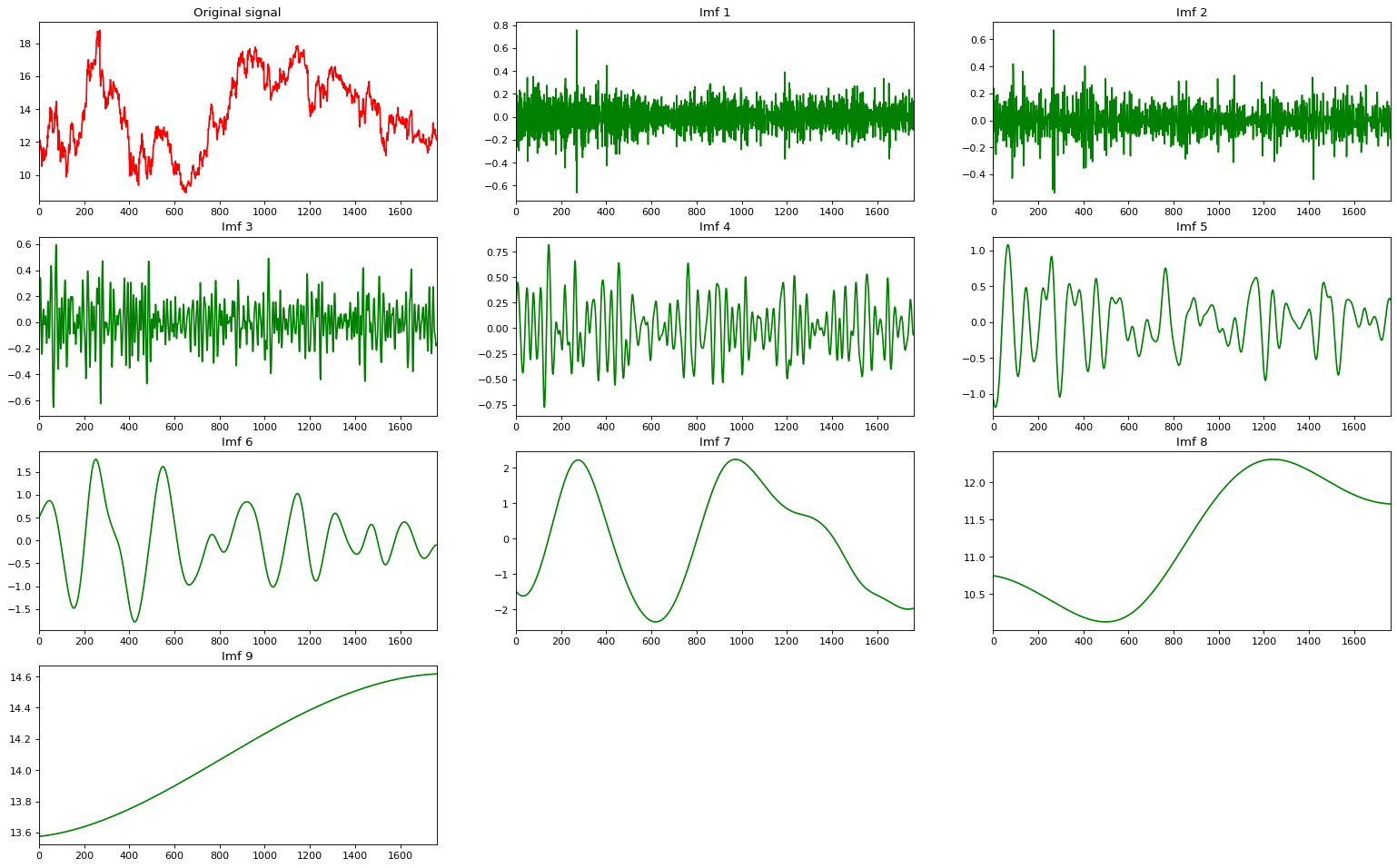
课题的目的是预测股票的价格，主要原理是应用LSTM（long short-term memory）算法，对于时间序列数据进行预测。

