Bike Sharing System

-Subba Rao

Overview

Bike sharing systems are a means of renting bicycles where the process of obtaining membership, rental and bike return is automated via a network of kiosk locations throughout the city. With this setup people can rent a bike from one place and return it in a different location based on the need.

Client

Capital Bikeshare in Washington DC.

Benefits to our Client

- The idea is to analyze the current bike sharing demand and to study historical usage patterns to forecast the demand for bike sharing and predict the usage to improve on the kiosk locations.
- To optimize and use the bikes to full potential so that no bike is free and customers should have the bike available whenever it is required. This can be achieved by predicting the demand for every hour using the historic demand and the factors that have influenced the demand using the data.
- The data used is hosted by UCI Machine Learning repository. Using the data the trends of bike sharing demand can be analyzed using factors like peak times, weather, traffic, weekday/weekend.

Understanding the dataset

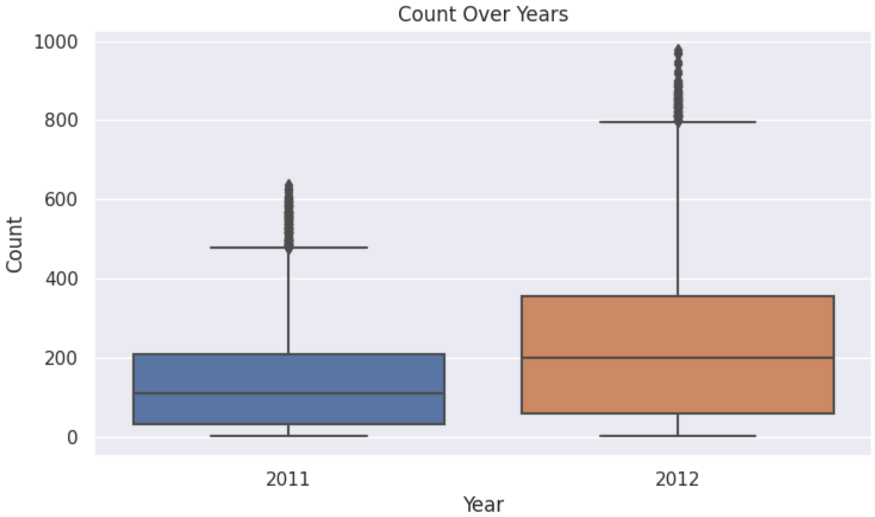
Features included in the dataset with explanation

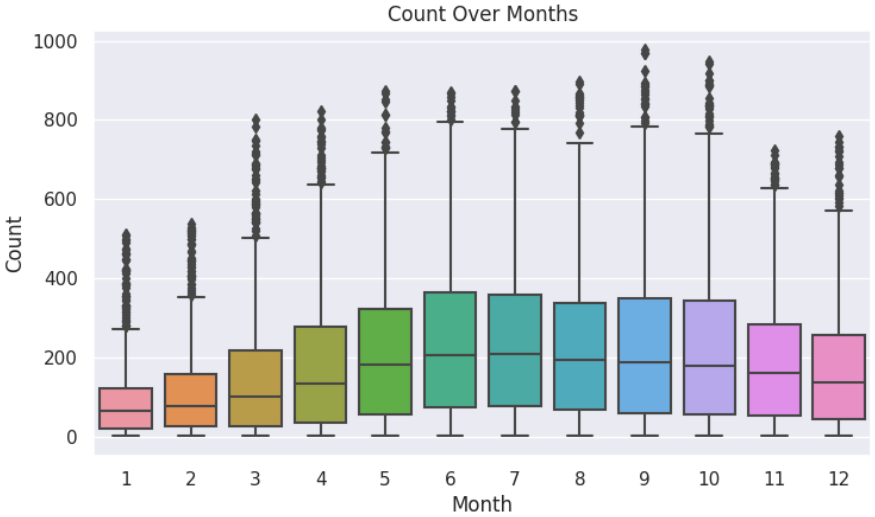
- datetime - hourly date + timestamp
- season - 1 = spring, 2 = summer, 3 = fall, 4 = winter
- holiday - whether the day is considered a holiday
- workingday - whether the day is neither a weekend nor holiday
- weather -
  - 1: Clear, Few clouds, Partly cloudy, Partly cloudy
  - 2: Mist + Cloudy, Mist + 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 - temperature in Celsius
- atemp - "feels like" temperature in Celsius
- humidity - relative humidity
- windspeed - wind speed
- casual - number of non-registered user rentals initiated
- registered - number of registered user rentals initiated
- count - number of total rentals

Exploratory Data Analysis

Data Wrangling - The data is extracted in csv format, so it can be easily loaded into pandas dataframe. Categorical data is represented as numbers which needs to be converted as category using pandas. Date time is also converted to datetime object in pandas library. The data has train and test in separate files, so train data is used for analysis. No Nans are present in the data.

