

**Assel Murzalimova** – master of Ecology, senior teacher of the Department «Chemical Technology and ecology»; Shakarim University of Semey, Republic of Kazakhstan; e-mail: murzalimova78@mail.ru. ORCID: 0000-0002-8341-9899.

**Bakyt Imamova** – master of Hydrology, senior teacher of the Department «Chemical Technology and ecology»; Shakarim University of Semey, Republic of Kazakhstan; e-mail: bakimamova@mail.ru. ORCID: 0000-0002-1050-5331.

*Материал 22.12.2022 ж. баспаға түсмі.*

DOI: 10.53360/2788-7995-2023-1(9)-7

MPHTI: 20.23.25

**Ж.М. Ордабаева\*, А.Н. Молдагулова**

Казахский национальный исследовательский технический университет имени К.И. Сатпаева  
050013, Республика Казахстан, г. Алматы, ул. Сатпаева 22а  
\*e-mail: zhannaordabayeva@gmail.com

## **ПОСТРОЕНИЕ МОДЕЛИ КРЕДИТНОГО СКОРИНГА НА ОСНОВЕ ТИПА ЦЕЛЕВОЙ ПЕРЕМЕННОЙ**

**Аннотация:** С быстрым развитием больших данных и интернет-технологий компании, занимающиеся финансовыми платформами больших данных, собирают и систематизируют массивные данные через свои собственные платформы, улучшают параметры кредитного скоринга и используют методы машинного обучения для проведения комплексных и научных оценок кредитного скоринга. Таким образом, банки при построении кредитного скоринга сталкиваются с большими проблемами. Исходя из ограничений существующей системы и методов персонального кредитного рейтинга, необходимо изучить персональный кредитный рейтинг на основе методов машинного обучения, усовершенствовать параметры и систему балльной оценки персонального кредитного рейтинга, уточнить каналы сбора данных, использовать технологию динамической десенсибилизации. для снижения чувствительности данных, метод тестирования LOF для проверки данных о выбросах и метод случайного леса для заполнения отсутствующих значений данных. Затем вы используете метод дерева решений с повышением градиента, чтобы просмотреть важные индикаторы, обработать проверенные индикаторы с помощью модели системы показателей, основанной на логистической регрессии, и получить личный кредитный рейтинг. Наконец, модель тестируется с помощью нейронной сети BP, и модель используется для прогнозирования уровня личного кредита. Исследование показывает, что машинное обучение может еще больше повысить точность кредитных рейтингов отдельных лиц и обеспечить научную основу и справочную информацию для кредитных рейтингов коммерческих банков.

**Ключевые слова:** большие данные, кредитный скоринг, логистическая регрессия, машинное обучение, десенсибилизация данных, Дерево решений, нейронная сеть БП.

### **Введение**

С развитием интернет-финансов и наступлением эры больших данных традиционная персональная оценка кредитоспособности коммерческих банков сталкивается с огромными проблемами. Во-первых, источники данных более обширны, а структуры данных разнообразны. Структурированные данные, в основном основанные на банковских кредитных данных, не могут точно отражать аккредитив субъекта кредита, а неструктурированные данные также являются важной информацией, отражающей кредитный статус субъекта кредита. Таким образом, структура данных включает структурированные данные и неструктурированные данные. Во-вторых, модель оценки более сложна. Модель оценки кредитоспособности коммерческих банков, основанная на модели логистической регрессии, не смогла обработать и проанализировать большой объем сложных данных. Поэтому для оценки необходимы сложные модели. В-третьих, процесс одобрения кредита в коммерческих банках относительно громоздкий, и при подаче заявки на необеспеченные кредиты возникнут

такие проблемы, как длительное одобрение, недостаточная квота и длительные интервалы между кредитами [1]. Компании, занимающиеся финансовыми платформами больших данных, собирают и систематизируют огромные объемы данных с помощью своих собственных платформ, улучшают параметры оценки кредитоспособности и используют методы машинного обучения для проведения всесторонних и научных оценок личного кредита. Оцените предмет кредита и решите, следует ли выдавать необеспеченные кредиты на основе его кредитного рейтинга. В то же время это экономит время и стоимость кредитных ссуд за счет электронных платежей и ликвидации. Таким образом, изучение личной кредитной оценки на основе методов машинного обучения имеет относительно большую теоретическую ценность и практическое значение.

