Инженерный вестник Дона, №5 (2023)
ivdon.ru/ru/magazine/archive/n5y2023/8394

Машинное обучение прогнозных моделей на несбалансировнных данных по опасным астероидам

Д.В. Горлатов

Санкт-Петербургский государственный архитектурно-строительный университет

Аннотация: Анализируется набор данных по потенциально опасным для Земли астероидам. С помощью моделей машинного обучения, астероиды из базы классифицируются на опасные и неопасные. Используются методы логистической регрессии, k-ближайших соседей; дерева решений и другие. С помощью перекрестной проверки находится наилучший метод, затем определяются его оптимальные гиперпараметры. Качество работы модели-классификатора оценивается по метрикам полноты и ее стандартного отклонения, а также с помощью матрицы ошибок и средней абсолютной ошибки в процентах. Приведены результаты анализа и моделирования в Python, демонстрирующие высокую точность прогнозирования полученной модели.
Ключевые слова: машинное обучение, прогнозная модель, анализ данных, несбалансированные данные, логистическая регрессия, метод k-ближайших соседей, дерево решений, случайный лес, метод опорных векторов, перекрестная проверка.

Введение

Применение машинного обучения и прогнозных нейросетевых моделей является очень удобным и востребованным способом решения задач классификации. Данные методы применяются в самых разных сферах научно-исследовательской деятельности. Например, в [1] изображения разделяются на классы с помощью гибридной нейросети. В [2] содержится обзор применения технологий искусственного интеллекта в космосе, в частности, для предупреждения о ракетном нападении с использованием баллистических ракет, что является близким к задачам, решаемым в данной статье. В [3], как и в настоящей работе, применяется бинарная классификация и настройка гиперпараметров модели, но в сфере защиты конфиденциальности данных. Приведенные примеры доказывают актуальность применяемых здесь методов.

Проблематика и актуальность

Околоземные объекты или объекты, сближающиеся с Землей (ОСЗ) (Near-Earth object (NEO)) – это любые небольшие тела Солнечной системы, орбиты которых иногда приближают их к Земле. Небесные тела классифицируют, как ОСЗ, если их перигелий (ближайшая точка орбиты небесного тела к Солнцу) составляет от 0,983 до 1,3 астрономических единиц (a.e.) (astronomical unit (AU)) [4].

В основном, ОСЗ – это астероиды и кометы, но также к ним могут относиться автономные межпланетные станции (АМС) и другие объекты искусственного происхождения. По данным от 1 декабря 2022 года [5], число обнаруженных астероидов составляет 30 789 из общего количества зафиксированных ОСЗ – 30 907, т. е. примерно 99,6 %.

Если ОСЗ приближается к орбите Земли на расстояние менее 0,05 a.e. (примерно 19,5 расстояния от Земли до Луны) и его диаметр превышает 100 – 150 метров [6], он считается потенциально опасным астрономическим объектом (ПОАО) (potentially hazardous object (PHO)).

ПОАО, в свою очередь, подразделяются на:

- потенциально опасные астероиды (ПОА) (potentially hazardous asteroids (PHA));
- потенциально опасные кометы (potentially hazardous comets (PHC)).

Наибольшую долю ПОАО занимают ПОА – примерно 94 % по данным NASA на октябрь 2008 года. Астероиды размером более 35 метров в диаметре могут представлять угрозу для города [7].

Исследование имеющихся данных по ОСЗ, оценка и прогнозирование их опасности являются актуальными задачами, поскольку:

1. На данный момент NASA разработала и испытала проект Double Asteroid Redirection Test (DART) по изменению направления движения астероидов [8]. Т. е. существует возможность предотвращения опасности.

2. Опасность астероида зависит от многих параметров. Астероид, помеченный как опасный, может не представлять опасности на текущий момент и находиться за пределами зоны функционирования DART. Нужно определить время и вероятность того, что DART придется задействовать.

3. Астероид может быть помечен, как неопасный, ошибочно, или может стать опасным в будущем, ввиду, например, изменения свой орбиты. Требуются модели, прогнозирующие потенциальную опасность астероидов и ранжирующие их по признакам опасности более детально.

