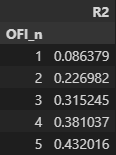
## 检验ofi因子和deltap的相关性

分每个股票每天的数据计算，按照原文做法，将每天的交易时间按照30min的时间区间分成8个大时间段，在每个大时间段内按照10s的小时间段计算OFI和delta\_p，得到每个大时间段内的OFI序列和delta\_p序列，并在大时间段内进行OLS回归得到这个大时间段内的OFI和delta\_p的回归R平方，再将8个大时间段内的R平方进行平均得到这一天的R平方。

接下来将每只股票每天的R平方数据进行平均得到多档OFI在样本内的解释力：



可见多档OFI拥有更好的解释力

## 网络搭建

* 1. **Input数据处理**

先按照原文的数据处理方式进行网络搭建，input数据的处理方式为：

**Input\_data:** sample\_shape= k\*14,其中k是时间窗口长度，取40-160分别测试，14是特征数量，此处取特征为：成交量、成交金额、成交量一阶差分、成交金额一阶差分、买一价一阶差分、卖一价一阶差分、买一量一阶差分、卖一量一阶差分、以及time\_window秒为窗口计算得出的五档OFI数据，使用z-score标准化数据，在测试集上用训练集的均值和标准差进行标准化

**Label：**对原始数据打标签的方法为三重边界法，以时间t对应的作为上下边界，为在每个时刻t之前的k秒内return标准差，若t之后d秒内return超过则label=1，若低于-则label=-1，若没有达到上下界，则label=0

**原始数据中可以调整的参数有4个：**

1. 计算五档OFI的时间长度time\_window；
2. 过去滑动时间窗口长度k；
3. 三重边界对应参数n和d，其中n为收益率标准差的倍数，d为未来时间窗口长度阈值

**具体数据选择：**选择code=1的trade数据，训练集时间范围是20191204-20200409共53天，测试集范围是20200414-20200831共53天，每天约4700个slice数据，若经过滑动窗口k=20的处理，sample十抽一，input\_shape=(24864, 20,14)

**不同标准差倍数n对label分布的影响：**

按照原文的描述，采用三重边界法在适当的n选择下，1，0，-1的数量分布相近，在此分别尝试不同n=1.0-1.5检验label分布：

* 1. **网络结构**

1. **不同层级结构**

原文的网络结构为两层LSTM，每个cell128个神经元，两层之间dropout=0.5，然后接32个神经元的全连接层，再接3个神经元的全连接层，最后softmax输出结果，对应0，-1，1的概率。

文章使用的bathsize为1024，epoch=5，此处使用的timstep长度为k/3，因为k是计算上下界阈值使用的时间窗口长度，而slice的时间间隔多为3s，所以按照原文的描述，使用的timstep长度为k/3，对每一个sample的最后一个label做预测，数据十抽一查看初步结果，以下是code==1的预测结果

* + - 1. 简单结构：

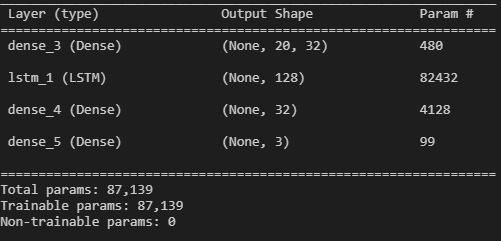
model1\_opt = optimizers.Adam(lr = 0.0005)

