CASE OF STUDY

Introducción a la Ciencia de Datos

Kaggle knowledge competition – Bike Sharing Demand

- participants are asked to forecast bike rental demand of Bike sharing program in Washington, D.C based on historical usage patterns in relation with weather, time and other data.

- Using these Bike Sharing systems, people rent a bike from one location and return it to a different or same place on need basis. People can rent a bike through membership (mostly regular users) or on demand basis (mostly casual users). This process is controlled by a network of automated kiosk across the city.

Sunil Ray , https://www.analyticsvidhya.com/blog/2015/06/solution-kaggle-competition-bike-sharing-demand/

How to approach a Dataset

- 1. Hypothesis Generation
- 2. Understanding the Data Set
- 3. Importing Data set and Basic Data Exploration
- 4. Feature Engineering
- 5. Hypothesis Testing (using multivariate analysis)
- 6. Model Building

1. Hyphotesis Generation

- Before exploring the data think about the problem and gain domain knowledge

Hourly trend: high demand during office timings. Early morning and late evening can have different trend. Low demand during 10:00 pm to 4:00 am.
Daily Trend: Registered users demand more bike on weekdays as compared to weekend or holiday.
Rain: The demand of bikes will be lower. Higher humidity will cause to lower the demand and vice versa.
Temperature: positive bike demand correlation with higher temperatures? Pollution: If the pollution level higher bike use? (influenced by government policies).
Time: Total demand should have higher contribution of registered user as compared to casual because registered user base would increase over time.
Traffic: It can be positively correlated with Bike demand.

Understanding the dataset

- The dataset shows hourly rental data for two years (2011 and 2012).
- The training data set is for the first 19 days of each month.
- The test dataset is from 20th day to month's end. We are required to predict the total count of bikes rented during each hour covered by the test set.
- In the training data set, they have separately given bike demand by registered, casual users and sum of both is given as count.

Understanding the dataset

Independent Variables

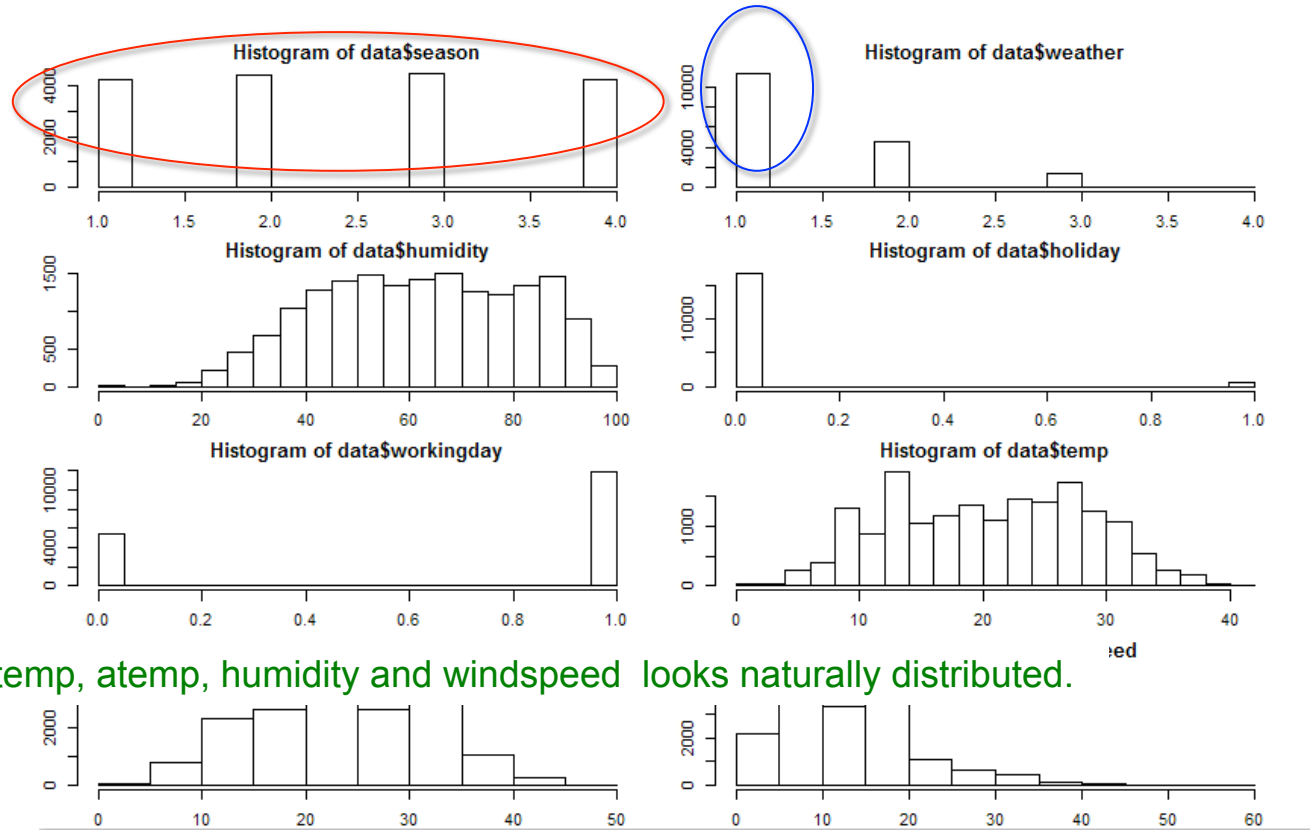
- datetime: date and hour in "mm/dd/yyyy hh:mm" format
- season: Four categories-> 1 = spring, 2 = summer, 3 = fall, 4 = winter
  holiday: whether the day is a holiday or not (1/0)
- workingday: whether the day is neither a weekend nor holiday (1/0)
- weather: Four Categories of weather
  - 1-> Clear, 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 Pallets + Thunderstorm + Mist, Snow + Fog
- temp: hourly temperature in Celsius
- atemp: "feels like" temperature in Celsius humidity: relative humidity
- windspeed: wind speed Dependent Variables
- registered: number of registered user casual: number of non-registered

Understanding the dataset

- Variable Type Identification str(data)
- Find missing values table(is.na(data))
- Understand the distribution of numerical variables

Weather 1 has higher contribution i.e. mostly clear weather.

4 categories with equal distribution



Variables temp, atemp, humidity and windspeed looks naturally distributed.

Understanding the dataset

- Convert discrete variables into factor (season, weather, holiday, workingday)

Feature Eingeneering

- In addition to existing independent variables, we will create new variables to improve the prediction power of model

Follow…in code

Useful Packages for Data Analysis

Pre-modeling stage

- Data visualization: ggplot2, googleVis
- Data Transformation: plyr, dplyr, data.table
- Missing value Imputations: Missforest, MissMDA
- Outliers Detection: Outliers, EVIR
- Feature selection: Features, RRF, Boruta
- Dimension Reduction: FactoMineR, CCP

Modeling stage

- Continuous regression: car, randomforest
- Ordinal Regression: Rminer, CoreLearn
- Classification: Caret, BigRF
- Clustering: CBA, RankCluster
- Time Series: forecast, LTSA
- Survival: survival, Basta

Post-modeling stage

- General Model Validation: LSMeans, Comparison
- Regression Validation: RegTest, ACD
- Clasification Validation: ClustEval, SIgClust
- ROC Analysis: PROC, TimeROC