VLADIMÍR LÖFFLER BARBORA ŠTĚTINOVÁ LUKÁŠ BERNAT

BIG DATA
A UMĚLÁ INTELIGENCE PRO MANAŽERY

Praktický návod, jak držet krok s dobou v 21. století

Big data a umělá inteligence pro manažery

Konverze do elektronických formátů: Ing. Milan Vilímek Jihlavský

© 2021, nakladatelství Vladimír Löffler 1. vydání

ISBN 978-80-908226-4-1 (ePub)

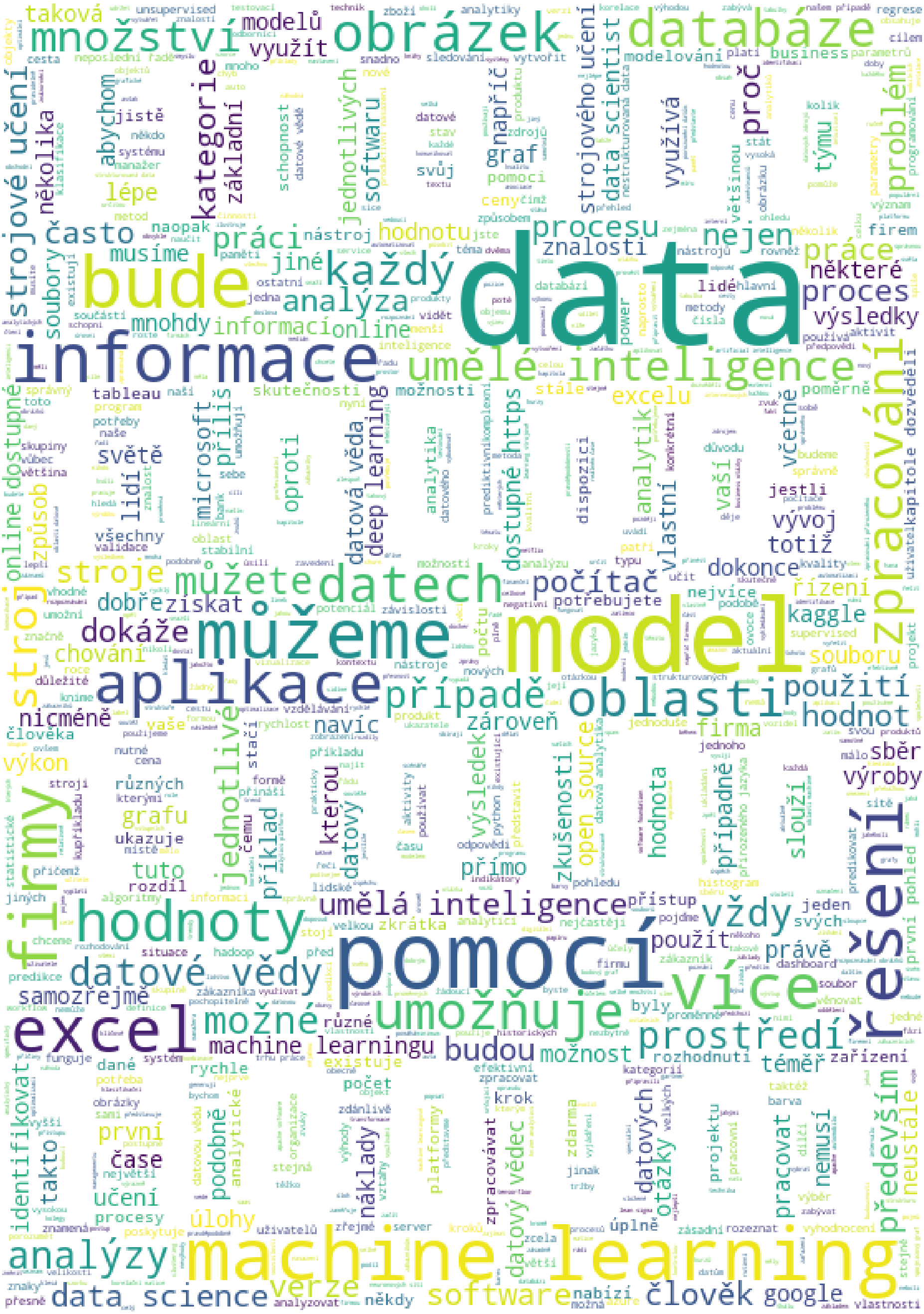
ISBN 978-80-908226-3-4 (PDF)

ISBN 978-80-908226-5-8 (MOBI)

Karlovarským krajem byla poskytnuta v roce 2020 dotace na realizaci projektu Populárně naučná kniha „Big data a umělá inteligence pro manažery“ ve výši 40 000 Kč, v rámci dotačního programu Kreativní vouchery.

Big data a umělá inteligence pro manažery

Praktický návod, jak držet krok s dobou v 21. století


data
model
pomocí
informace
můžeme
aplikace
machine learning
databáze
obrázek
množství
bude
datech
hodnoty
excel
řešení
zpracování
umělé inteligence
firmy
stroj
oblasti
případě
umožňuje
více
prostředí
datové vědy
analýzy
data science
software
člověk
google
verze
každý
proces
analýza
umělá inteligence
strojové učení
často
kategorie
základní
možné
machine
learningu
budou
možnost
vždy
můžete
počítač
dokáže
stroje
procesu
znalosti
práce
nejen
výsledky
zpracování
informací
online
mnohdy
musíme
práci
jiné
excelu
microsoft
datová věda
deep learning
dostupné
https
jednotlivé
výsledek
zkušenosti
open source
hodnota
případně
slouží
použití
hodnot
firma
výroby
sběr
předevší
neustále
identifikovat
takto
náklady
datový vědec
otázky
datových
pracovat
nemusí
úlohy
učení
první
čase
samozřejmě
dokonce
vývoj
totiž
kaggle
řízení
souboru
nicméně
navíc
zároveň
přímo
přístup
použít
právě
téměř
zařízení
rozhodnutí
dobře
získat
chování
grafu
tuto
rozdíl
výkon
světě
lidí
příliš
oproti
analytik
vlastní
dispozici
jestli
vaší
včetně
tableau
graf
týmu
problém
proč
využívá
data scientist
strojového učení
modelů
využít
taková
lépe
abychom
několika
soubory
způsob
online dostupné
různých
ukazuje
příklad
který
datový
různé
počet
krok
zkrátka
nejčastěji
byly
zákazník
existuje
rychle
dané
kategorií
datum
nabízí
některé
poměrně
dozvěděli
kapitole
uživatel
aktivit

O autorech

Ing. Barbora Štětinová MBA
Datový analytik a Data scientist (oblast automotive, telekomunikace), spoluzakladatel Elderberry data, mezinárodní instruktor pro Machine Learning a datovou analytiku na platformách Udemy, Packt Publishing, a dalších. Instruktor Knime Analytics Platform. Člen vítězného týmu soutěže Data Cup 2019 pořádaných Insiders40.

Ing. Lukáš Bernat
Data scientist a RPA specialista (oblast automotive, media), nadšenec do online vzdělávání, především z oblasti data science, doktorand Národohospodářské fakulty VŠE, člen vítězného týmu soutěže Data Cup 2019.

