美赛

1. 多公式：至少30个公式
2. 图表美观，无模糊锯齿
3. 摘要之中针对不同的模型说明问题，优先指出核心模型和核心数据

算法

随机森林：RandomForest.监督学习方法

`sklearn.ensemble.RandomForestClassifier`

- 分类
- 预测
- 特征降维

XGBoost

XGBoost（eXtreme Gradient Boosting）是一种基于决策树的集成学习算法，被广泛用于各种机器学习和数据挖掘任务。下面列举了一些XGBoost可以处理的任务和应用：

1. **二分类和多分类问题：** XGBoost支持二分类和多分类问题，并且在这些任务中表现出色。在多分类任务中，XGBoost使用一种称为“one-vs-all”的策略来实现多个二分类器。
2. **回归问题：** XGBoost也可以用于回归问题，通过训练一个回归树模型，预测连续目标变量的值。
3. **特征选择：** XGBoost可以帮助选择最重要的特征，通过计算特征重要性得分，并通过这些得分选择最相关的特征。
4. **处理缺失值：** XGBoost可以处理缺失值，自动将其放到一个子树中，在不丢失太多信息的情况下进行预测。
5. **处理高维度数据：** XGBoost可以很好地处理高维度数据，通过剪枝和正则化来防止过拟合，提高泛化能力。
6. **异常检测：** XGBoost也可以用于异常检测，通过将异常样本看作是弱分类器，然后将其加入到XGBoost的集成模型中。
7. **时间序列预测：** XGBoost也可以用于时间序列预测任务，通过将时间序列数据转换为监督学习问题，并训练一个集成模型来预测未来的值。

评估算法的数字特征

`sklearn.metircs`提供了相关的函数

- 平均绝对误差：`mean_absolute_error`
- 平均平方误差：`mean_squared_error`
- 平均标准误差：`np.sqrt(mean_squared_error)`
- 假设检验（检验预测的准确性）
- AUC曲线，评估分类模型的性能指标

数学规划

使用lingo进行求解

lingo

1. lingo程序的结构

2023年美赛

1. C题，爬虫，twitter上有一个wordle stats 机器人，可以尝试爬取其数据。1）开发一种模型，并且使用该模型来预测2023年3月1日结果的可能区间。不同属性的单词是否会影响到困难模式下得分的百分比？并且解释原因　2） 开发一个模型，预测未来日期报告结果的分布。预测模型和结果有哪些uncertainties?对于2023年3月1日的单词*EERIE*，该模型有多少可信度？　3） 开发并总结一个模型，根据难度对单词进行分类，在该模型之中，单词*ERRIE*有多大的难度？讨论该模型的准确率。　4）列出并且描述该数据集的一些interesting features.
2. A题：开发一种模型来预测植物种群在不同的不规则的天气环境下随时间变化的情况。包括干旱时期，考虑不同物种之间的相互作用。种群需要多少种物种（how many different species are required for the community to benefit.),当物种数量增加时，会对种群产生什么影响。物种的类型是如何影响种群的？干旱发生愈加频繁、范围更广时，会有什么影响？如果相反，物种的数量还会对种群有同样的影响吗？
3. B题：
4. D题：注意各个问题之间是相互影响的。建立一个数学模型来确定联合国可持续发展目标的优先次序。可能是用到的数据集：世界统计年鉴。 1）构建各个指标之间的关系网络（*协方差矩阵，典型联相关分析*）
5. E题：光污染问题。确定一种指标，来确定一个地区的光污染水平。将指标应用于如下地区：1）保护区 2）农村地区 3） 郊区 4）城市地区。 描述三种可能的措施，来抑制光污染，讨论每项措施的具体措施与具体影响。选择两个地区，使用建议的评价指标来确定干预策略之中哪一项最为有效，讨论风险水平（副作用）。 最后，针对选定的地点，制作一张宣传单

*珞珈一号：夜光遥感卫星。获取光污染指标

6. GGDP（绿色GDP），选择GGDP作为世界经济发展的衡量指标，那么，会产生什么影响？会对环境产生什么影响？ 1） 目前已经开发了许多计算GGDP的方法，选择一个可以对缓解气候变化的指标 2） 创建一个模型，来预测估计对全球气候的影响。（回归模型） 3） 判断使用GGDP来代替GDP的潜在优势

E题

1. Develop a broadly applicable metric to identify the light pollution risk level of a location.

   建立一种广适用性的评价指标，来评估一个地区的灯光污染水平

2. Apply your metric and interpret its results on the following four diverse types of locations:
   o a protected land location,
   o a rural community,
   o a suburban community, and o an urban community.

