Bike-Sharing Data Analysis: Prediction of Daily Bike Rental Counts Based on Multiple Linear Regression

Final Project Report · MA 575 Fall 2021 · C3 · Team #2

Ali Taqi, Hsin-Chang Lin, Huiru Yang, Ryan Mahoney, Yulin Li

12/10/2021

In this project, the following question is to be answered: If we have the past history of bike rental counts as well as environmental and seasonal conditions, how and how well could we predict the bike rental counts in the future? In this project, such questions are approached by predictive modeling of daily bike rental counts from a 2011-2012 Bike Sharing dataset [1]. The daily bike rental counts are predicted with models based on Multiple Linear Regression (MLS) using the environmental and seasonal variables as predictors. The initial goal of this project is to train the model using only the 2011 data, and then validate the prediction power of the model on the 2012 data. Given the limited time span of available training data, issues are found in the validation process using the 2012 data; the impact of user base on the future predictions is brought to our attention. The initial models are then revisited and corrected to account for the effect of user base. The refined models are expected to have better prediction powers than the initial MLS models, but a full validation would require further availability of bike rental data.

1 Introduction

Bike sharing has become a world-wide phenomenon. Optimization of inventories and dynamic reallocation of bike-sharing resources are of growing interests from both a business and an environmental point of view. Both of these tasks require accurate predictions of bike rental behaviors at least on the daily level.

(further motivates & applications?)

In this project, we strive to answer the following question:

- If we have the past history of bike rental counts as well as environmental and seasonal conditions, how and how well could we predict the bike rental counts in the future?
- In particular, how and how well could we predict for the next whole year, and what about for the next few days?

Such questions are approached by predictive modeling of daily bike rental counts from a 2011-2012 Bike Sharing dataset [1]. The modeling approach is based on Multiple Linear Regression (MLS), and the daily bike rental counts are predicted using the environmental variables (e.g., weather conditions) and seasonal variables (e.g., holiday schedules) as predictors.

2 Background

The aim of this project is to achieve the best model(s) that can be obtained from past data for the use of predictions for the future, preferably one year ahead. To validate the prediction power of models under this setting, the basic goal of this project is to train all models using only the 2011 data, and then test them on the 2012 data.

The response variable to be predicted is the **daily** bike rental count. In the dataset we are studying [1], the following 3 types of bike rental counts are recorded:

1. the count of bike rentals by **casual** users
2. the count of bike rentals by **registered** users
3. the **total** count, which is the sum of casual count and registered count.

Two main types of predictors are included in the dataset, the environmental ones and the seasonal ones:

1. environmental variables

(Table 1: A sample of the data - variable names, meanings, units, sample values)

2. seasonal variables

(Table 2: A sample of the data - variable names, meanings, units, sample values)

3 Modeling & Analysis

3.1 Preprocessing

To be noticed, the value of categorical variables indicates type labels and has minimal physical meaning in the scale of numbers, which thus cannot be used in the same way as the numeric variables in MLS models. The categorical variables therefore needs to be carefully recognized before the actual modeling process.

The below variables are interpreted as Boolean variables and are transformed into `logical`-type variables in `R`:

- `holiday` (holiday or not)
- `workingday` (working day or not)

The below variables are interpreted as categorical variables and are transformed into `factor`-type variables in `R`:

- `season` (season, from 1 to 4)
- `yr` (year, from 0 to 1)
- `mnth` (month, from 1 to 12)
- `weekday` (weekday, from 0 to 6)
- `weathersit` (weather type, from 1 to 4)

3.2 Variable Selection

3.2.1 Predictors Selection

3.2.2 Response Transformation

3.3 Initial Modeling

3.3.1 Beginning Model

3.3.2 Final Model

Model interpretation

3.4 Diagonostic Analysis

3.5 Validation and Problemshooting

3.6 Refined Model

3.6.1 Prediction of the Yearly Growth Ratio

3.6.2 Prediction without the Yearly Growth Ratio

4 Prediction

4.1 Unadjusted Model

4.2 Refined Model

5 Discussion

Models for both long-term and short-term predictions are included.

To be noticed, at least one more year's data is needed for a final validation of the refined model, which is not available for the moment. This is to be left for the future work.

Time series

6 Appendix