**Сара Сапунџија (ИН14/2021)**

**Ранка Радуловић (ИН16/2021)**

Процена кредитне способности кандидата на основу анализе више параметара  
(Кредитна способност)

Извештај за практично истраживање

# Увод

## Предмет истраживања

Један од фундамената економије савременог друштва су кредити; било да су у питању краткорочни или дугорочни, издавање кредита је процес који захтева детаљну обраду одређене групе података везане за финансијске способности клијента који подноси захтев за кредит. Како је смер који похађамо, информациониг инжењеринг, синтеза рачунарских наука и појединих грана економије, сматрамо да је истраживање ове теме добар избор, јер захтева познавање обе од поменутих наука. Подаци коришћени у раду су из 1994. године и преузети са следећег линка: <https://www.openml.org/search?type=data&sort=runs&status=active&id=31>.

## Циљеви истраживања

Циљ истраживања је да направимо предиктивни модел употребом алгоритама машинског учења који ће за задати улаз моћи да класификује клијента као адекватног или неадекватног за одобрење кредита. Основни циљ нашег истраживања је да идентификујемо и анализирамо факторе који највише утичу на процену кредитне способности купаца. Истраживањем података о кредитном одобрењу, укључујући параметре као што су статус запослености и кредитна историја, планирамо да пронађемо најбоље приступе за предвиђање успешног исхода кредитног захтева.

## Задаци истраживања

Главни задаци нашег истраживања укључују анализу и разумевање сваког од параметара кредитне способности, примену различитих алгоритама за предвиђање исхода кредитних захтева.

Задаци нашег истраживања могу се поделити у 4 групе:

1. идентификација предиктивних алгоритама – за почетак, неопходно је утврдити који су то алгоритми уопште предиктивни;
2. погодни алгоритми за имплементацију – потом, од предходне групе алгоритама треба издвојити оне који су погодни за имплементацију, спрам датог скупа података;
3. претпроцесирање – даље морамо средити улазне податке, тако да буду погодан улаз за моделе машинског учења;
4. визуализовање истраживања – за крај, неопходно је приказати резултате добијеног истраживања преко графикона.

## Очекивани резултати истраживања

Очекујемо да ћемо идентификовати кључне факторе који утичу на одобрење кредита клијентима. Претпостављамо да ће међу најзначајнијим бити приходи, статус запослености и кредитна историја, кредитно задужење, некретнине. Очекујемо да ће наша анализа потврдити важност ових фактора у процесу одобравања кредита и да ће наши резултати бити од користи за унапређење процеса одобравања кредита.

# Методологија

## Коришћени подаци

За наше истраживање користили смо податке из горе наведеног скупа (поднаслов Предмет истраживања). Овај скуп података је садржао комбинацију нумеричких (float) и категоричких (object) атрибута, због чега смо морали извршити трансформацију података како бисмо припремили улаз за моделе машинског учења. Важно је напоменути да у нашем скупу података није било недостајућих вредности, што је олакшало процес припреме података за анализу и моделовање. На крају, извршили смо кодирање категоричких атрибута користећи технику one-hot енкодирања како бисмо омогућили да модели као што су логистичка регресија, К-најближих суседа, алгоритам екстремног појачања градијента и други могу успешно обрадити те податке.

## Претходна истраживања других особа над коришћеним подацима

...

## Методе истраживања

Приликом нашег истраживања, користиле смо неколико алгоритама за предикцију. Пре саме имплементације било је потребно упознати се детаљно са подацима, припремити их за коришћење и проучити који алгоритми су погодни за наше истраживање.

Припрема података

Наш скуп података садржи 21 колону које су или типа Object или типа float. Дакле садржи и нумеричке и категоричке вредности. Међутим, већина модела за машинско учење не ради са категоричким подацима и да би их уклопили у модел морају бити претворени у нумеричке податке. Како бисмо решиле овај проблем, користиле смо енкодирање, конкретно One-hot encoding. Предности су: омогућава коришћене категоричких варијабли у моделима који захтевају нумерички улаз, може да побољша перформансе модела пружајући више информација о категоричкој варијабли, може помоћи да се избегне проблем ординалности, који може да се јави када категоричка варијабла има природан поредак (нпр. „мала“, „средња“, „велика“).

K-најближих суседа (KNN алгоритам)

Алгоритам к-најближих суседа је популарна техника машинског учења која се користи за задатке класификације и регресије. Ослања се на идеју да блиске тачке података обично имају сличне ознаке или вредности. Непараметарски је алгоритам, што значи да не прави никакве претпоставке о основним подацима. Не учи из скупа за обуку одмах, већ складишти скуп података и у време класификације врши акцију на скупу података. Када се убаци нови податак, мери се еуклидска удаљености између њега и осталих тачака у скупу, а потом се узима к најближих. За сваку категорију се преброји колико јој од к изабраних тачака припада и додељујемо нови податак категорији која има највише тачака.

Логистичка регресија

Логистичка регресија предвиђа излаз категоричке зависне променљиве. Према томе, исход мора бити категоричка или дискретна вредност. Може бити или да или не, 0 или 1, тачно или нетачно, итд. али уместо да даје тачну вредност као 0 и 1, даје вероватноће које се налазе између 0 и 1.

