Data Science

Final Report

105703057 資科五王靖文 106304003 統計五 鄭以湉 107207438 統計四 黃大瑋

Outline

1. Problem Background
2. Data Source
3. Method Introduction
4. Model Processing
5. Conclusion

Problem Background

How to manage bike and employee?

Data Source

datetime - hourly date + timestamp

season - 1 = spring, 2 = summer, 3 = fall, 4 = winter

holiday - whether the day is considered a holiday/ workingday - whether the day is neither a weekend nor holiday

weather - 1: Clear,Partly cloudy/ 2: Mist + Cloudy, Mist/ 3: Light Snow, Light Rain + Scattered clouds/ 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

temp - temperature in Celsius

atemp - "feels like" temperature in Celsius

humidity - relative humidity

windspeed - wind speed

count - number of total rentals (Dependent Varible)

Method Introduction

random forest :隨機森林其實就是進階版的決策樹，所謂的森林就是由很多棵決策樹所組成。隨機森林是使用 Bagging 加上隨機特徵採樣的方法所產生出來的整體學習演算法。

lasso: 迴歸的變形是引入正則化 regularization (i.e., shrinkage)的技巧，將迴歸的權重和給予限制，藉此「限制模型的複雜度」，解決 overfitting的問題。

Xgboost: Gradient Boosting Decision Tree的改良版。其中Gradient Boosting Machine是以Tree-based 為主，將數百個弱決策樹(CART)，跟梯度下降法和 Boosting 結合在一起。

Model Processing

First step

Null Model: Predict all testing data by the mean of training data.

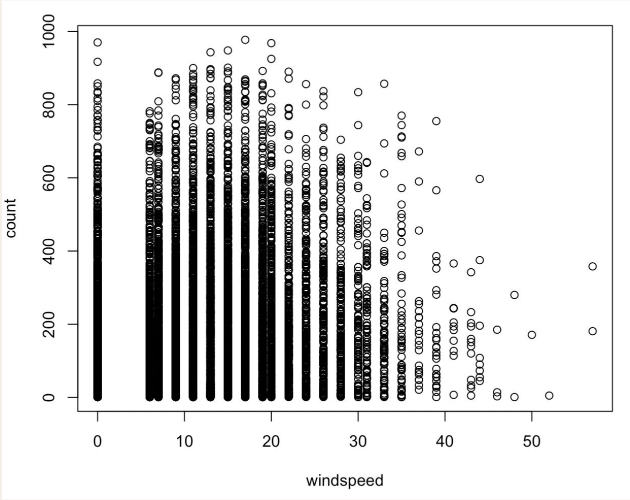
Way 1 (Spilt time):

Deal with the “date” variable by separating it to three variables, month, weekday and hour. Also, we use 3 times standard deviation to recongnize outliner and delete them.

Model Processing

Way 2 (Outliner):

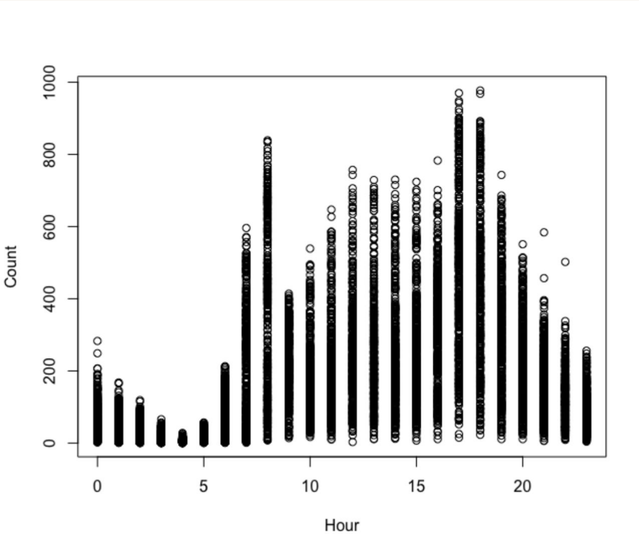
By EDA finding, we find there some strange value in windspeed variable. Since, windspeed is impossible to be equal to 0, we decided to train a model by “windspeed >0” and replace “windspeed=0” by model’s prediction.



Model Processing

Way 3 (Count hourly rank):

By EDA finding, we find there are obvious changes between different hours. There we decided to group hour by their rank of average count.