#### **Методы исследования**

Теоретические: сравнение и анализ полученных данных.

Математические: методы математической статистики, метод логистической регрессии.

#### **Результаты исследований**

Целями персональной кредитной оценки являются точность, универсальность и объяснимость. Среди них точность является основой оценки кредитоспособности, которая требует кредитного скоринга, чтобы точно определить, будет ли объект кредита дефолтным. Универсальность требует, чтобы кредитный скоринг был не только хорошо различим на выборочных данных, но и сохранял высокую точность при применении к разным группам людей и в разных сценариях. С одной стороны, интерпретируемость требует способности соответствовать нормативным требованиям, а с другой стороны, она должна соответствовать бизнес-опыту и суждениям здравого смысла, чтобы модель оставалась точной в различных сценариях [2]. Существующие ученые имеют недостатки в исследованиях оценки личного кредита, в основном в трех аспектах: во-первых, при построении модели используется только модель логистической регрессии, а результатам оценки не хватает точности; во-вторых, невнимание к обработке пропущенных значений, в основном с использованием среднего или режима Дополнения пропущенных значений, может привести к возникновению шума и сделать результаты оценки неточными; в-третьих, методы оценки в основном основаны на экспертных методах оценки, таких как процесс аналитической иерархии, метод нечеткой оценки и т.д., а результаты оценки лишены объективности. Таким образом, изучение личной кредитной оценки на основе методов машинного обучения имеет большую теоретическую ценность и практическое значение.

(1) Ограничения существующих методов оценки кредитоспособности физических лиц. Для улучшения системы оценки кредитоспособности физических лиц и методов оценки коммерческих банков. Следовательно, необходимо обсудить ограничения существующих методов личной кредитной оценки.

А. Однократное использование модели логистической регрессии, отсутствие точности

Основной моделью персональной кредитной оценки в стране и за рубежом является модель FICO в Соединенных Штатах. Его теоретической основой является логистическая регрессия, которая подходит для решения задач классификации с меньшим количеством переменных. В основном используются кредитные данные, а количество переменных составляет 10-30 [3]. В контексте Интернета при оценке личной кредитоспособности необходимо учитывать кредитные данные, поведенческие привычки и социальные отношения. Структура данных разнообразна и содержит тысячи переменных. Однако модель логистической регрессии сама по себе не может извлечь характеристики тысяч переменных, поэтому она не может точно оценить объект кредита.

В. Дополнить пропущенные значения средним значением или модой, отсутствие точности

При проведении оценки кредитоспособности кредитных организаций часто отсутствуют данные, и отсутствие различных данных повлияет на качество данных. Обработка пропущенных данных в основном основана на статистических методах, таких как дополнение пропущенных значений средним значением или модой. Использование этого метода для интерполяции отсутствующих данных в определенной степени искусственно увеличивает шум (данные интерференции), что влияет на конечный результат. Достоверность результатов оценки.

С. Метод экспертной оценки необъективен

Традиционным методом исследования личной кредитоспособности является метод экспертной оценки, такой как процесс аналитической иерархии, метод нечеткой оценки и т. Д. Недостатком этого метода является то, что на мнения экспертов будут влиять мнения авторитетных экспертов на собрании. В то же время следователи и респонденты Психологическое состояние эксперта также будет влиять на результаты оценки экспертов, так что результаты оценки лишены объективности. Из приведенного выше анализа видно, что основные проблемы в исследовании оценки кредитоспособности населения сосредоточены на таких аспектах оценки кредитоспособности, как построение модели, обработка данных и методы оценки. Оценка личной кредитоспособности с использованием методов машинного обучения может эффективно компенсировать недостатки и повысить объективность и качество оценки личной кредитоспособности. точность.

(2) Преимущества использования методов машинного обучения для личной кредитной оценки

В эпоху финансовых технологий машинное обучение постоянно меняет жизнь людей, а использование машинного обучения для предотвращения финансового мошенничества привлекает всеобщее внимание крупных банков, финансовых компаний и интернет-компаний. Оценка кредитоспособности является эффективным средством предотвращения и контроля рисков, а модель персональной оценки кредитоспособности, основанная на машинном обучении, может эффективно прогнозировать риск дефолта субъектов кредита.

На основе комплексной статистики, информации, компьютерных и других дисциплин методы машинного обучения автоматически улучшают компьютерные алгоритмы и оптимизируют производительность алгоритмов, изучая, как компьютеры имитируют поведение человека при обучении. В основном он включает в себя модель случайного леса, модель дерева решений с градиентным усилением, модель нейронной сети BP ит. Д. С ее помощью можно реализовать точное прогнозирование уровня личного кредита.

А. Использование модели случайного леса для дополнения отсутствующих значений данных

Модель случайного леса можно использовать для классификации, прогнозирования, добавления пропущенных значений и т. д., и она является широко используемым методом машинного обучения. Его основная идея состоит в том, чтобы случайным образом выбрать k выборочных данных с заменой в больших выборочных данных в качестве обучающих выборок и случайным образом выбрать d индикаторов из каждого набора данных в качестве разделенных функций, чтобы обучить k базовых учащихся. Базовые обучающиеся объединяются для реализации классификации или регрессии (рис. 1) [4].

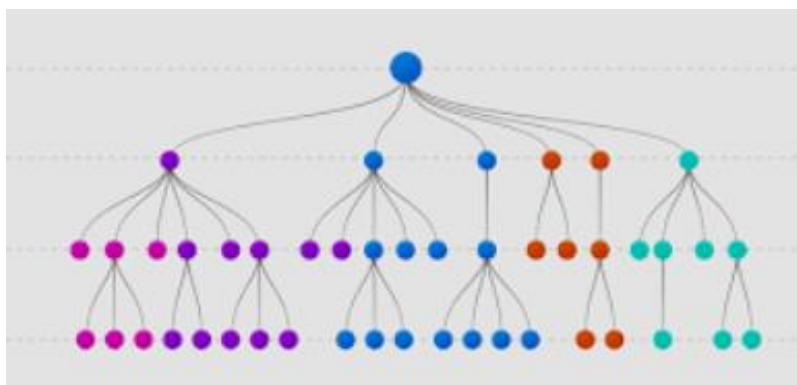


Рисунок 1 – Модель случайного леса

Модель случайного леса может обрабатывать многомерные данные и подходит для обработки большого количества выборочных данных субъекта кредита, дополняя недостающие значения в выборочных данных субъекта кредита.

В. Использование модели дерева решений с повышением градиента для проверки важных переменных.

Модель Gradient Boosting Decision Tree (GBDT) представляет собой модель кумулятивного анализа оценок, основанную на ошибках. Введите набор выборочных данных в модель дерева решений с повышением градиента и выберите индекс с наименьшим индексом Джини в качестве индекса разделения с помощью алгоритма. После разделения

корневого узла дочерние узлы повторяют тот же метод разделения до тех пор, пока нижний дочерний узел не выполнит заданные условия и не остановится. разделить и, наконец, вывести двоичное дерево [5].

Модель дерева решений с градиентным усилением эффективно предотвращает влияние изменений выборки и различных атрибутивных данных на стабильность модели и точность прогнозирования, а точность прогнозирования высока. В то же время модель может гибко обрабатывать линейные и нелинейные выборочные данные. Таким образом, он подходит для важной тестовой переменной больших выборок и многомерных выборочных данных субъекта кредита в контексте Интернета.

