Part 1 - Data Analysis and Predictive model

The following steps were performed to complete the analyze of Bike Sharing Dataset and build the predictive model.

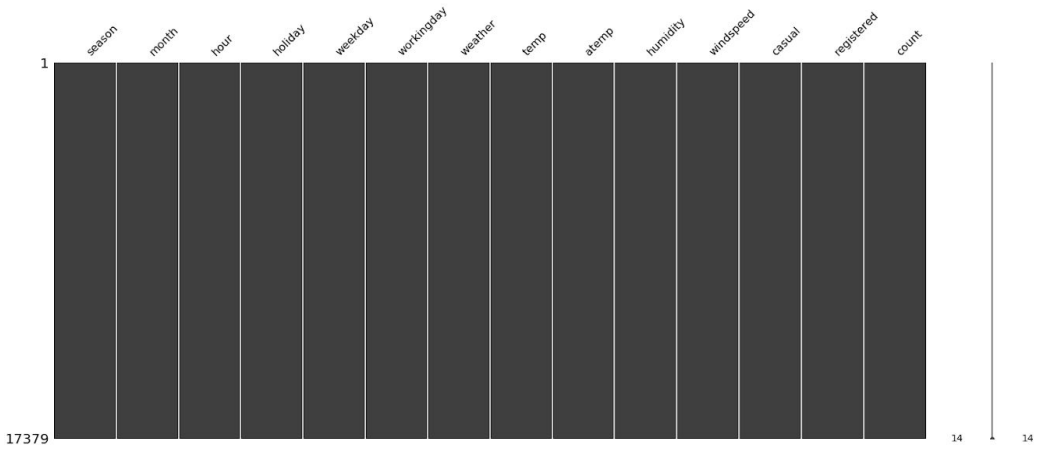
- Descriptive analysis
- Missing value analysis
- Outliers analysis
- Correction analysis
- Model Selection
- Random Forest Training

1. Business case

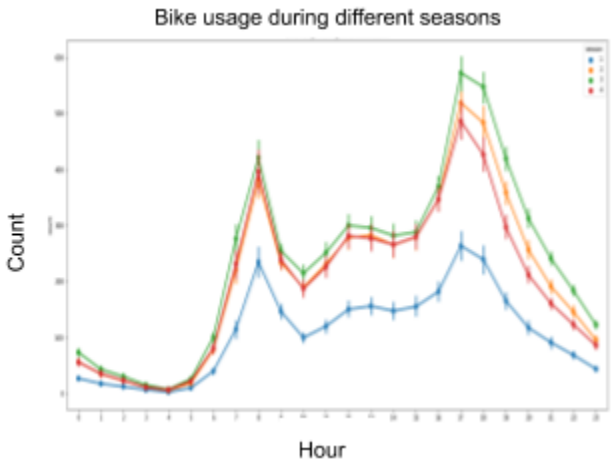
The prediction of Bikes sharing on hourly basis helps us understand the requirements of the bike at a particular station. We can learn if the allotted bike at that station is enough for example, if the bike returned at a particular station is more than the slot then there is a need to build another station nearby. The data can help us understand the number of bikes being rented from the different station and the prediction can help us to refill new bikes. The task focuses on predicting the hourly usage of bike considering day hours, weather conditions and seasons.

