

订阅DeepL Pro以编辑此演示文稿。  
访问[www.DeepL.com/pro](https://www.deepl.com/pro?cta=edit-document)，了解更多信息。

**RMDS 2021年第二季度数据科学竞赛**

后科维德时期的加州房价预测分析

**穆罕默德-赛米尔-伊克尔-阿卜杜勒-巴尔-尤吉-安格拉**

**Ergidya Liviana Riyanna Shabrina**

**零点团队**

# 简介

Covid-19是自2019年底开始的冠状病毒疾病的持续大流行。 该疾病是一种严重急性呼吸道综合征（因此被命名为SARS-CoV 2）。该疫情于2019年12月在中国武汉省首次发现。 这次大流行停止了社会和经济活动的许多方面，因为对身体活动有限制。 一个季度后，该大流行病开始显示其严重的影响。对经济的干扰甚至造成了全球经济衰退，大多数国家出现了经济负增长。

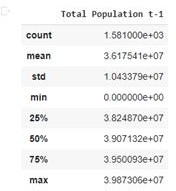
随着经济活动的减少，价格水平也受到影响，包括在房地产行业。 在大多数情况下，由于商品和服务的需求急剧下降，通货紧缩发生在2020年的第一和/或第二季度。尽管如此，随着房地产价格的积极或增加的趋势，特别是在像加州这样的密集和高需求的地方，大流行的影响可能会被消除，或至少与其他方面相比没有那么强大。 关于这个事实，作者用房价指数作为代理目标变量，对大流行如何影响加州的房价进行了研究。

# 变量和数据集来源

|  |  |
| --- | --- |
| **变化的** | **数据来源和描述** |
| 房价指数 | 基于网络的加州刑事司法统计数据的展示。来源于CSV格式的[https://openjustice.doj.ca.gov/exploration/crime- 统计数据。](https://openjustice.doj.ca.gov/exploration/crime-statistics) |
| 美国抵押贷款利率 | 美国30年固定利率抵押贷款的平均利率。来源于CSV格式的<https://fred.stlouisfed.org/series/MORTGAGE30US> |
| Covid-19 感染率 | Covid-19感染率数据由洛杉矶时报收集。这些数据是实时数据，以CSV格式存储在[https://github.com/datadesk/california-coronavirus- 数据中。](https://github.com/datadesk/california-coronavirus-data) |
| 家庭收入中位数 | 家庭收入中位数是指某一家庭所赚取的收入水平，在感兴趣的地理区域内，有一半的家庭收入较高，一半的家庭收入较低。来源于[加州](https://fred.stlouisfed.org/series/MORTGAGE30US) |

|  |  |
| --- | --- |
|  | [国家家庭收入 | 数字部 (deptofnumbers.com)](https://www.deptofnumbers.com/income/california/) |
| 家庭收入中位数 | 家庭收入中位数是衡量家庭满足食物、衣服、住房、保健费用护理、交通、儿童护理和高等教育的能力 来源于[加州家庭收入|](https://fred.stlouisfed.org/series/MORTGAGE30US)[数字部（deptofnumbers.com）。](https://www.deptofnumbers.com/income/california/) |
| 消费者信心指数 | 这个消费者信心指标提供了一个家庭消费和储蓄的未来发展的指示，基于他们对预期财务状况的回答，他们对总体经济形势的情绪，失业和储蓄能力。 该指标高于100，表明消费者对未来经济形势的信心增强，因此，他们不太愿意储蓄，而更倾向于在未来12个月内花钱购买大件物品。 低于100的数值表示对未来经济发展的悲观态度，可能会导致更多的储蓄和更少的消费。来源于经合组织，以CSV格式提供。 |
| 消费者物价指数 | 这个变量显示从2017年到2021年的CPI数据。来源于[局长办公室-研究部门。](https://www.dir.ca.gov/OPRL/capriceindex.htm)CSV格式的[加州消费者价格指数](https://www.dir.ca.gov/OPRL/capriceindex.htm) |
| 罪行 | 加州城市的犯罪数据集。这些数据经过汇总，形成了加州的犯罪情况。来源于[https://ucr.fbi.gov/crime-in-the-u.s/](https://ucr.fbi.gov/crime-in-the-u.s/2019/crime-in-the-u.s.-2019/tables/table-8/table-8-state-cuts/california.xls)的CSV格式。 |
| 人口 | 居住在一个国家、城市或任何区或地区的人的总数。来源于CSV格式的[搜索结果](https://www.census.gov/content/census/en/search-results.html?stateGeo=none&q=california%20population%20ca&searchtype=web&page=1) |

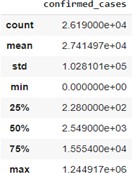
1. 解释型数据和变量分析
   1. 数据描述和基本统计分析
      * 人口



解释。

通过使用.describe()函数，我们得到2017年至2021年4月人口数据的平均值（3.617541e+07）、标准差（1.043379e+07）、最小值（0.00）和最大值（3.987306e+07）的值。