Ing. Vladimír Löffler
IT manažer a ERP/ BI specialista (oblast automotive), spoluzakladatel Elderberry data, instruktor na platformě Udemy.com. Autor publikace „Automatické zpracování dat pomocí Knime Analytics Platform“. Big Data and Data Science nadšenec.

Obsah

Úvod

Dostává se vám do rukou publikace, jež vznikla z rozmaru tří nadšenců, kteří se dost možná ocitli ve stejné situaci, v jaké se nacházíte právě teď vy, a zmateně se škrábali na čele, nevěda, kde začít. Ať už je vaší pohnutkou aktivně se zapojit do rozjetého technologického vlaku 21. století, nebo se zkrátka chcete dozvědět, jak to funguje, na počátku budete tápat. Internet vám sice poskytne tisíce návodů, článků i kurzů s přímou aplikací dané oblasti, ale ne kuchařku, která by vám poskytla komplexní nadhled nad takto složitým tématem.

Proto se zrodila myšlenka využít praktické i teoretické zkušenosti autorů a předat je dál. Naším cílem je ušetřit čtenáře tápání ve spleti slepých uliček, pracných a drahých chyb, a naopak dodat mu odvahu pustit se na pole umělé inteligence a datové vědy po hlavě a bez ostychu. Nalijme si čistého vína, zaspali jsme. Celá Evropa. To ovšem neznamená, že by nám vlak ujel úplně. Každý dílčí krůček k „chytrým firmám", „chytrým městům", „chytrému zdravotnictví", apod. zvyšuje konkurenceschopnost v dnešním dynamickém světě.

Pojďme nahlédnout pod pokličku. Koncept strojového učení (*machine learning*) a umělé inteligence (*artificial intelligence*) existuje již více než 50 let. Rozmach tohoto konceptu však umožnil až obrovský nárůst výpočetního výkonu současných počítačů, výsledky výzkumu v oblasti neuronových sítí a pokročilé datové analytiky a také objem dat, která jsou každou vteřinu generována stroji, lidmi a organizacemi.

Než zdlouhavě rozebírat, proč se téměř každé firmě vyplatí věnovat tématům, jako jsou *big data*, *machine learning* (strojové učení) a *artificial intelligence* (umělá inteligence, dále jen AI), posuďte prosím sami následující informace a zvažte, zda je téma atraktivní také pro vás a vaši firmu.

Informace vycházejí ze studie, kterou provedla poradenská firma McKinsey (BAUER, 2017) pro německý trh, jenž je s naším trhem intenzivně propojen:

- minimálně 30 % aktivit v 62 % německých podniků lze automatizovat (stejná čísla platí i pro trh USA).
- 2 % německých podniků mohou být kompletně automatizována.
- AI použitá v oblasti prediktivní údržby pomáhá zvýšit produktivitu výrobních zařízení až o 20 % a snížit celkové náklady na údržbu až o 10 %.
- AI umožňuje realizovat kontroly kvality výrobků (například pomocí počítačového vidění – *computer vision*) s nárůstem produktivity až o 50 % a zvýšením kvality vizuálních kontrol až o 90 % v porovnání s kontrolami prováděnými člověkem.
- Použití AI v řízení dodavatelských řetězců (*supply chain management*) dokáže zlepšit přesnost plánování o 20 až 50 %, snížit ztráty prodejů z důvodu nedostupnosti zboží až o 65 % a snížit zásoby v rámci řetězce o 20 až 50 %.
- Aplikace strojového učení při vývoji nových výrobků urychlí nejen samotný proces vývoje a uvedení výrobku na trh až o 10 %, ale redukuje i náklady na vývoj o 10–15 %.
- Automatizace podpůrných procesů pomocí AI přispívá ke zvýšení jejich efektivity a kvality, například IT service desk dovede automatizovat až 90 % aktivit.
- Ve výrobním sektoru je možné automatizovat až 55 % aktivit, které v současné době vykonávají lidé.
- Predikovatelné činnosti ve stabilním prostředí (např. svařování a balení) lze automatizovat až v 90 % případů.
- Ostatní pracovní aktivity (kromě aktivit vyžadujících kreativní lidskou činnost) mají potenciál pro automatizaci mírně přesahující 50 %.
- Největší potenciál pro automatizaci je v oblasti výroby, logistiky, ubytování a stravování, prodeje (retail) a zemědělství.

Všechna tato čísla poukazují na fakt, že aplikace AI může většině firem přinést skokové snížení nákladů, zvýšení výnosů, případně zcela nové, dříve netušené tržní příležitosti. Technologie pro strojové učení a umělou inteligenci zaznamenaly obrovský krok vpřed v řádu několika málo let a jsou

navíc dostupné (mnohdy zdarma) i dobře popsané. Zavedení takto přelomových technologií ve firmách tak brání pouze nedostatek informací a chybějící odpovědi na otázky jako:

- Co AI přinese mojí firmě?
- Je použití AI pro moji firmu vhodné?
- Kde a jak mám se zavedením AI začít?
- Co k zavedení AI potřebuji (technologie, lidé, informace)?
- Kolik mě zavedení AI bude stát?
- Jak dlouho zavedení AI asi trvá?
- Kdo mi se zavedením AI pomůže?
- Koho mám hledat na trhu práce?

Proč bychom se měli zabývat umělou inteligencí nebo datovou vědou? Datová věda se zabývá využitím pokročilých datových analytických nástrojů a nástrojů umělé inteligence, jakými jsou *machine learning* (strojové učení) a *deep learning* (hluboké učení pomocí neuronových sítí), ke zpracování dat (*big data*).

V posledních několika letech jsme si zvykli v médiích pozorovat senzační zprávy z oblasti datové vědy typu: „*Vědci z Googlu vytvořili program AlphaGo pro svou AI platformu Deep Mind, a tento program pak porazil 4:1 18násobného mistra světa ve hře Go.*“, nebo „*Superpočítač IBM Watson 3× za sebou zvítězil v kvízové hře Jeopardy, kdy mu protivníky byli 74násobný a 20násobný šampión v této hře.*“ (DeepMind, 2020)

Jsou to jistě skvělé úspěchy z oblasti datové analýzy, strojového učení a umělé inteligence. Populární zprávy však mnohdy způsobují jeden negativní efekt – po vyslechnutí, zhlédnutí nebo přečtení podobných zpráv si člověk představí týmy vědců v bílých pláštích s tlustými brýlemi, jak kdesi v podzemní laboratoři několik let zkoumají a za obrovských nákladů vyvíjí komplikované programy, a občas se některému z těchto týmů něco podaří, třeba porazit velmistra ve hře Go.

Taková představa může vést k mylnému závěru, že zavedení datové vědy, strojového učení nebo umělé inteligence v mé firmě je zhola nemožné,

protože nemám peníze na partu vědců v bílých pláštích a vlastně ani ve výrobě nehraji Go. Vždyť dělám opravdový byznys, který mě živí, a nemám čas zabývat se hrami. Myšlenka na umělou inteligenci je zapovězena. Chyba!

Budeme rádi, pokud se po přečtení naší knihy přesvědčíte, že aplikace datové vědy (*big data*, *advanced data analytics*, *machine learning*) je ve vaší firmě nejen možná, ale je i technologicky a cenově dostupná a může být pro vaši firmu velkým přínosem (nižší náklady, vyšší výnosy, nebo zcela nová tržní příležitost). Možná nakonec dospějete k názoru, že je pro vaši firmu naprosto nepostradatelná.

***

Jste-li ostřílenými harcovníky internetových diskusí zapálenými do aktuálních trendů, zvyklí vše si vyhledat přes Google, nebude pro vás tématika knihy velkou překážkou. Zkrátka se zakousněte a přejeme příjemnou jízdu. Málokdo má však na takový přístup vlohy nebo čas.

Pro vás, kteří se rádi dozvíte něco nového, ale připadá vám příliš náročné se zabývat některými tématy pomalu až na úrovni experta, jsme připravili několik tipů, jak pracovat s naší knihou, abyste si toho odnesli co nejvíce:

- **Pojmy** jsou pro tuto oblast stěžejní. Neobejdete se bez nich. Je jich hodně. Nezoufejte, připravili jsme pro vás slovníček pojmů, k němuž doporučujeme se neustále vracet. Obavy nejsou na místě – všechny pojmy se hravě zažijí.

- Není vyprávění bez příběhu. Aby bylo téma lépe uchopitelné, připravili jsme pro vás **boxy s příklady** a hlubším **vysvětlením pojmů**. Jedná se o pomůcku na dovysvětlení, ale můžete je přeskočit, aniž by vám něco uniklo.

- Kniha není míněna coby návod „Jak se rychle a účinně stát datovým vědcem“. Za tím stojí dřina a desítky měsíců učení a práce. Po přečtení byste měli být schopni vydat se na dlouhou, ale zábavnou a užitečnou cestu aplikace umělé inteligence ve vaší firmě. Nebudete-li tedy všemu rozumět, nebo vám některé znalosti budou připadat příliš povrchní, nezoufejte a ponořte se do hlubšího studia tématu nad rámec knihy. **I cesta může být cíl**.

- Na konci každé kapitoly jsme pro vás připravili oddíl „Co jsme se v kapitole dozvěděli?“, kde jsou dílčí **témata shrnuta do několika vět**. Zkuste se zamyslet, jestli jste se to skutečně dozvěděli a sami pro sebe si téma interpretovat, případně se ke kapitole znovu vraťte.
- Konfrontujte se s aktuálními fakty. Svět se vyvíjí bleskovou rychlostí a než se vám tato kniha dostane do ruky, mnohé již nemusí platit. Navíc se mohou i některé přístupy lišit – ne dramaticky, ale přece. Stojí za to mít přehled.

K čemu to je: machine learning

Než se pustíme do hlubšího vysvětlování pojmů a principů, zaměřme se na užitečnost tématu, jemuž se nadále budeme věnovat. Podívejme se na praktické využití metody *machine learning* (strojové učení, dále jen ML). Ta se řadí do oblasti umělé inteligence (AI), umožňuje automatické učení a zlepšování na základě zkušeností a historických dat bez explicitního programování.

Později si ukážeme ve větším detailu, že procesy *machine learningu* jsou klasifikovány na *supervised* a *unsupervised* (a někdy i *semi-supervised*), tedy tzv. s učitelem a bez učitele. V praxi to znamená, že *supervised* techniky pracují s historickými označenými daty a na nich se učí. Po naučení (tzv. natrénování) jsou schopny statisticky odhadnout výsledek neznámých vzorků a přiřadit jim označení (v terminologii používáno „*label*"). Naopak *unsupervised* techniky pracují s neoznačenými daty a hledají tak mezi nimi asociace, relace a různé logické prvky, které lze pak na nových datech aplikovat a určit právě tyto asociace či je nějak zařadit dle naučeného algoritmu.

Rozdíl mezi *supervised* a *unsupervised* je zřejmý již ve chvíli, kdy se podíváme na data, která budou vstupovat do našeho modelu. V níže znázorněné tabulce 1 jsou zobrazena data použitelná pro *supervised* ML (ve skutečnosti se jedná o malý vzorek použitelný pro strojové učení). Každý řádek obsahuje jednotlivý záznam k jednomu produktu. Ve sloupcích se uvádějí nezávislé proměnné, jako například barva, datum výroby, stav či cena produktu. V posledním sloupci jsou zaznamenány labely – informace, zda daný produkt byl prodán či nikoli.

Tabulka 1 – Data o produktech

Značka	Barva	Vyrobeno	Cena [EUR]	Vlastník	Použité / nové	Prodáno
Alfa	Silver	06/10/2017	54901	Company	Použité	Ano
Beta	Blue	01/21/2018	64180	Private	Nové	Ne
Gamma	Silver	04/28/2017	11985	Private	Použité	Ne
Gamma	Silver	09/15/2016	42359	Company	Nové	Ano
Alfa	Blue	02/4/2018	63414	Private	Nové	Ne
Gamma	Silver	12/12/2016	88929	Private	Nové	Ano
Gamma	Red	01/29/2018	48609	Company	Použité	Ano
Gamma	Blue	09/13/2017	42245	Private	Nové	Ano
Beta	Silver	01/21/2018	56233	Company	Nové	Ano
Alfa	Silver	09/25/2016	94462	Private	Nové	Ano

V tabulce 2 jsou znázorněny informace o zákaznících konkrétní firmy. Každý řádek tak představuje jednoho zákazníka a v jednotlivých sloupcích jsou informace jako věk, kraj, pohlaví nebo měsíční příjem v CZK. Tento seznam však neobsahuje žádný label, tedy výstupní informaci, která by pro nás byla hodnotná a podle níž bychom mohli vykonat akci výhodnou pro náš business (např. segment zákazníka nebo informaci, zda naši firmu zákazník neopustil či opustil a přešel ke konkurenci). Tato data zjevně slouží k variantě *unsupervised learning*.

Tabulka 2 – Data o zákaznících

ID zákazníka	Věk	Pohlaví	Bydliště kraj	Měsíční příjem [Kč]
2650	38	Muž	Karlovarský	21180
2613	55	Muž	Olomoucký	22500
2921	65	Žena	Moravskoslezský	18000
2910	35	Muž	Praha	43130
2473	24	Žena	Vysočina	17200
2276	63	Muž	Plzeňský	56500
2492	36	Muž	Ústecký	40100
2943	46	Muž	Středočeský	38120
2803	28	Žena	Liberecký	21450
2253	20	Muž	Královéhradecký	17650

Tyto dva příklady nám pomohly pochopit, jak snadno na první pohled rozeznat, zda je data možné použít pro metodu s, respektive bez učitele. Pozorný čtenář si jistě všimne, že i data bez labelů lze ručně uzpůsobit tak, abychom byli schopni použít metodu s učitelem. Nicméně, cesta ruční úpravy je velmi pracná, a právě metody bez učitele ji dostatečně nahrazují.

Využití strojového učení s učitelem

Pro řešení úlohy pomocí *supervised learning* volíme mezi dvěma typy řešení – klasifikační a regresní. Jaký typ zvolíme, nám určuje povaha dat. Je-li label vstupních dat hodnotou kategorickou (každý výskyt je přiřazen do určité kategorie), pak se jedná o kategorický způsob. Avšak pokud je hodnota číselná, volíme regresi.

