Slovenská technická univerzita v Bratislave
Fakulta informatiky a informačných technológií

FIIT-0000-00000

Michal Fašánek

Pokročilé spracovanie sekvenčných dát pomocou umelých neurónových sietí

Diplomová práca

Vedúci práce: Ing. Michal Barla, PhD

Január, 2016

Slovenská technická univerzita v Bratislave
Fakulta informatiky a informačných technológií

FIIT-0000-00000

Michal Fašánek

Pokročilé spracovanie sekvenčných dát pomocou umelých neurónových sietí

Diplomová práca

Študijný program:	Informačné systémy
Študijný odbor:	Informačné systémy
Miesto vypracovania:	Ústav Aplikovanej Informatiky
Vedúci práce:	Ing. Michal Barla, PhD

Január, 2016

>>>> ASSIGNMENT <<<<
>>>> ZADANIE <<<<

Anotácia

Slovenská technická univerzita v Bratislave
FAKULTA INFORMATIKY A INFORMAČNÝCH TECHNOLÓGIÍ

Študijný program:	Informačné systémy
Diplomová práca:	Pokročilé spracovanie sekvenčných dát pomocou umelých neurónových sietí
Autor:	Michal Fašánek
Vedúci práce:	Ing. Michal Barla, PhD

Január, 2016

Analýza dát z používateľského správania k zdrojom na webe je v súčasnosti populárna téma, vzhľadom na svoj potenciál zlepšovať služby poskytované návštevníkom webu. Najnovšie prístupy skúmajú aj možnosti aplikácie metód strojového učenia. V tejto práci sa zameriavam na možnosti využitia rekurentných neurónových sietí s dlhou krátkodobou pamäťou(LSTM) pri analýze dát z platobných brán pre používanie online spravodajských portálov. Takáto analýza poskytuje náhľad do používateľského správania a z toho vyplývajúce možnosti spätnej väzby voči návštevníkom online spravodajských portálov. Manažment biznisu orientovaného na služby si čoraz viac uvedomuje cenu verného zákazníka na trhu. Je preto nutné odhaliť zákazníka uvažujúceho o prechode ku konkurencii pomocou jeho správania a pozitívne motivovať jeho vernosť.

Annotation

Slovak University of Technology Bratislava
FACULTY OF INFORMATICS AND INFORMATION TECHNOLOGIES

Study program:	Informačné systémy
Master thesis:	Advanced processing of sequential data by artificial neural networks
Author:	Michal Fašánek
Supervisor:	Ing. Michal Barla, PhD
2016, January	

Data analysis of user behaviour in accessing web sources has become a popular topic, due to its potential in improvement of services offered to the visitors of web. Most recent approaches examine possibilities of applying methods of machine learning. In this paper, I focus on using reccurent neural networks with long short-term memory(LSTM) to analyze data from paywall of online news portals. Such analysis provides insight into the user behaviour and resulting feedback possibilities to the users of online news portals. Management of service-oriented business is starting to realize the value of loyal customer on the market. It is necessary to detect customer evaluating churning and use positive motivation to keep his loyalty.

Obsah

Kapitola 1

Úvod

Ľudský mozog je zložitý a ešte v dnešných dobách z veľkej časti nepochopený orgán. Od prvého okamihu je trénovaný riešiť problémy, ktorých formálna špecifikácia presahuje naše možnosti. Snahou neurónových sietí je napodobniť takéto schopnosti a možnosti simuláciou architektúry mozgu.
Jednou z možných aplikácií takýchto schopností je prenikanie do mysle zákazníka na trhu služieb. Online biznis má prostriedky, ktoré mu dovoľujú zhromaždiť obrovské množstvo dát o aktivite svojich zákazníkov. Na získanie pridanej hodnoty z týchto dát je však nutné identifikovať a pochopiť vzory, ktoré sa v týchto dátach nachádzajú. Takáto úloha je pre neurónové siete adekvátnou výzvou.

1.1 Použité skratky

- **NN, ANN** - Neurónová sieť (Neural Network, Artificial Neural Network)
- **FNN** - Dopredná neurónová sieť (Feedforward Neural Network)
- **RNN** - Rekurentná neurónová sieť (Recurrent Neural Network)
- **LSTM** -Dlhá krátkodobá pamäť (Long Short-Term memory)
- **CRM** - Manažment vzťahov so zákazníkmi (Customer Relationship Management)
- **VAW** - Web s pridanou hodnotou (Value Added Web)
- **TINA** - Informačná sieťová architektúra v telekomunikáciach (Telecommunication Information Network Architecture)
- **IP** - Internetový protokol (Internet Protocol)
- **CEC** - Konštantné cyklenie chyby (Constant Error Carousel)
- **AUC** - Oblasť pod krivkou (Area under curve)

Kapitola 2

Analýza

V tejto časti sa venujeme dôkladnej analýze podkladov. Jednotlivé časti sú popísané v rozsahu relevantnom pre túto prácu. Analýza je štrukturovaná na nasledovné časti:

- Problémová oblasť
- Dáta sprístupnené pre prácu
- Neurónové siete
- Výskum v danej oblasti

2.1 Problémová oblasť

V tejto práci sa zameriavame na predikciu úbytku zákazníkov(churn rate) pri predplatiteľských službách. V súčasnej dobe sa do popredia biznis prístupov stále viac dostávajú prístupy riadenia vzťahov zo zákazníkmi(CRM). Ukazuje sa totiž, že na trhu s dostatočným pokrytím poskytovateľov cieľovej služby je niekoľkonásobne drahšie získať nového ako udržať si existujúceho zákazníka. Tento prístup však vyžaduje rozsiahlu znalosť dostupnej zákazníckej základne, ktorou poskytovateľ disponuje [16].