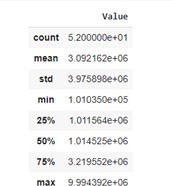
* + - Covid-19 感染率



解释。

根据输出结果，我们可以得出结论，平均值为2.741497e+04，标准差为1.028101e+05，最小值为0.00，最大值为1.244917e+06。

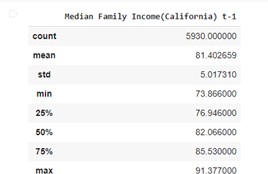
* + - CCI（消费者信心指数）数据



解释。

根据使用.describe()函数后的输出，我们可以得出结论：平均值为3.092162e+06，标准差为3.975898e+06，最小值为1.010350e+05，最大值为9.994392e+06。

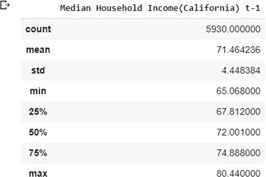
* + - 家庭收入中位数



解释。

通过使用.describe()函数，我们得到家庭收入中位数数据的平均值（81.402659）、标准差（5.017310）、最小值（73.866000）和最大值（91.377000）。

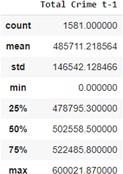
* + - 家庭收入中位数



解释。

根据使用.describe()函数后的输出，我们得到家庭收入中位数数据的平均值（71.464236）、标准差（4.448384）、最小值（65.068000）和最大值（80.440000）。

* + - 罪行



解释。

通过使用.describe()函数，在输出中我们得到了平均值（485711.218564），标准差（146542.128466），最小值（0.00），和

最大（600021.870000）。

* + - 房价指数



解释。

根据输出结果，我们可以得出结论：平均值为648.835294，标准差为37.389765，最小值为581.120000，最大值为718.340000。

1. 方差分析和Tukey HSD
   * 方差分析CASTHPI 平古因试验

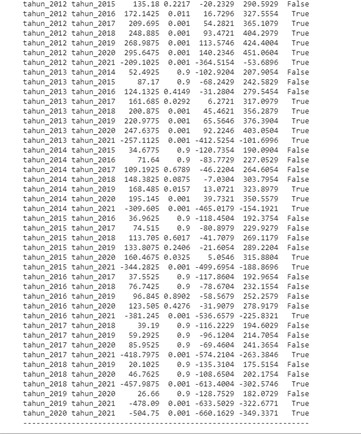
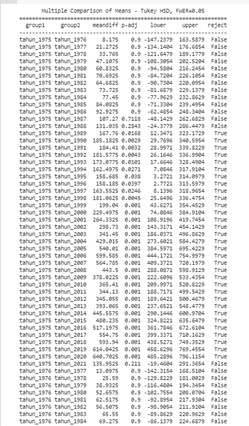
OLS法



解释。

因为P值小于α，所以结论是拒绝H0，这样就可以得出结论，CASTHPI几年的平均影响存在差异。

Tukey测试

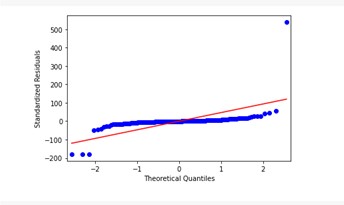


解释。

使用Tukey HSD来检查组间差异，显示存在统计学上的显著差异。我们可以看到，在大多数情况下，2015-2021、2016-2021、2017-2021、2018-2021、2019-2021和2021-2021表明

2021年，CASTHPI的平均年份将始终不同。

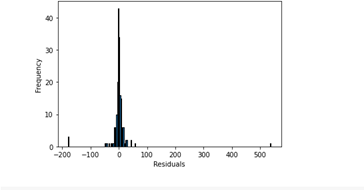
QQ计划



解释。

因为这些点都位于线性线周围或沿着对角线，所以可以得出结论，残值是正态分布。

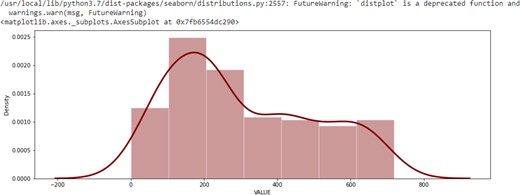
柱状图



解释。

在直方图中，可以看到它没有形成一个完美的钟形，所以可以得出结论，数据不是正态分布。

检查测试数据的正态性



解释。

从上面的密度图可以看出，数据并没有完全形成一个钟形，所以有可能是假设这里的测试数据不是正态分布。为了确定这一点，应该用Kolmogorov Smirnov Uji Test进一步检验