首先对于数据进行预处理：

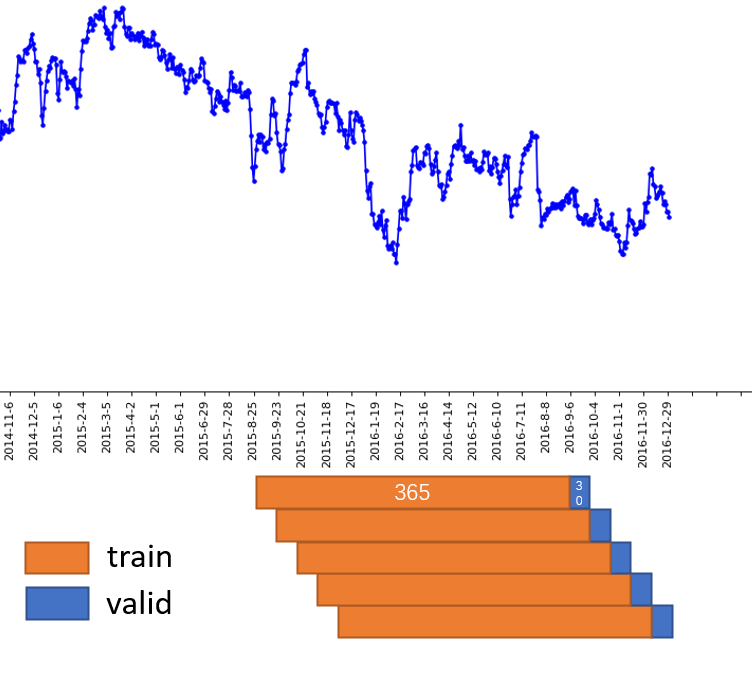
* 读取文件
* 在securities中提取出行业与股票代码的对应关系，即通过股票代码'Ticker Symbol'能够知道其行业sector
* 在fundamentals中，首先根据股票代码标注其行业，然后选择行业，我们选择了Consumer Discretionary，找到该行业中总资产最高的股票Ford，并从prices-split-adjusted中提取其数据
* 提取出来的数据有：日期和每日收盘价
* 做出Ford股价走势图（2010~2016）



* 由于信号比较混乱，我们对于收盘价格close信号进行EMD（经验模态分解，依据数据自身的时间尺度特征来进行信号分解，无需预先设定任何基函数，是一种时频域信号处理方式。EMD在处理非平稳及非线性数据上具有明显的优势，适合分析非线性非平稳的信号序列，具有较高的信噪比）以得到更加平稳的信号，我们得到有限个本征模函数IMF和一个残差Residue（原始信号等于所有IMF与残差Residue之和）
* 我们分解得到9个IMF信号如下：



* 由于股票价格数据是序列数据，不可以直接进行随机划分训练集验证集的k折交叉验证，因此我们采用滑动时间窗的划分方法：分成5块，训练集大小为365,验证集大小为30，分割方法如下图所示：



* 整合训练集和验证集，并找到对应的IMF函数和残差Residue

下面来进行LSTM预测，预测步骤如下：

* 对于每一个imf，他们都是股票价格的一个组成部分
* 我们首先计算后一天与前一天之差，将差值作为我们要预测的变量
* 我们将每天的差值下移一天并附在原数据左侧，这样每一个差值就有一个前一天的差值与之对应
* 例子：

原数据

|  |
| --- |
| 1 |
| 5 |
| 3 |
| 6 |

计算与前一天的差值

|  |
| --- |
| 4=5-1 |
| -2=3-5 |
| 3=6-3 |

将数据下移一格，NaN用0补上，并附在左侧

|  |  |
| --- | --- |
| 0 | 4 |
| 4 | -2 |
| -2 | 3 |

用左边的值（前一天的差值）预测右边的值（当天与前一天的差值）

* 接着划分训练集和测试集并且进行尺度缩放（-1到1之间）
* 将训练集放入LSTM模型中
* LSTM模型第一层LSTM网络，激活函数为默认双曲正切函数，神经元个数为neurons，stateful=True表示批次中每个样本的最后状态将用作下一批中样本的初始状态

