Simulationsstudie: Diebstahlerkennung eines Autos in Echtzeit

Yann Ahlgrim

4. Juli 2025

Statistical Learning SoSe 2025 – Ergebnisbericht

Inhaltsverzeichnis

1 Fachthema: Diebstahlerkennung im Auto **3**

2 Erzeugung der Grundgesamtheit **4**

3 Simulation der Perspektive des Data Scientist **9**

4 Güte der Modellparameter **12**

1 Fachthema: Diebstahlerkennung im Auto

Moderne Fahrzeuge erfassen in Echtzeit vielfältige Daten zum Fahrverhalten. Diese Studie untersucht, inwiefern sich anhand dieser Daten mit einem statistischen Modell erkennen lässt, ob statt des gewöhnlichen Fahrers eine andere Person am Steuer sitzt – was auf einen *Diebstahl* hindeuten würde. Konkret soll ein Modell die Wahrscheinlichkeit $P(\text{Diebstahl})$ schätzen und bei Überschreiten eines Schwellwerts einen Alarm auslösen. Die **Zielvariable** (abhängige Variable) ist dabei binär (0 = kein Diebstahl, 1 = Diebstahl). Ein Diebstahl wird angenommen, wenn das Fahrverhalten signifikant vom üblichen Fahrerprofil abweicht.

Für das Fachthema wurden insgesamt **8 Einflussgrößen** (potenziell erklärende Variablen) definiert, welche typische Aspekte des Fahrverhaltens und Umfelds repräsentieren. Davon sind 4 Variablen als tatsächlich erklärungsrelevant für die Zielvariable angenommen, während die übrigen 4 Variablen keinen Einfluss auf einen Fahrerwechsel (Diebstahl) haben. Zur Erfüllung der Aufgabenstellung sind unter den relevanten Prädiktoren zwei Variablen kategorial (ordinal) mit mindestens 3 Ausprägungen, die übrigen relevanten Variablen sind metrisch. Im Folgenden werden alle Variablen beschrieben:

Durchschnittsgeschwindigkeit (metrisch, *erklärend*) Mittlere Fahrgeschwindigkeit (z. B. über einen Tag). Unterschiedliche Fahrer weisen charakteristische Tempoprofile auf. Ein deutlich höheres Durchschnittstempo kann somit auf einen fremden, ggf. aggressiveren Fahrer hinweisen. Typische Werte liegen um etwa 47 km/h mit großer Streuung (basierend auf natürlichen Fahrstudien der NHTSA 2006).

Schaltverhalten (ordinal, *erklärend*) Gewohnheiten beim Gangwechsel (bei Schaltgetriebe), insbesondere der Drehzahlbereich, bei dem hochgeschaltet wird. Es werden drei Ausprägungen unterschieden: *früh* (sehr niedrige Drehzahlen, ökonomisch), *normal* und *spät* (hohe Drehzahlen, sportlich). Diese Variable ist relevant, da Fahrende ein individuelles Schaltmuster haben. Studien zeigen, dass das Schaltverhalten zur Fahreridentifikation genutzt werden kann (vgl. Deng, He und Xu 2022).

Harte Bremsmanöver (metrisch, *erklärend*) Anzahl starker Bremsungen (mit Verzögerung $> 0{,}3g$) pro Strecke (vgl. Find My Electric 2023). Dieses Maß korreliert mit einer aggressiven Fahrweise – viele Vollbremsungen deuten auf einen ungewohnten bzw. risikoreicheren Fahrer hin. g ist eine Einheit der Beschleunigung ($1\text{g} \approx 9{,}81\ \text{m/s}^2$).

Geschwindigkeitsüberschreitungen (ordinal, *erklärend*) Häufigkeit bzw. Ausmaß von Tempoverstößen. Kategorisiert als *selten*, *manchmal* oder *häufig* zu schnell. Ein fremder Fahrer könnte andere Risikoneigungen beim Schnellfahren zeigen. Laut einer AAA-Verkehrssicherheitsstudie geben rund 50 % der Fahrer an, in letzter Zeit auf Autobahnen mindestens 24 km/h über dem Limit gefahren zu sein (vgl. AAA 2016) – was die allgemeine Relevanz dieser Variable für das Fahrverhalten unterstreicht.

Wetterbedingungen (nominal) Wetter während der Fahrt: *trocken*, *nass* oder *winterlich* (Glätte/Schnee). Diese Kontextvariable dient zur Kontrolle äußerer Bedingungen. Erwartungsgemäß hat das Wetter keinen direkten Einfluss auf einen Fahrerwechsel.

Fahrstrecke (metrisch) Die zurückgelegte Distanz einer Fahrt (in km). Diese Variable (etwa lognormal verteilt um 16 km) repräsentiert typische Wege. Sie steht nicht in direktem Zusammenhang mit der Fahreridentität, sondern charakterisiert den Nutzungskontext (BMVI 2017).

