УДК 519.862
ББК 22.172

МОДЕЛИ КРЕДИТНОГО СКОРИНГА НА ОСНОВЕ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ

Низамитдинов Ахлитдин Илёситдинович *- доктор философии (PhD), старший преподаватель кафедры цифровой экономики Политехнического института Технического университета Таджикистана им. академика М.С. Осими, e-mail: ahlidin@mail.com*

МОДЕЛИ СКОРИНГИ ҚАРЗИ ДАР АСОСИ АЛГОРИТМҲОИ ОМӮЗИШИ МОШИНӢ

Низомитдинов Ахлитдин Илеситдинович *- доктори фалсафа (PhD), муаллими калони кафедраи иқтисоди рақамӣ Донишкадаи политехникии Донишгоҳи техникии Тоҷикистон ба номи академик М.С. Осимӣ, e-mail: ahlidin@mail.com*

CREDIT SCORING MODELS BASED ON MACHINE LEARNING ALGORITHMS

Nizamitdinov Akhlitdin Ilyositdinovich *- Doctor of Philosophy (PhD), The Senior Teacher of the Digital Economy Department at KPITTU, e-mail: ahlidin@mail.com*

Ключевые слова: *кредитование, кредитный скоринг, риск дефолта, алгоритмы машинного обучения, логистическая регрессия, дерево решений.*

Кредитование играет важную роль в финансовом мире на протяжении многих лет. Данный вид финансовой деятельности стал популярным в финансовом секторе Республики Таджикистан в последнее десятилетие.

В статье рассматриваются методы построения моделей кредитного скоринга в мировой практике. Эксперты и исследователи данной сферы в основном присваивают людям числовые баллы, известные как кредитные баллы для измерения риска и кредитоспособности клиентов. Хотя этот вид финансовой деятельности является довольно выгодным, однако он несет большой риск, который в сфере ссудного кредитования называется кредитным риском. Поэтому, прогнозирование кредитного дефолта клиента является наиболее важным аспектом в финансовой деятельности организаций. В последние годы стали часто использоваться алгоритмы машинного обучения в задачах классификации кредитного дефолта. Такими алгоритмами являются логистическая регрессия, к-ближайших соседей (KNN), дерево решений (decision tree), случайный лес (Random Forest) и др.

Калидвожаҳо: *қарздиҳӣ, баҳодиҳии қарзӣ, хавфи пешфарз, алгоритмҳои омӯзиши мошинӣ, регрессияи логистикӣ, дарахти қарорҳо*

Қарздиҳӣ солҳои зиёд дар ҷаҳони молиявӣ нақши муҳим дорад. Ин намуди фаъолияти молиявӣ дар даҳсолаи охир дар бахши молияи Ҷумҳурии Тоҷикистон низ маъмул гаштааст. Дар мақола дар бораи усулҳои сохтани моделҳои баҳодиҳии қарзӣ дар таҷрибаи ҷаҳонӣ сухан меравад. Барои муайян кардани хавф ва қобилияти қарздиҳии муштариён, коршиносон ва муҳаққиқони соҳа пеш аз ҳама ба одамон холҳои ададӣ медиҳанд. Гарчанде ки ин намуди фаъолияти молиявӣ хеле фоидаовар аст, аммо он хавфи калон дорад, ки онро дар соҳаи қарздиҳӣ хавфи кредитӣ меноманд. Аз ин рӯ, пешгӯии қарзи муштарӣ ҷанбаи муҳимтарин дар фаъолияти молиявии ташкилотҳо мебошад. Дар солҳои охир алгоритмҳои омӯзиши мошинӣ дар мушкилоти таснифоти пешфарз бештар истифода мешаванд. Чунин алгоритмҳо регрессияи логистикӣ, ҳамсояҳои наздиктарин (KNN), дарахти қарор, ҷангали тасодуфӣ ва ғайра мебошанд.

Key words: *lending, credit scoring, default risk, machine learning algorithms, logistic regression, decision tree*

Lending has played an important role in the financial world for recent years. This type of financial activity has become popular in the financial sector of the Republic of Tajikistan in the last decade. The article discusses methods for constructing credit scoring models in world practice. Experts and researchers in this field primarily assign numerical scores to people, known as credit scores, to measure the risk and their creditworthiness. Although this type of financial activity is quite profitable, it carries a large risk, which in the field of loan lending is called

credit risk.

