MACHINE LEARNING MODEL FOR CUSTOMER CHURN

Dushko Todevski
UGD Shtip, Republic of North Macedonia, duskotodevski@gmail.com
Vesna Georgieva Svrtinov
UGD Shtip, Republic of North Macedonia, vesna.svrtinov@ugd.edu.mk

Abstract: The subject of this paper is the presentation of a case for using artificial intelligence to increase the client base of companies. The data used in this research is the data from KAGGLE, more precisely they are taken from the IBM customer loyalty database. Part of this database is presented at the data mining competition PAKDD 20061. In this paper we present three models for artificial intelligence, logistic regression, gradient boosting and random forest. The results show that the best model can predict with 93% probability whether a potential customer would remain a customer of the company. Application of this model or similar models can be found in creating a policy for managing customer relations, but also for the overall sales policy.
Keywords: customer retention, machine learning, artificial intelligence

МОДЕЛ СО МАШИНСКО УЧЕЊЕ ЗА ЗАЧУВУВАЊЕ НА КЛИЕНТСКАТА БАЗА

Душко Тодевски
УГД Штип, Република Северна Македонија, duskotodevski@gmail.com
Весна Георгиева Свртинов
УГД Штип, Република Северна Македонија, vesna.svrtinov@ugd.edu.mk

Апстракт: Предмет на овој труд е презентирање на случај на користење на вештачка интлигенција за зголемување на клиентската база на компаниите. Податоците кои се користат во ова истражување се податоците од од KAGGLE, поточно истите се превземени од ИБМ базата за лојалност на клиенти. Дел од оваа база се податоци е претставена на натпреварот за податочно рударење PAKDD 20061. Во овој труд презентираме три модели за вештачка интелигенција, логистичка регресија, gradient boosting и random forest . Резултатите покажуваат дека кај најдобриот модел може со веројантост од 93 проценти да се предвиди дали еден потенцијален клиент би останал клиент на компнанијата. Апликативна примената на овој модел или слични модели можат да најдат при креирање на политика за менаџирање на односите со клиените, но и на целокупната продажна политика.
Клучни зборови: клиентска база, машинско учење, вештачка ителигенција

1. ВОВЕД

Глобалниот и локалниот пазар во современи услови се одликуваат со динамичен амбиент на дигитализација и со голема зависност од софтверски апликации и огромни бази на податоци. Во оваа смисла и современотото претприемништво е под огромно влијание на овие процеси. Едната насока на влијание кај претприемачите овозможува креирање на бизниси кои нудат решенија за подршка на компаниите и потрошувачите во дигиталната економија, додека другата насока на влијание е во подршка на веќе постоечките претприемачки бизниси со нова технологија, софтвер, вештачка интелигенција и deep learning и во нивна комерцијалализација кај посточеките бизнис процеси. Во суштина користењето на новите технологии е во голема мерка комплементарно со постоечките постулати на претприемништвото, особено во делот на комерцијализација на иновации како одговор на реалните потреби на бизнисот и потрошувачкото општество.

Слични турдови на оваат проблематика се и на (Umayaparvathi and Iyakutti, 2012) со наслов “Applications of data mining techniques in telecom churn prediction “ во кој се користат decision tree и neural network модели. Ова истражување треба да даде оценка дали со користење на понаков пристап на ова проблематика, пред се со користење на поинаков софтвер и на поинакви модели при предвидувањето може да се добие поголема предвидувачка моќ за дадениот пример.

Во овој труд се презентира еден вид на решение за продобрувње на перформансите на компниите со помош на модели кои ќе помогната во зачувување на клинетската база на телоком компаниите, но како пример овој модел може да се користи и во развивање на слично решение за подршка на продажбата во било која индустрија. Овој модел со соодветна апликативна поддршка може да биде предмет на имплементација како

софтверски модул во потоечките оперативни системи, но може да се користи и како додаток во некој од постоечките сервиси за продажба и односи со клиенит кои ги користат бизнисите. Претприемачкиот потеницијалот на овој случај на коиристење на вештачка интелигеницијa во продажбата е и во можноста во целост да се понуди како веб сервис на бизнси кои имаат потреба од ваков тип на порддршка.
Основна цел на ова истражување е да предложи можност на користење на најновите модели на вештачка интелигенија за целите на маркетингот. Во ова истражување се презентирааат податоци кои во голем дел се на располагање на голем број на мали, средни и голем компании во оваа или слична форма. Основната идеја и централна хипотеза на ова истражување е да се испита можноста за креирање и користење на модел за зачуваување на клиентската база на компанијата. Во случајов се користат податоците од orange телеком операторот од САД. За потребите на ова истражување се користат 20 варијабла

2. ПОДАТОЦИ И МЕТОДОЛОГИЈА

Податоците користени во овој модел се од KAGGLE, поточно истите се превземени од ИБМ базата за лојалност на клиенти. Дел од оваа база се податоци е претставена на натпреварот за податочно рударење PAKDD 20061. Податоците се објавени од азиска телеком компанија која успешно лансираше 3G мобилна телекомуникациска мрежа. Компанијата сакаше да идентификува кои клиенти најверојатно ќе се префрлат на користење на нивната 3G мрежа, со користење на постојната употреба на клиентите и демографските податоци. Изворната база се состои од 24.000 клиенти додека овде се користат податоци за 7043 клиенит. Секој клиент е опишан со 20 атрибути од кои дел се демографски атрибути, дел се податоци за клиентот, додека останатио дел се податоци со детали за деловниот однос на клиентот со телеком операторот
Некои од податоците кои се користет вклучува информации за:
Клиенти кои заминаа во последниот месец, потоа услуги на коишто се пријавил секој клиент – телефон, повеќе линии, интернет, безбедност на Интернет, резервна копија на Интернет, заштита на уредот, техничка поддршка и стриминг ТВ и филмови. Исто така има и информации за сметката на клиентите - колку долго тие биле клиенти, договор, начин на плаќање, наплата без хартија, месечни трошоци и вкупни трошоци.
Демографски информации за клиенти - пол, возрасен опсег и ако тие имаат партнери и зависни лица, но и колкав работн стаж имаат
Влезни варијабли:

1. Држава
2. Должина на сметка
3. Повикувачки код
4. Меѓународен план за разговор
5. План за говорна пошта
6. Број на пораки за пошта
7. Вкупно дневни минути
8. Вкупно дневни повици
9. Вкупно дневна наплата
10. Вкупно минути во вечерна тарифа
11. Вкупно вечерни повици
12. Вкупно вечерна наплата
13. Вкупно ноќни минути
14. Вкупно ноќни повици
15. Вкупно ноќни наплата
16. Вкупно интенационални минути
17. Вкупни интенационални повици
18. Вкупно меѓународно наплата
19. Повици за услуги на клиенти
20. Клиенти кои заминаа во последниот месец

За потребите на предвидувањето на овој модел се користат пред селектирани три модели на машинско учење и тоа логистичка регресија, random forest и gradient boosting. Примарно беа тестирани повеќе модели достапни во Orange data mining софтверот (Demšar et al., 2013; Demšar and Zupan, 2013), но при тестирањето со почетните параметри на модели претходно споменатите три модели се издвојуваа со резултатите на тренинг и тест податоците.
Логистичката регресија е класификационен алгоритам со примена на логистичка регресија и во овој случај користиме ласо регресија како параметар за класификација на независните ваирјабли. Random forest (Случајна шума) е комплексен метод на учење кој се развива со надградба не некој поедноставен модел

најчесто моделот на дрва на одлучување. Овој модел најчесто се користи за класификација, регресија и други задачи. Прво беше предложен од Тин Кам Хо и понатаму го развија Лео Брејман (Брејман, 2001) и Адел Катлер. Овде ќе се користи моделот од Orange верзијата на софтверот (Demšar et al., 2013; Demšar and Zupan, 2013). Random forest (Случајна шума) генерално се базиран на градење на пакет дрва на одлучување. Секое дрво е развиено од примерок за подигање од тренираните податоци. При развивање на одделни дрвја, се извлекува произволна подгрупа на атрибути (оттука и терминот „Случаен“), од кој е избран најдобриот атрибут за поделбата. Конечниот модел се заснова на мнозинството гласови од индивидуално развиените дрвја во шумата. Како параметри овде користиме 300 дрва и ограничување на под примероците да не бидат помали од 30 единици.

Gradient boosting е техника на машинско учење креирана да оговори на проблеми поврзани со регресија и класификација. Со овој алгоритам се генерира модел на предвидување кој е надградба на поедноставни модели на предвидување, обично дрва на одлуки (Friedman, 1999). Параметри кое се користеа во овој модел се: броеви на дрва 200, стапка на учење 0,1 со репликативен тренинг. За контрола на растот се користеа лимити од 5 на индивидуални дрва и ограничување на подпримероци со прараметар не помали од 3. Делот на тренинг истанциите беше сетиран на 0.85. а пресметки се користеше Gradient boosting од Scikit learn библиотеката

За потребите на предвидувањето на за останување во деловен одоно со телекомот ќе се користат повеќе техники за конструкција, тренирање и тестирање на моделите.

