МАШИН СУРГАЛТЫН ЗАРИМ АРГААР КРЕДИТ СКОРИНГИЙН ЗАГВАР БОЛОВСРУУЛАХ НЬ: ББСБ-ЫН ЖИШЭЭН ДЭЭР

Б.Оюундарь*, Д.Баянжаргал**, Ё.Чимэдцогзол***, М.Амарбаясгалан****

Хураангуй: Кредит скорингийн шинжилгээ нь санхүүгийн үйлчилгээ үзүүлэгч байгууллагууд зээлдэгчдэд эрсдэлгүй буюу хамгийн бага эрсдэлтэй зээл олгоход тусалдаг. Энэхүү ажлаар зээл олгох шийдвэр гаргахад эрсдэл багатай байх кредит скорингийн загвар боловсруулахыг зорьсон. Сүүлийн үеийн судалгааны ажлуудын үр дүнгээс харахад машин сургалтын аргууд түүн дотроо холимог сургалтын /ensemble learning/ аргаар боловсруулсан загварууд энэ салбарт тэргүүлэх байр суурь эзэлж байна.. Бид энэхүү судалгааны ажлараа А банк бус санхүүгийн байгууллагын 1650 зээлдэгчийн ерөнхий мэдээлэл болон зээлийн түүхийнөгөгдлийг ашиглан холимог сургалтын хоёр /XGBoost, Catboost/ алгоритмаар кредит скорингийн загвар боловсруулж, харьцуулах оролдлого хийсэн. Судалгааны ажлын үр дүнгээс харахад XGBoost алгоритм ашиглан боловсруулсан загварын үр дүн алдааны матриц (confusion matrix), нарийвчлал 0.93% (accuracy), ерcision, recall, f1-score, ROC муруй зэрэг үзүүлэлтүүд хүлээн зөвшөөрөгдөхүйц гарсан. Цаашид энэхүү боловсруулсан загварыг улам сайжруулахын тулд зарим нэмэлт мэдээлэл оруулах хэлбэрээр хувьсагчийн тоог нэмэгдүүлж, турших шаардлагатай.

Түлхүүр үг: Кредит скоринг, машин сургалт, холимог сургалт, Catboost, XGboost

DEVELOPING A CREDIT SCORING MODEL USING SOME MACHINE LEARNING METHODS: A CASE STUDY OF A NON-BANKING FINANCIAL INSTITUTION

Abstract: Credit scoring analysis helps financial service providers offer borrowers loans with zero or minimal risk. Our research aims to develop a credit scoring model to reduce the risk of making loan decisions. According to recent research results, machine learning methods, including ensemble learning models, are leading in this field. In this research work, we have tried to develop and compare credit scoring models using two supervised learning algorithms /XGBoost and Catboost/ using the data of 1650 borrowers' of non-bank financial institution A. According to the research results, the model evaluation matrix developed using the XGBoost algorithm, accuracy value 0.93% (ACC), precision, recall, F1-Score, and ROC curve were acceptable. To further improve this developed model, it is necessary to increase the number of variables in the form of adding some additional information.

Keywords: Credit scoring, machine learning, ensemble learning, catboost, XGboost

* МУИС, Бизнесийн сургууль, (E-mail): oyundari.b@num.edu.mn
** МУИС МУИС-ийн Мэдээллийн Технологи, Электроникийн сургууль, (E-mail): bayanjargal@num.edu.mn
*** МУИС, Бизнесийн сургууль, (E-mail): chimedtsogzolyo@num.edu.mn
**** МУИС, Бизнесийн сургууль, (E-mail): amarbayasgalan.m@num.edu.mn

Оршил

Зээлийн скорингийн загвар нь зээл хүсэгчийн санхүүгийн мэдээлэл болон зан төлөвийн үзүүлэлтэд оноо өгөх байдлаар тухайн зээлдэгчийн зээлийн эрсдэлийг үнэлдэг загвар юм. Ингэж харилцагчийн зээлжих чадварыг тоон хэлбэрээр хэмжсэнээр тухайн хүн зээлээ хэр найдвартай, цаг хугацаанд нь төлөх чадвартайг хялбар харах, харьцуулах боломжтой. Мөн түүнчлэн зээл хүсэж буй харилцагчдыг эрсдэлийн түвшнээр нь ангилах, зээл авах хэмжээг тодорхой болгох, зээлийн багцын эрсдэлийг хэмжих, удирдах боломжтой болно.

Скорингийн загвар нь зээлдэгчийн мэдээлэлд дүн шинжилгээ хийн зээл олгох эсэхийг шийдэхээс гадна цаашлаад зээлийн явцыг хянах, олгосон зээлийн эрсдэлийн шалтгааныг мэдээллийн баазаас түүвэрлэн авч шинжилгээ хийх, засварлах, сайжруулах тасралтгүй процесс юм. Эрсдэлийн шинжилгээнд ашиглагдах боломжтой статистикийн аргачлалуудын хөгжил хурдсаж, эрсдэлтэй уялдуулан зээлийн хүүг тогтоох чиг хандлага бий болсноос гадна хямд, чанартай мэдээлэл боловсруулах боломж нээгдсэн нь зээлийн скорингийн хэрэглээ хурдацтай өссөний шалтгаан болсон. Тэр дундаа хиймэл оюун ухааны хэрэглээ нь энэ салбарын өнөөгийн гол чиг хандлага болоод байна.

Зээлийн скорингийн үнэлгээнд статистикийн арга зүйн үндсийг ашиглах санааг анх 1936 онд Английн статистикч Рональд Фишер боловсруулсан (Fisher, 1936). Тэрээр бие даасан хувьсагчидад үндэслэн бүлгүүдийг ангилах шугаман дискриминант шинжилгээг санал болгосон. (Fisher, 1936). Үүний дараа 1941 онд Дюранд (Durand, 1941) “сайн” болон “муу” зээлийг ангилах судалгааг хийсэн. Энэхүү судалгаа нь банкны харилцагчийн мэдээлэлд хийсэн анхны судалгаануудын нэг байсан бөгөөд судлаач нь хувь хүмүүсийн зээлээ төлөх хандлага, хэрэглэгчдийн эрсдэлийн хүчин зүйлсийн нөлөөг тодорхойлсон байдаг. Санхүүгийн эрсдэлийн гол хүчин зүйлсэд хэрэглэгчийн орлого, зээлийн хэмжээ, зээлийн гэрээний хугацаа, зээлийн баталгаа, бэлэн мөнгөний урсгал, урьдчилгаа төлбөр, зээлдэгчийн хөрөнгө, өр төлбөр зэргийг багтаасан. (Durand) Судалгаанд санхүүгийн бус эрсдэлт хүчин зүйлсэд ажлын байрны тогтвор суурьшил, оршин суугаа газар, ажил мэргэжил, хувь хүний онцлог (хүйс, нас, гэр бүлийн байдал), зээлийн зориулалт зэргийг тодорхойлсон. Өнөө үед Дюрандын санал болгосон аргууд нь орчин үеийн зээлийн онооны загварт өргөн хэрэглэгдэж байна. (Solemne, 18.12.2000) (Division, 2024).

