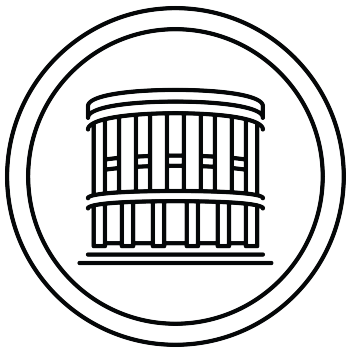
UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY



Kvantifikácia vplyvu časového rozpätia predikcie a udalosti na kvalitu predikcie

DIPLOMOVÁ PRÁCA

2025 Bc. Filip Blaho

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

Kvantifikácia vplyvu časového rozpätia predikcie a udalosti na kvalitu predikcie

DIPLOMOVÁ PRÁCA

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

Bratislava 2025

Bc. Filip Blaho





Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta:	Bc. Filip Blaho
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: Kvantifikácia vplyvu časového rozpätia predikcie a udalosti na kvalitu predikcie
Quantification of the impact of delay between prediction and event on prediction quality

Anotácia: Súčasťou nastavovania marketingovej komunikácie je aj určenie vhodného časového rozpätia medzi časom predikcie a časom realizácie predikovanej udalosti. So zväčšujúcim sa časovým rozpätím sa náročnosť problému zvyšuje. Miera degradácie kvality predikcie významne ovplyvňuje spôsob nastavenia pravidiel pre oslovovanie klientov. V práci sa budeme venovať určovaním veľkosti časovej degradácie pre rôzne typy udalostí ako napríklad nákup produktu, odchod klienta z banky a pod.

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

Poďakovanie: Rád by som touto cestou vyjadril svoje poďakovanie všetkým, ktorí ma podporovali pri písaní tejto diplomovej práce. Moje poďakovanie patrí predovšetkým môjmu vedúcemu práce Mgr. Martinovi Hurbanovi, PhD. za odborné vedenie a cenné rady. Ďalej by som chcel poďakovať spoločnosti ČSOB za poskytnutie možnosti spolupracovať. V neposlednom rade som vďačný svojej rodine za ich podporu počas celého obdobia môjho štúdia.

Abstrakt

BLAHO Filip: Kvantifikácia vplyvu časového rozpätia predikcie a udalosti na kvalitu predikcie [Diplomová práca], Univerzita Komenského v Bratislave, Fakulta matematiky, fyziky a informatiky, Katedra aplikovanej matematiky a štatistiky; vedúci práce: Mgr. Martin Hurban, PhD., Bratislava, 2025, 79 s.

V marketingovej komunikácii s klientmi je nevyhnutné predvídať ich potreby a záujmy ešte skôr, ako ich sami oznámia banke. Pri tvorbe predikčného modelu zohráva kľúčovú úlohu správne nastavenie závislej premennej, ktorá musí zahŕňať primerane dlhý časový interval. Tento interval umožňuje banke efektívne reagovať na možné zámery klienta, či už prostredníctvom adresnej ponuky produktov alebo opatrení zameraných na prevenciu odchodu zákazníka. Hlavnou výzvou je nájsť rovnováhu medzi kvalitou predikcie a horizontom predikcie, aby lepšie zohľadňoval potreby a umožnil vhodnú stratégiu oslovenia klienta. Čiastočná degradácia kvality predikcie, ktorá môže vzniknúť pri použití dlhšieho časového intervalu, môže byť vyvážená schopnosťou skoršieho reagovania na zmeny v správaní klientov. Horizont predikcie nie je jediný časový faktor ovplyvňujúci kvalitu predikčných modelov. Identifikácia vhodného horizontu predikcie by marketingovým oddeleniam umožnila efektívnejšie prispôsobovať svoje stratégie, čo by viedlo k optimalizácii interakcie s klientmi. Je taktiež potrebné zohľadniť aj ďalšie časové faktory ovplyvňujúce kvalitu modelov, ktoré zároveň vplývajú na nákladovú stránku spoločnosti, ako sú náklady na ľudský kapitál, dátové úložiská či výpočtové zdroje.

Kľúčové slová: závislá premenná, vektor, črty, model, metóda, LightGBM, pozitívna vzorka, odchodovosť, klesanie, rozpätie, predikčné okno, horizont predikcie, čas, história udalostí

Abstract

BLAHO Filip: Quantification of the impact of delay between prediction and event on prediction quality [Diploma thesis], Comenius University in Bratislava, Faculty of Mathematics, Physics and Informatics, Department of Applied Mathematics and Statistics; supervisor: Mgr. Martin Hurban, PhD., Bratislava, 2025, 79 pages

In marketing communication with clients, it is essential to anticipate their needs and interests before they explicitly express them to the bank. When developing a predictive model, the proper definition of the dependent variable plays a crucial role, and it must encompass a sufficiently long time interval. This interval enables the bank to respond effectively to potential client intentions, either through targeted product offers or measures aimed at customer retention. The main challenge lies in striking a balance between prediction quality and the prediction horizon, in order to better reflect client needs and enable an appropriate engagement strategy. A partial decline in prediction quality, which may occur with the use of a longer time interval, can be offset by the ability to react earlier to changes in client behavior. The prediction horizon is not the only temporal factor affecting the performance of predictive models. Identifying an appropriate prediction horizon would allow marketing departments to adapt their strategies more effectively, leading to optimized client interaction. It is also necessary to consider additional time-related factors influencing model quality, which simultaneously impact the company's cost structure, such as expenses related to human capital, data storage, and computational resources.

Keywords: dependent variable, vector, features, model, method, LightGBM, positive sample, churn, decline, range, prediction window, prediction horizon, time, target history

Obsah

Obsah

Úvod **8**

1 Úloha času v predikčných modeloch **10**

2 Využívané metódy strojového učenia **13**

2.1 Gradient boosting metóda 13

2.2 Metóda LightGBM 16

2.2.1 Vylepšenia v LightGBM 17

2.3 Metriky sledujúce kvalitu predikčného modelu 19

2.4 Logistická regresia 23

2.4.1 Použitie logistickej regresie v kontexte práce 24

3 Postup tvorby dátovej sady pre predikčný model **26**

3.1 Črty klientov 26

3.1.1 Embedding 28

3.1.2 FastText: Embeddingový model 28

3.2 Časové rozpätie modelu - tvorba závislej premennej 29

3.2.1 Kontinuálne zväčšujúce sa rozpätie predikcie 32

3.2.2 Konštantne veľké časové rozpätie predikcie 35

3.2.3 Porovnanie časového rozpätia predikčných modelov 42

3.3 Doplňenie dátovej sady o negatívnu vzorku 44

4 Konštrukcia predikčného modelu **46**

4.1 Selekcia čŕt klientov 46

4.2 Rozdelenie dátovej sady 47

4.3 Zakódovanie kategorických čŕt klientov 48

4.4 Optimalizácia hyperparametrov metódy LightGBM 48

4.4.1 Zoznam optimalizovaných hyperparametrov metódy LightGBM . . 48

4.4.2 Použitie k-násobnej krížovej validácie 49

4.4.3 Účelová funkcia optimalizácie 50

5 Klesanie kvality predikčných modelov **52**

5.1 Merania na 36-mesačnej histórii udalostí 52

5.2 Merania na 24-mesačnej histórii udalostí 58

5.3 Merania na 12-mesačnej histórii udalostí 63

5.4 Syntéza výsledkov klesania kvality modelov 66

Záver **73**

Zoznam použitej literatúry **75**

Prílohy **78**

A Vzájomná informácia 78

B Rozdiel medzi mriežkovým vyhľadávaním a Optuna 78

Úvod

Spoločnosti pri nastavovaní stratégie komunikácie s klientmi musia efektívne využívať svoje zdroje. Z tohto dôvodu je potrebné nastaviť predikcie tak, aby dosahovali čo najväčšiu presnosť predpovede správania sa klientov s dostatočne veľkým časovým horizontom. Napríklad, keď ide o odchodovosť klientov, môže byť výhodné, ak spoločnosť dokáže takéto správanie odhaliť skôr. Potom sa môže nastaviť správny spôsob oslovenia klienta a predísť tejto udalosti. Podobnému problému spojenému s predikciou čelia aj v iných odvetviach.

V zdravotníckom prostredí sa lekári stretávajú aj so zraneniami, ktoré vzniknú až po hospitalizácii pacienta. Príkladom takéhoto zranenia môže byť akútne poškodenie obličiek. Presnosť predikcie, či u pacienta dôjde k spomínanému zraneniu, skúmajú autori v štúdii [1]. Autori na porovnanie kvality modelov používajú metriku AUC–ROC. Jedným zo skúmaných vplyvov na presnosť predikcie bola práve dĺžka časového horizontu predikcie a jej dopady na kvalitu. Z tejto štúdie vyplýva, že predikcie vytvorené na dlhšie časové okno medzi tvorbou predikcie a nastaním zranenia (vstupné dáta predikcie pochádzali z časového horizontu: deň hospitalizácie až $k - dní$ pred nastaním poškodenia obličky, hodnota k sa pohybovala od 1 až 5) dosahovali nepresnejšiu klasifikáciu pacienta.

Metódy strojového učenia sa objavujú aj v oblasti investovania. V štúdii [2] sa autori zaoberali krátkodobou predikciou pohybu ceny bitcoinu pre horizonty: 1 minúta, 5 minút, 15 minút a 60 minút. V tejto štúdii sa dostávame k opačnému pozorovaniu. S rastúcim rozpätím časového horizontu sa predikcie stávali jemne presnejšími, vzhľadom na sledovanú metriku presnosti (anglický výraz: *accuracy*). Taktiež prediktory, ktoré sa priamo neodvíjali od ceny aktíva, dosahovali vyššiu významnosť s rastúcim predikčným horizontom.

Na začiatku diplomovej práce v kapitole č. 1 sa venujeme predstaveniu a objasneniu skúmanej problematiky s cieľom poskytnúť prehľad o jej kontexte.

Následne v kapitole č. 2 prechádzame k opisu metódy strojového učenia, ktorá je v práci využívaná. V tejto časti sú podrobne rozobrané aj jednotlivé metriky slúžiace na hodnotenie kvality binárnych klasifikačných modelov, pričom sa zameriavame aj na metodiku sledovania závislosti medzi horizontom predikcie a kvalitou predikčných modelov.

Kapitola č. 4 sa zameriava na tvorbu predikčného modelu, s dôrazom na prípravu dátových sád použitých pri jeho vývoji. Zvláštnu pozornosť venujeme aj spôsobu definovania závislej premennej, ktorá určuje časový horizont predikcie.

V tejto časti práce predstavíme spôsob tvorby závislej premennej, ktorý v súčasnosti využíva spoločnosť ČSOB. Zároveň uvedieme aj návrhy alternatívnych spôsobov tvorby závislej premennej, ktoré budú v rámci tejto práce ďalej analyzované a porovnávané pri rôznych dĺžkach predikčného obdobia.

Proces tvorby predikčného modelu zahŕňa viacero krokov. Dáta, ktorými disponujeme, podstúpia pred samotným modelovaním úpravu, ktorú opisujeme v kapitole č. 3. Upravené dáta predstavujú súbor dát, ktorý slúži ako vstupná sada pre trénovanie hyperparametrov metódy LightGBM, pričom proces ich optimalizácie je podrobne rozobraný v kapitole č. 4.

V kapitole č. 5 sú prezentované výsledky analýzy vplyvu časových faktorov na výkonnosť predikčných modelov. Pozornosť je primárne venovaná analýze vývoja kvality modelov v závislosti od predlžovania časového horizontu predikcie a aplikácie rôznych princípov tvorby závislej premennej, ktoré boli v práci navrhnuté.

1 Úloha času v predikčných modeloch

Metódy požívané na predikovanie správania klientov sú väčšinou modely binárnej klasifikácie. Takéto metódy majú interpretačnú výhodu, lebo ich odpoveď na otázku, či nastane udalosť je áno alebo nie. V práci budeme pracovať iba s metódami binárnej klasifikácie.

Dáta, ktoré spoločnosti zbierajú o klientoch alebo používateľoch sú závislé od času, pretože predstavujú sériu udalostí s konkrétnymi časmi nastatia. Reprezentácia údajov do formy vhodnej pre klasifikačný algoritmus zväčša vyžaduje agregáciu údajov do časových rezov v primeranej frekvencii. V ideálnom svete všetky potrebné údaje na modelovanie by boli vypočítané na dennej alebo menšej frekvencii, ale to nemusí byť ekonomicky efektívne. Spoločnosť ČSOB napríklad vypočítava agregované črty klienta z rôznych systémov na mesačnej báze a pomocou týchto údajov modeluje správanie klienta. Častejšia frekvencia by znamenala zvýšenie nákladov na úložisko, ale aj ľudskú prácu. Aby sme dokázali vytvoriť predikčný model s využitím týchto dát, k črtám klienta sa musí ešte vytvoriť závislá premenná – binárna premenná, ktorá hovorí o tom, či u klienta nastala istá udalosť alebo nie.

Dátovú sadu používanú na tvorbu predikčného modelu binárnej klasifikácie môžeme rozdeliť do dvoch kategórií, ktoré budeme nazývať ako **pozitívnu vzorku** (črty, u ktorých nastala udalosť a majú tým pádom hodnotu 1 v závislej premennej) a **negatívnu vzorku** (črty, u ktorých nenastala udalosť a majú tým pádom hodnotu 0 v závislej premennej).

Nastavenie modelu, ktorý má za cieľ odhadovať správanie klienta, vždy čelí trom najväčším otázkam:

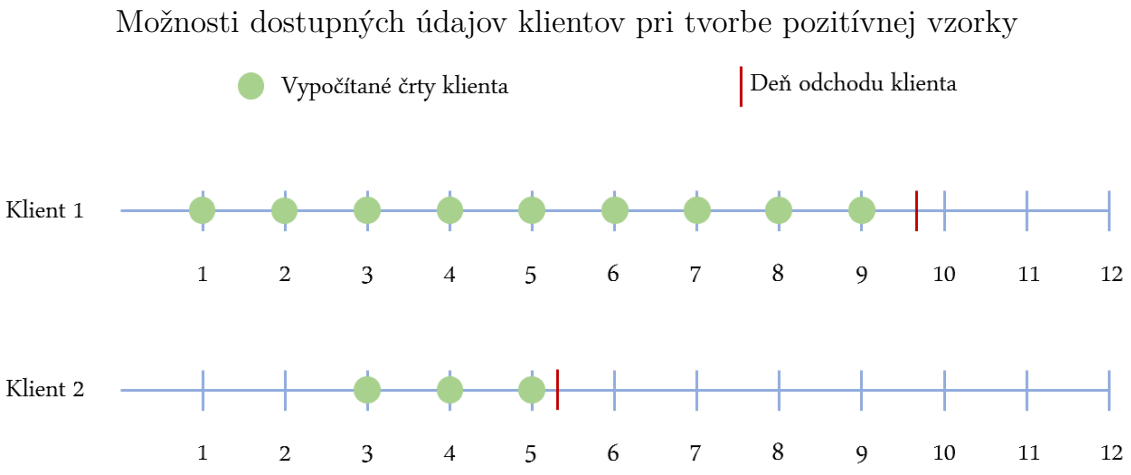
1. Ako vytvoriť závislú premennú: $Y(t) = f(X(t-1)) + \epsilon$?
2. Čo budú vstupné črty v modeli?
3. Ako zvoliť klasifikátor $f(X(t-1))$, ktorý modeluje správanie klientov?

Naša práca je práve zameraná na prvú otázku. Aj keď sa to nemusí zdať, podľa spôsobu tvorby závislej premennej a teda označenie dátových bodov, ktoré budú vstupovať do modelu v pozitívnej vzorke je kľúčové. Na základe tohto výberu vzniká otázka, na ktorú model zodpovedá.

Črty, pomocou ktorých modelujeme správanie klienta sú závislé od času a to tým spôsobom, že vznikajú z dostupnej histórie pozorovaní. Taktiež v každom čase v minulosti mohol mať klient iné hodnoty vypočítaných čŕt a teda máme aj viac pozorovaní o tom istom klientovi s rôznymi hodnotami čŕt.

Pre lepšie pochopenie uvádzame príklad. Na obrázku č. 1 je znázornená časová os dvoch klientov počas jedného roka, pričom číselne sú označené konce jednotlivých mesiacov. Zelenou farbou sú vyznačené dátumy, ku ktorým boli vypočítané črty klienta na základe historických údajov (vždy k dátumu konca mesiaca). Bordová farba označuje dátum vzniku udalosti, v tomto príklade si môžeme predstaviť odchod klienta.

Z obrázku je zrejmé, že črty používané na modelovanie udalostí sú obmedzené dátumami výpočtu, konkrétne poslednými dňami jednotlivých mesiacov. V tejto súvislosti vzniká otázka, ktorý konkrétny dátum je vhodné zvoliť na výber čŕt klienta, ktorým bude v závislej premennej priradená hodnota 1, čím zároveň indikujeme, že pre daný vektor čŕt nastala príslušná udalosť. Na základe rozdielu medzi dátumami vybraných vektorov čŕt klientov a dátumami ich odchodov bude model odpovedať nielen na otázku, či klient odíde, ale aj v akom časovom horizonte odchod nastane.

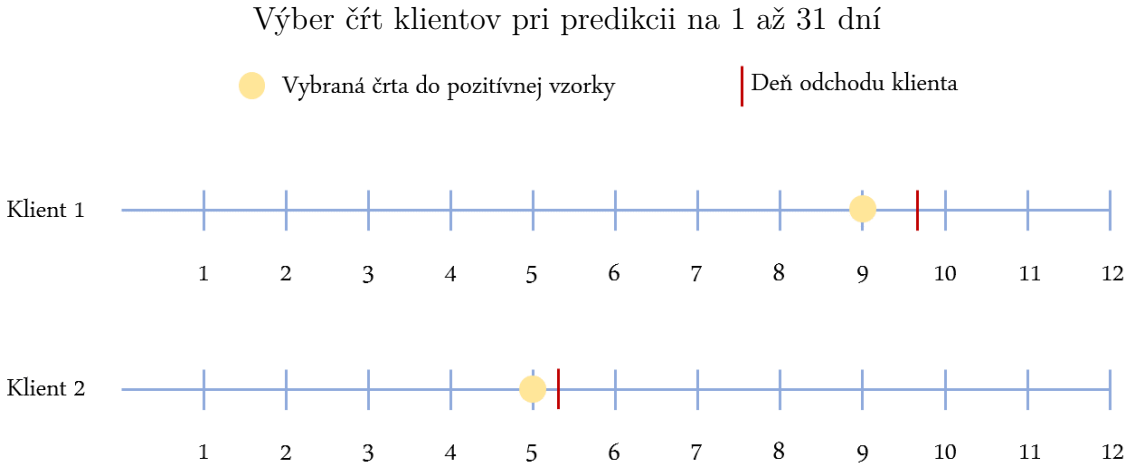


Obr. 1: Obrázok znázorňuje dostupnosť údajov potrebných na modelovanie správania klienta. Môžeme pozorovať, že pri Klientovi 2 sú dostupné údaje iba z troch časových rezov, ktoré môžu byť využité pri tvorbe pozitívnej vzorky. *Zdroj: Vlastné spracovanie.*

Na obrázku č. 1 môžeme pozorovať aj bežnú vec v reálnych dátach. A to je nedostupnosť dát. Kým u Klienta 1 máme 9 pozorovaní, tak u Klienta 2 máme len 3.

Ak by sme označili vektor čŕt s najbližším dátumom k výskytu udalosti ako pozitívny,

natrénovaný model by odpovedal na otázku, či klient odíde v časovom horizonte 1 až 31 dní. Tento princíp je graficky znázornený na obrázku č. 2.



Obr. 2: Obrázok znázorňuje výber najaktuálnejšieho vektora čŕt, pri ktorom predpokladáme, že došlo k udalosti. Tento vektor vybraný z príslušného dátumu je následne zaradený do pozitívnej vzorky počas procesu tvorby modelu, pričom má hodnotu 1 v závislej premennej. *Zdroj: Vlastné spracovanie.*

Takéto riešenie však nemusí byť efektívne, najmä pri odchodovosti, pretože spoločnosti chcú identifikovať takýchto klientov čo najskôr, aby mali možnosť ovplyvniť ich rozhodnutie. Navyše, pri určitých typoch produktov môže správanie klienta už v skoršom období naznačovať záujem o konkrétny produkt. Príkladom môže byť hypotéka, na ktorú sa klient musí dlhodobo pripravovať, aby mohol o ňu požiadať.

2 Využívané metódy strojového učenia

V práci využívame klasifikačnú metódu strojového učenia LightGBM, ktorá patrí do rodiny modelov gradientného boostingu. V tejto kapitole predstavíme túto kategóriu modelov a zároveň priblížime výhody samotnej metódy LightGBM. Autori tejto metódy na záver svojej práce hodnotia jej vylepšenia nasledovne:

„Vykonali sme teoretickú analýzu aj experimentálne štúdie týchto dvoch techník. Experimentálne výsledky sú v súlade s teóriou a ukazujú, že s pomocou GOSS a EFB môže LightGBM výrazne prekonať XGBoost a SGB z hľadiska výpočtovej rýchlosti a spotreby pamäte.“ [3]

V tejto kapitole zároveň predstavíme metriky, ktoré slúžia na sledovanie kvality modelu a podrobne rozoberieme metodiku použitú pri kvantifikovaní vzťahu medzi kvalitou predikčného modelu a predlžovaním horizontu predikcie.

2.1 Gradient boosting metóda

Gradient boosting je metóda strojového učenia používaná na úlohy regresie aj klasifikácie, ktorá vytvára robustný predikčný model ako súčet jednoduchých modelov. Táto metóda spadá do širšieho rámca funkčnej aproximácie, kde cieľom je nájsť funkciu $\widehat{f}(\mathbf{x})$, ktorá minimalizuje hodnotu stratovej funkcie $L(y, \hat{f}(\mathbf{x}))$ na trénovacích dátach. Informácie k tejto podkapitole čerpáme zo zdrojov [4, 5].

Základný princíp metódy gradient boosting

Nech máme trénovaciu množinu $\{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, kde $\mathbf{x}_i \in \mathbb{R}^d$ sú vstupné vektory čŕt a y_i sú hodnoty závislej premennej, pričom N je počet pozorovaní. Cieľom je nájsť aproximáciu $\hat{f}(\mathbf{x})$ funkcie $f^* : \mathbf{x} \to y$, ktorá minimalizuje očakávanú hodnotu stratovej funkcie $L(y, \hat{f}(\mathbf{x}))$:

$$\widehat{f}(\mathbf{x}) \approx f^*(\mathbf{x}), \tag{1}$$

$$\widehat{f}(\mathbf{x}) = \arg\min_{f} E_{\mathbf{x}}[E_y(L(y, f(\mathbf{x}))) \mid \mathbf{x}], \tag{2}$$

kde:

- E_y je stredná hodnota stratovej funckie pre jedno pozorovanie,

- $E_{\mathbf{x}}$ je celková stredná hodnota stratových funkcií vzhľadom na celý dataset,
- $L(y, f(\mathbf{x}))$ je stratová funkcia,
- $\widehat{f}(\mathbf{x})$ je odhad funkcie $f^*(\mathbf{x})$.

Závislá premenná y môže pochádzať z rôznych rozdelení, od ktorého závisí predpis stratovej funkcie $L(y, f(\mathbf{x}))$. Aby bolo možné odhadnúť funkciu, pri výpočte sa obmedzuje len na parametrickú rodinu funkcií: $f(\mathbf{x}, \theta)$. Následne pre túto funkciu sa optimalizačný problém mení na nasledujúci tvar:

$$\widehat{f}(\mathbf{x}) = f(\mathbf{x}, \widehat{\theta}), \tag{3}$$

$$\widehat{\theta} = \arg\min_{\theta} E_{\mathbf{x}}[E_y(L(y, f(\mathbf{x}, \theta))) \mid \mathbf{x}]. \tag{4}$$

Gradient boosting metóda využíva na aproximáciu funkcie $f^*(\mathbf{x})$ iteratívne nasčítavanie jednoduchých parametrizovaných metód. V práci používame metódu, ktorá používa na tento účel rozhodovacie stromy, ale vo všeobecnosti to môže byť hocijaká funkcia. Z vyššie uvedených informácií môžeme hľadanú funkciu zapísať v nasledujúcom tvare:

$$\widehat{f}(\mathbf{x}) = \hat{f}_M(\mathbf{x}) = f_M(\mathbf{x}, \widehat{\rho}_{1:M}, \widehat{\theta}_{1:M}) = f_{M-1}(\mathbf{x}, \widehat{\rho}_{1:M-1}, \widehat{\theta}_{1:M-1}) + \widehat{\rho}_M T(\mathbf{x}, \widehat{\theta}_M), \tag{5}$$

kde:

- $T(\mathbf{x}, \widehat{\theta}_M)$ je jednoduchá metóda s aproximáciou jej parametrov $\widehat{\theta}_M$,
- $\widehat{\rho}_M$ je odhadnutá váha M-tej jednoduchej metódy $T(\mathbf{x}, \widehat{\theta}_M)$ v aproximácii funkcie $f^*(\mathbf{x})$,
- $f_{M-1}(\mathbf{x}, \widehat{\rho}, \widehat{\theta}_{1:M-1})$ je aproximácie skutočnej funkcie $f^*(\mathbf{x})$ v $M-1$ iterácii s parametrami: $\widehat{\rho}_{1:M-1}$, $\widehat{\theta}_{1:M-1}$.

Následne pre M-tú iteráciu dostávame optimalizačný problém v tvare:

$$(\widehat{\rho}_M, \widehat{\theta}_M) = \arg\min_{\rho,\theta} \sum_{i=1}^{N} L(y_i, \widehat{f}_{M-1}(\mathbf{x}_i) + \rho T(\mathbf{x}_i, \theta)), \tag{6}$$

kde $\mathbf{x}_i$ reprezentuje i-te pozorovanie v trénovacej sade.

V prípade, že je daný tvar stratovej funkcie $L(y, f(\mathbf{x}))$ a jednoduchá metóda $T_M(\mathbf{x}, \theta)$, je odhad parametrov $\hat{\theta}$ a $\hat{\rho}$ výpočtovo náročný. Metóda gradient boosting túto náročnosť zmierňuje tak, že odhaduje parametre $\hat{\rho}_M$ a $\hat{\theta}_M$ s cieľom, aby vektor:

$$\mathbf{T}_M = [T(\mathbf{x}_1, \hat{\theta}_M), T(\mathbf{x}_2, \hat{\theta}_M), \ldots, T(\mathbf{x}_N, \hat{\theta}_M)], \tag{7}$$

bol čo najviac paralelný s vektorom negatívneho gradientu:

$$\mathbf{g}^{(M)} = [g_1^{(M)}, g_2^{(M)}, \ldots, g_N^{(M)}], \tag{8}$$

ktorý má na i-tej zložke hodnotu:

$$g_i^{(M)} = -\left[\frac{\partial L(y_i, f(\mathbf{x}_i))}{\partial f(\mathbf{x}_i)}\right]_{f(\mathbf{x}_i)=\hat{f}_{M-1}(\mathbf{x}_i)}, \quad \forall i \in \{1, 2, \ldots, N\} \tag{9}$$

kde $g_i^{(M)}$ reprezentuje i-tú zložku negatívneho gradientu (pseudo-rezíduá). Namiesto hľadania všeobecného riešenia pre ideálny prírastok $T_M(\mathbf{x}, \widehat{\theta})$ k funkcii $\widehat{f}_{M-1}(\mathbf{x})$ sa metóda gradient boosting zameriava na výber parametrov jednoduchej metódy $T_M(\mathbf{x}, \theta)$ tak, aby bol výsledný prírastok čo najbližšie k hodnote záporného gradientu, ktorý udáva smer poklesu hodnoty stratovej funkcie. Týmto spôsobom sa výpočtovo náročná optimalizácia parametrov θ môže transformovať na jednoduchšiu úlohu:

$$\widehat{\theta}_M = \arg\min_{\theta} \sum_{i=1}^{N} (g^M(\mathbf{x}_i) + T(\mathbf{x}_i, \theta))^2 \tag{10}$$

a následne v pôvodnom optimalizačnom probléme sa optimalizuje váha jednoduchej metódy:

$$\widehat{\rho}_M = \arg\min_{\rho} \sum_{i=1}^{N} L(y, f_{M-1}(\mathbf{x}) + \rho T(\mathbf{x}_i, \hat{\theta})) \tag{11}$$

$$\widehat{\rho}_M, \tag{12}$$

Následne dostávame aproximáciu funkcie $f^*(\mathbf{x})$ v M-tej iterácii:

$$f^*(\mathbf{x}) \approx \hat{f}(\mathbf{x}) = \hat{f}_M(\mathbf{x}) = \hat{f}_{M-1}(\mathbf{x}) + \hat{\rho}_M T(\mathbf{x}, \hat{\theta}_M) \tag{13}$$

Gradient Boosting je základom mnohých moderných metód `XGBoost`, `LightGBM` alebo `CatBoost`.

2.2 Metóda LightGBM

LightGBM (skrátené z anglického výrazu: *Light gradient boosting machine*) je efektívna metóda z rodiny gradientného boostingu založená na rozhodovacích stromoch. Je navrhnutá tak, aby bola rýchla a efektívna na veľkých dátových sadách. LightGBM upravuje tradičný gradient boosting pomocou inovatívnych techník, ktoré znižujú čas trénovania a zlepšujú presnosť modelu.

Ako sme uviedli v podkapitole 2.1, gradient boosting používa na odhad funkcie $f^* : \mathbf{x} \to y$ postupné nasčítavanie jednoduchých metód $T(\mathbf{x}, \hat{\theta})$ s váhou $\hat{\rho}$. Gradient boosting je základom metódy LightGBM, ale táto metóda obsahuje vylepšenia, ako je napríklad parameter učenia η. Parameter učenia zabraňuje pretrénovaniu modelu na trénovacej sade. Aproximáciu funkcie $f^*(\mathbf{x})$ pomocou metódy LightGBM môžeme zapísať nasledovne:

$$f^*(\mathbf{x}) \approx \hat{f}_M(\mathbf{x}) = \hat{f}_{M-1}(\mathbf{x}) + \eta\hat{\rho}T_M(\mathbf{x}, \hat{\theta}). \tag{14}$$

Nízka hodnota parametra učenia η spomaľuje učenie sa modelu, čo môže viesť k dlhšej výpočtovej náročnosti, avšak takéto nastavenie vedie aj k viac robustnejším výsledkom z dôvodu väčšieho počtu iterácií.

V práci od tohto momentu budeme používať označenie vstupných parametrov metódy LightGBM ako hyperparametre metódy. Pod týmto pojmom označujeme všetky parametre, ktoré slúžia na výpočet parametrov funkcie $\hat{f}(\mathbf{x})$. Takýmito hyperparametrami sú napríklad:

- M - počet iterácií,
- η - parameter učenia,
- maximálna hĺbka rozhodovacieho stromu.

Pri tvorbe modelu optimalizujeme hodnoty hyperparametrov, aby čo najpresnejšie odhadnutá funkcia $\hat{f}(\mathbf{x})$ aproximovala reálnu funkciu $f^*(\mathbf{x})$. Zoznam všetkých hyperparametrov a postup optimalizácie je uvedený v podkapitole č. 4.4.

Informačné kritérium

V rozhodovacom strome sa uzly delia na základe hodnoty informačného kritéria a vyberá sa črta, ktorá má v uzle najväčšiu hodnotu kritéria, teda s najväčším informačným ziskom

pre rozhodovací strom. Pri tvorbe klasifikačných modelov, metóda LightGBM používa zisk z variancie (anglický výraz: *gain variance*). V prípade črty j pre hodnotu d v danom uzli je definovaný zisk z variancie nasledovne [3]:

$$V_{j|O}(d) = \frac{1}{n_O}\left(\frac{\left(\sum_{x_i \in O: x_{ij} \leq d} g_i\right)^2}{n^j_{l|O}(d)} + \frac{\left(\sum_{x_i \in O: x_{ij} > d} g_i\right)^2}{n^j_{r|O}(d)}\right), \tag{15}$$

kde:

- j je index črty,
- O je množina dát v j-tej črte klienta,
- $x_{i,j}$ je i-ta hodnota v rámci j-tej črty,
- g_i je hodnota negatívneho gradientu,
- n_O je celkový počet pozorovaní v množine O
- $n^j_{l|O}(d)$ je počet pozorovaní, pre ktoré platí $x_{i,j} \leq d$,
- $n^j_{r|O}(d)$ je počet pozorovaní, pre ktoré platí $x_{i,j} > d$.

Následne pre črtu j algoritmus rozhodovacieho stromu vyberie hodnotu rozdelenia uzla s hodnotou d^*

$$d^*_j = \arg\max_d V_j(d) \tag{16}$$

Keď je pre črtu j vypočítaný optimálny deliaci bod d^*_j, v rozhodovacom strome sa pre deliaci uzol vyberie tá črta, ktorá dosahuje maximálnu hodnotu zisku z variancie.

2.2.1 Vylepšenia v LightGBM

LightGBM zavádza niekoľko algoritmických inovácií, ktoré ho odlišujú od tradičných metód gradientného boostingu.

Na rozdiel od úrovňového rastu stromov (anglický výraz: *level-wise*) používaného v iných algoritmoch, LightGBM používa uzlový rast (anglický výraz: *leaf-wise*). To znamená, že strom expanduje najskôr v uzle s najvyšším ziskom. Tým sa minimalizuje stratová funkcia efektívnejšie.

Gradientne založené jednostranné vzorkovanie (GOSS)

GOSS (skrátené z anglického výrazu: *gradient-based one-side sampling*) predstavuje efektívne riešenie pre optimalizáciu času potrebného na trénovanie metódy LightGBM. Pri spracovaní rozsiahlych dátových súborov dokáže LightGBM znižovať počet vzoriek použitých na tréning tým, že selektívne vyberá podmnožinu dát na základe hodnoty negatívnych gradientov. Vzorky s výrazne vysokými gradientmi sú ponechané pre výpočet nasledujúceho rozhodovacieho stromu, zatiaľ čo vzorky s nižšími gradientmi sú náhodne podvzorkované [3].

Exkluzívne zoskupovanie čŕt (EFB)

EFB (skrátené z anglického výrazu: *Exclusive feature bundling*) predstavuje techniku v LightGBM, ktorá je navrhnutá na efektívne spracovanie vysoko-dimenzionálnych dát. Kľúčovým aspektom tejto metódy je zlučovanie čŕt, pričom sa využíva pojem exkluzívna črta.

Exkluzívne črty sú také črty, ktoré majú nenulové hodnoty na odlišných pozíciách v rámci dátovej sady, pričom prekryv medzi týmito pozíciami je minimálny [3].

Jednoduchým príkladom využitia tohto vylepšenia je situácia, keď v dátovej sade disponujeme binárnymi črtami – aké produkty má klient zakúpené. V takomto prípade máloktorý klient má všetky produkty. Vzhľadom na to, že niektoré produkty väčšina klientov nemá súčasne, LightGBM tieto črty s nízkym prekryvom dokáže efektívne zlúčiť do jednej komplexnej črty.

Táto transformácia umožňuje zníženie počtu čŕt v dátovej sade, čo prispieva k zníženiu výpočtovej náročnosti modelu a zároveň zachováva čo najväčšie množstvo informačného obsahu. Redukcia rozmernosti vstupných dát tak zefektívňuje tréning modelu a skracuje čas potrebný na odhad parametrov.

Delenie založené na histograme

Histogramové delenie (z anglického výrazu: *Histogram-based split*) predstavuje techniku, ktorá značne zrýchľuje výpočet pri tvorbe predikčného modelu. Táto metóda spočíva v predbežnej príprave dát, ktoré sú následne využité na výpočet parametrov v metóde LightGBM. V rámci tejto techniky dochádza k transformácii číselných hodnôt čŕt do

preddefinovaných intervalov, s ktorými algoritmus pracuje pri tvorbe rozhodovacích stromov [3].

Aplikácia histogramového delenia umožňuje výrazné zvýšenie efektivity algoritmu. Namiesto výpočtu informačného kritéria pre každú jednotlivú číselnú hodnotu v každom uzle stromu sa výpočty redukujú na prácu s niekoľkými preddefinovanými deliacimi bodmi, ktoré reprezentujú intervaly danej črty. Takýmto spôsobom dochádza nielen k urýchleniu výpočtového procesu, ale aj k zefektívneniu pri hľadaní optimálnych hodnôt v uzloch rozhodovacích stromov.

2.3 Metriky sledujúce kvalitu predikčného modelu

Pri hodnotení kvality modelov strojového učenia je kľúčové zvoliť vhodné metriky, ktoré umožňujú objektívne posúdenie presnosti predikcie. V tejto kapitole sa venujeme metrikám určeným na hodnotenie kvality binárnych klasifikačných modelov, ktoré využívame pri tvorbe predikčného modelu v kapitole č. 4, ako aj pri porovnávaní vývoja kvality modelov v závislosti od zvoleného časového horizontu predikcie v kapitole č. 5. V tejto kapitole uvádzame definície relevantných metrík.

Matica zámien

Matica zámien (anglický výraz: *confusion matrix*) poskytuje základný prehľad o správnosti a chybovosti binárneho klasifikačného modelu. V prípade binárnej klasifikácie obsahuje štyri hodnoty [6]:

- **TP** (Skutočne pozitívne, anglický výraz: *True positive*): Počet správne klasifikovaných pozitívnych prípadov.
- **TN** (Skutočne negatívne, anglický výraz: *True negative*): Počet správne klasifikovaných negatívnych prípadov.
- **FP** (Falošne pozitívne, anglický výraz: *False positive*): Počet nesprávne klasifikovaných negatívnyh prípadov ako pozitívne.
- **FN** (Falošne negatívne, anglický výraz: *False negative*): Počet nesprávne klasifikovaných pozitívnych prípadov ako negatívne.

Matematicky je matica zámien štruktúrovaná nasledovne:

$$\text{Matica zámien} = \begin{bmatrix} TP & FP \\ FN & TN \end{bmatrix} \tag{17}$$

V tejto časti je potrebné doplniť to, že metódy strojového učenia modelujú propenzitu výskytu udalosti, ktorá nadobúda hodnotu v intervale: $[0, 1]$. Keďže cieľom je rozhodnúť, či k udalosti došlo alebo nie, je potrebné túto spojitú predikciu previesť na binárne rozhodnutie. Na tento účel sa zavádza prahová hodnota klasifikácie, označovaná symbolom α, ktorá slúži na rozlíšenie medzi pozitívnou a negatívnou triedou. Konkrétne, ak predikovaná hodnota propenzity prekročí prah α, je príslušné pozorovanie klasifikované ako pozitívne, v opačnom prípade ako negatívne.

Precíznosť

Precíznosť (anglický výraz: *precision*) vyjadruje presnosť pozitívnych predikcií. Metrika vyjadruje pravdepodobnosť, že inštancia je skutočne pozitívna za predpokladu, že ju model klasifikoval ako pozitívnu [6].

$$\text{Precíznosť} = \text{Precision} = P(y = 1 \mid \hat{y} = 1) = \frac{TP}{TP + FP}. \tag{18}$$

Senzitivita

Senzitivita (anglický výraz: *recall*, *sensitivity* alebo *true positive rate* - *TPR*) vyjadruje schopnosť modelu identifikovať všetky pozitívne prípady. Senzitivita predstavuje pravdepodobnosť, že model správne predikuje pozitívnu kategóriu za predpokladu, že pozorovanie skutočne patrí k pozitívnej vzorke [6].

$$\text{Senzitivita} = \text{TPR} = \text{Sensitivity} = \text{Recall} = P(\hat{y} = 1 \mid y = 1) = \frac{TP}{TP + FN}. \tag{19}$$

P-R krivka

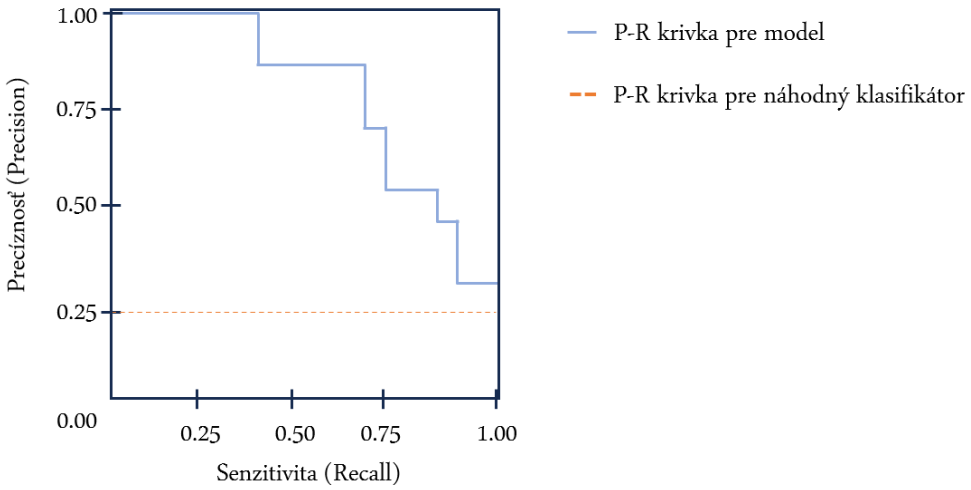
P-R krivka (skrátené z anglického výrazu: *Precision-Recall curve*) zobrazuje závislosť ukazovateľov precíznosti (angl. *precision*) a senzitivity (angl. *recall*) v závislosti od klasifikačného prahu α, ktorým je krivka parametrizovaná pri určovaní zaradenia pozorovania modelom do pozitívnej kategórie.

Táto krivka vzniká spojením bodov $(\text{Recall}_\alpha, \text{Precision}_\alpha)$, pričom α predstavuje hodnotu klasifikačného prahu.

$$\text{Precision}_\alpha = \frac{TP_\alpha}{TP_\alpha + FP_\alpha}, \tag{20}$$

$$\text{Recall}_\alpha = \frac{TP_\alpha}{TP_\alpha + FN_\alpha}. \tag{21}$$

Na obrázku č. 3 je zobrazený ilustračný príklad vývoja P-R krivky, pričom je porovnaná s P-R krivkou modelu náhodného klasifikátora. P-R krivka náhodného klasifikátora je umiestnená na úrovni pomeru počtu prípadov, v ktorých nastala udalosť ku celkovému počtu dát v sade. V ilustrovanom príklade je táto hodnota 0.25.



Obr. 3: Ilustračný vývoj P-R krivky s porovnaním ku náhodnému klasifikátorovi. *Zdroj: Vlastné spracovanie.*

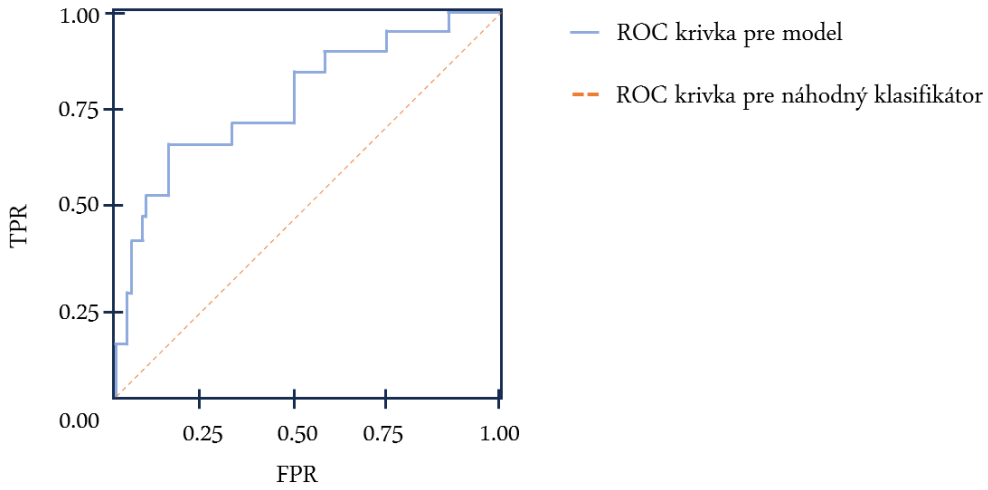
ROC krivka

ROC krivka (skrátené z anglického výrazu: *Receiver Operating Characteristic*) zobrazuje vzťah medzi skutočne pozitívnou mierou (TPR, vzťah (23)) a falošne pozitívnou mierou (FPR, vzťah (22)) pri všetkých nastaveniach prahovej hodnoty α pre rozhodovanie, či ide o pozitívnu alebo negatívnu predikciu. Spojením jednotlivých bodov $(\text{FPR}_\alpha, \text{TPR}_\alpha)$ vzniká ROC krivka, ktorá slúži na znázornenie kvality modelu.

$$\text{FPR}_\alpha = \frac{FP_\alpha}{FP_\alpha + TN_\alpha} \tag{22}$$

$$\text{TPR}_\alpha = \frac{TP_\alpha}{TP_\alpha + FN_\alpha}, \tag{23}$$

Na obrázku č. 4 uvádzame príklad pre vývoj ROC krivky pre ilustračný model v porovnaní s ROC krivkou pre model náhodného klasifikátora.



Obr. 4: Na obrázku je znázornený ilustračný vývoj ROC krivky s porovnaním ku náhodnému klasifikátorovi. *Zdroj: Vlastné spracovanie.*

AUC

AUC (skrátené z anglického výrazu *Area Under Curve*) predstavuje metriku, ktorá vyjadruje plochu pod krivkou. V rámci našej práce je najpodstatnejšou metrikou plocha pod P-R krivkou, pretože poskytuje informácie o celkovej presnosti modelu pri rôznych úrovniach prahovej hodnoty. Táto metrika dáva možnosť simultáne zohľadniť všetky prahové hodnoty α, čo by pri pohľade na krivku nebolo možné. AUC-PR môžeme definovať nasledovne:

$$\text{AUC-PR} = \int_0^1 Precision(Recall(\alpha))dRecall(\alpha) = \int_0^1 \text{Precision}(\alpha) \cdot \left| \frac{d}{d\alpha}\text{Recall}(\alpha) \right| d\alpha, \tag{24}$$

v prípade P-R krivky s optimálnym tvarom by plocha pod krivkou pokrývala celý jednotkový štvorec. To by znamenalo, že pre určitú prahovú hodnotu α by platilo $Precision_\alpha = 1$ a zároveň $Recall_\alpha = 1$, čo by signalizovalo maximálnu možnú precíznosť (angl. *Precision*) a 100% senzitivitu (angl. *Recall*). Na druhej strane, v prípade predikčného modelu s najhorším možným tvarom PR krivky, plocha pod krivkou by mala nulovú hodnotu.

Pre náhodný klasifikátor platí, že hodnota metriky zodpovedá hodnote pomeru počtu prípadov v pozitívnej vzorke k celkovému počtu prípadov v dátovej sade. V práci budeme túto hodnotu označovať ako teoretické minimum.

Okrem toho budeme v tejto práci používať aj metriku AUC-ROC, ktorá vyjadruje plochu pod ROC krivkou. Túto metriku aplikujeme pri výbere hyperparametrov metódy LightGBM, ako je uvedené v podkapitole č. 4.4. Hodnota AUC-ROC nadobúda hodnoty v intervale: $[0, 1]$.

2.4 Logistická regresia

V podkapitole 2.3 sme predstavili metriky používané na hodnotenie kvality modelov. Aby sme mohli analyzovať ich hodnoty a vývoj v súvislosti s predlžovaním horizontu predikcie, resp. parametra t v modeloch, ktorý vysvetľujeme v podkapitole č. 3.2, je potrebné zaviesť metriku, ktorá dokáže tento vzťah zachytiť.

Logistická regresia je štatistický model, ktorý sa využíva na modelovanie binárnej závislej premennej Y s alternatívnym rozdelením. Matematicky môžeme túto závislosť vyjadriť ako [7]:

$$Y \sim \text{Alt}(1, p_{\mathbf{x}}), \tag{25}$$

kde Y je náhodná premenná z alternatívneho rozdelenia: $\text{Alt}(1, p_{\mathbf{x}})$ a jej realizácie patria množine: $\{0, 1\}$ s pravdepodobnosťou $p_{\mathbf{x}}$ nadobudnutia realizácie s hodnotou 1. Tento model sa používa v úlohách binárnej klasifikácie, kde sa predpovedá, či inštancia patrí do jednej z dvoch kategórií. V logistickej regresii sa na odhad pravdepodobnosti $p_{\mathbf{x}}$ môže použiť špeciálny prípad logistickej funkcie, ktorá transformuje lineárnu kombináciu nezávislých premenných na hodnotu v intervale: $(0, 1)$. Táto funkcia má tvar [8]:

$$\text{Logistic}(z) = \frac{1}{1 + e^{-z}}. \tag{26}$$

Logistická regresia modeluje pravdepodobnosť, že dané pozorovanie patrí do kategórie 1, na základe hodnoty nezávislých premenných. Tento vzťah možno vyjadriť nasledovne [8]:

$$P(Y = 1 \mid \mathbf{x}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)}}, \tag{27}$$

kde:

- $P(Y = 1 \mid \mathbf{x})$ predstavuje pravdepodobnosť, že náhodná premenná Y nadobúda hodnotu 1, čiže dané pozorovanie patrí do skúmanej kategórie,
- $\mathbf{x}$ je vektor známych hodnôt charakterizujúci náhodnú premennú Y,
- $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ sú koeficienty logistickej regresie.

2.4.1 Použitie logistickej regresie v kontexte práce

Logistická regresia je využívaná na modelovanie klasifikačných problémov, kde cieľová premenná môže nadobudnúť hodnoty 0 alebo 1. Metrika na hodnotenie kvality modelu, ktorú v práci používame je AUC-PR. Táto metrika má obor hodnôt v intervale: $[0, 1]$. V prípade, že hodnota metriky dosiahne hodnotu 1, znamená to, že model dosiahol perfektné výsledky v predikcii. Naopak, ak hodnota metriky dosiahne hodnotu 0, ide o model s najhoršími možnými predikciami. Výsledky modelov vyjadrené pomocou metriky AUC-PR možno interpretovať ako pravdepodobnosti, že daný model bude efektívne predikovať správanie klienta, pričom hodnota 1 hovorí o tom, že model predikuje správanie klientov bez chyby.

V našom prípade pri rôznych hodnotách horizontu predikcie, ktorý je určený parametrom t pri tvorbe závislej premennej, získavame pre každý model rôzne hodnoty týchto metrik v intervale $(0, 1)$. Tieto metriky kvality modelov nám umožňujú hodnotiť, ako kvalitné sú modely pri rôznych nastaveniach časového horizontu. V zásade, tieto výsledky kvality modelov môžeme považovať za pravdepodobnosti, že model v danej metrike je perfektný $P(metrika = 1 \mid t)$. Na základe výsledkov môžeme následne odhadnúť závislosť medzi kvalitou modelov a dĺžkou predikcie.

Model logistickej regresie používaný na tento účel má nasledovný tvar:

$$P(\text{AUC-PR} = 1 \mid t) = \frac{1}{1 + \exp^{-(\beta_0 + \tau t)}}, \tag{28}$$

kde:

- $P(\text{AUC-PR} = 1 \mid t)$ je pravdepodobnosť, že model je perfektný vzhľadom na využívanú metriku AUC-PR,
- t je parameter, ktorý určuje horizont predikcie (viď podkapitola. č. 3.2).

Model logistickej regresie nám prináša aj možnosť kvantifikovania vplyvu dĺžky predikcie vzhľadom na kvalitu modelu. Pomocou parametra τ vieme vypočítať koľko násobne narastú alebo poklesnú šance na perfektný model v prípade zväčšenia parametra t nasledovne:

$$\text{OR} = \frac{ODDS'}{ODDS} = e^{\tau} \quad \text{, pričom} \tag{29}$$

$$ODDS = \frac{p}{1-p} = e^{\beta_0 + \tau t} \quad , \quad ODDS' = \frac{p'}{1-p'} = e^{\beta_0 + \tau(t+1)}, \tag{30}$$

kde:

- p je pravdepodobnosť že model je perfektný v rámci metriky vzhľadom na časový horizont t tj.: $P(\text{AUC-PR} = 1 \mid t)$,
- p' je pravdepodobnosť že model je perfektný v rámci metriky, vzhľadom na časový horizont $t+1$ tj.: $P(\text{AUC-PR} = 1 \mid t+1)$,
- $ODDS$ je šanca, že model je perfektný v rámci metriky,
- OR je pomer šancí.

3 Postup tvorby dátovej sady pre predikčný model

Predtým, než môžeme začať s tvorbou predikčného modelu je potrebné vytvoriť samotnú dátovú sadu. V tejto práci využívame už existujúce črty klientov, ktoré boli vytvorené v spoločnosti ČSOB. Okrem toho sa však venujeme aj tvorbe nových čŕt klienta, ktoré dopĺňajú existujúce informácie a umožňujú detailnejšie modelovanie jeho správania.

Dátová sada tvorená vektormi čŕt klientov, ktoré slúžia na trénovanie a validovanie binárnej klasifikačnej metódy strojového učenia LightGBM, je rozdelená na dve vzorky dát: pozitívnu a negatívnu vzorku.

Pozitívna vzorka pozostáva z vektorov čŕt klientov, u ktorých závislá premenná nadobúda hodnotu 1, čo signalizuje výskyt udalosti. V negatívnej vzorke sa nachádzajú vektory čŕt klientov, u ktorých nenastala udalosť a z tohto dôvodu majú hodnotu 0 v závislej premennej. V rámci tejto časti prípravy dátovej sady určenej pre tvorbu predikčného modelu vzniká konkrétna otázka, na ktorú model následne poskytuje odpoveď.

3.1 Črty klientov

Ako sme uviedli v kapitole č. 1, spoločnosť ČSOB využíva dáta zo systémov, z ktorých generuje tabuľky čŕt pre svojich klientov na mesačnej báze. Spolu ide o viac ako 2000 čŕt, ktoré umožňujú modelovať správanie klientov. Črty sledujú historické správanie klienta v rámci určitého časového obdobia, napríklad mesiaca, kvartálu alebo roka, alebo historický vývoj ekonomickej situácie.

Príkladom črty historického vývoja ekonomickej situácie môže byť aktuálna miera inflácie alebo výška úrokovej sadzby. Medzi používané črty historického správania klienta patrí napríklad NACE kód sféry podnikania v prípade právnickej osoby, vek alebo medzimesačné pohyby na účte a mnohé ďalšie v prípadoch fyzickej osoby.

V rámci našej práce sme vytvorili ďalších 100 čŕt z ďalšieho systému, ktoré môžu byť použité na modelovanie správania klienta. Systém, z ktorého pochádzajú tieto dáta je chatbot Kate od spoločnosti ČSOB. Nespracované dáta z tohto systému obsahujú informácie o každej správe, ktorú klient napísal počas celej existencie Kate. Táto tabuľka bude v budúcnosti vypočítavaná na mesačnej báze, podobne ako ostatné tabuľky čŕt

klientov.

Hlavnými údajmi, ktoré sme použili na tvorbu tabulky čŕt, boli dátum a čas správy, zámer správy a jej predmet. Zámer a predmet správy sú odhadované jazykovým modelom Kate na základe obsahu správy od klienta. Tieto hodnoty sú kategorické, čo znamená, že existuje pevne stanovený zoznam možných zámerov a predmetov. Tento zoznam však môže byť v budúcnosti aktualizovaný a rozšírený o nové zámery alebo predmety.

Aby sme tento potenciálny vývoj jazykového modelu Kate zohľadnili, pri práci s týmito dátami sme využili embeddingový model od spoločnosti Facebook AI Research s názvom FastText [11]. Embedding umožňuje priradiť slovám numerické hodnoty - vektory, ktoré sa v prípade podobných slov nachádzajú v priestore blízko seba. Detailnejšie vysvetlenie princípu embeddingu je uvedené v podkapitole 3.1.1.

Spôsob akým sme črty vytvárali, vychádza z marketingovej stratégie RFM, ktorá sleduje 3 typy ukazovateľov [9]:

- **Aktuálnosť** (anglický výraz: *Recency*) – Črty sledujúce, koľko času uplynulo od istej udalosti. Príkladom takýchto čŕt sú novovytvorené črty, ktoré sledujú: počet dní od poslednej správy klienta, že si chcel niečo zakúpiť a počet dní od poslednej konverzácie.

- **Frekvencia** (anglický výraz: *Frequency*) – Črty sledujúce frekvenciu udalostí. Príkladom takýchto čŕt sú novovytvorené črty, ktoré sledujú: priemerný počet správ, ktorým Kate nerozumela za posledný štvrťrok, alebo priemer počtu konverzácií, v ktorých potreboval klient s niečím pomôcť.

- **Hodnota** (anglický výraz: *Monetary value*) – V pôvodnom koncepte stratégie sa táto štatistika zameriava na množstvo peňazí, ktoré klient minie. V našom prípade by sme sem mohli zaradiť ukazovatele, ako je množstvo času, ktorý klient strávi konverzáciou s Kate počas určitého časového obdobia. Príkladom takýchto čŕt sú novovytvorené črty, ktoré sledujú: koľko času strávil klient konverzáciou s Kate za predchádzajúci mesiac alebo rok.

3.1.1 Embedding

Embedding je technika používaná v strojovom učení na reprezentáciu slov respektíve textových reťazcov vo forme numerických vektorov v konečno rozmernom priestore. Tento prístup má ambíciu zachytiť syntaktické a sémantické vzťahy medzi slovami a výrazmi a reprezentovať ich vektormi, ktoré sú v priestore blízko seba [10].

Matematicky môžeme chápať ako postup, že pre každé slovo alebo diskrétny objekt w z množiny $\mathcal{V}$ (slovnej zásoby) je mapované na vektor $\mathbf{v}_w \in \mathbb{R}^d$, kde d je rozmer embeddingového priestoru. Formálne:

$$\mathbf{v}_w = \text{embedding}(w), \quad \mathbf{v}_w \in \mathbb{R}^d. \tag{31}$$

Embedding umožňuje:

- kompaktnú reprezentáciu slov aj s nízkou dimenziou,
- zachytenie sémantických vzťahov (napr. $\mathbf{v}_{\text{kráľ}} - \mathbf{v}_{\text{muž}} + \mathbf{v}_{\text{žena}} \approx \mathbf{v}_{\text{kráľovná}}$).

Embedding slov môžeme zakategorizovať do dvoch hlavných skupín a to na: embedding založený na frekvencii výrazov (anglický výraz: *frequency based embedding*) a embedding založený na predikcii (anglický výraz: *prediction based embedding*). V rámci tvorby čŕt z konverzácií s Kate sme používali model z druhej spomínanej skupiny.

Modely z tejto skupiny vznikajú ako vedľajší produkt predikčných modelov, najčastejšie neurónových sietí, ktoré riešia zdanlivo umelo vytvorené problémy. Príkladom takéhoto problému je predikcia nasledujúceho slova v texte na základe predchádzajúcich slov. V princípe ide o úlohu klasifikácie, pretože slovník má konečný počet možných slov.

Po natrénovaní takejto neurónovej siete sa ukázalo, že váhy vychádzajúce z neurónov na vstupnej vrstve alebo vchádzajúce do neurónov na výstupnej vrstve dokážu dobre numericky reprezentovať význam a sémantiku týchto slov v prirodzenom jazyku.

3.1.2 FastText: Embeddingový model

FastText je model navrhnutý spoločnosťou Facebook AI Research [11], ktorý rozširuje Word2Vec [12] model od spoločnosti Alphabet Inc. a zohľadňuje pod-slová (anglický výraz: *sub-words*). Pri vytváraní nových čŕt klientov sme použili už natrénovaný model v

anglickom jazyku zo zdroja [13]. Tento embedingový model reprezentuje každé slovo vo vektorovom priestore s dimenziou 300.

FastText pracuje s pod-slovami namiesto celých slov. Každé slovo je rozdelené na pod-slová a embedding každého slova je reprezentovaný ako súčet vektorov pod-slov vydelený počtom pod-slov [11].

$$\mathbf{v}_w = \frac{1}{|\mathcal{G}_w|} \sum_{g \in \mathcal{G}_w} \mathbf{v}_g, \tag{32}$$

kde $\mathcal{G}_w$ je množina pod-slov slova w, a $\mathbf{v}_g$ je embedding vektor pod-slova g.

Výhodou použitia práve tohto modelu bola jeho voľná dostupnosť, pričom bol natrénovaný aj na veľkej dátovej sade. Z dôvodu, že model je trénovaný na pod-slovách, tak v princípe nemôže vypadnúť zo slovnej zásoby, čo je možné v prípade modelu Word2Vec.

3.2 Časové rozpätie modelu - tvorba závislej premennej

Ako sme načrtli v práci v kapitole 1, pri tvorbe predikčného modelu v reálnom svete dochádza k tomu, že agregované črty klientov sú vypočítané v časových rezoch. V našom prípade ide o mesačné časové rezy. Na druhej strane máme presný časový záznam o nastatí udalostí. Ak takéto dáta spojíme do jednej dátovej sady, ktorú použijeme na tvorbu predikčného modelu, samotný model bude odpovedať na otázku, v akom intervale nastane udalosť. Preto môžeme definovať závislú premennú nasledovne:

$$Y_{L,U} = P\left(\varepsilon \in [L, U] \mid \mathbf{x}_j\right) = f(\mathbf{x}_j), \tag{33}$$

kde:

- ε je modelovaná udalosť (čas odchodu klienta), pričom $\varepsilon \in \mathbb{R}^+$,
- $\mathbf{x}_j$ je vybraný vektor čŕt, ktorý je použitý na modelovanie udalosti,
- j je časová jednotka reprezentujúca rez, v ktorom sú vypočítané črty klienta, ktorými predikujeme jeho správanie, pričom $j \in \mathbb{N}$,
- $f(\mathbf{x}_j)$ je klasifikátor pri použití hodnôt čŕt klienta $\mathbf{x}_j$ z časového rezu j,
- $[L, U]$ je interval, v ktorom sa predikuje udalosť,
- U je horná hranica nastatia udalosti, pričom $U \in \mathbb{R}$,

- L je dolná hranica nastatia udalosti, pričom $L < U$ a $L \in \mathbb{R}$.

Zjednodušene povedané, vyjadrenie (33) opisuje, že pravdepodobnosť odchodu klienta v intervale $[L, U]$ je modelovaná pomocou klasifikátora $f(\mathbf{x_j})$, ktorý využíva vektor čŕt klienta $\mathbf{x}_j$ vypočítaný v j-tom časovom reze, pričom v našom prípade to je koniec mesiaca.

Hornú hranicu nastatia udalosti môžeme zapísať ako:

$$U = \mu_j + H_t, \tag{34}$$

pričom H_t reprezentuje časový horizont predikcie. Parameter t určuje vzdialenosť horizontu predikcie, pričom v práci sa venujeme len celočíselným hodnotám, ktoré predstavujú maximálnu časovú vzdialenosť od času výpočtu vektora čŕt k času nastatia udalosti v mesiacoch. μ_j reprezentuje čas, ktorý prešiel do časového rezu j.

Spodnú hranicu intervalu predikcie môžeme určiť pomocou veľkosti predikčného okna ${}_m\delta_t$:

$$L = U - {}_m\delta_t \tag{35}$$

Veľkosť predikčného okna môžeme vyjadriť ako rozdiel hraníc intervalu: ${}_m\delta_t = U - L$, pričom ľavý dolný index m reprezentuje to, že hodnota je v mesiacoch. V práci používame ako hlavnú časovú jednotku jeden mesiac. V prípade, že v niektorých častiach práce túto hodnotu prevedieme do inej časovej jednotky, dôkladne na to upozorníme.

Definícia závislej premennej v tvare (33) nám umožňuje zakategorizovať spôsob tvorby závislej premennej do dvoch skupín na základe formy veľkosti rozpätia predikcie ${}_m\delta_t$. Prvým spôsobom je predikcia s konštantne veľkým časovým rozpätím predikcie. Tento spôsob nastáva v prípade, že rozdiel funkcií časových hraníc nastatia udalosti je konštanta: ${}_m\delta = c$, pričom konštanta $c \in \mathbb{R}^+$. Druhým spôsobom je predikcia s kontinuálne rozširujúcim sa rozpätím predikcie. Tento spôsob vzniká v prípade, že rozdiel funkcií časových hraníc nastatia udalostí je výsledok závislý od času vstupujúceho do horizontu predikcie: ${}_m\delta_t = C(t)$, pričom $\mathrm{H}(C(t)) \in \mathbb{R}^+$.

Nastavenie parametra t pri tvorbe závislej premennej ovplyvňuje časový horizont predikcie. Spoločnosť ČSOB aktívne modeluje správanie svojich klientov, pričom pre každý obchodný segment aplikuje špecifický spôsob nastavenia príslušných funkcií a parametra od čoho závisí závislá premenná. Tento spôsob môžeme zaradiť do druhej spomínanej skupiny spôsobov tvorby závislej premennej.

V rámci práce sme sa venovali aj druhej kategórii spôsobov tvorby závislej premennej – predikciám s konštantnou veľkosťou časového rozpätia predikcie. Predikčné modely vytvorené na základe dátovej sady, ktorá využíva tento spôsob tvorby pozitívnej vzorky, poskytujú predikciu s vyššou časovou presnosťou nastatia udalosti. Zároveň však kladú prísnejšie nároky na históriu dát o klientoch, čo môže byť problém pri predikciách na dlhý čas do budúcna.

Črty klientov, ku ktorým máme prístup, pochádzajú z obdobia od januára 2021 až do augusta 2024. Udalosťou, ktorou sa v práci zaoberáme je odchodovosť klientov – fyzických osôb. Otázkou, ktorá sa pri analýze týchto dát prirodzene vynára je, akú históriu udalostí zohľadniť pri vytváraní modelov. S väčšou históriou je síce k dispozícii viac dát, čo znamená aj viac udalostí, ale zároveň môže vzniknúť riziko, že model bude príliš prispôsobený historickým vzorcom správania, ktoré už nemusia byť aktuálne.

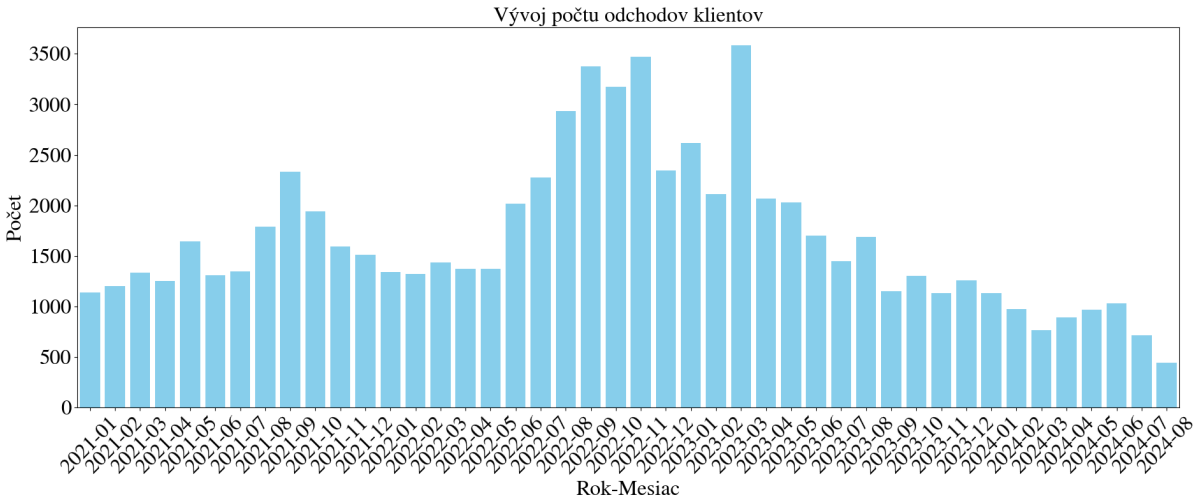
Z tohto dôvodu v práci budeme skúmať tri prípady: prvý, ktorý zohľadňuje históriu udalostí za posledných 36 mesiacov, druhý prípad zahŕňa udalosti z posledných 24 mesiacov a posledný prípad používa na tvorbu predikčného modelu za posledných 12 mesiacov.

Na obrázku č. 5 je znázornený vývoj počtu odchodov klientov v sledovanom období. Z grafu je možné pozorovať, že v roku 2022 dochádzalo k výraznému nárastu v počte odchádzajúcich klientov. Tento trend môže byť dôsledkom viacerých faktorov, pričom jedným z kľúčových aspektov bola akvizícia OTP Banky spoločnosťou ČSOB.

Súkromné inštitúcie sa usilujú o kontinuálny rast a expanziu svojho trhového podielu. V rámci tejto stratégie ČSOB uskutočnila akvizíciu OTP Banky, čím si rozšírila svoje klientské portfólio. Takýto krok však nevyhnutne prináša aj viaceré výzvy v rámci dátoveho spracovania. Medzi ne patrí napríklad skutočnosť, že časť klientov pôvodnej OTP Banky nemusí byť ochotná pokračovať v spolupráci s novou bankou, ktorá prevzala správu nad ich finančnými produktami. Okrem toho je potrebné vykonať proces optimalizácie klientského portfólia, ktorý zahŕňa identifikáciu a prípadné vyradenie klientov, ktorí nespĺňajú aktualizované kritériá a požiadavky banky na správu produktov.

Z uvedeného obrázku je zároveň zrejmé, že analyzované dáta nie sú stacionárne, ale podliehajú vplyvom interných ale aj externých faktorov. Tieto faktory môžu byť dôsledkom interných rozhodnutí a strategických krokov banky, ktoré nemajú priamy a

okamžitý dopad na individuálnych klientov, avšak v širšom kontexte významne ovplyvňujú dynamiku ich správania, ako môže byť nová stratégia smerovania spoločnosti.



Obr. 5: Graf s vývojom počtu odchodov klientov podľa mesiaca a roka. *Zdroj: Vlastné spracovanie.*

V tejto kapitole podrobne analyzujeme základné spôsoby tvorby závislej premennej v predikčnom modeli. Pre každý z prezentovaných prístupov predstavíme jeho potenciálne výhody a nevýhody. Definícia a konštrukcia závislej premennej predstavuje kľúčový krok v procese tvorby predikčného modelu, pretože priamo ovplyvňuje interpretáciu výsledkov modelovania. Formulácia otázky, na ktorú má model poskytnúť odpoveď, vyplýva práve z tvorby závislej premennej. Tieto dáta následne tvoria pozitívnu vzorku v dátovej sade pre tvorbu predikčného modelu.

V tejto práci používame ako časovú jednotku mesiac. Pre jednoduchosť zápisu v prípade potreby prevodu počtu mesiacov na počet dní alebo naopak, pracujeme v rámci práce s konštantnou hodnotou že jeden mesiac má 30 dní.

3.2.1 Kontinuálne zväčšujúce sa rozpätie predikcie

Názov tejto kategórie spôsobov tvorby závislej premennej je odvodený od skutočnosti, že so zväčšujúcim sa horizontom predikcie H_t sa zároveň rozširuje aj celkové rozpätie predikcie ${}_m\delta_t$

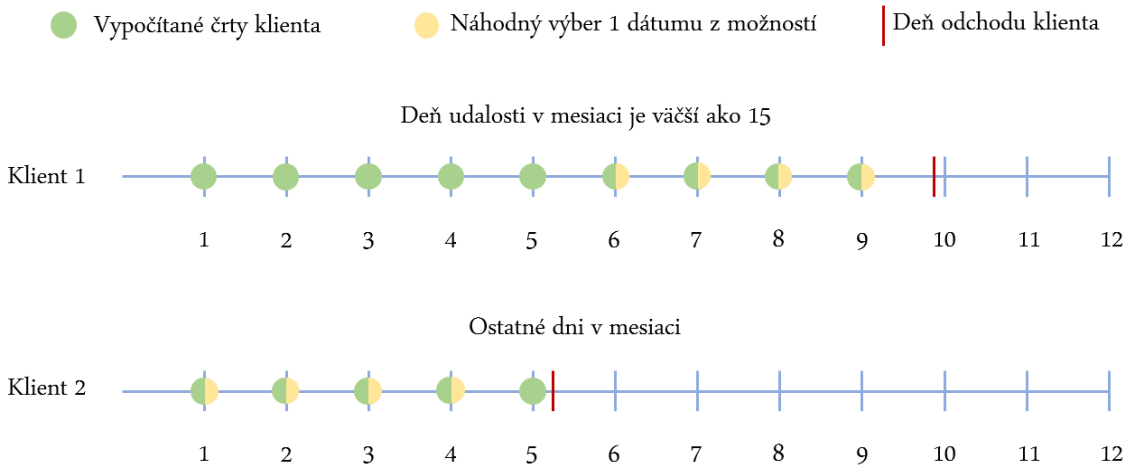
Aktuálny spôsob k tvorbe závislej premennej využíva tento typ tvorby závislej premennej. Rozlišuje dva scenáre na základe dňa v mesiaci, kedy došlo k sledovanej

udalosti.

Ak deň, v ktorý udalosť nastala presahuje hodnotu 14, do pozitívnej vzorky sa náhodne vyberie jeden vektor čŕt vypočítaný z t najbližších koncomesačných dátumov k dátumu vzniku udalosti. V prípade ostatných dátumov sa náhodne vyberie tiež jeden vektor čŕt klienta z t koncomesačných dátumov, ktoré ale sú najbližšie k dátumu udalosti zníženého o jeden mesiac.

V prípade modelov, ktoré odhadujú odchodovosť klientov je parameter t nastavený momentálne na hodnotu 3, pri tvorbe ostatných predikčných modelov je tento parameter nastavený na hodnotu 2. Pri tomto nastavení parametra t modely odchodovosti odpovedajú na otázku: *Mohol by klient odísť v najbližších 15 až 104 dňoch?*. Ostatné modely, ktoré odhadujú napríklad kúpu produktu, majú menší horizont predikcie a teda odpovedajú na otázku: *Mohol by si klient kúpiť produkt v najbližších 15 až 74 dňoch?*

Pre lepšie pochopenie uvádzame grafické znázornenie spôsobu tvorby závislej premennej pre model odchodovosti a dvoch klientov na obrázku č. 6, pri nastavení horizontu predikcie s $t = 4$. Prvý klient odišiel z banky po 15. dni v mesiaci, druhý pred 15. dňom v mesiaci. V prípade tohto nastavenia závislej premennej by sme do pozitívnej vzorky zaradili jeden zo štyroch vektorov čŕt klienta vybraný náhodne pre každého klienta.



Obr. 6: Obrázok zobrazuje spôsob tvorby závislej premennej pomocou kontinuálne zväčšujúcemu sa rozpätiu predikcie. Na obrázku je zobrazená časová os, dátumy v ktorých sú vypočítané vektory čŕt klienta, dátum odchodu klienta a možnosti pre náhodný výber čŕt klienta vstupujúcich do pozitívnej vzorky. *Zdroj: Vlastné spracovanie.*

Ako môžeme vidieť z obrázku č. 6 pozorovania v dátovej sade môžu mať rôzne rozdiely v dĺžke času medzi dátumom odchodu a dátumom vybraných čŕt klientov. Preto výsledný natrénovaný predikčný model bude predikovať odchod klienta v nasledovnom tvare:

$$Y_{L,U} = P\left(\varepsilon \in [L, U] \mid \mathbf{x}_j\right) = f(\mathbf{x}_j), \tag{36}$$

pričom:

- $U = \mu_j + t + 0.5$ je horná hranica intervalu predikcie,
- ${}_m\delta_t = t$ je veľkosť rozpätia predikcie v mesiacoch,
- $L = U - {}_m\delta_t = \mu_j + 0.5$ je spodná hranica predikcie,
- $\mathbf{x}_j$ je vektor čŕt vypočítaný v časovom reze j.

Horizont $H_t = t + 0.5$ vznikol na základe maximálneho rozdielu medzi časom výpočtu vektora čŕt klienta a časom nastatia jeho odchodu. V prípade použitia parametra horizontu $t = 4$, tento rozdiel môže byť až 4.5 mesiaca.

Dolná hranica $L = \mu_j + 0.5$ vznikla na základe toho, že minimálny rozdiel medzi časom výpočtu vektora čŕt klienta a časom jeho odchodu v dátovej sade môže byť 15 dní, čo reprezentuje pol mesiaca od dátumu výpočtu vektora čŕt použitého na predikciu.

Veľkosť predikčného okna je v tomto prípade tvorby závislej premennej ${}_m\delta_t = t$. Teda so zväčšujúcim sa časovým parametrom horizontu predikcie sa bude taktiež zväčšovať predikčné okno o tú istú hodnotu.

Predikčné modely, pri ktorých bol použitý spôsob tvorby závislej premennej pomocou kontinuálne sa zväčšujúceho rozpätia predikcie, budeme v ďalšom texte označovať ako:

$$\mathrm{CONT}_t, \tag{37}$$

pri danej hodnote parametra t. V práci budeme skúmať vplyv zmeny parametra t na kvalitu modelov a následne porovnávať tieto výsledky s alternatívnymi prístupmi tvorby pozitívnej vzorky pri rôznych dĺžkach predikcií, ktoré závisia od hodnoty parametra t v procese tvorby závislej premennej.

3.2.2 Konštantne veľké časové rozpätie predikcie

Ako sme uviedli v predchádzajúcej podkapitole č. 3.2.1, veľkosť predikčného rozpätia závisí od parametra horizontu predikcie t. Pri tvorbe závislej premennej s použitím konštantne veľkého časového rozpätia predikcie však veľkosť tohto rozpätia nezávisí od t vstupujúceho do horizontu predikcie. Rozdiel medzi hraničnými funkciami predikčného intervalu teda bude konštanta: $U - L = c$.

Použitie tvorby závislej premennej s konštantne veľkým časovým rozpätím predikcie znižuje neistotu ohľadom času, kedy nastane udalosť pri predikciách smerujúcich ďaleko do budúcnosti. Na druhej strane, aby sme takéto spôsoby mohli aplikovať, musíme disponovať dostatočne dlhými časovými radmi vektorov čŕt klientov. Tieto spôsoby sú zároveň striktnejšie pri výbere dátumov pre črty klienta, čo môže viesť k zmenšovaniu pozitívnej vzorky.

V rámci tejto časti práce sme navrhli 5 spôsobov tvorby závislej premennej s tromi rôznymi veľkosťami rozpätia predikcie. Tieto spôsoby zabezpečujú výrazne vyššiu presnosť, kedy udalosť nastane pri veľkých časových horizontoch, na rozdiel od spôsobu s kontinuálne sa zväčšujúcim rozpätím predikcie.

V práci budeme modely vytvorené na základe dátovej sady, v ktorej bolo použitá tvorba závislej premennej s konštantne veľkým časovým rozpätím predikcie, označovať ako:

$$\mathrm{FIX}_{{}_d\delta,\, n,\, t}, \tag{38}$$

pri danej hodnote parametra t, kde:

- ${}_d\delta$ predstavuje hodnotu veľkosti rozpätia predikcie v dňoch: $L - U$,
- n je počet vektorov čŕt klienta, ktoré vstupujú do pozitívnej vzorky.

V ďalších častiach tejto kapitoly podrobne opíšeme jednotlivé spôsoby tvorby závislej premennej. Taktiež zachováme rozdelenie tvorby závislej premennej na dva prípady vzhľadom na deň vzniku udalosti, tak ako to bolo v prípade spôsobu CONT_t

$\mathbf{FIX}_{30,1,t}$

Ako prvý spôsob s konštantne veľkým časovým rozpätím definujeme spôsob tvorby závislej premennej, pri ktorom je dĺžka časového rozpätia predikcie stanovená na jeden mesiac,

teda 30 dní.

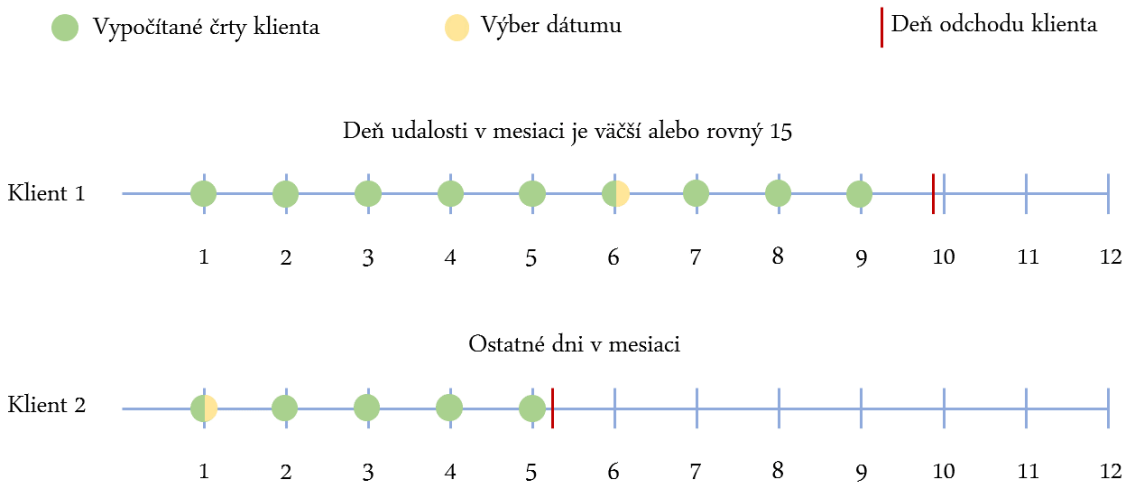
V prípade modelov, ktoré využívajú tvorbu závislej premennej s kontinuálne zväčšujúcim sa časovým rozpätím predikcie sa dátum vektora čŕt charakterizujúci nastatie udalosti vyberie náhodne jeden z t dátumov, ktoré sú najbližšie ku dňu vzniku udalosti u klienta. Výber konkrétnych dátumov závisí od dňa v mesiaci, kedy k udalosti došlo.

Spôsob $\mathbf{FIX}_{30,1,t}$ zachováva rovnaké rozdelenie na dva prípady na základe dňa v mesiaci vzniku udalosti, avšak na rozdiel od náhodného výberu dátumu z t možností, ako je to pri predchádzajúcom prístupe, spôsob $\mathrm{FIX}_{30,1,t}$ vždy volí najstarší dátum. Týmto spôsobom je pozitívna vzorka čŕt klientov vytvorená na základe striktnejšieho pravidla určujúceho rozdiel medzi dátumom vzniku udalosti a dátumom vektora čŕt, ktorý je vybraný pre modelovanie udalosti.

Ak by bol parameter t nastavený na hodnotu 4, model odchodovosti by odpovedal na otázku: *Mohol by klient odísť v najbližších 105 až 134 dňoch?* Je zrejmé, že neistota ohľadom času nastatia udalosti, ktorú model poskytuje, je výrazne nižšia v porovnaní s spôsobom $\mathbf{CONT}_t$.

Na druhej strane, prísnym výberom dátumu sme do značnej miery eliminovali náhodnosť pri tvorbe pozitívnej vzorky. Táto náhodnosť však môže byť do istej miery výhodná, pretože pri použití len najstaršej črty klienta pre daný horizont predikcie nemusí byť klient správne charakterizovaný, ak jeho rozhodnutie o odchode vzniká po alebo pred týmto dátumom výpočtu vektora čŕt.

Aj v tomto prípade uvádzame grafický príklad na obrázku č. 7, kde je parameter t nastavený na hodnotu 4. Na časovej osi sú zobrazení dvaja klienti a dostupné dáta ich vypočítaných vektorov čŕt, vyznačené zelenou farbou. Bordovou farbou je zobrazený ich odchod. Žltou farbou je znázornený výber dátumu pre vektor čŕt klienta, ktorý bude vstupovať do pozitívnej vzorky.



Obr. 7: Obrázok zobrazuje spôsob tvorby závislej premennej $\text{FIX}_{30,1,t}$. Na obrázku je zobrazená časová os, dátumy v ktorých sú vypočítané vektory čŕt klienta, dátum odchodu klienta a vybratý dátum do pozitívnej vzorky. *Zdroj: Vlastné spracovanie.*

Oproti spôsobu tvorby závislej premennej CONT uvedenému v podkapitole 3.2.1 je tento spôsob striktnejší. V prípade, že klient nebol klientom ČSOB v danom dátume, strácame informáciu o tomto klientovi, čo vedie k zmenšeniu pozitívnej vzorky.

Ako môžeme vidieť z obrázku č. 6 v pozorovaniach v dátovej sade môžu mať rôzne rozdiely v dĺžke času medzi dátumom odchodu a dátumom výpočtu vektora čŕt klientov. Preto výsledný natrénovaný predikčný model bude predikovať odchod klienta v nasledovnom tvare:

$$Y_{L,U} = P\left(\varepsilon \in [L, U] \mid \mathbf{x}_j\right) = f(\mathbf{x}_j), \tag{39}$$

pričom:

- $U = \mu_j + t + 0.5$ je horná hranica intervalu predikcie,
- ${}_m\delta = 1$ je veľkosť rozpätia predikcie v mesiacoch, pričom v dňoch je táto hodnota 30,
- $L = U - {}_m\delta = \mu_j + t - 0.5$ je spodná hranica predikcie,
- $\mathbf{x}_j$ je vektor čŕt vypočítaný v čase j.

Horizont $H_t = t + 0.5$ vznikol na základe maximálneho rozdielu medzi časom výpočtu možného vybraného vektora čŕt klienta a časom nastatia jeho odchodu. V prípade použitia parametra horizontu $t = 4$, tento rozdiel môže byť až 4.5 mesiaca.

Dolná hranica $L = \mu_j + t - 0.5$ vznikla na základe toho, že minimálny rozdiel medzi časom výpočtu vektora čŕt klienta a jeho odchodom v dátovej sade môže byť minimálne $t - 0.5$ mesiaca. j je časový rez výpočtu vektora čŕt použitý na modelovanie.

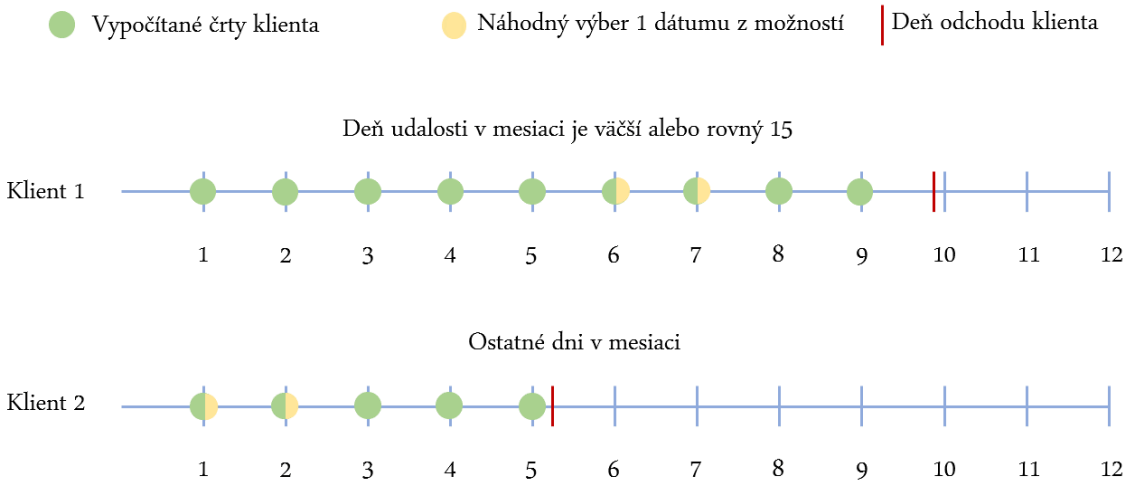
V ďalších častiach práce vysvetlíme použité spôsoby tvorby závislej premennej s väčšími rozpätiami predikcie.

$\mathbf{FIX}_{60,1,t}$ a $\mathbf{FIX}_{60,2,t}$

Ako ďalšie spôsoby tvorby závislej premennej uvedieme tie, ktoré majú veľkosť rozpätia predikcie 60 dní. Pri tomto spôsobe tvorby závislej premennej sa vektory čŕt klienta vyberajú z dvoch možných koncomesačných dátumov, ktoré predchádzajú dňu vzniku udalosti o t a $t-1$ mesiacov v prípade, že deň udalosti v mesiaci je viac ako 14. V prípade ostatných dní sa vektory čŕt vyberajú z koncomesačných dátumov, ktoré predchádzajú $t+1$ a t mesiacov k dňu vzniku udalosti.

Označenie $\mathbf{FIX}_{60,1,t}$ bude označovať modely, pri ktorých bude použitý len jeden vektor čŕt klienta v rámci tvorby pozitívnej vzorky. Tento vektor sa náhodne vyberie z dvoch spomínaných možností. $\mathbf{FIX}_{60,2,t}$ bude označenie pre modely, u ktorých boli použité oba vektory čŕt klientov ak boli dostupné.

Pre lepšiu ilustráciu spôsobu $\mathrm{FIX}_{60,1,t}$ uvádzame príklad, ako je zobrazený na obrázku č. 8. V prípade nastavenia horizontu predikcie s parametrom $t = 4$ dochádza k výberu jedného dátumu pre vektor čŕt klienta, pričom najstarší dátum vektora čŕt je vzdialený maximálne 4.5 mesiace od dátumu nastatia udalosti.



Obr. 8: Obrázok zobrazuje spôsob tvorby závislej premennej $\text{FIX}_{60,1,t}$. Na obrázku je zobrazená časová os, dátumy v ktorých sú vypočítané vektory čŕt klienta, dátum odchodu klienta a možnosti dátumov do pozitívnej vzorky. *Zdroj: Vlastné spracovanie.*

Pre spôsob tvorby závislej premennej $\mathbf{FIX}_{60,2,t}$ by bol rozdiel oproti obrázku č. 8 v tom, že by nedochádzalo k náhodnému výberu, ale vždy by sa vybrali oba dátumy, ak sú dostupné. Zároveň je potrebné dať na zreteľ, že tento spôsob umožňuje zväčšiť pozitívnu vzorku pridaním viacerých pozitívnych vektorov čŕt klienta, ktoré pochádzajú z rôznych dátumov. Tento prístup by mohol mať pozitívny vplyv na kvalitu modelu, najmä v prípadoch, keď sú dáta výrazne nevyvážené. Na druhej strane môže tento postup viesť aj k negatívnym dopadom, konkrétne k zvýšenej náchylnosti modelu na pretrénovanie. Preto je dôležité dôkladne zvážiť nastavenie parametrov a vyhodnotiť ich vplyv na výkonnosť modelu.

V prípade použitia tvorby závislej premennej spôsobom $\text{FIX}_{60,1,t}$ by bola modelovaná pravdepodobnosť v nasledovnom tvare:

$$Y_{L,U} = P\left(\varepsilon \in [L, U] \mid \mathbf{x}_j\right) = f(\mathbf{x}_j), \tag{40}$$

pričom:

- $U = \mu_j + t + 0.5$ je horná hranica intervalu predikcie,
- ${}_m\delta = 2$ je veľkosť rozpätia predikcie v mesiacoch, pričom v dňoch je táto hodnota 60,

- $L = U - {}_m\delta = \mu_j + t - 1.5$ je spodná hranica predikcie,
- $\mathbf{x}_j$ je vektor čŕt vypočítaný v časovom reze j.

Horizont $H_t = t + 0.5$ vznikol na základe maximálneho rozdielu medzi časom možného vybraného vektora čŕt klienta a časom nastatia jeho odchodu. V prípade použitia parametra horizontu $t = 4$, tento rozdiel môže byť až 4.5 mesiaca.

Dolná hranica $L = \mu_j + t - 1.5$ vznikla na základe toho, že minimálny rozdiel medzi časom výpočtu čŕt klienta a jeho odchodom v dátovej sade môže byť minimálne $t - 1.5$ mesiaca. j je časový rez výpočtu vektora čŕt použitý na modelovanie.

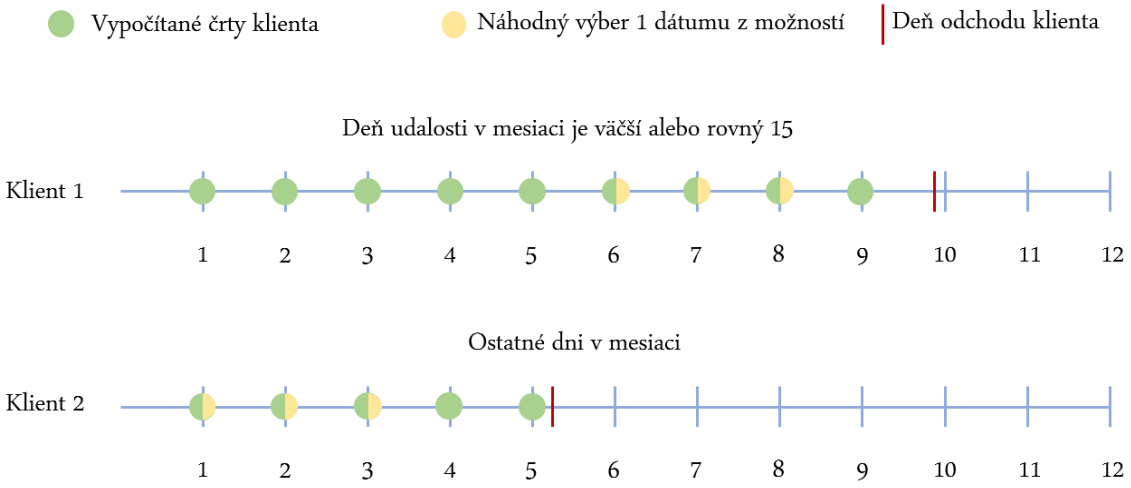
Pri tvorbe závislej premennej $\mathbf{FIX}_{60,2,t}$ platí rovnaká definícia pre predikčný model.

$\mathbf{FIX}_{90,1,t}$ a $\mathbf{FIX}_{90,2,t}$

Ako ďalšie spôsoby tvorby závislej premennej uvádzame tie, ktoré majú veľkosť rozpätia predikcie 3 mesiace a teda 90 dní. Pri tomto spôsobe tvorby závislej premennej sa vektory čŕt klienta vyberajú z troch možných koncomesačných dátumov, ktoré predchádzajú dňu vzniku udalosti o $t - 2$ až t mesiacov v prípade, že deň udalosti v mesiaci je viac ako 14. V prípade ostatných dní sa vektory čŕt vyberajú z koncomesačných dátumov, ktoré predchádzajú k času nastatia udalosti o $t - 1$ až $t + 1$ mesiacov.

Označenie $\mathbf{FIX}_{90,1,t}$ bude označovať modely, pri ktorých bude použitý len jeden vektor čŕt klienta v rámci tvorby pozitívnej vzorky. Tento vektor sa náhodne vyberie z troch spomínaných možností. $\mathbf{FIX}_{90,2,t}$ bude označenie pre modely, u ktorých dôjde k náhodnému výberu bez vrátenia 2 vektorov čŕt klientov.

Pre lepšiu ilustráciu spôsobu $\mathbf{FIX}_{90,1,t}$ uvádzame príklad, ktorý je zobrazený na obrázku č. 9. V prípade nastavenia parametra horizontu predikcie $t = 4$. Pri výbere dátumov vektorov čŕt klientov by dochádzalo k náhodnému výberu jedného dátumu z možných troch možností, pričom najstarší dátum je vzdialený maximálne 4.5 mesiaca od dátumu nastatia udalosti.



Obr. 9: Obrázok zobrazuje spôsob tvorby závislej premennej $\text{FIX}_{90,1,t}$. Na obrázku je zobrazená časová os, dátumy v ktorých sú vypočítane vektory čŕt klienta, dátum odchodu klienta a možnosti dátumov do pozitívnej vzorky. *Zdroj: Vlastné spracovanie.*

V prípade použitia tvorby závislej premennej spôsobom $\text{FIX}_{90,1,t}$ by bola modelovaná pravdepodobnosť v nasledovnom tvare:

$$Y_{L,U} = P\left(\varepsilon \in [L, U] \mid \mathbf{x}_j\right) = f(\mathbf{x}_j), \tag{41}$$

pričom:

- $U = \mu_j + t + 0.5$ je horná hranica intervalu predikcie,
- ${}_m\delta = 3$ je veľkosť rozpätia predikcie v mesiacoch, pričom v dňoch je táto hodnota 90,
- $L = U - {}_m\delta = \mu_j + t - 2.5$ je spodná hranica predikcie,
- $\mathbf{x}_j$ je vektor čŕt vypočítaný v čase j.

Horizont $H_t = t + 0.5$ vznikol na základe maximálneho rozdielu medzi časom možného vybraného vektora čŕt klienta a časom nastatia jeho odchodu. V prípade použitia parametra horizontu $t = 4$, tento rozdiel môže byť až 4.5 mesiaca.

Dolná hranica $L = \mu_j + t - 2.5$ vznikla na základe toho, že minimálny rozdiel medzi časom výpočtu vektora čŕt klienta a jeho odchodom v dátovej sade môže byť minimálne $t - 2.5$ mesiaca. j je časový rez výpočtu vektora čŕt použitý na modelovanie.

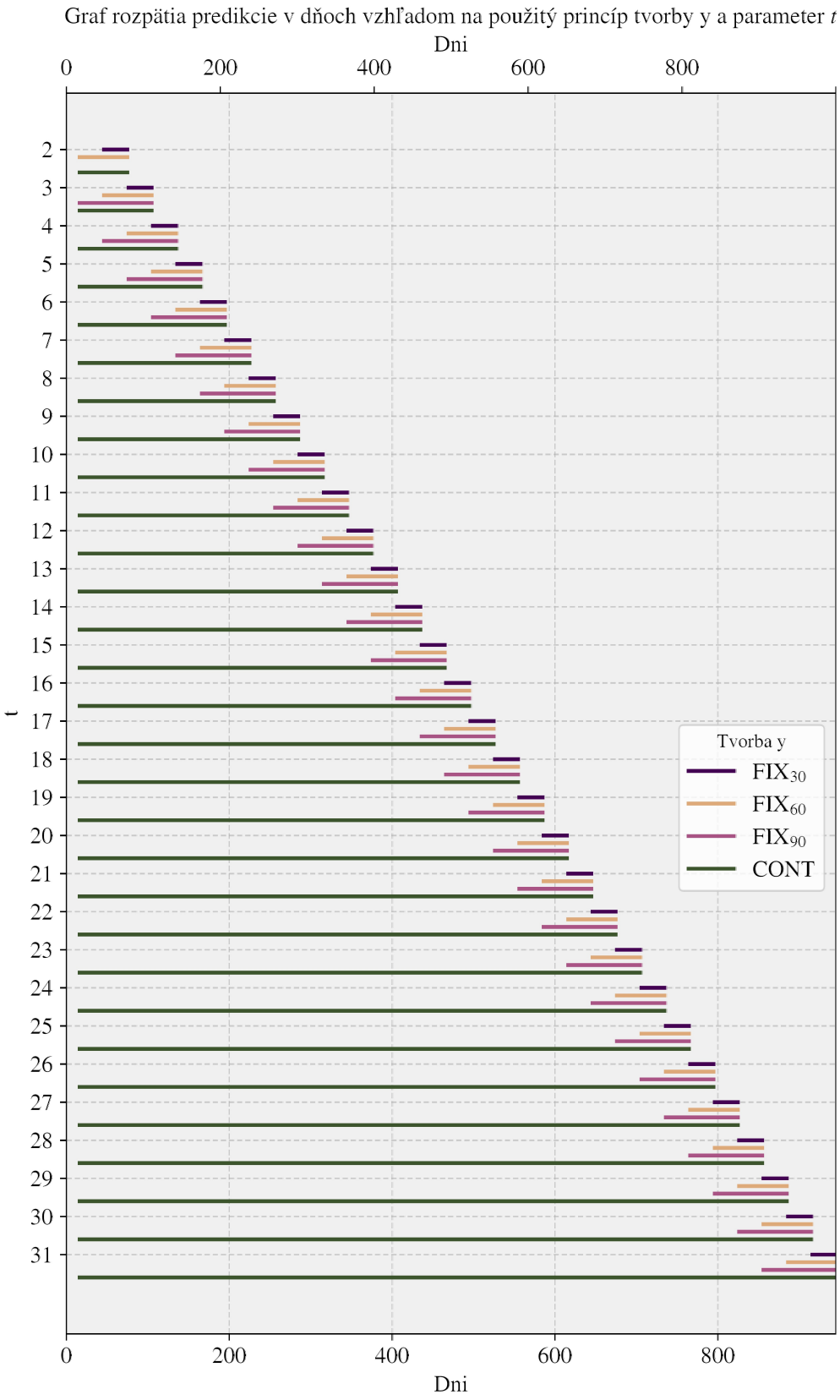
Pri tvorbe závislej premennej $\mathbf{FIX}_{90,2,t}$ platí rovnaká definícia.

3.2.3 Porovnanie časového rozpätia predikčných modelov

Vo vyššie uvedených podkapitolách sme uviedli spôsoby tvorby závislej premennej, ktoré budeme používať v práci. Pre lepšiu predstavu, uvádzame na obrázku č. 10 grafické znázornenie v akom časovom rozpätí modely predikujú nastanie udalosti, vzhľadom na hodnotu parametra t určujúceho horizont predikcie a použitia spôsobu tvorby závislej premennej.

Do pozornosti dávame, že spôsoby používané v tejto práci, ktoré umožňujú využitie viacerých vektorov čŕt klienta v pozitívnej vzorke, nemenia veľkosť rozpätia predikcie. Tieto spôsoby majú rovnaké predikčné okno. Z tohto dôvodu v grafickom znázornení, uvedenom nižšie, označujeme spôsoby $\mathbf{FIX}_{60,1,t}$ a $\mathbf{FIX}_{60,2,t}$ jednotne ako $\mathbf{FIX}_{60}$, pričom analogicky postupujeme pre spôsoby $\mathbf{FIX}_{90,1,t}$ a $\mathbf{FIX}_{90,2,t}$, ktoré značíme ako $\mathbf{FIX}_{90}$.

Ako je možné pozorovať z obrázka, predikčné horizonty pri krátkodobých predikciách sa značne nelíšia. Avšak pri strednodobých a dlhodobých predikciách sú rozdiely výrazné. Pri spôsobe tvorby závislej premennej CONT_t je neistota ohľadom presného času nastatia udalosti oveľa väčšia, v porovnaní so spôsobmi tvorby závislej premennej s konštantným časovým rozpätím.



Obr. 10: Graf zobrazuje predikované rozpätie modelu vzhľadom na hodnotu parametra t a spôsobu tvorby závislej premennej. *Zdroj: Vlastné spracovanie.*

3.3 Doplňenie dátovej sady o negatívnu vzorku

Negatívna vzorka predstavuje doplnok dátovej sady, ktorá sa používa na trénovanie a validáciu modelu strojového učenia.

V prípade pozitívnej vzorky má závislá premenná hodnotu 1, čo znamená, že u klienta nastala určitá udalosť. Naopak, negatívna vzorka obsahuje vektory čŕt klientov, u ktorých daná udalosť nenastala a preto má závislá premenná hodnotu 0. Tento rozdiel medzi pozitívnou a negatívnou vzorkou je kľúčový pri tvorbe binárnych klasifikačných modelov, kde je cieľom naučiť model rozlišovať medzi týmito dvoma populáciami.

Model by zároveň mal byť schopný čo najpresnejšie predikovať výsledky na nových, doposiaľ nevidených dátach. Ako sme uviedli na začiatku kapitoly č. 3.2, v tejto práci budeme porovnávať modely vytvorené na základe 36, 24, 12-mesačnej histórie udalostí. Pre všetky tri prípady určíme počet vektorov čŕt klientov v celej dátovej sade (počet dát v pozitívnej a negatívnej vzorke spoločne). V prípade 36-mesačnej histórie udalostí použijeme dátovú sadu s celkovým počtom 300 000 záznamov. Pre 24-mesačnú históriu budeme pracovať s dátovou sadou obsahujúcou 200 000 záznamov, zatiaľ čo pre 12-mesačnú históriu použijeme len 50 000 záznamov. Tento postup sme zvolili s cieľom zabezpečiť, aby pri krátkodobých predikciách zostal približne rovnaký pomer medzi počtom dát v pozitívnej vzorke a celkovým počtom dát v súbore pre všetky prípady použitej veľkosti histórie udalostí.

Pri tvorbe negatívnej vzorky sa zameriavame na zachovanie medzimesačného vývoja počtu výskytov udalostí. To znamená, že pre každý mesiac z pozitívnej vzorke určíme požadovaný celkový počet vektorov čŕt (na základe pomeru celkového požadovaného počtu vektorov čŕt a počtu mesiacov). Následne náhodne vyberieme vektory čŕt klientov, u ktorých udalosť nenastala, v takom počte koľko pozorovaní chýba do požadovanej hodnoty pre daný mesiac. Tieto vektory čŕt klientov tvoria negatívnu vzorku v dátovej sade. Tento prístup nám umožňuje zabezpečiť vyváženosť pomerov medzi pozitívnymi a negatívnymi prípadmi v rámci jednotlivých mesiacov, pričom zároveň zachováme informácie o prípadných nezvyčajných nárastoch alebo poklesoch výskytu udalosti v daných mesiacoch.

Tento prístup má za cieľ presnejšie zachytiť reálny vývoj udalostí v dátach, čím sa líši od jednoduchého náhodného výberu z celkového množstva dostupných dát.

Týmto spôsobom sa minimalizuje riziko, že črty klientov z konkrétneho mesiaca budú v dátovej sade neprimerane zastúpené, čo by mohlo viesť k nesprávnym záverom modelu spôsobeným atypickým správaním klientov v danom období.

4 Konštrukcia predikčného modelu

V predchádzajúcich kapitolách sme opísali proces tvorby dátových sád, ktoré slúžia na trénovanie modelov a porovnávanie ich kvality. V tejto kapitole sa zameriame na aspekty spojené so samotným trénovaním modelov. Predstavíme postup úpravy a selekcie čŕt klientov, ktoré vstupujú do výpočtov. Následne sa budeme venovať opisu postupu zakódovania kategorických premenných do numerických hodnôt a rozdelenia dátovej sady. V poslednej časti kapitoly opíšeme použitý postup optimalizácie hyperparametrov metódy LightGBM.

4.1 Selekcia čŕt klientov

Ako uvádzame v predchádzajúcich častiach práce, spoločnosť ČSOB sleduje viac ako 2000 rôznych čŕt pre každého klienta. Z tohto dôvodu nie je praktické pri tvorbe modelu postupovať spôsobom, ktorý by zahŕňal použitie všetkých dostupných čŕt, následné vyhodnotenie ich dôležitosti na základe určitého kritéria a následné opätovné pretrénovanie modelu po odstránení menej dôležitých čŕt. V tejto časti práce sa zameriame na opis procesu predselekcie čŕt, ktoré budú vybrané pre optimalizáciu hyperparametrov metódy LightGBM. Tento postup vychádza z metodiky spoločnosti ČSOB pre tvorbu dátovej sady použitej na tvorbu predikčného modelu.

Prvý krok

V prvom kroku sa zameriavame na hodnoty, ktoré nadobúdajú jednotlivé črty. V prípade, že hodnota črty na 0,25% a 99,75% kvantile je rovnaká, túto črtu klientov vylúčime z ďalšieho procesu tvorby modelu. Približne 450 čŕt bolo v tomto kroku vylúčených pri tvorbe každého predikčného modelu. Je potrebné zdôrazniť, že sa nemusí jednať o tie isté črty, pri každom modeli, keďže do tvorby dátovej sady, z ktorej sa model trénuje vstupuje náhodnosť a odlišné dátumy čŕt klientov pre každý parameter. Preto ide o približné číslo, ktoré naznačuje, že v používaných dátach nemusí vždy existovať veľká disperzia a samotná informácia, z ktorej by sa dalo čerpať.

Zvyšné črty prejdú druhým krokom selekcie, v ktorom vyberáme **300** čŕt klientov, ktoré sú následne použité pri optimalizácii hyperparametrov metódy a tvorbe predikčného

modelu.

Druhý krok

V druhom kroku selekcie čŕt klientov najprv vypočítame korelačnú maticu medzi črtami klientov a závislou premennou. Ak niektoré črty klientov vykazujú medzi sebou vysokú hodnotu Pearsonovej korelácie a presahujúcu nami stanovenú hraničnú hodnotu 0,9, danú črtu vylúčime z ďalšieho procesu. Črtu ktorú ponecháme z dvojice vyberieme na základe veľkosti absolutnej hodnoty korelácie, so závislou premennou.

Po odstránení redundantných čŕt vypočítame hodnotu informačného kritéria vzájomnej informácie (anglický výraz: *mutual information*) medzi jednotlivými črtami klientov a závislou premennou [14]. Na tento účel využijeme knižnicu `sklearn`, konkrétne funkciu `mutual_info_classif` [15]. Definíciu informačného kritéria vzájomnej informácie pre diskrétny prípad uvádzame v prílohe A.

Toto informačné kritérium zachytáva aj nelineárne závislosti medzi premennými. Po výpočte hodnôt vzájomnej informácie medzi črtami klientov a závislou premennou vyberieme 300 čŕt klientov s najvyššou hodnotou vzájomnej informácie na ďalšie použitie pri tvorbe predikčného modelu.

4.2 Rozdelenie dátovej sady

Pred tvorbou modelu rozdelíme dátovú sadu, ktorá obsahuje pozitívnu a negatívnu vzorku na trénovaciu (80%) a validačnú (20%) sadu. Výber dát do týchto vzoriek bude realizovaný náhodne, avšak so stratifikáciou [1], ktorá zohľadňuje pomer pozitívnych a negatívnych prípadov v celej dátovej sade. Vzhľadom na využitie optimalizácie hyperparametrov metódy LightGBM, podrobne popísanej v kapitole 4.4, nebudeme dátovú sadu ďalej deliť na testovaciu sadu.

[1] Stratifikácia je proces, pri ktorom sa zabezpečuje, že rozdelenie dát do rôznych vzoriek (napr. trénovacia, validačná) bude zachovávať rovnaký pomer kategórií (pozitívne a negatívne prípady) ako v pôvodnej dátovej sade.

4.3 Zakódovanie kategorických čŕt klientov

Črty klientov, ktoré používame nemusia byť striktne numerické; niektoré z nich sú aj kategorické. Príkladom môže byť NACE kód podniku, ktorý reprezentuje odvetvie, v ktorom podniká právna osoba. Pre zefektívnenie výpočtov sme sa rozhodli zakódovať tieto kategorické črty do numerických hodnôt. Aj keď model LightGBM dokáže pracovať s kategorickými hodnotami, pretože využíva rozhodovacie stromy ako jednoduché metódy, v našom prípade sme sa rozhodli pre numerickú reprezentáciu. Kategorické premenné môžeme zakódovať rôznymi spôsobmi. Napríklad, v prípade mesta bydliska by sme mohli využiť počet obyvateľov. V našom prípade sme však zvolili prístup, pri ktorom kategorickej premennej priradíme informáciu o početnosti výskytu udalosti (závislej premennej) v dátach.

V tejto práci budeme na účel zakódovania kategorických premenných používať metódu CatBoost [16], ktorá je jedným z modelov strojového učenia. Model natrénujeme na našej trénovacej sade dát a následne ho použijeme na priradenie hodnôt ku kategorickým premenným vo validačnej sade. Pôvodné kategorické hodnoty čŕt budú odstránené a nahradené novými numerickými hodnotami v oboch sadách, čím sa umožní ich efektívne spracovanie metódou LightGBM. Podrobnejšie informácie o tejto metóde v článku [16].

4.4 Optimalizácia hyperparametrov metódy LightGBM

Pri optimalizácii hyperparametrov metódy LightGBM na účely tvorby predikčného modelu využívame metodiku spoločnosti ČSOB. Tento prístup zahŕňa použitie knižnice `Optuna` [17] a k-násobnej krížovej validácie (anglický výraz: *k-fold cross validation*[18]). Tvorba dátovej sady a optimalizácia jedného predikčného modelu trvá približne 1.5 hodiny.

V prílohe B uvádzame základný rozdiel medzi použitím algoritmu `Optuna` v rámci hľadanie ideálnych hyperparametrov a spôsobu mriežkového vyhľadávania (anglický výraz: *grid search*).

4.4.1 Zoznam optimalizovaných hyperparametrov metódy LightGBM

Najskôr identifikujeme súbor hyperparametrov, ktoré budú predmetom optimalizácie. Nasledujúci zoznam uvádza jednotlivé hyperparametre optimalizované v rámci metódy LightGBM:

- **learning_rate**: Koeficient učenia, ktorý ovplyvňuje veľkosť kroku pri optimalizácii modelu.
- **min_child_samples**: Minimálny počet vzoriek potrebných v uzle pred jeho rozdelením v rozhodovacom strome.
- **max_depth**: Maximálna hĺbka rozhodovacieho stromu, ktorá obmedzuje počet jeho úrovní.
- **reg_lambda**: Regularizačný parameter, ktorý penalizuje veľké váhy čŕt, čím redukuje riziko pretrénovania.
- **reg_alpha**: Regularizačný parameter, ktorý indukuje riedkosť modelu tým, že vynucuje nulové váhy pri nepodstatných črtách.
- **min_split_gain**: Minimálna hodnota potrebná na rozdelenie uzla v rozhodovacom strome.
- **scale_pos_weight**: Škálovací faktor, ktorý upravuje váhu pozitívnych vzoriek, čím vyvažuje nerovnomerne distribuované triedy.
- **subsample**: Podiel vzoriek použitých pri konštrukcii jednotlivých stromov.
- **colsample_bytree**: Podiel čŕt vybraných pre trénovanie každého stromu.
- **num_leaves**: Maximálny počet listových uzlov v rozhodovacom strome.
- **n_features**: Počet vybraných čŕt z pôvodnej dátovej množiny použitých na trénovanie modelu. Tento parameter nie je natívny pre LightGBM, ale je zavedený na redukciu dimenzionality vstupných dát.

4.4.2 Použitie k-násobnej krížovej validácie

Pri optimalizácii hyperparametrov využívame metodiku k-násobnej krížovej validácie (anglický výraz: *k-fold cross validation*), ktorá predstavuje štandardný postup na objektívne hodnotenie výkonnosti modelu. Táto technika spočíva v rozdelení trénovacej množiny na k disjunktných podmnožín. V každej iterácii je model trénovaný na dátovej sade pozostávajúcej z $k - 1$ podmnožín a následne testovaný na zostávajúcej jednej

podmnožine. Tento postup sa iteratívne opakuje pre všetkých k podmnožín, pričom konečná výkonnosť modelu je vyjadrená ako priemer hodnôt vybraných výkonnostných metrík naprieč všetkými iteráciami k-násobnej krížovej validácie [18].

V každej iterácii optimalizačný algoritmus `Optuna` vyberie hodnoty jednotlivých hyperparametrov, ktoré sú následne aplikované pri definícii metódy LightGBM. Následne dôjde ku k-ásobnej krížovej validácii s nastavenými hyperparametrami pre metódu LightGBM v rámci iterácie algoritmu `Optuna`. Na vyhodnotenie kvality výberu hyperparametrov v danej iterácii sa využívaná metodika pozerá na výsledky metrík vzhľadom na podmožinu určenu na testovanie, ale aj trénovanie v rámci iterácie v krížovej validácii.

Metodika optimalizácie hyperparametrov, využívaná spoločnosťou ČSOB, sa zameriava na nasledujúce metriky kvality modelov v rámci krížovej validácie:

- AUC-PR$_{test}$ na testovacej množine v rámci jednej iterácie k-násobnej krížovej validácie,
- AUC-ROC$_{train}$ na trénovacej množine v rámci jednej iterácie k-násobnej krížovej validácie,
- AUC-ROC$_{test}$ na testovacej množine v rámci jednej iterácie k-násobnej krížovej validácie.

Celý proces sa aplikuje pre každú iteráciu krížovej validácie a následne sa pre každú iteráciu algoritmu `Optuna` vypočíta hodnota účelovej funkcie, ktorá slúži ako kritérium pre výber optimálnej konfigurácie hyperparametrov. V našom prípade sa hodnota k odvíja od použitej dĺžky histórie udalostí v mesiacoch.

4.4.3 Účelová funkcia optimalizácie

Po vykonaní všetkých iterácií v rámci krížovej validácie sa vypočíta hodnota účelovej funkcie, ktorá má nasledujúci tvar:

$$F_{opt} = \overline{\text{AUC-PR}}_{test} \cdot \max\left(0.999484^{|n_features|}, 0.1\right) \cdot \min\left(1, \frac{\overline{\text{AUC-ROC}}_{test}}{\overline{\text{AUC-ROC}}_{train}}\right) \tag{42}$$

kde:

- $\overline{\text{AUC-PR}}_{test}$ je priemerná hodnota AUC-PR z k-násobnej krížovej validácie na testovacích množinách,
- $\overline{\text{AUC-ROC}}_{test}$ je priemerná hodnota AUC-ROC z k-násobnej krížovej validácie na testovacích množinách,
- $\overline{\text{AUC-ROC}}_{train}$ je priemerná hodnota AUC-ROC z k-násobnej krížovej validácie na trénovacích množinách,
- 0.999484 odborný odhad konštanty zodpovednej za penalizáciiu v prípade veľkého počtu premnných v dátatch,
- $|n_features|$ je počet vybraných čŕt použitých v danej iterácii optimalizačného algoritmu.

Účelová funkcia je navrhnutá tak, aby primárne reflektovala priemernú hodnotu metriky AUC-PR, ktorá je obzvlášť vhodná pre identifikáciu pozitívnych vzoriek, najmä v prípadoch, keď sú tieto vzorky v dátach výrazne nedostatočne zastúpené. Okrem toho zohľadňuje pomer priemerov metriky AUC-ROC, ktorý je vypočítaný na trénovacích a testovacích množinách dát v rámci krížovej validácie. Tento pomer slúži ako indikátor možného pretrénovania modelu – v prípade, že by k pretrénovaniu dochádzalo, príslušný člen účelovej funkcie by výrazne znižoval jej celkovú hodnotu. Nakoniec je do účelovej funkcie zahrnutý aj člen, ktorý penalizuje nadmerný počet vybraných čŕt počas jednej iterácie optimalizácie, čo prispieva k zníženiu dimenzionality. Člen $0.999484^{|n_features|}$ je navrhnutý tak, aby za každých 100 čŕt navyše bola penalizácia približne 5%.

Optimalizačný algoritmus `Optuna` následne hľadá konfiguráciu hyperparametrov, ktorá maximalizuje hodnotu tejto účelovej funkcie.

5 Klesanie kvality predikčných modelov

Výsledky prezentované v tejto práci sa týkajú modelov odchodovosti klientov (fyzických osôb). Ako bolo v práci uvedené, budeme porovnávať modely vytvorené na základe 36-mesačnej, 24-mesačnej a 12-mesačnej histórie udalostí odchodov klientov. Pri každom modeli sa využíva jednotný prístup k tvorbe dátovej sady aj samotného predikčného modelu, pričom sa postupne zvyšuje hodnota parametra horizontu predikcie t. V nasledujúcej časti práce skúmame vzťah medzi týmto parametrom a kvalitou predikčných modelov. Prezentované výsledky sú vypočítané na validačnej vzorke modelov.

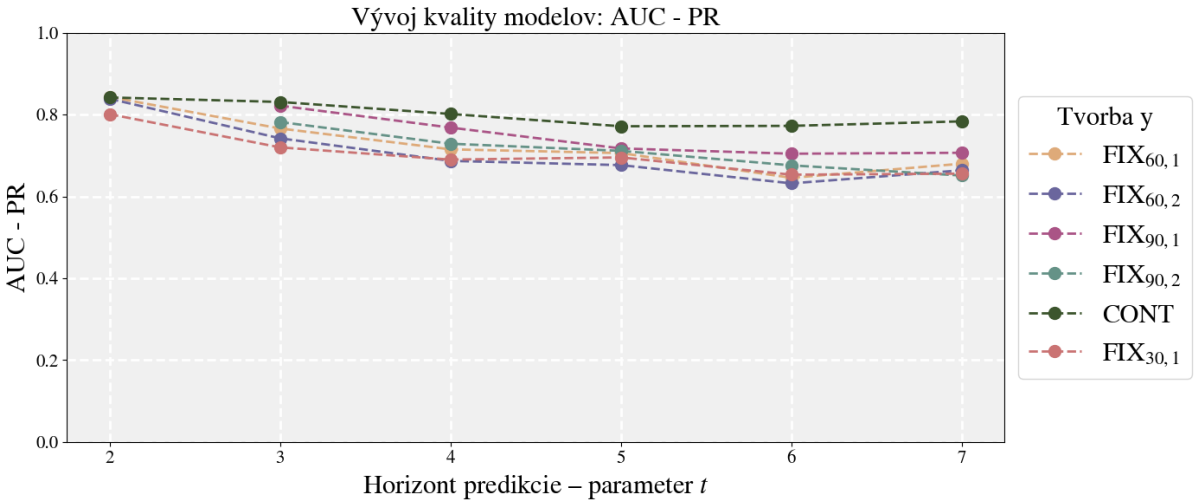
5.1 Merania na 36-mesačnej histórii udalostí

Pri tvorbe predikčných modelov zameraných na predikciu odchodov klientov (fyzických osôb) prvotne využívame 36-mesačnú históriu dostupných údajov o odchodoch klientov. Z dôvodu obmedzeného rozsahu dát bolo možné vytvoriť modely s maximálnym horizontom predikcie s hodnotou $t = 7$.

Na obrázku č. 11 je znázornený vývoj kvality modelov (vyjadrený pomocou metriky AUC-PR) v závislosti od rastúceho horizontu predikcie t. Tento graf poskytuje dôležitý pohľad na vplyv horizontu predikcie na kvalitu modelov pre rôzne spôsoby tvorby závislej premennej.

Z výsledkov možno pozorovať, že pri všetkých typoch modelov, ktoré sa odlišujú spôsobom tvorby závislej premennej pre črty klientov vstupujúcich do pozitívnej vzorky, dochádza s predlžovaním horizontu predikcie k poklesu ich kvality. Tento trend je v súlade s predpokladom, že vzdialenú budúcnosť predpovedať ťažšie ako bližšiu.

Zároveň je možné identifikovať, že modely pri ktorých bola závislá premenná vytvorená spôsobom CONT_t, dosahujú najlepšie výsledky pre každú hodnotu parametra t. Táto skutočnosť naznačuje, že voľba spôsobu tvorby závislej premennej má vplyv na výkonnosť predikčného modelu a mala by byť starostlivo zvážená pri tvorbe predikčného modelu.



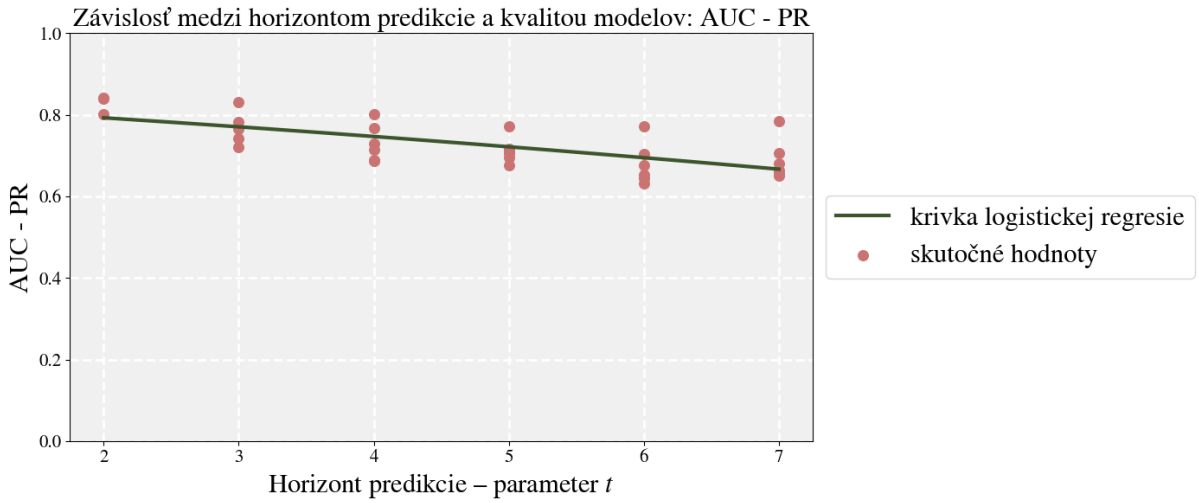
Obr. 11: Na grafe je znázornený vývoj hodnoty metriky AUC-PR v závislosti od rastúceho časového parametra t pre modely odchodovosti klientov s použitím 36-mesačnej histórie odchodov klientov. *Zdroj: Vlastné spracovanie.*

Na druhej strane, pri využití tohto spôsobu tvorby závislej premennej, dochádza s rastúcim parametrom t k výraznému nárastu neistoty ohľadom presného momentu odchodu klienta. Tento dôsledok naznačuje, že hoci vyššia hodnota časového parametra umožňuje dlhodobejšiu predikciu, zároveň znižuje spoľahlivosť modelu v určovaní presného okamihu, kedy klient prestane využívať služby spoločnosti. V tejto súvislosti vzniká zásadná otázka: čo je pre nás dôležitejšie – vyššia celková kvalita modelu alebo jeho časová presnosť pri predikcii odchodu klienta?

Pri analýze predchádzajúceho grafu možno pozorovať trend vo vývoji kvality modelov v závislosti vzdialenosti horizontu predikcie. Tento trend vykazuje klesajúcu tendenciu, čo naznačuje, že so vzdialenejším horizontom predikcie dochádza k zhoršeniu výkonnosti modelov. Aby sme tento vývoj nielen kvalitatívne opísali ale aj kvantifikovali, je potrebné presnejšie analyzovať vzťah medzi parametrom horizontu t a kvalitou modelov.

V podkapitole č. 2.4 podrobne opisujeme metodológiu, ktorú pri tejto analýze používame. Na základe všetkých získaných výsledkov modelov konštruujeme model logistickej regresie, ktorý umožňuje lepšie pochopiť a formalizovať vzťah medzi horizontom predikcie a kvalitou modelov. Po odhadnutí parametrov logistickej regresie získavame krivku, ktorá reprezentuje túto závislosť a poskytuje formálnu interpretáciu pozorovaného poklesu kvality.

Na obrázku č. 13 sú znázornené všetky výsledky kvality predikčných modelov a logistická krivka, ktorá opisuje vývoj kvality modelov s horizontom predikcie. Tento graf poskytuje vizuálny náhľad na to, ako sa presnosť modelov mení s rastúcim horizontom predikcie a umožňuje lepšie pochopenie limitácií modelov pri predikcii dlhodobých odchodov klientov. V rámci odhadu logistickej krivky sme nepoužili výsledky modelov s tvorbou závislej premennej $\text{FIX}_{60,1}$ s hodnotou parametra $t = 2$ a $\text{FIX}_{90,1}$ s hodnotou parametra $t = 3$. Je to z toho dôvodu, aby sme v dátach nemali redundantné pozorovania, vzhľadom na to, že tieto modely sú totožné s modelmi s tvorbou pozitívnej premennej CONT pri tých istých hodnotách parametra t.



Obr. 12: Závislosť kvality modelov ku zväčšujúcemu sa horizontu predikcie (parameter t). *Zdroj: Vlastné spracovanie.*

Odhadnutý záporný koeficient $\hat{\tau}$ potvrdzuje opticky pozorovanú klesajúcu tendenciu vývoja kvality modelov. Tento výsledok naznačuje, že s rastúcou hodnotou časového parametra t dochádza k systematickému znižovaniu presnosti modelov, čo je v súlade s našimi predpokladmi a predchádzajúcou vizuálnou analýzou.

V tabuľke č. 1 uvádzame odhadnuté hodnoty koeficientov získané z logistickej regresie. Tieto hodnoty poskytujú kvantitatívny pohľad na vzťah medzi horizontom predikcie a kvalitou modelov, čím umožňujú presnejšiu interpretáciu trendov.

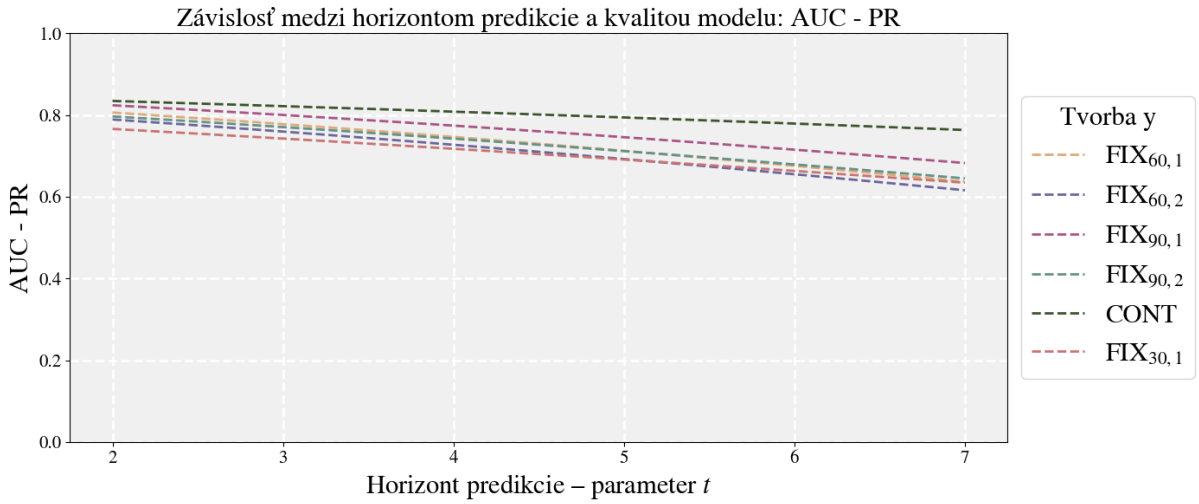
Koeficient	Hodnota	Štandardná odchylka	P-value
$\hat{\beta}_0$	1.600	1.307	0.221
$\hat{\tau}$	-0.129	0.253	0.608

Tabuľka 1: Tabuľka obsahuje hodnoty odhadnutých koeficientov logistickej regresie, pri použití všetkých výsledkov kvality modelov. Celkový počet dátových bodov je 32. *Zdroj: Vlastné spracovanie.*

Odhadnutý koeficient $\hat{\tau}$ nadobúda hodnotu -0.129, čo indikuje, že pri zväčšení horizontu predikcie o jeden mesiac sa šance dosiahnutia perfektného modelu znižuje približne 0.88-násobne $= \mathbf{e}^{\tau}$.

Časový parameter t určuje vzdialenosť horizontu predikcie, avšak presný časový interval, v ktorom klient odíde, závisí od definície spôsobu tvorby závislej premennej. Modely natrénované na dátovej sade, v ktorej bol použitý spôsob s kontinuálne zväčšujúcim sa rozpätím predikcie CONT_t, majú najširšie predikčné okno. V jednoduchosti tieto modely odpovedajú na otázku, či klient odíde v priebehu nasledujúcich t mesiacov. Na druhej strane, modely trénované na dátovej sade, v rámci ktorej sa použil spôsob tvorby závislej premennej s konštantne veľkým časovým rozpätím $\text{FIX}_{30,1,t}$, predikujú, či klient odíde presne v t-tom mesiaci. Tento zásadný rozdiel v definovaní predikčného intervalu môže mať významný vplyv na vzťah medzi kvalitou modelov a horizontom predikcie. Pri širšom časovom intervale predikcie môže model lepšie zachytiť celkový trend odchodu klientov, avšak za cenu vyššej neistoty ohľadom presného okamihu odchodu. Naopak, pri užšom intervale predikcie sa zvyšuje presnosť v čase, avšak môže dochádzať k zníženiu celkovej výkonnosti modelu z dôvodu obmedzeného množstva pozitívnych prípadov v trénovacej sade.

Ako možno pozorovať na obrázku č. 13, logistická krivka vykazuje klesajúcu tendenciu pre všetky modely bez ohľadu na použitý spôsob tvorby závislej premennej. Tento trend naznačuje, že s rastúcim časovým parametrom dochádza k systematickému znižovaniu výkonnosti modelov, čo poukazuje na prirodzenú degradáciu kvality informácie z dát.



Obr. 13: Závislosť kvality jednotlivých modelov ku predlžujúcemu sa horizontu predikcie (parameter t). *Zdroj: Vlastné spracovanie.*

Je však potrebné zdôrazniť, že uvedená závislosť bola vypočítaná na základe obmedzeného počtu dátových bodov. Pre každý spôsob tvorby závislej premennej disponujeme len 5 alebo 6 dátovými bodmi, čo ovplyvňuje spoľahlivosť odhadnutých parametrov modelov logistickej regresie.

Okrem toho výsledná krivka závislosti vykazuje značnú citlivosť na hodnotu prvého dátového bodu, čo môže viesť k skresleniu celkovej interpretácie trendu. Z tohto dôvodu je potrebné pristupovať k výsledkom s primeranou mierou opatrnosti.

V tabuľke č. 2 uvádzame odhadnuté hodnoty koeficientov logistickej regresie pre jednotlivé spôsoby tvorby závislej premennej, ktoré poskytujú detailnejší pohľad na vplyv časového parametra na kvalitu modelov aj vzhľadom na veľkosť predikčného intervalu. Taktiež sa v nej nachádzajú odhady štandardných odchyliek $\hat{\sigma}_\beta$ pre jednotlivé koeficienty logistickej regresie, ich p-value a počet dátových bodov, ktoré boli na tento odhad použitý.

Tvorba y	$\hat{\beta}_0$	$\hat{\sigma}_{\beta_0}$	$\hat{\tau}$	$\hat{\sigma}_{\tau}$	P-value	OR: $e^{\hat{\tau}}$	# dát
$FIX_{60,1}$	1.771	2.744	-0.173	0.547	0.752	0.841	6
$FIX_{60,2}$	1.659	2.675	-0.169	0.535	0.752	0.844	6
$FIX_{90,1}$	1.854	3.911	-0.155	0.733	0.832	0.856	5
$FIX_{90,2}$	1.672	3.745	-0.154	0.704	0.827	0.858	5
CONT	1.796	2.981	-0.089	0.601	0.882	0.915	6
$FIX_{30,1}$	1.440	2.616	-0.127	0.529	0.810	0.881	6

Tabuľka 2: Koeficienty logistickej regresie, jej základné štatistiky a pomer šancí OR pre rôzne spôsoby tvorby závislej premennej pri použití 36-mesačnej histórie udalostí. *Zdroj: Vlastné spracovanie*

Na základe odhadnutých hodnôt koeficientov $\hat{\tau}$ môžeme pozorovať, že najmenší pokles kvality modelov pri predlžujúcom sa horizonte predikcie nastáva pri $CONT_t$ spôsobe tvorby závislej premennej. Naopak, najrýchlejší pokles kvality modelov je zaznamenaný pri tvorbe závislej premennej spôsobom $FIX_{60,1,t}$. Tento výsledok môže na prvý pohľad pôsobiť prekvapivo, keďže najstriktnejšie modely vzhľadom na interval predikcie sú tie, ktoré vznikajú pomocou $FIX_{30,1,t}$ spôsobu tvorby závislej premennej. Je však potrebné zohľadniť aj rozdiel medzi maximálnou a minimálnou hodnotou metriky kvality modelov AUC-PR, ktorú jednotlivé modely dosahujú v závislosti od zvoleného spôsobu tvorby závislej premennej.

Na obrázku č. 11 možno pozorovať, že model vytvorený použitím $FIX_{30,1,t}$ spôsobom tvorby závislej premennej dosahuje viditeľne nižšiu kvalitu už pri horizonte predikcie $t = 2$. Hoci s predlžujúcim sa horizontom dochádza k ďalšiemu poklesu kvality modelu pri tomto spôsobe, tento pokles nie je taký výrazný ako pri ostatných metódach tvorby závislej premennej. Tento jav možno pripísať skutočnosti, že pri $FIX_{30,1,t}$ spôsobe tvorby závislej premennej je celková úroveň výkonnosti modelu už na začiatku nižšia, a preto jej relatívny pokles v priebehu predikčného horizontu nie je taký strmý.

Ďalším zaujímavým pozorovaním je skutočnosť, že zväčšenie pozitívnej vzorky pri spôsoboch $FIX_{60,2,t}$ a $FIX_{90,2,t}$ nemalo výrazný vplyv na klesajúci trend kvality modelov

v porovnaní s ich základnými variantmi, ktoré zahŕňali len jeden vektor čŕt klienta do pozitívnej vzorky. Tento fakt naznačuje, že zväčšenie pozitívnej vzorky pomocou údajov od toho istého klienta nemusí viesť k zlepšeniu modelu a jeho schopnosti presnejšie predikovať celkový odchod klientov.

5.2 Merania na 24-mesačnej histórii udalostí

V predchádzajúcej podkapitole sme identifikovali klesajúcu tendenciu kvality predikčných modelov pri zväčšovaní hodnoty horizontu predikcie t u všetkých spôsoboch tvorby závislej premennej. Táto degradácia kvality modelov môže byť pripísaná znižovaniu informačnej hodnoty historických údajov pri predlžovaní predikčného horizontu. Napriek snahe o dôslednú analýzu tohto javu sme boli obmedzení dostupnosťou dát, čo nám znemožnilo testovanie modelov s vyššími hodnotami časového parametra t.

S cieľom prekonať tieto obmedzenia sme pristúpili k modifikácii rozsahu historických údajov, pričom sme pre históriu udalostí použili iba posledných 24 mesiacov. Táto úprava nám umožnila zvýšiť maximálnu vzdialenosť horizontu predikcie t až na hodnotu 19, čím získavame širší priestor na skúmanie predikčných modelov a horizontu predikcie.

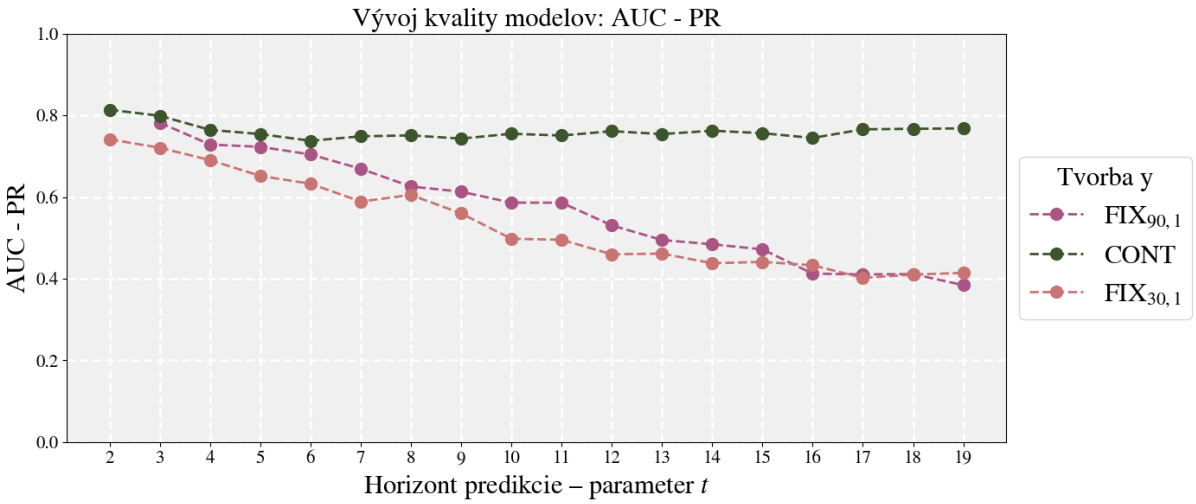
Táto metodická zmena však prináša aj určité nevýhody. Aby sme zachovali približne rovnaký pomer medzi pozitívnou a negatívnou vzorkou, bolo nevyhnutné redukovať celkový rozsah dátovej sady určenej na trénovanie a validáciu modelov o jednu tretinu. Napriek tejto redukcii zostáva výsledná dátová množina dostatočne rozsiahla, keďže obsahuje až 200 000 pozorovaní.

V tejto časti sa zameriavame na komparatívnu analýzu vybraných spôsobov tvorby závislej premennej, pričom prezentujeme výsledky pre spôsoby: CONT_t, $\text{FIX}_{30,1,t}$ a $\text{FIX}_{90,1,t}$. Hlavným cieľom tejto časti je detailne preskúmať, akým spôsobom predlžovanie predikcie ovplyvňuje kvalitu predikčných modelov.

Aj pri použití menšej dátovej sady dochádza k poklesu kvality modelov, čo je zreteľne viditeľné na obrázku č. 16. Pozorujeme však výrazne odlišnú dynamiku poklesu kvality modelov v závislosti od konkrétneho spôsobu tvorby závislej premennej. Kým pri modeloch s konštantne veľkým časovým rozpätím, ako sú $\text{FIX}_{30,1,t}$ a $\text{FIX}_{90,1}$, možno sledovať takmer lineárny trend zhoršovania výkonu modelov, pri modeloch využívajúcich kontinuálne zväčšujúce sa rozpätie predikcie (CONT_t) dochádza

k poklesu kvality len do hodnoty horizontu predikcie s $t = 8$. Následne pozorujeme stabilizovaný vývoj, pričom hodnoty oscilujú mierne okolo úrovne 0.75.

Oproti spôsobu CONT_t, modely vytvorené pomocou konštantnej veľkosti rozpätia predikcie naďalej vykazujú klesajúci trend aj po prekročení tejto hodnoty parametra horizontu predikcie. Pri vzájomnom porovnaní týchto dvoch metód je zrejmé, že modely využívajúce spôsob $\text{FIX}_{90,1,t}$ dosahujú mierne vyššiu úroveň kvality predikcie až do hodnoty horizontu predikcie s $t = 15$. Tento fakt naznačuje, že uvedený spôsob môže byť vhodnejší pri modelovaní udalostí, kde je kľúčové dosiahnuť vyššiu presnosť predikcie nastatia budúcich udalostí.



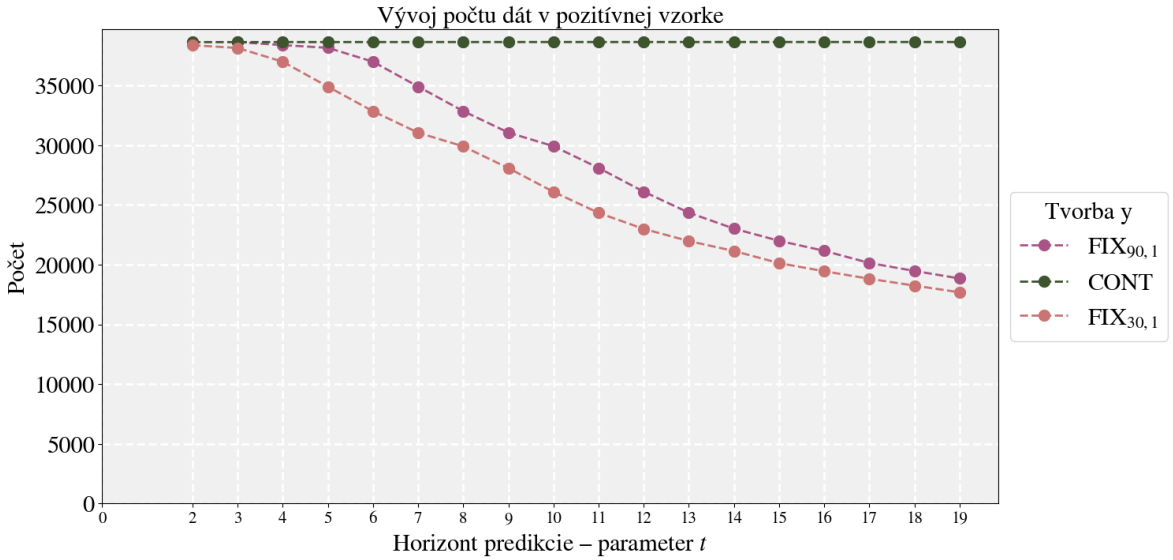
Obr. 14: Na grafe je znázornený vývoj hodnoty metriky AUC-PR v závislosti od rastúceho časového parametra t pre modely odchodovosti klientov s použitím 24-mesačnej histórie odchodov klientov. *Zdroj: Vlastné spracovanie.*

V tejto súvislosti sa objavuje otázka, prečo dochádza k tak výraznému rozdielu v dynamike kvality modelov, ktorý nebol až tak nápadný pri použití 36-mesačnej histórie odchodov. Tento rozdiel možno pripísať zmenám v spôsobe tvorby závislej premennej, najmä pri použití spôsobu CONT_t, kde sa začína prejavovať flexibilita pri výbere čŕt klientov. V prípade ostatných spôsobov dochádza k výraznému zmenšeniu pozitívnej vzorky, čo vedie k obmedzeniu dostupných informácií o klientoch, ktorí odišli. Tieto informácie sú nevyhnutné na to, aby sa model mohol naučiť rozpoznať rôzne kategórie a predpovedať odchod klientov. Naopak, spôsob CONT_t vďaka svojej definícii a možnosti výberu čŕt umožňuje zahrnúť aj novšie dáta v prípade, že staršie údaje nie sú k dispozícii.

Tento spôsob poskytuje širšie možnosti pri tvorbe dátovej sady a môže viesť k lepšiemu rozpoznaniu a modelovaniu odchodov v dlhodobých predikciách. Je preto evidentné, že flexibilita pri výbere čŕt zohráva kľúčovú úlohu v zlepšení kvality modelov, najmä v prípade, keď sa predpovedá na dlhšie obdobie. Tento jav sa stáva obzvlášť zrejmým pri porovnaní s inými spôsobmi, ktoré nezohľadňujú nové dostupné dáta a obmedzujú sa na staršie informácie.

Na obrázku č. 15 je zreteľne viditeľný rozdiel v dynamike vývoja počtu dát v pozitívnej vzorke. Z grafu je možné pozorovať postupné ubúdanie dát, ktoré sú dostupné na tvorbu a validáciu modelov pre spôsoby $\text{FIX}_{30,1,t}$ a $\text{FIX}_{90,1,t}$. Modely vytvorené týmito spôsobmi prichádzajú približne o 55% dát v pozitívnej vzorke pri predikčnom horizonte s parametrom $t = 19$, v porovnaní s hodnotou $t = 2$.

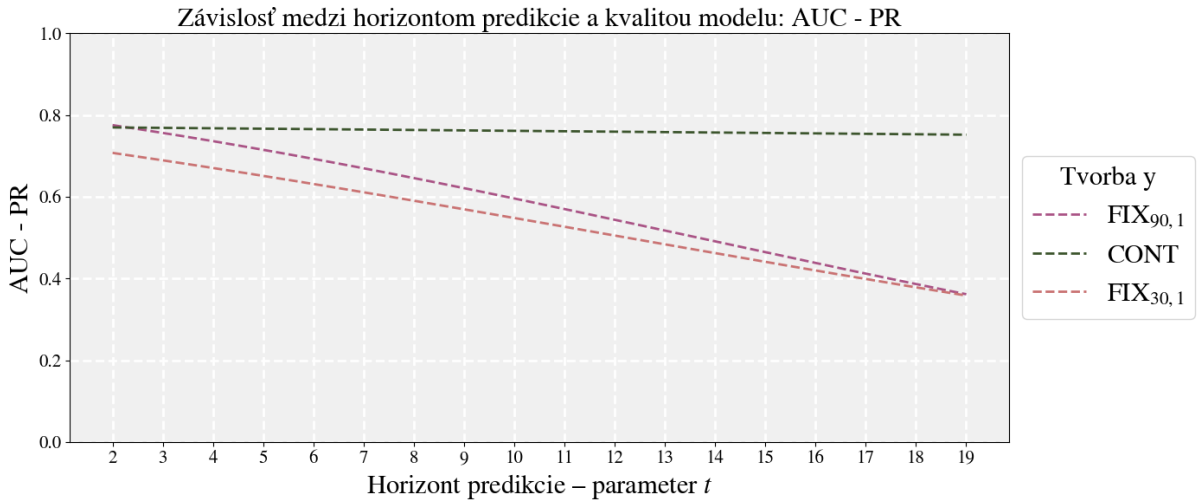
Ďalšou výhodou, z ktorej profitujú modely využívajúce spôsob CONT_t, je fakt, že aj keď klient mohol byť spoločnosťou evidovaný po celé obdobie 19 mesiacov, CONT_{19} spôsob môže stále vybrať črty, ktoré boli vypočítané tesne pred jeho odchodom. Týmto spôsobom môžu byť údaje v pozitívnej vzorke tvorené primárne črtami, ktoré sú bližšie k dátumu odchodu ako k maximálnej hodnote predikcie. Inými slovami to znamená, že len malé množstvo vektorov čŕt klientov v pozitívnej vzorke môže pochádzať z veľmi vzdialenej histórie klienta, čo ďalej prispieva k lepšiemu modelovaniu odchodov.



Obr. 15: Na grafe je znázornený vývoj počtu dát v pozitívnej vzorke pri použití 24-mesačnej histórie udalostí, vzľadom na použitý spôsob tvorby závislej premennej. *Zdroj: Vlastné spracovanie.*

S horizontom predikcie sa pri použití 24-mesačnej histórie udalostí dostávame do bodu, kde sa začínajú výrazne prejavovať rozdiely v definíciach časových rozpätí modelov. Z tohto dôvodu považujeme za nevhodné skúmať závislosť medzi horizontom predikcie a výsledkami všetkých modelov súčasne. Na druhej strane je však zaujímavá závislosť medzi horizontom predikcie a kvalitou modelov, ktorá je podmienená použitým spôsobom tvorby závislej premennej. Krivky týchto závislostí sú zobrazené na obrázku č. 16.

Z obrázku je možné pozorovať, že pri použití 24-mesačnej histórie udalostí a aplikácii spôsobu CONT_t tvorby závislej premennej sa závislosť medzi horizontom predikcie a kvalitou modelu javí ako konštantná. Tento jav potvrdzuje aj odhadnutá hodnota $\hat{\tau}$, ktorú uvádzame v tabuľke č. 3 pre všetky spôsoby tvorby závislej premennej.



Obr. 16: Závislosť kvality modelov a horizontom predikcie (parameter t). *Zdroj: Vlastné spracovanie.*

Podobne ako v prípade použitia 36-mesačnej histórie, aj kvalita modelov vytvorených pomocou spôsobu $\text{FIX}_{90,1,t}$ klesá rýchlejším tempom v porovnaní s modelmi vytvorenými pomocou $\text{FIX}_{30,1,t}$ spôsobom. Tento jav je však opäť spôsobený tým, že spôsob $\text{FIX}_{90,1,t}$ dosahuje lepšie výsledky pri kratších predikciách. Až následne, keď dochádza k výraznému zníženiu počtu dát v pozitívnej vzorke, kvalita modelov vytvorených týmto spôsobom klesá rýchlejšie. Tým pádom je tento spôsob z pohľadu poklesu kvality náchylnejší na rastúcim horizontom predikcie, avšak samotne dosahuje kvalitnejšie výsledky pri predikciách na kratšie obdobia.

Tvorba y	$\hat{\beta}_0$	$\hat{\sigma}_{\beta_0}$	$\hat{\tau}$	$\hat{\sigma}_{\tau}$	**P-value**	**OR:** $e^{\hat{\tau}}$	**# dát**
$\text{FIX}_{90,1}$	1.451	1.302	-0.106	0.106	0.316	0.899	17
CONT	1.218	1.255	-0.006	0.107	0.957	0.994	18
$\text{FIX}_{30,1}$	1.055	1.124	-0.086	0.095	0.365	0.917	18

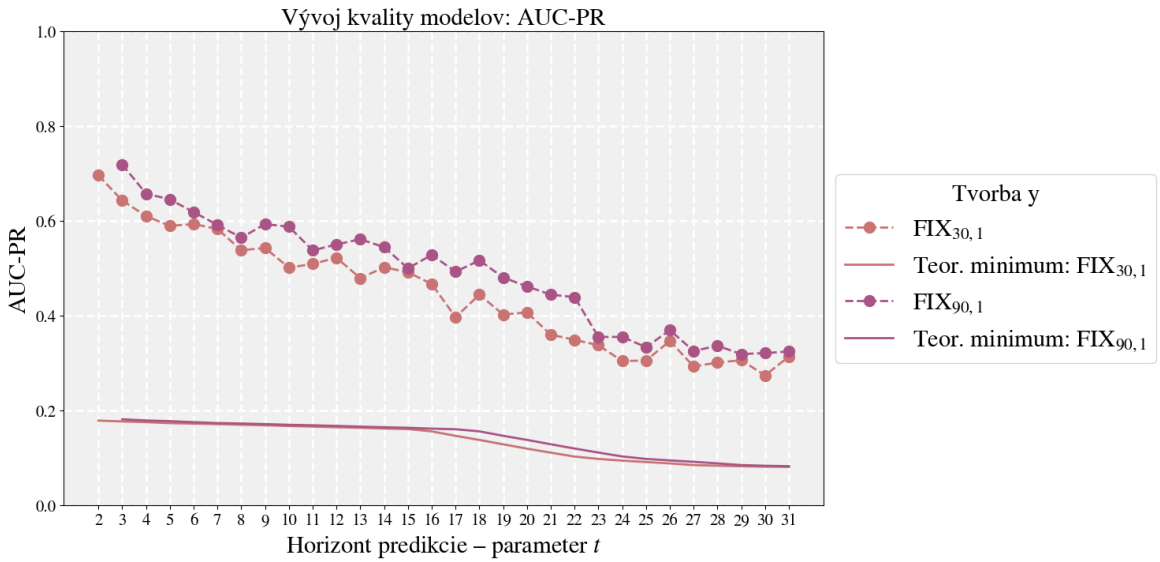
Tabuľka 3: Koeficienty logistickej regresie, jej základné štatistiky a pomer šancí OR pre rôzne spôsoby tvorby závislej premennej pri použití 24-mesačnej histórie udalostí. *Zdroj: Vlastné spracovanie*

5.3 Merania na 12-mesačnej histórii udalostí

Ako sme mohli doposiaľ pozorovať, kvalita predikčných modelov vykazuje klesajúcu tendenciu so vzdialenejším horizontom predikcie. Napriek tomu ani pri horizonte predikcie s parametrom $t = 19$ nedochádza k úplnej degradácii modelu na úroveň náhodnej predikcie. Na základe tohto pozorovania sme implementovali ďalšiu úpravu, pri ktorej sme skrátili historické obdobie použitých údajov na 12 mesiacov. Tento krok nám umožnil zvýšiť hodnotu parametra t až na 31, pričom analýza bola v tejto fáze obmedzená na modely využívajúce spôsoby: $\text{FIX}_{30,1,t}$ a $\text{FIX}_{90,1,t}$. Spôsob tvorby závislej premennej CONT_t dosahuje výraznú neistotu nastatia udalosti pri takýchto dlhých predikciách a výsledky modelov sa odrážajú skôr od nových pozorovaní o klientovi ako od starších informácií.

Ako bolo uvedené v podkapitole 2.3, metrika AUC-PR má teoretické minimum, ktoré reprezentuje hodnotu dosiahnutú náhodným klasifikačným modelom. Obrázok 17 znázorňuje pokles hodnoty AUC-PR v závislosti od horizontu predikcie, pričom súčasne zobrazuje aj teoretické minimum pre každý spôsob tvorby závislej premennej a príslušnú hodnotu parametra t. Z grafu je zrejmé, že hoci sa kvalita modelov s predlžovaním horizontu predikcie znižuje, stále nedosahuje svoje teoretické minimum. Na druhej strane, rozdiel medzi skutočnou výkonnosťou modelov a ich teoretickým minimom sa s rastúcim horizontom predikcie postupne zmenšuje.

Hodnota teoretického minima pre AUC-PR je definovaná ako podiel počtu pozitívnych vzoriek k celkovému počtu údajov v dátovej sade. Z grafu môžeme pozorovať, že pri hodnote $t = 31$ sa počet pozitívnych vzoriek znížil približne o viac ako polovicu, v porovnaní s hodnotami $t = 2$ a $t = 3$. Tento trend sa odzrkadľuje aj na samotnej kvalite modelov, ktorá sa v tomto prípade taktiež znížila približne na polovicu. Táto skutočnosť jasne demonštruje súvislosť medzi kvalitou dostupných dát a presnosťou predikčných modelov.



Obr. 17: Na grafe je znázornený vývoj hodnoty metriky AUC-PR v závislosti od rastúceho parametra horizontu t pre modely odchodovosti klientov s použitím 12-mesačnej histórie odchodov klientov s porovnaním výsledkov pre náhodný model. *Zdroj: Vlastné spracovanie.*

Hoci sme pri vyšších hodnotách parametra t nedosiahli teoretické minimum, z grafu možno pozorovať, že kvalita modelov pre $t = 31$ je len mierne vyššia ako kvalita náhodného klasifikačného modelu pri $t = 2$ alebo $t = 3$. Táto skutočnosť vedie k zásadnej otázke: *Je takýto model dostatočne kvalitný a má vôbec význam?* Odpoveď závisí od definovanej obchodnej stratégie, pre ktorú je predikčný model tvorený.

Ak je cieľom predikcie identifikovať presný okamih udalosti v budúcnosti, nie je možné vyhnúť sa využívaniu historických údajov klienta, ktoré neobsahujú takú informačnú hodnotu ako novšie pozorovania. Napriek tomu sa ukazuje, že pri správnej metodike tvorby predikčných modelov dokážu metódy strojového učenia extrahovať relevantné informácie aj z dávnej histórie. Tieto pozorovanie vychádzajú z faktu, že ani pri predikcii správania klienta na viac ako 2 roky, predikčné modely boli lepšie ako náhodný klasifikátor.

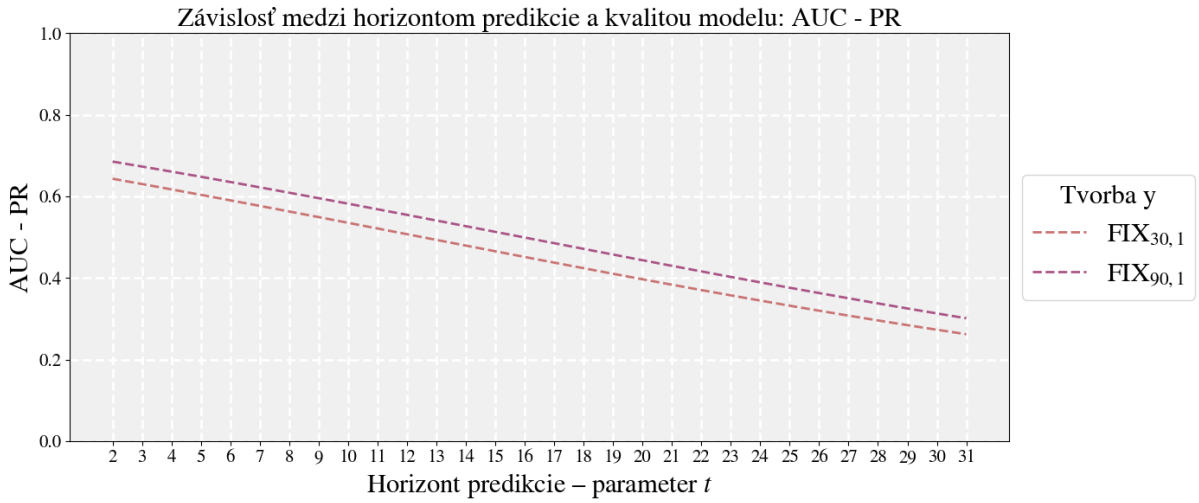
Na druhej strane, ak nie je nevyhnutné vytvárať predikčné modely pre vzdialenú budúcnosť, napríklad v kontexte marketingových stratégií, je vhodnejšie uprednostniť pozorovania bližšie k okamihu vzniku udalosti a teda zmenšiť časový horizont predikcie. Výsledky analýzy ukazujú, že modely určené na

predikciu vzdialeného odchodu klienta dosahujú nízku kvalitu.

Podobne ako v predchádzajúcich častiach práce, aj v tejto sekcii uvádzame na obrázku č. 18 vzťah medzi kvalitou modelov a horizontom predikcie vzhľadom na použitý spôsob tvorby závislej premennej.

Ak by sme sa snažili dosiahnuť pravdepodobnosť blížiacu sa hodnote 1, bolo by potrebné nastaviť parameter t na výrazne záporné hodnoty. To by však znamenalo, že ani pri predikovaní minulosti by model nedosahoval 100% kvalitu. Tento výsledok nám reprezentuje to, že kvalita modelov pri použití 12-mesačnej histórie udalostí je výrazne nižšia, čo sa odzrkadľuje na konštantných koeficientoch logistickej regresie β_0. Môžeme taktiež pozorovať výrazný paralelný pokles kvality modelov, ktorý je zrejme zapríčinený použitou veľkosťou histórie udalostí.

Z tohto dôvodu dávame do pozornosti ďalší dôležitý časový aspekt pri tvorbe predikčných modelov, ktorý nemusí byť na prvý pohľad evidentný – a to otázku, akú dlhú historickú periódu udalostí je vhodné zahrnúť do analýzy. Výber historického obdobia môže výrazne ovplyvniť kvalitu predikcií a preto je nevyhnutné venovať tejto problematike primeranú pozornosť pri tvorbe modelu.



Obr. 18: Závislosť kvality modelov a horizontom predikcie (parameter t). *Zdroj: Vlastné spracovanie.*

Ako možno pozorovať v tabuľke č. 4, pri použití 12-mesačnej histórie udalostí odchodov klientov vykazujú oba aplikované spôsoby tvorby závislej premennej takmer identickú závislosť od horizontu predikcie.

V porovnaní s výsledkami prezentovanými v predchádzajúcich častiach práce však možno pozorovať všeobecný negatívny posun kvality modelov. Tento pokles je zreteľný pri hodnotách parametrov $\hat{\beta}_0$, ktoré sú v tomto prípade výrazne nižšie. Uvedené skutočnosti naznačujú, že skrátenie historického obdobia použitých údajov má vplyv na kvalitu predikčných modelov, čo je dôležité zohľadniť pri ich návrhu a interpretácii výsledkov.

Tvorba y	$\hat{\beta}_0$	$\hat{\sigma}_{\beta_0}$	$\hat{\tau}$	$\hat{\sigma}_{\tau}$	**P-value**	**OR:** $e^{\hat{\tau}}$	**# dát**
$FIX_{30,1}$	0.702	0.811	-0.056	0.045	0.210	0.946	30
$FIX_{90,1}$	0.890	0.874	-0.056	0.047	0.230	0.946	29

Tabuľka 4: Koeficienty logistickej regresie, jej základné štatistiky a pomer šancí OR pre rôzne spôsoby tvorby závislej premennej pri použití 12-mesačnej histórie udalostí. *Zdroj: Vlastné spracovanie*

5.4 Syntéza výsledkov klesania kvality modelov

V doterajších častiach práce sme skúmali vzťah medzi horizontom predikcie a jej kvalitou postupne vzhľadom na použitú históriu udalostí, ako aj vzhľadom na samotné rozpätie predikcie. Model logistickej regresie, ktorý sme predstavili v podkapitole 2.4, má len jeden vstup a to je hodnota parametra t, ktorý je používaný v rámci tvorby závislej premennej.

V rámci nášho výskumu sme natrénovali 146 predikčných modelov odchodovosti klientov. Pri použití 36-mesačnej histórie udalostí sme využili všetky prezentované prístupy tvorby závislej premennej. Na analýzu tejto časti bolo natrénovaných celkovo 34 modelov. Pri použití 24-mesačnej histórie udalostí sme natrénovali celkovo 53 predikčných modelov. V rámci tejto časti sme zväčšili horizont predikcie t na hodnotu 19. Bolo zreteľné, že rozdiely v spôsobe tvorby závislej premennej sa v tejto časti výrazne prejavili. Pri použití 12-mesačnej histórie udalostí sme sa obmedzili len na modely so spôsobom tvorby závislej premennej s konštantne veľkým časovým rozpätím $FIX_{30,1,t}$ a $FIX_{90,1,t}$. Ako možno v tejto časti práce pozorovať, **ani pri horizonte predikcie $t = 31$ sa kvalita modelov nedostala na úroveň náhodného klasifikátora.** V tejto časti analýzy vplyvu času na kvalitu predikcie bolo natrénovaných 59 predikčných modelov. Tvorba dátovej sady a optimalizácia hyperparametrov jedného predikčného modelu trvala približne jeden a pol

hodiny.

Ako sme doposiaľ mohli pozorovať, kvalita a početnosť informácií, z ktorých sa model trénuje, nezávisí len od horizontu predikcie, ale aj od jej veľkosti časového rozpätia predikcie a použitej veľkosti histórie udalostí.

Jedným z kľúčových faktorov ovplyvňujúcich kvalitu predikčných modelov je rozsah historických údajov použitých pri ich trénovaní. V predchádzajúcich analýzach sme mali možnosť pozorovať, že znižovanie početnosti historických dát vedie k systematickému poklesu kvality modelov, pričom tento pokles možno sledovať ako paralelný posun predikčných schopností smerom nadol.

Pre ilustráciu tohto efektu môžeme porovnať výsledky modelov natrénovaných na historických údajoch s rôznou dĺžkou záznamov. Pri použití 36-mesačnej histórie udalostí dosahovali modely pri krátkodobých predikciách (výsledky pre $t = 2$ a $t = 3$) priemernú kvalitu na úrovni 0.80. Ak však rozsah historických dát zredukujeme na 24 mesiacov, kvalita predikcie klesne na hodnotu 0.77. Pri ďalšom skrátení histórie na 12 mesiacov dochádza k výraznejšiemu poklesu kvality až na úroveň 0.68.

Táto skutočnosť má významné dôsledky nielen z hľadiska presnosti predikčných modelov, ale aj z pohľadu ekonomických aspektov správy dát. Ukladanie a spracovanie rozsiahlych historických údajov predstavuje značnú finančnú záťaž pre organizácie a preto je dôležité skúmať, do akej miery je rozsiahla historická databáza skutočne nevyhnutná pre dosiahnutie dostatočne presných predikcií.

Na základe uvedených výsledkov môžeme konštatovať, že rozdiel v kvalite predikcií medzi modelmi trénovanými na 36-mesačných a 24-mesačných historických dátach nie je výrazný, čo naznačuje, že redukcia rozsahu historických údajov na 24 mesiacov môže predstavovať efektívnu stratégiu na zníženie nákladov bez podstatného zníženia kvality modelov. Na druhej strane, pokles kvality modelov pri použití 12-mesačnej histórie udalostí je už natoľko výrazný, že môže negatívne ovplyvniť ich praktickú využiteľnosť. Tieto zistenia poskytujú dôležitý podklad pre rozhodovacie procesy týkajúce sa správy historických údajov a optimalizácie predikčných modelov v rôznych aplikáciách.

Horizont predikcie a veľkosť časového rozpätia predikcie sú významne ovplyvnené degradáciou informácií v dátach. S predlžujúcim sa horizontom predikcie dochádza k narastajúcej absencii relevantných údajov o jednotlivých klientoch.

Tento fenomén je dôsledkom prirodzených zmien v správaní zákazníkov, ktorí sa v čase vyvíjajú a fluktuujú medzi rôznymi spoločnosťami, a to aj napriek snahám o ich dlhodobú retenciu.

Pri dlhodobých predikciách sa tak vzorka klientov, ktorí zostávajú verní spoločnosti a na základe ktorých je možné zostaviť robustný predikčný model, neustále zmenšuje. Táto postupná strata informácií môže viesť k oslabeniu presnosti predikcie a zníženiu spoľahlivosti modelov.

Jedným z možných prístupov k spomaleniu tejto informačnej degradácie je rozšírenie časového rozpätia využívaného pri tvorbe závislej premennej, podobne ako spôsob s kontinuálne rozširujúcim sa rozpätím predikcie, ktorý bol v práci prezentovaný. Avšak tento postup prináša aj negatívne dôsledky – najmä zvyšovanie neistoty.

V konečnom dôsledku je teda potrebné starostlivo vyvážiť horizont predikcie a rozsah historických údajov, aby bolo možné dosiahnuť optimálny kompromis medzi kvalitou predikčných modelov a dostupnosťou relevantných informácií. Tento aspekt je obzvlášť dôležitý pri návrhu modelov využívaných v dynamických prostrediach, kde dochádza k častým zmenám v správaní zákazníkov.

Ako sme už uviedli, horizont predikcie nie je jediným časovým faktorom, ktorý ovplyvňuje predikčné modelovanie. Aby sme kvantifikovali vplyv ďalších časových faktorov v predikčnom modelovaní a zároveň očistili doterajšie výsledky od možných skreslení spôsobených týmito vplyvmi, pristúpime k tvorbe rozšíreného modelu logistickej regresie.

Informáciu o spôsobe tvorby závislej premennej zakódujeme pomocou indikátorových premenných. Pri tomto prístupe bude spôsob $\text{FIX}_{30,1,t}$ zahrnutý v koeficiente β_0 a nebudeme pre neho vytvárať premennú navyše.

Takýmto spôsobom budeme schopní analyzovať dĺžku časového rozpätia predikcie na kvalitu modelu. Zároveň budeme môcť skúmať, či jednotlivé prístupy k tvorbe závislej premennej významne ovplyvňujú presnosť predikcie. Model logistickej regresie bude mať nasledovný tvar:

$$P(\text{AUC-PR} = 1 \mid \mathbf{x}) = \frac{1}{1 + \exp^{-\boldsymbol{\beta}^\intercal \mathbf{x}}} \quad \text{, pričom :} \tag{43}$$

$$\boldsymbol{\beta}^\intercal \mathbf{x} = \beta_0 + \tau t + \beta_{hist} hist + \beta_{CONT} x_{CONT} + \beta_{FIX_{90}} x_{FIX_{90}} + \beta_{FIX_{60}} x_{FIX_{60}}, \tag{44}$$

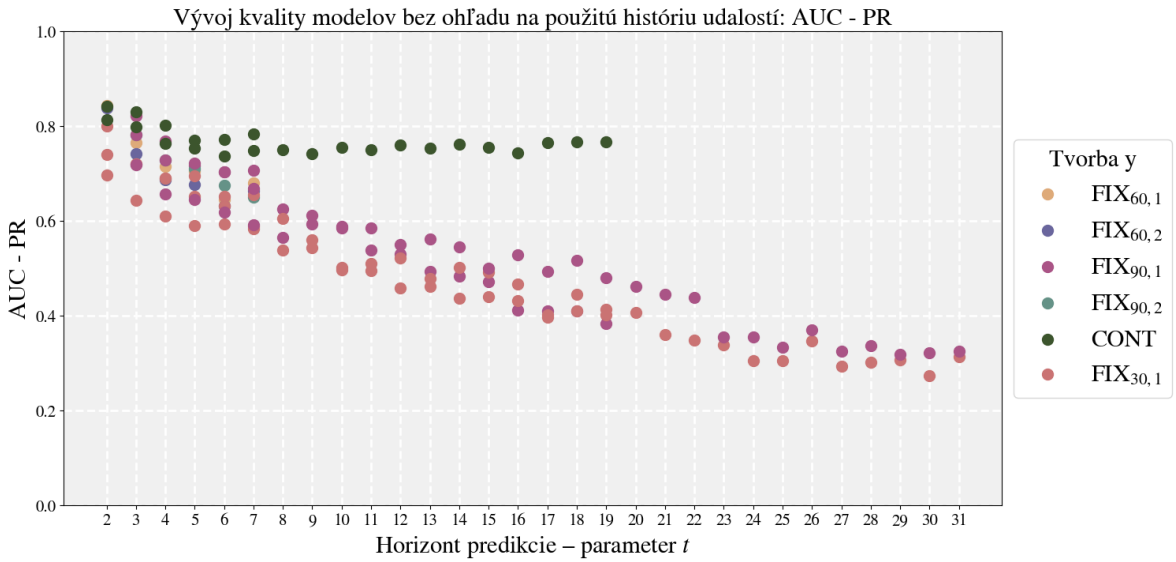
kde:

- $\boldsymbol{\beta}$ je vektor koeficientov logistickej regresie,
- t je hodnota parametra zodpovedného za vzdialenosť horizontu predikcie,
- $hist$ je dĺžka histórie udalostí použitej v rámci tvorby dátovej sady v mesiacoch (v práci sme používali hodnoty: 36, 24 a 12),
- x_{CONT} je binárna premenná. Dosahuje hodnotu 1 v prípade, že sme použili CONT_t spôsob tvorby závislej premennej,
- $x_{FIX_{90}}$ je binárna premenná. Dosahuje hodnotu 1 v prípade, že sme použili spôsob $\text{FIX}_{90,1,t}$ alebo $\text{FIX}_{90,2,t}$ tvorby závislej premennej,
- $x_{FIX_{60}}$ je binárna premenná. Dosahuje hodnotu 1 v prípade, že sme použili spôsob $\text{FIX}_{60,1,t}$ alebo $\text{FIX}_{60,2,t}$ tvorby závislej premennej.

Ako sme už spomenuli v podkapitole č. 3.2.3, modely vytvorené na základe spôsobov $\text{FIX}_{60,1,t}$ a $\text{FIX}_{60,2,t}$, ako aj $\text{FIX}_{90,1,t}$ a $\text{FIX}_{90,2,t}$, majú totožné časové rozpätie predikcie. Ich jediným rozdielom je počet vybraných čŕt klienta pri tvorbe pozitívnej vzorky. Z tohto dôvodu tieto modely agregujeme do spoločnej premennej, čím zabezpečíme zjednodušenie analýzy a elimináciu redundantných premenných v analýze dát.

Model logistickej regresie bude slúžiť na zachytenie vzťahov v rámci predikčného modelovania odchodovosti klientov, pričom zohľadní všetky tri identifikované časové faktory. Konkrétne sa bude zaoberať horizontom predikcie, použitou históriou udalostí a dĺžkou časového rozpätia predikcie.

Doterajšie výsledky hodnotenia kvality modelov sme prezentovali s ohľadom na veľkosť použitej histórie udalostí. Priebeh vývoja kvality aj tvar logistických kriviek naznačovali takmer lineárny charakter. Pri ich spoločnom zobrazení však možno pozorovať sigmoidný tvar s postupným spomalením poklesu kvality, najmä po odmyslení dlhodobých predikcií pri použití spôsobu CONT_t, ktorého vývoj bol podrobne opísaný v predchádzajúcich častiach.



Obr. 19: Agregovaný graf vývoja kvality modelov vzhľadom na rastúci horizont predikcie a bez ohľadu na použitú históriu udalostí. Celkový počet dátových bodov: 146. *Zdroj: Vlastné spracovanie.*

V tabuľke č. 5 prezentuje odhadnuté hodnoty koeficientov modelu logistickej regresie a pomer šancí na perfektný model v prípade zväčšenia hodnoty o 1.

Koeficient	**Hodnota**	**Štandardná odchylka $\hat{\sigma}_\beta$**	**P-value**	**OR: e^{Koef}**
$\hat{\beta}_0$	0.644	0.799	0.420	–
$\hat{\tau}$	-0.060	0.028	0.033	0.942
$\hat{\beta}_{hist}$	0.009	0.026	0.737	1.009
$\hat{\beta}_{\text{CONT}}$	0.902	0.593	0.128	2.463
$\hat{\beta}_{\text{FIX}_{90}}$	0.186	0.397	0.639	1.204
$\hat{\beta}_{\text{FIX}_{60}}$	0.236	0.802	0.769	1.266

Tabuľka 5: Tabuľka s odhadnutými koeficientami logistickej regresii, základnými štatistikami a vypočítanou hodnotou pomeru šancí OR na dosiahnutie perfektného modelu v prípade zväčšenia hodnoty v premennej o 1. Celkový počet dátových bodov: 146. *Zdroj: Vlastné spracovanie.*

Odhadnutý koeficient pre horizont predikcie $\hat{\tau}$ je záporný, čo je konzistentné s výsledkami pozorovanými v neagregovaných dátach. Hodnota tohto koeficientu dosiahnutá pri použití 12-mesačnej histórie udalostí sa pohybovala na podobnej úrovni, pričom pri využití 24-mesačnej histórie sa výsledky posunuli na mierne vyššie hodnoty. Na druhej strane, pri analýze 36-mesačnej histórie udalostí dochádzalo k výraznej zmene v hodnote koeficientu, čo naznačuje, že rozšírenie historických dát malo vplyv na výsledky predikcie.

Tento očistený výsledok možno interpretovať nasledovne: **pri zväčšení horizontu predikcie o jednu jednotku času v parametri t dochádza k zníženiu šancí na dosiahnutie perfektného modelu 0.94-násobne.** Tento pokles kvality modelov je obzvlášť významný v kontexte dlhodobých predikcií. Napríklad pri predĺžení horizontu o 12 mesiacov by šance na perfektný model poklesli o 55% z pôvodnej hodnoty. To podčiarkuje dôležitosť starostlivej voľby vzdialenosti horizontu predikcie pri tvorbe predikčných modelov.

Naopak, paralelný posun, ktorý sme evidovali pri zmenšovaní použitej histórie udalostí, sa prejavuje aj v hodnote koeficientu β_{hist}. Tento koeficient je kladný, čo indikuje, že modely využívajúce rozsiahlejšiu históriu majú vyššiu šancu dosiahnuť lepšiu presnosť, teda byť „perfektné“ v širšom zmysle. Je však potrebné podotknúť, že v rámci nášho experimentálneho nastavenia sme s týmto parametrom nehýbali do extrémov a preto nemožno definitívne generalizovať jeho správanie na všetky prípady.

Pri používaných nastaveniach zahŕňajúcich väčší počet dát, sa prejavuje pozitívny vplyv rozšírenej histórie. **Konkrétne, ak zväčšíme rozsah použitej histórie udalostí o 12 mesiacov, šanca dosiahnuť perfektný model sa zvýši 1.112-násobne.** Tento výsledok jednoznačne poukazuje na dôležitosť zahrnutia rozsiahlejšej histórie klientských dát pri tvorbe predikčných modelov.

Treba však zohľadniť aj praktické aspekty implementácie takéhoto prístupu. Zvýšenie objemu historických dát je spojené s vyššími nárokmi na úložné kapacity a spracovateľský výkon. Navyše, tvorba predikčného modelu s väčšou históriou vyžaduje náročnejšiu optimalizáciu hyperparametrov metód strojového učenia, čo predstavuje finančné náklady na výpočtovú techniku. Preto je pre spoločnosti dôležité, aby mali vyhradené dostatočné prostriedky na pokrytie týchto dodatočných nákladov, čo je kľúčové pre efektívne využitie rozšírenej histórie pri predikčnom modelovaní.

Pri pohľade na koeficienty β_{CONT}, $\beta_{\text{FIX}_{90}}$ a $\beta_{\text{FIX}_{60}}$ pozorujeme, že všetky hodnoty sú pozitívne. **Z toho vyplýva, že každý z uvedených spôsobov tvorby pozitívnej vzorky prispieva k zvýšeniu šancí na dosiahnutie perfektného modelu v porovnaní s $\mathbf{FIX}_{30,1}$.** Striktnosť výberu dátumu čŕt klientov v rámci spôsobu $\text{FIX}_{30,1}$ sa odrazila aj v tejto časti analýzy.

Pri analýze časového rozpätia a vplyvu striktnosti pri výbere dátumu črty klienta možno pozorovať výrazný prínos použitia $\mathbf{CONT}_t$ spôsobu tvorby závislej premennej. **Zvýšenie šancí sa pohybuje na úrovni 2.5-násobného zlepšenia, nezávisle od nastavenia horizontu predikcie t a veľkosti histórie udalostí pre dátovú sadu použitú na tvorbu predikčného modelu.**

Záver

Horizont predikcie má zásadný vplyv na kvalitu predikčných modelov, no nejde o jediný časový faktor, ktorý ju ovplyvňuje. V práci boli predstavené alternatívne prístupy tvorby závislej premennej s konštantným rozpätím predikcie. Modely využívajúce tieto prístupy dosahovali nižšiu predikčnú výkonnosť v porovnaní s modelmi s kontinuálne rozširujúcim sa rozpätím predikcie, čo bolo viditeľné na obrázku č. 19. Ich využitie je však opodstatnené v prípadoch, keď je potrebné predikovať udalosť v presne stanovenom časovom období.

Modely s rozširujúcim sa rozpätím ťažia z charakteru trénovacích dát, kde historická informácia predstavuje len malú časť. Skutočné rozpätie, v ktorom sú tieto modely efektívne, môže byť v dôsledku toho zošikmené.

Smerovanie výsledkov korešponduje s trendmi prezentovanými v [1]. Na rozdiel od prác [1] a [2] sme analyzovali širšie spektrum časových aspektov pri tvorbe závislej premennej a pracovali sme s údajmi z výrazne dlhších časových úsekov. Z toho možno usudzovať, že predlžovanie dĺžky predikcie má pozitívny vplyv len pri krátkodobých predikciách, kde závislé premenné majú až náhodný charakter, ako je to v prípade trhovej ceny aktíva. Naopak, v prípadoch, kde sa v dátach môžu nachádzať opakujúce sa vzorce správania (napr. v tejto práci v rámci odchodovosti alebo v [1]), je výhodnejšie tieto udalosti predikovať z novších dát.

Z uvedeného vyplýva, že pri návrhu predikčného modelu je nevyhnutné stanoviť optimálny časový horizont, ktorý zároveň umožní spoľahlivú identifikáciu vzorcov v historických údajoch a minimalizuje informačnú degradáciu spôsobenú nadmerným časovým odstupom. V prípade využitia modelu založeného na tvorbe závislej premennej $\mathrm{FIX}_{90,1,7}$ by pri použití 36-mesačnej histórie udalostí nedošlo k signifikantnému poklesu výkonnosti modelu. Zároveň by však takto definovaný model poskytol marketingovým oddeleniam výrazne väčší časový priestor na prípravu cielenej kampane, s cieľom ovplyvniť rozhodnutie klienta ešte pred jeho potenciálnym odchodom.

V oblasti predikcie odchodovosti klientov sme preukázali, že dĺžka rozpätia predikcie a využitá história udalostí sú dôležitým aspektom v rámci tvorby predikčného modelu. Rozšírenie historického obdobia má potenciál zlepšiť kvalitu predikcie, avšak len do určitého bodu. V našom prípade sa tento bod neprejavil. Avšak tempo zlepšovania výkonnosti modelov pri pridávaní historických dát sa postupne spomaľovalo. Zväčšenie

použitej histórie udalostí o jeden rok by znamenalo nárast šancí na perfektný model o 11%, čo sme prezentovali v podkapitole č. 5.4.

Zväčšovanie rozpätia predikcie malo taktiež pozitívny vplyv pri modelovaní odchodovosti klientov. Výsledky ukazujú, že časové rozhodnutie klientov o odchode nie je úplne deterministické. Tento jav je zreteľný pri modeloch s kontinuálne sa rozširujúcim rozpätím predikcie, ktoré dokážu flexibilnejšie reagovať na neurčitosť v správaní klientov, čo sme videli v podkapitole č. 5.2. U modelov s použitím kontinuálne sa zväčšujúcim sa rozpätím predikcie sa zastavil pokles kvality pri hodnote parametra $t = 8$. V prípade, že nie je potrebné vedieť presný okamih nastania udalosti, tak modely s kontinuálne rozširujúcim spôsobom tvorby závislej premennej dosahujú výrazne kvalitnejšie výsledky.

Ako pozitívum a prínosom môžeme taktiež považovať to, že ani pri dlhodobom predikovaní, ktoré sa môže zdať ako nereálne, kvalita modelov sa nedostala na úroveň modelu náhodnej klasifikácie. Aj v prípade predikovania na viac ako 31 mesiacov, kvalita predikčných modelov bola lepšia ako u náhodného klasifikátora.

Zoznam použitej literatúry

[1] Cheng, P. a kol.: *Predicting Inpatient Acute Kidney Injury over Different Time Horizons: How Early and Accurate?* Dostupné na internete (9.4.2024): https://pmc.ncbi.nlm.nih.gov/articles/PMC5977670/.

[2] Jaquart, P., Dann, D., Weinhardt, C.: *Short-term bitcoin market prediction via machine learning*, The Journal of Finance and Data Science, vol. 7, november 2021, s. 45–66. Dostupné na internete (9.4.2024): https://www.sciencedirect.com/science/article/pii/S2405918821000027?ref=pdf_download&fr=RR-2&rr=86f2395b7c3fc270.

[3] Ke, G., Meng, Q., Finley, T. a kol.: *LightGBM: A Highly Efficient Gradient Boosting Decision Tree.* Dostupné na internete (9.4.2024): https://www.researchgate.net/publication/378480234_LightGBM_A_Highly_Efficient_Gradient_Boosting_Decision_Tree.

[4] Natekin, A., Knoll, A.: *Gradient boosting machines, a tutorial*, Frontiers in neurorobotics, vol. 7, december 2013. Dostupné na internete (7.4.2025): https://www.researchgate.net/publication/259653472_Gradient_Boosting_Machines_A_Tutorial.

[5] Hastie, T., Tibshirani, R., Friedman, J.: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2. vyd., Springer, 2009. ISBN 978-0-387-84858-7.

[6] Vujović, Z.-D. *Classification Model Evaluation Metrics*, International Journal of Advanced Computer Science and Applications, vol. 12, jún 2021, s. 1–8. Dostupné na internete (30.3.2025): https://www.researchgate.net/publication/352902406_Classification_Model_Evaluation_Metrics.

[7] Muller, M.: *Generalized Linear Models*, Fraunhofer Institute for Industrial Mathematics, január 2004. Dostupné na internete (7.4.2025): https://www.researchgate.net/publication/228396785_Generalized_Linear_Models.

[8] Jurafsky, D., Martin, J. H.: *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3. vyd., 2024, s. 77–80.

[9] Bult, J.-R., Wansbekk, T.: *Optimal selection for direct mail*, Marketing science, vol. 14, november 1995, s. 378–394. Dostupné na internete (10.1.2025): https://www.researchgate.net/publication/227442110_Optimal_Selection_for_Direct_Mail.

[10] Birunda, S., Devi, R.-K.: *A Review on Word Embedding Techniques for Text Classification*, Innovative Data Communication Technologies and Application, vol. 59, 2020, s. 267–281. Dostupné na internete (7.4.2025): https://www.researchgate.net/publication/348989232_A_Review_on_Word_Embedding_Techniques_for_Text_Classification.

[11] Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: *Enriching Word Vectors with Subword Information*, 2016. Dostupné na internete (19.12.2024): https://arxiv.org/pdf/1607.04606.

[12] Mikolov, T., Chen, K., Corrado, G., Dean, J.: *Efficient Estimation of Word Representations in Vector Space*, Innovative Data Communication Technologies and Application, január 2020. Dostupné na internete (7.4.2025): https://www.researchgate.net/publication/234131319_Efficient_Estimation_of_Word_Representations_in_Vector_Space.

[13] Grave, E., Bojanowski, P., Gupta, P., Joulin, A., Mikolov, T.: *Learning Word Vectors for 157 Languages*, Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018), 2018. Dostupné na internete (1.6.2025): https://fasttext.cc/docs/en/crawl-vectors.html.

[14] Gencaga, D., Malakar, N., Lary, D.-J.: *Survey On The Estimation Of Mutual Information Methods as a Measure of Dependency Versus Correlation Analysis*, 2016. Dostupné na internete (19.12.2024): https://www.researchgate.net/publication/259743124_Survey_On_The_Estimation_Of_Mutual_Information_Methods_as_a_Measure_of_Dependency_Versus_Correlation_Analysis.

[15] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: *Scikit-learn: Machine Learning in Python*, Journal of Machine Learning Research, vol. 12, 2011, s. 2825–2830.

[16] Dorogush, A.-V., Ershov, V., Gulin, A.: *CatBoost: gradient boosting with categorical features support*, CoRR, vol. abs/1810.11363, 2018. Dostupné na internete (14.4.2025): https://arxiv.org/abs/1706.09516.

[17] Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M.: *Optuna: A Next-generation Hyperparameter Optimization Framework*, Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2019.

[18] Berrar, D.: *Cross-validation*, Encyclopedia of Bioinformatics and Computational Biology, vol. 1, január 2018, s. 542–545. Dostupné na internete (4.1.2025): https://www.researchgate.net/publication/324701535_Cross-Validation.

[19] Belete, D.-M., Huchaiah, D.-M.: *Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results*, International Journal of Computers and Applications, september 2021. Dostupné na internete (19.4.2025): https://www.researchgate.net/publication/354541720_Grid_search_in_hyperparameter_optimization_of_machine_learning_models_for_prediction_of_HIVAIDS_test_results.

[20] Bergstra, J., Bardenet, J., Bengio, Y., Kégl, B.: *Advances in Neural Information Processing Systems 24*, International Journal of Computers and Applications, 2011. Dostupné na internete (19.4.2025): https://papers.nips.cc/paper_files/paper/2011/hash/86e8f7ab32cfd12577bc2619bc635690-Abstract.html.

Prílohy

A Vzájomná informácia

Vzájomná informácia je matematicky definovaná nasledovne pre diskrétny prípad [14]:

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} P(x,y) \ln \left(\frac{P(x,y)}{P(x)P(y)} \right), \tag{A.1}$$

kde:

- $I(X,Y)$ – vzájomná informácia medzi premennou X a Y
- $P(x,y)$ – združená pravdepodobnosť X a Y
- $P(x)$ – marginálna pravdepodobnosť premennej X
- $P(y)$ – marginálna pravdepodobnosť premennej Y

B Rozdiel medzi mriežkovým vyhľadávaním a Optuna

Optimalizácia hyperparametrov pomocou mriežkového vyhľadávania (anglický výraz: *grid search*) [19] spočíva v definovaní konečnej množiny hodnôt pre každý hyperparameter modelu. Nech $H_1, H_2, \ldots, H_k$ predstavujú hyperparametre a nech $V_i = \{v_{i1}, v_{i2}, \ldots, v_{in_i}\}$ je množina hodnôt, ktoré môže nadobúdať hyperparameter H_i, pričom n_i reprezentuje počet hodnôt v množine. Mriežkové vyhľadávanie následne skúma kartézsky súčin $V_1 \times V_2 \times \cdots \times V_k$, pričom pre každú kombináciu $\mathbf{v} = [v_1, v_2, \ldots, v_k] \in V_1 \times \cdots \times V_k$ sa trénuje model a hodnotí jeho výkonnosť na základe vopred definovanej metrík kvality. Najlepšia kombinácia hyperparametrov je určená ako tá, ktorá maximalizuje (resp. minimalizuje) zvolenú metriku. Hlavnou nevýhodou tejto metódy je jej výpočtová neefektívnosť v prostredí s vysokou dimenzionalitou, keďže počet trénovaných modelov rastie exponenciálne s počtom hyperparametrov.

`Optuna` je rámec na optimalizáciu hyperparametrov, ktorý v našom prípade používa metódu TPE (skrátené z anglického výrazu: *Tree-structured Parzen Estimator*)[17]. Tento prístup sa odlišuje od tradičných metód ako mriežkové vyhľadávanie tým, že sa nezaoberá len systematickým testovaním všetkých možných hodnôt hyperparametrov, ale inteligentne sa zameriava na hodnoty, ktoré majú vyššiu šancu na lepší výkon.

Začína sa náhodným vzorkovaním, a následne sa pri každom novom pokuse vyberajú hyperparametre na základe historických dát, ktoré ukazujú, ktoré hodnoty priniesli najlepšie výsledky v rámci optimalizovanej funkcie F_{opt}. Hyperparametre $\mathbf{v}$ sa vyberajú v rámci iterácie j tak, aby sa maximalizoval výraz [20]:

$$\mathbf{v}_j^* = \max_{\mathbf{v}} \frac{g(\mathbf{v})}{l(\mathbf{v})}, \tag{B.1}$$

kde:

- $\mathbf{v}_j^*$ je vektor optimálnych hyperparametrov v rámci iterácie `Optuna`,
- $g(\mathbf{v}) = P\left(\mathbf{v} \mid F_{opt}^j \geq q\left(F_{opt}^{1:j-1}\right)\right)$ pravdepodobnosť výskytu vektora v, ak vieme, že pri nej bola vysoká hodnota funkcie F_{opt},
- $l(\mathbf{v}) = P\left(\mathbf{v} \mid F_{opt}^j < q\left(F_{opt}^{1:j-1}\right)\right)$ pravdepodobnosť výskytu vektora v, ak vieme, že pri nej bola nízka funkcie F_{opt},
- $q\left(F_{opt}^{1:j-1}\right)$ predstavuje kvantil predchádzajúcich výsledkov na zvolenom percentile.

Pravdepodobnosti $l(\mathbf{v})$ a $g(\mathbf{v})$ sa odhadnú neparametricky pomocou Parzenovho odhadu (anglický výraz: *Parzen estimator*). Na konci algoritmu `Optuna` sa vyberie vektor hyperparametrov $\mathbf{v}^*$, ktorý prislúchal najväčšej hodnote F_{opt}.

`Optuna` algoritmus taktiež obsahuje prvky, ktoré šetria čas výpočtu a to na základe toho, že dokážu v rámci jednej iterácie algoritmu `Optuna` zastaviť výpočet, ak sa neposúva dostatočne rýchlo. Toto šetrí výpočtovú náročnosť.