**Übersicht Fehlerklassenanalyse Price Prediction**

**Strukturierte Explanations - Importance**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Attribute** | **Count** | **Mean Importance** | **Variance** | **Std.** |
| Year | 400 | 0.26 | 0.0023 | 0.048 |
| Km\_driven | 400 | 0.28 | 0.0048 | 0.070 |
| Fuel | 399 | 0.16 | 0.0005 | 0.023 |
| Transimission | 309 | 0.13 | 0.0008 | 0.029 |
| Engine | 255 | 0.11 | 0.0011 | 0.033 |
| Owner | 254 | 0.10 | 0.0006 | 0.025 |
| Max\_Power | 160 | 0.09 | 0.0011 | 0.033 |
| Mileage | 69 | 0.11 | 0.0007 | 0.026 |
| Seats | 38 | 0.08 | 0.0015 | 0.038 |
| Seller\_type | 7 | 0.06 | 0.0006 | 0.024 |
| Torque | 2 | 0.08 | 0.0013 | 0.035 |

**Strukturierter Prompt**:

***System****: You are a helpful AI*

***Human****: {user\_prompt}*

***ai****: {ai\_answer}*

***human****:  Explain your prediction in a structured format, listing attributes of the car you used and how important you deemed them for your prediction. Each attribute should accompanied by the attribute value and a score between 0 and 1 that shows the importance of the attribute for the prediction. All scores in total must add up to 1. Only include attributes that had an influence on your decision.*

*Return the final result as JSON like in this example:*

*[{{"attribute":"name","importance":"0.50","value":"Mahindra XUV500 W10 2WD"}},*

*{{"attribute":"km\_driven","importance":"0.45","value":"70000"}},*

*{{"attribute":"mileage","importance":"0.05","value":"16 kmpl"}}]*

*ONLY return the JSON.*

**Unstrukturierter Prompt**:

***System****: You are a helpful AI*

***Human****: {user\_prompt}*

***ai****: {ai\_answer}*

***human****:  Now explain concisely how you made your prediction and explicitly mention the attributes and values that had a high influence on your decision.*

**Error Class Analysis with 10 false and 2 correct examples**

**Struktur von Examples:**

{Prompt}

{AI Prediction}

{Explanation} – Strukturiert & Unstrukturiert

{Actual Price}

**Baseline Prompt – GPT-3.5-Turbo:**

*Prompt: In the following I will give you a few price prediction tasks together with a prediction decision, details about the decision and the actual price.*

*A wrong prediction is a prediction that deviates by more than 20 percent from the actual price. The Prediction was made by an LLM.*

*Can you please group the wrong decisions into 3 fault categories? Please also indicate how often each one occurs.*

*There are also some correct decisions (deviation of less than 20 percent) in the examples. Please just use them as a reference and don't categorize them.*

*False Predictions:*

*{10 examples}*

*Correct Prediction:*

*{2 examples}*

Baseline Answer:

*Fault Categories:*

*1. Overestimation of Price:*

*- Task 2: Prediction deviates by 50% from the actual price.*

*- Task 3: Prediction deviates by 40% from the actual price.*

*- Task 4: Prediction deviates by 30% from the actual price.*

*- Task 5: Prediction deviates by 36.36% from the actual price.*

*- Task 6: Prediction deviates by 25% from the actual price.*

*- Task 7: Prediction deviates by 23.53% from the actual price.*

*- Task 8: Prediction deviates by 25.93% from the actual price.*

*- Task 9: Prediction deviates by 53.85% from the actual price.*

*- Task 10: Prediction deviates by 25% from the actual price.*

*2. Underestimation of Price:*

*- Task 1: Prediction deviates by 50% from the actual price.*

*3. Inconsistent Weighting of Attributes:*

*- None in the given examples.*

🡪 GPT-3.5-Turbo erkennt Grundlegende Fehlerklassen, ordnet die Beispiele aber vollkommen falsch zu, kann also nicht zwischen überschätzen / unterschätzen unterscheiden

🡪 Zuordnung erscheint zufällig

🡪 Aufgrund der schlechten Antwortqualität wurden weitere Fehlerklassenexperimente nur noch mit GPT-4 gemacht

**Baseline Prompt – GPT-4-Turbo-Preview:**

🡪 Siehe oben

Baseline Answer:

*Based on the provided examples, the wrong predictions can be grouped into three fault categories:*

*1. \*\*Underestimation of Market Value\*\*:*

*- This occurs when the predicted price is significantly lower than the actual market price.*

*- Examples: Task 2 (Prediction: 180000, Actual: 270000), Task 3 (Prediction: 150000, Actual: 210000), Task 4 (Prediction: 350000, Actual: 500000), Task 6 (Prediction: 750000, Actual: 1000000), Task 7 (Prediction: 650000, Actual: 850000), Task 8 (Prediction: 200000, Actual: 270000), Task 10 (Prediction: 900000, Actual: 1200000).*

*- Occurrences: 7*

*2. \*\*Overestimation of Market Value\*\*:*

*- This occurs when the predicted price is significantly higher than the actual market price.*

*- Examples: Task 1 (Prediction: 60000, Actual: 40000), Task 5 (Prediction: 450000, Actual: 330000), Task 9 (Prediction: 100000, Actual: 65000).*

*- Occurrences: 3*

*3. \*\*Incorrect Valuation Due to Attribute Misinterpretation\*\*:*

*- This category is a bit nuanced and could overlap with the first two, but it specifically refers to cases where the prediction might have gone wrong due to incorrect weighting or misunderstanding of the importance of certain attributes (e.g., overvaluing the impact of low mileage or undervaluing the impact of the car's age or condition based on the owner).*

*- However, based on the provided examples, all wrong predictions seem to stem from either overestimation or underestimation directly, without clear evidence of misinterpretation beyond these two reasons. Thus, this category might not be applicable without more detailed insight into how the LLM interpreted each attribute beyond the importance scores provided.*

*Given the examples and explanations, it appears that all incorrect predictions fall into the first two categories: underestimation and overestimation of market value. There isn't explicit evidence in the provided data to categorize any errors as due to incorrect valuation from attribute misinterpretation beyond these two reasons.*

🡪 Antwortqualität sehr viel besser und ausführlicher

🡪 Zuordnung zu Overestimations / Underestimations hier tendenziell korrekt

🡪 Vorschlag: Quantitative Analyse wie oft GPT-4 bei Erkennung der Overestimations / Underestimations Fehler macht (?)

🡪 Fehlerklassen sind trotzdem noch sehr oberflächlich und gehen nicht über Overestimations/Underestimations hinaus 🡪 Explanations werden bisher noch nicht genutzt

**Prompt mit mehr Kategorien – GPT-4-Turbo-Preview:**

🡪 Prompt siehe oben – nur dass nach 5 und nicht 3 Fehlerklassen gefragt wird

Antwort:

*Based on the provided examples of false predictions, we can categorize the wrong decisions into five fault categories:*

*1. \*\*Underestimation of Market Value\*\*:*

*- Task 2: Prediction significantly lower than actual price (180000 vs. 270000).*

*- Task 3: Prediction lower than actual price (150000 vs. 210000).*

*- Task 4: Prediction significantly lower than actual price (350000 vs. 500000).*

*- Task 6: Prediction lower than actual price (750000 vs. 1000000).*

*- Task 7: Prediction lower than actual price (650000 vs. 850000).*

*- Task 8: Prediction lower than actual price (200000 vs. 270000).*

*- Occurrences: 6*

*2. \*\*Overestimation of Market Value\*\*:*

*- Task 1: Prediction higher than actual price (60000 vs. 40000).*

*- Task 5: Prediction significantly higher than actual price (450000 vs. 330000).*

*- Task 9: Prediction higher than actual price (100000 vs. 65000).*

*- Occurrences: 3*

*3. \*\*Misjudgment of Vehicle Condition Impact\*\*:*

*- Task 5: Overestimated impact of high mileage on a diesel vehicle.*

*- Task 7: Failed to adequately account for the high mileage's impact on a relatively new car.*

*- Occurrences: 2*

*4. \*\*Misjudgment of Market Demand for Specific Models\*\*:*

*- Task 10: Prediction significantly lower than actual price for a model with potentially high demand (900000 vs. 1200000).*

*- Occurrences: 1*

*5. \*\*Incorrect Assessment of Seller Type Influence\*\*:*

*- Task 6: Underestimated the price premium associated with a Trustmark Dealer.*

*- Occurrences: 1*

*These categories highlight the challenges in accurately predicting used car prices based on various attributes, including the vehicle's condition, market demand for specific models, and the influence of the seller type.*

**Prompt mit Auschluss von Overestimation / Underestimation**

🡪 Prompt siehe oben – mit zusätzlichem Satz:

*Please go deeper with your analysis and don't use Overestimation or Underestimation as categories.*

Antwort:

*Analyzing the provided data, the wrong predictions can be grouped into three fault categories based on the nature of the errors made by the LLM. These categories are:*

*1. \*\*Misjudgment of Market Value\*\*:*

*- This category includes predictions where the LLM seems to have inaccurately assessed the market value of the car based on its attributes. This could be due to a lack of understanding of how certain attributes like the car's brand, model, and specific features (e.g., engine size, max power) influence its price in the market.*

*- Occurrences: Task 2 (Hyundai EON Era Plus), Task 3 (Chevrolet Optra Magnum 2.0 LT), Task 4 (Hyundai i20 1.4 CRDi Asta), Task 6 (Toyota Yaris V BSIV), Task 7 (Ford EcoSport 1.5 Petrol Titanium Plus BSIV), Task 8 (Honda City 2017-2020 GXi), Task 10 (Toyota Fortuner 3.0 Diesel).*

*2. \*\*Overemphasis on Mileage and Engine Capacity\*\*:*

*- In this category, the LLM's predictions seem to be overly influenced by the car's mileage and engine capacity, potentially at the expense of considering other crucial factors such as the car's age, condition, and market demand for the model.*

*- Occurrences: Task 5 (Toyota Innova 2.5 G (Diesel) 7 Seater BS IV).*

*3. \*\*Underestimation of Depreciation Impact\*\*:*

*- Predictions in this category appear to underestimate the impact of depreciation, particularly for older models or cars with higher mileage. The LLM fails to account adequately for the significant loss in value that cars undergo as they age and as their mileage increases.*

*- Occurrences: Task 1 (Maruti 800 Std), Task 9 (Maruti Wagon R LXI).*

*Summary of Occurrences:*

*- Misjudgment of Market Value: 7 occurrences*

*- Overemphasis on Mileage and Engine Capacity: 1 occurrence*

*- Underestimation of Depreciation Impact: 2 occurrences*

*This analysis suggests that the most common fault in the LLM's predictions is a misjudgment of the market value, indicating a potential area for improvement in understanding how various attributes contribute to a car's price in the market. Additionally, there's a need for better modeling of depreciation effects and a more balanced consideration of all relevant attributes beyond just mileage and engine capacity.*

🡪 Antwortqualität wird besser und Antworten gehen tiefer und beschreiben nicht mehr nur noch offentsichtliche Fehlerklassen

🡪 Funktioniert auch sehr gut mit bswp. 5 Fehlerklassen – Qualität wirkt hier sehr viel besser