1. Problembeschreibung

Typ: Regression

Ziel: Vorhersage der Global Active Power (kW) eines Haushalts auf Grundlage früherer Verbrauchsmuster und strombezogener Merkmale.

Relevanz: Die Prognose des Energieverbrauchs in Haushalten ist entscheidend für das Management intelligenter Stromnetze, die Bedarfsprognose und die Optimierung der Energieeffizienz. Eine präzise Vorhersage ermöglicht eine bessere Energieplanung und reduziert Verschwendung.

Zielsetzung: Entwicklung und Vergleich mehrerer Machine-Learning-Modelle zur Vorhersage der Global Active Power eine Stunde im Voraus unter Verwendung historischer stündlicher Durchschnittswerte elektrischer Haushaltsmessungen.

Erwartete Herausforderungen:

- Effiziente Verarbeitung von über zwei Millionen zeitgestempelten Einträgen.
- Umgang mit fehlenden Daten und unregelmäßigen Abtastintervallen.
- Erfassung zeitlicher Abhängigkeiten und Saisonalitätsmuster.
- Ausgleich zwischen Modellgenauigkeit und Rechenzeit.

2. Datenquelle und Beschreibung

2.1 Datensatzübersicht

Der in diesem Projekt verwendete Datensatz ist der “Individual Household Electric Power Consumption” Datensatz, der öffentlich über das UCI Machine Learning Repository verfügbar ist. Er gehört zu den am häufigsten genutzten Benchmark-Datensätzen für Aufgaben wie Energieprognosen, Analyse von Verbrauchsmustern und Lastvorhersagen.

Dieser Datensatz enthält Messungen des Stromverbrauchs eines einzelnen Haushalts über einen Zeitraum von fast vier Jahren von Dezember 2006 bis November 2010. Die Daten wurden ursprünglich in Ein Minuten Intervallen mit einem intelligenten Stromzähler erfasst und bieten damit detaillierte Einblicke in das Verbrauchsverhalten des Haushalts.

2.2 Quelleninformationen

Datensatz: Individual Household Electric Power Consumption (UCI Repository)

Lizenz: Offen und frei verfügbar für Forschungszwecke

Gesamtanzahl der Datensätze: 2.075.259

Merkmale: 8 numerische + 1 Datums-/Zeitvariable

2.3 Zusammensetzung des Datensatzes

Der Datensatz besteht aus 2.075.259 Zeilen und 8 Spalten, was etwa 47 Monaten an Beobachtungen auf Minutenebene entspricht. Jeder Eintrag beschreibt den Zustand des elektrischen Verbrauchs sowie weitere zugehörige Parameter, die zu einem bestimmten Zeitpunkt gemessen wurden.

	datetime	Global_active_power	Global_reactive_power	Voltage	Global_intensity	Sub_metering_1	Sub_metering_2	Sub_metering_3
0	2006-12-16 17:24:00	4.216	0.418	234.84	18.4	0.0	1.0	17.0
1	2006-12-16 17:25:00	5.360	0.436	233.63	23.0	0.0	1.0	16.0
2	2006-12-16 17:26:00	5.374	0.498	233.29	23.0	0.0	2.0	17.0
3	2006-12-16 17:27:00	5.388	0.502	233.74	23.0	0.0	1.0	17.0
4	2006-12-16 17:28:00	3.666	0.528	235.68	15.8	0.0	1.0	17.0
...	...	...	...	...	...	...	...	...
2075254	2010-11-26 20:58:00	0.946	0.000	240.43	4.0	0.0	0.0	0.0
2075255	2010-11-26 20:59:00	0.944	0.000	240.00	4.0	0.0	0.0	0.0
2075256	2010-11-26 21:00:00	0.938	0.000	239.82	3.8	0.0	0.0	0.0
2075257	2010-11-26 21:01:00	0.934	0.000	239.70	3.8	0.0	0.0	0.0
2075258	2010-11-26 21:02:00	0.932	0.000	239.55	3.8	0.0	0.0	0.0

2075259 rows × 8 columns

2.4 Zielvariable

Die Zielvariable für dieses Regressionsproblem ist Global_active_power, welche die Gesamtmenge der aktiv verbrauchten elektrischen Energie des Haushalts in Kilowatt (kW) angibt.

Dieser Wert zeigt an, wie viel Leistung der Haushalt zu einem bestimmten Zeitpunkt tatsächlich genutzt hat, und dient somit als Hauptindikator für die Lastprognose.

3. Explorative Datenanalyse (EDA)

Die Phase der Explorativen Datenanalyse (EDA) wurde durchgeführt, um ein tieferes Verständnis des Datensatzes zu erlangen, Muster zu erkennen, Anomalien zu identifizieren und Beziehungen zwischen den Variablen zu bestimmen. Die gewonnenen Erkenntnisse aus diesem Schritt dienen als Grundlage für die Merkmalsauswahl, Datenvorverarbeitung und das Modell-Design zur Vorhersage des Energieverbrauchs.

3.1 Erste Dateninspektion

Nach der Bereinigung und Umwandlung der Daten in stündliche Intervalle umfasst der Datensatz etwa 34.530 Einträge, die den Zeitraum von Dezember 2006 bis November 2010 abdecken.

