PUN Predictive Analysis: A Comprehensive Statistical Exploration

Vittorio Balestrieri

NUS Consulting

Introduction

This statistical analysis aims to explore and understand the dynamics of key variables through various data sources. The data used in this analysis includes several relevant variables in the context of energy and financial analysis. The main variables considered are as follows:

1. **PUN (Single National Price):** PUN represents the price of electricity in Italy. This parameter is of fundamental importance for understanding the dynamics of the energy market and can influence various aspects of the analysis.
2. **CBOE Crude Oil Volatility Index:** The Crude Oil Volatility Index (Oil VIX) is an indicator of the implied volatility of oil prices. This parameter is relevant for assessing the stability and variability of oil prices, which can impact energy costs and the economy in general.
3. **CBOE Emerging Markets ETF Volatility:** This volatility index represents volatility in emerging markets. Volatility in emerging markets can provide insight into global economic trends and market interconnections.
4. **Actual Generation Quantity Split by Source:** This variable represents the actual quantity of energy generated, divided by source. Sources include, but are not limited to, hydropower, photovoltaic, and other renewable sources. Exploring energy generation from different sources can provide information on the energy mix and environmental impact.

ANALYSIS PERIOD: From 01-01-2019 to 12-31-2023

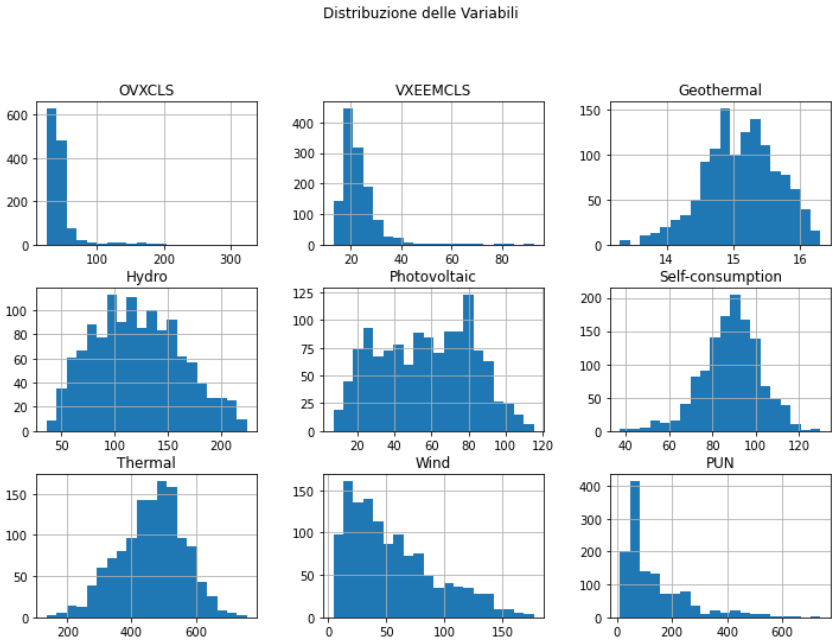

Figure 1: Distribution of Variables

Methodologies Used in the Analysis

1. **Multiple Regression:** Multiple regression is a statistical technique used to analyze the relationship between a dependent variable and two or more independent variables. This method is useful for accurately modeling complex relationships between variables.

   Results:

   Being the simplest model, it is not sufficiently performant to capture a likely non-linear dependency among the considered variables. In the following image (Figure 2), the lack of predictive capability of the model can be observed.

   Mean Squared Error: 7793.27

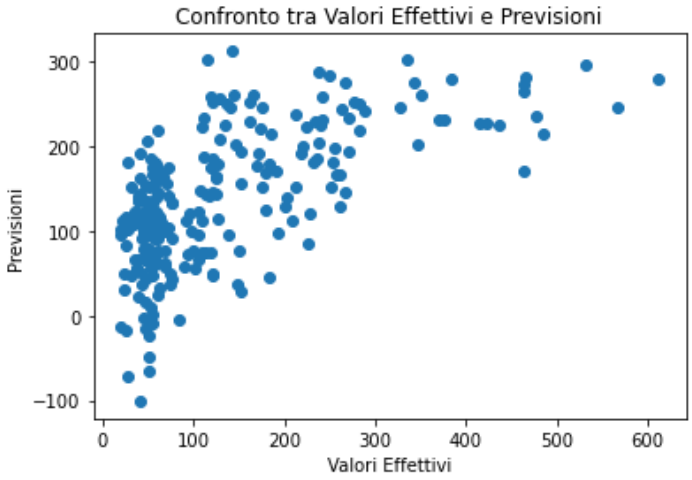

Figure 2: Comparison between Actual Values and Predictions

2. **Correlation Matrix:** The correlation matrix is a table that displays correlation coefficients between multiple variables. This tool helps identify linear relationships between variables and assess the strength and direction of these relationships.

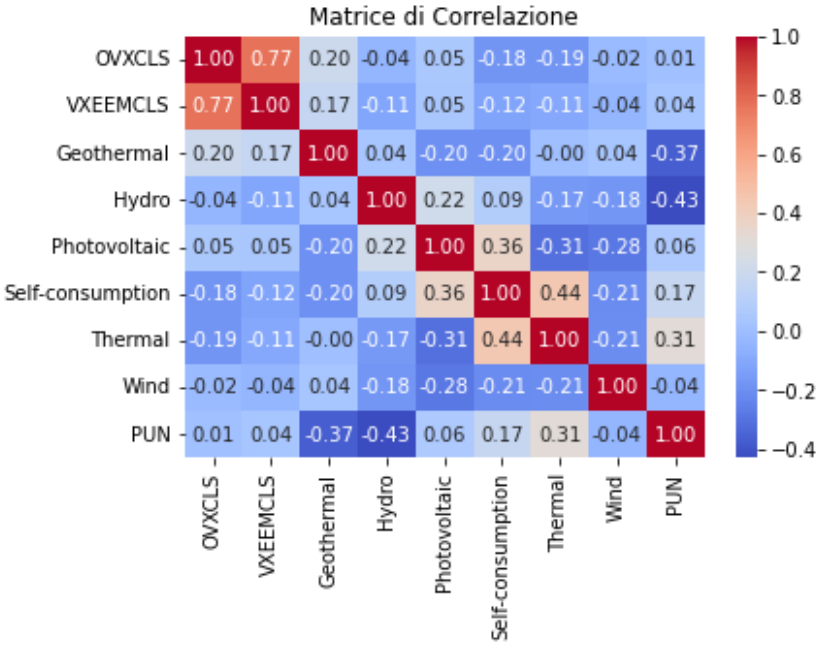

Figure 3: Correlation Matrix

3. **Feature Importance:** Feature Importance identifies which variables contribute most to the model's prediction. This can be useful for selecting the most relevant features and identifying which of the examined variables contributes most to explaining the variability of PUN.

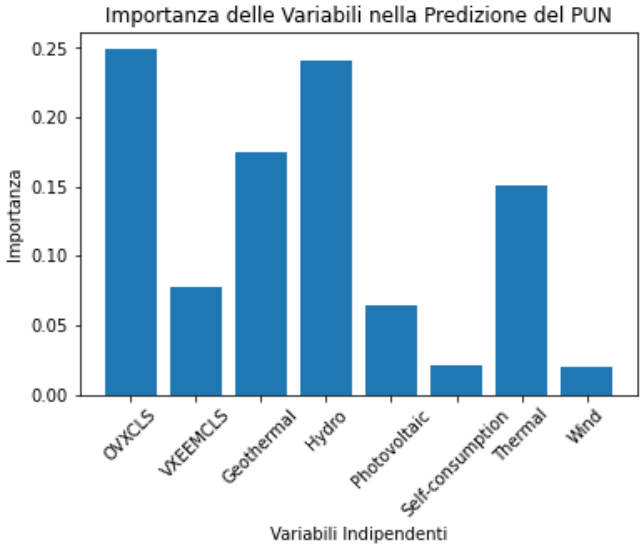

Figure 4: Feature Importance

4. **Multi-Layer Perceptron Regressor:** The Multi-Layer Perceptron (MLP) is an artificial neural network with one or more hidden layers. The MLP Regressor is employed for regression problems, learning complex representations of relationships in the data. If the relationship between independent and dependent variables is nonlinear or complex, an MLP may be able to capture more intricate patterns than a simple linear regression.

   Results: The model performs better than a simple regression model, especially regarding its ability to capture the variability of PUN (see R-squared value). However, regarding the predictive ability of the model, we are still far from having a model capable of predicting the trend of PUN (see Mean Squared Error and Residual Plot).

   Mean Squared Error (MLP): 5781.037

   R-squared (MLP): 0.565

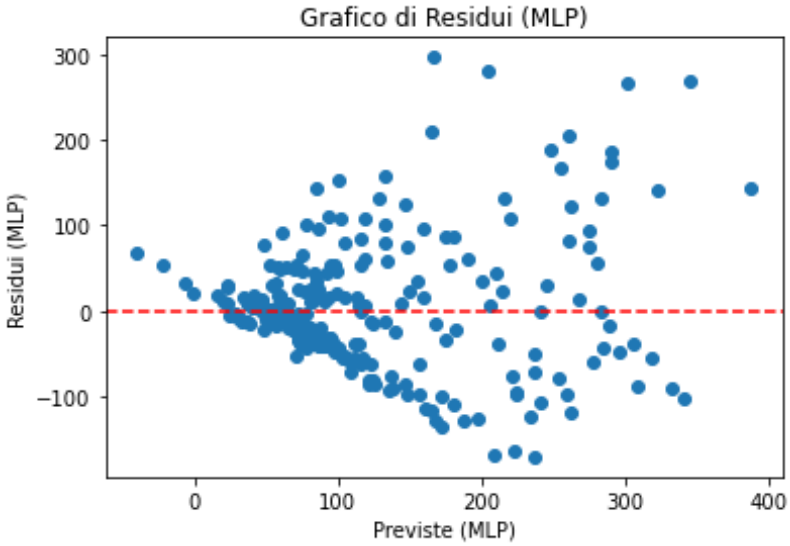

Figure 5: MLP Residual Plot

5. **XGBoost:** XGBoost is a gradient boosting-based machine learning algorithm. This algorithm is effective in regression and classification problems, combining multiple weak models to obtain a robust model. XGBoost can handle complex relationships between input variables and adapt to nonlinear patterns in the data. Its ability to build complex decision trees makes it suitable for problems where the relationship between variables cannot be effectively modeled by linear models or simpler algorithms.

   Results:

   As expected, a model capable of handling nonlinear relationships performs better than "traditional" regression models. As evident from the values below and the plot, the model exhibits significant ability to follow the general pattern of the variable of interest. However, observing the plot and the R-squared value, we can notice a consistent difficulty in capturing the structural variability of PUN.

   Although this model is much more informative compared to those previously analyzed, it exhibits a structural flaw that prevents it from being used for predictive analysis.

   The model remains relevant, especially in the event of obtaining more data for the analyzed variables, as it could help the model converge towards a solution more conducive to predictive activity.

   Mean Squared Error (XGBoost): 512.111

   RMSE: 22.630

   R^2: 0.185

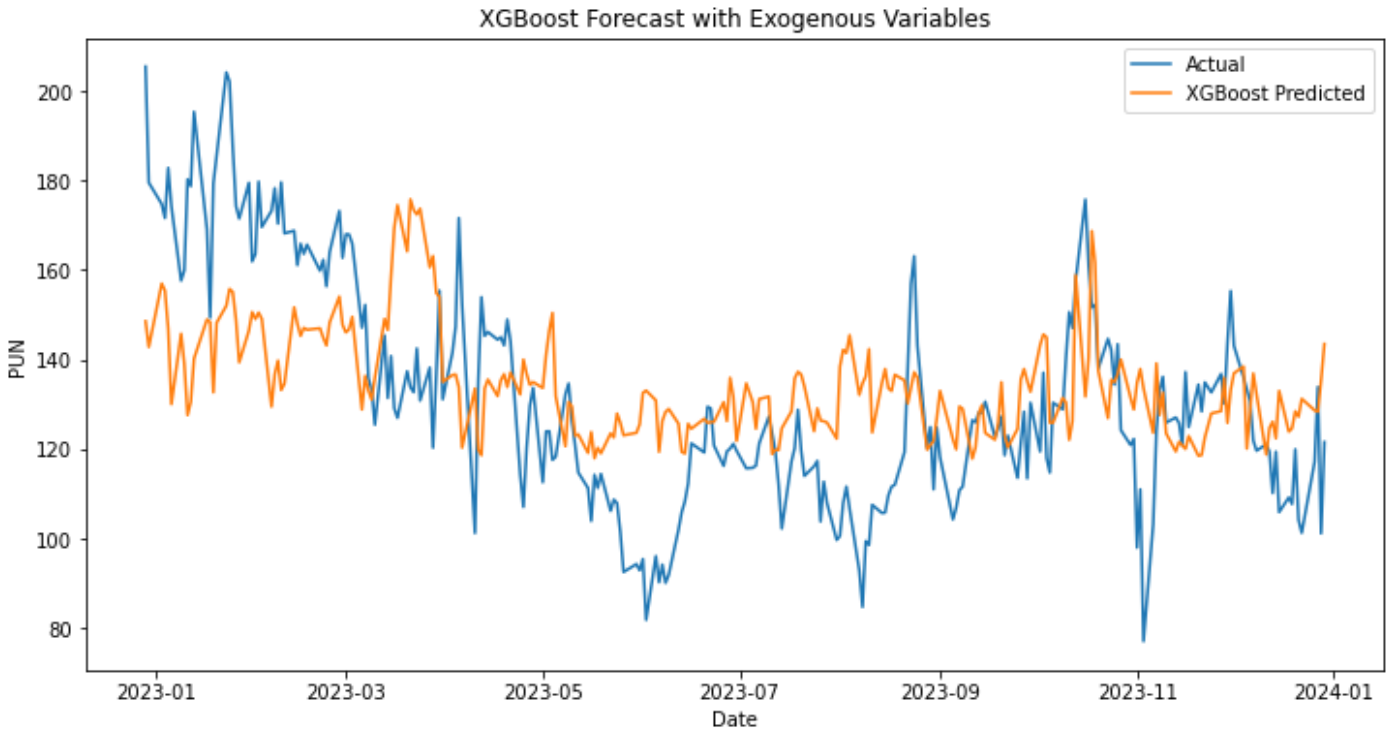

Figure 6: XGBoost Forecast Model

6. **LSTM (Long Short-Term Memory):** LSTM is a type of recurrent neural network (RNN) designed to handle long-term dependencies. LSTMs are designed to manage time sequences by maintaining long-term memory of information. This feature makes them suitable for problems where the time sequence of data is critical, such as in the case of time series.
   LSTM models can handle complex and nonlinear data, adapting well to more intricate relationships in the data. For time series regression problems, where the goal is to predict a future value based on the data's history, LSTMs can offer remarkable performance.
   Additionally, they exhibit two critical features in our case:

   1. LSTMs are designed to prevent the "vanishing gradient" problem, allowing them to preserve important long-term information.
   2. Handling Irregular Sequences: LSTMs can handle sequences of variable length, making them suitable for situations where the sampling frequency or sequence length may vary.

   Results:

   The LSTM model demonstrates the best performance in successfully predicting the trend of PUN using the variables introduced earlier.

   As shown in Fig. 7, the model's behavior during testing correctly follows most of the PUN oscillations. Despite the limited available data, the LSTM model manages to capture more than 50% of the variability of PUN and predicts with an average error of +/- 15.

   With the possibility of expanding the number of significant variables and the search for a larger dataset, this model can be refined and used as an indicator in the challenging task of predicting PUN on a daily basis, especially in less tumultuous and volatile periods than those considered.

   Mean Squared Error (LSTM): 255.593

   RMSE: 15.987

   R^2: 0.545

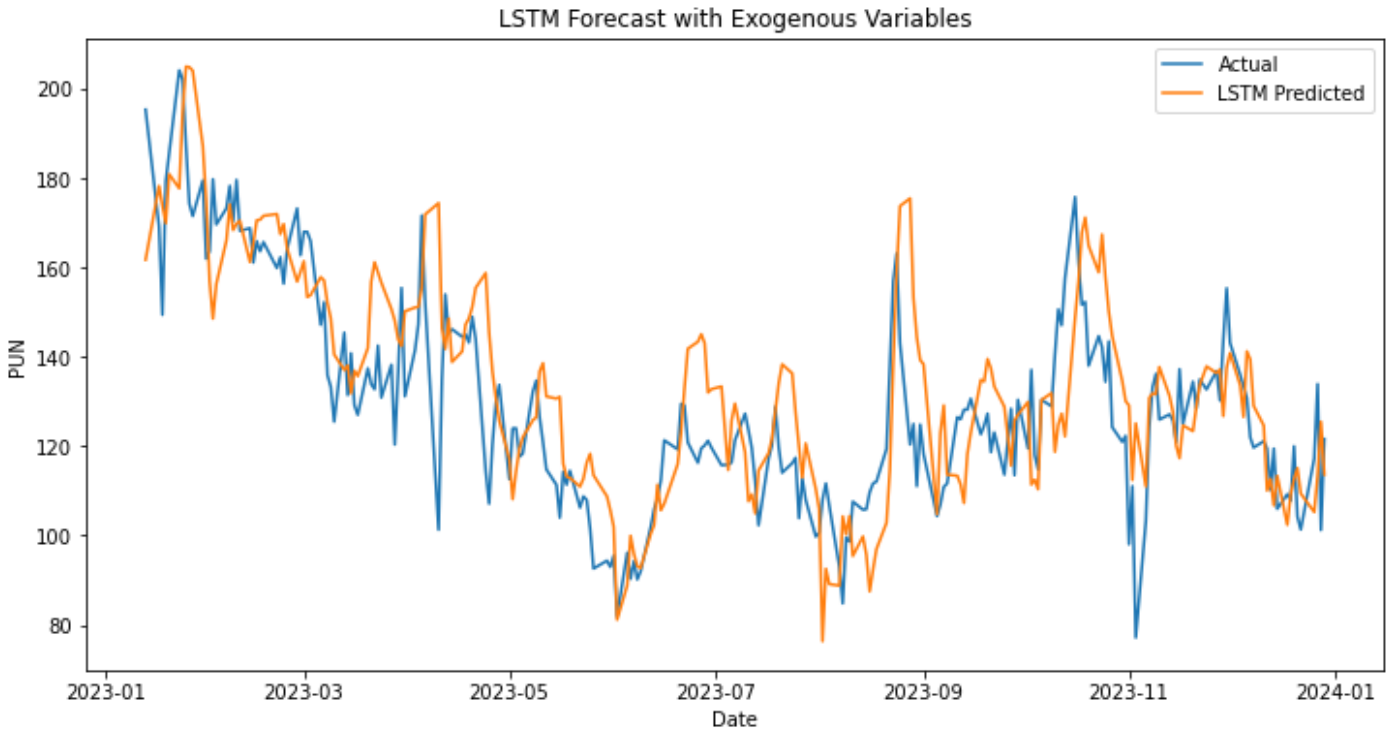

Figure 7: LSTM Forecast Model

7. **LSTM Prediction Model:** Given the compatibility with the data of the LSTM model and its ability to capture non-linear relationships, it appears to be the best starting point for building a Day-ahead prediction model.
In creating this model, it was preferred to use variables at time t-1 as input, as making predictive analyses with inputs at time t would be impractical.
After a brief discussion with industry experts, it was decided to include the variable related to the gas price at time t-1 in this model, as it could carry important informational content in explaining fluctuations in the PUN (Prezzo Unico Nazionale - National Single Price).

Results:

This predictive model demonstrates the ability to capture over 70% of the target variable 'PUN,' a significant result considering the historical period analyzed characterized by high volatility and significant fluctuations, as highlighted in Figure 8. The indicators (MSE and RMSE) show an average error margin of around +-12.
Although still a significant error margin, it represents an excellent starting point for the creation of a true predictive model that can provide insights to analysts, consultants, and/or clients about the direction of the PUN in the next day.
To enhance this model, the addition of further data examined at different times and/or the consideration of incorporating new variables or indices is necessary.
Expanding the time window could be crucial, especially in creating different models, such as a monthly PUN index capable of having a "wide" predictive window to extrapolate and anticipate future trends.

Mean Squared Error (LSTM): 160.186

RMSE: 12.656

R^2: 0.737

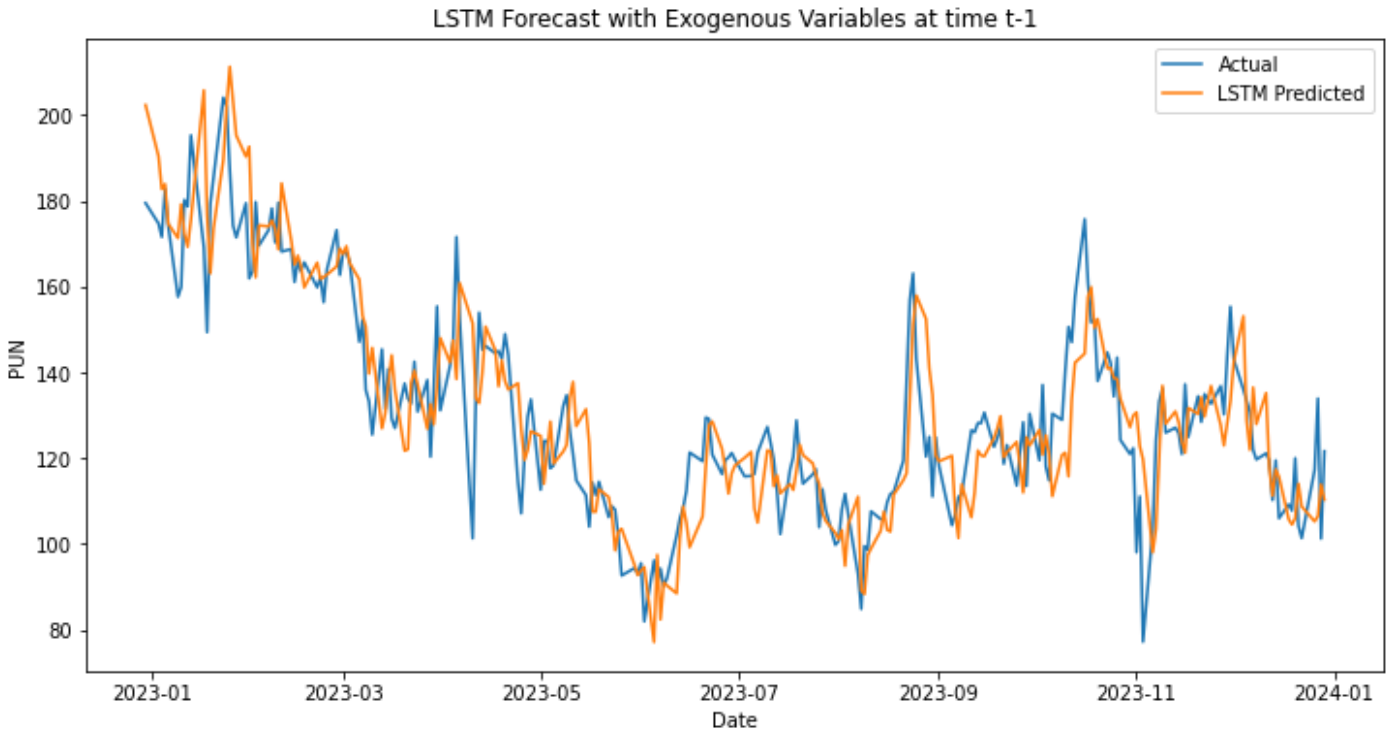

Figure 8: LSTM Prediction Model