Casual Bike Rental Volume via Artificial Neural Network

By Jenny Shan

Background

Bike sharing systems are new generation of traditional bike rentals where whole process from membership, rental and return back has become automatic. Through these systems, user is able to easily rent a bike from a particular position and return back at another position. Currently, there are about over 500 bike-sharing programs around the world which is composed of over 500,000 bicycles. The number of major cities that are becoming bike-friendly is growing in recent years. It is expected that in a near future, most major cities provide this service along their other public transport services. How to better predict the casual rental volume is a key challenge to the business.

In this study, we built a predictive model for casual bike rental volume using neural network and compared its performance with a more popular approach, linear regression. This study suggests that it is possible to develop a reproducible and transportable predictive instrument for casual bike rental volume.

Objective

This study aimed to build a predictive model for casual bike rental volume using artificial neural network and compare its performance with traditional regression method, linear regression.

Data

The data set under study is related to 2-year usage log of a bike sharing system namely Capital Bike Sharing (CBS) at Washington, D.C.,USA. There were some external sources that corresponding historical environmental values such as weather conditions, weekday and holidays are extractable. The sample size is 4322 in the test sample and 4323 in training sample, a total of 8645 records from year 2011.

All the records were randomly assigned into 2 groups: training sample (50%) and testing sample (50%). After using R to classify the weather data and to display the distribution of the testing and the training samples, we observed similar trend in data sets of two groups after the randomization (Figure 1,2). To avoid error caused by discrepancy in two data sets that may affect the performance and mean squared error, this validation is necessary.

Table 1: Descriptive Information Of Test Sample And Training Sample

	Training sample		Test sample	
Season				
1	1000	48.36	1068	51.64
2	1129	51.25	1074	48.75
3	1128	50.36	1112	49.64
4	1065	49.91	1069	50.09
Holiday				
No	4210	50.08	4196	49.92
Yes	112	46.86	127	53.14
Week day				
0	609	49.47	622	50.53
1	596	48.26	639	51.74
2	597	48.85	625	51.15
3	642	52.24	587	47.76
4	628	51.27	597	48.73
5	624	50.36	615	49.64
6	626	49.53	638	50.47
Working day				
No	1347	49.27	1387	50.73
Yes	2975	50.33	2936	49.67
weathersit	1.44	0.66	1.44	0.65
temp	0.49	0.20	0.48	0.20
atemp	0.47	0.17	0.46	0.18
hum	0.65	0.20	0.64	0.20
windspeed	0.19	0.12	0.19	0.12
Casual rentals	28.67	39.19	28.53	38.49

Figure 1: Distribution of Casual Bike Rentals in Training Sample

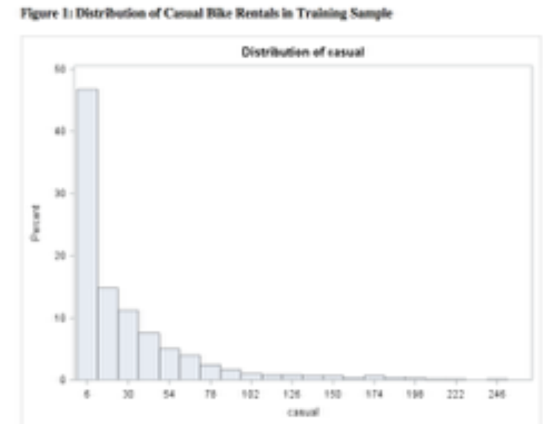
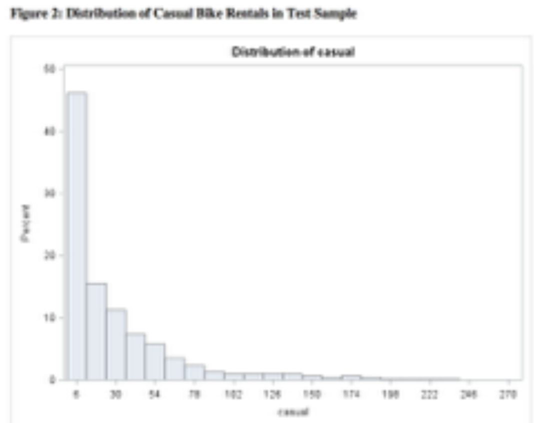


