Analyzing Bike Sharing Systems: A Comparison of Machine Learning Models

Prince Joseph Erneszer Javier, Reynaldo Tugade Jr.

Asian Institute of Management, Makati, Philippines
pjavier@aim.edu, rtugade@aim.edu

Abstract

In performing data analysis, a common task is to search the most appropriate algorithm(s) to best resemble a given system. In this report, we demonstrate the suitability of using a neural network in predicting the potential number of bicycle-sharing users using combined historical rental and weather information. The idea is to augment previous machine learning models and discover the possibility of getting better test accuracy. We used K-Nearest Neighbor, Linear Regression, Ridge Regression, Lasso Regression, Linear Support Vector Machine, Decision Trees, Random Forest, and Gradient Boosting Method as baseline machine learning models. We used a 3 layer fully-connected feed-forward network with 30 hidden nodes. This report shows that such configuration works well the most with 73.2% r^2 and 52.7% MAPE. In comparison, RF could predict with 67.9% r^2 and 75.9% MAPE while GBM could predict with 67.2% r^2 and 68.5% MAPE.

1 Introduction

Bicycle-sharing systems are paid services where an individual can rent an available bicycle on a short-term basis. Used to be considered as a service only available in small and closed communities (e.g. campuses, private subdivisions), bicycle-sharing systems are becoming mainstream modes for public-transport in several countries. Few of these systems include Paris' "Vellib" which started operating in 2005, Hangzhou's bicycle hub in China which houses more than 50,000 bicycles and even locally with Asian Development Bank's (ADB) Sustainable Transport Initiative program.

From a sustainability perspective, bicycle-sharing systems have its benefits. One, it promotes better flexible mobility. Bicycle stations can be placed anywhere especially in areas where there's a perceived concentration of people traveling. Second, it impacts emission reduction due to no fuel use and reduces congestion. Third, it's relatively cheap and very convenient specifically since it helps improve multimodal transport connections.

In this regard, we wish to explore interesting behavior found in bicycle-sharing systems. The richness of data involved in bicycle-sharing systems can provide more information from a mobility sensing perspective. In this report, we will explore potential users based on previous rentals and weather data. We will leverage on learned Machine Learning and Neural network techniques to contrast and compare which among these models best resemble the bicycle-sharing system.

2 Data

The original dataset comes from the Capital Bikeshare website which is compiled by the Laboratory of Artificial Intelligence and Decision Support (LIAAD), University of Porto and placed in UCL. The dataset contains time-series rental and weather information.

Table 1: Available features found in the dataset.

Feature Variable	Possible Values	Description
instant	0 to 17378	Index of the record
dteday	2011-01-01 to 2012-12-31	Date
season	1,2,3,4	Season (Spring, Summer, Fall, Winter)
yr	0,1	Year of occurrence
mnth	1,2,3,...,12	Month
hr	0,1,2,...,23	Hour
holiday	0,1	Whether current day is a holiday. Based from (dc.gov)[Holidays]
weekday	0,1,2,...,6	Day of the week
weathersit	1,2,3,4	Weather information based meteorological events
_	_	(1) Clear, Few clouds, Partly cloudy
_	_	(2) Misty plus still generally cloudy environment
_	_	(3) Light Snow, Light Rain with occasional Thunderstorms,Light Rain with scattered clouds
_	_	(4) Heavy Rain with Ice pellets, Thunderstorm with Mist, Snow with Fog
temp	0.02 to 1.00	Normalized feeling temperature in Celsius.
atemp	0.0000 to 1.0000	Normalized feeling temperature in Celsius.
humz	0.00 to 1.00	Normalized humidity
windspeed	0.0000 to 0.8507	Normalized windspeed
casual	0 to 367	Count of casual users
registered	0 to 886	Count of registered users
cnt	1 to 977	Count of total rental bikes including both casual and registered

* Temperature variable *(temp)* is computed using the following equation:

$$\frac{t - t_{min}}{t_{max} - t_{min}}, t_{min} = -8, t_{max} = +39 \tag{1}$$

* Absolute temperature variable *(atemp)*:

$$\frac{t - t_{min}}{t_{max} - t_{min}}, t_{min} = -16, t_{max} = +50 \tag{2}$$

3 Methodology

3.1 Data Preprocessing

Features selected were year, holiday, temp, hum, windspeed, season, weathersit, mnth, hr, and weekday. The target variable was cnt. One-hot encoding was applied on season, weathersit, mnth, hr, and weekday. A bias was added as an additional column having a single value of 1.00. The resulting features data including bias contained 56 features. The features were then scaled using min-max scaling given by:

$$X_{scaled} = \frac{X - min(X)}{max(X) - min(X)} \tag{3}$$

where X is the feature matrix. Since the maximum value of cnt was found to be 977, cnt was divided by 1000 to scale it to values between 0 and 1. The last 20 days were set for testing set. In the remaining dataset, the last 60 days were set as the validation set. The remaining dataset was used to train the machine learning models. The validation set was used to evaluate the model during training while the testing set was used to test the accuracy of the model after training using the best parameters.

