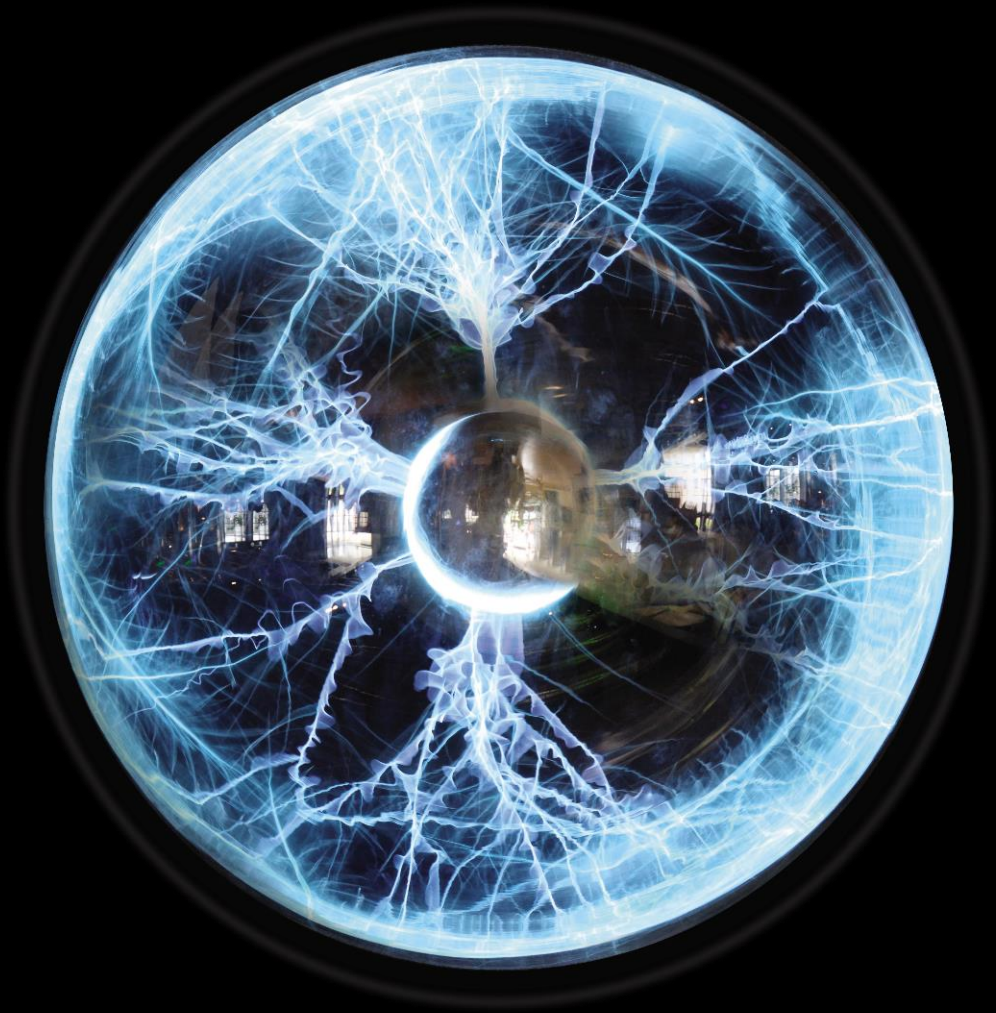
Deloitte.



Odchodovost klientů: Modelování pomocí rozhodovacích stromů & Voice to Text analýza

Veronika Počerová

Motivace

Proces reaktivní retence – zachraňovat chceme pouze klienty:

kteří mají vysokou pravděpodobnost, že budou zachráněni (akceptují retenční nabídku)

kteří zůstanou profitabilní i po poskytnutí retenční slevy

Voice to Text a Text Mining– chceme se rozkoukat v nestrukturovaných datech abychom:

mohli kategorizovat příchozí a odchozí hovory

poznat nejlepší operátory a využít jejich schopností i pro ostatní

zlepšit fungování callcentra na základě identifikovaných patternů v úspěšných hovorech

Rozhodovací stromy

Teorie

Rozhodovací stromy

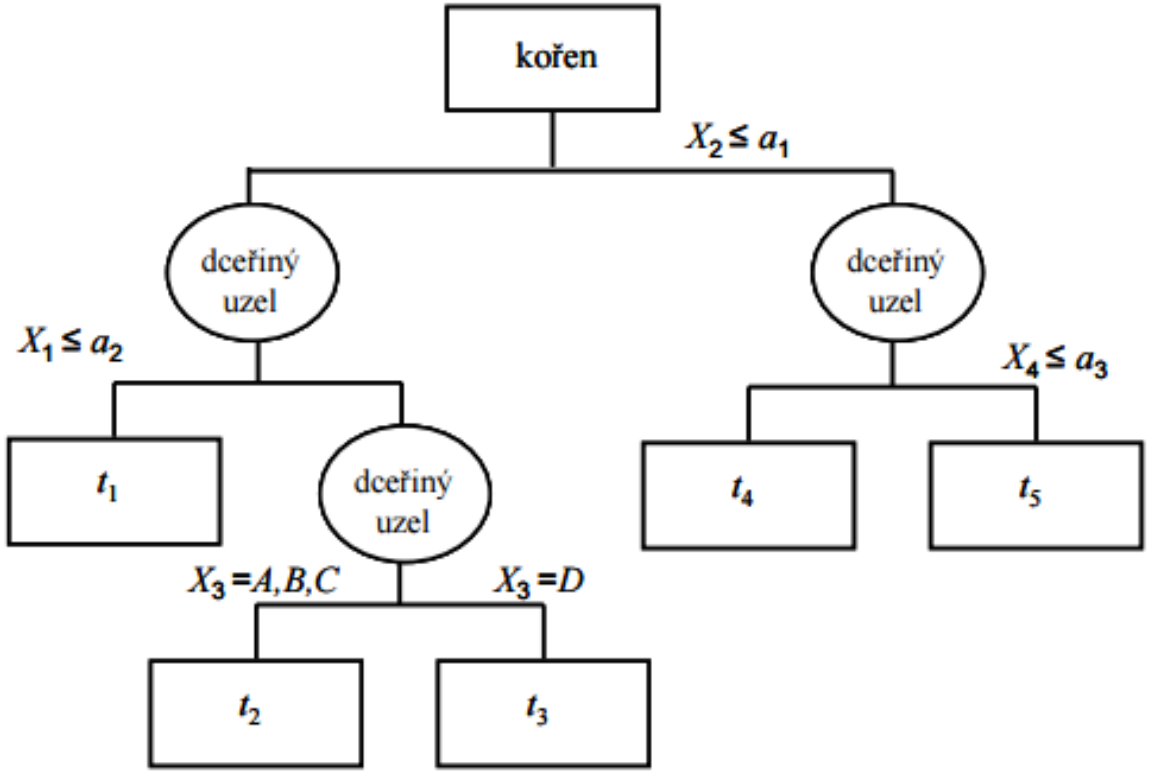
Teorie

- Hierarchicky uspořádaná rozhodovací pravidla
- Kořen, hrany, uzly, listy
- Binární x nebinární
- Klasifikační x regresní

Mějme strom T s listy $t = (t_1, \dots, t_N)$.

