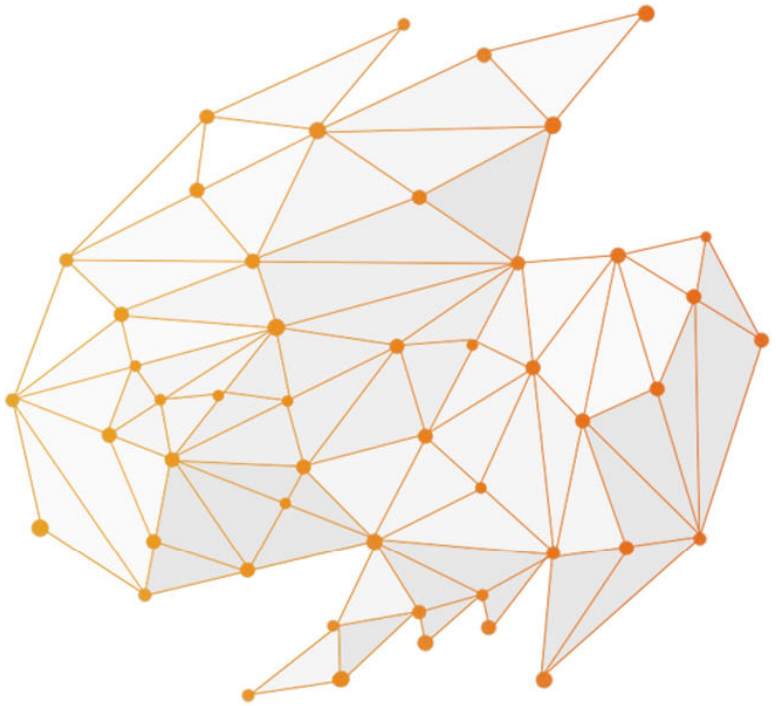
SUSTech Southern University of Science and Technology



