北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

《数据工程》实践报告

项目名称 个股资金流数据分析实践

学　　院 计算机学院

专业班级 07152101

姓　　名 王凯渤

学　　号 1120210487

2022 年 6 月

个股资金流数据分析实践

摘要：本实践爬取东方财经个股资金流数据，并应用数据处理及分析方法对所获得的数据进行分析。使用爬虫爬取并保存了东方财经个股资金流数据，对数据进行了初步处理，包括删除无用数据以及标准化等。计算了数据的 PEARSON 相关系数以及 SPEARMAN 相关系数，汇出了热力图，得到了数据之间的关系。使用常见机器学习方法对今日涨跌额进行了回归及分类预测，包括神经网络、决策树以及 PCA，取得了较好的结果。

一、 问题描述

资金流向是指资金在股市中主动选择的方向。资金流向的判断对于分析股市的走势和个股的操作有着非常重要的作用。使用机器学习方法可以挖掘数据中潜在的信息，对于判断股票涨跌情况有积极的作用。分析个股资金流的相关数据可以分析验证股市涨跌情况，对于投资者选择股票有积极意义。本实践爬取了一日内的数据，分析当日各数据之间的关联，寻找今日涨跌额与其他数据之间的联系。

二、 数据获取

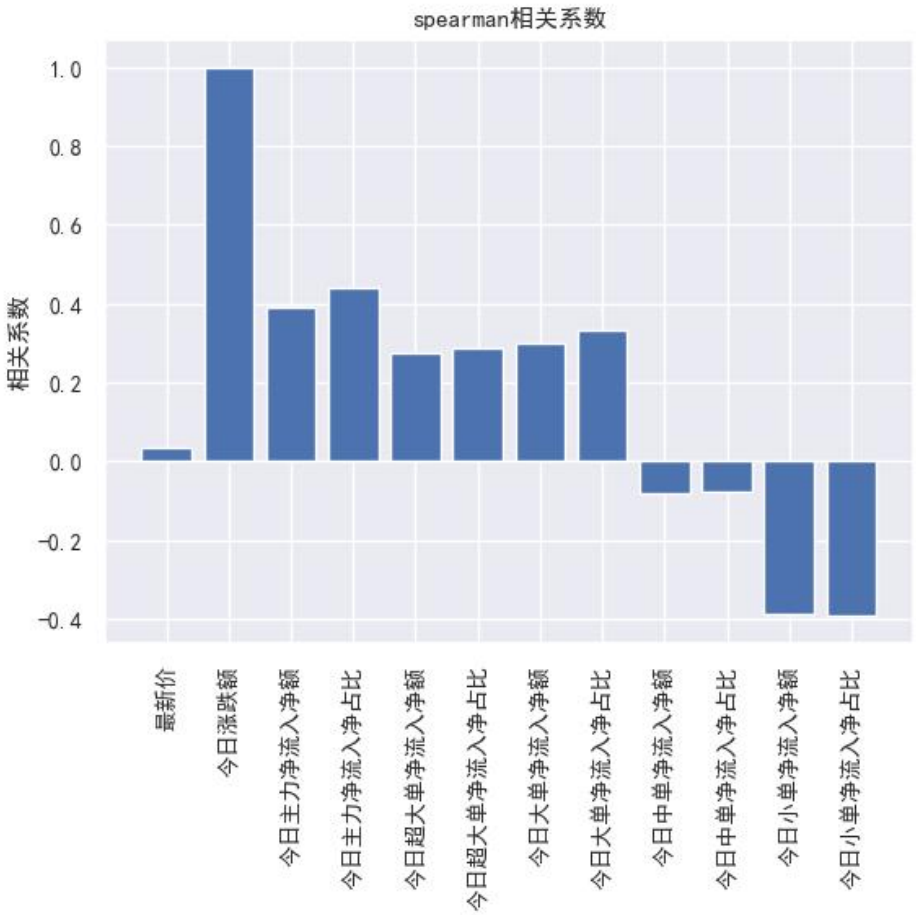
实验数据来源于沪深两市实时资金流向排行_ 数据中心 _ 东方财富网 (eastmoney.com)，使用爬虫技术获取了 5 月 29 日个股资金流数据。

数据集共有 5046 行，每行代表一只股票；14 列，分别为股票代码，股票名称，最新价，今日涨跌额，今日主力净流入净额，今日主力净流入净占比，今日超大单净流入净额，今日超大单净流入净占比，今日大单净流入净额，今日大单净流入净占比，今日中单净流入净额，今日中单净流入净占比，今日小单净流入净额，今日小单净流入净占比。