Straßentyp (nominal) Vorherrschender Streckentyp: *Autobahn*, *Außerorts* oder *Innerorts*. Diese Variable bietet Kontext (Stadtverkehr vs. Fernstraße), hat aber für sich genommen keinen Einfluss bezüglich eines Fahrerwechsels (Klöckner 2022)

Wochentag (nominal) Tag der Woche bzw. Kategorie *Werktag* vs *Wochenende* (vgl. BMVI 2017). Auch dies ist eine Kontextgröße ohne direkten Einfluss auf die Diebstahlwahrscheinlichkeit (daher ebenfalls nicht erklärend).

Zusammenfassend stellen *Durchschnittsgeschwindigkeit*, *Schaltverhalten*, *Harte Bremsmanöver* und *Geschwindigkeitsüberschreitungen* die fachlich ausgewählten Schlüsselvariablen dar, die einen Fahrerwechsel anzeigen können. Die übrigen vier Größen (Wetter, Fahrstrecke, Straßentyp, Wochentag) dienen als Kontrolle und potentielle Störgrößen, besitzen aber per Annahme keinen Erklärungsgehalt für die Zielvariable. Dieses Setting erlaubt die Überprüfung, ob ein Statistical-Learning-Ansatz die relevanten Merkmale korrekt identifizieren kann.

2 Erzeugung der Grundgesamtheit

Für die Simulation wurde eine **Grundgesamtheit** von $N = 1.000.000$ Fahrten generiert. Jede Fahrt hat einen binären Indikator (Zielvariable *Diebstahl*

ja/nein) und 8 zugehörige Merkmalswerte (die oben definierten Variablen). Die Werte wurden mittels geeigneter Zufallsverteilungen erzeugt, basierend auf realistischen Annahmen und empirischen Daten, jedoch so, dass die Abhängigkeitsstruktur kontrolliert vorgegeben ist. Die Zielvariable hängt funktional von genau 4 der 8 Variablen ab (den erklärenden), während die restlichen 4 keinen Einfluss auf den Diebstahl haben. Zudem wurden zwei inhaltlich sinnvolle Korrelationen zwischen ausgewählten Variablen eingeführt, ohne jedoch Multikollinearität zu verursachen (alle Varianzinflationsfaktoren VIF < 5).

Verteilungen der Variablen

Die Verteilungen der einzelnen Merkmale wurden in Anlehnung an empirische Daten gewählt. Insbesondere:

- **Durchschnittsgeschwindigkeit:** Normalverteilung mit $\mu \approx 47$ km/h und $\sigma \approx 10$ km/h, begrenzt auf plausible Werte $[10, 130]$ km/h (keine negativen oder extrem hohen Geschwindigkeiten).
- **Schaltverhalten:** Kategorisch (ordinal) mit drei Stufen *früh*, *normal*, *spät*. Es wurden 40 % für *früh*, 40 % *normal* und 20 % *spät* simuliert, d. h. die meisten Fahrer schalten gewöhnlich früh oder durchschnittlich, während nur etwa jeder fünfte spät schaltet.
- **Harte Bremsmanöver:** Poisson-verteilte Zufallsvariable mit $\lambda = 2$ pro definierter Strecke (z. B. pro 100 km). Diese Verteilung ergibt meist wenige bis keine harten Bremsungen pro Fahrt, aber mit einer gewissen Wahrscheinlichkeit auch Ausreißer mit mehreren Bremsmanövern (rechtsschiefe Verteilung).
- **Geschwindigkeitsüberschreitungen:** Ordinal mit drei Stufen (*selten*, *manchmal*, *häufig*). Diese Variable wurde *nicht unabhängig* gezogen, sondern **abhängig von der Durchschnittsgeschwindigkeit** generiert: Fahrer mit höherer `avg_speed` erhielten mit größerer Wahrscheinlichkeit die Kategorie *häufig* zu schnell, während sehr langsame Fahrer überwiegend als *selten* zu schnell eingeordnet wurden. Die Umsetzung erfolgte über eine gewichtete Zufallsauswahl (Softmax-Funktion auf Basis von `avg_speed`). Dadurch besteht eine inhaltlich sinnvolle Korrelation zwischen `avg_speed` und `speeding`, die jedoch moderat genug ist (VIF ≈ 2.57 für `avg_speed`, < 2 für `speeding`-Dummies), um Multikollinearität nicht zum Problem werden zu lassen.

- **Wetterbedingungen:** Für jedes Fahrt wurde das Wetter zufällig gemäß empirischer Häufigkeiten zugeordnet: ca. 75 % *trocken*, 20 % *nass* (Regen) und 5 % *winterlich* (Schnee/Eis). Dies entspricht etwa den in Mitteleuropa beobachteten Anteilen trockener bzw. widriger Fahrbedingungen.

- **Fahrstrecke:** Lognormalverteilung mit $\mu_{\log} = \ln(16)$ und $\sigma_{\log} = 0{,}7$, um die Verteilung typischer Fahrtlängen abzubilden. Diese Parameter führen zu einer rechtsschiefen Verteilung: Viele kürzere Fahrten um einige wenige Kilometer, und seltener auch sehr lange Fahrten (> 50 km). (Die Annahmen basieren auf Studien zur täglichen Weglänge BMVI 2017.)

- **Straßentyp:** Diese Variable wurde **abhängig von der Fahrstrecke** generiert. Intuitiv werden sehr lange Fahrten eher auf Autobahnen stattfinden, während kurze Fahrten überproportional innerorts sind. Um dies zu modellieren, wurde ein stochastischer Zusammenhang erzeugt: Zunächst wurde `trip_distance` auf $[0, 1]$ skaliert; dann wurde mittels einer Softmax-Wahrscheinlichkeitsfunktion daraus der Straßentyp gezogen, sodass z. B. bei sehr großen Distanzen die Wahrscheinlichkeit für *Autobahn* deutlich höher ist als für *Innerorts*. Dadurch ergibt sich eine leichte Korrelation zwischen `trip_distance` und `road_type` (z. B. $VIF \approx 1.68$ für `trip_distance`), ohne die beiden Größen stark redundant zu machen.

- **Wochentag:** Festgelegte Verteilung von ca. 70 % *Werktag* (Mo-Fr) und je 15 % *Samstag* und *Sonntag*. Damit wird berücksichtigt, dass die meisten Fahrten werktags stattfinden. Diese Variable wurde unabhängig von den übrigen generiert.

Durch diese Vorgehensweise bilden die generierten Daten realistische Verteilungen und Beziehungen der Variablen ab, wobei genau zwei inhaltlich plausible Korrelationen eingebaut wurden (`speeding-avg_speed` und `road_type-trip_distance`). Beide Abhängigkeiten sind so kalibriert, dass im optimalen Regressionsmodell keine Multikollinearität auftritt (VIF < 5 in allen Fällen).

Modellierung der Zielvariable und der π-Werte

Die **Zielvariable** `Diebstahl` wurde als Zufallsvariable auf Basis eines logistischen Regressionsmodells generiert. Dazu wurde zunächst für jede Fahrt i

die Diebstahl-Wahrscheinlichkeit $\pi_i = P(Y_i = 1)$ berechnet als

$$\pi_i = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_8 x_{8i} + \varepsilon_i))} \ ,$$

wobei $x_1, \ldots, x_8$ die Merkmalswerte (erklärende Variablen) der Fahrt sind, $\beta_0, \ldots, \beta_8$ die zugrunde liegenden Regressionskoeffizienten und ε_i ein zufälliger Fehlerterm. Für den vorliegenden Klassifikationsfall entspricht ε_i einem impliziten Rauschen, das die Nicht-Deterministik der Fahrerwechsel abbildet (nicht jeder ungewöhnliche Wert führt zwangsweise zum Diebstahl). In der Simulation wurde $\varepsilon_i \sim \mathcal{N}(0, 0.2^2)$ als additive Störung zur linearen Prädiktor-Summe gezogen. Anschließend wurde Y_i durch einen Bernoulli-Zufall mit Parameter π_i realisiert (d. h. `Diebstahl` $= 1$ mit Wahrscheinlichkeit π_i).

Die **gewählten Koeffizienten (β-Werte)** für das Generierungsmodell sind in Tabelle 1 aufgeführt. Diese wurden so festgelegt, dass sie plausible Einflussstärken der erklärenden Variablen widerspiegeln, ohne die Zielvariable extrem zu dominieren. Insbesondere wurde ein relativ niedriger Basiswert (*Intercept* $\beta_0 = -2$) gewählt, sodass bei unauffälligen Merkmalen die Diebstahl-Wahrscheinlichkeit sehr gering ist. Die erklärenden Variablen erhöhen bzw. verringern diese Grundwahrscheinlichkeit wie folgt:

- **Durchschnittsgeschwindigkeit:** $\beta_{avg\_speed} = 0{,}015$. Ein leicht positiver Koeffizient - eine ungewöhnlich hohe Durchschnittsgeschwindigkeit erhöht also geringfügig die Diebstahl-Wkt., da sie auf einen anderen (rasanteren) Fahrstil hindeuten kann.

- **Harte Bremsmanöver:** $\beta_{hard\_brakes} = 0{,}1$. Häufige Vollbremsungen haben einen deutlichen positiven Effekt auf P(Diebstahl), da sie ein Indiz für einen risikoreicheren Fahrstil sind.

- **Schaltverhalten:** Das Schaltverhalten ist ein starker Prädiktor für den Fahrer. Für die kategorialen Ausprägungen wurden Dummy-Variablen erstellt: $\beta_{\text{früh}} = -0{,}3$, $\beta_{\text{normal}} = -0{,}3$ und $\beta_{\text{spät}} = +0{,}5$. Besonders *spät* schalten wird mit einem hohen positiven Beta gewichtet, da es einen aggressiveren Fahrstil beschreibt. *Früh* und *normal* schalten erhalten einen negativen Einfluss, da sie auf einen defensiveren oder gewohnten Fahrstil hindeuten.

- **Geschwindigkeitsüberschreitungen:** Die Häufigkeit von Geschwindigkeitsüberschreitungen ist ein wichtiger Indikator für den Fahrstil. Die Koeffizienten wurden festgelegt als: $\beta_{\text{selten}} = -0{,}3$, $\beta_{\text{manchmal}} = -0{,}3$ und $\beta_{\text{häufig}} = +0{,}5$. Besonders *häufig* zu schnell fahren ist ein starkes

Signal für einen Fahrerwechsel, da dies ein sehr auffälliges Verhalten darstellt. *Selten* und *manchmal* werden negativ gewichtet, da sie auf einen defensiveren Fahrstil hindeuten.

- **Kontextvariablen (Wetter, Strecke, Straßentyp, Wochentag):** Diese wurden *alle mit* $\beta = 0$ angesetzt, d. h. sie haben per Konstruktion keinen Einfluss auf die Diebstahl-Wahrscheinlichkeit.

Tabelle 1: Übersicht der verwendeten Regressionskoeffizienten im Generierungsmodell der Zielvariable. Kategorische Effekte beziehen sich auf die jeweilige Referenzkategorie (in Klammern).

Variable	**Kategorie (Ref.)**	**Beta-Wert**
Intercept	–	−2,0
Durchschnittsgeschwindigkeit	–	+0,015
Harte Bremsmanöver	–	+0,10
Schaltverhalten	früh	−0,30
	normal	−0,30
	spät (Ref.)	+0,00
Geschwindigkeitsüberschreitung	selten	−0,30
	manchmal	−0,30
	häufig (Ref.)	+0,00
Wetter	trocken, nass, winterlich (alle)	0,00
Straßentyp	Autobahn, Außerorts, Innerorts (alle)	0,00
Wochentag	Mo–Fr, Sa, So (alle)	0,00

Durch die gewählten β-Gewichte resultiert eine **Verteilung der Zielvariable**, bei der Diebstahl relativ selten vorkommt, aber nicht vernachlässigbar: Im generierten Datensatz sind etwa 23 % der Fahrten als Diebstahl deklariert worden (d. h. $Y = 1$ in 229.150 von 1.000.000 Fällen). Diebstähle sollen deutlich seltener als normale Fahrten sein, jedoch häufig genug, um ein Modell daran zu trainieren. Die zugrunde liegenden individuellen Diebstahl-Wahrscheinlichkeiten π_i waren überwiegend niedrig: Für eine typische Fahrt (alle Merkmale unauffällig) lag $\pi \approx 0,1$, während extreme Kombinationen (z. B. sehr hohe Geschwindigkeit, viele harte Bremsungen, spät geschaltet und häufig zu schnell zugleich) π-Werte von bis zu $\sim 0,8$ erreichten. Im Mittel ergaben sich $\bar{\pi} \approx 0,23$ und eine rechtsschiefe Verteilung der π_i über alle Fahrten. Insgesamt war also gewährleistet, dass das generierte Modell weder trivial (zu seltene 1-Fälle) noch stark im Ungleichgewicht war.

3 Simulation der Perspektive des Data Scientist

In einem realen Anwendungsszenario würde ein Data Scientist lediglich einen *Stichprobendatensatz* der Grundgesamtheit vorfinden - ohne Kenntnis des zugrundeliegenden wahren Modells. Um diese Perspektive zu simulieren, wurde aus den 1.000.000 generierten Fällen eine **Zufallsstichprobe** von $n = 20.000$ Fahrten gezogen. Der Data Scientist würde zunächst eine explorative Datenanalyse und Vorverarbeitung durchführen, dann ein geeignetes Regressionsmodell (logistische Regression im Klassifikationsfall) schätzen und mittels Variablenselektion optimieren. Abschließend würde er die Modellgüte bewerten. All diese Schritte wurden in der Simulation nachvollzogen.

