Классификация потока финансовых новостей с целью выявления динамики цен биржевых инструментов

Выпускная квалификационная работа магистра

Кулага Роман Александрович

Московский Физико-Технический Институт (Государственный Университет)
Кафедра «Интеллектуальные системы», ФУПМ

06.06.2018

Актуальность

- Анализ новостей широко используется компаниями, лидирующими в области торговли финансовыми инструментами
- Разрабатываемые инструменты и результаты успешных исследований подвержены NDA

Цель

Создать модель классификации, пригодную для использования в области биржевой торговли.

Задача

Провести исследование и понять, как в данной области можно применить алгоритмы машинного обучения.

- Новостные заголовки из ряда крупных англоязычных источников (в т.ч. «CNN», «CNBC», «Yahoo Finance») с точным временем публикации
- Исторические данные по ценам и объемам сделок с секундной точностью из «Yahoo Finance». Из цен вычитается трендовая составляющая.

Базовая предобработка новостей

1. Удаление пунктуации
2. Приведение к нижнему регистру, лемматизация
3. Фильтрация стоп-слов
4. Токенизация

По множеству новостных заголовков строится коллекция документов D. В качестве объекта $d \in D$ может выступать:

- Разреженный вектор высокой размерности: $d \in \mathbb{R}^{N_{feat}}$
- Вектор категориальных признаков небольшой размерности: $d \in \mathbb{Z}^{N_{cat}}$

Возможные ответы классификатора:

- **Класс «1»**: имеет место движение («рывок») цены
- **Класс «0»**: значительных отклонений цены нет

Формирование объектов и автоматическая разметка производятся несколькими альтернативными способами.

Пусть t_0 — некоторый момент времени, ΔT — длительность временного интервала, $P(t)$, $V(t)$ — цена и объем сделок в момент времени t.

В качестве меры динамичности торговли в интервале между t_0 и $t_0 + \Delta T$ предлагаются:

- $\Delta P_1(t_0, \Delta T) = \max\limits_{\delta t = 1, \ldots, \Delta T} \frac{|P(t_0+\delta t) - P(t_0)|}{P(t_0)}$
- $\Delta P_2(t_0, \Delta T) = \left| \sum\limits_{\delta t=1}^{\Delta T} \frac{[P(t_0+\delta t) - P(t_0)]}{P(t_0)} \cdot V(t_0 + \delta t) \right|$

