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?
    R-squared value is 1. This model is overfitting and hence not a good model.
    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)?
    With 1 unit increase in humidity, there is an extremely small change in the total bike count, keeping all other variables constant. Humidity has almost no effect on the total count. This does not align with the findings in part (a) as in an overfitting model one would expect a DV to be highly dependent on the variable.

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.
    The total count is the most on days when the "feels like" temperature is pleasant and decreases both when the temperature is too cold or too hot. Additionally, in general the count is low when there is light or heavy rain or snow.
    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?
        More number of casual users ride on days with a pleasant weather around 20 degree Celsius. Registered users show a similar trend but more

number of registered users ride even on slightly colder days or warmer than the casual users.

ii) How is *bad weather* associated with casual usage? Is that different from how it is associated with registered usage?
There are extremely low number of casual riders during bad weather. Registered users' numbers are low too but these are greater than the casual usage numbers.

iii) Do your answers in (i) and (ii) make logical sense? Why or why not?
Yes, it does make sense. The registered users may have ease of access to their rides or they might have paid a subscription fee, leaving them more encouraged to use the bikes even on cold days or during bad weather whereas the casual users have no motivation to do so.

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?
The relationship looks almost like a bell curve. This is similar to lab 2 when we plotted weekly sales against temperature. Since both shopping and riding a bike are outdoor activities, people avoid doing them on cold or hot days, giving us a similar graph in both cases.

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?
0.521. It means that the model with its given variables is slightly good at predicting the total count. It cannot precisely predict the total count for some combinations of variables values. This is still better than the previous model which was overfitting.

b) State precisely how BADWEATHER is associated with the predicted COUNT.
Given there is bad weather on a day, the total count of rides will be on an average 1955 less than for a day with not a bad weather, given all the other 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%?
2520.497
~2520

d) Do you have any concerns about this model or your predicted COUNT in **Q3-c**? Why or why not?
The temperature and feels like temperature are similar to each other which can lead to multicollinearity which can hamper with the model's results.

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() look fine to me!
There is no curved pattern in the diagnostics, showing that the model is adequate in capturing the nonlinear relationship. The normality assumption holds true as residuals line up well on the dotted line of the plot. The points are almost equally spread out and the red line is almost flat, hence the heteroskedasticity assumption is taken of. Lastly, there are no points outside the Cook's distance lines meaning there are no outliers influencing the regression coefficients.
Further, the residuals are mostly centered on zero with no obvious trend, meaning that one error is not affecting the other. The model has multicollinearity in the case of TEMP and ATEMP, which is evident by the high VIF values. This can be mitigated by removing one of these variables from the model.

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?
    The coefficient shows that bad weather affects the total count more in this model than that in Q4. This is because bad weather is the only independent variable here, whereas Q4 considers all the variables that can affect total count, thus reducing the contribution by bad weather.
    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?
    This knowledge can be added by adding an interaction term containing WEEKDAY and BADWEATHER. This is done when the effect of an independent variable on a DV changes, depending on the values of another independent variable. The new regression model will show us how total count is affected by bad weather depending on whether it is a weekday or not.
    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?
    This is a worse model as adjusted r-squared has only slightly increased while the r-squared has gone down and the interaction term is not statistically significant; p value is large.
    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
        The total count is on an average 201 less during bad weather on a weekday as compared to on a weekend.
        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]*

BADWEATHER	WEEKDAY	COUNT(rounded off)
Yes	Yes	1799
Yes	No	1815
No	Yes	4638
No	No	4452

According to the model, the count is the highest on a weekday with no bad weather and the lowest on a weekday with bad weather.

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*.

   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)`**

Predictive analytics -

fitOrg –

.metric	.estimator	.estimate
<chr>	<chr>	<dbl>
rmse	standard	1361.129
mae	standard	1154.461

fitNew –

.metric	.estimator	.estimate
<chr>	<chr>	<dbl>
rmse	standard	1317.429
mae	standard	1136.023

The second model is better for predictive analysis because its RMSE and MAE are lower giving us a more efficient predictive model. Additionally, the r-squared value of the second model is larger suggesting that it is a better model.

