MODELOVÁNÍ VYBRANÝCH UKAZATELŮ O FINANČNÍ SITUACI DOMÁCNOSTÍ V ČESKÉ REPUBLICE

Hana Řezanková*

Úvod

Jedním z šetření, která sledují ekonomickou a sociální situaci obyvatelstva, je statistické zjišťování životních podmínek domácností. Český statistický úřad od roku 2005 zajišťuje výběrové šetření nazvané Životní podmínky, což je národní modul celoevropského zjišťování EU-SILC (European Union – Statistics on Income and Living Conditions). Jeho cílem je získat přehled o příjmech a životních podmínkách domácností v České republice. Podrobněji viz (ČSÚ, 2012a).

Uvedeným šetřením je kromě příjmů a výdajů domácností získáno velké množství dalších ukazatelů jak o domácnostech, tak o osobách. U těchto ukazatelů lze zkoumat různé souvislosti a závislosti. V minulých letech byla pozornost při analýzách věnována například finančnímu potenciálu (Bartošová a Bína, 2011) a charakteristikám bydlení (Řezanková a Löster, 2011). Vývojem příjmů na základě šetření z let 2005 a 2008 se zabývaly Bartošová a Forbelská (2011) a Bílková a Malá (2012).

Příspěvek je zaměřen na analýzu dat získaných o domácnostech v rámci šetření Životní podmínky 2010. Cílem článku je na základě vybraných vysvětlujících ukazatelů odhadnout vhodné modely, s jejichž pomocí by bylo možné vyjádřit zvolené ukazatele o finanční situaci domácností. Těmito zvolenými ukazateli jsou: možnost domácnosti dovolit si týdenní dovolenou mimo domov, možnost domácnosti dovolit si zaplatit neočekávaný výdaj 8 500 Kč¹ a hodnocení, jak domácnost vycházela s příjmy, překódované do tří kategorií. K modelování jsou využity klasifikační stromy a logistická regrese, získané výsledky jsou porovnávány z hlediska celkové úspěšnosti predikce a F-míry.

* Vysoká škola ekonomická v Praze, Fakulta informatiky a statistiky (hana.rezankova@vse.cz).
Práce na tomto článku byla podpořena grantem P202/10/0262.

1 Výše neočekávaného výdaje se do roku 2010 každoročně zvyšovala o 500 Kč. Při prvním šetření v roce 2005 byla tato částka 6 000 Kč, v roce 2010 pak 8 500 Kč.

1. Charakteristika analyzovaných ukazatelů

Datový soubor zakoupený od Českého statistického úřadu obsahuje hodnoty 170 ukazatelů zjištěných u 9 098 domácností. Popis, jak bylo šetření provedeno, je uveřejněn na webové stránce ČSÚ (2010).

Kategorie ukazatelů, jejichž hodnoty mají být v modelech vysvětlovány, byly pozměněny z důvodů přehlednější interpretace výsledků, případně dosažení kvalitnější predikce. Byl upraven jednak způsob kódování, jednak byl u některých ukazatelů vytvořen menší počet kategorií. U ukazatelů týkajících se *dovolené mimo domov* a *neočekávaného výdaje* bylo změněno pořadí kategorií (0 – *ne*, 1 – *ano*). U otázky, *jak domácnost vycházela s příjmy*, byly vytvořeny nové kategorie „s obtížemi“ (vytvořena sloučením původních kategorií „s velkými obtížemi“ a „s obtížemi“) a „snadno“ (vytvořena sloučením původních kategorií „docela snadno“, „snadno“ a „velmi snadno“). Ponechána byla kategorie „s menšími obtížemi“.

Procentní zastoupení nově uspořádaných, případně vytvořených, kategorií jsou uvedena v tabulce 1. U prvních dvou ukazatelů převažují kladné odpovědi (58,5 % a 60,6 %). Na otázku, jak domácnost vycházela s příjmy, vybralo odpovědi obsahující hodnocení „snadno“ pouze 34,3 % domácností. Nejčtenější odpověď byla „s menšími obtížemi“ (38,3 % domácností).

Tabulka 1

Procentní podíly kategorií sledovaných vysvětlovaných ukazatelů

Týdenní dovolená mimo domov		Zaplacení neočekávaného výdaje		Domácnost vycházela s příjmy	
Název kategorie	Procentní podíl	Název kategorie	Procentní podíl	Název kategorie	Procentní podíl
ne	41,5 %	ne	39,4 %	s obtížemi	27,4 %
ano	58,5 %	ano	60,6 %	s menšími obtížemi	38,3 %
				snadno	34,3 %

Vysvětlující proměnnou by mohl být ukazatel hrubých peněžních příjmů (v Kč za rok) nebo čistých příjmů, případně ukazatel odvozený. Z dostupných údajů lze vypočítat hrubý či čistý příjem na osobu nebo spotřební jednotku, lze zohlednit celkové náklady na bydlení, splátky hypotéky, půjček a úvěrů. Nicméně hodnocení, jak domácnost vycházela s příjmy a co si mohla po finanční stránce dovolit, je ovlivněno dalšími faktory, jako jsou úspory, a do jisté míry též subjektivním názorem. To je zřejmé z tabulek 2 a 3, kde jsou uvedeny minimální a maximální hodnoty a dolní a horní kvartil hrubých peněžních příjmů za celou domácnost a čistých příjmů na spotřební jednotku podle definice OECD (osoba v čele domácnosti má váhu 1, děti ve věku 0 až 13 let mají váhu 0,5 a ostatní děti a osoby pak váhu 0,7), neboť minimální hodnoty příjmů pro kladné odpovědi (resp. lepší hodnocení finanční situace) jsou podstatně nižší než maximální hodnoty příjmů pro záporné odpovědi (resp. horší hodnocení finanční situace).

Tabulka 2

Minimální a maximální hodnoty a kvartily hrubých peněžních příjmů (v Kč za rok) pro sledované kategorie ukazatelů

Týdenní dovolená mimo domov		Zaplacení neočekávaného výdaje		Domácnost vycházela s příjmy	
Název kategorie	Min. / max. dolní kvartil / horní kvartil	Název kategorie	Min. / max. dolní kvartil / horní kvartil	Název kategorie	Min. / max. dolní kvartil / horní kvartil
ne	5 518 / 1 838 231 140 847 / 364 109	ne	5 518 / 1 838 231 140 970 / 366 018	s obtížemi	5 518 / 2 405 692 141 936 / 366 246
ano	10 000 / 6 070 588 244 509 / 570 000	ano	31 365 / 6 070 588 240 000 / 567 268	s menšími obtížemi	31 365 / 3 094 376 210 600 / 483 742
				snadno	55 000 / 6 070 588 240 684 / 611 060

Tabulka 3

Minimální a maximální hodnoty a kvartily čistých peněžních příjmů na spotřební jednotku (podle definice OECD) pro sledované kategorie ukazatelů

Týdenní dovolená mimo domov		Zaplacení neočekávaného výdaje		Domácnost vycházela s příjmy	
Název kategorie	Min. / max. dolní kvartil / horní kvartil	Název kategorie	Min. / max. dolní kvartil / horní kvartil	Název kategorie	Min. / max. dolní kvartil / horní kvartil
ne	5 515 / 588 192 113 885 / 159 114	ne	5 515 / 619 764 112 800 / 162 714	s obtížemi	5 515 / 1 410 256 110 641 / 160 630
ano	10 000 / 2 838 459 141 826 / 230 397	ano	31 100 / 2 838 459 140 930 / 225 669	s menšími obtížemi	31 100 / 1 393 078 130 200 / 191 428
				snadno	45 290 / 2 838 459 145 000 / 246 131

Pokud porovnáme kvartily u prvních dvou ukazatelů, tak jak u hrubých, tak čistých příjmů na spotřební jednotku dolní kvartil ve druhé skupině (kladné odpovědi) není vyšší než horní kvartil v první skupině (záporné odpovědi), ale je vyšší než medián v první skupině. U třetího ukazatele dolní kvartil ve skupině „snadno“ není vyšší než horní kvartil ve skupině „s obtížemi“, je pouze vyšší než medián v této první skupině.