На практике выбирается $\Delta P(t_0, \Delta T) = \Delta P_1(t_0, \Delta T)$

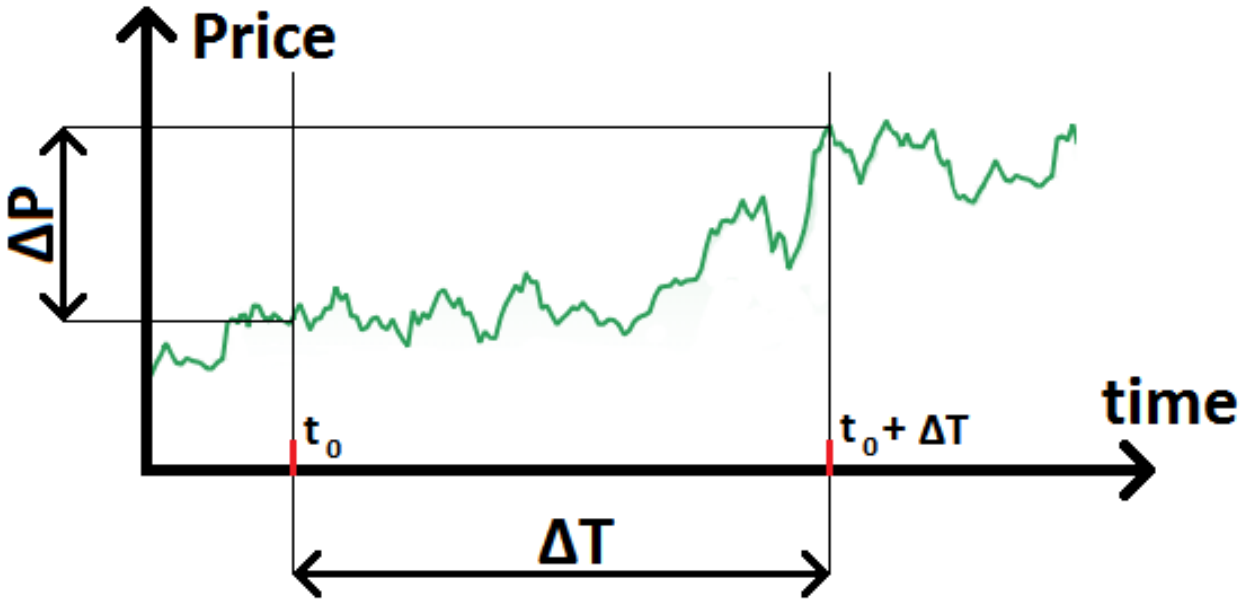


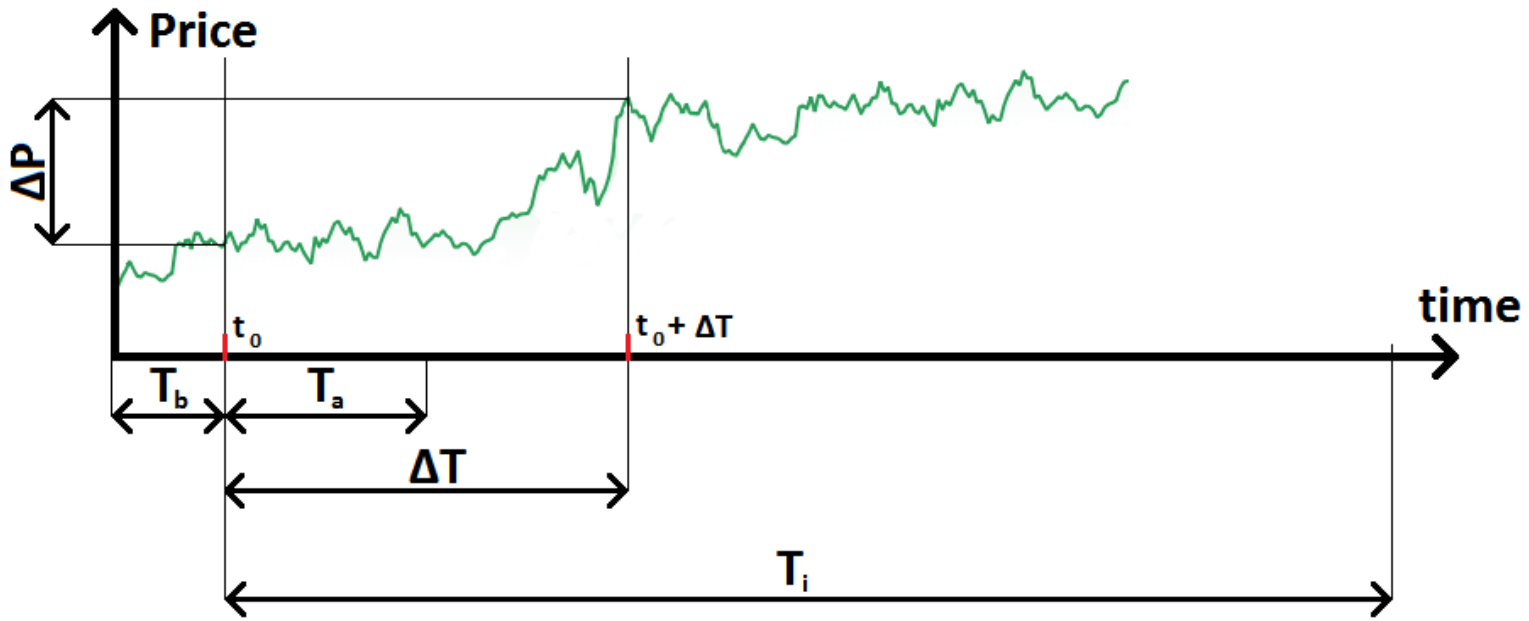
Figure: Наглядная иллюстрация величины рывка $\Delta P(t_0, \Delta T)$

Метрики бинарной классификации

Были выбраны:

- ROC (Receiver Operating Characteristic) кривая и ROC-AUC
- PR (Precision Recall) кривая и PR-AUC

Для получения численных оценок используется кросс-валидация K-Fold.



