Project of Introduction to Big Data Science

Yan Huang & Yifan Chen

Department of mathematics, SUSTech

2022.6.2

- Yan Huang: task1、task5、task2: plot tendency of pm2.5 variation、task3: imputation method
- Yifan Chen: task2 周一、12 月和春节绘图部分全部预测部分, task3 重新运行 task2 部分
- 占比：50%：50%

1. Task1&4
2. Task2
3. Task3&4
4. Task5

1 Task1&4
数据预处理
模型建立
特征选择

2 Task2

3 Task3&4

4 Task5

1 Task1&4
数据预处理
模型建立
特征选择

2 Task2

3 Task3&4

4 Task5

- 在原数据集增加“date”属性，以格式‘%Y-%m-%d’记录日期，再通过 pd.date_range() 获取测试数据日期。将数据集中的‘NA’转换为 np.nan，再用 pd.dropna 去除
- 从 pm2.5 概率密度图中发现数据严重左偏

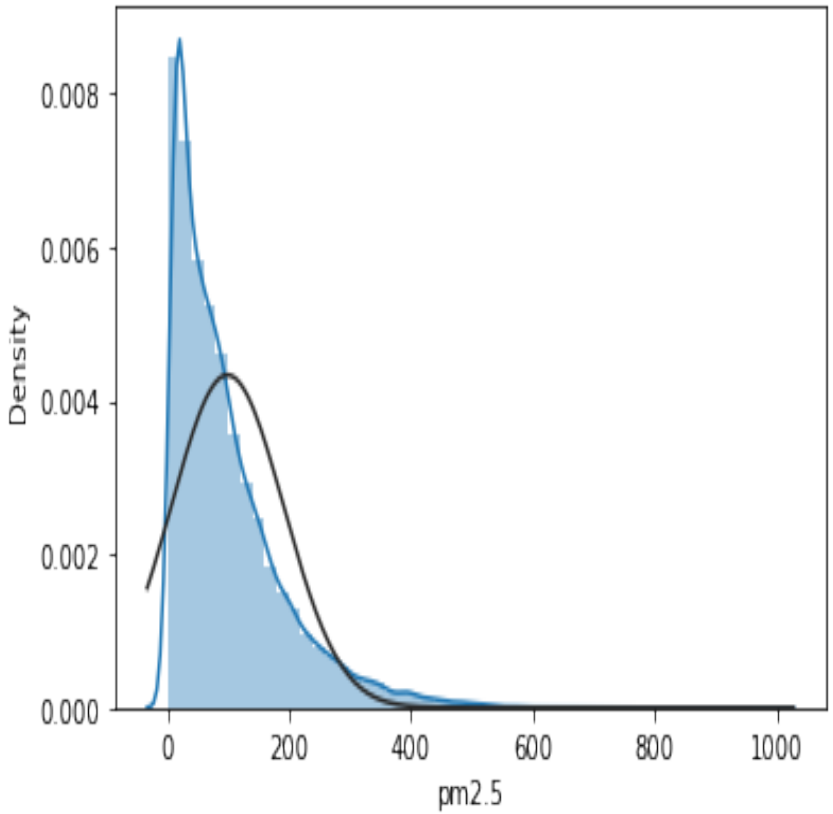

图 1: pm2.5 概率密度图

- 对数转换 pm2.5，我们得到一个近似正态分布的数据。log pm2.5 概率密度图的 $Skewness = -0.356563$，$Kurtosis = -0.563840$

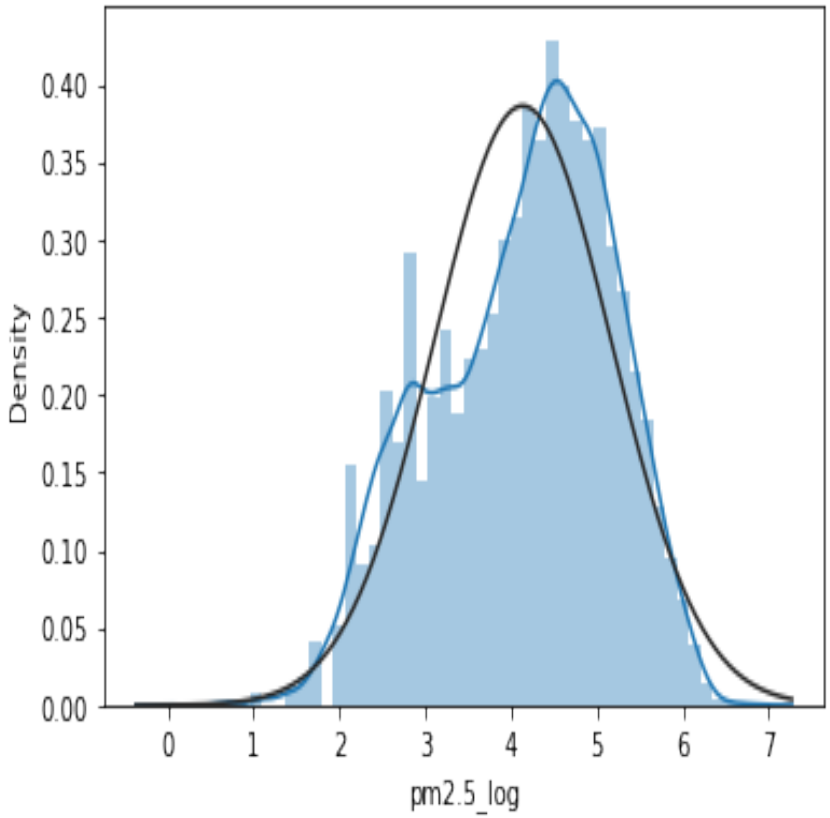

图 2: log pm2.5 概率密度图

- 从 p-p 图中也可以看出对数转换后的 pm2.5 近似服从正态分布

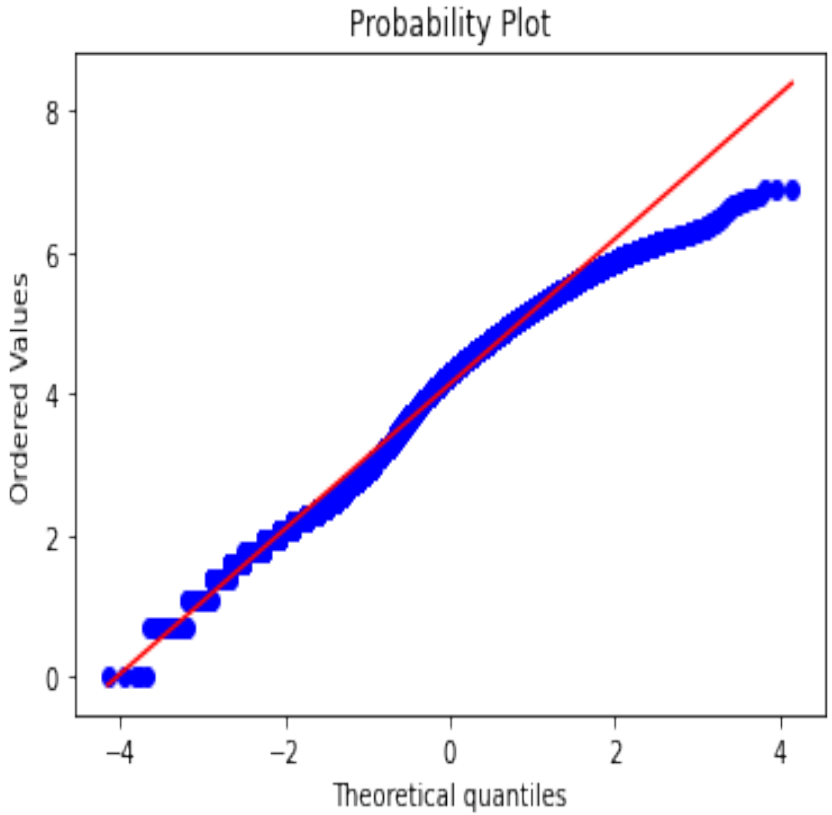

图 3: p-p 图

1 Task1&4
- 数据预处理
- 模型建立
- 特征选择

2 Task2

3 Task3&4

4 Task5

- 如果特征中只含气象信息
- 我们选择模型 XGBRegressor(learning_rate=0.03, n_estimators=400, max_depth=6) 拟合对数转换后的数据
- 我们在测试集上预测 pm2.5，然后将得到的预测数据做指数变换得到 y_pred；将 y_pred 与真实数据 y_test 作 min_max 归一化，再计算它们的 MSE，我们得到的结果是 0.0281，该模型的 $R^2 = 0.5603$

- 如果特征中包含时间信息
- 我们选择模型 XGBRegressor(learning_rate=0.1, n_estimators=600, max_depth=5) 拟合对数转换后的数据
- 我们在测试集上预测 pm2.5，然后将得到的预测数据做指数变换得到 y_pred；将 y_pred 与真实数据 y_test 作 min_max 归一化，再计算它们的 MSE，我们得到的结果是 0.0084，该模型的 $R^2 = 0.7100$

1 Task1&4
数据预处理
模型建立
特征选择

2 Task2

3 Task3&4

4 Task5

- 通过 XGB.feature_importances_ 选择特征
- 从图中我们看出在所有特征中 day 的重要性最高；在气象特征中，DEWP、PRES、Iws、TEMP 的重要性较高

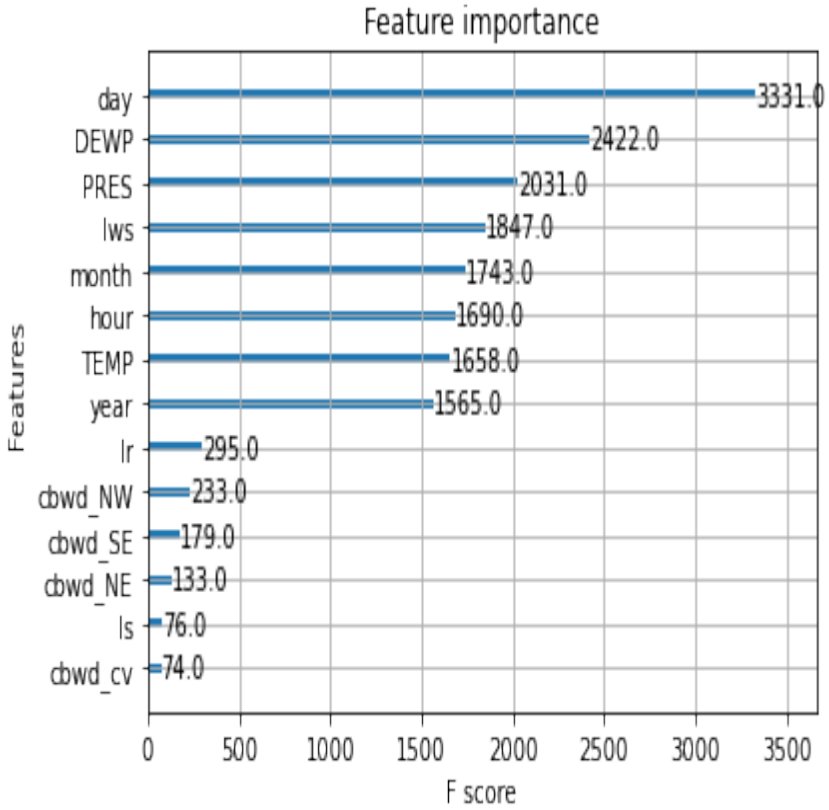

图 4: 特征重要性

- 通过相关系数矩阵选择特征
- 从图中可以看出 DEWP、TEMP、Iws、cbwd_NW、cbwd_SE、cbwd_cv 与 pm2.5 的相关性较高

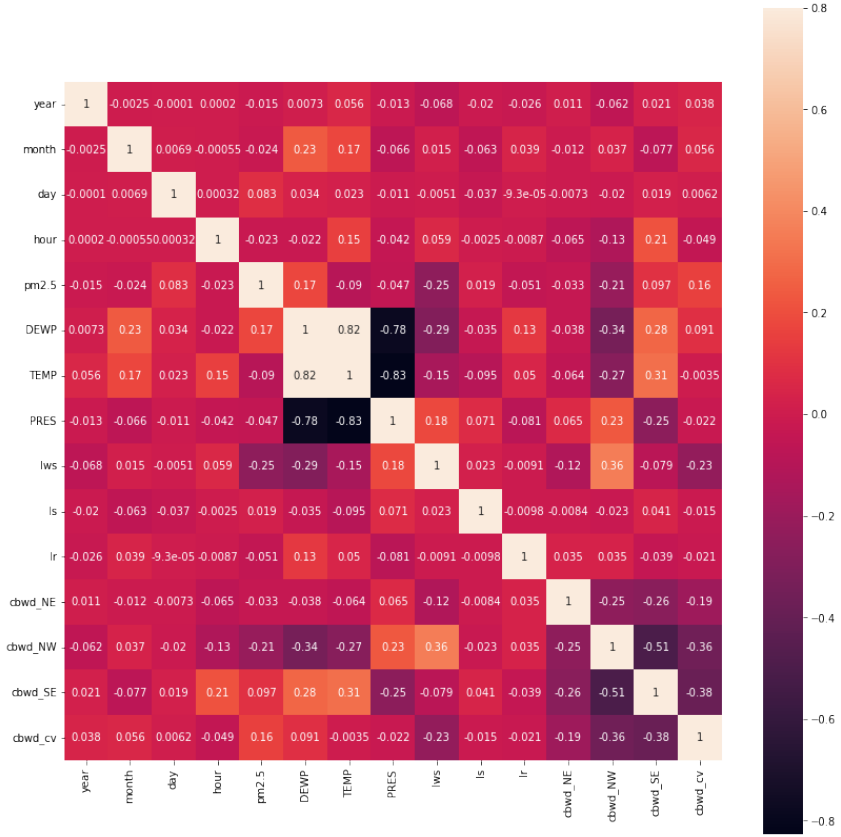

图 5: 相关系数矩阵

- 通过 PCA.explained_variance_ratio_ 选择特征
- 从图中我们可以看出可以取前 3 个主成分作为特征，其中第一主成分主要反映了对 pm2.5 的影响

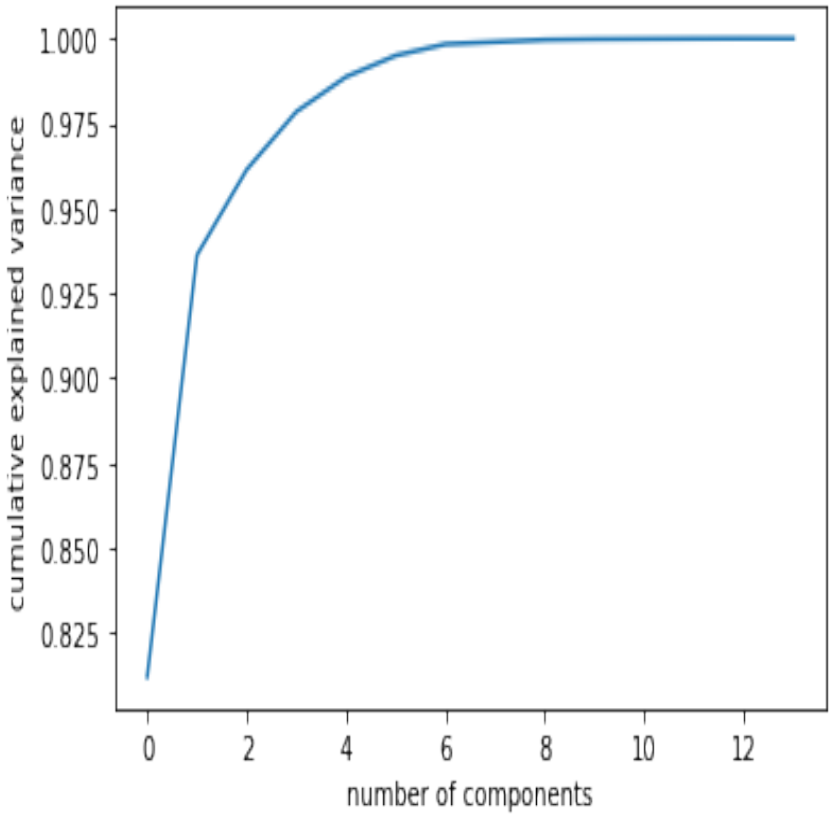

图 6: cumulative explained variance

1 Task1&4

2 Task2

用折线图记录变化趋势
举例: 每年每月 pm2.5 的变化趋势
举例: 每年春节期间（除夕前两天、除夕、正月初一至初七、初七后两天）每小时平均 pm2.5 的变化趋势
运用上述特征进行对 pm2.5 的预测
举例: 分别使用每年每月数据的时间信息和天气信息，对每月 pm2.5 平均浓度进行预测

3 Task3&4

4 Task5

1 Task1&4

2 Task2

用折线图记录变化趋势

举例: 每年每月 pm2.5 的变化趋势

举例: 每年春节期间（除夕前两天、除夕、正月初一至初七、初七后两天）每小时平均 pm2.5 的变化趋势

运用上述特征进行对 pm2.5 的预测

举例: 分别使用每年每月数据的时间信息和天气信息，对每月 pm2.5 平均浓度进行预测

3 Task3&4

4 Task5

- 每年一天 24 小时 pm2.5 的变化趋势
- 每年一周每天 pm2.5 的变化趋势
- 每年每月 pm2.5 的变化趋势
- 每年四季 pm2.5 的变化趋势
- 每年 pm2.5 的变化趋势
- 每年周一早晨（6 点至 9 点）每小时平均 pm2.5 的变化趋势
- 每年 12 月周末（周六周日）晚上（18 点至 24 点）每小时平均 pm2.5 的变化趋势
- 每年春节期间（除夕前两天、除夕、正月初一至初七、初七后两天）每小时平均 pm2.5 的变化趋势
- 每年春节期间（除夕前两天、除夕、正月初一至初七、初七后两天，以除夕前第二天为第 0 天）每天平均 pm2.5 的变化趋势

1 Task1&4

2 Task2

用折线图记录变化趋势

举例: 每年每月 pm2.5 的变化趋势

举例: 每年春节期间（除夕前两天、除夕、正月初一至初七、初七后两天）每小时平均 pm2.5 的变化趋势

运用上述特征进行对 pm2.5 的预测

举例: 分别使用每年每月数据的时间信息和天气信息，对每月 pm2.5 平均浓度进行预测

3 Task3&4

4 Task5

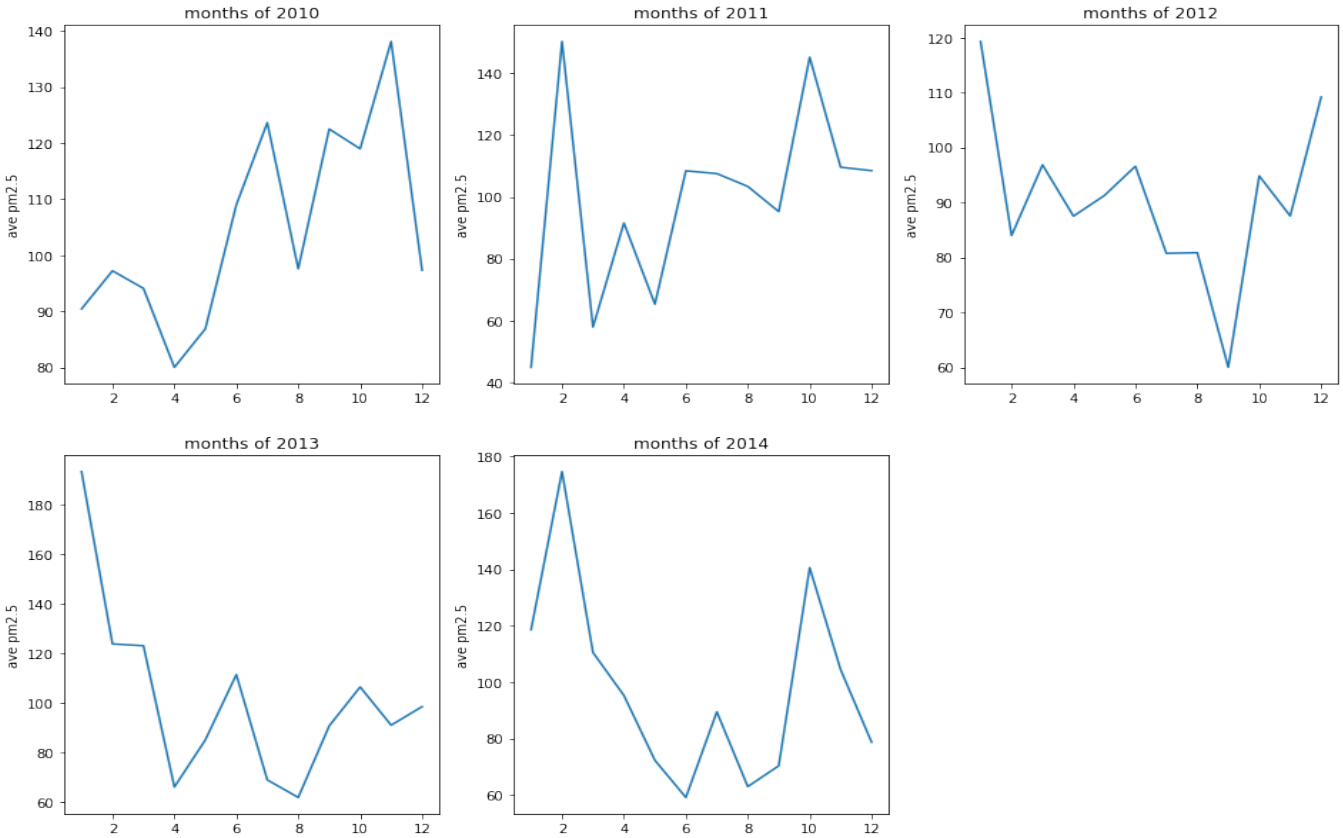

图 7: 每年每月 pm2.5 的变化趋势

- 每年每月 pm2.5 平均浓度变化趋势不同，但从 2011 年后，基本从 2 月到 10 月浓度维持在相对较低水平，而其余月份基本维持在相对较高水平，可能与冬季居民供暖有关

