**Einreichung Projektvorschlag**

**für**

**Data Analytics**

**(28.07.25 – 22.08.25)**

**Projektersteller/Gruppe:**

**Wu, Guangfa**

**Daten:**  [**https://www.kaggle.com/datasets/hellbuoy/car-price-prediction**](https://www.kaggle.com/datasets/hellbuoy/car-price-prediction)

**Projektidee, Beschreibung:**

**Problembeschreibung:**

Ein chinesisches Automobilunternehmen, Geely Auto, plant den Eintritt in den US-amerikanischen Markt, indem es dort eine Produktionsstätte errichtet und Autos lokal herstellt, um mit amerikanischen und europäischen Konkurrenten zu konkurrieren. Dazu hat Geely ein Beratungsunternehmen für die Automobilbranche beauftragt, um die Faktoren zu analysieren, die den Preis von Autos beeinflussen. Insbesondere möchten sie verstehen, welche Faktoren den Preis von Autos auf dem amerikanischen Markt bestimmen, da diese sich deutlich vom chinesischen Markt unterscheiden könnten.

**Das Unternehmen möchte wissen:**

* Welche Variablen sind signifikant für die Vorhersage des Autopreises?
* Wie gut erklären diese Variablen den Preis eines Autos?

Basierend auf verschiedenen Marktforschungen hat das Beratungsunternehmen einen umfangreichen Datensatz mit unterschiedlichen Fahrzeugtypen auf dem US-amerikanischen Markt zusammengetragen.

**Lösung der Aufgabestellung:**

Ein Modell soll entwickelt werden, das den Preis von Autos basierend auf den verfügbaren unabhängigen Variablen am besten vorhersagt. Da ein Auto aus verschieden Komponenten besteht, ist es nur sinnvoll, alle Komponenten zusammen zu betrachten und ein Preismodell zu erstellen. Gleichzeitig kann auch untersucht werden welche einzelnen Faktoren den Preis beeinflussen.

**Anwendung:**

Das Modell soll dem Management dabei helfen zu verstehen:

* Wie sich der Preis je nach Veränderung einzelner Merkmale verändert
* Welche Stellschrauben (Design, Strategie etc.) genutzt werden können, um bestimmte Preisniveaus zu erreichen.
* Zusätzlich ist das Modell ein hilfreiches Werkzeug, um die Preisdynamik in einem neuen Markt wie den USA zu verstehen.

**Projektschritte, Aufgaben zum Projekt (falls schon vorhanden):**

Folgende Schritte sind zu unternehmen, um das definierte Ziel zu erreichen

Es werden Arbeitsschritte nach CRISP-DM vorgesehen.

1. Business Understanding (Aufgabendefinition)
2. Data Understanding (Auswahl der relevanten Datenbestände)

2.1 Deskriptive Analyse

2.2 Explorative Datenanalyse (EDA)

1. Data Preparation (Datenaufbereitung)

3.1. Datenaufbereitung

3.2. Umwandlung von kategorischen Daten in numerischen Daten

1. Modeling (ML-Maschinelles Lernen)

4.1 Ein lineares Modell mit One-Hot Encoding für kategorische Spalten

4.2 Ein lineares Modell mit ordinal Encoding für kategorische Spalten

4.3 Modell mit zusätzlichen Spalten aus quadratischen Werten von numerischen Spalten

4.4 Modell mit flexibler Spaltenauswahl durch Dashboard-Anwendung

1. Evaluation (Bewertung und Interpretation der Ereignisse)

5.1 Prädiktive Analyse

5.2 Modellbewertung

1. Deployment (Anwendung der Ergebnisse)

Dashoard für Preis-Vorhersage

**3. Data Preparation (Datenaufbereitung)**

Für Machine-Learning ist sehr wichtig, eine gute aufbereitete Daten zu haben, um ein allgemein aussagkräftiges Modell zu tranineren.

**3.1. Datenaufbereitung**

citympg und hightwaympg: die Verteilung wird noch mal genau angeschaut. Die Aussereisser werden bereinigt.

**3.2. Umwandlung von kategorischen Daten in numerischen Daten**

Anzahl von Türen und Zylindern sind als kategorische Daten dargestellt. sie werden in Zahlen umgewandelt(schon erledigt in der Deskriptiven Analyse)

**3.3. Zerlegung von kategorischen Daten in einer Spalten**

In der Spalte 'CarName' sind Information über Marke und Modelle zusammengefasst. Sie sollen in zwei Spalten dargestellt werden

**3.4. Analyse von Ausreißern in der Spalten citympg und hightwaympg**

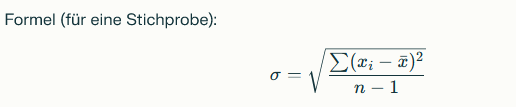
**Statistische Betrachtung:**

Was ist die Standardabweichung (σ oder *std*)?

Die Standardabweichung misst, wie stark die einzelnen Werte einer Datenmenge im Durchschnitt von deren Mittelwert (Durchschnitt) abweichen.

* Kleine Standardabweichung → die Werte liegen eng um den Mittelwert herum (geringe Streuung).
* Große Standardabweichung → die Werte sind weit vom Mittelwert verteilt (hohe Streuung).

💡 Formel (für eine Stichprobe):

****

wobei  
xi = einzelner Wert,  
xˉ = Mittelwert,  
n = Anzahl der Werte.

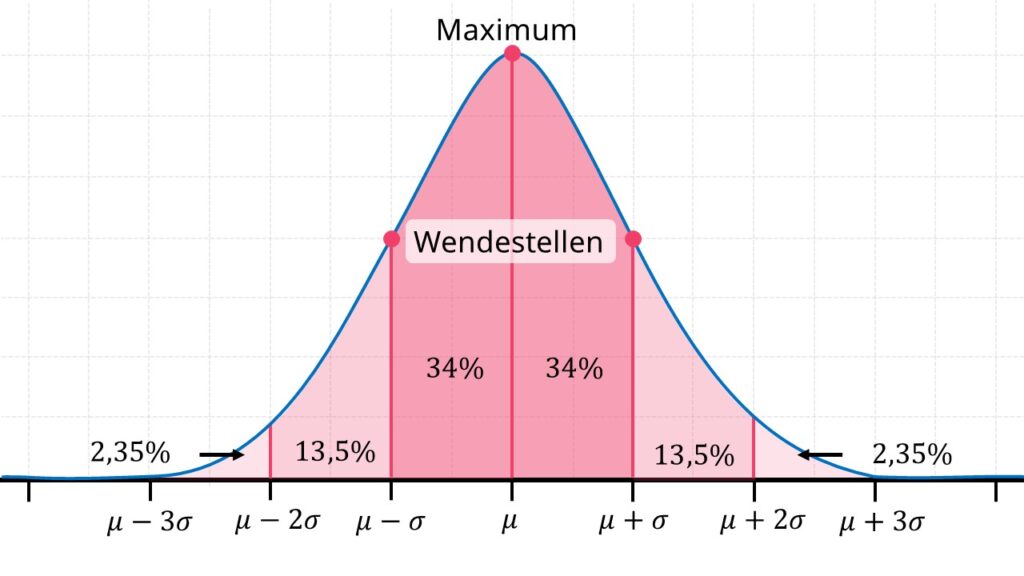
Was bedeutet „3-fache Standardabweichung“?

Wenn die Daten annähernd normalverteilt (Glockenkurve) sind, gilt die sogenannte 68–95–99,7-Regel:

* ±1 σ → ca. 68 % aller Werte liegen innerhalb einer Standardabweichung vom Mittelwert.
* ±2 σ → ca. 95 % aller Werte liegen innerhalb von zwei Standardabweichungen.
* ±3 σ → ca. 99,7 % aller Werte liegen innerhalb von drei Standardabweichungen.

Das heißt:  
Werte, die mehr als 3 Standardabweichungen vom Mittelwert entfernt sind, treten sehr selten auf (~0,3 % Wahrscheinlichkeit). In der Statistik werden diese oft als Ausreißer betrachtet.

die Werte außer ±3σ werden beseitigt.



Bbbb