# 数据科学导论大作业报告-LMP 短期预测

王铮澄 林啸 牛腾腾

2019 年 1 月 16 日

## 一、 选题意义

节点边际电价（Locational marginal price）是一种基于最优潮流的电价算法，这种方式得到的节点电价会受到 4 个因素影响：发电机边际成本，系统容量，网损和线路阻塞情况。而节点电价电力市场在国外已经有多年的发展和使用历史。不同于我国电力市场统一定价的方式，在基于节点电价的电力市场中，每个节点的电价都可能不相同，受到上述 4 个因素影响。一般来说，当线路出现阻塞或是某区域电力负荷较高时，节点电价会被抬高。典型的使用节点电价的电力市场：美国 PJM 电力市场、加州电力市场 CAISO、澳洲电力市场、欧洲电力市场。我国即将在广东试行节点电价电力市场，若能准确预测次日 LMP，发电商、售电商就能调整竞价策略，获取更高的利润，有其实用意义。此外，对节点电价的预测本质上是对时间序列的预测，可以在这个项目中很好地结合本课程所学。

## 二、 项目目标

我们的预测目标是基于历史数据，对次日 24 小时每小时的日前市场节点电价进行预测。在前期文献调研中，国外对于节点电价预测的准确度在 MAPE 误差之下大都在为 10% 左右，并且对尖峰点的预测效果普遍不好，因此我们将这次大作业的目标定为预测 MAPE 误差在 12% 以内，并且通过图像来判断尖峰点的预测效果，尽量优化尖峰点处的预测精度。

## 三、 数据分析

项目数据来源是美国 PJM 市场官网（www.dataminer2.pjm.com），在前期数据分析中，发现这个网站的数据种类较多，但是数据质量参差不齐，而且数据格式并不统一，例如每日发电容量、每小时负荷测量数据等更新不及时，最新的数据一般是几天甚至几个月之前的，这对于我们需要的次日预测目标来说是不合用的，因此在选取训练数据的时候需要对数据的实时性加以分析。此外网站上数据的组织形式不统一，电价类数据是根据时间和节点进行划分的，每个节点会给出 24 小时的电价，而负荷类数据是根据区域划分的，并不精确到节点，发电类数据则是根据能源种类和更大的区域来划分的，因此都需要专门处理，得到可以训练的向量形式。

# 四、 符号说明

| 符号 | 含义 |
|---|---|
| $load_f^t$ | 预测日当天时间 t 的负荷 |
| $lmp_{f-n}^t$ | 预测日前 n 天时间 t 的总电价 |
| $cop_{f-n}^t$ | 预测日前 n 天时间 t 的电价拥堵分量 |
| $mlp_{f-n}^t$ | 预测日前 n 天时间 t 的电价线损分量 |
| $t$ | 时间 t 标记，取值为 1 到 24 之间的整数 |
| $spike\_mark_f^t$ | 预测日时间 t 的电价尖峰标记，非尖峰为 0 |

# 五、 第一阶段

## (一) 特征选择与误差评估

[$load_f^t$、$lmp_{f-1}^t$、$lmp_{f-2}^t$、$lmp_{f-7}^t$、$t$] 共 5 个特征，在尖峰优化设想当中加入了‘尖峰电价标记’特征。

误差函数的选择为 MAPE。

## (二) 数据选取与划分

本次选择的数据为美国 PJM 电力市场 5021071 号节点的 daily ahead lmp 数据，时间跨度为 2018.7.2—2018.10.1，其中选择 7 月 2 日到 9 月 17 日的数据作为训练集，9 月 18 日到 10 月 1 日的数据作为测试集。

## (三) 模型评估

### 1 MLP

MLP 模型的准确率不是特别稳定，在参数固定的情况下每次运行得到的准确度有 1% 左右的波动。

基本参数为：

| 参数类型 | 参数详情 |
|---|---|
| 隐层数量 | 2 |
| 隐层节点 | (100，80) |
| 激活函数 | ‘relu’ |
| 迭代次数 | 50000 |
| 学习率 | 0.1 |
| 优化器 | ‘Adam’ |

