CS559-B: Machine Learning Name: Yanyan Jiang CWID: 10430965

Project Final Report

I. Introduction

Over 400 cities around the world have implemented or are planning to implement a bicycle share program. Bicycling, in general, has numerous benefits. Some of those benefits are: decreasing CO2 emissions; reducing various diseases such as diabetes and obesity; reducing traffic congestion and noise pollution by providing alternatives to auto commuting; and increasing public transit use.

we attempt to examine the effect of weather, date time, and temperature on Bike Share in two years. Using real-time bicycle ridership data provided by Bike Share. Finally, a multilevel model is employed to examine the number of bike share rentals throughout the different day and different weather. The objective of the study is to provide a planning that can be used by transportation planners and city officials to predict the total trip rates of various temporal scales generated at or attracted to assess the effect of potential bicycle infrastructure development on bicycle share ridership.

II. Background

The first public bicycle share system was introduced in Amsterdam, Netherlands, in the 1960s. The “White Bikes” were distributed around the city for the public’s use. The program was aborted as the bicycles were repeatedly stolen and vandalized. In 1996, Portsmouth University in England introduced an IT based public bicycle share system. The program’s users accessed the bicycles using electronic swipe cards. This new generation of bicycle shared systems allowed the operator to identify the customer and provided him or her with the ability to track its use. This addition to the bicycle share concept led to a significant reduction of bicycle thefts as well as vandalism. Today, we witness the evolution of “Multi-modal Systems” that provide: GPS tracking of bicycle use; real time ridership data; moveable bike stations; and system integration with public transit modes

III. Your approach

To predict the total count of bikes rented. Predict the demand for shared bicycles in the future to determine our volume to avoid unnecessary expenses. It even can help transportation planners and city officials with urban planning.

I try to find a multilevel model to examine the rented count throughout the different day and different weather.

IV. Experimental design

Data from Kaggle: https://www.kaggle.com/c/bike-sharing-demand

It’s hourly bike share rental data spanning two years. The training set is comprised of the first 19 days of each month, while the test set is the 20th to the end of the month.

Data Fields

datetime : hourly date + timestamp
season :   1 = spring, 2 = summer, 3 = fall, 4 = winter
holiday : whether the day is considered a holiday
workingday : whether the day is neither a weekend nor holiday
weather : 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 : temperature in Celsius
atemp : "feels like" temperature in Celsius
humidity : relative humidity
windspeed : wind speed
casual : number of non-registered user rentals initiated
registered : number of registered user rentals initiated
count : number of total rentals

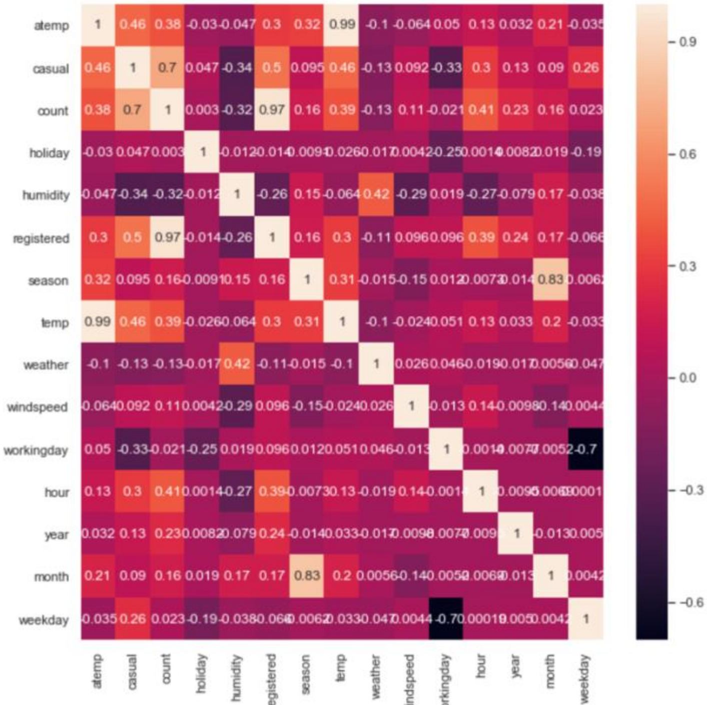
Data Preprocessing

Once we get hang of the data and columns, next step we generally is to find out whether we have any missing values in our data. Luckily I don‘t have any missing value in the dataset.

At first look, "count" variable contains lot of outlier data points, so remove outliers from more than three standard deviations. It is desirable to have Normal distribution as most of the machine learning techniques require dependent variable to be Normal. So take log transformation on "count" variable to approach it.

Correlation Analysis

For understand how a dependent variable is influenced by features, we want to find a correlation matrix between them. Plot a correlation plot between "count” and other fields.