Therefore, forecasting a client's credit default is the most important aspect in the financial activities of organizations. In recent years, machine learning algorithms have become frequently used in credit default classification problems. Such algorithms are logistic regression, k-nearest neighbors (KNN), decision tree, random forest, etc.

1. Введение

Кредитный скоринг – это статистический анализ, проводимый кредиторами и финансовыми учреждениями для определения кредитоспособности клиента. Кредиторы используют кредитный скоринг, для принятия решения о предоставлении кредита или отказе в нем.

Для вычисления кредитного балла (скора) прежде использовались традиционные методы подсчета по весам переменных, которые определялись традиционными статистическими методами. Представление о моделировании кредитного риска изменились по мере роста количества клиентов, когда традиционные методы не могли находить взаимосвязи переменных в динамическом изменении характеристик. Поэтому, появился спрос на создание модели машинного обучения в этой области. Однако многие регулирующие органы по-прежнему очень осторожно относятся к переходу на методы машинного обучения. Возможно, что на этом этапе трансформации алгоритмы машинного обучения будут работать вместе с традиционными методами.

Доверие со стороны регулирующих органов может быть достигнуто после того, как будет установлено, что алгоритмы машинного обучения, бросая вызов принятым в данной области правилам, также дают более надежные результаты, чем традиционные методы. Более того, новые методы интерпретации алгоритмов машинного обучения могут помочь создать более прозрачный процесс.

Кредитование стало важной частью повседневной жизни как для организаций, так и для частных лиц. С постоянно растущей конкуренцией в финансовом мире и из-за огромного количества финансовых ограничений в структуре кредитования, взятие долгового кредита стало более или менее неизбежным [1]. Физические лица по всему миру зависят от активности взятия кредитов и кредитования по таким причинам, как преодоление финансовых ограничений для достижения каких-то личных целей.

Хотя кредитование выгодно как для кредиторов и получателей и считается важным для финансовой организации, она несет в себе определенные риски [3]. Этот тип риска представляет собой неспособность клиента выплачивать кредит в назначенное время, которое было принято решением между кредитором и заемщиком в течение возникновения кредита и именуется кредитным риском [1].

Кредитный риск, как известно, вызывает серьезные опасения среди финансовых институтов, поскольку это может привести к ситуации, известной как дефолт по кредиту, который может оказаться серьезным [4]. Даже при определенном риске, финансовые институты по всему миру рассматривают деятельность кредитования как основную возможность для получения прибыли и необходимым для бесперебойного функционирования их бизнеса [5].

1. Алгоритмы машинного обучения

На протяжении последнего десятилетия алгоритмы машинного обучения используются для расчета и прогнозирования кредитного риска путем оценки исторических данных о физических лицах. В данной статье рассматривается наиболее часто используемые алгоритмы моделирования и прогнозирования оценки риска, в которых используются машины алгоритмы обучения.

1.1. Логистическая регрессия

Логистическая регрессия - один из самых популярных статистических методов среди других в финансовом мире для модели оценки кредитного риска. Основные сильные стороны модели логистической регрессии заключаются в ее простом понимании, высокой производительности и простоте реализации [8]. Более того, логистическая регрессия превосходит линейную регрессию, поскольку преодолевает множество проблем, таких как линейная регрессия-результат регрессии может быть отрицательным или большим чем значение 1, что невозможно для вероятности.

Логистическая регрессия решает эту проблему, обеспечивая непрерывный диапазон оценок от 0 до 1 и сохранение вывода ограничено значениями от 0 до 1 [9].

В логистической регрессии используется логистическая функция, которая определяется по следующей формуле.

$$p(X) = \frac{e^{\beta_0+\beta_1 x+\cdots+\beta_m}}{1+e^{\beta_0+\beta_1 x+\cdots+\beta_m}} \quad (1)$$

Для подгонки модели к набору независимых переменных используется метод наименьших квадратов или метод максимального правдоподобия. В результате оценки неизвестных коэффициентов регрессии выбирается оптимальная модель для вычисления вероятности. Логистическая функция строит S-форменный график вероятности с диапазоном значений от 0 до 1. График построения модели линейной и логистической регрессии показан на рис. 1. [10].

