17-基于数据挖掘的电力负荷预测 最终报告

成员：蔡静轩(3220180783), 王峤(3220180867)

此报告记录我们所采用的模型、方法以及最终评估结果。

我们取最终得到的DataSet7.0.csv为实验数据集，随机划分 2016-2017 年的数据，取 30%的数据集为测试数据集，通过暴力寻参（6 折交叉验证）测试多种模型的效果。评价指标为R^2 与测试集上总的MAPE。

一、线性模型

线性模型主要包括了线性回归、Lasso回归、岭回归和ElasticNet回归。结果显示在以半小时为刻度区分电机的数据集上，4 种回归的效果几乎一样。R^2 接近 0.86，MAPE更是只有 0.5%。而在以天为刻度合并负荷的数据集上，效果更是几乎能够百分之百的预测准确。模型的最佳参数、最佳系数如下图所示。结果显示Hour、Equsid、F_half、F_day、F_week、rh_10 为主要特征。我们也依照电机序号将数据集划分为 6 堆，单独做模型测试，结果几乎一致。

以半小时为刻度区分电机的模型结果如下图所示：

```
! ! !  Welcome to Linear Model discovery hall ! ! !

The features used are :
['Month' 'Day' 'Hour' 'Half' 'Equsid' 'F_half' 'F_day' 'F_week'
 'dayOfWeek' 'isHoliday' 'isWorkday' 'Season' 't_10' 'rh_10']

######### The result of Linear regression #########
Best parameters set found: {'normalize': True}
Best score found:  0.8566484967073135
Optimized Score R2: 0.8526998856862216
Optimized Score MAPE: 0.005292581550465456
Best linear model parameter: [ 1.81353545e-01  1.30958519e-01 -7.28537105e-01 -2.96691183e-01
  9.46484990e-01  1.00061161e+02  2.95445379e+00  3.22909836e+00
 -2.05883279e-02 -1.84352304e-01  3.53569026e-01  3.89634248e-01
 -2.61736312e-02 -1.47179174e+00]

######### The result of ElasticNet regression #########
Best parameters set found: {'alpha': 0.01, 'l1_ratio': 0.9, 'normalize': False}
Best score found:  0.856648399574725
Optimized Score R2: 0.8527074625150409
Optimized Score MAPE: 0.00528352541012588
Best linear model parameter: [ 1.69702606e-01  1.22918355e-01 -7.13771256e-01 -2.87358190e-01
  9.50320775e-01  9.99270236e+01  2.98711792e+00  3.25491709e+00
 -2.79705840e-02 -1.81102717e-01  3.42914041e-01  3.84473721e-01
 -2.17029623e-02 -1.46667582e+00]
Actual number of iterations: 8
```

下图为以天为刻度合并负荷的模型结果：

```
! ! !  Welcome to Linear Model discovery hall ! ! !

The features used are :
['Month' 'Day' 'F_day' 'F_week' 'dayOfWeek' 'isHoliday' 'isWorkday'
 'Season' 'Tem_max' 'Tem_min' 'RH_max' 'RH_min' 'Tag']

######### The result of Linear regression #########
Best parameters set found: {'normalize': True}
Best score found:  1.0
Optimized Score R2: 1.0
Optimized Score MAPE: 1.943696621960276e-16
Best linear model parameter: [ 8.88885856e-13  1.24523548e-12  8.52861667e+03 -1.56807384e-12
  2.04439823e-12 -3.16522883e-12  5.95139893e-13  5.06839427e-12
  5.92406178e-12 -3.87428130e-13 -9.88992934e-13 -6.08862237e-13
  3.81574023e-12]

######### The result of ElasticNet regression #########
Best parameters set found: {'alpha': 0.01, 'l1_ratio': 0.9, 'normalize': False}
Best score found:  0.9999973544051367
Optimized Score R2: 0.9999976299678625
Optimized Score MAPE: 6.34554531832429e-05
Best linear model parameter: [ 3.29097760e-01  1.38080151e-02  8.51011807e+03 -2.23919898e-01
  1.03285474e+00  6.03059406e-01 -1.37404170e+00 -1.36821186e+00
 -1.36959830e+01  9.59031979e+00 -8.72683986e-01 -4.17328172e+00
  1.26410365e+01]
Actual number of iterations: 67
```

