Прогнозиране на електронни спортове

Никола Шахпазов

6/5/2021

Въведение

Последното десетилетие, електронните спортове са набрали огромна популярност, особено сред младите. Дори на компютрите за развлечение във ФМИ може да се видят играещи и разгорещено спорещи студенти, а в международен план се провеждат състезания с наградни фондове за десетки милиони. Електронният характер позволява събиране на данни до най-малката детайлност на игрите/играта, както преди самото започване на мачовете, така и от реално време.

В съответната теза ще разгледаме кратко описание на електронните спортове (играта дота), подхода при извличане и обработка на данни, както и няколко възможни модела които могат да се приложат за решаването на изхода от мача.

За играта Dota 2

Dota 2 (Защита на Древните) е може би най-популярния електронен спорт, а предшественика му - DOTA е първият такъв.

Двата отбора водят териториална военна битка един с друг, като целта на всеки е да унищожи древните храмове противника си. Освен това, върху картата, двата отбора имат и военни крепости от които на помощ им идват компютърно симулирани войници, чието унищожаване дава злато и точки опит за противниковия отбор.

В една класическа партия участват два отбора, всеки с по пет играча, като всеки играч играе с определен герой.

Броят на героите надвишава 112, като всеки герой има предопределена роля в играта и в този смисъл, правилната комбинация на героите има значителен принос върху вероятността за победа на единия или другия отбор. Героите са предпоределени от ролите за които са подходящи да участват в играта, техни свойства, атрибути и способности.

Свойствата и особеностите на героите, техните способности в битка предоставят голямо разнообразие и различие между героите, давайки на всеки герой негово уникално тактическо предимство за ситуации в играта и различни местоположения по картата. Поради това, за да сформират добър отбор, играчите трябва да са добре запознати със силните и слабите страни на всеки герой.

Съответните роли от които всеки играч може да бъде са

- **Носач(Carry)** - отговорен за носенето на повечето събрано злато от битките с което прави и съотборниците си отговорни за неговото опазване. Докато носачите не участват толкова активно в битките, те трупат злато и опит с който компенсират в късните етапи на играта, където имат далеч повече сила. Името **износван** идва оттам, че макар и в по-късните етапи на играта да са изключително опасни, в началото на играта са немощни и износвани от своите съотборници.
- **Подкрепа (Support)** - героите с подкрепяща роля са отговорни за опазване на своите съотборници живи, така че да могат да събират спокойно злато и точки опит. Това се случва посредством способности за лекуване и защитни магии които героите с подкрепяща роля имат. Също така се занимават с разузнаване на картата и покупка на боеприпаси. Също така, имат способности като лечение на съотборниците си или зашеметяване на противниците си. За тях златото не е такъв приоритет като за носачите. Често се случва да жертват себе си за спечелят време за бягство от съотборниците си.
- **Бомбардировач (Nuker)** - героите от този тип имат експлозивни магии с много високо ниво на щети и бавно възстановяване. Може да нарани и погуби по няколко вражески героя наведнъж.
- **Деактиватор (Disabler)** - деактиваторите имат способност да зашеметяват вражески герои, както и да спират техни магии и удари.
- **Джунглър (Jungler)** - събират злато и артефакти, повечето случаи в джунглата на картата.
- **Странично атакуващи (Off-laner)** - Това е една от най-застрашените роли в цялата игра. Тяхна отговорност е самостоятелно да обезопасят страничните алеи на играта, с което са под голям риск да бъдат атакувани от няколко противникови героя наведнъж.

- **Други**

Използвани данни за Dota 2 и техните предиктори

Данни за предварителни прогнози

За предварителни прогнози, един набор от данни предоставен от Yang et al.[2] съдържа 78362 игри с 19790 участващи в тях играча с висок рейтинг.

Всеки мач има информация за побеждаващия отбор, уникален номер на всеки мач, играчите и избраните герои в съответните мачове. Освен това са събрани хакатеристики и статистики за всеки герой, като способности, силни и слаби страни, препоръчана роля за игра, бързина, сила, интелект и много други.

За играчите имаме исторически статистики като рейтинг оценяващ силата на играча, както и информация за брой победи и загуби с често играни от него герои. Използваните данни са извлечени от сайта http://www.opendota.com (http://www.opendota.com).

Подобни данни използваме и от работата на Conley et al. [1], както и от състезание за прогнозиране на Dota 2 игри проведено в сайта http://www.Kaggle.com (http://www.Kaggle.com).

Данни за предварителни прогнози

За данни от реално време, използваме набор от 50 хиляди игри, предоставени от споменатото по-горе състезание в Kaggle.com При него имаме пълна информация за играчите, избраните герои, победителя, както и информация за различни статистики събрани в реално време на интервал от една минута. Някои такива характеристики са

- Събрано злато за всеки играч
- Натрупан опит за всеки играч
- Брой финализиращи удари
- Разрушаване на вражески кули
- Купени артефакти и избрани подобрения при качване на нива опит
- И други

Предварителни прогнози за победа върху Dota 2 мачове

Една от задачите които представляват интерес за нас, е да успеем да направим точна прогноза върху изхода от мача, единствено с информация за участващите играчи и избраните от тях герои. В този смисъл, целта е да моделираме изхода от мача като Бернулиева Случайна величина

$$Y := \begin{cases} 1, & \text{ако отбор Сияние печели} \\ 0 & \text{иначе} \end{cases}$$

както и вероятността $P(Y = 1)$.

