

Эквивиальность в задачах гидрологического прогнозирования

Эльдар Хузин

Московский физико-технический институт

Консультант: Д. Абрамов

2025

Цель исследования

Цель

Исследовать эквивалентность моделей RF, XGBoost, LSTM, а также их способность выявлять гидрологические закономерности в данных о стоке рек.

Задачи

1. Сравнить модели по метрикам NSE, RMSE, MAE, MAPE.
2. Построить “карту эквивалентности” — области, где ошибки моделей статистически неразличимы.
3. Оценить вклад признаков через SHAP-значения.
4. Проанализировать компромисс “точность\вычислительная сложность\интерпретируемость”.

Решение

Провести на обширном датасете временных рядов и статических признаков SHAP-анализ моделей.



Beven, Keith и Andrew Binley (июль 1992). “The Future of Distributed Models: Model Calibration and Uncertainty Prediction”. В: *Hydrological Processes* 6.3, с. 279—298. DOI: [10.1002/hyp.3360060305](https://doi.org/10.1002/hyp.3360060305).



Janssen, Joseph, Ardalan Tootchi и Ali Ameli (2 мая 2024). *Tackling Water Table Depth Modeling via Machine Learning: From Proxy Observations to Verifiability*. DOI: [10.31223/X5939N](https://doi.org/10.31223/X5939N).



“Spatial-Temporal Behavior of Precipitation Driven Karst Spring Discharge in a Mountain Terrain” (1 сент. 2022). В: *Journal of Hydrology* 612, с. 128116. DOI: [10.1016/j.jhydro1.2022.128116](https://doi.org/10.1016/j.jhydro1.2022.128116).

Вычислительный эксперимент

1. Предобработка данных

1.1 Корреляционный анализ и отбор статических признаков

1.2 Заполнение пропусков

1.3 Добавление лагов и статических атрибутов

2. Обучение

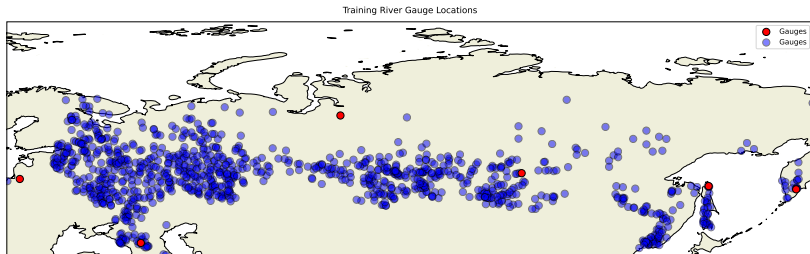
2.1 Выделение выборок (12/3)

2.2 Подбор гиперпараметров

2.3 Обучение

3. SHAP-анализ и другие метрики

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|! (|N| - |S| - 1)!}{|N|!} [f(S \cup \{i\}) - f(S)]$$



Ежедневные данные за период с 2008 по 2022 год на 1072 бассейнов.

prcp – количество осадков (мм/сут)

lvl_sm – уровень воды в реке (м)

t_min – минимальная температура (°C)

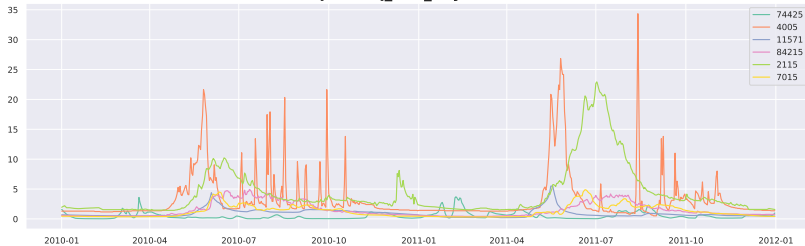
t_max – максимальная температура (°C)

t_mean – средняя температура (от почасовой) (°C)

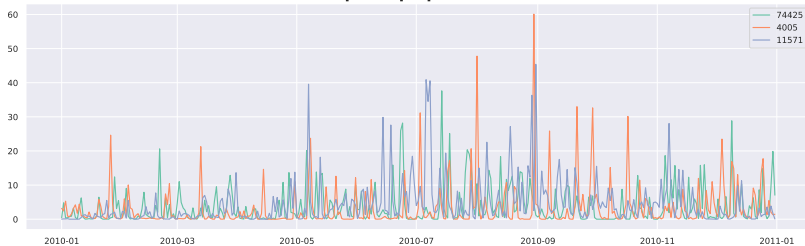
q_mm_day – расход воды (мм/сут)

Периодичность: суточная.

