Ален Калач

УНАПРЕДУВАЊЕ НА ПРОЦЕСОТ НА КРЕДИТНА АНАЛИЗА КАЈ БАНКИТЕ ПРЕКУ УПОТРЕБА НА МЕТОДИ ОД МАШИНСКО УЧЕЊЕ

Скопје, 2020 година

Апстракт

Зголемениот обем на кредитирање создава потреба од прецизна процена на кредитниот ризик. Поради тоа, финансиските институции чии активности го опфаќаат кредитирањето се во континуирана потрага по што поуспешни методи за кредитно оценување. Во овој труд, направен е осврт врз управувањето со кредитниот ризик во македонскиот банкарски систем, а потоа се истражени можностите за примена на методи од вештачката интелигенција, поконкретно машинското учење, во процесите на кредитна анализа и кредитно оценување. Тестирани се перформансите на релевантни модели на класификација од машинското учење при проценувањето на веројатноста за отплата на даден кредит. Добиените резултати укажуваат дека методите на машинско учење, пред се модерните методи базирани на нагласени дрва постигнуваат задоволителни перформанси при класификацијата на кредитните барања. Линеарните модели, иако претставуваат соодветен пристап при оценувањето на кредитниот ризик, сепак заостануваат зад методите базирани на нагласени дрва.

Клучни зборови: кредитен ризик, кредитна анализа, кредитно оценување, машинско учење, вештачка интелигенција, модели за класификација, нагласени дрва

JEL класификација: C45, C61, C63, G21, G32

Содржина

Вовед 4
1. Преглед на емпириската литература 5
2. Осврт врз управувањето со кредитен ризик во македонскиот банкарски сектор 10
2.1. Регулаторна рамка во областа на кредитниот ризик 10
2.2. Традиционални методи на кредитна анализа 11
2.3. Кредитниот ризик во македонскиот банкарски сектор 13
3. Спецификација на користените модели 16
3.1. Модели за класификација 17
3.1.1. Линеарни модели 17
3.1.2. Алгоритми базирани на најблиски соседи 19
3.1.3. Наивен Баесов алгоритам 19
3.1.4. Машини со носечки вектори (SVM) 19
3.1.5. Модели базирани на дрва 20
3.1.6. Вештачки невронски мрежи 21
3.1.7. Други ансамбли од класификатори 23
3.2. Показатели за перформансите на моделите за класификација 23
4. Преглед на користените променливи и податоци 25
5. Анализа на резултати 28
Заклучни согледувања 35
Библиографија 36
Прилог 1. Опис на поважни независни променливи 39
Прилог 2. Дескриптивна статистика на поважните нумерички варијабли и корелациона топлотна карта 40
Прилог 3. ROC-криви на избрани модели за класификација 42
Прилог 4. Граници на одлучување на избрани модели за класификација 44
Прилог 5. Матрици на конфузија на моделите за класификација 46

Вовед

Процесот на кредитна анализа и управувањето со кредитниот ризик се од исклучително значење за финансиските институции чија основна дејност е кредитирањето. Високите трошоци кои произлегуваат од нефункционалните кредити создаваат потреба за што попрецизна процена на кредитниот ризик поврзан со одредено кредитно барање. Во таа насока, финансиските институции употребуваат разни статистички техники преку кои ја оценуваат кредитоспособноста на кредитобарателот. Преку користење на овие техники финансиските институции се стремат кон зголемување на обемот на дадени кредити при истовремено минимизирање на загубите од ненаплатени кредити. Со оглед на се повисокиот износ на кредитните портфолија, дури и најситни подобрувања на процесот на кредитно оценување можат да придонесат кон намалување на кредитниот ризик и значителни заштеди за овие институции (West, 2000).

Исклучителниот раст на обемот на кредитирање во последните децении резултираше со создавање и унапредување на бројни квантитативни методи насочени кон подобрување на кредитното оценување (Crook et al., 2007), а обидите за развивање на нови и подобри методи траат континуирано. Ова е особено изразено во услови на брзи промени во индустријата. Финансиските институции се соочуваат со брз раст на кредитирањето и на кредитните портфолија, значително зголемени обем на податоци и пресметковна моќ, како и се поголема употреба на нови технологии во оваа сфера. Поради сето горенаведено, традиционалните техники на кредитната анализа постепено застаруваат, а се создава потреба од софистицирани методи кои можат посоодветно да одговорат на новите предизвици. Во такви услови, вештачката интелигенција со својот огромен потенцијал се издвојува како соодветна алтернатива за адресирање на предизвиците со кои се соочува кредитната индустрија.

Иако финансиските институции се почесто користат методи од вештачката интелигенција во своето работење, нивниот потенцијал е далеку од целосно искористен. Направени се голем број емпириски истражувања фокусирани на употребата на вештачката интелигенција, поконкретно машинското учење како нејзин интегрален дел, во процесот на управување со кредитен ризик. Истражувањата сугерираат на супериорноста на овие техники во однос на традиционалните (Huang et al., 2004; Ong et al., 2005; Crook et al., 2007). Сепак, најголем дел од истражувањата се изработени со користење на мали и субоптимални множества на податоци, како резултат на

недостапноста на квалитетни и обемни податоци. Ова секако негативно влијае врз веродостојноста на нивните резултати.

Новите трендови во кредитната индустрија овозможија значително олеснет пристап до квалитетни множества на податоци, што го отвора патот кон нови и поверодостојни истражувања. Токму ова е мотивацијата за изработка на овој труд. Достапноста на квалитетни сетови на микро-податоци овозможува релевантно тестирање на софистицирани методи од машинското учење, насочено кон изнаоѓање и идентификување на модели кои што попрецизно ќе ги проценат квалитетот на кредитното барање и веројатноста за отплата на кредитот доколку е одобрен. Во овој труд ќе бидат тестирани релевантните модели за класификација кои денес се употребуваат во машинското учење, со цел евалуирање и споредба на нивните перформанси при проценатa на кредитен ризик. Целта на истражувањето е преку добиените резултати и наоди да се идентификува насоката во која треба да бидат вложени напорите на македонските банки и останатите кредитни институции за подобрување на процесот на кредитна анализа.

Трудот е организиран во 5 дела. Во Дел 1 ќе биде разгледана постоечката емпириска литература од оваа област, главните согледувања во неа, но и главните недостатоци кои ја карактеризираат. Овој труд, меѓу другото, ќе биде фокусиран на надминување на тие недостатоци. Во Дел 2 е направен осврт кон кредитниот ризик во македонската економија, кој се состои од преглед на регулаторната рамка, традиционалните методи на кредитна анализа, како и експлоративна анализа на движењата во земјата во оваа област. Во Дел 3 се опишани моделите чии перформанси ќе бидат оценети и споредувани, како и показателите преку кои ќе се мерат перформансите. Во Дел 4 е претставено множеството на податоци кои ќе се користат, независните променливи и начинот на претпроцесирање на податоците. Конечно, во Дел 5 се анализирани добиените резултати, претставени се перформансите на различните модели, а извршена е и компаративна анализа помеѓу нив.

1. Преглед на емпириската литература

Емпириската литература на полето на примена на машинско учење и вештачка интелигенција за процена на кредитниот ризик иако релативно обемна, главно е фокусирана на споредба на перформансите на мал број индивидуални алгоритми во

однос на традиционалните методи на кредитна анализа, а нешто помалку на меѓусебна споредба на алгоритмите. Дополнителен проблем претставува исклучително брзиот технолошки развој во оваа област и континуираното подобрување на постоечките и создавање нови методи и алгоритми, поради што најголем дел од направените истражувања се застарени и оставаат простор за значително подобрување. Овие трендови создаваат потреба за постојано надополнување на емпириската литература. Сепак, главниот недостаток на постоечката литература е последица на недостапноста на податоци, со оглед на тоа што за овој вид истражувања се потребни податоци по поединечни кредити, поради што најголем дел од истражувањата се направени врз основа на мали и субоптимални множества на податоци кои не ја отсликуваат доволно добро реалната ситуација. И покрај одредените недостатоци, во разгледуваната литература речиси и да постои консензус дека методите за кредитна анализа базирани на машинско учење се супериорни во однос на традиционалните методи.

Како што е споменато погоре, емпириската литература во најголем дел е фокусирана на истражување на перформансите на мал број поединечни методи. Притоа, најиспитувани се релативно постарите методи. Според обемот на истражувања, предничат вештачките невронски мрежи (Galindo и Tamayo, 2000; Angelini et al., 2008; Rafiei et al., 2011; Oreski et al., 2012; Blanco et al., 2013), SVM (Huang, 2007; Bellotti и Crook, 2009; Cao et al., 2013; Harris, 2013; Yao et al., 2017), но и релативно понови методи базирани на дрва за одлучување или ансамбли од класификатори (Bastos, 2008; Khandani, 2010; Zhang, 2017; Addo et al., 2018; Hamori et al., 2018).

Hamori et al. (2018) во својата анализа имплементираат нагласени (англ. *boosted*) модели базирани на дрва, кои се во широка употреба генерално во машинското учење. Во овој труд се споредува точноста на 11 методи на машинско учење при класифицирање на кредитобарателите. Притоа, 3 од имплементираните методи претставуваат ансамбли од класификатори, и тоа случајни шуми (англ. *Random Forest*), „собирање“ (англ. *bagging*) и нагласување (англ. *boosting*), додека останатите 8 се различни имплементации на невронски мрежи. Перформансите на различните методи при проценувањето на кредитниот ризик се мерат преку повеќе индикатори соодветни за дадената проблематика. Добиените резултати покажуваат дека способноста за класификација на нагласените дрва е супериорна во однос на останатите методи, вклучувајќи ги и невронските мрежи, кои во бројни претходни истражувања постигнуваа најдобри перформанси. Нагласувањето дава подобри резултати кај сите

главни индикатори за споредба. Ова истражување сугерира дека нагласените дрва кои генерално се во подем можат да се користат и во финансиската индустрија во процесот на кредитна анализа. Иако наодите на трудот се корисни, постои простор за подобрување преку имплементирање на поголем број на методи со различни карактеристики, испитување на овие методи врз небалансирано множество на податоци, користење на множество на податоци со поголем број на опсервации и имплементирање на уште понови и пософистицирани модели.

Друго истражување кое имплементира некои од најсофистицираните алгоритми од машинското учење е тоа на Petropoulos et al. (2018). Ова истражување е насочено кон подобрување на ефикасноста на супервизијата на кредитниот ризик од страна на грчката централна банка. За потребите на трудот користени се податоци по поединечни кредити во грчкиот банкарски сектор од датабазите на грчката централна банка. При анализата, имплементирани се длабоки невронски мрежи и методот на нагласени дрва XGBoost, а нивните перформанси се споредени со традиционални методи како логистичка регресија и линеарна дискриминантна анализа. Очекувано, модерните алгоритми покажуваат значително подобри резултати при класифицирањето на кредитобарателите. За споредба, имплементирани се и методите на случајни шуми и плитки (англ. *shallow*) невронски мрежи, чии перформанси се послаби од XGBoost и длабоките невронски мрежи, но сепак подобри од традиционалните методи.

Уште еден понов труд е тој на Addo et al. (2018), во кој авторите обучуваат 7 бинарни класификатори базирани на машинско и длабоко учење (англ. *deep learning*) со цел процена на веројатноста за неврaќање на кредитот. 4 од овие класификатори се модели на длабоко учење, надополнети со модел на случајни шуми, модел на градиентно нагласување (англ. *gradient boosting*), како и модел на еластична мрежа (англ. *elastic net*) кој се користи како бенчмарк. При евалуирањето на моделите се користат критериумите AUC (*area under the curve*) и RMSE (*root mean squared error*). Потоа, авторите ги тестираат моделите на одделно множество податоци користејќи ги 10-те најважни атрибути добиени по првичното обучување. Во сите сценарија, моделите базирани на дрва (случајни шуми и градиентно нагласување) покажуваат најдобри перформанси. Овој труд го потенцира значењето на изборот на алгоритми, параметри и релевантни независни варијабли, улогата на критериумите за евалуација, но и исклучителната важност на човечкиот фактор при донесување на конечна одлука.

Bastos (2008) исто така ги тестира перформансите на неколку модели при оценувањето на кредитниот ризик, при што моделот базиран на нагласени дрва покажува подобри резултати во однос на невронската мрежа и SVM. Освен ова, нагласените дрва овозможуваат едноставно рангирање на најважните атрибути при класифицирањето, односно на атрибутите кои се најтесно поврзани со веројатноста за неврaќање на дадениот кредит. Недостаток на овој труд, како и на голем дел од останатите трудови во областа е користењето на множества на податоци со мал број на опсервации за машинско учење (1.000 и помалку), поради што произлезените заклучоци треба да се земат со резерва.

Tsai и Chen (2010) тестираат повеќе хибридни модели за класификација или кластерирање со цел оценување на кредитниот ризик. Овој пристап се состои од комбинирање на повеќе модели во насока на добивање резултат подобар од резултатот на индивидуалните модели. Иако хибридните методи се често користени генерално во машинското учење, овој труд е помеѓу првите кој споредува повеќе различни хибридни модели за оценување на кредитен ризик. Обучени се 4 различни хибридни модели, при што најдобри резултати покажува хибридниот модел кој претставува комбинација на логистичка регресија и невронски мрежи. Голем недостаток на истражувањето, препознаен и од авторите е невклучувањето на софистицирани хибридни техники како што се ансамблите на класификатори и каскадирање (англ. *stacking*), чија имплементација бара за тоа време значителни пресметковни напори. Со оглед на технолошкиот развој во областа, овие супериорни методи денес се полесно изводливи и се имплементирани во рамки на овој труд.

Една од најсеопфатните компаративни анализи на различните класификатори е изработена од Lessmann et al. (2015). Во трудот се анализираат бројни претходни истражувања на ова поле и се идентификуваат нивните главни недостатоци. Анализата покажува дека претходните трудови користат мал број на множества на податоци со мал број на опсервации и независни променливи, најчесто употребуваат еден индикатор за перформансите на моделите и обучуваат мал број на слични модели. Дополнително, само две студии тестираат ансамбли на класификатори. Lessmann et al. ги адресираат овие недостатоци. Во овој труд се споредуваат 41 метод на класификација, при што некои се прв пат употребени во контекст на кредитното оценување. Класификаторите се обучени врз 8 различни множества на податоци, а евалуирањето е направено преку 6 различни споредбени индикатори. Истражувањето покажува дека бројни класификатори

го проценуваат кредитниот ризик подобро од традиционалните методи, при што најдобри перформанси покажуваат хетерогените ансамбли на класификатори.

За разлика од останатите истражувања фокусирани на емпириска анализа, Bazarbash (2019) го анализира потенцијалот на финансиските технологии (Финтек) за кредитирање во економиите во развој. Финтек кредитирањето е ветувачко решение за зголемување на финансиската инклузија. Притоа, една од главните предности на финтек кредитирањето е употребата на техники од машинското учење за кредитно оценување. Во овој труд се истражени главните предизвици во оценувањето на кредитниот ризик и анализирани се начините преку кои методи на машинско учење можат да бидат применети за надминување на овие предизвици. Истражени се и главните предности и слабости на користењето на финтек и машинско учење во доменот на кредитниот ризик и зголемувањето на финансиската инклузија.

