Bike Sharing 大数据分析

自行车租赁行为关联分析

自在随行，乐享心动

Comfort your life, move your heart

CONTENT

项目简述

instant	dteday	season	yr	mnth	hr	holiday	weekday	workingda	weathersi	temp	atemp	hum	windspeed	casual	registere	cnt
1	2011/1/1	1	0	1	0	0	6	0	1	0.24	0.2879	0.81	0	3	13	16
2	2011/1/1	1	0	1	1	0	6	0	1	0.22	0.2727	0.8	0	8	32	40
3	2011/1/1	1	0	1	2	0	6	0	1	0.22	0.2727	0.8	0	5	27	32
4	2011/1/1	1	0	1	3	0	6	0	1	0.24	0.2879	0.75	0	3	10	13
5	2011/1/1	1	0	1	4	0	6	0	1	0.24	0.2879	0.75	0	0	1	1
6	2011/1/1	1	0	1	5	0	6	0	2	0.24	0.2576	0.75	0.0896	0	1	1
7	2011/1/1	1	0	1	6	0	6	0	1	0.22	0.2727	0.8	0	2	0	2
8	2011/1/1	1	0	1	7	0	6	0	1	0.2	0.2576	0.86	0	1	2	3
9	2011/1/1	1	0	1	8	0	6	0	1	0.24	0.2879	0.75	0	1	7	8
10	2011/1/1	1	0	1	9	0	6	0	1	0.32	0.3485	0.76	0	8	6	14
11	2011/1/1	1	0	1	10	0	6	0	1	0.38	0.3939	0.76	0.2537	12	24	36
12	2011/1/1	1	0	1	11	0	6	0	1	0.36	0.3333	0.81	0.2836	26	30	56
13	2011/1/1	1	0	1	12	0	6	0	1	0.42	0.4242	0.77	0.2836	29	55	84
14	2011/1/1	1	0	1	13	0	6	0	2	0.46	0.4545	0.72	0.2985	47	47	94
15	2011/1/1	1	0	1	14	0	6	0	2	0.46	0.4545	0.72	0.2836	35	71	106
16	2011/1/1	1	0	1	15	0	6	0	2	0.44	0.4394	0.77	0.2985	40	70	110
17	2011/1/1	1	0	1	16	0	6	0	2	0.42	0.4242	0.82	0.2985	41	52	93

