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

Abstract

In this project, the following question is to be answered: If we have the past history of bike rental counts as well as records of 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.

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

- If we have the past history of bike rental counts as well as records of 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 predictions 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 being studied [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

```
##       dteday weathersit     temp    atemp      hum windspeed
## 1 2011-01-01          2 0.344167 0.363625 0.805833 0.160446
```

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

2. **seasonal** variables

```
##       dteday season yr mnth holiday weekday workingday
## 1 2011-01-01      1  0    1       0       6          0
## 2 2011-01-02      1  0    1       0       0          0
```

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

3 Modeling & Analysis

3.1 Pre-processing

3.1.1 Type Conversion

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

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)

3.1.2 Value Conversion

The recorded values of `temp` (measured temperature), `atemp` (feeling temperature), `hum` (measured humidity) and `windspeed` (measured wind speed) in the data set being studied here are the normalized ones; all recorded values are the ones that have been divided by the maximum of measured values [1]. For example, the recorded values of `temp` (measured temperature) are obtained by dividing the original measured values by 41 (max) and are thus all less than or equal to 1.

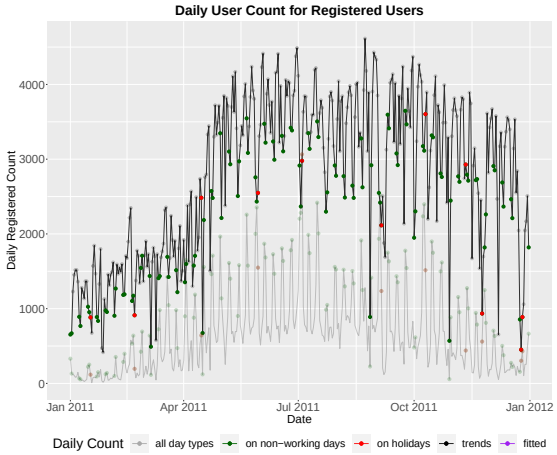
In this project, these normalized records are scaled back to their original values for the sake of easier interpretations. For example, the recorded values of `temp` (measured temperature) are multiplied by 41 (max) in the pre-processing process, which recovers the original scale of temperatures in Celsius.

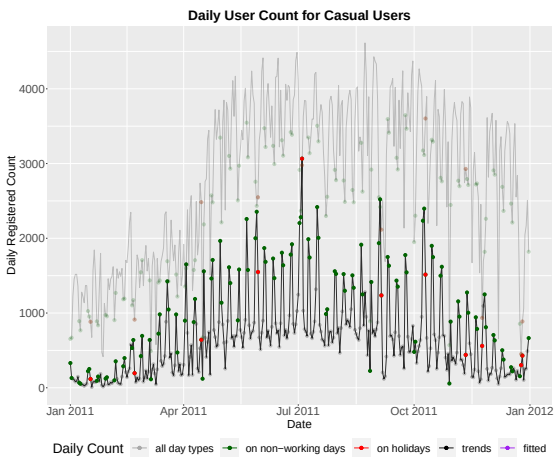
3.2 Variable Selection

3.2.1 Response Transformation

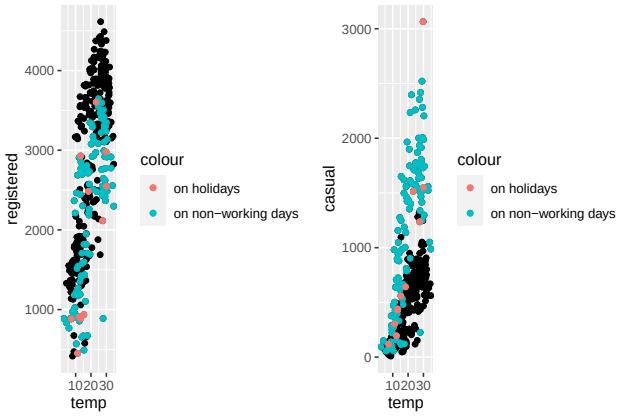
Notably, the behaviors of rental counts from different user types are considerably different.

1. **Patterns with weekdays** (see Figure 1,2): Over the time span of a week, the casual count usually reaches its minimum in the middle of a week (grey dots mostly) and its maximum on weekends (green dots mostly), while the registered count does the opposite.





2. **Patterns with temperatures** (see Figure 3): On working days, the casual count seems more linear in temperature (`atemp` and `temp`), while the registered count seems to be (at least) quadratic.



