**ISE 243**

**Report**

**Supply Chain Analytics**

**[](https://www.google.com/url?sa=i&rct=j&q=&esrc=s&source=images&cd=&cad=rja&uact=8&ved=2ahUKEwjEotKp1vfdAhXmHDQIHXm7CP8QjRx6BAgBEAU&url=https://techcrunch.com/2013/07/19/san-jose-states-bold-experiment-in-online-ed-disappoints-suspends-pilot-with-udacity/&psig=AOvVaw1PiEQrTriKUo8qxKXlAOWu&ust=1539116601226785)**

**Submitted To**

**Professor Hongrui Liu**

**on**

**December 17, 2019**

**by**

**Vivek Shah**

**TABLE OF CONTENTS**

1. **Executive Summary.**
2. **Introduction**
3. **Data Understanding and Exploring**
4. **Preprocessing and Data Visualization**
5. **Training**

* **Linear Regression**
* **Holt – Winter**
* **ARIMA**
* **Neural Network**

**6. Conclusion**

**7. References**

**1. Executive Summary:**

The basic principle of Supply Chain is to help companies that face challenges when they buy, manufacture, move, sell and service products. Its is important that all these processes take place in a smooth manner with minimal cost and effort. The main objective of my project revolves around a specific task i.e. forecasting demand of bike sharing rental process. The tasks that were performed was based on the knowledge imbibed during ISE 243 course of Supply Chain Analytics. Software skills were used such as Python for coding and different algorithms like regression, holt-winter, arima and neural network were performed to check which gave the best results. Trying out different results, the Neural network model was the best in predicting the demand for bike sharing rental as it gave the least rsme value.

**2. Introduction**

Bike sharing systems are new generation of traditional bike rentals where whole process from membership, rental and return has become automatic. Through these systems, user can easily rent a bike from a position and return at another position. Currently, there are about over 500 bike-sharing programs around the world which is composed of over 500 thousand bicycles. Today, there exists great interest in these systems due to their important role in traffic, environmental and health issues.

Apart from interesting real-world applications of bike sharing systems, the characteristics of data being generated by these systems make them attractive for the research. Opposed to other transport services such as bus or subway, the duration of travel, departure and arrival position is explicitly recorded in these systems. This feature turns bike sharing system into

a virtual sensor network that can be used for sensing mobility in the city. Hence, it is expected that most of important events in the city could be detected via monitoring these data.

**3. Data Understanding and Exploring**

The bike sharing rental process compromised of 731 rows of data where each row is a day between year 2011 and year 2012. The data set determines the count of bikes rented during the day. It also consists information of date, year, month, season, weekday, working day, holiday, type of weather, temperature, windspeed, humidity, registered members, casual bike renters. With the help of these variables we can get a better understanding of the demand of bikes during a day and relate the effect of these variables on the count of bikes rented in a day. We can develop an infrastructure to provide immense understanding on bike renting patterns and the way it can in be helpful for variety of reasons.

Also, there are several other features that can improve our predictive model like the type of bike available of rental, the cost of bike rental, etc. The sole purpose of our model is to improve the predictions of bike rentals during a day.

**4. Preprocessing and Data Visualization**

In order to understand the bike rental data better and the contributions of the different variables, data preprocessing was done. Initially the data was imported using pd.read\_csv() syntax. Once the data was imported, different attributes about the data were discovered. The dimension of the data was found and different variables like season , weekday, holiday, working day, type of weather was converted into categorical data by setting bins for each column. Variables like weekday, holiday, working day were converted to binary variables in their respective columns whereas season was converted into categorical variable having four categories namely 1:springer, 2:summer, 3:fall, 4:winter.Even the type of weather was divided into four categories, any string in that column having text like {Clear, Few clouds, Partly cloudy, Partly cloudy} = 1, {Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist} = 2 , {Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds} = 3, {Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog} = 4.

Once few columns were converted into categorical variables then correlation analysis was carried out using scatter plots to see the relation of different variables . If two variables were highly correlated, then one of them was eliminated . Variables like temp : Normalized temperature in Celsius and atemp: Normalized feeling temperature were highly correlated so atemp was the only one used. For most of the models we converted it into time series data and divided the data and used 80-20 stratified split. 80% of the data was used for training and 20 was used as testing data. Many preprocessing activities were carried out on the data depending on the model.

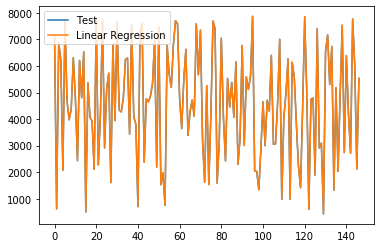
**5. Training**

A set of 5 models were built to forecast the bike sharing rental counts per day. The models built were;

* Linear Regression
* Holt – Winter
* ARIMA
* Neural Network

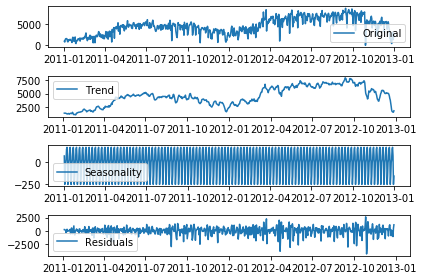
**Linear Regression**

Before performing linear regression, the data was scaled, and the output y and input variables X were separated, and the data was split into train and test set. The linear regression model was carried out and it gave a very high accuracy because it was used to predict the count of bikes rented (y) for the period where the input variables were given. This model was conducted without converting the data into time series because it is easy to know the input variables before hand as the variables included season , temperature, windspeed, etc which can be predicted for the coming days and can be used in the model. I managed to get a very high accuracy for this model which was 0.99.

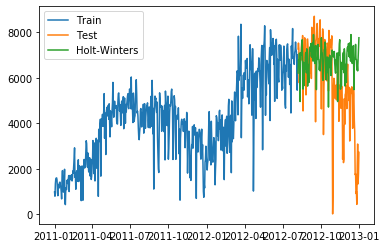


**Holt – Winter**

For Holt – Winter the data was converted into time series .Once the data was converted into time series decomposition of the data was carried out to check the trend seasonality and residual errors. The result of the decomposed data is shown below.

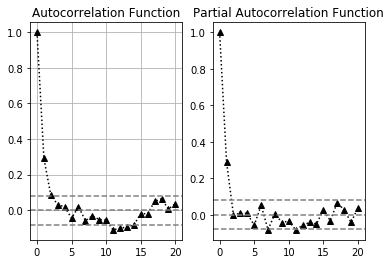


Looking at the trend, seasonality I applied the exponential smoothening model and managed to achieve a rsme value of 2182.791. The holt winter model managed to predict the test set to a large extent but could not incorporate the trend very well.



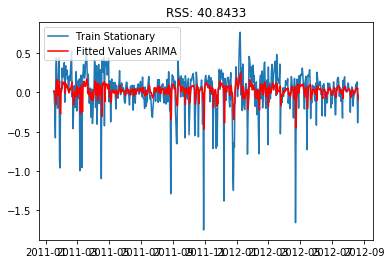
**ARIMA**

Built a univariate arima model where initially the data was converted to time series and made stationary before applying the model. In order to make it stationary the data was transformed into log to reduce the trend and then moving average with a window of 15 was applied. As the data was yet not stationary differencing of the moving average data was done to get the p value down to 2.214463538905168e-30 which is much less than 0.05 making the data stationary. For the model Partial Autocorrelation and Autocorrelation function was determined by plotting it ( p = 2, q = 2)



The ARIMA model of order (2, 0, 2) was applied to predict the test set results and the log of original test set and predicted test set were plotted to get a rsme value of 0.549059.

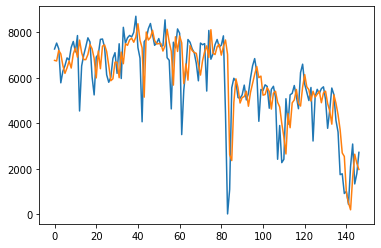
Below is the plot of stationary trained set with the fitted ARIMA set.



The reason the rsme value is so low because the data is stationary and in its log form. When the data was converted into its original form a very high rsme value was achieved.

**Neural Network**

Built a univariate and multivariate neural network model to forecast the count of bikes. As the data was not stationary it was converted into stationary. This was done because stationary data is easier to model and will very likely result in more skillful forecasts. Then created a differenced series and then divide data into input (X) and output (y) components with time steps. Also, created a function to transform the differenced data back to its original form. The data was then spit into train and test sets. The activation function used for this data was hyperbolic tangent (tanh) as it gives the output between 1 and -1, which is a preferred output for time series data. Again, it is inverted so the values are back to its original form. In order to fit the univariate model, we used the neurons to be 4 and a batch size of 1 as it yields the best result. On plotting the predicted and original data for a univariate model it gave the best result compared to others any other model and it gave an RSME value of 1247.02 which the best between holt winter, arima and neural network.



A multivariate model was also attempted but the result obtain was not optimal due to constraint of time.

**Conclusion :**

The conducted analysis proved that according to linear regression where the time series data is not considered. Independent variables are used to predict the dependent variable and give a very high level of accuracy . So, the count of bike rented each day depends many of the features affecting it . From the time series data , the model that best represented the data or rather predicted the data was the neural network model and that can be seen in the graph where predicted, and original data are plotted. As a result, though the neural network model was the best, but it was computationally very time consuming.

**References**

[1] 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, doi:10.1007/s13748-013-0040-3.

@article{year={2013}, issn={2192-6352}, journal={Progress in Artificial Intelligence}, doi={10.1007/s13748-013-0040-3}, title={Event labeling combining ensemble detectors and background knowledge}, url={http://dx.doi.org/10.1007/s13748-013-0040-3}, publisher={Springer Berlin Heidelberg}, keywords={Event labeling; Event detection; Ensemble learning; Background knowledge}, author={Fanaee-T, Hadi and Gama, Joao}, pages={1-15} }