АНАЛИЗ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ПРОГНОЗИРОВАНИЯ КОНЦЕНТРАЦИИ СЕРЕБРА В ГЕОЛОГИЧЕСКОМ РАЙОНЕ ЦЕНТРАЛЬНОГО РУДНОГО УЗЛА (ЧУКОТСКОГО АВТОНОМНОГО ОКРУГА)

Фигероа Б.Ф.К.[1], Савинова О.В.[2]
[1]ТПУ, ИШПР, 2ЛМ21, студент, e-mail: ffigueroa.balvin@gmail.com
[2]ТПУ, ИШПР, к.г.-м.н. доцент, e-mail: logvinenkoov@tpu.ru

Введение

Центральный золото-сереброрудный узел расположен в Канчалано-Амгуэмской металлогенической зоне, являющейся частью внутренней зоны Охотско-Чукотского вулканического пояса (ОЧВП). Полезные ископаемые района образовались в течение мелового минерагенического этапа, совпадающего с этапом формирования ОЧВП. Наиболее характерным для Канчалано-Амгуэмской металлогенической зоны является оруденение золото-серебряной формации. В 25-ти километрах северо-западнее находится эксплуатируемое золото-серебряное месторождение Валунистое, выбранное в качестве эталонного объекта-аналога с установленной промышленной значимостью для Центральной перспективной площади (Ширшов С.А, 2020).

Основная часть

Целью работы является анализ и сравнение точности нейронной сети (ANN) и трех методов машинного обучения (ML) для прогнозирования содержания серебра (Ag) на основе геохимических данных, полученных методом приближенно-количественного спектрального анализа на 16 элементов (Pb, As, Cr, W, Ni, Co, Bi, Mn, Ba, Be, Li, Mo, Sn, Cu, Ag и Zn). Исходный материал представляет собой набор данных, содержащий 8920 образцов и 19 переменных, включающих координаты, глубину и концентрации элементов. На начальном этапе анализа оценивалось распределение собранных данных. Этот шаг имеет решающее значение для понимания того, как распределяются концентрации серебра в геохимических образцах. При проверке распределения данных было обнаружено, что они не соответствуют стандартному нормальному распределению. Поскольку данные не были распределены нормально, были применены методы нормализации. Цель нормализации – сделать данные более сопоставимыми и пригодными для дальнейшего анализа (James et al., 2013). В этом случае робастная нормализация помогла уменьшить влияние выбросов и улучшить распределение данных. Помимо нормализации, для преобразования данных использовался метод натурального логарифма (ln). Логарифмическое преобразование является распространенный метод обработки асимметричных данных, если они не соответствуют нормальному распределению (Carranza, E.J.M., 2011). Применение натурального логарифма уменьшает дисперсию и облегчает статистический анализ и моделирование. Для дальнейшего улучшения качества данных был реализован процесс удаления выбросов. Выбросами называют необычные наблюдения, которые могут исказить результаты анализа (Pearson, R. K., 2002). Удаление этих значений гарантирует, что аномальные данные не окажут негативного влияния на модели машинного обучения.

Данные были разделены на обучающую и тестовую выборки для обучения и оценки производительности моделей, соответственно. Такое разделение гарантирует, что модели смогут обобщать невидимые данные. С помощью перекрестного поиска по сетке были выбраны оптимальные параметры модели. Этот процесс позволяет настроить гиперпараметры модели машинного обучения для достижения максимально возможной производительности. Взаимосвязь между геохимическими переменными устанавливалась с помощью четырех основных подходов: факторного анализа, агломеративной кластеризации, кластерного анализа и метода k-means. Факторный анализ выявил основные закономерности в переменных, упрощая интерпретацию. Агломеративная кластеризация группирует элементы со схожими профилями, а кластерный анализ группирует связанные переменные. Алгоритм k-means сгруппировал элементы на основе общих характеристик. Эти подходы выявили закономерности и взаимосвязи между переменными, что повлияло на выбор модели машинного обучения для оценки концентраций серебра.