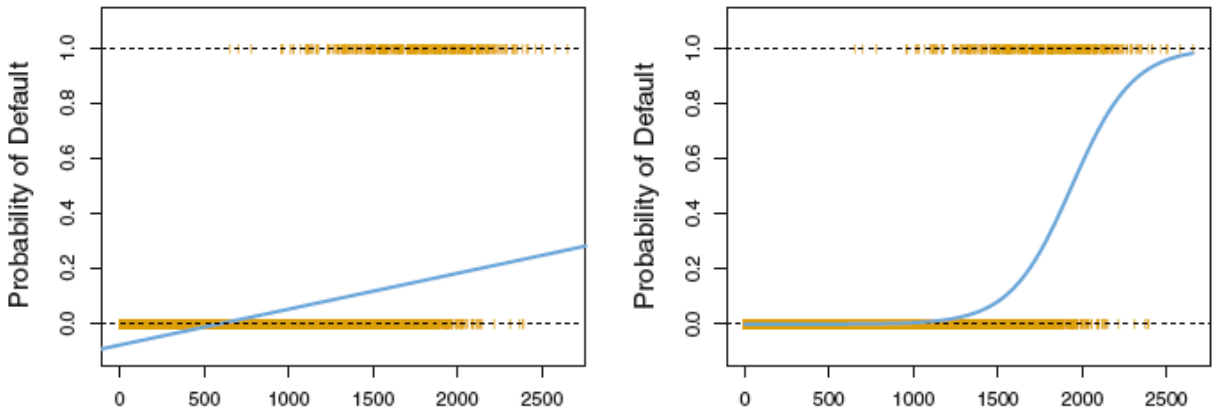

Рис.1. График построения модели а) линейной регрессии б) логистической регресии

1.2. Дерево решений (Decision tree)

Один из самых популярных методов классификации используемый в области кредитного скоринга известен как дерево решений, состоящего из нескольких ветвей, корневых узлов и листовых узлов. Как следует из названия, методика генерирует структуру похожую на дерево, классифицируя экземпляры и с использованием рекурсивного алгоритма разделения (recursive partitioning algorithm) [9].

Каждый листовой узел в дереве представляет метку класса и все ветви в дереве представляют результаты теста, эти тесты представлены внутренними узлами для характеристик [8].

Деревья решений могут быть использованы как для регрессионных, так и для классификационных проблем. Классификационные деревья предназначены для прогнозирования качественных зависимых переменных. В классификационных деревьях решений прогнозируется как каждое наблюдение относится к наиболее частой встречаемости класса наблюдений предназначенных для обучения[10].

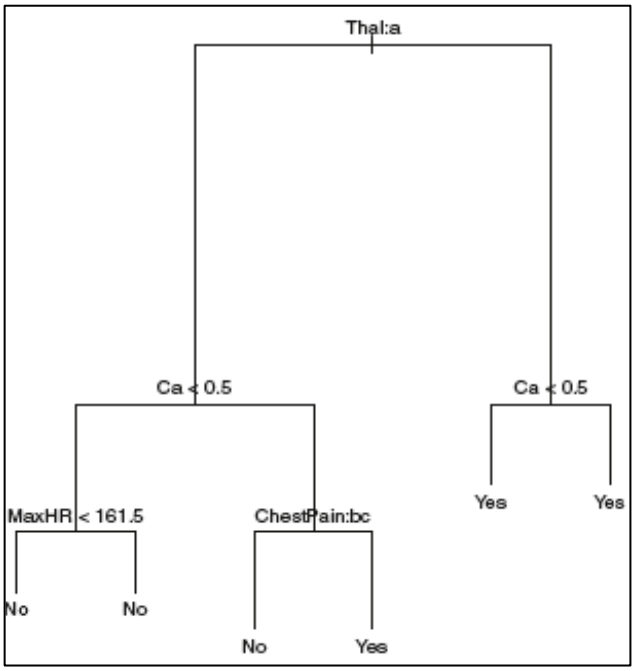

Рис.2. Пример построения дерева решений

1.3. Случайный лес (Random forest)

Классификация алгоритмом случайного леса представляет собой алгоритм который собирает

деревья решений используя случайные подмножества характеристик и выбирая наиболее повторяющиеся ветви между ними для классификации. Этот алгоритм лучше работает для предотвращения сверхподгонки(overfitting) и увеличивает общую аккуратность модели [10]. Вид построения модели случайного леса приводится на рисунке 3.

