Evaluation of Machine Learning Models for Forecasting Water Consumption: A Case Study in New York City

Abstract.

To research forecasting water consumption in New York City, this paper uses daily water consumption (based on million gallons per day) from 1979 to 2017 as sample data, and so there are 39 observations in this data set. Firstly, the paper applies two machine learning models, Simple Linear Regression model and Polynomial Regression model using Python. In both models, the water consumption is predicted by the feature “Year”. As a result, the evaluation of these two models shows that the polynomial regression model is more accurate on this data set and fits more better on real data values. Secondly, in order to further explaining, the Multiple Linear Regression is examined considering two independent features, “Year” and “Population”. This model acts well on real test values too. Lastly, Hierarchical Clustering is quoted in this research. Regarding five clusters, the proportion of each year in water consumption is analysed.

Keywords: Forecasting, Linear Regression, Water Consumption, Machine Learning Models.

Introduction.

Since water is a vital resource for human life, economic development and environmental quality, there are a great number of reaserches who pay attention on it. Over the past three decades, many countries have made major strides in management of water resources. Water consumption is one can be challenging to rectify. There are a grand attention to foreacasting water demand (see [1, 2, 3, 5, 6, 8]). Today, most countries are placing unprecedented pressure on water resources. The global population is growing fast, and estimates show that with current practices, the world will face a shortfall between forecast demand and available supply of water by 2030. Furthermore, water scarcity and extreme water events (floods and droughts) are perceived as some of the biggest threats to global prosperity and stability. Feeding 9 billion people by 2050 will require a 60% increase in agricultural production, and a 15% increase in water withdrawals. Estimates indicate that by 2025, about 1.8 billion people will be living in regions or countries with absolute water scarcity.

Due to NYC Dept. of Environmental Protection, in 1979, the New York City consumed over 1.5 billion gallons of water each day. Since then, the total daily consumption has dropped by about a third. In 1019, the daily total decreased to below a billion gallons each day. This includes all water used by households, businesses, industry, and lost through leaks in the water supply system. This decrease is rather surprising, because NYC[1] has experienced steady population growth throughout the same time period. NYC’s population has increased by over a million people since 1980. Only in the last few years, since 2016, has population growth begun to level off and slightly decline. Even though there are more people, less water is consumed per person. On a per capita basis, New Yorkers consume about 45% less water

[1] New York City

today than in 1980. What has caused this decrease in NYC's total water demand? A combination of successful government initiatives deserve credit for the decreased water demand, including metering, regulation, and conservation programs. Most of the efforts began in the mid-1980s, spurred by significant droughts in 1980, 1985, and 1989 that critically threatened NYC's water supply. In order to appreciate the reasoning behind the city's different initiatives, it's helpful to first understand the breakdown of how water is used in NYC: The largest demand segment is residential and mixed-use residential. Next, about a fifth of the city's total water demand goes to "non-revenue" usage. Non-revenue water is water that is produced by the water supply system but never reaches a customer due to leaks and other losses.

The used Dataset is provided from the last updated information https://data.cityofnewyork.us/data.json of data.world dated on 2020.05.22. Dataset shows water consumption in NewYork city based on it's population during 1979 to 2017.

Section two begins with an explanation of the dataset. As some of reasons have stated above, Fig. 1 and Fig. 2 emphasis a true management in water consumption. We are going to predict water consumption in NYC for the future years. The applied models are "Simple Linear Regression" and "Polynomial Linear regression" which have done all in Python. As a result, the polynomial linear regression fit better on this dataset.

In [2], Calvo et al. evaluated the performance of multiple linear regression and feed forward computational neural networks (CCNs). We have tried the multiple linear regression model. This is performed based on "Year" and "Population" as two independent features. Numbers state that in most cases of this model on the real test values, multiple linear regression acts accurately.

To reach more elaboration for water demend, we have applied Hierarchical Clustering with five clusters. This clusters have made a clarity in proportion of each year in consumption of this vital resource.

The last but not least, conclusion of this paper can help goverments and people make a strategy in order to saving this vital resource, because this paper finds that water consumption is affected by various reasons and its amount of consumption, is controlable by organised programms.

Results

Dataset Explanation

Fig. 1 and Fig.2 show the plots of population and water consumption both throughout the thirty nine years respectively. While population has increasing trend in this period of time, water consumption is reversely decreasing. It should be noted that this descending trend of water demand is at the first years almost volatile, but a constant descending trend can be has seen after around 1990.

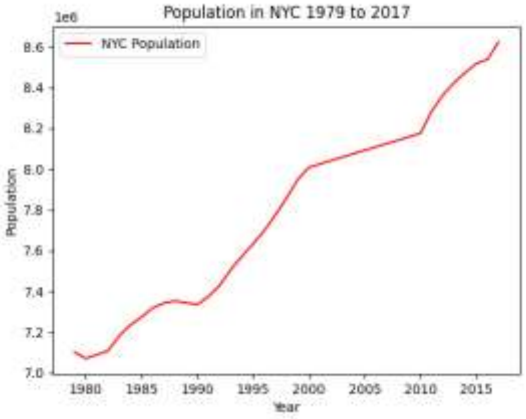

Fig. 1

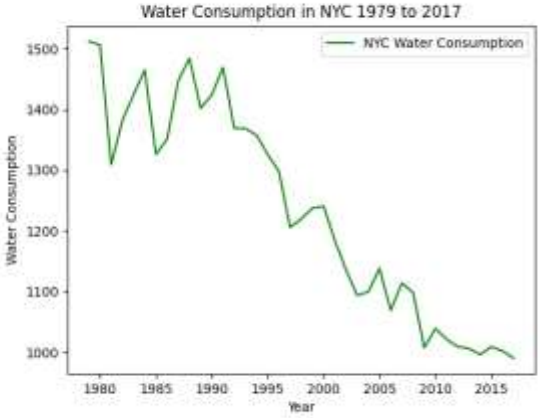

Fig. 2

But how the paper can access to an appropriate prediction for water consumption at the future years. This paper examines some Machine Learning Models to answer this question. Firstly, Simple Linear Regression and Polynomial Regression models have applied.

Simple Linear Regression Model.

Here we allocated 1/3 of whole dataset to the test set. As a result of visualising the Training and Test set, we have the following:

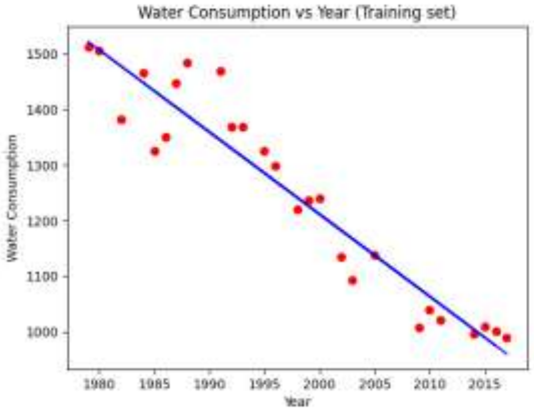

Fig. 3

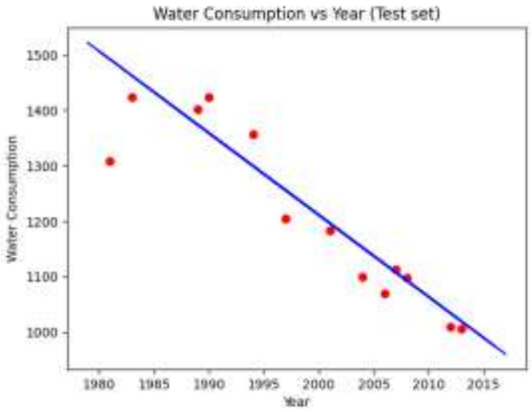

Fig. 4

Now, we are going to get the final linear regression equation with the value of coefficients:

$$y = b_0 + b_1 * x$$

$$\Downarrow$$

$$water\_consumption = b_0 + b_1 * Year$$

$$water\_consumption = 4013.12 - 0.0003556 * Year$$

Ordinary Least Squares, $Sum\ (y-\hat{y})^2\ \rightarrow min$, is the best fitting line.

In this case, we can make a single prediction. For instance, water consumption when the Year is 2020, is equal to 916.267 (Million gallons per day).

Polynomial Linear Regression Model.

The polynomial linear regression model is as

$$y = b_0 + b_1 x_1 + b_2 {x_1}^2 + \cdots + b_n {x_1}^n.$$

When we say linear and non-linear, we are not talking about x variable and it's about coefficients b_i. In order to obtain the best polynomial regression model, we have tried different degrees of n on our training model on the whole dataset. One can supposed to be as best for our model is degree=6.

Fig. 5 shows the visualisation of the simple linear regression model. The red points are the real value for water consumption and the blue line expresses the simple linear regression model. First of all, we can see that indeed the linear regression model is not well-adapted to this data set. Because for many values of feature "Year", the prediction level is far and even for some of values, the prediction level is super far.

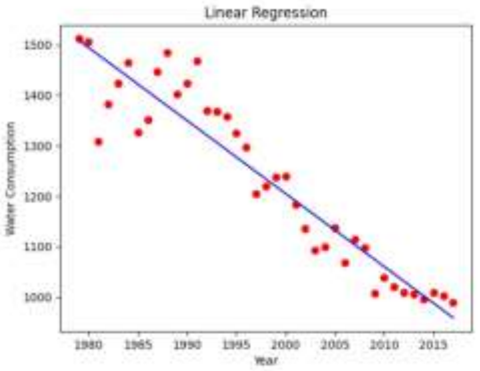

Fig. 5

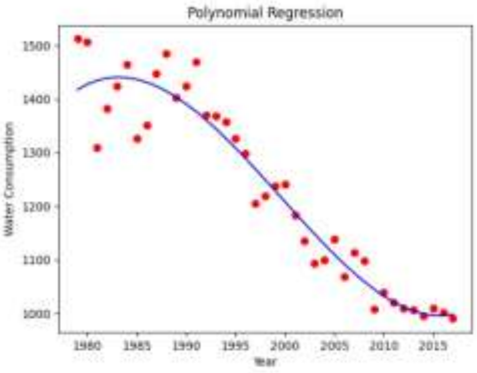

Fig. 6

So as the simple regression model does not work efficient on this data set, we will apply Polynomial regression model. According to Fig. 6, we have indeed a well more adapted regression curve which is much closer to the real results. Hence, we get a better result applying Polynomial regression model with degree six on this data set. We have checked these two models for the year 2010. The predicted values by simple linear regression model and polynomial linear regression model are 1060 and 1032 respectively. While the real value was 1039 in the year of 2010. Thus it can be obviously have seen that the polynomial regression model is more accurately works here and is super close to consumption amount that mentioned in real. But the prediction does not work true in simple linear regression model as we can see it in above in Fig. 5 too.

Now if we check the water consumption in the year 2021 with the Simple linear regression, we get 901.529 (million gallon per day) as a result. But the predicted value using the polynomial regression model is equal to 1035. 178.

Multiple Linear Regression Model.

We expand our simple linear regression to multiple linear regression with two independent variables which are "Year" and "Population". The applied model is

$$y = b_0 + b_1 x_1 + b_2 x_2$$

We have allocated 0.2 of the whole dataset to the test set result. Hence the prediction results on this proportion is as follows.

Prediction result	Real Test-value
197.19	198.
137.89	137.
135.43	135.
119.6	121.
118.01	119.
137.17	136.
190.89	191.
149.29	148.

table. 1

Hierarchical Clustering

The following figure represents the proportion of each year in water consumption. Although, there is volatility during this period of time, the general trend in the last years is descending. Here, the role of each year in the obtained Fig. 7 has showed as [1 1 0 0 0 1 0 0 1 1 0 0 1 0 0 0 0 0 4 4 4 4 4 3 3 3 3 2 3 3 2 2 2 2 2 2 2 2 2]. In fact, in year 2006 and during 2009-2017, the water consumption has it's least amount of owns. Where as, the most water consumption in this period of time occurs in years 1981, 1982, 1983, 1985, 1986 and 1989. Moreover, Fig. 7 shows that the least water consumptions are in cluster 3 and the most consumption are in cluster 2.

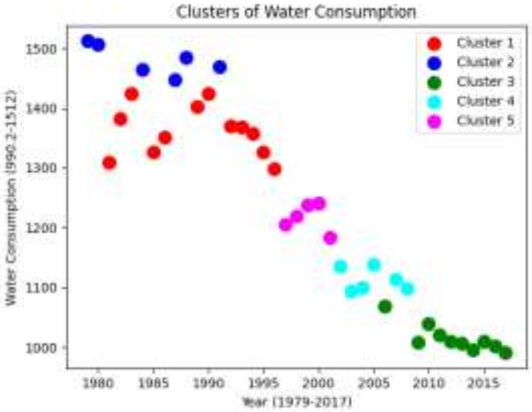

Fig. 7

For future Works, our models can be improved using more features and some other machine learning models such as non-linear models. Neural approach is as an instrument may improve the result of prediction.

References

1. M. Bata et al, Short-term water demand forecasting using hybrid supervised and unsupervised machine learning model, Smart Water, https://doi.org/10.1186/s40713-020-00020-y, 2020.
2. Calvo et al., Linear regression and neural approaches to water demand forecasting in irrigation districts with telemetry systems. Biosystems Eng.97, 283-293, 2007.
3. D. Kofinas, N. Mellios, E. Papageorgiou and C. Laspido, Urban Water Demand Forecasting for the Island of Skiathos, 16th Conference on Water Distribution System Analysis, No. 89, 1023-1030, 2014.
4. F. Nielsen, Introduction to HPC with MPI for Data Science, 1st ed, 2016.
5. Danielle C. M. Ristow, Elisa Henning, Andreza Kalbusch and Cesar E. Petersen, Models for forecasting water demand using time series analysis: a case study in Southern Brazil, Journal of Water, Sanitation and Hygiene for Development, 11, 2, 231-240, 2021.
6. M. Sakizadeh, Comparision of Time Series Forecasting Techniques Applied for Water Quality Prediction in Southwest Iran, Tarbiat Modares University Press, Vol. 8, Issue. 4, 199-208, 2020.
7. Valdimir N. Vapnik, The Nature of Statistical Learning Theory, Springer, 1992.
8. A. Yaser, M. Biligili and E. Simsek, Water Demand Forecasting Based on Stepwise Multiple Nonlinear Regression Analysis, Arab J Sci Eng, Vol. 37, No. 8, 2012.