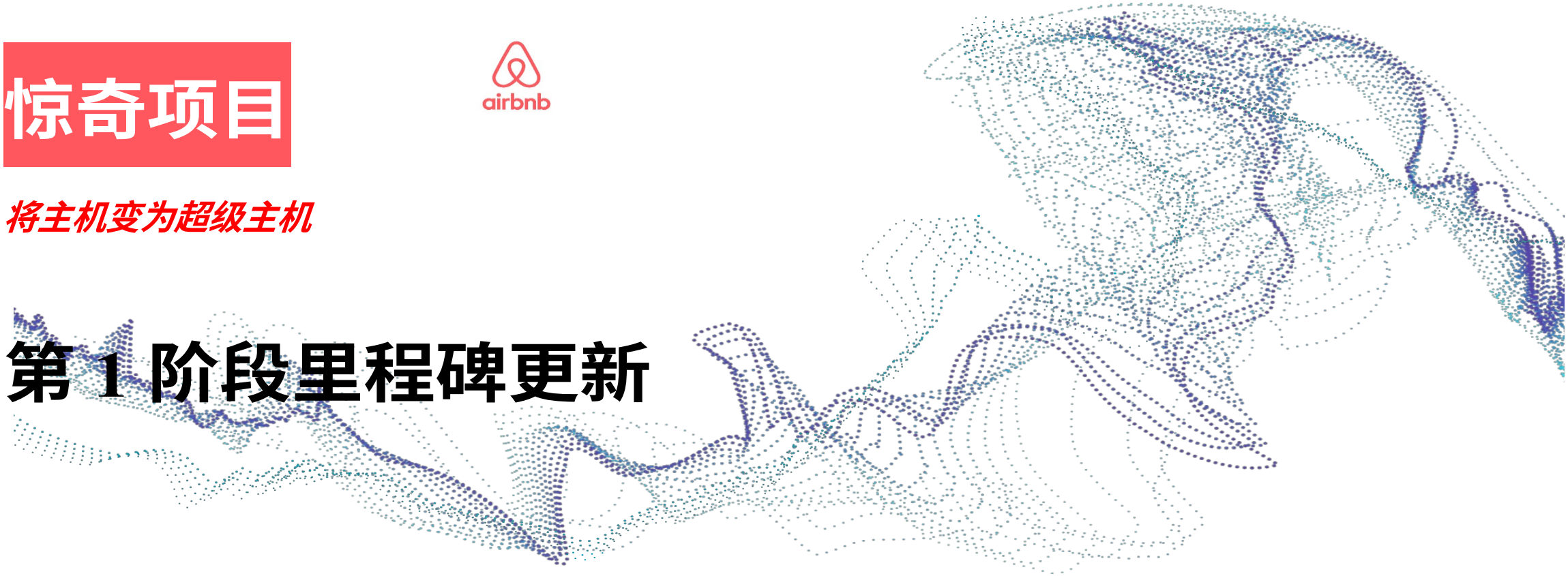
Group18 Consulting Co.

惊奇项目

airbnb

将主机变为超级主机



第 1 阶段里程碑更新

G18 咨询项目组：

乔纳森-约克 黄炳耀

Jisoo KIM Elsa ZHAN

苏道清 Terence ZHANG

G18 CONSULTING
WE EAT YOUR PROBLEM FOR BREAKFAST

1.
导言

- 项目定义
- Airbnb "超级房东 "计划的背景
- 香港市场研究


G18 CONSULTING
WE EAT YOUR PROBLEM FOR BREAKFAST

1.1 项目定义

有哪些简单可行的因素可以帮助香港的房东成为 "超级房东"？

关于**帮助香港房东成为 "超级房东 "的**影响因素的定量分析

G18 CONSULTING
WE EAT YOUR PROBLEM FOR BREAKFAST

1.2 背景 - "超级主机"

什么是 "超级主机"?

为什么是 "超级主机"?

Airbnb 有什么好处?

Airbnb **表现最出色的房东**。

东道主必须拥有一个信誉良好的账户，并在过去 12 个月中符合以下标准：

- 完成**至少 10 次旅行**或 3 次预订，总计至少 100 晚；
- 保持 **90% 以上的回复率**；
- 保持**低于 1%的取消率**，但符合 Airbnb 情有可原政策的情况除外；
- **总体评分**保持在 **4.8 分**。

获得潜在客人的更多关注、额外的盈利潜力、专属奖励，并获得 Airbnb 的优先支持

- **每周浏览量**增加 5
- **入住率**提高 81
- **每日收入**比平均固定主机
- AirBnB 提供的**现金奖励**用于导师

鉴于 Airbnb 的收入主要来自两个方面，它也有动力鼓励更多房东成为 "超级房东"：

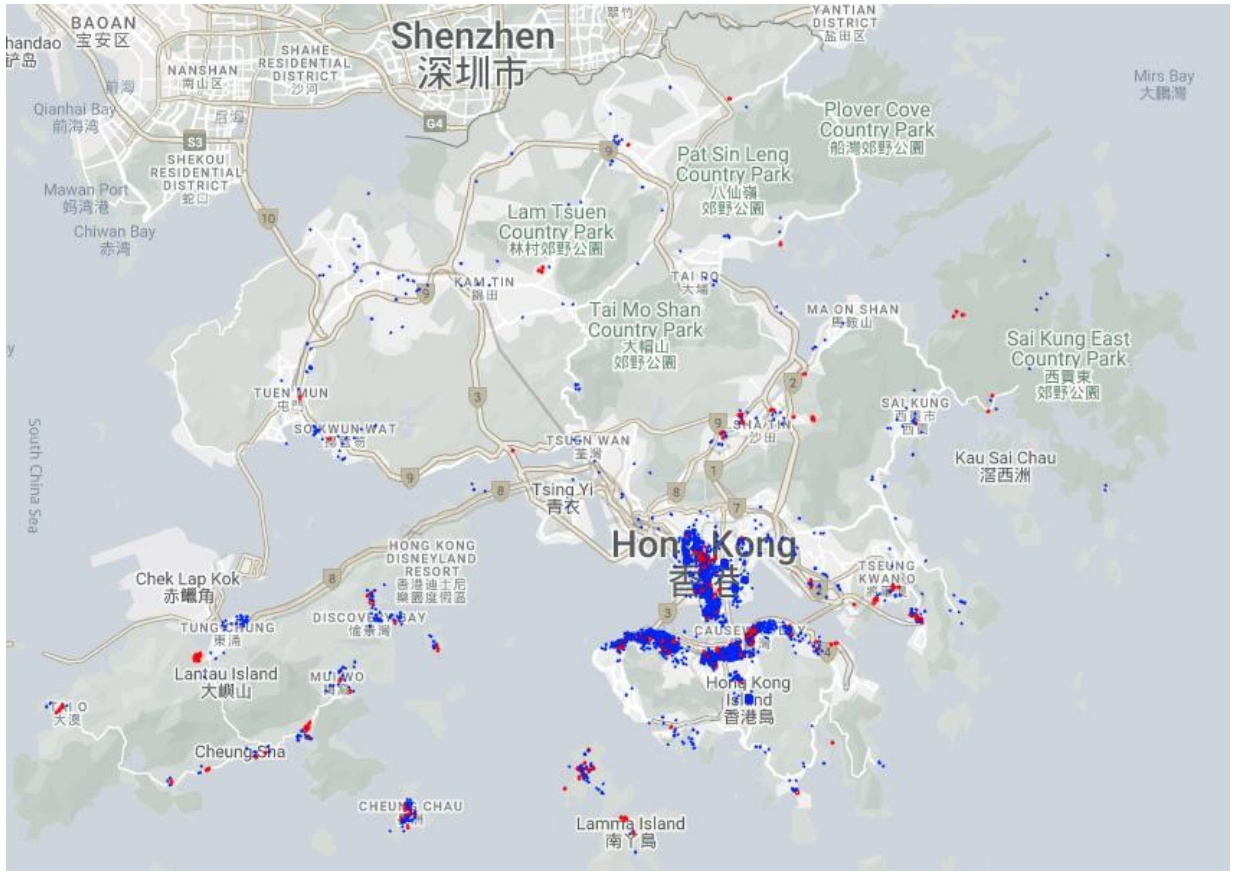
- **房东佣金：**每次有人选择房东的房产并付款时，Airbnb 都会从付款金额中抽取 10% 作为佣金。
- **旅客支付的交易费：**当旅客为住宿付款时，他们会被收取 3% 的交易费。这笔费用会增加 Airbnb 的收入。

1.2 背景--市场研究

19.4%

10.5%

1.2 背景 - 香港市场研究



对邻里的影响?

沙田、西贡和**香港岛的**房源通常由 "超级房东 "提供。

葵青、黄大仙和**深水埗**的 "超级房东 "比率最低。

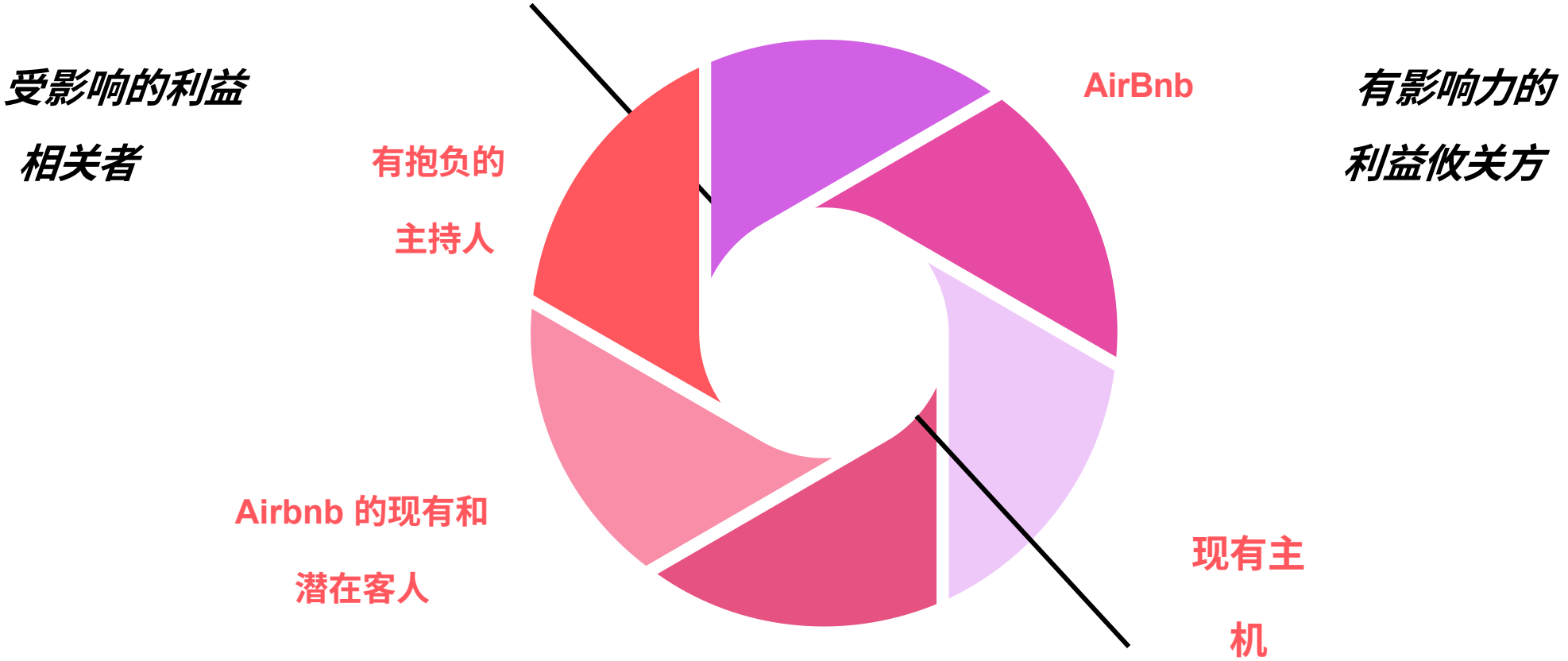
5056 个列表，532 个超级主机

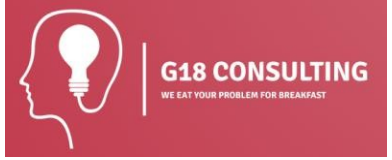
数据最后一次采集于 2022 年 9 月 16 日

包括 2009 年至 2022 年的东道主

1.3 利益相关者分析

提供实用技巧，帮助有抱负的东道主和现有东道主成为 "超级东道主"。




G18 CONSULTING
WE EAT YOUR PROBLEM FOR BREAKFAST

2.
分析与结论

- 数据清理和准备
- 模型和研究结果


G18 CONSULTING
WE EAT YOUR PROBLEM FOR BREAKFAST

2.1 数据准备

数据集：

- 香港地区 5056 个列表，75 个栏目
- 变量概述：
    - 主机信息：位置、ID、验证状态
    - 房源信息：最多/最少可入住天数、价格、设施、房型、位置
    - 审查评分信息：审查评分细目及其数值

数据准备步骤

1. 将 "对象 "数据类型的数据转换为适当的数据类型；
2. 删除缺失值超过 **75% 的**列；
3. 用适当的条目填补缺失的数据；

4. 删除变量；以及
5. 创建分类/虚拟变量。

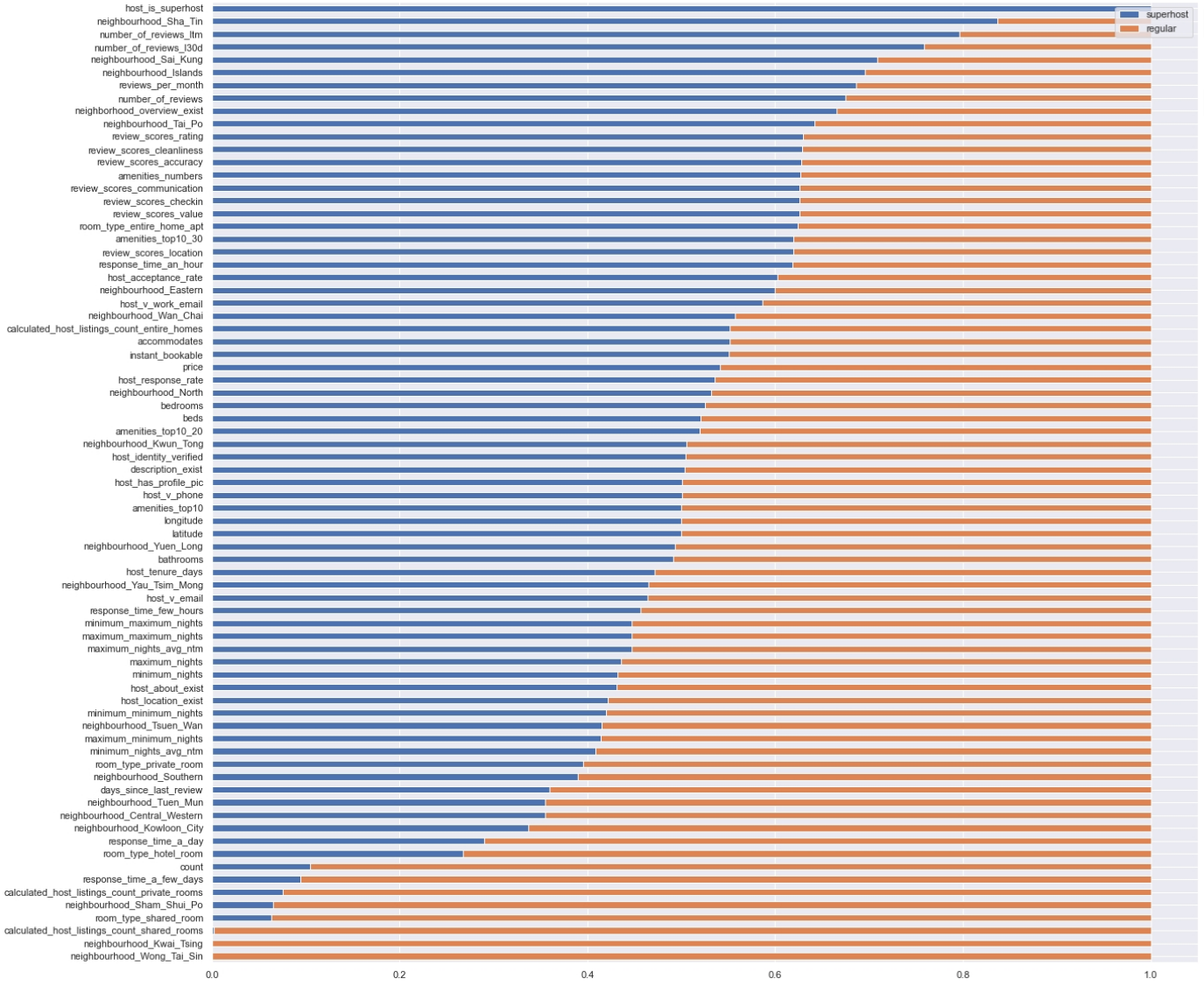
2.2 了解主机与超级主机

了解差异

1. 邻里事项!
2. 评论与超级主机相关
3. 评论数量也很重要

   超级主机
4. ...
5. 有很多...

终结语：成为超级主机有很多相关因素，但



并非所有因素都是可操作的，也并非所有因素都很简单。

2.2 模型和结果

模型 1：特征数据的逻辑回归

- Host_is_Superhost：从属变量（二进制）
- 使用 AIC 前向选择模型进一步缩小变量数量。
- 进行 VIF 分析，检查是否存在多重共线性问题。我们决定放弃 host_has_profilepic
- 我们最终的逻辑回归模型包含 9 个变量。

	variables	VIF
0	amenities_numbers	5.237735
1	host_acceptance_rate	3.853621
2	host_v_email	17.219852
3	neighborhood_overview_exist	2.294009
4	host_identity_verified	2.728111
5	response_time_a_day	1.161903
6	response_time_a_few_days	1.102534
7	beds	2.695326
8	instant_bookable	1.845464
9	host_about_exist	5.039361
10	host_has_profile_pic	20.956781

	variables	VIF
0	amenities_numbers	5.072229
1	host_acceptance_rate	3.780598
2	host_v_email	7.852591
3	neighborhood_overview_exist	2.290653
4	host_identity_verified	2.721392
5	response_time_a_day	1.157507
6	response_time_a_few_days	1.091049
7	beds	2.661999
8	instant_bookable	1.844996
9	host_about_exist	4.616332

Generalized Linear Model Regression Results

Dep. Variable:	host_is_superhost	No. Observations:	4050
Model:	GLM	Df Residuals:	4039
Model Family:	Binomial	Df Model:	10
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-932.70
Date:	Sun, 11 Dec 2022	Deviance:	1865.4
Time:	18:30:38	Pearson chi2:	2.81e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-2.7300	0.216	-12.648	0.000	-3.153	-2.307
amenities_numbers	0.0698	0.006	10.891	0.000	0.057	0.082
host_acceptance_rate	2.7683	0.242	11.427	0.000	2.294	3.243
host_v_email	-1.9317	0.196	-9.856	0.000	-2.316	-1.548
neighborhood_overview_exist	1.1165	0.144	7.736	0.000	0.834	1.399
host_identity_verified	-0.8830	0.142	-6.217	0.000	-1.161	-0.605
response_time_a_day	-1.5392	0.300	-5.124	0.000	-2.128	-0.950
response_time_a_few_days	-1.9418	0.597	-3.251	0.001	-3.112	-0.771
beds	-0.1859	0.050	-3.706	0.000	-0.284	-0.088
instant_bookable	-0.5843	0.147	-3.976	0.000	-0.872	-0.296
host_about_exist	-0.2795	0.136	-2.050	0.040	-0.547	-0.012

Generalized Linear Model Regression Results

Dep. Variable:	host_is_superhost	No. Observations:	4050
Model:	GLM	Df Residuals:	4039
Model Family:	Binomial	Df Model:	10
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-932.70
Date:	Sun, 11 Dec 2022	Deviance:	1865.4
Time:	18:30:38	Pearson chi2:	2.81e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-2.7300	0.216	-12.648	0.000	-3.153	-2.307
amenities_numbers	0.0698	0.006	10.891	0.000	0.057	0.082
host_acceptance_rate	2.7683	0.242	11.427	0.000	2.294	3.243
host_v_email	-1.9317	0.196	-9.856	0.000	-2.316	-1.548
neighborhood_overview_exist	1.1165	0.144	7.736	0.000	0.834	1.399
host_identity_verified	-0.8830	0.142	-6.217	0.000	-1.161	-0.605
response_time_a_day	-1.5392	0.300	-5.124	0.000	-2.128	-0.950
response_time_a_few_days	-1.9418	0.597	-3.251	0.001	-3.112	-0.771
beds	-0.1859	0.050	-3.706	0.000	-0.284	-0.088
instant_bookable	-0.5843	0.147	-3.976	0.000	-0.872	-0.296
host_about_exist	-0.2795	0.136	-2.050	0.040	-0.547	-0.012

Generalized Linear Model Regression Results

Dep. Variable:	host_is_superhost	No. Observations:	4050
Model:	GLM	Df Residuals:	4039
Model Family:	Binomial	Df Model:	10
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-932.70
Date:	Sun, 11 Dec 2022	Deviance:	1865.4
Time:	18:30:38	Pearson chi2:	2.81e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-2.7300	0.216	-12.648	0.000	-3.153	-2.307
amenities_numbers	0.0698	0.006	10.891	0.000	0.057	0.082
host_acceptance_rate	2.7683	0.242	11.427	0.000	2.294	3.243
host_v_email	-1.9317	0.196	-9.856	0.000	-2.316	-1.548
neighborhood_overview_exist	1.1165	0.144	7.736	0.000	0.834	1.399
host_identity_verified	-0.8830	0.142	-6.217	0.000	-1.161	-0.605
response_time_a_day	-1.5392	0.300	-5.124	0.000	-2.128	-0.950
response_time_a_few_days	-1.9418	0.597	-3.251	0.001	-3.112	-0.771
beds	-0.1859	0.050	-3.706	0.000	-0.284	-0.088
instant_bookable	-0.5843	0.147	-3.976	0.000	-0.872	-0.296
host_about_exist	-0.2795	0.136	-2.050	0.040	-0.547	-0.012

2.2 模型和结果

模型 2：市容数据的逻辑回归

- 筛选出香港列表中最常见的设施。选取了 35 个变量，占香港上市设施的 80% 以上。
- 再进一步看 "便利设施"。
- 再进行一次逻辑回归，看看哪些因素与 "超级主机 "身份最相关？

- 使用 AIC 正向选择模型和 VIF 检验多重共线性问题。

	variables	VIF
0	Shampoo	6.252371
1	Iron	2.925847
2	Hot_water_kettle	1.430071
3	First_aid_kit	1.788546
4	Elevator	2.708379
5	Coffee_maker	1.603831
6	TV	3.687530
7	Cable_TV	1.756567
8	Dedicated_workspace	1.419282
9	Kitchen	4.788073
10	Dryer	1.460525
11	Dishes_and_silverware	3.743478
12	Hot_water	3.334319
13	Extra_pillows_and_blankets	1.721110
14	Fire_extinguisher	2.999285
15	Cooking_basics	2.539328
16	Hair_dryer	6.879310
17	Refrigerator	3.522900
18	Air_conditioning	14.585210
19	Essentials	6.893271
20	Hangers	4.298193
21	Long_term_stays_allowed	15.029286
22	Luggage_dropoff_allowed	1.835552
23	Carbon_monoxide_alarm	1.640830
24	Lock_on_bedroom_door	1.853370

前 3 名

洗发熨斗

咖啡机

Generalized Linear Model Regression Results

Dep. Variable:	host_is_superhost	No. Observations:	5056
Model:	GLM	Df Residuals:	5031
Model Family:	Binomial	Df Model:	24
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1256.7
Date:	Mon, 12 Dec 2022	Deviance:	2513.4
Time:	19:37:25	Pearson chi2:	5.36e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-4.5907	0.346	-13.260	0.000	-5.269	-3.912
Shampoo	1.2221	0.179	6.831	0.000	0.871	1.573
Iron	0.8694	0.131	6.656	0.000	0.613	1.125
Hot_water_kettle	0.4263	0.169	2.519	0.012	0.095	0.758
First_aid_kit	0.5245	0.121	4.329	0.000	0.287	0.762
Elevator	-1.0319	0.115	-8.942	0.000	-1.258	-0.806
Coffee_maker	0.8132	0.155	5.233	0.000	0.509	1.118
TV	0.7333	0.152	4.825	0.000	0.435	1.031
Cable_TV	0.7780	0.188	4.148	0.000	0.410	1.146
Dedicated_workspace	0.5728	0.129	4.447	0.000	0.320	0.825
Kitchen	-0.8450	0.149	-5.684	0.000	-1.136	-0.554
Dryer	0.2279	0.123	1.855	0.064	-0.013	0.469
Dishes_and_silverware	1.0339	0.202	5.108	0.000	0.637	1.431
Hot_water	-0.4396	0.149	-2.940	0.003	-0.733	-0.147
Extra_pillows_and_blankets	-0.3403	0.152	-2.237	0.025	-0.639	-0.042
Fire_extinguisher	0.4785	0.130	3.681	0.000	0.224	0.733
Cooking_basics	0.3323	0.154	2.159	0.031	0.031	0.634
Hair_dryer	0.8791	0.207	4.251	0.000	0.474	1.285
Refrigerator	-0.5541	0.168	-3.303	0.001	-0.883	-0.225
Essentials	-0.2636	0.203	-1.299	0.194	-0.661	0.134
Hangers	0.1213	0.159	0.761	0.447	-0.191	0.434
Long_term_stays_allowed	0.2415	0.275	0.879	0.379	-0.297	0.780
Luggage_dropoff_allowed	-0.0908	0.141	-0.643	0.520	-0.368	0.186
Carbon_monoxide_alarm	0.3394	0.130	2.607	0.009	0.084	0.594
Lock_on_bedroom_door	-0.1268	0.129	-0.980	0.327	-0.380	0.127


G18 CONSULTING
WE EAT YOUR PROBLEM FOR BREAKFAST

负面影响：为什么？厨房电梯

Generalized Linear Model Regression Results

Dep. Variable:	host_is_superhost	No. Observations:	5056
Model:	GLM	Df Residuals:	5031
Model Family:	Binomial	Df Model:	24
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1256.7
Date:	Mon, 12 Dec 2022	Deviance:	2513.4
Time:	19:37:25	Pearson chi2:	5.36e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-4.5907	0.346	-13.260	0.000	-5.269	-3.912
Shampoo	1.2221	0.179	6.831	0.000	0.871	1.573
Iron	0.8694	0.131	6.656	0.000	0.613	1.125
Hot_water_kettle	0.4263	0.169	2.519	0.012	0.095	0.758
First_aid_kit	0.5245	0.121	4.329	0.000	0.287	0.762
Elevator	-1.0319	0.115	-8.942	0.000	-1.258	-0.806
Coffee_maker	0.8132	0.155	5.233	0.000	0.509	1.118
TV	0.7333	0.152	4.825	0.000	0.435	1.031
Cable_TV	0.7780	0.188	4.148	0.000	0.410	1.146
Dedicated_workspace	0.5728	0.129	4.447	0.000	0.320	0.825
Kitchen	-0.8450	0.149	-5.684	0.000	-1.136	-0.554
Dryer	0.2279	0.123	1.855	0.064	-0.013	0.469
Dishes_and_silverware	1.0339	0.202	5.108	0.000	0.637	1.431
Hot_water	-0.4396	0.149	-2.940	0.003	-0.733	-0.147
Extra_pillows_and_blankets	-0.3403	0.152	-2.237	0.025	-0.639	-0.042
Fire_extinguisher	0.4785	0.130	3.681	0.000	0.224	0.733
Cooking_basics	0.3323	0.154	2.159	0.031	0.031	0.634
Hair_dryer	0.8791	0.207	4.251	0.000	0.474	1.285
Refrigerator	-0.5541	0.168	-3.303	0.001	-0.883	-0.225
Essentials	-0.2636	0.203	-1.299	0.194	-0.661	0.134
Hangers	0.1213	0.159	0.761	0.447	-0.191	0.434
Long_term_stays_allowed	0.2415	0.275	0.879	0.379	-0.297	0.780
Luggage_dropoff_allowed	-0.0908	0.141	-0.643	0.520	-0.368	0.186
Carbon_monoxide_alarm	0.3394	0.130	2.607	0.009	0.084	0.594
Lock_on_bedroom_door	-0.1268	0.129	-0.980	0.327	-0.380	0.127


G18 CONSULTING
WE EAT YOUR PROBLEM FOR BREAKFAST

2.2 模型和结果

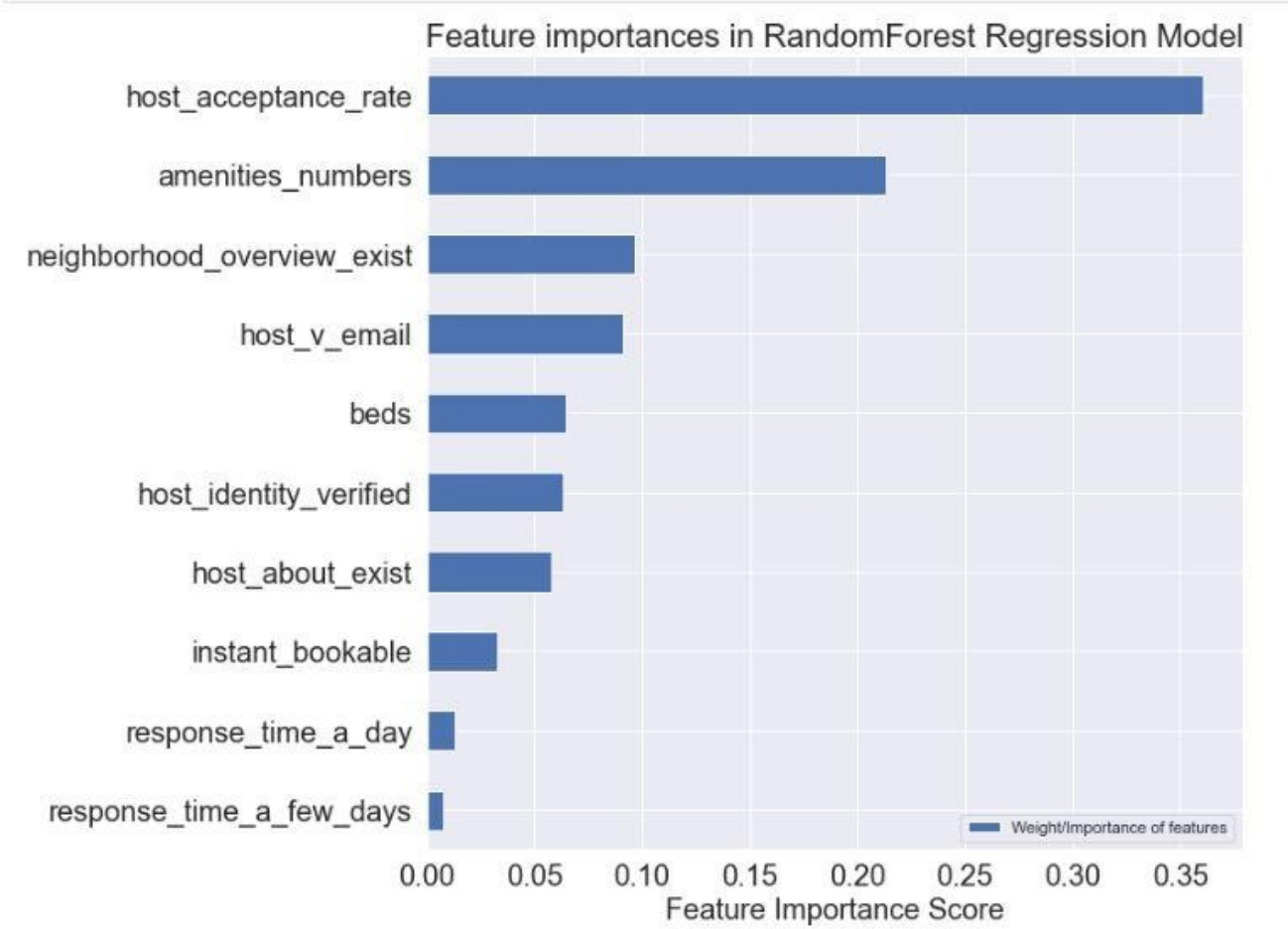
模型 3：对特征数据进行随机森林分类

- 对上述最终模型 1 确定的变量进行随机森林回归，从另一个角度验证每个变量的重要性。

- 两种模型得出的结果相似，即可用设施的数量和房东的**接受率**以及是否**存在 "**房东 "。

  影响最大的**街区概览**

  成为 "超级主机"。




G18 CONSULTING
WE EAT YOUR PROBLEM FOR BREAKFAST

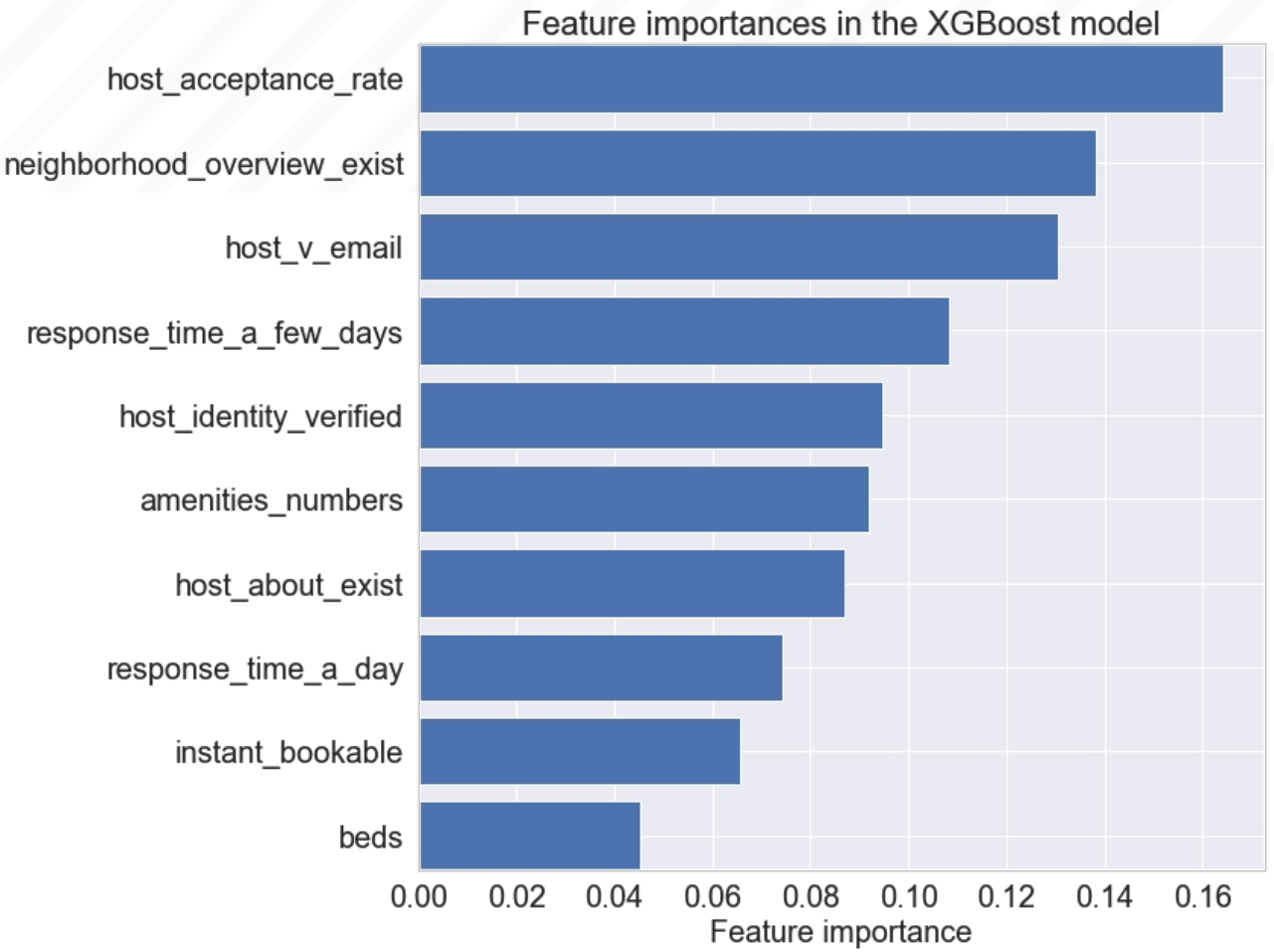
2.2 模型和结果

型号 4：XGBoost

- XGBoost 模型来验证每个最终模型 1 中确定的变量

- 有**验证电子邮件**和**响应时间慢**的主机重要性得分最高。

- 我们必须警惕的是，这一模型并不能说明某项功能对成为 "超级主机 "的前景有积极还是消极的影响。




G18 CONSULTING
WE EAT YOUR PROBLEM FOR BREAKFAST

3.
建议

- 对东道主的建议


G18 CONSULTING
WE EAT YOUR PROBLEM FOR BREAKFAST

3.建议

Airbnb 应鼓励香港房东...

在

天

尽可能多地接受预订

可能

如果不能接受预订，提前删除

未来日期的列表

包括概述

列表中的街区

提供更多便利设施。不仅要

提供基本设施，还要为客人

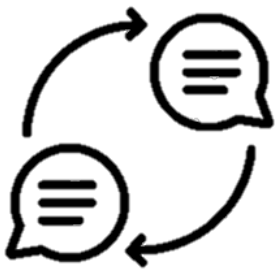
提供额外的服务

顶级设施

熨斗、电视、设备齐全的厨

房、设备齐全的浴室、

工作区、安全设备



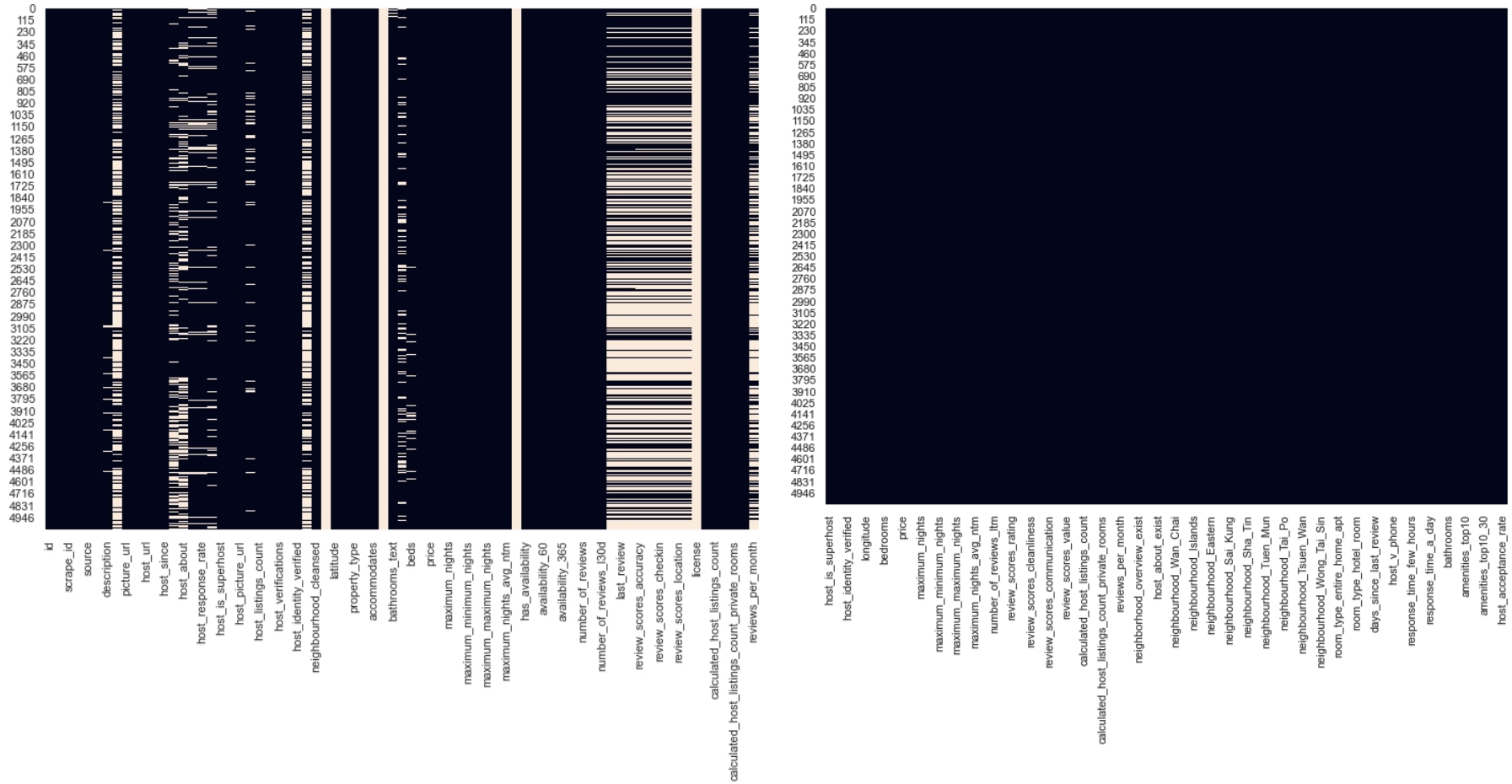
谢谢！

有问题吗？

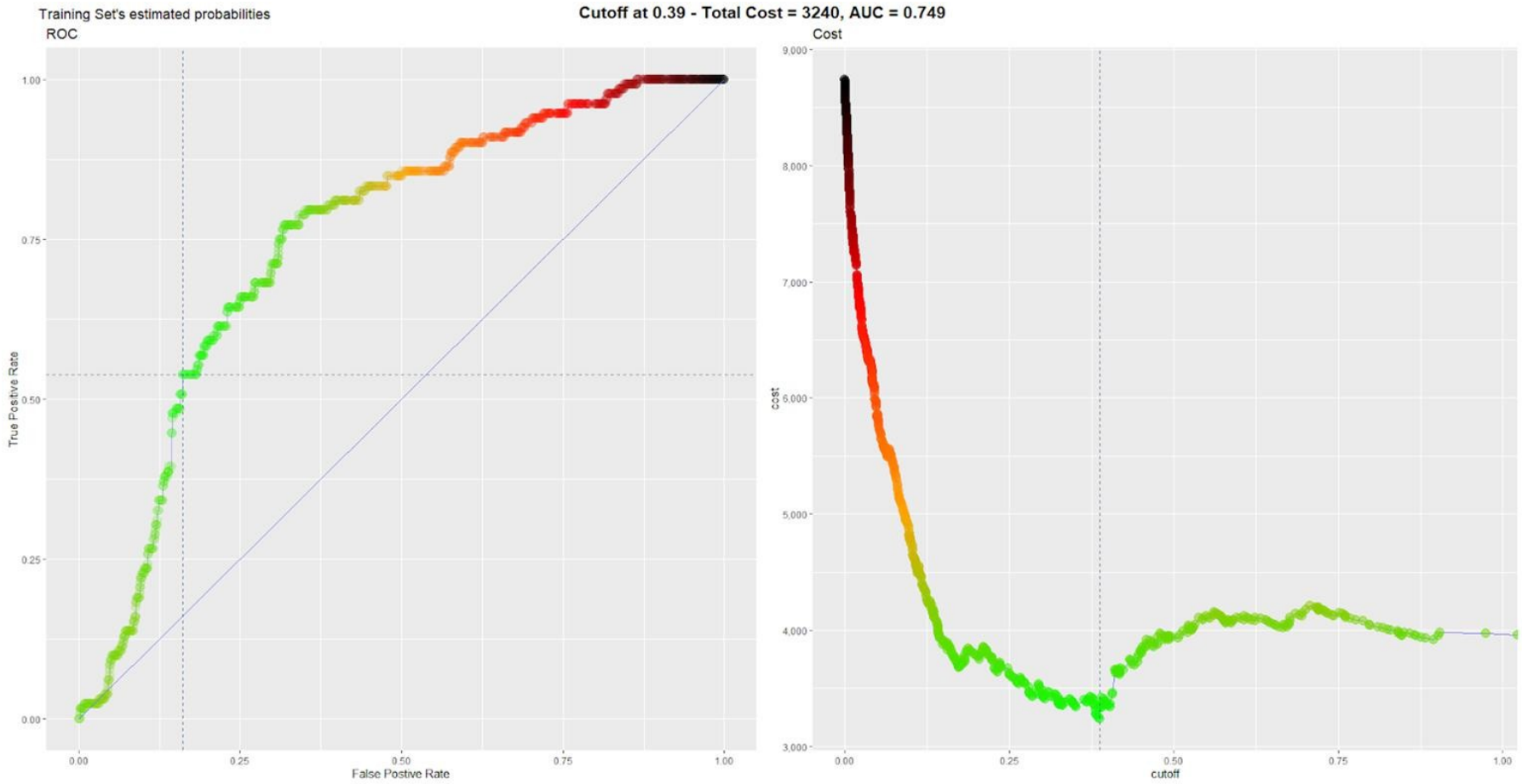
请访问 ProjectMarvel@G18Consulting.co

4.
附录

- 数据清理过程
- 逻辑回归的分类
- 模型评估指标



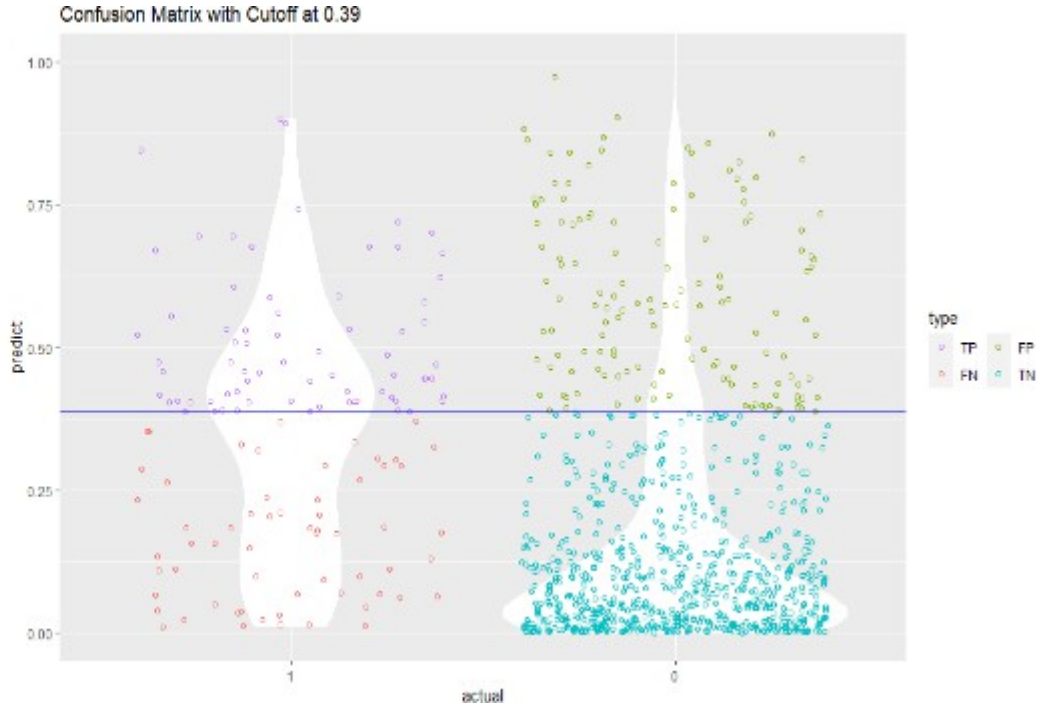
显示清理前后的缺失数据图。



根据 5 倍交叉验证方法，并假设错误预测主机成为超级主机的成本（假阴性）是未提前预测某些超级主机的成本（假阳性）的三倍，我们确定最佳截断点为 0.39，这可以使总成本最小化。这将使 AUC 达到 0.749。

	FALSE	TRUE
0	733	141
1	65	67

[1] The sensitivity is 0.5379
[2] The specificity is 0.8387
[3] The Misclassification Rate is 0.2008



可以得出的结论是，该模型具有合理的预测准确性，TP 率（灵敏度）为 53.79%，而总体误分类率为 20.08%。

```
Goodness Fit on the Models (Train/Test Split) with all cleaned variables:

Performance Metrics for Test Set
--------------------------------


Model 1: Logistic Regression on Feature Data (MSE): 0.11076
Model 1: Logistic Regression on Feature Data (R^2): 0.02835

Model 3: RandomForest Classification on Feature Data (MSE): 0.06917
Model 3: RandomForest Classification on Feature Data (R^2): 0.31843

Model 4: XGBoost Classification on Feature Data (MSE): 0.05237
Model 4: XGBoost Classification on Feature Datat (R^2): 0.31843


Performance Metrics for Train Set
---------------------------------
Model 1: Logistic Regression on Feature Data (R^2): 0.19819
Model 3: RandomForest Classification on Feature Data (R^2): 0.54083
Model 4: XGBoost Classification on Feature Data (R^2): 0.76639
```

我们使用 MSE 和 R^2 对本报告中使用的每个模型进行了评估。XGBoost 分类显示了最高的 R^2，最低的 MSE。