Hall of Fame predictor

Die **Baseball Hall of Fame** ist die Ruhmeshalle der größten US-amerikanischen Baseballspieler. Dieses Museum ist der Geschichte des Baseballs und der Helden der Major Leagues gewidmet.

In einem Baseballspiel gibt es drei grundlegende Aufgaben, die erfüllt werden müssen:

1. **Werfen des Balls (Pitcher)**
2. **Schlagen des Balls (Batter)**
3. **Abwehren und Fangen des Balls (Fielder)**

Die **Lahman Baseball-Datenbank** bietet umfassende MLB-Statistiken seit 1871.
Sie enthält Daten zu Spielern, Teams und Spielen und wird oft für Analysen und Data-Science-Projekte genutzt.

1

Tabellen der Lahman Datenbank als CSV heruntergeladen

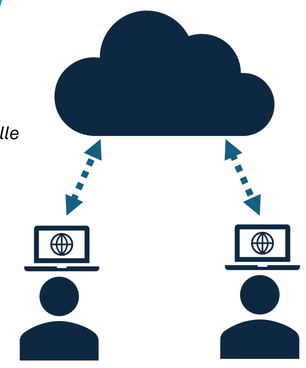
2

playerID	yearID	stint	teamID	lgID	W	L	G	GS	CG	SHO	SV	IPouts	H	ER	HR	BB	SO	BAOpp	ERA	IBB	WP	HBP	BK	BFP	GF	R	SH	SF	GIDP
aardsda01	2004	1	SFN	NL	1	0	11	0	0	0	0	32	20	8	1	10	5	0,417	6,75	0	0	2	0	61	5	8	0	1	1
aardsda01	2006	1	CHN	NL	3	0	45	0	0	0	0	159	41	24	9	28	49	0,214	4,08	0	1	1	0	225	9	25	1	3	2
aardsda01	2007	1	CHA	AL	2	1	25	0	0	0	0	97	39	23	4	17	36	0,3	6,4	3	2	1	0	151	7	24	2	1	1
aardsda01	2008	1	BOS	AL	4	2	47	0	0	0	0	146	49	30	4	35	49	0,268	5,55	2	3	5	0	228	7	32	3	2	4
aardsda01	2009	1	SEA	AL	3	6	73	0	0	0	38	214	49	20	4	34	80	0,19	2,52	3	2	0	0	296	53	23	2	1	2
aardsda01	2010	1	SEA	AL	0	6	53	0	0	0	31	149	33	19	5	25	49	0,198	3,44	5	2	2	0	202	43	19	7	1	5
aardsda01	2012	1	NYA	AL	0	0	1	0	0	0	0	3	1	1	1	1	1	0,25	9	0	0	0	0	5	1	1	0	0	0
aardsda01	2013	1	NYN	NL	2	2	43	0	0	0	0	119	39	19	7	19	36	0,257	4,31	6	1	4	1	178	7	20	2	1	2
aardsda01	2015	1	ATL	NL	1	1	33	0	0	0	0	92	25	16	6	14	35	0,223	4,7	3	1	1	0	129	9	17	0	1	4
aasedo01	1977	1	BOS	AL	6	2	13	13	4	2	0	277	85	32	6	19	49	0,244	3,12	1	0	1	0	373	0	36	2	3	7

Beispiel Pitching Tabelle

Für das Projekt zählen insbesondere persönliche Statistiken, weswegen beispielsweise die Tabelle der geworfenen Bälle besonders wichtig war. Tabellen wie die der Manager war für uns nicht von Bedeutung. Da diese keinen Einfluss auf die Spieler haben wurden nicht weiter berücksichtigt.