(Figure 3: casual vs temp & atemp, reg vs temp & atemp)

We therefore expect that the registered counts and casual counts will follow different distributions and should thus be predicted by separate models. Furthermore, for the casual count, avoiding unnecessary higher other terms has the benefit of more stable computations and model structures. The prediction of total counts will then be obtained by adding the predicted registered counts and predicted casual counts together.

3.2.2 Predictor Selection

Given the predictive nature of modeling in the current problem setting, the predicted response is of greater interests than the actual value of the parameter estimates, as opposed to that in an inference task. This, to some degree, relaxes the constraint forbidding colinearity in the predictors, since colinearity will only lead to instability in the parameter estimates but not in the predictions; however, we should still seek to minimize colinearity at least in our beginning model, which would lead to clearer model structures as well as better interpretability of model statistics at the early stage of modeling, which could provide us clearer directions in the improvement process that follows.

With the above considerations in mind, the predictors in the beginning model are selected following the 2-step approach below:

1. The scatter plot matrix for the whole set of variables are plotted for the 2011 training dataset, and all predictors that seem to be significant, i.e., predictors with which the response variable (daily rental count, `cnt`) exhibits a notable visual pattern, are selected.

2. From the selected predictors above, all the highly correlated predictors are removed. Within a group of correlated predictors, only the one that has the largest correlation coefficient with the response variable as well as having the strongest causal relation with the response (in the intuitive sense) will be kept.

It is important that the investigation is done for all predictors for the sake of minimal loss of information. Note that in practice, the whole set of predictors is divided into two groups, environmental and seasonal, and plotted separately, for better readability of the large scatter plot matrices. The separation is justified by the fact that most environmental variables, such as weathers, are expected to be independent of the seasonal variables, such as weekdays and holiday schedules.

At last, the above process leaves us with a small subset of the very core predictors for our beginning model: `weathersit`, `atemp` and `weekday`.

3.3 Initial Modeling

In the model building and selection process, we start from the simplest models, which have the minimal number of predictors all in the additive form, as the beginning models.

3.3.1 Beginning Models

1. For **total count**:

   $$\text{cnt} \sim \text{wkngday} + \text{weathersit} + \text{atemp} + \text{atemp}^2$$

```
##       name   rmse nrmse prc_err CV_rmse
## 1 2011 tot 717.93  0.52   25.83   730.8
```

(Table 3: eval table for total model)

2. For **registered count**:

   $$\text{reg} \sim \text{wkngday} + \text{weathersit} + \text{atemp} + \text{atemp}^2$$

3. For **casual count**:

   $$\text{cas} \sim \text{wkngday} + \text{weathersit} + \text{atemp}$$

```
##       name   rmse nrmse prc_err CV_rmse
## 1 2011 cas 309.97  0.56   72.87  314.86
## 2 2011 reg 584.55  0.55   25.37  594.36
## 3 2011 tot 722.15  0.52   25.44  734.18
```

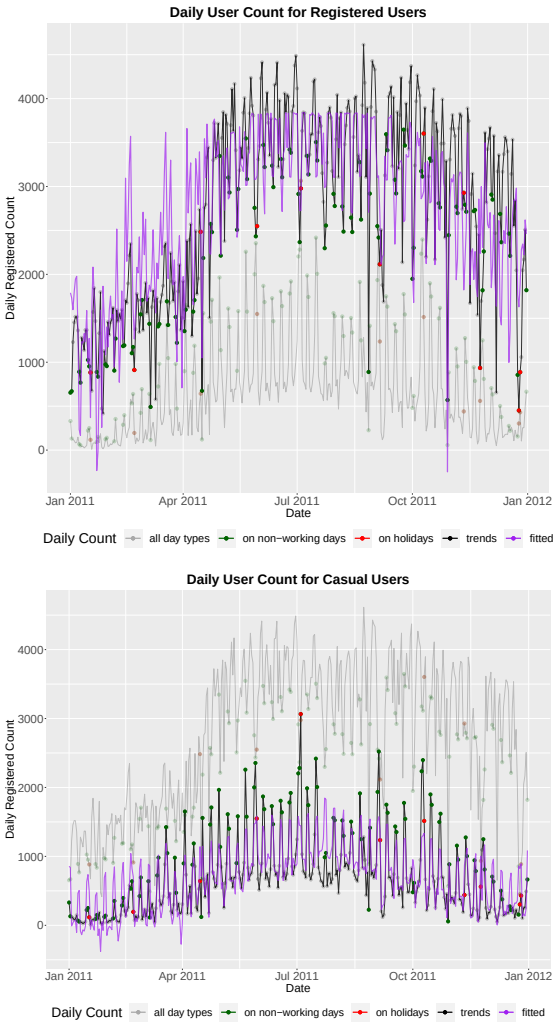
(Table 4: eval table for registered + casual model)

