KAPITOLA 4

Učící se stroj

Pohled do útrob predikce hypotečního rizika banky Chase

Jaký typ rizika je nelépe maskovaný? Jak se díky predikci stává z rizika příležitost? Co by si měly vzít všechny firmy za příklad z pojišťoven? Proč strojové učení vyžaduje kromě vědy i kus umění? Jaký typ prediktivního modelu je každému srozumitelný? Kdy můžeme počítačovým predikcím věřit? Proč predikce nemohly zabránit globální finanční krizi?

Tato kapitola je především romantickým příběhem o člověku jménem Dan a bance jménem Chase a o tom, jak se společně naučili vzdorovat nepřízni osudu – přesněji řečeno, jak pomocí strojového učení posílili roli predikce, která pak na oplátku snižuje riziko. Poznáme-li tento příběh, uvidíme, jak strojové učení pod pokličkou opravdu funguje.[1]

BANKA A VĚDEC

Kdysi dávno obdržel vědec jménem Dan Steinberg telefonát, z něhož plynulo, že největší americká banka čelí novému stupni rizika. Pro jeho podchycení byla ochotna vsadit na tohoto znalce strojového učení.

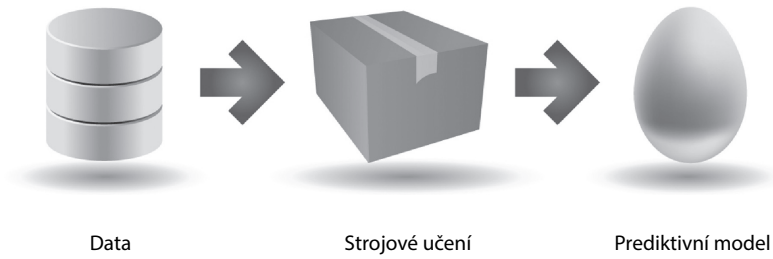
Bylo to šťastné spojení, protože Dan měl ty správné nástroje a znal i metodu, jak bance pomoci. Podnikavý vědec sestrojil komerční prediktivněanalytický systém, který převáděl špičkové výsledky výzkumů z laboratoře do komerční sféry. Banka zase přicházela s věnem

[1] Další podrobnosti k případové studii banky Chase lze najít v prezentaci z konference z roku 2005 zmiňované v poznámkách k této kapitole.

digitální kořisti: s nekonečnou řadou jedniček a nul, v nichž byly zaznamenány zkušenosti, obrovský zdroj k učení.

Banka tedy měla palivo a Dan motor. Božské spojení. Někdy, když se ve dne zasním, načrtávám obrázek:

Dospělejší profesionál by mohl své srdce otevřít formálnějším způsobem a načrtnout toto spojení tak, jak jsme ho v knize již dříve popsali:



Strojovým učením se data zpracovávají v prediktivní model.

BANKA ČELÍ RIZIKU

Finanční riziko se ke každé organizaci plíží nepozorovaně, dokonale a jednoduše maskované: z mnoha drobných ztrát, z nichž každá sama o sobě vypadá neškodně, se postupně naakumuluje jedna velká ztráta. Pod úrovní dosahu radarů proplouvají jednotlivé nešťastné případy, nudně a zcela nevzrušivě. Jsou v podstatě neviditelné.

Brzy po megafúzi, jíž se v roce 1996 stala Chase největší americkou bankou, rozpoznal její tým pro domácí finance nový stupeň rizika: obrovský nárůst počtu hypoték. Z portfolia původně šesti bank nyní měla banka Chase najednou miliony majitelů hypoték. Každý z nich představoval miniaturní kousek rizika: *mikroriziko*. A tak zavolali Danovi.

Existují paradoxně dvě zdánlivě opačné problémové varianty, jak se může držitel hypotéky vůči bance zachovat. Může se buď dostat do platební neschopnosti, nebo může hypotéku splatit plně, ale příliš brzy:

Mikroriziko A: Zákazník není schopen splácet hypotéku.

Mikroriziko B: Zákazník hypotéku splatí *dříve* – refinancováním u konkurenční banky, nebo prodejem domu. Předčasné splacení hypotéky představuje pro banku ztrátu, protože přichází o výběr plánovaných budoucích úroků.

