UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

ASSIGNMENT 2

The goal of this assignment is to get you started with predictive analytics. You will first prepare and explore the data, and run a basic regression. You will then predict the variable COUNT as a function of the other variables. You will also determine the effect of bad weather on the number of bikes rented. Finally, you will build alternative models, measure and compare their predictive performance, make *data-informed* and *data-driven* inferences for a business case.

Assignment Instructions

You will use data from DC's Capital Bikeshare (also serves Maryland and Virginia). Capital Bikeshare has about 30K members, and served about 23.6 million trips through its 550 stations. In this dataset, we combined the Capital Bikeshare data with weather data to gather insights.

Data Dictionary:

1. DATE -*You'll also create a* MONTH *variable using this*
2. HOLIDAY: Whether the day is a U.S. holiday or not.
3. WEEKDAY: If a day is neither a weekend nor a holiday, then WEEKDAY is YES.
4. WEATHERSIT: The values are (1) Clear/Few clouds (2) Misty (3) Light snow or light rain (4) Heavy rain, snow, or thunderstorms.
5. TEMP: Average temperature in Celsius.
6. ATEMP: "Feels like" temperature in Celsius.
7. HUMIDITY: Humidity out of 100 (not divided by 100).
8. WINDSPEED: Wind speed in km/h.
9. CASUAL: Count of bikes rented by casual bikeshare users.
10. REGISTERED: Count of bikes rented by registered bikeshare members.
    COUNT: Total count of bikes rented by both casual users and members **-You'll create this**

Before you start:

- Load the following four libraries in the given order: *tidyverse, tidymodels, plotly, skimr*
- Load the bikeshare data and call it *dfb0rg*
- Explore the dataset using `skim()` etc.

1) **Data preparation**
    a) **Create the additional variables:**
        i) Create the COUNT variable and add it to the data frame.
        ii) Extract MONTH from the DATE variable and add it to the data frame. **This time, do NOT use lubridate. Use the base `months()` function instead.**
    b) **Scale the data (and save it as *`dfbStd`*):** Start by standardizing the four variables, TEMP, ATEMP, HUMIDITY, WINDSPEED. If you don't remember what it means to standardize a variable, see the link. Surely, you don't need to do this manually!

2) **Basic regression in R:** In `dfbStd`, run a regression model `fitAll` using COUNT as the DV, and all the variables as independent variables. [Don't forget to use summary(fitAll)]
    a) Does this appear to be a good model? Why or why not?
    This seems to be a good model as the correlation with dependent variable is high. The value of multiple R square and adjusted R square is 1 which is very good. This means that the model explains 100% of the variance and is perfectly fit. The adjusted R square is 1 which means that the independent variables accounted for are highly accurate. Also some of the variable have negligible correlation with count because of the fact that correlation values are in the negative power of 10.
    b) According to your model, what is the effect of humidity on the total bike count in a formal interpretation? Does this finding align with your answer to Part (a)?
    The effect of humidity seems to be positive. A unit increase in humidity would cause 1.4*10^-13 increase in bike count. Even if the value is negligible, it still has an effect on total bike count. Therefore, yes it does align with the answer in part a.

In the rest of the assignment, use the original data frame `dfbOrg`:

3) **Working with data and exploratory analysis:**
    a) Add a new variable and call it **BADWEATHER**, which is "YES" if there is light or heavy rain or snow (if WEATHERSIT is 3 or 4), and "NO" otherwise (if WEATHERSIT is 1 or 2). You know what functions to use at this step.
    b) Present a scatterplot of COUNT (y-axis) and ATEMP (x-axis). Use different colors or symbols to distinguish "bad weather" days. Briefly describe what you observe.
    We observe that when the temperature feels very low, it causes the bike count to go down till it reaches the minimum value of 10. Also, when there are harsh weather conditions, the count is low.
    Pleasant weather conditions correspond to a higher number of bikes till 25. Also as the temperature reaches too high, we see less count of rentals.
    c) Make two more scatterplots (and continue using the differentiated coloring for BADWEATHER) by keeping ATEMP on the x-axis and changing the variable on the y-axis: One plot for CASUAL and another for REGISTERED. Given the plots:

i) How is *temperature* associated with casual usage? Is that different from how it is associated with registered usage?

