შავი ზღვის საერთაშორისო უნივერსიტეტი
კომპიუტერული ტექნოლოგიებისა და საინჟინრო საქმის ფაკულტეტი
დოქტორანტურის პროგრამა

მონაცემთა დამუშავების გამოყენება კომერციულ ბანკებში კრედიტების დათვლის ოპტიმიზაციისთვის R პროგრამული ენის გამოყენებით

დილმუროდჯონ ზაკიროვ
ინჟინერიის დოქტორი ინფორმატიკაში სადოქტორო დისერტაციის ავტორეფერატი

თბილისი-2016

სამეცნიერო ხელმძღვანელი: პროფესორი ნოდარ მომცელიძე

(შავი ზღვის საერთაშორისო უნივერსიტეტის პროფესორი, დოქტორი)

(ხელმძღვანელის ხელმოწერა)

ექსპერტები (სახელი და გვარი, აკადემიური ხარისხი)

1. პროფ. დოქტ. ალექსანდრე მილნიკოვი
2. ასოც. პროფ. დოქტ. მიხეილ რუხაია

ოპონენტები (სახელი და გვარი, აკადემიური ხარისხი)

1. ასოც.პროფ. აბზეთდინ ადამოვი
2. პროფ.დოქტ. ზურაბ ბოსიკაშვილი
3. პროფ.დოქტ. სერგო ცირამუა

შესავალი

მონაცემთა ინტელექტუალურმა ანალიზმა ან მონაცემთა ბაზებში მონაცემთა მოპოვებამ (KDD), მიიპყრო უკვე მრავალი სამეცნიერო სფეროს ყურადღება. ბოლო ორი ათწლეულის განმავლობაში დიდი მონაცემები განსაკუთრებით განვითარდა. მონაცემთა ასეთი გაფართოება არ იქნებოდა შესაძლებელი რომ არ გამოყენებულიყო მონაცემთა ინტელექტუალური ანალიზის მეთოდი და ალგორითმები, რომლებიც შემუშავებულ იქნა აქამდე. ყველა ეს მეთოდი და ალგორითმი გამოიყენება სწრაფად მზარდი მონაცემთა უზარმაზარი ბაზიდან ცოდნის მისაღებად.

მონაცემთა ბაზები შედგება მილიონობით ჩანაწერისაგან და უკვე ჩატარებული იქნა მრავალი კვლევა ამ მონაცემების მიხედვით იმისათვის, რომ დადგენილიყო არის თუ არა რამე მალული ფაქტები, რომლებიც ჯერ კიდევ უცნობია. რამდენადაც ეს ჩანაწერები და მონაცემები მნიშვნელოვნად იზრდება და ამავე დროს მონაცემთა ბაზების ზრდას მივყავართ ახალი მეთოდებისა და კვლევების შემუშავებისაკენ, რათა მიღწეულ იქნას უკეთესი შედეგები ისეთი მსხვილი კომპანიების მიმდინარე და სამომავლო მოთხოვნილებების შესასრულებლად, რომლებიც ეძებენ ასეთ გადაწყვეტილებებს. არის რამდენიმე სფერო, სადაც ეს მოთხოვნილება განსაკუთრებით იგრძნობა, მაგალითად გაყიდვა/მარკეტინგი, მყიდველთა ქცევები, თაღლითობის შემთხვევების გამოვლენა, საკრედიტო სკორინგი და ა.შ.

მონაცემთა შენახვისა და ანალიზში პროგრესი ზემო აღნიშნულ სფეროებში ორგანიზაციები მიჰყავს გარკვეულ სირთულეებამდე, როდესაც ხდება დიდი მოცულობის მონაცემთა დამუშავება და ინტერპრეტაცია, რაც მას აქცევს სასარგებლო ინფორმაციად და ცოდნად.

კვლევის საკმაო პერიოდის შემდეგ ნათელი გახდა, რომ მონაცემთა ინტელექტუალური ანალიზი წარმოადგენს იმ მიდგომას, რომელიც საჭიროა ამ რთული მოთხოვნების დასაკმაყოფილებლად. ამ პროცესის წამოჭრა უშუალოდ იყო დაკავშირებული ყველა ამ ფარულ რისკებთან და უცნობ არააშკარა ინფორმაციასთან, რომელმაც იცრუა მონაცემთა დიდი მოცულობის ბაზების შემთხვევაში.

მონაცემთა ინტელექტუალური ანალიზი ძირითადად ეხმარება კლასიფიკაციასთან კლასტერიზაციასთად და ასოციაციის წესებთან დაკავშირებული

პრობლემების გადაჭრაში. რამდენადაც რეალურ დროში წამოიჭრა მონაცემთა მოპოვების საკითხი, უამრავი მეთოდი და ალგორითმი იქნა აგებული მონაცემთა ბაზებიდან ცოდნის გამოსაყოფად. მოცემული ნაშრომი აგრეთვე კონცენტრირებული იქნება კლასიფიკაციის მოდელზე, როგორც უმრავლესობა მეთოდებისა და ალგორითმებისა შემუშავებულ იქნა მონაცემთა ბაზებიდან ცოდნისა და ინფორმაციის მისაღებად. კლასიფიკაცია წარმოადგენს მონაცემთა ინტელექტუალური ანალიზის ფუნქციას, რომელიც მიიკუთვნებს კოლექციაში მიზნობრივი კატეგორიის ელემენტებს ან კლასებს.

(http://docs.oracle.com/cd/B28359_01/datamine.111/b28129/classify.htm#i1005746)

კლასიფიკაციის მიზანს წარმოადგენს მონაცემებში თითოეული შემთხვევისათვის ზუსტად განისაზღვროს მიზნობრივი კლასი. მაგალითად: კლასიფიკაციის მოდელი შეიძლება გამოყენებულ იქნას როგორც დაბალი , ისე საშუალო ან მაღალი საკრედიტო რისკების მქონე მსესხებლის საიდენტიფიკაციოდ, რომელიც მოცემულია ჩვენს ახლანდელ კვლევებში. კლასიფიკაციაში ძირითადად გამოყენებულ იქნა ასეთი სახის მიდგომა: მონაცემების კლასები წინასწარ არის განსაზღვრული, მონიშნული ობიექტების კვლევების ნაკრები გამოყენებულია მოდელის ფორმირებისათვის კლასიფიკატორის მეშვეობით მომავალი დაკვირვებისათვის, ასევე ცნობილი როგორც ტესტური ნაკრები. მაგალითად: კლასიფიკაციის მოდელი, რომელიც პროგნოზირებს საკრედიტო რისკებს, შესაძლებელია შემუშავდეს დაკვირვების ქვეშ მყოფი მსესხებელთა მონაცემების საფუძველზე დროის გარკვეული პერიოდის განმავლობაში. ისტორიულ საკრედიტო რეიტინგზე მონაცემებს დამატებით შეუძლიათ თვალი ადევნონ დასაქმების ისტორიას, საკუთრებაში არსებულ საცხოვრებელს ან საიჯარო გადასახადს, ცხოვრების წლებს, ინვესტიციის ნომერს და სახეს და ა.შ. საკრედიტო რეიტინგი იქნება მიზანი, სხვა ატრიბუტები იქნება წინასწარ განმსაზღვრელი და თითოეული კლიენტისათვის მონაცემები იქნება შემთხვევა. დამუშავების ამ სახეს უწოდებენ კონტროლირებად სწავლებას. კლასიფიკაციის ამოცანების ამოსახსნელად როგორც წესი იყენებენ სწავლების კონტროლირებად მეთოდებს .

დისერტაციის სტრუქტურა

ნაშრომი შედგება შემდეგი ნაწილებისაგან :

პირველად ჩვენ გავაკეთეთ - ჩვენი თემის შესავალი

პირველი ნაწილი შედგება მე-1 და მე-2 თავებისაგან.