Логистична регресия

Един подход за прогнозирането на Y който е използван в [1], е чрез метода на Логистичната Регресия. Посредством него, можем да представим логистичната трансформация (свързваща функция) на вероятността за победа на определен мач, спрямо вероятността за загуба като линейна комбинация на това дали определен герой участва в някой от отборите.

$$log[\frac{P(Y=1)}{1-P(Y=1)}] = \beta_0 + \sum_{i=1}^{113} \beta_i \mathbb{1}(\text{герой } i \in \text{отбор 1}) + \sum_{i=113}^{226} \beta_i \mathbb{1}(\text{герой } i - 112 \in \text{отбор 2}) =$$

$$= \beta_0 + < \vec{\beta}, \vec{X} >$$

където

$$X_i = \begin{cases} 1, & \text{ако герой i участва в отбор Сияние за i=1,...,113} \\ 1, & \text{ако герой i-112 участва в отбор Мраколес за i=113,...,226} \\ 0 & \text{иначе} \end{cases}$$

Изразявайки спрямо P(Y=1), получаваме

$$P(Y=1|\vec{\beta}) = \frac{1}{1+e^{-(\beta_0 + \sum_{i=1}^{p} \beta_i x_i)}}$$

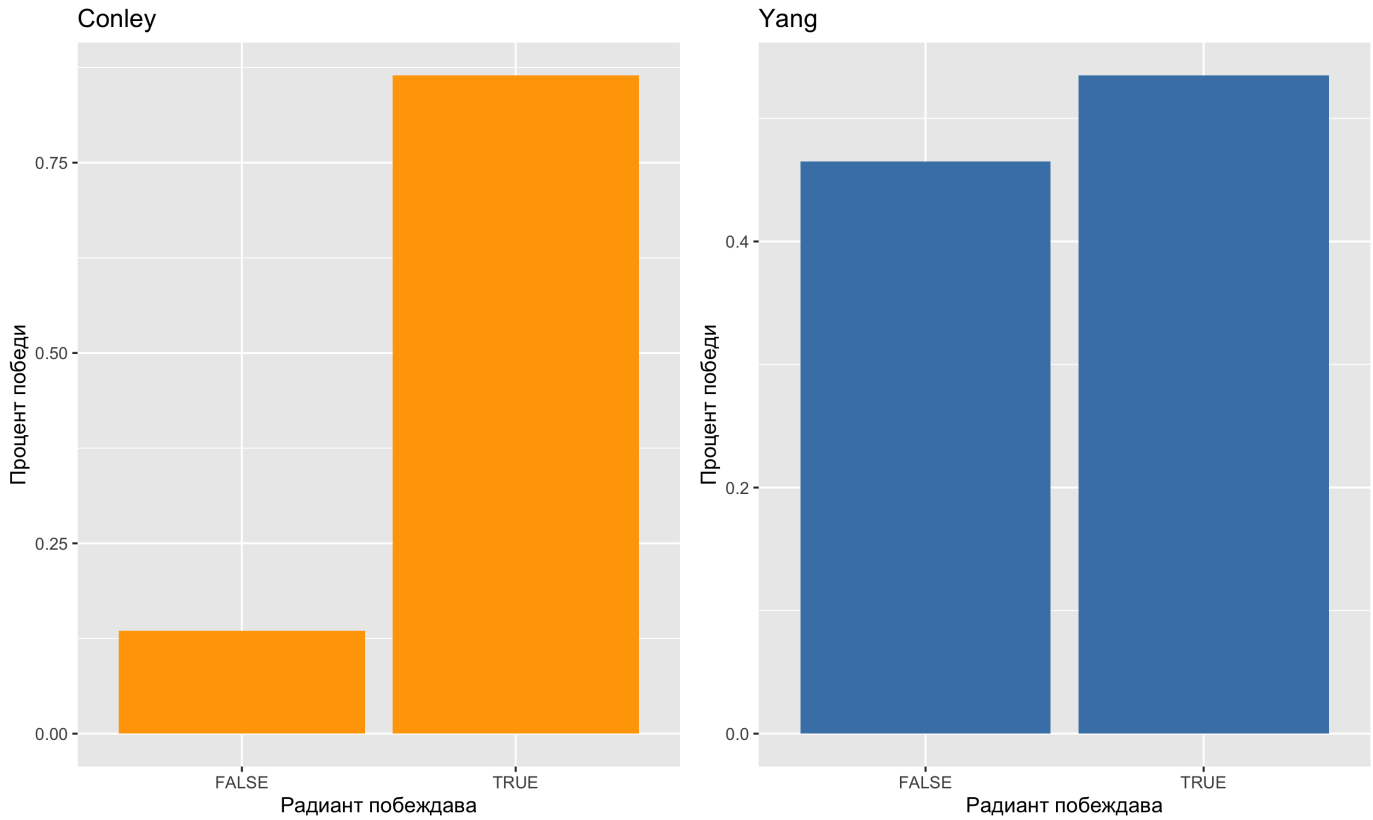
Параметрите $\beta_0, \ldots, \beta_{226}$ оценяваме с метода на **максимално правдоподобие**, при който максимизираме функцията на правдоподобие.

