Prediktivní analytika v podnikové informatice

Jan Pour
Fakulta informatiky a statistiky
Vysoká škola ekonomická v Praze
nám. W. Churchilla 4, 130 67 Praha 3
pour@vse.cz

Abstrakt: *Prediktivní analytika (predictive analytics) přestavuje jeden z významných trendů rozvoje podniků, podnikového řízení i IT. Tvoří součást jak komplexní disciplíny data science, tak business analytiky. Řešení prediktivních analýz představuje několik specifických úloh s úzkou vazbou na aplikace datových skladů, datových tržišť a data miningu. Aplikace prediktivních analýz se postupně rozšiřují na širokou škálu podniků a jejich oblastí řízení. Pro úspěšné řešení je účelné dobře pochopit jejich efekty i potenciální problémy, které jsou shrnuty v závěru příspěvku.*

Klíčová slova: prediktivní analytika, datová věda, dolování dat, prediktivní modely, podniková informatika, řízení podnikové informatiky, business intelligence, datové zdroje.

Abstract: *Predictive analytics represents one of the most important trends in the development of the business, management and IT. It creates the parts both disciplines as data science and business analytics. The predictive analytics solutions are strongly related to the data warehouses, data marts and data mining. Predictive analytics applications are step by step established in the big and medium enterprises and their fields of management. The successful implementation of predictive analytics requests a good understanding of their possible effects and potential problems that are summarized at the end of paper.*

Keywords: predictive analytics, data science, data mining, predictive models, business informatics, management of business informatics, business intelligence, data sources.

Prediktivní analytika (*predictive analytics*) patří k jednomu z klíčových témat dolování dat a v širším kontextu i do oblastí datové vědy (*data science*), a podnikové analytiky (*business analytics*). Úzce tak souvisí i s dnes již běžně aplikovanými koncepty, technologiemi a řešeními business Intelligence a jejími mnoha komponentami. Ukazuje se, že v případě prediktivní analýzy nejde již o teoretické konstrukce, ale v praxi využívaná řešení, byť s ohledem na jejich relativní vyšší náročnost, jde o jejich uplatňování spíše v organizacích většího rozsahu.

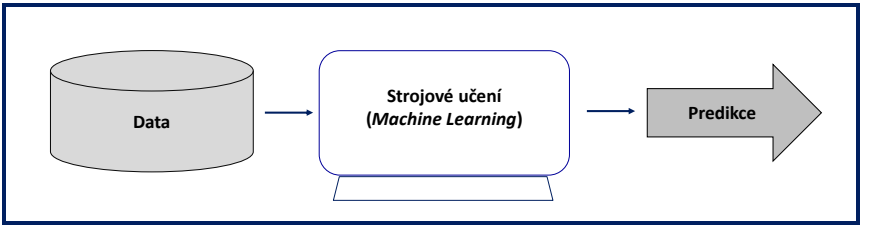
Mluvíme-li o reálném a **praktickém využití** tohoto typu úloh a aplikací, pak jeho nezbytným předpokladem je jejich pochopení především podnikovými manažery, analytiky a specialisty a prosazení do reálného života. V tomto případě nejde ani tak o pochopení jejich technických detailů, ale zejména jejich principů, efektů a oblastí řízení podniku, kde se skutečně uplatní (Siegel, 2016). Cílem tohoto příspěvku je proto alespoň rámcově ukázat tento potenciál, jeho základní principy, očekávané efekty a současně i nezbytné nároky a aktivity, které s jeho využitím jsou spojené.

Podstata prediktivní analytiky

Podstatou a smyslem prediktivní analytiky je na základě velkých objemů dostupných dat předpovídat zatím neznámé nebo nepoznané jevy, stavy nebo situace. Slouží tak k objektivizaci, zpřesňování a obecně zkvalitňování rozhodovacích aktivit v řízení podniků a institucí, a to na základě vysoké komplexity nalezených vztahů mezi zkoumanými objekty. Podstatným předpokladem pro uplatňování prediktivních analýz je existence datových skladů a tržišť obvykle obsahujících obrovské objemy dat s již provedenými analytickými a čistícími operacemi s daty na úrovni řešení business intelligence.

