ҒТАХР 62.30.20

10.51889/2959-5894.2025.89.1.016

Д.К. Даркенбаев[1*], Н.О. Мекебаев[2]

[1]әл-Фараби атындағы Қазақ ұлттық университеті, Алматы қ., Қазақстан
[2]Қазақ ұлттық қыздар педагогикалық университеті, Алматы қ., Қазақстан
[*] *e-mail: dauren.kadyrovich@gmail.com*

МАШИНАЛЫҚ ОҚЫТУ АЛГОРИТМДЕРІН НЕСИЕЛІК ТӘУЕКЕЛДІ БАҒАЛАУДА ҚОЛДАНУ

Аңдатпа

Мақалада банктік несиелік тәуекелді бағалаудың өзекті мәселелері, сондай-ақ қарыз алушының несиелік қабілетін бағалау үшін деректерді талдау әдістерін пайдалана отырып, машиналық оқыту алгоритмдері зерттелді. Деректерді өңдеу процесін көрсету үшін мысал ретінде несиелік рейтингтер таңдалды. Несиелік тәуекелді бағалау үшін несиелік скоринг қаржылық институттардағы төлем қабілеттері бар немесе төлем қабілеттері жоқ клиенттерден ажыратудың маңызды құралы екені талданды. Мақалада, машиналық оқыту алгоритмдері кредиттік скорингке сәтті қолданылды. Несиелік тәуекелді азайту – қаржылық дағдарыстардан кейін қызығушылықтың артатын саласы, сондықтан қаржы институттары көптеген деректер жинақтайды. Бұл тәуекел сарапшыларына үлкен деректерді өңдеу және жеке тұлғаның төлем қабілетін анықтау сияқты қиын тапсырма берді. Қаржы институттары жоғары тиімді несиелік скорингтік модельдерді табу үшін күрделі машиналық оқыту әдістерін пайдалана алады. Мақалада салыстырмалы талдау үшін машиналық оқытуды жіктеу әдістері қолданылды. Нәтижелер көрсеткендей, регрессия дефолтты жақсы болжағанын, содан кейін кездейсоқ орман алгоритмі екенін көрсетті. Кеңінен қолданылатын логит моделі тірек векторлық машинаға қарағанда жақсы нәтиже көрсетті. Сонымен қатар, Колмогоров-Смирновтың сынағы арқылы біз машиналық оқытудың басқа әдістерінің кластарды қаншалықты жақсы жіктей алатындығын және логит моделінен асып түсетінін дәлелдедік.

Түйін сөздер: деректер, алгоритмдер, өңдеу, технология, әдістер, машиналық оқыту.

Д.К.Даркенбаев[1], Н.О.Мекебаев[2]
[1]Казахский национальный университет им. Аль-Фараби, г. Алматы, Казахстан
[2]Казахский национальный женский педагогический университет, Алматы, Казахстан

ИСПОЛЬЗОВАНИЕ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ ПРИ ОЦЕНКЕ КРЕДИТНОГО РИСКА

Аннотация

В статье исследуются актуальные вопросы оценки банковского кредитного риска, а также алгоритмы машинного обучения, применяемые для анализа данных с целью оценки кредитоспособности заемщика. В качестве примера для иллюстрации процесса обработки данных были выбраны кредитные рейтинги. Кредитный скоринг стал важнейшим инструментом оценки кредитного риска, позволяющим финансовым учреждениям отличать платежеспособных клиентов от неплатежеспособных. Для решения задач кредитного скоринга успешно применяются алгоритмы машинного обучения. Снижение кредитного риска представляет собой область повышенного интереса, особенно в условиях финансовых кризисов. В связи с этим финансовые учреждения собирают значительные объемы данных, что ставит перед аналитиками сложную задачу обработки больших данных и точного определения платежеспособности заемщиков. В поисках высокоэффективных моделей кредитного скоринга финансовые учреждения могут использовать современные методы машинного обучения. В статье для сравнительного анализа были применены методы классификации машинного обучения. Результаты исследования показали, что регрессия обеспечивает наилучшую оценку вероятности дефолта, за ней следует модель случайного леса. Широко используемая логит-модель продемонстрировала более высокие результаты по сравнению с методом опорных векторов. Кроме того, с помощью теста Колмогорова-Смирнова было доказано, что другие методы машинного обучения превосходят логит-модель по точности классификации классов.

Ключевые слова: данные, алгоритмы, обработка, технология, методы, машинное обучение.

D.K. Darkenbayev[1], N.O.Mekebayev[2]
[1]Al-Farabi Kazakh National University, Almaty, Kazakhstan
[2]Kazakh National Women's Pedagogical University, Almaty, Kazakhstan

USING MACHINE LEARNING ALGORITHMS IN CREDIT RISK ASSESSMENT

Abstract

The article investigated topical issues of assessing bank credit risk, as well as machine learning algorithms using data analysis methods to assess the creditworthiness of the borrower. Credit ratings were chosen as an example to illustrate the data processing process. To assess credit risk, credit scoring has become an essential tool for distinguishing solvent clients from non-solvent clients in financial institutions. Accordingly, machine learning algorithms have been successfully applied to credit scoring. Reducing credit risk is an area of increased interest in connection with financial crises and therefore financial institutions collect a lot of data. This has presented risk analysts with the difficult task of processing big data and adequately determining a person's ability to pay. In the search for highly effective credit scoring models, financial institutions can use sophisticated machine learning techniques. In the article, machine learning classification methods were used for comparative analysis. The results show that regression provides the best estimate of default, followed by a random forest model. The widely used logit model has shown better results than the support vector machine. Moreover, using the Kolmogorov-Smirnov test, we proved that other machine learning methods outperform the widely used logit model in how well the model can classify classes.

