Prediction Model for Bike Rental System

Supervised Machine Learning Capstone

Jimin Shin

August, 2019

Introduction

- **Bike sharing systems** are new generation of traditional bike rentals where whole process from membership, rental and return back has become automatic.
- Today, there exists great interest in these systems due to their important role in traffic, environmental and health issues.
- Predicting the amount of bike rental is directly related to **maintenance cost**

Research Question

- Bike-sharing rental process is highly correlated to the environmental and seasonal settings.
    - weather conditions, precipitation, day of week, season, and hour of the day can affect the rental behaviors.
- Q1: How many extra bikes do we need to prepare?
    - Prediction of the number of **unexpected customers** (casual customers) by environmental or weather condition should be required.
- Q2: Which daily condition create situations to be needed extra bikes?

Data Source

- Rental Log: Capital Bikeshare system, Washington D.C., USA
  - Bike sharing counts aggregated on hourly basis. Records: 17379 hours
- Weather information: http://www.freemeteo.com
- Duration: two-year historical log from 2011 to 2012

Description of Variables

- *Continuous variables*
  - **temp** : Normalized temperature in Celsius. T
  - **atemp**: Normalized feeling temperature in Celsius.
  - **hum**: Normalized humidity.
  - **windspeed**: Normalized wind speed.
  - **casual**: count of casual users
  - **registered**: count of registered users
  - **cnt**: count of total rental bikes including both casual and registered

Description of Variables

- *Categorical and time variables*
  - **dteday** : date
  - **yr** : year (0: 2011, 1:2012)
  - **mnth** : month (1 to 12)
  - **hr** : hour (0 to 23)
  - **weekday** : day of the week (0: Sunday ~ 6: Saturday)
  - **season** : season (1:winter, 2:spring, 3:summer, 4:fall)
  - **holiday** : weather day is holiday or not
  - **workingday** : if day is neither weekend nor holiday is 1, otherwise is 0.
  - **weathersit**: (1: clear, 2: cloudy, 3: light snow or rain, 4: heavy snow or rain)

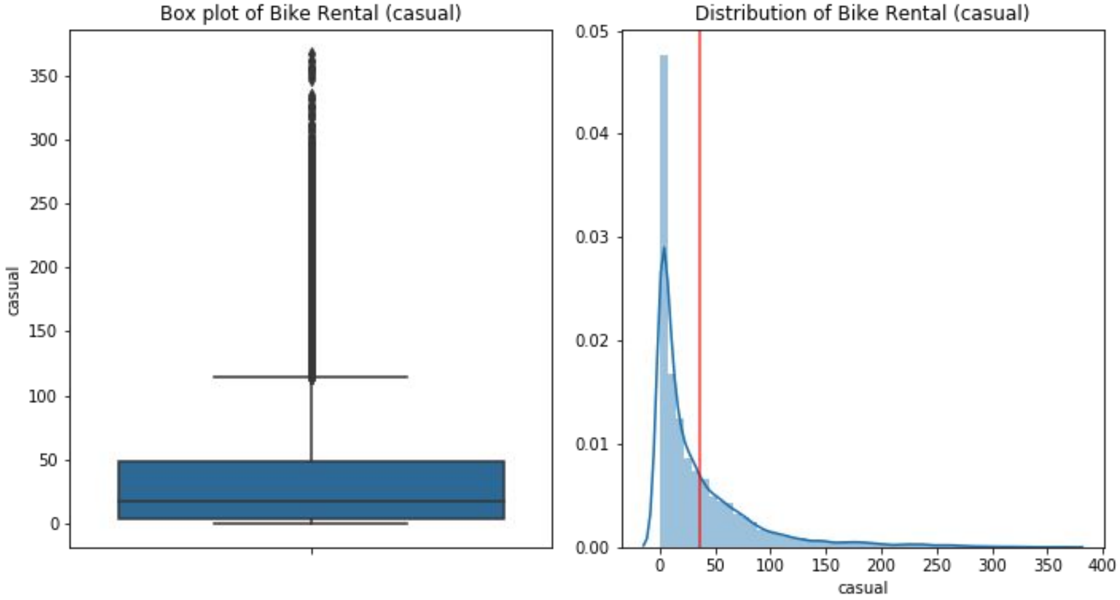
Data Cleaning

- There is no missing value
- Add **part_day** variables using **hr** variable
  - 0 am ~ 6 am: ‘night’
  - 6 am ~ 12 pm: ‘morning’
  - 12pm ~ 6 pm: ‘afternoon’
  - 6pm ~ 12am: ‘evening’

