An)[Hae Dong Kim](https://scholar.cnki.net/home/search?sw=6&sw-input=Hae%20Dong%20Kim). Characteristics of Mountain and Valley Winds Observed at Palgongsan Mountain and their Effects on Reducing High Temperatures in Summer