Aлгоритам екстремног појачања градијента (XGBoost)

XGBoost је моћан алгоритам машинског учења базиран на Gradient Boosted Decision Trees (GBT) принципу. Користи се за широк спектар проблема као што су класификација, регресија, рангирање и предвиђања. Алгоритам ради тако што се итеративно гради серија стабала одлучивања, при чему свако стабло исправља грешке претходног. Основна идеја је додавање тежина свакој променљивој у складу са њеним значајем у предвиђању резултата. XGBoost је популаран због своје ефикасности, скалабилности и способности да постиже високу тачност у предвиђањима. За његово успешно коришћење, корисно је разумети основе стабала одлучивања, на којима се заснива његова архитектура и функционалност. У пракси, XGBoost често доминира у такмичењима у области података због своје способности да ефикасно обрађује велике скупове података и пружа робустне резултате.

Класификатор гласанја (Voting Classifier)

Класификатор гласања је модел машинског учења који искуство стиче тренирањем на колекцији различитих модела и предвиђа излаз (класу) на основу класе са највећом вероватноћом да постане излаз. Идеја је изградња модела који учи од других модела и комбинује их како би побољшао перформансе. Овај модел се користи за предвиђање класне ознаке гласањем већине, а користан је када желимо да узмемо у обзир резултате различитих модела и сагледамо целокупну слику. Крајњу одлуку доносимо на основу већине или просека предвиђања.

Идеја комбиновања је настала због претпоставке да различити модели праве различите грешке на истом скупу података, тако да ће се грешке „исправити“ уколико сагледавамо већину или просек.

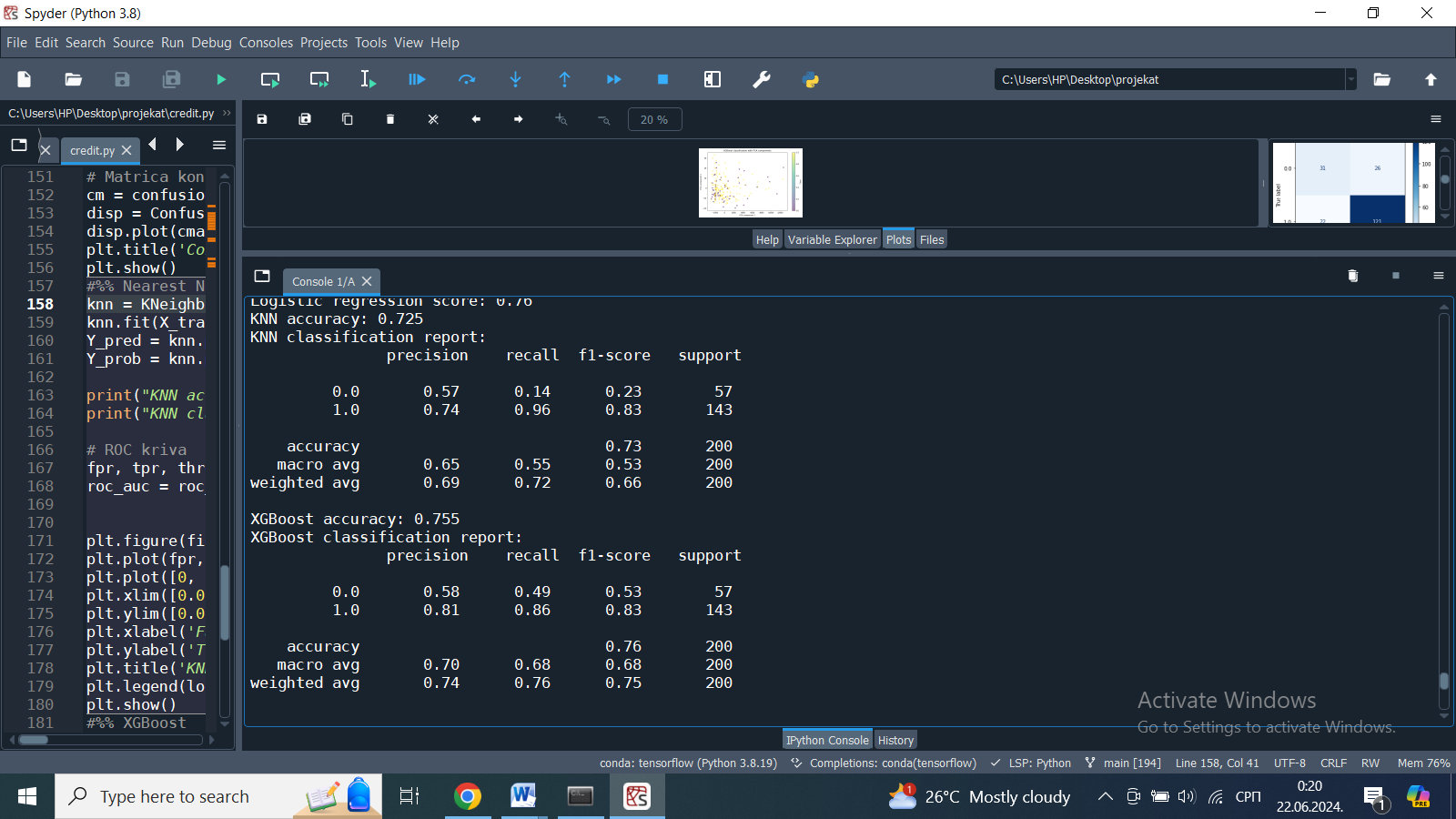
У нашем истраживању користиће се обе методе: гласање већине и гласање просека. Где се као финално предвиђање користи класа која добије већину гласова, односно прослеђује просечна вероватноћа класа предложених од стране основних класификатора у случају гласања просека.