Keywords: data, algorithms, processing, technology, methods, machine learning.

Негізгі ережелер

Мақаланың негізгі мақсаты жеке тұлғалардың деректерін өңдеу негізінде несиелік тәуекелді бағалау. Деректердің күн санап өсуі, оларды өңдеу мен сақтау мәселелерін туындатқаны белгілі. Оларды өңдеуде тиімді әдістер ретінде машиналық оқыту алгоритмдері таңдалып алынып, жартылай құрылымданған деректерді өңдеудегі нәтижелері салыстырыла зерттелді. Жалпы машиналық оқыту алгоритмдерін таңдауда олардың үлкен көлемді деректерді өңдеудегі нәтижелері салыстырылады, кейбір модельдер аз ғана деректерде жақсы нәтижелер көрсетсе, ал кейбіреулері үлкен көлемді деректерді өңдеуде жақсы нәтижелер көрсетеді. Деректерді өңдеуде олардың шынайылығы модельдің дұрыс жұмыс істеп, нақты болжамдар алуына оң әсерін тигізеді. Аталмыш мақаладағы негізгі идеясы жеке тұлғалардың жартылай құрылымданған деректерін өңдеп, несиелік тәуекелді бағалау яғни несие беруге болатын немесе несие беруге болмайтын клиенттерін анықтау болып табылады.

Кіріспе

Технология күн санап дамыған сайын, машиналық оқыту алгоритмдері де қаржы институттарының несие қабілеттілігін және несиені мақұлдау процестерін бағалауға қолданылуда [1].

Машиналық оқыту - бұл деректерден үйрену және адамдар анықтауы қиын үлгілерді тану үшін алгоритмдерді пайдаланатын жасанды интеллект түрі [2].

Қаржы институттары қазір несие қабілеттілігін дәлірек бағалау, сондай-ақ қолмен жасалатын процестер мен қолмен жіберілетін қателерді азайту үшін машиналық оқытуды пайдаланады. Машиналық оқыту несие берушілерге үлкен көлемдегі деректерді дәстүрлі әдістерге қарағанда тезірек және дәлірек талдауға көмектеседі. Бұл несие берушілерге тәуекелді азайтуға және несиені мақұлдау мөлшерлемелерін жақсартуға көмектесетін әлеуетті қарыз алушылар туралы неғұрлым саналы шешім қабылдауға мүмкіндік берді. Сондай-ақ несиеге өтініш беру процесін автоматтандыру үшін машиналық оқыту қолданылады. Өтінімдерді өңдеу үшін алгоритмдерді пайдалану арқылы несие берушілер несиені мақұлдауға кететін уақытты қысқарта алады және талап етілетін құжат айналымын азайтады. Бұл несиені мақұлдауды жылдамдатуға және несиені өңдеу уақытын қысқартуға әкеледі [3].

Сонымен қатар, дәлірек несие ұпайларын беру үшін машиналық оқыту қолданылады.

Әртүрлі көздерден алынған деректерді пайдалана отырып, несие берушілер несие тәуекелін жақсы болжап, дәлірек несие ұпайларын жасай алады. Жалпы алғанда, машиналық оқыту қаржы институттарына несиені мақұлдау және несиелік скоринг туралы көбірек негізделген шешімдер қабылдауға қажетті құралдармен қамтамасыз етеді. Деректерді тиімдірек пайдалану арқылы несие берушілер өздерінің несиелерімен байланысты тәуекел деңгейін төмендете алады және әлеуетті қарыз алушыларға дәлірек ақпарат бере алады. Несиелік тәуекелдерді болжау үшін машиналық оқытуды пайдалану қаржы институттары арасында танымал бола түсуде. Бұл дәлірек және сенімді болжамдар жасау үшін күрделі алгоритмдерді пайдалануға мүмкіндік беретін технология жетістіктерінің арқасында мүмкін болды [4].

Машиналық оқыту қаржы институттарына деректердің үлкен көлемін талдау арқылы несиелік тәуекелді дәлірек болжауға мүмкіндік берді. Машиналық оқыту алгоритмдерін пайдалана отырып, жеке тұлғаларға несие берумен байланысты тәуекелді бағалау үшін пайдаланылады. Бұл қаржы саласына үлкен әсер етуі мүмкін несиелік тәуекелді болжау дәлдігін арттыруға әкеледі. Несиелік тәуекелді болжау үшін машиналық оқытуды пайдаланудың артықшылықтары өте көп. Бұл қаржы институттарына несиелерін өтей алмау қаупі ең төмен, ең жақсы несие алушыларды анықтауға көмектеседі. Бұл нашар несиелермен байланысты шығындарды азайтуға көмектеседі және пайданың артуына әкелуі мүмкін. Сонымен қатар, машиналық оқыту банктерге ықтимал алаяқтарды анықтауға және күдікті әрекетті белгілеуге көмектеседі. Жалпы алғанда, машиналық оқыту несиелік тәуекелді болжауға қатысты қаржы институттары үшін барған сайын маңызды құралға айналуда. Күрделі алгоритмдерді пайдалана отырып, банктер дефолттар мен алаяқтық әрекеттерден болатын шығындарды айтарлықтай азайтуға мүмкіндік береді және сенімді болжамдар жасай алады. Бұл пайданың артуына және банктер мен олардың клиенттерінің қаржылық қауіпсіздігін арттырады [5].