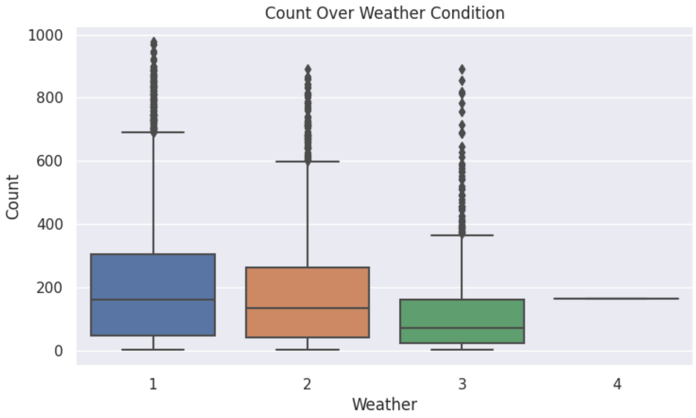
Feature Selection - Data is divided into categorical and number features which are used for evaluation and analysis. Categorical and Number features description is analyzed separately. For each categorical feature, the number of unique categories, frequency of top feature and top feature is given. For each numerical feature, the statistical numbers like mean, average, std deviation, minimum, 25% quartile, 50% quartile, 75% quartile, maximum is given.