Kolmogorov Smirnov



解释。

从KS检验可以得出结论，残差数据不是正态分布，或拒绝H0，因为P值（0.000）小于α（0.05）。

莱文测试



解释。

为了检查来自人口的样本是否是同质的，由于数据不是正态分布，所以使用了Levene检验，结论是未能拒绝H0或来自人口的样本是同质的。

* + - 方差分析 群体 平古因



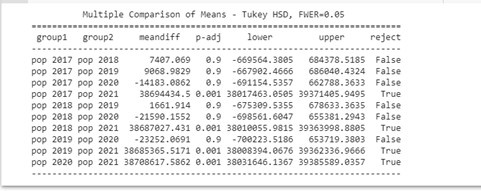
OLS法



解释。

因为p值小于α（0.05），所以结论是拒绝H0，所以可以得出结论，每年的人口平均数是有差异的。

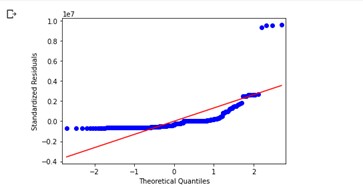
Tukey检验



解释。

使用Tukey HSD来检查组间的差异，显示有统计学上的显著差异。 我们可以看到在大多数情况下，对于2017-2021年的人口，2018-2021年的人口，2019-2021年的人口，2020-2021年的人口表明，2021年的人口平均值总是不同的。

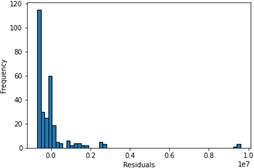
QQ计划



解释。

因为这些点大多不在直线周围，所以可以得出结论，残值不是正态分布。

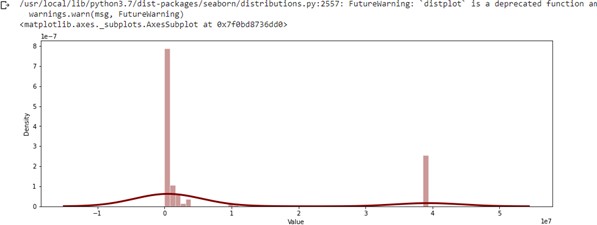
柱状图



解释。

在直方图中，可以看到它没有形成一个完美的钟形，所以可以得出结论，数据不是正态分布。

检查测试数据的正态性



解释。

从上面的密度图可以看出，数据并没有完全形成一个钟形，所以有可能是假设这里的测试数据不是正态分布。为了确定这一点，应该用Kolmogorov Smirnov Uji Test进一步检验

Kolmogorov Smirnov



解释。

从KS检验可以得出结论，残差数据不是正态分布或拒绝H0，因为P值（0.000）小于alpha（0.05）。

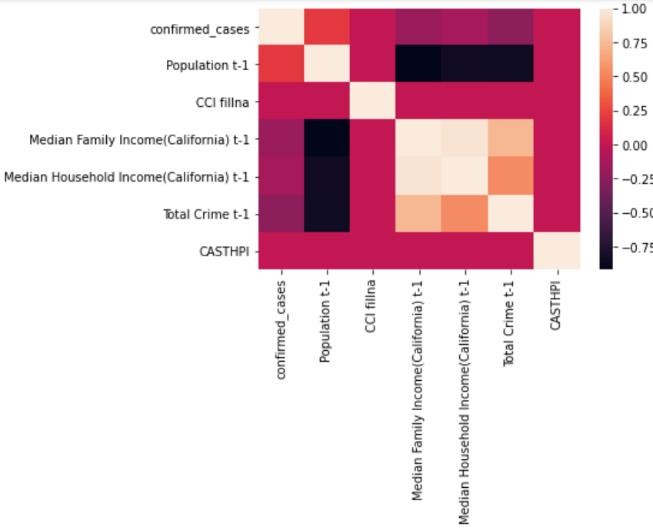
李文检验



解释。

为了检查来自人口的样本是否是同质的，使用Levene检验，因为数据不是正态分布，结论是拒绝H0或来自人口的样本是异质的。

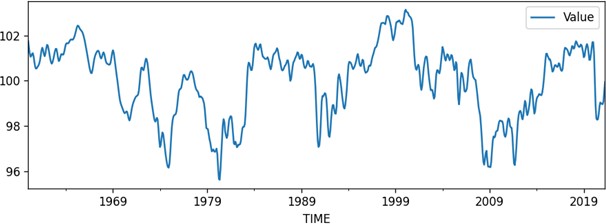
1. 矩阵关联性

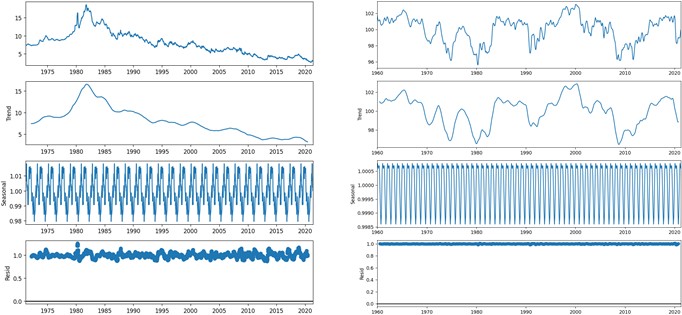


# 时间序列分解

在我们进行时间序列预测之前，我们必须明白，根据文献回顾，有一些变量对房价有很大的影响，在这里我们用HPI作为房价的代表。在之前的文献中经常提到的变量是消费者信心指数，作为消费者预期的代表，以及抵押贷款利率，作为直接影响房屋购买可行性的因素。

