UDC: 005.934:347.961.3 Info M str. 51-58

АНАЛИЗА И ЕВАЛУАЦИЈА ПРОМЕНЕ КОНЦЕПТА ИЗЛАЗНЕ ПРОМЕНЉИВЕ У МАШИНСКОМ УЧЕЊУ
ANALYSIS AND EVALUATION OF CONCEPT DRIFT IN MACHINE LEARNING

Анђела Ристивојевић

РЕЗИМЕ: Модели машинског учења се сусрећу са променама у расподелама података, које могу утицати на перформансе и поузданост модела. Промене у подацима су најчешће узроковане догађајима из окружења и променама у корисничким преференцијама. Дешава се да су у реалном окружењу промене изненадне и ефикасно прилагођавање модела на исте изостаје. Описани проблем се назива промена концепта излазне променљиве (енг. Concept drift), док уочавање, праћење и решавање проблема промене концепта излазне променљиве побољшава стабилност модела. Технике за уочавање промена омогућавају континуалну анализу, али и прилагођавање модела променама.

У раду је дата свеобухватна дефиниција промене концепта излазне променљиве, уз истицање разлике између ове промене и промене у расподели улазних података, циљног атрибута и предикција, као и јасна дистинкција типова промене концепта излазне променљиве.

Истраживање је спроведено над скуповима података банке Banca Intesa ad Beograd, припремљеним тако да прате различите значајне догађаје у протекле четири године, са циљем симулације динамичног окружења у ком банка послује. Додатно, решаван је проблем бинарне класификације предвиђања продаје готовинских кредита, као једног од најзначајнијих производа банке. Истраживање се темељи на коришћењу одабраних метода за уочавање промене у расподелама улазних података, циљног атрибута, предикција и промене концепта излазне променљиве, које одговарају датом проблему бинарне класификације, моделу машинског учења и јавно су доступне.

Циљ истраживања је испитати постојање промена концепта излазне променљиве за одговарајући скуп података и одговарајући модел машинског учења. Уочавање свих врста промена омогућава боље разумевање животног циклуса модела и временског оквира у ком користимо модел машинског учења. За крај, дат је предлог корекције уочених промена.

КЉУЧНЕ РЕЧИ: промена концепта излазне променљиве, промене у расподелама улазних података, методе и алгоритми за уочавање промене концепта излазне променљиве, продаја готовинских кредита

ABSTRACT: Machine learning models often encounter changes in data distributions, which can affect their performance and reliability. These changes are typically driven by environmental events and shifts in user preferences. In real-world environments, changes are often abrupt, causing models to fail to adapt effectively, which is known as concept drift. Detecting, monitoring, and addressing concept drift can significantly enhance model stability. Detecting techniques enable continuous analysis and model adaptation.

This paper provides a comprehensive definition of concept drift, emphasizing the distinction between concept drift and data, target and prediction drift. It also offers a clear differentiation of concept drift types.

The research was conducted on datasets from Banca Intesa ad Beograd, prepared to reflect various significant events over the past four years, simulating the dynamic environment in which bank operates. Additionally, the study addresses the binary classification problem of predicting cash loan production, as one of the bank's most important products. The research focuses on the application of selected open-source methods for detecting data, target, prediction and concept drift. These methods correspond to the binary classification problem and the chosen machine learning model.

The study aims to examine the existence of concept drift for the given dataset and machine learning model. Detecting all types of changes enables a better understanding of the model's lifecycle and the time frame in which the machine learning model operates. Finally, the paper proposes approaches to correct the detected changes.

KEY WORDS: concept drift, data drift, methods and algorithms for detecting concept drift, personal/cash loan production

1. ОПИС ПРОБЛЕМА И ПРЕДМЕТА ИСТРАЖИВАЊА

Динамично и непредвидиво окружење, као и промене корисничких преференција и понашања, могу значајно утицати на моделе машинског учења. Имајући у виду *CRISP-DM* (енг. *CRoss Industry Standard Process for Data Mining*) парадигму и посматрање животног циклуса модела машинског учења, јасно је да није довољно само имплементирати модел у реалном окружењу, већ и посматрати понашање модела у продукцији [21, 23]. Добро обучени модели машинског учења могу временом давати све лошије предикције, а самим тим и пословне резултате [18, 3].

Поменут пад у перформансама модела током времена, се неретко темељи на променама у расподелама података. Наиме, уколико се нови подаци удаље од статистичке расподеле података над којима је модел научен, модел машинског учења неће бити у стању да пружи валидне предикције. Главна премиса модела машинског учења је да тренинг подаци над којима је модел обучен осликавају податке из реалног окружења веродостојно, те је на тај начин примена модела оправдана и сврсисходна. [7]

Истраживање је спроведено над подацима, тренутно, водеће банке на тржишту Републике Србије – *Banca Intesa ad Beograd*, чланице *Intesa Sanpaolo* групације. Конкретно, посматран је бинарни класификациони проблем продаје готовинских кредита клијентима, односно вероватноћу да клијент купи готовински кредит. Посматрајући различите демографске и бихевиоралне карактеристике клијената у погледу коришћења банчиних производа, модел ће уочити одређене патерне у понашању и предвиђати (не)куповину кредита.