	Total	Percent
cnt	0	0.0
weekday	0	0.0
dteday	0	0.0
season	0	0.0
yr	0	0.0
mnth	0	0.0
hr	0	0.0
holiday	0	0.0
workingday	0	0.0
registered	0	0.0
weathersit	0	0.0
temp	0	0.0
atemp	0	0.0
hum	0	0.0
windspeed	0	0.0
casual	0	0.0
instant	0	0.0

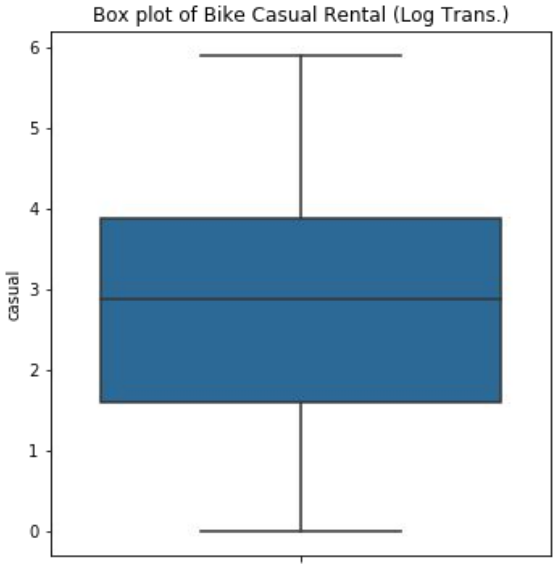
Exploring Target

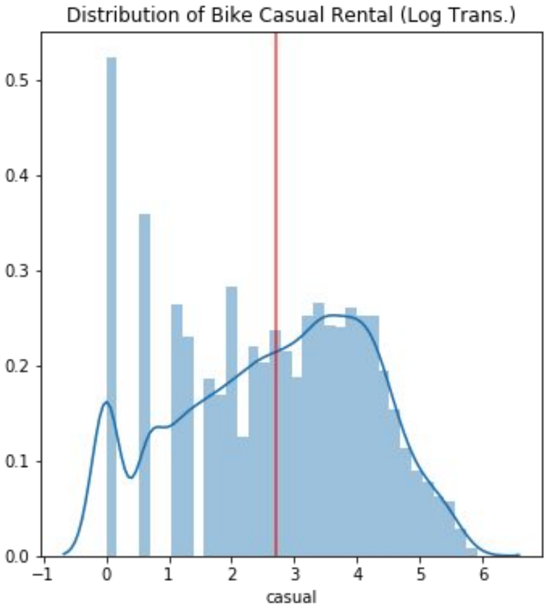
- Target: **casual,** count of casual users



