Deloitte.

Analytics Applied.

Propenzitní modelování

Veronika Počerová

10. 4. 2015

motivace

definice

Prediktivní analytika je disciplína, která využívá metod Data Miningu k tomu, aby na základě historického chování sledovaného jevu mohla předpovídat jeho budoucí průběh.

propenzitní modelování

Metoda prediktivní analytiky

Statistická analýza zákazníků/pacientů/klientů/zaměstnanců

- Kdo jsou
- Jak se chovají

typické problémy

- Jak je pravděpodobné, že si koupí produkt?
- Chystá se v nejbližší době odejít?
- Jaká je šance, že ho přesvědčím, aby neodešel?
- Jaká je šance, že přestane splácet?
- Jak je pravděpodobné, že má nemoc?
- Vyplatí se ho najmout?
- Jaká je šance, že odejde během prvního roku?

teoretická část

konstrukce modelu

1. Stanovení cílové proměnné (targetu), stanovení segmentů bázové množiny
2. Příprava dat
3. Sampling
4. Binování
5. Weight of evidence
6. Modelování
7. Vyhodnocení

stanovení cílové proměnné (targetu)
stanovení segmentů bázové množiny

- Interpretace obchodního požadavku do modelovací řeči.
- Target je událost, kterou chceme predikovat. Vyjadřuje se jako úspěch (jednička) nebo neúspěch (nula) na úrovni zákazníka.

- Pokud existují přirozené segmenty v bázové množině, na kterých se budou prediktory chovat odlišným způsobem, pak tyto segmenty popíšeme a vyvineme pro ně samostatné modely.
- Rozhodnutí o segmentaci pro modely je klíčové s ohledem na budoucí výkonnost PtB modelu.

příprava dat

- Identifikace a napočtení prediktorů
  - Každý prediktor představuje více či méně komplexní výpočet, který transformuje data ze zdrojových tabulek do výsledné spojité či kategoriální proměnné, která na úrovni pozorování (tj. pro konkrétního klienta v konkrétním čase) popisuje realitu. Prediktor by měl být relevantní vzhledem k tomu, co chce model predikovat
  - poměry, trendy, rozdíly, agregované atributy atd.
- Napočítání cílové proměnné (targetu) – důležitá časová souslednost
- KO kritéria