Имајући у виду све велике промене у окружењу које су задесиле тржиште и саме клијенте у претходне четири године, користићемо неколико скупова података распоређених у различитим временским тренуцима, како бисмо симулирали динамичност и непредвидивост окружења у ком банка послује и указали на потенцијалну појаву промене концепта излазне променљиве. Почетни скуп података обухвата период јаке погођености вирусом *COVID-19* у периоду од 1. фебруара до 31. маја 2020. године, када су навике клијената биле изузетно специфичне услед увођења полицијског часа и немогућности напуштања кућа у периоду од касних поподневних до раних јутарњих часова. Управо овај период биће наш скуп за обучавање модела уз претпоставку да ће он забележити другачије понашање клијената од нама очекиваног у односу на периоде када пандемија вируса није алармантна.

Потом, тако обучен модел ћемо користити у реалном окружењу над неколико скупова података, те ћемо пратити постојање концептуалне промене излазне променљиве. Скупови над којима ћемо уочавати и решавати наш проблем биће:

- скуп података који осликава период јаке погођености COVID-19 вирусом у току 2021. године;
- скуп података који ће пратити почетак инфлације у току 2022. године; и
- најсвежији доступан скуп података из 2024. године.

Додатно, користићемо модел машинског учења, прилагођен и развијен за решавање проблема бинарне класификације. Посматраћемо понашање модела и његове евалуативне метрике у различитим временским тренуцима.

Сврха описаног истраживања је испитати постојање промене концепта излазне променљиве, што ће бити пропраћено и испитивањем постојања промена у статистичким расподелама улазних података. За овај корак истраживања биће коришћени алати за детектовање промене концепта излазне променљиве, односно промене у улазном скупу података. Потом, уз разумевање промена у скуповима података код којих су промене уочене, биће дат предлог третирања модела одговарајућим техникама. Дакле, истраживање се темељи на испитивању модела у погледу његове стабилности, робусности и респонзивности на промене. Додатно, циљ истраживања је и упоредити моћ алата за детектовање и дати предлог техника за кориговање проблема промене концепта излазне променљиве.

2. ТЕОРИЈСКИ КОНЦЕПТИ

Промена концепта излазне променљиве (енг. *Concept drift*) подразумева да се статистичка обележја излазне променљиве (енг. *target variable*), коју модел машинског учења покушава да предвиди, непредвидиво мењају током времена [22]. У случају да дође до промене концепта излазне променљиве, уочене законитости у старим подацима, неће бити релевантне за нове податке, што доводи до лошијих, односно мање прецизних предикција модела машинског учења [7].

Промену (енг. *Drift*) поред концепта излазне променљиве, можемо уочити и у расподели улазних података, лабела и предикција, што представља значајну информацију и може нам бити сигнал за постојање промене концепта излазне променљиве. Углавном је комбинација различитих промена у расподелама та која и узрокује крајњу промену концепта излазне променљиве. [7]

Ако су улазни подаци означени као X, стварне лабеле опсервација као y, предикције одговарајућих опсервација као $\hat{y}$, онда, математички можемо дефинисати [7]:

- промене у расподели улазних података (енг. *Data/ Feature drift*) подразумева промену $P(X)$;
- промене у расподели лабела (енг. *Target/Label drift*) подразумева промену $P(y)$;
- промене у расподели предикција (енг. *Prediction drift*) подразумева промену $P(\hat{y})$;
- промене концепта излазне променљиве (енг. *Concept drift*) подразумева промену $P(y|X)$.

Како модели машинског учења банке користе различите улазне податке клијента, јасно је да ће било какве драстичне промене у расподели улазних података – демографских, бихевиоралних или трансакционих, значајно утицати на сам модел. Примера ради, промена расподела улазних података може бити промена дистрибуције зараде клијената, што значајно може утицати на перформансе модела будући да она представља значајан улазни податак разних модела.

Пример промене у расподели лабела може бити заступљен код модела машинског учења за предвиђање кредитног ризика. Одељење ризика банке може променити дефиницију немогућности отплате кредита (енг. *loan default*) због промене кредитне политике или економске ситуације на тржишту. Дакле, лабеле додељене опсервацијама за обучавање модела у том случају нису исправне, те модел обучен на старим вредностима лабела неће бити у стању да предвиђа поштујући нову дефиницију излазних атрибута и самим тим је поузданост и тачност модела нижа.

Промене у расподели предикција можемо уочити код модела за предикцију напуштања банке (енг. *churn prediction*) или код модела за превенцију екстерног рефинанса кредита. Наиме, законитости које је модел научио се могу временом мењати, те су онда предикције модела од све мањег значаја.

У литератури се истиче и термин промене у оквиру самог модела (енг. *Model drift*). Овај термин се односи на пад

перформанси модела, без одређивања специфичног разлога, односно само истичемо уочену чињеницу да модел више нема задовољавајуће перформансе и да губи своју сврху. Разлог за лошије перформансе модела може бити органски, али исто тако често може бити и промена концепта излазне променљиве, па се ова два појма неретко у литератури и пракси сусрећу заједно [12]. Додатно, органски пад у перформансама модела је очекиван, али га је могуће и ублажити уколико спроводимо поновно тренирање модела.

Као што је већ речено, промену концепта излазне променљиве углавном везујемо за екстерне факторе. На основу учесталости промена из окружења, можемо дефинисати различите типове проблема промене концепта излазне променљиве. У литератури најзаступљенија је подела на четири главна типа: изненадни, градуални, инкрементални и рекурентни тип [18].