Chceme-li vytvořit model, který by mohl být využit při analýze dat z jiných průzkumů, pak je lépe zaměřit se na snadněji získatelné ukazatele, než je příjem domácnosti. Ten se často zjišťuje pouze pomocí intervalů hodnot a bývá ovlivněn neochotou respondentů odpovídat pravdivě, což vede buď k nesprávným údajům, nebo k tomu, že údaje zcela chybí. Budeme proto brát v úvahu ukazatele, které obvykle mají vliv jednak na výši příjmu, jednak na odpovědi na dotazy týkající se finanční situace domácnosti.

Zaměříme se na možné vysvětlující ukazatele týkající se typu domácností a charakteristiky osoby v jejím čele². Pro účely tohoto článku byly vybrány takové kategoriální ukazatele, u nichž byl zjištěn procentní podíl jednotlivých kategorií alespoň 10 %. Pokud byly zvažovány dva ukazatele s podobným obsahem (vzdělání osoby v čele a typ domácnosti podle vzdělání), byla dána přednost ukazateli s více kategoriemi (dostatečně zastoupenými). Dále budou do analýz zahrnuty následující kategoriální ukazatele:

- *pohlaví osoby v čele,*
- *rodinný stav osoby v čele,*
- *vzdělání osoby v čele* (překódované do 4 kategorií, viz níže),
- *druh domácnosti – typ OECD.*

Navíc bude uvažován *věk osoby v čele*, tj. kvantitativní ukazatel. Pro něj budou vytvářeny věkové kategorie – buď je tento postup součástí některých metod, nebo bude některý z navržených způsobů překódování převzat do jiné klasifikační metody. Problém u původního ukazatele je ten, že neovlivňuje příjmy lineárně (s věkem se finanční situace nejprve zlepšuje, ale od určitého věku se pak zhoršuje). Navíc ukazatel z šetření Životní podmínky je kvantitativní pouze do věku 90 let, osoby starší než 90 let jsou označeny kódem 90.

V tabulce 4 jsou uvedeny relativní četnosti (v procentech) kategorií výše uvedených ukazatelů. Pro *věk osoby v čele* jsou uvedeny pouze dvě kategorie – jde o jednu z mnoha možností, které byly navrženy v rámci modelů získaných pomocí klasifikačních stromů. Kategorie zahrnující věk nad 70 let se sice částečně překrývá s důchodovou domácností, ale ukazatel byl zařazen z důvodu komplexnějšího pohledu na problematiku. Kategorie *rodinného stavu osoby v čele* jsou uvedeny podle znění v dřívějších šetřeních; nyní kategorie zahrnují i registrované partnerství (včetně zaniklého rozhodnutím a zaniklého smrtí).

Ukazatel udávající *vzdělání osoby v čele* byl překódován z původních 10 kategorií do čtyř následujících:

- základní (zahrnuje neukončený první stupeň ZŠ, první stupeň ZŠ, druhý stupeň ZŠ),
- vyučení, střední (odpovídá pouze jedné původní kategorii nazvané „vyučení, nižší střední /bez maturity/“),
- úplné střední (zahrnuje vzdělání úplné střední s maturitou, nástavbové studium, pomaturitní kurzy a vyšší odborné vzdělání) a
- vysokoškolské (bakalářské, magisterské či inženýrské a doktorské vzdělání).

2 Podle (ČSÚ, 2012b) je osobou v čele domácnosti v úplných rodinách (manžel – manželka, druh – družka) vždy muž, bez ohledu na jeho ekonomickou aktivitu. U neúplných rodin (jen jeden rodič s dětmi) a nerodinných domácností (osoby nespojené manželstvím nebo partnerským svazkem ani vztahem rodič – dítě) je prvním hlediskem pro určení osoby v čele ekonomická aktivita a druhým výše peněžního příjmu jednotlivých členů domácnosti. Tato zásada je uplatňována také u složitějších typů hospodářských domácností (např. při společném hospodaření více úplných rodin).

Pokud jde o kategorie *domácností*, pak podle klasifikace OECD (jak je uvedeno v dokumentaci k datovému souboru) je za *plně zaměstnanou* považována buď domácnost s jedním dospělým³, který je ekonomicky aktivní (ve smyslu pracujících⁴), nebo se dvěma a více dospělými, z nichž alespoň dva jsou pracující. *Nezaměstnaná* je taková domácnost, v níž nikdo z dospělých není pracující. V *důchodecké* domácnosti jsou jen nepracující osoby ve věku 65 a starší. *Částečně zaměstnaná* je domácnost, která nevyhovuje žádným z výše uvedených kritérií, tj. např. domácnost se dvěma a více dospělými, přičemž pouze jeden z nich je pracující.

Tabulka 4**Procentní podíly kategorií sledovaných vysvětlujících ukazatelů**

Pohlaví osoby v čele		Věk osoby v čele (2 kategorie)	
Název kategorie	Procentní podíl	Název kategorie	Procentní podíl
muž	73,7%	do 70 let včetně	81%
žena	26,3%	71 let a více	19%

Tabulka 4 – pokračování**Procentní podíly kategorií sledovaných vysvětlujících ukazatelů**

Rodinný stav osoby v čele		Vzdělání osoby v čele		Druh domácnosti – typ OECD	
Název kategorie	Procentní podíl	Název kategorie	Procentní podíl	Název kategorie	Procentní podíl
svobodný(á)	12,7%	základní	11,9%	plně zaměstnaná	44,9%
ženatý, vdaná	55,6%	vyučení, střední	43,8%	nezaměstnaná	14,8%
rozvedený(á)	15,6%	úplné střední	30,9%	částečně zaměstnaná	17,2%
ovdovělý(á)	16,1%	vysokoškolské	13,4%	důchodecká	23,1%

1.1 Vztahy vysvětlujících a vysvětlovaných ukazatelů

V tabulkách 5 až 9 jsou charakterizovány vztahy tří vysvětlovaných ukazatelů a výše popsaných vysvětlujících ukazatelů. Z důvodů snadnější interpretace byly dichotomické (alternativní) ukazatele (*pohlaví osoby v čele* a *věk osoby v čele*) překódovány

3 Osoba ve věku 18–64 let nebo ve věku 15–17 let či starší 65 let, která je ekonomicky (pracovně) aktivní.

4 Za pracující byly považovány osoby, u nichž v referenčním roce převažovala pracovní činnost, tj. osoby v běžném pracovním poměru, členové produkčních družstev, osoby samostatně výdělečně činné a osoby pomáhající v rodinném podniku viz (ČSÚ, 2012b).

na hodnoty 0 (odpovídající vyššímu podílu u kategorie *ne* alespoň u jednoho z vysvětlovaných alternativních ukazatelů) a 1 (druhá alternativa). Pořadí kategorií bylo v podobném smyslu změněno u ukazatelů *rodinný stav osoby v čele* a *druh domácnosti podle OECD*.

Jsou uvedena jednak řádková procenta, podle nichž lze sledovat, jak se s měnícími se kategoriemi vysvětlujícího ukazatele mění procentní zastoupení kategorií ukazatele vysvětlovaného. Dále je pro všechny dvojice ukazatelů uvedena hodnota *Pearsonova kontingenčního koeficientu*. Podle něj lze pro dvojice ukazatelů se stejným počtem kategorií usuzovat, kde je závislost větší a kde menší.

Pro dvojice ordinálních ukazatelů je navíc uvedena hodnota asymetrického *Somersova d*, které vyjadřuje intenzitu jednostranné závislosti proměnné vysvětlované na proměnné vysvětlující. Pokud se zvyšujícími se hodnotami jedné proměnné se zvyšují také hodnoty druhé proměnné, pak je hodnota Somersova *d* kladná. Jestliže se hodnoty druhé proměnné naopak snižují, je hodnota tohoto koeficientu záporná. Somersovo *d* je uvedeno též u některých čtyřpolních tabulek (oba ukazatele mají pouze dvě kategorie). Jsou-li kategorie označeny pomocí kódů 0 a 1, pak lze sledovat vztah, zda změně z 0 na 1 u jedné proměnné odpovídá spíše také změna z 0 na 1, nebo naopak spíše změna z 1 na 0, případně jsou kombinace kategorií přibližně stejně zastoupeny.

U čtyřpolních tabulek je také uveden *poměr šancí*, který udává, kolikrát se zvýší šance, že vysvětlovaná hodnota nabude hodnoty 1 místo 0, jestliže se hodnota vysvětlující proměnné změní z 0 na 1. Z tabulky 5 tedy usuzujeme, že pokud je osobou v čele domácnosti muž, pak existuje 2,644krát větší šance, že si domácnost bude moci dovolit týdenní dovolenou mimo domov. Obdobně za stejné situace existuje 2,918krát větší šance, že si domácnost bude moci dovolit zaplatit neočekávaný výdaj ve výši 8 500 Kč.

Podrobněji o analýzách kategoriálních dat pojednává např. Agresti (2002; 2007), Hebák et al. (2007), Pecáková (2011), Řehák a Řeháková (1986) a Řezanková (2011).

Tabulka 5

Charakteristiky závislosti vysvětlovaných ukazatelů na pohlaví osoby v čele

Pohlaví osoby v čele	Týdenní dovolená mimo domov		Zaplacení neočekávaného výdaje (8 500 Kč)		Domácnost vycházela s příjmy		
	ne	ano	ne	ano	s obtížemi	s menšími obtížemi	snadno
žena	59,0 %	41,0 %	58,5 %	41,5 %	40,4 %	36,4 %	23,2 %
muž	35,2 %	64,8 %	32,6 %	67,4 %	22,8 %	39,0 %	38,2 %
<i>kontingenční koeficient</i>	0,208		0,227		0,184		
<i>Somersovo d</i>	0,238		0,259		0,225		
<i>poměr šancí</i>	2,644		2,918		x		

Tabulka 6

**Charakteristiky závislosti vysvětlovaných ukazatelů na věku osoby v čele
(dvě věkové kategorie)**

Věk osoby v čele	Týdenní dovolená mimo domov		Zaplacení neočekávaného výdaje (8 500 Kč)		Domácnost vycházela s příjmy		
	ne	ano	ne	ano	s obtížemi	s menšími obtížemi	snadno
71 let a více	58,6%	41,4%	45,1%	54,9%	28,2%	39,6%	32,2%
do 70 let včetně	37,5%	62,5%	38,1%	61,9%	27,3%	38,0%	34,8%
<i>kontingenční koeficient</i>	0,166		0,056		0,022		
<i>Somersovo d</i>	0,211		0,070		0,025		
<i>poměr šancí</i>	2,359		1,334		x		

Tabulka 7

Charakteristiky závislosti vysvětlovaných ukazatelů na vzdělání osoby v čele

Vzdělání osoby v čele	Týdenní dovolená mimo domov		Zaplacení neočekávaného výdaje (8 500 Kč)		Domácnost vycházela s příjmy		
	ne	ano	ne	ano	s obtížemi	s menšími obtížemi	snadno
základní	72,5%	27,5%	65,2%	34,8%	43,7%	34,7%	21,6%
vyučení, střední	48,1%	51,9%	44,8%	55,2%	31,6%	40,0%	28,4%
úplné střední	31,5%	68,5%	32,1%	67,9%	22,3%	39,7%	38,0%
vysokoškolské	15,5%	84,5%	15,7%	84,3%	11,2%	32,8%	56,0%
<i>kontingenční koeficient</i>	0,308		0,267		0,241		
<i>Somersovo d</i>	0,250		0,211		0,205		

Tabulka 8

Charakteristiky závislosti vysvětlovaných ukazatelů na rodinném stavu osoby v čele

Rodinný stav osoby v čele	Týdenní dovolená mimo domov		Zaplacení neočekávaného výdaje (8 500 Kč)		Domácnost vycházela s příjmy		
	ne	ano	ne	ano	s obtížemi	s menšími obtížemi	snadno
ovdovělý(á)	60,7%	39,3%	52,5%	47,5%	33,5%	38,5%	28,0%
rozvedený(á)	49,4%	50,6%	53,6%	46,4%	40,5%	34,6%	24,9%
svobodný(á)	40,8%	59,2%	48,5%	51,5%	32,8%	33,7%	33,5%
ženatý, vdaná	33,8%	66,2%	29,5%	70,5%	20,8%	40,3%	38,9%
<i>kontingenční koeficient</i>	0,201		0,222		0,18		

Tabulka 9**Charakteristiky závislosti vysvětlovaných ukazatelů na druhu domácnosti podle OECD**

Druh domácnosti – typ OECD	Týdenní dovolená mimo domov		Zaplacení neočekávaného výdaje (8 500 Kč)		Domácnost vycházela s příjmy		
	ne	ano	ne	ano	s obtížemi	s menšími obtížemi	snadno
nezaměstnaná	55,7 %	44,3 %	54,0 %	46,0 %	41,3 %	34,6 %	24,1 %
důchodecká	57,0 %	43,0 %	45,9 %	54,1 %	29,5 %	39,2 %	31,3 %
částečně zaměstnaná	41,4 %	58,6 %	40,9 %	59,1 %	31,2 %	36,8 %	32,0 %
plně zaměstnaná	28,8 %	71,2 %	30,7 %	69,3 %	20,3 %	39,6 %	40,1 %
<i>kontingenční koeficient</i>	0,247		0,175		0,172		

Z tabulek 5 až 9 lze na základě hodnot kontingenčního koeficientu usoudit, že pohlaví a rodinný stav osoby v čele mělo větší vliv na možnost zaplacení neočekávaného údaje, než na možnost dovolené mimo domov, v porovnání s ostatními sledovanými vysvětlujícími ukazateli, které měly naopak větší vliv na možnost dovolené mimo domov. Tu si podle hodnocení dílčích vztahů mohla častěji dovolit plně nebo částečně zaměstnaná domácnosti, kde byl v čele muž do 70 let s vyšším než základním vzděláním, který nebyl ovdovělý. K obdobnému závěru lze dospět v případě možnosti zaplacení neočekávaného údaje s tím rozdílem, že si ho navíc častěji mohly dovolit zaplatit domácnosti důchodecké (zřejmě z úspor, s nimiž se na tyto výdaje počítá), a spíše nemohly dovolit zaplatit domácnosti s rozvedenou osobou v čele. Nejmenší vliv na vysvětlované ukazatele má věk (vyjádřený dvěma kategoriemi), u něhož se liší převažující četnosti u kategorií vysvětlovaných proměnných pouze v případě dovolené mimo domov.

