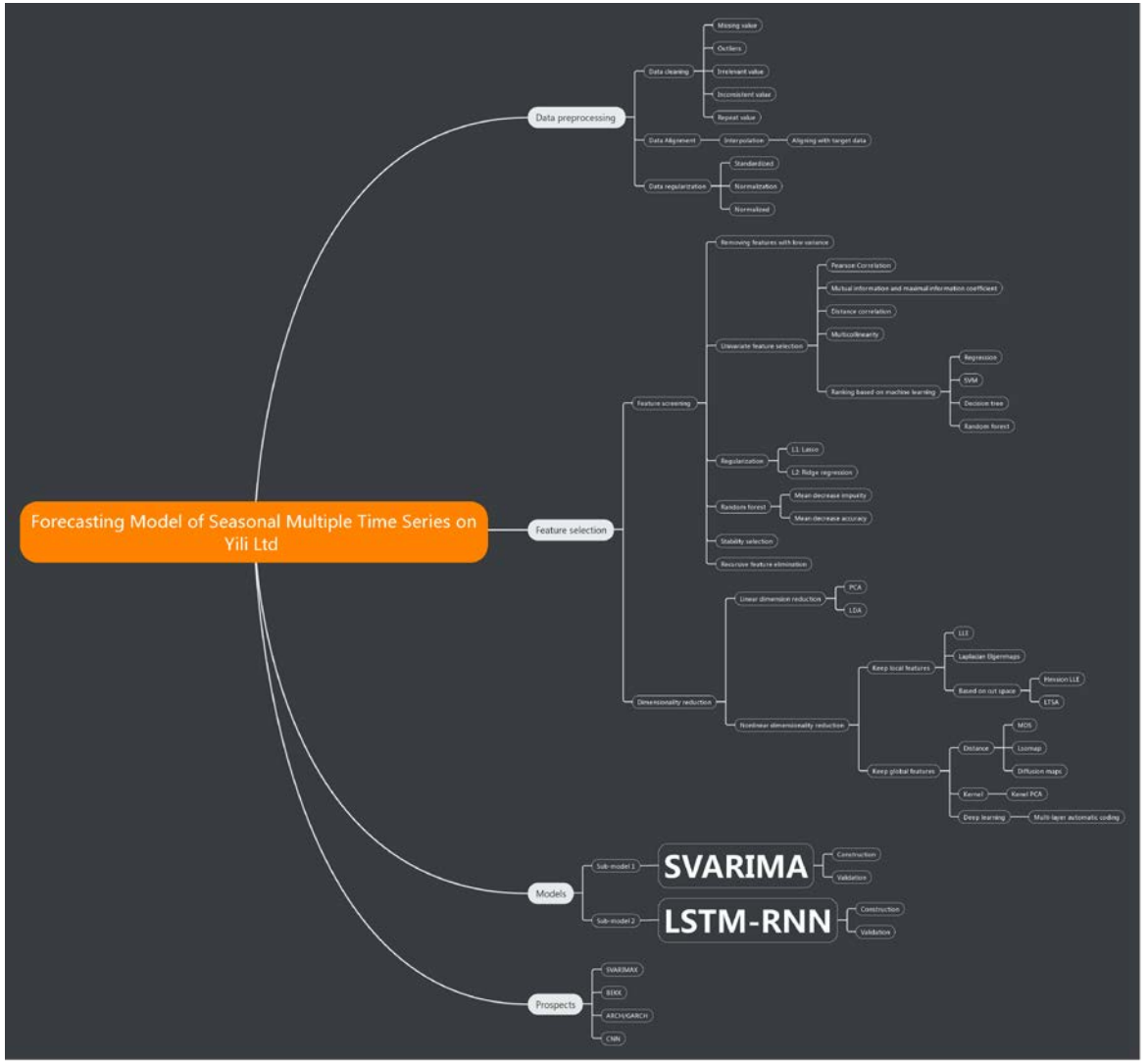
Forecasting Model of Seasonal Multiple Time Series on Yili Ltd

Presented by Wentao Zheng

May 27, 2018

CONTENTS

Mind Map


Forecasting Model of Seasonal Multiple Time Series on Yili Ltd
Data preprocessing
Data cleaning
Missing value
Outliers
Irrelevant value
Inconsistent value
Repeat value
Data Alignment
Interpolation
Aligning with target data
Data regularization
Standardized
Normalization
Normalized
Feature selection
Feature screening
Removing features with low variance
Univariate feature selection
Pearson Correlation
Mutual information and maximal information coefficient
Distance correlation
Multicollinearity
Ranking based on machine learning
Regression
SVM
Decision tree
Random forest
Regularization
L1: Lasso
L2: Ridge regression
Random forest
Mean decrease impurity
Mean decrease accuracy
Stability selection
Recursive feature elimination
Dimensionality reduction
Linear dimension reduction
PCA
LDA
Nonlinear dimensionality reduction
Keep local features
LLE
Laplacian Eigenmaps
Based on cut space
Hessian LLE
LTSA
Keep global features
Distance
MDS
Isomap
Diffusion maps
Kernel
Kernel PCA
Deep learning
Multi-layer automatic coding
Models
Sub-model 1
SVARIMA
Construction
Validation
Sub-model 2
LSTM-RNN
Construction
Validation
Prospects
SVARMAX
BEKK
ARCH/GARCH
CNN

整体概要

对目标进行分析，需要对伊利营业成本、营业收入、利润和股价多个因变量建立预测模型，此过程可以分为数据预处理，因子筛选与合成降维，模型搭建与展望四个部分。考虑到前三个变量均为季度频，而股价为日频，因而预测模型初步可以分为两个档次。

首先是对数据的预处理，预处理可以分为三个阶段：1、数据清洗：利用python进行对原始数据异常值、缺失值、非相关值、重复值的清洗，保证数据的稳定性；2、数据对齐：考虑到所提供数据中基本均为不同发布频率的原始数据，为了保证数据的有效性和合理性，这里采用了近邻值的方法、分别以两个档次因变量数据为参考序列实现了数据的对齐。3、数据转换：考虑到数据之间数量级有相当大的差异，为有效的构建预测模型，在分别对归一化、规范化和z分位数标准化尝试进行数值变换的过程之后，发现z分位数标准化的预测效果更显著，因而采用z分位数标准化的方式进行数值转换。

其次是对因子的筛选与合成降维：因子筛选分别采用了筛除低波动，单变量筛选，Lasso正则化、随机Lasso和贪心算法等方式进行因子筛选，同时思维导图内刻画了多种因子合成降维的方式，但由于时间有限，尚未完成这一部分，后续可补充。

整体概要

然后是分别对两个档次进行预测模型的搭建：

一、考虑对营业成本、营业收入、利润建立预测模型：结合对实际数据的初步判断，发现伊利的营业成本、营业收入、利润具有很强的季节效应，拟建立具有季节效应的时间序列模型。前期，尝试利用深度学习神经网络建立预测模型，发现预测结果并不理想，分析原因很有可能是由于季节效应和过小的样本量所致，降低深度学习神经网络用来构造预测模型的优先级；同时由于数据具有强的前后时间效应，故暂时不考虑忽略时间效应的多元回归模型，因而对这个档次建立预测模型时，优先考虑自回归移动平均时间序列模型，同时考虑到不同时间序列的相关性和潜在有效信息，综合考虑为利用季节序列单整向量自回归移动平均时间序列sVARIMA建立预测模型。（在这里考虑到，）

二、考虑对伊利股价建立预测模型：结合对实际数据的初步判断，伊利的股价序列为日频，同时观察到伊利股价5000+的样本量，考虑到机器学习在大样本量学习优化建立预测模型的优势，同时考虑到深度学习balbalbalbala，这里拟采用长短期记忆神经网络 RNN-LST来建立预测模型。

最后是对模型的展望，上述模型存在如下可以拓展的地方：在构建时间序列模型当中，可以考虑加入到其它一些外生变量作为因子，构造SVARIMAX时间序列预测模型，同时后续可以考虑运用改进的CNN神经网络的方式建立预测模型等。

Note

数据说明

- 为方便表述，后文将因子用数值标签代替，参考右侧表格
- 考虑到营业收入、营业成本、利润三个变量均为季度频，而股价为日频，因而本文预测模型分为两个档次，季度频为Plate1，日频为Plate2

Factor	因子名称
1	进口数量:原奶(0401):当月值
2	新西兰:原奶产量
3	国际现货价:玉米
4	国际现货价:豆粕
5	玉米:产量
6	M2
7	预测平均值:CPI:当月同比
8	城镇居民人均消费性支出:累计同比
9	伊利股份:产量:液体乳
10	平均到岸价:苜蓿草:当月值
11	美国:牛:市场存栏量
12	进口数量:奶粉:当月值
13	进口金额:奶粉:当月值
14	存栏:奶牛:全国
15	产量:奶类:全国
16	产量:奶粉:中国
17	产量:牛奶:全国
18	人口数:全国
19	进口数量:种牛冻精
20	进口数量:苜蓿草
21	零售价:牛奶

Data preprocessing

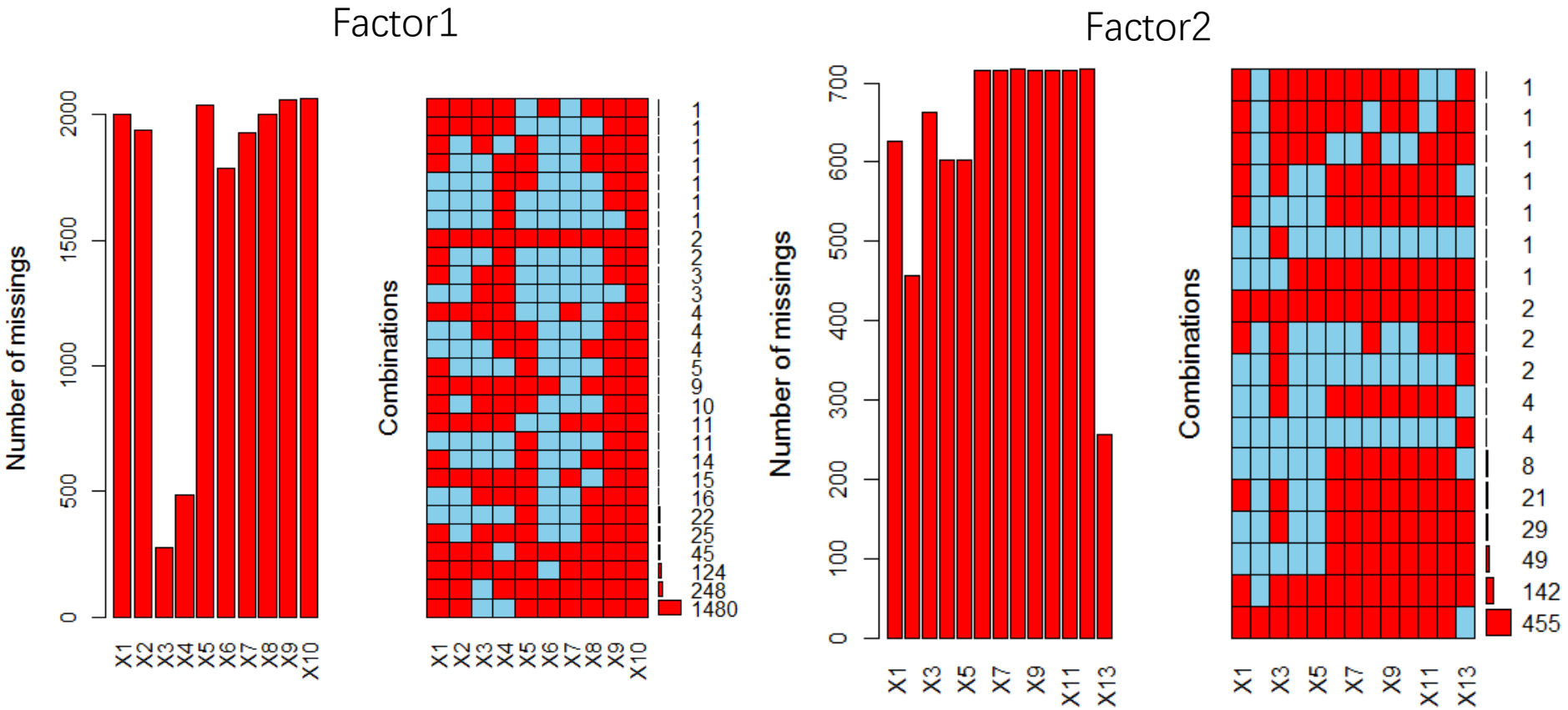
数据预处理

- 数据清洗
- 数据对齐
- 数据标准化

Data cleaning

数据清洗

- 缺失值
- 异常值
- 非相关数据
- 不一致数据
- 重复数据



Data alignment

数据清洗

- 插值法
- 移动平均
- 均值回归
- 近邻值(采纳)