大数据导论

—北京PM2.5回归 分类分析

目录

CATALOGUE

SUSTech Southern University of Science and Technology

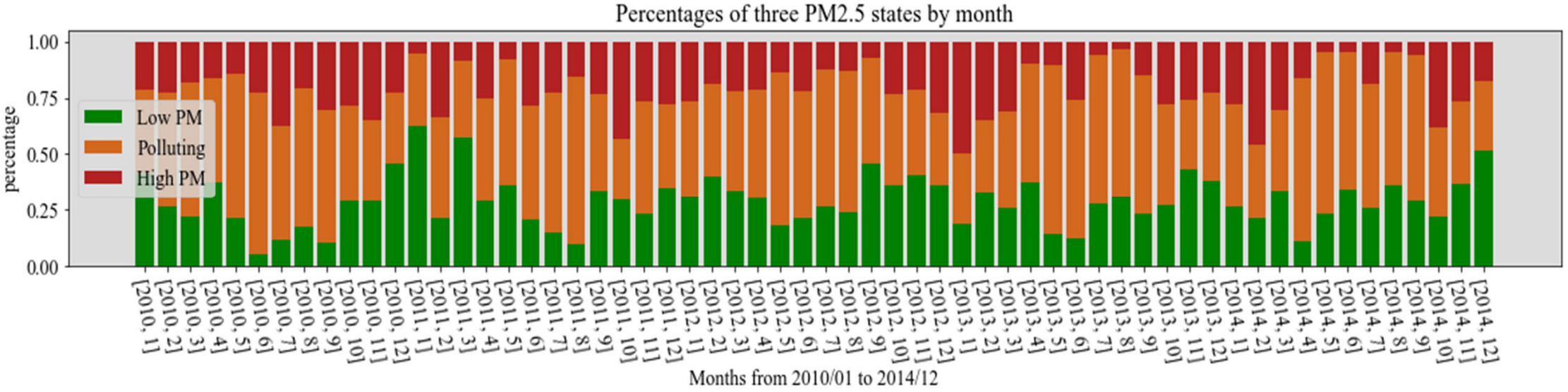
01

统计与分析

Statistics & Analysis

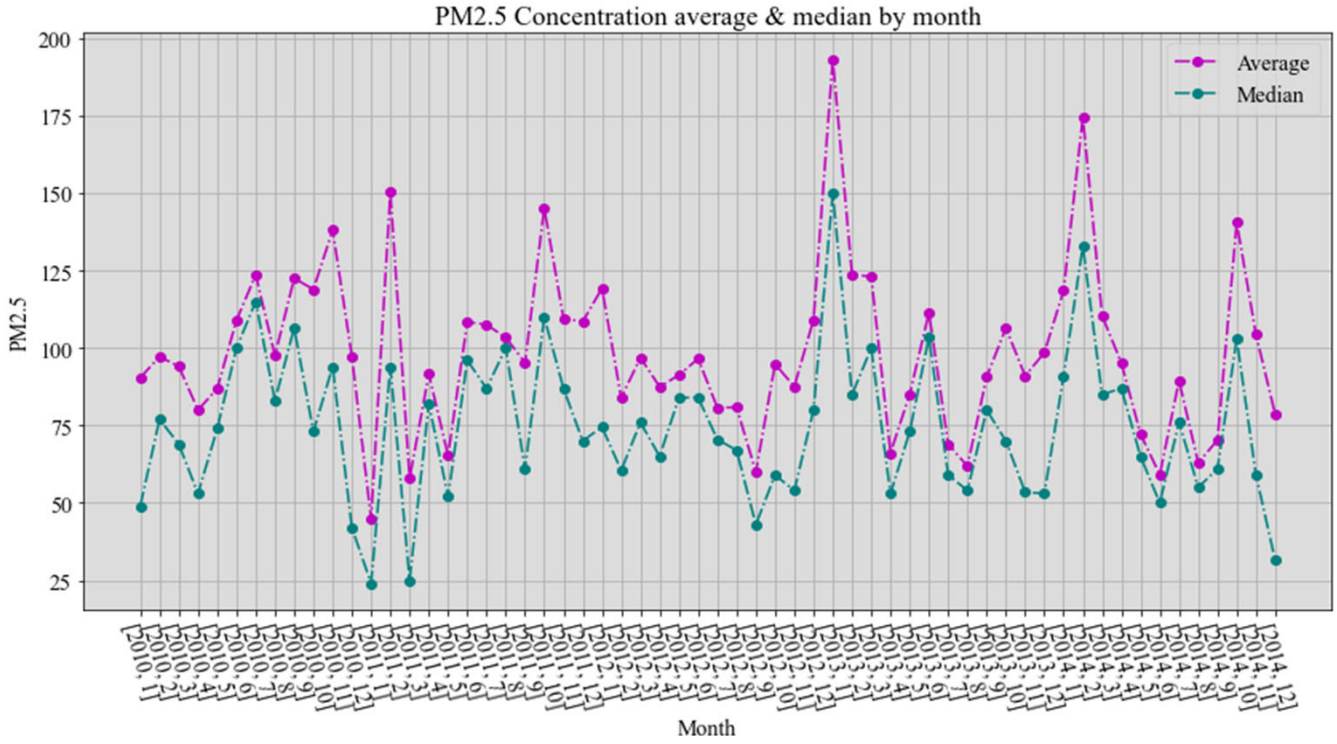
01/统计与分析：总体趋势（按月）

大致满足周期性，需进一步探究与具体月份的关系



01/统计与分析

SUSTech Southern University of Science and Technology

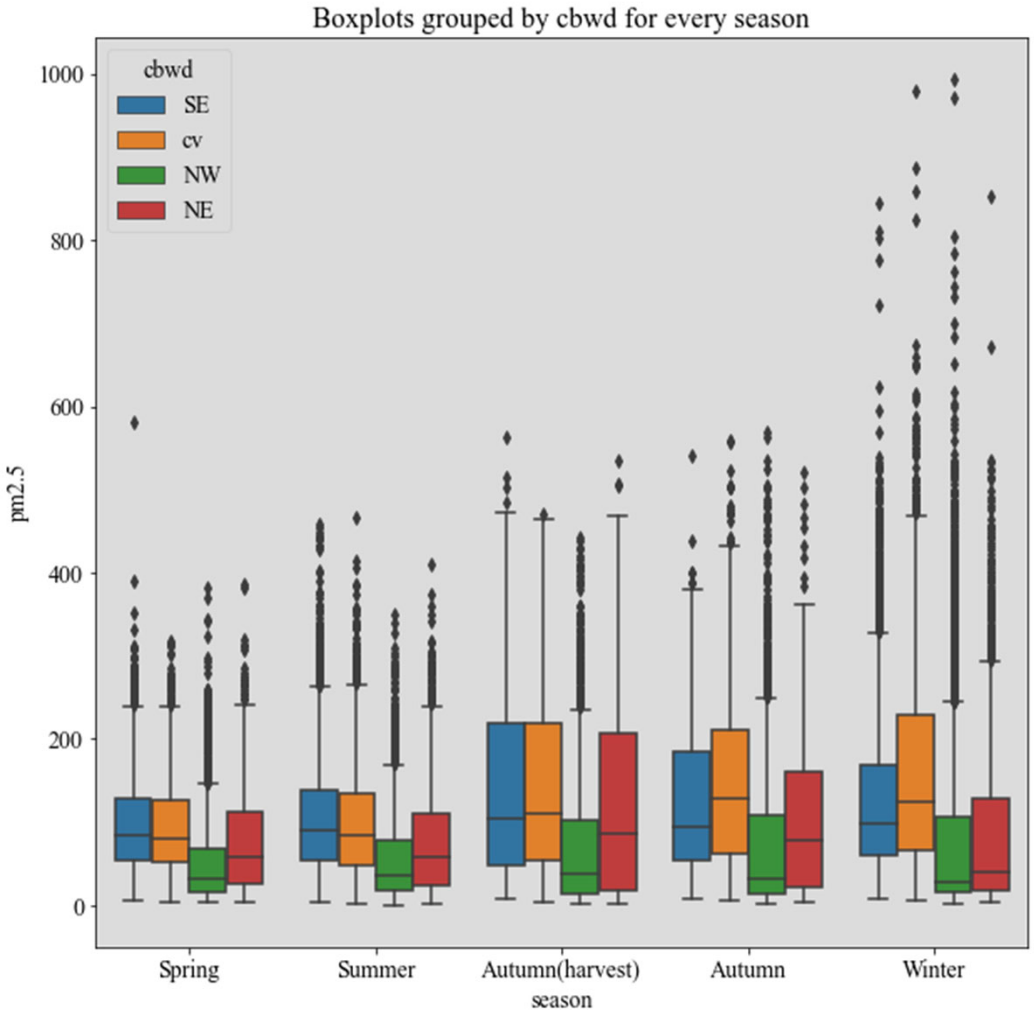


峰值解读：供暖与收获季

农民收获季：10月

北京供暖季：11月至次年3月

01/统计与分析

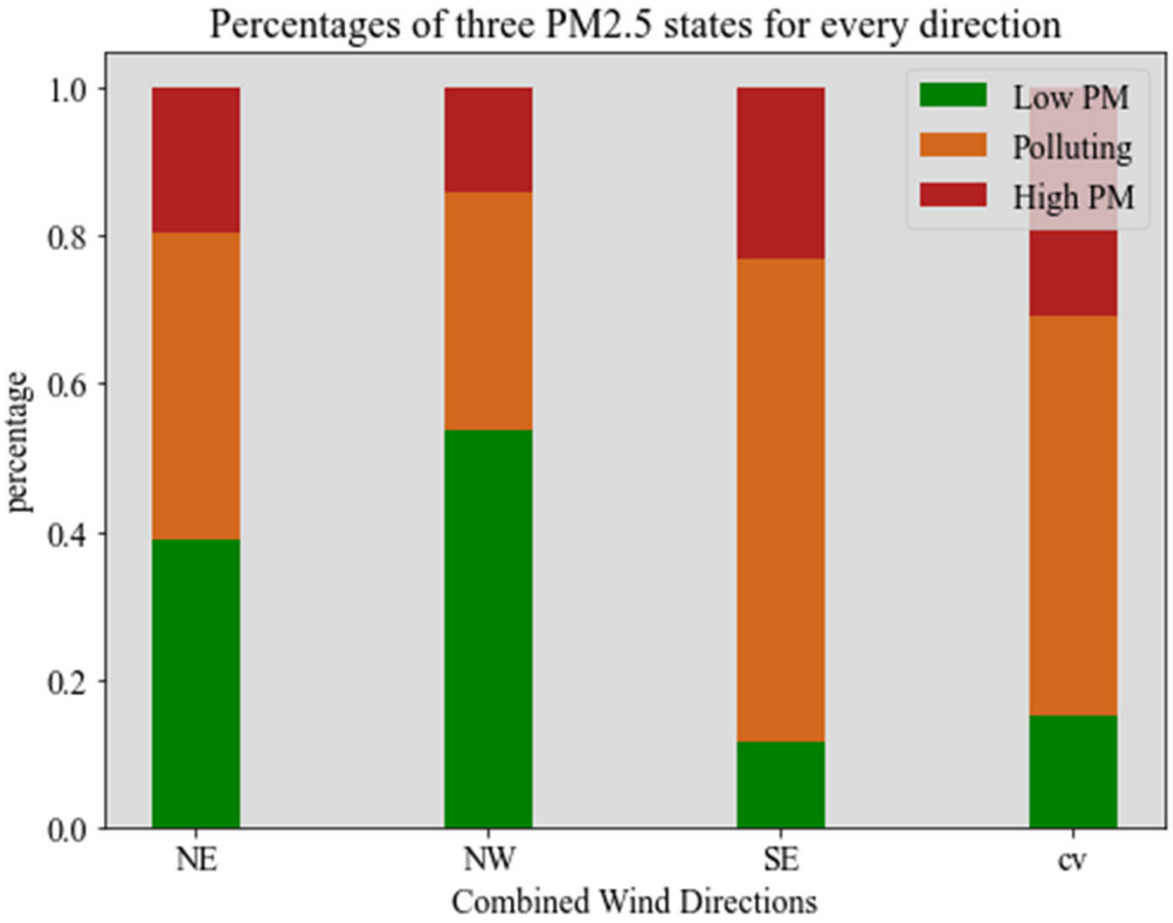


季节影响：

Autumn（harvest）：农民收获季

秋冬：供暖时节

01/统计与分析



风向：决定性因素

cv（calm variable）：

PM2.5均值最高

NW：PM2.5均值最低

（方向包含*W, WNW, NW, NNW 以及 N*）

SUSTech Southern University of Science and Technology

02

回归模型

Regression Models

Task 1、Task 2、Task 3

Task1： 删掉缺失值的PM2.5回归预测

SUSTech Southern University of Science and Technology

数据预处理

- 异常值处理：LOF算法（Local Outlier Factor）
  - 根据LOF密度来判断当前点是否为异常值
- 对预测变量PM2.5数值进行**Log变换**（PM2.5数值大于0）
  - 效果：
    - 减小绝对值，稳定方差，方便计算。
    - 取对数后，可以将乘法计算转换称加法计算。
- 训练集与测试集划分
  - 依据：周四划分为测试集，其余为训练集
  - 评价：待改进（在后续Task5中再提及）

