УДК 004.89

МАШИННОЕ ОБУЧЕНИЕ В КРЕДИТНОМ СКОРИНГЕ

Киселев В.В.

к.т.н., доцент,

Московский Государственный Технический Университет им. Н.Э. Баумана,

Москва, Россия

Борисов И.Д.

магистрант,

Московский Государственный Технический Университет им. Н.Э. Баумана,

Москва, Россия

Аннотация

В статье рассматривается задача кредитного скоринга. Исследуются разные подходы к решению этой задачи. Используются различные методы машинного обучения. В статье предлагается сравнить модели, и предложен новый подход, позволяющий улучшить существующие решения.

Ключевые слова: кредитный скоринг, машинное обучение, метрика качества GINI, DPD, PD.

MACHINE LEARNING IN CREDIT SCORING

Kiselev V. V.

Ph. D., associate Professor, Bauman

Moscow State Technical University,

Moscow, Russia

Borisov I.D.

Master's student,

Moscow State Technical University. N.E. Bauman,

Moscow, Russia

Annotation

The article discusses the problem of credit scoring. Different approaches to solving this problem are being investigated. Various machine learning methods are used. The article proposes to compare models, and a new approach is proposed to improve existing solutions.

Key words: credit scoring, machine learning, GINI quality metric, DPD, PD.

В данной статье рассматриваются плюсы и минусы основных моделей машинного обучения для задачи кредитного скоринга, а также новый подход, в котором предложено научить модель адаптироваться к изменениям экономической ситуации в стране.

Задача кредитного скоринга является частным случаем задачи бинарной классификации в машинном обучении [1]. Целевой переменной, которую пытается предсказать модель, чаще всего является DPD_n_m (Days past due – количество дней просрочки по кредиту) – уйдет ли клиент в просрочку больше, чем m дней, за n платежей. Особенностью такого выбора является то, что целевая переменная «созревает» довольно долго, например, для dpd_9_90, который используется чаще всего, нужно ждать целый год (9 месяцев и еще 90 дней), чтобы получить данные для обучения модели. Также модели предсказывают не метку класса, а PD (Probability of default) – вероятность дефолта в выбранный временной диапазон. Именно PD получается на выходе модели.

Такое представление результатов модели позволяет расположить клиентов, которые хотят взять кредит, в порядке убывания относительно PD. С помощью полученной таблицы составляю скоркарту – таблица, в которой клиенты банка будут располагаться в зависимости от их надежности. На рис. 1 (рисунок автора) изображен пример такой скоркарты.

Рис. 1 – Пример скоркарты

Аналитики оценивают допустимые для банка риски, после чего выставляется некий порог p_* уверенности модели (на рис. 1) p=0.47 и после него уже не выдает кредит.

У истоков внедрения машинного обучения в кредитный скоринг стояла модель логистической регрессии [2].

Плюсы данной модели:

1) Высокая скорость обучения, поэтому можно использовать для обучения модели выборки большого размера;
2) При большом количестве разреженных признаков не уступают более продвинутым моделям;
3) При использовании полиномиальных признаков можно построить нелинейную модель.

Минусы:

1) Если связь признаков с целевой переменной нелинейная, модель будет плохо работать;
2) Почти никогда не выполняются в реальных задачах предположения теоремы Маркова-Гаусса, из-за этого линейные методы сильно уступают более продвинутым моделям.

Следующий метод машинного обучения, который начал применяться в кредитном скоринге – деревья решений [3]. Если раньше итоговое решение принималось на основании правил, выведенных опытным путем, то данный

алгоритм формирует эти правила, максимизируя прирост информации (information gain, IG). Почти все реализации данного типа моделей обучаются, максимизируя прирост информации на каждом шаге, пока энтропия не дойдет до минимума. Энтропия вычисляется по следующей формуле [4]:

$$E = -\sum_{i=1}^{n} p_i \log_2 p_i ,$$

где i – текущий класс из выборки, n – количество классов в выборке, p_i – это вероятность встретить i-ый класс в выборке.

Следовательно, идеальной классификация будет при энтропии равной 0, ибо в этом случае все выборки в узле принадлежат к одному классу, а также энтропия максимальна, если у нас равномерное распределение классов в этом узле. Но почти всегда необходимо ограничивать эту минимизацию энтропии, чтобы избежать ситуации, когда модель выучила обучающую выборку, в следствии чего переобучилась [5].

Используя энтропию, прирост информации вычисляется по следующей формуле:

$$IG = E - \sum_{i=1}^{n} \frac{|D_i|}{|D|} E_i,$$

где i – текущее разбиение, n – количество разбиений, $|D_i|$ – количество элементов в текущем разбиении, $|D|$ – количество элементов во всей выборке, H_i – энтропия в текущем разбиении [6].

Рассмотрим плюсы деревьев решений:

1) Модель довольно легко интерпретировать;
2) Относительно быстрое обучение, а также получение предсказаний;
3) Быстрый подбор гиперпараметров;
4) Может обрабатывать числовые и категориальные признаки.

Минусы:

1) Так как модель выдает предсказания на основе правил, составленных в

процессе обучения, она становится очень чувствительна к выбросам, а также к данным, которые модель не встречала в процессе обучения, то есть она не умеет экстраполировать;

2) Модель склонна к переобучению, если не ограничивать глубину дерева;
3) Модель плохо обрабатывает пропуски в данных.

Следующая модель, которая на данный момент является лучшей в задаче кредитного скоринга – градиентный бустинг [7]. Суть данной модели заключается в последовательном построении линейной комбинации алгоритмов (чаще всего деревьев решений). Каждый следующий алгоритм старается уменьшить ошибку текущего ансамбля [8].

Плюсы градиентного бустинга:

1) Не зависит от выбора функции потерь;
2) Обрабатывает пропуски в данных.

Минусы:

1) Необходимо ограничивать количество моделей в ансамбле, ибо в ином случае итоговая модель будет переобучена;
2) Требует больших вычислительных мощностей, а также времени на его обучение.

Последний подход, который на данный момент только зарождается в банковском секторе – нейросетевой подход [9]. Такие модели применяются в связке с градиентным бустингом, но обучаются на неструктурированных данных, которые бустинги не могут обрабатывать. Это решение помогает улучшить качество, но занимает очень много времени и вычислительных ресурсов на разработку и обучение модели.

Плюсы нейросетей:

1) Позволяют получить полезную информацию из неструктурированных данных, которые не могут обработать другие модели;

Минусы:

1) Разработка и обучение модели требуют много времени,

вычислительных мощностей и большого количества данных;

2) Не всегда удается значительно повысить качество, добавив ее в модель градиентного бустинга.

На данный момент, это все подходы к решению задачи кредитного скоринга. В данной работе предлагается новая архитектура итоговой модели, которая основана на блендинге – использование предсказаний нескольких моделей для обучения основной модели. Цель данного подхода заключена в следующем: модели, обученные на предсказании целевой переменной dpd_9_90, хорошо работают только в том случае, если экономическая ситуация в стране стабильна. За последние пару лет ситуация в стране и мире очень сильно и быстро изменялась, из-за чего истинное распределение целевой переменной постоянно изменялось, а качество предсказаний моделей значительно ухудшалось. Поэтому предлагается обучать помимо основной (базовой) модели, которая предсказывает dpd_9_90, вспомогательные модели, обученные более «короткой» целевой переменной. Предсказания вспомогательных моделей будем использовать в качестве признаков для базовой модели. Предполагается, что так как целевая переменная у вспомогательных моделей «созревает» раньше, то с их помощью можно корректировать предсказания основной модели в зависимости от изменения экономической ситуации в стране (кризисы, инфляция и т.д.). Также предполагается, что при стабильной ситуации в стране будет высокая корреляция между вспомогательными целевыми переменными и базовой. Поэтому, при неожиданном изменении экономической ситуации в стране, предсказание самой «короткой» модели будет сообщать базовой модели о возможном повышении рисков, при этом более «длинные» вспомогательные модели будут сообщать о текущем состоянии кризиса (ситуация ухудшается, стабилизировалась или улучшается). Также можно добавлять веса каждой из вспомогательных моделей, чтобы была возможность регулировать их влияние на базовую модель самостоятельно. На рисунке 2 (рисунок автора) изображены 6 вспомогательных моделей с более короткими целевыми переменными (1_10,

