Predicting outcomes of Bundesliga matches

Predstavitev seminarske naloge
pri predmetu Strojno učenje

Timur Kulenović, januar 2021

Cilji seminarske naloge

- Klasifikacijski problem → **H** (dom. ek.), **D** (neodločeno), **A** (gost. ek.)
- Sestaviti **atributni prostor** za napovedovanje izidov tekem
- Učenje različnih algoritmov strojnega učenja

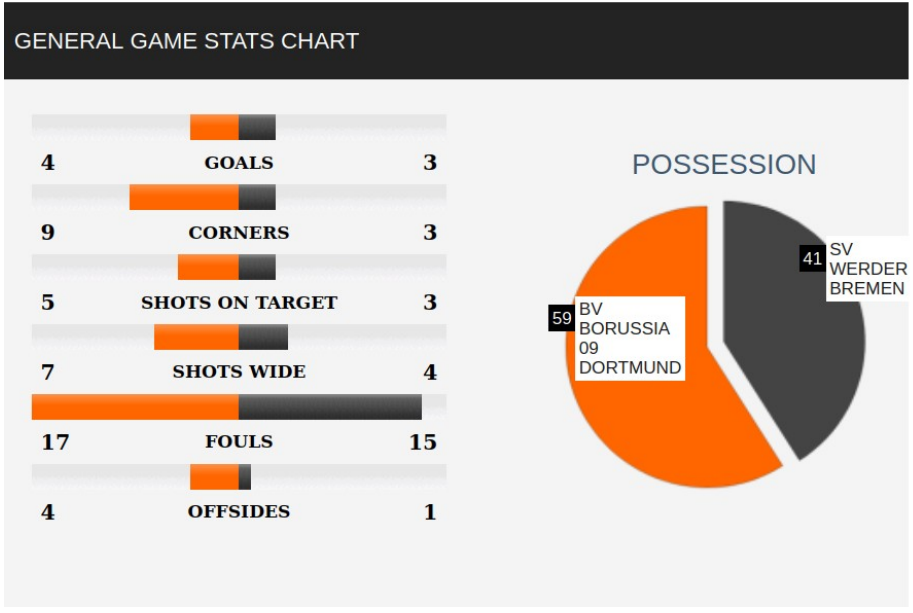
na 7 sezonah (2011/2012 – 2017/2018) nemške nogometne lige

in testiranje na sezoni **2018/2019**

- Primerjava s stavnimi kvotami s pomočje metrike **RPS** (ranked probability score)

Spletno luščenje podatkov

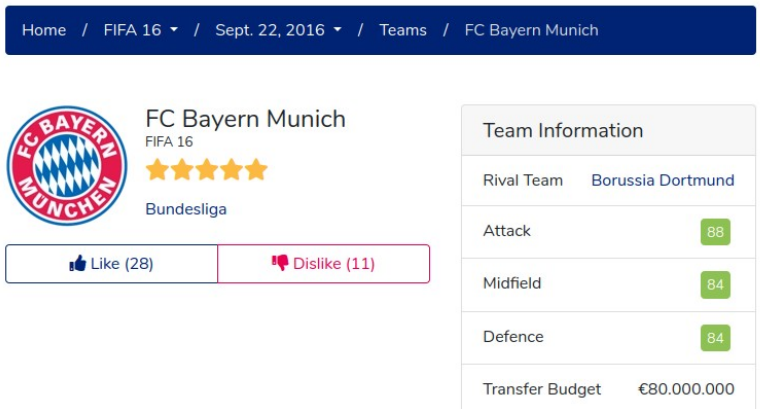
Podatki o posameznih tekmah – soccerway.com