Task1： 删掉缺失值的PM2.5回归预测

SUSTech Southern University of Science and Technology

回归分析结果

models(Task1)	R^2(train)	R^2(test)
OLS	0.471006363	0.475256873
Ridge	0.471006362	0.475256140
Ridge with CV	0.471006058	0.475246136
LASSO	0.319759069	0.334565029
LASSO with CV	0.471005222	0.475271069
XGBoost	0.634981228	0.604012641
Gradient Boost	0.632314424	0.603615480
CatBoost	0.653400236	0.609629450

Task1： 删掉缺失值的PM2.5回归预测

SUSTech Southern University of Science and Technology

调参方法

- 参数调整方法：

```
#梯度boosting算法(含优化函数调参)
from hyperopt import hp, fmin, rand, tpe, space_eval
from sklearn.ensemble import GradientBoostingRegressor
space = [hp.uniform('x', 300, 700), hp.uniform('y', 2, 4),hp.uniform('z',0.1,0.2)]
def q (args) :
    x, y,z= args
    gbr=GradientBoostingRegressor(n_estimators=int(x),max_depth=int(y),learning_rate=z)
    gbr.fit(train_data_task2_X,train_data_task2_y['pm2.5'])
    return -1*gbr.score(test_data_task2_X,test_data_task2_y['pm2.5'])
best = fmin(q, space, algo=rand.suggest,max_evals=100)

gbr=GradientBoostingRegressor(n_estimators=int(best['x']),max_depth=int(best['y']),learning_rate=best['z'])
gbr.fit(train_data_task2_X,train_data_task2_y['pm2.5'])
gbr.score(train_data_task2_X,train_data_task2_y['pm2.5']),gbr.score(test_data_task2_X,test_data_task2_y['pm2.5'])
```

Task1： 删掉缺失值的PM2.5回归预测

SUSTech Southern University of Science and Technology

CatBoost介绍

Catboost

From Wikipedia, the free encyclopedia

CatBoost[6] is an open-source software library developed by Yandex. It provides a gradient boosting framework which among other features attempts to solve for Categorical features using a permutation driven alternative compared to the classical algorithm.[7] It works on Linux, Windows, macOS, and is available in Python,[8] R,[9] and models built using catboost can be used for predictions in C++, Java,[10] C#, Rust, Core ML, ONNX, and PMML. The source code is licensed under Apache License and available on GitHub.[6]

InfoWorld magazine awarded the library "The best machine learning tools" in 2017.[11] along with Tensorflow, Pytorch, XGBoost and 8 other libraries.

Kaggle listed CatBoost as one of the most frequently used Machine Learning (ML) frameworks in the world. It was listed as the top-8 most frequently used ML framework in the 2020 survey[12] and as the top-7 most frequently used ML framework in the 2021 survey.[13]

As of April 2022, CatBoost is installed about 100000 times per day from PyPI repository[14]

Task2：加入时间特征的PM2.5回归预测

SUSTech Southern University of Science and Technology

新加入的特征

- Cbwd：独热编码成四个boolean，后序新创建的变量同样使用独热编码
- **Hour**：分成night（0-5）、morning（6-11）、afternoon（12-17）、evening（18-23）
- **周末、周内**分成两个状态：考虑到工厂排放、交通尾气等在两种状态下不同的强度；节假日也考虑在周末类别里
- 政策：将2014年设为有环保政策影响，其余年份不受政策影响（国务院于13年10月发布了《污染治理办法》，在此假设政策落地至少在两个月之后）
- 月份：分为5个类别，分别为 春季（4、5月）、夏季（6、7、8、9月）、秋季&收获季节（10月）、秋季&供暖季（11月）、冬季&供暖季（12、1、2、3月）

Task2：加入时间特征的PM2.5回归预测

SUSTech Southern University of Science and Technology

回归分析 效果对比 （Task1 不考虑时间）

- 整体效果有了一定提升

models(Task1)	R^2(train)	R^2(test)	models(Task2)	R^2(train)	R^2(test)
OLS	0.471006363	0.475256873	OLS	0.560275110	0.544333954
Ridge	0.471006362	0.475256140	Ridge	0.560275067	0.544351889
Ridge with CV	0.471006058	0.475246136	Ridge with CV	0.560274973	0.544365860
LASSO	0.319759069	0.334565029	LASSO	0.319499949	0.335050235
LASSO with CV	0.471005222	0.475271069	LASSO with CV	0.557739243	0.545857148
XGBoost	0.634981228	0.604012641	XGBoost	0.714439076	0.650345634
Gradient Boost	0.632314424	0.603615480	Gradient Boost	0.714062106	0.651403287
CatBoost	0.653400236	0.609629450	CatBoost	0.745576336	0.656212890