Анализ данных

Анализ данных и машинное обучение моделей-классификаторов проводилось на основе набора данных, сгенерированных по результатам наблюдений NASA [9], полученных в период с 1 января 2001 по 14 июня 2022. В рассматриваемой базе содержатся данные только по всем астероидам – потенциально опасным и неопасным.

Исследуемые данные налагают 2 ограничения, которые являются несущественными для решения поставленной задачи, но которые необходимо озвучить:

1. Данные не содержат информации об искусственных объектах на земной орбите и на низкой околоземной орбите (НОО) (Low Earth Orbit (LEO)) [10]. Несмотря на свою многочисленность, эти объекты обладают слишком малыми размерами и не могут рассматриваться как потенциально опасные.

2. Данные не содержат информации о кометах по причине их малой доли в ОСЗ – менее 1 % [5].

В таблице №1 приведена детализация параметров набора данных по столбцам с пояснениями.

Таблица № 1

Детализация данных по столбцам

Название столбца	**Пояснение**
id	уникальный идентификатор астероида
name	название астероида, данное Национальным управлением по аэронавтике и исследованию космического пространства (NASA)
est_diameter_min	минимальный предполагаемый диаметр в километрах
est_diameter_max	максимальный предполагаемый диаметр в километрах
relative_velocity	скорость относительно Земли в км/ч
miss_distance	расстояние до объекта в километрах
orbiting_body	планета, вокруг которой вращается астероид
sentry_object	отметка о вхождении в базу автоматизированной системы контроля столкновений «Часовой» (Sentry)
absolute_magnitude	абсолютная звёздная величина (блеск, свечение)
hazardous	логическая переменная, показывающая, опасен астероид или нет (True – опасен; False – нет)

В таблице №2 содержится статистическая информация по набору данных.

В соответствии с результатами статистического анализа, приведенными в таблице №2, можно сделать следующие выводы:

1. Незаполненные ячейки отсутствуют. Об этом свидетельствует одинаковое число значений, содержащихся в столбцах – 90 836. Т. е. предобработка удалением лишних строк не требуется.

2. Столбец **orbiting_body** (планета, вокруг которой вращается астероид) содержит единственное (уникальное) значение во всех строках – Earth (Земля). Поскольку все объекты из набора данных вращаются вокруг интересующего нас объекта – Земли, это значит, что все они должны подвергнуться более пристальному изучению. Сам столбец можно удалить, так как для дальнейшего анализа он уже не важен.

3. Столбец **sentry_object** (отметка о вхождении в базу автоматизированной системы контроля столкновений «Часовой» (Sentry))

содержит единственное (уникальное) значение во всех строках – False (ЛОЖЬ). Т. е. ни один из представленных объектов не входит в базу «Часовой», а, значит, все они интересны для исследования и возможного дальнейшего включения в данную базу. Сам столбец также можно удалить.

4. Столбцы **est_diameter_min** и **est_diameter_max** содержат соответственно минимальный и максимальный предполагаемые диаметры объекта в км. Можно использовать среднее арифметическое значение диаметра.

Таблица № 2

Описательная статистика

Столбец	Число заполненных ячеек в столбце	Уникальные значения	Наиболее часто встречающееся значение	Количество наиболее часто встречающихся значений (из предыдущего столбца)	Среднее значение
id	90836	NaN*	NaN	NaN	14382878,05
name		27423	469219 Kamo`oalewa (2016 HO3)	43	NaN
est_diameter_min		NaN	NaN	NaN	0,127432106
est_diameter_max		NaN	NaN	NaN	0,284946852
relative_velocity		NaN	NaN	NaN	48066,91892
miss_distance		NaN	NaN	NaN	37066546,03
orbiting_body		1	Earth	90836	NaN
sentry_object		1	ЛОЖЬ	90836	NaN
absolute_magnitude		NaN	NaN	NaN	23,52710347
hazardous		2	ЛОЖЬ	81996	NaN

* NaN – Not-a-Number.