Výchozím momentem řešení prediktivních analýz je určení cíle predikcí (target) a ve vztahu k tomuto cíli je pak třeba navrhnout jejich řízenou segmentaci do odpovídajících skupin (*supervised segmentation*) a ze všech dostupných vybrat důležité informativní proměnné (*important informative variables*). To znamená ty, které jsou relevantní vzhledem k definovanému cíli prediktivních analýz (Provost, Fawcett, 2013). To umožňuje realizovatelnost těchto úloh, které při obrovských objemech dat by byly jinak časově i finančně mimořádně náročné.

Prediktivní modely jsou v současném pojetí postavené zejména na principech strojového učení (*machine learning*), které analyzují a učí se z dat v datovém skladu v relativně dlouhé časové řadě. Na jejich základě určují významné vztahy a důležité proměnné vztahující se k cíli predikce, a tedy i k cílové proměnné. To jsou data, která v existujících databázích již existují, a jejichž hodnoty je v tomto smyslu účelné nebo nezbytné predikovat. Tento princip v jednoduché formě vyjádřil E. Siegel (Siegel, 2016) následujícím schématem:



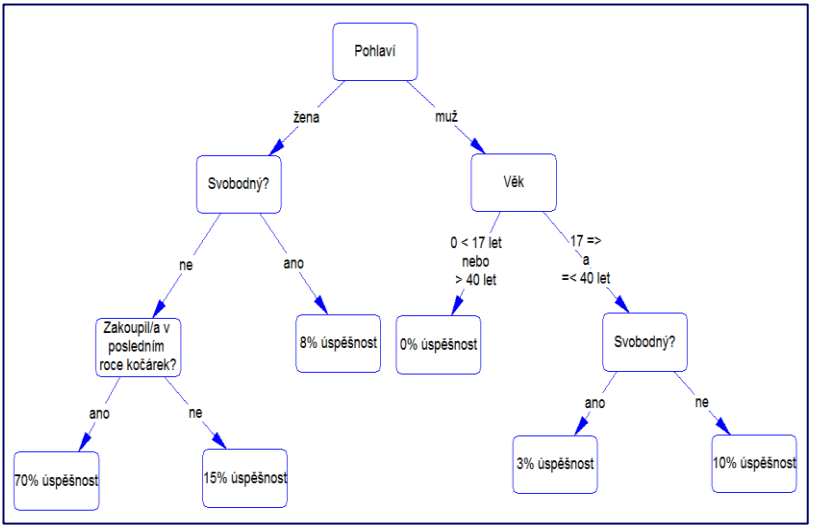
Obrázek 1: Základní princip prediktivních analýz *(Siegel, 2016)*

Komplex řešení prediktivní analýzy je založen na prediktivních modelech. Jejich podstatou je generalizace, tj. schopnost učit se v existujících datech (datového skladu nebo tržiště) hledat pouze ty charakteristiky zkoumaných jevů nebo objektů, které jsou pro realizaci predikcí významné a určit naopak ty jevy a charakteristiky, které jsou v daném kontextu nevýznamné.

V současnosti existuje množství **prediktivních modelů** a jsou založeny na principech data miningu (Berry, Linoff, 2016). Tyto modely se často skládají do větších celků, které nejlépe odpovídají definovanému cíli predikce a cílové proměnné. V prediktivní analýze se využívají např. následující typy modelů (Praus, 2013):

- **Rozhodovací stromy** (*decision trees*) - tento typ modelu je jedním z nejčastěji aplikovaných prediktivních modelů, což je dáno jeho jednoduchostí a kvalitními výsledky.

