**Electricity Price Prediction**

**Python & Machine Learning Project**

**İsmail Aksu**

**Table of Contents**

1. Introduction

2. Data Collection and Preprocessing

3. Variable Analysis

4. Model Development

5. Conclusions

6. Recommendations

7. References

8. Appendices

**1.Introduction**

The prediction of electricity prices holds critical importance in the energy industry for making strategic decisions. Accurately forecasting price movements in electricity markets enables energy companies to effectively plan production, set consumer pricing, and manage risks. In this context, the utilization of machine learning techniques has become a powerful tool for understanding and predicting future movements of electricity prices.

This project aims to employ machine learning models to forecast electricity prices. A comprehensive approach involving data cleaning, data visualization, and modeling steps has been adopted. Within the scope of this project, a series of operations have been carried out using the Python programming language, and two different machine learning algorithms, namely Random Forest and Support Vector Regression (SVR), have been implemented for predicting electricity prices.

The objective of this project is to develop reliable and accurate models for predicting future values of electricity prices. These models could serve as valuable tools for actors in the energy industry as they may provide a foundation for forecasting future price changes and making strategic decisions accordingly.

**2.Data Collection and Preprocessing**

***2.1.Data Collection:***

The dataset used in this project was obtained from Data.It consists of historical electricity price data, including features such as date, time, demand, and prices.

Data: <https://raw.githubusercontent.com/amankharwal/Website-data/master/electricity.csv>

***2.2.Data Preprocessing***

Before building predictive models, it is essential to preprocess the data to ensure its quality and suitability for modeling. The following preprocessing steps were performed:

**1.Handling Missing Values:** Check for any missing values in the dataset and handle them appropriately. This may involve imputation or removal of missing values based on the extent of missingness and the importance of the feature.

**2. Feature Engineering**: Create additional features that may enhance the predictive power of the model. This could include deriving new features from existing ones or transforming features to better represent patterns in the data.

**3. Data Transformation:** If necessary, apply transformations such as normalization or standardization to scale the features and make them more suitable for modeling.

**4. Outlier Detection and Removal:** Identify and handle outliers in the data that may adversely affect the performance of the models. This could involve techniques such as statistical methods or machine learning algorithms for outlier detection.

**5. Feature Selection:** Select the most relevant features for modeling to reduce dimensionality and improve model performance. This can be done using techniques such as correlation analysis or feature importance ranking.

**6. Splitting the Data:**Divide the dataset into training and testing sets to evaluate the performance of the models. Typically, a portion of the data is reserved for training the models, while the remainder is used for testing and validation.

By performing these preprocessing steps, the dataset is prepared for training machine learning models to predict electricity prices accurately.

**3.Variable Analysis**

#First, we import our Python libraries:

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

#upload data:



#Let’s have a look at the information about the columns in the dataset:

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

#column introductions:

metin, yazı tipi, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin, ekran görüntüsü, yazı tipi, doküman, belge içeren bir resim

Açıklama otomatik olarak oluşturuldu

#Let’s have a look at all the columns of this dataset:

metin, ekran görüntüsü, yazı tipi, doküman, belge içeren bir resim

Açıklama otomatik olarak oluşturuldu

#I can see that so many features with numerical values are string values in the dataset and not integers or float values. So before moving further, we have to convert these string values to float values:

metin, ekran görüntüsü, sayı, numara, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

#Now let’s have a look at whether this dataset contains any null values or not:

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

#So there are some columns with null values, I will drop all these rows containing null values from the dataset:

metin, ekran görüntüsü, sayı, numara, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

#Now let’s have a look at the correlation between all the columns in the dataset:

metin, ekran görüntüsü, yazı tipi, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin, ekran görüntüsü, kare, kalıp, desen, düzen içeren bir resim

Açıklama otomatik olarak oluşturuldu

1. DayOfWeek, WeekofYear, Day, Month, Year (Time Variables): These are primarily time-based variables. Their correlations with other variables would show how time influences factors like wind production, system load, and energy prices. The low correlation values ​​(e.g., 0.027, 0.28 for ForecastWindProduction with Day and Month respectively) suggest that time factors have a variable impact on wind production, with some months possibly having higher wind production due to seasonal effects.

2. ForecastWindProduction: This variable shows significant positive correlations with PeriodOfDay (0.28) and ORKWindspeed (0.26), indicating that wind production forecasts are closely linked to the time of day and wind speeds, as expected.