以下为连续三次运行得到的训练集误差率和测试集误差率：

| 训练集误差/% | 测试集误差/% |
|---|---|
| 6.13 | 13.01 |
| 6.36 | 12.42 |
| 6.06 | 13.75 |

预测图像为：

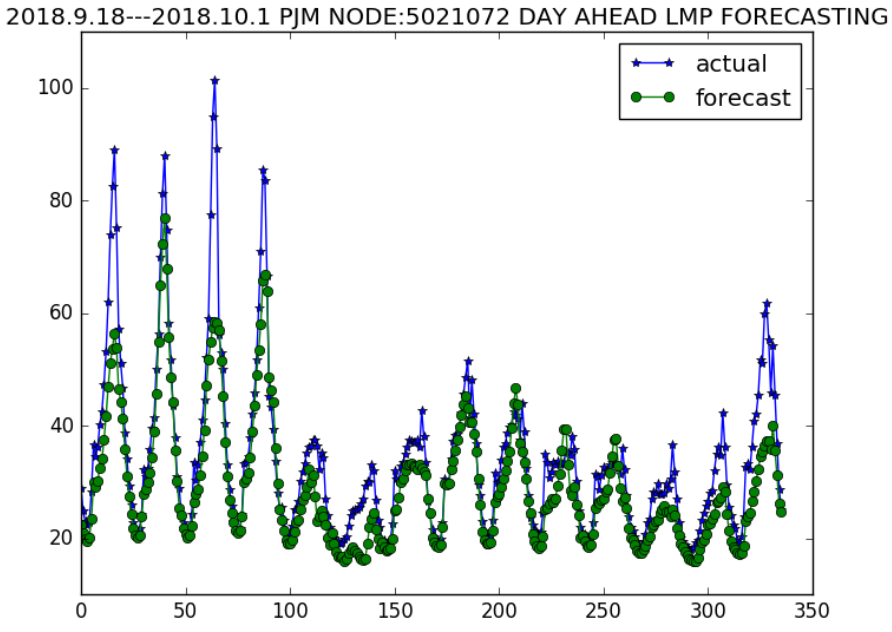


图 1: MLP 预测

## 2 SVM

基本参数为：

| 参数类型 | 参数详情 |
|---|---|
| C | 0.1 |
| gamma | 0.01 |
| 核函数 | ‘rbf’ |

误差率稳定

| 训练集误差/% | 测试集误差/% |
|---|---|
| 7.44 | 11.93 |

预测图像为：

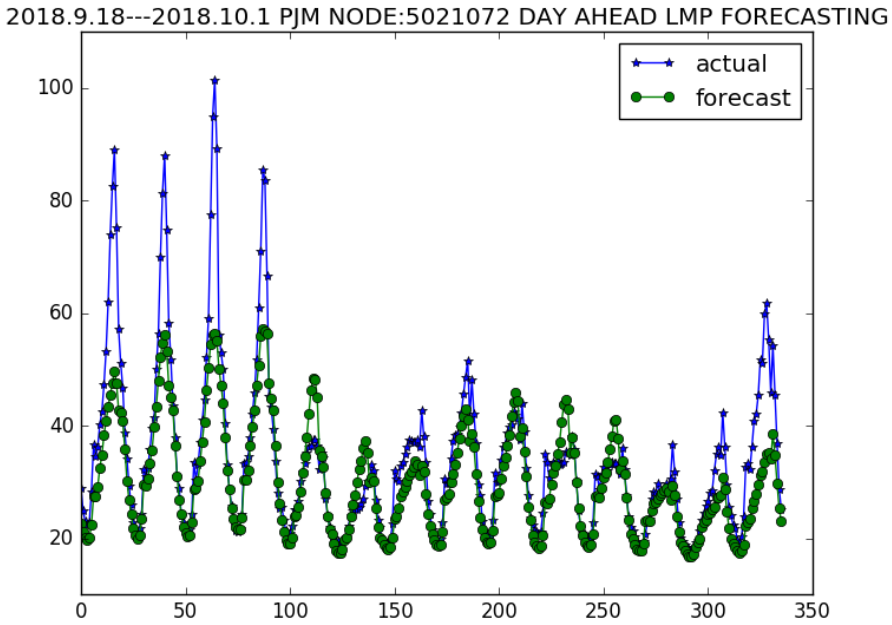


图 2: SVM 预测

## 3 RF

基本参数为：

| 参数类型 | 参数详情 |
|---|---|
| n_estimators | 60 |
| max_depth | 5 |
| max_features | 2 |

误差率较为稳定，测试集误差波动在 0.5% 以内，以下为连续 3 次训练的误差：

| 训练集误差/% | 测试集误差/% |
|---|---|
| 6.45 | 10.23 |
| 6.48 | 10.68 |
| 6.50 | 10.48 |

预测图像为：

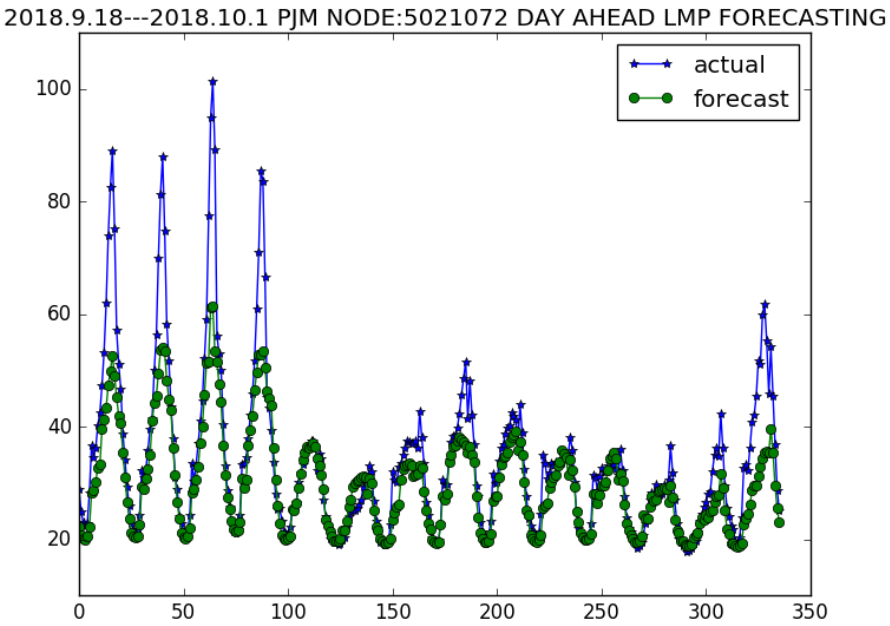


图 3: RF 预测

## 4 LGBM

基本参数为：

| 参数类型 | 参数详情 |
|---|---|
| num_leaves | 4 |
| max_depth | 3 |
| reg_lambda | 7 |