2.1.1 Predikcia úbytku zákazníkov

Predikcia úbytku zákazníkov sa venuje spracovaniu dostupných dát o zákazníckej aktivite, službách ktoré využívajú a vývoja ich správania v čase. Výsledkom analýzy je štatistika poskytujúca informácie o jednotlivých zákazníkoch a ich šanci na presun k inému poskytovateľovi. Z týchto dát je následne odvoditeľné, aké percento zákazníkov odíde ku konkurencii a aký to bude mať dopad na finančné príjmy od ktorých je poskytovateľ závislý.

2.1.2 Moderovanie úbytku zákazníkov

Vo vzťahu k úbytku zákazníkov definuje CRM dva základné prístupy, ktorými je možné moderovať úbytok [3].

2.1.2.1 Reaktívny prístup

Motivácia zákazníka pre zotrvanie s pôvodným poskytovateľom služby nastáva, až keď sa zákazník explicitne rozhodne pre prechod ku konkurenčnému poskytovateľovi. V tomto okamihu začína poskytovateľ na svojho zákazníka apelovať výhodnými ponukami, zľavami alebo inými spôsobmi motivácie pre zotrvanie u poskytovateľa. Takýto prístup sa ukazuje ako ľahko zneužiteľný ostatnými zákazníkmi, ktorí by inak nemali motiváciu pre prechod ku konkurencii. Predikcia úbytku zákazníkov v tomto prístupe nemá nijakú významnú úlohu.

2.1.2.2 Proaktívny prístup

Pri úspešnej predikcii záujmu zákazníka o prechod ku konkurenčnému poskytovateľovi je možné efektívne jeho zámer smerovať pozitívnou motiváciou. Tento prístup však predpokladá vysokú úspešnosť predikčných metód. Pri nesprávnej identifikácii zákazníckeho správania je totiž nielen možné nezabrániť zákazníkom v presune ku konkurenčnému modelu, ale aj investícii finančných

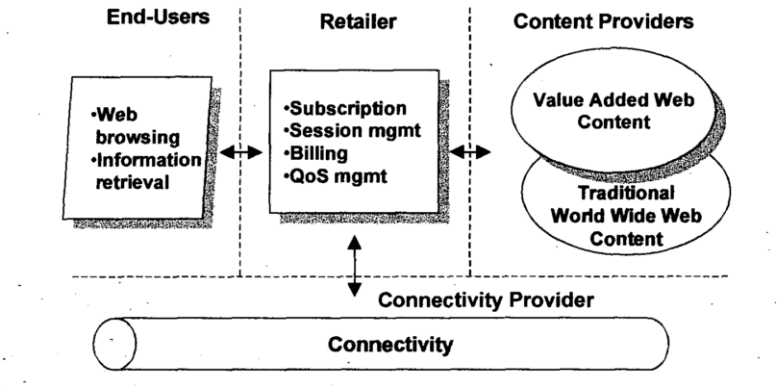
prostriedkov do skupiny zákazníkov, ktorá by naďalej generovala zisk aj bez významnejšej motivácie, resp. nevrátila by rozdielom v úbytku motivačné náklady, ktoré na ňu daný poskytovateľ vynaložil.

2.2 Dáta sprístupnené pre prácu

Pre túto prácu boli sprístupnené dáta z platobnej brány portálu pre online spravodajské denníky. Platobný portál poskytuje platformu pre periodiká, ktoré majú záujem o online funkcionalitu ale nemajú záujem implementovať vlastný platobný systém. Zákazníci tohto portálu tak získavajú rýchle riešenie pre možnosť vyhradenia exkluzívneho obsahu zo svojich online materiálov.

2.2.1 Exkluzívny obsah

Exkluzívny obsah je nástroj, ktorý množstvo poskytovateľov služieb využíva pri prechode na web. Umožňuje prístup k väčšiemu počtu potenciálnych zákazníkov, pričom poskytovateľovi ostáva možnosť oddeliť, čo bude prístupné každému od exkluzívneho obsahu určeného pre predplatiteľov.
Realizáciu exkluzívneho obsahu pomocou platobnej brány tretej strany umožňuje špecifikácia VAW(value added web). VAW aplikuje TINA(Telecommunications Information Networking Architecture) biznis model do klasického WWW(world wide web) prostredia. Určuje tak vzťahy medzi jednotlivými právnymi subjektami podľa obr. 2.1. Poskytovateľ služieb(spravodajské periodikum) tak môže poskytovať nielen klasický ale aj exkluzívny obsah bez toho, aby sa vo väčšej miere muselo zaoberať správou poskytovaných služieb a finančnou administratívou. Za tú zodpovedá sprostredkovateľ(platobný portál), ktorého úloha spočíva v správe exkluzívneho obsahu vo vzťahu ku koncovému používateľovi [18].



Obr. 2.1: *Základná schéma VAW*

[18].