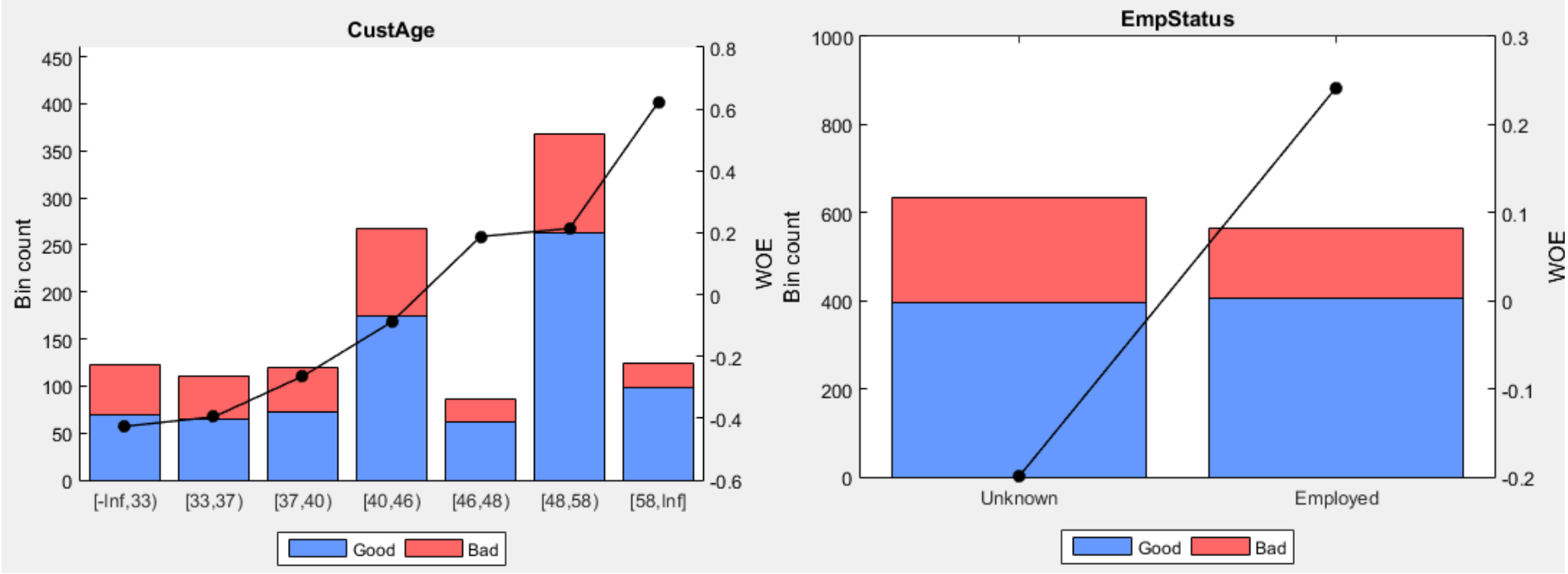
sampling

- Náhodné rozdělení bázové množiny na tréninkovou (learning) a testovací (testing) množinu, zde obecně platí pravidlo 80:20
  - 80% dat využijeme pro trénování modelu a 20% dat pro testování.

- Pokud jsou data příliš veliká, může se přistoupit k technice známé jako **target dependent sampling:**
  - Do vzorku se vezmou všechna pozorování prodeje (event) a náhodný výběr části pozorování bez prodeje (non-event). Každému pozorování prodeje se potom přiřadí váha 1, zatímco pozorování bez prodeje se přiřadí váha odpovídající podílu mezi skutečným a zahrnutým počtem pozorování.

binování

- Binování je kategorizace proměnných – pro spojité atributy vytváříme vhodný počet skupin (intervalů), diskrétní hodnoty můžeme spojovat do logických celků. Binování řeší mj chybějící hodnoty a extrémní hodnoty v datech.
- Základním kritériem pro rozdělení do skupin (binů) je vztah k targetu. Binování by mělo vystihovat trend.



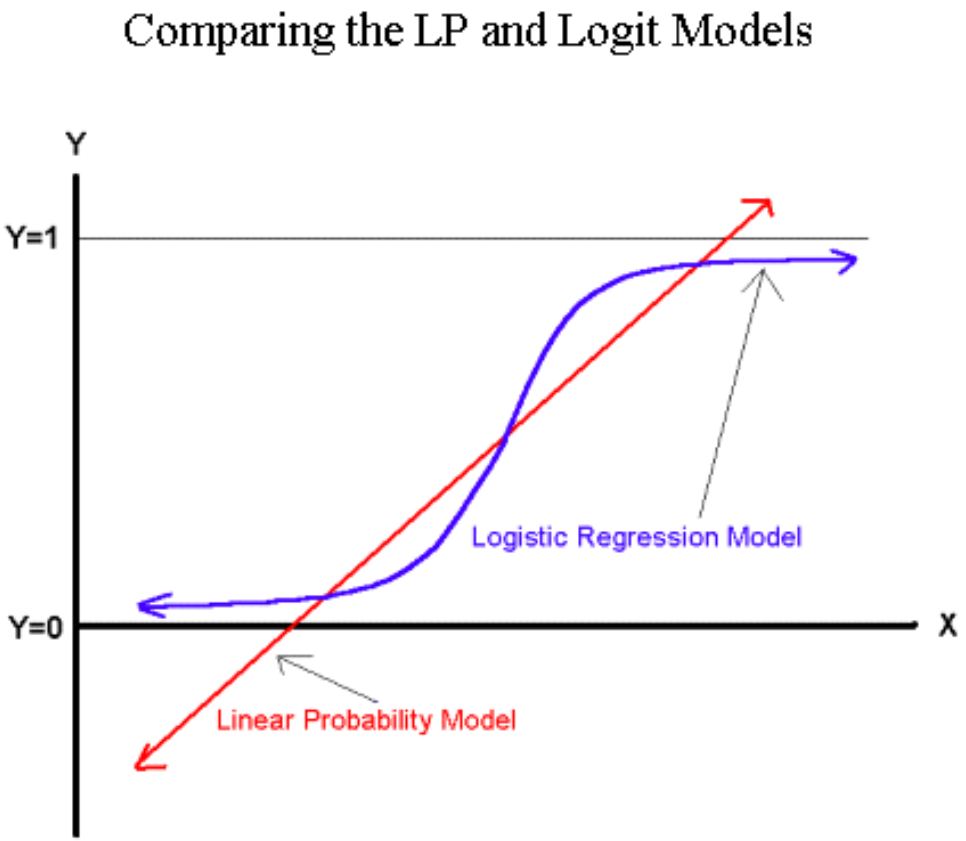
weight of evidence

- Weight of evidence (WoE) je transformace kategoriální proměnné (v našem případě nabinované proměnné), která každé kategorii (každému binu) přiřazuje reálné číslo.
- Transformace WoE zkoumá prediktivní sílu jednotlivých binů jedné proměnné vzhledem k výsledku (targetu). WoE bere v úvahu nejen podíl pozitivních výsledků, ale také rozdělení pozitivních a negativních výsledků všech skupin (binů).
- Proměnné po transformaci WoE dále vstupují jako spojité proměnné do logistické regrese.

$$WoE_i = \left[\ln\left(\frac{n_{1i}}{n_{0i}}\right) - \ln\left(\frac{n_1}{n_0}\right)\right]$$

modelování

- Logistická regrese
- Neuronové sítě
- Rozhodovací stromy



logistická regrese

náhodná veličina *Y(**x**) s alternativním rozdělením* - nabývá hodnot 0 a 1

$$\pi(\boldsymbol{x}) = \boldsymbol{E}\big(Y(\boldsymbol{x})\big)$$

$$odds(\boldsymbol{x}) = \frac{\pi(\boldsymbol{x})}{1 - \pi(\boldsymbol{x})}$$

$$logit(\boldsymbol{x}) = \ln\big(odds(\boldsymbol{x})\big) = \ln\left(\frac{\pi(\boldsymbol{x})}{1 - \pi(\boldsymbol{x})}\right) = \boldsymbol{\beta x}$$

$$\pi(\boldsymbol{x}) = \frac{e^{\boldsymbol{\beta x}}}{1 + e^{\boldsymbol{\beta x}}}$$

$\boldsymbol{\beta}$ značí vektor neznámých parametrů

pro odhad se tradičně používá metoda maximální věrohodnosti

odhad parametrů metodou maximální věrohodnosti

Y_i: $i = 1, \dots, n$

β_j: $j = 1, \dots, k$

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} Y_i\left(\ln\left(\pi(\boldsymbol{x_i})\right)\right) + (1 - Y_i)\left(\ln\left(1 - \pi(\boldsymbol{x_i})\right)\right)$$

Parciální derivace podle β_j:

$$\sum_{i=1}^{n} {x_i}^{(j)}(Y_i - \pi(\boldsymbol{x_i})) = 0 \quad j = 1, \dots, k$$

Newton-Raphsonův iterační algoritmus

$\boldsymbol{\beta}^{(0)}$ *libovolně*

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \left(H^{(t)}\right)^{-1} * q^{(t)}$$

$$q^{(t)} = \nabla L\left(\boldsymbol{\beta}^{(t)}\right),\ {h_{i,j}}^{(t)} = \frac{\partial L^2\left(\boldsymbol{\beta}^{(t)}\right)}{\partial \beta_i \beta_j} \quad i, j = 1, \dots, k$$

výběr prediktorů

- Existují tři základní typy výběru prediktorů:
  - Forward selection začíná s modelem bez prediktorů a testuje, jestli přidání konkrétního prediktoru model zlepší. V každém kroku je vybrán prediktor, který model zlepší nejvíce. Takto se postupuje, dokud existují prediktory, jejichž přidání model zlepšuje.
  - Backward elimination začíná s modelem, který obsahuje všechny prediktory a postupně testuje, jestli ubrání konkrétního prediktoru model zlepší.
  - Kombinace forward-backward provádí v každém kroku obojí – nejdříve přidá nejlepší prediktor a pak testuje všechny prediktory v modelu, jestli by vyřazení některého z nich nepomohlo model nadále vylepšit. Takto se pokračuje, než dospějeme do fáze, kdy přidání nebo odebrání žádného prediktoru model nezlepší.

Testování nulovosti subvektoru $\boldsymbol{\beta}'$:

$H_0: \beta'_1 = \beta'_2 = \cdots = \beta'_e = 0$

$G = 2(\mathrm{L}(\boldsymbol{\beta}) - L(\boldsymbol{\beta}'))\sim\chi^2(e)$

$P\ value = P(\chi^2(e) > \mathrm{G})$

vyhodnocení modelu - ROC křivka, Giniho koeficient

- Mějme data $(\boldsymbol{x_i}, y_i)$
- Každému prvku $(\boldsymbol{x_i}, y_i)$ přiřadíme skóre, tj. číslo $s_i \in R$. Čím větší očekáváme pravděpodobnost $P\ (Y_i = 1)$, tím větší skóre pro prvek $(\boldsymbol{x_i}, y_i)$.
- Seřadíme prvky sestupně podle jejich skóre. Sestavíme dvě distribuční funkce – jednu pro prvky s $y_i = 0$, druhou pro prvky s $y_i = 1$.

$$F_0(s) = \frac{1}{n_o} \sum_{i=1}^{n} \boldsymbol{I}_{(-\infty < s_i \leq s)}\ (1 - y_i)$$

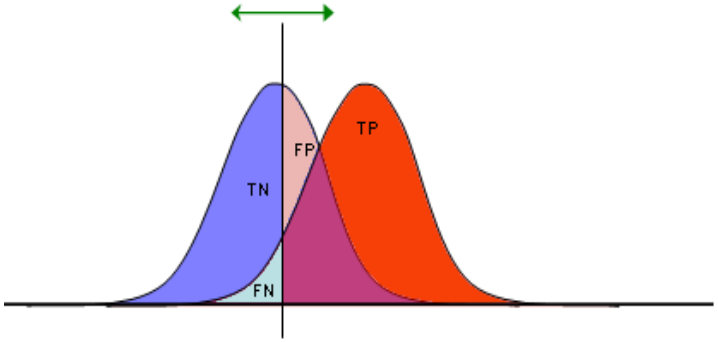
$$F_1(s) = \frac{1}{n_1} \sum_{i=1}^{n} \boldsymbol{I}_{(-\infty < s_i \leq s)}\ y_i$$

n_0… počet prvků, pro které $y_i = 0$

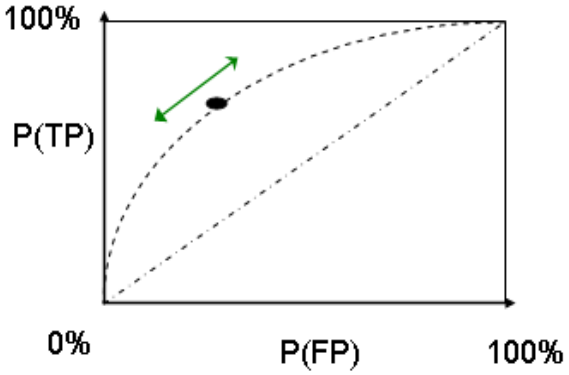
n_1… počet prvků, pro které $y_i = 1$

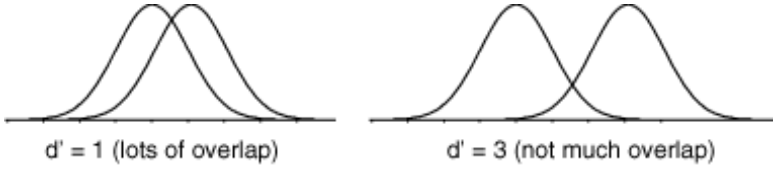
vyhodnocení modelu - ROC křivka, Giniho koeficient

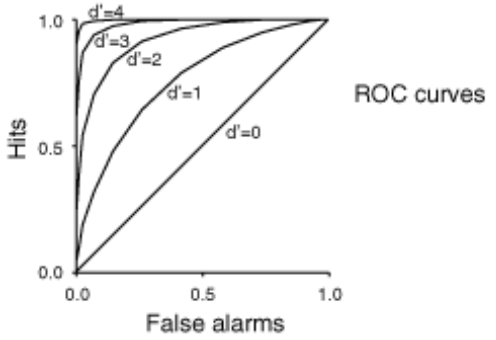
- Distribuční funkce nám říkají, jaká je pravděpodobnost, že náhodně vybraný klient, co si produkt nekoupí, bude mít skóre horší než s, respektive, že náhodně vybraný klient, který si produkt koupí, bude mít skóre menší než s.



