Modelování pozitivní response marketingové kampaně

Vhodným nástrojem optimalizace marketingové kampaně je tzv. cílený marketing. Úkolem je nalézt podskupinu (segment) klientů, kteří s vysokou pravděpodobností kladně odpoví na učiněnou marketingovou nabídku. Nezbytností je v tomto případě konstrukce modelu odhadujícího pravděpodobnost pozitivní response. Osloveni jsou pak ti klienti, u nichž model predikuje pozitivní responsi vyšší než je mezní hodnota (cut-off).

Metodologií konstrukce zmíněného modelu existuje celá řada, stejně tak jako různých typů těchto modelů. Mezi nejznámější a v praxi nejpoužívanější patří model logistické regrese.

Logistická regrese

Logistická regrese je označení metody matematické statistiky zabývající se problematikou odhadu pravděpodobnosti nějakého jevu (závisle proměnné) na základě určitých známých skutečností (nezávisle proměnných), které mohou ovlivnit výskyt jevu. V našem případě bude daným jevem skutečnost, zda klient pozitivně reaguje na marketingovou kampaň. Událost, zda zkoumaný jev nastal, se modeluje pomocí náhodné veličiny, která nabývá hodnoty 0, pokud jev nenastal, tj. klient na kampaň nereaguje (negativní response), nebo 1, pokud jev nastal, tj. klient na kampaň reaguje (pozitivní response). O náhodné veličině, která nabývá dvou hodnot 0 a 1 se říká, že má alternativní rozdělení. Formálně lze psát

$$Y = \begin{cases} 1, & \textit{pozitivní response} \\ 0, & \textit{negativní response.} \end{cases}$$

Metoda logistické regrese předpokládá, že za podmínek, které určuje vektor $\mathbf{x} = (1, x_1, x_2, \ldots, x_s)$, bude náhodná veličina Y rovna 1 s pravděpodobností, jejíž závislost na $\mathbf{x}$ můžeme vyjádřit ve tvaru

$$P(Y = 1|\mathbf{x}) = \frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1 + \exp(\mathbf{x}'\boldsymbol{\beta})}.$$

Vektor $\boldsymbol{\beta}' = (\beta_0, \beta_1, \ldots, \beta_s)$ je vektorem neznámých parametrů, jehož odhadem získáme i odhad hledané pravděpodobnosti výskytu zkoumaného jevu, tj. pozitivní response na marketingovou kampaň. Vzhledem k volbě první složky vektoru $\mathbf{x}$, tj. $x_0 = 1$, určuje parametr β_0 vliv tzv. absolutního členu.

K logistickému regresnímu modelu můžeme dojít např. ze zobecněného lineárního modelu (GLM)

$$g[E(Y|\mathbf{x})] = \mathbf{x}'\boldsymbol{\beta},$$

ve kterém nějaká funkce g podmíněné střední hodnoty náhodné veličiny Y je vyjádřena jako lineární funkce vektoru regresorů $\mathbf{x} = (1, x_1, x_2, \ldots, x_s)$ s regresními koeficienty

$\boldsymbol{\beta}' = (\beta_0, \beta_1, \ldots, \beta_s)$. Pokud má náhodná veličina Y alternativní rozdělení, tedy $Y \sim A(p)$, které má, jak známo, střední hodnotu $E(Y) = p$, a jako spojovací (tzv. link) funkci ve zobecněném lineárním modelu zvolíme logit,

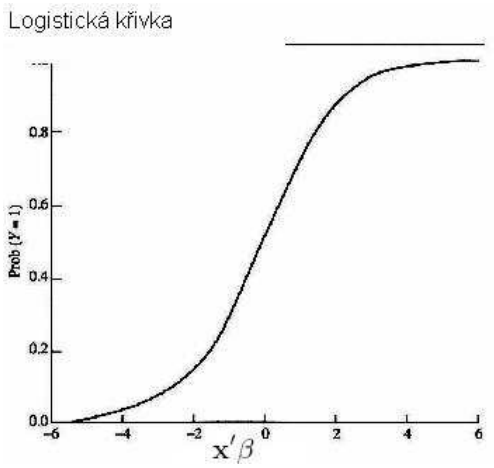
$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right),$$

dojdeme k logistickému regresnímu modelu

$$\ln\left(\frac{p}{1-p}\right) = \mathbf{x}'\boldsymbol{\beta},$$

ve kterém logit podmíněné střední hodnoty je vyjádřen jako lineární funkce regresorů.