   将该指标应用在以下四种不同的地区：保护区；农村地区；郊区；城市地区

3. Describe three possible intervention strategies to address light pollution. Discuss specific actions to implement each strategy and the potential impacts of these actions on the effects of light pollution in general.

   描述三种可以抑制光污染的有效策略。讨论每项策略的具体措施以及潜在影响。

4. Choose two of your locations and use your metric to determine which of your intervention strategies is most effective for each of them. Discuss how the chosen intervention strategy impacts the risk level for the location.

   选择两个地区，并且使用指标，来得出对于以上地区的最有效措施。讨论

5. 制作一张海报来推广你所提出的策略。

1. 第一题：自变量：GDP、人口密度、城镇化率、地区经济类型 因变量：光照强度
   1. 初步建立评价公式：

令GDP、人口密度等为自变量：X_i

令光照强度为：D_n

建立指标：$Z = \frac{X_i}{D_n}$，Z的数值越大越好，说明：产生单位经济效益，消耗的光污染强度越低。

2. 模型：评价模型：熵权法、TOPSIS方法、随机森林回归（SVR），XGBoost
3. 能否将*risk level of a location*分为几类：1）低 2）中 3）高，这样问题转化为一个分类问题。或者，利用回归模型预测光污染具体数值，之后再进行分类。

2. 第二题：
    1.
3. 第三题：

潜在影响因素：[biodiversity]

XGBoost的mae参数，在KFold之中进行：

5 行 × 1 列

	0
0	-254.874880
1	-154.983975
2	-246.522150
3	-186.091384
4	-174.964730

```
from sklearn.model_selection import cross_val_score,cross_validate
score = cross_val_score(xgb_model,x,y,cv=cv,scoring='neg_mean_squared_error')
```

```
score
```

5 行 × 1 列 CSV

	0
0	-165058.345587
1	-40851.292698
2	-169524.581315
3	-175018.312066
4	-77878.623443

```
# xgb_model.score()
xgb_model.score(x_train_fold_1,y_train_fold_1),xgb_model.score(x_test_fold_1,ans)
```

```
(0.9988684734564929, 1.0)
```

```
cv = KFold(n_splits=5,shuffle=True,random_state=100)
```

cv:cross_validation generator

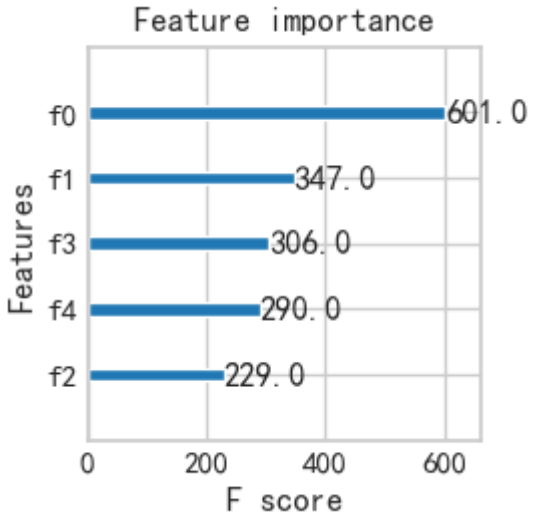
```
from sklearn.model_selection import cross_val_score,cross_validate
score = cross_val_score(xgb_model,x,y,cv=cv,scoring='r2')
```

```
score
```

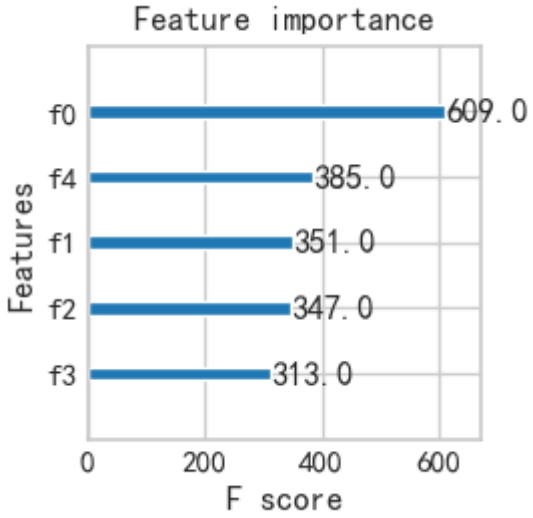
5 行 5 行 × 1 列 CSV

	0
0	-165058.345587
1	-40851.292698
2	-169524.581315
3	-175018.312066
4	-77878.623443