Pokud jde o hodnocení, jak domácnost vycházela s příjmy, pak podle dílčích vztahů snadno nejčastěji vycházela plně zaměstnaná domácnost, v níž osoba v čele měla vysokoškolské vzdělání. S menšími obtížemi nejčastěji vycházela domácnost částečně zaměstnaná nebo důchodecká, přičemž osobou v čele byl muž se střední úrovní podle vzdělání (vyšší než základní a nižší než vysokoškolské), u něhož navíc mohlo rozhodovat, aby nebyl rozvedený. S obtížemi nejčastěji vycházela domácnost nezaměstnaná, přičemž osobou v čele byla žena se základním vzděláním, u níž navíc mohlo rozhodovat, zda je rozvedená.

Výše uvedené závěry lze formulovat i jinak, například, že osobou v čele byla rozvedená osoba, u níž mohlo rozhodovat, zda je to žena, apod. Z toho důvodu je vhodné konstruovat komplexní modely, které by zohledňovaly vysvětlující ukazatele buď podle určité hierarchie (dané statistickými postupy), nebo současně. Odhadnuté modely lze navíc hodnotit, například podle celkové úspěšnosti predikce.

2. Odhady modelů

Na základě dostupných dat lze tedy vytvářet modely, pomocí nichž by bylo možné pro známé hodnoty vysvětlujících ukazatelů odhadovat (predikovat) neznámé hodnoty vysvětlovaných ukazatelů. To znamená, že domácnosti je možné zařazovat do skupin (charakteristické např. tím, zda si domácnost může či nemůže dovolit určitý finanční výdaj). K tomuto účelu jsou určeny různé metody. Je vhodné používat více metod, a to z důvodů rozdílných přístupů k analýzám, a tím i různých detailních rozborů analyzovaných vztahů.

2.1 Použité metody

K názornému zobrazení vztahů lze využít *klasifikační stromy*. Ty postupně vybírají vysvětlující ukazatele, které při určitém počtu kategorií (v případě nevýznamnosti určité kategorie se zkoumají různé možnosti jejího sloučení s jinou, případně se zkoumají všechny možné kombinace) vykazují nejvyšší intenzitu závislosti s vysvětlovanou proměnnou. V prvním kroku se berou v úvahu všechny objekty, dále pak je závislost zkoumána pouze ve skupinách objektů, vytvořených podle kategorií (resp. skupin kategorií) vysvětlujících ukazatelů vybraných v předchozích krocích. Na základě výsledného grafu (stromu) lze vytvořit pravidla, která odhadují zařazení objektů (zde domácností) do některé z předem známých skupin. Pořadí intenzit závislosti je posuzováno podle některého z používaných postupů, může to být například p-hodnota získaná na základě některého chí-kvadrát testu o nezávislosti (buď s použitím Pearsonovy statistiky chí-kvadrát, nebo věrohodnostního poměru). Podrobněji viz např. (Han a Kamber, 2001) nebo (Berka, 2003).

Při analýzách byla použita jednak metoda CHAID (*Chi-squared Automatic Interaction Detection*), založená na výše uvedených chí-kvadrát testech (klasifikace byly provedeny jak pomocí Pearsonovy statistiky, tak pomocí věrohodnostního poměru), jednak metoda CRT (*Classification and Regression Trees*), která je založena pouze na binárním štěpení (u vícekategoriálních vysvětlujících ukazatelů se slučují kategorie za účelem vytvoření dichotomického ukazatele), a to s cílem vytvořit skupiny co nejvíce homogenní z hlediska vysvětlovaného ukazatele.

Klasifikace pro účely tohoto článku byly v rámci metody CRT provedeny s použitím Giniho míry, která je obvykle také používána pro konstrukci Goodmanova-Kruskalova τ (tau) hodnotícího intenzitu jednostranné závislosti u nominálních proměnných. Koeficient tau je založen na principu hodnocení závislosti v analýze rozptylu, neboť je počítán jako podíl meziskupinové a celkové variability vysvětlovaného ukazatele, přičemž variabilita je vyjádřena pomocí Giniho koeficientu. Při porovnání vlivu různých vysvětlujících ukazatelů je tato celková variabilita vždy stejná, proto pro výběr vysvětlujícího ukazatele je postačující hodnota meziskupinové variability (vybírání se podle nejvyšší hodnoty).

Klasifikační stromy poskytují detailní analýzu vztahů vysvětlujících a vysvětlovaného ukazatele, názorný graf a poměrně jednoduchý návod pro odhady hodnot vysvětlovaného ukazatele. K odhadům nemusí být využity všechny ze zadaných vysvětlujících ukazatelů. Nastavením jiných způsobů ukončení větvení stromu je

možné ovlivnit úspěšnost predikce, to ovšem může mít za následek nepřehledný strom a složitá pravidla pro odhady hodnot vysvětlovaného ukazatele.

Odlišnou metodou z hlediska interpretace je *logistická regrese*, jejíž podstatou je odhad logitu, tj. přirozeného logaritmu šance, že vysvětlovaný ukazatel nabude konkrétní hodnoty ze dvou možných. Každý vysvětlující ukazatel s k kategoriemi je převeden na skupinu pomocných proměnných. Kromě výše uvedených autorů pojednávajících ve svých publikacích o analýze kategoriálních dat se na logistickou regresi zaměřuje Pecáková (2009), Řeháková (2000; 2008) a Stankovičová (2007). Při analýzách pro účely tohoto článku bylo použito vždy $k - 1$ indikátorových proměnných, přičemž jako referenční kategorie byla stanovena první. Cílem logistické regrese je získat odhady parametrů modelu, tj. hodnoty b_j . Při analýzách byla použita kroková metoda *forward* (dopředná), spočívající v postupném výběru ukazatelů v pořadí podle jejich významnosti pro model (jako kritérium byla použita věrohodnostní míra).

Součástí výstupů z programových systémů jsou transformace $\exp(b_j)$. Interpretace těchto získaných hodnot je analogická interpretaci poměru šancí. Hodnota udává, kolikrát se zvýší šance, že vysvětlovaná hodnota nabude hodnoty 1 místo 0, jestliže se hodnota vysvětlující proměnné změní z referenční na příslušnou kategorii. Tato šance se samozřejmě může také snížit. V případě hodnot menších než 1 je potřeba vyjádřit toto snížení.

Nabývá-li vysvětlovaný ukazatel více než dvou kategorií, pak se analýza provádí pomocí *multinomické logistické regrese*. Postupuje se analogicky jako v případě vícekategoriálních vysvětlujících ukazatelů, to znamená, že se jedna z kategorií stanoví jako referenční a ostatní se vůči ní porovnávají, tj. úloha se rozdělí na několik úloh binární logistické regrese (jejich počet je roven počtu kategorií sníženému o hodnotu 1). Při analýzách byla jako referenční stanovena první kategorie.

Protože jde o regresní model, je vhodné vybrat takový, aby všechny jeho parametry byly statisticky významné. Aby mohl být ukazatel zařazen do modelu, měly by být významné všechny parametry odpovídající všem indikátorovým proměnným. Není-li podmínka významnosti parametrů splněna, může pomoci slučování kategorií. Na rozdíl od klasifikačních stromů je však na analytikovi, aby různé kombinace vyzkoušel.

K porovnání úspěšnosti predikcí byla použita jednak celková úspěšnost, jednak F -míra. *Celková úspěšnost* vyjadřuje podíl počtu správně klasifikovaných domácností na celkovém počtu domácností. F -míra je harmonickým průměrem přesnosti a úplnosti, přičemž *přesnost* je podíl počtu správně klasifikovaných domácností do skupiny *ano* (resp. konkrétní kategorie u vysvětlovaného ukazatele s více než dvěma kategoriemi) na celkovém počtu domácností zařazených modelem do sledované skupiny a *úplnost* je podíl počtu správně klasifikovaných domácností do sledované skupiny na celkovém počtu domácností, pro něž vysvětlovaná proměnná nabývá hodnoty *ano* (resp. jiné sledované).

Data pro účely tohoto článku byla analyzována v programovém systému IBM SPSS Statistics, verze 20, některé výpočty (z důvodu dostupných licencí) byly realizovány v předchozí verzi 18, která byla distribuována pod názvem PASW Statistics (analýzy pomocí logistické regrese a klasifikačních stromů). F -míra byla dopočtena

na základě počtů správně a chybně zařazených domácností, které jsou součástí výstupů jednotlivých procedur.

2.2 Ukazatel možnosti dovolené mimo domov

Charakteristika modelů pro ukazatel možnosti dovolit si dovolenou mimo domov, získaných pomocí klasifikačních stromů a logistické regrese, jsou prezentovány v tabulce 10. Z důvodu přehlednosti jsou v této tabulce i dalších částech použity pro metody, ukazatele a některé kategorie následující zkratky:

- CHAID-PS – klasifikační strom CHAID s Pearsonovou statistikou chí-kvadrát,
- CHAID-LR – klasifikační strom CHAID s věrohodnostním poměrem (likelihood ratio),
- CRT-Gini – klasifikační strom CRT s Giniho mírou,
- LOGREG-LR – binární logistická regrese s využitím dopředné krokové metody s věrohodnostním poměrem,
- LOGREG-SIG – binární logistická regrese s ukazateli, jejichž všechny kategorie jsou podle Waldova testu statisticky významné (na 5% hladině významnosti),
- POHL* – pohlaví osoby v čele,
- VZD* – vzdělání osoby v čele,
- STAV* – rodinný stav osoby v čele,
- VEK* – věk osoby v čele (původní kvantitativní vysvětlující ukazatel, použitý pouze u klasifikačních stromů, které pro každý model navrhuji specifické kategorie podle jejich přínosu pro odhad kategorií vysvětlované proměnné),
- VEK_70* – věková kategorie osoby v čele (ukazatel překódovaný do dvou kategorií),
- OECD* – druh domácnosti – typ OECD,
- DOV* – možnost domácnosti dovolit si týdenní dovolenou mimo domov,
- VYD* – možnost domácnosti dovolit si zaplatit neočekávaný výdaj,
- VYCH* – hodnocení, jak domácnost vycházela s příjmy,
- ZŠ* – osoba v čele se základním vzděláním,
- SŠ* – osoba v čele buď vyučená, nebo se středním vzděláním,
- SŠM* – osoba v čele s úplným středním vzděláním,
- VŠ* – osoba v čele s vysokoškolským vzděláním.

Jak již bylo naznačeno ve výše uvedeném výčtu, pro analýzu pomocí klasifikačních stromů byl zadán ukazatel *VEK* (mohou být vytvářeny různé množiny kategorií v závislosti na modelu) a pro analýzu pomocí logistické regrese ukazatel *VEK_70*. V tabulce 10 je u každé metody uveden nejvýznamnější ukazatel (který byl vybrán do modelu jako první) a nevlivný ukazatel (u klasifikačních stromů jde o ukazatel, který byl sice do stromu zařazen, ale neměl vliv na změnu kategorie vysvětlované proměnné

odhadnuté v předchozí úrovni; v logistické regresi vyšla indikátorová proměnná odpovídající dané kategorii statisticky nevýznamná). Dále tabulka obsahuje sloupec pro nezařazený ukazatel (nebyl metodou vybrán z důvodu jeho nevýznamnosti pro model). Takový ukazatel nebyl při v případě analýzy dovolené mimo domov nalezen, ale sloupec je ponechán z důvodu porovnání s modelováním dalších ukazatelů.

Tabulka 10**Charakteristiky modelů pro možnost dovolené mimo domov**

Metoda	Nejvýznamnější ukazatel	Nezařazený ukazatel	Nevlivný ukazatel (kategorie)	Úspěšnost odpovědi ano	Celková úspěšnost
CHAID-PS CHAID-LR	VZD	–	STAV VEK	70,9%	67,3%
CRT-Gini	VZD	–	VEK	90,1%	67,9%
LOGREG-LR	VZD	–	STAV (ženatý, vdaná)	79,5%	68,2%
LOGREG-SIG	VZD	x	–	81,0%	68,2%

Z uvedené tabulky je zřejmé, že žádná z metod nevyřadila žádný z vysvětlujících ukazatelů. Metodami CHAID-PS a CHAID-LR byla navržena shodná pravidla, to znamená, že bylo dosaženo stejné úspěšnosti. Pravidla získaná pomocí klasifikačních stromů jsou uvedena v tabulce 11 (metodou CRT byly některé ukazatele zařazeny do stromu opakovaně s podrobnějším členěním podle jednotlivých kategorií na nižších úrovních).

Tabulka 11**Pravidla pro odhad kategorií ukazatele dovolené mimo domov**

Metoda	Kategorie ukazatele dovolené mimo domov	
	ano	ne
CHAID-PS CHAID-LR	<ol style="list-style-type: none"> 1. VŠ 2. SŠM, plně nebo částečně zaměstnaná domácnost 3. SŠM, důchodická nebo nezaměstnaná domácnost, muž 4. SŠ, muž, plně zaměstnaná domácnost 	<ol style="list-style-type: none"> 1. ZŠ 2. SŠ, žena 3. SŠ, muž, jiná než plně zaměstnaná domácnost 4. SŠM, důchodická nebo nezaměstnaná domácnost, žena
CRT-Gini	<ol style="list-style-type: none"> 1. VŠ nebo SŠM 2. ZŠ nebo SŠ, žena, plně zaměstnaná domácnost, jiný stav než rozvedená 3. (ZŠ nebo SŠ), muž, plně zaměstnaná domácnost SŠ 4. (ZŠ nebo SŠ), muž, plně zaměstnaná domácnost, ZŠ, ženatý nebo ovdovělý 	<ol style="list-style-type: none"> 1. ZŠ nebo SŠ, žena, plně zaměstnaná domácnost, rozvedená 2. ZŠ nebo SŠ, žena, jiná než plně zaměstnaná domácnost 3. (ZŠ nebo SŠ), muž, plně zaměstnaná domácnost, ZŠ, svobodný nebo rozvedený