*Әдебиетке шолу***.** Машиналық оқыту алгоритмдерін тиімді пайдалану және үлкен көлемді деректерді өңдеу бойынша ғылыми зерттеу жұмыстарды шетелдік ғалымдар Крис Филлипс, Пол Ежилчелван, Чунг, Х.М., Джойс Джексон, Сринивасан В., КимЁнг, Хенли В.Э., Десай В.С., Конвей Д.Г., Крук Дж. Ресей ғалымдары Н.В. Бабина, А.А. Земцова, Т.Ю. Осипов, В.Расторгуев және отандық ғалымдар М.Н.Қалимолдаев, Е.Н.Әмірғалиев, Г.Т.Балақаева, О.Ж.Мамырбаев, М.Е.Мансурова сияқты жетекші ғалымдардың еңбектерінен көруімізге болады.

Зерттеу әдіснамасы

Ғылыми мақаланың мақсаты машиналық оқыту алгоритмдерін қолдану арқылы үлкен көлемді деректерді өңдеп, болжамдар жасау және шешім қабылдауды автоматтандыру. Жеке тұлғалардың деректері заңмен қорғалады, көптеген қаржылық ұйымдар деректерді құпия ұстайды. Сондықтан, шынайы деректерді тауып, олардың негізінде болжамдар жасау аса маңызды. Бұл қолданылатын модельдің дұрыс жұмыс жасауына ықпал етеді.

Несиелік скорингті автоматтандыру үшін машиналық оқытуды пайдалану банктік және қаржылық қызметтер индустриясында кең таралған. Шешім қабылдаудың автоматтандырылған әдістерін пайдалана отырып, банктер және басқа қаржы институттары адамның несие қабілеттілігін жылдам бағалай алады және олардың несиелік өтінімдері немесе несие лимитін ұлғайту туралы шешім қабылдай алады [6].

Машиналық оқыту алгоритмдері деректердің үлкен көлемін жылдам және дәл өңдей алады, содан кейін адамның несие қабілеттілігін автоматты түрде анықтау үшін пайдалануға болады. Болжам жасауға негізделген модельдер адамның несиелік ұпайы, кірісі, несие тарихы және басқа қаржылық ақпарат сияқты әртүрлі факторларды ескере алады. Бірнеше көздерден алынған деректерді қосу арқылы машиналық оқыту алгоритмдері дәстүрлі кредиттік скоринг үлгілеріне қарағанда дәлірек болжам жасай алады. Несиелік скорингті автоматтандыру үшін машиналық оқытуды пайдалану адамның несие қабілеттілігін бағалауға байланысты уақыт пен шығындарды айтарлықтай қысқартуы мүмкін. Қаржы институттарына шешімді тез және

дәл қабылдауға мүмкіндік бере отырып, банктер мен басқа несие берушілер өз клиенттеріне бәсекеге қабілетті мөлшерлемелер мен шарттарды ұсынуға жақсырақ жабдықталған. Сонымен қатар, шешім қабылдау процесін автоматтандыру арқылы қаржы институттары үлкен қателіктерін азайта алады. Кредиттік скорингті автоматтандыру үшін машиналық оқытуды пайдалану әлі бастапқы кезеңдерінде деп айтуға болады, бірақ ықтимал артықшылықтар бар. Технологиялар күннен күнге дамып келе жатқандықтан, ол банктік және қаржылық қызметтер индустриясының маңызды бөлігіне айналады деп есептеледі [7].

Бұл мақалада біз клиенттің дефолт мүмкіндігін көрсететін сипаттамаларын зерттеуге тырысамыз, сондай-ақ ашық қолданыстағы Хоум банк несие деректерін пайдаланамыз, ол деректер ашық қолданыста бар және несиенің мерзімін өтеу мен несие қабілеттілігін тиімді модельдейтін ең жақсы несиелік модельдеу алгоритмін табуға тырысамыз. Жартылай құрылымданған деректерді машиналық оқыту алгоритмдерін қолдану арқылы өңдейміз және дефолттық жағдайға болжамдар жасаймыз және несиелерін қайтара алатындар, қайтара алмайтындар деп нәтижелерді екі класқа жіктейміз.

Логистикалық регрессия. Логит моделі несиелік скорингте ең көп қолданылатын алгоритмдердің бірі болып табылады. Бұл жалпы сызықтық модельдің бірегей жағдайы және белгілі бір жағынан сызықтық регрессиямен салыстыруға болады. Дегенмен, жалпы сызықтық регрессиядан айырмашылығы, логит модельдері ең алдымен үздіксіз нәтижелерді емес, дихотомиялық тәуелді нәтижелерді болжауға арналған.Дегенмен, жалпы сызықтық регрессиядан айырмашылығы, логит модельдері ең алдымен үздіксіз нәтижелерді емес, дихотомиялық тәуелді нәтижелерді болжауға арналған. Бұл логистикалық түрлендірудің арқасында [−∞, + ∞] мәнінен 0 мен 1 арасындағы ықтималдыққа дейін шығысты шектеу арқылы қол жеткізіледі. Логистикалық функция кері логитпен беріледі [8].

$$\log it^{-1}(\alpha)=\frac{1}{1+e^{(-\alpha)}}=\frac{e^{\alpha}}{e^{\alpha}+1} \tag{1}$$

N деректер нүктелерінің оқу жинағы үшін, $H=\{(x_i,y_i)\}_{i=1}^{N}$ және $x_i\in R^n$ кіріс айнымалылар. Бинарлы нәтижеге $y_i\in\{0,1\}$ сәйкес келетін логистикалық регрессияның мақсаты $P(y=1|x)$ мәнін келесідей бағалау.

$$P(y=1|x)=\frac{1}{1+e^{\left(-(\omega_0+\omega^T x)\right)}} \tag{2}$$

және

$$P(y=0|x)=\frac{e^{-(\omega_0+\omega^T x)}}{1+e^{-(\omega_0+\omega^T x)}} \tag{3}$$

бұл жерде, ω_0 кесінді, ω вектордың параметрі және $x\in R^n$ n өлшемді вектор болып табылады.

Біз, ω_0 және ω параметрлерін бағалау үшін максималды ықтималдық техникасын қолданамыз. Байқау ықтималдығымен нәтиже төмендегідей берілген.

$$P(y|x)=P(y=1|x)^{y}\left(1-P(y=1|x)^{1-y}\right) \tag{4}$$

Бұл процедураның негіздемесі мынаны сипаттайды бақылау ықтималдығын барынша арттыру деректер жинағы H бақылауларды ескере отырып дербес сызылады, бұл төмендегідей

нәтиже нәтиже береді.

$$\prod_{i=1}^{N} P\left(y_i = 1 \middle| x_i\right)^{y_i} \left(1 - P\left(y_i = 1 \middle| x_i\right)^{1-y_i}\right) \tag{5}$$

Логарифдік сенімділік келесідей анықталады:

$$LL = \sum_{i=1}^{N} y_i \log\left(P\left(y_i = 1 \middle| x_i\right)\right) + \left(1 - y_i\right) \log\left(1 - P\left(y_i = 1 \middle| x_i\right)\right) \tag{6}$$

Логарифмдік сенімділік статистикасы үлгіні орнатудан кейін түсіндірілмеген ақпараттың өнімділік өлшемі болып табылады, бұл оны квадраттардың қалдық сомасымен салыстыруға мүмкіндік береді. Өнімділік көрсеткішінің критерийі үлкенірек логарифмдік сенімділік шамасы арқылы беріледі, мұнда статистика неғұрлым үлкен болса, соғұрлым түсініксіз ақпарат болады.

Кездейсоқ орман (Random Forest). Кездейсоқ орман - бұл классификацияның ансамбльдік оқыту әдістемесі, онда классификатор ағаш құрылымдарының жиынтығы түрінде болады.

$$\left\{h\left(X, \theta_k\right), k = 1, \ldots\right\} \tag{7}$$

Кездейсоқ векторлар θ_k тәуелсіз және бірдей таралған, әрбір ағаш X кірісінде және θ шығысында ең танымал санат бойынша шешімді тіркейді.

Кездейсоқ кіріс векторы $X \subset x \subset R^p$ белгілі және кіріс деректерінің көмегімен регрессия функциясын бағалау арқылы квадрат интегралдың $Y = R$ кездейсоқ жауабын болжау:

$$m(x) = E\left[Y \middle| X = x\right] \tag{8}$$

Тәуелсіз және бірдей берілген, $D_n((X_1, Y_1), \ldots (X_n, Y_n))$ оқу деректер жинағы бар делік. D_n деректер жинағының көмегімен келесідей функция алынады: $m_n = x \to R$ Бұл жағдайда m_n регрессия функциясының дәйектілігі сақталады егер, $E\left[m_n(X) - m(X)\right]^2 \to 0$ және $n \to \infty$ ұмтылса. Орташа мән, X және D_n деректер жиынтығы бойынша есептеледі. Кездейсоқ шешім ормандарын рандомдалған негізгі регрессия ағаштарының жиынтығы ретінде қарастыруға болады:

$$\left\{r_n\left(x, \theta_m, D_n\right), m \geq 1\right\} \tag{9}$$

мұндағы θ_k -тәуелсіз және бірдей бөлінген кездейсоқ векторлар.

$$r_n\left(x, \theta_j\ D_n\right) = \sum_{i \in D_n^*\left(\theta_j\right)} \frac{I_{X_I \in A_n\left(x, \theta, D_n\right)^{Y_i}}}{N_n\left(x, \theta_j, D_n\right)} \tag{10}$$

Бұл жердегі, $D_n^*\left(\theta_j\right)$жаттығу деректер жиынтығы. $A_n\left(x; \theta_j, D_n\right)$ кіріс деректері бар ұяшық,Содан кейін ағаштар соңғы орманды бағалау үшін біріктіріледі.

$$m_n\left(x, \theta_1, \theta_M, D_n\right) = \frac{1}{M} \sum_{j=1}^{m} m_n\left(x, \theta_j, D_n\right) \tag{11}$$

мұндағы E_θ, D_n -шарты бойынша θ кездейсоқ мүмкіндігіне қатысты математикалық күту.

Тірек вектор машинасы. Тірек векторлық машинасы (SVM) – алгоритм графикте бөлу сызығына жақын нүктелерді іздейді. Бұл нүктелер тірек векторлары деп аталады. Содан кейін, алгоритм тірек векторлары мен бөлетін жазықтық арасындағы қашықтықты есептейді. Бұл қашықтық саңылау деп аталады. Алгоритмнің негізгі мақсаты - тазарту қашықтығын барынша арттыру. Ең жақсы гипержазықтық бұл алшақтық мүмкіндігінше үлкен болатын гипержазықтық болып саналады [9].

Тірек векторлық машиналары (SVM) тарихи қаржылық деректер жиынтығы бойынша модельді оқыту арқылы несиелік ұпайларды болжау үшін пайдаланылуы мүмкін машиналық оқыту әдісі. SVM алгоритмі деректер нүктелерінің кластарын ең үлкен маржамен бөлетін ең жақсы гипержазықтықты табуға тырысады. Несиелік баллды болжау жағдайында, алгоритм жақсы несие ұпайлары бар қарыз алушыларды нашар несиелік ұпайлары бар қарыз алушылардан бөлуге бағытталған. Оқытылған SVM моделін жаңа қарыз алушылардың қаржылық деректеріне негізделген несиелік ұпайын болжау үшін пайдалануға болады. Дегенмен, SVM негізіндегі несиелік балды болжау өнімділігі гиперпараметрлерді таңдауға сезімтал болуы мүмкін. Сондықтан, SVM негізіндегі несиелік ұпайды болжауда максималды өнімділікке қол жеткізу үшін гиперпараметрлерді мұқият таңдау және реттеу маңызды [10].

Лассо регрессиясы. Лассо регрессиясы реттеуші әдіс болып табылады. Ол дәлірек болжау үшін регрессия әдістерінің орнына қолданылады. Бұл модельді қателікті қысқарту мақсатында пайдалануға болады [11]. Қысқарту - бұл деректер мәндері орташа мән ретінде орталық нүктеге қысылған кезде. Лассо қарапайым, сирек модельдерді (яғни параметрлері аз модельдерді) қолдайды. Регрессияның бұл ерекше түрі мультиколлинеарлықтың жоғары деңгейлерін көрсететін үлгілер үшін немесе айнымалы таңдау/параметрді алып тастау сияқты үлгі таңдаудың белгілі бір бөліктерін автоматтандыру қажет болғанда өте қолайлы [12].

Енгізу кеңістігі үшін $X \in R^N$ және өлшенетін $Y \in R$ сызықтық гипотезалар тобын $G = \{x \to w \cdot x + b : w \in R^N, b \in R\}$ қарастырамыз. Жаттығу деректері $S = \{(x_1, y_1),..(x_m, y_m) \in (X \times Y)^m\}$ болсын делік. Лассоның мақсаты L_1 салмақ векторының нормасымен реттелетін реттеу мүшесі S бойынша эмпирикалық квадраттық қатені азайту болып табылады.

$$\min_{(w,b)} F(w,b) = \lambda \|w\|_1 + \sum_{i=1}^{m} (w \cdot x_i + b - y_i)^2 \tag{12}$$

мұндағы λ оң параметр. 12-ші теңдеуді оңтайландыру есебі деп айтуға болады, себебі $\|w\|_1$ және эмпирикалық қате дөңес. Сондықтан, 12-теңдеу үшін оңтайландыруды былай жазуға болады:

$$\sum_{i=1}^{m} (w \cdot x_i + b - y_i)^2 \ \|w\|_1 \prec \psi \tag{13}$$

мұндағы ψ - оң параметр болып табылады [13].

Өнімділік өлшемдері. Модельдің өнімділігін бағалау мен мүмкіндігін тексеретін оқытылатын деректерге сүйенетін бірнеше әдістер бар [14]. Модельдің өнімділігін бағалау үшін қабылдағыштың жұмыс сипаттамасы қисығы астындағы ауданды (AUROC), Джини статистикасын және Андерсон Дарлинг статистикасын салыстырамыз. ROC қисығы екілік жіктеу мәселелерін бағалау үшін кеңінен қолданылатын құрал болып табылады [15]. Бұл әртүрлі шекті параметрлердегі сезімталдық пен ерекшеліктің графикалық көрінісі. AUROC қисығы неғұрлым үлкен болса, модель соғұрлым жақсы болады. Әдетте, өте жақсы модельдің ауданы 0,80-0,89 болады. 0,6-дан жоғары Джини коэффициенті жақсы модель екенін білдіреді. Колмогоров-Смирнов тесті (KS тесті) алгоритмнің моделінің екі класты қаншалықты ажырата

алатынын бағалайды.

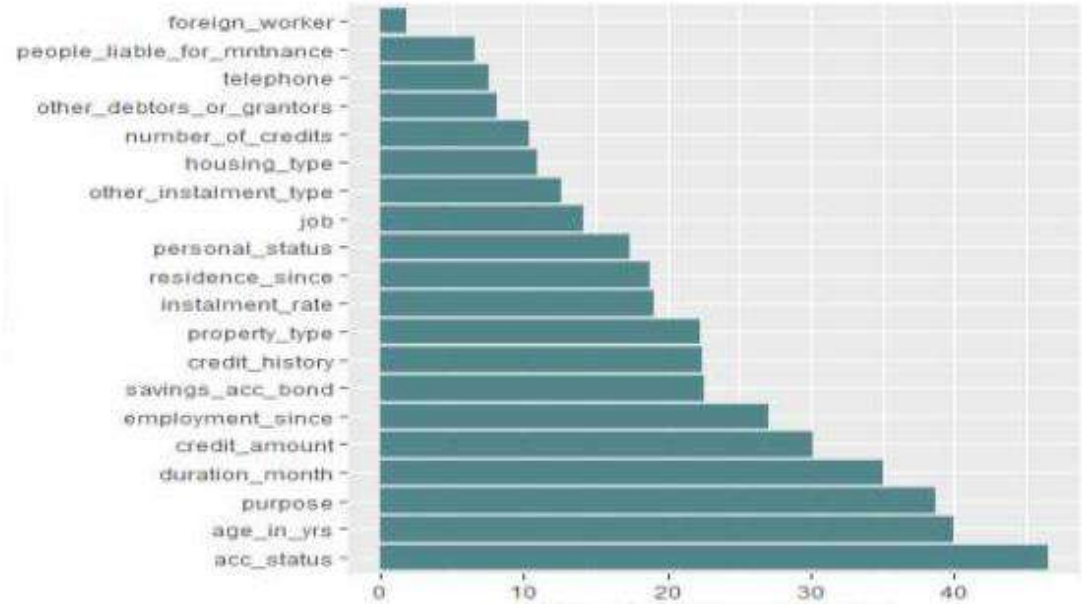
Деректердің сипаттамасы. Деректер жинағы 1000 жазбадан тұрады, оның 70%-ын машиналық оқыту алгоритмдерін жаттықтыруға бөлдік, ал 30%-ын тесттік сынамаға бөлдік. Негізгі мақсатымыз құрылған моделіміздік дұрыс жұмыс жасауына көз жеткізу. Жалпы моделді оқытқанда деректермен шамадан тыс оқыту болмауы керек, ондай жағдайда модельдің дұрыс жұмыс жасауына күмін келтіретін боламыз. Қазіргі қоғамда несие алушы азаматтар көп, олардың төлем қабілеттеріне болжам жасау аса маңызды, сондықтан біз мақалада әрбір несие алушының 20 маңызды сипаттамалары бойынша өңдедік. Жалпы өңдеу процесін сипаттар болсақ жартылау құрылмалған деректер деректер қорында сақталса да болады немесе csv форматтан бірден өңдесекте болады, машиналық оқыту алгоритмдерінің дұрыс жұмыс жасауы үшін модельді жаттықтырамыз жоғарыда айтып өткендей содан кейін ғана тестік деректерді пайдаланып өңдейміз. Толық тізімі төменде 1-суретте берілген.

Сурет 1. Өңделетін деректер жиыны

Зерттеу нәтижелері

Шамадан тыс сәйкестікті болдырмау үшін, біз 10 рет крос-валидацияны орындаймыз. Модельдерде жеке тұлғалардың тек өте маңызды сипаттамалары ғана алынып өңделді. Өңдеу кезінде жалпы машиналық оқытудың 4 алгоритмі алынып ансамбль ретінде қолданылды және қайсы алгоритмдердің қанша көлемді деректерді өңдеудегі беретін өнімділігі зерттелді. Жеке тұлғалардың маңызды сипаттамаларын бағалау 2-суретте көрсетілген.

Ерекшеліктердің маңыздылық диаграммасы



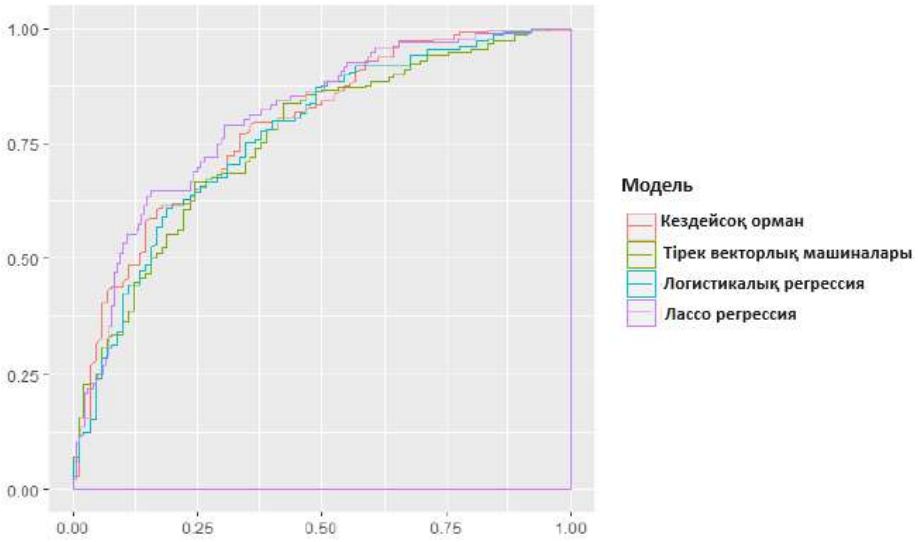
Сурет 2. Өңдеуге қажетті ең маңызды атрибуттар жиынтығы.

Джини коэффициентінің көп төмендеуі алынған сипаттамалардың (атрибуттардың) деректерді өңдеуде және оларды жіктегенде өте маңызды екенін көрсетеді. Біз алдыңғы 20 атрибуттың ішінен тағы өте маңызды деген 10 атрибутты таңдап алдық, себебі ол несие алушының деректерін өңдеп, шешім қабылдауда аса қажет және маңызды. Өңдеу нәтижелері төменде 1-кестеде берілген.

Кесте 1. Модель нәтижелерінің кестесі

Модель	*AUC*	*KS*	*Gini*
Логистикалық регрессия	*74.76*	*42.07*	*53.55*
Кездейсоқ орман	*77.61*	*43.80*	*57.37*
Тірек векторлық машинасы	*74.80*	*42.20*	*51.60*
Лассо регрессия	*78.48*	*47.80*	*60.95*

Қисық астындағы ауданды (AUC) пайдалана отырып, Lasso регрессия моделі жағдайлардың 78.48%-ын дұрыс жіктегені анық болды. Бұл жоғары пайыз регрессиялық модель несие алушыларды жіктеуде өте жақсы модель екенін білдіреді. Лассо регрессиядан кейін кездейсоқ орман моделі 77.61% көрсетті және дәстүрлі жіктеу моделі, логистикалық регрессия моделі арқылы бақыланды. Машиналық оқыту алгоритмдері дәстүрлі модельдерге қарағанда 4% жақсы нәтиже көрсетіп отыр. Қолданылған модельдердің сапасын төменде 3-суреттен бағалауымызға болады.