2.2.2 Získavanie dát

Pri pokuse o prístup k exkluzívnemu obsahu stojí medzi používateľom a obsahom platobná brána portálu. Používateľovi bez predplatenej služby je zobrazená ponuka na platený prístup. Predplatitel prechádza cez bránu a je mu sprístupnený exkluzívny obsah. Pri všetkých aktivitách na portáli sú zaznamenávané používateľské údaje. Dostupné údaje sú vo forme záznamov - textových súborov priebežne generovaných používateľskou činnosťou. Bežná činnosť pri analýze záznamov z činnosti a práci s velkými dátami všeobecne je predspracovanie dát. Pri sledovaní činnosti používateľov sa generujú súbory so stovkami miliónov až miliardami záznamov. V súčasnosti nie je možné klasickými prístupmi spracovať takéto objemy dát bez predspracovania - filtrovania, segmentácie a čistenia dát. Spôsob predspracovania dát je z podstatnej časti ovplyvnený metódami, ktorými chceme dáta spracovať. Pri práci so záznamami je bežné deliť dáta na tzv. používateľské prístupy (user sessions). Používateľský prístup modeluje aktivitu - jeden prístup jedného používateľa. Takto rozdelené záznamy poskytujú elasticitu pri spracovaní podľa špecifického času alebo podľa používatelov.

Medzi najdôležitejšie dostupné údaje z platobného portálu patria:

- IP adresa
- Používateľský účet
- Časový rozsah prístupu
- Prehliadaný obsah
- Aktivácia/prerušenie predplatného

2.3 Neurónové siete

Koncept neurónových sietí vznikol v 40. rokoch minulého storočia inšpiráciou biologickými neurónovými sieťami v mozgu [14]. Cieľom bolo prekonať bariéru medzi tým, čo je pre ľudský mozog ľahko riešitelné ale ťažko formálne definovatelné matematickými pravidlami. Tieto problémy, ktoré riešime intuitívne, pri pokuse o formálnu špecifikáciu ukazujú, aké množstvo znalostí používame v každodennom živote. Ako vhodný príklad slúži vizuálne rozoznávanie objektov, ktoré je pre osobu samozrejmé, no až v posledných rokoch zaznamenávame prvé úspechy v tejto problematike pri použití NN [10].

2.3.1 Štruktúra

Podobne ako v mozgu, základ neurónovej siete tvoria neuróny a prepojenia medzi nimi. Neuróny sú organizované vo vrstvách, ktoré sa delia na 3 základné typy.

Vstupná vrstva - reprezentuje dáta, ktoré podsúvame sieti pre interpretáciu. Dáta musia byť pred posunutím vstupnej vrstve často predspracované, aby bola sieť schopná interpretovať ich. Počet neurónov na vstupnej vrstve je ovplyvnený množstvom dát, ktoré máme na vstupe. V sieti existuje iba jediná vstupná vrstva.

Výstupná vrstva - interpretácia dát sieťou. Výstupnú vrstvu je možné nazvať „výsledok"siete.

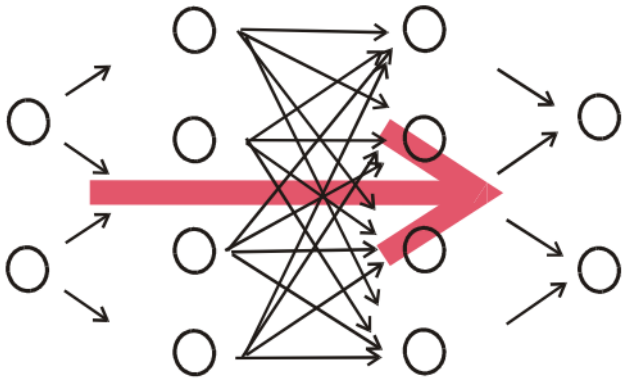
Skrytá vrstva - nachádzajú sa medzi vstupnou a výstupnou vrstvou. Ich počet určuje hĺbku siete. NN nemusí mať ani jednu skrytú vrstvu, no takáto sieť dokáže modelovať iba lineárnu závislosť. Všeobecne platí, že čím viac skrytých vrstiev má sieť, tým zložitejšie vzťahy dokáže simulovať. Zvyšujú sa však aj nároky na učenie a výpočtové nároky. Jediná skrytá vrstva vytvára pozoruhodný rozdiel v aplikovatelnosti modelu, keďže prekonáva hranicu lineárnej závislosti funkcie, ktorú model pokrýva. Pri vysokej zložitosti modelu je možné naraziť na problém preučenia, ktorý bráni sieti korektne generalizovať. Neexistuje nijaký spoľahlivá metóda pre správny počet alebo veľkosť skrytých vrstiev. Empiricky sa vyvinulo niekoľko odhadov, ale v praxi je nutné overovať správnosť modelu praktickou evaluáciou. Odhadové pravidlá najčastejšie padajú na neschopnosti integrovať vo svojom rozhodnutí komplexitu úlohy a redundanciu v tréningových dátach [10].

Prepojenia - Váhované prepojenia medzi neurónmi fungujú ako pamäť neurónovej siete. V jednoduchom modeli neurónovej siete sú prepojenia iba medzi neurónmi navzájom susediacich vrstiev. Prepojenie existuje medzi každým neurónom *n*-tej do *n+1* vrstvy. Neuróny jednej vrstvy pritom medzi sebou nie sú prepojené. Signál sa šíri týmito prepojeniami od vstupnej vrstvy smerom k výstupnej vrstve v jednom smere, ako je to ilustrované na obr. 2.3. Takéto siete sa volajú *dopredné*. Hlavný účel prepojenia je niesť váhu. Váha prepojenia určuje, aký významný je vzťah medzi dvomi danými neurónmi, ktoré spája. Korektná váha daného prepojenia je na začiatku neznáma, jej korektné nastavenie je výsledkom procesu učenia [10].