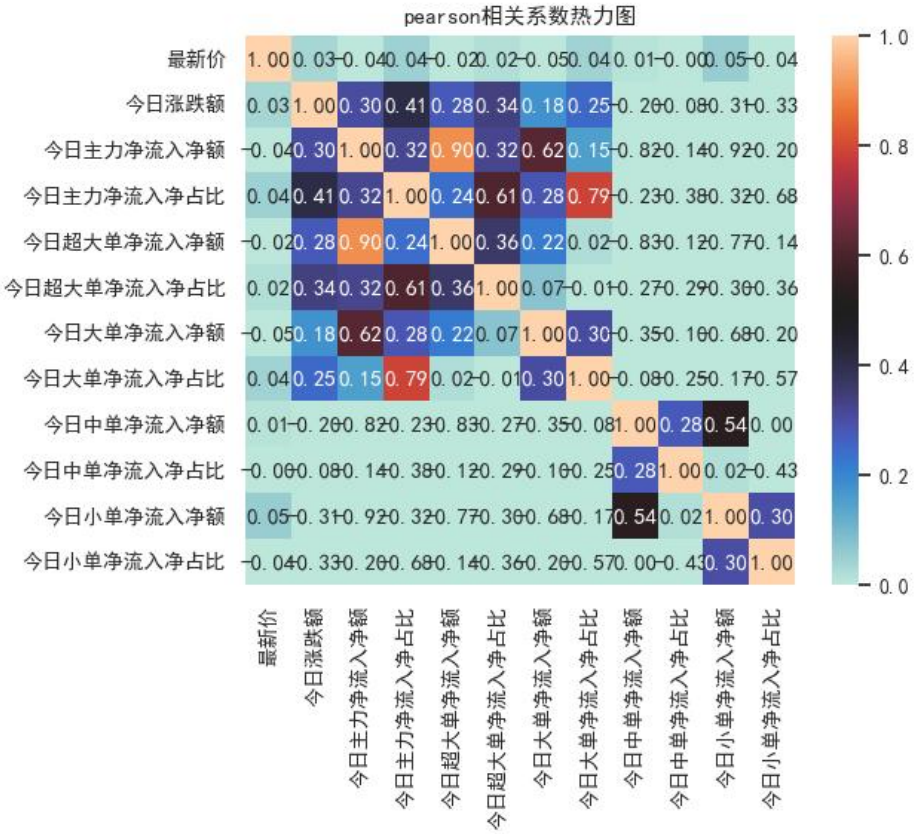
本实践主要分析今日涨跌额与其他数据之间的联系。

三、 数据处理（指标分析、特征工程）

1. 首先去掉无用数据。由于原始数据已经足够好，所以不需要过多处理。显然股票代码、股票名称两列与预测值并无直接关联，仅用于识别股票，将其删除。
2. 查看数据集缺失值，发现爬取数据规整齐全，没有缺失值。由于从现有网站上爬取数据，可以认为数据已经经过预处理，得到的数据均真实有效且无过多异常。
3. 对数据进行规范化，标准化化为均值为 0，标准差为 1 的数据。
4. 划分测试集与训练集。标签数据为今日涨跌额，80%划分到训练集，得到 X_TRAIN, X_TEST, Y_TRAIN, Y_TEST。
5. 计算今日涨跌额与其他数据的 SPEARMAN 相关系数，得到下图。

