цифровой прорыв

сезон: ИИ

Задача “Радар тенденций новостных статей”, RBK Group

Николай Мошков

telegram: @Affernus
e-mail: n_moshkov@mail.ru
git: https://github.com/Affernus/RBC

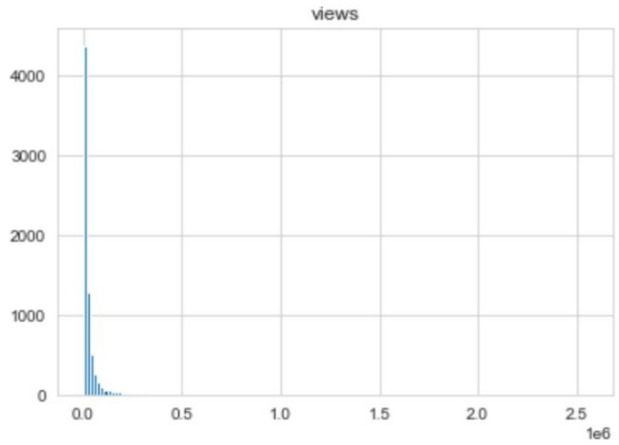
Цель работы

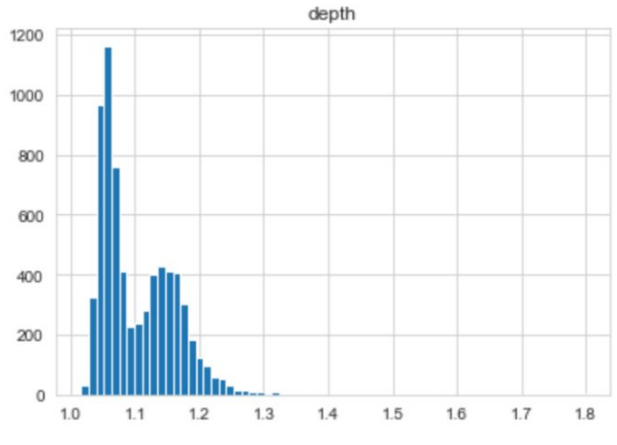
Проанализировать новости российских СМИ и научиться предсказывать их популярность. Ожидается, что для этого будут использованы NLP модели

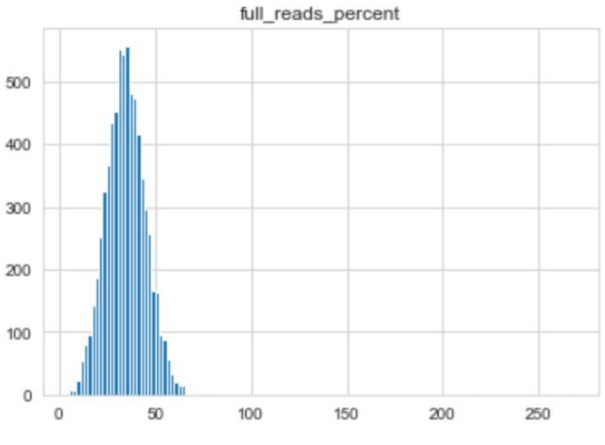
Цель модели — предсказать 3 численные характеристики, которые в полной мере показывают популярность статьи: views, full reads percent, depth.
Для оценки качества решения используется метрика R2.

Анализ данных: целевые переменные

Видно, что по всем целевым признакам есть выбросы. По проценту прочтения есть значения с цифрой больше 100, глубина прочтения - бимодальное распределение. Количество просмотров - сильно асимметричное распределение. Среди новостей очевидно есть “вирусные”

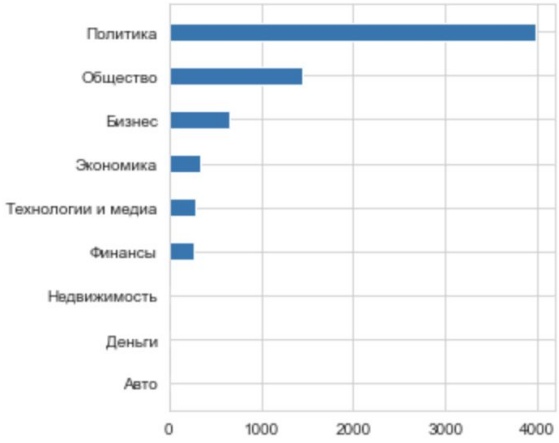


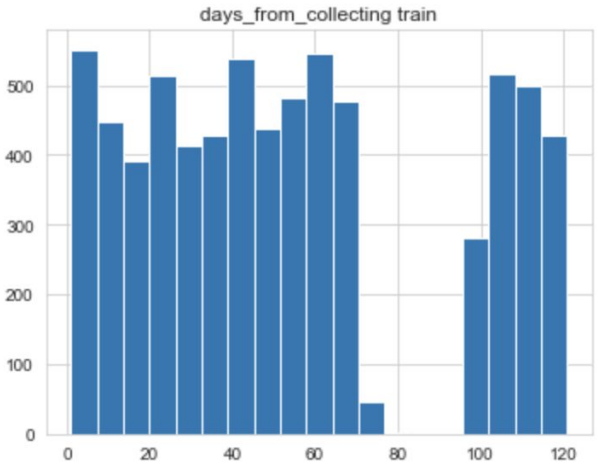


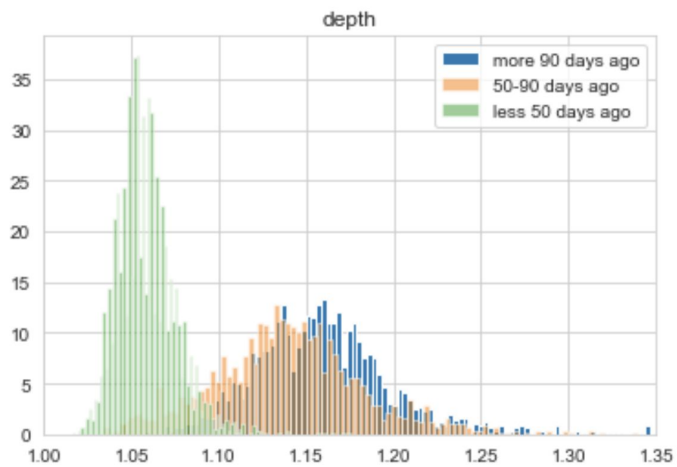


Анализ данных: наблюдения и инсайты

- Политика, общество и бизнес - топ-3 категорий по количеству статей
- Из данных вырезан кусок, совпадающий с началом СВО
- Интересный факт: глубина просмотра новостей снизилась за последние 50 дней до сбора данных, поведение читателей сместилось в сторону скрининга новостей







Анализ данных: наблюдения и инсайты

В данных обнаруживаются схожие записи от разных дат на одинаковых url. Бывает, что немного меняется заголовок и текст, но общая суть новости сохраняется. Также во времени меняется ctr. Вероятно, связано с тем, что актуальные и популярные новости периодически обновлялись уже после публикации.

publish_date	title	ctr	views
2022-03-23 11:29:10	Какие места на Украине взяли под контроль российские военные. Карта	5.907	616365
2022-03-31 18:58:21	Какие места на Украине взяли под контроль российские военные. Карта	5.907	616365
2022-03-31 18:58:21	Какие места на Украине взяли под контроль российские военные. Карта	0.862	317039
2022-04-01 17:54:10	Какие места на Украине взяли под контроль российские военные. Карта	0.862	317039
2022-01-30 08:55:18	Выявили более 120 тыс. заболевших. Актуальное о COVID на 1 февраля	3.759	518294
2022-02-01 08:38:12	Вирус разучился проникать в легкие. Актуальное о COVID на 2 февраля	3.759	518294
2022-02-04 08:41:41	Снижение коллективного иммунитета. Актуальное о COVID на 5 февраля	3.759	518294
2022-02-05 08:26:49	Коронавирусные ограничения смягчили. Актуальное о COVID на 7 февраля	3.759	518294

Анализ данных: гипотезы

Популярность новостной статьи потенциально может быть связана с:

- **особенностями конкретной статьи** (стиль текста, объем, читабельность, тематика, соответствие заголовка содержанию и т.д.; перечисленное так или иначе зависит от того, кто является автором статьи и сколько всего авторов)
- **особенностями отнесения к теме, подбора тегов и ключевых слов** (к какой категории отнесена статья, какие теги и ключевые слова из нее выделены, в какой части новостной ленты она размещена и как выделена/помечена)
- **насколько статья соответствует общему пулу статей за последнее время**, несет ли новую информацию по теме и т.д.
- **обстановкой в стране и в мире** (экономической, политической, эпидемиологической и т.д; определяет интересы читателей и горячие темы)
- **временем выпуска** (выходной или будний день, утренние или вечерние часы и т.д.)

Разработка ML-моделей: создание признаков

- парсинг текстов статей и сбор дополнительных данных: погода, заболеваемость и смертность от covid-19, макроэкономические индикаторы (цена на нефть, золото, курс доллара к рублю, индекс ММВБ-РТС, ключевая ставка ЦБ, фондовые индексы S&P 500, NASDAQ, FTSE, Nikkei, DAX)
- обучение модели word2vec на текстах статей и расчет эмбеддингов для статьи в целом, заголовка, авторов, тегов, первых и последних 100 слов статьи
- данные о статье (показатели читабельности - среднее количество слов в предложении, слогов в слове, доля уникальных слов, многосложных слов, объем, количество авторов, тегов, ключевых слов, размер заголовка и т.д.)
- соответствие общему пулу предшествующих статей (как часто теги, авторы, категории, ключевые слова встречались в статьях за прошедший период, сколько всего статей выпущено за прошедшие сутки, неделю, месяц и т.д.)

Разработка ML-моделей: подбор и валидация

Для каждой из трех целевых переменных построена своя модель

Не взлетели: LSTM, линейная регрессия

Показали хороший результат: feature selection + градиентные бустинги LightGBM, CatBoost и их комбинации

Методика подбора признаков: топ N признаков по усредненному на 50 бутстрап-разбиениях (70% трейн + 30% валидация) feature importance на валидации. Модель для получения feature importance - LGBMRegressor с небольшим количеством estimators и неглубокими деревьями (n_estimators=100, learning_rate=.1, num_leaves=7) (подробности в коде проекта)

Методика кросс-валидации и подбора гиперпараметров: бутстрап-разбиение обучающего набора данных 50 раз на трейн 70% + валидация 30%, ручной fine-tuning с целью максимизировать среднее значение R2 на валидации и минимизировать среднее значение RMSE

Описание выбранных моделей

Для получения векторного представления слов:

gensim word2vec, алгоритм на основе нейронных сетей, обученный на текстах статей РБК, гиперпараметры: *min_count=5, window=9, vector_size=100, negative=5, epochs=20*

Для прогноза каждой из целевых переменных views, depth, full_reads_percent:

Градиентный бустинг, пара (стэк) LightGBM + Catboost, итоговый прогноз - усреднение прогнозов двух моделей на объекте;

Дополнительно - эвристика: если аналогичная статья (аналогичный или близкий ctr) встречалась ранее (в обучающем наборе), то прогноз - усредненное по уже известным статьям значение

Эвристика в прототипе модели затрагивает всего несколько десятков статей, но может быть расширена за счет поиска похожих статей, например, по косинусному расстоянию между текстами. В прототипе использование эвристики можно при необходимости отключить.

Обоснование точности решения

Для прогноза использовали мощные современные ансамблевые алгоритмы. При настройке подбирали гиперпараметры и признаки (fine-tuning + feature selection). Оценивали качество на кросс-валидации: 50 бутстрап разбиений (70% трейн + 30% валидация). Для каждого из 50 разбиений модель обучали на трейне, метрику считали на валидации. Пересечение валидации и трейна внутри фолда исключалось. Анализировали средние значения и СКО метрик R2 и RMSE

- оценка RMSE дополнительно к R2 позволяет исключить низкую прогнозную способность модели при высоком R2
- оценка R2 по среднему значению на 50 фолдах минимизирует влияние случайности при выборе гиперпараметров модели
- оценка СКО позволяет заметить overfitting модели (маркер - рост СКО на валидации при снижении СКО на трейне)

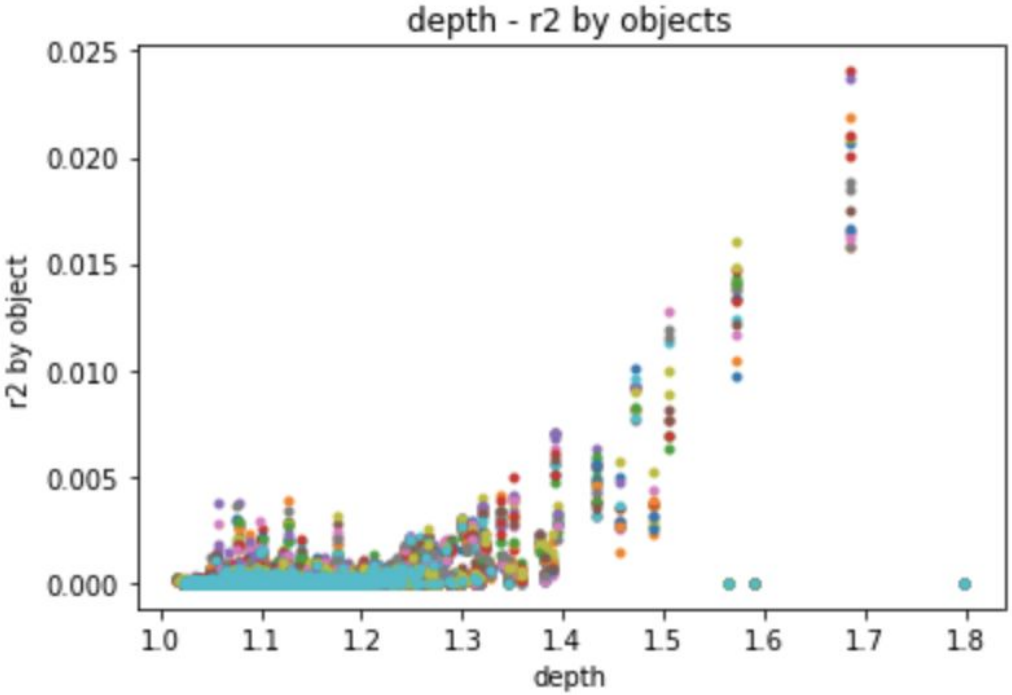
Дополнительный анализ - вклад в итоговое значение R2 для каждого отдельного объекта:

*R2_i = (true_i - predicted_i)**2 / sum((true_i - mean)**2)*

Обоснование точности решения

Пример анализа вклада в итоговое значение R2 для каждого отдельного объекта: видны потенциальные проблемы в области высоких значений depth, повторяющиеся от валидации к валидации.

Визуальный анализ упрощает подбор гиперпараметров и признаков, подсвечивает проблемные зоны



Обоснование точности решения

Комбинирование в решении современных ансамблевых алгоритмов, подбор признаков и гиперпараметров на кросс валидации, визуальный контроль ошибок на отдельных объектах обеспечили высокое качество итоговых моделей (низкий R2 view для train связан с исключением из расчета вирусных статей с высокими views).

Высокая вариабельность значений R2 связана, предположительно, с наличием выбросов (небольшого количества объектов с очень высокими значениями целевого признака)

`depth`

index	mean	std
cv_score_val	0.8549	0.0177
cv_score_train	0.9396	0.0014
cv_rmse_val	0.0249	0.0013
cv_rmse_train	0.0147	0.0002

`full_reads_percent`

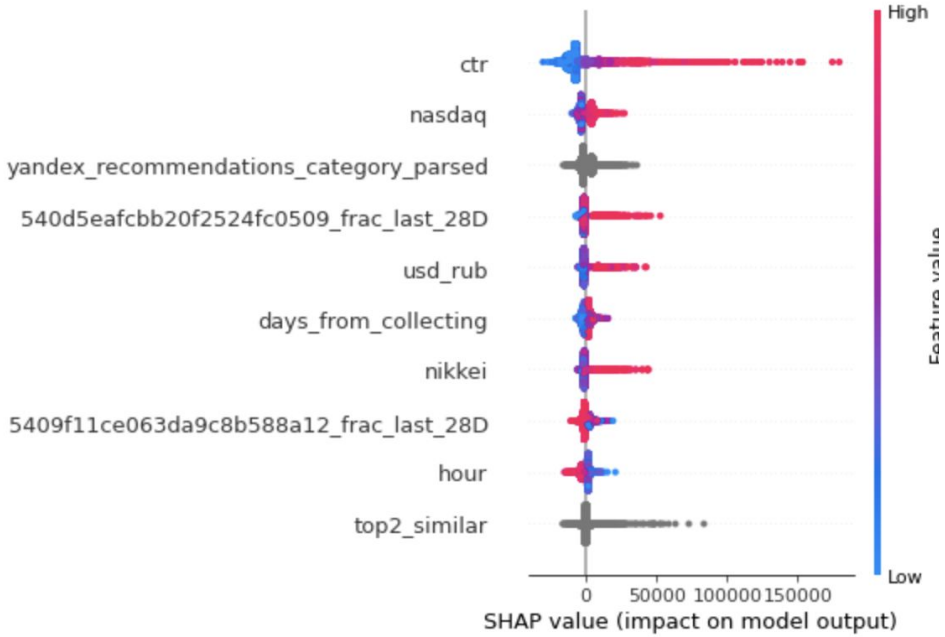
index	mean	std
cv_score_val	0.5368	0.0685
cv_score_train	0.6822	0.0411
cv_rmse_val	7.2742	0.9686
cv_rmse_train	6.0603	0.5545

`views`

index	mean	std
cv_score_val	0.8382	0.0913
cv_score_train	0.6089	0.0332
cv_rmse_val	37662.5975	9896.8175
cv_rmse_train	42477.6072	4139.7029

Влияние признаков на прогноз моделей: SHAP анализ, пример интерпретации

views



full reads percent

Спасибо за внимание

Контактные данные:

Николай Мошков
telegram: @Affernus
e-mail: n_moshkov@mail.ru
git: https://github.com/Affernus/RBC