Bike Sharing项目

聚焦美国华盛顿地区，自行车**共享租赁频数**与**环境**、**季节**因素密切关联

数据可视化？

租赁数据！**时间**序列！**散点**展示！

预测模型构建？

非线性问题！**决策树**模型！**评估**变量重要性！

项目简述

数据可视化

多因素租赁高峰时段分布图

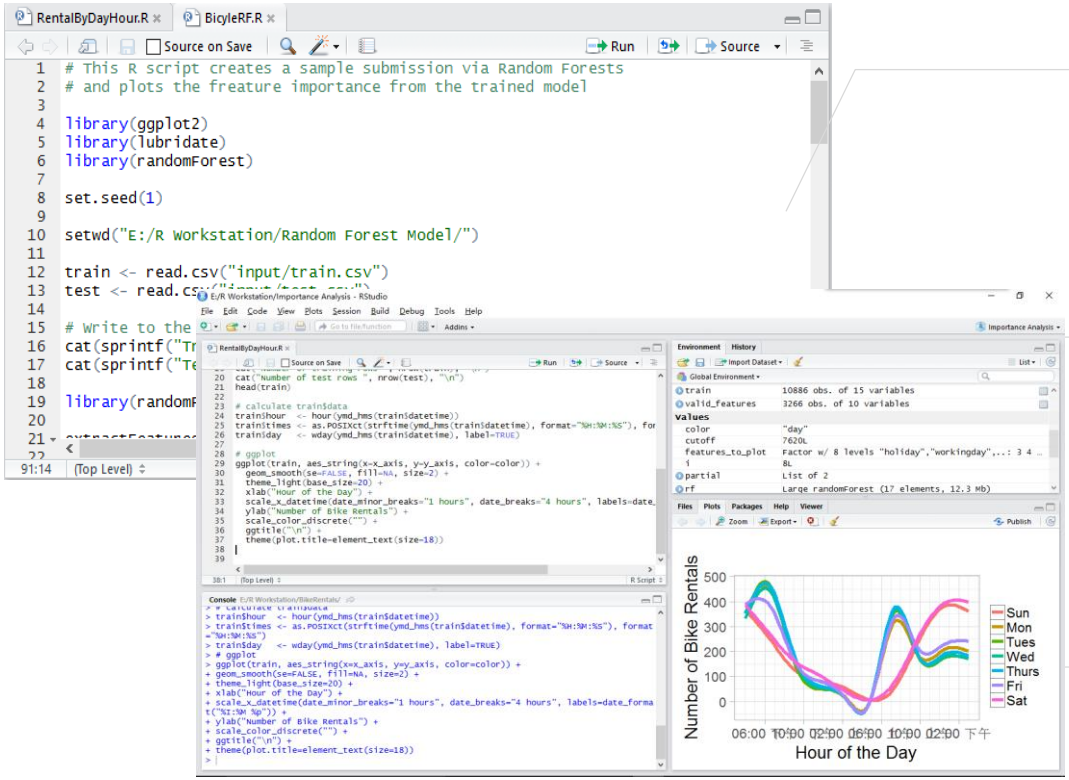