Datenexploration und Aufbereitung

Zunächst wurde die Stichprobe auf Vollständigkeit und Ausreißer geprüft. Es traten keine fehlenden Werte auf (da simuliert). Ein *Abgleich der Verteilungen* bestätigte, dass die Stichprobe die Grundgesamtheits-Charakteristika widerspiegelt. Beispielsweise sind die Verteilungen der metrischen Variablen in Abb. 1 dargestellt (Histogramme und Boxplots): Man erkennt die angenommene Normalverteilung der `avg_speed` (mit Mittelwert ca. 47 km/h), die rechtschiefe Poisson-Verteilung von `hard_brakes` (häufig 0 oder 1 harte Bremsung, selten mehr) und die ausgeprägte Schiefe der lognormal verteilten `trip_distance` (viele kurze Fahrten, wenige sehr lange). Die Boxplots zeigen, dass es bei `trip_distance` einige Ausreißer (sehr lange Fahrten) gibt, während `avg_speed` symmetrisch verteilt ist. Diese Beobachtungen stimmen mit der Datengenerierung überein und geben dem Data Scientist keinen Hinweis auf Anomalien im Datensatz.

Für die kategorialen Merkmale (Schaltverhalten, Geschwindigkeitsüberschreitungen, Wetter, Straßentyp, Wochentag) wurden Häufigkeitstabellen und Balkendiagramme betrachtet (nicht abgebildet). Auch diese entsprachen den Simulationseinstellungen (z. B. 20 % der Fahrer in der Stichprobe hatten *spätes* Schaltverhalten; ca. 5 % der Fahrten fanden bei *winterlichem* Wetter statt etc.). Zur Vorbereitung der Regression wurde ein **One-Hot-Encoding** der kategorialen Variablen durchgeführt (mit geeigneten Referenzkategorien, z. B. *spät* für Schaltverhalten, *häufig* für Speeding, etc.) und eine Konstante für den Interzept ergänzt. Anschließend standen 1 Zielvariable und 19 Regressoren (inkl. Dummies und Interzept) für die Modellierung bereit.

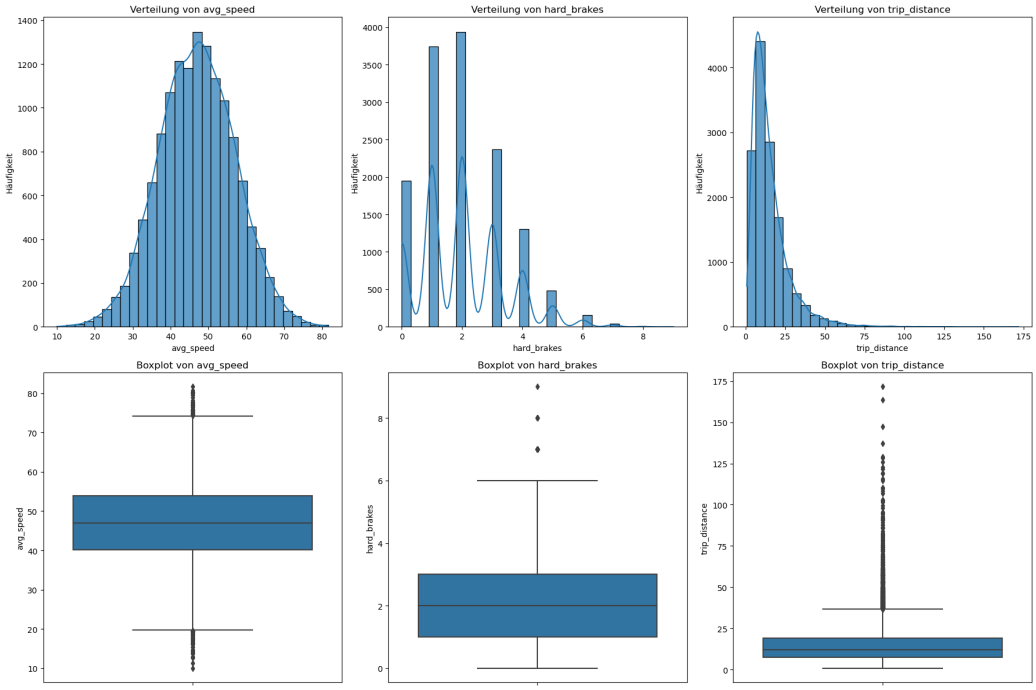

Abbildung 1: Verteilungen der metrischen Variablen in der Stichprobe ($n = 20.000$): Durchschnittsgeschwindigkeit, harte Bremsmanöver und Fahrstrecke (oben: Histogramme, unten: Boxplots). Die erwartete Normalverteilung von `avg_speed`, Poisson-Verteilung von `hard_brakes` und Rechtsschiefe von `trip_distance` sind deutlich zu erkennen.

