北京師範大學

数据分析报告

部　　院　系：　统计学院

专　　　业：　经济与金融统计

学　　　号：　202322011053

学 生 姓 名：　欧宇翔

指 导 教 师：　谢传龙

2023 年 12 月

数据分析报告

摘　要

本文利用老师上课所给的数据集进行案例分析，实现端到端的机器学习框架。其中主要包括数据的初步探索，数据预处理，特征工程，模型建立，超参数调优等步骤。通过对该数据集进行练习，使得对机器学习项目有了更深入，更系统的了解。此外，本文所展示的内容均使用 Python 实现。

关键词：特征工程，搭建模型，模型调优

目　录

1 探索性数据分析

1.1 数据集描述

首先利用 Pandas 读取该数据集，观察可知，该组数据集可知一共有 66 个样本，30 个特征，3 个预测变量。部分变量和预测目标含义如下：

Festival& Public Holiday：该月内是否有节假日和公众假期

Summer & Winter Break：该月内是否有寒暑假

Exchange Rate：汇率

month_visit：每月的访问人数(已做预处理)

hotel_occupancy：酒店房间入住率

hotel_price：酒店房间价格

weather：平均气温

enviroment_pm2.5：平均环境质量

Hong Kong-Mainland China Border Crossing Policy：是否封关

此处样本量较少，因此在后续搭建模型时有可能导致模型欠拟合。此外，该组数据集自变量个数较多，有可能产生高维问题，如数据集可能存在多重共线性等问题。

1.2 变量分布信息可视化

利用 Seaborn 和 Matplotlib 对 30 个自变量分布进行可视化，结果如下。

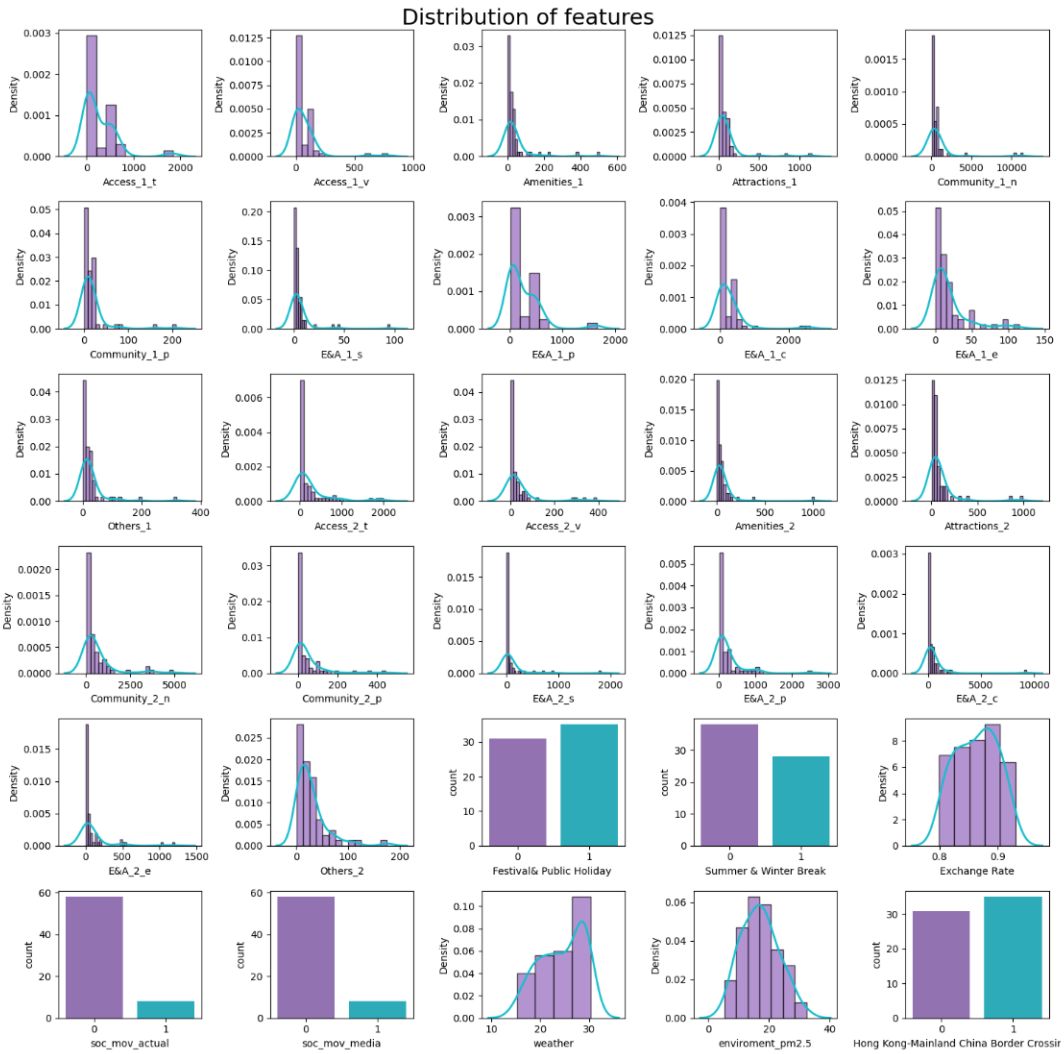

图 1.1 自变量分布图

观察上述分布可以初步得出结论

1.该组变量既包含数值型变量，也包含分类型变量。

2.数值型变量中，绝大部分变量分布并非对称分布，因此后续直接使用线性回归模型或其他模型进行拟合，极有可能导致模型拟合效果较差。

1.3 描述性统计

对数值型变量和分类型变量分别进行描述性统计，结果如下表所示，

对于数值型变量而言，其中不少变量的标准差较大，说明数据集分布较为异常，但由于样本量较少，若选择剔除异常值点会导致信息损失较大，因此初步考虑选择能有效抵抗异常值干扰的模型。

对于分类型变量而言，包含两个分类型变量分布不均衡，有可能是样本量较少所导致

的。

	count	mean	std	min	25%	50%	75%	max
Access_1_t	66.0	273.272727	358.643708	1.000000	39.250000	105.500000	482.500000	1854.0000
Access_1_v	66.0	70.151515	126.140367	0.000000	6.250000	20.500000	106.500000	781.0000
Amenities_1	66.0	39.151515	81.957445	0.000000	5.250000	18.000000	31.750000	503.0000
Attractions_1	66.0	89.363636	180.351591	6.000000	13.250000	37.000000	99.250000	1153.0000
Community_1_n	66.0	799.590909	1893.856952	46.000000	147.000000	294.000000	715.250000	11381.0000
Community_1_p	66.0	18.136364	33.510910	0.000000	2.000000	10.500000	20.000000	208.0000