Task3：填补缺失值

SUSTech Southern University of Science and Technology

填补方法

- 均值填补：效果一般
- KNN（K紧邻算法）填补：效果一般
- **极端森林（Extremely randomized trees）填补：效果较好**

◆ 用不含缺失值的数据训练模型，再来预测含缺失值的样本

◆ 原理：每个决策树采用所有的样本，随机选取特征以训练决策树模型

◆ 效果：见下表

Task3： 极端随机森林填补

SUSTech Southern University of Science and Technology

回归分析 效果对比（极端随机森林填补、考虑时间）

- 相比Task1、2效果有了略微提升

	Task1		Task2		Task3	
models	R^2(train)	R^2(test)	R^2(train)	R^2(test)	R^2(train)	R^2(test)
OLS	0.471006363	0.475256873	0.560275110	0.544333954	0.563981589	0.545857636
Ridge	0.471006362	0.475256140	0.560275067	0.544351889	0.563981551	0.545874334
Ridge with CV	0.471006058	0.475246136	0.560274973	0.544365860	0.563981469	0.545887343
LASSO	0.319759069	0.334565029	0.319499949	0.335050235	0.320516883	0.333600026
LASSO with CV	0.471005222	0.475271069	0.557739243	0.545857148	0.563954819	0.546231473
XGBoost	0.634981228	0.604012641	0.714439076	0.650345634	0.716364923	0.653759181
Gradient Boost	0.632314424	0.603615480	0.714062106	0.651403287	0.707597397	0.651900258
CatBoost	0.653400236	0.609629450	0.745576336	0.656212890	0.747465507	0.660447298

SUSTech Southern University of Science and Technology

03

降维模型与特征选取

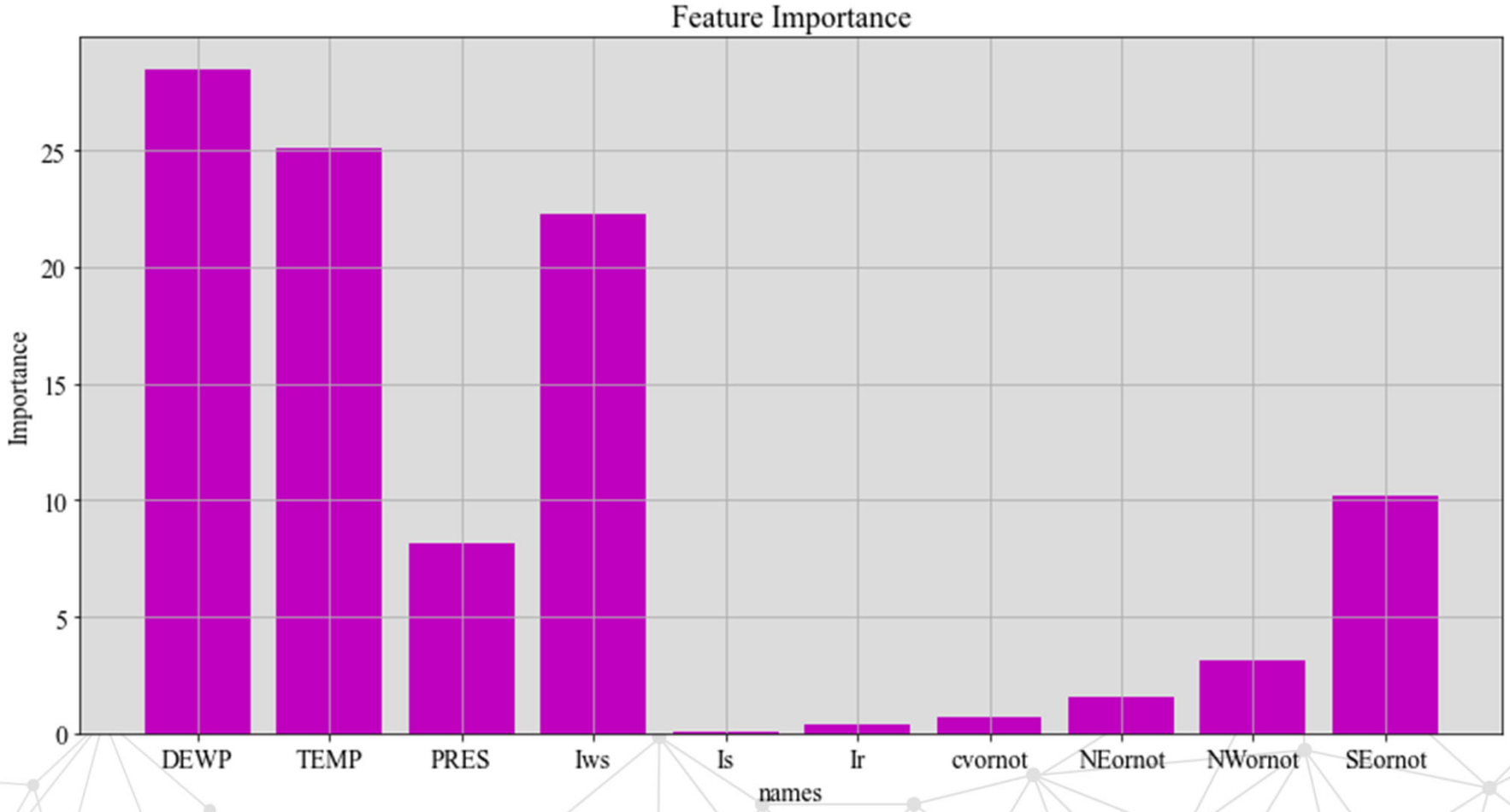
Dimension Reduction & Feature Selection

Task4

Task4：Feature Importance

SUSTech Southern University of Science and Technology

特征重要性：预览



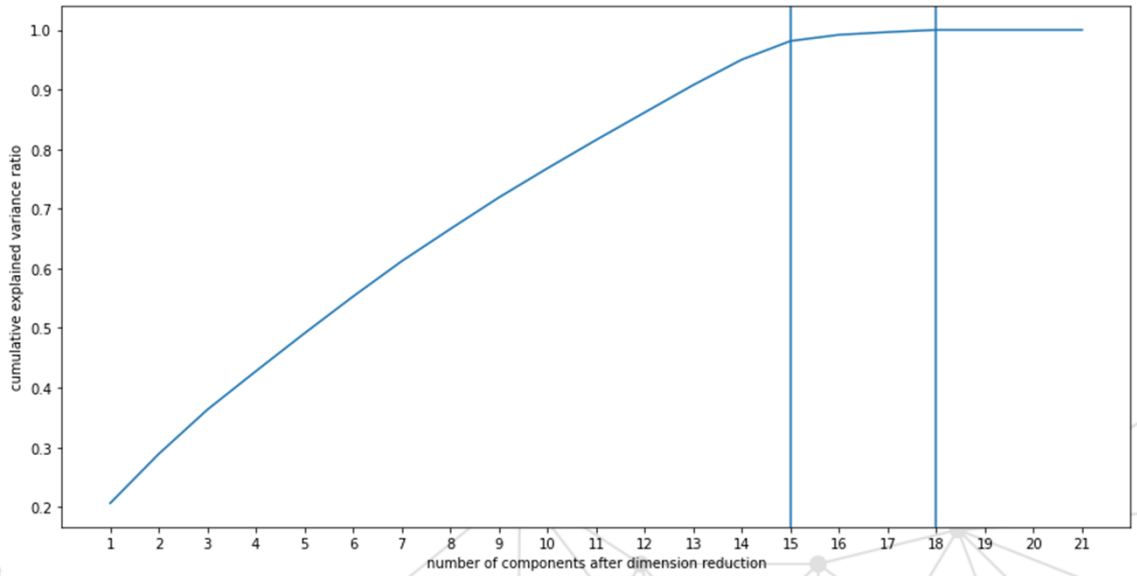
Task4：数据降维与特征选择

SUSTech Southern University of Science and Technology

降维方法

- Kernel PCA(效果不理想)、LLE（效果不理想）、ISOMAP（效果不理想）：占用内存过多、计算时间过长
- **PCA（主要讨论）**

◆ 原本有21维特征，先画出其PCA线性变换后的累计可解释方差比率图，发现可选取前15、18维作主要特征，从而降低计算复杂度，且或许可略微降低过拟合误差



Task4：数据降维与特征选择

SUSTech Southern University of Science and Technology

PCA降维后的拟合结果

	21维特征（参考）		18维特征		15维特征		4维特征（参考）	
	21 features(all features)		18 selected features		15 selected features		4 selected features	
models	R^2(train)	R^2(test)	R^2(train)	R^2(test)	R^2(train)	R^2(test)	R^2(train)	R^2(test)
OLS	0.560275110	0.544333954	0.560275110	0.544333954	0.510975727	0.517566311	0.406231452	0.408603403
Ridge	0.560275067	0.544351889	0.560275067	0.544351889	0.510975724	0.517565716	0.406231452	0.408603428
LASSO	0.319499949	0.335050235	0.307651804	0.324599921	0.307651804	0.324599921	0.307651804	0.324599921
CatBoost	0.745576336	0.656212890	0.774361244	0.660230320	0.757425619	0.655118404	0.595220054	0.537921794

