什么样的餐厅更受顾客的欢迎?

成员：张晓宇、陈锟、刘懋霖、刘妍、高伦

CONTENT

PART 01

数据获取与数据处理

01.数据获取与处理

Type	ID	Name	Address	ReviewNum	LN(Review	Level	P
粤菜	2954893	炳胜品味(	珠江新城冼	7145	8.874168	4	
粤菜	580743	阿一鲍鱼(	天河北路4	2328	7.752765	4.5	
粤菜	3500059	麓苑轩酒家	麓苑路36号	2934	7.984122	4.5	
粤菜	96379038	阿呆雷州洋	黄石东路4	16	2.772589	3.5	
粤菜	93559683	禄运茶居(	珠江新城花	293	5.680173	5	
粤菜	80614370	和苑酒家(	花城大道6	305	5.720312	5	
粤菜	27162190	點都德・德	中山四路2	6155	8.72502	5	
粤菜	56985236	點都德(花	珠江新城花	4026	8.300529	4.5	

基本属性

数据获取

北京大学开放研究数据平台

数据采集时间为2017年11月。
数据格式为csv。
共33个字段。

LN(Busine	Label anc	FunctionL	FlavorLab	Environme	DishesLab	ServiceLa	Cost-eff
7.96346	回头客（3	0.800591	0	0.059084	0	0.140325	C
7.963112	回头客（9	0.795918	0	0.077551	0	0.126531	C
6.079933	菜品健康	0.356061	0.510606	0.109091	0	0.024242	C
3.89182	味道赞（6	0	0.285714	0	0.52381	0.190476	C
3.73767	干净卫生	0.577465	0.070423	0.225352	0	0.126761	C
5.572154	回头客（2	0.5	0.170455	0.272727	0	0.056818	C
6.687109	回头客（3	0.815682	0	0.069246	0	0.115071	C
6.593045	回头客（2	0.802486	0	0.08011	0	0.117403	C
5.231109	回头客（5	0.813492	0	0.079365	0	0.107143	C
7.986165	回头客（3	0.808511	0	0.141844	0	0.049645	C
6.55108	高大上（1	0.546667	0.106667	0.293333	0	0.053333	C
7.839526	回头客（2	0.801724	0.077586	0.025862	0	0.094828	C
5.278115	味道赞（3	0.029412	0.294118	0.257353	0	0.25	0.169118
5.141664	回头客（3	0.729452	0	0.171233	0	0.099315	C
6.98379	回头客（8	0.838147	0	0.045319	0	0.116534	C
4.51086	回头客（8	0.913669	0	0.061151	0	0.02518	C
5.247024	回头客（4	0.576613	0.274194	0.084677	0	0.064516	C

01.数据获取与处理

A	B	C	D	E	F	G	H	I	J	K	L	M	N	
ApplauseF	ReviewNu	Level	FlavorScor	Environme	ServiceScc	5StarRevie	High-qua	PopularAr	PictureNu	ParkingNu	ParkingInf	GroupPur	Promotior	Ac
37.76	7145	4	7.8	8.6	7.3	2666	0	1	1995	193	1	0	0	
57.44	2328	4.5	8.6	8.2	8.5	1308	1	1	344	44	1	1	0	
56.05	2934	4.5	8.8	7.9	8.1	1644	1	0	723	15	1	1	1	
43.75	16	3.5	8.2	7.2	7.6	7	0	0	14	1	1	1	0	
54.95	293	5	9.1	9.1	8.8	161	1	1	238	0	0	1	0	
66.56	305	5	9.1	9.1	9.2	203	1	0	197	0	0	1	0	
58.57	6155	5	9.1	8.8	8.5	3605	1	1	2253	6	1	1	0	
55.49	4026	4.5	9	8.9	8.3	2234	1	1	1786	1	1	1	0	
48.1	1187	4.5	9	9	8.4	571	1	1	894	0	0	0	0	
59.14	1085	5	9	9.3	9.1	608	1	0	645	21	1	0	0	
66.55	281	5	9.1	9.3	9.1	187	1	1	204	1	1	0	0	
49.64	663	5	8.7	8.7	8.5	279	1	1	258	30	1	0	0	
57.86	140	5	9.2	8.7	8.9	81	1	1	128	0	0	1	0	
51.95	1588	4.5	8.5	9.1	8.8	825	1	1	1272	2	1	0	0	
57.31	13002	4.5	9	8.6	8.3	7452	1	1	3261	18	1	1	0	
54.67	1747	5	8.8	8.7	8.7	580	1	0	449	218	1	0	0	

数据处理

删除就非结构化数据，数据集共包含3124条餐厅数据，18个字段。

PART 02

数据探索

逐步回归法

逐步回归就是从自变量x中挑选出对y有显著影响的变量，已达到最优.
逐步回归分析是以AIC信息统计量为准则，通过选择最小的**AIC**信息统计量，来达到删除或增加变量的目的。

02. 数据探索

```
data1<-read.csv(file.choose())
a<-step(lm(ApplauseRate~.,data1))
mylm<-lm(formula = ApplauseRate ~ ReviewNum + FlavorScore
          + EnvironmentScore + ServiceScore + X5StarReviewNum
          + High.qualityMerchant + PopularArea + PictureNum
          + ParkingNum + ParkingInfo + GroupPurchase + TakeOut
          + PerConsumption + BusinessDay, data = data1)
summary(mylm)
```

代码截图

02. 数据探索

```
Call:
lm(formula = ApplauseRate ~ ReviewNum + FlavorScore + EnvironmentScore +
    ServiceScore + X5StarReviewNum + High.qualityMerchant + PopularArea +
    PictureNum + ParkingNum + ParkingInfo + GroupPurchase + TakeOut +
    PerConsumption + BusinessDay, data = data1)

Residuals:
    Min      1Q  Median      3Q     Max
-25.447  -5.937  -0.002   5.255  34.482

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         -3.667e+01  2.479e+00 -14.790  < 2e-16 ***
ReviewNum           -4.762e-03  4.341e-04 -10.968  < 2e-16 ***
FlavorScore          3.109e+00  4.507e-01   6.899 6.31e-12 ***
EnvironmentScore     1.793e+00  3.982e-01   4.503 6.95e-06 ***
ServiceScore         4.972e+00  5.229e-01   9.508  < 2e-16 ***
X5StarReviewNum      1.185e-02  8.938e-04  13.254  < 2e-16 ***
High.qualityMerchant 1.407e+01  3.777e-01  37.252  < 2e-16 ***
PopularArea         -9.507e-01  3.141e-01  -3.027  0.00249 **
PictureNum          -2.663e-03  6.336e-04  -4.202 2.72e-05 ***
ParkingNum           2.760e-02  4.337e-03   6.363 2.26e-10 ***
ParkingInfo         -1.040e+00  3.908e-01  -2.660  0.00784 **
GroupPurchase        3.158e+00  3.337e-01   9.463  < 2e-16 ***
TakeOut              8.474e-01  3.339e-01   2.538  0.01119 *
PerConsumption       1.392e-02  2.772e-03   5.020 5.44e-07 ***
BusinessDay         -1.876e-03  2.221e-04  -8.448  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

运行结果

02. 数据探索