Exploring Target

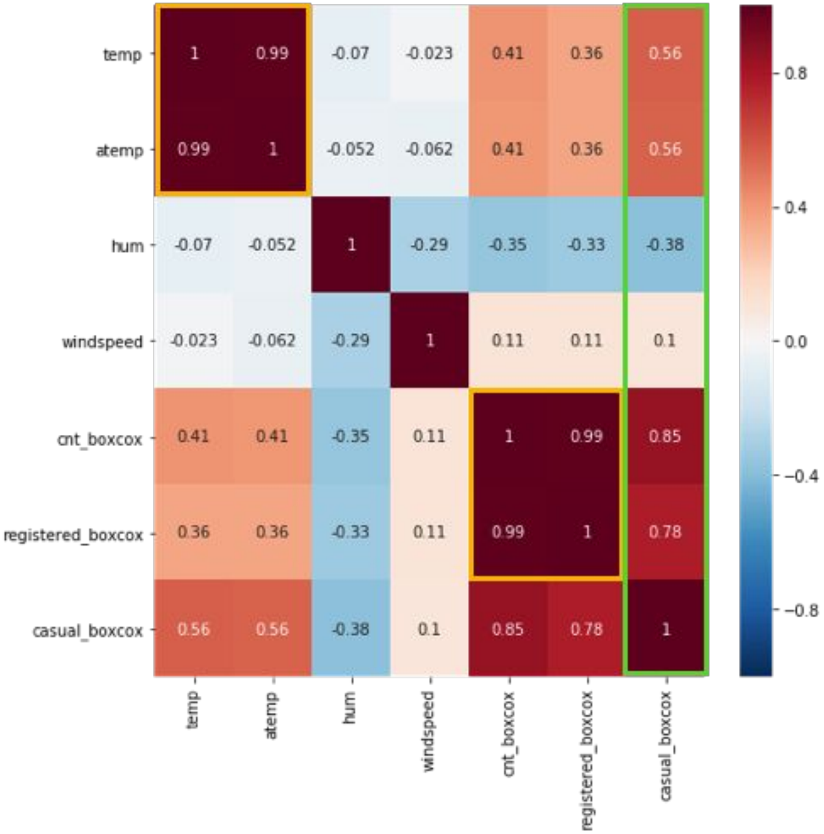
- Handling Non-normality with **Box-Cox Transformation**
- **Use transformed values for the prediction model**





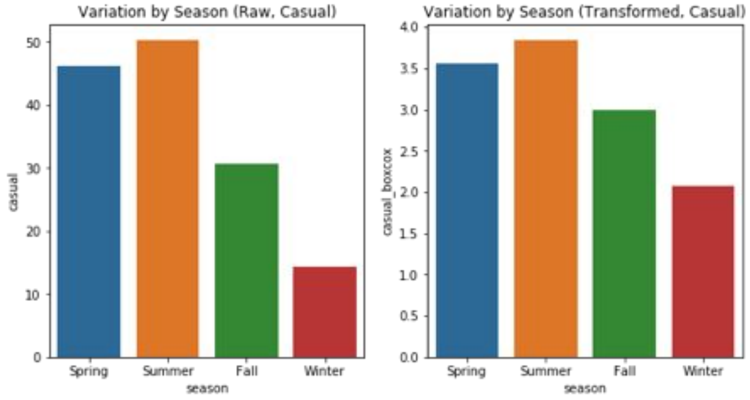
Exploring Continuous Variables

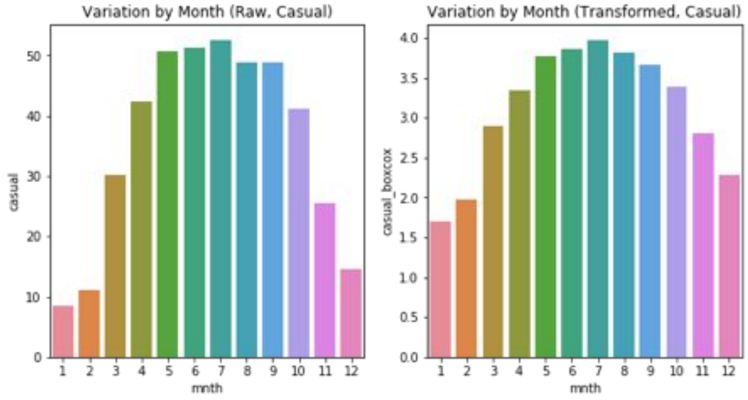
- Temperature, Feeling Temperature, and Humidity are correlated with casual rental (> 0.30)
- Windspeed has relatively small correlation with casual rental (0.1)
- Temperature and Feeling Temperature are highly correlated (0.99)
    - exclude Feeling Temperature



