Coding Challenge - Data Scientist

HUK COBURG

Moritz Wendhausen - 19.08.2024

Bilddaten

Hood: 0.36
Backdoor Left: 0.00

Hood: 0.00
Backdoor Left: 0.00

Hood: 0.00
Backdoor Left: 0.90

Hood: 0.00
Backdoor Left: 0.91

Hood: 0.00
Backdoor Left: 0.00

Hood: 0.90
Backdoor Left: 0.10

Hood: 0.00
Backdoor Left: 0.00

Hood: 0.00
Backdoor Left: 0.72

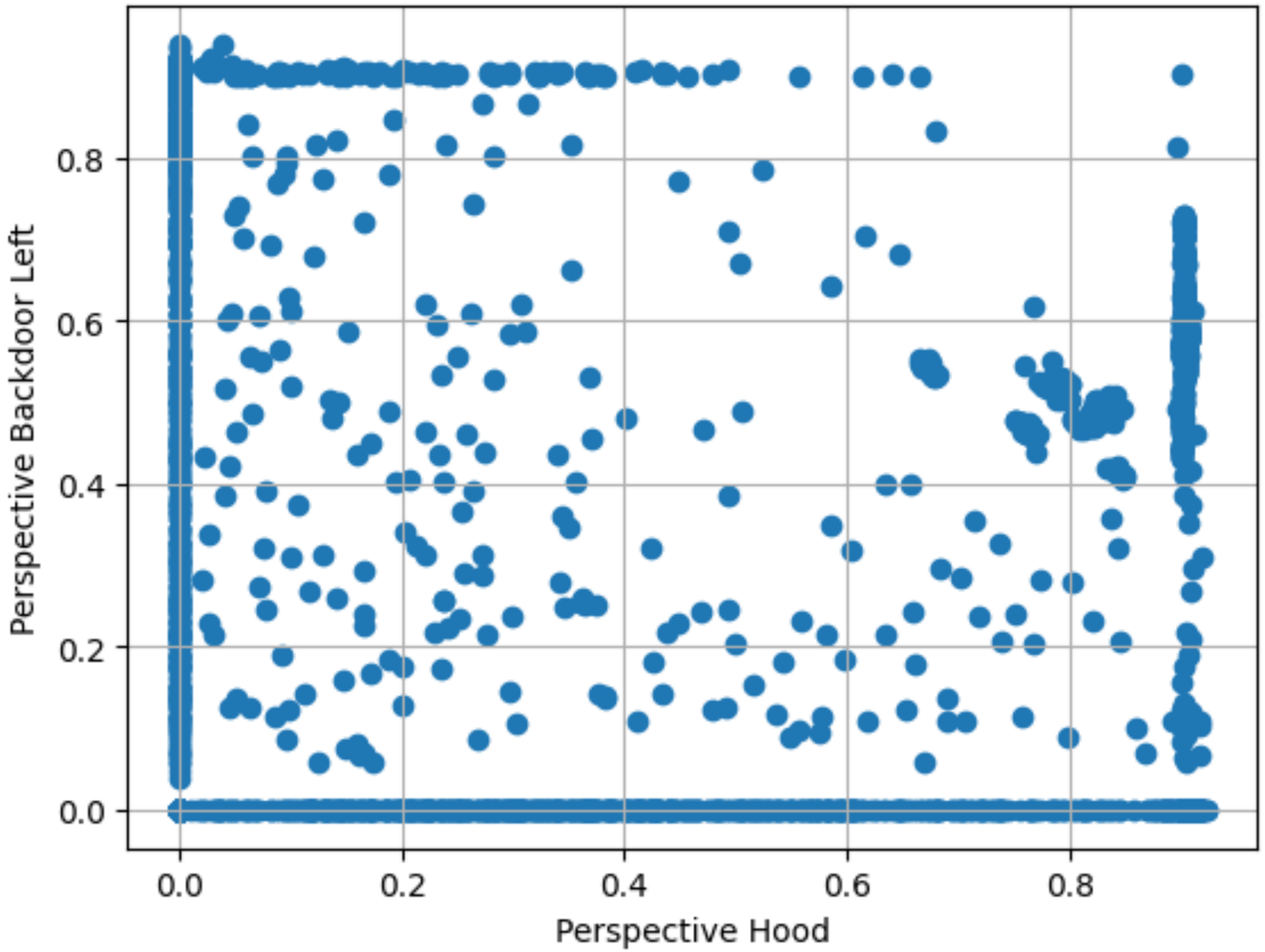
Hood: 0.91
Backdoor Left: 0.00

Hood: 0.00
Backdoor Left: 0.00

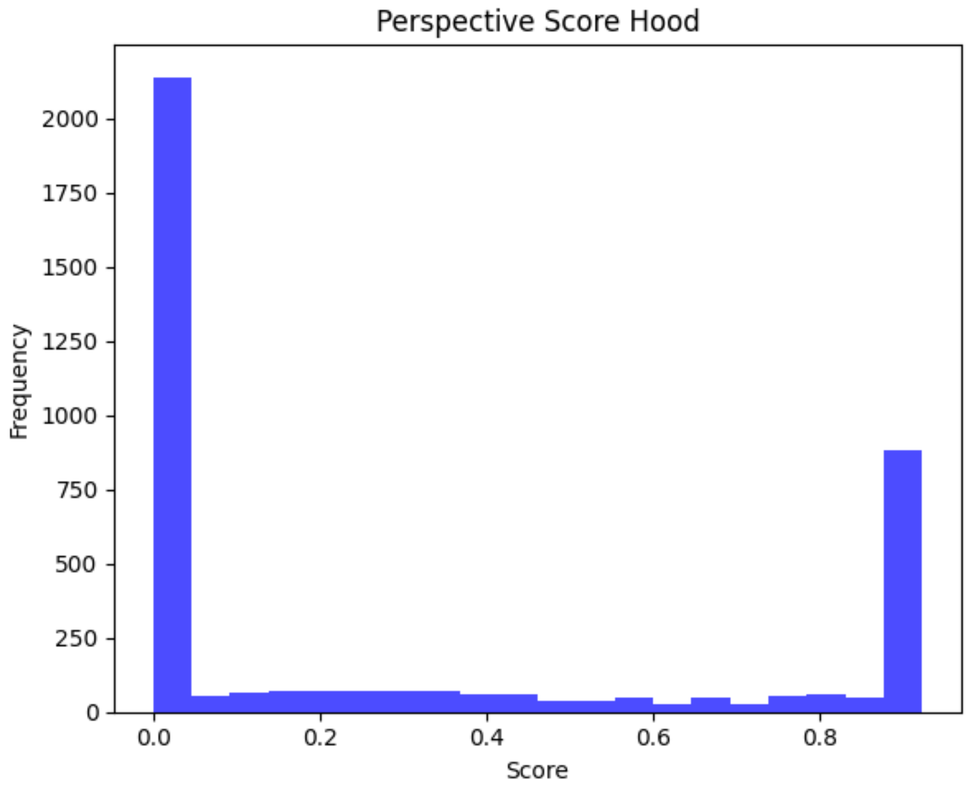
Scores - Descriptive Statistiken

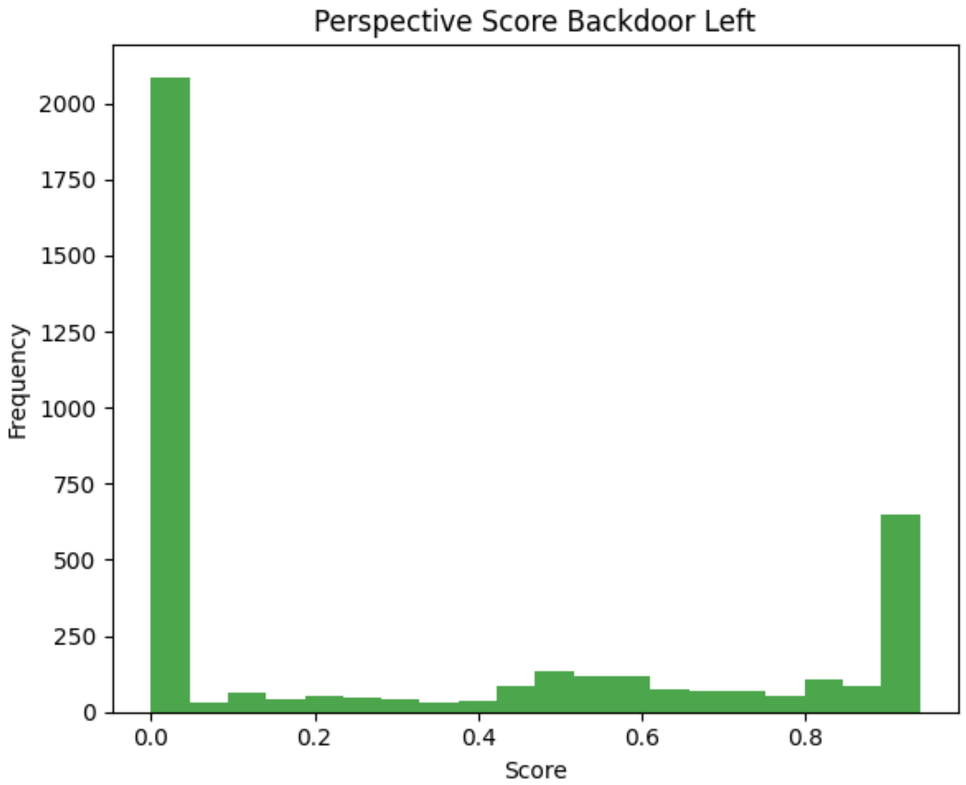
	Mean	Median	Standard Deviation	Minimum	Maximum	Korrelation
Perspective Hood	0.3030	0.0	0.3825	0.0	0.9224	-0.2120
Perspective Backdoor Left	0.3134	0.0	0.3723	0.0	0.9400	

Scores


Perspective Backdoor Left
0.0
0.2
0.4
0.6
0.8
Perspective Hood
0.0
0.2
0.4
0.6
0.8

Klassen Imbalance





Maßnahmen

- Mehr Daten sammeln
  - → Eventuell teuer
- Entfernen von Datenpunkten der Mehrheit
  - → Datensatz wird kleiner
- Vortrainierte Modelle benutzen

Maßnahmen

- Kontinuierliche Daten diskretisieren (Binning)
  - → Informationsverlust
- Trainings-Techniken: Loss functions, Under-/Oversampling
- Data Augmentation: Bilddaten drehen, spiegeln

Scores - Diskretisiert

	Weder Motorhaube noch Hintertür	Motorhaube	Linke Hintertür	Motorhaube und Hintertür
Anzahl an Datenpunkten	1692	1067	920	321

* Binning mit Threshold: > 0.5

Scores - Diskretisiert

	Motorhaube	Keine Motorhaube
Anzahl an Datenpunkten	2759	1241

	Linke Hintertür	Keine linke Hintertür
Anzahl an Datenpunkten	2612	1388

* Binning mit Threshold: > 0.5

Ist es sinnvoll, zwei unabhängige Modelle zu trainieren?

Jupyter Notebook

Problem Formulierung

- Regression
- Multi-label Klassifikation
- Binäre Klassifikation

Modell Architektur

- Vortrainiert → Transferlernen
- Convolutional Neural Network: ResNet, VGG16, YOLO..
- Vision Transformer
- Segment-Anything?
- **Abwägen von Kosten und Nutzen der Modellkomplexität**

Jupyter Notebook

Metriken

Regression

- Mean Squared Error → Vorhersage von kontinuierlichen Scores

Klassifikation

- Binary Cross Entropy → Vorhersage von binären Labeln (mit Wahrscheinlichkeiten)
  - Confusion Matrix, Precision, Recall, F1 Score, AUC
- Cross Entropy

Wie Weiter Optimieren?

- Mehr Daten
- Hyperparameter-Suche (Modell Größe, Learning Rate…)
- Regularisierung
  - L1, L2
  - Dropout
  - Normalization…
- Introspection und Feature Visualisierung

Weitere Anwendungsfälle

- Fahrzeugtyp Bestimmung: Hersteller, Modell, Baujahr
- Einschätzung der Schadenshöhe
- Betrugserkennung

Vielen Dank für eure Aufmerksamkeit