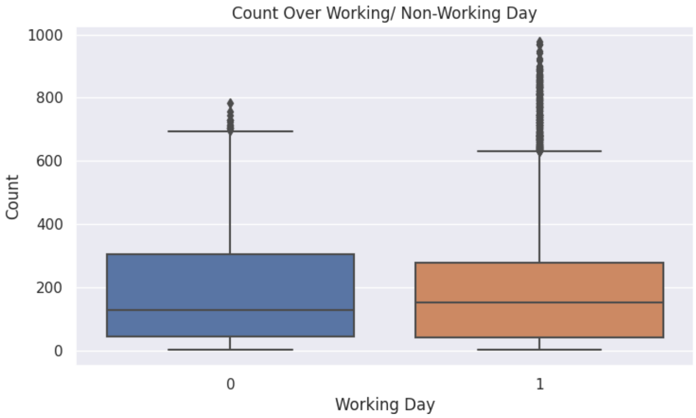


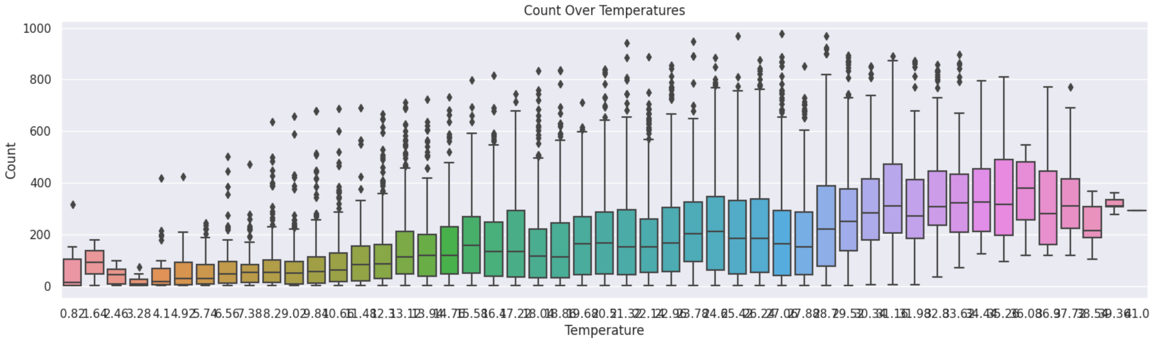


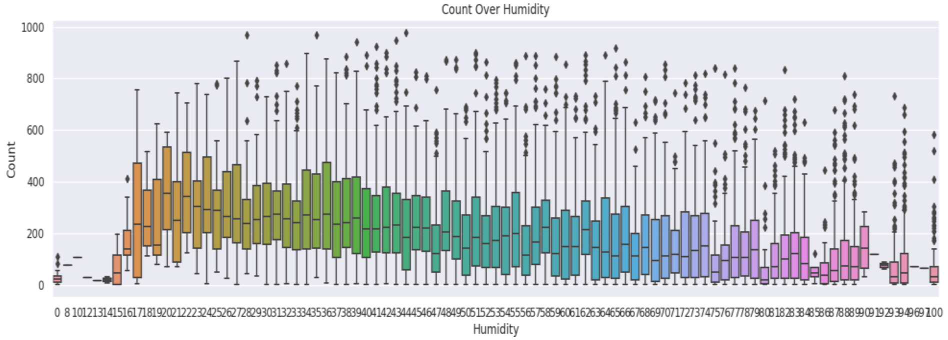
- The demand for bikes has been increasing over the years 2011 – 2012
- Number of bikes rented are high during June and July