U klasifikačního stromu jsou pozorování kategoriální závisle proměnné Y *s* J *kategoriemi zařazeny do některé z kategorií* $c = (c_1, \dots, c_J)$, *kde* $J \geq 2$.

Pokud je závisle proměnná spojitá $Y = (y_1, \dots, y_N)$, *pozorováním je přiřazena hodnota predikovaná modelem* $\hat{y}_i$ *a výsledný strom bude regresní.*



Grafická struktura rozhodovacího stromu CART. Indexy u terminálních uzlů udávají, v jakém pořadí došlo k oddělení jednotlivých terminálních uzlů. Prediktory X_1, X_2 a X_4 jsou spojité, prediktor X_4 je kategoriální s kategoriemi A,B,C,D.

Rozhodovací stromy
Teorie

Pozorování proměnné Y jsou rozdělena do uzlů hodnotami ***vysvětlujících proměnných (prediktorů)*** $\boldsymbol{X_1, \ldots, X_M}$***.***

Kategorické prediktory

Odpovídáme na otázku, které pozorování y_i patří do množiny, kde $x_i \in A$, přičemž A je neprázdná vlastní podmnožina množiny všech hodnot veličiny X.

Spojité prediktory

Y rozdělujeme pomocí hodnoty daného prediktoru X. V tomto případě pozorování y_i patří do prvního uzlu, pokud je $x_i \geq a$ a do druhého uzlu pokud je $x_i < a$.

- K danému větvení stromu je použito vždy jen jednoho prediktoru. Stejný prediktor však může být využit v dalším větvení.
- Každé pozorování y_i tak patří pouze do jednoho terminálního uzlu a je mu přiřazena kategorie (klasifikační strom) nebo průměr hodnot (regresní strom) závisle proměnné Y tohoto uzlu.
- Stromy nekladou nároky na rozložení dat (např. konstantní rozptyl, normální rozložení, nezávislost prediktorů).

Rozhodovací stromy

Konstrukce stromu – kritérium pro větvení (kriteriální statistika)

Gini index

$$GI = \sum_{c=1}^{J} p_{tc}(1 - p_{tc})$$

Entropie

$$H = -\sum_{c=1}^{J} p_{tc} \log_2 p_{tc}$$

Informační zisk (GAIN)

$$GAIN_{celk} = H - \left(\sum_{i=1}^{k} \frac{N_i}{N_t} H(i) \right)$$

kde p_{tc} je podíl pozorování y_i s kategorií c v uzlu t z celkového počtu všech pozorování y_i v tomto uzlu neboli pravděpodobnost kategorie c v uzlu t.

Rozhodovací stromy

Konstrukce stromu – kritérium pro větvení (kriteriální statistika)

Chí-kvadrát

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{\left(p_{ij} - o_{ij}\right)^2}{o_{ij}}$$

$$o_{ij} = \frac{R_i S_j}{n}$$

kde i a j je označení řádků (resp. sloupců) v kontingenční tabulce, p_{ij} je pozorovaná frekvence, o_{ij} očekávaná frekvence, n je celkový počet pozorování, R_i je počet pozorování v řádku i, S_j je počet pozorování ve sloupci j.

Minimum kvadratické chyby (Least Square Deviation)

$$\bar{y_t} = \frac{1}{N_t}\sum y_{i(t)}$$

$$Q_t(T) = \frac{1}{N_t}\sum_{i=1}^{N_t}(y_i - \bar{y_t})^2$$

kde N_t je počet pozorování v uzlu t a $y_{i(t)}$ jsou hodnoty závisle proměnné v uzlu t.

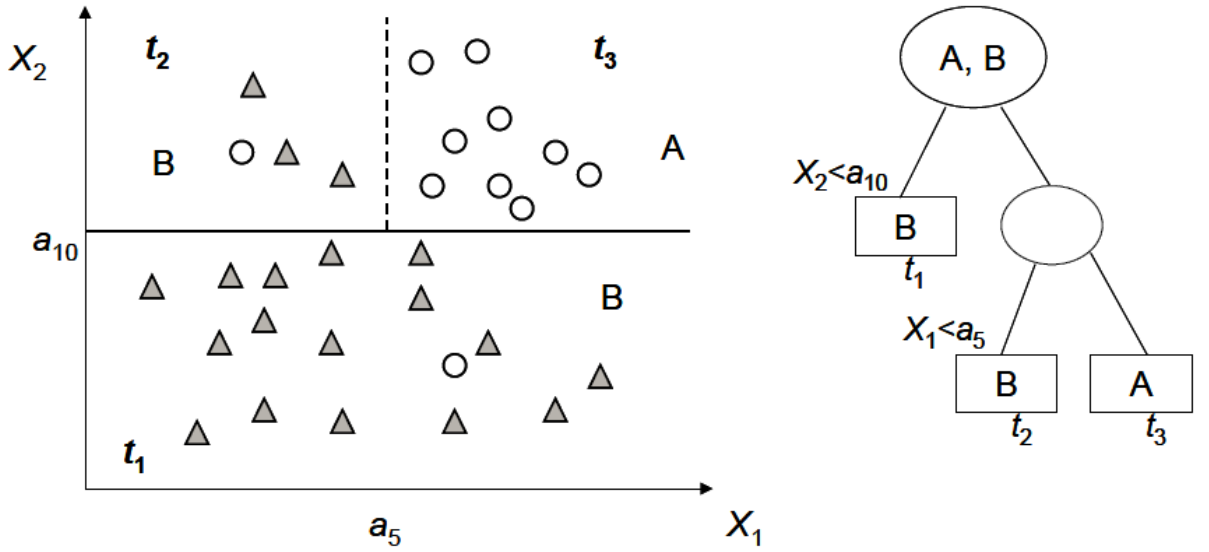
Binární stromy – algoritmus CART (Classification and Regression Trees)
Úvod

Kategoriální i regresní úlohy

Růst na základě rekurzivního binárního dělení

Problém: Hledáme rozdělení závisle proměnné Y prediktorem X, aby hodnoty proměnné Y byly uvnitř uzlu co nejhomogennější a zároveň mezi uzly co nejrozdílnější.

Homogenitu měříme pomocí kriteriální statistiky – nejčastěji se používá Gini index nebo entropie.



Obr. 2.3 *Grafické znázornění rozdělení pozorování do kategorií A a B závisle proměnné Y s použitím dvou spojitých prediktorů X_1, X_2.*

Binární stromy – algoritmus CART (Classification and Regression Trees)

Algoritmus růstu stromu

1. Rozděl soubor na trénovací a testovací. Tento poměr se určuje na základě počtu pozorování a účelu studie.

2. Najdi nejlepší rozdělení každého z prediktorů:

   – Pro spojité proměnné - seřaď hodnoty každého prediktoru (spojitého nebo ordinálního) od nejmenší po největší. Projdi všechny hodnoty prediktoru *X* a spočítej kriteriální statistiku všech možných rozdělení proměnné *Y* na dva potenciální dceřiné uzly. Hledáme minimum.

   – Pro kategoriální prediktor se za účelem nalezení nejlepšího rozdělení projdou všechny možné kombinace, tvořené jednotlivými kategoriemi prediktoru a hodnot nebo kategorií závisle proměnné. Opět se použije dělení s nejnižší hodnotou kriteriální statistiky.