Výsledný model z *logistické regrese* typu SIG byl získán překódováním ukazatele *STAV* do ukazatele *STAV_ZVD* se dvěma kategoriemi (1 – *ženatý, vdaná*, 0 – *ostatní*) a ukazatele *OECD* do ukazatele *OECD_PZ* se dvěma kategoriemi (1 – *plně zaměstnaná*, 0 – *ostatní*). Odhadnutý model lze zapsat ve tvaru (pořadí ukazatelů odpovídá pořadí jejich významnosti pro model)

$$\ln \frac{\hat{\pi}(DOV=1)}{1-\hat{\pi}(DOV=1)} = -1,586 + 0,568 \cdot VZD(S\check{S}) + 1,343 \cdot VZD(S\check{S}M) + 2,17 \cdot VZD(V\check{S}) + \\ + 0,645 \cdot OECD_PZ + 0,571 \cdot POHL + 0,255 \cdot STAV_ZVD + 0,23 \cdot VEK_70,$$

kde $\hat{\pi}$ znamená odhad pravděpodobnosti (že ukazatel *DOV* nabude hodnoty 1). To znamená, že pokud není domácnost plně zaměstnaná a v jejím čele je žena se základním vzděláním, která starší než 70 let a není vdaná, je výsledná hodnota rovna konstantě, což je $-1,586$. Odhad pravděpodobnosti, že si tato domácnost bude moci dovolit dovolenou mimo domov, je $e^{-1,586}/(1 + e^{-1,586}) = 0,17$. Naopak pro plně zaměstnanou domácnost, v jejímž čele je ženatý muž do 70 let s vysokoškolským vzděláním, je tato pravděpodobnost $e^{-1,586+2,17+0,645+0,571+0,255+0,23}/(1 + e^{-1,586+2,17+0,645+0,571+0,255+0,23}) \cong 0,91$, tj. téměř 91 %.

Ve všech modelech je tedy nejvýznamnějším ukazatelem vzdělání osoby v čele, následovaný buď ukazatelem, zda je domácnost plně zaměstnaná, nebo ukazatelem pohlaví.

2.3 Ukazatel možnosti zaplatit neočekávaný výdaj

Charakteristiky modelů pro ukazatel možnosti zaplatit neočekávaný výdaj, získaných pomocí klasifikačních stromů a logistické regrese, jsou prezentovány v tabulce 12. Metodami CHAID-PS a CHAID-LR byla navržena shodná pravidla, to znamená, že bylo dosaženo stejné úspěšnosti. Pravidla získaná pomocí klasifikačních stromů jsou uvedena v tabulce 13 (pokud byl v případě metody CRT zařazen ukazatel do stromu opakovaně, je uvedena pouze kategorie na nejnižší úrovni členění).

Tabulka 12

Charakteristiky modelů pro možnost zaplatit neočekávaný výdaj

Metoda	Nejvýznamnější ukazatel	Nezařazený ukazatel	Nevlivný ukazatel (kategorie)	Úspěšnost odpovědi ano	Celková úspěšnost
CHAID-PS CHAID-LR	VZD	–	–	84,6%	68,2%
CRT-Gini	POHL	–	–	84,2%	69,0%
LOGREG-LR	VZD	VEK_70	STAV (ženatý, vdaná)	86,7%	68,6%
LOGREG-SIG	VZD	x	–	86,0%	68,4%

Tabulka 13**Pravidla pro odhad kategorií ukazatele možnosti zaplatit neočekávaný výdaj**

Metoda	Kategorie ukazatele možnosti zaplatit neočekávaný výdaj	
	ano	ne
CHAID-PS CHAID-LR	1. VŠ 2. SŠM, muž 3. SŠM, žena, jiná než nezaměstnaná domácnost 4. SŠ, muž, starší než 38 let 5. ZŠ, ženatý/vdaná, jiná než nezaměstnaná domácnost	1. ZŠ, jiný stav než ženatý/vdaná 2. ZŠ, ženatý/vdaná, nezaměstnaná domácnost 3. SŠ, žena 4. SŠ, muž, do 38 let 5. SŠM, žena, nezaměstnaná domácnost
CRT-Gini	1. muž, SŠM nebo VŠ 2. muž, SŠ, svobodný nebo rozvedený, plně zaměstnaná domácnost 3. muž, ženatý nebo ovdovělý, nad 48 let, SŠ 4. muž, ZŠ nebo SŠ, ženatý nebo ovdovělý, do 48 let, plně zaměstnaná domácnost 5. žena, SŠM nebo VŠ, jiná než nezaměstnaná domácnost	1. žena, ZŠ nebo SŠ 2. žena, SŠM nebo VŠ, nezaměstnaná domácnost 3. muž, ZŠ nebo SŠ, svobodný nebo rozvedený, jiná než plně zaměstnaná domácnost 4. muž, ženatý nebo ovdovělý, nad 48 let, ZŠ 5. muž, ZŠ nebo SŠ, ženatý nebo ovdovělý, do 48 let, jiná než plně zaměstnaná domácnost

Do výsledného modelu z *logistické regrese* typu SIG byly zařazeny ukazatele *STAV_ZVD* a *OECD_PZ*, stejně jako při modelování ukazatele dovolené mimo domov. Odhadnutý model lze zapsat ve tvaru (pořadí ukazatelů odpovídá pořadí jejich významnosti pro model)

$$\ln \frac{\hat{\pi}(VYD=1)}{1-\hat{\pi}(VYD=1)} = -0,937 + 0,427 \cdot VZD(SS) + 1,087 \cdot VZD(SŠM) + 1,919 \cdot VZD(VŠ) + \\ + 0,601 \cdot POHL + 0,56 \cdot STAV_ZVD + 0,498 \cdot OECD_PZ - 0,396 \cdot VEK_70.$$

Na rozdíl od možnosti dovolit si dovolenou mimo domov je u možnosti dovolit si neočekávaný výdaj v uvedeném modelu zřejmý opačný vliv u věkových kategorií – pro hodnotu *ano* svědčí kategorie domácností s osobou v čele starší než 70 let.

V modelech je nejvýznamnějším ukazatelem buď vzdělání osoby v čele, nebo její pohlaví. V pravidlech vytvořených pomocí klasifikačních stromů jsou využity všechny ukazatele. Pokud jde o logistickou regresi, tak při zařazení překódovaných ukazatelů pro typ domácnosti a rodinného stavu byly využity všechny ostatní ukazatele.

2.4 Ukazatel hodnotící úroveň vycházení s příjmy

V tabulce 14 jsou prezentovány charakteristiky modelů pro hodnocení úrovně vycházení s příjmy, přičemž modely byly získány pomocí klasifikačních stromů a logistické regrese. V tomto případě pro multinomickou logistickou regresi (NOMREG) je uveden pouze model, v němž jsou všechny ukazatele statisticky významné. Pravidla získaná pomocí klasifikačních stromů CHAID-PS a CHAID-LR jsou uvedena v tabulce 15.

