Seminární úkol č. 3

Veronika Machová (UČO 372035), Matěj Konštacký (UČO 391150)

Prostřednictvím dat z výzkumu EU Kids Online jsme predikovali šanci, že daná osoba bude dívka. K predikci jsme použili dvě sady prediktorů. První z nich obsahovala proměnné zjišťující vlastnictví elektronických přístrojů sloužících k přístupu a k chování na internetu. Konkrétně jsme zařadili položky vlastnictví herních konzolí, laptopů a televizních setů. Vycházeli jsme z našeho přesvědčení, že chlapci hrají více her po internetu (tzv. multiplayery) a mohou se za tímto účelem připojovat z více různých přístrojů. Druhou sadu prediktorů tvořily vlastní aktivity na internetu. V návaznosti na předpoklad, že chlapci hrají více her, jsme zvolili jako první prediktor právě hraní s kamarádem přes internet. U dvou dalších prediktorů – práce s e-mailem a používání aplikací typu instant messaging – jsme očekávali vyšší zastoupení dívek, protože si myslíme, že ty častěji internet využívají jako prostředek komunikace například s kamarádkami.

Ve zkoumané skupině bylo 17 873 osob ve věku od jedenácti do šestnácti let, přičemž obě pohlaví byla zastoupena téměř stejně (chlapců bylo 8 954, dívek 8 919). Celkem 836 osob z původního souboru dat jsme z modelu logistické regrese museli vyřadit, protože neodpověděly na všechny otázky zjišťující naše proměnné (v následující tabulce četností však uvedeny jsou).

Tabulka 1. *Četnosti jednotlivých proměnných v závislosti na pohlaví dítěte*

proměnná	pohlaví	NE	ANO	celkem
TV set	*chlapci*	6 091 (65,90%)	3 149 (34,10%)	9 240
	dívky	6 476 (70,00%)	2 782 (30,00%)	9 258
	celkem	12 567 (67,90%)	5 931 (32,10%)	18 498
Laptop	*chlapci*	6 591 (71,30%)	2 657 (28,70%)	9 248
	dívky	6 087 (65,60%)	3 187 (34,40%)	9 274
	celkem	12 676 (68,40%)	5 844 (31,60%)	18 522
Konzole	*chlapci*	6 057 (65,70%)	3 168 (34,30%)	9 225
	dívky	7 659 (82,80%)	1 586 (17,20%)	9 245
	celkem	13 716 (74,30%)	4 754 (25,70%)	18 470
Hry on-line	*chlapci*	3 351 (36,20%)	5 918 (63,80%)	9 269
	dívky	6 025 (65,30%)	3 203 (34,70%)	9 228
	celkem	9 376 (50,70%)	9 121 (49,30%)	18 497
E-mail	*chlapci*	3 049 (33,00%)	6 180 (67,00%)	9 229
	dívky	2 565 (27,80%)	6 663 (72,20%)	9 228
	celkem	5 614 (30,40%)	12 843 (69,60%)	18 457
Instant messaging	*chlapci*	2 777 (30,00%)	6 484 (70,00%)	9 261
	dívky	2 334 (25,20%)	6 925 (74,80%)	9 259
	celkem	5 111 (27,60%)	13 409 (72,40%)	18 520

Jedním z předpokladů logistické regrese je diagnostika vlivu multikolinearity. Pomocí analýzy VIF jsme její velký vliv vyloučili (prediktory spolu navzájem příliš nesouvisí). Protože v našem modelu pracujeme pouze s kategoriálními dichotomickými prediktory, nelze ověřovat další předpoklad logistické regrese, a to lineární vztah mezi jednotlivými prediktory a jejich logaritmovanými transformacemi. Pomocí Durbin-Watsonova testu (1,97) jsme potvrdili nezávislost chyb měření.

Metodou ENTER jsme nejprve vytvořili model, do nějž jsme zahrnuli první tři prediktory týkající se vlastnictví přístrojů. Test tohoto modelu se ukázal být statisticky signifikantní, χ^2 (3) = 865,91, p < 0,01. S jeho pomocí se nám povedlo správně predikovat pohlaví u 34,2 % chlapců a u 82,9 % dívek (celkově u 58,5 % dětí).

Tabulka 2. *Výsledky regresního modelu při zapojení tří prediktorů týkajících se vlastnictví elektronických přístrojů*

prediktor	*B (SE)*	*Wald χ2 (df* = 1)	*p*	95 % interval spolehlivosti		
				spodní hranice	poměr šancí	horní hranice
KROK 1						
TV set	0,25 (0,04)	45,14	< 0,01	1,20	1,29	1,38
laptop	0,38 (0,03)	127,72	< 0,01	1,37	1,46	1,56
konzole	-1,10 (0,04)	716,37	< 0,01	0,31	0,33	0,36
konstanta	0,07 (0,02)	11,94	< 0,01		1,08	

Poznámka. R^2 = 0,05 (Cox & Snell), 0,06 (Nagelkerke).

Poté jsme do regresního modelu vložili i zbývající prediktory týkající se aktivit na internetu. Model byl statisticky signifikantní, χ^2 (6) = 2 440,47, p < 0,01. Tento model správně predikoval pohlaví u 68,80 % chlapců a u 62,90 % dívek. Celková úspěšnost modelu byla 65,90 %.

Tabulka 3. *Výsledky regresního modelu při zapojení všech šesti prediktorů*

prediktor	*B (SE)*	*Wald χ2 (df* = 1*)*	*p*	95 % interval spolehlivosti		
				spodní hranice	poměr šancí	horní hranice
KROK 2						
TV set	0,34 (0,04)	73,78	< 0,01	1,30	1,40	1,50
laptop	0,32 (0,04)	80,54	< 0,01	1,28	1,37	1,47
konzole	-1,02 (0,04)	565,42	< 0,01	0,33	0,36	0,39

hry on-line	-1,25 (0,03)	1 425,54	< 0,01	0,27	0,29	0,31
e-mail	0,38 (0,04)	75,77	< 0,01	1,29	1,39	1,49
instant messaging	0,32 (0,04)	68,42	< 0,01	1,28	1,38	1,49
konstanta	0,20 (0,04)	27,75	< 0,01		1,22	

Poznámka. R^2 = 0,13 (Cox & Snell), 0,17 (Nagelkerke).

Zařazení dalších třech prediktorů do modelu zlepšilo jeho schopnost predikovat ženské pohlaví, nicméně ukazatele úspěšnosti predikce (65,90 %) a R^2 (Cox & Snell; Nagelkerke) ukazují spíše na slabší shodu modelu s daty.

Z regresního modelu vyplývá, že jestliže dítě nevlastní TV set, šance na to, že jde o dívku, se zvýší 1,40 krát; jestliže nevlastní laptop, 1,37 krát; a jestliže nemá herní konzoli, 0,36 krát. Pokud dítě na internetu nehraje hry s někým dalším, šance na to, že je to dívka, se sníží 0,29 krát; jestli dítě nepoužívá e-mail, tato šance se zvýší 1,39 krát; a jestliže dítě nepoužívá aplikace typu instant messaging, šance, že je to dívka, se zvýší 1,38 krát.

Dobrá práce, polevili jste až na konci u interpretace.