The percentage error of 2011 predicted total counts computed by the sum of the registered and casual models is less than that computed by the single total model (see Table 3,4), which demonstrates the power of separate modeling for registered and casual

Model	rmse	n-rmse	% Error	cv-rmse
2011 cas	309.97	0.56	72.87	314.86
2011 reg	584.55	0.52	25.37	594.36
2011 tot	722.15	0.52	25.44	734.18

Table 1: Diagnostics for Model 1

users. Therefore, we will continue with the scheme of separate modeling in the later part.



(Figure ?,?: Fitted vs Actual for 2011 cas, reg)

3.3.2 Procedures

Starting from the beginning models, we perform model building and selection in an iterative manner, for the registered and casual counts separately:

- Starting from a (relatively) simpler version of the model, the 2011 fitted response using this model is plotted along with the actual response against time (see, for example, Figure ?,?); the numeric metrics (RMSE, normalized RMSE, percentage error and LOOCV RMSE) are obtained as well (see, for example, Table ?,?).

- From the 2011 fitted versus actual plot, patterns in the biases can usually be visually recognized; combined with commonsense, this provides us with ideas of new variables potentially needed to account for the unexplained biases.

For example, for the beginning models of registered and casual counts, both response variables are being consistently overestimated in the 2011 spring and underestimated in the later part of the year (see Figure ?,?), which indicates that the users' response to weather and temperature might differ across seasons. This makes sense because each label of the weather type includes several different kinds of weathers (e.g., slightly snowy and slightly rainy days are both labeled as 3), which could cast different level of difficulties on biking activities, and it is still insufficient to fully distinguish between those weathers given only temperature information. Additionally, the level of biases also differ considerably between workdays (grey dots) and weekends (mostly green dots), which indicates that the weekday variables might also be needed.

- The new variables, one at a time, are then added to the simple model to create a more complex model. Both additive terms and interactive terms will be attempted. Then the version with the most significant improvement, according to the fitted versus actual plot and the numeric metrics, will be kept for the next round.

In this iterative process of modeling, we look at the Leave-One-Out-Cross-Validation (LOOCV) RMSE as a proxy for the extent of overfitting. The model building process stops when the LOOCV RMSE starts to ramp up as model complexity increases, typically becoming considerably larger than the RMSE (compared to the previous models).

Note that in this process, the model statistics (e.g., p-values) are also checked but are not relied on as much, out of the following two considerations:

1. There is no guarantee in the normality of residual distributions as well as correct model forms, in which case the summary statistics might thus be invalid at all, especially in the intermediate stage of modeling with the incomplete models.

2. As the model becomes more and more complex in this process, the colinearity issues worsen, which might weaken the significance of inter-correlated predictors (i.e., it might turn out that none of

those predictors are indicated as significant in the summary table), while this is not to say that none of those predictors are necessary for the model to have more accurate predictions.

3.3.3 Final Models

At the end of the iterative modeling process, we arrive at the following final models:

1. For **registered count**:

   reg ~ holiday + season:weathersit + season:workingday:atemp + atemp^2

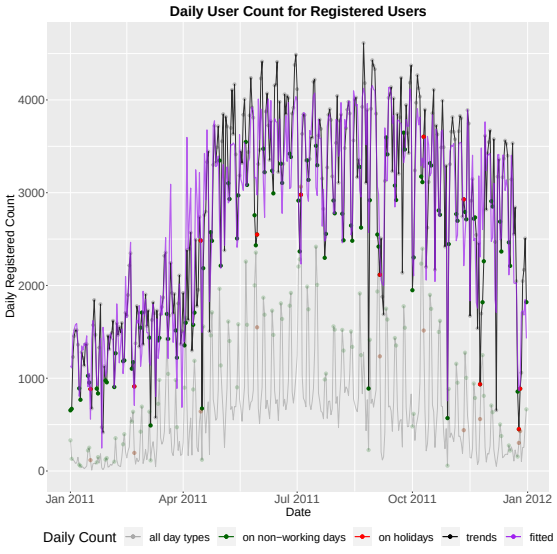
2. For **casual count**:

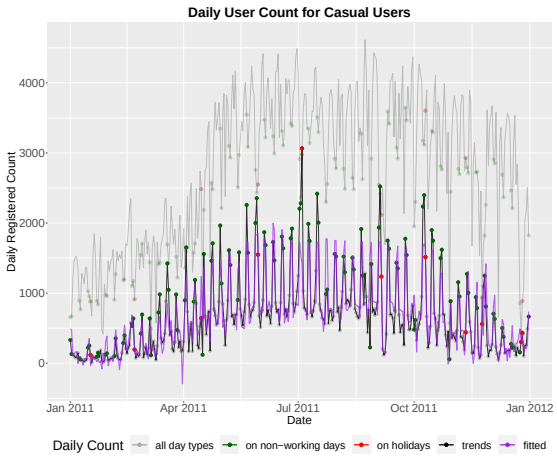
   cas ~ holiday + season:weathersit + season:workingday:atemp + season:workingday:atemp^2

```
##       name   rmse nrmse prc_err CV_rmse
## 1 2011 cas 238.01  0.43   42.57  264.81
## 2 2011 reg 392.24  0.37   14.81  441.33
## 3 2011 tot 516.80  0.37   15.72  589.74
```

(Table 5: eval table for the final models)

(For a complete list of attempted models, see Appendix.)





The final models are built with considerations of real-life experience and commonsense. For example, the interaction terms between feeling temperature, working day and seasons indicate people's different reaction to temperature change under different weathers on working days and on weekends.

From this point on, adding any other variable to the model will result in a rise in the LOOCV RMSE, which indicates an overfitting issue that weakens the predictive power of the model. Notably, using the `weekday` variable in place of the `workingday` variable also worsens the performance.

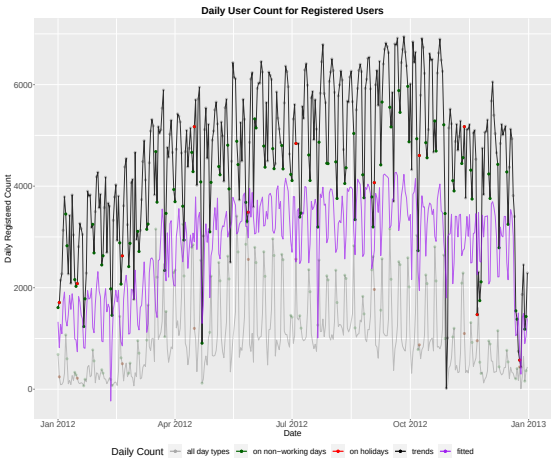
The final models achieve a ~10% improvement in the prediction percentage error in 2011 daily total counts as well as significant improvements in all the other metrics. The LOOCV RMSE is only ~15% higher than the training RMSE, indicating that the overfitting issue is still at a moderate level. Also, the systematic biases in the beginning models, as mentioned in the former section, have been alleviated (see Figure ?).

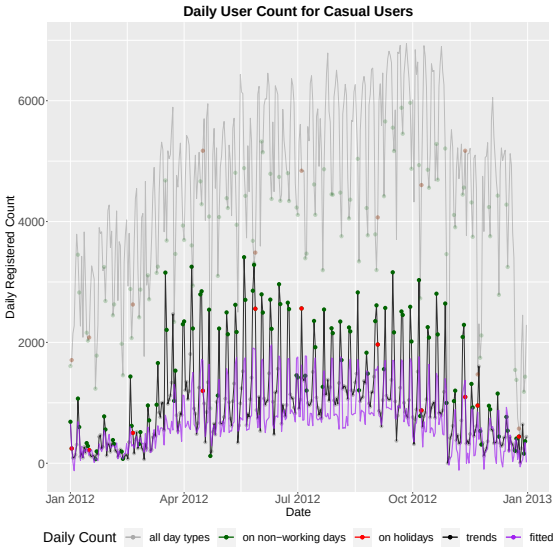
A diagnostic analysis indicates that the constant-variance assumption and normality assumption of the registered model are mostly sound, while those of the casual model are more in doubt. Accordingly, the prediction errors of the casual model is also at a higher level than the registered model (see normalized RMSE and percentage error in Table ?). This situation can be understood intuitively, as the behavior of registered users is generally more regular and can be better captured by the simple MLS model, thus more predictable.