误差率稳定

| 训练集误差/% | 测试集误差/% |
|---|---|
| 6.51 | 10.73 |

预测图像为：

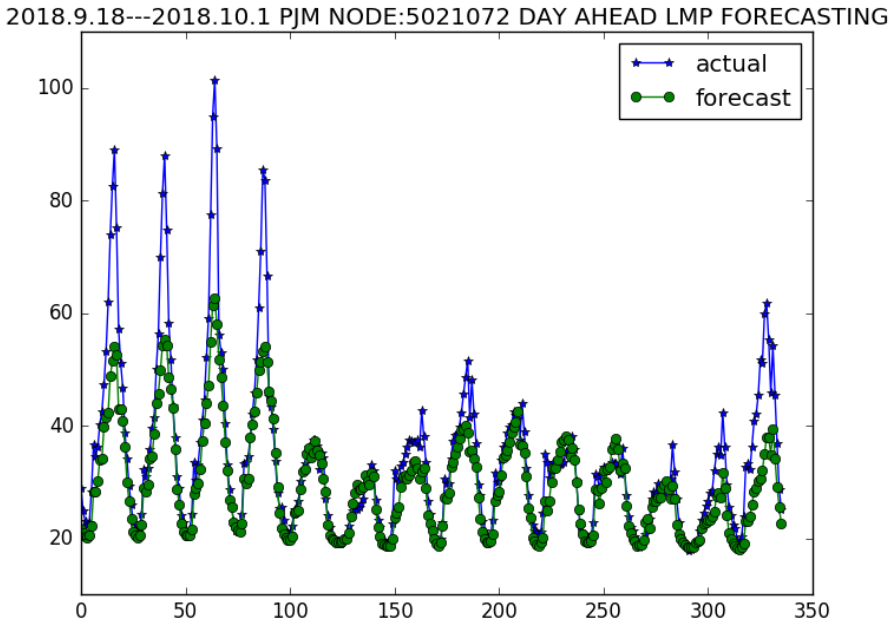


图 4: LGBM 预测

## 5 XGB

基本参数为：

| 参数类型 | 参数详情 |
|---|---|
| n_estimators | 150 |
| max_depth | 2 |
| reg_lambda | 7 |
| gamma | 0.01 |
| colsample_bytree | 0.4 |
| subsample | 0.6 |
| learning_rate | 0.1 |

误差率稳定

| 训练集误差/% | 测试集误差/% |
|---|---|
| 6.71 | 10.02 |

预测图像为：

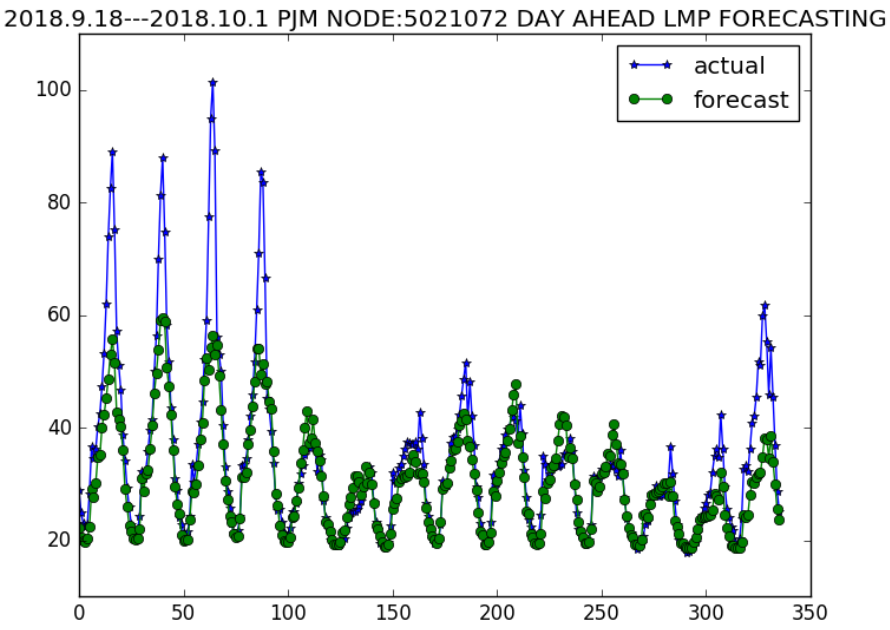


图 5: XGB 预测

### 6 STACK

将以上五种模型进行 stacking，将表现最好的 XGB 模型作为第二层的回归模型。参数选择与每种模型单独的参数一样。以下列出每种模型单独的 MSE Score 和其标准差，并将每种模型的训练集、测试集误差汇总。不稳定模型的误差取三次均值。