The registered users seem to use a greater number of bikes during bad weather conditions at zero temperatures whereas non registered don't. Both plots touch 0 count when the bad weather has a temperature of around 10. Also, in general registered users have more usage than non-registered users.

ii) How is *bad weather* associated with casual usage? Is that different from how it is associated with registered usage?
During bad weather, the registered users seem to have more usage. At approx. 10 both the type of users reaches the zero. Registered users reach 3000 as compared to casual that reach 1000. In conclusion, Registered users have more usage in general.

iii) Do your answers in (i) and (ii) make logical sense? Why or why not?
They do make logical sense since extremely low or extremely high temperatures would discourage people from going on a bike. But also, since registered users feel obligated to use the service their numbers are high whereas casual users would prefer other comfortable means since they aren't obligated to use the service.

iv) Keep ATEMP in the x-axis, but change the y-axis to COUNT. Remove the color variable and add a geom_smooth() without any parameters. How does the overall relationship between temperature and bike usage look? Does this remind you of Lab 2? Why do you think the effects are similar?
In lab 2, we used geaom_smooth to visualize the relationship between CPI and weekly sales. When CPI was low the sale was high, similarly when the temperature is good enough the count is the maximum. Also, like individual stores had unique line, each temperature has a unique Count.

4) **More linear regression:** Using dfbOrg, run another regression for COUNT using the variables MONTH, WEEKDAY, BADWEATHER, TEMP, ATEMP, and HUMIDITY.

a) What is the resulting adjusted R^2? What does it mean?
The resulting adjusted R-squared is 0.521.
There is 52.1% correlation of COUNT with the variables which are given above. COUNT can change depending on correlation between the independent and dependent variables. The value of adjusted R-square is completely dependent on the independent variables like TEMP, ATEMP, etc.

b) State precisely how BADWEATHER is associated with the predicted COUNT.
BADWEATHER is negatively associated with the predicted COUNT since its coefficient is -1954.835. This means that an unit increase in bad weather value will cause an average of 1954.835 decrease in Count assuming that rest of the variables are constant.

c) What is the predicted count of rides on a weekday in January, when the weather is BAD, and the temperature is 20° and feels like 18°, and the humidity is 60%?
We can calculate this as follows :
COUNT = (3967.981) + (-858.3)*MONTH + (69.745)*WEEKDAY + (-1954.835)* BADWEATHER + (184.596)*TEMP + (-48.640)*ATEMP + (-25.341)*HUMIDITY
COUNT = 2520.497

d) Do you have any concerns about this model or your predicted COUNT in **Q4-c**? Why or why not?
The adjusted R square and the R square value have improved but not as significantly. This is not a good value though considering the addition of new variables. The independent variables should have improved the R square value but the model has not improved.

5) **Regression diagnostics:** Run the regression diagnostics for the model developed in **Q4**. Discuss whether the model complies with the assumptions of multiple linear regression. ***If you think you can mitigate a violation, take action,*** and check the diagnostics again.
Hint: The Q-Q plot and the other diagnostics from the plot() function look fine to me!
In the residual vs. fitted graph, we can say the model is appropriate since the main regression curve is intercepting the y = 0 residuals multiple times.
The combination of the points in the Q-Q plot is intercepting the 45 degree slope in the middle range which suggests validity of the above model.
In the scale location plot, there are clusters near the regression curve which suggests validity of the above model as well.
The clusters in the residual vs. leverage graph have been formed around the least leverage which is good and the regression curve intercepts y = 0 multiple times.

6) **Even more regression:** Run a simple linear regression to determine the effect of bad weather on COUNT when **none** of the other variables is included in the model.
a) Compare the coefficient with the corresponding value in **Q4**. Are they different? Why or why not?
Q4 had the coefficient of the value -> (-1954.835)
Here, the bad weather has the coefficient (--2780.95) whereas Q4 had the coefficient of the value -> (-1954.835) which was better. There has been only one independent variable in this case whereas in the previous Q4 case there were more. Since there is only one independent variable and we know that bad weather is most negatively co related with count.

b) A consultant has indicated that bike use is affected differently by bad weather on weekdays versus non-weekdays, as people go to work on weekdays. How can you add this domain knowledge to the regression model you built in (a)? Why?