TP	FP
FN	TN
1	1



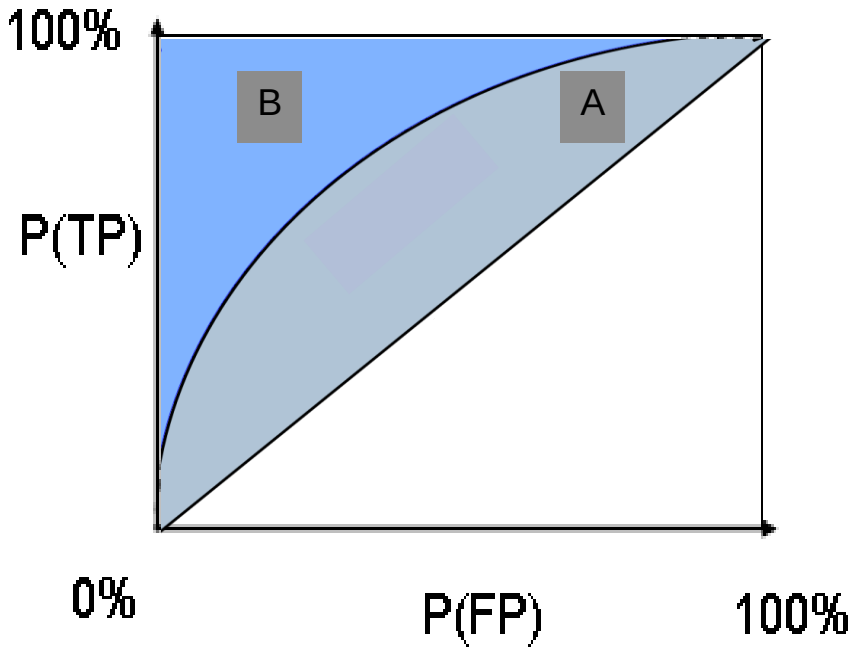




vyhodnocení modelu - ROC křivka, Giniho koeficient

$$GINI = \frac{A}{A+B} = 1 - 2B$$

$$GINI = 1 - 2\int_0^1 F_1(s)dF_0(s)$$

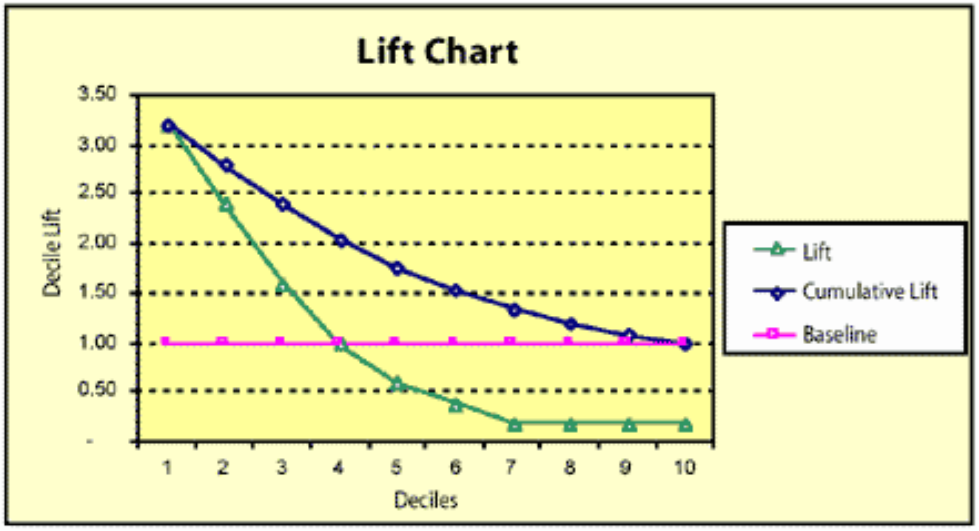


Giniho koeficient nabývá hodnot v intervalu [-1,1]

- Záporný Giniho koeficient znamená, že vysoké skóre indikuje spíše y = 0, tedy je model postaven obráceně. Giniho koeficient v intervalu [0;0,3] označuje podprůměrný model. Giniho koeficient v intervalu [0,3;0,6] značí průměrný model. Vyšší Giniho koeficient než 0,6 znamená velmi dobrý model.

