Data Analytics with Statistics (WS 24/25)
Data Science (WS 24/25)
Dozent: Jan Kirenz

Projekt: Fußballer Marktwert Vorhersage

Antonino Piloro

INHALT

Einführung

- Profifußball ist globales Wirtschaftsphänomen mit Milliardenumsätzen
- Spielertransfers und Marktwerte sind zentrale wirtschaftliche Faktoren
- Komplexe Bewertung durch viele Einflussfaktoren wie:
  a. Leistungsdaten
  b. Alter und Position
  c. Internationale Erfahrung
  d. Transfermarktdynamik

Einführung

Projektziel

- Entwicklung eines präzisen Vorhersagemodells für Spielermarktwerte
- Einsatz von maschinellem Lernen und statistischer Analyse

Einführung

Wissenschaftliche Relevanz

- Interdisziplinärer Ansatz verbindet:
  a. Statistik und Informatik
  b. Ökonomie und Sportwissenschaft
-
- Praktischer Nutzen für Vereine und Scouts

- Tiefere Einblicke in einen der größten Sportmärkte weltweit

Einführung

Hypothese

- Marktwertvorhersage durch Baumdiagramm-Modell möglich

- Basierend auf systematischer Analyse von:
  a. Leistungsdaten
  b. Physischen Eigenschaften
  c. Altersfaktoren
  d. Markttrends
  e. Transferhistorien

Daten

Datengrundlage

- Quelle: Transfermarkt.com (via Kaggle)
- 10 verknüpfte CSV-Dateien
- Automatische wöchentliche Aktualisierung durch API
- Speicherung als Parquet-Dateien

Daten

Datengrundlage

Players (Spielerprofil)

- Umfassende persönliche und professionelle Daten
- Beinhaltet:
  - Persönliche Details (Alter, Herkunft)
  - Sportliche Merkmale (Position, Größe)
  - Vertragsinformationen
  - Aktueller und höchster Marktwert

Appearances (Einsatzdaten)

- Detaillierte Spielstatistiken pro Match
- Erfasst wichtige Leistungsindikatoren:
  - Tore und Vorlagen
  - Gelbe/Rote Karten
  - Gespielte Minuten

Player Valuations (Marktwertentwicklung)

- Zeitliche Entwicklung der Spielermarktwerte
- Enthält für jeden Zeitpunkt:
  - Aktuellen Marktwert in Euro
  - Vereinszugehörigkeit
  - Liga-Information

Transfers (Transferhistorie)

- Dokumentiert alle Spielerwechsel
- Wichtige Informationen wie:
  - Transfer-Zeitpunkt und Saison
  - Abgebender und aufnehmender Verein
  - Ablösesumme und Marktwert

Daten

Datenaufbereitung - Zwei zentrale Datensätze

Combined Data Aggregiert

- Enthält aktuelle Werte pro Spieler
- Fokus auf letzten verfügbaren Marktwert
- Basis für Momentaufnahme-Analysen

Combined Data Historisch

- Komplette Zeitreihe der Spielerdaten
- Ermöglicht Analyse von Entwicklungen

Daten

Datenaufbereitung - Zwei zentrale Datensätze

Combined Data Aggregiert

	name	current_market_value	height_in_cm	foot	position	age	total_goals
0	Miroslav Klose	1000000.0	184.0	right	Attack	47.0	48
1	Roman Weidenfeller	750000.0	190.0	left	Goalkeeper	45.0	0
2	Dimitar Berbatov	1000000.0	NaN	None	Attack	44.0	38
3	Lúcio	200000.0	NaN	None	Defender	47.0	0
4	Tom Starke	100000.0	194.0	right	Goalkeeper	44.0	0

Combined Data Historisch

	name	current_market_value	height_in_cm	foot	position	age	goals_per_game	transfer_fee
54	Kevin De Bruyne	45000000.0	181.0	right	Midfield	21.0	0.222222	8000000.0
82	Dusan Tadic	3200000.0	181.0	left	Attack	24.0	0.343750	5500000.0
198	Steven Davis	200000.0	172.0	right	Midfield	27.0	0.066667	1000000.0
238	Nikolaos Karelis	500000.0	173.0	left	Attack	20.0	0.000000	20000.0
309	Filip Kostić	6500000.0	184.0	left	Defender	20.0	0.000000	1250000.0