| 模型名称 | MSE Score | 标准差 | 训练集误差/% | 测试集误差/% |
|---|---|---|---|---|
| XGB | -12.25 | +/- 0.0702 | 6.71 | 10.02 |
| RF | -13.43 | +/- 0.0751 | 6.48 | 10.46 |
| LGBM | -13.60 | +/- 0.0891 | 6.51 | 10.73 |
| SVM | -15.18 | +/- 0.0861 | 7.44 | 11.93 |
| MLP | -13.04 | +/- 0.0839 | 6.18 | 13.06 |
| STACK | -12.28 | +/- 0.0787 | 5.58 | 11.08 |

预测图像为：

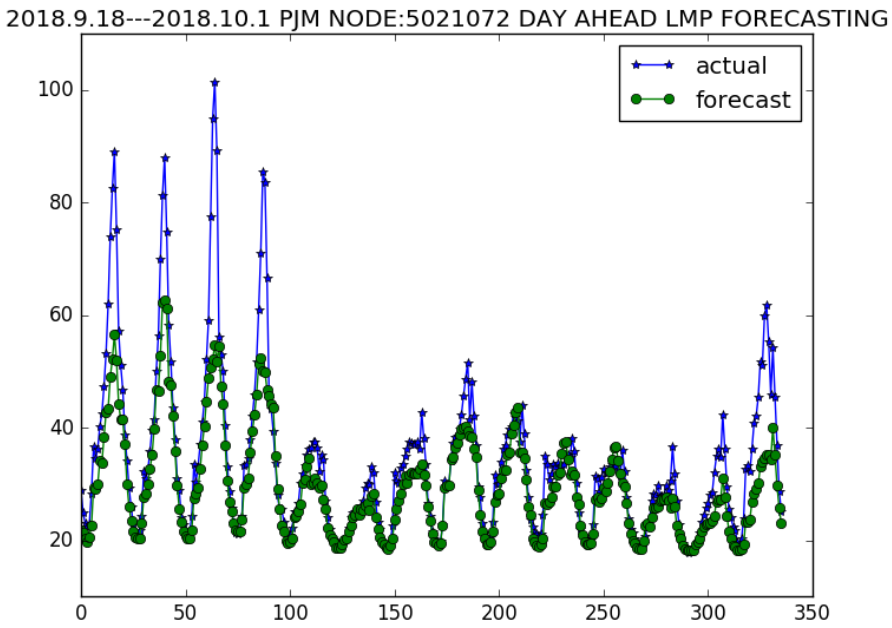


图 6: STACK 预测

## (四) 尖峰优化

观察到以上模型得到的图像中，对于尖峰点的预测普遍不理想，因此我考虑加入其它特征尝试优化尖峰处的预测。我将训练集中的电价进行分段，先计算出训练集电价的均值 $p$ 和标准差 $\delta$，然后以

$p+\delta$、$p+1.5\delta$、$p+2.5\delta$ 作为分界点将电价分为三个等级，分别标记为 0,1,2,3。而对于测试集，由于电价事先未知，但是根据预测结果可以看出趋势大致是准确的，所以我考虑先进行第一次训练，然后将预测结果重复对训练集的操作，给预测的电价也进行标号，然后将得到的标号作为特征再次训练，也就是说，在第二次训练的时候我们才将标号作为新特征加入。由于第一次训练之后预测的峰值较小，可能不能被很好地识别和增强，因此我考虑采用多次训练、重标号的办法，使峰值处的电价得以增强。