还不清楚参数如何进行解释。



使用手工KFlod交叉检验时的feature importance



此图为在使用`train_test_split()`方法时，所得到的结果

第一问，首先进行数据清洗，整理。之后进行皮尔逊相关性系数分析与**显著性检验矩阵。**证明因子与结果之间具备相关性。

注意，必须要进行显著性检验，否则无法直接说明数据之间的相关性（具体原因我也不懂）

这是我们进行回归分析的基础。

接下来，建立Xgboost的回归分析过程，XGBoost是否有**decision_function**?从这几种指标来进行评价：

- 进行拟合残差分析。
- R^2
- $mean\quad absolute\quad error$与$mean\quad squred\quad error$
- 显著性检验。

1. 整理数据，清洗数据
2. 首先进行皮尔逊相关性分析，得出相关系数矩阵（corr_matrix.csv).也有热力图：

我们发现，即TNL与GDP等因素之间的相关系数为：

$$r = (0.52 \quad 0.08 \quad 0.47 \quad 0.66 \quad 0.66)$$

同时获得对应的显著性检验矩阵为：*对于该数值的解释，问下hyz*

	VIIRS（W）	总 gdp（亿元）	人口密度（平方公里/人）	城镇化率（%）	汽车保有量（万辆）	总用电量（亿千瓦时）
VIIRS（W）	1					
总 gdp（亿元）	.524**	1				
人口密度（平方公里/人）	.080	.028	1			
城镇化率（%）	.469**	.298**	.619**	1		
汽车保有量（万辆）	.663**	.581**	-.038	.406**	1	
总用电量（亿千瓦时）	.655**	.577**	-.011	.299**	.489**	1

**.·在·0.01·级别（双尾），相关性显著。

> 1 显著性检验的目的是为了 将从样本中得到的结论推广到总体中，通过“小概率事件是不可能事件”这一原理进行推断。一般而言是对总体做出原假设，然后通过对随机的样本数据对原假设进行分析，判断其与原假设是否存在显著性的差异。

接下来，我们使用了以下几种模型，来试图建立回归分析：

- 支持向量机回归：支持向量机(SVR,Support Vecotr Machine)是一种传统的机器学习方法。其主要思想为：按照某种事先选择的非线性映射$x \to f(x)$，将输入量映射到一个高维的特征空间，在这个空间构造最优分类超平面。为了得到最优分类超平面，SVM引入了核函数(kernel funciton)，例如：高斯核(rbf)，线性核(linear kernel)，多项式核(poly kernel)等等。（*注意术语不要翻译错*）通过引入特定的损失函数，从而支持向量机从分类算法推广到了回归分析。
- 随机森林回归：随机森林是一种基于决策树的集成算法，其核心思想是一个有多颗随机生成的决策树组成的森林，每一个数据输入以后，有各个不相关的决策树做分类或回归，并且投票决定该数据该如何分类或者回归。随机森林主要用到的决策算法是CART(classification and regression tree)算法，同时引入了随机采样的概念，能高效的避免过拟合的出现。