Neurón - predstavuje základnú stavebnú jednotku neurónovej siete. Skladá sa z *aktivačnej funkcie* a *prahovej hodnoty*. Prahová hodnota neurónu ϑ_i^{k+1} je odpočítaná od sumy vstupných váhovaných hodnôt $w_{ij}^k \,.\, o_j^k$.
Na výsledok o_i^{k+1} sa následne aplikuje aktivačná funkcia f podľa obr. 2.2. Takýto výstup je následne prepojeniami posielaný do ďalších neurónov. Špeciálny prípad je neurón vstupnej a výstupnej vrstvy. Na vstupe totiž neurón hodnotu iba posiela ďalej a na výstupe po spracovaní nie je zasielaná nikam - predstavuje výsledok siete.

$$o_i^{k+1} = f\left(\sum_{j=1}^{N} w_{ij}^k \,.\, o_j^k \;-\; \vartheta_i^{k+1}\right)$$

Obr. 2.2: *Výstupná hodnota neurónu [13]*



Obr. 2.3: *Štruktúra doprednej neurónovej siete (FNN) [11]*

2.3.2 Učenie neurónovej siete

Učenie predstavuje kľúčovú aktivitou pre schopnosť siete produkovať požadované výsledky. Spočíva vo vystavovaní neurónovej siete tréningovým dátam, ktoré sa sieť snaží interpretovať.

Učenie s učiteľom je metóda, pri ktorej je dostupná sada tréningových dát „označená". Pri interpretovaní výsledku je možné okamžite určiť, aká chyba nastala a následne ju propagovať do siete. Na toto sa využíva tzv. *spätná propagácia*(backpropagation), ktorá upravuje váhy siete v rozsahu chyby, ktorá nastala - rozdiel medzi správnym výsledkom pre daný vstup a samotným výsledkom siete.

Učenie bez učiteľa predstavuje alternatívnu metódu, pri ktorej tréningové dáta nemajú dostupné výsledky. Neurónová sieť sa sama učí rozhodnúť, čo je

pre ňu relevantné. Učenie bez učiteľa predstavuje možnosť ako získať takmer neobmedzené množstvá tréningových dát tam, kde učenie s učiteľom vyžaduje manuálne a kvôli časovej náročnosti nedostupné označovanie.

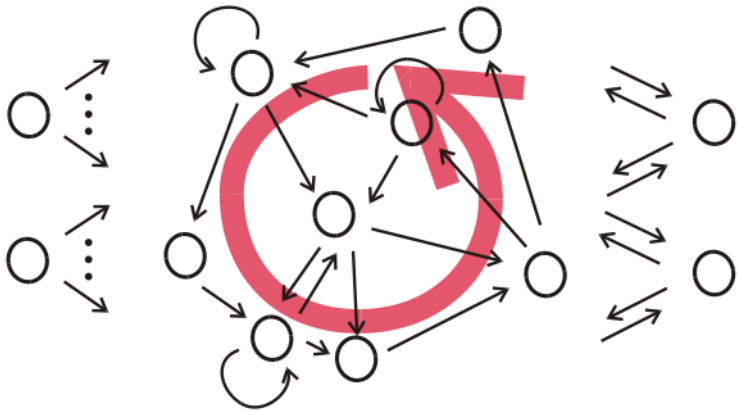
2.3.3 Hyperparametre

Nastavenia, pomocou ktorých kontrolujeme správanie neurónových sietí sa nazývajú *hyperparametre*. Tieto hodnoty nie sú získané učením siete pokiaľ nemodelujeme vnorený systém za týmto účelom. Príkladom hyperparametra je počet skrytých vrstiev NN. Pri nízkom počte nebude model schopný naučiť sa funkciu definovanú problémom. Pri vysokom počte je možné, že sieť v sebe uloží menší tréningový dataset, nazývané tiež ako problém *preučenia*(overfitting). Pri preučení sieť nezíska schopnosť generalizácie problému kvôli sledovaniu tréningového datasetu. Je zjavné, že zvolenie správnych hyperparametrov má pre výsledky metódy kľúčovú úlohu [10]. Medzi ďalšie významné hyperparametre patria:

- Šírka jednotlivých vrstiev
- Rýchlosť učenia
- Momentum
- Aktivačné funkcie neurónov

2.3.4 Rekurentné neurónové siete

Do popredia výskumu sa v súčasnosti dostávajú pokročilé modely, ktoré už nie sú obmedzené na jednoduchý dopredný prístup. Vďaka rapídnemu zvyšovaniu výkonu grafických kariet sa čoraz častejšie aplikujú *rekurentné modely neurónových sietí*(RNN) [11]. Špecializáciou rekurentných sietí je práca so sekvenčnými dátami. Tieto siete predstavujú generalizáciu dopredných modelov ich rozšírením o cyklické prepojenia [10]. Takýmto spôsobom je možné využiť súčasnú hodnotu premennej na ovplyvnenie vlastnej hodnoty v budúcnosti. Cyklický charakter rekurentného modelu je zobrazený na obr. 2.4.



Obr. 2.4: *Štruktúra rekurentnej neurónovej siete [11]*

2.3.5 Siete s dlhou krátkodobou pamäťou - LSTM

LSTM predstavuje vylepšený model RNN. Vnútorná štruktúra ako doplnok ku externej rekurencii medzi jednotlivými neurónmi obsahuje aj *internú rekurenciu*, zobrazenú v štruktúre LSTM neurónu na obr. 2.5. Medzi najdôležitejšie súčasti tohto modelu patria sigmoidné brány, ktoré rozhodujú o tom, ako sa signál bude širiť. LSTM tak prekonáva problém strácajúceho sa gradientu, ktorým trpí klasická RNN architektúra [9].