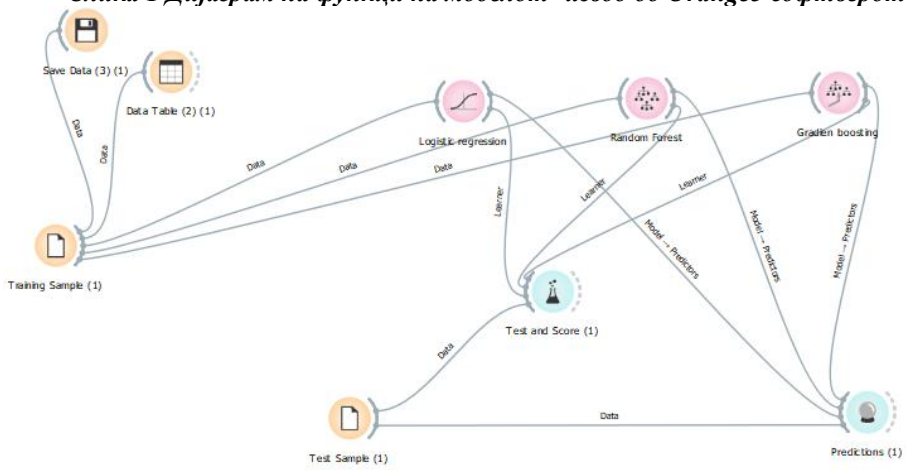
За потребите на крерање на примероци за тренирање и тестирање од почетниот податоци за корисниците 2666 индивидуални сметки на Orange клиенит, со што се прави поделба на почетниот примерок на два посебни дата сетови и тоа на тренинг и тест дата сет. Поделбата е направена во 80 проценти во корист на тренирањето на податоците, додека 20 проценти од податоците ќе се користат кај тест податоци за да се види успешноста на моделот со податоци кои не биле претходно користени за тренирање на моделите.

Фазата на тренирање на моделите се врши на тренинг податоците и тие по правило се секогаш во што е поголем обем за моделите да можат што е можно подобро да ги опфатат скирените релации и шаблоните кои настанале помеѓу вредностите за варијаблите. Ова практично значи дека моделите секој според капацитетот и перформансот на својот алгоритам се обидува да ја пронајде релацијата помеќу осумнаесетте варијабли во однос податоците од нивните кориснички сметки.

Логистичка регресија, random forest и gradient boosting моделите се користет за предвидување на податоците на у таргетирана променливата и кај тренираните и кај тестираните податоци при што се добиваат различни оценки за секој од моделите при тренирање и тестирањет на податоците.

Оценки на моделите ќе се врши во две фази од кои во едната фаза ќе се врши тестирање само на тренираните модели со претходно тренираните подататоци(примерок од 80%) со стратифицирана крос-валидација на 10 различни примери. Другата фаза на финално оценување ќе се реализира со оценување на моќта на предвидување на моделите на 20 проценити од податоците кои не се користеа за тренирање и се одвоија од првичниот примерок со цел да се провери прецизноста на конечно тренираните податоци.

Слика 1 Дијаграм на функци на моделот- извод од Orange3 софтверот



Оваа слика ги прикажува визуелно сите мапирани процеси од реализација на истражувањето од фазата на користење на податоците, предпроцесирање, креирање на тренинг и тест примерок, па се до користење на

моделите во тест и финалната фаза.

3. РЕЗУЛТАТИ И ДИСКУСИЈА

По спореведената конструкција на моделот и сите неопходни релации кои беа предходно дефинирани се чекори пред фазата на интерпретација на разултатите. При оценувањето на двете фази користење на моделите се користат следните параметри за оценување на спецификацијата на моделот:

- AUC (простор под ROC[6])
- CA (classification accuracy) е пропорција на точно класифицирани инстанци
- F1 е пондерирана хармонична средина на прецизност и recall параметарот
- Precision (прецизноста) е процентот на вистински позитиви предвидувања меѓу инстанците класифицирани како позитивни
- Recall процентот на вистински позитиви меѓу сите позитивни примери во податоците, на пр. бројот на болни кај сите дијагностицирани како болни.

За тест фазата на оценки на предвидувањата се користи примерок од одговорите на 667 испитаници, и предвидувањата ќе се вршат со моделите тренирани на овој обем на податоци.