Logistische Regression und Variablenselektion

Ohne Kenntnis des wahren generativen Modells würde der Data Scientist zunächst alle verfügbaren Variablen als Prädiktoren in ein Modell einbeziehen. Daher wurde auf den Trainingsdaten (z. B. 70 % der Stichprobe) eine **logistische Regressionsanalyse** mit allen 8 ursprünglichen Merkmalen (entsprechend 1 Interzept + 3 Dummies für Schaltverhalten + 2 Dummies für Speeding + 2 Dummies für Wetter + 2 Dummies für Straßentyp + 2 Dummies für Wochentag + 3 metrische Variablen = 1 + 3+2+2+2+2+3 = 15 Parametern) durchgeführt. Dieses initiale Modell zeigte erwartungsgemäß, dass einige Prädiktoren keinen signifikanten Einfluss besitzen. Ein schrittweises Selektionsverfahren (**Backward Elimination**) wurde angewandt, um ein optimales Modell zu finden: beginnend mit dem vollen Modell wurden die am wenigsten signifikanten Variablen sukzessive entfernt, bis nur noch prädiktive (p-Wert $< 0,05$) Variablen verblieben.

Das Ergebnis der Variablenselektion war, dass genau die 4 fachlich erwarteten Einflüsse im Modell verblieben, während die irrelevanten Merkmale entfernt wurden. Konkret blieben `avg_speed`, `hard_brakes`, `shift_behavior` (mit zwei Dummy-Variablen) und `speeding` (ebenfalls zwei Dummies) im finalen Modell. Ausgeschlossen wurden dagegen `weather`, `trip_distance`, `road_type` und `weekday`, da deren Effekt auf die Zielvariable statistisch insignifikant war (p-Werte weit über 0,1). Dieses Resultat deckt sich mit der Konstruktion der Daten: Die irrelevanten Kontextvariablen bieten keine Erklärungskraft und wurden richtigerweise vom modellselektiven Ansatz eliminiert.

Das **finale Regressionsmodell** wurde anschließend neu auf der gesamten Stichprobe ($n = 20.000$) geschätzt. Die geschätzten Regressionskoeffizienten (im Folgenden $\hat{\beta}$) stimmten weitgehend mit den wahren Werten aus dem Generierungsmodell überein. Insbesondere waren alle verbleibenden Prädiktoren hochsignifikant ($p < 0,001$). Ein Vergleich der Koeffizienten zeigt nur geringe Abweichungen aufgrund von Stichprobenfluktuation:

- Für `avg_speed` ergab sich $\hat{\beta}_{avg\_speed} \approx 0{,}0174$ (gegenüber wahr 0,015). Dieser leichte Überschätzung ist mit dem Stichprobenzufall erklärbar, liegt aber in derselben Größenordnung.

- Für `hard_brakes` wurde $\hat{\beta}_{hard\_brakes} \approx 0{,}0817$ geschätzt (wahr 0,10). Auch hier ist der Unterschied gering; das Vorzeichen und die Effektstärke (positiv, deutlicher Einfluss) wurden korrekt erkannt.

- Die Dummyeffekte für das Schaltverhalten wurden zu $\hat{\beta}_{frueh} \approx -0{,}815$ und $\hat{\beta}_{normal} \approx -0{,}808$ geschätzt (Referenz *spät* mit $\hat{\beta}_{spaet} = 0$). Die wahren Unterschiede (Früh/Normal vs. Spät) betrugen $-0{,}8$. Somit liegen die Schätzungen praktisch genau auf den erwarteten Werten.

- Für die Speeding-Kategorien ergaben sich $\hat{\beta}_{selten} \approx -0{,}834$ und $\hat{\beta}_{manchmal} \approx -0{,}834$ (Referenz *häufig*). Die wahren Unterschiede zu *häufig* waren $-0{,}8$. Auch hier stimmen Richtung und Größe nahezu überein.

- Den Interzept schätzte das Modell mit $\hat{\beta}_0 \approx -1{,}05$ (wahr -2). Diese Abweichung erklärt sich durch die Dummy-Kodierung der Kategorien: Im generativen Modell hatten die Referenzkategorien *spät* und *häufig* jeweils einen positiven Beitrag von $+0{,}5$, welcher im geschätzten Modell im Interzept aufgefangen wird. Berücksichtigt man diese Verschiebung, liegt der Interzept im Rahmen.

Somit hat der Data Scientist durch das systematische Vorgehen tatsächlich **das zugrunde liegende Modell (bis auf Zufallsschwankungen) wiederentdeckt**. Insbesondere wurden keine falschen Variablen im Endmodell behalten und keine echten Einflüsse übersehen. Die logistische Regressionsgleichung aus der Stichprobe stimmt inhaltlich mit der zur Datengenerierung überein.