მე-1 თავი წარმოადგენს ამ ნაშრომთან დაკავშირებული ლიტერატურის მიმოხილვას, მე-2 თავი მოიცავს კონტროლირებადი სწავლების არსებული მოდელების დეტალურად შესწავლას.

მეორე ნაწილი შედგება მე-3 და მე-4 თავებისაგან.

მე-3 თავი იქნება კონცეფციის მტკიცებულება. ის იქნება დაფუძნებული მონაცემთა ტესტურ კრებულზე ცვლადის ზუსტი ოდენობით, როგორც მონაცემთა საწარმოო ნაკრებში, მაგრამ უფრო ნაკლები რაოდენობის მონაცემებით. ეს ერთ-ერთი გაცილებით მნიშვნელოვანი ნაწილია, რადგან ყველა ჩვენი მოთხოვნა აქ იქნება განხილული პრაქტიკული თვალსაზრისით. კონცეფციის დადასტურების განმავლობაში ჩვენი მონაცემთა წყარო იქნება მძიმით გამოყოფილი ნიშანი. შედეგი იმოქმედებს მთლიან კვლევაზე.

მე-4 თავი იქნება მე-3 თავში მუშაობის შედეგის რეალიზაცია. ჩვენ გამოვიყენებთ მონაცემთა საწარმოო ნაკრებს კლიენტების შესახებ საკმარისი რაოდენობის მონაცემებით. ეს მონაცემები შეინახება მონაცემთა ბაზა - Oracle-ში. და ჩვენ განვახორციელებთ მხოლოდ გაცილებით შესაბამის და ზუსტ მოდელს, რომელიც შევარჩიეთ მე-3 თავში. ეს თავი მოიცავს აგრეთვე ექსპერიმენტულ შედეგებს და მათ შეფასებებს, რომლებიც გამომდინარეობს თვითონ ამ ნაწილიდან. ბოლოს ჩვენ შევაფასებთ პროგნოზებს და შევხედავთ საკრედიტო სკორინგის არსებული სისტემის ოპტიმიზაციას, რამდენადაც ეს წარმოადგენს ამ ნაშრომის ერთ-ერთ მთავარ მიზანს. ოპტიმიზაციის ამოცანის ფარგლებში, ჩვენ შეგვიძლია განმეორებით განვიხილოთ საკრედიტო პროცესებში ის ცვლილებები, რომელთაც შეუძლიათ დაეხმარონ კრედიტის მიხედვით უფრო მეტი წარმატების მიღწევაში შემოსავლების მხრივ.

მესამე ნაწილი არის დასკვნა.

ის მოიცავს სამეცნიერო- კვლევით სამუშაოებს, რომელშიც დართულია ნაშრომის მომავალი მიმართულებები და შეთავაზებები. ის აგრეთვე მოიცავს დანართებს და მთელი კვლევის მანძილზე გამოყენებულ წყაროებს.

მეთოდოლოგია

მოცემული ნაშრომი წარმოადგენს ემპირიულ კვლევას, რომელიც მოიცავს დიდი რაოდენობით ექსპერიმენტულ სამუშაოებს, რომლებიც სრულდება კონტროლირებადი სწავლების რამდენიმე მეთოდის შედარებით.

მოკლედ რომ ვთქვათ, ჩვენ შეგვიძლია ჩამოვთვალოთ ყველა კვლევა შემდეგი სახით:

#	სამუშაოს აღწერა:
	წინასწარი კვლევებისა და სამუშაოების შეფასება: კვლევები მონაცემთა ინტელექტუალური ანალიზის სფეროში კომბინირებული R პროგრამირებასთან ჯერ კიდევ იმყოფება საკრედიტო სკორინგის ძალიან ახალ და წინასწარი მომზადების ეტაპზე. ჩვენ ვიკვლევდით ლიტერატურაში არსებულ მეთოდებს. არსებული ინფორმაცია მოცემული კვლევებისათვის განხილულია მე-2 თავში.
	საკრედიტო სკორინგის ახალი მოდელის კონცეფცია და დიზაინი მსესხებლის შესახებ ინფორმაციის მონაცემთა ნაკრები იქმნება ისეთი სახის, რომელიც შეიცავს სრულ ინფორმაციას, მაგრამ არ გამოიყენება ისე, როგორც ის არის. ეს იქნება მონაცემთა ნაკრების ჩვენი პროდუქცია. მთელი მონაცემთა ნაკრები მდებარეობს Oracle მონაცემთა ბაზაში. შესაბამისად მონაცემები იკითხება Oracle -ის მონაცემთა ბაზაში და გადაეცემა R-ს. მოგვიანებით მონაცემთა ეს ნაკრები გაიყოფა ორ ტიპად: მონაცემთა ნაკრების მომზადება და ტესტირება. მონაცემთა ეს ნაკრებები

	გამოყენებულ იქნება პროგნოზირების მიზნით.
	რეალიზაცია და ტესტირება: თითოეულ მოდელის განხილვის პერიოდში, გამოყენებულ იქნება მონაცემთა რამდენიმე აქტივობა, როგორიცაა მომზადება და გაწმენდა. ცვლადები განსაზღვრულია შესაბამის კლასებში.
	შეფასება: ყველა განხილული კონტროლირებადი შესწავლის მოდელისა და შერჩეული საუკეთესო მეთოდის შეფასების შემდეგ ჩვენ განვახორციელებთ ამ მეთოდს უკანასკნელ და ყველაზე სრულ მონაცემთა ნაკრებზე. ჩვენ შევადარებთ განვითარების ყოველ ეტაპზე შერჩეულ ყველა მონაცემთა ინტელექტუალური ანალიზის მოდელების შედეგებს, რომლის შემდეგაც დავინახავთ შედეგის ხარისხსა და სიზუსტეს. ყველა შეფასება კეთდება R პროგრამირების ენაზე.
	გავრცელება: ჩვენ ვავრცელებთ ამ კვლევის შედეგებს რამდენიმე სტატიის წარმოდგენითა და ანალიტიკურ სფეროებში ექსპერტებთან ანალიზის გზით, რაც დაკავშირებულია საკრედიტო სკორინგის უკუკავშირთან. განხილვის პროცესი საშუალებას მოგვცემს გავაუმჯობესოთ ამ მოდელის შემდგომი განვითარება.

ცხრილი 1: დისერტაციაში გამოყენებული კვლევის მეთოდოლოგიის პრაქტიკული მიდგომა:

ზემოთ ჩვენ მოკლედ აღვწერეთ წარმოდგენილ დისერტაციაში სამუშაოსადმი პრაქტიკული მიდგომა. მომდევნო თავებში ჩვენ განვიხილავთ უფრო დეტალურად თითოეულ მოდელისა და მეთოდს აღწერით და შედეგების ინტერპრეტაციას.

კვლევის მიზანი

დისერტაციის მიზნის დასამტკიცებლად, გაანალიზებული იქნება კონტროლირებად სწავლებისათვის სხვადასხვა პროცესები და ახალი განვითარებული მოდელები წარმოდგენილი იქნება ზუსტი შედეგებით, რომლებიც აღებულია რეალური ცხოვრების მონაცემთა ნაკრებიდან.

იმისათვის, რომ უკეთესად მოხდეს ზოგიერთი მოდელირების შედეგის ვიზუალიზაცია , ყველა გაანალიზებულ და წარმოდგენილ მოდელს თან დაერთვება გრაფიკები.

შესწავლის სიახლე

იმისათვის, რომ შეთავაზებულ იქნას ისეთი ახალი ალგორითმები და / მოდელები, რომლებიც ორიენტირებული იქნებიან მონაცემთა ინტელექტუალური ანალიზის მეთოდზე და R პოგრამირების ენაზე და Oracle მონაცემთა ბაზა განვითარებისა და ოპტიმიზაციის მიზნით KDD (Knowledge Discovery in Databases) ბანკის საკრედიტო სკორინგის სისტემის დამუშავება. ეს არის სიახლე, რომელსაც კონკრეტული სწავლება მოიტანს.

სამეცნიერო და პრაქტიკული მნიშვნელობა

ჩვენ შეგვიძლია დავყოთ ამ სადისერტაციო ნაშრომის წვლილი შემდეგ კატეგორიებად:

წვლილი მონაცემთა ინტელექტუალურ ანალიზში

ამ ნაშრომში სხვადასხვა მეთოდები გამოყენებულია მთელ მსოფლიოში კონტროლირებადი სწავლებისა და სხვადასხვა ექსპერიმენტებში, გადაწყვეტილებათა ხე (კლასიფიკაციისა და რეგრესიის ხე, C5.0, Random Forests, CHAID, ხეების ჯგუფი), ნერვული ქსელები (მრავლობითი დამალული ფენები, SVM და ღრმა სწავლება), ლოჯისტიკური რეგრესია, kNN კლასიფიკაცია, Bayesian კლასიფიკაცია (Tree Augmented Network), ექსპერტთა ჯგუფი. ექსპერიმენტის განმავლობაში მთელი ძალისხმევა მიმართული იყო ამ მოდელებთან დაკავშირებულ სხვადასხვა ფაქტორების გამოკვლევაზე. აგრეთვე სხვადსხვა მოდელები, რომელიც დაფუძნებულია სხვადასხვა განზომილებებზე და მიღებულია დიდი ექსპერიმენტების შედეგად.

შედეგი გვიჩვენებს, რომ მონაცემთა ინტელექტუალური ანალიზის სხვადასხვა მოდელებს თან ერთვის მონაცემთა ნაკრები, რომელიც შეიცავს რამდენიმე ცვლადს, შეუძლია გვაჩვენოს განსხვავებული სიზუსტის კურსები. ამრიგად შეიძლება ვივარაუდოთ, რომ მონაცემთა ნაკრების ცვლადები და გამოყენებული მოდელის სირთულეებს შეიძლება ჰქონდეს სხვადასხვა შედეგი გამოსავალში.

წვლილი ბანკებისა და ფინანსური დაწესებულებებისათვის

როგორც მითითებულია „პრობლემა და მოტივაციაში“ ნაშრომის ნაწილში ნებისმიერი ფინანსურ დაწესებულებასა და ბანკს აქვს უზარმაზარი მოთხოვნილება საკრედიტო სკორინგის სისტემაში. მზა დანართის ყიდვა ყოველთვის შესაძლებელია, მაგრამ აქვს გარკვეული უარყოფითი მხარეები, როგორიც არის : ფასწარმოქმნა და ლიცენზირება, დაყენების და გამოყენების სირთულე. თუკი ფინანსური დაწესებულება ან ბანკი არც ისე დიდია და ა.შ. **სიზუსტის უმაღლესი ხარისხით მოდელის შექმნისას პროგნოზირებისათვის ის შეიძლება მოგვიანებით გამოყენებულ იქნას ნებისმიერი ზომის ორგანიზაციაში მომავალი გამოყენებისათვის , როგორც ცვლადები, რომლებიც გამოიყენება სატრენინგო და სატესტო მონაცემთა ნაკრებებში, თითქმის ერთნაირია ყველა ფინანსური ინსტიტუტებისა და ბანკისათვის.** სხვა მნიშვნელოვანი ფაქტორი, რომელიც აღსანიშნავია მონაცემთა ნაკრები ინახება და

ითვლება Oracle-ს მონაცემთა ბაზიდან, რაც ძალიან მნიშვნელოვანია, რადგან მონაცემთა ბაზა Oracle წარმოადგენს დეფაქტო სტანდარტს, რომელიც გამოყენებულია მთელი მსოფლიოს ბანკებში. **ამრიგად, მოდელი შეიძლება განვიხილოთ, როგორც ზოგადი დანიშნულების და შეიძლება გამოყენებულ იქნას ნებისმიერ საბანკო დაწესებულებაში საკითხების გადასაწყვეტად ან არსებული ოპტიმიზაციისათვის.**

სამუშაოს სტრუქტურა და მოცულობა

ნაშრომი 120 გვერდია და შედგება ხუთი თავისაგან, ლიტერატურის , ნახატებისა და კოდების ჩამონათვლებისაგან.

თავი 1 ლიტერატურის მიმოხილვა

მოცემული ლიტერატურის მიმოხილვის მთავარი მიზანი იყო მონაცემთა ინტელქტუალური ანალიზის, R დაპროგრამირების ენასა და კრედიტების დათვლის ტექნოლოგიების შესახებ ინფორმაციის მოძიება და ანალიზი.მონაცემთა ინტელექტუალურის ანალიზს გააჩნია უამრავი შესაძლებლობები რომლებიც შეიძლება გამოყენებული იქნეს პრაქტიკულად, R დაპროგრამირების ენა იმყოფება თავის საწყის ეტაპზე და უახლოეს მომავალში აჩვენებს მის შესაძლებლობებს , ასევე კრედიტების დათვლის ტექნოლოგიებიც ვითარდება ყოველწლიურად

თავი 2 კვლევაში გამოყენებული მონაცემთა მოპოვების მეთოდები და მოდელები :

დღეისათვის არსებობს მრავალი ალგორითმი და მეთოდი, რომელიც გამოიყენება კონტროლირებადი სწავლებისათვის. ძირითადი პრობლემა სპეციფიური პრობლემის შემთხვევაში მონაცემთა ამოღებისათვის შესაფერისი ალგორითმის პოვნაა. ფინანსურ დაწესებულებებშიც მსგავსი სიტუაციაა საკრედიტო სკორინგთან მიმართებაში. ამან ჩვენ მოგვცა მოტივაცია, გამოგვეკვლია და გაგვეანალიზებინა ის ფაქტორები, რომლებიც ზეგავლენას ახდენენ კონტროლირებადი სწავლების შესაბამისი ალგორითმის შერჩევაზე. აგრეთვე მოგვეხდინა საკრედიტო სკორინგის ზოგადი

პროცესის ოპტიმიზაცია მონაცემთა ინტელექტუალური ანალიზის ამ მეთოდების გამოყენებით პროგრამულ ინტერფეისებთან ერთად. *საკრედიტო სკორინგი წარმოადგენს ბანკის მყარი განვითარების ერთ-ერთ მთავარ ფაქტორს კლიენტების რაოდენობის გასაზრდელად და სარგებლის მისაღებად.* მაგრამ იმისათვის, რომ მიაღწიო წარმატებას ამ მხრივ, უნდა იყოს ეფექტური საშუალება და /ან მექანიზმი, რომელიც განიხილავს კლიენტის საკრედიტო განაცხადის ყველა შესაძლო შედეგს და ეს სკორინგი, როგორც არაეფექტური სკორინგი, გამოიწვევს უარყოფით შედეგს, როგორიცაა რეალურად არფუნქციონირებადი სესხი, მომხმარებლის ერთგულების დაკარგვა და რა თქმა უნდა, შემოსავლებისა და სარგებლის შემცირება. მიუხედავად იმისა, რომ ბანკი არ არის ისეთი დიდი ზომით და საკრედიტო განაცხადებიც არც ისე ბევრია, მაგრამ ეს არ ნიშნავს იმას, რომ არ არის მოთხოვნა ეფექტური კრედიტუნარიანობის შეფასებაზე. რამდენადაც იზრდება მომხმარებელთა რაოდენობა და სესხის სახეები, საკრედიტო სკორინგის გაზრდა ნებისმიერი ბანკისათვის სავალდებულო ხდება მათი ზომებისაგან დამოუკიდებლად. სისტემა არ უნდა იყოს დამოკიდებული ადამიანურ ფაქტორზე. *რა თქმა უნდა, გადაწყვეტილება მიიღება ადამიანის მიერ და არა პროგრამული სისტემით, ადამიანს უნდა ჰქონდეს კარგი პროგრნოზირების უნარი, რომელიც დაფუძნებული იქნება კლიენტის საკრედიტო განაცხადიდან მიღებულ რაოდენობრივ და ხარისხობრივ მონაცემებზე.* აუტონომიური პროგრამული სისტემის გამოყენება არ იძლევა ასეთ შედეგს იმ სისტემებთან შედარებით, რომლებიც შემუშავებულია სტატისტიკური სისტემების საფუძველზე და რეალიზებულია ანალიტიკური ინსტრუმენტების გამოყენებით.

