Datenbasierte Startup Auswahl

Startup Growth & Funding Trends

Engel, Nolepa, Schneider, Trölenberg | 19.11.25

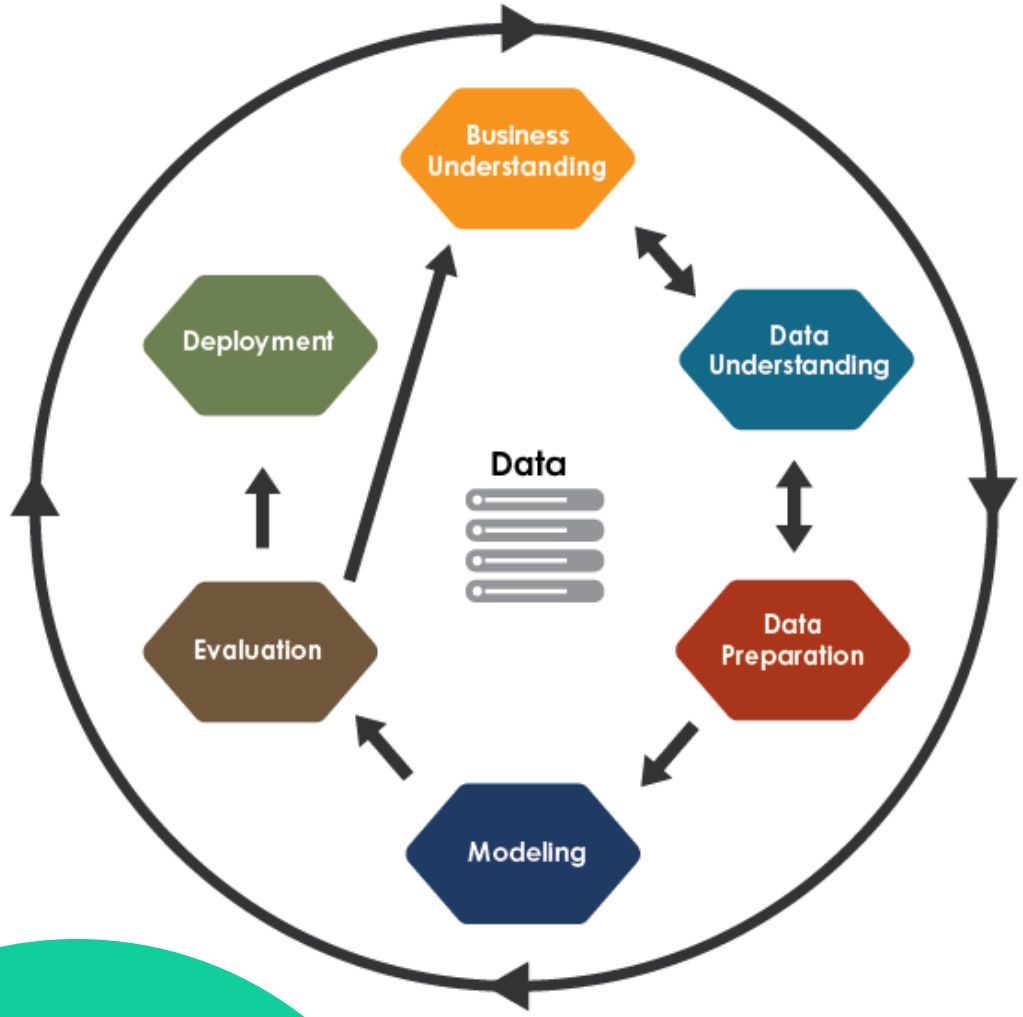

Abbildung 1

Agenda

1. Business Understanding & Datenexploration
    - 1.1 Business Context
    - 1.2 Relevanz von Data Science
    - 1.3 Explorative Datenanalyse
    - 1.4 Statistische Analyse
    - 1.5 potenzielle Verzerrungen
2. Datenaufbereitung
3. Modellierung & Evaluation
4. Geschäftsempfehlungen & Kritische Reflexion
    - 4.1 Geschäftsempfehlungen
    - 4.2. Kritische Reflexion