С. Используйте модель нейронной сети BP для точного прогнозирования кредитного статуса субъектов кредита.

Нейронная сеть BP (Back Propagation) является наиболее широко используемой нейронной сетью прямого распространения [6]. При прямом распространении информация обрабатывается слой за слоем от входного слоя к скрытому слою и, наконец, передается на выходной уровень. Если ожидаемый результат не получен в выходном слое, ошибка распространяется в обратном направлении в соответствии с исходным каналом, а вес нейронной сети корректируется таким образом, чтобы функция потерь была минимизирована, чтобы минимизировать ошибку нейронной сети.

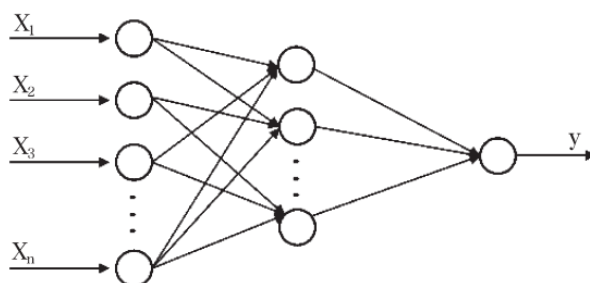


Рисунок 2 – Модель нейронной сети BP

В структуре нейронной сети BP (как показано на рис. 2) левый слой является входным слоем, средний – скрытым слоем, а правый – выходным слоем. По сравнению с моделью логистической регрессии модель нейронной сети BP лучше подходит для нелинейных сложных задач классификации. Среди них входной слой и скрытый слой нейронной сети BP представляют собой процесс извлечения признаков, который подходит для извлечения признаков многомерных выборочных данных для достижения цели уменьшения размерности, что полезно для нейронной сети выходного слоя для вывод более точной личной кредитной оценки.

(3) Создайте личную систему оценки кредитоспособности.

В контексте финансовых технологий персональная оценка кредитоспособности компаний, занимающихся платформой финансовых больших данных, имеет преимущества, которые нельзя сравнить с традиционной персональной кредитной оценкой коммерческих банков, такие как исчерпывающие источники данных, передовые методы анализа, широкий спектр услуг и разнообразие. сценарии применения. Изучая параметры оценки личной кредитоспособности существующих компаний, работающих на финансовых платформах больших данных, можно увидеть, что параметры оценки личной кредитоспособности включают в себя пять типов параметров оценки: основная личная информация, личная кредитная история, способность к работе, социальные отношения и личное поведение при транзакциях. По сравнению с американской FICO (Fair Isaac Corporation, американская компания по оценке личных потребительских кредитов), которая фокусируется на кредитной истории и производительности, отечественные финансовые компании-платформы больших данных вводят такие параметры оценки, как социальные отношения и поведение при совершении личных сделок, в соответствии с национальными условиями Казахстана, чтобы установить собственная система оценки кредитоспособности.

### Обсуждение научных результатов

Перед построением модели персональной кредитной оценки требуется сбор и обработка данных. Существует четыре основных канала для сбора персональных данных об оценке кредитоспособности: данные, предоставленные самими пользователями, интернет-

данные, собранные компаниями-платформами финансовых больших данных, данные с торговых платформ электронной коммерции, принадлежащие компаниям-платформам финансовых больших данных, и данные, собранные финансовой платформой больших данных. компании. Данные внешних агентств [7]. На основе сбора обширных данных обработайте и организуйте собранные данные, такие как десенсибилизация, добавление недостающих значений и т. д., для подготовки к дальнейшему моделированию.

### **Заключение**

Изучая процесс личной оценки кредитоспособности компаний-платформ финансовых больших данных для сбора и обработки массивных данных и использования методов машинного обучения для создания личных кредитных рейтингов, было обнаружено, что систему личной кредитной оценки коммерческих банков можно улучшить по трем аспектам.

