Bike Sharing

Part -1 -Data Analysis and predictive model

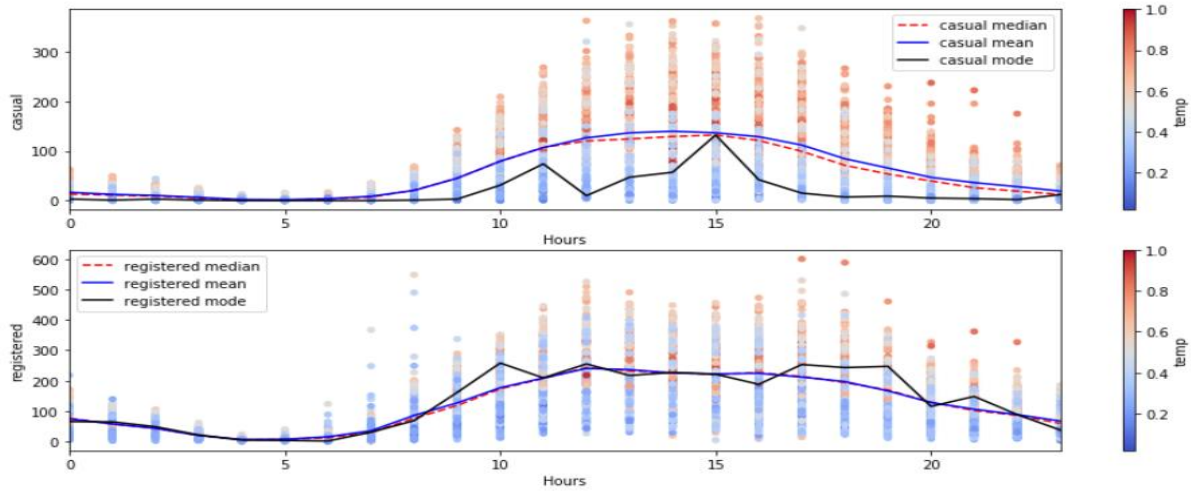
Introduction:

Bike sharing systems is playing and important and essential role in today's transportation system it has gained increased attention around the world. As the users tend to rent a bike whenever they need one. Thus, bike provider companies need to allocate bikes efficiently according to the demand. Prediction of these bike demands across different areas over different times is crucial. The goal of this project is to predict the number of cycles rented for the presented features in the months November and December.

Analysis and Preparation of the data

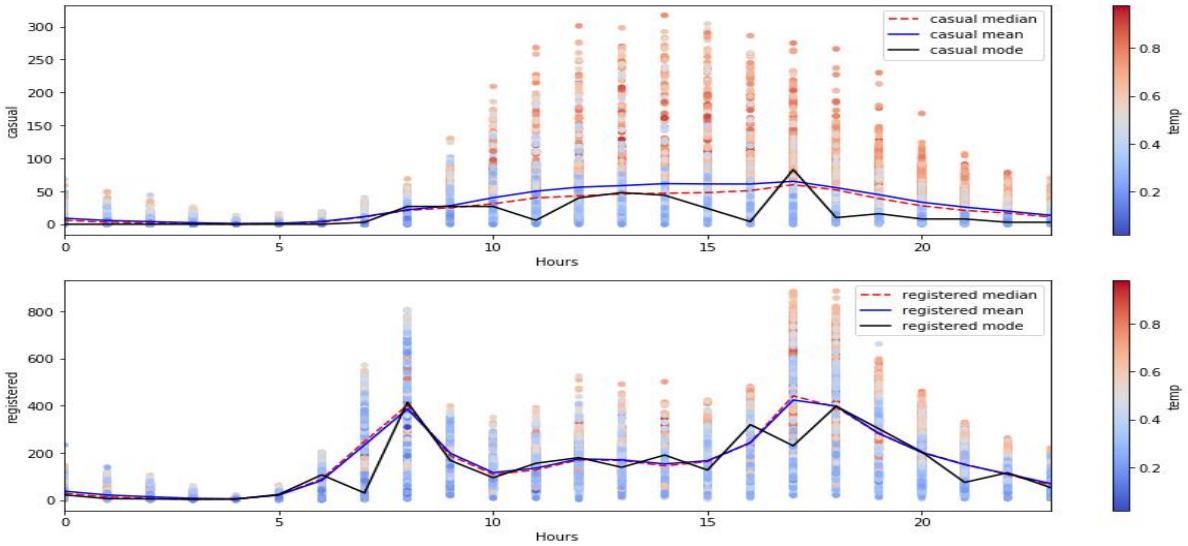
Pre-processing the data:

Firstly, I checked for the nan values in the dataset as there are no nan values, I had to look for outliers as they tend to cause a problem with the machine learning algorithm as it is sensitive to outliers in the features, so removing them could increase the accuracy. Through exploratory data analysis I found that mean and median of the data is almost equal so there aren't any potential outliers which might change our result.

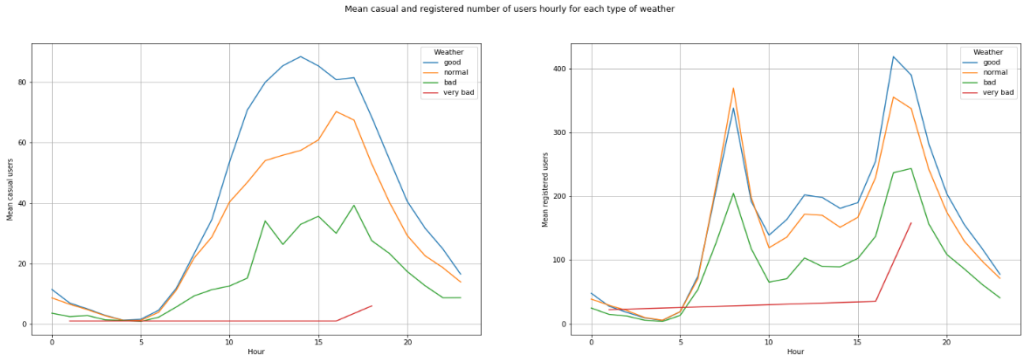


The above plot for non-working days. We can see the trend for non-working days the demand by both casual and registered users is higher during mid-day compared to working days.

