BIKE SHARING DEMAND PREDICTION

INDERJEET SINGH (173190001)
SOURAV MONDAL (173190025)
VHATKAR MANJUNATH SHRINIWAS (173190026)

IEOR, IIT BOMBAY

April 29, 2018

Outline

Introduction

- What is the Bike Sharing System ?
  Bike sharing systems are innovative ways of renting bicycles for use without the a disagreeable necessity of ownership . A pay per use system, the bike sharing model either works in two modes:
  1. Users can get a membership for cheaper rates.
  2. Users can pay for the bicycles on an ad-hoc basis.

  The users of bike sharing systems can pick up bicycles from a kiosk in one location and return them to a kiosk in possibly any location of the city.

Problem Statement

- Predicting the number of bikes which will be rented at any given hour in a given city.
- Inventory management in bike sharing system.
- The efficacy of standard machine learning techniques namely Regression, Random Forests, Boosting by implementing and analyzing their performance with respect to each other

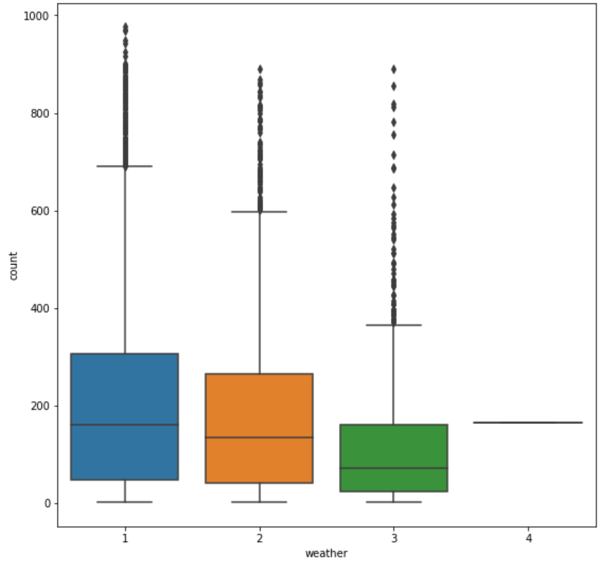
Feature Generation

- We had very low number of features (9) as compared to the number of examples (17379).
- So our training algorithm may be susceptible to very high bias.
- So from date-time time series three new features were generated as follows :
  1. Month
  2. Hour
  3. Weekday

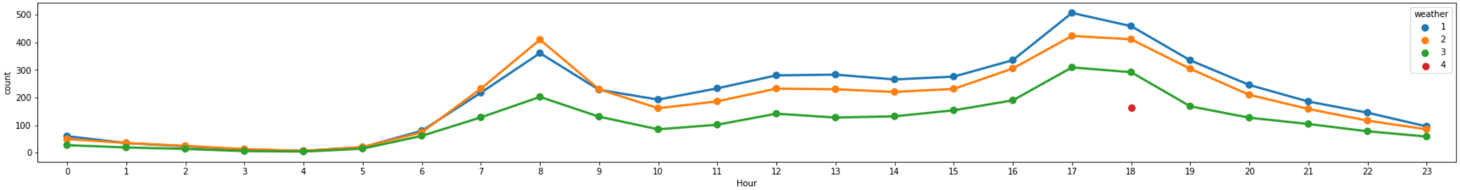
FEATURE ANALYSIS

Weather

- Demand for bike changes with weather.
- During rough weather conditions like snowfall demand is negligible as can be seen from the box plot for weather 4.