5. Столбец **name** (название астероида, данное NASA) содержит 27 423 уникальных имени, т. е. число дубликатов 90 836 – 27 423 = 63 413. Дубликаты содержат данные по наблюдениям в различные моменты времени с разными результатами по параметрам **relative_velocity** (скорость относительно Земли в км/ч); **miss_distance** (расстояние до объекта в километрах); **est_diameter_min** (минимальный предполагаемый диаметр в километрах); **est_diameter_max** (максимальный предполагаемый диаметр в километрах); **absolute_magnitude** (абсолютная звёздная величина (блеск, свечение)). Такой вывод можно сделать, исходя из отсутствия уникальных значений в этих столбцах таблицы №2. Также число отметок об опасности / безопасности (**hazardous**) не совпадает с числом уникальных ПОА (**name**), следовательно те же самые объекты в разные моменты наблюдения могут оказаться как опасными, так и безопасными.

6. Данные в столбце **hazardous** несбалансированы, так как содержат 81996 пометок о неопасных объектах со значением False (ЛОЖЬ) из 90836. Т. е. количество наблюдений, когда объекты были отмечены как опасные, всего около 10 % от общего числа наблюдений. Поскольку основная задача в работе состоит в обучении моделей различать эти 2 типа объектов, требуется предобработка выборок, чтобы распределение опасных объектов было более равномерным.

Далее нужно проанализировать, какие параметры коррелируют друг с другом и с опасностью (значением True в столбце **hazardous**). Для этого используем тепловую карту корреляции (рис. 1).

Самую высокую корреляцию с параметром опасности / безопасности (**hazardous**) показали:

- абсолютная звёздная величина (**absolute_magnitude**): −0,37 (отрицательная корреляция);
- скорость относительно Земли в км/ч (**relative_velocity**): 0,19;

- средний предполагаемый диаметр в километрах (**est_diameter_av**): 0,18.

Этим параметрам нужно уделить повышенное внимание в дальнейшем.

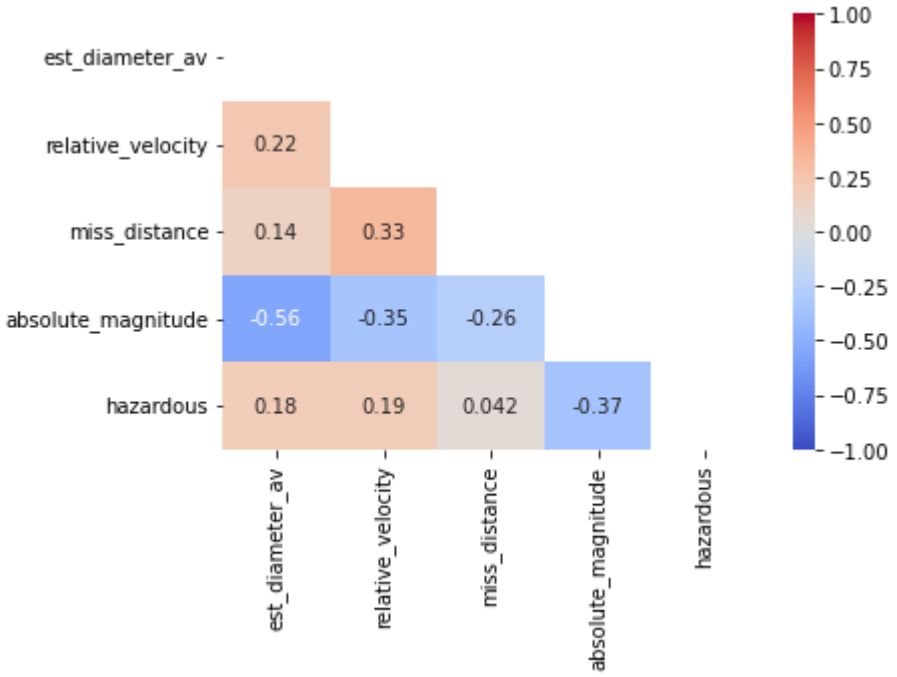

Рис. 1. – Тепловая карта корреляции

Предобработка данных, подготовка к моделированию

Данные не требуют очистки, так как не выявлено пропусков, противоречий и экстремальных значений. Дубликаты по столбцу **name** решено оставить, как могущие представлять интерес для прогнозирования опасности.

По результатам предварительного анализа решено:

1) исключить столбцы: **id**, **orbiting_body**, **sentry_object**;

2) заменить столбцы **est_diameter_min** и **est_diameter_max** на столбец **est_diameter_av** со средним арифметическим значением диаметра.

