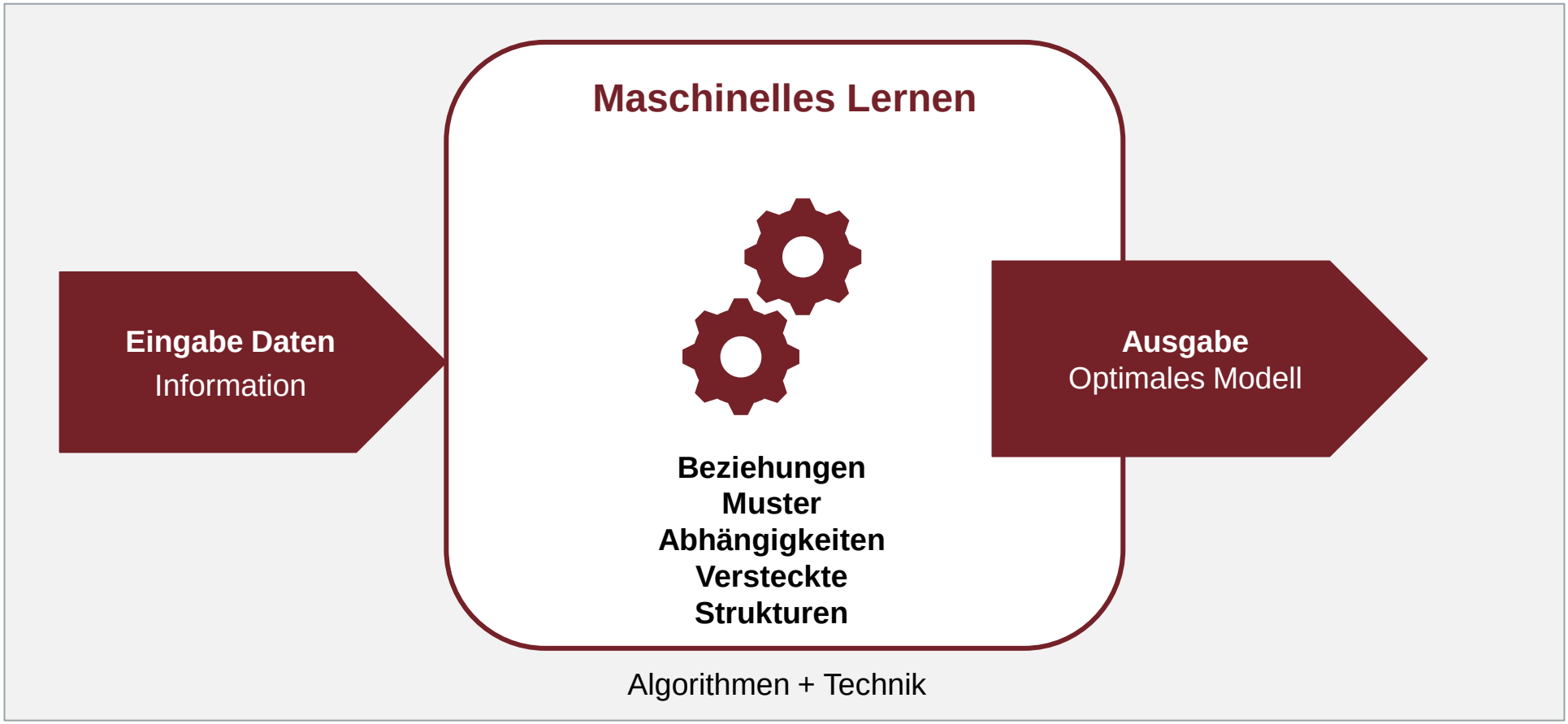
LEUPHANA
UNIVERSITÄT LÜNEBURG

KI-Werkstatt - Anwendung von maschinellen Lernverfahren

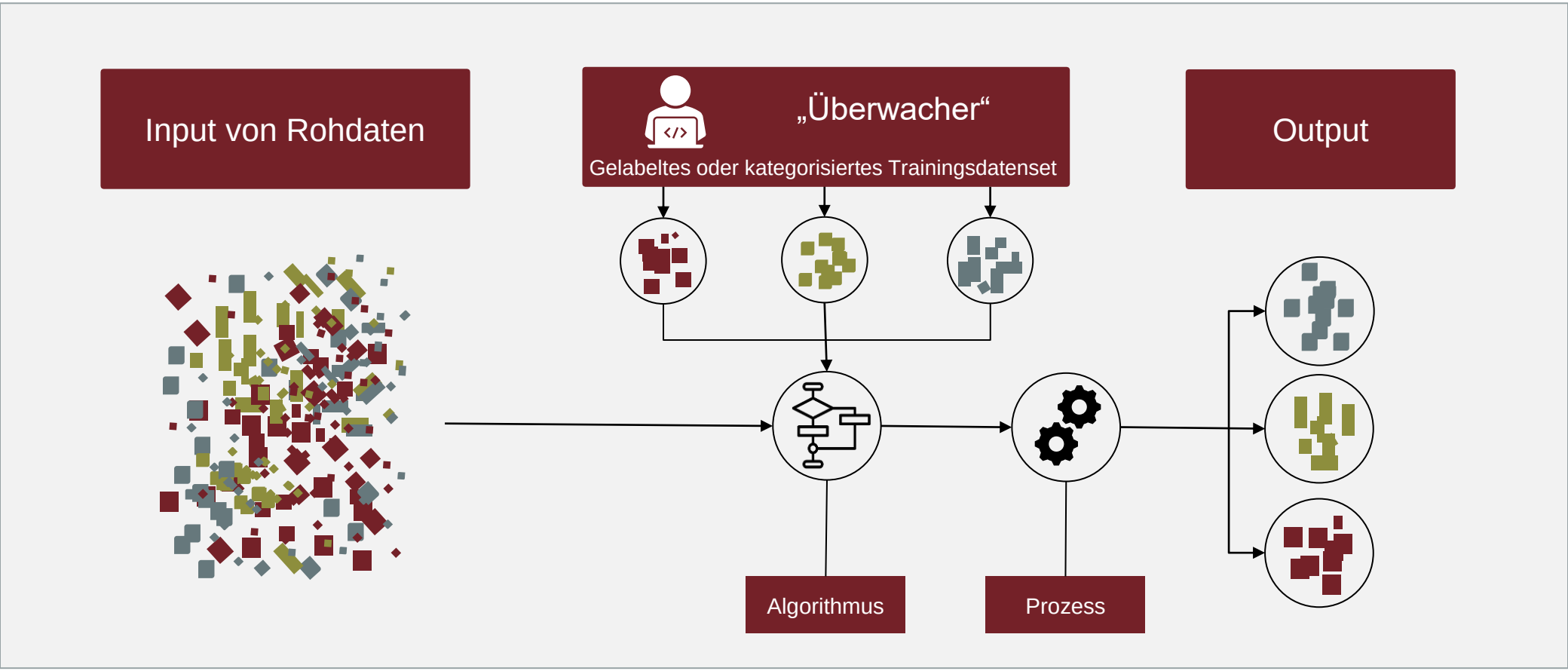
LEUPHANA
UNIVERSITÄT LÜNEBURG

Schematischer Ablauf



LEUPHANA
UNIVERSITÄT LÜNEBURG

Überwachtes Lernen (Supervised Learning)