ამიტომ მიღებულ იქნა გადაწყვეტილება, დანერგილ იყო ქულების დათვლის სისტემა, რომელიც აგებულ იქნება მონაცემთა ინტელექტუალური ანალიზის მეთოდების სახის მიხედვით და შემუშავდება პროგრამირების ენის სტატისტიკურ საფუძველზე.

ეს გამოკვლევები, პირველ რიგში, ფოკუსირებულია კონტროლირებადი სწავლების სხვადასხვა მეთოდების ანალიზზე. მეორე რიგში კი, არსებული სატრენინგო და

სატესტო მონაცემთა ნაკრებზე მუშაობის შემდეგ, ის გვთავაზობს ახალ, გაუმჯობესებულ და ოპტიმიზირებულ პროცესს ცოდნის ამოსაღებად. ამ პერიოდის განმავლობაში მონაცემთა ნაკრებებით ექსპერიმენტთა დიდი რაოდენობა ტარდება იმისათვის, რათა გაირკვეს რომელ მოდელს როგორი სიზუსტის ხარისხი აქვს. მესამე რიგში, ეს კვლევა ფოკუსირებულია ცოდნის განვითარებაზე და გადაწყვეტილება დაფუძნებულია საკრედიტო სკორინგის სისტემაზე.

თავი 3 კონცეფტის დამტკიცება- სატესტო მონაცემებზე დაფუძნებით შერჩეული მეთოდების შემოწმება და ექსპერიმენტების ჩატარება:

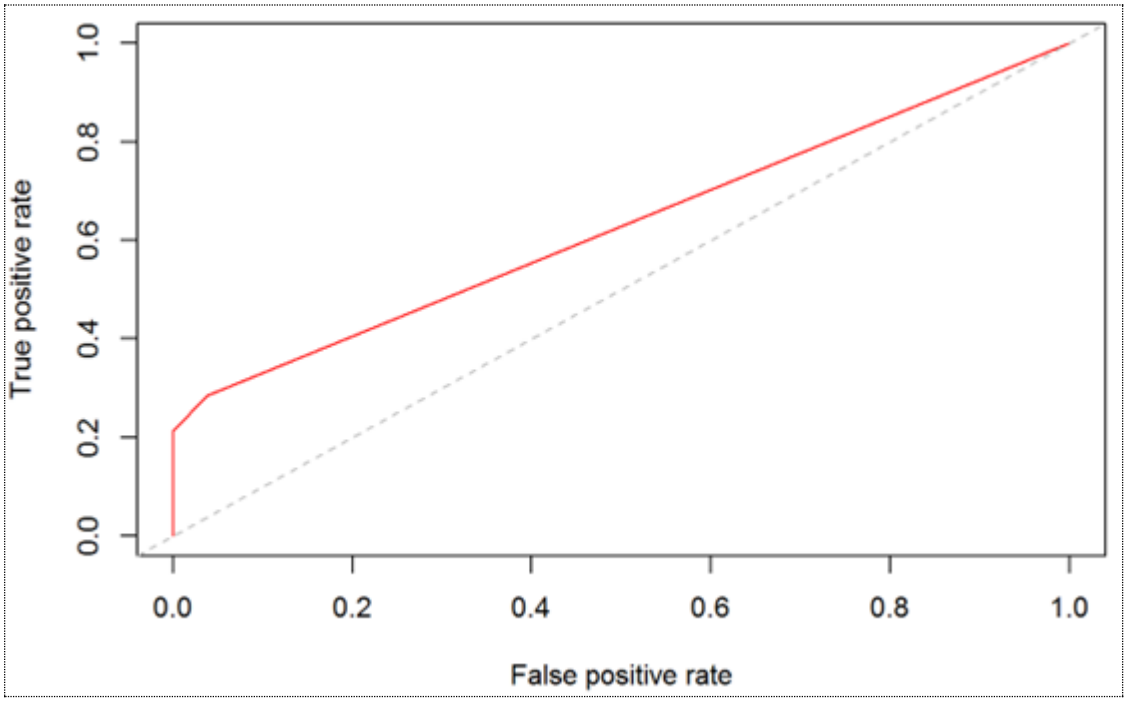
მოდელირებისას ჩვენ ვიყენებთ უნიკალურ მიდგომას; მონაცემთა წაკითხვა CSV ფაილიდან, მონაცემთა ნაკრების აუცილებელი ნაწილების გაწმენდა, მონაცემთა მზა ნაკრების დაყოფა სატრენინგო და სატესტო მონაცემებად, R ენის სპეციფიური ბიბლიოთეკების გამოყენებით მოდელის აგება, შედეგების ვიზუალიზირება გრაფიკის სხვადასხვა უბნების გამოყენებით, არეული მატრიცის გამოყენება, სადაც ეს შესაძლებელია და ბოლოს შედეგის ინტერპრეტაცია და გადაწყვეტა , გამოდგება თუ არა მონაცემთა ინტელექტუალური ანალიზის ეს მეთოდი ჩვენი მოდელისათვის.
პირველი, რასაც ჩვენ ვაკეთებთ ყველა მოდელში, არის მონაცემთა წაკითხვა. ყველა მოდელის ნაწილების შედარებითი ანალიზისათვის ვაპირებთ წავიკითხოთ მონაცემები CSV ფაილიდან. შემდეგ ჩვენ აღვნიშნავთ თითოეულ სვეტში მონაცემს სპეციფიური ცვლადით, ამით ჩვენ შეგვიძლია მათი იდენტიფიცირება მოდელში. შემდგომი ეტაპები არის მონაცემთა გაწმენდა და მოსამზადებელი მოქმედებები. აქ მუშავდება შეუსაბამო მონაცემები ვალუტური ცვლადების დონის დასადგენად. შეიქმნა ორი ახალი ცვლადი **“Pldg_Cur”** და **“Pldg_Pledge”,** რომელიც მიღებულია **“Pledge”**- ცვლადიდან და შედგება რიგი თავდებისა და გირაოსაგან.

კრედიტის მიზანი შედგება მრავალი დონისაგან და მნიშვნელობები მიეკუთვნა კატეგორიებს ახალ მიღებულ ცვლად **“loanPurpose”**-ის ქვეშ. მაშასადამე, **loanPurpose** - ცვლადში არის მხოლოდ რამდენიმე დონე, საკრედიტო ისტორიისათვის

მონაცემები, მსესხებლის ცუდი გადამხდელუნარიანობის შესახებ აღინიშნება 1-ით, ხოლო სხვები 0-ით. ეს მონაცემები ითვლება კრედიტის გაცემაზე უარის ან თანხმობის თქმის მთავარ მიზეზად.

იმისათვის, რომ მონაცემები ხელმისაწვდომი იყოს მოდელის შესაქმნელად, ის იყოფა სატრენინგი და სატესტო მონაცემთა ნაკრებებად. სატრენინგო მონაცემები გამოიყენება მოდელის გამოსაცდელად (მაგ.მონაცემთა ინტელექტუალური ანალიზის მოდელი) და სატესტო მონაცემთა ნაკრები, რომელიც გამოიყენება მოდელის უტყუარობის შესამოწმებლად.

ყველა ზემოთ აღნიშნული საფეხური საერთოა ყველა ტიპის მოდელისათვის, არ აქვს მნიშვნელობა, ავტონომიური იქნება ის თუ სხვა ჯგუფთან გაერთიანებული. მონაცემთა გავრცელების და სიზუსტის მნიშვნელობის ვიზუალიზაციისათვის ჩვენ ვიყენებთ რამდენიმე ტიპის სქემას, ჩვენ შემთხვევაში ძირითადად ვიყენებთ ROC მრუდებს, რადგან ისინი სპეციალურად ამ მიზნებისთვისაა შემუშავებული. მაგალითი ასეთი ROC მრუდისა არის მონაცემთა ინტელექტუალური ანალიზით გადაწყვეტილებების ხისათვის შეიძლება იხილოთ ქვემოთ.



