МОВС-2022
ВШЭ

Предсказание волатильности валютного рынка Московской Биржи методами NLP

Вячеслав Бучков

Эдуард Ачикян

Татьяна Панютина

Куратор:

Светлана Щербакова

https://github.com/v-buchkov/hse_volatility_project

06.03.2023

СОДЕРЖАНИЕ

ИДЕЯ ПРОЕКТА

Волатильность = Неопределённость

- Волатильность выражает неопределённость участников рынка по поводу будущих цен
- Участники рынка выражают свои мысли и ожидания **при помощи текстов, в которых можно проследить их настрой**

Волатильность => опционы

- Для проверки гиптоезы №1 мы используем данные по опционам и спотовым ценам
- **Идея проекта — построить устойчивый прогноз для реализации настоящей торговой стратегии**

Кризис => выше неэффективность

- В 2022 году были введены ограничения на доступ к рынкам РФ для нерезидентов
- Количество институционалов, торгующих на рынках РФ, стало меньше => **больше стратегий для построения предиктов**

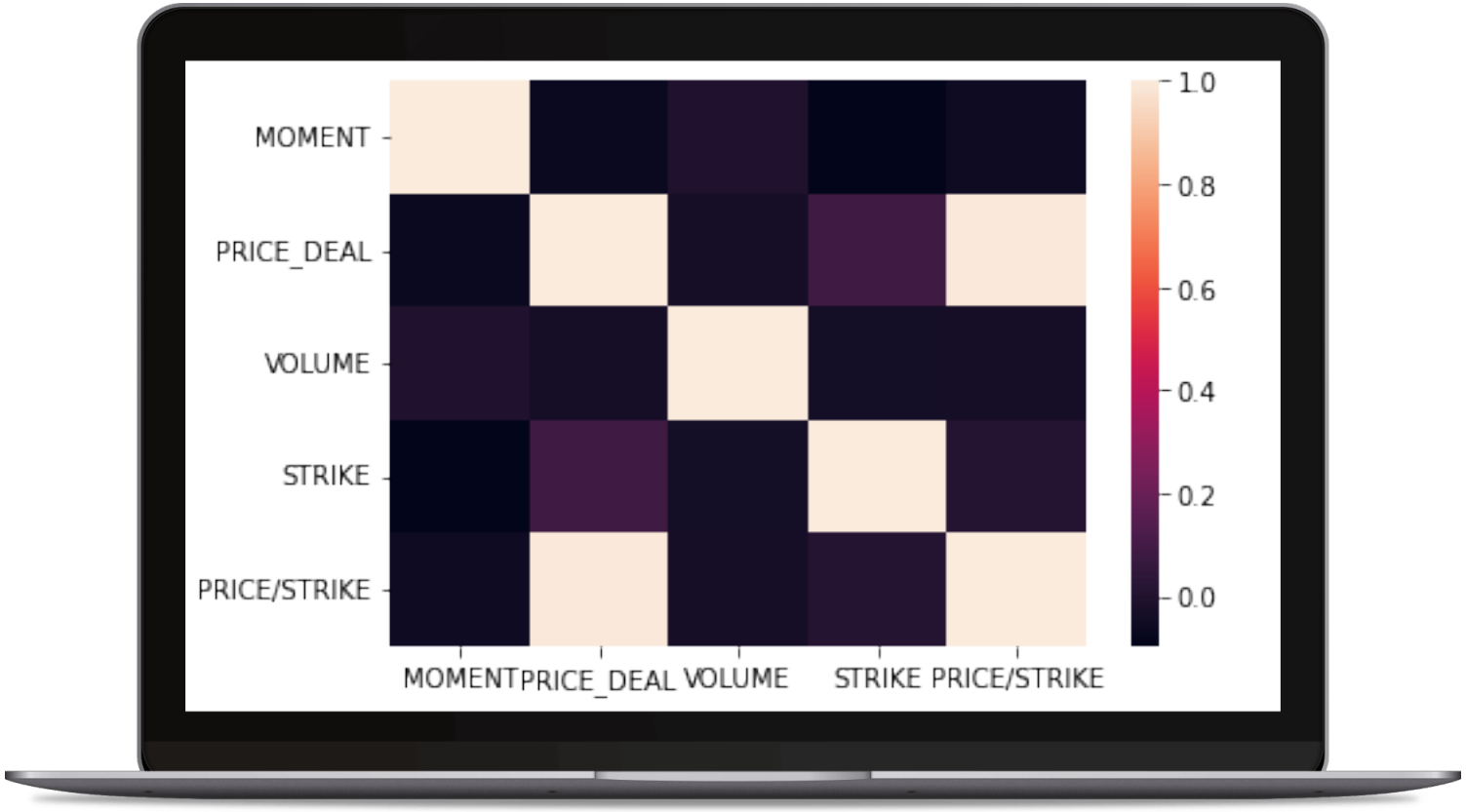
РЕШАЕМ ЗАДАЧУ КЛАССИФИКАЦИИ ЗНАКА PNL ПО ВЫБРАННОЙ ОПЦИОННОЙ ПОЗИЦИИ

* Также можно решать и задачу регрессии, но для реальной стратегии размер ожидаемого PnL важен уже не так сильно

ДАННЫЕ ОПЦИОНОВ

Используем данные Московской Биржи и оффшорных источников цен

- Источники: MOEX FORTS, MOEX FX, An offshore price source
- Период с 01.01.2020 по 01.11.2022
- Собраны данные по USDRUB, EURUSD, CNHRUB, а также всем сделкам по опционам на MOEX
