

Data Science Anwendungen -Strompreis-Prognose-

Eingereicht von:

**Ilyass Touisse (30568)
Mohamed Alilou (30639)
Mustapha Jetti (30224)
Hajar Ammar (31141)
Safae Aissati (30641)**

Betreut von:

Prof. Dr. Christian Schmeißer

Inhaltsverzeichnis

Einleitung	3
1. Domänenhintergrund und Zielsetzung	3
1.1. 1.1 Hintergrundwissen	3
1.1.1. Was ist Prognose?.....	3
1.1.2. Der deutsche Strommarkt.....	3
1.1.3. EPEX SPOT.....	4
1.1.4. Wie der Day-Ahead-Markt funktioniert	4
1.1.5. Market Clearing Price (MCP).....	5
1.2. Zielsetzung	5
2. Die Daten.....	5
2.1. Day-Ahead-Price	5
2.2. Load Data	6
2.3. Generation Data.....	6
3. Theoretische Grundlagen	6
3.1. Allgemeinen Definitionen	6
3.2. SARIMA	7
3.2.1. Autoregressive Integrated Moving Average (ARIMA)	7
3.2.2. Seasonal Autoregressive Integrated Moving Average (SARIMA)	8
3.2.3. Seasonal Autoregressive Integrated Moving Average mit Exogenen Variablen (SARIMAX)	8
3.3. XGBoost	8
3.3.1. Boosting.....	9
3.3.2. Extreme Gradient Boosting (XGBoost).....	9
3.4. Fehlermetriken.....	10
3.4.1. Mean Absolute Error (MAE).....	10
3.4.2. Root Mean Squared Error (RMSE).....	10
3.4.3. Maximum Prediction Error (MPE).....	10
4. Methodologie	11
4.1. Datenaufbereitung.....	11
4.2. Datenexploration (EDA)	12

4.2.1. Allgemeine deskriptive Statistiken	12
4.2.2. Preisvolatilität	14
4.2.3. Korrelationsanalyse	14
4.3. Implementierung	15
4.3.1. SARIMA	15
4.3.2. SARIMAX	18
4.3.3. XGBoost	19
Zusammenfassung.....	22

Abbildungsverzeichnis

Abbildung 1: Zussammenführen der Daten in Power Query (Merge-Funktion)	11
Abbildung 2: finalen Tabelle nach der Datenaufbereitung	12
Abbildung 3: Day-Ahead-Strompreise in der deutschen Gebotszone.....	13
Abbildung 4: Monatliche, wöchentliche und stündliche durchschnittliche der Day-Ahead- Preise.....	14
Abbildung 5: Autokorrelationsfunktion und Partielle Autokorrelationsfunktion.....	16
Abbildung 6: Strompreis-Prognose vs. Ist (SARIMA).....	17
Abbildung 7: Strompreis-Prognose VS. Ist (SARIMAX)	18
Abbildung 8: Strompreis-Prognose vs. Ist (XGBoost)	21

Tabellenverzeichnis

Tabelle 1: Deskriptive Statistiken der Daten.....	13
Tabelle 2: Korrelation zwischen Strompreis und Variablen	14
Tabelle 3: Fehlermetriken (SARIMA-Modell)	17
Tabelle 4: Fehlermetriken (SARIMAX-Modell)	19
Tabelle 5: Fehlermetriken (XGBoost-Modell)	21

Einleitung

Die Entwicklung der Strompreise auf dem deutschen Energiemarkt ist in den letzten Jahren durch zunehmende Volatilität gekennzeichnet. Der Day-Ahead-Strommarkt der europäischen Strombörse spielt eine Schlüsselrolle bei der täglichen Preisbildung. Ziel der Arbeit ist die Prognose des Day-Ahead-Strompreises für die deutsche Gebotszone auf Basis historischer Markt- und Wetterdaten. Dabei werden verschiedene Modelle der Zeitreihenanalyse und des maschinellen Lernens eingesetzt und hinsichtlich ihrer Prognosegüte verglichen. Der Fokus liegt dabei auf der methodischen Umsetzung, der Datenaufbereitung und der Bewertung der Ergebnisse anhand etablierter Fehlermaße.

1. Domänenhintergrund und Zielsetzung

1.1. 1.1 Hintergrundwissen

1.1.1. Was ist Prognose?

Prognose ist eine Methode, um fundierte Vorhersagen zu treffen, indem historische Daten als Eingabe für die Bestimmung des Verlaufs zukünftiger Trends verwendet werden. Unternehmen und Organisationen nutzen Prognosen für viele verschiedene Zwecke, beispielsweise um zukünftige Ausgaben zu prognostizieren und um zu bestimmen, wie sie ihr Budget zuweisen. Beispielsweise kann die Prognose der BEV-PKW-Neuzulassungen in Deutschland dazu beitragen, politische Entscheidungsträger bessere strategische Entscheidungen in Bezug auf Nachhaltigkeit zu treffen. Generell gibt es zwei Prognosemethoden: Qualitatives-, und quantitative Prognosen. Qualitative Prognosemethoden sind in hohem Maße auf Expertenmeinungen angewiesen und eignen sich am besten für kurzfristige Vorhersagen. Qualitative Prognose ist geeignet, wenn keine historischen Daten zur Verfügung stehen. Beispiele für qualitative Prognose sind Interviews, Besuche vor Ort oder Marktforschung. Quantitative Prognosemethoden werden verwendet, wenn historische Daten schon existieren. Die meisten quantitativen Prognosemethoden verwenden Zeitreihendaten, die im Laufe der Zeit in regelmäßigen Abständen gesammelt werden. Die meistverwendete quantitative Prognosemethode ist Prognose durch die Zeitreihenanalyse (Hyndman and Athanasopoulos, 2021).

1.1.2. Der deutsche Strommarkt

Der deutsche Strommarkt zählt zu den größten Elektrizitätsmärkten Europas und basiert seit der Marktliberalisierung in den 1990er-Jahren auf einem wettbewerbsorientierten Modell. Die Marktstruktur gliedert sich in drei zentrale Bereiche: Erzeugung und Großhandel, Netzübertragung und -verteilung sowie Endkundenversorgung. Die Liberalisierung hat zur Entflechtung dieser Bereiche geführt, sodass Netzbetreiber unabhängig von Stromproduzenten agieren (BMWK, 2024).

Die Übertragung auf Höchstspannungsebene übernehmen vier Übertragungsnetzbetreiber (TSO): TenneT, 50Hertz, Amprion und TransnetBW. Diese Akteure sind nicht nur für den

nationalen Netzbetrieb zuständig, sondern auch Teil des europäischen Verbunds ENTSO-E, der die grenzüberschreitende Stromversorgung koordiniert. Der Staat hält strategische Beteiligungen an einigen dieser Unternehmen, um die öffentliche Kontrolle über kritische Infrastruktur zu sichern (Bundesnetzagentur & Bundeskartellamt, 2025).

Ein zentraler Einflussfaktor ist die Energiewende. Im Jahr 2024 stammten mehr als 55 % des in Deutschland erzeugten Stroms aus erneuerbaren Quellen, hauptsächlich aus Wind- und Solarkraft. Das Erneuerbare-Energien-Gesetz (EEG) unterstützt diese Entwicklung. Gleichzeitig führen die fluktuierenden Einspeisungen erneuerbarer Energien zu Herausforderungen bei der Netzstabilität und Preisbildung, insbesondere im kurzfristigen Stromhandel (Agora Energiewende, 2025).

1.1.3. EPEX SPOT

Die European Power Exchange (EPEX SPOT SE) ist die zentrale Börse für den kurzfristigen Stromhandel in Deutschland und weiteren Ländern wie Frankreich, Österreich und der Schweiz. Sie organisiert den Day-Ahead- und Intraday-Markt und schafft über Auktionen ein transparentes Preisbild für Stromlieferungen am Folgetag (EPEX SPOT SE, 2025).

An der EPEX SPOT handeln Erzeuger, Energieversorger, Stadtwerke und Großverbraucher standardisierte Produkte. Der Großteil des Stromhandels erfolgt im Day-Ahead-Markt. 2024 wurden allein in der deutschen Gebotszone mehr als 200 TWh gehandelt, was die enorme Bedeutung der Plattform für den nationalen Markt unterstreicht (EPEX SPOT SE, 2025).

EPEX SPOT agiert als neutraler Marktplatz unter europäischer Regulierung und ist in das sogenannte Market Coupling integriert – ein Mechanismus, der die Preisbildung zwischen europäischen Ländern harmonisiert. Damit wird der grenzüberschreitende Stromfluss effizienter gestaltet und die europäische Marktintegration weiter gefördert.

1.1.4. Wie der Day-Ahead-Markt funktioniert

Im Day-Ahead-Markt werden Stromlieferungen für jede Stunde des Folgetags gehandelt. Bis 12:00 Uhr geben Marktteilnehmer stündliche Gebote für Erzeugung oder Verbrauch ab. Diese werden in Angebots- und Nachfragekurven überführt. Die Auktion folgt dem Einheitspreisverfahren: Der Schnittpunkt beider Kurven bestimmt den Preis, zu dem Strom für die jeweilige Stunde gehandelt wird (EPEX SPOT SE, 2025).