E&A_1_s	66.0	5.515152	13.415539	0.000000	1.000000	2.000000	5.000000	95.0000
E&A_1_p	66.0	254.272727	320.194823	10.000000	35.000000	118.500000	428.750000	1654.0000
E&A_1_c	66.0	263.681818	456.021024	10.000000	22.750000	90.000000	350.000000	2634.0000
E&A_1_e	66.0	18.090909	24.449463	0.000000	3.000000	10.500000	19.000000	115.0000
Others_1	66.0	28.136364	47.911420	2.000000	6.250000	16.000000	26.750000	319.0000
Access_2_t	66.0	245.515152	430.323894	9.000000	26.000000	51.000000	205.250000	1997.0000
Access_2_v	66.0	38.045455	78.597403	0.000000	4.000000	11.500000	29.500000	391.0000
Amenities_2	66.0	57.272727	132.094785	1.000000	7.250000	25.500000	53.750000	1006.0000
Attractions_2	66.0	89.227273	164.623467	2.000000	21.000000	41.000000	80.500000	990.0000
Community_2_n	66.0	621.787879	958.526506	41.000000	140.250000	252.000000	601.500000	4922.0000
Community_2_p	66.0	46.303030	83.042338	0.000000	6.250000	13.000000	44.750000	443.0000
E&A_2_s	66.0	95.121212	266.369467	0.000000	3.250000	12.500000	53.000000	1811.0000
E&A_2_p	66.0	246.090909	398.453268	16.000000	48.250000	78.000000	273.750000	2539.0000
E&A_2_c	66.0	484.333333	1183.554104	24.000000	78.250000	174.500000	454.500000	9238.0000
E&A_2_e	66.0	99.272727	217.432398	1.000000	7.250000	17.500000	69.750000	1198.0000
Others_2	66.0	28.545455	29.923586	1.000000	9.000000	19.000000	34.500000	175.0000
Exchange Rate	66.0	0.863479	0.035663	0.799200	0.832025	0.865000	0.892325	0.9302
weather	66.0	24.421212	4.479853	15.200000	20.850000	25.150000	28.775000	30.3000
enviroment_pm2.5	66.0	17.078651	6.098001	5.277778	12.901042	16.583333	21.140625	32.3125

图 1.2 数值型变量描述统计

	Festival& Public Holiday	Summer & Winter Break	soc_mov_actual	soc_mov_media	Hong Kong-Mainland China Border Crossing Policy
count	66	66	66	66	66
unique	2	2	2	2	2
top	1	0	0	0	1
freq	35	38	58	58	35

图 1.3 分类型变量描述统计

1.4 目标变量可视化

对 3 个目标变量分布可视化结果如下

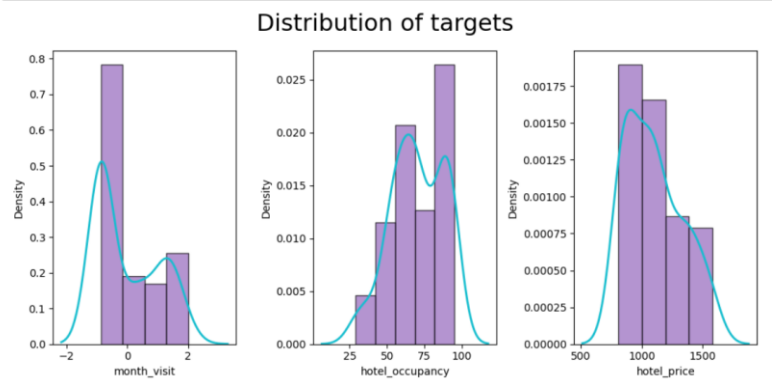

图 1.4 目标变量分布

上述 3 个目标变量也并非呈现对称分布，其中 month_visit 已进行预处理，而 hotel_occupancy 和 hotel_price 并未经过预处理。若每月的旅游人数较高，则有可能导致酒店入住率和酒店价格升高，因此推测目标变量之间存在一定的相关性。

