CSDA 1010

Lab 2

Capital Bikeshare

***BDA Group***

Adeel Javed

Alan Yung

Lucas DiMonte

Mariam Daoud

Nadir Khan

Vivek Chaubey

June 27, 2018

# Business Problem Framing

## The bikeshare background

Capital Bikeshare is a bikeshare service in metro Washington, DC, United States. It has 4,300 bikes and more than 500 stations across Washington, DC.; Arlington, VA; Alexandria, VA; Montgomery, MD; Prince George's County, MD; and Fairfax County, VA. Bikeshare was designed for a quick and affordable way to get around.



Figure 1: a Capital Bikeshare’s station

## Capital Bikeshare’s marketing strategy

Capital Bikeshare conducts biannual membership surveys to assess how people use the system and what impact it has on their lives and their communities. To improve their marketing strategy the managers of Capital Bikeshare need to answer the following questions:

1. Where do Capital Bikeshare riders go?
2. When do they ride?
3. How far do they go?
4. Which stations are most popular?
5. What days of the week are most rides taken on?

## Capital Bikeshare data analytics team’s task

The data analytics team must build a data-driven model to answer the questions above by using data that have been collected since Capital Bikeshare was launched in 2010.

# Analytical Problem Framing

Based on the stated needs of Capital Bikeshare managers, the solution must separate those events when customers rent a bicycle from those who will not.

# Data Understanding

## Data Source

The raw data for modeling Capital Bikeshare customers’ rental behavior were retrieved from the UCI Machine Learning Repository at https://archive.ics.uci.edu/ml/datasets/bike+sharing+dataset. This set of raw data was captured from the first bike trip, and the data are stored in the “day.csv” file.

## Data Dictionary

The retrieved raw data contains the attributes as shown in Table 1 below.

Table 1: description of the raw set of data retrieved from Capital Bikeshare’s database for analysis and modeling

|  |  |  |
| --- | --- | --- |
| Attribute | Description | Data type |
| instant | record index | numeric |
| 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) | numeric |
| casual | count of casual users | numeric |
| cnt | count of total rental bikes including both casual and registered | numeric |
| dteday | date | datetime |
| holiday | weather day is holiday or not (extracted from [[Web Link]](http://dchr.dc.gov/page/holiday-schedule)) | numeric |
| hum | Normalized humidity. The values are divided to 100 (max) | numeric |
| mnth | month ( 1 to 12) | numeric |
| registered | count of registered users | numeric |
| season | season (1:springer, 2:summer, 3:fall, 4:winter) | numeric |
| 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) | numeric |
| 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 | numeric |
| weekday | day of the week | numeric |
| windspeed | Normalized wind speed. The values are divided to 67 (max) | numeric |
| workingday | if day is neither weekend nor holiday is 1, otherwise is 0. | numeric |
| yr | year (0: 2011, 1:2012) | numeric |

## Data feature engineering and selection for the classification model

The raw data from launch in 2010 was collected by the company. The raw dataset contained observations and variables. The selected target variable for prediction is the “cnt” as shown in Table 2 below.

Table 2: the target variable indicating a bike rental event

|  |  |  |
| --- | --- | --- |
| Attribute | Description | Data type |
| cnt | count of total rental bikes including both casual and registered | numeric |

# Exploratory Data analysis

## Data visualization

We conducted an exploratory data analysis (EDA) in order to gain insights into the dataset.

### Scatter plots

We plotted pairwise scatter plots and inspected the result for relationships between the dependent target variable “cnt” and the numerical independent variables. Based on the attributes’ relationships, we excluded attributes that have correlation with other input attributes.

Scatter plots for the variables “cnt” with “season”, “mnth”, “holiday”, “weekday”, and “workingday” are shown in Figure 2 below. As these attributes were originally categorical, we did not seek to find a linear relationship among them. Therefore, we show only the scatter plots between the target variable “cnt” and the time attributes.

