Campus Gym Crowd Prediction

Objective

- Campus Gyms have limited number of workout equipment, to find a perfect time slot when gym is less crowded is challenging.

- The objective is to predict how much crowded the Gym is at given point of time and what are the factors affecting number of people using the gym.

- Predicting how crowded gym will be in future and accordingly can be used by the students at the campus.

Introduction

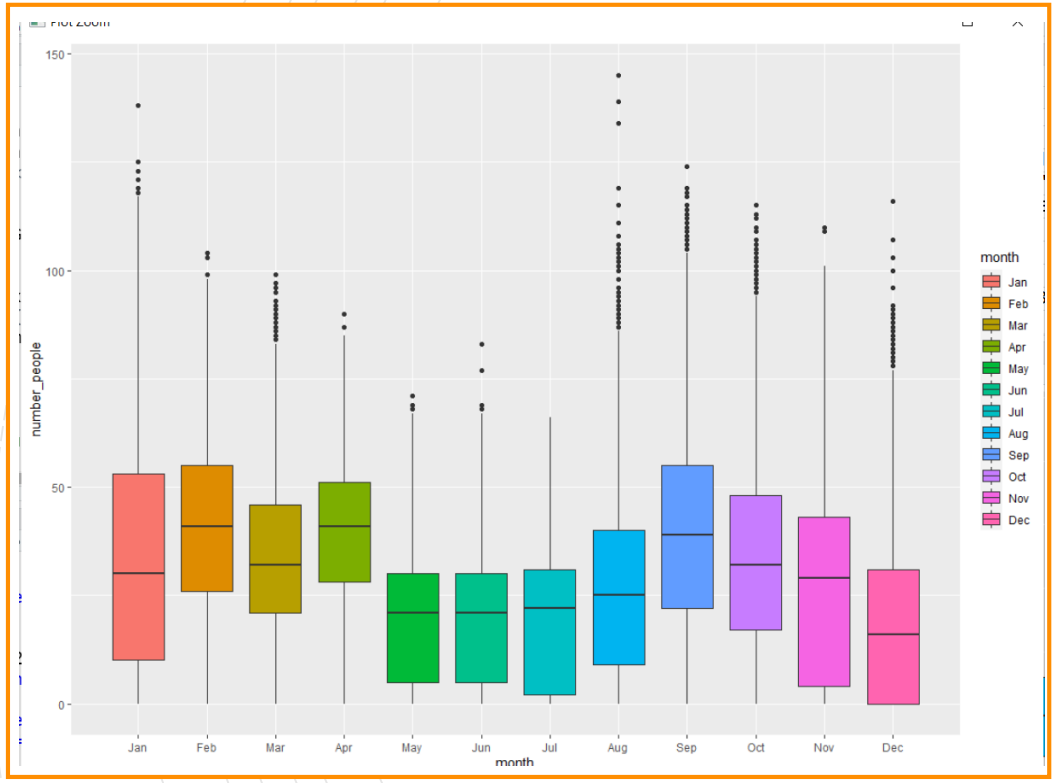
- The dataset consists of 26,000 people counts (about every 10 minutes) over the last year.
- 62,000+ observations has been collected which also includes weather and semester-specific information that might affect how crowded it is.
- The label is the number of people, which is to be predicted from the given some subset of the features and it is the dependent variable.
- The independent variables in the data-set are as follows:

  date (string; datetime of data)
  timestamp (int; number of seconds since beginning of day)
  day*of*week (int; 0 [monday] - 6 [sunday])
  is_weekend (int; 0 or 1) [boolean, if 1, it's either saturday or sunday, otherwise 0]
  is_holiday (int; 0 or 1) [boolean, if 1 it's a federal holiday, 0 otherwise]
  temperature (float; degrees fahrenheit)
  is*start*of_semester (int; 0 or 1) [boolean, if 1 it's the beginning of a school semester, 0 otherwise]
  month (int; 1 [jan] - 12 [dec])
  hour (int; 0 - 23)

Data Cleaning and Preparation

- At first, we have identified any possible blank spaces within the character variables in the data-set and replaced them with NA.
- Then we removed the duplicate entries within the data-set using duplicate() function in R.
- We further extracted the minutes from the 'date and time' variable using minute() function and assigned the Universal Coordinated Time Zone to parsed date value using ymd_hms() function.
- For proper scaling we converted the values of temperature variable into degrees from Fahrenheit.