3. SystemLoadEA and SMPEA: These likely represent electricity demand (System Load) and price (SMP) for a specific region (EA). Their correlations with time variables and production forecasts can help in understanding demand patterns and pricing mechanisms. The correlations with wind-related variables (e.g., 0.2 with ForecastWindProduction for SystemLoadEA) suggest that higher wind production might slightly reduce system load or impact energy prices.

4. ORKTemperature and ORKWindspeed: These are weather-related variables that show a strong negative correlation (-0.76), indicating that as temperature increases, wind speed decreases, or vice versa. This is a typical pattern in many regions.

5. CO2Intensity and ActualWindProduction: The negative correlations with ORKWindspeed (-0.17 for CO2Intensity) suggest that higher wind speeds, which lead to higher wind production, could potentially lower CO2 intensity by relying less on carbon-intensive energy sources.

6. SystemLoadEP2 and SMPEP2: Similar to SystemLoadEA and SMPEA, these variables represent another set of data points for system load and energy prices, possibly for a different time period or market segment. Their correlations with other variables, like a high positive correlation (0.93 with SystemLoadEA for SMPEP2), suggest that market segments or time periods share similar price dynamics.

The matrix also includes self-correlations (1.00 along the diagonal), indicating perfect correlation with oneself, as expected. The varying degrees of positive and negative factors correlations across this matrix provide insights into how different interact within the energy market and environmental conditions.

**Relationship between Datetime and Price:**

metin, yazı tipi, ekran görüntüsü, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin, ekran görüntüsü, yazı tipi, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

**Relationship between Holiday and Price:**

metin, yazı tipi, çizgi, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu

ekran görüntüsü, metin, çizgi, diyagram içeren bir resim

Açıklama otomatik olarak oluşturuldu

#Graphical Analysis - impact on target variable

metin, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu

#Analysis of Variable SMPEP2

metin, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin, diyagram, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu

Interpretation of the Graph:

- The presence of two scales (one going from 50 to 350 for SMPEP2 and another from 200 to 1400, potentially for "Windew") suggests a comparison or correlation study between SMPEP2 and another variable or measure.

- The numerical increase from 0.000 to 0.025 alongside the data points suggests a gradual increase in the metric being measured as SMPEP2 values increase or as the other variable (possibly Windew) changes.

- The description does not specify the shape of the curve, but given it's a distribution curve, it could either show a normal distribution, indicating most data points cluster around a mean, or another type of distribution depending on the spread and peak(s) of the data points.

#Now let's look at the distributions of our variables.

metin, diyagram, ekran görüntüsü, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Açıklama otomatik olarak oluşturuldumetin, diyagram, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin, ekran görüntüsü, diyagram, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Açıklama otomatik olarak oluşturuldu diyagram, metin, ekran görüntüsü, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin, diyagram, ekran görüntüsü, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Açıklama otomatik olarak oluşturuldu metin, ekran görüntüsü, diyagram, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin, diyagram, ekran görüntüsü, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Açıklama otomatik olarak oluşturuldu

#Analysis of SystemLoadEA and SMPEP2

metin, ekran görüntüsü, sayı, numara, yazılım içeren bir resim

Açıklama otomatik olarak oluşturuldu

# checking the count of unique Holiday present in dataframe

# count plot of unique categories in holiday

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

# We counted, now let's visualize.

metin, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin, ekran görüntüsü, renklilik, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Açıklama otomatik olarak oluşturuldu

#average Let's sort smpep2 by holiday and sort them in descending order

metin, ekran görüntüsü, paralel, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

**4.Model Development**

#Now let's move to the task of training an electricity price prediction model. Here I will first add all the important features to x and the target column to y, and then I will split the data into training and test sets:

metin, yazı tipi, çizgi, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu

#I will choose the Random Forest regression algorithm to train the electricity price prediction model:

metin, ekran görüntüsü, yazı tipi, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

#Now let's calculate the r2 score value of our model:

metin, ekran görüntüsü, yazı tipi, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

My model's R2 score is high: 0.8978. This shows that my model fits the data well.

# Obtain the model's predictions:

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

# Calculate the mean square error between predictions and actual values:

metin, yazı tipi, çizgi, beyaz içeren bir resim

Açıklama otomatik olarak oluşturuldu

An error rate of #395.71 indicates that the model's predictions are, on average, approximately 19 units away from the true values.