Tyto ztráty se označují jako „mikro", protože pro banku není jeden zákazník u hypotéky zase tak velký obchod. Mikroztráty se však mohou nahromadit. Ve finančním světě se slovo *riziko* nejčastěji kryje se slovem *úvěrové riziko*, tj. mikroriziko A, při němž se nesplacený dluh stává nevymahatelným a je jednou provždy odepsán. Pokud jsou však vaší potravou platby úroků, nejde v případě naplnění mikrorizika B také o žádnou hostinu. Vaše banka nestojí o to, abyste jí přestali dlužit.[2]

PREDIKCE SNIŽUJE RIZIKO

Při většině diskuzí o rozhodování se předpokládá, že rozhodnutí činí pouze vyšší management nebo že záleží pouze na jeho rozhodnutích. To je nebezpečný omyl.
– Peter Drucker, americký pedagog a spisovatel narozený v roce 1909

Hypoteční portfolio banky Chase čelilo rizikovým faktorům s dopadem jdoucím do stovek milionů dolarů. Je to jako na pláži: každé zrnko písku je jedním z milionu mikrorizik. Je-li žádost o hypotéku označena známkou „nízké riziko" a schválena, proces řízení rizika tím ve skutečnosti teprve začal. Portfolio uzavřených hypoték banky se musí chovat jako stádo dojných krav na farmě. Důvod? Vždy někde číhá riziko. Na světě čekají miliony hypoték na rozhodnutí, které z nich je vhodné prodat jiným bankám, které se pokusit udržet při životě a u kterých přistoupit na refinancování s nižší úrokovou sazbou.

Prediktivní analytika slouží jako protijed na ničivou akumulaci mikrorizik. Prediktivní analytika bdí a výhledově známkuje každé mikroriziko tak, aby s ním daná organizace mohla něco podniknout.

[2] Podobně nemají společnosti vydávající kreditní karty radost, pokud vždy včas splácíte jistinu a neplatíte jim žádný úrok.

To není nic nového. Jde o běžný postřeh, známý již od samých počátků prediktivní analytiky. Predikování zákaznického rizika je dobře známo pod tradičním pojmem kreditní skóre, které vydává firma FICO či kreditní agentury, například Experian. Původ kreditního skóre se datuje do roku 1941, dnes se tento pojem stal součástí běžného žargonu. Jeho zavedení bylo podkladem pro vznik prediktivní analytiky a jeho úspěch pomohl prediktivní analytiku zviditelnit. Dnešní riziková skóre se mnohdy generují pomocí stejné prediktivní modelovací metody, na jaké stojí prediktivněanalytické projekty.

Výhody boje s rizikem pomocí prediktivní analytiky lze snadno demonstrovat. Ačkoli predikce může být součástí projektu, sama o sobě pouze aritmeticky počítá hodnotu realizovanou, když na predikci dojde. Představte si, že máte banku s tisíci nesplácených půjček, z nichž deset procent považujete za nedobytné. U každého z deseti nesplácejících dlužníků je budoucnost zahalena obvyklým závojem: nevíte, kteří z nich se zrovna ukážou jako ti špatní.

Řekněme, že riziko z každého úvěru ohodnotíte pomocí efektivního prediktivního modelu. Někteří dlužníci obdrží vysoce riziková skóre, jiní zase nízká. Jsou-li tato riziková skóre dobře propočtena, může horní polovina predikovaná jako rizikovější vyjevovat oproti průměru téměř dvakrát vyšší pravděpodobnost, že se z nich vyklubou neplatiči – abychom byli realističtější, řekněme o sedmdesát procent více, než je celková míra nesplácení.

To musí znít vašim bankéřským uším jako rajská hudba. Špetkou aritmetiky jste rozdělili portfolio na dvě poloviny, jednu se sedmnáctiprocentní mírou nesplácení (o 70 % více než celková míra 10 %) a druhou s tříprocentní mírou nesplácení (jelikož 17 % a 3 % dávají v průměru 10 %).

Vysoce rizikové půjčky: nebude schopno splácet 17 % dlužníků.
Nízko rizikové půjčky: nebudou schopny splácet 3 % dlužníků.

