



DOI 10.51885/1561-4212_2025_3_206

FTAXP 28.23.29

КРЕДИТТІК ӨНІМДЕРДІҢ ӘРТҮРЛІ ТҮРЛЕРІ БОЙЫНША СКОРИНГТІК КАРТАЛАР

СКОРИНГОВЫЕ КАРТЫ ДЛЯ РАЗЛИЧНЫХ ТИПОВ КРЕДИТНЫХ ПРОДУКТОВ

SCORING CARDS FOR DIFFERENT TYPES OF CREDIT PRODUCTS

Ж.М. Ордабаева^{1*}, А.Н. Молдагулова¹

¹Қ.И. Сәтбаев атындағы Қазақ ұлттық зерттеу техникалық университеті, Алматы қ., Қазақстан

*Жауапты автор: Ордабаева Жанна Муратовна, e-mail: zhanmaordabayeva@gmail.com

Түйінді сөздер:

кредиттік скоринг
жүйелері, машиналық
оқыту алгоритмдері,
үлкен деректерді өңдеу,
түсіндірілетін жасанды
интеллект (Explainable AI),
модельдердің
тұрақтылығы, деректерді
алдын ала өңдеу, фича
инженериясы, баламалы
деректер, банктік
аналитика

ТҮЙІНДЕМЕ

Жұмыс машиналық оқыту әдістерін және баламалы деректер көздерін пайдалана отырып, несиелік өнімдердің әртүрлі түрлері үшін баллдық үлгілерді жасаудың заманауи әдістерін ескереді. Модельдер (логистикалық регрессия, Random Forest, XGBoost және LSTM) бірінші рет Қазақстанның екінші деңгейлі банктері үшін олардың түсіндірмелілігіне, ұзақ мерзімді тұрақтылығына және әртүрлі несиелік нарықтар үшін жарамдылығына назар аудара отырып, салыстырылды. Зерттеудің ғылыми жаңалығы оның түсіндірілетін AI әдістерін (SHAP, PDP) және балама деректерді (сандық мінез-құлық, транзакциялар және телекоммуникация деректері) бағалау карталарын жасау және халық тұрақтылығы индексін (PSI) пайдалана отырып, уақытша тұрақтылық үлгілерін бағалау үшін қалай пайдаланатынында жатыр. Нәтижелер XGBoost ең жоғары дәлдікті (AUC = 0,88) көрсететінін, Random Forest сапа мен түсіндіру арасындағы тепе-теңдікті (AUC = 0,82) ұсынатынын және логистикалық регрессия оның ашықтығы мен ережелерді сақтауына байланысты реттелетін өнімдер үшін негізгі үлгі болып қала беретінін көрсетеді.

Ключевые слова:

системы кредитного
скоринга, алгоритмы
машинного обучения,
обработка больших
данных, объяснимый
искусственный интеллект
(Explainable AI),
устойчивость моделей,
предварительная
обработка данных, фича-
инжиниринг,
альтернативные данные,
банковская аналитика

АННОТАЦИЯ

В работе учтены современные методики создания скоринговых моделей для различных видов кредитных продуктов с использованием методов машинного обучения и альтернативных источников данных. Впервые проведено сравнение моделей (логистическая регрессия, случайный лес, XGBoost и LSTM) для банков второго уровня Казахстана с акцентом на их интерпретируемость, долгосрочную устойчивость и пригодность для различных рынков кредитования. Научная новизна исследования заключается в использовании методов объяснимого ИИ (SHAP, PDP) и альтернативных данных (цифровое поведение, транзакции и данные телекоммуникаций) для создания скоринговых карт и оценки моделей временной стабильности с использованием индекса стабильности населения (PSI). Результаты показывают, что XGBoost демонстрирует наивысшую точность (AUC = 0,88), случайный лес обеспечивает баланс между качеством и интерпретируемостью (AUC = 0,82),



а логистическая регрессия продолжает оставаться основополагающей моделью для регулируемых продуктов благодаря своей прозрачности и соблюдению нормативных требований.

Keywords:

credit scoring systems, machine learning algorithms, big data processing, explainable AI, model stability, data preprocessing, feature engineering, alternative data, banking analytics

ABSTRACT

The work takes into account contemporary techniques for creating scoring models for different kinds of credit products using machine learning techniques and alternative data sources. The models (logistic regression, Random Forest, XGBoost, and LSTM) were compared for the first time for Kazakhstan's second-level banks, with a focus on their interpretability, long-term stability, and suitability for various lending markets. The study's scientific novelty lies in how it uses Explainable AI techniques (SHAP, PDP) and alternative data (digital behavior, transactions, and telecom data) to create scoring cards and assess temporal stability models using the population stability index (PSI). The findings indicate that XGBoost exhibits the highest accuracy (AUC = 0.88), Random Forest offers a balance between quality and interpretability (AUC = 0.82), and logistic regression continues to be the fundamental model for regulated products because of its transparency and adherence to regulations.

КІРІСПЕ

Қаржы жүйесінде тәуекелдерді басқару құралдарының бірі ретінде кредиттік скоринг ерекше орын алады. Қарыз алушылардың төлем қабілеттілігін дәл бағалау банктердің қаржылық тұрақтылығын қамтамасыз ету үшін ғана емес, сонымен бірге бүкіл қаржы нарығының сенімді жұмыс істеуі үшін маңызды. Дәстүрлі скорингтік әдістер негізінен қарапайым статистикалық тәсілдерге сүйеніп, қарыз алушының табысы, жасы, несиелік тарихы сияқты құрылымдалған көрсеткіштерді қолданды. Бұл тәсілдер ұзақ уақыт бойы тиімділігін дәлелдегенімен, қазіргі қаржы нарығының күрделілігі олардың шектеулерін айқын көрсетті.

Соңғы онжылдықта цифрландыру үдерісінің жеделдеуі нәтижесінде клиенттік база едәуір ұлғайып, деректердің жаңа түрлері пайда болды. Қаржы ұйымдары тек дәстүрлі құрылымдалған мәліметтермен шектелмей, транзакциялық деректерді, пайдаланушылардың цифрлық белсенділігін, әлеуметтік желілердегі мінез-құлқын және телеком-операторлардан алынған қосымша ақпаратты пайдалана бастады. Мұндай баламалы деректер кредиттік скоринг модельдерінің болжамдық қабілетін арттырып, қарыз алушылардың мінез-құлқын анағұрлым кешенді сипаттауға мүмкіндік береді.

Сонымен қатар, деректердің көлемі де экспоненциалды өсуде. Банктер мен финтех-компаниялардың ақпараттық жүйелерінде миллиондаған жазбалар жинақталып, дәстүрлі өңдеу әдістерінің өнімділігі жеткіліксіз бола бастады. Мұндай жағдайда Big Data технологияларын (Apache Spark, Hadoop, бұлттық инфрақұрылымдар) пайдалану қажеттілігі туындайды. Олар үлкен көлемдегі деректерді параллельді өңдеуді қамтамасыз етіп, нақты уақыт режимінде скоринг жүргізуге мүмкіндік береді.

Ғылыми әдебиеттерде кредиттік скорингке қатысты көптеген зерттеулер бар. Классикалық еңбектерде логистикалық регрессия негізгі модель ретінде қолданылды (Thomas, 2009; Hand & Henley, 1997). Кейінгі зерттеулерде ансамбльдік әдістер – Random Forest, Gradient Boosting және олардың нұсқалары кеңінен қолданыла бастады (Baesens et al., 2016). Соңғы жылдары терең оқыту архитектуралары, соның ішінде рекурренттік нейрондық желілер (LSTM) транзакциялық және уақыттық деректерді талдау үшін пайдаланылып жүр. Дегенмен, Қазақстандағы және жалпы Орталық Азиядағы жарияланымдарда баламалы деректерді интеграциялау, Explainable AI (мысалы, SHAP,



PDP) құралдарын қолдану және модельдердің уақыт бойынша тұрақтылығын PSI арқылы бағалау сирек қарастырылады.

Осыған байланысты, бұл зерттеу бірқатар ғылыми және практикалық мәселелерді шешуге бағытталған:

- құрылымдалған, құрылымдалмаған және баламалы деректерді салыстыра отырып, әртүрлі машиналық оқыту модельдерінің сапасын бағалау;
- алгоритмдердің интерпретациясын қамтамасыз ету үшін Explainable AI әдістерін енгізу;
- модельдердің уақыт бойынша тұрақтылығын Population Stability Index (PSI) арқылы талдау;
- әртүрлі несиелік өнімдерін тек мысал ретінде қолданып, әдістердің әмбебаптығын көрсету.

Зерттеудің мақсаты – кредиттік скоринг жүйелерін жетілдіру үшін заманауи машиналық оқыту алгоритмдерінің (Logistic Regression, Random Forest, XGBoost, LSTM) болжамдық қабілетін, интерпретациясын және тұрақтылығын кешенді бағалау. Несиелік өнімдері (тұтынушылық, ипотекалық, автонеие, қайта қаржыландыру, ШОБ, P2P) тек тәжірибелік сынақ алаңы ретінде пайдаланылып, басты назар деректердің түрлері мен алгоритмдердің сапасына аударылды.