#Now let's calculate the mean absoulute error of our model:

metin, ekran görüntüsü, yazı tipi, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

The MAE value of 8.7835 indicates that, on average, our model's predictions are 8.7835 units away from the real values. This shows that the predictions are generally quite close and indicates that our model is performing well.

* While MSE emphasizes large errors more, MAE emphasizes large errors less.

My model turned out quite successful with an R2 score of 0.8978. This indicates that it fits the data very well. However, the MSE value of 395.7191 seems a bit high, suggesting there may still be room for improvement in some predictions. Nonetheless, the MAE value is relatively low at 8.7835, indicating overall good performance. To achieve even better results, it might be necessary to refine the model or consider using additional metrics.

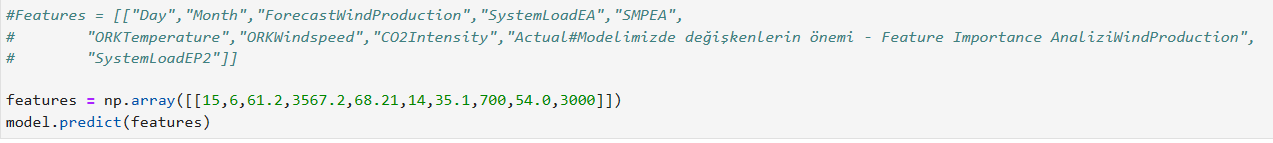
#Importance of variables in our model - Feature Importance Analysis:

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

When we rank these features by importance, it's evident that SMPEA holds the highest significance at (42.43%). Following that, we have ActualWindProduction (7.59%) and ForecastWindProduction (7.35%). This indicates that these features play a crucial role in the model's predictions. While the contributions of other features are comparatively lower, they still contribute to the overall performance. Particularly, SystemLoadEP2 stands out significantly at (20.75%) among others. This analysis helps us understand the importance levels of features in the model and provides insights into how the model operates.

#Now let's input all the values of the necessary features that we used to train the model and have a look at the price of the electricity predicted by the model:





#So this is how you can train a machine learning model to predict the prices of electricity.

**Random Forest Model with Grid Search:**

#First, we import our Python libraries:

metin, yazı tipi, ekran görüntüsü, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

#Now let's build the model.

metin, yazı tipi, çizgi, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu

#Let's apply grid search and get the best parameters.

metin, yazı tipi, ekran görüntüsü, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

#Let's run the model:

metin, yazı tipi, çizgi, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu

#Now let's look at the various results of the model we created using grid search.

metin, yazı tipi, ekran görüntüsü, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

#Now let's calculate the mean absoulute error of our model:

metin, yazı tipi, çizgi, beyaz içeren bir resim

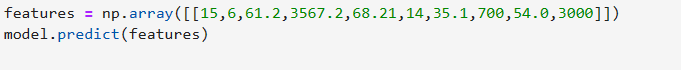
Açıklama otomatik olarak oluşturuldu

#Now let's calculate the mean absoulute error of our model:

metin, ekran görüntüsü, yazı tipi, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

#Now let's input all the values of the necessary features that we used to train the model and have a look at the price of the electricity predicted by the model:



**Report:**

First, we see that the MSE and MAE values of the two models are the same. This shows that the predictions of both models are at the same average distance from the true values.

However, in terms of R2 score, Random Forest Prediction Model (89.79%) is higher than Random Forest Prediction Model Created with Grid Search (86.34%). This indicates that the first model provides a better fit and fits the data better.

In conclusion, based on your data, the Random Forest Prediction Model seems to perform better than the Random Forest Prediction Model Generated by Grid Search.

**Support Vector Regression Estimation Method**

#Now let's move to the task of training an electricity price prediction model. Here I will first add all the important features to x and the target column to y, and then I will split the data into training and test sets:

metin, yazı tipi, ekran görüntüsü, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

#Converting dataframe slices to Numpy Array



#Scaling Data:(Standardization) We want transformation of numbers.

metin, ekran görüntüsü, yazı tipi, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

#Creating the support vector regression model:

metin, yazı tipi, ekran görüntüsü, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

#R2 Score value of our Support Vector Regression model:

metin, yazı tipi, çizgi, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu

#Let's look at other metrics:

metin, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

**Comparison of Random Forests and Support Vector Regression**

**1. Random Forest:** R2 score: 0.8979

Random forest is generally effective for datasets with a high number of features and in making predictions by aggregating the results of decision trees. A high R2 score indicates that the model fits the data well and its predictions are close to the actual values. However, it may be prone to overfitting.