Své podnikání jste právě rozdělili na dva zcela oddělené světy: jeden bezpečný a jeden hazardní. Nyní víte, kam zaměřit svou pozornost.

S touto premisou vypočítala banka Chase *makroriziko* ve velkém měřítku. Vložila svou důvěru do predikce, jíž tak svěřila do rukou rozhodování o stovkách milionů dolarů. A její příběh skončí šťastně pouze tehdy, pokud tato predikce vyjde – pokud data přinesou své ovoce v té obrovské nejistotě, kterou představuje budoucnost.

Predikce představuje vrcholné dilema. Jak lze i při významné znalosti minulosti ospravedlnit důvěru v to, že technologie správně rozpozná neznámou budoucnost?

Než se dostaneme k tomu, jak predikce fungují, povězme si ještě něco málo o riziku.

RISKANTNÍ PODNIKÁNÍ

Revoluční myšlenkou, oddělující moderní dobu od minulosti, je zvládnutí rizika: uvědomění si, že budoucnost není jen rozmar bohů a že lidé nejsou jen pasivní hříčkou přírody. Do doby, než lidské bytosti odhalily průchod přes tuto hranici, byla budoucnost zrcadlem minulosti či jen temnou doménou zázraků a jasnovidců, jež měli monopol na znalost toho, co přijde.
– Peter Bernstein, z knihy *Against the Gods: The Remarkable Story of Risk*

Nic takového jako špatné riziko neexistuje; existuje jen špatná cenotvorba.
– Stephen Brobst, CTO Teradata

Ovšemže banky nenesou břímě řízení rizika celé společnosti. Významná je také role pojišťoven. Hlavní úlohou bank je spíše analýza dat umožňující kvantifikovat riziko tak, aby mohlo být efektivně rozloženo. Eric Webster, viceprezident pojišťovny State Farm Insurance, to hodnotí brilantně takto: „Pojištění není nic jiného než správa informací. Je to sdílení rizika a ten, kdo dokáže s informacemi zacházet nejlépe, má významnou konkurenční výhodu." Jednoduše řečeno, takové firmy vkročily na pole predikce.

Pojišťovnictví učinilo z řízení rizika umění. Douglas Hubbard v knize *The Failure of Risk Management* poukazuje na to, co je bolestné pro všechny organizace mimo pojišťoven: „Kromě oblasti, která je striktně považována za pojišťovnictví, neexistuje žádná tak certifikovaná a regulovaná profese, jako je pojistněmatematická praxe."

Navzdory tomu může kterákoli organizace riziko řídit tak, jak to dělají pojišťovny. Jak? Aplikováním prediktivní analytiky pro predikci nepříznivých událostí. Prediktivní model plní u každé organizace v zásadě stejnou funkci jako *pojistná matematika* u pojišťovny: oceňuje jednotlivce podle pravděpodobnosti negativního výsledku. Prediktivní analytiku můžeme proto definovat těmito základními podmínkami.[3]

Zde je původní definice:

***Prediktivní analytika** je technologie, která na základě minulých zkušeností (dat) učí predikovat budoucí chování jednotlivců a napomáhat tak kvalitnějšímu rozhodování.*

Organizace se díky prediktivní analytice efektivně učí, *jak snížit riziko předvídáním mikrorizik*. Zde je alternativní definice zmiňující riziko:

[3] Funguje to i opačně: zatímco klasické pojistněmatematické metody obsahují manuální prvky, například tabulky a analýzu, pojišťovny tyto praktiky výrazně rozšiřují o prediktivní modelování, a dosahují tak lepších prediktivních výsledků. Používají automatizovanější a zdokonalené metody prediktivního modelování, a ty jsou i náplní této kapitoly.

Prediktivní analytika *je technologie, která se na základě minulých zkušeností (dat) učí řídit mikroriziko.*

Platí obě definice, protože každá z nich implikuje tu druhou.

Stejně jako oportunistický mladý podnikatel zaměstnávající Toma Cruise rozjíždí svůj byznys ve filmu *Riskantní podnik* z roku 1983, jsou všechny podniky rizikové. A stejně jako pojišťovny, všechny organizace profitují z ohodnocování a predikování rizika špatného chování, kam patří platební neschopnost, vypovídání smluv, nehody, podvody a zimní kalamity. Prediktivní analytika tímto způsobem přetavuje riziko v příležitost.

Kde by mohlo být v masové ekonomii řízení rizika důležitější než ve světě hypoték? Hypoteční byznys, měřený v bilionech dolarů, slouží jako finanční základ vlastnictví domů, jako punc rodinné prosperity.

A hypotéky jsou při svém významu všeobecně považovány za ústřední katalyzátor nedávné finanční krize neboli Velké deprese. I mikrorizika jsou podstatná. Nejsou-li ošetřena, hrozí, že se na sebe nabalí jako sněhová koule. Naší nejlepší možností je naučit se je predikovat.

UČÍCÍ SE STROJ

Proces učení se z dat není až tak složitý, jak se může zdát.[4]

Začněme skromnou otázkou: Jakým způsobem lze od sebe nejsnadněji začít odlišovat vysoce rizikové a nízko rizikové hypotéky? Který z faktorů o hypotékách nejvíce vypovídá?

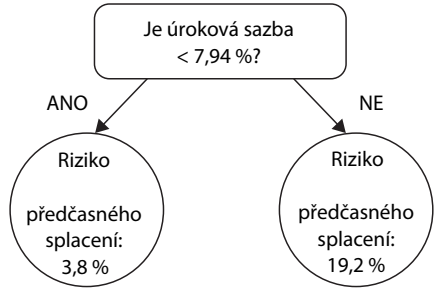
Danův učící se systém učinil v datech banky Chase objev: *Je-li úroková míra hypotéky pod 7,94 %, je riziko předčasného splacení 3,8 %; v opačném případě je riziko 19,2 %.*[5]

Překresleme si situaci do obrázku.

Jaký rozdíl! Pouze na základě úrokových sazeb jsme rozdělili balík hypoték na dvě skupiny: jednu pětkrát rizikovější než druhou, pokud jde o pravděpodobnost předčasného

[4] Všeobecné povědomí o strojovém učení je na vzestupu. Katedra počítačových věd Stanfordovy univerzity (jedna ze tří špičkových ve Spojených státech) poprvé zpřístupnila zdarma svůj kurz Strojové učení v roce 2011 a počet přihlášených z celého světa přesáhl sto tisíc. Tento úspěch inspiroval profesora tohoto projektu, Andrewa Nga, aby se podílel na založení projektu Coursera, nabízejícího zdarma internetové kurzy z různých předmětů.

[5] Studie popsaná v této kapitole byla pořízena na základě 21 816 hypoték se sazbou fixovanou minimálně na patnáct let, které byly relativně nové, mezi prvním a čtvrtým rokem trvání, a proto s nadprůměrným rizikem předčasného splacení, jelikož dlužníci splácející již déle než čtyři roky pravděpodobněji dodrží plánované splátky. Všimněte si, že úrokové míry odpovídají konci devadesátých let, kdy se tento projekt uskutečnil a kdy byla shromážděna data.



splacení (zákazník neočekávaně splatí celý dluh, čímž banku připraví o budoucí výdělek z plateb úroku).

Tento objev je cenný, ačkoli není úplně překvapivý. Majitelé domů platící vyšší úroky mají vyšší tendenci hypotéku refinancovat nebo prodat dům než ti, kdo platí nižší úroky. Celkem jsme to mohli očekávat, nyní to však máme potvrzeno empiricky a daný efekt je přesně kvantifikován.

Strojové učení provedlo svůj první krok.

TVORBA UČÍCÍHO SE STROJE

Už jsme v polovině cesty. Věřte tomu nebo ne, už jsme jen krok od toho, abychom se stali svědky plné podstaty strojového učení – schopnosti generovat z dat prediktivní model, učit se z příkladů a vytvářet elektronického Sherlocka Holmese, jenž si člověka měří přísným okem a předvídá. Už jsme pouhý kousek od jádra jedné z nejvíce fascinujících věcí ve vědě a nejodvážnějších lidských ambicí: *automatizace učení*.

Není k němu potřeba žádné složité matematiky ani počítačového kódu; celý zbytek se dá vysvětlit v podstatě dvěma slovy. Nejprve se však na okamžik podívejme na to, jak je tento vědecký problém plně definován.