1.1 Business Context

Unser Ziel:

→ Wir möchten anhand des Datensatzes, große Mengen an Startups ressourceneffizient für Investoren bewerten, um mögliche Unicorns zu identifizieren

1. Welche Kennzahlen beeinflussen die Profitabilität für Investoren?

2. Wie gut kann ein datenbasiertes Modell Vorhersagen über die Profitabilität von Start-ups treffen?

1.2 Inwiefern ist Data Science relevant?

Weniger psychologische Verzerrungen

Confirmation Bias

Halo-Effekt

Recency Bias

Überblick über große Start-up Mengen verschaffen

Besonders relevante und risikoarme Start-ups rausfiltern

→ Chancen und Risiken werden faktenbasiert und objektiv bewertet

1.3 Explorative Datenanalyse

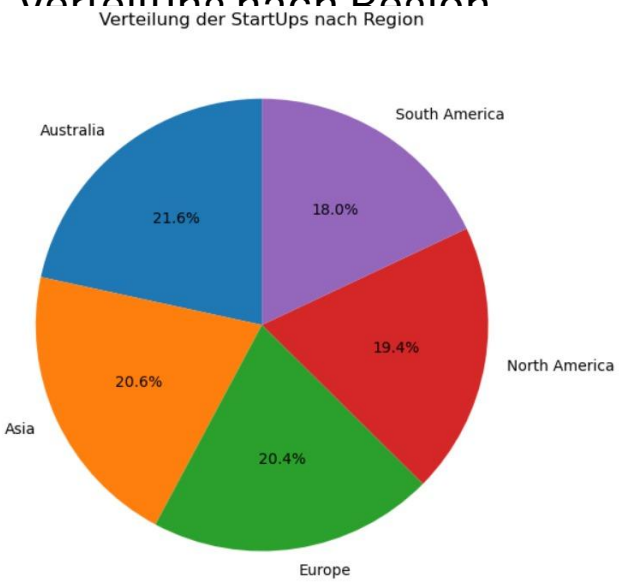

Abbildung 2

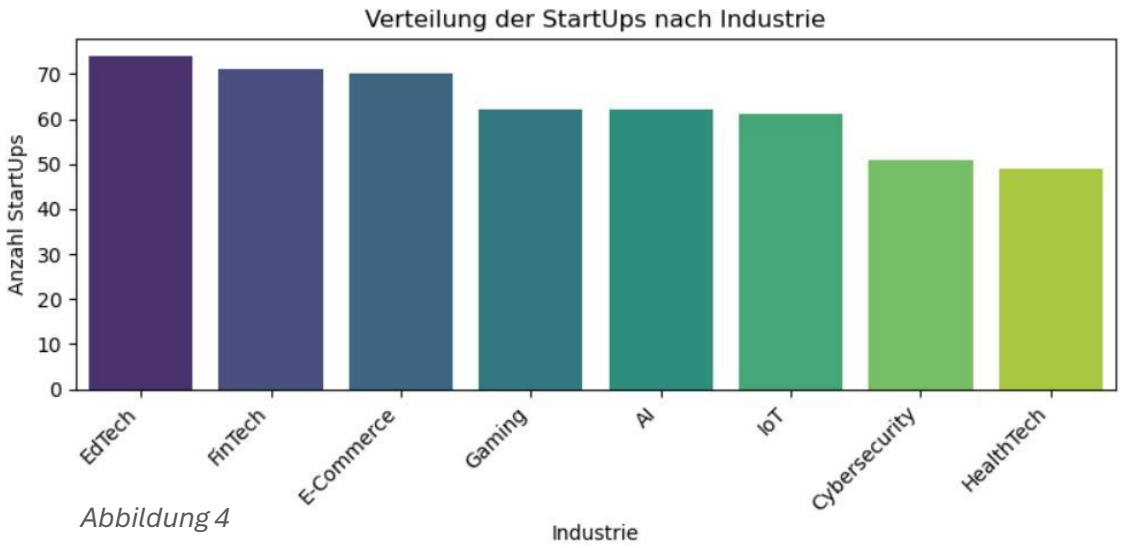

Abbildung 4

Umsatzeffizienz nach Branche

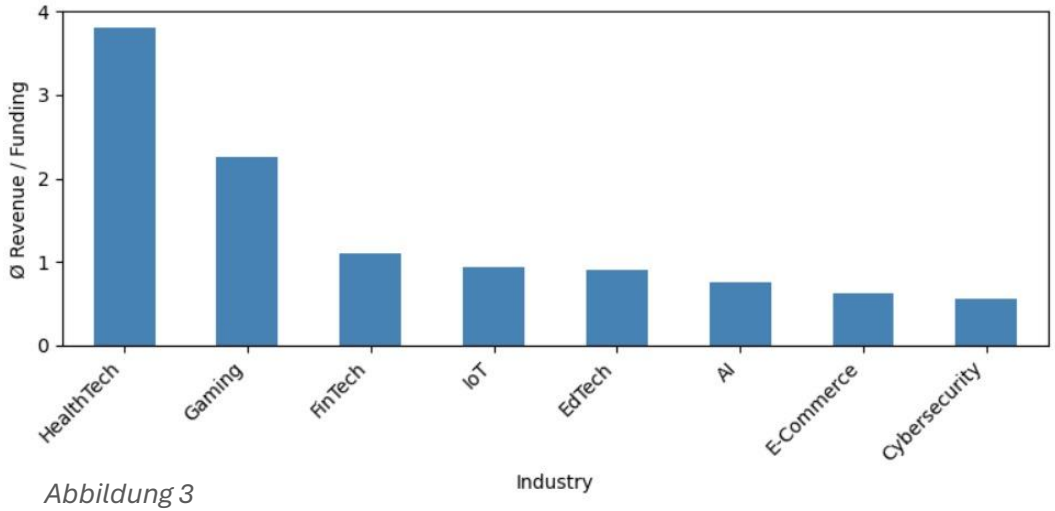

Abbildung 3

1.4 Statistische Analyse

- Korrelationsmatrix zeigt, wie 2 Variablen zusammenhängen
- Dabei gilt: Korrelation ≠ Kausalität
- Bei 500 Start-ups sind Werte > 0,088 mathematisch „echt“
- Bei Korrelationswerten von…
  - …0,3 und kleiner ist der Effekt klein
  - … 0,5 und größer ist de Effekt stark