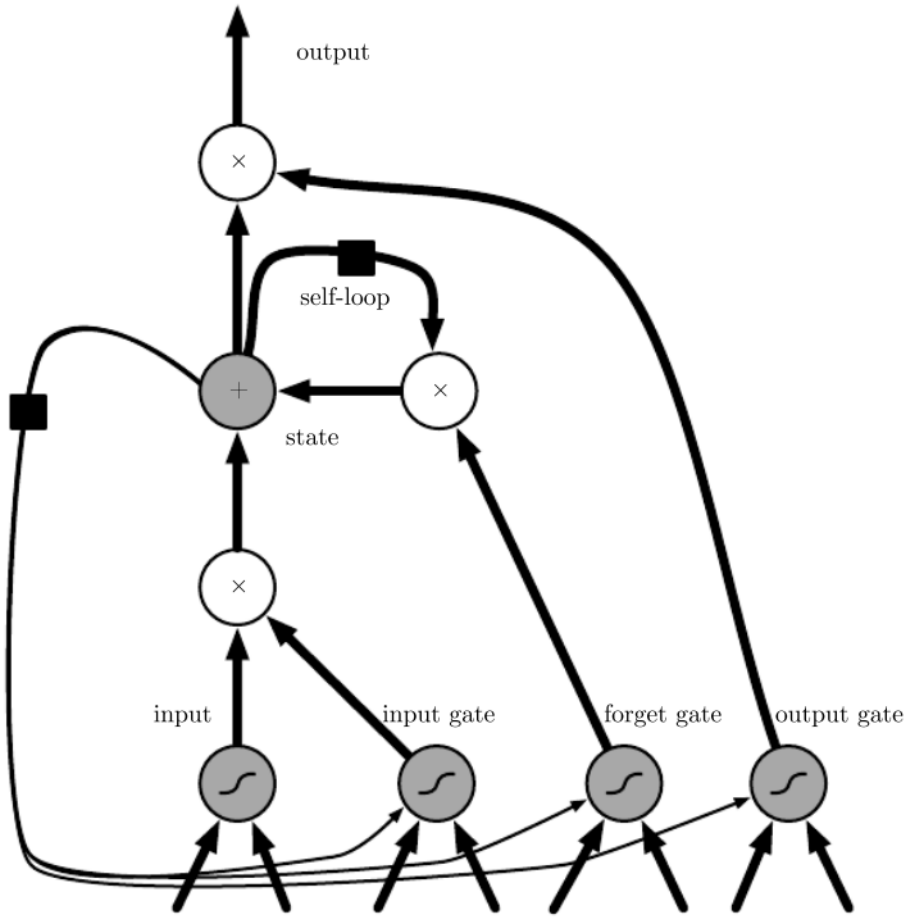
Brána zabudnutia ovplyvňuje, či nastáva vnútorná rekurencia neurónu. Stav tak môže ale nemusí byť faktorom ovplyvňujúcim nasledujúcu iteráciu výpočtu v sieti. Významné zlepšenie v LSTM sieťach prišlo s myšlienkou *kontextom podmieneného zabúdania*. Takýto model sa ukazuje extrémne výhodným pri riešení problémov zahŕňajúcich *časové pauzy*(lags) [4]. Dôležitý prvok na obr. 2.5 predstavuje čierna kocka. Označuje pauzu o veľkosti jednej iterácie. Hodnota signálu tak ovplyvňuje nasledujúcu iteráciu, tj. vplýva na neskoršie udalosti.

Nazeracie diery(peepholes) predstavujú vylepšenie LSTM. Rieši problémy, ktoré vznikajú na základe faktu, že brána nedostáva priame informácie o

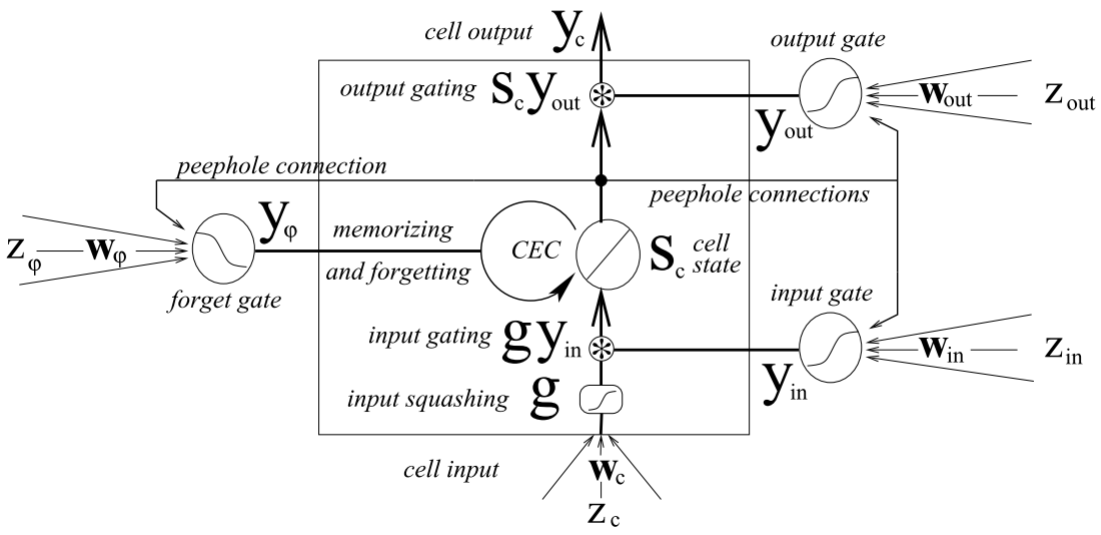
stave jadra LSTM bloku(CEC). Táto situácia nastáva, keď je výstupná brána zatvorená. *Nazeranie* predstavuje techniku váhovaného prepojenia CEC s bránami bloku daného jadra. Prepojenia sú štandardné s výnimkou časovej pauzy. Schéma nazerania v LSTM bloku je zobrazená na obr. 2.6.