- Объекты — отдельные новости
- Разметка на классы: обходим коллекцию новостей окном ширины ΔT; если $\Delta P(t_0, \Delta T) > P_{threshold}$, помечаем все новости в интервалах T_b и T_a классом «1» и переносим окно на $T_i + T_b$ вперед, удаляя новости между t_0 и $t_0 + T_i$. Остальные новости помечаем классом «0».

Дальнейшая обработка факторов:

1. Удаляем из признакового описания униграммы w с документной частотой $N_d(w) < N_d^{min}$
2. К оставшимся факторам для каждого документа $d \in D$ применяем TF-IDF взвешивание:
   $tfidf(w,d) = tf(w,d) \cdot idf(w)$, где $tf(w,d)$ - частота w в документе d, $idf(w) = \log(\frac{|D|}{|D_w|})$, $D_w = \{d : w \in d\}$
3. Нормируем вектор каждого документа d

Параллельно строим описание каждого объекта d на основе категориальных признаков d_{cat}:

1. Сортируем униграммы по убыванию $tfidf(w,d)$
2. Берем не более N_{cat} первых и заполняем вектор d_{cat} их токенами, а пустые места — специальным значением -1.

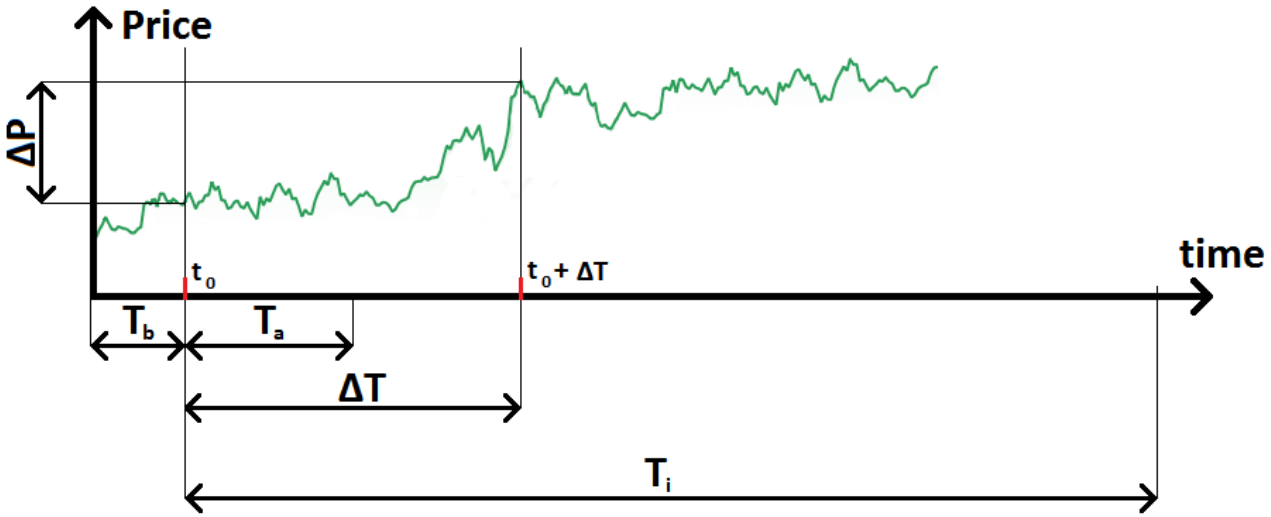
Параметры модели

$P_{threshold} = 3.5\sigma(\Delta T)$, где $\sigma(T)$ — волатильность данного инструмента на интервале времени T.
$N_{cat} = 25$
Временны́е параметры разбиты на две конфигурации:

- $\Delta T = 10$ мин, $T_b = 3$ мин, $T_a = 6$ мин, $T_i = 60$ мин
- $\Delta T = 3$ мин, $T_b = 3$ мин, $T_a = 6$ мин, $T_i = 60$ мин

Результаты эксперимента

ΔT	N_{obj}	N_1/N_{obj}	Алгоритм	ROC-AUC	PR-AUC
10	1186095	0.164	Random Forest	0.58	0.24
			LogRegression	0.58	0.23
			Linear SVM	0.58	0.23
			XGBoost	0.61	0.26
			CatBoost	0.60	0.31
3	937689	0.259	Random Forest	0.58	0.40
			LogRegression	0.59	0.40
			Linear SVM	0.59	0.40
			XGBoost	0.63	0.44
			CatBoost	0.64	0.48



- Объекты — агрегация новостей в интервале $T_b + T_a$
- Порождение объектов: обходим коллекцию новостей окном ширины ΔT; если $\Delta P(t_0, \Delta T) > P_{threshold}$, агрегируем все новости в интервалах T_b и T_a в один объект и помечаем его классом «1», затем переносим окно на $T_i + T_b$ вперед.
- Из случайных неразмеченных интервалов длиной $T_b + T_a$ формируем новые объекты и помечаем классом «0»

Параметры модели

$P_{threshold} = 3.5\sigma(\Delta T)$, где $\sigma(T)$ — волатильность данного инструмента на интервале времени T.

$N_{cat} = \mathbf{60}$

Используем **только** CatBoost Classifier

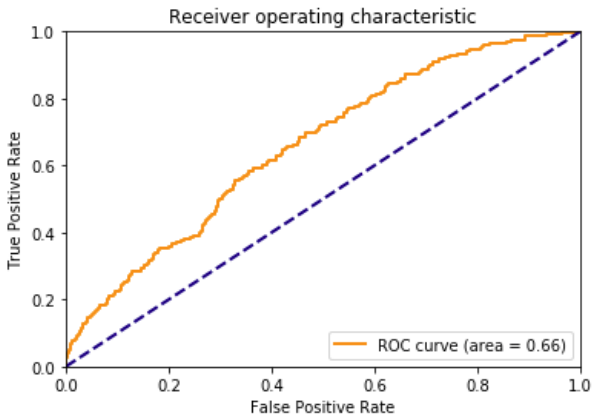
Временны́е параметры разбиты на две конфигурации:

- $\Delta T = 10$ мин, $T_b = 3$ мин, $T_a = 6$ мин, $T_i = 60$ мин
- $\Delta T = 3$ мин, $T_b = 3$ мин, $T_a = 6$ мин, $T_i = 60$ мин

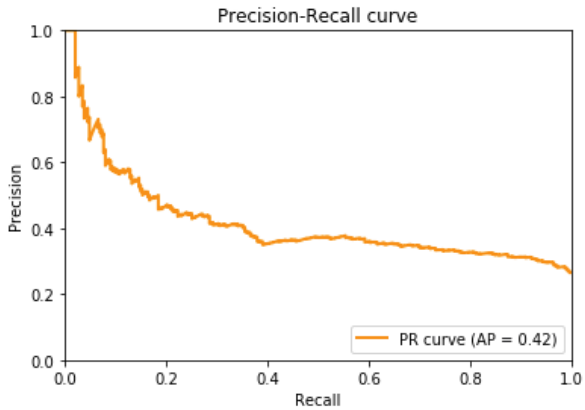
Результаты эксперимента

В таблице приведены результаты для CatBoost Classifier:

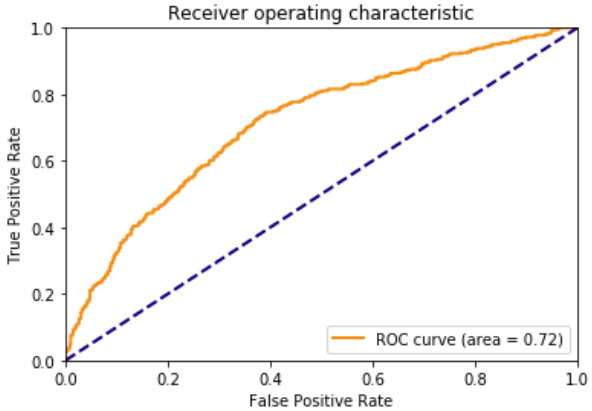
ΔT	N_{obj}	N_1	N_1/N_{obj}	N_{cat}	ROC-AUC	PR-AUC
10	4838	1338	0.277	60	0.66	0.42
3	5696	1696	0.298	60	0.72	0.54



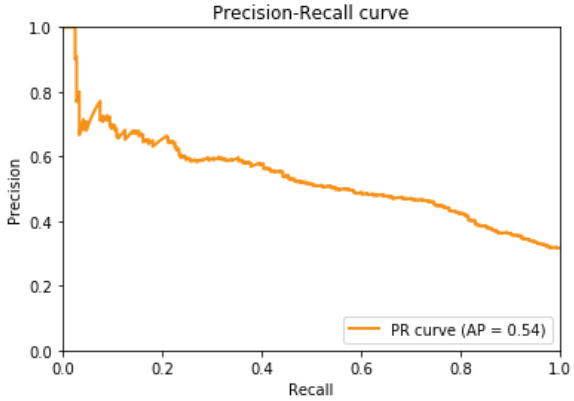
(a) ROC, $\Delta T = 10$ мин



(b) PR, $\Delta T = 10$ мин



(c) ROC, $\Delta T = 3$ мин



(d) PR, $\Delta T = 3$ мин

Битермы

Определение

Битермы представляют собой все различные пары слов в рамках одной фразы (в нашем случае - коминации слов внутри одного новостного заголовка).
Например, предложение вида «aaa bbb ccc ddd» содержит в себе битермы «aaa; bbb», «aaa; ccc», «aaa; ddd», «bbb; ccc», «bbb; ddd», «ccc; ddd».

- Использование битермов требует тщательной предварительной фильтрации.

Significance Score

Дальнейшая фильтрация производится с помощью Significance Score.

Определение

Пусть две униграммы A и B образуют битерм $(A;B)$, $f(A,B)$ — случайная величина, характеризующая документную частоту битерма, $\mu_0(A,B)$ — мат ожидание величины $f(A,B)$ в предположении нулевой гипотезы, что термы A и B встречаются в тексте независимо.
Будем говорить, что $sig(A,B) = \frac{f(A,B)-\mu_0(A,B)}{\sqrt{f(A,B)}}$ — значение **Significance Score** для битерма $(A;B)$.

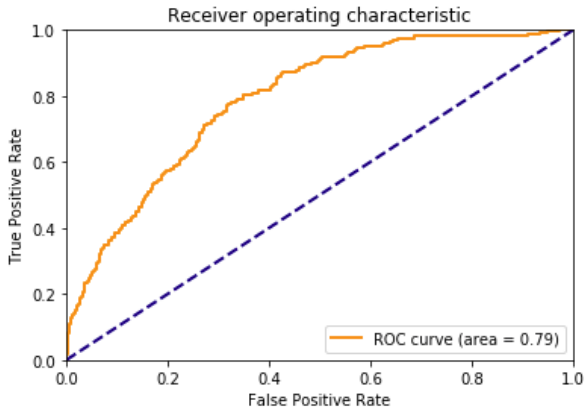
Построение категориальных факторов

1. Фильтруются битермы с низкой документной частотой: $N_d(b) < N_d^{min}$
2. Фильтруются неинформативные битермы $b = (A; B)$, для которых $sig(b) = sig(A, B) < S^{min}$
3. Формируются агрегированные документы
4. Формируются категориальные факторы: битермы внутри документа сортируются по убыванию $sig(b)$ и берутся не более первых N_{cat}