(Figure ?: residual histogram, residual vs fitted, normal qq-plot for 2011 reg)

(Figure ?: residual histogram, residual vs fitted, normal qq-plot for 2012 reg)

3.4 Validation and Problemshooting






```
##       name    rmse nrmse prc_err
## 1 2012 cas  506.02  0.67   39.36
## 2 2012 reg 1912.08  1.36   39.71
## 3 2012 tot 2262.83  1.28   37.94
```

```
##       name   rmse nrmse prc_err
## 1 2012 cas 422.99  0.56   52.06
## 2 2012 reg 774.86  0.55   17.41
## 3 2012 tot 977.65  0.55   17.39
```

However, the 2012 rental rounts are being consistent

3.5 Refined Model

3.5.1 Prediction of the Yearly Growth Ratio

The modeling is based on the assumption that the growth trend will remain the same in the future years as that in the year of 2011. Note that this is NOT saying that the user base is supposed to remain unchanged throughout the entire year; the fact that the same scaling factor works at all points in the entire year is due to the fact that the MLS model in the later part of the year, e.g., in fall and winter, are already trained to compensate for rental count growth due to user growth using the environmental and seasonal variables.

3.5.2 Prediction without the Yearly Growth Ratio

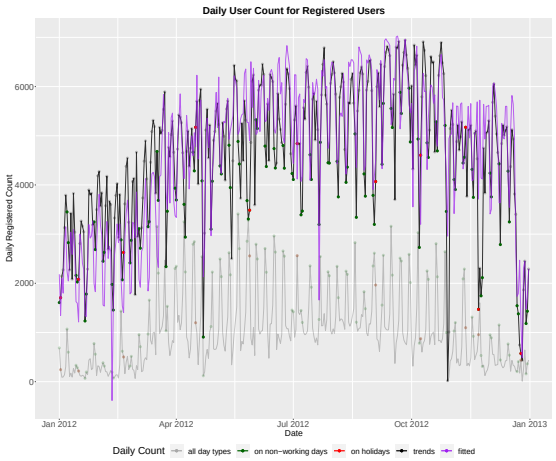
4 Prediction

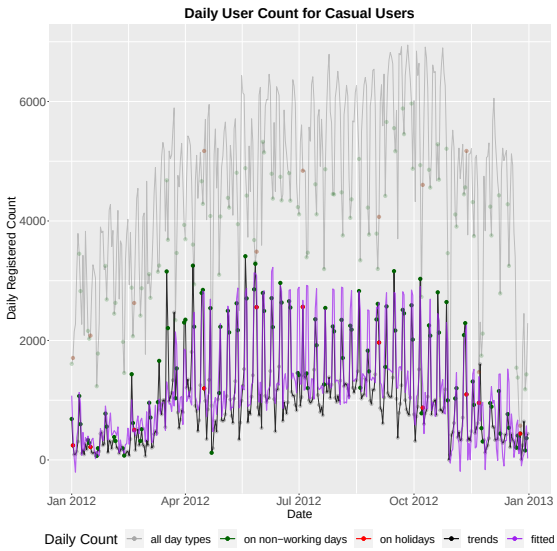
4.1 Unadjusted Model

(For the 2012 Fitted versus Actual Plot, see Figure ? in Section ?.)

(Table ?: 2012 Metrics)

4.2 Refined Model





5 Discussion

(Figure ?: 2011-2012 residual plot by time for registered counts) (Figure ?: 2011-2012 residual plot by time for casual counts)

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 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, doi:10.1007/s13748-013-0040-3.

7 Appendix

7.1 Preprocessing

7.1.1 Type Conversion

(codes here)

7.1.2 Value Conversion

(codes here)

7.2 Variable Selection

7.2.1 Predictors Selection

(Figure ?: Scatter plot matrix for Group 1 predictors)

(Figure ?: Scatter plot matrix for Group 2 predictors)

7.2.2 Response Transformation

7.3 Initial Modeling

7.3.1 Beginning Model

7.3.2 Final Model

7.4 Diagonostic Analysis

(Figure ?: 2011 fitted vs residual plot and normal qq-plot for registered counts)

(Figure ?: 2011 fitted vs residual plot and normal qq-plot for casual counts)

7.5 Validation and Problemshooting

7.6 Refined Model

7.6.1 Prediction of the Yearly Growth Ratio

7.6.2 Prediction without the Yearly Growth Ratio