Daten

Feature Engineering & Datenbereinigung

Neue Features erstellt

- Transfer-Gap: Differenz zwischen Ablöse und Marktwert
- Retired-Status: Spieler gilt als retired nach 5 Jahren Inaktivität

```
## Feautre Engineering: transfer gap
combined_data_hist['transfer_fee_gap'] = combined_data_hist['transfer_fee']
                                       - combined_data_hist['transfer_market_value']
## Aktiv oder Rente
current_year = 2025
combined_data_hist['is_retired'] = combined_data_hist['last_season'] <= (current_year - 5)
combined_data_agg['is_retired'] = combined_data_agg['last_season'] <= (current_year - 5)
```

Daten

Feature Engineering & Datenbereinigung

Datenbereinigung

- Entfernung unwichtiger Spalten (URLs, Agenten, etc.)
- Filterung: Nur Spieler über 15 Jahre
- Grund: Jüngere Spieler haben oft unrealistische Marktwerte

```
## Feautre Engineering: Alter filtern
combined_data_hist = combined_data_hist[combined_data_hist['age'] > 15]
combined_data_agg = combined_data_agg[combined_data_agg['age'] > 15]
```

EDA

auf Streamlit

Modell

Methodik

Datenstrukturierung Zwei Kategorien-Ansatz

- Aggregierte Daten erfassen Momentaufnahmen und statische Zusammenhänge
- Historische Daten ermöglichen Analyse von Entwicklungen und Trends
- Jede Kategorie wird in drei Analyseebenen untersucht

```
# für aggregiert combined_data_agg
features_agg = [
    "height_in_cm",
    "country_of_citizenship",
    "foot", "position",
    'total_goals','total_assists',
    'total_yellow_cards',
    'total_red_cards',
    'total_minutes_played',
    'age',
    'is_retired']
features_hist = [
   "historical_to_date__market_value",
   "year", "height_in_cm",
   "country_of_citizenship","foot",
   "position", 'goals_per_game',
'assists_per_game','yellow_cards_per_game',
'red_cards_per_game','minutes_played_per_ga
me','age', 'is_retired',"transfer_fee_gap",
"transfer_fee"]

target = "current_market_value"
```

Modell

Methodik

Drei analytische Perspektiven

- Gesamtmarktanalyse durch vollständigen Datensatz
- Stabilisierte Analyse durch bereinigten Datensatz ohne Ausreißer
- Premium-Segment-Fokus durch Top-5-Ligen-Datensatz

```
q1 = combined_data_hist["current_market_value"]
.quantile(0.25)
q3 = combined_data_hist["current_market_value"]
.quantile(0.75)
iqr = q3 - q1
lower_bound = q1 - 1.5 * iqr
upper_bound = q3 + 1.5 * iqr

# Filtern der Daten ohne Ausreißer
combined_data_hist_no_outliers =
combined_data_hist[
   (combined_data_hist["current_market_value"] >=
lower_bound) &
   (combined_data_hist["current_market_value"] <=
upper_bound)]
```

Modell

Methodik

Modellierungsstrategien

- Lineare Regression
  a. als Fundament
- Random Forest
  a. für komplexe Muster
- Gradient Boosting
  a. hohe Anpassungsfähigkeit
- Hist Gradient Boosting
  a. schneller als GB bei großen Datensätze

Modell

Methodik

Multi-Metrische Bewertung

- MSE
  a. absolute Fehlerquadrate
- R^2
  a. Modellgüte und Varianzaufklärung
- MAPE
  a. relative Abweichungen

=== Modellvergleich Aggregiert===

Modell	MSE	R^2	MAPE
Linear Regression	35689279745879.8	0.1769	9.7341
Linear Regression ohne Außreiser	143645224685.5	0.2472	1.4729
Linear Regression Top Spieler	136960341106665.3	0.2328	8.1965
Random Forest	19223612414564.2	0.5566	1.7499
Random Forest ohne Außreiser	123137631693.8	0.3547	1.1806
Random Forest Top Spieler	59370630822828.3	0.6674	1.2308
Gradient Boosting	21260003210651.5	0.5097	2.8145
Gradient Boosting ohne Außreiser	124454997763.5	0.3478	1.2632
Gradient Boosting Top Spieler	58421343068698.2	0.6727	1.9378
Hist Gradient Boosting	17642493737218.6	0.5931	1.9981
Hist Gradient Boosting ohne Außreiser	115399112748.4	0.3953	1.1544
Hist Gradient Boosting Top Spieler	57557462167448.0	0.6776	1.4613