LSTM siete v praxi dokázali svoje schopnosti pri aplikácií na rôzne netriviálne dátové problémy. Pozornosť je kladená na frekventovanú časovú závislosť v dátach:

- Rozoznávanie rukopisu [7]
- Rozoznávanie reči [6]
- Označovanie obrázkov [12]



Obr. 2.5: *Štruktúra LSTM bunky [5]*



Obr. 2.6: *Schéma nazerania v LSTM bloku [5]*

2.4 Výskum v danej oblasti

V tejto časti sa zaoberáme štúdiou dostupných riešení pre problém úbytku zákazníkov. Zameriavame sa na metódy strojového učenia. Analýza poskytuje náhľad do konkrétneho biznis odvetvia, v ktorom bol skúmaný úbytok zákazníkov pre lepšie pochopenie stratégie, s ktorou bolo spracovanie dát a použité metódy optimalizované.

2.4.1 Markovove reťazce, náhodné lesy

Štúdia aplikuje štatistické modely, medzi nimi najúspešnejšie *Markovove reťazce* a *náhodné lesy* pri predikcii úbytku zákazníkov spoločnosti poskytujúcej káblovú televíziu. Skúmaná spoločnosť v minulosti rýchlo získala veľký podiel zákazníkov na trhu, o ktorý začala stabilne počas rokov prichádzať. Táto spoločnosť za použitia metód na predpovedanie úbytku zákazníkov dokázala takmer zdvojnásobiť svoje zisky zameriavaním sa na pozitívnu motiváciu

najrizikovejších zákazníkov [3].
Spoločnosť ponúka stabilné ročné predplatné bez možnosti prerušenia zmluvy. Nahlasovanie plánovaného nepredlžovania zmluvy je povinné v poslednom mesiaci zmluvy, pričom neohlásenie automaticky predlžuje zmluvu na ďalší rok. V štúdii boli sledovaní zákazníci, ktorí mali aktívnu zmluvu vo *vzorkovacom dátume*(28. 2. 2002) a neboli vylúčení kvôli neplateniu predplatného.

2.4.1.1 Dataset

Dataset zákazníkov obsahuje nasledovné informácie o zákazníkoch:

- Zmluvné - počet mesiacov trvania predplatného, mesiac ukončenia, typ produktu, špeciálne balíčky(šport, filmové, ...), spôsob platby
- Socio-demografické - vek, pohlavie, región, biznis
- Finančné - upomienky, typ upomienok, čas od poslednej upomienky
- Historické - počet obnov predplatného, získané zľavy

Evidovaní zákazníci, ktorí opustili spoločnosť tvoria 15% z datasetu. Pri využití metód bola vyhradená časť dát(60%) pre kalibráciu a časť(40%) pre testovanie úspešnosti. Stratifikácia bola aplikovaná kvôli udržaniu pomeru 15% odchádzajúcich zákazníkov pre obe časti [15].

2.4.1.2 Markovove reťazce

Markovove reťazce sú pravdepodobnostná technika pre reprezentáciu korelácie medzi za sebou idúcimi pozorovaniami. Táto štúdia poukazuje na vplyv sekvencie v odoberanom type produktu na predpoveď, viď. tabuľku 2.7. Kvôli tomuto javu boli využité Markovove reťazce.

	$t-10$	$t-9$	$t-8$	$t-7$	$t-6$	$t-5$	$t-4$	$t-3$	$t-2$	$t-1$
Customer 1	A	A	B	B	B	B	B	B	B	B
Customer 2	A	A	A	A	A	A	A	A	A	B

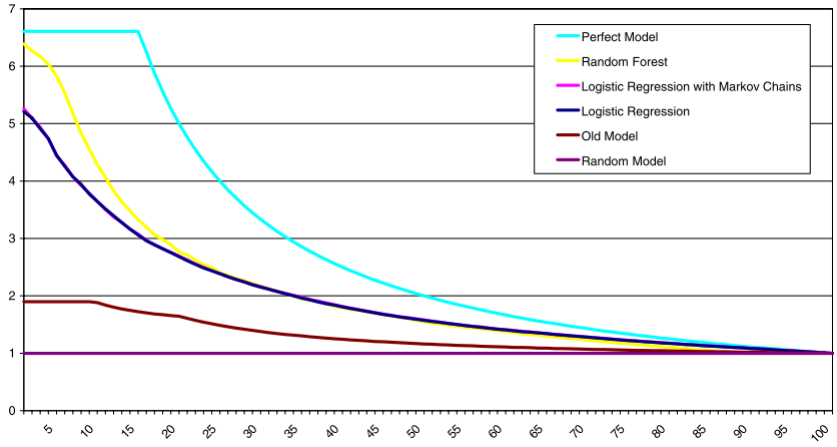
Obr. 2.7: *Sekvencia typu odoberanej služby dvoch zákazníkov [3]*

2.4.1.3 Náhodné lesy

Kvôli svojej jednoduchosti použitia a interpretácie a schopnosti práce s ukazateľmi na rôznych úrovniach merania sa stali rozhodovacie stromy populárnou metódou predikcie. Ich nevýhody(ako napr. nedostatok robustnosti) úspešne rieši vytváranie veľkého počtu stromov so samostatným hlasovaním - lesov. Tento experiment využil náhodné lesy podľa štúdie L. Breimana [2].