$$\hat{\vec{\beta}} = \underset{\vec{\beta}}{argmax} \prod_{i=1}^{n} P(Y_i = 1|\vec{\beta})$$

за определени $1, \ldots, n$ игри които имаме. За целта, параметрите ги оценяваме върху заделено от нас множество от игри, а проверката за точността на прогнозата я правим върху остатъка (train-test split). Пропорцията която използваме за двете множества е 70% за трениране на модела и 30% за тестване на прогнозата.

За фунцкията на правдоподобие е трудно да се намери аналитично решение и в почето случаи, софтуерните пакети които използваме прилагат числена апроксимация за да намерят параметрите които максимизират функцията.

Оценявайки параметрите и прогнозирайки върху множеството за тестване, Conley at al. [1] са достигнали около 83% оценка за точността върху прогнозата. Извършвайки по детайлно разглеждане на използвните от тях данни, може да се види, че така или иначе, в 86% от случаите, отбор Сияние побеждава и съответната точност не е сравнена с точност от базов модел (**benchmark**). Същия модел е използван в Yang et al. [2], където авторите са получили 60% оценка за точност на прогнозиране на отбор Сияние.



Получаваме около 60% точност на прогнозата, която можем да сравним с точността на базов модел (**baseline*), който прогнозира победителя като отбора който просто има повече победи и дава 51% точност.

Друг подход използван в Yang et al. [2] е отново с Логистична Регресия, но с далеч повече предиктори за изразяването на отклика Y. Освен използваният 226-мерен вектор с индикатори за играните герои, те са използвали и още няколко групи предиктори, като характеристики на играните герои (hero attributes), като

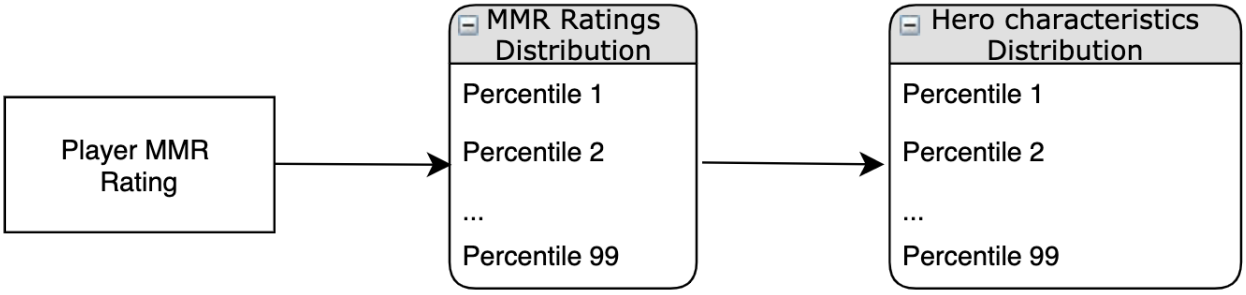
- базова жизненост
- базова сила
- базова регенерация
- базови умения в стрелба
- други.

Друга група която са добавили е статистики от сайта www.dotabuff.com за това колко победи и загуби имат всеки от играчите с избраните от тях герои.

Като финал, авторите са използвали изчисляван от opendota.com рейтинг на играчите, наричан MMR (Matchmaking Rating) и взето от същия сайт разпределение на съответните MMR оценки и разпределение на 7 специфични статистики за играни герои като

- среден количество натрупан опит на минута (**xp_per_min**)
- среден брой убийства на минута (**kills_per_min**)
- средно нанесени щети на минута (**hero_damage_per_min**)
- среден нанесени финализиращи удари на минута (**last_hits_per_min**)
- лекуване на съотборници (**hero_healing_per_min**)
- брой потрошени вражески кули (**tower_damage**)
- средно количество злато събирано на минута (**gold_per_min**)

Посредством рейтинга на играча, разпределението на рейтинга и разпределенията на гореописаните статистики, авторите взимат процентила в който попада рейтинга на всеки играч, и в същия процентил взимат стойността на всяка от статистиките по-горе.



Комбинирайки всички тези предиктори във един 611-мерен вектор от предиктори, авторите твърдят за достигане на 71% точност. Едно по-детайлно разглеждане на ефекта върху точността на всяка група предиктори, стигнахме до съответните количества.