Inferences from the above heatmap:

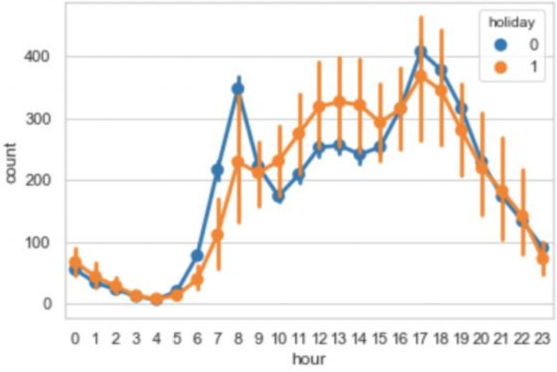
temp and atemp are highly related as expected. So I only pick one as the feature.

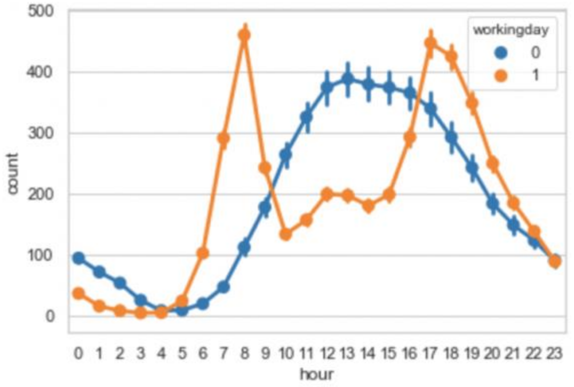
humidity is inversely related to count as expected as the weather is humid people will not like to travel on a bike. The same temp(or atemp) , holiday weather and count are highly inversely related.

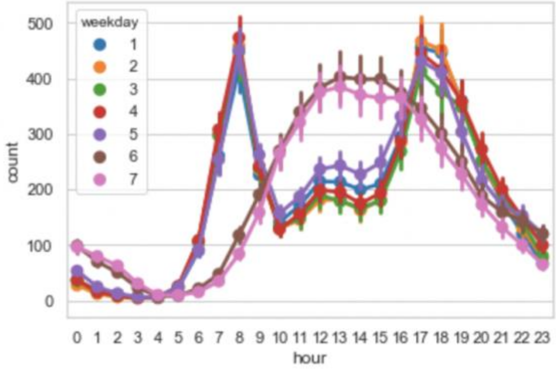
registered/casual and count are highly related which indicates that most of the bikes that are rented are registered

Analyze the impact between field and count one by one to determine the choice of feature. The complete process is in the code.

Working day or holiday influence:







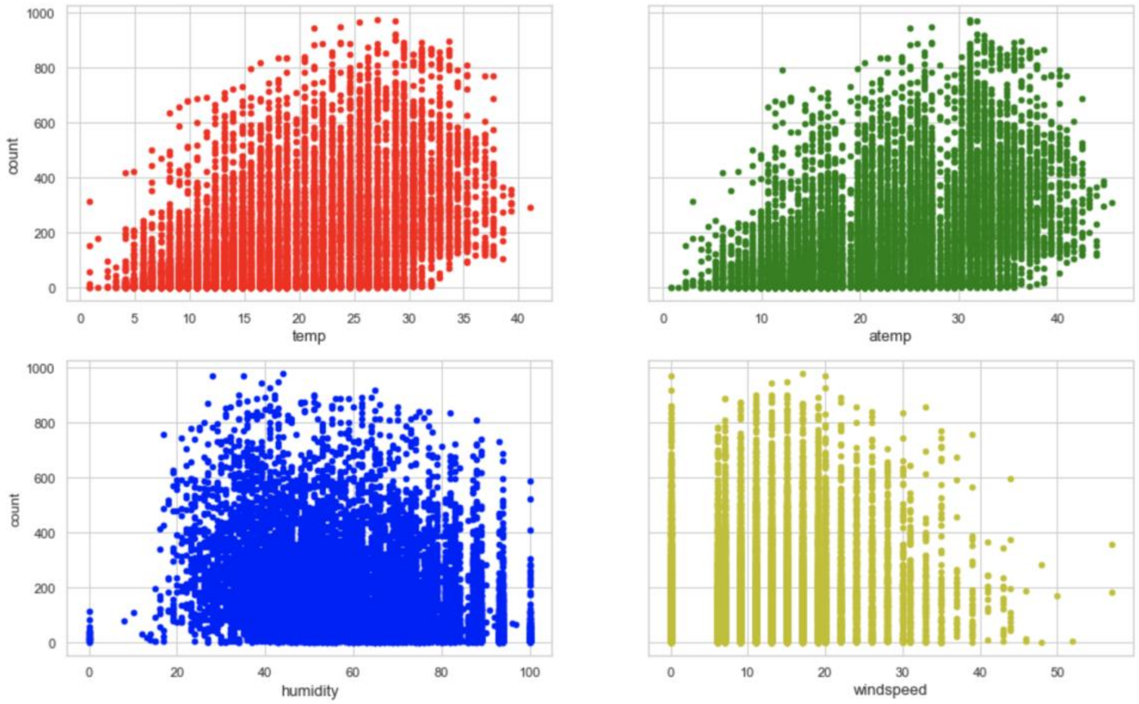
The trend of this three plot is very similar. We can divide the data into two situations: people go to work and not go to work for specific analysis.