2.4.1.4 Evaluácia

Štúdia pri evaluácii využíva počítanie zdvihu - pomeru zákazníkov predikovaných ako náchylných k prechodu a z nich zákazníkov, ktorí prešli inam, relatívne k percentu všetkých ušlých zákazníkov. Vysoký zdvih teda indikuje úspešný model. Pri 15,13% úbytku zákazníkov teda perfektná predpoveď predstavuje $100/15,13 = 6,61$ zdvih. Na grafe 2.8 je možné vidieť úspešnosť jednotlivých metód pri vybratí daného percenta najohrozenejších zákazníkov. Vidno tu úspešnosť náhodných lesov aj takmer nulové zlepšenie logistickej regresie Markovovými reťazcami.



Obr. 2.8: *Výsledky použitých metód pre dané % najohrozenejších zákazníkov [3]*

2.4.2 Neurónová sieť

Neurónová sieť bola využitá pri štúdii zaoberajúcej sa predikciou úbytku zákazníkov u mobilného operátora. Zdanlivo jednoduchý model využívajúci učenie s učiteľom priniesol prekvapivo dobré výsledky pre verejne dostupný dataset [16].

2.4.2.1 Dataset

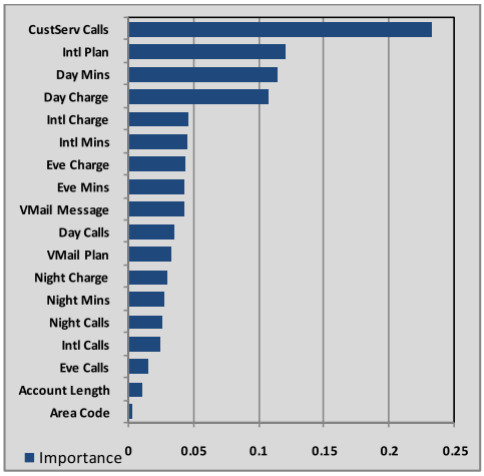
Záznamy 20 rôznych premenných od 2427 zákazníkov obsahujú okrem samotnej informácie o úbytku nasledovné informácie:
štát, doba aktívnosti účtu, kód oblasti, telefónne číslo, medzinárodný plán (áno/nie), služba hlasového záznamu, počet hlasových záznamov, prevolané minúty (deň/večer/noc), počet hovorov(deň/večer/noc), (deň/večer/noc) platba, medzinárodné služby, počet volaní na zákaznícku linku.

2.4.2.2 Neurónová sieť

Táto štúdia pracuje s klasickým modelom FNN pri predikcii. Ako prevencia problému preučenia funguje vyberanie náhodných záznamov do tréningového datasetu. Zbytok dát je využitých pre evaluáciu schopností siete predikovať úbytok.

Výsledná úspešnosť siete dosahuje 92,35%. Architektúra tejto siete využíva jeden neurón pre každý vstup na prvej vrstve. Informácie o čísle a štáte neboli zahrnuté, lebo plnili iba identifikačnú funkciu. Po rozsiahlejších experimentoch so skrytými vrstvami sa ako najlepšie riešenie ukazuje využitie jedinej skrytej vrstvy s 3 neurónmi. Na výstupe neurónová sieť poskytuje informáciu o prechode(áno/nie) ale aj istotu, s ktorou tento výsledok určila.
Na obr. 2.9 je tiež vidno, ako boli vyhodnotené jednotlivé vstupné parametre z hľadiska dôležitosti. Tá je určená na intervale $< 0; 1 >$, pričom však zriedka prekročí $0,35$. Ako sa ukázalo, najdôležitejším indikátorom prechodu zákazníka je počet volaní na zákaznícku podporu a množstvo medzinárodných

služieb.

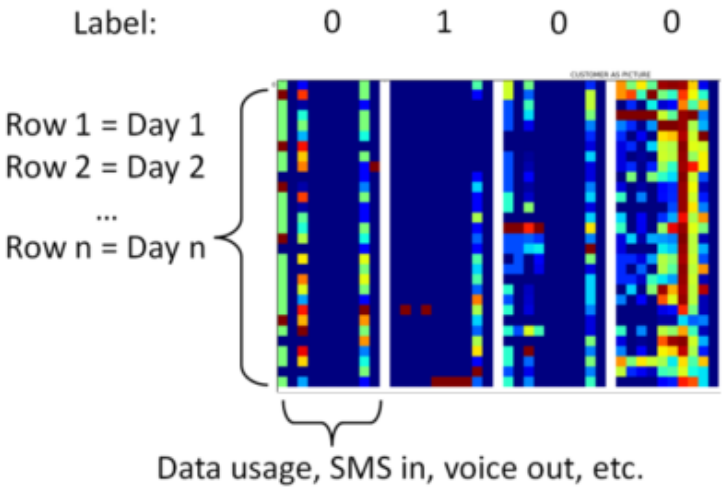


Obr. 2.9: *Vplyv parametrov na stratu zákazníka [16]*

Tento model správne predikuje až 97% ostávajúcich zákazníkov. Správne však určí iba 66% strácaných zákazníkov. Evaluácia tohto výsledku je teda ťažko interpretovatelná hodnotou 92%.