Произведем группировку дубликатов по уникальным значениям столбца **name**. Добавим столбец **prev_hazardous**, в котором будем помечать, какой объект (**name**) содержит более одной отметки об опасности в столбце

hazardous. Будем считать такие ПОА «превалирующими» по опасности и ставить им пометку True в столбце **prev_hazardous**.

Для определения корреляции между ОСЗ, чаще помечаемыми как опасные (**prev_hazardous**) и опасными (**hazardous**) используем корреляцию Спирмена, как более универсальную. Результат корреляции – 0,97168 близок к единице. Теперь остается выбрать один из трех параметров с наибольшей корреляцией для обучения моделей. Наиболее наглядно это можно сделать с помощью визуализации распределения значений соответствующих параметров по опасности / безопасности (рис. 2 – 4).

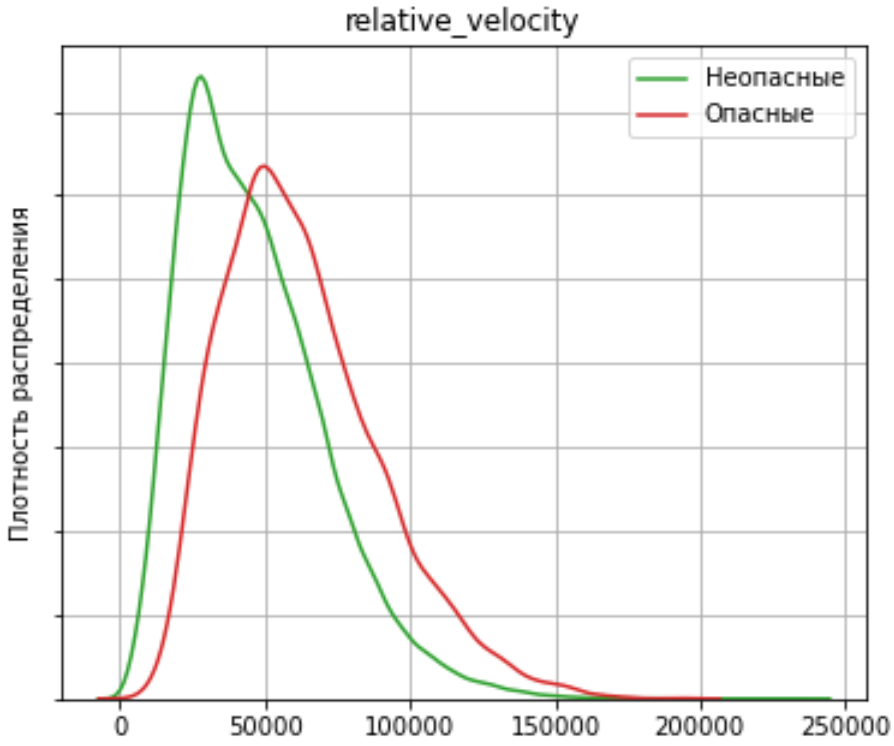

Рис. 2. – Плотность распределения значений скорости относительно Земли в км/ч, сгруппированных по опасности / безопасности

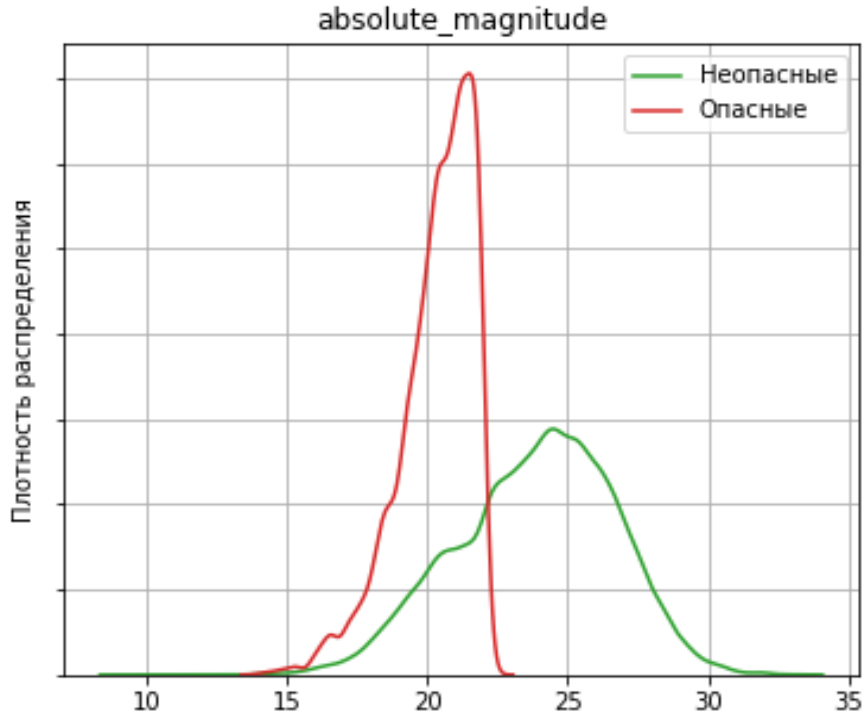

Рис. 3. – Плотность распределения значений абсолютной звёздной величины (блеска, свечения), сгруппированных по опасности / безопасности

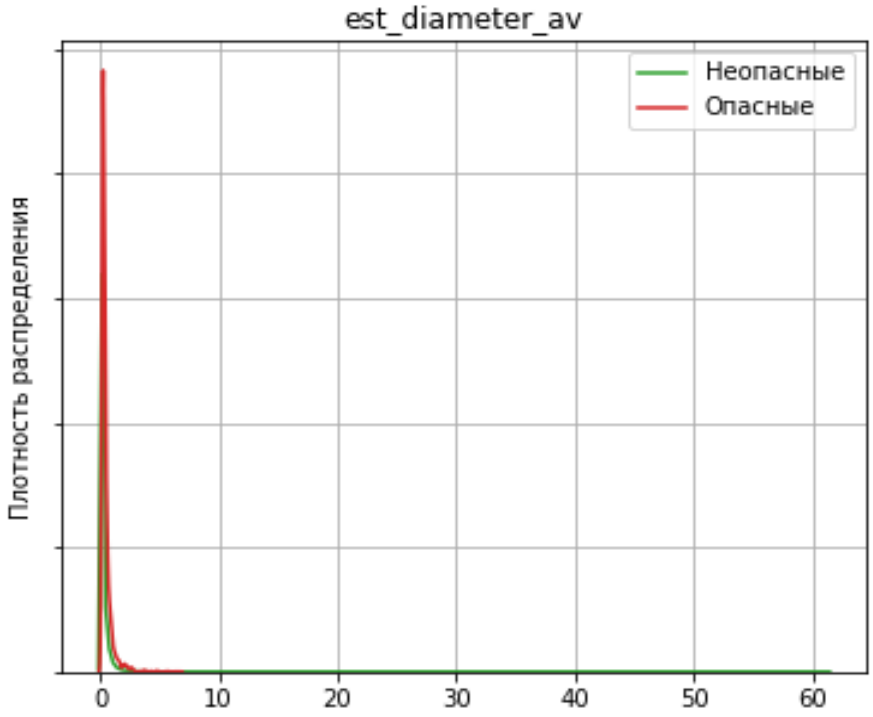

Рис. 4. – Плотность распределения значений среднего предполагаемого диаметра в километрах, сгруппированных по опасности / безопасности

Из рис. 3 – 5 можно понять, что для классификации по опасности / безопасности наилучшим образом подходит параметр абсолютной звёздной величины (блеска, свечения) (**absolute_magnitude**). Для обучения модели будем использовать именно его. В метеорной астрономии принято альтернативное определение абсолютной звёздной величины: абсолютной называется та звёздная величина метеора M, которую он имел бы, наблюдаясь в зените на расстоянии 100 км [11].

$$M = m - 5 \lg R - K,$$

где K – поправка на поглощение в атмосфере (редукция к зениту), R – расстояние до метеора, m – его видимая звёздная величина.

Для получения дополнительной информации о характеристиках модели будем использовать матрицу ошибок (confusion matrix). Она поможет визуализировать ошибки модели при различении двух классов. Это матрица размерности 2×2. Названия строк представляют собой эталонные метки, а названия столбцов – предсказанные (таблица №3).