Tabulka 14

Charakteristiky modelů pro úroveň vycházení s příjmy

Metoda	Nejvýznamnější ukazatel	Nezařazený ukazatel	Úspěšnost odpovědi snadno	Úspěšnost odpovědi s menšími obtížemi	Celková úspěšnost
CHAID-PS	VZD	–	42,2 %	53,7 %	45,2 %
CHAID-LR	VZD	STAV	42,2 %	60,4 %	45,0 %
CRT-Gini	VZD	–	44,6 %	59,3 %	46,0 %
NOMREG	VZD	–	42,5 %	50,5 %	45,1 %

Tabulka 15

Pravidla pro odhad kategorií ukazatele úrovně vycházení s příjmy

Metoda	Kategorie ukazatele úrovně vycházení s příjmy		
	snadno	s menšími obtížemi	s obtížemi
CHAID-PS	1. VŠ 2. SŠM, muž, plně zaměstnaná nebo důchodická domácnost	1. SŠM, muž, částečně zaměstnaná nebo nezaměstnaná domácnost 2. SŠM, žena, jiná než nezaměstnaná domácnost 3. SŠ, muž, ženatý nebo ovdovělý 4. ZŠ, starší než 60 let	1. ZŠ, do 60 let 2. SŠ, žena 3. SŠ, muž, svobodný nebo rozvedený 4. SŠM, žena, nezaměstnaná domácnost
CHAID-LR	1. VŠ 2. SŠM, muž, plně zaměstnaná nebo důchodická domácnost	1. SŠM, muž, částečně zaměstnaná nebo nezaměstnaná domácnost 2. SŠM, žena, jiná než nezaměstnaná domácnost 3. SŠ, muž 4. ZŠ, starší než 60 let	1. ZŠ, do 60 let 2. SŠ, žena 3. SŠM, žena, nezaměstnaná domácnost

Protože výsledkem aplikace klasifikační stromu CRT je poměrně hodně pravidel, budou uvedena ve výčtu mimo tabulku.

Pravidla pro kategorii *snadno*:

1. VŠ, plně zaměstnaná domácnost,
2. VŠ, jiná než plně zaměstnaná domácnost, jiný stav než rozvedený(á),
3. SŠM, muž, plně zaměstnaná nebo důchodecká domácnost,
4. SŠM, muž, částečně zaměstnaná nebo nezaměstnaná domácnost, starší než 60 let.

Pravidla pro kategorii *s menšími obtížemi*:

1. VŠ, jiná než plně zaměstnaná domácnost, rozvedený(á),
2. SŠM, muž, částečně zaměstnaná nebo nezaměstnaná domácnost, do 60 let,
3. SŠM, žena, plně zaměstnaná nebo důchodecká domácnost,
4. SŠM, žena, částečně zaměstnaná nebo nezaměstnaná domácnost, starší než 56 let,
5. ZŠ nebo SŠ, ženatý/vdaná, plně zaměstnaná nebo důchodecká domácnost,
6. ZŠ nebo SŠ, ženatý/vdaná, částečně zaměstnaná nebo nezaměstnaná domácnost, nad 54 let,
7. ZŠ nebo SŠ, jiný stav než ženatý/vdaná, plně zaměstnaná nebo důchodecká domácnost, muž,
8. ZŠ nebo SŠ, plně zaměstnaná nebo důchodecká domácnost, žena, svobodná nebo ovdovělá,
9. ZŠ nebo SŠ, částečně zaměstnaná nebo nezaměstnaná, nad 59 let, muž.

Pravidla pro kategorii *s obtížemi*:

1. ZŠ nebo SŠ, plně zaměstnaná nebo důchodecká domácnost, žena, rozvedená,
2. ZŠ nebo SŠ, částečně zaměstnaná nebo nezaměstnaná, do 59 let,
3. ZŠ nebo SŠ, částečně zaměstnaná nebo nezaměstnaná, nad 59 let, žena,
4. ZŠ nebo SŠ, ženatý/vdaná, částečně zaměstnaná nebo nezaměstnaná domácnost, do 54 let.

Do analýzy pomocí *multinomické logistické regrese* byly zařazeny ukazatele *STAV_ZVD* a *OECD_PZ*, stejně jako při modelování předchozích ukazatelů. Vzhledem k tomu, že na rozdíl od binární logistické regrese je jako referenční kategorie brána automaticky poslední, bylo u vysvětlujících ukazatelů změněno pořadí kategorií. Odhadnuté modely jsou dva, přičemž v prvním je vyjádřen logaritmus šance, že domácnost bude vycházet s příjmy s menšími obtížemi, v porovnání se situací, kdy domácnost vychází s příjmy s obtížemi. Druhý model vyjadřuje logaritmus šance, že

domácnost bude vycházet s příjmy snadno, v porovnání se situací, kdy domácnost vychází s příjmy s obtížemi. Získané modely lze zapsat ve tvaru

$$\ln \frac{\hat{\pi}_2}{\hat{\pi}_1} = -0,234 + 0,212 \cdot VZD(SS) + 0,623 \cdot VZD(SSM) + 1,045 \cdot VZD(VŠ) + \\ + 0,531 \cdot OECD_PZ + 0,49 \cdot STAV_ZVD + 0,263 \cdot POHL - 0,543 \cdot VEK_70,$$

$$\ln \frac{\hat{\pi}_3}{\hat{\pi}_1} = -0,887 + 0,196 \cdot VZD(SS) + 0,935 \cdot VZD(SSM) + 1,907 \cdot VZD(VŠ) + \\ + 0,685 \cdot OECD_PZ + 0,429 \cdot STAV_ZVD + 0,72 \cdot POHL - 0,659 \cdot VEK_70,$$

kde $\hat{\pi}_k$ je odhad pravděpodobnosti, že ukazatel *domácnost vycházela s příjmy* nabude k -té kategorie. Pro oba modely je nejvýznamnějším ukazatelem vzdělání osoby v čele (z hlediska porovnání kategorií *VŠ* a *ZŠ*, příp. *SŠM* a *ZŠ* ve druhém modelu). V prvním modelu za ním následuje typ domácnosti, rodinný stav a pohlaví, ve druhém modelu to je pohlaví a věkové kategorie a teprve pak typ domácnosti a rodinný stav. V obou modelech výslednou hodnotu logitu snižuje kategorie domácností s osobou v čele do 70 let. Příznivější pro model je tedy kategorie nad 70 let, což je zřejmě částečně ovlivněno větší skromností seniorů.

2.5 Hodnocení získaných modelů

V odhadnutých modelech má největší vliv na hodnotu vysvětlovaného ukazatele vzdělání osoby v čele, s výjimkou jediného případu, kdy tímto nejvýznamnějším ukazatelem bylo pohlaví. V modelech získaných s využitím logistické regrese byly při zařazení překódovaných ukazatelů pro typ domácnosti a rodinného stavu využity všechny ostatní ukazatele.

Podle předpokladu se problematickým vysvětlujícím ukazatelem ukázal věk, který měl spíše podružný vliv při podrobnějším členění v případě klasifikačních stromů. V některých případech nebyl pro konečnou klasifikaci vůbec použit, v krokové logistické regresi byl v jednom případě ukazatel věkových kategorií z modelu vyrazen. V modelech byly věkové kategorie ponechány z důvodu, že přece jen vykazují určitý vliv na hodnotu celkové úspěšnosti modelu.

Hodnocení modelů podle F-míry je uvedeno v tabulce 16. V případě modelů s binárním vysvětlovaným ukazatelem je F-míra vyjádřena vzhledem ke kategorii *ano* a navíc je uvedena i celková úspěšnost. U těchto modelů byly nejvyšší hodnoty zjištěny buď u CRT stromu, nebo u logistické regrese.