Група Предиктори	Точност
Избрани Герои	59%
Характеристики на Героите	55.67%
Честота на победи с Герой	54%
MMR рейтинги	55%
Трансформация върху MMR	54.35%
Честота на победа на играч с герой	69.72%
Избрани герои + Честота на победи	70.37%
Избрани герои + Честота на победи + Трансформация	70.58%

Виждаме, че предложената трансформация, всъщност е сред най слабата група предиктори - оценката за точността е около 54%, докато най-силна е групата от честотата на победи на играч с герой. 611-мерния вектор спокойно може да се сбие до група от 10 предиктора и това малко да промени или намали точността на модела.

Основен проблем на този подход е, че честотата на победи на определен играч с определен герой, не е съпоставена с определен мач (няма времеви компонент), а е взета в момента на извличане на всички данни, тоест след самите мачове. Това създава предпоставка за теч на информация/данни (**Data Leakage**) и надвишена прогноза за точността на модела, поради използването на предиктори от бъдещето.

Друг проблем който може да се забележи при [2], е че авторите са направили оценка на точността чрез метод на крос-валидация, който пренебрегва времевия компонент в използваните рейтинги. Можем да твърдим, че по-правилен подход би бил със стандартно разделяне на множествата, където първите 70% от мачовете по време са заделени за извличане на закономерност, а последните 30% по време мачове се използват за оценка на точността на прогнозата.

Освен това, използваните MMR рейтинги са взети след края на всички мачове в данните. Използването на рейтинги след съответните мачове е възможно да доведе до нереалистична оценка за точността на прогноза при нови мачове, дори при правилно разбиване на множествата за

трениране и тестване.

Като алтернатива на Логистичната Регресия, Conley et al. са използвали и метод на **К най-близки съседи (KNN)**.

К най-близки съседи (K-Nearest Neighbhours)

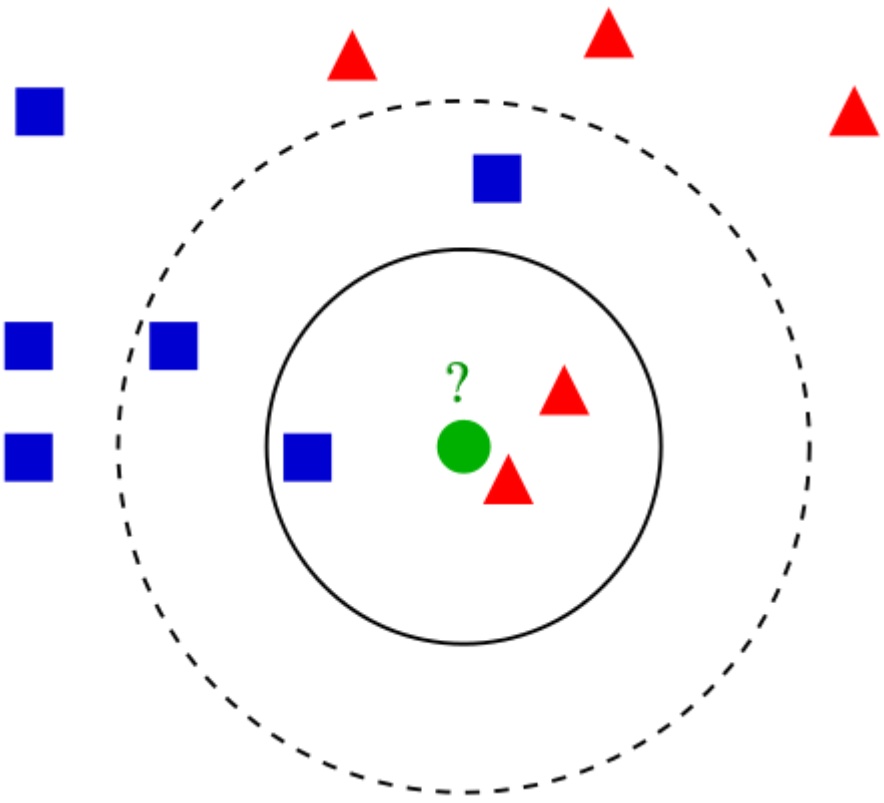
A drawback of the basic “majority voting” classification occurs when the class distribution is skewed.

Класическия метод на най-близките съседи е непараметричен подход, при който ново наблюдение бива класифицирано като най-често срещания клас сред K на брой най-близки елементи, според някаква метрика $\rho(x_1, x_2)$ дефинирана върху пространството от независими променливи $\mathfrak{X}$ (най-често Евклидова).

По-точно, нека $D := (\overrightarrow{X_1}, Y_1), (\overrightarrow{X_2}, Y_2), \ldots, (\overrightarrow{X_n}, Y_n)$ са двойки наблюдения от независими променливи X_i и отклици Y_i (множество за трениране на модела). Ако дефинираме метрика $\rho : \mathfrak{X} \times \mathfrak{X} \to \mathbb{R}$. По този начин, за ново наблюдение (X, Y), можем да конструираме наредба $X_1 \leq X_2 \leq \ldots \leq X_n$ по такъв начин, че $\rho(X, X_1) \leq \rho(X, X_2) \leq \ldots \leq \rho(X, X_n)$. Така за избран от нас хиперпараметър k, можем да изберем множество D_k с първите к на брой елемента от наредбата и да вземем най-често срещания сред тяк клас за съответните отклици $Y_1, \ldots, Y_k$.

$$\hat{Y} = \underset{c}{argmax} \sum_{(x_i, y_i) \in D_k} I(c = y_i)$$

Изборът на k може да се направи чрез, примерно метод на крос валидация, като се изпробват различни стойности за k и се избере съответната, даваща най-висока точност на прогноза.