Сурет 3. Қолданылған модельдер үшін ROC қисықтарының салыстырмасы

ROC қисығы нақты шешімге сәйкес келетін сезімталдық пен ерекшелікті білдіреді. AUROC қисығы алгоритмнің ықтимал топтарды (жақсы және жаман) қаншалықты жақсы ажырата алатынын бағалайды. Нашар және жақсы несиенің екі класы арасында модельдің айтарлықтай айырмашылығы бар-жоғын тексеру үшін Колмогоров-Смирнов тесті жүргізіледі. Егер олар толығымен шашыраңқы болса, 100 мәні қайтарылады, бұл ұпай неғұрлым жоғары болса, модель екеуін жақсырақ ажыратады. 1-кестеде Лассо регрессиясының ең жоғары Колмогоров-Смирнов ұпайы 47,80 болғаны көрсетілген. Тірек вектор машиналары сызықты ядролы, Колмогоров-Смирнов тесті бойынша логистикалық регрессияға қарағанда жақсы нәтиже көрсетті. Джини - екілік классификация моделінің сәйкестік дәрежесін бағалайтын метрика. Басқа сынақтар сияқты, мән неғұрлым жоғары болса, модель соғұрлым жақсы болады. Бағалаудың барлық әдістеріне қарасақ, Лассо ең жақсы, одан кейін кездейсоқ орман, одан кейін сызықты регрессия және соңында тірек векторлық машина (сызықты) моделі екені анықталып отыр.

Дискуссия

Бүгінгі таңда адам санымен қоса деректер көлемі де артуда, сәйкесінше оларды сақтау мен қатар олардың деректерін өңдеу мәселелері өзекті болып отыр. Қаржы секторында машиналық оқыту алгоритмдерін қолдану жартылай құрылымданған деректерді өңдеуде сонымен қатар нәтижелердің дәлдігін анықтауда аса маңызды екенін көрсетіп отыр. Көп жағдайда алынған несиелер қайтарылмай қаржы ұйымдары зиян шегуде. Жеке тұлғалардың сипаттамалары негізінде болжам жасау қазіргі таңда аса маңызды болып отыр. Қаржы саласында оның ішінде ұзақ мерзімге ипотекалық несие алушы жеке тұлғаларға несие беру немесе бермеу туралы шешім шығаруда осы аталмыш машиналық оқыту алгоритмдерін пайдалану болжам нәтижелерінің дәлдігін көрсетіп отыр, сондай-ақ шешім болжам негізінде шешім қабылдауға оң ықпалын тигізіп отыр.

Қорытынды

Зерттеу жұмысының негізгі мақсаты деректерді интеллектуалды талдау негізінде әдістерді қолдану және бағалау болатын. Деректерді талдау негізінде несиелік скоринг үшін ең жақсы жіктеу алгоритмі ретінде Лассо регрессияны айтуға болады. Біздің нәтижелеріміз машиналық оқыту әдістері дәстүрлі логистикалық регрессия әдісінен асып түсетінін көрсетеді. Қаржылық институттарға жеке тұлғалардың деректерңн өңдеу негізінде шешім қабылдау өте қажет екенін зерттедік. Машиналық оқыту алгоритмдері бүгінгі таңда шешім қабылдау да таптырмас құрал деп айтуымызға негіз бар. Біз осы әдістерді пайдалануды ұсынамыз және алгоритмдердің өнімділігін жақсарту үшін гибридті модельдерді де қолдануға болады.

Пайдаланылған дереккөздер тізімі

[1] Machine Learning-Based Application fo rPredicting Risk of Type 2 Diabetes Mellitus (T2DM) inSaudiArabia: A Retrospective Cross-Sectional Study / A.Syed; T. Khan - IEEE Access 30 October 2020 https://ieeexplore.ieee.org/document/9245498

[2]A data analytics approachto building a clinicaldecisionsupportsystemfordiabeticretinopathy: Developinganddeploying a modelensemble / M. Zolbanin, S. Piri, D. Delen, T. Liu - Decision Support Systems Volume 101, September 2017, Pages 12-27 https://www.sciencedirect.com/science/article/abs/pii/S0167923617300908

[3] Quality of Life and Glucose Control After 1 Year of Nationwide Reimbursement of Intermittently Scanned Continuous Glucose Monitoring in Adults Living With Type 1 Diabetes (FUTURE): A Prospective Observational Real-World Cohort Study / S. Charleer, B. Broos, S. Fieuws - Diabetes Care 2020;43(2):389–397 https://diabetesjournals.org/care/article/43/2/389/36133/Quality-of-Life-and-Glucose-Control-After-1-Year

[4] Implementation of a Web Application to Predict Diabetes Disease: An Approach Using Machine Learning Algorithm / S. Dey, A. Hossain, Md. Rahman - Published in 2018 21st International Conference of Computer and Information Technology (ICCIT) https://ieeexplore.ieee.org/document/8631968

[5] Classification Of Diabetes Disease Using Support Vector Machine / V. Anuja Kumari, R.Chitra - March - April 2013 Iternational Journal of Engineering Research and Applications

[6] Improve Classification Performance In Diabetes Prediction / C. Chauhan, S. Karvande - http://www.oaijse.com/VolumeArticles/FullTextPDF/449_8.IMPROVE_CLASSIFICATION_PERFORMANCE_IN.pdf

[7] Real-time crash prediction on freeways using data mining and emerging techniques / J. You, J. Wang, J. Guo - Journal of Modern Transportation volume 25, pages116–123 (2017) https://link.springer.com/article/10.1007/s40534-017-0129-7

[8] Кондрашов Ю.Н. Анализ данных и машинное обучение на платформе MS SQL Server: учебное пособие / Ю.Н. Кондрашов. – Москва : РУСАЙНС, 2020. – 304 р.