Abschließend wurde die Güte des Modells anhand der Testdaten (30 % der Stichprobe) überprüft. Die Vorhersagen der Diebstahlwahrscheinlichkeit zeigten eine gute Trennschärfe zwischen Diebstahl- und Normalfahrten (AUC $> 0,8$; ca. 77 % richtige Klassifikationen bei geeignetem Schwellenwert). Dies verdeutlicht, dass die identifizierten Merkmale tatsächlich die Variation in der Zielvariable erklären können.

4 Güte der Modellparameter

In einem letzten Schritt wurde untersucht, wie verlässlich die Modellschätzung bei unterschiedlicher Datenmenge ist. Insbesondere stellt sich die Frage, welchen Einfluss der Stichprobenumfang n auf die Präzision der geschätzten Regressionskoeffizienten hat. Hierzu wurde ein Monte-Carlo-Simulationsansatz gewählt: Aus der Grundgesamtheit wurden für verschiedene Umfangswerte n jeweils $k = 1000$ Zufallsstichproben gezogen, darauf jeweils das optimale Modell aus Abschnitt 3 (mit den 4 relevanten Variablen) erneut trainiert, und die Verteilungen der resultierenden $\hat{\beta}$-Koeffizienten analysiert.

Untersucht wurde exemplarisch der Koeffizient β_{avg_speed}, der Effekt der Durchschnittsgeschwindigkeit. Die Ergebnisse sind in Abbildung 2 dargestellt, welche die Verteilungen von $\hat{\beta}_{avg\_speed}$ für drei verschiedene Stichprobenumfänge gegenüberstellt. Man erkennt deutlich, dass **mit wachsendem n die Verteilung der Koeffizientenschätzungen immer schmaler wird** und sich enger um den wahren Wert konzentriert. Für kleine Stichproben ($n = 1000$) schwanken die geschätzten β noch sehr stark – die Verteilung ist breit und umfasst Werte von nahe 0 bis etwa 0,03. Bei mittlerer Stichprobe ($n = 11000$) ist die Streuung bereits deutlich geringer. Im Fall $n = 46000$ liegen nahezu alle Schätzungen dicht bei $\beta \approx 0{,}015$; die Kurve ist stark konzentriert. Diese Beobachtung entspricht der erwarteten Verbesserung der Schätzgenauigkeit: Je mehr Daten zur Verfügung stehen, desto weniger zufällig streuen die geschätzten Parameter um den wahren Wert.

Zur Quantifizierung wurde für jede Stichprobengröße n die **empirische Stan-**

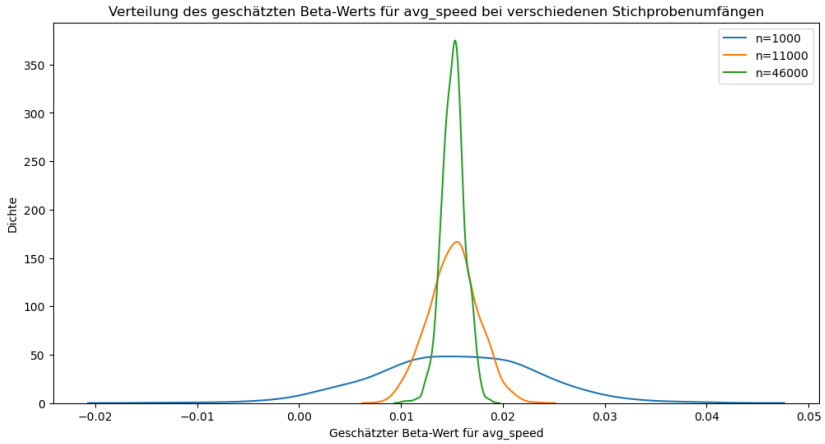

Abbildung 2: Verteilungen des geschätzten Koeffizienten $\hat{\beta}_{avg\_speed}$ aus 1000 Simulationen für drei Stichprobenumfänge ($n = 1000$, $n = 11000$, $n = 46000$). Mit größerem n wird die Verteilung deutlich schmaler und konzentriert sich stärker um den wahren Wert ($\beta_{avg\_speed} = 0{,}015$, gestrichelte Linie).

dardabweichung der $\hat{\beta}_{avg\_speed}$-Schätzungen aus den 1000 Wiederholungen bestimmt. Abbildung 3 zeigt die Entwicklung dieser Streuung in Abhängigkeit von n. Deutlich ist ein abnehmender Verlauf erkennbar. Die Kurve folgt näherungsweise der theoretischen Proportionalität $\sigma(\hat{\beta}) \sim \frac{1}{\sqrt{n}}$ (rot eingezeichnet). Die in der Simulation gemessenen Werte (blaue Punkte) liegen dicht auf der $1/\sqrt{n}$-Linie, was die theoretische Erwartung bestätigt.