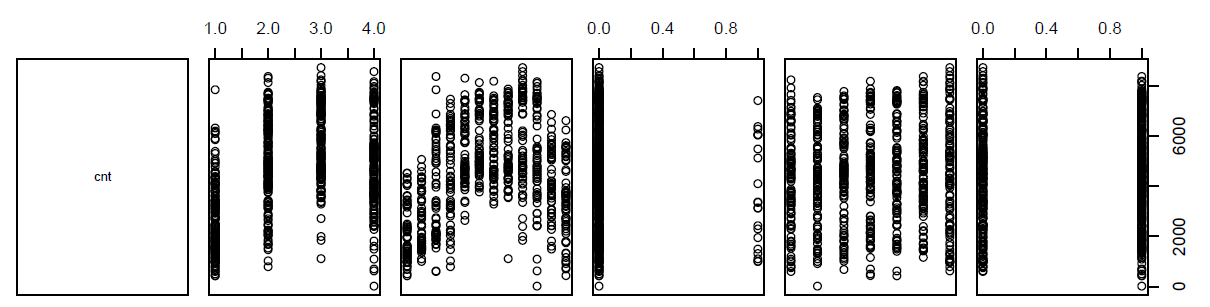


Figure 2 shows the scatter plots for the target variable with the input variables

Based on the scatter plots for “cnt” with “season”, “mnth”, “holiday”, “weekday”, and “workingday”, we see that “cnt” increases when season is summer, where 1 represents spring, 2 represents summer, 3 represents fall, and 4 represents winter. The variable “Cnt” also increases when the month starts with 5. This means that when spring starts, the temperature is high up until month 10 or mid-fall season. “Cnt” is much higher when it is a holiday. For the last two scatter plots, “Cnt” seems to be invariant with respect to weekday, from Monday to Sunday, and workingday.

The scatter plot in Figure 4 below shows the relationship between count and the weather related attributes, that is, pairs(~cnt + weathersit + temp + atemp + hum + windspeed).