Емпириската литература во оваа област е релативно богата. Сепак, најголем дел од објавените трудови се соочуваат со одредени недостатоци. Недостапноста на обемно множество на податоци со голем број на опсервации и независни варијабли е сериозен проблем кај постарите трудови, којшто е адресиран во овој труд благодарение на зголемувањето на достапноста на богати множества на податоци во последните години. Друг проблем е и малиот број на класификатори и споредбени показатели кои се користат, како и отсуството на одредени важни видови класификатори, како што се модерните ансамбли од класификатори. Овој проблем е исто така адресиран преку користење на различни класификатори од повеќе хетерогени групи и имплементирање на софистицирани ансамбли од класификатори. Сепак, главната потреба од надополнување на постоечката емпириска литература произлегува од брзиот развој на нови и поквалитетни техники и алгоритми. Во овој труд, освен стандардните класификатори, имплементирани се и најнови алгоритми кои малку или воопшто не се застапени во постоечката литература. Дополнително, во овој труд се земени во предвид карактеристиките на македонскиот банкарски сектор, така што претпроцесирањето на истражуваното множество на податоци и имплементирањето на алгоритмите се извршени на начин да се лесно применливи од страна на банките во земјата.

2. Осврт врз управувањето со кредитен ризик во македонскиот банкарски сектор

2.1. Регулаторна рамка во областа на кредитниот ризик

НБРСМ има централна улога во регулирањето на управувањето со кредитниот ризик на банките во земјава. Клучниот акт во регулацијата на кредитниот ризик претставува Одлуката за методологијата за управување со кредитниот ризик, која ги регулира класификацијата, начинот на утврдување на исправка на вредноста и износот на посебната резерва, супервизорските стандарди за пристигнатите за наплата а ненаплатени побарувања и опфатот и елементите на управувањето со кредитниот ризик. Одлуката ги дефинира поимите кредитен ризик, изложеност на кредитниот ризик, очекувана кредитна загуба, стапка на веројатност за ненаплата, реструктурирање на кредитна изложеност, нефункционална кредитна изложеност и други. Согласно Одлуката, за нефункционална кредитна изложеност се подразбира кредитната изложеност којашто по која било основа не е наплатена подолго од 90 дена од рокот на достасување или кредитната изложеност за која е утврдено дека клиентот нема да може да ги намири своите обврски кон банката.

Банките во македонскиот банкарски сектор се должни најмалку еднаш месечно да извршат класификација на кредитната изложеност по поединечен договор, врз основа на кредитната способност на клиентот, квалитетот на проектот и уредноста во намирувањето на обврските. Кредитната изложеност се класифицира во категориите на ризик „А“, „Б“, „В“, „Г“ и „Д“, при што најсигурните побарувања се класифицирани во категоријата „А“, а најризичните во категоријата „Д“.

Врз основа на утврдената очекувана кредитна загуба, банката најмалку на месечна основа утврдува исправка на вредноста и посебна резерва за кредитните изложености. Утврдувањето на очекуваната кредитна загуба се врши на поединечна и/или групна основа. Износот на издвоената посебна резерва зависи од категоријата на ризик, така што за категоријата на ризик „А“ е најниска и изнесува до 5%, додека за категоријата „Д“ е највисока и изнесува над 70%.

Согласно Одлуката за методологијата за управување со кредитниот ризик, банките не смеат да вршат директно или индиректно одобрување нова кредитна изложеност за затворање постоечка кредитна изложеност, освен при реструктурирање на кредитот. При менување на условите на кредитот, банките се должни да изработат

анализа на клиентот. Доколку со анализата се утврди влошена финансиска состојба или сигнали за влошување на финансиската состојба на клиентот, банката врши реструктурирање на кредитот. Во случај банката да не очекува наплата на кредитната изложеност, врши целосен или делумен отпис на кредитот. Овие кредитни изложености банките ги пренесуваат на сметките за вонбилансна евиденција.

Банките се должни да воспостават и применат Политика за управување со кредитниот ризик или друг интерен акт од областа на кредитниот ризик, кој меѓу другото ќе содржи критериуми за одобрување кредити и други облици на кредитна изложеност. Методите и техниките истражени во овој труд можат значително да го подобрат управувањето со кредитниот ризик кај банките, кои ќе придонесат кон поквалитетно донесување на одлуки по кредитни барања, но и подобрување на останатите фази во процесот на управување со кредитниот ризик, регулирани со Одлуката за методологијата за управување со кредитниот ризик.

2.2. Традиционални методи на кредитна анализа

Анализата на кредитното барање и оценувањето на кредитната способност на клиентот отсекогаш претставувале критичен момент во процесот на кредитирање. Поради тоа, континуирано се прават обиди за подобрување на овој процес, како во светот, така и во македонскиот банкарски сектор. Трендовите во кредитната анализа се насочени кон се поголема употреба на статистички техники во овој процес, вклучувајќи и методи на машинско учење и вештачка интелигенција. Познавањето на потрадиционалните техники за кредитна анализа и нивните слабости ќе придонесе за подобро разбирање на потребата и корисноста од воведувањето на модерни и софистицирани техники силно засновани на статистичкиот пристап.

Без разлика на начинот на кој се врши анализата на кредитно барање, од голема важност се податоците на кои се заснова конечната одлука по барањето. Тука клучна улога имаат информациите, документите и извештаите што ги обезбедува кредитобарателот на барање на банката. Покрај овие, на македонските банки им се достапни податоци од институции специјализирани за прибирање информации за кредитната способност на физичките и правните лица. Главни институции од ваков тип во земјава се Кредитниот регистар на НБРСМ и Македонското кредитно биро (МКБ). Кредитниот регистар на НБРСМ прибира податоци за изложеноста на кредитен ризик на правните и физичките лица спрема банките и штедилниците. При оценката на

кредитната способност на кредитобарателот или постоен корисник на кредит, банките и штедилниците можат да ги употребуваат податоците од Кредитниот регистар. МКБ дава информации за задолженоста и редовноста во подмирување на обврските на правните и физичките лица.

По поднесувањето на кредитното барање и обезбедувањето податоци, следува обработката на кредитното барање. Еден од најраспространетите традиционални методи на кредитна анализа претставува методот 5Ц (*5Cs*), кој анализира 5 различни области. Првата од нив е карактерот, поим кој ги опфаќа аспектите кои се поврзани со намерата на клиентот да го употреби кредитот согласно неговата намена и неговата подготвеност за отплата на кредитот. Важни извори за процена на карактерот на клиентот се почетните разговори со него, неговата кредитна историја, искуствата на други банки во соработката со него и други. Второто Ц го означува капацитетот, односно финансиската состојба и способноста на кредитобарателот да го отплати кредитот. При проценатa на капацитетот на правно лице се користат финансиските извештаи на кредитобарателот. Преку финансиските извештаи се добиваат информации и за капиталот, кој ја претставува способноста на кредитобарателот да го поднесе финансискиот товар од отплатата на кредитот. Важен аспект на методот 5Ц е колатералот, односно средството кое банката може да го преземе за да го наплати побарувањето во случај кредитокорисникот да не може уредно да го врати земениот кредит. Последен аспект на овој метод се кредитните услови, кои ги опфаќаат надворешните фактори кои можат да влијаат на отплатата на кредитот, но на кои кредитобарателот не може да влијае.

При кредитната анализа на правните лица, од голема важност е финансиската анализа, која подразбира длабинска анализа на финансиските извештаи на претпријатието. Во рамки на финансиската анализа, се разгледуваат пред се билансот на состојба, билансот на успех и извештајот за готовински текови. Притоа, освен износите непосредно искажани во овие извештаи, значајна е пресметката на финансиски показатели кои се добиваат од вредностите на ставките во финансиските извештаи. Показатели кои е важно да се анализираат се задолженоста на субјектот, покриеноста на каматните расходи, приносот на средства (*ROA*), приносот на капитал (*ROE*), профитната маржа, ликвидноста, обртот на средствата, просечниот период на наплата на побарувањата, како и било кои показатели кои според банката се важни индикатори за способноста на клиентот да го врати бараниот кредит.

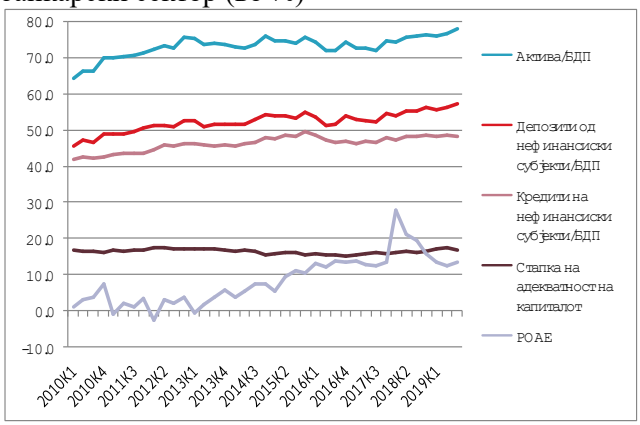
Традиционален метод на квантитативна кредитна анализа претставува кредитното бодување. Преку овој метод, на кредитобарателот му се доделува одреден број на поени за сет од претходно одредени атрибути и карактеристики. При агрегирањето на овие поени секој атрибут има пондер, односно тежина во определувањето на конечниот број на бодови. Токму начинот на определување на овие пондери има клучно влијание во прецизноста на методот на кредитно бодување. Меѓу останатото, резултатите добиени од ова истражување се фокусирани и на одредување на овие пондери за анализираното множество на податоци. Врз основа на вкупниот број на бодови и на поставениот минимален праг за одобрување на кредитот, банката одлучува за прифаќање или одбивање на кредитното барање.

Со оглед на квантитативниот пристап, овој метод има повеќе сличности со пристапот на машинското учење кон кредитната анализа во споредба со претходно споменатите. Главната сличност е што и двата метода користат историски податоци за претходно одобрени кредити, и со примена на математички техники врз овие податоци се обидуваат да ги идентификуваат карактеристиките на кредитобарателите кои се клучни за враќањето или невраќањето на кредитот. Сепак, софистицираните математички методи на машинското учење се далеку понапредни во однос на традиционалното кредитно бодување.

2.3. Кредитниот ризик во македонскиот банкарски сектор

Кредитниот ризик има централна улога кај македонските банки, со оглед на тоа што е блиску поврзан со нивната главна активност, кредитирањето на домашните претпријатија и домаќинства.

Графикон 2.1. Показатели за македонскиот банкарски сектор (во %)



Извор: НБРСМ

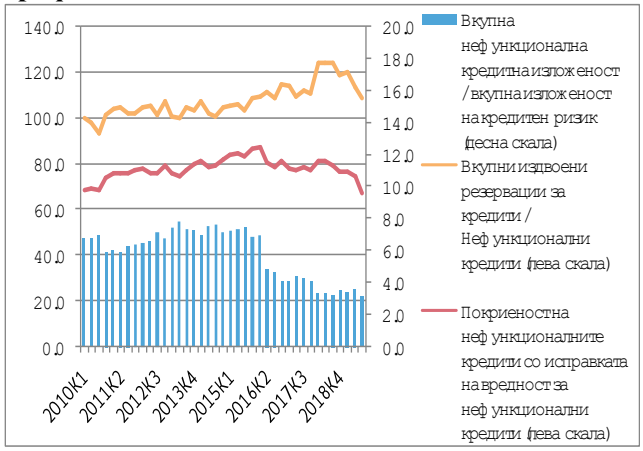
Во периодот до почетните години на XXI век, во услови на големи структурни промени во економијата и општеството, кредитниот пазар беше неразвиен и главно фокусиран на претпријатијата. Нивоата на нефункционални кредити беа исклучително високи. На почетокот од новиот век, како резултат меѓу

другото и на направените реформи на финансискиот систем и трансформацијата на сопственичката структура на банките, учеството на нефункционалните кредити во вкупните кредити започна позначително да се намалува. Истовремено е забележан и висок раст на вкупните кредити кој достигна ниво од околу 40%. Со појавувањето на глобалната финансиска криза, кредитниот раст значително забави, а нефункционалните кредити почнаа да се зголемуваат. Сепак, во споредба со останатите европски и земји од регионот, растот на нефункционалните кредити беше побавен, како резултат на конзервативноста на македонските банки при кредитирањето, претпазливоста при сметководственото признавање на загубите за кредитен ризик, солидните макроекономски движења итн. Стапката на нефункционални кредити во пост-кризниот период достигна ниво од над 12%, по што започна да се намалува достигнувајќи рекордно ниски 5,0%. Ваквиот пад во голема мера е резултат на обврската за задолжителен отпис на нефункционалните побарувања на банките кои се целосно покриени со исправка на вредноста подолго од две години, воведена во 2016 година.

Графикон 2.2. Вкупни и нефункционални кредити

Извор: НБРСМ

Графикон 2.3. Показатели за квалитет на активата

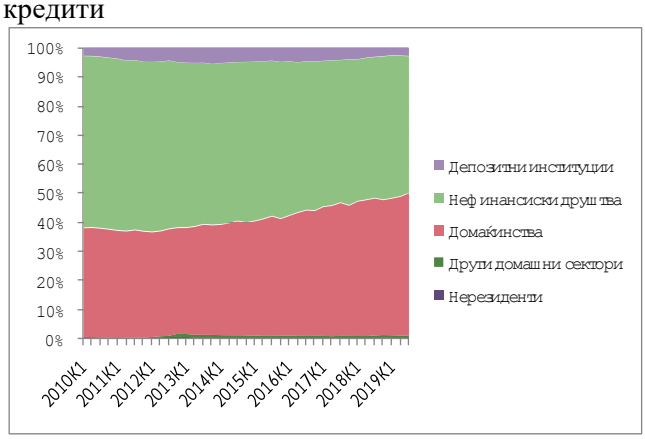


Извор: НБРСМ

На крајот на третиот квартал од 2019 година, нефункционалните кредити изнесуваа 5,0% од вкупните кредити кон нефинансиските субјекти, при што се намалија за 11,1% или 0,5 п.п. на квартална основа. Кај домаќинствата, учеството на нефункционалните во вкупните кредити изнесува 2,1%, додека кај нефинансиските друштва изнесува 8,0%. Значителниот пад во најголем дел е резултат на стапувањето на сила на одредбите од новата Одлука за методологијата за управување со кредитниот ризик со кои се намалува периодот за задолжителен отпис на целосно

резервираните изложености од две на една година. Доколку се изолира ефектот од отписите, нефункционалните кредити се зголемени за 9,1% во однос на претходниот квартал, како резултат на растот на нефункционалните кредити на домаќинствата. Сепак, ефектот од оваа регулаторна промена е далеку помал во споредба со регулаторните измени од 2016 година.