Ze vstupních proměnných vybírá ty, které jsou statisticky nejvýznamnější a vytváří pravidla, kterými segmentuje bázi dat. Vytvořená pravidla modelu lze schematicky zobrazit jako strom s kořenem a listy (Obrázek 2). Rozhodovací stromy umožňují zpětnou interpretaci a vyvození dalších závěrů. Oproti jiným modelům lze na nich zkoumat jednotlivá rozhodnutí a pravidla z nich vyplývající. Každý rozhodovací strom vychází z jednoho kořenového uzlu (root node), který představuje všechna data. Kořenový uzel je generace 0. Jeho přímí potomci (uzly) jsou generace 1 a každý další uzel obsahuje podmnožinu báze omezenou na pravidla předcházející danému uzlu. Strom je zakončen listy – uzly, které se už dále nevětví. Způsob dělení uzlů a výběr proměnných probíhá na základě statistických metod, které určují důležitost každé proměnné. Pro dělení je vždy vybrána v danou chvíli nejdůležitější proměnná.

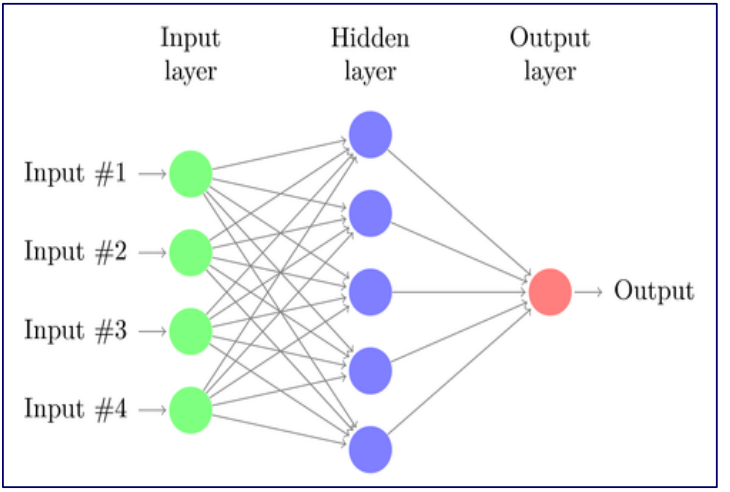


Obrázek 2: Příklad rozhodovacího stromu *(Zdroj: Praus, 2013)*

- **Neuronové sítě** (*neural networks*) – se v prediktivním modelování skládají z neuronů, které jsou navzájem propojeny a jsou schopné přijímat vstupy a odesílat výstupy. Každý neuron je aktivován, tedy produkuje výstup, pouze pokud hodnoty do něj vstupující (po vynásobení s váhami a sečtení) překročí definovanou, prahovou hodnotu. Neurony jsou složené v několika vrstvách: **vstupní vrstva** *(input layer*) – představuje proměnné, které vstupují do modelu, **vnitřní schovaná vrstva** (*hidden layer*), kde hodnoty ze vstupní vrstvy jsou transformovány a propagovány dále, **výstupní vrstva** (*output layer*) – představuje výstupní modelem predikované hodnoty cílové proměnné, resp. proměnných (Obrázek 3).

  Všechny neurony vnitřní a výstupní vrstvy jsou propojeny se všemi neurony vrstvy předchozí. Určení počtu skrytých vrstev a počtu neuronů v nich je jedním z nejdůležitějších rozhodnutí, které ovlivňuje schopnost predikce a generalizaci neuronové sítě. Rozhodnutí závisí na počtu vstupních

proměnných a vlastnostech a velikosti učících dat. Nevýhodou neuronových sítí je fakt, že produkované výstupy nejsou zpětně refaktorovatelné – není možné s určitostí říci, proč je výsledek takový, jaký je.