- The day people go to work: (Like working day and “Monday” to “Friday”) more people tend to rent bicycle around 7AM-8AM and 5PM-6PM. As we mentioned earlier this can be attributed to regular school and office commuters. There will also be a small peak at noon, and I guess may be someone who is out for lunch.

- The day people not go to work: (Like holiday and "Saturday" and "Sunday".)

More people tend to rent bicycle between 10AM and 4PM. People may be cycling out on vacation or exercise.

weather factors influence:



Although the influence of the weather factors on the number of leases is relatively scattered, it can be clearly seen that the temperature and wind speed are positively correlated with the count, and the humidity is negatively correlated with it.

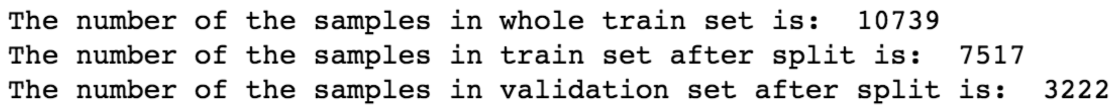
Feature Engineering

the columns like "season", and "weather" should be of "categorical" data type. But the current data type is "int" for those columns. Since the CART decision tree uses a two-category, multi-category data is converted into multiple dichotomous categories using one-hot.

After that, remove redundant columns based on our selected feature.

Cross-validation

the original train is divided into two sets of train set and validation set by a ratio of 7:3.



```
The number of the samples in whole train set is:  10739
The number of the samples in train set after split is:  7517
The number of the samples in validation set after split is:  3222
```

Evaluate Metric

The mean squared error is a risk metric corresponding to the expected value of the squared (quadratic) error or loss.

If $\widehat{y}_i$ is the predicted value of the i-th sample, and y_i is the corresponding true value, then the mean squared error (MSE) estimated over n_{samples} is defined as

$$\text{MSE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} (y_i - \hat{y}_i)^2.$$

"mean squared error", often referred to as mean squared prediction error or "out-of-sample mean squared error", can refer to the mean value of the squared deviations of the predictions from the true values.

V. Result analyze and Conclusion

Compare the prediction results of each model on the validation data, select the optimal model.

	Linear Regression	Ridge Regression	Lasso Regression	Decision Tree Regression	Random Forest Regression	XgBoost Regression	Multi-layer Perceptron Regression
MSE	0.385	0.385	0.384	0.189	0.124	0.245	0.144

Based on observation and analysis, Random Forest Regression model achieved the Minimum Mean square error.

Because the prediction of count is after log transformation to a approach Normal distribution. (Transform count from around 0~1000 to -1~7). So it should do exponent movement to get the final predicted result.

VI. Future work

- To find some processing method for the variables of the training set instead of delete it.
- To find other methods to convert some features like wind speed and humidity variables.
- Use Ensemble Learning-model fusion skills instead of a single model.

VII. Reference

https://scikit-learn.org/stable/modules/model_evaluation.html#mean-squared-error

https://www.imooc.com/article/69542

https://www.jianshu.com/p/2ce2b21f67bf

https://www.kaggle.com/viveksrinivasan/eda-ensemble-model-top-10-percentile/notebook

https://zhuanlan.zhihu.com/p/42428514

https://www.kaggle.com/rajmehra03/bike-sharing-demand-rmsle-0-3194

https://www.researchgate.net/profile/Khandker_Nurul_Habib2/publication/286379583_Effects_of_Built_Environment_and_Weather_on_Bike_Sharing_Demand_A_Station_Level_Analysis_of_Commercial_Bike_Sharing_in_Toronto/links/56684b7c08ae7dc22ad1c992/Effects-of-Built-Environment-and-Weather-on-Bike-Sharing-Demand-A-Station-Level-Analysis-of-Commercial-Bike-Sharing-in-Toronto.pdf

https://scholarcommons.usf.edu/cgi/viewcontent.cgi?referer=http://scholar.google.com/&httpsredir=1&article=1196&context=jpt