```
Residual standard error: 8.361 on 3109 degrees of freedom
Multiple R-squared:  0.7402,    Adjusted R-squared:  0.739
F-statistic: 632.6 on 14 and 3109 DF,  p-value: < 2.2e-16
```

运行结果

02. 数据探索

决策树

决策树(Decision Tree）是在已知各种情况发生概率的基础上，通过构成决策树来求取净现值的期望值大于等于零的概率，评价项目风险，判断其可行性的决策分析方法，是直观运用概率分析的一种图解法。由于这种决策分支画成图形很像一棵树的枝干，故称决策树。在机器学习中，决策树是一个预测模型，他代表的是对象属性与对象值之间的一种映射关系。
Entropy = 系统的凌乱程度，使用算法ID3, C4.5和C5.0生成树算法使用熵。

决策树是一种树形结构，其中每个内部节点表示一个属性上的测试，每个分支代表一个测试输出，每个叶节点代表一种类别。

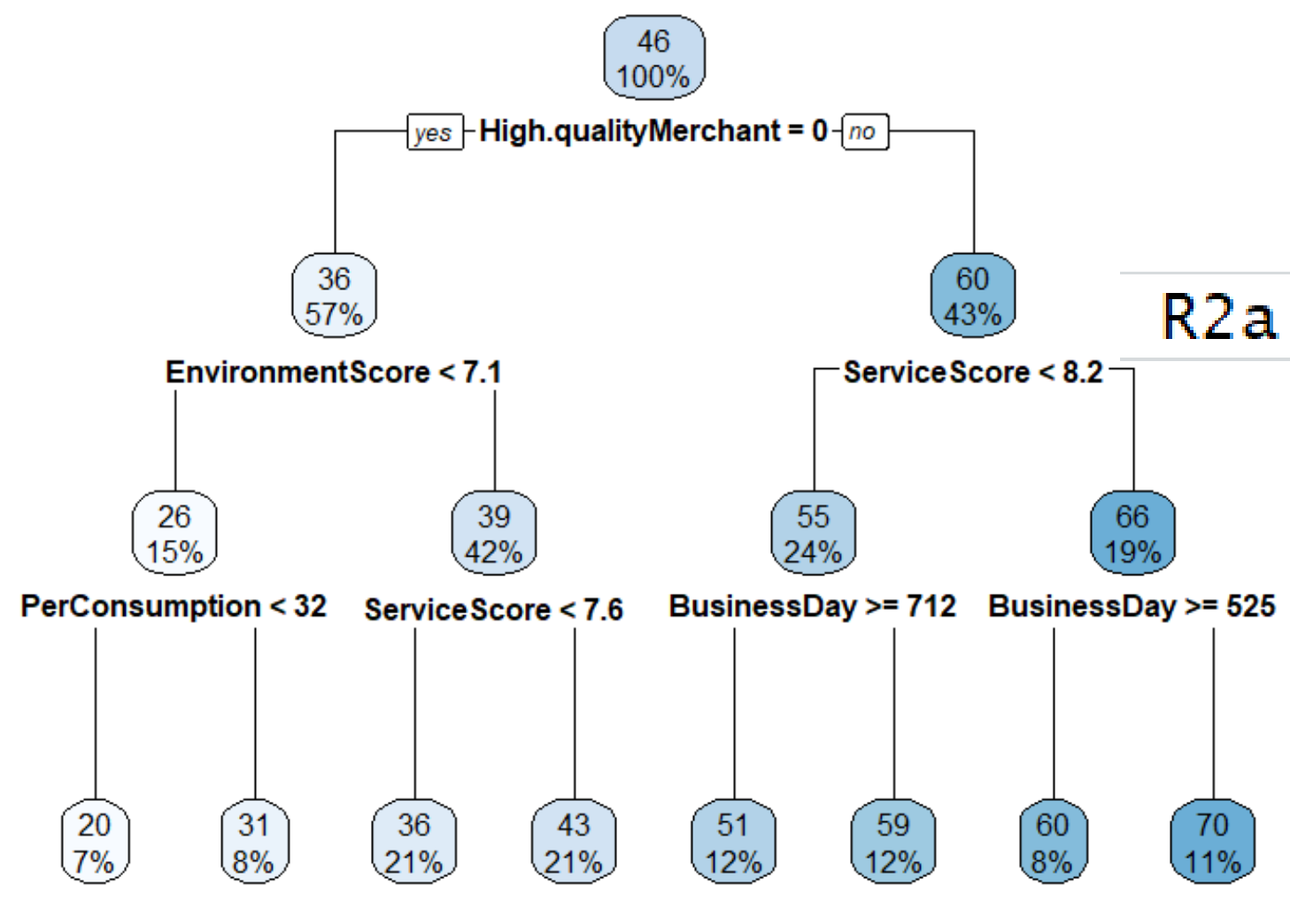
1. 将预测变量空间（$X_1, X_2, X_3, \cdots, X_p$）的可能取值构成的集合分割成$J$个互不重叠的区域$\{R_1, R_2, R_3, \cdots, R_J\}$
2. 对落入区域R_j的每个观测值作同样的预测，预测值等于R_j上训练集的各个样本取值的算术平均数。

02. 数据探索

```
10 library(rpart.plot)
11 a2<-rpart(ApplauseRate~.,data1)
12 rpart.plot(a2)
13 SST<-sum((data1[,1]-mean(data1[,1]))^2)
14 resa<-data1[,1]-predict(a2,data1[,-1])
15 SSEa<-sum(resa^2)
16 R2a<-1-SSEa/SST
```

代码截图

02. 数据探索



R2a	0.7167

运行结果

随机森林

随机森林实际上是一种特殊的bagging方法，它将决策树用作bagging中的模型。首先，用bootstrap方法生成m个训练集，然后，对于每个训练集，构造一颗决策树，在节点找特征进行分裂的时候，并不是对所有特征找到能使得指标（如信息增益）最大的，而是在特征中随机抽取一部分特征，在抽到的特征中间找到最优解，应用于节点，进行分裂。随机森林的方法由于有了bagging，也就是集成的思想在，实际上相当于对于样本和特征都进行了采样（如果把训练数据看成矩阵，就像实际中常见的那样，那么就是一个行和列都进行采样的过程），所以可以避免过拟合。

02. 数据探索

```
18 library(randomForest)
19 a3<-randomForest(ApplauseRate~.,data1,importance=T,localImp=T,proximity=T)
20 resa3<-data1[,1]-predict(a3,data1[,-1])
21 SSEa3<-sum(resa3^2)
22 R2a3<-1-SSEa3/SST
23 a3$importance
24 a3$rsq
25 varImpPlot(a3)
```

代码截图

02. 数据探索

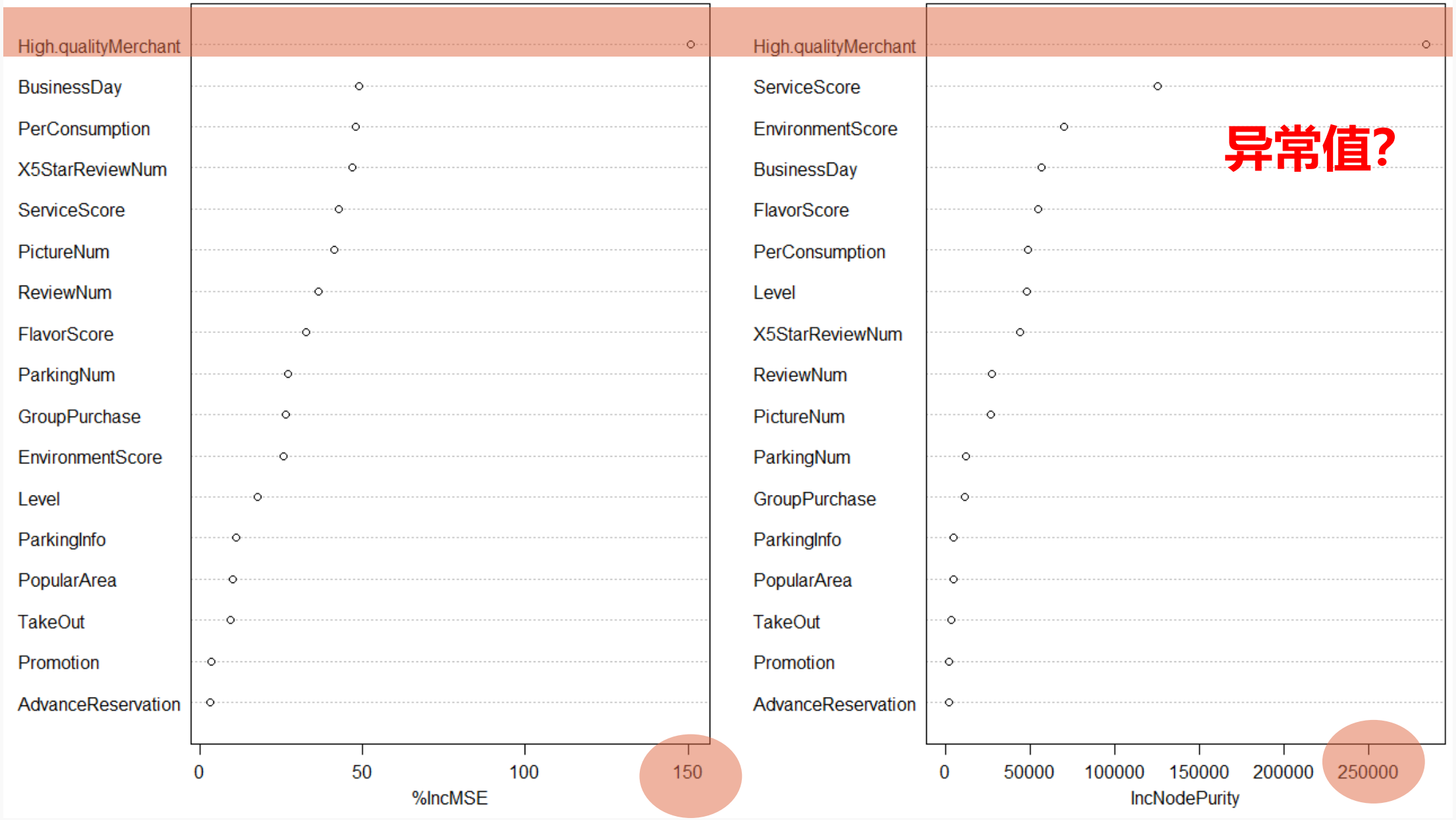
```
> a3$importance
                          %IncMSE IncNodePurity
ReviewNum               18.4002078     27112.355
Level                   12.3205198     47731.647
FlavorScore             14.1221755     54107.400
EnvironmentScore        15.2217935     69747.607
ServiceScore            38.7495988    125236.943
X5StarReviewNum         41.3581218
High.qualityMerchant   145.4085403
PopularArea              1.0055157
PictureNum              13.2799118
ParkingNum               6.4426378
ParkingInfo              2.5250448
GroupPurchase            4.1273328
Promotion                0.2344481
AdvanceReservation       0.1092980
TakeOut                  0.7716763
PerConsumption          12.4425928
BusinessDay             17.6126585     56384.299
```

```
> a3$rsq
  [1] 0.6266758 0.6615959 0.6734057 0.6705903 0.6854050 0.6850530 0.7097861
  [8] 0.7260304 0.7392103 0.7538398 0.7571637 0.7606196 0.7684423 0.7722690
 [15] 0.7749938 0.7802254 0.7836025 0.7866660 0.7879512 0.7910483 0.7919512
 [22] 0.7924317 0.7934278 0.7949615 0.7962869 0.7972582 0.7980035 0.7991154
 [29] 0.8001932 0.8011380 0.8015214 0.8023544 0.8035214 0.8037069 0.8041869
 [36] 0.8049500 0.8045845 0.8053515 0.8061276 0.8070709 0.8077929 0.8084038
 [43] 0.8082863 0.8086223 0.8091710 0.8093979 0.8095099 0.8091962 0.8094955
 [50] 0.8100698 0.8103435 0.8105650 0.8107273 0.8106756 0.8106997 0.8114677
 [57] 0.8119782 0.8119648 0.8121136 0.8122238 0.8122680 0.8122891 0.8123461
 [64] 0.8121375 0.8119249 0.8119273 0.8121939 0.8124976 0.8126002 0.8130760
 [71] 0.8132942 0.8136393 0.8137036 0.8137045 0.8136951 0.8137055 0.8138229
 [78] 0.8138434 0.8138256 0.8141375 0.8147946 0.8149429 0.8151461 0.8152057
```

R2a3 0.9645

运行结果

02. 数据探索


a3
High.qualityMerchant
BusinessDay
PerConsumption
X5StarReviewNum
ServiceScore
PictureNum
ReviewNum
FlavorScore
ParkingNum
GroupPurchase
EnvironmentScore
Level
ParkingInfo
PopularArea
TakeOut
Promotion
AdvanceReservation
0
50
100
150
%IncMSE
High.qualityMerchant
ServiceScore
EnvironmentScore
BusinessDay
FlavorScore
PerConsumption
Level
X5StarReviewNum
ReviewNum
PictureNum
ParkingNum
GroupPurchase
ParkingInfo
PopularArea
TakeOut
Promotion
AdvanceReservation
0
50000
100000
150000
200000
250000
IncNodePurity
异常值?

PART 03
数据分析

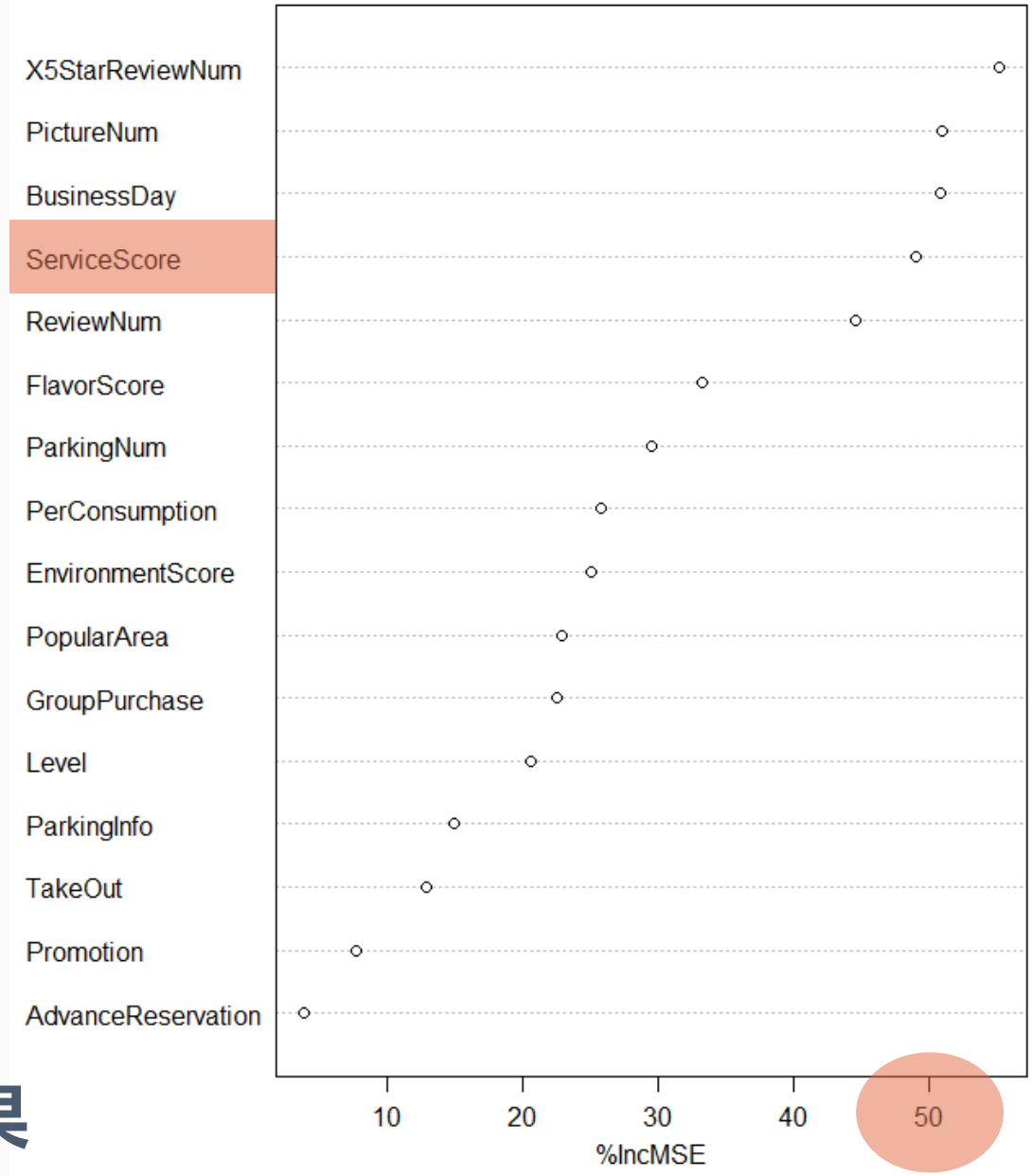
03. 数据分析

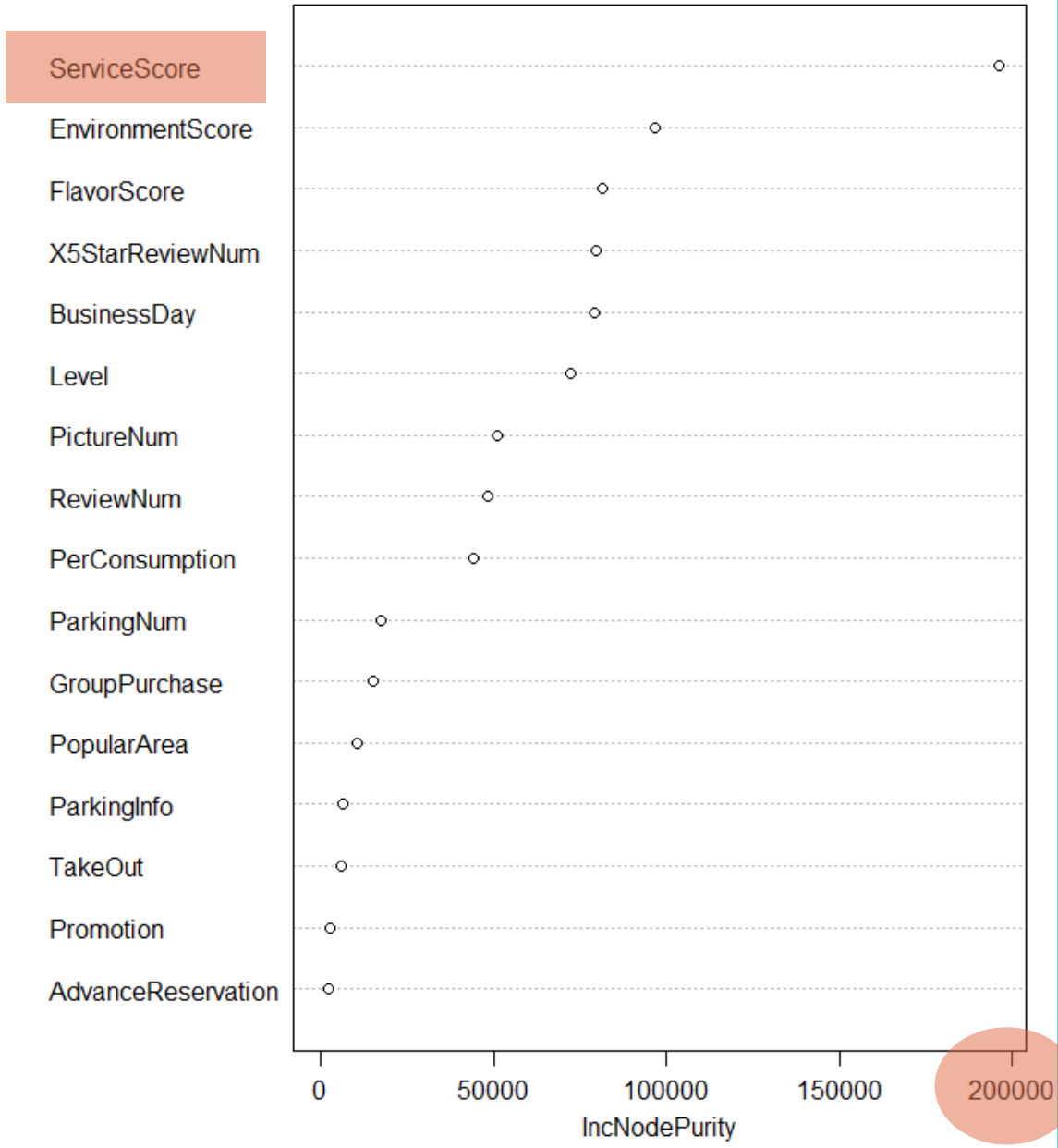
- 剔除异常值
- 随机森林回归

```
27 data2<-data1[,-8]
28 a4<-randomForest(ApplauseRate~.,data2,importance=T,localImp=T,proximity=T)
29 resa4<-data2[,1]-predict(a4,data2[,-1])
30 SSEa4<-sum(resa4^2)
31 R2a4<-1-SSEa4/SST
32 a4$importance
33 a4$rsq
34 varImpPlot(a4)
```

代码截图

03. 数据分析


X5StarReviewNum
PictureNum
BusinessDay
ServiceScore
ReviewNum
FlavorScore
ParkingNum
PerConsumption
EnvironmentScore
PopularArea
GroupPurchase
Level
ParkingInfo
TakeOut
Promotion
AdvanceReservation
10
20
30
40
50
%IncMSE


ServiceScore
EnvironmentScore
FlavorScore
X5StarReviewNum
BusinessDay
Level
PictureNum
ReviewNum
PerConsumption
ParkingNum
GroupPurchase
PopularArea
ParkingInfo
TakeOut
Promotion
AdvanceReservation
0
50000
100000
150000
200000
IncNodePurity

运行结果

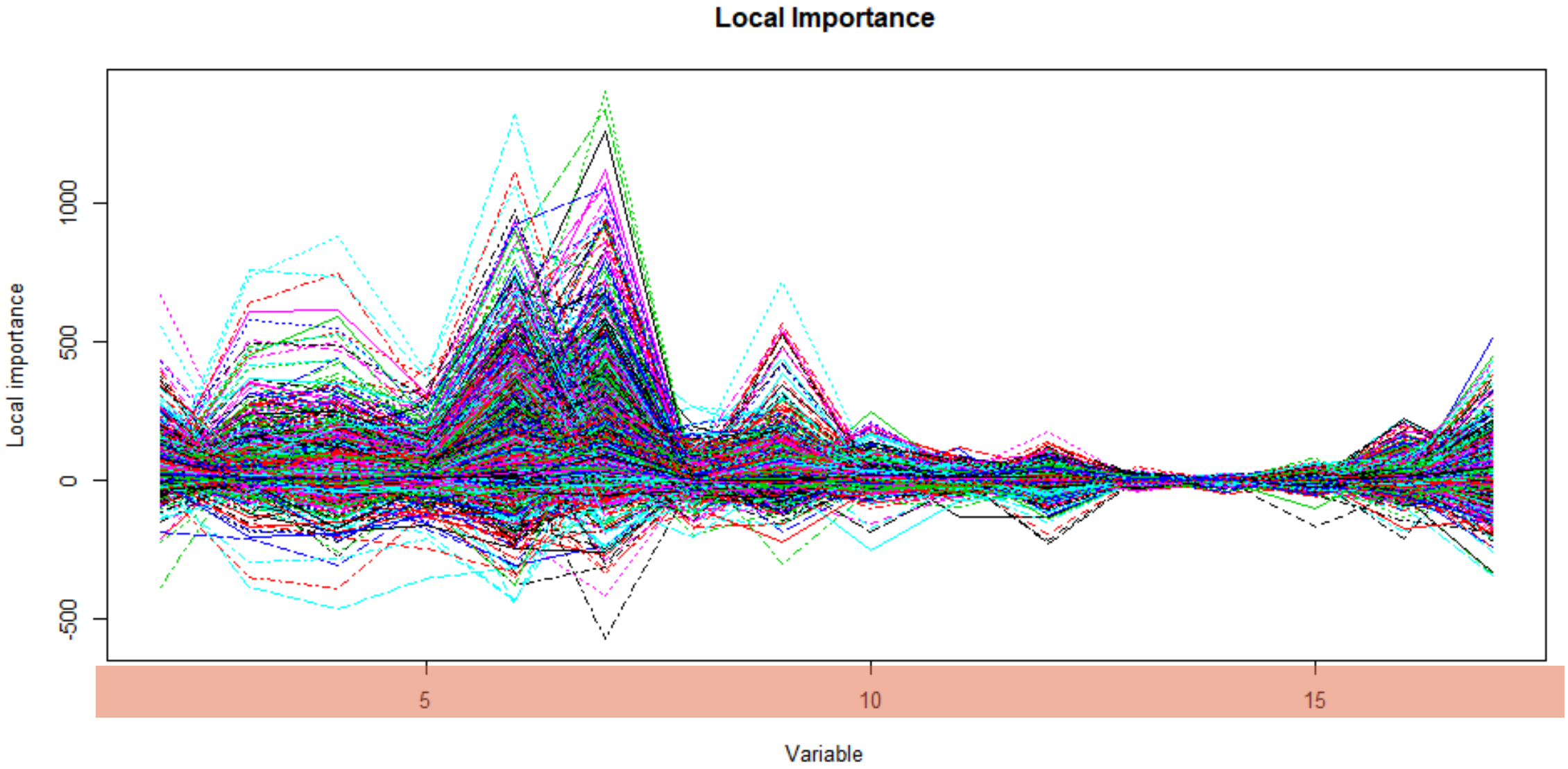
03. 数据分析

拟合变量的局部重要性

```
36 layout(matrix(c(1,2,3,3),nrow = 2,b=T))
37 for(i in 1:2)
38 {
39   title(colnames(a4$importance)[i])
40 }
41 matplot(2:17,a4$local,type = "l",xlab = "Variable",
42         ylab = "Local importance",main="Local Importance")
43
```

代码截图

03. 数据分析


Local Importance
Local importance
1000
500
0
-500
5
10
15
Variable

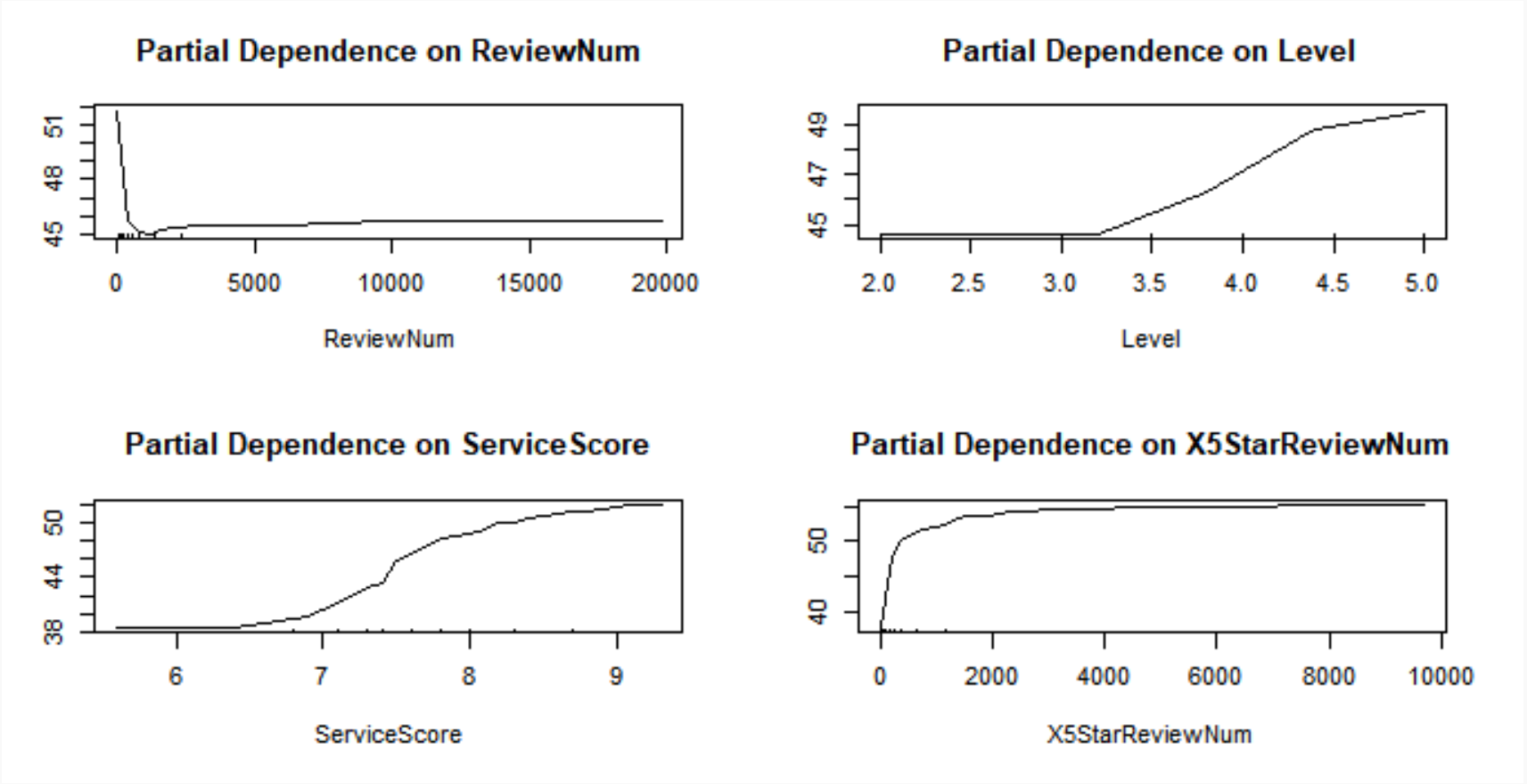
03. 数据分析

- 部分依赖图

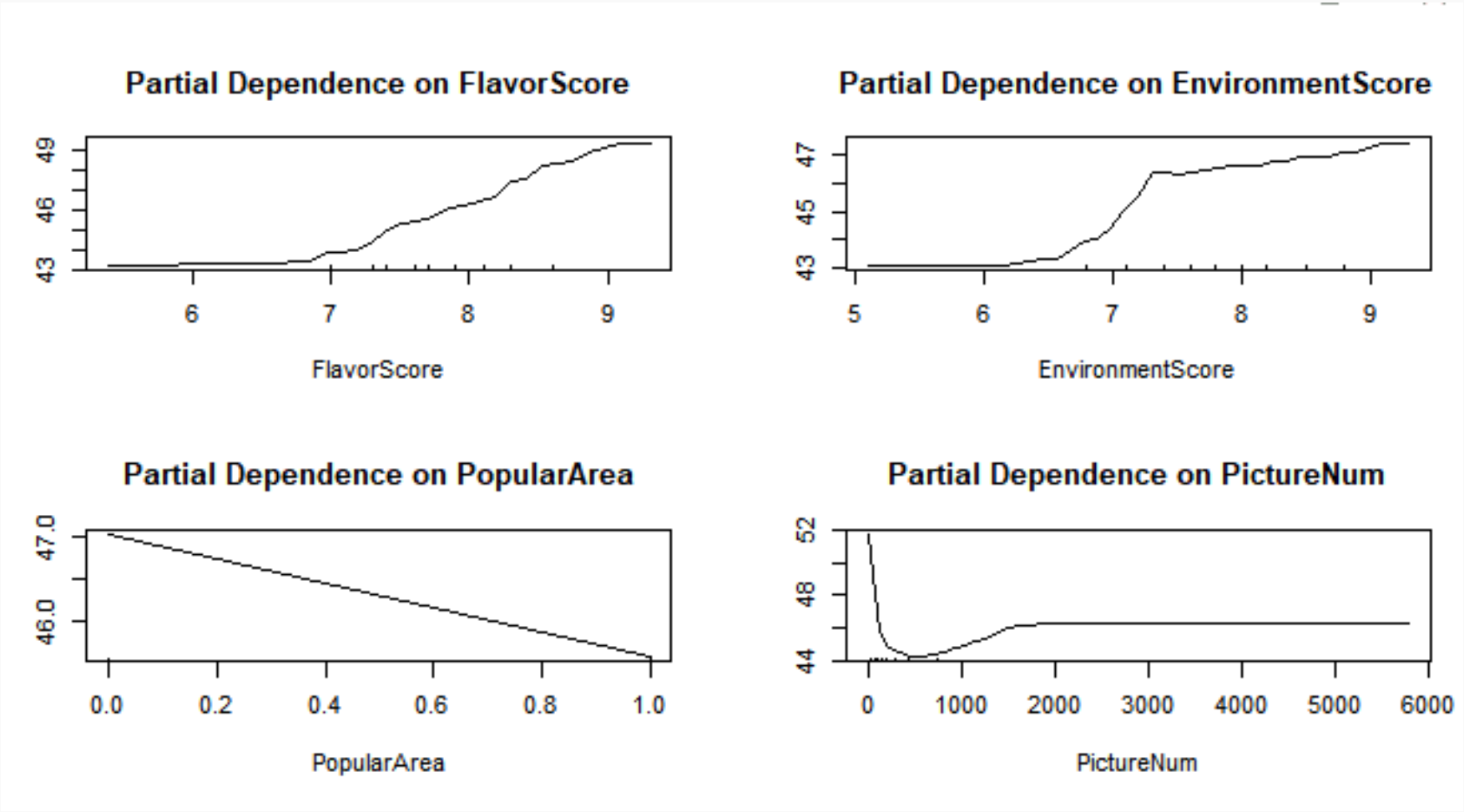
```
42
43 par(mfrow=c(4,4))
44 partialPlot(a4,pred.data = data2,ReviewNum);partialPlot(a4,pred.data = data2,Level)
45 partialPlot(a4,pred.data = data2,FlavorScore);partialPlot(a4,pred.data = data2,EnvironmentScore)
46 partialPlot(a4,pred.data = data2,ServiceScore);partialPlot(a4,pred.data = data2,X5StarReviewNum)
47 partialPlot(a4,pred.data = data2,PopularArea);partialPlot(a4,pred.data = data2,PictureNum)
48 partialPlot(a4,pred.data = data2,ParkingNum);partialPlot(a4,pred.data = data2,ParkingInfo)