- **Проведён EDA для данных опционов — построены основные метрики для моделирования (анализ для разного % moneyness)**



МОДЕЛИРОВАНИЕ ОПЦИОНОВ

- В самой простой реализации мы **решаем проблему с пропущенными точками в ценах опционов**
- Используем **построение динамического дельта-хеджирования** на x днях, симулируя реальный опцион
- Дальнейший план — **построение факторной позиции для обособления эффекта волатильности** (гамма- и вега- стратегии)

ДАННЫЕ ТЕКСТОВ ИЗ ПОЧТЫ

Используем открытые данные новостей и аналитических записок из почтовых рассылок

- Источники : 'Открытие', 'Movchan', 'ATON RESEARCH', 'Bloomberg', 'ROSBANK Research'
- Рассматривался период с 06.05.2020 по 03.11.2022
- Собрано 1305 писем
- **Главный вывод: из-за упадка в экономике, вызванной пандемией, в новостях чаще говорили про ставку (rate) и инфляцию (inflation)**

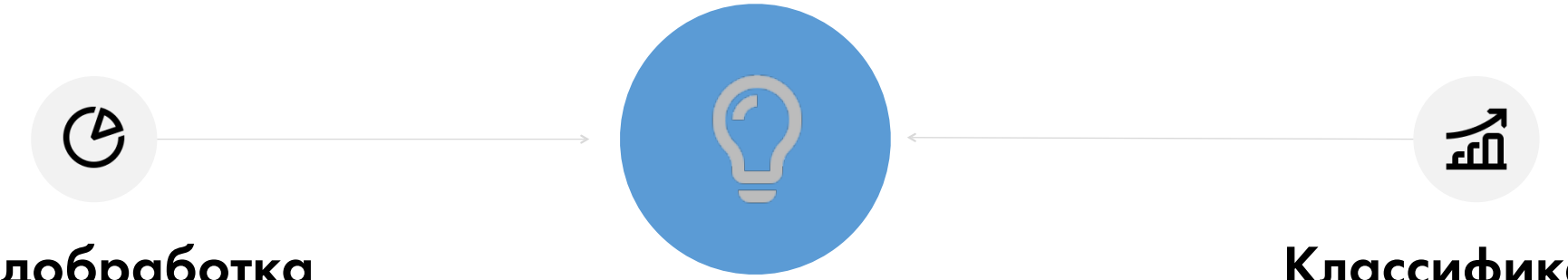
ДАННЫЕ ТЕКСТОВ ИЗ ТЕЛЕГРАМ И САЙТОВ

Используем открытые данные телеграм-каналов и новостей с сайтов

- В ходе проекта были собраны данные 11 телеграм каналов и новости с сайта tinkoff
- Собрана вся история постов с начала ведения канала или сайта
- В рамках EDA проведена лемматизация слов и выявлены наиболее употребляемые слова среди существительных, глаголов и прилагательных по всем собранным текстам, а также частота употребления ключевых слов из подготовленного списка
- **Получили готовый датасет для использования методов предобработки**

ПРОЦЕСС ЭКСПЕРИМЕНТОВ

В процессе экспериментов мы работали с предобработкой текстов и разными классификаторами



Предобработка

- sklearn’s CountVectorizer
- sklearn’s TfidfVectorizer
- gensim’s Word2Vec (from scratch)
- pretrained GloVe (“glove-twitter-100”)
- pretrained FastText (“cc.ru.300.bin”)

Основные идеи

- Классифицируем каждый текст как “к какому знаку прибыли он привёл на выбранном интервале”
- Перебираем не только модели, но и валютные пары, а также инструменты (опцион и спот)

Классификатор

- sklearn’s LogisticRegression
- sklearn’s RandomForestClassifier
- CatBoost
- LightGBM
- FastText embedded classifier

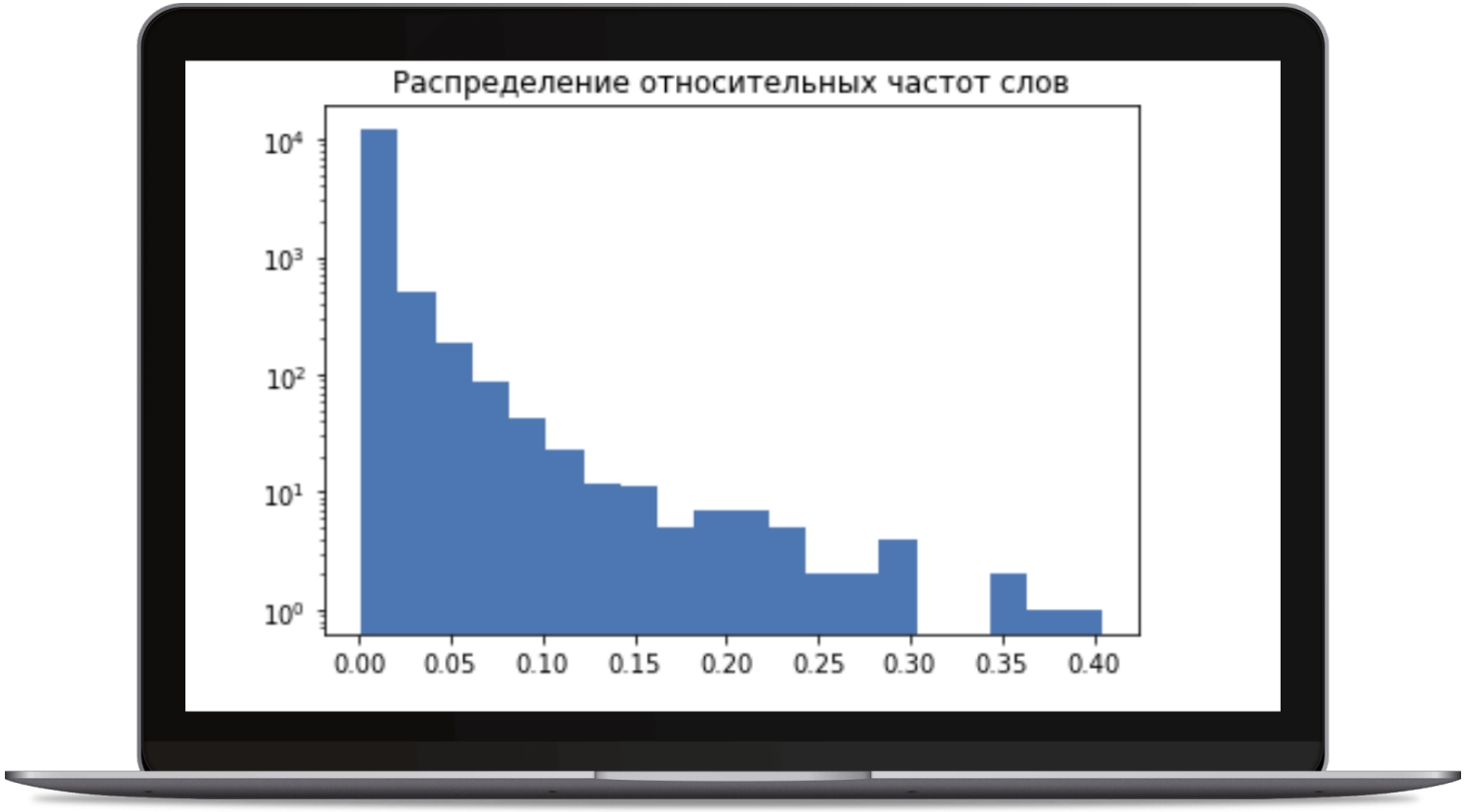
ПРЕДОБРАБОТКА ТЕКСТОВ

Используем открытые данные новостей и аналитических записок из почтовых рассылок

- Токенизируем тексты, выбрасывая stopwords("russian") по nltk

- Был проведён эксперимент по замене чисел на слова (например, "+10%" => "большой", чтобы "рост +10%" был заменён на "рост большой"). Использование замен не помогло