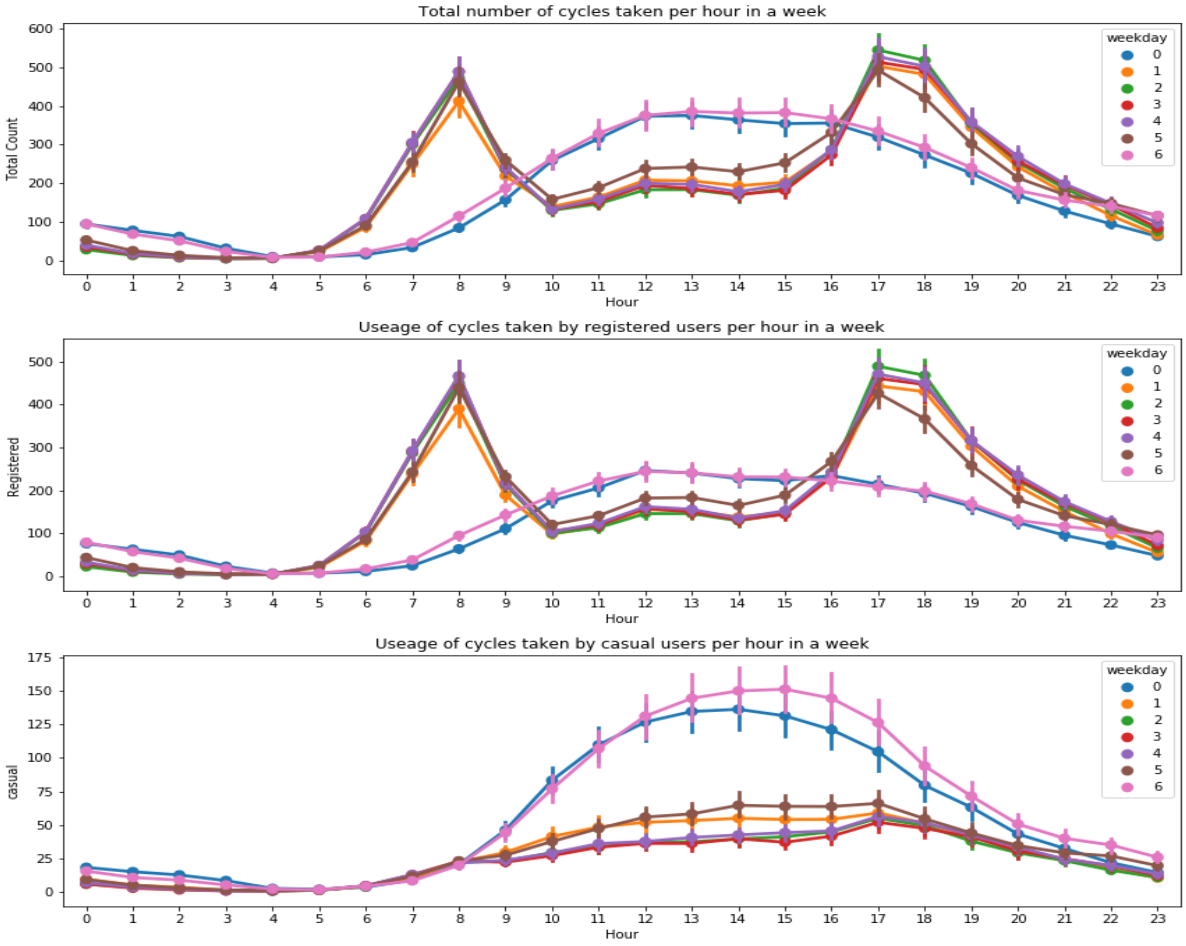
During Weekdays. The median of the casual user demand is low, and the demand is increasing with temperature. As the mean and median are almost in the same line, we could also say that there is no potential outlier in the data frame.



Bike Sharing

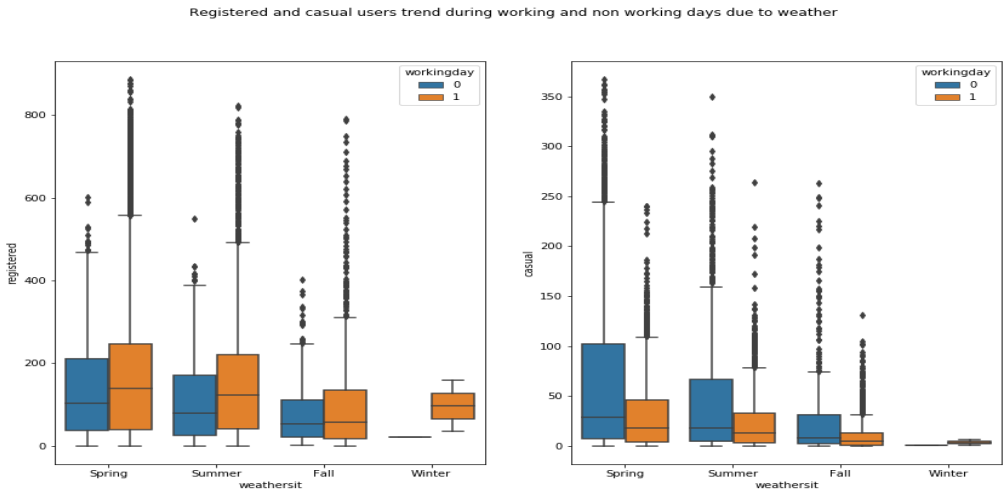


As we can see both casual and registered users prefer a good or normal weather to go for a ride
The weather is directly proportional to the count