Task4：数据降维与特征选择

SUSTech Southern University of Science and Technology

基于PCA的发散：

- 选取几乎全部离散特征并进行独热编码，再用PCA降维，以找出更多潜在的关键特征
- 结果：

33维特征（选取全部特征并独热编码后）　28维特征　16维特征

	33 features(all features)		28 selected features		16 selected features	
models	R^2(train)	R^2(test)	R^2(train)	R^2(test)	R^2(train)	R^2(test)
OLS	0.586067672	0.561340261	0.586062221	0.561340003	0.491220049	0.495999412
Ridge	0.586060347	0.561465610	0.586060347	0.561465610	0.491220047	0.495998906
LASSO	0.306909999	0.323666932	0.306909999	0.323666932	0.306909999	0.323666932
CatBoost	0.829964096	0.681110826	0.830086807	0.682637157	0.759052816	0.623852754

- 分析：选取28维时，略微降低维度与复杂度，并减小了略微提高了R2（可能减小了过拟合误差）

SUSTech Southern University of Science and Technology

04

分类模型

Classification Models

Task5

Task5：分类问题

SUSTech Southern University of Science and Technology

预处理

- 缺失值填补方式：极端森林
- 光滑化Smoothing：连续三个小时的均值，非PM2.5的所有连续变量
- PM2.5分成三个labels：low（PM2.5 $\leqslant 35\mu g/m^3$）、medium（$35\ \mu g/m^3 <$ PM2.5 $\leqslant 150\ \mu g/m^3$）、high(PM2.5 $> 150\ \mu g/m^3$)
- 特征：‘year’, ‘month’, ‘day’, ‘DEWP’, ‘TEMP’, ‘PRES’, ‘Iws’, ‘Is’, ‘Ir‘, 风向（独热编码）, 凌晨、早上、下午、傍晚（独热编码）

Task5：分类问题

分类结果（题目要求的测试/训练集划分）

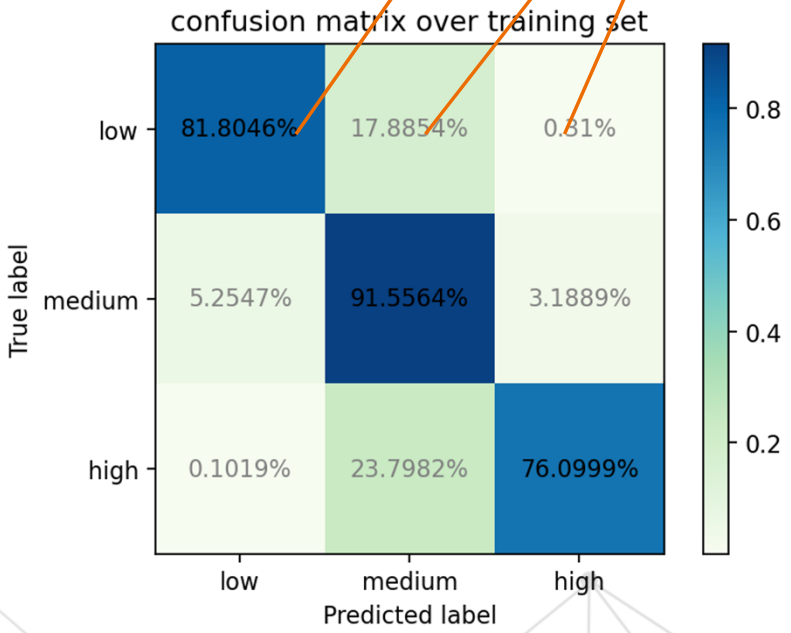
- CatBoost 分类结果Score（分类算法准确率）：

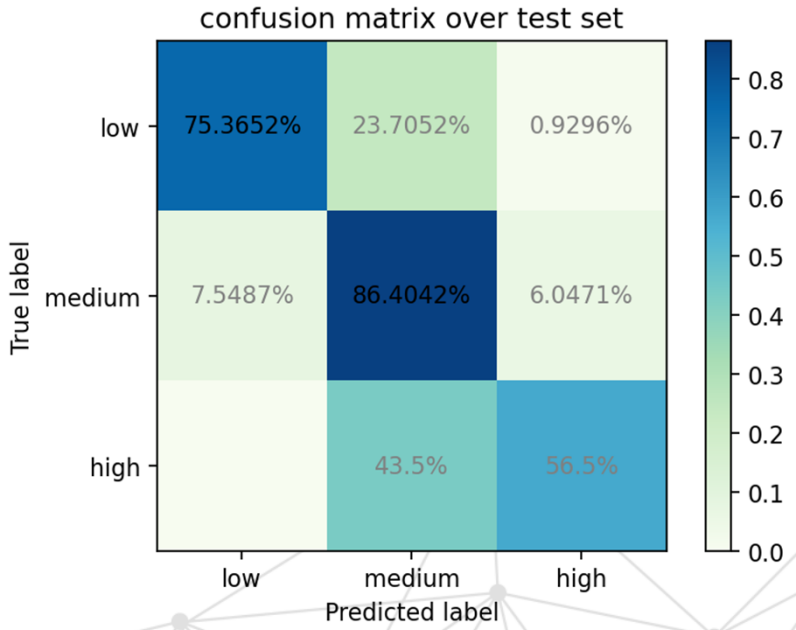
```
(0.8570906311276546, 0.7704225352112676)
```