Графикон 2.4. Секторска структура на вкупните кредити

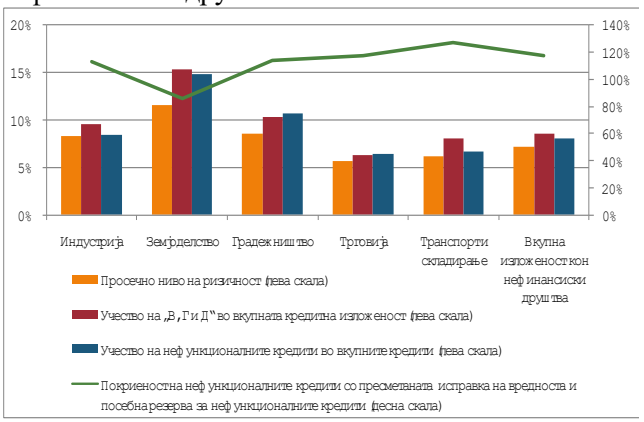


Извор: НБРСМ

Гледано по групи на банки, стапката на нефункционални кредити е највисока кај средните банки (5,7%), додека најниска е кај големите банки (4,7%). Според валутната структура, кај денарските кредити се забележува највисока стапка на нефункционални кредити (6,4%), кај девизните кредите таа изнесува 3,9%, додека најниска е кај денарските кредити со девизна клаузула (2,6%).

Покриеноста на нефункционалните кредити со исправката на вредноста за нефункционални кредити на крајот на К-3 2019 година изнесува 66,9%. Притоа, највисока е кај големите банки (73,5%), кај малите изнесува 67,2%, додека кај средните банки е најниска и изнесува 49,2%. Оваа стапка е значително намалена во однос на изминатиот период, главно како последица на задолжителните отписи. И покрај тоа, покриеноста на нефункционални кредити е висока, така што негативните ефекти од нивната евентуалната ненаплатливост не би ја загрозило значително солвентноста на банкарскиот систем. Покриеноста е повисока кај нефинансиските друштва (68,1%) во однос на домаќинствата (62,5%).

Графикон 2.5. Показатели за секторот нефинансиски друштва

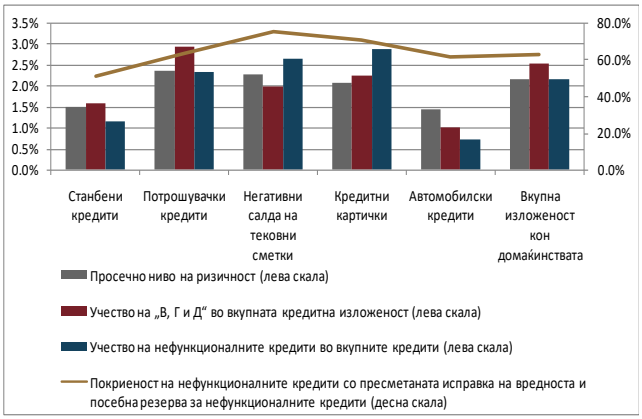


Извор: НБРСМ

Износот на вкупните издвоени резервации за кредити е повисок од износот на нефункционалните кредити, и во третиот квартал од 2019 година изнесува 108,8% од

нивната вредност. Како резултат на отписите се намали учеството на „В“, „Г“ и „Д“ во вкупната кредитна изложеност (4,1%), додека само учеството на „Д“ категоријата изнесува 1,7%.

Графикон 2.6. Показатели за секторот домаќинства



Извор: НБРСМ

Процентот на обезбеденост на кредитите е на задоволително ниво, при што кај претпријатијата овој процент е нешто под 100%. Поволен е и соодносот помеѓу износот на кредитот и проценетата вредност на обезбедувањето, кој изнесува речиси 50%. Обезбеденоста на кредитите на домаќинствата е пониска поради кредитните картички, потрошувачките кредити и негативните салда на тековните сметки, и тој изнесува речиси 70%.

3. Спецификација на користените модели

Во овој дел од трудот ќе бидат опишани користените модели и алгоритми, како и генералниот концепт на машинско учење, врз кое се базирани сите употребени модели. Со цел трудот да биде поразбирлив, фокусот ќе биде на поинтуитивно разгледување на моделите и нивна практична примена, додека описот на нивната математичка имплементација ќе биде сведен на неопходниот минимум.

Машинско учење е подмножество на вештачката интелигенција кое им дава на компјутерите способност самостојно да учат од обезбедено множество на податоци, без притоа системот да е експлицитно програмиран. Машинското учење употребува бројни алгоритми кои итеративно учат од дадените податоци, ги опишуваат податоците и предвидуваат исходи. Модел на машинско учење е добиениот резултат на обучување на алгоритам преку податоци. По обучувањето, овој модел е способен да прими податоци како инпут, да ги обработи и да даде одреден аутпут. Вештачката интелигенција и машинското учење датираат од 1950-тите, но вистинскиот пробив доаѓа со зголемувањето на моќта на компјутерите која овозможи нивна употреба при обработката

на податоци. Исклучително брзиот раст на обемот на достапни податоци исто така значително влијаеше на зголемување на популарноста на машинското учење.

Техниките на машинско учење во зависност од природата на проблемот и достапното множество на податоци можат да се поделат во неколку категории, при што важна е поделбата на надгледувано и ненадгледувано учење. Кај надгледуваното учење, множеството на податоци кое се користи за обучување на моделот освен влезните податоци го содржи и излезниот податок (англ. *label*). Доколку излезниот податок е категорична варијабла од одредено конечно множество, дадениот проблем претставува класификација. Од друга страна, кога излезниот податок е во форма на континуирана варијабла која може да изнесува било која вредност, дадениот проблем претставува регресија. Со оглед на тоа што целта на трудот е идентификување на квалитетот на кредитно барање, природата на проблемот ја наметнува класификацијата како пристап, при што двете категории во кои кредитното барање може да се класифицира се „добрите“ и „лошите“ кредитни барања.

Во овој труд ќе бидат имплементирани и споредени 19 модели за класификација од 7 различни групи, кои се образложени во продолжение. Дел од овие модели, како што се некои од линеарните модели имаат традиционална употреба во процесот на кредитна анализа. Од друга страна, моделите базирани на дрва, особено најновите алгоритми кои припаѓаат на оваа група се уште не се често применувани при проценувањето на веројатноста за враќање на кредит. Освен овие, ќе бидат разгледани и споредени неколку други групи на алгоритми, како што се методот на најблиски соседи, наивен Баесов класификатор, машините со носечки вектори (SVM), невронски мрежи, како и различни ансамбли на класификатори.

3.1. Модели за класификација

3.1.1. Линеарни модели

Линеарните модели класифицираат одредена опсервација врз основ на линеарна комбинација на нејзините атрибути. Во овој дел тестирани се 4 линеарни класификатори.

Логистичката регресија, и покрај тоа што го содржи зборот регресија во називот се користи за класификациски проблеми. Со овој алгоритам обично се проценува веројатноста за одредена опсервација да припаѓа на дадена класа, односно која е веројатноста одредено кредитно барање да стане нефункционален кредит. При

проценатa на веројатноста се применува логистичка трансформација, која го ограничува аутпутот од [-∞,+∞] на веројатност помеѓу 0 и 1. Формата на бинарната логистичка регресија е:

$$g(x) = \ln\left(\frac{p_g}{1-p_g}\right) = b_0 + b_1x_1 + \cdots + b_nx_n \quad (3.1)$$

каде што p_g е веројатноста опсервацијата да припаѓа на „добрата“ класа, $g(x)$ е логит трансформацијата, а b_0 до b_n се регресионите коефициенти добиени преку методот на проценка на максимална подобност (веројатност). Од (3.1) може да се пресмета веројатноста опсервацијата да припаѓа на „добрата“ класа:

$$p_g = \frac{e^{b_0+b_1x_1+\cdots+b_nx_n}}{1+e^{b_0+b_1x_1+\cdots+b_nx_n}} \quad (3.2)$$

Друг линеарен модел кој ќе биде имплементиран е **ridge регресијата**, која се добива така што при обучувањето на моделот се додава дополнителен член за регуларизација на трошочната функција на линеарната регресија. Трошочната функција на Ridge регресијата е дадена преку:

$$J(\theta) = MSE(\theta) + \alpha\frac{1}{2}\sum_{i=1}^{n}\theta_i^2 \quad (3.3)$$

каде што θ ги претставува векторите на атрибутите на множеството податоци, а α е хипер-параметар преку кој се контролира регуларизацијата на моделот (при $\alpha = 0$, регуларизираните модели се сведуваат на линеарна регресија). На овој начин се ограничуваат пондерите на моделот. Ridge регресијата ја намалува веројатноста за пренагодување (англ. *overfitting*) на моделот.

Линеарната дискриминантна анализа (ЛДА) е линеарен класификациски алгоритам, со кој при обучувањето моделот ги препознава дискриминативните оски помеѓу класите. Овие оски потоа се користат за дефинирање на хипер-рамнина на која се врши проекција на податоците, на која растојанието помеѓу класите ќе биде максимизирано.

Стохастичкото градиентно спуштање (англ. *Stochastic Gradient Descent - SGDClassifier*) е едноставен и ефикасен пристап за дискриминантно обучување. Овој пристап имлементира регуларизирани линеарни модели преку стохастичко градиентно спуштање, и е еднакво соодветен за мали и обемни линеарно сепарабилни множества на податоци. Регуларизацијата се врши со додавање член на трошочната функција кој ги намалува параметрите на моделот.

3.1.2. Алгоритми базирани на најблиски соседи

Од групата алгоритми базирани на најблиски соседи ќе биде имплементиран **KNN** алгоритамот (k-nearest neighbors). KNN ги класифицира опсервациите врз основа на класата на нејзините најблиски k соседи, каде што k е бројот на соседи врз чија основа се класифицира опсервацијата. Притоа, опсервацијата ќе биде класифицирана во класата во која се класифицирани најголем број од нејзините k соседи. Најблиските k соседи ќе бидат одредени брз база на Minkowski растојанието, согласно следното равенство:

$$d(x, y) = \left(\sum_{i=1}^{k}(|x_i - y_i|)^p\right)^{\frac{1}{p}} \tag{3.4}$$

каде што со x_i и y_iсе претставени координатите на точките x и y, додека p е параметар кој влијае на растојанието (при имплементацијата на моделот $p = 2$, со што Minkowski растојанието се сведува на Евклидово растојание).

3.1.3. Наивен Баесов алгоритам

Класификацијата преку алгоритамот на **наивен Баес** е базирана врз Баесовата теорема, која користејќи претходни веројатности ги пресметува веројатностите опсервацијата да припаѓа на дадените класи. Потоа, алгоритамот ја класифицира опсервацијата во класата со најголема веројатност. Преку Баесовата теорема се пресметува веројатноста одредена опсервација y да припаѓа на класа k при дадено множество на податоци X со атрибути $\widetilde{x_j}$ (3.5). Недостаток на алгоритамот е тоа што претпоставува дека атрибутите се меѓусебно независни.

$$P(y = k|X) = \frac{P(y = k)\prod_{j=1}^{p} P(\widetilde{x_j}|y = k)}{P(X)} \tag{3.5}$$

3.1.4. Машини со носечки вектори (SVM)

Пристапот на **машини со носечки вектори** (англ. *SVM – Support Vector Machines*), предложен од Vapnik (1995), користејќи ги опсервациите од множеството податоци за обука наоѓа хипер-рамнина во векторскиот простор која ги дели двете класи. Со оглед на тоа што совршено одвојување на класите во пракса е речиси секогаш неизводливо, целта на SVM е да ја пронајде хипер-рамнината која ја максимизира маргината помеѓу носечките вектори на двете класи, и истовремено ги минимизира грешките во класифицирањето. При имплементирањето се користат „меки“ маргини, со кои се дозволуваат грешки во класифицирањето, со цел да се избегне проблемот на пренагодување на моделот.

Во случаите кога податоците не се линеарно сепарабилни, односно кога одреден нелинеарен регион може попрецизно да ги одвои двете класи, SVM користи кернел функција (англ. *kernel function*) или кернел трик (англ. *kernel trick*), со чија помош ги мапира податоците во векторски простор со повеќе димензии. На овој начин податоците стануваат линеарно сепарабилни.

SVM се исклучително популарни и имаат бројни примени, вклучувајќи и за процена на кредитен ризик. Тие покажуваат одлични перформанси на повеќедимензионални множества на податоци, како и на множества со релативно мал број на опсервации. SVM имаат важна примена кај податоци кои не се линеарно сепарабилни. Во ова истражување имплементиран е SVM кој користи кернел со радијална основа. Одредени слабости на пристапот на машини со носечки вектори се потешкотиите при интерпретирање и разбирање на моделот, како и високата пресметковна интензивност на моделот во случај на обемни множества на податоци, што го прави моделот непрактичен во тие случаи.

3.1.5. Модели базирани на дрва

Моделите базирани на дрва се широка група модели чија основа е во **дрвата на одлучување**, како нивна наједноставна форма. Овие модели ги расчленуваат големите одлуки на серија едноставни прашања со кои се покриваат сите можни исходи, каде во хиерархиска структура подоцнежните решенија зависат од претходните и конечниот исход е базиран на сите претходно донесени одлуки. Дрвата за одлучување имаат форма на дијаграм, така што започнуваат од почетен јазол (англ. *node*), продолжуваат кон следните јазоли преку серија одлуки и завршуваат со финален јазол кој е резултат на сите претходно донесени одлуки. Овој пристап е едноставен, лесен за интерпретација и може да се користи и за класификација и за регресија. Сепак, голем недостаток на дрвата за одлучување е нивната нестабилност, со оглед на тоа што ситни флуктуации во податоците можат да резултираат во големи варијации при класификацијата. Поради тоа, развиени се бројни модели кои се базирани на дрвата за одлучување, но ги надминуваат нивните слабости.

Случајни шуми (англ. *Random Forest*) е модел базиран на дрва предложен од Breiman (2001). Оваа техника комбинира бројни дрва за одлучување, а конечниот резултат го пресметува преку упросечување на исходите (за регресиони проблеми) или преку идентификување на најчестиот исход (за класификациони проблеми) од

индивидуалните дрва. При формирањето на дрвата, моделот не ги користи сите независни варијабли, туку само нивен случајно избран примерок. На овој начин се спречува доминантна независна варијабла да има преголемо влијание врз крајниот резултат. Методот на случајни шуми е поимун на пренагодување, подобро генерализира и изработува поробустни проценки во однос на индивидуалните дрва. Меѓутоа, моделот е покомплициран и потежок за интерпретација и објаснување, посебно доколку содржи голем број на дрва и/или независни варијабли.

Иако методот на случајни шуми претставува значителен напредок во споредба со дрвата за одлучување, развиени се уште понапредни методи базирани на дрва, меѓу кои најважните се методите на **нагласување** (англ. *boosting*). Слично како и кај случајни шуми, преку нагласување се создава ансамбл на бројни дрва со послаба моќ за проценување со цел создавање на конечен класификатор со што подобра моќ за процена. Сепак, за разлика од алгоритамот на случајни шуми кој паралелно обучува и потоа комбинира меѓусебно независни дрва, кај нагласените алгоритми дрвата се меѓусебно зависни, односно се обучуваат секвенционално. Притоа, секое наредно обучено дрво ги адресира слабостите на претходното. По бројни формирани дрва, конечниот исход е пондериран просек на исходите на индивидуалните дрва.

Нагласените методи базирани на дрва за одлучување се помеѓу најуспешните класификациски методи општо во машинското учење. Освен тоа, кај овие методи се забележува исклучително брз развој, при што во последните години се развиени и унапредени бројни софистицирани алгоритми од оваа група, кои се малку или воопшто користени во литературата за процена на кредитен ризик. Поради тоа, при ова истражување имплементирани се повеќе сродни нагласени методи базирани на дрва, меѓу кои се **AdaBoost**, **XGBoost** (Chen и Guestrin (2016), претставува напредна имплементација на алгоритмите на градиентно нагласување, се карактеризира со зголемена ефикасност, флексибилност и прецизност), **Catboost** (претставен од страна на компанијата Yandex во 2017 година) и **LightGBM** (претставен од компанијата Microsoft во 2017 година).

3.1.6. Вештачки невронски мрежи

Невронските мрежи во литературата за процена на кредитниот ризик се често истражувани, при што нивната супериорност во однос на традиционалните методи на кредитна анализа е сугерирана во бројни наврати. Иако вештачките невронски мрежи се

присутни со децении, нивното значење и влијание драстично се зголемуваат како резултат на се поголемиот обем на достапни податоци за обучување на невронските мрежи, со оглед на тоа што нивните перформанси во однос на останатите техники од машинското учење се генерално пропорционални на обемот на податоци. Дополнително, растот на моќта на компјутерите овозможи обучување на големи невронски мрежи за релативно кратко време.

Стандардната невронска мрежа се состои од голем број едноставни и меѓусебно поврзани неврони (јазли), кои обработуваат податоци и продуцираат серија на активации. Едноставна невронска мрежа се состои од три дела, односно влезен слој, скриен слој или слоеви и излезен слој. Влезниот слој се состои од невроните кои ги содржат податоците преку кои се обучува невронската мрежа. Скриените слоеви примаат инпути од влезниот или од претходен скриен слој кои ги обработуваат и трансформираат преку избрана функција на активација (при имплементацијата на невронската мрежа, користена е ReLu функцијата за скриените слоеви и логистичката функција за излезниот слој). Аутпутот на неврон од скриениот слој се пресметува на следниот начин:

$$h_i = f\left(b_i + \sum_{j=1}^{M} W_{ij}x_j\right) \tag{3.6}$$

каде што f е функцијата на активација, b_i го содржи сигналот за пристрасност (англ. *bias*), W е матрица која ги содржи пондерите, W_{ij} е пондерот кој го поврзува инпутот j со скриениот неврон i. Бројот и големината на скриените слоеви ја одредуваат длабочината на невронската мрежа. Пристапот на длабоко учење се состои во додавање повеќе слоеви во невронската мрежа. Аутпутот на вештачката невронска мрежа е содржан во излезниот слој.

Главната предност на невронските мрежи е нивната флексибилност и способност да препознаат различни врски во множеството на податоци. Ова е особено важно при обемно множество податоци, карактеристично за неструктурирани податоци, кај кои класифицирањето од страна на невронските мрежи генерално е поточно во споредба со други алгоритми на машинско учење. Слабоста на овој модел произлегува од неможноста за интерпретирање и разбирање на моделот од страна на човек, со оглед на тоа што невронските мрежи се во групата на црна кутија модели (англ. *black-box model*).

3.1.7. Други ансамбли од класификатори

Базирано на идејата дека проценитe агрегирани од повеќе извори покажуваат подобри перформанси дури и од најдобрите индивидуални резултати, развиени се различни видови ансамбли од класификатори. Некои од овие, како што се случајни шуми и нагласени дрва се претходно споменати, но покрај нив имплементирани се и други видови ансамбли. При имплементирањето на овие методи, потребно е прво да се обучат повеќе индивидуални класификатори, за потоа преку користење на ансамблите да се направи обид за подобрување на перформансите на најдобриот од индивидуалните класификатори. Невклучувањето на ансамблите е еден од недостатоците на најголем дел од трудовите во оваа област. Во овој дел од трудот тестирани се гласачките класификатори (англ. *voting classifiers*) и каскадирањето (англ. *stacking*).

Гласачките класификатори се ансамбли кои комбинираат повеќе процени и едноставно го земаат просечниот исход. Ваквиот начин на агрегирање на проценитe може да резултира со подобри перформанси во однос на сите алгоритми користени во ансамблот, особено доколку се вклучени поголем број индивидуални класификатори. Врз ансамблот позитивно влијае и доколку користените класификатори припаѓаат на различни групи модели. На овој начин, нивните процени би биле повеќе диверзифицирани и независни, а грешките би биле помалку корелирани. Класификаторите користени при формирањето ансамбли можат да имаат и различни пондери во конечниот класификатор, со што на одреден модел може да му се даде поголемо влијание.

Нешто покомплексен начин на комбинирање модели е нивно **каскадирање**. Овој пристап комбинира повеќе класификатори и ги користи како атрибути при обучување на нов модел. Првиот чекор, како и кај гласачките класификатори е обучувањето на неколку класификатори. Потоа, обучените класификатори го предвидуваат исходот на нов примерок од множеството на податоци. Овие предвидувања се користат како инпут при обучувањето на конечниот класификатор. Перформансите на вака обучениот модел можат да ги надминат оние на првично обучените поединечни класификатори.

3.2. Показатели за перформансите на моделите за класификација

За да може да се споредат перформансите на обучените модели и да се идентификува најдобриот, неопходно е да се одреди конкретен и мерлив статистички показател кој објективно ќе ја прикаже успешноста на моделот. Постоењето на бројни

показатели дополнително го отежнува одредувањето на најсоодветниот за даден проблем. Во ова истражување моделите се тестирани и споредени според неколку показатели, како што се точноста, балансираната точност, матрицата на конфузија и показателите кои произлегуваат од неа, како и ROC-AUC резултатот.

Наједноставен показател за перформансите на класификациски модел е **точноста,** која се пресметува како бројот на точно предвидени опсервации во однос на вкупниот број опсервации. Предноста на овој показател е што е крајно едноставен, разбирлив и лесен за интерпретација и споредба. Сепак, точноста се карактеризира и со неколку големи недостатоци. Пред се, точноста не е релевантен показател кај множествата на податоци со небалансирани класи, односно кај проблеми во кои учеството на една класа е значително поголемо во однос на другата (или останатите). Со оглед на тоа што кај класифицирањето на кредитните барања генерално постои висока небалансираност помеѓу класите, така што бројот на отплатени кредити е обично далеку поголем од бројот на отпишаните, точноста кај овие проблеми сама по себе не е соодветен показател.

Друг голем недостаток на точноста е нејзината зависност од дискрециони одлуки на заемодавателите, односно од критериумите што тие ги поставиле за одобрување или одбивање на кредитното барање. Имено, моделите опишани во овој Дел ја проценуваат веројатноста за враќање на заемот, но од самиот заемодавател зависи која е најниската прифатлива веројатност за одобрување на заемот. И поради ова, перформансите на моделите не можат да се споредуваат само по точноста.

Со цел надминување на недостатокот на точноста при евалуирање на небалансираните множества на податоци, изведен е показателот **балансирана точност**. Овој показател претставува просек од точностите на двете класи. Кај податоци со совршено балансирани класи, вредностите на точноста и балансираната точност се идентични, но кај множествата на податоци со доминантна класа балансираната точност е посоодветен показател.

Други показатели поврзани со точноста кои ќе бидат анализирани се **прецизноста**, **чувствителноста** и **Ф1 резултатот**. Прецизноста го претставува процентот на точно предвидени опсервации од одредена класа во вкупниот број на предвидувања на класата (конкретно, процентот на одобрени „добри“ кредити во однос на вкупниот број на одобрени кредити). Чувствителноста го претставува учеството на

точно предвидени опсервации од одредена класа во вкупниот број на опсервации во таа класа (конкретно, учеството на одобрените „добри“ кредити во вкупниот број „добри“ кредити). Ф1 резултатот претставува хармониска средина на прецизноста и чувствителноста. И овие показатели, како и претходно споменатите, зависат од субјективните одлуки на одлучувачот. Доколку кредитодавателот има понизок критериум за одобрување кредити, тоа ќе ја зголеми чувствителноста, но истовремено може негативно да влијае врз прецизноста, и обратно. Поради зависноста од надворешни фактори, овие показатели сами по себе не можат да бидат објективна мерка за перформансите на моделите.

Последен показател кој ќе се користи за евалуација и споредба на моделите е **ROC-AUC** (Area under ROC curve) резултатот, кој ја претставува површината под ROC (Receiver Operating Characteristic) кривата. ROC - кривата графички ја прикажува успешноста на моделите за класификација, со користење на чувствителноста и лажната позитивна стапка (англ. *false positive rate*). Површината која ја зазема оваа крива се употребува како показател за перформансите на даден модел. Вредноста на ваквиот показател се движи помеѓу 0,5 (модел кој врши случајна класификација на опсервациите) и 1 (совршен модел). За разлика од претходните показатели, на ROC-AUC резултатот не влијаат балансираноста на множеството податоци и дискреционите одлуки на заемодавателот, што овој показател го прави најсоодветен за евалуација и споредување на обучените модели. Поради тоа, ROC-AUC е примарен статистички показател според кој ќе се споредуваат перформансите на моделите.

4. Преглед на користените променливи и податоци

За потребите на овој труд и тестирање на моделите опишани во Дел 3 неопходни се податоци за поединечни одобрени кредити. Множеството податоци може да содржи податоци за одобрениот кредит (износ, вид, обезбедување и сл.), за кредитобарателот (висина на приходи, вработување, возраст, кредитна историја и сл.), податоци за окружувањето доколку се сметаат за релевантни (макроекономски движења и сл.), статусот на кредитот (отплатен, нефункционален, тековен и сл.), како и било кои податоци за кои се смета дека имаат влијание на веројатноста за враќање на кредитот. Податоци од овој тип за македонскиот банкарски систем не се јавно достапни, но депозитните институции поседуваат обемни бази на податоци, како и Кредитниот

регистар на НБРСМ. Со помош на методите анализирани во овој труд, овие бази на податоци можат да се искористат за подобрување на процесот на кредитна анализа и попрецизно кредитно оценување кај депозитните институции, но и при извршувањето на супервизорската функција и анализата на кредитниот ризик од страна на НБРСМ.

Иако потребните податоци за македонскиот банкарски сектор од разбирливи причини не се јавно достапни, неколку глобални депозитни и останати финансиски институции дозволуваат слободен пристап до множества податоци за поединечни кредити. За потребите на овој труд, користени се податоци од американската финтек компанија Lending Club, која во последните години јавно објавува микро-податоци за дадените заеми. Ова множество податоци е избрано поради обемот (вкупното множество на достапни податоци содржи милиони опсервации), како и големиот број (стотици) на независни променливи (атрибути), од кои голем број се слични на податоците што ги поседуваат македонските банки. Токму оваа карактеристика на сетот на податоци на Lending Club овозможува прилагодување на дадените податоци на карактеристиките на македонскиот банкарски сектор, така што преку внимателна селекција на независни променливи може да се создаде множество на податоци слично, но сепак помало во однос на она кое го поседуваат банките во земјава и Кредитниот регистар. Дополнително, големиот број на опсервации овозможува поголема веродостојност на истражувањето и добиените резултати. Најголем дел од останатите достапни сетови на податоци се значително помалубројни и со помалку атрибути, што е и еден од главните проблеми со кои се соочувале авторите на разгледуваните истражувања од областа.

Целокупното обезбедено множество на податоци содржи околу 900.000 опсервации и 74 променливи за заеми одобрени во периодот од јуни 2007 година до декември 2015 година. Откако ќе се елиминираат променливи кои не се релевантни, не се достапни при оценување на кредитното барање или не се достапни на македонските банки и ќе се креираат вештачки (англ. *dummy)* варијабли за категоријалните променливи, остануваат 57 независни варијабли. Притоа, како најважни меѓу нив се издвојуваат годишните приходи на заемобарателот, рокот на отплата на кредитот, годината на поднесување на барањето (бидејќи разгледуваниот период е особено турбулентен, оваа варијабла е значајна), категоријата на ризик, работниот стаж, кредитната историја, показателот вкупен долг/вкупни приходи итн. Листа на најважните независни променливи е дадена во Прилог 1, додека дескриптивна статистика на нумеричките променливи е претставена во Прилог 2.

Со оглед на тоа што целта на истражувањето е класификација на кредитните барања на „добри“ и „лоши“, од вкупниот број на опсервации неопходно е да се исклучат тековните кредити. По бришењето на овие инстанци, конечното множество на податоци содржи 57 независни променливи и 252.971 опсервација. Зависната променлива која се предвидува е веројатноста за враќање на кредитот, како и статусот на кредитот, односно бинарната променлива која може да изнесува 1 („добар“ кредит) или 0 („лош“ кредит). Бројот на променливи и опсервации е значително повисок во однос на најголем дел од претходните истражувања од оваа област.

По гореопишаното првично претпроцесирање на податоците, неопходно е множеството на податоци да се подели на множество за обучување и множество за тестирање. Притоа, множеството за обучување ќе биде користено за обучување на моделите и пресметка на коефициентите, додека множеството за тестирање ќе се користи за независна евалуација на моделите. На овој начин моделите се тестираат на нови и за нив непознати податоци, што е единствен начин за добивање веродостојни резултати. Вкупниот сет на податоци ќе биде поделен така што 70% од податоците (177.079 опсервации) ќе се користат за обучување на моделите, додека нивните перформанси ќе бидат тестирани на преостанатите 30% од податоците (75.892 опсервации). Како исклучок, методите базирани на најблиски соседи и машините со носечки вектори се обучени и тестирани на помал обем на податоци, со оглед на тоа што нивно обучување на поголем обем на податоци е непрактично.

Дополнителен проблем кој се јавува при класифицирањето на кредитните барања е небалансираноста на категориите на зависната променлива. Имено, поради тоа што учеството на „добри“ кредити е значително повисоко од учеството на „лошите“ кредити кај вкупните одобрени кредити поради природата на проблемот, класификаторите имаат тенденција далеку почесто да ја предвидуваат позастапената категорија, доколку овој проблем не е соодветно адресиран. Поради тоа во множеството податоци за обучување се врши балансирање на категориите, така што вештачки се зголемува бројот на „лоши“ кредити додека не се изедначи со бројот на „добри“ кредити.

Последен чекор од подготовката на податоците претставува стандардизацијата на податоците, односно сведување на независните променливи на слична скала и блиску до нормална дистрибуција. Ова придонесува за подобрување на перформансите и зголемување на брзината на некои модели. Врз податоците ќе се изврши робустна стандардизација, со цел намалување на влијанието на отстапувањата (англ. *outliers*). При

овој вид на стандардизација, од вредноста на опсервациите за дадена променлива се одзема средната вредност на променливата и се дели со нејзиниот интерквартален ранг.