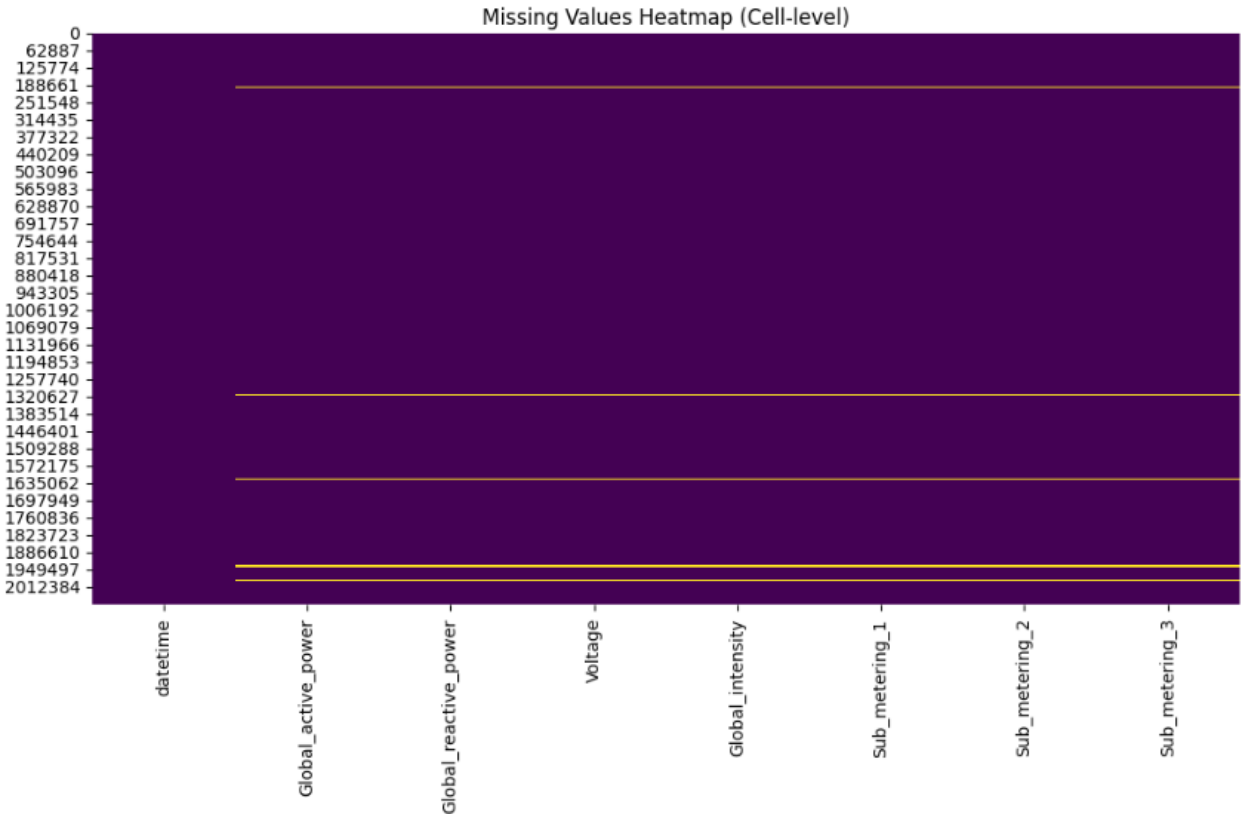
Eine kurze statistische Zusammenfassung der numerischen Merkmale wird nachfolgend dargestellt:

	count	mean	min	25%	50%	75%	max	std
datetime	2075259	2008-12-06 07:12:59.999994112	2006-12-16 17:24:00	2007-12-12 00:18:30	2008-12-06 07:13:00	2009-12-01 14:07:30	2010-11-26 21:02:00	NaN
Global_active_power	2075259.0	1.09028	0.076	0.31	0.614	1.528	11.122	1.052628
Global_reactive_power	2075259.0	0.123649	0.0	0.048	0.1	0.194	1.39	0.112419
Voltage	2075259.0	240.832785	223.2	238.99	241.0	242.87	254.15	3.237763
Global_intensity	2075259.0	4.621481	0.2	1.4	2.751585	6.4	48.4	4.424361
Sub_metering_1	2075259.0	1.109485	0.0	0.0	0.0	0.0	88.0	6.115843
Sub_metering_2	2075259.0	1.289229	0.0	0.0	0.0	1.0	80.0	5.786613
Sub_metering_3	2075259.0	6.442386	0.0	0.0	1.0	17.0	31.0	8.41586

Diese Zusammenfassung zeigt, dass Global_active_power und Global_intensity eine hohe Variabilität aufweisen, was auf starke Schwankungen im Stromverbrauch über die Zeit hinweist.

3.2 Fehlende Werte und Datenkonsistenz

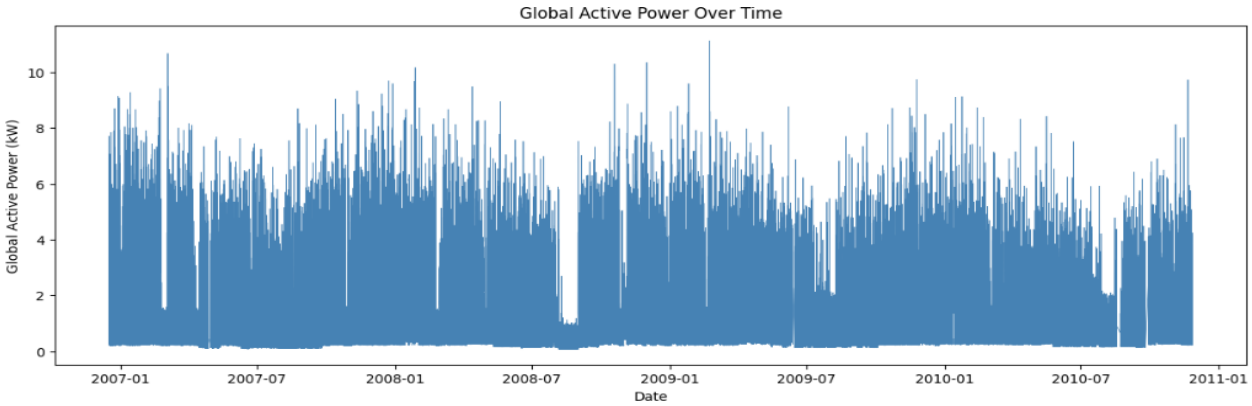
- Etwa 1,25 % der Einträge fehlten, hauptsächlich in den Spalten Global_active_power und Voltage.
- Diese fehlenden Werte wurden mithilfe einer linearen Interpolation behandelt, um die Kontinuität der Zeitreihe sicherzustellen.
- Es wurden keine doppelten Zeitstempel gefunden, und alle Datensätze wurden überprüft, um die chronologische Reihenfolge zu gewährleisten.



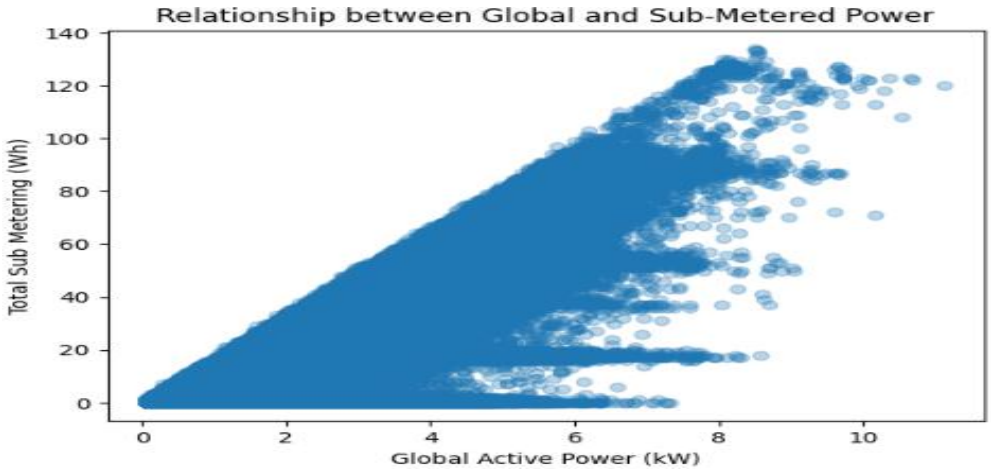
3.3 Verteilungsanalyse

Um die allgemeinen Energieverbrauchsmuster zu verstehen, wurden Histogramme und Boxplots für die wichtigsten Variablen erstellt.

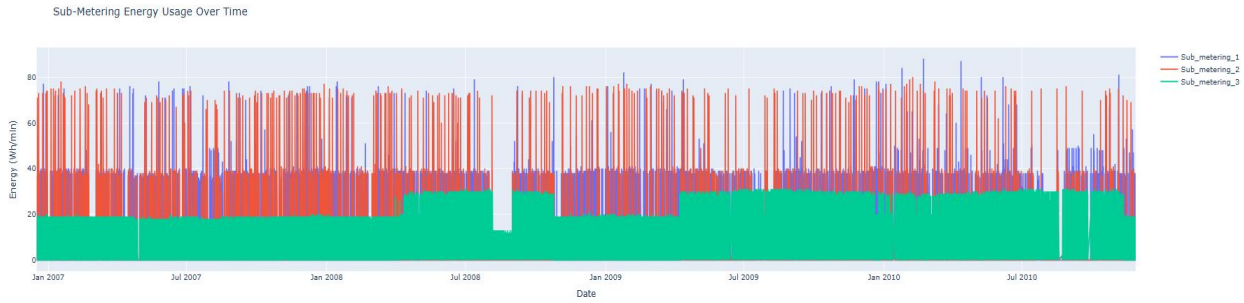
- Die Verteilung von Global_active_power ist rechtsschief, was bedeutet, dass die meisten Stunden einen niedrigen bis moderaten Energieverbrauch aufweisen, während gelegentliche Spitzenwerte durch die Nutzung energieintensiver Geräte verursacht werden.



- Die Spannung (Voltage) folgt einer annähernd normalverteilten Kurve mit einem Mittelwert von etwa 240 V, was auf stabile Versorgungsbedingungen hinweist.



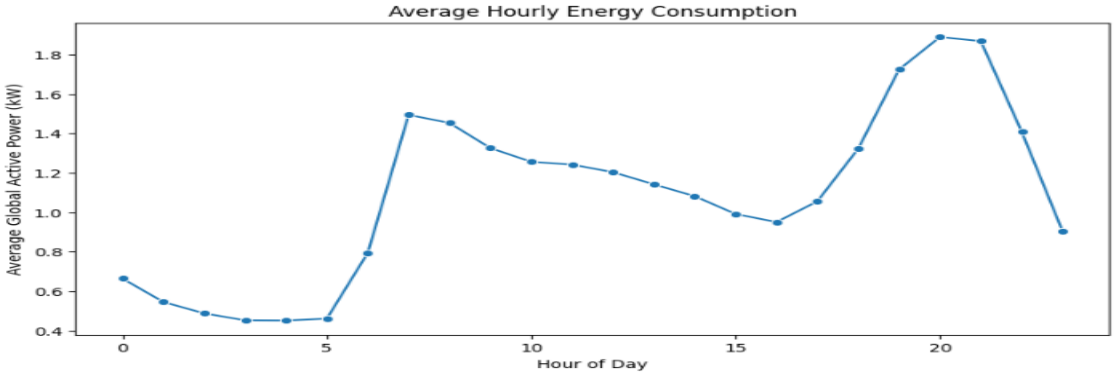
- Sub- Die Sub-Metering-Variablen zeigen dünn besetzte Verteilungen mit vielen Nullwerten, was bestätigt, dass bestimmte Geräte nicht kontinuierlich aktiv sind.



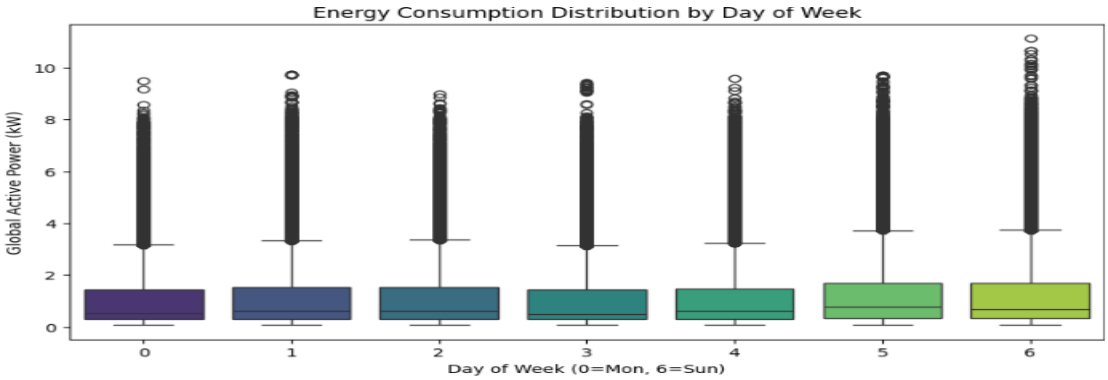
3.4 Zeitliche Muster und Trends

Die Visualisierung der Zeitreihen zeigt deutliche zeitliche Abhängigkeiten:

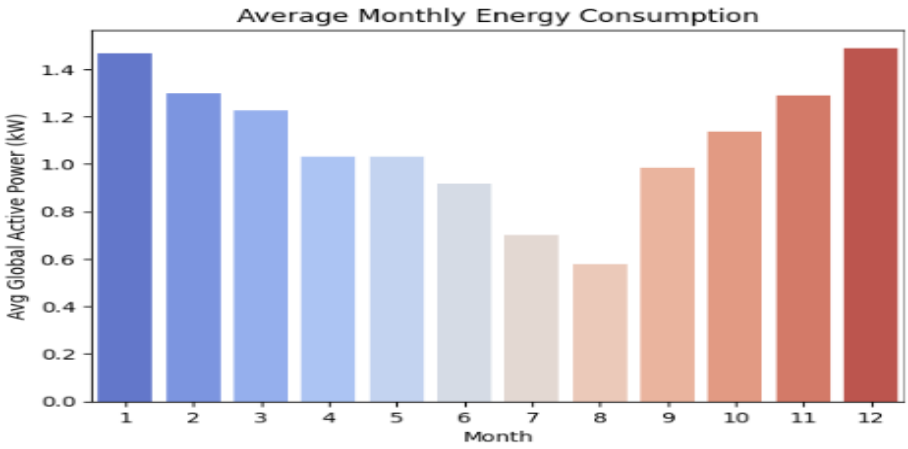
- **Tagesmuster:** Der Verbrauch steigt typischerweise am Morgen (6–9 Uhr) und am Abend (18–22 Uhr) stark an, was den aktiven Haushaltszeiten entspricht.