ЗЕРТТЕУ МАТЕРИАЛДАРЫ МЕН ӘДІСТЕРІ

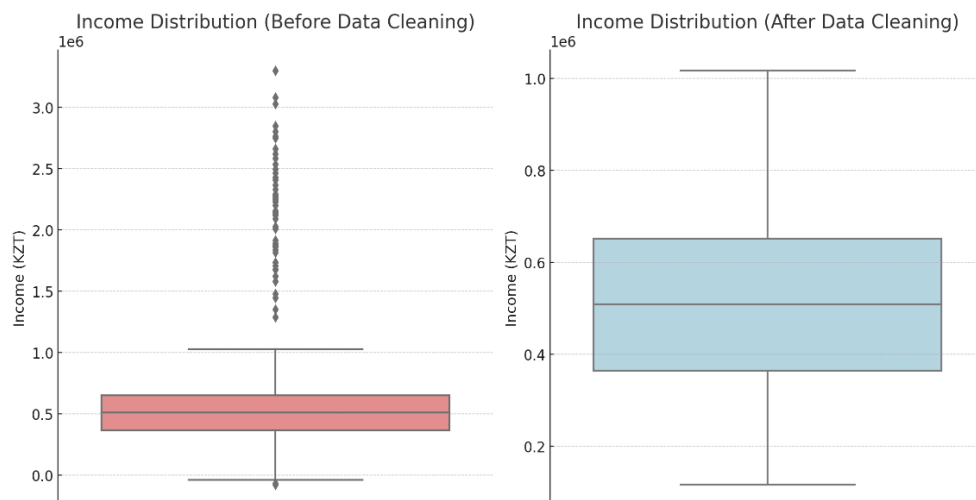
Бірнеше несиелік өнім түрлерінің қарыз алушылары туралы ақпаратты қамтитын мәліметтер жиынтығы осы зерттеу үшін жиналды. Деректерді жинау және дайындау бірнеше шешуші кезеңдерді қамтыды (Khandani, Kim & Lo, 2010).

Бірнеше көздерден, соның ішінде несиелік бюролардан, финтех платформалар мен коммерциялық банктерден несиелік өтінімдердің жазбаларын жинау бастапқы кезең болды. Біз несиелік өнімдердің кең спектрін қамтитын іріктеменің репрезентативтілігін ескере отырып, деректерді таңдадық (Baesens et al., 2016). Оларға тұтынушылық несиелер, қайта қаржыландыру, шағын және орта фирмаларға (ШОБ) несиелер беру, автокредиттер, ипотека және жеке-жеке несиелер жатады. Зерттеудің өзектілігін арттыру үшін зерттеуде несиелік скорингтің әдеттегі және баламалы элементтері қолданылды (Malekipirbazari & Aksakalli, 2015).

Деректерді одан әрі өңдеу ақпараттың толықтығы мен сапасын арттыру үшін жиналған алғашқы мәліметтерден кейін орындалды. Біз қарыз алушылардың несиелік қабілеттілігінің көрсеткіштерін, демографиялық ақпаратты және қаржылық ерекшеліктерін алдық. Әсіресе P2P несиелеу үлгілері үшін зерттеу сонымен қатар басқа деректер көздерінің арасында цифрлық мінез-құлық шараларын және транзакция деректерін қамтыды (Bastani et al., 2019; Berg et al., 2020).

1-кесте. Деректерді таңдаудың сипаттамасы

Несие түрі	Қарыз алушылардың саны	Орташа жасы	Орташа табыс (теңге)	Дефолттардың пайызы (%)
Тұтынушылық несиелер	10 000	35,2	450 000	5,8
Қайта қаржыландыру	7 500	39,8	520 000	4,5
ШОБ несиелеу	5 000	42,1	1 200 000	12,3
Автокредиттер	3 500	37,5	600 000	3,7
Ипотека	4 200	40,2	750 000	2,1
P2P несиелеу	6 000	29,8	300 000	15,4
Ескерту: авторлар ашық дереккөзден алынған мәліметтер негізінде құрастырған				



1-сурет. Деректерді тазалаудың қарыз алушылардың кірістерін бөлуге әсері.

Ескерту: авторлар синтетикалық деректер негізінде құрастырған

Деректерді түрлендіру және тазарту әдістері – олқылықтармен жұмыс істеу, қаржылық айнымалыларды қалыпқа келтіру және категориялық ерекшеліктерді кодтау – соңғы кезеңді белгіледі. Бұл деректерді келесі модельдеу мен талдауда орынды қолдануға кепілдік берді (Zhou et al., 2017; Chen et al., 2016).

Деректерді тазаламас бұрын, бірінші деректер сапа үлгісіне нұқсан келтіруі мүмкін кірістері өте жоғары және төмен шығарындыларды көрсетеді. Келесі деректерді тазарту барысында табыс ауқымы теңдестіріліп, деректер таза және модельдеуге жарамды күйге келтірілді.

Мәліметтерді алдын ала өңдеу

Несиелік тәуекел моделінің болжау мүмкіндіктерін жақсарту үшін мұқият деректерді алдын ала өңдеу құбыры орнатылды (Kelleher et al., 2020). Деректерді тазалау, мүмкіндіктерді жобалау, айнымалы түрлендіру, кодтау, масштабтау, деректерді бөлу және сыныптың теңгерімсіздігін шешу – процестегі дәйекті қадамдардың бірі болды (Zhang et al., 2019). Әрбір қадам машиналық оқыту алгоритмдері үшін жоғары сапалы кірістерді шығару кезінде қиғаштықты азайту және түсіндіруді қолдау үшін жасалған.

Деректерді тазарту

Деректер жинағы алдымен шектен тыс мәндерден және қайталанатын жазбалардан құтылу үшін тазартылды. Айырмашылықтарды табу үшін кватильаралық диапазон (IQR) және Z-балл әдістері қолданылды (Liu et al., 2004). Қималанған үлестірімдердің әсерін азайту үшін сандық айнымалылардағы жетіспейтін мәндер медиана мәнімен толықтырылды. Деректердің болмауына байланысты ықтимал нұсқаушы үлгілерді сақтау үшін категориялық айнымалылар үшін нақты «жоқ» санаты белгіленді (Saar-Tsechansky & Provost, 2007).

Зерттеу барысында Қазақстанның екінші деңгейлі банктерінен, кредиттік бюролардан және финтех-платформаларынан алынған қарыз алушылар жөніндегі кешенді деректер жиынтығы пайдаланылды. Жалпы деректер көлемі **31 200 жазбаны** құрайды, бұл нәтижелердің репрезентативтілігі мен статистикалық сенімділігін қамтамасыз етеді.

Іріктеменің құрылымы төмендегідей:

- 10 000 тұтынушылық несие;
- 7 500 қайта қаржыландыру жағдайлары;



- 5 000 ипотекалық несие;
- 4 200 автонесие;
- 2 500 шағын және орта бизнес (ШОБ) клиенттері;
- 2 000 P2P және финтех-платформалар деректері.

Зерттеу барысында қарыз алушыларға қатысты деректер бірнеше көзден алынды:

- Кредиттік бюро – қарыз алушының несие тарихы, бюроға жасалған сұраулар саны;
- Екінші деңгейлі банктер – табыс, жас, қарыз сомасы, қарыздық міндеттемелер;
- Финтех-компаниялар – транзакциялық белсенділік, төлем жиілігі, цифрлық мінез-

құлық индикаторлары;

- P2P платформалар – баламалы деректер (онлайн-активтілік, әлеуметтік желілердегі із қалдыру, телеком-операторлардан алынған ақпарат).

2-кесте. Модель айнымалылары және олардың сипаттамасы

Айнымалы	Сипаттамасы	Деректер түрі	Дерек көзі
Табыс	Қарыз алушының ай сайынғы расталған табысы	Құрылымдалған	Банк
Жасы	Өтініш берген сәттегі толық жас	Құрылымдалған	Банк
Несие сомасы	Сұратылған несие мөлшері	Құрылымдалған	Банк
Кредиттік міндеттемелер	Қарыз алушының барлық ағымдағы несие қарыздарының жиынтығы	Құрылымдалған	Банк / Кредиттік бюро
Бюроға сұраулар саны	Соңғы 12 ай ішінде кредиттік бюроға жасалған сұраулар саны	Құрылымдалған	Кредиттік бюро
DTI (Debt-to-Income)	Қарыздық жүктеменің табысқа қатынасы	Есептелген метрика	Банк / Финтех
PTI (Payment-to-Income)	Ай сайынғы төлемнің табысқа қатынасы	Есептелген метрика	Банк
Цифрлық мінез-құлық	Клиенттің онлайн-банкінг, төлемдер және мобильді қосымшалардағы белсенділігі	Құрылымдалмаған	Финтех
Телеком-деректер	Қоңыраулар жиілігі, геолокациялық үлгілер	Баламалы	Телеком-оператор
Әлеуметтік желілер	Қарыз алушының белсенділігі, желілік қауымдастықтарға қатысуы, мінез-құлық индикаторлары		

Ескерту: авторлар ашық дереккөзден алынған мәліметтер негізінде құрастырған

Жиналған ақпарат үш негізгі категорияға бөлінді:

Құрылымдалған деректер – қарыз алушының табысы, жасы, несие сомасы, кредиттік міндеттемелері, кредиттік бюроға жасалған сұраулар саны.

Құрылымдалмаған деректер – транзакциялық белсенділік, төлемдердің жиілігі мен тұрақтылығы, цифрлық мінез-құлық ерекшеліктері.

Баламалы деректер – телеком-операторлардан алынған ақпарат (қоңырау жиілігі, геолокациялық үлгілер), әлеуметтік желідегі белсенділік, салықтық және коммуналдық төлемдер.

Осындай деректердің үйлесімі дәстүрлі скоринг әдістерін заманауи цифрлық іздер мен мінез-құлық индикаторларына негізделген тәсілдермен салыстыруға мүмкіндік береді.