# Резултати

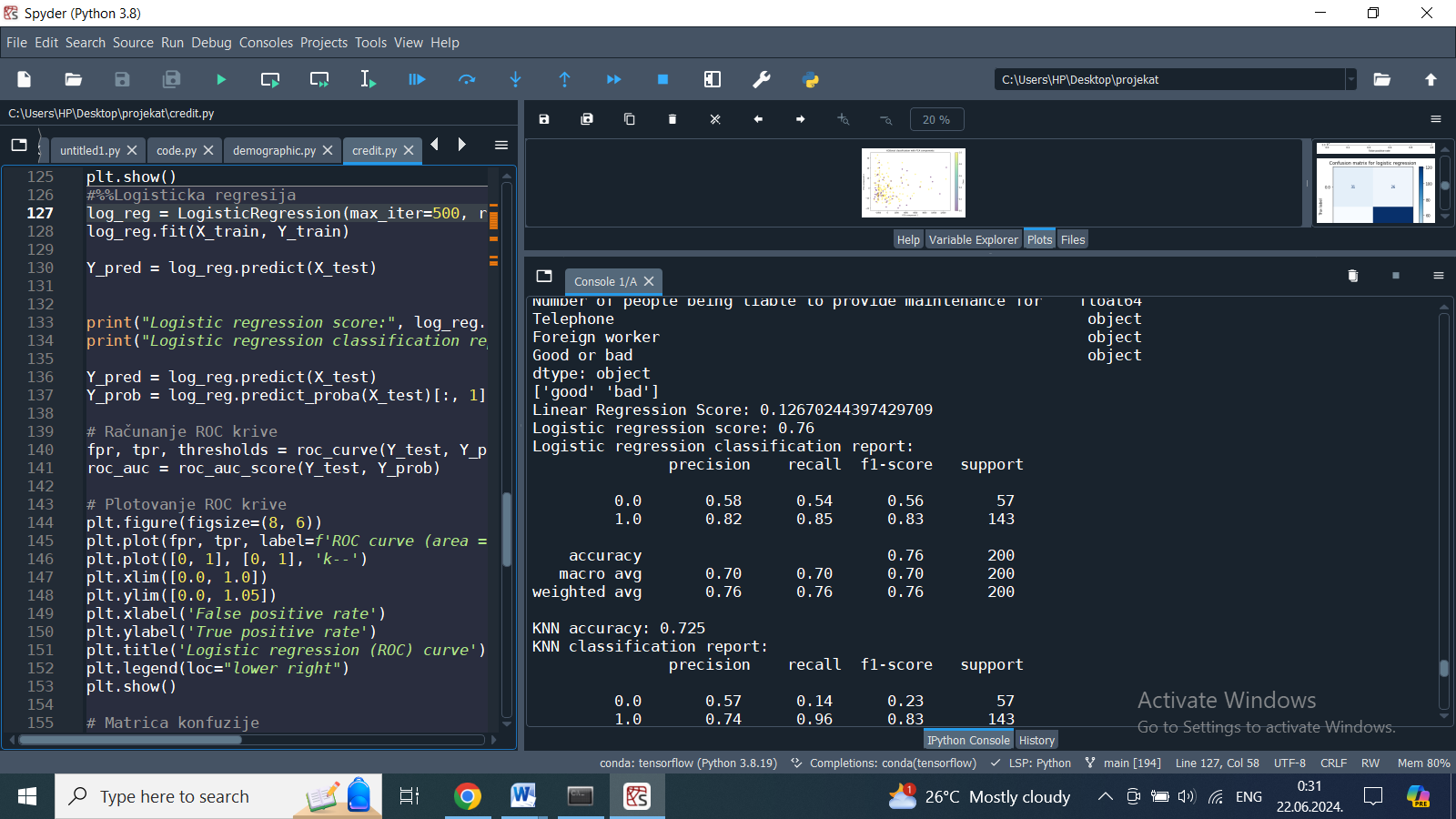
## Приказ резултата

У наставку су приложени резултати нашег истраживања. За сваки алгоритам за предвиђање, приказана је тачност, прецизност и осетљивост.