Exploratory Data Analysis



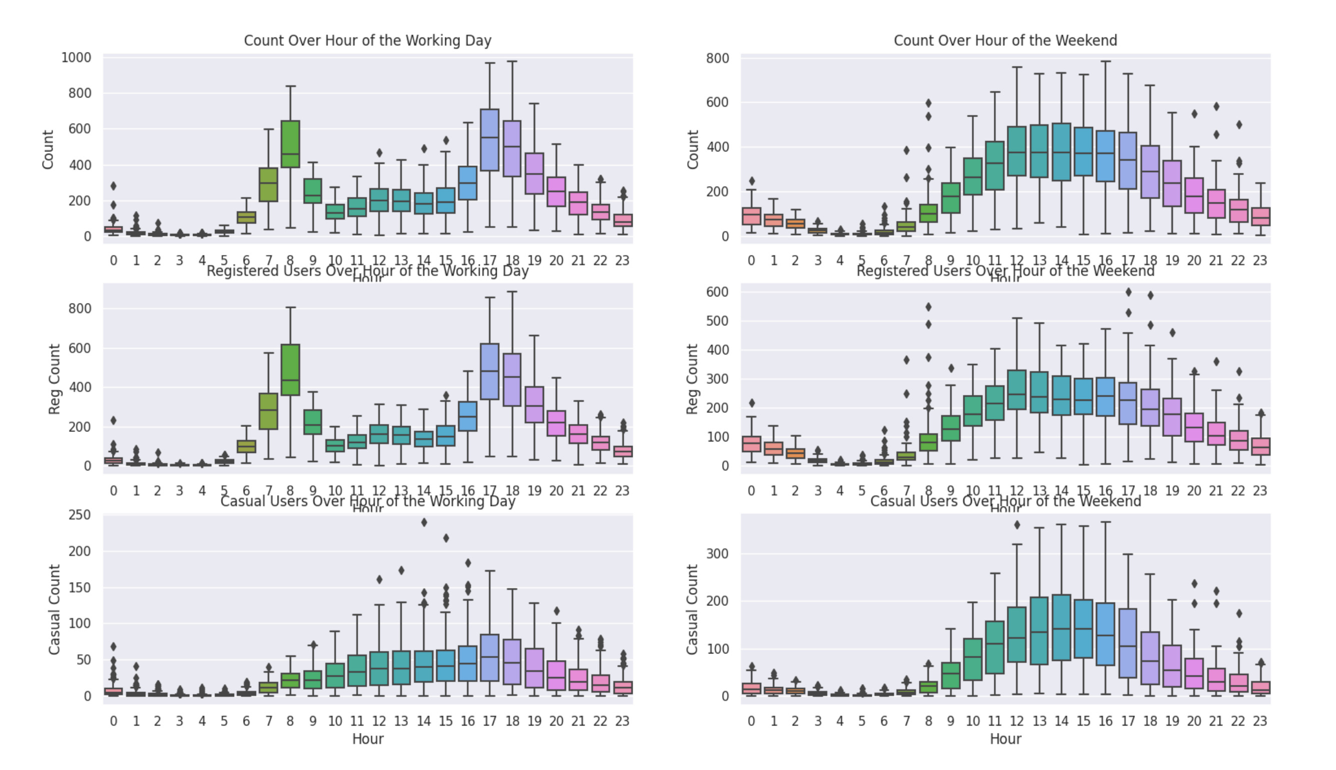






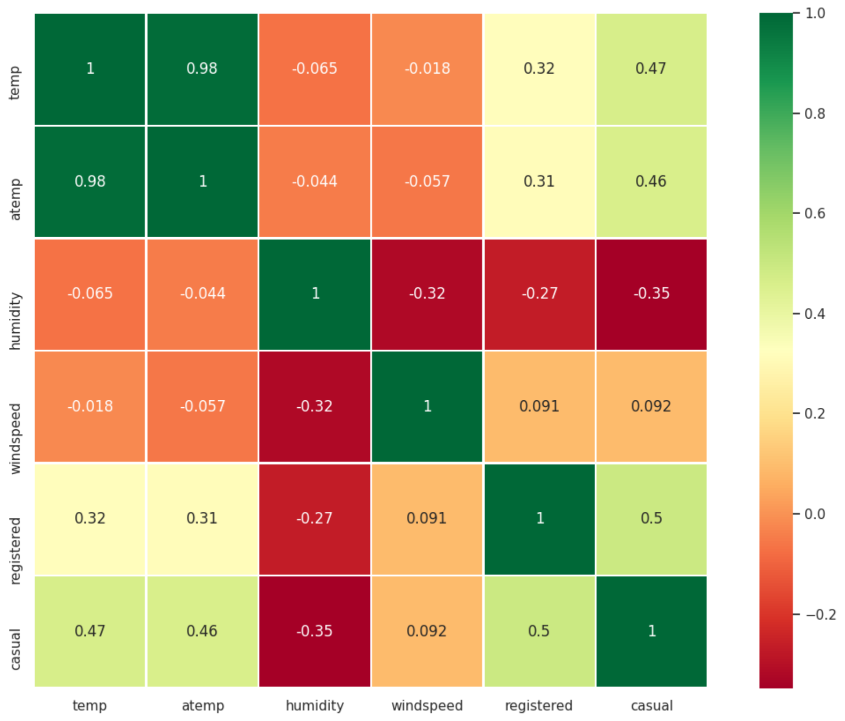
- Bike demand increases with temperature
- Humidity has a negative impact on demand

Exploratory Data Analysis



- The trends of count over the hour on a working day and weekend are different.
- The peak times during working day are 7 AM - 8 AM & 5 PM - 6 PM, this means that office commuters and school going kids are the main users during working days.
- During weekends the demand is high from 10 AM - 4 PM. The trend over a working day is majorly driven by registered users (compare registered users on working and count).

Exploratory Data Analysis

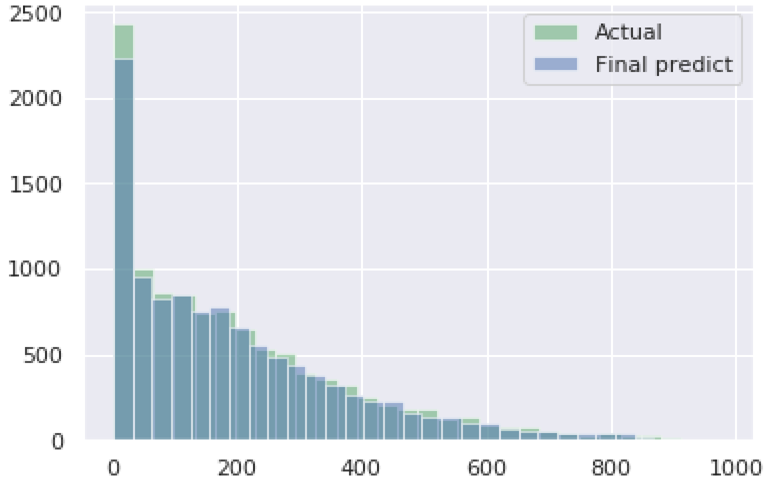


	temp	atemp	humidity	windspeed	registered	casual
temp	1	0.98	-0.065	-0.018	0.32	0.47
atemp	0.98	1	-0.044	-0.057	0.31	0.46
humidity	-0.065	-0.044	1	-0.32	-0.27	-0.35
windspeed	-0.018	-0.057	-0.32	1	0.091	0.092
registered	0.32	0.31	-0.27	0.091	1	0.5
casual	0.47	0.46	-0.35	0.092	0.5	1