3



Hochladen der CSV Dateien in Cloud für kollaboratives Arbeiten

4

```
player_IDs = set(HallOfFame_copy.query('inducted == "Y"').playerID)
player_IDs

Batting_copy['hof'] = Batting_copy['playerID'].isin(player_IDs).astype(int)
BattingPost_copy['hof'] = BattingPost_copy['playerID'].isin(player_IDs).astype(int)

Pitching_copy['hof'] = Pitching_copy['playerID'].isin(player_IDs).astype(int)
PitchingPost_copy['hof'] = PitchingPost_copy['playerID'].isin(player_IDs).astype(int)

Fielding_copy['hof'] = Fielding_copy['playerID'].isin(player_IDs).astype(int)
FieldingPost_copy['hof'] = FieldingPost_copy['playerID'].isin(player_IDs).astype(int)
```

Lediglich 1,7% aller Spieler haben es bisher in die Hall of Fame geschafft

Hall of Fame Spieler identifizieren um die restlichen Daten Anhand der Spieler-ID zu labeln

5

Identifizieren aller irrelevanten Spalten

playerID	yearID	stint	teamID	lgID	W	L	G	GS	CG	SHO	SV	IPouts	H	ER	HR	BB	SO	BAOpp	ERA	IBB	WP	HBP	BK	BFP	GF	R	SH	SF	GIDP
aardsda01	2004	1	SFN	NL	1	0	11	0	0	0	0	32	20	8	1	10	5	0,417	6,75	0	0	2	0	61	5	8	0	1	1
aardsda01	2006	1	CHN	NL	3	0	45	0	0	0	0	159	41	24	9	28	49	0,214	4,08	0	1	1	0	225	9	25	1	3	2
aardsda01	2007	1	CHA	AL	2	1	25	0	0	0	0	97	39	23	4	17	36	0,3	6,4	3	2	1	0	151	7	24	2	1	1
aardsda01	2008	1	BOS	AL	4	2	47	0	0	0	0	146	49	30	4	35	49	0,268	5,55	2	3	5	0	228	7	32	3	2	4
aardsda01	2009	1	SEA	AL	3	6	73	0	0	0	38	214	49	20	4	34	80	0,19	2,52	3	2	0	0	296	53	23	2	1	2
aardsda01	2010	1	SEA	AL	0	6	53	0	0	0	31	149	33	19	5	25	49	0,198	3,44	5	2	2	0	202	43	19	7	1	5
aardsda01	2012	1	NYA	AL	0	0	1	0	0	0	0	3	1	1	1	1	1	0,25	9	0	0	0	0	5	1	1	0	0	0
aardsda01	2013	1	NYN	NL	2	2	43	0	0	0	0	119	39	19	7	19	36	0,257	4,31	6	1	4	1	178	7	20	2	1	2
aardsda01	2015	1	ATL	NL	1	1	33	0	0	0	0	92	25	16	6	14	35	0,223	4,7	3	1	1	0	129	9	17	0	1	4
aasedo01	1977	1	BOS	AL	6	2	13	13	4	2	0	277	85	32	6	19	49	0,244	3,12	1	0	1	0	373	0	36	2	3	7

Beispiele für irrelevante Spalten:
L: Anzahl verlorener Spiele
SHO: Shoutouts (fast überall 0)
lgID (Liga-ID) und teamID
WP: Wild Pitches (Besondere Fehlwürfe)

6

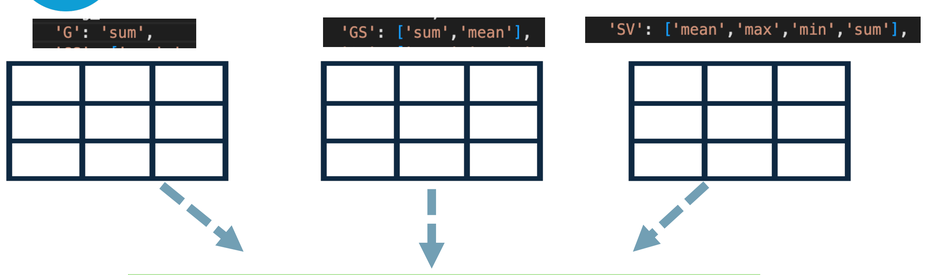
Entfernen der irrelevanten Spalten

playerID	yearID	G	GS	SV	IPouts	H	ER	HR	BB	BAOpp	BFP	GF	R	hof
aardsda01	2004	11	0	0	32	20	8	1	10	0,417	61	5	8	0
aardsda01	2006	45	0	0	159	41	24	9	28	0,214	225	9	25	0
aardsda01	2007	25	0	0	97	39	23	4	17	0,3	151	7	24	0
aardsda01	2008	47	0	0	146	49	30	4	35	0,268	228	7	32	0
aardsda01	2009	73	0	38	214	49	20	4	34	0,19	296	53	23	0
aardsda01	2010	53	0	31	149	33	19	5	25	0,198	202	43	19	0
aardsda01	2012	1	0	0	3	1	1	1	1	0,25	5	1	1	0
aardsda01	2013	43	0	0	119	39	19	7	19	0,257	178	7	20	0
aardsda01	2015	33	0	0	92	25	16	6	14	0,223	129	9	17	0
aasedo01	1977	13	13	0	277	85	32	6	19	0,244	373	0	36	0
aasedo01	1978	29	29	0	536	185	80	14	80	0,27	773	0	88	0
aasedo01	1979	37	28	2	556	200	99	19	77	0,277	817	4	104	0

7

Erstellung von Aussagekräftigen Kennzahlen für jeden Spieler in der Datenbank.
Beispielsweise: Maximale Anzahl an getroffenen Schlägen, Minimale Anzahl, Durchschnittliche Anzahl, aufsummierte Anzahl, Anzahl aller gewonnen Awards, etc..

8

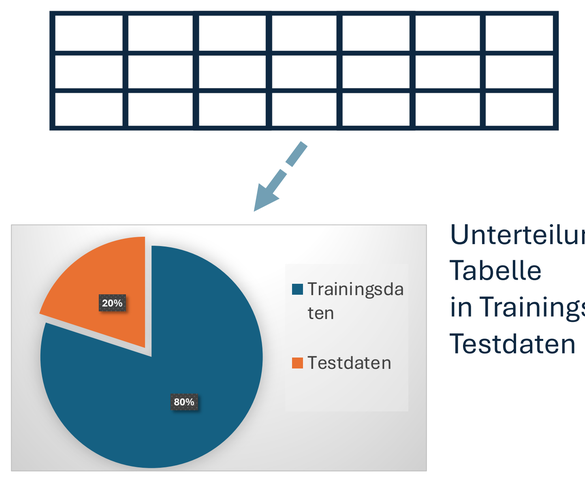


G	GS	GS	SV	SV	SV	SV
sum	sum	mean	mean	max	min	sum
331	0	0.0	7.666666666666667	38	0	69
448	91	7.0	6.307692307692307	34	0	82
406	6	0.5	0.16666666666666666	1	0	2
79	65	10.833333333333334	0.16666666666666666	1	0	1

Die Originale Tabelle umfasst 45 Spalten.
Zur Veranschaulichung wurde hier nur ein kleiner Teil dargestellt

Erstellung einer Tabelle, welche alle Spielerstatistiken für jeden Einzelnen Spieler beinhaltet.