Изненадни тип (енг. *Sudden concept drift*) се базира на идеји да до промене концепта излазне променљиве долази на основу до сада невиђених и изненадних промена у окружењу [15]. То су углавном неочекивани догађаји, чије последице имају драстичне ефекте на окружење. Познато је да су готово све индустрије биле погођене изненадном пандемијом *COVID-19* вируса. Наравно, саме преференције корисника су се драстично промениле, што је утицало на све сфере пословања. Примера ради, услед пандемије клијенти су због неизвесности сталног запослења били мање склони куповини стамбених кредита.

Градуални тип (енг. *Gradual concept drift*) подразумева да се промене дешавају након дужег временског периода, односно да се како време пролази у пословању неке промене дешавају органски [9]. Наиме, клијенти ће временом мењати своје навике, поготово ако на њих полако утичу спољни фактори. Пример може бити утицај инфлације, која не долази нагло, већ постепено, и утиче на пословање и навике клијената. Додатно, истиче се и инкрементални тип (енг. *Increment concept drift*), који можемо сматрати посебном врстом градуалног типа, где се промене како и сам назив каже, дешавају инкрементално, са могућношћу бољег праћења тренда у односу на класичан градуални тип [15].

Рекурентни тип (енг. *Recurring/Reoccuring concept drift*) карактеришу периодичне промене, односно одређена сезоналност. Познато нам је да су навике клијената другачије око празника и периода годишњих одмора [9]. Модел машинског учења можда не узима у обзир сезоналности, уколико није учен на довољно дугом временском периоду да се те сезоналности испоље. Примера ради, модел је могао бити над подацима који су у стабилнијем временском периоду попут септембра и октобра, те да се сезоналност у погледу празника у новембру и децембру не испољи и из угла модела не очекује. Управо зато се клијентима и нуде различите понуде у таквим периодима године. У банкарском сектору нуде се специјалне каматне стопе клијентима у периодима новогодишњих празника, дана државности, црног петка, те треба имати у виду и који период године ће се одабрати као полазни скуп за обучавање, како би што боље осликао понашање клијената.

2. МЕТОДОЛОГИЈА ИСТРАЖИВАЊА

Као што је већ речено, истраживање је спроведено над подацима банке *Banca Intesa ad Beograd*, те је почетна фаза истраживања била, најпре, прикупљање података, коришћењем релационог упитног језика *SQL (Structured Query Language).*

За само истраживање коришћена су четири скупа података, при чему је битно разликовати за који период су прикупљени подаци о понашању клијената, а за који подаци о реализацији куповине готовинских кредита. Податке о понашању клијената ћемо прикупљати за четири везана месеца. Потом ћемо, будући да је модел потребно користити у датом месецу, прескочити један месец који се односи на период употребе модела (енг. *black period*). За крај циљни период за који ћемо прикупљати податке о реализацији куповине готовинских кредита ће бити три везана месеца након месеца употребе модела. За овакав начин прикупљања података смо се одлучили како бисмо симулирали коришћење модела машинског учења у реалном окружењу у оквиру различитих продајних кампања.

Слика 1 - *Временска одредница скупова података*

Упоредни скуп ће нам бити почетни скуп података – период понашања клијената током карантина и строгог неизлажења, и ту расподелу података ћемо поредити са осталим расподелама. Наравно, сам модел ће бити обучен над периодом понашања у току карантина уз информацију о томе каква је реализација у одговарајућем циљном периоду. Касније, модел ће за остала три задата скупа бити примењен над периодима понашања клијената, где ћемо онда посматрајући циљне периоде моћи да пратимо стопу грешке и покушамо да уочимо промене у концепту излазне променљиве.

За потребе истраживања су осмишљене 34 варијабле, које описују понашање сваког клијента. Варијабле можемо поделити у три кључне групе – социо-демографске варијабле, варијабле које описују клијентово коришћење банчиних производа и трансакционе варијабле. На тај начин можемо стећи увид у то како клијент користи већ постојеће услуге које му банка пружа, да ли има услова

за куповину готовинског кредита, али и које су његове преференције у погледу потрошње у оквиру кључних индустрија – здравства, телекомуникација, прехрамбених ланаца и компјутерске опреме. Поменуте индустрије одабране су као кључне будући да је референтна тачка период пандемије и карантина. Варијабле су већински нумеричког типа – 31 варијабла од чега су 3 нумеричке бинарне варијабле (енг. *flag*) и 3 категоричке варијабле.

Почетни скуп података над којим ће модел бити обучаван – период понашања од 01.02.2020. године до 31.05.2020. садржи 60.000 клијената *Banca Intesa ad Beograd*, одабраних тако да су селектовани сви запослени клијенти са реализацијом односно куповином готовинског кредита у циљном периоду, а остатак клијената је одабран насумично од клијената без реализације, те је битно истаћи и да је скуп података небалансиран. За почетни скуп података однос класа излазне варијабле је је 80:20 у корист већинске класе 0 – клијенти који нису имали куповину готовинског кредита. На идентичан начин, креирана су и остала три скупа података, уз напомену да се однос класа излазне варијабле временом мења, достижући и однос од 65:35 у последњем доступном скупу података.

Тачни називи варијабли, као и њихови начини израчунавања, ће бити изостављени из рада због поштовања приватности података клијената *Banca Intesa ad Beograd*.

Спроведена је анализа недостајућих вредности и недостајуће вредности су третиране у складу са типом варијабле и пословном логиком банке. Како се испитивање расподела улазних података базира искључиво над доступним подацима, присуство одређеног броја недостајућих вредности, ни на који начин не ремети разумевање и уочавање промена [10], али задржава стварну слику окружења без додавања шума у подацима.