If we add the weekday as an independent variable in the model, when we observe the coefficient of weekday (YES) we are able to see the bike usage being affected due to bad weather, since more people would need bikes on a weekday to travel to work while not on weekends as not so many people need to travel for work.

c) Run a new model with your addition from (b). Is this a better or worse model than your original model in (a)? How do you decide?
Our model is similar to the old one, the r-squared value has increased by 0.01. Therefore, it hasn't improved much.

d) Using your model from (c),
   i) Interpret the average effect of bad weather on the COUNT depending on whether it is a weekday or not, and
   interpret the average effect of bad weather on the COUNT depending on whether it is a weekday or not, and

   COUNT = 4452.5 - 2637.1*BADWEATHERYES + 185.3*WEEKDAYS - 201.2*BADWEATHERYES:WEEKDAYYES

   Case 1 : BADWEATHER AND WEEKEND
   COUNT= 4452.5 - 2637.1*BADWEATHERYES + 185.3*WEEKDAYS - 201.2*BADWEATHERYES:WEEKDAYYES
   = 4452.5 - 2637.1
   = 1815.4 ~ 1815

   Case 2 : BADWEATHER AND WEEKDAY

   COUNT= 4452.5 - 2637.1*BADWEATHERYES + 185.3*WEEKDAYS - 201.2*BADWEATHERYES:WEEKDAYYES
   = 4452.5 - 2637.1 + 185.3 - 201.2
   = 1799.5 ~ 1800

   ii) quantify the effect of bad weather on the COUNT in different scenarios (be sure to calculate *all* effect sizes for the **four alternatives (2x2)** here).
   *[In calculating the effects here, do **not** worry about the statistical significance]*
   Case 1 : No BADWEATHER AND WEEKEND

   COUNT= 4452.5 - 2637.1*BADWEATHERYES + 185.3*WEEKDAYS - 201.2*BADWEATHERYES:WEEKDAYYES
   = 4452.5

   Case 2 : No BADWEATHER AND WEEKDAY

COUNT= 4452.5 - 2637.1*BADWEATHERYES + 185.3*WEEKDAYS - 201.2*BADWEATHERYES:WEEKDAYYES
= 4452.5 + 185.3
= 4637.8 ~ 4638

Case 3 : BADWEATHER AND WEEKEND

COUNT= 4452.5 - 2637.1*BADWEATHERYES + 185.3*WEEKDAYS - 201.2*BADWEATHERYES:WEEKDAYYES
= 4452.5 - 2637.1
= 1815.4 ~ 1815

Case 4 : BADWEATHER AND WEEKDAY

COUNT= 4452.5 - 2637.1*BADWEATHERYES + 185.3*WEEKDAYS - 201.2*BADWEATHERYES:WEEKDAYYES
= 4452.5 - 2637.1 + 185.3 - 201.2
= 1799.5 ~ 1800

7) **Predictive analytics:** Follow the steps below to build two predictive models. Which model is a better choice for predictive analytics purposes? Why? Does your conclusion remain the same for explanatory analytics purposes? Please copy and paste the predictive and explanatory performance levels of both models into your response.
    a) Set the seed to **333** (Always set the seed and split your data in the same chunk!).
    b) Split your data into two: 80% for the training set, and 20% for the test set
        i) Call the training set *dfbTrain* and the test set *dfbTest.*
    c) Build two different models, calculate, and compare performance.
        i) The first model will include the variables in **Q4 with any adjustments you may have made during the diagnostics tests in Q5** (call this one *fitOrg*). The second model will add WINDSPEED to this model -*Call it fitNew*.
        Including the WINDSPEED, the new model is markedly better for the task of predictive analysis.
        The RMSE and MAE values for the fitNew model are lower than the corresponding values for the fitOrg model.
        In the fitNew model, we can see that the R-squared value has also improved compared to the fitOrg model.

    Hint: Remember, every time you build a new model, there are three steps you need to follow to be able to calculate the predictive performance of the model:
        i. Build the model and store it as *fitXxx*
        ii. Create a new copy of the test dataset *dfbTest* by adding the predicted values as a new column. Name this new dataframe as *resultsXxx*

iii. Calculate the performance measures (RMSE and MAE) using the actual and predicted values stored in the results dataframe *resultsXxx*

-*You'll replace* Xxx *with the model names you use (*Org *&* New *are suggestions)*

You may have trouble with the metric_set() function if you used `modelr` in Q5 for the diagnostics test. Trouble means learning. If you run the following code, you can simply ask R to unload `modelr` and you'll be fine: **`detach('package:modelr', unload=TRUE)`**

8) **More predictive analytics:** In this final question, experiment with the time component. In a way, you will almost treat the data as a time series. We will cover time series data later, so this is just a little experiment. Taking into account date, you can't split your data randomly (well, evidently, you would not want to use future data to predict the past). Instead, you have to split your data by time. Start with `dfbOrg` and **use the variables you used in `fitOrg` from Q7c**. Split your data into training using the year "2011" data, and test using the "2012" data. Has the performance improved over the random split that assumed cross-sectional data (which you did in the previous questions)? Why do you think so? Split again by assigning 1.5 years of data starting from January 1st, 2011 to the training set and the remaining six months of data (the last six months) to the test set. Does this look any better? Discuss your findings.
Since our RMSE and MAE have increased, the new model is not better than the original model.
The error being higher nullifies the effect of the R-squared value being better, therefore, the new model is not better than the original model for predictive analysis.