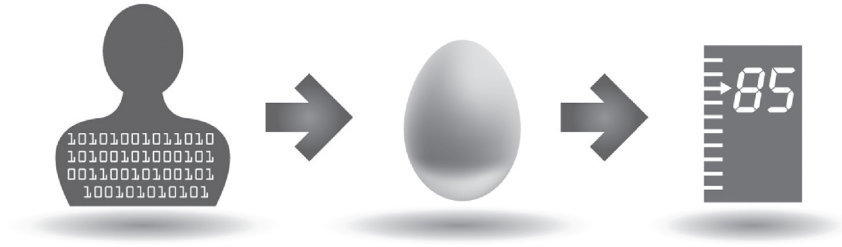
Náš dosavadní poznatek, že úroková míra predikuje riziko, vyzýval k vytvoření velmi striktního prediktivního modelu. Rozděluje každou hypotéku do jedné ze dvou prediktivních kategorií: na hypotéku s vysokým rizikem a na hypotéku s nízkým rizikem. Protože bere o jednotlivci v úvahu jediný faktor, neboli predikční proměnnou, můžeme ho nazývat modelem *s jednou proměnnou* (neboli *univariantní model*). Všechny příklady v tabulkách v předcházející kapitole, přinášející bizarní a překvapivé poznatky, jsou univariantní – každý z nich závisí na jedné proměnné, jako je například plat, e-mailová adresa nebo kreditní skóre.

Musíme se však posunout *k více proměnným* (*k multivariantnímu modelu*). Proč? Efektivní prediktivní model musí jistě brát do úvahy současně více faktorů, ne jen jeden. A v tom spočívá háček.

Zde je pro připomenutí definice:

Prediktivní model *je mechanismus predikující chování jednotlivce, například zda klepne myší, koupí, bude lhát nebo zemře (nebo splatí předčasně hypotéku). Jako vstup přijímá charakteristické vlastnosti (proměnné) jednotlivce a jako výstup dodává prediktivní skóre. Čím je toto skóre vyšší, tím pravděpodobnější je, že daný jednotlivec bude vykazovat predikované chování.*

Po implementaci formou strojového učení prediktivní model predikuje výsledky pro každého jednotlivce zvlášť:



Charakteristické vlastnosti jednotlivce | Prediktivní model | Prediktivní skóre

Představme si následujícího zákazníka splácejícího hypotéku:

Dlužník: Sally Smithersová
Hypotéka: 174 000 USD
Hodnota nemovitosti: 400 000 USD
Typ nemovitosti: Rodinný domek (jednogenerační)
Úroková sazba: 8,92 %
Roční příjem dlužníka: 86 880 USD
Čisté jmění: 102 334 USD
Kreditní skóre: vysoké
Pozdní splátky: 4
Věk: 38
Stav: vdaná/ženatý
Vzdělání: vysokoškolské
Jak dlouho bydlel(a) v předchozím bydlišti: 4 roky
Profese: obchodní manažer
Samostatně výdělečně činný: ne
V dosavadním zaměstnání: 3 roky

Toto jsou predikční proměnné, charakteristiky vstupující do prediktivního modelu. Úlohou tohoto modelu bude brát všechny takové proměnné v úvahu a vyprodukovat z nich jediné prediktivní skóre. Nazývejme to výpočtem nové, *vyšší proměnné*. Model tak vzájemným skloubením všech těchto dílčích informací vyprodukuje jedno výsledné skóre.

To je podstatou strojového učení. Vaším posláním je pak naprogramovat nemyslící notebook tak, aby zanalyzoval data o jednotlivcích a automaticky z nich vytvořil prediktivní model o více proměnných. Pokud se vám to podaří, bude se váš počítač učit predikovat.

UČENÍ ZE ŠPATNÝCH PŘÍKLADŮ

Zkušenost je pojem, kterým každý nazývá své chyby.
– Oscar Wilde

Mé renomé roste s každým neúspěchem.
– George Bernard Shaw

Na strojové učení klademe další požadavek. Pro získání znalostí ze směsice dobrých i špatných příkladů je nutno vymyslet vhodnou metodu, která si vezme poučení jak z kladných, tak i ze záporných výsledků, jež jsou součástí dat. Některé dřívější hypotéky byly splaceny podle plánu, jiné byly bohužel (pro banku) splaceny předčasně. Obou těchto typů dat je však potřeba využít.