Klasifikace

Vraťme se k našemu příkladu s prodejem produktu, u něhož jsme identifikovali label „Prodáno“. Vidíme zde dvě skupiny „ano“ a „ne“. Zároveň vidíme, že data mají popsanou vlastnost, či jsou zařazena do skupiny, tedy klasifikována, proto použijeme metodu klasifikace, aby nám umělá inteligence pomohla odhadnout na základě předchozí zkušenosti, jak budou budoucí data klasifikována. V našem příkladu nás zajímá, zda se bude produkt prodávat, či nikoli.

V *machine learningu* se nejčastěji setkáváme s klasifikačními labely typu ANO / NE, PRAVDA / NEPRAVDA, PES / KOČKA / MORČE / HAD / ŽELVA, DAL VÝPOVĚĎ / NEDAL VÝPOVĚĎ, NÁDOR ZHOUBNÝ / NÁDOR NEZHOUBNÝ apod. Tyto labely nám označují zařazení dat do skupin a na základě nich tak můžeme u nového vzorku např. predikovat, že daný zákazník pravděpodobně vypoví smlouvu na základě parametrů, na kterých jsou trénovací data naučena a testovací data otestována.

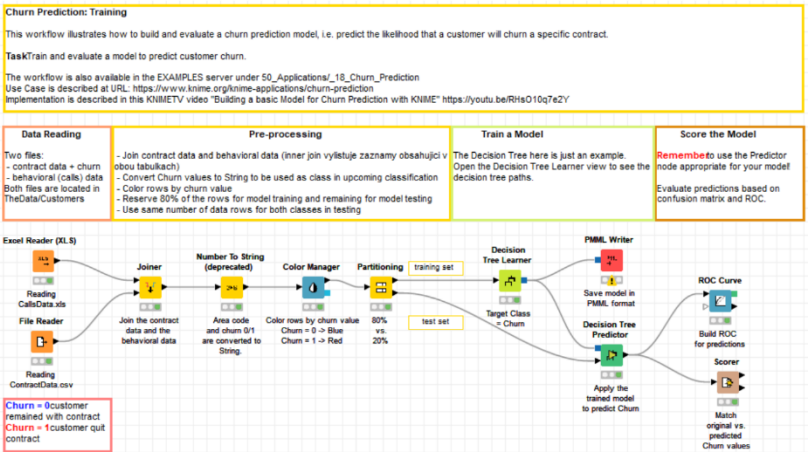
Mám tu zůstat, nebo jít?
Jako typický příklad klasifikačního prediktivního problému se v literatuře a případových studiích uvádí tzv. *churn modelling, který* slouží k predikci výpovědí klientů z firmy, resp. vede k identifikaci konkrétních zákazníků, kteří mají nebo budou mít tendenci opustit naši firmu a přejít ke konkurenci. *Churn model* tedy umožní tyto rizikové zákazníky identifikovat, my bychom je poté měli kontaktovat a snažit se je nadále udržet např. pomocí změny cenových nabídek nebo jiných podmínek, například výhodnějšího nákupu produktu. To má za následek snížení rizika ztráty klientů, a tím dochází k udržení tržeb firmy.

Tyto modely jsou využívány především v bankovnictví a telekomunikacích, kde jsou zákazníci dlouhodobě a trvale klienty dané firmy, která tak disponuje velkým množstvím dat o každém zákazníkovi, a to ve stejné struktuře, což umožňuje tvořit *machine learningové* prediktivní modely.

Jestliže k těmto datům existují i informace, zda konkrétní zákazník vypověděl smlouvu s danou firmou, pak lze vytvořit prediktivní klasifikační model, pomocí kterého na základě vybraných

klasifikačních metod firma dokáže identifikovat rizikové klienty, které by mohla ztratit a rozhodnout o dalších krocích na udržení těchto klientů (kontaktování, diskuse, nabídka speciálních programů a podmínek apod.).

A jak takový model vypadá? Vydejme se tou neschůdnější cestou a pojďme si ukázat, jak jednoduše lze model vytvořit a znázornit pomocí již existujícího vzoru v softwaru KNIME Analytics Platform. Tento software je snadno dostupný (zdarma) a je jednou z často používaných platforem pro tvorbu *machine learning* a *deep learning*. Jeho výhoda tkví v jednoduchosti a možnosti tvořit modely bez nutnosti použití programovacího jazyka. Obrázek 1 nám ilustruje schéma jednotlivých kroků, které model používá. V této podobě model jednoduše sestavíte sami.



Obrázek 1 – Churn analýza (KNIME Analytics Platform, 2020)

Software KNIME Analytics Platform je, jak již bylo řečeno, z kategorie freeware, tudíž jej můžete stáhnout a používat zcela zdarma. Byl vyvinut v akademickém prostředí za účelem rychlé aplikace *machine learning* a *deep*

learning metod a rychle našel aplikační využití i v komerční sféře. Jeho velkou výhodou je jednoduchost, s níž můžete vytvořit model „na klik“, neboť jednotlivé kroky vkládáte pomocí ikon ze seznamu předdefinovaných kroků do společného *workflow*.

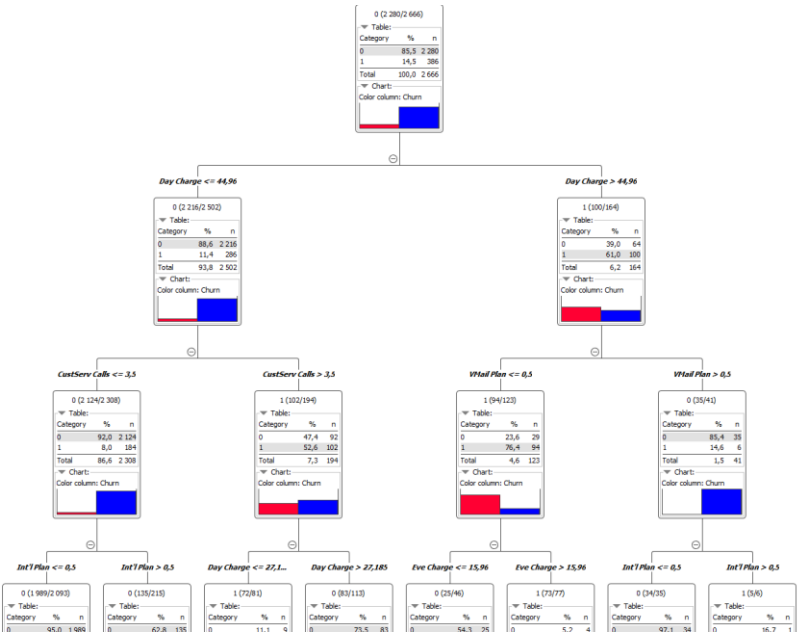
Pokud je vám programování a kódování cizí, nemusíte si nutně zoufat, že datová věda není pro vás. KNIME vám umožní celý model „naklikat“. U každého uzlu představujícího aktivitu máte možnost nastavení specifického dané aktivitě. Sám program má značné množství ukázkových *workflow*, jako například zmiňovaný *churn model*, z něhož můžete doslova okoukat, jak se softwarem pracovat. Na druhou stranu nelze předpokládat, že bez znalosti datových modelů opanujete datovou vědu pomocí KNIME. Pouze vám usnadní práci bez nutnosti kódovat.

Jak KNIME funguje?

Schéma na obrázku 1 první ikonou (uzlem) ukazuje, že celý proces začíná sběrem dat, které načte v našem případě ze souboru MS Excel. V něm jsou historické údaje o klientech – ID, které „odpersonalizuje“ nejen klienty a další údaje o jejich chování (měsíční paušál, délka kontraktu, apod.), ale také zásadní informaci, zda klient zrušil kontrakt (1), či nikoli (0).

Takto získaná data před použitím v modelu musíme nejprve poupravit, zformátovat sloupce do správných datových typů, vyčistit, spojit datové soubory apod. Mějme na paměti, že na pozadí umělé inteligence stojí matematické modely, které potřebují přesně dané a strukturované vstupy, jinak nefungují.

Pro náš model jsme použili prediktivní nástroj *decision tree* (rozhodovací strom). Nejprve je třeba rozdělit data na tréninková (obvykle 80 %) a testovací, na nichž si ověříme, že se správně natrénoval. Výsledkem je pak graf ve stromové struktuře (viz obrázek 2 – Ukázka *decision tree* v programu KNIME), kde lze přímo vyčíst, s jakou přesností model předpovídá. Každé „patro“ stromu nám říká, jak se rozhodujeme u dílčích parametrů (např. zákazník má kontrakt delší než 5 let). Každé takové rozhodnutí nám ukazuje cestu od nejnižších pater stromu napříč parametry s pravděpodobností hrozby, že zákazníka ztratíme.



Obrázek 2 – Ukázka decision tree v programu KNIME

Regrese

Příklad *churn analýzy* demonstroval zhodnocení chování zákazníka redukovaného na dva stavy. Co se však týče výsledných číselných hodnot, na ty statistika pamatuje metodou zvanou regrese. Strojové učení si osvojilo tuto metodu, takže umí naučit model číselné hodnoty tak, aby výsledkem byla i číselná hodnota. Regrese nám tak umožňuje predikovat kupříkladu ceny akcií, kurzy měn, ceny nemovitostí, množství napadaného sněhu, vývoj počtu obyvatel apod.

Mezi regresní metody se řadí lineární regrese, exponenciální regrese, polynomiální regrese, logaritmická regrese a další metody. Rozdíl v těchto typech je dán vztahem mezi proměnnými a výsledkem.

Pojďme se zaměřit na častý příklad *machine learningu* pro regresní problém na trhu s komoditami, kde dochází k prodeji a nákupu různých

typů komodit, jako jsou např. domy, pozemky, podnikatelské provozovny apod. Ceny těchto komodit závisí na řadě faktorů, které určují atraktivitu pro potenciální kupce. Regresní analýza je statistický proces, který studuje závislosti mezi jednotlivými faktory a cenou dané komodity. Pomocí těchto technik lze porozumět, jak se tato cena pohybuje, dochází-li ke změnám ovlivňujících faktorů.

Unsupervised machine learning (tzv. strojové učení bez učitele)

Podobně jako lidská mláďata se i stroje učí dvojím způsobem – s učitelem a bez něj. Začněme nejprve bez něj, tedy technikou *unsupervised learning*, kterou obvykle rozdělujeme na dva základní typy, a to:

- asociační analýzu (*association rules*),
- shlukování (*clustering*).

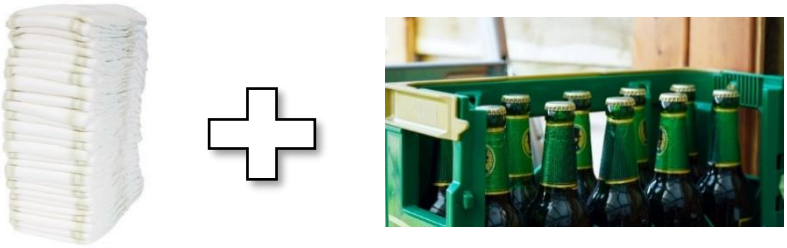
Asociační analýza

Asociační analýza je metoda, která odhaluje zajímavé relace, tedy objevuje vztahy mezi jednotlivými prvky v datech. Je zaměřená na identifikaci pravidel a definování zmíněných relací mezi daty.

Asociační analýza nejčastěji nachází využití v produktové a marketingové strategii, především v retailu nebo internetových obchodech. Jako neuvěřitelný příklad asociační analýzy se často uvádí asociace mezi plenkami a pivem. Jak spolu tyto dvě na první pohled rozdílné věci mohou souviset a co stálo za zrodem této kuriózní myšlenky?

Tato analýza byla kdysi zpracována a využita v nadnárodní retailové společnosti. Pomocí asociační analýzy jistý obchodní řetězec získal velmi zajímavou a hodnotnou informaci ohledně vztahu mezi určitými produkty. A to, že každou neděli v podvečer vzrostl prodej plenek a piv v jednotlivých nákupních košících.

Otázka proto zněla: Opravdu malé děti rády pijí pivo?



Obrázek 3 – Pijí kojenci více piva?

Po dalším průzkumu a dotazování firma zjistila, že v podvečer často chodí nakupovat „tatínci" plenky pro své děti a při té příležitosti si rádi koupí pivo domů. Obchodní řetězec této informace využil a vedle regálu s plenkami umístil regály s pivem. Tímto aktem řetězec nejlépe potvrdil svou hypotézu, neboť se firmě výrazně zvýšily tržby.

Pro další příklad nemusíme chodit daleko. Většina z nás má zkušenosti s nákupy přes internet. Kupujete-li např. sportovní boty, internetový obchod vám ve většině případů nabídne další zboží, které nějakým způsobem souvisí s vybranými botami, a to buďto jako doplňkové zboží nebo jako zboží, které ostatní zákazníci zakoupili právě s těmito botami. K botám vám tedy nabídnou běžecké tričko, ponožky, impregnační ochranu na boty apod.

Zásadní otázkou je, jak to obchodníci dělají? Obchodníci totiž sbírají data prodaných produktů a zjišťují nejčastější kombinace produktů, které jednotliví zákazníci zakoupili v jednom nákupním košíku. Poté novým zákazníkům doporučují tyto produkty k již vybraným produktům.

Prokoukat se k úsporám

Firma Netflix uspořila nemalé částky pomocí asociační analýzy a doporučovacího systému. Netflix je americký placený poskytovatel online filmů sídlící v Kalifornii. Tato firma ušetřila jednu miliardu amerických dolarů za rok na udržení zákazníků. A jak se jí to povedlo? Firma Netflix jako jedna z velkých firem využila investic do sběru dat,

aby expandovala svůj obchod. Digitální uživatelé, jichž získal Netflix na začátku roku 2019 139 milionů, jsou tím hlavním nástrojem *big dat*.

Netflix sbírá data o jejich vyhledávání, hodnocení zhlédnutých programů, o opakovaně zhlédnutých programech atd. Toto pomáhá Netflixu navrhnout každému uživateli doporučení zábavného programu na míru. V sílící konkurenci znamená každý ztracený zákazník finanční ztrátu, efektivní využití dat je proto esencí konkurenceschopnosti v oblasti zábavního průmyslu.

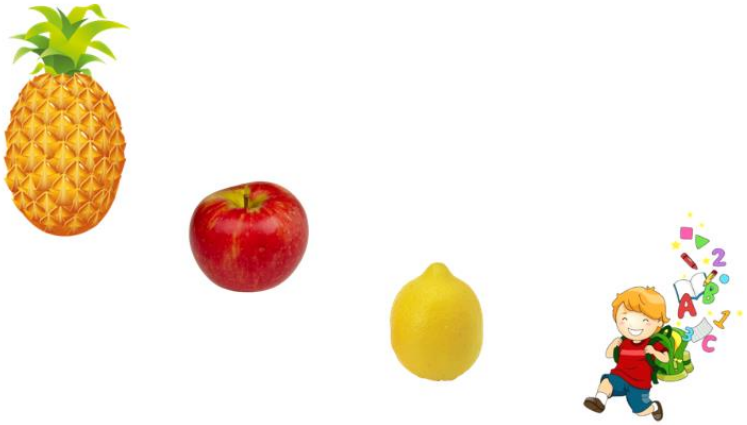
Clustering

Clustering, neboli metoda shlukování se soustřeďuje na rozdělení dat do určitého počtu skupin tak, aby spolu tyto skupiny na základě podobnosti v datech nějak souvisely. Data v jedné skupině jsou pak odlišná od dat ve skupině jiné. Metodou, která se nejčastěji využívá, je tzv. *k-means clustering*, kde K vyjadřuje počet skupin.

Metoda se využívá např. pro analýzu zákazníků, kdy firma nemá k dispozici jednotlivé segmenty neboli clustery a potřebuje mít zákazníky roztříděné do skupin. *machine learning* tak vytvoří skupiny dle parametrů zákazníků, kde hledá podobné a rozdílné vlastnosti napříč všemi zákazníky a poté zařadí zákazníky s podobnými vlastnostmi do jedné skupiny, další do druhé atd. S každou skupinou pak vede obchodník odlišné obchodní jednání, které přinese nejlepší výsledek právě u dané skupiny.

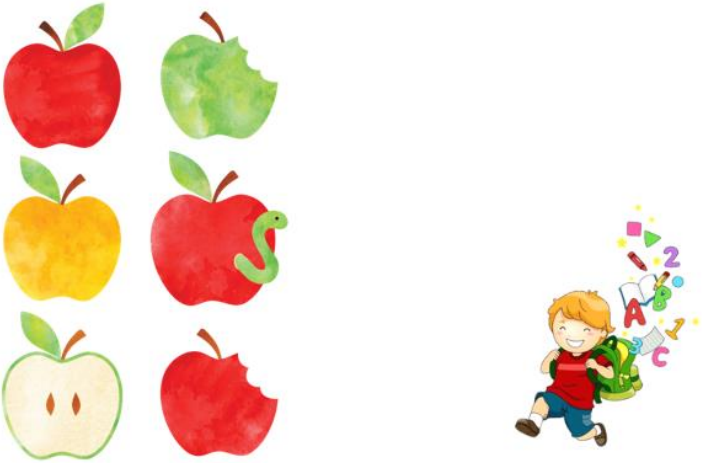
Jak tedy funguje *machine learning*? Pojďme se podívat na klasifikační problém učení s učitelem.

Dítě řeší klasifikační problém – jak rozpoznat tři typy ovoce: jablko, citrón a ananas. V dětském světě existuje někdo nebo něco (rodič, učitel, atlas ovoce atd.), jež klasifikaci ovoce již zná a dítěti tuto znalost předá (proto *supervised* technika – s učitelem).



Obrázek 4 – Rozeznávání ovoce

Na základě dříve naučené klasifikace ovoce je dítě schopné rozpoznat např. jablka, i když konkrétní typ jablka, nebo jeho stav nikdy předtím nevidělo. Zkrátka jsou mu sděleny znaky, které daný objekt má, a pro větší efektivitu jsou často dávány do kontrastu další objekty, které se ve znacích odlišují. Nesmíme opomenout, že každý objekt je jiný a vzájemně zapadají do kategorií jen těmi znaky. Čím více objektů pak uvidí v rámci procesu učení, tím lepší rozeznávací schopnost bude mít.



Obrázek 5 – Rozeznávání jablek

Na podobném principu funguje i *machine learning*, kdy se počítač učí na označených datech, jak poznat daný druh ovoce na základě parametrů. Stejně jako školákovi poskytneme prostor, aby získal vhled do toho, co je jablko a co ananas, a dokonce aby udělal chybu, poskytneme tento prostor i počítači v podobě dat. Zdánlivě musíme „nakrmit" počítač větším počtem dat, ale nezapomínejme na fakt, že školák už má za ty roky vyvinutý slušný kognitivní systém jak rozeznávat tvary, barvy a mnohem víc, zatímco počítač začíná od píky.

Obrázek 6 – Klasifikace rozeznaného ovoce

Co jsme se dozvěděli?

Jak dělíme *machine learning* a jak slouží k predikci v reálném světě.

Machine learning umožňuje *supervised* (data již předem označená) a *unsupervised* (data bez označení) *learning*, na jehož základě se učí a dokáže předpovídat nová data dle naučeného algoritmu.

Co mají společného *machine learning* a školák.

Aplikace v příkladech

„Z příkladu jiných lidí se nauč, co máš dělat a co ne. Život jiných je naším učitelem.“

Cato starší

Datová věda přináší své ovoce již značnou dobu a nezanedbatelně promlouvá do světa byznysu. Pojďme se nyní podívat na některé úspěšné projekty zapojení datové vědy ve firmách, kterým přinesly značnou úsporu, nebo pomohly vyřešit problémy. Významným posunem pro změnu přístupu firem k datové analytice není jen inspirace přímo v řešení problému, ale i samotná identifikace toho, že se vůbec o problém řešitelný pomocí datové vědy jedná. To je pravděpodobně častější případ, jelikož „provozní slepota“ širšímu nahlížení na řešení mnohdy brání.

Kaggle

Vítejte ve světě Kaggle
Kaggle.com je webová platforma pro analytiky. Hlavním tématem je zde *machine learning*, *big data* a pokročilá datová analytika. Analytici zde mohou získat cenné znalosti a zkušenosti. Jsou zde řešeny nejen hypotetické příklady, ale i příklady z reálného světa.

Kaggle je také velmi zajímavý tím, že jakákoli firma nebo organizace může vyhlásit soutěž s cílem vyřešit nějaký svůj analytický problém. Do soutěže o odměnu pro nejlepší řešení problému se mohou zapsat jednotlivci, nebo řešitelské týmy. Svá řešení pak v prostředí Kagglu publikují, diskutují a dále vylepšují. Výherci získají vypsanou odměnu, která může být až v řádu milionu amerických dolarů.

Řešení, která jsou publikovaná či diskutovaná na Kagglu, je možné zkopírovat, upravit a spustit – většina řešení je publikovaná jako aktivní Jupyter Notebook (viz informace o Jupyter Notebooku). Notebooky jsou psány buď v pythonu, nebo v jazyce R, a bývají bohatě komentované.

Jak s Kagglem (pro začátek) pracovat?

Můžete hledat téma „predikce fluktuace zaměstnanců“ a najít k němu řadu publikovaných řešení. Pak si v *„rank listu“* snadno vyberete tři nejlépe hodnocená řešení, ta si otevřete, spustíte, a pokud pro vás budou zajímavá, můžete si je zkopírovat do vlastní verze a na té pak pracovat. Výsledek si můžete stáhnout a použít ve svém lokálním Jupyter Notebooku, případně zavést do svého analytického scénáře, projektu či *workflow*.

Naší první zastávkou je platforma Kaggle, kde se zaměříme na zajímavé soutěže v řešení datových úloh na reálných datových sadách.

Mercedes – zkrácení doby nutné pro testování konfigurací vozů

Aby byla zajištěna bezpečnost a spolehlivost automobilu předtím, než se poprvé objeví na silnici, vyvinuli inženýři společnosti Daimler (vyrábějící vozy pod obchodní značkou Mercedes-Benz) robustní testovací systém. Moderní automobily jsou nabízeny zákazníkům v několika výrobních řadách s bohatou možností konfigurace a individualizace pro každý typ.

Firma Daimler se snažila testovací systém vytvořit co nejjednodušší a nejrychlejší. Časem však přestala být spokojena s délkou testovacího procesu, který podstatně zpomaluje vývoj nových vozů, což ve světě automotive znamená značnou konkurenční nevýhodu. Optimalizace rychlosti jejich testovacího systému pro tolik možných konfigurací vozů se však ukázala složitá a časově náročná. Protože pro Daimler je bezpečnost prvořadá, zadání na optimalizaci požadovalo zkrátit testovací proces při zachování jeho současného vysokého standardu.

V Daimleru proto vytvořili datovou sadu, která obsahovala jednotlivé konfigurace vozů a také čas, který byl třeba pro otestování každé konfigurace (jejím obsahem bylo 377 možných komponent jako např. klimatizace, pohon 4×4 apod.) a jejich možné kombinace. Tato datová sada posloužila jako základ vypsané analytické soutěže s dotací 25.000 USD na platformě Kaggle.com.

Výsledky analýzy datového souboru mimo jiné umožnily firmě Daimler identifikovat klíčové body testování, ukázaly jim možnosti pro optimalizaci pořadí testů, shlukování testů apod. Predikce délky testů pro jednotlivé konfigurace umožnila také lepší plánování směn, čímž výrazně zefektivnila celý proces. Redukce délky testování vozů také vedla ke snížení emisí CO_2 (Kaggle, 2017).

Ford – detektor bdělosti řidiče

Pokud během jízdy autem provádíte činnosti, které odvádějí vaši pozornost od toho, co se děje na silnici před vámi a okolo vás (telefonujete, přijímáte nebo odesíláte textové zprávy, svačíte, nebo živě konverzujete s pasažéry), nebo pokud jedete již dlouhou dobu bez přestávky a jste unavení, může vaše jízda skončit tragicky. Toho jsou si vědomi i inženýři z Fordu.

Moderní automobily jsou vybaveny informačním systémem, širokou paletou čidel a centrální jednotkou, která data z čidel a další informace o stavu vozidla shromažďuje a vyhodnocuje. Pracovníky Fordu napadlo, jestli data, jež jsou během jízdy shromažďována, bude možné využít pro vyhodnocení, zda se řidič plně soustředí nebo nesoustředí na jízdu.

Inženýři Fordu připravili datový soubor, který obsahoval 33 charakteristik (fyziologická data, data z vnějšku vozu, data z vnitřku vozu) uspořádaných do cca dvouminutových sekvencí při frekvenci sběru dat každých 100 ms. Data pocházela od přibližně 100 řidičů, obou pohlaví, různého věku a etnického původu. Z hlediska modelu byla zásadní klasifikace obsahující stav řidiče: 1 – je bdělý (pozorný, soustředí se na jízdu), 0 – není bdělý (nesoustředí se na jízdu).

Takto připravená data byla základem jedné z prvních analytických soutěží na platformě Kaggle.com, s dotací 950 USD (v roce 2011 byly odměny na Kagglu v řádu stovek USD, v roce 2019 jsou to desítky a někdy i stovky tisíc USD) a vstupenkami na konferenci o neuronových sítích. Cílem této soutěže/výzvy bylo navrhnout detektor/klasifikátor, který zjistí, zda se řidič soustředí na jízdu, s využitím jakékoli kombinace dat o řidiči, voze, nebo prostředí, která jsou získána během jízdy.

Výsledky soutěže Fordu pomohly s vývojem systému, který upozorní řidiče, že by se měl více věnovat řízení, nebo může doporučit, aby řidič zastavil a odpočinul si. Aplikace *machine learningu* přinesla novou technologii, která zvýšila bezpečnost na silnicích. Tato technologie se začala používat i v automobilech dalších značek a stala se standardní součástí informačních systémů moderních dopravních prostředků (Kaggle, 2011).

Další aplikace ve firmách

Svět datové vědy se netočí jenom kolem platformy Kaggle a automobilového průmyslu. Podívejme se, jak je možné usnadnit tzv. prediktivní údržbu pomocí strojového učení.

Lennox International Inc.

Firma Lennox International Inc. vyrábí topné a klimatizační systémy pro firmy i domácnosti. Zákazníkům poskytuje jako službu také údržbu svých zařízení. V minulosti byla údržba zařízení prováděna preventivně, nebo na základě odhadů, kdy by mohlo dojít k selhání. Mnohdy tak docházelo k falešným poplachům, což bylo frustrující jak pro prodejce, tak pro zákazníky.

Firma Lennox se rozhodla tuto situaci zlepšit a přišla s konceptem prediktivní údržby svých zařízení, který byl součástí její digitalizační strategie (v celkové hodnotě 4 miliard dolarů). Základem je monitorování výkonu stroje v reálném čase – sběr dat a jejich vyhodnocování pomocí *machine learningového* modelu.

Model průběžně vyhodnocuje data a umožňuje s 90% pravděpodobností určit, kdy dojde k selhání stroje. Lennox tak může svým zákazníkům od majitelů domů až po manažery obchodních center poskytnout 4 hodiny předem informaci o tom, že k selhání s vysokou pravděpodobností dojde. A co je podstatné, dříve Lennox platil svým lokálním servisním partnerům několik servisních návštěv u svých zákazníků ročně a nyní (díky digitalizaci a datové vědě) platí pouze servisní zásahy, které jsou opravdu opodstatněné. Přesná data z mnoha zařízení také umožňují Lennoxu cíleně zdokonalovat své produkty a zvyšovat jejich spolehlivost.

Lennox International Inc. používá platformu MS Azure, na které provozuje systém Spark od Databricks – jako svůj hlavní *machine learning* software a jako sjednocenou platformu, na které spravuje a vyhodnocuje stovky terabytů dat ze stovek databází (Manohararaj & Chandravihar, 2020).

