基于 R 语言的贵州茅台股票分析与预测

课程名称：　数据挖掘

小组成员：　张晓宇、刘懋霖

联系方式：　zxynju2017@outlook.com

指导老师：　石进

报告日期：　2020 年 7 月

摘要：时间序列和随机森林是数据挖掘中用于分析和预测的两种常用方法，本文借助于 R 平台，采用随机森林、时间序列的研究方法，从横截面与纵向两个维度来对股票数据进行预测分析，探求股票每日收盘价的重要影响因素，并对未来的收盘价进行预测，借此加深对数据挖掘的理解以及相关理论知识的实践，便于后续对数据挖掘方法的应用。

关键词：时间序列、随机森林、R 语言、股票

一、研究背景

股票是股份公司发给股东证明其所入股份的一种有价证券，它可以作为买卖对象和抵押品，是资金市场主要的长期信用工具之一，代表着其持有者（即股东）对股份公司的所有权。购买股票也是购买企业生意的一部分，即可和企业共同成长发展。因此，对股票的走势甚至未来数据进行预测是投资者十分关心的问题。

中国贵州茅台酒厂（集团）有限责任公司（简称：茅台集团）是一家中国特大型国有酒业公司。公司涉足产业包括白酒、保健酒、葡萄酒、金融、文化旅游、教育、酒店、房地产及白酒上下游等。其主导产品贵州茅台酒被称为世界三大（蒸馏）名酒之一。

本文借助于 R 平台，采用随机森林、时间序列的研究方法，从横截面与纵向两个维度来对股票数据进行预测分析，以此探求股票每日收盘价的重要影响因素，并对未来的收盘价进行预测。

二、研究方法

2.1 随机森林

随机森林是一种组成式的有监督学习方法。在随机森林中，我们同时生成多个预测模型，并将模型的结果进行投票汇总确定最终预测结果，以提高分类结果的正确率。

随机森林的算法涉及对样本单元和变量进行抽样，从而生成大量决策树。对每个样本单元来说，所有决策树预测类别中的众数类别即为随机森林所预测的这一样本单元的类别。同理，随机森林也可以用于回归。每棵树输出一个回归值，以所有回归值的加权均值作为最终预测结果。

2.2. 时间序列分析

时间序列（或称动态数列）是指将同一统计指标的数值按其发生时间的先后顺序排列而成的数列。时间序列分析主要对时间序列进行观察、研究，找寻它变化发展的规律，预测它将来的走势，其目的是根据已有历史数据对未来进行预测。

根据观察时间的不同，时间序列中的时间可以是年份、季度、月份或其他任何时间形式。简言之，时间序列分析的基本思想是根据系统有限长度的历史记录（即已知的观察数据），建立能够比较精确地反映时间序列中所包含的动态依存关系的数学模型，并借以对系统的未来行为进行预测。

三、具体研究过程

本文的研究和分析对象来自网易财经沪深个股行情获取的茅台股票数据，该数据集包含 15 个字段，且具有趋势性等特点，是一种非平稳的时间序列。因此大致思路是通过差分使得数据变得平稳，从横截面与纵向两个维度，采用随机森林、时间序列的研究方法，借助 ARIMA 模型，来对股票数据进行分析和预测。

需要说明的是，横截面数据，是在一个给定的时间点测量数据值，并对自变量和因变量之间的关系进行分析，能够进行分类预测以及探寻因变量对于自变量的重要程度。在股票预测中，可以探求其他因素对于收盘价的影响程度，从而为投资者根据当日数据变化做的决策提供依据。

纵向数据，则是随着时间的变化反复测量变量值，持续跟踪某一现象，获取更多的了解，对该数据的未来值进行预测。在股票预测中，可以持续追踪每日收盘价的变化，探求收盘价根据时间的变化情况。同时还可以使用纵向回归，将前 N 天的收盘价作为因变量来预测当天的收盘价。

此外，本文还将数据集进行了一定的划分，使得训练和验证能够独立进行，能够检验模型的泛化能力，使得整个预测能够达到比较好的效果。

3.1. 股票行情数据获取

本次预测数据来源于网易财经沪深个股行情（http://quotes.money.163.com/stock），通过搜索股票名称“茅台”，在详情页下载历史交易数据。数据包含股票从上市日（2001.8.27）至下载时（2020.7.17）的所有交易数据，具有交易日期、股票代码、名称、收盘价、最高价、最低价、开盘价、前收盘、涨跌额、涨跌幅、换手率、成交量、成交金额、总市值、流通市值共计 15 个字段。

日期	股票代码	名称	收盘价	最高价	最低价	开盘价	前收盘	涨跌额	涨跌幅	换手率	成
2020/7/15	'600519	贵州茅台	1752.53	1782.6	1744.71	1744.88	1762	-9.47	-0.5375	0.3241	
2020/7/14	'600519	贵州茅台	1762	1782.03	1748.4	1781.99	1781.99	-19.99	-1.1218	0.3669	
2020/7/13	'600519	贵州茅台	1781.99	1787	1714.68	1714.68	1713.85	68.14	3.9758	0.4834	
2020/7/10	'600519	贵州茅台	1713.85	1736.6	1688.94	1689	1706	7.85	0.4601	0.2953	
2020/7/9	'600519	贵州茅台	1706	1712	1682	1691	1681.34	24.66	1.4667	0.2789	
2020/7/8	'600519	贵州茅台	1681.34	1719	1660	1680	1688	-6.66	-0.3945	0.4233	

图 1 原始数据集

3.2. 数据读取与预处理

读取并观察数据集，了解到数据集中有交易日期、开盘价、最高价、最低价、

收盘价、成交量、成交金额等多个变量。其中，数据集中的开盘价和收盘价代表股票在某一天交易的起始价和最终价；最高价、最低价和最后交易价表示当天股票的最高价、最低价和最后交易价格；成交量是指当天买卖的股票数量，而营业额是指某一特定公司在某一特定日期的营业额。

需要注意的是，收盘价是个股一天价格走势中最重要的指标，由于股市中损益的计算通常由股票当日的收盘价决定，通俗的说，只有收盘价才是判断赚钱或赔钱的基准，因而我们选取收盘价作为预测的目标，也就是因变量。

在数据与处理方面，对数据集进行算法分析之前，需要去除掉对分析无用的列，并检查每列的数据类型是否都是数值型，避免带有符号的数值被解析为字符串变量而影响回归树的分析，之后按年份，拆分数据集为训练集和测试集。

```
> stock
         日期   收盘价   最高价   最低价
1   2020/7/15 1752.53 1782.60 1744.71
2   2020/7/14 1762.00 1782.03 1748.40
3   2020/7/13 1781.99 1787.00 1714.68
4   2020/7/10 1713.85 1736.60 1688.94
5    2020/7/9 1706.00 1712.00 1682.00
6    2020/7/8 1681.34 1719.00 1660.00
7    2020/7/7 1688.00 1744.82 1601.00
8    2020/7/6 1600.00 1616.00 1531.81
9    2020/7/3 1541.79 1552.50 1516.88
10   2020/7/2 1544.00 1550.90 1495.00
11   2020/7/1 1494.27 1506.10 1464.02
12 2020/6/30 1462.88 1468.98 1455.12
```

图 2 去除无用列，并转化数据类型为数值型

此外，由于股票市场在周末和公共假期休市，因此数据中缺失了一些日期值如 2/10/2018、6/10/2018、7/10/2018，其中 2 号是国庆节，6 号和 7 号是周末。

对于时间序列分析而言，若以天为粒度，则数据不连续，因此需要将缺失的天进行填补。通过对股市调查，我们了解到，若是股市当日无成交，则以前一天的收盘价作为当日收盘价。因此，在填补数据时，若遇到缺失的天，则以前一天的数据作为填补内容。同时，离待预测数据时间越近的点，价值越高，故选取 2017 年以来的数据作为分析样本。

对于随机森林算法而言，不要求数据存在时间上的连续。经检验，数据没有缺失，因而使用该算法时不用再对数据进行额外的清理。故本次分析的流程为：将 2020 年的数据作为测试集，2017-2019 年的数据作为训练集，通过随机森林算法提取出重要变量，对提取后的数据集补充差值，对收盘价进行时间序列分析并做出预测（代码详见 stock.R 或 stock.txt）。

Data	
stock	861 obs. of 12 variables
stock_data	4581 obs. of 15 variables
stock.test	5 obs. of 12 variables
stock.train	856 obs. of 12 variables

图 3 分离训练集与测试集

3.3. 随机森林探索分析

随机森林算法需要加载包 randomForest。本次算法将 2017 年至 2019 年的数据作为训练集，2020 年的数据（截止到 7 月 15 日）作为测试集，构建随机森林

模型。生成各变量的重要性如下图所示：

```
> importance(stock.forest)
          %IncMSE IncNodePurity
最高价   15.683763    8170281.45
最低价   16.209242    8882424.99
开盘价   12.833969    6828773.41
前收盘   12.814091    6418435.80
涨跌额    2.511047      48710.20
涨跌幅    3.342220      12769.38
换手率    3.454626      85749.98
成交量    2.917633     106171.96
成交金额  4.283364    1210212.49
流通市值 18.184689    9117061.30
```

图 4 各变量的重要性

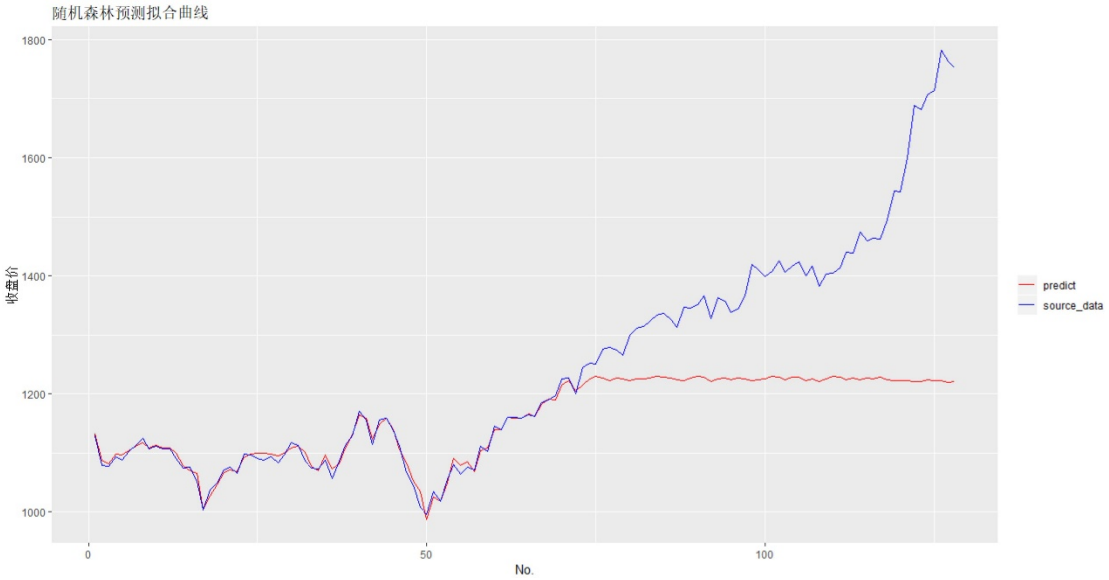

图 5 随机森林预测拟合曲线

由该输出结果可知，重要程度最高的变量降序排列应为流通市值、最低价、最高价、开盘价、前收盘五个变量，变量的权重可近似参考字段在上述重要度衡量指标中所占的百分比。

使用测试集对该模型进行验证，得到结果如图 5（随机森林拟合曲线）所示。由图 5 可知，该随机森林模型对前 75 天的数据预测较为准确，而对于 75 天以后的数据预测误差很大。由此可知，在 2020 年二月中旬后，其他变量对于收盘价的重要程度关系发生了较大变化，以至于基于前三年数据训练出的模型失效。

为了探寻更准确的模型，我们将测试集的数据归入训练集，以求训练出更符合当前情形的模型，最终得到变量的重要程度如下图 6 所示。由图 6 可知，筛选出的最重要的变量和上一模型一致，但各变量的相对重要性都有 1-2 个百分点的浮动，且节点纯度提升了，其中流通市值的相对重要性和节点纯度都提升极大。

综上，对收盘价影响最大的变量降序排列及其相对重要程度分别为：流通市值（20%），最高价（16%），最低价（14%），开盘价（12%），前收盘（10%）。

```
> importance(stock.forest.new)
           %IncMSE IncNodePurity
最高价    16.801357    18386301.34
最低价    14.369583    14065582.15
开盘价    12.566352    11218564.00
前收盘    10.018522     7381882.58
涨跌额     1.322822      179587.47
涨跌幅     3.800295       29293.44
换手率     4.579708      116123.70
成交量     4.114986      166690.44
成交金额   4.287694     1931510.30
流通市值  20.038280    23816889.29
```

图 6 变量重要程度（new）

3.4. 时间序列分析预测

在时间序列分析预测方面，我们进行了数据补齐、平稳性判断、消除时间序列不平稳性等多个部分的探索和处理，并在后文通过建立 ARIMA 模型预测茅台股票未来五天的走势。

3.4.1. 数据补齐

由于股市节假日休市，所以该数据以天为粒度时是不连续的。为此，在进行时间序列分析之前，需将其补齐。本次探究中数据补齐的算法如下：使用 seq 函数生成完整的日期序列，将该数据与收盘价序列全连接，生成新表。倒序扫描新表，若遇空值则用前一天数据补齐。由于时间序列进行短期预测效果最佳，且股市受外界因素影响较大很难长时间使用同一个模型，故留最新的 5 条数据作为验证，其余数据都作为训练集。

```
> all_days<-seq(as.Date("2017/01/03"),as.Date("2020/07/17"),by="days")
> all_days<-data.frame("日期"=all_days)
> all_days[,1]<-as.character(all_days[,1])
> stock_all_time<-merge(x=all_days,y=stock_time,by.x = "日期",by.y = "日期",
all = TRUE)
> for(i in 1:1292){
+   if(is.na(stock_all_time$收盘价[i]))stock_all_time[i,c(2:6)]<-stock_all_
time[i-1,c(2:6)]
+ }
```