[9] Robert L. Learning Data mining with Python. / L. Robert. – Birmingham : Packt Publishing, 2015. – 317 б.

[10] Burns A. Real—Time Systems and Programming Language Ada, Rea|~Time Java and C, Real-Time POSIX / A. Burns, A. - Wellings University of Yor, 2009. – 602 б.

[11] Williams R. Real-Time Systems Development / R. Williams - Waltham: Elsevier, 2006. – 450 б.

[12] Ian H. Data mining Practical Machine Learning Tools and Techniques. / H. Ian, F. Eibe, H. Mark, P. Christopher – Изд. 4-е – Cambridge : Todd Green, 2017. – 622 p.

[13] Jared D. Big Data, Data mining and Machine Learning / D. Jared – Hoboken : John Wiley & Sons, 2014. – 265 б.

[14] Darius M. Data mining for genomics and proteomics Analysis of Gene and Protein Expression Data / M. Darius - Hoboken: John Wiley & Sons, 2010. – 349

[15] Darkenbaev D. Big data processing on the example of credit scoring // Journal of Problems of Informatics and Information Technology. 2023. Vol. 3, No. 1. P. 50–61. doi:10.26577/1i32jpcsit230

References

[1] Machine Learning-BasedApplicationforPredicting Risk of Type 2 Diabetes Mellitus (T2DM) in Saudi Arabia: A RetrospectiveCross-SectionalStudy / A.Syed; T. Khan - IEEE Access 30 October 2020 https://ieeexplore.ieee.org/document/9245498

[2] A data analytics approach to building a clinical decision support system for diabetic retinopathy: Developing and deploying a model ensemble / M. Zolbanin, S. Piri, D. Delen, T. Liu - Decision Support Systems Volume 101, September 2017, Pages 12-27 https://www.sciencedirect.com/science/article/abs/pii/S0167923617300908

[3] Quality of Life and Glucose Control After 1 Year of Nationwide Reimbursement of Intermittently Scanned Continuous Glucose Monitoring in Adults Living With Type 1 Diabetes (FUTURE): A Prospective Observational Real-World Cohort Study / S. Charleer, B. Broos, S. Fieuws - Diabetes Care 2020;43(2):389–397 https://diabetesjournals.org/care/article/43/2/389/36133/Quality-of-Life-and-Glucose-Control-After-1-Year

[4] Implementation of a Web Application to Predict Diabetes Disease: An Approach Using Machine Learning Algorithm / S. Dey, A. Hossain, Md. Rahman - Published in 2018 21st International Conference of Computer and Information Technology (ICCIT) https://ieeexplore.ieee.org/document/8631968

[5] Classification Of Diabetes Disease Using Support Vector Machine / V. Anuja Kumari, R.Chitra - March - April 2013 Iternational Journal of Engineering Research and Applications

[6] Improve Classification Performance In Diabetes Prediction / C. Chauhan, S. Karvande - http://www.oaijse.com/VolumeArticles/FullTextPDF/449_8.IMPROVE_CLASSIFICATION_PERFORMANCE_IN.pdf

[7] Real-time crash prediction on freeways using data mining and emerging techniques / J. You, J. Wang, J. Guo - Journal of Modern Transportation volume 25, pages116–123 (2017) https://link.springer.com/article/10.1007/s40534-017-0129-7

[8] KondrashovYu.N. (2020) Analyz dannyh I mashinnoe obushenie na platforme MS SQL Server [Data analysis and machine learning on the MS SQL Server platform: textbook].Yu.N. Kondrashov. Moscow: RUSAINS, 304.(In Russian)

[9] Robert L. Learning Data mining with Python. / L. Robert. – Birmingham : Packt Publishing, 2015. – 317 p.

[10] Burns A. Real—Time Systems and Programming Language Ada, Rea|~Time Java and C, Real-Time POSIX / A. Burns, A. - Wellings University of Yor, 2009. – 602 б.

[11] Williams R. Real-Time Systems Development / R. Williams - Waltham: Elsevier, 2006. – 450 б.

[12] Ian H. Data mining Practical Machine Learning Tools and Techniques. / H. Ian, F. Eibe, H. Mark, P. Christopher – Изд. 4-е – Cambridge : Todd Green, 2017. – 622 p.

[13] Jared D. Big Data, Data mining and Machine Learning / D. Jared – Hoboken : John Wiley & Sons, 2014. – 265 б.

[14] Darius M. Data mining for genomics and proteomics Analysis of Gene and Protein Expression Data / M. Darius - Hoboken: John Wiley & Sons, 2010. – 349.

[15] Darkenbaev D. Big data processing on the example of credit scoring // Journal of Problems of Informatics and Information Technology. 2023. Vol. 3, No. 1. P. 50–61. doi:10.26577/1i32jpcsit230