Exploring Categorical Variables

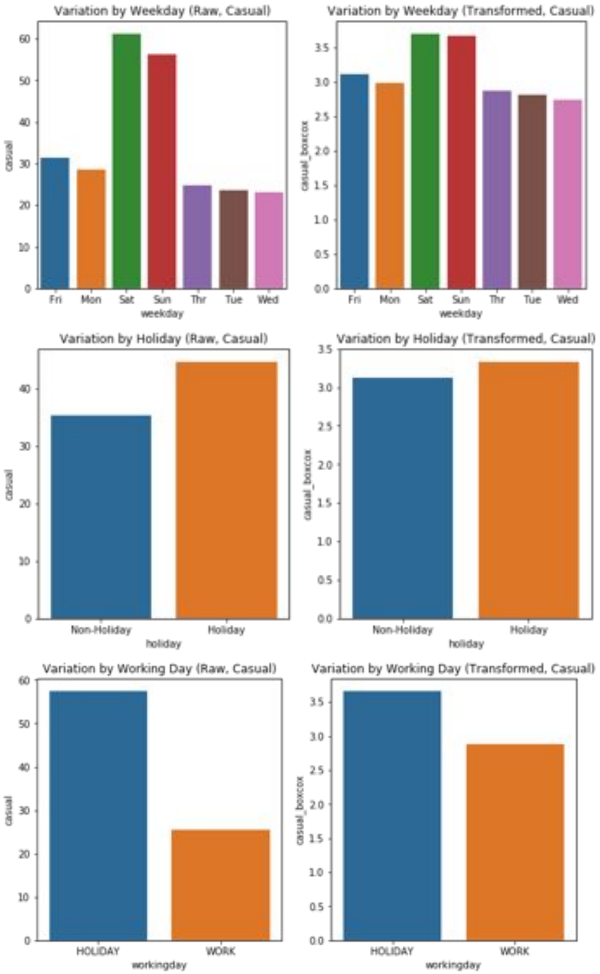
- **season** variable is the similar concept with **mnth** variable
- exclude **mnth** variable out of the feature set.





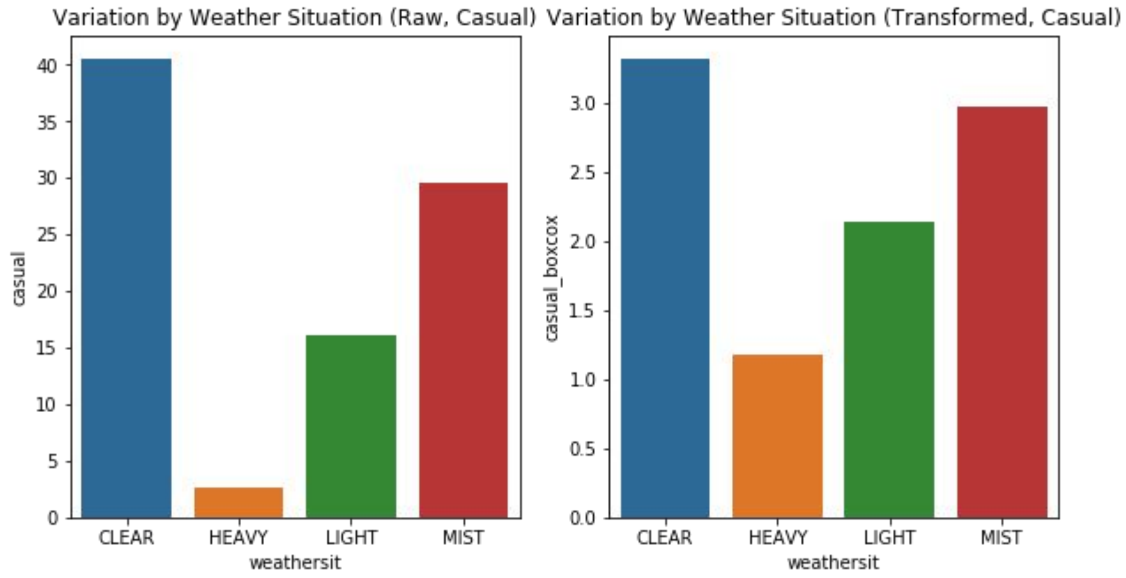
Exploring Categorical Variables

- **Holiday** and **Working day** have a very similar concept and trend.
- In the weekday plot, casual bike rental increased in the weekend
- exclude **weekday** and **holiday** variable out of the feature set