Мэдээллийн технологийн үсрэнгүй хөгжил, АНУ зэрэг томоохон улсуудад санхүүгийн үйлчилгээ үзүүлж буй байгууллагуудад тавигдсан хууль эрх зүйн зохицуулалт, дүрэм журмуудын шаардлага зэрэг нь санхүүгийн салбарын байгууллагуудыг илүү өргөн хүрээнд ажиллахад түлхэц өгсөн. (Furletti,

2002). Ихэнх зээлийн оноо тооцох загваруудын хувьд зээлдэгчийн өнгөрсөн хугацааны төлбөрийн мэдээлэл нь тэдний зээлээ эргэн төлөх хандлага болон чадварыг олж тогтооход ашиглагдах суурь болж байсан (Register, 2011). 1990-ээд онд хэрэглээний зээл, зээлийн карт, моргейжийн зээл зэрэг зээлийн бүтээгдэхүүнүүдэд зээлдэгч эргэн төрөлх үүргээ биелүүлээгүй байх эсвэл хугацаа хэтрүүлэх магадлалыг урьдчилан таамаглахад статистикийн аргыг өргөн ашиглаж байсан (Kenton, 2019). 2000 онд Н.Круук, П.Росс болон Б.Юубас нарын эрдэмтэд анх параметрийн бус аргуудаар банкны кредит картын харилцагчийг сайн болон муу гэж ангилах оролдлого хийсэн. Энэхүү судалгаандаа linear discriminant analysis, neural network, decision tree, genetic algorithms гэсэн 4 загварыг ашигласан байна (M.B.Yobas). 2017 онд FICO компани машин сургалтыг ашиглаж боловсруулсан скорингийн загварыг уламжлалт загварын үр дүнтэй харьцуулан судлахад машин сургалтын гүйцэтгэлийн чадамж илүү өндөр гарсан (FICO, 2018). Мэдээллийн хүртээмж, тооцоолох технологийн өсөлт, зохицуулалтын шаардлага, үр ашиг, эдийн засгийн өсөлтийн эрэлт зэргээс шалтгаалан кредит скорингийн аргуудын хэрэглээ сүүлийн жилүүдэд өссөн (Demirguc-Kunt, 2017) (World Bank). Дэлхийн банкны 2019 оны тайланд дурдсанаар зээлийн скоринг дараах төрлүүдэд хуваагдан хөгжиж байна. Үүнд зээлдэгчийн зээлийн өргөдлийн мэдээлэлд үндэслэсэн, зан төлөвт үндэслэсэн, зээл төлөлтийн, эрсдэлийг эрт илрүүлэх дохионы, залилан илрүүлэлтийн скорингууд тус тус багтаж байна. (Worldbank, n.d.)

Хиймэл оюун ухааны (AI) технологи нь хүний үйл ажиллагааг орлохуйц шийдлүүдийг тооцоолох хэрэгслүүдийн тусламжтайгаар орлуулах боломжийг олгож байна. (Mohamed Ali Mestikou, 2020) Кредит скорингийг боловсруулахад ашигладаг уламжлалт аргуудад шугаман регрессийн загвар, дискриминант шинжилгээ, ложит ба пробит загвар гэх мэт загварууд орно (Fisher, 1936) (Altman, 1968). Банк санхүүгийн салбарын байгууллагын уламжлалт зээлийн бүтээгдэхүүн өөрчлөгдөж, улмаар зээлийн эрсдэлийг бууруулах машин сургалтын алгоритмууд түүн дотроо хяналттай сургалтын (supervised learning) - шийдвэрийн мод, санамсаргүй ой, gradient boosting, deep neural networks, хяналтгүй сургалтын (unsupervised learning) - кластер, к хамгийн ойр хөрш, шаталсан кластер, бататгах сургалт (reinforcement learning) - natural language processing, блокчейн технологид суурилсан төвлөрсөн бус кредит скоринг зэрэг аргууд ашиглагдах болсон (ШУТИС, 2021) (World bank, 2019). AI машин сургалтын дээрх аргуудаас гадна сүүлийн үеийн судалгааны ажлуудын үр дүнгээс харахад кредит скорингийн үнэлгээг хийх түгээмэл бөгөөд илүү ач холбогдол өндөртэй аргуудад санамсаргүй ой, AdaBoost, XGBoost, Catboost, LightGBM болон Stacking гэх мэт холимог сургалтын (ensemble learning) алгоритмууд ашиглагдаж байна (Yiheng Li, 2020) (Bhilare, 2018).

Кредит скоринг нь эдийн засгийн өсөлтийг дэмжих нэг төрлийн нөөц бөгөөд аливаа санхүүгийн байгууллагын үр ашгийг нэмэгдүүлэх үнэ цэнтэй хэрэгсэл болж чадна. Кредит скорингийн хэрэглээ, арга зүй, боломжууд сайжирснаар санхүүгийн хүртээмж нэмэгдэх (Peter Carroll, 2017) аливаа үйл ажиллагаа автоматжих, харилцагчийн үйлчилгээ сайжрах, шийдвэр гаргалтын хугацаа богиносох, оновчтой болох, шударга болох зэрэг давуу талууд бий болдог (Proudman, 2018). Үүний зэрэгцээ тоон өгөгдлийн нууцлал зэрэг асуудлаас шалтгаалж, мэдээллийн нууцлал алдагдах, технологийн талын мэдлэг ур чадвартай ажилтнуудын хомсдолтой байдал үүсэх, хууль эрх зүйн зохицуулалт дутагдах зэрэг анхаарах асуудлууд үүсэж байна (Bloom, 2024). Кредит скорингийн шинэлэг аргуудыг дэмжих технологиуд хөгжиж байгаатай холбоотой тэдгээрийг ашиглах, хэрэглэхтэй холбоотой зохицуулалт, хууль эрх зүй, ёс зүйн тогтолцоог цаг хугацааны явцад боловсронгуй болгож, төлөвшүүлэх нь улс орон бүрийн анхаарах асуудал юм.

Зээлийн эрсдэлийн удирдлагад хиймэл оюун ухааныг ашиглахын зорилго нь компанийн төлбөрийн чадваргүй болох магадлал болон бусад холбогдох тоон өгөгдлөөс үүдэн гарах алдагдлын хэмжээг урьдчилан таамаглахад оршино.