3-кесте. Деректердің түрлері мен мысалдары

Деректер категориясы	Айнымалылар мысалдары	Қолдану саласы
Құрылымдалған	Табысы, жасы, несие сомасы, міндеттемелері, бюро сұраулары	Барлық несие өнімдері
Құрылымдалмаған	Транзакциялар, төлем жиілігі, цифрлық мінез-құлық	P2P, финтех, тұтынушылық несиелер
Баламалы	Телеком деректері, әлеуметтік желілер, салықтық төлемдер	ШОБ, P2P, қайта қаржыландыру
<i>Ескерту: авторлар ашық дереккөзден алынған мәліметтер негізінде құрастырған</i>		

Үлкен деректерді өңдеу

Зерттеу барысында қолданылған деректер көлемі 31 200 жазбамен шектелгенімен, өндірістік жағдайда екінші деңгейлі банктер мен финтех-платформаларда скорингтік жүйелердің дерекқорлары миллиондаған жазбаны қамтиды. Мұндай жағдайда дәстүрлі әдістер деректерді өңдеу жылдамдығы мен есептеу қуаттылығына байланысты шектеулерге тап болады.

Big Data технологияларын пайдалану ірі көлемдегі деректерді өңдеу тиімділігін арттырады және нақты уақыт режимінде (real-time) скоринг жүргізуге мүмкіндік береді. Атап айтқанда, төмендегі тәсілдер ұсынылады:

Apache Spark және Hadoop – таратылған есептеулерді ұйымдастыру, параллельді деректерді өңдеу және үлкен көлемді транзакциялық ағындарды талдау үшін.

Облачные решения (AWS, Google Cloud, Azure) – масштабталатын инфрақұрылымды қамтамасыз етіп, модельдерді онлайн-режимде оқытуға және қолдауға мүмкіндік береді.

Таратылған дерекқорлар (HDFS, NoSQL) – құрылымдалмаған және жартылай құрылымдалған деректермен (телеком, әлеуметтік желілер) жұмыс істеу үшін.

Бұл тәсілдер скорингтік модельдердің өнімділігін арттырып қана қоймай, деректердің уақытылы жаңартылуын, болжамдардың жедел есептелуін және бизнес-процестердің автоматтандырылуын қамтамасыз етеді.

Осылайша, Big Data инфрақұрылымын енгізу қаржы институттарына үлкен көлемдегі клиенттік деректерді тиімді өңдеуге, модельдердің сапасын тұрақты бақылауға және тәуекелдерді басқарудың жаңа деңгейіне шығуға мүмкіндік береді.

Деректерді бөлу және валидация

Зерттеу барысында деректер жиынтығы оқыту (train) және тест (test) жиынтықтарына 80/20 қатынасында бөлінді. Модельдің нәтижелерін тұрақты бағалау үшін қатпарлы кросс-валидация (k-fold cross-validation) әдісі қолданылды, мұндағы $k = 5$. Бұл тәсіл әрбір модельдің өнімділігін бірнеше итерацияда тексеруге және орташа нәтижелерді алуға мүмкіндік берді.

Маңызды ерекшелік ретінде стратификация принципі сақталды: яғни, деректерді бөлу кезінде мақсатты айнымалының (дефолт / дефолт емес) пропорциясы барлық жиынтықтарда бірдей деңгейде қамтамасыз етілді. Бұл модельдердің теңгерімді оқытуына және бағалау нәтижелерінің объективтілігіне жағдай жасады.

Инженерлік ерекшеліктері

Деректер жиынының ақпараттық мәнін жақсарту үшін жаңа болжамды айнымалылар жасалды (Chakraborty et al., 2022). Олардың арасында қарыздың кіріске қатынасы (DTI) және төлемнің кіріске қатынасы (PTI) сияқты несие қабілеттілігінің өлшемдері болды. Несие лимитінің көрсеткіші де қарыз алушының ағымдағы қарыздары мен орташа айлық табысын пайдалана отырып есептелді. Қаржылық тәртіп пен өтімділік тәртібін сипаттау



үшін транзакциялық деректерге негізделген айнымалылар – төлемдердің жиілігі, транзакцияның орташа сомасы және шағын несиелердің болуы – қолданылды (Zhang et al., 2020).

Өзгермелі түрлендіру

Модельдің тұрақтылығын жақсарту және үлестірімнің қиғаштығын азайту үшін кіріс және несие сомасы сияқты үздіксіз айнымалылар логарифмдік түрлендіруге ұшырады (Tukey, 1977). Қарыз алушылар кіріс деңгейіне байланысты алдын ала белгіленген шектермен топтарға бөлінді - бұл тәсіл түсіндіруді жеңілдетіп, сызықтық емес тәуелділікті жақсырақ қамтуға мүмкіндік берді (Kuhn & Johnson, 2013).

Категориялық айнымалыларды кодтау

Номиналды категориялық айнымалылар бір ыстық кодтау (ONE) әдісі арқылы түрлендірілді (Pedregosa et al., 2011). Реттік немесе монотонды категориялық айнымалылар үшін дәлелдердің салмағы (WoE) кодтауы пайдаланылды – бұл логистикалық регрессияда интерпретацияны сақтау үшін маңызды болды (Hand & Henley, 1997).

Стандарттау және нормалау

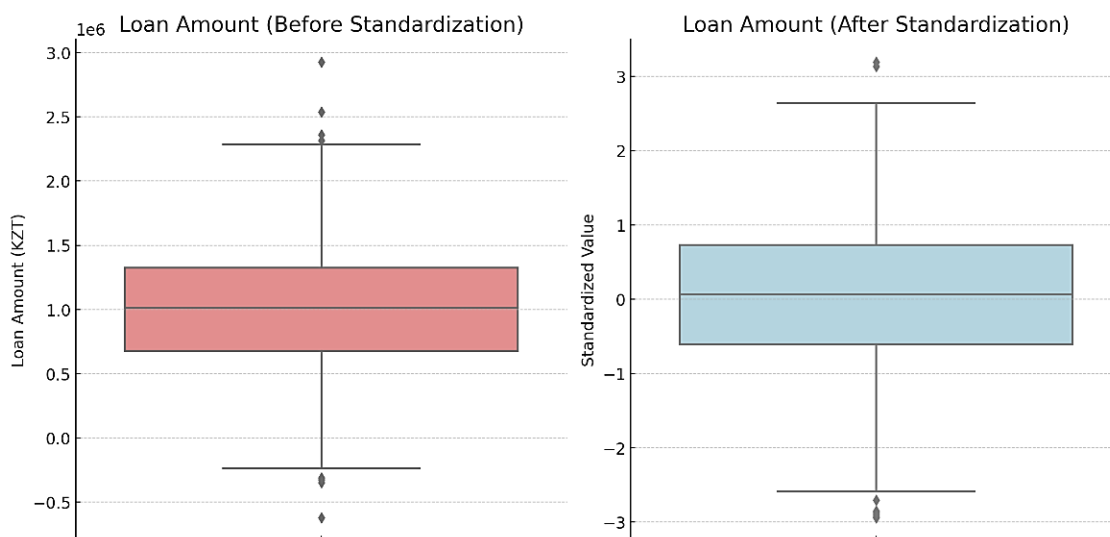
Айнымалылардың әртүрлі масштабтарын біріздендіру үшін Min-Max шкаласы немесе Z-балл (стандартты) нормализация қолданылды (Juszczak et al., 2002). Бұл масштабқа сезімтал модельдер (мысалы, логистикалық регрессия немесе нейрондық желілер) үшін маңызды болды.

Сынақ-поезды бөлу

Мәліметтер 80:20 қатынасында оқыту және сынау жиындарына бөлінді. Мақсатты айнымалы - әдепкі - бойынша стратификацияланған іріктеу қолданылды, бұл үлгілердің сынып бойынша бастапқы таралуын сақтауға көмектесті (Kohavi, 1995).

Сыныптағы теңгерімсіздікті шешу

Төленбеген несиелер 10 %-дан аз болғандықтан, теңгерімсіздікті шешу үшін бірқатар тәсілдер қолданылды. Оларға SMOTE сияқты артық іріктеу (Chawla et al., 2002) және жеткіліксіз іріктеу кірді. Сонымен қатар, үлгілердің жоғалту функциясына сынып салмағын қосу арқылы сирек кездесетін дефолт оқиғаларға сезімталдықты арттыру көзделді (He & Garcia, 2009).



2-сурет. Деректерді стандарттаудың несие сомасына әсері

Ескерту: авторлар ашық дереккөзден алынған мәліметтер негізінде құрастырған



Несие сомасы: Стандарттау алдында көрсеткіштердің кең спектрі және мүмкін ауытқулары бар. Несие сомасы (Стандарттаудан кейін) – Стандарттаудан кейін (Z-балл) мәндер төмендетіледі нөлдік орташа және бірлік дисперсиясы бар қалыпты үлестірімге.

Бағалау үлгілерін құру

Бұл зерттеу тең дәрежелі (P2P) несиелеу, тұтынушылық несиелер, қайта қаржыландыру бағдарламалары және ШОБ қаржыландыру сияқты әртүрлі несие өнімдері үшін мамандандырылған несиелік скоринг үлгілерін жасады. Модельдеу тәсілдерін таңдаған кезде қолда бар деректердің сапасы, несие өнімдерінің түрі және үлгіні түсіндірудің қажетті дәрежесі ескерілді (Thomas et al., 2017).

Түсіндірудің жоғары дәрежесіне, салыстырмалы түрде шағын деректер жинақтары бар тұрақтылыққа және нормативтік сәйкестікке байланысты логистикалық регрессия тұтынушылық және ипотекалық несиелер үшін базалық үлгі ретінде таңдалды (Hosmer et al., 2013).