Предложената модификация от Perry&Conley[1] е, чрез използване на персонализирана метрика $\rho : X \to [0, 1]$ върху предикторното пространство X от възможни комбинации от герои, т.ч. за нов вектор $\vec{q}$ от избрани герои и вектор x от множеството за тренировка

$$\rho(\vec{x}, \vec{q}) := \sum_{i=1}^{226} I(q_i = x_i)$$

добавяне на тегла към функцията

- READ the paper again
- run their code (maybe with a downsampled set since its too slow)
- Compare algorithm complexities of the different algorithms as well
- reproduce their approach in R on the kaggle dataset and see the metrics
- experiment with the parameters, metric functions, etc and plot the results
- explain the model
- read in ESL about the model and reproduce some plots

Прогнози върху данни от реално време

При събраните данни, притежаме информация за събраното злато и натрупания опит от играчите за всяка една минута от съответните партии. Един подход използван в [2] за трениране на прогнозиране върху данни от реално време, е отново да напаснем логистична регресия, но този път върху характеристики като злато, опит и убийства събрани върху интервали от пет минути. Нека първо въведем следните функции

$$G(i, t, T) := \text{Златото на i-тия играч от отбор T в минута t}$$
$$L(i, t, T) := \text{Брой нокаутиращи удари на i-тия играч от отбор T в минута t}$$

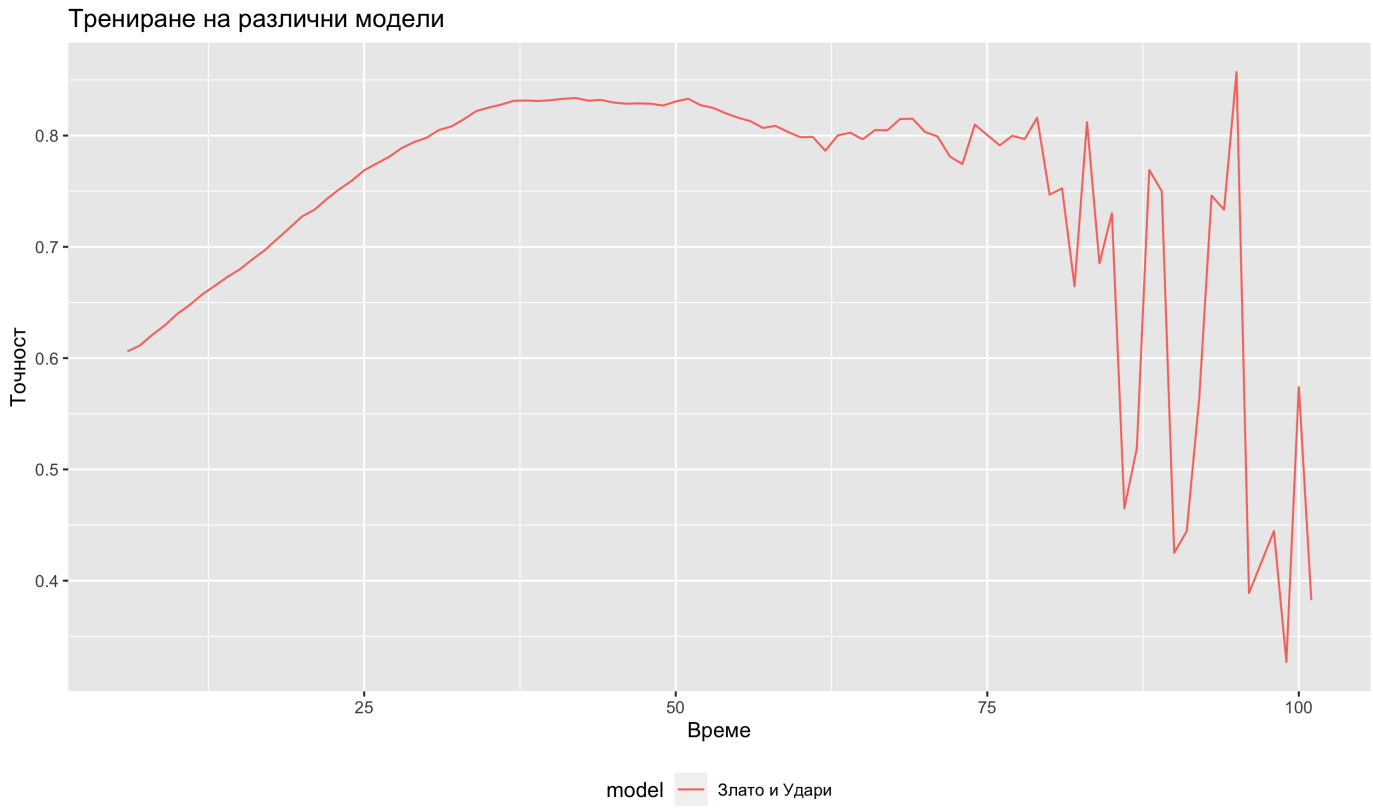
$$P(Y = 1 | t = j; \vec{\beta}) = \frac{1}{1 + exp(-(\beta_0 + \sum_{i=j-6}^{j-1} \beta_i x_i + \sum_{i=j-6}^{j-1} \beta_i' x_i')}, \text{за j=6, ..., T}$$

