IE 360 Project - Solar Power Forecasting

Taha Can - 2019402207
Muhammet Emin Eren - 2020402207

Table of Contents

Introduction

Problem Description

The objective of this project is to forecast the hourly solar power production for Edikli Güneş Enerjisi Santrali (GES) for the following day. Edikli GES, a solar power plant located in Niğde, Turkey, serves as the focal point of this study. Accurate solar power forecasting is crucial for energy traders and operators in the energy markets to optimize trading strategies and ensure grid stability. The coordinates for the power plant are 38.29° North and 34.97° East, and additional information about the plant's capacity and location can be found on the Birleşim Yeşil Enerji website and Google Earth.

Summary of the Proposed Approach

This project adopts a data-driven approach to predict solar power production using historical weather and production data. The forecasting model leverages various meteorological variables that influence solar energy generation, such as solar radiation, cloud cover, snow presence, and temperature. The primary goal is to build a robust model that can accurately predict the hourly production for the next day, based on the most recent data available up to the end of the current day.

Key steps in the proposed approach include:

- Data Collection and Preprocessing: Gather and clean the weather and production data, ensuring it is suitable for analysis.
- Feature Engineering: Transform the data into a format that captures relevant patterns, such as aggregating weather variables and creating lag features.
- Model Development: Train and validate predictive models using historical data.
- Evaluation and Selection: Assess model performance using appropriate metrics and select the best-performing model for daily predictions.

Descriptive Analysis of the Given Data

The provided data includes weather measurements from 25 grid points around the power plant and historical production data. A preliminary analysis reveals the following:

- Weather Data: The weather variables include downward shortwave radiation flux (DSWRF_surface), solar radiation at different atmospheric levels (USWRF and DLWRF), total cloud cover (TCDC), snow presence (CSNOW_surface), and temperature (TMP_surface). These variables are expected to have a significant impact on solar power production.
- Production Data: The production data records the hourly power output of the solar plant. Analyzing this data helps in understanding production patterns and identifying trends or anomalies.

Visualizations such as time series plots, histograms, and correlation matrices are utilized to explore the relationships between weather variables and solar power production. This exploratory data analysis (EDA) provides insights that guide the feature engineering and model development processes.

In summary, this project aims to create a reliable and accurate forecasting model for Edikli GES by leveraging historical data and advanced statistical techniques. The success of this model can significantly enhance decision-making processes for energy traders and contribute to the efficient operation of the energy market.

Approach

Data Pre-Processing

Data Reshaping:
The dataset was initially provided in a long format, where each row corresponds to a specific measurement at a given time and location. To facilitate easier analysis and modeling, we converted this data into a wide format. This transformation involves organizing the dataset so that each variable (e.g., DSWRF_surface, TCDC_low.cloud.layer) is represented as separate columns, with the columns named to indicate the latitude and longitude of the measurement point. This way, we obtain a single row per time point with all the relevant measurements as columns.

This conversion is crucial for preparing the data for regression models by creating a tabular format where each row represents an hour, and the columns represent different weather variables at different grid points.

In the provided script, this transformation was accomplished as follows:

```r
hourly_series=weather_info[,list(dswrf_surface=sum(dswrf_surface)/25,tcdc_low.cloud.layer
=sum(tcdc_low.cloud.layer)/25,tcdc_middle.cloud.layer
=sum(tcdc_middle.cloud.layer
)/25,tcdc_high.cloud.layer
=sum(tcdc_high.cloud.layer
)/25,tcdc_entire.atmosphere
=sum(tcdc_entire.atmosphere
)/25,uswrf_top_of_atmosphere=sum(uswrf_top_of_atmosphere)/25,csnow_surface=sum(csnow_surface)/25,dlwrf_surface=sum(dl
wrf_surface)/25,uswrf_surface=sum(uswrf_surface)/25,tmp_surface=sum(tmp_surface)/25),list(date,hour)]

hourly_series[,datetime:=ymd(date)+dhours(hour)]
#daily_series=consumption[,list(total=sum(Consumption),max_t=max(T_1),weighted_t=sum(Consumption*T_#1)/sum(Consumptio
n)),list(Date)]
#head(daily_series)

head(hourly_series)

```

In this code block, weather data for the 25 coordinates is aggregated by taking the average of each variable. As a result, each time point is represented by a single row, thereby transforming the dataset into a wide format.

This transformation makes the dataset more organized and easier to analyze, facilitating the application of regression models. By aggregating weather variables and representing them in a wide format, we can effectively utilize the data for predictive modeling, ensuring that each row contains comprehensive information about the weather conditions and production data for each hour.