此外，考虑到尖峰电价的形成与线路拥堵状况和损耗状况密切相关，我将这两项也作为特征纳入考虑，在尝试了几种特征组合后，以下的特征表现较好: [$load_f^t$、$lmp_{f-1}^t$、$lmp_{f-2}^t$、$lmp_{f-7}^t$、$t$、$cop_{f-1}^t$、$cop_{f-7}^t$、$mlp_{f-1}^t$、$mlp_{f-1}^t$、$spike\_mark_f^t$]　　由于 XGB 模型的效果较好，我选择 XGB 进行尖峰优化。为了对比，XGB 模型的参数没有调整。在进行了 4 轮迭代之后，得到的结果如下：

| 训练集误差/% | 测试集误差/% |
|---|---|
| 5.35 | 8.86 |

预测图像为：

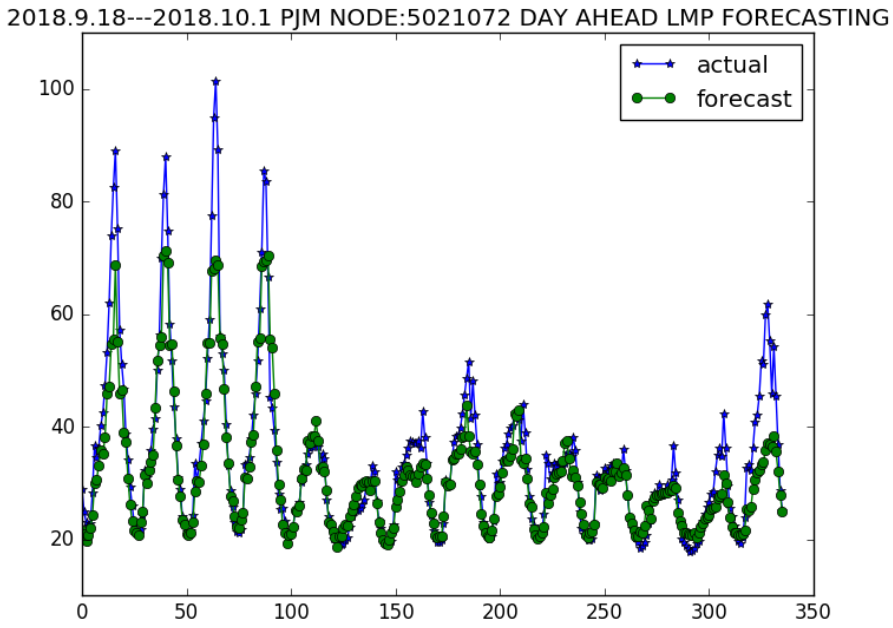


图 7: XGB 引入电价标记之后的预测结果

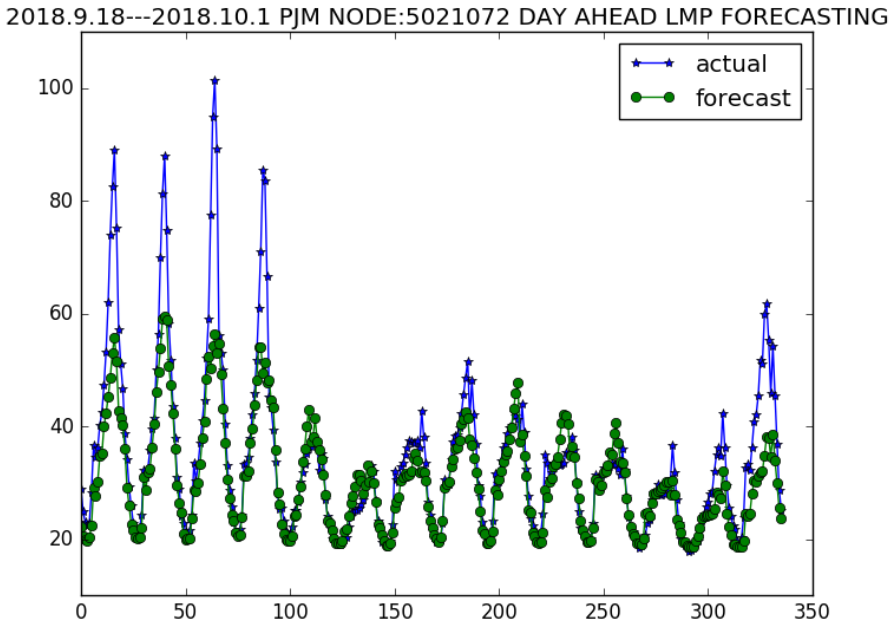


图 8: XGB 未引入电价标记的预测结果

从结果可以发现，训练集和测试集上的误差均下降，测试集误差下降大约 1.2%，并且在图像上看出尖峰点有所改善，可以认为该方法确实起到了一定作用。

### (五) 结果评估

虽然在调参以及优化之后准确度较高，但是在测试其它节点的时候，我们发现这个模型的效果并不好，误差甚至可以达到 15% 左右。分析原因，是因为选取的测试数据不够有代表性，这个区间内波形的周期性较为明显，并且几乎没有低洼下陷的波形，因此对特征更丰富的数据预测效果较差。因此我们考虑进行第二阶段的探究，使用更多的数据进行训练，测试集也选用更长时间尺度的数据来模拟多种情况的预测。

## 六、 第二阶段

第二阶段的具体操作步骤与第一阶段大同小异，这里就不再详细描述。具体区别是，我们选用了包括电价分量，负荷，各类能源发电量，电力断供量等更多数据，组成了长度为 44 的特征向量。并且我们使用了接近一年的数据作为训练，两个月的数据进行测试。使用的模型是 3 层 MLP。

由于数据量较大，特征较多，我使用了自编码器进行主成分分析和特征提取，自编码器的结构设计是 4 层神经网络，每层以 75% 的比例缩减。具体见下图：

```python
self.encoder = nn.Sequential(
    nn.Linear(44, 37),
    nn.ReLU(),
    nn.Linear(37, 31),
    nn.ReLU(),
    nn.Linear(31, 26),
    nn.ReLU(),
    nn.Linear(26, 22),
)
```

图 9: 编码器结构

## (一) 无编码器结果

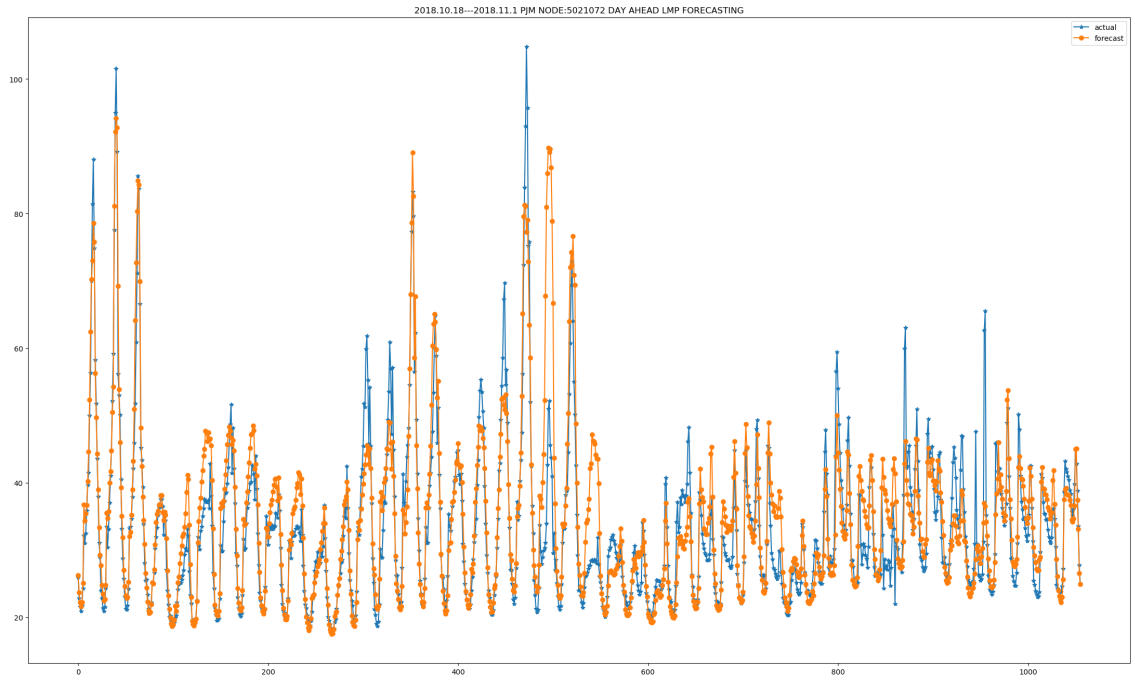


图 10: 无编码器

无编码器时训练误差为 3.93%，测试误差为 10.64%，怀疑有过拟合迹象。

## (二) 有编码器结果

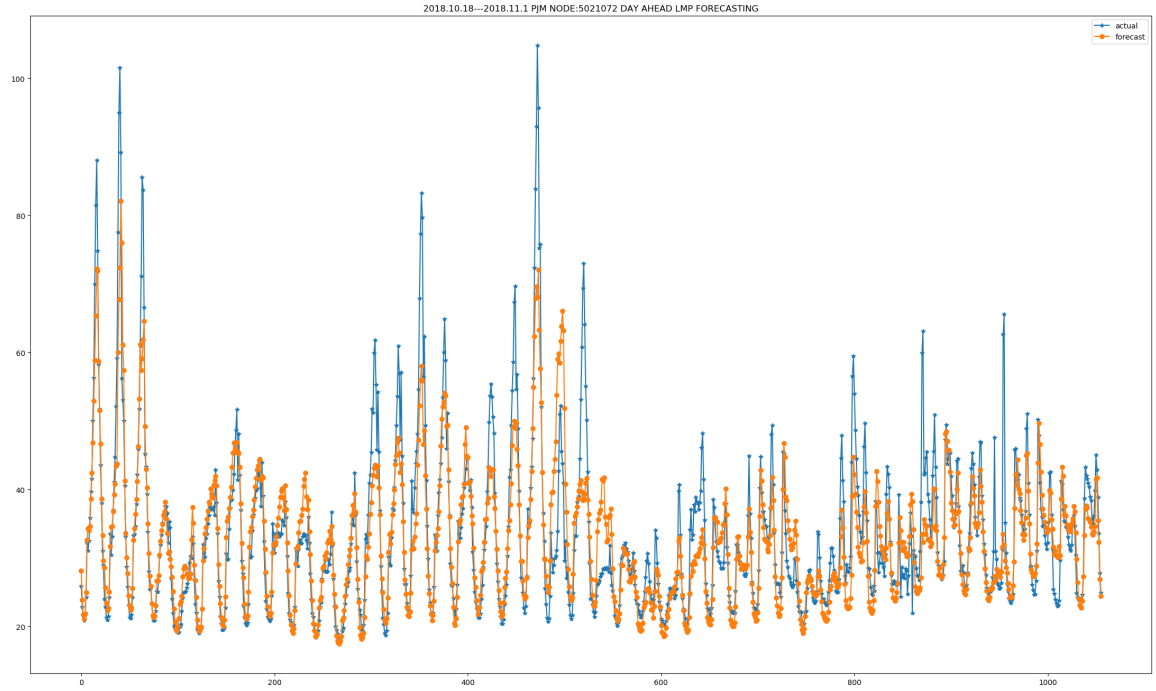


图 11: 有编码器

有编码器时训练误差为 7.03%，但是测试误差降至 10.19%，训练误差接近测试误差，模型泛化能力变强。

## (三) 结果分析

从误差结果和预测图象上看，数据量加大之后，预测效果明显变好，并且对尖峰点的预测效果大大增强，自编码器的使用降低了模型运算量和复杂度，并且提升了模型的泛化能力。预测误差已经达到目标，这次大作业基本得到完成。