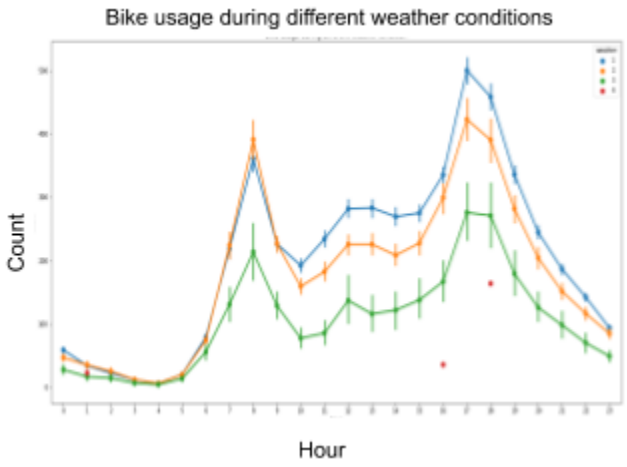
2. Data Analysis

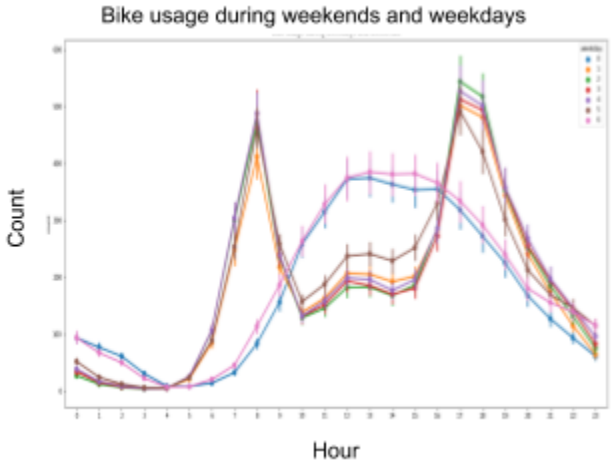
The data consists of 17379 rows and 17 columns of which some of them are dropped as it does not contains important information and would not contribute much towards model prediction. In

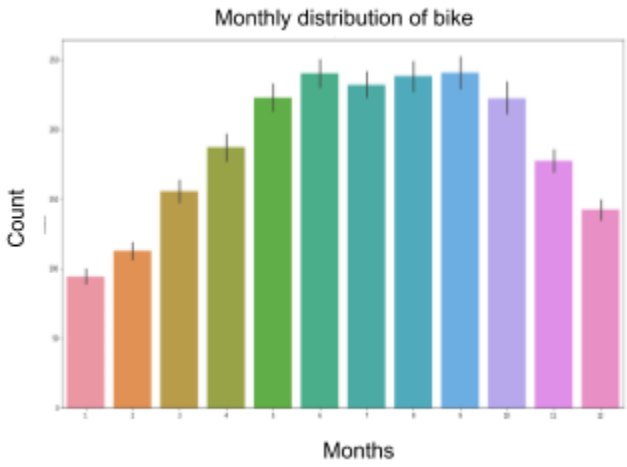


the missing value analysis the data does not consist of any NULL or NAN values. The matrix shows data with 14 features and the line plot represents the hourly pattern. The weekday and weekend plot shows that the bike is rented more during office hours and the weather plot shows the bikes are rented more in summer and fall and less during the winter and spring.

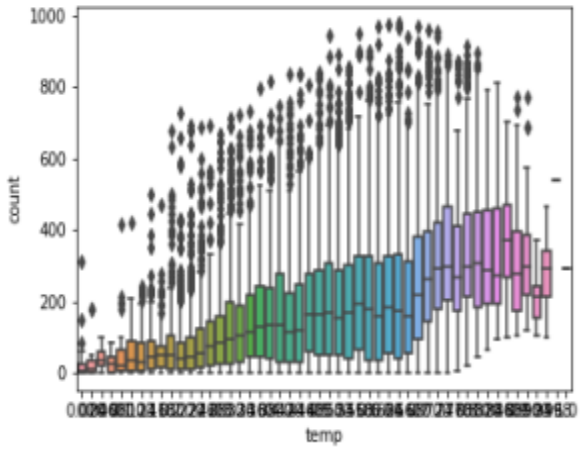


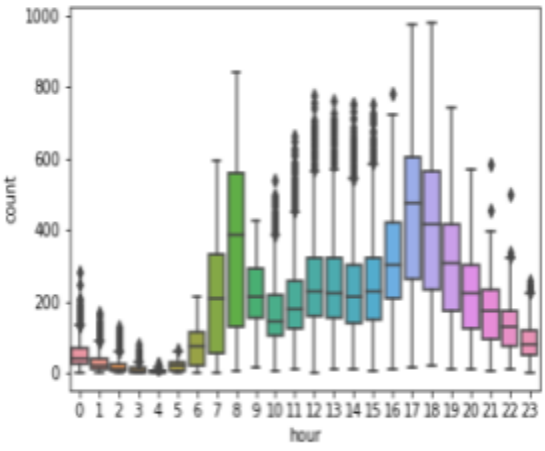


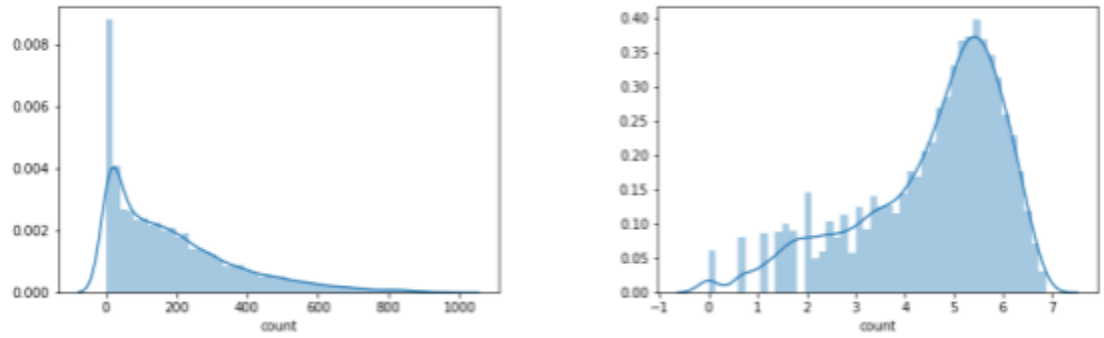




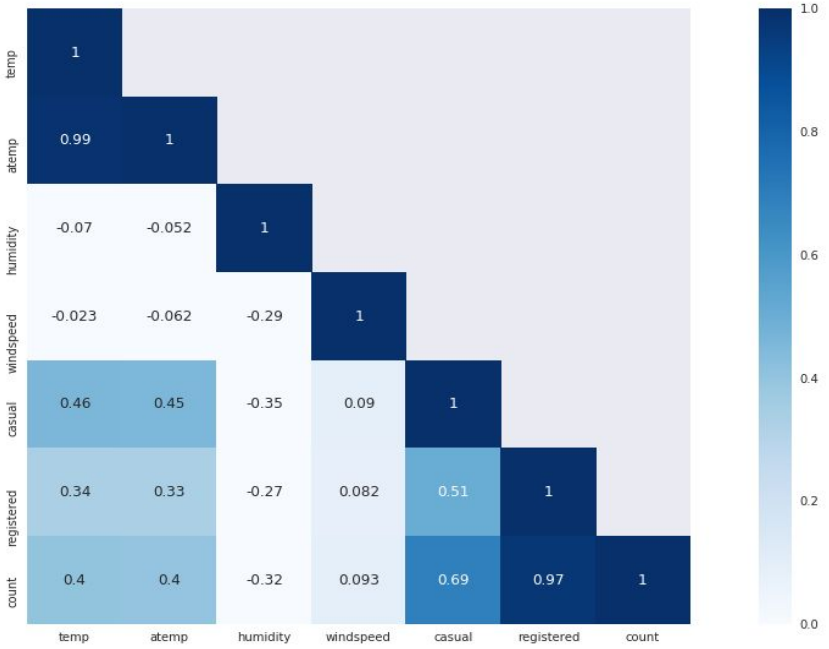
The box plot shows bike renting distribution with changing temperature and as the temperature increases bikes renting increases with less outliers while less temperature consist of outliers.







The count in the data was being normalized and we can see in the histogram plot. The correlation analysis represents temperature and feeling temperature are highly correlated along with it the sum of casual and registered users gives us count. Thus to prevent data leakage and collinearity, we drop those columns.



3. Model Selection

The data consists of both categorical and numerical features, the task is to further predict the bikes count which can be categorized as a regression problem. The data is small and as few features are significant. I have used Random forest Classifier for feature selection and Random forest regressor and evaluated the model performances. The metrics used are ranking, mean accuracy, mean absolute error and root mean square log error. The reason for using Random Forest is that it uses decision trees, it can easily handle categorical and numerical data with little

preprocessing. I have tried Logistic regression and random forest regression of which random forest performed better on the Bike Sharing Dataset and hence I chose it for my final model. The random forest model consists of 300 decision trees trained on the subsamples of the dataset. To generalize and prevent overfitting I have used cross-validation and used mean to improve the predictive accuracy.

Model	**Split**	**Mean Squared Error**	**Split**	**Mean Squared Error**
Random Forest Regressor	1	0.8437337	6	0.82030546
	2	0.84785369	7	0.84930008
	3	0.84230213	8	0.83367517
	4	0.819506	9	0.8514789
	5	0.83905532	10	0.82455922
Mean AUC				**0.837176967**

The model mean accuracy is **0.837176967** using 10 split cross validation. The mean absolute error is **45.889** and RMSLE error is **0.46**.

Mean absolute error	45.889
Mean Accuracy	0.837176967
Root mean squared log error	0.46

4. Feature Ranking

To evaluate the feature performance I have used random forest classifier and the results can be seen as follows:

Ranking	1	2	3	4	5	6	7
Features	'hour'	'temp'	'atemp'	'humidity'	'windspeed'	'casual'	'registered'

Among the 14 features these 7 features are considered best, however some of these are highly correlated which was removed later on.

5. Source file

The source files consists of ProCode folder and ipython notebook.

ProCode: The ProCode folder consists of python files and the dataset. The python file does the same analysis as one can find in the ipython notebook. The objective is to display the reusability of the code that can be maintained in the production environment. The analysis are broken into functions and can be used for daily prediction task. It makes it easier to understand and debug. The code is reusable and can be maintained and also can be extended by the potential colleagues. It also consists of documentation, submission file and the unittest file.

Bike_Rental_Task.ipynb: This file consists of stepwise analysis such as descriptive analysis, missing value analysis, feature importance, outliers analysis, model selection and conclusion. For better visualization I would suggest using notebook. The notebook consists of detailed explanation of each result.

Requirements:

IPython Notebook
Python 3.7, Python notebook 3
Sklearn, Pandas, NumPy, missingno, seaborn

To run the notebook use following command:

jupyter notebook

To run the python code use the following command:

python main.py

To run the unittest use the following command:

python -m unittest

Part 2 - Scale Up/ Large-scale dataset

The Bike sharing dataset is small with just 17k samples for which the model works fine. In case of increase in data samples the model would require high computation thus increasing computation cost. The model would ultimate crash and would not be suitable for the large dataset.

To deal with large dataset where the data is in terabytes, a best solution would be using Apache Hadoop/ Apache Mahout/ Apache spark ML that uses distributed and parallel computing. It can

access data easily from apache databases and can further speed up the processing. These big data tools have proven to be successful.
The new approach comes with some limitations:

- The Hadoop file system is not fully developed and hence is rough in manner.
- Cluster management is tough, such as debugging and collecting logs.
- The MapReduce concept makes it slower and as it is not developed in C/C++ it is not so fast.

Another alternative approach can be use of Neural network and Deep Learning techniques which might require high computation but are capable of working on big dataset. The problem with deep learning is that it mostly focuses on classification and dimensionality reduction and less on regression.

In the first case, I have completed an online course on Big data Hadoop from Udemy where I learned the basics of Big data and its usage. I have subjective knowledge on distributed computing, MapReduce. I do not have any hands-on experience on Big data framework so far.

In the second case, I have worked with deep learning model where my dataset size was in Gigabytes. The data consisted of tweets which was used to predict the stock value using Long short term memory (LSTM) and Convolution neural network (CNN). I have worked with deep learning models for almost a year. I have taken subjects such as very deep learning, applications of Artificial Intelligence and completed Andrew NG course deeplearning.ai from coursera.