У оквиру истраживања решавамо проблем бинарне класификације продаје готовинских кредита, при чему класа 1 означава да ће клијент купити готовински кредит у периоду предвиђања, док 0 означава да неће. Додатно, како класификациони модели враћају вероватноћу за остваривање одређеног догађаја, користићемо праг одлучивања од 0,5, те ћемо свим клијентима који имају вероватноћу за куповину кредита већу од 0,5 доделити 1, док ћемо им у супротном доделити 0.

За решавање овог проблема одабран је један од тренутно најпопуларнијих алгоритама машинског учења за табеларне податке [8, 19] и стандард у индустрији - ***XGBoost (eXtreme Gradient Boosting).*** *XGBoost* је скалабилна, дистрибуирана библиотека машинског учења базирана на градијентном појачавању стабала одлучивања (енг. *Gradient Boosting Decision Tree - GBDT*) [19]. Овај алат је посебно познат по својој ефикасности и способности да се носи са великим скуповима података, подржава различите оптимизације и регуларизације стабала, као и дистрибуирано тренирање на више процесорских језгара [8]. *XGBoost* је популаран у индустрији због своје високе тачности и брзе имплементације, чинећи га кључним алатом за комплексне задатке машинског учења [8].

Као и сваки алгоритам машинског учења, *XGBoost* поседује различите хипер-параметре које је потребно прилагодити решаваном проблему [24]. Дефинисане су различите вредности које хипер-параметри модела могу имати и то коришћењем библиотеке *HyperOpt*, која нам помаже у проналажењу оптималних вредности и саме конфигурације хипер-параметара. Хипер-параметар може узети било коју вредност из унапред дефинисаног интервала [5] уз дефиницију начина на који ће се вршити претрага оптималне вредности у интервалу. Одабрано је *tree-structured prazen estimator – TPE* претраживање, односно претраживање на основу информација из претходних итерација, будући да у односу на случајну претрагу брже конвергира ка оптималном решењу.

Додатно, како бисмо пронашли оптималну комбинацију хипер-параметра користимо унакрсну валидацију (енг. *cross-validation*) [6]. Унакрсна валидација помаже при решавању проблема пренаучености модела (енг. *over-fitting*) и обезбеђује бољу генерализацију модела [6]. Најпре, скуп података смо поделили у односу 75:25, где ћемо 75% почетног скупа (45,000 инстанци) користити за учење модела и спровођење унакрсне валидације. За сваку комбинацију хипер-параметара, спроводимо унакрсну валидацију која дели скуп на 5 делова (енг. *fold*). Затим, тај модел са одговарајућом комбинацијом хипер-параметара, учимо над 4 дела нашег скупа, а над петим тестирамо, што итеративно спроводимо 5 пута, сваки пут користећи по један различит скуп за тестирање и преостала 4 скупа за тренирање одговарајућег модела [6].

Прецизније, у оквиру функције циља, коју ће *HyperOpt* техника користити дефинисана је унакрсна валидација. Наравно, како бисмо проценили успешност на унакрсној валидацији неопходно је дефинисати метрику којом меримо колико је модел, научен над 4 дела, успешно предвиђао над петим делом скупа података. За ове потребе дефинисане су две метрике – техничка и пословна метрика.

Техничка метрика, која ће бити директно везана и са претрагом хипер-параметра и која ће нам пружити информацију о перформансама модела је површина испод ROC *(Receiver Operating Characteristics*) криве (енг. *Area Under the Curve – AUC*). Интеграл, односно површина испод *ROC* криве дефинише способност модела да прави разлику између позитивне и негативне излазне класе. *AUC* мера, дакле, осликава успешност модела да врши сепарацију између ове две класе. Домен над којим се дефинише *AUC* је [0,1], при чему је правило да што је вредност *AUC* метрике виша, то је модел бољи са аспекта раздвајања класа. Идеал који је у пракси углавном недостижан, је да површина испод криве износи 1. Иако су вредности близу 0 сигнал за ниску сепарабилност модела, то су ипак корисне информације, јер значи да модел заправо показује инверзан резултат – негативној класи додељује вредност позитивне и обратно. Најгори случај је када метрика износи 0.5, будући да тада модел готово случајно и без правила одређује класе. [13, 11, 20]

Прилагођена пословна метрика је уведена, како би се поред перформанси модела, пратила и његова корисност

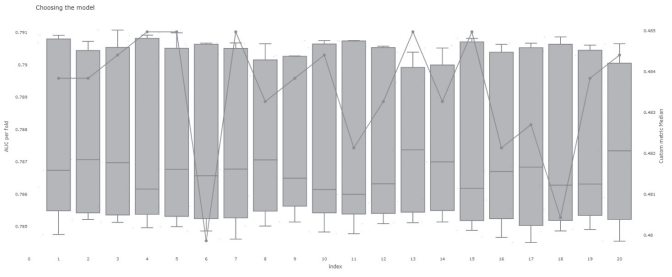
у каснијој употреби, али и како би модел касније био интерпретабилнији интересним странама. Као што знамо, у итерацији унакрсне валидације за део скупа, предвиђамо за сваког клијента колика је вероватноћа за куповину готовинског кредита. Одабиром 20% клијената са највишом вероватноћом, желимо да видимо колики проценат стварне продукције кредита покривамо. Математички, метрика би била дата као:

$$\text{Пословна метрика} = \frac{\text{20\% кл—-–ијената са највишом вероватноћом за куповину кредита}}{\text{укупан број свих реализованих куповина кредита}}$$

Од укупних 500 комбинација, најпре је издвојено 20 најбољих модела по средњој вредности *AUC* метрике и за тих 20 комбинација је урађен графички приказ за одабир модела. За сваку комбинацију хипер-параметара памтимо 5 вредности *AUC* метрике и 5 вредности прилагођене пословне метрике (које се добијају по једна у свакој итерацији унакрсне валидације), што ће нам помоћи при финалном одабиру модела.