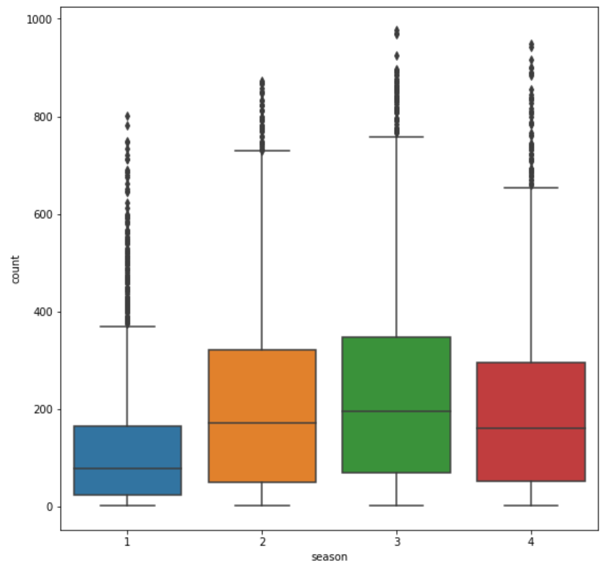
Weather



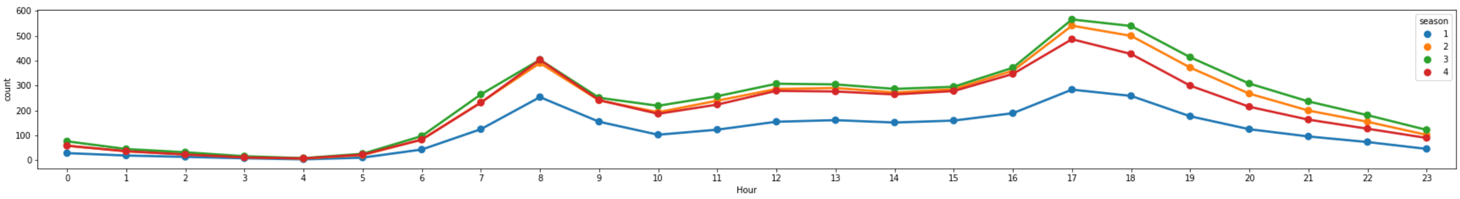
- As can be seen from the plot maximum demand occurs at 8 a.m. in the morning and 5-6 p.m. in the evening in all the seasons.
- This is so because these are office timings and during these hours there is maximum demand.

Season

- Demand for bike changes with season but it does not change drastically as in case of weather.
- During rainy season it is more likely that demand will go off.



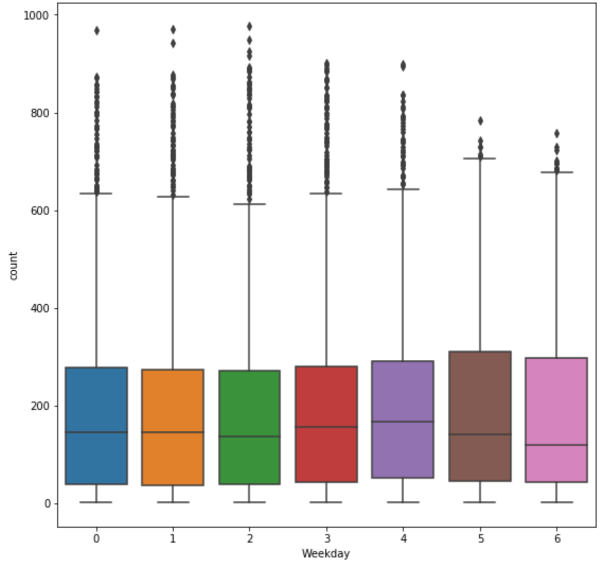
Season



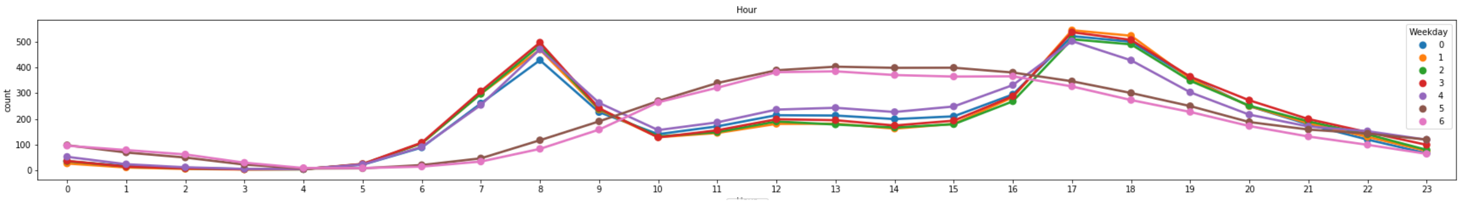
- Time for maximum demand does not change with weather as well as season
- So we can say office timings does not change with season or weather

Weekdays

- Demand for bike is nearly same throughout the week except a slight increase in two days
- Which are these two days?



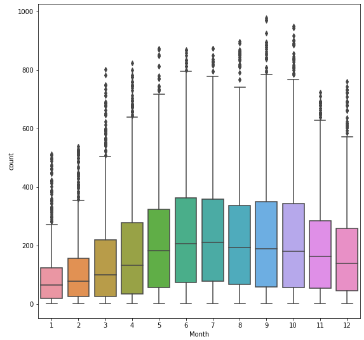
Weekdays

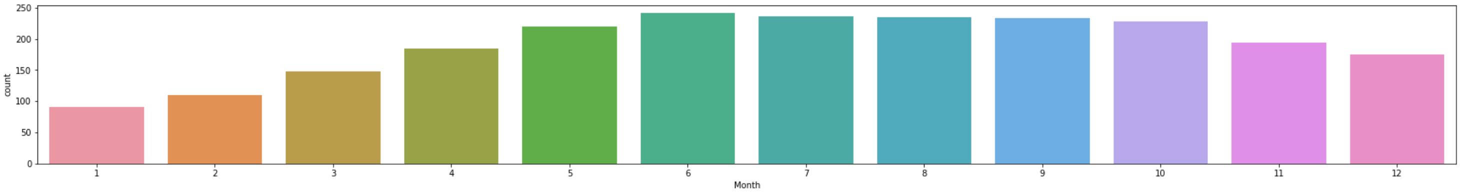


- Two weekdays have different demand pattern during the day.
- These weekdays are Saturday and Sunday as the demand is high during afternoon when people go for outing during weekends