5. Анализа на резултати

Главната цел на трудот, пронаоѓањето на оптимален модел за процена на веројатноста за враќање на кредит е разработена во овој дел. Перформансите на моделите се мерат преку нивната способност да ги разделат „добрите“ и „лошите“ кредитни барања преку предвидената веројатност за враќање на кредитот. Анализата на ова прашање е заснована на моделите опишани во Дел 3.1 на овој труд и податоците опишани во Дел 4. Клучните резултати и сознанија од истражувањето се претставени во Табела 5.1.

Главен показател преку кој ќе се споредуваат моделите е ROC-AUC показателот. ROC-AUC е објективен статистички показател чија вредност не зависи од дискрециони одлуки на депозитните институции, што го прави идеален за споредба на различни модели. Според ROC-AUC, методите кои најдобро ја предвидуваат веројатноста за враќање на кредитот се методите базирани на дрва и ансамблите од методите базирани на дрва. Алгоритмот LightGBM постигна најдобри перформанси на даденото множество податоци за тестирање, со ROC-AUC вредност од 0,7180, нешто над сродните алгоритми CatBoost (0,7177) и XGBoost (0,7145). На графикон 5.1 е дадена ROC кривата на најуспешниот класификатор LightGBM, додека на Графикон

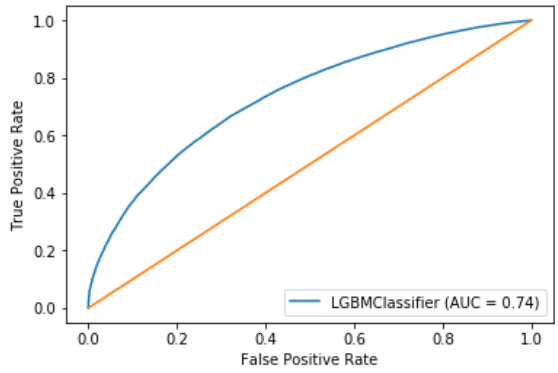
Табела 5.1. Резултати добиени преку обучување на моделите за класификација

Модел за класификација	ROC-AUC Резултат	Балансирана точност	Точност	Категорија „лоши" кредитни барања			Категорија „добри" кредитни барања		
				Прецизност	Чувствителност	Ф1 резултат	Прецизност	Чувствителност	Ф1 резултат
Линеарни модели									
Линеарна дискриминантна анализа	0,6988	0,6457	0,6351	0,2800	0,6622	0,3935	0,8953	0,6292	0,7390
Логистичка регресија	0,6929	0,6411	0,6406	0,2798	0,6418	0,3897	0,8914	0,6403	0,7453
SGD класификатор	0,6647	0,6233	0,6747	0,2851	0,5431	0,3739	0,8761	0,7034	0,7803
Ridge класификатор	0,6988	0,6447	0,6364	0,2800	0,6576	0,3927	0,8945	0,6318	0,7405
KNN	0,5998	0,5734	0,6285	0,2011	0,4946	0,2859	0,8794	0,6522	0,7489
Наивен Баесов алгоритам	0,6562	0,6213	0,6833	0,2882	0,5247	0,3720	0,8740	0,7179	0,7883
Машини со носечки вектори (SVC (rbf))	0,6461	0,6067	0,5713	0,2378	0,6609	0,3497	0,8852	0,5524	0,6803
Модели базирани на дрва									
Случајни шуми (RF)	0,6986	0,5443	0,8167	0,4527	0,1202	0,1900	0,8349	0,9684	0,8967
AdaBoost	0,7049	0,6499	0,6527	0,2890	0,6455	0,3993	0,8945	0,6543	0,7558
XGBoost	0,7145	0,6571	0,6661	0,2986	0,6432	0,4079	0,8963	0,6711	0,7675
CatBoost	0,7177	0,6591	0,6592	0,2963	0,6589	0,4087	0,8987	0,6593	0,7606
LightGBM	**0,7180**	**0,6598**	0,6534	0,2940	0,6697	0,4086	0,9004	0,6499	0,7549
Вештачки невронски мрежи									
MLPClassifier	0,6947	0,6400	0,6656	0,2898	0,6002	0,3909	0,8865	0,6798	0,7695
Длабока невронска мрежа	0,7013	0,6474	0,6593	0,2907	0,6289	0,3976	0,8918	0,6659	0,7625
Други ансамбли од класификатори									
Гласачки ансамбл 1 (NB+LGBM+XGB)	0,7014	0,6443	0,6606	0,2898	0,6191	0,3948	0,8898	0,6696	0,7642
Гласачки ансамбл 2 (RF+LGBM+XGB)	0,7167	0,6392	0,7509	0,3515	0,4654	0,4005	0,8748	0,8130	0,8428
Гласачки ансамбл 3 (CB + LGBM + XG)	0,7178	0,6586	0,6603	0,2966	0,6560	0,4085	0,8982	0,6612	0,7617
Каскадирачки ансамбл 1 (CB+LGBM+RF=XG)	0,7146	0,5006	**0,8213**	0,5526	0,0015	0,0031	0,8214	0,9997	0,9018
Каскадирачки ансамбл 2 (XG+AB+CB=LGB)	0,7176	0,6586	0,6566	0,2949	0,6618	0,4080	0,8990	0,6554	0,7581

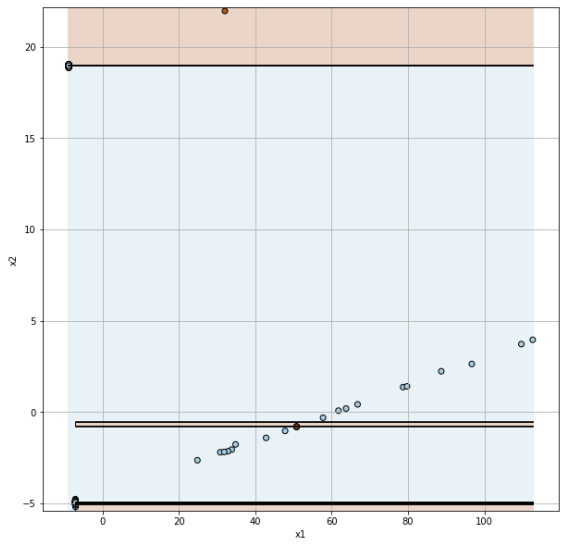
5.2 се претставени неговите граници на одлуки. ROC кривите[1] и границите на одлуки[2] на повеќе избрани класификатори се дадени во Прилозите 3 и 4 соодветно.

Од останатите ансамбли на класификатори, најдобри перформанси постигна гласачкиот класификатор во кој индивидуалните естиматори се CatBoost, LightGBM и XGBoost (0,7178), а блиску е каскадирачкиот класификатор составен од XGBoost, AdaBoost, CatBoost и LightGBM како конечен естиматор со ROC-AUC резултат од 0,7176. Со оглед на тоа што овие ансамбли се составени од нагласените модели базирани на дрва, нивните резултати се прилично блиски. Од моделите кои не припаѓаат на методите засновани на нагласени дрва и ансамблите од класификатори, ROC-AUC резултат повисок од 0,7 постигна единствено вештачката невронска мрежа (0,7013).

Графикон 5.1. ROC крива на LightGBM



Графикон 5.2. Граници на одлука на LGBM



Од линеарните модели, Линеарната дискриминантна анализа и Ridge регресијата забележаа ROC-AUC вредност од 0,6988, следени од Логистичката регресија (0,6929). Сите останати групи класификатори постигнаа значително послаби перформанси, при што ROC-AUC вредноста на наивниот баесов алгоритам изнесува 0,6562, на машините со носечки вектори 0,6461 и на методот на најблиски соседи 0,5998. [3]

При споредбата на ROC-AUC вредностите треба да се има во предвид дека тие пред се зависат од анализираното множество податоци, односно споредба на ROC-AUC резултати може да се врши единствено помеѓу модели обучени на исто множество на

[1] Површината под графички прикажаната крива не мора нужно да соодветствува со ROC-AUC резултатот претставен во Табела 5.1.

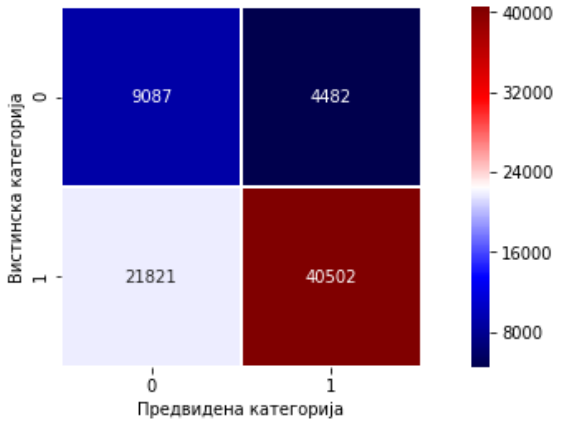
[2] Поради тоа што визуелизација на множеството податоци е невозможна со оглед на тоа што содржи 57 димензии, преку имплементирање на Анализа на главни компоненти множеството се сведува на 2 димензии чии граници на одлуки се визуелизирани на Графикон 5.2 и во Прилог 4.

[3] Поради нивната пресметковна интензивност, машините со носечки вектори и методите на најблиски соседи се обучени и тестирани на помал сет податоци во споредба со останатите модели, но сепак на сет податоци доволен тоа да не влијае значително на нивните перформанси.

податоци. Исто така, поради тоа што вредноста првенствено зависи од податоците, треба да се има во предвид дека разликите помеѓу поединечните класификатори кои на прв поглед можат да изгледаат ситни, всушност можат да имаат значително влијание врз способноста на моделот за предвидување, во зависност од карактеристиките на множеството податоци.

Освен ROC-AUC показателот, анализирана е и точноста на моделите, матрицата на конфузија и показателите изведени од неа. На Графикон 5.3 е претставена матрицата на конфузија за LightGBM, додека матриците на конфузија за останатите модели се претставени во Прилог 5. Иако точноста на моделите како резултат на недостатоците наведени во Дел 3.2 е инфериорен показател при споредба на моделите во однос на ROC-AUC, таа поседува одредени карактеристики поради кои тука е анализирана, пред се едноставната интерпретација и разбирање на показателот. Наспроти точноста, вредноста на ROC-AUC показателот е релативно потешка за интерпретирање.

Графикон 5.3. Матрица на конфузија на LightGBM



Доколку се анализира точноста изолирано, без да се земат во предвид останатите показатели поврзани со неа, лесно може да се добие погрешна претстава за перформансите на моделот. Согласно добиените резултати, далеку најголема точност имаат методот на случајни шуми (81,7%) и каскадирачкиот ансамбл на класификатори во чиј состав е вклучен методот на случајни шуми (82,1%). Меѓутоа, доколку се погледнат останатите показатели во прилог на точноста, како што се чувствителноста, специфичноста и Ф1 резултатот, ќе се увиди дека двата наведени модели имаат исклучително ниска чувствителност и Ф1 резултат при предвидување на категоријата „лоши“ кредити. Високиот процент на точност е резултат на тенденцијата на овие два модела да ја предвидуваат почестата категорија во најголем број од случаите, поради што почесто ја погодуваат категоријата на која и припаѓа опсервацијата, но истовремено методот на случајни шуми точно предвидел само 12% од „лошите“ кредити.

Со цел надминување на еден од главните недостатоци на точноста, конкретно високата вредност на овој показател кај моделите кои пречесто ја предвидуваат

почестата категорија, можеме да ја анализираме балансираната точност. Овој показател ги изолира ефектите од небалансираноста на множеството податоци за тестирање и дава појасна слика во овие случаеви. Точноста и балансираната точност се идентични единствено кај множествата податоци каде учеството на категориите е еднакво.

При анализирање на балансираната точност се доаѓа до заклучоци кои се очекувани и пологични во споредба со точноста. Најдобри перформанси и по овој показател покажува методот базиран на нагласени дрва LightGBM (65,98%), додека балансирана точност од 65% надминуваат и нему сродните алгоритми CatBoost (65,91%) и XGBoost (65,71%), како и гласачкиот и каскадирачкиот модел составени од овие алгоритми, кои постигнаа балансирана точност од 65,86%). За околу 1-2 п.п. послаба балансирана точност покажуваат невронските мрежи и линеарните модели, додека KNN, SVC и наивниот Баесов класификатор не покажуваат задоволителни резултати. Двата алгоритма кои имаа највисока точност, случајни шуми и каскадирачкиот ансамбл, истовремено постигнаа најниска балансирана точност, поради тоа што нивната пристрасност кон предвидување на почестата категорија води кон голем број на грешки кај балансираните множества на податоци.

Треба да се има во предвид дека изборот на најдобар модел преку точноста и балансираната точност зависи од природата на проблемот и субјективниот пристап на истражувачот, односно неговите преференции во однос на специфичноста и чувствителноста на моделот. Поради тоа, за разлика од ROC-AUC показателот, не секогаш може да се донесе конечна одлука за перформансите на различни модели само преку анализирање на точноста, без да се земат во предвид околностите и преференциите на истражувањето.

И покрај различниот пристап и начин на пресметка, од Табела 5.2 видлива е корелацијата помеѓу рангот на моделите според балансираната точност и ROC-AUC показателот. Кај двата показателa, моделите базирани на дрва и другите ансамбли се со најдобри перформанси, следени од невронските мрежи, линеарните модели, наивниот Баесов класификатор, KNN и машините со носечки вектори.

Табела 5.2. Ранг на моделите за класификација според релевантни показатели

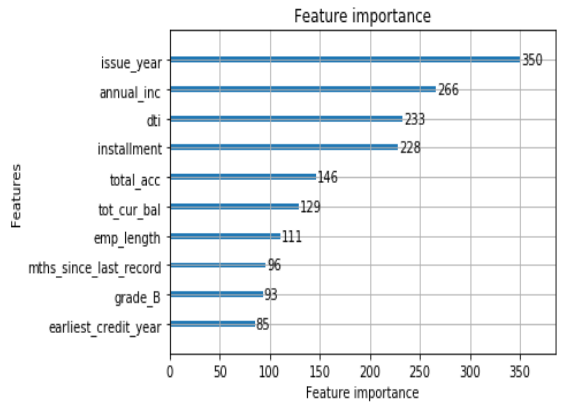
Модел за класификација	Ранг според ROC-AUC резултат	Ранг според балансирана точност
Линеарни модели		
Линеарна дискриминантна анализа	12	8
Логистичка регресија	15	11
SGD класификатор	16	14
Ridge класификатор	11	9
KNN	19	17
Наивен Баесов алгоритам	17	15
Машини со носечки вектори (SVC (rbf))	18	16
Модели базирани на дрва		
Случајни шуми (RF)	13	18
AdaBoost	8	6
XGBoost	7	5
CatBoost	3	2
LightGBM	**1**	**1**
Вештачки невронски мрежи		
MLPClassifier	14	12
Длабока невронска мрежа	10	7
Други ансамбли од класификатори		
Гласачки ансамбл 1 (NB+LGBM+XG)	9	10
Гласачки ансамбл 2 (RF +LGBM+XG)	5	13
Гласачки ансамбл 3 (CB + LGBM + XG)	2	4
Каскадирачки ансамбл 1 (CB+LGBM+RF=XG)	6	19
Каскадирачки ансамбл 2 (XG+AB+CB=LGB)	4	3

При обучувањето, освен параметрите кои се користат за имплементирање, некои алгоритми ја квантифицираат и важноста на поединечните независни променливи, според која можат и да ги рангираат. Ова е исклучително важно при интерпретирање на моделите, особено на black box моделите чиј начин на класифицирање е невозможен за интерпретирање и објаснување. Имајќи квантифицирани податоци за важноста на променливите, заемодавателот може полесно да ги увиди факторите кои влијаеле одредено кредитно барање да биде класифицирано како „добро“ или „лошо“. Најважните независни променливи според кои моделите LightGBM и XGBoost носат одлука за класификација на кредитно барање, рангирани според вредноста на F-тестот се претставени на графиконите 5.4 и 5.5.