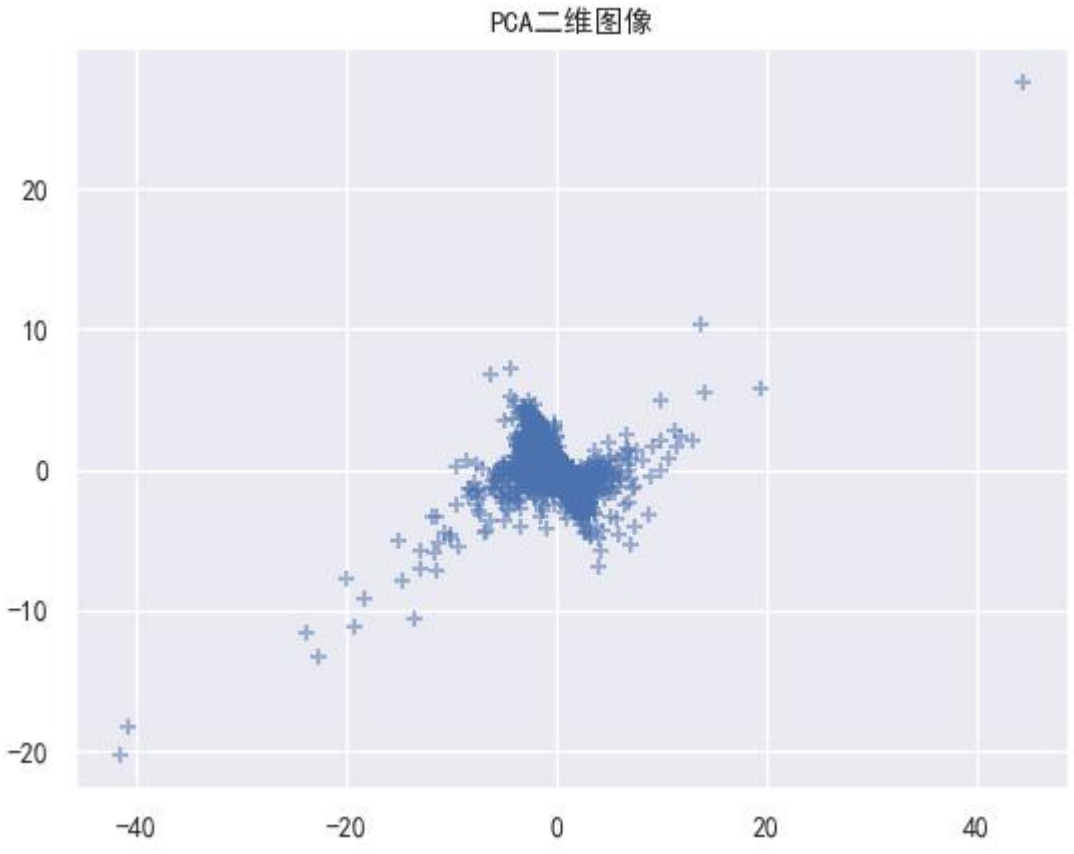


6. 绘出 PEARSON 相关系数热力图，得到热力图如下所示。



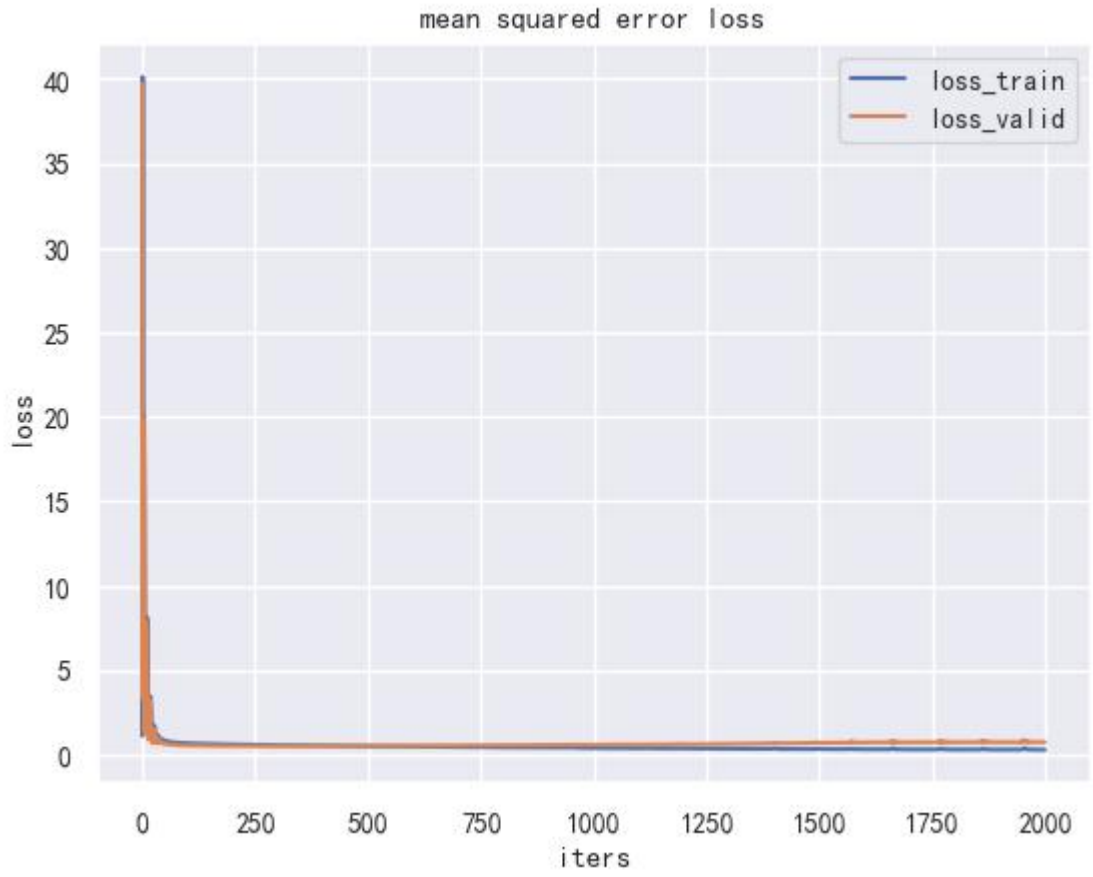
