UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY




# MODELOVANIE PODMIENENEJ HUSTOTY DENNEJ AKTIVITY KLIENTOV BANKY

DIPLOMOVÁ PRÁCA

2022 Bc. Jozef GAGO

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

# MODELOVANIE PODMIENENEJ HUSTOTY DENNEJ AKTIVITY KLIENTOV BANKY

## DIPLOMOVÁ PRÁCA

| | |
|---|---|
| Študijný program: | Ekonomicko-finančná matematika a modelovanie |
| Študijný odbor: | 1114 Aplikovaná matematika |
| Školiace pracovisko: | Katedra aplikovanej matematiky a štatistiky |
| Vedúci práce: | Mgr. Martin Hurban, PhD. |

Bratislava 2022 **Bc. Jozef GAGO**






Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

# ZADANIE ZÁVEREČNEJ PRÁCE

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

**Názov:** Modelovanie podmienenej hustoty dennej aktivity klientov banky
*Modeling of conditional density of daily activity for bank clients*

**Anotácia:** Modelovanie podmienenej hustoty je zovšeobecnením regresnej úlohy, kde sa modeluje podmienená stredná hodnota. Modelovanie hustoty je zaujímavé najmä v prípadoch, kedy je podmienená hustota multimodálna. V práci sa budeme zaoberať modelovaním intenzity dennej aktivity klientov banky podmienenej mnohorozmernou charakteristikou klienta, pričom dôraz bude kladený na implementáciu algoritmov uvažujúcich veľký počet premenných.

**Vedúci:** Mgr. Martin Hurban, PhD.
**Katedra:** FMFI.KAMŠ - Katedra aplikovanej matematiky a štatistiky
**Vedúci katedry:** prof. RNDr. Marek Fila, DrSc.

**Dátum zadania:** 13.01.2021

**Dátum schválenia:** 17.01.2021 prof. RNDr. Daniel Ševčovič, DrSc.
garant študijného programu

.....
študent

.....
vedúci práce

# Abstrakt v štátnom jazyku

GAGO, Jozef: Modelovanie podmienenej hustoty dennej aktivity klientov banky [Diplomová práca], Univerzita Komenského v Bratislave, Fakulta matematiky, fyziky a informatiky, Katedra aplikovanej matematiky a štatistiky; školiteľ: Mgr. Martin Hurban, PhD., Bratislava, 2022, 49s.

Táto práca obsahuje 2 rôzne modely optimálneho oslovenia klientov banky v závislosti od času v priebehu dňa implementovateľné v softvéri, ktorý je dostupný v call centre. Uvažujeme pritom veľký počet ďalších premenných. Vzhľadom na možnosti pracovných nástrojov operátorov call centra to znamená, že v konkrétnom časovom okne budeme hľadať klienta, ktorý má najvyššiu pravdepodobnosť, že bude dostupný. Túto pravdepodobnosť určíme na základe predpočítaných pravdepodobností v rovnomerne zvolených diskrétnych bodoch daného časového okna. Celkové výsledky ukazujú možné zvýšenie efektivity call centra o približne 19%. Základ nášho modelu je postavený na logistickej regresii, pomocou ktorej modelujeme pravdepodobnosť, že danom čase bude klient, ktorý splna nejaké ďalšie charakteristiky, zastihnuteľný. V práci využívame splajnovú regresiu na modelovanie nelineárnej závislosti a interakcie dvoch premenných. Na príklade transakčných premenných ukazujeme ako možno implementovať funckionálne premenné v regresnom modeli.

**Kľúčové slová:** Logistická regresia, Splajnová regresia, Model obsluhy klienta

# Abstract

GAGO, Jozef: Modeling of conditional density of daily activity for bank clients [Master Thesis], Comenius University in Bratislava, Faculty of Mathematics, Physics and Informatics, Department of Applied Mathematics and Statistics; Supervisor: Mgr. Martin Hurban, PhD., Bratislava, 2022, 49p.

This thesis contains 2 different models of optimal addressing of the bank's clients depending on the time during the day, implementable in the software available in the call center. We consider many of other variables. Considering the capabilities of call center operators' work tools, we will look for the client who is most likely to be available in a particular time interval. We will determine this probability on the basis of the read probabilities in evenly selected discrete points of the given time interval. Our results show that the possible increase of call center efficiency is approximately 19%. Our model is based on logistic regression, which we use for model the probability that at a given time a client who meets some other characteristics will be reachable. Our model is based on logistic regression, which we use for model the probability that at a given time a client who meets some other characteristics will be reachable. In this work we use spline regression to model the nonlinear dependence and interaction of two variables. Using the transaction variables, we show how to implement functional variables in a regression model.

**Keywords:** Logistic regression, Spline regression, Client servicing

# Obsah

**Úvod** **8**

**1 Model efektívneho oslovovania klientov v priebehu dňa** **10**

1.1 Formulácia prolému . . . . . 10

**2 Príprava dát** **12**

2.1 Závislá premenná . . . . . 12

2.2 Prediktory - funkcionálne premenné . . . . . 14

2.2.1 Identifikácia aktivít, ktoré odlišujú denný harmonogram klientov . . 18

2.2.2 Spôsob použitia v modeloch . . . . . 19

2.3 Prediktory - spojité premenné . . . . . 21

**3 Logistická regresia a jej varianty** **24**

3.1 Nelineárna závislosť v regresnom modeli . . . . . 24

3.2 Viacrozmerné splajny . . . . . 25

3.3 Transformácia funkcionálnych premenných . . . . . 27

3.4 Implementácia v jazyku Python . . . . . 28

**4 Model dennej intenzity klienta** **30**

4.1 Základný demografický model . . . . . 30

4.2 Model používajúci denné intenzity nakupovania . . . . . 31

4.2.1 Integrácia hustôt . . . . . 31

4.2.2 Redukcia dimenzie použitím bázy kubických splajnov . . . . . 32

4.3 Model používajúci spojité premenné . . . . . 34

4.4 Model využívajúci spojité a funkcionálne premenné . . . . . 34

**5 Výsledky modelu efektívneho oslovenia klientov** **36**

**Záver** **40**

**Zoznam použitej literatúry** **42**

**Príloha A** **44**

**A.1 Kullback-Leiblerova divergencia** **44**

**Príloha B** **46**

**B.1 Základné metódy numerického integrovania** **46**

B.1.1 Lichobežníková metóda . . . . . 46

B.1.2 Simpsonova metóda . . . . . 47

**Príloha C** **49**

**C.1 ROC krivka a AUC** **49**

# Úvod

Pre banky aj pre klientov je výhodné identifikovať časy, kedy je možné nadviazať kontakt s vyššou pravdepodobnosťou úspešnosti. Zastihnuteľnosť klienta sa behom dňa mení a možno ju charakterizovať ako podmienenú hustotu jeho dennej aktivity, respektíve dennej intenzity. Modelovanie dennej intenzity je možné robiť pri binárnych dátach pomocou nástrojov binárnej klasifikácie, ako pri aj funkcionálnych dátach pomocou modelovania podmienenej hustoty [7]. Keďže každý telefonát má v čase priradenú informáciu o úspešnosti, dennú intenzitu klienta možno modelovať pomocou logistickej regresie [6].

V tomto prípade má logistická regresia ako nástroj binárnej klasifikácie má tú výhodu, že nám umožňuje vyjadriť predpoveď úspešnosti hovoru ako pravdepodobnosť. Takto sa dá v každom čase nájsť klient, ktorý má najvyššiu pravdepodobnosť, že bude práve zastihnuteľný.

V lineárnych regresných modeloch sa predpokladá lineárna závislosť cieľovej premennej od prediktorov. Ako je ukázané v [6], tento predpoklad nie je obmedzujúci a možno ho obísť tým že, je uvažovaná vhodná nelineárna transformáciu vstupných prediktorov, napríklad projekcia prediktorov do priestoru bázových funkcií kubického splajnu.

Podobne podľa [6] využijúc tento typ bázového rozšírenie prediktorov sa dá do lineárneho modelu zakomponovať interakcia dvoch prediktorov pomocou tenzorového súčinu ich bázových funkcií. Tento spôsob umožňuje priamo zohľadniť previazanosť dvoch prediktorov.

Implementácia funkcionálnych premenných do regresného modelu je popísaná v [6], kde bolo ukázané, že aj tu sa dá využiť báza kubických splajnov. Definičný obor funkcionálnych premenných sa dá rozdeliť na ľubovoľne veľa diskrétnych bodov a hodnoty v týchto bodoch následne použiť ako prediktory v lineárnom modeli. Počet prediktorov sa však dá výrazne zredukovať práve projekciou do priestoru kubických splajnov.

Tieto rozšírenia lineárneho modelu boli implementované do programovacieho jazyka Python prostredníctvom knižnice PyGAM [11]. Táto implementácia umožňuje jednoduchým spôsobom voliť spôsob, ako chceme konkrétny prediktor zahrnúť do regresného modelu.

Cieľom práce je navrhnúť model, maximalizuje efektivitu call centra kontrolovaním času kontaktovania jednotlivých klientov. V našej práci sformulujeme modely, ktoré na

základe modelu logistickej regresie v danom časovom okne v priebehu dňa určia, ktorý klient má najvyššiu pravdepodobnosť, že bude zastihnuteľný. Tento model budeme trénovať na dátach predstavujúcich záznamy hovorov za obdobie 10 mesiacov a následne vyhodnotíme na sade dát za jeden ďalší mesiac. K dispozícii máme dáta od Československej obchodnej banky (ČSOB) ktoré obsahujú informácie o čase a dĺžke hovoru, základné demografické údaje o klientoch, ktorým bolo volané, časové údaje o transakciách platobnými kartami a širokú škálu ďalších premenných týkajúcich sa využitia rozličných bankových služieb v rámci ČSOB.

Práca je rozdelená do piatich kapitol. V prvej predstavíme a formálne popíšeme úlohu, ktorá zodpovedá problému call centra. Navrhneme tu algoritmus oslovovania klientov, ktorý v danom čase určí klienta s najvyššou pravdepodobnosťou zastihnutia. Druhá kapitola bude zameraná na popis a analýzu dát, ktoré budeme v modeli používať. Tretia kapitola obsahuje popis techník spracovania jednotlivých premenných pre účely vývoja nášho modelu dennej intenzity klienta. Budeme sa tu venovať logistickej regresii a jej rozšíreniam, pomocou ktorej budeme v konkrétnom čase schopný určiť pravdepodobnosť úspešnosti hovoru konkrétnemu klientovi. V štvrtej kapitole sa venujeme trénovaniu modelu dennej intenzity na bankových dátach za obdobie 10 mesiacov. Tieto dáta obsahujú záznamy z približne 215 000 hovorov. V poslednej, piatej kapitole sa budeme venovať vyhodnoteniu nášho modelu na základe jednoduchých Monte Carlo simulácií.

# 1 Model efektívneho oslovovania klientov v priebehu dňa

Model oslovenia je náš navrhovaný nástroj, pomocou ktorého budeme vedieť určiť optimálny čas kontaktovania klienta. Z pohľadu banky ide o snahu vyhnúť sa zbytočnému volaniu ľuďom, ktorí sú práve zaneprázdnení, čo môže spôsobiť nespokojnosť zo strany klienta a v neposlednom rade aj plytvanie zdrojmi banky. Budeme teda hľadať pravdepodobnosti, že klienti sú v danom časovom okne počas dňa zastihnuteľní. Následne môžeme v každom časovom okne tieto pravdepodobnosti zoradiť a nájsť klientov, ktorí sú v danom časovom okne zastihnuteľní s najväčšou pravdepodobnosťou.

## 1.1 Formulácia prolému

Tento model sme navrhli ako funkciu intenzity klienta $I_i$, t.j. ako pravdepodobnosť, že klient bude zastihnuteľný v konkrétnom čase $t$ za podmienky, že spĺňa ďalšie charakteristiky. Toto môžeme formálne napísať ako

$$I_i(t) = P(Y_i = 1 | T = t, X = x_i) \tag{1}$$

kde $i$ je index pre klienta, $i = 1, \ldots, K$, $Y$ ako závislá premenná hovorí o úspešnosti kontaktovania klienta pracovníkom banky, $T$ je čas kontaktovania v priebehu dňa a $X \in \mathbb{R}^n$ sú pozorované vlastnosti klienta do času $t$, ako napríklad vek, pohlavie, preferované časy nakupovania v potravinách, preferované časy prihlasovania do aplikácií elektronického bankovníctva a podobne. Dajú sa navrhnúť pomerne jednoduché pravidlá, ako optimálne rozdeliť klientov podľa toho, kedy ich osloviť v priebehu dňa. Vzhľadom k tomu, že operátor call centra nevie dopredu ako sa deň bude vyvíjať, uvažujeme optimálne zoradenie klientov v najbližšom časovom okne $\langle t_{\text{start}}, t_{\text{end}} \rangle$ podľa pravdepodobnosti, že ak bude realizovať jeden telefonát, tak bude úspešný. Formálne budeme hľadať riešenie teoretickej úlohy

$$\arg\operatorname*{sort}_i \left( \frac{\int_{t_{\text{start}}}^{t_{\text{end}}} I_i(s) ds}{d} \right), \tag{2}$$

kde $d = t_{end} - t_{start}$ je dĺžka príslušného časového intervalu. Podobne celý deň vieme rozdeliť na $D$ podintervalov dĺžky $d$, t,j dostaneme podintervaly $\langle t_j, t_{j+1} \rangle$, pre $j = 0, \ldots, D$. Pre efektívnu aproximáciu funkcie $I_i(t)$ môžeme pre $i$-teho klienta v $j$-tom časovom okne

predpočítať sadu hodnôt $\hat{I}_{ij}$

$$\hat{I}_{ij} = \frac{\int_{t_j}^{t_{j+1}} I_i(s)ds}{t_{j+1} - t_j}. \tag{3}$$

Potom môžeme urobiť odhad

$$\frac{\int_{t_{\text{start}}}^{t_{\text{end}}} I_i(s)ds}{d} \approx \sum_{j \in \tau} \hat{I}_{ij} = \hat{I}_i, \tag{4}$$

kde $\tau$ je množina indexov spínajúca $\tau = \{j | t_{start} \leq t_j \leq t_{end}, \quad j = 0, \ldots, D\}$.

Paralelný postup pri použití na reálnych dátach je definovať aproximáciu funkcie $I_{ij}$ pomocou vierohodnostnej funkcie nasledovne:

$$\hat{I}_{ij}^L = I_i \left( \frac{t_j + t_{j+1}}{2}. \right) \tag{5}$$

Potom pre každého klienta môžeme hľadať optimálny interval ako

$$\max_{j \in \tau} (\hat{I}_{ij}^L) = \hat{I}_i^L \tag{6}$$

a na každom intervale $\langle t_j, t_{j+1} \rangle$ budeme riešiť úlohu pomocou aproximácie

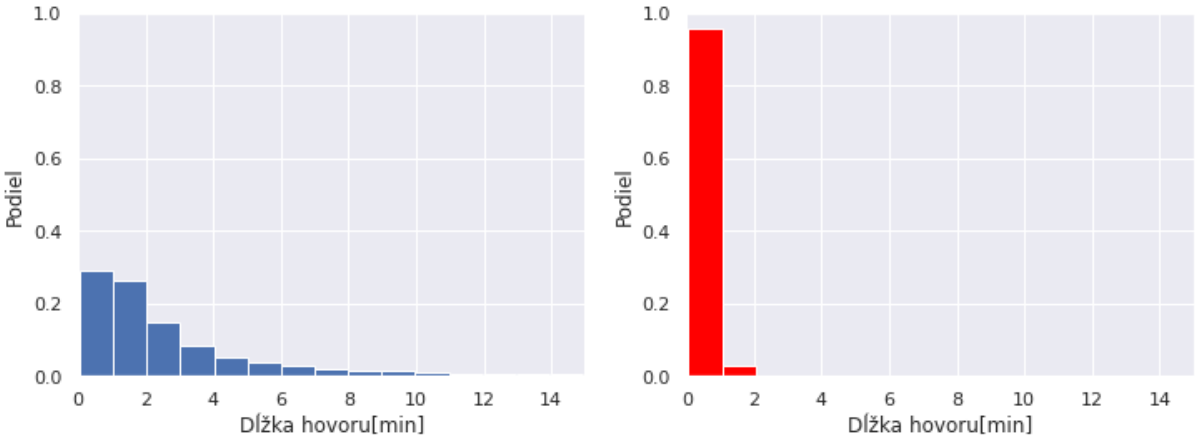
$$\arg\operatorname*{sort}_i \left( I_i^L \right). \tag{7}$$

# 2 Príprava dát

V našom modeli sú ako sledované objekty klienti banky. O každom klientovi máme k dispozícii základné demografické dáta ako sú vek a pohlavie. Okrem nich budeme používať množstvo ďalších dát, ktoré priblížime v tejto kapitole. Všetky pochádzajú z interných databáz ČSOB. Predstavíme dátové sady pre všetky vysvetľujúce premenné, ktoré použijeme v našom modeli.

## 2.1 Závislá premenná

Záznamy hovorov z call centra predstavujú závislú premennú v našom modeli. Sú to hovory, ktoré realizovali zamestnanci banky smerom k zákazníkom. Každý záznam obsahuje informáciu o dĺžke trvania hovoru a výsledku daného hovoru, t.j. či bol hovor úspešný, alebo nie. Túto informáciu sa budeme snažiť našim modelom predikovať. V tejto práci budeme pracovať s dátami za obdobie od 1.2.2021 do 31.12.2021. Spomedzi týchto záznamov je v databáze približne 55% klasifikovaných ako úspešné hovory. Na obrázku 1 môžeme vidieť histogramy trvania úspešných a neúspešných hovorov. Dá sa vidieť, že úspešné hovory majú tendenciu trvať dlhšie ako neúspešné.

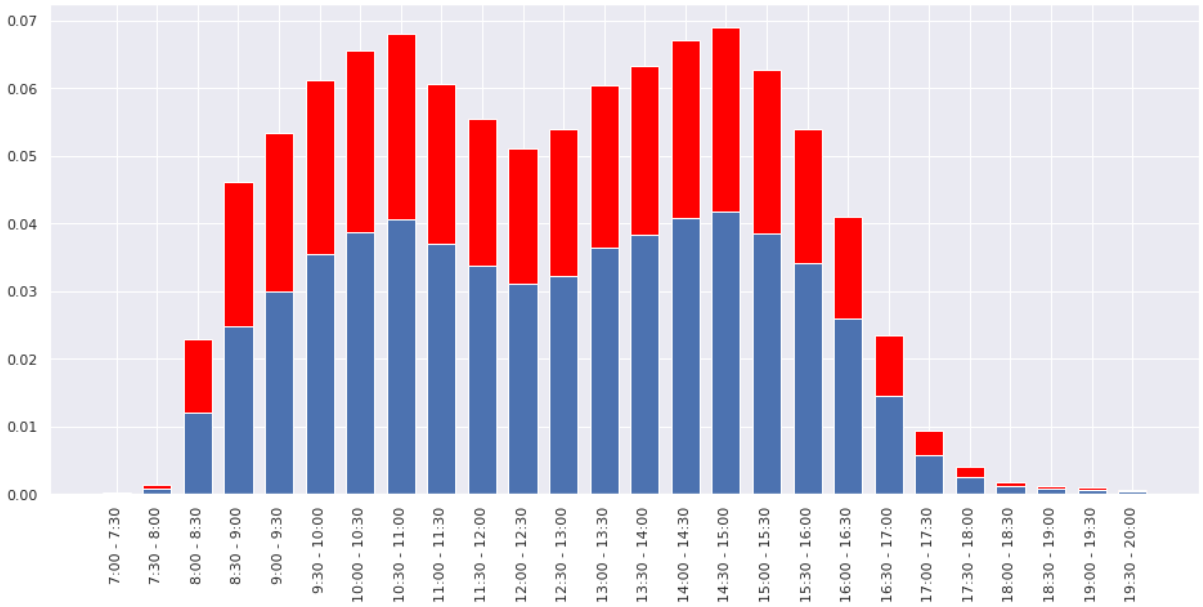



**Obr. 1:** *Histogramy dĺžky trvania hovorov. Vľavo histogram pre úspešné hovory, vpravo histogram pre neúspešné hovory.*

Hlavná črta, ktorú budeme používať na predikciu úspešnosti hovoru, je čas kedy bol realizovaný. Predovšetkým nás bude zaujímať, kedy bol hovor realizovaný v priebehu dňa. Táto informácia nám môže poslúžiť na lepšie odhadnutie správania sa väčšej skupiny ľudí,

ktorí si môžu byť v zmysle nejakých špecifických čŕt podobní.

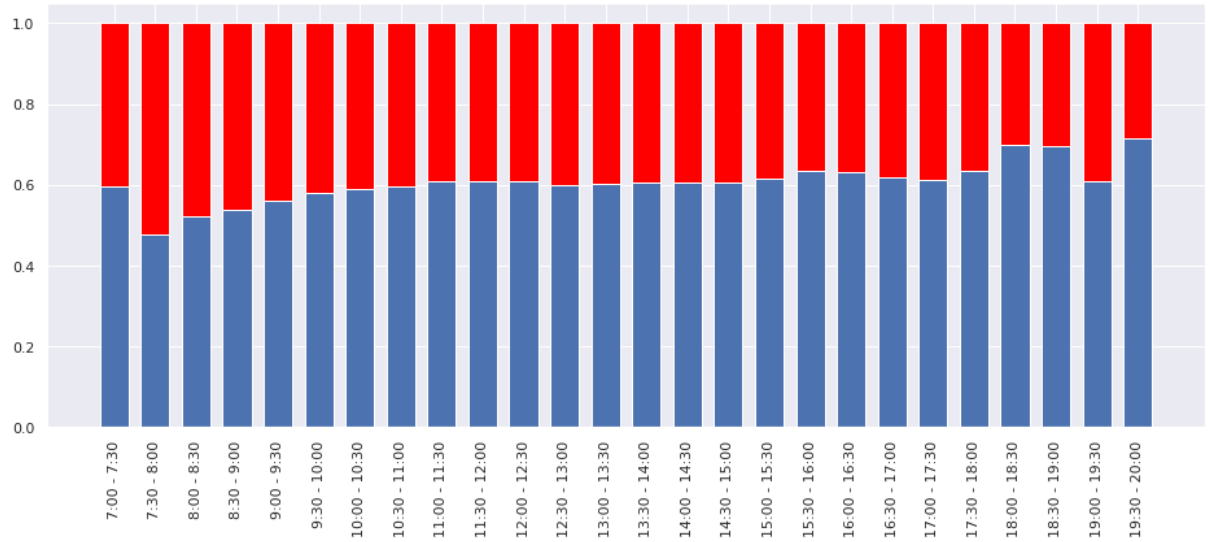
Tento čas bude do modelu vstupovať ako desatinné číslo a bude uvedený v hodinách. Na obrázku 2 je histogram realizovaných hovorov, ktoré sú rozdelené do polhodinových intervalov. Môžeme vidieť, že najpodstatnejšia časť hovorov je realizovaná od 8:30 do 16:30, ďalej môžeme vidieť dva vrcholy dopoludnia a popoludní a údolie v čase obeda od 12:00 do 12:30. Na obrázku 3 je zobrazený podiel úspešných a neúspešných hovorov v



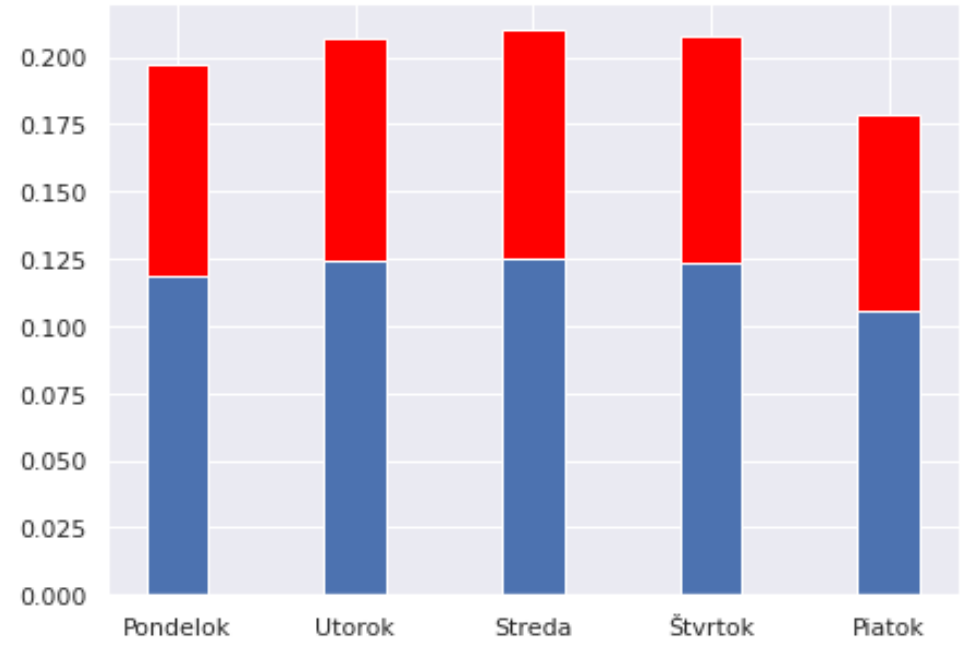

**Obr. 2:** *Histogram časov kontaktovania klientov počas dňa. Každý stĺpec zodpovedá príslušnému polhodinovému intervalu. Modrá časť stĺpcov zodpovedá hovorom, ktoré boli úspešné a červená časť zodpovedá neúspešným hovorom.*

rovnakých polhodinových intervaloch. V časovom okne od 8:30 do 16:30 kedy bolo realizovaných jednoznačne najviac hovorov môžeme vidieť, že podiel úspešných a neúspešných hovorov sa až na pár menších rozdielov priebehu dňa nemení.

Podobne zaujímavá informácia môže byť deň v týždni, kedy bol realizovaný hovor. Na obrázku 4 môžeme vidieť, že najviac hovorov bolo realizovaných v strede týždňa a najmenej piatok. Ďalej na obrázku 5 môžeme vidieť, že podiel úspešných a neúspešných hovorov sa v priebehu týždňa mení len nepatrne.



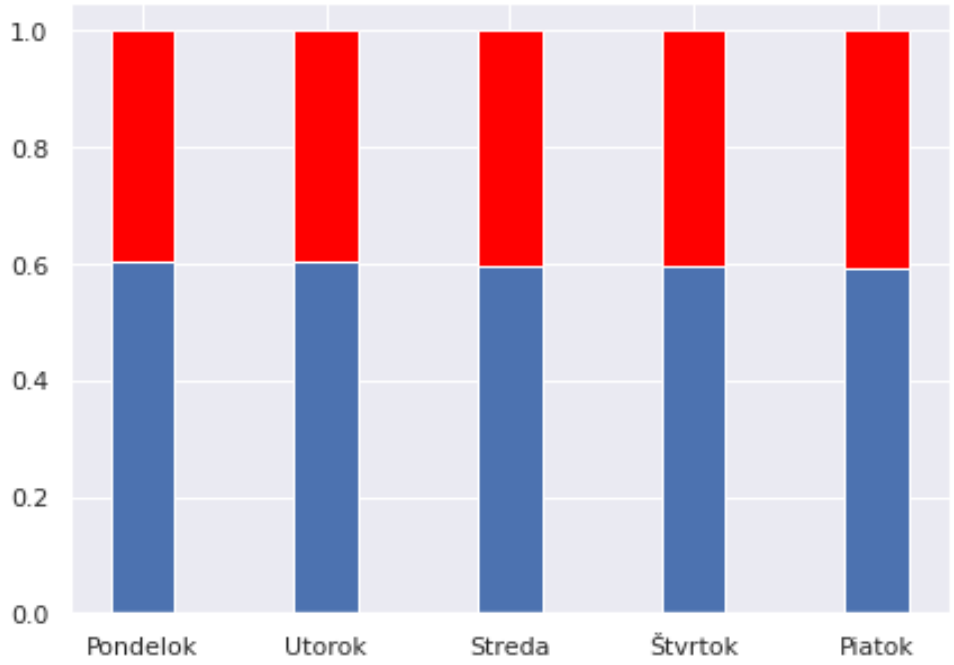

**Obr. 3:** *Podiel úspešných a neúspešných hovorov počas dňa v polhodinových intervaloch. Môžeme si všimnúť mierny nárast úspešnosti od* 8:30 *do* 11:30 *a od* 15:30 *do* 16:00*.*




**Obr. 4:** *Rozdelenie úspešných a neúspešných hovorov podľa dní v týždni. Modrá časť stĺpcov zodpovedá hovorom, ktoré boli úspešné a červená časť zodpovedá neúspešným hovorom.*

## 2.2 Prediktory - funkcionálne premenné

Záznamy kartových transakcií môžu predstavovať zaujímavú črtu pri hľadaní optimálneho času kontaktovania klienta. Využívame pritom logickú úvahu, že ak by sme poznali



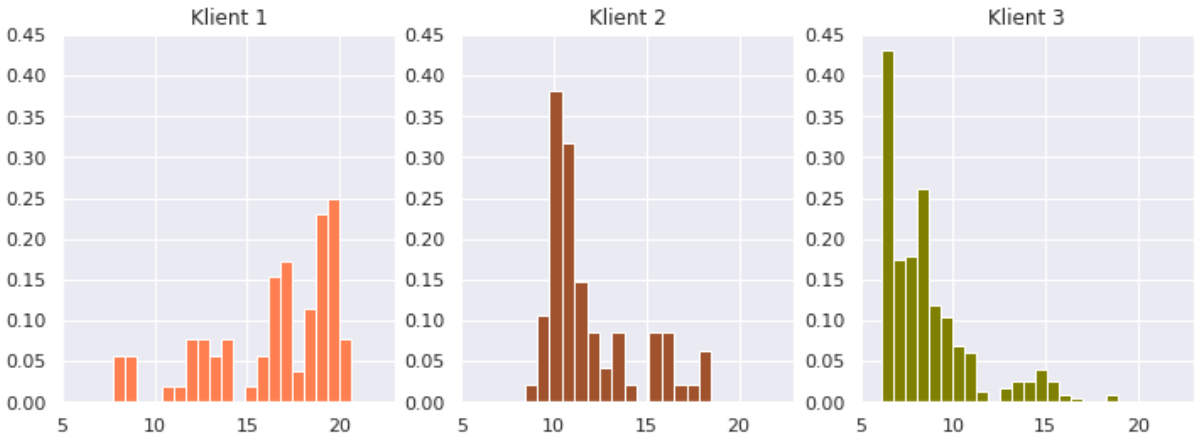

**Obr. 5:** *Podiel úspešných a neúspešných hovorov podľa dní v týždni. V závislosti od jednotlivých dní sa výrazne nemení.*

pravidelné časy, kedy daný klient obvykle nakupuje v potravinách, drogériách, lekárňach, obchodoch s oblečením a obuvou alebo tankuje na čerpacích staniciach, mohlo by nám to pomôcť lepšie odhadnúť, či je v danom čase vhodné klientovi volať, alebo nie. Aj keď sú tu klienti, o ktorých máme z rôznych dôvodov takýchto záznamov málo, pri značnej skupine klientov máme takýchto záznamov dostatočné množstvo na to, aby sme vedeli odhadnúť, v ktorých časoch nakupujú najčastejšie.

K dispozícii máme záznamy časov autorizácie platby na POS termináloch, ktoré klienti realizovali prostredníctvom kreditnej alebo debetnej karty. Tieto údaje budeme používať ako jednu z charakteristických čŕt pri predikcii, či je v nejakom konkrétnom čase vhodné volať klientovi alebo nie. Na modelovanie úspešnosti konkrétneho hovoru použijeme záznamy kartových transakcii, ktoré boli realizované v polročnom časovom intervale pred realizáciou hovoru.

Z týchto údajov nás budú najviac zaujímať časy realizácií transakcií v priebehu dňa. Tak dostaneme ku každému klientovi sadu časov v rozpätí $\langle 0, 24)$. Toto môžeme chápať aj ako realizácie náhodnej premennej s neznámou hustotou, ktorá predstavuje náhodný čas kedy konkrétny klient navštívi obchod konkrétnej kategórie. Ďalej tomu budeme hovoriť

denná intenzita nakupovania.




**Obr. 6:** *Histogramy dennej intenzity nakupovania pre troch rôznych klientov, pre ktorých máme aspoň 30 záznamov.*

Na obrázku 6 môžeme vidieť ukážku histogramov denných intenzít nakupovania. Vybrali sme tu troch klientov, ktorí v polročnom časovom intervale realizovali aspoň 30 platieb v potravinách a supermarketoch. Ako sa dalo predpokladať, tento obrázok ukazuje, že každý klient môže preferovať iné časy nákupov. Vystáva tu otázka, že do akej miery sa v priemere odlišuje denná aktivita klientov banky v zmysle preferovaných časov nakupovania. Na to sme sa pozreli pomocou Kullback-Leiblerovej divergencie dvoch denných intenzít nakupovania. Aby sme mohli použiť tento spôsob porovnania dvoch distribúcii, použili sme na odhad hustoty dennej intenzity nakupovania jadrový odhad hustoty (KDE) v gaussovským jadrom. Parameter *bandwidth* sme pri týchto odhadoch nastavili na hodnotu 0.8.

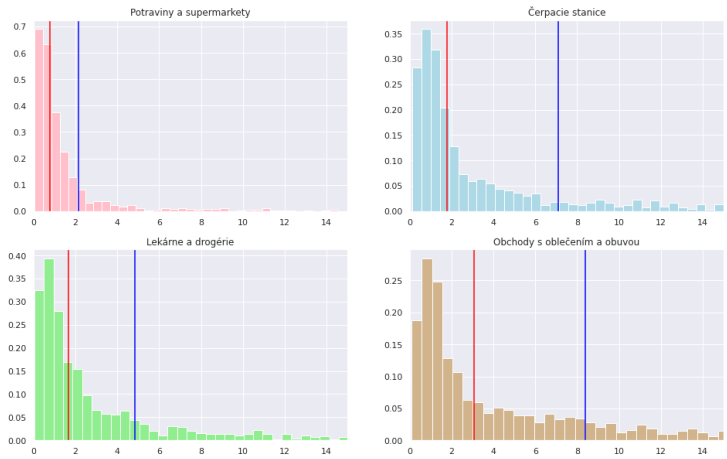
Okrem platieb v potravinách a supermarketoch budeme používať aj ďalšie kategórie, ktoré nám môžu tiež pomôcť odhadnúť denné správanie klientov. Konkrétne sú to kategórie čerpacie stanice, lekárne a drogérie a obchody s oblečením a obuvou. V tabuľke 1 sme zhrnuli základné informácie o týchto kategóriach.

Na odhad KL divergencie pre denné intenzity nakupovania našich klientov sme z každej každej kategórie vybrali 1000 náhodných dvojíc pozostávajúcich z aspoň jedného pozorovania, pre ktoré sme vypočítali najprv odhad hustoty pomocou KDE a následne pre tieto odhadnuté hustoty sme vypočítali KL divergencie. Histogramy Výsledných hodnôt pre náhodný výber 1000 dvojíc v každej kategórii spolu s priemermi a mediánmi môžeme

| | Podiel klientov s aspoň 1 záznamom | Podiel klientov s aspoň 30 záznamami | Priemerný počet záznamov na klienta | Priemerná hodnota času nákupu | Medián KL divergencie | Priemer KL divergencie |
|---|---|---|---|---|---|---|
| Potraviny a supermarkety | 0.7415 | 0.4355 | 46.57 | 13.8502 | 0.8147 | 2.0182 |
| Čerpacie stanice | 0.6057 | 0.0854 | 15.88 | 13.7975 | 1.8897 | 6.7965 |
| Lekárne a drogérie | 0.6315 | 0.0311 | 9.89 | 13.8237 | 1.6215 | 4.6005 |
| Obchody s oblečením a obuvou | 0.5385 | 0.0080 | 6.19 | 15.0393 | 2.8296 | 8.6615 |

**Tabuľka 1:** *Porovnanie denných intenzít nakupovania podľa jednotlivých kategórii vrátane štatistických ukazovateľov vypočítaných odhadov KL divergencií pre 1000 náhodných intenzít, ktoré pozostávajú z aspoň jedného pozorovania.*

vidieť na obrázku 7.




**Obr. 7:** *KL divergencie pre KDE 1000 náhodných dvojíc denných intenzít nakupovania pre jednotlivé kategórie. Červené vertikálne čiary predstavujú priemer a modré medián.*

### 2.2.1 Identifikácia aktivít, ktoré odlišujú denný harmonogram klientov

Na rozumné zhodnotenie vypočítaných hodnôt KL divergencie sme použili sme dve jednorozmerné normálne normálne rozdelenia s parametrami $(\mu_1, \sigma_1)$ a $(\mu_2, \sigma_2)$. Keďže pre tieto normálne rozdelenia poznáme hustoty, môžeme pre ne jednoducho odvodiť hodnotu KL divergencie. Označme $p(x) = \frac{1}{\sqrt{2\pi}\sigma_1} \exp^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}$ a $q(x) = \frac{1}{\sqrt{2\pi}\sigma_2} \exp^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}}$. KL divergenciu týchto rozdelení môžeme počítať nasledovne:

$$D_{KL}(p\|q) = E_p[\ln(p) - \ln(q)] =$$

$$= E_p\left[\ln\left(\frac{1}{\sigma_1\sqrt{2\pi}} \exp^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}\right) - \ln\left(\frac{1}{\sigma_2\sqrt{2\pi}} \exp^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}}\right)\right] =$$

$$= E_p\left[\ln\left(\frac{1}{\sqrt{2\pi}}\right) + \ln\left(\frac{1}{\sigma_1}\right) - \frac{(x-\mu_1)^2}{2\sigma_1^2} - \ln\left(\frac{1}{\sqrt{2\pi}}\right) - \ln\left(\frac{1}{\sigma_2}\right) + \frac{(x-\mu_2)^2}{2\sigma_2^2}\right] =$$

$$= E_p\left[\ln\left(\frac{\sigma_2}{\sigma_1}\right) - \frac{(x-\mu_1)^2}{2\sigma_1^2} + \frac{(x-\mu_2)^2}{2\sigma_2^2}\right] =$$

$$= E_p\left[\ln\left(\frac{\sigma_2}{\sigma_1}\right)\right] - \frac{1}{2}E_p\left[\frac{(x-\mu_1)^2}{\sigma_1^2}\right] + \frac{1}{2}E_p\left[\frac{(x-\mu_2)^2}{\sigma_2^2}\right] =$$

$$= \ln\left(\frac{\sigma_2}{\sigma_1}\right) - \frac{1}{2}E_p\left[\frac{(x-\mu_1)^2}{\sigma_1^2}\right] + \frac{1}{2}E_p\left[\frac{(x-\mu_2)^2}{\sigma_2^2}\right].$$

Vieme, že výraz $\frac{(x-\mu_1)^2}{\sigma_1^2}$ má rozdelenie $\chi_1^2$, preto $E_p\left[\frac{(x-\mu_1)^2}{\sigma_1^2}\right] = 1$, kde $E_p$ predstavuje strednú hodnotu pri pravdepodobnostnej miere $p$. Ďalej môžeme počítať

$$E_p\left[\frac{(x-\mu_2)^2}{\sigma_2^2}\right] = E_p\left[\frac{(x-\mu_1+\mu_1-\mu_2)^2}{\sigma_2^2}\right] =$$

$$= \frac{1}{\sigma_2^2}E_p\left[(x-\mu_1)^2 + 2(x-\mu_1)(\mu_1-\mu_2) + (\mu_1-\mu_2)^2\right] =$$

$$= \frac{1}{\sigma_2^2}\left[\sigma_1^2 + 0 + (\mu_1-\mu_2)^2\right] = \frac{(\mu_1-\mu_2)^2}{\sigma_2^2} + \frac{\sigma_1^2}{\sigma_2^2}. \tag{8}$$

Na záver môžeme výsledok sformulovať ako tvrdenie vety 2.1.

**Veta 2.1.** *Nech $p$ je hustota normálneho rozdelenia $N(\mu_1, \sigma_1^2)$ a $q$ je hustota normálneho rozdelenia $N(\mu_2, \sigma_2^2)$. Potom KL divergenciu rozdelení $p$ a $q$ môžeme počítať pomocou vzorca*

$$D_{KL}(p\|q) = \ln\left(\frac{\sigma_2}{\sigma_1}\right) - \frac{1}{2} + \frac{(\mu_1-\mu_2)^2}{2\sigma_2^2} + \frac{\sigma_1^2}{2\sigma_2^2}. \tag{9}$$

*Dôkaz.* Dôkaz tejto vety je uvedený pred jej formuláciou. □

Teraz sa môžeme pozrieť ako sa mení hodnota KL divergencie dvoch normálnych rozdelení v závislosti od ich parametrov. Aby sme vedeli pomocou tejto závislosti zhodnotiť naše vypočítané odhady KL divergencie denných intenzít nakupovania v obchodoch jednotlivých kategórii, urobili sme dve pozorovania.

Pozorovanie 1

V prvom prípade sme porovnali KL divergenciu dvoch normálnych rozdelení $N(\mu_1, \sigma_1)$ a $N(\mu_2, \sigma_2)$, pričom parametre $\sigma_1$ a $\sigma_2$ sme zvolili rovnaké. Presnejšie, zobrali sme hodnotu výberovej disperzie denných intenzít nakupovania pre všetkých klientov v kategórii potraviny a supermarkety, čo bola hodnota približne 3.8. Na obrázku 8 vľavo môžeme vidieť, ako sa mení hodnota KL divergencie v závislosti od rozdielu stredných hodnôt $\mu_1 - \mu_2$.

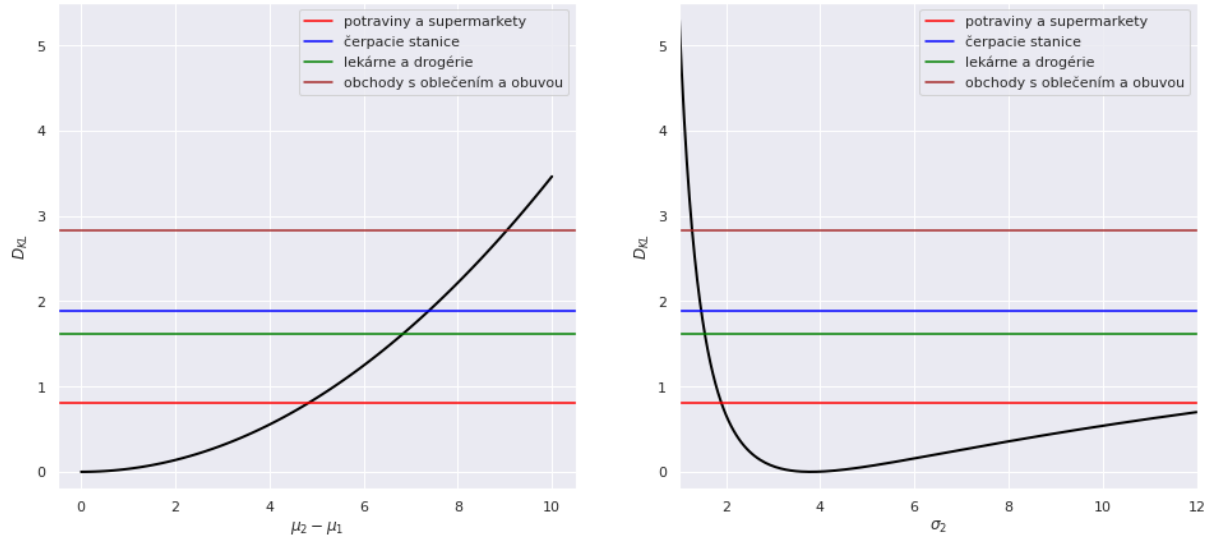
Pozorovanie 2

V druhom prípade sme použili rovnaké stredné hodnoty, t.j. $\mu_1 = \mu_2$. Smerodajnú odchýlku sme použili rovnakú, ako v predošlom pozorovaní, $\sigma_1 = 3.8$ a pozreli sme sa na závislosť od $\sigma_2$. Výsledok môžeme vidieť na obrázku 8 vpravo.

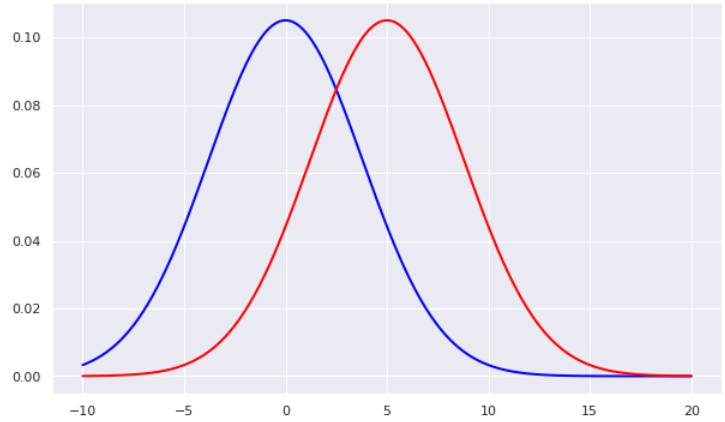
Na obrázku 8 sú horizontálnymi čiarami znázornené mediánové hodnoty odhadnutých KL divergencii pomocou KDE pre 1000 náhodných dvojíc denných intenzít nakupovania pre jednotlivé kategórie. Na základe týchto obrázkov môžeme povedať, že pravdepodobnostné rozdelenie dennej intenzity nakupovania dvoch náhodne vybraných klientov sa v priemere líši podobne, ako dve normálne rozdelenia s rovnakou smerodajnou odchýlkou 3.8 a rozdielom stredných hodnôt približne 5. Pre lepšiu predstavu uvádzame obrázok 9.

### 2.2.2 Spôsob použitia v modeloch

Na modelovanie úspešnosti konkrétneho hovoru budeme používať dennú intenzitu nakupovania. Náš zámer je používať tieto dátové sady ako funkčné premenné. Takéto funkcie môžeme dostať napríklad použitím jadrového odhadu hustoty, ako môžeme vidieť na obrázku 10. Definičným oborom takýchto funkcií bude časové okno jedného dňa, t.j. $\langle 0, 24 \rangle$. Definičný obor rozdelíme na na intervaly dĺžky $d$ a pozrieme sa, koľkokrát klient platil v danom type obchodov v danom časovom intervale za posledný polrok. Takýmto spôsobom na spôsob histogramu dostaneme obor hodnôt. Tieto funkčné hodnoty následne použijeme



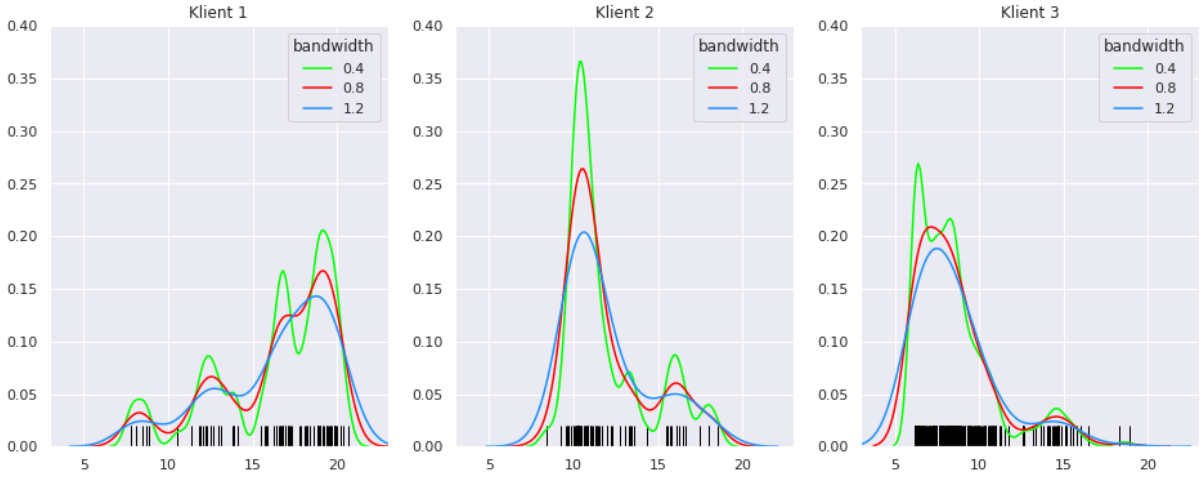

**Obr. 8:** *Vľavo závislosť KL divergencie od rozdielu stredných hodnôt dvoch normálnych rozdelení s rovnakou smerodajnou odchýlkou* $\sigma_1 = \sigma_2 = 3.8$. *Vpravo závislosť KL divergencie dvoch normálnych rozdelení s rovnakou strednou hodnotou, kde smerodajná odchýlka prvej distribúcie bola konštantná,* $\sigma_1 = 3.8$ *a smerodajná odchýlka druhej distribúcie sa pohybuje od 1 po 12. Horizontálne čiary ukazujú úrovne odhadnutej KL divergencie denných intenzít nakupovania pre jednotlivé kategórie.*



**Obr. 9:** *Hustoty dvoch normálnych rozdelení so strednými hodnotami* $\mu_1 = 0$, $\mu_2 = 5$ *a s rovnakými smerodajnými odchýlkami* $\sigma_1 = \sigma_2 = 3.8$.

na modelovanie úspešnosti hovorov.

Veľkosť intervalov $d$ sme nastavili na hodnotu 1/3 hodiny, t.j. 20 minút. Takýto interval môžeme považovať za dostatočne presný z hľadiska presnosti nášho navrhovaného modelu a zároveň výpočtovo zvládnuteľný.



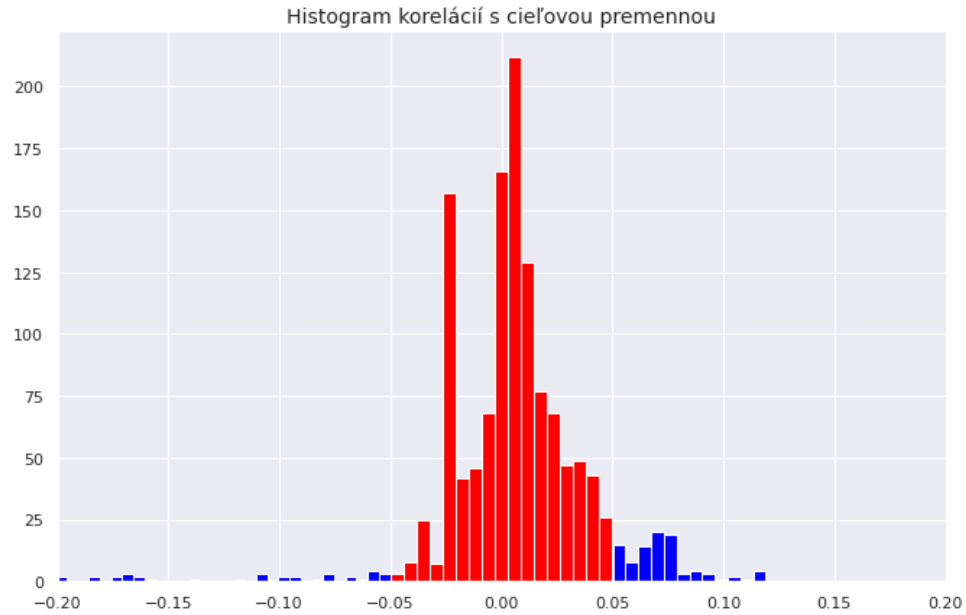

**Obr. 10:** *Jadrový odhad hustoty pre dáta z obrázka 6. Použili sme gaussovské jadro a hodnoty parametra bandwidth postupne* 0.4, 0.8, 1.2.

## 2.3 Prediktory - spojité premenné

V tejto časti v skratke predstavíme sadu spojitých premenných, ktoré budeme používať na modelovanie pravdepodobnosti úspešnosti hovoru. Tieto premenné sme vyberali z veľkého množstva premenných, ktoré nejakým spôsobom charakterizujú klienta. Pri výbere týchto premenných sme zohľadňovali 2 dôležité kritériá.

**Korelácia so závislou premenou:** V prvom rade sme vyberali tie premenné, ktoré mali najvyššiu absolútnu hodnotu Pearsonovho korelačného koeficientu s cieľovou premennou, t.j. informáciou o úspešnosti hovoru. V tomto prípade sme ako minimálnu hodnotu Pearsonovho korelačného koeficientu zobrali 0.05.

**Vzájomná korelácia premenných:** Druhé kritérium výberu premenných bolo zamerané na korelácie medzi vybranými premennými navzájom. Keďže budeme využívať lineárny parametrický model logistickej regresie, potrebujeme aby boli tieto premenné nezávislé. Pri praktickom použití ak označíme $\mathbf{X}$ ako maticu pozostávajúcu z realizácii prediktorov vstupujúcich do modelu, pre numerickú stabilitu výsledkov potrebujeme zabezpečiť, aby matica $\mathbf{X}^T\mathbf{X}$ mala nízky spektrálny polomer. Preto v prípade, že hodnota Pearsonovho korelačného koeficientu nejakých dvoch premenných presiahla hodnotu 0.95, tak sme zobrali len tú, ktorá mala vyššiu absolútnu hodnotu s cieľovou premennou.

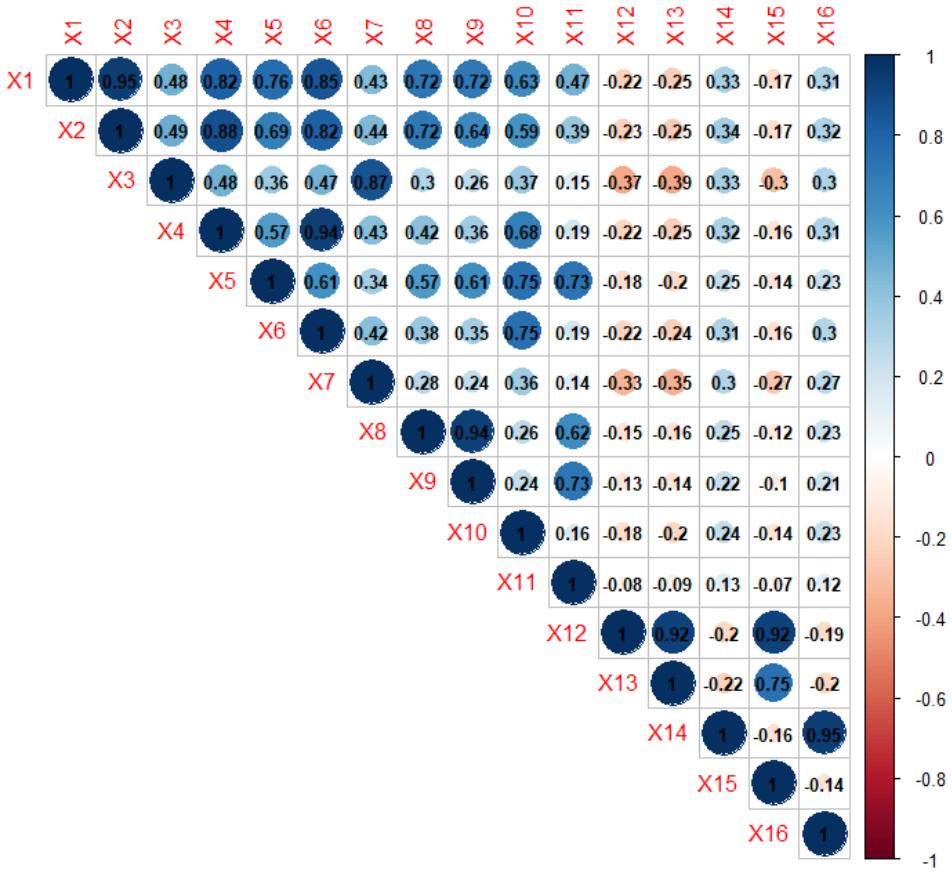



**Obr. 11:** *Histogram korelácií všetkých dostupných premenných s cieľovou premennou. V strede červenou farbou sú znázornené tie, ktoré majú hodnotu Pearsonovho korelačného koeficientu s cieľovou premennou menšiu ako* 0.05.

Premenné, ktoré majú najvyššiu absolútnu hodnotu korelačného koeficientu s cieľovou premennou sú charakterizované v tabuľke 2. Pre stručnosť uvádzame len tie, pri ktorých Pearsonov korelačný koeficient v absolútnej hodnote presiahol 0.1. Na obrázku 12 je znázornená ich korelačná matica. Z celkového počtu približne 1300 dostupných premenných vyššie spomenuté kritériá spĺňalo 89. Na obrázku 11 sme zobrazili histogram Pearsonovho korelačného koeficientu s cieľovou premennou pre všetkých 1300 premenných premenných, kde červenou farbou sme vyznačili tie, ktoré vyššie uvedené kritériá nespĺňajú.

| Označenie premennej | Korelácia s cieľovou premennou (Pearsonov korelačný koeficient) | Popis |
|---|---|---|
| X1 | -0.20045765 | Počet odchádzajúcich telefonátov za posledných 6 mesiacov |
| X2 | -0.19726312 | Počet odchádzajúcich telefonátov za posledných 12 mesiacov |
| X3 | -0.18259011 | Priemerný mesačný počet nepovolených prečerpaní |
| X4 | -0.17201941 | Počet relevantných hovorov za posledných 12 mesiacov |
| X5 | -0.17111782 | Celkový počet odchádzajúcich telefonátov |
| X6 | -0.17032257 | Počet relevantných hovorov za posledných 6 mesiacov |
| X7 | -0.16704839 | Počet nepovolených prečerpaní |
| X8 | -0.16342708 | Počet krátkych telefonátov za posledných 12 mesiacov |
| X9 | -0.15711401 | Počet krátkych telefonátov za posledných 6 mesiacov |
| X10 | -0.13896464 | Celkový počet relevantných hovorov |
| X11 | -0.11916293 | Celkový počet krátkych telefonátov |
| X12 | 0.11888214 | Normalizovaný počet trvalých príkazov |
| X13 | 0.11638304 | Počet pohyblivých trvalých príkazov |
| X14 | -0.10810781 | Priemerný normalizovaný počet exekúcií |
| X15 | 0.1058795 | Celkový počet trvalých príkazov |
| X16 | -0.10041727 | Normalizovaný počet exekúcií |

**Tabuľka 2:** *Prehľad základných spojitých premenných.*




**Obr. 12:** *Hodnoty korelačnej matice pre základné premenné z tabuľky 2. Popis premenných X1,X2,... X16 je rovnaký ako v tabuľke 2.*

# 3 Logistická regresia a jej varianty

Základný prvok nášho navrhovaného modelu je logistický aditívny regresný model. Pomocou logistickej regresie budeme modelovať pravdepodobnosť, že hovor, ktorý realizuje pracovník banky smerom ku klientovi bude úspešný, pričom výsledok bude závisieť od známych charakteristík klientov. Formálne zapísané

$$\ln \frac{P(\text{hovor bude úspešný})}{P(\text{hovor nebude úspešný})} = X^T \beta, \tag{10}$$

kde $X$ je $K \times n$ matica premenných charakterizujúcich klientov a $\beta \in \mathbb{R}^n$ je vektor koeficientov.

## 3.1 Nelineárna závislosť v regresnom modeli

V aditívnom regresnom modeli vo všeobecnosti modelovaná závislá premenná nezávisí od prediktorov lineárne. Lineárny model je neraz len jednoduchá aproximácia, ktorá z

realizácii prediktorov dokáže zobrať len malú informáciu. V [6], kapitola 5.2, môžeme nájsť spôsob, ako možno reflektovať nelineárny príspevok prediktorov.

Tento prístup funguje na princípe transformácie prediktora $X$ pomocou nejakých bázových funkcií. Ako príklad uvažujme jednoduchý model logistickej regresie s jednorozmerným prediktorom $x_1$ v tvare

$$\ln\left(\frac{P(Y=1)}{P(Y=0)}\right) = \beta_1 x_1 \tag{11}$$

a bázové funkcie $h_1(X) = X$, $h_2(X) = X^2$ a $h_3(X) = X^3$. Nelineárnu závislosť môžeme dosiahnuť, ak odhadneme model v tvare

$$\ln\left(\frac{P(Y=1)}{P(Y=0)}\right) = \theta_1 x_1 + \theta_2 x_1^2 + \theta_3 x_1^3, \tag{12}$$

kde nelineárna závislosť je vyjadrená pomocou súčinu bázových funkcií $h_1$, $h_2$ a $h_3$, ktoré prenásobíme príslušnými hodnotami odhadnutých koeficientov $\theta_1$, $\theta_2$ a $\theta_3$.

Všeobecne môžeme označiť $h_m(X) : \mathbb{R}^p \longrightarrow \mathbb{R}$ ako $m$-tú transformáciu $X \in \mathbb{R}^p$, $m = 1, \ldots, M$. Potom môžeme uvažovať model, ktorý predstavuje lineárne bázové rozšírenie v $X$

$$\ln\left(\frac{P(\text{hovor bude úspešný})}{P(\text{hovor nebude úspešný})}\right) = \sum_{m=1}^{M} \beta_m h_m(X), \tag{13}$$

kde $M$ je počet bázových funkcií.

Výhoda tohto prístupu je, že keď definujeme bázové funkcie $h_m$, tak model je už potom lineárny. Transformáciu $h_m$ môžeme zvoliť ľubovoľne. Budeme uvažovať konkrétnu voľbu $h_m$, ktorá bude pokrývať pomerne rozsiahlu triedu závislostí.

Rovnako v [6], kapitola 5.2, bolo ukázané, že ideálny kandidát je voľba $h_m$ ako bázové funkcie priestoru kubických splajnov. Ako príklad môžeme uvažovať model s nejakou jednorozmernou premennou $X$ v tvare

$$\ln\left(\frac{P(Y=1)}{P(Y=0)}\right) = \theta_0 + h_1(X)\theta_1 + \cdots + h_M(X)^T\theta_M, \tag{14}$$

kde $\theta_0, \ldots, \theta_M \in \mathbb{R}$ sú koeficienty, $h_i$ sú vektory bázových funkcií prirodzeného kubického splajnu.

## 3.2 Viacrozmerné splajny

V predošlej časti sme sa venovali modelu logistickej regresie s jednorozmernou nezávislou premennou $X \in \mathbb{R}$. Podľa [6], kapitola 5.7, sa tento postup dá rozšíriť aj pre viacrozmerné

premenné. Najprv sa pozrieme na model, kde $X \in \mathbb{R}^{K\times 2}$, kde $K$ je počet pozorovaní. V tomto prípade máme bázové funkcie $h_{1k}(X_1)$, $j = 1, \ldots, M_1$, pre premennú $X_1$ a $h_{2k}(X_2)$, $k = 1, \ldots, M_2$, pre premennú $X_2$. Potom môžeme definovať $M_1 \times M_2$ rozmernú bázu tenzorového súčinu $X_1 \otimes X_2$ ako

$$g_{jk}(X_1, X_2) = h_{1j}(X_1)h_{2k}(X_2), \quad j = 1, \ldots, M_1, \quad k = 1, \ldots, M_2 \tag{15}$$

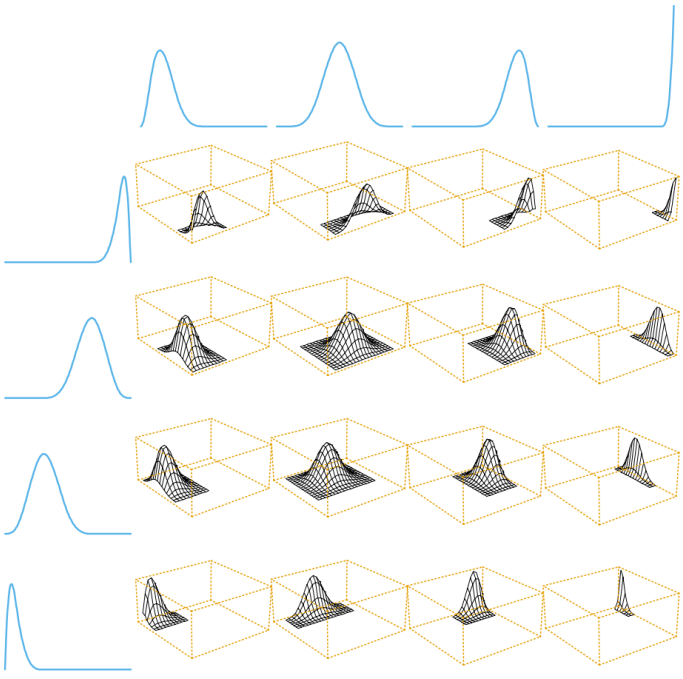
pomocou ktorého dostaneme dvojrozmernú funkciu

$$g(X_1, X_2) = \sum_{j=1}^{M_1}\sum_{k=1}^{M_2} \theta_{jk} g_{jk}(X_1, X_2). \tag{16}$$

Potom môžeme uvažovať model logistickej regresie v tvare

$$\ln\left(\frac{P(Y=1)}{P(Y=0)}\right) = \theta_0 + g(X_1, X_2). \tag{17}$$

Na obrázku 13 je ukážka tenzorového súčinu pre bázové funkcie prirodzeného kubického splajnu. Podobne by sme mohli uvažovať aj model pre $X \in \mathbb{R}^{K\times n}$, avšak je potrebné



**Obr. 13:** *Ukážka tenzorových súčinov bázových funkcii prirodzeného kubického splajnu pre niekoľko párov. Každá dvojrozmerná funkcia je tenzorový súčin dvoch marginálnych funkcii. (Zdroj: [6])*

uvedomiť si, že počet aditívnych členov s rastúcim $n$ rastie exponenciálne.

## 3.3 Transformácia funkcionálnych premenných

V ďalšej časti budeme diskutovať implementáciu spojitých premenných ako vstupy do modelu. Charakteristiky, ktoré budeme v našom modeli používať ako spojité premenné sú denné intenzity nakupovania. Pre tieto intenzity nakupovania vieme odhadnúť hustotu pomocou jadrového dohadu (*angl. kernel density estimator*, KDE), a tak dostaneme funkcie pomocou ktorých sa budeme snažiť modelovať pravdepodobnosť úspešnosti hovoru.

Denné intenzity nakupovania sú funkcie závislé od času v priebehu dňa, a teda môžeme ich označiť $f(T)$. Keďže regresory sú v tomto prípade nejaké spojité funkcie, aj koeficienty budú vyjadrené pomocou spojitej funkcie závislej od času $T$. Môžeme teda uvažovať spojitý model logistickej regresie v tvare

$$\ln\left(\frac{P(\text{hovor bude úspešný})}{P(\text{hovor nebude úspešný})}\right) = \int f(T)\cdot\beta(T)dT. \tag{18}$$

Najjednoduchší a najpriamočiarejší spôsob ako takýto model reálne použiť je rozdeliť vstupné funkcie na $N$ intervalov rovnakej dĺžky. Na týchto intervaloch môžeme spočítať integrál a tak dostaneme sadu diskrétnych hodnôt, ktoré môžeme následne použiť na modelovanie úspešnosti hovoru. Takto môžeme funckionálne premenné použiť v modeli logistickej regresie, ktorá má tvar

$$\ln\left(\frac{P(\text{hovor bude úspešný})}{P(\text{hovor nebude úspešný})}\right) = \sum_{i=1}^{N} x_i\cdot\beta_i, \tag{19}$$

kde $x_i$ predstavuje hodnotu integrálu dennej intenzity nakupovania v $i$-tom časovom okne. Koeficienty takéhoto modelu už nie je problém odhadnúť použitím vhodného softvéru.

Vo všeobecnosti tento prístup je náročný na výpočet, pretože odhadnúť model pre rádovo desiatky či stovky premenných môže byť náročné na operačnú pamäť. V [6], kapitola 5.2.3, je predstavený zaujímavý spôsob, ako znížiť počet regresorov, ktorý môže dokonca zlepšiť presnosť modelu. Princíp je založený podobne ako v 3.1, avšak v tomto prípade nám to umožní zredukovať počet premenných v modeli.

Aproximáciu funkcie $\beta(T)$ vieme vyjadriť pomocou bázových funkcii prirodzeného kubického splajnu $h_m$, t.j.

$$\beta(T) = \sum_{m=1}^{M} h_m(T)\theta_m. \tag{20}$$

Tieto funkcie majú všeobecný tvar

$$h_1(X) = 1, \quad h_2(X) = X, \quad h_{k+2}(X) = d_k(X) - d_{K-1}(X), \tag{21}$$

kde

$$d_k(X) = \frac{(X - \xi_k)_+^3 - (X - \xi_K)_+^3}{\xi_K - \xi_k}. \tag{22}$$

V praxi to znamená, že $\beta = H\theta$, kde $\beta \in \mathbb{R}^N$, $\theta \in \mathbb{R}^M$ a $H$ je $N \times M$ matica bázových funkcii prirodzeného kubického splajnu definovaného na definičnom obore našich funkcii intenzít s $M$ uzlami rovnomerne rozloženými na definičnom obore intenzít. Keďže $X^T\beta = X^T H\theta$, $X$ môžeme nahradiť $X^* = H^T X$ a odhadnúť lineárny model logistickej regresie v tvare

$$\ln\left(\frac{P(\text{hovor bude úspešný})}{P(\text{hovor nebude úspešný})}\right) = \sum_{m=1}^{M} X_m^* \theta_m. \tag{23}$$

Môžeme si všimnúť že $M$ je rádovo menšie ako $N$, čo znamená, že odhad parametrov modelu je výpočtovo jednoduchšia úloha.

## 3.4 Implementácia v jazyku Python

*Pygam* je knižnica v jazyku Python navrhnutá na tvorbu zovšeobecnených aditívnych modelov (*angl. generalised additive models, GAM*) ([11]), ktoré majú všeobecný tvar

$$E[Y|X] = \beta_0 + f_0(X_0) + f_1(X_1, X_2) + \cdots + f_m(X_{n-1}) \tag{24}$$

kde $X^T = [X_0, X_1, \ldots, X_{n-1}]$ sú nezávislé premenné, $Y$ je závislá premenná a $g$ je *link function*. Je to spôsob, ktorý umožňuje použiť nelineárne funkcie nezávislých premenných $X_i$ pri zachovaní aditivity. Pri našom modeli použijeme skôr jeho klasifikačnú verziu, kde je na mieste *link function* funkcia *logit*, t.j.

$$\ln \frac{P(Y=1)}{P(Y=0)} = \beta_0 + f_0(X_0) + f_1(X_1, X_2) + \cdots + f_m(X_{n-1}). \tag{25}$$

Knižnica *pygam* podporuje nasledovné transformácie nezávislých premenných $X_i$, ktoré môžeme definovať nasledovne:

- `l()` lineárny,
- `s()` splajn,
- `f()` diskrétny faktor,
- `te()` tenzorový súčin,
- `intercept`.

Takto môžeme jednoducho a kompaktne vytvoriť model napríklad `LogisticGAM(s(0, n_splines=200) + te(3,1) + s(2))`, čo znamená, že na premennú $X_0$ chceme použiť splajn pozostávajúci z 200 bázových funkcii, splajn na tenzorový súčin premenných $X_1$ a $X_3$ a splajn na premennú $X_2$.

# 4 Model dennej intenzity klienta

V tejto časti sa budeme venovať aplikáciám prístupov z častí 3.1-3.3 na dátach popísaných v kapitole 2. Postupne pre jednotlivé kategórie dát urobíme niekoľko modelov, ktorých kvalitu predpovedí porovnáme pomocou ROC krivky a AUC (C.1).

## 4.1 Základný demografický model

Keďže budeme modelovať pravdepodobnosti, že klient je v danom časovom okne v priebehu dňa zastihnuteľný, za najdôležitejšiu črtu môžeme považovať čas realizácie hovoru. Túto črtu budeme používať v každom modeli spolu s ďalšími dvoma, ktoré sú vek a pohlavie.

Môžeme označiť $p$ ako pravdepodobnosť, že hovor bude úspešný. Potom z týchto čŕt môžeme zostrojiť model logistickej regresie v tvare

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \sum_{m=1}^{M} \beta_m h_m(X_T) + g(X_T, X_A) + \gamma_1 X_S, \tag{26}$$

kde $\beta_0$ je intercept, $\beta_1, \ldots, \beta_M$, $\theta_{jk}$ a $\gamma_1$ sú koeficienty, $X_T$ predstavuje čas kontaktovania klienta, $h_m$ sú rovnice priestoru kubických splajnov (13), $M$ je počet bázových funkcií, $X_A$ predstavuje vek, ktorý vstupuje do modelu ako tenzorový súčin bázových funkcii premenných $X_T$ a $X_A$, ktorý je definovaný v rovnici (16). $X_S$ predstavuje pohlavie, čo je binárna premenná. V ďalších častiach budeme pre zjednodušenie používať označenie

$$B := \beta_0 + \sum_{m=1}^{M} \beta_m h_m(X_T) + g(X_T, X_A) + \gamma_1 X_S. \tag{27}$$

Premenná čas realizácie hovoru teda nevstupuje do modelu (26) len priamo, ale aj ako interakcia s premennou vek. Tento prístup nám umožňuje modelovať pravdepodobnosť úspešnosti hovoru pomocou novej premennej, ktorá vyjadruje súvislosť medzi časom realizácie hovoru a vekom klienta.

Navyše táto premenná veľmi úzko súvisí s modelom oslovenia, ktorý sme navrhli v kapitole 1, preto má pre nás zmysel sledovať, ako táto premenná interaguje s ostatnými premennými.

## 4.2 Model používajúci denné intenzity nakupovania

Model denných intenzít nakupovania je model, v ktorom okrem základný čŕt použijeme funkcie denných intenzít nakupovania ako funkcionálne premenné. Využijeme pri to poznatky z kapitoly 3.3. K pravej strane modelu (26) pridáme ďalšie členy, ktoré budú zahŕňať vhodne transformované funkcie denných intenzít nakupovania (2.2).

### 4.2.1 Integrácia hustôt

V našom prípade sme hustoty denných intenzít nakupovania uvažovali na časovom intervale od rána 5:00 do večera 23:00. Tento interval sme rozdelili na 56 intervalov dĺžky 20 minút, teda hustoty denných intenzít nakupovania sme rozdelili na 56 oblastí. Na každej takejto oblasti sme spočítali integrál.

Numerický výpočet tohoto integrálu sme skúsili realizovať lichobežníkovou metódou (B.1.1) a Simpsonovou metódou (B.1.2). V tabuľke 3 uvádzame porovnanie rýchlosti týchto dvoch metód, ktoré sme realizovali na vzorke hustôt denných intenzít nakupovania v rôznych kategóriach obchodov pre náhodný výber 10000 klientov.

V ďalších výpočtoch sme uprednostnili lichobežníkovu metódu, nakoľko je výrazne rýchlejšia a hodnoty týchto integrálov nepotrebujeme s veľmi vysokou presnosťou.

| | Hustoty denných intenzít nakupovania | | | |
|---|---|---|---|---|
| | Potraviny a supermarkety | Čerpacie stanice | Lekárne a drogérie | Obchody s oblečením a obuvou |
| Lichobežníková metóda | 44.8 s | 34.2 s | 38.9 s | 29.7 s |
| Simpsonova metóda | 3 min 28 s | 2min 39s | 2min 58s | 2min 20s |

**Tabuľka 3:** *Porovnanie rýchlosti lichobežníkovej a Simpsonovej metódy numerického integrovania. Porovnanie sme robili na denných intenzitách nakupovania pre náhodných 10000 klientov (2.2). V kategórii potraviny a supermarkety je najviac záznamov, a teda aj najviac nenulových hustôt, preto integrovanie v tejto kategórii trvá najdlhšie.*

### 4.2.2 Redukcia dimenzie použitím bázy kubických splajnov

Na modelovanie pravdepodobnosti úspešnosti hovoru sme takto dostali 54 premenných. Najpriamočiarejší spôsob, ako z takto pripravených dát urobiť model je priame použitie logistickej regresie. Takto dostaneme model v tvare

$$\ln\left(\frac{p}{1-p}\right) = \sum_{i=1}^{54} X_i\beta_i, \tag{28}$$

kde $X_i$ sú hodnoty integrálu funkcie dennej intenzity nakupovania v $i$-tom časovom okne. Podľa kapitoly 3.3 môžeme zadefinovať $54 \times M$ maticu $H$, čo je matica bázových funkcií prirozeného kubického splajnu definovaných na sade hodnôt numerického integrálu funkcie dennej intenzity nakupovania. M predstavuje počet týchto bázových funkcií. Sade týchto hodnôt prislúcha definičný obor diskrétnych hodnôt $5, 5\frac{1}{3}, 5\frac{2}{3}, \ldots, 22\frac{2}{3}$, z ktorých rovnomerne vyberieme $M$ uzlových bodov, a teda dostaneme $M$ bázových funkcií.

Potom môžeme urobiť transformáciu $X^* = H^T X$, a následne odhadneme model lineárnej logistickej regresie v novej premennej $X^*$. Touto transformáciou sme teda dosiahli redukciu premenných z pôvodných 54 na $M$.

Keďže tieto premenné charakterizujú klienta v zmysle jeho denného časového harmonogramu, považovali sme za zmysluplné použiť tieto premenné ako interakcie s premenou čas realizácie hovoru. Preto sme funnkcionálne premenné použili v modeli ako tenzorový súčin s premennou čas realizácie hovoru podľa rovnice (16).

Počet bázových funkcií $M^*$ sme zvolili na základe odhadu modelu

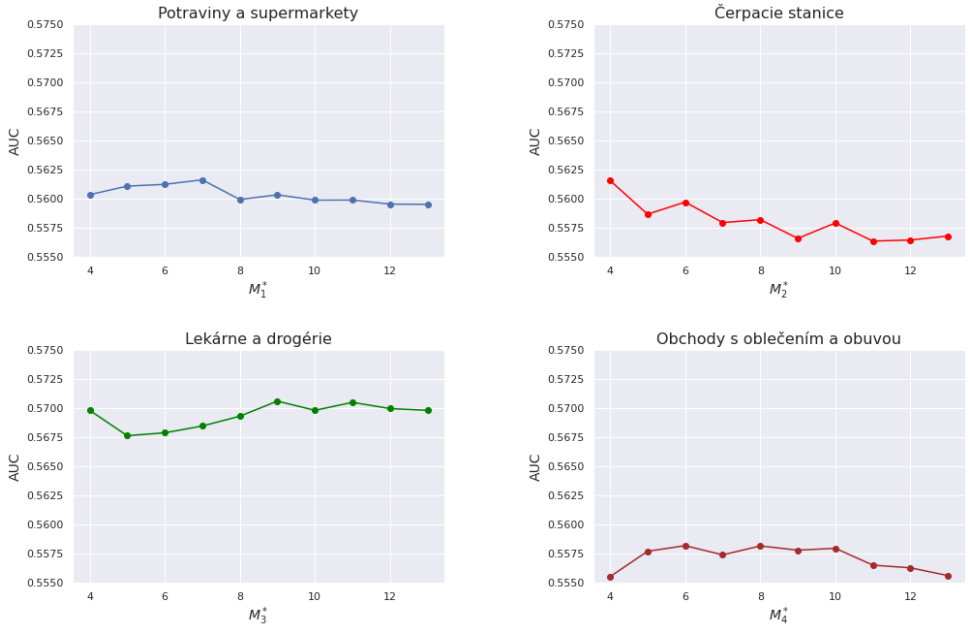
$$\ln\left(\frac{p}{1-p}\right) = \sum_{i=1}^{M} g(X_T, X_i^*) \tag{29}$$

kde sme postupne volili hodnotu $M$ od 4 do 12. Za optimálne $M^*$ sme zvolili také $M$, pri ktorom mal model (29) najvyššiu hodnotu AUC.

Tento postup sme využili pri všetkých štyroch funkcionálnych premenných, ktoré predstavujú denné intenzity nakupovania v štyroch kategóriach obchodov. Výsledky sú znázornené na obrázku 14. Takto sme dostali hodnotu $M^*$ pre všetky kategórie. Výsledné hodnoty sú uvedené aj v tabuľke 4.

Takto sme sa dostali k modelu v tvare

$$\ln\left(\frac{p}{1-p}\right) = B + \sum_{i=1}^{7} g(X_T, X_{1i}^*) + \sum_{i=1}^{4} g(X_T, X_{2i}^*) + \sum_{i=1}^{9} g(X_T, X_{3i}^*) + \sum_{i=1}^{8} g(X_T, X_{4i}^*), \tag{30}$$




**Obr. 14:** *Výsledky modelu (29) v zmysle AUC podľa počtu bázových funkcií $M^*$ pre jednotlivé kategórie. $M$ pre ktoré sme dostali najvyššie AUC sme zobrali ako optimálne. Takto sme zvolili postupne pre všetky kategórie $M_1^*, \ldots, M_4^*$ (tabuľka 4).*

| | $M_1^*$ (Potraviny a supermarkety) | $M_2^*$ (Čerpacie stanice) | $M_3^*$ (Lekárne a drogérie) | $M_4^*$ (Obchody s oblečením a obuvou) |
|---|---|---|---|---|
| Počet bázových funkcii $M_j^*$ | 7 | 4 | 9 | 8 |

**Tabuľka 4:** *Voľba počtu bázových funkcií pre denné intenzity nakupovania pre všetky kategórie obchodov.*

kde $X_{j\cdot}^*$ predstavuje transformované hodnoty funkcií denných intenzít nakupovania postupne pre všetky štyri kategórie. V tomto výslednom modeli nám vyšla hodnota AUC 0.572.

## 4.3 Model používajúci spojité premenné

V tejto časti popíšeme model pozostávajúci z premenných, ktorých popis sa nachádza v kapitole 2.3. V tomto prípade sme odhadli model, ktorý umožňuje uvažovať nelineárnu závislosť od premenných na princípe, ktorý sme opísali v kapitole 3.1. Použitím všetkých 89 premenných spomenutých v kapitole 2.3 dostaneme model v tvare

$$\ln\left(\frac{p}{1-p}\right) = B + \sum_{i=1}^{89}\sum_{j=1}^{M} h_{ij}(x_i)\theta_{ij} \tag{31}$$

kde funkcie $h_{ij}$ sú bázové funkcie kubického splajnu, $\theta_{ij}$ sú koeficienty modelu a $x_i$ predstavuje $i$-tu premennú. Za $M$ sme dosadili hodnotu 12, čo znamená, že sme použili 12 bázových funkcií pre každú premennú.

Nevýhodou tohto modelu je, že takýmto spôsobom nám výrazne narástol počet premenných vstupujúcich do modelu, a teda je potrebné odhadnúť výrazne viac koeficientov, čo sa prejavilo na dĺžke trvania celého výpočtu. V tomto prípade sme dostali hodnotu AUC 0.655, čo je viac, ako v prípade predošlého modelu.

## 4.4 Model využívajúci spojité a funkcionálne premenné

V tejto časti sa pokúsime k modelu (31) pripojiť funkcionálne premenné, t.j. denné intenzity nakupovania, aby sme dosiahli čo najlepší model v zmysle AUC. Tieto premenné sme v modeli použili rovnako ako v modeli (30). Vzhľadom na náročnosť výpočtov sme postupovali tak, že sme modelu (31) postupne pridávali funkcionálne premenné.

V prvom kroku sme pridali k modelu (31) postupne len jednu funkcionálnu premennú, t.j. odhadli sme model

$$\ln\left(\frac{p}{1-p}\right) = B + \sum_{i=1}^{89}\sum_{j=1}^{M} h_{ij}(x_i)\theta_{ij} + \sum_{i=1}^{M_k} g(X_T, X^*_{ki}), \tag{32}$$

postupne pre $k = 1, \ldots, 4$ a $X_{ji}$ sú prvky funkcionálnej premennej podobne ako v modeli (30).

Pre štyri funkcionálne premenné sme odhadli štyri takéto modeli a výsledné hodnoty AUC sme zaznamenali do tabuľky 5.

Ako môžeme vidieť v tabuľke 5, hodnoty AUC sa v závislosti od použitej funkcionálnej premennej líšia len minimálne. Je potrebné si však uvedomiť, že denné intenzity nakupovania v potravinách a supermarketoch sú určené na základe výrazne vyššieho počtu dát

| | $X_{1.}$ (Potraviny a supermarkety) | $X_{2.}$ (Čerpacie stanice) | $X_{3.}$ (Lekárne a drogérie) | $X_{4.}$ (Obchody s oblečením a obuvou) |
|---|---|---|---|---|
| AUC | 0.656 | 0.656 | 0.660 | 0.659 |

**Tabuľka 5:** *Hodnoty AUC pre model (32), ktorý zahŕňa postupne po jednej funkcionálnej premennej, čo sú vlastne denné intenzity nakupovania klientov v jednotlivých kategóriach obchodov.*

(kapitola 2.2, tabuľka 1), preto príslušnú funkcionálnu premennú považujeme za najspoľahlivejšiu.

Nakoľko sme limitovaný operačnou pamäťou, vo finálnom modeli sme navrhli použiť funkcionálne premenné prislúchajúce denným intenzitám nakupovania v potravinách a supermarketoch a lekárňam a drogériam, nakoľko táto funkcionálna premenná sa ukázala ako najlepšia podľa tabuľky 5. Tvar finálneho modelu teda je

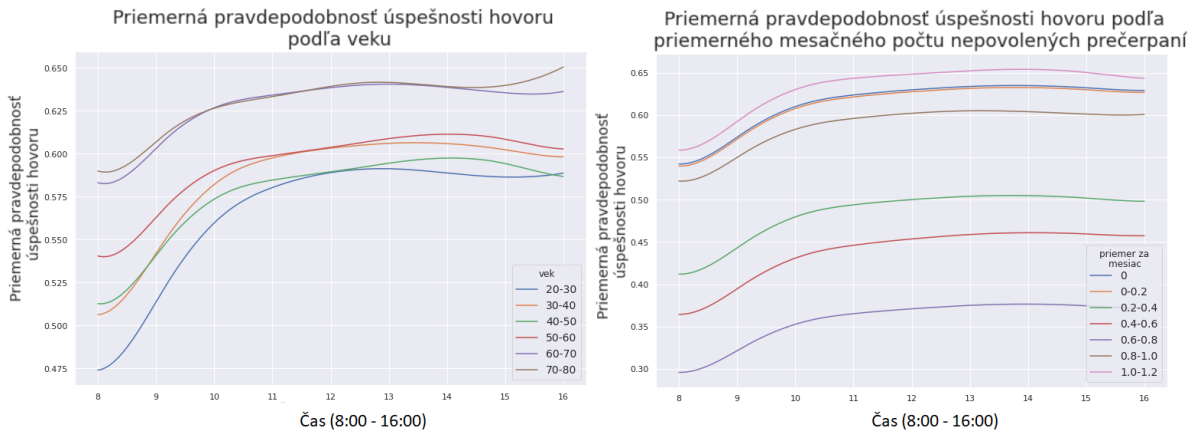
$$\ln\left(\frac{p}{1-p}\right) = B + \sum_{i=1}^{89}\sum_{j=1}^{M} h_{ij}(x_i)\theta_{ij} + \sum_{i=1}^{M_1} g(X_T, X_{1i}^*) + \sum_{i=1}^{M_3} g(X_T, X_{3i}^*), \tag{33}$$

kde $X_{1.}$ a $X_{3.}$ zodpovedajú funkcionálnym premenným, rovnako ako v kapitole 4.2. V tomto prípade sme dostali hodnotu AUC 0.682.

# 5 Výsledky modelu efektívneho oslovenia klientov

V tejto časti sa budeme venovať vyhodnoteniu modelu efektívneho oslovovania klientov v priebehu dňa tak, ako bol navrhnutý v kapitole 1. Využijeme pri tom model logistickej regresie (33) z kapitoly 4.4, ktorý mal pri testovacích dátach najvyššiu hodnotu AUC. Zatiaľ čo sme model (33) trénovali a testovali na sade dát predstavujúcich záznamy hovorov z obdobia 1.2.2021 - 31.12.2021, modely efektívneho oslovovania klientov v priebehu dňa budeme validovať na sade dát za obdobie 1.3.2022 - 31.3.2022.

Najprv sme dopočítali modelové pravdepodobnosti pre záznamy hovorov po jednej minúte v časovom rozpätí od 8:00 do 16:00. Na obrázku 15 sme znázornili, ako sa menia tieto pravdepodobnosti v priemere pre všetkých klientov v priebehu dňa podľa premennej zodpovedajúcej priemernému mesačnému počtu nepovolených prečerpaní a podľa veku. Môžeme si všimnúť, že zatiaľ čo v prípade premennej zodpovedajúcej priemernému mesačnému počtu nepovolených prečerpaní mali tieto krivky rovnaký tvar a líšili sa len o konštantu, tak v prípade veku môžeme pozorovať rozdielne tvary týchto kriviek. Je to spôsobené tým, že premenná vek vstupovala do modelu ako tenzorový súčin s premennou čas realizácie hovoru, teda interakcia týchto dvoch premenných.




**Obr. 15:** *Vlavo priemerná pravdepodobnosť v priebehu dňa podľa vekových kategórií, vpravo priemerná pravdepodobnosť v priebehu dňa podľa priemerného mesačného počtu nepovolených prečerpaní.*

Samotný model efektívneho oslovovania klientov v priebehu dňa sme vyhodnotili na

základe nasledovnej zjednodušenej simulácie. Uvažovali sme situáciu, že pracovník call centra má 100 klientov, ktorých potrebuje osloviť. Pracovný čas od 8:00 do 16:00 si rozdelí na 48 desaťminútových intervalov a v každom takomto časovom okne vyberie na oslovenie klienta, ktorý má najvyššiu pravdepodobnosť zastihnutia. Potom tohoto klienta vyradí a ďalšom priebehu dňa ho už neuvažuje.

V rámci našej simulácie sme vybrali 100 klientov a v zmysle odhadnutého modelu logistickej regresie (33) sme vypočítali pravdepodobnosť zastihnutia po minúte počas celého dňa. Takto sme pre každého klienta dostali odhad funkcie intenzity klienta $\hat{I}$, definovanej v kapitole 1.1.

V kapitole 1.1 sme navrhli dva spôsoby, ako v danom časovom okne vybrať optimálneho klienta. V prvom prípade sme v každom desaťminútovom okne spočítali hodnotu $\hat{I}_i$ podľa vzorca

$$\hat{I}_i = \sum_{j \in \tau} \hat{I}_{ij} = \sum_{j \in \tau} \frac{\int_{t_j}^{t_{j+1}} I_i(s)ds}{t_{j+1} - t_j}, \tag{34}$$

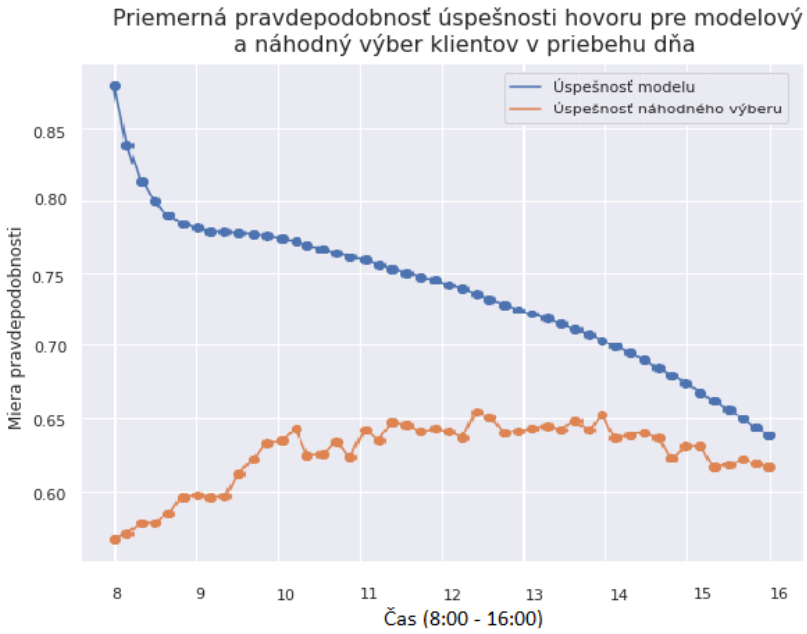
kde $\tau$ je množina indexov spĺňajúca $\tau = \{j | t_{start} \leq t_j \leq t_{end}, \quad j = 0, \ldots, 10\}$ a následne sme vybrali klienta s najvyššou hodnotou.

Alternatívny postup bol v každom desaťminútovom okne zoradiť klientov na základe maximálnej pravdepodobnosti, ktorú v danom časovom okne dosahujú, t.j. pomocou vierohodnostnej funkcie

$$\hat{I}_i^L = \max_{j \in \tau}(\hat{I}_{ij}^L) = \max_{j \in \tau} \left( I_i \left( \frac{t_j + t_{j+1}}{2} \right) \right). \tag{35}$$

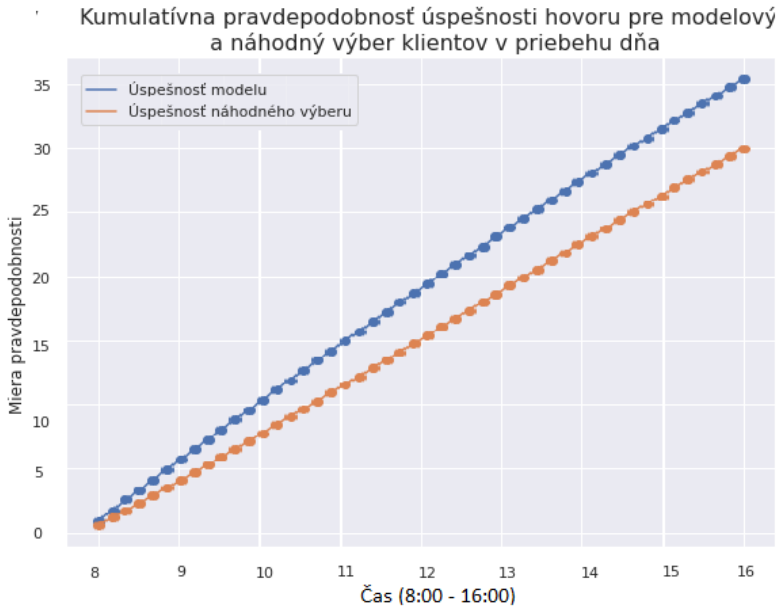
Takto sme dostali v každom časovom okne jedného klienta, ktorého nám určil model ako zastihnuteľného s najvyššou pravdepodobnosťou. Toto sme porovnali s náhodným výberom klientov v jednotlivých časových oknách v zmysle modelovej pravdepodobnosti úspešnosti hovoru. Inými slovami, v každom časovom okne sme vypočítali modelovú pravdepodobnosť úspešnosti hovoru pre klienta vybraného pomocou modelu a náhodne vybraného klienta.

Podľa tohto postupu sme urobili Monte Carlo simuláciu ([5]), kde sme tento postup opakovali 500 krát a porovnali sme priemerné pravdepodobnosti pre klientov vybraných pomocou modelu a náhodne vybraných klientov. Na obrázku 16 sme zobrazili graf priemerných pravdepodobností v priebehu dňa pre modelový a náhodný výber klientov. Vo všeobecnosti tieto metódy boli schopné v každom časovom okne nájsť rovnakých klientov, preto ich z hľadiska výsledkov môžeme považovať za rovnocenné. Na obrázku 17 zase




**Obr. 16:** *Porovnanie priemerných pravdepodobností úspešnosti oslovenia klientov podľa modelu a náhodné oslovenie.*

môžeme vidieť kumulatívne priemerné pravdepodobnosti. Ako tu môžeme vidieť, táto simulácia tu naznačuje, že zatiaľ čo pri náhodnom výbere klientov bol operátor call centra schopný vybaviť v priemere 30.00 klientov, v prípade použitia nášho modelu to bolo 35.63, čo predstavuje zlepšenie o približne 18.77%.




**Obr. 17:** *Porovnanie priemerných kumulatívnych pravdepodobností úspešnosti oslovenia klientov podľa modelu a náhodné oslovenie. V rámci simulácie z náhodných hovorov bolo v priemere úspešných* 30.00*, zatiaľ čo v prípade použitia modelu bolo úspešných v priemere* 35.63*, čo predstavuje zlepšenie úspešnosti o* 18.77%*.*

# Záver

Cieľom práce bolo navrhnúť model, pomocou ktorého by sa pracovník call centra v banke mohol vedieť, ktorý klient je v danom čase najpravdepodobnejšie zastihnuteľný, čo by mohlo zefektívniť jeho prácu. Z pohľadu banky môže tento model pomôcť optimalizovať využitie ľudských zdrojov a z pohľadu klienta minimalizovať nevyžiadané telefonické interakcie.

Náš navrhovaný model využíva lineárny model logistickej regresie, pomocou ktorého modelujeme pravdepodobnosť, že daný klient je v danom časovom okne zastihnuteľný. Využívame pri tom známe poznatky o implementácii funkcionálnych prediktorov (3.3), nelineárnej závislosti prediktorov so závislou premennou (3.1) a interakciám dvoch prediktorov (3.2).

V prvej kapitole uvádzame základnú charakteristiku problému a formálny popis úlohy. Je to naša vlastná formulácia, ktorá zodpovedá technickým možnostiam softvéru používanom v call centre. Navrhujeme tu dva rôzne algoritmy, ako nájsť optimálny čas kontaktovania klienta.

V druhej kapitole sa venujeme popisu dát, ktoré sme využili na trénovanie modelu. Všetky tieto dáta pochádzajú z interných databáz ČSOB. Uvádzame tu podrobnejší popis základných demografických premenných a funkcionálnych premenných týkajúcich sa klienta. Prezentujeme tu vzájomné korelácie prediktorov, ako aj vplyv prediktorov na vzájomnú nepodobnosť správania sa klientov.

V tretej kapitole uvádzame poznatky o modeli logistickej regresie potrebné v nasledujúcich kapitolách. Využívame tu predovšetkým poznatky z [6], kapitola 5.1 – 5.3. Bližšie tu rozoberáme nelineárnu závislosť v lineárnom modeli, implementáciu funkcionálnych premenných v lineárnom modeli, ako aj implementáciu interakcie dvoch prediktorov.

V štvrtej kapitole uvádzame naše výsledky pri trénovaní modelu dennej intenzity na dátach popísaných v kapitole (2). Uvádzame tu postupne sériu modelov, budovanú spôsobom zdola nahor, t.j. postupne pridávame jednotlivé premenné do modelu. Tieto modely porovnávame pomocou ROC krivky a metriky AUC.

V piatej kapitole sa venujeme porovnaniu modelov algoritmov, ktoré môžu byť využité pre voľbu optimálneho okna a ich porovnaniu na simulovaných dátach. Tu sme ako základ použili model logistickej regresie, ktorý dosahoval najlepšie výsledky podľa kapitoly (4).

Uvádzame tu porovnanie priemerných pravdepodobností pre 500 Monte Carlo simulácií. Porovnávame modelové pravdepodobnosti úspešnosti hovoru klientom, ktorých vybral model s náhodným výberom. Výsledky naznačujú, že vďaka nášmu modelu môže operátor call centra dosiahnuť vyššiu úspešnosť pri kontaktovaní klientov o približne 19%.

Tento model bol navrhnutý pre potreby banky nájsť najvhodnejší čas kontaktovania klientov. Možné rozšírenie tejto práce vidíme v modeli, ktorý by umožnil pracovníkom call centra plánovať na celý deň, prípadne týždeň.

# Zoznam použitej literatúry


[1] Atkinson, K.E.: *An introduction to numerical analysis*, John Wiley Sons, Inc., Iowa City, 1978, dostupné na internete (10.03.2022): http://www.math.science.cmu.ac.th/docs/qNA2556/ref_na/Katkinson.pdf

[2] Bentley, J. L.: *Multidimensional binary search trees used for associative searching*, Communications of the ACM 18 (1975), 509–517, dostupné na internete (25.02.2021): http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.160.335&rep=rep1&type=pdf

[3] Dalmasso, N., Pospisil, T., Lee, A. B., Izbicki, R., Freeman, P. E., Malz, A. I.:*Conditional density estimation tools in python and R with applications to photometric redshifts and likelihood-free cosmological inference*, Astronomy and Computing 30 (2020), dostupné na internete (09.03.2021): https://arxiv.org/pdf/1908.11523.pdf

[4] Gooijer, J. G. D., Zerom, D.: *On Conditional Density Estimation*, Statistica Neerlandica 57 (2003), 159-176

[5] Harman, R.: *Stochastické simulačné metódy*, učebné texty, FMFI UK, Bratislava, 2019, dostupné na internete (31.01.2022): http://www.iam.fmph.uniba.sk/ospm/Harman/SSMp.pdf

[6] Hastie, T., Tibshirani, R., Friedman, J. H.: *The Elements of Statistical Learning; Data Mining, Inference, and Prediction*, Springer, 2001, 139-167, dostupné na internete (13.03.2022): https://hastie.su.domains/Papers/ESLII.pdf

[7] Holmes, M. P., Gray, A. G., Isbell, C. L.: *Fast Nonparametric Conditional Density Estimation*, Proceedings of the Conference on Uncertainty in Artificial Intelligence (2007), 175-182, dostupné na internete (20.02.2021): https://arxiv.org/ftp/arxiv/papers/1206/1206.5278.pdf

[8] Izbicki, R., Lee, A. B.: *Nonparametric Conditional Density Estimation in a High-Dimensional Regression Setting*, Journal of Computational and Graphical Statistics 25 (2016), 1297–1316, dostupné na internete (20.02.2021): https://

www.researchgate.net/publication/283337172_Nonparametric_Conditional_Density_Estimation_in_a_High-Dimensional_Regression_Setting

[9] Kullback, S., Leibler, R. A.: *On Information and Sufficiency*, Annals of Mathematical Statistics 22 (1951), s. 79 - 86, dostupné na internete (31.01.2022): https://doi.org/10.1214/aoms/1177729694

[10] Perez-Cruz, F.:*Kullback-Leibler Divergence Estimation of Continuous Distributions*, Proceedings of IEEE International Symposium on Information Theory (2008), s. 1666-1670 dostupné na internete (31.01.2022): http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.422.5121

[11] Servén D., Brummitt C.: *pyGAM: Generalized Additive Models in Python.*, Zenodo, 2018, dostupné na internete (02.04.2022): https://readthedocs.org/projects/pygam/downloads/pdf/latest/

[12] Tauchmanová, M.: *ROC křivka - nástroj pro hodnocení klasifikačních algoritmů*, bakalárska práca, Přírodovědecká fakulta MU, Brno, 2007, dostupné na internete (28.03.2022): https://is.muni.cz/th/d99tt/bakalarka.pdf

# Príloha A

## A.1 Kullback-Leiblerova divergencia

Kullback-Leiblerova divergencia (KL divergencia) je jeden z najznámejších spôsobov, ako merať rozdiel medzi dvoma pravdepodobnostnými rozdeleniami. Túto metódu predstavili jej autori S. Kullback a R. A. Leibler v článku [9]. KL divergencia bola definovaná ako

$$D_{KL}(P\|Q) = \int_X \ln \frac{dP}{dQ} dP, \tag{A.1}$$

kde $P$ a $Q$ sú pravdepodobnostné miery nad množinou $X$ a miera $P$ je absolútne spojitá vzhľadom k miere $Q$. Výraz $\frac{dP}{dQ}$ je Radon–Nikodymova derivácia $P$ vzhľadom k $Q$. V prípade spojitých distribúcií vieme rovnicu (A.1) napísať nasledovne:

$$D_{KL}(P\|Q) = \int_{-\infty}^{\infty} p(x) \ln \frac{p(x)}{q(x)} dx, \tag{A.2}$$

kde $p(x)$ a $q(x)$ sú hustoty pravdepodobností $P$ a $Q$.

V prípade empirických rozdelení nie je KL divergencia jednoznačne definovaná, nakoľko nepoznáme ich hustoty. Jedným zo spôsobov, ako tento problém obísť je odhadnúť neznáme hustoty pomocou jadrového odhadu (*angl. kernel density estimation*). V tomto prípade je však potrebné nastaviť hodnotu parametra *bandwidth*, čo sa môže predstavovať ďalšiu nejednoznačnosť.

V [10], rovnica (4), bol predstavený o niečo priamočiarejší prístup ako odhadnúť KL divergenciu dvoch empirických rozdelení. Tento prístup predpokladá, že $\mathcal{X} = \{x_i\}_{i=1}^n$ sú realizácie z rozdelenia $p(x)$ a $\mathcal{X}' = \{x'_i\}_{i=1}^m$ sú realizácie z rozdelenia $q(x)$. Ďalej predpokladá, že $P(x)$ a $Q(x)$ sú absolútne spojité kumulatívne distribučné funkcie rozdelení $p(x)$ a $q(x)$ a bez ujmy na všeobecnosti realizácie v týchto sadách sú zoradené v rastúcom poradí. Dá sa zadefinovať empirická kumulatívna distribučná funkcia v tvare

$$P_e = \frac{1}{n} \sum_{i=1}^{n} U(x - x_i), \tag{A.3}$$

kde $U$ je definovaná ako

$$U(x) = \begin{cases} 0, & x < 0 \\ \frac{1}{2}, & x = 0 \\ 1, & x > 0. \end{cases} \tag{A.4}$$

Pre funkciu $P_e$ sa dá jednoducho definovať spojitá, po častiach lineárna funkcia

$$P_c(x) = \begin{cases} 0, & x < x_0 \\ a_i x + b_i, & \\ 1, & x > x_{n+1} \end{cases} \tag{A.5}$$

pričom $x_o < \inf\{\mathcal{X}\}$ a $x_{n+1} > \sup\{\mathcal{X}\}$ koeficienty $a_i$ a $b_i$ sú definované tak, aby platilo $P_e(x_i) = P_c(x_i)$, pre všetky $i = 1, \ldots, n$. Presné hodnoty $x_0$ a $x_{n+1}$ nemajú vplyv na tento odhad divergencie.

Samotný odhad KL divergencie má potom tvar

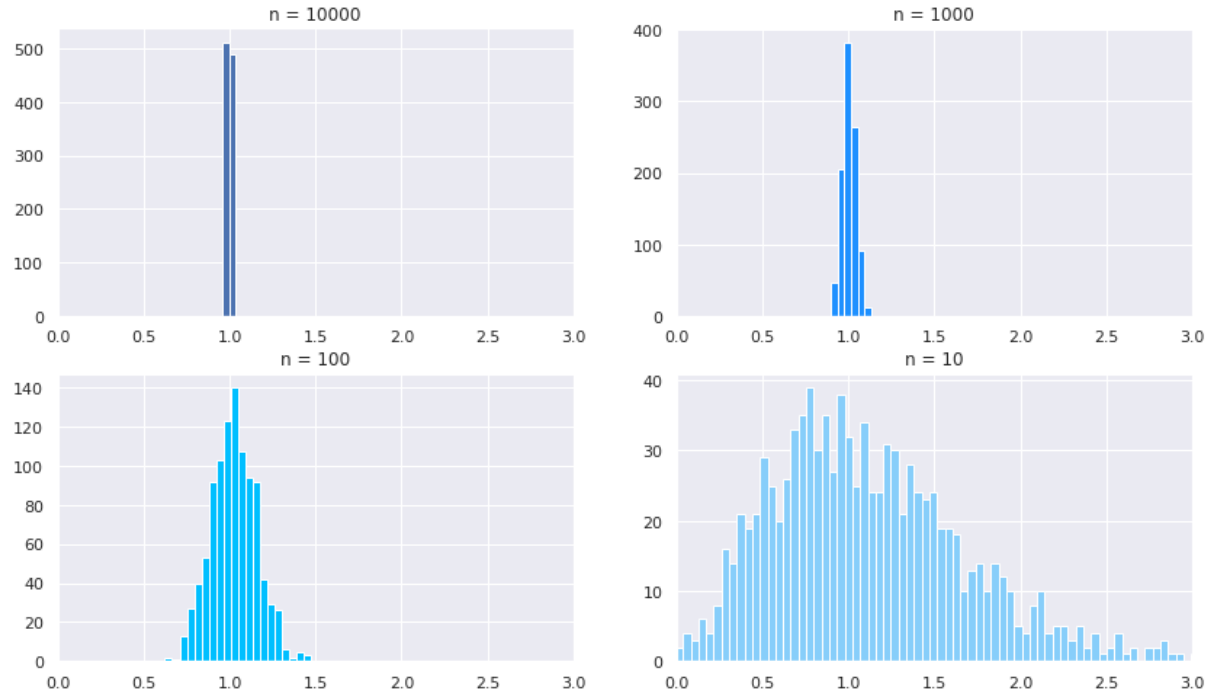
$$\hat{D}_{KL}(P\|Q) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{\delta P_c(x_i)}{\delta Q_c(x_i)} \tag{A.6}$$

kde $\delta P_c(x_i) = P_c(x_i) - P_c(x_i - \epsilon)$ pre nejaké $\epsilon < \min_i\{x_i - x_{i-1}\}$.

V [10] bolo tiež dokázané, že ak teoretické distribúcie P a Q sú absolútne spojité a ich KL divergencia je konečná, tak potom pre sady $\mathcal{X} = \{x_i\}_{i=1}^n$ a $\mathcal{X}' = \{x_i'\}_{i=1}^m$ pochádzajúce z týchto rozdelení platí, že

$$\lim_{n\to\infty} \hat{D}_{KL}(P\|Q) - 1 = D_{KL}(P\|Q). \tag{A.7}$$

Hlavná nevýhoda tohto odhadu je, že s klesajúcim počtom dát výrazne rastie jeho odchýlka. Toto môžeme vidieť aj na jednoduchom simulačnom príklade, kde sme počítali $\hat{D}_{KL}$ pre dve sady dát generované z rovnakého normovaného normálneho rozdelenia $N(0, 1)$ pozostávajúce z $n$ pozorovnaní. Pre každé $n$ sme urobili 1000 simulácii a dostali 1000 hodnôt $\hat{D}_{KL}$. Na obrázku A.18 môžeme vidieť, že aj keď priemerná hodnota $\hat{D}_{KL}$ je vo všetkých prípadoch blízka 1, odchýlka narastá s klesajúcim $n$. Môžeme tu vidieť, že v prípade porovnávania dvoch empirických rozdelení s počtom pozorovaní menej ako 10 bude tento odhad Kullback-Leiblerovej divergencie s vysokou pravdepodobnosťou výrazne vychýlený.




**Obr. A.18:** *Hodnoty $\hat{D}_{KL}$ pre 1000 dvojíc simulovaných sád z normovaného normálneho rozdelenia pozostávajúcich z n pozorovaní. V priemere je hodnota $\hat{D}_{KL}$ blízka 1, avšak s klesajúcim n môžeme pozorovať nárast kolísavosti odhadu $\hat{D}_{KL}$.*

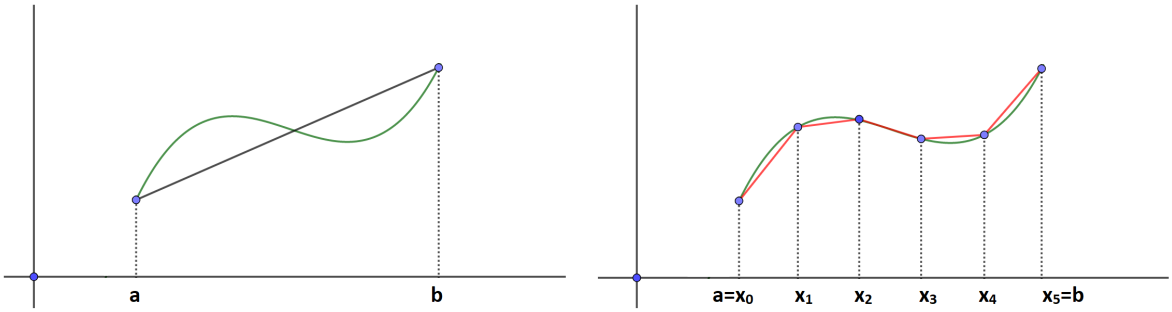
# Príloha B

## B.1 Základné metódy numerického integrovania

### B.1.1 Lichobežníková metóda

Lichobežníková metóda je jedna z najjednoduchších a najstarších metód numerického integrovania ([1]). Myšlienku tejto metódy poznali už starovekom Babylone. Postavená je na tom, že ak máme nejakú spojitú funkciu $f(x)$ definovanú na nejakom intervale $\langle a, b\rangle$ a chceme na tomto intervale spočítať určitý integrál, tak funkciu $f(x)$ aproximujeme lineárnou funkciou $l(x)$, ktorá spĺňa $f(a) = l(a)$ a $f(b) = l(b)$ (obrázok B.19 vľavo). Potom odhad integrálu môžeme vypočítať ako obsah pod grafom lineárnej funkcie, čo je vlastne obsah lichobežníka:

$$\int_a^b f(x) \approx (b-a) \cdot \frac{1}{2} \cdot ((f(a) + f(b)) \,. \tag{B.1}$$




**Obr. B.19:** *Náčrt výpočtu integrálu lichobežníkovou metódou ([1], vlastné spracovanie).*

Pre lepší odhad integrálu môžeme interval $\langle a, b\rangle$ rozdeliť na viac podintervalov použitím $n$ deliacich bodov, ktoré spĺňajú $a = x_0 < x_1 < x_2 < \cdots < x_{n-1} < x_n = b$ ako je znázornené na obrázku B.19 vpravo. Potom sa bude odhad integrálu počítať

$$\int_a^b f(x)dx \approx \sum_{k=1}^{n} \frac{f(x_{k-1}) + f(x_k)}{2} \cdot \Delta x_k \tag{B.2}$$

kde $\Delta x_k = x_k - x_{k-1}$.

### B.1.2 Simpsonova metóda

Simpsonova metóda integrovania predstavuje istú formu vylepšenia lichobežníkovej metódy ([1]). V tomto prípade sa na aproximáciu funkcie $f(x)$ použije kvadratický splajn, t.j. kvadratickú funkciu $p_2(x)$. Nech $c = \frac{a+b}{2}$, potom $p_2(x)$ bude mat tvar

$$p_2(x) = \frac{(x-c)(x-b)}{(a-c)(a-b)} f(a) + \frac{(x-a)(x-b)}{(c-a)(c-b)} f(c) + \frac{(x-a)(x-c)}{(b-a)(b-c)} f(b) \tag{B.3}$$

Potom odhad integrálu funkcie $f(x)$ na intervale $\langle a, b\rangle$ môžeme počítať ako integrál kvadratickej funkcie $p_2(x)$. Náčrt tejto metódy môžeme vidieť na obrázku B.20.

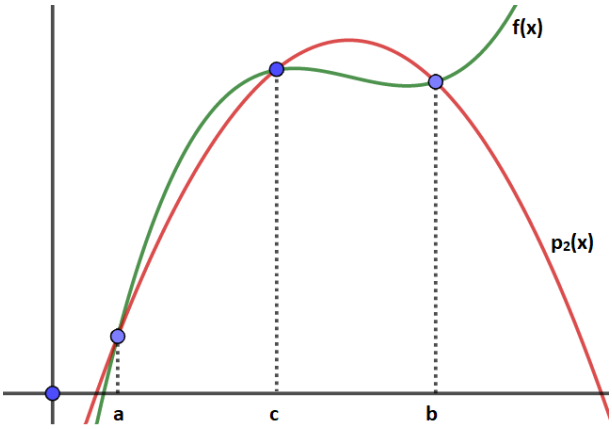
$$\int_a^b f(x) \approx (b-a) \cdot \frac{1}{2} \cdot ((f(a) + f(b)) \,. \tag{B.4}$$

Podľa [1] môžeme tento integrál upraviť do tvaru

$$\int_a^b p_2(x)dx = \frac{h}{3}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right] \tag{B.5}$$

Aby sme dosiahli presnejší odhad $\int_a^b f(x)dx$ môžeme $\langle a, b\rangle$ rozdeliť na $n$ podintervalov. Potom vzorec na výpočet tohto integrálu bude mať tvar

$$\int_a^b f(x)dx \approx \frac{h}{3} \sum_{j=1}^{\frac{n}{2}} [f(x_{2j-2}) + 4f(x_{2j-1}) + f(x_{2j})] = \tag{B.6}$$




**Obr. B.20:** *Náčrt výpočtu integrálu Simpsonovou metódou ([1], vlastné spracovanie).*

$$= \frac{h}{3}\left[f(x_0) + 4\sum_{j=1}^{\frac{n}{2}} f(x_{2j-1}) + 2\sum_{j=1}^{\frac{n}{2}-1} f(x_2 j) + f(x_n)\right] \tag{B.7}$$

kde $x_j = a + jh$ pre $j = 0, 1, \ldots, n-1$, $h = \frac{b-a}{n}$, $x_0 = a$ a $x_n = b$.

# Príloha C

## C.1 ROC krivka a AUC

ROC krivka je známa metóda vyhodnotenia kvality predikcií modelov binárnej klasifikácie. Podrobnejšie odvodenie ROC krivky a AUC môžeme nájsť v [12]. V nasledujúcich riadkoch opíšeme princíp fungovania tejto metriky na porovnanie kvality predpovedí.

Modely strojového učenia vrátane logistickej regresie ako nástroje binárnej klasifikácie nám pre každého klasifikovaného jedinca $i$ vráti nejakú hodnotu $p_i \in \langle 0, 1 \rangle$. Ďalej môžeme uvažovať deliaci bod $c \in \langle 0, 1 \rangle$, ktorý rozdelí klasifikovaných jedincov do dvoch kategórii, t.j. *pozitívni*, $\{i, p_i > c\}$ a *negatívni* $\{i, p_i \leq c\}$.

Na základe tohto rozdelenia môžeme uvažovať dve kategórie:

- True Positive (TP), správne zaradené pozitívne prípady
- False Positive (FP), nesprávne zaradené negatívne prípady.

Pravdepodobnosti zodpovedajúce týmto možnostiam môžeme označiť

- $TP(c)$, pravdepodobnosť, že jedinec patrí do triedy *pozitívni* je pri danom deliacom bode $c$ správne zaradený do triedy *pozitívni*.
- $FP(c)$ pravdepodobnosť, že jedinec patrí do triedy *negatívni* je pri danom deliacom bode $c$ nesprávne zaradený do triedy *pozitívni*.

ROC krivka je potom definovaná ako graf pravdepodobnosti $FP(c)$ na osi $x$ a graf pravdepodobnosti $TP(c)$ na osi $y$. Každý deliaci bod $c$ určuje práve jeden bod $(FP(c), TP(c))$ na ROC krivke. Hodnota AUC predstavuje plochu pod grafom ROC krivky.