Tonnage-Prediction-AI

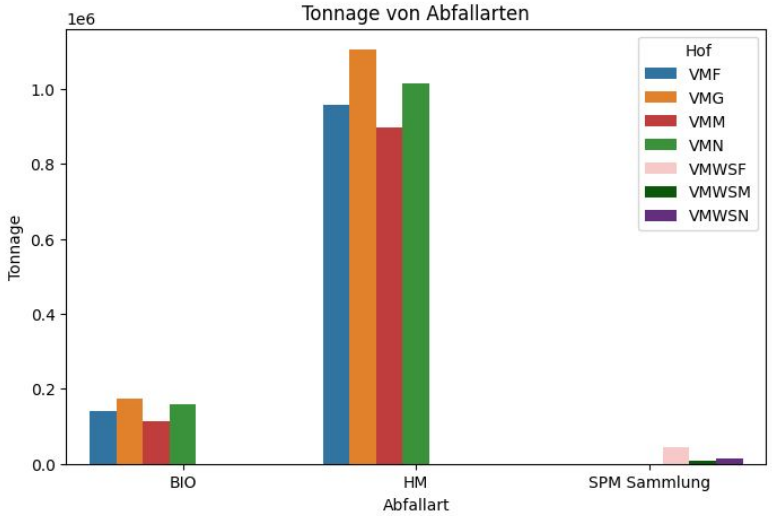
Abschlusspräsentation

Agenda

1. Business Understanding
2. Data Understanding
3. Plan
4. Data Preparation
5. Model Engineering
6. Model Evaluation

Business Understanding

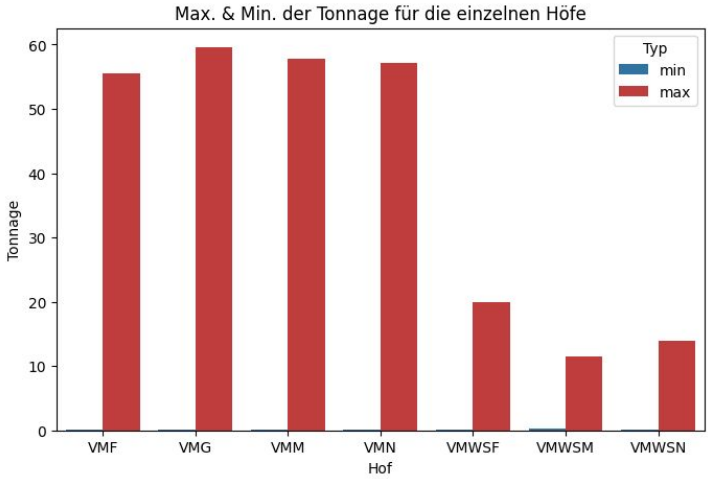
Data Understanding

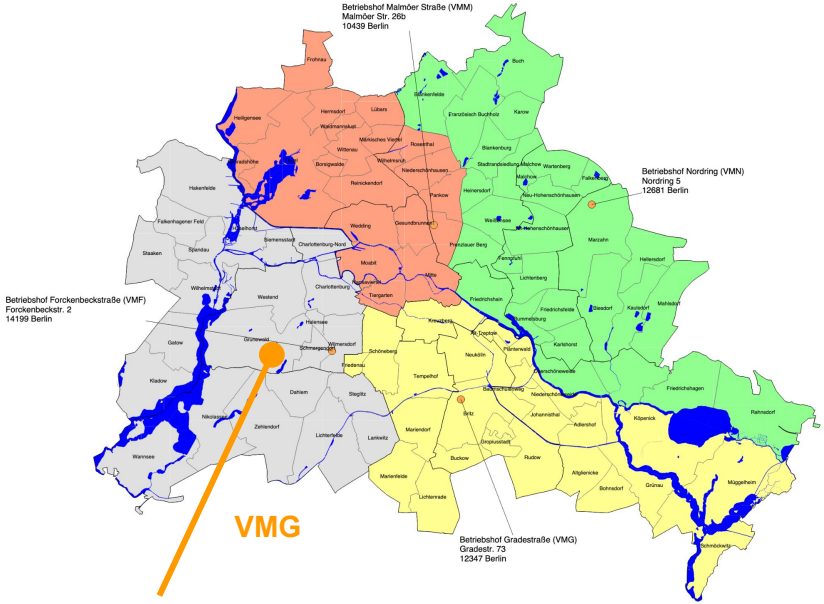


Delimiter: ;

	Monat	KW	Jahr	Datum	Hof	Schicht	Tour	Tonnage
1	1	1	2019	02.01.19	VMF	1	1	5,59
2	1	1	2019	02.01.19	VMF	1	4	3,23
3	1	1	2019	02.01.19	VMF	1	5	5,68
4	1	1	2019	02.01.19	VMF	1	6	5,48
5	1	1	2019	02.01.19	VMF	1	7	7,84
6	1	1	2019	02.01.19	VMF	1	9	4,63
7	1	1	2019	02.01.19	VMF	1	10	4,38
8	1	1	2019	02.01.19	VMF	1	12	9,50
9	1	1	2019	02.01.19	VMF	1	1	24,00
10	1	1	2019	02.01.19	VMF	1	2	20,70
11	1	1	2019	02.01.19	VMF	1	3	22,30
12	1	1	2019	02.01.19	VMF	1	4	17,16
13	1	1	2019	02.01.19	VMF	1	5	11,60
14	1	1	2019	02.01.19	VMF	1	6	19,10
15	1	1	2019	02.01.19	VMF	1	7	15,94
16	1	1	2019	02.01.19	VMF	1	8	20,94
17	1	1	2019	02.01.19	VMF	1	9	16,94
18	1	1	2019	02.01.19	VMF	1	10	11,28
19	1	1	2019	02.01.19	VMF	1	11	10,54
20	1	1	2019	02.01.19	VMF	1	12	22,76
21	1	1	2019	02.01.19	VMF	1	13	18,34

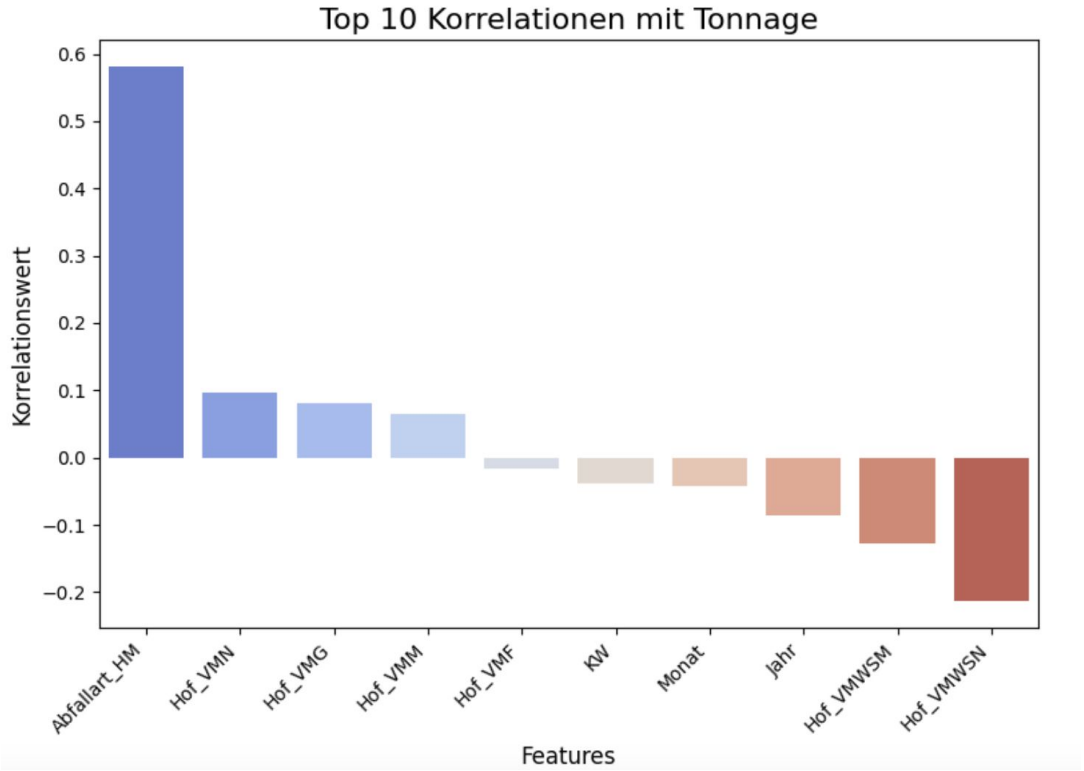
Data Understanding



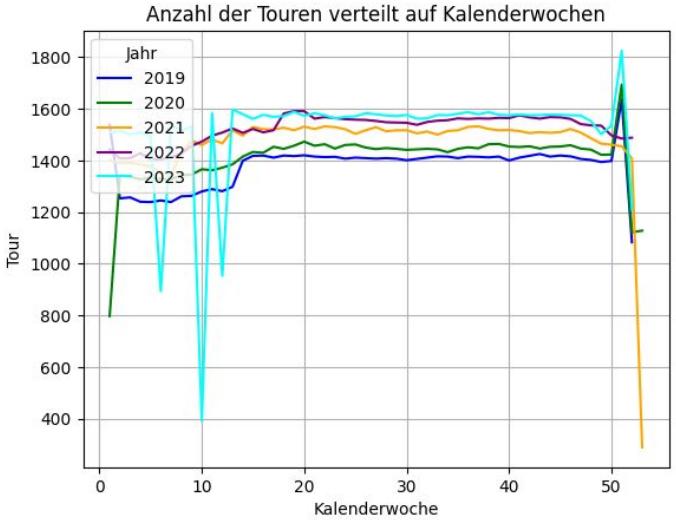


max Wert Tonnage: 59.58
min Wert Tonnage: 0.10

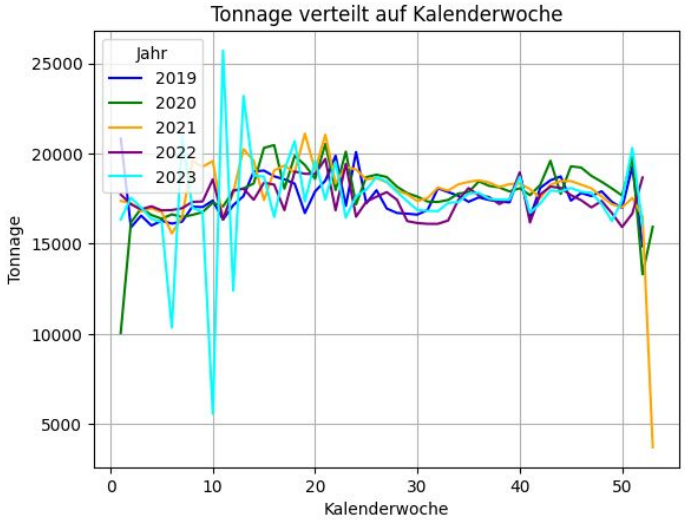
Data Understanding



Data Understanding



Anzahl Touren auf Kalenderwochen & Jahr



Tonnage auf Kalenderwochen & Jahr

Plan

Geplante Datensätze:

- Wetterdaten (mit rolling) -> Free-Weather-API
- Inflationsdaten zu Nahrungsmittelgruppen -> statistisches Bundesamt
- Wahlumfragen -> wahlumfragen.org
- Feiertage -> https://date.nager.at/
- Ferienzeiten -> Schulkalender

- Saison
- Wochentag

new data / OpenMeteo - FreeWeatherAPI Historical

Daily Parameter Definition

Aggregations are a simple 24 hour aggregation from hourly values. The parameter

Variable	Unit	Description
weather_code	WMO code	The most severe weather condi
temperature_2m_max **temperature_2m_min**	°C (°F)	Maximum and minimum daily a
apparent_temperature_max **apparent_temperature_min**	°C (°F)	Maximum and minimum daily a
precipitation_sum	mm	Sum of daily precipitation (inclu
rain_sum	mm	Sum of daily rain
snowfall_sum	cm	Sum of daily snowfall
precipitation_hours	hours	The number of hours with rain
sunrise **sunset**	iso8601	Sun rise and set times

new data / Inflationsdaten - Nahrungsmittel

inflation

Jahr	Monat	Brot	Fleisch	Fisch	Molkerei-Eier	Speisefette-öle	Obst	Gemüse	Zucker, Marmelade, Honig	Fertigerichte	Kaffee
2020	Januar	100.4	97.9	100.7	99.8	102.4	98.1	99.9	100.2	100.5	100.9
2020	Februar	100.4	99.6	100.8	99.7	101.7	101.3	105.9	100.4	100.7	101.0
2020	März	100.5	100.3	100.8	99.9	101.3	100.6	103.9	100.8	100.8	101.0
2020	April	101.1	101.0	101.0	100.4	101.2	101.8	109.4	100.4	100.9	100.9
2020	Mai	100.9	101.5	101.1	101.7	101.3	102.2	104.8	100.7	101.2	101.4
2020	Juni	101.2	101.7	101.0	101.7	100.2	102.5	101.9	101.3	101.5	101.3
2020	Juli	99.0	99.9	99.1	99.6	99.4	99.4	95.6	98.5	99.5	98.9
2020	August	99.3	99.8	99.1	99.8	99.6	98.7	93.2	99.1	99.1	99.2
2020	September	99.2	99.5	99.3	99.4	98.5	98.9	93.0	99.2	98.7	98.8
2020	Oktober	99.0	99.6	98.9	99.3	97.4	98.6	96.2	99.6	99.0	99.3
2020	November	99.4	99.7	99.0	99.4	98.8	98.3	99.8	99.9	99.2	99.0
2020	Dezember	99.5	99.5	99.2	99.3	98.2	99.5	96.4	99.7	99.0	98.4
2021	Januar	101.2	101.0	100.9	101.4	100.6	100.9	104.2	101.2	100.4	101.2
2021	Februar	102.0	101.3	100.9	102.2	99.8	102.2	107.1	103.3	101.3	101.8
2021	März	102.3	101.7	101.3	102.2	102.1	102.7	104.7	103.3	101.5	102.3
2021	April	102.5	102.4	101.6	102.6	103.6	102.8	112.5	103.7	101.7	103.3
2021	Mai	102.8	102.4	101.5	103.2	104.7	103.2	106.4	102.9	101.3	102.8
2021	Juni	102.9	102.1	102.0	104.0	105.8	102.3	101.3	103.4	102.0	102.8
2021	Juli	103.2	103.0	102.3	104.4	106.6	100.9	103.1	103.5	102.0	102.9
2021	August	103.7	103.1	102.4	104.7	107.3	101.2	101.6	103.0	101.9	103.0
2021	September	104.1	103.3	102.5	105.0	105.4	100.6	101.5	103.4	102.1	102.9
2021	Oktober	104.3	103.9	102.6	105.1	105.2	100.1	100.0	104.2	102.0	103.1
2021	November	104.8	104.2	102.9	105.6	110.6	99.8	102.1	103.6	102.0	103.6
2021	Dezember	105.6	104.6	102.7	106.1	113.6	102.2	105.9	103.9	102.5	103.2
2022	Januar	106.2	105.2	103.5	107.4	117.9	104.2	113.1	104.6	102.0	104.5
2022	Februar	107.2	105.3	103.9	108.5	119.2	104.6	118.1	104.8	102.9	104.6
2022	März	108.5	106.7	105.2	109.5	120.5	105.1	118.6	105.3	103.3	105.3
2022	April	111.5	114.1	108.0	112.3	132.6	105.9	122.8	104.6	106.2	106.9

new data / Wahlumfragen

wahlumfragen.org/berlin/

Ableitungsrechner | Integralrechner | Moodle | HTW Mail | ChatGPT | Kalender

Veröffentlichung	SPD	Grünen	CDU	Linke	AfD	FDP	Piraten	Sonstige
09.02.2023	21,0 %	17,0 %	25,0 %	11,0 %	10,0 %	6,0 %	N/A	10,0 %
09.02.2023	19,0 %	18,0 %	25,0 %	12,0 %	10,0 %	6,0 %	N/A	10,0 %
06.02.2023	17,0 %	18,0 %	26,0 %	12,0 %	10,0 %	5,0 %	N/A	12,0 %
03.02.2023	21,0 %	18,0 %	24,0 %	11,0 %	10,0 %	6,0 %	N/A	10,0 %
02.02.2023	19,0 %	18,0 %	25,0 %	12,0 %	10,0 %	6,0 %	N/A	10,0 %
18.01.2023	18,0 %	21,0 %	23,0 %	11,0 %	11,0 %	6,0 %	N/A	10,0 %
09.01.2023	19,5 %	20,0 %	22,5 %	12,5 %	11,0 %	4,0 %	N/A	10,5 %
21.12.2022	21,0 %	20,0 %	21,0 %	12,0 %	10,0 %	6,0 %	N/A	10,0 %
23.11.2022	19,0 %	22,0 %	21,0 %	11,0 %	10,0 %	5,0 %	N/A	12,0 %
16.11.2022	20,0 %	20,0 %	21,0 %	12,0 %	10,0 %	7,0 %	N/A	10,0 %
21.09.2022	17,0 %	22,0 %	21,0 %	12,0 %	10,0 %	6,0 %	N/A	12,0 %
12.07.2022	20,0 %	21,0 %	20,0 %	12,0 %	8,0 %	7,0 %	N/A	12,0 %
19.06.2022	21,0 %	20,0 %	21,0 %	12,0 %	8,0 %	8,0 %	N/A	10,0 %
23.03.2022	20,0 %	21,0 %	20,0 %	12,0 %	8,0 %	8,0 %	N/A	11,0 %

new data / Feiertage + Ferienzeiten

Daten zu Feiertagen

Feiertage von 2019 bis 2023 über API abrufen

- über API: https://date.nager.at/
- API URL: https://date.nager.at/api/v3/PublicHolidays/{year}/DE
- Filterung für Berlin (DE-BE)

Ferienzeiten aus Schulkalender

Data Preparation - BSR

Prüfung:

- Null-Werte
- Duplikate
- Korrekte Datenbereiche (2019-2023)
- Datentypen
- Outlier und ungewöhnliche Werte

Data Preparation - weitere Datensätze

- Extraktion relevanter Merkmale
- Zu numerische Datentypen casten
- One-Hot-Encoding
- Datum vereinheitlichen

Feature Engineering

- **Tonnage_delay**: Die Menge an Abfall, die in den letzten Tagen vor der aktuellen Tour geliefert wurde.
- **Tage_vorher_nicht_geliefert**: Zeigt an, wie viele Tage zuvor keine Lieferung erfolgte (z. B. aufgrund von Feiertagen oder anderen Unterbrechungen).
- **Wetter letzte drei Tage (rolling):** gibt jeweils von den letzten drei Tagen die Temperatur, Regen in mm, Schnee in mm und Windgeschwindigkeit an

Delimiter: ,

	Datum	TageZuvorNichtGeliefert
22	2019-01-22	0
23	2019-01-23	0
24	2019-01-24	0
25	2019-01-25	0
26	2019-01-26	0
27	2019-01-27	1
28	2019-01-28	2
29	2019-01-29	0
30	2019-01-30	0
31	2019-01-31	0
32	2019-02-01	0
33	2019-02-02	0
34	2019-02-03	1

finaler Datensatz

- **Alle Tabellen über Datum Feature gemerged**

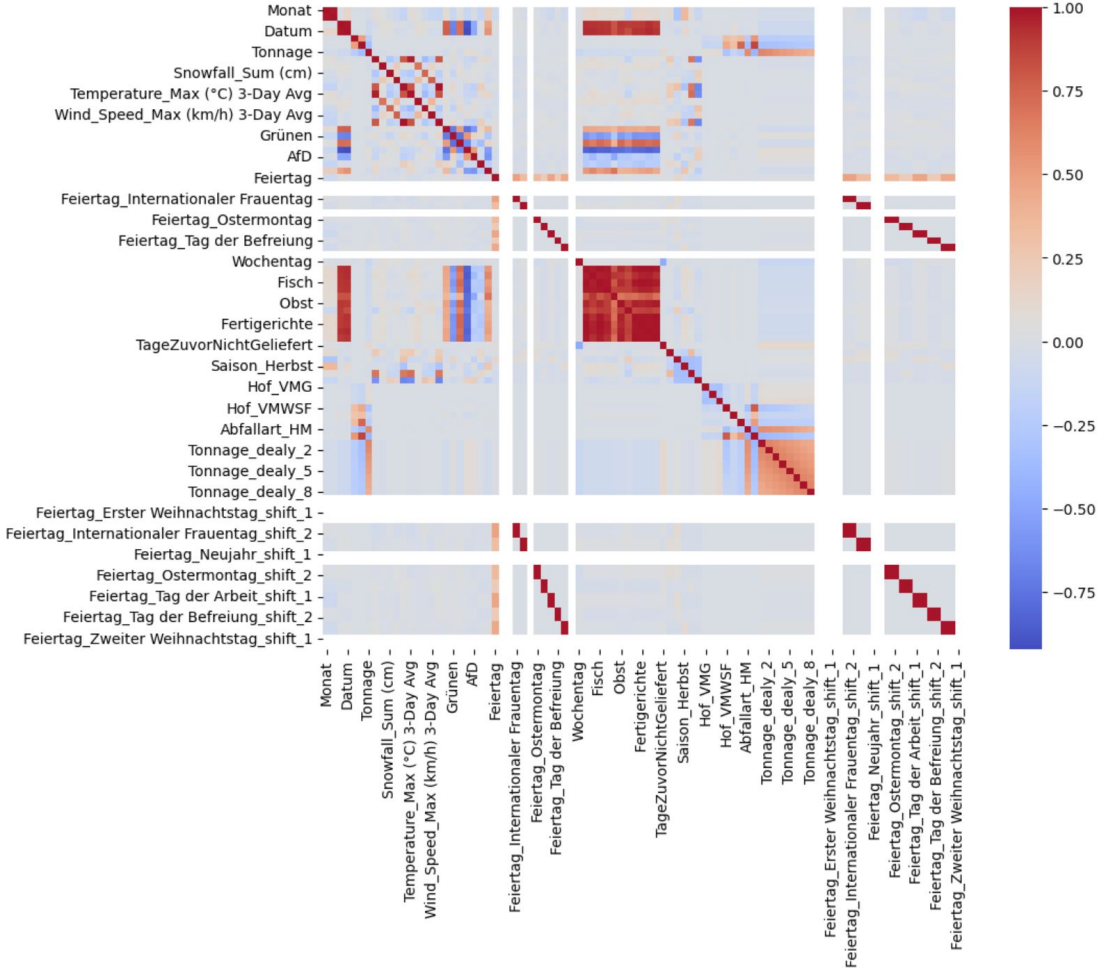
```
df.shape
```

```
(381661, 91)
```

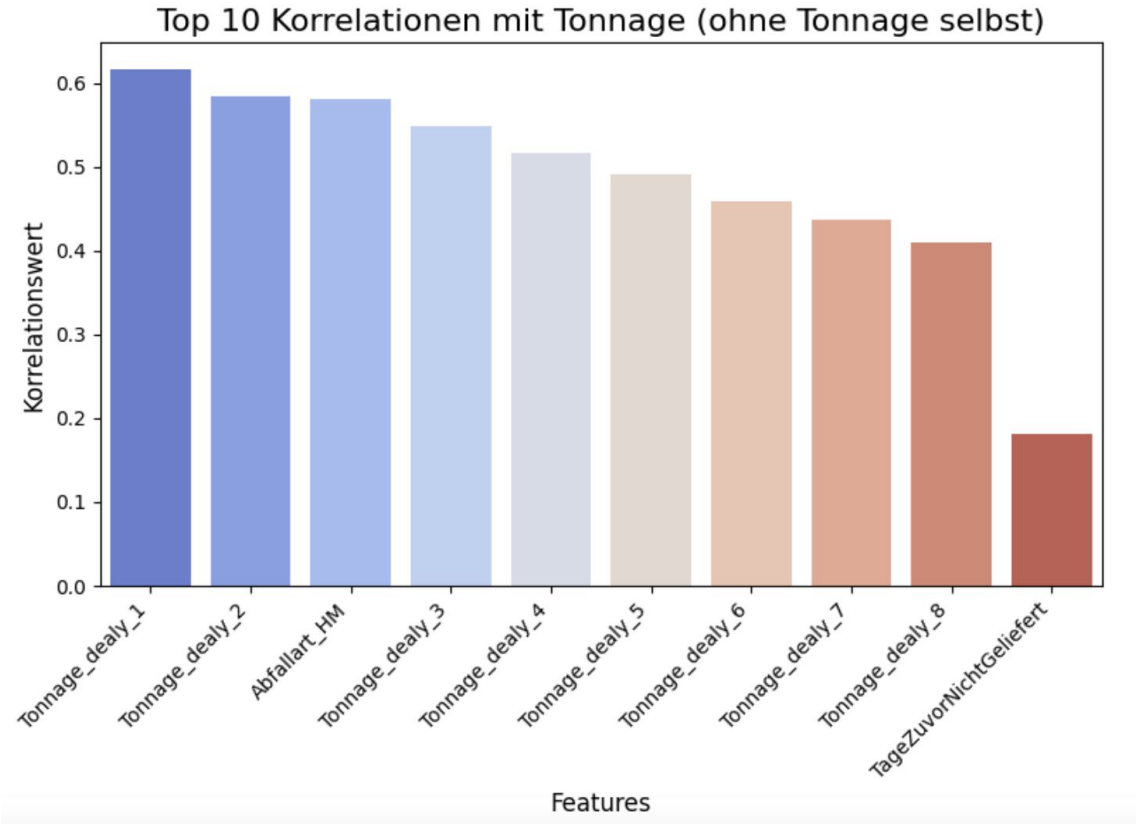
```
df.head()
```

	Schicht	Tour	Tonnage	Temperature_Max (°C)	Rain_Sum (mm)	Snowfall_Sum (cm)	Wind_Speed_Max (km/h)	Daylight_Duration (s)	Temperature_Max (°C) 3-Day Avg	Rain_Sum (mm) 3-Day Avg	...
0	1	1	5.59	3.6	0.0	0.0	31.2	27967.86	6.133333	0.633333	...
1	1	4	3.23	3.6	0.0	0.0	31.2	27967.86	6.133333	0.633333	...
2	1	5	5.68	3.6	0.0	0.0	31.2	27967.86	6.133333	0.633333	...
3	1	6	5.48	3.6	0.0	0.0	31.2	27967.86	6.133333	0.633333	...
4	1	7	7.84	3.6	0.0	0.0	31.2	27967.86	6.133333	0.633333	...

finaler Datensatz

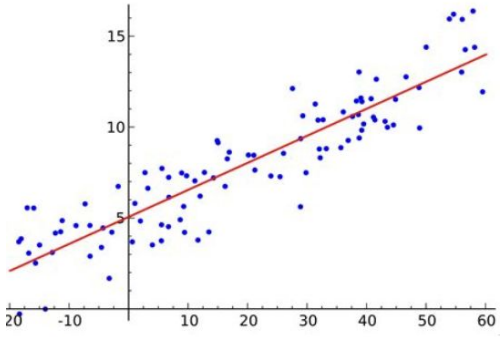


finaler Datensatz

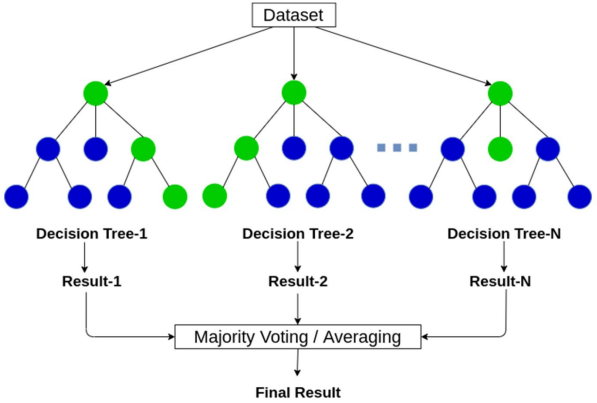


Modelling

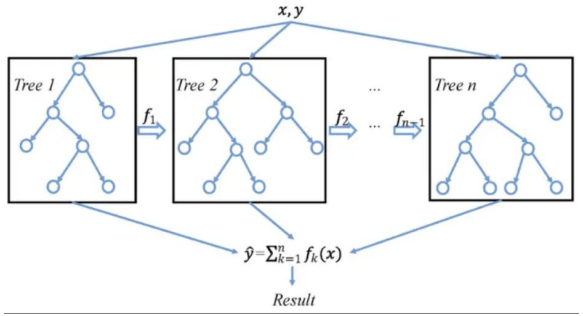
genutzte Modelle



lineare Regression



Random Forest

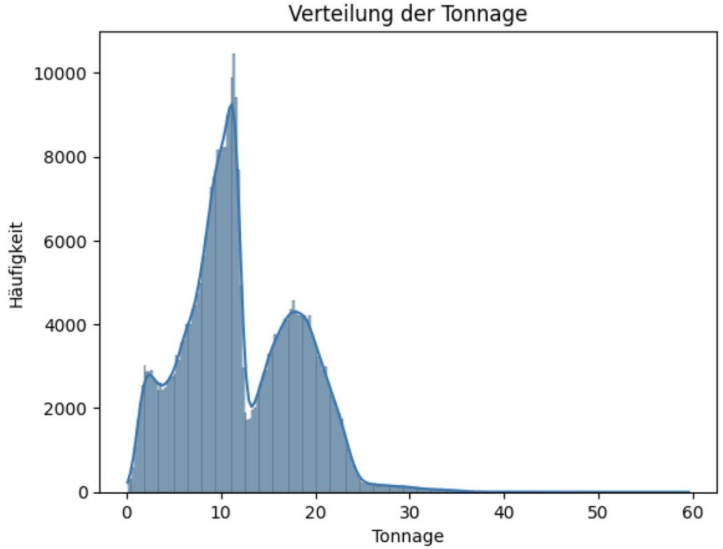


XGBoost

Lineare Regression

- Trainings-, Validierungs-, Testdaten
- Entscheidung:
  mehr features vs. mehr Datenpunkte
- Standardisierung
- Log Transformation
- Ridge & Lasso Regression

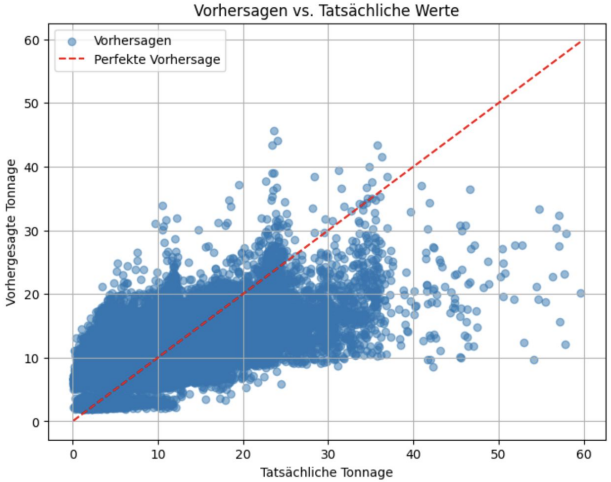
- Cross-Validation
- Feature Selection

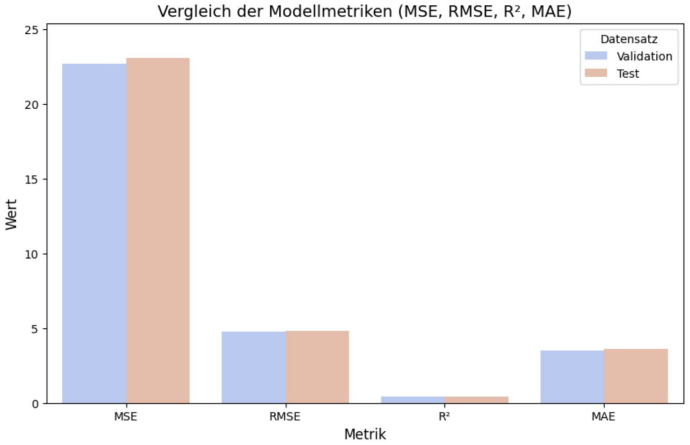


Schiefe-Wert von 0.42 für die Tonnage zeigt eine l
eichte positive Schiefe (Rechtsschiefe) an.

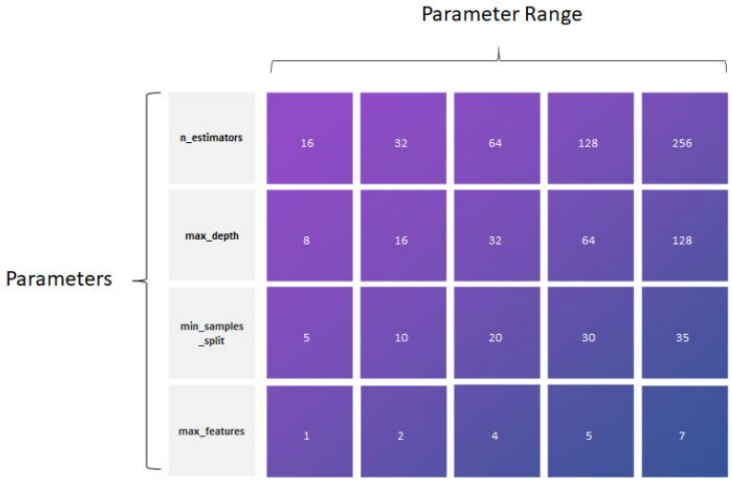
Lineare Regression

Metrik	Original Datensatz - Nach Tuning	Finaler Datensatz - Nach Tuning
Mean Absolute Error (MAE)	3.88	3.21
Root Mean Squared Error (RMSE)	5.27	4.38
Mean Squared Error (MSE)	27.79	19.17
R^2-Score	0.27	0.497





Random Forest Regressor



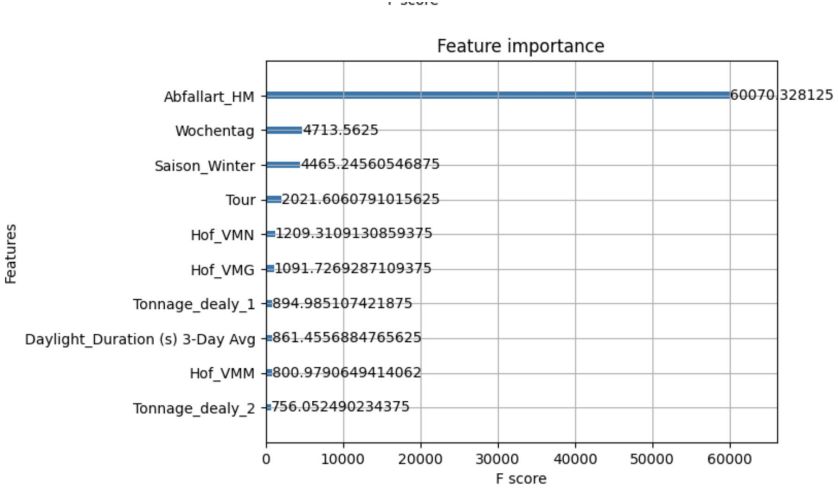

```
[52] search = HalvingRandomSearchCV(
        estimator=model,
        param_distributions=param_dist,
        factor=2,
        cv=tscv,
        scoring='neg_mean_squared_error',
        n_jobs=-1,
        random_state=42,
        verbose=3
    )

    search.fit(X, Y)
```

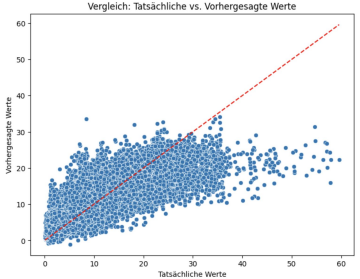
Bestes MSE: 15.682

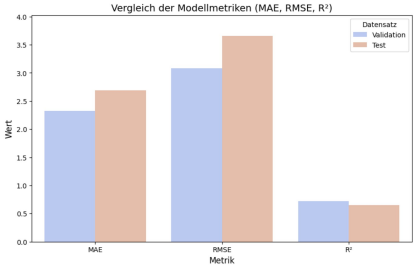
Feature	Importance
Abfallart_HM	0.504637
Tour	0.213131
Wochentag	0.155236
Tonnage_dealy_2	0.027692
Tonnage_dealy_1	0.024392

XGBoost Regression



Metrik	Original Datensatz - Vor Tuning	Original Datensatz - Nach Tuning	Finaler Datensatz - Vor Tuning	Finaler Datensatz - Nach Tuning
Mean Absolute Error (MAE)	3.61	3.55	2.75	2.69
Root Mean Squared Error (RMSE)	4.52	4.33	3.73	3.66
R^2-Score	0.2256	0.2885	0.6335	0.6478
Trainingszeit	0.43 sek	0.52 sek	1.09 sek	1.56 sek
Validierungs(fehler) MAE/RMSE R^2	3.25/4.15 0.4912	3.31/4.11 0.5021	2.44/3.21 0.6960	2.32/3.08 0.7193





Bestes MSE: 13.3956

Optimierung:

- Feature Selection
- GridSearch

Ausblick

Mögliches weiteres Vorgehen:

- Data Preparation verbessern
- weitere Feature Selection
- weitere Datensätze
- tieferes Verständnis von Modellen und Training
- weiteres Spektrum an Modellen trainieren
- Deployment