Классификация потока финансовых новостей с целью выявления динамики цен биржевых инструментов

Кулага Роман Александрович

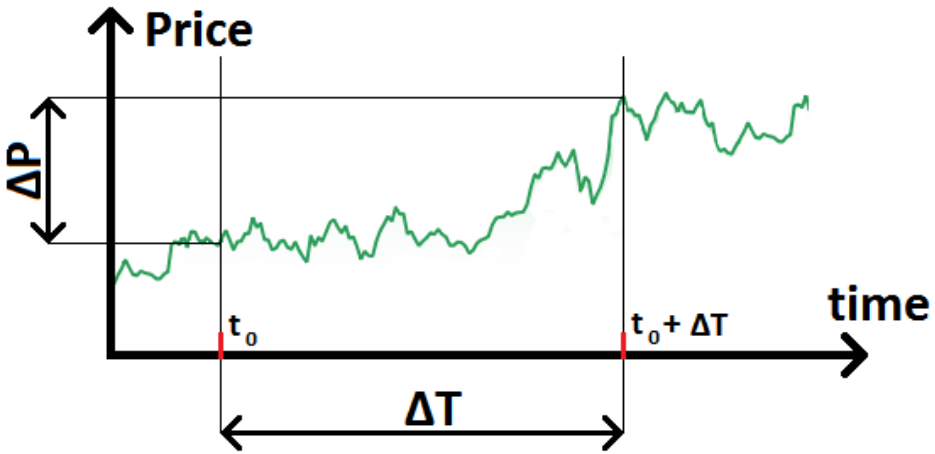
Московский Физико-Технический Институт (Государственный Университет)
Кафедра «Интеллектуальные системы», ФУПМ
Научный руководитель: д.ф.-м.н. К.В. Воронцов

Москва 2018

Определение понятия скачка цены

Пусть t_0 — момент времени внутри торговой сессии (например, время выхода новости), ΔT — длительность временного интервала, $P(t)$ — цена инструмента в момент времени t.
Введем функцию, характеризующую величину скачка цены в интервале между t_0 и $t_0 + \Delta T$:

$$\Delta P(t_0, \Delta T) = \max_{t=t_0,\ldots,t_0+\Delta T} \frac{|P(t) - P(t_0)|}{P(t_0)}$$



Исходные данные и их предобработка

Источники данных

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

Предобработка текста

1. Удаление пунктуации
2. Разбиение на отдельные слова
3. Приведение к нижнему регистру
4. Лемматизация
5. Фильтрация стоп-слов
6. Токенизация

Формирование обучающей выборки

В качестве объекта может выступать:

- Отдельно взятая новость
- Объект, полученный агрегированием новостей за определенный временной интервал

Каждому объекту выборки автоматической разметкой ставится в соответствие:

- В модели классификации — класс («1», если наблюдается скачок цены и «0» иначе)
- В модели регрессии — величина скачка цены $\Delta P(t, \Delta T)$

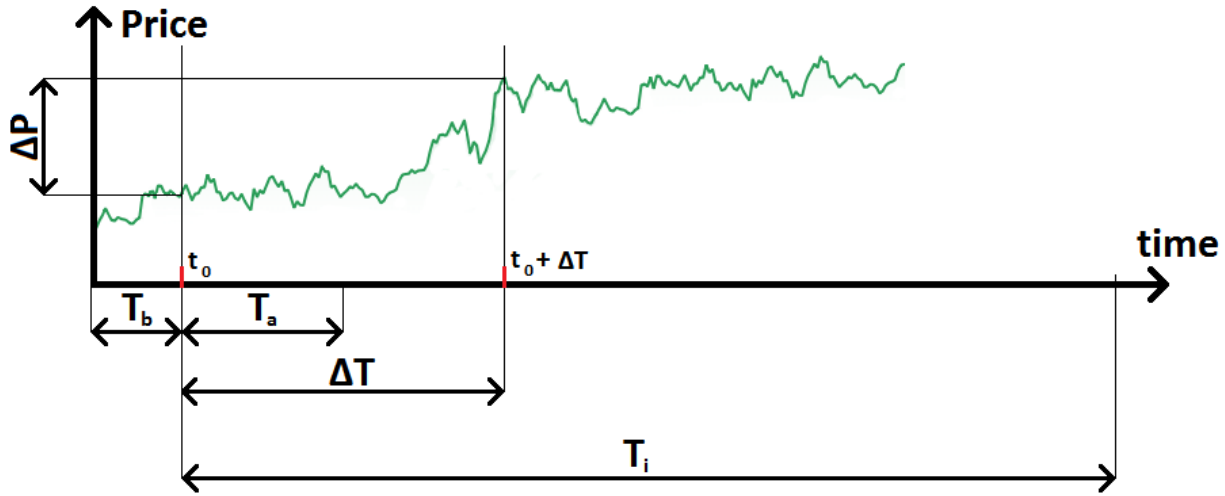
Метрики бинарной классификации

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.

Базовая модель



Объекты (документы) — отдельные новости.
Формирование обучающей выборки:

1. Обходим все новостные заголовки; пусть t_0 - время публикации текущей новости
2. Если $\Delta P(t_0, \Delta T) > P_{threshold}$, помечаем все новости в интервалах T_b и T_a классом «1» и удаляем из выборки новости между $t_0 + T_a$ и $t_0 + T_i$.
3. Оставшиеся новости помечаем классом «0».

Признаковое описание

Обработка факторов:

1. Удаляем слова w с низкой документной частотой: $N_d(w) < N_d^{min}$
2. Для каждого документа $d \in D$ применяется TF-IDF взвешивание

Формируем категориальные признаки d_{cat} каждого документа d:

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.

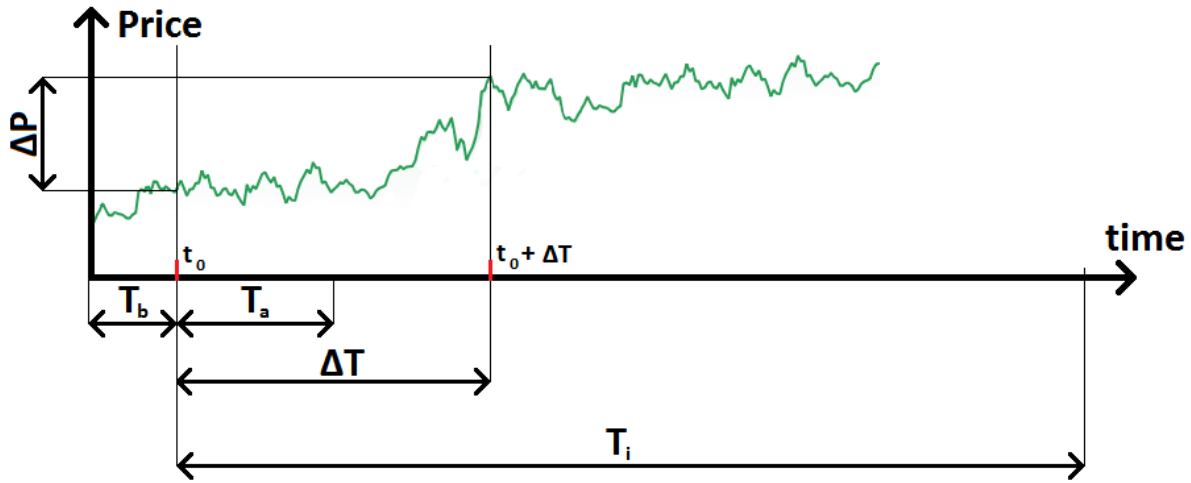
Результаты эксперимента

Общие параметры: $T_b = 3$ мин, $T_a = 5$ мин, $T_i = 60$ мин; $N_{cat} = 25$.
Для ширины окна ΔT рассматриваем две конфигурации:

- $\Delta T = 10$ мин (имеет предсказательную силу)
- $\Delta T = 3$ мин (не имеет предсказательной силы)

ΔT, мин	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**

Модель с агрегированием документов

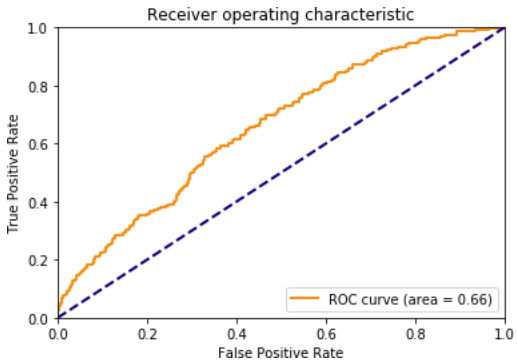


Формирование объектов:

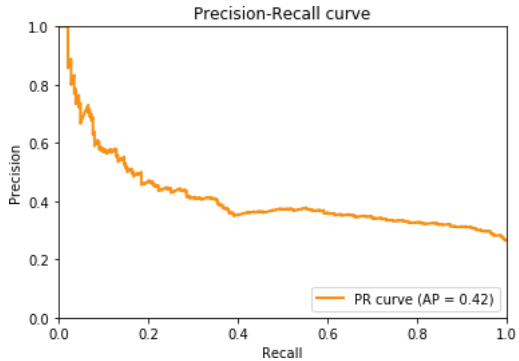
1. Обходим всю коллекцию новостей; пусть t_0 — время публикации текущей новости.
2. Если $\Delta P(t_0, \Delta T) > P_{threshold}$, агрегируем все новости в интервалах T_b и T_a в один объект, помечаем его классом «1» и удаляем из выборки новости между $t_0 + T_a$ и $t_0 + T_i$.
3. Из случайных неразмеченных интервалов длиной $T_b + T_a$ формируем новые объекты и помечаем классом «0».

Результаты эксперимента (Catboost Classifier)

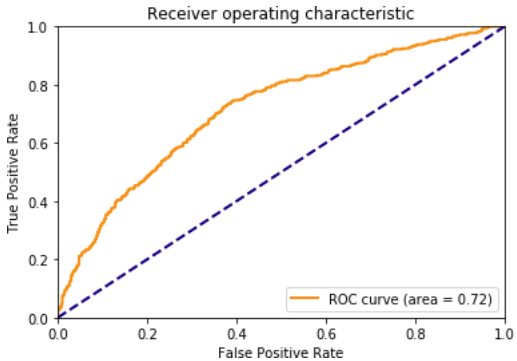
ΔT, мин	N_{obj}	N_1	N_1/N_{obj}	N_{cat}	ROC-AUC	PR-AUC
10	4838	1338	0.277	60	**0.66**	**0.42**
3	5696	1696	0.298	60	**0.72**	**0.54**



(a) ROC, $\Delta T = 10$ мин



(b) PR, $\Delta T = 10$ мин



(c) ROC, $\Delta T = 3$ мин

(d) PR, $\Delta T = 3$ мин

Модель на основе битермов

Определение битермов

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

Определение Significance Score

Пусть две униграммы A и B образуют битерм $(A;B)$, $f(A,B)$ — документная частота битерма, а $\mu_0(A,B)$ — мат ожидание случайной величины $f(A,B)$ в предположении, что термы A и B встречаются вместе 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

Фильтруются битермы b:

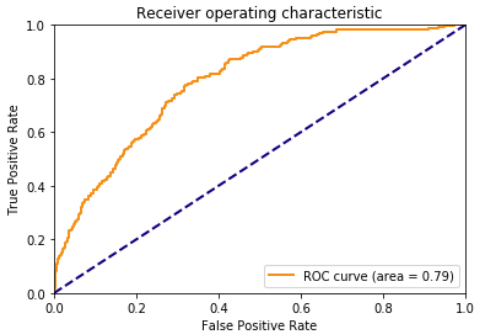
1. С низкой документной частотой: $N_d(b) < N_d^{min}$
2. Неинформативные: $sig(b) < S^{min}$

Формирование категориальных факторов:

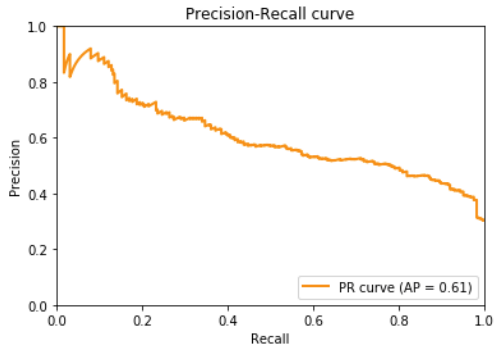
1. Сортируем битермы внутри документа по убыванию $sig(b)$
2. Берем не более N_{cat} первых

Результаты эксперимента (CatBoost Classifier)