Data	
all_days	1292 obs. of 1 variable
pre_data	5 obs. of 3 variables
stock	861 obs. of 12 variables
stock_all_time	2153 obs. of 6 variables
stock_data	4581 obs. of 15 variables
stock_time	861 obs. of 6 variables
stock.forest	Large randomForest.formula (18 elements, 9.2 Mb)
stock.forest.new	Large randomForest.formula (18 elements, 9.2 Mb)
stock.test	5 obs. of 12 variables
stock.train	856 obs. of 12 variables
Values	
i	1292L
stock.pred	Named num [1:5] 1587 1604 1631 1580 1580

图 7 数据补齐

3.4.2. 平稳性判断

拿到一个时间序列，在对时间序列数据的预处理和分析之后，需要确定一个时间序列的平稳性。所谓平稳，是指统计值在一个常数上下波动并且波动范围是有界限的。如果有明显的趋势或者周期性，那么就是不稳定的。只有非白噪声的稳定性时间序列才有分析的意义以及预测数据的价值。在确保时间序列平稳性的基础上，后续的分析和建模才能顺利展开。

对时间序列平稳性的判断，一般有三种方法：

1) 画出时间序列的趋势图，看趋势判断
2) 画自相关图和偏相关图，平稳时间序列的自相关图和偏相关图，要么拖尾，要么截尾。
3) 检验序列中是否存在单位根，如果存在单位根，就是非平稳时间序列。

对数据集进行时间序列的绘制，以观察数据的总体趋势，时间序列图如下图所示：

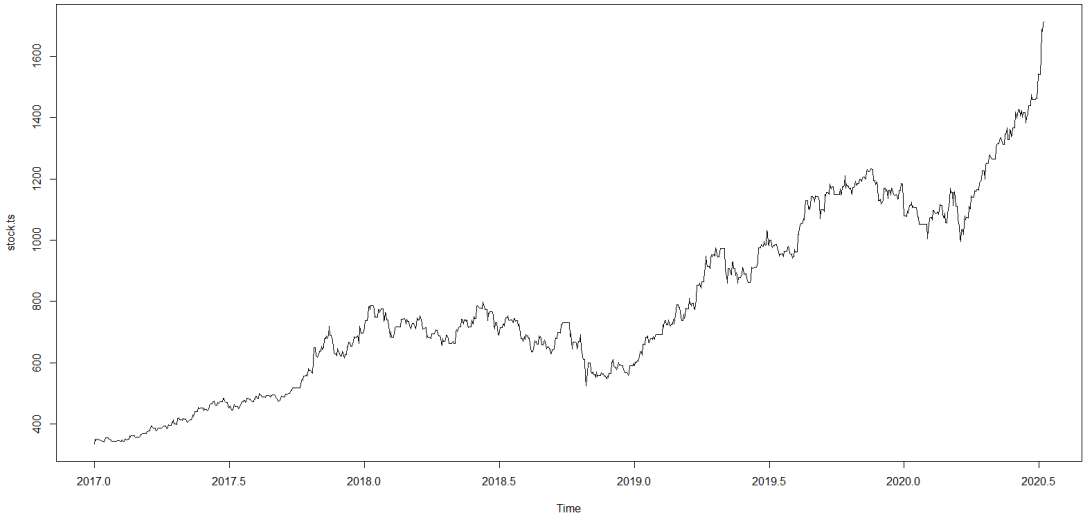

图 8 时间序列图

增加线性拟合曲线，lm 函数中“~”表示左边为因变量，右边为自变量，得到时间序列图及其线性拟合曲线。通过这条拟合曲线可以大致看出序列的趋势性，拟合直线如下图所示：

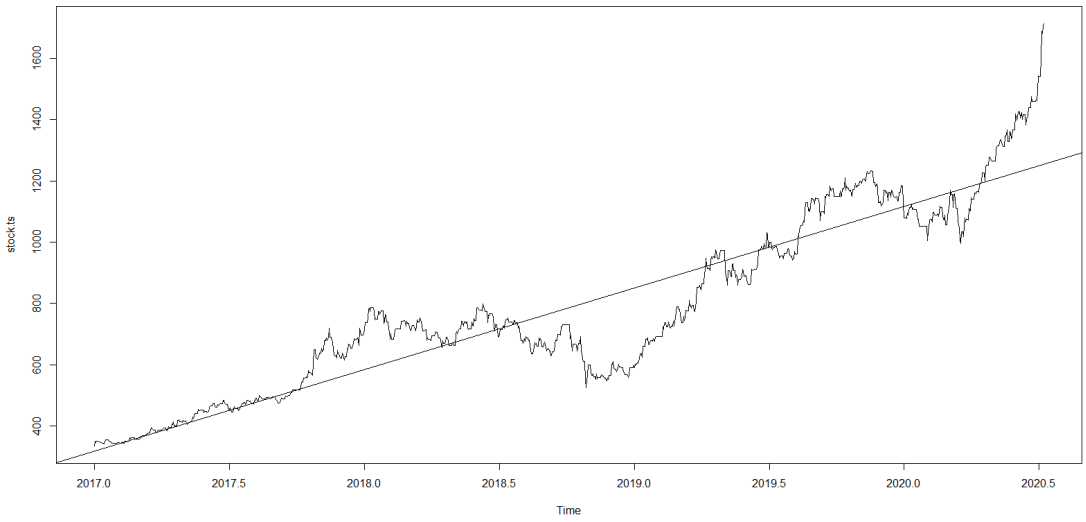

图 9 时间序列图及其线性拟合曲线

再绘制以年为单位的时间序列图，结果如下图所示：

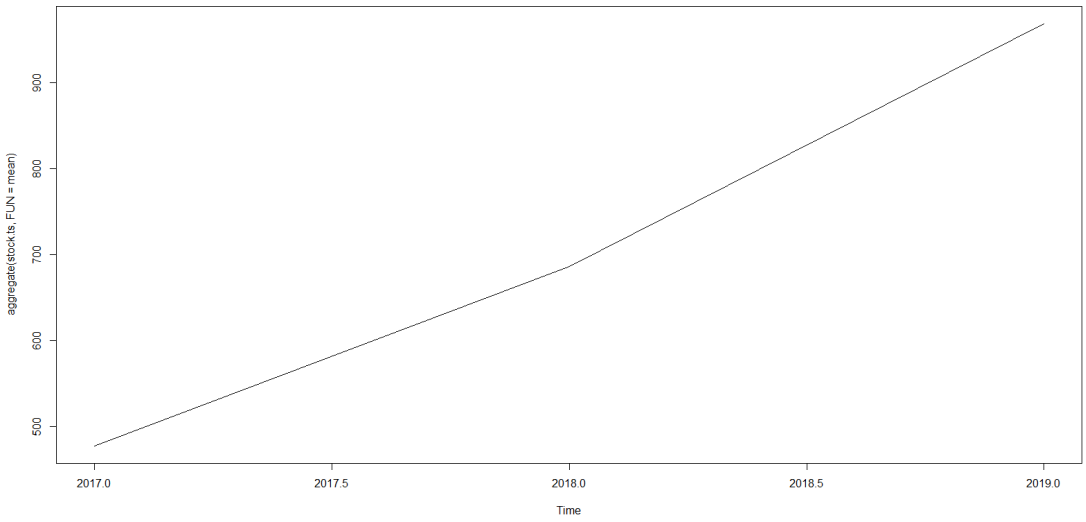

图 10 以年为单位的时间序列图

也可以使用 R 的用 ACF 和 PACF 指令分别画出自相关函数和偏自相关函数图，结果如下图所示：

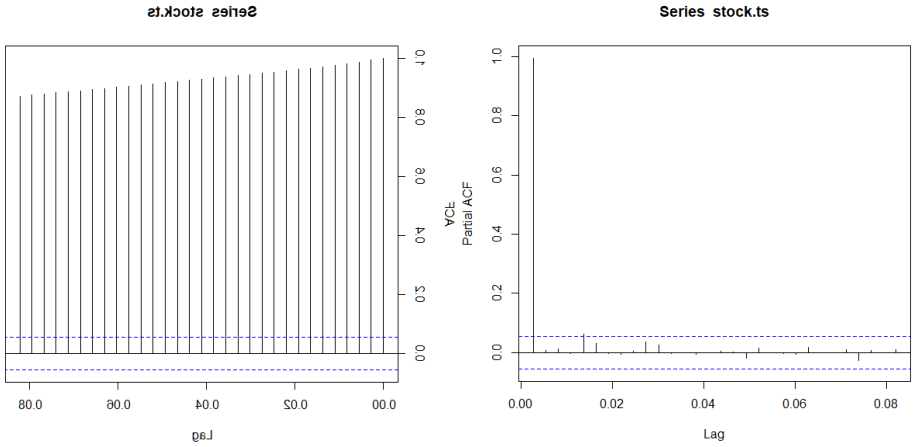

图 11　ACF、PACF 偏相关图

对于时间序列，我们通过简单的绘图及对比，观察图 8、图 9、图 10、图 11，可以得出以下结论：

1) 从以日为单位和以年为单位的时间序列图可以看出，随着时间轴的延伸，收盘价呈上升趋势，数据的方差也逐渐增大。
2) 从以日为单位时间序列图可以看出，每 12 个月为一个周期。此外，在每 12 个月的周期内部，收盘价的变化也呈现出一定的规律性，即每年都会有一个明显的收盘价峰值，在年底时收盘价会呈下降趋势。
3) 根据 ACF、PACF 相关图，可以看出该时间序列不是平稳的，因为其自相关函数和偏相关函数并未呈典型的指数衰减，不能断定其是截尾的还是拖尾的。

