

АРХИТЕКТУРА ML-СИСТЕМЫ ПРОГНОЗИРОВАНИЯ РИСКА САМОВОЗГОРАНИЯ УГЛЯ НА СКЛАДАХ

1. Введение

Самовозгорание угля представляет собой значительную техногенную угрозу на угольных складах, приводя к экологическим, экономическим и безопасностным последствиям. Традиционные методы мониторинга, основанные на пороговых значениях температуры, часто не учитывают комплексное взаимодействие факторов риска. В данной работе представлена архитектура системы машинного обучения, предназначенной для прогнозирования вероятности самовозгорания угля на основе мультимодальных данных.

2. Методология

2.1. Архитектура системы

Разработанная система реализована в виде пайплайна, состоящего из следующих компонентов (табл. 1).

Таблица 1 – Архитектура системы

| Компонент системы | Описание |
|---------------------|--|
| Загрузчик модели | Выполняет десериализацию предварительно обученной модели RandomForestClassifier и вспомогательных объектов, в частности, LabelEncoder для кодирования категориальных признаков. |
| Модуль предсказания | Реализован в виде класса CoalFirePredictor; предоставляет метод add_predictions_to_df, который добавляет к входному датафрейму две колонки: <ul style="list-style-type: none">• fire_proba — вероятность самовозгорания (от 0 до 1)• fire_pred — бинарный прогноз (1 — высокий риск, 0 — низкий риск) |
| Источники данных | Система использует данные из следующих файлов: <ul style="list-style-type: none">• supplies.csv — информация о поставках угля• temperature.csv — температурные измерения в штабелях• weather_data_*.csv — метеорологические данные• fires.csv — реестр зарегистрированных пожаров |

2.2. Загрузчик модели представляет собой осуществление десериализации предварительно обученной модели RandomForestClassifier и сопутствующих объектов, напр., LabelEncoder для кодирования категориальных признаков).

2.3. Модуль предсказания инкапсулирован в классе `CoalFirePredictor`, предоставляющем метод `add_predictions_to_df`, который расширяет входной датафрейм двумя колонками:

- `fire_proba` — вероятностная оценка риска (в диапазоне `[0, 1]`),
- `fire_pred` — бинарное решение (1 — высокий риск, 0 — низкий риск).

2.4. Источники данных представляет собой систему, которая интегрирует данные из пяти типов источников:

- поставки угля (`supplies.csv`),
- температурные измерения в штабелях (`temperature.csv`),
- метеоданные (`weather_data_*.csv`),
- реестр зарегистрированных пожаров (`fires.csv`)

2.5. Признаковое пространство

Модель оперирует 10 признаками, включая категориальные и числовые переменные:

- Категориальные:
 - Марка — тип угля.
- Числовые:
 - возраст штабеля (Возраст_дн),
 - масса (`mass`),
 - максимальная температура в штабеле (Максимальная температура),
 - скорость изменения температуры (Темп_изменение),
 - временные и метеорологические переменные (`weekday`, `month`, `t`, `p`, `humidity`).

Категориальные признаки преобразуются с использованием `LabelEncoder`, обученного на обучающей выборке.

3. Этапы анализа

3.1. Валидация на синтетических данных

Для верификации корректности предсказаний выполнена проверка модели на наборе сценариев с экстремальными и типичными значениями признаков. Результаты подтверждают адекватную чувствительность модели к сочетанию критических факторов, напр., высокая температура + возраст штабеля + низкая влажность.

3.2. Прогноз на реальных данных

Для лет 2019 и 2020 годов:

- отобраны по 25 штабелей;
- для каждого сгенерировано 24 наблюдения (2 раза в месяц);

- для каждого наблюдения рассчитаны признаки на основе исторических данных и синтетических аугментаций (масса, темп изменения температуры).

Модель применяется к полученным датафреймам с последующим сохранением предсказаний в формате CSV.

3.3. Оценка качества

Оценка эффективности выполнена по двум метрикам:

- Точность по событиям (1):

$$Acc_{\text{события}} = \frac{\min(N_{\text{pred}}, N_{\text{real}})}{\max(N_{\text{pred}}, N_{\text{real}})} \cdot 100\% \quad (1)$$

- Точность по штабелям (2):

$$Acc_{\text{штабели}} = \frac{|S_{\text{pred}} \cap S_{\text{real}}|}{|S_{\text{real}}|} \cdot 100\% \quad (1)$$

- Дополнительно анализируются:
 - распределение вероятностей,
 - месячная динамика рисков
 - топ-штабели по частоте предсказанных инцидентов

4. Визуализация и интерпретация

Для каждого года выведены четыре графика:

- Сравнение числа предсказанных и реальных пожаров по месяцам;
- Среднемесячная вероятность риска;
- Гистограмма распределения fire_proba;
- Топ-10 штабелей с наибольшим числом предсказанных пожаров.

Графики сохранены в файлы analysis_2019.png, analysis_2020.png и предназначены для последующей интеграции в отчётность операторов складов.

5. Заключение

Представленная ML-архитектура позволяет осуществлять проактивный мониторинг рисков самовозгорания угля, объединяя оперативные сенсорные данные, метеорологическую информацию и машинное обучение. Использование вероятностного подхода обеспечивает гибкость интерпретации и возможность раннего вмешательства до достижения критических температур. Дальнейшее развитие системы может включать онлайн-обновление модели, использование ансамблей и интеграцию с системами IoT-мониторинга.