Data regularization

数据归整

- 规范化
- 归一化
- 标准化（采纳）

Result-Pred

规范化

Plate 1

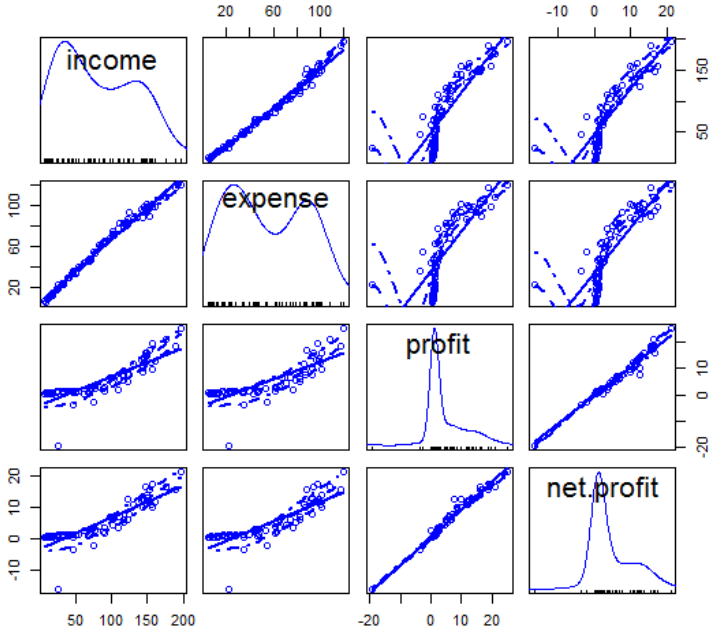
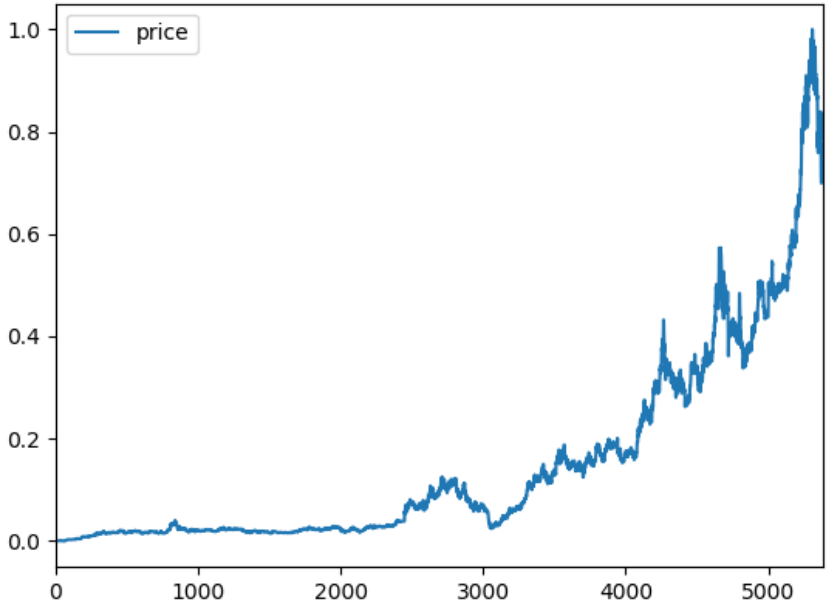


Plate 2



Result-Factor

规范化

Plate 1

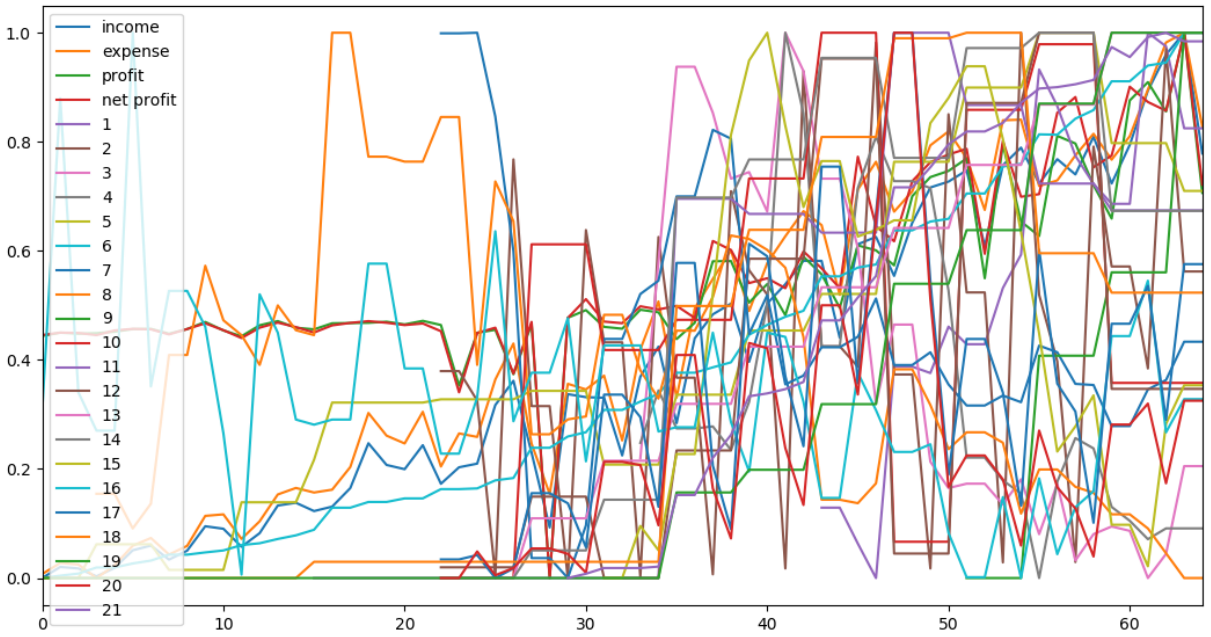
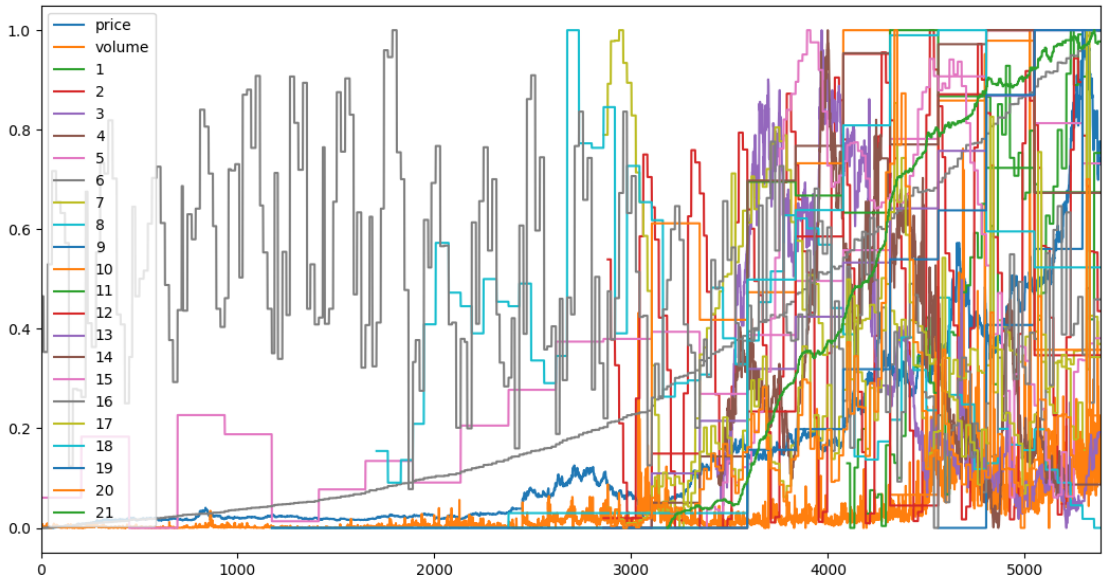