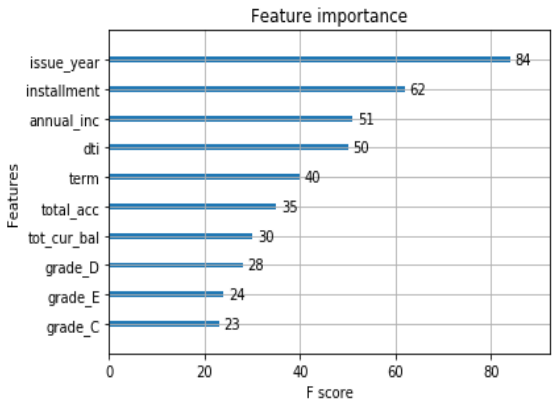
Од презентираните резултати, може да се донесат неколку заклучоци за примената на техниките од машинско учење во областа на кредитната анализа. Методот базиран на најблиски соседи, машините со носечки вектори и наивниот Баесов класификатор не постигнуваат задоволителни резултати при проценувањето на веројатноста за враќање на кредитот. Дополнително, методот на најблиски соседи и машините со носечки вектори се исклучително пресметковно интензивни, така што

нивното обучување, подесување на хипер-параметрите и тестирање на обемни множества податоци може да биде исклучително долго и непрактично.

Графикон 5.4. Најважни независни променливи според LightGBM



Графикон 5.5. Најважни независни променливи според XGBoost



Линеарните модели, како што се логистичката регресија и линеарната дискриминантна анализа и невронските мрежи покажуваат подобри перформанси во споредба со наведените во претходниот пасус, и тие се потенцијална техника за проценка на кредитниот ризик. Всушност, линеарните модели имаат широка употреба во процесот на кредитна анализа. Невронските мрежи, иако покажуваат слични перформанси, се соочуваат со одредени недостатоци. За разлика од линеарните модели, невронските мрежи се покомплицирани за имплементација и бараат специјализирани познавања од областа. Дополнително, тие се целосно black box модел, така што процесот кој е користен од страна на вештачките невронски мрежи и нивните резултати не можат да се интерпретираат и објаснат.

Ансамблите од класификатори покажуваат уште подобри резултати, меѓутоа тие не можат да се оценуваат изолирано, со оглед на тоа што нивните перформанси зависат пред се од класификаторите кои се вклучени во ансамблот.

Конечно, методите засновани на нагласени дрва за одлучување покажуваат најдобри перформанси при проценувањето на веројатноста за неврaќање на кредитот и нивната класификација. Ова е и очекувано, со оглед на тоа што се работи за исклучително нови, модерни и софистицирани алгоритми, од кои некои се развиени од страна на најголемите светски технолошки компании. Дополнително, овие алгоритми се брзи и лесно изводливи во пракса, далеку пофлексибилни во однос на другите модели, а постигнуваат подобри резултати и при помали и помалку претпроцесирани множества на податоци. Нивните резултати се лесно разбирливи, со оглед на тоа што ги

квантифицираат важностите на независните променливи, и со тоа овозможуваат едноставно објаснување и оправдување на одлуката за прифаќање или одбивање на кредитното барање.

Заклучни согледувања

Развивањето на попрецизен модел за кредитно оценување придонесува кон подобро одлучување по кредитните барања и помали загуби за финансиските институции по основ на ненаплатени кредити. Во овој труд се обучени и тестирани 19 модели за класификација поделени во 7 различни групи, со цел идентификување на соодветни модели за оценување на кредитниот ризик. Моделите се евалуирани примарно преку ROC-AUC резултатот, но и според показателот балансирана точност.

Линеарните модели и невронските мрежи се покажаа како соодветни методи за оценување на кредитниот ризик, од аспект на успешноста при предвидувањето. Меѓутоа, перформансите на невронските мрежи не се доволно квалитетни да компензираат за недостатоците на овој метод, како што се неможноста за интерпретирање и времето потребно за обучување на моделот. Линеарните модели, освен што покажаа солидни перформанси, се и едноставни за имплементирање и разбирање, што ги прави посоодветни во однос на невронските мрежи. Сепак, моделите базирани на дрва за одлучување, пред се модерните алгоритми кои користат нагласени дрва за одлучување, се супериорни во однос на линеарните и сите останати анализирани модели. Дополнително, овие модели се едноставни за имплементација, брзи, флексибилни и овозможуваат разбирање на одлуките што ги носат преку идентификување на најважните независни променливи. Ансамблите на модели за класификација не се покажаа како поуспешни од моделите базирани на нагласени дрва на даденото множество податоци, но за целосна слика за нивните перформанси потребни се дополнителни истражувања. Сите други групи на модели за класификација се покажаа како несоодветни за класификација на кредитни барања.

Со цел ограничување на должината на трудот, моделите за класификација се тестирани врз само едно множество на податоци, иако се работи за едно од најквалитетните достапни множества на податоци. Сепак, за целосна слика на

успешноста на класификаторите, потребно е дополнително тестирање врз нови сетови податоци, со оглед на значајното влијание што го имаат карактеристиките на податоците врз перформансите на моделите.

И покрај горенаведеното, наодите од овој труд можат да послужат како патоказ за македонските финансиски институции чија дејност е кредитирањето каде да бидат насочени нивните напори за подобрување на моделите за кредитно оценување. Врз база на историските податоци со кои располагаат, финансиските институции во земјава можат да направат обиди за развивање на модели базирани на нагласени дрва. Вака создадените модели треба да се споредат со постоечките модели кои ги користат, според критериуми кои се соодветни на потребите на институцијата. Засновано на овие анализи, финансиските институции можат да ги подесат своите постоечки методи за оценување на кредитен ризик.

Библиографија

Addo, P., Guegan, D. and Hassani, B. (2018), "Credit risk analysis using machine and deep learning models", *Risks,* Vol. 6, Paper No. 38.

Angelini, E., Tollo, G. and Roli, A. (2008), "A neural network approach for credit risk evaluation", *The Quarterly Review of Economics and Finance,* Vol. 48, No. 4, pp. 733–55.

Bastos, J. A. (2008), "Credit scoring with boosted decision trees", MPRA Paper No. 8156

Bazarbash, M. (2019), "FinTech in Financial Inclusion: Machine Learning Applications in Assessing Credit Risk", IMF Working Paper No. 19/109.

Bellotti, T. and Crook, J. (2009), "Support vector machines for credit scoring and discovery of significant features", *Expert System Applications,* Vol. 36, No. 2, pp. 3302–8.

Blanco, A., Pino-Mejias, R., Lara, J. and Rayo, S. (2013), "Credit scoring models for the microfinance industry using neural networks: evidence from Peru", *Expert System Applications,* Vol. 40, No. 1, pp. 356–64.

Breiman, L. (2001), "Random Forests", *Machine Learning,* Vol. 45, No. 1, pp. 5–32.

Cao, J., Hongke, L., Weiwei, W. and Jian, W. (2013), "A Loan Default Discrimination Model Using Cost-Sensitive Support Vector Machine Improved by PSO", *Information Technology and Management*, Vol. 14, pp. 193–204.

Chen, T. and Guestrin, C. (2016), "XGBoost: A Scalable Tree Boosting System", In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco.

Crook, J. N., Edelman, D. B. and Thomas, L. C. (2007), "Recent developments in consumer credit risk assessment", *European Journal of Operational Research,* Vol. 183, No. 3, pp. 1447–65.

Galindo, J. and Tamayo, P. (2000), "Credit Risk Assessment Using Statistical and Machine Learning: Basic Methodology and Risk Modeling Applications", *Computational Economics,* Vol. 15, No. 1-2, pp. 107–43.

Hamori, S., Minami, K., Takahiro, K., Yuji, M. and Chikara, W. (2018), "Ensemble Learning or Deep Learning? Application to Default Risk Analysis", *Journal of Risk and Financial Management,* Vol. 11, No.1, Paper No. 12.

Harris, T. (2013), "Quantitative credit risk assessment using support vector machines: Broad versus Narrow default definitions", *Expert System Applications*, Vol. 40, No. 11, pp. 4404–13.

Huang, C. L., Mu, C. C. and Chieh, J. W. (2007), "Credit Scoring with a Data Mining Approach Based on Support Vector Machines", *Expert System Applications,* Vol. 33, No. 4, pp. 847–56.

Huang, Z., Chen, H., Hsu, C. J., Chen, W. H. and Wu, S. (2004), "Credit rating analysis with support vector machines and neural networks: a market comparative study", *Decision Support Systems,* Vol. 37, No. 4, pp. 543–58.

Khandani, A. E., Adlar J. K. and Andrew, W. Lo. (2010), "Consumer credit-risk models via machine-learning algorithms", *Journal of Banking & Finance,* Vol. 34, No. 11, pp. 2767-87.

Lessmann, S., Baesens, B., Seow, H. V. and Thomas, L. C. (2015), "Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring: An Update of Research", *European Journal of Operational Research,* Vol. 247, No. 1, pp. 124–36.

Ong, C. S., Huang, J. J. and Tzeng, G. H. (2005), "Building credit scoring models using genetic programming", *Expert Systems with Applications,* Vol. 29, No. 1, pp. 41–47.

Oreski, S., Oreski, D. and Oreski, G. (2012), "Hybrid system with genetic algorithm and artificial neural networks and its application to retail credit risk assessment", *Expert System Applications,* Vol. 39, No. 16, pp. 12605–17.

Petropoulos, A., Siakoulis, V., Stavroulakis, E. and Klamargias, A. (2018), "A robust machine learning approach for credit risk analysis of large loan level datasets using deep learning and extreme gradient boosting", Ninth IFC Conference on "Are post-crisis statistical initiatives completed?, Bank for International Settlements.

Rafiei, F. M., Manzari, S. and Bostanian, S. (2011), "Financial health prediction models using artificial neural networks, genetic algorithm and multivariate discriminant analysis: Iranian evidence", *Expert System Applications,* Vol. 38, No. 8, pp. 10210–7.

Tsai, C. F. and Chen, M. L. (2010), "Credit rating by hybrid machine learning techniques", *Applied Soft Computing,* Vol. 10, No. 2, pp. 374–80.

Vapnik, V. (1995), *The Nature of Statistical Learning Theory*, Springer, New York.

West, D. (2000), "Neural network credit scoring models", *Computers and Operations Research*, Vol. 27, No. 11, pp. 1131–52.

Yao, X., Crook, J. and Andreeva, G. (2017), "Enhancing two-stage modelling methodology for loss given default with support vector machines", *European Journal of Operational Research,* Vol. 263, No. 2, pp. 679–89.

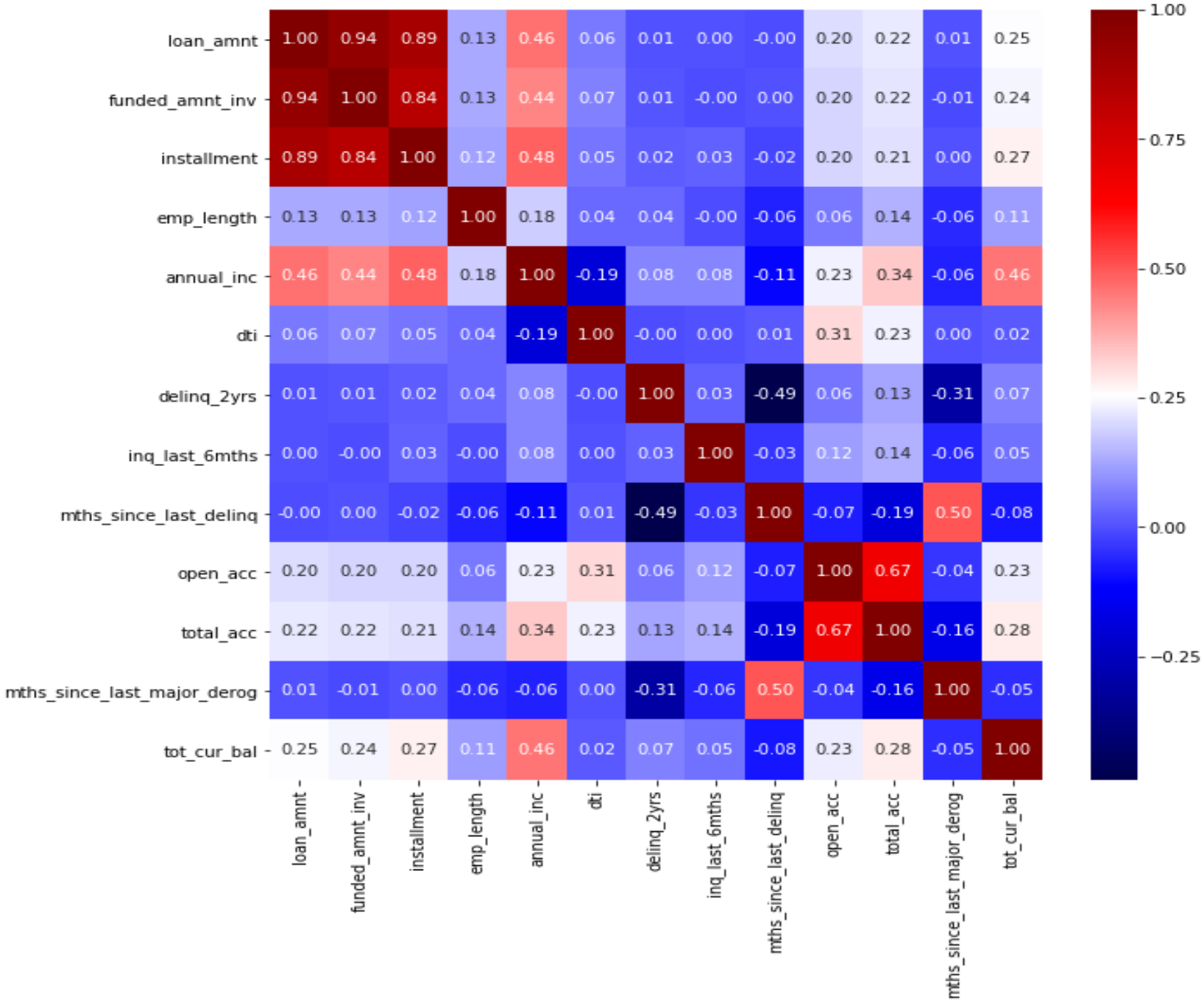
Zhang, W. (2017), "Machine Learning Approaches to Predicting Company Bankruptcy", *Journal of Financial Risk Management,* Vol. 6, No. 4, pp. 364-74.