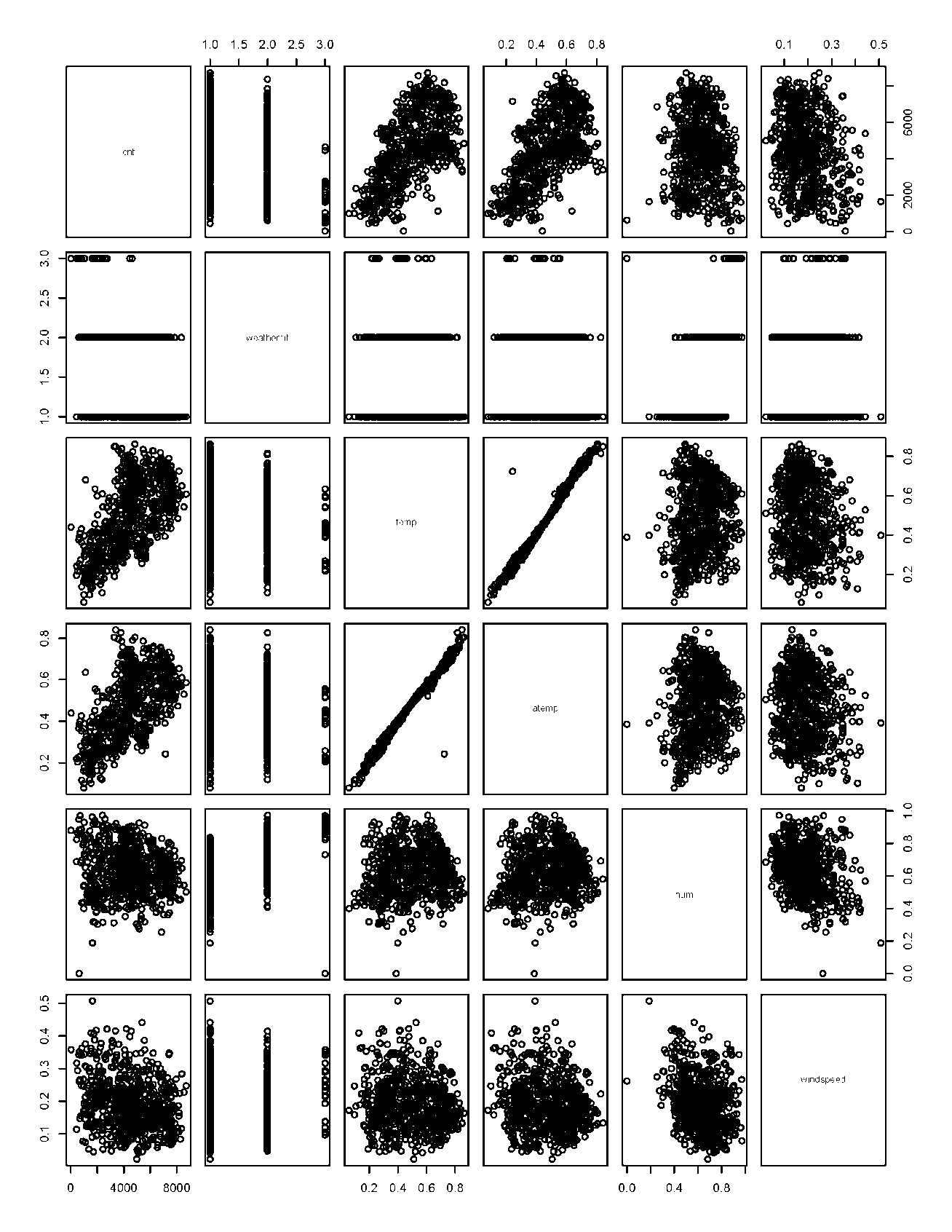


Figure 3 shows the relationship between count and the weather related attributes

We found that there is a linear relationship between the attributes “temp” and “atemp”. We derived these attributes are follows:

|  |
| --- |
| 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) |

Based on the scatterplot between temp and atemp, there is a linear relationship between the two attributes. This is referred to as the Multicollinearity. As regression modeling is based on the assumption that there must be no correlation among independent variables, we need to remove one of these two variables. We decided to remove the “atemp” attribute.

In terms of the relationship between “cnt” and the weather attributes, by looking at the first row of plots, we concluded the following relationships:

In relation with weathersit, “cnt” is high when weathersit = 1 or 2. Weathersit takes one of the following values:

|  |
| --- |
| 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 clouds4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog |

When there is Heavy Rain, Ice Pallets, Thunderstorm, Mist, Snow, Fog, the count of rental bikes is not zero. Moreover, in relation with the temperature, “cnt” increases when the temperature increases. In relation with humidity and windspeed, there is no noticeable linear increase in the count with respect to humidity and windspeed.

### Correlation within input variables

We found that correlation exists within the input variables. Correlations within the input variables are shown below in Figure 4. The correlations are:

1. Holiday is correlated with weekday and workingday.
2. Winspeed is correlated with hum, temp, atemp, season, and mnth.
3. Weathersit is correlated with hum, temp, and atemp.
4. Hum is correlated with temp, atemp, season, and mnth.
5. Temp is correlated with atemp.
6. Atemp is correlated with season and mnth.
7. Season is correlated with mnth.

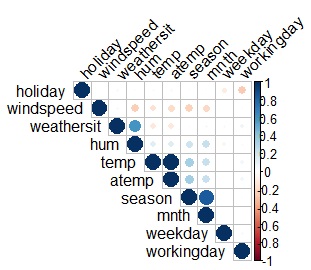


Figure 4 shows correlation within the input variables

# Data Preparation

The dataset day.csv contains 731 rows and 16 columns with the following attributes:

1. Instant: record index.
2. Datetime: date and hour in "mm/dd/yyyy and hh:mm" format.
3. Season: Four categories-> 1 = spring, 2 = summer, 3 = fall, 4 = winter. Integer
4. Year (0: 2011, 1:2012)
5. Month (1 to 12)
6. Hour (0 to 23) Holiday: whether the day is a holiday or not (1/0).
7. Weekday: day of the week. Integer.
8. Working day: whether the day is neither a weekend nor holiday (1/0). Integer.
9. Weather: Four Categories of weather. Integer

* Clear, Few clouds, Partly cloudy, Partly cloudy.
* Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
* Light Snow and Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
* Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

1. Temp: hourly temperature in Celsius. Numeric.
2. ATemp: "feels like" temperature in Celsius. Numeric.
3. Humidity: relative humidity. Integer.
4. Windspeed: wind speed. Numeric.
5. Registered: number of registered user. Numeric.
6. Casual: number of non-registered user. Numeric.
7. Count: number of total rentals (registered + casual). Numeric.

The following non-influential attributes were removed:

1. Instant: record index.
2. Datetime: date and hour
3. Year (0: 2011, 1:2012)

## Derived attributes

The following two attributes casual and registered added up to the total daily count of rental bikes (count). As our goal is to predict the total daily count of rental bikes (cnt), we excluded those two attributes for the scatter plots.