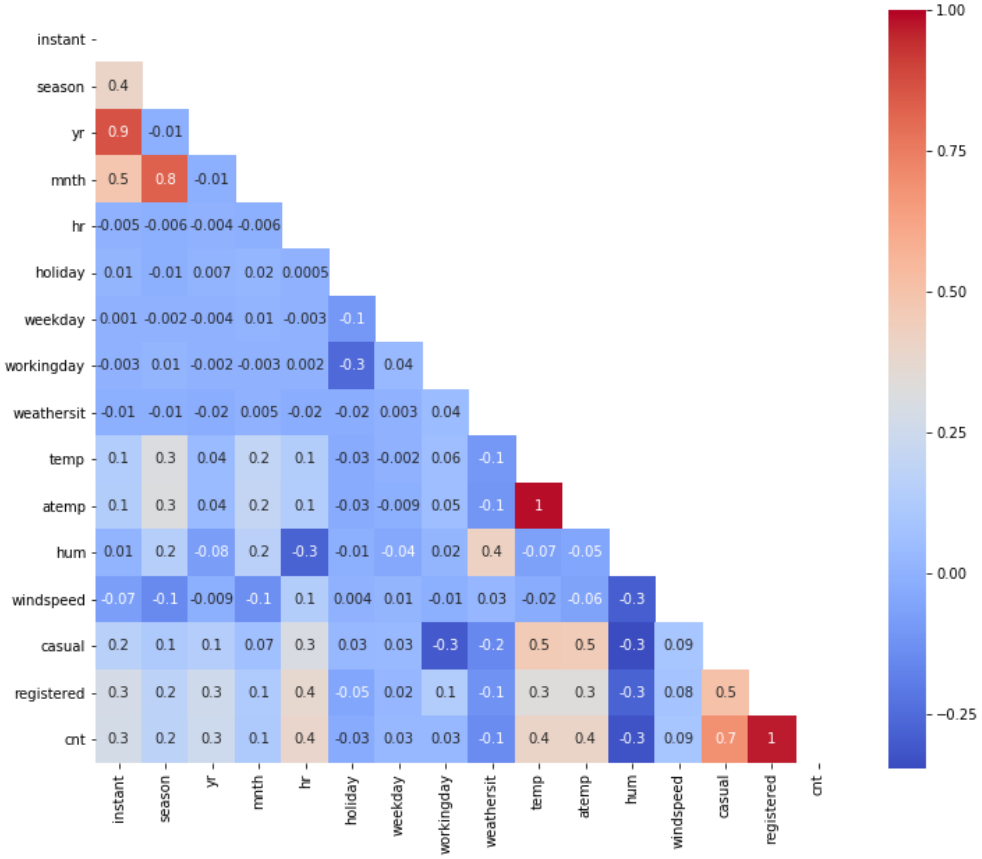


We can see a trend here as registered users tend to rent the bikes during weekdays that is from 1 to 5 at a time period of 7:00 to 9:00 am maybe riding to work and in the evening at 17:00 to 19:00 after work riding to home. as for the casual users they tend to rent the bikes during weekends for leisure.

Bike Sharing



we can observe that on working day we have higher demand rather than non-working day even due to bad weather. We had some categorical features which had been changed from numbers to denote whether or not the bike is used for that particular feature.



the correlation matrix of the numerical features revealed that the hour and temperature are some promising feature variables to predict the hourly count value. The correlation analysis also revealed that temperature and feeling temperature are highly correlated.

Bike Sharing

Model Selection:

The features of the given problem are:

1. Small dataset: which has less than 50k values
2. Regression: The target feature 'cnt' is a continuous feature or a quantity

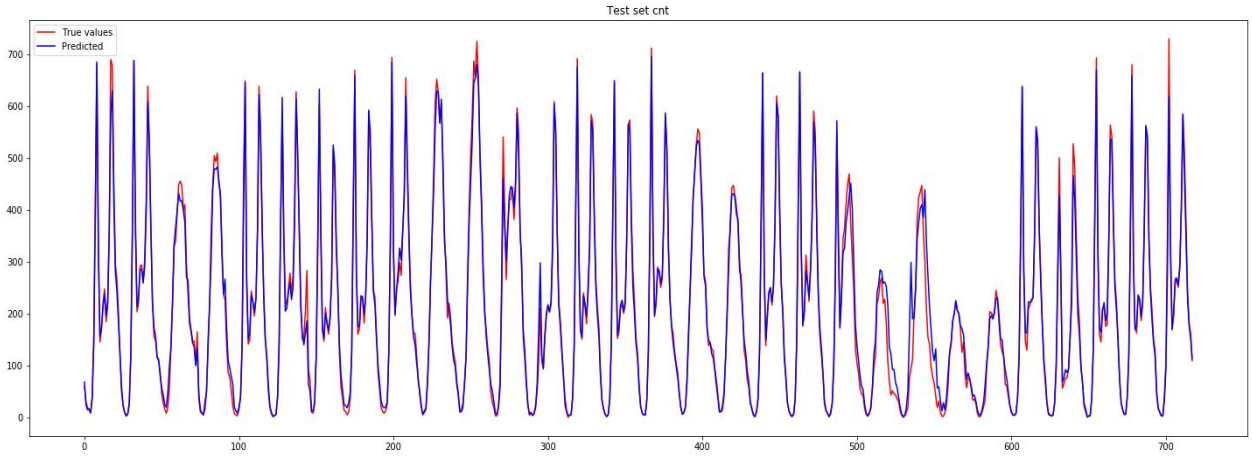
These features make the following methods such as Support Vector Regression, CatBoost Regressor, Random Forest Regressor promising. I used Random Forest Regressor for this dataset as it was more favourable.

Random forest regressor:

Random forest is an ensemble learning model used for classification and regression. The general idea of random forest is to combine large number of decision trees. Each of which are individually built on bootstrapped samples of the data. The predictions are performed by taking the mean of outputs from each individual decision tree [1].

1. One of the main reasons for using the Random forest regressor is it can deal with both categorical and numerical variables without normalization.
2. It provides importance of features based on the training dataset, which sheds light on the checkout patterns. For example, meteorology features may have higher importance on stations near tourist attractions.

The random forests showed the most promising results on the Bike Sharing Dataset and were picked for the final result.



Using all the data of 0^{th} year and excluding the 11^{th} and 12^{th} months from the 1^{st} year as training set and as for the test set, I predicted the total count of the 1^{st} year 11^{th} month with a mean absolute error of 14.404. The above graph shows the true values vs predicted values for the total count.

Using Pickle:

To save the trained models in a file and restore them in order to reuse it to compare the models with other models, to test the model on new data. When we need the same trained data in some different project or later sometime, to avoid the waste age of the training time, store trained model so that it can be used anytime in the future.

The pickle module implements a fundamental, but powerful algorithm for serializing and de-serializing a Python object structure.

Pickle.dump() to serialize an object hierarchy we use *dump()*

Pickle.load() to deseralize a data stream we call the *loads()*

Bike Sharing

Part 2

Random forest regression will slow down for large datasets this is due to the costly computations and the data can't store completely in the main memory. In some cases, the scikit learn implementation of Random forest regression can be crashed and it is not usable for large scale datasets.

A good approach would be using the distributed computation like Apache Spark or Hadoop. Apache spark is a distributed general-purpose cluster computing. It is a machine learning framework highly optimized for distributed computation.

Batch learning is also a promising method in handling large datasets. It is basically dividing the data into one or more batches. When all training samples are used to create one batch.

Another approach would be using Bag of random forests with resampling where we take a subsample of x number of rows from the total dataset and then fit a random forest on the subsample. Repeat n times. Then apply the estimator.

I have theoretical knowledge about distributed computation and cluster architectures. I have knowledge in frameworks like Hadoop and Apache Spark but I never had an opportunity to use them in my projects.

References:

[1] https://www.microsoft.com/en-us/research/wp-content/uploads/2016/07/mobisys16bike.pdf