Ансамбльге негізделген машиналық оқыту үлгілері, атап айтқанда Random Forest, XGBoost және LightGBM, ШОБ несиелері және қайта қаржыландыру схемалары сияқты күрделірек өнімдер үшін пайдаланылды (Breiman, 2001; Chen & Guestrin, 2016; Ke et al., 2017). Әртүрлі қарыз алушы профильдерінде жиі кездесетін кіріс мүмкіндіктері арасындағы жоғары дәрежелі өзара әрекеттесулер мен сызықтық емес қатынастарды осы үлгілер түсіруі мүмкін.

P2P несиелеу үлгілерінде терең нейрондық желілер және ұзақ қысқа мерзімді жады (LSTM) желілері сияқты нейрондық архитектуралар пайдаланылды (Hochreiter & Schmidhuber, 1997; Goodfellow et al., 2016). Баламалы және құрылымдалмаған деректерді тиімді пайдалану және мінез-құлық және транзакциялық мүмкіндіктерді біріктіру үшін бұл модельдер орталықтандырылмаған несие платформаларында қарыз алушы тәуекелін анықтау үшін маңызды болды.

Үлгінің өнімділігін арттыру және шамадан тыс орнатуды азайту үшін мұқият гиперпараметрлерді баптау және кросс-тексеру тәсілі қолданылды (Bergstra & Bengio, 2012). Random Forest және XGBoost үшін гиперпараметрлерді оңтайландыру үшін Grid Search және Random Search әдістері қолданылды. Байес оңтайландыруы оқу уақытын қысқарту және терең оқыту және градиентті арттыру үлгілері үшін дәлдікті арттыру үшін пайдаланылды (Snoek et al., 2012).

Модель өнімділігінің сенімді және объективті бағасын қамтамасыз ету үшін, әсіресе теңгерілмеген деректерде, k еселенген және стратификацияланған k-fold валидация әдістері қолданылды (Kohavi, 1995).

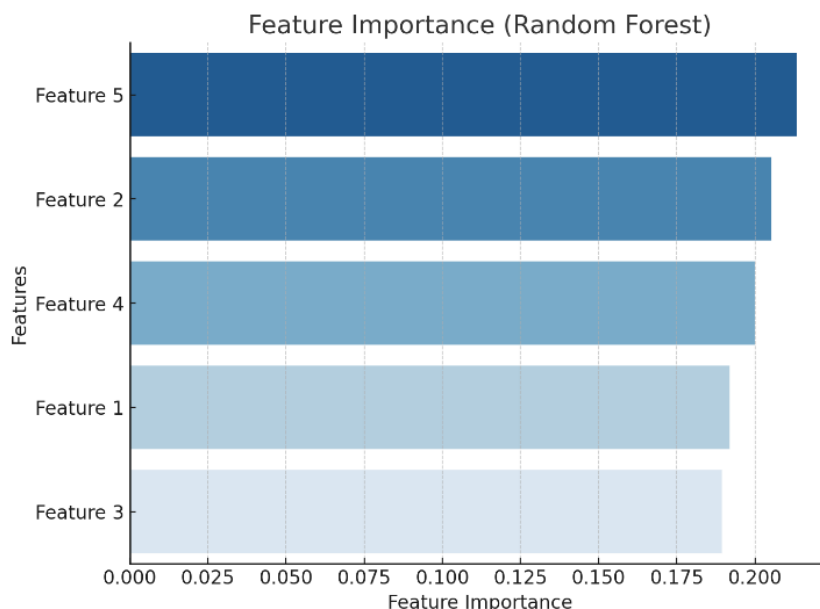
Шешім қабылдау мен ашықтықты жеңілдету үшін бірнеше түсіндірілетін AI (XAI) әдістері қолданылды. Әрбір мүмкіндіктің болжау нәтижесіне үлесін көрсету үшін SHAP (SHapley Additive Explanations) мәндері пайдаланылды (Lundberg & Lee, 2017). Ағашқа негізделген модельдер үшін дәстүрлі Gini importance және feature gain сияқты метрикалар да қолданылды. Сонымен қатар, ішінара тәуелділік графиктері (PDPs) айнымалылар өзгерісі әдепкі ықтималдыққа қалай әсер ететінін визуализациялау үшін қолданылды (Molnar, 2020).

Модельдің рейтингтік тиімділігін бағалау үшін Джини коэффициенті, ал классификация қуатын бағалау үшін ROC-AUC пайдаланылды (Hand & Till, 2001). Класстардың бөлінуі теңгерімсіз болған жағдайда, модельдің азшылық класына сезімталдығын тексеру үшін Precision-Recall қисықтары қарастырылды. Модельдің калибрленуін бағалау үшін калибрлеу сызбалары қолданылды (Niculescu-Mizil & Caruana, 2005).

Түсіндіру, болжамды өнімділік және іскерлік сәйкестік арасындағы тепе-теңдік соңғы модельді таңдауда басты рөл атқарды. Өзінің сенімді өнімділігі мен оңай түсіндірілуіне



байланысты логистикалық регрессия тұтынушылық және ипотекалық портфельдер үшін таңдаулы нұсқаға айналды. ШОБ және қайта қаржыландыру сегменттерінде градиентті күшейту алгоритмдері (XGBoost және LightGBM) жоғары болжам дәлдігін көрсетті. Терең нейрондық желілер, әсіресе мінез-құлық және балама деректер қосылғанда, P2P несиелеу үшін ең жақсы нәтиже көрсетті. Бұл оларды цифрлық негіздегі несиелеу платформалары үшін тиімді шешім етеді.



3-сурет. Кездейсоқ орман үлгісіндегі ерекшеліктердің маңызы

Ескерту: авторлар ашық дереккөзден алынған мәліметтер негізінде құрастырған

Графикте қарыз алушының дефолт ықтималдығын бағалау үшін қолданылатын кездейсоқ орман моделіндегі ерекшеліктердің өзектілігі көрсетілген. Модельге енгізілген факторлар Y осінде көрсетілген; олардың тәуекелді болжауға қосқан үлесі X осінде көрсетілген (мән неғұрлым жоғары болса, соғұрлым бұл сипаттаманың әсері көп болады). Графикте модельді таңдауға қандай факторлар көбірек әсер ететіні көрсетілген. Скорингтік картаны әзірлеу ең маңызды элементтерді ескеруі керек, өйткені олар қарыз алушылардың несиелік қабілетін бағалауға үлкен көмек көрсетеді. Маңыздылығы азырақ элементтер олардың құндылығы бойынша жойылуы немесе зерттелуі мүмкін.

Бұл зерттеу модельдің жұмысын нақтылайды және төлем қабілеттілігін болжау үшін қарыз алушылардың ең маңызды белгілері туралы гипотезалар жасайды (Baensens et al., 2016).

Үлгінің сапасын бағалау әдістері

Әзірленген баллдық үлгілердің өнімділігі мен сапасын бағалау үшін несиелік тәуекелді бағалауда жиі қолданылатын өлшемдер жинағы пайдаланылды (Thomas et al., 2017). Бұл көрсеткіштер үлгілердің тұрақтылығын, дискриминациялық күшін және әртүрлі несиелік өнім санаттарында пайдалануға жарамдылығын жан-жақты бағалауға мүмкіндік береді.

Модельдің жоғары тәуекелді және төмен тәуекелді қарыз алушыларды ажырату қабілеті Қабылдаушының жұмыс сипаттамасы қисығы астындағы аумақ (ROC-AUC) арқылы өлшенді. Жоғары ROC-AUC мәні үлгінің жақсы классификациялық өнімділігін білдіреді (Fawcett, 2006).



Кемсітушілік қуатты бағалаудың тағы бір маңызды құралы – Джини коэффициенті, ол ROC қисығына негізделген және модельдің қарыз алушыларды әдепкі ықтималдықтарына қарай қаншалықты дәл реттей алатынын көрсетеді (Hand & Henley, 1997; Crook et al., 2007). Джини жиі несиелік скорингте қолданылады, өйткені оның интерпретациясы түсінікті және тиімді.

Колмогоров-Смирнов (КС) статистикасы болжамды ұпайлардың жинақталған үлестіріміндегі дефолт жасағандар мен жасамағандар арасындағы максималды айырмашылықты көрсетеді. Бұл метрика екі сыныпты қаншалықты жақсы бөлетінін бағалауға мүмкіндік береді (Lem, 2019).

Халық тұрақтылығы индексі (PSI) модель болжамдарының уақыт бойынша өзгеруін бақылауға көмектеседі. PSI деректер дрейфінің бар-жоғын көрсету арқылы үлгіні қайта оқыту қажеттігін көрсетеді (Bailey, 2020).

Бизнес сценарийлерінің ауқымында үлгілердің дәлдігін, сенімділігін және жалпыламалық қабілетін тексеру үшін бұл бағалау көрсеткіштері әртүрлі несиелік өнімдерге жүйелі түрде қолданылды (Malekipirbazari & Aksakalli, 2015).

4-кесте. Модельді бағалау әдістері

Үлгі	РОК-АУК	Джини коэффициенті	К5 Статистикалық	PSI
Логистикалық регрессия	0,76	0,52	0,32	0,08
Кездейсоқ орман	0,82	0,64	0,41	0,12
XGBoost	0,88	0,76	0,52	0,15
<i>Ескерту: авторлар ашық дереккөзден алынған мәліметтер негізінде құрастырған</i>				

Бұл кестеде қарыз алушылардың дефолтын болжау үшін қолданылатын бірнеше бағалау жүйелерін бағалау нәтижелері келтірілген. Көрсеткіштер үлгілердің тиімділігін бағалауға және несиелік қабілеттілікті бағалаудың оңтайлы алгоритмін таңдауға мүмкіндік береді.