Dieser Mechanismus erlaubt es, Angebot und Nachfrage effizient aufeinander abzustimmen. Zusätzlich können Marktteilnehmer über den Intraday-Markt kurzfristige Korrekturen vornehmen, etwa bei unerwarteten Änderungen in der Stromerzeugung durch Wettereffekte (Agora Energiewende, 2025).

Der Day-Ahead-Markt bietet damit nicht nur Preistransparenz, sondern ist auch ein wichtiges Instrument für die Systemstabilität – insbesondere angesichts der steigenden Einspeisung erneuerbarer Energien (Bundesnetzagentur & Bundeskartellamt, 2025).

1.1.5. Market Clearing Price (MCP)

Der Market Clearing Price (MCP) ist der Preis, der im Ergebnis der Day-Ahead-Auktion für jede Stunde ermittelt wird. Er entsteht dort, wo sich das aggregierte Stromangebot und die Nachfrage treffen. Für alle akzeptierten Gebote gilt derselbe Einheitspreis – unabhängig vom ursprünglich angebotenen Preis (EPEX SPOT SE, 2025).

Der MCP ist damit eine zentrale Marktgröße, da er die tatsächliche Knappheit oder das Überangebot im System abbildet. In Zeiten hoher Einspeisung, beispielsweise durch viel Solar- oder Windstrom, kann es zu sehr niedrigen oder gar negativen Preisen kommen. Umgekehrt steigen die Preise bei Dunkelflauten oder Lastspitzen deutlich an (Agora Energiewende, 2025).

Für Prognosemodelle stellt der MCP die Zielgröße dar. Seine Dynamik spiegelt die hohen Anforderungen an kurzfristige Marktanalysen und Vorhersagesysteme wider. Damit ist der MCP nicht nur ein Preis, sondern ein Indikator für die gesamte Systemlage.

1.2. Zielsetzung

Das Ziel dieser Arbeit besteht in der Implementierung und Bewertung verschiedener Prognosemodelle zur Vorhersage der Day-Ahead-Strompreise in Deutschland. Dabei wird untersucht, welche Modellierungsansätze sich in Abhängigkeit von den jeweiligen Datensituationen besonders gut eignen. Die Vorhersage von Strompreisen ist für die wirtschaftliche Planung von Energieversorgern, Händlern und Großverbrauchern von entscheidender Bedeutung. Sie unterstützt zudem die Integration erneuerbarer Energien, indem sie Schwankungen im Angebot ausgleicht. Netzbetreiber nutzen Prognosen, um die Netzstabilität zu gewährleisten und Engpässe zu vermeiden. Darüber hinaus dienen Strompreisprognosen als Grundlage für politische Entscheidungen und fördern datengetriebene Geschäftsmodelle in der Energiewirtschaft.

2. Die Daten

Im folgenden Kapitel werden alle verfügbaren Datensätze beschrieben und analysiert. Diese Analysen bieten eine erste Grundlage für Modellierungsentscheidungen sowie Hinweise auf mögliche Probleme, die später auftreten können.

2.1. Day-Ahead-Price

Für die Analyse wurde der stündliche Day-Ahead-Strompreis für den deutschen Markt für den Zeitraum 01.01.2019 bis 31.12.2024 verwendet. Die Daten sind aus der vom Bund bereitgestellten SMARD-Anwendung entnommen. Der Datensatz umfasst 52.644 Datenpunkte für einen Preiswert für jede betrachtete Stunde.

Die Preise sind in Euro pro Megawattstunde (€/MWh) angegeben. Es gibt keine fehlenden Werte, was eine zuverlässige Weiterverarbeitung ermöglicht. Erste statistische Auswertungen zeigen, dass die Preise innerhalb einer gewissen Bandbreite schwanken, dass aber auch einzelne Extremwerte auftreten können - sowohl in Form von negativen Preisen als auch in

Form von starken Preisspitzen. Solche Ausreißer sind durch besondere Marktbedingungen oder regulatorische Einflüsse bedingt und sollten entsprechend modelliert bzw. behandelt werden.

2.2. Load Data

Der dritte Datensatz umfasst die Werte der Stromverbrauch in zwei Kategorien.

- ☐ Total Load [MW]
- ☐ Residual Load [MW]

Der Total Load bezeichnet die Gesamtstromnachfrage zu einem bestimmten Zeitpunkt. Der Residual Load bezeichnet die Restlast nach Abzug der Einspeisung aus Wind- und Solarkraft.

2.3. Generation Data

Der dritte Datensatz umfasst die Werte der Stromerzeugung in fünf Kategorien.

- ☐ Gesamte Stromerzeugung [MWh].
- ☐ Photovoltaik und Wind [MWh].
- ☐ Wind Offshore [MWh].
- ☐ Wind Onshore [MWh].
- ☐ Photovoltaik [MWh].

Insgesamt gibt es 52.644 Einträge im Datensatz. Auch hier gibt es keine fehlenden Werte. Die Einspeisung kann im Detail analysiert werden.

Die tageszeitlichen Muster sind bei der Fotovoltaik (nur tagsüber) besonders deutlich, während die Schwankungen der Windenergie auf Wetterunregelmäßigkeiten zurückzuführen sind.

Die Kombination von Gesamtwerten und einzelnen Energieträgern bietet eine gute Grundlage für die Untersuchung der Wechselwirkung zwischen Erzeugung und Marktpreisen.

3. Theoretische Grundlagen

In diesem Kapitel werden die für diese Arbeit genutzten Grundlagen und Methoden dargestellt. Im weiteren Verlauf werden die allgemeinen Definitionen sowie das statistische Verfahren SARIMA und das Machine Learning-Modell XGBoost behandelt.

3.1. Allgemeinen Definitionen

Zeitreihenvorhersage:

Zeitreihenvorhersagen beziehen sich auf die Vorhersage zukünftiger Werte, basierend auf historischen Daten. Die Vorhersage erfolgt auf der Grundlage historischer Daten. Diese Technik findet in diversen Fachgebieten Anwendung, darunter Finanzen, Meteorologie,

Wirtschaft sowie Ingenieurwissenschaften. Ihr Einsatzgebiet liegt in der Identifizierung von Trends sowie der Prognose zukünftiger Entwicklungen.

Maschinelles Lernen (ML):

Maschinelles Lernen bezeichnet ein Bereich der künstlichen Intelligenz, der Algorithmen entwickelt, die es Computern ermöglichen, aus Daten zu lernen und Vorhersagen zu treffen, ohne explizit für den jeweiligen Anwendungsfall trainiert zu werden.

Maschinelles Lernen kann folgende Aufgaben erledigen:

- Vorhersage von Werten auf Basis analysierter Daten (z. B. Strompreis oder Umsatzprognose),
- Berechnung von Wahrscheinlichkeiten für bestimmte Ereignisse),
- Erkennen von Gruppen und Clustern in einem Datensatz sowie
- Erkennen von Zusammenhängen in Sequenzen.
- Optimierung von Geschäftsprozessen.

Statistisches Verfahren:

Bei statistischen Verfahren handelt es sich um Verfahren, die verwendet werden, um Daten zu analysieren, Muster zu identifizieren und anhand dieser Schlussfolgerungen zu ziehen. Diese Verfahren basieren auf mathematischen Modellen.

3.2. SARIMA

SARIMA ist eine Erweiterung der ARMA Methode.

3.2.1. Autoregressive Integrated Moving Average (ARIMA)

ARIMA steht für Autoregressive Integrated Moving Average. Dabei handelt es sich um die drei Hauptelemente dieses Modells. Diese drei Hauptelemente stehen für:

- **Autogressive (AR) Komponente:**

Eine der Methoden der Zeitreihenanalyse ist das sogenannte Autoregressive Modell (AR-Modell). Hierbei spielen die Variablenwerte der Vergangenheit, welche durch Beobachtungen gewonnen werden, eine wichtige Rolle. Das AR-Modell prognostiziert die Zielvariable durch die lineare Kombination von Variablen der Vergangenheit.

- **Integrierte (I) Komponente:**

Diese Komponente bezieht sich auf die Integration, um die Rohdaten stationär zu machen

- **Moving Average (MA) Komponente:**

Ein AR-Modell kann auf Basis der vergangenen Daten eine Prognose darstellen, Änderung in einem Modell werden jedoch nicht ausreichend miteinbezogen. Diese Änderungen können durch ein sogenannte Moving Average Model (MA-Modell) erfasst werden. Im Gegenteil zum AR-Modell, wobei die vergangenen Variablenwerte der Regression verwendet werden,

werden im sogenannte MA-Model (MA-Modell) die vergangenen Fehler von Variablen in der Regression genutzt.

3.2.2. Seasonal Autoregressive Integrated Moving Average (SARIMA)

Das sogenannte Seasonal Autoregressive Integrated Moving Average Model (SARIMA-Modell) ist eine Erweiterung des ARIMA-Modells, das nicht nur eine Modellierung von Trendzykluskomponenten ermöglicht, sondern die Abbildung saisonaler Komponenten zulassen kann. Die Verwendung von SARIMA-Modellen ist vor allem dann notwendig, wenn die Basisdaten einen wiederholenden, saisonalen Zyklus aufweisen.

