1. Введение

Данный отчет представляет результаты разведочного анализа данных, проведенного в рамках хакатона "Flood Challenge". Целью анализа было выявление ключевых факторов, влияющих на уровень воды в реках Восточно-Казахстанской области, и разработка модели для прогнозирования паводковых ситуаций.

1. Анализ важности признаков

2.1 Распределение важности признаков

На основе анализа важности признаков (Image 1 и Image 2) были выявлены следующие ключевые факторы:

1. water\_level\_critical - критический уровень воды (важность 0.7)
2. water\_level\_rolling\_mean - скользящее среднее уровня воды (важность 0.12)
3. water\_level\_rolling\_std - стандартное отклонение скользящего среднего (важность 0.09)
4. water\_flow - водный поток (важность 0.04)
5. water\_level\_diff - разница в уровнях воды (важность 0.03)

Остальные признаки (месяц, день недели, сезон, год) имеют относительно низкую важность для модели.

2.2 Интерпретация результатов

Критический уровень воды (water\_level\_critical) оказался наиболее значимым фактором для прогнозирования паводковых ситуаций. Это логично, так как данный показатель, вероятно, учитывает географические и исторические особенности каждого гидропоста.

Скользящее среднее и стандартное отклонение уровня воды также играют важную роль, что указывает на значимость динамики изменения уровня воды во времени.

Водный поток и разница в уровнях воды имеют меньшую, но все же заметную важность, что говорит о необходимости учета этих факторов при прогнозировании.

1. Анализ распределения и сезонности

3.1 Распределение текущего уровня воды

Анализ показал, что распределение текущего уровня воды не является нормальным и имеет правостороннюю асимметрию. Это означает, что большинство наблюдений сосредоточено в области низких и средних значений, но присутствуют выбросы в области высоких значений. Такое распределение может указывать на редкие, но потенциально опасные случаи высокого уровня воды.

3.2 Сезонность

Выявлена четкая сезонность в уровне воды. Различия между сезонами могут быть связаны с такими факторами, как таяние снега весной, летние дожди или осенние паводки. Учет сезонности важен для повышения точности прогнозов.

1. Корреляционный анализ

Анализ корреляций между различными параметрами выявил следующие значимые взаимосвязи:

* Сильная положительная корреляция между текущим уровнем воды и критическим уровнем воды (0.81)
* Умеренная корреляция между текущим уровнем воды и водным потоком (0.41)

Эти корреляции подтверждают важность учета критического уровня воды и водного потока при прогнозировании паводковых ситуаций.

1. Моделирование и оценка результатов

5.1 Выбор модели

В ходе анализа были протестированы две модели: Random Forest и Gradient Boosting. Лучшие результаты показала модель Gradient Boosting Regressor с следующими параметрами:

* max\_depth: 10
* n\_estimators: 200

5.2 Оценка эффективности модели

Модель продемонстрировала высокую эффективность на валидационной выборке:

* R2 score: 0.98
* MSE (Mean Squared Error): 402.29
* MAE (Mean Absolute Error): 11.67

Высокий R2 score (0.98) указывает на то, что модель объясняет 98% вариации в данных, что является отличным результатом. Низкие значения MSE и MAE также подтверждают высокую точность модели.

5.3 Прогнозирование

Разработанная модель способна прогнозировать уровень воды на 7 дней вперед, что соответствует поставленной задаче.

1. Выводы и рекомендации

6.1 Основные выводы

1. Критический уровень воды является ключевым фактором для прогнозирования паводковых ситуаций.
2. Динамика изменения уровня воды (скользящее среднее и стандартное отклонение) играет важную роль в прогнозировании.
3. Сезонность оказывает значительное влияние на уровень воды.
4. Разработанная модель демонстрирует высокую точность прогнозирования.

6.2 Рекомендации

1. Улучшение сбора данных: Рассмотреть возможность сбора дополнительных данных, таких как информация о погоде (осадки, температура), таянии снега, состоянии почвы и ландшафтных особенностях местности.
2. Расширение временного горизонта: Исследовать возможность увеличения горизонта прогнозирования до 14 или 30 дней для более эффективного планирования превентивных мер.
3. Учет экстремальных событий: Разработать дополнительные методы для лучшего прогнозирования редких, но потенциально катастрофических паводковых ситуаций.
4. Интеграция с ГИС: Рассмотреть возможность интеграции модели с геоинформационными системами для визуализации прогнозов и улучшения пространственного анализа.
5. Мониторинг и обновление модели: Регулярно проводить переобучение модели на новых данных для поддержания ее актуальности и точности.
6. Разработка системы раннего оповещения: На основе прогнозов модели создать систему раннего оповещения для своевременного информирования соответствующих служб и населения о потенциальных паводковых ситуациях.

Заключение

Разработанная модель демонстрирует высокую эффективность в прогнозировании уровня воды и может служить надежной основой для системы предупреждения паводковых ситуаций в Восточно-Казахстанской области. Реализация предложенных рекомендаций позволит дальше улучшить точность прогнозов и расширить возможности системы, что в конечном итоге приведет к повышению безопасности населения и снижению экономического ущерба от наводнений.