总体而言，该数据满足周期性和趋势性的特点，是一种典型的非平稳数据。需要消除其不平稳性，使得整体的序列呈现平稳性，便于进行后续的研究和分析。因此，需要对原始数据进行差分或其他处理。

3.4.3. 消除时间序列不平稳性

平稳时间序列是时间序列分析中最重要的特殊类型，到目前为止，时间序列分析基本上是以平稳时间序列为基础的。对于非平稳时间序列的统计分析，其方法和理论都很有局限性，需要消除其不平稳性，使时间序列具有分析的意义以及预测未来数据的价值。

消除时间序列不平稳性的方法有以下几种：

1) 对时间序列取对数log，消除方差的变化；
2) 对时间序列求一阶差分，消除时间序列中的趋势；
3) 用季节差分，去除时间序列的季节性趋势。

本文使用 ndiff 函数来尝试消除原序列的不平稳性。时间序列 Z_t的一阶差分就是相邻两个观测的差：$\mathrm{diff}(Z) = Z_t - Z_{t-1}$，对时间序列进行差分可以去除线性趋势，R 中diff()函数即计算时间序列的一阶差分。对时间序列进行差分，去除其

线性趋势，结果如下图所示：

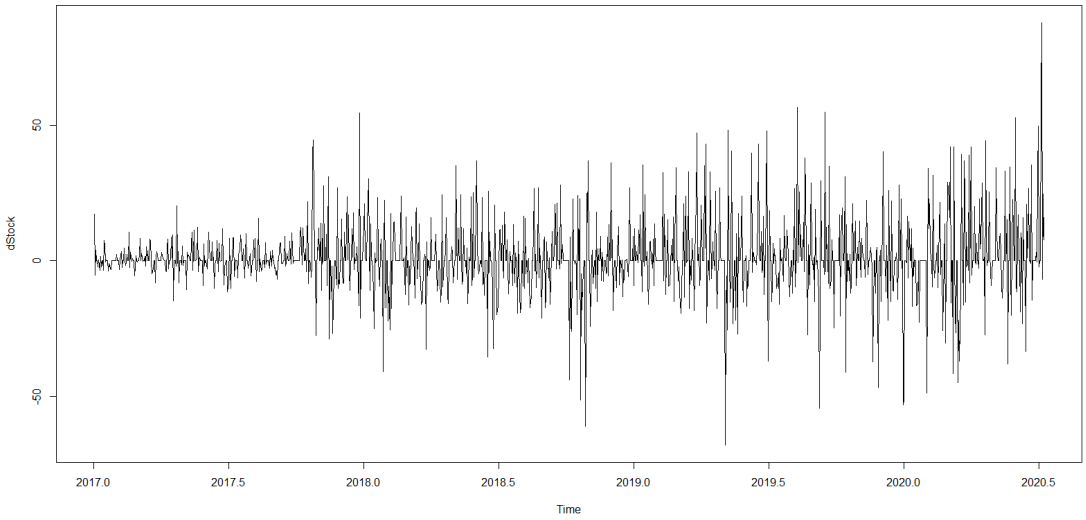

图 12 差分后的时间序列

对一阶差分后的序列使用 ACF 和 PACF 指令，绘制出该序列的自相关函数和偏自相关函数的相关图，结果如下图所示：

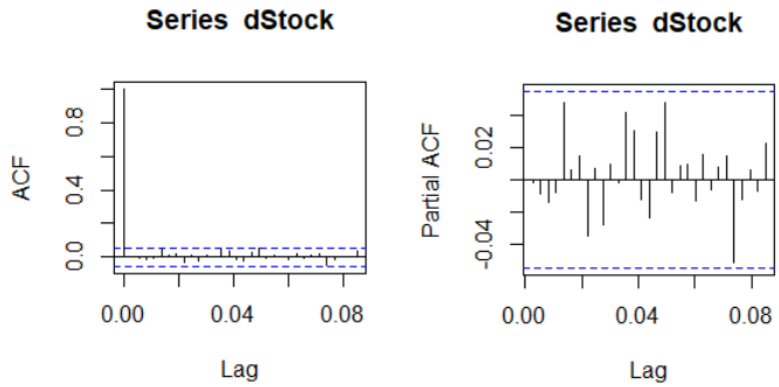

图 13 ACF、PACF 偏相关图

观察图 12、图 13 可知，一阶差分后的时序图基本围绕常数“0”振动，说明时间序列基本平稳。根据 ACF、PACF 相关图，ACF 在 0 阶以后迅速减 0，PACF 始终在接近 0 的范围内波动。进一步使用 adf 检验平稳性，判断是否存在单位根，若存在单位根，则该序列不平稳。结果显示，单位根检验通过（p<0.05，显著拒绝存在单位根），说明该时间序列平稳。

```
        Augmented Dickey-Fuller Test

data:  dStock
Dickey-Fuller = -10.27, Lag order = 10,
p-value = 0.01
alternative hypothesis: stationary
```