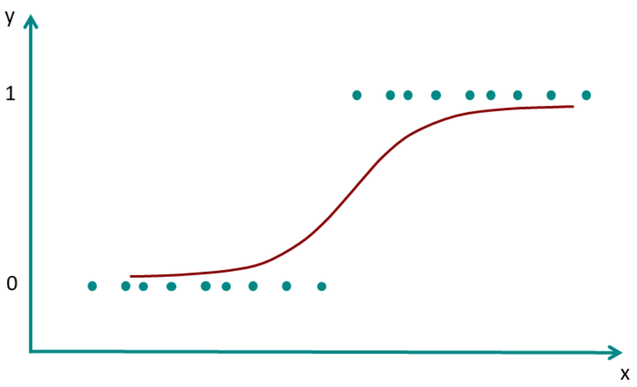
9



Unterteilung der Tabelle in Trainings- und Testdaten

1 Logistische Regression

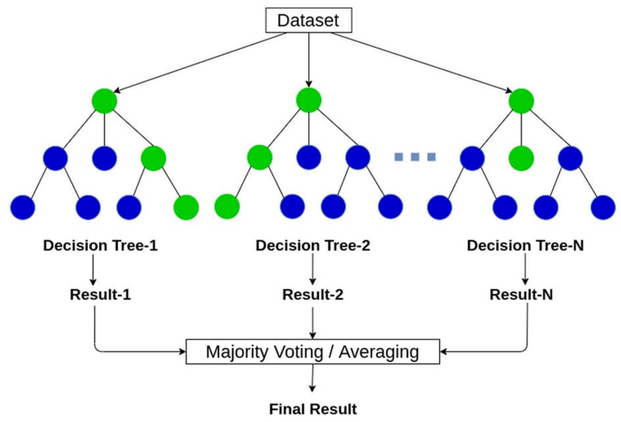
Accuracy: **0.9884**
Precision: **0.7174**
Recall: **0.4853**
F1-Score: **0.5789**



Die logistische Regression bietet eine einfache, interpretierbare Basis für die Vorhersage.
Sie quantifiziert die Wahrscheinlichkeit der Aufnahme anhand linearer Zusammenhänge.

2 Random Forest

Accuracy**: 0.9889**
Precision**: 0.775**
Recall**: 0.4559**
F1-Score: **0.5741**



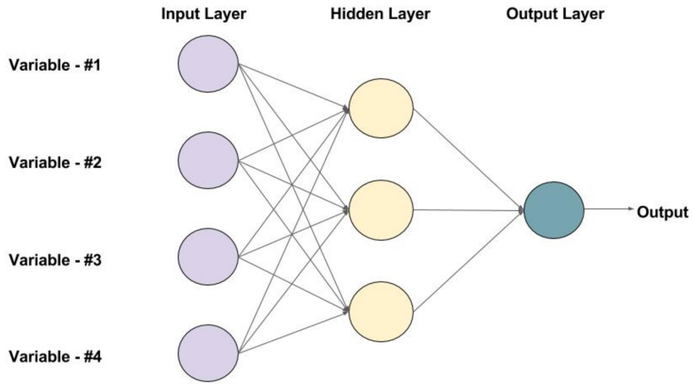
Der Random Forest wurde genutzt,
da er robuste Vorhersagen liefert und
wichtige Merkmale identifizieren kann.
Er ist ideal für strukturierte Daten und unbalancierte Klassen.

3 Deep Learning

Accuracy: **0.958**
Precision: **0.5042**
Recall: **0.61**
F1-Score: **0.548**

Loss: **0.0908**
Epochen: **500** (Batch**: 32)**
Neuronen: **128 – 64 – 32 - 1**
Dropout-Rate: **30%**
Lern-Rate: **Adam-Optimizer**

Feedforward Neural Network (FNN)



Das FNN wurde gewählt, um komplexe und nicht-lineare Zusammenhänge in den Spielerstatistiken und Karriereverläufen zu erkennen. Es eignet sich besonders für große und vielschichtige Datensätze wie die Lahman DB.

Fazit

Unsere Analyse zeigt, dass sowohl die logistische Regression als auch der Random Forest **hohe Genauigkeiten** erzielen. Allerdings haben beide Modelle deutliche Schwächen im **Recall**, was bedeutet, dass sie viele potenzielle Hall-of-Fame-Spieler übersehen. Das FNN hingegen erreicht einen höheren Recall, jedoch auf Kosten der **Präzision** und **Gesamtgenauigkeit**.
Diese Ergebnisse machen deutlich, dass eine **rein statistische Betrachtung** der Spielerleistungen an ihre Grenzen stößt, wenn es darum geht, Hall-of-Fame-Aufnahmen zuverlässig vorherzusagen.
Die Entscheidungen scheinen von weiteren Faktoren beeinflusst zu sein, die in den vorhandenen Daten nicht erfasst wurden – etwa Karrierekontext, **Medienpräsenz** oder **subjektive Wahrnehmungen**.
Insgesamt zeigt das Projekt die Machbarkeit eines datengetriebenen Ansatzes, macht aber auch die **Herausforderungen** deutlich.
Besonders die Auswahl relevanter Features spielt eine große Rolle, ebenso wie der Umgang mit unausgeglichenen Daten.
Zukünftige Arbeiten könnten sich auf die Integration zusätzlicher Datensätze und die Entwicklung komplexerer Modelle konzentrieren, um die Vorhersageleistung weiter zu verbessern.