预测模型评估

预测自行车实时租赁信息，对调度、保养工作提供建议

影响因子联系

评估各影响因子的重要性，为模型分类提供直观表述

项目简述

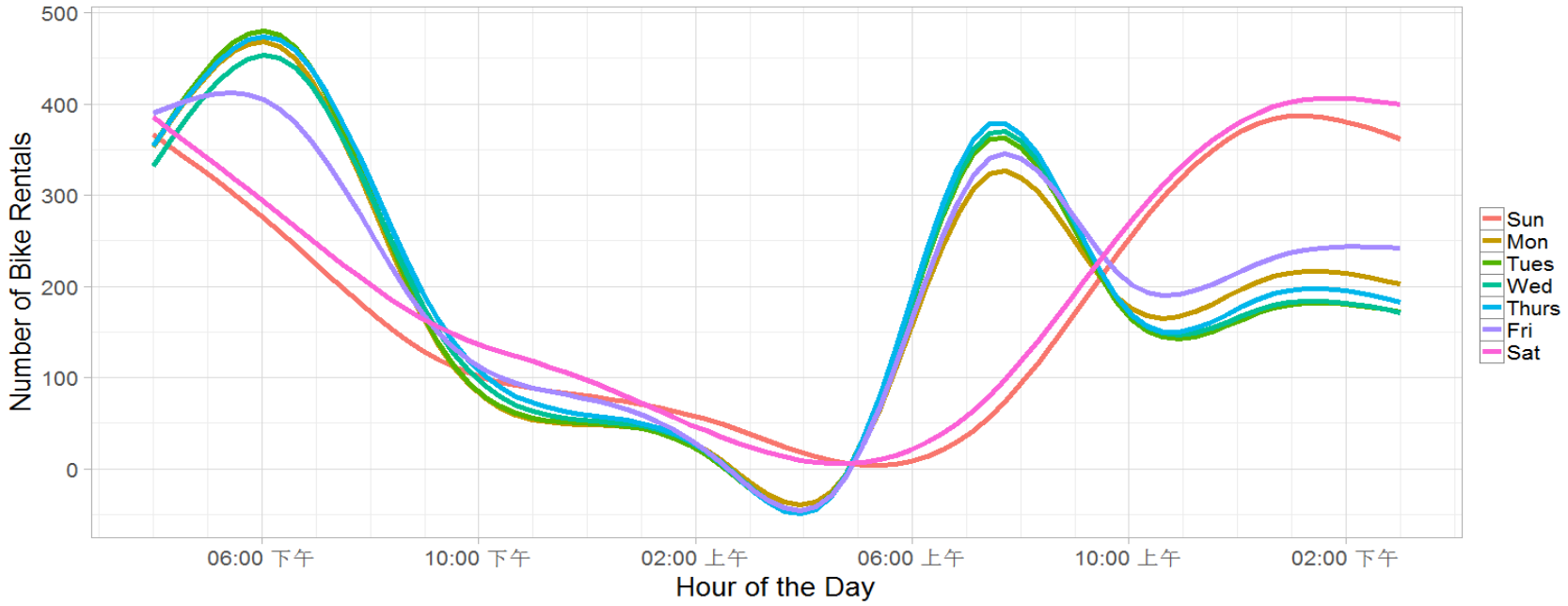


- R Studio
- 随机森林算法
- ggplot2制图

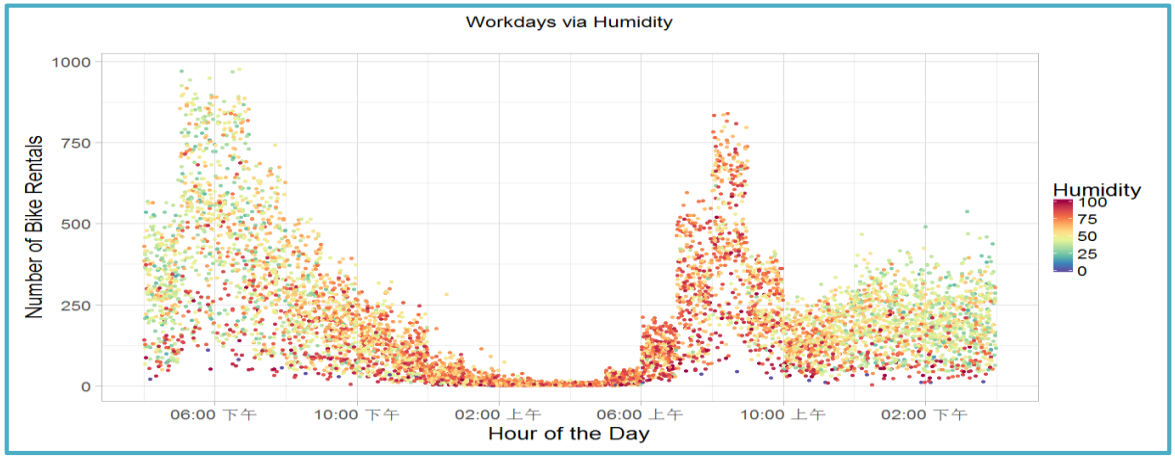
分析概述

工作日：早晚通勤租赁高峰

节假日：白天游玩租赁高峰



分析概述

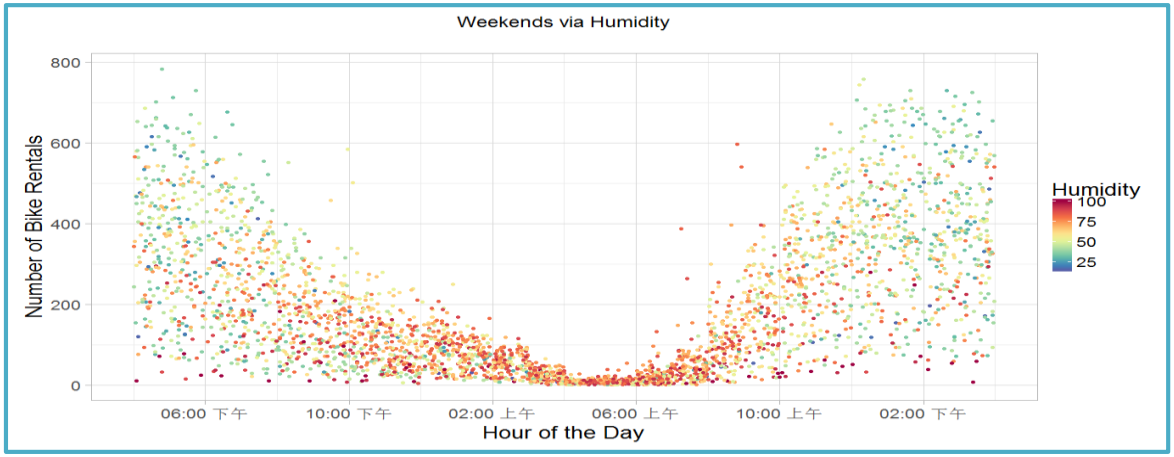


与体感温度不同，散点聚类

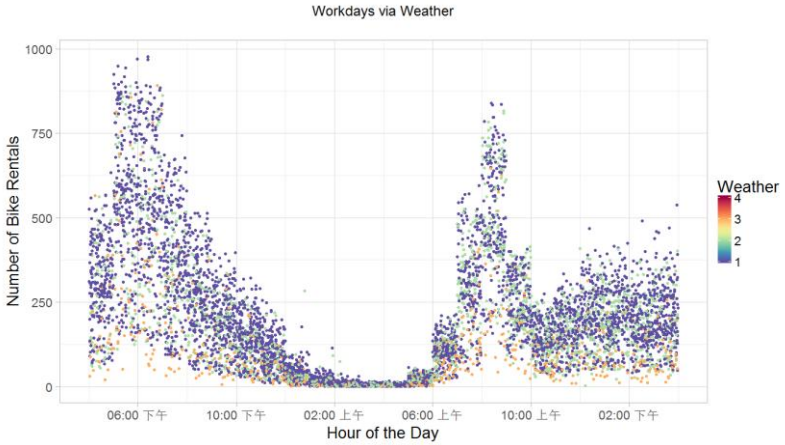
并未存在明显的“分界”