където x_i е средната разлика в събраното злато между двата отбора в i-тата минута, а x_i' е средната разлика довършващи удари между двата отбора в i-тата минута, т.е.

$$x_i := \frac{1}{5} \sum_{j=1}^{5} [G(j, i, \text{Сияние}) - G(j, i, \text{Мраколес})]$$

$$x_i' := \frac{1}{5} \sum_{j=1}^{5} [L(j, i, \text{Сияние}) - L(j, i, \text{Мраколес})]$$

Независимите променливи злато и финализиращи удари са избрани от общо 3, заедно с натрупан опит, като съответните са дали най-добра крива на прогнозата по време.



Виждаме, че кривата на точността започва силно да се дестабилизира след 75-тата минута, поради малкия брой мачове които имаме с дължина над 75 минути.

TODO: Интерпретация на модела

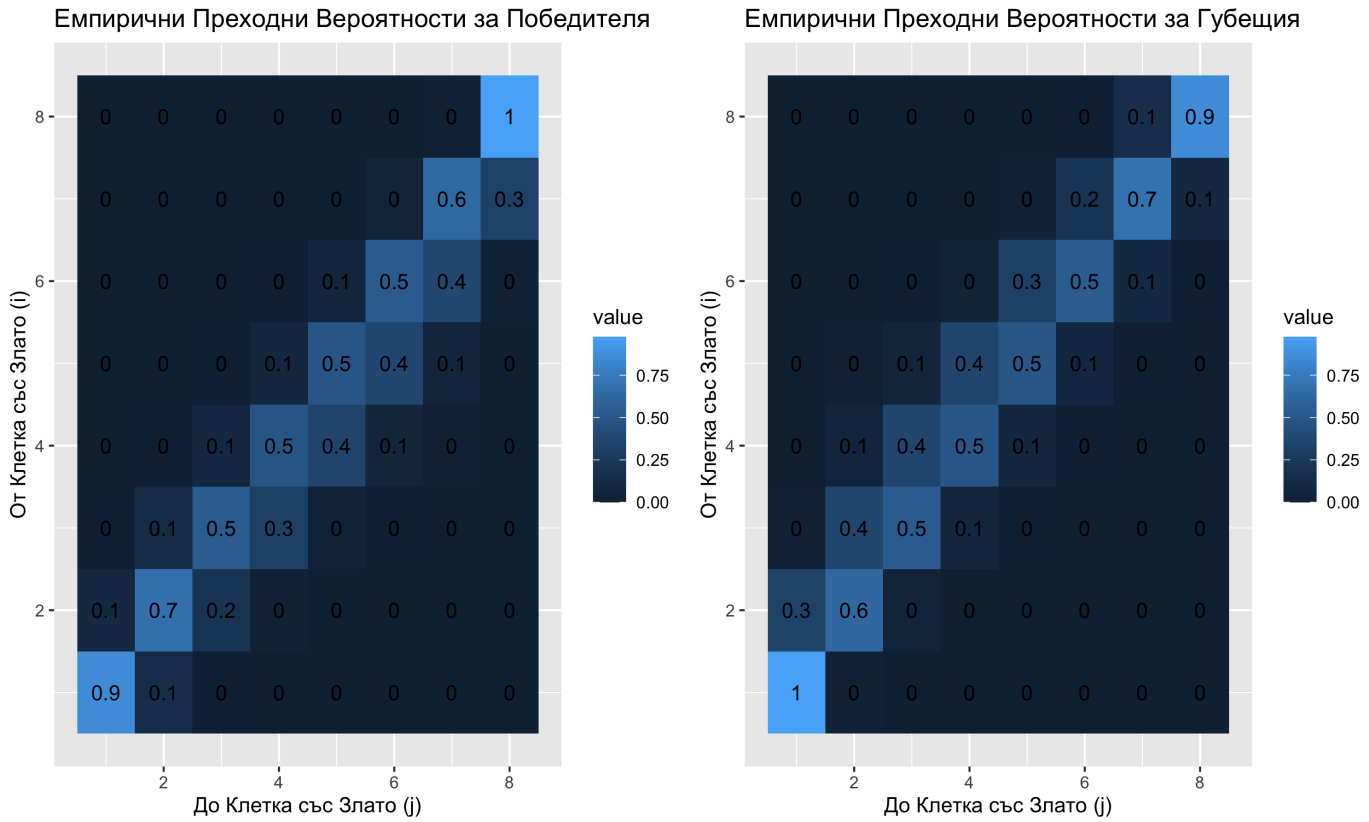
Освен Логистичната Регресия, друг подход разгледан в [1] е изграждане на Марковска Верига.