Zdroj: http://www.texample.net/media/tikz/examples/PDF/neural-network.pdf

Obrázek 3: Struktura jednoduché neuronové sítě s jednou skrytou vrstvou

Aktivity spojené s řešením prediktivní analýzy

S řešením prediktivních analýz je spojena řada úloh analytického, organizačního a provozního charakteru. Následující přehled obsahuje podstatné charakteristiky hlavních z nich, dle (Praus, 2013):

- **Definování cíle prediktivních analýz** – které musí být založeno na obsahu a charakteru podnikového řízení, jeho strategických cílech a aktuálních, resp. očekávaných potřebách managementu podniku. V souvislosti s definovanými cíli musí být jasně určeno, co je předmětem prediktivních analýz (resp. co má být cílovými proměnnými) a následně, jak se budou jejich výsledky aplikovat v řídících a rozhodovacích aktivitách podniku,
- **Analýza dostupných zdrojů dat** – je činnost, která je obvyklá prakticky i u všech úloh business intelligence, v případě prediktivních analýz má však svoje zvláštnosti. Dobrým předpokladem úspěchu predikcí je co nejvyšší úplnost dat a vysoká granularita, tedy podrobnost relevantních dat, které zvyšují přesnost a objektivitu výsledků získaných na bázi prediktivních modelů. To znamená, že je nezbytné analyzovat možnosti užití jak interních podnikových systémů (obvykle datového skaldu a datových tržišť), tak dat z externích systémů (např. z demografických databází, marketingových a analytických databází atp.).

  Musí se zde vyhodnotit kvalita identifikovaných datových zdrojů a jejich dostupnost, konsistence a konsolidace. V tomto kontextu je nezbytné vyhodnotit,

zda prediktivní analýza je vůbec na dostupných datech proveditelná, zda budou pro realizaci prediktivní analýzy nutné změny přímo na úrovni datových zdrojů, nebo až v rámci jejich transformací a další aspekty.

- **Organizace, integrace a čištění dat** – což je obdobně jako u většiny analytických úloh (BI a další) nejpracnější část celého projektu prediktivních analýz.

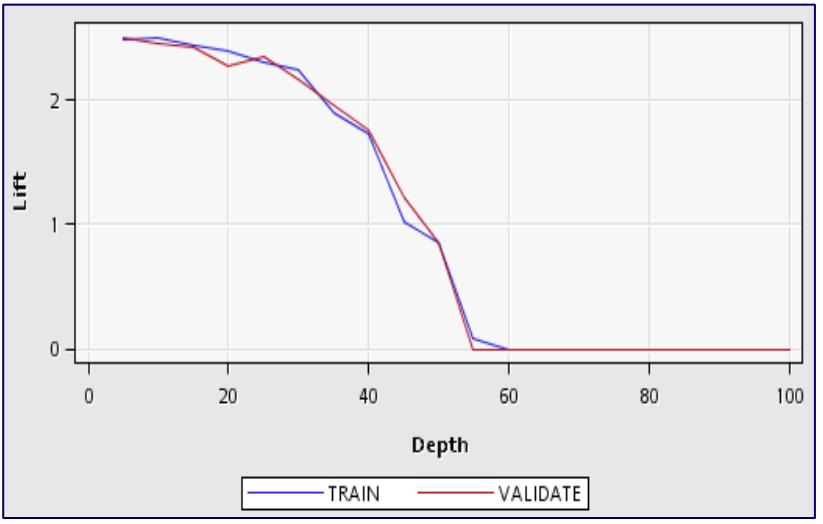
  Hlavní aktivitou je segmentace, shlukování, třídění dat podle jejich business významu a zejména podle potřeb predikce a cílových proměnných. Úspěšnost prediktivních modelů tak rozhodujícím způsobem ovlivňuje datová kvalita, resp. kvalita datového základu a datových tržišť. Součástí této úlohy je i příprava a zpracování dalších kalkulovaných hodnot jednotlivých ukazatelů, opět definovaných na základě analýz požadavků na cílové proměnné prediktivních analýz. Tyto hodnoty mohou mít v prediktivních modelech potencionálně velice vysoké váhy a je nezbytné je posuzovat zejména se business specialisty,
- **Identifikace relevantních vztahů** – realizovaných na bázi data miningu a clusterových analýz., tj. skrytých vazeb, vzorů a vztahů. Data mining je významná součást prediktivních analýz, protože data a vztahy, která identifikuje jako relevantní jsou také relevantní z pohledu očekávaných výsledků prediktivního modelu.