=== Modellvergleich Historisiert===

Modell	MSE	R^2	MAPE
Linear Regression	43957324172981.0	0.4517	7.0864
Linear Regression ohne Außreiser	562450087542.4	0.3284	1.8511
Linear Regression Top Spieler	124279373378924.2	0.4644	4.1891
Random Forest	21334905511297.5	0.7339	1.4094
Random Forest ohne Außreiser	285220509528.6	0.6594	0.9363
Random Forest Top Spieler	72165262598932.0	0.6890	1.3455
Gradient Boosting	28999361567938.7	0.6383	3.0771
Gradient Boosting ohne Außreiser	353662557446.8	0.5777	1.1543
Gradient Boosting Top Spieler	87557730241633.9	0.6227	2.0314
Hist Gradient Boosting	25249665107947.7	0.6851	1.8114
Hist Gradient Boosting ohne Außreiser	313399233645.2	0.6258	1.0157
Hist Gradient Boosting Top Spieler	75740548721061.5	0.6736	1.6188

Modell

Ergebnisse

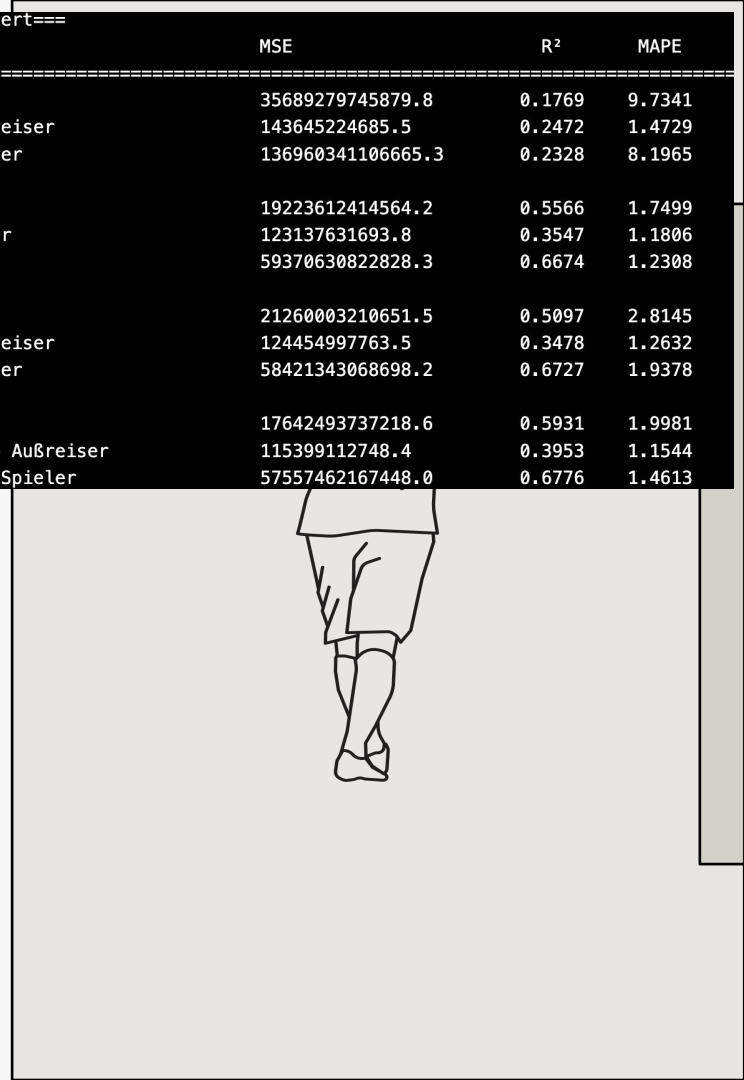
```
=== Modellvergleich Aggregiert===
Modell                                  MSE                    R²       MAPE
==========================================================================================
Linear Regression                       35689279745879.8       0.1769   9.7341
Linear Regression ohne Außreiser        143645224685.5         0.2472   1.4729
Linear Regression Top Spieler           136960341106665.3      0.2328   8.1965

Random Forest                           19223612414564.2       0.5566   1.7499
Random Forest ohne Außreiser            123137631693.8         0.3547   1.1806
Random Forest Top Spieler               59370630822828.3       0.6674   1.2308

Gradient Boosting                       21260003210651.5       0.5097   2.8145
Gradient Boosting ohne Außreiser        124454997763.5         0.3478   1.2632
Gradient Boosting Top Spieler           58421343068698.2       0.6727   1.9378

Hist Gradient Boosting                  17642493737218.6       0.5931   1.9981
Hist Gradient Boosting ohne Außreiser   115399112748.4         0.3953   1.1544
Hist Gradient Boosting Top Spieler      57557462167448.0       0.6776   1.4613
```