vyhodnocení modelu - Lift

- Mějme data $(\boldsymbol{x_i}, y_i)$
- Každému prvku $(\boldsymbol{x_i}, y_i)$ přiřadíme skóre, tj. číslo $s_i \in R$. Čím větší očekáváme pravděpodobnost $P\ (Y_i\ =\ 1)$, tím větší skóre pro prvek $(\boldsymbol{x_i}, y_i)$.
- Seřadíme prvky sestupně podle skóre a rozdělíme je do decilů. Hodnoty $\frac{TP\prime}{TP}$ se pro každý decil zanesou do grafu.
- Častěji se používá Cumulative Lift – v tomto případě nezaneseme do grafu bod $\frac{TP\prime}{TP}$ pro klienty pouze v druhém decilu, ale hodnotu společnou pro klienty v obou prvním a druhém decilu. Takto bychom postupovali až do posledního decilu.



vyhodnocení modelu – Kolmogorov-Smirnov

- Vyjdeme-li z definice ROC, KS definujeme jako supremum rozdílu výše zavedených distribučních funkcí:

$$KS = \sup_{s \in R} \{F_0(s) - F_1(s)\}.$$

- Kolmogorov-Smirnovova statistika běžně nabývá hodnot z intervalu [0,1], kde hodnota KS = 1 značí dokonalou diverzifikační schopnost modelu, KS = 0 značí nulovou diverzifikační schopnost.
- Neexistuje žádný pevný vztah mezi Kolmogorov-Smirnovovou statistikou a Giniho koeficientu. Lze dokázat, že za předpokladu, že skóre prvků jsou normálně rozdělená (tvoří náhodný výběr z normálního rozdělení), platí mezi Kolmogorov-Smirnovovou statistikou a Giniho koeficientem téměř lineární závislost.

$$KS \approx \frac{\sqrt{2}}{2} GINI$$

praktická část

model na kreditní karty pro středně velkou českou banku

- Z časového hlediska bereme v potaz posledních 12 měsíců (pokrytí sezónnosti). Bázová množina je tedy souborem takovýchto kliento-měsíců. Kritéria pro vyřazení jsou v našem případě: a) má kreditní kartu, b) neprojde na KO kritéria risku.

model na kreditní karty pro středně velkou českou banku

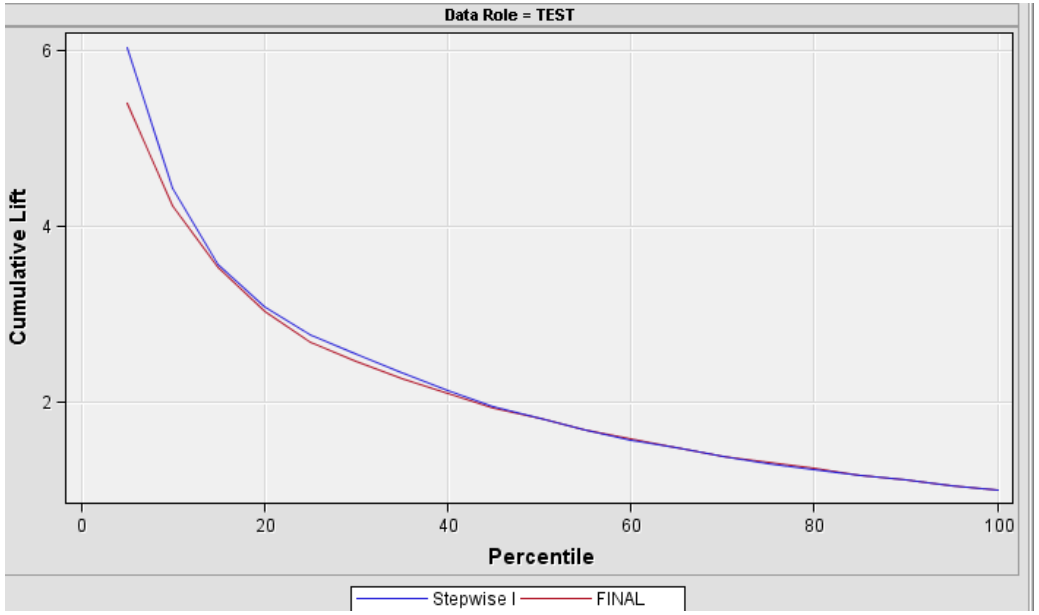
- Cílová proměnná (target) pro pozorování (K, M) je definovaná jako prodaná kreditní karta v měsíci M+2. V časovém období po pozorování bude případná nabídka komunikována klientovi v měsíci M+1 a klient bude mít minimálně 30 dnů na to, aby prošel schvalovacím procesem. Nová smlouva bude pozorována nejpozději na konci měsíce M+2.
- Segmenty pro pozorování (K, M) jsme stanovili dva. S předschválenou nabídkou: klient K má v měsíci M předschválenou nabídku na kreditní kartu. Bez předschálenky: klient K nemá v měsíci M předschválenou nabídku na kreditní kartu. Chování prediktorů předpokládáme natolik odlišné, že vyúsťuje v potřebu na samostatné modely.