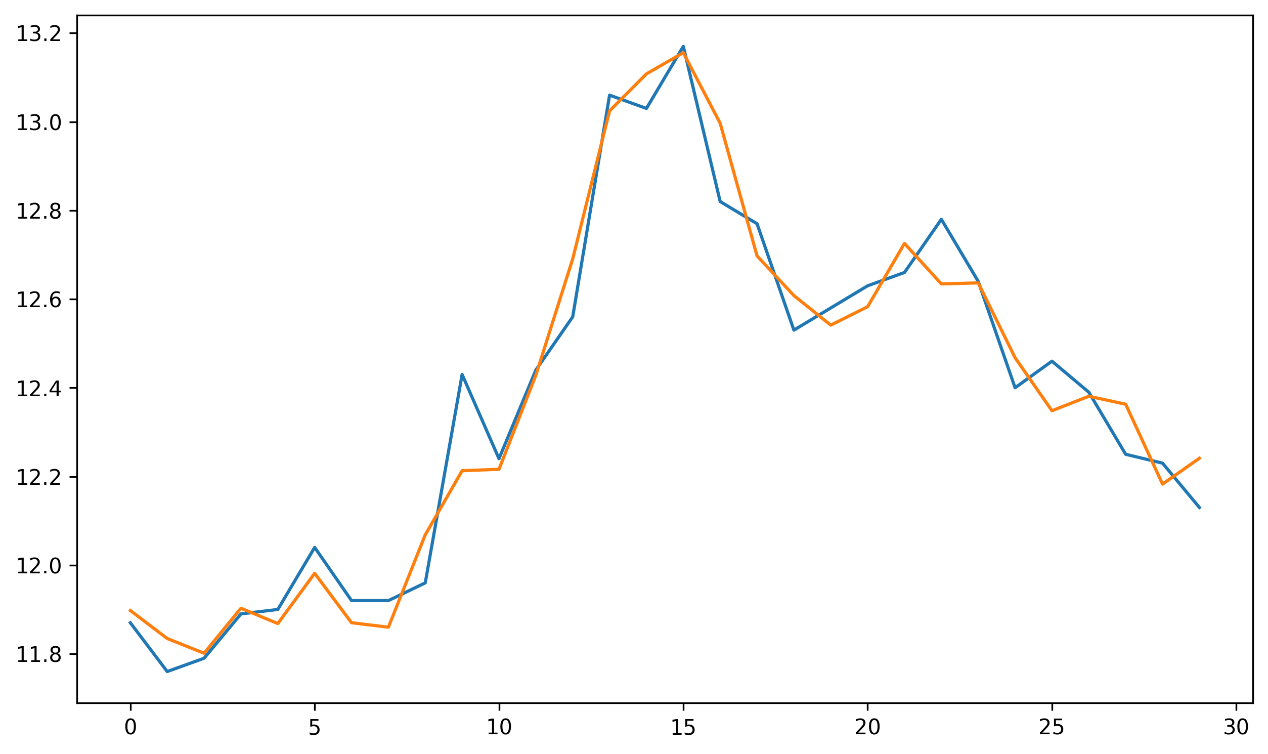
第二层密集层，节点个数node，激活函数默认线性激活函数

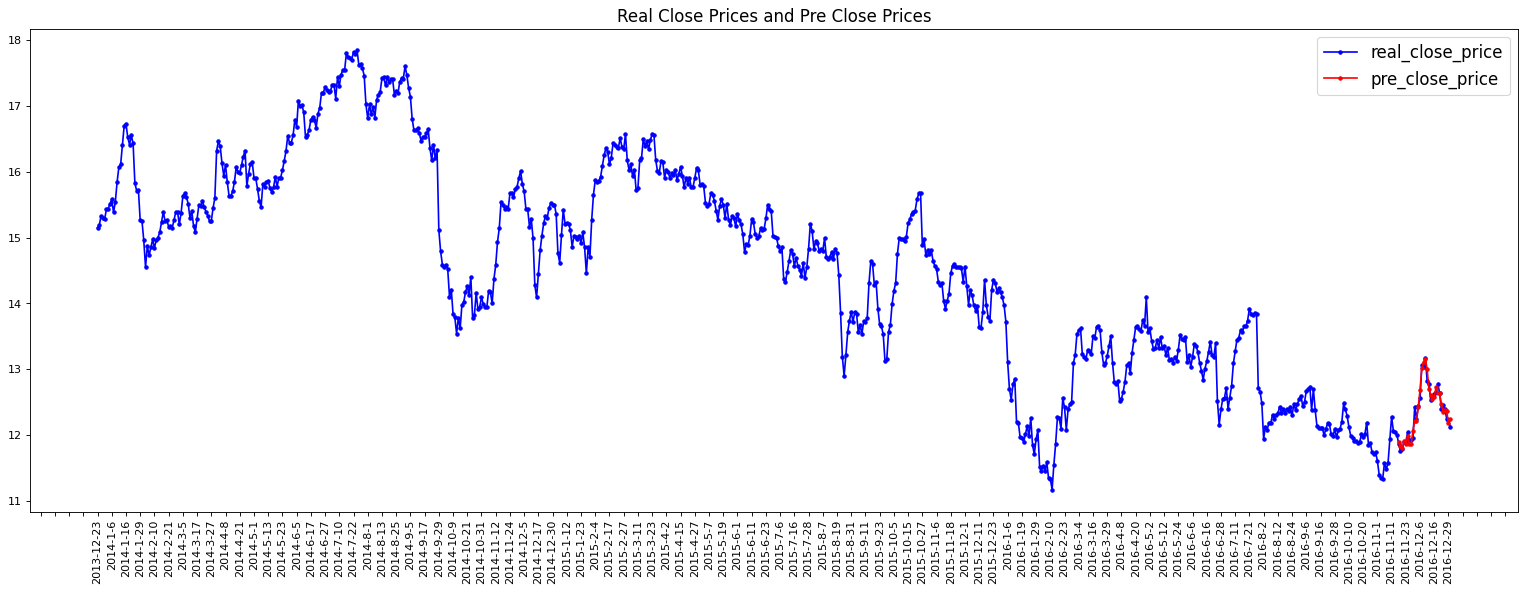
以mse为损失loss，adam为优化器进行优化

* 训练完成之后，对训练集中的变量进行预测，首先根据前一天的值当天的值，然后反缩放到原来的尺度，根据之前保存的量加回差值即为预测值
* 对于每个IMF预测完成后，全部相加再加上Residue即为每日收盘价的预测值
* 函数定义完成之后需要进行参数优化（调参 regler les hyperparametres），para\_lstm函数返回五个滑动时间窗验证后的mape平均值的相反数，因为采用贝叶斯优化器BayesianOptimization，该优化器寻找函数的最大值，由于我们的误差为正值且我们要寻找最小值，因此函数返回值设置为误差的相反数
* BayesianOptimization优化三个变量：网络训练次数epochs，LSTM层中的神经元数neurons，Dense层中的节点数nodes。
* 找到最优参数组合为{'target': -0.005956112811066365, 'params': {'epochs': 12.663114889828307, 'neurons': 21.259850686653415, 'node': 17.00723097290962}}

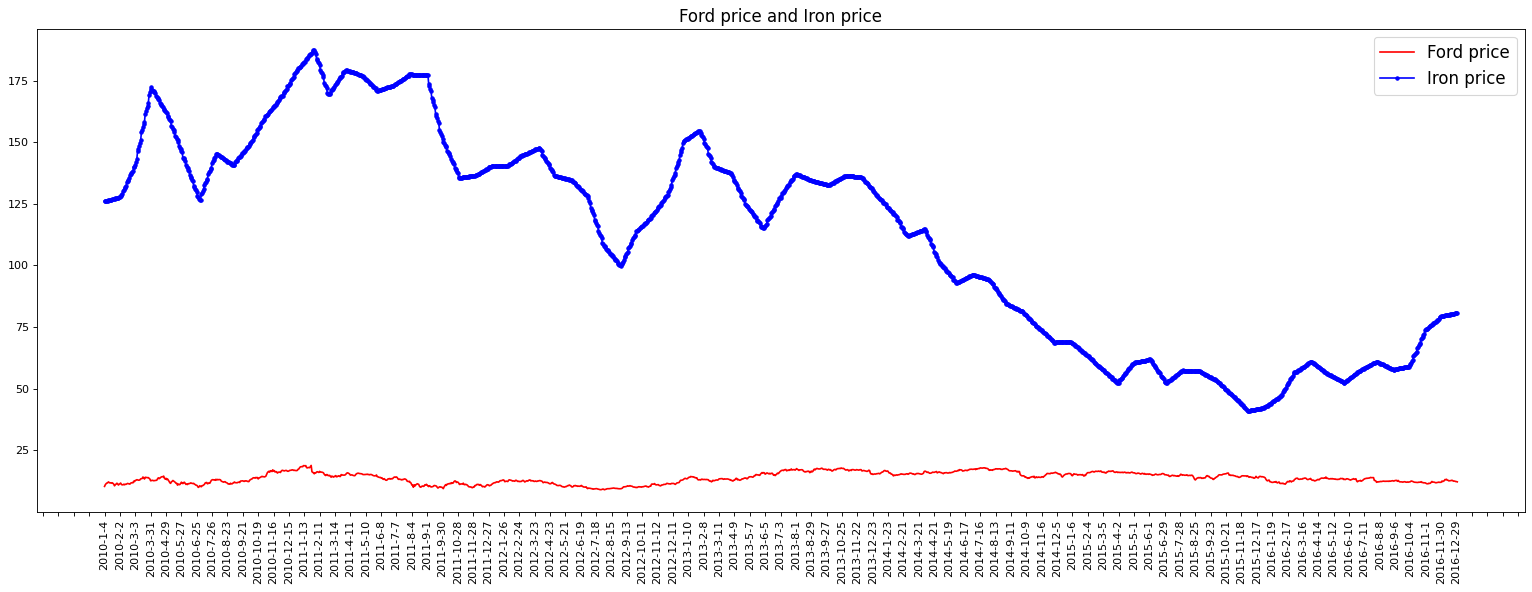
，取整为(13,21,17)