- Пороговая частота слов определялась экспериментально. Были выбраны min_count=5, max_df=0.8



ПОСТРОЕНИЕ МОДЕЛИ

Как мы строили датасет?

	date_start	pnl	pnl_sign	cumm_text
0	2022-03-02	-8967.190672	0	‼Банк России принял решение не возобновлять ...
1	2022-03-03	-12548.743088	0	‼ РОССИЯ – РЕЙТИНГ – МУСОР Это, конечно, по...
2	2022-03-04	-15486.634276	0	‼- Банк России принял решение не возобновля...
3	2022-03-09	12709.989084	1	Прошедшая неделя кардинально изменила жизнь к...

- date_start - дата входа в позицию, закрытие которой происходит через 5 дней
- pnl - исторический PnL за 5 дней
- pnl_sign - знак PnL
- cumm_text - текст собранный за дату date_start и за предыдущие даты до последней сделки

Как мы строили саму модель?

- Источники выбирали жадным перебором, выкидывая из наиболее полного датасета
- Предобработка текстов
- Обучение модели
- Перебор на валидации (для бустингов использовали Optuna)

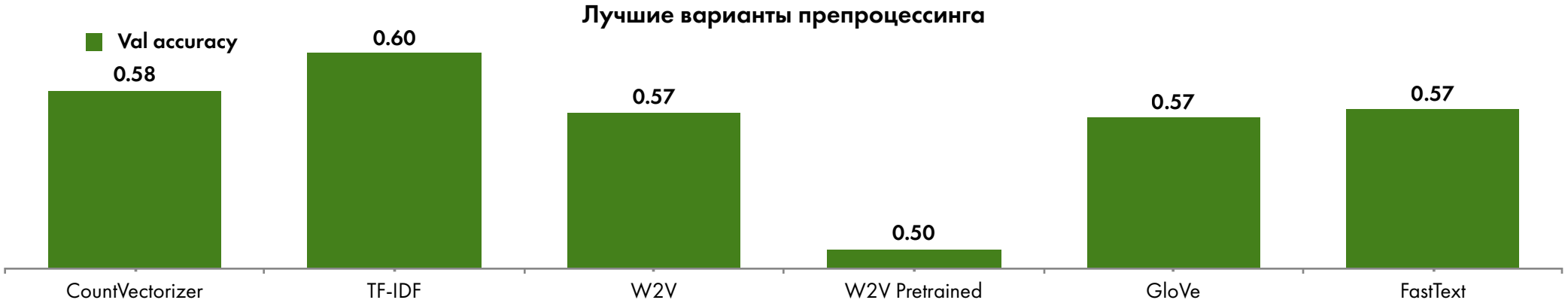
Обучили 19 моделей поочерёдно.

