Bike Sharing Demand Prediction
In- Depth Analysis Report

Name: Xingkai Wu

Python Code(Github) Link:
https://github.com/xwu0223/Bike-Sharing-Prediction-Project/blob/master/Bike_Sharing_Further_Analysis.ipynb

Independent Variables：

dteday: date and hour in "yyyy-mm-dd" format
season: Four categories-> 1 = spring, 2 = summer, 3 = fall, 4 = winter
holiday: whether the day is a holiday or not (1/0)
workingday: whether the day is neither a weekend nor holiday (1/0)
weathersit: Four Categories of weather
1-> Clear, Few clouds, Partly cloudy, Partly cloudy
2-> Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
3-> Light Snow and Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
4-> Heavy Rain + Ice Pellets + Thunderstorm + Mist, Snow + Fog
temp: hourly temperature in Celsius
hum: relative humidity
windspeed: wind speed

Dependent Variables：

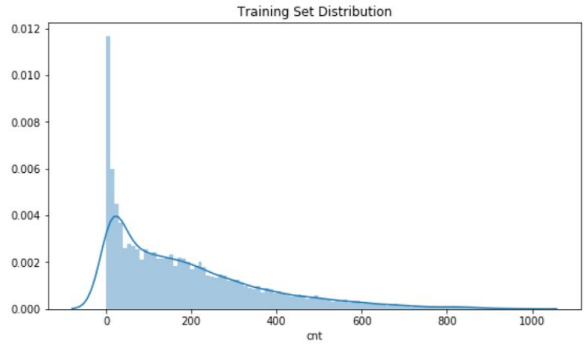
registered: number of registered user
casual: number of non-registered user
count: number of total rentals (registered + casual)

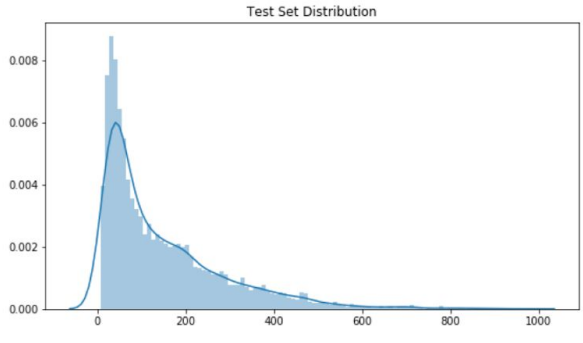
Linear Regression Model

RMSLE (Root Mean Squared Logarithmic Error) is to evaluate the performance in predicting.

Visualizing Distribution Of Train And Test

By splitting the data into training data and testing data, the following plots illustrated the distribution of the training(left) and testing(right) data.
The RMSLE value for linear regression in validation is 0.991, this is not anywhere close to ideal.





Regularization

Regularization is extremely useful in any of the following cases. We do not face all the cases mentioned below but overfitting and multicollinearity may pose some issues for us.

Overfitting.

Large number of variables.

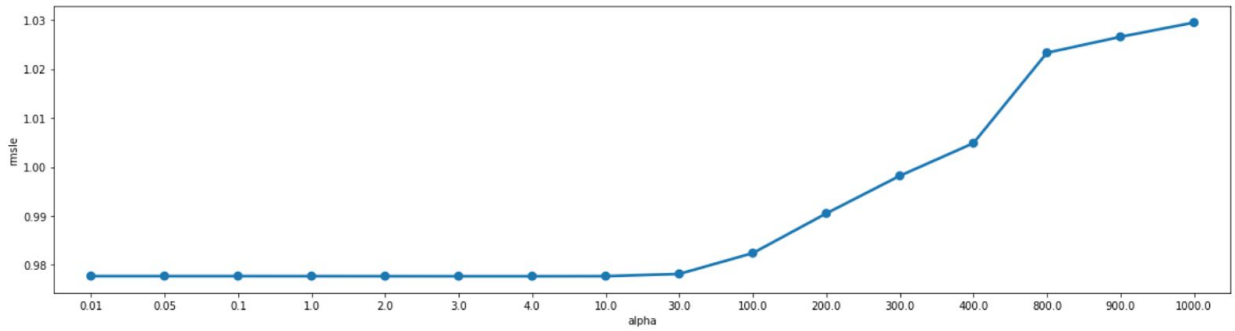
Low ratio of number of observations to the number of variables.

Multicollinearity.

L2 regularization (Ridge Regression) is extremely helpful for third case where we have more number of attributes than observation. But we are doing fine with that as of now since we have just 12 attributes and dataset has 10886 records. *Ridge Regression* is also quite useful when there is high *collinearity* between predictor variable. It may happen in our dataset since we have highly correlated variables such as temp-atemp and month-season.

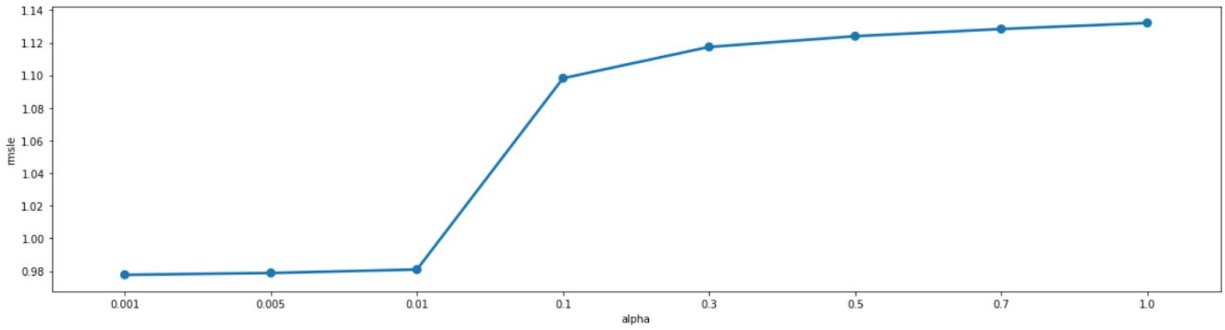
Both Ridge and Lasso give RMSLE value of 0.99

Regularization Model - Ridge :



Having really large number of variable may again result in overfitting. This is because when we have more variables model becomes more complex and sometimes lowers its

predicting and generalization power. L1 regularization (Lasso Regression) comes handy in these situations by reducing the coefficients to zero thereby producing simpler. Models.



Ensemble Models - Random Forest

Ensemble models are nothing but art of combining diverse set of individual weak learners(models) together to improve stability and predictive capacity of the model. Ensemble Models improves the performance of the model by

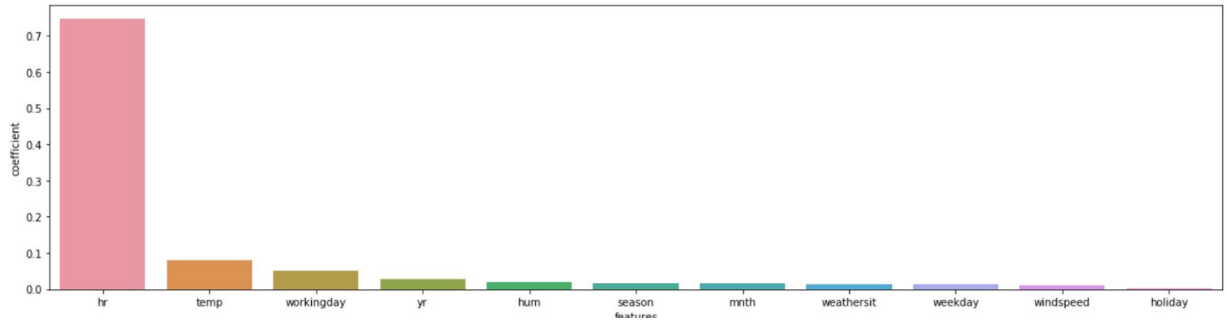
Averaging out biases.

Reducing the variance.

Avoiding overfitting.

The RMSLE value for Random Forest is 0.299, this is much better than the linear regression model.

Feature Importance By Random Forest



From the bar plot above, it is obvious that time of the day is the most important feature in the data set that influence the number of user count, temperature and working- day have relatively smaller influences than time of the day.

Ensemble Model - Gradient Boost

The RMSLE value for Random Forest is 0.273, this value has reached the top 10% percentile of the Kaggle scoreboard.

By using the Gradient Boost model to make the prediction with testing data, the comparison between training data and predicted data. The distribution of the predicted result is almost identical to the training data.