Conclusion-RMSE Table

RMSE Table

	dataset1	dataset2	dataset3
Random Forest	0.65195	0.64199	0.53741
Xgboost	0.84648	0.77997	0.78114
Lasso	2.65298	1.22353	1.04907

dataset1= split-time
dataset2=split-time, outliner
dataset3=split-time, outliner, time-rank

Conclusion-Xgboost factor contribution

```
> xgb.importance(colnames(dtrain1), model = xgb.model1)
```

	Feature	Gain	Cover	Frequency
1:	hour	0.615860013	0.149390264	0.188677845
2:	temp	0.093897934	0.150789130	0.147944799
3:	workingday	0.083391603	0.011555129	0.034055498
4:	atemp	0.051783455	0.123972105	0.075233714
5:	humidity	0.044005438	0.227609021	0.170425879
6:	season	0.032044820	0.027754237	0.042736311
7:	month	0.022320642	0.087474437	0.072043330
8:	weather	0.020773696	0.028921204	0.037542662
9:	wday	0.020122691	0.070243483	0.095934115
10:	windspeed	0.012818493	0.113148505	0.125612109
11:	holiday	0.002981215	0.009142486	0.009793738

Model1

```
xgb.importance(colnames(dtrain2), model = xgb.model2)
```

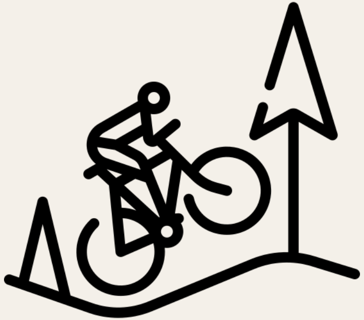
	Feature	Gain	Cover	Frequency
1:	hour	0.633683307	0.165721982	0.19702685
2:	temp	0.105467079	0.146540507	0.14414614
3:	workingday	0.068045635	0.013100942	0.03010132
4:	humidity	0.047603537	0.229691188	0.17151098
5:	season	0.036527228	0.027604354	0.04112122
6:	atemp	0.025890977	0.121720167	0.07669551
7:	weather	0.023774908	0.031590930	0.03875453
8:	wday	0.021046517	0.064937872	0.09525923
9:	month	0.019784188	0.084815188	0.06826418
10:	windspeed	0.015780387	0.104835675	0.12691369
11:	holiday	0.002396237	0.009441197	0.01020635

Model2

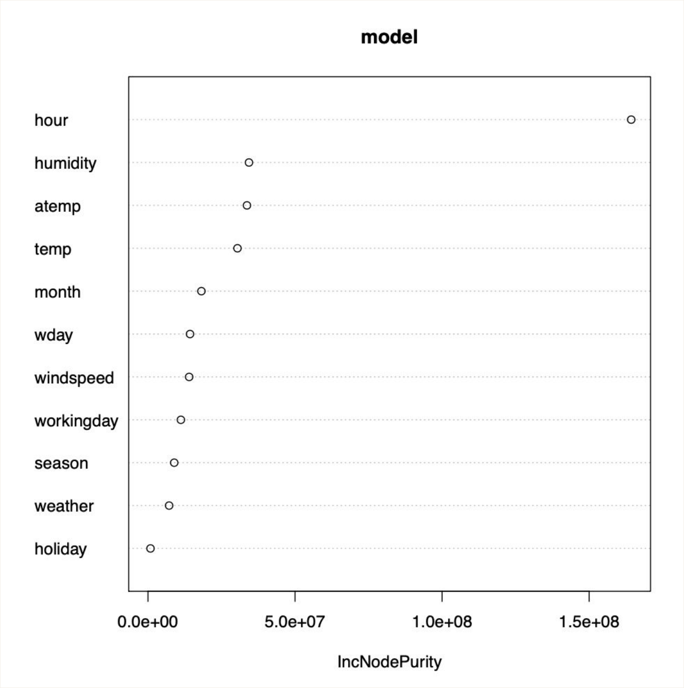
```
> xgb.importance(colnames(dtrain6), model = xgb.model6)
```

	Feature	Gain	Cover	Frequency
1:	hour	0.524250981	0.128843714	0.162998881
2:	hour_group	0.118397270	0.026742783	0.041700858
3:	temp	0.100193295	0.141316188	0.143304737
4:	workingday	0.051388542	0.016757564	0.032077583
5:	humidity	0.047925959	0.233204823	0.175680716
6:	atemp	0.033495270	0.132336048	0.072510257
7:	month	0.028836641	0.085533790	0.068183514
8:	weather	0.026542299	0.026539817	0.036031332
9:	season	0.025437447	0.029450226	0.039686684
10:	wday	0.022700827	0.063155580	0.088922044
11:	windspeed	0.017656024	0.108337428	0.129130921
12:	holiday	0.003175445	0.007782038	0.009772473

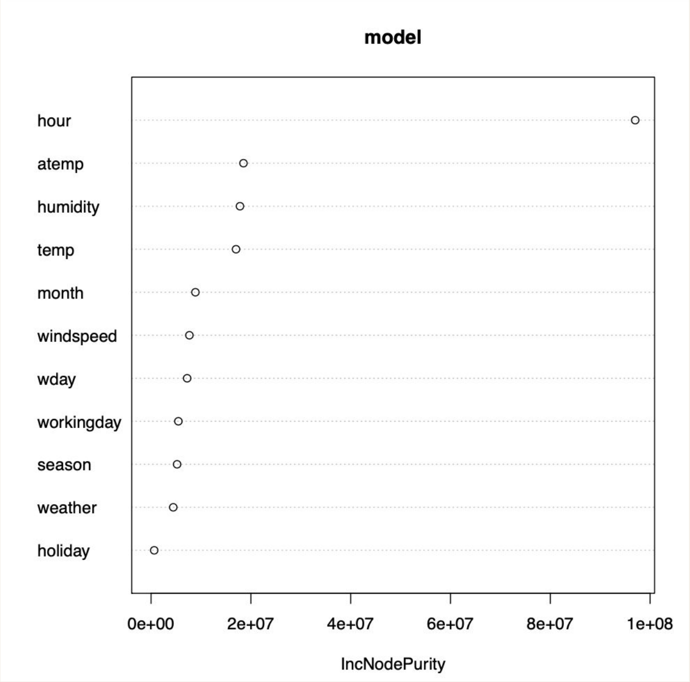
Model3



Conclusion-RF factor contribution

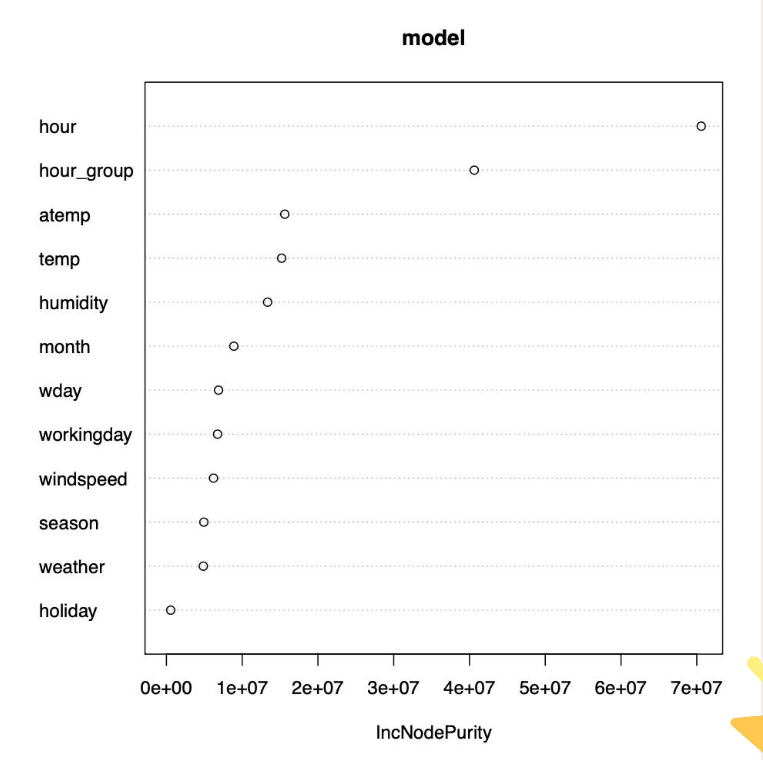


Model1



Model2

Conclusion-RF factor contribution



Model3

Conclusion-Lasso factor selection

```
12 x 1 sparse Matrix of class "dgCMatrix"
                    s1
(Intercept) 37.8935136
season       .
holiday      .
workingday   .
weather     -2.0794993
temp         2.3348893
atemp        4.3259506
humidity    -2.2496609
windspeed    0.3340296
month        7.2756514
wday         1.9244552
hour         7.5264667
```

Model1

```
12 x 1 sparse Matrix of class "dgCMatrix"
                    s1
(Intercept) 39.9144267
season       .
holiday      5.6974644
workingday  -8.4286541
weather     -0.9683099
temp         1.0460630
atemp        3.9875752
humidity    -1.7403809
windspeed    0.3194968
month        5.9027526
wday         1.4727930
hour         6.7564946
```

Model2

Conclusion-Lasso factor Selection

```
13 x 1 sparse Matrix of class "dgCMatrix"
                     s1
(Intercept) -17.1953294
season        .
holiday       5.9994683
workingday   -3.2767709
weather     -15.6479946
temp          0.9724485
atemp         3.4623013
humidity     -0.8280259
windspeed    -0.3570003
month         5.4669801
wday          1.8207767
hour          0.8868155
hour_group   75.5510028
```

Model3

Thank for Your Listening