Kartographierung des Sonnensystems

Projektpräsentation Gruppe F

Team: Natalia Beller, Maximilian Behr, Matthias Faß

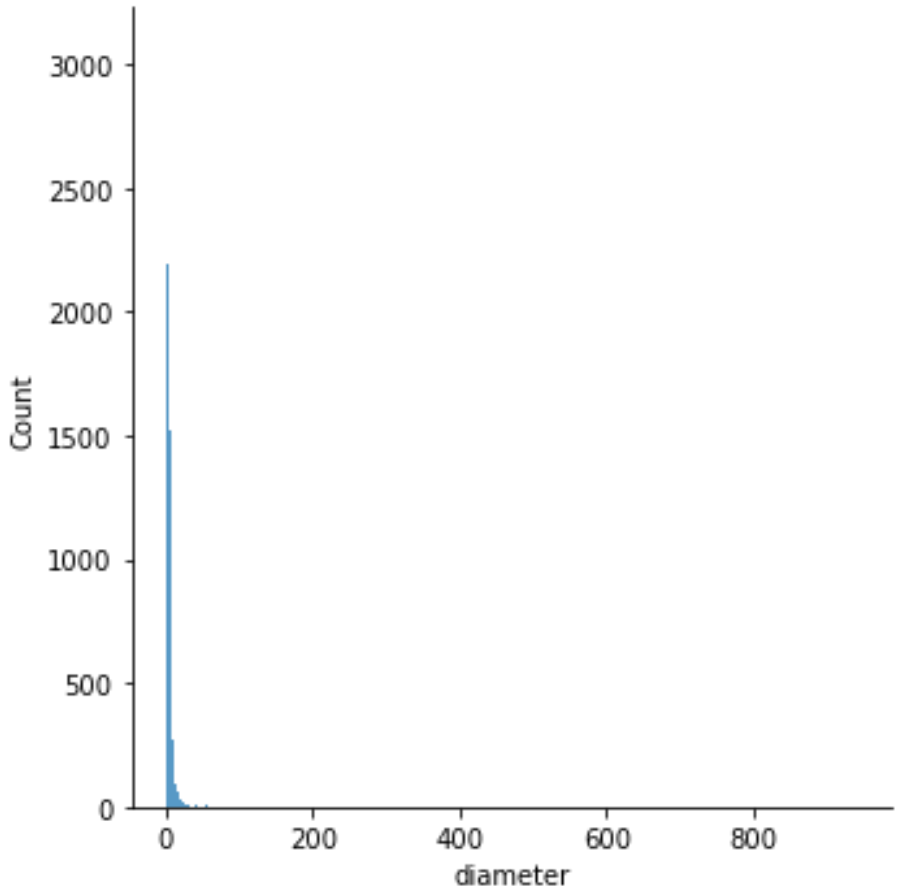
Problembeschreibung

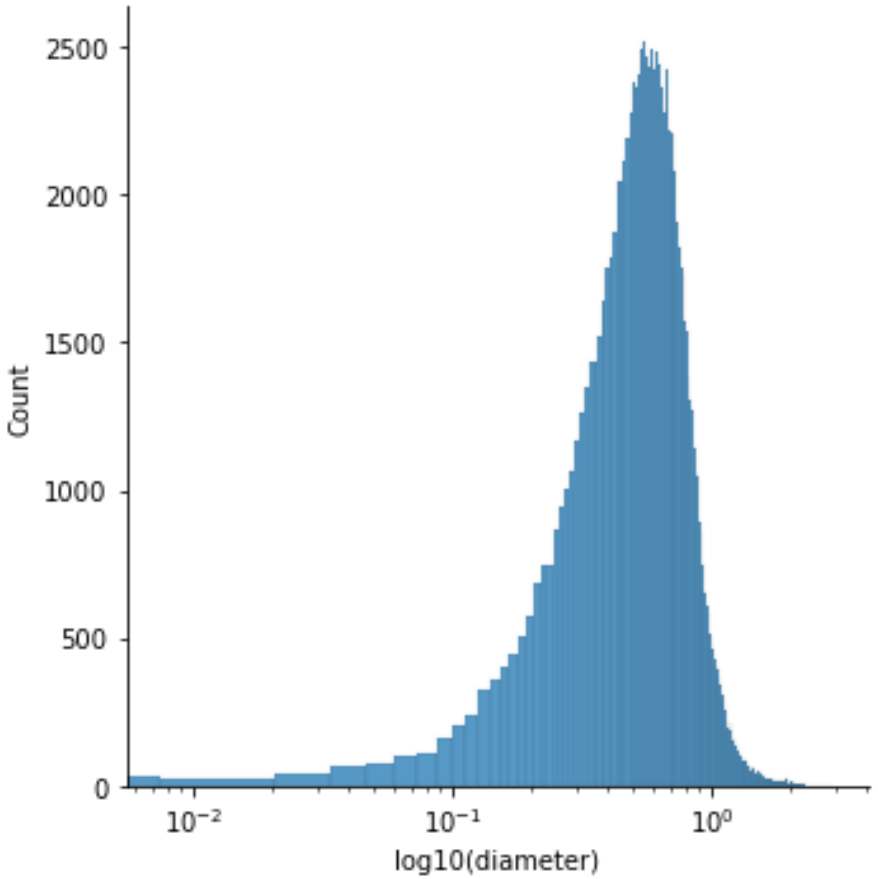
- Ziel: Aufgrund eines Datensatzes mit Features aus der Astronomie soll man die Zielvariable („diameter“) vorhersagen können
- Herausforderungen:
  - Vorverarbeitungsprozess
    - Sichtung der Daten
    - Missing Values
    - Skalierung, Transformation
    - Feature Analyse
  - Visualisierungen
  - Modellsuche und –auswahl für Vorhersagen
  - Hoher Zeitaufwand für Hyperparametersuche

Datenexploration

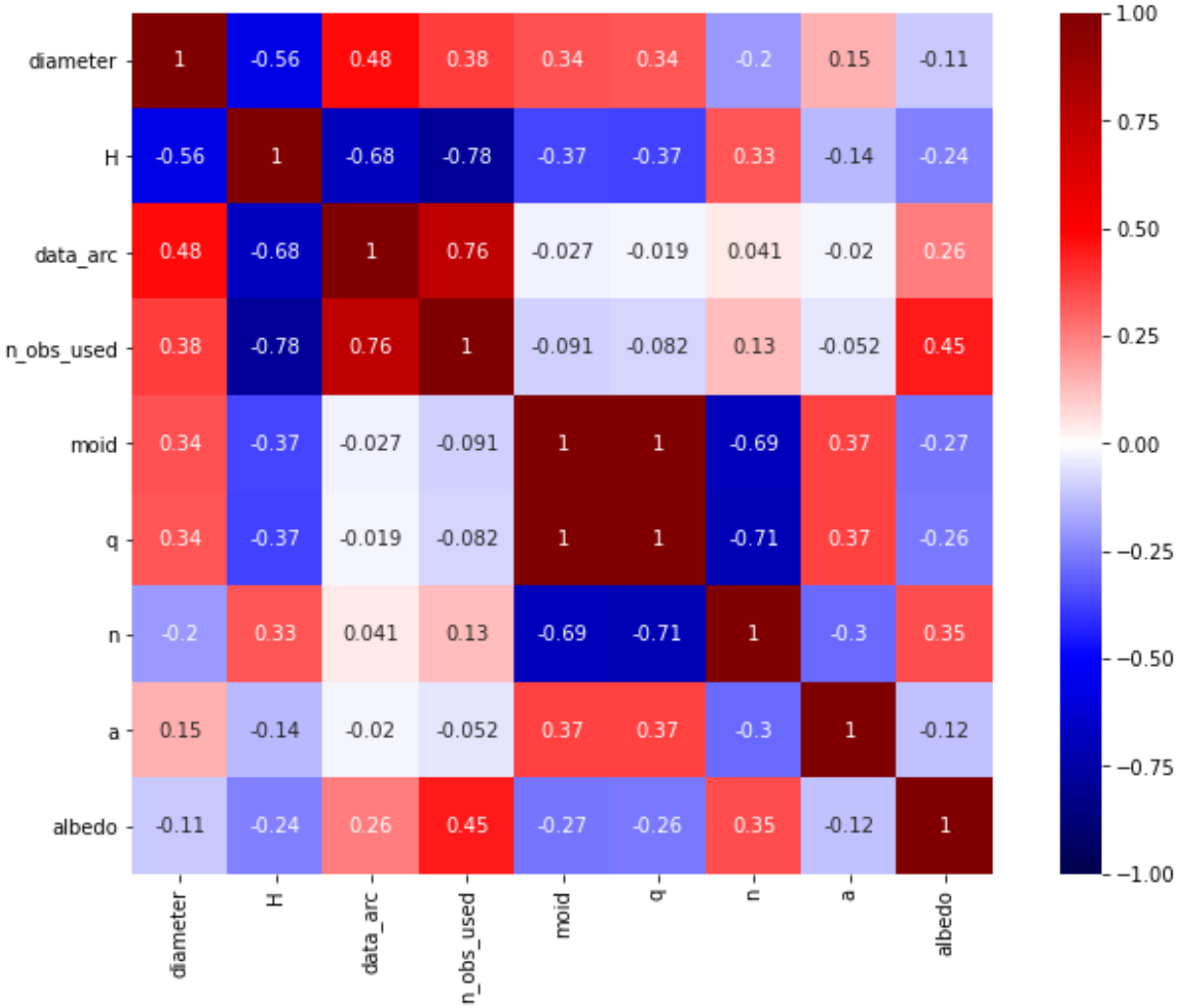
- Erste Sichtung
  - Vertraut machen mit Features (Beschreibung)
  - Datensatz laden und zwei Teile „mit“ und „ohne“ Zielvariable trennen
- Zweite Sichtung
  - Analyse der Daten hinsichtlich benötigter Vorverarbeitungsschritte
  - Missing Values; nicht relevante Spalten; nominale, ordinale numerische Features
  - Skalierung
  - Zielvariable (Umwandlung in log10, Zusammenhänge mit Features…)

Transformation des Durchmessers (y)





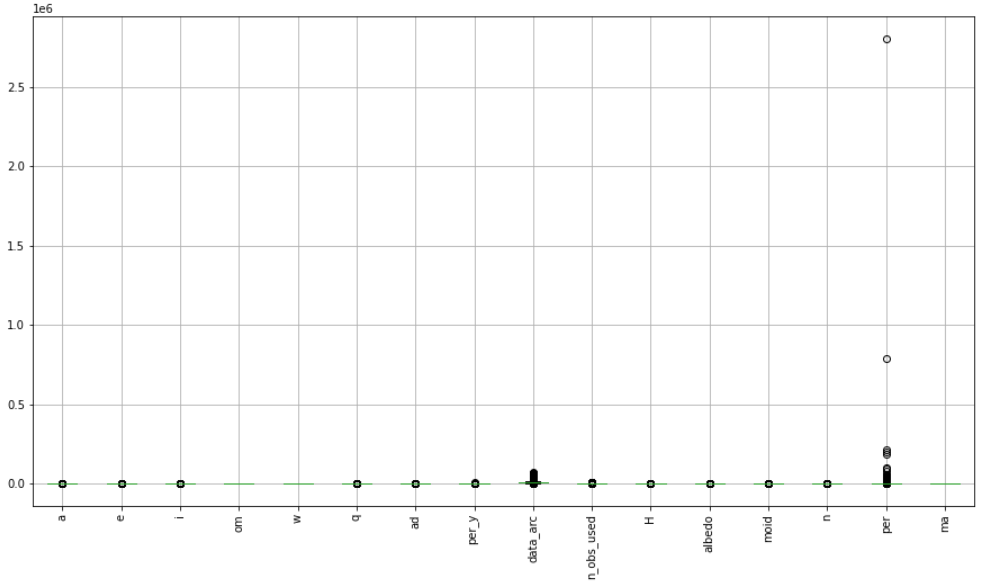
Korrelationsmatrix (reduziert, nur lineare Zusammenhänge)

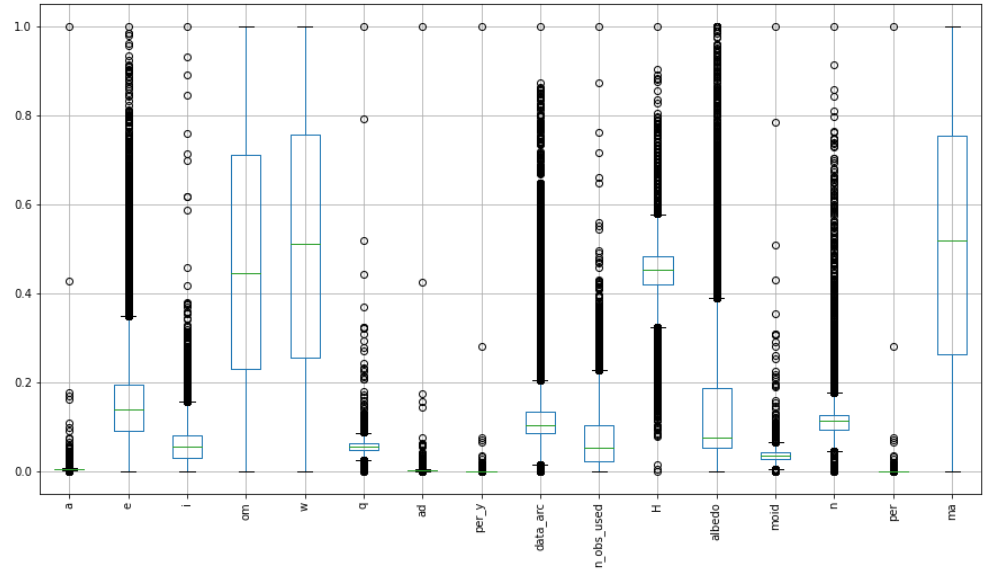


Preprocessing

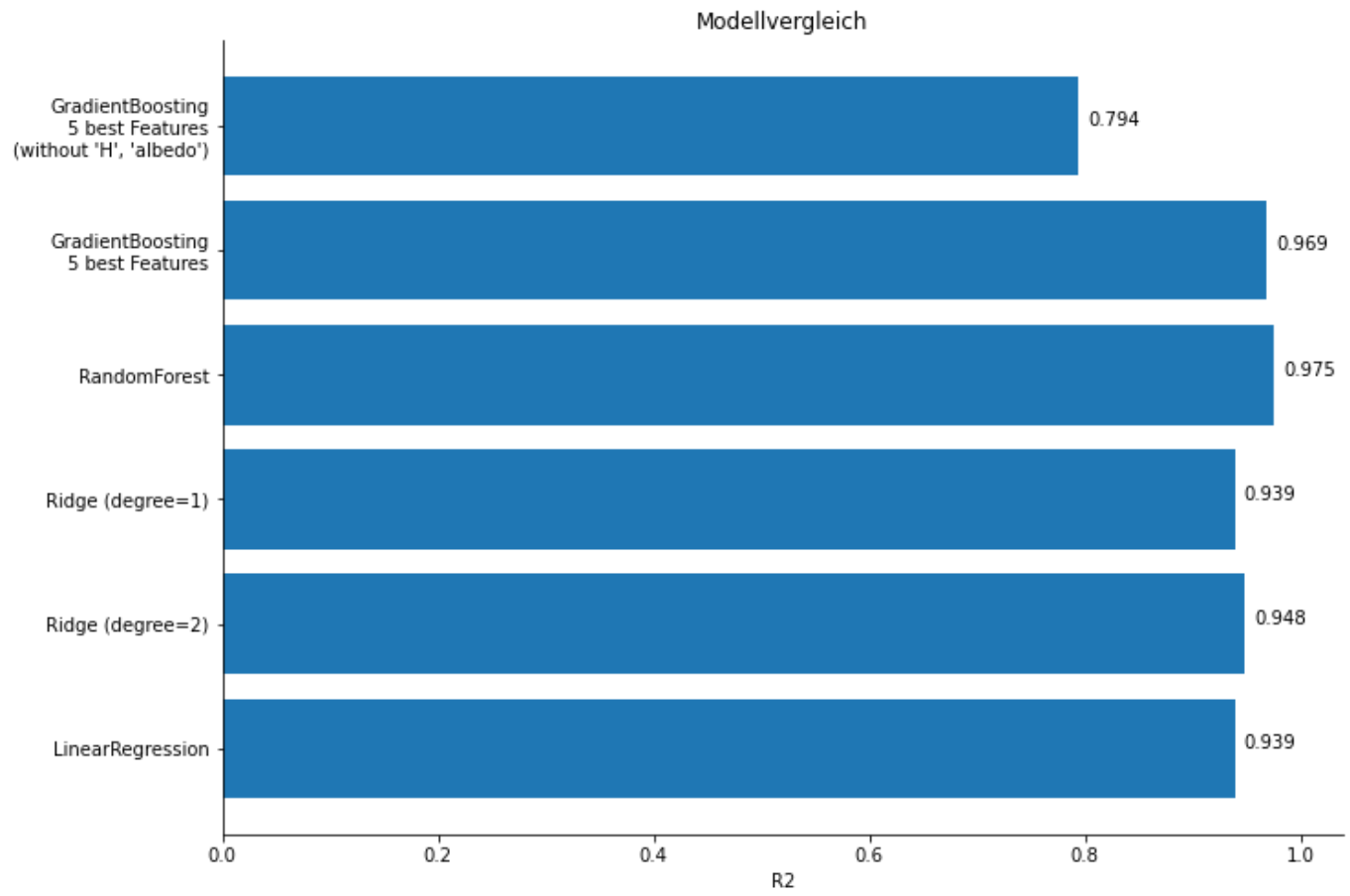
- Grundidee: Pipeline mit ColumnTransformer bauen für Vorverarbeitung
- Missing Values:
  - SimpleImputer benutzen mit Ersetzungsstrategie median, most-frequent
- Ordinale und Nominale Daten enkodiert:
  - OrdinalEncoder und OneHotEncoder benutzen um in numerische Werte umzuwandeln
- Daten skalieren:
  - MinMaxScaler benutzen

Beispiel: Numerische Features skaliert





Modellvergleich



Lineare Regression

- Erster Schritt: Nochmal Datensplitting der Trainingsdaten in Trainings- und Validierungsdaten für alle Modelle

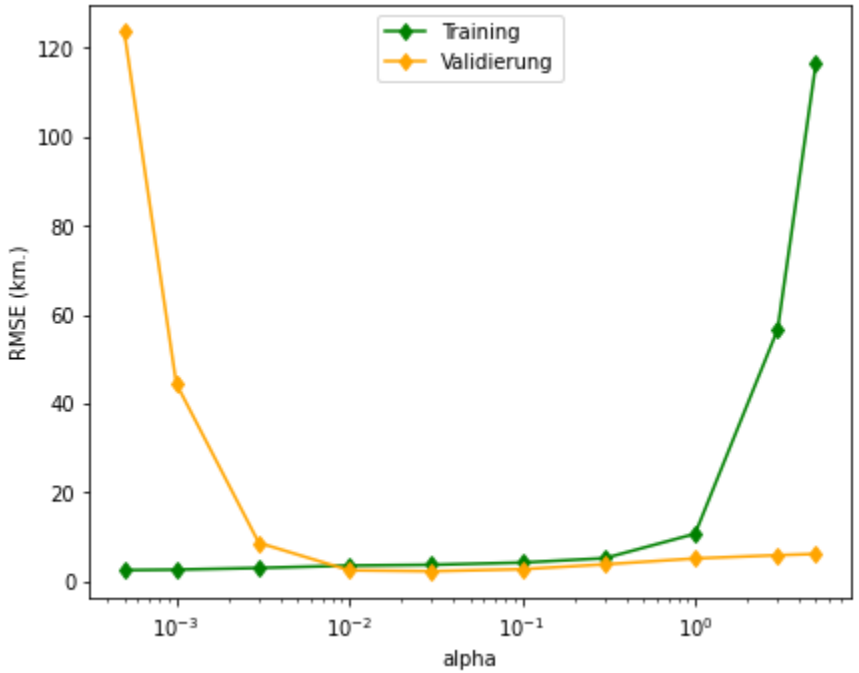
- einfaches Einstiegs-Modell zum ersten Test
- RMSE auf Test und Validierungsdaten war akzeptabel und rel. klein
- kein Overfitting bzw. Underfitting
- R2 score für Trainings-,Validierungs- und Testdaten ungefähr 0.94.

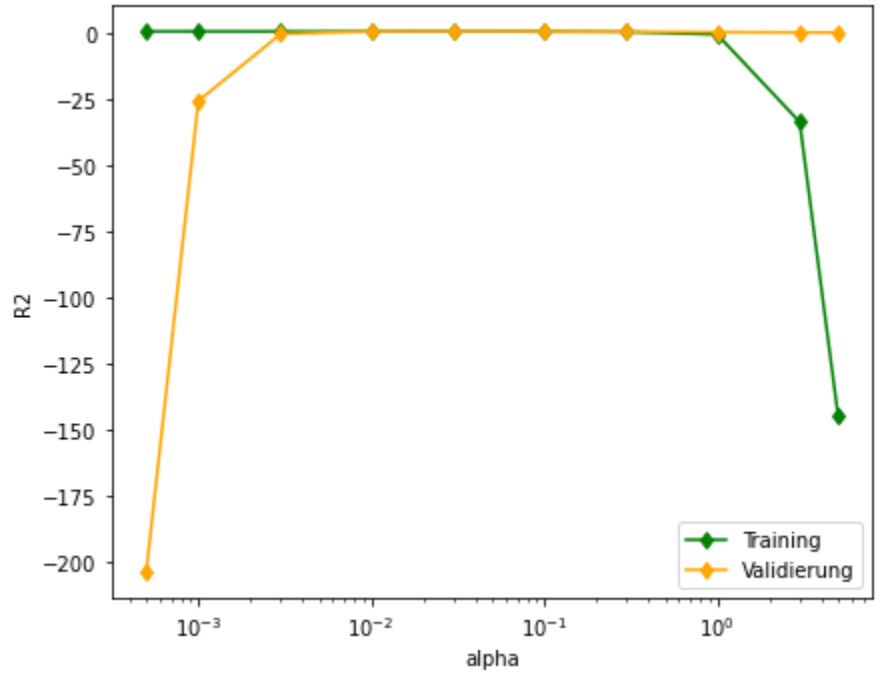
Feature Expansion für Quadratische Regression

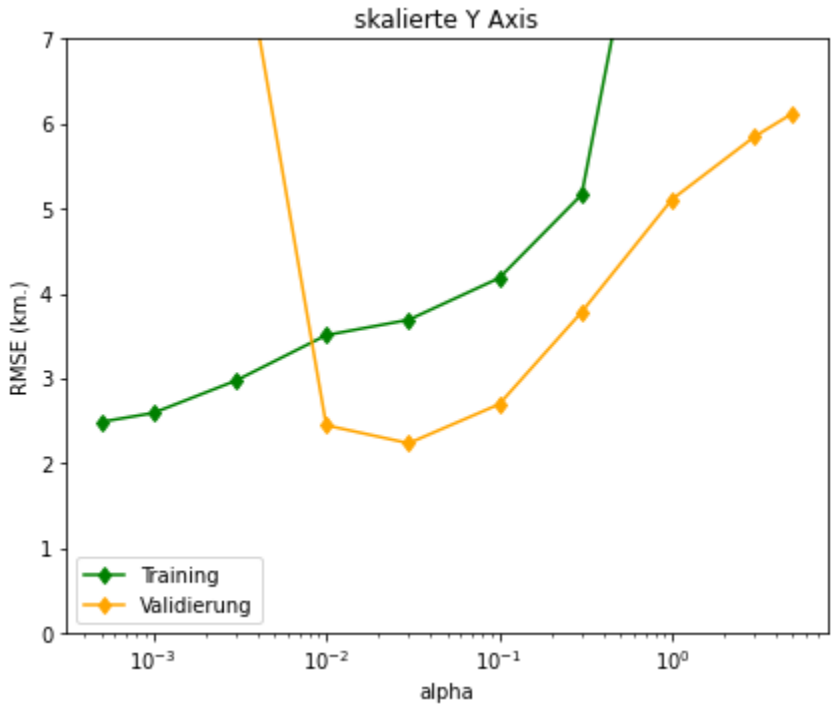
- etwas komplexeres Model
- Ergebnis: Overfitting -> Fehler auf Validierungsdaten riesig
  - R2 Train = 0.963
  - R2 Validierung= -8.899067976653561e+20
- Idee: Ridge Regularisierung

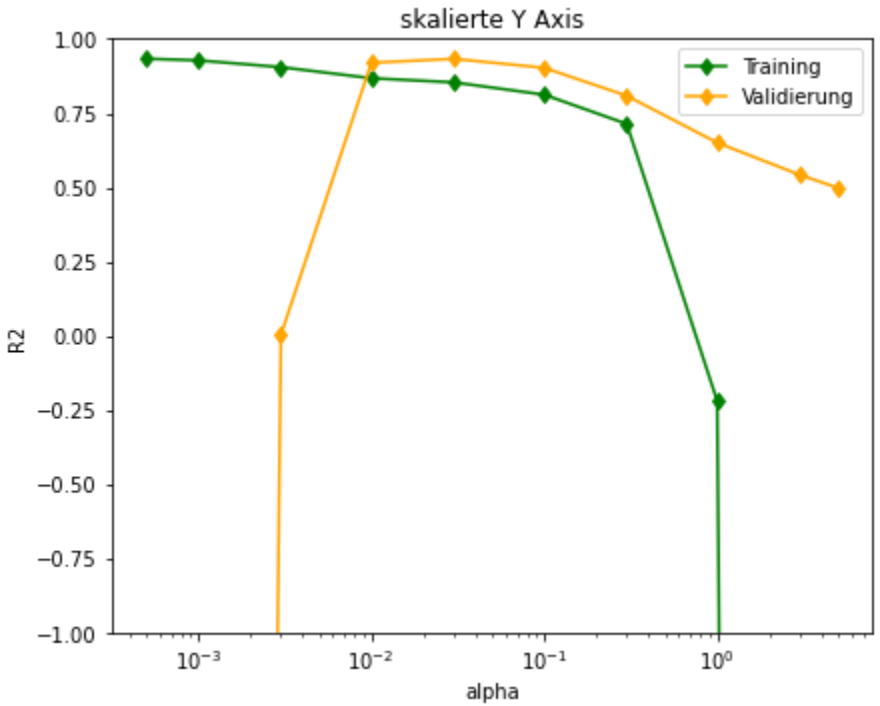
Ridge Regularisierung

- Hyperparametersuche manuell α=[0.0005,0.001,...,3,5]
  - ungefähr α =0.03 (Grad 2)
- Nochmal Kreuzvalidierung mit zusätzlicher Variation des degrees
  - beste Hyperparameter: α=0.003 und Grad 1


Training
Validierung
RMSE (km.)
alpha


R2
alpha
Training
Validierung


skalierte Y Axis
RMSE (km.)
alpha
Training
Validierung


skalierte Y Axis
R2
alpha
Training
Validierung

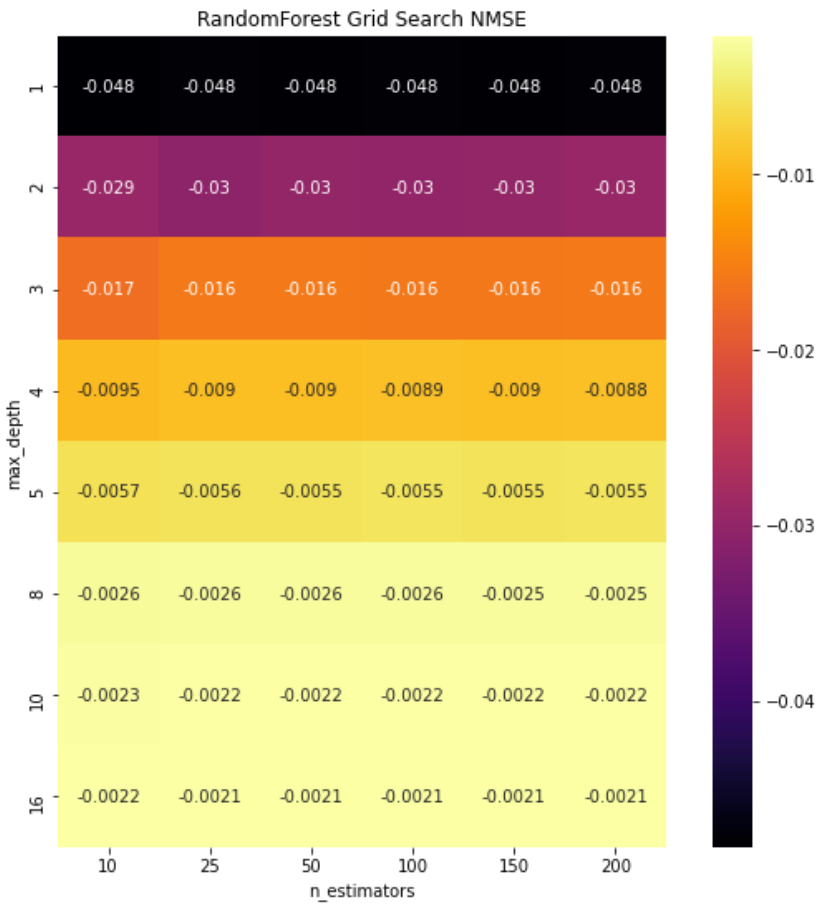
Random Forest

- 100 Bäume zum schnellen Testen
- gute RMSE und R2 Werte auf Trainings- und Testdaten
- Idee: Hyperparametersuche Kreuzvalidierung GridSearchCV

Hyperparametersuche Kreuzvalidierung GridSearchCV

- Anzahl der Bäume
- Tiefe der Bäume
- Ergebnis: {'max_depth': 16, 'n_estimators': 200}
  - leichte Verbesserung erzielt

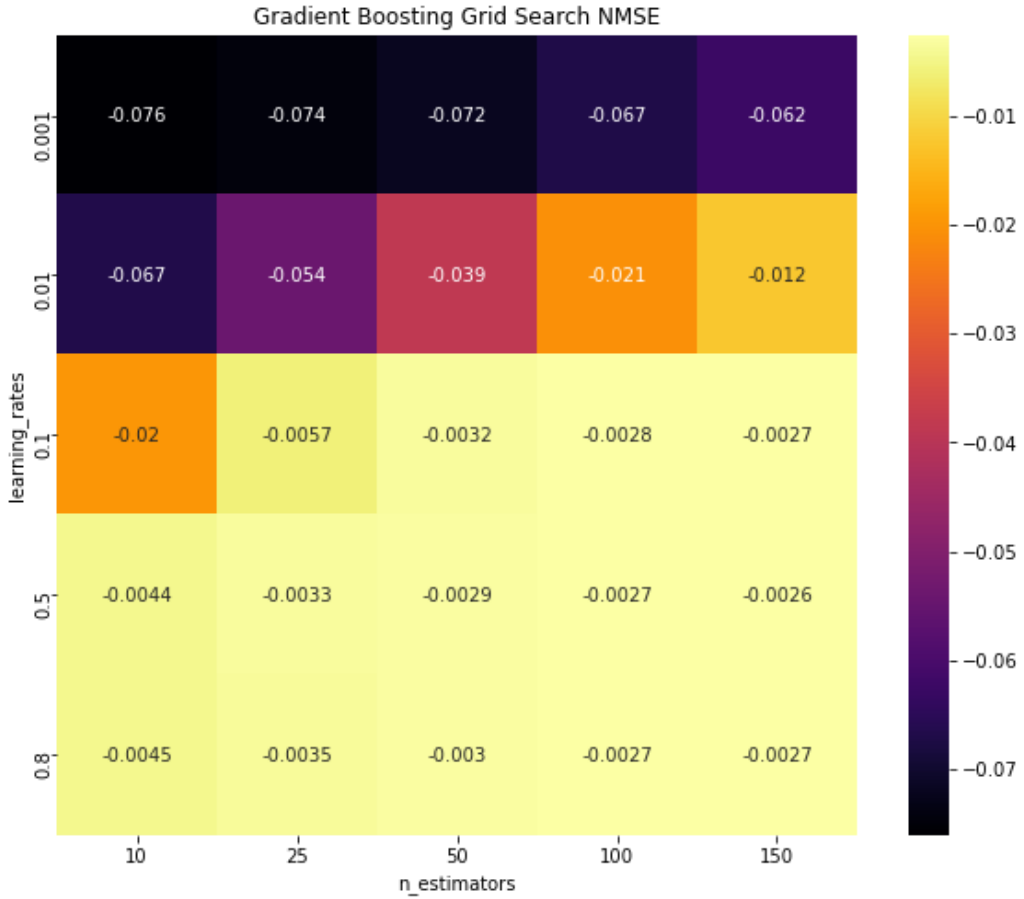
Diagramm (Heatmap, Kreuzvalidierung)



Gradient Boosting

- Kreuzvalidierung über learning rate und Anzahl der Bäume
  - für '5 best Features‘
  - Mit und ohne „H2“ und „albedo“
- Ergebnis: keine wesentliche Veränderungen mehr

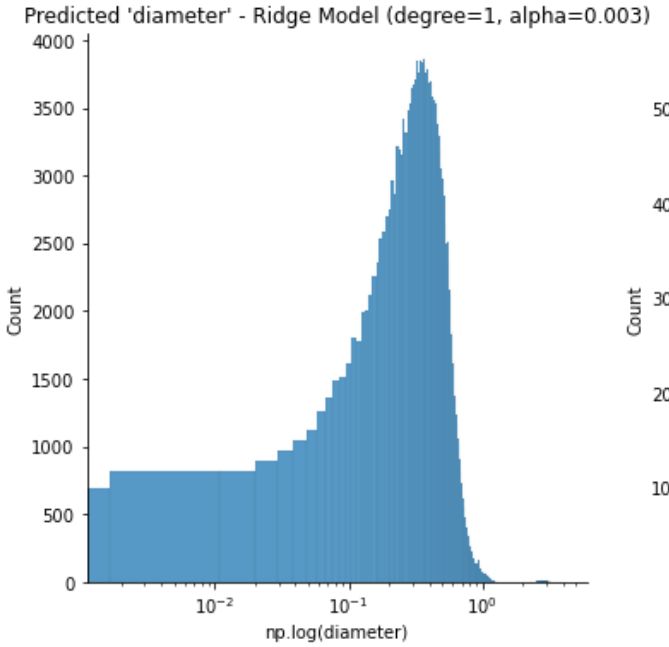
Gradient Boosting (5 best Features)



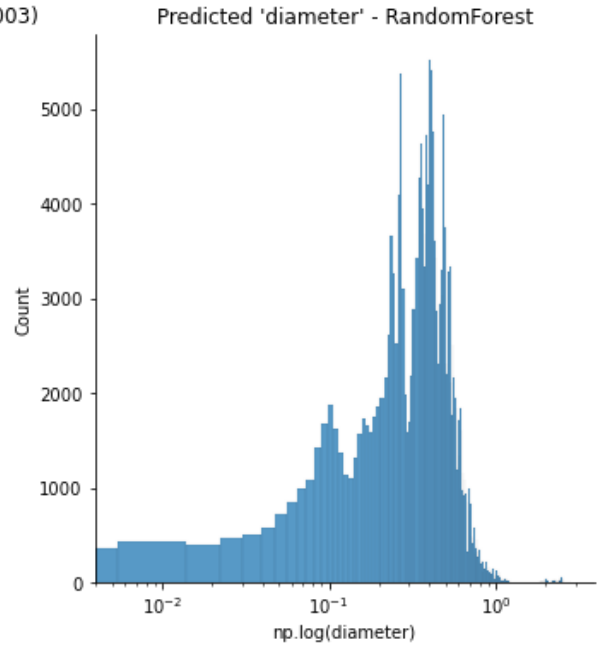
Experimental 1

- Wichtigste Features für Regression (Grundlage: RandomForest)
- Ergebnisse vom RandomForest feature_importances_
- 'H', 'albedo', 'n', 'moid', ... (von wichtig zu unwichtig geordnet)
- Hier wird bei der Best Feature Suche erkannt, dass H und Albeldo sehr wichtig sind. Dies war anfänglich nicht aus der Korrelationsmatrix ersichtlich, weil zwischen albedo (a) und diameter (d) ein nichtlinearer Zusammenhang besteht
- Beweis:
  - $$d = 10^{(3.1236 - 0.5 \log_{10}(a) - 0.2H)}$$

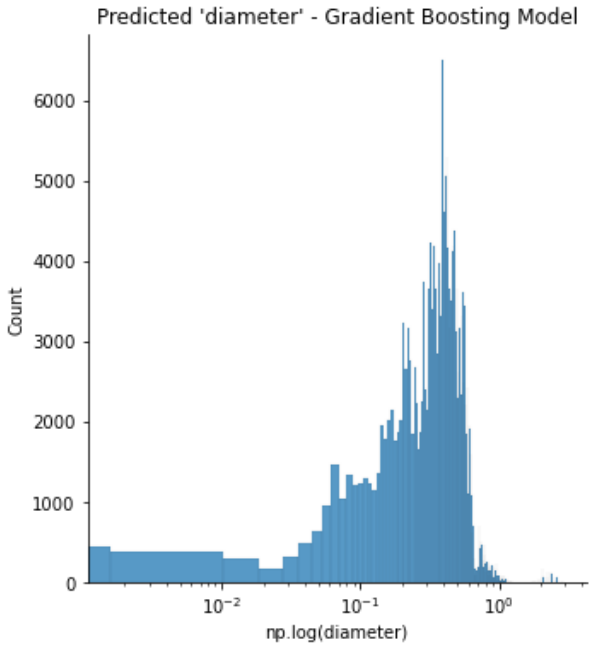
Experimental 2



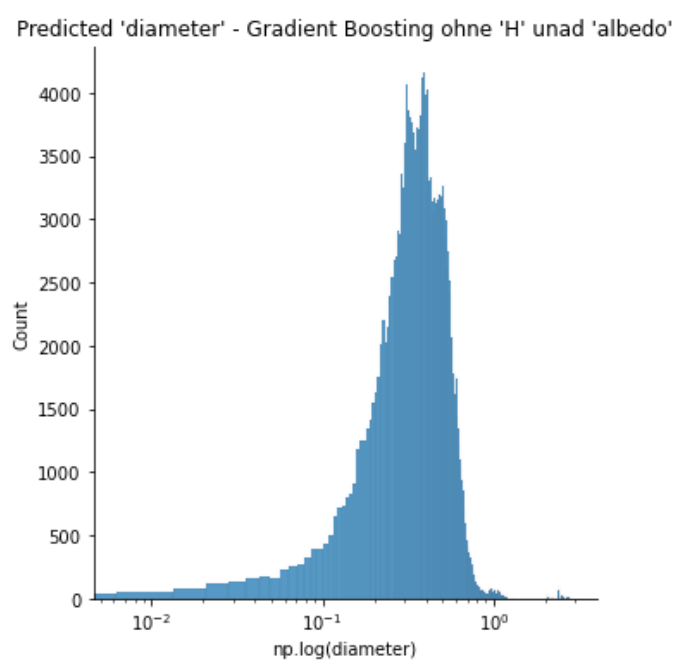
Mean 'diameter' = 7.515 km.
Std 'diameter' = 702.79 km.



Mean 'diameter' = 3.52 km.
Std 'diameter' = 15.615 km.



Mean 'diameter' = 3.442 km.
Std 'diameter' = 15.357 km.



Mean 'diameter' = 3.962 km.
Std 'diameter' = 20.558 km.

Experimental 3

- Features ohne H und Albedo (Grundlage: Gradient Boosting)
- drastische Verschlechterung
  - 5 best Features
  - (without 'H', 'albedo') ---
  - R2 Test = 0.794
- wichtigste Features nicht mehr vorhanden -> deutlich spürbar