3. Rozděl soubor na dva dceřiné uzly *t*1 a *t*2 podle kroku 2.

4. Opakuj krok 2 a 3, dokud se dělení nezastaví na předem definované hodnotě (dokud není dosaženo některého z pravidel pro zastavení růstu stromu – minimálního počtu pozorování v listu, maximálního únosného počtu větví, maximální hloubky stromu, velikosti chyby v potenciálních uzlech).

5. Použij testovací soubor k ověření vhodné velikosti stromu, a pokud je strom příliš velký, prořež strom.

Binární stromy – algoritmus CART (Classification and Regression Trees)
Výběr optimálního stromu

Optimální velikost stromu (underfitting, overfitting)

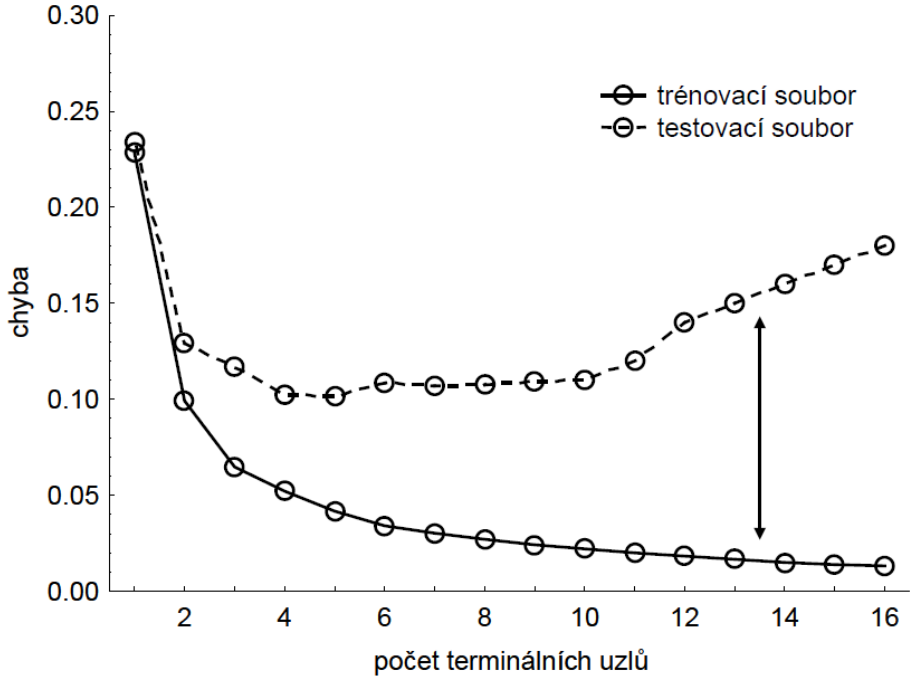
Trénovací & testovací soubor

Prořezávání & křížová validace

1. Nechat narůst libovolně veliký strom
2. Rozdělit soubor na k podsouborů: k-1 využít pro trénování a jeden pro testování, sestavit k modelů
3. Odhadnout α tak, aby strom měl co největší přesnost, ale zároveň měl co nejmenší rozdíl v chybě mezi testovacím a trénovacím souborem

$$C_\alpha(T_1) = DT_1 + \alpha|T_1|$$

kde $|T_1|$ je počet terminálních uzlů a DT_1 je chyba stromu. Parametr α vyjadřuje kompromis mezi velikostí a přesností stromu



Rozdíl ve velikosti chyby mezi testovacím a trénovacím souborem při různé velikosti stromu, dané počtem terminálních uzlů. Nejprve byla spočítána chyba (procento chybně zaklasifikovaných pozorování) na testovacím a trénovacím souboru pro strom s 16 terminálními uzly. Postupně bylo vždy zpětně odstraněno poslední rozdělení uzlů, čímž se snížil počet terminálních uzlů o jedna. Pro takto zmenšený strom byla opět spočítána chyba pro oba soubory. Takto se postupně strom zmenšoval, až zbyl pouze jeden uzel –kořen stromu.

Binární stromy – algoritmus CART (Classification and Regression Trees)
Měření přesnosti

Celková správnost

$$OA = \frac{n_p}{n}$$

kde n_p je počet správně klasifikovaných pozorování a n je celkový počet pozorování

Korekce na velikost kategorií: $$OA_{kateg} = \frac{1}{J}\sum_{c=1}^{J}\frac{n_{pc}}{n_c}$$

kde J je celkový počet kategorií, n_{pc} je počet správně klasifikovaných pozorování v kategorii c a n_c je počet všech pozorování v kategorii c

Koeficient determinace

$$R^2 = \frac{variabilita\ vysvětlená\ modelem}{celková\ varianbilita\ Y} = 1 - \frac{residuální\ variabilita}{celková\ variabilita\ Y} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

kde $\hat{y}_i = \overline{y_t}$ je průměr v příslušných terminálních uzlech a odchylka od průměru uzlu t je spočítána vždy pro pozorování y_i zařazené do tohoto terminálního uzlu

Nebinární stromy - Algoritmus CHAID (Chi-squared Automatic Interaction Detector)
Úvod

Zejména pro kategoriální prediktory

Strom nemusí být binární

Kriteriální statistika: Pearsonův chí-kvadrát statistický test

Vhodnější pro větší datové soubory

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(p_{ij} - o_{ij})^2}{o_{ij}}$$

$$o_{ij} = \frac{R_i S_j}{n}$$

kde i a j je označení řádků (resp. sloupců) v kontingenční tabulce, p_{ij} je pozorovaná frekvence, o_{ij} očekávaná frekvence, n je celkový počet pozorování, R_i je počet pozorování v řádku i, S_j je počet pozorování ve sloupci j.

		kategorie prediktoru X				
	j / ***i***	**1**	**2**	**…**	***s***	**Celkem**
kategorie Y	**1**	p_{11}	p_{12}	…	p_{1s}	R_1
	2	p_{21}	p_{22}	…	p_{2s}	R_2
	…	…	…	…	…	**…**
	r	p_{r1}	p_{r2}	…	p_{rs}	R_r
	Celkem	S_1	S_2	**…**	S_s	***n***

Nebinární stromy - Algoritmus CHAID (Chi-squared Automatic Interaction Detector)
Algoritmus růstu

1. Vytvoř kontingenční tabulku kategorií závisle proměnné a prediktoru

2. Pokud je počet kategorií prediktoru > 2, utvoří se dvojice z kategorií prediktoru. Najde se taková dvojice, která si je co do hodnot závisle proměnné Y nejvíce podobná neboli dvojice, jejíž χ^2 test má nejvyšší *p* hodnotu.

3. Dvojice s nejvyšší *p* hodnotou, která není statisticky významná nebo jejichž *p* hodnota je větší než α, se sloučí do jedné skupiny. Pokud je i po sloučení počet kategorií > 2, algoritmus se vrátí do kroku 2. Pokud ne, algoritmus pokračuje krokem 5.

4. Každá kategorie, která má velmi málo pozorování je spojena s nejpodobnější kategorií

5. Vybere se prediktor s nejmenší adjustovanou *p* hodnotou. Tento prediktor s optimálně sloučenými kategoriemi je použit k rozdělení uzlu. Pokud významný prediktor nelze nalézt, uzel se již dále nedělí a je považován za terminální.

Bonferroniho multiplikátor

Ordinální proměnné: $B_{ordinal} = \binom{s-1}{r-1}$

Kategoriální proměnné: $B_{categorial} = \sum_{i=0}^{r}(-1)^i \frac{(r-i)^s}{i!(r-i)!}$

Rozhodovací stromy
Výhody a nevýhody

Výhody

Snadné grafické znázornění, jednoduchá interpretace

Neklade žádné podmínky na typ rozdělení závisle proměnné ani prediktorů

Závisle proměnná i prediktory mohou být všech typů (kategoriální, ordinální i spojité) – *CART*

Nevýhody

Nestabilita

Účinky jednotlivých prediktorů nelze „sčítat“

Stromy jsou nevhodné pro malý počet vzorků a velký počet kategorií závisle proměnné

Výsledkem je nespojitá plocha (převažující kategorie nebo průměr pro každý terminální uzel)

Použití rozhodovacích stromů v praxi

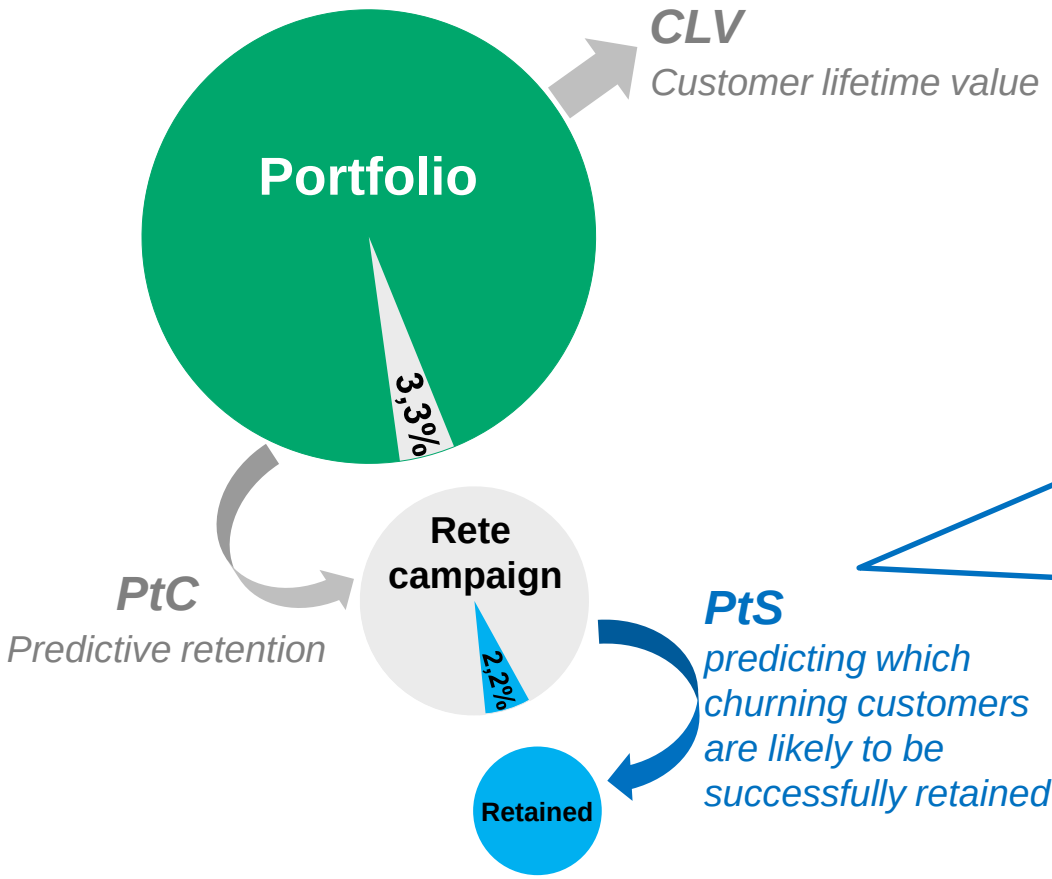
Propensity to Save model

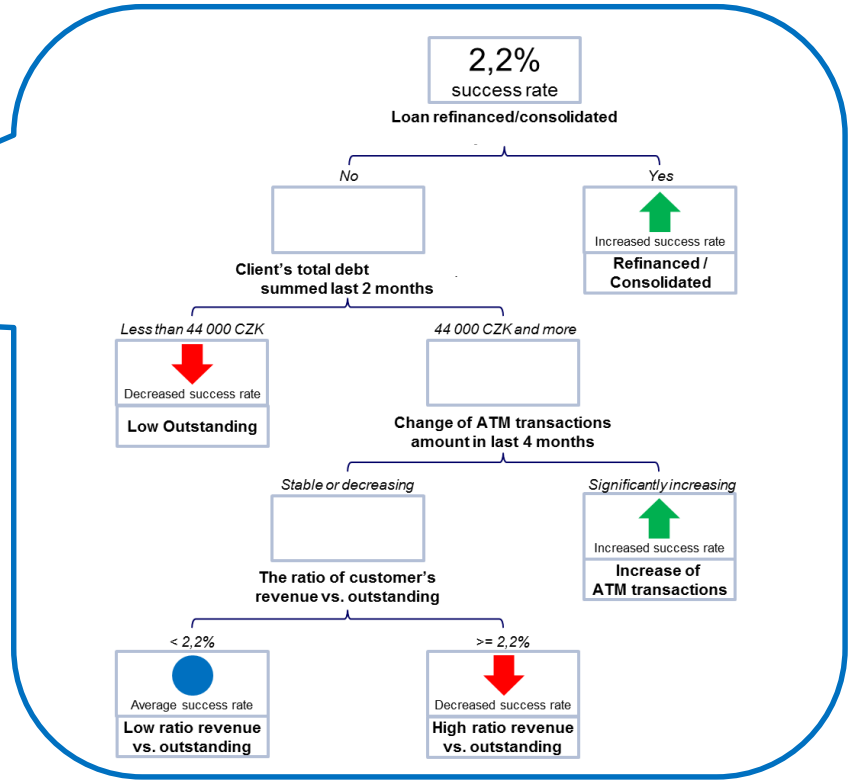
Propensity to Save model pro klienty s úvěrem

Specifikace modelu

Propensity to Save Model (PtS)

- Identifying **segments worth specific treatment** in Reactive retention
- Focusing on customers with potential to be retained and maintain revenue
- Result of an objective statistical analysis of over 2000 derived attributes (Successful retention = customer has been retained and maintained revenue at least 100CZK after the retention)





Propensity to Save model pro klienty s úvěrem

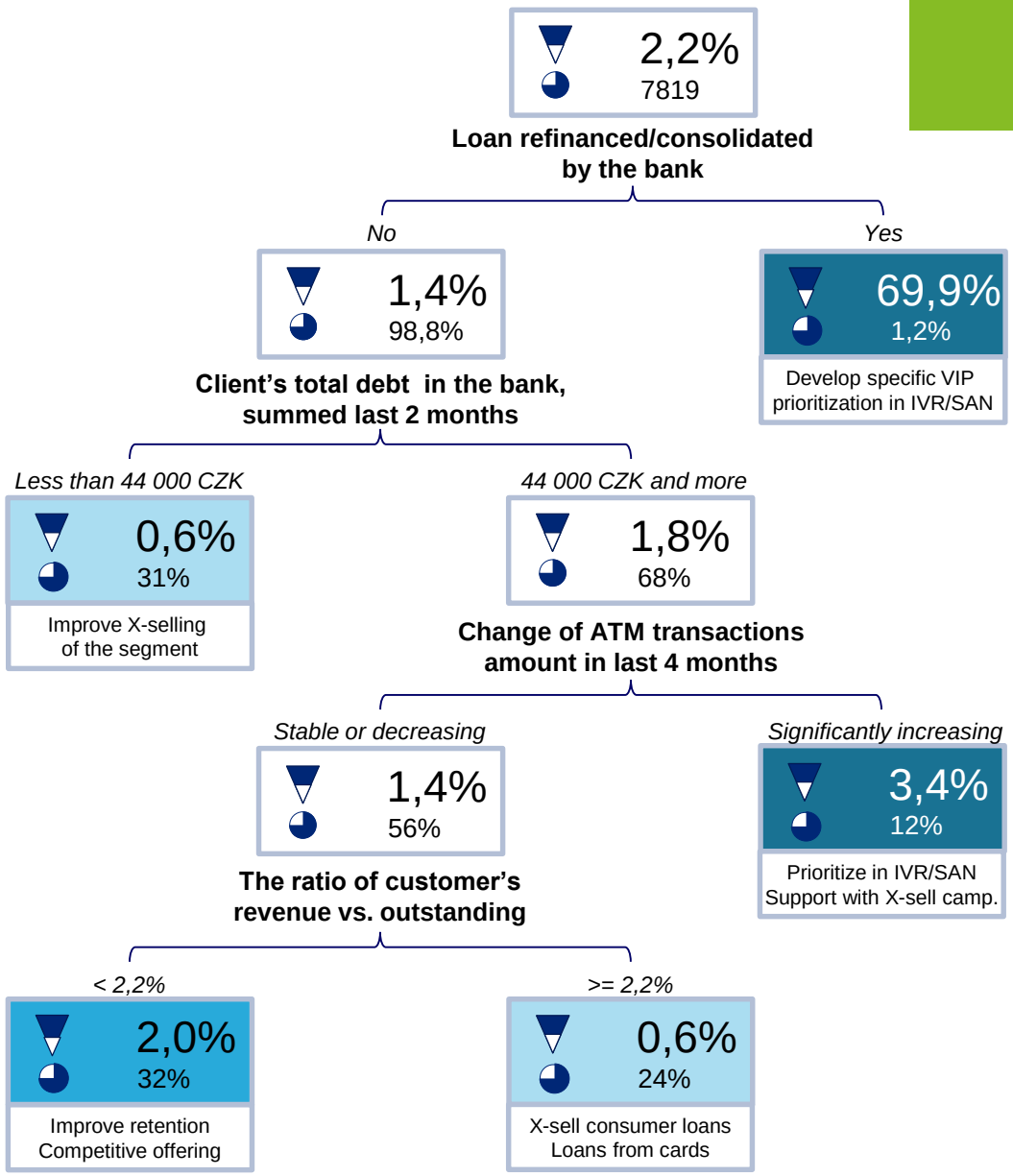
Výsledný model

2000 prediktorů
Gini=0,49

Legend:

success rate
% of retention camp.

- **Retention campaign** = customers entering the retention campaign, without current delinquency (as of the period 07/2015 – 11/2015)
- **Success rate** = % of customers in retention campaign, marked as retained, and maintaining revenue at least 100CZK after the retention
- The PtS model is a result of statistical analysis, relevant to a **static point in time** (evaluating portfolio 07/2015 – 11/2015). Thus strong monitoring of the model is needed to be implemented and adequate model maintenance to be in place.



Propensity to Save model pro klienty s úvěrem

Segment: "Increase of ATM transactions"

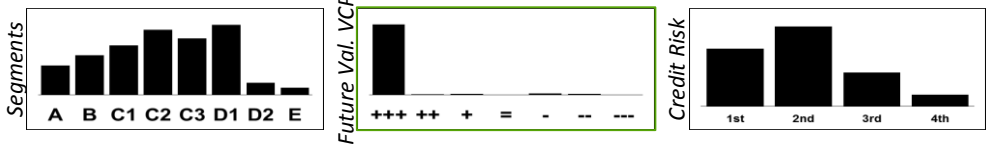
"Increase of ATM transactions" 🎯 3,4%

52% 48%

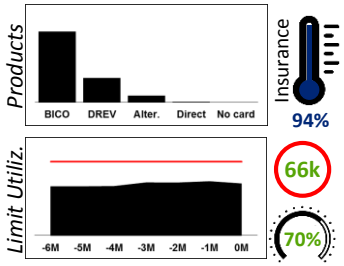
Age Ø44

Opt-out 1,6%

Ø outstanding	88 000 CZK
∑ outstanding	17mio CZK
Ø revenue/month	1 600 CZK
Number of clients	192

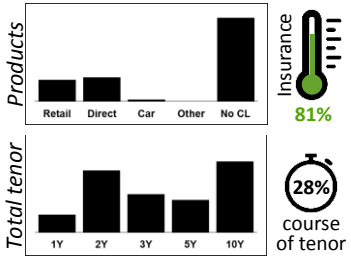


Credit Card

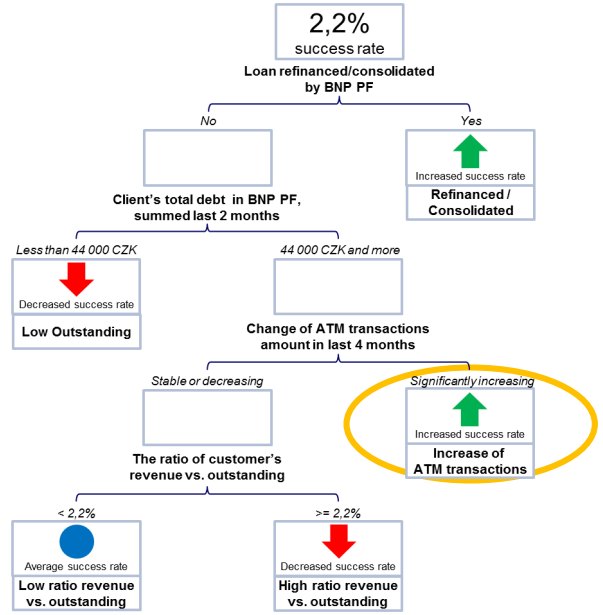


Ø outstanding	46 000 CZK
∑ outstanding	9,5mio CZK
Ø credit limit	66 000 CZK
Ø months on books	55
# of credit cards	206

Consumer Loan



Ø outstanding	99 000 CZK
∑ outstanding	7,5mio CZK
Ø disbursed amt.	130 000 CZK
Ø months on books	14
# of consumer loans	75