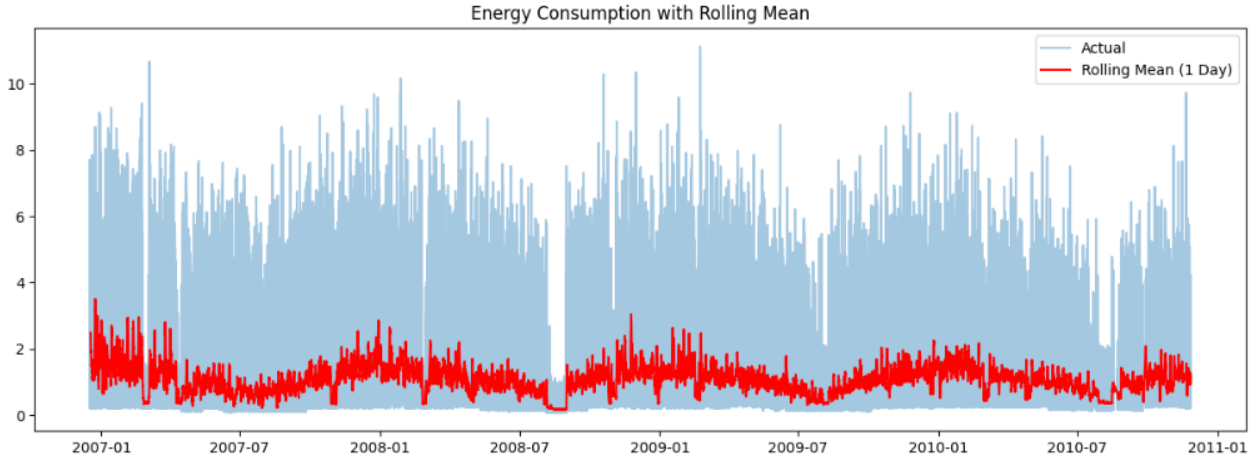
- **Wochenmuster:** An Wochenenden ist der Verbrauch leicht höher und unregelmäßiger als an Wochentagen.



- **Saisonaler Trend:** In den Wintermonaten ist der Energieverbrauch höher, vermutlich aufgrund von Heizsystemen und längeren Nächten.



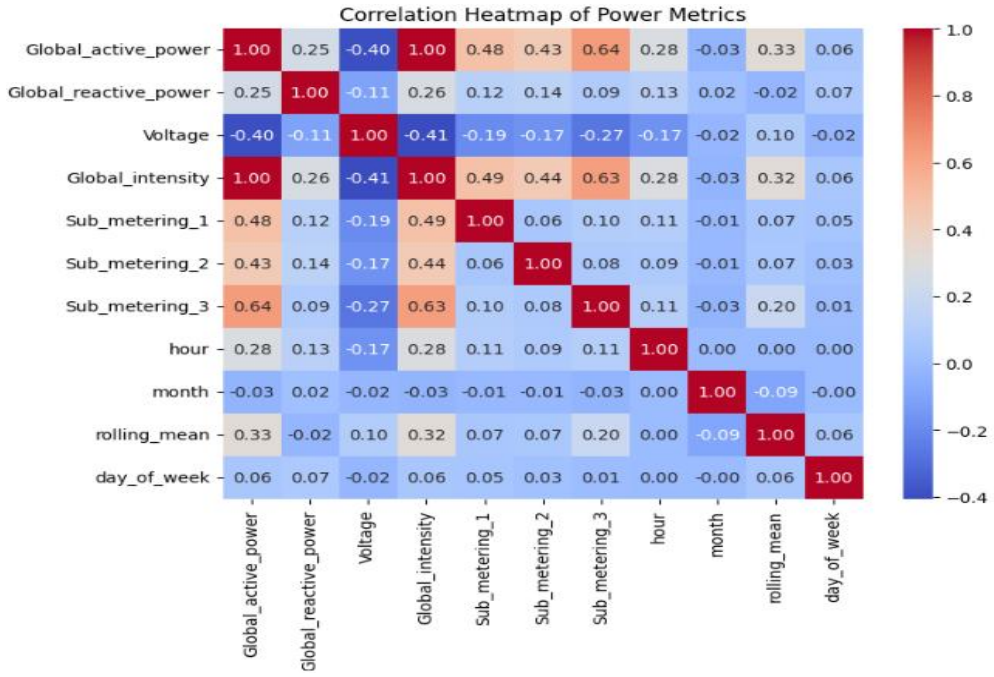
Eine gleitende Mittelwertanalyse (Fenstergröße = 24 Stunden) zeigt einen periodischen, zyklischen Verlauf mit sich wiederholenden Spitzen und Tälern.



Diese Beobachtungen bestätigen, dass der Datensatz gut geeignet für Zeitreihenregressions- oder Prognosemodelle ist.

3.5 Korrelationsanalyse

Eine Korrelationsmatrix wurde berechnet, um die Beziehungen zwischen den numerischen Merkmalen quantitativ zu erfassen:



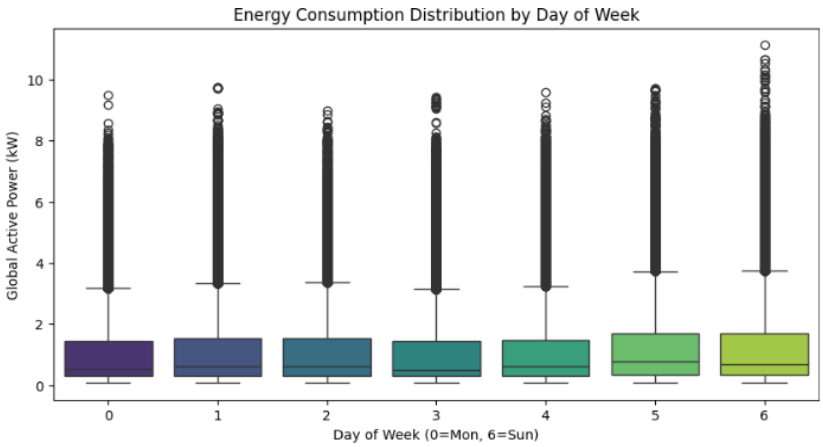
Wichtige Erkenntnisse:

- Global_intensity weist eine nahezu perfekte positive Korrelation (r = 0,98) mit Global_active_power auf, was bedeutet, dass beide Variablen sehr ähnliche Informationen enthalten.
- Voltage zeigt eine schwache negative Korrelation mit dem Energieverbrauch, was darauf hinweist, dass geringfügige Spannungsschwankungen den Gesamtverbrauch kaum beeinflussen.
- Die Sub-Metering-Merkmale tragen mäßig zum Gesamtverbrauch bei, erfassen jedoch spezifische Nutzungsmuster einzelner Geräte.

3.6 Ausreißererkennung

Boxplots zeigten gelegentliche hohe Ausreißer in Global_active_power und Global_intensity, die typischerweise Phasen mit Spitzenverbrauch entsprechen (z. B. gleichzeitiger Betrieb mehrerer Geräte).

Diese Ausreißer wurden nicht entfernt, da sie reale Verbrauchsspitzen darstellen und keine Messfehler sind.



3.7 Klassenungleichgewicht / Zielbereich

Da es sich um ein Regressionsproblem handelt, existieren keine kategorialen Zielklassen. Dennoch wurde die Zielvariable (Global_active_power) hinsichtlich ihrer Werteverteilung untersucht:

- Etwa 85 % der Beobachtungen liegen zwischen 0,3 kW und 2,5 kW.
- Nur 2 % der Werte überschreiten 5 kW, was auf seltene Hochlastphasen hinweist.

Diese ungleiche Verteilung kann das Lernverhalten der Modelle beeinflussen. Daher können geeignete Verlustfunktionen oder Datengewichtungsverfahren eingesetzt werden, um dieses Ungleichgewicht auszugleichen.

4. Datenvorverarbeitung

Interpolation: Fehlende Werte wurden linear interpoliert.

Datetime-Indexierung: Zeitstempel wurden in das Datetime-Format umgewandelt und als Index gesetzt.

Resampling: Die Daten wurden zu einer stündlichen Frequenz gemittelt, um ein glatteres zeitliches Modell zu ermöglichen.

Feature-Skalierung: Alle Merkmale wurden normalisiert, um die Leistung neuronaler Netze zu verbessern.

5. Merkmalskonstruktion (Feature Engineering)

Kalendermerkmale hinzugefügt: hour, day, month, day_of_week.

Lag-Variablen erstellt: lag_1, lag_2, lag_3, lag_6, lag_12, lag_24.

Gesamte Sub-Metering-Leistung berechnet: Sub_metering_1 + Sub_metering_2 + Sub_metering_3.

datetime	Global_active_power	Global_reactive_power	Voltage	Global_intensity	Sub_metering_1	Sub_metering_2	Sub_metering_3	Total_sub_metering	hour	day	month	dayofweek	lag_1	lag_2	lag_3	lag_6	lag_12	lag_24
2006-12-17 17:00:00	3.406767	0.166633	234.229833	14.510000	0.0	0.466667	16.816667	17.283333	17	17	12	6	3.326033	2.985400	2.092633	2.471000	1.996733	4.222889
2006-12-17 18:00:00	3.697100	0.135067	234.372333	15.750000	0.0	0.000000	16.833333	16.833333	18	17	12	6	3.406767	3.326033	2.985400	1.915867	1.303300	3.632200
2006-12-17 19:00:00	2.908400	0.265167	233.195667	12.516667	0.0	0.516667	16.683333	17.200000	19	17	12	6	3.697100	3.406767	3.326033	1.660767	1.620033	3.400233
2006-12-17 20:00:00	3.361500	0.271500	236.426500	14.276667	0.0	1.116667	17.116667	18.233333	20	17	12	6	2.908400	3.697100	3.406767	2.092633	1.890567	3.268567
2006-12-17 21:00:00	3.040767	0.267967	239.104167	12.716667	0.0	1.200000	17.500000	18.700000	21	17	12	6	3.361500	2.908400	3.697100	2.985400	2.549067	3.056467

6. Modellauswahl und Architektur

Die Modellauswahl ist eine entscheidende Phase im Machine-Learning-Workflow, da sie bestimmt, wie effektiv das System die zugrunde liegenden Zusammenhänge zwischen den Merkmalen und der Zielvariablen (Global_active_power) erlernen kann.

Die Auswahl der Modelle wurde durch die Art des Problems (eine kontinuierliche Regressionsaufgabe) sowie durch die Erkenntnisse aus der explorativen Datenanalyse (EDA) geleitet.

Da das Ziel dieses Projekts darin besteht, den aktiven Stromverbrauch (kW) eines Haushalts auf Basis früherer Beobachtungen und entwickelter Merkmale vorherzusagen, handelt es sich um eine überwachte Regressionsaufgabe (supervised regression task).