ნახ.1 გადაწყვეტილებების ხის მონაცემთა ROC გრაფიკი

რამდენადაც ეს სისტემა წარმოადგენს მანქანური სწავლების ტიპს, ამდენად უნდა ვასწავლოთ ჩვენს მოდელს, შემდეგ კი ის დამოუკიდებლად გათვლის დანარჩენ მონაცემებს. ჩვენ შემთხვევაში ვიყენებთ CSVფაილის მონაცემებს ჩვენი მოდელის ტესტირებიათვის და დამაკმაყოფილებელი შედეგის მიღების შემდეგ ჩვენ გამოვიყენებთ Oracle ბაზის მონაცემებს შესამოწმებლად.

ამრიგად, ამ თავში მონაცემთა წყაროსთნ ერთად ჩვენ გვაქვს ახალი დამატებითი პროცესი ჩვენი მოდელირებისათვის. ვალიდაციის პროცესისათვის. ვალიდაცია იგივეა, რაც საწარმოო გარემო. ყველა ტრენინგი და ტესტი ტარდება გარეგან ფაილზე და შემოწმება ტარდება მონაცემთა ბაზაში რეალურ მონაცემებზე. ჩვენ ავირჩიეთ Oracle მონაცემთა ბაზა არაშემთხვევითად. ეს მსოფლიო პრაქტიკაა, რომ Oracle მონაცემთა ბაზა გამოიყენება უმრავლესობა კრიტიკულად მნიშვნელოვან ტრანზაქციულ სისტემებში. ასეთი სისტემები ძირითადად რეალიზდება ფინანსურ, ტელესაკომუნიკაციო კომპანიებსა და საბანკო სექტორებში. მანქანური სწავლების ასეთი რთული და მაღალწარმოებადი სისტემების კომბინაცია და მონაცემთა წყაროს შექმნა Oracle მონაცემთა ბაზაში იქნება ღრმა სწავლების სისტემისათვის უფრო მძლავრ გადაწყვეტილება.

მალე სწავლებისა და ტესტირების პროცესი იქნება შემდეგი სახის:

- ჩვენ ვასწავლით ჩვენს მოდელს, დაფუძნებულს Random Forest და Random Forest UnderSampled- ზე მონაცემთა ინტელექტუალურ ანალიზს.
- შემდგომში ჩვენ შევქმნით მოდელთა ჯგუფს, რომლებიც შემდგარი იქნება მონაცემთა ინტელექტუალური ანალიზის იმ მეთოდებით, რომლებიც განხილულია მე-4 თავში.
- ბოლო საფეხურზე ჩვენ შევადარებთ Random Forest მოდელებს და ჯგუფის მოდელებს იმისათვის, რომ განვსაზღვროთ რომელს აქვს საუკეთესო სიზუსტე და შეესაბამება ჩვენს მოდელს.

შემდეგ ჩვენ ვაგებთ ორ მოდელს; Random Forest მოდელი და Random Forest UnderSampled მოდელი.

ახლა ჩვენ შევქმნით ფუნქციას, რომელიც უზრუნველყოფს კოლექციის, ანალიზისა და ვიზუალიზაციის მეთოდებს საერთო მონაცემთა ბაზიდან.

ფუნქციის რეზიუმე გამოითვლის თითოეული მოდელის მიდედვით ჯამურ სტატისტიკას/ მეტრული კომბინაციები. იმავე მონაცემთა ნაკრებიდან ის გარდაქმნილია და ჩამონათვალში აერთიანებს ყველა მოდელს. **სტატისტიკური მონაცემები გენერირდება სიზუსტის მიხედვით, მაგალითად როგორი იყო საშუალო სიზუსტე თითოეული მოდელის ყველა 10 მაგალითისათვის.**

```
results <-
resamples(list(ENSEMBLE=ensemble_experts,RF=rf_model,RF_US=rf_model_US,CART=c
art_model,GBM=gbmFit2, SVM=svm_model))
summary(results)

## Models: ENSEMBLE, RF, RF_US, CART, GBM, SVM
## Number of resamples: 10
##
## Accuracy
##             Min. 1st Qu. Median   Mean 3rd Qu.   Max. NA's
## ENSEMBLE 0.8261  0.8696 0.8748 0.8891  0.9158 0.9565    0
## RF       0.8000  0.8696 0.8723 0.8813  0.9130 0.9565    0
## RF_US    0.8261  0.8279 0.8723 0.8757  0.9130 0.9583    0
## CART     0.7917  0.8350 0.8696 0.8813  0.9130 1.0000    0
## GBM      0.7826  0.8474 0.8940 0.8853  0.9130 0.9565    0
## SVM      0.8261  0.8424 0.8723 0.8808  0.9035 1.0000    0
##
## Kappa
##                Min. 1st Qu. Median   Mean 3rd Qu.   Max. NA's
## ENSEMBLE  0.28360  0.5712 0.6363 0.6243  0.6902 0.8321    0
## RF        0.00000  0.3551 0.4343 0.4563  0.6230 0.8321    0
## RF_US     0.00000  0.2553 0.4343 0.4133  0.6230 0.8636    0
## CART      0.00000  0.2841 0.3551 0.4358  0.6230 1.0000    0
## GBM      -0.07477  0.4372 0.6049 0.5328  0.6788 0.8321    0
## SVM       0.00000  0.3015 0.3561 0.4394  0.5956 1.0000    0
```

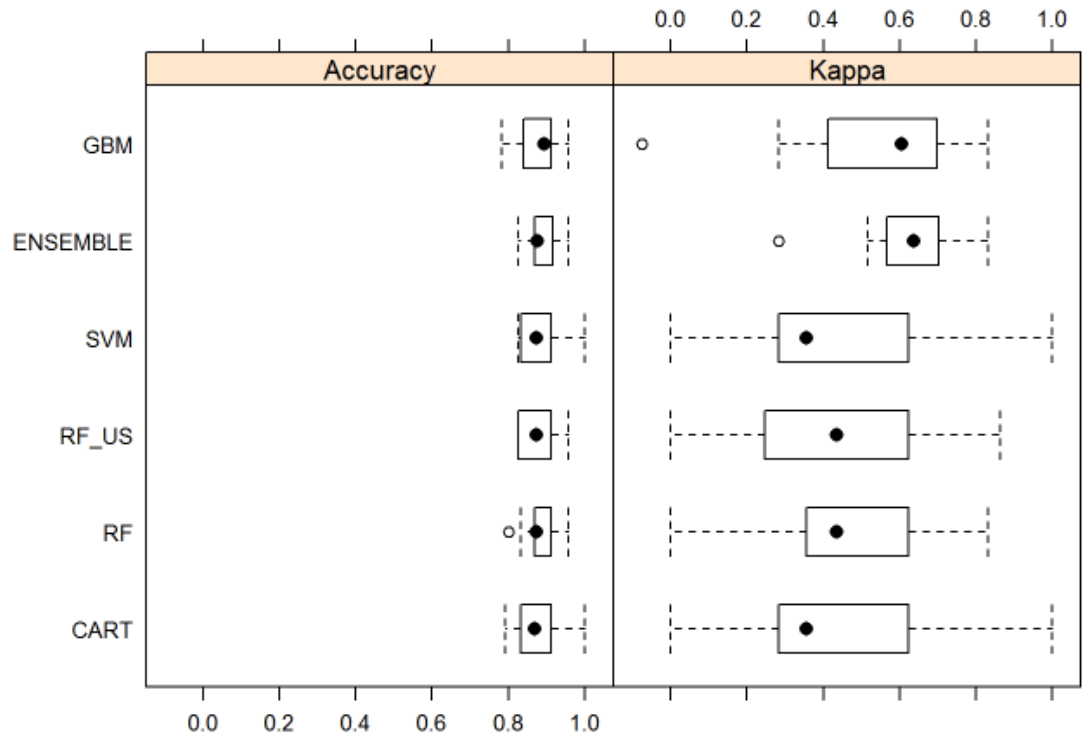
ნახ 2 ფუნქცია, რომელიც გამოიყენება მონაცემთა ნაკრებიდან კოლექციის, ანალიზის და ვიზუალიზაციისთვის

ვნახოთ როგორ ჯდება მოცემული შედეგები box-and-whisker plot-ში და dotplot-ში. რესემფლინგი არის მოდელების შემოწმება შემთხვევითი ქვესიმრავლეების გამოყენებით. ე.ი. ჯვარედინი შემოწმება.