Schéma č. 1: Logistická křivka



Parametry $\boldsymbol{\beta}' = (\beta_0, \beta_1, \ldots, \beta_s)$ lze odhadovat metodou maximální věrohodnosti. Algoritmy pro nalezení těchto odhadů, označíme je $b' = (b_0, b_1, \ldots, b_s)$, jsou již řadu let implementovány v běžně dostupných statistických programech. Mezi ty nejčastěji v praxi používané lze zařadit SAS, SPSS nebo Statistica.

Známe-li hodnoty koeficienty $b' = (b_0, b_1, \ldots, b_s)$, definuje skóre klienta vztahem

$$s = \frac{\exp(\mathbf{x}'b)}{1 + \exp(\mathbf{x}'b)}.$$

Toto skóre tedy vyjadřuje pravděpodobnost response daného klienta. Funkční předpis přiřazující vektoru $\mathbf{x} = (1, x_1, x_2, \ldots, x_s)$ hodnotu skóre s pak představuje tzv. skóringovou funkci.

Kvalita modelu

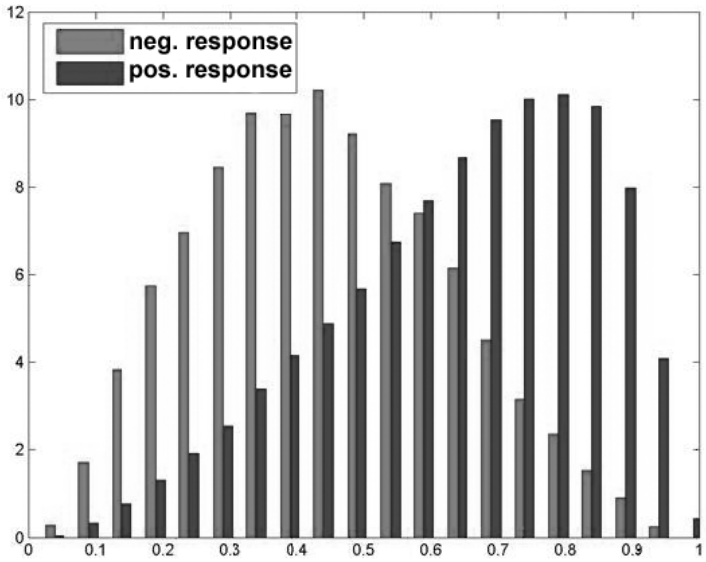
Jaké regresory, tj. jaké informace o klientovi, tvoří skóringovou funkci, zásadně ovlivňuje kvalitu této funkce. Existuje celá řada metod, jak vybrat regresory tak, aby výsledná funkce co nejefektivněji rozlišila klienty s pozitivní a negativní responsí. Mezi ty základní patří metody forward, backward a stepwise, které iterativně pracují s množinou všech daných regresorů a na základě nějakého kritéria přidávají nebo ubírají regresory z/do skóringové funkce. Obecně lze ovšem říci, že nelze vytvořit kvalitní funkci z nekvalitních regresorů. Právě množina regresorů, tj. množina údajů známých o klientovi, je tím nejzásadnějším faktorem ovlivňujícím výslednou kvalitu skóringové funkce.

Předpokládejme, že máme k dispozici skóringovou funkci, tj. ke každému klientovi jeho skóre *s*. V tomto okamžiku je třeba nějakým způsobem kvantifikovat kvalitu této funkce.

Histogram

Histogram je nejjednodušší a nejrychlejší způsob, jak posoudit kvalitu skóringové funkce. Vykreslíme-li do jednoho obrázku histogramy pozitivně a negativně respondujících klientů, a to na škále skóre, je zřejmé, že čím méně se tyto histogramy překrývají, tím kvalitnější je posuzovaná skóringová funkce.

