Analysis and Prediction of Produced Water Data using Machine Learning Algorithms

Raiel Amesquita[1,2]; Bibian Ogbuji, PhD[1,2,3]
[1]Texas Water and Energy Institute, [2]Department of Mechanical Engineering, [3]University of Texas Permian Basin

THE UNIVERSITY OF TEXAS PERMIAN BASIN
COLLEGE OF ENGINEERING

Abstract

The distinguished features of machine learning (ML) based modeling provides a deep insight pertaining to the prediction of total dissolved solids (TDS) in produced water (PW). The current study investigates the predictive performance of Linear Regression (LR) model and Time Series Forecasting (TSF) for modeling yearly TDS. The USGS and CLIENT datasets were used for the models training and testing. The results were evaluated using various performance and accuracy indicators such as RMSE, MAE, and MSE. The model outcome indicated that both Linear Regression and Time Series Forecasting are reliable techniques in predicting TDS.

Method

Model Building Process:

1. Import the Libraries and Dataset
2. Data Preprocessing (Imputation, Feature Selection)
3. Split the Dataset
   a. Split the dataset into a 70:30 % ratio
   b. Training set: a subset of the entire data 70% (TDS, Na, Ca, and Cl)
   c. Testing set: a subset of the whole data 30% (TDS, Na, Ca, and Cl)
4. Build the Model
   a. Identify and set the dependent variable (y) in the model (TDS)
   b. Identify the independent variables (X) in the model (Ca, Cl, Na)
   c. Fit the training set to the model
   d. Fit the model in the test set to predict TDS values
5. Evaluate the Performance and Accuracy of the Model
6. Parameter Tuning
7. Visualize the actual test data and predicted data

Data

- The Permian Basin Data from USGS database consists of 46,268 data points, containing information on sample location, pH, resistivity, conductivity, and various constituents such as TDS, cations, and anions for produced water from 58,706 oil and gas wells.
- The CLIENT database is a compilation of produced water geochemical analysis (WA) reports from 2003 to 2018 collated from 45 oil and gas companies consisting of 34,924 data points.
- **Figure(1)** displays the statistical parameters of each variables used in the model. The parameters include mean, standard deviation, min value, and max value, etc.

	TDS	Ca	Cl	Na
count	11565.000000	11561.000000	11565.000000	11565.000000
mean	105454.754920	5463.751275	63058.282662	35548.294837
std	78148.680001	6393.116488	48669.534131	41663.171128
min	666.000000	1.010000	8.000000	0.000000
25%	44335.000000	1460.000000	24760.000000	15225.000000
50%	87307.000000	2966.000000	51644.020000	27152.610000
75%	154778.000000	7323.000000	94300.000000	45775.470000
max	397572.000000	66381.000000	245700.000000	333752.570000

Figure(1) Statistical Parameters

Models and Accuracy Test

Models:

1. **Linear Regression:** The goal is to predict a dependent variable (y) based on a given independent variable (X). The USGS dataset was used.

Y: Dependent Variable
A: Population Y Intercept
Beta1: Population Slope Coefficient
X1: Independent Variable

$$Y = a + \beta_1 X_1 + \beta_2 X_2 + \cdots\cdots + \beta_i X_i$$

Equation (1) Mathematical form of Linear Regression

The best performance and accuracy test for the Linear Regression model include:
Mean Absolute Error (MAE): A measure of errors between paired observations, in this study it is the difference between the predicted value and the actual value.

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - y_j|$$

Equation (2) Mathematical form of MAE

Root Mean Squared Error (RMSE): Is the standard deviation of the residuals. Residuals are a measure of how far the actual values are from the regression values.

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

Equation (3) Mathematical form of RMSE

2. **Time Series Forecasting:** Will make predictions and forecast predicted values based on historical time stamped data. The model chosen for this study to forecast TDS is Autoregressive Integrated Movement Average (ARIMA). Each component in the ARIMA model is specified as a different parameter in the model which are (p, d, q). The CLIENT dataset was used.

AR: Autoregression uses the dependent relationship between observation and some number of lagged observations. **(p)** is the amount of lag observations included in the model.
I: Integrated is the use of differencing of raw observations to make the time series stationary. **(d)** is the number of times that the raw observations are differenced.
MA: Moving Average uses the dependency between an observations and a residual error from a moving average model applied to a lagged observation. **(q)** is the size of the moving average window

$$\hat{y_t} = \mu + \sum_{i=1}^{p}\alpha_i y_{t-i} + \sum_{i=1}^{q}\theta_j \epsilon_{t-j}$$

Equation (4) Mathematical form of ARIMA

The best performance and accuracy test for Time Series Forecasting model are:
Mean Forecast Error: Is the calculated mean of difference between an observed values (X) and the predicted values (y) at a specific time-period (t).

$$\frac{1}{N}\sum_{t=1}^{N}(Actual_t - Forecast_t)^2$$

Equation (5) Mathematical form of Mean Forecast Error

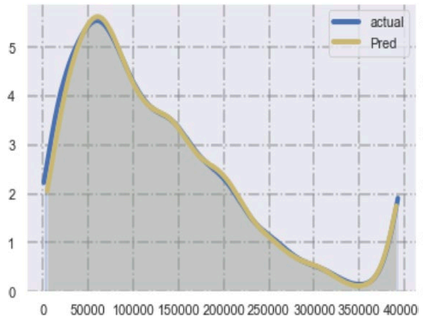
Weighted Mean Absolute Percentage Error (wMAPE): Gives a percentage score of how forecasts deviate from the actual values. The weighted overcomes the infinite error issue of the standard MAPE method.

$$\frac{\sum_{t=1}^{N} ABS(Actual_t - Forecast_t)}{\sum_{t=1}^{N} Actual_t} * 100\%$$

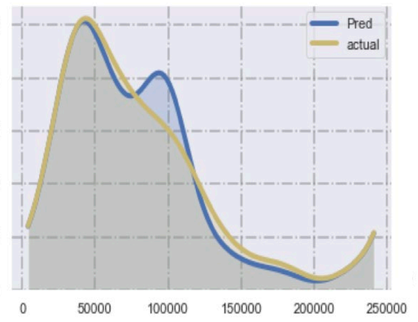
Equation (6) Mathematical form of Weighted Mean Absolute Percentage Error

Visualization of Models

Time Series



Linear Regression



	Pred	actual
0	34037	56153
1	20022	33553
2	21023	36421
3	5006	9583
4	41045	67177
5	38042	62158
6	39043	64087
7	40044	65356

	Pred	actual
count	291.000000	291.000000
mean	75526.951890	76635.426117
std	45981.816287	47724.255843
min	4041.000000	4041.000000
25%	40143.000000	39936.500000
50%	67100.000000	67534.000000
75%	98465.000000	100728.000000
max	241075.000000	241075.000000

Figure (2) Data Table and Distribution of Time Series Forecasting

	actual	Pred
0	33533.0	34970.496502
1	131000.0	130763.804246
2	106359.0	108143.499638
3	77677.0	80020.669521
4	216538.0	214453.109297
5	1764.0	6217.801097
6	271000.0	316128.044534
7	14149.0	15242.435767

	actual	Pred
count	2314.000000	2314.000000
mean	108912.346327	108849.765825
std	78540.259129	77895.393007
min	1091.000000	4907.659758
25%	47492.750000	47697.644026
50%	91948.000000	91492.376809
75%	158975.500000	158395.502272
max	392600.000000	390765.331412

Figure (3) Data Table and Distribution of Linear Regression

Discussion and Results

Linear Regression: The results for Linear Regression are presented in visualization in **Figure 2** for predicting TDS in the USGS dataset with a satisfactory estimated output. The **MAE** and **RMSE** values were found to be (0.541) and (0.3) for the predicted values of TDS (y_pred) and the testing values of TDS (y_test) respectively. The visualization results shows the declining accuracy drawback of regression-based modeling techniques. . The values of MAE and RMSE range from zero to infinity and an acceptable value will be closer to zero and less than one.

Time Series Forecasting: The results of the ARIMA algorithm are presented in the visualization in **Figure (3)** for predicting the forecast of TDS in the CLIENT dataset with a satisfactory estimated output. The **MFE** and **wMAPE** values were found to be (0.127) and (1.4%) for the predicted values of TDS (y_pred) and the testing values of TDS (y_test) respectively. The visualization results show how the data is sampled by year and how the future years are sampled and forecasted.

Conclusion

Linear Regression and Time Series Forecasting models are presented in this study to predict future values of TDS in PW using the values of Ca, Cl, and Na. These models have the capability to predict TDS values given the set of constituents. Moreover, the performance of Time Series Forecasting model turned out to be the most accurate due to the additional feature variables and parameter tuning that provides relevant information leading to an increase in forecasting precision. The modeling techniques applied in this study could assist engineers in developing an effective strategy for effective management, treatment and designing of treatment technology.

References

1.) Faraj, F., & Shen, H. (n.d.). *Forecasting the Environmental Parameters of Water Resources using Machine Learning Methods* . Retrieved November 29, 2021, from https://www.researchgate.net/profile/Farshid-Faraji-3/publication/328671400_Forecasting_the_Environmental_Parameters_of_Water_Resources_Using_Machine_Learning_Methods/links/5be98c92a6fdcc3a8dd0e32a/Forecasting-the-Environmental-Parameters-of-Water-Resources-Using-Machine-Learning-Methods.pdf.
2.) Cheng, C., Sa-Ngasaongsmg, A., Beyca, O., Le, T., Yang, H., Kung, Z., & T.S. Bukkapatnam, S. (n.d.). *Time series forecasting for nonlinear and non-stationary processes: A review and comparative study*. Taylor & Francis. Retrieved November 29, 2021, from https://www.tandfonline.com/doi/abs/10.1080/0740817X.2014.999180.
3.) Su, X., Yan, X., & Tsai, C.-L. (2012, February 10). *Linear regression*. Wiley Interdisciplinary Reviews. Retrieved November 29, 2021, from https://wires.onlinelibrary.wiley.com/doi/full/10.1002/wics.1198.