图 14 单位根检验

3.5. ARIMA 模型建立

3.5.1. 模型的定阶

一般来说，消除数据的不平稳性后，就可以看出大部分序列的衰减，因而可以大致确定模型的阶数。根据上文中的差分阶数、ACF 与 PACF 图，最终确定使用模型为 ARIMA(0,1,0)。

模型	ACF	PACF
AR(p)	衰减趋于零（几何型或振荡型）	p阶后截尾
MA(q)	q阶后截尾	衰减趋于零（几何型或振荡型）
ARMA(p，q)	q阶后衰减趋于零（几何型或振荡型）	p阶后衰减趋于零（几何型或振荡型）

图 15　Arima 模型定阶判定

3.5.2. 模型的拟合优度检验

使用 tsdiag 检验模型，观察模型的结果，结果如图 16 所示：

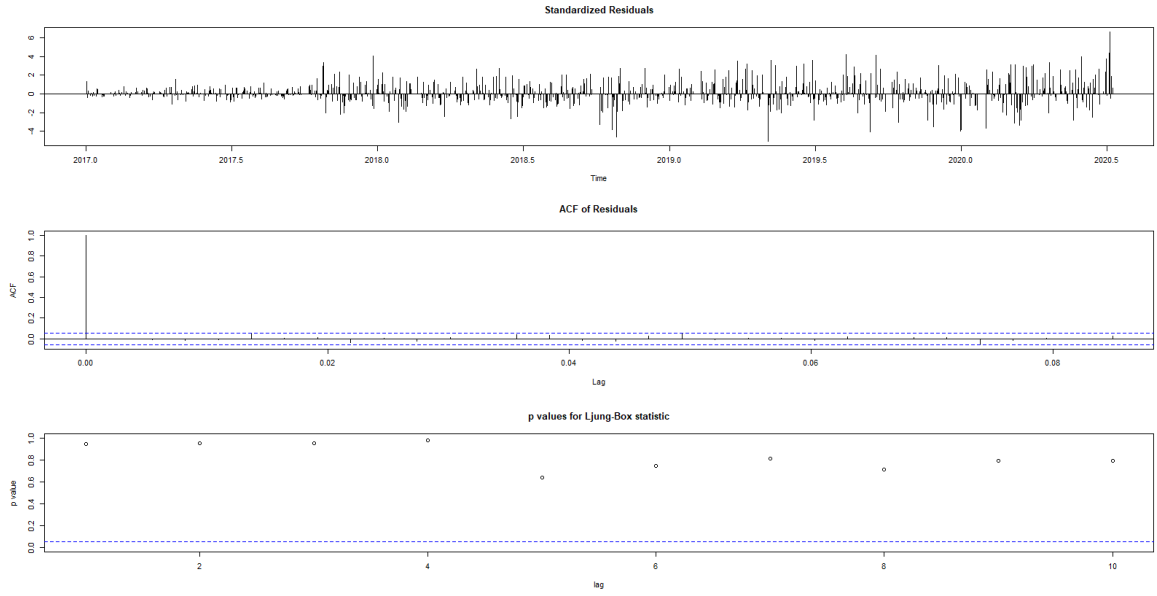

图 16　tsdiag 检验 ARIMA(0,1,0)

还可以用 Ljung-Box 检验模型，结果如图 18 所示：

```
> Box.test(arima_fit$residuals,type = "Ljung-Box")

        Box-Ljung test

data:  arima_fit$residuals
X-squared = 0.0039144, df = 1, p-value = 0.9501
```

图 17　Ljung-Box 检验 ARIMA(0,1,0)

综上，检验结果表明：

1) 模型残差、标准差在[-1,1]之间；
2) Acf 相关图说明，残差的自回归为 0（两虚线内），没有明显的自相关性；
3) Ljung-Box 检验的 p 值在 0.05 之上，说明残差大致符合白噪声特征。

综上，模型合格，且模型效果不错。

3.6. 应用模型对未来进行预测

在 arima_fit 模型中，预测值表示为由最近的真实值和最近的预测误差组成的线性函数。接下来，尝试应用 arima_fit 模型对未来五天的收盘价进行预测，结果如下：

```
> forecast(arima_fit,5)
          Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
2020.5205        1713.85 1696.730 1730.970 1687.668 1740.032
2020.5233        1713.85 1689.639 1738.061 1676.823 1750.877
2020.5260        1713.85 1684.198 1743.502 1668.501 1759.199
2020.5288        1713.85 1679.611 1748.089 1661.486 1766.214
2020.5315        1713.85 1675.569 1752.131 1655.305 1772.395
```