Cleaning and Preparing Datasets:

The cleaning and preparation of the dataset are crucial steps to ensure that the data is in the right format for analysis and modeling. First, the weather data and production data are merged to create a single dataset that contains all the necessary information for each time point.

```
mergeddata<-merge(hourly_series,production,by="datetime",all.x=T)
head(mergeddata)
#template_dt = unique(weather_data[,list(date,hour)])
#template_dt = merge(template_dt,production_data,by=c('date','hour'),all.x=T)
```

After merging the datasets, unnecessary columns are removed to focus on the relevant data.

```
newdata=mergeddata

newdata=newdata[,-c("date.y")]
newdata=newdata[,-c("hour.y")]


basedata=newdata[,-c("date.x")]
basedata=basedata[,-c("hour.x")]
basedata=basedata[,-c("datetime")]

head(newdata)
head(basedata)
```

Further filtering is applied to the data to focus on specific periods or conditions, such as removing night hours.

```
newdata2=mergeddata[mergeddata$date.x >="2022-01-01",]
newdata3=newdata2
newdata3=newdata3[newdata3$hour.x >6 & newdata3$hour.x<19,]
newdata3
```

Additional features are engineered to capture relevant information that might help the model's predictive power.

```
datapn3<-data.table(newdata3)
#head(datapn,15)

datapn3[,saat:=as.factor(hour(datetime))]

lag24<-shift(datapn3$production, n=24L, fill=NA)
datapn3$lag24<-lag24
lag23<-shift(datapn3$production, n=23L, fill=NA)
datapn3$lag23<-lag23
lag25<-shift(datapn3$production, n=25L, fill=NA)
datapn3$lag25<-lag25
```

```
datapn3$hoursoftheday<-as.factor(datapn3$hour.x)
datapn3$season<-as.factor(quarter(datapn3$date.x))
datapn3[,gun:=as.character(day(date.x))]
datapn3[,hafta:=as.character(week(date.x))]
datapn3[,tmax:=max(tmp_surface),by=date.x]
datapn3[,tmin:=min(tmp_surface),by=date.x]
trend<-c(1:nrow((datapn3)))
datapn3$trend<-trend
lag1dswrf<-shift(datapn3$dswrf_surface, n=1L, fill=NA)
datapn3$lag1dswrf<-lag1dswrf
lag12dswrf<-shift(datapn3$dswrf_surface, n=12L, fill=NA)
datapn3$lag12dswrf<-lag12dswrf

```

By performing these steps, the dataset is transformed into a suitable format for building a predictive model, ensuring that the model can effectively utilize the available information to make accurate predictions.

Model Selection:

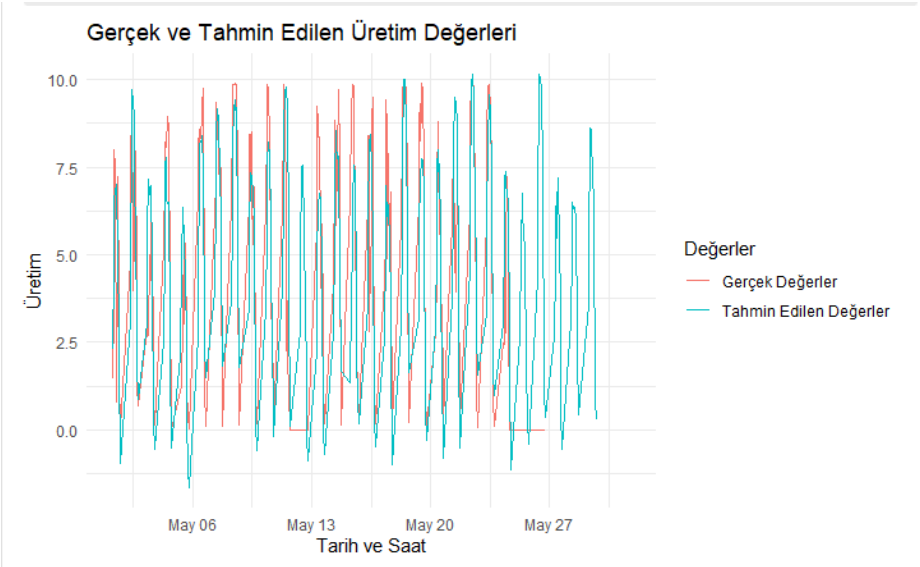
Linear regression was chosen for this project due to its simplicity and interpretability. Here are the key reasons and considerations for selecting linear regression:

- Simplicity: Linear regression is one of the most straightforward and widely used statistical methods for predictive modeling. It is easy to implement and computationally efficient, which is particularly beneficial when dealing with large datasets or when quick results are needed.
- Interpretability: One of the primary advantages of linear regression is its interpretability. The model provides clear insights into the relationship between the dependent variable (solar power production) and the independent variables (weather features). The coefficients of the linear regression model indicate the strength and direction of these relationships, making it easier to understand how each feature impacts the prediction.