Судалгааны арга зүй

Судалгааны ажлын хүрээнд тоон өгөгдөл ашиглан, машин сургалтын загвар сургах, нэгтгэн дүгнэх, харьцуулсан дүн шинжилгээ хийх буюу тоон судалгааны аргыг ашигласан. Энэ чиглэлийн судалгааны ажлуудад машин сургалтын ангилал хийх параметрийн загварууд болох шугаман регрессийн (linear regression model), пробит болон ложит загвар (probit models and logit models), дискриминант загвар (discrminant analysis) зэрэг загварууд түлхүү ашиглагддаг. Мөн параметр бус шийдвэрийн мод (decision tree), хиймэл оюун ухааны сүлжээ (artificial neural network), олон хувьсагчийн регресс (multivariate adaptive regression splines), математик программчлалын загвар (mathematical hierarchy process), шинжилгээний шатлалын үйл явц (analytical hierarchy process), К хамгийн ойрхон хөршүүд (K-nearest neighbors), эксперт систем (expert system), бейсийн сүлжээ (bayesian network), тулгуур вектор машин (support vector machine), амьдрах чадварын шинжилгээ (survival vector machine), генетик программчлалын загварууд (genetic programming models), опцион үнийн загвар (option pricing model), ангилалд тулгуурласан дүрмүүд (rule-based classification) санамсаргүй ой (random forest) зэрэг загваруудаас гадна сүүлийн жилүүдэд эдгээр аргуудыг нэгтгэж, сайжруулж, холимог сургалтын / ensemble learning/ загваруудыг ашиглах болсон. Холимог сургалтын загварууд нь уламжлалт ангиллын загваруудтай харьцуулахад таамаглалын хувь илүү

байх хандлагатай байна. Тиймээс бид энэхүү судалгаанд XGboost, Catboost аргуудыг зээлдэгчийн зэрэглэл тогтоох загвар сургахад ашигласан.

Зээлийн скорингийн дараах төрлүүд байдгийг дэлхийн банкны судалгааны ажилд дурдсан байна. Үүнд: Өргөдлийн мэдээлэлд үндэслэсэн (application scoring), зан төлөвийн мэдээлэлд үндэслэсэн (behavioral scoring), зээл төлүүлэлтийн мэдээлэлд үндэслэсэн (collection scoring), эрт анхааруулах бусад мэдээлэлд үндэслэсэн (early warning scoring), залилан буюу бусад хууран мэхлэлтийн мэдээлэлд үндэслэсэн (fraud detection scoring) хэмээн ангилжээ. Бидний боловсруулсан загварт зээлдэгчийн анкетийн мэдээлэл, зан төлөвийн мэдээлэл, зээл төлүүлэлтийн мэдээлэл, эрт анхааруулах бусад мэдээллүүд ашиглагдсан. Үүнд хүр системийн мэдээлэл харилцагчийн нас, хүйс, боловсрол, үл хөдлөх хөрөнгөтэй эсэх, тээврийн хэрэгсэлтэй эсэх, цалингийн орлогын хэмжээ, орлого буурсан эсэх мэдээлэл, зээлийн мэдээллийн сангаас идэвхтэй зээлийн тоо, зээлийн ангиллын түүх, харилцагчийн өгсөн мэдээлэл орлогын төрөл, А банк бус санхүүгийн байгууллагын тооцооллоор өр орлогын харьцаа, кредит скоринг, зээлийн эргэн төлөх дүн, үргэлжлэх хугацаа, хугацаа хэтэрсэн зээлтэй эсэх хувьсагчдыг тус тус авч ашигласан.

Энэхүү судалгааны ажлын зорилго нь машин сургалтын аргуудаар А банк бус санхүүгийн байгууллагын харилцагчдын мэдээлэлд үндэслэн зээлдэгчдийн кредит скорингийг урьдчилан таамаглах загвар боловсруулахад оршино. Энэхүү зорилгын хүрээнд дараах зорилтуудыг дэвшүүлсэн. Үүнд:

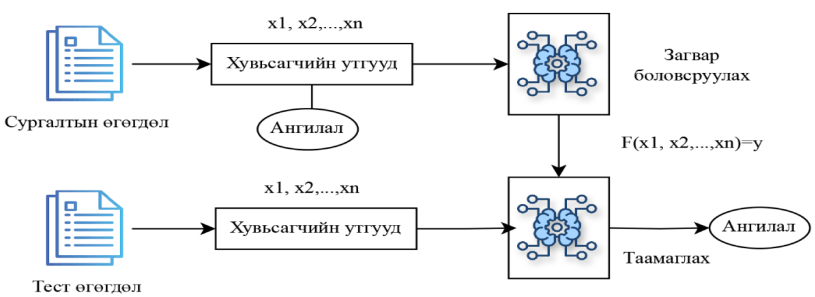
- Холбогдох судалгааны ажлуудыг гадаад болон дотоод эх сурвалжуудаас судлах
- Зээлийн скоринг тооцоход ашиглагдах өгөгдлийг боловсруулах, бэлтгэх
- Машин сургалтын загвар боловсруулж, сургана.

Бид энэхүү судалгааны ажилдаа А банк бус санхүүгийн байгууллагын 1650 зээлдэгчийн өгөгдлийг ашиглан машин сургалтын хяналттай сургалтын аргуудаас XGradient boosting, Catboost гэсэн хоёр аргыг сонгон авч кредит скорингийн загвар боловсруулж, үнэлгээ хийхийг зорьсон. Машин сургалтын алгоритмыг ашиглан кредит скорингийн загварыг дараах үе шатуудтай боловсруулна. Үүнд:

- Түүхий өгөгдөл олж авах, оруулах
- Оруулсан өгөгдлийг шинжлэх, боловсруулах
- Өгөгдлийг хөрвүүлэх
- Шаардлагатай функц, загваруудыг сонгох
- Машин сургалтын алгоритм, загварыг сургах
- Сургалтын үр дүнг тооцох
- Үр дүнг тайлбарлах, загвараа үнэлэх
- Сургасан загвараа ашиглан ирээдүйн тооцоолол хийх

А банк бус санхүүгийн байгууллагын нийт зээлдэгчдийн мэдээллээс 132 чанаргүй зээлдэгчийн тоог нийт зээлийн 8 хувьд тооцож нийт 1650 зээлдэгчийн мэдээлэлд үндэслэн машин сургалтын аргаар загвар сургасан. Нийт зээлдэгчийн мэдээллийг 60:40 хувийн харьцаатай сургалтын болон тестийн өгөгдөл бэлтгэсэн. Сургалтын өгөгдөл нь 990 зээлдэгчийн мэдээлэл агуулсан үүний 79 нь чанаргүй зээлдэгч. Тестийн өгөгдөл нь 660 зээлдэгчийн өгөгдөл ба тэдгээрийн 53 нь чанаргүй зээлдэгч байсан. Сургалтын болон тестийн өгөгдлийг Python программ дээр шинжилгээ хийж, статистикийн зарим үзүүлэлтүүдийг тооцон, хувьсагчийн утгуудын найдвартай байдал болон корреляцийн шинжилгээ хийсэн. Судалгааны загвар нь машин сургалтын XGBoost, Catboost гэсэн төрлийн алгоритм ашиглах бөгөөд үр дүнг confusion matrix, accuracy, percision, recall, f1-score, ROC муруй ашиглан үнэлсэн. Мөн загварыг сургахад хамгийн өндөр ач холбогдолтой хувьсагчийн утгуудыг илрүүлсэн. Машин сургалтын алгоритмуудыг өгөгдсөн даалгавар болон зорилгоос хамааран хяналттай, хяналтгүй, хүч нэмэгдүүлсэн сургалт гэж ангилдаг. Машин сургалтын хяналттай сургалтын арга нь сургалтын өгөгдлийг боловсруулах замаар зорилтот хувьсагчийг таамаглахад чиглэгддэг. Зураг 1-д хяналттай сургалтын алгоритмуудыг сургах бүтэцийг дүрслэв.