Figure 2: Distribution of Casual Bike Rentals in Test Sample

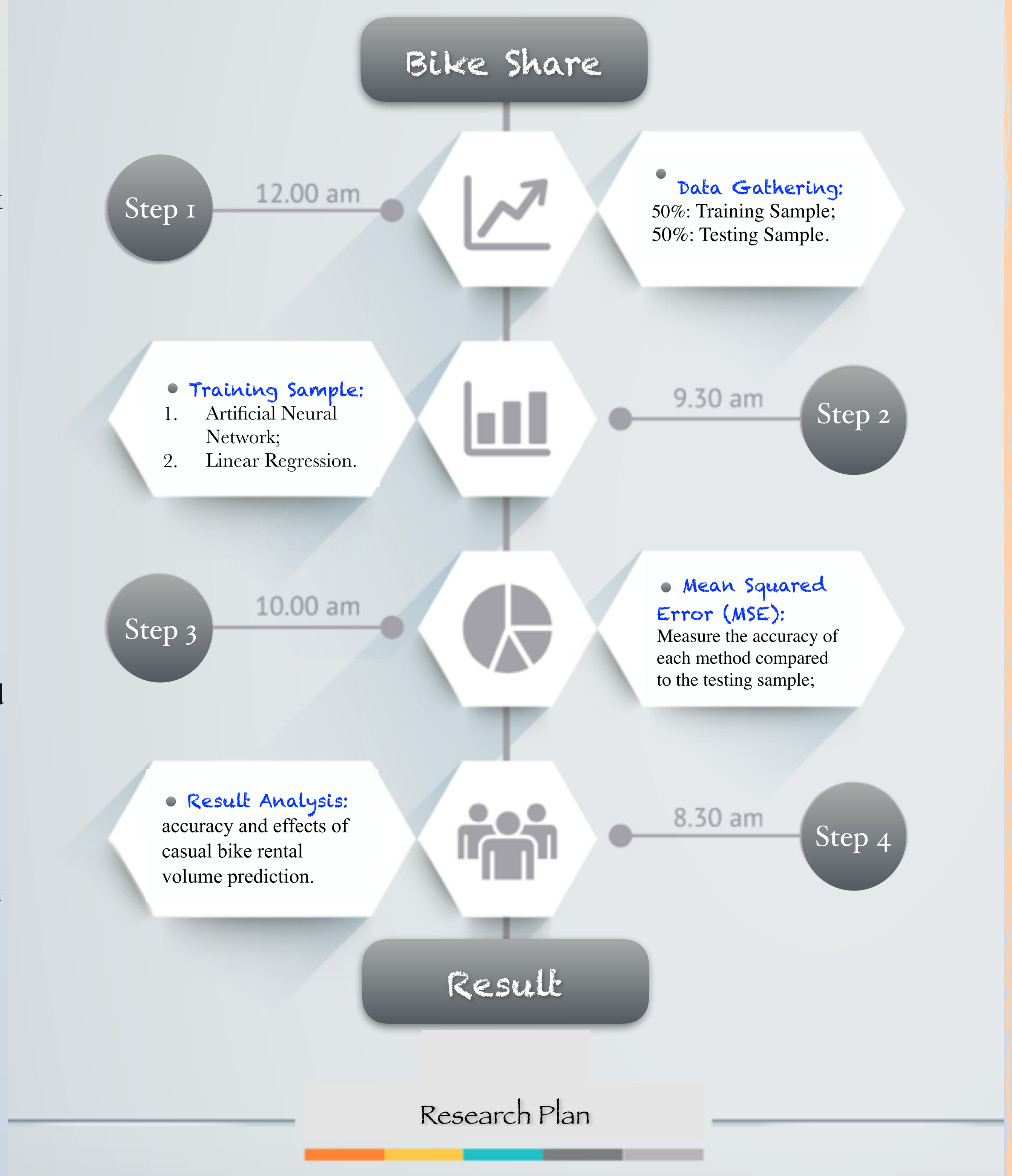


Method

Parameters:

season	1:springer, 2:summer, 3:fall, 4:winter)
mnth	month (1 to 12)
hr	hour (0 to 23)
holiday	weather day is holiday or not
weekday	day of the week
workingday	if day is neither weekend nor holiday is 1, otherwise is 0
weathersit	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	Normalized temperature in Celsius. The values are derived via (t-t_min)/(t_max-t_min), t_min=-8, t_max=+39 (only in hourly scale)
atemp	Normalized feeling temperature in Celsius. The values are derived via (t-t_min)/(t_max-t_min), t_min=-16, t_max=+50 (only in hourly scale)
hum	Normalized humidity. The values are divided to 100 (max)
windspeed	Normalized wind speed. The values are divided to 67 (max)
casual	count of casual users

All the records were randomly assigned into 2 groups: training sample (50%) and testing sample (50%). Two models were built using training sample: artificial neural network and linear regression. For artificial neural network, the input layer has 11 inputs, the two hidden layers have 3 and 2 neurons and the output layer has a single output. Mean squared errors (MSE) were calculated and compared between both models. A cross validation was conducted using a loop for the neural network and the cv.glm() function in the boot package for the linear model. A package called "neuralnet" in R was used to conduct neural network analysis.



Mean squared error (MSE) : MSE is a risk function, corresponding to the expected value of the squared error loss or quadratic loss.Both linear regression techniques such as analysis of variance estimate the MSE as part of the analysis and use the estimated MSE to determine the statistical significance of the factors or predictors under study. The goal of experimental design is to construct experiments in such a way that when the observations are analyzed, the MSE is close to zero relative to the magnitude of at least one of the estimated treatment effects. MSE is also used in several stepwise regression techniques as part of the determination as to how many predictors from a candidate set to include in a model for a given set of observations.

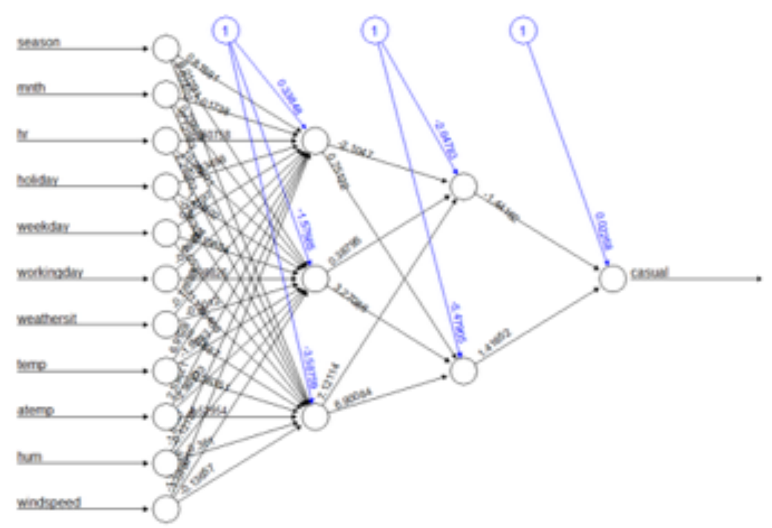
Results

Table 3: Linear Regression To Predict The Volume Of Casual Bike Rental

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	21.0	2.7	7.7	0.000	***
season	1.9	0.7	2.5	0.011	*
mnth	-0.3	0.2	-1.4	0.150	
hr	1.1	0.1	17.0	< 2e-16	***
holiday	-8.6	2.7	-3.1	0.002	**
weekday	0.1	0.2	0.5	0.645	
workingday	-29.6	1.0	-30.7	< 2e-16	***
weathersit	0.3	0.8	0.5	0.643	
temp	78.8	18.3	4.3	0.000	***
atemp	11.0	20.5	0.5	0.593	
hum	-49.2	2.7	-18.2	< 2e-16	***
windspeed	-2.3	3.9	-0.6	0.550	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Figure 3: Neural Network of Casual Bike Rental Volume in Training Sample



The black lines show the connections between each layer and the weights on each connection while the blue lines show the bias term added in each step. The bias can be thought as the intercept of a linear model. The net is essentially a black box so we cannot say that much about the fitting, the weights and the model. Suffice to say that the training algorithm has converged and therefore the model is ready to be used.

For testing sample, the MSE was 798 for the linear regression and 265 for the artificial neural network. Artificial neural network performed better clearly.

By visually inspecting the plot we can see that the predictions made by the neural network are (in general) more concentrated around the line (a perfect alignment with the line would indicate a MSE of 0 and thus an ideal perfect prediction) than those made by the linear model.