Abychom dokázali predikovat, potřebujeme zodpovědět otázku: Jak lze předem rozeznat kladné a záporné případy? Naučit se replikovat minulé úspěchy tím, že budeme brát v úvahu pouze pozitivní případy, to nebude fungovat.[6] Zásadní význam mají negativní příklady. Při učení se predikci stojíme o chyby.

JAK FUNGUJE STROJOVÉ UČENÍ

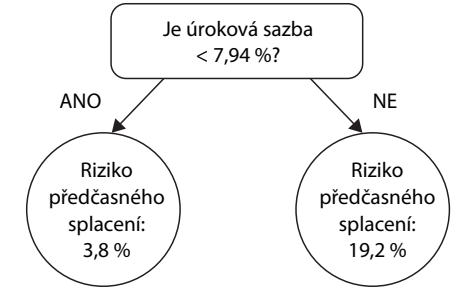
A nyní zde máme intuitivní, elegantní odpověď na toto velké dilema, další krok učení, jímž se dostaneme od prediktivního modelování o jedné proměnné k více proměnným, vedeni jak pozitivními, tak negativními případy: *pokračujte dále*.

Zatím jsme vytvořili dvě skupiny podle výše rizika. Nyní najděte ve skupině s nízkým rizikem další faktor, který nejlépe tuto skupinu dále dělí do dvou podskupin s navzájem odlišným rizikem. Poté udělejte totéž se skupinou s vysokým rizikem. A pak pokračujte stejným způsobem se vzniklými podskupinami. Rozdělujte a panujte a pak znovu rozdělujte ještě chvíli, na menší a menší skupiny. Ale nezacházejte s rozdělováním zase příliš daleko.

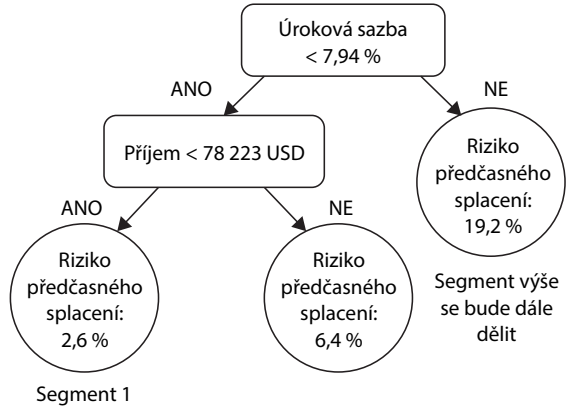
[6] Analýza pouze pozitivních případů se někdy nazývá profilování a klonování zákazníků.

Tato učicí metoda, zvaná *rozhodovací strom*, není jediným způsobem, jak vytvořit prediktivní model, bývá však v praxi stabilně používána jako nejčastější nebo druhá nejčastější, protože je v poměru ke své účinnosti relativně jednoduchá. Neprodukuje sice vždy ty nejpřesnější prediktivní modely, ale protože jednodušší modely bývají přijatelnější než nesrozumitelné matematické vzorce, může být výborným východiskem nejen pro studium prediktivní analytiky, ale i pro začátek každého prediktivněanalytického projektu.

Pojďme náš rozhodovací strom rozvíjet. Dosud jsme dospěli do následujícího stavu:



Nyní najděme predikční proměnnou, která bude dále dělit skupinu zákazníků s nízkým rizikem na levé straně diagramu. V této datové množině bere Danův softwarový rozhodovací strom v úvahu dlužníkův příjem:[7]



[7] Tento rozhodovací strom, stejně jako dále uvedené, predikující rovněž předčasné splátky hypotéky, jsou zjednodušené ve smyslu, že se nezabývají zpracováním neznámých hodnot. Nemusejí být například známy příjmy některých majitelů hypoték. Pro tyto neznámé hodnoty se v rozhodovacím stromu používá alternativní zástupná proměnná, s jejíž pomocí se metoda rozhoduje, zda půjde v rozhodovacím bodě doleva či doprava. I když ukázkové rozhodovací stromy v této kapitole vycházejí z reálných dat, nejsou přímo z projektu hypoték banky Chase.