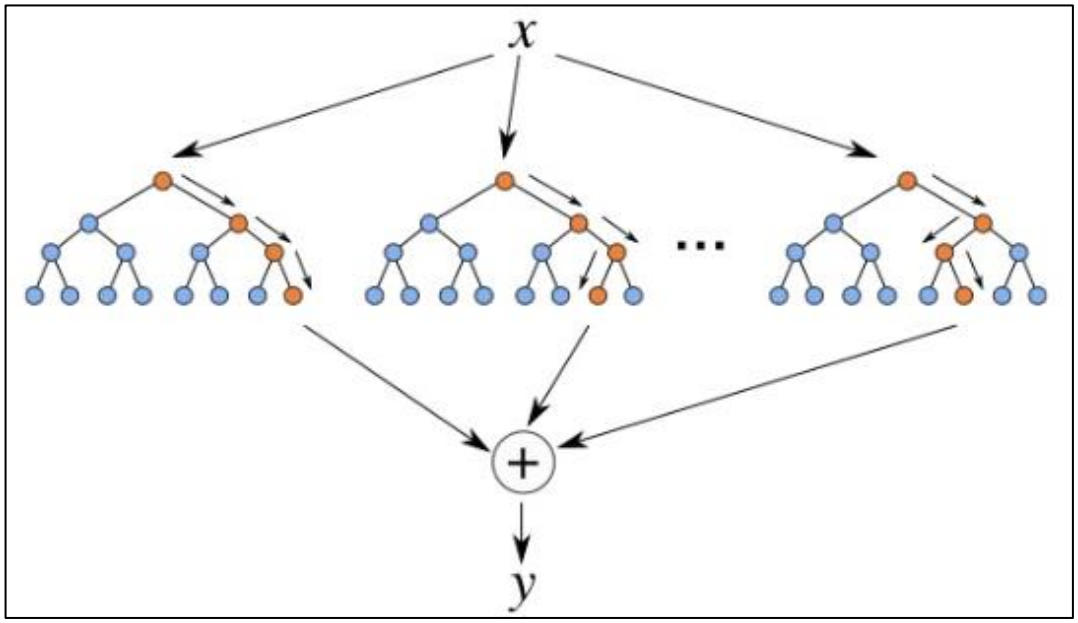

Рис.3. Пример построения модели алгоритмом случайный лес

2. Набор данных и выбор переменных

Для построения моделей кредитного скоринга необходим набор исторических данных по наиболее значимым переменным для прогнозирования кредитного дефолта клиента. Во многих изученных исследованиях наиболее часто встречаемыми характеристиками клиентов являются возраст клиента, пол, семейное положение, вид деятельности, образование, тип кредита, собственность жилья, количество детей, цель кредита, сумма кредита, вид залога, длительность кредита.

Для практического проведения эмпирического анализа используются несколько наборов данных которые доступны в общем доступе. К ним относятся UC Irvine Machine Learning repository (UCI - ML) – репозиторий машинного обучения который содержит различные наборы данных для проведения анализа с использованием алгоритмов машинного обучения. Для кредитного скоринга используется два набора данных из данного репозитория, набор данных Австралийского кредитного скоринга (Australian credit dataset) [11] и набор данных кредитного скоринга Германии (German credit dataset) [12].

Набор данных Австралийского кредитного скоринга имеет 14 характеристик клиентов и 690 наблюдений. Набор данных кредитного скоринга Германии включает 20 характеристик и 1000 наблюдений. Зависимая переменная представляет собой бинарное решение, является ли клиент кредитоспособным или нет. Оба набора данных имеют общие характеристики, такие как, кредитный балл, цель кредита, информацию клиента (должность, возраст, длительность кредита и т.д.).

На примере данных по кредитам Германии далее описывается переменные, которые были выбраны для анализа. Оригинальный набор данных содержит 1000 наблюдений с 20 количественными и качественными переменными которые были подготовлены профессором Хоффманом. В этом наборе данных каждое наблюдение представляет клиента, который взял кредит в банке. Каждый клиент классифицируется как хороший или плохой риск кредита в зависимости от набора характеристик клиента.

Представляемый набор данных не подготовлен для построения модели и его невозможно понять без предварительной обработки, так как набор данных представляется в виде сложной системы категорий и символов. Соответственно необходимо написать код программы на Python или R для того, чтобы построить читаемый файл в виде таблицы и сохранить в виде файла *.csv. Построенный набор данных содержит основные характеристики клиента, которые приводятся в таблице 1.

