УДК 004.89:631.44

Применение методов машинного обучения для определения категории грунта

Буткина А.А., Гущина О.А., Коржов А.С., Шамаев А.В.

Национальный исследовательский Мордовский государственный университет

Аннотация: В статье описывается применение методов машинного обучения для определения категории грунта. Был выполнен анализ нескольких мето-дов предсказательной аналитики, и в качестве лучшего был выбран метод случайного леса.

Ключевые слова: классификация грунтов, категории грунта, предсказательная аналитика, машинное обучение, случайный лес.

Грунты, представляющие собой комплекс природных материалов, различаются по своему происхождению, химическому составу, текстуре, физическим свойствам и другим характеристикам. Важность определения категорий грунта в инженерной сфере невозможно переоценить, поскольку геологические и геотехнические особенности грунтов напрямую влияют на проектирование и строительство различных объектов. Целью данной работы является разработка программного обеспечения (ПО), которое реализует методы предсказательной аналитики, применяемые для повышения точности прогнозирования и уменьшения вероятности получения ошибочного прогноза категории грунта.

Актуальность работы заключается в том, что применение существующих методов анализа данных о грунтах зачастую требует значительных временных и людских ресурсов, а также может быть подвержено возникновению ошибок, связанных с человеческим фактором. Разработанное ПО позволяет автоматизировать процесса анализа данных, что значительно повышает его объективность, точность и эффективность.

В качестве исходных данных в работе были использованы результаты анализа проб грунта, собранных на территории Московской области. Они были представленные в формате CSV-файла, который содержит информацию о координатах залегания, химическом составе, физических свойствах и других параметрах грунта, необходимых для проведения оценки качества почвы и принятия решений по её использованию.

В качестве основных инструментов разработки использовались интерактивная среда `Jupyter Lab` и язык программирования `Python`, предоставляющий большое количество библиотек для эффективной реализации алгоритмов машинного обучения, визуализации данных и построения графических программных интерфейсов. В частности, в процессе разработки использовались следующие библиотеки:

· `pandas` – предоставляет удобные инструменты для работы с данными, включая чтение и запись данных, манипуляции с таблицами, фильтрацию и агрегацию данных;

· `Numpy` – для выполнения математических операций и работы с массивами данных;

· `Scikit-learn (sklearn)` – предоставляет реализацию различных алгоритмов машинного обучения, включая используемые при анализе: `Random Forest Classifier`

(RFC), Naive Bayes, Multi-Layer Perceptron (MLP) и Decision Tree;

· Matplotlib и Seaborn – предоставляют различные инструменты для построения стати-стических графиков и диаграмм для анализа и визуализации данных;

· Imblearn – позволяет выполнять работы с дисбалансом классов;

· Tkinter – позволяет создавать графический интерфейс пользователя (GUI).

Исходный CSV-файл был импортирован с помощью библиотеки pandas и ее метода read_csv в среду Jupyter Lab и преобразован в DataFrame. После этого выполнена стандартная последовательность действий по обработке данных с помощью моделей машинного обучения, которая включает подготовку данных, корректную обработку пропущенных значений, выбор метрики качества, удаление аномалий из выборки, удаление дублирующей информации и устранение дисбаланса классов, обучение модели и аналитику результатов. Далее опишем наиболее значимые этапы обработки данных из числа вышеупомянутых.

Сбор и анализ данных являются одним из ключевых этапов научной и производственной деятельности. Однако любая выборка данных может содержать аномальные значения, известные также как выбросы, которые могут искажать результаты анализа. Необходимость обнаружения таких аномалий является критически важной для обеспечения точности и достоверности выводов, основанных на результатах анализа. Для обнаружения выбросов в данной работе были использованы визуальные методы, такие как:

· диаграммы разброса (scatter plots), которые позволяют визуально выявить выбросы, отображая точки данных на графике,

· диаграммы размаха (box plots), которые визуально показывают медиану, квартили и размах данных, что позволяет выявлять выбросы.