Exploratory Data Analysis



- The following Box-plot shows us the number of people visiting Gym in different months.

Methodology

We have applied Decision tree and linear regression models to interpret the influence of each variable on the targeted variable.

Decision Tree

- All 8 Variables are used
- 12 Terminal Nodes
- Total RMSE:14.32

Multiple Linear Regression

- All 9 Variables are used
- Total RMSE:14.34

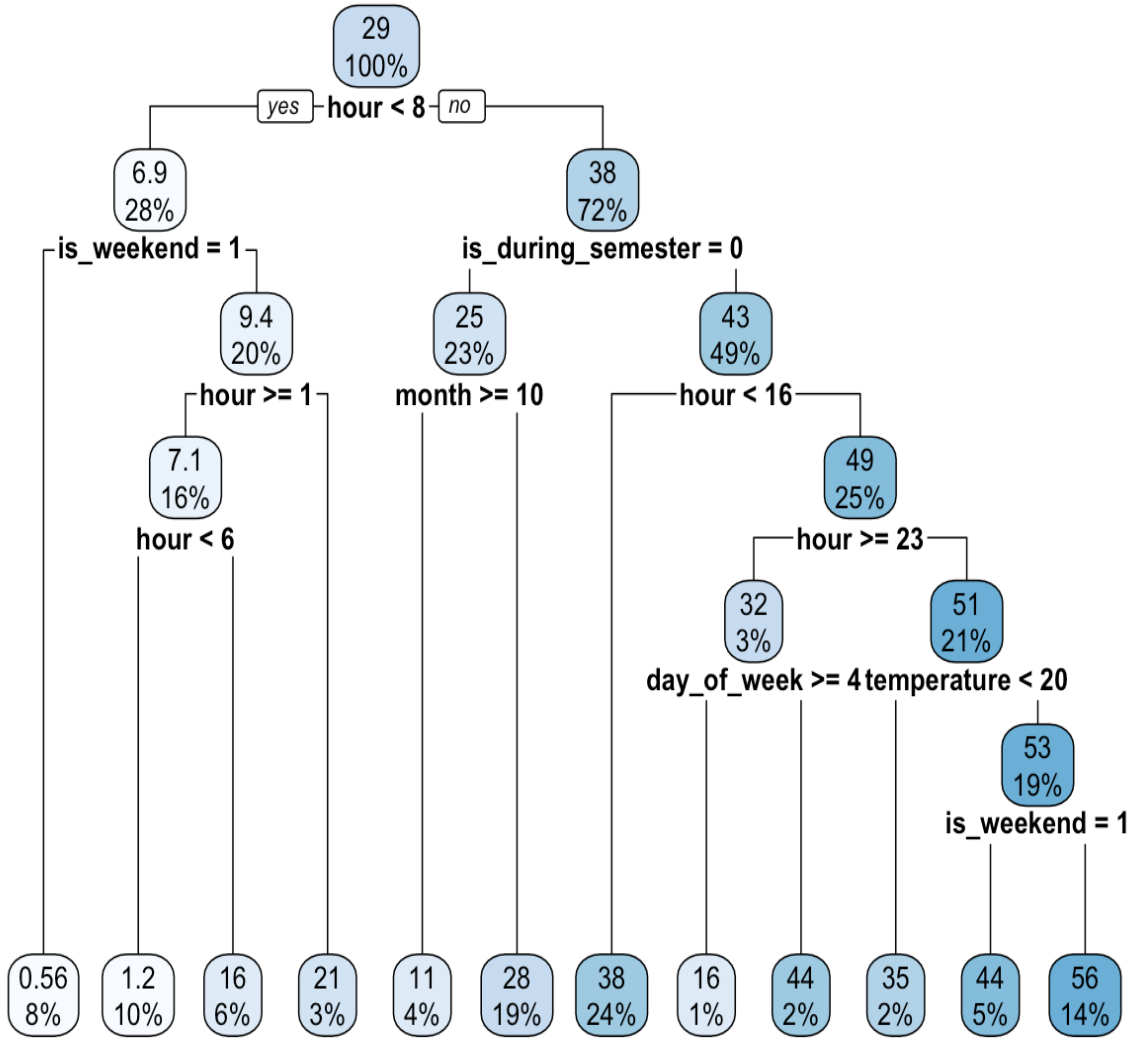
Random Forest

- All 9 Variable are used
- Total RMSE:7.41

Results and Discussion

Decision Tree

In decision tree it looks like people go more to the gym during the semester, around 22h, when weekend is near and temperature is greater than 20º C.



Linear Regression

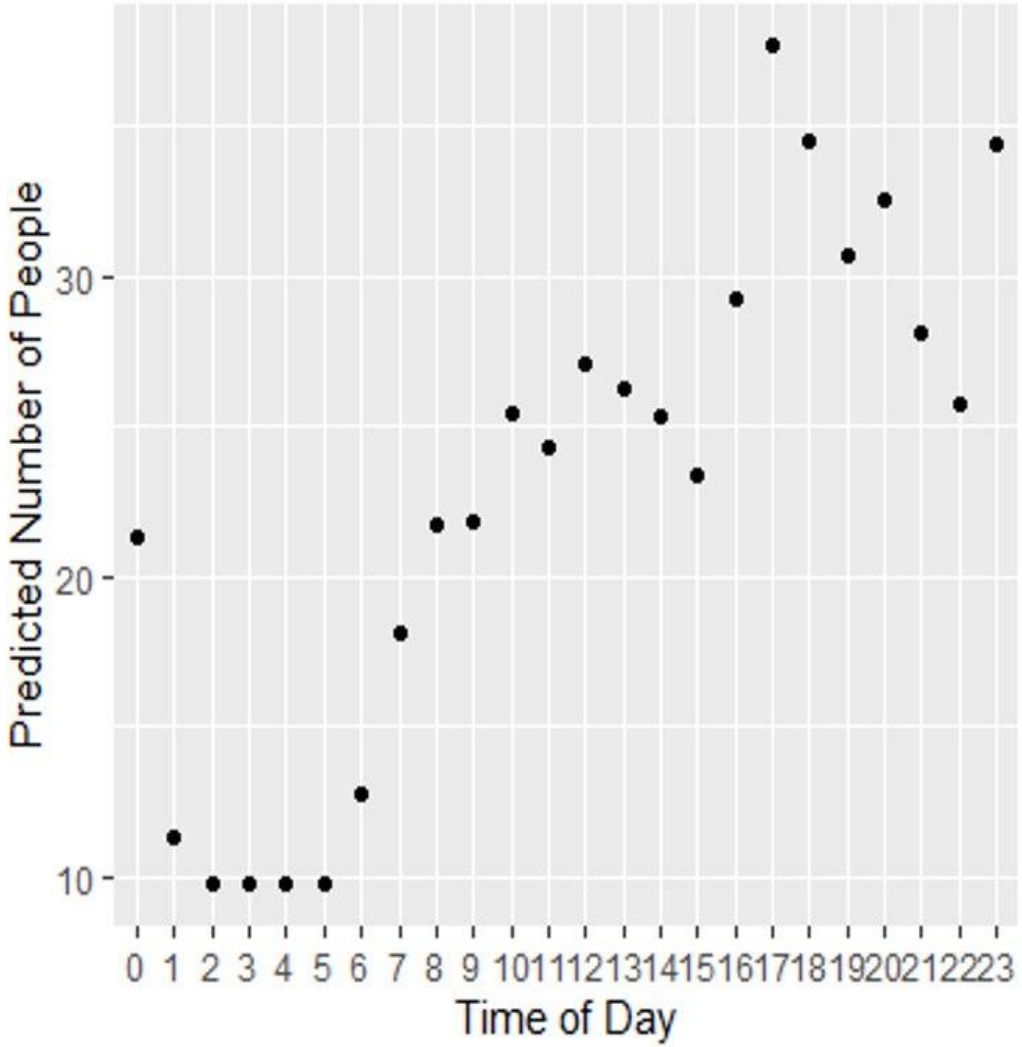
LR model shows more people go to the gym during semester (at the beginning of it) and less on holidays. At the end of the week (during the weekend) people go less, and in general, people go more at later (greater) hours.

Random Forest

To reduce the Root Mean Sqaure Error Rate of both the decision and linear regression, as it use multiple decision tree to till the result.

Final Recommendation

- Students can see when the gym is least busy.
- Help to find the perfect time slot .
- The gym can schedule workout classes appropriately.
- The gym will know how many employees to have on hand
- Identify maintainance time of Gym.



THANK YOU