Табела 1 Оценки од тест фаза на предвидувања

Модел	AUC	CA	F1	Precision	Recall	Просек AUC/CA
Random Forest	0.881	0.892	0.869	0.888	0.892	0.887
Logistic Regression	0.794	0.865	0.834	0.836	0.865	0.8295
Gradient Boosting	0.893	0.945	0.942	0.943	0.945	0.919

Согласно презентираното највисоки оцени кај моделите во финалната фаза на оценка на моделот највисока прецизност има кај Gradient Boosting моделот со 89,3% прецизност според AUC и 94,5% според CA, или во просек прецизност во однос на двата параметри од 91,9%. Просекот кај Logistic Regression е 83% додека кај Random Forest моделот е 88,7%. Овој резултат значи дека овој модел може да ја предвиди вредноста на таргетираната у (во случајот дали корисникот ќе го прекине деловниот однос со 91,9%, 83% и 88,7 % кај Gradient Boosting, Logistic Regression и Random Forest моделот.

За финалната фазата оценки на предвидувањата се користи претходно одвоениот примерок кој ги содржи податоците со одговорите на 667 испитаници (20% како тест примерок, додека 80% или 2666 инстатнци се користеа за тренирање на моделот). Мора да се напомени дека предвидувањата и во тест и во финалнат фаза се вршат со моделите тренирани на 80% (дата сетот од 2666 инстанци) од тренинг примерок на податоци.

Табела 2 Оценки од финалната фаза на предвидувања со тест податоци

Модел	AUC	CA	F1	Precision	Recall	Просек AUC/CA
Random Forest	0.926	0.934	0.927	0.935	0.934	0.927
Logistic Regression	0.822	0.862	0.832	0.831	0.862	0.8425
Gradient Boosting	0.920	0.963	0.961	0.962	0.963	0.945

Согласно презентираното највисоки оцени кај моделите во тест фазата на оценка на моделот на трениранитe податоци имаме највисоки прецизност кај Random forest моделот со 92% прецизност според AUC и 96,3% според CA, или во просек прецизносност во однос на двата параметри од 94,5%. Просекот кај Logistic Regression е 84,25% додека кај Random Forest моделот е 92,7%. Овој резултат значи дека овој модел може да ја предвиди вредноста на таргетираната у (во случајот дали корисникот ќе остане клиент или не со 94,5%, 82,2% и 92.7% кај Gradient Boosting, Logistic Regression и Random Forest моделот.

Во случајот на финалното оценување имаме за значително подоби резултати од оценувањето со трениранитe податоците од тест фазата, и може да се забележи дека Random forest моделот во финалната фаза на оценување на моделот го завзема второто место за разлика од тест фазата на оценување.

Значајно во двата случаи е што трениранитe модели можат без никаков проблем да се користат за апликација

[6] ROC крива е график што ја прикажува работата на моделот на класификација на сите прагови на класификација. Оваа крива исцртува два параметри вистинска позитивна стапка на предвидувања и лажна позитивна стапка на предвидувања

на овој тип на случаи и со овој тип на податоци. Препорака за имплементација на овие модел е дека прецизноста од во просек од 92-95% кај gradient boosting и random forest моделите може да има големо значење за оценување на континуитетот на делониот однос на клитенот на Orange телекомот. Овие модели даваат силна препорака дека вака конструирните модели можат со голема веројатност да предвидат кој од постоечките клиенти најверојатно би го продолжил деловниот однос со банката или пак за брзо време би го прекинал деловниот однос со банката.. Ова практично значи дека Orange како телеком бренд мора да се фокусира на деловните корисници кај кои е голема веројатноста да го прекинат деловниот однос со Orange телекомуникацискиот оператор.

4. ЗАКЛУЧОК

Ова истражување користи модели на вештачка инелигенција и машинско учење за предвидување на потенициjалот за предвидување на континуитетот на деловниот однос со клиентот. Идејата на овој труд е преку креирање на модел кој ќе прозилезе од постоечките податоци поврзани со клиентот и неговото користење на услуги и тарифи до Orange телеком операторот во САД, да се истражи можноста и потенцијалот за предвидување на континуитетот на деловниот однос со слични типови на податоци кои би биле на располагање како во оценуваните модели при нивното тренирање.

Во овој труд се користат податоци како на Orange телеком операторот од САД и главнта идеја е да се испита предвидувачкиот капацитет на овој дата примерок со помош на Orange 3 софтверот 3.29.3 и да се согледа предвидувачкиот капацитет на овие модели и тренинг податоци.

