Bike Share Project

Author: Prachi Kharat

Email : p_kharat@u.pacific.edu

1. Objective:

The aim of this project is to produce a model which predicts bike usage, based on other features of the data. Data provides hourly and daily rental data spanning two years. Our objective is to do predication of bike rental count hourly or daily based on the environmental and seasonal settings. This is a regression problem.

2. Dataset:

Bike-sharing rental process is highly correlated to the environmental and seasonal settings. For instance, weather conditions, precipitation, day of week, season, hour of the day, etc. can affect the rental behavior. The core data set is related to the two-year historical log corresponding to years 2011 and 2012 from Capital Bikeshare system, Washington D.C., USA. There are two datasets named hourly.csv and daily.csv.

- hourly.csv : bike sharing counts aggregated on hourly basis. Records: 17379 hours

- daily.csv - bike sharing counts aggregated on daily basis. Records: 731 days

Both hourly.csv and daily.csv have the following fields, except hr which is not available in daily.csv

3. Explanatory Data Analysis:

The Objective of EDA was to find relation of target variable count with other feature of data.

- People tend to rent bike during summer season since it is really conducive to ride bike at that season. Therefore, June, July and August has got relatively higher demand for bicycle.
- On weekdays more people tend to rent bicycle around 7AM-8AM and 5PM-6PM. This can be attributed to regular school and office commuters.
- Above pattern is not observed on "Saturday" and "Sunday". More people tend to rent bicycle between 10AM and 4PM.
- The peak user count around 7AM-8AM and 5PM-6PM is purely contributed by registered user.
- Demand is higher on work days rather than holidays.

4. Data Preprocessing:

The target variable was not normally distributed. "count" variable is skewed towards right. It is desirable to have Normal distribution as most of the machine learning techniques require dependent variable to be Normal. We normalised the data to bring it in range. This doesn't correct the right skewness but brings data in range.

Outliers: Target variable contained lot of outliers. Box plot use the IQR method for finding display data and outliers. Using the IQR method, we found that data contains 2.9% outliers. We removed these using standard deviation.

Splitting of data: we splitted the data in train and test with 0.25 test proportion.

Correlation analysis: To understand how a dependent variable is influenced by features (numerical) I found a correlation matrix between them. temp and humidity features has got positive and negative correlation with count respectively. Although, the correlation between them are not very prominent still the count variable has got little dependency on "temp" and "humidity"."atemp" is variable is not taken into since "atemp" and "temp" has got strong correlation with each other. "Casual" and "Registered" are also not taken into account since they are leakage variables in nature and need to dropped during model building.

5. Model :

I used random forest regressor() to create the model. While it is not necessary, I normalised the test and train data. After fitting the model, it was used on test set to evaluate and predict.

6. Result:

Metrics for evaluation : R^2 score, Mean Squared Error

R^2 score of the model is 0.93. This means that the model explains 93% of variance in the observed data which is explained by the model.

MSE is 0.0033.

Ideal MSE is 0. Minimizing MSE is equivalent to maximizing the likelihood of the data under the assumption that target comes under the assumption that target comes from a normal distribution.

Plot of our prediction to actual value: