УДК 336:338.27

СРАВНИТЕЛЬНЫЙ АНАЛИЗ ТЕХНОЛОГИЙ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ЗАДАЧ КРЕДИТНОГО СКОРИНГА

Татаринцев М.А., Никитин П.В., Горохова Р.И., Долгов В.И.

Финансовый университет при правительстве РФ, Москва, e-mail: PVNikitin@fa.ru

Машинное обучение позволяет такому банковскому бизнесу, как кредитование, нести минимальные потери. Так, правильная классификация заемщика при его обращении с запросом на получение кредита сводит вероятность предоставления банком денежных средств под проценты сомнительному клиенту к минимуму, а безопасному – к максимуму. Такой исход, в свою очередь, гарантирует как преувеличение капитала за счет выдачи займов тем, кто их сможем погасить, так и его сохранение благодаря возможности применить «фильтр» и избежать кредитования ненадежных клиентов. Подобное стало реализуемо благодаря решению задачи, которое получило название «кредитый скоринг». В данной статье будут рассмотрены три технологии, реализованные авторами для достоверной классификации заемщиков: Artificial neural network, XGBoost classifier и Random forest classifier. Применимость всех методов была проверена экспериментально на реальных данных. В качестве данных выступает история кредитования граждан США, опубликованная компанией Lending Club за 9 месяцев 2022 г. Грамотная предобработка данных в совокупности с корректно реализованными моделями позволила добиться высокой точности предсказания, что может свидетельствовать об удачно выбранных технологиях.

Ключевые слова: кредитный скоринг, машинное обучение, искусственная нейронная сеть, метод случайного леса, градиентный бустинг

COMPARATIVE ANALYSIS OF TECHNOLOGICAL MACHINE LEARNING FOR CREDIT SCORING PROBLEMS

Tatarintsev M.A., Nikitin P.V., Gorokhova R.I., Dolgov V.I.

Financial University under the Government of the Russian Federation, Moscow, e-mail: PVNikitin@fa.ru

Machine learning allows such banking business as lending to suffer minimal losses. Thus, the correct classification of the borrower when applying for a loan reduces the probability of the bank providing funds at interest to a doubtful client to a minimum, and to a safe one – to a maximum. Such an outcome, in turn, guarantees both an exaggeration of capital due to the issuance of loans to those who can repay them, and its preservation due to the possibility of apply a «filter» and avoid crediting unreliable customers. This became feasible thanks to the solution of the problem, which was called «credit scoring». This article will use three technologies implemented by the authors for reliable classification of borrowers: Artificial neural network, XGBoost classifier and Random forest classifier. The applicability of all methods was tested experimentally on real data. The data is the history of lending to US citizens, published by Lending Club for 9 months of 2022. Competent data preprocessing in combination with correctly implemented models allowed to achieve high prediction accuracy, which may indicate well-chosen technologies.

Keywords: credit scoring, machine learning, artificial neural network, random forest method, gradient boosting

Машинное обучение является одной из самых популярных тем для обсуждения в области науки о данных последнего десятилетия. И такая актуальность вполне обоснована – специально разработанные алгоритмы научились решать даже самые нетривиальные задачи с выгодой для бизнеса. Минимизация простоев на производствах, распознавание фотографий, текста или музыкальных произведений, выявление угроз безопасности, принятие правильных управленческих решений в сфере маркетинга – все это перестало быть чем-то удивительным сегодня. Спектр областей применения машинного обучения огромен, и каждая индустрия имеет свою специфику. В данной статье внимание авторов будет сосредоточено на решении задачи кредитного скоринга.