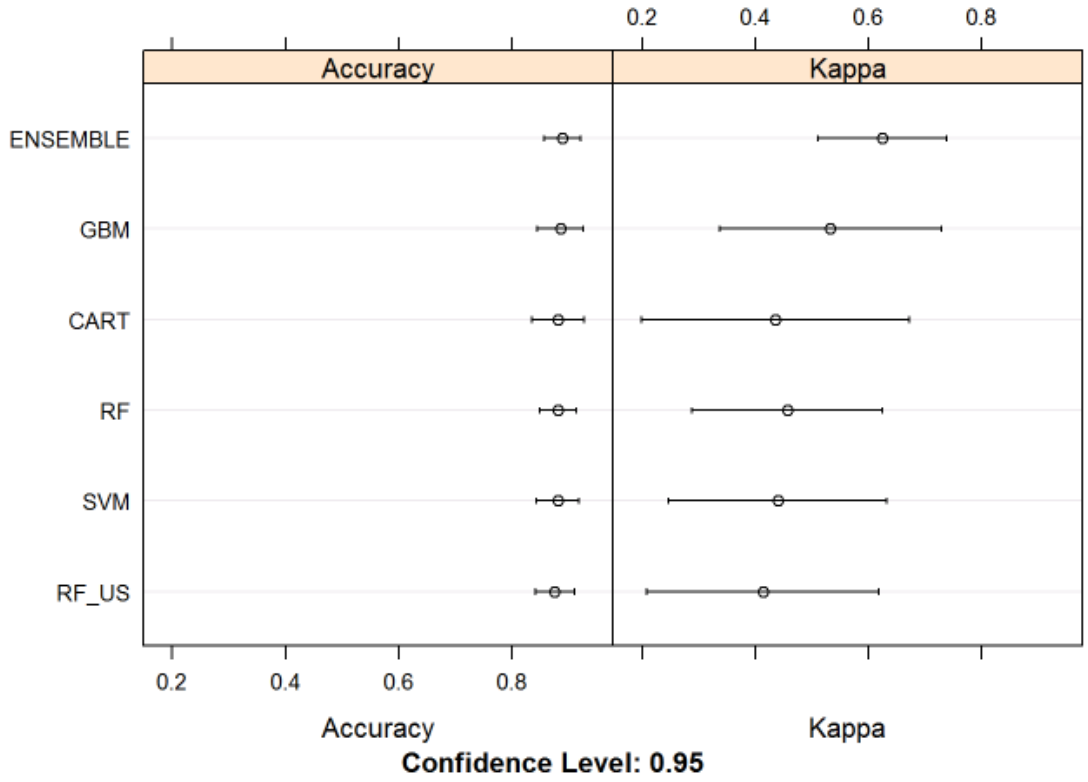
Bwplot ახდენს გადათვლის შემდეგ სიზუსტის შედეგების ვიზუალიზაციას. კარგი მოდელი უნდა ახდენდეს ძალიან მცირე დისპერსიის დემონტრირებას ყველა მაგალითში. ჩვენ შეგვიძლია დავინახოთ, რომ Random Forest (RF)-ს აქვს სუსტი outlier

მოდელი - არც ისე კარგი შედეგით. **RF_US** -ს არ აქვს ქვედა საზღვარი, შესაძლოა უმცირესი დისპერსიით.

```
bwplot(results)
```




```
dotplot(results)
```



ნახ. 3 ვიზუალიზაციის ფუნქცია კოლექციის, ანალიზის მეთოდებისა და მონაცემთა ნაკრებიდან შედეგების ვიზუალიზაციისათვის.

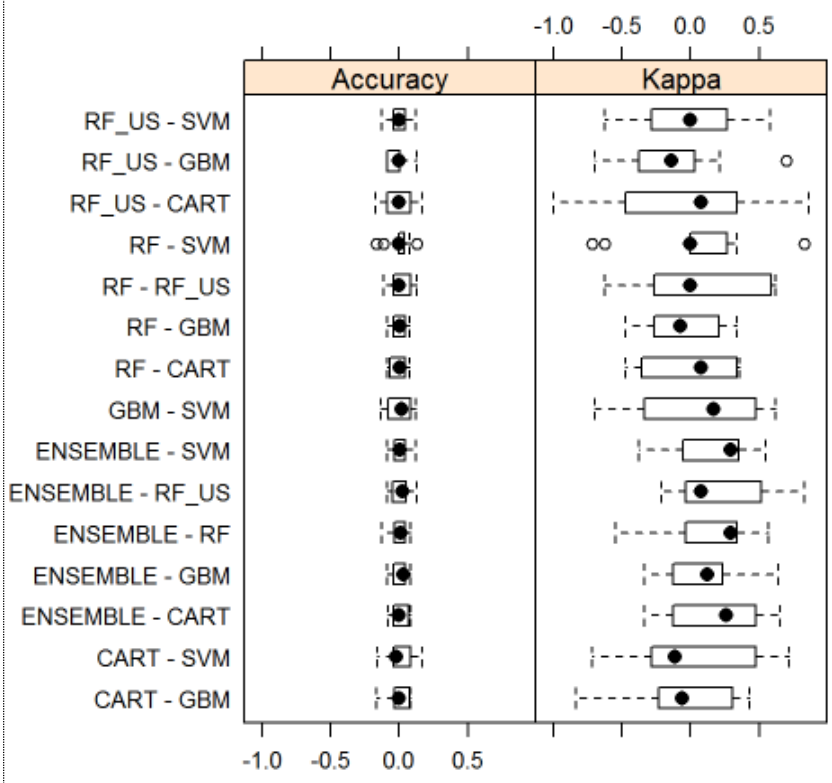
და ახლა ჩვენ შგვიძლია ვიპოვოთ განსხვავება მოდელებს შორის ჩვენს მიერ შექმნილი ფუნქციის გამოყენებით. ამ გრაფიკის მეშვეობით ხდება მოდელების შედარება. ჩვენს შემთხვევაში ის არ არის ძალიან გამჭრიახი, მაგრამ შეიძლება გამოყენება, თუ გვექნებოდა ორი ძალიან მსგავსი შემთხვევა.