四、 数据分析

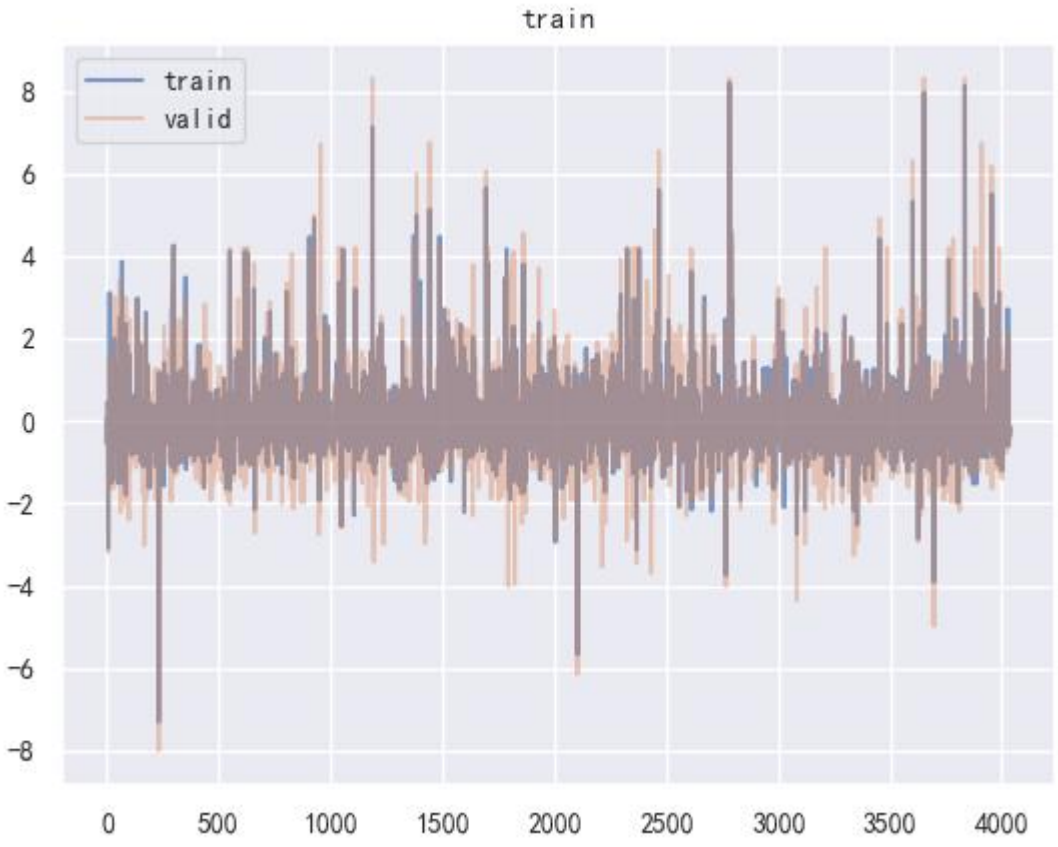
1. 首先使用主成分分析（PCA），将处理后的 12 个维度压缩到 2 个维度上，得到散点图。