Во-первых, в процессе оценки личного кредита нужно использовать метод теста LOF, чтобы проверить выбросы данных выборки и устранить их, а затем использовать модель случайного леса, чтобы дополнить пропущенные значения исходных данных выборки и пропущенные значения после удаления выбросы, чтобы заменить среднее или моду. Статистические методы для дополнения пропущенных значений для повышения точности прогноза.

Во-вторых, требуется отфильтровать характеристические переменные, которые влияют на оценку личного кредита, с помощью модели дерева решений с градиентным усилением, а затем создайте модель системы показателей на основе логистической регрессии или модели нейронной сети BP для оценки и прогнозирования отфильтрованных характеристических переменных и вывода личных кредитных баллов. Компенсируйте неточность кредитного рейтинга, вызванную однократным использованием кредитных данных и модели логистической регрессии для вывода личного кредитного рейтинга.

В-третьих, на основе метода машинного обучения личный кредитный рейтинг может быть выведен более объективно и точно, а также могут быть устранены дефекты односторонности и субъективности традиционного метода экспертной оценки.

### **Список литературы**

1. Сорокин, А.С. Применение законов распределения случайных величин для моделирования экономических явлений и процессов [Текст]: монография / Н.Я. Бамбаева, А.С. Сорокин – М.: МЭСИ. – 2010. – 156 с.
2. Ляо Биин. Исследование модели персонального кредитного рейтинга коммерческих банков [D]. Наньчан: Университет финансов и экономики Цзянси. – 2016.
3. ЮЯ. Сравнительное исследование оценки кредитоспособности физических лиц кредитной компании «Жима» [D]. Чанша: Хунаньский университет. – 2017.
4. Чэнь Чжипэн. Сравнительный анализ систем управления персональной кредитной информацией в стране и за рубежом [D]. Харбин: Хэйлунцзянский университет. – 2017.
5. Сорокин, А.С. К вопросу оценки согласованности мнений экспертов при использовании методов экспертной оценки в кредитном рейтинге. [Текст] /А.С. Сорокин // Роль бизнеса в преобразованиях общества – 2014: Сб. по мат. IX междунар. научн. конгр. – М.: «Иедитус». – 2014. – с. 281-283.
6. Цао Цзе, Шао Сяосяо. Исследование модели оценки личного кредита, основанной на интегрированном алгоритме обучения сбора и сбора информации [J]. Практика и понимание математики. – 2016. – 46 (8): 90-98.
7. Тан Чжунмин, Се Кун, Пэн Яопэн. Исследование оценки кредитного риска заемщиков онлайн-кредитов P2P на основе модели дерева решений с градиентным повышением [J]. Мягкая наука. – 2018, 32 (12): 136-140.

### **References**

1. Sorokin, A.S. Primenenie zakonov raspredelenija sluchajnyh velichin dlja modelirovanija jekonomicheskikh javlenij i processov [Tekst]: monografija. / N.Ja. Bambaeva, A.C. Sorokin – M.: MJeSI, 2010. – 156 c. (in Russian).
2. Liao Biying. Research on Personal Credit Rating Model of Commercial Banks [D]. Nanchang: Jiangxi University of Finance and Economics. – 2016. (in Russian).
3. Yu Ya. Comparative Research on Personal Credit Evaluation of Zhima Credit Company [D]. Changsha: Hunan University. – 2017. (in Russian).

4. Chen Zhipeng. Comparative analysis of personal credit information management systems at home and abroad [D]. Harbin: Heilongjiang University. – 2017. (in Russian).
5. Sorokin, A.S. K voprosu ocenki soglasovannosti mnenij jekspertov pri ispol'zovanii metodov jekspertnogo ocenivaniya v kreditnom skoringe. [Tekst] /A.C. Sorokin // Rol' biznesa v transformacii obshhestva – 2014: Sb. st. po mat. IX mezhdunar. nauchn. kongr. – M.: «Jeditus». – 2014. – s. 281-283. (in Russian).
6. Cao Jie, Shao Xiaoxiao. Research on Personal Credit Evaluation Model Based on Information Gain and Bagging Integrated Learning Algorithm [J]. Practice and Understanding of Mathematics. – 2016, 46 (8): 90-98. (in Russian).
7. Tan Zhongming, Xie Kun, Peng Yaopeng. Research on Credit Risk Evaluation of P2P Online Loan Borrowers Based on Gradient Boosting Decision Tree Model [J]. Soft Science. – 2018, 32 (12): 136-140. (in Russian).

**Ж.М. Ордабаева\*, А.Н. Молдагулова**

Қ.І. Сәтбаев атындағы Қазақ ұлттық зерттеу техникалық университеті  
050013, Қазақстан Республикасы, Алматы қ., көш. Сәтпаева 22а  
\*e-mail: zhannaordabayeva@gmail.com

### **МАҚСАТТЫ АЙНЫМАЛЫ ТҮРІНЕ НЕГІЗДЕЛГЕН НЕСИЕЛІК СКОРИНГ МОДЕЛІН ҚҰРУ**

*Үлкен деректер мен интернет технологияларының қарқынды дамуымен үлкен деректер қаржы платформалары компаниялары өз платформалары арқылы жаппай деректерді жинайды және ұйымдастырады, несиелік скоринг параметрлерін жақсартады және несиелік скорингтің кешенді және ғылыми бағалауларын жүргізу үшін машиналық оқыту әдістерін қолданады. Осылайша, Банктер Несиелік скорингті құру кезінде үлкен қиындықтарға тап болады. Қолданыстағы жүйенің шектеулеріне және дербес кредиттік рейтинг әдістеріне сүйене отырып, Машиналық оқыту әдістері негізінде дербес кредиттік рейтингті зерделеу, дербес кредиттік рейтингтің параметрлері мен балдық бағалау жүйесін жетілдіру, деректерді жинау арналарын нақтылау, динамикалық десенсибилизация технологиясын пайдалану қажет. деректердің сезімталдығын төмендету үшін шығарындылар туралы деректерді тексеру үшін LOF тәсілдеу әдісі және жетіспейтін деректер мәндерін толтыру үшін кездейсоқ орман әдісі. Содан кейін сіз маңызды индикаторларды көру, логистикалық регрессияға негізделген индикаторлар жүйесінің моделі арқылы тексерілген индикаторларды өңдеу және жеке несиелік балл алу үшін градиентті жоғарылататын шешімдер ағашының әдісін қолданасыз. Соңында, модель BP нейрондық желісі арқылы тексеріледі және модель жеке несие деңгейін болжау үшін қолданылады. Зерттеу көрсеткендей, Машиналық оқыту жеке тұлғалардың несиелік ұпайларының дәлдігін одан әрі жақсарта алады және коммерциялық банктердің несиелік ұпайлары үшін ғылыми негіз мен анықтамалық ақпарат береді.*

***Түйін сөздер:** үлкен деректер, несиелік скоринг, логистикалық регрессия, Машиналық оқыту, деректерді десенсибилизациялау, шешім ағашы, ПД нейрондық желісі.*

**Z.M. Ordabayeva\*, A.N. Moldagulova**

Kazakh National Research Technical University named after K. I. Satpaev  
050013, Republic of Kazakhstan, Almaty, st. Satpaeva 22a  
\*e-mail: zhannaordabayeva@gmail.com

### **BUILDING A CREDIT SCORING MODEL BASED ON THE TYPE OF TARGET VARIABLE**

*With the rapid development of big data and Internet technologies, companies engaged in big data financial platforms collect and systematize massive data through their own platforms, improve credit scoring parameters and use machine learning methods to conduct complex and scientific credit scoring assessments. Thus, banks face big problems when building credit scoring. Based on the limitations of the existing system and methods of personal credit rating, it is necessary to study personal credit rating based on machine learning methods, improve the parameters and scoring system of personal credit rating, clarify data collection channels, use dynamic desensitization technology. To reduce the sensitivity of the data, the LOF test method is used to verify the emission*