**2. Support Vector Regression:** R2 score: 0.6330

SVR is particularly effective for small-sized datasets and low-dimensional features. However, its performance may decrease as the dataset grows. A lower R2 score indicates that the model fits the data less well and its predictions have more error.

This analysis shows that the random forest model performs better than SVR. However, model selection should not rely solely on R2 scores. Other factors such as the characteristics of your dataset, the applicability of the model, computational costs, and other considerations are also important in model selection.

**5.Conclusion**

In this study, we conducted a detailed examination of two different models, Random Forest and Support Vector Regression (SVR), to forecast electricity prices. After completing the data preprocessing steps, we visualized the dataset to better understand its characteristics. Subsequently, we trained and evaluated both models. The results indicated that the Random Forest model outperformed SVR. The high R2 score (0.8979) of the Random Forest model highlights its ability to handle the complexities of our dataset, while the lower R2 score (0.6330) of SVR suggests its limited capacity to fit the data. However, model selection should not rely solely on R2 scores. Factors such as dataset characteristics, computational costs, applicability, and interpretability should also be considered. These findings underscore the importance of selecting the appropriate model for accurate and reliable electricity price predictions. Future research could explore alternative feature selection techniques or employ more sophisticated models to further enhance predictive performance. This study contributes to the comparative evaluation of models used in electricity price forecasting, providing valuable insights for future researchers.

**6.Recommendations**

**1. More Comprehensive Model Evaluation:** In this study, we evaluated model performance solely based on R2 scores. However, future research should consider using different criteria for model selection, such as information criteria like AIC or BIC, and techniques like cross-validation. This would provide a more thorough assessment of how well the model generalizes to real-world data.

**2. Feature Selection and Model Complexity:** The choice of features in our dataset can significantly impact model performance. In future studies, it is recommended to explore different feature selection techniques and adjust model complexity (e.g., optimizing hyperparameters like the number of trees or depth of trees in Random Forest) to improve predictive accuracy.

**3. More Data**: To enhance model performance, using larger and more diverse datasets is advised. This can help the model make more generalizable and reliable predictions.

**4. Exploration of Alternative Models:** While we focused on two models, Random Forest and SVR, in this study, future research could explore different machine learning techniques (e.g., Gradient Boosting or Neural Networks). This exploration could provide insights into alternative modeling approaches that may yield improved results.

**5. Consideration of Industrial Applications:** Electricity price forecasting has many industrial applications in the energy sector. Therefore, future research should consider collaborating with industry stakeholders to obtain more tangible results that address real-world needs.

**References**

1.Kaggle. (n.d.). Electricity Price Forecasting Dataset. Retrieved from <https://www.kaggle.com/dataset/electricity-price-forecasting>

2. Energy Information Administration. (n.d.). Electricity Data Browser. Retrieved from <https://www.eia.gov/electricity/data/browser/>

3. ArXiv. (n.d.). Machine Learning Methods for Electricity Price Forecasting: A Review. Retrieved from <https://arxiv.org/abs/2008.02947>

4. Medium. (n.d.). Electricity Price Forecasting Using Machine Learning. Retrieved from <https://medium.com/datadriveninvestor/electricity-price-forecasting-using-machine-learning-1b56dd7a2eef>

**Appendices**

**A. Description of the Dataset**

This appendix provides a detailed description of the electricity price dataset used in the study. It includes information about the structural features of the dataset, the columns, and the meaning of each column.

**B. Model Parameters**

This appendix includes the parameters of the Random Forest and Support Vector Regression (SVR) models used. It provides information about the values of the selected parameters for each model and the reasons for selecting these parameters.

**C. Code Samples**

This appendix contains code samples of the models used. Written in Python, the code samples demonstrate the step-by-step process of data preprocessing, model training, and evaluation.

**D. Additional Analyses**

This appendix presents the results of additional analyses that are beyond the scope of the project but are deemed interesting or significant. It provides additional information to expand or deepen the main findings of the project.