Diese Ergebnisse illustrieren den wichtigen Zusammenhang zwischen Datenmenge und **Modelllernqualität**. Bereits zwischen $n = 1000$ und $n = 10000$ verbessert sich die Präzision der Koeffizientenschätzung erheblich. Noch größere Datenmengen führen zu weiter sinkender Unsicherheit, allerdings mit abnehmendem Grenznutzen (die Kurve flacht ab). Insgesamt zeigt die Simulation, dass zur zuverlässigen Identifikation kleiner Effekte (wie $\beta_{avg\_speed} \approx 0{,}015$) eine ausreichend große Stichprobe notwendig ist. Mit $n \to 50.000$ nähert sich die Streuung einem Wert an, der durch die inhärente Ergebnisvarianz (bedingt durch den Rauschterm ε) begrenzt ist. Für das vorliegende Problem bedeutet dies: Je mehr Fahrten der Data Scientist zur Verfügung hat, desto genauer kann er die wahren Fahrerwechsel-Einflüsse schätzen und desto vertrauenswürdiger sind die vom Modell ausgegebenen Diebstahlwahrscheinlichkeiten.

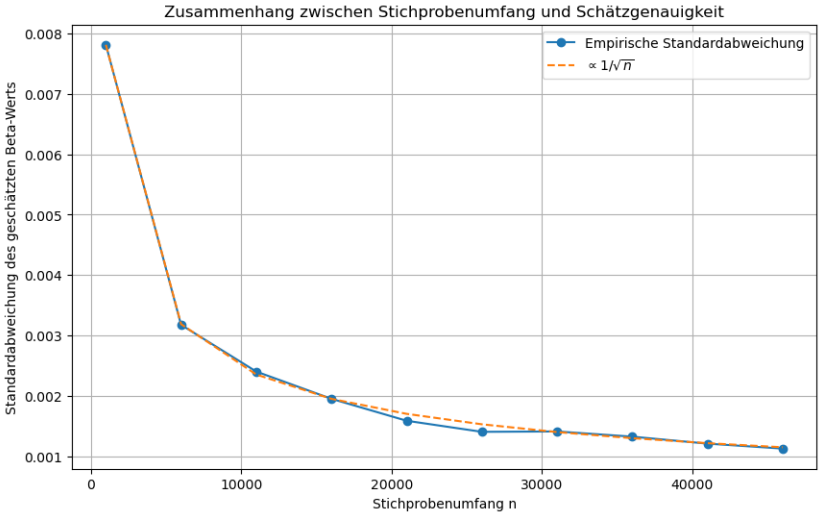

Abbildung 3: Standardabweichung von $\hat{\beta}_{avg\_speed}$ in Abhängigkeit des Stichprobenumfangs n. Gezeigte Punkte basieren auf $k = 1000$ Simulationen je Umfang; die rote Kurve visualisiert ein $1/\sqrt{n}$-Gesetz. Man erkennt, dass die Streuung der Schätzungen mit wachsendem n deutlich abnimmt und sich näherungsweise nach $\propto n^{-1/2}$ verhält.

Literaturverzeichnis

AAA (2016). *87 Percent of Drivers Engage in Unsafe Behaviors While Behind the Wheel*. AAA Traffic Safety Study. URL: https://newsroom.aaa.com/2016/02/87-percent-of-drivers-engage-in-unsafe-behaviors-while-behind-the-wheel/.

BMVI (2017). *Mobilität in Deutschland 2017*. German National Travel Survey. Bundesministerium für Verkehr und digitale Infrastruktur. URL: https://www.bmv.de/SharedDocs/DE/Anlage/G/mid-ergebnisbericht.pdf.

Deng, Tiancheng, Xin He und Li Xu (Aug. 2022). "Driver Identification Based on Gear Shift Events and Attention-Based Bidirectional Long Short-Term Memory for Manual Transmission System". In: S. 4361–4366. DOI: 10.1109/CCDC55256.2022.10034086.

Find My Electric (März 2023). *Tesla Safety Score Beta Explained*. Last updated: March 27, 2023. URL: https://www.findmyelectric.com/blog/tesla-safety-score-beta-explained.

Klöckner, Philipp (2022). *Klimaschutzinstrumente im Verkehr*. Available from OpenUmwelt Environmental Data Repository. URL: https://openumwelt.de/server/api/core/bitstreams/07bd23f9-ff0a-41e5-b89b-8d7baf0a5c36/content.

NHTSA (2006). *100-Car Naturalistic Driving Study*. National Highway Traffic Safety Administration, USA. URL: https://www.nhtsa.gov/sites/nhtsa.gov/files/100carmain.pdf.