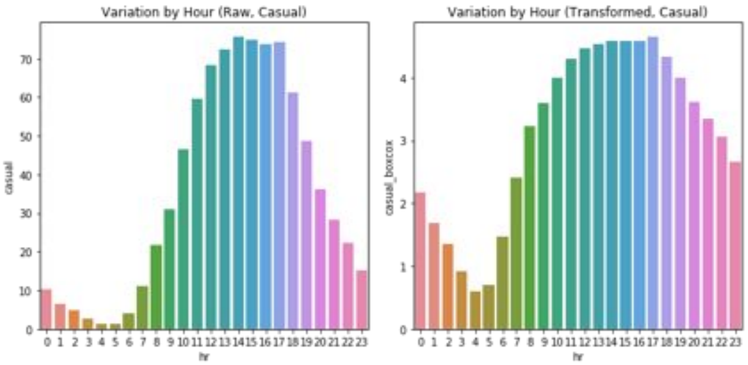
Exploring Categorical Variables

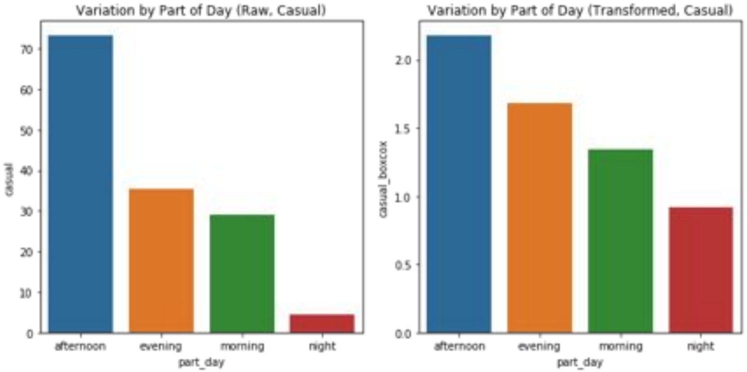
- casual bike rental count is changed by the weather situation



Exploring Categorical Variables

- **Hour** variable can be explained by **Parts of Day**





Chosen Features

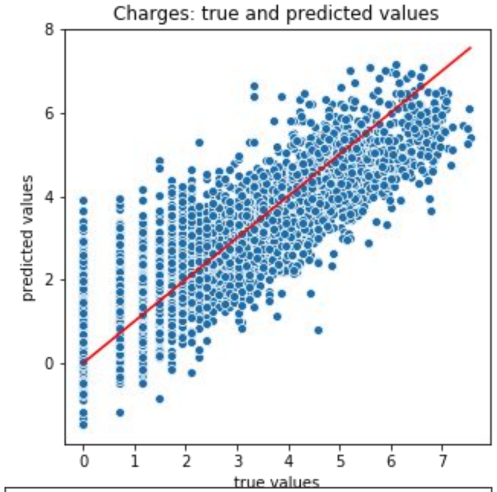
- Continuous: temp, hum, windspeed
- Categorical: season, weathersit, workingday, part_day

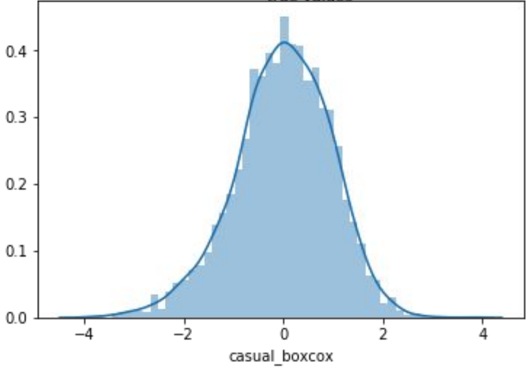
Q1: How many extra bikes do we need to prepare?