1.5 自变量与目标变量散点图

分别绘制所有特征对目标变量的散点图结果如下。

1.所有特征与所有目标变量并没有显著的线性关系，若使用多元线性回归模型进行拟合，可能导致模型拟合效果较差。

2.存在不同特征与同一变量的散点图趋势大致相同，说明特征之间可能存在高度相关性。

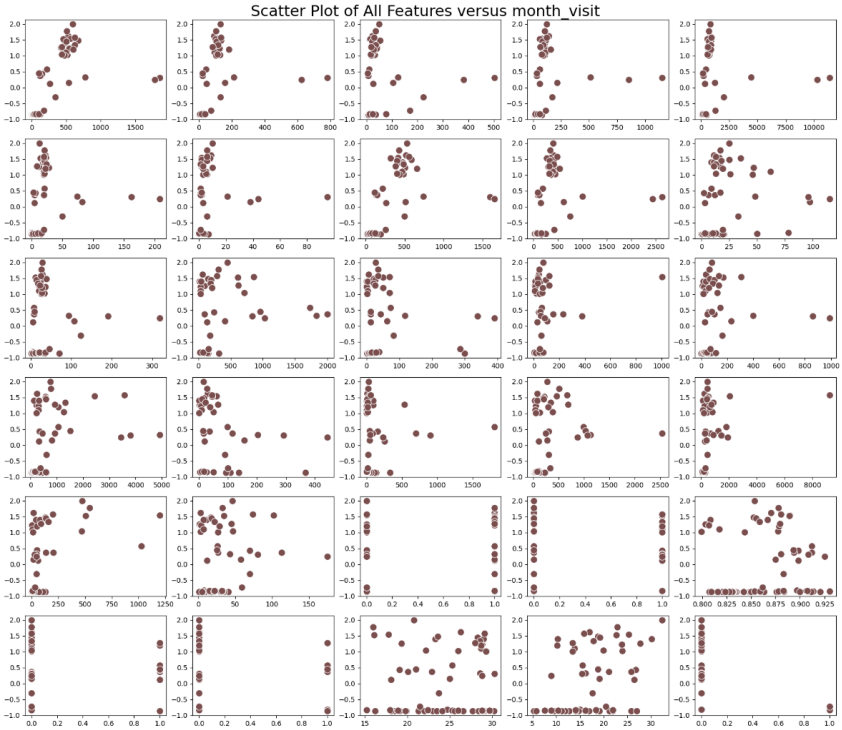

图 1.5 自变量与 month_visit 散点图

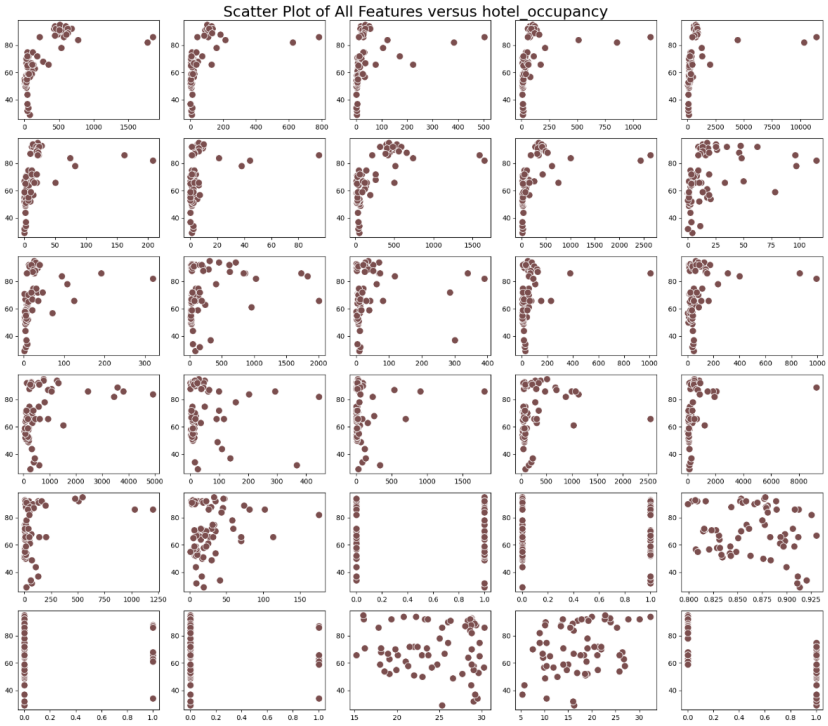

图 1.6 自变量与 hotel_occupancy 散点图

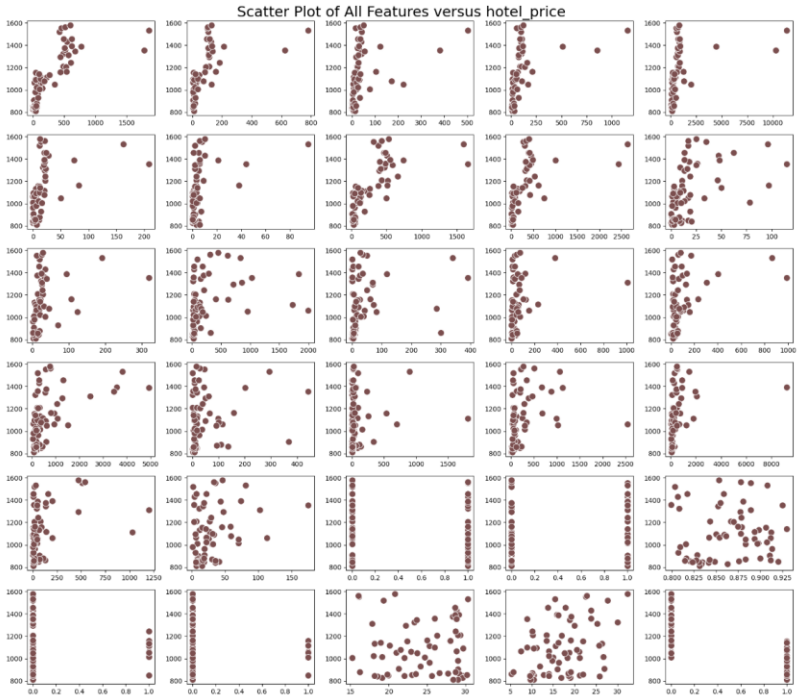

图 1.7 自变量与 hotel_price 散点图

1.6 自变量相关系数图

自变量相关系数图展示如下，前 11 个自变量之间高度相关，与前文初步结论一致。

此外，Attractions_2、Community_2_n、Community_2_p 也与其他变量高度相关。

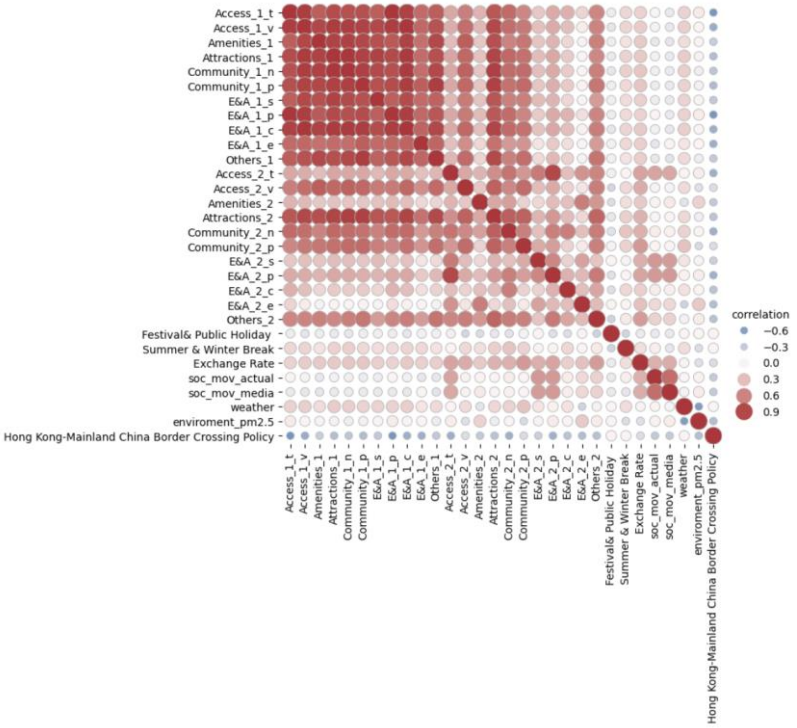

图 1.8 自变量相关系数热力图

2 特征工程

2.1 缺失值处理

利用 Pandas 对该组数据集探查，观察是否存在缺失值，若数据集中存在缺失值，可考虑使用平均数或中位数进行填充。

观察下图可知该组数据集不存在缺失值，因此不需要对缺失值进行处理。

```
: Access_1_t                                        0
  Access_1_v                                        0
  Amenities_1                                       0
  Attractions_1                                     0
  Community_1_n                                     0
  Community_1_p                                     0
  E&A_1_s                                           0
  E&A_1_p                                           0
  E&A_1_c                                           0
  E&A_1_e                                           0
  Others_1                                          0
  Access_2_t                                        0
  Access_2_v                                        0
  Amenities_2                                       0
  Attractions_2                                     0
  Community_2_n                                     0
  Community_2_p                                     0
  E&A_2_s                                           0
  E&A_2_p                                           0
  E&A_2_c                                           0
  E&A_2_e                                           0
  Others_2                                          0
  Festival& Public Holiday                          0
  Summer & Winter Break                             0
  Exchange Rate                                     0
  month_visit                                       0
  soc_mov_actual                                    0
  soc_mov_media                                     0
  hotel_occupancy                                   0
  hotel_price                                       0
  weather                                           0
  enviroment_pm2.5                                  0
  Hong Kong-Mainland China Border Crossing Policy   0
  dtype: int64
```

图 2.1 缺失值探查

2.2 数据标准化与归一化

数据标准化可以将数据转换为均值为 0，标准差为 1 的分布。而归一化可以将数据缩放到 0 和 1 之间。

本例中对数值型变量选择使用数据标准化，一方面考虑到数据样本量较少，且数据集偏差较大，若使用归一化有可能导致模型过拟合，泛化能力较差，另一方面使用数据标准化，可以较好的保留原始数据信息。

此外，对于数据变换还可以采用对数变换，幂次变换，Box-cox 变换等。需要注意的是原始数据集必须全为正数，否则需要引入一个较小的正数来保证数据能够正常变换。

经过数据标准化后的部分结果如下

	Access_1_t	Access_1_v	Amenities_1	Attractions_1	Community_1_n	C
0	0.875281	0.278385	0.035022	0.059428	-0.110985	
1	0.980993	0.310338	0.022727	0.065015	-0.071080	
2	1.036417	0.454130	-0.087928	0.081776	-0.020001	
3	1.079882	0.438153	-0.001863	0.042666	0.015116	
4	0.997182	0.198500	-0.173992	0.098538	-0.029578	
...	...	...	...	...	...	
61	0.724943	0.470107	2.260409	0.450533	0.641361	
62	1.062150	0.701771	0.797309	0.685196	0.231136	
63	1.364405	1.133146	1.030913	2.361360	1.944399	
64	2.128623	5.678560	5.702995	5.942764	5.630042	
65	2.091208	4.424377	4.215306	4.272187	5.055407	

图 2.2 标准化后部分结果展示

2.3 特征编码

在建立模型之前，必须对分类型变量进行编码，以便计算机能够识别分类变量。在本例中的 5 个分类变量均为二分类变量，且已转变为 0-1 编码，因此此处无需对分类变量进行编码。

2.4 特征选择

特征选择是指从原始数据中选择最有用的特征，用于训练模型。选择正确的特征可以提高模型的准确性和效率，并降低过拟合的风险。

由于本案例数据量较少，为了最大化保留模型的原始信息，此处选择保留所有变量。若在实际生活中样本量较大，则为了提高模型的计算效率和准确性，进行特征选择是非常有必要的。

2.5 特征组合

特征组合是指将原始特征进行组合，创建新的特征以提高模型性能或更好地捕获数据

中的复杂关系。特征组合主要包括多项式特征组合，交叉特征组合，特征转换等处理方法。

特征组合的目标是将原始特征转化为更具信息量的新特征，从而改善模型的性能。但特征组合也可能导致过拟合和维度过高等问题，由于本例数据量较少且特征数量较多，因此不考虑进行特征组合。

3 模型建立

3.1 随机森林

如先前所述，数值型变量已经经过标准化处理，变量间仍然存在高度相关性。此时可考虑两个大方向，第一便是对变量进行降维，消除变量相关性带来的影响。第二则是在模型中加入正则项来防止模型过拟合。

考虑到样本量数据较小，若进行降维处理，可能导致信息损失较大。部分变量含义不明确，若采用特征组合方式，有可能导致模型不具有可解释性，且产生过拟合等问题。

综上所述，本文选择保留所有变量特征，并选择能一定程度上解决变量之间高度相关性问题的模型。

随机森林模型在一定程度上可以解决变量之间高度相关性的问题。这是因为随机森林模型中的每棵树都是基于随机子集的数据和特征进行训练，而不是使用全部数据和全部特征，因此可以减少过拟合的风险。此外，随机森林还具有以下两个特点：

随机选择特征：随机森林在每次分裂节点时，会从所有特征中随机选择一部分特征用于建立决策树，这样可以避免使用高度相关的特征。

集成多个树：随机森林由多棵决策树组成，可以通过投票或平均值来确定最终的预测结果，这样可以降低由于某一棵树的误差导致的整体预测错误的可能性。

在本案例中，选择使用随机森林作为基准模型。

将样本按 7：3 划分训练集和测试集，使用随机森林模型观察三个目标变量在训练集和测试集上的均方误差（MSE）

由于样本数量较小，在训练集和测试集分别使用交叉验证提高精确度，其中折数 k=5。

显然使用随机森林模型预测三个目标变量，在测试集上表现较差，原因可能为

1.样本数量过少，测试集只有 20 个样本，可能存在异常值干扰

2.训练集样本量较少，目标函数可能未达到最优解

3.模型过拟合，泛化能力较差

4.该模型异常值较多

表 3.1 随机森林拟合结果

	rf_train_mse	rf_test_mse
month_visit	0.013911	0.073932
hotel_occupancy	10.413949	34.104972
hotel_price	2632.570052	12606.257075

3.2 Xgboost

Boosting 是一种集成学习方法，旨在通过将多个弱分类器组合成一个强分类器来提高模型性能。该算法的核心思想是通过反复训练弱分类器并加权调整样本权重，来逐步提升整个模型的性能。Boosting 算法的优点在于，它可以有效地减少模型的偏差，提高模型的泛化能力和准确度。

Xgboost 模型由多棵树组成，每棵树都是一个回归树或分类树。对于回归问题，每个叶节点对应一个实数值。在训练过程中会不断添加新的树，并将新的树的输出与前面的树相加，得到最终预测值。该过程可以表示为下式

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i)\ ,\ f_k(x_i) \in F$$

其中，$\hat{y}_i$表示样本i的预测值，K表示树的数量，$f_k(x_i)$表示第k棵树的输出，F表示所有可能的树的集合，关于 Xgboost 数学表达式具体推导过程参考其他文献，此处不进行详细展开。它的特点主要有以下几点：

1.梯度提升树：XGBoost 使用梯度提升树作为基础模型。梯度提升树通过逐步拟合残差来改善模型的预测能力，每一棵树都尝试修正前一棵树的预测结果。

2.正则化：XGBoost 引入了正则化项，以控制模型的复杂度和过拟合的风险。通过在目标函数中添加正则化项，可以有效地防止模型过度拟合训练数据。

3.特征重要性评估：XGBoost 提供了对特征重要性的评估，可以帮助你理解各个特征对模型预测的贡献程度。这对于特征选择和特征工程非常有用。由于

4.灵活性：XGBoost 提供了丰富的超参数调节选项，可以根据具体问题进行调优。它还支持分类、回归、排序等不同类型的问题，并可以处理缺失值。

由于本例数据量较少，为了最大化保留信息特征，故不进行特征重要性评估。当样本量足够大时，进行特征选择从而提高模型计算效率是非常重要的。

利用 Xgboost 模型计算在训练集和测试集上的均方误差，和前文一样均使用交叉验证来提高准确度。

可以发现使用 Xgboost 模型，在测试集上表现也很差，大致原因如上。

表 3.2 Xgboost 拟合结果

	xgb_train_mse	xgb_test_mse
month_visit	0.021066	0.262726
hotel_occupancy	4.352665	129.628401
hotel_price	2679.808739	13589.428314

3.3 超参数调优

超参数调优是指通过选择最佳的超参数组合来优化机器学习模型的性能。超参数是在模型训练之前设置的参数，不同的超参数组合可能导致不同的模型表现。

常用的超参数调优方法主要包括网格搜索，随机搜索，贝叶斯优化，集成优化等。本例中对 Xgboost 利用 Pyhon 中的 Optuna 筛选最优参数组合，将改进后的 Xgboost 模型和上述模型进行对比，结果如下。

经过改进后的模型，在训练集的表现不如原始模型，但在测试集上的表现要优于原始模型，表明偶然误差影响因素很大，下列结果仅供参考。

将代码反复运行几次，每一次得到的结果都截然不同，说明数据集量较少，可解释性较差。

表 3.3 所有模型拟合结果

	rf_train_mse	rf_test_mse	xgb_train_mse	xgb_train_mse	optuna_xgb_train_mse	optuna_xgb_test_mse
month_visit	0.01	0.07	0.02	0.26	0.05	0.19
hotel_occupancy	10.41	34.10	4.35	129.63	20.49	61.73
hotel_price	2632.57	12606.26	2679.81	13589.43	5364.56	8457.39

3.4 模型预测

分类型变量主要包括 Festival& Public Holiday，Summer & Winter Break，Hong Kong-Mainland China Border Crossing Policy，soc_mov_actual，soc_mov_media。对于前三个特征

而言，可以通过网上查询或是结合先前经验可以得到真实值，对于后二个特征而言，尽管含义不明确，但结合数据分析可知仅在 2019 年 7 月到 2020 年 6 月存在 1 取值，因此推测可能与疫情封控有关。结合 2020 年 7 月至 2023 年 7 月数据，可以认为 2023 年 8 月该两个变量取值特征为零。

对于数值型变量而言，观察 25 个数值型变量均与时间序列有关，因此考虑使用时间序列分析来预测每个指标。绘制偏自相关系数图如下，可以观察到所有变量均与滞后一阶数高度相关，大部分变量与滞后二阶数高度相关。

