Analytics Foundations: Problem Set 13

Today's dataset comes from a bike sharing company (Capital Bike Share). Each *hour*, the number of riders (**cnt**) is given, along with various other attributes as shown in the table below:

cnt	Count of total rental bikes including both casual and registered
dteday	Date
instant	Record index (ID)
season	Season (1:winter, 2:spring, 3:summer, 4:fall)
yr	Year (0:2011, 1:2012)
mnth	Month (1 to 12)
hr	Hour (0 to 23)
holiday	Whether 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. Values are divided to 41 (max)
atemp	Normalized feeling temperature in Celsius. Values are divided to 50 (max)
hum	Normalized humidity. Values are divided to 100 (max)
windspeed	Normalized wind speed. Values are divided to 67 (max)
casual	Count of casual users
registered	Count of registered users

Source: http://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset#

Data:

You can obtain the dataset by running the following code:

```
bike <- read.csv('https://raw.githubusercontent.com/IAA-
Faculty/statistical_foundations/master/bike.csv')
```

Questions:

1. Run the following code to get the training and test split:

```
set.seed(123)

bike <- bike %>% mutate(id = row_number())

train <- bike %>% sample_frac(0.7)

test <- anti_join(bike, train, by = 'id')
```

2. There are abnormal times where the number of casual users is greater than or equal to the number of registered users. You can use the following code to create a variable **casual_high** that captures this:

```
train$casual_high <- train$casual >= train$registered
```

3. Build a logistic regression model to predict the probability that we have these abnormally high number of casual users. Feel free to use any of the other *predictor* variables in your data set to do so. (HINT: **cnt, casual,** and **registered** are NOT predictor variables). Use p-value backward selection with a significance level of 0.001. What variables do you end up with at the end? Interpret one of the odds ratios from your result. (HINT: Careful about using variables that would be perfectly correlated. For example, if I know the month of the year, then I automatically know which season as well.)

SIDE NOTE: This data set actually has a problem with a rare number of events. Anything less than 5% of a target category typically requires us to do *rare-event sampling* as well as model adjustments to account for this. DO NOT WORRY ABOUT THIS FOR THIS PROBLEM! We will address all these things in the Fall semester.