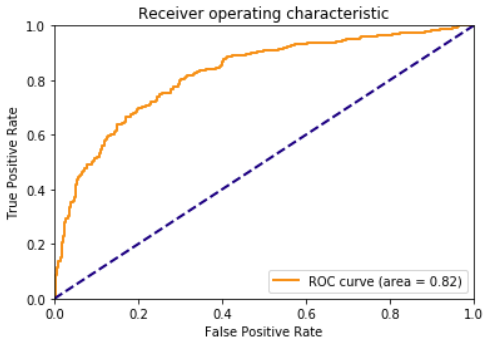
ΔT, мин	N_{obj}	N_1	N_1/N_{obj}	N_{cat}	ROC-AUC	PR-AUC
10	4838	1338	0.277	60	**0.79**	**0.61**
3	5696	1696	0.298	60	**0.82**	**0.72**



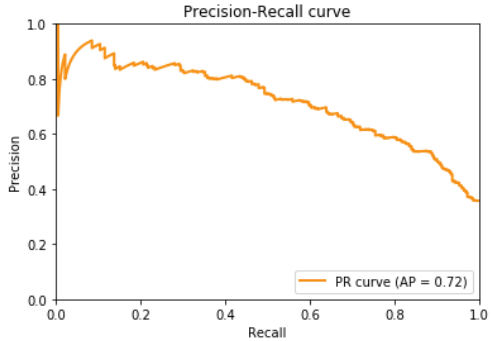
(e) ROC, $\Delta T = 10$ мин



(f) PR, $\Delta T = 10$ мин



(g) ROC, $\Delta T = 3$ мин



(h) PR, $\Delta T = 3$ мин

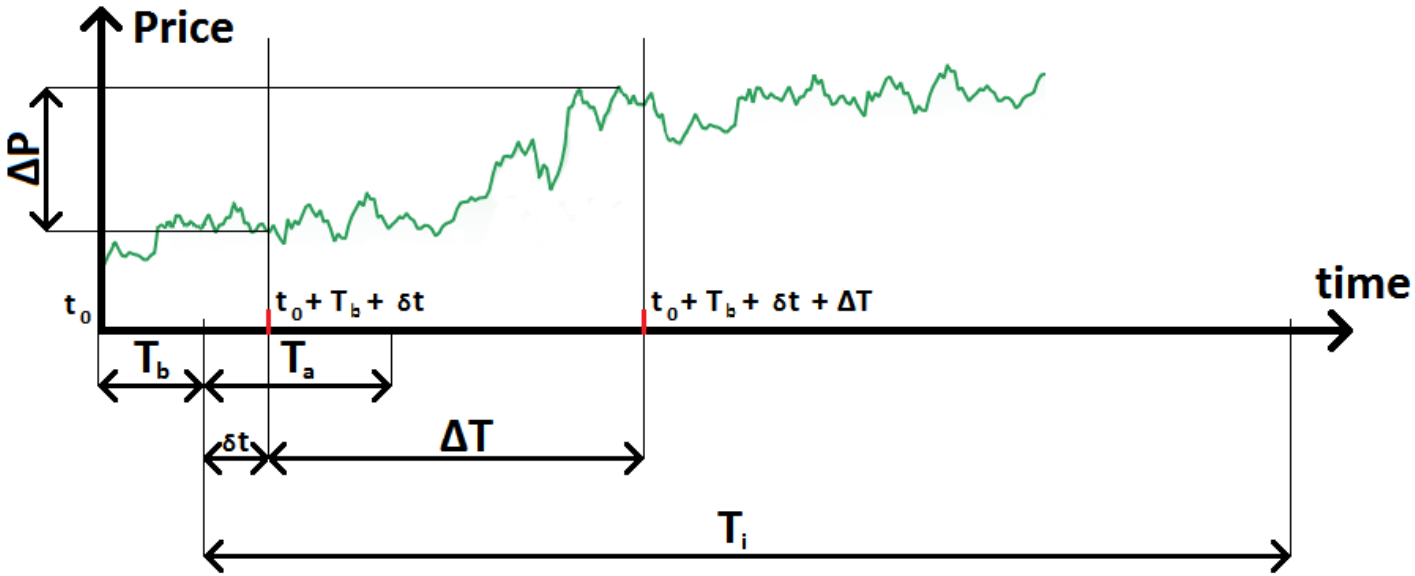
Модель регрессии

Предлагается разбить модель классификации документа на 2 этапа:

1. Модель регрессии предсказывает для документа величину скачка цены $\Delta\hat{P}(t_0, \Delta T)$
2. Документ относится к классу «1», если $\Delta\hat{P}(t_0, \Delta T) > P_{threshold}$, иначе — к классу «0».