U.S. Bank

Není ilustrativnější sektor, kde potřebujete mít detailní přehled o svých zákaznících, než je bankovnictví. Stejně jako mnoho velkých bank i U.S. Bank shromažďuje velké množství údajů o zákaznících. A stejně jako většina světových bank i U.S. Bank dlouho neúspěšně bojovala s tím, jak tato data efektivně využít. Vedení firmy se rozhodlo tuto situaci změnit. Analytický tým banky vyvinul a produktivně nasadil modely *machine learningu*, které jí umožňují odhalovat a efektivně využívat hodnotu ukrytou v datech.

Pokud například zákazník hledá na webových stránkách U.S. Bank informace o hypotečních úvěrech, pak si bankéř nebo operátor zákaznického servisu může s tímto zákazníkem promluvit o úvěrech při jeho příští návštěvě pobočky nebo prostřednictví telefonu. *Machine learning* v U.S. Bank pomáhá také odhalit vzorce chování, kterých si lidé zde zaměstnaní nemusí všimnout.

Model může například doporučit, aby agenti zavolali potenciálnímu klientovi pracujícímu v určitém oboru v úterý mezi 9:00 a 11:00, protože je pravděpodobnější, že zvedne telefon a nebudete jej nijak zásadně vyrušovat. Na základě dat může banka předpovídat i další potřeby svých zákazníků – jejich bonitu, míru spokojenosti, schopnost splácet úvěry, nebo i kanály, kterými chtějí s bankou komunikovat.

U.S. Bank používá technologii Einstein AI smart CRM assistant od společnosti Salesforce.com (Tarique, 2017).

Akershus University Hospital

Nejcennější komoditou je zdraví, a i zde může datová věda pomoci. Univerzitní nemocnice využila *machine learning* pro klasifikace CT skenů u dětských pacientů a pro vyhodnocení, zda jsou CT skeny využívány způsobem, který balancuje případné pozitivní efekty ve vztahu k možným vedlejším efektům, aby byla léčba co nejšetrnější. Při analýze příslušných lékařských zpráv byly pro automatickou klasifikaci pozitivních a negativních nálezů použity rovněž metody rozpoznávání přirozeného jazyka (v norštině). Přesnost klasifikačního modelu dosáhla neuvěřitelných 99 %.

Úspěch pilotního projektu pomohl nastartovat další projekt, který je zaměřen na správnou klasifikaci pacientů s rakovinou prostaty. Nemocnice bude analyzovat data od 1800 pacientů (lékařské zprávy, CT a MRI snímky) s cílem vyhodnotit správnost diagnózy a opodstatněnost zařazení pacienta do léčebného programu. Vyloučení pozitivní diagnózy pak zamezí plýtvání cennými zdroji a traumatizování zdravých pacientů.

Nemocnice použila *machine learningovou* platformu IBM Watson Explorer (IBM, 2017).

Pojmy, pojmy, pojmy

„Dějiny člověka jsou dějinami pojmů, o které postupně rozšířil své znalosti.“

Antoine de Saint-Exupéry

Bez pojmů se neobejdeme, neboť bychom si správně nerozuměli. Pojďme si tedy představit ty nejzásadnější z nich.

Big data

Big data představují soubory dat, jejichž velikost znemožňuje schopnost je načítat, spravovat a analyzovat běžně používaným softwarem v rozumném čase.

Big data se tak stala IT disciplínou, která se zabývá nalezením cest, jak získávat informace, různě zpracovávat a analyzovat datové sady, které jsou extrémně objemné a komplexní a je tak nemožné je zpracovávat tradičními aplikačními softwary v požadované rychlosti a čase. Oblast *big dat* sestává z aktivit, mezi něž se řadí sběr dat, ukládání dat a aktualizace, datové analýzy, přesun dat, jejich vizualizace a transformace.

Využití *big dat* směřuje k prediktivním analýzám a ostatním pokročilým analytickým metodám pro získání hodnoty „*VALUE*“ z dat za pomocí *machine learningu* (strojového učení) a *deep learningu* (hlubokého učení) algoritmů a dalších pokročilých technik z oblasti *artificial intelligence* (umělá inteligence).

Data mohou mít formu strukturovaných, semi-strukturovaných a nestrukturovaných dat, kde záleží na tom, co je zdrojem generovaných dat a v jaké jsou struktuře.

Strukturovaná data představují různé databáze, firemní data, která mají jasnou strukturu, většinou formátovanou v tabulkách do sloupců se záznamy v jednotlivých řádcích. Mezi nestrukturovaná data lze řadit textové datové sady generované např. Twitterem, Facebookem, blogy na internetových stránkách a dále do této kategorie spadají také obrázky. Semi-strukturovaná data tak sestávají z dat na pomezí strukturovaných a nestrukturovaných dat, a přestože nejsou organizovaná do databází nebo podobných

strukturovaných aplikací, obsahují informace jako metadata, která umožňují jednodušší procesování než data nestrukturovaná.

Nejsnazší způsob uchopení pojmu *big data* je ustálená charakteristika pomocí „šesti + 1 V“:

Volume: svět generuje extrémní množství dat.

Variety: data mají různou strukturu.

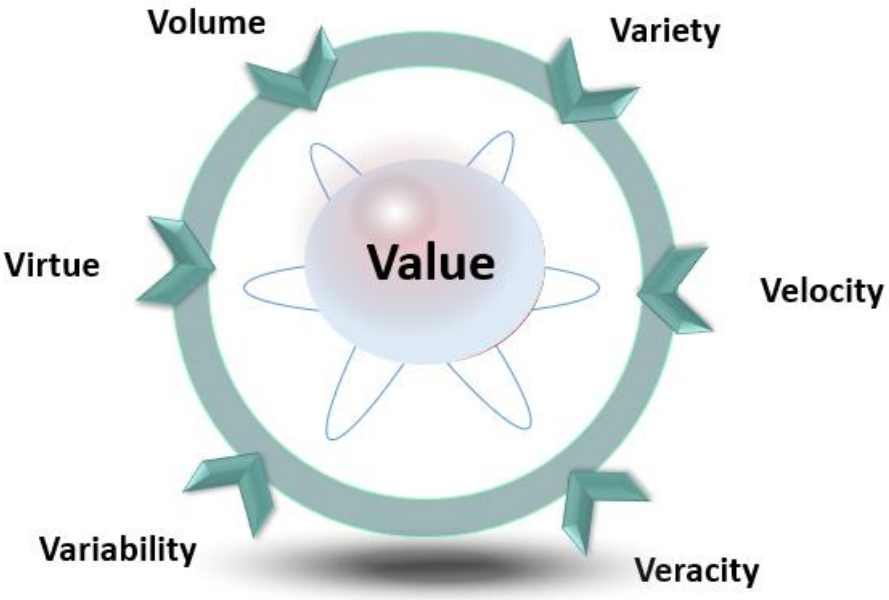
Velocity: rychlost, jakou jsou data generována, enormně roste.

Veracity: data nejsou vždy důvěryhodná a kvalitní.

Variability: množství možností využití dat je různorodé.

Virtue: dodržování etického kodexu.

Value: data přinášejí hodnotu a napomáhají k rozhodnutím.



Obrázek 7 – „6 +1 V“ pro big data

Proč jsou big data důležitá, proč se jimi máme zabývat?