Month

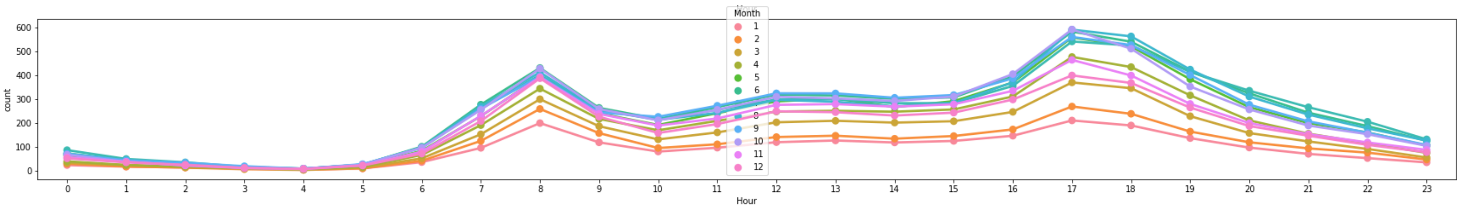
- Demand for bike varies throughout the year





- Month 1,2,11 and 12 have low demand so these are the months when snowfall takes place.

Month



- Hourly demand for bike remains the same throughout the year.

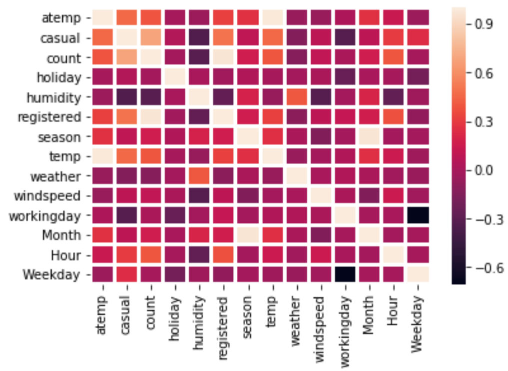
Correlation Analysis

- Correlation Analysis is used to study the strength of relationship between two features.

$$r = \frac{N \sum XY - (\sum X \sum Y)}{\sqrt{[N \sum x^2 - (\sum x)^2][N \sum y^2 - (\sum y)^2]}} \tag{1}$$

- HIghly correlated features not contributing in the learning of the algorithm are need to be dropped.

Correlation Analysis



- Following pairs of features have a high correlation values
  1. Count and Registered
  2. Atemp and Temp
  3. Season and Month
- These were eventually dropped.

Dropped Features

- Following features were dropped as they were highly correlated:
  1. Average temperature
  2. Casual (booking)
  3. Registered (booking)

Fine tuning

- To protect our models from over-fitting and for hyper parameter tuning we used Grid-search CV from sklearn.model-selection.
- There were some outliers in humidity and wind speed data, they were replaced by relatively feasible values.
- To train our models data was split into training set (80% of data) and test set (20% of data) using train-test-split of sklearn.model-selection.

MODELS AND THEIR ANALYSIS

Root Mean Squared Logarithmic Error Value (RMSLE)

- RMSLE is a technique to evaluate the quality of regression model
- RMSLE evaluate the model on the basis of difference between log of actual and predicted regression values.
- The formulae for RMSLE is as follows:

$$RMSLE = \sqrt{\frac{1}{n}\sum_{i}(\log(P_i+1)) - (\log(a_i+1)))^2} \tag{2}$$

LINEAR REGRESSION

- Root Mean Squared Logarithmic Error Value (RMSLE) = 1.048

RIDGE REGRESSION

- Best Parameter $\alpha = 400$,max iteration = 3000.
- Root Mean Squared Logarithmic Error Value (RMSLE) = 1.047

LASSO REGRESSION

- Best Parameter $\alpha = 0.005$,max iteration = 3000.
- Root Mean Squared Logarithmic Error Value (RMSLE) = 1.049

MODELS AND THEIR ANALYSIS

RANDOM FOREST REGRESSION

- Best Parameter n-estimators = 497
- Root Mean Squared Logarithmic Error Value (RMSLE) = 0.3687

GRADIENT BOOSTED REGRESSION

- Best Parameter n-estimators = 2600, learning-rate = 0.1
- Root Mean Squared Logarithmic Error Value (RMSLE) = 0.3418

THINGS TRIED THAT DID NOT WORKED

We tried contacting Zoomcar for getting dataset for PEDAL bi-cycles in our institute but we did not got any reply.

THANK YOU