ADVANCED DATA SCIENCE IBM CAPSTONE PROJECT

Muhammad Abdur Rehman Khan

BUSINESS PROBLEM

Case

Energy Consumption in Netherland, an open case

Context

In Netherland, the three companies are responsible for providing energy in terms of electricity and gas to the whole country. The three companies are Enexis, Liander and Stedin. Every company has shared data of ten years from 2009 to 2018. The data is splitted into two categories, Electricity and Gas. Each year data is comprised of twelve features. Among features, five are providing statistical information.

Objective

Total Energy consumption in 2019

ARCHITECTURAL DECISION

Architectural Choices

IBM & Kaggle Cloud

Data Source & Initial Data Exploration

The three companies data are available in csv format. There are almost twenty csv files of each company, showing electricity and gas data.

Extract Transform and Load, ETL

- ETL is an import tool to convert data into readable format. There are three phases. In first phase, we will extract each company's data. In other two phases, we will first merge the data according to our requirements and then load the data frame. We will make three data frames, one for electricity data, the other for gas data and the last one consisting of all the data.

ARCHITECTURAL DECISION

Feature Creation

We have tried two different methods to design our model to predict Annual Consumption, the target variable. First, we have selected our features, ‘delivery_perc’, ‘annual_consume_lowtarif_perc’, smartmeter_perc’ and ‘type_connn_perc’, and correlated with the target variable ‘annual_consume’.

Correlation Information

The correlation is very weak among the features, so it won't work to train the model.

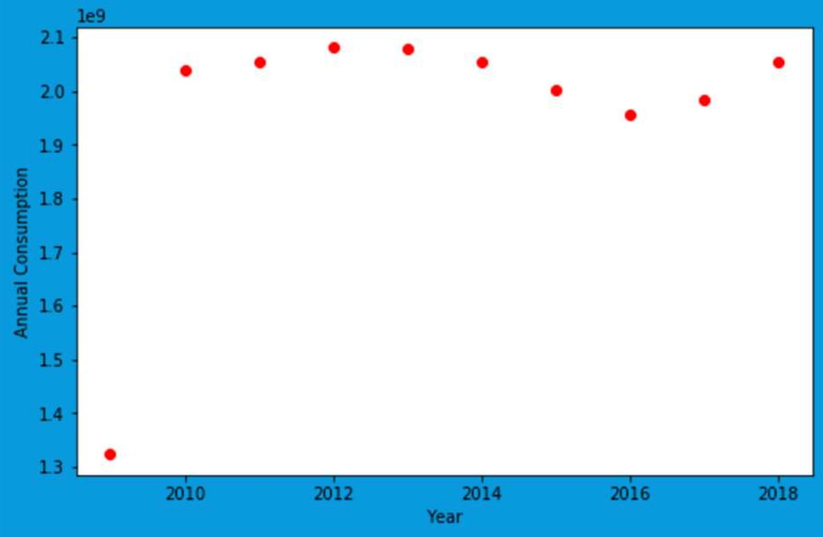
Feature Engineering

Now its time for feature engineering. We have created a new feature called ‘year’ against the ‘annual_consume’ This created variable will give every year information which will be suitable for model designing. Now we will design our model against this feature.

MODEL

Model Definition

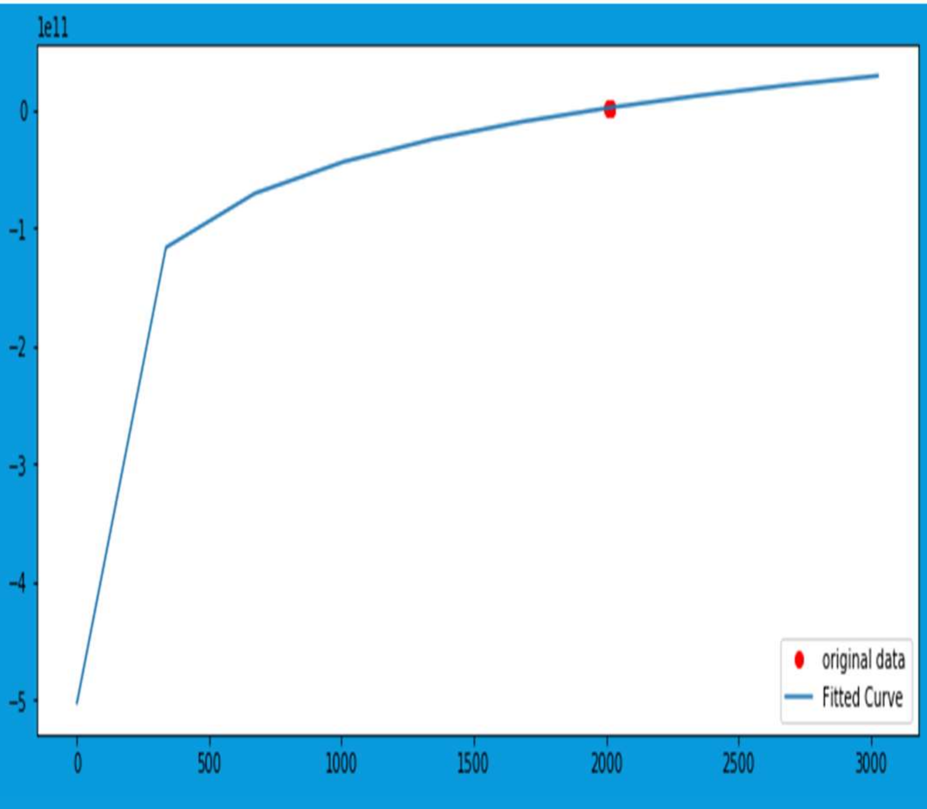
After visualizing the feature ‘year’ against the predicting variable. It is obvious that the trend is non linear and the shape is Logarithmic. We have to go for machine learning nonlinear regression algorithm.



MODEL

Model Algorithm

As the data shape is nonlinear and the trend is showing Logarithmic. We have created a function with two parameters for controlling the curve steepness and slides the curve on the x-axis. Through SciPy optimize function curve fit, we found the value of the parameters. From the image , the fitted curve response is excellent.



MODEL

Model Training

After building the algorithm , we train the model with the training set and the testing is done with the test set. Normalization is must before training the data, otherwise the evaluation will be not be appropriate.

MODEL EVALUATION

We have applied different tests to calculate Mean Absolute Error (MAE) and Mean Squared Error for error (MSE). For accuracy, we have applied coefficient of determination, R2-score.

The results are amazing. R2 score is 1, while MAE and MSE is zero.