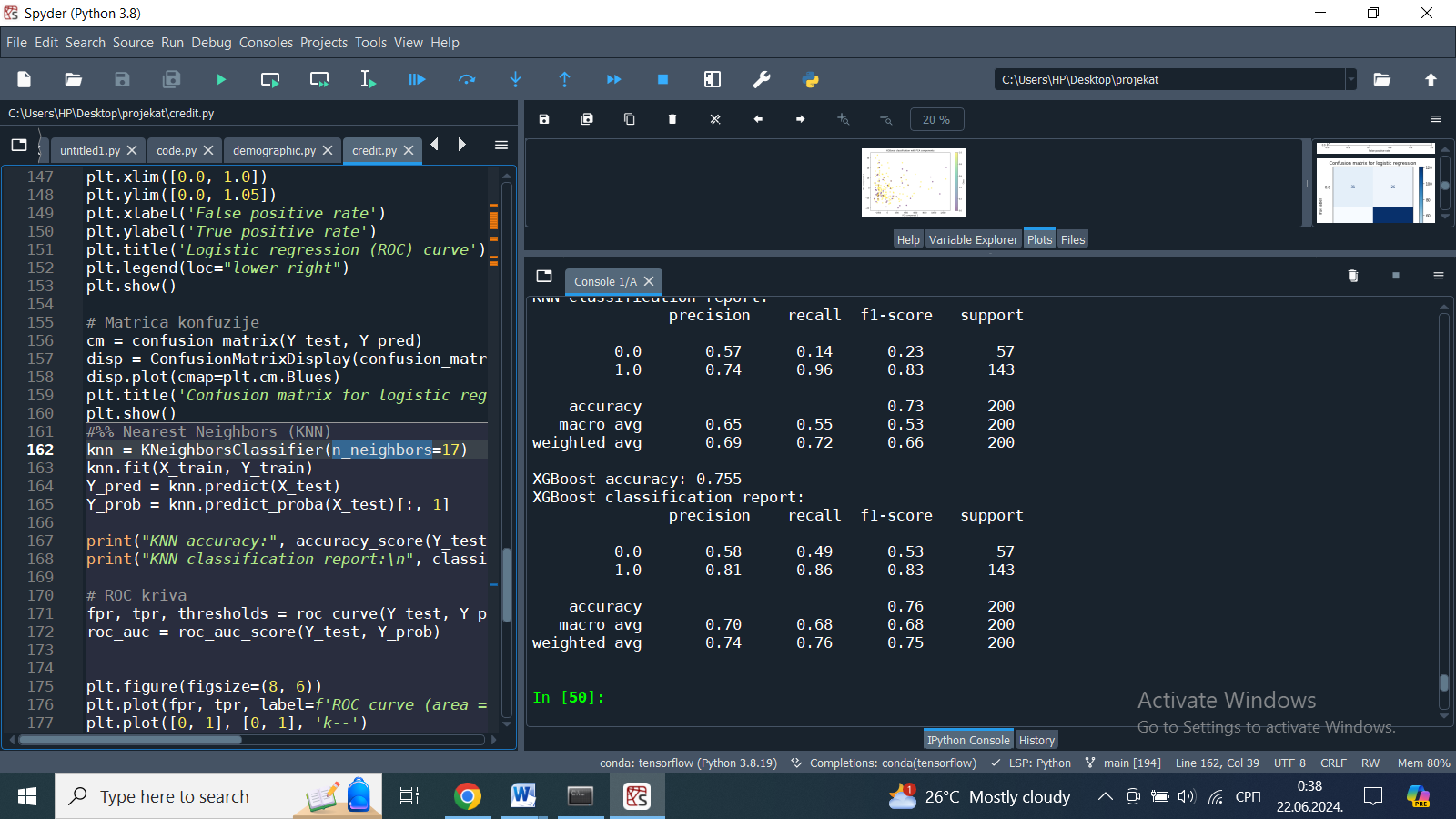
К-најближих суседа



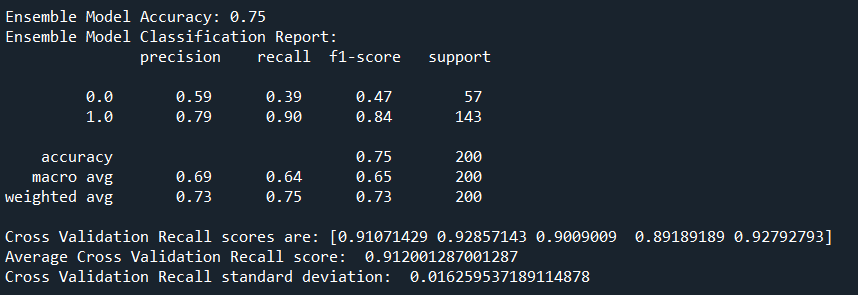
Логистичка регресија



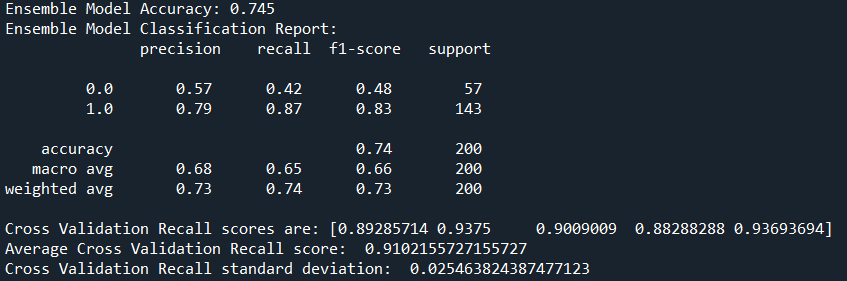
Aлгоритам екстремног појачања градијента (XGBoost)



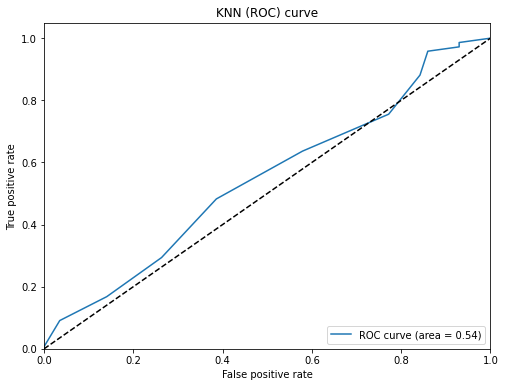
Ансамбл метод



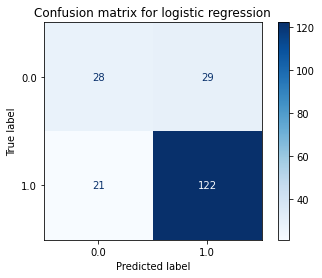
Ансамбл метод (soft voting)



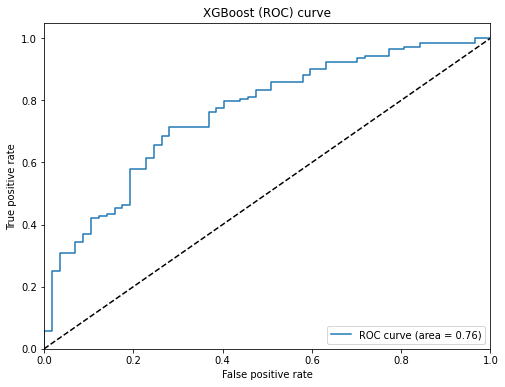
ROC крива за к-најближих суседа



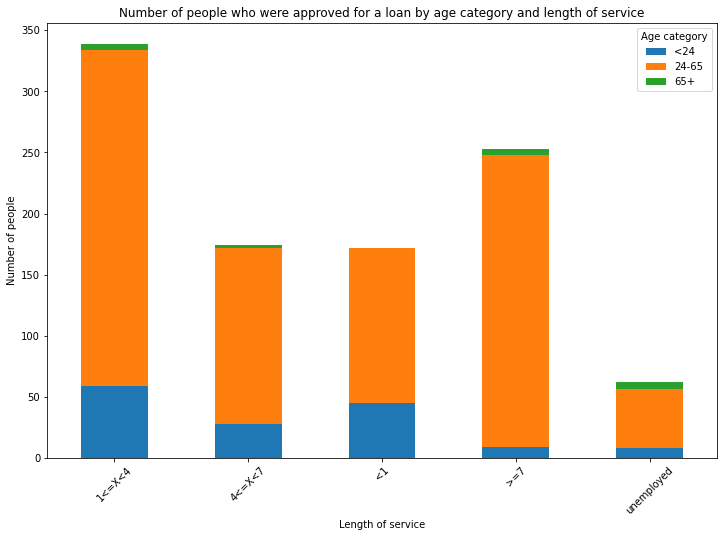
Матрица конфузије за логистичку регресију



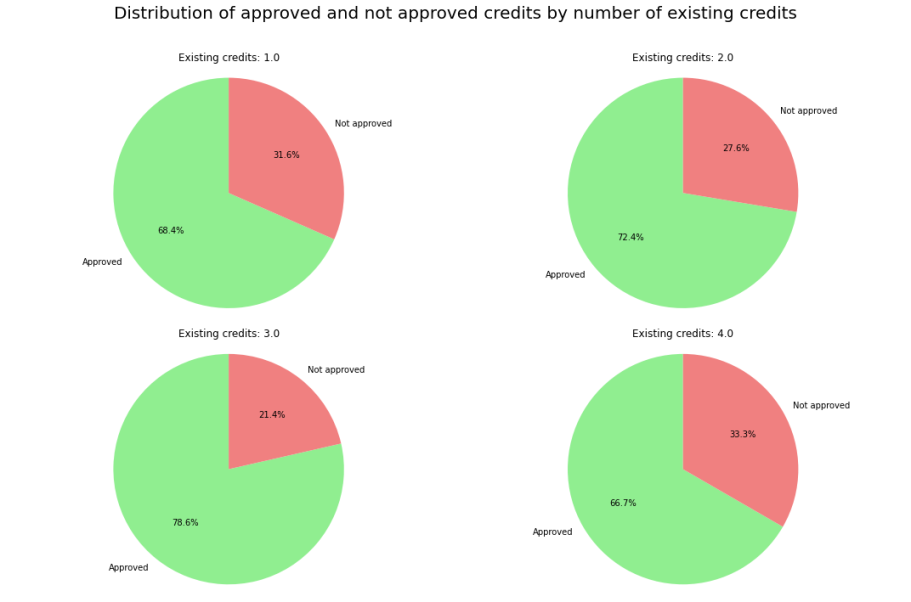
ROC крива за к-најближих суседа aлгоритам екстремног појачања градијента (XGBoost)



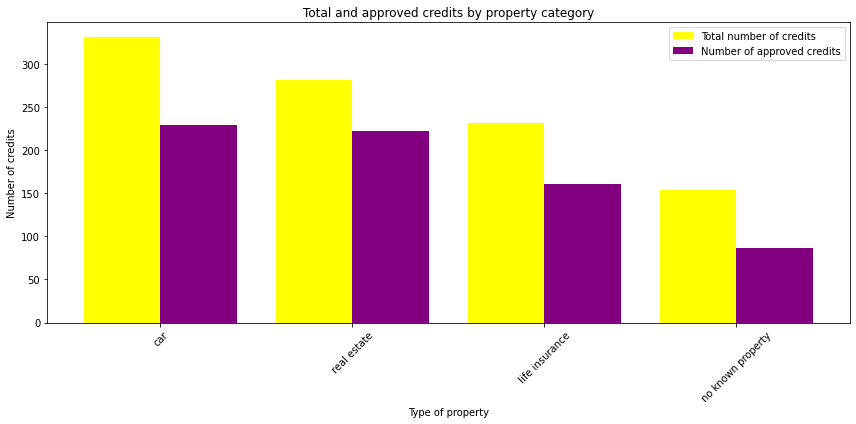
Број људи којима је дозвољен кредит према старосној категорији и радном стажу