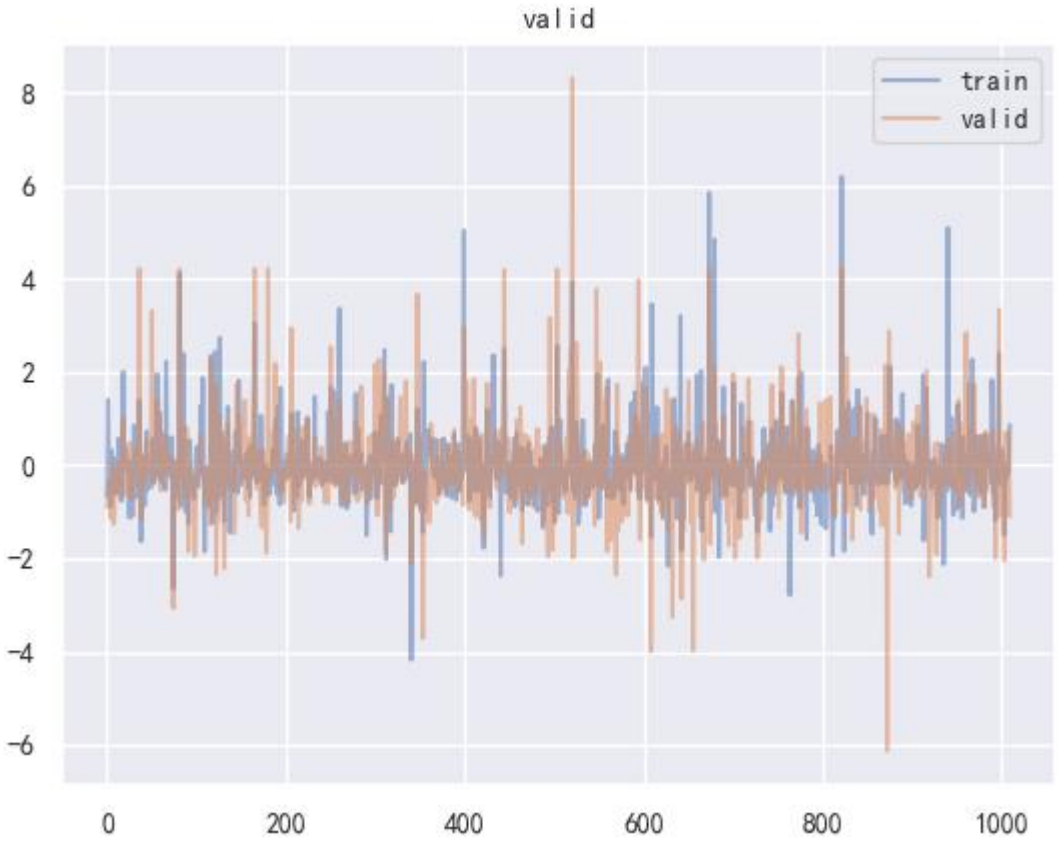


压缩到两个维度时，保留的信息仅为 0.56。经测试，当压缩到 5 个维度时，可以保留 0.86 的信息，可以认为保留了较为充足的信息。由于本实践数据集较小，仅有 12 个维度，所以后续分析不使用压缩后的数据，尽可能保留所有信息。

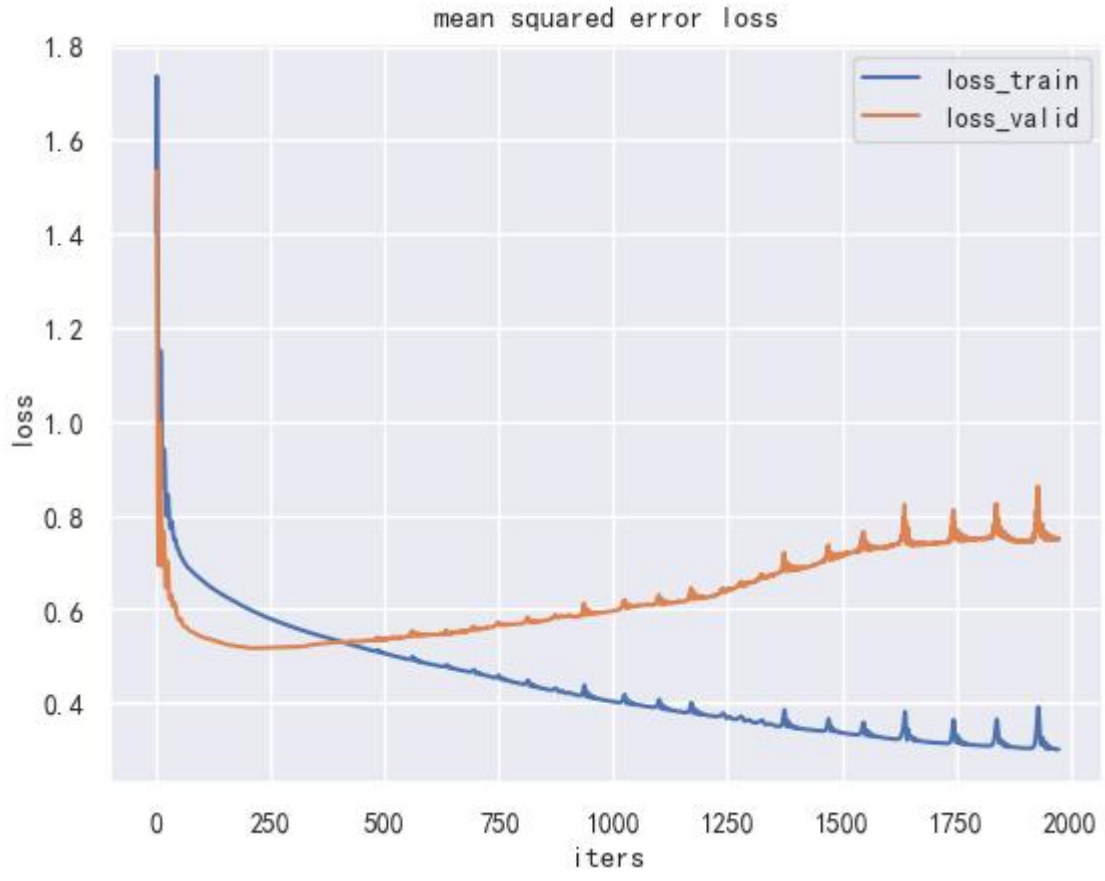
2. FFN 回归预测今日涨跌额。使用简单的前馈神经网络对今日涨跌额进行预测，隐层 128 维，ADAM 算法，MSELOSS 迭代 2000 轮[1]。得到如下训练结果。


mean squared error loss
loss_train
loss_valid
loss
iters
0
250
500
750
1000
1250
1500
1750
2000
40
35
30
25
20
15
10
5
0