Зураг 1. Хяналттай сургалтын загваруудын ерөнхий бүдүүвч.



Эх сурвалж: Судлаачийн боловсруулснаар

Өгөгдлийн хувиргалт

Чанарын хувьсагчдын хувиргалт /One hot encoding/

Бидэнд ***m*** анги бүхий чанарын хувьсагч байна гэж үзье. Эдгээр хувьсагчдыг ижил жинтэйгээр кодлох хэрэгтэй. Вектор v нь m урттай вектор бөгөөд зөвхөн 1 ба 0 гэсэн хоёр утгатай. Үүнийг математикаар илэрхийлбэл

$$\upsilon \in \{0,1\}^m$$
$$\sum_{i=1}^{m} \upsilon_i = 1 \qquad (1)$$

байна. Энд υ_i нь υ векторын нь і-р координат ба хэрэв тухайн хувьсагч і-р ангид орж буй тохиолдолд 1, бусад үед 0 утгатай байна.

Тоон хувьсагчдын хувиргалт /StandardScaler/

StandardScaler нь масштабаар ялгаатай тоон хувьсагчийг нормлодог. Тухайлбал цалин болон нас зэрэг ихээхэн ялгаатай утгатай хувьсагчуудыг машин сургалтын загварт бэлтгэж нормчлохдоо дараах томьёонуудыг голчлон хэрэглэдэг (Overflow, 2019). Үүнд:

1. Стандарт нормаль тархалтын томьёо

$$z = \frac{x-\mu}{\sigma} \qquad (2)$$

Энд μ нь дундаж утга, σ стандарт хазайлт.

2. Min-Max scaling

$$X_{norm} = \frac{X-X_{min}}{X_{max}-X_{min}} \qquad (3)$$

Энд X_{min} нь хамгийн бага утга, X_{max} нь хамгийн их утга юм.

Холимог сургалтын алгоритмууд

XGBoost ангилагч (Extreme Gradient Boost): Энэ алгоритм нь 2014 онд танилцуулагдсан хяналттай сургалтын алгоритм. Энэ нь уламжлалт градентийн алгоритмтай харьцуулахад олон сайжруулалтыг санал болгодог бөгөөд ангиллын болон регрессийн шинжилгээнд аль алинд нь тохиромжтой алгоритм юм. Энэ төрлийн уламжлалт алгоритмуудаас ялгаатай нь XGBoost-ын алдаа тооцох системд загвар хэт суралцахаас (overfiting problem) сэргийлэх зохицуулалтыг нэвтрүүлсэн (Liu et al., 2024). Үүний тулд дараах томьёог ашигладаг.

$$\mathcal{L}(x) = \sum_{i=1}^{\infty} \ell\big(y_i, f(x_i)\big) + \Omega(f) \qquad (4)$$

Catboost ангилагч: CatBoost нь “Categorical Boosting” гэсэн үгний товчлол бөгөөд ангиллын олон регрессийн загваруудад тохиромжтой (Ibrahim et al.,

n.d.; Qi et al., 2021). Энэ алгоритм нь градент бүүстийн аргын алдааг багасгах замаар хоёртын шийдвэрийн модны нэгдлийг давталттайгаар байгуулдаг. Давталт бүрд одоогийн таамаглалтай холбоотой алдааны функцийн градентийн сөрөг утгыг тооцоолж шинэ модыг тохируулна. Сурах хурд (learning rate) нь градентийг багасгах алхмын хэмжээг тодорхойлдог. Урьдчилан тогтоосон тооны мод нэмэх эсвэл нэгдэх шалгуурыг хангах хүртэл процесс давтагдана. Үр дүнг таамаглахдаа CatBoost нь тооцоолсон бүх модны таамаглалыг нэгтгэдэг ба ингэснээр өндөр нарийвчлалтай, найдвартай загварууд бий болдог. Энэхүү загвар нь дараах математик томьёогоор тодорхойлогддог.

Оролтын өгөгдлийг $D = \{(X_j, y_j)\}_{j=1,\dots,m}$ гэвэл, $X_j = (x_j^1, x_j^2, \dots, x_j^n)$ нь n хувьсагчийн вектор, $y_j \in \mathbb{R}$ нь 0,1 гэсэн бинар утгатай гаралтын хувьсагч юм. (X_j, y_j) нь үл хамаарах, санамсаргүй хэмжигдэхүүнүүд $p(.,.)$ нэгэн ижил тархалттай. Загварын үндсэн зорилго нь H: $\mathbb{R}^n \to \mathbb{R}$ функцийг дараах *(5)* томьёогоор өгөгдсөн хүлээгдэж буй алдааг хамгийн бага байхаар сургах юм.

$$\mathcal{L}(H) := \mathbb{E}L(y, H(X)) \qquad (5)$$

$L(.,.)$ ь алдааны гөлгөр функц, (X, y) нь D сургалтын нийт өгөгдлөөс түүвэрлэсэн тестийн өгөгдөл юм. Градент бүүстийн арга нь $H^t: \mathbb{R}^n \to \mathbb{R}, t = 0,1, \dots,$ функцийн дөхөлтүүдийг байгуулдаг. Өмнөх дөхөлт H^{t-1} -ийг ашиглан дараагийн дөхөлт H^t-ийг гарган авахдаа

$$H^t = H^{t-1} + \alpha g^t \qquad (6)$$

томьёо ашигладаг. Энд α алхмын хэмжээ, $g^t: \mathbb{R}^n \to \mathbb{R}$ нь суурь хувьсагчийн функц бөгөөд түүнийг түүнийг (7) томьёогоор илэрхийлэгдэх хүлээгдэж буй алдааг бууруулах эсвэл хамгийн бага утгыг нь олох зорилгоор g функцүүдийн багцаас сонгож авдаг.

$$g^t = \arg min_{g \in G} \mathcal{L}(H^{t-1} + g) = \arg min_{g \in G} \mathbb{E}\mathcal{L}(y, H^{t-1}(X) + g(X)) \qquad (7)$$

Ерөнхийдөө минимумчлах бодлогыг $H^t = H^{t-1} + \alpha g^t$ функцийн 2-р эрэмбийн уламжлалын H^{t-1} цэг дээр утгыг ашиглан Ньютоны аргаар эсвэл градиантын сөрөг утгыг олох замаар боддог (Ibrahim, 2020) (Nguyen, 2022)

Үнэлэх аргууд

Загварын үнэлгээнд алдааны матриц (confision matrix), загварын нарийвчлал (accuracy), precision, recall, F1-Score болон ROC муруйн үнэлгээ зэргийг ашигладаг.