1 Task1&4

2 Task2

用折线图记录变化趋势

举例: 每年每月 pm2.5 的变化趋势

举例: 每年春节期间（除夕前两天、除夕、正月初一至初七、初七后两天）每小时平均 pm2.5 的变化趋势

运用上述特征进行对 pm2.5 的预测

举例: 分别使用每年每月数据的时间信息和天气信息，对每月 pm2.5 平均浓度进行预测

3 Task3&4

4 Task5

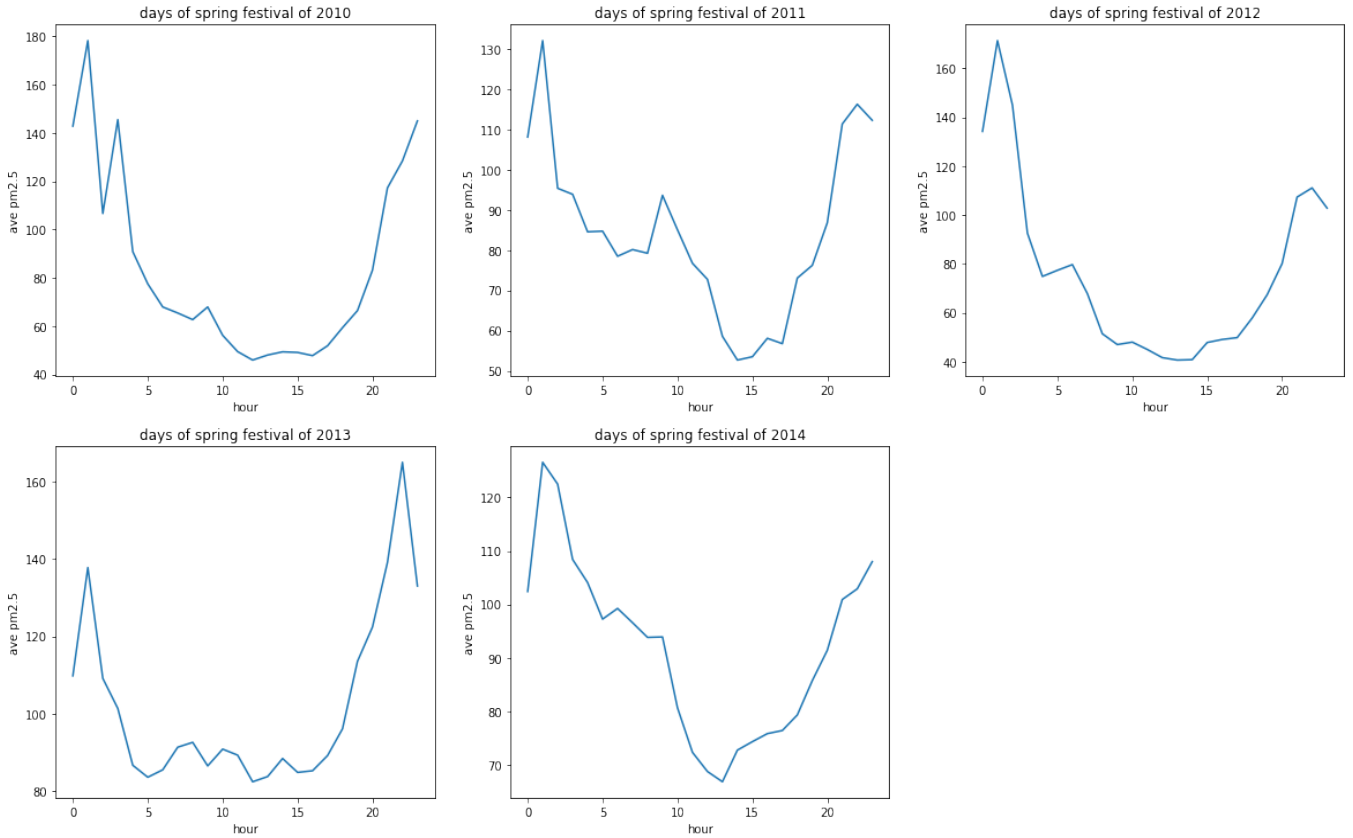

图 8: 每年春节期间每小时平均 pm2.5 的变化趋势

- 每天每小时的变化趋势与全年相似，均为 15 点后上升，凌晨 1 点后慢慢下降

1 Task1&4

2 Task2

用折线图记录变化趋势

举例: 每年每月 pm2.5 的变化趋势

举例: 每年春节期间（除夕前两天、除夕、正月初一至初七、初七后两天）每小时平均 pm2.5 的变化趋势

运用上述特征进行对 pm2.5 的预测

举例: 分别使用每年每月数据的时间信息和天气信息，对每月 pm2.5 平均浓度进行预测

3 Task3&4

4 Task5

- 运用 task1 中提出的 xgboost 模型、训练集和测试集
- 利用上述与时间有关的 pm2.5 平均浓度信息，分别使用每条数据的时间信息和天气信息
- 使用 score 作为预测结果好坏的测量标准，认为越接近 1 预测效果越好

- 分别使用每年一天 24 小时数据的时间信息和天气信息，对每小时 pm2.5 平均浓度进行预测
- 分别使用每年每月数据的时间信息和天气信息，对每月 pm2.5 平均浓度进行预测
- 分别使用每年每季数据的时间信息和天气信息，对每季 pm2.5 平均浓度进行预测

1 Task1&4

2 Task2

用折线图记录变化趋势
举例: 每年每月 pm2.5 的变化趋势
举例: 每年春节期间（除夕前两天、除夕、正月初一至初七、初七后两天）每小时平均 pm2.5 的变化趋势
运用上述特征进行对 pm2.5 的预测
举例: 分别使用每年每月数据的时间信息和天气信息，对每月 pm2.5 平均浓度进行预测