Прилог 1. Опис на поважни независни променливи

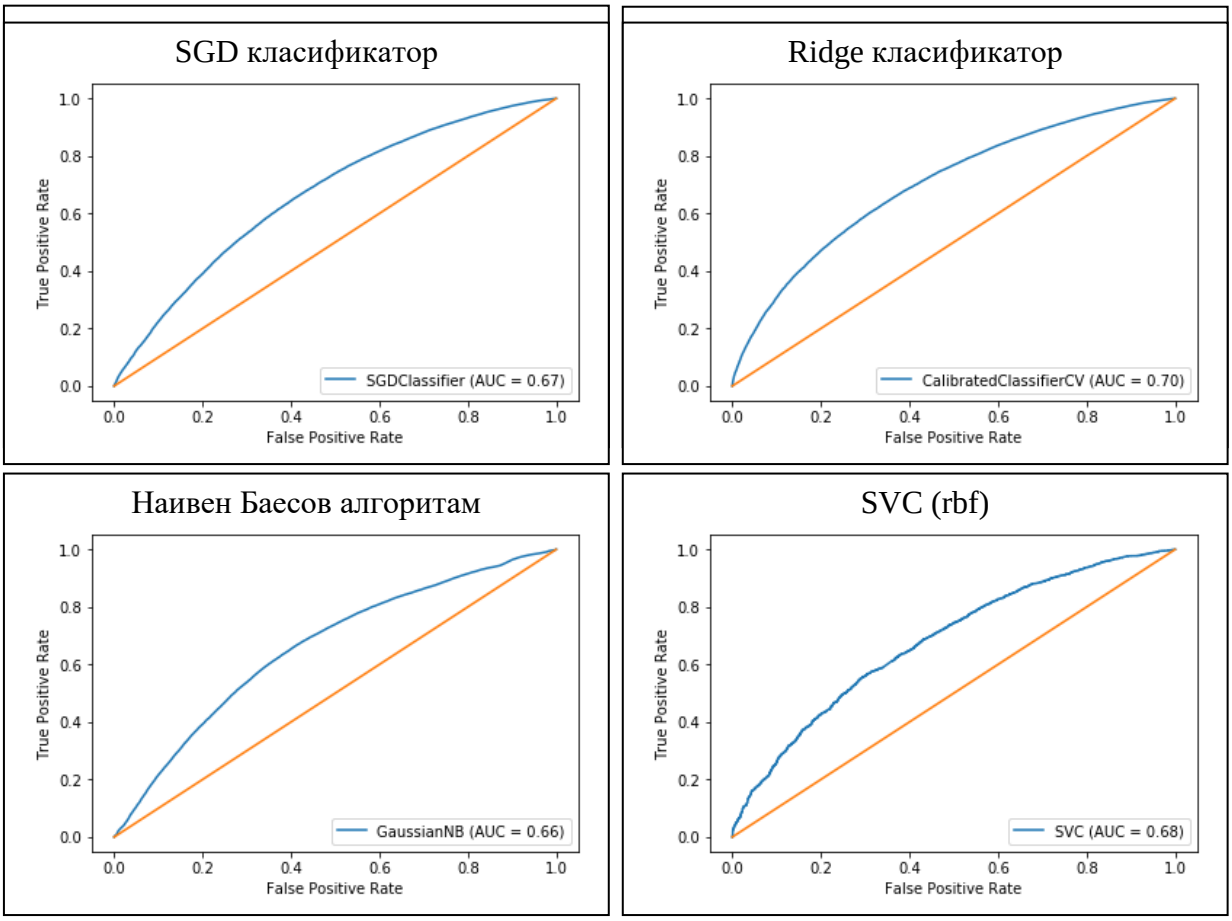
Независна променлива	Опис
log(loan_amnt)	Логаритам од износот на бараниот кредит
term	Рок на отплата на бараниот кредит
installment	Месечна рата на кредитот доколку е одобрен
emp_length	Должина на работниот стаж
log(annual_inc)	Логаритам од вкупниот годишен приход на кредитобарателот
dti	Показател вкупна месечна отплата по основ на долг/вкупни месечни приходи, во %
delinq_2yrs	Колку пати во последните 2 години кредитобарателот доцнел при плаќање на обврските повеќе од 30 дена
mths_since_last_delinq	Пред колку месеци кредитобарателот последен пат доцнел при плаќање на обврските (максимална вредност 152)
open_acc	Број на вкупно отворени кредитни линии на кредитобарателот
total_acc	Број на тековни кредитни линии на кредитобарателот
mths_since_last_major_derog	Пред колку месеци кредитобарателот последен пат доцнел при плаќање на обврските повеќе од 90 дена (максимална вредност 159)
tot_cur_bal	Вкупно салдо на сите сметки на кредитобарателот
grade	Категорија на ризик на кредитобарателот
home_ownership	Категориска променлива, го содржи станбениот статус на кредитобарателот
issue_month	Категориска променлива, го содржи месецот во кој е поднесено кредитното барање
purpose	Категориска променлива, ја содржи намената на кредитот
earliest_credit_year	Година во која кредитобарателот за прв пат користел кредит

Прилог 2. Дескриптивна статистика на поважните нумерички варијабли и корелациона топлотна карта

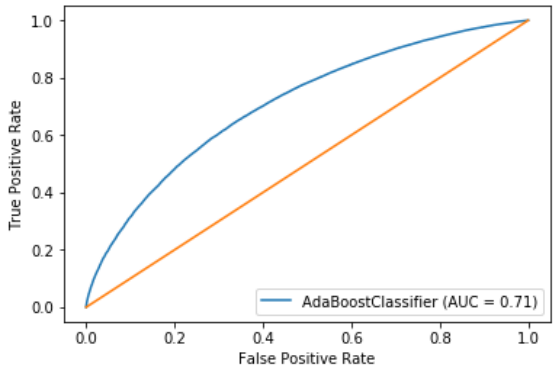
Променлива	Број на опсервации	Просечна вредност	Стандардна девијација	Минимална вредност	25%	50%	75%	Максимална вредност
loan_amnt	252.971	13.562,8	8.131,7	500	7.200	12.000	18.250	35.000
installment	252.971	418,1	244,9	15,7	239,6	365,2	547,6	1.424,6
annual_inc	252.971	72.538,3	58.811,8	3.000	45.000	62.000	87.000	8.706.582
dti	252.971	16,5	7,8	0	10,8	16,2	22,0	57,1
delinq_2yrs	252.971	0,2	0,7	0	0	0	0	29
mths_since_last_delinq	252.971	99,1	58,9	0	37	150	150	152
mths_since_last_major_derog	252.971	130,0	42,7	0	150	150	150	159
tot_cur_bal	252.971	123.893,4	134.179,8	0	38.429,5	81.002,0	164.900,5	8.000.078,0


loan_amnt 1.00 0.94 0.89 0.13 0.46 0.06 0.01 0.00 -0.00 0.20 0.22 0.01 0.25
funded_amnt_inv 0.94 1.00 0.84 0.13 0.44 0.07 0.01 -0.00 0.00 0.20 0.22 -0.01 0.24
installment 0.89 0.84 1.00 0.12 0.48 0.05 0.02 0.03 -0.02 0.20 0.21 0.00 0.27
emp_length 0.13 0.13 0.12 1.00 0.18 0.04 0.04 -0.00 -0.06 0.06 0.14 -0.06 0.11
annual_inc 0.46 0.44 0.48 0.18 1.00 -0.19 0.08 0.08 -0.11 0.23 0.34 -0.06 0.46
dti 0.06 0.07 0.05 0.04 -0.19 1.00 -0.00 0.00 0.01 0.31 0.23 0.00 0.02
delinq_2yrs 0.01 0.01 0.02 0.04 0.08 -0.00 1.00 0.03 -0.49 0.06 0.13 -0.31 0.07
inq_last_6mths 0.00 -0.00 0.03 -0.00 0.08 0.00 0.03 1.00 -0.03 0.12 0.14 -0.06 0.05
mths_since_last_delinq -0.00 0.00 -0.02 -0.06 -0.11 0.01 -0.49 -0.03 1.00 -0.07 -0.19 0.50 -0.08
open_acc 0.20 0.20 0.20 0.06 0.23 0.31 0.06 0.12 -0.07 1.00 0.67 -0.04 0.23
total_acc 0.22 0.22 0.21 0.14 0.34 0.23 0.13 0.14 -0.19 0.67 1.00 -0.16 0.28
mths_since_last_major_derog 0.01 -0.01 0.00 -0.06 -0.06 0.00 -0.31 -0.06 0.50 -0.04 -0.16 1.00 -0.05
tot_cur_bal 0.25 0.24 0.27 0.11 0.46 0.02 0.07 0.05 -0.08 0.23 0.28 -0.05 1.00
loan_amnt
funded_amnt_inv
installment
emp_length
annual_inc
dti
delinq_2yrs
inq_last_6mths
mths_since_last_delinq
open_acc
total_acc
mths_since_last_major_derog
tot_cur_bal
1.00
0.75
0.50
0.25
0.00
−0.25

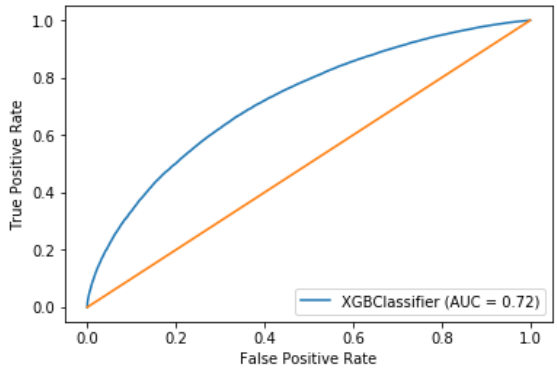
Прилог 3. ROC-криви на избрани модели за класификација



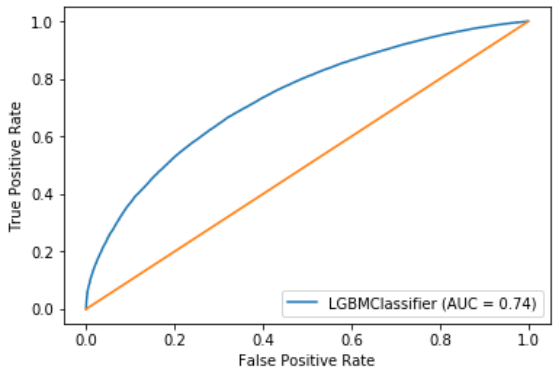
AdaBoost



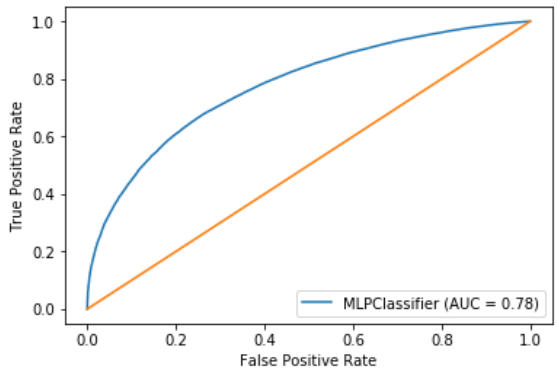
XGBoost



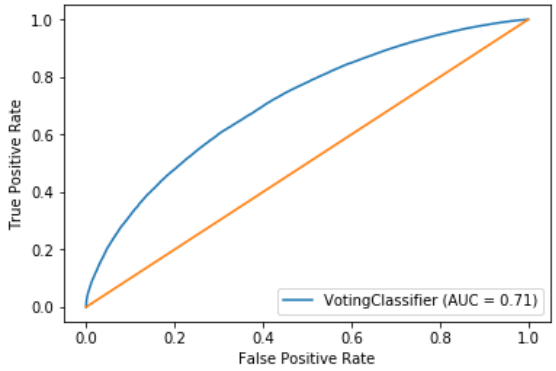
LightGBM



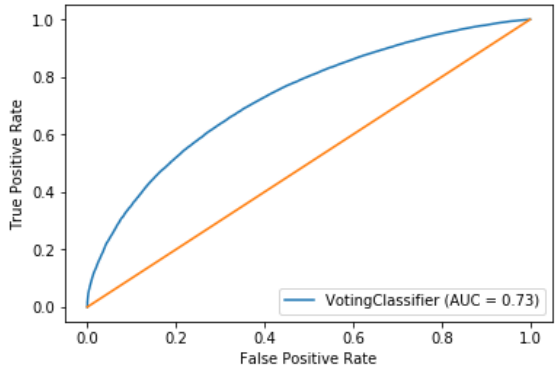
MLPClassifier



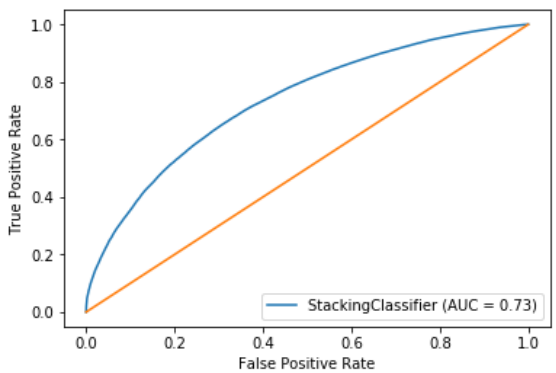
Гласачки ансамбл 1 (NB+LGBM+XGB)