Дискретна Марковска Верига от първи ред

TODO: Описание на МС

В нашия случай, за да възпроизведем резултатите, взехме средната разлика в златото между отборите по време и дискретизирахме стойностите в 8 отделни клетки за да конструираме пространството на състоянията. Дискретизацията изградихме чрез емпиричните квантили на златото.

Итерирайки по време и броейки честотата на смяна на състоянията на златото (оценка по метод на максимално правдоподобие) за побеждаващия отбор и за губещия отбор, получихме следните две преходни матрици.



Ако искаме да направим прогноза с Марковската Верига във време t, използваме формулата на Бейс $P(Y = r|D)$, където Y - е бернулевата случайна величина описваща победа или загуба, а $D = \{x_{t-5}, \ldots, x_{t-1}\}$ е множеството от средните разлики в златото между двата отбора за предходните 5 минути.

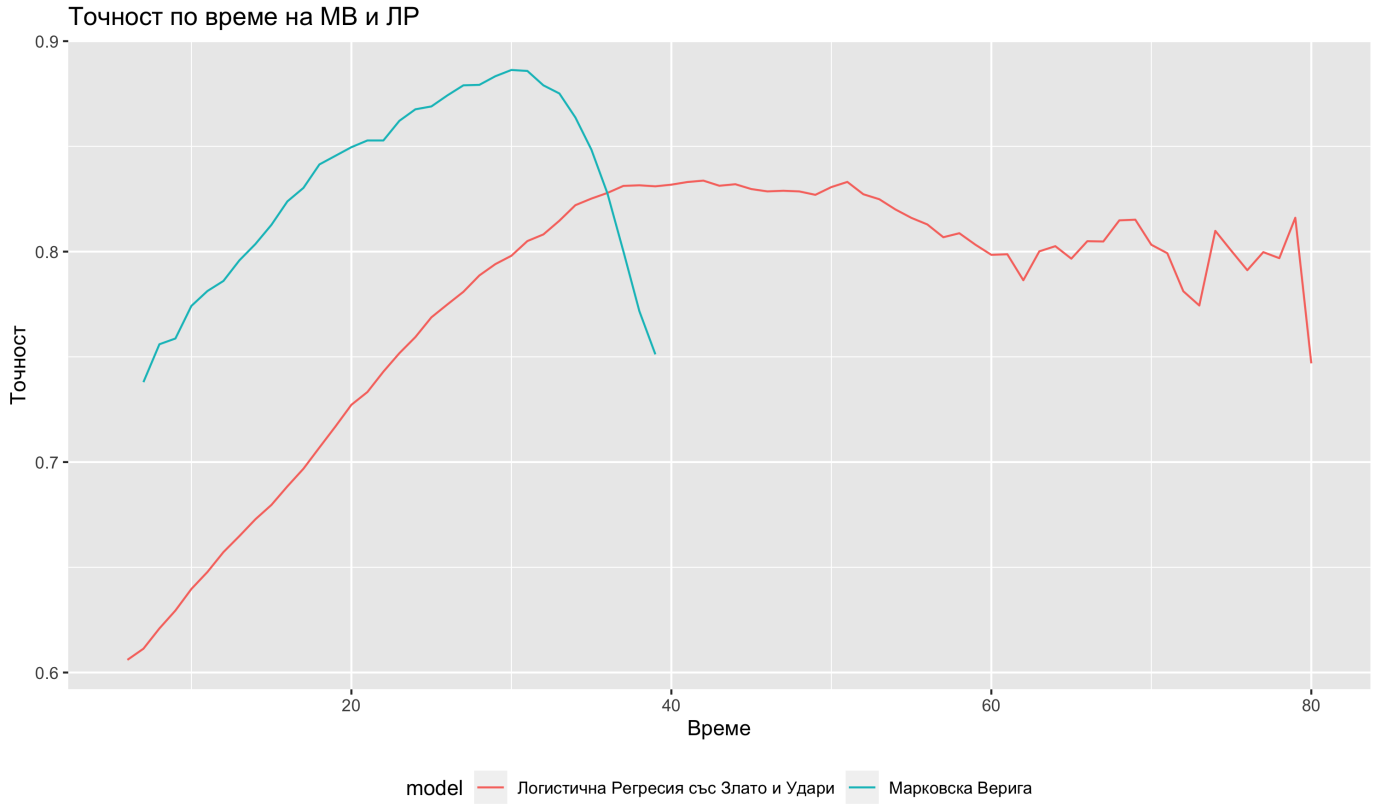
Така получаваме

$$P(Y = r|D) = \frac{P(D|Y = r)P(Y = r)}{P(D)} = \frac{\prod_{i=1}^{5} P(x_{t-i}|Y = r)P(Y = r)}{P(D|Y = 1)P(Y = 1) + P(D|Y = 1)P(Y = 1)}$$

В основата на горната формула стои траекторията на движение $P(D|Y = r)$ за $r \in \{0, 1\}$, тоест при съответно победа и загуба за съответния отбор който ще прогнозираме. За априорната вероятност $P(Y = r)$ бихме могли да използваме и вероятността от прогноза върху данни от преди самото започване на мач (Предварителна Прогноза).

