Exploratory Data Analysis of Bike Sharing Dataset

Problem Statement: Bike sharing rental system has become a growing business all over the world due to its low-cost service, easy availability and reduction in congestion on roads. This is an automated system which has got IoT devices that generate data related to the performance of bike & environmental condition. The intent is to analyze data through Artificial Intelligence by adopting Machine Learning techniques to predict the demand of the bikes & run the business profitably.

Descriptive Analysis: The dataset is retrieved from UCI Machine Learning Repository & is analyzed in Jupyter notebook with Python 3.4 version. It has got 17389 records with 16 variables out of which 'cnt' is the target variable i.e. count of total rental bikes which is a summation of registered & casual bikes. It is a combination of continuous variables like registered & casual user's count, temperature, humidity, windspeed etc. & discrete variables like season, year, weather situation, working day etc.

Missing/Outliers & Distribution Analysis: There were no missing values & a histogram (Fig1) represented that the target variable count is positively (right) skewed indicating that there are times where the demand of bikes is very less. Thus, outliers are removed to ensure normality. It is prudent to mention that bike demand is similar during all the seasons (winter, summer, spring, fall). However, demand is highly influenced by the "hour" variable as it follows a trend over hours e.g. there is high demand from 7-9 & 17-19 hrs. and low demand from 0-6 & 20-23 hrs. during weekdays whereas the demand is high in the afternoons (11-17 hrs.) during weekends. It is observed that the demand of casual bikes increases during weekend. It is to be noted that the usage of bikes during holidays is negligible confirming that the bike is used mostly for work purpose. Variables temp, atemp, humidity & windspeed are naturally distributed but 'atemp' is highly correlated with temp thereby leading to dropping of variable. Around 65.7% of the times, the weather was clear as per the dataset and hardly it rained heavily. Usage of bikes month wise is found to be normally distributed with high usage of > 200 bikes used per month between May & October. Also, bikes usage has increased from 2011 to 2012 by 45% indicating that there is a scope for improving.

Feature Engineering: A correlation matrix between continuous variables using Pearson method suggest that 'atemp' & 'registered' variables are highly correlated with that of 'temp' & 'cnt'. Also, Spearman technique is adopted to evaluate the correlation between categorical variables as they have ordinal values attributing to dropping of 'month' variable in view of high correlation with 'season'. Creation of dummies for the discrete variables is done which ensures label encoding & one hot encoding of different levels.

Training/Evaluation/Pipeline: The dataset is split into 80% of training & 20% of test & was standardized in view of retaining variance in the data. A data pipeline is created with several Regression models like Linear, Lasso, Ridge, Linear SVR, Kernel SVR, Random Forest Regressor & Gradient Boosting Regressor as the target variable is continuous & has <100K variables. It has been observed that Random Forest Regressor gave the highest accuracy (91%) with lowest Mean squared error (2245) which means the data is correctly fit. Temperature, humidity, working day & hour variables have high feature importance.