- Confusion Matrices：

真实为low的样本中，预测为low的比率

预测错，但跳过medium直接为high的较少（连续性）





Task5：分类问题

SUSTech Southern University of Science and Technology

训练/测试集选择的改进

- 原划分：
  - 划分方式：测试集为周四数据，训练集为非周四数据
  - 隐患：如果周四数据和非周四数据真实分布相差较大，则用非周四数据去预测周四数据的效果较差
- 新的划分：
  - 目的：保持训练/测试集样本数，及二者数量的比例不变，同时取得更好的一般性（模型泛化能力）
  - 划分方式：将周四（原测试集）的一半数据放入新的训练集，同时将周日（原训练集）的一半数据放入新的测试集；其他数据保持不变

Task5：分类问题

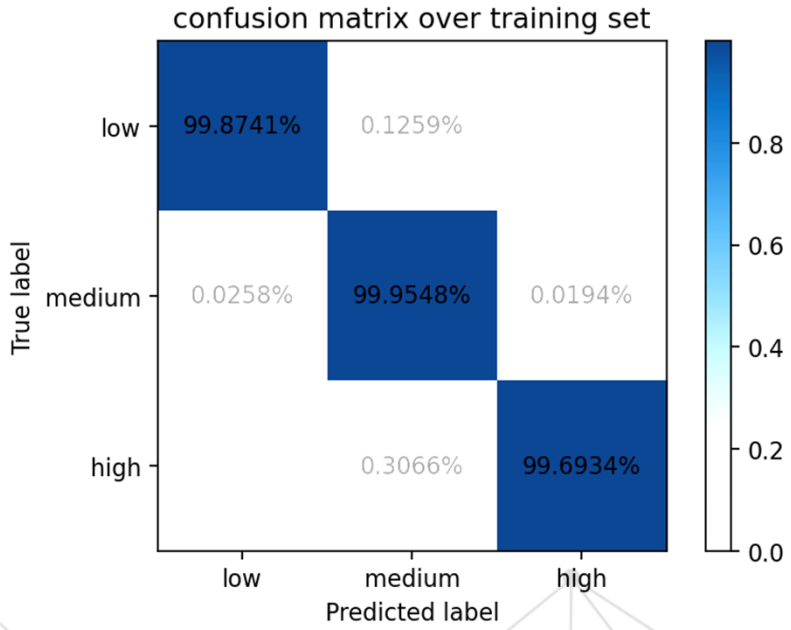
分类结果（新划分方式）

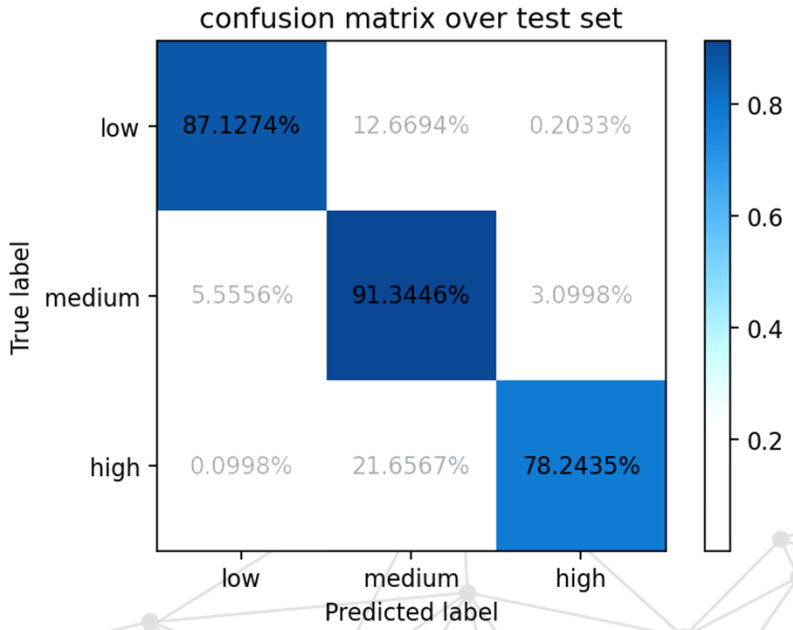
从0.770提高到0.874

- XGBoost 分类结果Score（分类算法准确率）：

```
(0.9988038674950992, 0.8744457879887142)
```

- Confusion Matrices：





总结

C O N C L U S I O N

R^2

66.0%

回归

填补：极端随机森林

添加自定义特征

模型：CatBoost

测试集划分：原划分

R^2

68.3%

PCA降维回归

填补：极端随机森林

添加所有特征后PCA，选取前28个特征

模型：CatBoost

测试集划分：原划分

Accuracy

87.4%

分类（新划分）

填补：极端随机森林

添加自定义特征（加入年、月、日）

模型：XGBoost

测试集划分：新划分

SUSTech Southern University of Science and Technology

SUSTech Southern University of Science and Technology

感谢观看