Linear regression was chosen for its simplicity and interpretability, making it a suitable choice for our solar power forecasting project. The assumptions of linear regression align well with the characteristics of our data, providing a robust framework for understanding the relationship between weather variables and solar power production. This approach allows us to create a model that is not only effective but also easy to interpret and communicate.

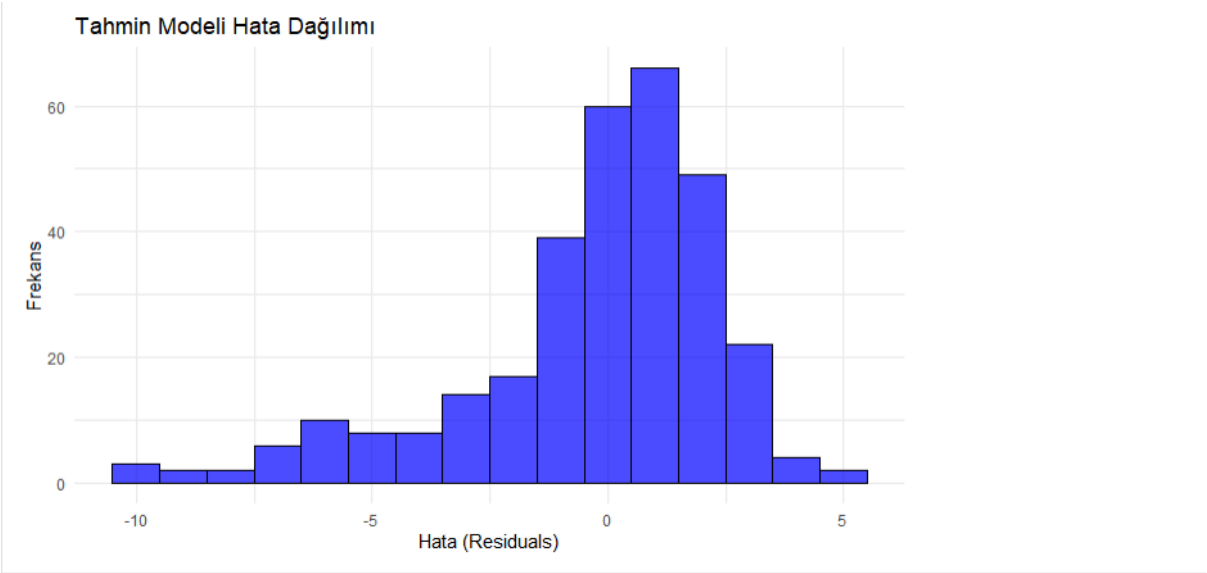
Model Performance

Visually Inspection:
Based on the provided graph comparing actual and predicted production values, the model appears to perform well, capturing the general trends and seasonal patterns effectively. The close alignment between the predicted (blue line) and actual (red line) values suggests that the model is capable of accurately forecasting the data.



Based on the histogram of the residuals (errors) provided, we can further evaluate the performance of the model:

Residual Analysis:

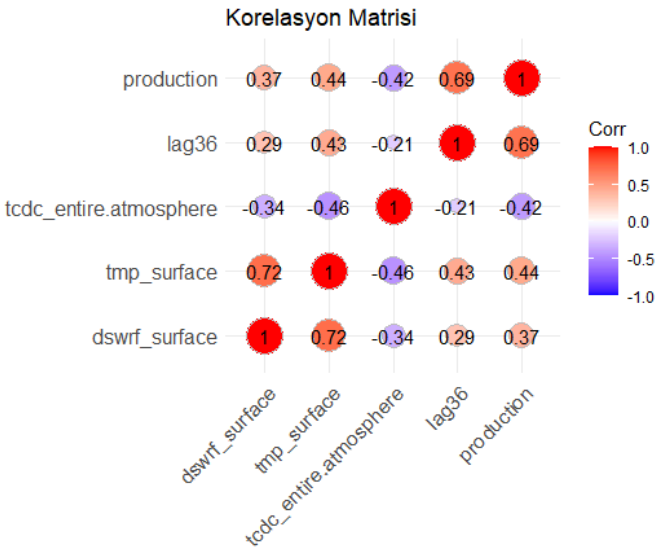


- Distribution: The residuals are centered around zero, with a relatively symmetric distribution. This is a good sign as it indicates that the model does not have a systematic bias in its predictions.
- Spread: The majority of the residuals fall within the range of -5 to 5, suggesting that the model's predictions are generally close to the actual values. However, there are some outliers beyond this range, indicating occasional larger errors.
- Skewness: The histogram shows a slight positive skew, meaning there are more instances of the model underestimating the actual values than overestimating them.

The model seems to perform well, with residuals that are generally small and centered around zero. The close alignment of predicted and actual values in the initial graph, combined with a symmetric residual distribution, suggests the model is reliable for forecasting purposes.