* casual: count of casual users
* registered: count of registered users

# Model Development

## Feature selection

We need to manage the potential predictor variables, fine tune the model, and choose the best predictor variables from the available attributes in the dataset. To remove the least significant variable, we decided to use the stepwise linear regression (Automated F-test-based backward selection). We found that the most important attributes that would contribute to predict the daily number of rental bikes are “holiday”, “temp”, “hum”, “windspeed”, “jan”, “feb”, “mar”, “june”, “jul”, “aug”, “sept”, “oct”, “nov”, “weather1”, and “weather2”.

## Regression tree modeling

The regression tree was fitted to the dataset contained in the “day.csv” file. It shows the top nodes on which the tree was split. Those top nodes are the most important variables within the dataset on which decisions were made. They are the “temp”, ‘hum”, and “windspeed”.

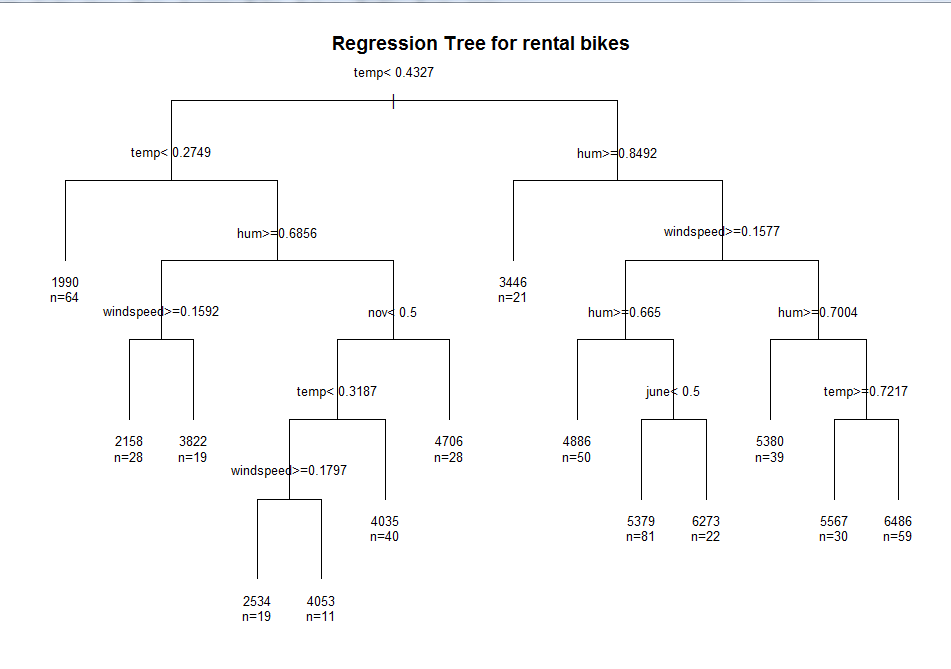


Figure 5 shows the regression tree model

The regression tree shows that the “temp” variable is the most important attribute to branch at the root node of the tree. The tree is split at the root node. The left of the tree is for “temp” less than 0.4327 and the right side of the tree is for “temp” greater than 0.4327. The IF-THEN-ELSE rules are followed to predict the total number of bikes rented daily. An example of the IF-THEN-ELSE rules is shown below.

|  |
| --- |
| Starting at the root node of the tree, IF temp is less than 0.4327, go to the left side of the tree.  IF temp is less than 0.2749 THEN go to the left side of the node, a leaf of the tree will be reached. The predicted total number of bikes rented would be 1990.  IF temp is greater than 0.4327 and hum is less than 0 .8492 THEN a leaf will be reached where the total number of bikes rented is predicted to be 3446. |

The regression tree model was fitted by using the train data set and then tested by using the test data set. The Root Mean Square Error (RMSE) was found to be 1387.381. The Mean Absolute Percentage Error (MAPE) was found to be 0.36275, and the correlation was found to be 0.7481815.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Root-mean-square error (RMSE)** | **Mean absolute percentage error (MAPE)** | **Correlation** |
| Regression tree | 1387.381 | 0.36275 | 0.7481815 |

To get a better look at what's happening; we plotted the predicted and the actual data points on a scatter plot. Figure 6, shows a strong correlation between the actual and the predicted values in the test data set. The actual versus predicted plot shows that there is a constant linear pattern at several estimated levels of the daily rate of rental bikes. For example, there is a constant daily rate of bikes rented at the 2000, 4000, and 5500.