За свих 20 комбинација на основу вредности *AUC*-а конструисан *box plot*, који ће нам помоћи да визуелно тумачимо стабилност модела. *Box plot* је начин да се прикаже расподела нумеричких података, која даје преглед централне тенденције, распона и расподеле података, тако што се графика јасно очитавају максимална и минимална вредност (дате у виду линија), први и трећи квартил (дати као ивице кутије) и медијана расподеле (дата као линија између квартила односно у оквиру кутије) [4, 17].

Додатно, приказан је и линијски график (енг. *line plot*) заснован на вредности медијане пословне метрике за сваку комбинацију хипер-параметара. На овај начин, желимо да изаберемо модел који је и стабилан и даје добре перформансе са техничког аспекта, али који даје и добре резултате значајне за пословање банке.



Слика 2 - Одабир модела победника

Уколико гледамо *box plot*, стабилност модела осликава позиција медијане [4] AUC-а , те је циљ одабрати моделе где је медијана позиционирана што је ближе могуће половини кутије [4]. Додатно, желимо да је *box plot* мањег распона, јер то значи да је стабилнији [4]. С друге стране, посматрајући линијски график, са аспекта медијане пословне метрике, желимо да вредност буде што виша. Због свега наведеног одабран је модел по редним бројем 13, као модел победник. Медијана пословне метрике модела победника износи 0,48497, што значи да избором 20% клијената највероватнијих за куповину кредита покривамо чак 48,497% целокупне реализације кредита. С друге стране, медијана *AUC* метрике износи 0,7874, што указује на добре перформансе модела. Модел победника ћемо обучити над 75% тренинг скупа и сачувати као обучен модел, који ћемо касније користити за предикције над осталим скуповима података.

У наставку истраживања коришћене су технике за уочавање промена у расподелама улазних података, у расподели излазне променљиве и у расподели предикција, при чему постојање наведених промена може бити чест узрочник промене концепта излазне променљиве. Дакле, ове технике су спроведене како бисмо стекли целокупну слику, али и знали шта да очекујемо од примене нашег модела. За крај истраживања спроведене су и технике за уочавање промене концепта излазне променљиве.

За уочавање промене концепта излазне променљиве (енг. *concept drift*) коришћене су две методе засноване на праћењу стохастичке стопе грешке.

За сва три периода од интереса, прво је примењен алгоритам ***Drift Detection Method (DDM)***, који прецизно уочава изненадни и градуални тип промене (нормалне брзине ка бржем) [1].

Drift Detection Method (DDM) приликом пристизања нове инстанце, рачуна стохастичку стопу грешке, а потом ажурира границу за упозорење и границу за аларм тј. границу за уочавање промене концепта излазне променљиве [14]. Алгоритам користи правило два и три сигма, при чему су границе за упозорење и уочавање промене концепта излазне променљиве, респективно, дефинисане као [14]:

$$p_i + s_i \geq p_{min} + 2 * s_{min}$$
$$p_i + s_i \geq p_{min} + 3 * s_{min}$$

где је p_i стохастичка стопа грешке, а s_i припадајућа стандардна девијација, док су p_{min} и s_{min} минимална стохастичка стопа грешке и минимална стандардна девијација забележене до датог тренутка.

Како желимо испитати и постепену градуалну промену, односно постојање споре градуалне промене, за сва три периода од интереса је примењен алгоритам ***Early Drift Detection Method (EDDM)***, који је у односу на Drift Detection Method (DDM) ефикаснији у овом погледу [1].

Early Drift Detection Method (DDM) приликом пристизања нове инстанце, такође, рачуна стохастичку стопу грешке, а потом ажурира границу за упозорење и границу за аларм односно уочавање промене концепта излазне променљиве. При чему су границе за упозорење и уочавање промене концепта излазне променљиве, респективно, дефинисане као [2]:

$$\frac{p_i + 2 * s_i}{p_{\max} + 2 * s_{\max}} \leq \alpha$$
$$\frac{p_i + 2 * s_i}{p_{\max} + 2 * s_{\max}} \leq \beta$$

где је стохастичка стопа грешке, а припадајућа стандардна девијација, док су и максимална стохастичка стопа греш-

ке и стандардна девијација забележене до датог тренутка. Подразумеване вредности за параметре су 0.95 за параметар α и 0.9 за параметар β. [2]

3. РЕЗУЛТАТИ ИСТРАЖИВАЊА

Први део истраживања подразумева испитивање постојања промена у расподелама улазних података (енг. *data drift*), коришћењем скупова података који се односе на понашање клијената. Креирана су три извештаја, где је као референтни (енг. *benchmark*) период коришћен скуп података од 01.02.2020. – 31.05.2020. године, односно период карантина услед пандемије вируса. Периоди од значаја, које изучавамо, су преостала три скупа који се односе на понашање клијената, наравно допуњени уз излазну променљиву која се односи на одговарајући циљни период.

Сва три извештаја пореде расподеле улазних података за све колоне у два временска периода, при чему се у извештају проверава и статистичка расподела излазне променљиве. Подразумевана референтна вредност за проглашавање статистички значајне промене целокупног скупа података је 0,5, односно, ако се за 50% и више варијабли уочи да постоји промена у расподелама улазних података, онда проглашавамо да се скуп података статистички значајно променио (енг. *Dataset drift*).

За период пика пандемије није детектована промена целокупног скупа података (енг. *Dataset drift*)**,** будући да је промена у расподелама улазних података (енг. *Data drift*) уочена код **16 атрибута**, од укупних 35 атрибута, односно код **45,714%** атрибута. Такође, није ни уочена промена у расподели излазне променљиве (енг. *Target/Label drift*). Имајући у виду да промене у расподелама нису статистички значајне, очекујемо да ће модел давати задовољавајуће резултате у посматраном периоду, као и да је употреба овог модела валидна. Наравно, промена концепта излазне променљиве као и промене у перформансама модела, не морају нужно бити изазване променама у расподелама улазних података и излазне променљиве, али ипак су то чести узрочници пада перформанси.

Из извештаја за период инфлације видимо да је детектована промена целокупног скупа података (енг. *Dataset drift*), будући да је промена у расподелама улазних података (енг. *Data drift*) уочена код **19 атрибута**, од укупних 35 атрибута, односно код **54,286%** атрибута. Међутим, није уочена промена у расподели излазне променљиве (енг. *Target/Label drift*). Промене у расподелама улазних података су статистички значајне, док промене у расподели излазне променљиве нису, те не можемо са сигурношћу рећи какве ће перформансе модел имати. У случају да чак и употреба модела буде валидна и да не буде пада у перформансама модела и промене концепта излазне променљиве, ипак би требало размотрити кориговање модела, будући да промене у расподелама онда можемо гледати као аларм односно упозорење на будуће проблеме. Наравно, поменуте промене могу пак утицати на генерисање лоших предикција и резултата, те би тада дефинитивно било потребно кориговати модел машинског учења.

За крај, из извештаја за последњи доступан период видимо да је детектована промена целокупног скупа података (енг. *Dataset drift*), будући да је промена у расподелама улазних података (енг. *Data drift*) уочена код **25 атрибута**, од укупних 35 атрибута, односно код **71,429%** атрибута. Такође, уочена је и промена у расподели излазне променљиве (енг. *Target/Label drift*). Због статистички значајних промена, очекујемо да употреба модела неће бити валидна, као и да ћемо уочити промену концепта излазне променљиве и пад у перформансама модела. Посебно је алармантно што је проценат промене целокупног скупа података 71%, те је кориговање модела и његово поновно тренирање крајње извесно.

Модел победника који смо обучили над 75% тренинг скупа, ћемо најпре применити на до сада неискоришћених 25% почетног тренинг скупа и добићемо расподелу предикција. Управо, ова расподела ће нам бити упоредна и користићемо је као полазну тачку за уочавање промена у расподели предикција (енг. *prediction drift*). Модел је примењен и над преостала три скупа података и на тај начин ћемо добити расподеле предикција у различитим временским тренуцима. Све три расподеле понаособ поредимо са упоредном расподелом предикција (енг. *benchmark*) и за сваки период тумачимо да ли је дошло до промене у расподели предикција.

Примећујемо да за циљни период пика пандемије промена у расподели предикција није статистички значајна, ни из угла вредности вероватноћа, ни из угла додељених класа. Ова чињеница додатно оснажава наша очекивања да ће модел давати задовољавајуће резултате у посматраном периоду, као и да је употреба модела у посматраном периоду валидна.

За циљни период инфлације промена у расподели предикција није статистички значајна из угла додељених класа, међутим, уколико посматрамо расподелу предвиђених вероватноћа, онда примећујемо статистички значајну промену. С тим у вези, не можемо са сигурношћу рећи да ли ће модел давати задовољавајуће резултате, као и да ли је употреба модела у периоду инфлације валидна. Будући да су уочене парцијалне промене у расподели предикција, требало би ипак разматрати кориговање модела, ако не у посматраном, онда у скоријем периоду.

Примећујемо статистички значајну промену у расподели предикција за последњи доступан период понашања и циљни период. Промена је статистички значајна и у погледу вредности вероватноћа, и у погледу додељених класа. Имајући у виду и претходно уочене статистички значајне промене у расподелама улазних података и расподели циљаног атрибута, очекујемо да употреба модела више није валидна, као и да ћемо уочити статистички значајну промену у концепту излазне променљиве.

За крај, осврнућемо се на резултате приликом спровођења метода за уочавање промене концепта излазне променљиве.

У наставку, приказани су сумарни резултати за сва три периода, као и одговарајући проценти уочених упозорења

или аларма у случају коришћења технике ***Drift Detection Method (DDM)***, која нам омогућава откривање нагле односно изненадне промене концепта излазне променљиве.

Такође, у другој табели, приказани су резултати и у случају коришћења ***Early Drift Detection Method (EDDM)*** методе, која нам омогућава уочавање споре градуалне промене концепта излазне променљиве. Детаљније тумачење ових резултата, дато је у оквиру поглавља Анализа резултата и дискусија.

Период	Догађај од интереса	Уочена упозорења (број)	Уочени аларми (број)	Уочена упозорења/ аларми (%)
период понашања ***01.05.2021. – 31.08.2021.*** и циљни период ***01.10.2021. – 31.12.2021.***	пик *COVID-19* пандемије	5	0	0.01%
период понашања ***01.09.2021. – 31.12.2021.*** и циљни период ***01.02.2022. – 30.04.2022.***	инфлација	577	1	0.96%
период понашања ***01.10.2023. – 31.01.2024.*** и циљни период ***01.03.2024. – 31.05.2024.***	последњи доступан скуп	8,543	1	14.24%

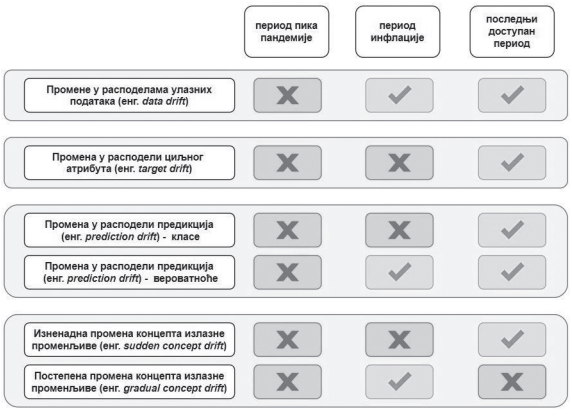
Слика 3 - Резултати коришћења Drift Detection Method (DDM) методе

Период	Догађај од интереса	Уочена упозорења (број)	Уочени аларми (број)	Уочена упозорења/ аларми (%)
период понашања ***01.05.2021. – 31.08.2021.*** и циљни период ***01.10.2021. – 31.12.2021.***	пик *COVID-19* пандемије	706	104	1.35%
период понашања ***01.09.2021. – 31.12.2021.*** и циљни период ***01.02.2022. – 30.04.2022.***	инфлација	6,333	138	10.79%
период понашања ***01.10.2023. – 31.01.2024.*** и циљни период ***01.03.2024. – 31.05.2024.***	последњи доступан скуп	827	54	1.47%

Слика 4 - Резултати коришћења Early Drift Detection Method (EDDM) методе

4. АНАЛИЗА РЕЗУЛТАТА И ДИСКУСИЈА

Ради лакшег прегледа резултата и разумевања истих, можемо погледати графички приказ, који за све посматране периоде истиче да ли су одређене промене уочене или нису.



Слика 5 - Графички приказ свих резултата истраживања за све посматране периоде

Резултати истраживања добијени за први скуп података, односно период када се десио пик пандемије вируса *COVID-19* у 2021. години, указују на свеукупно непостојање статистички значајних промена. За овај период нису уочене статистичке значајне промене ни у расподелама улазних података, ни у расподели излазног атрибута, нити у расподели предикција. Такође, коришћењем метода за уочавање промена концепта излазне променљиве није уочена ни изненадна, али ни спорија градуална промена. Овакви резултати сугеришу валидност и релевантност употребе модела машинског учења у току посматраног периода. Почетни период, над којим је тренирaн модел и који је коришћен као упоредни за све анализе, карактерише карантин и пословање у периоду јаке погођености вирусом, те стога није изненађење што за период пика пандемије који је релативно близак поменутом нисмо уочили статистички значајне промене. Прецизније, разлика између ова два периода је нешто више од годину дана, што је значајно мање од разлике између осталих скупова и почетног скупа, али и додатно описују јако сличан догађај од интереса – карантин и јаку погођеност вирусом у пику пандемије.

С друге стране, за други период, где је догађај од интереса почетак инфлације, немамо тако јасну слику. За почетак, уочене су промене у расподелама улазних података и расподели вероватноћа самих предикција, али нису уочене промене у расподели излазне варијабле и расподели додељених класа путем предикција. Због уочених парцијалних промена и непредвидивости која карактерише овај период, као и уочене постепене градуалне промене концепта излазне променљиве, потребно је предузети одређене корективне мере.

Наиме, прва могућност би била поновно тренирање модела на скупу података, који ће временски бити ближи овом посматраном транзиционом периоду, како би се умањио ефекат те градуалне промене.

С друге стране овакав транзициони период можемо посматрати и као одличну прилику за припрему израде новог модела или поновно тренирање постојећег модела уз модификацију тежинских коефицијената скупа. Наиме, можемо посебан акценат ставити на клијенте који су купили кредит непосредно пре транзиционог периода и тренирати модел над свим подацима, али уз акценат на новијим инстанцама.

Такође, у току периода градуалне промене, можемо користити и адаптивне ансамбл алгоритме који се могу сами прилагодити променама. Један пример ових алгоритама су адаптивне случајне шуме (енг. *Adaptive Random Forest*), које имају могућност динамичног прилагођавања. Овај модел машинског учења може самостално заменити стара стабла одлучивања оним релевантнијим стаблима, која боље осликавају тренутну дистрибуцију података. [16]

Наравно, корисници модела, још увек немају алармантну слику у датом тренутку, те је и употреба постојећег модела, док се он не замени одговарајућим новим моделом, прихватљива у датом тренутку. Ипак, препорука је, имајући овакве резултате у виду, са великом важношћу размотрити и одабрати оптималну корективну меру. Односно, било избором и изградњом новог модела, било поновним тренирањем постојећег модела, пожељно је реаговати на уочену постепену градуалну промену.