二、KNN

KNN的结果确实不理想，原因是KNN本身无法很好的利用分类特征。

```
######### The result of KNN regression #########
Best parameters set found: {'n_neighbors': 10, 'p': 2, 'weights': 'distance'}
Best score found:  0.788470257753118
Optimized Score R2: 0.8226290135393847
Optimized Score MAPE: 0.09251028818179605
```

三、集成学习

可以看出集成学习的效果也相当不错。

```
######### The result of Adaboost regression #########
Best parameters set found: {'learning_rate': 1, 'loss': 'linear', 'n_estimators': 260}
Best score found:  0.5639756409858186
Optimized Score R2: 0.6895773707904562
Optimized Score MAPE: 0.07036464509688001
Best adaboost model parameter : [2.36226894e-02 4.07345777e-02 7.21488859e-02 7.11987503e-02
 5.01435985e-02 7.14383063e-02 2.70317121e-02 2.21428004e-02
 8.03164388e-03 1.07857310e-02 9.28068706e-02 7.00137529e-02
 7.58993651e-02 8.59939099e-02 2.68216290e-01 1.58807698e-04
 9.63230838e-03]

######### The result of Random Forest regression #########
Best parameters set found: {'n_estimators': 230}
Best score found:  0.5149277069225098
Optimized Score R2: 0.6634557651485685
Optimized Score MAPE: 0.009244939143775438
Best RF model parameter : [2.49684718e-02 3.29730181e-02 7.35473357e-02 5.33438251e-02
 3.95886698e-02 6.05303491e-02 2.72459327e-02 1.35701735e-02
 2.28687127e-02 1.29848224e-02 7.62698885e-02 7.44230734e-02
 8.09883322e-02 9.87415926e-02 2.99074081e-01 2.13761232e-04
 8.66795957e-03]

######### The result of GBDT regression #########
Best parameters set found: {'learning_rate': 0.1, 'loss': 'huber', 'n_estimators': 55}
Best score found:  0.526180750023044
Optimized Score R2: 0.6515486248738318
Optimized Score MAPE: 0.07730897160701472
Best GBDT model parameter : [0.00500614 0.00699185 0.03470632 0.0120593  0.03031395 0.02031912
 0.02532315 0.03195597 0.04415288 0.00327689 0.04075306 0.0408889
 0.06442619 0.10403641 0.53046796 0.00068415 0.00463777]
```

四、基于树的模型

从理论上分析，线性模型和KNN都很难利用到分类特征的信息，而基于树的模型则恰恰相反，因此使用Adaboost、Random Forest、GBDT的最佳参数重新训练 2016~2017 年的数据，并在 2018 年的数据集上做验证。