LEUPHANA
UNIVERSITÄT LÜNEBURG

Cross Industry Standard Process for Data Mining (CRISP-DM)

Material und Software zur Durchführung des Übung

Datensatz

https://ki-werkstatt.info/upload/frachtdaten.zip

Software

https://rapidminer.com/get-started/

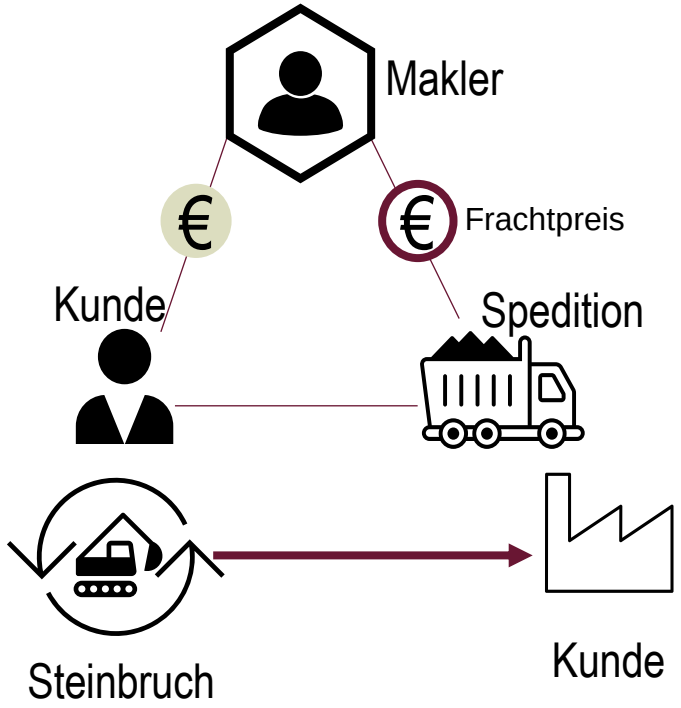
Literatur

- Chapman, et al. (2000)
- Kuhn & Johnson (2013)
- Witten, Frank & Hall (2011)
- Suthaharan (2016)
- James, Witten, Hastie & Tibshirani (2013)

LEUPHANA
UNIVERSITÄT LÜNEBURG

Prognose des Frachtpreises

- Geschäfts-verständnis
- Datenverständnis
- Daten-vorbereitung
- Modellierung
- Evaluation
- Bereitstellung



Ausgangslage:

- Frachtpreismarkt sehr dynamisch
- regionale, produkt- und streckenbezogene Einflussfaktoren
- verhandelte Preise beruhen derzeit auf der jahrelangen Erfahrung der Geschäftsführung
- Präzise Preisbestimmung in der Vermakelung von hoher wirtschaftlicher Relevanz

Aufgabe:

Entwickeln Sie eine KI-basierte Frachtpreisbestimmung durch Anwendung von Supervised Learning.
Beantworten Sie hierzu die nachstehenden Teilaufgaben entlang des CRISP-DM.

LEUPHANA
UNIVERSITÄT LÜNEBURG

Geschäftsverständnis

Sie haben das Unternehmen nun in groben Zügen kennengelernt. Nun wird Ihnen ein umfangreicher Datensatz zur Verfügung gestellt. Sie machen sich daher zunächst ein eigenes Bild von den aktuellen Frachtdaten.

Aufgabenstellung:

- Öffnen Sie den Datensatz und machen Sie sich mit dem Umfang vertraut.
- Welche (auch externe) Informationen könnten zur Vorhersage des Frachtpreises noch nützlich sein?
- Sammeln Sie Ideen, wie aus den vorhanden Daten weitere Attribute extrahiert werden können.

Datei: frachtprognose.xlsx

(https://ki-werkstatt.info/upload/tutoring/frachtprognose.zip)

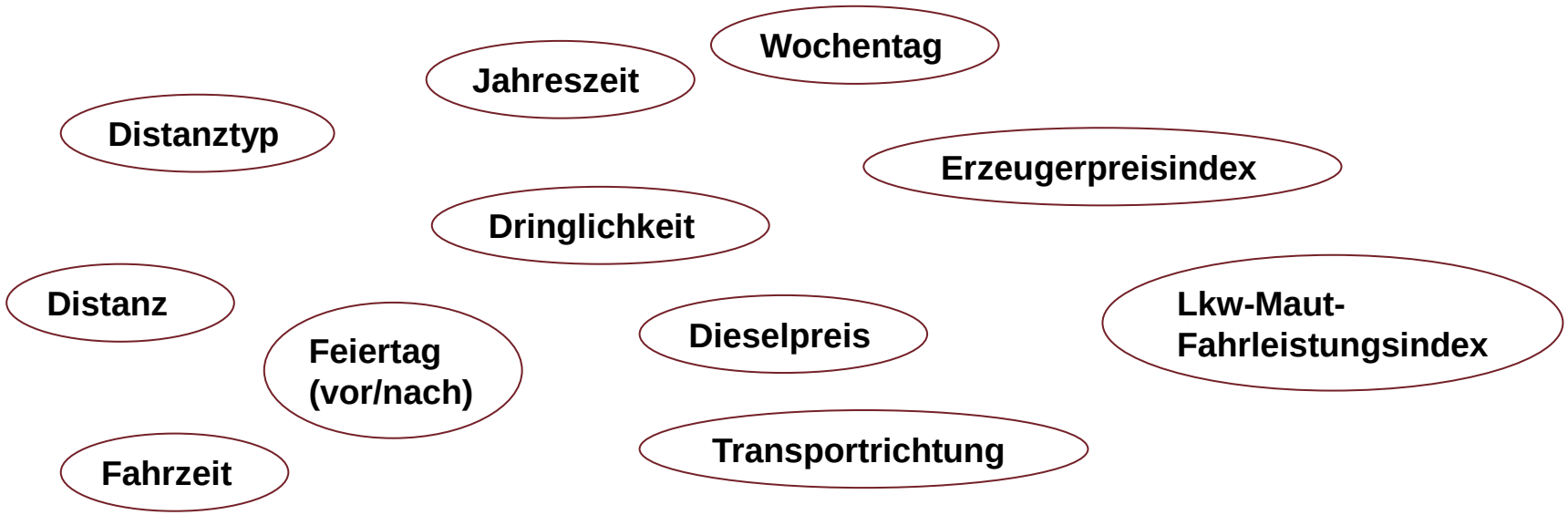
Gemeinsame Ideensammlung im Plenum

LEUPHANA
UNIVERSITÄT LÜNEBURG

Geschäftsverständnis

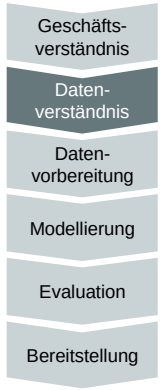
Sie haben das Unternehmen nun in groben Zügen kennengelernt. Nun wird Ihnen ein umfangreicher Datensatz zur Verfügung gestellt. Sie machen sich daher zunächst ein eigenes Bild von den aktuellen Frachtdaten.

Aufgabenstellung:
- Welche (auch externe) Informationen könnten zur Vorhersage des Frachtpreises noch nützlich sein?
- Sammeln Sie Ideen, wie aus den vorhanden Daten weitere Attribute extrahiert werden können.



LEUPHANA
UNIVERSITÄT LÜNEBURG

Datenverständnis



Sie haben ein erstes Verständnis über das Geschäft und relevante Daten erlangt. Um ein neues Prognoseverfahren entwickeln zu können, ist es darüber hinaus bedeutend, die verfügbaren Daten zu verstehen.

Aufgabenstellung:
Untersuchen Sie bitte den Datensatz und beantworten Sie dazu die nachstehenden Leitfragen.

- Wie sind die Frachtpreise verteilt? Betrachten Sie dazu das Histogramm.
- Gibt es Ausreißer? Fehlen Daten? Gibt es fehlerhafte Werte?
- Was könnten mögliche Ursachen für fehlerhafte Werte sein?

10 min Bearbeitungszeit

Bilden Sie ggf. kleine Gruppen

LEUPHANA
UNIVERSITÄT LÜNEBURG

Ergebnisdiskussion zum Datenverständnis

LEUPHANA
UNIVERSITÄT LÜNEBURG

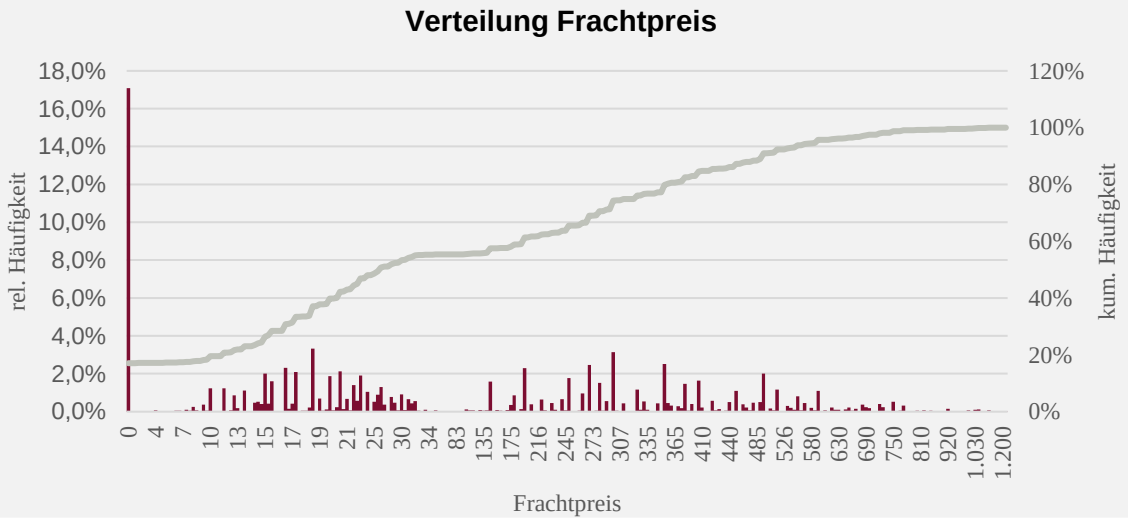
Lösungsskizze zum Datenverständnis

Label-Attribut „Frachtpreis“

- unbereinigter Frachtpreis eines Auftrags in Euro
- Datentyp: Real

- Spanne: 0 – 1400 Euro
- Mittelwert: 177,95 Euro
- Standardabweichung: 216,65 Euro
- 55 % der Aufträge des Betrachtungszeitraums lagen bei weniger als 40 Euro
- 17 % der Aufträge haben einen Frachtpreis von 0 Euro

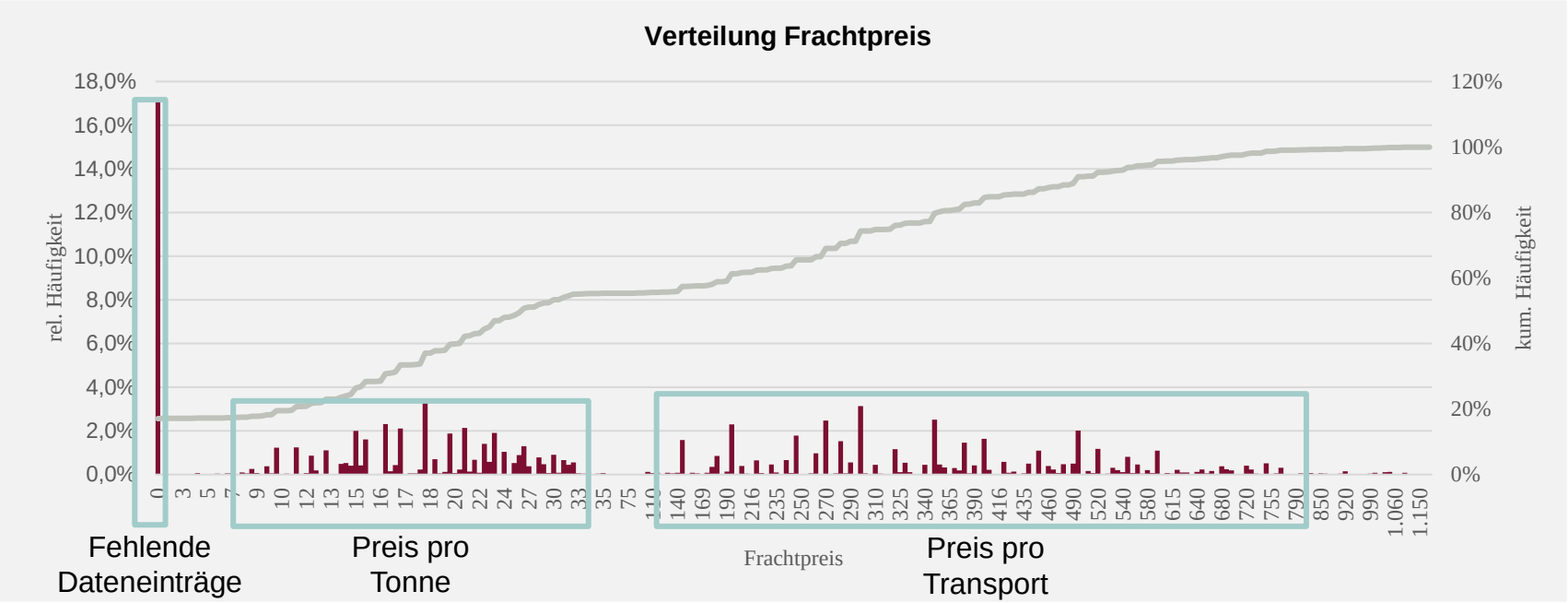
→ Die Verteilung des Frachtpreises muss bereinigt werden!



- Große Ausreißer beim Frachtpreis
- Viele Werte fehlen, die mit 0 Euro angenommen wurden

LEUPHANA
UNIVERSITÄT LÜNEBURG

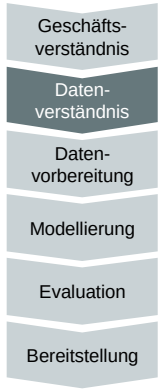
Lösungsskizze zum Datenverständnis



Aufbereitung der Daten ist notwendig.

LEUPHANA
UNIVERSITÄT LÜNEBURG

Datenverständnis



Im nächsten Schritt ist es erforderlich anhand der gewonnenen Erkenntnisse den Datensatz zu erweitern und erneut zu untersuchen. Sie erhalten nun einen aufbereiteten Datensatz.

Aufgabenstellung:
Untersuchen Sie bitte den aufbereiteten Datensatz und beantworten Sie dazu die nachstehenden Leitfragen.

- Wie sind die Frachtpreise pro km verteilt?

- Wie viele Einträge hat der Datensatz und über welche Attribute/ Features werden diese beschrieben?
- Gibt es Attribute, welche Sie zum Zeitpunkt der Prognose nicht kennen können? Wenn ja, welche?

- Erkennen Sie in Bezug auf die Distanz der Frachtpreise pro km relevante Muster in den Auftragsdaten? Erstellen Sie ggf. eine geeignete Grafik.

Datei: frachtprognose_aufbereitet.xlsx

(https://ki-werkstatt.info/upload/tutoring/frachtprognose_aufbereitet.zip)

10 min Bearbeitungszeit

Bilden Sie ggf. kleine Gruppen

Datenverständnis

- Ein **Histogramm** ist eine Grafik, in der Sie Häufigkeiten oder die Häufigkeitsdichte bestimmter Ausprägungen einer Variable ablesen können. Trotz der Ähnlichkeit handelt es sich hierbei nicht um ein Säulendiagramm. Bei einem Säulendiagramm liegen die Daten bereits gruppiert vor (z.B. „männlich“, „weiblich“, „divers“), bei einem Histogramm müssen sie hingegen erst in Gruppen eingeteilt werden (z.B. Einteilung des Alters in 10-Jahres-Klassen). Die verschiedenen Gruppen werden anschließend nebeneinander in Rechtecken dargestellt. Sie können dabei entscheiden, ob Sie die absolute oder die relative Häufigkeit der Merkmalsgruppen in deinem Histogramm abbilden.

- RapidMiner Histogramm erstellen:
  1. Im RapidMiner klicken Sie “Blank process“ und importieren Sie den gewünschten Datensatz per Drag and Drop.
  2. Wählen Sie die Zellen aus die Sie importieren möchten, hier alle.
  3. Überprüfen Sie, ob der Richtige Datentyp ausgewählt wurde.
  4. Wählen Sie einen Speicherplatz für das Projekt aus.
  5. Klicken Sie nun auf „Visualization“. Nehme Sie nun folgende Einstellungen vor:

  Histogramm Plot:
  Plot type: Histogramm
  Value Columns: Frachtpreis pro km
  Colour: -

  Zusatz Scatter Plot:
  Plot type: Scatter / Bubble
  X-Axis: Distanze
  Value Columns: Frachtpreis pro km
  Colour und Size: -
  Über den Plot style können Sie zudem noch die Größe der Point marker size einstellen (2 ist optimal)

Datei: frachtprognose_aufbereitet.xlsx

(https://ki-werkstatt.info/upload/tutoring/frachtprognose_aufbereitet.zip)

10 min Bearbeitungszeit

Bilden Sie ggf. kleine Gruppen

LEUPHANA
UNIVERSITÄT LÜNEBURG

Ergebnisdiskussion zum Datenverständnis

LEUPHANA
UNIVERSITÄT LÜNEBURG

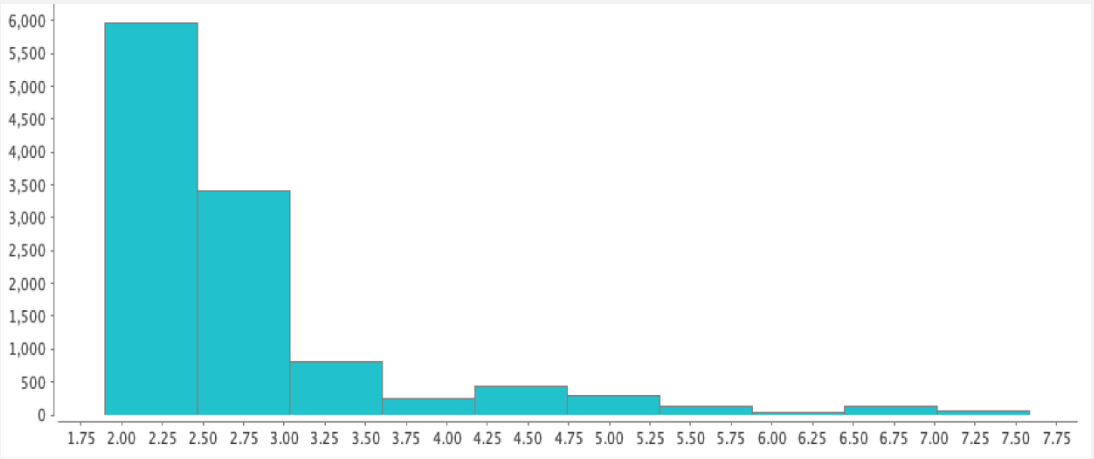
Lösungsskizze zum Datenverständnis

Label-Attribut „Frachtpreis pro km“

- Bereinigter Frachtpreise pro km eines Auftrags in Euro
- Datentyp: Real

- Spanne: 1,89 – 7,59 Euro
- Mittelwert: 2,73 Euro
- Standardabweichung: 0,91 Euro
- 75% der Aufträge des Betrachtungszeitraums lagen bei weniger als 2,75 Euro pro km
- Es gibt wesentlich mehr Aufträge mit geringeren Frachtpreisen pro km als mit hohen Frachtpreisen pro km

→ Die Verteilung des Frachtpreises pro km sieht nun realistischer aus.



- Es fehlen keine Werte.
- Elemente des Attributs variieren stark.

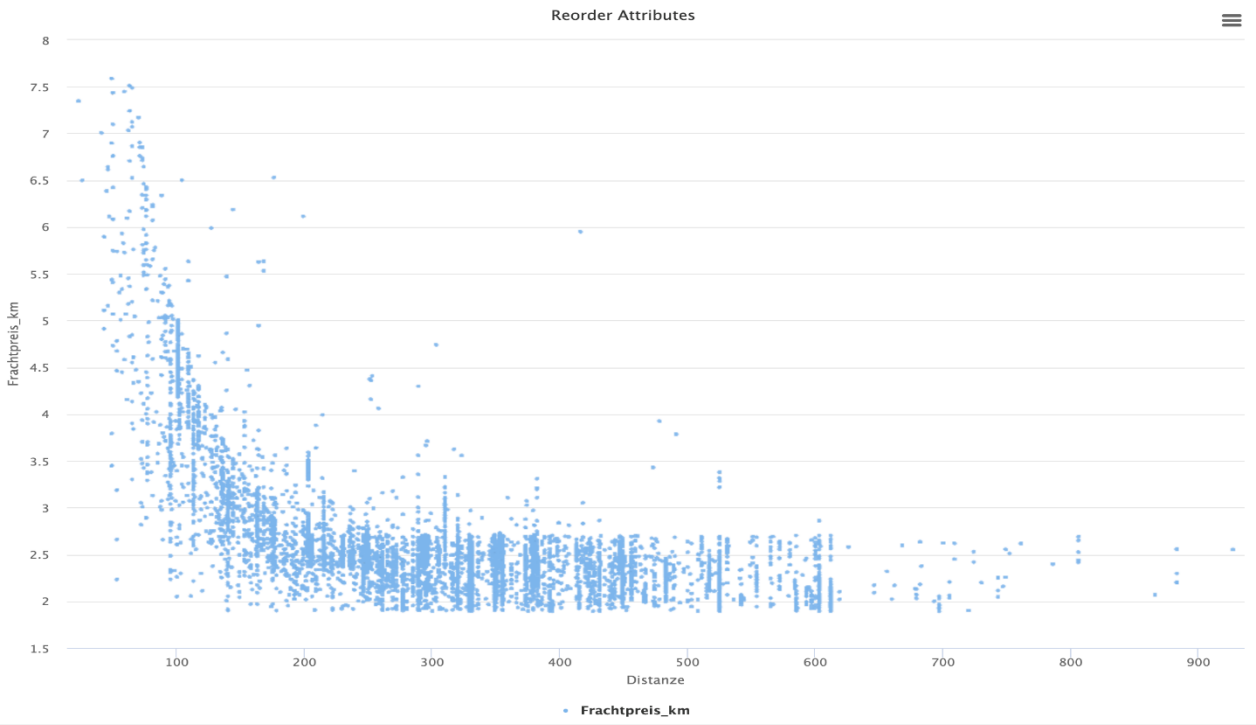
- Anzahl beschreibender Attribute: 24
- 11.472 Aufträge

LEUPHANA
UNIVERSITÄT LÜNEBURG

Lösungsskizze zum Datenverständnis

Frachtpreis pro km – über die Distanz

- Distanz ist das wichtigste beschreibende Attribut für den Frachtpreis



- Wenige Fahrten unter 100 km
- Starke Streuung vor allem auf kürzeren Distanzen
- Zunehmende Annäherung an einen stabilen Mittelwert für mittlere bis längere Strecken

LEUPHANA
UNIVERSITÄT LÜNEBURG

Datenvorbereitung

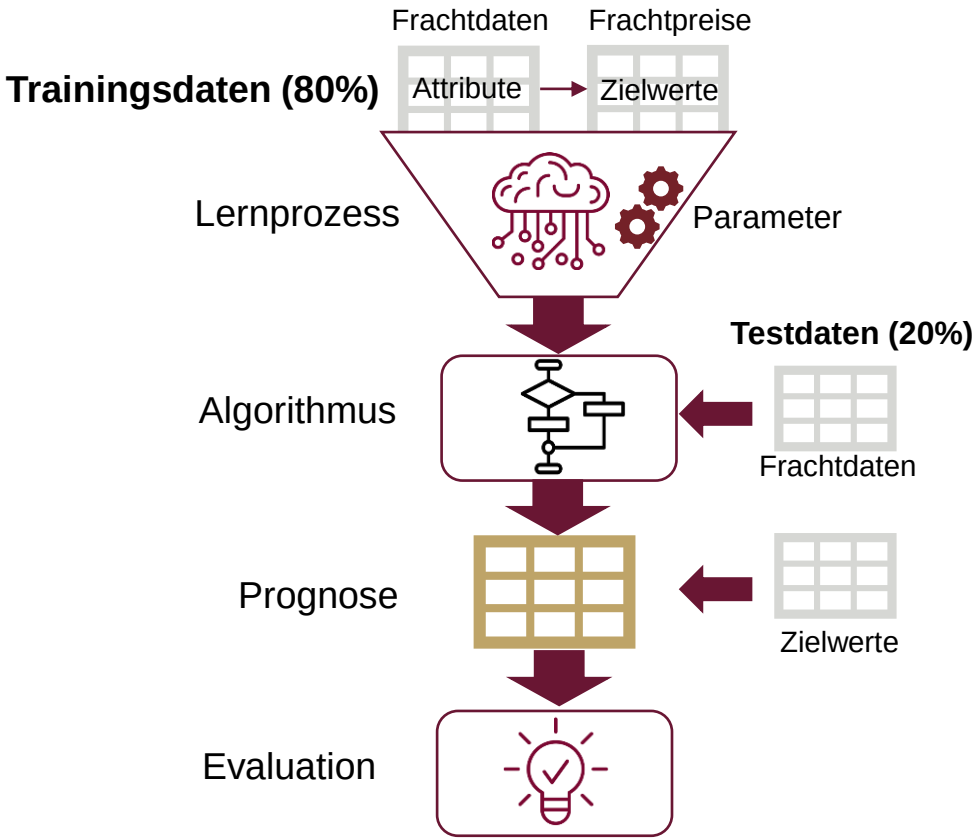
Nachdem Sie sich mit dem verfügbaren Datensatz vertraut gemacht haben, gilt es diesen für die Modellierung vorzubereiten.

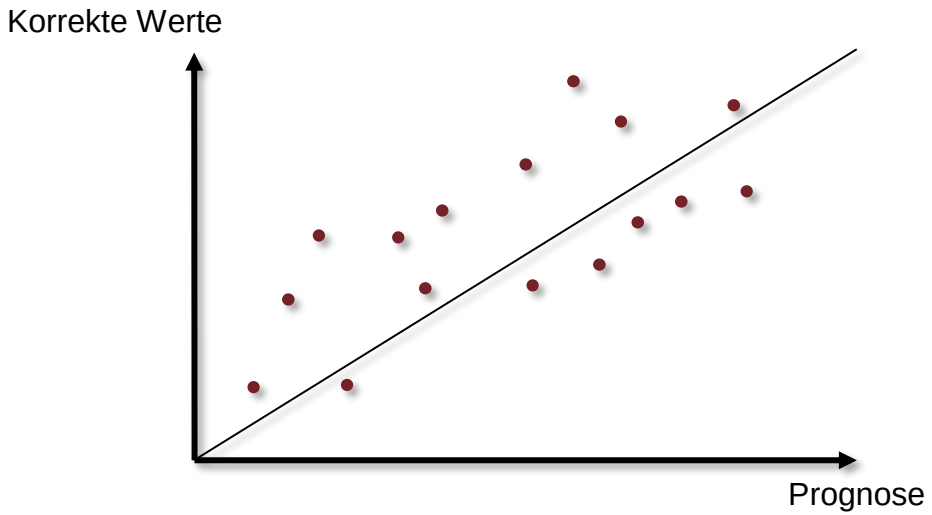
Aufgabenstellung: Wie würden Sie bei der Datenvorbereitung vorgehen?

- Würden Sie Datenpunkte oder Attribute ausschließen oder ersetzen?
- Welche Attribute sind aus ihrer Sicht die wichtigsten für eine Prognose des Frachtpreises?
- Was müssen Sie hinsichtlich der Formatierung der Daten beachten? (Datentypen, Skalen)

Gemeinsame Diskussion (5 min)

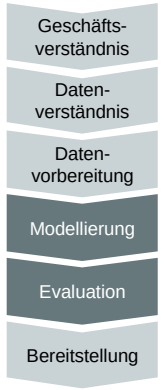
Modellierung: Prinzip einer Regression anhand von Supervised Learning





LEUPHANA
UNIVERSITÄT LÜNEBURG

Modellierung & Evaluation



Die vorbereitenden Schritte des CRISP-DM sind abgeschlossen. Es folgt der eigentliche Modellierungsprozess. Nutzen Sie hierzu die Software Rapid Miner und den aufbereiteten Datensatz.

Aufgabenstellung:
Öffnen Sie das Auto-ML Tool Rapid Miner
Folgen Sie den unten genannten Schritten zu Ihrer ersten Prognose.

- Laden Sie den Modell-Datensatz in Rapid Miner hoch.
- Ersetzen Sie dabei Fehlerwerte als „Missing Values“
- Kontrollieren Sie, ob die richtigen Datentypen ausgewählt sind
- Erstellen Sie eine Prognose und wählen Sie das entsprechende Label-Attribut aus.
- Nutzen Sie für Ihre Prognose mehrere Algorithmen.

Datei: frachtprognose_aufbereitet.xlsx

(https://ki-werkstatt.info/upload/tutoring/frachtprognose_aufbereitet.zip)

30 min gemeinsame Bearbeitungszeit

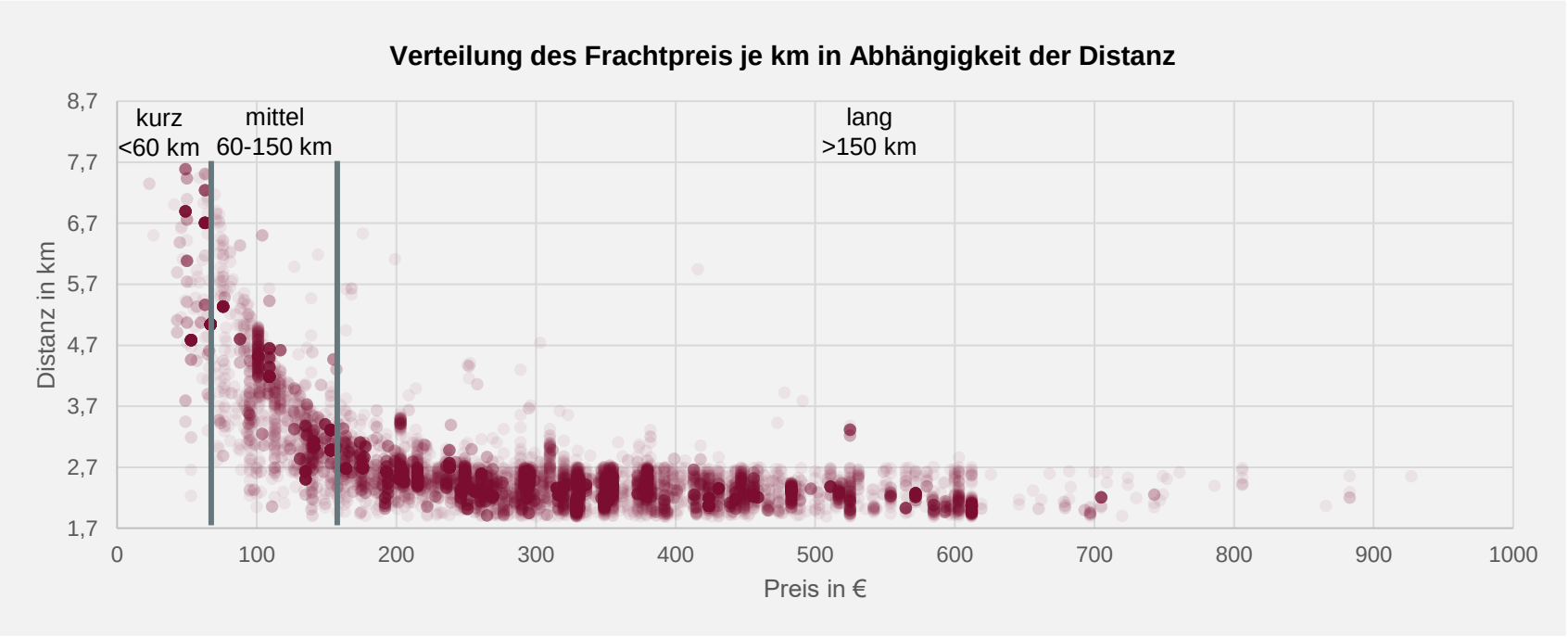
Quelle: Mierswa, et al. (2020).

LEUPHANA
UNIVERSITÄT LÜNEBURG

Bonusaufgabe: Benchmark

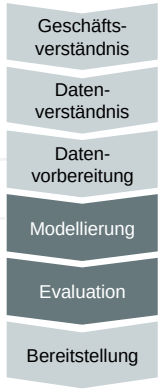
Aufgabenstellung:
Erstellen Sie in Excel ein Benchmark-Modell und ermitteln Sie den mittleren, absoluten Fehler

- Bilden Sie Mittelwerte des Frachtpreises/km für kurze, mittlere und lange Fahrten der letzten 1500 Aufträge
- Ordnen Sie diese Mittelwerte den entsprechenden Fahrten gemäß Ihrer Distanz in einer neuen Spalte zu
- Berechnen Sie die Differenz zum tatsächlichen Frachtpreis/km und bilden Sie den absoluten Mittelwert



LEUPHANA
UNIVERSITÄT LÜNEBURG

Modellierung & Evaluation



Bearbeitungshinweise für die Modellierung mit RapidMiner

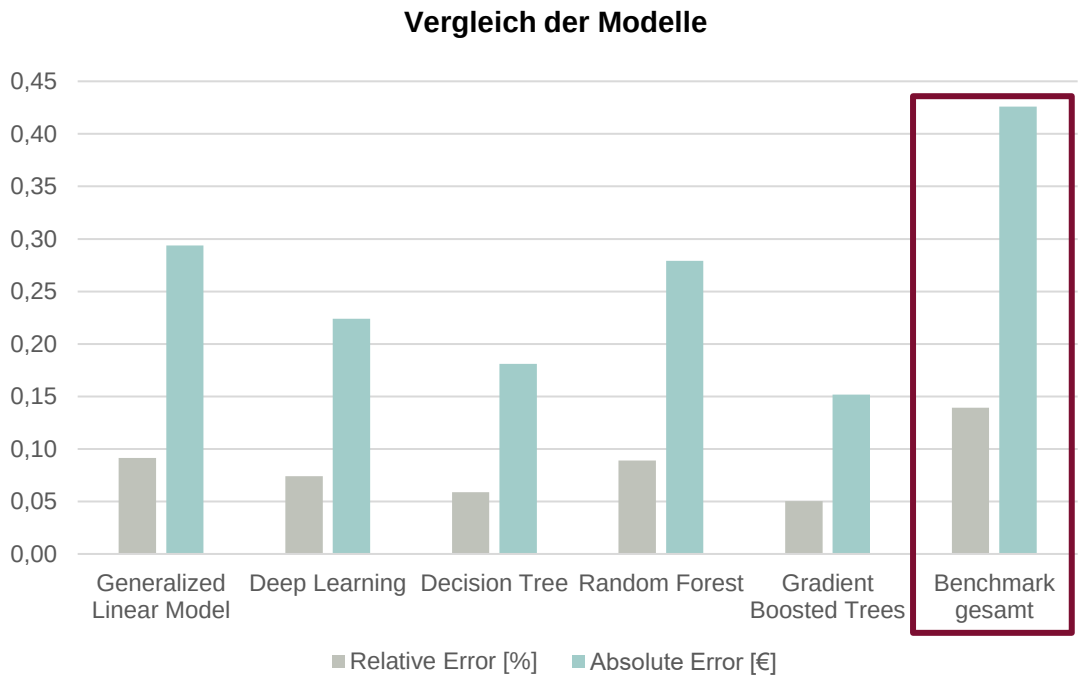
Split des Datensatzes in Trainings und Testdaten	▪ Durch RapidMiner voreingestellt
Regressionsmodelle (Supervised Learning)	▪ Lineare Regression ▪ Entscheidungsbaum ▪ Random Forest ▪ Gradient Boosted Trees ▪ Deep Learning
Einstellung der Modellparameter	▪ Durch RapidMiner automatisch optimiert
Automatische Feature Selektion und Generierung	▪ Ausgeschaltet
Evaluation	▪ Abweichung zwischen Ist-Frachtpreis pro km und Prognosewert des Testdatensatzes

LEUPHANA
UNIVERSITÄT LÜNEBURG

Ergebnisdiskussion zur Modellierung

LEUPHANA
UNIVERSITÄT LÜNEBURG

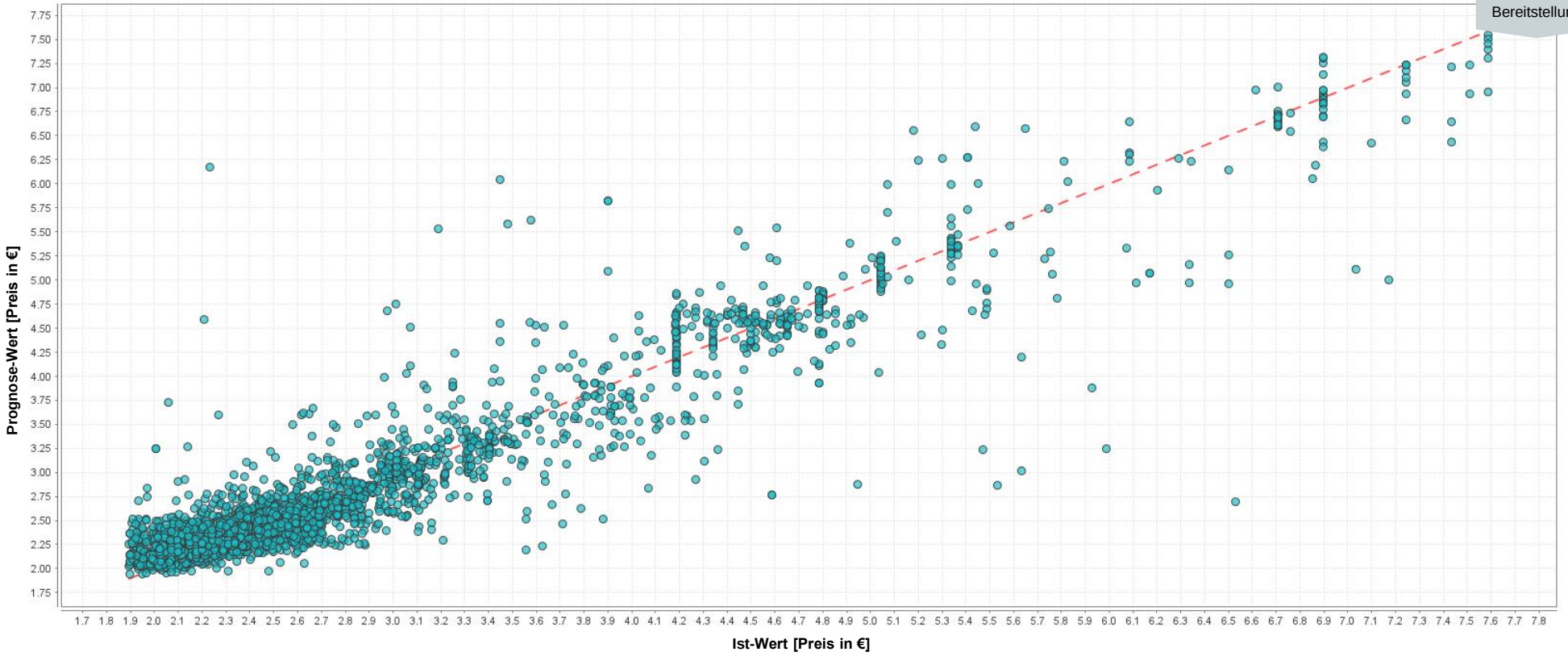
Evaluation



LEUPHANA
UNIVERSITÄT LÜNEBURG

Evaluation

Kritische Betrachtung der Modellperformance am Beispiel des Gradient Boosted Tree



LEUPHANA
UNIVERSITÄT LÜNEBURG

Lösungsvideo für die Modellierung in RapidMiner

https://ki-werkstatt.info/upload/frachtdaten_video.zip

Schlussfolgerung

- Insbesondere in komplexen Logistikumfeldern können ML-Verfahren im Vergleich zu klassischen Verfahren präzisere Frachtpreise bestimmen.
- Excel basiert auf einer generischen Programmierung und besitzt somit allgemeine Funktion, wobei RapidMiner speziell für die ML-Anwendung entworfen wurde.
- Datenauswahl und -qualität ist entscheidend. Hierbei ist Prozessverständnis von großer Bedeutung
- Die ML-Modelle übertreffen die manuell kalkulierte Benchmark Methode.
- Einfache Lösungen und Anwendungsprogramme existieren.

Arbeitsgruppe Produktionsmanagement

Leuphana Universität Lüneburg
Institut für Produkt- und Prozessinnovation (PPI)

Universitätsallee 1
21335 Lüneburg
www.leuphana.de/ppi

Literaturverzeichnis

- **Capgemini Research Institute (2019):** *Scaling AI in Manufacturing Operations: A Practitioners' Perspective.* In Capgemini Research Institute.
- **Chapman, Pete; Clinton, Julian; Kerber, Randy; Khabaza, Thomas; Reinartz, Thomas; Shearer, Colin; Wirth, Rüdiger (2000):** CRISP-DM 1.0. Step-by-step data mining guide. Hg. v. The CRISP-DM consortium.
- **Denkena, B., Dittrich, M.-A., Noske, H., Kramer, K., Schmidt, M. (2021):** Anwendungen des maschinellen Lernens in der Produktion aus Auftrags- und Produktsicht – ein Überblick, ZWF 05/2021.
- **Gallina, V., Lingitz, L., Karner, M. (Eds.) (2019)**: A New Perspective of the Cyber-Physical Production Planning System.
- **Green, T., Rokoss, A., Kramer, K., Schmidt, M. (2022)**:Application of Machine Learning on Transport Spot Rate Prediction in the Recycling Industry. In: Herberger, D.; Hübner, M. (Eds.): Proceedings of the Conference on Production Systems and Logistics: CPSL 2022.
- **James, Gareth; Witten, Daniela (2013):** An introduction to statistical learning: with applications in R. New York.
- **Kuhn, Max; Johnson, Kjell (2016):** Applied predictive modeling. Corrected at 5th printing. New York: Springer.
- **Kramer, K., Wagner, C., Schmidt, M. (2020):** Machine Learning-Supported Planning of Lead Times in Job Shop Manufacturing. In: Advances in Production Management Systems: The Path to Digital Transformation and Innovation of Production Management Systems. Springer, Band 1, S. 363-370.
- **Mierswa, Ingo; Klinkenberg, Ralf (2020):** RapidMiner Studio (9.3). Hg. v. RapidMiner Inc. Online verfügbar unter https://rapidminer.com/, zuletzt geprüft am 27.11.2020.
- **Schmidt, M.; Maier, J. T.; Grothkopp, M. (2020):** Eine bibliometrische Analyse: Produktionsplanung und -steuerung und maschinelles Lernen. wt Werkstattstechnik online 110 (2020) 4.
- **Suthaharan, Shan (2016):** Machine Learning Models and Algorithms for Big Data Classification. Thinking with Examples for Effective Learning. 1st ed. 2016. Boston, MA: Springer US (Integrated Series in Information Systems, volume 36).
- **Witten, Ian H.; Frank, Eibe (2011):** Data Mining. Practical Machine Learning Tools and Techniques. San Diego, CA, USA: Elsevier Science & Technology Books.