图 18　arima_fit 预测

对未来五天的收盘价进行预测图绘制，结果如下：

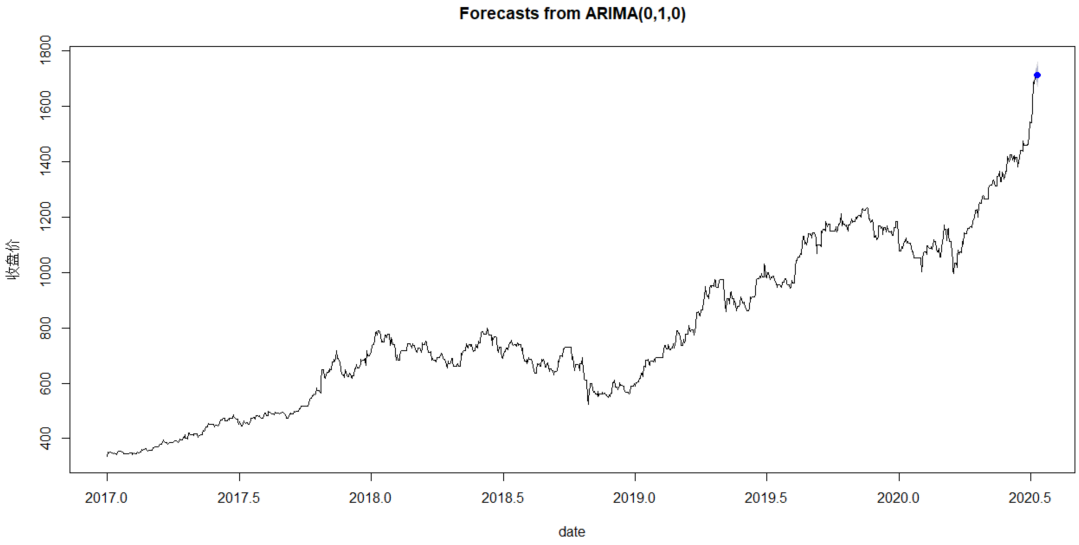

图 19　arima_fit 模型预测图

观察预测图并结合实际数据进行分析，我们发现预测数据较为贴合实际值。在 95%的置信区间内完全包含预测值，在 80%的置信区间与预测值较为接近，说明拟合效果和预测效果都很好。

此外，由预测结果可知，结合最近数据与误差分析的该时间序列模型所预测值为平均值，只能代表短时间内的大致趋势，而无法精确预测数值的涨落，若要进行较准确的预测，还需结合收盘价以外的其他变量，根据时间序列预测

的置信区间来进行判断。

四、总结

本次预测采用了随机森林与时间序列分析两种方法，分别从横向和纵向对股票数据集进行探索。在分析和预测过程中，得出以下结论：

1) 随着时间轴的延伸，茅台股票的收盘价呈上升趋势，数据的方差也逐渐增大，并且每年都会有一个明显的收盘价峰值，在年底时收盘价会呈下降趋势。
2) 对数据集进行一定的划分，可以使得训练和验证能够独立进行，能够检验模型的泛化能力，使得整个预测能够达到比较好的效果。
3) 从横向上，通过建立随机森林模型，得到了流通市值（20%），最高价（16%），最低价（14%），开盘价（12%），前收盘（10%）五个字段对当天的收盘价的影响程度最高，括号中的值代表变量的相对重要性。
4) 从纵向上，借助时间序列分析，预测出未来五天的收盘价会在 1713.85 左右波动，其中 95%的置信区间包含了所有真实值，结合股票递增的走势，可以发现 80%置信取间虽然未包含所有预测值，但其上界与预测值较为接近。

通过本次对茅台股票的分析和预测，我们发现，对股票的未来值进行预测其实是一件非常困难的事情。数据集中的流通市值、最高价等等字段虽然可以对收盘价造成较大影响，却不是最直接的影响来源。股票的走势不仅和数据集中的各字段相关，更和各种各样的政策、经济、社会因素密不可分，仅仅依靠数据集对股票进行预测是无法反映这些外在因素的，例如社会负面舆论会直接导致股票下跌，而数据是无法对此进行预测的。

因此，基于数据挖掘对股票进行预测可以增加一些预测依据，但不能精确预测股市数据和股票涨跌。若要进行预测，还需要关注与该股票相关的政治、经济、社会效应等多种因素，在结合数据的基础上综合分析，才能得出更为可信的结论。