结果如下所示：

```
! ! !  Welcome to Verify Model discovery hall ! ! !

The features used are :
['Month' 'Day' 'F_day1' 'F_day2' 'F_day3' 'F_week' 'dayOfWeek' 'isHoliday'
 'isWorkday' 'Tem_max' 'Tem_min' 'RH_max' 'RH_min' 'Tag']
Time of adaboost model for fitting:  0.7357570999999999
Time of adaboost model for predicting:  68.77130989999999

######### The result of adaboost model #########
The parameters are:  {'base_estimator__criterion': 'mse', 'base_estimator__max_depth': None, 'base_estimator__max_features':
 'sqrt', 'base_estimator__max_leaf_nodes': None, 'base_estimator__min_impurity_decrease': 0.0,
 'base_estimator__min_impurity_split': None, 'base_estimator__min_samples_leaf': 1, 'base_estimator__min_samples_split': 2,
 'base_estimator__min_weight_fraction_leaf': 0.0, 'base_estimator__presort': False, 'base_estimator__random_state': None,
 'base_estimator__splitter': 'best', 'base_estimator': DecisionTreeRegressor(criterion='mse', max_depth=None, max_features='sqrt',
           max_leaf_nodes=None, min_impurity_decrease=0.0,
           min_impurity_split=None, min_samples_leaf=1,
           min_samples_split=2, min_weight_fraction_leaf=0.0,
           presort=False, random_state=None, splitter='best'), 'learning_rate': 0.8, 'loss': 'exponential', 'n_estimators': 260,
 'random_state': None}
Optimized Score R2:  0.7550649387847259
Optimized Score Total MAPE:  0.006572157023600601
Optimized Score Mean MAPE:  0.31802068559149016
Ensemble model features selection:  [0.02342727 0.0411357  0.06538262 0.06464674 0.04500791 0.06038918
 0.02613479 0.01570085 0.01010595 0.09076511 0.07147859 0.07768453
 0.10404253 0.30409823]
```

```
! ! !  Welcome to Verify Model discovery hall ! ! !

The features used are :
['Month' 'Day' 'F_day1' 'F_day2' 'F_day3' 'F_week' 'dayOfWeek' 'isHoliday'
 'isWorkday' 'Tem_max' 'Tem_min' 'RH_max' 'RH_min' 'Tag']
Time of rf model for fitting:  1.0256346000000036
Time of rf model for predicting:  73.70138170000001

######### The result of rf model #########
The parameters are:  {'bootstrap': True, 'criterion': 'mse', 'max_depth': None, 'max_features': 'sqrt', 'max_leaf_nodes': None,
 'min_impurity_decrease': 0.0, 'min_impurity_split': None, 'min_samples_leaf': 1, 'min_samples_split': 2,
 'min_weight_fraction_leaf': 0.0, 'n_estimators': 240, 'n_jobs': None, 'oob_score': False, 'random_state': None, 'verbose': 0,
 'warm_start': False}
Optimized Score R2:  0.6918786472238587
Optimized Score Total MAPE:  0.09562418728601405
Optimized Score Mean MAPE:  0.44383947787347605
Ensemble model features selection:  [0.02387685 0.03471238 0.06404552 0.04820721 0.04343014 0.05311927
 0.02895269 0.01255174 0.02458419 0.08422578 0.0662677  0.08458858
 0.10462656 0.3268114 ]
```

```
! ! !  Welcome to Verify Model discovery hall ! ! !

The features used are :
['Month' 'Day' 'F_day1' 'F_day2' 'F_day3' 'F_week' 'dayOfWeek' 'isHoliday'
 'isWorkday' 'Tem_max' 'Tem_min' 'RH_max' 'RH_min' 'Tag']
Time of gbdt model for fitting:  0.11984830000000102
Time of gbdt model for predicting:  61.97301519999999

######### The result of gbdt model #########
The parameters are:  {'alpha': 0.9, 'criterion': 'friedman_mse', 'init': None, 'learning_rate': 0.2, 'loss': 'lad', 'max_depth':
 3, 'max_features': None, 'max_leaf_nodes': None, 'min_impurity_decrease': 0.0, 'min_impurity_split': None, 'min_samples_leaf':
 1, 'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0, 'n_estimators': 80, 'n_iter_no_change': None, 'presort': 'auto',
 'random_state': None, 'subsample': 1.0, 'tol': 0.0001, 'validation_fraction': 0.1, 'verbose': 0, 'warm_start': False}
Optimized Score R2:  0.7860011949738495
Optimized Score Total MAPE:  0.02252399083763418
Optimized Score Mean MAPE:  0.3099722314785605
Ensemble model features selection:  [0.01049289 0.04242693 0.0729443  0.04950494 0.04804042 0.06719487
 0.01236728 0.02478481 0.01425423 0.06324431 0.05346933 0.07617238
 0.10428284 0.36082048]
```

Adaboost模型的R^2达到了 0.74，总MAPE为 0.66%，而日MAPE绝对值的均值为 32%，这表明模型对整体的拟合较好，但是存在一些局部的过拟合或欠拟合。

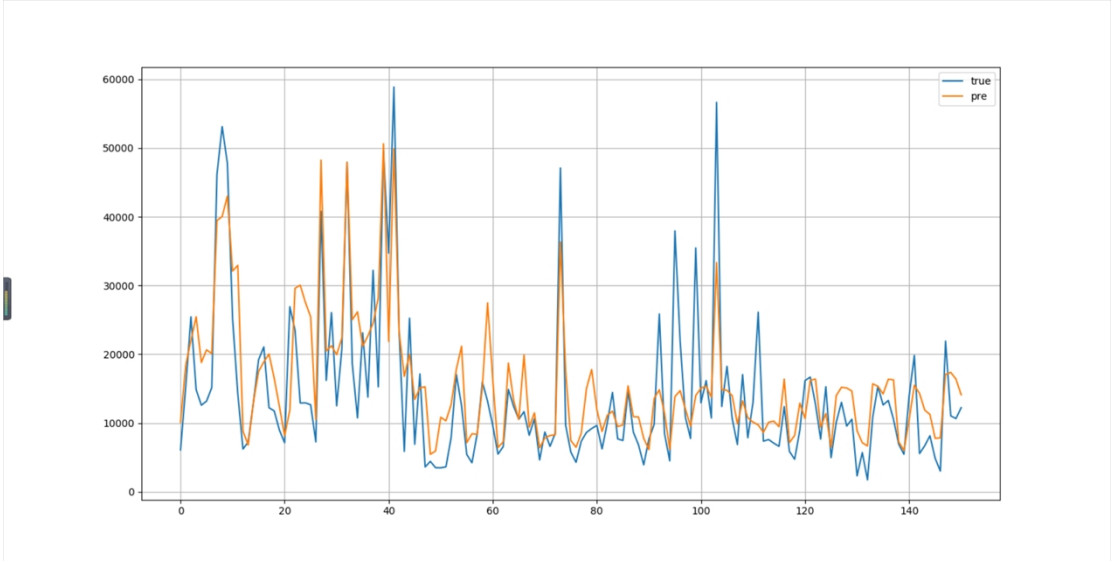
同样的RF与GBDT模型得出的结果与Adaboost类似。

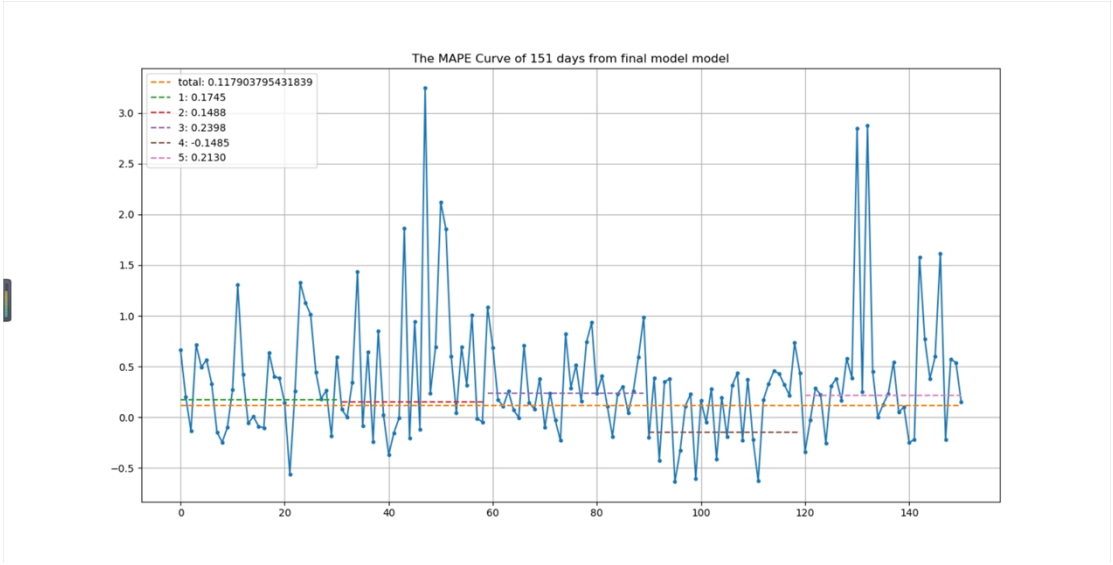
模型融合

考虑用线性回归或均值拟合 2018 年的负荷曲线。

1、 均值拟合

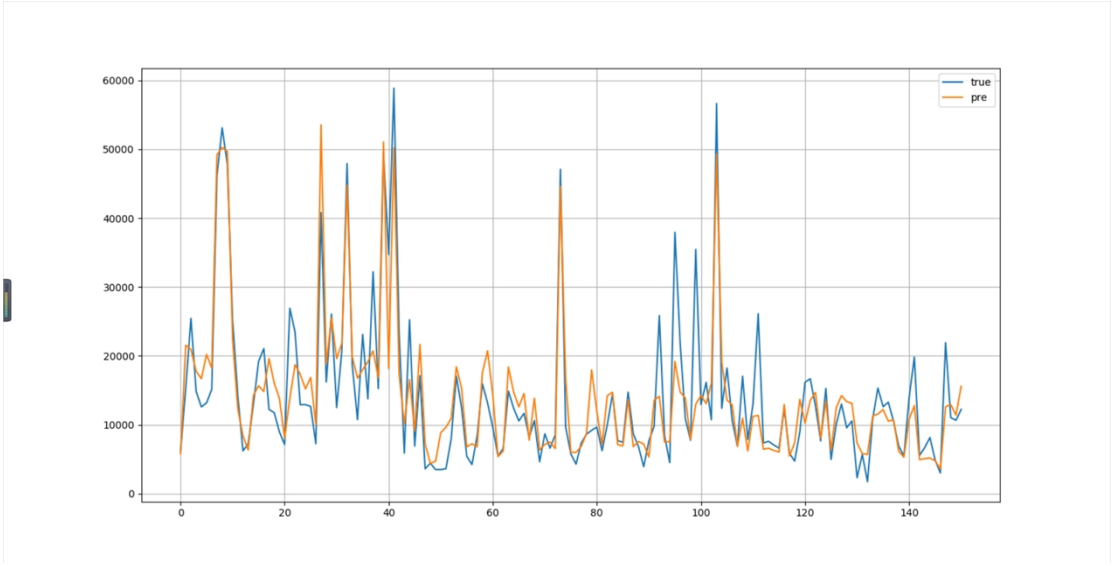
均值拟合的效果似乎变得更加不好了。