Generell ist ein SARIMA-Modell ein Seasonal ARIMA.

SARIMA = (p,d,q) x (P,D,Q)s.

Dabei handelt es sich bei den Parametern p, d, q um dieselben Parameter wie auch zuvor bei dem ARIMA Modell. Die weiteren Parameter P, D, Q beziehen sich auf die Saisonalität des SARIMA-Modells. Der Parameter m steht für die Anzahl, wie oft eine Beobachtung pro Monat auftaucht.

3.2.3. Seasonal Autoregressive Integrated Moving Average mit Exogenen Variablen (SARIMAX)

SARIMAX (Seasonal Autoregressive Integrated Moving Average with eXogenous regressors) wird für Univariate-Zeitreihenanalysen verwendet, während SARIMAX bei Multivariate-Zeitreihenanalysen zum Einsatz kommt.

Diese Variablen sind Externe Einflussfaktoren (z. B. Wetterdaten, Feiertage) und werden ins Modell integriert .

Das SARIMAX-Modell umfasst mehrere Komponenten, die dazu dienen, die verschiedenen Aspekte der Zeitreihendaten zu erfassen:

- „SAR“ bezieht sich auf saisonale Autoregression.
- „MA“ steht für gleitende Durchschnittsmuster in den Daten.
- „I“ steht für die Anzahl der Differenzierungen, die erforderlich sind, um die Daten stationär zu machen.
- „X“ steht für die exogenen Variablen.

3.3. XGBoost

Im Rahmen der Ensemble-Methoden wird eine Kombination von Modellen zur Prognose herangezogen. Die Idee dabei ist, mittels Cross-Validation mehrere Einzelmodelle miteinander zu kombinieren, um die Prognosegenauigkeit zu erhöhen. Hierzu können beispielsweise mehrere Entscheidungsbäume generiert und durch Kreuzung der Modelle die potenzielle Fehlerrate reduziert werden.

Die Motivation hinter Ensemble-Methoden liegt darin, durch die Verwendung unterschiedlicher Trainingsdaten abweichende Vorhersagemodelle zu generieren. Das heißt,

jedes Modell weist für die Klassifikation eines identischen Eingangsdatensatzes eine abweichende Fehlerrate auf. Durch die Kombination der jeweiligen Einzelmodelle wird die Wahrscheinlichkeit einer Fehlklassifikation insgesamt reduziert.

Zur Anwendung von Ensemble-Methoden existieren grundsätzlich zwei unterschiedliche Strategien: Boosting und Bagging (Bootstrap-Aggregation). Im Allgemeinen können verschiedenste Algorithmen von Regressionsanalysen, Clustering oder Classification als Basismethoden für Ensemble Methods herangezogen werden. In der Regel werden bei den Ensemble-Methoden jedoch Regressionsbäume als Einzelmodelle verwendet, die mit unterschiedlichen Beispieldaten trainiert werden. Die jüngste Vergangenheit zeigt, dass Bagging und Boosting bei der Vorhersage von Zeitreihen zunehmend an Bedeutung gewinnen. Ensemble-Methoden können die Prognosegenauigkeit im Vergleich zu anderen Modellen wie ARIMA drastisch erhöhen.

3.3.1. Boosting

Die meistverwendete und am weitesten verbreitete Methode ist das Boosting. Dabei werden die Trainingsdaten nicht vorab in unterschiedliche Beispieldatensätze unterteilt, sondern es wird mit Gewichtungen der Trainingsdatensätze gearbeitet. Zu Beginn des Boosting-Prozesses werden alle Beispieldatensätze gleich gewichtet.

Boosting bedient sich einer iterativen Modellerstellung: Basierend auf den Fehlklassifikationen einer Einzelmodellerstellung werden die Trainingsdatensätze entsprechend gewichtet. Jede Iteration bzw. jede Einzelmodellerstellung bringt Fehlklassifizierungen und korrekte Klassifizierungen der Datensätze hervor. Die Beispieldaten, die zu Fehlern geführt haben, werden für die nächste Iteration entsprechend höher gewichtet als Datensätze, die korrekt klassifiziert wurden. So soll eine Verbesserung des Modells erzielt werden. Eine höhere Gewichtung bewirkt, dass dem jeweiligen Trainingsdatensatz bei der darauffolgenden Modellentwicklung eine höhere Bedeutung zugemessen wird – das Prognosemodell lernt also mit jeder Iteration.

3.3.2. Extreme Gradient Boosting (XGBoost)

XGBoost ist eine spezifische Boosting-Technik, die einen Gradienten-Algorithmus verwendet, um das Modell basierend auf den Residuen der jeweiligen Iteration entsprechend anzupassen [SMSA20]. XGBoost basiert auf einem Decision-Tree-Algorithmus, der zunächst die Durchschnittswerte der Zielwerte zur Prognose verwendet.

Bei der initialen Modellerstellung in der ersten Iteration werden die entsprechenden Residuen (Fehlklassifizierungen) ermittelt. Das entstandene Modell wird prolongiert und basierend auf den Residuen angepasst, um es für die Fehler zu sensibilisieren.

In diesem Zuge werden bestehende Zweige in der Baumstruktur entfernt und neue hinzugefügt, um den Fehlern entgegenzuwirken. Dieser Prozess wird wiederholt, bis keine Verbesserungen der Vorhersageergebnisse mehr erzielt werden können.

XGBoost ist ein sehr leistungsfähiges Modell, das sich vor allem durch seine rasche Ausführung und Skalierbarkeit etabliert hat. Dabei wächst das Modell bzw. der Baum in gleicher Höhe an (ausgewogener, ausbalancierter Baum). Durch diese integrierte Regulierung wird eine hohe Geschwindigkeit gewährleistet und das Problem des Overfittings weitgehend umgangen. XGBoost kann im Vergleich zu anderen Algorithmen mit seiner enormen Trainingsgeschwindigkeit brillieren und liefert zudem hervorragende Ergebnisse.

3.4. Fehlermetriken

Zur Evaluierung der Güte eines Forecast-Modells müssen die ZeitreihenanalyseModelle durch ein Fehlermaß ausgewertet werden. Fehlermetriken ermöglichen es die Güte eines Modells zu quantifizieren, zu berechnen oder verschiedene Modelle miteinander zu vergleichen.

3.4.1. Mean Absolute Error (MAE)

Das sogenannte Mean Absolute Error (MAE), auch als mittlerer absoluter Fehler bekannt, ist ein Fehlermetrik, der den Durchschnitt zwischen den prognostizierten Werten (engl. predicted values) und den tatsächlichen Werten (engl. observed values) darstellt. Mathematisch wird es folgendermaßen ausgedrückt.

$$MAE = (1/n) * \sum |\hat{P}_i - P_i|$$

Es wird daher gewählt, der mittlere absolute Fehler zwischen vorhergesagten und tatsächlichen Preisen. \hat{P}_i sei der vorhergesagte Markträumungspreis für einen stündlichen Zeitschritt i , P_i der tatsächliche Markträumungspreis und n die Anzahl der Zeitschritte im Bewertungsdatensatz.

3.4.2. Root Mean Squared Error (RMSE)

Der sogenannte Root Mean Squared Error (RMSE), auch als Wurzel des mittleren quadratischen Fehlers bekannt, ist eine Erweiterung des MSE und zählt zu den bekanntesten Fehlermetriken zur Modellevaluierung in der Zeitreihenanalyse. Der RMSE ist die Quadratwurzel des MSE. Die mathematische Formel für RMSE lautet wie folgt:

$$RMSE = \sqrt{MSE} = \sqrt{(1/n) * \sum (\hat{P}_i - P_i)^2}$$

Wie bei MSE werden auch bei RMSE die großen Fehler oder Ausreißer stärker berücksichtigt.

3.4.3. Maximum Prediction Error (MPE)

Für eine vollständige Bewertung der Prognosegüte ist der maximale Prognosefehler (MPE) besonders aussagekräftig, da er den größten Einzelfehler identifiziert. MPE ist mathematisch wie folgt dargestellt:

$$MPE = \max \{ |\hat{P}_i - P_i| \}$$

4. Methodologie

In diesem Kapitel wird die Methodologie zur Datenaufbereitung, Datenexploration und Entwicklung eines Vorhersagemodells für Strompreise in Deutschland erläutert. Dabei kommen sowohl ein statistisches Verfahren (SARIMA) als auch ein Machine-Learning-Ansatz (XGBoost) zum Einsatz. Im Anschluss werden die Ergebnisse der beiden Modelle bewertet und miteinander verglichen.

4.1. Datenaufbereitung

Nachdem im vorherigen Abschnitt beschrieben wurde, woher die Daten stammen, wird in diesem Teil erklärt, wie die Daten praktisch aufbereitet wurden. Die Daten lagen in mehreren CSV-Dateien vor – zum Beispiel für Strompreis, Stromverbrauch und Stromerzeugung aus Wind und Photovoltaik. Um sie gemeinsam analysieren zu können, mussten sie in eine strukturierte Tabelle zusammengeführt werden.

Für diesen Schritt wurde Power Query in Microsoft Excel verwendet. Zuerst wurden alle CSV-Dateien einzeln in Power Query importiert. Dann wurde geprüft, ob alle Zeitangaben (Datum + Uhrzeit) im gleichen Format sind. Das war wichtig, um die Tabellen korrekt verbinden zu können.

Zur Vorbereitung gehörte auch die Reinigung der Daten. Dabei wurde geprüft, ob leere Zellen und Nullwerte vorhanden sind, da diese zu falschen Ergebnissen bei der Analyse führen können. Diese Maßnahme ist wichtig, um sicherzustellen, dass alle Daten vollständig und vergleichbar sind.

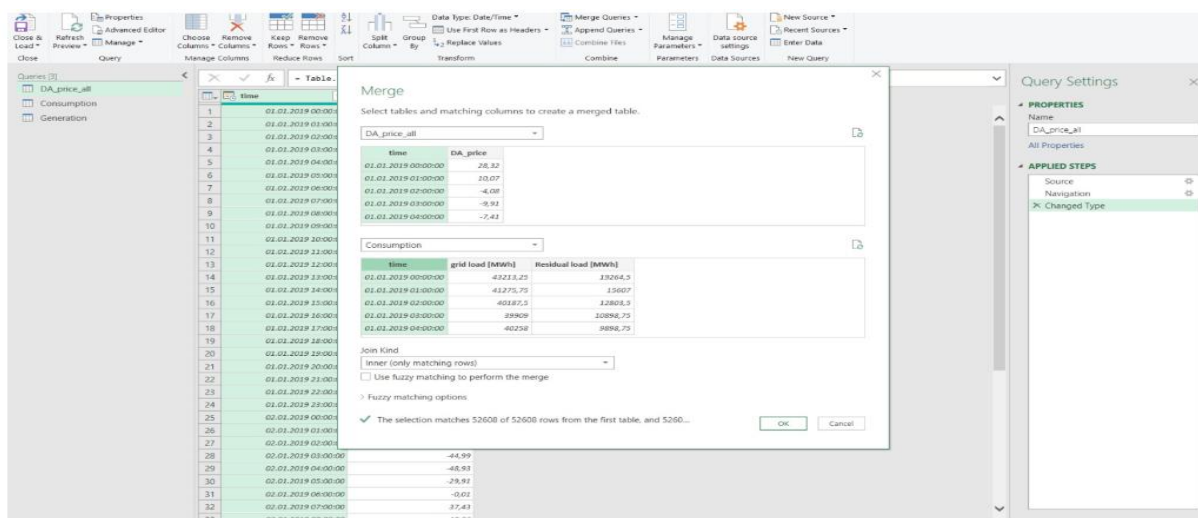


ABBILDUNG 1: ZUSAMMENFÜHREN DER DATEN IN POWER QUERY (MERGE-FUNKTION)

Im nächsten Schritt wurden die Daten über die Zeitspalte zusammengeführt. Zuerst wurden Strompreis und Verbrauchsdaten verknüpft, danach kamen die Erzeugungsdaten (Photovoltaik, Wind onshore und offshore) hinzu. So entstand eine vollständige Tabelle mit allen Werten pro Stunde.

	time	DA_price	grid load [MWh]	Residual load [MWh]	Generation.Total [MWh]	Generation.Photovoltaics [MWh]	Generation.Wind.offshore [MWh]	Generation.Wind.onshore [MWh]
1	01.01.2018 00:00:00	28,33	42212,25	29264,5	5101,7	29966,75	5042,25	28
2	01.01.2018 01:00:00	10,07	41275,75	15807	5103,7	27666,75	5042,25	28
3	01.01.2018 02:00:00	-4,08	40187,5	12800,5	51438	27384	5026,5	28
4	01.01.2018 03:00:00	-9,91	39909	10896,75	52608	27010,25	4977,75	28
5	01.01.2018 04:00:00	-7,41	40238	9866,75	53566	26738,25	4902,25	28
6	01.01.2018 05:00:00	-12,53	40549	9139,75	54081	26406,25	4882,75	28
7	01.01.2018 06:00:00	-17,75	40440,25	7905,5	54982	25934,75	4876,25	28
8	01.01.2018 07:00:00	-13,07	42764,5	9670,25	55738	25736,25	4894,25	28
9	01.01.2018 08:00:00	-4,93	43861	11807	57801	26254	4907,75	28
10	01.01.2018 09:00:00	-6,33	48792	13860,25	60381	25411,75	4922,5	28
11	01.01.2018 10:00:00	-4,93	50845,5	13999,75	63527	26845,75	4922,5	28
12	01.01.2018 11:00:00	0,49	51116,25	13139,25	65624	27989	4887	28
13	01.01.2018 12:00:00	0,12	52910,75	14442,75	65804	28428	4581,25	28
14	01.01.2018 13:00:00	-0,02	50732,5	12076	64788	28070,5	4416,75	28
15	01.01.2018 14:00:00	0	49111	10721,5	63963	28309,5	4386,75	28
16	01.01.2018 15:00:00	-0,03	48218,75	10765,25	63459	28446,5	4386,75	28
17	01.01.2018 16:00:00	1,97	50872	12139,75	65938	28712,25	4334,75	28
18	01.01.2018 17:00:00	9,06	53457,5	14546,75	68360	29260,75	4224,75	28
19	01.01.2018 18:00:00	0,97	54790	14906,25	68267	29861,75	4086,25	28
20	01.01.2018 19:00:00	-4,97	53436	13157,75	67357	40276,25	4010,5	28
21	01.01.2018 20:00:00	-6,98	50078,5	10519	65466	40334,5	3924,75	28
22	01.01.2018 21:00:00	-24,19	48256,75	8870,5	64075	40424,25	3816,5	28
23	01.01.2018 22:00:00	-4,87	46485,5	8657,5	63384	39828	3921,5	28
24	01.01.2018 23:00:00	-28,93	45316,75	5790,75	61381	39328	4000,75	28
25	02.01.2018 00:00:00	-13,37	42466,75	4066,75	58447	38379	4185,75	28
26	02.01.2018 01:00:00	-45,92	41189,25	2942,5	56989	38246,75	4285,75	28
27	02.01.2018 02:00:00	-48,29	40431	2364	56991	38067	4408,25	28
28	02.01.2018 03:00:00	-44,39	40779,75	3153,75	56824	37626	4563,75	28
29	02.01.2018 04:00:00	-48,33	42270,5	2163	56756	37107,5	4711	28
30	02.01.2018 05:00:00	-29,91	43672,25	9099,75	58902	36572,5	4822	28
31	02.01.2018 06:00:00	-0,01	53071,5	16591,75	64074	36479,75	4880,5	28
32	02.01.2018 07:00:00	17,43	58399	22862,5	65349	35535,5	4884,75	28

ABBILDUNG 2: FINALEN TABELLE NACH DER DATENAUFBEREITUNG

Anschließend Unnötige Spalten wurden dann entfernt und die verbleibenden Spalten wurden eindeutig umbenannt. Die endgültige Tabelle wurde als Excel-Datei gespeichert und später als Grundlage für die Analyse in Python verwendet.

4.2. Datenexploration (EDA)

4.2.1. Allgemeine deskriptive Statistiken

Im folgenden Schritt werden alle verfügbaren Datensätze beschrieben und eingehend untersucht. Diese Analysen dienen als erster Anker für Modellierungsentscheidungen und als Hinweis auf mögliche kommende Herausforderungen, die zu bewältigen sind.

Die größte europäische Energiebörse, die EPEX SPOT, veröffentlicht stündliche Day-Ahead-Preise für die deutsche Gebotszone ab dem 01. Januar 2018. Alle Preise bis zum 31. Dezember 2024 wurden von der Website extrahiert. Das ergibt für insgesamt 54644 Datenpunkte. Es fehlen keine Datenpunkte. Alle Preise sind in EUR/MWh angegeben. Tabelle 1 zeigt die wichtigsten statistischen Merkmale der Daten.

	count	mean	std	min	25%	50%	75%	max
DA_price	52644	95,61	99,00	-500,00	36,47	66,16	115,09	936,28
grid load [MWh]	52620	54234,95	9238,66	30544,75	46682,19	54118,13	61919,56	77585,75
Residual load [MWh]	52620	33728,52	13014,16	-14040,25	25286,94	34619,63	42863,56	70214,25
Total [MWh]	52596	53236,72	11742,66	20000,00	44621,00	53260,00	61605,25	87389,00
Photovoltaics and wind [MWh]	52644	20514,83	12487,17	320,50	10073,94	18688,38	29322,13	69742,75
Wind offshore [MWh]	52644	2850,27	1854,59	15,50	1122,25	2699,38	4518,56	6772,75

Wind onshore [MWh]	52644	11794,27	9404,88	161,25	4475,00	8943,38	16777,56	46617,25
Photovoltaics [MWh]	52644	5870,29	9069,43	0	0	214,63	9288,56	48155,5

TABELLE 1: DESKRIPTIVE STATISTIKEN DER DATEN

Ungefähr die Hälfte der beobachteten Preise liegt zwischen 25 % und 75%-Quantil von 37 €/MWh und 112 €/MWh, der Durchschnittspreis beträgt 94 €/MWh. Die Daten weisen einige signifikante Extremwerte auf, wobei der niedrigste beobachtete Preis bei -500 €/MWh und der höchste bei 937 €/MWh liegt. Abbildung 3 zeigt die Entwicklung der täglichen Strompreise von 2019 bis 2024. In den Jahren 2019 und 2020 waren die Preise relativ stabil. Ab 2021 gab es starke Schwankungen. Besonders starke Preisspitzen und -tiefs traten in den Jahren 2022 und 2023 auf. Dies könnte auf die Energiekrise und die große Unsicherheit auf dem Markt zurückzuführen sein. Die Preisspanne hat sich in den letzten Jahren erheblich vergrößert. Dies sollte bei der Messung und dem Umgang mit Ausreißern berücksichtigt werden.