Признаковое описание — категориальные признаки на основе битермов.

Автоматическая разметка величины скачка цены



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

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

Формирование объектов

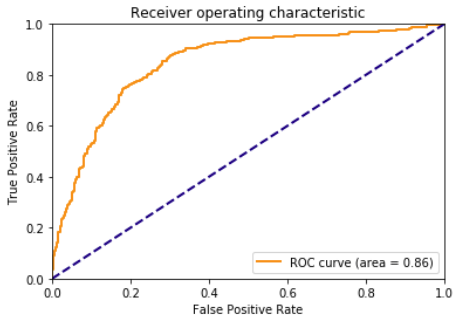
Обходим все новости; пусть t_0 - время публикации текущей новости:

1. Агрегируем новости в интервале $T_b + T_a$
2. Присваиваем новому объекту ответ $\delta P_\tau(t_0, \Delta T)$
3. Если $\delta P_\tau(t_0, \Delta T) > P_{threshold}$, удаляем из выборки все новости между $t_0 + T_a + T_b$ и $t_0 + T_b + T_i$

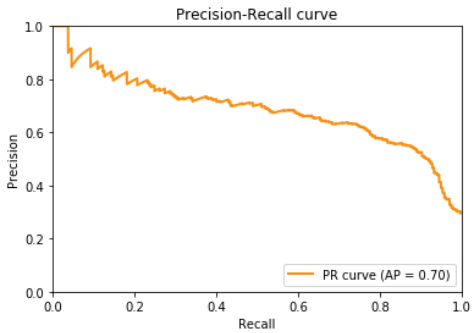
Затем считаем $P_{0.3}$ — 30-ый процентиль по всем $\delta P_\tau(t, \Delta T)$ — и удаляем объекты, для которых $P_{0.3} \leq \delta P_\tau(t, \Delta T) \leq P_{threshold}$

Результаты эксперимента (Catboost Regressor)

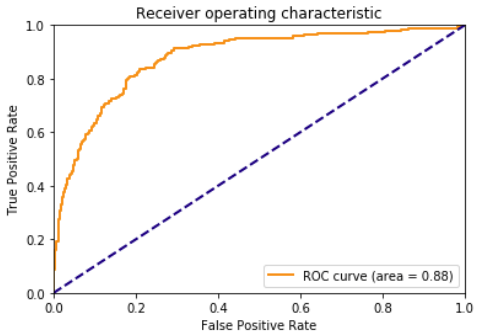
ΔT, мин	τ, с	N_{obj}	N_1	N_1/N_{obj}	N_{cat}	ROC-AUC	PR-AUC
10	20	3799	1014	0.267	60	**0.86**	**0.70**
3	20	4016	1160	0.289	60	**0.88**	**0.74**



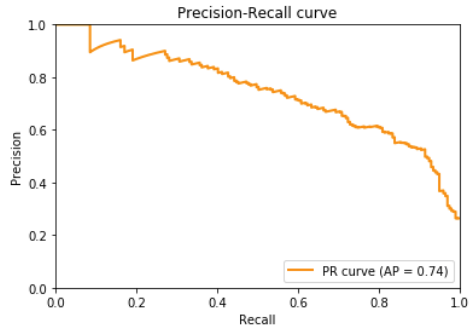
(i) ROC, $\Delta T = 10$ мин



(j) PR, $\Delta T = 10$ мин



(k) ROC, $\Delta T = 3$ мин



(l) PR, $\Delta T = 3$ мин

Результаты, выносимые на защиту

1. Предложена формализация задачи предсказания скачка цены финансового инструмента по новостному потоку.
2. Предложены три модели (на основе агрегирования документов, на основе битермов, а также модель регрессии), показано их преимущество по сравнению с базовой моделью.
3. Сделана реализация, пригодная для практической эксплуатации.