train
train
valid
8
6
4
2
0
−2
−4
−6
−8
0
500
1000
1500
2000
2500
3000
3500
4000

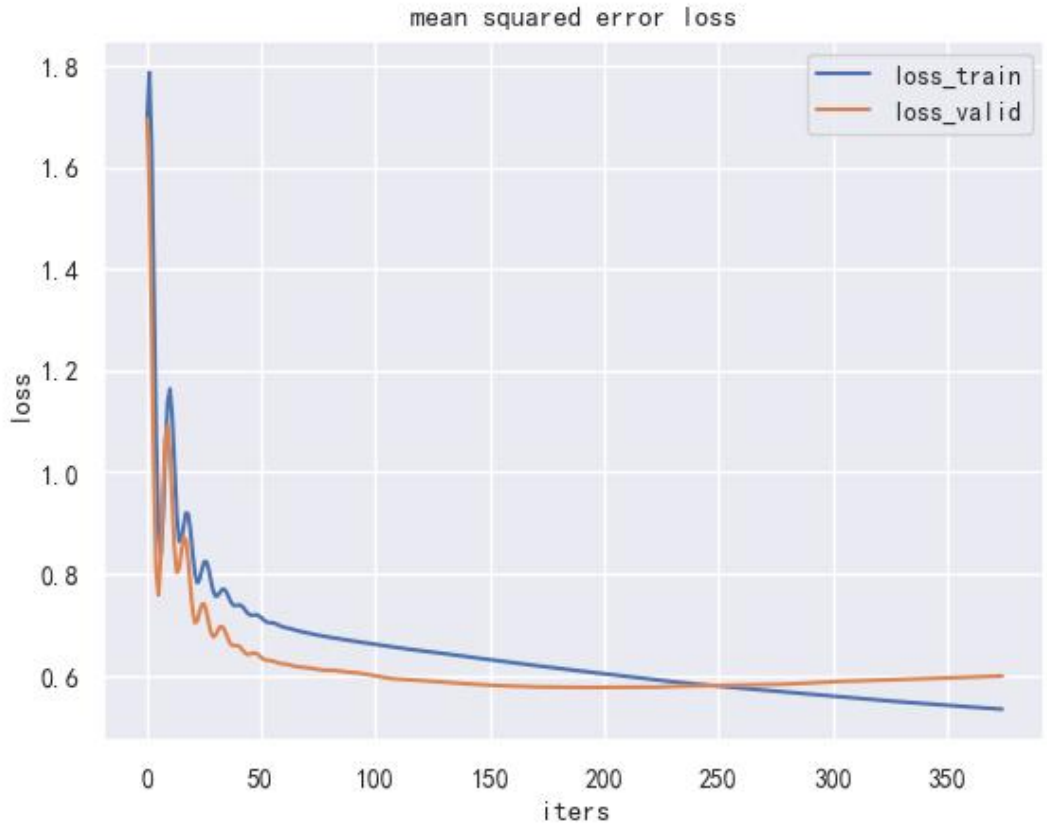


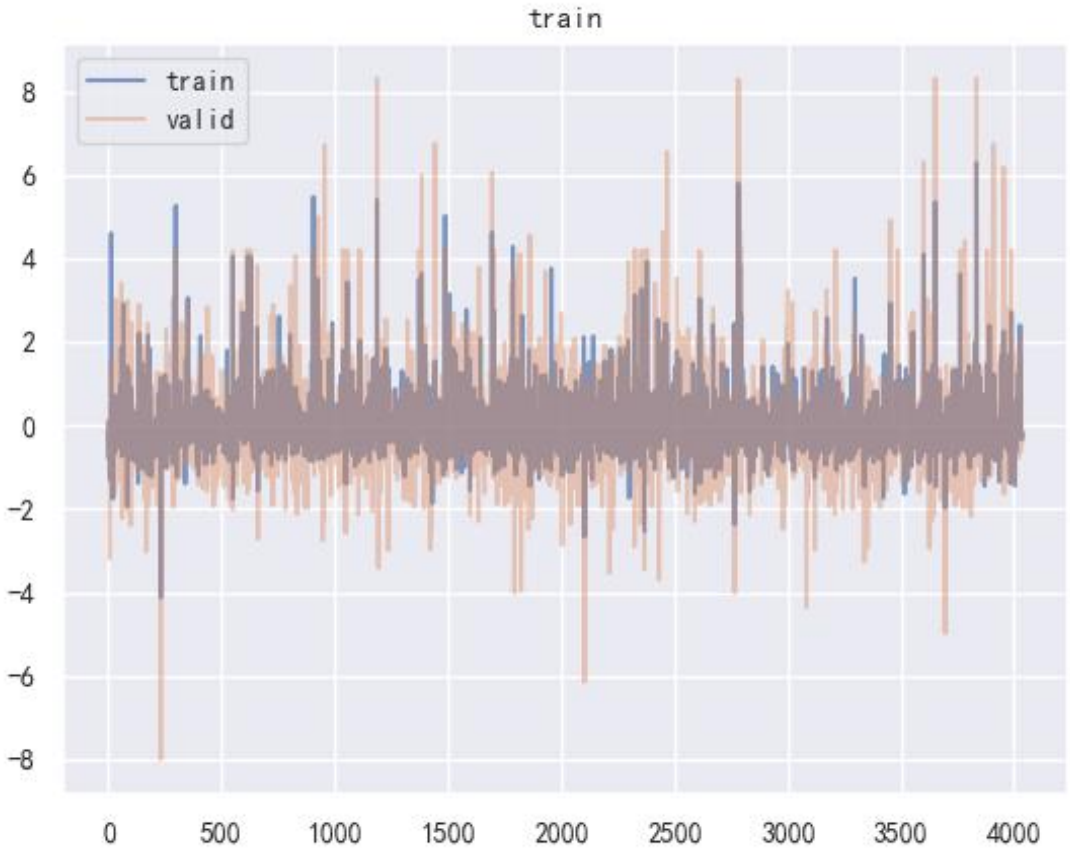
为了更直观地看到 LOSS 的变化情况，从第 25 轮的 LOSS 开始绘制 LOSS 下降图像。