Schéma č. 2: Histogram



Většinou je histogram definován jakožto graf zobrazující absolutní četnosti nějakých tříd. V případě, kdy chceme do jednoho grafu umístit dva histogramy, není tento způsob zobrazení vhodný. Pokud jsou rozsahy dvou datových výběrů výrazně odlišné, je nanejvýš vhodné v histogramu zobrazovat četnosti relativní.

Na schématu 2 je uveden příklad histogramu zobrazující relativní četnosti dvaceti tříd skóre.

Distribuční funkce

Distribuční funkce skóre klientů s pozitivní, resp. negativní, responsí jsou dány vztahem

$$F_{POS}(a) = P(s \le a | Y = 1),$$
$$F_{NEG}(a) = P(s \le a | Y = 0),\ a \in [0,1].$$

Jejich empirické podoby jsou dány vztahy

$$F_{n.POS}(a) = \frac{1}{n}\sum_{i=1}^{n} I(s_i \le a \wedge Y = 1),$$
$$F_{m.NEG}(a) = \frac{1}{m}\sum_{i=1}^{m} I(s_i \le a \wedge Y = 0),\ a \in [0,1],$$

kde s_i je skóre i-tého klienta, n, resp. m, je počet klientů s pozitivní, resp. negativní, responsí.

Pro úplnost ještě definujeme distribuční funkci skóre všech klientů

$$F_{ALL}(a) = P(s \le a)$$
$$F_{N.ALL}(a) = \frac{1}{N}\sum_{i=1}^{N} I(s_i \le a),\ a \in [0,1],$$

kde $N = n + m$ je počet všech klientů.

Další často používanou charakteristikou je Kolmogorovova-Smirnovova statistika (K-S nebo KS). Ta je definována jako

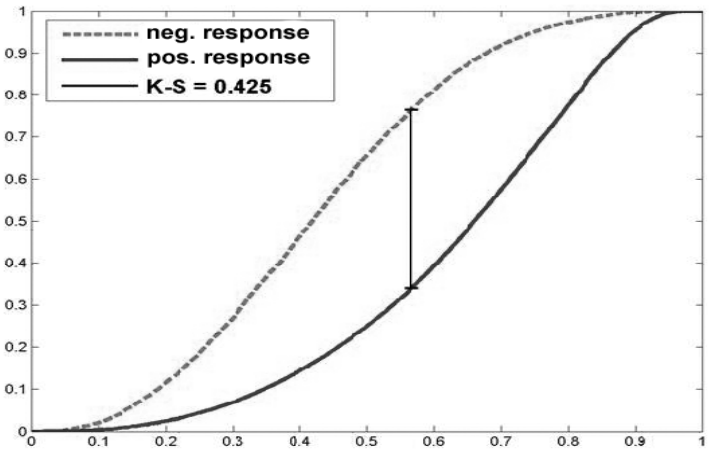
$$KS = \sup_{a \in [0,1]} |F_{NEG}(a) - F_{POS}(a)|.$$

Z definice přímo plyne, že její hodnoty leží v intervalu [0,1]. Její odhad lze snadno spočítat pomocí vztahu

$$KS = \max_{a \in [0,1]} |F_{m,NEG}(a) - F_{n,POS}(a)|.$$

Na následujícím schématu je uveden příklad empirických distribučních funkcí klientů s pozitivní a negativní responsí včetně odhadu KS statistiky. Vyčíst z něj lze například to, že skóre cca 0,57 a vyšší má v populaci přibližně 24 % negativně a 67 % pozitivně respondujících klientů.

Schéma č. 3: Empirická distribuční funkce



Lorenzova křivka, Giniho index

Pojem Lorenzovy křivky je znám již řadu let. Její vznik je motivován zobrazením rovnoměrnosti, resp. nerovnoměrnosti, příjmů dané skupiny lidí. Lorenzovu křivku lze v tomto případě obecně zapsat jako funkci $L(F)$, kde F reprezentuje horizontální osu a L osu vertikální.

Nechť je dána neklesající řada čísel $y_{i,}$, $i = 1,\ldots,n$, které reprezentují příjmy nějaké populace. Pak Lorenzova křivka je po částech spojitá lineární funkce spojující body $[F_i, L_i]$, $i = 0,\ldots,n$, kde $F_0 = 0$, $L_0 = 0$ a pro $i = 1,\ldots,n$:

$$F_i = \frac{i}{n},\ S_i = \sum_{j=1}^{i} y_j,\ L_i = \frac{S_i}{S_n}.$$

Tuto křivku lze však stejně dobře použít i pro zobrazení diskriminační síly skóringové funkce, tj. schopnosti rozpoznat klienty s pozitivní responsí od klientů s responsí negativní. Za Lorenzovu křivku v tomto případě považujeme křivku zadanou parametricky:

$$x = F_{NEG}(a)$$
$$y = F_{POS}(a),\ a \in [0,1],$$

respektive v empirické podobě

$$x = F_{m.NEG}(a)$$
$$y = F_{n.POS}(a),\ a \in [0,1].$$

Pro takto definovanou křivku například platí, že se nachází pod osou prvního kvadrantu, je konvexní a čím více se blíží bodu [1,0], tím lépe daná skóringová funkce rozlišuje mezi pozitivně a negativně respondujícími klienty.

Giniho index, nebo také Giniho koeficient, je globální míra kvality skóringové funkce. Nabývá hodnot mezi 0 a 1, přičemž ideální model, tj. skóringová funkce, má Giniho index roven 1. Na druhou stranu model, který klientovi přiřazuje skóre náhodně, má tento index rovný 0. Ideálním modelem myslíme takový model, který se 100% přesností rozpozná klienta s pozitivní responsí na marketingovou kampaň od klienta s responsí negativní. Samozřejmě jde jen o teoretický ideál a praxi se spokojíme s vědomostí, že čím je tento index vyšší tím lépe.

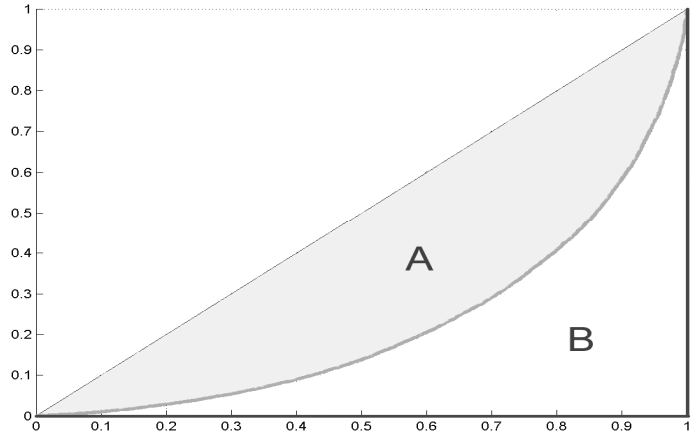
Graficky lze Giniho index definovat jakožto poměr plochy mezi Lorenzovou křivkou a osou prvního kvadrantu a plochy Lorenzovy křivky ideálního modelu a osy prvního kvadrantu. S využitím schématu 4 lze pak definovat

$$Gini = \frac{A}{A+B}.$$

Protože platí, že plocha $A+B$ je rovna $\frac{1}{2}$, můžeme psát

$$Gini = 2A.$$

Schéma č. 4: Giniho index



Vlastní výpočet Giniho indexu lze, vzhledem k předchozím označením, provést například pomocí vztahu

$$Gini = 1 - \sum_{k=2}^{n+m} \left(F_{m.NEG\,k} - F_{m.NEG\,k-1}\right) \cdot \left(F_{n.POS_k} + F_{n.POS\,k-1}\right),$$

kde $F_{m.NEG\,k}$, respektive $F_{n.POS_k}$, je k-tá hodnota vektoru empirické distribuční funkce negativně, resp. pozitivně, respondujících klientů. Více viz [80].

Lift

Dalším možným ukazatelem kvality skóringového modelu muže být graf navýšení *(cumulative lift chart)* demonstrovaný na schématu 5, jenž říká, kolikrát je při daném poměru vybraných a oslovených klientů skóringový model lepší než náhodný výběr. Formálně to lze vyjádřit vztahem:

$$Lift(a) = \frac{PosRate(a)}{PosRate} = \frac{\frac{\sum_{i=1}^{n} I(s_i \geq a \wedge Y = 1)}{\sum_{i=1}^{n} I(s_i \geq a)}}{\frac{\sum_{i=1}^{n} I(Y = 1)}{\sum_{i=1}^{n+m} I(Y = 0 \vee Y = 1)}} = \frac{\frac{\sum_{i=1}^{n} I(s_i \geq a \wedge Y = 1)}{\sum_{i=1}^{n} I(s_i \geq a)}}{\frac{n}{N}}$$

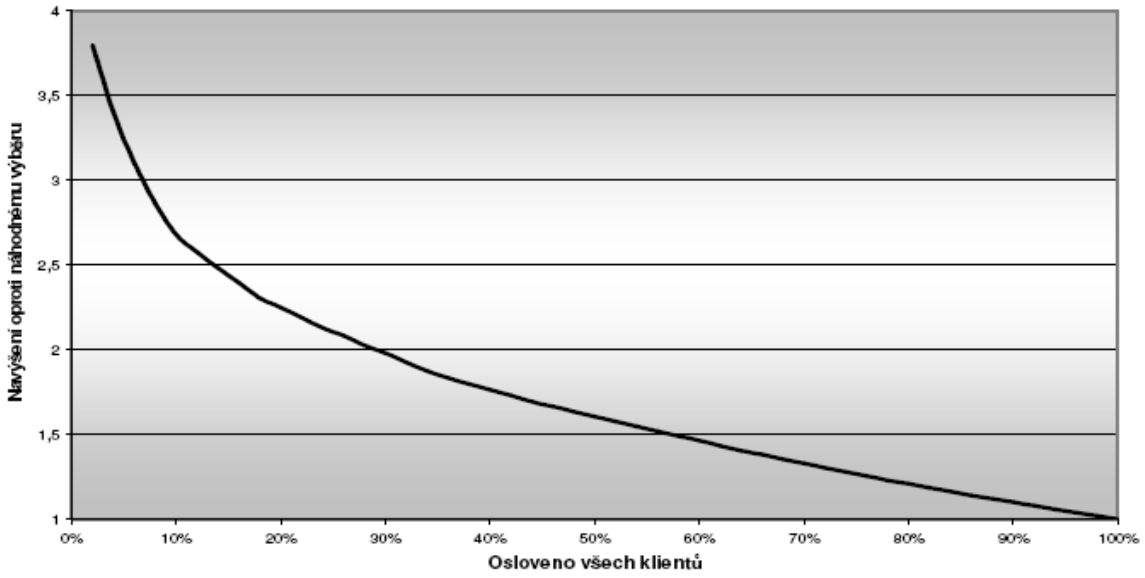
Snadno lze ověřit , že Lift můžeme ekvivalentně definovat jako

$$Lift(a) = \frac{1 - F_{POS}(a)}{1 - F_{ALL}(a)}, \; a \in [0,1],$$

tj. v empirické podobě jako

$$Lift(a) = \frac{1 - F_{n.POS}(a)}{1 - F_{N.ALL}(a)}, \; a \in [0,1].$$

Schéma č. 5: Graf navýšení (Lift)



Pramen: Data mining v praxi, Miloslav NEPIL, MU Brno, 2007.

V praxi si tento výpočet často zjednodušujeme tím, že hledáme pouze hodnoty Liftu odpovídající 10 %, 20 %, …, 100 % oslovených klientů. Demonstrujme tento postup na následujícím příkladu.

Předpokládejme, že máme k dispozici skóre 1000 klientů, z nichž 50 kladně odpoví na naši marketingovou nabídku. Poměr respondujících klientů je tedy 5 %. Seřadíme klienty podle skóre a rozdělíme do deseti skupin, tj. rozdělíme je podle decilů skóre. V každé skupině, v našem případě čítající 100 klientů, pak spočteme respondující klienty. Tím získáme jejich podíl ve skupině (Response Rate). Absolutní lift v každé skupině je pak dán poměrem podílu respondujících klientů v dané skupině ku podílu respondujících klientů celkem. Kumulativní lift je dán poměrem podílu respondujících klientů ve skupinách do dané skupiny včetně ku podílu respondujících klientů celkem. Více viz tabulka 1.

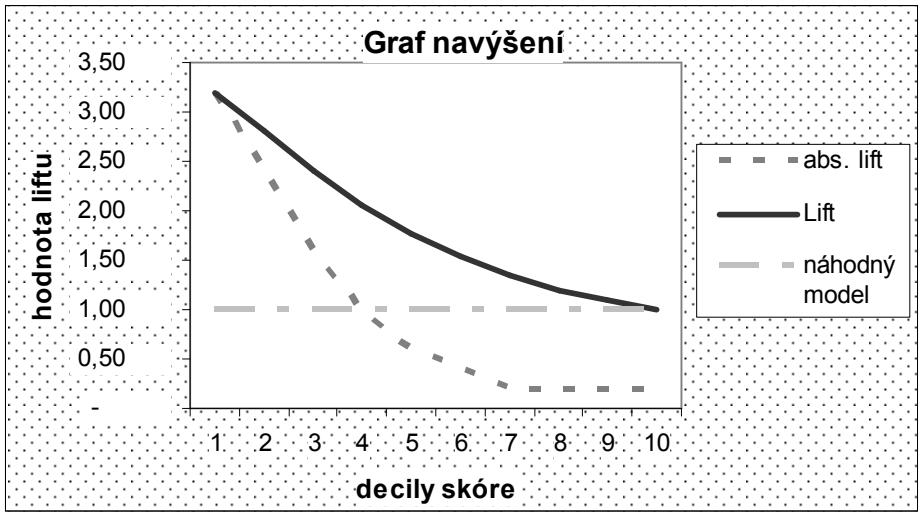
Tabulka č. 1: Absolutní a kumulativní lift

		absolutně			kumulativně		
decil	**# nabídek**	**# pozitivně resp. klientů**	**rezposní poměr**	**absolutní lift**	**# pozitivně resp. klientů**	**rezposní poměr**	**Lift**
1	100	16	16,0%	3,20	16	16,0%	3,20
2	100	12	12,0%	2,40	28	14,0%	2,80
3	100	8	8,0%	1,60	36	12,0%	2,40
4	100	5	5,0%	1,00	41	10,3%	2,05
5	100	3	3,0%	0,60	44	8,8%	1,76
6	100	2	2,0%	0,40	46	7,7%	1,53
7	100	1	1,0%	0,20	47	6,7%	1,34
8	100	1	1,0%	0,20	48	6,0%	1,20
9	100	1	1,0%	0,20	49	5,4%	1,09
10	100	1	1,0%	0,20	50	5,0%	1,00
celkem	**1000**	**50**	**5,0%**				

Pramen: Why Lift?, David S. Coppock, DM Review Online, 2002.

Schéma 6 zobrazuje získané údaje graficky. Z předchozí tabulky i z uvedeného schématu je patrné, že při oslovení například 10 % klientů s nejlepším skóre bude pozitivní response 3,2krát lepší než v případě náhodného oslovení stejného počtu klientů.

Schéma č. 6: Graf navýšení (Lift)



Pramen: Why Lift?, David S. Coppock, DM Review Online, 2002.

Finanční kvantifikace modelu

Porovnáváme-li vzájemně kvalitu několika různých modelů, srovnáváme obvykle jejich Giniho index, případně index navýšení (lift) na desátém percentilu. Obecně je velmi těžké stanovit nějakou mez, od které lze model považovat za dobrý, protože odhad response u odlišných typů marketingových kampaní muže být velmi různě obtížný.