* 应用这组参数对于整个数据集进行训练，预测最后一个月的数据，得到误差为Test MAPE:0.546，即平均0.546%的误差，效果较好

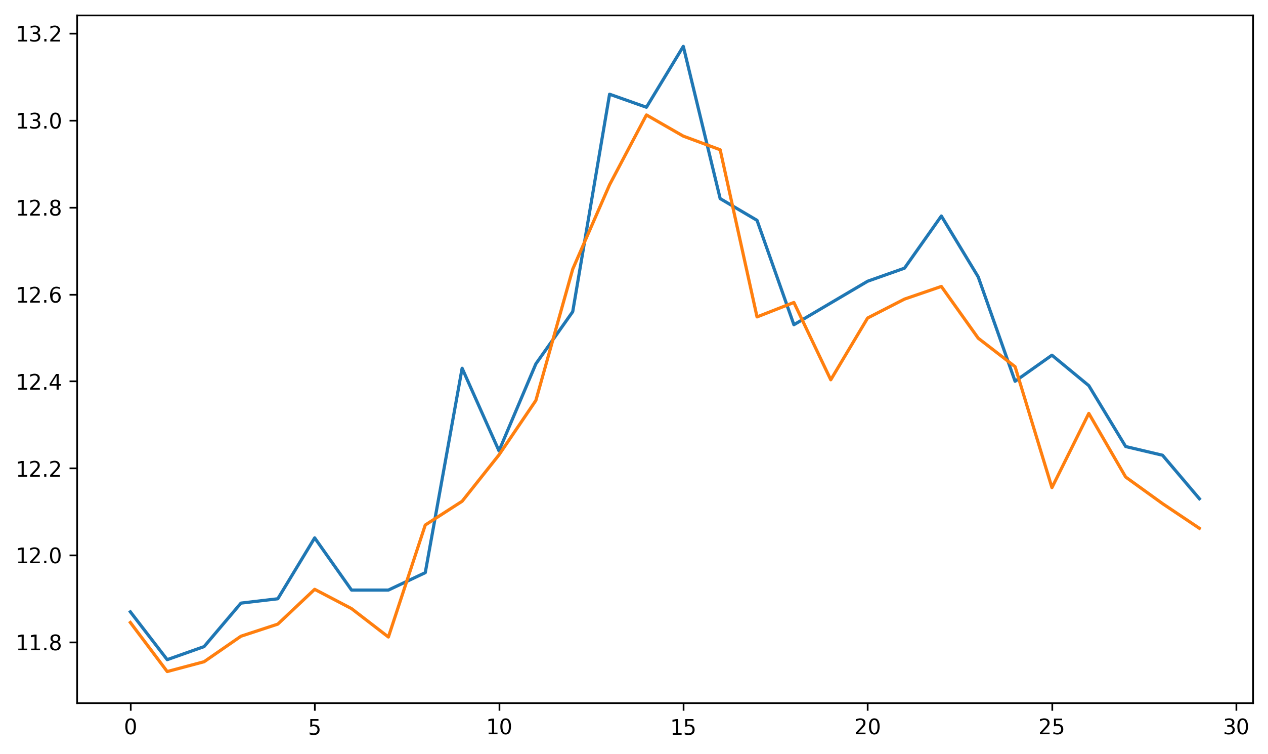


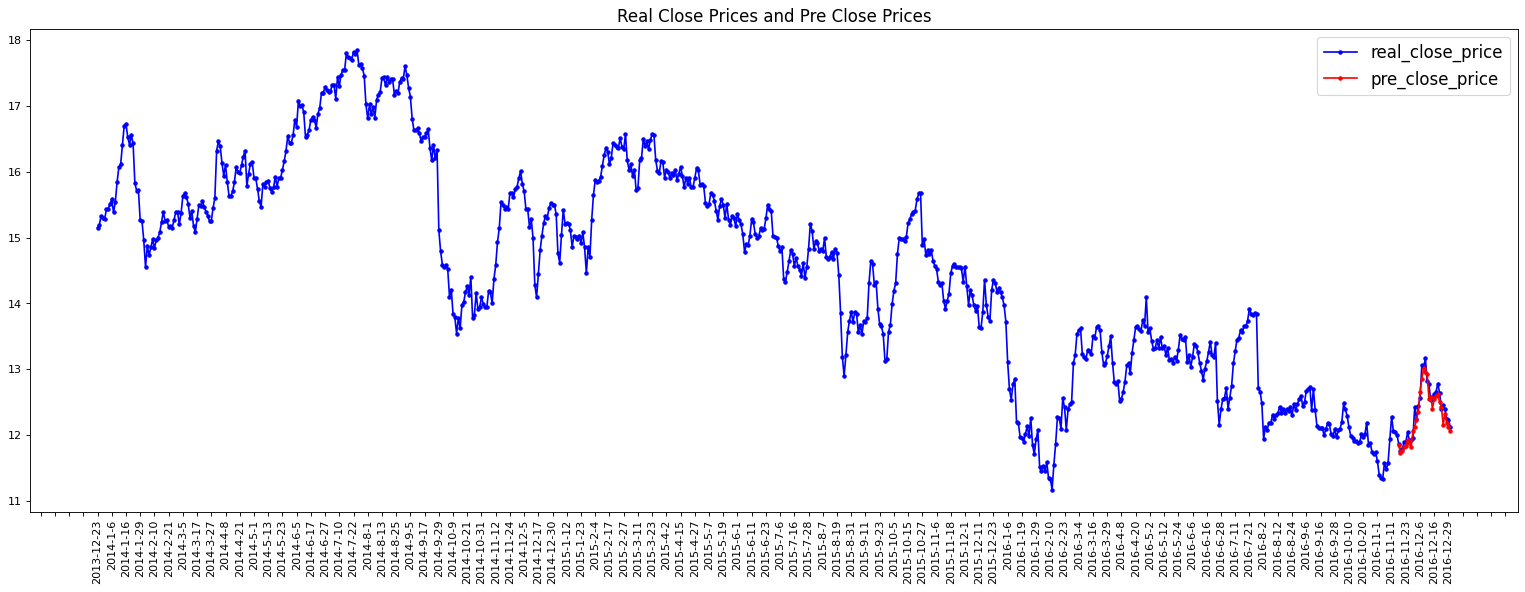


* 接下来考虑加入协变量辅助预测：
* 考虑到制造汽车的原材料主要为钢铁，选择铁矿石期货的价格为协变量，由于免费数据中，只有每个月月初的数据，因此对于每个月中间的数据进行了线性插值处理，即每个月的价格都呈一条直线
* 我们将铁矿石期货的价格与福特每日股价放到一起



* 预测过程与前面基本相同，唯一的区别在于由单元变量变为了多元变量，多了前一天铁矿石期货价格作为另一个变量，即现在通过前一天的股价和前一天的铁矿石期货价格预测当天的股价。
* 预测结果如下：
* 最优参数组合：{'target': -0.0070268598776048975, 'params': {'epochs': 16.636680071410716, 'neurons': 9.839458914511063, 'node': 14.805460587478368}}，取整为(17,10,15)
* 对于整个数据集进行训练，预测最后一个月的数据为Test MAPE:0.854
* 与真实值比较：





* 误差为0.854%，发现准确率不如直接用前一天的股价预测高，这可能是由于对于期货价格的处理采用了直接插值的方式，需要更加准确的数据，或者铁矿的期货价格与汽车的股价相关性并没有非常高。