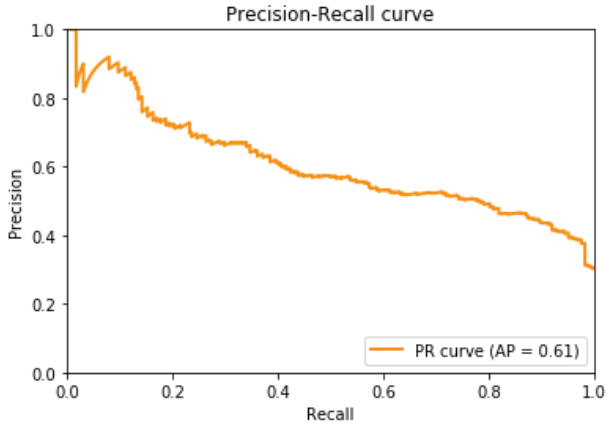
Результаты эксперимента

В таблице приведены результаты для CatBoost Classifier:

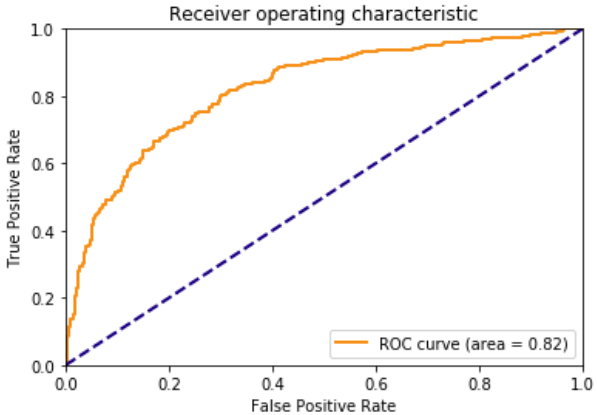
ΔT, мин	N_{obj}	N_1	N_1/N_{obj}	N_{cat}	ROC-AUC	PR-AUC
10	4838	1338	0.277	60	0.79	0.61
3	5696	1696	0.298	60	0.82	0.72



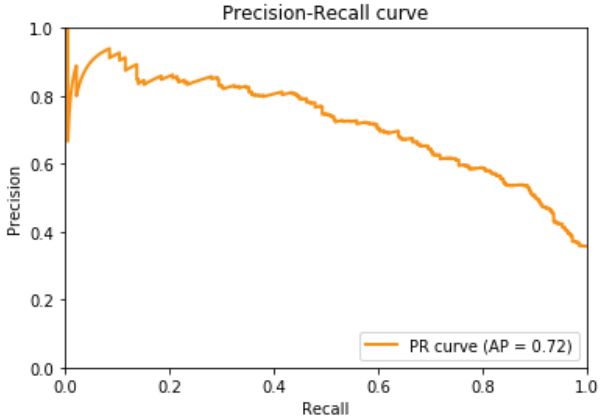
(e) ROC, $\Delta T = 10$ мин



(f) PR, $\Delta T = 10$ мин



(g) ROC, $\Delta T = 3$ мин



(h) PR, $\Delta T = 3$ мин

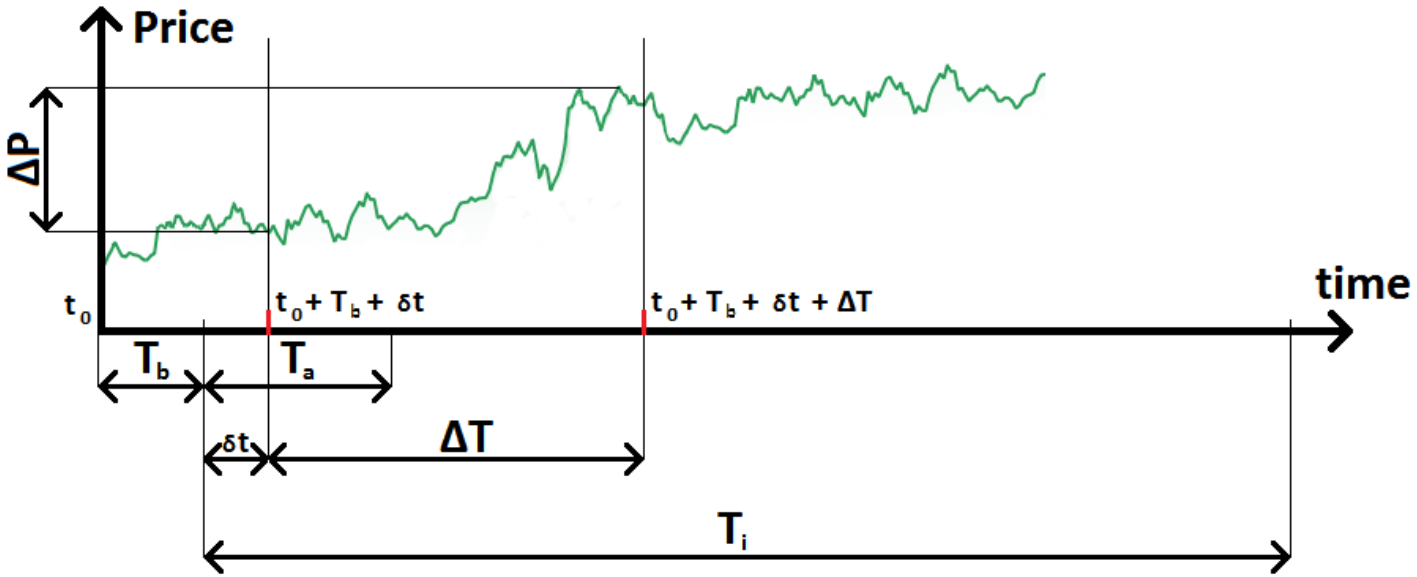
Задача регрессии

Предлагается разбить модель классификации на 2 этапа:

1. Модель регрессии предсказывает величину $\Delta\hat{P}(t_0, \Delta T)$
2. Пороговым правилом результат относится либо к классу «1», если $\Delta\hat{P}(t_0, \Delta T) > P_{threshold}$, либо к классу «0» иначе.

В качестве алгоритма регрессии выбран CatBoost Regressor. Признаковое описание — категориальные признаки на основе битермов.

Автоматическая разметка величины рывка



Введем новый временной параметр τ, задающий границы для небольшого сдвига δt.

Пусть $\delta P(t_0, \Delta T) = \max\limits_{\delta t \in [-\tau, \tau]} [\Delta P(t_0 + T_b + \delta t, \Delta T)]$

Формирование объектов

Обход новостей производится не окном, а непересекающимися интервалами длительностью $T_b + T_a$ с началом в t_0:

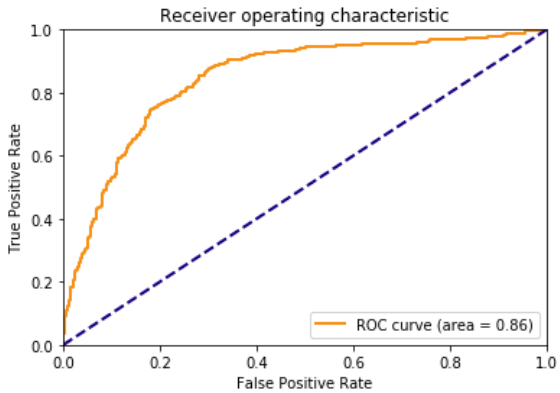
1. Вычисляем $\delta P(t_0, \Delta T)$
2. Агрегируем новости в интервале $T_b + T_a$ и присваиваем новому объекту соответствующий ответ $\delta P(t_0, \Delta T)$
3. Если $\delta P(t_0, \Delta T) > P_{threshold}$, начало очередного рассматриваемого интервала переносим на время не менее $t_0 + T_b + T_i$, иначе — на следующий момент времени после $t_0 + T_b + \max(T_a, \tau + \Delta T)$

Затем считаем $P_{0.3}$ — 30-ый перцентиль по всем $\delta P(t, \Delta T)$ и удаляем те объекты, для которых $P_{0.3} \leq \delta P(t, \Delta T) \leq P_{threshold}$

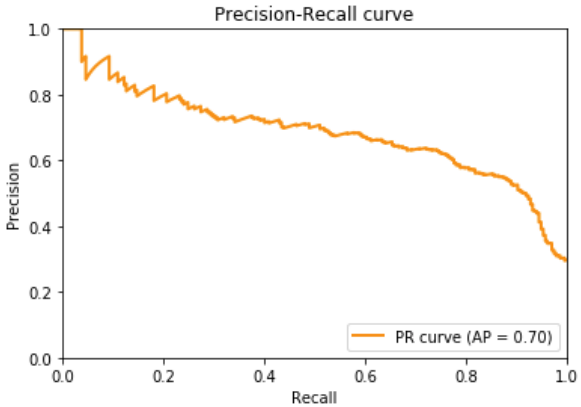
Результаты эксперимента

В таблице приведены результаты метрик качества для модели как классификатора в целом:

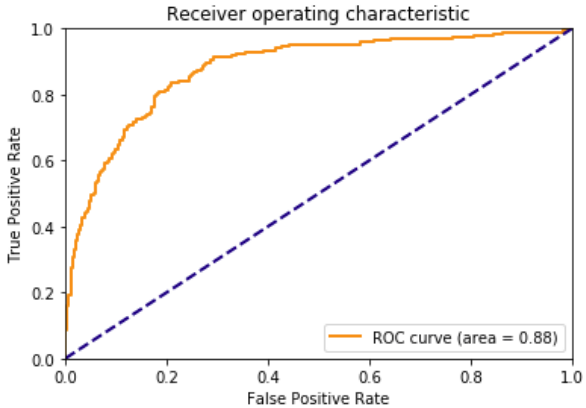
ΔT	τ, с	N_{obj}	N_1	N_1/N_{obj}	N_{cat}	ROC-AUC	PR-AUC
10	20	3799	1014	0.267	60	0.86	0.70
3	20	4016	1160	0.289	60	0.88	0.74



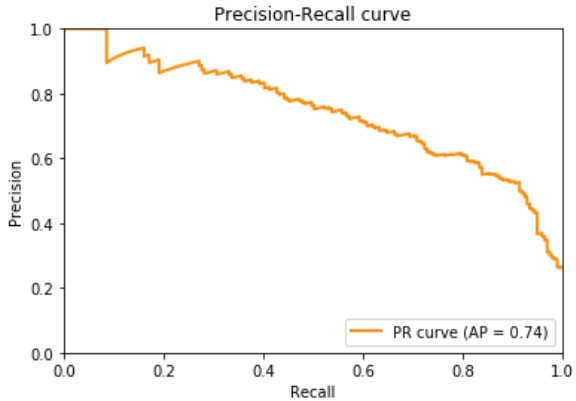
(i) ROC, $\Delta T = 10$ мин



(j) PR, $\Delta T = 10$ мин



(k) ROC, $\Delta T = 3$ мин



(l) PR, $\Delta T = 3$ мин

Заключение

Результаты, полученные в данной работе:

1. Поставлена общая задача классификации и предложено практическое применение результатов в рамках предметной области.
2. Предложен и реализован ряд моделей признакового описания объектов, в том числе с применением агрегации новостных заголовков и на основе битермов.
3. Проведены соответствующие эксперименты, подтверждающие улучшение качества классификации на каждом этапе модификации модели.
4. В результате построена модель, значительно превосходящая базовые и тривиальные решения по качеству классификации.