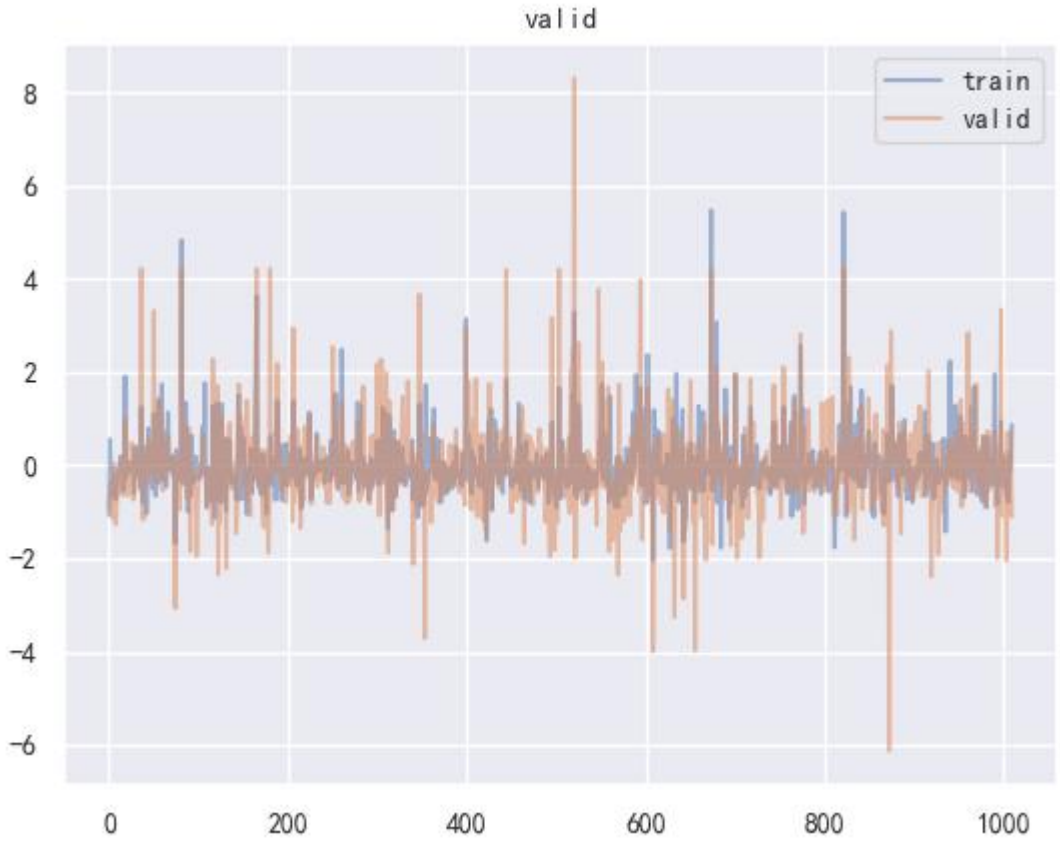


可以看到在 400 轮左右达到最优，之后进入过拟合。

重新训练 400 轮。得到如下结果。







3. 决策树分类今日涨跌情况，进行分类任务，对数据进一步处理。对于今日涨跌额中大于等于 0 的数据即“涨”置为 1，小于 0 的数据即“跌”置为 0。

   使用网格搜索，搜索 SPLITTER、CRITERION 及 MAX_DEPTH 三个参数，寻找决策树最佳参数。得到结果如下所示。

   网格搜索最佳参数设置： {'CRITERION': 'GINI', 'MAX_DEPTH': 5, 'SPLITTER': 'BEST'}

   网格搜索最佳得分： 0.6875605776191002

   网格搜索决策树最佳得分 0.692079207920792

   混淆矩阵：

   [[484 88]