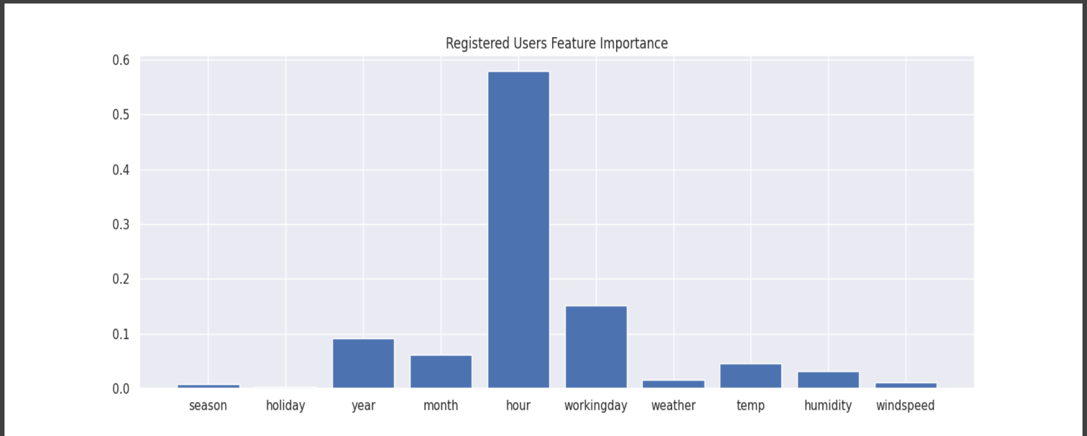
- 'atemp' and 'temp' are highly correlated. So 'atemp' feature will be removed.

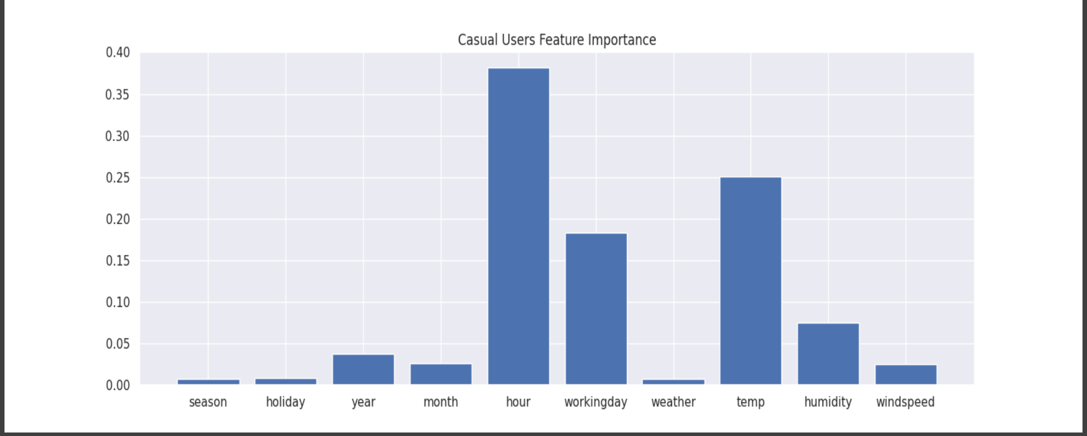
Model Selection

- Box plots show that linear relationship between features and target is not present.
- Two models are tested 1) Random Forest 2) Gradient Boost with scikitlearn in python.
- For each model, there are two targets i.e., casual and registered bike count predictions are run independently and added later.
- 10% of training data is split for testing the model and remaining data is passed for cross validation using GridSearchCV. This avoids overfitting of model.
- RMSLE & Accuracy of the model on training data are used as performance metrics.
- Random forest is selected for prediction on test data.
- The histogram comparison of actual count vs predicted count is shown.



Model	RMSLE	Accuracy (%)
RandomForestRegressor	0.2155	83.33
GradientBoostRegressor	0.5174	60.29





Feature Importance

- In both models hour feature has the highest importance.
- In case of casual users, temperature feature has good importance meaning the usage of bikes might not be preferred in colder temperatures.

Thank You