Regression task

Ordinary Least Square Regression

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           casual_boxcox   R-squared:                       0.726
Model:                             OLS   Adj. R-squared:                  0.726
Method:                  Least Squares   F-statistic:                     3349.
Date:                 Tue, 13 Aug 2019   Prob (F-statistic):               0.00
Time:                         22:35:15   Log-Likelihood:                -19179.
No. Observations:                13903   AIC:                         3.838e+04
Df Residuals:                    13891   BIC:                         3.847e+04
Df Model:                           11
Covariance Type:             nonrobust
======================================================================================
                         coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------
const                  0.7162      0.036     20.129      0.000       0.646       0.786
temp                   4.5434      0.045    100.588      0.000       4.455       4.632
hum                   -1.4869      0.055    -27.218      0.000      -1.594      -1.380
windspeed             -0.3645      0.071     -5.102      0.000      -0.504      -0.224
season_Fall            0.5005      0.021     24.127      0.000       0.460       0.541
season_Spring          0.4939      0.020     24.718      0.000       0.455       0.533
weathersit_CLEAR       0.5651      0.034     16.649      0.000       0.499       0.632
weathersit_MIST        0.5939      0.034     17.454      0.000       0.527       0.661
workingday_HOLIDAY     0.7907      0.020     39.024      0.000       0.751       0.830
workingday_WORK       -0.0745      0.019     -3.836      0.000      -0.113      -0.036
