

Backpack Price Prediction – Kaggle Projekt

1. Beschreibung des Datensatzes

Der Datensatz umfasst Informationen zu verschiedenen Rucksack- und Taschenmodellen. Jede Beobachtung enthält mehrere numerische sowie kategoriale Merkmale. Ziel dieser Aufgabe ist es, ein Vorhersagemodell zu entwickeln, das den Preis einer Tasche basierend auf ihren Eigenschaften möglichst genau schätzt.

Die zentralen Merkmale im Datensatz sind:

- Numerisch: Compartments, Weight Capacity (kg)
- Kategorial: Brand, Material, Size, Laptop Compartment, Waterproof, Style, Color

2. Baseline-Modell – Lineare Regression

Als Ausgangspunkt wurde ein einfaches Modell auf Basis der linearen Regression erstellt. Diese Methode nimmt an, dass zwischen den Eingangsvariablen und dem Zielwert (Preis) ein linearer Zusammenhang besteht.

Datenverarbeitung:

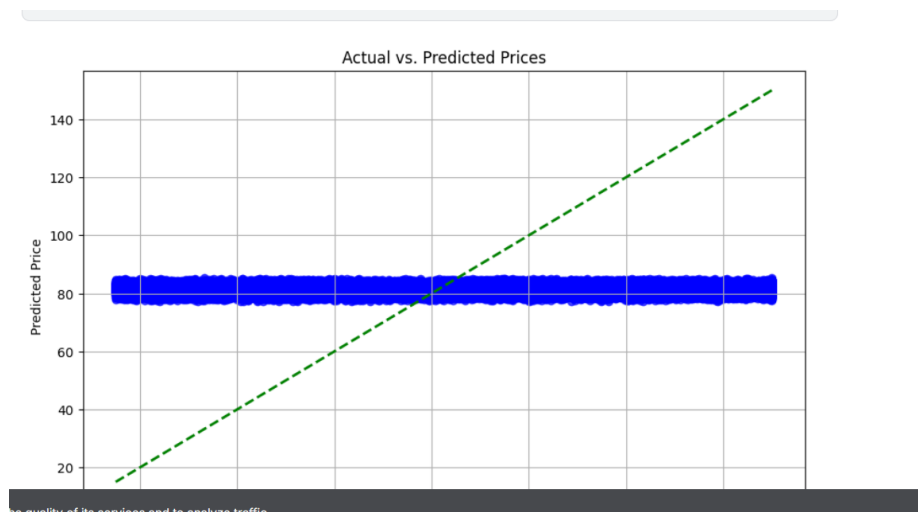
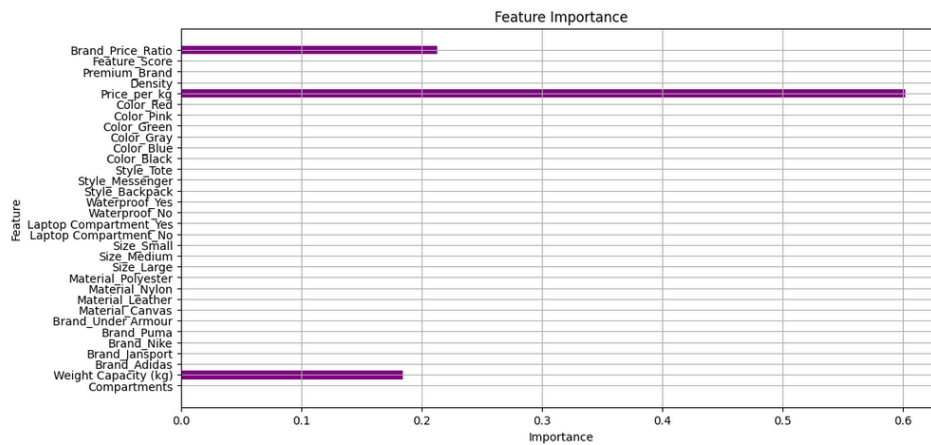
- Fehlende numerische Werte wurden durch den Mittelwert der jeweiligen Spalte ersetzt.
- Für fehlende Werte bei kategorialen Merkmalen wurde die häufigste Ausprägung verwendet.
- Kategoriale Merkmale wurden per One-Hot-Encoding in numerische Form gebracht.

Ergebnisse des Modells:

- MSE: 1514.83
- MAE: 33.65
- R^2 : 0.001

Die sehr geringe R^2 -Zahl zeigt, dass das Modell kaum in der Lage ist, Preisunterschiede sinnvoll zu erklären. Eine Verbesserung durch andere Ansätze ist notwendig.

Visualisierung – Baseline-Modell:



Kaggle-Notebook: Lineare Regression

<https://www.kaggle.com/code/mlinadumitrescu/uebung-i>

3. Verbesserungen – Erweiterung der Merkmale und Modellwahl

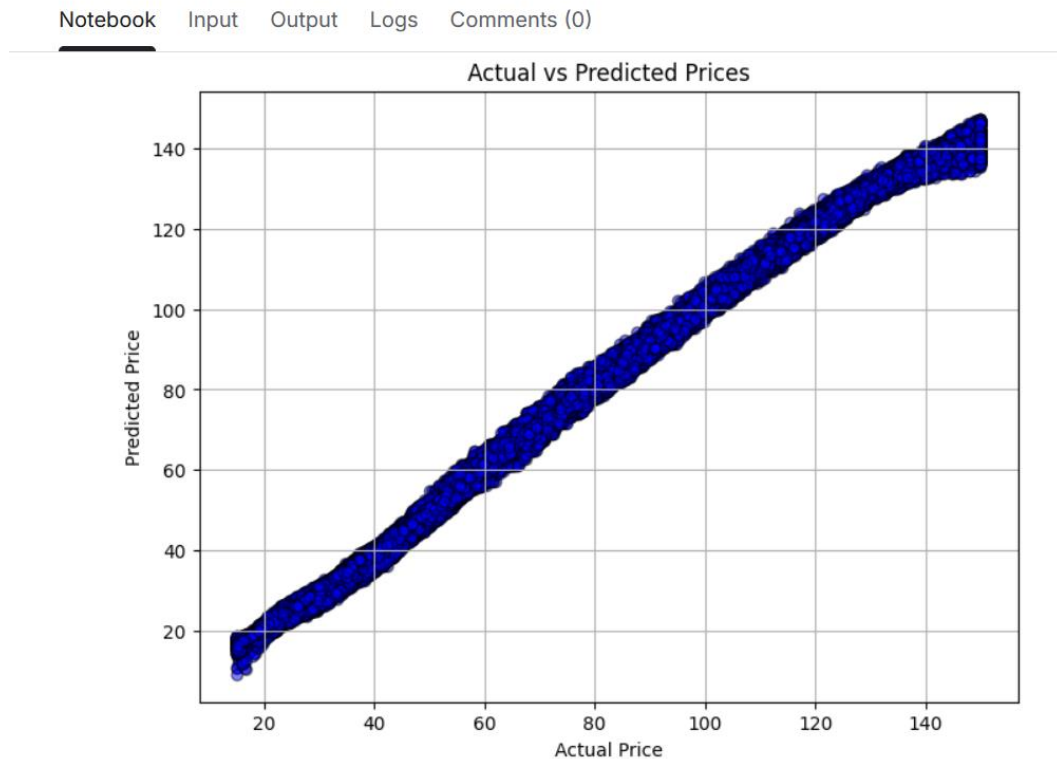
Um die Vorhersagegenauigkeit zu steigern, wurden zusätzliche Merkmale generiert, die komplexere Zusammenhänge besser abbilden können.

Neue Merkmale:

1. Price per kg – gibt Aufschluss über das Preis-Leistungs-Verhältnis bezogen auf die Kapazität
2. Density – Verhältnis von Tragkraft zur Anzahl der Fächer
3. Premium_Brand – binäre Variable für hochwertige Marken
4. Feature_Score – Anzahl spezieller Eigenschaften wie z. B. Laptopfach, wasserdicht etc.
5. Brand_Price_Ratio – Verhältnis des durchschnittlichen Markenpreises zur Price-per-kg-Metrik

Diese neuen Merkmale sollen zusätzliche Informationen liefern, die für die Preisabschätzung relevant sein können.

Visualisierung – Feature Importance:



4. Alternativmodell – Gradient Boosting Regressor

Anstelle der linearen Regression wurde ein Gradient Boosting Regressor verwendet, da er besser mit nichtlinearen Zusammenhängen umgehen kann und in vielen praktischen Anwendungsfällen eine höhere Genauigkeit liefert.

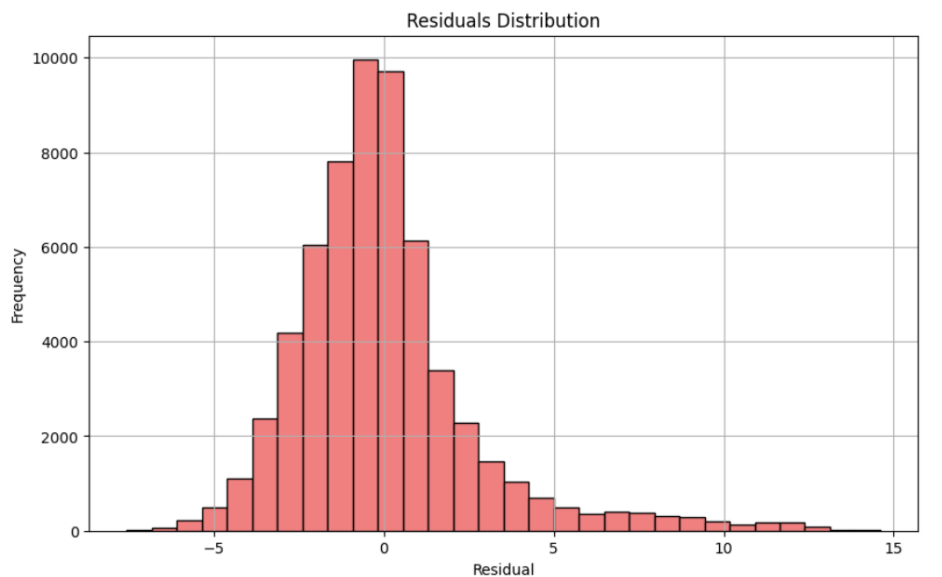
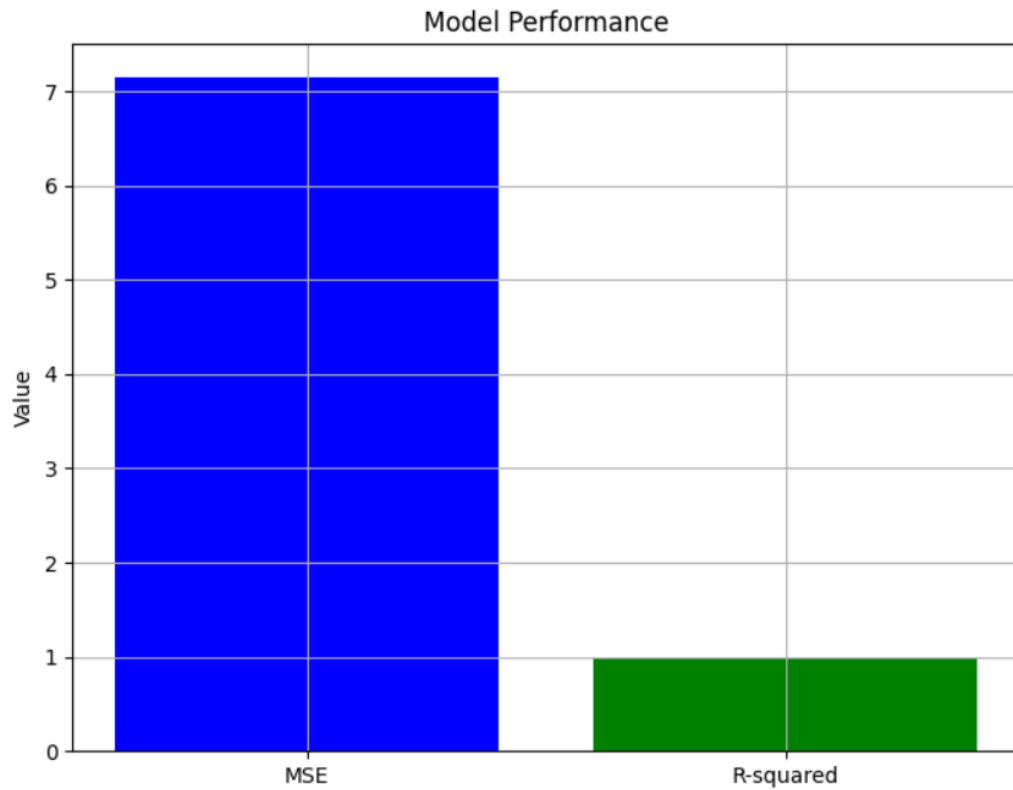
Vorteile des Gradient Boosting:

- Kombination vieler schwacher Modelle zu einem starken Modell
- Sehr gute Genauigkeit bei strukturierten Daten
- Flexible Anpassung durch Hyperparameter

Ergebnisse des Gradient Boosting Modells:

- MSE: 7.14
- MAE: 1.84
- R^2 : 0.995

Visualisierung – Gradient Boosting:



Kaggle-Notebook: Gradient Boosting

<https://www.kaggle.com/code/mliadumitrescu/uebung-ii>

Vergleichung zwischen den Modellen im Preisvergleich:



1. Lineare Regression – Vorhersagen waren ungenau und zeigten kaum Korrelation mit echten

Preisen.

2. Gradient Boosting – Punkte lagen sehr nah an der Diagonalen, was auf hohe Vorhersagegenauigkeit hinweist.

5. Fazit

Das erste Modell war nicht zuverlässig. Erst durch neue Merkmale und den Einsatz eines anderen Modells (Gradient Boosting Regressor) konnte die Vorhersage deutlich verbessert werden. Damit die Ergebnisse wirklich sicher sind, sollte man das Modell auch mit neuen Daten testen.

All		Successful	Selected	Errors	Recent ▾		
Submission and Description					Private Score ⓘ	Public Score ⓘ	Selected
	Uebung-II - Version 2				74.89601	74.93073	<input type="checkbox"/>
	Complete (after deadline) · 16s ago						
	Uebung-I - Version 2				38.94376	39.14692	<input type="checkbox"/>
	Complete (after deadline) · 6m ago						