Отново, разбивайки множеството на такова за извличане на закономерности и такова за оценка на точността на прогнозиране, получаваме следните точности по време.

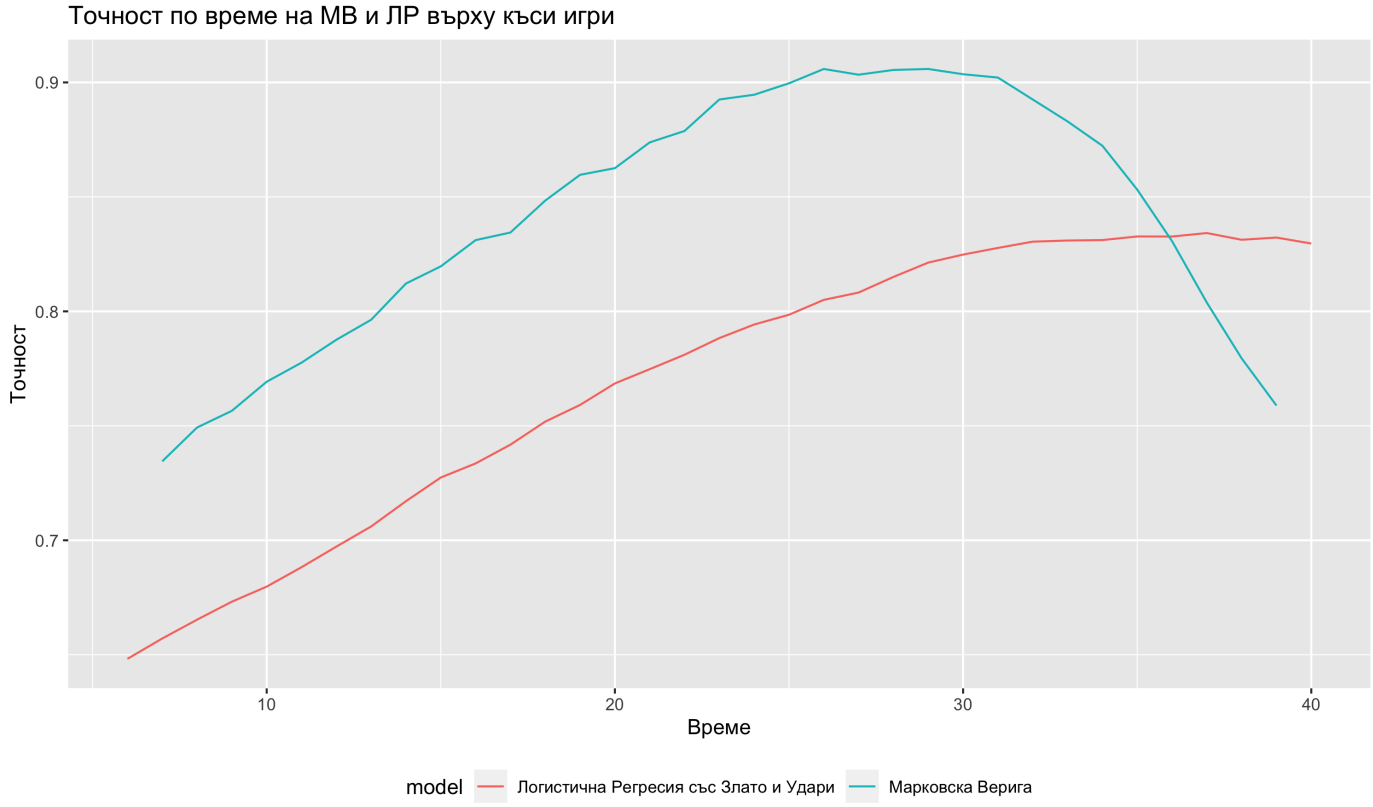
TODO: add accuracy plots for different durations of matches



Виждаме, че Марковската Верига има значително по-добра прогноза в първите 25 минути, след което Логистичната Регресия повежда като точност.

Точността която получаваме съвсем не отговаря на тази предоставена във Yang et al.[2], но пък затова е върху множество от мачове с дължина на някои от тях до 80 минути, а не само до 40 минути.

Взимайки по-къси мачове, успяваме да възпроизведем и даже получим по-висока точност до близо 40-тата минута, като между 25-тата и 30-тата минута, точността е над 90%.



Интуитивно това може да се обясни с това, че при по-късите игри има силен моментум за побеждаващия отбор, който да даде по-добър тренд за движение по състоянията на Марковската Верига. След 40-тата минута, мачовете са далеч по-оспорвани, а и в далеч по-малка наличност за извличане на закономерности. Като модификация на подхода предложен във [2], след известно експериментиране, сме направили дискретизация на златото не на 24 или 8, а на 60 клетки. Съответната дискретизация е дала и горепосочената точност.