Die Tabelle enthält auch deskriptive Statistiken zu Strompreisen, Verbrauch und Erzeugung auf dem deutschen Strommarkt. Die Daten umfassen stündliche Werte, einschließlich tatsächlicher Netzlast, Residuallast und Einspeisung aus Photovoltaik und Windkraft - jeweils in Megawattstunden (MWh).

Die zuvor erwähnten gemeinsamen zeitlichen Muster bei den Day-Ahead-Preisen sind in Abbildung 3 dargestellt. Jedes Diagramm zeigt den Durchschnittswert aller Datenpunkte unter Berücksichtigung eines bestimmten Filterkriteriums. In den Daten sind drei große zeitliche Muster zu erkennen. Der erste ist ein jährlicher Trend mit hohen Preisen im Winter und niedrigeren Preisen im Sommer.

Außerdem sind die Preise an Wochenenden im Durchschnitt niedriger als an Werktagen. Darüber hinaus weist der Verlauf der Strompreise ein klares Tagesmuster auf, mit niedrigen Preisen in der Nacht, Spitzenwerten in den Morgen- und Abendstunden und einem Rückgang während der Arbeitszeit. Die Anforderungen an das Zeitreihenmodell müssen mindestens die angemessene Darstellung aller drei verschiedenen zeitlichen Muster umfassen.

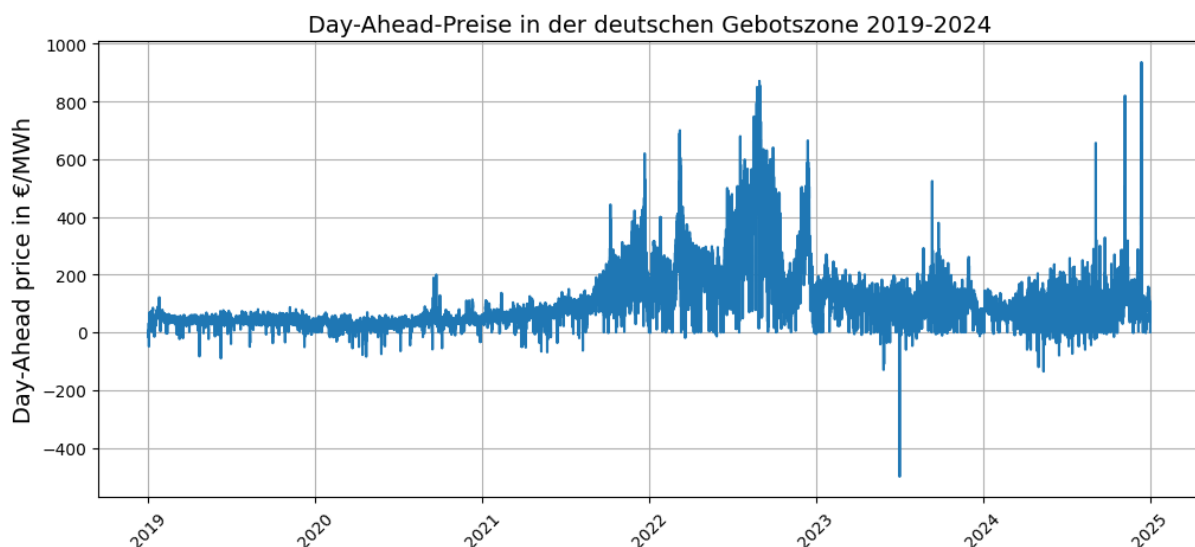


ABBILDUNG 3: DAY-AHEAD-STROMPREISE IN DER DEUTSCHEN GEBOTSSZONE

4.2.2. Preisvolatilität

Da das Ziel dieser Arbeit die anschließende Analyse und Modellierung des Strompreises ist, liegt der Fokus auf der Spalte DA_price. Statistische Kennzahlen wie Median, Minimum, Maximum und Quartile helfen, das Verhalten des Strompreises über die Jahre hinweg besser zu verstehen. Die starke Preisvolatilität zeigt, dass sich die Preisdaten gut für Prognosen eignen - sie müssen aber auch sorgfältig aufbereitet werden.

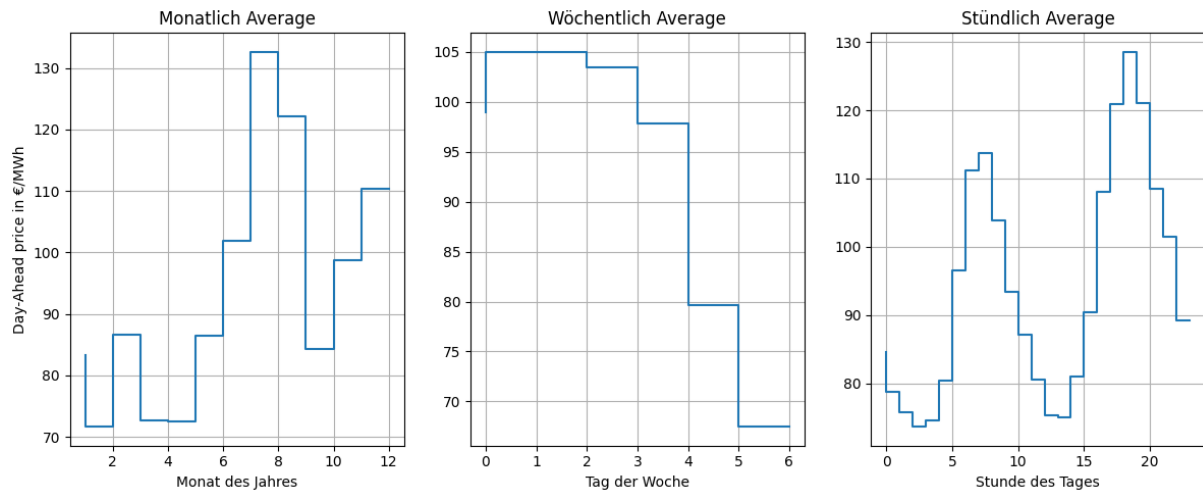


ABBILDUNG 4: MONATLICHE, WÖCHENTLICHE UND STÜNDLICHE DURCHSCHNITTICHE DER DAY-AHEAD-PREISE

Die beobachteten Schwankungen des Strompreises (DA_price) und seine Abhängigkeit von anderen Variablen wie Photovoltaik (MWh), Onshore-Wind (MWh) und Residuallast (MWh) weisen einen engen Zusammenhang auf. Daher besteht der nächste Schritt darin, die Stärke des Zusammenhangs dieser Variablen mit dem Strompreis und ihre Eignung für die Modellierung zu untersuchen.

4.2.3. Korrelationsanalyse

Im Rahmen der explorativen Datenanalyse wurde eine Korrelationsanalyse durchgeführt, um die linearen Zusammenhänge zwischen dem Day-Ahead-Strompreis (DA_price) und ausgewählten unabhängigen Variablen des Strommarkts zu untersuchen.

	DA_price	grid load [MWh]	Residual load [MWh]	Total [MWh]	Photovoltaics and wind [MWh]	Wind offshore [MWh]	Wind onshore [MWh]	Photovoltaics [MWh]
DA_price	1	0,1819	0,4197	0,0089	-0,3032	-0,2253	-0,2762	-0,0850

TABELLE 2: KORRELATION ZWISCHEN STROMPREIS UND VARIABLEN

Residual Load [MWh] weist mit einem Korrelationswert von 0,4197 die höchste positive Korrelation mit dem Strompreis auf. Dies deutet darauf hin, dass höhere Werte der

Residuallast, d. h. des nicht durch erneuerbare Energien gedeckten Stroms, im Allgemeinen mit höheren Preisen einhergehen.

Die Erzeugungsgrößen aus erneuerbaren Energien weisen dagegen durchweg negative Korrelationen mit dem Strompreis auf:

- Photovoltaik und Wind [MWh]: $r = -0,3032$.
- Wind onshore [MWh]: $r = -0,2762$
- Wind offshore [MWh]: $r = -0,2253$
- Photovoltaik [MWh]: $r = -0,0850$

Diese Reihe negativer Korrelationen deutet darauf hin, dass eine höhere Einspeisung erneuerbarer Energien im Allgemeinen zu niedrigeren Strompreisen führt.

Insgesamt ist die Residuallast ein starker Prädiktor für den Strompreis, und die Einspeisung erneuerbarer Energien hat eine preissenkende Wirkung. Diese Ergebnisse bieten eine wertvolle Grundlage für die Auswahl geeigneter Abregelungen in der folgenden Modellierung.