Таблица 1. Переменные, используемые в наборе данных

№	Имя переменной	Тип переменной	Значения
1	Age (Возраст)	Числовой	18-80
2	Gender	Текстовый	Мужчина, женщина
3	Job	Числовой	Неквалифицированный, Иногородний
4	Housing	Текстовый	Владелец, Аренда, Бесплатный
5	Saving accounts	Текстовый	маленький, умеренный, довольно богатый, богатый
6	Checking account	Числовой	В немецких марках
7	Credit amount	Числовой	В немецких марках
8	Duration	Числовой	В месяцах
9	Purpose	Текстовый	машина, мебель/техника, радио /ТВ, бытовая техника, ремонт, образование, бизнес, отдых/ прочее

Для создания и внедрения модели используются наиболее распространенные программные обеспечения бизнес-решений, такие как SPSS, SAS, STATISTICA на основе пользовательских интерфейсов. В последнее время также стали популярны программные обеспечения как услуги (Software as a Service), такие как Tableau, SAP, Oracle BI и другие. Но для самостоятельного создания моделей кредитного скоринга и последующего его внедрения в системы финансовых секторов наиболее часто используемыми языками программирования являются Python и R.

ЗАКЛЮЧЕНИЕ

В данной статье рассмотрены перспективы использования алгоритмов машинного обучения для построения модели кредитного скоринга в Таджикистане на основе анализа исследований, проведенных в последние годы. Было определено, что наиболее часто используемыми алгоритмами машинного обучения для построения моделей кредитного скоринга в мировой практике являются модели регрессионного характера и моделей, построенных на основе древовидных алгоритмов. Приводится информация о логистической регрессии, алгоритма дерева решений и случайного леса.

Было выявлено, что для построения модели необходимы наборы исторических данных финансовых институтов. Приводится пример использования двух наборов данных-данные Австралийского кредитного скоринга и набор данных кредитов Германии. Оба набора данных имеют сравнительно одинаковые характеристики клиентов для анализа, что свидетельствует о том, что существуют такие характеристики клиентов, которые являются базовыми для построения кредитного скоринга.

Следует отметить, что для внедрения кредитного скоринга в финансовых учреждениях Республики Таджикистан необходимо изучить мировой опыт внедрения средств построения моделей кредитного скоринга.

ЛИТЕРАТУРА
REFERENCES

1. S. Huang and M. Day, A comparative study of data mining techniques for credit scoring in Banking, in Proc. 14th IEEE Int. Conf. Information Reuse and Integration, 2013, pp. 684–691, doi: 10.1109/IRI.2013.6642534.
2. E. Kambal, I. Osman, M. Taha, N. Mohammed and S. Mohammed, Credit scoring using data mining techniques with particular reference to Sudanese banks, in Proc. Int. Conf. Electrical and Electronics Engineering, 2013, pp. 373–383, doi: 10.1109/ ICCEEE.2013.6633966.
3. Perera, H.A.P.L., Premaratne, S.C., 2016. An Artificial Neural Network Approach for the Predictive Accuracy of Payments of Leasing Customers in Sri Lanka.
4. Marqués, A.I., García, V. and Sánchez, J.S., 2012. Exploring the behaviour of base classifiers in credit scoring ensembles. Expert Systems with Applications, 39(11), pp.10244–10250.
5. Adewusi, A.O., Oyedokun, T.B. and Bello, M.O., 2016. Application of artificial neural network to loan recovery prediction. International Journal of Housing Markets and Analysis, 9(2), pp.222–238.
6. Choudhary, G., Garud, Y., Shetty, A., Kadakia, R. and Borase, S., 2019. Loan Default Identification and Its Effect.

7. Atiya, A.F., 2001. Bankruptcy prediction for credit risk using neural networks: A survey and new results. IEEE Transactions on Neural Networks, 12(4), pp.929–935.
8. Baesens, B., Roesch, D. and Scheule, H., 2016. Credit Risk Analytics: Measurement Techniques, Applications, and Examples in SAS. United States, John Wiley & Sons.
9. Nali´c, J. and Švraka, A., 2018. Using Data Mining Approaches to Build Credit Scoring Model: Case Study—Implementation of Credit Scoring Model in Microfinance Institution. 2018. 17th International Symposium Infoteh-Jahorina (INFOTEH), IEEE. pp.1–5.
10. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (1st ed.) [PDF]. Springer.
11. UC Irvine Machine Learning Repository Australian Credit Data. http://archive.ics.uci.edu/ml/datasets/Statlog+%28Australian+Credit+Approval%29
12. UC Irvine Machine Learning Repository German Credit Data. https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data).