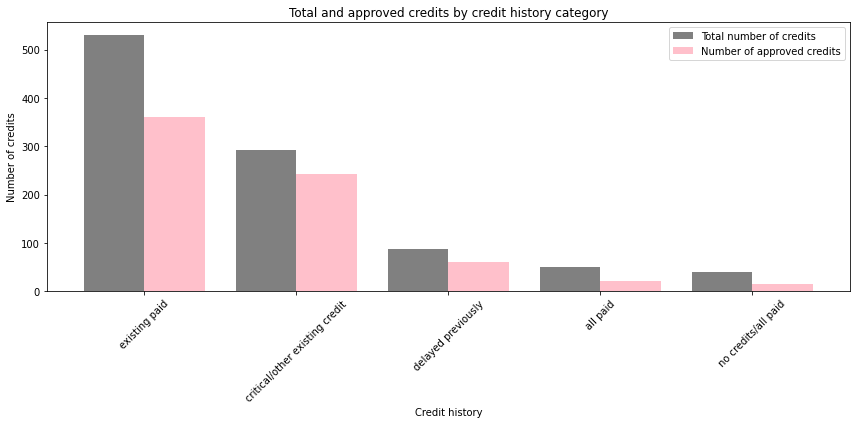
Проценат људи којима је одобрен кредит према броју постојећих кредита



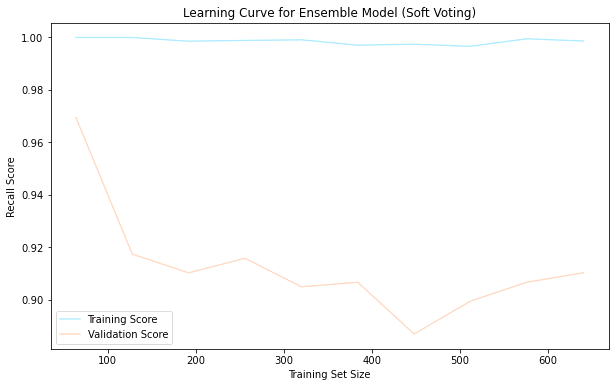
Укупан и одобрен број кредита према типу власништва



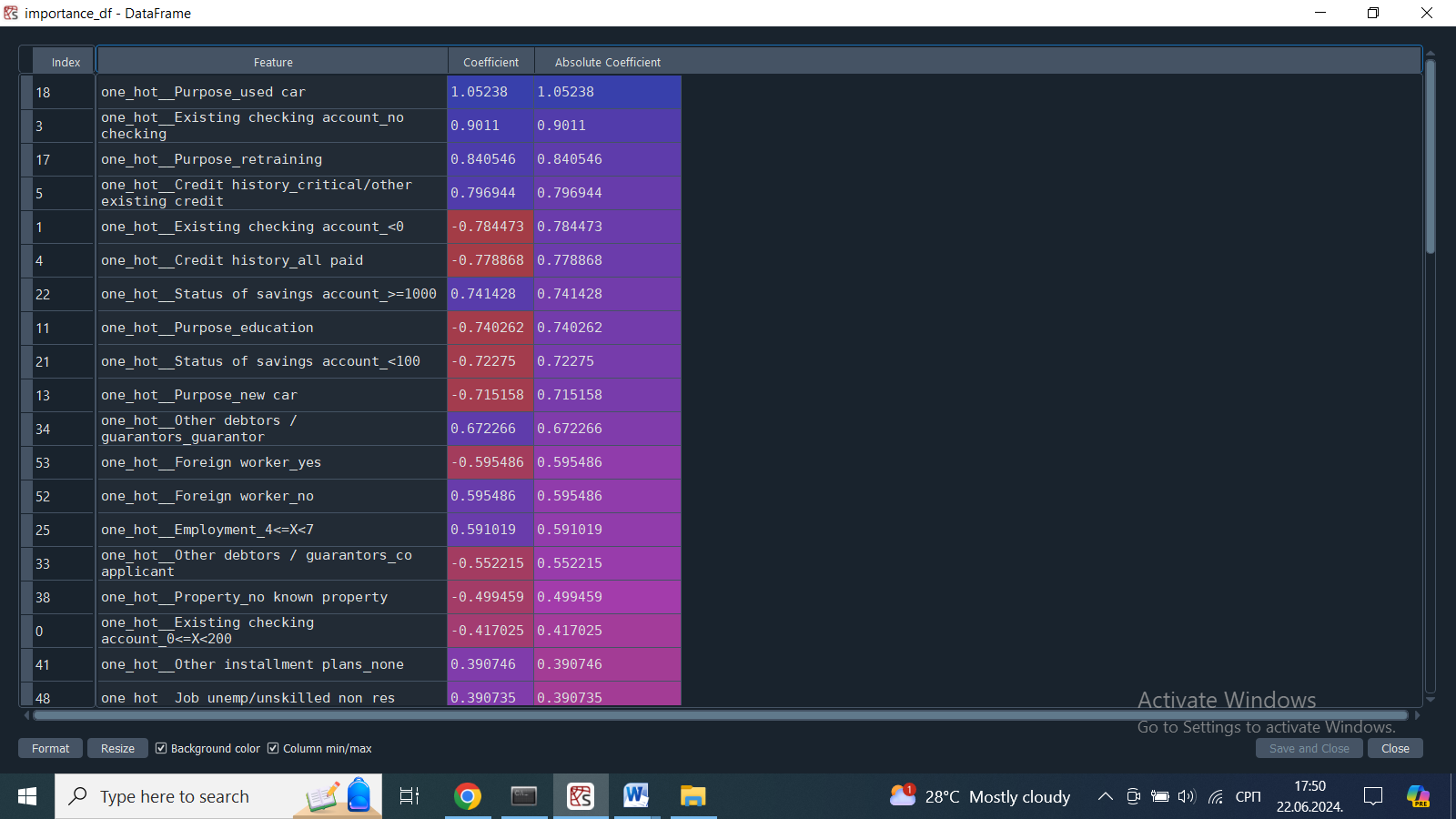
Укупан и одобрен број кредита према историји кредита



Крива учења ансамбл метода



Коефицијенти у логистичкој регресији



## Тумачење резултата

За сваки предиктивни алгоритам је дато тумачење резултата.