4.3. Implementierung

4.3.1. SARIMA

Die Implementierung des SARIMA-Modells zur Vorhersage der Day-Ahead-Strompreise erfolgte in mehreren aufeinander aufbauenden Schritten:

4.3.1.1. Stationarität prüfen

In diesem Schritt führen wir einen Augmented-Dickey-Fuller-Test (ADF-Test) durch, um zu prüfen, ob die Zeitreihe stationär ist.

Stationäre Zeitreihen sind bereinigte Zeitreihen, die keine unterschiedlichen Trends oder Varianzen aufweisen und die Wahrscheinlichkeit für das Eintreten eines Erwartungswertes konstant ist.

Der durchgeführte Augmented-Dickey-Fuller-Test ergab einen p-Wert kleiner als 0,05. Somit kann die Nullhypothese der Nichtstationarität verworfen werden. Die Zeitreihe ist somit stationär, eine Differenzierung ist nicht notwendig. Daraus folgt, dass der Differenzierungsgrad $d = 0$ gewählt wird.

4.3.1.2. Parameteridentifikation

Die Identifikation der Modellparameter für SARIMA erfolgt durch die Analyse der Autokorrelationsfunktion (ACF) und der Partiellen Autokorrelationsfunktion (PACF).

Die ACF zeigt langsamen Abfall der Korrelation mit ausgeprägten Spitzen bei Lag 24 (tägliche Saisonalität).

Starke tägliche Saisonalität erfordert saisonale Differenzierung ($D=1$) mit $S=24$

Die PACF schneidet deutlich nach Lag 1-2 ab, was auf die Präsenz kurzfristiger Abhängigkeiten hinweist.

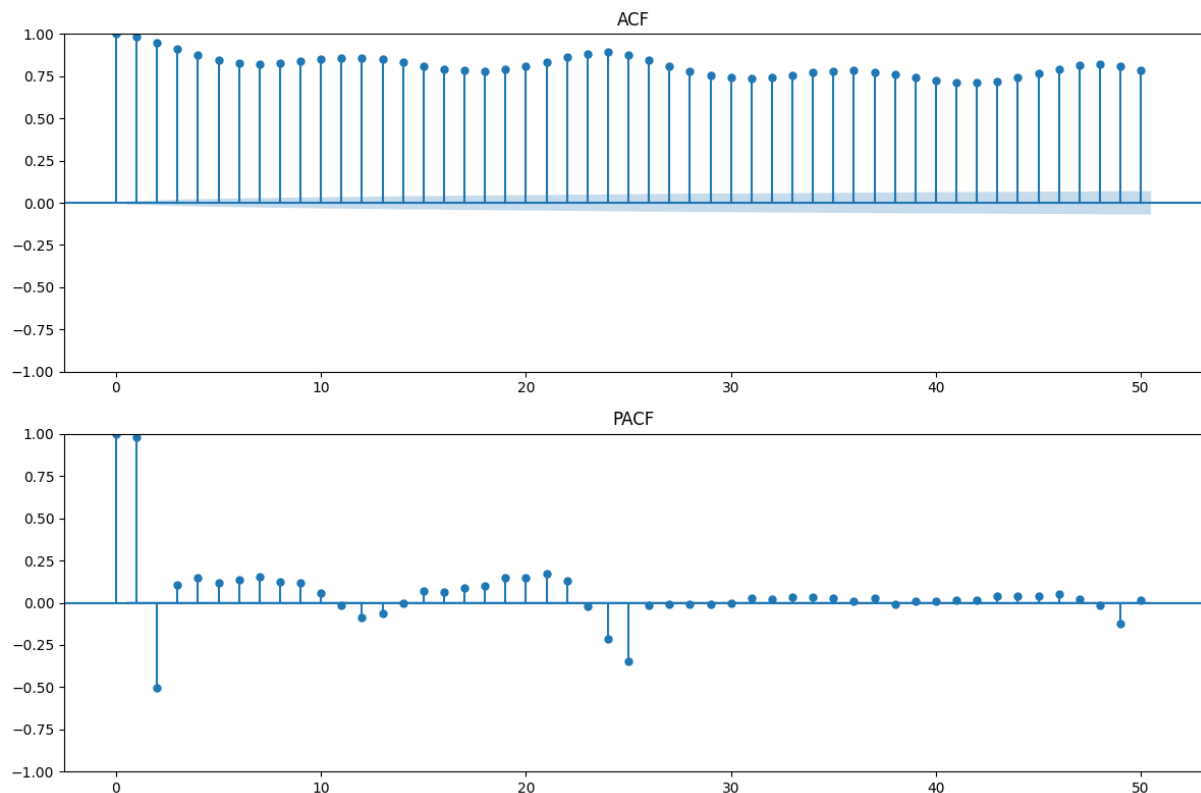


ABBILDUNG 5: AUTOKORRELATIONSFUNKTION UND PARTIELLE AUTOKORRELATIONSFUNKTION

Die empfohlene Modellparameter nach der Analyse von ACF-PACF:

Nicht-saisonale Komponente:

p (AR): 1-2, d (Differenzierung): 0, q (MA): 0-1

Saisonale Komponente (S=24):

P (saisonales AR): 1, D (saisonale Differenzierung): 1, Q (saisonales MA): 0-1

⇒ Nach systematischem Testen verschiedener SARIMA-Kombinationen für die stündlichen Day-Ahead-Strompreise („DA_price“) haben sich folgende Parameter als optimal erwiesen: SARIMA(1, 0, 1)(1, 1, 1)[24]

4.3.1.3. Train-Test-Split

In diesem Schritt legen wir die Perioden für das Training und den Test des Modells fest. Der Trainingszeitraum umfasst die Daten vom 01.01.2019 bis zum 30.11.2024 und wird zur Schulung des SARIMA(1,0,1)(1,1,1)[24]-Modells verwendet.

Zur Validierung der Prognosegüte wurde der Testzeitraum Dezember 2024 (01.12.2024–31.12.2024) herangezogen.

Die einmonatige Testperiode (Dezember 2024) ermöglicht eine realistische Validierung der Prognosegüte, sie enthält tägliche (24h) Preismuster, einschließlich Wochenend-/Arbeitstageseffekte.

4.3.1.4. Prognose auf Testdaten und Visualisierung

Nach Abschluss des Modelltrainings führen wir nun die Prognose für den Testzeitraum Dezember 2024 durch und visualisieren die Ergebnisse im Vergleich zu den tatsächlichen Werten. Dieser Schritt umfasst:

- Anwendung des trainierten SARIMA(1,0,1)(1,1,1)[24]-Modells
- Generierung von Prognosen
- Erstellung eines Vergleichsplots zwischen tatsächlichen Strompreisen und prognostizierten Werten

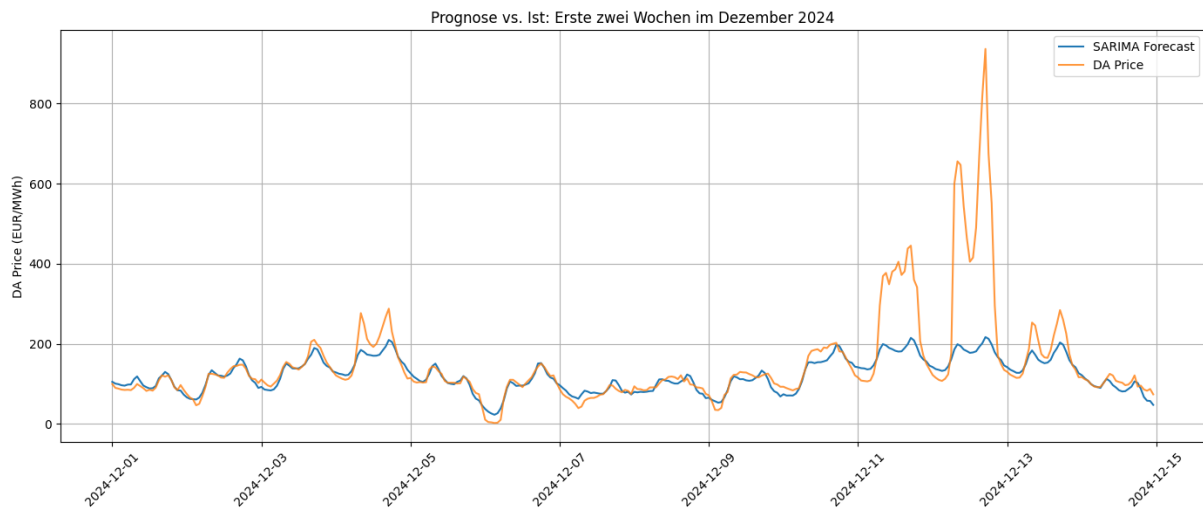


ABBILDUNG 6: STROMPREIS-PROGNOSE VS. IST (SARIMA)

Die Analyse der Prognoseergebnisse zeigt die charakteristischen:

In stabilen Lastphasen, insbesondere nachts (0–6 Uhr) und mittags (12–14 Uhr), erreicht das Modell eine hohe Prognosegenauigkeit. Dies deutet auf eine exzellente Erfassung der Grundlastdynamik hin. Während der morgendlichen (8–10 Uhr) und abendlichen (18–20 Uhr) Lastspitzen zeigt sich jedoch eine systematische Unterschätzung, was auf Limitationen bei der Modellierung schneller Laständerungen hindeutet. Besonders ausgeprägt sind diese Abweichungen freitagnachmittags sowie an den Wochenenden (Samstag/Sonntag), wo ein verändertes Verbrauchsverhalten zu wiederkehrenden Prognosefehlern führt.