→ Lösung der Geschäftsfragen durch Korrelationsanalyse nicht möglich

→ trotzdem relevant in Verbindung mit Feature Engineering

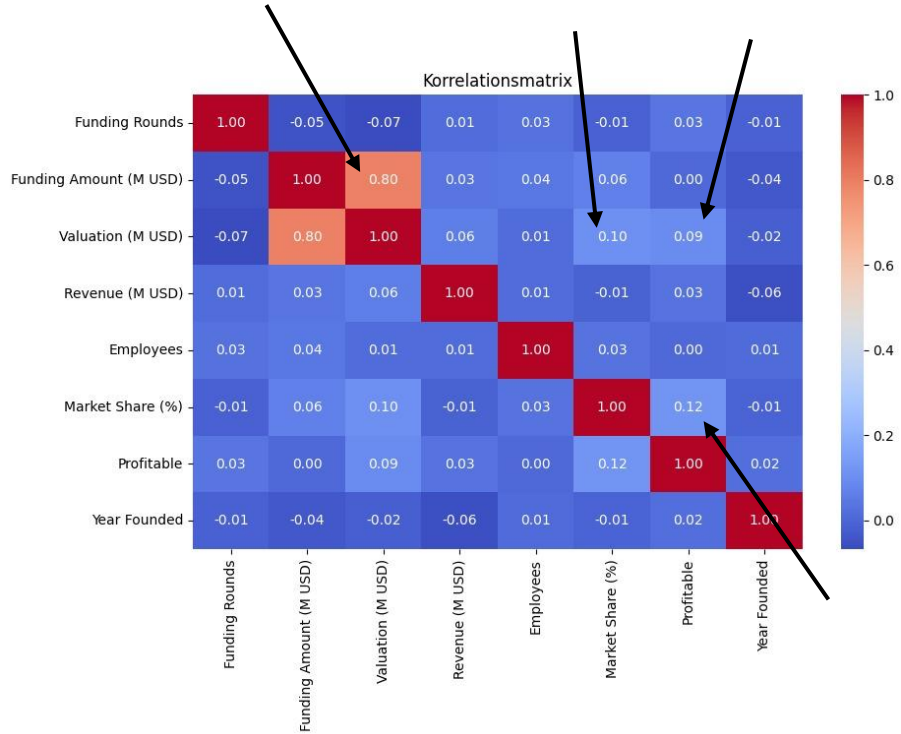

Abbildung 5

1.5 Potenzielle Verzerrungen im Datensatz

- Kleine Stichprobe
- Survivorship Bias
- Regional/Branchen-Bias
- Zeitlicher Bias
- Erfolgskonzentration
- Qualitative Daten
- Vage Definitionen

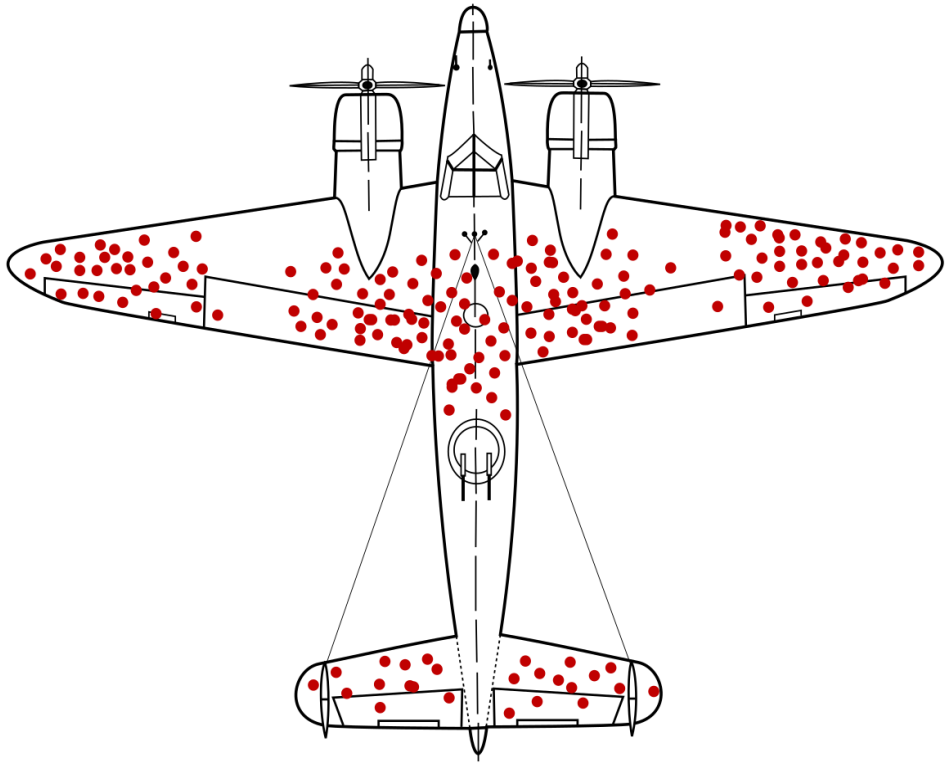

Abbildung 6

2. Fehlende Werte & Ausreißer

Fehlende Werte

- Ergebnis der Analyse:
  - Keine fehlenden Werte
- Interpretation & Bedeutung:
  - Außergewöhnliche Qualität
  - Hohe Zuverlässigkeit
- Konsequenzen für die Analyse:
  - Keine Imputation nötig
  - Kein Informationsverlust
  - Vereinfachte Modellierung

