An application of the R Forecast Package in the Kaggle Competition Bike Sharing Demand

John Arhin

j_arhin@yahoo.co.uk

https://github.com/jarhin/Kaggle-Competition-Bike-Sharing-Demand

Monday 13th October 2014

Abstract

We show how the Forecast package in R can be used to solve a Kaggle Problem predicting the use of a city bike share system.

1 Introduction

1.1 Kaggle Problem

We acknowledge that the data for this problem is attributed to [1].

For the benefit of the reader, we reproduce the details of the Bike Sharing Demand Kaggle Competition as described in the URL `http://www.kaggle.com/c/bike-sharing-demand`.

Bike sharing systems are a means of renting bicycles where the process of obtaining membership, rental, and bike return is automated via a network of kiosk locations throughout a city. Using these systems, people are able rent a bike from a one location and return it to a different place on an as-needed basis. Currently, there are over 500 bike-sharing programs around the world.

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 Kaggle competition, we 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.

1.1.1 Bike Data

We are provided hourly rental data spanning two years. For this Kaggle 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. We predict the total count of

Variable	Value	Meaning
datetime		hourly date + timestamp
season	1	Spring
	2	Summer
	3	Fall
	4	Winter
holiday		whether the day is considered a holiday
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
temp		temperature in Celsius
atemp		"feels like" temperature in Celsius
humidity		relative humidity
windspeed		wind speed
casual		number of non-registered user rentals initiated
registered		number of registered user rentals initiated
count		number of total rentals

Table 1: Data Fields

bikes rented during each hour covered by the test set, using only information available prior to the rental period.

We show the variables of the data-set together with their meaning in Table 1.

1.1.2 R Setup

In this report, we use an an Intel i7 PC with 16GB RAM. The operating system is the Linux distribution OpenSUSE.

We use R [1] Version 3.1.1.

We use R together with Project Template [3], ggplot2 [4] and the Forecast [5] packages.

1.2 Git Repository

The work of this analysis can be found in the git repository that is located at the URL `https://github.com/jarhin/Kaggle-Competition-Bike-Sharing-Demand`

2 Exploratory Data analysis

3 Methodology

4 Results

5 Conclusion

References

[1] R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

[2] Fanaee-T, Hadi, and Gama, Joao, Event labeling combining ensemble detectors and background knowledge, Progress in Artificial Intelligence (2013): pp. 1-15, Springer Berlin Heidelberg.

[3] ProjectTemplate: Automates the creation of new statistical analysis projects. R package; Version 0.6 http://projecttemplate.net

[4] H. Wickham. ggplot2: elegant graphics for data analysis. Springer New York, 2009. http://had.co.nz/ggplot2/book

[5] RJ Hyndman, Y Khandakar, Automatic time series for forecasting: the forecast package for R, Monash University, Department of Econometrics and Business Statistics, 2007, http://robjhyndman.com/software/forecast/