UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY



# MODELOVANIE KAUZÁLNYCH VPLYVOV MARKETINGOVÝCH OSLOVENÍ NA PREFERENCIE KLIENTA

DIPLOMOVÁ PRÁCA

2025 Bc. Olha Koshova

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

# MODELOVANIE KAUZÁLNYCH VPLYVOV MARKETINGOVÝCH OSLOVENÍ NA PREFERENCIE KLIENTA

## DIPLOMOVÁ PRÁCA

| | |
|---|---|
| Študijný program: | Ekonomická a finančná matematika a modelovanie |
| Študijný odbor: | 1113 Matematika |
| Školiace pracovisko: | Katedra aplikovanej matematiky a štatistiky |
| Vedúci práce: | Mgr. Martin Hurban, PhD. |

Bratislava 2025 **BC. OLHA KOSHOVA**

**Poďakovanie:** Rada by som vyjadrila poďakovanie školiteľovi mojej diplomovej práce Mgr. Martinovi Hurbanovi, PhD., za odborné vedenie, čas, ktorý mi venoval, a cennú pomoc počas vypracovania tejto práce. Zároveň by som mu chcela poďakovať aj za všetky vedomosti a skúsenosti, ktoré som vďaka spolupráci s ním získala.






Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

# ZADANIE ZÁVEREČNEJ PRÁCE

**Meno a priezvisko študenta:** Bc. Olha Koshova
**Študijný program:** ekonomicko-finančná matematika a modelovanie (Jednoodborové štúdium, magisterský II. st., denná forma)
**Študijný odbor:** matematika
**Typ záverečnej práce:** diplomová
**Jazyk záverečnej práce:** slovenský
**Sekundárny jazyk:** anglický

**Názov:** Modelovanie kauzálnych vplyvov marketingových oslovení na preferencie klienta
*Modelling of causal influences of marketing actions on client preferences*

**Anotácia:** V oblasti marketingu je predikcia zákazníckych preferencií fundamentálnou časťou práce marketingových oddelení. Zákaznícke preferencie závisia od marketingových akcií, ktoré im predchádzajú. Odhady vplyvov marketingových akcií na jednotlivcov sú dôležitým vstupom pre vytváranie alokačných politík marketingových akcií. Efektívna alokácia marketingových akcií (napr. poslanie emailu, telefonátu, SMS, notifikácie) je pre firmy zaujímavá z dôvodu zvýšenia efektivity financií vynakladaných na marketing.

**Vedúci:** Mgr. Martin Hurban, PhD.
**Katedra:** FMFI.KAMŠ - Katedra aplikovanej matematiky a štatistiky
**Vedúci katedry:** doc. Mgr. Igor Melicherčík, PhD.

**Dátum zadania:** 11.01.2024

**Dátum schválenia:** 25.01.2024 prof. RNDr. Daniel Ševčovič, DrSc.
garant študijného programu

.....
študent

.....
vedúci práce

# Abstrakt v štátnom jazyku

KOSHOVA, Olha: Modelovanie kauzálnych vplyvov marketingových oslovení na preferencie klienta [Diplomová práca], Univerzita Komenského v Bratislave, Fakulta matematiky, fyziky a informatiky, Katedra aplikovanej matematiky a štatistiky; školiteľ: Mgr. Martin Hurban, PhD. , Bratislava, 2025, 93s.

V našej práci sa zaoberáme modelovaním kauzálneho vplyvu oslovení na preferencie klienta a jeho nákupné správanie prostredníctvom nástrojov strojového učenia. Najprv uvádzame teoretický základ, vysvetľujeme základy kauzality, ideu meta-učiteľov (S-learner, T-learner, X-learner), pomocou ktorých vieme odhadnúť individuálny efekt oslovenia zákazníkov. Taktiež uvádzame relevantné techniky strojového učenia, na ktorých je modelovanie postavené.

Najprv sa zameriavame na modelovanie vplyvu samotného oslovenia na nákupné správanie zákazníka. Taktiež v danom prípade sa pozeráme na to, ako rôzne prístupy výberu čŕt môžu ovplyvniť výsledky a výkonnosť nástroja. Pri modelovaní tak berieme do úvahy aj črty korelujúce so závislou premennou, aj s príčinnou. Ukazujeme, že výber čŕt ovplyvňuje výsledky modelovania, a že rozdielom medzi črtami korelujúcimi so závislou premennou a s príčinnou vieme presnejšie identifikovať kauzálny efekt.

Práca následne rozširuje tento prístup tým, že do modelovania pridáva ďalšiu príčinnú premennú, ktorá hovorí, cez aký komunikačný kanál bol zákazník oslovený. Tým sa modelovanie stáva komplexnejším a umožňuje odhadnúť efekt jednotlivých kanálov vzhľadom na referenčnú skupinu.

V záverečnej časti práce sa venujeme porovnaniu vybraných modelov na základe viacerých kritérií. Výsledky danej práce slúžia podkladom pre analýzu výkonnosti už existujúcich procesov a návrh nových prístupov pri tvorbe ďalších kampaní a výberu cieľových skupín.

**Kľúčové slová:** Kauzálna ineferencia, strojové učenie v marketingu, nákupné správanie, personalizované oslovenia, komunikačné kanály, meta-učitelia (S -learner, T -learner, X -learner), analýza výkonnosti modelov, optimalizácia marketingových stratégií

# Abstract

KOSHOVA, Olha: Modeling of causal influences of marketing actions on client preferences [Master's Thesis], Comenius University in Bratislava, Faculty of Mathematics, Physics and Informatics, Department of Applied Mathematics and Statistics; Supervisor: Mgr. Martin Hurban, PhD., Bratislava, 2025, 93 pages.

This thesis focuses on modeling the causal influences of targeted messages on client preferences and purchasing behavior using machine learning models. We present the theoretical background, explain the basic concepts of causality, and introduce the framework of meta-learners (S-learner, T-learner, X-learner), by means of which the conditional average treatment effect (CATE) can be estimated for each customer. We also describe the relevant machine learning techniques on which our modeling approach is based.

Initially, we estimate the impact of being targeted on customer purchasing behavior. In this context, we also assess how various feature selection strategies affect model outcomes and performance. We consider both features correlated with the outcome variable and those with the treatment. We show that different strategies of feature selection can affect model performance. The difference between the sets of features correlated with the outcome and treatment can contribute to a more accurate estimation of the treatment effect.

The thesis further extends this model by adding an additional causal variable that incorporates the communication channel through which the customer was approached. The obtained results allow us to identify the preferred channel of customer engagement. This increases the complexity of the modeling and enables the estimation of the effect of each channel relative to a reference group.

Then, we focus on comparing models using different criteria. The insights gained from this thesis provide a basis for evaluating the effectiveness of current marketing processes and for developing new strategies for future campaigns and target selection.

**Keywords:** Causal inference, machine learning in marketing, purchasing behavior, personalized targeting, communication channels, meta-learners (S-learner, T-learner, X-learner), model performance analysis, optimization of marketing strategies

# Obsah

**Úvod** **9**

**1 Prehľad techník modelovania kauzálnej súvislosti** **11**

1.1 Kauzalita . . . . . 11

1.2 Odvodenie CATE rovnice . . . . . 13

1.3 Metódy odhadu CATE pri dvoch typoch ošetrenia . . . . . 15

1.3.1 Metriky výkonnosti pri dvoch typoch ošetrenia . . . . . 19

1.4 Metódy odhadu CATE pri viacerých ošetreniach . . . . . 20

1.4.1 Metriky výkonnosti meta učiteľov pri viacerých ošetreniach . . . . . 22

**2 Prehľad relevantných techník strojového učenia** **24**

2.1 Klasifikačné modely . . . . . 24

2.1.1 Klasifikačné stromy . . . . . 24

2.2 Logistická regresia . . . . . 25

2.3 Gradient Boosting . . . . . 26

2.4 Gradient Boosted Rozhodovacie Stromy . . . . . 27

2.4.1 LightGBM . . . . . 29

2.5 Zakodóvanie kategorických premenných . . . . . 32

2.6 Metriky výkonnosti základných modelov . . . . . 34

2.6.1 Metriky výkonnosti základných učiteľov . . . . . 34

**3 Formulácia problému oslovenia z pohľadu banky** **39**

3.1 Úvod do problematiky . . . . . 39

3.2 Formulácia optimalizačnej úlohy pre banky . . . . . 39

3.3 Výber klientov z pohľadu banky . . . . . 42

**4 Príprava dát pre trénovanie** **45**

4.1 Vysvetľujúce črty . . . . . 45

4.1.1 Črty z webovej aktivity . . . . . 46

4.2 Závislá premenná . . . . . 47

4.3 Výber tréningových pozorovaní . . . . . 47

4.4 Definovanie príčinnej premennej pri dvoch typoch ošetrenia . . . . . 51

4.5 Definovanie príčinnej premennej pri viacerých typoch ošetrenia . . . . . 53
4.6 Výber čŕt pre trénovanie . . . . . 55
4.6.1 Výber čŕt na základe korelácie so závislou premennou . . . . . 55
4.6.2 Výber čŕt na základe korelácie s príčinnou premennou . . . . . 56
4.6.3 Prístupy ku kombinácii výberu čŕt . . . . . 56

**5 Kvantifikácia efektu oslovenia 58**
5.1 Aplikácia S-learnera . . . . . 58
5.2 Aplikácia T-learnera . . . . . 62
5.3 Aplikácia X-learnera . . . . . 65
5.4 Porovnanie použitých nástrojov . . . . . 67

**6 Kvantifikácia efektu oslovenia prostredníctvom rôznych kanálov 70**
6.1 Aplikácia S-learnera na viacúrovňové ošetrenie . . . . . 71
6.1.1 Zakódovanie komunikačného kanála pomocou dummy premenných . 71
6.1.2 Výsledky modelu . . . . . 71
6.2 Aplikácia T-learnera na viacúrovňové ošetrenie . . . . . 73
6.3 Aplikácia X-learnera na viacúrovňové ošetrenie . . . . . 74
6.4 Stratégia priraďovania ošetrenia . . . . . 76
6.5 Porovnanie nástrojov pri viacúrovňovom ošetrení . . . . . 77

**Záver 82**

**Zoznam použitej literatúry 83**

**Príloha A. Detaily použitých metód 87**

**Príloha B. Doplňujúce technické údaje 93**

# Úvod

V súčasnosti je pre firmy dôležité spravovať marketingové aktivity z ohľadom na preferencie ich zákazníkov. Vzhľadom na konkurenciu a potrebu efektívneho vynakladania financií na marketing personalizovaná komunikácia nadobúda čoraz väčší význam. Preto odhady efektov oslovenia zákazníka na jeho správanie sú dôležitým vstupom pri tvorbe marketingových stratégií. Z pohľadu zákazníkov je modelovanie ich preferencií dôležité, nakoľko im poskytuje relevantné služby a produkty, ktoré zodpovedajú ich potrebám, čo by následne mohlo viesť k vyššej zákazníckej spokojnosti a lojalite.

Oslovenie je jedným z kľúčových marketingových nástrojov, ktorý spočíva v cielenej komunikácii so svojimi zákazníkmi. Pod oslovením rozumieme proces, pri ktorom spoločnosť cielene informuje zákazníkov o svojich produktoch a službách, aby ich motivovala k určitej akcii, napríklad k nákupu alebo k využitiu služby.

Kľúčovou úlohou danej práce je porozumieť tomu, aký vplyv majú jednotlivé oslovenia na rozhodnutia zákazníkov, aby boli investície do kampaní čo najefektívnejšie využité. Inými slovami, budeme sa snažiť odhadnúť rozdiel v pravdepodobnostiach nastania modelovanej udalosti za podemienky oslovenia a za podmienky, že sa oslovenie neuskutočnilo. Tento prístup poskytuje možnosť identifikovať zákazníkov, u ktorých sa pravdepodobnosť pozitívnej odpovede v dôsledku oslovenia výrazne zvyšuje, čím sa detailnejšie zaoberá aj [14]. Navyše, nezameriavame sa len na samotnú skutočnosť, či bol zákazník oslovený, ale zohľadňujeme aj vplyv komunikačného kanála, prostredníctvom ktorého bolo oslovenie realizované. Tento prístup nám umožňuje skúmať vplyv rôznych komunikačných kanálov na zákaznícke správanie.

V rámci danej práce sa zameriavame na modelovanie vplyvov oslovení zákazníkov na ich správanie prostredníctvom meta-učiteľov, ktoré umožňujú odhadovať efekty oslovení v závislosti od individuálnych charakteristík zákazníkov. Tento prístup bol rozpracovaný v práci [6], kde sú predstavené základné nástroje ako S-learner, T-learner a X-learner. Dané modely využívajú základné predikčné algoritmy na to, aby pomocou rôznych stratégií odhadli rozdiel v očakávaných výsledkoch medzi ošetrenými a neošetrenými jednotkami.

V Kap.1 predstavíme základné princípy kauzality a metódy odhadu individuálnych efektov ošetrenia (CATE). Venujeme sa nielen odhadu efektu pri dvoch typoch ošetrenia, ale aj tomu, ako sa tento prístup modifikuje v prípade, že uvažujeme viacero typov ošetre-

nia. Zároveň uvádzame metriky vhodné na vyhodnotenie výkonnosti uvedených prístupov v oboch scenároch.

V Kap.2 sa venujeme prehľadu relevantných techník strojového učenia. Predstavujeme základné klasifikačné modely vrátane logistickej regresie, spôsoby zakódovania kategorických premenných, ako aj modely typu gradient boosting, s dôrazom na algoritmus LightGBM.

Následne v Kap. 3 sformulujeme praktický problém z pohľadu banky. Taktiež zadefinujeme optimalizačnú úlohu pri osloveniach bankou zákazníkov. Zohľadňujeme pritom nielen očakávaný prínos, ale aj nákladovosť vzhľadom na rôzne typy oslovení.

V Kap. 4 uvádzame ako pripraviť dáta pre trénovanie modelu, konkrétne zadefinujeme závislú a príčinné premenné a vysvetlíme postup tvorby dátovej vzorky.

V Kap.5 prezentujeme aplikáciu meta-učiteľov a kvantifikujeme kauzálny efekt marketingového oslovenia na správanie zákazníkov. Následne v Kap.6 rozširujeme modelovanie o viacúrovňové ošetrenia, pričom modelujeme efekt komunikačných kanálov ako aj ich kombinácií. Taktiež porovnáme výkonnosť použitých modelov z hľadiska vlastností ich odhadov, schopnosti rozlišovať medzi jednotkami s vyšším a nižším efektom oslovenia, ako aj výpočtovej náročnosti jednotlivých prístupov.

V závere diplomovej práce vyhodnotíme výkonnosť modelov, ktoré zohľadňujú príčinné vzťahy medzi marketingovým oslovením a správaním zákazníkov. Poukážeme na aktuálnosť získaných výsledkov a možnosti ich praktického uplatnenia pri tvorbe marketingových stratégií banky.

# 1 Prehľad techník modelovania kauzálnej súvislosti

## 1.1 Kauzalita

V tejto kapitole vysvetlíme, čo je kauzalita a oboznámime sa s hlavnými pojmami, s ktorými sa v rámci práce budeme stretávať. Pri uvedení definícií a zavedení hlavných pojmov budeme vychádzať z [28].

Pod kauzalitou rozumieme vzťah medzi príčinou a následkom. Na rozdiel od korelácie, ktorá iba identifikuje závislosť medzi dvoma premennými, kauzalita skúma mechanizmy, ktoré podmieňujú túto súvislosť. Inak povedané, kauzalita odpovedá na otázku, ako zmena jednej premennej (príčiny) priamo ovplyvňuje druhú premennú (následok). Kauzálna inferencia je proces vyvodzovania záverov o príčinnej súvislosti na základe podmienok, za ktorých daný jav alebo efekt nastáva. Následne uvedieme definície pojmov, s ktorými v rámci práce sa budeme stretávať.

**Definícia 1.1.** *Pod jednotkou rozumieme výskumný objekt. V rámci kauzality jednotka môže byť fyzický objekt, firma, pacient, fyzická osoba.*

**Definícia 1.2.** *Ošetrenie označuje akciu, ktorá sa aplikuje na jednotku. Nech* $\mathcal{Z} = \{z_0, z_1, \ldots, z_m\}$ *označuje množinu možných ošetrení, kde* $m + 1$ *je ich celkový počet. Najčastejším je prípad binárneho ošetrenia, kde skupina jednotiek s* $Z = 1$ *je označená ako ošetrená skupina (angl. treatment group), zatiaľ čo skupina s* $Z = 0$ *ako kontrolná skupina (angl. controlled group) .*

V našej práci sa však zameriavame na fyzické osoby (zákazníkov banky), preto na označenie jednotky v ďalších kapitolách, budeme používať pojem zákazník.

**Definícia 1.3.** *Potenciálny výsledok predstavuje hypotetický výsledok, ktorý by sme pozorovali, ak by konkrétnej jednotke bolo priradené určité ošetrenie. Potenciálny výsledok ošetrenia s hodnotou* $z_k, \quad kde \quad k \in \{0, .., m\}$ *sa označuje ako* $Y(Z = z_k)$.

Pričom výskumné objekty v rôznych časových bodoch sa považujú za odlišné jednotky, lebo rovnaký objekt sa v rôznych časových bodoch môže správať inak. V kauzálnej inferencii skúmame, ako ošetrenie ovplyvňuje výsledky. Ak sa rovnaký objekt pozoruje v rôznych časových bodoch, jeho charakteristiky alebo prostredie sa môžu meniť, čo následne môže

viesť k zmene výsledkov. Pomocou tejto úvahy vieme zachytiť zmeny v správaní, prostredí alebo reakcii jednotky na ošetrenie. V kontexte danej práce to môžeme interpretovať nasledovne: aj keď ide o toho istého zákazníka banky, jeho charakteristiky, preferencie a správanie sa môžu v čase meniť. Napríklad v januári by si zákazník úver nezobral, ale v júni v dôsledku neočakávanej životnej situácie už by sa mohol rozhodnúť inak.

**Definícia 1.4.** *Pozorovaný výsledok je výsledok ošetrenia, ktoré bola skutočne jednotke priradené. Pozorovaný výsledok sa nazýva faktický výsledok a označujeme ho ako $Y_F$, kde $F$ znamená „faktický“. Pričom vzťah medzi potenciálnym a pozorovaným výsledkom je:*

$$Y_F = Y(Z = z), \tag{1}$$

*kde $z$ je ošetrenie, ktoré bolo skutočne jednotke priradené.*

**Definícia 1.5.** *Kontrafaktuálny výsledok je výsledok, ak by jednotka dostala iné ošetrenie. V jednom časovom bode jednotka môže dostať len jedno ošetrenie, teda len jeden potenciálny výsledok je pozorovateľný. Zvyšné nepozorované potenciálne výsledky voláme kontrafaktuálne výsledky. Pri binárnom ošetrení je kontrafaktuálny výsledok definovaný nasledovne:*

$$Y_{CF} = Y(Z = 1 - z), \tag{2}$$

V praxi to znamená, že zákazník nemôže byť v tom istom čase zároveň oslovený aj neoslovený, a teda môžeme pozorovať len jeden z potenciálnych výsledkov, zatiaľ čo druhý musí byť odhadnutý.

Klasické propenzitné modely sa zameriavajú na predikciu pravdepodobnosti správania zákazníkov (napríklad nákup), zatiaľ čo uplift modelovanie sa snaží odhadnúť, ako sa daná pravdepodobnosť zmení v dôsledku marketingového zásahu. Uplift je v tejto súvislosti používané označenie pre individuálny efekt marketingového zásahu. Uplift modelovanie predstavuje aplikačný rámec pre odhad efektu oslovenia.

Cieľom kauzálnej inferencie je odhadnúť efekty ošetrenia na základe dostupných dát, a to takým spôsobom, aby sme získali vplyv príčinnej črty na výsledok. V kontexte našej problematiky tento prístup umožňuje odpovedať na otázky, ako napríklad: „Ako by sa zmenil výsledok jednotky, keby nebola oslovená?“. Na túto otázku odpovedá tzv. uplift modelovanie, ktoré sa zameriava na odhad rozdielu medzi dvoma potenciálnymi výsled-

kami. Úlohou je teda odhadnúť Podmienený Priemerný Efekt Ošetrenia (angl. *Conditional Average Treatment Effect*) - CATE, ktorý odvádzame v nasledujúcej časti.

## 1.2 Odvodenie CATE rovnice

V danej kapitole odvodíme rovnicu na odhad CATE, vysvetlíme ideu a predpoklady, na ktorých je daný odhad postavený. Pri odvodení budeme vychádzať z [6].

Nech máme dátový súbor $\mathcal{D}_i = \{X_i, Z_i, Y_i\}$, $i \in \{1, \ldots, N\}$, kde $X_i \in \mathbb{R}^p$ označuje $p$-rozmernú množinu pozorovaných vstupných čŕt pre jednotku $i$, $Z_i$ je príčinná premenná, ktorá v prípade dvoch typov ošetrenia nadobúda hodnotu 1, ak jednotka bola vystavená ošetreniu, a 0, ak nebola. $Y_i$ je potenciálnym výstupom zákazníka. Potom individuálny efekt ošetrenia v danom prípade (angl. *Individual treatment effects*) je definovaný nasledovne:

$$\tau_i = Y_i(1) - Y_i(0). \tag{3}$$

Daný efekt je vyjadrený ako rozdiel medzi dvoma hypotetickými situáciami: medzi potenciálnym výsledkom zákazníka, ak by bol ošetrený $Y_i(1)$, a potenciálnym výsledkom zákazníka, ak by nebol ošetrený $Y_i(0)$. Dané situácie sú pre tú istú jednotku navzájom vylučujúce, teda pozorujeme len jeden z týchto potenciálnych výsledkov. Potom vieme reprezentovať potenciálny výstup nasledovne:

$$Y_i = Z_i Y_i(1) + (1 - Z_i) Y_i(0). \tag{4}$$

Úlohou je teda odhadnúť Podmienený Priemerný Efekt Ošetrenia (angl. *Conditional Average Treatment Effect*) - CATE, ktorý je definovaný nasledovne:

$$\tau(x_i) = \mathbb{E}[Y_i(1) - Y_i(0) | X_i = x_i]. \tag{5}$$

Taktiež môže zohrávať pravdepodobnosť, že jednotka bude vystavená ošetreniu. Formálne zapísané:

$$e(x_i) = \mathbb{P}[Z_i = 1 | X_i = x_i], \tag{6}$$

pričom platí, že $0 < e(x_i) < 1$. Použitie ostrej nerovnosti zaručuje, že jednotka s črtami $x_i$ sa môže s nenulovou pravdepodobnosťou nachádzať v oboch skupinách. Ak by $e(x_i) = 0$ alebo $e(x_i) = 1$, pre daný typ jednotiek by sme nemali pozorovania v jednej zo skupín a tým pádom by nebolo možné kvantifikovať efekt ošetrenia.

Navyše podľa [6] na získanie spoľahlivého odhadu očakávaného individuálneho účinku ošetrenia (CATE) musia byť splnené nasledujúce tri predpoklady:

- **Neporušenosť** (angl. *unconfoundedness*) - neexistujú žiadne nezaznamenané charakteristiky, ktoré by mohli ovplyvniť, či jednotka bude vystavený ošetreniu.
  Možné porušenie: Keď boli určití zákazníci manuálne vylúčení z kampane na základe demografických charakteristík (napr. na základe etnickej príslušnosti), ktoré sa v dostupných dátach nenachádzajú.

- **Spoločná podpora** (angl. *common causal support*) hovorí, že priraďovanie ošetrenia nie je deterministické a jednotka s vektorom charakteristík $X_i$ s nenulovou pravdepodobnosťou môže byť pozorovaný v akejkoľvek zo skupín.
  Možné porušenie: Niektoré segmenty zákazníkov môžu byť cielene vynechané z oslovenia na základe obchodnej stratégie banky alebo interných pravidiel. Napríklad, banka na základe vlastných politík neoslovuje maloletých. Tým pádom, určité skupiny zákazníkov sú vylúčené z kampaní.

- **Predpoklad stabilnej hodnoty účinku ošetrenia pre jednotku** (angl. *Stable Unit Treatment Value Assumption - SUTVA*), čo hovorí, že výsledok ošetrenia jedného nie je ovplyvnený pridelením ošetrenia iným objektom.
  Možné porušenie: Môže dôjsť k situácii, kde interakcia medzi zákazníkmi ovplyvní výsledky. Napríklad, ak jeden zákazník dostane oslovenie s konkrétnou ponukou a následne to zdieľa s inými ľuďmi, títo môžu byť nepriamo ovplyvnení aj napriek tomu, že banka ich priamo neoslovila.

Pričom **Neporušenosť** by sme formálne vedeli zapísať nasledovne:

$$(Y_i(0), Y_i(1)) \perp\!\!\!\perp Z_i \mid X_i, \tag{7}$$

kde symbol $\perp\!\!\!\perp$ označuje pravdepodobnostnú nezávislosť. Samotný predpoklad sa interpretuje tak, že ak poznáme všetky relevantné črty, priradenie ošetrenia je podmienečne nezávislé od potenciálnych výsledkov. Za predpokladu platnosti rovnosti (7), platí aj:

$$(Y_i(0), Y_i(1)) \perp\!\!\!\perp Z_i \mid e(X_i). \tag{8}$$

Teda ak priradenie ošetrenia po zohľadnení všetkých čŕt nie je ovplyvnené žiadnou nepozorovanou premennou, potom platí, že nie je ovplyvnené ani po ich zohľadnení pomocou

pravdepodobnosti zaradenia do ošetrenej skupiny $e(x_i)$ [21]. Daná pravdepodobnosť slúži ako jednorozmerné zhrnutie pôvodnej $d$-rozmernej množiny čŕt $x_i \in \mathcal{X}$.

Za predpokladu neporušenosti 7 a spoločnej podpory je možné odhadnúť CATE z pozorovaných dát $\mathcal{D}_i = \{X_i, Z_i, Y_i\}$ pre $i \in \{1, \ldots, N\}$. Definujeme podmienenú očakávanú hodnotu potenciálneho výstupu $\mu_z(x_i)$, teda očakávanú hodnotu, ak by jednotka $i$ s črtami $x_i$ dostala ošetrenie $z$:

$$\mu_z(x_i) = \mathbb{E}[Y_i^{(z)} \mid X_i = x_i], \tag{9}$$

kde $z \in \{0, 1\}$. Za predpokladu neporušenosti (7) môžeme do podmienky v (9) pridať $Z_i = z$. Dostaneme:

$$\mathbb{E}[Y_i^{(z)} \mid X_i = x_i] = \mathbb{E}[Y_i^{(z)} \mid Z_i = z, X_i = x_i]. \tag{10}$$

Následne v dôsledku platnosti (4) dostaneme:

$$\mu_z(x_i) = \mathbb{E}[Y_i(z) \mid Z_i = z, X_i = x_i] = \mathbb{E}[Y_i \mid Z_i = z, X_i = x_i]. \tag{11}$$

Nakoniec z rovnice (11) vyplýva:

$$\tau(x_i) = \mathbb{E}[Y_i(1) - Y_i(0) \mid X_i = x_i] = \mathbb{E}[Y_i \mid Z_i = 1, X_i = x_i] - \mathbb{E}[Y_i \mid Z_i = 0, X_i = x_i]. \tag{12}$$

Rovnica (12) poskytuje základ pre odhad CATE, pretože obe podmienené stredné hodnoty na jej pravej strane môžeme odhadnúť z dát.

Existuje viacero nástrojov na odhad podmieneného priemerného efektu ošetrenia (CATE). Jedným z nástrojov je použitie meta-učiteľov (angl. *meta-learners*). Ich idea je založená na použití a následnom kombinovaní výstupov rôznych základných učiteľov (angl. *base-learners*). Inak povedané, trénovaním základných modelov odhadujeme pravdepodobnosť výsledku v závislosti od toho, či bola jednotka oslovená. Následne sú ich výstupy kombinované prostredníctvom meta-učiteľa, ktorý slúži na odhad konečného efektu [6], [17]. Rôzne spôsoby kombinovania výstupov zo základných modelov môžu viesť k odlišným odhadom efektu. Z tohto dôvodu sa zameriame na viacero prístupov k odhadu CATE a porovnáme ich. V nasledujúcej kapitole predstavíme metódy, ktoré budeme v práci používať.

## 1.3 Metódy odhadu CATE pri dvoch typoch ošetrenia

### S-learner pri dvoch typoch ošetrenia

Metóda S-learner spočíva v tom, že pridá črtu $Z_i$, ktorá hovorí o tom, či jednotka $i$ bola ošetrená, ako ďalšiu vstupnú črtu do $X_i$. S-learner teda zodpovedá jednému spoločnému modelu pre ošetrenú a kontrolnú skupinu, potom odhady *CATE* sa získajú pomocou rozdielu medzi dvoma podmienenými potenciálnymi výsledkami. Teda postup pri S-learneri by sme vedeli zapísať nasledovne:

1. Odhadneme jeden spoločný model pre všetky skupiny:

$$\mu(x, z) := \mathbb{E}[Y^{\text{obs}} \mid X = x, Z = z]. \tag{13}$$

   Pričom $\hat{\mu}$ je odhad pre funkciu (13).

2. Odhad CATE je potom definovaný ako [17]:

$$\hat{\tau}(x) = \hat{\mu}(x, 1) - \hat{\mu}(x, 0), \tag{14}$$

pričom $\hat{\mu}(x, 1)$ sú odhady očakávaného výsledku pre jednotku, ak by dostala ošetrenie, zatiaľ čo $\hat{\mu}(x, 0)$ sú odhady očakávaného výsledku pre jednotku, ak by nedostala ošetrenie. Tým pádom, pre každú jednotku $i$ dostaneme odhad pravdepodobnosti za podmienky, že príčinná črta bola $Z_i = 1$ a taktiež za podmienky $Z_i = 0$. Následne sa dané odhady odčítajú medzi sebou, ako je to znázornené v (14), čím dostaneme výsledný odhad.

S-learner používa rovnaké regularizačné podmienky pre obe skupiny a taktiež nepredpokladá odlišnú distribúciu vstupných čŕt aj pre ošetrenú, aj pre kontrolnú skupinu, čo môže mať vplyv na presnosť pri odhade efektu ošetrenia. Výhodou daného nástroja je jeho ľahšia implementácia a nižšia výpočtová náročnosť [6].

**T-learner pri dvoch typoch ošetrenia**

Na rozdiel od S-learnera T-learner odhaduje dva samostatné regresné modely, jeden pre ošetrenú skupinu a druhý pre kontrolnú skupinu. Postup vieme zapísať nasledovne:

1. Rozdelíme dátovú vzorku na dve skupiny na základe priradeného ošetrenia:

   - $(X^0, Y^0)$: jednotky, ktoré ošetrenie nedostali $(Z = 0)$,
   - $(X^1, Y^1)$: jednotky, ktoré ošetrenie dostali $(Z = 1)$.

2. Natrénujeme model pre kontrolnú skupinu:

$$\mu_0(x) = \mathbb{E}[Y(0) \mid X = x]. \tag{15}$$

   Odhadnutú funkciu označujeme ako $\hat{\mu}_0$.

3. Natrénujeme model pre ošetrenú skupinu:

$$\mu_1(x) = \mathbb{E}[Y(1) \mid X = x]. \tag{16}$$

   V tomto prípade odhadnutú funkciu označujeme ako $\hat{\mu}_1$.

4. Na základe odhadov (15), (16) vypočítame odhad podmieneného priemerného efektu ošetrenia (CATE) [17]:

$$\hat{\tau}(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x). \tag{17}$$

V danom prípade rozdelíme dátovú sádu na základe pridelenej skupiny a následne natrénujeme dva samostatné modely. Potom použitím oboch modelov na celej dátovej sade získame odhady výsledkov pre každú jednotku v oboch skupinách – s ošetrením aj bez neho. Rozdiel týchto dvoch odhadov predstavuje odhad individuálneho kauzálneho efektu ako znázorňuje (17).

T-learner odhaduje samostatný model pre každú skupinu, čím zachováva distribučné rozdiely medzi dvoma skupinami a dokáže brať do úvahy aj rôzne úrovne hladkosti a riedkosti, ktoré môžu byť prítomné v dátach. Na druhej strane, spoločná informácia medzi skupinami sa môže zanedbať. Taktiež pri rozdelení dát do kontrolnej a ošetrenej skupiny sa môže stať, že v ošetrenej skupine, ktorá je obyčajne menšia, sa ponechá príliš málo pozorovaní na odhad. Zároveň T-learner predpokladá, že vzťahy medzi vstupnými črtami a výsledkom môžu byť pre každú skupinu nezávislé. Ak však tieto vzťahy obsahujú spoločné štruktúry, T-learner ich nedokáže efektívne zachytiť, čo môže viesť k suboptimálnym výsledkom v porovnaní s prístupmi, ktoré explicitne zdieľajú informácie medzi skupinami, ako napríklad X-learner, ktorý uvádzame nižšie [6].

**X-learner pri dvoch typoch ošetrenia**

Ďalším nástrojom je X-learner, ktorý pozostáva z dvoch fáz. V prvej fáze postupujeme analogicky ako pri T-learneri, čiže odhadneme modely rovnako ako v (15), (16). V druhej

fáze pomocou nich odhadneme hodnotu jednotky opačnej skupine, čím dostaneme hodnoty $\hat{\tau}_1$, $\hat{\tau}_0$. Taktiež v danom prístupe potrebujeme zvoliť vhodnú funkciu, ktorou vieme získané výsledky vyvážiť, čím dostaneme výsledný odhad pre *CATE*. Uvedený postup by sme teda vedeli zapísať nasledovne [17]:

1. Odhadneme modely pre obe skupiny:

$$\mu_1(x) = \mathbb{E}[Y(1) \mid X = x], \quad \mu_0(x) = \mathbb{E}[Y(0) \mid X = x]. \tag{18}$$

   Analogicky odhady označujeme ako $\hat{\mu}_1$, $\hat{\mu}_0$.

2. Pomocou odhadov (18) vypočítame imputované efekty pre jednotku $i$:

$$\tilde{D}_i^1 = Y_i^1 - \hat{\mu}_0(X_i^1), \quad \tilde{D}_i^0 = \hat{\mu}_1(X_i^0) - Y_i^0. \tag{19}$$

3. Natrénujeme model pomocou ktorého odhadneme imputované efekty pre obe skupiny:

$$\tau_1(x) = \mathbb{E}[\tilde{D}^1 \mid X = x], \quad \tau_0(x) = \mathbb{E}[\tilde{D}^0 \mid X = x]. \tag{20}$$

4. Skombinujeme odhadnuté imputované efekty $\hat{\tau}_1(x)$, $\hat{\tau}_0(x)$ z (20) do výsledného odhadu CATE:

$$\hat{\tau}(x) = g(x)\hat{\tau}_0(x) + (1 - g(x))\hat{\tau}_1(x), \tag{21}$$

kde $g(x)$ je nejaká váhová funkcia, napr. odhad pravdepodobnosti zaradenia do ošetrenej skupiny $e(x)$, ktorá bola zavedená v (6). Môžu sa však využíť aj iné spôsoby ako napríklad, vyváženie podľa veľkosti vzoriek alebo rovnomerné váhy. X-learner vykazuje spoľahlivé výsledky najmä v prípadoch, keď jedna zo skupín obsahuje výrazne viac pozorovaní ako druhá, pretože využíva aj imputované efekty uvedené v (19), ktoré zohľadňujú rozdiel medzi pozorovanými a odhadovanými hodnotami. Pričom $\tilde{D}^1$ vyjadruje, o koľko sa pozorovaný výstup $Y^1$ pre ošetrenú skupinu líši od očakávaného výstupu, ak by jednotky nachádzajúce sa v danej skupine neboli ošetrené ($\hat{\mu}_0(X_1)$). Analogicky $\tilde{D}^0$ vyjadruje, o koľko by sa zmenil očakávaný výsledok pre jednotky z kontrolnej skupiny ($\hat{\mu}_1(X_0)$) v porovnaní s tým, ak by boli ošetrené [17].

### 1.3.1 Metriky výkonnosti pri dvoch typoch ošetrenia

Pri hodnotení modelov predikujúcich individuálny efekt ošetrenia je dôležité si uvedomiť, že pre tú istú jednotku nie je možné súčasne pozorovať výsledok s ošetrením a bez neho. Je však ale možné odhadnúť účinok oštrenia na úrovni skupiny výpočtom tzv. lokálneho priemerného účinku ošetrenia. Táto myšlienka je základom metriky AUUC(*Area Under the Uplift Curve – AUUC*), ktorú uvádzame nižšie. Pričom vychádzať budeme z [4], [29].

**AUUC**

Na vyhodnotenie výkonnosti výsledného modelu použijeme uplift krivku (angl. *uplift curve*). Konkrétne by nás zaujímala plocha pod krivkou upliftu (angl. *Area Under the Uplift Curve – AUUC*). Pričom $AUUC$ by sme formálne vedeli definovať:

**Definícia 1.6.** *Nech $f(D,k)$ označuje prvých $k$ jednotiek v množine $D$, zoradených podľa predikcie modelu $f$. Nech $|T|$ a $|C|$ sú počty pozorovaní v ošetrenej ($T=1$) a kontrolnej ($T=0$) skupine. Potom $AUUC$ pre model $f$ je definovaná nasledovne:*

$$AUUC(f) = \int_0^1 V(f,x)\,dx \approx \sum_{k=1}^{n} V(f,k), \tag{22}$$

*kde:*

$$V(f,k) = \frac{1}{|T|}\sum_{i\in f(D,k)} y_i \cdot 1[z_i = 1] - \frac{1}{|C|}\sum_{j\in f(D,k)} y_j \cdot 1[z_j = 0]. \tag{23}$$

Funkcia $V(f,k)$ vyjadruje rozdiel medzi priemerným výstupom ošetrenej a kontrolnej skupiny v rámci prvých $k$ pozorovaní zoradených podľa modelu $f$. Inými slovami, $AUUC$ predstavuje plochu pod uplift krivkou.

Táto metrika sa počíta nasledovne: najprv výsledky CATE z testovacej množiny zoradíme zostupne. Následne zoradené hodnoty rozdelíme do $n$ rovnakých skupín (tzv. kvantilov). V nasledujúcom kroku pre každú skupinu podľa (23) odhadneme priemerný efekt ošetrenia v skupine. Model, ktorý dosahuje vyššiu hodnotu $AUUC$ dokáže lepšie identifikovať jednotky s vyšším prínosom z oslovenia a je preto považovaný za výkonnejší.

Samotná uplift krivka zobrazuje kumulatívny efekt oslovenia pre každú zo skupín. Na vodorovnej osi sa označuje poradie $k = 1,\ldots,n$, ktorá vznikla zostupným zoradením predikovanej hodnoty CATE, zatiaľ čo na zvislej osi sú hodnoty $V(f,k)$, ktoré predstavujú odhad rozdielu v priemernom výstupe medzi ošetrenou a kontrolnou skupinou v danom

kvantile $k$. Pri vizualizácii výkonnosti uplift modelov musíme brať do úvahy aj referenčnú krivku, ktorá predstavuje výkonnosť referenčného modelu (angl. *baseline*) modelu. Pričom jej hodnotu v kvantile $k$ môžeme vyjadriť nasledovne:

$$V_{\text{ref}}(k) = \frac{k}{n} \cdot \text{ATE}, \tag{24}$$

kde

$$\text{ATE} = \mathbb{E}[Y \mid Z = 1] - \mathbb{E}[Y \mid Z = 0] \tag{25}$$

je priemerný efekt ošetrenia v celom dátovom súbore.

Pričom čím vyššie sa uplift krivka nachádza nad referenčnou krivkou, tým vyššiu rozlišovaciu schopnosť daný model vykazuje. Inými slovami, model vie lepšie rozlíšiť jednotky s vysokým odhadovaným efektom ošetrenia od tých, ktorých očakávaný prínos z ošetrenia je nízky alebo zanedbateľný.

## 1.4 Metódy odhadu CATE pri viacerých ošetreniach

Odhad CATE za prítomnosti viacerých typov ošetrenia rozširuje štandardný prístup. Ošetrenie už teda nie je binárnou premennou, ktorá indikuje prítomnosť ošetrenia, ale nadobúda viacero možných hodnôt. Cieľom je identifikovať, ktorý typ ošetrenia má najvyšší očakávaný efekt pre konkrétnu jednotku. Výsledné modely tak umožňujú rozhodovanie nielen o tom, **či** zasiahnuť, ale aj **akým spôsobom** zasiahnuť.

Pri odhade CATE jedno z možných ošetrení zvolíme ako *referenčné* a voči nemu budeme odhadovať rozdiely účinkov ostatných ošetrení. Uvažujme o prostredí s viacerými možnými ošetreniami, reprezentovanými množinou $\mathcal{Z} = \{z_0, z_1, \ldots, z_m\}$, kde $z_0$ predstavuje referenčné (napr. žiadne) ošetrenie. Potom odhadované individuálne efekty vieme interpretovať nasledovne [1]:

$$\tau_k(x) = \mathbb{E}[Y(z_k) - Y(z_0) \mid X = x], \quad \text{pre } k = 1, \ldots m, \tag{26}$$

kde $Y(z_k)$ predstavuje výstup, ktorý by sme pozorovali, keby bolo jednotke priradené ošetrenie $z_k$, a $Y(z_0)$ výstup pri referenčnom ošetrení. Rovnica vyjadruje CATE pre ošetrenie $z_k$ pre jednotku s črtami $x$. Pričom nás bude zaujímať pre aké ošetrenie jednotka dostala najvyššiu hodnotu CATE.

Pri viacerých typoch ošetrenia sa teda odhad CATE pomocou modelovacích techník popísaných vyššie modifikuje. V nasledujúcej časti sa preto podrobnejšie venujeme vysvetleniu, ako sa mení implementácia pri odhade individuálnych efektov v prípade viacerých typov ošetrenia.

**S-learner pri viacerých ošetreniach**

V prípade viacerých typov ošetrenia pridávame do množiny čŕt všetky príčinné premenné z množiny $\mathcal{Z} = \{z_0, z_1, \ldots, z_k\}$ a následne trénujeme jeden spoločný model.

$$\mu(x, z) = \mathbb{E}[Y \mid Z = z, X = x]. \tag{27}$$

Potom z daného modelu pre každé ošetrenie $z_k$ voči referčnému ošetreniu $z_0$ odhadneme CATE [1]:

$$\tau_k(x) = \mu(x, z_k) - \mu(x, z_0), \quad \text{pre } k = 1, \ldots m. \tag{28}$$

**T-learner pri viacerých ošetreniach**

Pri použití T-learnera pre každé ošetrenie $z_k \in \mathcal{Z}$ trénujeme samostatný model $\mu_{z_k}(x) = \mathbb{E}[Y(z_k) \mid X = x]$, ktorý znázorňuje očakávaný výstup pri tomto ošetrení. Individuálny efekt ošetrenia $z_k$ oproti referenčnému ošetreniu $z_0$ sa následne vypočíta ako:

$$\tau_k(x) = \mu_{z_k}(x) - \mu_{z_0}(x), \quad \text{pre } k = 1, \ldots m. \tag{29}$$

Pri použití T-learnera s viacerými ošetreniami trénujeme samostatný model pre každú unikátnu hodnotu ošetrenia. Každý model je vytvorený len na podmnožine dát, v ktorej dané ošetrenie skutočne nastalo.

**X-learner pri viacerých ošetreniach**

1. Pre každé ošetrenie $z_k \in \mathcal{Z}$ odhadujeme očakávaný výstup analogicky ako pri T-Learneri.

2. Pomocou získaných odhadov vypočítame imputované efekty, ktoré vyjadrujú kontrafaktuálne rozdiely medzi ošetreniami a sú uvedené v (19) nasledovne:

$$D_i^{(z_0)} = \mu_{z_k}(x_i) - Y_i, \qquad \text{ak } Z_i = z_0, \tag{30}$$

$$D_i^{(z_k)} = Y_i - \mu_{z_0}(x_i), \qquad \text{ak } Z_i = z_k. \tag{31}$$

3. Podobne ako v (20) trénujeme modely na imputované efekty $D^{(z_0)}$ a $D^{(z_k)}$ ako závislé premenné, čím získame odhady $\tau^{(z_0)}(x)$ a $\tau^{(z_k)}(x)$ pre $k = 1, \dots m$.

4. Finálny odhad CATE pre ošetrenie $z_k$ voči referenčnému ošetreniu $z_0$ sa potom určuje [29]:

$$\tau_{z_k}(x) = \frac{e_{z_k}(x)}{e_{z_k}(x) + e_{z_0}(x)} \cdot \tau^{(z_0)}(x) + \frac{e_{z_0}(x)}{e_{z_k}(x) + e_{z_0}(x)} \cdot \tau^{(z_k)}(x), \tag{32}$$

kde $e_{z_k}$ je pravdepodobnosť zaradenia jednotky do skupiny $z_k \in \mathcal{Z}$ a je definovaná:

$$e_{z_k}(x) = \mathbb{P}(Z = z_k \mid X = x) \quad \text{pre } k = 1, \dots m. \tag{33}$$

### 1.4.1 Metriky výkonnosti meta učiteľov pri viacerých ošetreniach

**AUUC pri v prípade viacerých ošetrení**

Danú metriku sme predstavili v Kap. 1.3.1, kde sme uvažovali o prípade dvoch typov ošetrenia. V tejto časti vysvetlíme ako sa mení jej výpočet, keď uvažujeme o viacerých typoch ošetrenia. Pričom budeme vychádzať z [29].

Pre každú jednotku v testovacej množine získame individuálne efekty (CATE) pre všetky ošetrenia $z_k \in \mathcal{Z}$. Pre každú jednotku sa však vyberie iba tá hodnota, ktorá má najvyšší predikovaný efekt. Ak sú všetky odhady záporné, tak sa jednotke priradí referenčné ošetrenie. Formálne zapísané:

$$\tau^*(x) = \max_{k,0}(\tau_k(x), z_0). \tag{34}$$

Rovnako ako pri dvoch typoch ošetrenia hodnoty $\tau^*(x)$ sa zoradia zostupne a následne sa rozdelia do určeného počtu kvantilov. Zmena nastáva vo výpočte skutočného priemerného efektu. V každom kvantile sa porovnávajú iba tie jednotky, ktoré:

- dostali referenčné ošetrenie
- dostali presne to ošetrenie, ktoré im model odporučil.

Na základe týchto dvoch skupín sa vypočíta priemerný rozdiel vo výstupe ako odhad reálneho efektu modelom odporúčaného rozhodovania. Potom rovnicu (23) vzhľadom na úvahy vyššie prepísať nasledovne:

Nech $m_i$ je ošetrenie odporučené modelom a $z_i$ je skutočne priradené ošetrenie. Potom kumulatívny uplift v top-$k$ jednotkách je definovaný ako:

$$V(f,k) = \frac{1}{|z_k|} \sum_{i \in z_k} y_i \cdot 1[z_i = a_i] - \frac{1}{|C_k|} \sum_{j \in C_k} y_j \cdot 1[z_j = z_0], \tag{35}$$

kde $z_k$ je množina jednotiek v top-$k$ s priradeným odporúčaným ošetrením a $C_k$ sú jednotky s priradeným referenčným ošetrením.

# 2 Prehľad relevantných techník strojového učenia

V danej kapitole zavedieme techniky strojového učenia, ktoré v rámci modelovania budeme používať.

## 2.1 Klasifikačné modely

V danej podkapitole sa budeme venovať vysvetleniu klasifikačných modelov, pričom vychádzať budeme z [10].

Hlavným cieľom klasifikačnéj analýzy je na základe hodnôt vektorov vlastností vytvoriť klasifikátor $C$, ktorý sa použije na predikciu tried objektov. Na dosiahnutie tohto cieľa je potrebné špecifikovať $m$ tried. Často je klasifikácia binárna, teda $m = 2$, alebo multiklasifikačná, teda $m > 2$.

Klasifikátor $C$ sa obvykle konštruuje na trénovacej vzorke dát, ktorá pozostáva z $n$ objektov, ktoré sú charakterizované vektormi $x_1, \ldots, x_n \in \mathbb{R}^p$, teda každý objekt obsahuje $p$ čŕt, ktoré vyjadrujú jeho charakteristiky. Trénovacia vzorka dát obsahuje aj známe správne klasifikácie objektov $c_1, \ldots, c_n \in \{1, \ldots, m\}$, na základe ktorých sa model učí identifikovať vzory a priraďovať klasifikátory.

Klasifikátor je reprezentovaný pomocou diskriminačných funkcií $l_1, \ldots, l_m$ definovaných v $\mathbb{R}^p$. Diskriminačná funkcia je funkcia, ktorá sa používa na vyhodnotenie toho, do ktorej triedy by mal byť objekt $j$ s vektorom čŕt $x_j$ priradený. Pričom každá trieda má vlastnú diskriminačnú funkciu. Systém diskriminačných funkcií $l_1, \ldots, l_m$ generuje klasifikačné regióny $R_1, \ldots, R_m$, ktoré určujú do ktorej triedy bude nový objekt priradený. Klasifikáciu objektu do jednotlivých tried na základe vektoru vlastností $x \in \mathbb{R}^p$ a diskriminačných funkcií $l_i(x)$ môžeme definovať nasledovne:

$$x \in R_j \Leftrightarrow j = \min\left(\underset{i}{\operatorname{argmax}}\ l_i(x)\right). \tag{36}$$

### 2.1.1 Klasifikačné stromy

Klasifikačné stromy sú stromové štruktúry, používané na klasifikáciu dát. Každý neterminálny uzol v strome predstavuje rozhodovaciu podmienku, na základe ktorej sa dáta rozdeľujú. Končia sa terminálnymi uzlami, teda tými, pro ktorých ďalšie delenie už nie je možné a dostávame konečnú klasifikáciu, resp. výsledok. Pričom hranica sa určuje na

základe určitej zvolenej podmienky na jednu zo vstupných čŕt. Na konštrukciu klasifikačného stromu sa bežne používa postup rekurzívneho delenia. Daný postup spočíva v tom, že v každom uzle na základe určených kritérii (hĺbka, podiel objektov patriacích do rôznych tried) sa rozhodneme, či je daný uzol terminálny alebo nie. Ak vyhlásime uzol za terminálny, priradíme mu na základe *majority vote* označenie triedy. V opačnom prípade priradíme podmienku na jednu zo vstupných čŕt, na základe ktorej sa dáta priradia potomkom daného uzlu. Vytvoríme konečný počet možných podmienok, ktoré by mohli byť uzlu priradené a na základe vopred stanoveného kritéria vyberieme najlepšiu podmienku.

## 2.2 Logistická regresia

Logistická regresia je štatistický nástroj, ktorý slúži na modelovanie pravdepodobnosti klasifikácie objektov do $m$ tried na základe hodnôt vektorov vlastností $x$. Podľa [11] sa definuje nasledovne:

$$\forall j = 1, ..., m-1: \ \ln \frac{\Pr(G = j \,|\, X = X_i)}{\Pr(G = m \,|\, X = X_i)} = \beta_{i0} + \beta_i^T x. \tag{37}$$

Pričom model je špecifikovaný pomocou $(m-1)$ logitových transformácií, ktoré zohľadňujú, že súčet pravdepodobností je jedna. Posledná trieda $m$ je v tomto prípade braná ako referenčná. Potom pravdepodobnosti zariadenia k jednej z tried môžeme vyjadriť nasledovne:

$$\forall j = 1, ..., m-1: \ \ P(G = j|X = X_i) = \frac{\exp(\beta_0 + \beta^T X)}{1 + \sum_{i=1}^{m-1} \exp(\beta_0 + \beta^T X)}, \tag{38}$$

$$P(G = m|X = X_i) = \frac{1}{1 + \sum_{n=1}^{m-1} \exp(\beta_0 + \beta^T X)}. \tag{39}$$

V prípade, že počet tried $m > 2$, ide o multinomiálnu logistickú regresiu. Pri jej implementácii sa často využíva *softmax* kódovanie, ktoré umožňuje modelovať všetky triedy symetricky bez potreby určovania referenčnej triedy [13]. Detaily k danému prístupu sú uvedené v Prílohe A.

Pre binárny prípad, kde $m = 2$, $G \in \{0, 1\}$:

$$P(G = 1|X = X_i) = \frac{\exp(\beta_0 + \beta^T X)}{1 + \sum_{i=1}^{m-1} \exp(\beta_0 + \beta^T X)}, \tag{40}$$

$$P(G = 0|X = X_i) = \frac{1}{1 + \sum_{n=1}^{m-1} \exp(\beta_0 + \beta^T X)}. \tag{41}$$

## 2.3 Gradient Boosting

V tejto časti predstavíme metódu gradient boosting, ktorá patrí do triedy boostingových algoritmov. Pričom budeme vychádzať z [5], [7]. Cieľom boostingu je vytvoriť silný klasifikátor zo slabých klasifikátorov. Pri boostingu sa jednotlivé stromy netrénujú nezávisle, ale sekvenčne, čiže každý nový strom sa učí na základe už získanej informácií z predchádzajúcich stromov.

Nech máme trénovaciu množinu $D = \{(X_i, y_i)\}_{i=1}^{N}$, cieľom gradient boostingu je nájsť takú aproximáciu $\hat{F}(x)$ k neznámej cieľovej funkcii $F^*(x)$, ktorá mapuje vstupné črty $X$ na výstupné hodnoty $y$, a to tým, že minimalizuje očakávanú hodnotu zvolenej stratovej funkcie (angl.*loss function*) $L(y, F(x))$. Gradient boosting teda vytvára aditívnu aproximáciu $F^*(x)$ ako vážený súčet slabých modelov. Formálne, pre každú iteráciu $m \in \{1, \ldots, M\}$, môžeme danú aproximáciu vyjadriť nasledovne:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x). \tag{42}$$

Výslednú aproximáciu cieľovej funkcie $F(x)$ potom vieme zapísať v nasledovnom tvare:

$$\hat{F}(x) = \sum_{m=1}^{M} \gamma_m h_m(x), \tag{43}$$

kde $h_m(x)$ sú slabé modely a $\gamma_m$ sú váhy modelu $h_m$ v danej iterácii. Na začiatku sa zvolí konštantný odhad:

$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{N} L(y_i, \gamma). \tag{44}$$

V každej iterácii sa model aktualizuje hľadaním novej funkcie $h_m$ a váh $\gamma_m$, ktoré minimalizujú chybu:

$$(\gamma_m, h_m) = \arg\min_{\gamma, h} \sum_{i=1}^{N} L(y_i, F_{m-1}(x_i) + \gamma h(x_i)). \tag{45}$$

Pričom daná oprimalizáčna úloha sa rieši pomocou gradientného zostupu na dátovom súbore $\mathcal{D} = \{(x_i, r_{mi})\}_{i=1}^{N}$, kde pseudoreziduá $r_{mi}$ sú vypočítané ako:

$$r_{mi} = \left. \frac{\partial L(y_i, F(x))}{\partial F(x)} \right|_{F(x)=F_{m-1}(x)}. \tag{46}$$

V prípade nedostatočnej regulácie gradient boosting môže byť náchylný na preučenie (angl. *overfitting*), inak povedané, nízku generalizáčnu schopnosť na nových datách. Jedným z regularizačných parametrov je počet komponentov $M$, teda počet slabých modelov

(napr. rozhodovacích stromov), ktoré sú postupne pridávané do aditívnej rovnice (42). Ďalším nástrojom regulácie modelu je napríklad zmenšenie kroku (angl.*shrinkage*). Zmenšením kroku sa spomaľuje učenie modelu, čím sa vie znižit riziko preučenia a tým pádom sa zlepšuje generalizáčna schopnosť modelu. Daný paramater by sa dal zaviesť modifikáciou (42) tak, že sa každý nový prírastok do modelu bude teraz násobiť parametrom rýchlosti učenia (označ. `learning_rate`).

$$F_m(x) = F_{m-1}(x) + \nu \cdot \gamma_m h_m(x), \quad \nu \in (0, 1], \tag{47}$$

kde $\nu$ označuje zavedený parameter rýchlosti učenia. Ďalšou formou regularizácie sú isté obmedzenia na stromy pomocou ktorých sa model učí. Do takých obmedzení patria napríklad:

- maximálna hĺbka stromu (`max_depth`);
- počet čŕt pri hľadaní najlepšieho rozdelenia (`max_features`);
- miera náhodného výberu dát (`subsample`)– akú časť dát použije každý nový strom na učenie;
- minimálny počet pozorovaní na rozdelenie uzla (`min_samples_split`).

## 2.4 Gradient Boosted Rozhodovacie Stromy

Gradient Boosted Rozhodovacie Stromy (angl. *Gradient Boosted Decision Trees – (GBDT)* je aplikácia gradient boostingu, kde sa za slabé modely používajú rozhodovacie stromy. Pričom rozhodovacie stromy sú učené sekvenčne a každý nový strom sa učí na základe gradientov stratovej funkcie z predchádzajúcich iterácií. Algoritmus GBDT využíva rozhodovacie stromy na naučenie funkcie z priestoru vstupov $\mathcal{X}^s$ do gradientového priestoru $\mathcal{G}$. Nech máme trénovaciu dátovú sádu $O$ , ktorá obsahuje $n$ nezávislých a rovnako rozdelených pozorovaní $\{x_1, \ldots, x_n\}$, kde $\forall i = 1, \ldots, n$ každý vektor $x_i$ je vektor z priestoru $\mathcal{X}^s$. V každej iterácii sa pre všetky pozorovania na základe predikcie modelu v tejto iterácii počítajú záporné gradienty stratovej funkcie, ktoré označujeme ako $\{g_1, \ldots, g_n\}$. Strom následne rozdeľuje dáta v každom uzle podľa črty, ktorá maximalizuje informačný zisk, ktorý je bežne vyjadrený varianciou po rozdelení. Variancia po rozdelení podľa črty $j$ v bode $d$ je definovaná nasledovne [16]:

$$V_{j|O}(d) = \frac{1}{n_O} \left( \frac{\left( \sum\limits_{x_i \in O : x_{ij} \leq d} g_i \right)^2}{n^j_{l|O}(d)} + \frac{\left( \sum\limits_{x_i \in O : x_{ij} > d} g_i \right)^2}{n^j_{r|O}(d)} \right), \tag{48}$$

kde:

- $n_O$ je počet všetkých pozorovaní v uzle $O$,
- $n_{j_l|O}(d)$ označuje počet pozorovaní, pre ktoré platí $x_{ij} \leq d$ a ktoré sú priradené do ľavého potomkového uzla,
- $n_{j_r|O}(d)$ je počet pozorovaní, pre ktoré platí $x_{ij} > d$ a ktoré sú priradené do pravého potomkového uzla.

Pre črtu $j$ rozhodovací strom vyberie bod rozdelenia $d^*_j = \arg\max_d V_j(d)$, teda taký bod, pri ktorom je variancia maximálna. Následne sa dáta v aktuálnom uzle rozdelia podľa hodnoty črty $j$ v bode $d^*_j$ na dve vetvy – do ľavého a pravého potomkového uzla. V tomto kontexte variancia označuje mieru rozdelenia súčtov gradientov po tom, ako sa dáta v uzle rozdelia podľa konkrétneho bodu rozdelenia.

Z hľadiska výpočtovej náročnosti najzložitejšou úlohou GBDT je učenie rozhodovacích stromov, a to konkrétne hľadanie najlepších bodov rozdelenia (angl. *split points*) pre jednotlivé črty vstupujúce do modelu. Existuje niekoľko algoritmov pomocou ktorých sa to dá dosiahnúť. Jedným z nich je *pre-sorted* algoritmus, ktorý je založený na predtriedených zoznamoch hodnôt jednotlivých čŕt (angl. *attribute lists*). Inými slovami, každý atribút je reprezentovaný zoradeným zoznamom, v ktorom sa nachádzajú hodnoty samotného atribútu, trieda, do ktorej daná hodnota patrí a identifikátor záznamu *(record ID)*, ktorý vyjadruje index pôvodného pozorovania v trénovacej množine. Výhodou danej metódy je to, že pre každú črtu sa udržiava predtriedený zoznam hodnôt, a teda pri ďalších deleniach stromu nepotrebujeme dáta opäť triediť. Pre spojité črty algoritmus postupne prechádza susedné hodnoty v zozname a pre každú dvojicu určí kandidáta na bod rozdelenia. Pre každý takýto bod sa pomocou určenej metriky vyhodnotí kvalita rozdelenia a zvolí sa najlepšie rozdelenie. V prípade kategorických čŕt sa najskôr vytvorí frekvenčná matica výskytov hodnôt a tried, na základe ktorej sa testujú rôzne kombinácie rozdelenia

hodnôt. Tento algoritmus je výpočtovo a pamäťovo náročný, najmä pri veľkých dátových vzorkách [16], [22].

Ďalším nástrojom je histogramový algoritmus (angl. *histogram-based algorithm*), ktorý vytvára rozhodovací strom iteratívne po jednotlivých uzloch. Namiesto hľadania najlepších bodov rozdelenia priamo na usporiadaných hodnotách čŕt, histogramový algoritmus najskôr rozdelí spojité hodnoty čŕt do diskrétnych intervalov tzv. *binov*. Pre každé pozorovanie a každú črtu sa vytvára histogram, ktorý zachytáva informácie o počte pozorovaní a o súčtoch gradientov v jednotlivých intervaloch hodnôt. Na základe takto skonštruovaných histogramov sa následne vyberú kandidátne body rozdelenia. Pomocou zvolenej metriky sa potom určí ten bod, ktorý najlepšie spĺňa jej kritériá. Po určení optimálnych bodov rozdelenia pre každý uzol sa aktualizujú množiny záznamov a vytvoria sa tak potomkové uzly, ktoré predstavujú ďalšie vetvy rozhodovacieho stromu. Tento proces sa opakuje až po dosiahnutie maximálnej hĺbky stromu. Zložitosť výpočtu pre konštrukciu histogramov je $O(n \cdot m)$, pričom ďalších $O(b \cdot m)$ operácií je potrebných na prehľadanie histogramov a vyhodnotenie, ktorý bod rozdelenia je na základe zvolenej metriky najvhodnejší. Pričom $n$ je počet pozorovaní, $m$ je počet čŕt a $b$ je počet intervalov. Z čoho vyplýva, že pri znížení počtu čŕt alebo pozorovaní sa trénovanie modelu vie výrazne zrýchliť [16]. Práve tento problém rieši metóda LightGBM, ktorej sa v nasledujúcej časti podrobnejšie venujeme.

### 2.4.1 LightGBM

Pri trénovaní našich modelov budeme používať nástroj *LightGBM*, ktorý je rozšírením gradient boostingu. Preto v tejto časti priblížime danú metódu a vysvetlíme, na čom je postavená. Okrem štandardnej implementácie algoritmu gradient boostingu LightGBM obsahuje dve kľúčové optimalizačné techniky, ktoré sú implementované s cieľom zvýšiť výpočtovú rýchlosť a znížiť pamäťovú náročnosť. Konkrétne je to nástroj *Gradient-based One-Side Sampling* (GOSS), ktorý sa venuje zníženiu počtu pozorovaní a *Exclusive Feature Bundling* (EFB), ktorý sa zameriava na zníženie počtu čŕt. Následne detailnejšie popíšeme každú z nich. Pričom vychádzame z [16].

#### Gradient-based One-Side Sampling (GOSS)

Metóda Gradient-based One-Side Sampling (GOSS) predstavuje optimalizáciu algoritmu LightGBM, ktorej cieľom je výrazne znížiť počet dátových pozorovaní, a to bez straty

presnosti pri odhade bodov rozdelenia v rozhodovacích stromoch. GOSS vychádza z úvahy, že pozorovania s vyššími hodnotami gradientov významnejšie prispievajú k tomu, podľa ktorej črty a v ktorom bode má rozhodovací strom rozdeliť dáta, aby sa zlepšila presnosť modelu. Vysoké hodnoty gradientu poukazujú na oblasti, kde sa model mýli. Naopak, pozorovania s malými gradientmi sú tie, pri ktorých model už dosahuje nízku chybu, a teda ich ďalší vplyv je menší. GOSS sa preto snaží zameriavať na pozorovania s vysokým gradientom, pričom zároveň nezanedbáva ani tie s nižšími hodnotami. Používame pritom nasledovný postup:

- zoradíme pozorovania zostupne podľa absolútnych hodnôt ich gradientov,
- zachováme všetky pozorovania s vyššími hodnotami gradientu (horných $a \cdot 100\%$ príkladov),
- zo zvyšných pozorovaní ($(1-a) \cdot 100\%$) náhodne vyberieme len časť ($b \cdot 100\%$ zo zvyšku),
- pri výpočte variancie upravíme hodnoty gradientov náhodne vybratých pozorovaní násobením konštantou $\frac{1-a}{b}$.

V danom postupe označujeme ako množinu $A$ tie pozorovania, ktoré patria do horných $a \cdot 100\%$ s vyššími absolútnymi hodnotami gradientov. Zvyšné pozorovania tvoria množinu $A^c$, z ktorej náhodne vyberieme podmnožinu $B$ s veľkosťou $b \cdot |A^c|$. Na základe takto vytvorených množín $A$ a $B$ upravujeme výpočet variancie znázorneného v (48), a to tak, aby zohľadňoval nerovnomerný výber pozorovaní. Matematicky to vieme zapísať nasledovne:

$$\widetilde{V}_j(d) = \frac{1}{n}\left(\frac{\left(\sum\limits_{x_i \in A_l} g_i + \frac{1-a}{b}\sum\limits_{x_i \in B_l} g_i\right)^2}{n_j^l(d)} + \frac{\left(\sum\limits_{x_i \in A_r} g_i + \frac{1-a}{b}\sum\limits_{x_i \in B_r} g_i\right)^2}{n_j^r(d)}\right), \tag{49}$$

kde:

- $A_l = \{x_i \in A : x_{ij} \leq d\}$ a $A_r = \{x_i \in A : x_{ij} > d\}$ predstavujú rozdelenie množiny $A$ podľa hodnoty črty $j$ v bode $d$; $A_l$ obsahuje pozorovania, ktoré smerujú do ľavej vetvy stromu a zároveň $A_r$ do pravej,

- $B_l = \{x_i \in B : x_{ij} \leq d\}$ a $B_r = \{x_i \in B : x_{ij} > d\}$ analogicky predstavujú rozdelenie množiny $A$ podľa hodnoty črty $j$ v bode $d$; $B_l$ obsahuje pozorovania, ktoré smerujú do ľavej vetvy stromu a zároveň $B_r$ do pravej.

Tým pádom pri GOSS namiesto $V_{j|O}(d)$ vypočítaného na celej dátovej sáde použijeme $\widetilde{V}_j(d)$, ktorá je odhadnutá z menšieho počtu dát. Tým sa znižuje výpočtová náročnosť bez podstatného dopadu na presnosť. Aproximačná chyba $E(d) = |\widetilde{V}_j(d) - V_j(d)|$ je ohraničená výrazom, ktorý závisí od počtu pozorovaní v ľavej a pravej vetve po rozdelení. Pričom keď je rozdelenie dát relatívne vyvážené, asymptotická chyba klesá s rastúcim počtom pozorovaní $n$ rýchlosťou $O\left(\frac{1}{\sqrt{n}}\right)$.

**EFB**

Ďalším vylepšením je *Exclusive Feature Bundling* (EFB), ktorý sa zameriava na zníženie počtu čŕt bez výraznej straty informácií. Základná myšlienka EFB spočíva v tom, že veľa čŕt vstupujúcich do modelu je navzájom exkluzívnych. Pod exkluzívnymi črtami budeme rozumieť také črty, ktoré nikdy nenadobúdajú nenulové hodnoty v rámci toho istého pozorovania. Inými slovami, ak je jedna z týchto čŕt nenulová, ostatné v tejto skupine sú v danom riadku nulové. Takéto črty môžeme zlúčiť do jedinej novej črty – tzv. zväzku (angl. *exclusive feature bundle*). Počas trénovania sa následne pre celý zväzok vytvárajú histogramy, čím sa výrazne redukuje výpočtová náročnosť z $O(n \cdot m)$ na $O(n \cdot k)$, kde $k$ je počtom vytvorených zväzkov a zároveň platí: $k \ll m$.

Algoritmus optimálneho rozdelenia čŕt do čonajmenšieho počtu exkluzívnych skupín je NP-ťažký, čo znamená, že neexistuje algoritmus, ktorý by ho dokázal riešiť v polynomiálnom čase. Hovoríme, že algoritmus zbehne v polynomiálnom čase, ak čas jeho výpočtu nepresahuje $n^c$ pre nejakú konštantu $c > 0$. Ďalšie detaily vrátane podrobnejšej definície triedy NP-úloh uvádza [23].

Z uvedeného dôvodu pri EFB sa používa aproximačný algoritmus. Ten modeluje úlohu ako problém farebnosti grafu, kde každý vrchol predstavuje črtu a hrana vyjadruje konflikt medzi črtami. Konflikt medzi črtami nastáva vtedy, keď dve črty nadobúdajú nenulové hodnoty v tom istom pozorovaní. Váhy hrán zodpovedajú počtu nastania takýchto konfliktov. Následne sa črty zoradia zostupne podľa počtu konfliktov (tzv. stupňa vrcholu). Algoritmus potom prechádza jednotlivé črty v tomto poradí a každú z nich sa pokúsi

priradiť do existujúceho zväzku, a to tak, aby počet konfliktov nepresahoval vopred stanovený dovolený počet ich výskytov. Ak ju nie je možné priradiť do žiadneho existujúceho zväzku bez prekročenia povolenej miery konfliktu, vytvorí sa nový zväzok. Celková výpočtová zložitosť tohto algoritmu je $O(m^2)$, kde $m$ je počet čŕt, a vykonáva sa iba raz pred trénovaním modelu.

Druhou úlohou pri zoskupovaní čŕt je zabezpečiť, aby sa pri vytváraní histogramov nestratila informácia o pôvodných hodnotách jednotlivých čŕt. Keďže histogramový algoritmus pracuje s diskrétnymi intervalmi, je potrebné, aby každá črta v rámci jedného zväzku mala svoj vlastný rozsah hodnôt. To sa dosahuje priradením vhodného posunu (tzv. *offset*), ktorý zabezpečí, že hodnoty jednej črty sa nebudú prekryvať s hodnotami iných čŕt v tom istom zväzku.

Spôsob trénovania modelu prostredníctvom LightGBM, navrhnutý spoločnosťou ČSOB, je uvedený v prílohe A.

## 2.5 Zakodóvanie kategorických premenných

V praktických dátach sa často stretávame s črtami, ktoré nenadobúdajú numerické hodnoty, ale sú vyjadrené ako kategórie. Napriek tomu, že veľa metód strojového učenia dokáže pracovať s kategorickými črtami, prevodom týchto čŕt na numerické hodnoty vieme dosiahnuť lepšiu generalizačnú schopnosť modelu a jeho interpretovateľnosť. Preto také premenné chceme zakódovať do numerickej formy. V nasledujúcej časti preto predstavíme nástroje na zakódovanie.

### Vytváranie dummy premenných

Pod dummy premennými rozumieme binárnu premennú, ktorá vzniká prevodom kategorickej premennej do numerickej formy. Pre kategóriu s $G$ unikatnými hodnotami sa vytvára $G - 1$ dummy premenných. Každá z týchto premenných nadobúda hodnotu 1, ak pozorovanie patrí do príslušnej kategórie a 0 inak. Jedna z kategórií slúži ako referenčná kategória, ktorá nie je zastúpená samostatnou premennou, nakoľko príslušnosť k nej vieme odvodiť pomocou hodnôt zvyšných kategórií. Použitie $G - 1$ dummy premenných tak zabezpečuje lineárnu nezávislosť vstupných čŕt a zároveň znižuje pamäťovú náročnosť pri uchovávaní a spracovaní dát [9].

Vytváranie dummy premenných predstavuje základnú techniku zakódovania kategorických premenných, pri ktorej sa pre každú kategóriu vytvorí samostatná binárna črta. Pri danom prístupe sa výrazne navyšuje dimenzia vstupného priestoru, čo zvyšuje výpočtovú náročnosť. Preto daný prístup je užitočný iba pri črtách s malým počtom unikátnych hodnôt. Výsledná matica čŕt má riedku štruktúru, kde každý stĺpec reprezentuje jednu kategóriu a každý riadok zodpovedá jednému pozorovaniu [26].

**Zakódovanie pomocou CatBoost**

Týmto prístupom pri zakódovaní kategorických premenných každá katégoria sa nahradí priemerom závislej premennej pre danú kategóriu. Aby sme predišli úniku informácie o závislej premennej (angl. *target leakage*), kategórie zakódujeme na základe hodnôt závislej premennej vypočítaných z predchádzajúcich pozorovaní zoradených v náhodnom poradí.

Je daná trénovacia vzorka dát $\mathcal{D}$ a jej náhodná permutácia $\sigma$. Označme $\sigma(k)$ ako $k$-tý prvok množiny $\mathcal{D}$ podľa permutácie $\sigma$. Nech $\mathcal{D}k = x_1, x_2, \ldots, xk - 1$ predstavuje množinu všetkých predchádzajúcich pozorovaní vzhľadom na $x_k$. Funkcia $\mathbf{1}_{x_j^i = x_k^i}$ nadobúda hodnotu 1 v prípade, že hodnota $i$-tej kategórie pozorovania $x_j$ je rovná hodnote v $x_k$, a hodnotu 0 inak. Táto funkcia sa využíva na identifikáciu predchádzajúcich pozorovaní, ktoré mali rovnakú kategóriu ako aktuálne pozorovanie.

Zakódovanú hodnotu $i$-tej kategórie pre $k$-té pozorovanie potom definujeme ako [8]:

$$\hat{x}_k^i = \frac{\sum\limits_{x_j \in \mathcal{D}_k} \mathbf{1}_{x_j^i = x_k^i} \cdot y_j + a \cdot p}{\sum\limits_{x_j \in \mathcal{D}_k} \mathbf{1}_{x_j^i = x_k^i} + a}, \tag{50}$$

kde:

- $\mathcal{D}_k$ je množina predchádzajúcich pozorovaní použitých pri výpočte zakódovanej hodnoty pre $k$-té pozorovanie,
- $y_j$ je hodnota závislej premennej pre pozorovanie $x_j$,
- $p$ je tzv. prior, bežne nastavený ako priemer závislej premennej v trénovacej vzorke dát,
- $a$ je regularizačný parameter, ktorý zabezpečuje numerickú stabilitu výpočtu. Pričom pri implementácii hodnota daného parametra je súčasťou algoritmu a aktualizuje sa v procese trénovania.

Týmto spôsobom sa zabezpečí, že každé pozorovanie $x_k$ je zakódované na základe predchádzajúcich pozorovaní zoradených v náhodnom poradí a bez využitia jeho vlastnej výstupnej hodnoty.

## 2.6 Metriky výkonnosti základných modelov

Pri hodnotení modelov je dôležité určiť, do akej miery dokáže model správne predikovať očakávané výstupy na základe vstupných údajov. Ako bolo uvedené v Kap. 1.2, modelovanie upliftu je založené na kombinácii základného učiteľa a meta-učiteľa. Zatiaľ čo základný učiteľ trénuje model, meta-učiteľ spracováva a vyhodnocuje jeho výstupy. Z tohto dôvodu sa pri analýze výkonnosti zameriavame na oba komponenty a využívame rôzne metriky na ich vyhodnotenie.

### 2.6.1 Metriky výkonnosti základných učiteľov

V tejto časti uvedieme základné metriky na vyhodnocovaie výkonnosti modelov, pričom vychádzať budeme z [13].

Na úrovni základných učiteľov používame klasifikačné modely, ktoré predikujú pravdepodobnosť nastania modelovanej udalosti. Pri vyhodnocovaní výkonnosti klasifikačného modelu sa pozeráme na jeho schopnosť zaraďovať pozorovania do ich skutočných tried. Preto vychádzame z porovnania medzi skutočnými triedami a predikciami modelu. Pri binárnej klasifikácii môžu nastať štyri situácie v závislosti od kombinácie skutočnej triedy a triedy predikovanej modelom. Nasledujúca tabuľka uvádza tieto možnosti spolu s označeniami, ktoré sa bežne používajú pri hodnotení výkonnosti klasifikačných modelov a ktoré budeme následne využívať pri definovaní ďalších metrík.

| Predikcia / Skutočná trieda | Pozitívna | Negatívna | Spolu |
|---|---|---|---|
| Pozitívna | Skutočne pozitívne (TP) | Falošne negatívne (FN) | $P$ |
| Negatívna | Falošne pozitívne (FP) | Skutočne negatívne (TN) | $N$ |
| Spolu | $P^*$ | $N^*$ | $P + N$ |

**Tabuľka 1:** Možné situácie pri binárnej klasifikácii

Tab.1 ilustruje štyri možné scenáre pri binárnej klasifikácii do tried pozitívna a negatívna. V jej bunkách sú uvedené pomenovania týchto situácií spolu so skratkami, ktoré vysvetľujeme nižšie:

- **Skutočne pozitívne** (angl. *True Positive*) – model správne predikoval pozitívnu triedu, objekt je pozitívny a bol aj tak klasifikovaný.
- **Falošne negatívne** (angl. *False Negative*) – model nesprávne predikoval negatívnu triedu, objekt bol klasifikovaný ako negatívny, ale v skutočnosti je pozitívny.
- **Falošne pozitívne** (angl. *False Positive*) – model nesprávne predikoval pozitívnu triedu, objekt bol klasifikovaný ako negatívny, ale v skutočnosti je pozitívny.
- **Skutočne negatívne** (angl. *True Negative*) – model správne predikoval negatívnu triedu, objekt je negatívny a bol aj tak klasifikovaný.

Riadkové a stĺpcové súčty vyjadrujú celkové počty skutočne pozitívnych a negatívnych prípadov $(P, N)$ a taktiež celkové počty predikovaných pozitívnych a negatívnych prípadov $(P^*, N^*)$. Súčet všetkých hodnôt predstavuje celkový počet klasifikovaných prípadov $(P + N)$.

Na základe tejto tabuľky je možné definovať viacero metrík na vyhodnotenie výkonnosti modelu.

**Celková presnosť (angl. *accuracy*)**

Predstavuje podiel všetkých správne klasifikovaných prípadov a celkový počtet prípadov:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}. \tag{51}$$

**Presnosť pozitívnej predikcie (angl. *Precision*)**

Presnosť pozitívnej predikcie vyjadruje, aký podiel z pozorovaní, ktoré model označil ako pozitívne, bol skutočne pozitívny. Formálne to vieme zapísať nasledovne:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{52}$$

**Senzitivita (angl. *True Positive Rate* – *TPR*)**

Vyjadruje schopnosť modelu správne identifikovať pozitívne prípady:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \tag{53}$$

**Špecificita (angl. *True Negative Rate* – *TNR*)** Vyjadruje schopnosť modelu správne rozpoznať negatívne prípady:

$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}. \tag{54}$$

**Miera falošne pozitívnych prípadov (angl. *False Positive Rate* – FPR)**

Vyjadruje pravdepodobnosť, že klasifikátor nesprávne priradí objekt negatívnej triedy do triedy pozitívnej. Môžeme ju definovať nasledovne:

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}. \tag{55}$$

Táto miera je známa aj ako chyba prvého druhu a taktiež ako doplnok špecificity, teda platí nasledovné:

$$\mathrm{FPR} = 1 - \mathrm{TNR} = 1 - \frac{\mathrm{TN}}{\mathrm{FP} + \mathrm{TN}}. \tag{56}$$

**Matica zámien**

Ďalším nástrojom na vyhodnocovanie výkonnosti klasifikačných modelov je matica zámien (angl. *confusion matrix*). Daná matica obsahuje počty pozorovaní pre jednotlivé kombinácie skutočných a modelom predikovaných tried. Pre binárny prípad, by sme maticu zámien pomocou notácie zavedenej v Tab. 1 by sme mohli zapísať nasledovne:

| Predikcia / Skutočná trieda | Pozitívna | Negatívna |
|---|---|---|
| Pozitívna | TP | FN |
| Negatívna | FP | TN |

**Tabuľka 2:** Matica zámien pre binárnu klasifikáciu

Prvky na hlavnej diagonále matice zámien predstavujú počet prípadov, pri ktorých model správne priradil pozorovaniu jeho triedu. Naopak, prvky mimo diagonály predstavujú počet nesprávne klasifikovaných prípadov. V ideálnom prípade by sme očakávali vysoké hodnoty na diagonále a čo najnižšie hodnoty mimo nej.

**ROC krivka a metrika ROC-AUC**

ROC krivka (angl. *Receiver Operating Characteristic curve*) predstavuje graf závislosti medzi mierou falošne pozitívnych prípadov (*FPR*) predstavenú na (55) a senzitivitou (*TPR*) predstavenú na (53) pri rôznych prahových hodnotách. Teda na zvislej osi sa nachádza hodnota *TPR* a na vodorovnej osi hodnota *FPR*. ROC krivka zobrazuje správanie klasifikátora v závislosti od rozsahu prahových hodnôt. Výkonnosť klasifikátora vieme vyjadriť prostredníctvom plochy pod ROC krivkou, pričom táto metrika sa označuje **ROC-AUC** (angl. *Area Under the Receiver Operating Characteristic curve*). Ideálny klasifikátor má ROC krivku, ktorá prilieha k ľavému hornému rohu grafu, čo zodpovedá vysokej senzitivite. V takom prípade sa hodnota ROC-AUC blíži k 1. Metrika ROC-AUC predstavuje pravdepodobnosť, že model priradí náhodne vybranému negatívnemu prípadu nižšiu predikovanú pravdepodobnosť, že patrí do pozitívnej triedy ako náhodne vybranému pozitívnemu prípadu [10], [12].

Na základe ROC-AUC vieme nielenže vyhodnotiť predikčné schopnosti modelov, ale aj ich porovnať medzi sebou – vyššia hodnota indikuje lepší výkon.

Takto vytvorená krivka umožňuje vizuálne porovnať výkonnosť klasifikátora pri rôznych rozhodovacích prahoch.

Nakoľko pracujeme so skupinami, ktorých veľkosti sa môžu výrazne líšiť, tak to chceme zohľadniť aj pri vyhodnocovaní výkonnosti modelov. Nech máme $m$ skupín, pre každú z ktorých sú vypočítané hodnoty ROC-AUC. Označme $n_1, ..., n_m$ ako počet pozorovaní v jednotlivých skupinách. Potom vieme definovať vážené ROC-AUC (angl. weighted ROC-AUC), pomocou ktorého vieme získať celkovú výkonnosť modelu z ohľadom aj na veľkosť jednotlivých skupín [15]. Formálne to vieme zapísať nasledovne:

$$\text{Vážené ROC-AUC} = \frac{\sum_{i=1}^{m} n_i \cdot \text{ROC-AUC}_i}{\sum_{i=1}^{m} n_i}, \tag{57}$$

kde:

- $n_i$ - počet pozorovaní v $i$-tej skupine,

- ROC-AUC$_i$ - hodnoty ROC-AUC pre $i$-tú skupinu.

### PR krivka a metrika PR-AUC

V prípadoch, keď sú dáta výrazne nevyvážené a negatívna vzorka obsahuje výrazne viac pozorovaní je užitočné sa zameriavať aj na výkonnosť modelu pri odhade pozitívnych prípadov. Na dané účely sa zavádza PR krivka (angl. *Precision–Recall curve).*

Táto krivka porovnáva *presnosť*, ktorá je definovaná v (52) a *senzitivitu*, ktorá sa zvykne označovať aj ako *recall*, definovanej v (53). Pri vizualizácii PR krivky na vodorovnej osi sa zobrazí senzitivita, zatiaľ čo na zvislej osi sa zobrazí presnosť. Pričom čím bližšie sa PR krivka nachádza k pravému hornému rohu grafu, tým vyššiu výkonnosť model dosahuje. Na kvantifikáciu výkonnosti modelu sa používa plocha pod PR krivkou. Táto metrika sa označuje **PR-AUC** (angl. *Area Under the Precision–Recall curve*) [2].

# 3 Formulácia problému oslovenia z pohľadu banky

V tejto kapitole priblížime problém marketingového oslovenia zákazníkov z pohľadu banky, ktorá sa snaží čo najefektívnejšie zacieliť svoje kampane. Konkrétne sa venujeme formulácii daného problému pomocou optimalizačnej úlohy.

## 3.1 Úvod do problematiky

Pre banky je dôležité nielen propagovať svoje produkty a služby, ale si aj budovať vzťahy so zákazníkmi. Dôležité je zameriavať sa na tie skupiny zákazníkov, ktoré majú väčší potenciál na pozitívnu reakciu v dôsledku marketingového oslovenia. Tým sa dajú minimalizovať náklady tak, že banka nebude vynakladať prostriedky na zákazníkov, ktorí by svoje správanie nezmenili resp. reagovali by na oslovenie negatívne.

Jedným z marketingových nástrojov je oslovenie zákazníkov bankou. Pod oslovením rozumieme cielené sprostredkovanie relevantných ponúk zákazníkom. Jeho cieľom je motivovať ich k určitej udalosti, ako napríklad k nákupu produktu alebo k využitiu nejakej služby. Pričom stanovenie cieľovej skupiny si vyžaduje personalizovaný prístup, ktorý by zohladňoval individuálne charakteristiky a správanie zákazníkov. Oslovenie sa uskutočňuje cez určitý kanál. Pod kanálom rozumieme spôsob, akým firma uskutočňuje oslovenie. Existujú rôzne druhy kanálov, napríklad oslovenie môže byť uskutočnené cez SMS, notifikáciu, telefonát, sociálne siete, v pobočke banky a t.ď.

Na oslovenie zákazníkov banka vynakladá finančné prostriedky. Tieto prostriedky môžu byť investované do rôznych oblastí, akými sú napríklad náklady na pracovníkov, ktorí sa určitým spôsobom podieľajú na vytváraní obsahu kampane alebo napríklad uskutočňujú telefonáty. Taktiež sú náklady na technológie, napríklad na nejaké komunikačné platformy, softvéry. Preto z hľadiska optimalizácie nákladov je pri výbere zákazníkov na oslovenie dôležité efektívne využiť dostupné dáta a informácie.

## 3.2 Formulácia optimalizačnej úlohy pre banky

Cieľom banky je osloviť iba tých zákazníkov, pri ktorých sa očakáva najvyšší prínos. Preto odhad individuálného efektu je podkladom pri tvorbe marketingových stratégií. Pri rozhodovaní o tom, ktorých zákazníkov osloviť v marketingovej kampani, chce banka získať

čonajvyšší prínos, ale zároveň minimalizovať celkové straty spôsobené neefektívnym oslovením. Daný problém vieme sformulovať pomocou úlohy vo forme minimalizácie stratovej funkcie (angl. *loss function*).

Ako bolo uvedené v Kap.1 v (6) pravdepodobnosť, že jednotka s črtami $X = x$ bude zaradený do ošetrenej skupiny označujeme $e(x)$. Taktiež v (5) sme zaviedli CATE, ktoré označujeme $\tau(x)$. Pri formulácii tejto úlohy budeme pre premenné používať horný index $^*$, ktorým označujeme neznáme charakteristiky populácie, napríklad skutočné hodnoty očakávaných výsledkov.

Pre formuláciu úlohy potrebujeme zadefinovať podmienenú očakávanú hodnotu potenciálnych výsledkov (angl. *response surfaces*), ktorá reprezentuje očakávané správanie zákazníka so známymi črtami $X = x$ v závislosti od toho, či bol oslovený. Formálne to vieme zapísať nasledovne:

$$\mu^*_{(z)}(x) = \mathbb{E}[Y(z) \mid X = x], \quad \text{pre } z \in \{0, 1\}. \tag{58}$$

Rezíduum predstavuje rozdiel medzi skutočným (pozorovaným alebo potenciálnym) výsledkom a očakávanou hodnotou podľa modelu. Rezíduum pre jednotku $i$ a stav ošetrenia $z \in \{0, 1\}$ definujeme nasledovne:

$$\varepsilon_i(z) := Y_i(w) - \left\{\mu^*_{(0)}(X_i) + z \cdot \tau^*(X_i)\right\}, \tag{59}$$

Podmienenú očakávanú hodnotu výstupu definujeme nasledovne:

$$m^*(x) := \mathbb{E}[Y \mid X = x] = \mu^*_{(0)}(x) + e^*(x) \cdot \tau^*(x), \tag{60}$$

Potom rovnicu (5) vieme reformulovať pomocou vyššie zavedených pojmov:

$$Y_i - m^*(X_i) = \{Z_i - e^*(X_i)\} \cdot \tau^*(X_i) + \varepsilon_i(Z_i). \tag{61}$$

Motivácia danej reformulácie spočíva v tom, že uvedená reprezentácia umožňuje skonštruovať stratovú funkciu, ktorá zachytáva heterogénne efekty ošetrenia. Následne môžeme efekt ošetrenia odhadnúť hľadaním minima stratovej funkcie. Formálne to vieme definovať nasledovne [20]:

$$\tau^*(\cdot) = \arg\min_{\tau} \mathbb{E}\left[(Y - m^*(X)) - (Z - e^*(X))\,\tau(X))^2\right], \tag{62}$$

kde:

- $m^*(X)$ je podmienená stredná hodnota výstupu,
- $e^*(X)$ je pravdepodobnosť zaradenia do ošetrenej skupiny,
- $\tau(X)$ je podmienený priemerný efekt ošetrenia (CATE).

Rovnica (62) uvádza základný prístup, za predpokladu, že náklady na jednotlivé oslovenia sú rovnaké. V praxi rôzne oslovenia majú rôzne náklady, napríklad oslovenia cez SMS a telefonát má rôzne náklady. Z tohto dôvodu je vhodné optimalizačnú úlohu rozšíriť tak, aby zohľadňovala aj nákladovosť rôznych kanálov. Nech máme nasledovnú množinu možných ošetrení $\mathcal{Z} = \{z_0, z_1, \ldots, z_m\}$. V tomto prípade už neuvažujeme iba o tom, či bol zákazník oslovený, ale o $m$ rôznych ošetreniach, z ktorých je potrebné vybrať to, ktoré má pre konkrétnu jednotku najvyšší efekt.

Vo všeobecnosti rozlišujeme dva typy nákladov:

- Fixný náklad na oslovenie (označ. $c$) – keď za každé oslovenie je účtovaný určitý poplatok bez ohľadu na to, či zákazník následne reaguje. Napríklad kampane, ktoré sú realizované prostredníctvom SMS, e-mailov alebo notifikácie v aplikácii.
- Stimulačný náklad viazaný na konverziu (označ. $s$) – tento typ nákladu vzniká v prípade, že používateľ vykoná požadovanú akciu. Napríklad ak používateľ navštívil konkrétnu produktovú stránku a neskôr si ju opätovne vyhľadá, môže mu byť ponúknutá zľavnená cena ako motivácia k nákupu.

Preto definujeme tzv. čistý prínos (angl. *net value*), ktorý vyjadruje rozdiel medzi hodnotou konverzie a nákladom na oslovenia. Vtedy namiesto klasického CATE pracujeme s čistým prínosom podmieneného priemerného efektu ošetrenia (angl. *net uplift CATE*). V danom prípade kľúčovou úlohou je maximalizácia čistého prínosu. Vtedy berieme do úvahy nielen výnos z konverzie, ale aj náklady na oslovenie. Formálne zapísané:

$$\tau^{z_j}(x) = \mathbb{E}\left[(v - s_{z_j})Y_{z_j} - (v - s_{z_0})Y_{z_0} - (c_{z_j} - c_{z_0}) \mid X = x\right], \tag{63}$$

V našom prípade spoločnosť neponúka žiadne stimuly a preto o týchto nákladoch neuvažujeme, čiže platí $s = 0$. Na základe tejto úvahy vieme (63) zjednodušiť:

$$\tau^{z_j}(x) = \mathbb{E}\left[v(Y_{z_j} - Y_{z_0}) - (c_{z_j} - c_{z_0}) \mid X = x\right], \tag{64}$$

kde:

- $v$ je očakávaná hodnota konverzie (napr. zisk z predaja). Predpokladajme, že každá konverzia má rovnakú hodnotu, pričom táto hodnota je vopred známa,
- $c_{z_j}$ je náklad na oslovenie ošetrením $z_j$, kde $j \in \{0, 1, 2, .., m\}$,
- $z_0$ označuje referenčnú skupinu, v našom prípade ide o skupinu, ktorá nedostala oslovenie.

Tento prístup je užitočný pri modelovaní viacerých ošetrených skupín s rôznymi vynaloženými nákladmi na ich oslovenie, čím umožňuje optimalizovať alokáciu zdrojov a maximalizovať zisk kampane.

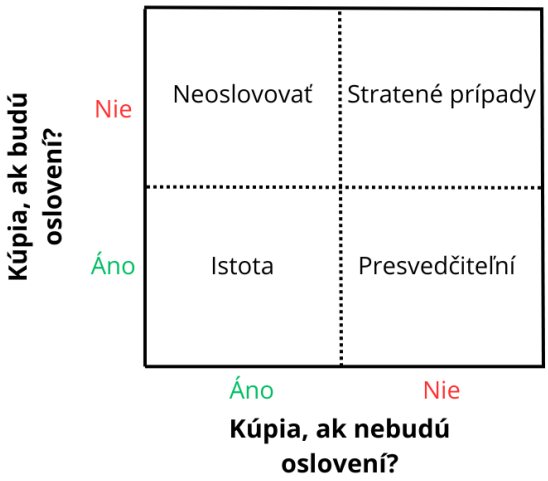
V prípade, že máme k dispozícii viacero možných ošetrení $\mathcal{Z} = \{z_0, z_1, \ldots, z_m\}$ a stimulačné náklady $s_{z_j}$ sú nulové. Optimalizačný problém teda vieme zjednodušiť nasledovne:

$$z^*(x) = \arg\max_{z_j \in \mathcal{Z}} \mathbb{E}\left[v(Y_{z_j} - Y_{z_0}) - (c_{z_j} - c_{z_0}) \mid X = x\right], \tag{65}$$

kde: $z^*(x)$ je optimálne ošetrenie pre zákazníka s črtami $x$. Pre každého zákazníka chceme vybrať taký typ ošetrenia z množiny $\mathcal{Z}$, ktorý maximalizuje očakávaný čistý prínos [29].

## 3.3 Výber klientov z pohľadu banky

Bežne používané propenzitné modely (angl. *propensity models*) sa zameriavajú na predpovedanie toho, aká je pravdepodobnosť, že zákazník vykoná modelovanú udalosť, teda, či si zákazník kúpi určitý produkt alebo využije určitú službu. Dôležité je nielen vedieť s akou pravdepodobnosťou zákazník vykoná modelovanú udalosť, ale aj pochopiť príčinné vzťahy, teda to, ako konkrétne oslovenie ovplyvní jeho pravdepodobnosť. Tento prístup nám umožňuje presnejšie odhadnúť, ktorí zákazníci majú vyšší efekt oslovenia. V rámci modelovania sa rozlišuje medzi rôznymi skupinami zákazníkov. Konkrétne na základe ich potenciálnej reakcie na oslovenie ich vieme rozdeliť do štyroch skupín, ako je znázornené nižšie [14]:




**Obr. 1:** Štyri skupiny zákazníkov v závislosti ich reakcie na oslovenie

Na Obr. 1 sú znázornené štyri základné typy zákazníkov v závislosti od ich reakcie na oslovenie. Tento prehľad pomáha lepšie porozumieť problematike a zároveň zdôrazňuje, prečo je pri modelovaní dôležité zohľadňovať aj príčinné vzťahy. Popíšeme skupiny detailnejšie:

- Presvedčiteľní (*angl. Persuadables*) - zákazníci, ktorí si skôr kúpia produkt alebo využijú službu v prípade, že budú oslovení(resp. dôjde ku skôršiemu nákupu alebo vo väčšom množstve). Inak povedané, po oslovení sú náchylní k nákupu.

- Istota (*angl. Sure Things*) – zákazníci, ktorí by uskutočnili modelovaciu udalosť, aj keby neboli oslovení. Sú to teda zákazníci, ktorí už mali záujem o produkt alebo službu aj bez oslovenia.

- Stratené prípady (*angl. Lost Causes*) – zákazníci, ktorí nie sú náchylní k nákupu ani v prípade oslovenia. Teda tá skupina zákazníkov, ktorá nereaguje na marketingové kampane.

- Neoslovovať (*angl. Do Not Disturbs*) – zákazníci, ktorí by produkt zakúpili iba v prípade, že by neboli oslovení. Akýkoľvek marketingový zásah má na nich negatívny vplyv, pretože ich od nákupu odrádza a naviac môže zhoršiť ich vzťah s firmou.

V tejto práci sa nebudeme zaoberať klasifikáciou zákazníkov do diskrétnych kategórií podľa typu odpovede na oslovenie. Namiesto toho sa sústredíme na kvantifikáciu presvedčiteľnosti každého zákazníka na spojitej škále. Konkrétne budeme využívať odhad individuálneho efektu ošetrenia (CATE), ktoré odhadneme nástrojmi uvedenými v Kap. 1. CATE slúži ako spojité skóre určujúce efekt oslovenia.

Konkrétne sa sústredíme na rozhodnutie založiť si spotrebný úver v dôsledku oslovenia. Cieľom je určiť, aký vplyv má na zákazníka oslovenie a či v dôsledku neho uskutoční modelovanú udalosť. Pod modelovanou udalosťou teda rozumieme založenie spotrebného úveru zákazníkom.

# 4 Príprava dát pre trénovanie

Nakoľko v tejto práci spolupracujeme s Československou obchodnou bankou (ČSOB), pre trénovanie modelu máme k dispozícii ich interné dáta. Naším cieľom je pomocou týchto údajov vytvoriť dátovú vzorku, ktorá sa následne použije na trénovanie modelu.

## 4.1 Vysvetľujúce črty

Dáta, ktoré máme obsahujú rôzne informácie o zákazníkoch. Konkrétne máme k dispozícii 15 tabuliek, ktoré spolu obsahujú 2152 rôznych čŕt pre každého zákazníka.

| **Tabuľka** | **Počet čŕt** | **Popis, k čomu sa vzťahujú** |
|---|---|---|
| ATTR_CLIENT_BASE | 22 | Demografické údaje |
| ATTR_CARD | 151 | Platobná karta |
| ATTR_TERMINAL | 73 | Terminál |
| ATTR_TXN | 287 | Transakcie |
| ATTR_BANK_ACCOUNT | 192 | Informácie k účtu |
| ATTR_ELB | 254 | Elektronické bankovníctvo |
| ATTR_FUND | 51 | Fondy |
| ATTR_INSURANCE | 141 | Poistenie |
| ATTR_ECONOMY_SITUATION | 95 | Ekonomická situácia |
| ATTR_TXN_CAT | 323 | Kategorizované transakcie |
| ATTR_ADDR | 14 | Adresa |
| ATTR_AFP | 128 | Analýza finančných potrieb |
| ATTR_INTERACTION | 263 | Interakcie banky so zákazníkom |
| ATTR_KATE | 102 | Používanie virtuálnej asistentky Kate |
| ATTR_WEB_TRACKING | 86 | Správanie klienta na webe |
| **Celkovo** | **2152** | |

**Tabuľka 3:** Kategorizácia atribútov klientskych dát

Dané tabuľky sa počítajú na mesačnej báze pre všetkých v tom mesiaci aktuálnych zákazníkoch. Pri výbere tréningových pozorovaní pracujeme s datámi za obdobie troch rokov (36 mesiacov). Každý zákazník má v rámci banky pridelený technický identifikátor

*customerid* a pre každý mesiac, v ktorom bol zákazníkom banky má ku svojmu identifikátoru priradené predpočítané črty. Črty, ktoré sa vzťahujú k určitému mesiacu sú označené prostredníctvom stĺpca *month_id*. Tento stĺpec obsahuje identifikátor vo formáte `YYYYMM`, kde prvé dve číslice označujú mesiac a nasledujúce štyri číslice rok. Napríklad hodnota 202412 predstavuje december 2024.

### 4.1.1 Črty z webovej aktivity

Taktiež v rámci práce bola vytvorená ďalšia tabuľka čŕt na základe aktivity zákazníkov banky na webovej stránke. K dispozícii sme mali údaje o interakciách návštevníkov s webovým rozhraním, pričom každý záznam obsahoval technický identifikátor *cookie_id*, ktorý umožňoval sledovanie správania jednotlivých používateľov pred ich prihlásením ako aj používateľov, ktorí nie sú zákazníkmi banky. Medzi zaznamenané údaje, s ktorými sme najviac pracovali patrili čas návštevy, odkaz, z ktorého bol používateľ presmerovaný a taktiež čas záznamu kliknutia na konkrétny odkaz. Pričom po prihlásení používateľa do internet bankingu sa *cookie_id* spätne prepojilo s identifikátorom zákazníka (*customerid*), čo umožňuje priradiť zaznámenanú aktivitu ku konkrétnemu zákazníkovi a následne ju využiť pri tvorbe čŕt pre modelovanie. Pri tvorbe čŕt sme použili štruktúru vychádzajúcu z ukazovateľov RFM [27]:

- **Aktuálnosť** (angl. *recency*) – predstavuje črty, ktoré vyjadrujú, ako dávno zákazník vykonal určitú aktivitu. Ide napríklad o počet dní od poslednej návštevy alebo interakcie.
- **Frekvencia** (angl. *frequency*) – predstavuje črty, ktoré popisujú, ako často zákazník vykonával aktivity počas definovaného časového obdobia, napríklad počet návštev alebo kliknutí za posledný mesiac.
- **Hodnota aktivity** (angl. *monetary value*) – predstavuje črty súvisiace s výškou finančného prínosu. Taktiež v predstavuje črty, ktoré približujú mieru angažovanosti používateľa, napríklad priemerný čas strávený na stránke, alebo počet interakcií počas jednej návštevy.

Na základe extrahovaných informácií z URL adries, časových údajvov a zdrojov presmerovania sme vytvorili črty, popisujúce rôzne vzorce správania zákazníkov, ich pre-

ferencie, navštívené podstránky. Taktiež sme mohli zachytiť aj zmeny v ich správaní v priebehu posledného mesiaca, troch mesiacov alebo roka. Tabuľka teda obsahuje rôzne metriky, ktoré sú ratané na mesačnej báze, pričom ich stručný popís je uvedený v Prílohe B. Tým pádom do modelovania budú zahrnuté aj črty popisujúce správanie zákazníka na webovej stránke. To poskytuje banke hlbší pohľad na správanie zákazníkov a umožňuje presnejšie modelovanie ich potrieb.

## 4.2 Závislá premenná

Závislá premenná v tomto prípade indikuje, či zákazník vykonal tzv. modelovanú udalosť. V kontexte danej práce ide o to, či si zákazník založil spotrebný úver.

Daná premenná nadobúda binárne hodnoty, pričom hodnotu 1 priradujeme v prípade, že zákazník realizoval modelovanú udalosť, zatiaľ čo hodnotu 0 priradujeme, ak zákazník udalosť nerealizoval.

Na základe hodnoty závislej premennej rozlišujeme dve skupiny zákazníkov:

- **Pozitívna vzorka:** obsahuje zákazníkov, ktorí vykonali modelovanú udalosť (t. j. $Y = 1$);
- **Negatívna vzorka:** obsahuje zákazníkov, ktorí modelovanú udalosť nevykonali (t. j. $Y = 0$).

Táto premenná slúži ako základ pre klasifikačné modely, pomocou ktorých sa predikuje pravdepodobnosť, že daný zákazník vykoná modelovanú udalosť na základe dostupných vstupných čŕt. V rámci našej práce nie je cieľom predikovať pravdepodobnosť nastania modelovanej udalosti, ale či na jej nastanie malo vplyv oslovenie. Preto sa závislá premenná posudzuje spoločne s príčinnou premennou, ktorú definujeme v nasledujúcej podkapitole.

## 4.3 Výber tréningových pozorovaní

Výber tréningových dát spočíva v zostavení vzorky, ktorá by reprezentovala správanie zákazníkov v rôznych časových obdobiach. Pri tvorbe výberovej vzorky je cieľom každému zákazníkovi, ktorý bol do nej vybraný priradiť zoznam už predpočítanych čŕt, ktoré predstavujú jeho charakteristiky agregované do konkrétneho dňa. Mesiac, za ktorý sa pre zákazníka priradia črty je určený podľa dátumu nastania modelovanej udalosti a istých

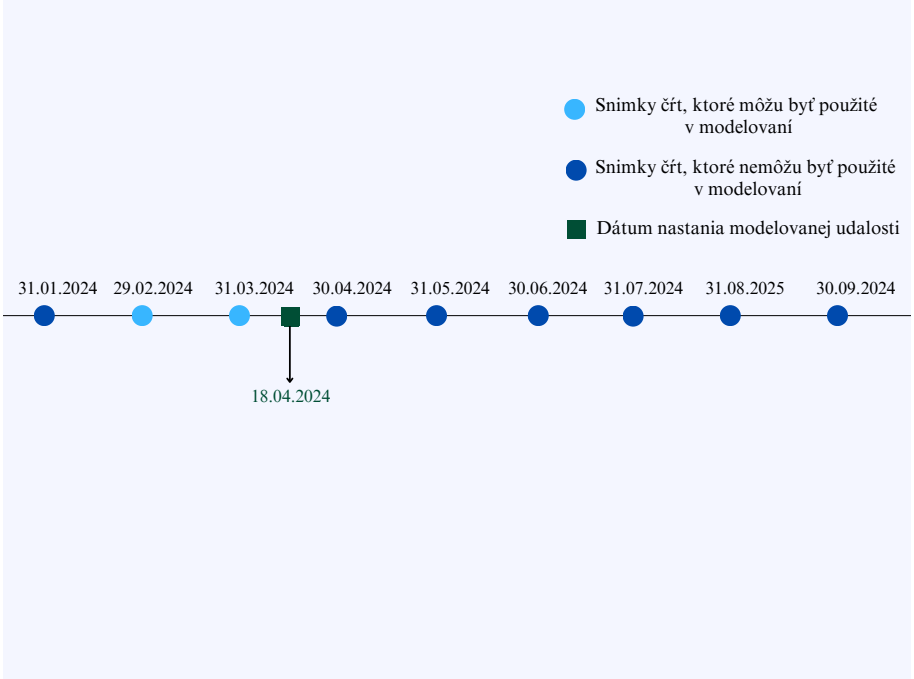
definovaných pravidiel, ktoré uvádzame v nasledujúcich častiach. V prvom kroku výberu tréningovej vzorky budú vylučení zákazníci, ktorí sú označení ako nevhodní na zahrnutie do modelu, ako aj zákazníci s nevyriešeným stavom nákupu, aby sa predišlo skresleniu. Zároveň platí, že každý zákazník môže byť zaradený do vzorky najviac raz. Taktiež výber zákazníkov do vzorky sa môže líšiť v závislosti od modelu, keďže rôzne modely môžu pracovať s odlišnými segmentmi zákazníkov.

**Konštrukcia pozitívnej vzorky**

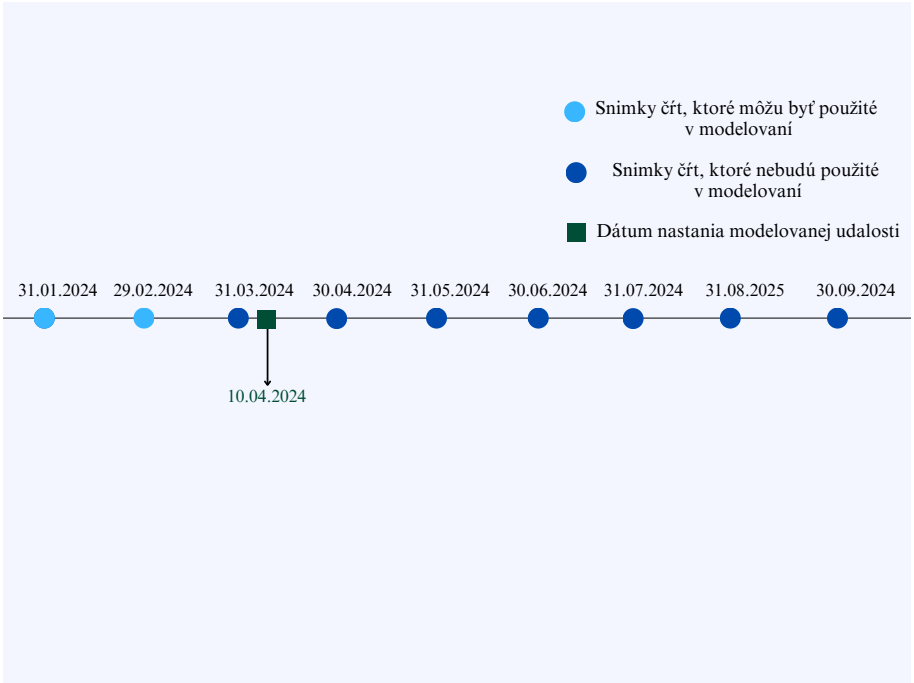
Pozitívna vzorka predstavuje skupinu zákazníkov, ktorí vykonali modelovanú udalosť, teda založili si spotrebný úver. Každému zákazníkovi v tejto vzorke pritom zodpovedajú vlastné črty, ktoré definujú jeho správanie a charakteristiky.

Množina pozitívnych jednotiek sa zostavuje z tých *customerids*, ktoré v danom časovom období mali aspoň jedno nastanie udalosti. Pričom každý z týchto zákazníkov je do pozitívnej vzorky zaradený najviac raz, a to aj v prípade, že viackrát vykonal modelovanú udalosť. Vtedy sa náhodne vyberie len jedna z týchto udalostí. Pri výbere čŕt pozitívnej vzorky sa zohľadňujú dva celé mesiace pred výskytom udalosti, pričom sa berie do úvahy aj 15-dňový posun. Posun o 15 dni slúži na zabezpečenie časového odstupu medzi dátami a modelovanou udalosťou, aby sa predišlo úniku informácií z obdobia tesne pred jej nastaním. Ak zákazník uskutoční udalosť po 15. dni v mesiaci, priradia sa mu črty z jedného z dvoch predchádzajúcich celých mesiacov. Napríklad, ak zákazník vykoná udalosť 18. apríla 2024, črty môžu byť priradené buď z marca 2024 alebo februára 2024.

Naopak, ak zákazník uskutoční udalosť pred 15. dňom v mesiaci, priradia sa mu črty z dvoch celých mesiacov pred aktuálnym mesiacom. Napríklad, ak zákazník vykoná udalosť 10. apríla 2024, jeho črty môžu pochádzať z februára 2024 alebo januára 2024. Pre lepšie pochopenie uvádzame nasledujúce schémy:




**Obr. 2:** Schéma priraďovania čŕt v procese tvorby vzorky (nákup po 15. dni mesiaca)




**Obr. 3:** Schéma priraďovania čŕt v procese tvorby vzorky (nákup pred 15. dňom mesiaca)

Na Obr.2 je znázornená schéma priraďovania čŕt na základe dátumu uskutočnenia modelovanej udalosti, keď sa uskutočnila po 15. dni mesiaca. Inými slovami, za akého časové obdobie sa vyberajú predpočítané črty, ktoré sa následne priraďujú k danému

zákazníkovi na základe dátumu realizácie udalosti.

Analogicky Obr.3 ilustruje schému priraďovania čŕt na základe dátumu uskutočnenia modelovanej udalosti v prípade, že sa uskutočnila pred 15. dňom mesiaca. Keďže predikčné obdobie pokrýva dva mesiace, ako aj vidíme na Obr.2 a Obr. 3 pozitívna vzorka môže byť priradená k viacerým hodnotám *month_id*. Preto pre každého zákazníka sa z dvoch mesiacov náhodne vyberie jeden konkrétny *month_id* z ktorého sa k zákazníkovi priradia črty.

### Konštrukcia negatívnej vzorky

Negatívna vzorka naopak predstavuje skupinu zákazníkov, ktorí nevykonali modelovanú udalosť. Množina negatívnych vzoriek je tvorená kombináciami *customerid* a *month_id*, ktoré nespĺňajú podmienky pozitívnej vzorky. Pričom pre každého z nich sa náhodne vyberie iba jeden *month_id*, nakoľko každý zákazník sa môže vo vzorke nachádzať najviac raz.

### Výsledná vzorka

Nakoniec dostávame vzorku, ktorá obsahuje pozorovania aj pozítivnej aj negatívnej vzorky. Daná vzorka má nasledovnú štruktúru:

| month_id | customerid | event_occured | event_dt |
|---|---|---|---|

**Tabuľka 4:** Štruktúra výberovej vzorky

Popíšeme aj stĺpce, ktoré daná dátová vzorka obsahuje:

- *month_id* - stĺpec určujúci, do ktorého posledného dňa boli informácie použité na tvorbu čŕt;
- *customerid*- identifikátor zákazníka v rámci banky;
- *event_occured* - binárna hodnota, ktorá napovedá o tom, či nastala modelovaná udalosť;
- *event_dt* - dátum nastania modelovanej udalosti.

Pričom pre zákazníkov patriacích do negatívnej vzorky dátum nastania modelovanej udalosti neexistuje, nakoľko žiadna udalosť nenastala. Napriek tomu je potrebné tento údaj zachovať, keďže zohráva úlohu pri tvorbe príčinnej premennej, a teda aj pre tieto záznamy musí byť jej hodnota definovaná, v tomto prípade ako *null*. Tým sme náhodne priradili dátum „neuskutočnenia“ udalosti. Tento dátum bol priradený tak, aby bol konzistentný s distribúciou dátumov uskutočnenia udalostí v pozítivnej vzorke a zároveň zodpovedal pravidlám tvorby vzorky, ktoré boli znázornené na Obr.2 a Obr.3. Inými slovami, aby dátum „neuskutočnenia“ nepresahoval predtým stanovené *month_id*, z ktorého boli pre daného zákazníka načítané črty.

## 4.4 Definovanie príčinnej premennej pri dvoch typoch ošetrenia

V danej podkapitole uvažujme o dvoch typoch ošetrenia, ktoré reprezentuje, či bol zákazník oslovený marketingovou ponukou alebo nie. Naším cieľom je preto analyzovať rozdiel v pravdepodobnostiach výsledku za podmienky, že zákazník bol oslovený, v porovnaní so situáciou, keď nebol oslovený. Z tohto dôvodu sme museli skonštruovať aj príčinnú premennú, ktorá by indikovala, či zákazník bol oslovený s istou ponukou. Chceme kvantifikovať vzťah medzi príčinnou a závislou premennou. V prípade binárnej klasifikácie a dvoch typov ošetrenia, kde príčinná premenná nadobúda hodnoty $Z \in \{0, 1\}$ môžeme všetky kombinácie týchto dvoch čŕt klasifikovať nasledujúcim spôsobom:

| $Z$ | $Y$ | **Interpretácia** |
|---|---|---|
| 1 | 1 | Bol oslovený a uskutočnil modelovanú udalosť |
| 1 | 0 | Bol oslovený, ale neuskutočnil modelovanú udalosť |
| 0 | 1 | Nebol oslovený, ale uskutočnil modelovanú udalosť |
| 0 | 0 | Nebol oslovený a ani neuskutočnil modelovanú udalosť |

**Tabuľka 5:** Možné stavy modelovanej udalosti a ošetrenia

V Tab.5 sú uvedené všetky možné kombinácie týchto dvoch čŕt spolu s ich interpretáciou. Pričom tieto kombinácie zodpovedajú štyrom typom zákazníkov ilustrovaných na Obr.1.

K dispozícii sme mali dáta o kontaktovaní zákazníkov s rôznymi ponukami, tabuľka má nasledovnú štruktúru:

| contact_dt | channel | p | cam_name |
|---|---|---|---|

**Tabuľka 6:** Štruktúra tabuľky kontaktovania zákazníkov

Samotná tabuľka obsahuje viac údajov. V Tab. 6 sme uviedli len tie, ktoré sú pre nás v rámci práce relevantné. Konkrétne sú to:

- *contact_dt* - dátumy kontaktovania zákazníka;
- *channel* - komunikačný kanál použitý na oslovenie zákazníka;
- *p* - typ produktu, na ktorý je kampaň zameraná;
- *cam_name*- názov kampane, v rámci ktorej kontaktovanie prebehlo.

Na základe dostupných a releventných údajov uvedených v Tab. 6 cieľom je vytvoriť binárnu premennú *treatment*, ktorá by indikovala, či bol zákazník oslovený.

Ako prvé na základe hodnôt stĺpca *cam_name* sme odstránili pasívne kampane. Pasívne kampane predstavujú typ marketingových kampaní, pri ktorých zákazník prichádza do kontaktu s ponukou bez priameho oslovenia zo strany banky. Následne sme pomocou stĺpca produktu vybrali iba relevantné kampane pre trénovaný model. V rámci danej práce skúmame vplyv oslovenia zákazníka s cieľom založiť si spotrebný úver. Z tohto dôvodu nás budú zaujímať kampane, ktoré boli zamerané na ponuku spotrebného úveru. V ďalších krokoch už s danými stĺpcami pracovať nebudeme, preto ich odstránime.

V tomto prípade sme sa zamerali iba na to, či bol zákazník oslovený v období 60 dní (približne dva mesiace) pred uskutočnením modelovanej udalosti. Tento časový rámec bol vybraný na základe predpokladu, že kontaktovanie zákazníka v tomto období môže mať významný vplyv na jeho rozhodovanie a následné nákupné správanie. Po vytvorení dátovej vzorky sme tabuľku kontaktovaní spojili so vzorkou pomocou operácie *left join* a to podľa stĺpca *customerid*, ktorý slúži ako identifikátor zákazníka. Tým pádom sme dostali tabuľku s nasledovnou štruktúrou: Niektorí zákazníci mohli byť zaradení do výberovej vzorky, avšak nemajú žiadny záznam v tabuľke kontaktovaní. Z tohto dôvodu môžu po

| month_id | customerid | event_occured | event_dt | contact_dt |
|---|---|---|---|---|

**Tabuľka 7:** Štruktúra tabuľky po spojení

spojení tabuliek niektoré hodnoty v stĺpci *contact_dt* chýbať. Na základe toho týmto záznamom priraďujeme hodnotu *treatment = 0*, čím explicitne vyjadrujeme, že zákazník nebol oslovený.

Pre zákazníkov, ktorí patria do pozitívnej vzorky, sme na základe dátumu uskutočnenia udalostí analyzovali, či v období 60 dní pred touto udalosťou bol zákazník oslovený. Skúmali sme, či dátum oslovenia( stĺpec *conctactdatetime*) predchádzal dátumu nastania udalosti (stĺpec *event_dt*) najviac o 60 dní, pričom zároveň platí, že oslovenie muselo nastať najneskôr v deň udalosti (t. j. rozdiel medzi dátumom udalosti a dátumom oslovenia sa pohybuje v rozsahu 0 až 60 dní). Pre zákazníkov, ktorí patria do negatívnej vzorky bol postup analogický, ale do úvahy sa bral náhodne vygenerovaný dátum neuskutočnenia udalosti a porovnával sa s dátumumom kontaktovania, ak už predtým nebola priradená hodnota *treatment = 0*. Na základe toho sme definovali binárnu príčinnú premennú $Z = \{0, 1\}$, kde $Z = 1$, ak zákazník bol oslovený v tomto časovom intervale, v opačnom prípade $Z = 0$, čo hovorí, že zákazník oslovený nebol. Po vykonaní vyššie uvedených operácií stĺpce *event_dt* a *conctactdatetime* boli z tabuľky odstránené.

## 4.5 Definovanie príčinnej premennej pri viacerých typoch ošetrenia

V praxi však často nastávajú situácie, kedy môžeme zákazníkovi poskytnúť viacero rôznych typov zásahov. Preto chceme zohľadniť aj typ kanála, prostredníctvom ktorého bol zákazník oslovený. Z tohto dôvodu sme v ďalšom kroku našu tabuľku rozšírili o stĺpec *channel_type*, ktorý bude obsahovať kategorické hodnoty. V danom prípade už by sme teda nemali dve skupiny: ošetrení (oslovení) a kontrolovaní (neoslovení).

Existuje viacero kanálov, cez ktoré zákazník môže byť oslovený, ktoré by sme vedeli na základe nákladov klasifikovať do dvoch nasledujúcich skupín:

- tvrdý kanál oslovenia (angl. *hard channel*)- vysokonákladný spôsob komunikácie so zákazníkmi, vyžadujúci vyššie výdavky;

- mäkký kanál oslovenia (angl. *soft channel*)- nízkonákladná forma komunikácie firmy so zákazníkmi.

Uvádzame aj tabuľku s príkladmi oslovení v závislosi od kanálu:

| **Typ kanála** | **Príklady** |
|---|---|
| Tvrdý | Telefonát, osobný mail |
| Mäkký | SMS, push notifikácia, správa v aplikácii, hromadný mail |

**Tabuľka 8:** Príklady kanálov v závislosti od ich typov

Môže však dôjsť aj k situácii, kedy zákazník bude oslovený aj cez mäkký, aj cez trvdý kanál, napríklad, ak by najprv dostal mail (oslovený cez mäkký kanál) a následne sa uskutoční telefonát (oslovený cez tvrdý kanál). Na to, aby sme priradili konkrétne oslovenia do typov kanálu sme zo stĺpca *channel* extrahovali unikátne textové hodnoty a na základe nákladovosti jednotlivých kanálov sme ich rozdelili do dvoch skupín uvedených v Tab. 8. Následne sme pomocou zoskupovania na úrovni zákazníka (operácia typu `group by`) určili, prostredníctvom ktorého kanála bol každý zákazník oslovený. Pričom ak zákazník oslovený nebol, tak sa mu v danom stĺpci priradí textová hodnota *channel_type* = "*none*". Nakoniec dostávame výslednú tabuľku, s ktorou budeme ďalej pri modelovaní pracovať a má nasledovnú štruktúru:

| month_id | customerid | event_occured | treatment | channel_type |
|---|---|---|---|---|

**Tabuľka 9:** Štruktúra finálnej tabuľky

V Tab. 9 je znázornené, aké stĺpce obsahuje výsledná tabuľka potom čo sme vytvorili príčinné premenné – *treatment* a *channel_type*. Pričom premenná *treatment* nadobúda hodnoty z množiny $\{0, 1\}$ a indikuje, či bol zákazník oslovený, zatiaľ čo premenná *channel_type* je kategorická a nadobúda textové hodnoty z množiny $\{\text{hard}, \text{soft}, \text{hard and soft}, \text{none}\}$, ktoré reprezentujú typ komunikačného kanála – tvrdý, mäkký, kombináciu oboch alebo žiadny.

## 4.6 Výber čŕt pre trénovanie

Každý zákazník v dátovej vzorke obsahuje 2152 čŕt popisujúcich jeho správanie. Pri modelovaní spravidla používame len podmnožinu týchto čŕt, ktoré vyberáme na základe vopred definovaných kritérií. Konkrétny počet zvolených čŕt závisí aj od použitej metódy výberu čŕt. Redukcia počtu vstupných čŕt prispieva k lepšej interpretovateľnosti modelu a znižuje riziko pretrénovania. Okrem toho týmto spôsobom môžeme efektívnejšie a rýchlejšie trénovať modely, čo je dôležité pri práci s veľkými množstvami dát, ktoré sú pamäťovo a výpočtovo náročné. V danej časti predstavíme postup pre výber čŕt do modelu, navrhnutý spoločnosťou ČSOB.

### 4.6.1 Výber čŕt na základe korelácie so závislou premennou

Výber čŕt prebieha typicky v dvoch krokoch:

1. **Odstránenie čŕt s nízkou variabilitou:** Črty, pre ktoré sa hodnota 0.99-kvantilu rovná hodnote 0.01-kvantilu, sú považované za nízko variabilné. Takéto črty sú označené za nerelevantné a sú odstránené z množiny čŕt.

2. **Výber čŕt podľa jednej z dvoch stratégií:**

   **Na základe vzájomnej informácie**

   - Vypočíta sa Pearsonova korelačná matica medzi všetkými vstupnými črtami.
   - Zo silno korelovaných dvojíc (takých, ktoré spĺňajú $|r| > \tau$, kde $\tau$ je vopred stanovený prah) sa ponechá črta s vyššou koreláciou voči závislej premennej, druhá sa odstráni.
   - Na zredukovanej množine čŕt sa vypočíta vzájomná informácia voči závislej premennej.
   - Vyberie sa $k$ prvých čŕt s najvyššou hodnotou vzájomnej informácie.

   Pričom definícia vzájomnej informácie a detaily k jej implementácii sú uvedené v Prílohe A.

**Na základe korelácie**

- Pre každú črtu sa vypočíta absolútna hodnota korelačného koeficienta vzhľadom na závislú premennú, pričom sa volí buď Pearsonov alebo Spearmanov korelačný koeficient.
- Črty sa zoradia zostupne podľa absolútnej hodnoty korelácie so závislou premennou.
- Črta sa zaradí do vybraných čŕt vtedy, ak jej korelácia so všetkými už vybranými črtami nepresahuje stanovený prah $\tau$.

Pri modelovaní sme sa rozhodli pre výber čŕt na základe korelácie a ako kritérium sme si zvolili Pearsonov korelačný koeficient. V rámci tejto práce bol prah $\tau$ stanovený na hodnotu 0.95. Na základe korelácie so závislou premennou sme pre trénovanie modelu ponechali 350 čŕt. Pričom túto množinu čŕt budeme označovať $\mathcal{F}_y$.

### 4.6.2 Výber čŕt na základe korelácie s príčinnou premennou

Naším cieľom je predikovať zmenu správania zákazníka v dôsledku oslovenia, preto pri výbere čŕt sme sa neobmedzili iba na koreláciu čŕt so závislou premennou, ale zobrali sme do úvahy aj koreláciu s príčinnou premennou (*treatment*). Pričom aplikujeme analogicky postup postavený na Pearsonovej korelácii popísaný v Kap. 4.6.1. Na základe korelácie s príčinnou premennou sme pre trénovanie modelu ponechali 100 čŕt. Pričom túto množinu čŕt budeme označovať $\mathcal{F}_t$.

### 4.6.3 Prístupy ku kombinácii výberu čŕt

Výber čŕt prebehol samostatne pre obe premenné, a tým pádom sme dostali dve množiny vybraných čŕt: $\mathcal{F}_t$, $\mathcal{F}_y$. Následne sme uvažovali tri stratégie, ako pri trénovaní modelu tieto množiny čŕt skombinovať. Nakoľko predikujeme efekt oslovenia zákazníka na jeho nákupné správanie, je dôležité zohľadniť spôsob, akým boli zákazníci vybraní na oslovenie. Výber zákazníkov na oslovenie v našom prípade neprebieha náhodne ani experimentálne, ale je determinovaný na základe iných dostupných čŕt. Preto by sme príčinnú premennú *treatment* chceli nejakým spôsobom izolovať. Z uvedeného dôvodu sme sa rozhodli preskúmať tri rôzne stratégie kombinácie čŕt vybraných samostatne pre premennú *treatment*

a závislú premennú s cieľom lepšie izolovať samotný efekt ošetrenia. Uvedieme zvolené prístupy:

1. **Výber čŕt korelovaných so závislou premennou** – do modelu vstupujú výhradne črty, ktoré vykazujú silnú koreláciu so závislou premennou. Danú množinu označujeme ako $\mathcal{F}_y$.

2. **Zjednotenie** oboch množín – zahŕňa všetky črty, ktoré sú silno korelované aspoň s jednou z čŕt. Túto množinu označujeme ako $\mathcal{F}_y \cup \mathcal{F}_t$.

3. **Rozdiel** množín – ponechá tie črty, ktoré vykazujú silnejšiu koreláciu so závislou črtou a zároveň sme odstránili črty, ktoré boli vysoko korelované s príčinnou črtou (*treatment*), čo nám dovoľuje izolovať príčinnú premennú. V tomto prípade danú množinu označujeme ako $\mathcal{F}_y \setminus \mathcal{F}_t$.

# 5 Kvantifikácia efektu oslovenia

V tejto kapitole porovnáme odhadnutý efekt ošetrenia pomocou nástrojov popísaných v Kap.1.

Výsledky budú reprezentované prostredníctvom husľových grafov (angl. *violin plot*), ktorými znázorníme rozdelenie odhadnutých individuálnych efektov ošetrenia (CATE). Taktiež pomocou prerušovanej čiary znázorníme nulovú hodnotu. Okrem vizualizácií budú výsledky doplnené o relevantné metriky hodnotiace kvalitu predikcie, konkrétne: AUUC (plocha pod krivkou upliftu), ako aj váženého ROC-AUC (plocha pod ROC krivkou) pre základné modely, ktoré slúžia na odhad jednotlivých pravdepodobností nastania modelovanej udalosti. Navyše, uvedieme hodnoty štandardných odchýlok (STD), ktoré sú vypočítané pomocou bootstrapu. Bootstrapový odhad bol realizovaný na základe 5000 opakovaní. Detaily k výpočtu sme uviedli v Prílohe A.

Taktiež uvedieme aj priemerný efekt ošetrenia (ATE). Pričom vzhľadom na postup popísaný v Kap. 4.6 sme mali tri množiny vybratých čŕt. Pre každú množinu vybratých čŕt sme natrénovali samostatný model, na základe ktorého boli odhadnuté CATE. Počet čŕt pre každý prístup uvádzame nižšie:

| Prístup | $\mathcal{F}_y \setminus \mathcal{F}_t$ | $\mathcal{F}_y$ | $\mathcal{F}_y \cup \mathcal{F}_t$ |
|---|---|---|---|
| Počet čŕt | 256 | 350 | 354 |

**Tabuľka 10:** Počet použitých čŕt pre jednotlivé prístupy

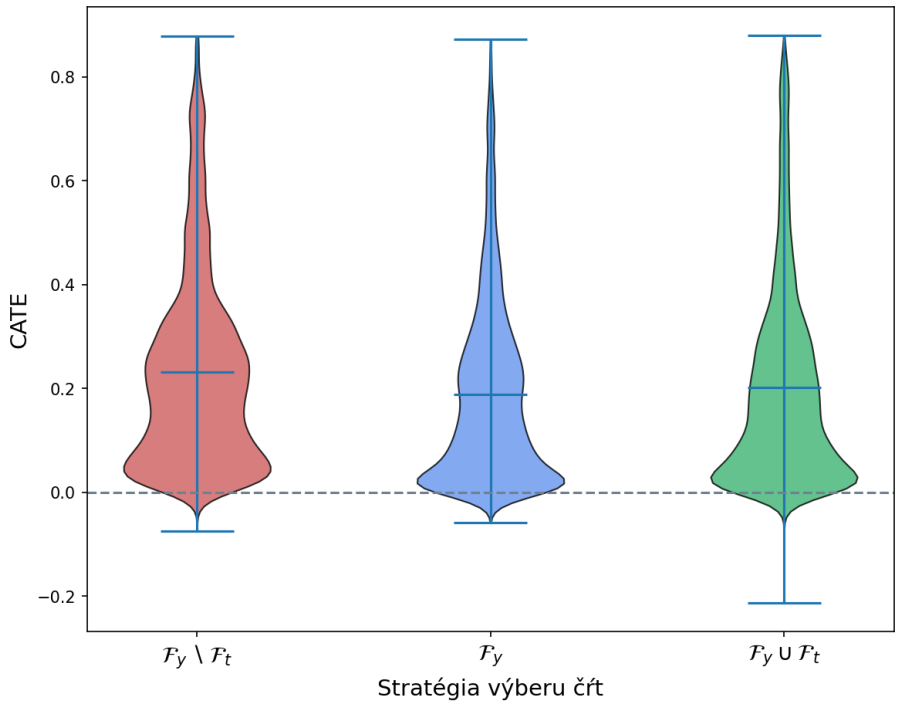
## 5.1 Aplikácia S-learnera

Pri trénovaní S-Learnera do množiny čŕt zákazníka sme pridali črtu ošetrenia. V tomto prípade okrem vybratých čŕt uvedených v Tab. 10 bol pridaný binárny stĺpec *treatment*. Z tohto dôvodu bol výsledný počet vstupných čŕt navýšený o jednotku a ich počty vieme vyjadriť pomocou nasledovnej tabuľky:

| Prístup | $\mathcal{F}_y \setminus \mathcal{F}_t$ | $\mathcal{F}_y$ | $\mathcal{F}_y \cup \mathcal{F}_t$ |
|---|---|---|---|
| Počet čŕt | 257 | 351 | 355 |

**Tabuľka 11:** Počet použitých čŕt pre jednotlivé prístupy

Rozdelenie CATE je znázornené na nasledujúcom obrázku.



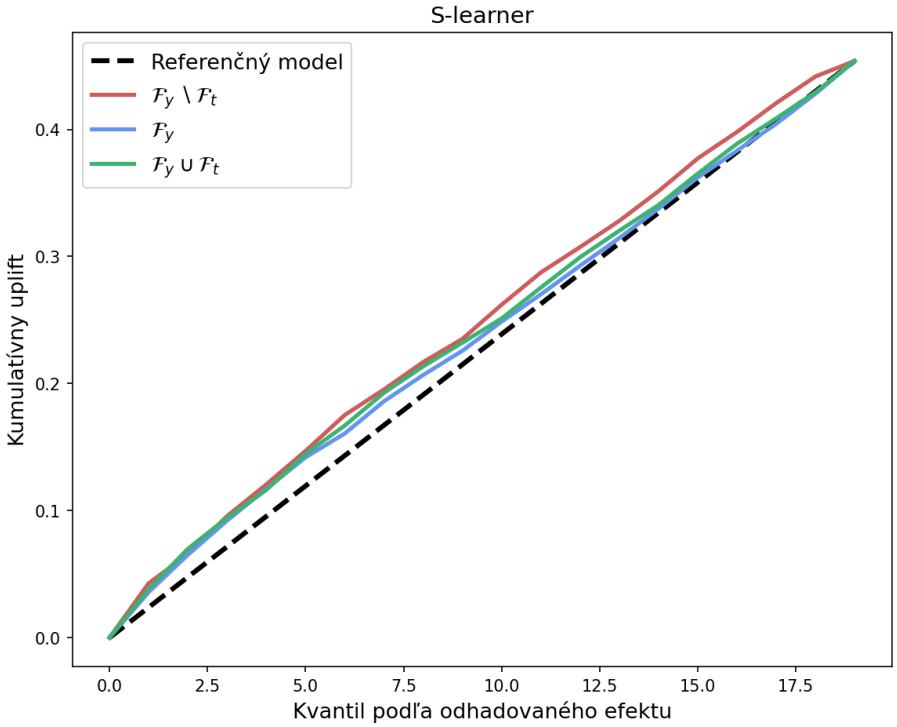

**Obr. 4:** Distribúcia odhadnutých CATE vzhľadom na výber čŕt

Na Obr. 4 sú znázornené výsledky odhadu CATE na testovacej vzorke.Sivá prerušovaná čiara označuje nulovú hodnotu, to platí aj pre ostatné grafy daného typu. Ako môžeme pozorovať, model len zriedkavo predikuje negatívny efekt oslovenia na správanie zákazníka, čo naznačuje, že prevažná väčšina predpokladaných účinkov je pozitívna alebo neutrálna. Taktiež vidíme, že väčšina odhadov efektu sa sústreďuje v blízkosti nulovej hodnoty.

| Množina čŕt | $\mathcal{F}_y \setminus \mathcal{F}_t$ | $\mathcal{F}_y$ | $\mathcal{F}_y \cup \mathcal{F}_t$ |
|---|---|---|---|
| ATE | 0.2321 | 0.2012 | 0.1886 |

**Tabuľka 12:** Priemerný efekt ošetrenia (ATE) pre S-learner vzhľadom na výber čŕt

Pomocou Tab. 12 sme pre každý z prístupov uviedli aj hodnoty ATE. Pričom najvyššia hodnota sa dosahuje pri množine $\mathcal{F}_y \setminus \mathcal{F}_t$.




**Obr. 5:** Uplift krivky pre S-learner vzhľadom na výber čŕt

Na Obr. 5 je zobrazená vizualizácia uplift kriviek pre každý z troch prístupov. Pri ich tvorbe za počet kvantilov $Q$ sme v rámci celej práce zvolili $Q = 20$. Tieto krivky poskytujú informáciu o schopnosti modelov rozlišovať medzi pozorovaniami s vyšším a nižším očakávaným prínosom z ošetrenia. Pričom vidíme, že krivky modelov ležia nad referenčným modelom, čo naznačuje, že model vie rozlíšiť medzi skupinami s vyšším a nižším efektom oslovenia. Všetky tri stratégie výberu čŕt vedú k modelom, ktoré dokážu identifikovať jednotlivcov s vyšším prínosom z ošetrenia. Zároveň však vidíme, že výber čŕt môže čiastočne ovplyvniť výsledky modelu.

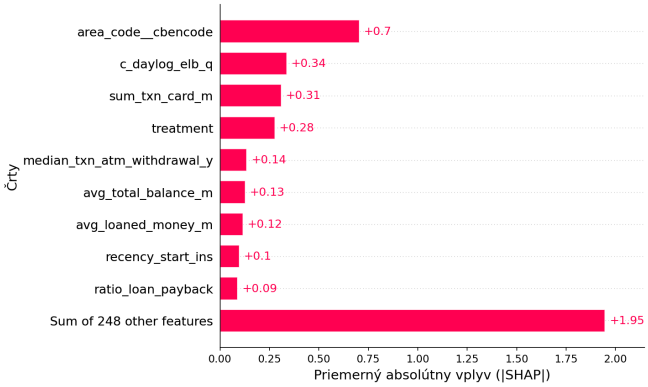
Taktiež k vizualizácii dopĺňame aj samotné AUUC hodnoty, pomocou ktorých môžeme vyhodnocovať rozlišovaciu schopnosť jednotlivých modelov.

| Množina čŕt | $\mathcal{F}_y \setminus \mathcal{F}_t$ | $\mathcal{F}_y$ | $\mathcal{F}_y \cup \mathcal{F}_t$ |
|---|---|---|---|
| AUUC | 0.2471 | 0.2368 | 0.2408 |

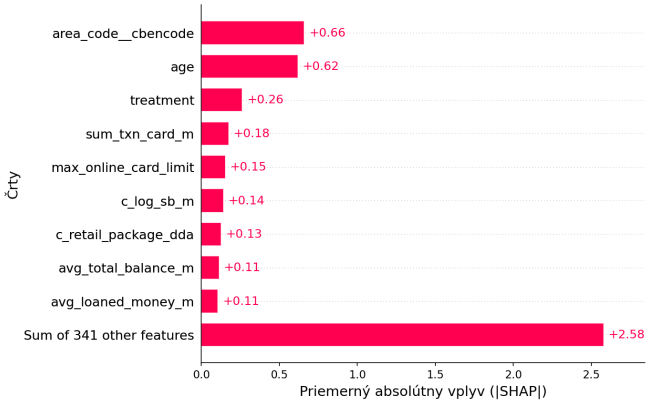
**Tabuľka 13:** AUUC hodnoty pre S-learner vzhľadom na výber čŕt

Tab. 13 uvádza hodnoty AUUC pre každý z modelov.

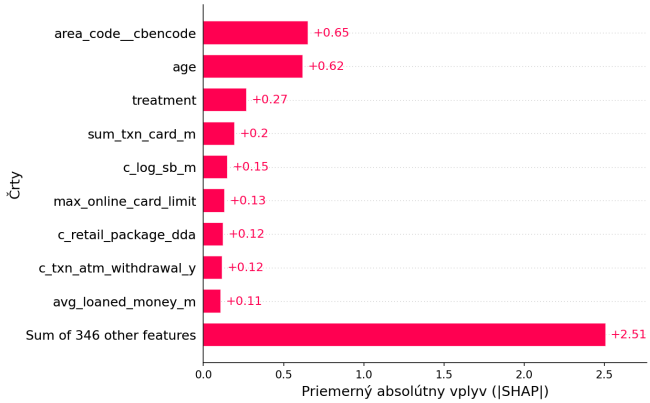
Keďže črta *treatment* bola súčasťou množiny čŕt, to nám umožňuje analyzovať jej významnosť v rámci výsledného modelu. Významnosť čŕt určujeme pomocou SHAP skóre. Detaily výpočtu SHAP skóre sú uvedené v Prílohe A. Nižšie znázorňujeme grafické výstupy, ktoré zobrazujú významnosť jednotlivých čŕt na základe hodnoty SHAP skóre:




**Obr. 6:** SHAP skóre pre $\mathcal{F}_y \cup \mathcal{F}_t$




**Obr. 7:** SHAP skóre pre $\mathcal{F}_y$




**Obr. 8:** SHAP skóre pre $\mathcal{F}_y \setminus \mathcal{F}_t$

Na Obr. 6, 7, 8 sú uvedené absolútne hodnoty SHAP skóre pre 10 najvýznamnejších čŕt v modeli. SHAP hodnota hovorí, aký je vplyv vstupnej črty na predikciu modelu. Napríklad na Obr.6 má absolútna hodnota SHAP skóre črty *treatment* hodnotu +0.28, čo napovedá, že daná črta ovplyvnila výsledok základného modelu (či zákazník vykoná modelovanú udalosť) približne o 0.28. Čím je absolútna hodnota SHAP skóre vyššia, tým väčší vplyv má daná črta na výstup modelu. Vidíme, že pri všetkých troch prístupoch na základe tejto metriky črta *treatment* je zaradená do prvých 10 najvýznamnejších čŕt a majú približne rovnaký vplyv na výsledok. Detaily k výpočtu SHAP skóre sú uvedené v Prílohe A.

V nasledujúcom kroku sa chceme pozrieť na výkonnosť základného modelu, ktorý je súčasťou S-learnera a slúži na predikciu výsledkov pri rôznych hodnotách ošetrenia. Pri S-learneri pre obidve skupiny trénujeme jeden spoločný model. Uvedieme hodnoty váženého ROC-AUC výpočet ktorého je uvedený v (57) a štandardnú odchýlku (STD):

| Množina čŕt | ROC-AUC ± STD |
|---|---|
| $\mathcal{F}_y \setminus \mathcal{F}_t$ | $0.83055 \pm 0.00331$ |
| $\mathcal{F}_y \cup \mathcal{F}_t$ | $0.84403 \pm 0.00298$ |
| $\mathcal{F}_y \cup \mathcal{F}_t$ | $0.84361 \pm 0.00331$ |

**Tabuľka 14:** Hodnoty váženého ROC-AUC a štandardnej odchýlky na testovacej vzorke pre základný učiteľ S-learnera

Tab. 14 uvádza hodnoty metriky váženého ROC-AUC, ktoré predstavuje vážený priemer na základe veľkostí skupín pri predikcii výstupu pre dva prípady: v každom riadku sa dosadí hodnota ošetrenej skupiny ($treatment = 1$) a hodnota kontrolnej skupiny ($treatment = 0$). Výsledky sú uvedené zvlášť pre každý z troch použitých prístupov výberu čŕt: ich rozdiel ($\mathcal{F}_y \setminus \mathcal{F}_t$), črty korelujúce so závislou premennou ($\mathcal{F}_y$) a zjednotenie čŕt ($\mathcal{F}_y \cup \mathcal{F}_t$). Z tabuľky vyplýva, že základný model pri S-learneri dosahuje vysokú hodnotu ROC-AUC v rámci testovacej vzorky.

## 5.2 Aplikácia T-learnera

Pri trénovaní T-Learnera sme trénovaciu dátovú vzorku rozdelili na základe hodnoty príčinnej črty na dve. V našom prípade išlo o rozdelenie na skupinu oslovených (ošetrená skupina) a neoslovených (kontrolná skupina) zákazníkov. Početné zastúpenie jednotlivých skupín uvádzame v tabuľkách nižšie: Následne pre každú skupinu bol natrénovaný samos-

| Skupina | Kontrolná skupina | Ošetrená skupina |
|---|---|---|
| Počet pozorovaní | 147813 | 12177 |

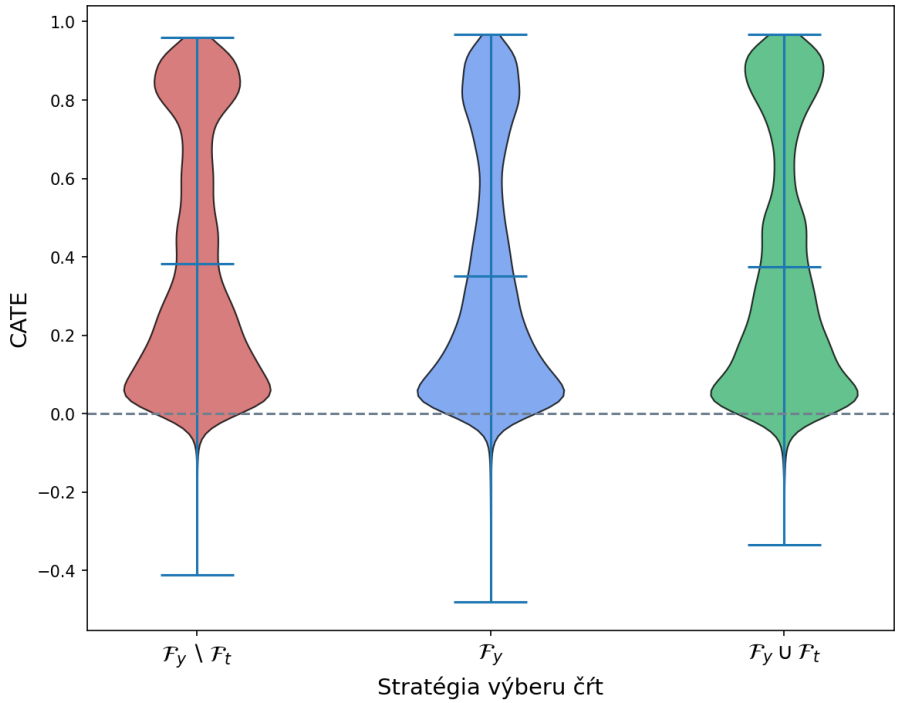
**Tabuľka 15:** Počet pozorovaní v každej skupine trénovacej sady

tatný model. Po natrénovaní samostatných modelov pre každého jednotlivca odhadneme výsledok v oboch scenároch: ak by bol oslovený(ošetrená skupina) a ak by nebol(oslovená

| Skupina | Kontrolná skupina | Ošetrená skupina |
|---|---|---|
| Počet pozorovaní | 36926 | 3072 |

**Tabuľka 16:** Počet pozorovaní v každej skupine testovacej sady

skupina). Tým pádom dostaneme dve predikcie: jednu z modelu, ktorý je trénovaný na oslovenej skupine, a druhú z modelu, trénovaného na neoslovenej skupine. Pričom počet použitých čŕt pri trénovaní zodpovedá uvedenému na Tab. 10. Odhadovaný efekt oslovenia zákazníka nájdeme ako rozdiel týchto predikcií. Potom výsledky pre CATE sú znázornené nižšie:



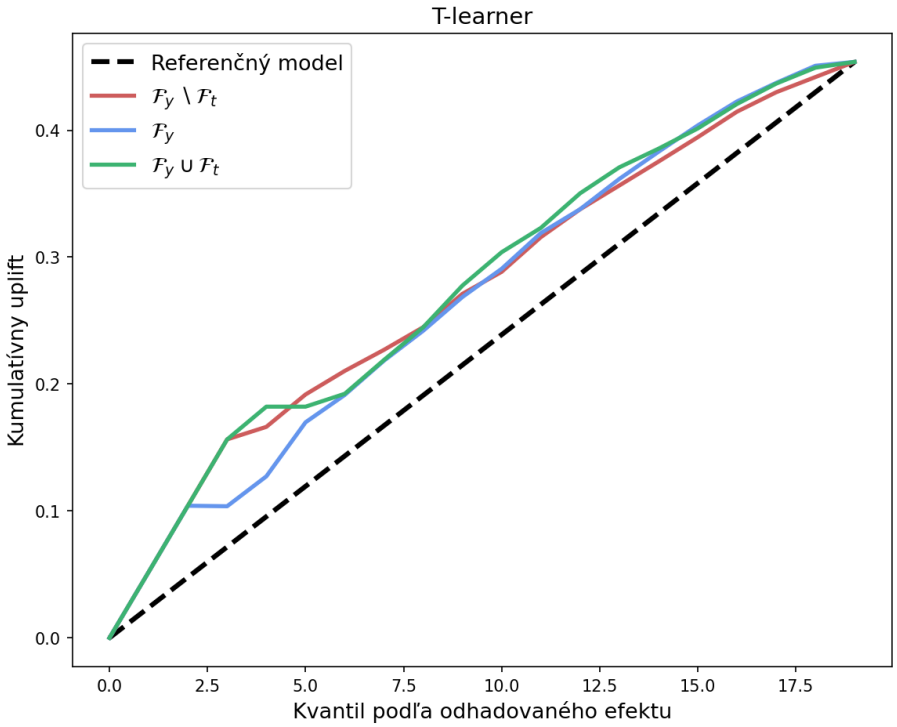

**Obr. 9:** Distribúcia odhadnutých CATE vzhľadom na výber čŕt pre T-learner

Na Obr.9 sú znázornené CATE hodnoty získané prostredníctvom T-learnera. Väčšina pozorovaní aj pre daný prístup nadobúda pozitívne hodnoty. Zároveň pozorujeme, že rozpätie hodnôt je pri všetkých prístupoch identické, avšak pri množinách výberu čŕt $\mathcal{F}_y \setminus \mathcal{F}_t$ a $\mathcal{F}_y \cup \mathcal{F}_t$ pozorujeme častejší výskyt vyšších hodnôt.

| Množina čŕt | $\mathcal{F}_y \setminus \mathcal{F}_t$ | $\mathcal{F}_y$ | $\mathcal{F}_y \cup \mathcal{F}_t$ |
|---|---|---|---|
| ATE | 0.3819 | 0.3511 | 0.3738 |

**Tabuľka 17:** Priemerný efekt ošetrenia (ATE) pre T-learner vzhľadom na výber čŕt

Prostredníctvom Tab. 17 sme uviedli hodnoty ATE získané daným prístupom. Následne pomocou grafu znázorníme uplift krivky pre daný prístup:




**Obr. 10:** Uplift krivky pre T-Learner vzhľadom na rôzne stratégie výberu čŕt

Na Obr. 10 je zobrazená vizualizácia uplift kriviek pre každý z troch prístupov. Pričom vo všetkých prípadoch krivky modelov ležia nad referenčným modelom, čo vykazuje dobrú rozlišovaciu schopnosť. V nasledujúcej tabuľke uvádzame hodnoty AUUC, znázorňujúce plochy pod týmito krivkami:

| Množina čŕt | $\mathcal{F}_y \setminus \mathcal{F}_t$ | $\mathcal{F}_y$ | $\mathcal{F}_y \cup \mathcal{F}_t$ |
|---|---|---|---|
| AUUC | 0.2739 | 0.269 | 0.2778 |

**Tabuľka 18:** AUUC hodnoty pre T-learner vzhľadom na výber čŕt

V danom prístupe pre model základného učiteľa sme pre každý prístup vypočítali

metriky váženého ROC-AUC spolu so štandardnými odchýlkami (STD):

| Množina čŕt | ROC-AUC ± STD |
|---|---|
| $\mathcal{F}_y \setminus \mathcal{F}_t$ | $0.81567 \pm 0.00358$ |
| $\mathcal{F}_y \cup \mathcal{F}_t$ | $0.83052 \pm 0.00313$ |
| $\mathcal{F}_y \cup \mathcal{F}_t$ | $0.0.83017 \pm 0.00358$ |

**Tabuľka 19:** Hodnoty váženého ROC-AUC a štandardnej odchýlky na testovacej vzorke pre základných učiteľov T-learnera

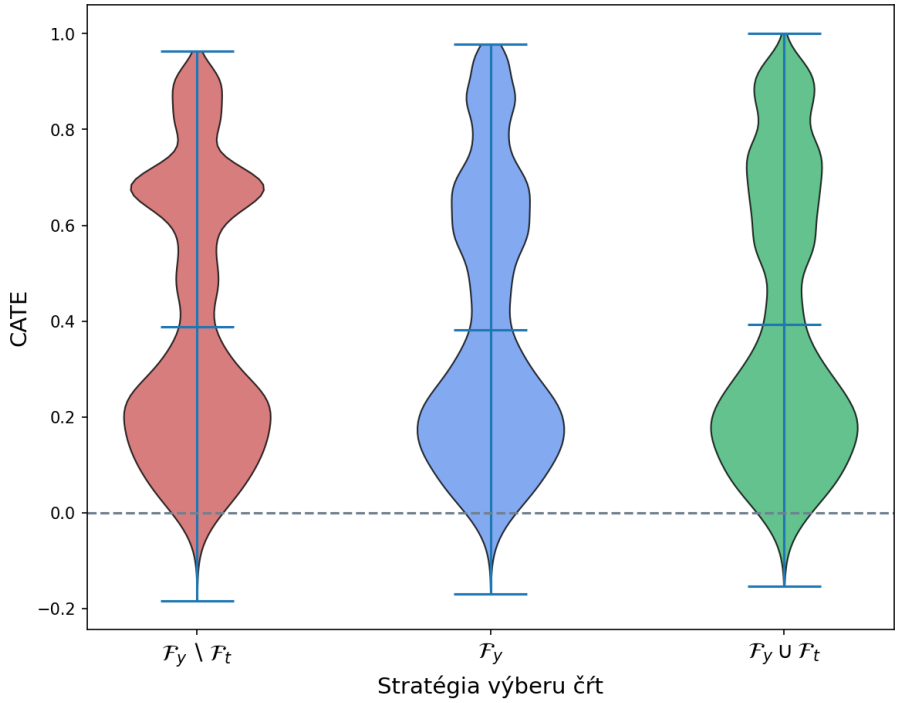
Tab. 19 uvádza hodnoty ROC-AUC vypočítané na testovacej vzorke ako vážený priemer hodnôt jednotlivých modelov podľa veľkosti skupín.

## 5.3 Aplikácia X-learnera

V prípade trénovania X-Learnera máme dve fázy, pričom v prvej fáze postupujeme ako pri T-Learneri, teda najskôr V prvej fáze aplikujeme rovnaký postup ako pri T-Learneri, teda trénujeme dva samostatné modely pre skupinu oslovených a neoslovených. Tieto modely sú prevzaté priamo z predchádzajúceho T-Learnera.

Na základe týchto odhadov boli vypočítané imputované efekty pre každého zákazníka v oboch skupinách.

Pri finálnom odhade sa používa váhová funkcia, ktorá slúži na vyváženie výsledkov z oboch modelov. V rámci trénovania modelu sme ju teda skonštruovali pomocou odhadu pravdepodobnosti zaradenia zákazníka do oslovenej skupiny. Pomocou logistickej regresie sme pre každého zákazníka vyrátali pravdepodobnosť zariadenia do oslovenej skupiny a následne sme danú hodnotu použili na výpočet odhadov CATE. Uvádzame získané výsledky pomocou nasledujúceho obrázku:



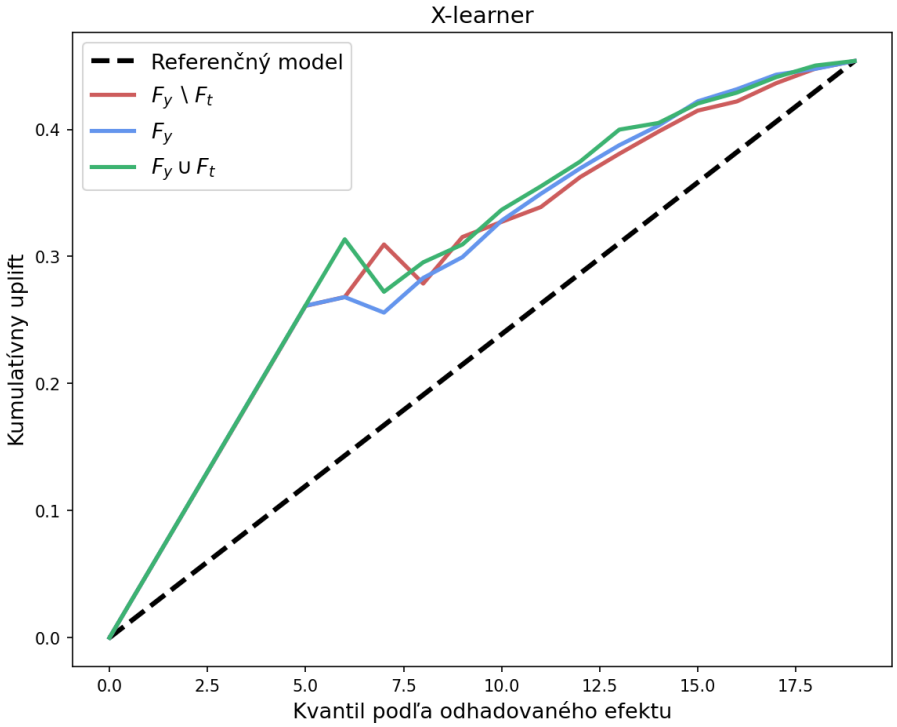

**Obr. 11:** Distribúcia odhadnutých CATE vzhľadom na výber čŕt pre X-learner

Na Obr. 11 sme pomocou husľového grafu znázornili distribúciu odhadov CATE. Pozorujeme, že pri všetkých prístupoch majú tieto odhady podobnú distribúciu. Zároveň daný prístup zachytáva aj negatívne efekty oslovenia, ako aj častejší výskyt vyšších hodnôt CATE.

| Množina čŕt | $\mathcal{F}_y \setminus \mathcal{F}_t$ | $\mathcal{F}_y$ | $\mathcal{F}_y \cup \mathcal{F}_t$ |
|---|---|---|---|
| ATE | 0.3883 | 0.3816 | 0.3931 |

**Tabuľka 20:** Priemerný efekt ošetrenia (ATE) pre X-learner vzhľadom na výber čŕt

Na základne hodnôt ATE uvedených v Tab. 20 vidíme, že pre všetky tri prístupy dosahujeme podobné hodnoty ATE. Pričom pri danom prístupe pozorujeme vyššie priemerné hodnoty ATE oproti ostatným prístupom.




**Obr. 12:** Uplift krivky pre X-Learner vzhľadom na rôzne stratégie výberu čŕt

Na Obr. 12 sú zobrazené uplift krivky, pomocou ktorých sa vieme pozrieť na rozlišovaciu schopnosť daného prístupu. Pričom vo všetkých prípadoch krivky sa nachádzajú nad referenčným modelom. Uvedieme aj hodnoty AUUC, ktoré vyjadrujú plochu pod krivkami:

| Množina čŕt | $\mathcal{F}_y \setminus \mathcal{F}_t$ | $\mathcal{F}_y$ | $\mathcal{F}_y \cup \mathcal{F}_t$ |
|---|---|---|---|
| AUUC | 0.3005 | 0.2999 | 0.306 |

**Tabuľka 21:** AUUC hodnoty pre X-learner vzhľadom na výber čŕt

## 5.4 Porovnanie použitých nástrojov

Pri porovnaní troch prístupov (S-learner, T-learner a X-learner) sa sústredíme na tri aspekty: rozlišovaciu schopnosť modelu, vyjadrenú prostredníctvom AUUC, výpočtovú a implementačnú náročnosť, charakter distribúcie odhadnutej metódou hodnôt CATE. Pričom dané metódy porovnáme na základe výsledkov na zvolenej množine čŕt $F_y \setminus F_t$. Dané rozhodnutie vieme zdôvodniť tým, že všetky tri prístupy vykazovali podobnú výkonnosť.

### Rozlišovacia schopnosť pomocou AUUC

Na porovnanie rozlišovacej schopnosti použijeme metriku AUUC, preto v nasledujúcej tabuľke uvádzame hodnoty pre každý z prístupov na tejto množine čŕt $F_y \setminus F_t$:

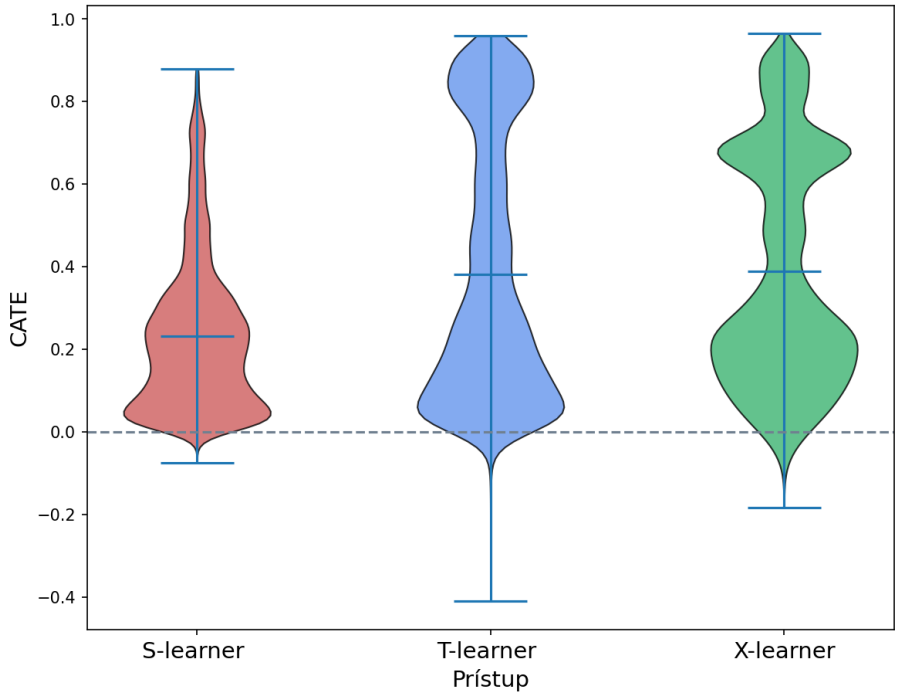
| Prístup | S-learner | T-learner | X-learner |
|---|---|---|---|
| AUUC | 0.2471 | 0.2739 | 0.3005 |

**Tabuľka 22:** AUUC hodnoty pre každý z prístupov na množine čŕt $F_y \setminus F_t$

Na základe hodnôt metriky AUUC v Tab. 22 vidíme, že X-learner dosahuje najvyššiu hodnotu.

### Porovnanie odhadov CATE medzi modelmi

Pre lepšiu prehľadnosť pomocou husľových grafov znázorníme distribúcie odhadov CATE pre všetky prístupy pre danú množinu čŕt.




**Obr. 13:** Odhady CATE pre každý z prístupov na množine čŕt $F_y \setminus F_t$

Z obrázku vidíme, že S-learner predikuje nižšie odhady CATE. T-learner a X-learner vykazujú väčší rozptyl v odhadoch a zároveň vyššie hodnoty ATE, ktoré sú uvedené v

Tab.12, 17, 20. Taktiež častejšie dosahujú záporné hodnoty odhadov CATE, a teda vedia viac zachytiť záporný efekt z oslovenia.

### Výpočtová a implementačná náročnosť

Z hľadiska implementačnej a výpočtovej náročnosti vieme usúdiť:

- S-learner sa javí ako najmenej náročný prístup, nakoľko spočíva v trénovaní jedného spoločného modelu do ktorého príčinná premenná vstupuje ako dodatočná črta.
- T-learner trénuje dva samostatné modely pre každú zo skupín, čo zvyšuje výpočtovú náročnosť.
- X-learner je z troch prístupov najkomplexnejší, nakoľko okrem trénovania samostatných modelov pre každú zo skupín berie do úvahy aj pravdepodobnosť zaradenia do danej skupiny.

Čas potrebný na trénovanie daných modelov spolu s parametrami prostredia, na ktorom boli operácie vykonané, sú uvedené v Prílohe B.

# 6 Kvantifikácia efektu oslovenia prostredníctvom rôznych kanálov

V danej kapitole uvedieme výsledky z nástrojov na odhad CATE, ktoré sme predstavili v Kap.1.4. V kontexte danej práce viacúrovňové ošetrenie predsatvuje rozšírenie modelov o premennú *channel_type*, ktorá hovorí cez aký kanál bol zákazník oslovený. Konkrétne máme štryi skupiny:

1. oslovený cez tvrdý kanál,
2. oslovený cez mäkký kanál,
3. oslovený súčasne cez mäkký aj tvrdý kanál,
4. neoslovený cez žiadny kanál.

Počty pozorovaní v jednotlivých skupinách sú uvedené nižšie, a to pre trénovaciu a testovaciu vzorku. Na Tab. 23, 24 vidíme početnosti jednotlivých skupín. Kapacita mäk-

| Skupina | 1. | 2. | 3. | 4. |
|---|---|---|---|---|
| Počet pozorovaní | 209 | 3633 | 8335 | 147813 |

**Tabuľka 23:** Počet pozorovaní v každej skupine trénovacej vzorky na základe kanálu oslovenia

| Skupina | 1. | 2. | 3. | 4. |
|---|---|---|---|---|
| Počet pozorovaní | 52 | 863 | 2157 | 36926 |

**Tabuľka 24:** Počet pozorovaní v každej skupine testovacej vzorky na základe kanálu oslovenia

kého kanálu je limitovaná len kontaktnými dátami, zatiaľ čo v prípade tvrdého kanálu sú fixne nastavené kapacity agentov v kontaktnom centre a pracovníkov na pobočke. Z tohto dôvodu skupina oslovených cez tvrdý kanál obsahuje najmenej pozorovaní. Taktiež v poslednej dobe sa ČSOB zameriava na kombinované oslovenie, napríklad najprv osloví zákazníka cez mail a neskôr uskutoční telefonát, preto daná skupina má väčšie zastúpenie oproti ostatným.

Štvrtú skupinu budeme považovať za referenčnú skupinu, voči ktorej budeme počítať odhady CATE pre jednotlivé typy ošetrení. Zvolili sme práve túto skupinu ako referenčnú, nakoľko sa pozeráme na efekt oslovenia cez daný kanál oproti tomu, ak by sa žiadne neuskutočnilo. Cieľom je porovnať účinnosť rôznych kanálov na základe odhadnutých CATE a zároveň interpretovať ich z ohľadom na špecifiká použitých nástrojov.

Ako množinu vstupných čŕt zvolíme $\mathcal{F}_y \setminus \mathcal{F}_t$, ktorá obsahuje 256 čŕt. Dané rozhodnutie vychádza z toho dôvodu, že pri porovnaní výsledkov s dvoma typmi ošetrenia všetky tri prístupy výberu čŕt vykazovali porovnateľnú výkonnosť. Preto predpokladáme, že zvolená množina bude postačujúca aj v tomto prípade.

## 6.1 Aplikácia S-learnera na viacúrovňové ošetrenie

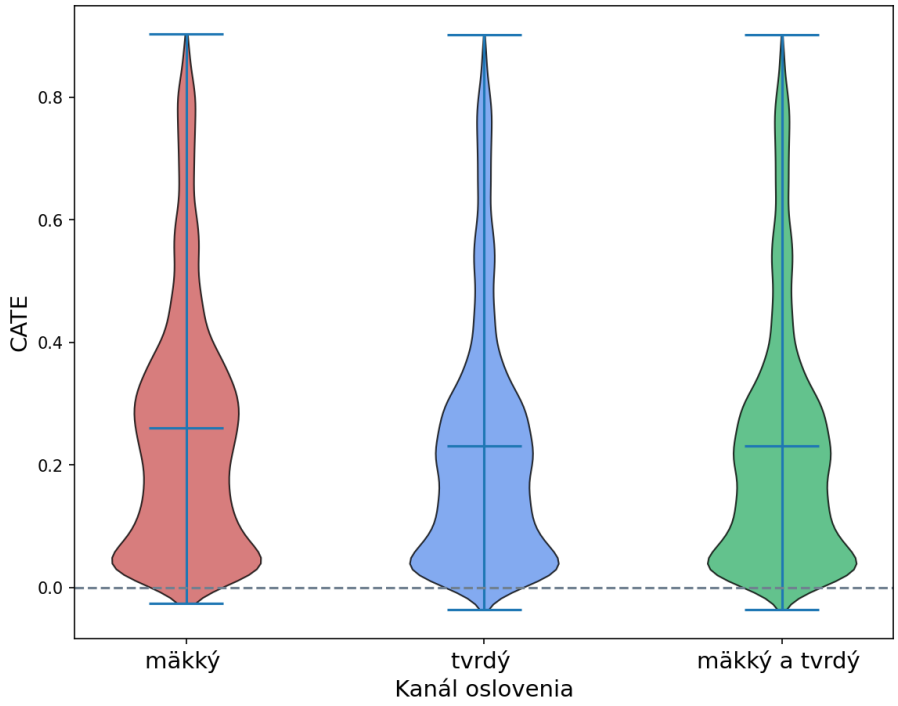
Pri trénovaní S-Learnera do množiny čŕt pridávame príčinnú premennú. V tomto prípade teda pridávame kategorickú premennú *channel_type*. Tým pádom ako aj všetky kategorické premenné zahrnuté do trénovania, sme túto premennú potrebovali transformovať do numerickej formy.

### 6.1.1 Zakódovanie komunikačného kanála pomocou dummy premenných

Premennú *channel_type* sme zakódovali pomocou dummy premenných, ktoré sme popísali v Kap.2.5, pričom každá kategória je reprezentovaná samostatným binárnym stĺpcom. Daná premenná pozostáva zo štyroch unikátnych hodnôt. Pri zakódovaní sme ju rozdielili na tri stĺpce pre tri typy kanálov. Daný prístup nám zabezpečuje jednoznačnú a interpretovateľnú reprezentáciu jednotlivých typov ošetrenia. CatBoost zakódovanie vychádza z priraďovania priemernej hodnoty závislej premennej, ktorá na základe podobnosti priemerov priradí rôznym zásahom podobné reprezentácie. To však môže viesť k strate kontrafaktuálneho rozdielu medzi ošetreniami, ktorý sa snažíme pomocou modelovania odhadnúť.

### 6.1.2 Výsledky modelu

Po zakódovaní sme natrénovali jeden model, ktorý sme následne použili na predikciu výsledkov pre všetky typy ošetrení. Na nasledujúcom obrázku uvádzame rozloženie hodnôt CATE pre tri rôzne kanály oslovenia:




**Obr. 14:** Distribúcia odhadnutých CATE vzhľadom na výber čŕt pre S-learner s viacerými ošetreniami

Na Obr.14 pozorujeme, že odhady CATE pre všetky tri kanály oslovení (mäkký, tvrdý, mäkký a tvrdý) majú väčšinou pozitívny efekt. Výsledky vykazujú podobnú distribúciu hodnôt medzi kanálmi. Navyše pozorujeme zriedkavý výskyt záporných hodnôt. Niektoré hodnoty sú sústredené okolo nuly, čo naznačuje slabší efekt z oslovenia cez ten konkrétny kanál.

| Kanál oslovenia | mäkký | tvrdý | tvrdý a mäkký |
|---|---|---|---|
| ATE | 0.2604 | 0.2307 | 0.2308 |

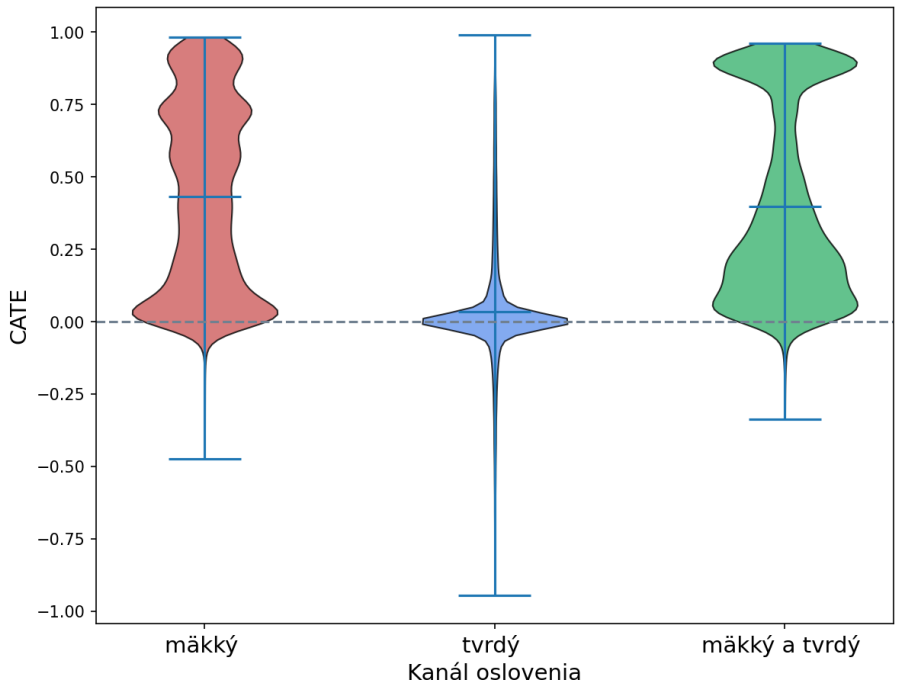
**Tabuľka 25:** Priemerný efekt ošetrenia (ATE) pre S-learner pre rôzne kanály

V Tab. 25 sú uvedené hodnoty ATE, pričom mäkký kanál má najvyššiu hodnotu a teda na základe daných výsledkov, môžeme tvrdiť, že dané oslovenie má v priemere najvyšší efekt.

Vypočítali sme aj hodnotu váženého ROC-AUC spolu so štandardnou odchýlkou (STD). V danom prípade ROC-AUC $\pm$ STD $= 0.8301 \pm 0.0357$.

## 6.2 Aplikácia T-learnera na viacúrovňové ošetrenie

Pri trénovaní T-learnera sme pre každú skupinu oslovenia vrátane referenčnej skupiny natrénovali samostatný model. Počty pozorovaní v jednotlivých skupinách trénovacej sady sú uvedené v Tab. 23. Následne sme pre každý z týchto modelov vypočítali rozdiel medzi jeho predikciou pravdepodobnosti výstupu a predikciou referenčného modelu, a to na celej testovacej vzorke. Tento rozdiel predstavuje odhad CATE pre príslušný typ ošetrenia, ktorý ilustrujeme nižšie:



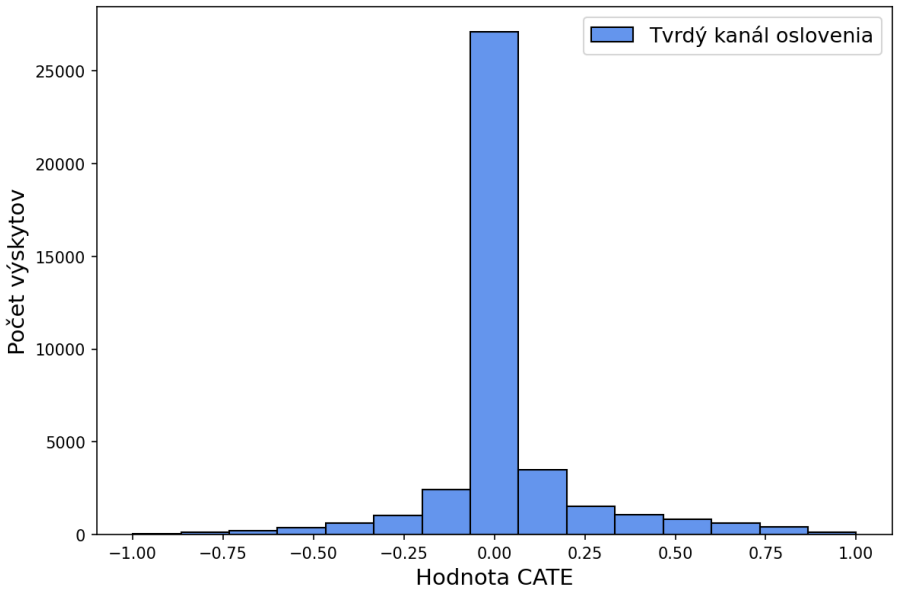

**Obr. 15:** Distribúcia odhadnutých CATE vzhľadom na výber čŕt pre T-learner s viacerými ošetreniami

Na Obr. 15 pozorujeme rozdiely v distribúciách odhadnutého efektu medzi jednotlivými kanálmi oslovenia pri použití T-learnera. Pričom vidíme, že odhady CATE sú občas aj záporné, čo poskytuje informáciu o tom, že oslovenie cez daný kanál má skôr negatívny efekt.

| Kanál oslovenia | mäkký | tvrdý | tvrdý a mäkký |
|---|---|---|---|
| ATE | 0.4313 | 0.0339 | 0.3977 |

**Tabuľka 26:** Priemerný efekt ošetrenia (ATE) pre T-learner pre rôzne kanály

V Tab. 26 sú uvedené hodnoty ATE pre každý z kanálov získaných T-learnerom. Pričom mäkký kanál a kombinácia mäkkého a tvrdého kanála vykazujú pomerne výsoku hodnotu priemerného efektu, zatiaľ čo pri tvrdom kanáli je priemerný efekt blízky nule. Zároveň ide o skupinu s najnižším zastúpením, čo mohlo ovplyvniť výsledky modelu. Pozrieme sa na distribúciu CATE:




**Obr. 16:** Distribúcia odhadnutých CATE pri oslovení tvrdým kanálom pre T-Learner

Na Obr. 16 pomocou histogramu sme znázornili distribúciu CATE pre tvrdý kanál oslovenia. Vidíme, že aj napriek nulovej hodnote ATE, niektoré jednotky dosahujú od nuly odlíšne hodnoty CATE. To naznačuje, že model dokáže rozlíšiť, pre koho dané oslovenie má efekt a pre koho nie.

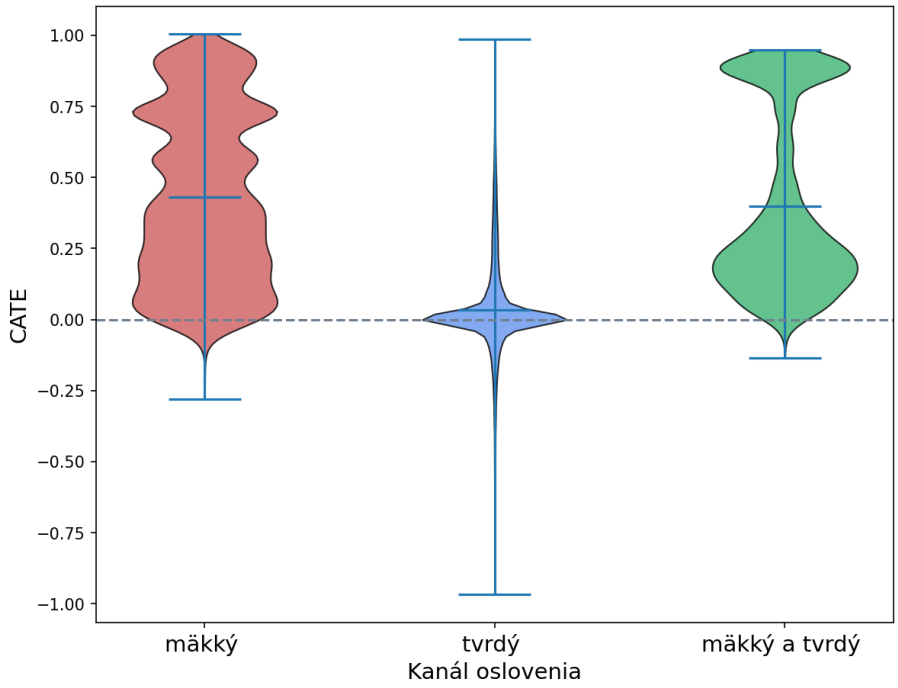
Aj v tomto prípade sme vypočítali hodnotu váženého ROC-AUC spolu so štandardnou odchýlkou (STD). Dostali sme ROC-AUC $\pm$ STD $= 0.8153 \pm 0.0358$.

## 6.3 Aplikácia X-learnera na viacúrovňové ošetrenie

V prvej fáze X-learnera sme využili už natrénované modely z T-learnera, ktoré odhadovali výsledky pre každé ošetrenie aj pre referenčnú skupinu. Pričom početné zastúpenie jednotlivých skupín je rovnaké. Následne na ich základe sme pre všetky ošetrenia ako aj pre referenčnú skupinu vypočítali imputované efekty a aj samotné CATE. Ako váhovú funkciu sme opäť zvolili pravdepodobnosť, že jednotka dostane určitý typ oslovenia, resp.

žiadne. Nakoľko máme viac ako dva typy ošetrení, tak v danom prípade sme na odhad pravdepodobností sme použili multinomiálnu logistickú regresiu.

Distribúcia odhadov CATE je znázornená na obrázku nižšie:



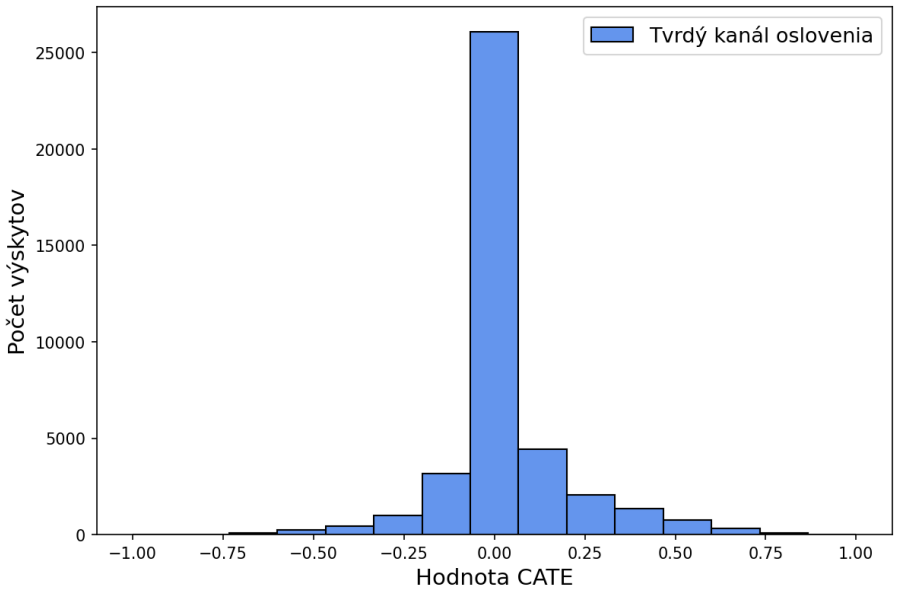

**Obr. 17:** Distribúcia odhadnutých CATE vzhľadom na výber čŕt pre X-learner s viacerými ošetreniami

Na Obr. 17 vidíme, že daný prístup vykazuje rozdiely v distribúcii hodnôt CATE medzi jednotlivými kanálmi ošetrenia. Pričom pri danom prístupe pozorujeme častejší výskyt záporných hodnôt.

| Kanál oslovenia | mäkký | tvrdý | tvrdý a mäkký |
|---|---|---|---|
| ATE | 0.4316 | 0.0344 | 0.3978 |

**Tabuľka 27:** Priemerný efekt ošetrenia (ATE) pre X-learner pre rôzne kanály

V Tab. 27 sú uvedené hodnoty ATE pre každý z kanálov získaných X-learnerom. Pričom hodnota ATE pre tvrdý kanál je blízka nulovej hodnote. Aj v tomto prípade si znázorníme rozdelenie samotných CATE pre daný kanál:




**Obr. 18:** Distribúcia odhadnutých CATE pri oslovení tvrdým kanálom pre X-Learner

Na Obr.18 je zobrazený histogram s hodnotami CATE pre tvrdý kanál oslovenia. Pomocou neho vidíme, že efekty oslovenia daným kanálom nadobúdajú aj hodnoty odlíšne od nuly, čo napovedá, že pri niektorých skupinách zákazníkov model vie zachytiť, pre koho je dané oslovenie efektívnym. Niektoré hodnoty CATE sú aj záporne, čo nám napovedá ktorých zákazníkov by daný typ oslovenia skôr odradil. V kontexte našej práce chceme zákazníkovi priradiť ten kanál, pri ktorom má oslovenie najvyšší efekt.

## 6.4 Stratégia priraďovania ošetrenia

Pri viacúrovňovom ošetrení sme pre každú skupinu dostali odhad CATE. Konečnou úlohou je každej jednotke prideliť typ kanálu, pre ktorý dosahuje najvyššiu hodnotu CATE.

| $CATE_1$ | $CATE_2$ | $CATE_3$ | Odporúčané ošetrenie | Priradené ošetrenie |
|---|---|---|---|---|
| 0.9473 | 0.0352 | 0.9271 | mäkký | žiadny |
| 0.3024 | 0.3883 | 0.1824 | tvrdý | mäkký |
| 0.2395 | 0.1736 | 0.2676 | tvrdý, mäkký | tvrdý,mäkký |
| -0.1806 | -0.5457 | -0.0594 | žiadny | tvrdý,mäkký |

**Tabuľka 28:** Princíp priraďovania ošetrenia s najväčším efektom

V Tab. 28 je znázornený princíp, ktorým sa na základe výsledkov z modelu zákazníkovi

priraďuje ošetrenie, kde:

- $CATE_1$ - odhad efektu oslovenia zákazníka cez mäkký kanál,
- $CATE_2$ - odhad efektu oslovenia zákazníka cez tvrdý kanál,
- $CATE_3$ - odhad efektu oslovenia zákazníka súčasne cez mäkký aj tvrdý kanál,
- odporúčané ošetrenie - kanál oslovenia, cez ktorý zákazník dosahuje najvyšší odhadovaný efekt,
- priradené ošetrenie - kanál, cez ktorý zákazník bol skutočne oslovený.

Farebne sme zvýraznili pri ktorých stĺpcoch sa nadobudla najvyššia hodnota. Pričom ak sú všetky hodnoty CATE záporné, tak mu nepriradíme žiadne ošetrenie. Inak povedané, pre neho je odporúčané nebyť osloveným cez žiadny kanál. Túto situáciu sme znázornili v poslednom riadku.

## 6.5 Porovnanie nástrojov pri viacúrovňovom ošetrení

Pri porovnaní výsledkov týchto troch nástrojov z viacerých kritérií, akými sú rozlišovacia schopnosť modelu, výpočtová a implementačná náročnosť. Taktiež sa pozrieme na distribúciu odhadov CATE a následné priradenie odporúčaného kanálu oslovenia.

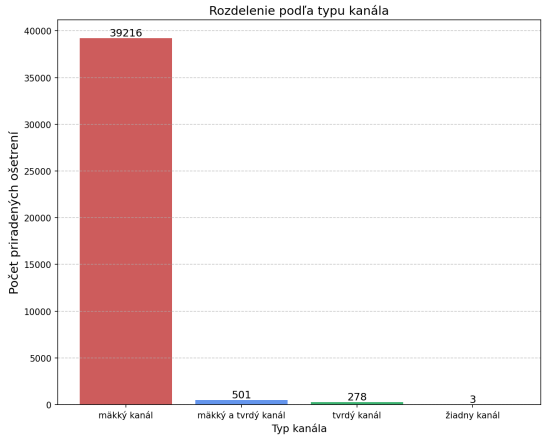
**Distribúcia odhadov CATE**

Na Obr.14, 15, 17 sú znázornené distribúcie odhadov CATE pre tri nástroje v závislosti od rôznych typov ošetrenia. Na ich základe môžeme usudzovať o nasledovnom:

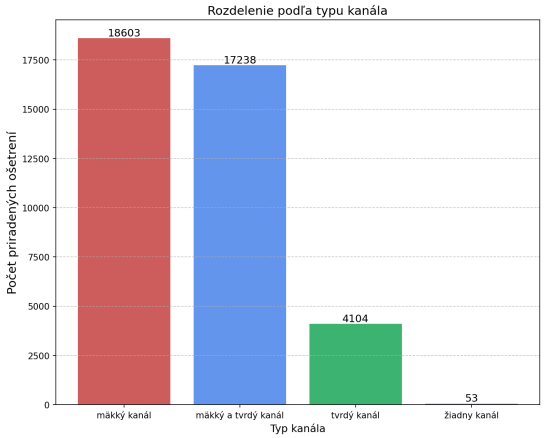
- S-learner vykazuje podobnú distribúciu CATE, zriedkavo predikuje záporný efekt z oslovenia a dosahuje nižší priemerný efekt oproti iným prístupom,
- T-learner a X-learner vykazujú výraznejšie rozdiely v distribúciách CATE pre jednotlivé kanály oslovenia a častejšie dosahuje záporné hodnoty odhadov efektu oslovenia,
- X-learner zo všetkých prístupov dosahuje najvyššie hodnoty priemerných efektov oslovenia pre jednotlivé kanály.

T-learner a X-learner sú založené na trénovaní viacerých modelov na základe rozdelenia podľa typu ošetrenia. Z praktických dôvodov, akými sú kapacity, technické obmedzenia, obmedzené náklady niektoré skupiny sú v dátach menej zastúpené, čo môže viesť k zníženiu spoľahlivosti odhadov modelov trénovaných na týchto menších skupinách. X-learner daný problém rieši vyvážením skupín na základe pravdepodobnosti, že do tej skupiny bude jedntotka zaradená. Daný problém sme pozorovali pri odhade CATE pre tvrdý kanál oslovenia, kde oba prístupy predikovali priemerný efekt ošetrenia blízky nule.

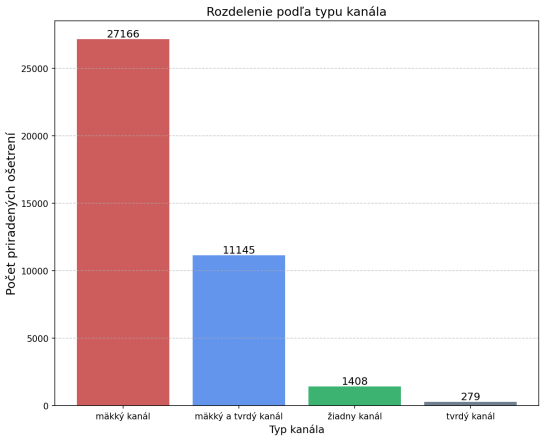
V nasledujúcej časti sa chceme pozrieť aj na to, ako často jednotlivé modely označujú konkrétny typ ošetrenia ako najviac efektívny.




**Obr. 19:** Frekvencia priradenia ošetrenia pre S-learner




**Obr. 20:** Frekvencia priradenia ošetrenia pre T-learner



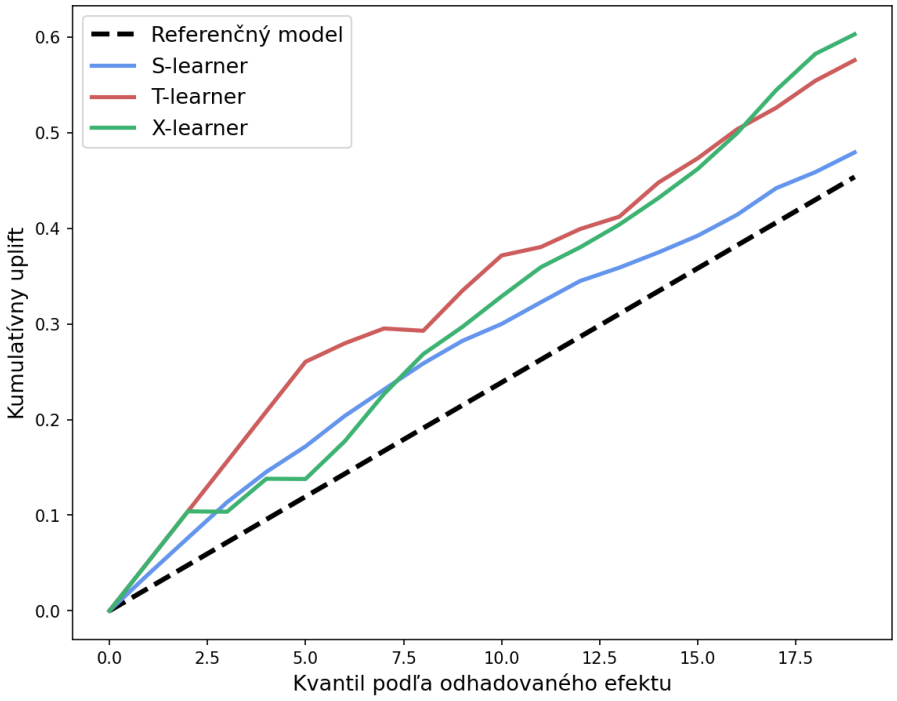

**Obr. 21:** Frekvencia priradenia ošetrenia pre X-learner

Na Obr. 19, 20, 21 prostredníctvom stĺpcového grafu sú znázornené počty priradení

modelmi konkrétny typ kanálu, teda ten, pri ktorom bola pre danú jednotku odhadnutá najvyššia hodnota CATE. S-learner priraďuje takmer všetkým mäkký kanál, zatiaľ čo T-learner a X-learner vykazujú vyváženejšie rozdelenie. Napriek tomu sa zákazníkom najčastejšie prideľuje mäkký efekt oslovenia, pre ktorý aj vo všetkých prístupoch hodnota ATE bola najvyššia.

### Rozlišovacia schopnosť pomocou AUUC

Pomocou uplift kriviek vieme kvantifikovať schopnosť modelu rozlišovať medzi pozorovaniami s vyšším a nižším očakávaným efektom z oslovenia. Postup sme uviedli v Kap.1.4.1. Pričom Tab.28 slúži ako podklad pre zostrojenie uplift kriviek a následný výpočet AUUC. Konkrétne nás budú zaujímať iba tie riadky, v ktorých sa ošetrenie priradené modelom zhoduje so skutočným.




**Obr. 22:** Uplift krivky pre tri použité prístupy

Na Obr. 22 je zobrazená vizualizácia uplift kriviek pre každý z troch prístupov. Pričom všetky tri prístupy sa nachádzajú nad referenčnou krívkou. V nasledujúcej tabuľke uvádzame hodnoty AUUC, znázorňujúce plochy pod týmito krivkami:

| Prístup | S-learner | T-learner | X-learner |
|---|---|---|---|
| AUUC | 0.2886 | 0.3549 | 0.3228 |

**Tabuľka 29:** AUUC hodnoty vzhľadom na použitý prístup pri odhade CATE s viacerými typmi ošetrenia

Pomocou Tab. 29 uvádzame hodnoty AUUC pre každý z nástrojov. Pričom vidíme, že na základe AUUC T-learner najlepšie rozlišuje medzi jednotkami s vyšším a nižším efektom ošetrenia.

### Výpočtová a implementačná náročnosť

Z hľadiska náročnosti medzi hlavné rozdiely teda patria:

- S-learner pre všetky skupiny trénuje jeden model, ktorý si vyžaduje iba pridanie príčinnej premennej do množiny čŕt, preto jeho trénovanie má najnižšiu implementačnú náročnosť.
- V prípade kategorickej príčinnej premennej, najmä, keď pozostáva z vyššieho počtu unikátnych hodnôt je potrebné zabezpečiť vhodné zakódovanie, čo zvyšuje náročnosť modelu a komplikuje interpretáciu výsledkov.
- T-learner trénuje pre všetky skupiny ošetrení samostatný model. Trénuje $n$ modelov pre $n$ typov ošetrení. Daný prístup je dobre interpretovateľný, ale je výpočtovo náročnejší oproti S-learneru.
- Pri X-learneri okrem trénovania modelov pre každú skupinu oslovenia sa v ďalšej fáze počítajú aj imputované efekty. Zároveň pomocou logistickej regresie sa odhaduje pravdepodobnosť, že konkrétna jednotka dostane konkrétne ošetrenie, čo slúži ako váha pri konečnom výpočte CATE.
- Zo všetkých troch použitých prístupov X-learner sa javí ako najkomplexnejší z hľadiska implementácie aj výpočtovej náročnosti.

Čas potrebný na trénovanie jednotlivých modelov spolu s parametrami prostredia, na ktorom boli dané operácie vykonané, sú uvedené v Prílohe B.

# Záver

V diplomovej práci sme skúmali kauzálny vplyv oslovenia zákazníka banky na jeho správanie. Konkrétne sme sa zamerali na to, aký vplyv má marketingové oslovenie na rozhodnutie zákazníka banky založiť si spotrebný úver.

Práca poskytuje úvod do problematiky z pohľadu banky a uvádza motiváciu zahrnutia príčinných vzťahov do modelovania. V práci boli taktiež priblížené základy kauzality, odvodená rovnica pre odhad efektu oslovenia zákazníka, ktorá je predstavená v rovnici (12). Takisto bol uvedený matematicky základ, na ktorom je odhad individuálnych efektov ošetrenia postavený.

Okrem samotného vplyvu oslovenia sme zohľadnili aj kanál, prostredníctvom ktorého bolo oslovenie realizované. Tento prístup poskytuje pohľad aj na efektivitu jednotlivých kanálov a umožňuje identifikovať optimálne stratégie oslovenia.

Prostredníctvom použitých nástrojov (S-learner, X-learner, T-learner) sme odhadli efekt oslovenia na správanie zákazníka, ktorý bol vyjadrený ako rozdiel pravdepodobností nastania modelovanej udalosti za podmienky, že zákazník bol oslovený a nebol oslovený. Pričom sme uvažovali tri prístupy výberu čŕt. Z našich výsledkov vieme usúdiť, že oslovenie vo väčšine prípadov malo pozitívny efekt. **Taktiež na odlíšenie efektu oslovenia pri veľkom počte čŕt sa ukazuje ako rozumný postup rozdielu čŕt korelujúcich so závislou premennou a príčinnou.** Daný prístup nám umožňuje izolovať samotný efekt oslovenia, ktorý je determinovaný na základe iných dostupných čŕt.

Následne sme dané nástroje aplikovali na viacúrovňové ošetrenie, čo v našom prípade predstavovalo kanál oslovenia, ktoré sme na základe jeho nákladovosti klasifikovali do dvoch skupín, ako uvádza Tab.8. Pri tomto prístupe cieľom bolo priradiť kanál, pri ktorom zákazník dosahuje najvyšší efekt z oslovenia. Pri danom prístupe sme pozorovali viac frekventovaný výskyt záporných hodnôt pri odhadoch efektu oslovenia konkrétnymi kanálmi. To nám umožňuje identifikovať, ktoré kanály majú skôr negatívny efekt pre konkrétneho zákazníka. Daný prístup nám ukazuje dôležitosť cielenia na zákazníkov cez správny kanál a navyše nám pomáha pochopiť, akým spôsobom by mali byť oslovení, aby dané oslovenie malo čonajvyšší efekt. **Na základe našich výsledkov najviac efektívnym je mäkký kanál oslovenia, čo potvrdzuje aj výstupy z už realizovaných spoločnosťou iných analýz korelácie digitálnej aktivity zákazníka a jeho prefe-**

**rencie mäkkého kanálu oslovenia.**

Výkonnosť použitých prístupov sme vyhodnotili na základe ich náročnosti, distribúcie odhadov a schopnosti rozlišovať medzi zákazníkmi s vyšším a nižším efektom oslovenia pomocou AUUC. **Z hľadiska schopnosti modelu rozlišovať medzi jednotkami s vyšším a nižším prínosom ošetrenia, X-learner dosahuje najlepšie výsledky pri dvoch typoch ošetrenia. Pri viacerých typoch ošetrenia vykazuje najvyššiu rozlišovaciu schopnosť T-learner.**

Prínosom je návrh nových stratégií, ktoré pomôžu pri planovaní a realizácii marketingových aktivít banke naplniť finančné ciele a zároveň ostať relevantní z pohľadu komunikáčneho kanálu a personalizovanej ponuky. Ako jedno z rozšírení by bolo možné do modelovania zahrnúť aj náklady spojené s oslovením cez konkrétny kanál a očakávaný zisk z predaja, čo vedie k efektívnejšiemu využitiu zdrojov pre marketingové aktivity. Takémuto prístupu sa venuje aj [29]. Aktuálny výstup našej práce tvorí súčasť novej obchodnej stratégie pre priamy marketing banky.

V rámci modelovania sme obmedzili analýzu iba na tie kampane, ktoré sú zamerané na spotrebný úver a neanalyzovali sme vplyv iných oslovení. To by však mohlo prispieť k hlbšiemu pochopeniu vzťahov medzi jednotlivými kampaňami a zároveň k tomu, ako oslovenie zamerané na jeden konkrétny produkt môže ovplyvniť záujem zákazníka o iné služby banky, čo predstavuje ďalšie rozšírenie tejto práce.

Rozšírením analýzy o komunikačné kanály, prostredníctvom ktorých bolo oslovenie realizované, modelovanie sa stáva komplexnejším a implementačne náročnejším. Napriek tomu, výsledky daných modelov tvoria základ pre podrobnejšiu analýzu správania zákazníkov a identifikáciu nových prístupov plánovania marketingových stratégií. Konkrétne, vieme analýzovať preferovaný typ oslovenia zákazníka alebo identifikovať tých na koho oslovenia nemá vplyv.

Záverom možno konštatovať, že zarhnutie príčinných premenných má zásadný význam pre optimalizáciu marketingových stratégií. Výsledky danej diplomovej práce sú podkladom pre ďalšie analýzy výkonnsostí jednotlivých kanálov, vytváranie cieľových skupín pre oslovenie, detekcia čŕt, ktoré majú vplyv na efekt oslovenia, čím prispievame k novým prístupom tvorby marketingových kampaní.

# Zoznam použitej literatúry


[1] Acharki, N., Lugo, R., Bertoncello, A., Garnier, J.: *Comparison of Meta-Learners for Estimating Multi-Valued Treatment Heterogeneous Effects*, In Proceedings of the International Conference on Machine Learning (ICML), PMLR, 2023, dostupné na internete (18.04.2025): https://proceedings.mlr.press/v202/acharki23a/acharki23a.pdf

[2] Ali, S., Rana, Z. A.: *Evaluating Performance of Software Defect Prediction Models Using Area Under Precision-Recall Curve (AUC-PR)*, In Proceedings of the 2019 2nd International Conference on Advancements in Computational Sciences (ICACS), IEEE, 2019, dostupné na internete (23.04.2025): https://www.researchgate.net/publication/332429410_Evaluating_Performance_of_Software_Defect_Prediction_Models_Using_Area_Under_Precision-Recall_Curve_AUC-PR

[3] Battiti, R.: *Using Mutual Information for Selecting Features in Supervised Neural Net Learning*, IEEE Transactions on Neural Networks, vol. 5, no. 4, pp. 537–550, IEEE, 1994, dostupné na internete (5.04.2025): https://doi.org/10.1109/72.298224

[4] Betlei, A., Diemert, E., Amini, M.-R.: *Treatment Targeting by AUUC Maximization with Generalization Guarantees*, Criteo AI Lab, Grenoble, 2020, dostupné na internete (08.05.2025): https://arxiv.org/pdf/2012.09897

[5] Bentéjac, C., Csörgő, A., Martínez-Muñoz, G.: *A Comparative Analysis of Gradient Boosting Algorithms*, Artificial Intelligence Review, Springer, 2021, dostupné na internete (5.04.2025): https://doi.org/10.1007/s10462-020-09896-5

[6] Caron, A., Baio, G., Manolopoulou, I.: *Estimating Individual Treatment Effects using Non-Parametric Regression Models: a Review*, Journal of the Royal Statistical Society Series A: Statistics in Society, 2022, dostupné na internete (10.04.2024): https://doi.org/10.1111/rssa.12824

[7] Friedman, J. H.: *Greedy Function Approximation: A Gradient Boosting Machine*, Annals of Statistics, JSTOR, 2001, dostupné na internete (5.04.2025): https://projecteuclid.org/journals/annals-of-statistics/volume-29/

issue-5/Greedy-function-approximation-A-gradient-boosting-machine/10.1214/aos/1013203451.full

[8] Hancock, J. T., Khoshgoftaar, T. M.: *CatBoost for Big Data: An Interdisciplinary Review*, Journal of Big Data, vol. 7, article no. 94, 2020, dostupné na internete (12.04.2025): https://doi.org/10.1186/s40537-020-00369-8

[9] Hardy, M. A.: *Regression with Dummy Variables*, Sage, 1993

[10] Harman, R.: *Multivariate Statistical Analysis*, Selected Lecture Notes, FMFI UK, 2023, dostupné na internete (09.04.2024): http://www.iam.fmph.uniba.sk/ospm/Harman/VSAp.pdf

[11] Hastie, T., Tibshirani, R., Friedman, J.: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Second Edition, Springer, 2009

[12] Huang, J., Ling, C. X.: *Using AUC and Accuracy in Evaluating Learning Algorithms*, IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 3, IEEE, 2005, dostupné na internete (08.05.2025): https://ieeexplore.ieee.org/document/1388242

[13] James, G., Witten, D., Hastie, T., Tibshirani, R.: *An Introduction to Statistical Learning with Applications in Python*, Springer, 2013, dostupné na internete (5.04.2025): https://link.springer.com/content/pdf/10.1007/978-3-031-38747-0.pdf

[14] Kane, K., Lo, V.S., Zheng, J.: *Mining for the truly responsive customers and prospects using true-lift modeling: comparison of new and existing methods*, Journal of Marketing Analytics, Springer, 82(4), 218–238, 2014, dostupné na internete (02.01.2025): https://link.springer.com/article/10.1057/jma.2014.18

[15] Katz, G., Shabtai, A., Rokach, L., Ofek, N.: *ConfDtree: A Statistical Method for Improving Decision Trees*, Journal of Computer Science and Technology, vol. 29, pp. 392–407, Springer, 2014, dostupné na internete (28.04.2025): https://link.springer.com/article/10.1007/s11390-014-1438-5

[16] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y.: *LightGBM: A Highly Efficient Gradient Boosting Decision Tree*, Advances

in Neural Information Processing Systems, vol. 30, 2017, dostupné na internete (5.04.2025): https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

[17] Künzel, S. R., Sekhon, J. S., Bickel, P. J., Yu, B.: *Meta-learners for Estimating Heterogeneous Treatment Effects using Machine Learning*, arXiv preprint, 2019, dostupné na internete (11.01.2025): https://arxiv.org/pdf/1706.03461

[18] Letoffé, O., Huang, X., Asher, N., Marques-Silva, J.: *From SHAP Scores to Feature Importance Scores*, IRIT, CNRS, ICREA, 2023, dostupné na internete (10.04.2025): https://www.researchgate.net/publication/380730227_From_SHAP_Scores_to_Feature_Importance_Scores

[19] Lundberg, S. M.: *SHAP Documentation*, dostupné na internete (17.04.2025): https://shap.readthedocs.io/en/latest/

[20] Nie, X., Wager, S.: *Quasi-Oracle Estimation of Heterogeneous Treatment Effects*, Biometrika, vol. 108, no. 2, pp. 299–319, Oxford University Press, 2021, dostupné na internete (17.04.2025): https://doi.org/10.1093/biomet/asaa076

[21] Rosenbaum, P. R., Rubin, D. B.: *The Central Role of the Propensity Score in Observational Studies for Causal Effects*, Biometrika, vol. 70, no. 1, pp. 41–55, Oxford University Press, 1983, dostupné na internete (20.04.2025): https://academic.oup.com/biomet/article/70/1/41/240879

[22] Shafer, J., Agrawal, R., Mehta, M.: *SPRINT: A Scalable Parallel Classifier for Data Mining*, In Proceedings of the 1996 International Conference on Very Large Data Bases (VLDB), Citeseer, 1996, dostupné na internete (18.04.2025): https://www.vldb.org/conf/1996/P544.PDF

[23] Sipser, M.: *Introduction to the Theory of Computation*, 3rd edition, Cengage Learning, 2013

[24] Scikit-learn developers: *mutual_info_classif*, Scikit-learn Documentation, dostupné na internete (10.04.2025): https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.mutual_info_classif.html

[25] Scikit-learn developers: *LogisticRegression*, Scikit-learn Documentation, dostupné na internete (29.04.2025): https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

[26] Udilă, A.-I.: *Encoding Methods for Categorical Data: A Comparative Analysis for Linear Models, Decision Trees, and Support Vector Machines*, Bakalárska práca, EEMCS, Delft University of Technology, 2023

[27] Wei, J.-T., Lin, S.-Y., Wu, H.-H.: *A Review of the Application of RFM Model*, African Journal of Business Management, vol. 4, no. 19, pp. 4199, Academic Journals, 2010, dostupné na internete (08.05.2025): https://www.researchgate.net/publication/228399859_A_review_of_the_application_of_RFM_model

[28] Yao, L., Chu, Z., Li, S., Li, Y., Gao, J., Zhang, A.: *A Survey on Causal Inference*, ACM Transactions on Knowledge Discovery from Data, 15(5), Article 74, Association for Computing Machinery, 2021, dostupné na internete (29.12.2024): https://dl.acm.org/doi/pdf/10.1145/3444944

[29] Zhao, Z., Harinen, T.: *Uplift Modeling for Multiple Treatments with Cost Optimization*, IEEE International Conference on Data Science and Advanced Analytics (DSAA), 2019, dostupné na internete (29.12.2024): https://arxiv.org/pdf/1908.05372

# Príloha A. Detaily použitých metód

## Softmax kódovanie

Softmax kódovanie nevyžaduje výber referenčnej triedy. V tomto prístupe sa každá z $m$ tried modeluje symetricky [13]:

$$\Pr(Y = j \mid X = x) = \frac{\exp(\beta_{j0} + \beta_{j1}x_1 + \cdots + \beta_{jp}x_p)}{\sum_{l=1}^{m} \exp(\beta_{l0} + \beta_{l1}x_1 + \cdots + \beta_{lp}x_p)}, \tag{66}$$

pričom koeficienty sa odhadujú pre všetky triedy $j = 1, \ldots, m$.

V jazyku Python sa na odhad multinomiálnej logistickej regresie používa knižnica `scikit-learn`, konkrétne `LogisticRegression`. Pre viacnásobnú klasifikáciu je potrebné nastaviť parameter `multi_class='multinomial'`, pri ktorej sa používa softmax kódovanie všetkých tried. Pričom jej použitie vyzerá nasledovne [25]:

```
from sklearn.linear_model import LogisticRegression
model = LogisticRegression(
    multi_class='multinomial',
    solver='lbfgs'
)
```

## Vzájomná informácia

Vzájomnú informáciu môžeme matematicky definovať pomocou entropie. Entropia $H(C)$ vyjadruje mieru neurčitosti (alebo informácie) spojenej s náhodnou črtou $C$, zatiaľ čo podmienená entropia $H(C|F)$ predstavuje očakávanú neurčitosť $C$, za podmienky, že poznáme hodnotu črty $F$. Na základe toho definujeme vzájomnú informáciu ako [3]:

$$I(C; F) = H(C) - H(C|F). \tag{67}$$

Táto rovnica vyjadruje množstvo informácie, ktorú nám premenná $F$ poskytuje o črte $C$. Formálne zapíšeme:

$$I(C; F) = \sum_{c,f} P(c, f) \log \left( \frac{P(c, f)}{P(c)P(f)} \right) \tag{68}$$

kde $P(c, f)$ je spoločné rozdelenie diskrétnych náhodných premenných $C$ a $F$, zatiaľ čo $P(c)$ a $P(f)$ predstavujú ich marginálne pravdepodobnostné rozdelenia.

Pričom na výpočet vzájomnej informácie sa v prostredí `Python` využíva funkcia `mutual_info_classif` z knižnice `sklearn.feature_selection`. Viac informácií o implementácii a parametroch tejto funkcie je dostupných v oficiálnej dokumentácii knižnice `scikit-learn` [24].

# SHAP Skóre

Uvažujme klasifikačný model definovaný množinou čŕt $\mathcal{F} = \{1, \ldots, m\}$, množinou tried $\mathcal{K} = \{c_1, c_2, \ldots, c_K\}$ a klasifikačnou funkciou $\kappa : \mathcal{F} \to \mathcal{K}$, ktorá každému vektoru čŕt $x = (x_1, \ldots, x_m)$ priraďuje triedu. Potom hodnoty SHAP skóre vychádzajú z charakteristickej funkcie $v_e : 2^{\mathcal{F}} \to \mathbb{R}$, ktorá je definovaná nasledovne:

$$v_e(S; \mathcal{E}) := \mathbb{E}[\kappa \mid x_S = v_S], \tag{69}$$

kde:

- $S \subseteq \mathcal{F}$ predstavuje podmnožinu vstupných čŕt z celej množiny $\mathcal{F}$,
- $x_S = v_S$ znamená, že hodnota čŕt v množine $S$ je fixovaná na konkrétne pozorovanie $v_S$,
- $\kappa$ je výstup klasifikátora alebo predikčného modelu,
- $v_e(S; \mathcal{E})$ vyjadruje očakávanú hodnotu výstupu modelu $\mathcal{E}$ za predpokladu, že poznáme len črty z množiny $S$.

Pričom $2^{\mathcal{F}}$ označuje všetky možné podmnožiny množiny čŕt $\mathcal{F}$ (tzv. potenčná množina) a funkcia $v_e$ každú takúto podmnožinu mapuje na reálne číslo, ktoré predstavuje očakávaný výstup modelu. Pre jednoduchosť značení definujeme marginálny prínos črty $i$:

$$\Delta_i(S; \mathcal{E}) := v_e(S; \mathcal{E}) - v_e(S \setminus \{i\}; \mathcal{E}) \tag{70}$$

t.j. rozdiel v očakávanej predikcii modelu, keď z množiny $S$ odstránime črtu $i$.

Pre úplnosť výpočtu SHAP skóre potrebujeme ešte zadefinovať váhový faktor, aby sme pri výpočte SHAP skóre zohľadnili všetky možné poradia, v ktorých môže byť črta pridaná do modelu:

$$\varsigma(|S|) := \frac{1}{|\mathcal{F}|} \cdot \binom{|\mathcal{F}| - 1}{|S| - 1}^{-1}, \tag{71}$$

kde $|\mathcal{F}|$ je počet všetkých čŕt a $|S|$ počet čŕt v množine $S$.

Nakoniec, SHAP skóre $\mathrm{Sc}_{\mathcal{E}} : \mathcal{F} \to \mathbb{R}$, pre črtu $i$ vzhľadom na model $\mathcal{E}$ definujeme ako [18]:

$$\mathrm{Sc}_{\mathcal{E}}(i) := \sum_{\substack{S \subseteq \mathcal{F} \\ i \in S}} \varsigma(|S|) \cdot \Delta_i(S; \mathcal{E}). \tag{72}$$

Týmto spôsobom SHAP skóre kvantifikuje očakávaný vplyv každej črty $i$ k výstupu modelu vzhľadom na všetky možné kombinácie ostatných čŕt. SHAP skóre tak určuje, do akej miery črta $i$ ovplyvňuje predikciu modelu, pričom vysledné hodnoty vieme interpretovať takto:

- kladná alebo záporná hodnota SHAP skóre naznačuje, že daná črta prispieva k zmene predikcie modelu,
- hodnota blízka nule naznačuje, že črta nemá na výstup modelu významný vplyv.

V prostredí `Python` je ich výpočet podporovaný knižnicou `shap`, pričom sa využíva predtrénovaný model (napr. `LightGBM`) a objekt `shap.Explainer`, ktorý zabezpečí výpočet SHAP hodnôt pre daný model. Detailnejší popis použitia tejto knižnice je uvedený v [19].

# Trénovanie LightGBM

LightGBM s cieľom zlepšenia výkonnosti vie byť trénovaný pomocou optimalizácie hyperparametrov. Na trénovanie modelu využívame navrhnutý postup spoločnosťou ČSOB. Pričom použijeme knižniu *Optuna*.

### Zavedenie hyperparametrov

V danej časti uvedieme použité hyperparametre a ich rozsahy, ktoré sa neskôr aplikujú na model:

- parameter, ktorý ovplyvňuje veľkosť krokov pri optimalizácii modelu. (*learning_rate*) – 0,01 až 0,3;
- maximálna hĺbka stromu (*max_depth*) – 3 až 10;
- počet listov v strome (*num_leaves*) – 20 až 150;
- počet stromov (*n_estimators*) – 100 až 1000;
- minimálny počet vzoriek v liste (*min_child_samples*) – 10 až 100;
- podiel náhodne vybraných vzoriek na tréning každého stromu (*subsample*) – 0,4 až 0,8;
- podiel náhodne vybraných čŕt pre každý strom (*colsample_bytree*) – 0,4 až 0,8;
- minimálna hodnota potrebná na rozdelenie uzla (*min_split_gain*) – 0 až 15;
- regularizačný parameter, ktorý slúži na penalizáciu absolútnych hodnôt koeficientov, čím môže viesť k úplnému vynulovaniu nepodstatných čŕt;(*reg_alpha*) – $0,1$ až 20;
- regularizačný parameter, ktorý slúži na penalizáciu veľkých váh čŕt a tým predchádza pretrénovaniu (*reg_lambda*) – $0,1$ až 50;

**Použitie krížovej validácie**

Na výber vhodných hyperparametrov používame metódu $k$-násobnej krížovej validácie, ktorá spočíva v tom, že množina pozorovaní sa náhodne rozdelí do $k$ približne rovnako veľkých, disjunktných podmnožín.(*folds*). Metóda prebieha tak, že v každej iterácii jedná z podmnožín sa zvolí ako validačná množina, zatiaľ čo zostávajúcich $k-1$ podmnožín slúžia na trénovanie modelu. Po natrénovaní modelu sa na validačnej množine vypočíta chyba predikcie. Tento postup sa opakuje $k$-krát, pričom v každom kroku za každým sa použije iná validačná množina, čím vznikne $k$ samostatných odhadov chyby. Finálny odhad testovacej chyby sa potom vypočíta ako priemer týchto hodnôt [13].

Pri použití daného prístupu spoločnosť ČSOB sa zameriava na nasledujúce metriky:

- PR-AUC - vyrátaná na testovacej množine v každej iterácii $k$-násobnej krížovej validácie,

- ROC-AUC - vyrátaná na testovacej množine v každej iterácii $k$-násobnej krížovej validácie,
- ROC-AUC - vyrátaná na trénovacej množine v každej iterácii $k$-násobnej krížovej validácie.

Úlohou je maximalizovať hodnotu nasledujúcej účelovej funkcie:

$$F_{\text{opt}} = \overline{\text{PR-AUC}}_{\text{test}} \cdot \max\left(0,999484^{|f_sel_n_features|};\ 0,1\right) \cdot \min\left(1,\ \frac{\overline{\text{ROC-AUC}}_{\text{test}}}{\overline{\text{ROC-AUC}}_{\text{train}}}\right), \quad (73)$$

kde:

- $|f_sel_n_features|$ – počet vybraných čŕt použitých v danej iterácii algoritmu
- $\overline{\text{PR-AUC}}_{\text{test}}$ – priemerná hodnota PR-AUC vyrátaná na testovacej množine pri $k$-násobnej krížovej validácie,
- $\overline{\text{ROC-AUC}}_{\text{test}}$ – priemerná hodnota ROC-AUC vyrátaná na testovacej množine pri $k$-násobnej krížovej validácie,
- $\overline{\text{ROC-AUC}}_{\text{train}}$ – priemerná hodnota ROC-AUC vyrátaná na trénovacej množine pri $k$-násobnej krížovej validácie na trénovacích množinách.

V každej iterácii teda *Optuna* navrhne kombináciu hyperparametrov, ktorá sa otestuje pomocou $k$-násobnej krížovej validácie. Pričom *Optuna* po vykonaní všetkých iterácií v rámci krížovej validácie spočíta hodnotu účelovej funkcie a tieto výsledky využije na rozhodnutia v ďalších iteráciach. Na konci *Optuna* vyberie tú kombináciu hyperparametrov, ktorá maximalizuje hodnotu účelovej funkcie. Voľba hodnoty $k$ je odvodená z počtu mesiacov, použitých pri modelovaní. V každej z $k$ iterácii algoritmus *Optuna* vyberie hodnoty hyperparametrov, na ktorých sa na trénovacej množine natrénuje model, a na testovacej množine na základe zvolených metrík sa vyhodnotí jeho výkonnosť.

# Bootstrap algoritmus pre odhad štandartnej odchýlky

Bootstrap je proces, pri ktorom sa z trenovacích dát náhodne s návratom vyberie zvolený počet dátových vzoriek. Potom algortimus výpočtu je nasledovný [11]:

Uvažujme trénovaciu vzorku dát $X = (x_1, x_2, \ldots, x_n)$ a vopred nastavený počet opakovaní $B$.

1. Vytvoríme $B$ nezávislých bootstrap vzoriek $x_1^*, \ldots, x_B^*$, získaných výberom s návratom z trénovacej vzorky. Pričom každá vzorka má rovnakú veľkosť ako pôvodná trénovacia vzorka.

2. Pre každú vytvorenú vzorku $b = 1, \ldots, B$ sa vypočíta sledovaná veličina $S(x_b^*)$, v našom prípade je to štandardná odchýlka.

3. Na základe vypočítaných hodnôt odhadneme štandardnú odchýlku ako výberovú štandardnú odchýlku:

$$\widehat{\mathrm{STD}} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(S(x_b^*) - \bar{S}^*\right)^2}, \tag{74}$$

kde:

$$\bar{S}^* = \frac{1}{B}\sum_{b=1}^{B} S(x_b^*). \tag{75}$$

# Príloha B. Doplňujúce technické údaje

## Popis čŕt z webovej aktivity

Tabuľka čŕt vytvorená na základe záznamov o aktivite používateľa na webovom rozhraní, okrem základných identifikačných údajov, ako sú identifikátor zákazníka a prísluchajúceho mesiaca, obsahuje aj nasledovné informácie:

- **Črty, súvisiace s návštevou**: priemerne trvanie návštevy, počet navštívených stránok.
- **Zdroje presmerovania**: na akej stránke sa návštevník nachádzal predtým ako sa presmeroval na stránku ČSOB (napr. vyhľadávače, sociálne siete, push notifikácie, e-mail, smart banking, médiá, noviny). Dané črty hovoria koľkokrát cez aký zdroj sa návštevník presmeroval na stránku banky.
- **Typ zariadenia**: najviac frekventované používané zariadenie (Apple, Android, iné).
- **Časové charakteristiky**: najviac frekventované denné obdobie (ráno, popoludní, večer, noc), či aktivity počas víkendov alebo pracovných dní, čas od poslednej návštevy. Počty presahovaní stanovenej maximálnej dĺžky návštevy (2 hodiny) a počty dosahovaní minimálnej dĺžky návštevy (30 sekúnd).
- **Preferencie a obsahové zameranie**: počty návštev stránok vzťahujúcich sa k istej tematike (poistenie, bankový účet, informačné stránky, osobný účet).
- **Zmeny v správaní**: pozorovania zmien v správaní zákazníka v rôznych časových obdobiach: mesačne, kvartálne, ročne.

Pričom medzi zmeny v správaní patria také črty ako zmena v počte navštívených stránok, zmeny v trvaní návštvevy alebo či došlo k zmene preferovanéhorovnaého zariadenia.

## Porovnanie časov trénovania

Pre hodnotenie efektivity výpočtových operácii porebujeme poznať vlastnosti výpočtového prostredia.

| Parameter | Počet CPU jadier | Veľkosť RAM | Architektúra |
|---|---|---|---|
| Hodnota | 8 | 32.93 GB | x86_64 |

**Tabuľka 30:** Parametre výpočtového prostredia použitého pri trénovaní modelov

Na Tab. 30 sú uvedené technické parametre prostredia, na ktorom boli vykonané výpočty, kde:

- Počet CPU jadier vyjadruje počet dostupných jadier pre výpočet.
- Veľkosť RAM vyjadruje maximálne dostupnú operačnú pamäť.
- Architektúra x86_64 označuje 64-bitový procesor.

Všetky modely boli trénovane na systéme s danými parametrami.

| Model | Dva typy ošetrení | Viacúrovňové ošetrenie |
|---|---|---|
| S-learner | 60 minút | 60 minút |
| T-learner | 120 minút | 150 minút |
| X-learner | 122 minút | 154 minút |

**Tabuľka 31:** Približný čas trénovania modelov

Na Tab. 31 sú uvedené časy v prípade dvoch typov ošetrenia (bol oslovený, nebol oslovený) a viacúrovňového, kde sme uvažovali oslovenie cez rôzne typy kanálov. Pričom čas trénovania X-Learnera obsahuje aj čas na odhad pravdepodobností pomocou logistickej regresie.