Ausreißer

- Hervorragende Datenqualität:
  - Extrem niedrige Ausreißerrate (0,2%)
  - Keine weiteren Anomalien
- Einzelausreißer Startup 385:
  - Valuation mit 5.357,49 Mio. USD
  - Abweichung: Marginal
- Handlung:
  - Beibehalten
  - Start-up-Ökosystem: Extremwerte sind Branchenüblich

3. Modellierung & Evaluation

Modellauswahl

- Decision Tree
  - Interpretierbarkeit
  - Umgang mit nichtlinearen Zusammenhängen
  - Klares Bias/Varianz-Verhalten

Evaluation

- ROC-AUC
- Confusion Matrix
- Kontextbezogene Metriken

Alternativen?

4.1 Geschäftsempfehlungen

Was bedeutet das?

- Pipeline zur Datenvorverarbeitung
- Decision Tree:
  - Baum aus Entscheidungsregeln
- Wie geht man mit den Werten um?

Was tun?

- NICHT! Blind investieren
- Empfehlungen sollen helfen
- Daten nach dem Invest nutzen

Wo sind die Grenzen?

- Datenqualität
- Features
- Modell
- Sonstiges

Startup Name	Industry	Region	Prob. Profit	Prognose (y_pred)	Realität (y_actual)
Startup_78	Cybersecurity	Australia	0.689	1	1
Startup_289	Gaming	Australia	0.689	1	1
Startup_330	Gaming	Asia	0.689	1	1
Startup_14	Cybersecurity	South America	0.689	1	0
Startup_234	Gaming	North America	0.689	1	0
Startup_8	HealthTech	Australia	0.689	1	1
Startup_208	Gaming	Asia	0.689	1	1
Startup_218	E-Commerce	Europe	0.689	1	1
Startup_418	E-Commerce	North America	0.689	1	1
Startup_192	Cybersecurity	Asia	0.689	1	0

Abbildung 7

4.2 Kritische Reflexion

Survivorship Bias? Fehlen gescheiterte Startups? Datensatz repräsentativ für gesamtes Marktgeschehen?

Modell-Performance? Sind 60% Accuracy / 35% Recall besser als zufälliges Raten? Wo liegt die Schwäche?

Kausalität vs. Korrelation? Führt Funding zu Erfolg oder erhalten erfolgreiche Startups mehr Funding (Selection Bias)?

Zeitliche Validität? Direkter Vergleich von Startups über 30 Jahre hinweg (Kohorteneffekte) ohne Normalisierung sinnvoll?

Umsetzbare Hebel? Sind Empfehlungen für Investoren oder Gründer überhaupt konkret steuer- und umsetzbar?

Qualitative Faktoren? Wie werden Faktoren wie Teamqualität, Vision und Anpassungsfähigkeit bewertet, die nicht im Datensatz sind?

Abbildungen

1. https://www.datascience-pm.com/wp-content/uploads/2021/02/CRISP-DM.png.webp
2. Verteilung nach Regionen mithilfe der Mathplotlib-Bibliothek erstellt
3. Verteilung nach Branchen mithilfe der Mathplotlib-Bibliothek erstellt
4. Umsatzeffizienz nach Branchen mithilfe der Mathplotlib-Bibliothek erstellt
5. Korrelationsmatrix mithilfe der Mathplotlib erstellt
6. https://upload.wikimedia.org/wikipedia/commons/thumb/b/b2/Survivorship-bias.svg/1280px-Survivorship-bias.svg.png
7. Tabelle