1_30, 3_30, 4_60, 6_60, 7_90) и основная модель, предсказывающая уже целевую переменную нужной длины (9_90). Светлыми квадратами обозначены признаки, на которых обучается модель, темными – целевая переменная, которую предсказывает модель. Обучать вспомогательные модели можно на таких же признаках, что и базовую модель, а также проводить отдельный отбор признаков для каждой из моделей.

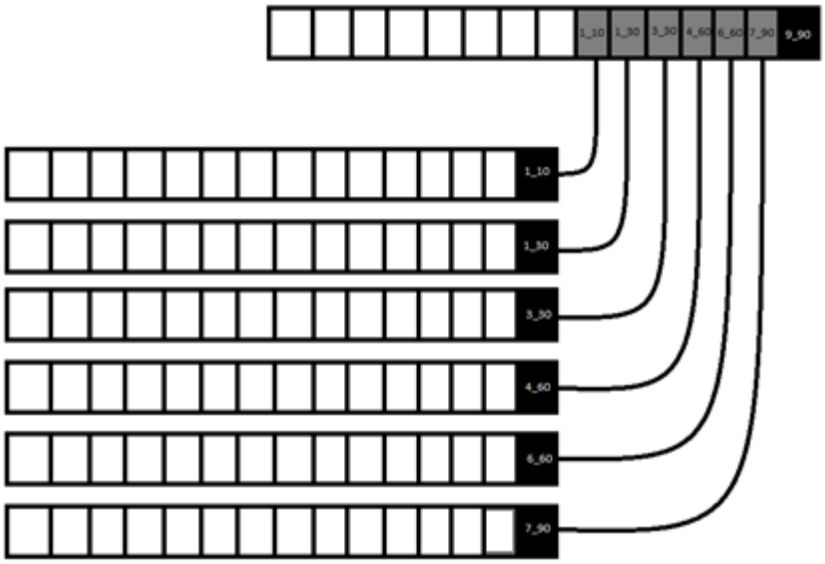

Рис. 2 – схема нового подхода к кредитному скорингу

Далее будут проведены практические исследования для всех предположений, выдвинутых выше.

Библиографический список:

1. Finlay S. Credit scoring, response modeling, and insurance rating: a practical guide to forecasting consumer behavior. – Springer, 2012.

2. Bischl B., Kühn T., Szepannek G. On class imbalance correction for classification algorithms in credit scoring //Operations Research Proceedings 2014: Selected Papers of the Annual International Conference of the German Operations Research Society

(GOR), RWTH Aachen University, Germany, September 2-5, 2014. – Springer International Publishing, 2016. – С. 37-43.

3. Brown I., Mues C. An experimental comparison of classification algorithms for imbalanced credit scoring data sets //Expert Systems with Applications. – 2012. – Т. 39. – №. 3. – С. 3446-3453.

4. Baesens B. et al. Benchmarking state-of-the-art classification algorithms for credit scoring //Journal of the operational research society. – 2003. – Т. 54. – С. 627-635.

5. Baesens B. et al. Benchmarking state-of-the-art classification algorithms for credit scoring //Journal of the operational research society. – 2003. – Т. 54. – С. 627-635.

6. Louzada F., Ara A., Fernandes G. B. Classification methods applied to credit scoring: Systematic review and overall comparison //Surveys in Operations Research and Management Science. – 2016. – Т. 21. – №. 2. – С. 117-134.

7. Friedman J. H. Greedy function approximation: a gradient boosting machine //Annals of statistics. – 2001. – С. 1189-1232.

8. Chen T., Guestrin C. Xgboost: A scalable tree boosting system //Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. – 2016. – С. 785-794.

9. Liaw A. et al. Classification and regression by randomForest //R news. – 2002. – Т. 2. – №. 3. – С. 18-22.

Оригинальность 85%