Kappa არის საზომი იმისა, თუ რამდენად კარგად შეასრულდა კლასიფიკაცია უბრალოდ შემთხვევით კლასიფიკაციასთან შედარებით. სხვა სიტყვებით რომ ვთქვათ მოდელს ექნება Kappa -ს მაღალი ქულა თუ არსებობს დიდი განსხვავება სიზუსტესა და შეცდომის გამოვლენის სიხშირეს შორის.

```
difValues<-diff(results)

summary(difValues)
## Accuracy
##           ENSEMBLE RF          RF_US       CART        GBM         SVM
## ENSEMBLE             7.819e-03  1.336e-02  7.804e-03  3.819e-03  8.290e-03
## RF        1                     5.543e-03 -1.449e-05 -4.000e-03  4.710e-04
## RF_US     1         1                     -5.558e-03 -9.543e-03 -5.072e-03
## CART      1         1          1                     -3.986e-03  4.855e-04
## GBM       1         1          1          1                      4.471e-03
## SVM       1         1          1          1          1
##
## Kappa
##           ENSEMBLE RF         RF_US      CART       GBM        SVM
## ENSEMBLE             0.167947  0.211028  0.188536  0.091453  0.184895
## RF        1.0000               0.043081  0.020589 -0.076494  0.016948
## RF_US     1.0000    1.0000               -0.022492 -0.119575 -0.026133
## CART      1.0000    1.0000     1.0000               -0.097083 -0.003641
## GBM       1.0000    1.0000     1.0000    1.0000                0.093442
## SVM       0.9654    1.0000     1.0000    1.0000     1.0000

bwplot(difValues, layout = c(3, 1))
```



ნახ. 4 მოდელებს შორის განსხვავების მითითება

თავი 4 მონაცემებზე და შედეგებზე შემუშავებული საბოლოო მოდელის განხორციელება, შეფასება და ინტერპრეტაცია

აქამდე ჩვენ ავაგეთ მოდელები და გავაანალიზეთ, როგორ კარგად ასრულებენ ისინი თავის ფუნქციებს ერთმანეთთან შედარებით. და ახლა ჩვენ უნდა გავაკეთოთ

გათვლები იმისათვის, რათა გადავწყვიტოთ, ჩვენი აგებული მოდელებიდან რომელია საუკეთესო ჩვენს კვლევებში გამოსაყენებლად.

ამისათვის ჩვენ შევქმენით valAtRisk ფუნქცია. ეს არის ფუნქცია, რომელიც ითვლის კლასიფიკატორებს უფრო ნაკლები ღირებულების საფუძველზე, რომელიც არ იყო რისკის ქვეშ.

```
valAtRisk<- function(cross_tab_obj){
  saved_money<-1200*cross_tab_obj[1]
  val_risk<-70000*cross_tab_obj[3]
  xtra_revision<-1200*(cross_tab_obj[2]+cross_tab_obj[4])
  total_var<-xtra_revision+val_risk-saved_money
}
```

ნახ. 5 რისკის რაოდენობის შეფასების გათვლის ფუნქცია :

ახლა, როდესაც ჩვენ გვაქვს ფუნქცია, უნდა გამოვიყენოთ ის ჩვენს მიერ აგებულ თითოეულ მოდელზე. შესაბამისად დაბრუნებულ მნიშვნელობაზე დაყრდნობით მივიღებთ გადაწყვეტილებას.

შედეგში გვაქვს RF_US მოდელი.

```
rf_var<-valAtRisk(rf_ct)
rf_US_var<-valAtRisk(rf_US_ct)
gbm_var<-valAtRisk(gbm_ct)
svm_var<-valAtRisk(svm_ct)
cart_var<-valAtRisk(cart_ct)
nb_var<-valAtRisk(nb_ct)
knn_var<-valAtRisk(knn_ct)
ensemble_var<-valAtRisk(ensemble_ct)
var_frame<-
data.frame("Model_Name"=c("RF","RF_US","GBM","SVM","CART","NB","KNN","ENSEMBL
E"),"VaR"=c(rf_var,rf_US_var,gbm_var,svm_var,cart_var,nb_var,knn_var,ensemble
_var))

best_model<-var_frame[which.min(var_frame$VaR),1]

print(paste0("The best model is ",best_model))

## [1] "The best model is RF_US"
```

ნახ. 6 მნიშვნელობები რისკების გათვლისას

უკეთ გასაგებად მოვიყვანოთ არეული მატრიცის ნაყოფიერების მაგალითები Random Forest -ის ორივე მოდელისათვის და ნათელი მოვფინოთ მათ. ჩვენ უკვე ვიცით, რომ არეული მატრიცა წარმოადგენს ცხრილს, რომელიც გამოიყენება კლასიფიკაციის მოდელის დახასიათებისათვის სატესტო მონაცემების ნაკრებზე, რომელთათვისაც ცნობილია რეალური ღირებულება.

```
print(rf_model$finalModel)

## No. of variables tried at each split: 2
##
##         OOB estimate of  error rate: 11.97%
## Confusion matrix:
##           Approved Rejected class.error
## Approved      191        0   0.0000000
## Rejected       28       15   0.6511628


print(rf_model_US$finalModel)

## No. of variables tried at each split: 2
##
##         OOB estimate of  error rate: 21.79%
## Confusion matrix:
##           Approved Rejected class.error
## Approved      159       32   0.1675393
## Rejected       19       24   0.4418605
```

ნახ. 7 არეული მატრიცა Random Forest მოდელებისათვის

rf_ მოდელი არის მრავალი random forest-ის ჩამონათვალი. Rf_model$finalModel - არის საუკეთესო ფორმირებული მოდელი. ჩვენ ვიყენებთ სატრენინგო კონსტრუქციას, სადაც ერთ ბრძანებას ჩვენ ვსწავლობთ ბევრ მოდელში ერთი ცდით. rf_model/rf_model_US ამ მოდელის წარმოება ჩამოთვლის ყველა სახის აგებული მოდელების სიზუსტეს და rf_model_US$finalModel მოგვცემს საუკეთესო მოდელის დეტალებს. ის აგრეთვე მოგვცემს სატრენინგო მონაცემების არეულ მატრიცას.

არეულ მატრიცაში სტრიქონი წარმოადგენს დანართის ფაქტიურ მდგომარეობას. სვეტი პროგნოზირდება მოდელიდან. ასე, რომ 178 დამტკიცებული განაცხადიდან 159 იყო სწორად განსაზღვრული, მოდელით დადასტურებული. ანალოგირუად, 56 უარყოფილი განაცხადიდან მხოლდ 24 იყო სწორად განსაზღვრული, მოდელით დადასტურებული.

class.error -ეს არის შეცდომების გამოვლენის სიხშირე სწორე კლასების განსაზღვრისას ე.ი. სიზუსტე. სხვა სიტყვებით რომ ვთქვათ რამდენად ხშირადაა ეს კლასიფიკატორი სწორი.

დასკვნა

ამ კვლევაში თავიდან ჩვენ დავიწყეთ მონაცემთა ინტელექტუალური ანალიზის მეთოდით. მოვიყვანეთ რამდენიმე მაგალითი იმ ლიტერატურისა, რომელიც ძირითადად მონაცემთა ინტელექტუალური ანალიზის ალგორითმს იყენებს და განვიხილეთ რამდენიმე მაგალითი. ჩვენ ვისაუბრეთ იმის შესახებ, თუ რას აკეთებს მონაცემთა ინტელექტუალური ანალიზი და მანქანური სწავლება. აგრეთვე საკითხები, რომელიც ჯერ კიდევ გამოტოვებულია. ჩვენ ჩამოვთვალეთ მონაცემთა ინტელექტუალური ანალიზის ამოცანები, ისეთები როგორიცაა : კლასიფიკაცია, რეგრესია და ა.შ. რამდენადაც ჩვენი პირველადი შესწავლა ეფუძნებოდა ფინანსურ სექტორს, ჩვენ ასევე მოკლედ განვიხილეთ მონაცემთა ინტელექტუალური ანალიზის გამოყენება ფინანსურ სექტორში.

შემდეგ ჩვენ განვიხილეთ ჩვენი ინსტრუმენტი, რომელიც გამოვიყენეთ R პროგრამირების ენის შემუშავებისათვის. ჩვენ შევადარეთ ის ბაზარზე არსებულ სხვა კონკურენტებს შესაძლებლობებისა და ძლიერი მხარეების მიხედვით. ვნახეთ, რომ R ენა წარმოადგენს ძლიერ მოთამაშეს სტატისტიკის, მონაცემთა ანალიზის ბაზარზე და მანქანური სწავლების სფეროში, მიუხედავად იმისა, რომ ის წარმოადგენს პროგრამირების შედარებით ახალ ენას.

