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)

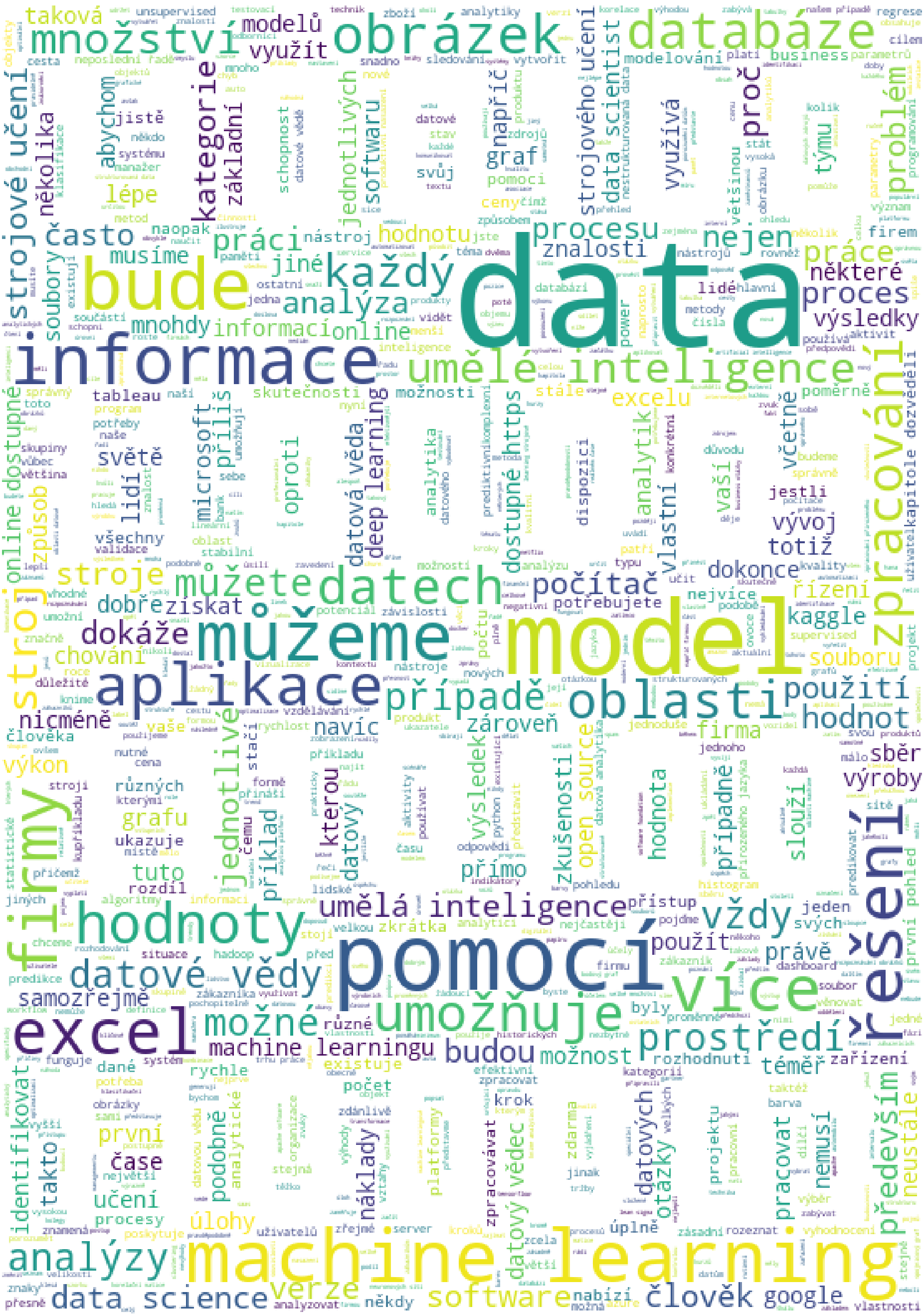
ISBN 978-80-908226-3-4 (PDF)

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
můžeme
informace
aplikace
machine learning
excel
hodnoty
obrázek
databáze
bude
řešení
oblasti
datech
firmy
umělé inteligence
umělá inteligence
zpracování
množství
umožňuje
datové vědy
prostředí
více
každý
analýzy
data science
software
člověk
google
případě
stroj
stroje
kategorie
analýza
procesu
proces
strojové učení
možné
machine
learningu
budou
vždy
hodnot
použití
počítač
dokáže
často
práci
práce
nejen
znalosti
informací
online
excelu
problém
strojového učení
data scientist
jednotlivých
jednotlivé
microsoft
datová věda
deep learning
open source
dostupné
zkušenosti
hodnota
datový vědec
identifikovat
především
neustále
náklady
verze
úlohy
učení
první
tuto
grafu
výkon
chování
získat
dobře
můžete
musíme
lépe
několika
abychom
základní
softwaru
graf
využívá
proč
týmu
některé
výsledky
modelů
využít
taková
firma
navíc
zároveň
samozřejmě
nicméně
totiž
vývoj
dokonce
řízení
kaggle
souboru
sběr
výroby
přímo
tableau
světě
lidí
příliš
skutečnosti
možnosti
analytik
vlastní
dispozici
včetně
jestli
slouží
případně
kterou
datový
příklad
způsob
soubory
modelování
business

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.

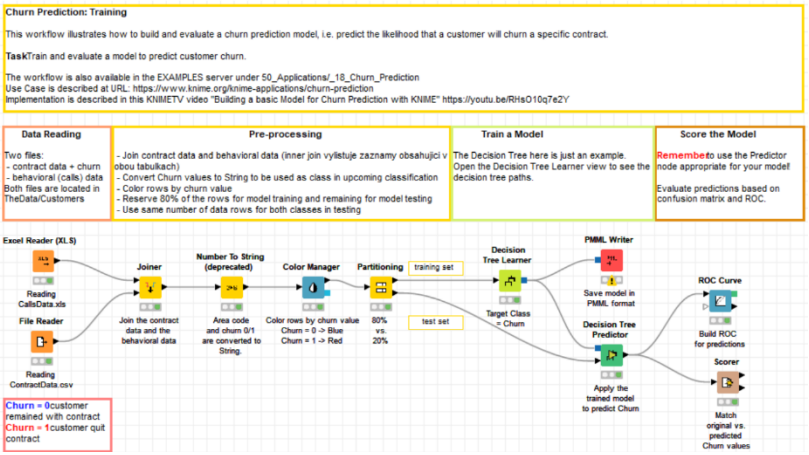
Mám tu zůstat, nebo jít?
Jako typický příklad klasifikačního prediktivního problému se v literatuře a případových studiích uvádí tzv. *churn modelling, který* slouží k predikci výpovědí klientů z firmy, resp. vede k identifikaci konkrétních zákazníků, kteří mají nebo budou mít tendenci opustit naši firmu 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*