Cross validation is another very important step of building predictive models. In cross validation, the average MSE for the neural network (268) is lower than the one of the linear model (806) although there seems to be a certain degree of variation in the MSEs of the cross validation. This may depend on the splitting of the data or the random initialization of the weights in the net.

Figure 4: Observed vs Predicted Casual Bike Rental Volume In Artificial Neural Network And Linear Regression Model

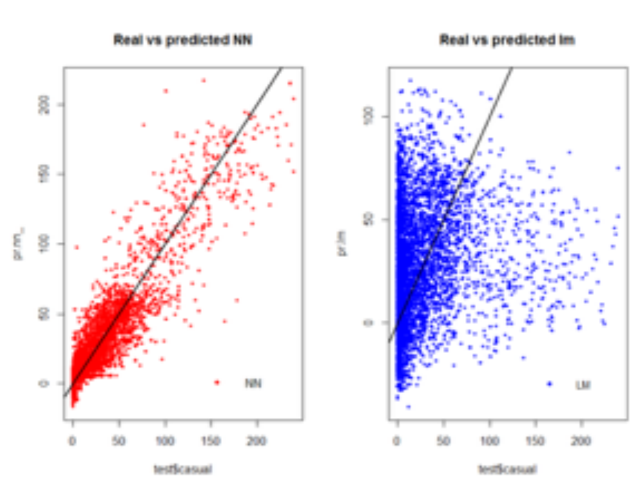
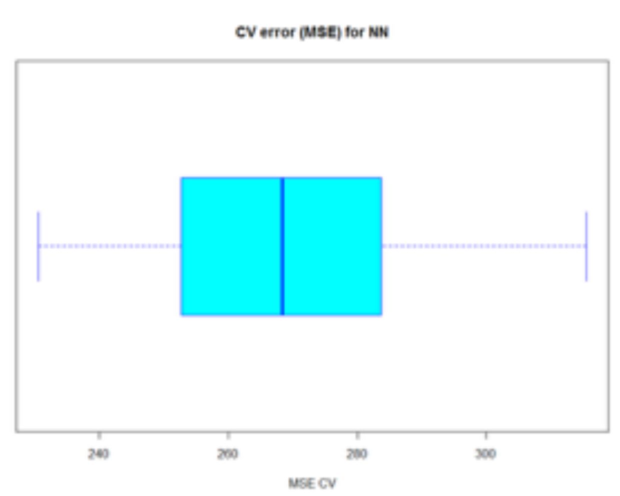


Figure 5: MSE for Artificial Neural Network for Testing Group



Discussions

There are limitations of this study. One of them was associated with artificial neural network method. This method employed deep machine learning method to explore the nonlinear association between casual bike rental volume and weather factors; however the nonlinear association make it very hard to interpreter the results, specially the association between the rate and individual predictors. Other predictors of casual bike rental volume were not available in this database.

Conclusion

In this study, we built a predictive model for casual bike rental volume using neural network and compared its performance with a more popular approach, linear regression.

This study suggests that it is possible to develop a reproducible and transportable predictive instrument for casual bike rental volume.

References

Data set: 2-year usage log of a bike sharing system namely Capital Bike Sharing (CBS) at Washington, D.C.,USA.

References:

http://www.ftchinese.com/story/001071840/ce
http://money.163.com/17/0306/10/CERBHHUV002580S6.html
http://www.news.com.au/finance/business/travel/bikeshare-cycles-dumped-en-masse-in-china/news-story/bd0dc411914fbbf1054d97a50d1ff2ed

Gerrard C., McCall J., Coghill G.M., Macleod C. (2012) Temporal Patterns in Artificial Reaction Networks. In: Villa A.E.P., Duch W., Érdi P., Masulli F., Palm G. (eds) Artificial Neural Networks and Machine Learning – ICANN 2012. ICANN 2012. Lecture Notes in Computer Science, vol 7552. Springer, Berlin, Heidelberg

Carner, Josep, et al. "Machine learning-based network modeling: An artificial neural network model vs a theoretical inspired model." Ninth International Conference on Ubiquitous and Future Networks IEEE, 2017:522-524.

Escolano, Andrés Yáñez, et al. "Statistical Ensemble Method (SEM): A New Meta-machine Learning Approach Based on Statistical Techniques." International Conference on Artificial Neural Networks: Computational Intelligence and Bioinspired Systems Springer-Verlag, 2005:192-199.