  Data mining v prediktivních analýzách představuje získávání znalostí o vztazích a prediktivní model je aplikací těchto znalostí. Zaznamenává všechny vztahy, které jsou v datech přítomny, bez ohledu na znalosti toho, co je zapříčinilo.

  Cluster analýza je v prediktivních analýzách využitá pro hledání podobností v datech. Pomocí algoritmů a metod seskupuje objekty podobných vlastností do skupin. Tato analýza může být využita k odhalení podstatných struktur v datech, ale neposkytuje interpretaci nebo vysvětlení proč tyto struktury existují,
- **Tvorba prediktivního modelu** - modely z dostupných dat analyzují stávající chování k posouzení pravděpodobnosti výskytu predikovaného jevu, např. že zákazník s určitými vlastnostmi zakoupí nějaký produkt. Jak již bylo uvedeno, nejdůležitější vlastností prediktivních modelů je generalizace. Ta zaručuje to, že model naučený z historických dat (in-sample) dokáže správně vyhodnotit data nová (out-of-sample), která do tvorby a učení modelu ještě nevstoupila. Používají se k tomu komplexní prediktivní modely využívající principy strojového učení, které se tvoří reltivně snadno v nástrojích, určených pro dolování dat. Vytvořené modely se v programech sami validují, optimalizují a vyhodnocují,
- **Vývoj a výběr prediktivního modelu** – modely mohou mít různý rozsah a formu v závislosti na jejich složitosti a využití, pro které jsou navrženy. Pro co nejvyšší přesnost může být použito více modelů, které jsou následně porovnávány a kombinovány.
- **Validace a ladění prediktivního modelu** - aby se zajistilo, že je naučený prediktivní model co nejpřesnější, musí být testován pomocí skupiny dat out-of-sample, testovacích dat, která nijak nevstoupila do vývoje a učení modelu. Ověřuje se schopnost predikce modelu na této skupině dat a porovnává se, jak moc se odchyluje od výsledků z učících dat. Pokud je odchylka velká, znamená to, že model není optimálně generalizován. Nabyté znalosti jsou poté aplikovány na model, který se podle nich upraví a následně opět testuje.

Využívané metriky modelu, pomocí kterých se vyhodnocuje úspěšnost modelu, jsou Lift, ROC a další:

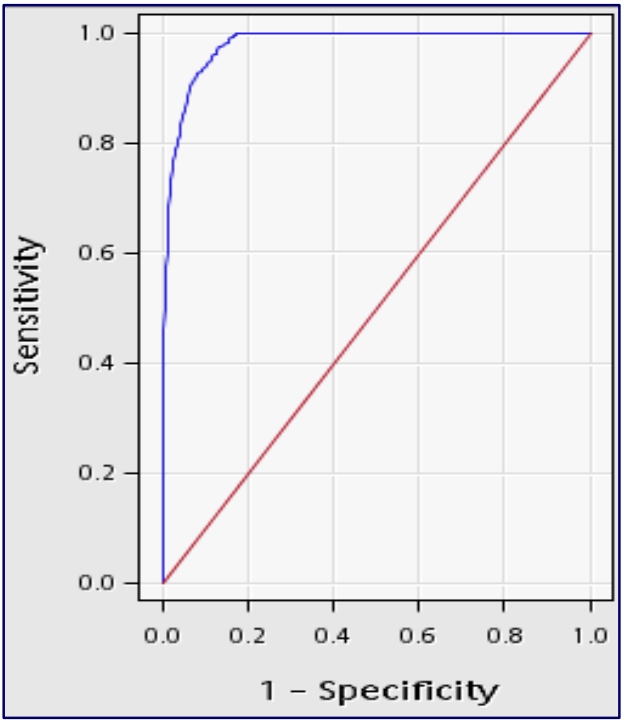
- **Lift** je ukazatel popisující úspěšnost a použitelnost prediktivního modelu. Definuje úroveň, jak moc je model úspěšnější než náhodný výběr. Úspěšnost prediktivních modelů se nad měnící se hloubkou báze dat různí. Pro usnadnění porovnání modelů nad různými hloubkami báze se zavedl pojem decil. Jeden decil obsahuje 10 % záznamů báze seřazené sestupně dle predikovaného skóre. Nejvyššího liftu modely zpravidla dosahují na začátku v prvním decilu a pak má klesající tendenci (Obrázek 4).



Obrázek 4: Graf liftu z programu SAS Eneterprise Miner

- **Graf ROC** (*Receiver operating characteristic*) je dalším typem grafického zobrazení úspěšnosti predikce modelu. Využívá se pro vyhodnocení úspěšnosti predikce binárního targetu. Graf ROC zobrazuje na ose Y relativní četnost správně klasifikovaných pozitivních případů – senzitivitu, na ose X relativní četnost správně klasifikovaných negativních případů – specificitu. Červená křivka znázorňuje náhodný výběr a modrá prediktivní model (Obrázek 5),
- **Misclassification Rate** (míra chybovosti) je poměr špatně vyhodnocených případů k celku. Výpočet hodnoty misclassification rate = počet špatně klasifikovaných (pozitivních i negativních targetů) / celkový počet klasifikací,
- **Overfitting (také overlearning)** přeučení modelu znamená, že model špatně vyhodnocuje náhodný šum v datech, určuje důležité vztahy na základě náhodných proměnných a postrádá schopnost generalizace. Úspěšnost predikce modelu na nových datech je oproti ideálnímu stavu snížena. Tento nežádoucí stav může být mimo jiné způsobený

následujícími případy - špatné nastavení modelu, například příliš velký (hluboký, rozvětvený) rozhodovací strom, příliš komplexní neuronová síť, příliš malý vzorek učících dat, nebo chyby ve vstupních proměnných, které nebyly řádně očištěny. Přeučení znamená, že model z dostupných dat předpokládá příliš mnoho,



Obrázek 5: Graf ROC z programu SAS Eneterprise Miner

- **Underfitting (nedoučení modelu)** znamená, že učení modelu bylo chybou nastavení modelu, nebo nedostatkem dat zastaveno příliš brzy a nebyly odhaleny všechny důležité vztahy. Model je příliš obecný a jednoduchý. Například nedoučený strom se skládá z příliš malého počtu pravidel a má málo listů,
- **Pruning (prořezávání)** je jedna z metod optimalizace modelů. Účelem je snížení komplexity modelu. Například u rozhodovacích stromů představuje pruning „prořezávání“ větví modelu, za účelem snížení počtu větví a listů a tím snížení rizika overfittingu, u neuronových sítí zase snížení počtu neuronů a vrstev,

- **Vyhodnocení prediktivního modelu** - po vytvoření prvních modelů se výsledky těch nejúspěšnějších testují v praxi na nových datech. Na základě predikce se uskuteční rozhodnutí a nastane odpovídající akce. Poté se vyhodnocuje, jak moc predikce odpovídá realitě, vyhodnocuje se úspěšnost modelu v praxi. Ta bývá

zpravidla řádově nižší, než model vykazuje na učících, či testovacích datech. V některých případech může být objektivní vyhodnocení problematické, protože provedená akce ovlivní chování jedince a není tak možné zjistit jeho chování, když by akce nenastala.

- **Aplikace prediktivního modelu** – představuje využití prediktivního modelu v praxi a jeho výstupů pro učinění rozhodnutí. V této fázi je model funkční a vyhodnocuje nová data. Na základě výsledků se provádějí akce a realizují rozhodnutí. Díky vyvinutému, naučenému a funkčnímu modelu se upraví procesy, zlepší rozhodnutí a model se dále validuje a vyhodnocuje se jeho úspěšnost na reálných datech.

Využití prediktivní analýzy v řízení podniků a organizací

Prediktivní analýzu momentálně využívají především **velké podniky** jako, např.