Модельдің дефолтқа ұшыраған және төленбеген қарыз алушыларды ажырату қабілеті бірнеше белгіленген жіктеу көрсеткіштері арқылы бағаланды. Қолданылатын негізгі шаралардың бірі Қабылдаушының жұмыс сипаттамасы – қисық астындағы аймақ (ROC-AUC) болды, ол модельдің тәуекелі жоғары және төмен тәуекелді қарыз алушыларды саралау мүмкіндігін сандық түрде көрсетеді. Жоғары ROC-AUC мәндері күшті дискриминациялық қуатты көрсетеді.

Тағы бір маңызды көрсеткіш Джини коэффициенті болды, ол несиелік скорингте жиі қолданылатын индикатор болып табылады, ол модельдің несиелік тәуекел бойынша қарыз алушыларды бағалау мүмкіндігін өлшейді. Ол стандартты формула бойынша есептелді:

$$Gini = 2 \times \left(ROC - AUC - \frac{1}{2} \right) \quad (1)$$

Колмогоров-Смирнов (КС) статистикасы дефолтқа ұшыраған және төленбеген қарыз алушылар арасындағы болжамды ұпай үлестіріміндегі алшақтықты бағалау үшін пайдаланылды. Жоғары KS мәндері осы екі топты бөлудің үлкен мүмкіндігін көрсетеді.

Модельдердің уақытша беріктігін бағалау үшін Популяция тұрақтылығы индексі (PSI) есептелді. Бұл көрсеткіш уақыт ішінде қарыз алушы сипаттамаларын бөлуде елеулі өзгерістер орын алғанын анықтайды. 0,20-дан жоғары PSI мәні әдетте үлгіні қайта калибрлеу қажеттілігін көрсетеді. Ағымдағы талдауда барлық үлгілер осы шекті мәннен төмен PSI мәндерін көрсетті, бұл олардың бағалау кезеңінде тұрақты болғанын көрсетеді.



Тексерілген алгоритмдердің ішінде XGBoost ROC-AUC = 0,88, Джини = 0,76 және KS = 0,52 арқылы ең жоғары жалпы өнімділікті қамтамасыз етті. Random Forest логистикалық регрессияға қарағанда жақсырақ, бірақ XGBoost-тан сәл төмен орындайтын ROC-AUC 0,82 орташа өнімділікті көрсетті. Логистикалық регрессия ең төменгі өнімділік көрсеткіштерін тіркеді (ROC-AUC = 0,76, Джини = 0,52), бірақ ол өзінің түсіндірмелілігі мен реттеу мөлдірлігінің арқасында практикада кеңінен қолданылады. XGBoost деректердің өзгеруіне (PSI = 0,15) жоғары сезімталдықты көрсеткенімен, үлгілердің ешқайсысы уақыт өте келе салыстырмалы беріктігін растайтын 0,20 шегінен аспады.

Модельдеу, алдын ала өңдеу және бағалау процедуралары бірқатар бағдарламалық құралдар мен бағдарламалау кітапханаларын қолдану арқылы жүргізілді. Python деректерді өңдеуге арналған пандалар, негізгі үлгілер үшін scikit-learn және кеңейтілген градиентті күшейту алгоритмдері үшін XGBoost және LightGBM сияқты маңызды пакеттері бар негізгі бағдарламалау тілі ретінде қызмет етті. SHAP кітапханасы (SHapley Additive Explanations) үлгі шығыстарын түсіндіру және мүмкіндіктің маңыздылығын бағалау үшін пайдаланылды. SQL реляциялық қарыз алушының дерекқорларына қол жеткізуді басқару үшін пайдаланылды және Jupyter Notebooks интерактивті талдау мен үлгіні әзірлеуге көмектесті. Осы құралдардың үйлесімі әрбір несиелік өнімнің ерекше сипаттамаларына бейімделген сенімді аналитикалық құбырды жүзеге асыруға мүмкіндік берді.

Зерттеуде бірнеше машиналық оқыту әдістері пайдаланылды: логистикалық регрессия, Random Forest, XGBoost және LSTM. Әрбір алгоритмнің математикалық негіздері, жұмыс принциптері, негізгі гиперпараметрлері және қолдану салалары төменде сипатталған.

1. Логистикалық регрессия (Logistic Regression)

Математикалық модель:

$$P(y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \dots + \beta_n X_n)}} \quad (2)$$

Принцип: ықтималдықтық функция негізінде қарыз алушының дефолт ықтималдығын бағалайды. Сызықтық тәуелділіктерге және түсіндіру мүмкіндігіне тиімді.

Гиперпараметрлер: penalty = l2; solver = liblinear; C = 1.0.

Қолдану саласы: тұтынушылық несиелер мен ипотека, себебі реттеу талаптарына сәйкес түсіндірілуі қажет.

2. Random Forest

Математикалық модель (жалпылама):

$$\hat{y} = \frac{1}{T} \sum_{t=1}^T h_t(x), \quad h_t(x) \in \{0,1\} \quad (3)$$

мұндағы T – ағаштар саны, $h_t(x)$ – әрбір ағаштың шешімі.

Принцип: шешім ағаштарының ансамблі (bagging), кездейсоқ ерекшеліктерді таңдау арқылы модельдің тұрақтылығын арттырады және сызықтық емес тәуелділіктерді жақсы қамтиды.

Гиперпараметрлер: n_estimators = 200; max_depth = 10; min_samples_split = 5; min_samples_leaf = 2.

Қолдану саласы: шағын және орта бизнес (ШОБ) және қайта қаржыландыру сегменттері.

3. XGBoost (Extreme Gradient Boosting)

Математикалық модель:

$$\widehat{y}_i^{(t)} = \sum_{k=1}^t f_k(x_i), \quad f_k \in F \quad (4)$$

мұндағы f_k – шешім ағаштары, F – барлық мүмкін ағаштар кеңістігі.



Принцип: градиентті бустинг әдісі, әрбір жаңа ағаш алдыңғы ағаштардың қатесін түзетеді. Жоғары дәлдік пен тұрақты оптимизацияны қамтамасыз етеді.

Гиперпараметрлер: $n_estimators = 300$; $learning_rate = 0.1$; $max_depth = 6$; $subsample = 0.8$; $colsample_bytree = 0.8$.

Қолдану саласы: қайта қаржыландыру, ШОБ және P2P сегменттері – күрделі деректер құрылымдары үшін.

4. LSTM (Long Short-Term Memory Neural Network)

Математикалық модель:

$$h_t = o_t \odot \tanh(c_t), \quad c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \quad (5)$$

мұндағы i_t, f_t, o_t — кіріс, ұмытушылық және шығыс қақпалары.

Принцип: уақыттық және реттелген деректерді тиімді өңдейтін рекуррентті нейрондық желі. Цифрлық іздер мен транзакциялық белсенділікті болжауда қолданылады.

Гиперпараметрлер: екі жасырын қабат (128 және 64 нейрон), $activation = ReLU$, $dropout = 0.3$, $optimizer = Adam$, $epochs = 50$, $batch_size = 128$.

Қолдану саласы: P2P және финтех сегменттері, құрылымдалмаған деректер мен пайдаланушы мінез-құлқын талдау үшін.

5-кесте. Алгоритмдердің ерекшеліктері мен қолдану салалары

Алгоритм	Ерекшеліктері	Қолдану саласы
Логистикалық регрессия	Қарапайым, түсіндірілетін, нормативтік талаптарға сай	Тұтынушылық несиелер, ипотека
Random Forest	Ансамбльдік әдіс, деректердегі шуды азайтады	ШОБ, қайта қаржыландыру
XGBoost	Жоғары дәлдік, тұрақты оптимизация	ШОБ, қайта қаржыландыру, P2P
LSTM	Уақыттық тәуелділікті ескереді, цифрлық іздерді талдайды	P2P және финтех өнімдері
<i>Ескерту: авторлар ашық дереккөзден алынған мәліметтер негізінде құрастырған</i>		

Кестеде логистикалық регрессия, Random Forest, XGBoost және LSTM сияқты негізгі машиналық оқыту алгоритмдерінің ерекшеліктері мен олардың екінші деңгейлі банктердің скорингтік модельдерінде қолдану салалары жүйеленген. Бұл салыстыру әрбір модельдің артықшылықтары мен шектеулерін анық көрсетеді.

Программалық қамтамасыз ету және қолданылған құралдар

Зерттеуде скорингтік модельдерді құру, оқыту және бағалау үшін заманауи бағдарламалық қамтамасыз ету мен ашық бастапқы кодты кітапханалар пайдаланылды.

Негізгі құралдар:

Python 3.11 – негізгі бағдарламалау тілі ретінде;

Scikit-learn – базалық алгоритмдер (Logistic Regression, Random Forest), метрикалар (ROC-AUC, Gini, KS) және деректерді алдын ала өңдеу үшін;

XGBoost және LightGBM – градиентті бустинг негізінде жоғары өнімді модельдер құру үшін;

SHAP (Shapley Additive Explanations) – модельдердің интерпретациясы мен Explainable AI талдауын жүзеге асыру үшін;

Pandas және NumPy – деректерді талдау және матрицалық есептеулер үшін;

TensorFlow/Keras – нейрондық желілерді (LSTM) жобалау және оқыту үшін;

SQL – реляциялық дерекқорлардан деректерді алу және алдын ала өңдеу үшін;



Jupyter Notebooks – эксперименттерді жүргізу және нәтижелерді құжаттау үшін;
Apache Spark – үлкен көлемді деректерді параллельді өңдеуді және модельдерді өндірістік масштабта енгізуді қолдау үшін.

Бұл құралдар кешені скорингтік модельдердің барлық кезеңдерін — деректерді жинаудан бастап модельдерді түсіндіруге дейін — толық қамтуға мүмкіндік берді.

Модельдерді баптау және интерпретация

Бұл зерттеу тек дайын кітапханаларды қолданумен шектелмей, бірқатар ғылыми-тәжірибелік қадамдарды қамтыды. Әрбір модельдің өнімділігін арттыру үшін гиперпараметрлерді мұқият баптау жүргізілді (Random Forest үшін ағаштар саны, тереңдігі, бөліну критерийлері; XGBoost үшін оқу жылдамдығы, ағаш тереңдігі, subsample және colsample параметрлері; LSTM үшін қабат саны, нейрондар, epochs және batch size).

Сонымен қатар, әртүрлі алгоритмдердің нәтижелері салыстырылып, олардың артықшылықтары мен шектеулері айқындалды. Бұл салыстырмалы талдау скорингтік модельдерді нақты сегменттерде қолдану тиімділігін бағалауға мүмкіндік берді.

Зерттеудің маңызды ерекшелігі – Explainable AI әдістерін (SHAP – Shapley Additive Explanations, PDP – Partial Dependence Plots) пайдалану арқылы модельдердің интерпретациясын қамтамасыз ету. Бұл тәсілдер модельдің ішкі логикасын ашуға, айнымалылардың әсерін бағалауға және регуляторлық талаптарға сәйкес түсіндірілуін қамтамасыз етуге мүмкіндік берді.

НӘТИЖЕЛЕР ЖӘНЕ ОЛАРДЫ ТАЛҚЫЛАУ

1. Модельдердің тиімділігін салыстыру

Бірнеше скорингтік модельдердің тиімділігін бағалау үшін ROC-AUC, Gini коэффициенті, Колмогоров-Смирнов статистикасы (KS) және халық тұрақтылығының индексі (PSI) қоса алғанда, негізгі статистика есептелді.

3-кестеде XGBoost, кездейсоқ орман және логистикалық регрессия нәтижелерін салыстыру көрсетілген.

- XGBoost барлық маңызды критерийлер бойынша үздік болып шықты. Тиімділік коэффициенті (ROC-AUC) = 0,88, бұл оның несиелік тәуекел негізінде қарыз алушыларды жіктеудің жоғары мүмкіндігіне кепілдік береді. Оның жоғары тану қабілетін 0,76 Джини коэффициенті дәлелдейді; 0,52 KS статистикасы дефолтқа жол берген қарыз алушылар мен дефолтқа жол бермеген қарыз алушылардың жақсы бөлінуін көрсетеді.

- Кездейсоқ орман орташа мәндерді көрсетті. ROC-AUC = 0,82 қатынасы XGBoost-қа қарағанда төмен болғанымен, логистикалық регрессияға қарағанда жоғары. Бұл интерпретация мен дәлдікті біріктіреді.

- Логистикалық регрессия ең әлсіз нәтижелерге қол жеткізгенімен (ROC-AUC = 0,76, Джини коэффициенті = 0,52), оның жоғары интерпретациялануы мен билік үшін ашықтығы оның неге әлі де жиі кездесетінін түсіндіруге көмектеседі.

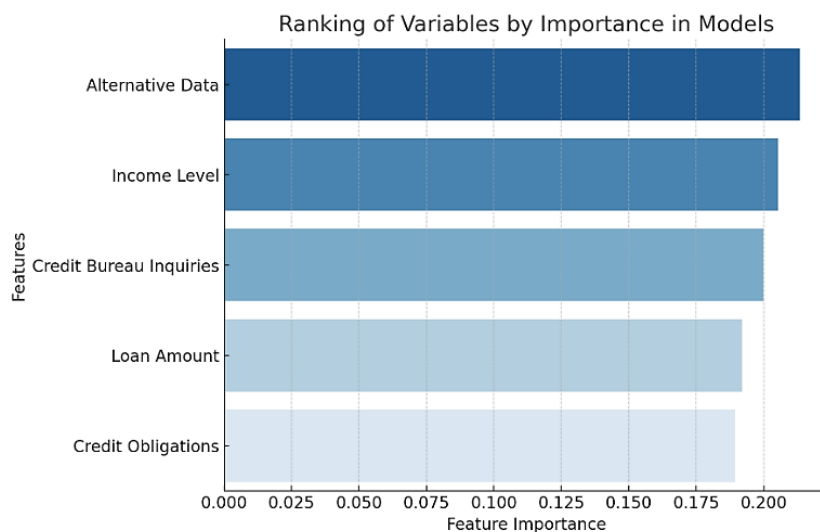
PSI зерттеулеріне сәйкес, барлық модельдер тұрақты болғанымен (PSI = 0,2), XGBoost (PSI = 0,15) қарыз алушының сипаттамаларының өзгеруіне сезімтал және жиі калибрлеуді қажет етеді.

2. Баллдық жүйелердегі ерекшеліктердің маңыздылығын тексеру

Пайдалану кездейсоқ орман үшін кіріктірілген маңыздылықты бағалау әдістері және XGBoost сондай-ақ ПШПН (Шапли Аддитивті түсініктемелер) ерекшеліктердің маңыздылығын талдау, модельдерді түсінуге болады.

Y осі атрибуттарды атайды: несие сомасы, кіріс деңгейі, несиелік міндеттемелер, несиелік бюронның сұраныстары, балама мәліметтер.

X осі: Мүмкіндіктердің маңыздылығы; айнымалы мән әдепкі болжамға олардың мәні неғұрлым жоғары болса, соғұрлым көбірек әсер етеді.



4-сурет. Модельдердің айнымалы дәрежелері олардың маңыздылығына қарай

Ескерту: авторлар ашық дереккөзден алынған мәліметтер негізінде құрастырған

Алғашқы талдаудың нәтижелері

Барлық модельдер үшін маңызды элементтер:

- Дефолттың негізгі көрсеткіштерінің бірі несие сомасы болып табылады. Несие сомасының жоғарылауы қарыз алушы үшін ықтимал қарыз жүктемесінің жоғарылауына әкеледі, сондықтан қайтарылмау тәуекелін арттырады.

- Қарыз алушыдағы төлем қабілеттілігінің негізгі белгілерінің бірі олардың табыс деңгейі болып табылады. Табыстың жоғарылауы дефолт тәуекелін төмендетсе де, оның әсерін қарыздың барлық жүктемесіне қатысты байқау керек.

- Несиелік міндеттемелер қарыз алушының қарыздық жүктемесін көрсетеді. Белсенді несиелердің көптігі қайтарылмау тәуекелін арттырады.

- Несиелік бюроға сұрау салудың жиілігі қарыз алушының қаржылық проблемалары мен тәуекелдің жоғары дәрежесін көрсетуі мүмкін.

Ауыстырылатын деректердің әсері

- P2P несиелеу кезінде дефолт болжамына мінез-құлық және транзакциялық деректер (төлемдердің жиілігі, шығыстардың құрылымы сияқты) қатты әсер етеді.

- Қайта қаржыландыруда маңызды мәселе несие қайта қаржыландырылған банк болып табылады, өйткені бастапқы банктегі несие шарттары қарыз алушының қаржылық жағдайын аша алады.

- Ипотекалық және автокредиттерде қарыз алушының қызмет ету мерзімі мен қамтамасыз ету құны активтің өтімділігіне және кірісті бағалаудың тұрақтылығына айтарлықтай әсер етеді.

Кездейсоқ орман ерекшеліктерінің өзектілігін түсіндіру

- Кездейсоқ орман моделінің ерекшеліктерінің маңыздылығын талдау әдепкі болжамға әсер етуі ықтимал айнымалыларды анықтады. Кесте X әрбір фактордың модель таңдауына сәйкес әсерін тізімдейді.

Қарапайым бақылаулар:

- Атап айтқанда, P2P несиелеу және финтех индустриясында баламалы деректер - мінез-құлық және транзакциялық көрсеткіштер – құрылымдалмаған деректерді біріктіру скоринг үлгілерінің дәлдігін арттыратынын дәлелдеуде үлкен өзектілігін көрсетті.



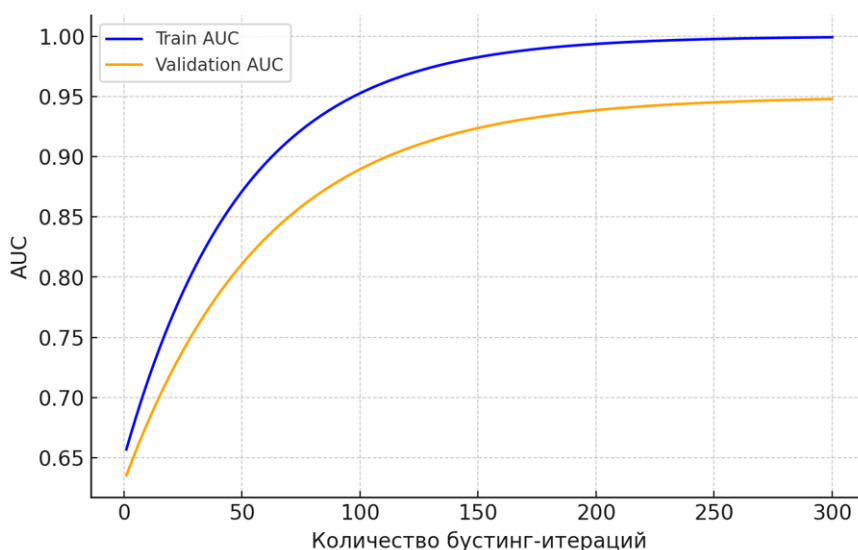
• Несиелік қабілеттіліктің негізгі детерминанттарының бірі бола тұра, қарыз алушының кірісі дефолт болжамына, әсіресе кәдімгі банктік несиелендіруге үлкен әсер етеді.

• Несиелік бюроға жасалған сұраулар қарыз алушының қаржылық жағдайы туралы анықтама береді: олардың көпшілігі тәуекелдің жоғары дәрежесін және ақшаның шұғыл қажеттілігін көрсетуі мүмкін.

• Несиелік міндеттемелер қарыз жүктемесінің қарыз алушының төлем қабілеттілігіне әсерін көрсетеді.

Модельдерді оқыту динамикасы

Зерттеу барысында модельдердің тек соңғы метрикалары ғана емес, сонымен бірге олардың оқыту процесі де талданды. Әсіресе XGBoost моделінде оқу қисықтары (learning curves) бақылауға алынды.



5-сурет. XGBoost training dynamics

Ескерту: авторлар ашық дереккөзден алынған мәліметтер негізінде құрастырған

Графикте training және validation жиынтықтары бойынша AUC мәндерінің өзгеруі көрсетілген. Алғашқы итерациялар кезінде модельдің дәлдігі біртіндеп өсіп, шамамен 150-бустинг ағашынан кейін тұрақтанды. Training және validation қисықтарының жақын орналасуы модельдің шамадан тыс үйреніп кетпегенін (overfitting жоқтығын) көрсетеді.

Random Forest және логистикалық регрессия жағдайында оқу процесі тұрақты болып, метрикалар бірден жоғары деңгейде сақталды. Ал LSTM моделінде loss және ассигасу бойынша айқын динамика байқалды: алғашқы 10 эпохада қателік жылдам төмендеді, ал 30-40 эпохадан кейін convergence деңгейіне жетті.

Модель интерпретациясы (SHAP importance)

Explainable AI әдістері, соның ішінде SHAP values, негізгі айнымалылардың модель шешіміне әсерін түсіндіруге мүмкіндік берді. SHAP-графиктерінде ең ықпалды факторлар ретінде табыстың деңгейі, DTI, PTI, несие сомасы және кредиттік бюроға сұраулар саны анықталды. Бұл факторлар дәстүрлі скоринг логикасына сәйкес келіп, машиналық оқыту алгоритмдерінің бизнес-заңдылықтармен үндес екенін көрсетті.

ҚОРЫТЫНДЫ

Бұл зерттеу кредиттік скоринг жүйелерінде машиналық оқыту әдістерін қолданудың тиімділігін жан-жақты көрсетті. Алынған нәтижелер дәстүрлі тәсілдердің (логистикалық



регрессия) қарапайымдылығы мен интерпретация мүмкіндігін сақтағанымен, олардың болжамдық қабілеті күрделі деректерді талдауда шектеулі екенін айқындады.

Ансамбльдік модельдер, атап айтқанда Random Forest және XGBoost, құрылымдалған және құрылымдалмаған деректерді өңдеуде жоғары өнімділік көрсетті. Random Forest тұрақтылық тұрғысынан тиімді болып шықса, XGBoost ең жоғары дәлдікке жетті (AUC = 0,88). Бұл көрсеткіш дәстүрлі логистикалық регрессиядан едәуір жоғары, алайда XGBoost деректердің уақыт бойынша өзгеруіне (population drift) сезімтал екенін атап өткен жөн.

LSTM нейрондық желілері P2P және финтех сегменттерінде цифрлық іздер мен уақыттық транзакциялық деректерді талдау үшін перспективалы екені анықталды. Бұл модельдер дәстүрлі құрылымдалған деректерді толықтырып, альтернативті деректердің (телеком, әлеуметтік желілер, онлайн-платформалар) кредиттік қабілетті бағалаудағы маңызын дәлелдейді.

Explainable AI құралдарын (SHAP, PDP) енгізу нәтижелері алгоритмдердің шешім қабылдау логикасын түсіндіруге мүмкіндік берді. Бұл екінші деңгейлі банктер үшін регуляторлық талаптарға сәйкес интерпретацияны қамтамасыз етудің маңызды тетігі болып табылады. Сонымен қатар, Population Stability Index (PSI) модельдердің уақыт бойынша тұрақтылығын бағалауға мүмкіндік беріп, олардың өндірістік жағдайда ұзақмерзімді қолдануға жарамдылығын көрсетті.

Зерттеудің ғылыми жаңалығы – құрылымдалған, құрылымдалмаған және баламалы деректерді кешенді қолдану арқылы әртүрлі машиналық оқыту модельдерінің сапасын салыстырмалы талдау, Explainable AI әдістерін енгізу және модельдердің уақыт бойынша тұрақтылығын PSI арқылы бағалау. Практикалық тұрғыдан алғанда, алынған нәтижелер екінші деңгейлі банктерге және финтех-компанияларға кредиттік скоринг жүйелерін жетілдіруде бағдар бола алады.

Қорытындылай келе, кредиттік скоринг жүйелерін жетілдіру үшін гибридік тәсіл ұсынылады:

- логистикалық регрессия – түсіндірілуі маңызды және реттеуші талаптарға сай өнімдер үшін (ипотека, тұтынушылық несиелер);
- Random Forest – тұрақтылық қажет жағдайларда;
- XGBoost – күрделі сегменттерде жоғары дәлдікті талап ететін процестер үшін;
- LSTM – цифрлық іздер мен транзакциялық деректерді талдауға бағытталған P2P және финтех өнімдері үшін.

Осылайша, зерттеу нәтижелері Қазақстандағы кредиттік скорингтің даму бағытын айқындап қана қоймай, заманауи машиналық оқыту әдістерін енгізудің ғылыми-тәжірибелік негізін қалыптастырады.

МҮДДЕЛЕР ҚАЙШЫЛЫҒЫ: Авторлар мүдделер қайшылығы жоқ екенін мәлімдейді.

ӘДЕБИЕТТЕР ТІЗІМІ

- Қоңыр I., Mues C. (2012). Теңгерімсіз несиелік скоринг деректерінің жіктеу алгоритмдерін эксперименттік салыстыру. Сараптамалық жүйелер қосымшалармен, Т. 39, № 3, 3446–3453. <https://doi.org/10.1016/j.eswa.2011.09.033> // Qonyr I., Mues C. Tengerimsiz nesielik scoring derekteriniñ zhikteý algoritmderin eksperimenttik salystyruý. Sarapnamalyq zhүйeler qosymshalarmen, T. 39, No. 3, 3446–3453. [An experimental comparison of classification algorithms for imbalanced credit scoring data. Expert Systems with Applications.] In Russ.
- Аррам А., Аюб М., Альбадр М.А.А., Сүлеймен А., Албашиш Д. (2023). Машиналық оқыту үлгілерін пайдалана отырып, несиелік рейтингті болжау: жаңа деректер жинағы.



- арХив баспаға дейінгі, арХив:2310.02956. <https://arxiv.org/abs/2310.02956> // Arram A., Ayub M., Albadar M.A.A., Süleymen A., Albashish D. Mashinalyq oqytý ulgilerin paidalana otyryp, nesielik reitingti bolzhaý: zhana derekter zhınaýy. arXiv preprint, arXiv:2310.02956. [Credit rating prediction using machine learning models: a new dataset.] In Russ.
- Дәулетханұлы Ү., Ойшынова Б. (2023). Монетарлық реттеушінің машиналық оқытуды және жасанды интеллектті қолдануы. Қазақстан Республикасы Ұлттық Банкінің экономикалық шолуы, № 4. <https://nationalbank.kz/file/download/98995> // Dauletkhanuly Y., Oyshynova B. Monetarlyq retteushinin mashinalyq oqytudy zhane zhasandy intellekti qoldanu. Qazaqstan Respublikasy Ulttyq Bankining ekonomikalıq sholuy, No. 4. [Application of machine learning and AI by the monetary regulator.] In Russ.
- Қадиев Ә., Чибисова Ә. (2022). Кредиттік скорингтің нейрондық желілік әдістері. Математикалық модельдеу және сандық әдістер, Т. 4, 81–92. https://mmcm.bmstu.ru/articles/download_article/295/ // Kadiev A., Chibisova A. Kredittik scoringting neirondyq zhelilik adisteri. Matematikalyq modeldeu zhane sandyq adister, T. 4, 81–92. [Neural network methods in credit scoring. Mathematical Modelling and Numerical Methods.] In Russ.
- Оғыз К., Уғур О., Кестел Ә.С. (2023). Кредиттік скорингті анықтауда мүмкіндіктерді таңдау мен түрлендірудің машиналық оқыту әдістеріне әсері. арХив баспаға дейінгі, арХив:2303.05427. <https://arxiv.org/abs/2303.05427> // Oguz K., Ugur O., Kestel A.S. Kredittik scoringti anyqtawda mumkindikterdi tandaw men turlendirudin mashinalyq oqytý adisterine aseri. arXiv preprint, arXiv:2303.05427. [Impact of feature selection and transformation methods on credit scoring.] In Russ.
- Шухардин Ә. (2024). Машиналық оқыту модельдеріне негізделген несиелік скоринг жүйесін дамыту. Орал Федералды университетінің электронды мұрағаты. https://elar.urfu.ru/bitstream/10995/140553/1/m_th_a.v.shukhardin_2024.pdf // Shukhardin A. Mashinalyq oqytý modelderine negizdelgen nesielik scoring zhuyesin damytu. Oral Federaldy universitetining elektronny muragaty. [Development of a credit scoring system based on machine learning models.] In Russ.
- Гриер Д.А. (2012). Кредиттік скоринг және кредиттік бақылау. Оксфорд: Оксфорд университетінің баспасы // Grier D.A. Kredittik scoring zhane kredittik baqylaw. Oxford: Oxford universitetining baspasy. [Credit scoring and credit control. Oxford University Press.] In Russ.
- Biecek P., Khlebus M., Gaida J., Gosevska A., Kozaq A., Ogonovskii D., Sztachelski J., Wojewnik P. (2023). Кредиттік скоринг үшін машиналық оқыту алгоритмдерін қосу: күрделі болжамды модельдерді нақты түсінуге арналған түсіндірілетін жасанды интеллект (ХАИ) әдістері. арХив баспаға дейінгі, арХив:2310.02956. <https://arxiv.org/abs/2310.02956> // Biecek P., Khlebus M., Gaida J., Gosevska A., Kozaq A., Ogonovskii D., Sztachelski J., Wojewnik P. Kredittik scoring ushun mashinalyq oqytý algoritmderin qosu: kurdeli bolzhamdy modelderdi tusinyge arналган tusindiriletin jasandy intellekt (ХАИ) adisteri. arXiv preprint, arXiv:2310.02956. [Explainable AI for complex predictive models in credit scoring.] In Russ.
- Хенли В.Е., Ханд Д.Ж. (1996). А к-ең жақын-көрші тұтынушылық несиелік тәуекелді бағалаудың жіктеуіші. Статист, Т. 45, № 1, 77–95. <https://doi.org/10.2307/2988486> // Henley V.E., Hand D.J. A k-en jakyn-korshi tutynushylyk nesielik tauyekeldi bagalau zhyiktewishi. Statist, T. 45, No. 1, 77–95. [A k-nearest neighbor classifier for consumer credit risk. The Statistician.] In Russ.
- Центральный банк Российской Федерации. (2022). Және әсері бастап теріс жаңалықтар күні қоғамдық қабылдаудың бұзылуы бастап инфляция. Рабочие материалы Банка



