1 Introduction

1.1 Problem Statement

Tracking the bike rental on the daily basis is a very tedious task. So we need to develop a model that makes this task simple.
The objective of this Case is to predict bike rental count on daily basis on the environmental and seasonal settings based on the previous dataset. We will use machine learning algorithm to predict bike rental considering various given factors.

1.2 Data

Given below is a sample of the data set that we are using to predict the bike rental count.
The details of the data attributes in the dataset is as follow:

instant: Record index
dteday: Date
season: Season (1:springer, 2:summer, 3:fall, 4:winter)
yr: Year (0: 2011, 1:2012)
mnth: Month (1 to 12)
hr: Hour (0 to 23)
holiday: weather day is holiday or not (extracted fromHoliday Schedule)
weekday: Day of the week
workingday: If day is neither weekend nor holiday is 1, otherwise is 0.
weathersit: (extracted fromFreemeteo)
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: Normalized temperature in Celsius. The values are derived via (t-t_min)/(t_max-t_min),
t_min=-8, t_max=+39 (only in hourly scale)
atemp: Normalized feeling temperature in Celsius. The values are derived via(t-t_min)/(t_maxt_min), t_min=-16, t_max=+50 (only in hourly scale)
hum: Normalized humidity. The values are divided to 100 (max)
windspeed: Normalized wind speed. The values are divided to 67 (max)
casual: count of casual users
registered: count of registered users
cnt: count of total rental bikes including both casual and registered
There are total 16 variables : 15=independent variables
1=dependent variable

2 Methodology

2.1 Pre Processing

Data preprocessing is an important step in the data mining process. The phrase "garbage in, garbage out" is particularly applicable to data mining and machine learning projects. Data-gathering methods are often loosely controlled.

Data preprocessing includes cleaning, Instance selection, normalization,

transformation, feature extraction and selection, etc. The product of data preprocessing is the final training set.
Tasks of data pre-processing:
Data cleansing
Data editing
Data reduction
Data wrangling

2.1.1 Outlier Analysis

In statistics, an outlier is a data point that differs significantly from other observations. An outlier may be due to variability in the measurement or it may indicate experimental error; the latter are sometimes excluded from the data set. An outlier can cause serious problems in statistical analyses.
There is no rigid mathematical definition of what constitutes an outlier; determining whether or not an observation is an outlier is ultimately a subjective exercise. There are various methods of outlier detection. Some are graphical such as normal probability plots. Others are model-based. Box plots are a hybrid.

2.1.2 Feature Selection

Feature engineering is the process of using domain knowledge of the data to create features that make machine learning algorithms work. Feature engineering is fundamental to the application of machine learning, and is both difficult and expensive.

2.2 Modeling

2.2.1 Linear Regression Model

The target variable count is continuous, so we have chose linear regression model.
In statistics, linear regression is a linear approach to modeling the

relationship between a scalar response and one or more explanatory variables. The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression.
The simplest form of the regression equation with one dependent and one independent variable is defined by the formula y = c + b*x, where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable.

3 Conclusion

3.1 Model Evaluation

Now that we have applied the linear regression model on the test data, we can check the acuracy of the model by applying error metrics.
Confusion metrics are meaningful for mostly regression studies.
Two dimensional confusion matrix should be applied to measure system performance.
There are 4 different outcomes based on confusion matrix.
True Positive: Ones classified as true and it is really true.
True Negative: Ones classified as false and it is really false.
False Positive: Ones classified as false but it is actually true.
False Negative: Ones classified as true but it is actually false.

The mean absolute percentage error (MAPE), also known as mean absolute percentage deviation (MAPD), is a measure of prediction accuracy of a forecasting method in statistics, for example in trend estimation, also used as a loss function for regression problems in machine learning. It usually expresses accuracy as a percentage.
Multiplying by 100% makes it a percentage error.
For the deployed model we have used MAPE.