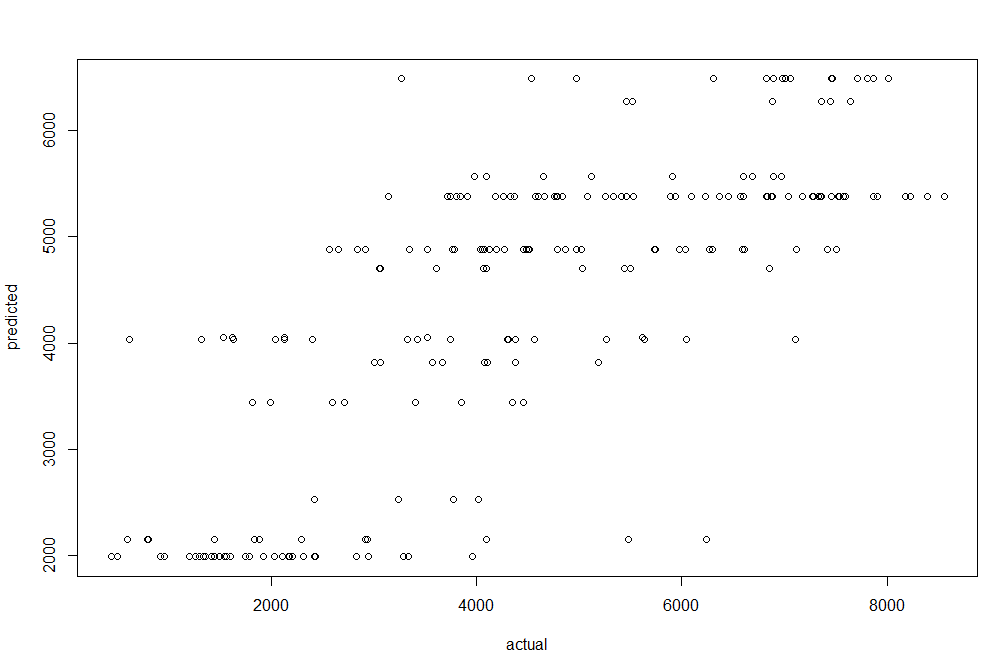


Figure 6 shows the scatter plot of actual versus predicted values of the test data set

# Linear Regression modeling

We selected the Linear Regression model for its simplicity, to easily show and understand the demand for rental bikes. The linear regression model shows the demand for the number of bikes (count) as follows:

The categorical attributes are weathersit (clear, mist, light rain, heavy rain/snow) and months with 12 variables (Jan-Dec).

We used the Linear Regression model to find the equation that fits the dataset. Once we have the regression equation, we can use the model to make predictions.

The correlation coefficient shows that the data is likely to be able to predict future bike rentals and a scatter plot of the data appears to form a straight line, as shown in Figure 7 below. There is a strong linear relationship between the predicted and actual bike rental demand. The simple linear regression can be used to find a predictive model. The equation for the line that fits the dataset is written in the form of count = β0 + β1 X + c. That is, the bike rental count is estimated as follows:

|  |
| --- |
| Count = β0+ β1\*holiday + β2\* temp + β3\* hum + β4\*windspeed + β5\* jan + β6\*feb + β7\*mar + β8\*june + β9\*jul + β10\*aug + β11\*sept + β12\*oct + β13\*nov + β14\*weather1 + β15\*weather2 + β16\*mon. |

|  |
| --- |
| Call: |
| lm(formula = cnt ~ holiday + temp + hum + windspeed + jan + feb + mar + june + jul + aug + sept + oct + nov + weather1 + weather2, data = train\_set) |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Residuals: | | | | |
| Min | 1Q | Median | 3Q | Max |
| -3514.7 | -904.2 | -195.3 | 1022.6 | 3284.2 |

## Coefficients:

Estimate Std. Error t value Pr(>|t|)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Count |  |  |  |
| Intercept | 2405.45 | 729.04 | 3.299 | 0.00104 \*\* |
| holiday | -795.52 | 335.17 | -2.373 | 0.01800 \* |
| temp | 7055.28 | 639.68 | 11.029 | < 2e-16 \*\*\* |
| hum | -3393.83 | 567.17 | -5.984 | 4.19e-09 \*\*\* |
| windspeed | -4282.22 | 865.63 | -4.947 | 1.04e-06 \*\*\* |
| jan | -889.38 | 271.14 | -3.28 | 0.00111 \*\* |
| feb | -916.08 | 244.49 | -3.747 | 0.00020 \*\*\* |
| mar | -289.65 | 224.05 | -1.293 | 0.19668 |
| june | -662.26 | 264.79 | -2.501 | 0.01271 \* |
| jul | -1325.53 | 291.07 | -4.554 | 6.64e-06 \*\*\* |
| aug | -798.14 | 266.91 | -2.99 | 0.00293 \*\* |
| sept | 36.08 | 264.51 | 0.136 | 0.89154 |
| oct | 653.61 | 238.77 | 2.737 | 0.00642 \*\* |
| nov | 361.55 | 232.15 | 1.557 | 0.12002 |
| weather1 | 1987.46 | 448.29 | 4.433 | 1.14e-05 \*\*\* |
| weather2 | 1847.59 | 420.82 | 4.39 | 1.38e-05 \*\*\* |

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1262 on 494 degrees of freedom

Multiple R-squared: 0.5621, Adjusted R-squared: 0.5479

F-statistic: 39.63 on 16 and 494 DF, p-value: < 2.2e-16

We can interpret the demand for bikes in based on a change in

|  |  |  |
| --- | --- | --- |
|  | **Mean absolute percentage error MAPE** | **Correlation** |
| **Linear regression** | 0.3402147 | 0.7513746 |

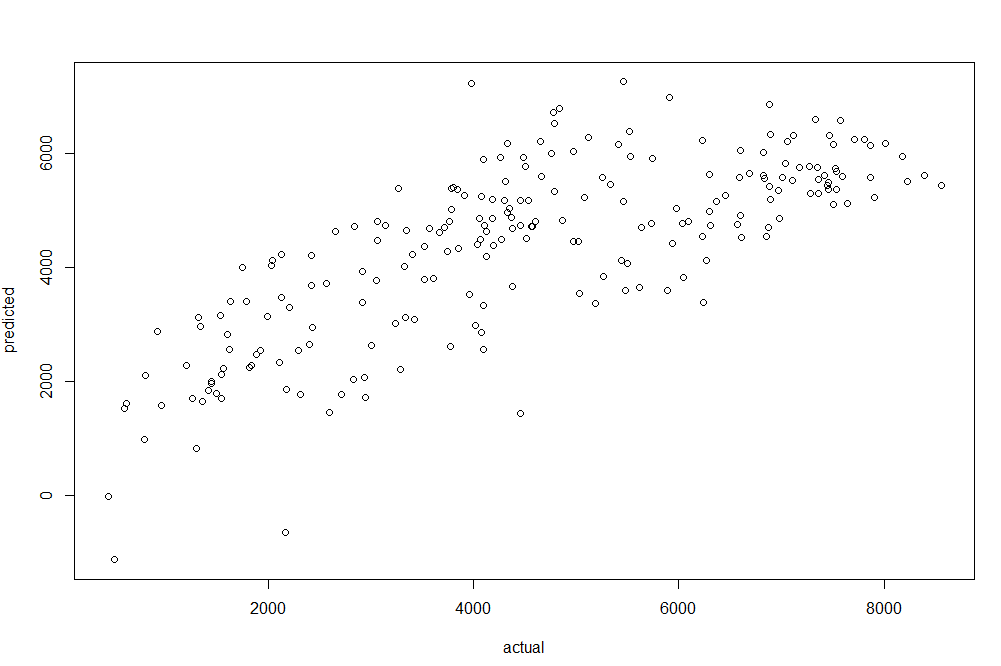


Figure 7 shows the scatter plot of the actual vs predicted bike rentals appears to form a straight line

# Model Deployment

Deployment and demonstration of the predictive model to the Capital Bikeshare would clarify:

1. purpose and goals,
2. limitations of the selected dataset, understanding, analysis, preparation and cleansing,
3. challenges in model selection, preparation, execution,
4. Interpretations, communication of results, learnings, and insights.

## Monitoring

Monitoring the performance of the classifier means communicating regularly with Capital Bikeshare stakeholders and obtaining feedback. Then to use the feedback received to fine tune the model.

## Review

Continuously review the outcome to the predictive results. Feed more up-to-date data into the model to retrain its algorithm and improve its accuracy. Then, update the model in production to improve the model’s success.