Bike Sharing Data Set Analytics Project

Prateek Chitpur

Dr. Liao Duoduo

George Mason University

AIT614-002 Big Data Essentials

05/15/2020

Bike Sharing Data Set Analytics

***Abstract* –** In today’s world, big data is a popular phenomenon which is driving attention of many scholars and researchers of various fields. The traditional methods of data analysis are stumbling to deal with variety of data that is humongous and generating at greater speed. So, there is an increase in demand of the cutting-edge technologies for data storage, processing, and its analysis. The hidden patterns and meaningful information can be easily derived from the analysis of big data. The results can accentuate in better decision making for the huge business organizations to gain better profit. The big data can also draw useful patterns for treatment of patients in the healthcare sector. Hence, a data analysis project is conducted on the bike sharing dataset where bike sharing organizations can make better decisions with the results of the analysis. The purpose of this project is to predict the count of bike rentals based on the consideration of various factors.

INTRODUCTION

Bike sharing is a system or a scheme where individuals share bikes (bicycles) for a short period that are publicly made available for certain amount of price. This new generation of renting bicycles made the entire process automatic. With the help of this system, any user can rent bike easily from a specific position and can return it back at different specific position. At present, there are over 500 bike sharing organizations consisting of over 500 thousand bicycles [1]. Hence, there is a vast interest in these systems as they have their major role in traffic controls, environmental and health related problems.

The data characteristics produced by bike sharing systems are driving attention for the research than the applications of these systems in real world, since the total time travel, time elapsed, arrival and departure locations are explicitly recorded in these systems. Therefore, this system acts as a sensor network, where we can study mobility and detect important events in the city. Also, bike rental demand can be predicted with the combination of historical usage pattern and weather data.

OBJECTIVES

The main objective of the study of analysis of bike sharing system data is to forecast the demand for bike rentals, as the renting process is highly correlated with the seasonal and environmental behaviors. Bike rental launchers can either increase or decrease the bikes with the help of prediction for demand. Also, count of rented bikes on specific date is correlated to the events in a city. Hence, event and anomaly can be traced with the help of search engine. For instance, search like “02/13/2011 Washington DC”, in search engine lists out important events in Washington DC.

THE DATASET

The Bike Sharing dataset is collected from UCI Machine Learning Repository for the visualization and analysis of the data, which can be found in this link. The dataset belongs to Hadi Fanaee-T [2], who collected bike share data from Capital Bikeshare, weather information from i-weather and holiday schedule data from DC.gov. This dataset contains both hourly and daily rental bike counts between years 2011 and 2012. The data folder contains two csv files: hour and day. Both the files contain same attributes, except ‘hr’ attribute is not present in day file.

Overall, the dataset contains 16 attributes and 17,389 number of instances. The total size of data is 1,130 kb. The following are the attributes present in the bike share dataset.

instant – it is the record of index of each observation. (discrete data type)

dteday – it represents date. (date data type)

season – it is the season of the year (1: winter, 2: spring, 3: summer, 4: fall). (discrete data type)

yr – it represents year (0: 2011, 1: 2012). (categorical data type)

mnth – it is the month of year from 1 to 12. (discrete data type)

hr – this field represents hour from 0 to 23. (discrete data type)

holiday – this field states whether day is holiday or not. 1 if holiday, otherwise 0. (categorical data type)

weekday – this represents day of the week from 0 to 6. (discrete data type)

workingday – if day is neither weekend nor holiday it represents as 1, otherwise 0. (categorical data type)

weathersit – it represents 1 if clear, few clouds, partly cloudy, 2 if mist and cloudy, mist and broken clouds, mist and few clouds, mist, 3 if light snow, light rain and thunderstorm and scattered clouds, light rain and scattered clouds and 4 if heavy rains and ice pallets and thunderstorm and mist, snow and fog. (discrete data type)

temp – it represents normalized temperature in Celsius scale. (continuous data type)

atemp – it represents normalized feeling temperature in Celsius scale. (continuous data type)

hum – it represents normalized humidity. (continuous data type)

windspeed – it represents normalized wind speed. (continuous data type)

casual – it is the count of casual users. (discrete data type)

registered – it gives count of registered users. (discrete data type)

cnt – it gives total count of rental bikes including both casual and registered users. (discrete data type)

For the analysis of data only hour.csv file will be used, since it has an extra attribute ‘hr’ which is not present in day.csv. Also, hour.csv file has many instances than day.csv, which can help in better analysis and visualization of data. All the columns of hour.csv file will be used for the analysis, except ‘instant’ column.

Pre-Processing

Initially, the Spark was not inferring the correct datatypes of the dataset. By default, all the datatypes of the dataset were read as string. For that a new schema is defined specifying the correct datatype. The columns ‘instant and ‘dteday’ are not selected in building machine learning models. Since the seasonal and environmental factors derive better characteristics of days in the date column. However, ‘dteday’ column is used for the visualization purpose.

THE SYSTEM

Architecture:

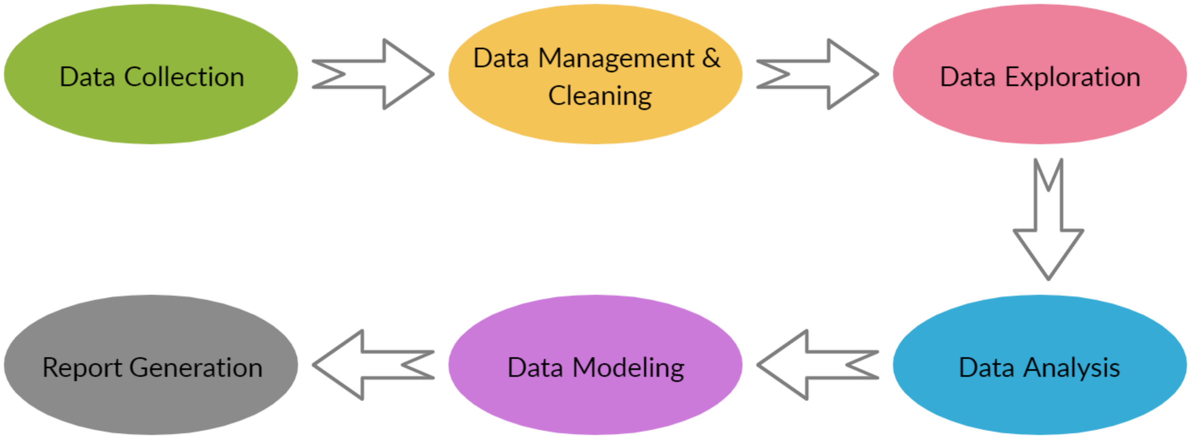


Figure 1: System Architecture

Data Collection: In this process, bike sharing dataset is collected from UCI Machine Learning Repository for the analysis.

Data Management & Cleaning: Spark is used for data managing and processing. Dataset does not contain missing values, only the 'instant' column is not required which will be deleted for further data analysis.

Data Exploration: This is an initial step of data analysis, where the characteristics of the data is explored visually. For instance, size, correctness, and completeness of data.

Data Analysis: In this stage, data is visualized with the help of Tableau and Python for data analysis. A bar graph can be plotted with attributes 'season' and 'cnt', so that minimum and maximum total count of bike rentals in each season can be obtained. A line graph plotted between 'mnth' and 'cnt' will show ups and downs of bike rental count throughout the months of respective years. Similarly, between 'dteday' and 'cnt'. Moreover, scatterplots of 'temp', 'atemp', 'hum', 'windspeed' against 'cnt' will show increasing or decreasing trend of bike rental count against environmental behaviors.

Data Modeling: In this stage a data model is created for the data. With the help of advanced analytics such as machine learning algorithms, regression and classification methods can be applied to predict the target variable. In this project, the target variable is 'cnt', which is a continuous variable, so a multiple liner regression model will be built considering all the columns of the dataset as predictors. Similarly, random forest algorithm, decision tree and gradient boosting are used as a regression task for predicting the target variable 'cnt'.

Report Generation: The results obtained from the data analysis and model buildings are collaborated into a separate final report.

Data Analytics Algorithms:

The major data analysis algorithms that are extensively used are Classification and Regression Trees (CART), K-Nearest Neighbors (KNN), K-means Clustering, Support Vector Machine (SVM) and Gradient Boosting Trees.

Classification and Regression Trees (CART) – This algorithm makes decision in categorizing the data. The decision made is based on the input variable. For each decision corresponding to response variable, the record of data moves nearer to the category that is being classified and form a tree-like structure. The end lines of decisions of classification tree called leaf nodes. The tree gets more complex and larger as it grows, which can be controlled by tree pruning.

Random Forest is one of the types of classification and regression trees. It creates many branches rather than a single tree. Each simple tree performs evaluation of data, at the end all the processes are collaborated to form single prediction.

K-Nearest Neighbors (KNN) – It is also a classification algorithm, where distance of each data point is calculated and makes decision to which class the data points fall into with respect to proximity of training observations. This algorithm can be expensive as each new observation is to be compared to all the observations of the training dataset. However, this algorithm is chosen as it is easy to use, train and interpret the outcomes.

K-means Clustering – This algorithm performs grouping of the attributes, which are termed as clusters. Once the clusters are formed, the other data points can be easily evaluated with the clusters formed based on how well they fit. Before it begins, the number of clusters to be formed is specified. This algorithm divides the data into number of clusters specified. The clusters created here are not same as classes since there were no business meanings in the beginning [3].

Support Vector Machine (SVM) – It is a supervised machine learning algorithm which is used for regression, classification and in the detection of outliers. The main aim of this algorithm to obtain the hyperplane in the n-dimensional space, where n is the number of features, that can classify the instances. There are many hyperplanes chosen to separate the classes of the instances, but the objective is to get the plane which has maximum distance in between the data points of the categories. The advantages of using this algorithm are it is effective in n-dimensional space, memory efficient as it makes use of subset of training data points in support vectors and versatile in nature [4].

Gradient Booting: It is an algorithm which converts weak learners to strong learners. In this, the new tree is a fit of the modified version the main dataset. This algorithm can be well explained by AdaBoost algorithm. In AdaBoost, it assigns equal weights to the instances by training the decision tree. After evaluation of first tree, the weights of instances are increased which are difficult in classifying. And the weights of observations are lowered that can be easily classified. The second tree grows with the newly weighted observations. The estimations of the ensemble final model are the sum of weights of previous predictions of models [5].

Existing Algorithms Usage:

The target variable in this project is a continuous variable, hence regression is performed with the machine algorithms such as Linear regression, Random Forest regression, Decision tree regression and Gradient Boosting. Linear regression easily finds the relationship between the independent variables and the dependent variable, whether it is linear or not. Random Forest regression is an ensemble algorithm, which is expected to give better prediction. Similarly, with the Decision tree regression and Gradient Boosting.

Data Visualization:

The data is visualized with the help of Tableau software, which is an interactive visualization platform.

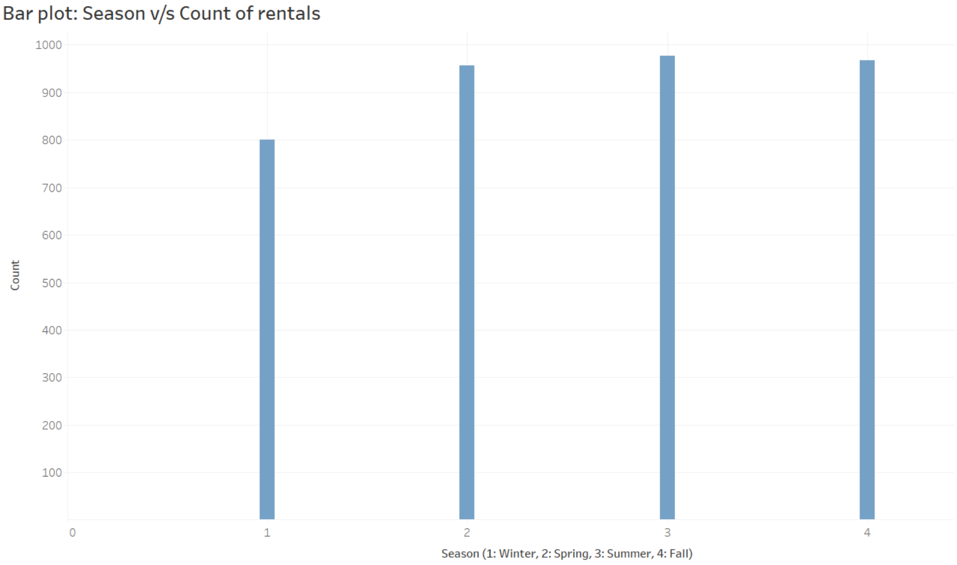


Figure 2: Bar plot of Season v/s Count

Figure 2 represents a bar plot. It is observable that season 2, 3 and 4 have larger count of bike rentals than season 1. Season 3 has the highest count compared to all the seasons.

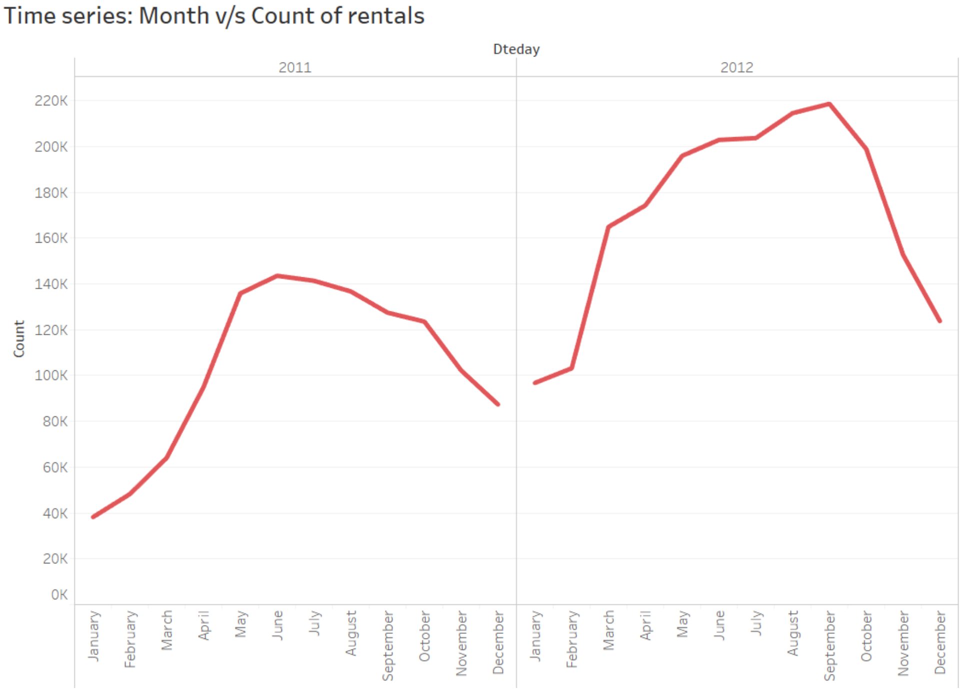


Figure 3: Time series plot of Month v/s Count

Time series plot in Figure 3 depicts the increasing and decreasing trend of rental counts over the entire months of both the years 2011 and 2012. Year 2012 has larger counts than year 2011. It is visible that counts are maximum in the month of June and September of years 2011 and 2012 respectively.

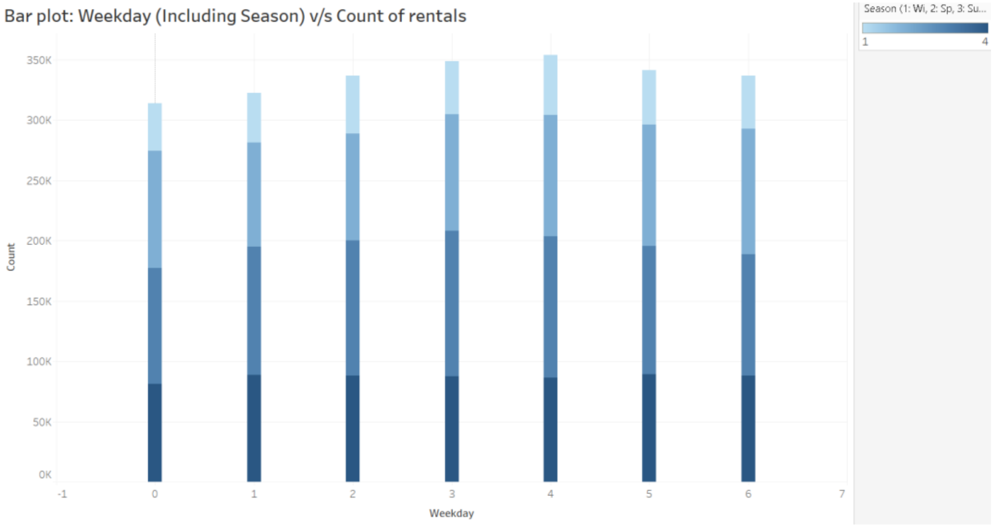


Figure 4: Bar plot Weekday v/s Count

Figure 4 illustrates count reaches maximum in the weekday 4. The weekday 0 has the lowest count of bike rentals.

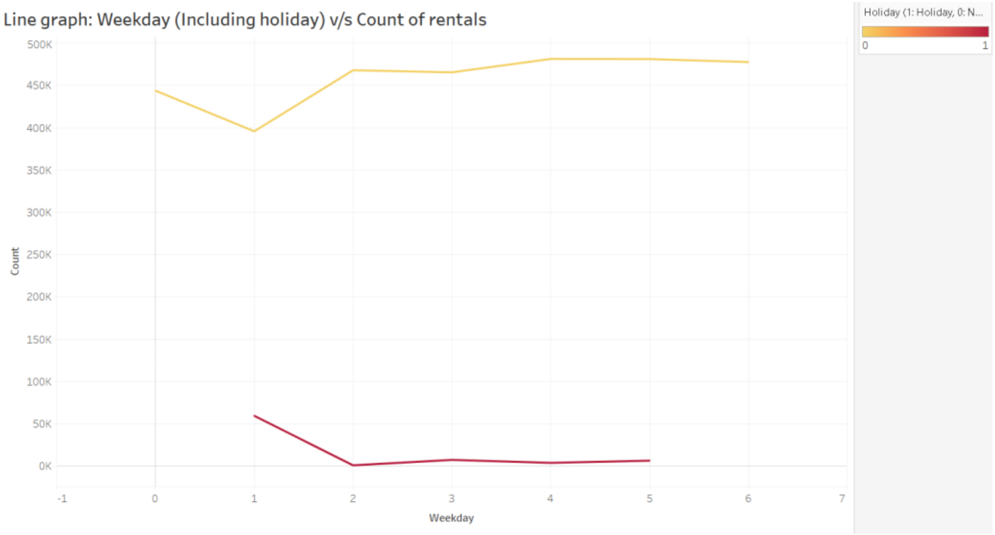


Figure 5: Line graph of Weekday v/s Count

From Figure 5, it is seen that the rental counts are maximum when it was not holiday i.e. 0.

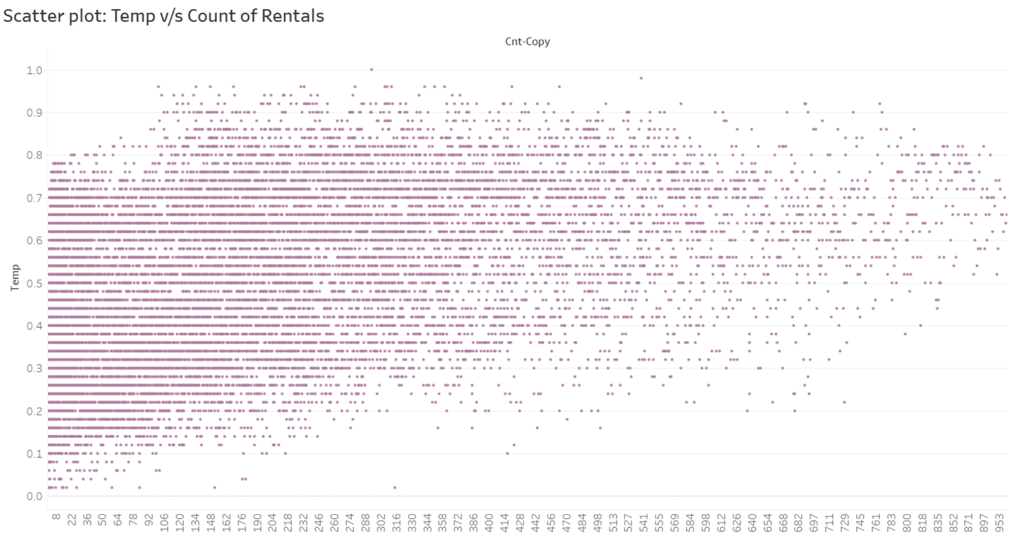


Figure 6: Scatter plot of Temp v/s Count

Scatter plot shown in Figure 6 depicts that both Temp and Count are weekly correlated. The count data points decrease with increase in temperature.

Data Analysis:

The mean value of 'temp' is 0.4969, maximum value is 1.0 and minimum value is 0.02. The mean of 'hum' is 0.6272, maximum of 1.0 and minimum of 0.0. Similarly, the mean value of 'windspeed' is 0.1900, maximum of 0.8507 and minimum of 0. The 'cnt' has maximum value of 977 and minimum value of 1.

The correlation plot is obtained with the features of dataset to show the correlation between the features. Below Figure 7 shows the correlation plot, plotted using heatmap.

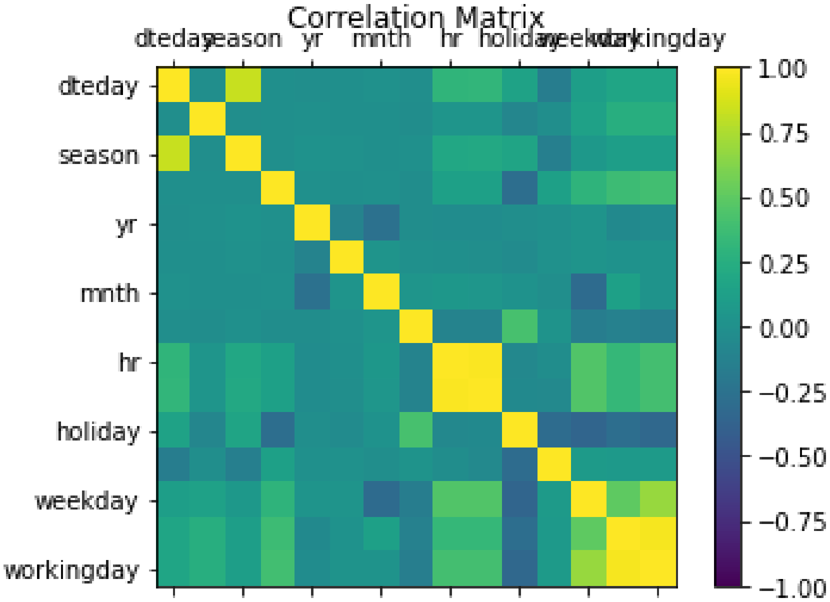


Figure 7: Correlation plot

From above plot it is obtained that the variable 'cnt' and 'temp' has positive weak correlation of 0.40. The correlation between 'cnt' and 'windspeed' has negative correlation. And the variables 'cnt' and 'holiday' has string positive correlation of around 0.80.

Regression models

To build machine learning models, a feature column and a label column is created. A feature column 'newcolumn' includes predictor variables such as season, yr, mnth, hr, holiday, weekday, workingday, weathersit, temp, atemp, hum, windspeed, casual, registered. And label column 'label' includes target variable cnt. Only the feature and label columns are selected to build regression models. The data frame consisting of feature and label is divided into training and testing data in the ratio 70:30.

Linear regression model:

The linear model is built and fitted with the training dataset with feature column and label column. In this, the maximum iteration is set to 10. The model is then evaluated with the testing dataset. The predicted values can be displayed to compare with the actual values. The Root Mean Squared Error (RMSE) and Residual squares (r2) are noted. The value of r2 shows the prediction accuracy of the model built.

Decision Tree Regression:

The decision tree regression is built and fitted with the training dataset. Both the datasets consist of feature column and label column only. The decision tree model is then evaluated with the testing dataset to test the prediction accuracy of the model. The accuracy can be determined with the r2 value.

Random Forest Regression:

The random forest regression model is built with the feature and label columns of training dataset. The number of trees set to 500. The model is then tested with the testing data. The accuracy is then determined with the r2 value.

Gradient Boosting:

The gradient boosting tree regression is built with the training dataset, with the maximum iteration of 10. The model is then tested with the testing dataset. The accuracy of the model is determined with r2 value.

Software and Hardware requirements:

Tableau is the fast growing and powerful data visualization tool, where raw data can be easily converted into readable format. Hence, this software tool is used to visualize data in the analysis phase. Python programming language is used for data analysis, instead of R programming. PySpark is a python API built to support Apache Spark. PySpark is used for everything from importing to processing the data. Apache Spark is a framework used to store the big data. It is a powerful computational storage that uses cluster computing in parallel. It performs in-memory computation which is 100 times faster than the Hadoop storage.

Amazon Web Service's (AWS) Elastic Compute Cloud (EC2) instance is created. The Ubuntu instance is created with 1GB of memory, 1 core processor and 8GB of data storage. The local system is connected to the EC2 instance with the Secure Shell Connection (SSH).

The Python, Jupyter Notebook and Spark are installed in the EC2 instance created.

EXPERIMENT RESULTS AND ANALYSIS

The plots obtained from the Tableau software in Figure 2 to Figure 6 show that the bike rental count is maximum in season 3 i.e. summer, appears to be peak in the third quarter of both years 2011 and 2012, highest in the weekday 4, larger count when there is no holiday and count decreases with increase in temperature.

The libraries used for the analysis of Bike Sharing dataset are matplotlib, MLlib. Matplotlib is used to plot the heatmap of correlation matrix. The regression models are built using the machine learning library supported by the Spark.

Below table depicts the regression model test results obtained on the training dataset.

Regression Models	**RMSE**	**R2**
Linear Regression	47.195	0.928
Decision Tree	50.652	0.917
Random Forest	54.203	0.905
Gradient Boosting	48.84	0.923

Table: 1

From the comparison of all the four models, the linear regression model gave the highest accuracy of 92.8%, followed by Gradient boosting with an accuracy of 92.3%.

CONCLUSION

The analysis clearly shows that the rental count is highly correlated with the environmental and seasonal behaviors. The prediction of the target variable is highly dependent on all the predictor variables considered. The machine learning model built using Linear regression gave the highest prediction accuracy in estimating count.

The Spark provides a very efficient machine learning library MLlib in building machine learning models. Spark performed faster in computations. Due to less availability of memory and processor in this project, the computations are harder in performing a huge dataset. This can be solved with increase in memory and processors. The accuracy of the models can further be enhanced with performing cross-validation and log transformations on the dataset.

As the bike rental count is correlated with the events in the city, event and anomaly can be easily traced with the search engine. The researchers can use this idea, where mobility acts as a sensor network. The traffic, environmental and health related issues can be addressed.

REFERENCES

[1] UCI Machine Learning Repository: Bike Sharing Dataset Data Set. Retrieved from: https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset

[2] Fanaee-T, Hadi, and Gama, Joao, 'Event labeling combining ensemble detectors and background knowledge', Progress in Artificial Intelligence (2013): pp. 1-15, Springer Berlin Heidelberg.

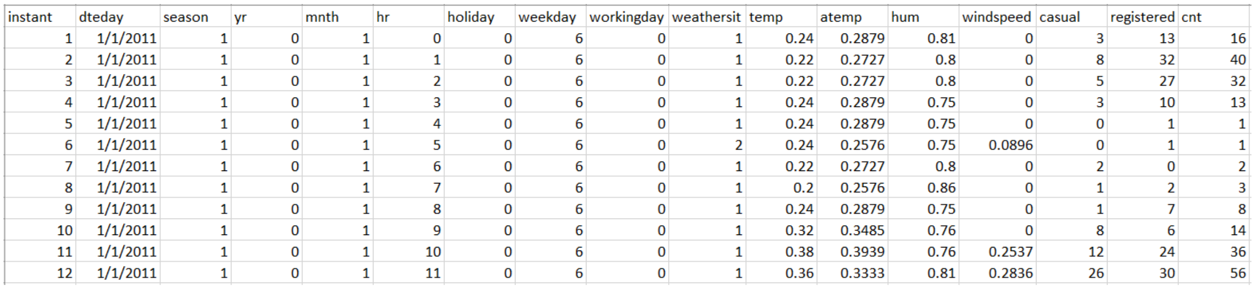
[3] HiltbrandJuly, Troy. "5 Advanced Analytics Algorithms for Your Big Data Initiatives." Transforming Data with Intelligence. Retrieved from https://tdwi.org/articles/2018/07/02/adv-all-5-algorithms-for-big-data.aspx

[4] "1.4. Support Vector Machines." (n.d). Scikit, Retrieved from https://scikit-learn.org/stable/modules/svm.html

[5] Singh, Harshdeep. "Understanding Gradient Boosting Machines." Medium, Towards Data Science, 4 Nov. 2018, Retrieved from https://towardsdatascience.com/understanding-gradient-boosting-machines-9be756fe76ab

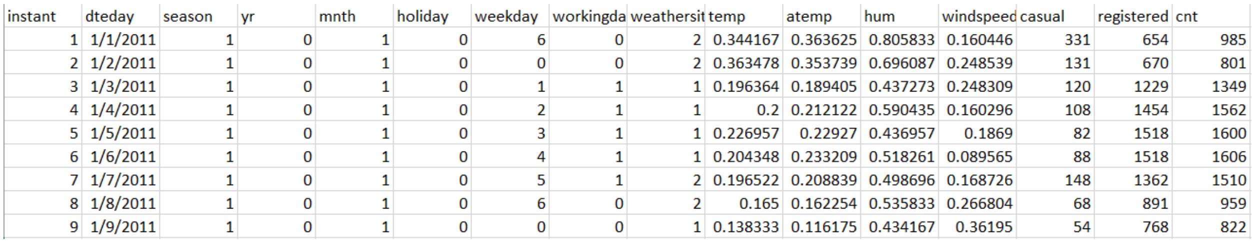
Appendix:

The dataset collected from the bike sharing data folder is shown here.



instant	dteday	season	yr	mnth	hr	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
1	1/1/2011	1	0	1	0	0	6	0	1	0.24	0.2879	0.81	0	3	13	16
2	1/1/2011	1	0	1	1	0	6	0	1	0.22	0.2727	0.8	0	8	32	40
3	1/1/2011	1	0	1	2	0	6	0	1	0.22	0.2727	0.8	0	5	27	32
4	1/1/2011	1	0	1	3	0	6	0	1	0.24	0.2879	0.75	0	3	10	13
5	1/1/2011	1	0	1	4	0	6	0	1	0.24	0.2879	0.75	0	0	1	1
6	1/1/2011	1	0	1	5	0	6	0	2	0.24	0.2576	0.75	0.0896	0	1	1
7	1/1/2011	1	0	1	6	0	6	0	1	0.22	0.2727	0.8	0	2	0	2
8	1/1/2011	1	0	1	7	0	6	0	1	0.2	0.2576	0.86	0	1	2	3
9	1/1/2011	1	0	1	8	0	6	0	1	0.24	0.2879	0.75	0	1	7	8
10	1/1/2011	1	0	1	9	0	6	0	1	0.32	0.3485	0.76	0	8	6	14
11	1/1/2011	1	0	1	10	0	6	0	1	0.38	0.3939	0.76	0.2537	12	24	36
12	1/1/2011	1	0	1	11	0	6	0	1	0.36	0.3333	0.81	0.2836	26	30	56

Figure 8: hour.csv



instant	dteday	season	yr	mnth	holiday	weekday	workingda	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
1	1/1/2011	1	0	1	0	6	0	2	0.344167	0.363625	0.805833	0.160446	331	654	985
2	1/2/2011	1	0	1	0	0	0	2	0.363478	0.353739	0.696087	0.248539	131	670	801
3	1/3/2011	1	0	1	0	1	1	1	0.196364	0.189405	0.437273	0.248309	120	1229	1349
4	1/4/2011	1	0	1	0	2	1	1	0.2	0.212122	0.590435	0.160296	108	1454	1562
5	1/5/2011	1	0	1	0	3	1	1	0.226957	0.22927	0.436957	0.1869	82	1518	1600
6	1/6/2011	1	0	1	0	4	1	1	0.204348	0.233209	0.518261	0.089565	88	1518	1606
7	1/7/2011	1	0	1	0	5	1	2	0.196522	0.208839	0.498696	0.168726	148	1362	1510
8	1/8/2011	1	0	1	0	6	0	2	0.165	0.162254	0.535833	0.266804	68	891	959
9	1/9/2011	1	0	1	0	0	0	1	0.138333	0.116175	0.434167	0.36195	54	768	822

Figure 9: day.csv