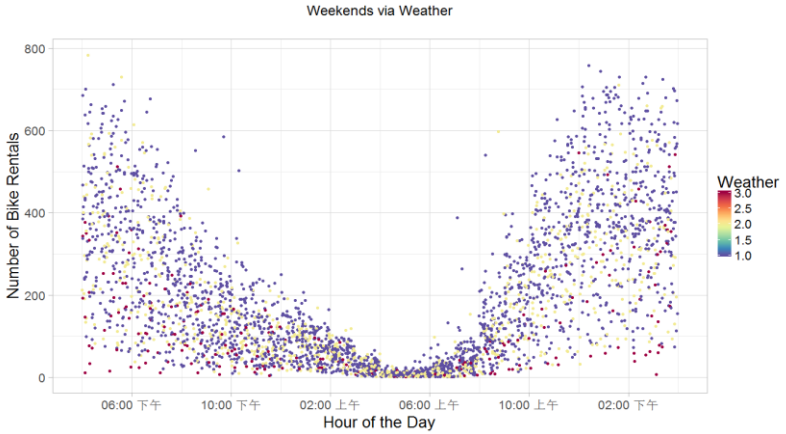
通勤影响依然存在

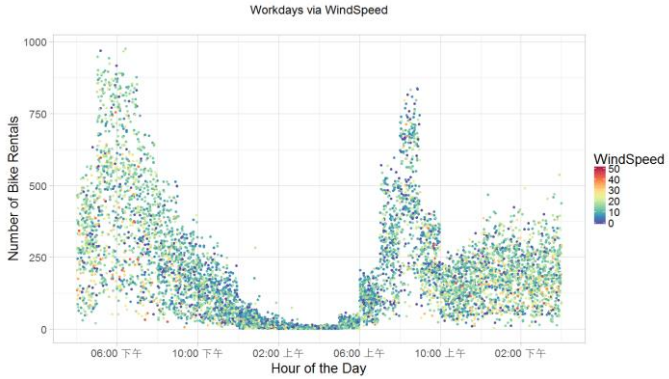
颜色对比度？调整筛选条件后，50~75数值较高人数

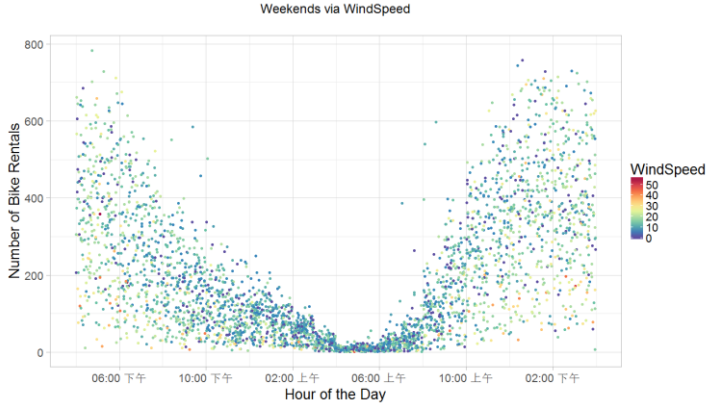


分析概述


Workdays via Weather
Number of Bike Rentals
1000
750
500
250
0
06:00 下午
10:00 下午
02:00 上午
06:00 上午
10:00 上午
02:00 下午
Hour of the Day
Weather
4
3
2
1


Weekends via Weather
Number of Bike Rentals
800
600
400
200
0
06:00 下午
10:00 下午
02:00 上午
06:00 上午
10:00 上午
02:00 下午
Hour of the Day
Weather
3.0
2.5
2.0
1.5
1.0


Workdays via WindSpeed
Number of Bike Rentals
1000
750
500
250
0
06:00 下午
10:00 下午
02:00 上午
06:00 上午
10:00 上午
02:00 下午
Hour of the Day
WindSpeed
50
40
30
20
10
0


Weekends via WindSpeed
Number of Bike Rentals
800
600
400
200
0
06:00 下午
10:00 下午
02:00 上午
06:00 上午
10:00 上午
02:00 下午
Hour of the Day
WindSpeed
50
40
30
20
10
0

数据模型

线性回归分析

类型变量特征提取，实数变量归一处理

回归模型构建，残差检测，置信预测评估

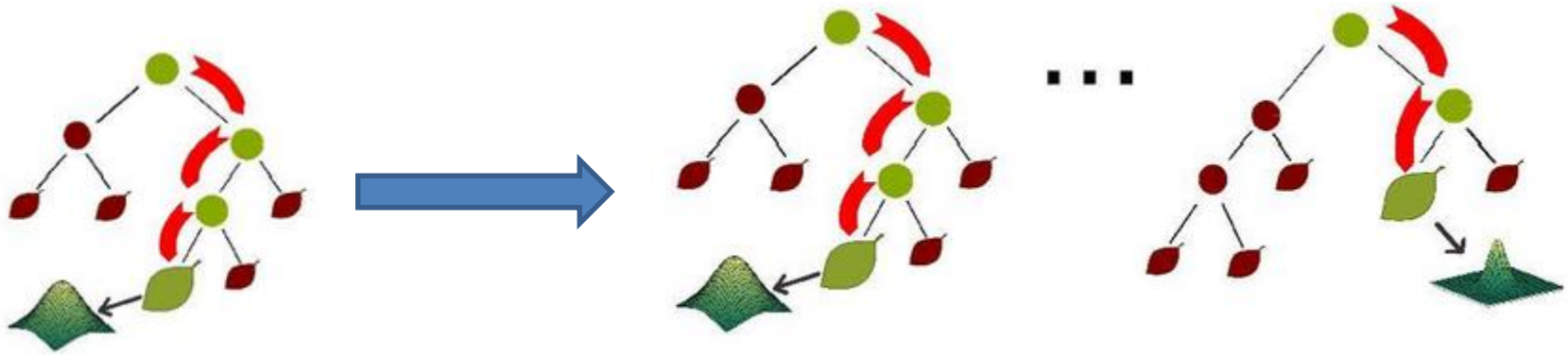
初期选取numpy提供的库，（非训练样本）手动参数调优

决策树随机森林

非线性问题，极值的抗干扰性，局部结构分析

分类过程，评估变量的重要性

直接选取R提供的randomForest包



Algorithm 6: $\mathbf{L}_K$-TreeBoost

$F_{k0}(\mathbf{x}) = 0, \quad k = 1, K$

For $m = 1$ to M do:

$p_k(\mathbf{x}) = \exp(F_k(\mathbf{x})) / \sum_{l=1}^{K} \exp(F_l(\mathbf{x})), \quad k = 1, K$

For $k = 1$ to K do:

$\tilde{y}_{ik} = y_{ik} - p_k(\mathbf{x}_i), \quad i = 1, N$

$\{R_{jkm}\}_{j=1}^{J} = J$-terminal node $tree(\{\tilde{y}_{ik}, \mathbf{x}_i\}_1^N)$

$\gamma_{jkm} = \frac{K-1}{K} \frac{\sum_{\mathbf{x}_i \in R_{jkm}} \tilde{y}_{ik}}{\sum_{\mathbf{x}_i \in R_{jkm}} |\tilde{y}_{ik}| (1 - |\tilde{y}_{ik}|)}, \quad j = 1, J$

$F_{km}(\mathbf{x}) = F_{k,m-1}(\mathbf{x}) + \sum_{j=1}^{J} \gamma_{jkm} 1(\mathbf{x} \in R_{jkm})$

endFor

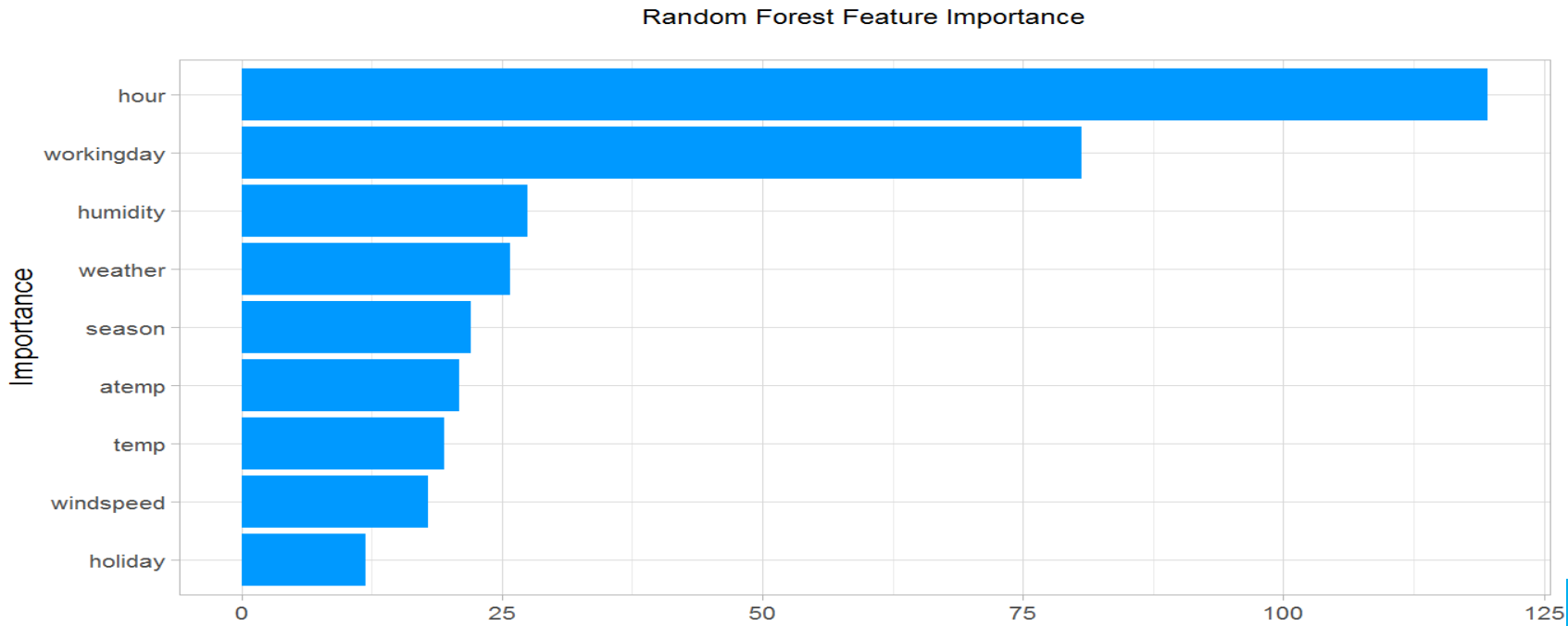
endFor

end Algorithm

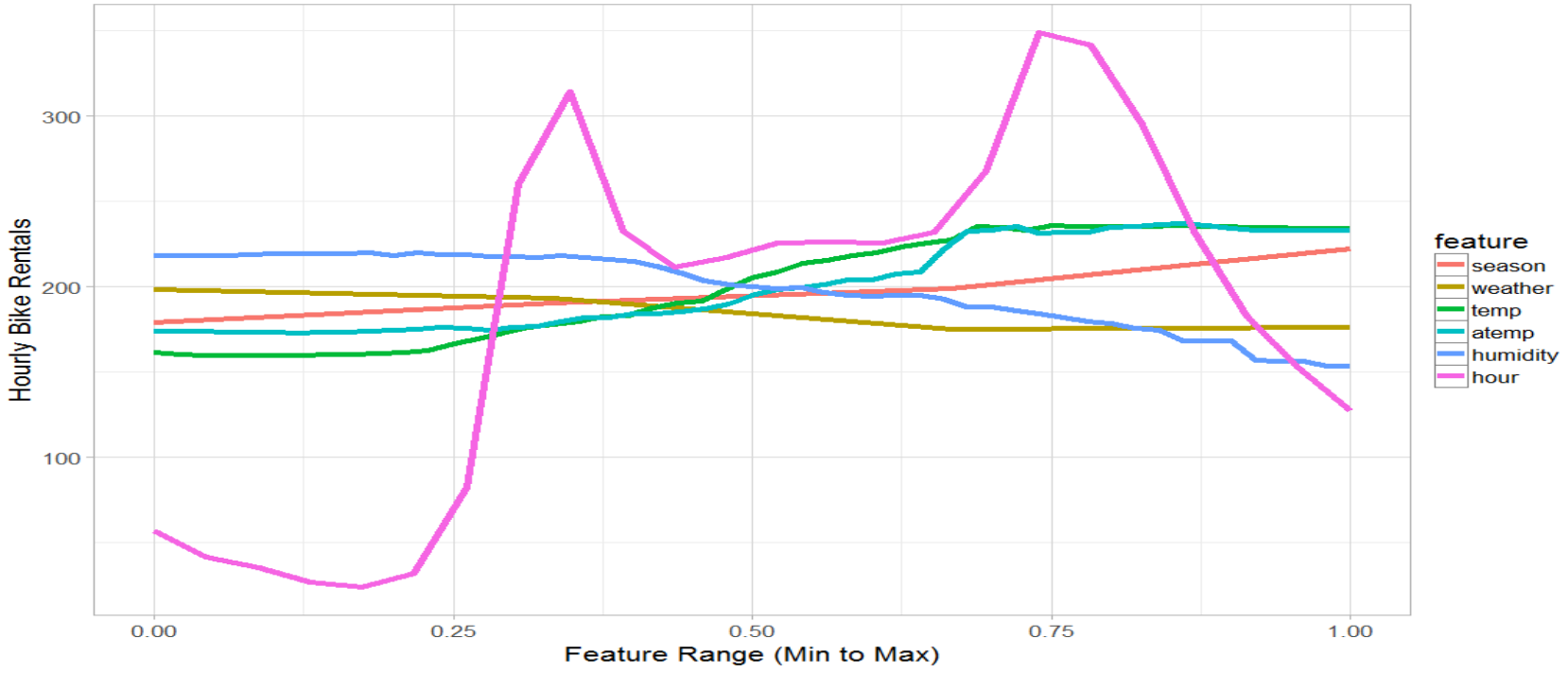
内部决策树，多决策树投票分类

行方向放回抽样　　：训练集
咧方向无放回抽样　：切分点