在此，作者将这两个变量分解为时间序列分解，以洞察和直观地了解这两个变量如何移动并可能影响房价指数。





根据上图，我们可以清楚地看到，数据中存在着季节和趋势，可以用低残差水平进行分解。然而，由于季节和趋势都不同，我们需要进一步分析，包括房价指数和Covid-19价差本身在内的其他变量，作为本研究的主要议题。

# 时间序列分析

* 1. ARIMA

[ARIMA模型](https://en.wikipedia.org/wiki/Autoregressive_integrated_moving_average)是一类用于分析和预测时间序列数据的统计模型。在统计学和计量经济学中，特别是在时间序列分析中，自回归整合移动平均模型（ARIMA）是自回归移动平均模型（ARMA）的一个概括。它明确地迎合了时间序列数据中的一系列标准结构，并因此提供了一种简单而强大的方法来进行熟练的时间序列预测。

ARIMA是AutoRegressive Integrated Moving Average的缩写，意思是自动回归整合移动平均线。它是对更简单的自动回归移动平均法的概括，并增加了整合的概念。

这个缩写是描述性的，抓住了模型本身的关键方面。简而言之，它们是

* **AR**：*自回归*。一个模型，使用一个观察值和一些滞后观察值𝑌𝑡之间的依赖关系，只取决于它自己的滞后期。 也就是说，𝑌𝑡是"𝑌𝑡的滞后期 "的一个函数。

𝑌𝑡 = 𝑎 + 𝑏1𝑌𝑡- 1 + 𝑏2𝑌𝑡 -2 + ⋯ + 𝑏𝑝 𝑌𝑡- ⋯ + 𝑝

* **I**：*综合*。使用原始观测值的差分（如用前一个时间步长的观测值减去一个观测值），以使时间序列静止。
* **MA**：*移动平均*。一种模型，使用观察值和应用于滞后观察值的移动平均模型的残余误差之间的依赖关系。𝑌𝑡只取决于滞后的预测误差。

𝑌𝑡 = 𝑎 + 𝑒𝑡 + 𝜃1 𝑒𝑡-1 + 𝜃2 𝑒𝑡-2 + ⋯+ 𝜃𝑞 𝑒𝑡-𝑞

ARIMA模型是指时间序列至少被差分一次以使其静止，你将AR和MA项结合起来。因此，该方程成为

𝑌𝑡 = 𝑎 + 𝑏1𝑌𝑡- 1 + 𝑏2𝑌𝑡- 2 + ⋯ + 𝑏𝑝 𝑌𝑡- ⋯ + 𝑒𝑡 + 𝜃1𝑒𝑡-1 + 𝜃2𝑒𝑡-2 + ⋯ + 𝜃𝑞 𝑒𝑡-𝑞

或

𝑌𝑡 = 𝑒𝑙𝑒𝑚𝑒𝑛𝑡𝑠 𝑜𝑓 𝑝 + 𝑒𝑙𝑒𝑚𝑒𝑛𝑡𝑠 𝑜 𝑞

因此，该模型被称为ARIMA，参数为p、d、q。此外，这些成分中的每一个都在模型中被明确指定为参数。一个标准的符号是ARIMA(p,d,q)，其中参数用整数值代替，以快速表示正在使用的特定ARIMA模型。

ARIMA模型的参数定义如下。

* **p**:模型中包含的滞后观测值的数量，也称为滞后顺序。
* **d**:原始观测值被差分的次数，也叫差分程度。
* **q**：移动平均线窗口的大小，也叫移动平均线的顺序。

换句话说，ARIMA是一个线性回归模型，包括指定的项的数量和类型，并通过一定程度的差分来准备数据，以使其静止，即去除对回归模型有负面影响的趋势和季节性结构。

𝑌𝑡 = 常数 + 𝑌的滞后线性组合（滞后至𝑝） + 滞后预测误差的线性组合（滞后至𝑞）。

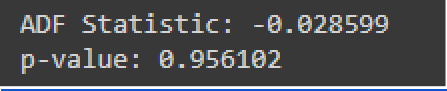
一个参数可以使用0的值，表示不使用该模型的元素。这样，ARIMA模型就可以被配置为执行ARMA模型的功能，甚至是一个简单的AR、I或MA模型。ARIMA(p,d,q)模型是一个关于回归的模型，包括ARIMA(p,0,0)和AR(p)，ARIMA(0,0,q)和MA(p)以及ARIMA(p,0,q)和ARMA（k,p）。

在这里，我们通过调整数据差分的参数来创建一个模型，以获得ARIMA模型中的适当参数。那么，如何确定正确的差分顺序？

正确的差分顺序是得到一个近乎稳定的序列所需的最小差分，该序列围绕一个确定的平均值游走，ACF图相当快地达到零。

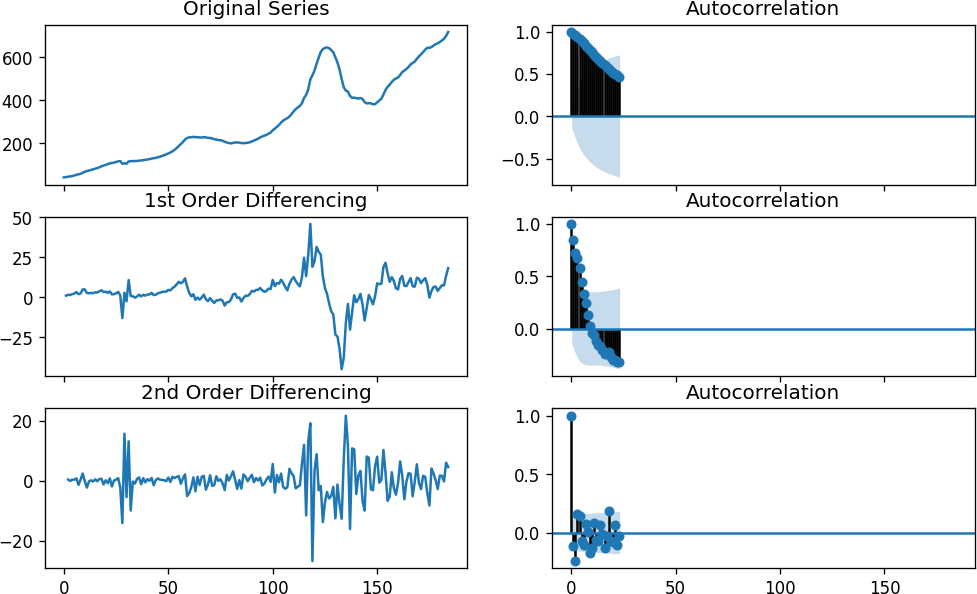
如果自相关在许多滞后期（10个或更多）都是正的，那么这个系列需要进一步差分。另一方面，如果滞后1期的自相关本身太负，那么该系列可能被过度差分了。在这种情况下，你无法在两级差分之间做出决定，那么就选择差分序列中标准差最小的那一级。

首先，我们需要用Augmented Dickey-Fuller检验来检查序列是否是静止的



由于P值大于显著性水平，让我们对该系列进行差分，看看发现（d）部分的自相关图是怎样的

* 1. 调查结果 (d)



通过上述可视化，作者得出结论，二阶差分是最好的（d）参数值，因为自相关变得稳定。否则，如果我们再进一步，可能会出现过度微分的情况

* 1. 发现 (p)

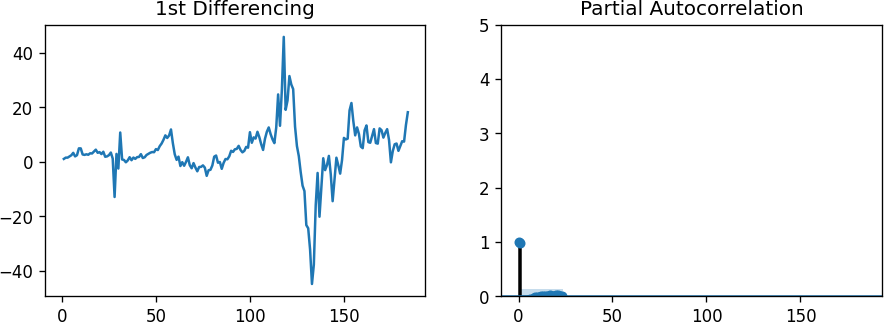
下一步是确定该模型是否需要任何AR项。 你可以通过检查部分自相关（PACF）图来找出所需的AR项的数量。 部分自相关可以被想象为在排除了中间滞后期的贡献后，序列和其滞后期之间的相关关系。因此，PACF有点像传达了纯粹的相关关系

滞后期和系列之间的关系。这样，你就会知道在AR项中是否需要该滞后期。

𝑌𝑡 = 𝑎0 + 𝑎1𝑌𝑡- 1 + 𝑎2𝑌𝑡- 2 + 𝑎3𝑌𝑡- 3 ...

一个序列的滞后期（k）的部分自相关是𝑌的自回归方程中该滞后期的系数。因此，假设𝑌𝑡是当前序列，𝑌𝑡-1是𝑌的滞后1，那么滞后3的部分自相关（𝑌𝑡-3 ）就是上述方程中𝑌𝑡-3的系数α。静止化序列中的任何自相关都可以通过增加足够多的AR项来纠正。 因此，我们最初认为 AR 项的阶数等于越过显著性限制的滞后期的数量

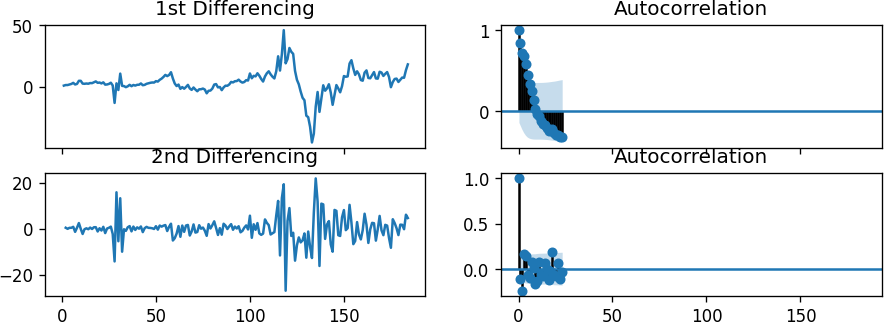
在PACF图中



据观察，PACF滞后1是相当显著的，因为它远远高于显著线。然而，这仍然是暂定的，因为参数还没有在模型中得到测试。

* 1. 发现(q)

就像我们看PACF图看AR项的数量一样，你可以看ACF图看MA项的数量。技术上讲，MA项是滞后预测的误差。ACF告诉我们需要多少个MA项来消除固定化序列中的任何自相关。



在这里，我们发现（q）的参数是2。然而，它就像（p）一样，如果它不能提供一个合适的时间序列预测，我们就需要在以后进行调整。

* 1. 测试参数

- 阿里马(1, 1, 2)

ARIMA模型结果

==============================================================================

|  |  |  |  |
| --- | --- | --- | --- |
| 存款变量。 | D.CASTHPI | 号的意见。 | 184 |
| 模型。 | ARIMA(1, 1, 2) | Log Likelihood | -566.567 |
| 方法。 | css-mle | 创新的SDD | 5.242 |
| 日期。 | 2021年6月12日，星期六 | AIC | 1143.134 |
| 时间。 | 13:48:50 | BIC | 1159.208 |
| 样本。 | 1 | HQIC | 1149.649 |

===================================================================================

coefstd errzP>|z|[0.0250.975] 。

------------------ -- const4.08532.8371. 4400.152 -1.4769.646

ar.L1.D.CASTHPI0.89270.04022.4560.0000.8150.971

ma.L1.D.CASTHPI-0.02630.084-0.3140.754-0.1910.138

ma.L2.D.CASTHPI-0.15340.072-2.1190.035-0.295-0.012

根部

=============================================================================

实数模数频率

----------------- AR.11.1202+0.0000j1.12020.0000

MA.12.4685+0.0000j2.46850.0000

MA.2-2.6400+0.0000j2.64000.5000

-- ---------------

- arima (4, 2, 2)

ARIMA模型结果

==============================================================================

|  |  |  |  |
| --- | --- | --- | --- |
| 存款变量。 | D2.CASTHPI | 号的意见。 | 148 |
| 模型。 | ARIMA(4, 2, 2) | Log Likelihood | -457.544 |
| 方法。 | css-mle | 创新的SDD | 5.150 |
| 日期。 | 2021年6月12日，星期六 | AIC | 931.087 |
| 时间。 | 14:10:14 | BIC | 955.065 |
| 样本。 | 2 | HQIC | 940.829 |

====================================================================================

coefstd errzP>|z|[0.0250.975] 。

------------------ const0.01660. 0220. 7630.447-0.0260.059

ar.L1.D2.CASTHPI1.68820.08220.5750.0001.5271.849

ar.L2.D2.CASTHPI-0.76860.158-4.8520.000-1.079-0.458

ar.L3.D2.CASTHPI0.31600.1582.0060.0470.0070.625

ar.L4.D2.CASTHPI-0.26120.080-3.2600.001-0.418-0.104

ma.L1.D2.CASTHPI-2.00000.039-51.4910.000-2.076-1.924

ma.L2.D2.CASTHPI1.00000.03925.7750.0000.9241.076

根部

=============================================================================

实数模数频率

----------------- AR.11.0239-0.1320j1.0323-0.0204

AR.21.0239+0.1320j1.03230.0204

AR.3-0.4189-1.8485j1.8954-0.2855

AR.4-0.4189+1.8485j1.89540.2855

MA.11.0000-0.0000j1.0000-0.0000

MA.21.0000+0.0000j1.00000.0000

-- ---------------

在调整参数后，我们发现（4，2，2）在预测房价指数时有更高的准确性。这种准确性可以就模型的大多数系数是显著的事实得出结论。此外，大多数指标说，该模型足以预测未来。

|  |  |
| --- | --- |
| 度量衡 | 价值 |
| MAPE | 0.092 |
| MAE | 57.93 |
| RMSE | 78.59 |

从指标值中，我们可以看到平均绝对误差和平均平方根误差的高值。这意味着该模型有很高的误差值。然而，即使ARIMA（4,4,2）的MAE和RMSE都很高，模型仍然是准确的，用平均绝对误差百分比来检查。 通过将误差放大到百分比，我们得到的误差只有9%，或者换句话说，我们可以说ARIMA（4,4,2）的准确度大约为91%，因此我们可以使用该模型。

# 机器学习模型

* 1. 多重线性回归

OLS回归结果

==============================================================================

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 存款变量。 | 辽宁沈阳 | R-squared。 | | 0.000 | | | |
| 模型。 | OLS | Adj. R-squared: | | -0.000 | | | |
| 方法。 | 最小二乘法 | F-统计学。 | | 0.02228 | | | |
| 日期。 | 2021年6月13日，星期日 | 概率（F-统计学）。 | | 1.00 | | | |
| 时间。 | 11:13:01 | Log-Likelihood: | | -34274. | | | |
| 号的意见。 | 25222 | AIC。 | | 6 .856e+04 | | | |
| Df残值。 | 25215 | BIC: | | 6 .862e+04 | | | |
| Df模型。  协方差类型。 | 6  非稳健 |  | |  | | | |
| ==============================================================  coefstd err | | | | ================  t | =======  P>|t| | ============ [0.025 | ==========  0.975] |
| 拘押 | | 648.7978 | 2.690 | 241.201 | 0.000 | 643.525 | 654.070 |
| 确认的案例 | | 2.491e- 16 | 6.65e-08 | 3.74e- 09 | 1.000 | -1.3e- 07 | 1.3e-07 |
| 人口t-1 | | - 1.187e- 17 | 4.98e-08 | -2.39e- 10 | 1.000 | -9.76e- 08 | 9 .76e-08 |
| CCI | | 1.213e- 08 | 3.32e-08 | 0.366 | 0.715 | -5.29e- 08 | 7 .71e-08 |
| 家庭收入中位数（加州）t-1 | | 4.123e- 12 | 0.052 | 7.97e- 11 | 1.000 | -0.101 | 0.101 |
| 家庭收入中位数（加州）t-1 | | - 4.912e- 12 | 0.064 | -7.7e- 11 | 1.000 | -0.125 | 0.125 |
| 犯罪总数t-1 | | - 4.055e- 18 | 2 .2e-07 | -1.85e- 11 | 1.000 | -4.31e- 07 | 4 .31e-07 |
| ====================  总集。 | ==========================================================  23623 . 934Durbin-Watson: 0.282 | | | | | | |
| Prob(Omnibus): | 0 .000 | Jarque-Bera（JB）。 | | 12868679704.675 | | | |
| 倾斜。 | -2 .408 | Prob(JB): | | 0.00 | | | |

Kurtosis:3502. 310Cond.No. 1.76e+10

==============================================================================

Y = 648.7978 + 2.491e-16 X1 - 1.187e-17 X2 + 1.213e-08 X3 + 4.123e-12 X4 - 4.912e-12 X5 - 4.055e-18 X6

𝑌 = 648.7978 + 2.491 ∗ 10-16 𝑋1 - 1.187 ∗ 10-17 𝑋2 + 1.213 ∗ 10-8 𝑋3 + 4.123

∗ 10-12 𝑋4 - 4.912 ∗ 10-12 𝑋5 - 4.055 ∗ 10-18 𝑋6

X1 = 确诊病例 X2 = 人口t-1 X3 = CCI

X4 = 家庭收入中位数（加州）t-1

X5 = 家庭收入中位数（加州）t-1 X6 = 犯罪总额t-1

解释。

|  |  |
| --- | --- |
| 参数 | 描述 |
| b0 = 648.7978 | CASTHPI 648.7978我们可以得出结论，如果变量（X1），（X2），（X3），（X4），（X5）和（X6）是零（0）。所以，如果CASTHPI不受其他变量的影响，它的价值将是648.7978。 |
| b1=2.491e-16。 | confirmed\_cases (X1)是正的，那么confirmed\_cases (X1)变量就有正的影响/与CASTHPI成正比，如果X1增加，那么CASTHPI也增加。参数值为2.491e-16，所以当X1增加1个单位时，CASTHPI增加2.491e-16。假设其他变量不变。 |
| b2 = - 1.187e-17。 | 人口t-1（X2）为负数。 如果人口t-1（X2）增加，那么CASTPHI实际上是下降的。当X2增加时，其值为0.1084，那么房价就会减少1.187e-17 |
| b3=1.213e-08。 | CCI（X3）是正的，那么CCI变量（X3）有正的影响/与CASTHPI成正比，如果X3增加，那么CASTHPI也增加。参数值为1.213e-08。 |

|  |  |
| --- | --- |
|  | 那么当X3增加1个单位时，假设其他变量不变，CASTHPI增加1.213e-08 |
| b4=4.123e-12。 | 家庭收入中位数（加州）t-1（X4）为正，则变量家庭收入中位数（加州）t-1（X4）对CASTHPI有正向影响/成正比，如果X4增加，则CASTHPI也增加。参数值为4.123e-12，那么当X4上升1000美元时，CASTHPI上升4.123e-12，假设其他变量不变。 |
| B5=-4.912e-12。 | 家庭收入中位数(加州)t-1(X5)为负值。如果家庭收入中位数(加州)t-1(X5)上升，那么CASTHPI实际上会下降。该值为  -4.912e-12 当X5增加时，CASTHPI减少了4.912e-12 |
| b6=-4.055e-18。 | 总犯罪率t-1（X6）为负值。如果总犯罪率t-1（X6）上升，那么CASTHPI实际上会下降。值为4.055e-18，当X6增加时，那么CASTHPI就会减少4.055e-18 |

同期测试结果

F统计量=234.72334155427433 P值 =2.7031835586322296e-150

解释。

因为P值（2.7031835586322296e-150）小于alpha（0.05），所以决定拒绝H0，那么X和Y变量之间存在线性关系。

决定性系数（R-Square）和调整后的R square

R2 =5.302363424330991e-06

调整后的R2 =-0.00023264997390737285

解释。

得到的R2和调整后的R2值表明模型不准确，因为Y变量不能被X变量所解释，而且r-平方值不接近于1。

线性回归特征重要性

特征:0, 分数: 60.32347

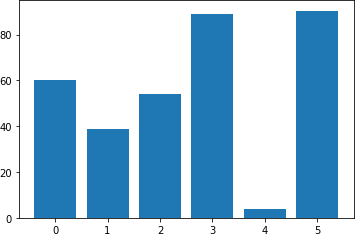
特征:1, 分数: 38.67115

特征: 2, 分数: 54.03354

特征:3, 分数: 88.82734

特征:4, 分数: 4.03574

特征:5, 分数: 90.44347



解释。

从线性回归特征重要性的结果来看，我们可以认为该分数表明模型找到了六个重要特征。

* 1. 梯度提升

XGBoost是一种算法，最近在结构化或表格化数据的机器学习应用中占据主导地位。XGBoost是一种梯度提升决策树的实现，旨在提高速度和性能。

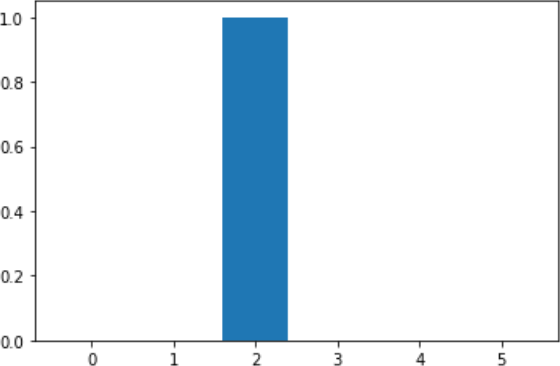
对于升压法参数，我们使用基于树的模型。原因是它是一个简单的模型，对规模差异不敏感。

从该模型中我们得到以下评价。

R2 : -24.94%

RMSE :1.79

MAE : 0.49

我们认为，R2的负值是因为目标变量的变化不大。因此，我们很难确定该特征对目标变化的描述程度。RMSE和MAE的得分分别为1.79和1.49。 这个数值表明模型的误差是相当小的。此外，我们介绍了模型中哪些特征是重要的。

|  |  |  |  |
| --- | --- | --- | --- |
| 特点。 | 0, | 分数。 | 0.00000 |
| 特点。 | 1, | 分数。 | 0.00000 |
| 特点。 | 2, | 分数。 | 1.00000 |
| 特点。 | 3, | 分数。 | 0.00000 |
| 特点。 | 4, | 分数。 | 0.00000 |
| 特点。 | 5, | 分数。 | 0.00000 |

由此可以看出，唯一对模型有影响的特征只有特征2，即CCI。

1. 参考文献

Prabhakaran, S. (n.a.).ARIMA模型--Python中时间序列预测的完整指南[。https://www.machinelearningplus.com/time-series/arima-model-time-series-](https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/)

[预测-python/](https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/)

杰森-布朗利。(2016年8月17日)。关于应用机器学习 的XGBoost的温和介绍[。https://machinelearningmastery.com/gentle-introduction-xgboost-applied-](https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/)

[机器学习/](https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/)

联邦调查局。(n.a.).美国的犯罪率[。https://ucr.fbi.gov/crime- in-the-u.s/](https://ucr.fbi.gov/crime-in-the-u.s/)

美国 社区 调查。 (n.a.)。 加州 家庭收入[。https://www.deptofnumbers.com/income/california/](https://www.deptofnumbers.com/income/california/)

美国人口普查局。 (n.a.). 人口[。https://www.census.gov/content/census/en/](https://www.census.gov/content/census/en/) DOJ。(n.a.).犯罪统计[。https://openjustice.doj.ca.gov/exploration/crime-statistics](https://openjustice.doj.ca.gov/exploration/crime-statistics)

圣路易斯联储。(n.a.).美国30年固定利率抵押贷款平均值[。https://fred.stlouisfed.org/series/MORTGAGE30US](https://fred.stlouisfed.org/series/MORTGAGE30US)

LATimes。 (n.a.)。 California-coronavirus-data [。https://github.com/datadesk/california-。](https://github.com/datadesk/california-coronavirus-data)

[冠状病毒-数据](https://github.com/datadesk/california-coronavirus-data)