2.4.2.3 Hlboká konvolučná neurónová sieť

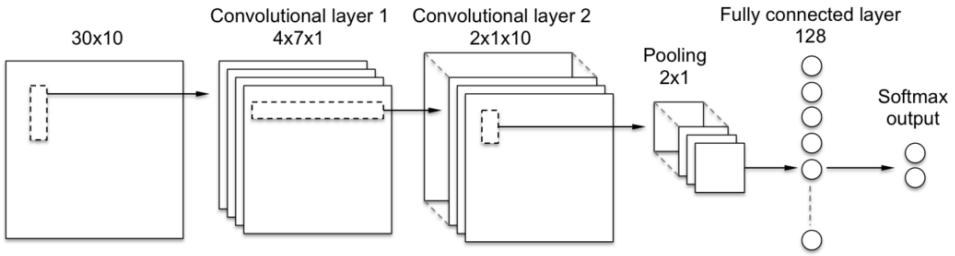
Pokročilé modely neurónových sietí ako hlboké konvolučné siete dosahujú veľmi dobré výsledky pri problémoch spracovania obrazu [17]. Za účelom využitia týchto vlastností štúdia skladá z používateľskej aktivity obrazovú mapu - dvojrozmerné pole normalizovaných pixelov. Za účelom učenia má každý obraz dostupné označenie, ktoré hovorí či daný zákazník prešiel ku konkurencii alebo nie. Obr. 2.10 zobrazuje ukážku aktivity zákazníka v dostupných službách za posledných n dní [19].



Obr. 2.10: *Aktivita zákazníka v mape pixelov. Hodnota pixelov sa zvyšuje od modrej k červenej [19].*

Experiment uvažuje 30-dňové okno predikcie, z ktorého sieť usudzuje aktivitu zákazníka. Okno sa nachádza 14 dní pred posledným registrovaný telefonátom. Pokiaľ sa posledný registrovaný telefonát nekonal v posledných 14 dňoch od aktuálneho dátumu, považujeme zákazníka za neaktívneho a neberieme ho do úvahy.

Po vytvorení obrazového datasetu z dostupných záznamov boli dáta podsunuté konvolučnej neurónovej sieti na obr. 2.11. Táto sieť má architektúru podobnú iným sieťam určeným pre spracovanie obrazu. Sieť analyzuje týždňové vzory v aktivite pomocou 7x1 filtra prvej konvolučnej vrstvy. Na konci siete je pomocou binárneho softmax klasifikátora vyhodnotený výsledok.



Obr. 2.11: *Architektúra konvolučnej siete pre klasifikáciu zákazníkov z pixelovej mapy aktivity [19].*

Pomocou metódy *oblasti pod krivkou*(AUC) bolo zistené, že konvolučná sieť dosahuje lepšie výsledku ako model CHAID rozhodovacieho stromu. AUC vyhodnocuje pravdivé aj nepravdivé pozitívne výsledky [8] [1].

Literatúra

[1] Andrew P Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159, 1997.

[2] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[3] Jonathan Burez and Dirk Van den Poel. Crm at a pay-tv company: Using analytical models to reduce customer attrition by targeted marketing for subscription services. *Expert Systems with Applications*, 32(2):277–288, 2007.

[4] Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. Learning to forget: Continual prediction with lstm. *Neural computation*, 12(10):2451–2471, 2000.

[5] Felix A Gers, Nicol N Schraudolph, and Jürgen Schmidhuber. Learning precise timing with lstm recurrent networks. *The Journal of Machine Learning Research*, 3:115–143, 2003.

[6] Alan Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. pages 6645–6649, 2013.

[7] Klaus Greff, Rupesh Kumar Srivastava, Jan Koutník, Bas R Steunebrink, and Jürgen Schmidhuber. Lstm: A search space odyssey. *arXiv preprint arXiv:1503.04069*, 2015.

[8] James A Hanley and Barbara J McNeil. The meaning and use of the

area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.

[9] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[10] Yoshua Bengio Ian Goodfellow and Aaron Courville. Deep learning. Book in preparation for MIT Press, 2016.

[11] Herbert Jaeger. *Tutorial on training recurrent neural networks, covering BPPT, RTRL, EKF and theëcho state networkäpproach*. GMD-Forschungszentrum Informationstechnik, 2002.

[12] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014.

[13] Vladimír Kvasnička, L'ubica Beňušková, Jiří Pospíchal, Igor Farkaš, Peter Tiňo, and Andrej Král'. *Úvod do teórie neurónovỳch sietí*. 1997.

[14] Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.

[15] Scott Neslin, Sunil Gupta, Wagner Kamakura, Junxiang Lu, and Charlotte Mason. Defection detection: improving predictive accuracy of customer churn models. *Tuck School of Business, Dartmouth College*, 2004.

[16] Anuj Sharma, Dr Panigrahi, and Prabin Kumar. A neural network based approach for predicting customer churn in cellular network services. *arXiv preprint arXiv:1309.3945*, 2013.

[17] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. pages 1–9, 2015.

[18] Aart T Van Halteren, Lambert JM Nieuwenhuis, Mike R Schenk, and

Maarten Wegdam. Value added web: Integrating www with a tina service management platform. pages 14–23, 1999.

[19] Artit Wangperawong, Cyrille Brun, Olav Laudy, and Rujikorn Pavasuthipaisit. Churn analysis using deep convolutional neural networks and autoencoders. *arXiv preprint arXiv:1604.05377*, 2016.