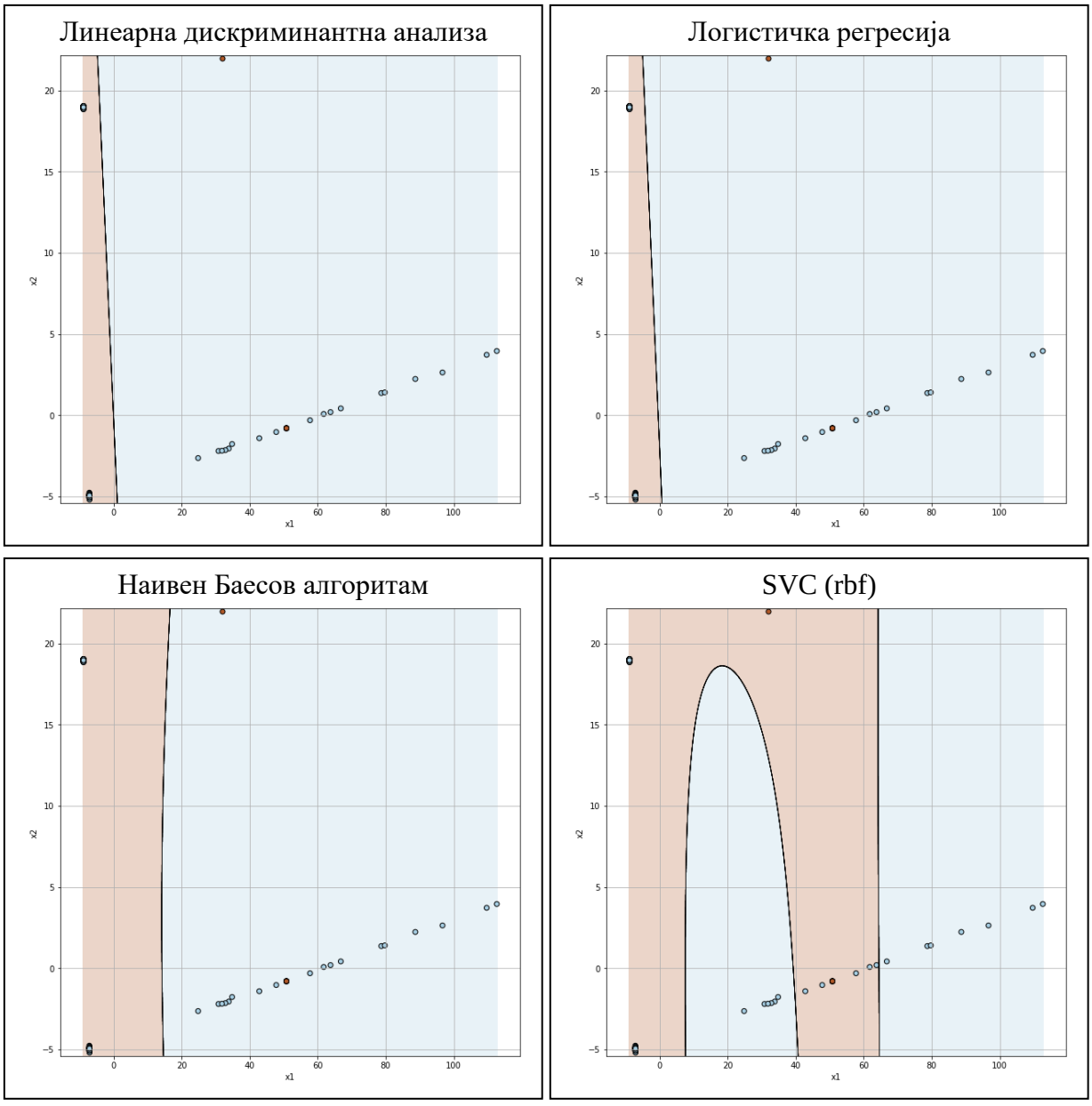
Гласачки ансамбл 3(CB + LGBM + XGB)



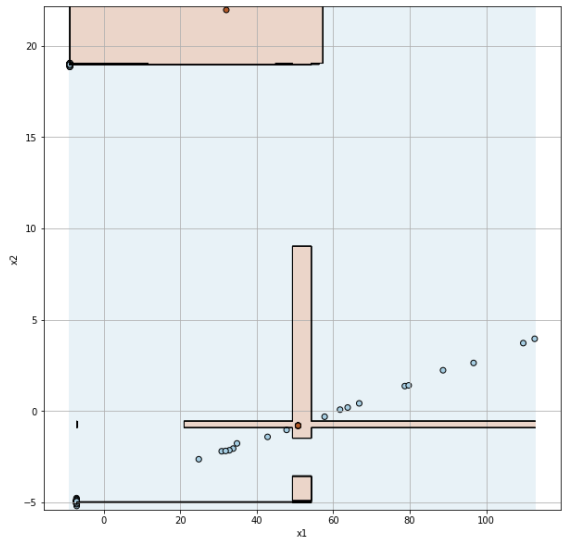
Каскадирачки ансамбл 2
(XG+AB+CB=LGB)



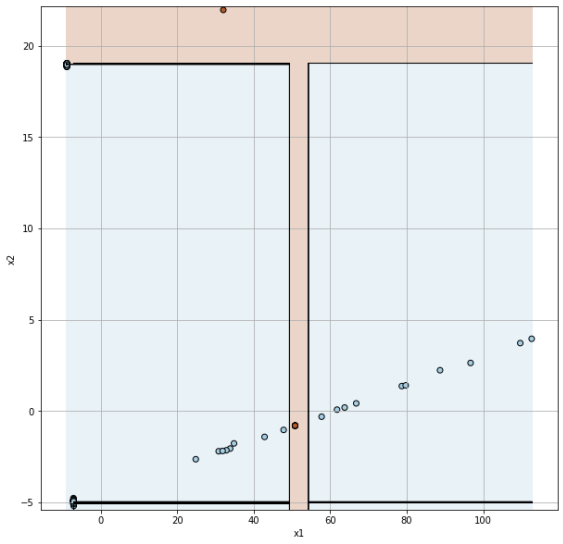
Прилог 4. Граници на одлучување на избрани модели за класификација



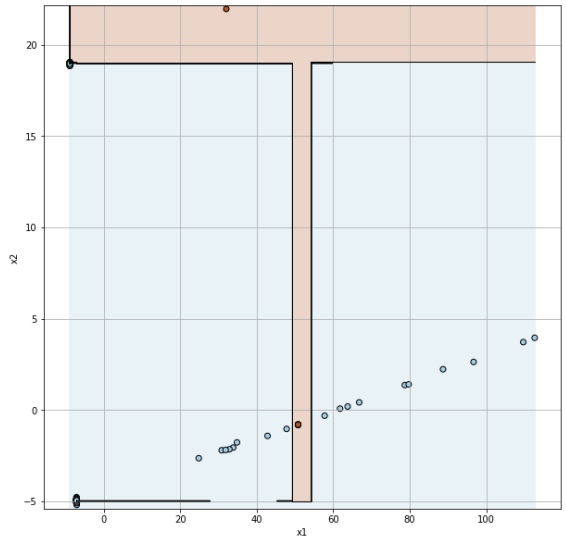
Случајни шуми (RF)



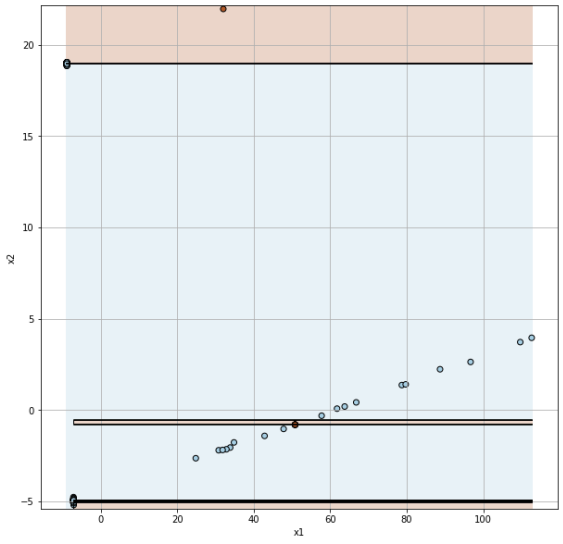
XGBoost



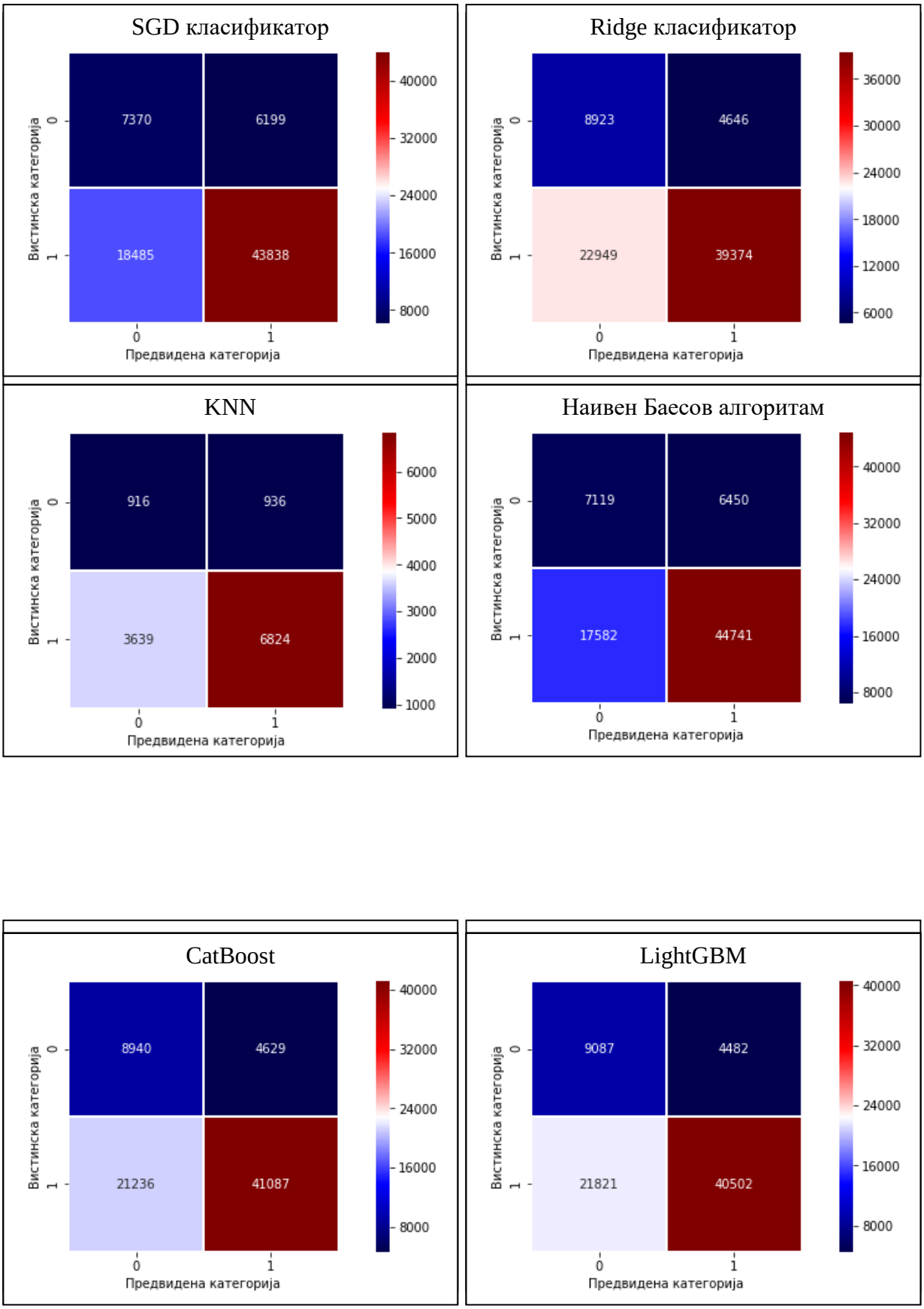
CatBoost



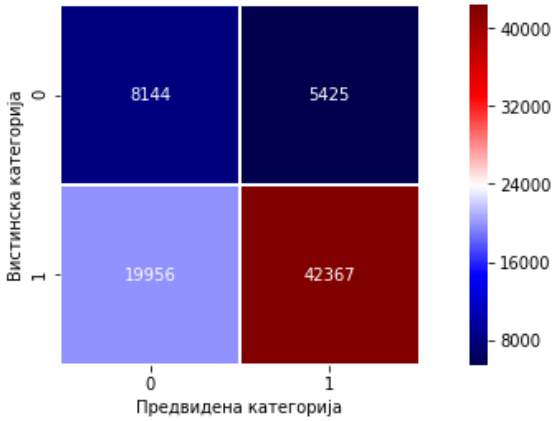
LightGBM



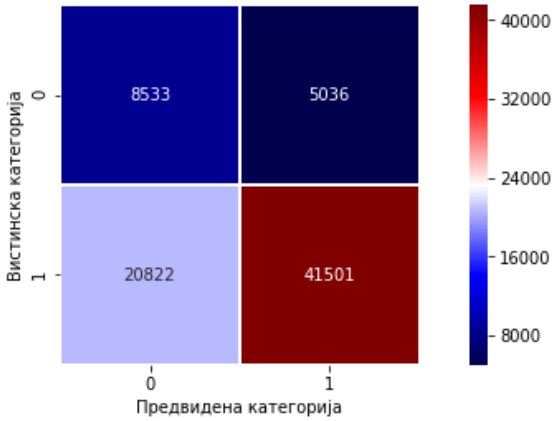
Прилог 5. Матрици на конфузија на моделите за класификација



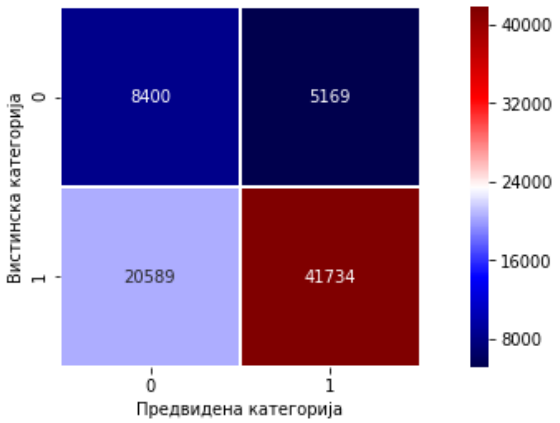
MLPClassifier



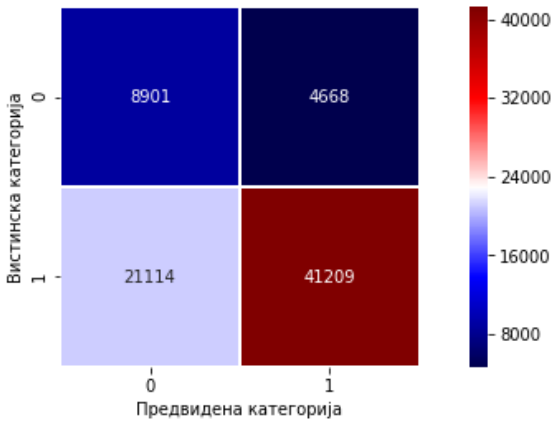
Длабока невронска мрежа



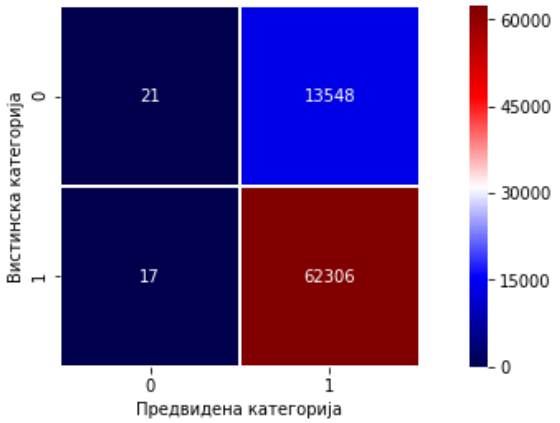
Гласачки ансамбл 1 (NB+LGBM+XGB)



Гласачки ансамбл 3(CB + LGBM + XGB)



Каскадирачки ансамбл 1
(CB+LGBM+RF=XG)



Каскадирачки ансамбл 2
(XG+AB+CB=LGB)