part_day_afternoon     1.1885      0.017     71.300      0.000       1.156       1.221
part_day_evening       0.5099      0.017     30.848      0.000       0.477       0.542
part_day_morning       0.3941      0.018     22.517      0.000       0.360       0.428
part_day_night        -1.3762      0.019    -73.337      0.000      -1.413      -1.339
==============================================================================
Omnibus:                      212.020   Durbin-Watson:                   2.004
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              221.497
Skew:                          -0.308   Prob(JB):                     7.99e-49
Kurtosis:                       3.048   Cond. No.                     1.98e+16
==============================================================================
```





Ordinary Least Square Regression

- Our model is not complex, it doesn't have overfit problem with small generalization gap.
- Casual bike rental is estimated by the factors below

	Coeffecient
temp	4.543439
hum	-1.486907
windspeed	-0.364458
season_Fall	0.500453
season_Spring	0.493905
weathersit_CLEAR	0.565076
weathersit_MIST	0.593935
workingday_HOLIDAY	0.432591
workingday_WORK	-0.432591
part_day_afternoon	1.009447
part_day_evening	0.330806
part_day_morning	0.215008
part_day_night	-1.555261

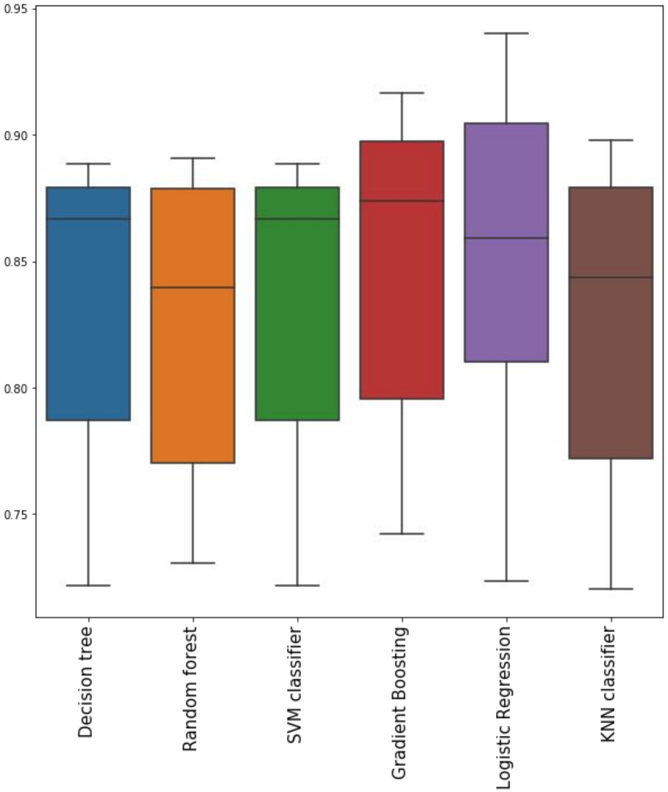
Q2: Which daily condition create situations to be needed extra bikes?

Classification task

Assumption: 50 bikes are always prepared for casual rental

When do we need to prepare extra bikes if casual bike rental is over 50?

Which classifier performs best?

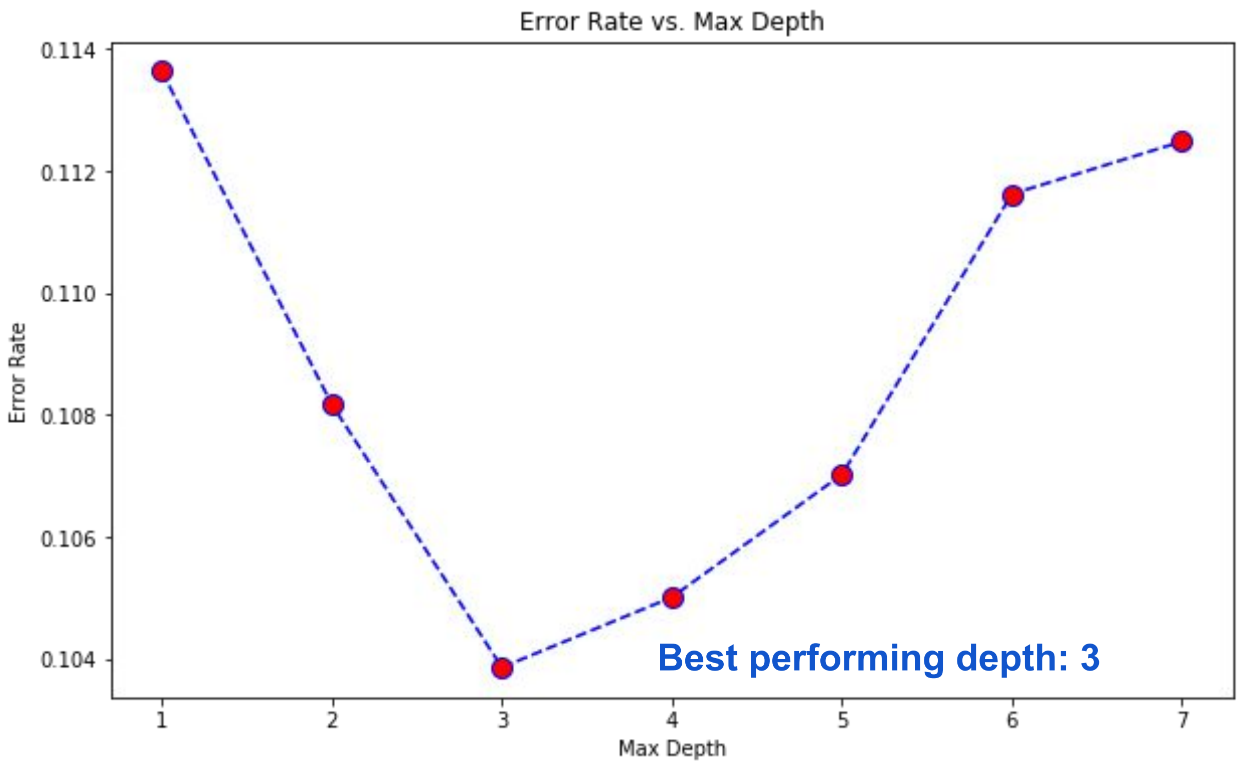


	Decision tree	Random forest	SVM classifier	Gradient Boosting	Logistic Regression	KNN classifier