В современном мире банки играют ключевую роль в развитии мировой экономики. Если производить сопоставление классических коммерческих банков с человеческим организмом, то мировой финансовый рынок – это физическая структура тела, а банки – его сердечно-сосудистая система. Выдавая кредиты и выступая посредником в инвестиционных сделках, банки осуществляют аккумуляцию и перераспределение денежных средств среди достойных кандидатов, тем самым обеспечивают оптимальное давление в импровизированной кровеносной системе [1]. Однако что будет происходить при нарушении такого баланса? Скажем, как повлияет на человеческий организм изредка повышающийся уровень холестерина? Проявления какой-то неожиданно экстремальной реакции организма в этом случае скорее всего ждать не стоит. Но если это будет носить закономерный и регулярный характер, то последствия могут быть катастрофическими. Похожим образом на «здоровье» банковского секто-

ра будет влиять выдача кредитов несостоятельным юридическим и физическим лицам. Если заемщик является недостаточно надежным, то риск невозврата выданных под процент денежных средств начинает резко возрастать [2]. Учащающиеся случаи дефолтов будут наносить серьезный удар по финансовому состоянию банков и, как следствие, по экономике страны. Чтобы описанная выше ситуация не случилась, ученые по данным начали исследовать потенциальные варианты для максимизации вероятности отличия надежного заемщика от недостаточно надежного. Если еще в конце XX в. решение об одобрении кредита принимали люди [3], опираясь на их экспертизу в решении этого вопроса, то сейчас даже в самых маленьких банках России используются технологии машинного обучения [4]. Так, специалисты в области анализа данных смогли придумать, реализовать и опытным путем опробовать решение задачи, которая позже получила название «Кредитный скоринг».

По своей сущности оценка кредитоспособности заемщиков сводится к принятию банком решения о выдаче кредита клиенту, т.е. всесторонне рассматривается способность исполнения заемщиком взятых на себя обязательств. Надежность при этом обеспечивается благодаря использованию скоринговых систем – специалисты банка при оформлении заявки на получение кредита фиксируют набор специальным образом подобранных параметров, на основании которых модель кредитного скоринга, обученная по данным (клиентский опыт), будет осуществлять прогноз, результатом которого будет ответ на вопрос: «Учитывая риск данного заемщика, целесообразно ли выдавать кредит?» [5] Рассматриваемая задача прогнозирования является задачей классификации. Например, модель может разделить людей, которые подали заявки на получение займа, на два крупных класса: надежные и сомнительные. Кажется, что такая сепарация выглядит слишком просто, однако такой подход позволяет наиболее точно ответить на вопрос, который был поставлен ранее.

Итак, на основании заранее собранных данных модель кредитного скоринга будет осуществлять прогнозирование класса, в который заемщиков можно будет отнести: «надежные» или «сомнительные».

Для того чтобы углубиться в специфику реализации алгоритма кредитного скоринга на конкретном примере, первоначально необходимо рассмотреть наиболее распространенные вариации задачи. Специалисты, которые работают в банковской сфере в области кредитования, выделяют следующие крупные виды скорингов: заявочный, поведенческий и мошеннический [6]. Первый тип может быть применен к двум категориям клиентов – тем, которые уже имеют кредитную историю (происходит сопоставление данных потенциального заемщика с данными по людям, в отношении которых в прошлом уже было принято решение о выдаче займа), и тем, кто такой истории не имеет. В общем случае это самая распространенная категория рассматриваемых задач.

Задача поведенческого скоринга сводится к прогнозированию поведения клиента. Часто банку требуется не только определение вероятности дефолта по кредиту конкретного человека, но и исследование частоты и объема выплат. Например, если вы собираетесь получать кредит в коммерческом банке, который является также вашим зарплатным банком, то вся история ваших трат и начислений будет служить точкой опоры для предсказания, будут ли осуществляться выплаты по заранее установленному плану равномерно.

Мошеннический тип предназначен для борьбы с недобросовестными заемщиками и используется преимущественно государственными силовыми структурами или службами безопасности банков.

Очевидно, что две последние разновидности могут быть рассмотрены только в условиях доступа к соответствующим данным. В текущем исследовании будет построена модель кредитного скоринга заявочного типа, поскольку для нее могут быть использованы открытые данные о выданных кредитах.

Целью исследования является применение методов машинного обучения на реальных данных о кредитовании с задачей максимизировать вероятность правильной классификации заемщиков.

Материалы и методы исследования

Для решения задачи кредитного скоринга мы будем использовать данные, которые содержатся в открытом доступе и при этом являются достоверными и достаточно полными. Lending Club – крупнейшая платформа по одноранговому кредитованию клиентов из США, которая выкладывает данные о своих пользователях в открытый доступ (URL: www.lendingclub.com). Ежеквартально датасеты дополняются и выкладываются на портале Kaggle.com с детальным описанием каждой переменной. Когда потенциальный заемщик отправляет заявку на получение кредита, то компания должна принять решение об одобрении займа на основе профиля заявителя. Для этого предоставляется и специальным образом заверя-

ется информация, необходимая для оценки кредитоспособности.

Мы будем использовать обновленные данные за 9 месяцев 2022 г., сгруппированные в 73 столбца (признак по каждому клиенту) и 400 тысяч строк (число одобренных заявителей). Именно эта пользовательская база станет основой для нашего исследования.

В первую очередь будет проведен исследовательский анализ данных, который позволит детально рассмотреть все ключевые переменные, представленные в датасете. Далее будет происходить предобработка данных, которая включает в себя удаление незначимых переменных, детектирование аномальных, пустых значений и их дальнейшую ликвидацию, преобразование категориальных признаков в численные, введение суррогатных переменных и нормализацию. После этого набор данных был разделен на обучающую и тестовую выборки.

Следующий этап будет заключаться в непосредственном решении задачи классификации. Для этого было решено выбрать модели, наиболее применимые на практике в кредитном скоринге, а именно: искусственную нейронную сеть (ANN), XGBoost классификатор и модель случайного леса. В качестве метрик качества были выбраны классические показатели – ROC, f1-score, Precision, Recall, Accuracy. Для всех трех моделей был произведен оптимальный подбор параметров в целях максимизации эффекта точности. Далее были произведены обучение моделей, сопоставление результатов и выбор наиболее удачной модели для решаемой нами задачи.

Подготовка данных к работе

После того как мы получили представление о том, какие переменные присутствуют в данных, и стали понимать их особенности, переходим к непосредственной работе с данными.

В первую очередь посмотрим, какие переменные содержат пустые значения. Для этого в цикле пройдем по каждому столбцу датасета и осуществим данную проверку. Обнаруживаем, что есть признаки, которые абсолютно неинформативны (например, open_il_6m), поскольку не представляют для наших задач никакой смысловой ценности. Кроме того, почти все такие характеристики содержат много значений типа NaN. С такими переменными работать далее не представляется возможным, поэтому мы можем удалить их. Осталось обработать только признаки, которые содержат небольшое число пропусков. Под небольшим числом пропусков в признаке понимается отсутствие не более 5% данных для данного признака. Для этого мы теперь будем удалять не целые столбцы (поскольку они содержат важную информацию для модели кредитного скоринга), а только выборочно строки, которые содержат пустые значения.

Теперь нам необходимо поработать с категориальными переменными. Всего среди 24 переменных 6 имеют тип Object. Признак term, как мы уже выяснили ранее, представляется числом месяцев, в которые необходимо делать выплату клиенту (36 или 60 месяцев). Заменим каждое строчное значение на соответствующее ему число. Адресные данные (переменные зип-код и непосредственный адрес) считаем необходимым удалить, поскольку эта информация не влияет на способность конкретного заемщика к выплате займа. Переменные grade (рейтинг, присвоенный клиенту на основании его кредитной истории) и sub_grade (так называемый детализированный рейтинг) также являются категориальными. Получаем, что grade – подфункция от sub_grade. Поэтому данную характеристику можно удалить. Последней категориальной переменной является целевая переменная loan_status, представляющая собой статус по кредиту для каждого заемщика. Уникальными значениями данного признака являются «полностью выплачен», «дефолт», «текущая выплата». Мы будем осуществлять предсказание для первого и второго вариантов.

Также на данном этапе обработки данных мы удалили дубликаты строк. Далее с помощью функции train_test_split осуществляем сепарацию данных на тестовую и тренировочную выборки в соотношении 1:3. Теперь обращаем внимание на то, что еще во время исследовательского анализа было обнаружено, что данные содержат выбросные значения. Устранять данную проблему будем с помощью ограничения всех данных 95% квантилем (так получится избежать негативного влияния выбросов). Теперь с помощью MinMaxScalar и fit_transform производим нормализацию получившихся выборок для дальнейшего применения моделей машинного обучения.

Построение моделей

D. Shashi, S.S. Handa и N.P. Singh в своей статье [7], посвященной изучению методологий решения задач кредитного скоринга, сравнивали эффективность 20 различных моделей на наборе данных Германии. Признанная экспертиза ученых по данным и их исследование побудили использовать три «лучших» с точки зрения получения наивысшей точности предсказания метода. Поскольку решаемая нами задача является классификацией, то в качестве метрик качества мы будем использовать Accuracy score, Precision, Recall, f1-score, ROC, AUC.

1. Artificial neural network (ANN)

ANN, или искусственная нейронная сеть, будет обучаться в 20 эпохах, со скоростью обучения 0.001, в качестве функции потерь будет использоваться бинарная кросс-энтропия. В результате обучения мы можем визуализировать AUC-кривую обучения, чтобы отследить динамику качества классификации, представленную на рисунке 1.

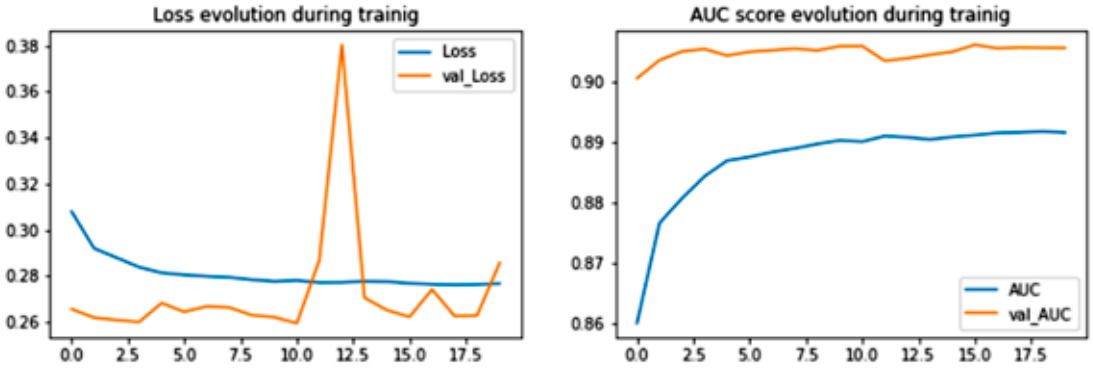


Рис. 1. AUC-кривая обучения для ANN

```
Test Result:
================================================
Accuracy Score: 88.86%

------------------------------------------------
CLASSIFICATION REPORT:
                0.0       1.0  accuracy  macro avg  weighted avg
precision      0.91      0.89      0.89       0.90          0.89
recall         0.48      0.99      0.89       0.73          0.89
f1-score       0.63      0.93      0.89       0.78          0.87
support   25480.00 104943.00      0.89  130423.00     130423.00
------------------------------------------------
Confusion Matrix:
 [[ 12195  13285]
 [  1239 103704]]
```

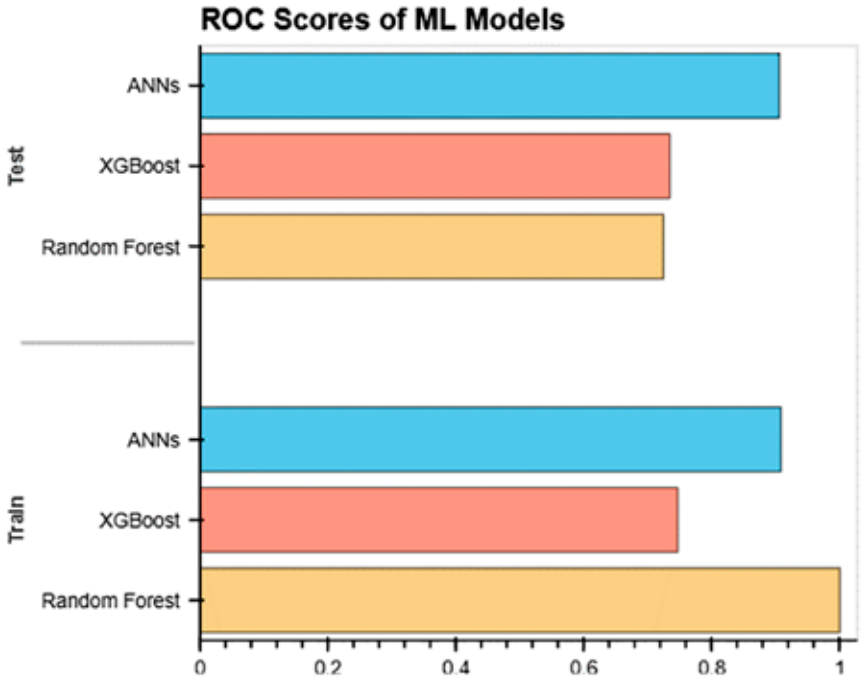
Рис. 2. Результаты точности предсказания модели 1

```
Test Result:
================================================
Accuracy Score: 88.92%

------------------------------------------------
CLASSIFICATION REPORT:
                0.0       1.0  accuracy  macro avg  weighted avg
precision      0.96      0.88      0.89       0.92          0.90
recall         0.45      1.00      0.89       0.72          0.89
f1-score       0.62      0.94      0.89       0.78          0.87
support   25480.00 104943.00      0.89  130423.00     130423.00
------------------------------------------------
Confusion Matrix:
 [[ 11557  13923]
 [   524 104419]]
```

Рис. 3. Результаты точности предсказания модели 2

```
Test Result:
================================================
Accuracy Score: 88.94%

------------------------------------------------
CLASSIFICATION REPORT:
              0.0       1.0  accuracy  macro avg  weighted avg
precision     0.91      0.89      0.89       0.90          0.89
recall        0.48      0.99      0.89       0.73          0.89
f1-score      0.63      0.94      0.89       0.78          0.88
support   25480.00 104943.00      0.89  130423.00     130423.00
------------------------------------------------
Confusion Matrix:
 [[ 12212  13268]
 [  1159 103784]]
```

Рис. 4. Результаты точности предсказания модели 3




```
RANDOM FOREST          roc_auc_score: 0.724
XGBOOST                roc_auc_score: 0.734
ANNS                   roc_auc_score: 0.906
```

Рис. 5. Результаты сопоставления точностей моделей

В среднем показатель AUC во время обучения был чуть выше 0,9, что для обучающейся выборки является довольно высоким показателем. На тестовой же выборке аналогичная метрика на несколько процентных пунктов меньше. Теперь обратим внимание на отчет о классификации (рис. 2).

Accuracy score составил 0,89. В задаче предсказания дефолтных случаев по кредиту Precision – 0,91, Recall – 0,48, f1-score – 0,63. Показатели метрик получились достаточно высокие, что свидетельствует о правильно подобранных параметрах модели и грамотно предобработанных данных.