Это делает возможным применение методов машинного обучения, такие как:

· алгоритмы кластеризации, которые позволяют выявлять аномалии на основе груп-пировки данных в кластеры и определять точки, значительно отличающиеся от остальных (DBScan, AgglomerativeClustering);

· алгоритмы обнаружения аномалий, которые используются для обучения моделей на нормальных данных и выявления аномальных значений (Isolation Forest).

Использование перечисленных методов позволило выполнить глубокий анализ данных о грунтах, выявить и исключить из выборки большинство выбросов. Как показало проведенное в работе исследование, наиболее подходящим методом для обнаружения аномалий является алгоритм DBSCAN, благодаря его способности эффективно идентифицировать плотные кластеры данных и отделять аномалии.

Рассмотрим примеры метрик качества, используемых для оценки моделей машинного обучения в задачах многоклассовой классификации, выбор которых играет важную роль при оценке производительности этих моделей:

· точность (Precision, P) – это метрика, которая отражает долю истинно положительных классификаций относительно всех положительных классификаций;

· полнота (Recall, R) – это метрика, которая отражает долю истинно положительных классификаций относительно всех действительно положительных примеров.

· F1-мера (F1-Score) – это гармоническое среднее между точностью и полнотой, которое позволяет оценить баланс между перечисленными выше метриками. Формула F1-меры:

$$F_1 = 2\frac{P \cdot R}{P + R} \tag{1}$$

Метрика `F1-score` была выбрана в качестве основной для оценки моделей классификации. Данный выбор обусловлен тем, что она учитывает как ложные положительные, так и ложные отрицательные классификации. Это особенно важно в случаях, когда баланс между точностью и полнотой критически важен, что делает `F1-score` наиболее подходящей метрикой для текущей задачи. Следует отметить, что данная метрика использовалась в качестве основной при решении схожих задач зарубежными авторами [1].

На следующем этапе выполнялось исключение дублирующихся признаков, который был реализован для того, чтобы избежать избыточных вычислений и повысить качество анализа данных. Для этого использовалась тепловая матрица корреляции, позволившая исключить наиболее взаимосвязанные признаки.

Далее для устранения дисбаланса классов была применена комбинация методов ре-семплинга. Был построен пайплайн из алгоритмов `SMOTE`, `RandomOverSampler` и `Nearmiss`, затем для разных заявленных нижних и верхних границ проводился ресемплинг для определения наилучшей пары границ выборки.

На следующем этапе категориальные признаки были преобразованы в числовые форматы для их последующего использования в моделях машинного обучения с помощью метода `get_dummies()`, а затем полученные числовые признаки были масштабированы для обеспечения нормализации данных. Этот этап важен для улучшения работы моделей машинного обучения, так как масштабирование помогает избежать проблемы несоответствия масштабов признаков. После этого данные были разделены на обучающую и тестовую выборки для оценки производительности модели. Это стандартный процесс, который позволяет оценить, как хорошо модель будет работать на новых данных.

Далее был реализован `uplift-test` для оценки влияния каждого фактора на точность предсказания. При этом факторы постепенно добавлялись в модель и фиксировалась ключевая метрика. Это помогает понять, какие факторы оказывают наибольшее влияние на целевую переменную или результат предсказания. Также данный тест позволяет потенциально выявить факт и момент начала переобучения модели под рабочую выборку, когда качество модели начинает ухудшаться или не меняется на независимой выборке при росте качества на рабочей выборке. Результаты `uplift-test` в нашем случае, говорят о том, что метрика качества монотонно возрастает (рис. 1), и не происходит переобучение модели при увеличении числа признаков, т.е. все признаки можно использовать для реализации модели. Однако количество признаков также можно и уменьшить, если ключевая метрика на тестовой выборке преодолела устанавливаемый заказчиком порог. Так, например, если пороговым значением для `F1-score` является 0.95, то можно оставить лишь 24 признака для обучения модели (см. красную черту на рис. 1), что положительно скажется на времени обучения модели, и что более важно – на скорости сбора признаков на новых данных.

Далее был выполнен подбор гиперпараметров модели `RFC` с использованием метода `Grid Search` – это процесс систематического перебора различных комбинаций значений гиперпараметров с целью определения наилучшей конфигурации для модели. Гиперпараметры – это настраиваемые параметры, позволяющие управлять процессом обучения модели. Наилучшие гиперпараметры модели `RFC` после проведения `grid_search`: 100 – количество деревьев (`n_estimators`), 20 – максимальная глубина деревьев (`max_depth`), 2 – минимальное количество образцов, необходимых для разделения внутреннего узла (`min_samples_split`), 1 – минимальное количество об-

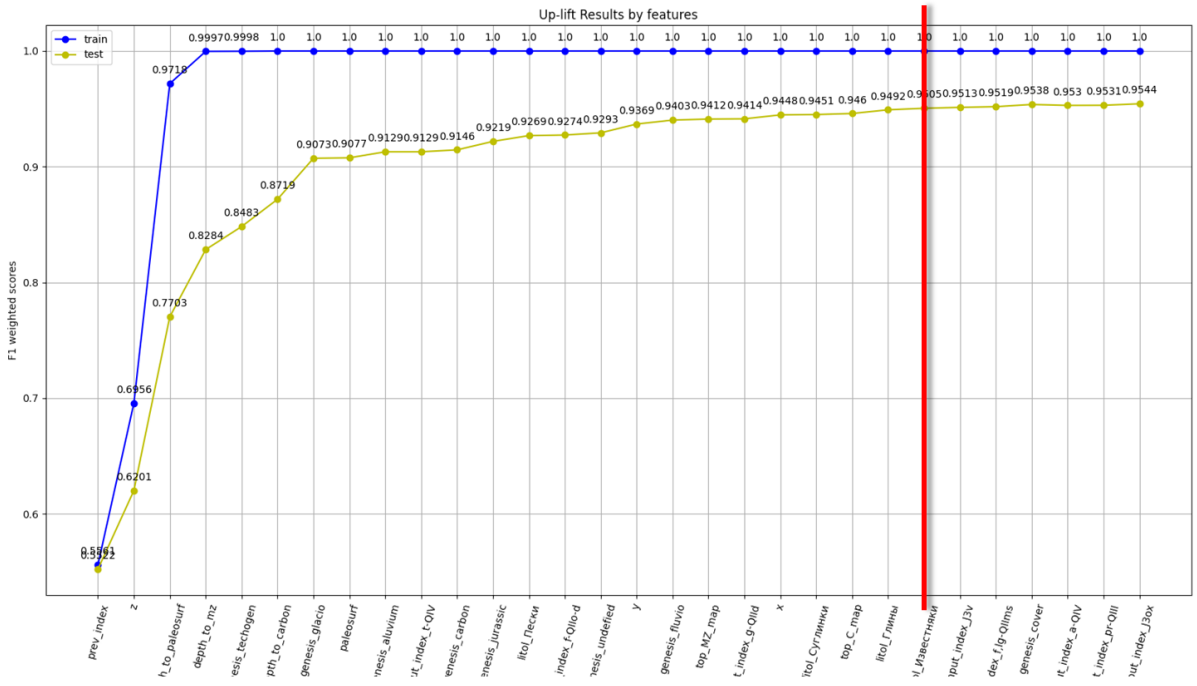

Рис. 1. Визуализация `uplift-test` для первых 30 признаков.

разцов, необходимых для листового узла (`min_samples_leaf`).

После подбора гиперпараметров и обучения модели было произведено сравнение `RFC` с другими классификаторами, которые можно было применить в этой задаче. На рис. 2 представлено сравнение производительности четырех различных моделей машинного обучения: `RFC`, `MLP`, `Naive Bayes` и `Decision Tree`.

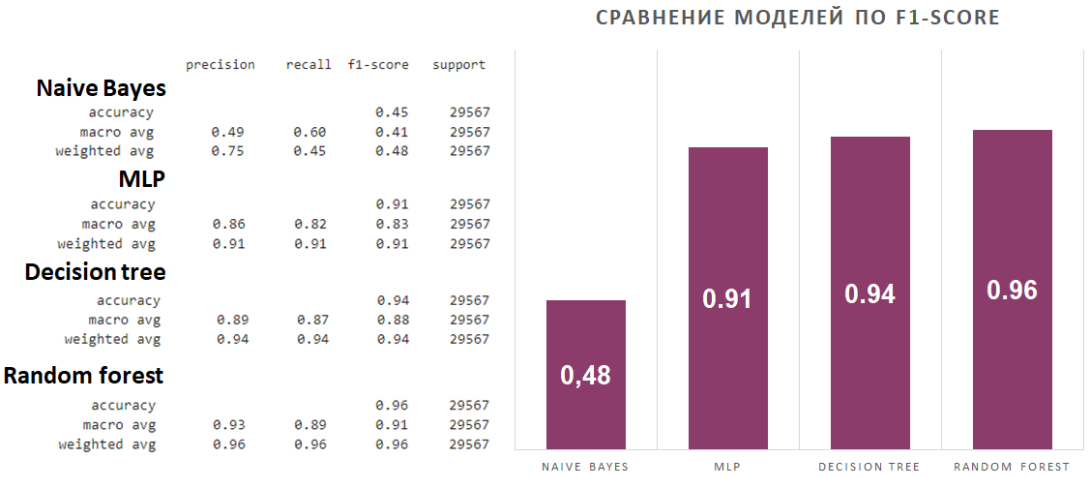

Рис. 2. Сравнение рассмотренных моделей по F1-score

Анализ результатов позволяет определить, ка-кая из указанных моделей лучше всего справляется с поставленной задачей:

1.`Naive Bayes`. *Преимущества*: проста в реализации, хорошо работает с категориальными данными, эффективен для задач классификации текста. *Результаты:*

точность предсказаний оставляет желать лучшего, её зачастую используют в качестве базовой модели, поскольку качество предсказаний основано на распределении классов.

2. `MLP`. *Преимущества*: способна моделировать сложные нелинейные зависимости, подходит для задач классификации и регрессии. *Результаты:* MLP достаточно неплохая модель, но для данной задачи показала более низкую точность, чем RFC.

3. `Decision Tree`. *Преимущества*: проста в интерпретации, не требует масштабирования данных, способна обнаруживать важные признаки. *Результаты*: по сути является ячейкой в `RFC`, поэтому она ожидаемо показала более низкий результат.

4. `RFC`. *Преимущества*: высокая точность благодаря ансамблевому подходу, устойчивость к переобучению. *Результаты:* гипотеза подтвердилась, `RFC` является лучшей моделью для решения поставленной задачи.

Практическим результатом проведенного в работе исследования является программное обеспечение, которое может использоваться в геотехнических исследованиях и строительных компаниях для оценки состояния грунта.

Литература

1. Nguyen M.D., Costache R., Sy A.H. et al. Novel approach for soil classification using machine learning methods // Bulletin of Engineering Geology and the Environment. 2022. 81, 468. https://doi.org/10.1007/s10064-022-02967-7

MSC 68T20 68-04

Application of machine learning methods for soil classification

A.A. Butkina, O.A. Gushina, A.S. Korzhov, A.V. Shamaev

National Research Mordovia State University

Abstract: The article describes how the authors apply of machine learning methods to classify soil types. An analysis of several predictive analytics methods was performed, and the «Random Forest Classifier» method was chosen as the best one.

Keywords: soil classification, soil types, predictive analytics, machine learning, random forest.

References

1. Nguyen M.D., Costache R., Sy A.H. et al. Novel approach for soil classification using machine learning methods // Bulletin of Engineering Geology and the Environment. 2022. 81, 468. https://doi.org/10.1007/s10064-022-02967-7