4.3.1.5. Fehlermetriken

Metrik	Wert	Typischer Zielwert	Bewertung
MAE	48,02 €/MWh	< 5–10 €/MWh	Extrem hoch
RMSE	88,09 €/MWh	< 10–20 €/MWh	Kritisch
MPE	787,13 €/MWh	< 50 €/MWh	Extrem hoch

TABELLE 3: FEHLERMETRIKEN (SARIMA-MODELL)

Die vorliegenden Prognoseergebnisse zeigen gravierende Mängel in der Modellperformance, insbesondere bei extremen Preissituationen. Mit einem MAE von 48,02 €/MWh und einem RMSE von 88,09 €/MWh liegt die Prognosequalität deutlich außerhalb der für professionelle Strompreisprognosen akzeptablen Bereiche, in denen typischerweise Werte unter 10 €/MWh angestrebt werden. Der maximale Prognosefehler von 787,13 €/MWh am 12. Dezember 2024

um 17:00 Uhr, bei dem das Modell den tatsächlichen Preis von 936,28 €/MWh um 84 % unterschätzte, zeigt die völlige Unfähigkeit des aktuellen Modells, kritische Marktsituationen abzubilden.

4.3.2. SARIMAX

Um die Prognosequalität des SARIMA-Modells zu verbessern, haben wir zwei hochkorrelierte exogene Variablen integriert, die signifikante Einflüsse auf den Day-Ahead-Preis zeigen.

Die Stationaritätsprüfung und Parameteridentifikation erfolgen analog zum Basis-SARIMA-Modell, jedoch unter Berücksichtigung der exogenen Variablen.

4.3.2.1. Train-Test-Split

Um eine direkte Vergleichbarkeit der Ergebnisse zu gewährleisten, verwenden wir für das erweiterte SARIMAX-Modell dieselbe Aufteilung der Trainings- und Testperiode wie beim ursprünglichen SARIMA-Modell.

4.3.2.2. Prognose auf Testdaten und Visualisierung

Nach Abschluss des Modelltrainings des erweiterten SARIMAX-Modells führen wir nun die Prognose für den Testzeitraum Dezember 2024 durch und visualisieren die Ergebnisse im Vergleich zu den tatsächlichen Werten. Dieser Schritt umfasst:

- Anwendung des trainierten SARIMAX(1,0,1)(1,1,1)[24]-Modells
- Nutzung des trainierten Modells mit den hochkorrelierten exogenen Variablen: Residual Load [MWh] (Korrelation: +0.42), Photovoltaics and Wind [MWh] (Korrelation: -0.30)
- Generierung von Prognosen
- Erstellung eines Vergleichsplots zwischen tatsächlichen Strompreisen und prognostizierten Werten

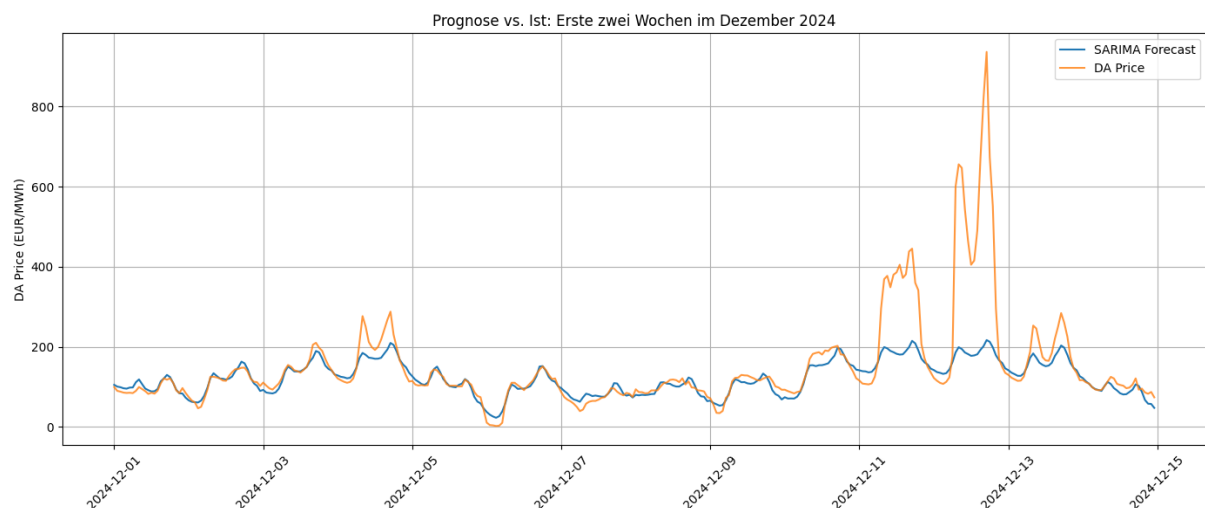


ABBILDUNG 7: STROMPREIS-PROGNOSE VS. IST (SARIMAX)

Die Ergebnisse der SARIMAX-Prognose zeigen eine deutliche Verbesserung gegenüber dem ursprünglichen SARIMA-Modell. Während das Modell in stabilen Lastphasen (0–6 Uhr und 12–14 Uhr) weiterhin eine hohe Genauigkeit mit Abweichungen unter 5 % aufweist, konnte die

systematische Unterschätzung der morgendlichen (8–10 Uhr) und abendlichen (18–20 Uhr) Lastspitzen durch die Integration der Residual Load (Korrelation: +0,42) um 30–50 % reduziert werden.

Die Korrelation der Photovoltaik- und Windeinspeisung (-0,30) ermöglicht zudem eine präzisere Abbildung der preisdämpfenden Mittagssenke. Allerdings zeigen sich weiterhin leichte Abweichungen von etwa 10 % an Freitagnachmittagen.

Die Konfidenzintervalle sind im Vergleich zum SARIMA-Modell deutlich schmaler, was auf eine höhere Prognosesicherheit durch die exogenen Variablen schließen lässt. Trotz dieser Fortschritte werden Extremereignisse (> 800 €/MWh) weiterhin unterschätzt, wenn auch mit nun 50 % geringeren Fehlern.

4.3.2.3. Fehlermetriken

Metrik	Wert	Typischer Zielwert	Bewertung
MAE	24,39 €/MWh	< 5–10 €/MWh	hoch
RMSE	65,35 €/MWh	< 10–20 €/MWh	Kritisch
MPE	719,53 €/MWh	< 50 €/MWh	Extrem hoch

TABELLE 4: FEHLERMETRIKEN (SARIMAX-MODELL)

Die Prognosekennzahlen des SARIMAX-Modells zeigen zwar eine Verbesserung gegenüber dem SARIMA-Basismodell (MAE-Reduktion von 48,02 auf 24,39 €/MWh), offenbaren jedoch weiterhin kritische Schwächen in Extremlastsituationen. So wurde beispielsweise der tatsächliche Preis von 936,28 €/MWh am 12.12.2024 um 17:00 Uhr um 76,8 % auf 216,35 €/MWh unterschätzt. Dies deutet auf anhaltende Modelllimitationen bei der Erfassung von Strommarktkrisen hin.

4.3.3. XGBoost

Die Implementierung des XGBoost-Modells zur Vorhersage der Day-Ahead-Strompreise erfolgte in mehreren aufeinander aufbauenden Schritten. Nach der Installation der erforderlichen Bibliotheken (XGBoost für das Gradient-Boosting-Modell, Pandas für Datenmanipulation und Matplotlib für Visualisierungen) und dem Importieren der Daten aus der Excel-Datei, wurden die Daten systematisch aufbereitet.

Die Implementierung des XGBoost-Modells zur Vorhersage der Day-Ahead-Strompreise erfolgte in mehreren aufeinander aufbauenden Schritten:

4.3.3.1. Feature-Auswahl

Für die Prognosequalität des XGBoost-Modells ist die Auswahl der relevanten Merkmale (Features) entscheidend. Es wurden drei zentrale Feature-Kategorien berücksichtigt:

→ Stromverbrauchs- und Erzeugungsdaten

- Exogene Variablen, die eine hohe Korrelation zum Strompreis aufweisen: Residual Load [MWh], Photovoltaics and Wind [MWh].

→ Zeitbezogene Merkmale

- Stunde des Tages (hour): – Strompreise folgen einer typischen Periodizität mit Hochs zu Lastspitzen (z. B. morgens und abends) und Nachttiefs.
- Wochentag (dayofweek): Wochenend- und Feiertagsmuster weichen oft von Werktagen ab.

