BIKE SHARING SYSTEM

63	↓11	Dhruv Singal	0.37011	6	Sat, 11 Apr 2015 23:34:25 (-9.1d)
-		**a.karwan**	**0.37040**	-	**Tue, 21 Jul 2015 20:24:47** Post-Deadline

Post-Deadline Entry
If you would have submitted this entry during the competition, you would have been around here on the leaderboard.

64	↓11	prashkr	0.37041	6	Thu, 09 Apr 2015 20:53:07 (-23.1h)

*final result **0.3704** that is **64** place out of **3252***

adam.karwan@gmail.com

Warsaw, 22 July 2015

Table of Contents

Table of Contents 2
Task description 3
Data Set Description 3
Data Fields with Description 4
Data Analysis 5
Time 5
- Hour 5
- Day 5
- Month, Year 5
- Hour & Day 6
- Weekday & Hour 6
- Weather 7
- Temperature 7
- Humidity 7
- Correlations 8

New Attributes 9
- dp_reg 9
- dp_cas 9
- temp_reg 9
- temp_cas 9
- day_type 9
- year_part 9

Final Solution 10
- R Model 10
- Python Model 10

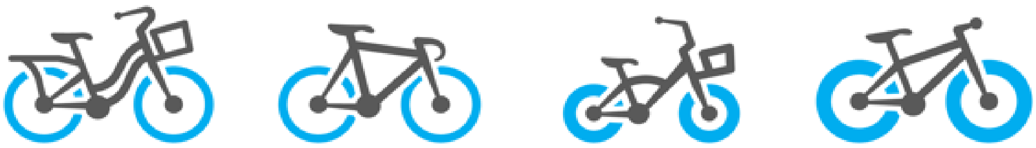
Links 11

Task description

[https://www.kaggle.com/c/bike-sharing-demand]

[http://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset]

You are provided hourly rental data spanning two years. For this competition, the **training set** is comprised of the **first 19 days of each month**, while the **test set** is the **20th to the end of the month**. You must predict the total count of bikes rented during each hour covered by the test set, using only information available prior to the rental period.



The data generated by these systems makes them attractive for researchers because the ***duration of travel***, ***departure location***, ***arrival location***, and ***time elapsed*** is explicitly recorded. Bike sharing systems therefore function as a **sensor network**, which can be used for studying **mobility in a city**. In this competition, participants are asked to combine historical usage patterns with weather data in order to forecast bike rental demand in the Capital Bikeshare program in Washington, D.C.

Data Set Description

Name	Capacity
Train Set Observation	10886
Test Set Observation	6493
Number of Attributes	12

No empty data

Data Fields with Description

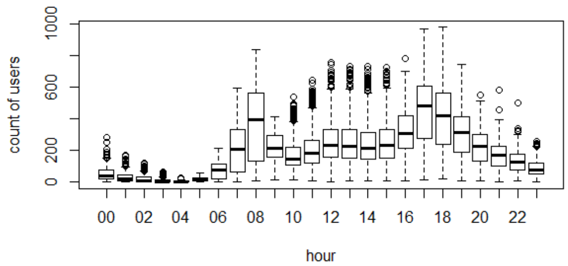
No	*Attribute*	Description
1	***datetime***	hourly date +timestamp
2	***season***	**1** = spring **2** = summer **3** = fall **4** = winter
3	***holiday***	whether the day is considered a holiday **(0; 1)**
4	***workingday***	whether the day is neither a weekend nor holiday **(0; 1)**
5	***weather***	**1**: Clear, Few clouds, Partly cloudy, Partly cloudy **2**: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist **3**: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds **4**: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
6	***temp***	temperature in Celsius
7	***atemp***	"feels like" temperature in Celsius
8	***humidity***	relative humidity
9	***windspeed***	wind speed
10	***casual***	number of non-registered user rentals initiated
11	***registered***	number of registered user rentals initiated
12	***count***	number of total rentals

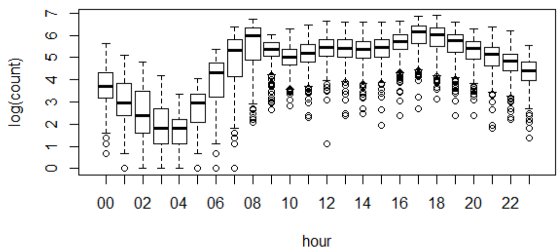
Count = Casual + Registered

Data Analysis

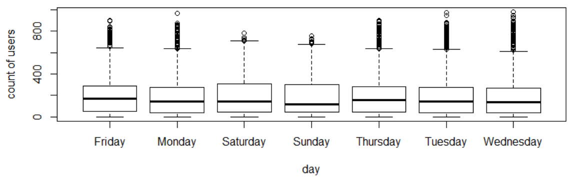
Time

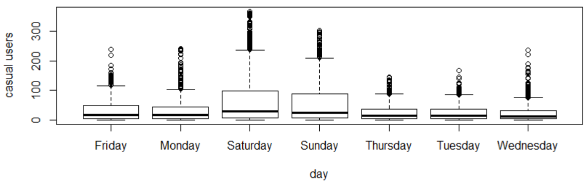
Hour



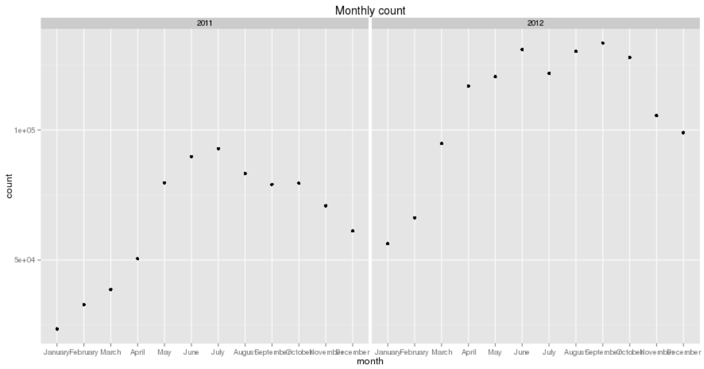


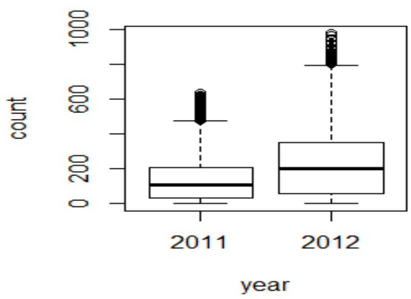
Day



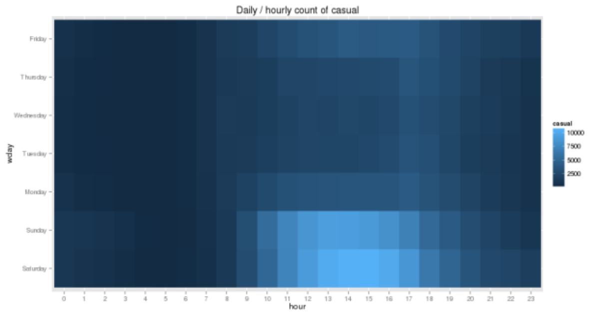


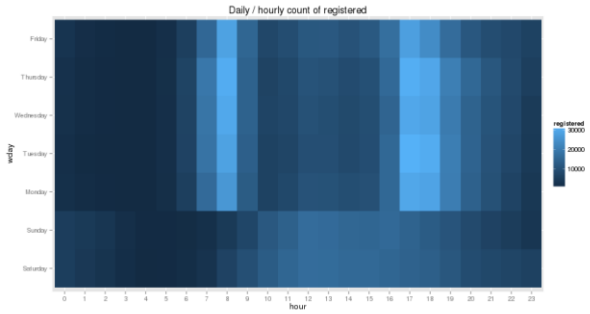
Month, Year

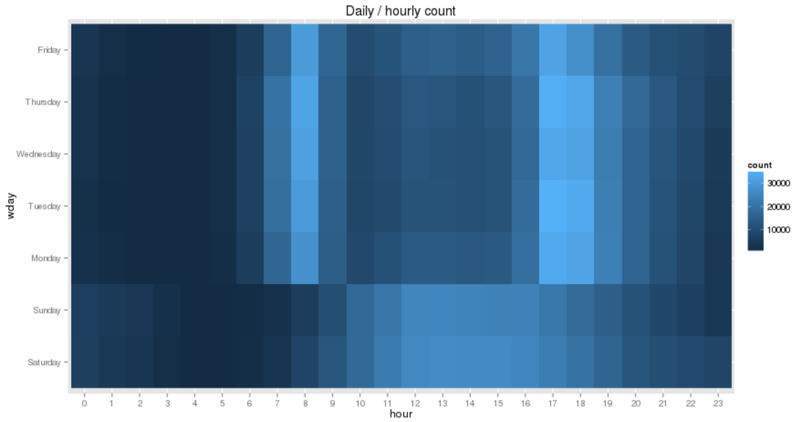




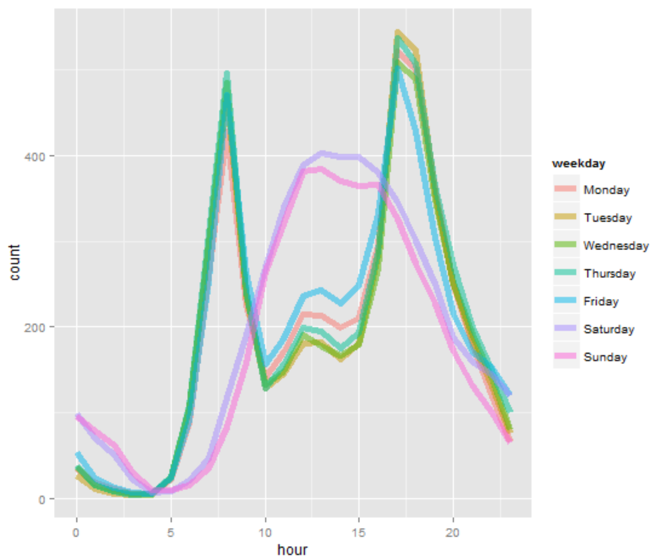
Hour & Day



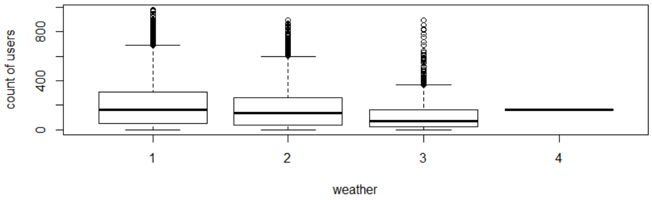


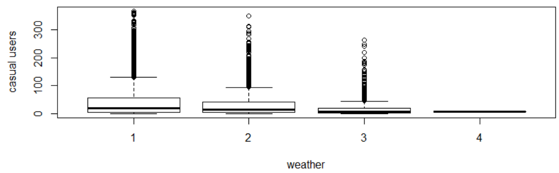


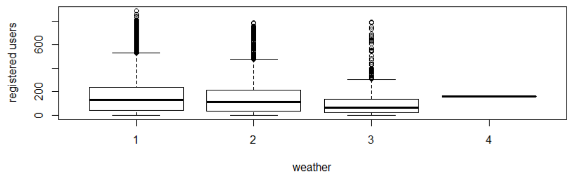
Weekday & Hour



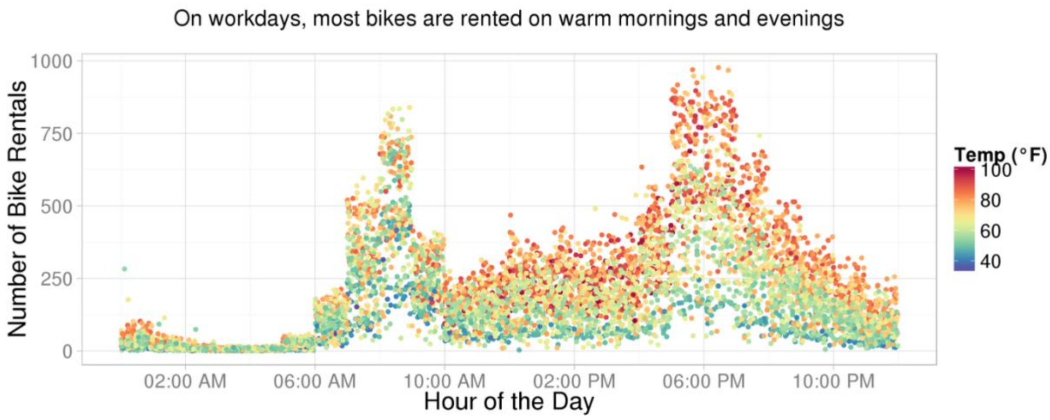
Weather



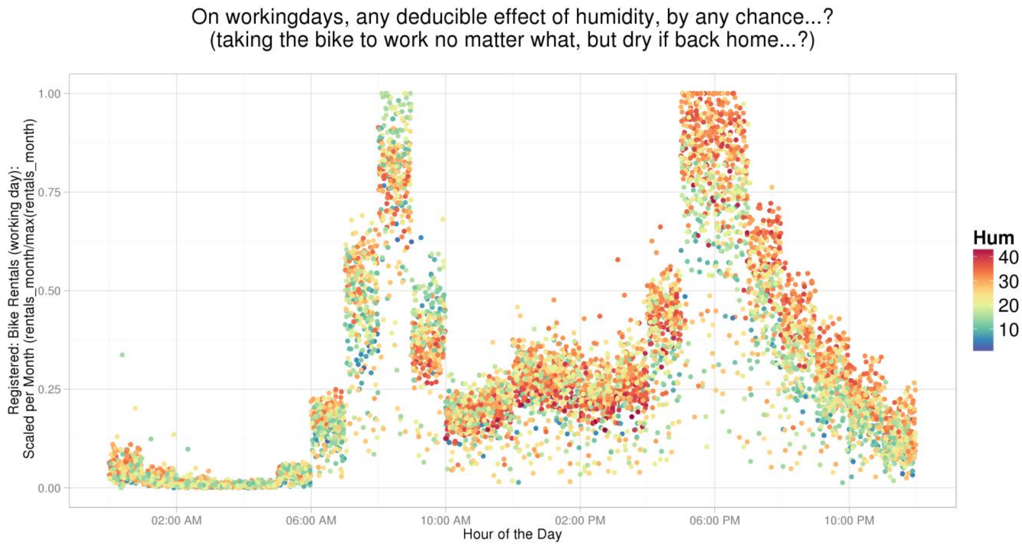




Temperature

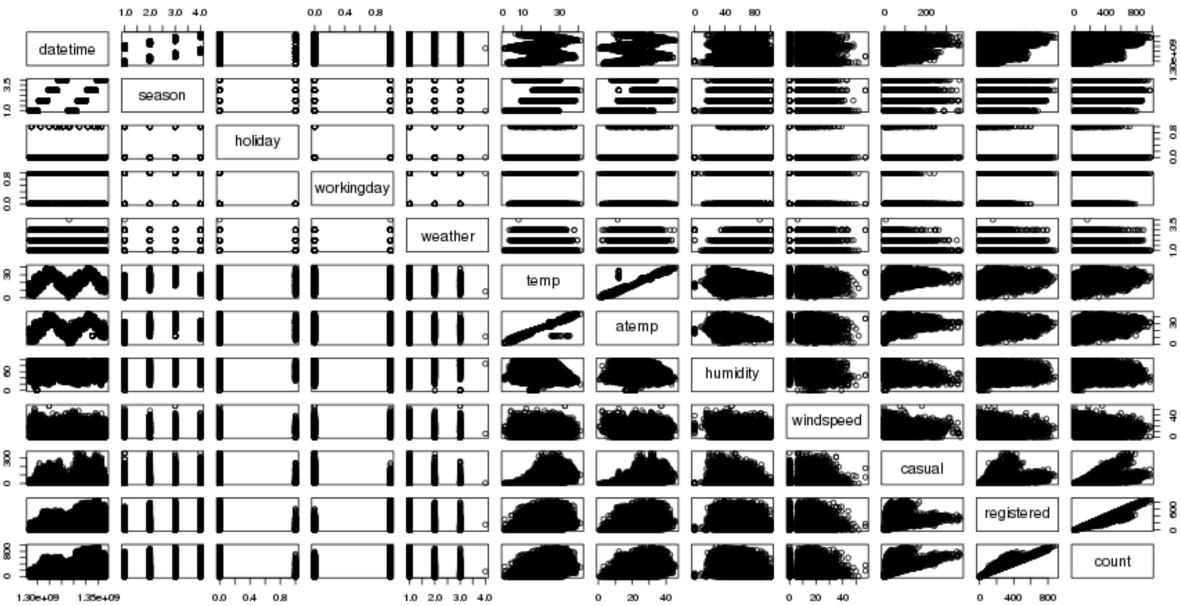


Humidity



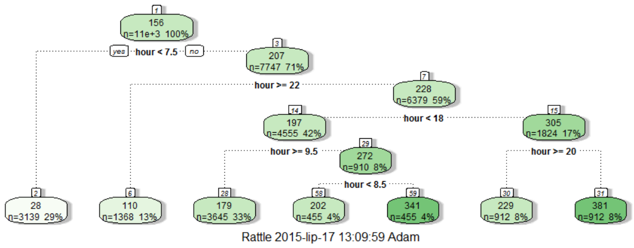
Correlations

	train.registered	train.casual	train.count	train.temp	train.humidity	train.atemp	train.windspeed
train.registered	1.00	0.50	0.97	0.32	-0.27	0.31	0.09
train.casual	0.50	1.00	0.69	0.47	-0.35	0.46	0.09
train.count	0.97	0.69	1.00	0.39	-0.32	0.39	0.10
train.temp	0.32	0.47	0.39	1.00	-0.06	0.98	-0.02
train.humidity	-0.27	-0.35	-0.32	-0.06	1.00	-0.04	-0.32
train.atemp	0.31	0.46	0.39	0.98	-0.04	1.00	-0.06
train.windspeed	0.09	0.09	0.10	-0.02	-0.32	-0.06	1.00

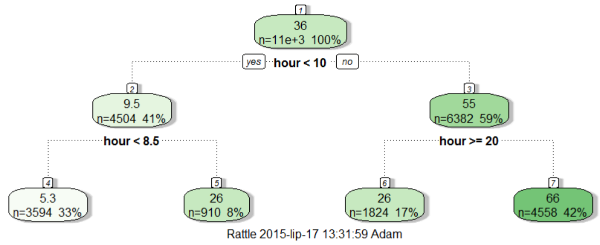


New Attributes

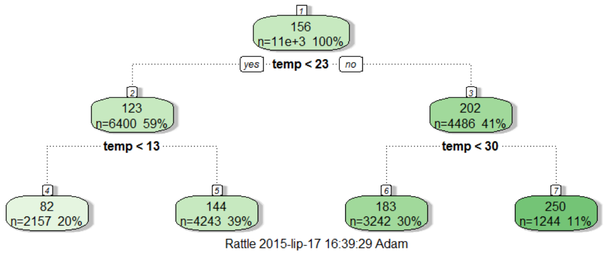
dp_reg



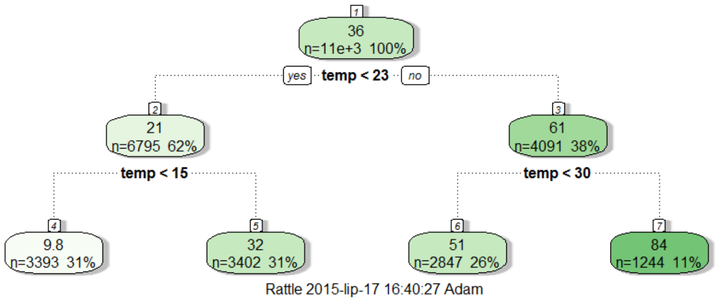
dp_cas



temp_reg



temp_cas



day_type

- *Holiday* [*holiday*=0 and *workingday*=0]
- *Weekend* [*holiday*=1]
- *Working Day* [*holiday*=0 and *workingday*=1]

year_part

- *1* to *8* (from first quarter of 2011 till fourth of 2012)

Final Solution

Final solution Is average of results from two models.

First one was computed in **R** and second in **Python**.

R Model

R model bases on two separately computed Random Forests for *casual* and *registered* users. And final result is a sum of predicted values daily. Due to differences in *count* of users *hourly* we use logarithm to normalize values. Moreover each model was trained on 250 trees. In first model we estimate number of registered users therefore *dp_reg* and *temp_reg* are used, analogously *dp_cas* and *temp_cas* for second one. Rest attributes used: *hour, day, day_type, holiday, season, year, year_part, weekend, workingday, atemp, humidity, weather, windspeed.*

Python Model

Python model bases on a combination of RF (Random Forrest) and GBDT (Gradient Boosting Decision Trees). Twelve attributes are used: *hour, day, holiday, season, weekday, workingday, year, atemp, temp, humidity, weather, windspeed.* Year is normalized by substraction 2011. GBDT is computed on 100 and RF on 1000 trees. Before computing final result we compute average regression of two instances for each approach RF and GBDT. For estimated variables logarithm is used to normalize results.

Links

1. http://www.analyticsvidhya.com/blog/2015/06/solution-kaggle-competition-bike-sharing-demand
2. https://github.com/adityashrm21/Kaggle/blob/master/Bike_Sharing_Demand.R
3. http://brandonharris.io/kaggle-bike-sharing
4. https://github.com/namebrandon/kaggle-bike-sharing
5. https://cran.r-project.org/web/packages/gbm/gbm.pdf
6. https://cran.r-project.org/web/packages/randomForest/randomForest.pdf
7. https://rstudio-pubs-static.s3.amazonaws.com/25177_bd95e70bb6bf4b26a2cc2d4ad1cb3c33.html
8. http://www.analyticsvidhya.com/blog/2015/07/guide-data-visualization-r
9. https://www.kaggle.com/c/bike-sharing-demand/forums/t/12809/python-scikit-learn-averaging-gbrt-and-random-forest-0-37108
10. https://github.com/dirtysalt/tomb/blob/master/kaggle/bike-sharing-demand/pub0.py
11. http://scikit-learn.org/stable/index.html