ლიტერატურული მიმოხლივის ბოლო ნაწილში ვისაუბრეთ ჩვენი კვლევის ძირითად თემაზე საკრედიტო სკორინგის შესახებ. ასევე ვისაუბრეთ მის ისტორიაზე და განვითარების საფეხურებზე. თუ რამდენად მნიშვნელოვანია ფინანსური ორგანიზაციებისათვის საკრედიტო სკორინგის გამოყენება. ვიმსჯელეთ, თუ რატომ შეიქმნა საკრედიტო სკორინგი, თუ რომელ პარამეტრებსა და ცვლადებს აქვთ დიდი მნიშვნელობა საკრედიტო სკორინგის დამუშავებისას.

ლიტერატურის მიმოხილვის შემდეგ ჩვენ ძირფესვიანად გავაანალიზეთ მონაცემთა ინტელექტუალი ანალიზის თითოეული ტექნიკა, რომელთაც ვგეგმავდით მომავალი

ექსპერიმენტისათვის. ჩვენ აღვწერეთ ყველა ის მეთოდი, რომლებიც ახორციელებენ მონაცემთა ინტელექტუალურ ანალიზს და მანქანურ სწავლებას.

მონაცემტა ინტელექტუალური ანალიზის მეთოდების თეორიულად განხილვის დასრულების შემდეგ, ჩვენ უნდა დაგვენახა როგორ გამოიყენება თითოეული მათგანი პრაქტიკულ ექსპერიმენტებში. რამდენადაც ჩვენი კვლევის მიზანი მდგომარეობს სკორინგის ოპტიმიზაციაში, ჩვენ პირველ რიგში უნდა გაგვეგო როგორ მუშაობდა თითოეული ალგორითმი შემდეგ კი შეგვერჩია საუკეთესო შესრულების ტექნიკა და შემდეგ კი გამოგვეყენებინა და აგვეწყო ის სკორინგის ოპტიმიზაციისათვის. სამართლიანობისათვის კონცეფციის ჭეშმარიტების დადგენისას ჩვენ ვიყენებდით მონაცემთა იგივე ნაკრებს მონაცემთა ინტელექტუალური ანალიზის თითოეულ ალგორითმში. ყველა ექსპერიმენტში მონაცემთა ინტელექტუალური ანალიზისა და მონაცემთა ნაკრებისათვისჩვენ ვიყენებდით R პაკეტებს მათი მოდელების შესწავლისა და ტესტირებისათვის.

რამდენადაც ჩვენ გავაკეთეთ ყველა შერჩეული ალგორითმის შედარებითი ანალიზი, უნდა შეგვერჩია ყველაზე საუკეთესო ტექნიკა ჩვენი მოდელის ასაგებად და მონაცემების შესამოწმებლად. ჩვენ შევარჩიეთ Random Forest Under Sampled - ალგორითმი საკრედიტო სკორინგის ჩვენი მოდელის ასაგებად. ბოლო ორ თავში ჩვენ გამოვიყენეთ კოდის ფრაგმენტები იმისათვის, რათა უკეთ გაგვეგო დასკვნის შედეგი და გამოვიყენეთ გრაფიკის რამდენიმე უბანი მომზადებისა და გამოცდის ვიზუალიზაციისათვის.

საბოლოოდ, ვისურვებდი აღმენიშნა ერთი მნიშვნელოვანი ასპექტი. კვლევის დასასრულს ჩვენ შევარჩიეთ საუკეთესო მოდელი, მაგრამ ეს არ ნიშნავს იმას, რომ სხვა მოდელები უვარგისია. ისინი ცუდი მოდელები არ არიან, მაგრამ მოცემულ მომენტში არ არის ისეთი კარგი, როგორიც ჩვენს მიერ შერჩეული მოდელი. მონაცემთა ინტელექტუალური ანალიზი და მანქანური სწავლება არის სიღრმისეული მეცნიერება და არ არსებობს საუკეთესო მოდელი ან მეთოდები ამ

კონკრეტული ამოცანის ამოსახსნელად. ყველაფერი იმაზეა დამოკიდებული, თუ რა ამოცანას ასრულებთ მონაცემთა ინტელექტუალური ანალიზის კონკრეტული ალგორითმის შესრულებაზე და როგორია თქვენი მონაცემთა ნაკრები. თუ არ გაქვთ სტანდარტიზირებული მონაცემთა ნაკრები თქვენ რჩებით არადამაკმაყოფილებელი შედეგის მიღების რისკის ქვეშ. ჩვენს შემთხვევაში მიზეზი, თუ რატომ ავირჩიეთ Random Forest Under Sampled, არის ის, რომ ამ მეთოდით ბანკები იღებენ უფრო უკეთეს ფინანსურ ღირებულებას, სხვა სიტყვებით რომ ვთქვათ ამ მეთოდის გამოყენება გულისხმობს იმას, რომ ბანკი კარგავს უმნიშვნელო რაოდენობით ფულს.
აქამდე ჩვენი გათვლები გვიჩვენებს, რომ Under Sampled Random Forest მოდელი საუკეთესოა ჩვენი მონაცემთა ნაკრების პროგნოზირებასა და მოდელირებაში.
ამის მიზეზი ისაა, რომ არის ხარჯები, რომლებიც დაკავშირებულია არასწორ კლასიფიკაციასთან, როგორც მიღებულია უარყოფილი იქნება, წლის ბოლოს კი ბანკი ან სხვა ნებისმიერი ფინანსური დაწესებულება თავის თავზე იღებს რისკს.

ამგვარად, მოდელი, რომელიც ამცირებს რისკს და უზრუნველყოფს უკეთეს ფინანსურ ღირებულებას ჩვენს კვლევებში, როგორც წესი **Under Sampled Random Forest** არის ეს და ჩვენი გათვლებით ის არის საუკეთესო და საბოლოო მოდელი. არსებული კლასიფიკაციის მოდელების უმრავლესობა არ მუშაობს კარგად, როდესაც მონაცემთა ნაკრები არაბალანსირებულია. ჩვენ შემთხვევაში ჩვენ გამოვიყენეთ შემთხვევითი **Forest** ალგორითმი და under Sampling მეთოდი დისბალანსის ამოსაფხვრელად და მივიღეთ ბევრად უკეთესი შედეგი სხვა მონაცემთა მოპოვების და კლასიფიკაციის მეთოდებზე.

რეკომენდაციები:

შემდგომი მუშაობისათვის აღნიშნული კვლევა შეიძლება გაფართოვდეს რამდენიმე გზით:

- იგივე მეთოდი შეიძლება იქნას გამოყენებულ სხვა ნებისმიერ სფეროში სხვადასხვა სახის ამოცანათა მოდელების შესაქმნელად, გარდა ფინანსურ სფეროში.
- ETL(გამოყოფა, გარდაქმნა, ჩატვირთვა) -ის გაცილებით ღრმა პროცესი შეიძლება გამოყენებულ იქნას უფრო სუფთა მონაცემებისათვის.

გამოყენებული ლიტერატურა:

1. Zakirov, D. and Momtselidze, N. (2015) Application of Data Mining in the Banking Sector, IBSU Journal of Technologies and Technical Science, 4(1), p.13-16, ISSN:2298-0032.
2. Zakirov, D., Bondarev, A., and Momtselidze N. (2015). A Comparison of Data Mining Techniques in Evaluating Retail Credit Scoring Using R Programming, 12th International Conference on Electronics Computer and Computation (ICECCO) IEEE, p.69-73, ISBN: 978-1-5090-0199-6
3. Zakirov, D., Bondarev, A., and Momtselidze N., (2015). Data Warehouse on Hadoop Platform for Decision Support Systems in Education, 12th International Conference on Electronics Computer and Computation (ICECCO) IEEE, p.73-77, ISBN: 978-1-5090-0199-6
4. Zakirov, D. (2016) Credit Scoring based on Random Forests Algorithm: An Effective Empirical example, Journal of Science, Innovation and New Technologies of Kyrgyzstan, 1(2016), p.44-49, ISSN:1694-7649.