Example of q_mm_day time series



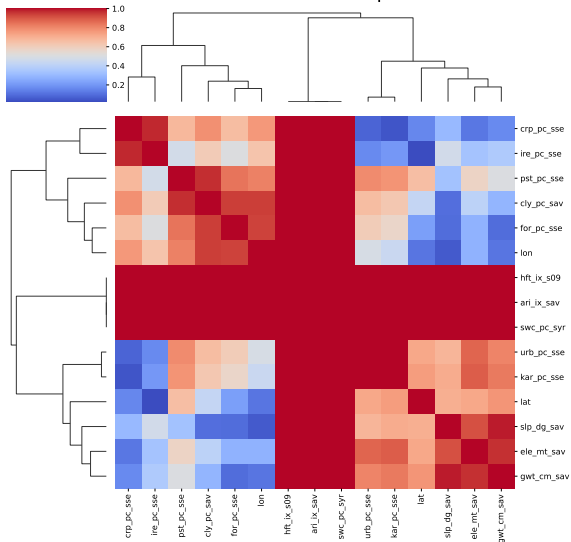
Example of prcp time series



Описание атрибутов

ID	Атрибут	
L07	Лесные территории (%)	for_pc_sse
L08	Сельскохозяйственные угодья (%)	crp_pc_sse
L09	Пастбища (%)	pst_pc_sse
L10	Орошаемые земли (%)	ire_pc_sse
A03	Урбанизированные территории (%)	urb_pc_sse
S07	Карстовые ландшафты (%)	kar_pc_sse
P01	Средняя высота над уровнем моря (м)	ele_mt_sav
P02	Средний уклон местности (°)	slp_dg_sav
C06	Аридность	ari_ix_sav
H10	Глубина залегания грунтовых вод (см)	gwt_cm_sav
S05	Среднегодовая влажность почвы (%)	swc_pc_syr
S01	Глина в почве (%)	cly_pc_sav
A06	Индекс антропогенного воздействия	hft_ix_s09
P04	Географическая широта (°)	lat
P05	Географическая долгота (°)	lon

Correlation Heatmap



Формальная постановка задачи

- ▶ Предсказать y на основе признаков x .
- ▶ Пусть $y \in \mathbb{R}^n$; $x \in \mathbb{R}^{n \times d}$;
- ▶ Набор данных: $\mathcal{D} = \{(x_i, y_i) \mid i = 1, \dots, n\}$;
- ▶ Класс моделей: $\mathcal{F}$;
- ▶ Оценка: $\hat{y}_i = f(x_i)$, $f \in \mathcal{F}$, $\hat{y} = (f(x_1), \dots, f(x_n))$.

$$f^* = \arg \min_{f \in \mathcal{F}} (\mathcal{L}(f, \mathcal{D}) + \Omega(f)) \quad (1)$$

Regression Trees

Модель:

$$f(x) = \sum_{l=1}^L \omega_l \mathbb{I}(x \in R_l).$$

Построение разбиений по критерию:

$$\mathcal{JG}(t,j) = \frac{1}{2} \left[\frac{(G_L)^2}{H_L + \lambda} + \frac{(G_R)^2}{H_R + \lambda} - \frac{G^2}{H + \lambda} \right] - \gamma,$$

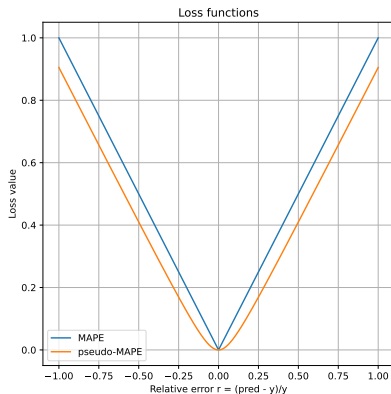
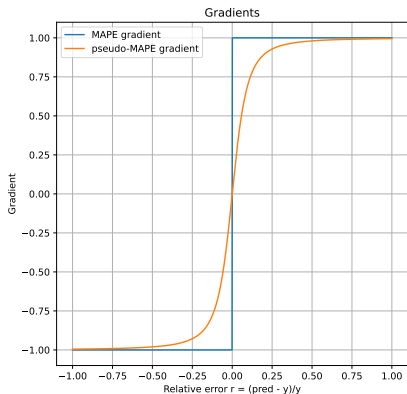
где $G_{L,R}, H_{L,R}$ — суммы градиентов и гессианов.

$$\mathbf{MAPE} = \frac{1}{n} \sum_{i=1}^n \left| \frac{\hat{y}_i - y_i}{y_i} \right|$$

$$\mathbf{Pseudo-MAPE}_{\delta} = \frac{1}{n} \sum_{i=1}^n \left(\sqrt{\left(\frac{\hat{y}_i - y_i}{y_i} \right)^2 + \delta^2} - \delta \right)$$

$$\mathbf{NSE} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

MAPE и Pseudo-MAPE



Random Forest и Gradient Boosting

Random Forest

$$f_{\text{RF}}(x) = \frac{1}{B} \sum_{b=1}^B f^{(b)}(x)$$

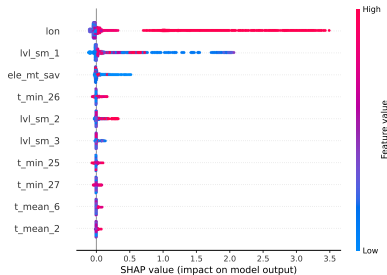
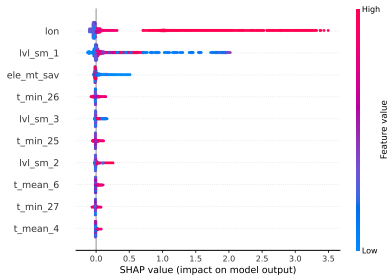
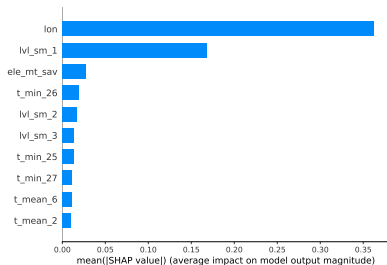
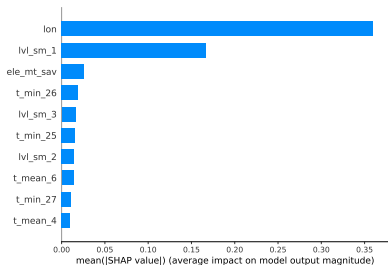
- ▶ Разделение времён года

Gradient Boosting

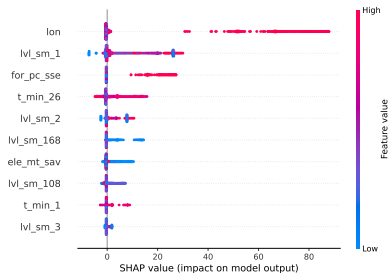
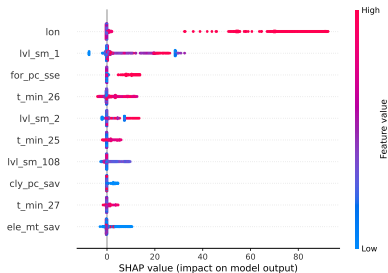
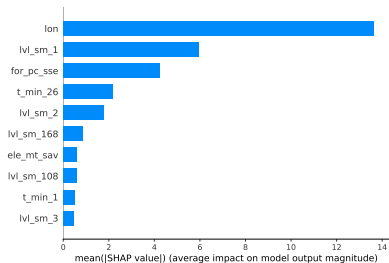
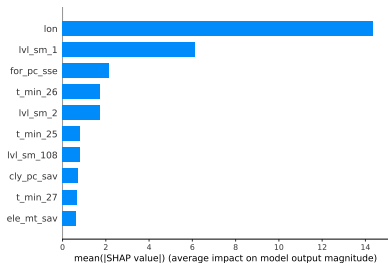
$$f_{(B)}(x) = f^{(0)}(x) + \nu \sum_{b=1}^B f^{(b)}(x)$$

- ▶ Последовательная обработка осадков из прошлого.

Результаты: Gradient Boosting



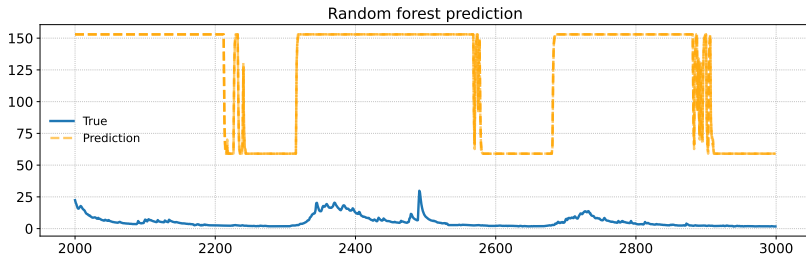
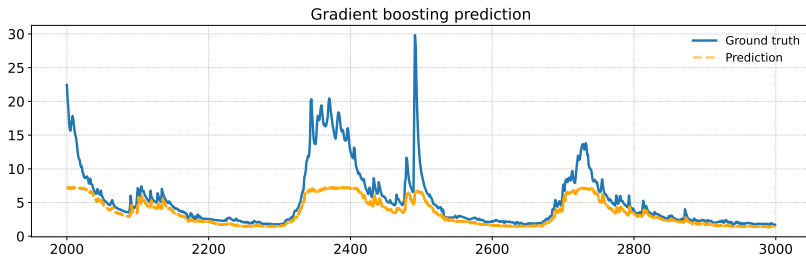
Результаты: Random Forest



Результаты

Model	NSE	MAPE	RMSE	MAE
Random Forest	-398.5	1355.1	60.654	36.755
Gradient Boosting	0.470	27.009	2.216	0.853

Результаты



Заключение

- ▶ Требуется анализ с использованием обширного датасета
- ▶ Модели внутри класса Gradient Boosting не проявляют эквивиальности
- ▶ Модели внутри класса Random Forest не проявляют эквивиальности
- ▶ Лучшие в классах Random Forest и Gradient Boosting модели выделяют разные признаки