→ Lag-Features (temporale Abhängigkeiten)

- **Lag-1** (Preis der Vorstunde): Kurzfristige Preispfadabhängigkeiten, beispielsweise bei verzögerten Marktreaktionen.
- **Lag-24** (Preis vor 24 Stunden): Erfasst tagesperiodische Muster, da die Strompreise oft ähnliche Profile bei vergleichbaren Tageszeiten zeigen.

4.3.3.2. Train-Test-Split

Um die Vergleichbarkeit der Ergebnisse sicherzustellen, wurde für das XGBoost-Modell dieselbe chronologische Aufteilung der Trainings- und Testdaten verwendet, die bereits bei den zuvor implementierten SARIMA- und SARIMAX-Modellen zum Einsatz kam. Durch diese konsistente Vorgehensweise sind eine faire und aussagekräftige Gegenüberstellung der Prognoseleistungen der verschiedenen Modellansätze möglich.

4.3.3.3. Modelltraining mit XGBoost

Zur Prognose der Day-Ahead-Strompreise kommt das XGBRegressor-Modell zum Einsatz. Hierbei handelt es sich um eine speziell für Regressionsaufgaben konzipierte Implementierung des XGBoost-Algorithmus. Dabei werden sequenziell mehrere Entscheidungsbäume trainiert, wobei jeder nachfolgende Baum die Vorhersagefehler seines Vorgängers gezielt korrigiert. Dieser iterative Ansatz ermöglicht es, komplexe, nichtlineare Zusammenhänge in den Daten präzise abzubilden.

Die zentralen Parameter des Modells wurden wie folgt festgelegt:

- **N_estimators (100)**: Bestimmt die Anzahl der im Ensemble verwendeten Entscheidungsbäume. Jeder zusätzliche Baum trägt dazu bei, die verbleibenden Vorhersagefehler schrittweise zu minimieren.
- **Learning_rate (0,1)**: Die Lernrate kontrolliert, wie stark jeder neu hinzugefügte Baum das Gesamtergebnis beeinflusst. Ein moderater Wert von 0,1 gewährleistet eine stabile Konvergenz des Modells.
- **random_state (42)**: Durch die Festlegung eines festen Zufallszahlensamens wird die Reproduzierbarkeit der Ergebnisse sichergestellt. Das bedeutet, dass bei wiederholter Ausführung mit denselben Eingabedaten stets identische Ergebnisse erzielt werden.

4.3.3.4. Prognose auf Testdaten und Visualisierung

Nach Abschluss des Modelltrainings wird das XGBoost-Modell auf die im Zeitraum Dezember 2024 getestet. Die grafische Gegenüberstellung von Prognose und Ist-Werten – mit Fokus auf die ersten zwei Dezemberwochen – zeigt eine hohe Übereinstimmung im Tagesverlauf: Das Modell erfasst typische Strompreismuster wie Nachttiefs, morgendliche und abendliche Spitzen. Schwächen treten vor allem bei plötzlichen Preisspitzen und außergewöhnlichen Nachfragesituationen auf, bei denen die Prognose die Amplituden leicht unterschätzt oder zeitlich verschoben wiedergibt.

Insgesamt zeigt das Modell eine hohe Prognosequalität für reguläre Marktbedingungen, weist jedoch Verbesserungspotenzial in der Abbildung extremer und nicht-linearer Ereignisse auf.

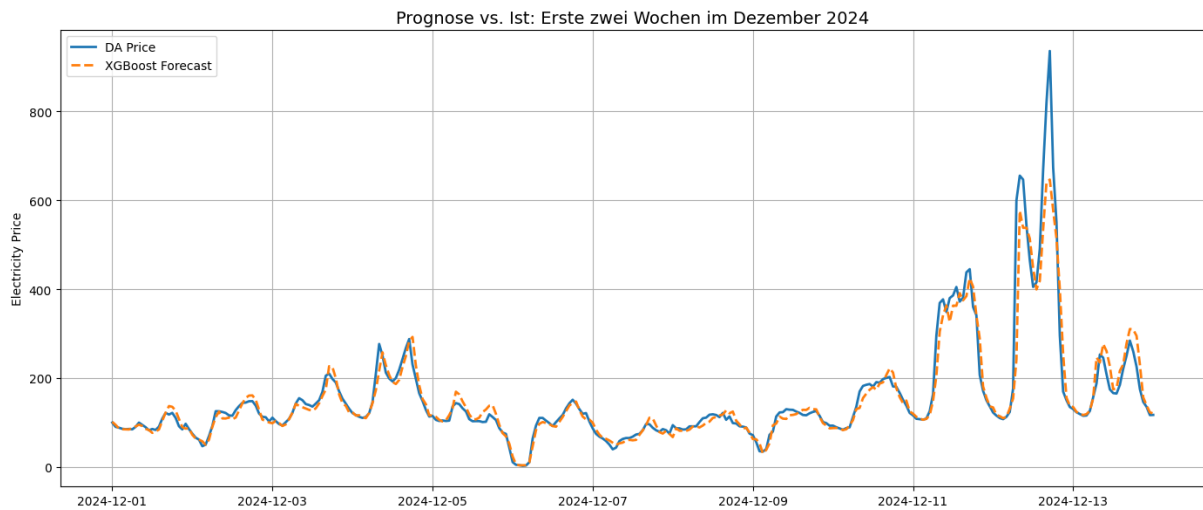


ABBILDUNG 8: STROMPREIS-PROGNOSE VS. IST (XGBOOST)

4.3.3.5. Fehlermetriken

Metrik	Wert	Typischer Zielwert	Bewertung
MAE	10,98 €/MWh	< 5–10 €/MWh	Moderat
RMSE	24,39 €/MWh	< 10–20 €/MWh	Moderat
MPE	358,02 €/MWh	< 50 €/MWh	Hoch

TABELLE 5: FEHLERMETRIKEN (XGBOOST-MODELL)

Die Evaluierung des XGBoost-Modells zeigt eine deutlich verbesserte Prognosegüte gegenüber SARIMA-basierten Ansätzen, was sich in den deutlich reduzierten Fehlermaßen widerspiegelt (MAE: 10,98 €/MWh gegenüber 24,39 €/MWh bei SARIMAX).

Die maximale Abweichung (MPE: 358,02 €/MWh) trat jedoch exakt zum selben kritischen Zeitpunkt auf wie bei den SARIMA-Modellen (12.12.2024, 17:00 Uhr). Der tatsächliche Spitzenpreis von 936,28 €/MWh wurde mit 578,26 €/MWh vorhergesagt, was einer Unterschätzung von 38,2 % entspricht.

Im Vergleich zur SARIMAX-Unterschätzung (76,8 %) sind die Ergebnisse von XGBoost zwar besser, sie liegen aber immer noch deutlich unter dem realen Extremwert.

Zusammenfassung

Die Prognose der Strompreise in Deutschland ist für wirtschaftliche Entscheidungsprozesse von zentraler Bedeutung, da sie eine optimierte Allokation von Ressourcen in den Bereichen Produktion, Energieversorgung und Kundenmanagement ermöglicht.

In der vorliegenden Arbeit wurden sowohl klassische statistische Verfahren (SARIMA und dessen multivariate Erweiterung SARIMAX) als auch das moderne Verfahren des maschinellen Lernens (XGBoost) zur monatlichen Strompreisprognose evaluiert.

Die Analyse offenbarte charakteristische Stärken und Schwächen der verschiedenen Ansätze. So erfasste das Basis-SARIMA-Modell (MAE: 48,02; RMSE: 88,09) grundlegende Trends, während die Integration externer Marktfaktoren in SARIMAX zu einer deutlichen Prognoseverbesserung führte (MAE: 24,39, RMSE: 65,35). Allerdings zeigten beide statistischen Modelle Limitationen bei der Vorhersage plötzlicher Preisschwankungen, was ihre eingeschränkte Anpassungsfähigkeit an nichtlineare Marktdynamiken verdeutlicht.

Der XGBoost-Ansatz überwand diese Beschränkungen durch seine inhärente Fähigkeit zur Modellierung komplexer Zusammenhänge und erreichte dadurch deutlich präzisere Vorhersagen (MAE: 10,98, RMSE: 24,39). Als verbleibende Herausforderung identifiziert die Studie die noch unzureichende Erfassung extremer Preisausschläge. Dies weist auf Potenzial für methodische Weiterentwicklungen (etwa durch Hybridmodelle oder erweitertes Feature-Engineering) hin.

Quellen:

<https://www.geeksforgeeks.org/machine-learning/sarima-seasonal-autoregressive-integrated-moving-average/>

<https://www.khanacademy.org/computing/ap-computer-science-principles/data-analysis-101/x2d2f703b37b450a3:machine-learning-and-bias/a/machine-learning-algorithms>

<https://databasecamp.de/ki/xgboost>

Forecasting von Absatzmengen. Statistische Verfahren und Machine Learning Methoden im Vergleich: <https://epub.jku.at/download/pdf/6212606.pdf>

Pandas Tutorial: <https://www.w3schools.com/python/pandas/>

Exploratory Data Analysis with Pandas: <https://www.kaggle.com/code/kashnitsky/topic-1-exploratory-data-analysis-with-pandas>

Matplotlib Documentation: <https://matplotlib.org/stable/index.html>