- России, № 111. https://www.cbr.ru/content/document/file/144918/wp_111.pdf // Tsentral'nyi bank Rossiiskoi Federatsii. Zhane aseri bastap teris zhanalyqtar kuni qogamdyq qabyldawdyn buzylyuy bastap inflyatsiya. Raboche materialy Banka Rossii, No. 111. [The effect of negative news and public perception shifts on inflation. Working Papers of the Bank of Russia.] In Russ.
- Мұнхдалай Л., Намсрай О.Б., Ли Ж.Ы., Рю Қ.Х. (2019). Банк клиентінің несиелік бағалауы бойынша машиналық оқыту әдістерін эмпирикалық салыстыру. Тұрақтылық, Т. 11, № 3, 699. <https://doi.org/10.3390/cy11030699> // Munkhdalai L., Namsrai O.B., Li Zh.Y., Ryu Q.H. Bank klientinin nesielik bagalauy boynsha mashinalyq oqytu adisterin emipirikalyk salystyry. Turaktylyk, T. 11, No. 3, 699. [Empirical comparison of machine learning methods for credit evaluation of bank clients. Sustainability.] In Russ.
- Расторгуев Ә., Микишанина Б. (2022). Қарыз алушының несиелік қабілеттілігін бағалаудың скорингтік моделін құру. Электронный научный журнал «Вектор экономики», № 5. https://www.vectoreconomy.ru/images/publications/2022/5/mathematicalmethods/Rastorguev_Mikishanina.pdf // Rastorguev A., Mikishanina B. Qaryz alushynyn nesielik qabiletiligin bagalaudyn scoringtik modelin quru. Elektronnyi nauchnyi zhurnal «Vektor ekonomiki», No. 5. [Building a scoring model for borrower creditworthiness assessment. Vector of Economics.] In Russ.
- Татаринцев М.Ә., Никитин П. В., Горохова Р. И., Долгов В. И. (2022). Кредиттік скоринг тапсырмалары бойынша машиналық оқыту технологияларын салыстырмалы талдау. Іргелі зерттеулер, Т. 2, 7-12. <https://doi.org/10.17513/фр.43419> // Tatarintsev M.A., Nikitin P.V., Gorokhova R.I., Dolgov V.I. Kredittik scoring tapsyrmalary boynsha mashinalyq oqyty tekhnologiyalaryn salystyrmaly taldaу. Irgeli zertteyler, T. 2, 7–12. [Comparative analysis of machine learning technologies in credit scoring tasks. Fundamental Research.] In Russ.
- XGBoost Құжаттама. (2025). Кездейсоқ ормандар™ ішінде XGBoost. <https://xgboost.readthedocs.io/en/stable/tutorials/rf.html> (қол жеткізу күні: 13.02.2025) // XGBoost Quzhattama. Kedzeisoq ormandar™ ishinde XGBoost. <https://xgboost.readthedocs.io/en/stable/tutorials/rf.html> (accessed: 13.02.2025). [XGBoost Documentation: Random Forests in XGBoost.] In Russ.
- Аддо П.М., Геган Д., Хасани В. (2018). Машиналық және тереңдетіп оқыту модельдерін қолдану арқылы несиелік тәуекелдерді талдау. Тәуекелдер, Т. 6, № 2, 38. <https://doi.org/10.3390/тауекелдер6020038> // Addo P.M., Guegan D., Khasani B. Mashinalyq zhane terengdetip oqytu modelderin qoldanu arqyly nesielik tauyekelderdi taldaу. Tauyekelder, T. 6, No. 2, 38. [Credit risk analysis using machine and deep learning models. Risks.] In Russ.
- Ху Л., Чен Д., Вонг Дж., Ян Х., Ван К., Суджианто А., Наир В.Н. (2020). Жетекшілік ететін машиналық оқыту әдістері: банктік қызметке қосымшалармен шолу. arXiv баспаға дейінгі, arXiv:2008.04059. <https://arxiv.org/abs/2008.04059> // Hu L., Chen D., Wong J., Yang H., Wang K., Sudjianto A., Nair V.N. Zhetekshilik etetin mashinalyq oqyty adisteri: banktik qyzmetke qosymshalarmen sholu. arXiv preprint, arXiv:2008.04059. [Supervised machine learning methods with applications to banking.] In Russ.
- Abedin, B., Meske, C., Junglas, I., & others. (2022). Designing and managing human-AI interactions. Information Systems Frontiers, 24, 691–697. <https://doi.org/10.1007/s10796-022-10313-1>
- Aggarwal, C. C. (2015). Outlier analysis (2nd ed.). Springer.
- Baesens, B., Roesch, D., & Scheule, H. (2021). Hybrid credit scoring models: Combining traditional and machine learning approaches. Springer.



- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). <https://doi.org/10.1145/2939672.2939785>
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.
- Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: A review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 160(3), 523–541. <https://doi.org/10.1111/j.1467-985X.1997.00078.x>
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.
- He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263–1284. <https://doi.org/10.1109/TKDE.2008.239>
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). Wiley.
- Kelleher, J. D., Mac Namee, B., & D'Arcy, A. (2020). *Fundamentals of machine learning for predictive data analytics* (2nd ed.). MIT Press.
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, 30. https://papers.nips.cc/paper_files/paper/2017/hash-6449f44a102fde848669bdd9eb6b76fa-Abstract.html
- Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)* (Vol. 2, Pp. 1137–1145). <https://www.ijcai.org/Proceedings/95-2/Papers/016.pdf>
- OECD. (2023). *AI in financial markets and credit scoring*. OECD Publishing. <https://www.oecd.org/finance/>
- Saar-Tsechansky, M., & Provost, F. (2007). Handling missing values when applying classification models. *Journal of Machine Learning Research*, 8, 1625–1657. <http://www.jmlr.org/papers/volume8/saar07a/saar07a.pdf>
- Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems*, 25. https://papers.nips.cc/paper_files/paper/2012/hash/05311655a15b75fab86956663e1819cd-Abstract.html
- Thomas, L. C., Crook, J. N., & Edelman, D. B. (2017). *Credit scoring and its applications* (2nd ed.). SIAM.
- Zhang, Z., Yang, L. T., Liu, Y., & Chen, J. (2019). A survey on deep learning for big data. *Information Fusion*, 42, 146–157. <https://doi.org/10.1016/j.inffus.2017.10.006>



Авторлар туралы мәліметтер
Информация об авторах
Information about authors



Ордабаева Жанна Мұратқызы – техника ғылымдарының магистрі, "Қ.И. Сәтбаев атындағы Қазақ ұлттық техникалық зерттеу университеті", Алматы, Қазақстан

Ордабаева Жанна Муратовна – магистр технических наук, НАО «Казахский Национальный Исследовательский технический университет имени К.И. Сатпаева», г. Алматы, Казахстан

Ordabayeva Zhanna Muratovna – Master of Technical Sciences, Kazakh National Research Technical University named after K.I. Satpayev, Almaty, Kazakhstan,

E-mail: zhannaordabayeva@gmail.com ,

ORCID: <https://orcid.org/0009-0009-8273-3147>



Молдағұлова Айма Николаевна – физика-математика ғылымдарының кандидаты, «Қ.И. Сәтбаев атындағы Қазақ ұлттық техникалық зерттеу университеті», Алматы қ., Қазақстан

Молдагулова Айма Николаевна – кандидат физико-математических наук, НАО «Казахский Национальный Исследовательский технический университет имени К.И. Сатпаева», г. Алматы, Казахстан

Aiman Nikolaevna Moldagulova – Candidate of Physico-Mathematical Sciences, Kazakh National Research Technical University named after K.I. Satpayev, Almaty, Kazakhstan

e-mail: aiman.moldagulova@gmail.com,

ORCID: <https://orcid.org/0000-0002-1596-561X>