Přesné porovnání kvality dvou modelů je sice důležité, v praxi je však mnohem důležitější vědět, jaký finanční dopad bude aplikace modelu mít. Máme-li prediktivní model hotový a známe-li pro něj funkci zisku $y=G(x)$, která pro dané procento oslovených klientů x udává procento pozitivně respondujících klientů $G(x)$, pak k výpočtu jeho finančního přínosu potřebujeme zjistit hodnotu těchto čtyř parametrů:

1. celkové náklady C na obeslání jednoho klienta,
2. čistý příjem I z jednoho respondenta (po odečtení rizikových nákladů),
3. předpokládanou apriorní responsi R, čímž máme na mysli takovou relativní responsi, jakou bychom dosáhli bez použití modelu,
4. počet klientů N, kteří potenciálně mohou být osloveni (velikost báze).

Jestliže jsme schopni tyto ctyři údaje získat, můžeme snadno sestrojit křivku profitu, která nám přesně řekne, jaké procento klientů bychom měli oslovit, abychom maximalizovali zisk z dané marketingové kampaně. Čistý zisk je totiž dán výrazem

$$\left(G(x)\cdot I\cdot R - x\cdot C\right)\cdot N,$$

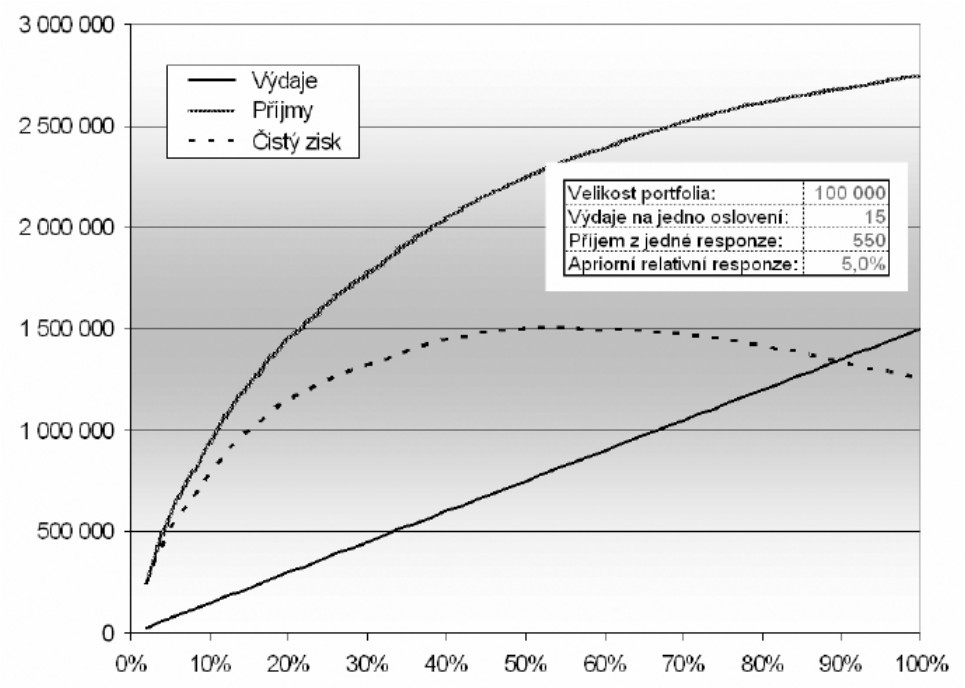
kde x značí procento oslovených klientů a $G(x)$ je výše zmíněná funkce zisku. Je zřejmé, že pro náhodný výběr platí $G(x) = x$, což znamená, že příjmy i čistý zisk rostou lineárně a zisk je nejvyšší v bodě x = 100 %.