К-најближих суседа

Након експериментисања са различитим вредностима параметра n\_neighbors у овом алгоритму, постигли смо најбоље резултате са вредношћу од 17 за број суседа. Почетна прецизност и тачност алгоритма су побољшане порастом овог параметра, достижући тачност од 72.5%, прецизност од 74% и осетљивост од 96% за позитивне класе, док је за негативне класе прецизност износила 57%, а осетљивост 14%. Ови резултати указују на то да је алгоритам боље препознао позитивне инстанце, док је тачност за негативне инстанце била нижа. Примећено је да је даљим повећањем параметра n\_neighbors прецизност алгоритма почела да опада након одређене вредности. за визуализацију перформанси класификационог модела, користили смо ROC кривз (Receiver Operating Characteristic curve). То је граф код којег x оса (False Positive Rate - FPR) представља удео инстанци које су погрешно класификоване као позитивне у односу на све стварне негативне инстанце, а y оса (True Positive Rate - TPR) представља удео инстанци које су тачно класификоване као позитивне у односу на све стварне позитивне инстанце. Рачунају се по следећим формулама:  
FPR= FP​ / (FP+TN), TPR= TP​ / (TP+FN), где су FP (False Positives) број инстанци које су погрешно класификоване као позитивне, а TN (True Negatives) број стварних негативних инстанци. Где су TP (True Positives) број инстанци које су тачно класификоване као позитивне, а FN (False Negatives) број инстанци које су погрешно класификоване као негативне. Површина испод криве представља AUC вредност: што је ближе 1, то је боља способност модела да разликује класе. Ако је 0.5, класификује инстанце насумично, а ако је мања од 0.5 модел је лошији од насумичног класификовања. У нашем примеру је ова вреност 0.54, што значи да модел има одређену способност да разликује између класа али не на високом нивоу.

Логистичка регресија

На основу анализе различитих solver параметара у контексту логистичке регресије за бинарну класификацију, закључујемо да избор solver параметра може значајно утицати на перформансе модела. За наш конкретан проблем, solver = ‘lbfgs’ је показао најбоље резултате у погледу тачности и прецизности предвиђања за бинарну класификацију. Овај solver је ефикасан за средње и велике скупове података, што га чини оптималним избором за нашу анализу кредитног одобрења. Матрица конфузије је предвидела тачно: 28 кандидата као неодговарајућих ѕа издавање кредита (True Negatives – TN) и 122 кандидата као одговарајућих (True Positives - TP). Нетачно је предвидела: 29 кандидата као одговарајућих, а нису били (False Positives - FP), и 21 као неодговарајућих, а требало је обрнуто (False Negatives - FN). На основу овога можемо израчунати тачност = (TP+TN) / (TP + TN + FP + FN) = 0.75, прецизност = (TP) / (TP+FP) = 0.81 и осетљивост = TP / (TP + FN) = 0.85. за позитивне класе (одобрен кредит), што је и приказано у претходном поднаслову, док је за негативне класе прецизност износила 57%, а осетљивост 49%.

Aлгоритам екстремног појачања градијента (XGBoost)

Mењањем параметара у овом алгоритму, успели смо да побољшамо перформансе модела. Повећањем броја стабала, дубине стабла као и стопе учења постигнута је тачност од 75.5%, прецизност од 81% и осетљивост од 86% за позитивне класе, док је за негативне класе прецизност износила 58%, а осетљивост 49%. Већим бројем стабала, као и већом дубином, модел је у стању да научи сложеније обрасце. Код овог алгоритма, вредност испод ROC криве је 0.76, што нам говори да наш модел може релативно добро да разликује класе.

Ансамл метод

Модел који користи гласање већине постиже тачност од 75% која је задовољавајућа. Детаљ који примећујемо је да је у скупу података који користимо заступљена неуравнотеженост класа која може у великој мери утицати на наш модел, посебно када се ради о методи тврдог гласања где може доћи до пристрасности и доводи до лоших перформанси на мањинској класи.

С обзиром да ће одзив на мањинску класу бити низак, многи случајеви ће бити пропуштени и смањиће у озбиљној мери тачност модела. Слично видимо и на Ф1-резултату, где је за мањинску класу резултат знатно нижи.

Ово резултује неконзистентним предикцијама, иако то не бисмо рекли гледајући стандардну девијацију скорова. Ово се дешава јер методом унакрсне валидације ми скуп података делимо на 5 делова, где мањинска класа вероватно није толико изражено мањинска, па се модел сналази боље.

Модел свакако у великој мери успешно идентификује већину позитивних инстанци (од 0.89189189 до 0.92857143).

Ансамбл метод (soft voting)

Модел који користи soft voting се мање сналази, са 74,5% тачности. И даље добро идентификује појединце који су кредитно способни, али и дање настаје проблем са класом 0 односно кредитно неспособним појединцима.

Меко гласање би требало боље да се носи са неуравнотеженошћу класа, али то не мора нужно да се огледа у самој тачности. Конкретно можемо упоредити тачности предвиђања класе 0:

Тврдо гласање: нижи одзив (0.39) али мало виша тачност (0.59)

Меко гласање: виши одзив (0.42) али нижа тачност (0.57)

Тврдо гласање ће можда пропустити нека појављивања класе 0, али то је зато што се ослања на глас већине. Меко гласање боље хвата овакве случајеве јер гледа вероватноће.