- telekomunikační společnosti využívají prediktivní analýzu pro – odhadování možných odchodů zákazníků od společnosti (churn analysis), zaměření marketingových kampaní na predikované události, nebo situace,
- velké bankovní ústavy využívající prediktivní analýzy pro hodnocení klientů, jejich spolehlivosti, churn analysis (viz výše), skórování úvěrů (*credit scoring*)
- většina finančních ústavů využívá preditkivní analýzy pro hodnocení a identifikaci trendů na finančních trzích, specifikaci nových finančních produktů a odhadování jejich úspěšnosti,
- pojišťovny aplikují prediktivní analýzy při odhalování podvodných pojistných událostí, pří upisování, při ocenění rizik a pojistného, pro zlepšení efektivity marketingových kampaní, pro tvorbu produktů,

V dalším vývoji lze ale očekávat (jako i u jiných typů aplikací) rozšíření aplikací prediktivních analýz **ve většině velkých a středních průmyslových, obchodních, dopravních a dalších typů podniků,** a to v hlavních oblastech řízení, zejména ve finčním řízení, řízení marketingu, řízení prodeje, řízení nákupu, v řízení výrby, případně v personálním řízení (Dohnal, Pour, 2016).

Závěr

Na závěr příspěvku je dobré shrnout hlavní přínosy a na druhé strany jistá omezení prediktivních analýz. **K evidentním efektům** patří:

- efekty ekonomického charakteru, tj. v pozitivních změnách klíčových ukazatelů, tj. zvýšení zisků, snížení nákladů v jejich druhovém členění i podle realizovaných činností,
- nejvýznamnější efekty predikce představují ty, které souvisejí s postavením podniku na trhu, tj. získání vyššího tržního podílu, získání konkurečních výhod v klíčových oblastech podnikání,
- obvykle dosahované efekty se váží také k marketingovým aktivitám, jako je efektivnější cílení marketingových kampaní, lepší a přesnější poznání zákazníků, jejich potřeb a očekávání,

- efekty prediktivních analýz znamejí i významná snížení finančních, obchodních, investičních a dalších rizik,
- souhrnně vede systematické uplatnění predikcí i k vyššímu zhodnocení existujících podniku.

Na druhé straně je s realizací prediktivních analýz spojena i **řada problémů** a otevřených otázek, zejména:

- existence, či neexistence zájmu vedení podniku na řešení a zejména využití prediktivních analýz v podnikovém řízení,
- nedostatky v kvalifikční přípravě uživatelů a analytiků, prediktivní analýzy znamenají specifické nároky na znalosti, zkušenosti a invenci jejích potenciálních řešitelů i uživatelů,
- rizikem je také to, že realizace prediktivních analýz je časově i finančně náročná akce a mnohdy s nejistým výsledkem.

Uvedené a případné další klady a problémy prediktivních analýz znamenají nutnost začlenění jejich vývoje i užití do systému řízení podniku i řízení IT podniku. To je zřejmě klíčový faktor jejich konečného úspěchu.

Literatura

Berry, M.J.A., Linoff, G.S., 2004: *Data Mining Techniques*, New York, John Wiley and Sons, ISBN 0-471-47064-3

Dohnal, J., Pour, J., 2016: *IT v řízení podniku*, Praha, Professional publishing, ISBN 978-80-7431-160-4

Kimball, R., Ross, M., 2010: *Relentlessly Practical Tools for Data Warehousing and Business Intelligence.* Indianopolis, John Wiley Publishing, ISBN 978-0-470-56310-6

Praus, O., 2013: *Prediktivní analýza – postup a tvorba prediktivních modelů*, VŠE, Praha

Provost, F., Fawcett, T., 2013: *Data Science for Business. What You Need to Know About Data Mining and Data-Analytic Thinking.* O´Reilly Media. Sebastopol, ISBN 978-1-449-36132-7

Siegel, E., 2016: *Predictive Analytics.* New York, John Wiley & Sons, ISBN 978-1-119-14567-7

JEL Classifications: M10, C88