Segment characteristics

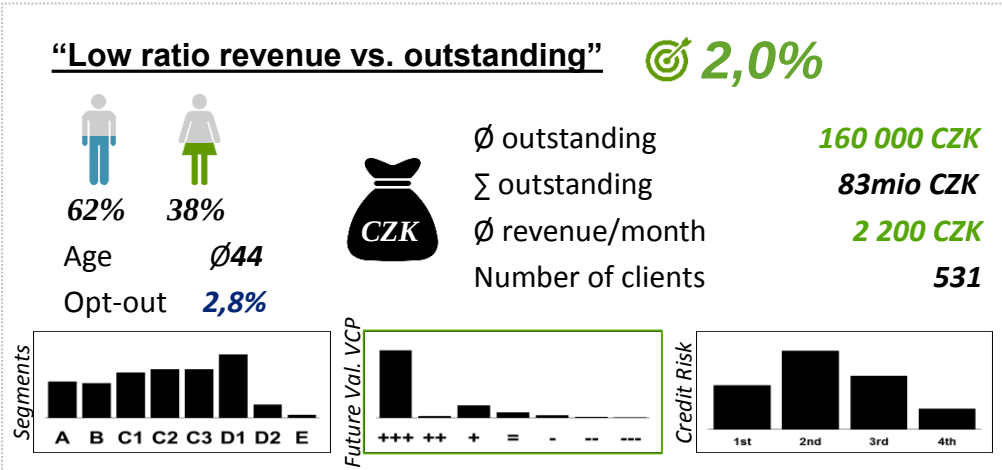
- Customers with increased amount of ATM transactions
- Increasing utilization of credit card
- Often without consumer loan

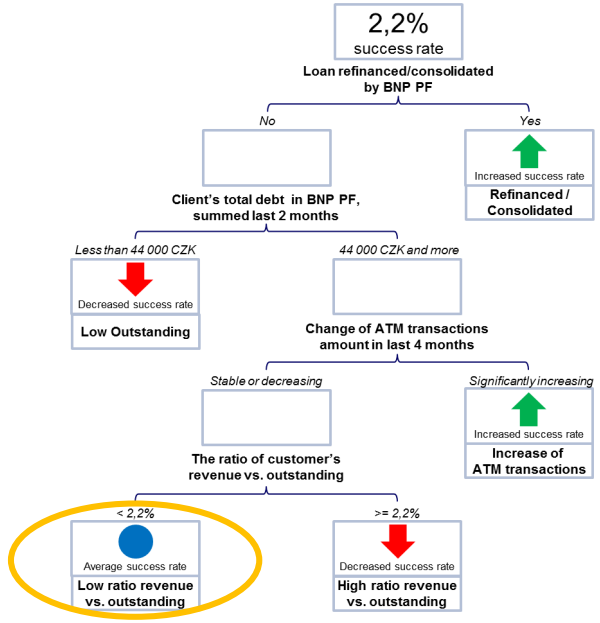
Typical treatment

- Prioritize segment in IVR/SAN
- X-sell consumer loans to the segment
- Focus on further credit card utilization

Propensity to Save model pro klienty s úvěrem

Segment: "Low ratio revenue vs. outstanding"





Segment characteristics

- Customers with low ratio between revenue and outstanding
- Long term consumer loans, with high disbursed amount
- Increasing utilization of credit card

Typical treatment

- Continuously make improvements in retention call-script and retention offer - allocated core capacity of retention team
- Standard priority in IVR/SAN

Voice to Text analýza

Teorie

V2T PROCESSING - OUTCOMES

textual file

- Textual transcription of call
- This is plain text file of transcribed text from recorded call
- Every single line of file contains transcribed text from one speaker.

token file

- tokens from call enriched with metadata:
  - timing,
  - speaker,
  - order,
  - probability

emotion file

- 4 identified emotions:
  - Apathy
  - Neutral
  - Low excitement
  - High excitement
- 4 call patterns:
  - Interruptions
  - Hesitations
  - Responses
  - Speeches

```
<s2> TAM SI KLIKNETE MÁTE TAM FORMULE KTERÝ
BY JAKO ANO PARDON </s2>
<s1> NO TAKY TĚ MÁME KDYŽTAK ZAVOLEJTE </s1>
```

```
T=383 ST=145.67 ET=146.25 W=ZAVOLÁ P=1 C=1
T=384 ST=146.25 ET=146.61 W=SDĚLÍ P=1 C=1
T=385 ST=146.61 ET=146.94 W=VŠECHNY P=1C=1
```

TEXT MINING PROCESSING

- **Tokenization** - In lexical analysis we will parse a document collection in order to quantify information about the terms that are contained therein. Lexical analysis of text is the process of breaking a stream of text up into words, phrases, symbols, or other meaningful elements called tokens (tokenization). Tokenization will simplify searching through all texts.
- **Lemmatization** - Lemmatization is the process of identifying a single term called lemma for different inflected forms of a token so that they could be analyzed as a single item. The objectives of lemmatization are to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form and to reduce computation complexity.
- **Document-Term Matrix** – Document-Term Matrix (DTM) is a mathematical matrix that describes a collection of analyzed documents in terms of the frequency (or different metric) of particular terms in the collection. In a DTM, rows correspond to documents in the collection and columns correspond to lemmas (terms) in the collection. The value w_{ij} that each entry of *i*-th row and *j*-th column in the DTM should take reflects the frequency (or different metric value) of the particular tokens in the documents. There is many options for choice of metric e.g. term frequency weighted by inverse document frequency (TF-IDF) in whole data set. The formula for TF-IDF is $\boldsymbol{tfidf_{ij} = w_{ij} \cdot \log\left(\frac{N}{n_i}\right)}$.
- **N-gram identification** - N-gram is a contiguous sequence of *n* tokens from a given collection of textual documents. The relevancy of n-gram can be evaluated by TF-IDF metric.
- **Collocation identification** – Collocation is a sequence of tokens that co-occur more often than by chance. Collocations are suitable for identification of topics in analyzed collection of documents.

Použití Voice to Text analýzy v praxi

Analýza hovorů Call Centra

V2T – AFTERSALES EMOTIONS – BEST OPERATOR

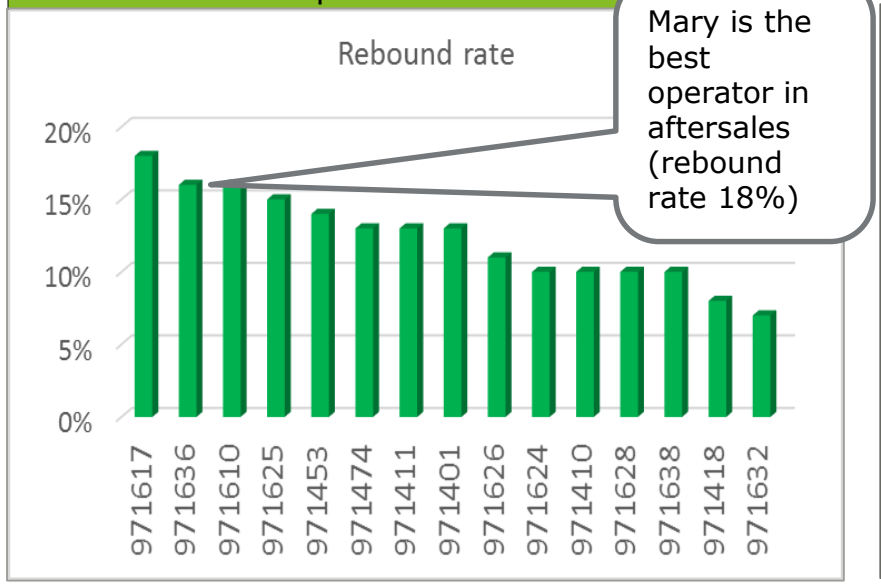
Finding

The best operator in aftersales is **Mary**. Her success influences her positive attitude and her arguments that she used in calls about early repayment. She could be the emotional pattern for others.