Čtvrtý údaj N (velikost portfolia) nemá vliv na průběh křivek a tedy ani na optimální procento oslovených klientů, tato informace však slouží k tomu, aby osa y vyjadřovala srozumitelný údaj, a to čistý zisk akce v peněžních jednotkách. Porovnáním schématu 7 a 8 snadno vidíme, že uplatnění prediktivních modelů je tím nevyhnutelnější, čím máme vyšší jednotkové náklady na oslovení, nižší příjem z respondujícího klienta a nižší předpokládanou apriorní responsi. Také je patrné, že poměrně malá změna těchto parametrů má velký vliv na profitovou křivku.

Jsou-li parametry nastaveny tak, jak ukazuje schéma 7, pak je čistý zisk maximalizován při oslovení 54 % klientské báze. Ovšem schéma 8 dává i po malé změně parametrů podstatně pesimističtější výsledky: Nejvyššího zisku dosáhneme při oslovení 21 % klientů a pokud jich oslovíme více než 68 %, dostaneme se dokonce do ztráty. Samozřejmě však mohou nastat situace, kdy je zisk maximalizován až při oslovení všech klientů – v tom případě prediktivní model není zapotřebí. Lze si ovšem představit, že nás tlaky konkurenčního prostředí zavedou do takové konstelace parametrů, kdy bez použití responzního modelu k rozumnému zisku nedospějeme.

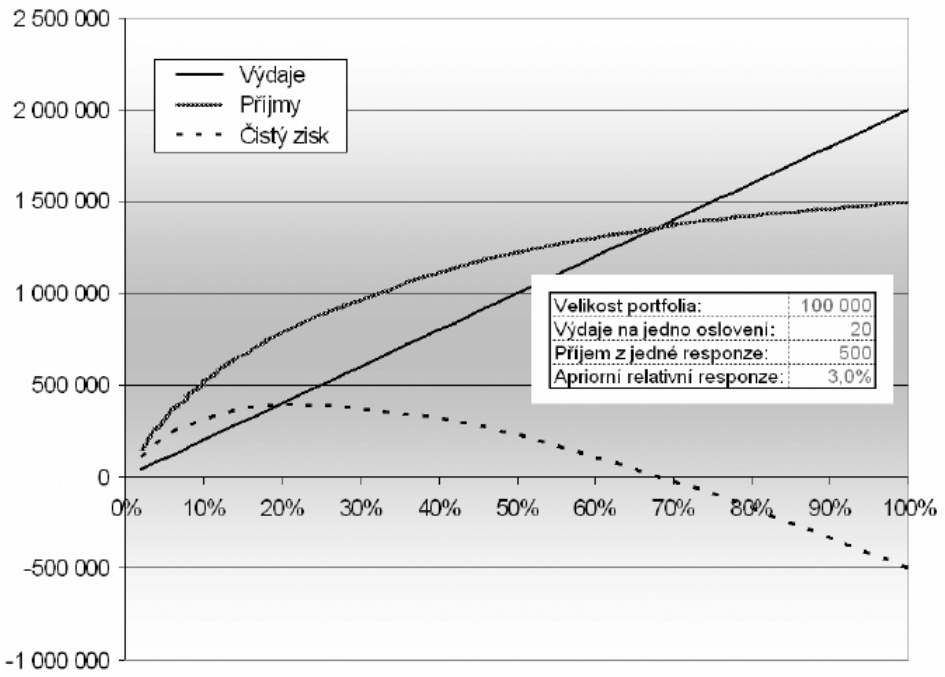
Tímto způsobem je tedy možné optimalizovat jednu marketingovou kampaň. Vyšší úrovně optimalizace ovšem docílíme tehdy, pokusíme-li se optimalizovat vytěžovací proces jako celek. Pokud máme například možnost oslovit klienty několika různými nabídkami, z nichž každá s sebou nese odlišné náklady na oslovení, jinou výši příjmu z response i rozdílnou předpokládanou apriorní responsi, stojíme před poměrně náročnou optimalizační úlohou.

Schéma č. 7: Graf čistého zisku marketingové kampaně 1



Pramen: Data mining v praxi, Miloslav NEPIL, MU Brno, 2007.

Schéma č. 8: Graf čistého zisku marketingové kampaně 2



Pramen: Data mining v praxi, Miloslav NEPIL, MU Brno, 2007.