Plate 2



Result-Factor

标准化

Plate 1

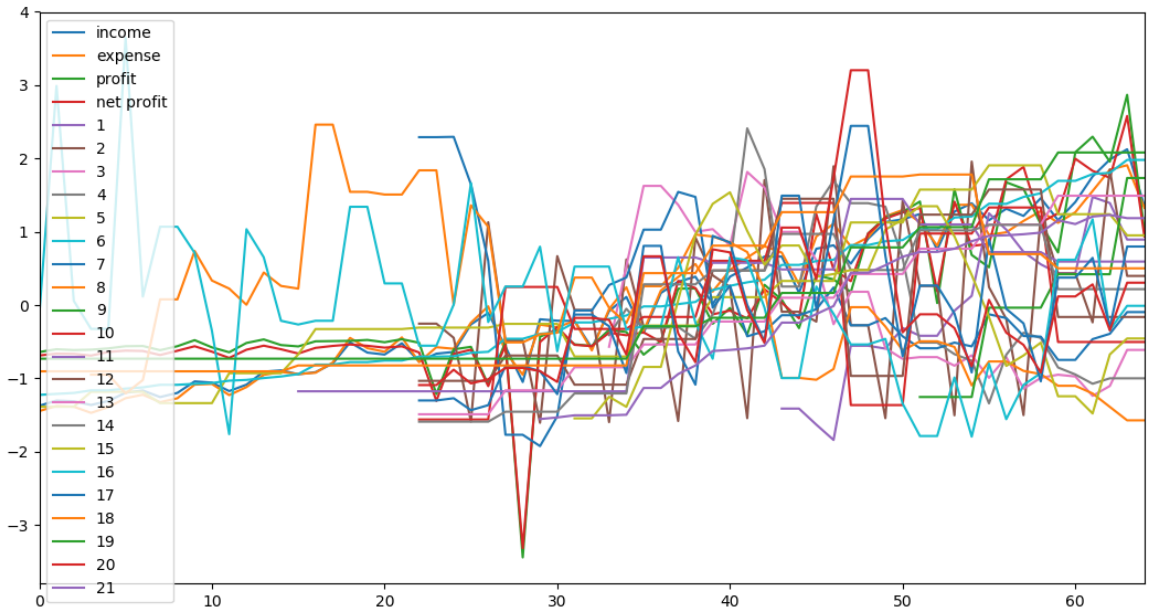
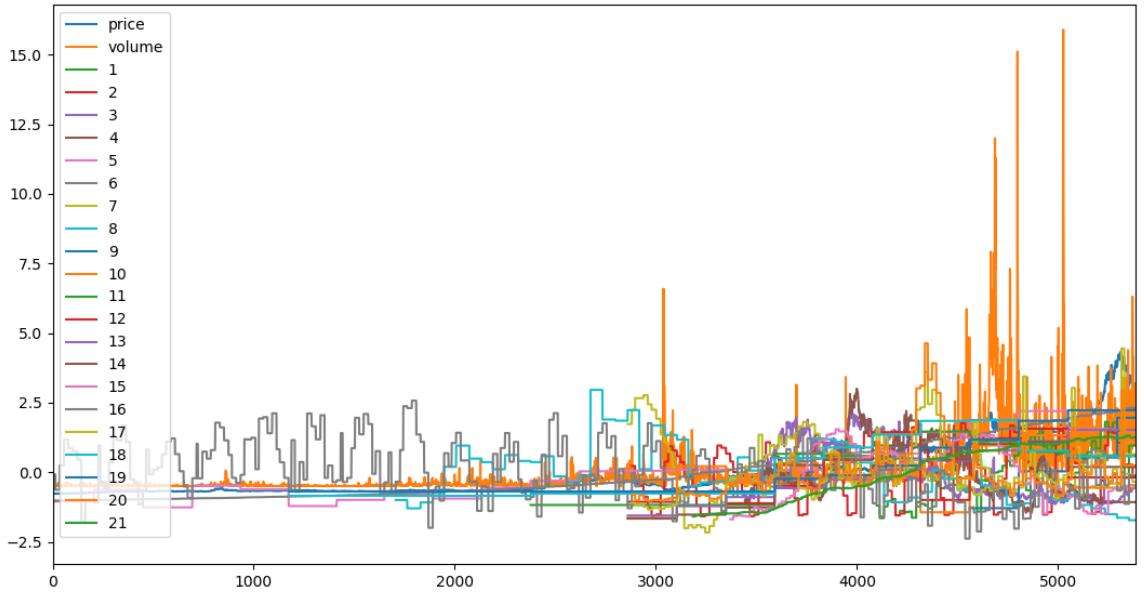


Plate 2



Feature selection

特征选择

- 特征筛选
- 降维

Feature screening

特征筛选

- 筛去取值变化幅度小的特征因子
- 单变量特征筛选
- 正则化特征筛选
- 随机森林特征筛选Random forest
- 稳定性选择
- 递归特征消除

Removing features with low variance

	X1	X2
n	5.385000e+03	5.385000e+03
mean	-4.081573e-18	7.730223e-18
stdev	4.887288e-01	6.818328e-01
skw	1.118778e-01	-1.981439e-01
kurtosis	4.511937e+00	9.312519e-01

	X3	X4
n	5.385000e+03	5.385000e+03
mean	5.833738e-16	-1.796795e-14
stdev	6.021260e-01	6.004270e-01
skw	9.873633e-01	1.246849e+00
kurtosis	1.953098e+00	4.226470e+00