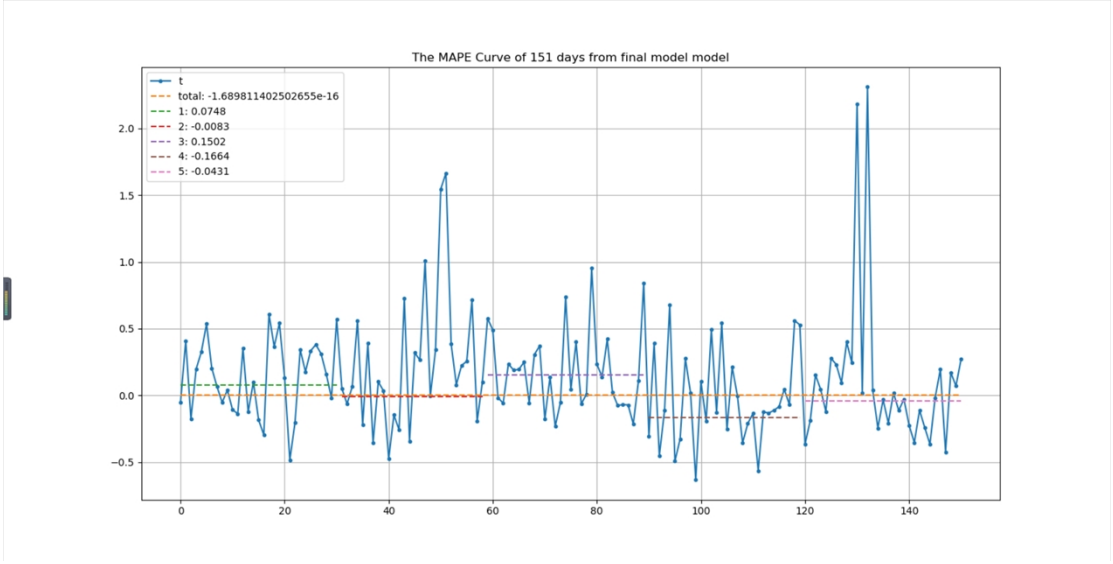




2、 线性回归

使用线性回归拟合负荷曲线的效果较好，5 个月总的MAPE几乎为 0。3/4 月份的MAPE也得到了一定的修正。





总结

在此项目过程中，结合课上所学内容，对数据做了大量的清洗和预处理工作，保证了后面实验过程的顺利进行。通过线性模型、KNN、集成学习及基于树的模型，得到了较好的效果，完成了预设目标。同时，对于相对应的算法、模型及工具等有了进一步的认识，得到了实际的经验。