Explanatory analytics –

```
Res.Df       RSS Df Sum of Sq      F    Pr(>F)
1    569 1.021e+09
2    568 9.700e+08  1  51006185 29.867 6.934e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Yes, the conclusion remains the same as the p-value here is very small suggesting that the full model is better at predicting COUNT then the reduced model.

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.
No the performance has gone down probably because being the trend has changed largely from 2011 to 2012. The model's predictions are based on the training data of the past which might show results very different from the actual results of the future.
Taking the 1.5 split also does not improve the model. The performance has decreased suggesting that either the trends in the total count is changing rapidly or the independent variables considered are not sufficient to predict the values.

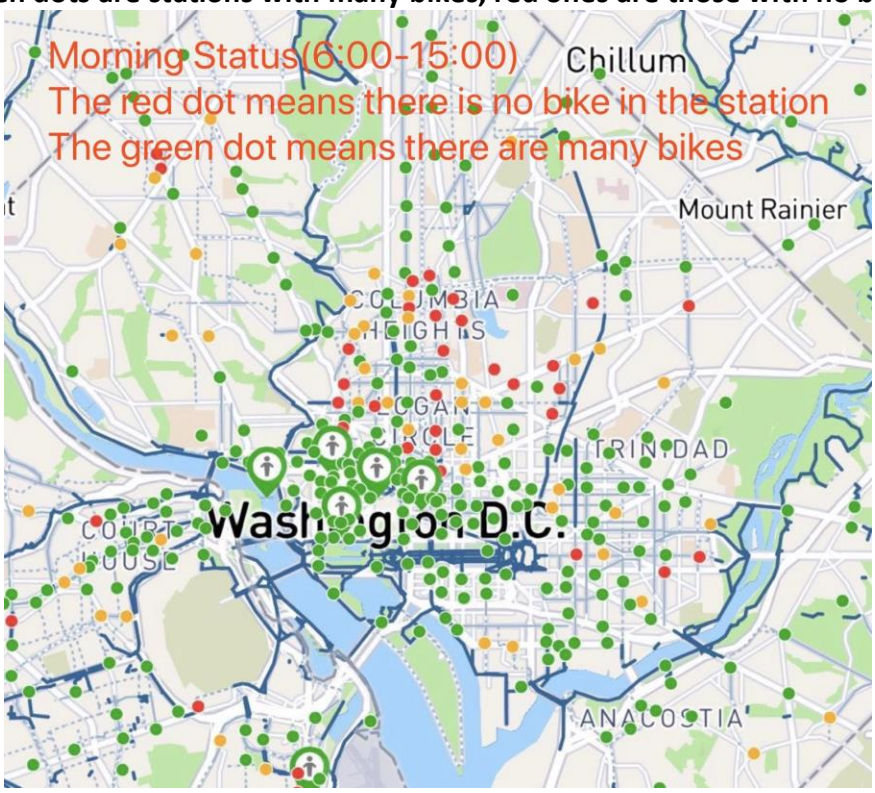
9) **Data-*informed* decision making:** Based on your quick analysis of the Capital Bikeshare data, what are some actions you would take if you were managing Capital Bikeshare's pricing and promotions? How do you think you would use your predictions?

I would vary the cost of cycles on a monthly basis depending on whether that month has more cold or hot days than pleasant days and how much of bad weather is predicted in the month. Further I would probably give special deals to casual bikers during the weeks that have a bad weather forecast as their count is the lowest during these conditions and during the weekend. I will use the predictions to give special prices on windy days as this has a negative effect on the total count.

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.

The company can collect route information with GPS and time on the routes leading to the dock stations which have the rebalancing problem. Using this data, it can use predictive analysis to apply surge charges on the bikes parked on those stations which often go towards stations with the rebalancing problem, at those times after predicting the trend for a particular day. Further, using this route data and the data collected from users, it can reward customers with free rides or discounts who fill up the empty stations more often during surge hours or who are going in the opposite direction to these routes. Again, it can use predictive analysis to predict the stations that are going to be empty soon and start the reward system even before the surge hours hit.

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: