

# Share Bike Analysis Report

Diwen Li

*Target: using 3 different model to predict Bike Rental data.*

First, analysis the data.

I use ggplot to plot the data group by hour, weekdays, and weather. I think these variable have the most significant influent to the prediction.

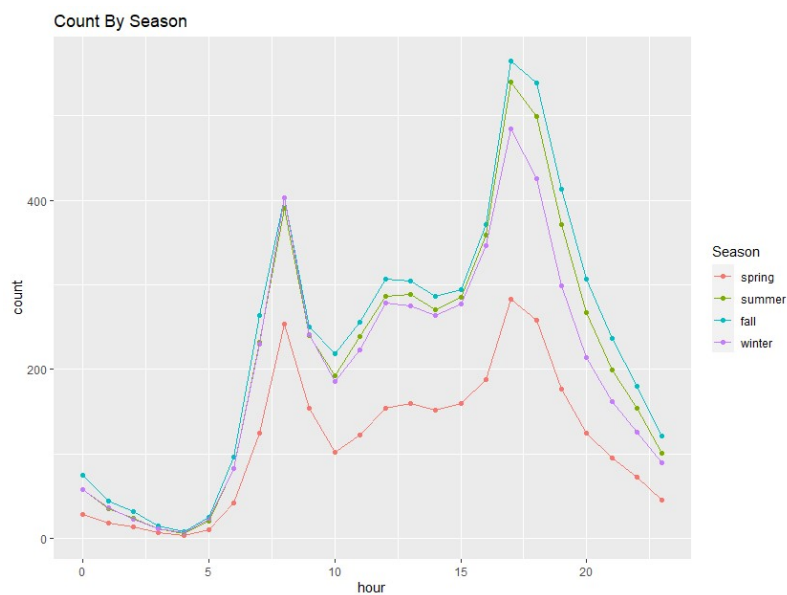


Figure1 above shows there are two peaks in morning and evening, and in noon have a small one. The season have influence most when its spring.

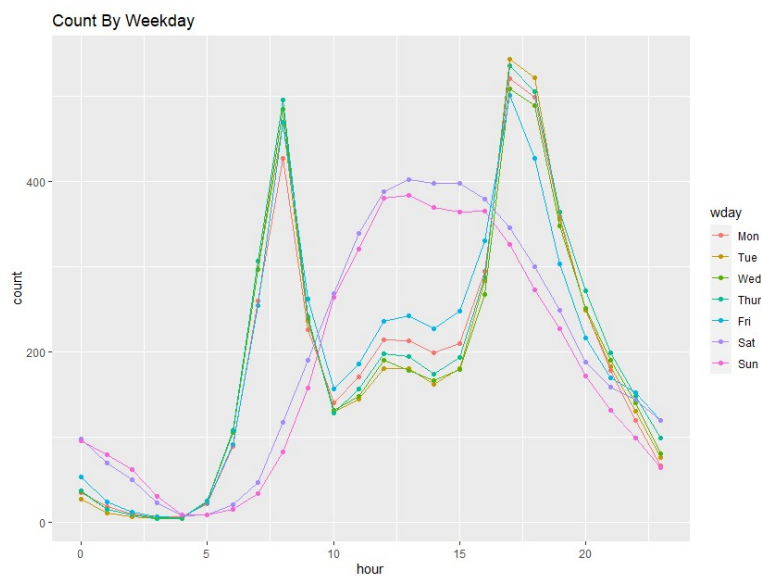


Figure2 shows Monday to Friday have almost the same count with two peaks in morning and

evening. But weekend have only one large peak in noon.

## Model1: Random Forest

Random Forest may be the second best model to predict.

At first, I use batches function to divided data in monthly, then for each month, find the rf model. Then use monthly rf model to predict monthly test. But in the same month, variable "month", "year", "season" are in same level. So the rf model can work with these variable well.

<a href="#">rf_ntree500_monthly.csv</a> 5 hours ago by Diwen Li <a href="#">add submission details</a>	0.88576	0.88576	<input type="checkbox"/>
<a href="#">rf_ntree250_monthly.csv</a> 5 hours ago by Diwen Li <a href="#">add submission details</a>	0.89351	0.89351	<input type="checkbox"/>

I use cross validation in each month to find the best ntree and mtry parameter to predict. But that takes too long to run and monthly data is not large enough to make a smaller cv test. So this process doesn't help much on my model.

<a href="#">rf_fullmodel_cv10.csv</a> 2 days ago by Diwen Li <a href="#">add submission details</a>	0.60460	0.60460	<input type="checkbox"/>
<a href="#">rf_model_fullmodel_CV10.csv</a> 2 days ago by Diwen Li rf model with all data and find ntree with cv10	0.59772	0.59772	<input type="checkbox"/>

The parameter choosing of ntree and mtry don't have much influence to the model. So I decide to pick some normal value to test is enough.

Then I change my train test. In the later month, I will use the earlier month data to help improve the model. (for example, in the second year, I will use the data in the first year to improve the model). That takes more time to run the model. But the result is much better.

```
for (m in 1:24){  
  cat("month: ",m,"\n")  
  monthlydata = Train[train_bat[1,1]:train_bat[2,m],-11]  
  monthlytest = Test[test_bat[1,m]:test_bat[2,m],-11]  
  monthlycount = count[train_bat[1,1]:train_bat[2,m]]  
  
  rf.fit = randomForest(monthlycount~.,  
                        data = monthlydata,  
                        trian = monthlycount,  
                        ntree = 500,  
                        mtry = 20)  
  monthlypred = predict(rf.fit,monthlytest)  
  result[test_bat[1,m]:test_bat[2,m],"count"] = monthlypred  
}  
output(result$datetime,result$count,"rf_ntree500_mtry20_countAllMonth.csv")]
```

*(changing the train\_bat index to use all data before that month)*

<a href="#">rf_ntree500_mtry20_monthly.csv</a> 3 hours ago by Diwen Li <a href="#">add submission details</a>	0.50025	0.50025	<input type="checkbox"/>
---	---------	---------	--------------------------

Final model with a 0.5 score, it works good!

Conclusion of model1:

Using tree to predict this dataset should be the best way. For the data has strong tree like influence. I mean, it can easily find the different in weekdays and weekend. And the count change when hour change is also significant. And other variables is much less important.

## Model2: boosting tree

Form the model 1 and data analysis, I can see using tree should be the best way to predict. So my second model is using boosting tree to predict.



```
for (m in 1:24){
  cat("month ",m,"\n")
  train_subset = train[train_bat[1,1]:train_bat[2,m],-c(1,2,11,12,13)]
  count_subset = count[train_bat[1,1]:train_bat[2,m]]
  test_subset = test[test_bat[1,m]:test_bat[2,m],-c(1,2,11,12,13)]
  fit.boost = gbm(log(count_subset)~.,
    data = train_subset,
    distribution = "gaussian",
    n.trees = 1000,
    interaction.depth = 20,
    cv.folds = 10)
  cv.num = gbm.perf(fit.boost)
  pred.boost = predict(fit.boost,test_subset,n.trees = cv.num)
  result[test_bat[1,m]:test_bat[2,m],"count"] = exp(pred.boost)
}
```

```
output(result$test.datetime,result$count,"boostingtree_cv10_depth20_cutModel.csv")
```

<a href="#">boostingtree_cv5_depth30.csv</a> 4 hours ago by Diwen Li <a href="#">add submission details</a>	0.71244	0.71244	<input type="checkbox"/>
<a href="#">boostingtree_cv5_depth25.csv</a> 4 hours ago by Diwen Li <a href="#">add submission details</a>	0.70567	0.70567	<input type="checkbox"/>
<a href="#">boostingtree_cv20_depth10.csv</a> 4 hours ago by Diwen Li <a href="#">add submission details</a>	0.70981	0.70981	<input type="checkbox"/>
<a href="#">boostingtree_cv10_depth20.csv</a> 4 hours ago by Diwen Li <a href="#">add submission details</a>	0.70681	0.70681	<input type="checkbox"/>
<a href="#">boostingtree_cv10_depth20_countAllMonth.csv</a> 2 hours ago by Diwen Li <a href="#">add submission details</a>	0.48725	0.48725	<input type="checkbox"/>

After uploading my result with different parameter, And using cv10 to train different model.

I found it don't influence much to my prediction. So I decide to pick some normal parameter is good enough.

<a href="#">logBoostingtree.csv</a> a day ago by Diwen Li <a href="#">add submission details</a>	0.42628	0.42628	<input type="checkbox"/>
<a href="#">boostingtree_cv5.csv</a> 2 days ago by Diwen Li <a href="#">add submission details</a>	Error 	Error 	<input type="checkbox"/>

I pick depth20 because its deep enough for this model. If going deeper, it will take too long to run and won't have much improve to the model. I still use monthly data in first, and soon, I found using previous data help a lot in improving the model. And each model will pick the best cv.num to predict the test.

And the result will have negative value when predicting, so I use log to help improve my mode. And it works very good!

### Conclusion:

**Using boosting tree help builds a better model and works better than random forest.**

## Modle3: Linear Regression

I pick linear regression as my third model because i think in some way it may be helpful to explain the model.

From the data analysis, I think using a polynomial regression for "hour" should be a good choice because it has two peaks. A degree 4 polynomial regression can fit this well. And for other variables, because they are not so significant and most of them may have a linear connection with count. For example, if temp and atemp goes higher, people may less like to ride a bike because its too hot.

Actually I would like to set up dummy variable for weekdays and weather in this model. But both don't work well and have a horrible score. So I give it up.

Finally, I use the same way as model 1and 2, I use the data in previous month to help improve the later month prediction. And I add "hour" with degree of 4. And the model just works not bad.

<a href="#">simplelm_fitlog_fulldata.csv</a> an hour ago by Diwen Li <a href="#">add submission details</a>	1.03236	1.03236	<input type="checkbox"/>
<a href="#">simplelm_fitlog_.csv</a> an hour ago by Diwen Li <a href="#">add submission details</a>	233.21408	233.21408	<input type="checkbox"/>

**Conclusion: using linear regression on this data is not a good choice. But the data do have some linear connection though I couldn't use it well in my model.**