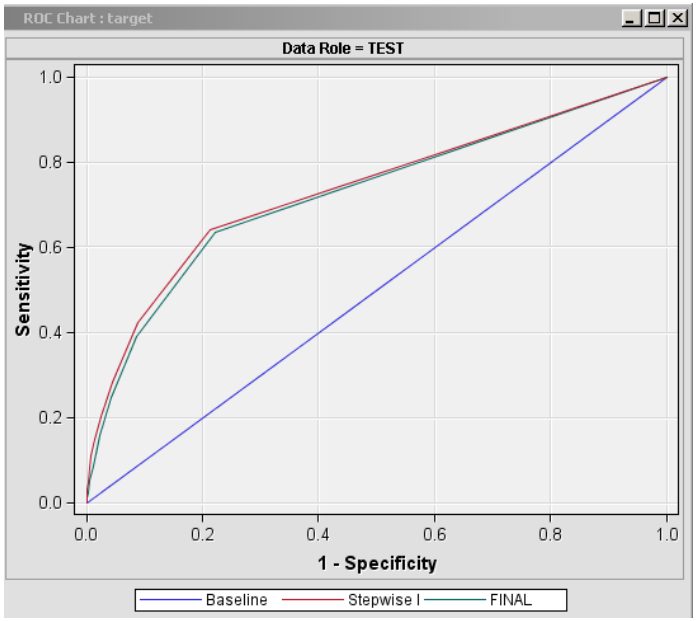
- Konverze na klientech s předschválenou nabídkou 1.317 %
- Konverze na klientech bez předschválené nabídky 0.759 %

výsledky modelu – bez předschválené nabídky

Giniho koeficient	0,59
Cumulative Lift	4,23

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept	1,000	-4,895	0,017	82431,640	<,0001		0,007
WOE_PROD_OVD_CONTRACT_AGE	1,000	0,197	0,024	66,840	<,0001	0,040	1,218
WOE_CUSTOMER_CLASS_CD	1,000	0,233	0,035	44,370	<,0001	0,073	1,262
WOE_PROD_CA_BALANCE_MIN_1M	1,000	0,274	0,030	85,700	<,0001	0,118	1,315
WOE_M0_DC_POS_AMT	1,000	0,341	0,033	108,790	<,0001	0,097	1,406
WOE_COMM_BRANCH_VISITS_CNT_L1M	1,000	0,404	0,028	206,090	<,0001	0,161	1,498
WOE_PARTY_INCOME_AMT_APPL	1,000	0,423	0,030	193,530	<,0001	0,125	1,527
WOE_M3_ALL_TRANS_CNT	1,000	0,430	0,024	335,390	<,0001	0,191	1,537
WOE_T_ACTIVE_LIMIT_AMT	1,000	0,441	0,030	214,200	<,0001	0,089	1,555
WOE_OPEN_ACCT_12MTH_CNT	1,000	0,471	0,023	422,630	<,0001	0,162	1,601
WOE_REGION_PSC	1,000	0,519	0,074	49,700	<,0001	0,078	1,680

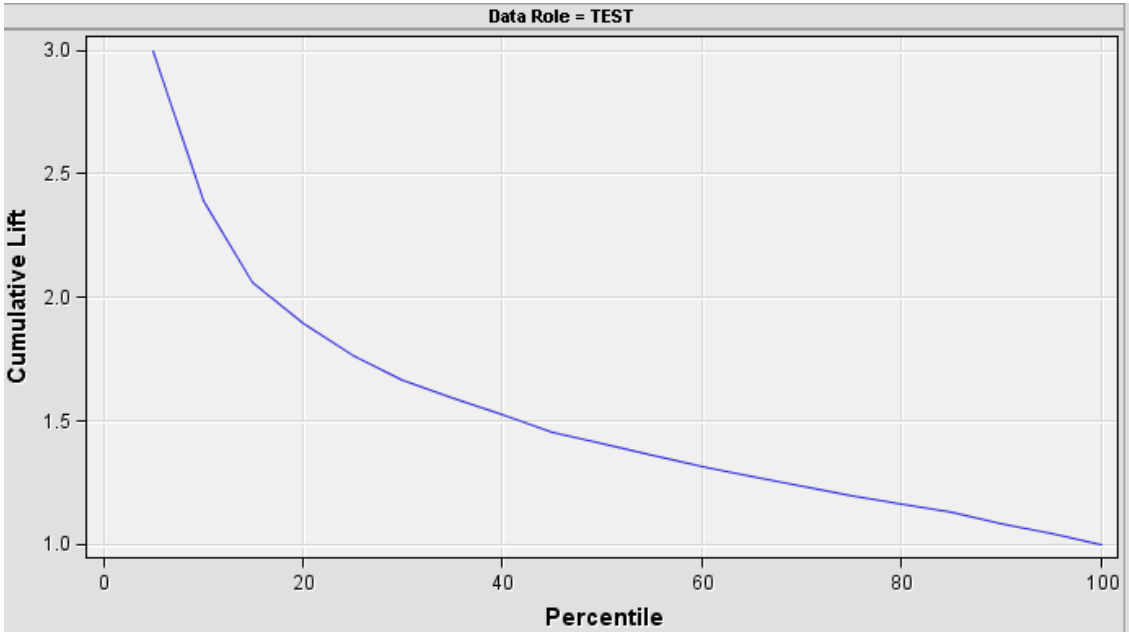


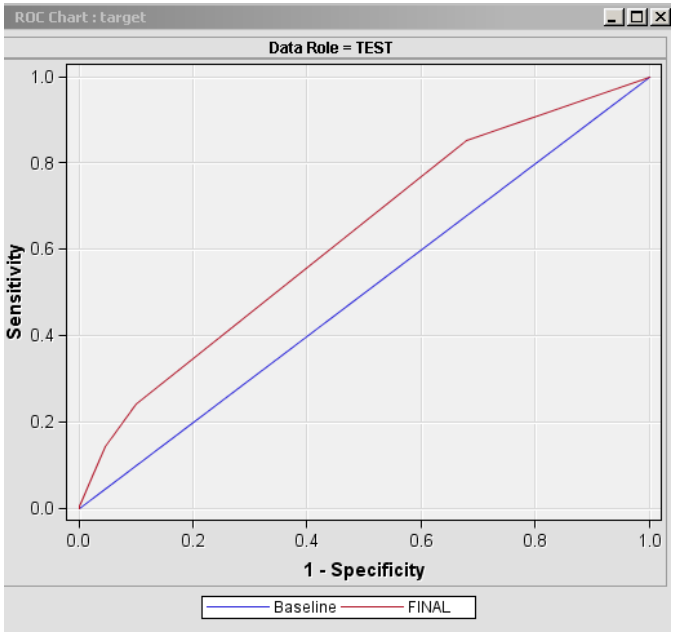


výsledky modelu – s předschválenou nabídkou

Giniho koeficient	0,31
Cumulative Lift	2,39

Parameter	Category	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimat	Exp(Est)
Intercept	1	-4,3225	0,0104	172627,3	<,0001		0,013
WOE_M0_PROD_ALL_FEE_AMT_M0	1	0,2412	0,0448	29,04	<,0001	0,0337	1,273
WOE_D6_TRANS_ALL_CNT	1	0,347	0,036	93,12	<,0001	0,0935	1,415
WOE_OFFER_CC_LIMIT	1	0,359	0,0494	52,87	<,0001	0,0399	1,432
WOE_CAMP_INFO_CNT_L12M	1	0,4106	0,0506	65,74	<,0001	0,0448	1,508
WOE_D3_DC_POS_AMT	1	0,432	0,036	144,3	<,0001	0,125	1,54
WOE_ACTIVE_DB_CNT	1	0,5108	0,0489	109,05	<,0001	0,0868	1,667
WOE_PROD_CA_CONTRACT_AGE	1	0,5279	0,0425	154,36	<,0001	0,144	1,695
WOE_T_ACTIVE_LIMIT_AMT	1	0,8977	0,0219	1672,55	<,0001	0,1415	2,454





děkuji za pozornost!

Mobilní aplikace Deloitte CZ

Android iOS

Zpravodaje I Studie I Semináře I Novinky I Videa

Deloitte.