# 课程预习报告

电气2011班 曹瑜 U202012613

1. 预习目标

1.1对这门课的期望

了解到这是学校开设的软件工程训练营，同时因为自己在负责一项与软件相关的大创项目，于是想通过较为专业的学习，为自己大创项目的顺利展开奠定基础，同时自己也参加过一些建模比赛，而在比赛过程中发现自身对于一些模型的掌握还不够熟练，还有许多模型未曾接触过，于是也是想通过这门课程，让自己在建模比赛中能有更好的表现。

1.2对人工智能的理解

在我眼中人工智能是科技史上一个具有重要意义的技术突破，是关于知识的学科，通过计算机等科技的应用，在现实中不断观察总结，在技术上不断创新，最终开发出一系列产品，从而达到在某些领域解放人力，节省人力资源的目的。希望人工智能能够今后在一些重大领域，一些人们未能充分涉猎或者限于一些自然条件停滞的领域，发挥出他的智能性，为人们的生活、生产、科研等领域做出巨大助力。

1.3预习计划

因为考虑到课程的展开是基于python语言展开的，于是与python编程相关的软件VScode需要先调配好环境，人工智能的基础离不开建模，所以我会先提前学习一些常见的模型比如微分方程、神经网络等，也算是一个预习过程。希望经过课程的学习把我之前对于机器学习这方面的短板补足，也能够在建模比赛过程中有机会选择更加适合的模型，提升自己的建模竞赛能力。因为自己的主要重心是针对建模这一类比赛的，所以我想通过做类似于建模题的预习项目。

1. 基础知识预习

2.1预习工具

我打算使用的编程工具是VScode，虽然之前在学院的建模课程学习过程中学过相关的使用，但是对于一些环境的配置的使用还不会，经常会出现小问题，于是我在网上查找资料，学习与VScode相配套的anaconda软件的使用，通过学习现在我更够针对过程中出现的问题及时做出调整，下载运行过程中缺失的库，增设不同的python环境应对不同的代码需求。对于之前所不熟悉的神经网络模型，也因着暑期的一次建模比赛，在比赛过程中向队友进行了初步学习，算是补足了自己在建模方面的一个短板。

1. 预习项目设计

3.1 项目选题

3.1.1 项目选题原因

因为自己算是有一定的编程基础，所以想通过预习的过程，巩固加强自己在这方面的能力，因为老师给的选题中有关于数据分析的，于是我结合暑期参加的一个数据分析的建模比赛完成了此次预习作业。

3.1.2 计划实现功能

根据历史数据，构建模型，对布伦特石油产品的open、close、high和low在未来一个月内的变化进行合理预测，并评估模型的效果

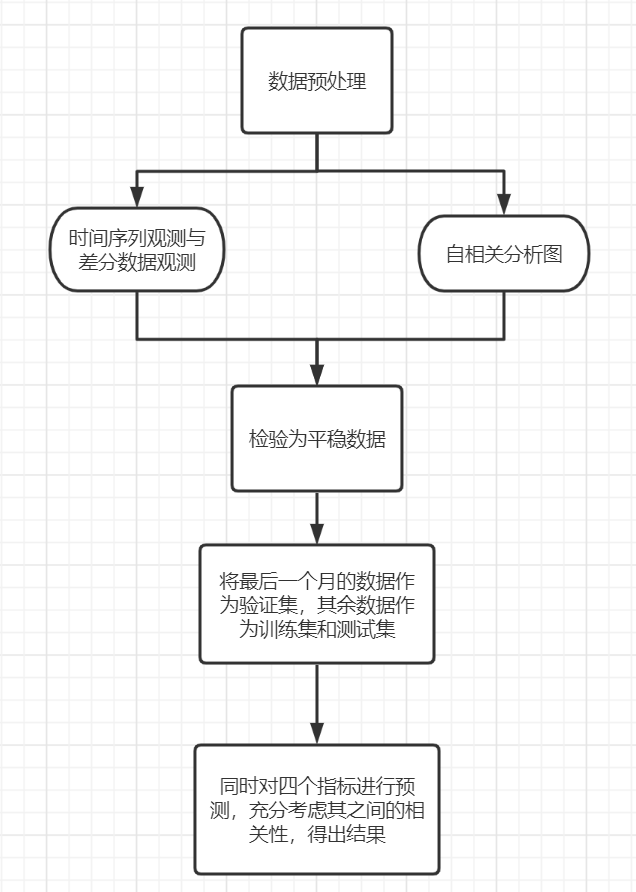
3.1.3 项目使用工具

本题采用向量自回归模型（VAR），利用VScode进行程序编写。

3.2 程序设计

3.2.1 程序设计思路

采用向量自回归模型（VAR），之所以不采用ARIMA模型是因为考虑到布伦特石油产品的open、close、high和low指标之间会存在一定的关系，这几个变量之间会有相互影响，因此采用适合多变量的VAR模型。使用后一个月的数据作为验证集，剩余的数据用于建立预测模型



3.2.2 问题复述

目前国际市场上比较大的主要是WTI和布伦特两种石油金融产品。我们收集了它们的历史信息（但两种产品的时间区间不同），其中open表示开盘价，close表示收盘价，high表示最高价，low表示最低价。这两种产品的交易周期和交易日不同，并不是每一天都会发生交易。价格的单位都是美元。

请构建模型，对布伦特石油产品的open、close、high和low在未来一个月内的变化进行合理预测，并评估你们模型的效果。

3.2.3 程序代码

**数据预测：**

#校赛

from matplotlib import pyplot as plt#导入库

import numpy as np

import pandas as pd

from statsmodels.tsa.api import VAR

exp\_8 = pd.read\_csv('布伦特石油数据.CSV').iloc[:, 1:]#读取石油数据，且读取除表格第一行外的所有数据

#exp\_8 = exp\_8[::6]

def rolling\_predict\_var(start, end, step):#定义预测函数

"""

To predict next day and update the model with observation.

:params start: point start to predict

:params end: point end to predict

"""

prediction = exp\_8[:start].values#设置循环，以设定的数据内容为训练集，预测下一时间段的值，再将预测值更新到训练集中，再进行预测

for i in range(start, end, step):

X = exp\_8[:i].values

model = VAR(X)

res = model.fit()

prediction = np.vstack((prediction, res.forecast(X, step)))

plt.plot(X[start:end, 0], label='close data')#画出所要求的数据对应的图像

plt.plot(X[start:end, 1], label='open data')

plt.plot(X[start:end, 2], label='high data')

plt.plot(X[start:end, 3], label='low data')

plt.plot(prediction[start:end, 0], label='close pre')

plt.plot(prediction[start:end, 1], label='open pre')

plt.plot(prediction[start:end, 2], label='high pre')

plt.plot(prediction[start:end, 3], label='low pre')

plt.legend()

plt.show()

return prediction[start:end]

prediction = rolling\_predict\_var(685, 715, 1)#设置训练集数量以及训练步长

**数据检验：**

import numpy as np

import pandas as pd#导入库

exp\_7 = pd.read\_excel('工作簿3.xlsx')#阅读表格数据

exp\_7.head()#

from statsmodels.graphics import tsaplots#导入库

from matplotlib import pyplot as plt

tsaplots.plot\_pacf(exp\_7['high'], zero=False) # zero=False即不画与当前的关系

plt.show()#显示图像

fig = plt.figure(dpi=150)#图像分辨率为150

import sys

import os

import pandas as pd#导入库

import matplotlib.pylab as plt

#%matplotlib inline

from statsmodels.tsa.stattools import adfuller

import statsmodels.api as sm

import statsmodels.formula.api as smf

import statsmodels.tsa.api as smt

from statsmodels.tsa.stattools import adfuller

from statsmodels.stats.diagnostic import acorr\_ljungbox

from statsmodels.graphics.api import qqplot

"""中文显示问题"""

plt.rcParams['font.family'] = ['sans-serif']

plt.rcParams['font.sans-serif'] = ['SimHei']

data = pd.read\_excel("工作簿4.xlsx",index\_col="Date",parse\_dates=True)#读取表格数据

data.head()

data["diff1"] = data["low"].diff(1).dropna()#画出差分图像

data["diff2"] = data["diff1"].diff(1).dropna()

data1 = data.loc[:,["low","diff1","diff2"]]

data1.plot(subplots=True, figsize=(18, 12),title="差分图")

print("单位根检验:\n")

