**УДК**

***Бучацкая В.В.,***

*кандидат технический наук, доцент, ФГБОУ ВО «Адыгейский государственный университет», г. Майкоп, buch_vic@mail.ru*

***Лобанов В.Е.,***

*студент 4 курса, ФГБОУ ВО «Адыгейский государственный университет», г. Майкоп, valery2698@mail.ru*

# АНАЛИЗ АЛГОРИТМОВ КЛАССИФИКАЦИИ

Аннотация. В статье проведен анализ наиболее используемых алгоритмов для решения задачи классификации. Представлены численные характеристики производительности моделей для решения задачи кредитного скоринга.

Ключевые слова: алгоритмы классификации, прогнозирование кредитоспособности, логистическая регрессия, решающие деревья.

***Buchatskaya V.V.***

*Candidate of Technical Sciences, Associate Professor, Adyghe State University, Maikop,*

***Lobanov V.E.,***

*4th year student Adyghe State University, Maikop,*

Abstract. The article analyzes the existing classification algorithms. The comparison of algorithms, based on various metrics, for credit scoring prediction is provided.

Key words: classification algorithms, credit-scoring forecasting, logistic regression, decision trees.

В настоящее время задачи Data-mining набирают большую популярность. К таким задачам относится задача классификации. Ее применение можно найти и в медицине, лингвистике, экономических

задачах, бизнес секторе и информационных технологиях. Условно задача классификации делится по типам классов [1]: двухклассовая, многоклассовая, с непересекающимися классами, с пересекающимися классами и нечеткие классы. В данной работе рассматривается задача двухклассовой классификации. На основании работ [2,5] были рассмотрены следующие наиболее используемые на практике методы классификации. Условно все алгоритмы можно разделить на статистический методы и интеллектуальные.

Так, классическим представителем статических методов являются решающие деревья. Этот классификатор разбивает данные на всё меньшие и меньшие подмножества на основе разных критериев, т. е. у каждого подмножества своя сортирующая категория. С каждым разделением количество объектов определённого критерия уменьшается. Классификация подойдёт к концу, когда сеть дойдёт до подмножества только с одним объектом. Если объединить несколько подобных деревьев решений, то получится так называемый *Случайный Лес* (англ. "*Random Forest"*). К основным достоинствам этого метода можно отнести простоту использования и интерпретации работы, широкая сфера применения, отсутствие строгих требований к данным, не нужна нормализация.

Однако основным недостатком таких моделей всегда являлось долгое время обучения алгоритма. В недавнее время особую популярность набирают алгоритмы, использующие градиентный бустинг [3]. В дальнейшем будем рассматривать градиентный бустинг на решающих деревьях модификации "XGBoost Classifier". Бустинг - альтернативный подход, в котором каждый специалист по подбору персонала основывается на оценке кандидата предыдущим специалистом. Это ускоряет процесс собеседования, так как не подходящие кандидаты сразу же отсеиваются. Градиентный бустинг является частным случаем бустинга, в котором ошибка минимизируется алгоритмом градиентного спуска. То есть, наименее квалифицированные кандидаты отсеиваются как можно раньше. К уже существующим плюсам деревьев

решений, градиентный бустинг добавляет аппаратную оптимизации, что сказывается на скорости работы; возможность отсечения ветвей дерева; встроенные регуляризацию, кросс-валидацию и работу с разреженными данными.

Еще одним выбранным алгоритмом является логистическая регрессия [1]. Логистическая регрессия выводит прогнозы о точках в бинарном масштабе: нулевом и единичном. Если значение чего-либо равно либо больше 0.5, то объект классифицируется в большую сторону (к единице). Если значение меньше 0.5 — в меньшую (к нулю). У каждого признака есть своя метка, равная только 0 или только 1. Логистическая регрессия является частным случаем обобщённой линейной модели регрессии. К главным плюсам данного алгоритма можно отнести его простоту и малые объемы вычислительного времени. Стоит учитывать, что данный алгоритм не решает задачи нелинейного характера.

Среди интеллектуальных методов наиболее известными являются нейронные сети [4]. Основным достоинством данного семейства алгоритмов можно считать масштабируемость, высокую адаптивность, нелинейность моделей и возможность применения во многих задачах. Даже если о данных ничего неизвестно, использование нейронных сетей является одним из лучших вариантов. Однако для решения задачи простой двухклассовой классификации на средней выборке данных, использование нейронных сетей будет неоправданным ввиду ресурсоемкости процесса обучения и непрозрачности алгоритма.

Таким образом для сравнительного анализа были выбраны алгоритмы решающих деревьев, логистическая регрессия и градиентный бустинг. Расчёты производились в среде “Jupyter Notebook”, язык программирования “Python 3.6”.

Для тестирования была выбрана таблица данных с характеристиками клиентов банка за фиксированный период времени. Необходимо по целевой

переменной определить подвержен ли данный клиент дефолту т.е. оценить бинарной классификацией его способность выплатить кредит. Для работы алгоритма нужно выделить свойства, характеристики клиента. По построенной тепловой карте признаков для дальнейшей работы были выделены следующие категории: возраст, образование, доход, трудовой стаж, другие задолжности, соотношения долга к доходу и кредита к долгу.

Предварительно исходные данные были подвергнуты обработке по критериям работы [6]. После осуществления теста на мультиколлинеарность, который показал, что все значения находятся в допустимых пределах, были обучены все алгоритмы. Для логистической регрессии полнота модели составила 54%, точность модели 81%. Алгоритм градиентного бустинга был задан параметрами “random state” = 123, “test size” = 0,2. Точность прогноза составила 77%, полнота модели 52%. Наконец, алгоритм дерева принятия решений строился по 5 складкам кросс-валидации и полученная точность модели составила 59%, точность прогноза 71%. Общие результаты алгоритмов приведены в табл.1.

Как говорилось ранее, задача двухклассовой классификации является наиболее простой из всех существующих. В данной задаче около 1/3 всех строк были размечены заранее, что и составило обучающую выборку для моделей. В качестве оценок для сравнительного анализа были выбраны: точность прогноза, полнота модели, точность модели, F1-критерий. Можно сделать вывод, что лучше всего себя показала логистическая регрессия, которая чаще всего и используется на практике при решении экономических задач классификации. Однако стоит учитывать, что потенциал алгоритма градиентного бустинга не был до конца раскрыт т.к. все параметры модели были выбраны по умолчанию. Для автоматизации этого процесса необходимо будет выполнить несколько итераций для подбора наилучшего значения лямбда, “random state” и “test size”.

Таблица 1.

Сравнительный анализ алгоритмов

| **Прогнозное значение** | **Факторы** | **Логистическая регрессия** | | | | **Решающие деревья** | | | | **XGBoost Classifier** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Точность модели | Полнота | Точность прогноза | F-тест | Точность модели | Полнота | Точность прогноза | F-тест | Точность модели | Полнота | Точность прогноза | F-тест |
| Дефолт клиента | Возраст<br>Образование<br>Доход<br>Трудовой стаж<br>долг/доход<br>кредит/долг<br>Другие долги | 81% | 54% | 70% | 0,7 | 77% | 52% | 64% | 0,6 | 71% | 40% | 59% | 0,44 |

На основании изученных методов классификации и проведенного сравнительного анализа планируется разработать практическую реализацию в виде автоматизированного модуля с гибким и простым интерфейсом взаимодействия.

## БИБЛИОГРАФИЧЕСКИЙ СПИСОК


1. David W. Hosmer, Stanley Lemeshow. Applied Logistic Regression, 2nd ed. New York, Chichester, Wiley. 2002. 392 P. ISBN 0-471-35632-8
2. Hastie, T., Tibshirani, R., Friedman, J. The Elements of Statistical Learning, 2nd edition. — Springer, 2009. — 533 p
3. Santhanam, Ramraj & Uzir, Nishant & Raman, Sunil & Banerjee, Shatadeep. (2017). Experimenting XGBoost Algorithm for Prediction and Classification of Different Datasets
4. Zhang, Peter. (2000). Neural Networks for Classification: A Survey. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on. 30. 451 - 462. 10.1109/5326.897072.

5. Айвазян С. А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д. Прикладная статистика: классификация и снижение размерности. — М.: Финансы и статистика, 1989

6. Горелова, Л.В. Основы прогнозирования систем: учеб. Пособие для инж.-экон. спец. ВУЗов / Л.В. Горелова, Е.Н. Мельникова. – М.: Высш. Шк., 1986. – 276 с.