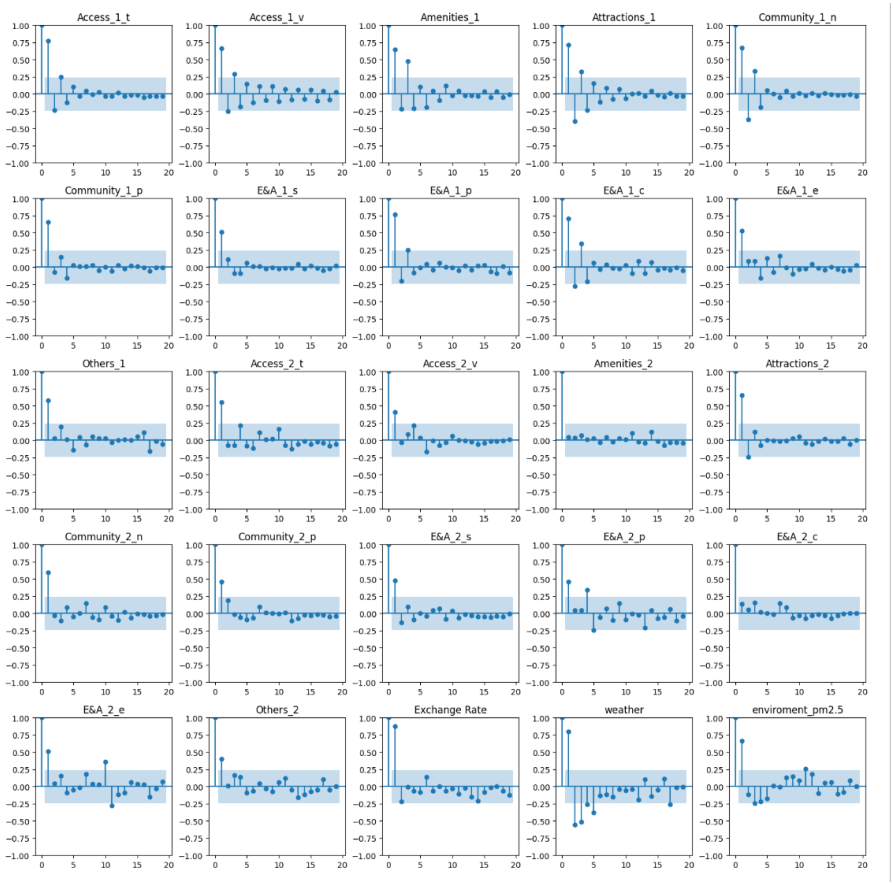

图 3.1 数值型变量偏自相关图

此处利用 Xgboost 筛选最优每个数值变量最优滞后阶数，然后利用 AR 模型计算出下一期的变量值，计算结果如下：

表 3.4 变量最优滞后期数和下一期的预测值

	lag	values
Access_1_t	2.0	2235.253303
Access_1_v	1.0	553.691750
Amenities_1	1.0	246.430858
Attractions_1	1.0	879.695770

Community_1_n	4.0	140054.492131
Community_1_p	1.0	829.187119
E&A_1_s	1.0	9.061125
E&A_1_p	1.0	2565.404751
E&A_1_c	1.0	3885.172975
E&A_1_e	1.0	31.963661
Others_1	1.0	1464.965505
Access_2_t	2.0	304.597101
Access_2_v	4.0	430.156223
Amenities_2	3.0	64.053167
Attractions_2	1.0	2850.579380
Community_2_n	4.0	1747.329884
Community_2_p	1.0	134.999005
E&A_2_s	1.0	107.697711
E&A_2_p	1.0	288.086868
E&A_2_c	1.0	509.585551
E&A_2_e	1.0	104.686889
Others_2	1.0	48.011142
Exchange Rate	1.0	0.911162
weather	1.0	25.992086
enviroment_pm2.5	3.0	17.364831
Access_1_t	2.0	2235.253303
Access_1_v	1.0	553.691750
Amenities_1	1.0	246.430858
Attractions_1	1.0	879.695770

将每个变量进行标准化处理后代入优化后的 Xgboost 模型中，最终计算结果如下

表 3.5 目标变量预测值

	month_visit	hotel_occupancy	hotel_price
0.395767	75.402931	1240.989624	0.395767

4 未来可行性分析

4.1 特征工程处理

受到样本量的限制，本文中仅展示了使用标准化对数据集进行处理，但当数据量较大时，可以考虑使用其他方法对数据集进行处理。同时样本量足够大时，可以考虑剔除异常值点等操作。此时需要注意对特征个数的限制，过多的特征容易使得计算成本大大增加，而模型精度却未得到有效提高。

4.2 模型选择

本案例选择使用随机森林作为基准模型，通过使用 Xgboost 模型并调优来进行对比。但在实践中可以发现，并非复杂模型效果会更好，在实际使用中还应注意模型的适用场景。

还可以对数据进行降维处理，然后使用决策树或者多元线性回归等其他模型来拟合对比分析。

本案例中使用一个模型来对三个目标变量同时进行处理，因此还可以考虑针对不同的目标变量分别构建相同或者不同的模型，并分别进行超参数调优，从而提高预测精度。例如本案例可以使用决策树模型预测 month_visit，使用随机森林预测 hotel_occupancy 和 hotel_price，或者是其他组合方式。

该组数据集还涉及到了时间序列，根据客观经验，酒店的价格，旅游人数，酒店入住率收到时间和季节的波动影响，因此可以考虑使用 ARIMA 模型等时间序列模型来对本数据集进行处理。

此外，还可以使用神经网络来对目标变量进行预测分析，例如使用 RNN 循环神经网络或者是 Transformer 等神经网络模型。但这要求数据量足够大，模型才具有泛化能力和可解释性。