Резултатите на оценуваните модели (Random forest и Gradient boosting) во тест и во финалната фаза покажуваат прецизност од околу 93 проценти во просек и како модел кој би бил најсоодветен за имплементација во некое решение го предлагаат gradient boosting моделот со просечна прецизност од 94,5 проценти.

Овој резултат и ова ниво на прецизност покажува дека доколку би имале податоци како во првичниот примерок, со 94,5 проценти прецизност би можеле да оцениме дали клиентот има намера да го раскине или да остане во постоечкиот деловен однос со телеком операторот. Со ова практично можеме да предвидиме дали еден клиенти има преференции кон друг оператор и постои причина за незадоволсто од постоечкиот операт. Со подоетална анализа на факторите на овој модел би можеле и да се добијат првични сознанија за главните двигатели за одлуката за престанок на деловниот однос, но и за факторите кои придонесуваат еден клиент да остане во деловниот однос со телеком операторот. Вакава анализа може да прокаже кои се клучните посакувани тарифи и услови кои еден клиент го поврзуваат со бренд операрот и неговата понуда.

Апликативна примената овој модел или слични модели можат да најдат при анализа на бренд преференциите на многу индустрии, и секако истиот модел со помали или поголеми модификации може да се аплицира кај секоја компанија која има малку поголема база на податоци.

БИБЛИОГРАФИЈА

Ahmad, A. et al. Customer churn prediction in telecom using machine learning in big data platform. *Springer*. [Online]. Available at: https://link.springer.com/article/10.1186/s40537-019-0191-6 [Accessed 3 July 2021].

Ahmad, A. K., Jafar, A. and Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. *Journal of Big Data*, 6 (1). [Online]. Available at: doi:10.1186/s40537-019-0191-6 [Accessed 3 July 2021].

Alboukaey, N. et al. Dynamic behavior based churn prediction in mobile telecom. *Elsevier*. [Online]. Available at: https://www.sciencedirect.com/science/article/pii/S0957417420306035 [Accessed 3 July 2021].

Azeem, M. et al. A churn prediction model for prepaid customers in telecom using fuzzy classifiers. *Springer*. [Online]. Available at: https://link.springer.com/article/10.1007/s11235-017-0310-7 [Accessed 3 July 2021].

Demšar, J. et al. (2013). Orange: Data Mining Toolbox in Python Tomaž Curk Matija Polajnar Laň Zagar. *Journal of Machine Learning Research*, 14. [Online]. Available at: https://www.jmlr.org/papers/volume14/demsar13a/demsar13a.pdf [Accessed 27 May 2021].

Hashmi, N. and Butt, N. A. (2013). Customer Churn Prediction in Telecommunication A Decade Review and Classification Educational Data Mining View project Behavioural Modeling View project Customer Churn Prediction in Telecommunication A Decade Review and Classification. *researchgate.net*. [Online]. Available at: https://www.researchgate.net/publication/257920014 [Accessed 3 July 2021].

Idris, A., Iftikhar, A. and ur Rehman, Z. (2019). Intelligent churn prediction for telecom using GP-AdaBoost learning and PSO undersampling. *Cluster Computing*, 22, pp.7241–7255. [Online]. Available at: doi:10.1007/s10586-017-1154-3 [Accessed 3 July 2021].

Kamalraj, N., Prof, A. and Malathi, A. (2013). A Survey on Churn Prediction Techniques in Communication Sector.

International Journal of Computer Applications, 64 (5). [Online]. Available at: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.278.4171&rep=rep1&type=pdf [Accessed 3 July 2021].

Li, F. and Zou, Y. (2014). *The Impact of Credit Risk Management on Profitability of Commercial Banks: A Study of Europe*. [Online]. Available at: http://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-92799 [Accessed 14 June 2021].

Pamina, J. et al. (2020). Inferring machine learning based parameter estimation for telecom churn prediction. In: *Advances in Intelligent Systems and Computing*. 1108 AISC. 2020. Springer. pp.257–267. [Online]. Available at: doi:10.1007/978-3-030-37218-7_30 [Accessed 3 July 2021].

Ullah, I. et al. A churn prediction model using random forest: analysis of machine learning techniques for churn prediction and factor identification in telecom sector. *ieeexplore.ieee.org*. [Online]. Available at: https://ieeexplore.ieee.org/abstract/document/8706988/ [Accessed 3 July 2021].

Umayaparvathi, V. and Iyakutti, K. (2012). Applications of Data Mining Techniques in Telecom Churn Prediction. *International Journal of Computer Applications*, 42 (20). [Online]. Available at: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.259.1368&rep=rep1&type=pdf [Accessed 3 July 2021].