www.volsu.ru

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ В БЕЗОПАСНОСТИ И ТЕЛЕКОММУНИКАЦИЯХ

DOI: https://doi.org/10.15688/NBIT.jvolsu.2024.3.4

УДК 519.7
ББК 22.18

РЕШЕНИЕ ЗАДАЧИ БИНАРНОЙ КЛАССИФИКАЦИИ С ПОМОЩЬЮ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

Александра Алексовна Даньшина

Студент, кафедра информационной безопасности,
Волгоградский государственный университет
BIT-211_164731@volsu.ru
просп. Университетский, 100, 400062 г. Волгоград, Российская Федерация

Алексей Александрович Бабенко

Кандидат педагогических наук, доцент,
кафедра информационной безопасности,
Волгоградский государственный университет
babenko.aleksey@volsu.ru
просп. Университетский, 100, 400062 г. Волгоград, Российская Федерация

Аннотация. В данном исследовании рассмотрены модели машинного обучения для решения задачи бинарной классификации. Представлен алгоритм обработки набора данных для обучения и тестирования, а также проведен сравнительный анализ предлагаемых моделей, по итогу которого была определена наиболее рациональная модель машинного обучения для выполнения поставленной цели.

Ключевые слова: бинарная классификация, модели машинного обучения, набор данных, распределение данных, корреляция.

В нынешнее время количество различных алгоритмов машинного обучения позволяет сделать выбор в пользу данного средства для решения разноплановых задач. Преимущество этих технологий обуславливается тремя важными факторами для большинства компаний, разрабатывающих системы информационной защиты: низкая стоимость, простота реализации, скорость обучения и предоставления результата.

Для обучения любой системы необходимо сформировать структурированный и обрабо-

танный массив данных – датасет. Классической задачей фильтрации и классификации набора информации является дихотомическая или же бинарная классификация. Определение принадлежности объекта к одному из двух возможных классов – основная цель рассматриваемой задачи [3]. Ее решение находят с помощью различных моделей машинного обучения. Для определения наиболее рациональной из них, сформируем датасет, на основе которого, в последствии, будет происходить обучение (рис. 1) [1].

Необходимо убедиться, что в сформированных столбцах нет пропущенных значений. Данный пул имеет 86 непустых записей по 7 критериям, описания которых представлены в таблице.

Так как в датасете присутствуют три категориальных поля, таких как location, merchant и gender, необходимо перевести их уникальные значения в числовые, используя метод One-hot encoding.

Таким образом получим перечень из следующих преобразований:

a) gender: M -1, F- 0;

b) location: New York – 1, Chicago – 2, Los Angeles – 3, San Francisco – 4;

c) merchant: ABC Corp – 0, XYZ Inc – 1.

Основываясь на преобразованных данных (рис. 2), произведем оценку корреляции между столбцами числовых признаков. Полученная матрица корреляция отображается в

	transaction_id	transaction_amount	location	merchant	age	gender	fraud_label
0	1	1000.0	New York	ABC Corp	35	M	0
1	2	500.0	Chicago	XYZ Inc	45	F	0
2	3	2000.0	Los Angeles	ABC Corp	28	M	1
3	4	1500.0	San Francisco	XYZ Inc	30	F	0
4	5	800.0	Chicago	ABC Corp	50	F	0
...	...	...	...	...	...	...	...
81	82	1500.0	Los Angeles	XYZ Inc	31	M	0
82	83	2800.0	San Francisco	ABC Corp	50	F	1
83	84	1350.0	Chicago	XYZ Inc	28	M	0
84	85	920.0	New York	ABC Corp	47	F	0
85	86	2000.0	Los Angeles	XYZ Inc	36	M	0

86 rows × 7 columns

Рис. 1. Сформированный массив данных

Описание критериев данных

Название критерия	Описание	Тип данных	Описание типа данных
transaction_id	Номер транзакции	Int64	Целочисленный 64х-битный
transaction_amount	Сумма транзакции	Float64	Нецелочисленный 64х-битный
location	Место выполнения транзакции	object	Набор свойств
merchant	Платёжная система	object	Набор свойств
age	Возраст	Int64	Целочисленный 64х-битный
gender	Половая принадлежность	object	Набор свойств
fraud_label	Метка мошенничества	Int64	Целочисленный 64х-битный

	transaction_id	transaction_amount	location	merchant	age	gender	fraud_label
0	1	1000.0	1	0	35	1	0
1	2	500.0	2	1	45	0	0
2	3	2000.0	3	0	28	1	1
3	4	1500.0	4	1	30	0	0
4	5	800.0	2	0	50	0	0
...	...	...	...	...	...	...	...
81	82	1500.0	3	1	31	1	0
82	83	2800.0	4	0	50	0	1
83	84	1350.0	2	1	28	1	0
84	85	920.0	1	0	47	0	0
85	86	2000.0	3	1	36	1	0

86 rows × 7 columns

Рис. 2. Данные в числовом формате

том случае, если значение ее модуля превышает отметку 0,3 (рис. 3).

После предобработки данных датасета и определения корреляционной матрицы, можем отобразить графики распределения данных (рис. 4).

Для дальнейшего определения наилучшей модели машинного обучения, необходимо разбить данные на две группы: тестовые и обучающие. Так как в нашем датасете число записей меньше 1000, то соотношение будет 70:30, где 70 – процентное количество данных для обучения, а 30 – для тестирования [4].

Для сравнительного анализа выберем следующие модели машинного обучения: логическая регрессия, метод опорных векторов, дерево решений, наивный байесовский классификатор, метод *k*-ближайших соседей, XGBoost, градиентный бустинг, метод случайного леса, AdaBoost. Метрикой оценки качества моделей будет выступать F1-Score – гармоническое среднее между точностью и пол-

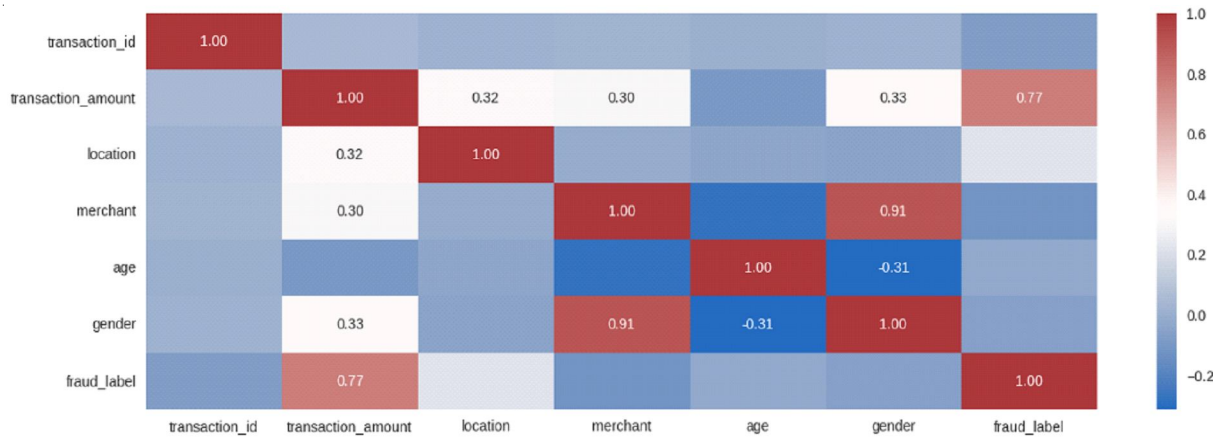

Рис. 3. Матрица корреляции столбцов числового признака

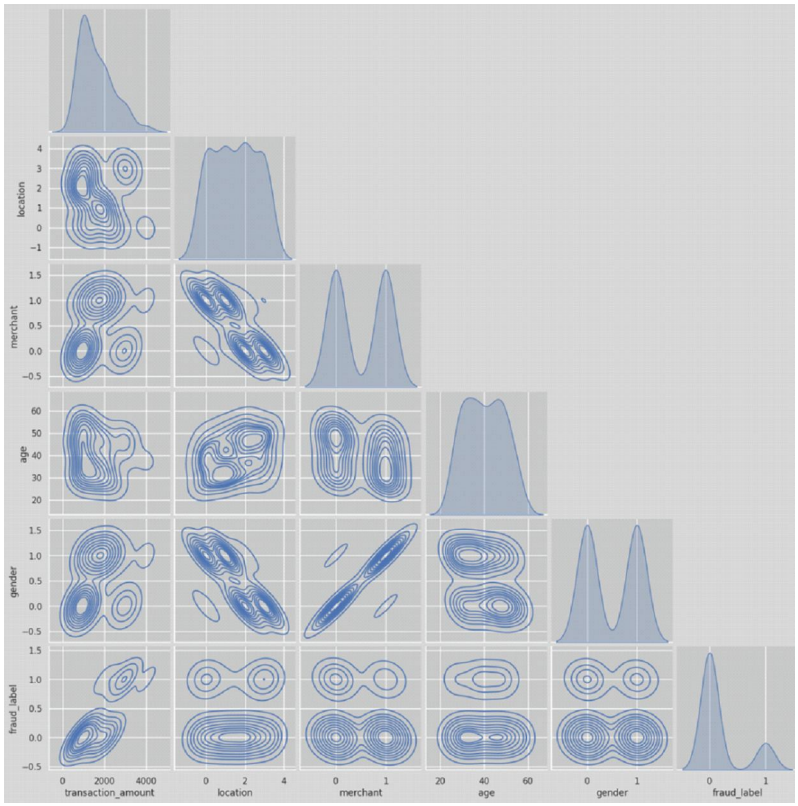

Рис. 4. Графики распределение данных

нотой [4]. Для каждой модели гиперпараметры были определены по умолчанию (рис. 5).

Сравнительный анализ показал, что наилучшей моделью машинного обучения для решения задачи бинарной классификации по метрике F1-Score и настроенных гиперпараметрах по умолчанию является AdaBoost.

Ключевой проблемой данной модели является склонность к переобучению при наличии значительного уровня шума данных [2]. Одним из методов решения является определение наиболее влиятельных гиперпараметров для предсказания алгоритма. Такими являются: максимальная глубина дерева решений {1,2,3,4,5,6} и количество деревьев в ансамбле {1,2,3…2500} [5].

Все рассмотренные модели машинного обучения справляются с поставленной целью – решение задачи бинарной классификации. Однако наиболее рациональной в использовании оказалась модель AdaBoost, применяющаяся в связке со слабым алгоритмом одноуровневых деревьев решений. Высокая эффективность достигается за счет бустинга слабых классификаторов, что также компенсирует такую проблему модели, как переобучение.

Дальнейшим вектором научных исследований является рассмотрение эффективности алгоритма AdaBoost при решении других задач классификации.

СПИСОК ЛИТЕРАТУРЫ

1. %94 Accuracy All Classifiers Fraud Detection // Kaggle. – Electronic text data. – Mode of access: https://www.kaggle.com/code/tayfundogruer/94-accuracy-all-classifiers-fraud-detection

2. Алгоритм AdaBoost // Хабр. – Электрон. текстовые дан. – Режим доступа: https://habr.com/ru/companies/otus/articles/503888

3. Как заставить работать бинарный классификатор чуточку лучше // Хабр. – Электрон. текстовые дан. – Режим доступа: https://habr.com/ru/articles/228963

4. Метрики оценки качества моделей и анализ ошибок в машинном обучении. Подробное руководство // Хабр. – Электрон. текстовые дан. – Режим доступа: https://habr.com/ru/articles/821547

5. Повороты признаков в алгоритме AdaBoost // OSP. – Электрон. текстовые дан. – Режим доступа: https://www.osp.ru/netcat_files/userfiles/Hadoop_TBD_2_2016/Goy_tbd_2.pdf

REFERENCES

1. %94 Accuracy All Classifiers Fraud Detection. *Kaggl*. URL: https://www.kaggle.com/code/tayfundogruer/94-accuracy-all-classifiers-fraud-detection