Таблица № 3

Пример матрицы ошибок

	Прогнозируемые отрицательные цели	Прогнозируемые положительные цели
Фактические отрицательные цели	**TN** (True Negative)	**FP** (False Positive)
Фактические положительные цели	**FN** (False Negative)	**TP** (True Positive)

Расшифровка сокращений в таблице №3:

- **TP** – число верно спрогнозированных положительных целей;
- **TN** – число верно спрогнозированных отрицательных целей;
- **FP** – число фактически отрицательных целей, которые были спрогнозированы как положительные;

- **FN** – Число фактически положительных целей, которые были спрогнозированы как отрицательные.

По причине несбалансированности данных и приоритета на поиск именно опасных объектов для оценки качества полученных моделей используем метрику полноты (Recall):

$$Recall = TP / (TP + FN).$$

Для оценки стабильности показателя Recall для разных методов определяем также его стандартное отклонение. Чем меньше стандартное отклонение Recall метода, тем более стабильную Recall он дает.

Для оценки точности прогноза применим среднюю абсолютную ошибку в процентах MAPE (Mean Absolute Percentage Error):

$$MAPE = (1/n) \times \Sigma(|факт – прогноз| / |факт|) \times 100,$$

где n – размер выборки.

Машинное обучение. Сравнение эффективности методов

Для задач классификации, подобных представленным в этой работе, лучше всего подходят такие методы, как:

- логистическая регрессия;
- метод k-ближайших соседей;
- дерево решений;
- метод опорных векторов (Support vector machine (SVM));
- случайный лес и др.

Сравним эффективность работы этих методов с помощью стратифицированной перекрестной проверки (Stratified K-Fold Cross Validation). По причине несбалансированности данных будем осуществлять их дополнительную подготовку масштабированием с помощью инструмента Pipeline из библиотеки sklearn.

Результаты перекрестной проверки методов с градацией от лучшего к худшему сверху-вниз по метрике Recall представлены в таблице №4.

Таблица № 4

Перекрестная проверка методов

№	Метод	Recall, %	Стандартное отклонение метрики Recall, %
1	Дерево решений	96,222	0,615
2	Случайный лес	95,950	0,570
3	K-ближайших соседей	95,486	0,533
4	Метод опорных векторов	94,932	0,466
5	Логистическая регрессия	94,921	0,467

Методы показали очень хорошие результаты с метрикой Recall > 94 % и стандартным отклонением метрики std(Recall) < 0,7 %. Любой их них применим для данной выборки. Однако лучшее значение метрики Recall продемонстрировал метод «дерево решений», поэтому для дальнейшего исследования используем его.

Создадим функцию настройки гиперпараметров метода, с помощью которой будем настраивать 3 параметра:

1. **Метрика tree criterion**. Будем выбирать из 2-х вариантов: неопределенность Джини (gini); энтропия Шеннона (entropy).

2. **Максимальная глубина (tree_max_depth)**. Этот параметр определяет максимальную глубину дерева. По умолчанию устанавливается значение **None**, что часто приводит к переопределенным деревьям решений. Параметр глубины является одним из способов, которыми можно упорядочить дерево или ограничить его рост. Будем выбирать из 3-х вариантов: None, 5, 10.

3. **Минимальное количество выборок (min_samples_split)**, необходимое на внутреннем узле для разделения. Например, если min_samples_split = 5 и в узле решения есть 8 выборок, тогда разделение

разрешено, в противном случае, если min_samples_split < 5, то не разрешено. Будем выбирать из 3-х вариантов: 2, 5, 10.

По итогам работы функции решено выбрать следующие оптимальные гиперпараметры:

- метрика tree criterion – неопределенность Джини (gini);
- максимальная глубина (tree_max_depth) – None;
- минимальное количество выборок (min_samples_split) – 2.

Настроенная таким образом модель дает метрику Recall, равную уже 96,335 %, что лучше предыдущего результата, полученного при перекрестной проверке.

Оценка веса параметра в модели «дерево решений» (рис. 5) показала преобладающую значимость введенного нами параметра **prev_hazardous**.

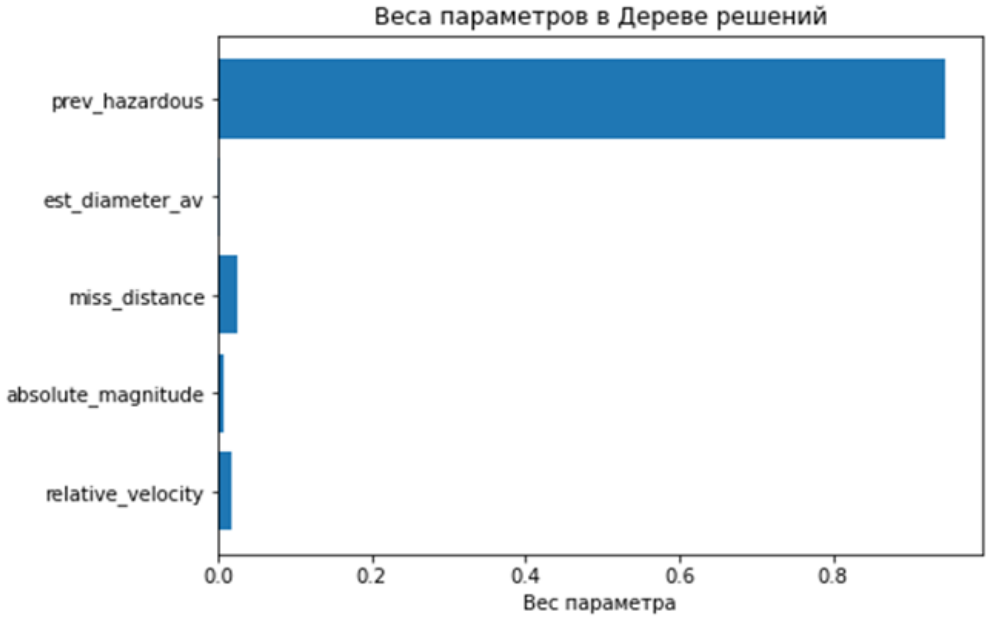

Рис. 5. – Веса параметров из датасета в дереве решений

Матрица ошибок для наилучшего из использованных методов классификации – «дерева решений» представлена в виде таблицы № 5.

Таблица № 5

Матрица ошибок прогнозной модели «дерева решений»

Прогноз / Факт	**Неопасные ОСЗ = 20522**	**Опасные ОСЗ = 2187**
Неопасные ОСЗ = 20499	**TN** = 20441	**FP** = 58 (Ошибка 2-го рода)
Опасные ОСЗ = 2210	**FN** = 81 (Ошибка 1-го рода)	**TP** = 2129

Точность прогноза можно оценить по средней абсолютной ошибке в процентах MAPE. В нашем случае MAPE = 0,576 % <1 %, что является отличным результатом.

Заключение

В работе был проанализирован и подготовлен к моделированию с помощью методов машинного обучения набор данных, содержащий сведения по потенциально опасным астероидам (ПОА). Использование в методике стратифицированной перекрестной проверки позволяет применять ее для несбалансированных данных, как в случае изучаемого датасета. Результаты работы программы показали эффективность работы всех методов, взятых к рассмотрению.

Разработанная программа, осуществляющая сравнение моделей-классификаторов, является универсальной и позволяет легко добавлять в перекрестную проверку новые модели, например, нейросетевые, что планируется реализовать в дальнейших исследованиях. Также, содержащиеся в программном коде пользовательские функции, можно успешно применять для более сложных задач классификации на других наборах данных.

Литература

1. Гушанский С.М., Буглов В.Е. Разработка гибридной нейросети для классификации изображений // Инженерный вестник Дона, 2023, №1. URL: ivdon.ru/ru/magazine/archive/n1y2023/8150.

2. Судьина Д.О., Петросян Л.Э., Зырянова С.А. Применение российских технологий с элементами искусственного интеллекта в космосе // Инженерный вестник Дона, 2023, №1. URL: ivdon.ru/ru/magazine/archive/n1y2023/8148.

3. Шадский В.В. Методика выбора конфигурируемых гиперпараметров интеллектуального классификатора неструктурируемых текстовых данных по степени конфиденциальности на основе метода анализа иерархий // Инженерный вестник Дона, 2023, №4. URL: ivdon.ru/ru/magazine/archive/n4y2023/8357.

4. International Astronomical Union, ed. (31 August 2012), RESOLUTION B2 on the re-definition of the astronomical unit of length, RESOLUTION B2, Beijing, China: International Astronomical Union // URL: iau.org/static/resolutions/IAU2012_English.pdf (дата обращения: 03.12.2022).

5. Discovery Statistics URL: translated.turbopages.org/proxy_u/en-ru.ru.817bc556-638ba4d4-93b19c3f-74722d776562//cneos.jpl.nasa.gov/stats/totals.html (дата обращения: 03.12.2022).

6. Task Force on potentially hazardous Near Earth Objects. Report of the Task Force on potentially hazardous Near Earth Objects (англ.): Journal. – 2000 // URL: web.archive.org/web/20161210142717///nss.org/resources/library/planetarydefense/2000-ReportOfTheTaskForceOnPotentiallyHazardousNearEarthObjects-UK.pdf (дата обращения: 03.12.2022).

7. Уилл Фергюсон (22 января 2013 г.). «Охотник за астероидами сообщает обновленную информацию об угрозе сближающихся с Землей объектов». Scientific American // URL: blogs.scientificamerican.com/observations/asteroid-

hunter-gives-an-update-on-the-threat-of-near-earth-objects/ (дата обращения: 03.12.2022).

8. NASA Confirms DART Mission Impact Changed Asteroid’s Motion in Space NASA // URL: nasa.gov/press-release/nasa-confirms-dart-mission-impact-changed-asteroid-s-motion-in-space (дата обращения: 03.12.2022).

9. NASA Open APIs URL: api.nasa.gov/ (дата обращения: 03.12.2022).

10. ARES _ Orbital Debris Program Office _ Photo Gallery URL: orbitaldebris. jsc.nasa.gov/photo-gallery/ (дата обращения: 03.12.2022).

11. Мартынов Д. Я. Курс общей астрофизики. – М.: Наука, 1979. – С. 591.

References

1. Gushanskij S.M., Buglov V.E. Inzhenernyj vestnik Dona, 2023, №1. URL: ivdon.ru/ru/magazine/archive/n1y2023/8150.

2. Sud'ina D.O., Petrosjan L.Je., Zyrjanova S.A. Inzhenernyj vestnik Dona, 2023, №1. URL: ivdon.ru/ru/magazine/archive/n1y2023/8148.

3. Shadskij V.V. Inzhenernyj vestnik Dona, 2023, №4. URL: ivdon.ru/ru/magazine/archive/n4y2023/8357.

4. International Astronomical Union, ed. (31 August 2012), Resolution B2 on the re-definition of the astronomical unit of length, Resolution B2, Beijing, China: International Astronomical Union. URL: iau.org/static/resolutions/IAU2012_English.pdf (data obrashhenija: 03.12.2022).

5. Discovery Statistics URL: translated.turbopages.org/proxy_u/en-ru.ru.817bc556-638ba4d4-93b19c3f-74722d776562/https/cneos.jpl.nasa.gov/stats/totals.html (data obrashhenija: 03.12.2022).

6. Task Force on potentially hazardous Near Earth Objects. Report of the Task Force on potentially hazardous Near Earth Objects (angl.): journal. 2000. URL: https://web.archive.org/web/20161210142717/http://nss.org/resources/library/planetarydefense/2000-

ReportOfTheTaskForceOnPotentiallyHazardousNearEarthObjects-UK.pdf (date accessed 03.12.2022).

7. Uill Fergjuson (22 janvarja 2013 g.). “Ohotnik za asteroidami soobshhaet obnovlennuju informaciju ob ugroze sblizhajushhihsja s Zemlej ob#ektov”. Scientific American. URL: blogs.scientificamerican.com/observations/asteroid-hunter-gives-an-update-on-the-threat-of-near-earth-objects/ (date accessed 03.12.2022).

8. NASA Confirms DART Mission Impact Changed Asteroid’s Motion in Space NASA. URL: nasa.gov/press-release/nasa-confirms-dart-mission-impact-changed-asteroid-s-motion-in-space (date accessed 03.12.2022).

9. NASA Open APIs URL: api.nasa.gov/ (date accessed 03.12.2022).

10. ARES _ Orbital Debris Program Office _ Photo Gallery URL: orbitaldebris. jsc.nasa.gov/photo-gallery/ (date accessed 03.12.2022).

11. Martynov D. Ja. Kurs obshhej astrofiziki [General Astrophysics Course]. M.: Nauka, 1979. P. 591.