Алдааны матриц (confision matrix) нь гаралтын хувьсагч у-ийн бодит болон таамагласан утгуудын тоон үр дүнг харуулдаг. Хүснэгт 1-д тусгасан хэлбэрээр ихэвчлэн матриц хэлбэрээр үр дүнг харуулдаг.

Хүсэгт 1. Загварын алдааны матриц

		Сөрөг – Negative	Эерэг – Positive
Бодит утга	Сөрөг – Negative	True Negative (TN)	False Negative (FN)
	Эерэг – Positive	False Positive (FP)	True Positive (TP)
		Таамагласан утга	

Нарийвчлал (accuracy) нь нийт зөв таамагласан тоог нийт таамагласан өгөгдлийн тоонд харьцуулсан харьцаа юм.

$$ACC = \frac{TP+TN}{TP+FP+TN+FN} \quad (8)$$

Precision нь үнэн эергийн тоог үнэн эерэг ба худал эерэг тоонуудын нийлбэрт хуваана.

$$Precision = \frac{TP}{TP+FP} \quad (9)$$

Recall нь үнэн эергийн тоог үнэн эерэг ба худал сөрөг тоонуудын нийлбэрт хувааж тооцоолно.

$$Recall = \frac{TP}{TP+FN} \quad (10)$$

F1-Score нь загварын гүйцэтгэлийн үнэлгээг нэг хэмжүүрт нэгтгэдэг. Ихэвчлэн ангиллын загварын үнэлгээнд хэмжүүр болгон ашигладаг. Дараах томьёоллоор тодорхойлогддог.

$$F1\ score = \frac{2 \times Precision \times Recall}{(Precision+Recall)} \quad (11)$$

Receiver operating curve (ROC) муруй нь хоёртын ангиллын зааглагч утгаас хамааруулан үнэн эерэг ба худал эерэг утгуудыг харьцуулсан график.

Судалгааны өгөгдлийн шинжилгээ

Бидний авч ашигласан А банк бус санхүүгийн байгууллагын өгөгдөлд нийт 7 төрлийн чанарын хувьсагч, 8 төрлийн тоон хувьсагчтай бөгөөд хүснэгт 2-т дэлгэрэнгүй харуулав.

Хүснэгт 2. А банк бус санхүүгийн байгууллагын зээлдэгчдийн мэдээлэлд агуулагдаж буй хувьсагчид

№	Хувьсагч	Тэмдэглэгээ	Төрөл	Утга
1	Кредит скоринг	creditscore	Тоо	0-100
2	Өр орлогын харьцаа	dti	Тоо	0-4900
3	Нас	Age	Тоо	21-65
4	Хүйс	Sex	Чанар	Male, Female
5	Боловсрол	Educ	Чанар	Phd, MBA, College_Degree, Middle_school, Highschool, other
6	Зээлийн эргэн төлөх дүн	paidLoanamount	Тоо	0-890,856.941
7	Зээлийн ангилал		Чанар	Performing, Special mention, Substandard
8	Зээлийн тоо	loanCount	Тоо	1-9
9	Автомашинтай эсэх	Car	Чанар	Withcar, Nocar
10	Үл хөдлөх хөрөнгөтэй эсэх	RealEstate	Чанар	Residence, land_and_house, Land,No-Estate,
11	Орлогын хэлбэр	IncomeType	Чанар	Salary, Business
12	Зээлийн хугацаа сараар	Duration	Тоо	12-24
13	Орлого буурсан эсэх	IncomeDecrease	Чанар	Yes, No
14	Сарын орлого	MonthlyIncome	Тоо	851.850-15,291.053
15	Зээлийн эрсдэл	Risk	Тоо	1, 0

Машин сургалтын алгоритмуудыг ашиглан загвар боловсруулахын өмнө өгөгдлийн шинжилгээг зайлшгүй хийх шаардлагатай байдаг. Шинжилгээний энэ хэсэгт өгөгдлийн зүй тогтол, орхигдсон, давхардсан болон алслагдсан утга байгаа эсэхийг тогтоох зэрэг өргөн боломжтой. Хэрэв шаардлагатай бол өгөгдлийг засах, сайжруулах боломж олгодог. Бидний оролтын хувьсагчид давхардсан болон орхигдсон, хэт алслагдсан утга байхгүй байсан.

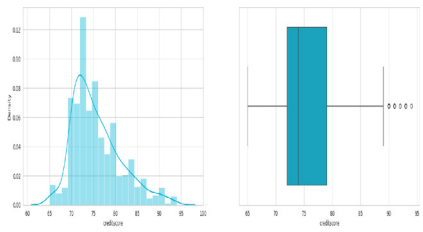
Загвар сургах сургалтын 990 мөр тоон хувьсагчдын статистикийн зарим үзүүлэлтүүдийг хүснэгт 3-т харуулав. Мөн зураг 2-9-д тоон хувьсагчдын тархалт болон гистограм харууллаа. Тухайлбал кредит скоринг үнэлгээг гаргахдаа А банк бус санхүүгийн байгууллага нь зээлдэгчдийг нас, хүйс, боловсрол, орлого зэрэг хэд хэдэн үзүүлэлтүүдээс хамааруулан 0-100 хүртэлх оноогоор үнэлж, 65 буюу түүнээс дээш оноотой зээлдэгчдэд зээл олгосон байна. Энэхүү үнэлгээний дундаж утга 75.7, хамгийн бага утга 65, хамгийн их утга нь 94 байсан. Зураг 3-т харуулсан өр орлогын харьцааны дундаж утга нь 65.1, хамгийн бага утга нь 0, хамгийн их утга нь 4744 байсан нь анхаарал татсан үзүүлэлт байсан.