2. *Algoritm AdaBoost* [AdaBoost Algorithm]. *Khabr* [Habr]. URL: https://habr.com/ru/companies/otus/articles/503888/

3. *Kak zastavit' rabotat' binarnyj klassifikator chutochku luchshe* [How to Make a Binary Classifier Work a Little Better]. *Khabr* [Habr]. URL: https://habr.com/ru/articles/228963/

4. *Metriki ocenki kachestva modelej i analiz oshibok v mashinnom obuchenii. Podrobnoe rukovodstvo* [Model Quality Assessment Metrics and Error Analysis in Machine Learning. A Detailed Guide]. *Khabr* [Habr]. URL: https://habr.com/ru/articles/821547/

5. *Povoroty priznakov v algoritme AdaBoost* [Feature Rotations in the AdaBoost Algorithm]. *OSP*. URL: https://www.osp.ru/netcat_files/userfiles/Hadoop_TBD_2_2016/Goy_tbd_2.pdf

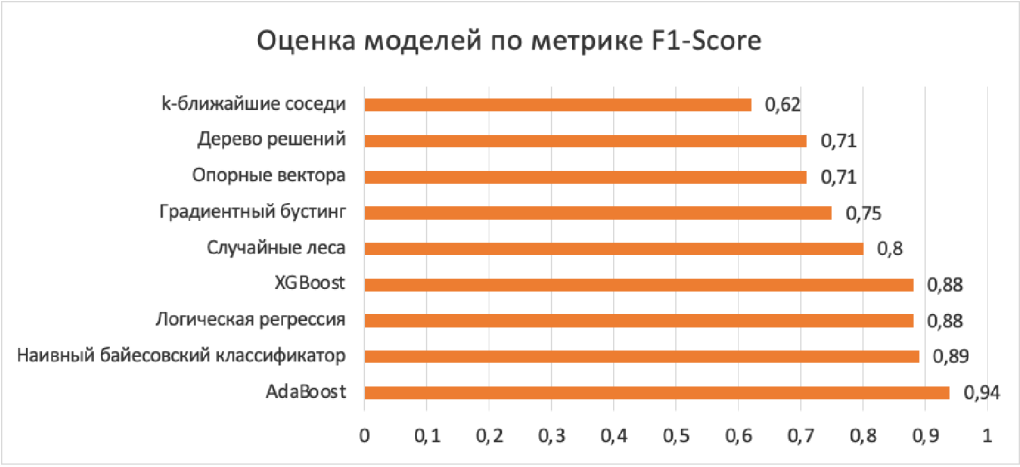

Рис. 5. Оценка моделей по метрике F1-Score

SOLVING BINARY CLASSIFICATION PROBLEM USING MACHINE LEARNING METHODS

Alexandra A. Danshina

Student, Department of Information Security,
Volgograd State University
BIT-211_164731@volsu.ru
Prosp. Universitetsky, 100, 400062 Volgograd, Russian Federation

Aleksey A. Babenko

Candidate of Sciences (Pedagogy), Associate Professor,
Department of Information Security,
Volgograd State University
babenko.aleksey@volsu.ru
Prosp. Universitetsky, 100, 400062 Volgograd, Russian Federation

Abstract. This study discusses machine learning models for solving binary classification problems. An algorithm for processing a data set for training and testing is provided, as well as a comparative analysis of the proposed models, based on the results of which the most rational one for achieving the stated goal was determined. A structured dataset consisting of 86 records on seven criteria was created, with categorical variables such as location, merchant and gender transformed using one-point coding. Correlation analysis was performed to assess the relationships between the numerical features. The dataset was then divided into training (70%) and test (30%) subsets for model evaluation. The different machine learning models were compared using F1-Score metric. All considered machine learning models cope with the objective of solving the binary classification problem. However, the AdaBoost model, used in conjunction with a weak single-level decision tree algorithm, turned out to be the most rational in use. High efficiency is achieved by bousting weak classifiers, which also compensates for such a problem of the model as overtraining.

Key words: binary classification, machine learning models, data set, data distribution, correlation.