model1.compile(loss='categorical\_crossentropy',

              optimizer='adam',

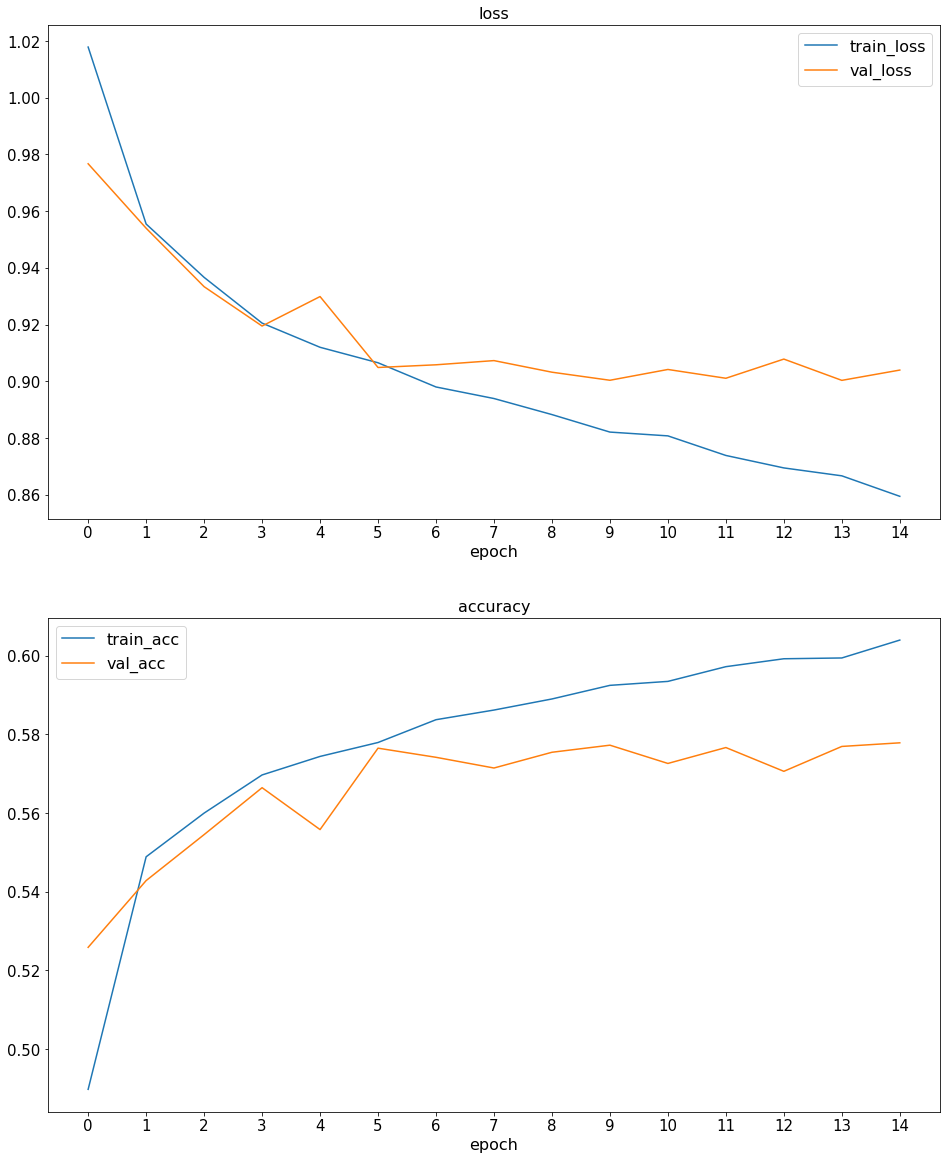
              metrics=['accuracy'])

**结构**：dense32 >> lstm 128 >>dense32 >> dense3 >> softmax (no dropout)



**路径**：model\_d\_l\_2d.h5

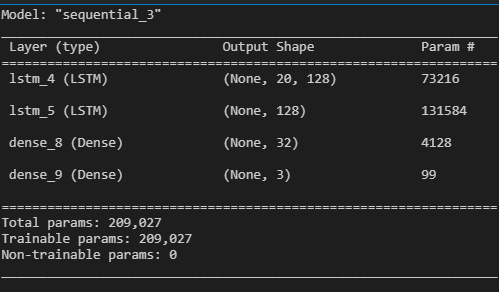
**表现：**





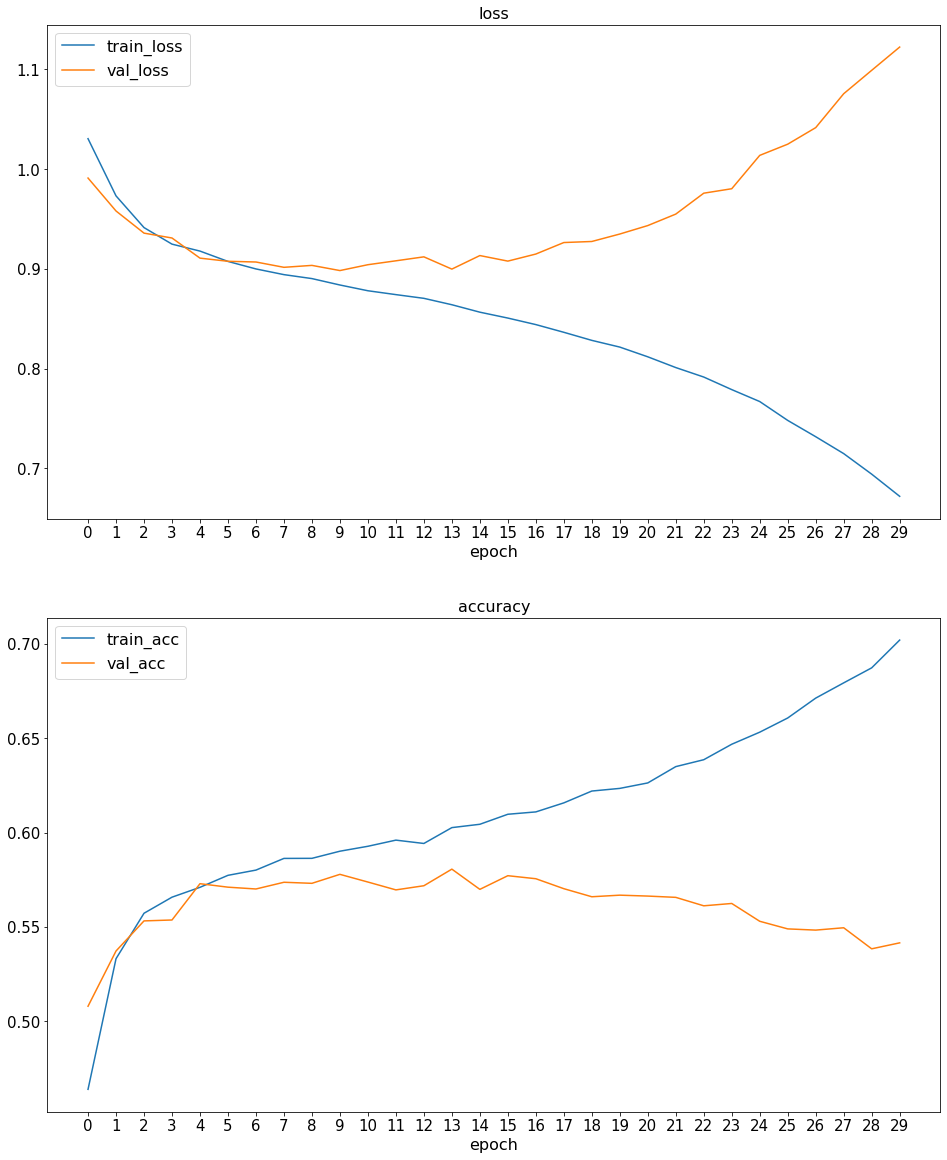
* + - 1. 文章使用的结构

**结构：** lstm 128 >> lstm 128>>dense32 >> dense3 >> softmax



**路径：**model\_l\_dp\_l\_2d.h5

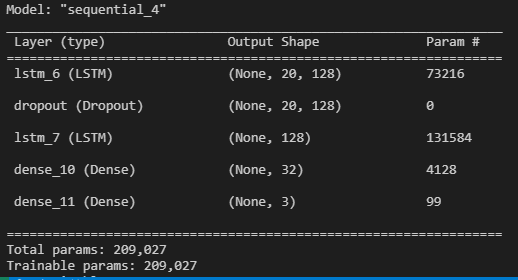