	X5	X6
n	5.385000e+03	5.385000e+03
mean	-2.636030e-16	-1.379810e-14
stdev	1.000000e+00	1.000000e+00
skw	7.811962e-01	8.383939e-01
kurtosis	-4.901558e-01	-6.370003e-01

	X7	X8
n	5.385000e+03	5.385000e+03
mean	3.981936e-15	5.037279e-14
stdev	6.846871e-01	8.265199e-01
skw	9.025773e-01	7.393425e-01
kurtosis	4.961200e+00	1.434400e+00

	X9	X10
n	5.385000e+03	5.385000e+03
mean	-2.265266e-14	-3.711597e-13
stdev	3.909736e-01	6.846871e-01
skw	5.947631e-01	-2.381365e-01
kurtosis	1.310610e+01	1.012726e+00

	X11	X12
n	5.385000e+03	5.385000e+03
mean	1.649247e-14	1.605372e-13
stdev	7.473330e-01	6.846871e-01
skw	-2.531693e-01	6.452078e-01
kurtosis	-6.030925e-01	4.222801e-01

	X13	X14
n	5.385000e+03	5.385000e+03
mean	-2.990052e-15	-3.992329e-14
stdev	6.846871e-01	6.846871e-01
skw	3.866158e-02	-9.628696e-01
kurtosis	7.624604e-01	9.362128e-01

	X15	X16
n	5.385000e+03	5.385000e+03
mean	-1.868327e-15	-5.413201e-16
stdev	6.116137e-01	1.000000e+00
skw	-2.666069e-01	2.634209e-01
kurtosis	1.235929e+00	-4.601947e-01

	X17	X18
n	5.385000e+03	5.385000e+03
mean	9.126395e-17	1.598565e-17
stdev	6.818328e-01	8.842774e-01
skw	2.505465e+00	7.563217e-01
kurtosis	1.164121e+01	-5.883540e-01

	X19	X20
n	5.385000e+03	5.385000e+03
mean	-2.497236e-15	2.211343e-17
stdev	8.842774e-01	6.818328e-01
skw	1.343265e+00	3.316559e+00
kurtosis	7.210068e-01	1.571696e+01

	X21
n	5.385000e+03
mean	4.122372e-17
stdev	6.392325e-01
skw	-3.715211e-01
kurtosis	7.473813e-01

左图为对21个因子的描述性统计数据，这里在于筛选出取值变化较小的特征因子，由于数据标准化之后可以对数据进行横向比较，发现X9的取值变化较小，因而初步筛除出因子9.

Univariate feature selection

单变量筛选

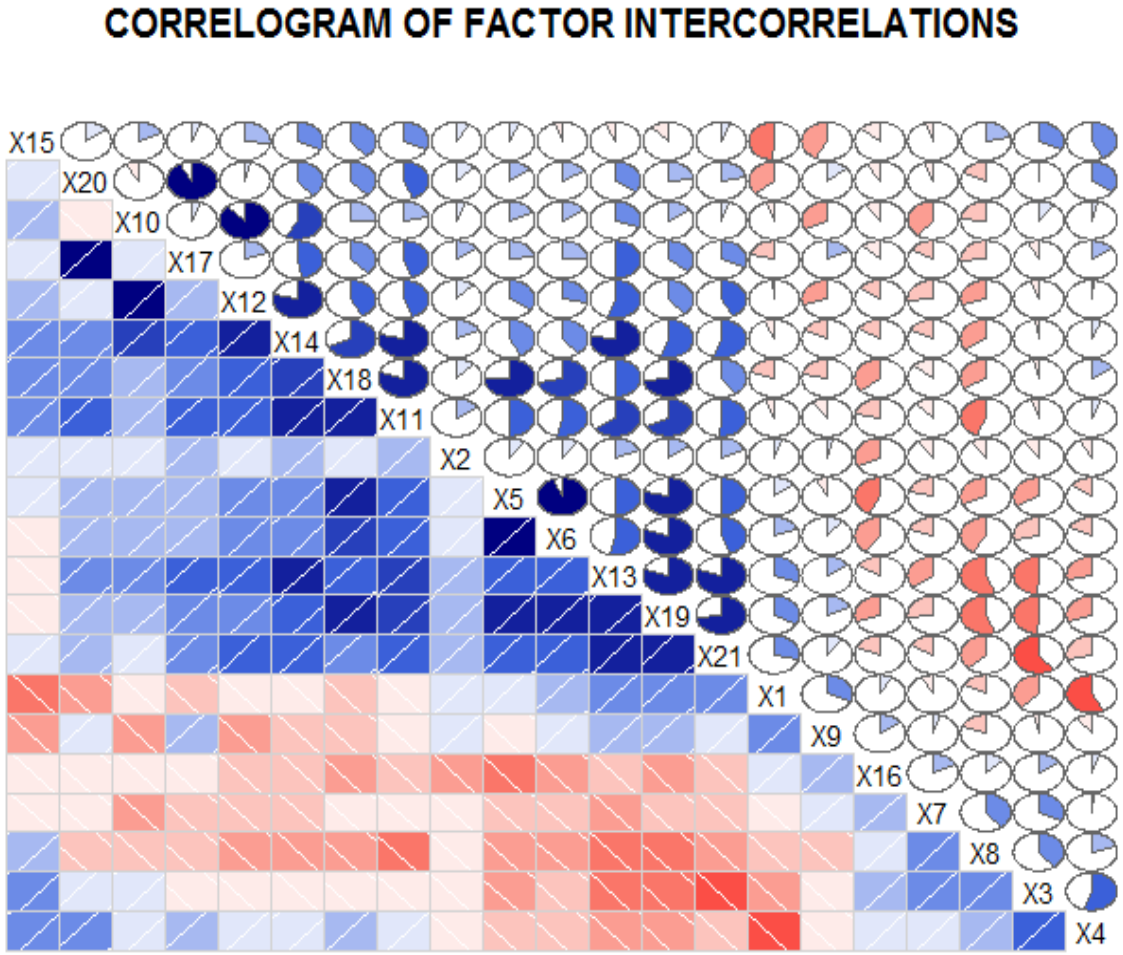
- Pearson相关系数（最直观，检验线性相关）
- 互信息和最大信息系数 $I(X,Y)=\sum_{y\in Y}\sum_{x\in X}p(x,y)\log\left(\frac{p(x,y)}{p(x)\,p(y)}\right)$
- 距离相关系数（检验非线性相关）
- 多重共线性（修正Frish逐步回归）
- 基于机器学习模型的特征排序（决策树、随机森林，检验非线性相关）