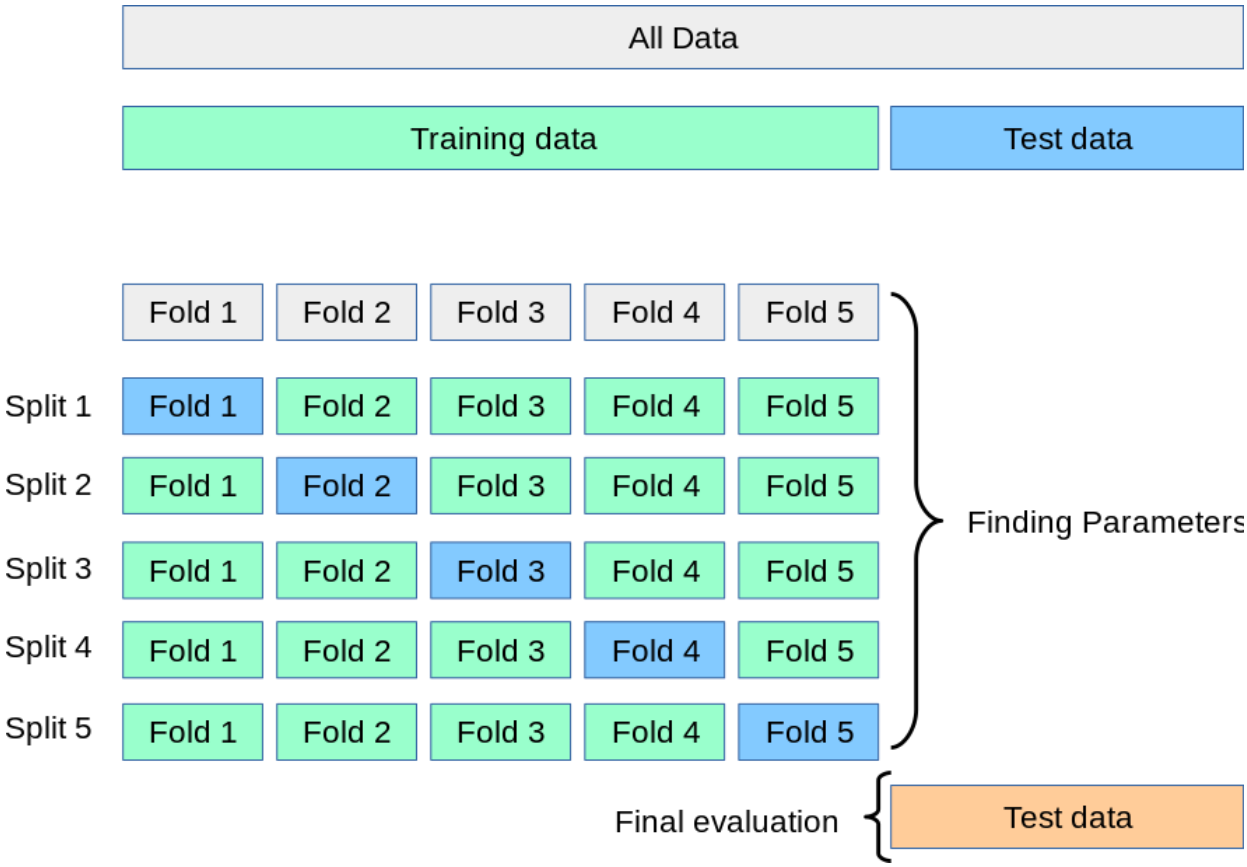
$CART$算法的优化方程如下：

$$MSE_{node} = \sum_{i \in node} (\hat{y} - y^{(i)})^2$$

$$\hat{y}_{node} = \frac{\sum_{i \in node} y^{(i)}}{m_{node}}$$

其中，$MSE = \frac{1}{N} \sum_{n=1} N(f(x_n) - y_n)^2$

- 基础的线性回归：
- XGBoost回归：XGBoost 是一种Boosting 型的树集成模型，在梯度提升决策树GBDT基础上扩展，能够进行多线程并行计算，通过迭代生成新树，即可将多个分类性能较低的弱学习器组合为一个准确率较高的强学习器。XGBoost 采用随机森林对字段抽样，将正则项引入损失两数中，从而防止模型过拟合，并降低模型计算量。



在`python`之中使用sklearn以及XGBoost库建立回归模型，各项模型未经参数优化，在K Fold交叉验证之中，分别以r^2以及mean absolute error还有$mean \quad squred \quad error$得出了如下结果：

- 决定系数，拟合优度，反映了y的波动有多少百分比能被x的波动所描述。拟合优度越大，自变量对因变量的结实程度越高。

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(f(x_i) - y_i)^2}{\sum_{i=1}^{m}(f(x_i) - \bar{y}_i)^2}$$

- mean squred error

$$MSE = \frac{1}{n} * \sum_{i=1}^{n}(y_{pred} - y_{actual})^2$$

- mean absulote error:

$$MAE = \frac{1}{n} * \sum_{i=1}^{n}|y_{pred} - y_{actual}|$$

回归模型	r^2	MSE	MAE
randomforest	0.52928	157000	248.47
xgboost	0.54	115810	205
svc(linear)	0.43323	150150	199
svc(poly)	-0.16374	344839	293
linearregression	0.4605854	142790	226

