Решение задачи «Радар тенденций новостных статей»

•••

июль 2022 г.

Обзор

Стоит задача научиться предсказывать популярность новостных статей для расширения аудитории компании РБК за счет добавления статей на актуальные темы.

Анализ проблемы

Гипотеза 1	Недостаток гипотезы 1	Возможное решение
Большая часть просмотров и др. целевых переменных формируется согласно привычкам пользователей и имеют паттерны и тенденции во времени (модель SARIMAX)	Предоставлено недостаточно данных для выявления долгосрочных тенденций, а также регулярных, но редких событий (выборы, чемпионаты мира и пр.). Множество нерегулярных событий (COVID)	Выявлять краткосрочные паттерны и тенденции. Для редких и нерегулярных событий формировать дополнительные признаки (например на основе заголовков)

Анализ проблемы

Гипотеза 2	Недостаток гипотезы 2	Возможное решение
Категории и авторы имеют свои собственные паттерны (не все пользователи читают все статьи; также не всем пользователям интересны темы о которых пишут авторы из интересующих их категорий)	Предоставлено недостаточно данных. Для некоторых авторов всего пара статей и 200 статей для категории. При обучении на столь малом количестве данных риск недообучения либо переобучения велик.	Использовать данные по авторам и категориям как вспомогательные вкупе с остальными признаками для расширения признакового пространства.

Анализ проблемы

Гипотеза 3	Недостаток гипотезы 3	Возможное решение
Заголовка статьи достаточно для определения требуемых целевых параметров.	Модели на таких данных потребуется определять присутствие паттернов и тенденций. Представленного количества данных может оказаться недостаточно.	Расширить представленную информацию о каждой из статей через парсинг статей с РБК и извлечение дополнительных данных из статей.

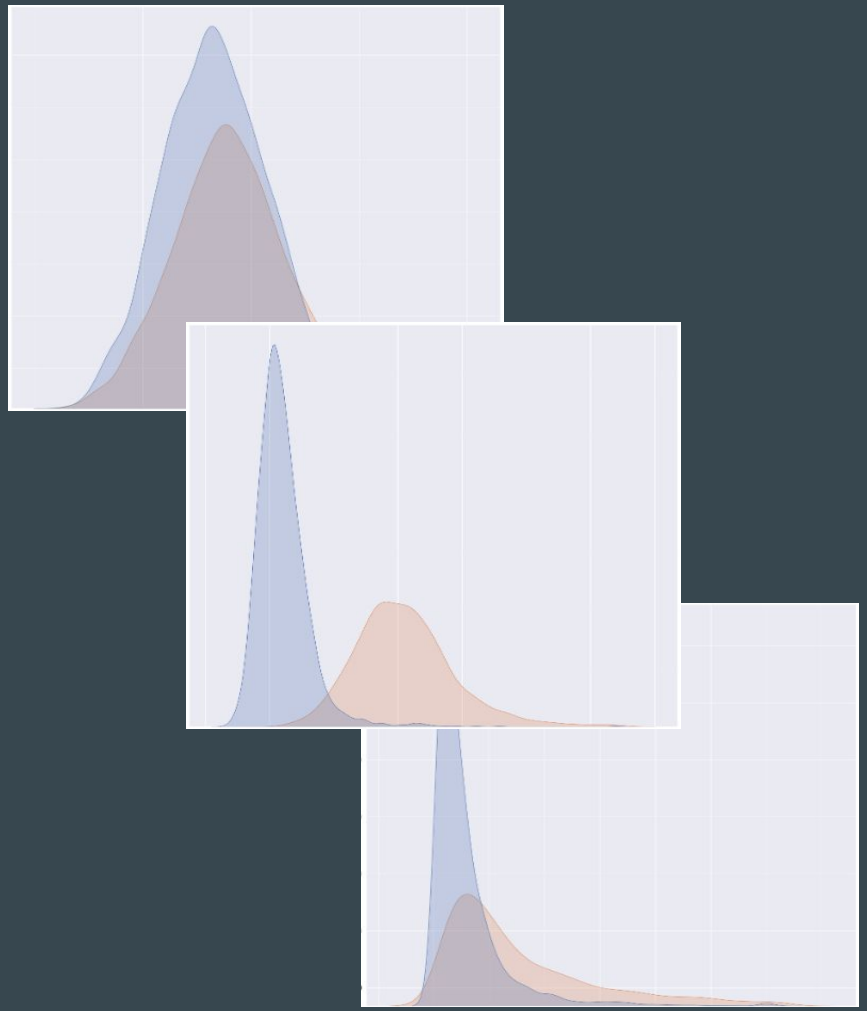
Итоговое решение

Выявлять паттерны в повседневном поведении пользователей.

Обогащать данные статистиками категорий и авторов.

Выявлять пики и провалы статей через NPL (заголовки статей и т.п.)

Решение



В процессе анализа выявил различие в распределениях целевых переменных (статистически не проверялось).

Граница изменения распределений- 2022-04-08.

Принято решение строить отдельные модели для данных до даты 2022-04-08 и после нее.

	title	ctr	text_len	views	depth	full_reads_percent
2424	Эскалация вокруг Украины. Что известно к 14:00	6.096	3191	2554204	1.799	4.978
4183	Байден попросил россиян не бояться США и НАТО....	6.096	3284	2554204	1.799	4.978
5086	Новое обострение ситуации вокруг Украины. Главное	6.096	3284	2554204	1.799	4.978
5634	ДНР и ЛНР объявили эвакуацию в Россию. Главное	6.096	3284	2554204	1.799	4.978
5951	Эскалация вокруг Украины. Что известно к 00:00	6.096	3284	2554204	1.799	4.978
6359	Россия возвращает военных с учений у границ. Г...	6.096	3284	2554204	1.799	4.978

	title	ctr	text_len
945	Новое обострение ситуации вокруг Украины. Главное	6.096	3191
1440	Новое обострение ситуации вокруг Украины. Главное	6.096	3191
2645	Эскалация вокруг Украины. Что известно к 03:00	6.096	3191

В процессе анализа выявил наличие константных значений целевых переменных при признаке ctr = 6.096.

При построении прогноза данные в тесте с ctr = 6.096 в соответственно также получили константные значения.

Парсинг РБК

Произвел парсинг статей с РБК с извлечением дополнительной информации: длина текста, количество изображений и т.п.

В дополнение сформировал ряд бинарных переменных: прямая трансляция, фоторепортаж, инфографика и т.п.

Собрал статистические данные (мин, макс, среднее, ско) по:

- временным параметрам*;
- категориям*;
- авторам;

По временным параметрам и категориям так же собирал лаги и разницу за 7 последних дней, использовал плавающее окно с распределением гаусса в качестве тренда на 2, 3 и 7 дней.

* с разделением на даты до 2022-04-08 и после;

NLP

Использовал NLP модель sbert_large_mt_nlu_ru от sberbank-ai для извлечения эмбеддингов из заголовков.

После чего, для борьбы с переобучением, полученные эмбеддинги длиною в 1024 ужимал через PCA в эмбеддинги длиною в 64.

Модель

Итоговая модель была Catboost.

Наилучшее количество итераций выбиралось по CV на 5 фолдов по средней минимальной RMSE на валидационных фолдах.

Количество итераций и модель выбирались и строились отдельно для каждой целевой переменной и интервалов до 2022-04-08 и после.

Итог

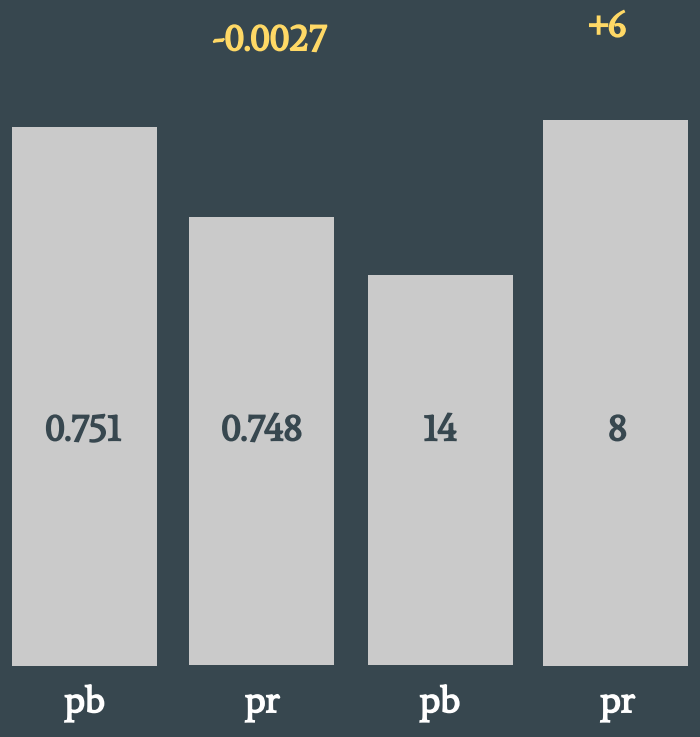
567 признаков

Catboost

Время обучения 214 минут.

Паблик (pb): 0.751067 и 14 место

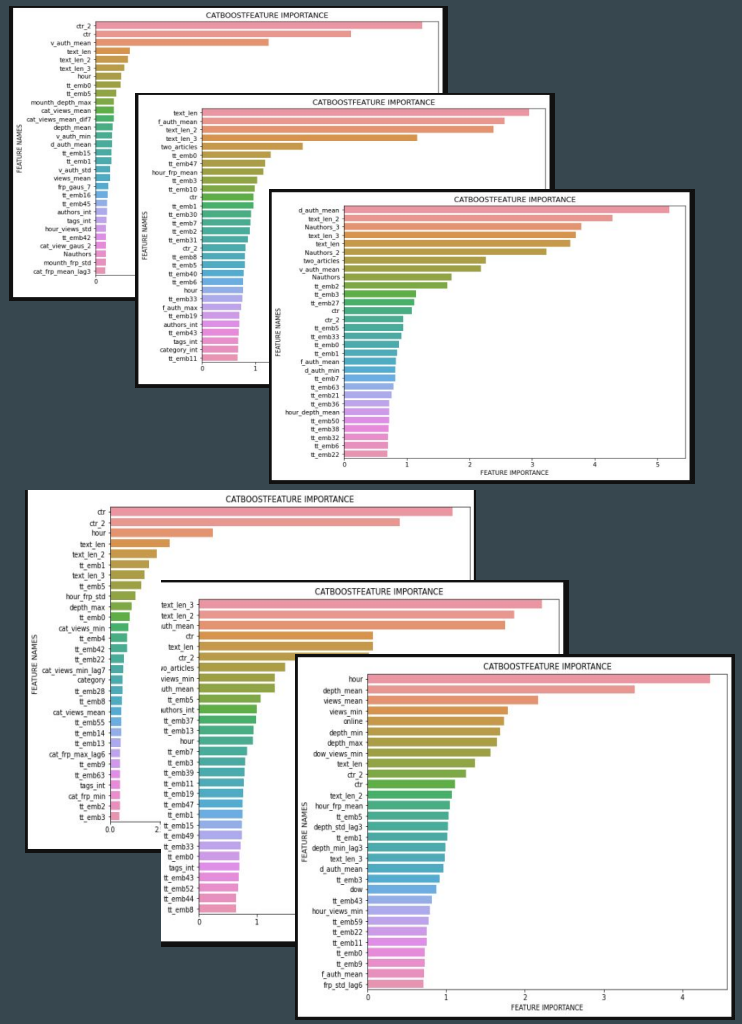
Приват (pr): 0.748285 и 8 место



Итог

Наиболее важными признаками оказались:

- длина текста;
- ctr;
- средние по целевым переменным;
- час публикации статьи;
- эмбеддинги заголовков;



Итог

Подтверждена концепция предсказания популярности новостных статей на основе выявления паттернов с обогащением признаков на основе NLP.

Разработана и протестирована модель на основе такой концепции.

Пространство улучшения

- Более широкое использование NLP в части заголовков и описания статьи;
- Использование эмбеддингов категорий и авторов.

Благодарю за внимание!

Ахременко Владимир