Tabulka 16**Hodnocení úspěšnosti predikce modelů podle celkové úspěšnosti a F-míry**

Název metody	Týdenní dovolená mimo domov		Zaplacení neočekávaného výdaje (8 500 Kč)		Domácnost vycházela s příjmy (F-míra)		
	úspěšnost	F-míra	úspěšnost	F-míra	s obtížemi	s menšími obtížemi	snadno
Strom CHAID-PS	0,673	0,717	0,682	0,763	0,414	0,469	0,457
Strom CHAID-LR	0,673	0,717	0,682	0,763	0,355	0,487	0,457
Strom CRT-Gini	0,679	0,767	0,690	0,767	0,381	0,487	0,471
Logistická regrese	0,682	0,749	0,684	0,767	0,439	0,457	0,454

Závěr

Je zřejmé, že na základě šetření Životní podmínky jsou získávány velmi cenné údaje. Ještě cennější jsou však výsledky zkoumání různých vztahů a souvislostí prováděných na jejich základě. Při modelování vztahů za účelem možnosti predikce hodnot sledovaných ukazatelů o finanční situaci domácností byly použity tři typy klasifikačních stromů a logistická regrese. Přestože klasifikační stromy poskytují pravidla založená pouze na přítomnosti určitých kategorií sledovaných ukazatelů, úspěšnost predikce je srovnatelná s úspěšností v případě logistické regrese, v některých případech může být i vyšší. Přítomnost kategorií však není nijak ohodnocena. Pomocí logistické regrese získáme konkrétní ohodnocení vztahů mezi vysvětlovanou a vysvětlujícími veličinami.

Odpovědi na otázky týkající se finanční situace domácnosti byly ze zkoumaných ukazatelů ovlivněny především vzděláním osoby v čele a dále jejím pohlavím a typem domácnosti podle klasifikace OECD. Nejlépe svou finanční situaci hodnotí domácnosti s vysokoškolsky vzdělanou osobou v čele. Vyšší šance mají také plně zaměstnané domácnosti, v jejichž čele je muž, nejlépe ženatý.

V článku nebyly aplikovány všechny existující klasifikační metody. Některými metodami, např. neuronovými sítěmi, lze dosáhnout vyšší úspěšnosti při predikci, ovšem bez objasnění závislostí. K charakteristice vztahů mezi analyzovanými ukazateli je třeba použít například postupy prezentované v tomto příspěvku.

Literatura

- AGRESTI, A. *Categorical Data Analysis. Second Edition*. Hoboken : John Wiley & Sons, 2002. ISBN 0-471-36093-7.
- AGRESTI, A. *An Introduction to Categorical Data Analysis. Second Edition*. Hoboken : John Wiley & Sons, 2007. ISBN 978-0-471-22618-5.
- BARTOŠOVÁ, J.; BÍNA, V. Dependence of expenditures of the Czech households on financial power. *Aplimat*. 2011, vol. 4, no. 3, s. 187–197.
- BARTOŠOVÁ, J.; FORBELSKÁ, M. Differentiation and dynamics of household incomes in the Czech Republic: EU-SILC survey in the years 2005–2008. *Aplimat*. 2011, vol. 4, no. 3, s. 198–208.
- BERKA, P. *Dobývání znalostí z databází*. Praha : Academia, 2003. ISBN 80-200-1062-9.

- BÍLKOVÁ, D; MALÁ, I. Application of the L-moment method when modelling the income distribution in the Czech Republic. *Austrian Journal of Statistics [online]*. 2012, vol. 41, no. 2, s. 125–132 [cit. 2012-08-15]. www.stat.tugraz.at/AJS/ausg122/122Bilkova1.pdf.
- ČSÚ. *Životní podmínky 2010* [online]. Praha : ČSÚ, 2010 [cit. 2013-02-15]. www.czso.cz/csu/tz.nsf/i/zivotni_podminky_2010.
- ČSÚ. 2012a. *Životní podmínky (EU-SILC)* [online]. Praha : ČSÚ, 2012 [cit. 2013-02-15]. [www.czso.cz/csu/redakce.nsf/i/zivotni_podminky_eu_silc/\\$File/zp_silc.pdf](http://www.czso.cz/csu/redakce.nsf/i/zivotni_podminky_eu_silc/$File/zp_silc.pdf).
- ČSÚ. 2012b. *Příjmy a životní podmínky domácností 2011 – Metodické vysvětlivky* [online]. Praha : ČSÚ, 2012 [cit. 2013-02-15]. www.czso.cz/csu/2012ediciplan.nsf/p/3012-12.
- HAN, J.; KAMBER, M. *Data Mining: Concepts and Techniques*. San Francisco : Morgan Kaufmann Publishers, 2001. ISBN 1-55860-489-8.
- HEBÁK, P. a kol. *Vícerozměrné statistické metody* (3). 2. vyd. Praha : Informatorium, 2007. ISBN 978-80-7333-001-9.
- PECÁKOVÁ, I. Kategoriaální vysvětlující proměnné v lineárním modelu. *Statistika*. 2009, vol. 89, no. 1, s. 72–80.
- PECÁKOVÁ, I. *Statistika v terénních průzkumech*. 2. vyd. Praha : Professional Publishing, 2011. ISBN 978-80-7431-039-3.
- ŘEHÁK, J.; ŘEHÁKOVÁ, B. *Analýza kategorizovaných dat v sociologii*. Praha : Academia, 1986.
- ŘEHÁKOVÁ, B. Nebojte se logistické regrese. *Sociologický časopis*. 2000, vol. 36, s. 475–492.
- ŘEHÁKOVÁ, B. Kontrasty v logistické regresi. *Sociologický časopis*. 2008, vol. 44, s. 745–765.
- ŘEZANKOVÁ, H. *Analýza dat z dotazníkových šetření*. 3. dopl. vyd. Praha : Professional Publishing, 2011. ISBN 978-80-7431-062-1.
- ŘEZANKOVÁ, H.; LÖSTER, T. Analysis of the dependence of the housing characteristics on the household type in the Czech Republic. In *Aplimat 2011* [CD-ROM]. Bratislava : Slovak University of Technology, 2011, s. 1629–1636. ISBN 978-80-89313-52-5.
- STANKOVIČOVÁ, I. Logistická regresia a jej využitie v ekonomickej praxi. *Forum Statisticum Slovakum*. 2007, no. 1, s. 42–54.

MODELLING SELECTED INDICATORS OF THE FINANCIAL SITUATION OF HOUSEHOLDS IN THE CZECH REPUBLIC

Abstract: The aim of the paper is to estimate models for household classification from the point of view of their financial situation. The models are constructed on the basis of data from the Living Conditions 2010 survey. The target indicators are the possibility of a household to afford a week-long vacancy outside home, the possibility of a household to afford paying an unplanned expenditure in a certain amount, and an evaluation of how a household is economical with its income. The explanatory indicators are the gender of the head of the household (HOH), the education level of the HOH, the marital status of the HOH, the age of the HOH, and the household type according to the OECD classification. For this purpose, classification trees and logistic regression were applied. The models obtained were evaluated according to the total success rate and the F-measure. The education level of the head of the household was found to be the most important indicator for the prediction.

Keywords: Czech households, financial indicator, household classification, classification tree, binary logistic regression, multinomial logistic regression, F-measure

JEL Classification: C25, C38, D19, I39