Хүснэгт 3. Тоон утгатай хувьсагчдын статистик үзүүлэлт

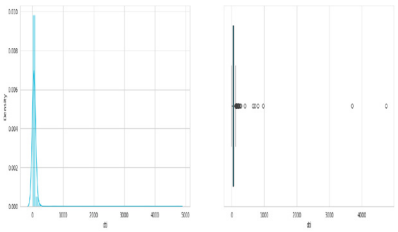
	Тоо үзүүлэлт	Кредит скоринг	Өр орлогын харьцаа	Нас	Төлөх дүн	Зээлийн тоо	Зээлийн эргэн төлөх хугацаа	Сарын орлого	Зээлийн эрсдэл
Mean	75.7	65.1	34.6	51,418.201.2	3	22.4	2,167.188.3	0.1
std	5.5	197.4	7.4	66,755.124.6	2	1.9	1,400.146.4	0.3
min	65	0	21	0	0	12	851,850	0
25%	72	35	29	14,379.401.75	2	22	1,372.397.75	0
50%	74	54	33	32,171.700	3	23	1,766.862.5	0
75%	79	67	38	66,585.479	4	23	2,501.163	0
max	94	4744	63	890,856.943	9	24	15,291.053	1

Зураг 6 д үзүүлснээр зээлийн тооны үзүүлэлтийн дундаж утга 3 хамгийн бага нь 0, их утга 9 байна. Зээлийн хугацааны хувьд дунджаар 22,4 сар хамгийн багадаа 12 сар, хамгийн урт хугацаатай зээл 24 сар хүртэлх хугацаатай байгааг зураг 7-д харуулав. Зээлдэгчдийн 92% нь хэвийн, 8% нь хугацаа хэтэрсэн зээлтэй байна (зураг 9)

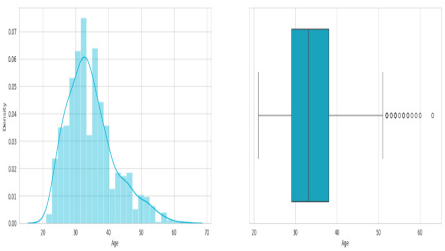
Зураг 2. Кредит скоринг



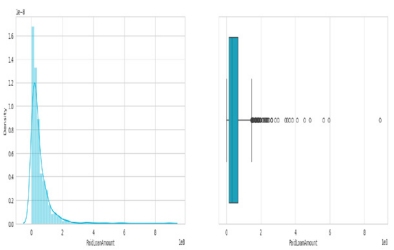
Зураг 3. Өр орлогын харьцаа



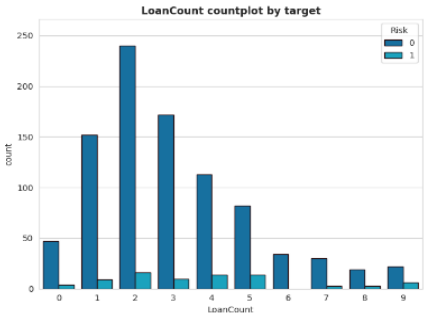
Зураг 4. Нас



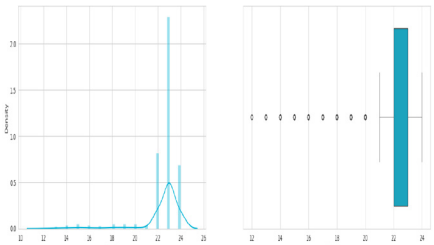
Зураг 5. Зээлийн эргэн төлөх дүн



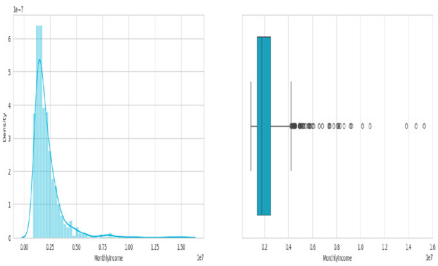
Зураг 6. Зээлийн тоо



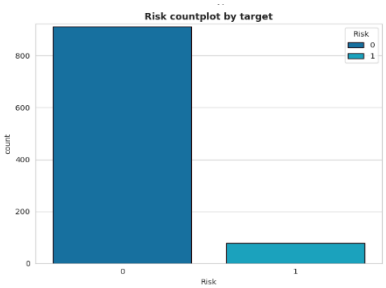
Зураг 7. Зээлийн эргэн төлөх хугацаа



Зураг 8. Сарын орлого



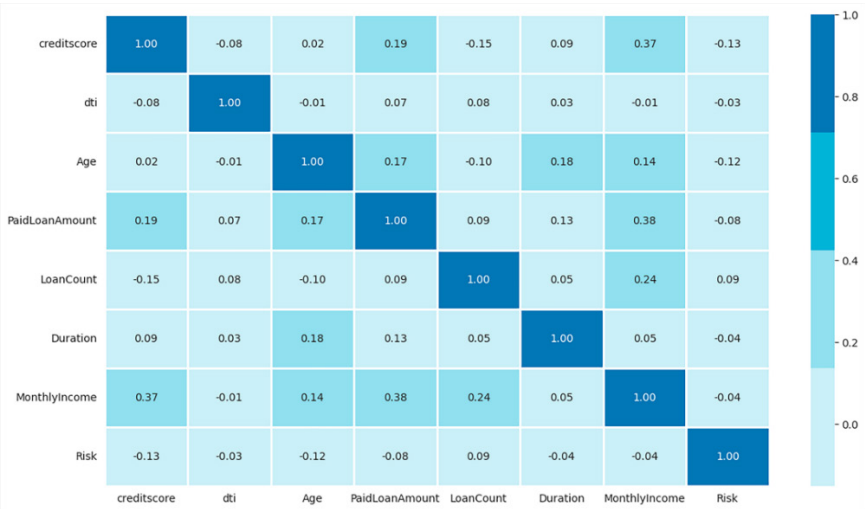
Зураг 9. Эрсдэлтэй эсэх



Нийт зээлдэгчдийн 574 буюу 58% нь эрэгтэй, 416 буюу 42% эмэгтэй байна. Боловсролын түвшингээр нь ангилж үзвэл 11.4% доктор, 49% магистр, 16.5% коллеж төгсөгч, 20.7% бүрэн дунд, 2.3% суурь боловсролтой, 0.1% бусад байна. Зээлийн түүх нь 98% хэвийн, 1.8% анхаарал хандуулах, 0.2% хэвийн бус ангилалтай байна. Зээлдэгчдийн 99.9% нь автомашинтай, 0.1% нь автомашингүй байсан. Үл хөдлөх хөрөнгийн хувьд 48.2% нь орон сууц, 10% хашаа байшин, 8% газар тус тус эзэмшдэг бол 33.8% нь үл хөдлөх хөрөнгөгүй байна. Орлогын хэлбэр нь 100% цалингийн эх үүсвэрээс байсан. Зээлдэгчдийн 88.5%- ийн орлого бууражгүй, 11.5%-ийн орлого буурсан үзүүлэлттэй байна.

Тоон хувьсагчдын корреляцийн шинжилгээ хийж үр дүнг дараах зураг 10-д харуулав. Корреляцийн матрицаас харвал тоон хувьсагчдын хамаарлын утга 0.4-оос бага буюу хувьсагчдын утгууд хоорондоо хэт өндөр хамааралгүй байна.

Зураг 10. Тоон хувьсагчдын корреляцийн матриц



Эх сурвалж: Судлаачийн тооцоолол

Өгөгдлийн хувиргалт

Өгөгдлийн шинжилгээ хийсний дараа чанарын хувьсагчийн утгуудаа OneHotEncoding функц ашиглан тоон утгууд бүхий багануудад, тоон утгатай хувьсагчдын утгыг StandardScaler ашиглан нормчилсон. Түүнчлэн зээлдэгчдийн насыг 10 насны интервалаар тус тусад нь тооцох хувьсагч шинээр нэмж оруулсан.

Загварын үр дүн

Сургалтын өгөгдөлд XGBoost ангилагчийн алгоритм сургаж шалгахад загварын нарийвчлан таамаглах (accuracy) хувь хамгийн ихдээ 0.92%-ийн үр дүн өгсөн. Мөн сургалтын өгөгдөлд Catboost ангилагчийн алгоритмын суралцах хувийг (learning rate) 0.002, сургалтын давталтын тоог 1000 байхаар тохируулан сургахад 0.94 хувьтай таамагласан. Давталтын тоог 1000 гэж өгсөн боловч хамгийн сайн үр дүн өгч буй давталтын утга 262 байсан. Сургасан загваруудад тестийн өгөгдлийг оруулж precision, recall, F1-score, accuracy, ROC муруй үнэлгээнүүдийг хийж дараах үр дүнд хүрсэн. XGboost ангиллын загварын таамаглалын үр дүнг хүснэгт 4 болон 5-д дэлгэрэнгүй харууллаа.

Catboost ангиллын загварын таамаглалын үр дүн *(Хүснэгт 4-5)*.

Хүснэгт 4. XGboost ангилалын загварын таамаглалын үр дүн

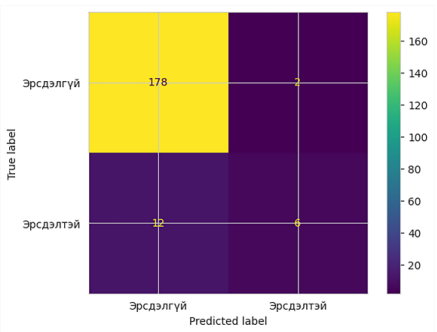
	precision	recall	f1-score	support
Эрсдэлгүй	0.94	0.99	0.96	180
Эрсдэлтэй	0.75	0.33	0.46	18
Accuracy			0.93	198
Macro avg	0.84	0.66	0.71	198
Weighted avg	0.92	0.93	0.92	198

Хүснэгт 5. Catboost ангилалын загварын таамаглалын үр дүн.

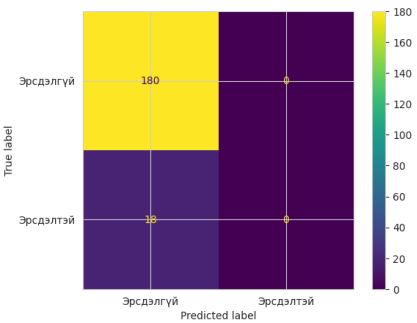
	precision	recall	f1-score	support
Эрсдэлгүй	0.91	1.00	0.95	180
Эрсдэлтэй	0.00	0.00	0.00	18
Accuracy			0.91	198
Macro avg	0.45	0.50	0.48	198
Weighted avg	0.83	0.91	0.87	198

Зураг 11 болон 12-т өмнө сургасан загваруудад тестийн өгөгдлийг ашиглан алдааны матрицийг харьцуулж харвал.

Зураг 11. XGBoost алдааны матриц



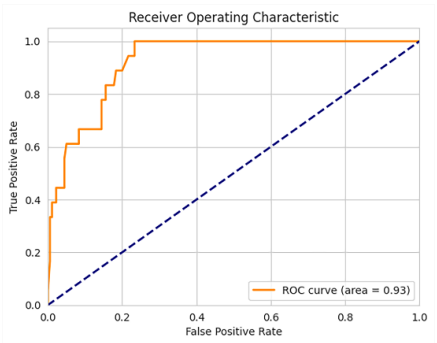
Зураг 12. Catboost алдааны матриц



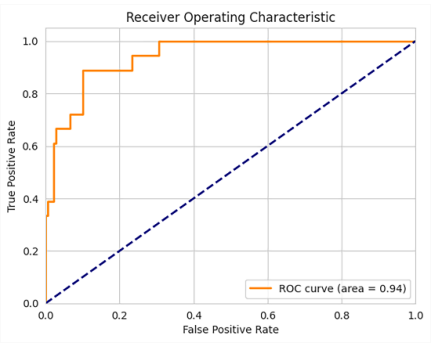
Эх сурвалж: Судлаачийн тооцоолол

Загварын үнэлгээний нэг үр дүн болох ROC муруйг зурж зураг 13-14 д харьцуулан үзүүлэв. Үнэлгээнээс харахад сургалтын өгөгдөлд XGBoost ангилагчийн оновчтой таамаглах хувь 0.93, Catboost ангилагчийн таамаглалын хувь 0.94 гарсан.

Зураг 13. XGBoost ROC curve



Зураг 14. Catboost ROC curve



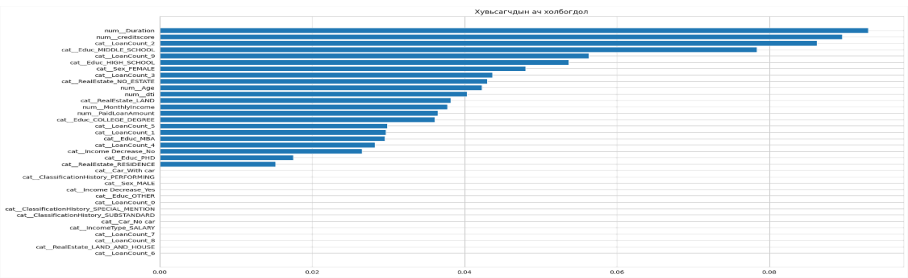
Эх сурвалж: Судлаачийн тооцоолол

Дээрх загваруудыг сургахад ашигласан оролтын утгуудын чухал үзүүлэлтүүдийг зураг 15-д хувьсагч тус бүрээр нь эрэмбэлж гаргасан. Хувьсагчдаас хамгийн өндөр ач холбогдолтой нь зээлийн үргэлжлэх хугацаа түүний дараа кредит скоринг, зээлийн тоо нь 2, дунд боловсролтой, зээлийн тоо нь 9, суурь боловсролтой, эмэгтэй, үл хөдлөх хөрөнгөгүй байх, нас, өр орлогын харьцааны үзүүлэлт, газартай эсэх, сарын орлого зэрэг үзүүлэлтүүд өндөр ач холбогдолтой байна.

Дээрх үр дүнгүүдээс харахад санхүүгийн үйлчилгээ үзүүлэгч байгууллагууд дараах оролдлогуудыг хийх боломжтой.

1. Зээлийн эрсдэл тооцоолоход хамгийн өндөр ач холбогдолтой хувьсагчийн утгуудыг анхаарч кредит скоринг.
2. Бидний сургасан загварт өөрийн харилцагчийн мэдээллийг оруулж, үр дүнг таамаглах оролдлого хийх.
3. Хувьсагчийн утгуудыг нэмэх болон хасах замаар шийдвэр дэмжих загвар боловсруулж турших боломжтой.