3 Task3&4

4 Task5

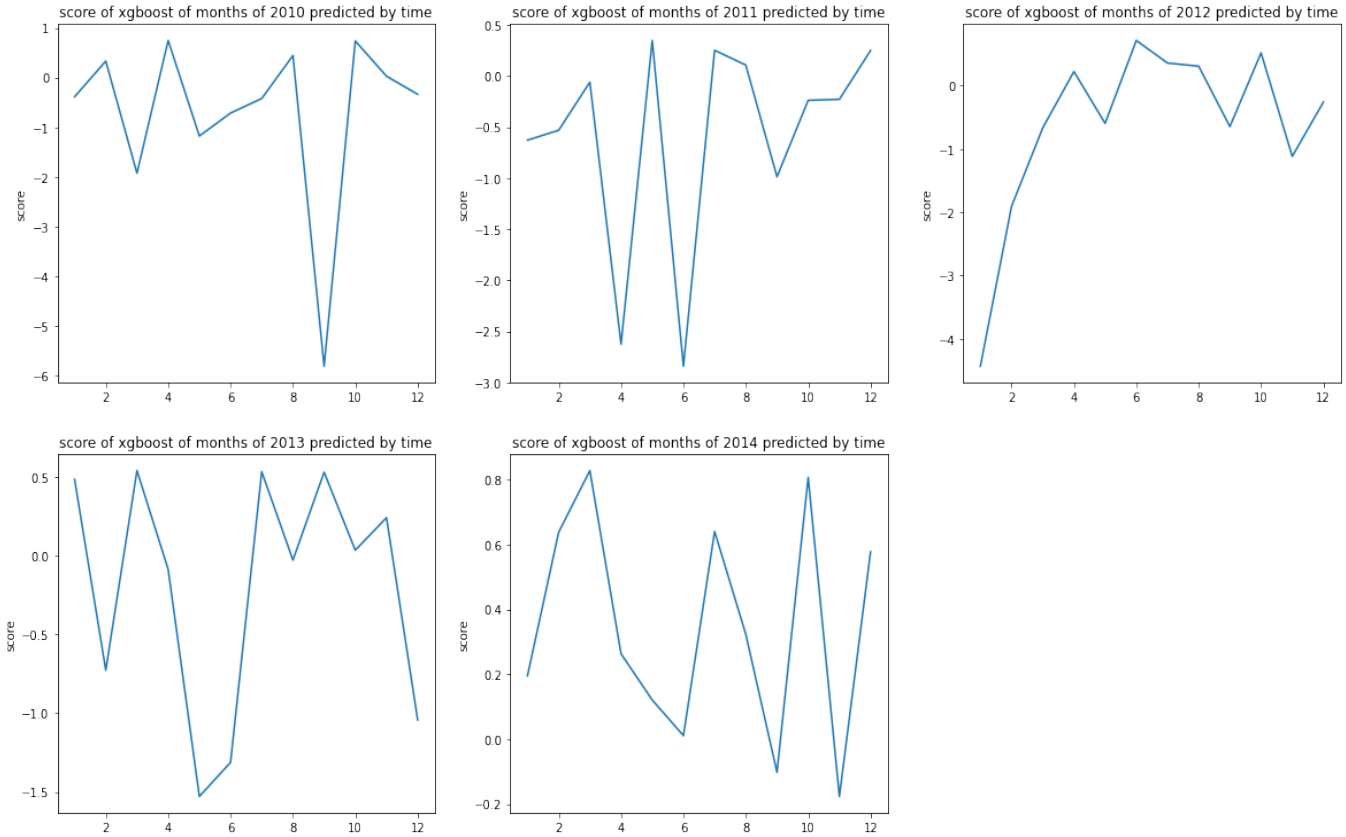

图 9: 使用每年每月数据的时间信息, 得到每月 pm2.5 平均浓度的预测 score 折线图

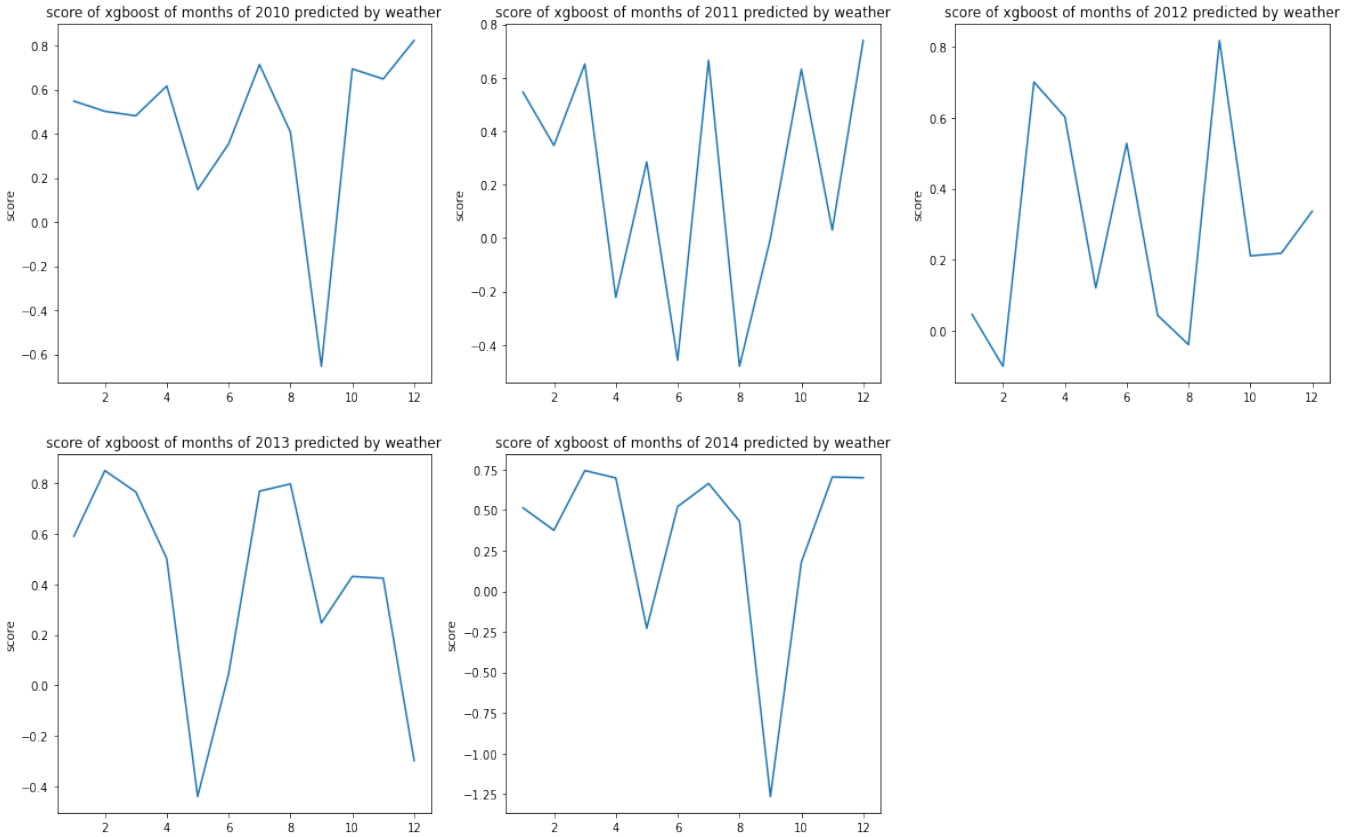

图 10: 使用每年每月数据的天气信息, 得到每月 pm2.5 平均浓度的预测 score 折线图

- 可以看出，虽然用每月平均浓度预测时只用天气信息可以达到更高精度，但预测准度更加不稳定，既可以达到 0.8，也可以达到-1.25

1. Task1&4
2. Task2
3. Task3&4
   - 平均填补法
   - K 近邻填补法, $K = 40$
4. Task5

1 Task1&4

2 Task2

3 Task3&4

平均填补法

填补说明

与 task1 类似，用 xgboost 预测模型拟合填补后的数据，并进行特征选择

与 task1 类似，用相关系数矩阵进行特征筛选

与 task1 类似，通过主成分分析选择特征并尝试降维

类似 task2，用由平均填补法得到的新数据记录变化趋势

类似 task2，用由平均填补法得到的新数据的上述新特征对平均 pm2.5 浓度进行预测

K 近邻填补法，$K = 40$

4 Task5

填补说明

- 用其他年份的相同日期时刻的平均值进行填补
- 注意到 2012 年 2 月 29 日也缺了数据，决定用 2012 年 2 月 28 日和 3 月 1 日的数据取平均作为 2012 年 2 月 29 日的数据

与 task1 类似，用 xgboost 预测模型拟合填补后的数据，并进行特征选择

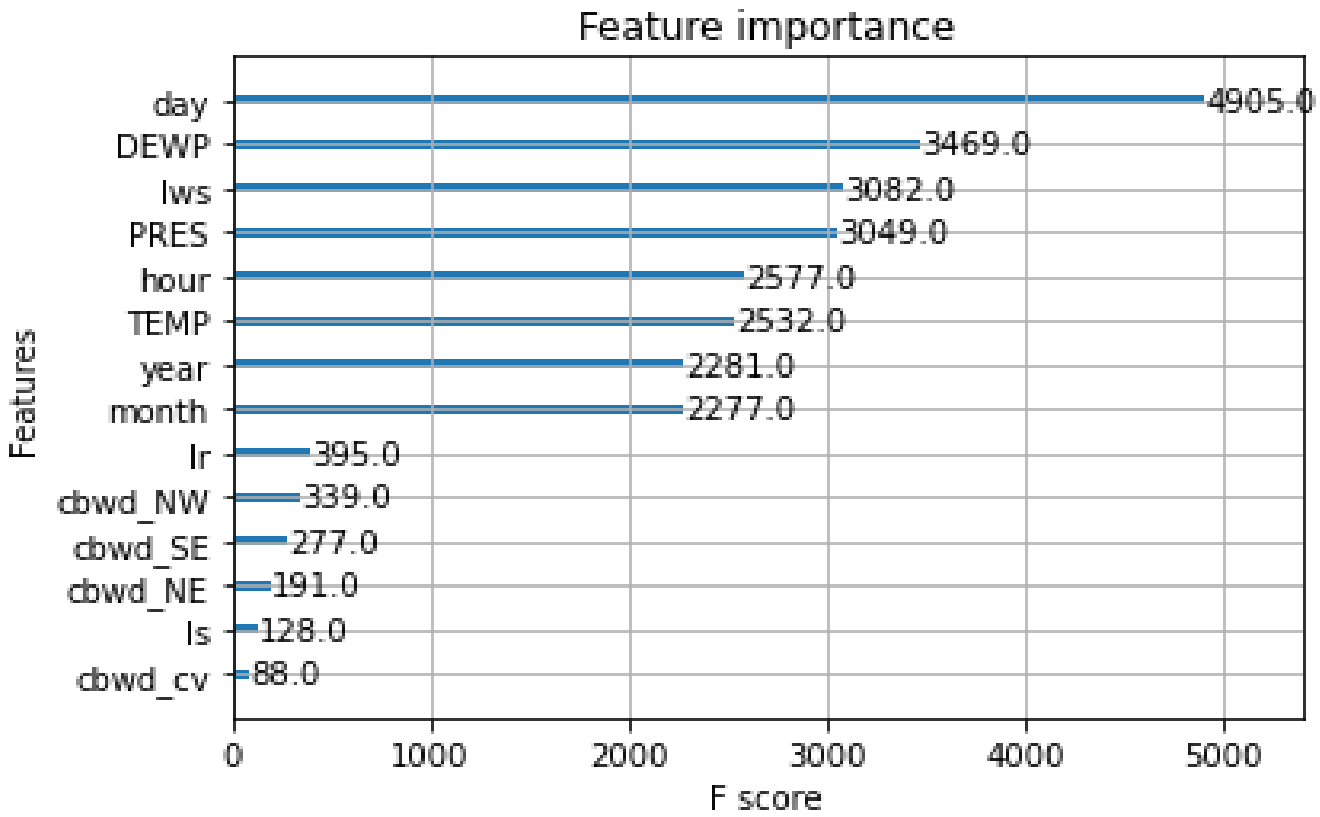

图 11: 通过 xgboost 进行特征选择

- 得到平均填补法补全的数据集的预测结果为 $R^2 = 0.6848$，略低于原数据
- 筛选后 day 仍是最重要的特征
- 气象因素中仍是 DEWP、lws、PRES、TEMP 最重要，但各个因素的重要程度有所改变，重要性依次为 DEWP、lws、PRES、TEMP、即: 湿度、累积风速、压强、温度

与 task1 类似，用相关系数矩阵进行特征选择

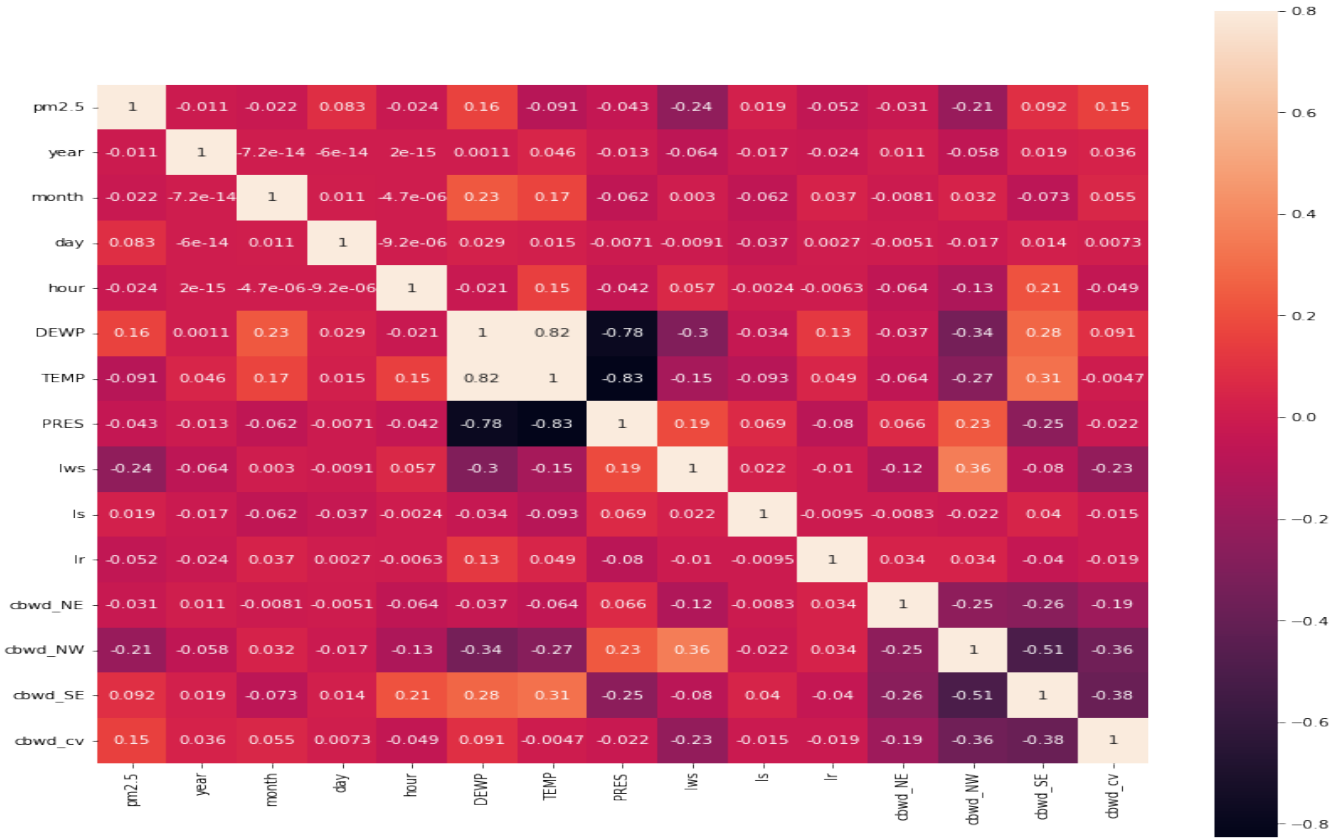

图 12: 用相关系数矩阵进行特征选择

- 筛选后结果发生变化，与 pm2.5 相关性较高的因素：DEWP、lws、cbwd_NW、cbwd_cv，即湿度、累积风速、累积风向--西北风、累积风向--平静多变

与 task1 类似，通过主成分分析选择特征并尝试降维

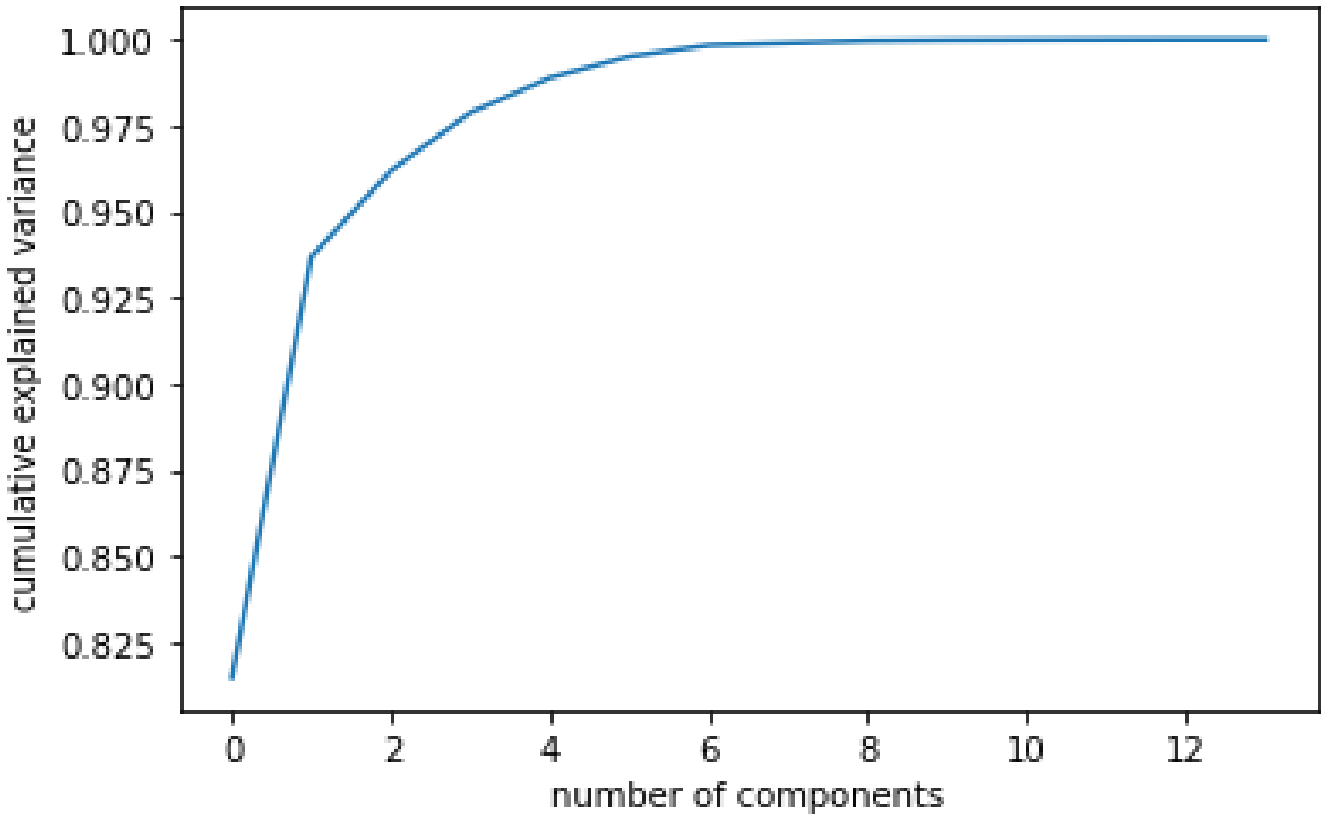

图 13: 通过主成分分析进行特征选择

- PCA 给出的结果与原数据类似，取前 3 个主成分作为特征，其中第一主成分主要反映了对 pm2.5 的影响

举例: 每年每月 pm2.5 的变化趋势

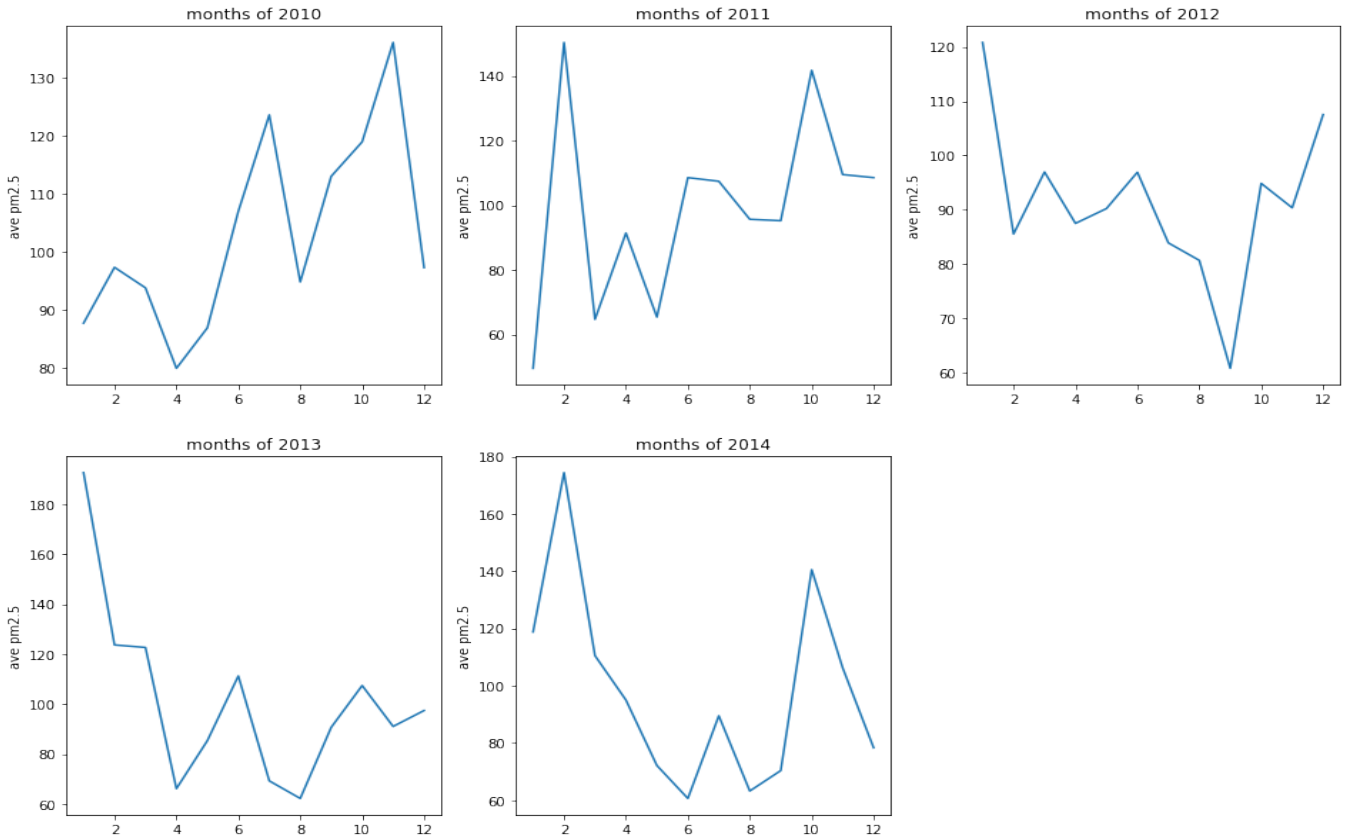

图 14: 每年每月 pm2.5 的变化趋势

- 结果与原数据的类似，每年每月 pm2.5 平均浓度变化趋势不同，但从 2011 年后，基本从 2 月到 10 月浓度维持在相对较低水平，而其余月份基本维持在相对较高水平

举例: 分别使用每年每月数据的时间信息和天气信息，对每月 pm2.5 平均浓度进行预测

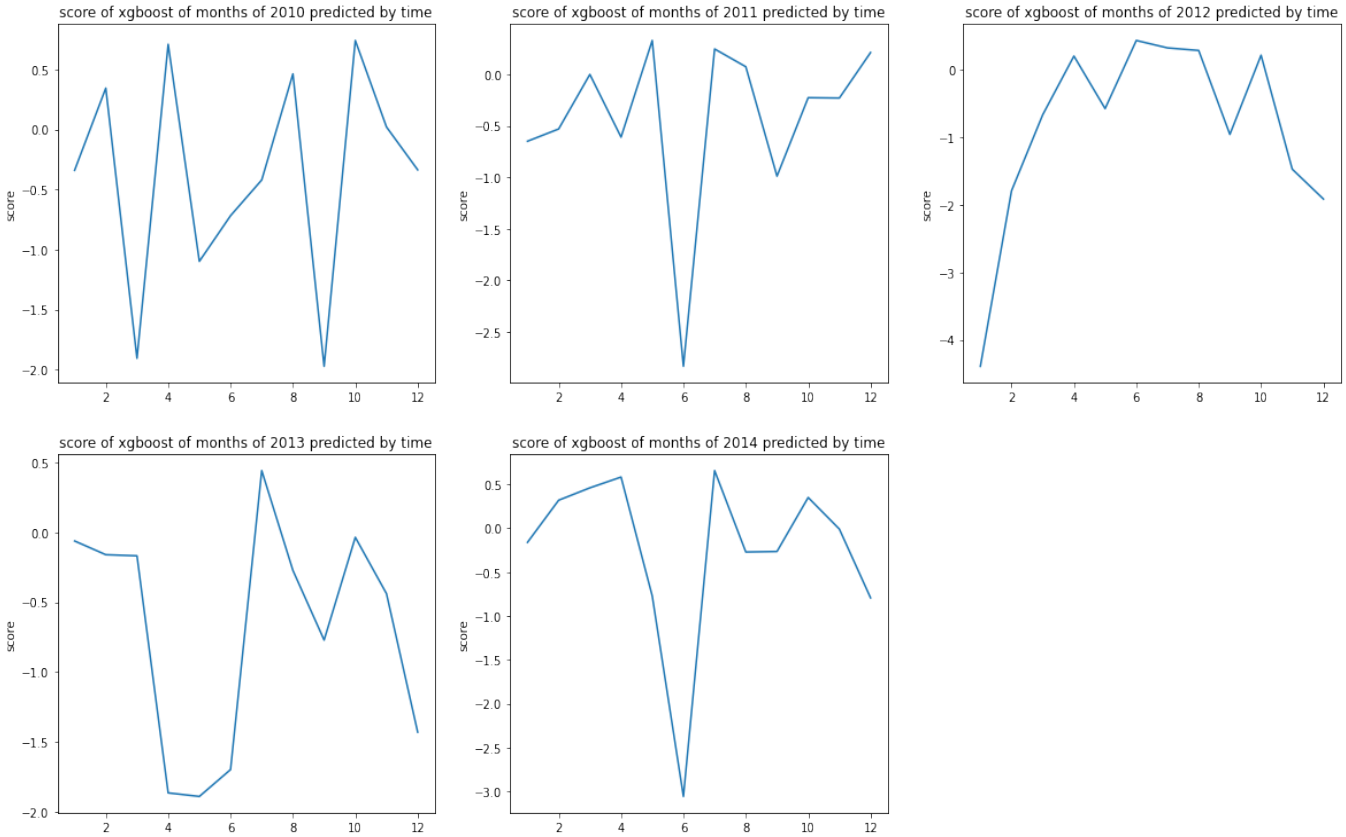

图 15: 使用每年每月数据的时间信息，得到每月 pm2.5 平均浓度的预测 score 折线图

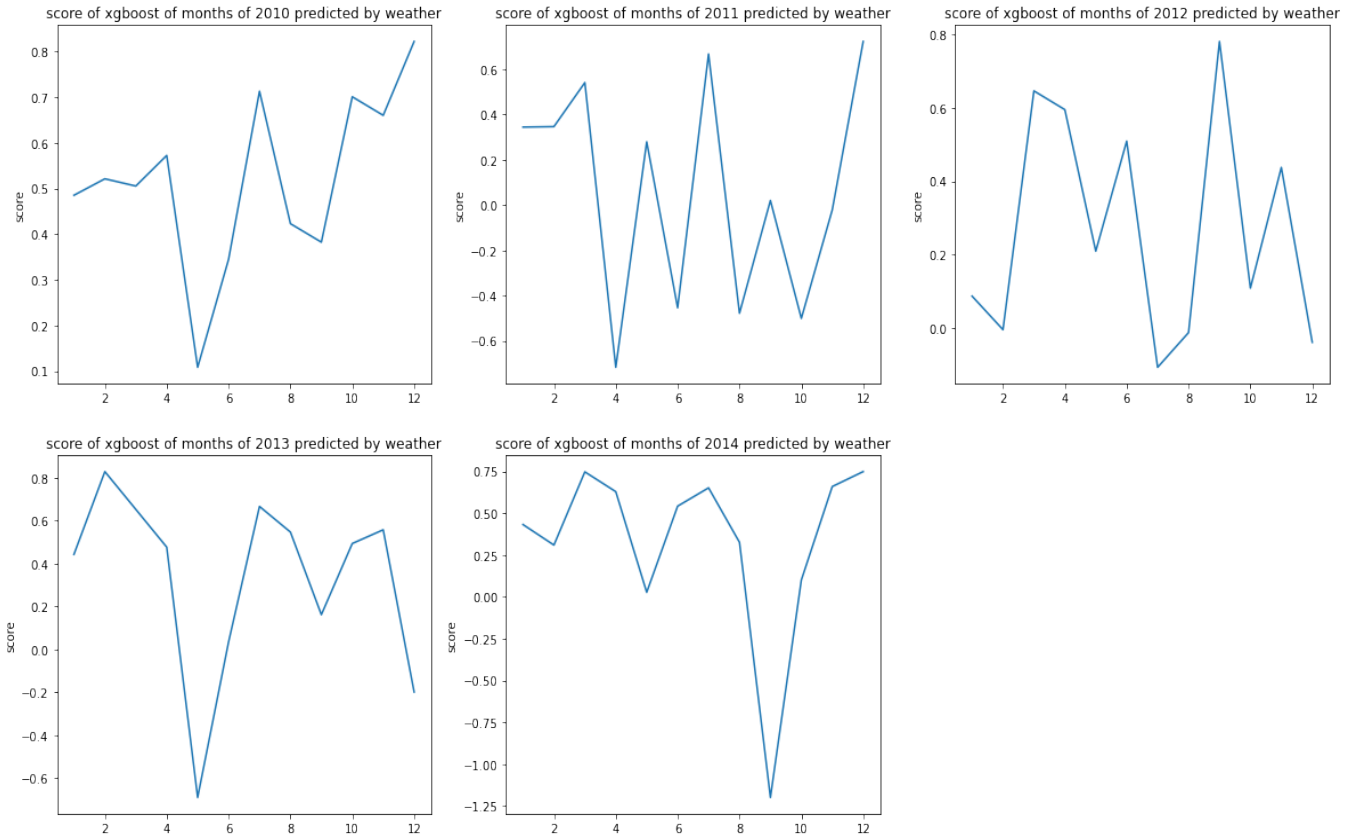

图 16: 使用每年每月数据的天气信息, 得到每月 pm2.5 平均浓度的预测 score 折线图

- 相比于原数据，两种方法给出的预测分数变化趋势和数值均与原来的相似，用每月平均浓度预测时只用天气信息可以达到更高精度，但预测准度同样更加不稳定

1 Task1&4

2 Task2

3 Task3&4

平均填补法

K 近邻填补法, $K = 40$

与 task1 类似，用 xgboost 预测模型拟合填补后的数据，并进行特征选择

与 task1 类似，用相关系数矩阵进行特征筛选

与 task1 类似，通过主成分分析选择特征并尝试降维

类似 task2, 用由 K 近邻填补法得到的新数据记录变化趋势

类似 task2，用由 K 近邻填补法得到的新数据的上述新特征对平均 pm2.5 浓度进行预测

4 Task5

与 task1 类似，用 xgboost 预测模型拟合填补后的数据，并进行特征选择

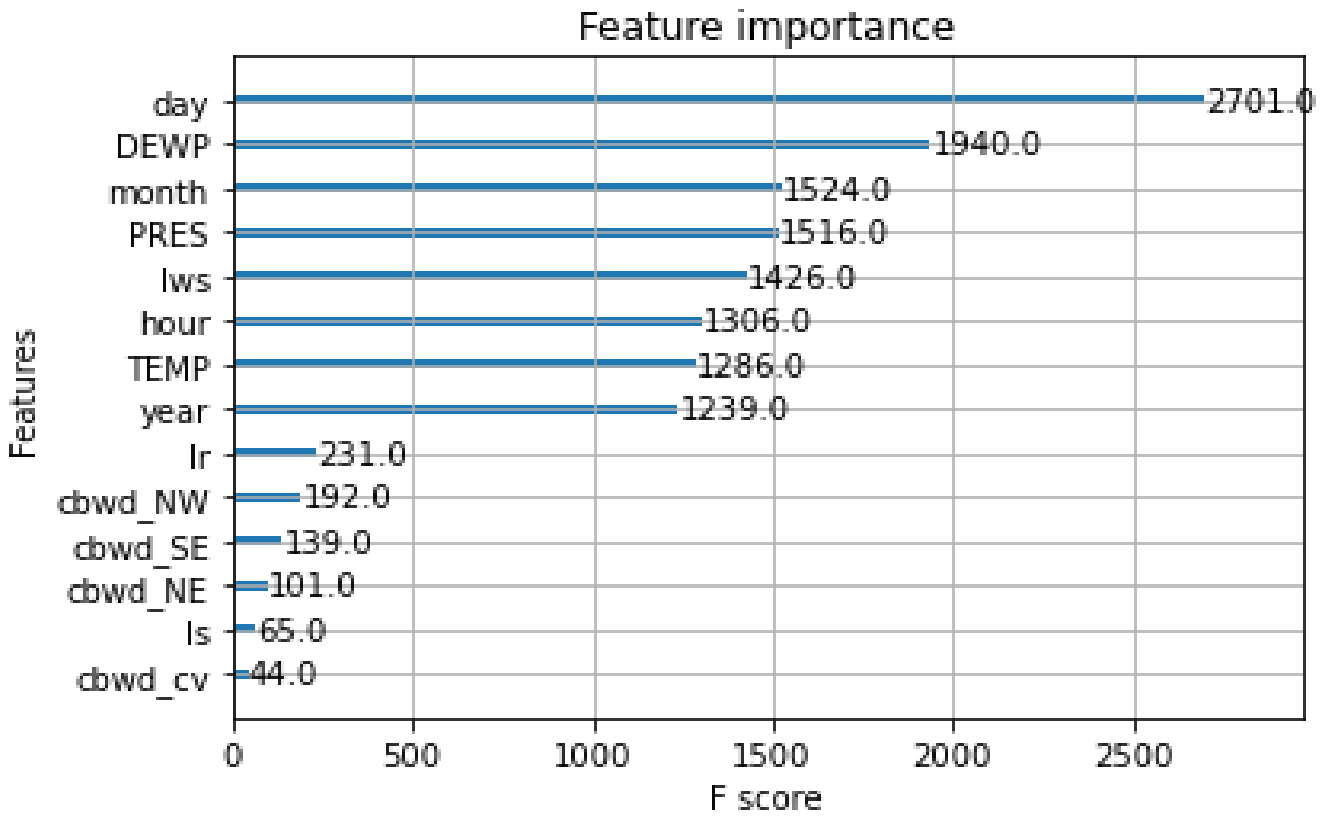

图 17: 通过 xgboost 进行特征选择

- 得到平均填补法补全的数据集的预测结果为 $R^2 = 0.6817$, 略低于原数据
- 筛选后 day 仍是最重要的特征
- 气象因素中仍是 DEWP,Iws,PRES,TEMP 最重要，各个因素的重要程度不变，重要性依次为 DEWP,PRES,Iws,TEMP, 即: 湿度、压强、累积风速、温度

与 task1 类似，用相关系数矩阵进行特征选择

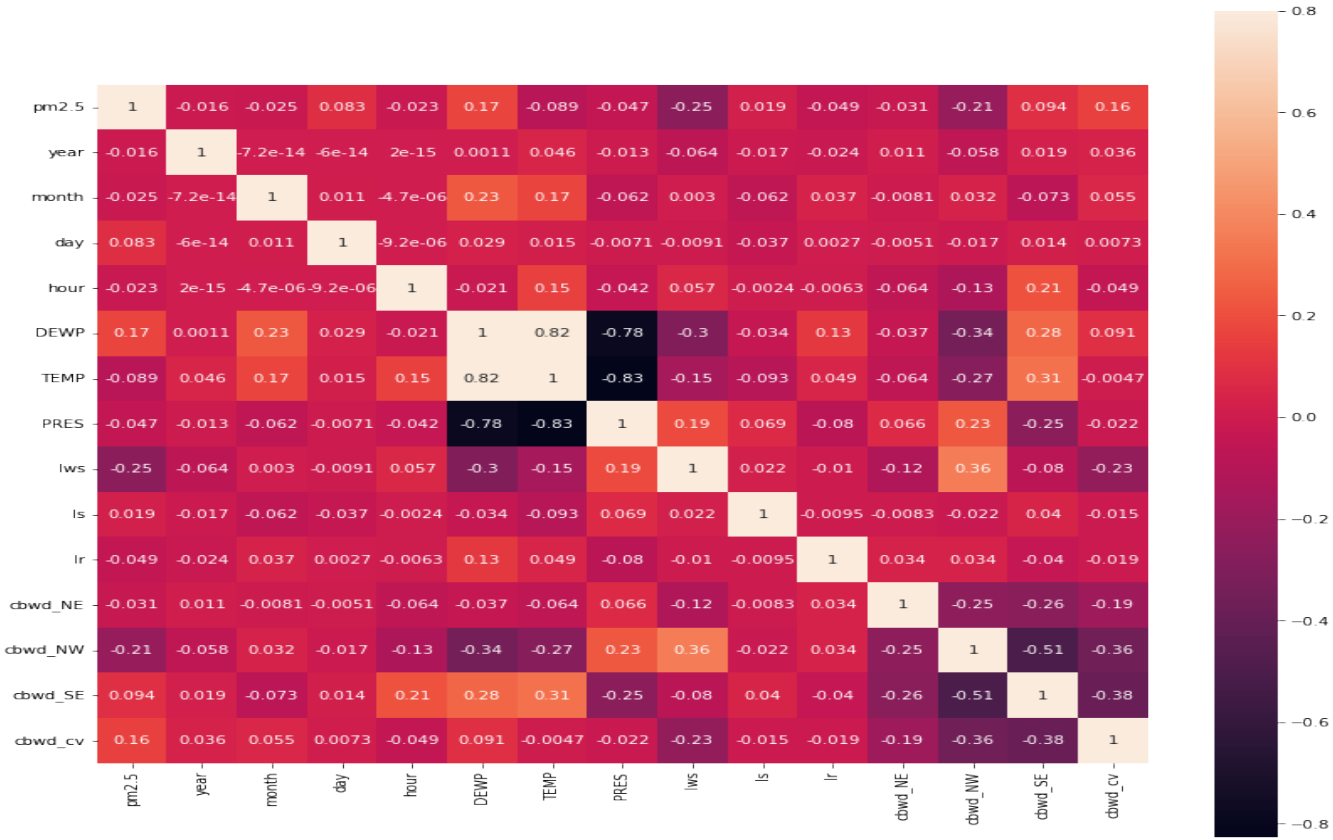

图 18: 用相关系数矩阵进行特征选择

- 筛选后结果发生变化，与 pm2.5 相关性较高的因素：DEWP、Iws、cbwd_NW、cbwd_cv，即湿度、累积风速、累积风向--西北风、累积风向--平静多变

与 task1 类似，通过主成分分析选择特征并尝试降维

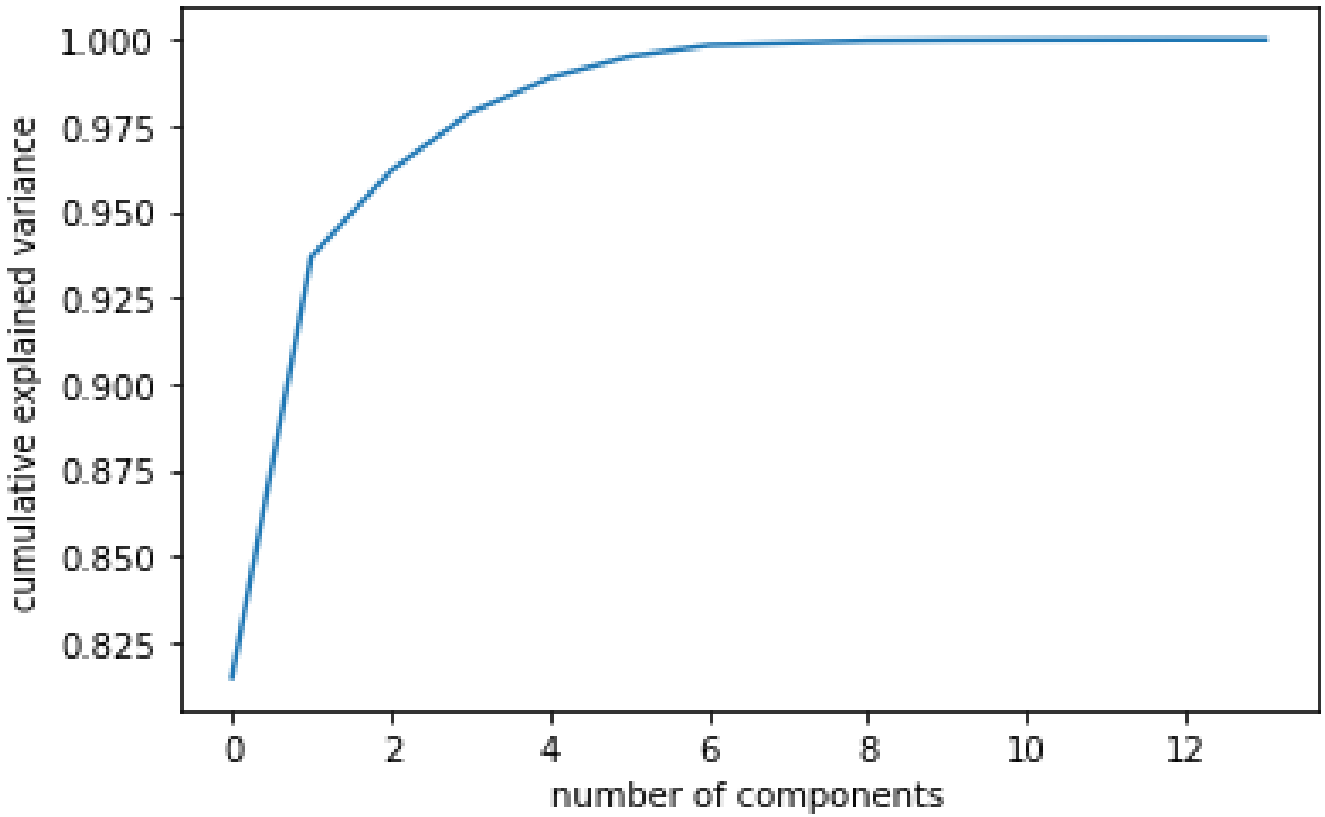

图 19: 通过主成分分析进行特征选择

- PCA 给出的结果与原数据类似，取前 3 个主成分作为特征，其中第一主成分主要反映了对 pm2.5 的影响

举例: 每年每月 pm2.5 的变化趋势

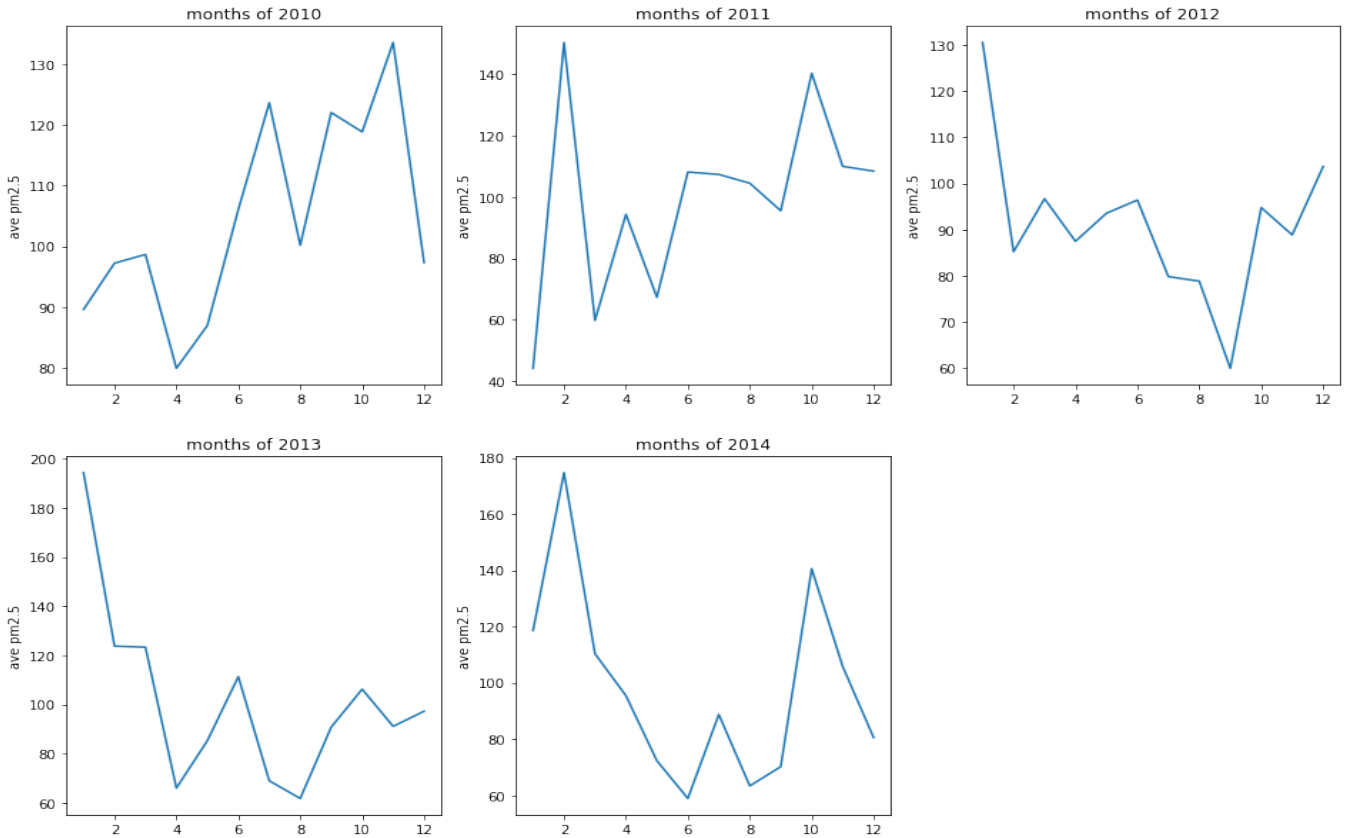

图 20: 每年每月 pm2.5 的变化趋势

- 结果与原数据的类似，每年每月 pm2.5 平均浓度变化趋势不同，但从 2011 年后，基本从 2 月到 10 月浓度维持在相对较低水平，而其余月份基本维持在相对较高水平