Top collocations on Rebound

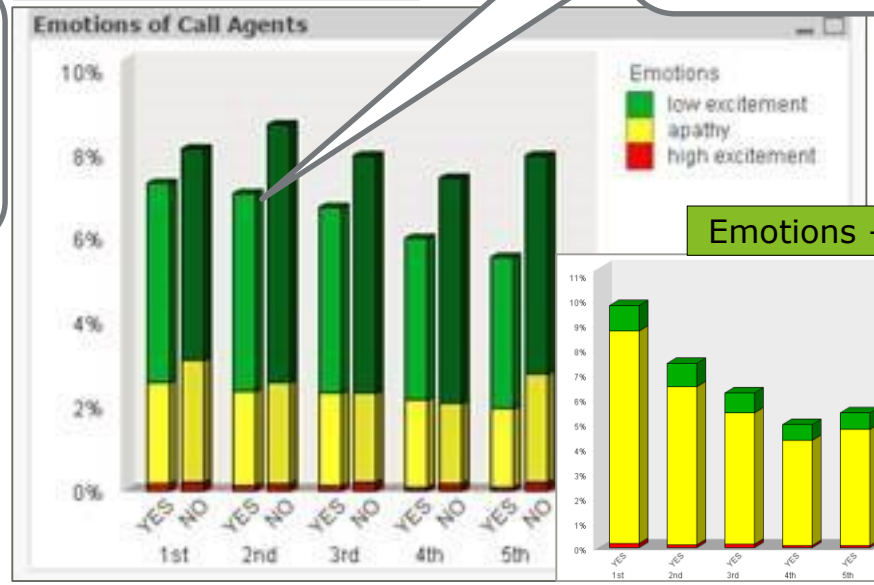
- Collocations: kreditní karta, osobní půjčka, dlužný částka, aktuální dlužný částka, kreditní karta – successfully solves the early repayment and consolidation.
- The main call purpose: consolidation

Rebound rate of operators in aftersal



Mary is the best operator in aftersales (rebound rate 18%)

Emotions - Mary



Emotionally happy: She is mainly happy during the call **(4% vs. 1% avg)** with a positive attitude.

Emotions - AVG

V2T – RETENTION TEXT MINING – TOP 5 FREQUENT COMPETITORS

Finding

Each operator is better in the rebound rate in the calls against a specific competitor. Operators may well complement in retention calls each othe

Rebound rate of operators

Operator	Competitor 1	Competitor 2	Competitor 3	Competitor 4	Competitor 5	Competitor 6
John	22,7%	28,4%	30,2%	47,4%	36,8%	22,7%
Angel	40,6%	25,4%	38,0%	56,9%	13,1%	56,9%
Peter	22,7%	37,9%	21,2%	37,9%	11,4%	37,9%
Ami	0,0%	34,1%	42,5%	0,0%	22,7%	38,9%
Charles	11,4%	0,0%	19,6%	35,5%	19,0%	23,7%
Veronica	56,9%	3,5%	27,0%	39,8%	36,2%	32,5%
Total Rebound	**25,7%**	**21,5%**	**29,7%**	**36,3%**	**23,2%**	**35,0%**
Source of knowledge	Veronica	Peter	Ami	Charles	John	Angel

Typical treatment

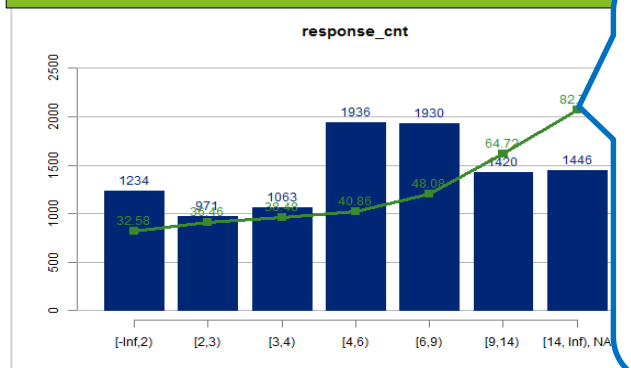
- The operator with the best rebound rate in the calls against a specific competitor may be a source of knowledge for others.
- Analysis of the rebound rate and the frequency of the specific competitors in retention calls, we can determine the best allocation of operators.

V2T – RETENTION – CUSTOMERS CALL PATTERNS (RESPONSE, SILENT, SPEECH, HESITATION)

Finding

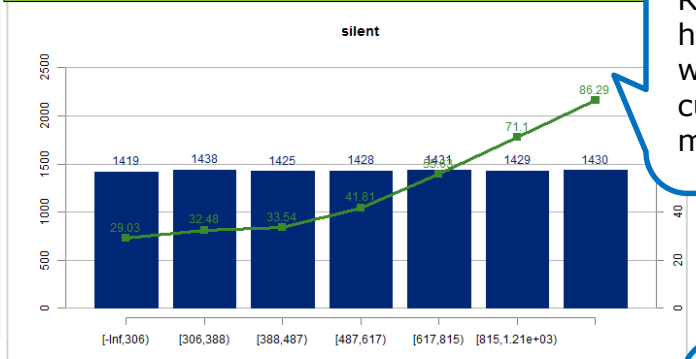
The more customer responses and hesitations are in the call it is more successful. = **The more short questions from the operator that the customer can shortly answer leads to success.**

Number of customer responses



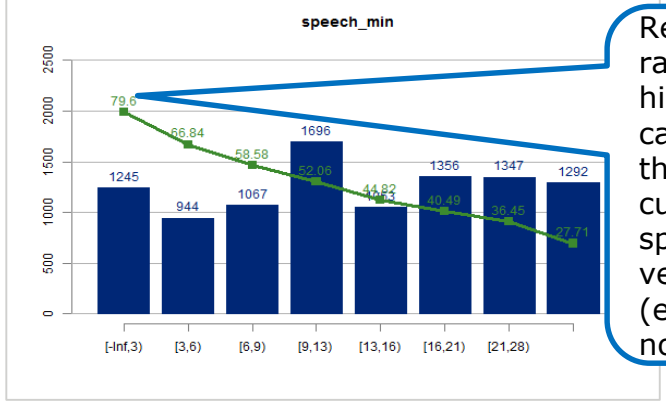
Rebound rate is highest in calls with more customer pauses before starting answer the question from the operator.