Резултати трећег, односно последњег доступног периода, указују на неопходно реаговање на промене. Као

што видимо, за овај период уочене су статистички значајне промене у расподелама улазних података, расподели циљног атрибута и предикција. Није уочена спорија градуална промена, већ нагла, изненадна промена у концепту излазне променљиве. Овакав сценарио указује нам на невалидност употребе овог модела у пословном окружењу банке. Дати модел не даје поуздане резултате, те његова употреба није оправдана.

Почетни корективни корак може бити поновно тренирање модела на скупу података, доста временски ближем од тренутног почетног скупа. Наиме, тренутно је модел трениран на скупу података из 2020. године, док предвиђа догађаје из 2024. године.

Ипак, због толике промене у окружењу сигурније би било применити радикалније мере. Изградња новог модела, уз додатне нове варијабле, које би додатно описале тренутно окружење у ком банка послује. Целокупну анализу, прикупљање и припрему података, потом тренирање и одабир модела победника би, требало поново спровести, што наравно захтева додатно време.

Овакви резултати су и очекивани, будући да је период пандемије давно завршен и да се пословање додатно изменило након исте, поготово у погледу дигиталне писмености клијената. Такође, баш зато што је период од неколико година након изласка из кризне ситуације, потреба клијената за производима банке је сада већа, него у периоду карантина.

5. ЛИТЕРАТУРА

[1] Agrahari, S., & Singh, A. K. (2022). Concept drift detection in data stream mining: A literature review. Journal of King Saud University-Computer and Information Sciences, 34(10), 9523-9540.

[2] Baena-Garcıa, M., del Campo-Ávila, J., Fidalgo, R., Bifet, A., Gavalda, R., & Morales-Bueno, R. (2006, September). Early drift detection method. In Fourth international workshop on knowledge discovery from data streams (Vol. 6, pp. 77-86).

[3] Bayram, F., Ahmed, B. S., & Kassler, A. (2022). From concept drift to model degradation: An overview on performance-aware drift detectors. Knowledge-Based Systems, 245, 108632.

[4] Benjamini, Y. (1988). Opening the Box of a Boxplot. The American Statistician, 42(4), 257–262.

[5] Bergstra, J., Yamins, D., Cox, D. D. (2013) Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. TProc. of the 30th International Conference on Machine Learning (ICML 2013), June 2013, pp. I-115 to I-23.

[6] Berrar, D. (2019). Cross-validation.

[7] Bressler, N. (2022). How to Detect Concept Drift with Machine Learning Monitoring. https://deepchecks.com/how-to-detect-concept-drift-with-machine-learning-monitoring/ (приступано: мај 2024)

[8] Brownlee J. (2018). XGBoost with Python. Gradient Boosting Trees With XGBoost and scikit-learn.

[9] Das, S. (2023). Best Practices for Dealing with Concept Drift. https://neptune.ai/blog/concept-drift-best-practices (приступано: мај 2024)

[10] Evidently AI. (n.d.). Data drift algorithm. Input data requirements. https://docs.evidentlyai.com/reference/data-drift-algorithm (приступано: јун 2024)

[11] Evidently AI. (n.d.). How to explain the ROC curve and ROC AUC score? https://www.evidentlyai.com/classification-metrics/explain-roc-curve (приступано: јул 2024)

[12] Evidently AI Team. (n.d.). What is concept drift in ML, and how to detect and address it https://www.evidentlyai.com/ml-in-production/concept-drift (приступано: мај 2024)

[13] Fawcett, T. (2006). An introduction to ROC analysis. Pattern recognition letters, 27(8), 861-874.

[14] Gama, J., Medas, P., Castillo, G., & Rodrigues, P. (2004). Learning with drift detection. In Advances in Artificial Intelligence–SBIA 2004: 17th Brazilian Symposium on Artificial Intelligence, Sao Luis, Maranhao, Brazil, September 29-Ocotber 1, 2004. Proceedings 17 (pp. 286-295). Springer Berlin Heidelberg.

[15] Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. ACM computing surveys (CSUR), 46(4), 1-37.

[16] Gomes, H. M., Bifet, A., Read, J., Barddal, J. P., Enembreck, F., Pfharinger, B., ... & Abdessalem, T. (2017). Adaptive random forests for evolving data stream classification. Machine Learning, 106, 1469-1495.

[17] Imozaik author. (2019). Box plot. https://imozaik.wordpress.com/2019/02/08/box-plot (приступано: јул 2024)

[18] Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., & Zhang, G. (2018). Learning under concept drift: A review. IEEE transactions on knowledge and data engineering, 31(12), 2346-2363.

[19] Nvidia. (n.d.) XGBoost. https://www.nvidia.com/en-us/glossary/xgboost (приступано: јул 2024)

[20] SAS Institute Inc. (n.d.). Model Performance Measures and Statistics. https://documentation.sas.com/doc/en/edmcdc/3.3/edmug/p0ljaafkbyw38tn14ojsct0ksnwy.htm (приступано: јул 2024)

[21] Suknović, M., & Delibašić, B. (2010). Poslovna inteligencija i sistemi za podršku odlučivanju. FON, Beograd.

[22] Widmer, G., & Kubat, M. (1996). Learning in the presence of concept drift and hidden contexts. Machine learning, 23, 69-101.

[23] Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. In Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining (vol. 1, pp. 29-39).

[24] XGBoost developers. (2022.) XGBoost Parameters. (приступано: јул 2024) https://xgboost.readthedocs.io/en/stable/parameter.html

Анђела Ристивојевић, Senior Data Scientist у Banca Intesa ad Beograd

Контакт: andjelaristivojevicr@gmail.com

Области интересовања: машинско инжењерство, data science