Что касается архитектуры модели искусственной нейронной сети (ANN), она была разработана с последовательной структурой, включающей несколько уровней: входной, скрытый уровень и выходной уровень. В процессе обучения ANN подвергалась 75 эпохам корректировки, при этом наблюдалась эффективная коррекция, минимизировавшая расхождения между предсказаниями модели и реальными

значениями. Максимальные значения коэффициента детерминации (R2) достиг значения 0,6187 в обучающей выборке и 0,6646 в тестовой выборке. Для оптимизации метода опорных векторов (SVM) учитывалось значение C, которое уравновешивает максимизацию точности и минимизацию ошибок классификации. Кроме того, была применена стандартизация данных с помощью Standard Scaler для обеспечения равномерного взвешивания функций в модели. Оценка выявила значение R2 равное 0,6555 на обучающем наборе и 0,6512 на тестовом наборе. С другой стороны, модель XG Boosting использовала модель XGB Regressor с корректировками нескольких ключевых параметров. Было выполнено 1000 обучающих итераций и установлена максимальная глубина 7 для контроля сложности модели. Параметр «eta» был установлен на 0,01, чтобы влиять на скорость обучения, а «colsample bytree» был установлен на 0,8, чтобы внести некоторую степень случайности в процесс.

Результаты, полученные с помощью модели XG Boosting, были идеальными. Коэффициент детерминации (R2) достиг максимального значения 0,8955 в обучающем наборе, что указывало на то, что модель способна объяснить примерно 89,55 % изменчивости обучающих данных. В тестовом наборе R2 составлял 0,7481, что позволяет предположить, что модель смогла объяснить около 74,81 % изменчивости невидимых данных (рис. 1).

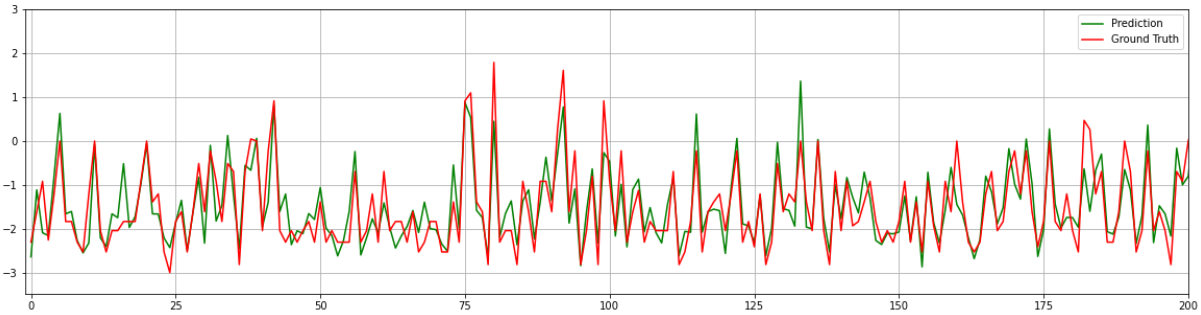

Рис. 1. Сравнение прогнозов (Prediction) с фактическими значениями (Ground Truth) в модели XGBOSST

Кроме того, была проведена перекрестная проверка, направленная на вычисление средней абсолютной ошибки (MAE). Результаты показали среднее значение MAE 0,354, что повысило эффективность и надежность модели XG Boosting при прогнозировании концентрации серебра. Наконец, модель случайного леса (RF). Был учтен параметр max_length, которому было присвоено высокое значение 1000 для управления максимальной глубиной деревьев. В обучающем наборе модель Random Forest достигла максимального значения R2 равного 0,9604. В тестовом наборе R2 составлял 0,7262, что позволяет предположить, что модель смогла объяснить около 72,62 % изменчивости невидимых данных.

Результаты были представлены в таблице, суммирующей баллы R2 для каждого метода в различных группах на основе их индекса корреляции (Таблица 1), где каждая группа представляла свой аналитический подход. Каждому методу был присвоен рейтинг на основе его эффективности.

Таблица 1

Сравнение эффективности прогнозирования на основе R2 по группам

Group	Relationship indices	ANN	SVM	XG BOOST	RF	Rank
1	All data	0.6620	0.6512	0.7481	0.7262	1
2	Factor analysis	0.6398	0.6065	0.6896	0.6960	4
3	Agglomerative cluster	0.4527	0.3320	0.5155	0.5801	5
4	Cluster by variables	0.6646	0.6386	0.7227	0.7222	2
5	K-means	0.6390	0.6107	0.7171	0.7032	3

С точки зрения сходства, все методы моделирования имели общий подход к оценке концентрации серебра. Все они продемонстрировали способность улавливать изменчивость данных, но с разным

уровнем точности. Более того, результаты отражают общую согласованность рангов, присвоенных методам, что предполагает определенную стабильность их относительной эффективности в различных группах.

Однако, между методами наблюдались и существенные различия. XGBoost получил наивысший балл R2 в большинстве групп, что указывает на его превосходство с точки зрения точности прогнозирования. SVM и RF также показали хорошие результаты, тогда как ANN и SVM в некоторых случаях показали несколько худшие результаты (рис. 2).

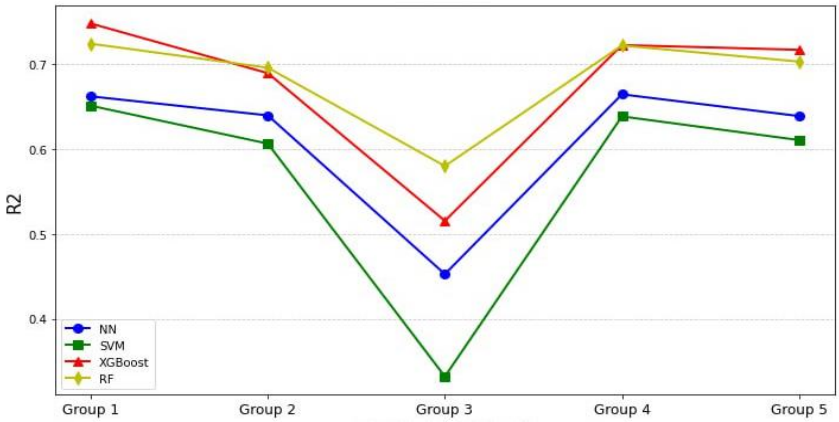

Рис. 2. Сравнение результатов прогнозирования коэффициента детерминации (R2) для каждой группы

Данное исследование подчеркивает важность применения и анализа не одного, а нескольких методов машинного обучения в области разведки полезных ископаемых и способствует более полной и точной оценке концентраций драгоценных металлов в месторождениях полезных ископаемых.

Список использованных источников

1. Ширшов С.А. Поисковые работы на золото и серебро в пределах Центрального рудного узла Канчалано-Амгуэмской рудной зоны (Чукотский АО) // Проектная документация. – Москва, 2020.
2. Carranza E.J.M. Analysis and mapping of geochemical anomalies using log ratio-transformed stream sediment data with censored values // Journal of Geochemical Exploration. – 2011. – V. 110. – P. 167–185. – doi: 10.1016/j.gexplo.2011.05.007.
3. Ibrahim B., Majeed F., Ewusi A., & Ahenkorah I. Residual geochemical gold grade prediction using extreme gradient boosting // Environmental Challenges. – 2022. – V. 6. – https://doi.org/10.1016/j.envc.2021.100421.
4. James G., Witten D., Hastie T., & Tibshirani R. An Introduction to Statistical Learning // Springer. – 2013.
5. Pearson R.K. Outliers in process modeling and identification // IEEE Transactions on Control Systems Technology. – 2002. – V. 10(1). – P. 55–63. – Doi: 10.1109/87.974338.