3.2 Neural Network Modeling

A feed-forward neural network having 56 input nodes, 30 nodes in one hidden layer, and 1 output node was developed. The learning rate from input to hidden and hidden to output were both 0.001. The input, hidden, and output activation functions were linear, sigmoid, and sigmoid respectively. These parameters were selected through a genetic algorithm optimizer that aimed to maximize validation accuracy.

Table 2: The ranges of parameter values fed into the genetic algorithm.

Parameter	Range of values
hidden nodes	20, 22, 24,...,54, 56
learning rates	0.001, 0.0001
activation functions	sin, relu, tanh, sigmoid

The loss function being minimized by the neural network is given by:

$$Error = \frac{1}{2}(\Psi_{NN} - \Psi_{true})^2 \tag{4}$$

where Ψ_{NN} is the predicted value and Ψ_{true} is the true value. The neural network was trained and validated using the training and validation sets over 5,000 iterations. The testing set was used to evaluate the predictive accuracy of the model using the coefficient of determination (r^2) and mean absolute percentage error (MAPE). The equations are shown below.

$$r^2 = \frac{\sum_i (\hat{y}_i - \bar{y})}{\sum_i (y_i - \bar{y})} \tag{5}$$

$$MAPE = \frac{100\%}{n} \sum_i \left| \frac{y_i - \hat{y}_i}{y_i} \right| \tag{6}$$

where $\hat{y}_i$ is the predicted value, y_i is the true value, $\bar{y}$ is the mean of true values, and n is the number of samples.

3.2 Machine Learning Modeling

Eight more regression models were trained on the same training dataset, namely k nearest neighbors (kNN), linear regression, lasso regression, ridge regression, linear support vector machines (LSVM), decision tree, random forest, and gradient boosting machines (GBM). Using the optimal parameter, the target values were predicted using the test set. The accuracy was measured as the r^2 and MAPE between the true values and predicted values.

Table 3: Summary of hyperparameters tweaked for each model.

Model	Parameters
Feed-forward NN	no. of nodes no. of hidden layers activation functions learning rates
k-Nearest Neighbor	no. of nearest neighbors
Linear Regression	-
Lasso Regression	alpha
Ridge Regression	alpha
Linear Support Vector Machine	C
Decision Tree	max-depth
Random Forest	max-depth
Gradient Boosting	max-depth

Each model was trained and validated on a range of parameters. The parameters that gave the highest r^2 on the validation set were identified as the optimal parameters.

4 Results

The feed forward neural network was able to predict bike-sharing counts on the test set with 73.2% accuracy, higher than the eight other machine learning models. The MAPE obtained was the lowest at 52.7%.

Table 4: Summary of the predictive accuracies and corresponding parameters of all models evaluated.

Model	Parameters	Values	r^2 Test Accuracy	Test MAPE
Feed-forward NN	no. of hidden nodes no. of hidden layers activation functions learning rates	30 1 (linear, sigmoid, sigmoid) (0.001, 0.001)	73.2%	52.7%
Random Forest	max-depth	33	67.9%	75.9%
Gradient Boosting Method	max-depth	22	67.2%	68.5%
k-Nearest Neighbor	no. of nearest neighbors	2	62.9%	134.9%
Lasso Regression	alpha	0.0001	48.1%	382.5%
Ridge Regression	alpha	10	48.0%	399.4%
Linear Regression	-	-	47.1%	424.4%
Decision Tree	max-depth	24	46.0%	75.3%
Linear Support Vector Machine	C	1	42.5%	303.6%

5 Conclusion

We demonstrated that a fully-connected feed-forward neural network could predict hourly bike rentals with the highest accuracy and lowest MAPE. Results showed that using a feed-forward neural network yields a 73.2% r^2 and 52.7% MAPE. The neural network was followed by the random forest regressor which could predict with 67.9% r^2 and 75.9% MAPE. And finally, GBM with 67.2% r^2 and 68.5% MAPE.

Further research can include using recurrent neural networks and ARIMA, which are specifically made for sequential data.

References

[1] 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, https://archive.ics.uci.edu/ml/datasets/bike+sharing+dataset