count	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000
mean	0.830833	0.823177	0.830833	0.844870	0.846365	0.824156
std	0.063483	0.061881	0.063483	0.067566	0.075188	0.067813
min	0.721519	0.730725	0.721519	0.742232	0.723245	0.720368
25%	0.787255	0.770138	0.787255	0.795311	0.810127	0.771864
50%	0.866475	0.839711	0.866475	0.873921	0.859240	0.843457
75%	0.879282	0.878722	0.879282	0.897194	0.904647	0.879009
max	0.888377	0.890679	0.888377	0.916571	0.940161	0.897642

Gradient Boosting model has the best performance regarding average accuracy with low variation

Logistic Regression model also has the best value of average accuracy, however, it has a high level of variation.

Gradient Boosting Classifier Detail - depth

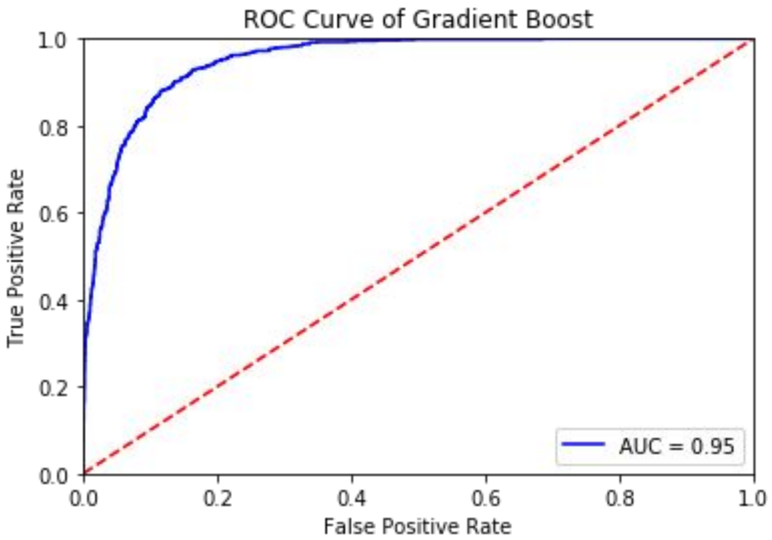


Gradient Boosting Classifier Detail - confusion matrix

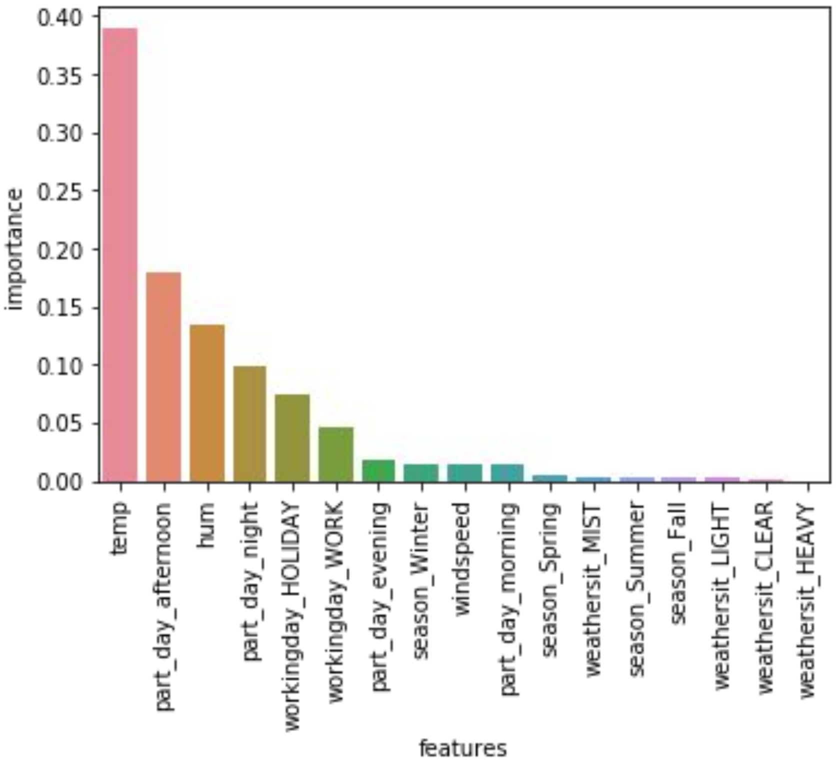
2482 (true negative)	**170** *(false positive)*
191 *(false negative)*	**633** (true positive)

Sensitivity = 0.77

Specificity = 0.94



Gradient Boosting Classifier Detail - feature importance



Summary

- Predicting the amount of bike rental is directly related to **maintenance cost**
- **Temperature** is the important factor that affects to casual bike rental.
- According to regression and classification modeling, in the holiday afternoon with high temperature and low humidity, casual bike rental is increased, therefore the bike rental system might need to prepare extra bikes. (regression)

Discussion and Future Work

- Target value, casual bike rental, is not normally distributed.
    - obtain more data sets as much as possible
- We can think about adding economic factors on the prediction models. This can affect people’s emotional or economical situation.