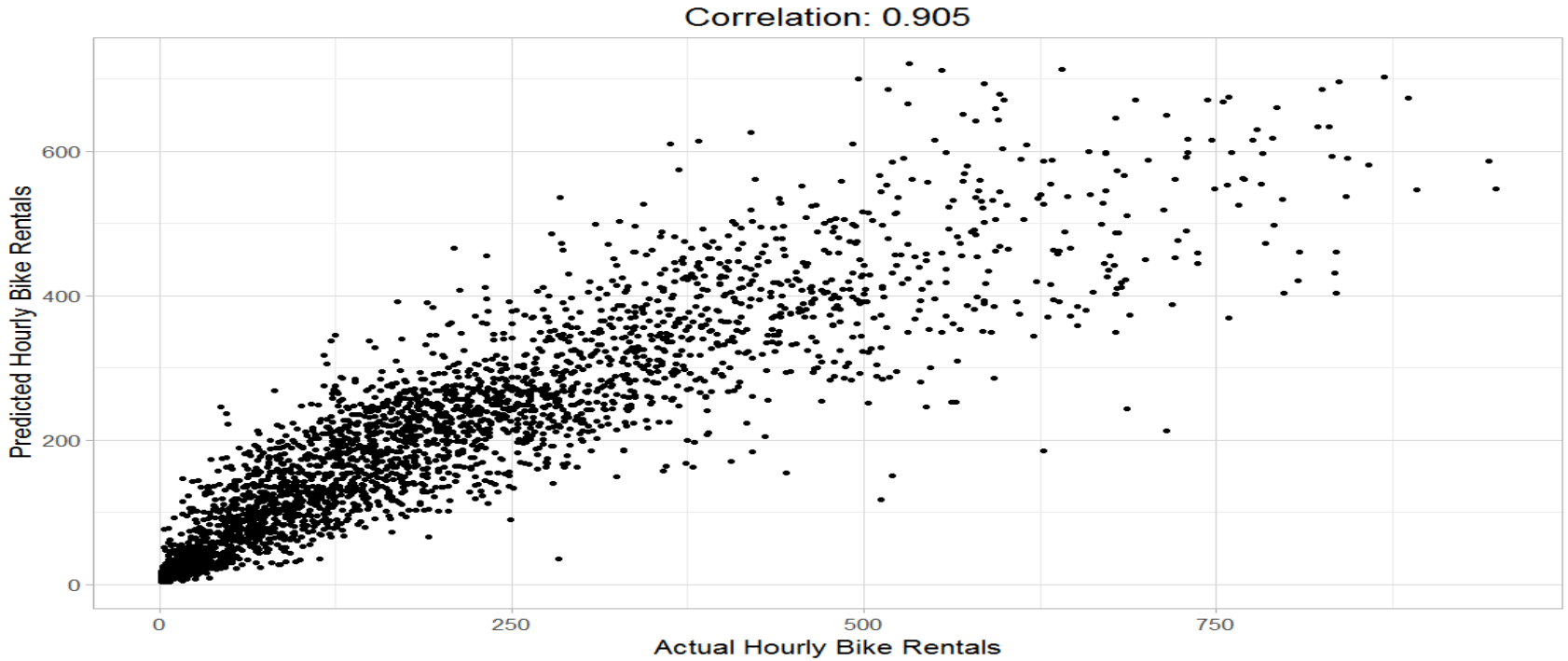
Random Forest Importance：评估自行租赁需求的影响因素



Marginal Effect：评估各自变量对因变量的影响



Marginal Effect：评估各自变量对因变量的影响



评估扩展

答辩分析提供了预测模型和影响因子的重要性评估

决策树在一定程度上上屏蔽了“噪声”数据，却可能使得部分“小数据结论”无法展示

答辩分析仍能继续扩展，数据集由华盛顿转移至北京校园的OFO模式，添加空间、骑行时间等特征，进一部完成模型的搭建和扩展

评估扩展

可视化租赁高峰时段分布图，直观感受租赁系统动态变化

有效预测车辆实时租赁信息，对公共自行车的调度工作、保养调整工作

动态调整租赁策略，规避租赁高峰