Univariate feature selection

单变量筛选

通过如上页PPT五种方法的综合判别，

根据综合结果初步筛除出因子5，因子12，因子18，因子19，因子20，因子21



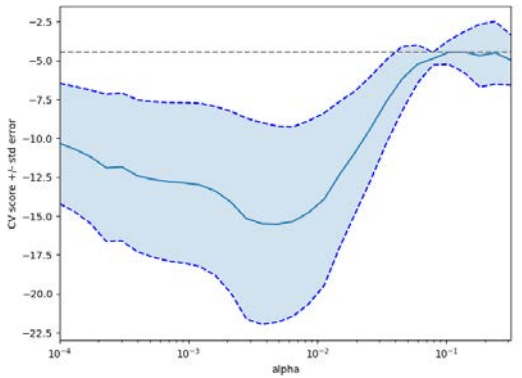
Regularization

- L1正则化：Lasso
- L2正则化：Rige

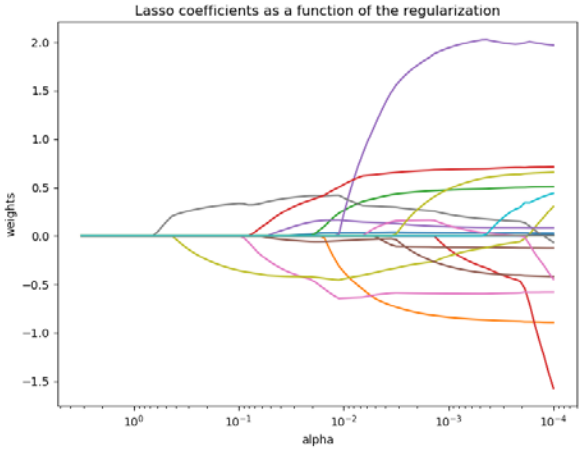
优先采用LASSO的方法进行变量筛选，这里以price作为因变量用于演示，通过计算出权重向量图P1，得到最优权重为8.4*10^-1，对应图2。

最终，综合多个结果，筛除出权重接近于0的因子：因子6，因子12，因子16，因子19

P1



P2



Random forest

- 以平均减少不纯度作为特征选择值
- 以平均精确度作为特征选择值

这里采取随机森林机器学习算法，以平均减少不纯度作为特征选择值，以价格序列作为预测序列，最终得到21个因子各因子打分值如下：
Features sorted by their rank:
[(0.8772, 5), (0.059, 1), (0.0254, 20), (0.0075, 2), (0.0067, 3), (0.0056, 14), (0.0055, 16), (0.0051, 6), (0.0032, 15), (0.0028, 19), (0.0012, 4), (0.0007, 7), (0.0001, 18), (0.0, 17), (0.0, 13), (0.0, 12), (0.0, 11), (0.0, 10), (0.0, 9), (0.0, 8)]
Note:其中数对第一个值表示因子打分值，第二个表示因子序数
这里拟初步筛除因子8，因子9，因子10，因子11，因子12，因子13，因子17因子打分值较低的序列

Advanced feature screening algorithm

- Stability selection

采用稳定性选择的方法，利用随机lasso和随机逻辑回归实现，结果如下：
Features sorted by their score:
[(1.0, 15), (1.0, 14), (1.0, 8), (1.0, 7), (1.0, 5), (1.0, 4), (1.0, 3), (1.0, 2), (1.0, 1), (0.99, 6), (0.985, 16), (0.715, 20), (0.685, 19), (0.12, 11), (0.1, 17), (0.095, 9), (0.05, 10), (0.035, 18), (0.035, 13), (0.02, 12)]
Note:其中数对第一个值表示因子打分值，第二个表示因子序数
根据结果拟筛除出因子12

- Recursive feature

利用寻找最优特征子集的贪心算法将递归特征消除，得到因子打分结果如下：
Features sorted by their score:
[(1, 13), (2, 9), (3, 18), (4, 17), (5, 12), (6, 10), (7, 11), (8, 4), (9, 8), (10, 5), (11, 2), (12, 19), (13, 16), (14, 14), (15, 7), (16, 3), (17, 20), (18, 6), (19, 15), (20, 1)]
Note:其中数对第一个值表示因子打分值，第二个表示因子序数
根据结果拟筛除出因子1和因子19

Models

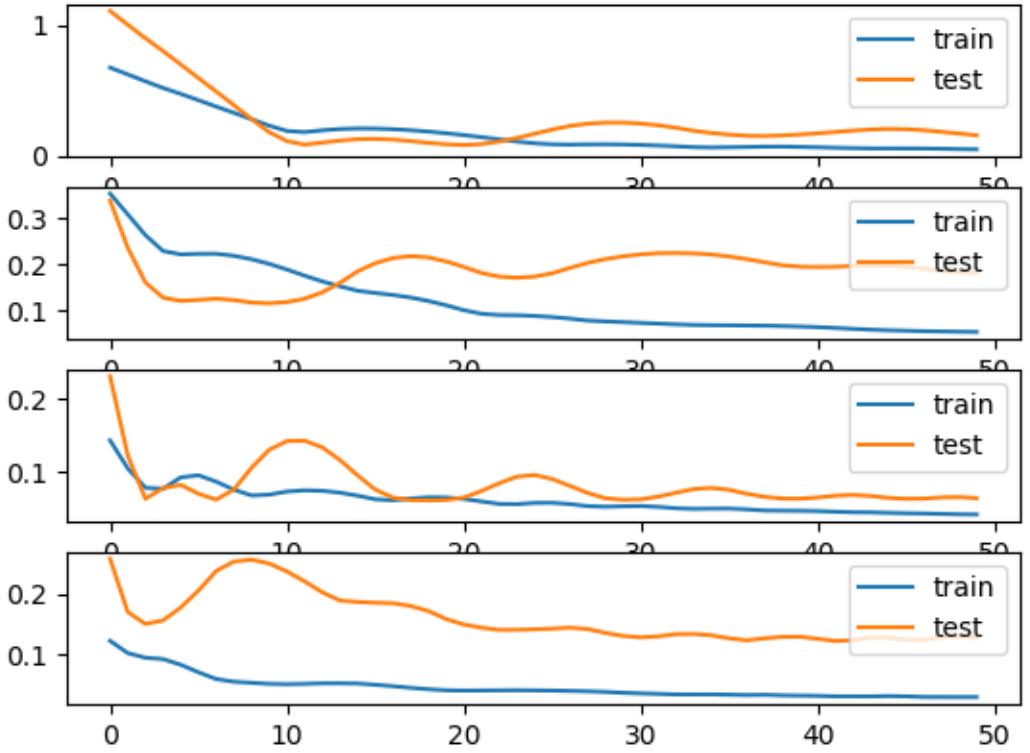
- 针对Plate 1：子模型一-季节序列单整向量自回归移动平均时间序列预测模型 SVARIMA
- 针对Plate 2：子模型二-长短期记忆神经网络 RNN-LST