*data and the random forest method is used to fill in missing data values. Then you use the gradient-boosting decision tree method to view important indicators, process proven indicators using a metric system model based on logistic regression, and get a personal credit score. Finally, the model is tested using a BP neural network, and the model is used to predict the level of personal credit. The study shows that machine learning can further improve the accuracy of individuals' credit ratings and provide a scientific basis and background information for commercial banks' credit ratings.*

**Key words:** *big data, credit scoring, logistic regression, machine learning, data desensitization, Decision tree, BP neural network.*

#### **Сведения об авторах**

**Жанна Муратовна Ордабаева** – PhD докторант кафедры «Программная инженерия»; Казахский национальный исследовательский технический университет имени К.И. Сатпаева, Алматы, Казахстан; e-mail: zhannaordabayeva@gmail.com. ORCID: 0009-0009-8273-3147.

**Айман Николаевна Молдагулова** – кандидат физико-математических наук кафедры «Физика и математика»; профессор кафедры «Программная инженерия»; Казахский национальный исследовательский технический университет имени К.И. Сатпаева, Алматы, Казахстан; e-mail: a.moldagulova@satbayev.university. ORCID: 0000-0002-1596-561X.

#### **Авторлар туралы мәліметтер**

**Жанна Муратовна Ордабаева** – «Бағдарламалық қамтамасыз ету» кафедрасының PhD докторанты; Қ.И.Сәтбаев атындағы Қазақ ұлттық зерттеу техникалық университеті, Алматы, Қазақстан; e-mail: zhannaordabayeva@gmail.com. ORCID: 0009-0009-8273-3147.

**Айман Николаевна Молдагулова** – физика-математика кафедрасының физика-математика ғылымдарының кандидаты; Бағдарламалық қамтамасыз ету инженериясы кафедрасының профессоры; Қ.И. Сәтбаев атындағы Қазақ ұлттық зерттеу техникалық университеті, Алматы, Қазақстан; e-mail: a.moldagulova@satbayev.university. ORCID: 0000-0002-1596-561X.

#### **Information about authors**

**Zhanna Ordabayeva** – PhD doctoral student of the department "Software Engineering"; Kazakh National Research Technical University named after K. I. Satpayev, Almaty, Kazakhstan; e-mail: zhannaordabayeva@gmail.com. ORCID: 0009-0009-8273-3147.

**Ayman Moldagulova** – Candidate of Physical and Mathematical Sciences of the Department of Physics and Mathematics; Professor of the Department of Software Engineering; Kazakh National Research Technical University named after K. I. Satpayev, Almaty, Kazakhstan; e-mail: a.moldagulova@satbayev.university. ORCID: 0000-0002-1596-561X.

*Материал поступил в редакцию 01.03.2023 г.*

DOI: 10.53360/2788-7995-2023-1(9)-8

MPHTI: 20.19.27

**K. Nursakitov\*, A. Bekishev, S. Kumargazhanova, A. Urkumbaeva**

D. Serikbayev East Kazakhstan technical university,  
070004, The Republic of Kazakhstan, Ust-Kamenogorsk, 69 Protozanov Street  
\*e-mail: nursakitov@bk.ru

#### **REVIEW OF METHODS FOR DETERMINING THE TONATION OF TEXTS IN NATURAL LANGUAGES**

**Annotation:** *The analysis of sentiment in user comments finds application in many areas, such as evaluating the quality of goods and services, analyzing emotions in messages, and detecting phishing advertisements. There are numerous methods for analyzing the sentiment of textual data in the Russian language, but automatic sentiment analysis of Russian-language texts is much less developed than for other major world languages. This article is part of a broader study on the creation*