49 partialPlot(a4,pred.data = data2,GroupPurchase);partialPlot(a4,pred.data = data2,Promotion)
50 partialPlot(a4,pred.data = data2,AdvanceReservation);partialPlot(a4,pred.data = data2,TakeOut)
51 partialPlot(a4,pred.data = data2,PerConsumption);partialPlot(a4,pred.data = data2,BusinessDay)
52
```

代码截图

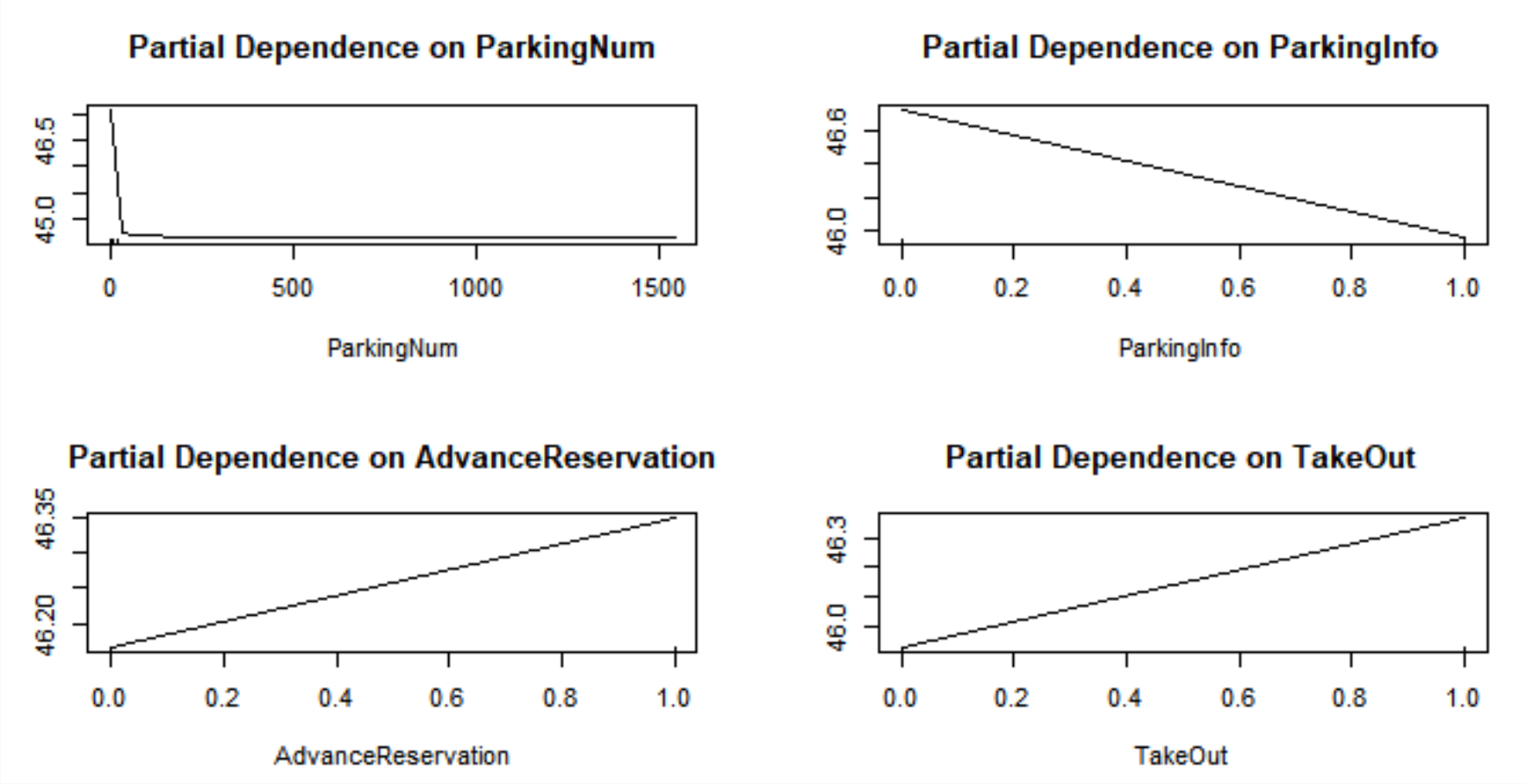
03. 数据分析


Partial Dependence on ReviewNum
51
48
45
0
5000
10000
15000
20000
ReviewNum
Partial Dependence on Level
49
47
45
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Level
Partial Dependence on ServiceScore
50
44
38
6
7
8
9
ServiceScore
Partial Dependence on X5StarReviewNum
50
40
0
2000
4000
6000
8000
10000
X5StarReviewNum

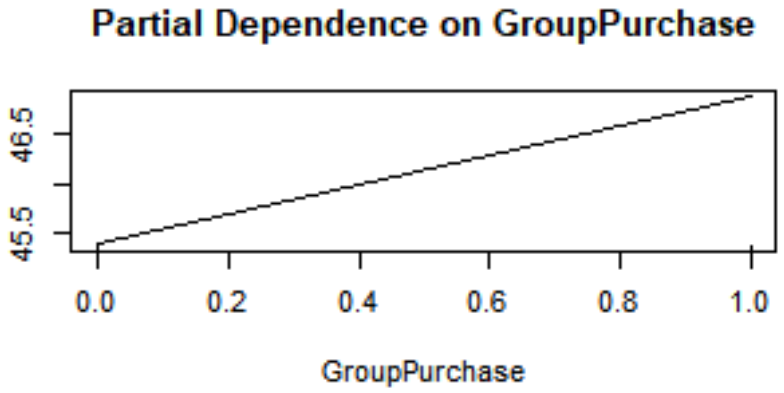
03. 数据分析

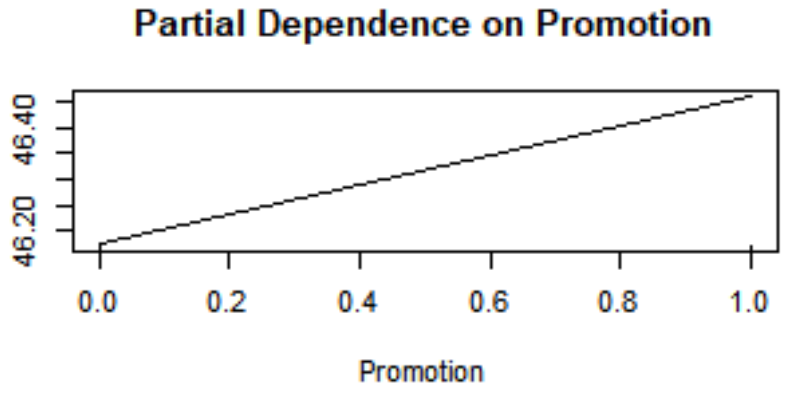

Partial Dependence on FlavorScore
43
46
49
6
7
8
9
FlavorScore
Partial Dependence on EnvironmentScore
43
45
47
5
6
7
8
9
EnvironmentScore
Partial Dependence on PopularArea
46.0
47.0
0.0
0.2
0.4
0.6
0.8
1.0
PopularArea
Partial Dependence on PictureNum
44
48
52
0
1000
2000
3000
4000
5000
6000
PictureNum

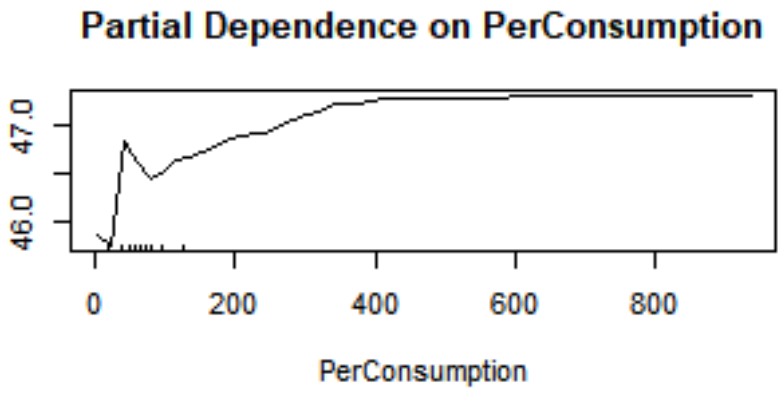
03. 数据分析

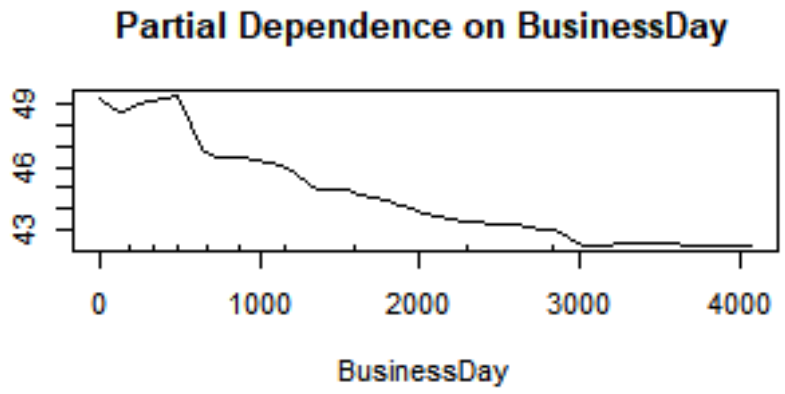

Partial Dependence on ParkingNum
46.5
45.0
0
500
1000
1500
ParkingNum
Partial Dependence on ParkingInfo
46.6
46.0
0.0
0.2
0.4
0.6
0.8
1.0
ParkingInfo
Partial Dependence on AdvanceReservation
46.35
46.20
0.0
0.2
0.4
0.6
0.8
1.0
AdvanceReservation
Partial Dependence on TakeOut
46.3
46.0
0.0
0.2
0.4
0.6
0.8
1.0
TakeOut

03. 数据分析


Partial Dependence on GroupPurchase
46.5
45.5
0.0
0.2
0.4
0.6
0.8
1.0
GroupPurchase


Partial Dependence on Promotion
46.40
46.20
0.0
0.2
0.4
0.6
0.8
1.0
Promotion


Partial Dependence on PerConsumption
47.0
46.0
0
200
400
600
800
PerConsumption


Partial Dependence on BusinessDay
49
46
43
0
1000
2000
3000
4000
BusinessDay

04. 分析结果

经过上述分析，最终得出以下四项最能代表餐厅受欢迎程度：

ServiceScore;

flavourScore;

businessday;

X5starReviewnum.

THANKS!