Aggregierte Daten

- Lineare Regression
  a. zeigte die schwächste Performance
  b. R^2 = 0,1769 bis 0,2472
  c. MAPE = 1,4729 bis 9,7341

- Gradient Boosting und Random Forest zeigten vergleichbar starke Performance
  a. R^2 = 0,6727 GB
  b. R^2 = 0,6674 RF

- Hist Gradient Boosting
  a. höchste Vorhersagegenauigkeit für Top-Spieler
     R^2 = 0,6776



Modell

Ergebnisse

```
=== Modellvergleich Historisiert===
Modell                                  MSE                  R²       MAPE
================================================================================
Linear Regression                       43957324172981.0     0.4517   7.0864
Linear Regression ohne Außreiser        562450087542.4       0.3284   1.8511
Linear Regression Top Spieler           124279373378924.2    0.4644   4.1891

Random Forest                           21334905511297.5     0.7339   1.4094
Random Forest ohne Außreiser            285220509528.6       0.6594   0.9363
Random Forest Top Spieler               72165262598932.0     0.6890   1.3455

Gradient Boosting                       28999361567938.7     0.6383   3.0771
Gradient Boosting ohne Außreiser        353662557446.8       0.5777   1.1543
Gradient Boosting Top Spieler           87557730241633.9     0.6227   2.0314

Hist Gradient Boosting                  25249665107947.7     0.6851   1.8114
Hist Gradient Boosting ohne Außreiser   313399233645.2       0.6258   1.0157
Hist Gradient Boosting Top Spieler      75740548721061.5     0.6736   1.6188
```

Historisierte Daten

- Alle nicht-linearen Modelle übertrafen die lineare Regression deutlich
  a. MAPE = 1,85%

- Random Forest dominierte mit bester Gesamtperformance
  a. R^2 = 0,7339
  b. MAPE: 0,94%

Modell

Ergebnisse

Weitere Erkenntnisse

- Lineare Regression zeigte durchgehend schwächste Performance, bestätigt Nichtlinearität
- Ausreißerbereinigung führte bei allen Modellen zu Reduktion des MSE, aber teilweise Verlust des R^2
- Hohe MSE-Werte aufgrund der hohen Skalierung der Zielvariable (Marktwerte)

Modell

Ergebnisse

Weitere Erkenntnisse

- Alle Tree-basierte Modelle erreichten MAPE-Werte unter 3%
- Historisierung führte bei allen Modellen zu Verbesserung der Vorhersagegenauigkeit

DISKUSSION & FAZIT

Ergebnisse

Modellgüte und Limitationen

- Konzentration auf Top-Spieler zeigte keinen signifikanten Vorteil
- Beste Modelle mit MAPE unter 1,5% und R^2 bis 0,7339 (73% Varianzaufklärung)
- Verbleibende 27% Varianz durch qualitative und emotionale Faktoren nicht erfasst

DISKUSSION & FAZIT

Ergebnisse

Emotionale Komponenten

- Emotionale Dimension des Fußballs entzieht sich Quantifizierung

- Aspekte wie Fanbindung, Charisma, Vereinstreue spielen bedeutende Rolle

DISKUSSION & FAZIT

Ergebnisse

Strukturelle Limitationen

- Nicht-Quantifizierbarkeit qualitativer Faktoren
  a. wie Teamchemie oder Markenwert
- Zeitliche Verzögerungen bei Datenverfügbarkeit
- Externe Einflüsse wie makroökonomische Faktoren oder Ereignisse
  a. Financial Fair Play

DISKUSSION & FAZIT

Ergebnisse

Zukünftige Forschungsansätze

- Integration qualitativer und emotionaler Faktoren
  a. Natural Language Processing für Medien- und Social-Media-Analysen

- Erweiterung der Datenintegration
  a. Analyse medizinischer Daten, Untersuchung Social-Media-Effekte

DISKUSSION & FAZIT

Ergebnisse

Fazit

- Ergebnisse liefern Erkenntnisse zur Marktwertbildung, zeigen aber auch Grenzen
- Entwickelte Modelle bieten Unterstützung
- Erkenntnisse bilden Basis für Transferansätze
- Tragen zum besseren Verständnis der Marktwertbildung im Profifußball bei

DANKE