Али за предвиђање класе 1 имамо другачију ситуацију:

Тврдо гласање: виши одзив (0.90) и иста тачност (0.79)

Меко гласање: мало нижи одзиб (0.87) и иста тачност (0.79)

Мало смањење одзива за меко гласање вероватно се дешава зато што се процењују вероватноће и одлука ка класи 1 није увек финална јер није поткрепљена када су вероватноће приближне.

Тврдо гласање је једноставније у односу на меко гласање које користи више информација из основних модела. У нашем случају, као тачнији модел показао се модел тврдог гласања, јер када се ради о оваквој категоризацији, нису нам у великој мери биле битне нијансе.

Ипак, прилагођавањем сета података, вероватно бисмо дошли до импресивнијих и тачнијих резултата, уколико бисмо у најмањој мери смањиле неуравнотеженост класа, или бар мањинску класу више пута укључиле у тренинг скуп података.

Број људи којима је дозвољен кредит према старосној категорији и радном стажу

Уочљиво је да за кредит аплицира највише радно способно становништво, тј. људи старости између 24 и 65 година. Знатно мање је људи млађих од 24 године и очекивано најмање пензионера. Такође примећујемо да је највише кандидата са радним стажом између једне и 7 година, а одмах потом кандидати који раде дуже од 7 година.

Проценат људи којима је одобрен кредит према броју постојећих кредита

Постојећи кредити у нашем скупу података могу бити 1, 2, 3 или 4. Графикон приказује, од укупног броја кандидата, проценат којима је дозвољен и одбијен захтев за издавање кредита. Без обзира на број постојећих кредита, удео одобрен је у свакој категорији сличан.

Укупан и одобрен број кредита према типу власништва

Највише кандидата који се пријављују за кредит као некретнину поседују аутомобил, одмах после њих су каднидати који имају неку некретнину, ѕатим људи са животним осигурањем и људи без својине. Интересантно је напоменути да смо пронашле податак да људи са животним осигурањем врло лако добијају кредит, без посебних провера.

Укупан и одобрен број кредита према историји кредита

Људи се према историји кредита, у нашем скупу података, разврставају у оне који тренутно отплаћују кредит, критичне клијенте, на оне који су претходне кредите одложили, који су све отплатили и оне који немају кредит/све плаћено. Највише кадидата тренутно отплаћује неки други кредит, а скоро дупло мање је критичних података, али упркос томе није велика разлика у броју одобрених кредита за ове две категорије.

Коефицијенти у логистичкој регресији

Такође смо посматрале коефицијенте логистичке регресије. Што је апсолутни коефицијент већи, то је утицај конкретне варијабле на резултат значајнији. Можемо уочити да уколико особе желе да купе половни аутомобил, или немају немају текући рачун или уколико су критични кандидати може значајно да утиче на то да ли ће им бити одобрен кредит или не.

Крива учења ансамбл метода

На графикону се види да модел добро предвиђа на мањим скуповима података, али са подацима са којима се није сусретао раније се не сналази. Ово се може дешавати из више разлога као што је комплексност модела (меко гласање јесте комплексан модел), али и чињеница да наши подаци немају једнак број кредитно способних и кредитно неспособних појединаца, о чему ћемо дискутовати касније.

# Закључак

## Анализа испуњења циљева истраживања

Успеле смо да развијемо неколико предиктивних алгоритама који имају релативно задовољавајућу прецизност, тачност и осетљивост. Утврдиле смо да кредитна историја, раѕлог издавања кредита и стање на текућем рачуну представљају неке од значајнијих фактора приликом одобрења кредита.

## Анализа остварења очекиваних резултата истраживања

Наша претпоставка је била да ће међу најзначајнијим факторима за одобрење кредита бити приходи, статус запослености и кредитна историја, кредитно задужење, некретнине. Резултати су показали да се кандидату повећавају шансе за одобрење уколико су му претходни кредити исплаћени и уколико има висока примања.

## Могућности за примену истраживања у пракси

Резултати истраживања се могу применити у банкарском сектору. Запосленим би била олакшана процена кандидата за издавање кредита. Анализа утицаја различитих променљивих на вероватноћу одобрења кредита омогућава банкама да боље разумеју факторе ризика и профитабилности. Ови резултати могу бити интегрисани у постојеће системе за доношење одлука како би се побољшала тачност процена кредитне способности кандидата. Разумевање које променљиве имају највећи утицај на вероватноћу одобрења кредита омогућава банкама да развију стратегије за смањење ризика од невраћања кредита. На основу анализе променљивих, банке могу креирати персонализоване кредитне понуде које боље одговарају потребама и профилу сваког клијента.

## Идеје за побољшање и разраду истраживања

Прикупљање већег скупа података, укључујући дужи временски период и већи број апликаната, може повећати репрезентативност и робусност резултата. Такође, интеграција података из различитих извора (нпр. кредитни бирои, социјални медији) може пружити свеобухватнији увид у профиле клијената. Испитивање додатних променљивих које могу утицати на одобрење кредита, као што су демографски подаци, понашање у потрошњи, и економски индикатори, може побољшати модел и пружити детаљније информације за доношење одлука.

# Литература

1. https://towardsdatascience.com/a-simple-interpretation-of-logistic-regression-coefficients-e3a40a62e8cf
2. https://stackoverflow.com/questions/38387913/reason-of-having-high-auc-and-low-accuracy-in-a-balanced-dataset
3. <https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html>
4. https://grading.rs/sta-je-kreditna-sposobnost-i-kako-je-pravilno-izracunati/