2. XGBoost классификатор

Следующая модель машинного обучения – XGBoost классификатор. На удивление, метрики качества в данном случае оказались почти полностью идентичны нейронной сети, построенной ранее. Обратим внимание на отчет о классификации (рис. 3).

3. Random Forest классификатор

Модель случайного леса (Random Forest) является также одной из самых мощных моделей в машинном обучении. Возможно, в решении задачи кредитного скоринга получится улучшить полученные ранее результаты. В процессе подбора оптимальных параметров модель пересчитывала результаты. В итоге получилось увеличить показатель Recall на 0,05 при эквивалентных остальных значениях метрик. Однако ROC-score здесь оказался значительно ниже, чем у нейронной сети. Чтобы удостовериться в этом, посмотрим на отчет о классификации модели (рис. 4).

Сравнение метрик качества по моделям

В настоящей статье были построены три модели машинного обучения, способные осуществлять классификацию заемщиков на надежных и сомнительных. Однако необходимо определиться с тем, насколько удовлетворительными являются полученные результаты. Сделать это можно в том числе с помощью сравнения метрик качества моделей на тестовых данных. В задаче кредитного скоринга эффективность построенной модели чаще всего анализируют на основании показателей ROC и AUC. Визуализируем получившиеся результаты для каждой модели, чтобы можно было сделать выводы.

На тестовых данных наилучшим образом показала себя искусственная нейронная сеть (обе метрики превосходят 0,9, что свидетельствует о довольно хорошем качестве модели классификации). При этом XGBoost и Random Forest также показали неплохую эффективность в решении задачи классификации.

Заключение

В данной работе была продемонстрирована реализация 3 методов машинного обучения (искусственная нейронная сеть (ANN), XGBoost классификатор и модель случайного леса) в решении задачи кредитного скоринга, которая, в свою очередь, была сведена к задаче классификации. Для всех трех моделей был произведен оптимальный подбор параметров в целях максимизации эффекта точности. Далее были реализованы обучение моделей, сопоставление результатов и выбор наиболее удачной модели для решаемой нами задачи. В результате мы получили ощутимо высокие результаты для всех выбранных моделей машинного обучения, однако наилучшим образом себя показала искусственная нейронная сеть с показателем ROC = 0,91.

Список литературы

1. Дробышевская М.Н., Кулякова М.Н. Проблемы и перспективы развития современной российской банковской системы // Концепт. 2017. Т. 18. С. 62–65. [Электронный ресурс]. URL: http://e-koncept.ru/2017/770385.htm (дата обращения: 30.12.2022).

2. Moradi S., Mokhatab Rafiei F. A dynamic credit risk assessment model with data mining techniques: evidence from Iranian banks. Financ Innov. 2019. Vol. 5. No. 15. DOI: 10.1186/s40854-019-0121-9.

3. Hutchins J. The US Farm credit system and agricultural development: Evidence from an early expansion, 1920–1940. American Journal of Agricultural Economics. 2023. Vol. 105. No. 1. P. 3-26. DOI: 10.1111/ajae.12290/.

4. Pushkareva L.V., Galochkina O.A., Bezgacheva O.L. Current trends in the banking system of Russia. Espacios. 2019. Vol. 40. No. 4. P. 22-29.

5. Chorzempa M., Triolo P., Sacks S. China's Social Credit System: A Mark of Progress or a Threat to Privacy? Peterson Institute for International Economics: Policy Brief. 2018. No. 18-14. P. 9.

6. Feng S., Xingchao Z., Gang K. Three-stage reject inference learning framework for credit scoring using unsupervised transfer learning and three-way decision theory. Decision Support Systems. 2020. Vol. 137. P. 113366. DOI: 10.1016/j.dss.2020.113366.

7. Shashi D., Handa S.S., Singh N.P. Credit Scoring Using Ensemble of Various Classifiers on Reduced Feature Set. Industrija. 2015. Vol. 43. No. 4. P. 163-174. DOI: 10.5937/industrija43-8211.