ПОЛУЧИЛИ ГОТОВЫЙ ПАЙПЛАЙН ДЛЯ ЭКСПЕРИМЕНТОВ

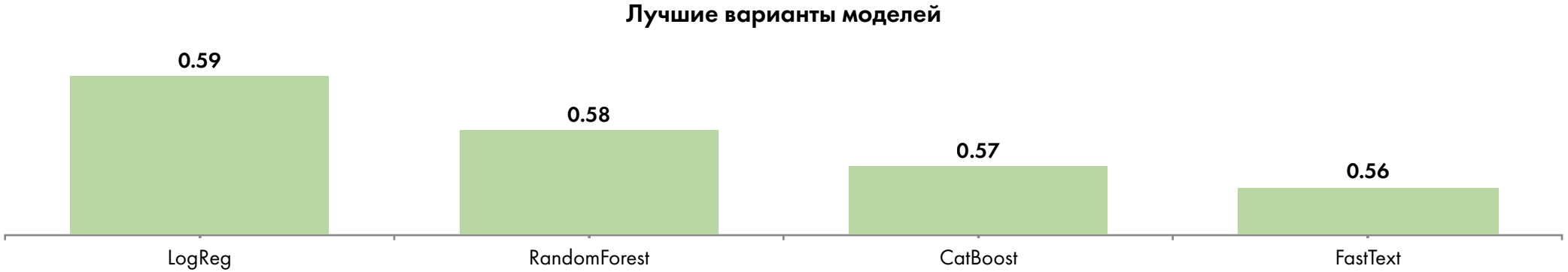
РЕЗУЛЬТАТЫ

ЛУЧШАЯ МОДЕЛЬ ДЛЯ КАЖДОГО АКТИВА

- Используем accuracy, так как выборка целевой переменной сбалансирована



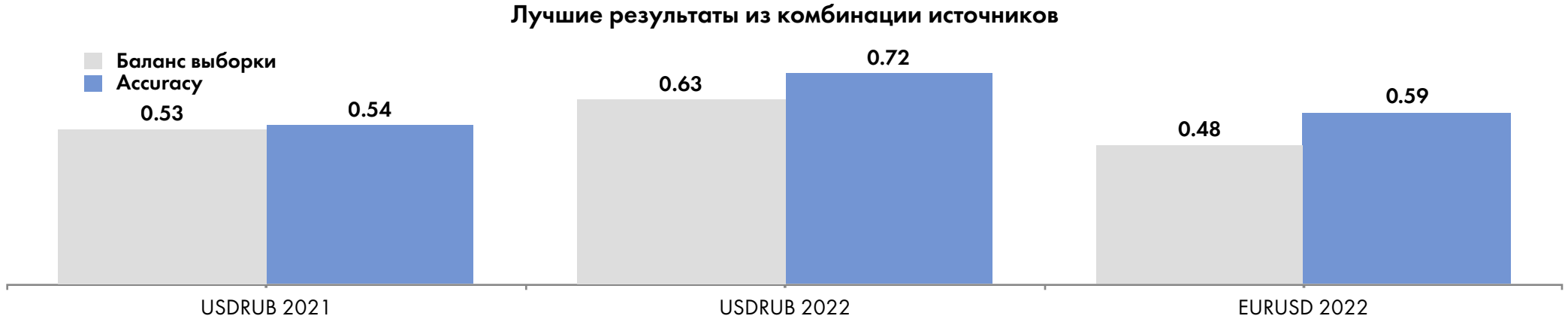
- Модели показывают интересную предсказательную силу, но далёкую от идеальной из-за шума



ЛУЧШИЙ ВАРИАНТ = TF-IDF + LOGREG

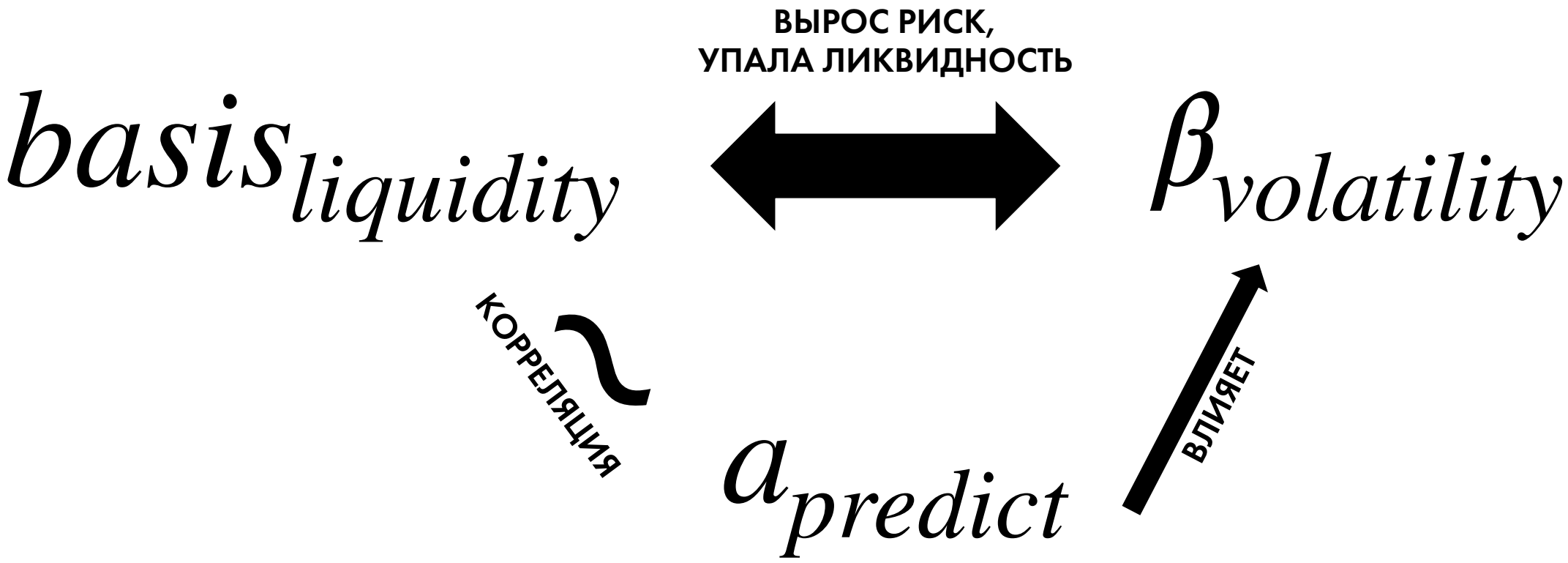
РЕЗУЛЬТАТЫ ЛУЧШЕЙ МОДЕЛИ

- Используем accuracy, так как выборка целевой переменной сбалансирована



- Модели показывают интересную предсказательную силу, но далёкую от идеальной из-за шума

ОБЪЯСНЕНИЕ РЕЗУЛЬТАТОВ



Неэффективность рынка выросла => предсказания улучшились

ЗНАЧИМОСТЬ & ДАЛЬНЕЙШИЕ ШАГИ

Значимость полученных результатов

- Увидели, что в простейших линейных моделях есть предсказание для знака PnL, то есть можно построить работающую торговую стратегию
- Нашли признаки того, что неэффективность (то есть возможность для предсказания) на российском рынке выросла с момента введения ограничений
- Построили готовую базу кода для дальнейших экспериментов с моделями, чтобы можно было сфокусироваться только на ML&DL частях

Дальнейшее исследование & Улучшение результатов

- Необходимо построить факторные стратегии для исключения эффекта влияния других финансовых переменных, кроме волатильности
- Нужно учитывать веса каждого источника данных, чтобы придавать больший вес наиболее надёжным источникам
- Включить больше реалистичных деталей в бэктест — в том числе, сделать классификацию на 3 класса, создав класс, где PnL “близок к нулю”

СПАСИБО ЗА ВНИМАНИЕ!

ДАЛЬШЕ — БОЛЬШЕ:)