Note:
Plate1指对季度频数据（营业收入、营业成本、利润）建立预测模型
Plate2指对日频数据（股价）建立预测模型

Sub-model 1 SVARIMA

Plate 1

- 首先初步尝试利用LSTM神经网络建立预测模型，但发现结果并不理想（如右图），考虑可能原因在于Plate1营业收入等季度频样本量过小以及存在季节效应，故转换为之后的时间序列模型建模



Sub-model 1 SVARIMA

Plate 1

结合对实际数据的初步判断，发现伊利的营业成本、营业收入、利润具有很强的季节效应，分别建立具有季节效应的时间序列模型。

$$(1-B)(1-B^S)z_t = (I_k - \theta B)(I_k - \Theta B^S)a_t$$

同时由于样本量大小的限制（65），为保证矩阵的可逆性，建立该模型只能选取单因子和预测变量构建sVARIMA模型，这里分别采用前文优选的因子和预测因子配对构建预测模型。

下文数据为以净利润和因子2构造的模型结果作为展示，在经历过模型阶数的判别之后，建立参数为VARIMA(1,1,2)-Sorder(1,0,1)的SVARIMA预测模型

Sub-model 1 SVARIMA

Coefficient(s):

	Estimate	Std. Error	t value	Pr(>|t|)	
[1,]	-0.72962	0.07587	-9.617	< 2e-16	***
[2,]	-1.00529	0.07245	-13.875	< 2e-16	***
[3,]	0.14720	0.07458	1.974	0.04840	*
[4,]	-0.96169	NA	NA	NA	
[5,]	0.75202	NA	NA	NA	
[6,]	0.28641	NA	NA	NA	
[7,]	0.17888	NA	NA	NA	
[8,]	0.87654	NA	NA	NA	
[9,]	-0.33281	NA	NA	NA	
[10,]	-0.73859	0.11686	-6.320	2.61e-10	***
[11,]	0.77340	0.49930	1.549	0.12139	
[12,]	0.34230	0.29916	1.144	0.25254	
[13,]	0.17209	NA	NA	NA	
[14,]	-0.39980	0.14730	-2.714	0.00664	**
[15,]	-0.16656	NA	NA	NA	
[16,]	0.61170	0.09649	6.340	2.30e-10	***
[17,]	0.50530	NA	NA	NA	
[18,]	0.31484	NA	NA	NA	
[19,]	-0.04505	NA	NA	NA	
[20,]	0.87193	NA	NA	NA	

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Estimates in matrix form:

Regular AR coefficient matrix

AR(1)-matrix

	[,1]	[,2]
[1,]	-0.730	-1.005
[2,]	0.147	-0.962

Seasonal AR coefficient matrix

AR(4)-matrix

	[,1]	[,2]
[1,]	0.752	0.286
[2,]	0.179	0.877

Regular MA coefficient matrix

MA(1)-matrix

	[,1]	[,2]
[1,]	-0.333	-0.739
[2,]	0.172	-0.400

MA(2)-matrix

	[,1]	[,2]
[1,]	0.773	0.342
[2,]	-0.167	0.612

Seasonal MA coefficient matrix

MA(4)-matrix

	[,1]	[,2]
[1,]	0.50530332	0.3148403
[2,]	-0.04504734	0.8719325

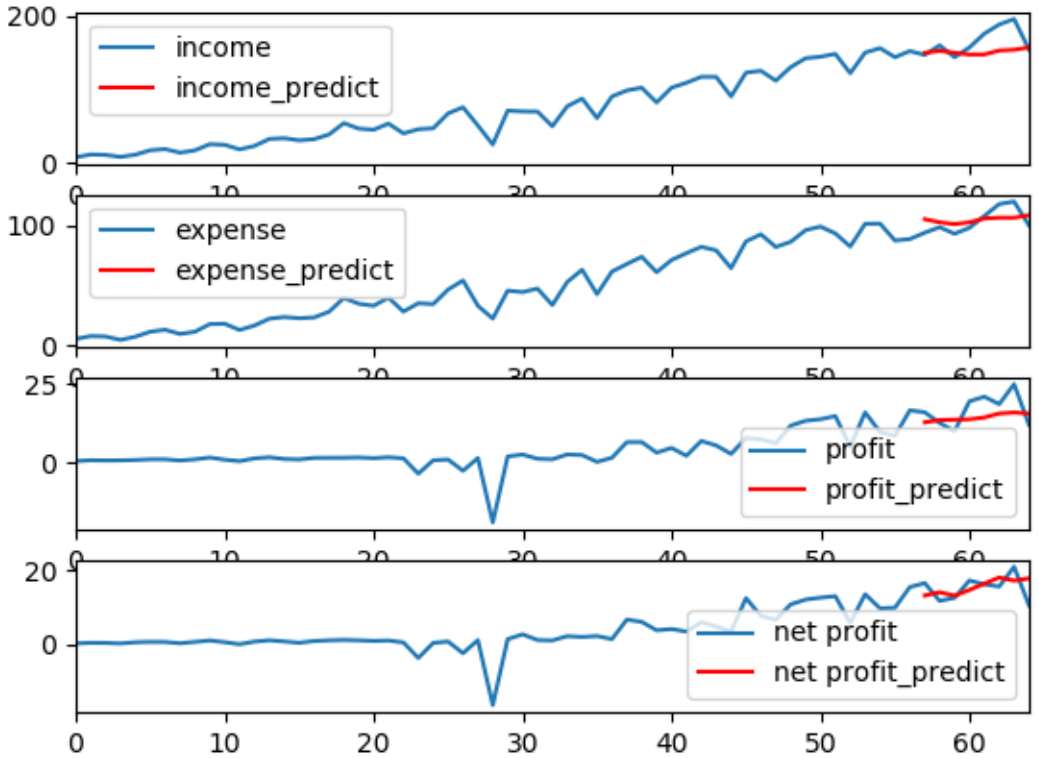
Residuals cov-matrix:

	resi	resi
resi	0.03963633	0.03864189
resi	0.03864189	0.04243695

---
aic= -7.8966
bic= -7.1923

Sub-model 1 SVARIMA

Plate 1-Results



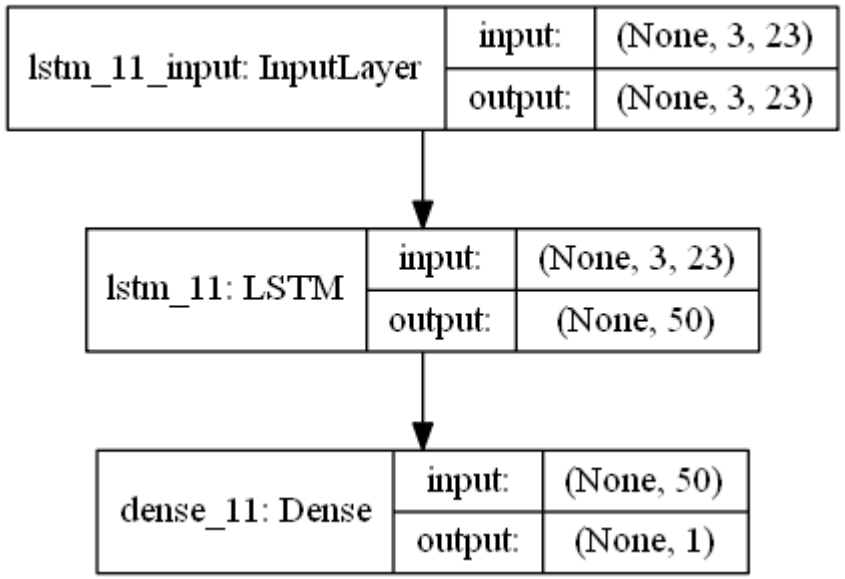
Sub-model 2 LSTM-RNN

Plate 2

- 首先将高维时间序列（5382*23）转换为可供
- 于监督学习的数据集（5382*92）

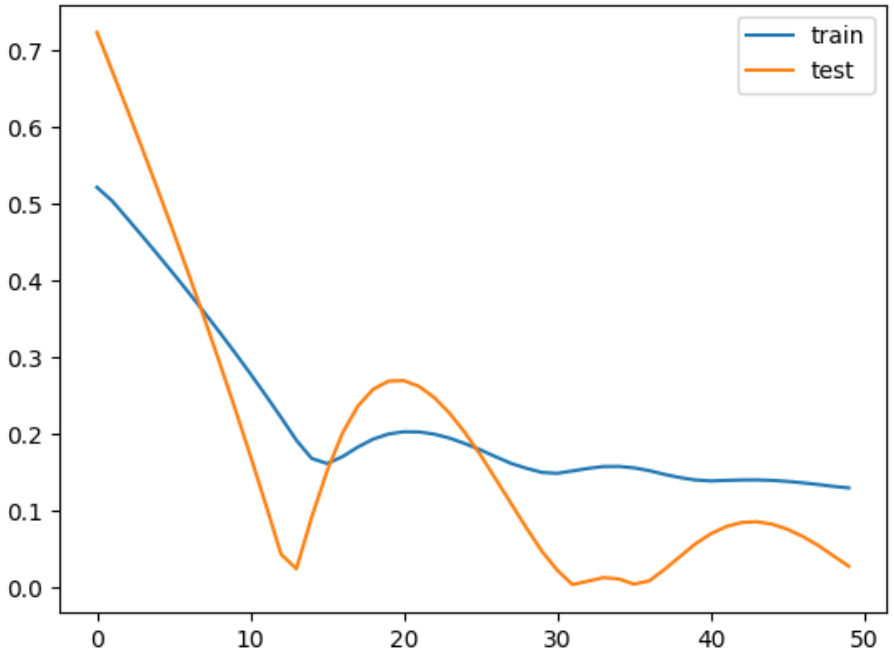
  其次将数据集分为训练数据和测试数据集，
- 用于预测（这里采用预测区间为3个月92天）

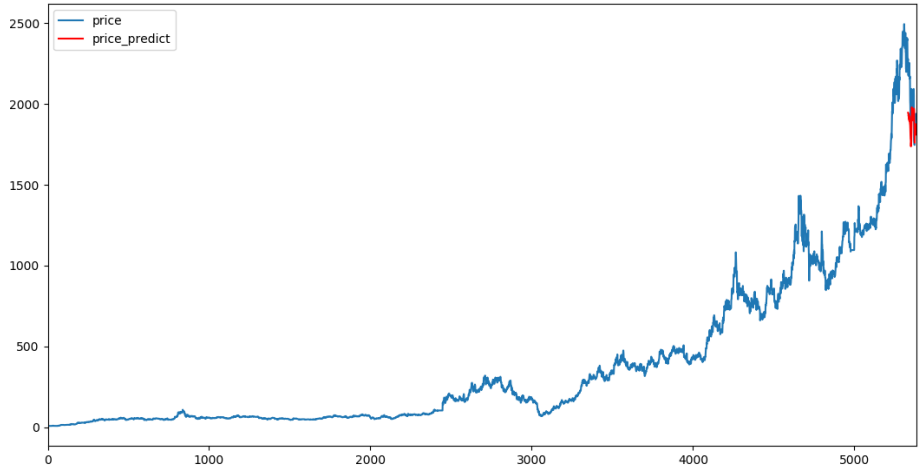
  构造拟合模型：这里选取MAE作为损失函数和梯度下降的参数，构造50个神经元的隐藏层和1个神经元的输出层，设置每个神经元处理大小为72，滞后期调整为3

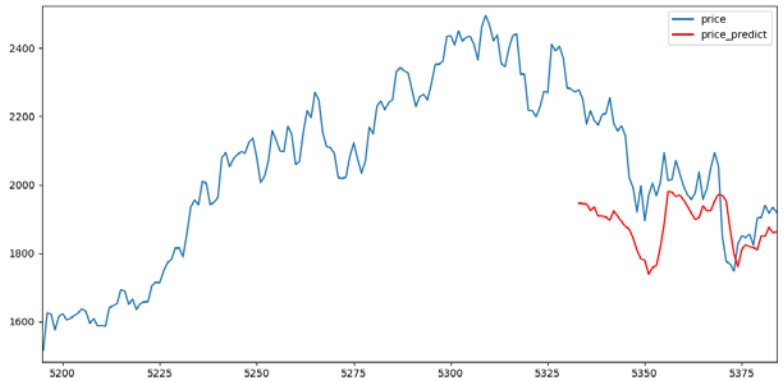


Sub-model 2 LSTM-RNN

结果







Prospects

Advantages

- 综合考虑到了数据的特殊性，分别通过时间序列和深度学习的方法建立预测模型
- 综合利用多种方法进行因子的筛选与合成

Disadvantages

- 后期可以考虑加入合理的外生变量X的方法来建立预测模型sVARIMAX同时也可以考虑到BEKK、GARCH、ARCH建立多元波动率模型；后期可以考虑CNN等深度神经网络的改进来建立预测模型.
- 后续可自行研究添加可供参考因子
- 由于时间有限，因子合成PCA等方式尚未完成，后续可补充.

THANK YOU