Total duration of customer silence (sec



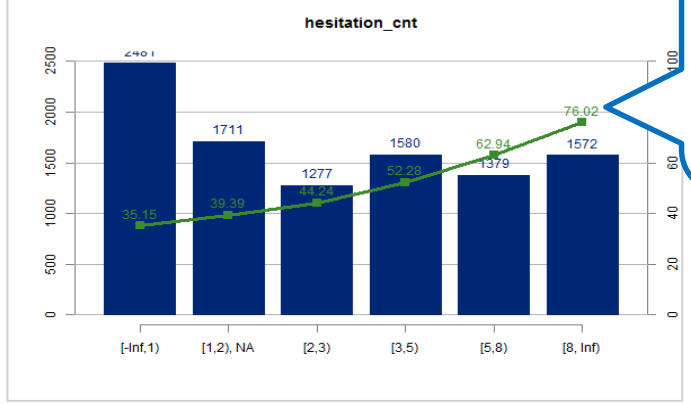
Rebound rate is highest in calls when the customer is more silent.

Min. duration of customer speech (seconds)



Rebound rate is highest in calls when the shortest customer speech is very short (eg; yes, no).

Number of customer hesitations

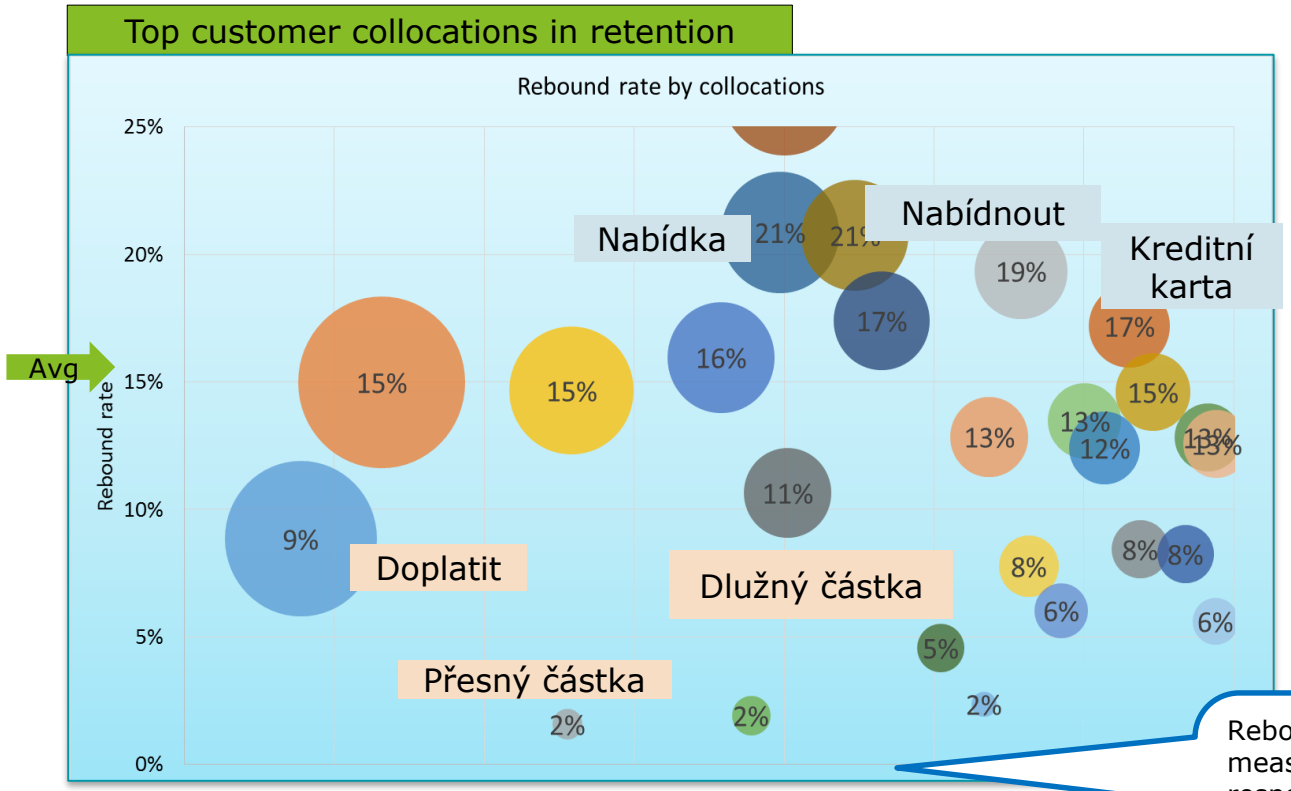


Rebound rate is highest in calls when the customer makes more hesitations in speech.

V2T – TOP CUSTOMERS COLLOCATIONS

Finding

When customer says in retention calls collocations: nabídka, nabídnout or kreditní karta the call have higher rebound rate than calls in which customer says collocations: doplatit, přesný částka or dlužný částka. Average rebound rate of retention calls is 15 %.



Typical treatment

- **Focus on offer in every retention calls.** Customer collocations that leads to success: nabídka (21% rr), nabídnout (19% rr), kreditní karta (17% rr).
- **Focus on care and usage campaigns that could avoid the customers reasons to early repayment.** Collocations that leads to lower rebound rate: doplatit (9% rr), přesný částka (2% rr), dlužný částka (5% rr).

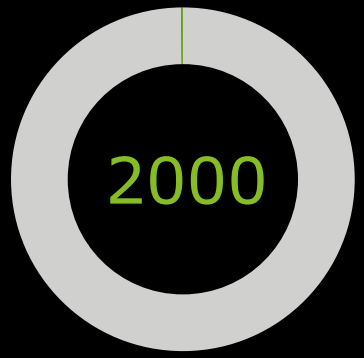
Rebound rate 95 % confidence interval width measures how accrued is the rebound rate with respects to the number of calls observed. Rebound rate of collocations on the left hand side are more accrued then rebound rate of collocations on the right side.

Shrnutí

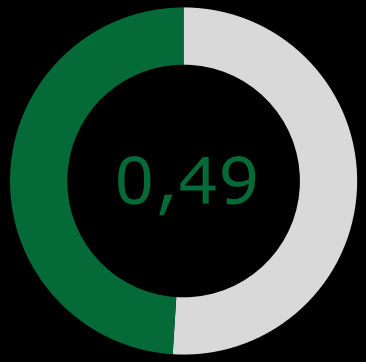
Propensity to Save model, který prioritizuje klienty pro reaktivní retenci:

Klienti, kteří mají vysokou pravděpodobnost, že budou zachráněni (akceptují retenční nabídku) & kteří zůstanou profitabilní i po poskytnutí retenční slevy

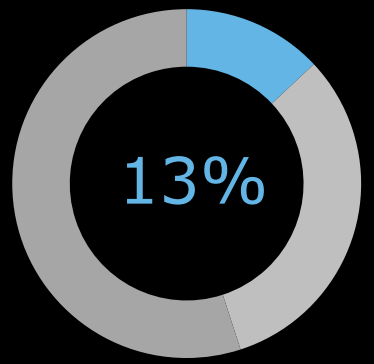
Počet prediktorů



Gini



Klienti s prioritou 1



Voice to Text a Text Mining:

Na základě analýzy proběhl workshop pro operátory na téma úspěšný hovor.

Pomocí textové analýzy se podařilo identifikovat největší konkurenci, kam klienti odcházejí.

Děkujeme za pozornost!

Otázky?

Mobilní aplikace Deloitte CZ

Android iOS

Zpravodaje | Studie | Semináře | Novinky | Videa

Deloitte.