举例: 分别使用每年每月数据的时间信息和天气信息，对每月 pm2.5 平均浓度进行预测

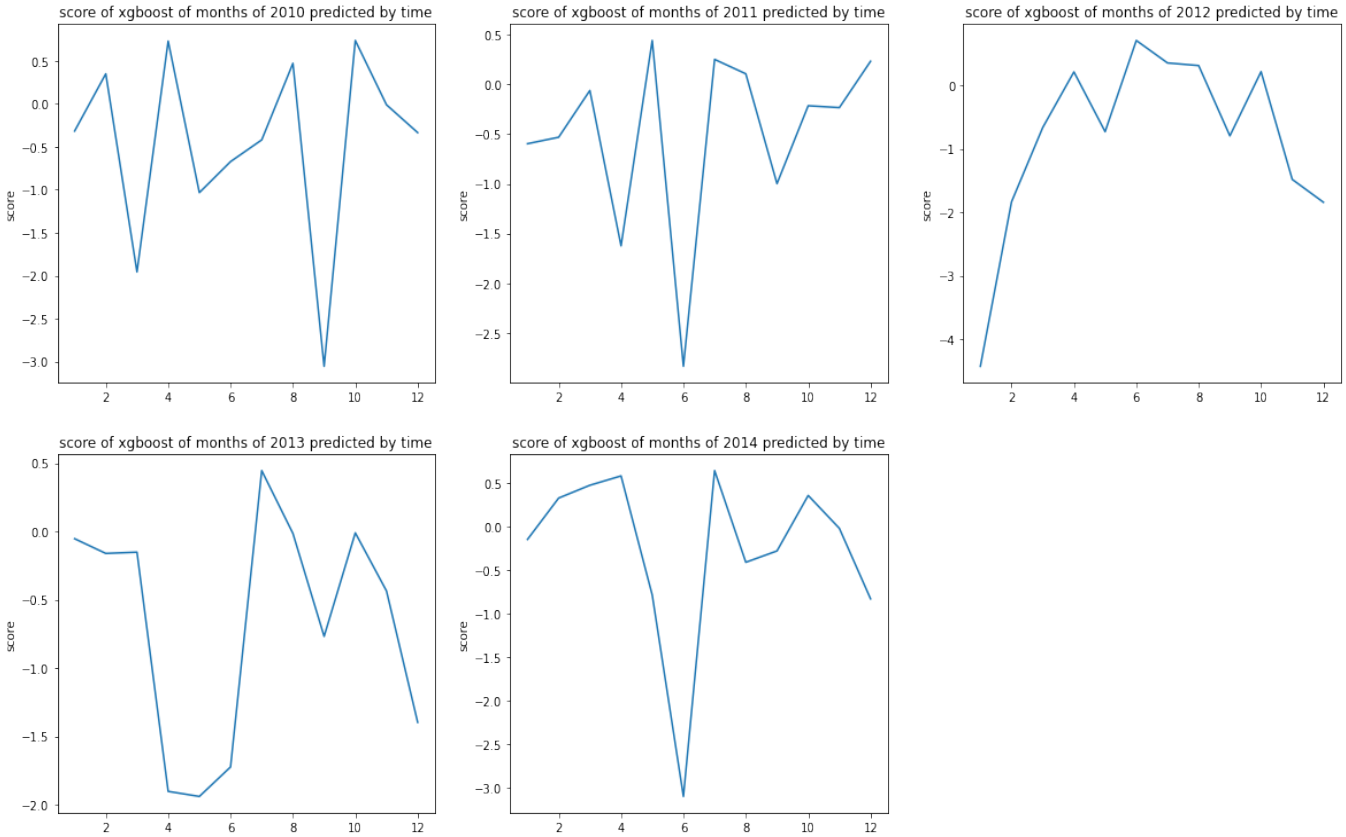

图 21: 使用每年每月数据的时间信息，得到每月 pm2.5 平均浓度的预测 score 折线图

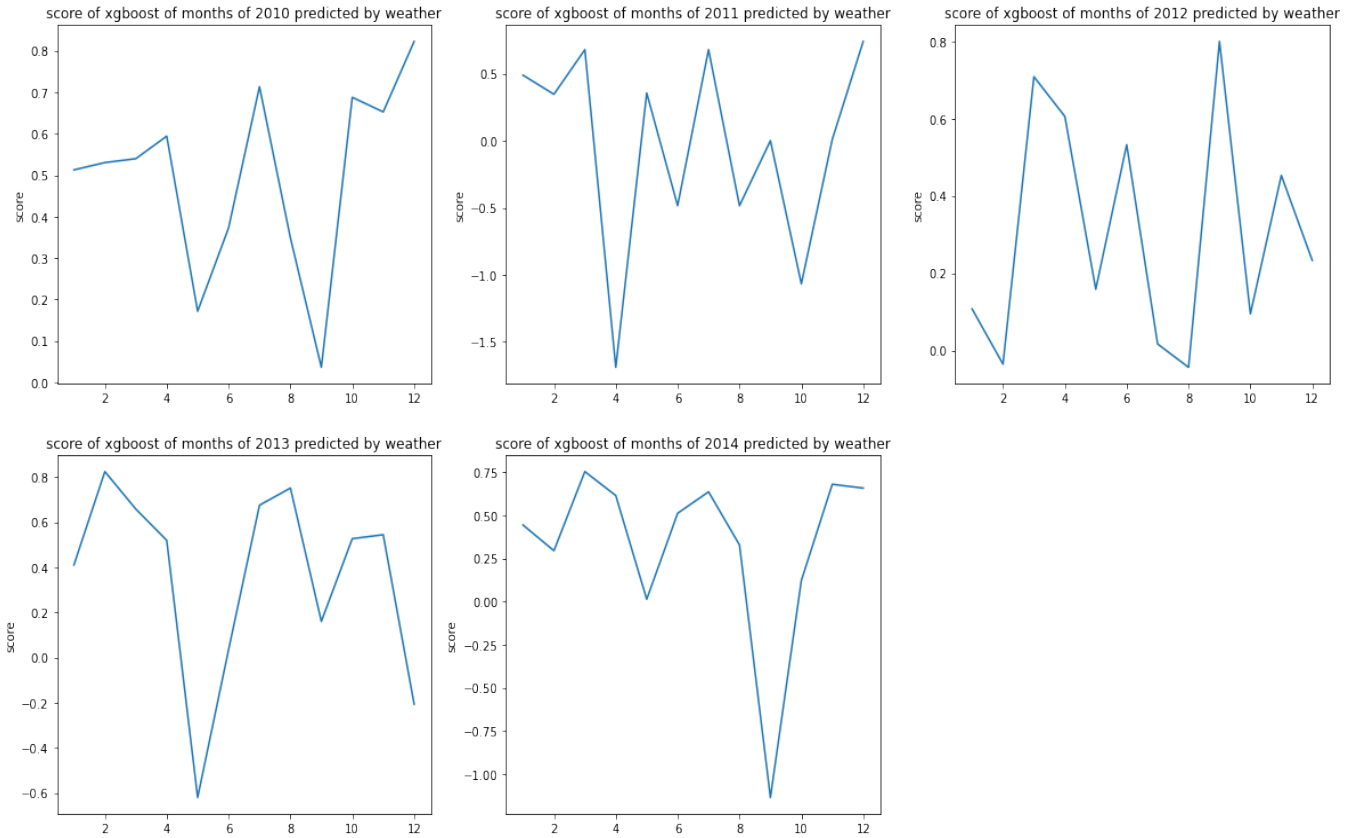

图 22: 使用每年每月数据的天气信息, 得到每月 pm2.5 平均浓度的预测 score 折线图

- 相比于原数据，除 2011 年外，两种方法给出的预测分数变化趋势和数值均与原来的相似，用每月平均浓度预测时只用天气信息可以达到更高精度，但预测准度同样更加不稳定

1. Task1&4
2. Task2
3. Task3&4
4. Task5
   - 平均填补法
   - K 近邻填补法，K=40

1 Task1&4

2 Task2

3 Task3&4

4 Task5

平均填补法

K 近邻填补法，K=40

- 使用 task3 中用平均填补得到的数据
- 将 $pm2.5$ 数据分为三类：当 $pm2.5 \leq 35$，我们记为 1 类；当 $35 < pm2.5 \leq 150$，我们记为 2 类；当 $pm2.5 > 150$，我们记为 3 类
- 用相邻 3 小时的平均数据来光滑数据
- 使用 XGBClassifier(learning_rate=0.1, n_estimators=600, random_state=0) 分类
- 模型的分类准确率为 74.55%

1. Task1&4
2. Task2
3. Task3&4
4. Task5
   - 平均填补法
   - K 近邻填补法，K=40

- 使用 task3 中用 KNN 填补得到的数据
- 将 $pm2.5$ 数据分为三类：当 $pm2.5 \leq 35$，我们记为 1 类；当 $35 < pm2.5 \leq 150$，我们记为 2 类；当 $pm2.5 > 150$，我们记为 3 类
- 用相邻 3 小时的平均数据来光滑数据
- 使用 XGBClassifier(learning_rate=0.1, n_estimators=600, random_state=0) 分类
- 模型的分类准确率为 74.60%