Зураг 15. Хувьсагчдын ач холбогдол



Эх сурвалж: Судлаачийн тооцоолол

Дүгнэлт

Судалгааны ажлын хүрээнд машин сургалтын хяналттай сургалтын аргууд болох XGBoost, Catboost ангиллын алгоритмуудыг А банк бус санхүүгийн байгууллагын зээлдэгчдийн өгөгдөл дээр сургаж, зээлийн эрсдэл урьдчилан таамаглах загвар боловсруулахыг зорилоо. А банк бус санхүүгийн байгууллагын нийт 7050 зээлдэгчийн мэдээллээс чанаргүй зээлийн тоо харьцангуй цөөн буюу 132 байсан үүнийг нийт зээлийн 8 хувьд тооцож үлдсэн 92 хувийн зээлдэгчдийг санамсаргүй сонголтын аргаар түүвэрлэж 1650 зээлдэгчийн өгөгдлийг ашиглан 2 төрлийн загварыг Python программ хангамж ашиглан сургасан. Нийт 15 хувьсагчийн 7 чанарын 8 тоон хувьсагч байв.

Эдгээр хувьсагчид нь хүр систем, зээлийн мэдээллийн нэгдсэн сан, харилцагчийн анкетын мэдээлэл болон А банк бус санхүүгийн байгууллагын дотоод тооцооллын үр дүнд бий болсон хувьсагчид байв. Тэдгээр хувьсагчдыг олон улсад хийгдсэн ижил төстэй судалгааны ажлуудад авч ашигласан хувьсагчидтай харьцуулж судлахад загвар сурган тооцоолох бүрэн боломжтой байсан.

Өгөгдөлд давхардсан, алслагдсан, хоосон болон корреляцийн хамаарал бүхий утга байгаа эсэхийг шинжилсний дараа чанарын хувьсагчийг машин сургалтын кодлогч хэрэгслийн тусламжтайгаар тоон хувьсагч руу шилжүүлж нийт 36 багана бүхий өгөгдөл сургалтад ашигласан. Сургалтад ашигласан XGboost, Catboost алгоритмуудын сургалтын үр дүнгийн нарийвчлалын үр дүн 0.01 хувийн зөрүүтэй ойролцоо утгатай гарсан.

Сургасан загвараа ашиглан тестийн өгөгдлийг боловсруулсан туршилтаар алдааны матриц, accuracy, precision, recall, f1-score, ROC муруй зэрэг үнэлгээний үзүүлэлтүүдийн үр дүн XGBoost моделд хүлээн зөвшөөрөгдөхүйц өндөр үзүүлэлтүүдтэй гарч байгаа тул цаашид шинээр зээлдэгчийн мэдээллийг оруулан эрсдэлтэй эсхийг таамаглах боломжтой. Загвараа улам сайжруулахын тулд санамсаргүй түүврийн аргыг олон удаа давтан хэрэглэж, шинээр түүврүүд үүсгэн машин сургалтын аргуудаа хэрэглэх боломжтой. Ингэснээр цаашид амьдралд хэрэглэж болохуйц оновчтой загвар гаргах бүрэн боломжтой юм.

Ашигласан материал:

Altman, E. (1968). Financial Ratios, Discriminant Analysis, and the Prediction of Corporate Bankruptcy. *Journal of Finance, 23 (4)*, 589–609.

Bhilare, A. C. (2018). Application of ensemble models in credit scoring models . *Business Perspectives and Research, vol. 6, no. 2*, 129–141.

Bloom. (2024). Decentralized credit scoring powered by Ethereum and IPFS. (https://www.fintastico.com/services/blockchain/bloom/).

Committee on Technology, N. (2016). Preparing for the Future of Artificial Intelligence. *National Science and Technology Council, Washington, DC*.

Demirguc-Kunt, A. L. (2017). Financial Inclusion and Inclusive Growth: A Review of Recent Empirical Evidence.

Division, C. R. (2024 оны 1 25). *The Equal Credit Opportunity Act*. Гарсан 2024 оны 5 14, https://www.justice.gov/crt/equal-credit-opportunity-act-3-aac

Durand, D. (огноо байхгүй). Risk Elements in Con sumer Instalment Financing . *National Bureau of Economic Research*, 163.

FICO. (2018). How Credit Scoring Helps Me. (https://www.myfico.com/credit-education/creditscores/how-scoring-helps-you).

Fisher, R. A. (1936). “The Use of Multiple Measurements in Taxonomic Problems.” . *Annals of Eugenics , 7 (2)*(https://onlinelibrary.wiley.com/doi/epdf/10.1111/j.1469-1809.1936.tb02137.x.), 179–88.

Furletti, M. (2002). An Overview and History of Credit Reporting. *Payment Cards Center Discussion Paper, Federal Reserve Bank of Philadelphia, Philadelphia* .

Ibrahim, A. M. (2020). Performance of CatBoost classifier and other machine learning methods.

Kenton, W. (2019). *Credit Scoring*. https://www.investopedia.com/terms/c/credit_ scoring.asp-ээс Гарсан

M.B.Yobas, J. C. (огноо байхгүй). Credit scoring using neural and evolutionary techniques.

Mohamed Ali Mestikou, K. E. (2020). Artificial intelligence and machine learning in financial services Market developments and financial stability implications.

Nguyen, N. D. (2022). A Proposed Model for Card Fraud Detection Based on CatBoost and Deep Neural Network. *IEEE Access, 10*(https://doi.org/10.1109/ACCESS.2022.3205416), 96852–96861.

Overflow, S. (2019 оны 01 07). *Stack Overflow*. Гаргасан 13 оны 05 2024, https://stackoverflow.com/questions/40758562/can-anyone-explain-me-standardscaler-aac

Peter Carroll, S. R. (2017). *Alternative data and the unbanked*. https://www.oliverwyman.com/our-expertise/insights/2017/may/alternative-data-and-the-unbanked.html-ээс Гаргасан

Proudman, J. (2018). Cyborg supervision – the application of advanced analytics in. *Research on bank supervision*.

Register, F. (2011). *Fair Credit Reporting* . https://www.federalregister.gov/documents/2011/12/21/2011-31728/fair-creditreporting-regulation-v.-ээс Гаргасан

Solemne, P. (18.12.2000). Charter of fundamental rights. *Official Journal of the European Communities*, 16-21.

World bank. (2019). *Credit scoring approaches guidelines*.

World Bank. (огноо байхгүй). *Policy Research Working Paper 8040*. Washington, DC: http://documents. worldbank.org/curated/en/403611493134249446/ pdf/WPS8040.pdf.

Worldbank. (огноо байхгүй). *worldbank*. https://thedocs.worldbank.org/en/doc/935891585869698451-0130022020/original/CREDITSCORINGAPPROACHESGUIDELINESFINALWEB.pdf-ээс Гаргасан

Yiheng Li, W. C. (2020). A Comparative Performance Assessment of Ensemble Learning for Credit Scorin.

ШУТИС. (2021). *Хиймэл оюун ба машин сургалт 2021*. УБ.