   [223 215]]

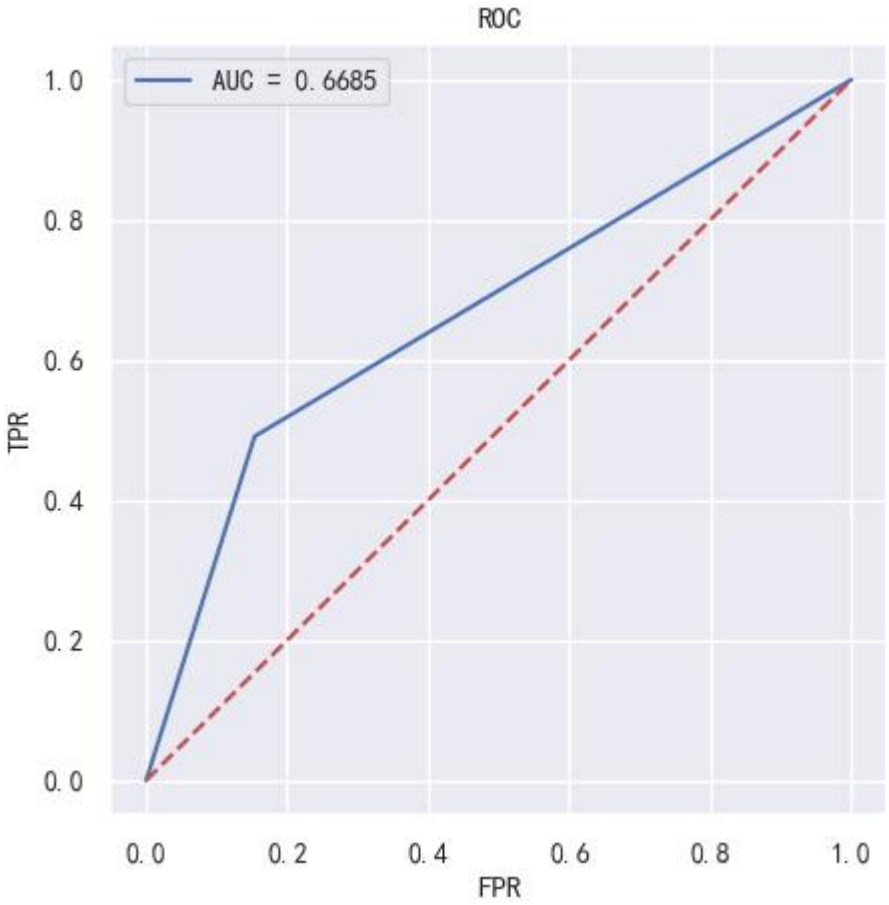
   ACCURACY SCORE = 0.69

   F1 SCORE = 0.58

   PRECISION SCORE = 0.71

   RECALL SCORE = 0.49

   ROC AUC SCORE = 0.67



五、 数据展示

可视化图像已在前文展示。决策树可视化图像位于可视化文件夹中。

从皮尔逊相关系数热力图以及斯皮尔曼相关系数中可以看到，今日涨跌额与今日主力净流入额、净流入占比、净流入净额，今日超大单净流入额、净流入占比、净流入净额相关性最大。

PCA 图像中也可以看到，大部分点分布较为集中，仅有个别点位置较偏僻。

FFN 回归预测中，可以看到，迭代轮数较多时，对数据集有明显过拟合趋势，对验证集的预测准确度明显降低。而迭代轮数较少时，预测值的偏差较大，但基本能够正确预测涨跌情况。

决策树分类中，可以看到各项指标不算差，基本能够达到 0.7 的准确率。

六、 结论

通过包括神经网络、决策树两种监督学习方法，PCA 一种无监督学习方法对个股资金流数据进行分析，可以得到较好的结果。神经网络得到的最终 MSE 可以低至 0.6，决策树取得的 F1 值也可以达到 0.58，虽然最终结果并不算太好，但是成功完成了数据工程的整套流程，并且实现了许多数据可视化工作。

七、 心得体会

机器学习作为一种从数据中提取信息的方法，已经广泛应用于各个领域并且取得了惊人的效果。通过数据工程实践整套流程，可以帮我们熟悉、掌握较为前沿、热门的技术方法。实际上获取的数据本身就有极强的相关性，个股资金流向本身就通过净流入额、净流入占比、净流入净额反映，而且由于数据集很小，数据量不够，得到的结果也差强人意。

并且没有使用很好的方法，对于股市分析缺乏专业知识，不了解各项指标具体含义，仅能猜测各项指标间的关联。

八、 参考文献

[1]pytorch 一个最简单的回归预测_pytorch 回归_1731064109 的博客-CSDN 博客