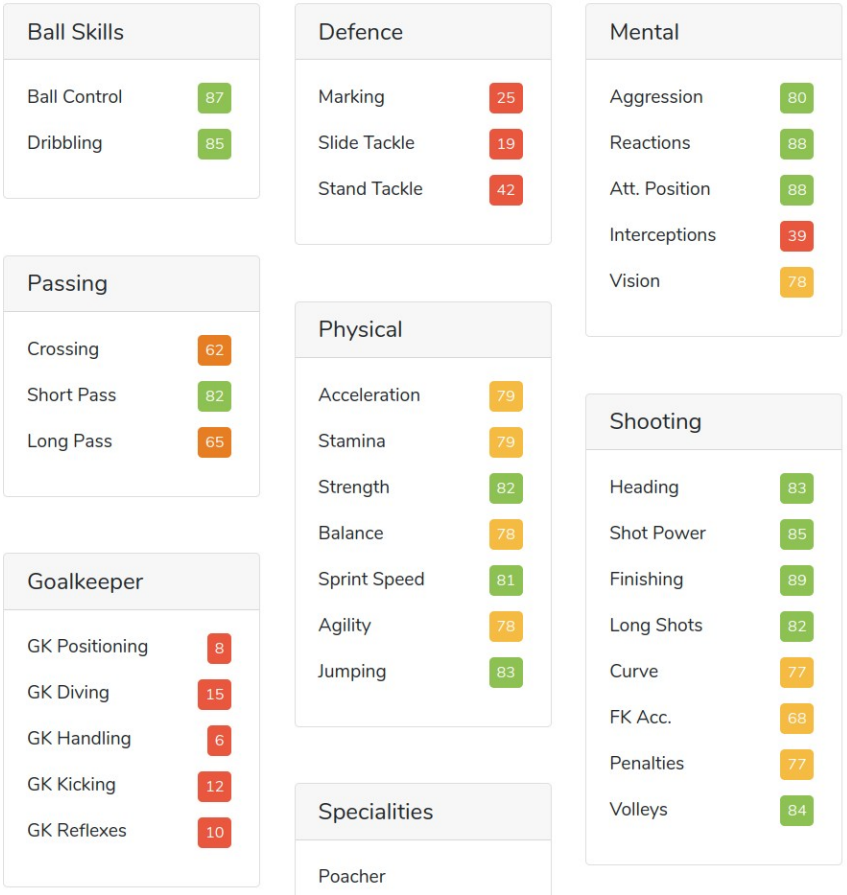


38	R. Bürki		42	F. Wiedwald	
25	S. Papastathopoulos	11'	18	N. Moisander	
5	Marc Bartra		23	T. Gebre Selassie	
13	Raphaël Guerreiro		26	L. Sané	
37	E. Durm		13	M. Veljkovic	90+2'
28	M. Ginter		20	U. Garcia	
8	N. Şahin	62'	16	Z. Junuzović	7'
23	S. Kagawa	87'	22	F. Bartels	46' 58'
17	P. Aubameyang	42' 89'	6	T. Delaney	
11	M. Reus	32' 75'	27	F. Grillitsch	
7	O. Dembélé	61'	10	M. Kruse	68'

Spletno luščenje podatkov

Podatki o ocenah za posamezno sezono – Fifaindex.com:





Feature engineering – pretekle predstave ekipe

Za vsako od obeh ekip na tekmi:

- AVG vrednosti za število strelov, golov, kotov itd. na zadnjih *k* tekmah (6 atributov)
- Število točk na zadnjih *k* tekmah
- Uteženo število točk na zadnjih *k* tekmah (starejša kot je tekma, manj šteje)
- Gol razlika in število točk na **vseh dosedanjih** tekmah (2 atributa, neodvisna od *k*)
- **Forma:**
  - zgornji atributi **ne upoštevajo nasprotnikov** ekipe na zadnjih obračunih
  - Začetek sezone → vse ekipe forma = 1
  - V primeru zmage → ekipi se forma poviša ter nasprotni ekipi zniža za isto vrednost (0.33)
  - V primeru neodločenega izida → boljša ekipa formo izgubi, slabša pa pridobi

- V našem primeru bo **k = 5** (napovedujemo lahko torej le od 6. kroga dalje)
- **V prostor atributov dodamo razlike atributov**: Att_H - Att_A

Feature engineering – ocene ekspertov (Fifaindex)

Za vsako od obeh ekip na tekmi:

- Ocene lastnosti ekip - Attack, Defence, Midfield, Overall, Budget (5 atributov)
- Za Physical in Mental lastnosti samo povprečne vrednosti (brez vratarjev, 14 atributov)
- 5 top Ball Skilled odštejemo vrednosti top 5 Tacklerjem naspr. ekipe (povprečje, 2 atributa)
- Podobno 5 najboljšim podajalcem odštejemo vrednosti top 5 Marker naspr. ekipe (povprečje, 3 atributi)
- Vrednostim Shooting lastnosti odštejemo kvaliteto vratarja naspr. ekipe (povprečje, 7 atributov)

Opozorilo: podatke o igralcih lahko dobimo **1 uro** pred tekmo
V prostor atributov ponovno dodamo razlike atributov: Att_H - Att_A

Rezultati algoritmov

- **Izbrani algoritmi:**

logistična regresija, random forest, gradient boosting, multilayer perceptron

- Modeli **učeni na podatkih iz sezon 2011/2012 - 2017/2018**

in **testirani na sezoni 2018/2019**

- Rezultati modelov precej podobni med seboj
- slabo napovedovanje neodl. izidov

LOGISTIC REGRESSION CONFUSION MATRIX.

	Pred. Home	Pred. Draw	Pred Away
Actual Home	97	0	20
Actual Draw	41	3	16
Actual Away	31	2	51

ACCURACY OF MODELS

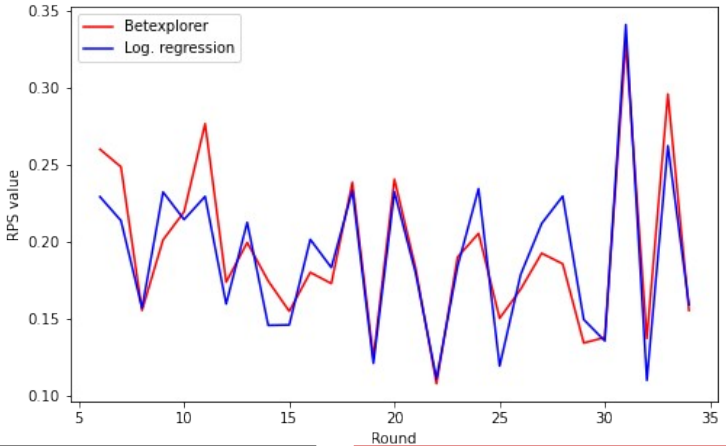
Model	Accuracy
Logistic Regression	0.5785
Random Forest	0.5875
Gradient Boosting	0.5985
MLP Classifier	0.5762

Primerjavo s stavnimi kvotami

- Stavne kvote za 2018/2019 → **BetExplorer**
- Kvote pretvorimo v verjetnosti (inverz in normalizacija)
- Metrika **Ranked probability Score:**
  - primerna zaradi ordinalnosti razredov (Če je izid H, je boljše napovedati D kot A)
  - nižje vrednosti so boliše (ocena napake modela)

RPS VALUES

Model	Value
Betexplorer	0.1931
Logistic Regression	0.1904
Random Forest	0.1912
Gradient Boosting	0.1918
MLP Classifier	0.1900



Zaključki

- Modeli za sezono 2018/2019 uspejo “premagati” kvote na Betxplorer
- Forma igralcev
- Izbira parametrov z rolling window pristopom
- Posplošitev na več sezon