**表现：**



****

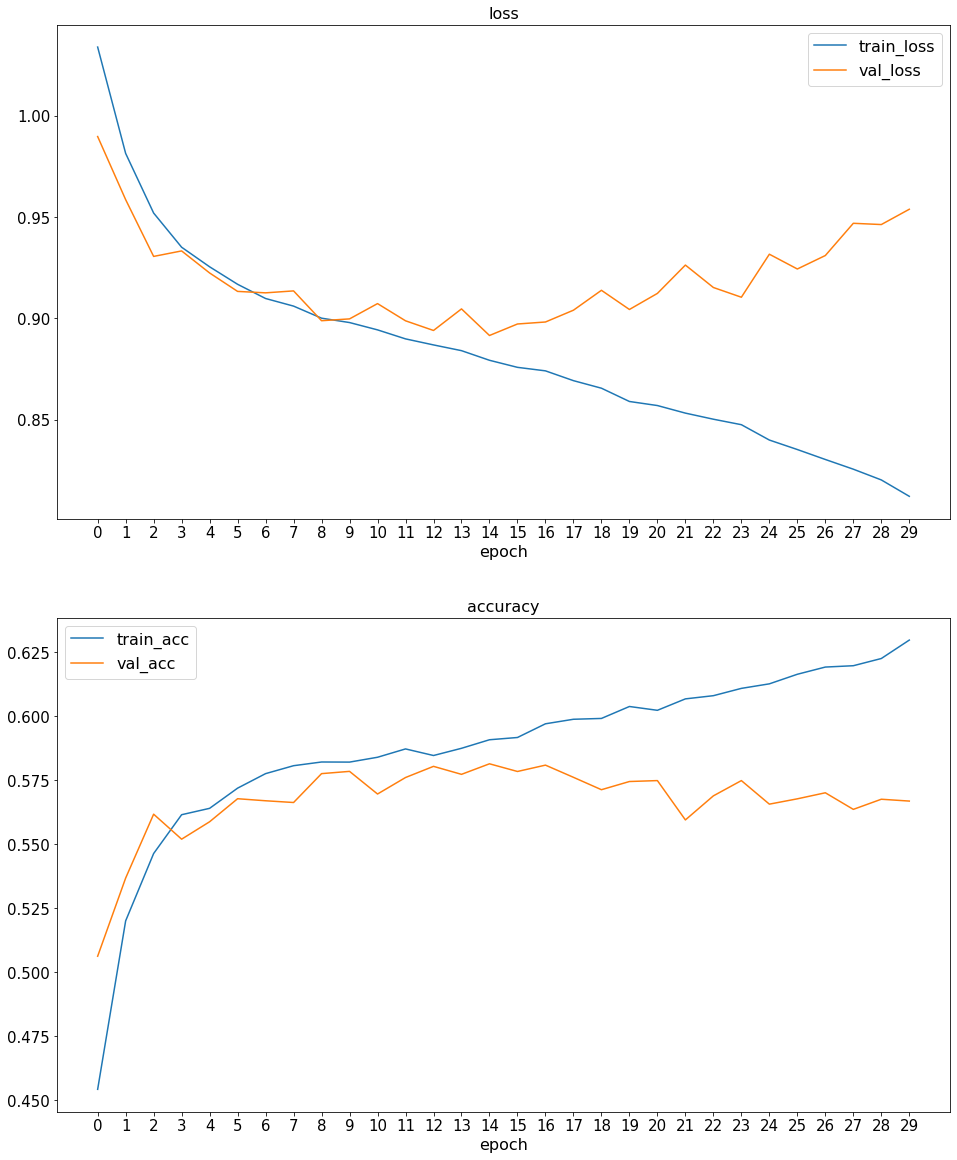
对比有dropout层的结果查看泛化能力的变化：

**结构**： lstm 128>>dropout0.5 >> lstm 128>>dense32 >> dense3 >> softmax



**路径**：model\_l\_ndp\_l\_2d.h5

**表现：**





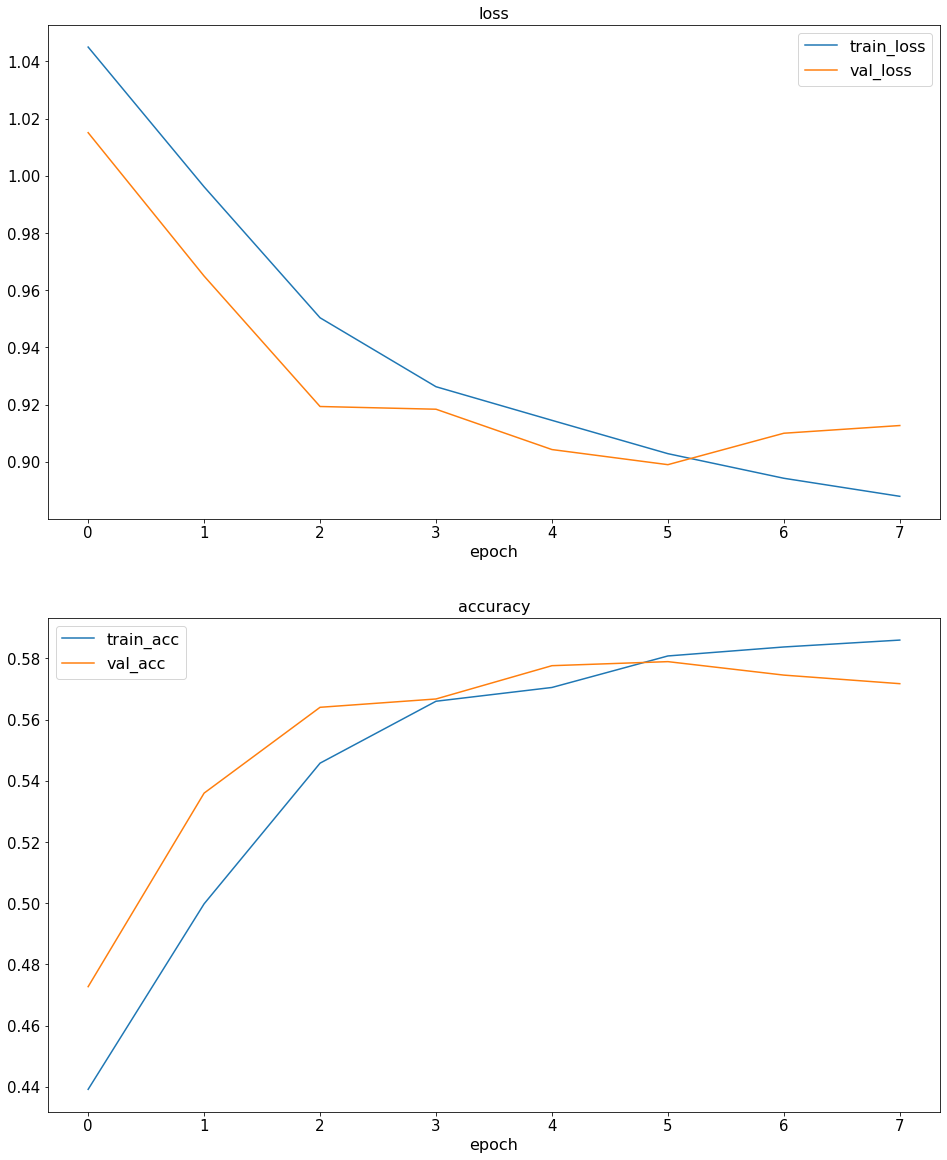
增加dropout层之后，训练集的收敛速度有下降，测试集的表现略有提升但不明显

* + - 1. 文章2使用的结构

结构：lstm512>>lstm256>>lstm128>>dense128>>dense6>>dense3(no dropout)

**路径：**model\_3l\_2d\_ndp.h5

**表现：**earlystop限制，若val\_accuracy在连续两个epoch中没有超过0.001的增长则停止来减少学习时间。





Todo：用过去180s的波动率预测未来120s的label：

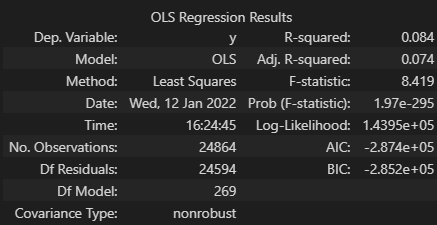
1. 检查OFI是否对预测有显著的帮助

用去掉五档ofi的9个因子进行lstm预测，得到结果：



多元线性回归检查ofi因子的引入对ytrue的影响：

使用20个滑动窗口的14维因子对y\_30s回归：



去掉5档ofi因子后的9维价量因子对y\_30s做回归：