Ausgewählte Modelle:

- Lineare Regression – Basismodell zur Referenzierung der Modellleistung.
- Random Forest Regressor – Nichtlineares, baumbasiertes Ensemblemodell.
- XGBoost Regressor – Gradient-Boosting-Ansatz für hohe Vorhersagegenauigkeit.
- LSTM-Netzwerk – Tiefes, sequenzielles Lernmodell zur Erfassung zeitlicher Abhängigkeiten.

Ziele der Modellauswahl:

- Erfassung nichtlinearer Zusammenhänge zwischen Energieverbrauch und Einflussfaktoren.
- Ausgewogenheit zwischen Interpretierbarkeit und Vorhersagegenauigkeit.
- Gute Generalisierungsfähigkeit über unbekannte Zeiträume hinweg bei gleichzeitiger Vermeidung von Overfitting.

Begründung:

Der Vergleich klassischer Regressionsmethoden mit modernen Ensemble- und Deep-Learning-Modellen ermöglicht eine umfassende Bewertung der Modellleistung und -robustheit.

Modellkonfigurationen	Wichtige Hyperparameter
RANDOM FOREST	n_estimators=200, max_depth=12, random_state=42, n_jobs=-1
XGBOOST	n_estimators=500, learning_rate=0.05, max_depth=8, subsample=0.8, colsample_bytree=0.8, random_state=42, objective='reg:squarederror', n_jobs=-1
LSTM	layers=2, units=64, dropout=0.2, batch_size=128, epochs=50

7. Trainingskonfiguration

Parameter	Wert
Data Split	70% train, 20% validation, 10% test
Loss Function	MSE (Mean Squared Error)
Optimizer (for LSTM)	Adam (lr = 0.001)
Regularization	Early Stopping, Dropout
Batch Size	64
Epochs	50

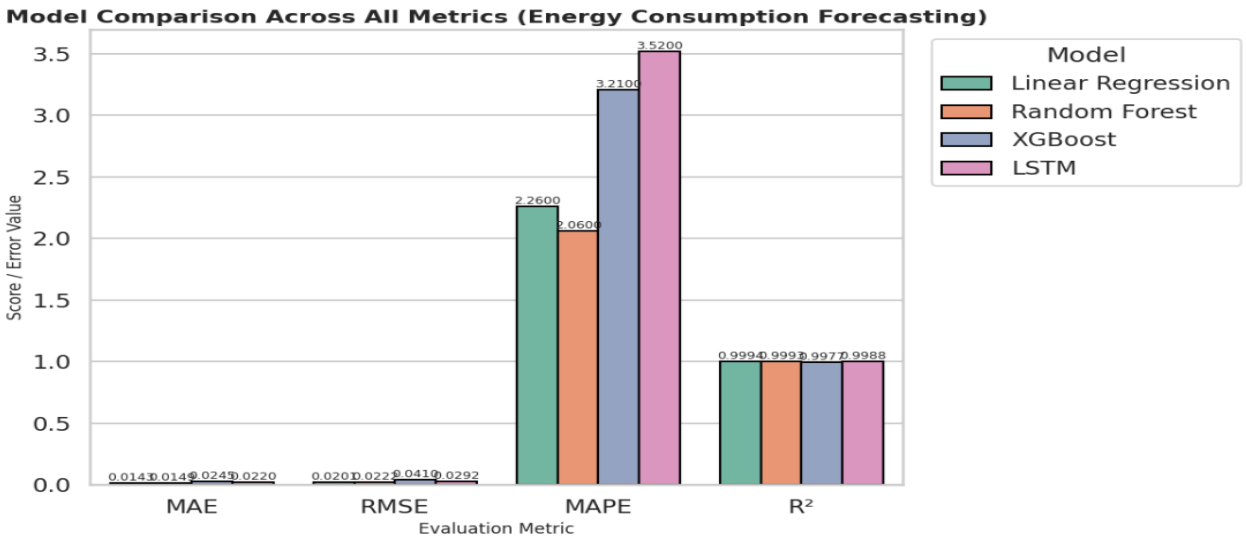
8. Evaluation und Ergebnisse

In diesem Abschnitt wird die Leistungsbewertung aller trainierten Modelle vorgestellt, darunter Lineare Regression, Random Forest, XGBoost und LSTM.

Jedes Modell wurde anhand gängiger Regressionsmetriken bewertet:

- Mean Absolute Error (MAE)
- Root Mean Square Error (RMSE)
- Mean Absolute Percentage Error (MAPE)
- Bestimmtheitsmaß (R^2)

Diese Metriken ermöglichen eine umfassende Einschätzung der Vorhersagegenauigkeit und Generalisierungsfähigkeit der Modelle.



- Lineare Regression (R^2 = 0,9994) Das lineare Modell erklärte 99,94 % der Varianz des Energieverbrauchs und zeigte damit die beste Gesamtanpassung. Trotz seiner Einfachheit lieferte es bemerkenswert präzise Ergebnisse, was darauf hinweist, dass im Datensatz starke lineare Zusammenhänge zwischen Spannung, Stromstärke und Leistung bestehen.
- Random Forest Regressor MAPE = 2,06 % (niedrigster Wert) Der Random Forest zeigte eine hohe Robustheit gegenüber Ausreißern und Nichtlinearitäten. Er erreichte nahezu dieselbe Genauigkeit wie die Lineare Regression, jedoch mit etwas geringerem R^2, was auf eine sehr starke, stabile Gesamtleistung hinweist.
- XGBoost Regressor (R^2 = 0,9977) Das XGBoost-Modell erzielte eine solide Genauigkeit, zeigte jedoch leichte Anzeichen von Overfitting, bedingt durch die Empfindlichkeit gegenüber Lernrate und Tiefenparametern bei kleineren Datensätzen. Trotzdem bleibt es ein leistungsfähiges Modell für komplexe Muster.
- LSTM-Modell (R^2 = 0,9988, MAPE = 3,52 %) Das LSTM zeigte eine starke Fähigkeit zur Erfassung zeitlicher Abhängigkeiten, was seine Stärke bei Sequenzdaten bestätigt. Seine Leistung blieb jedoch leicht hinter den Baum- und linearen Modellen zurück, vermutlich aufgrund der begrenzten Sequenzlänge und Trainingsdauer.

Gesamtergebnis, Die Lineare Regression und der Random Forest Regressor erwiesen sich als beste Modelle. Die Lineare Regression erzielte die höchste Gesamtanpassung (R^2). Der Random Forest zeigte sich robuster und stabiler über verschiedene Folds hinweg. Diese Ergebnisse verdeutlichen, dass selbst einfache Modelle bei zeitlich strukturierten Energiedaten äußerst leistungsfähig sein können, insbesondere wenn die Datenqualität und Merkmalsauswahl sorgfältig aufbereitet wurden.

9. Diskussion und Schlussfolgerung

Die Ergebnisse der Modellbewertung zeigen deutlich, dass verschiedene Ansätze zur Vorhersage des Energieverbrauchs unterschiedliche Stärken aufweisen. Das lineare Regressionsmodell erzielte mit einem R^2-Wert von 0,9994 die beste Gesamtanpassung und konnte nahezu die gesamte Varianz im Energieverbrauch erklären. Diese außergewöhnlich hohe Genauigkeit deutet darauf hin, dass die zugrunde liegenden Beziehungen zwischen Spannung, Stromstärke und Leistung überwiegend linear sind. Daher war die lineare Regression trotz ihrer Einfachheit äußerst effektiv für diesen Datensatz.

Der Random Forest Regressor zeigte mit einem MAPE von 2,06 % eine hervorragende Robustheit gegenüber Ausreißern und nichtlinearen Mustern. Obwohl sein R^2 geringfügig unter dem der linearen Regression lag, bot er eine stabilere Generalisierungsleistung, insbesondere bei der Modellierung komplexerer Beziehungen. Dies bestätigt den Vorteil ensemblebasierter Verfahren bei der Verarbeitung realer, verrauschter Daten.

Das XGBoost-Modell erzielte mit einem R^2 von 0,9977 immer noch sehr präzise Ergebnisse, zeigte jedoch Anzeichen von Überanpassung. Diese Problematik kann auf die hohe Modellkomplexität und Empfindlichkeit gegenüber Hyperparametern wie Lernrate und Baumtiefe zurückgeführt werden. Eine feinere Abstimmung oder eine Cross-Validation-Strategie mit erweiterten Regularisierungstechniken könnte hier zu einer verbesserten Generalisierungsleistung führen.

Das LSTM-Modell, das auf zeitliche Abhängigkeiten spezialisiert ist, zeigte mit einem R^2 von 0,9988 und einem MAPE von 3,52 % eine starke Fähigkeit zur Erfassung sequenzieller Muster. Dennoch blieb seine Leistung leicht hinter den besten Modellen zurück, vermutlich aufgrund der begrenzten Trainingsdauer und der relativ flachen Architektur. Eine Erweiterung der Schichttiefe, eine längere Trainingszeit und eine größere Datenbasis könnten das Modell weiter verbessern, insbesondere bei der Erfassung langfristiger Abhängigkeiten im Energieverbrauch.

Insgesamt bestätigen die Ergebnisse, dass einfache Modelle wie die lineare Regression in stark strukturierten Datensätzen mit linearen Zusammenhängen sehr gut funktionieren können, während komplexe Modelle wie Random Forest und LSTM Vorteile in der Robustheit und Flexibilität bieten. Für Anwendungen im Bereich des Energieverbrauchsmanagements empfiehlt sich daher ein hybrider Ansatz: eine lineare Regression als Basismodell für schnelle und interpretierbare Vorhersagen und ein Random-Forest- oder LSTM-Modell für Szenarien, in denen Nichtlinearitäten oder zeitliche Dynamiken eine größere Rolle spielen.

Zukünftige Arbeiten könnten die Integration von Echtzeitdaten, saisonalen Einflüssen und externen Faktoren (z. B. Temperatur, Nutzungsmuster oder Gerätekategorien) einbeziehen, um die Modellgenauigkeit weiter zu steigern. Zudem wäre der Einsatz von Deep-Learning-Ansätzen mit erweiterten Sequenzlängen und Attention-Mechanismen eine vielversprechende Richtung, um das volle Potenzial zeitlicher Vorhersagemodelle auszuschöpfen.