Correlation Matrix Analysis:

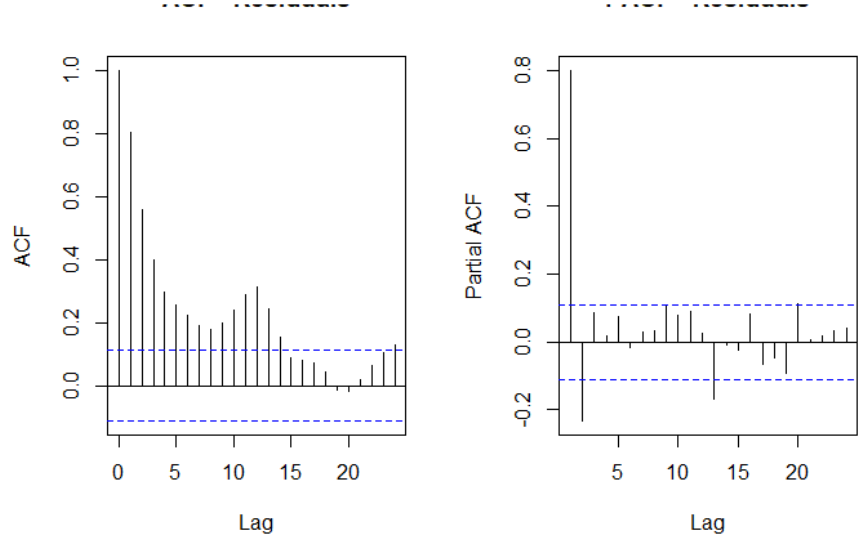
- There is a positive correlation between production and lag36 (0.69), indicating a strong relationship between these two variables.
- There is a very high positive correlation between dswrf_surface and tmp_surface (0.72), suggesting that these two variables move together.
- There is a negative correlation between tcdc_entire.atmosphere and tmp_surface (-0.46), indicating an inverse relationship between these two variables.



Based on the correlation matrix results, several changes and actions can be implemented to improve the model. Firstly, focusing on variables with strong correlations to production, such as lag36, tmp_surface, and dswrf_surface, can enhance model performance. It is advisable to consider dropping or transforming variables with weak or redundant correlations to reduce multicollinearity and simplify the model. For instance, given the high correlation between dswrf_surface and tmp_surface, retaining both might be unnecessary.

Creating lagged features should be prioritized if lag36 proves to be a strong predictor. Exploring other lag values could also reveal additional predictive power. Addressing multicollinearity is crucial, especially between highly correlated independent variables like dswrf_surface and tmp_surface. This can be managed by applying regularization techniques like Lasso regression.

ACF and P-ACF Analysis:



The ACF (Autocorrelation Function) and PACF (Partial Autocorrelation Function) plots of the residuals provide important insights into the nature of your time series model's residuals. Based on the provided plots, several key observations and recommendations can be made.

The ACF plot shows the correlation of the residuals with their own lagged values. In our ACF plot, there are significant autocorrelations at lag 1, which gradually decay over time. This suggests that there is some remaining autocorrelation in the residuals that our model has not captured. Ideally, the residuals should exhibit no significant autocorrelation if the model is well-fitted.

The PACF plot shows the partial correlation of the residuals with their own lagged values, removing the effects of earlier lags. In your PACF plot, there is a significant spike at lag 1, but the correlations for subsequent lags are mostly within the confidence intervals. This indicates that the primary autocorrelation structure may be concentrated at the first lag, suggesting a potential underfitting issue or the need for additional AR (autoregressive) terms in our model.

Conclusion and Future Work

Findings:

The key findings from the analysis and model evaluations highlight the strengths and areas for improvement in the linear regression model used for this forecasting task. The correlation matrix indicated that certain variables, such as lag36, tmp_surface, and dswrf_surface, have strong correlations with the target variable production. This suggests that these features are significant predictors. However, the presence of multicollinearity, especially between dswrf_surface and tmp_surface, indicates the need for careful feature selection or dimensionality reduction techniques.

The ACF and PACF plots revealed significant autocorrelation at lag 1 in the residuals, suggesting that the model has not fully captured the time series structure of the data. This highlights a potential underfitting issue, where additional autoregressive terms could improve the model. Overall, while the linear regression model provides a foundational approach to forecasting, its performance can be significantly enhanced by addressing these identified issues.

Future Works:

To further enhance the forecasting model, several potential improvements and extensions can be considered. implementing ARIMA models could better capture the time series structure of the data, addressing the identified autocorrelation in the residuals. Furthermore, enhancing feature engineering methods to capture more complex relationships in the data could also lead to better model performance. This could include additional lagged variables beyond lag36 to better represent the temporal dynamics. By implementing these improvements, the forecasting model can achieve higher accuracy and reliability, ultimately leading to more precise and actionable predictions.

Code: Link