print(adfuller(data.diff1.dropna())) #单位根检验

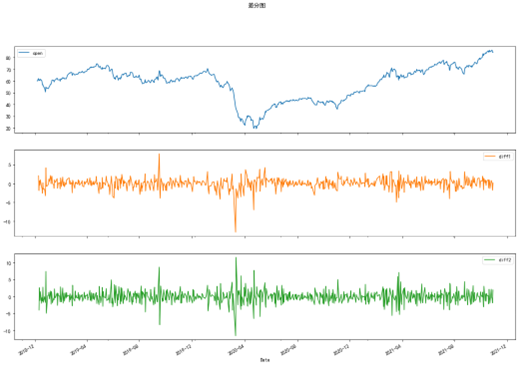
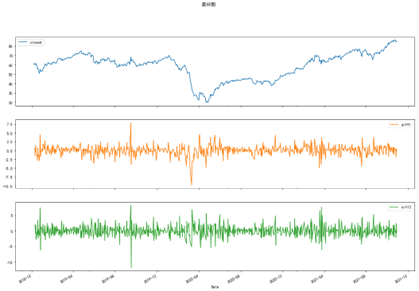
* 1. 运行结果

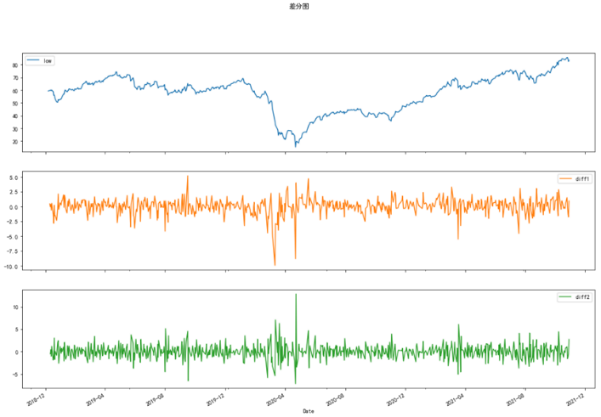
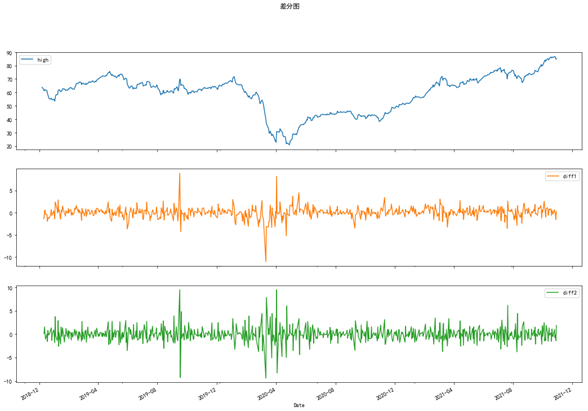
3.3.1 数据预处理

VAR模型的适用条件要求时间序列是平稳的，于是首先对变量进行平稳性检验。

方法一：时间序列观测与差分数据观测

从时间序列图以及一阶、二阶差分图可知，四个指标之间相差不大，于是我们这里分析就以其中一个指标为例，展开说明。对于每一个指标而言，从时序图可以大致判断其历史数据的走向，从一阶、二阶差分图可以观察到数据的相对变化情况，从图中可以看出每个指标的一阶差分与二阶差分图像都比较稳定，但是因为不同人对于图像的判断不同，且单纯从图像并不能够较准确地体现出时间序列的平稳性，于是我们采用对数据进行单位根检验的方法对数据进行进一步验证。





方法二：自相关分析图

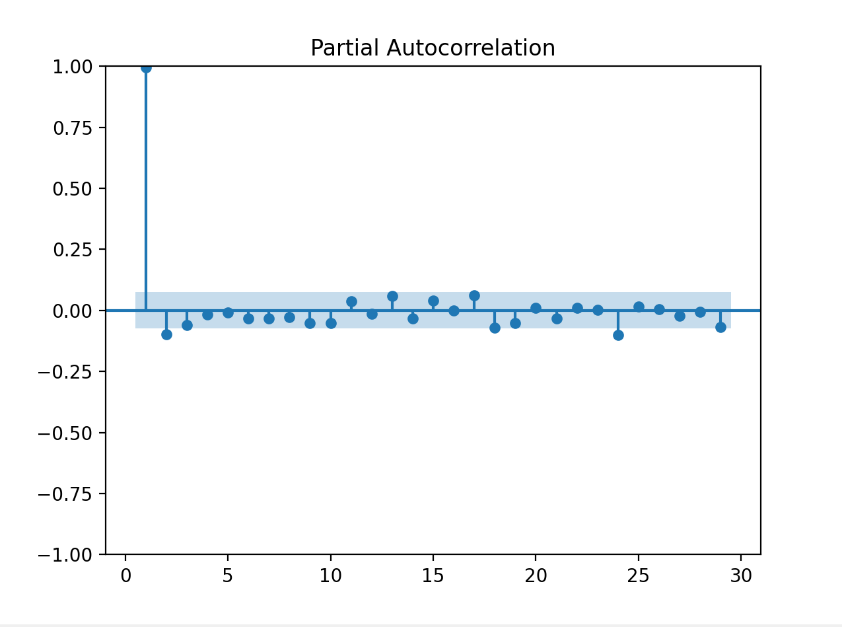


图3 偏自相关系数图

通过序列的自相关分析图来判断时间序列的平稳性，如果序列的自相关系数随着滞后阶数的增加很快趋于0，即落入随机区间，则序列是平稳的；反之，序列是不平稳的。此处通过观测可以发现时间序列的偏自相关系数都在95％的置信区间里面，且后面的点都未超出该置信区间，于是可以认为其是衰减到零的。这也说明我们要用来进行训练的时间序列是平稳的。

通过以上平稳性判断方法，我们直观、准确地判断出开盘价（open）、收盘价（close）、最高价（high）、最低价（low）四个指标的时间序列数据是平稳的，因此可以采用VAR模型进行其进行分析预测。

将所给布伦特石油产品历史数据作为训练集对模型进行训练，搜集未来一个月四个指标的数据作为验证集对模型的预测效果进行检验评估。将其用图表的形式进行呈现：

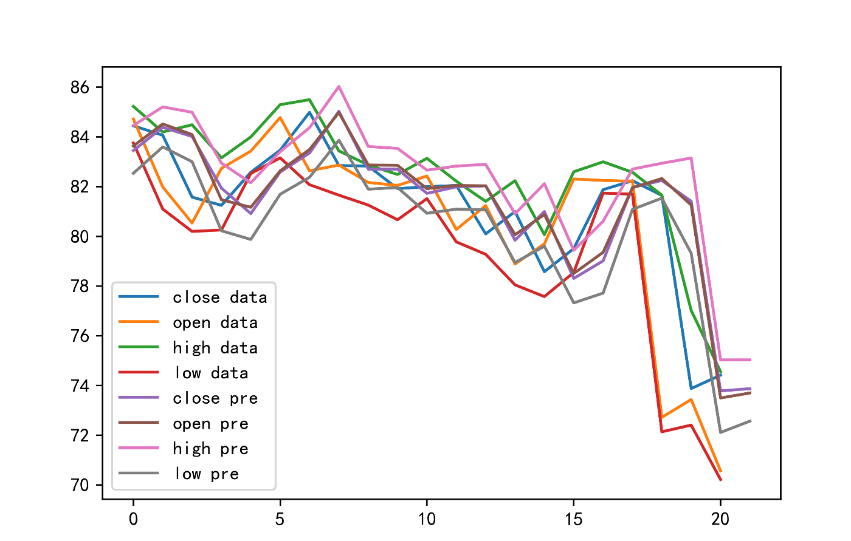


图4 VAR回归结果与真实价格变动图

图4为使用VAR模型，加入额外的数据作为样本验证集进行长期预测的得到的结果。从图形上看预测数据相对于真实数据会产生一定的时间滞后现象，这是因为在实际情况下数据序列中产生了变化趋势，而基于滑动时间窗口策略的对发生变化趋势的数据感知是滞后的，于是也就会产生上面的滞后现象。但是从预测值的趋势和变化上面看，运用VAR模型进行预测的结果与真实值的变动不仅趋势一样，而且能够较为实时地反应价格的变动，对于真实情况的拟合程度较好。

1. 预习总结

4.1 预习成果

经过课程的学习把我之前对于机器学习这方面的短板补足，也能够在建模比赛过程中有机会选择更加适合的模型，提升了自己的建模竞赛能力。对于一些基本的模型以及编程语言有了更加熟练的掌握。

4.2 预习不足

对于一些比较难下载的库，还是不能很好地进行处理，在这个过程中会浪费很多时间。