注意，该表格之中的各项数据均经过为Kfold交叉检验的结果。

可见其中随机森林与XGBoost表现结果最好。

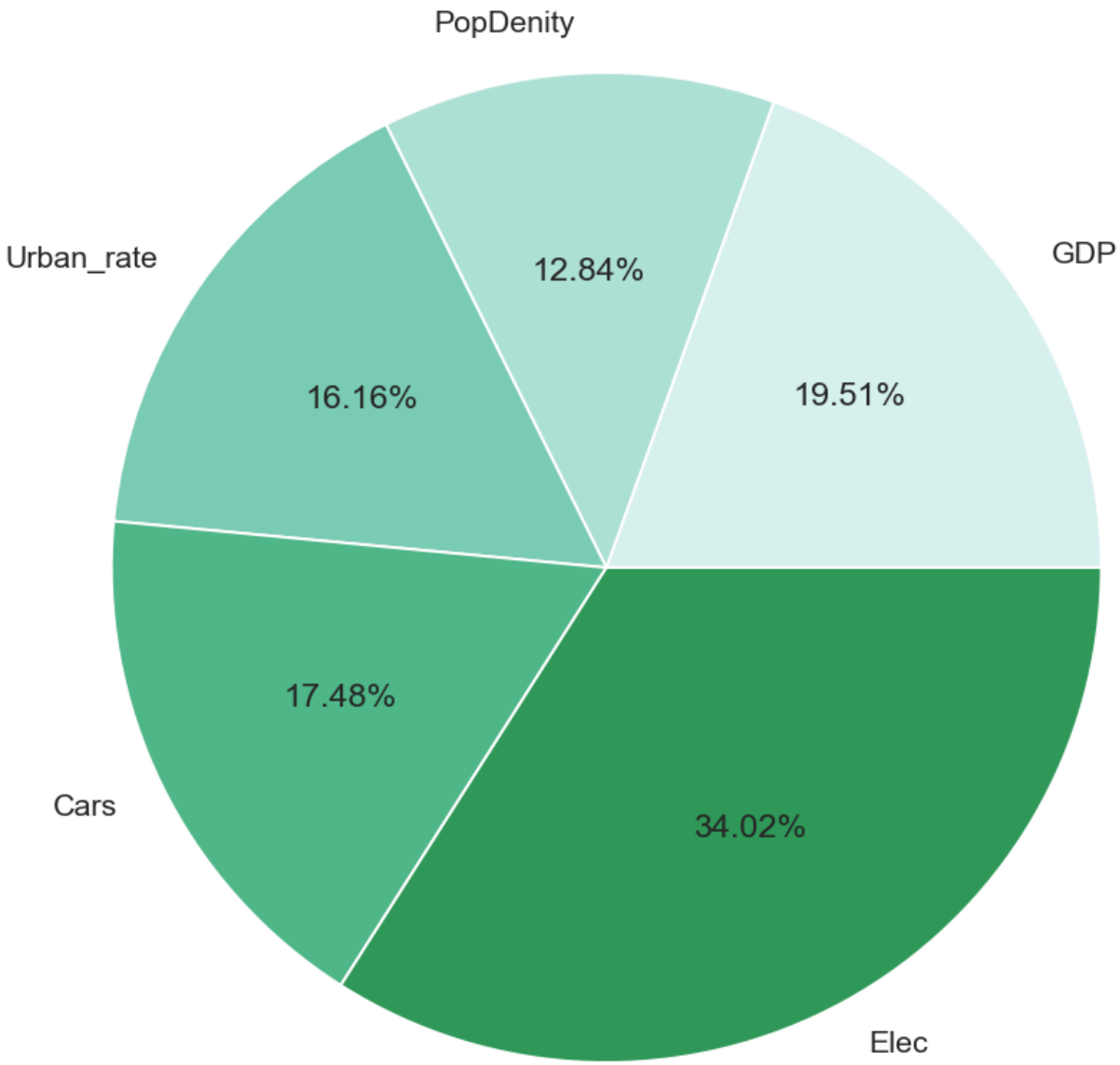
选择随机森林进行进一步的调参优化。

最终实现随机森林在本数据集上的最优参数：

```
min_samples_split=1,
min_samples_leaf=2,
max_features='sqrt',
n_estimators=60,
random_state=10,
max_depth=8,
oob_score=True
```

回归分析的各项指标	r^2	MAE	MSE
最优参数的随机森林回归表现结果	0.62	202	102878

各项指标的重要程度。利用`randomforestregressor._impotrance_`得出



验证集：

武汉市	**2368.89**	**15000**	**1050**	**81.7**	**260**	**364**
漳州市	462.78	603.72	410	51.1	31.22	60.36
宣城市	104.23	383.73	256	46.4	16.52	39.15
南川区	18.05	148.87	205	38.29	30.29	49.35

预测结果：