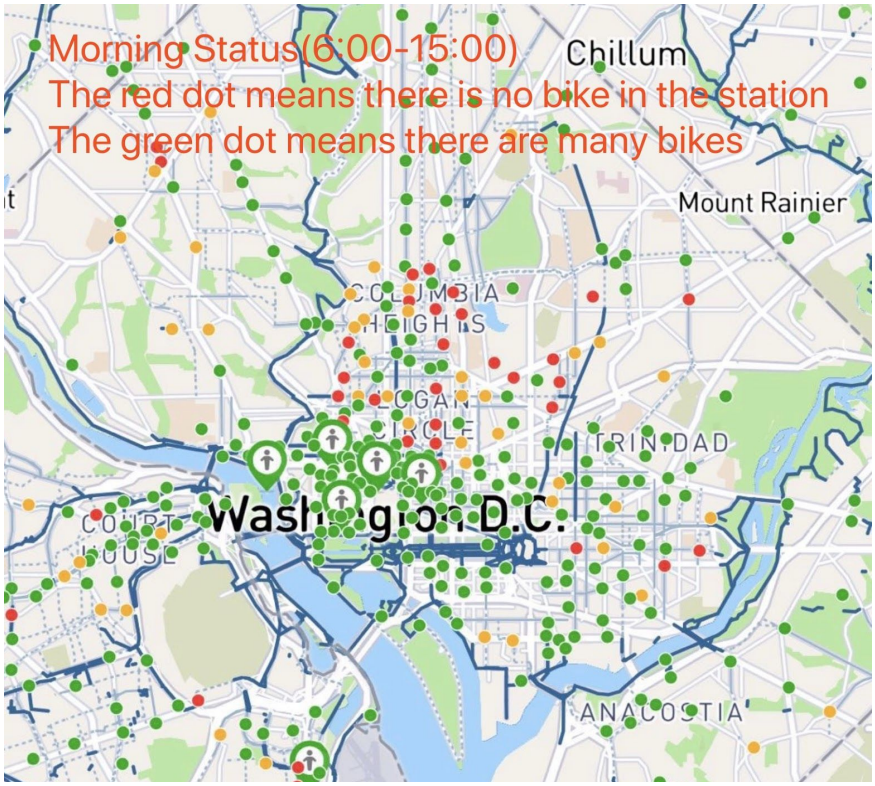
9) **Data-*informed* decision making:** Based on your quick analysis of the Capital Bikeshare data, what are some actions you would take if you were managing Capital Bikeshare's pricing and promotions? How do you think you would use your predictions?
I think that The Capital Bikeshare company could try to employ a policy that ensures that a bike station has bikes available based on the current weather including temperature.
Adverse weather suggests calling for a higher number of bikes to be stationed at the centers with very high rush hour traffic.
We can also suggest a price drop when there are adverse weather conditions to make sure that customers will use their bikes.
Therefore, we can have a high customer retention rate while acquiring new customers in the process.

10) **Data-*driven* solutions to "the" big challenge of bikeshare:** As shown in the visuals on the next page, Capital Bikeshare (like most other shared services) has an inherent challenge. In the morning, people use bikes to commute to their workplaces, leaving the

bike racks empty in residential areas (this is called *rush-hour surge*). In the evening, the same phenomenon repeats in the opposite direction. Shared-service companies attempt to resolve this problem by *rebalancing*, which is basically moving bikes manually during the off-peak hours using trucks (which you may have seen on the streets) and other means. **Assuming you have access to all the data Capital Bikeshare collects, and you can collect new data**, what is a data-driven solution you would pursue? Be specific about the data you would collect (if any) and the analytics project/model you would use.

We could incentivize people to transfer the bikes from one location to the other. We can target the regular users of the app, by providing them with incentives for their next bike rides or offer current rides at cheaper prices for destinations that need the bikes to be moved there.

Morning -Green dots are stations with many bikes, red ones are those with no bikes:



Evening -Green dots are stations with many bikes, red ones are those with no bikes: