



Факультет Экономических наук
Направление «Прикладная Экономика»

Предсказание волатильности фондового рынка с использованием данных социальных сетей

Подготовили:
Ан Е.
Миткинов В.
Шайдуллин А.

Научный руководитель:
Мамедли Мариам Октаевна



I. ВВЕДЕНИЕ

Цели и актуальность

Прикладная экономика

Цель

Построить модель прогнозирования волатильности фондового рынка на один день вперед методами машинного обучения, используя факторы на основе данных социальной сети Twitter.

В качестве основной модели была использована гетерогенная авторегрессия реализованной волатильности (The Heterogeneous Autoregressive model of the Realized Volatility, HAR-RV, HAR).

Актуальность

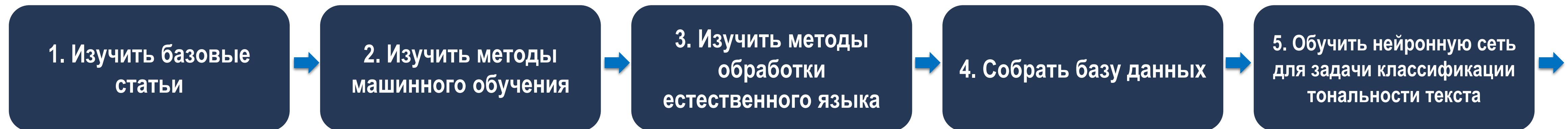
1. Необходимость снижения риска финансовых потерь методом построения прогнозов на основе различных баз данных.
2. Рост популярности социальных сетей, в т.ч. и как средства получения информации (баз данных).
3. Наличие корреляции между настроениями в социальных сетях и стоимостью акций компаний.



I. ВВЕДЕНИЕ

Задачи

Прикладная экономика



- Audrino F., Sigrist F., Ballinari D. The impact of sentiment and attention measures on stock market volatility //International Journal of Forecasting. – 2020.

- Giuseppe B., Paola C., Juri M., Giancarlo N. Twitter Sentiment and Banks' Financial Ratios: Is There Any Causal Link? – 2018.

Мы использовали более продвинутую модель для классификации текстов.

Мы анализировали российские компании, брали другой набор факторов.

Мы использовали не только регрессионную модель, но и более продвинутые модели.

- Stepik: Программирование на Python; KarrovCourses Аналитик данных.

- Coursera: Специализация – Машинное обучение и анализ данных.

- Open Data Science: Open Course "Machine Learning and Deep Learning".

- Соревнования по машинному обучению на платформе Kaggle.

- ВШЭ, ФКН: Курс «Машинное обучение».

3. Изучить методы обработки естественного языка

- MIPT Deep Learning School: курс по методам обработки естественного языка (NLP).

- Данные по 5-ти минутным внутридневным ценам акций были собраны с сайта investing.com.
- Данные для экономических и финансовых показателей были собраны с сайтов: investing.com, cb.ru. Данные ежедневные.

- Данные для факторов на основе социальных сетей были собраны с Twitter с помощью библиотеки Snsccape в python.

Все данные находятся в открытом доступе.

5. Обучить нейронную сеть для задачи классификации тональности текста

- Нейронная сеть для определения тональности текста была построена с помощью библиотеки Fast.ai в python. Работа проводилась в платной среде Google collab.

- Перед обучением классификатора была дообучена языковая модель на основе собранных нами данных из Twitter и других данных, находящихся в открытом доступе: Корпус Юлии Рубцовой, корпус SentiRuVal 2016.

- Обучена модель для прогнозирования тональности текста ULMFiT (архитектура этой модели - AWD-LSTM).

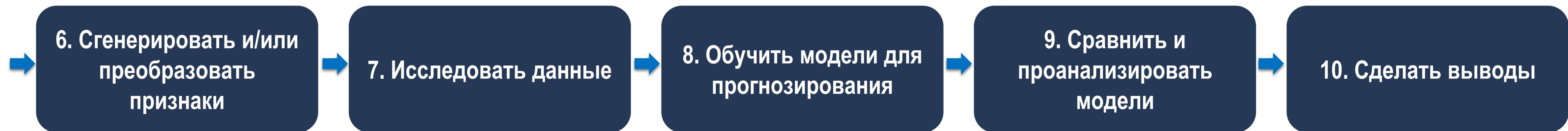




I. ВВЕДЕНИЕ

Задачи

Прикладная экономика



- Реализованная волатильность была вычислена с помощью оценки медианной реализованной волатильности (MedRV). Средние реализованные волатильности за неделю, месяц были усреднены за 5, 22 предыдущих дня соответственно.
 - Макроэкономические факторы были линейно интерполированы, т.к. доступны только с месячной периодичностью.
 - Факторы внимания на основе данных социальных сетей вычислялись как кол-во постов за день, факторы настроения – получены из нейронной сети.
 - Проведена описательная статистика всех переменных.
 - Проанализирована корреляция между целевой переменной и всеми признаками, корреляция между всему признаками друг с другом.
 - Визуализированы распределения всех признаков.
 - Отобран конечный набор признаков для построения модели прогнозирования.
 - Определена методика подбора гиперпараметров для моделей.
 - Проведена предобработка данных.
 - Построена базовая модель, экономическая модель и модель настроения.
 - Построен прогноз реализованной волатильности.
 - Приведена экономическая интерпретация.
 - Проанализировано качество прогноза каждой модели.
 - Проведен сравнительный анализ каждой модели.
 - Проведен анализ важности признаков, в т.ч. с помощью библиотеки Shap в python.
 - Выбрана наилучшая модель для каждой компании.
 - Результаты показали, что данные социальных сетей влияют на целевую переменную и могут улучшить качество прогноза.
- Использованные библиотеки в python: Numpy, Pandas, Matplotlib, SciKit-Learn, XGBoost, LightGBM, Seaborn, Statistics, StatsModels, Csv, SciKit Optimize.



I. ВВЕДЕНИЕ

Обзор литературы

Прикладная экономика

Автор	Год	Основная работа	Основной тезис
Fama E. F.	1970	Efficient capital markets: A review of theory and empirical work. <i>The Journal of Finance</i> , 25(2), 383–417.	Согласно гипотезе эффективного рынка: прогнозирование доходности акций не должно быть возможным, поскольку рыночные цены будут отражать всю доступную информацию.
Daniel, Hirshleifer и Teoh	2002	Investor psychology in capital markets: evidence and policy implications. <i>Journal of Monetary Economics</i> , 49(1), 139–209.	Сообщают о растущем количестве эмпирических данных, показывающих, что фондовый рынок управляет психологией инвесторов.
Tseng K.C.	2006	Behavioral finance, bounded rationality, neurofinance, and traditional finance. <i>Investment Management and Financial Innovations</i> , 3(4), 7–18.	Попытка дать объяснение психоэмоциональным факторам в экономике («поведенческие финансы»): неправильная интерпретация причинно-следственной связи, нарушение логической последовательности и др.
Sun L, Najand, M., & Shen, J.	2016	Stock return predictability and investor sentiment: A high-frequency perspective. <i>Journal of Banking & Finance</i> , 73, 147–164.	Анализируют предсказуемость 30-минутной доходности с помощью индекса Thomson Reuters MarketPsych - показателя настроений, основанного на новостных лентах, источниках новостей в Интернете и социальных сетях. Их результаты показывают, что изменения в настроениях инвесторов могут предсказывать доходность акций в течение дня.
Antweiler & Frank	2004	Is all that talk just noise? the information content of internet stock message boards.	Более скептически относятся к предсказательной способности социальных сетей. Анализ сообщений, размещенных на Yahoo!, показал, что онлайн - сообщения не имеют экономически значимой предсказательной силы для доходности акций. То есть, факторы настроения, включенные в модель с экономическими весами, не в достаточной степени влияют на ситуацию на фондовом рынке. Проще говоря, факторы настроения оказались незначимыми.
Oliveira et al.	2013	On the predictability of stock market behavior using StockTwits sentiment and posting volume.	



I. ВВЕДЕНИЕ

Обзор литературы

Прикладная экономика

Автор	Год	Основная работа	Основной тезис
Wang Y.-H., Keswani, A., & Taylor, S. J.	2006	The relationships between sentiment, returns and volatility. International Journal of Forecasting, 22(1), 109–123.	Подчеркивают важность контроля над другими экономическими и финансовыми переменными при анализе влияния запаздывающих настроений на реализованную волатильность. Но при этом, их исследование показывает, что индикаторы настроений, мало влияют на будущую волатильность после учета запаздывающей доходности.
Mao, H., Counts, S., & Bollen, J.	2011	Predicting financial markets: Comparing survey, news, Twitter and search engine data (arXiv preprint, arXiv:1112.1051).	Сравнили различные источники данных о настроениях (социальные сети, новости и данные поисковых систем) для прогнозирования доходности, объема и подразумеваемой волатильности, хотя и без учета финансовых ковариат. Они считают, что как доходность акций, так и подразумеваемая волатильность имеют статистически значимую связь с прошлым объемом поиска в Google и настроениями в Twitter.
Ho, K.-Y., Shi, Y., & Zhang, Z.	2013	How does news sentiment impact asset volatility? Evidence from long memory and regime-switching approaches. The North American Journal of Economics and Finance, 26, 436–456.	Используя набор данных по конкретным компаниям и макроэкономическим новостям, обнаружили, что новостные настроения оказывают значительное влияние на внутридневную волатильность ряда отдельных акций США (особенно это оказалось актуальным для акций компаний, действовавших в сфере IT).
Aouadi et al. Dimpfl and Jank Hamid and Heiden	2013 2015 2015	Investor attention and stock market activity: Evidence from France. Can internet search queries help to predict stock market volatility? Forecasting volatility with empirical similarity and Google Trends.	Их исследования показывают, что объем поисковых запросов можно использовать для прогнозирования волатильности фондового рынка. На основе оценок настроений, рассчитанных для финансовых статей, полученных с платформы NASDAQ, обнаружили, что рост настроений влияет на волатильность.



II. ДАННЫЕ Компании

Прикладная экономика

Исследуемый период: с 01.03.2016 по 28.02.2021.

Мотивация выбора:

1. Акции, торгуемые на Московской бирже из списка голубых фишек.
2. Компании разных отраслей, т.к. переменные на основе данных социальных сетей могут по-разному влиять на будущую реализованную волатильность в зависимости от сферы компании.

Все данные **ежедневные**.

Отрасль	Компания
IT	Яндекс
Банки	Сбербанк
Телеком	МТС
Нефть	Лукойл Роснефть
Газ	Газпром Новатэк

Яндекс  СБЕР  МТС  ЛУКОЙЛ 



ГАЗПРОМ  НОВАТЭК 



II. ДАННЫЕ

Факторы

Прикладная экономика

Реализованная волатильность
(3)

Экономические и финансовые факторы
(31)

Факторы настроения и внимания
(9)

Реализованная волатильность в
предыдущий день

Факторы фондового рынка
(8)

Факторы внимания к фондовому рынку
(2)

Средняя реализованная волатильность за
неделю

Факторы рынка облигаций
(5)

Факторы внимания к исследуемой компании
(1)

Средняя реализованная волатильность за
месяц

Факторы валютного рынка
(5)

Факторы настроения к фондовому рынку
(4)

Факторы ликвидности
(2)

Факторы настроения в исследуемой
компании
(2)

Макроэкономические факторы
(11)



II. ДАННЫЕ

Реализованная волатильность

Прикладная экономика

Для расчета дневной реализованной волатильности использована **оценка медианной реализованной волатильности** (MedRV), основанная на 5-ти минутной доходности акций, так как она показывает хорошую устойчивость к скачкам и «нулевой» доходности:

$$\text{MedRV}_N = \frac{\pi}{6 - 4\sqrt{3} + \pi} \left(\frac{N}{N - 2} \right) \sum_{i=2}^{N-1} \text{med}(|\Delta Y_{i-1}|, |\Delta Y_i|, |\Delta Y_{i+1}|)^2 = RV_t^{(d)}(1)$$

$Y = \{Y_t\}_{0 \leq t \leq 1}$, Y_t – логарифм цены в дискретный момент времени t внутри одного дня. $N + 1$ – количество наблюдений цен за день в дискретном наборе точек $0 \leq t_0 \leq \dots \leq t_N \leq 1$. $\Delta Y_i = Y_{t_i} - Y_{t_{i-1}}$, $i = 1, \dots, N$ – доходность для каждого 5-минутного интервала.

Таблица 1. Признаки реализованной волатильности. Источники данных: *investing.com*. Таблица построена автором.

Фактор	Расчет
Реализованная волатильность в предыдущий день	$\log RV_t^{(d)}$
Средняя реализованная волатильность за неделю	$\frac{1}{5} \sum_{i=1}^5 \log RV_{t-i+1}^{(d)}$
Средняя реализованная волатильность за месяц	$\frac{1}{22} \sum_{i=1}^{22} \log RV_{t-1+i}^{(d)}$



II. ДАННЫЕ

Экономические и финансовые факторы

Прикладная экономика

Таблица 2. Экономические и финансовые факторы. Источники данных: investing.com, cbr.ru. Таблица построена автором.

	Категория	Фактор	Общий / специфичный
1	Факторы фондового рынка	Индекс IMOEX	Общий
2	Факторы фондового рынка	Индекс MSCI Russia	Общий
3	Факторы фондового рынка	Доходность акции, %	Специфичный
4	Факторы фондового рынка	Объем торгов, тыс. руб.	Специфичный
5	Факторы фондового рынка	Индекс RVI	Общий
6	Факторы фондового рынка	Изменение в Индексе RVI, %	Общий
7	Факторы фондового рынка	Dow Jones Industrial Average	Общий
8	Факторы фондового рынка	Индекс РТС	Общий
9	Факторы рынка облигаций	Доходность облигации Россия 10-летние (годовая ставка), %	Общий
10	Факторы рынка облигаций	Доходность облигации Россия годовые (годовая ставка), %	Общий
11	Факторы рынка облигаций	Дневное изменение в доходности облигации Россия 10-летние (годовая ставка), %	Общий
12	Факторы рынка облигаций	Дневное изменение в доходности облигации Россия годовые (годовая ставка), %	Общий
13	Факторы рынка облигаций	Разница между доходностью 10-летней облигации и годовой	Общий
14	Факторы валютного рынка	Курс доллара к рублю, руб.	Общий
15	Факторы валютного рынка	Курс евро к рублю, руб.	Общий
16	Факторы валютного рынка	Курс швейцарского франка к рублю, руб.	Общий
17	Факторы валютного рынка	Курс фунта стерлингов к рублю, руб.	Общий
18	Факторы валютного рынка	Курс японской иены к рублю, руб.	Общий
19	Факторы ликвидности	Объем торгов индекса МосБиржи, тыс.руб.	Общий
20	Факторы ликвидности	Объем торгов Dow Jones Industrial Average	Общий
21	Макроэкономические факторы	Индекс потребительских цен (ИПЦ)	Общий
22	Макроэкономические факторы	Ожидаемый индекс потребительских цен (ИПЦ)	Общий
23	Макроэкономические факторы	Объём промышленного производства в России	Общий
24	Макроэкономические факторы	Денежная масса, млрд рублей	Общий
25	Макроэкономические факторы	Денежная масса сезонно скорректированная, млрд руб.	Специфичный
26	Макроэкономические факторы	Первая разность денежной массы, млрд руб.	Специфичный
27	Макроэкономические факторы	Первая разность денежной массы сезонно скорректированной, млрд руб.	Общий
28	Макроэкономические факторы	Доходность индекса CRB	Общий
29	Макроэкономические факторы	Уровень безработицы, %	Общий
30	Макроэкономические факторы	Индекс производственной активности PMI России	Общий
31	Макроэкономические факторы	Фьючерс на нефть Brent	Общий



II. ДАННЫЕ

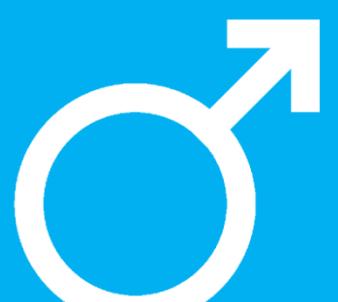
Факторы внимания и настроения

Прикладная экономика

Уникальных пользователей в месяц

317
млн

Аудитория преимущественно мужская



Аудитория представлена в данной возрастной категории

Возраст
18-29

Никогда не «постят» ничего

53%

Среднее время, проведенное через моб.прил.

2.7
минут



II. ДАННЫЕ

Факторы внимания и настроения

Прикладная экономика

Таблица 3. Источники данных по постам Twitter. Таблица построена автором.

Новостные паблики	(1.1) Общий	Финансы и инвестиции	<ul style="list-style-type: none"> • InvestingRu • Finam • Finam signals • RBC invest
	(1.2) Общий	Новости и события	<ul style="list-style-type: none"> • Meduza • RBC • Forbes • Ria
Пользовательские посты	(2.1) Общий	Посты, содержащие ключевые слова «московская биржа», «мосбиржа»	
	(2.2) Специфичный	Посты, содержащие ключевые слова по компаниям: «новатэк», «мтс», «газпром», «лукойл», «роснефть», «яндекс», «сбербанк», «сбер»	

Таблица 4. Страницы в Twitter, выбранные для проведения исследования. Таблица построена автором.

Страница	Количество твитов	Аудитория
InvestingRu	85 тыс.	7,7 тыс.
Finam	120 тыс.	12.5 тыс.
Finam signals	23 тыс.	6.8 тыс.
RBC invest	5 тыс.	31 тыс.
Meduza	72 тыс.	1.3 млн
RBC	109 тыс.	363 тыс.
Forbes	46 тыс.	16.6 млн
Ria	299 тыс.	2.7 млн



финам

meduza

РБК

Forbes



II. ДАННЫЕ

Факторы внимания и настроения

Прикладная экономика

Таблица 5. Факторы внимания и настроения. Источники данных: Twitter. Таблица построена автором.

	Общий / специфичный	Внимание / настроение	Источник	Наименование фактора	
1	Общий	Внимание	Финансы и инвестиции	Прирост логарифма количества постов	
2			Пользовательские посты	Логарифм количества постов	
3		Настроение	Финансы и инвестиции	Среднее настроение, сглаженное	
4				Стандартное отклонение настроения, сглаженное	
5			Новости и события	Среднее настроение, сглаженное	
6				Стандартное отклонение настроения, сглаженное	
7		Внимание	Пользовательские посты	Прирост логарифма количества постов	
8		Настроение		Среднее настроение, сглаженное	
9				Стандартное отклонение настроения, сглаженное	

- Переменные 3-6, 8-9 **сглаживались** следующим образом:

$$\tilde{x}_t = 0.7x_t + 0.2x_{t-1} + 0.1x_{t-2}$$

- Для оценки переменных настроения мы строили классификатор тональности текста, который имеет 2 класса: **положительное** и **отрицательное** настроения. Для всех переменных настроения использовались метки классов, а не вероятностные значения. В итоге получаем, что любому посту сопоставляется число из дискретного набора {0,1}.



III. КЛАССИФИКАЦИЯ ТОНАЛЬНОСТИ ТЕКСТА

UNIVERSAL LANGUAGE MODEL FINE-TUNING FOR TEXT CLASSIFICATION

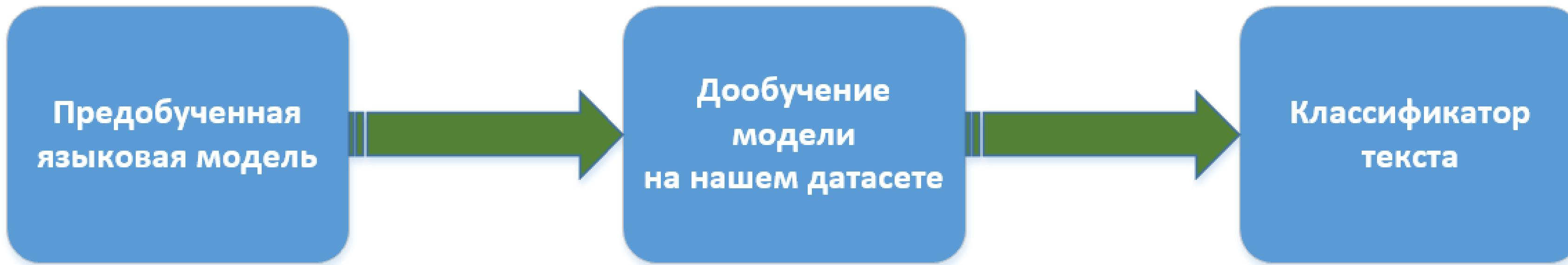
Прикладная экономика

ULMFiT - модель от разработчиков Fast AI; представлена в 2018 году.



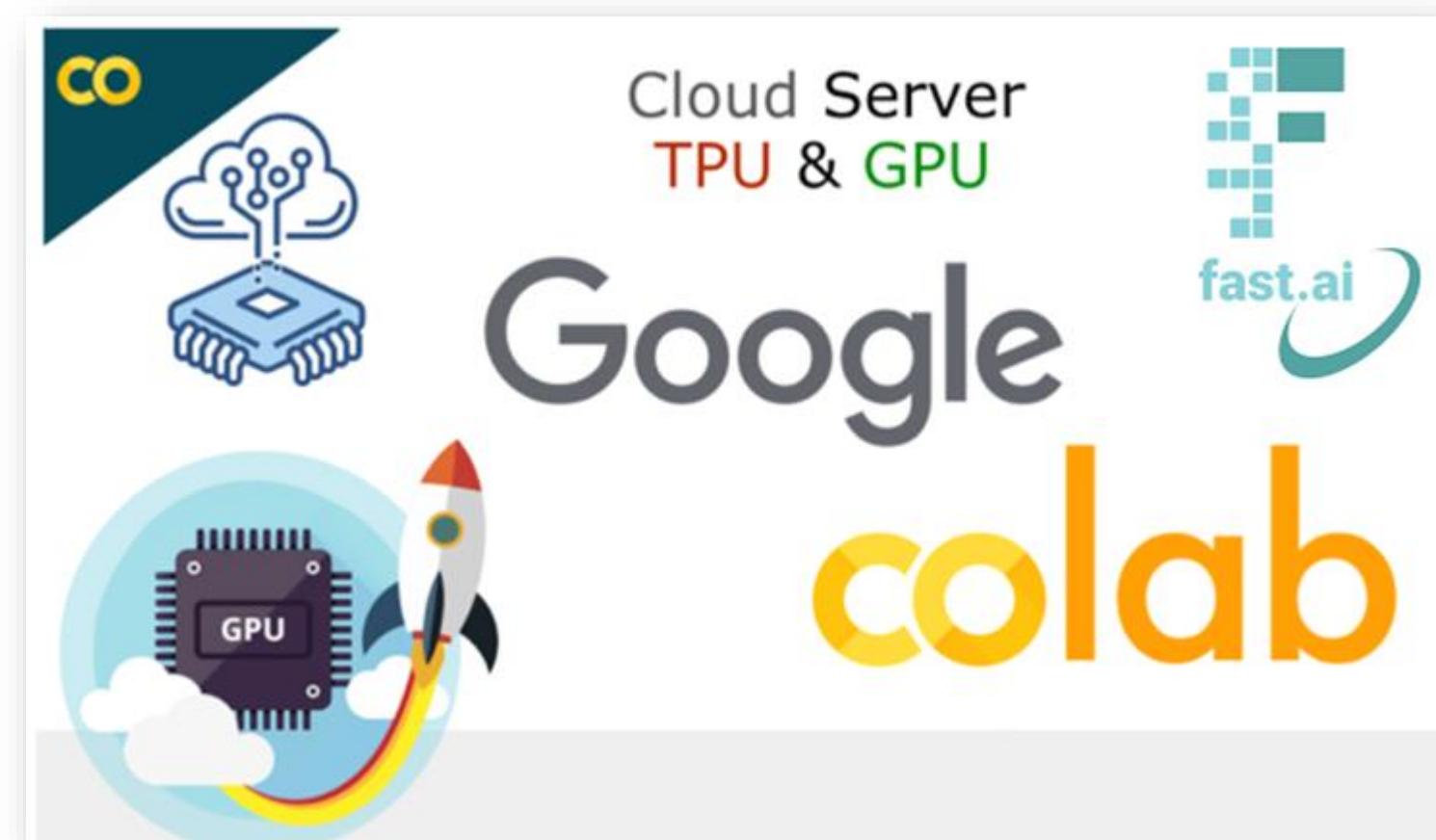
Особенность -Transfer Learning подход к задачам NLP.

Схема обучения модели:



Платформа для обучения:

Google Colaboratory – облачный сервис от Google на основе Jupyter Notebook, который предоставляет все необходимые ресурсы для машинного обучения – более мощные GPU и TPU.



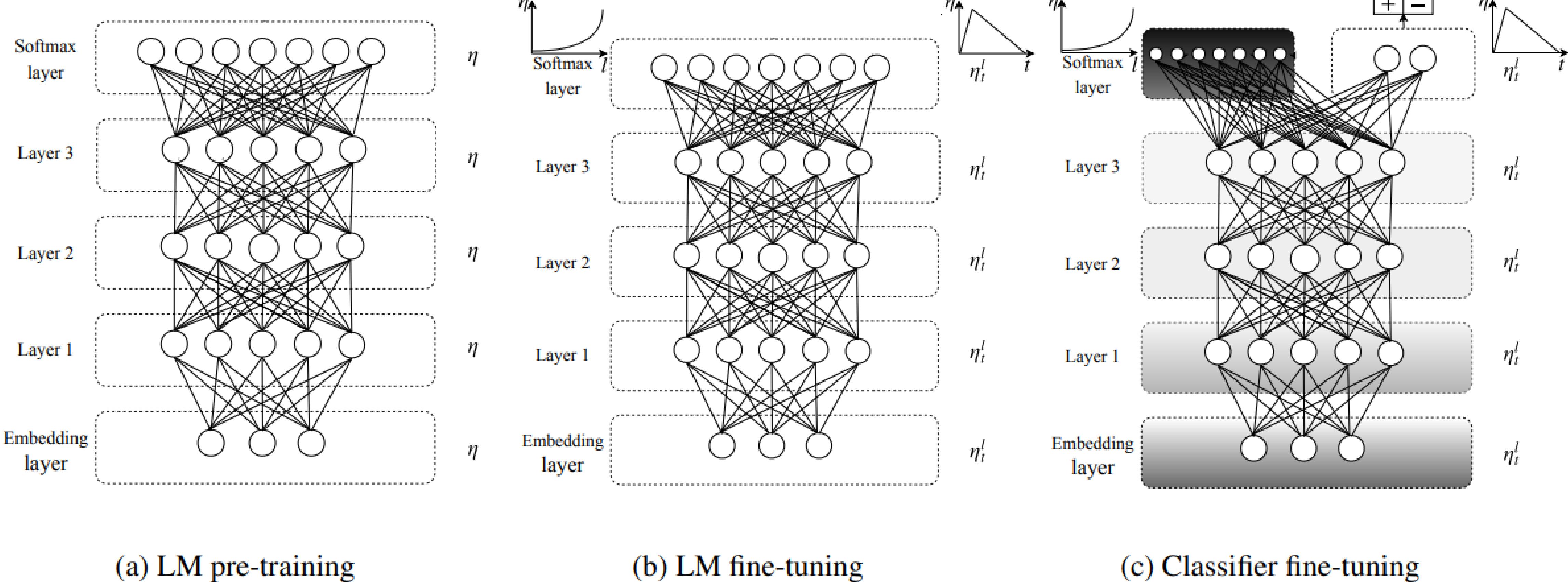


III. КЛАССИФИКАЦИЯ ТОНАЛЬНОСТИ ТЕКСТА

UNIVERSAL LANGUAGE MODEL FINE-TUNING FOR TEXT CLASSIFICATION

Прикладная экономика

Архитектура ULMFiT:





III. КЛАССИФИКАЦИЯ ТОНАЛЬНОСТИ ТЕКСТА

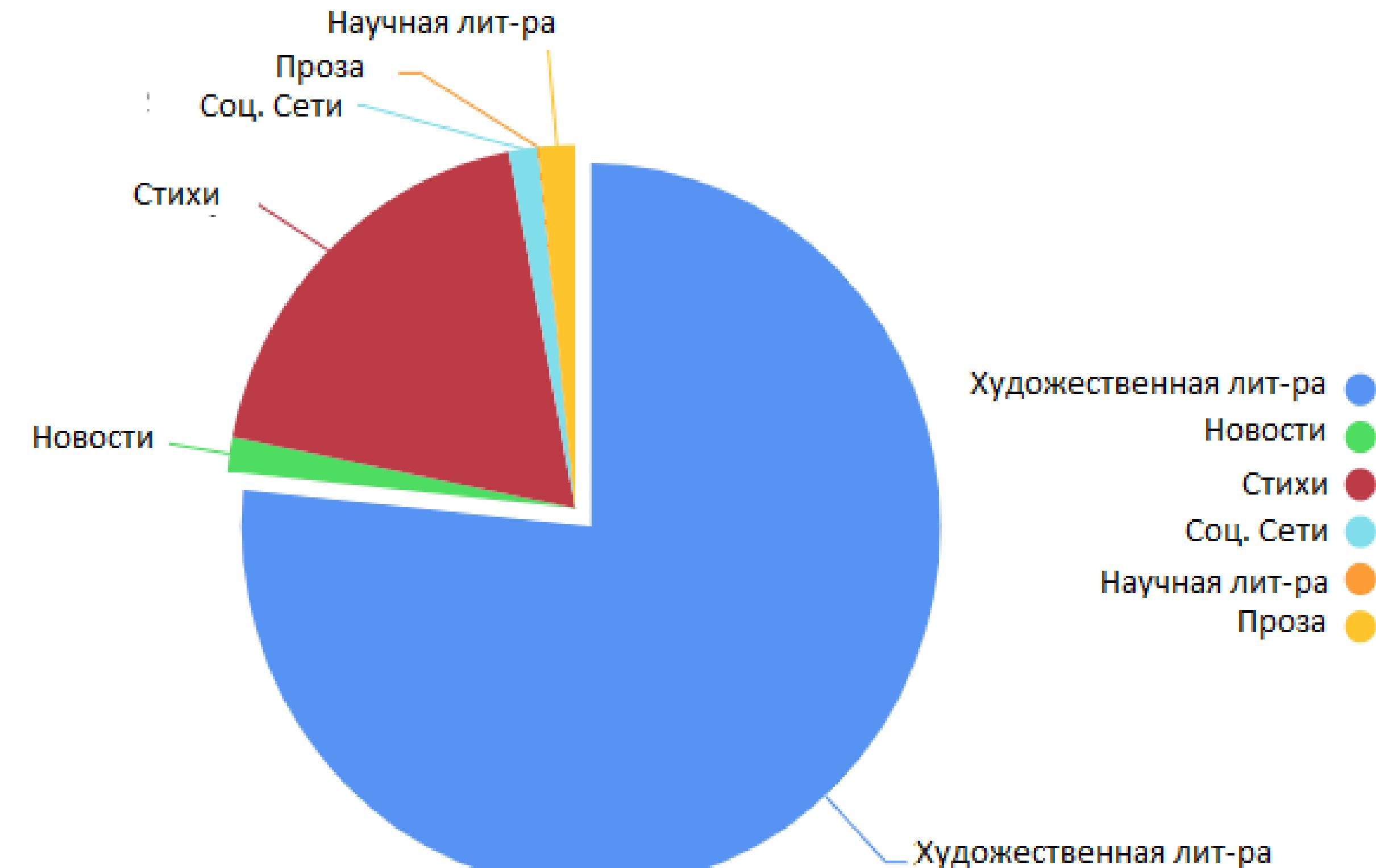
Прикладная экономика

Предобученная модель

Russian AWD LSTM Language model – предобученная на корпусе
Taiga языковая модель.

Taiga корпус:

- **5 миллиард слов;**
- 77 % - Художественный текст;
- 19 % - Стихи;
- 2 % - Новости;
- 2 % - Тексты из социальных сетей, научной литературы, любительских поэм и проз.



Для чего нужна предобученная модель:

- Веса модели используются для тонкой настройки под наши данные.
- Словарь слов.



III. КЛАССИФИКАЦИЯ ТОНАЛЬНОСТИ ТЕКСТА

ULMFIT Fine - Tuning

Прикладная экономика

Fine Tuning - тонкая настройка языковой модели под характер наших данных.

Данные для дообучения:

- Подвыборка запарсенных твитов.
- Корпус размеченных твитов от Юлии Рубцовой 2013 – 2014 год. 226914 твитов.
- Корпус размеченных твитов по банкам с соревнования SentRuEval 2016 год. 2000 твитов.

Предобработка твитов:

- Приведение текста к нижнему регистру;
- Замена буквы «ё» на «е»;
- Удаление ссылок на интернет ресурсы;
- Удаление упоминания пользователей;
- Удаление знаков пунктуации.



III. КЛАССИФИКАЦИЯ ТОНАЛЬНОСТИ ТЕКСТА

ULMFIT Fine - Tuning

Прикладная экономика

- Обучается нейронная сеть на основе минимизации функции потерь. В нашем случае – cross-entropy loss.

$$L(Y, \hat{Y}) = - \sum_{t=1}^n y \log \hat{y}$$

- Веса модели обновляются на основе алгоритма обратного распространения ошибки.
- Пример работы языковой модели:

```
learn_lm.predict("сегодня в стране ситуация", n_words=20)

'сегодня в стране ситуация с нормализации жизни в россии у нас все в порядке а бешеный карантин прошел на бали мюнхене сообщил дефицит государственного'
```

- Языковая модель после тонкой настройки дополняется двумя линейными блоками ReLU и SoftMax для дальнейшей классификации твита.



III. КЛАССИФИКАЦИЯ ТОНАЛЬНОСТИ ТЕКСТА

ULMFIT Обучение классификатора

Прикладная экономика

Данные для обучения классификатора:

- Корпус размеченных твитов от Юлии Рубцовой 2013 – 2014 год. 226914 твитов.
- Корпус размеченных твитов по банкам с соревнования SentRuEval 2016 год. 2000 твитов.
- Выборка поделена на тестовую 20%, обучающую 64% и валидационную 16%. Разделение случайное, при этом соблюдалась балансировка классов в каждой подвыборке.

Распределение твитов в подвыборках

Таргет	Обучающая	Тестовая	Валидационная
0 – Негативный	90926	22731	18186
1 – Позитивный	92490	23123	18498



III. КЛАССИФИКАЦИЯ ТОНАЛЬНОСТИ ТЕКСТА

ULMFIT Обучение классификатора

Прикладная экономика

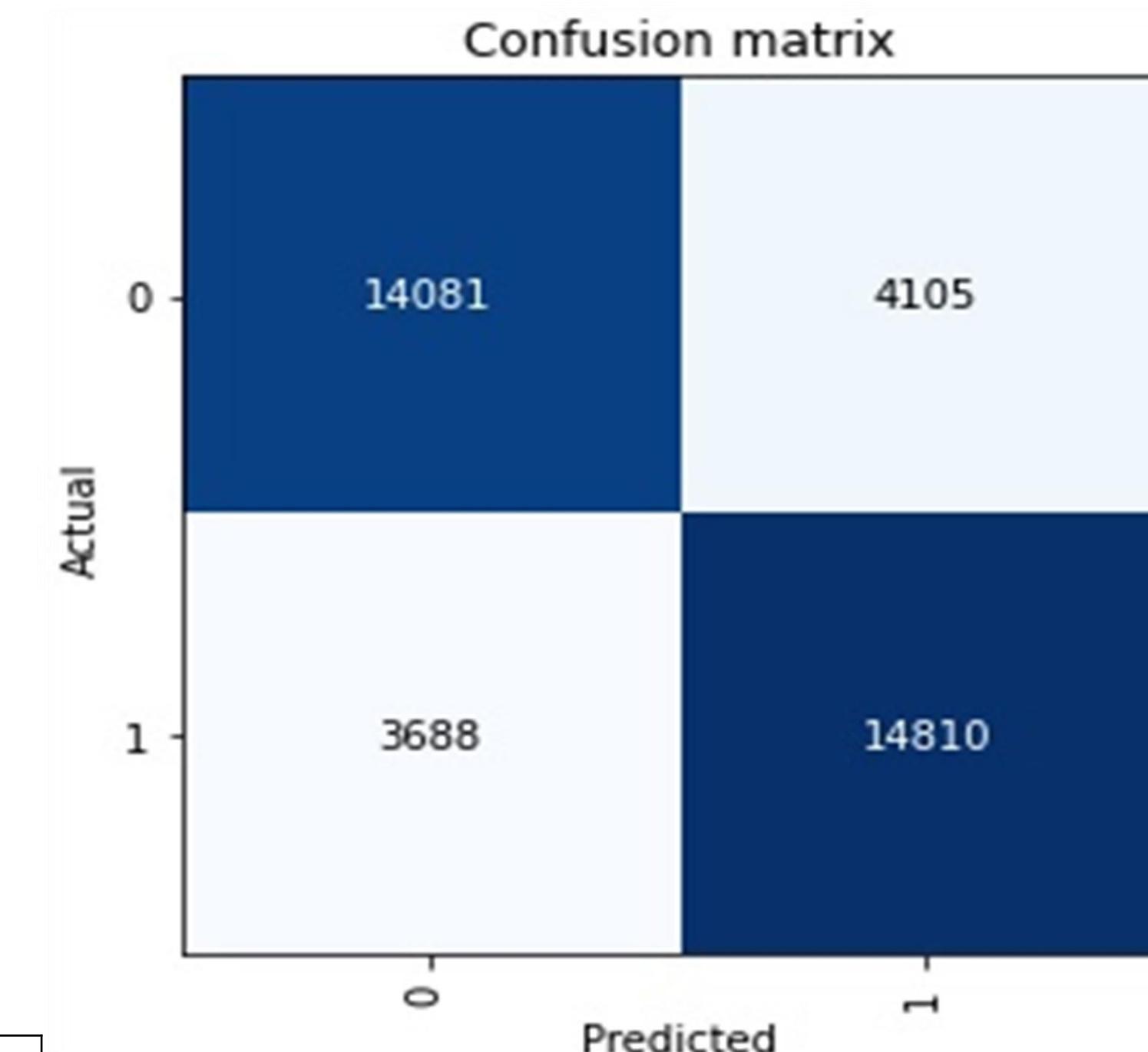
Оценка классификатора

$$\text{Точность} = \frac{TP}{TP + FP}; \quad \text{Полнота} = \frac{TP}{TP + FN}$$

$$F = 2 \frac{\text{Точность} * \text{Полнота}}{\text{Точность} + \text{Полнота}}$$

Значение метрик классификатора

Метка класса	Точность	Полнота	F – score
0 – Негативный	0.79	0.77	0.78
1 – Позитивный	0.78	0.80	0.79
Среднее	0.79	0.79	0.79





IV. ПРОГНОЗИРОВАНИЕ

Прикладная экономика

Модели

Функция потерь для всех моделей: **MSE**





IV. ПРОГНОЗИРОВАНИЕ

Основные модели

Прикладная экономика

Базовая модель

$$\log RV_{t+1}^{(d)} = c + \beta^{(d)} \log RV_t^{(d)} + \beta^{(w)} \log RV_t^{(w)} + \beta^{(m)} \log RV_t^{(m)} + \varepsilon_{t+1}$$
$$\log RV_{t+1}^{(d)} = c + (\log RV_t)' \beta_{RV} + \varepsilon_{t+1} \quad (2)$$

$\log RV_t^{(d)}$ – оценка реализованной волатильности в день t методом MedRV, ε_{t+1} – ошибка предсказания, $(\log RV_t)'$ – матрица признаков волатильности размера $n \times 3$, n – количество наблюдений, β_{RV} – трехмерный вектор-столбец весов.

Экономическая модель

E_t' – матрица размера $n \times 9$, в которой каждый вектор-столбец – экономический или финансовый фактор, γ_{eco} – 9-мерный вектор-столбец весов для экономических и финансовых признаков

Регуляризация:

$$L_{lasso}(X_t, RV_{t+1}^{(d)}, W) = (\log RV_{t+1}^{(d)} - c - W' X_t)^2 + \lambda ||W||_1 \rightarrow \min_{c, W} \quad (4)$$

X_t – матрица признаков размером $n \times 12$ (3 признака волатильности + 9 экономических признаков), W – вектор-столбец весов размера 12, λ – коэффициент регуляризации (гиперпараметр), $||W||_1 = \sum_{j=1}^m |w_j|$ – норма весов.

Модель настроения

S_t' – матрица размера $n \times 9$, в которой каждый вектор-столбец – фактор внимания или настроения, θ_{sent} – 9-мерный вектор-столбец весов для признаков внимания и настроения

Регуляризация:

$$L_{lasso}(X_t, RV_{t+1}^{(d)}, W) = (\log RV_{t+1}^{(d)} - c - W' X_t)^2 + \lambda ||W||_1 \rightarrow \min_{c, W} \quad (6)$$

где X_t – матрица признаков размером $n \times 21$ (3 признака волатильности + 9 экономических признаков + 9 признаков настроения и внимания), W – вектор-столбец весов размера 21, λ – коэффициент регуляризации (гиперпараметр), $||W||_1 = \sum_{j=1}^m |w_j|$ – норма весов



IV. ПРОГНОЗИРОВАНИЕ

Стандартизация для численных методов

Прикладная экономика

Все признаки приведены к распределению с нулевым матожиданием и единичной дисперсией:

$$\bar{\mu}_{\cdot j} = \frac{1}{n} \sum_{i=1}^n x_{ij},$$
$$\bar{\sigma}_{\cdot j} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_{ij} - \bar{\mu}_{\cdot j})^2},$$
$$X^{new} = \frac{X - \bar{\mu}}{\bar{\sigma}}$$



IV. ПРОГНОЗИРОВАНИЕ

Дополнительные модели

Прикладная экономика

Random Forest (Случайный Лес)

Алгоритм построения:

- Формируем N бутстррап подвыборок
- Для каждой подвыборки строится решающее дерево b_n со следующими условиями:
 - Признак при разбиении в каждом узле дерева выбирается из случайного подмножества признаков q (всего признаков d)
 - Деревья должны быть глубокие
- Деревья строятся независимо
- Ответ алгоритма на одном наблюдении :
$$a(X) = \frac{1}{N} \sum_{n=1}^N b_n(X)$$





IV. ПРОГНОЗИРОВАНИЕ

Дополнительные модели

Прикладная экономика

XGBoost (Экстремальный градиентный бустинг)

Алгоритм построения:

- Инициализируется первое дерево b_0
- На шаге N-1 алгоритм выглядит как композиция:

$$a_{N-1}(x) = \sum_{n=1}^{N-1} b_n(x)$$

- На шаге N добавляется алгоритм:

$$b_N(x) = \operatorname{argmin}_b \sum_{i=1}^l L(y_i, a_{N-1}(x) + b(x))$$

1. Деревья неглубокие и строятся последовательно
2. Регуляризация: количество листьев и значения в листе

dmlc
XGBoost



IV. ПРОГНОЗИРОВАНИЕ

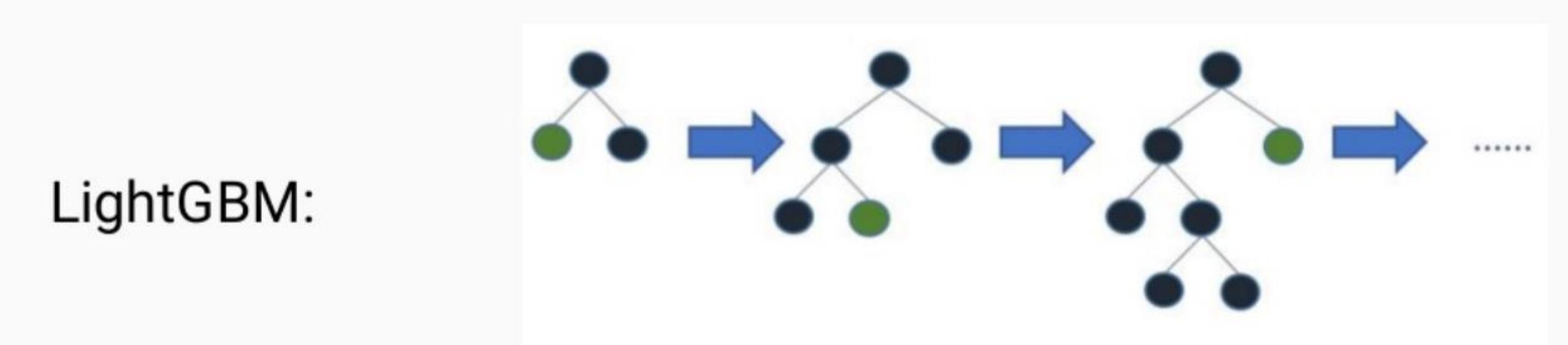
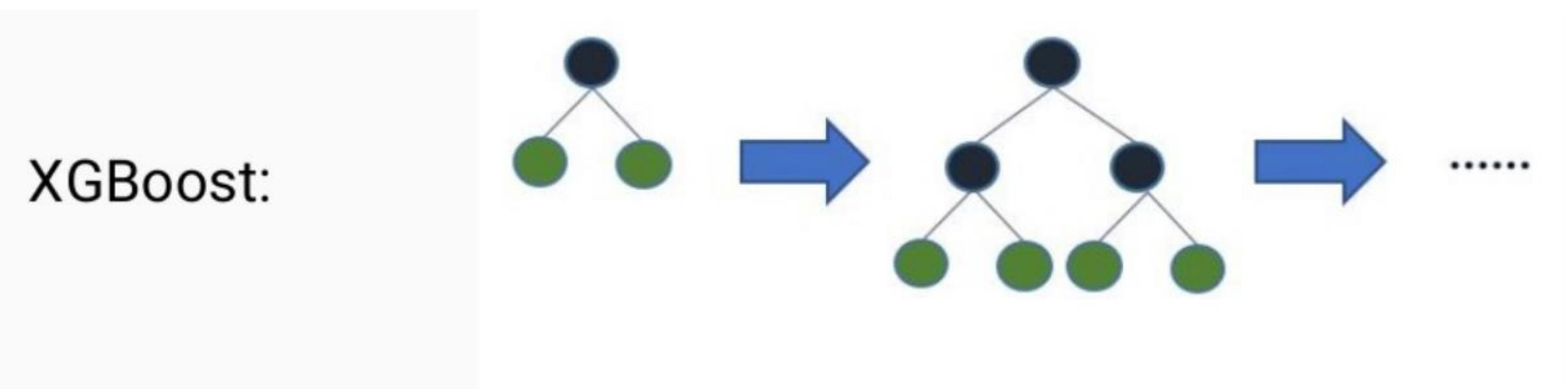
Дополнительные модели

Прикладная экономика

Расширенный градиентный бустинг (LightGBM):

Алгоритм построения композиции схож с XGBoost.

Принципиальное отличие от XGBoost это метод построения композиции:





IV. ПРОГНОЗИРОВАНИЕ

Подбор гиперпараметров и кросс-валидация

Прикладная экономика

Исследуемый период был сокращен: с 01.01.2020 по 28.02.2021, т.к.:

1. Коронавирус внес сильный вклад в несопоставимость новостей за 2016–2019 и 2019 – 2021 гг.
2. Очень большой промежуток, поэтому влияние факторов различно в разные годы.
3. Качество данных за 2016–2019 гг. сильно хуже, особенно данные социальных сетей (малое кол-во постов).

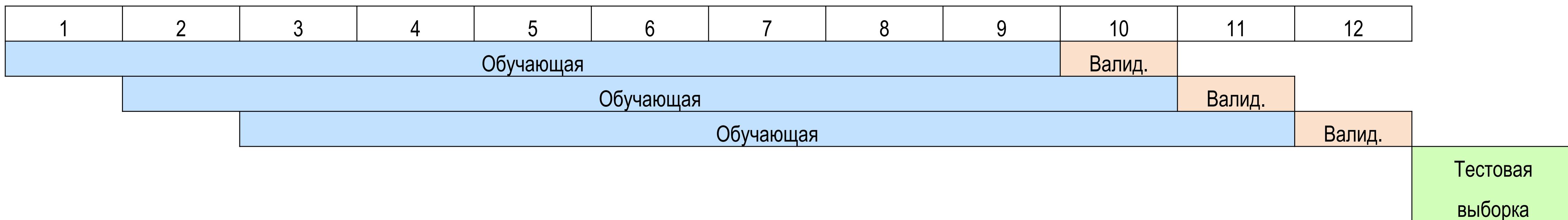
Поиск по сетке (Grid Search)

Поиск по сетке принимает на вход модель и различные значения гиперпараметров (сетку гиперпараметров). Далее, для каждого возможного сочетания значений гиперпараметров, метод считает ошибку и в конце выбирает сочетание, при котором ошибка минимальна.

Байесовский оптимизатор (Bayesian Optimization)

В отличии от поиска по сетке этот метод полагается на информацию, полученную моделью во время предыдущих оптимизаций, чтобы найти наиболее оптимизированный список параметров.

Рис. 5. Деление на обучающую и тестовую (отложенную) выборку. Обучающая выборка делилась в свою очередь на 12 фолдов для кросс-валидации, чтобы подобрать оптимальные гиперпараметры. Источник: нарисовано автором.



Так как подобной реализации кросс-валидации нет в библиотеках python, нами самостоятельно был реализован класс, выполняющий данную процедуру, в библиотеке SciKit-Learn.



IV. ПРОГНОЗИРОВАНИЕ

Реализованная волатильность

Прикладная экономика

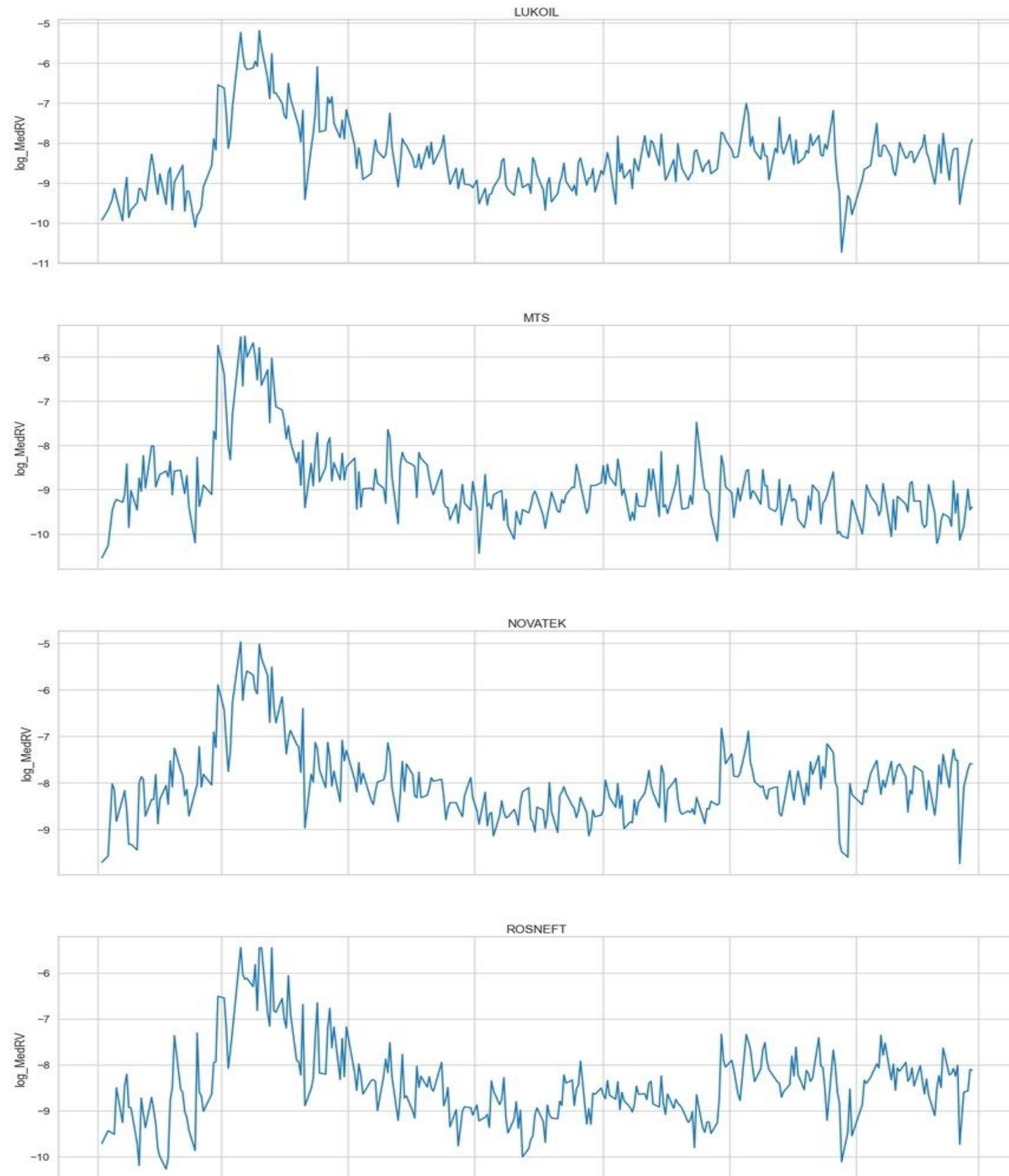
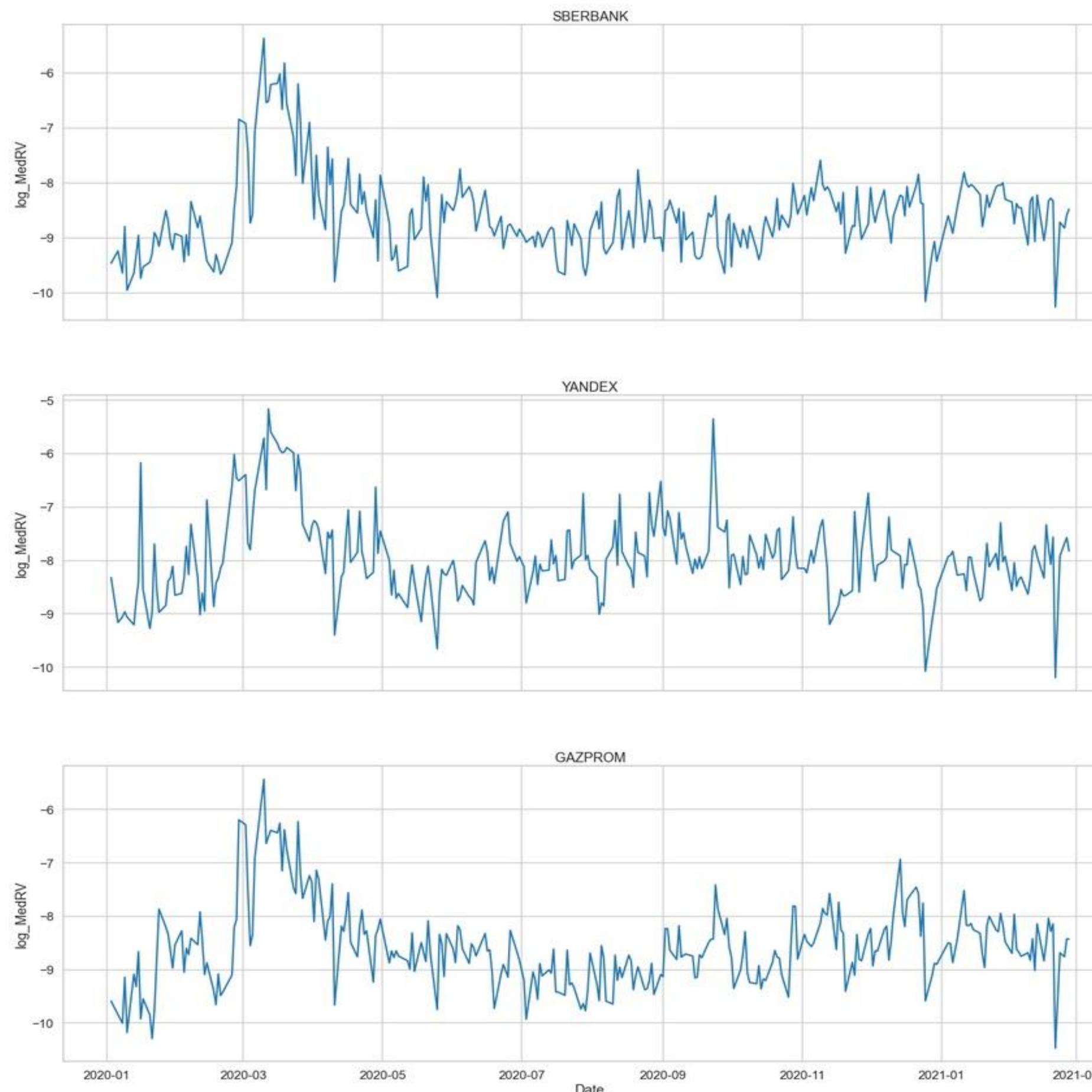


Рисунок 1. Временной ряд логарифма оценки реализованной волатильности для всех компаний.
Построено автором.





IV. ПРОГНОЗИРОВАНИЕ

Реализованная волатильность

Прикладная экономика

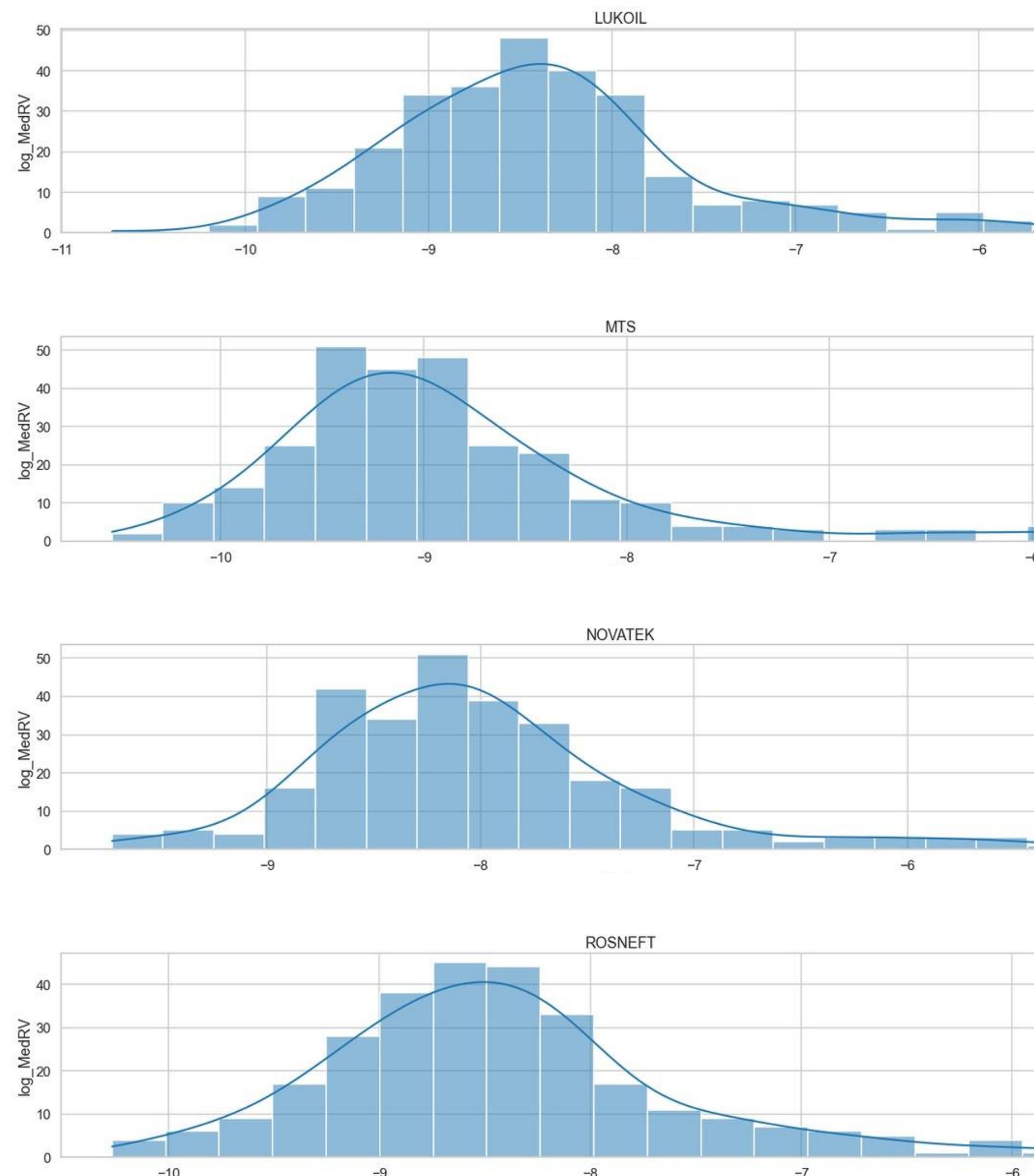
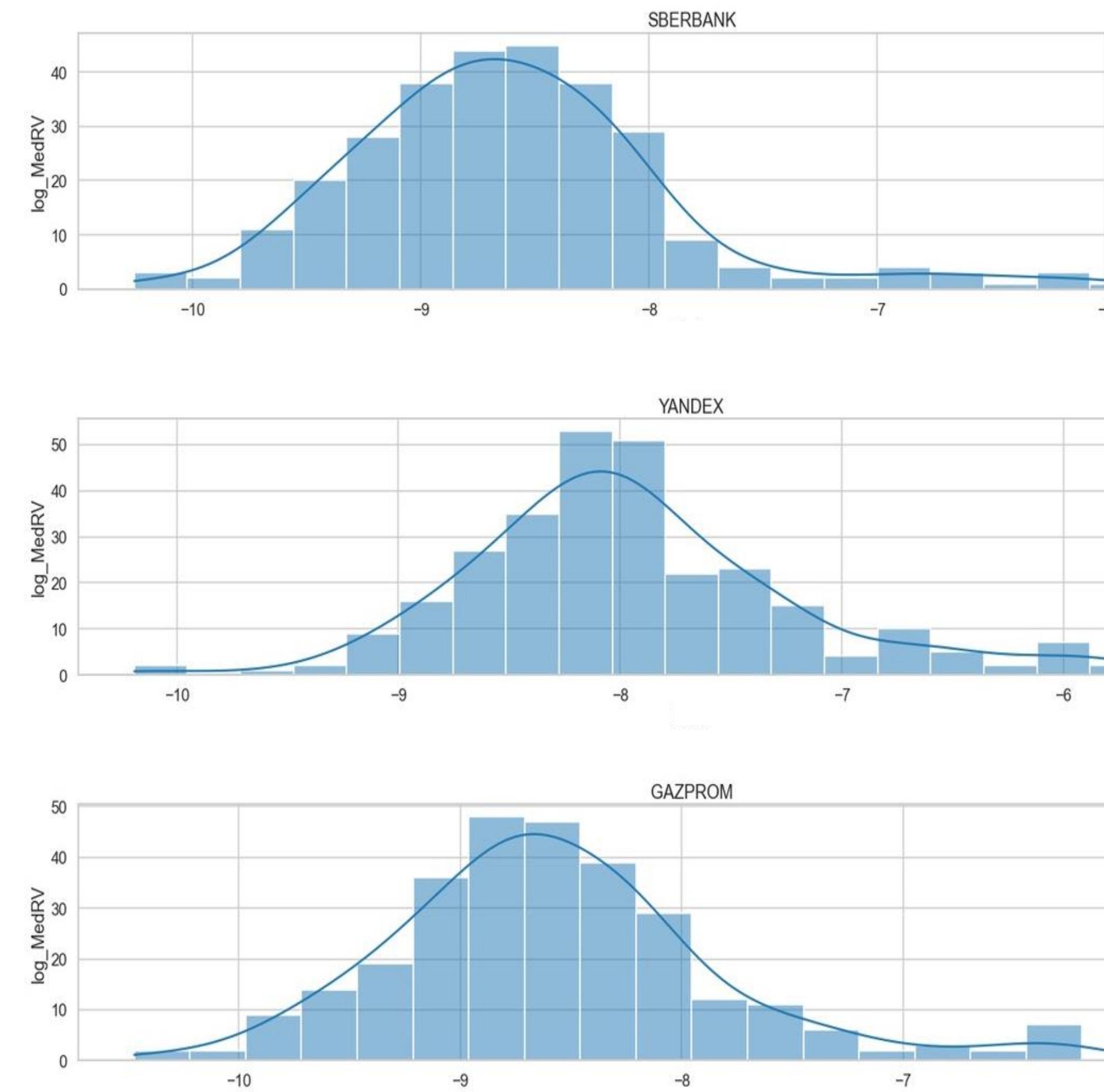


Рисунок 2. Распределение логарифма оценки реализованной волатильности для всех компаний.
Построено автором.





IV. ПРОГНОЗИРОВАНИЕ

Прикладная экономика

Описательная статистика факторов для «Яндекс»

Таблица 6. Описательная статистика переменных волатильности, экономических и финансовых показателей и показателей внимания и настроения. По строкам расположены слева направо: количество наблюдений, среднее, стандартное отклонение, минимум, 25%, 50%, 75% перцентили максимум. По столбцам расположены факторы. Таблица построена автором.

Фактор	count	mean	std	min	0,25	0,5	0,75	max
Логарифм реализованной волатильности за предыдущий день	278	-7,93	0,78	-10,19	-8,38	-8,04	-7,58	-5,16
Логарифм реализованной волатильности за предыдущую неделю	278	-7,93	0,63	-9,10	-8,30	-8,02	-7,72	-5,70
Логарифм реализованной волатильности за предыдущую месяц	278	-7,96	0,51	-8,94	-8,26	-8,02	-7,73	-6,32
Индекс IMOEX	278	0,01	0,01	0,00	0,00	0,01	0,01	0,09
Индекс MSCI Russia	278	0,00	0,03	-0,15	-0,01	0,00	0,01	0,13
Доходность облигации России 10-летние (годовая ставка), %	278	0,03	2,97	-12,07	-1,30	-0,18	1,01	16,12
Доходность облигации Россия годовые (годовая ставка), %	278	0,04	1,52	-5,88	-0,49	0,00	0,50	10,12
Курс доллара к рублю, руб	278	0,01	0,01	0,00	0,00	0,01	0,01	0,07
Индекс потребительских цен (ИПЦ)	278	0,00	0,01	-0,05	0,00	0,00	0,00	0,06
Доходность акции, %	278	0,16	2,80	-13,34	-1,34	0,39	1,77	11,80
Объем торгов, тыс. руб.	278	2612,71	1720,95	370,63	1612,50	2225,00	3200,00	19730,00
Объем торгов индекса МосБиржи, тыс. руб.	278	13155,94	8658,80	961,46	7765,00	11435,00	15492,50	69820,00
Среднее настроение, сглаженное (финансы и инвестиции)	278	0,53	0,09	0,25	0,45	0,53	0,58	0,82
Прирост логарифма количества постов (финансы и инвестиции)	278	2,36	1,15	0,00	2,56	2,83	3,04	3,50
Стандартное отклонение настроения, сглаженное (финансы и инвестиции)	278	0,48	0,04	0,14	0,47	0,49	0,50	0,52
Среднее настроение, сглаженное (новости и события)	278	0,60	0,03	0,52	0,58	0,60	0,63	0,70
Стандартное отклонение настроения, сглаженное (новости и события)	278	0,49	0,01	0,46	0,48	0,49	0,49	0,50
Среднее настроение, сглаженное (пользовательские посты по компании)	278	0,55	0,04	0,41	0,52	0,55	0,57	0,66
Стандартное отклонение настроения, сглаженное (пользовательские посты по компании)	278	0,50	0,01	0,47	0,50	0,50	0,50	0,50
Прирост логарифма количества постов (пользовательские посты по компании)	278	4,76	0,29	4,08	4,62	4,74	4,90	6,30
Логарифм количества постов (пользовательские посты по МосБирже)	278	0,40	0,55	0,00	0,00	0,00	0,69	2,71

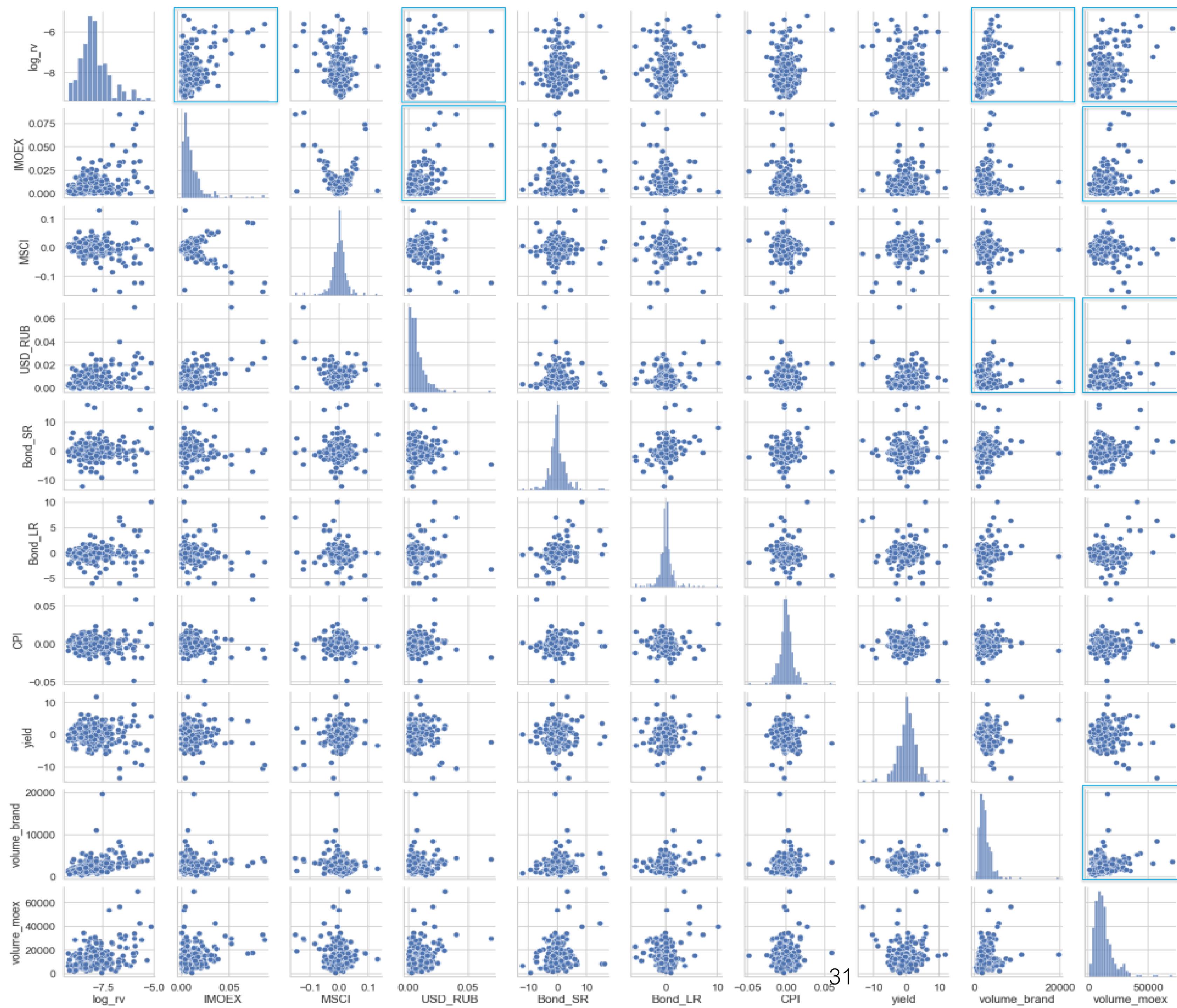
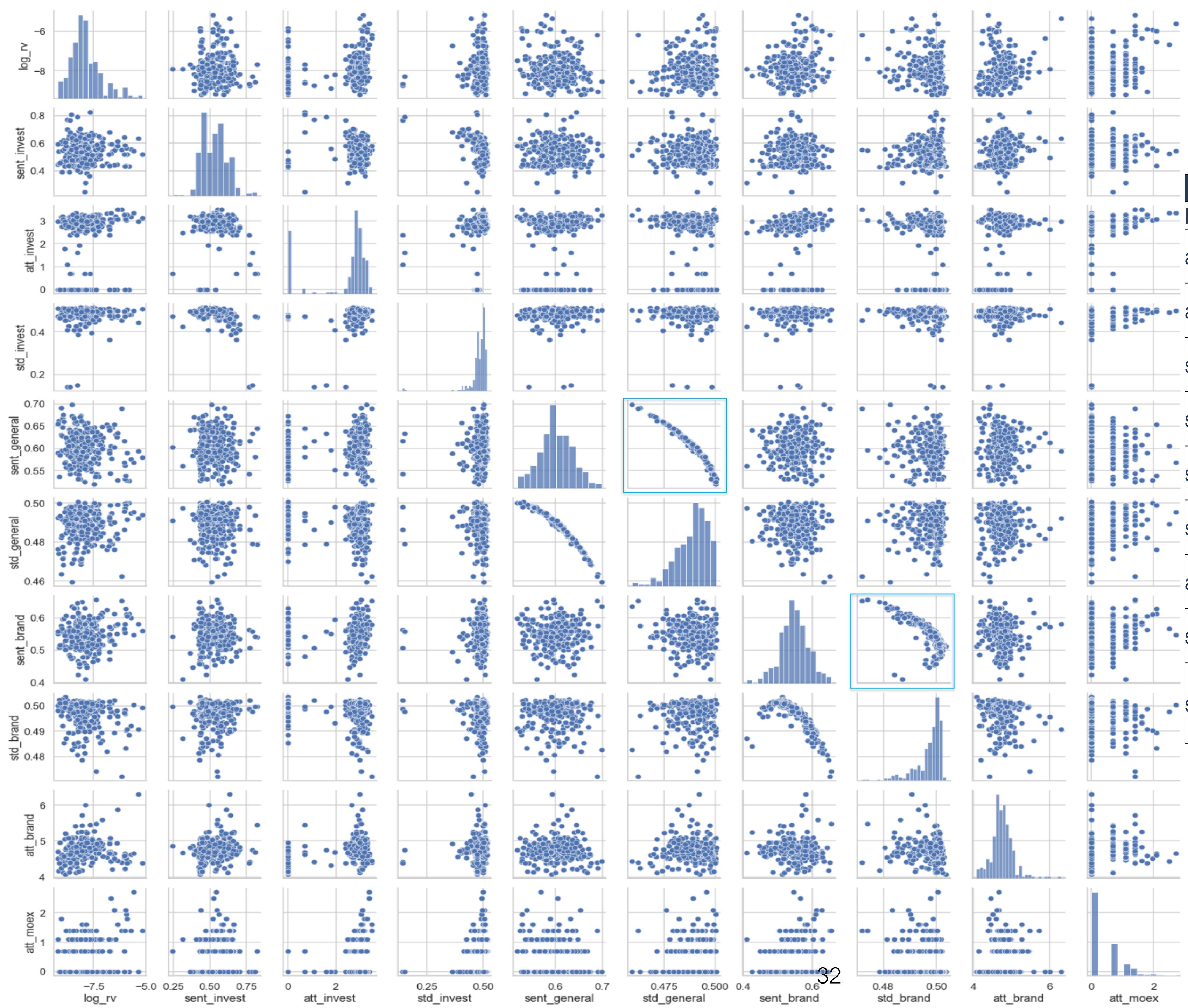


Рисунок 3. Графики зависимостей между экономическими и финансовыми переменными и распределения факторов для компании «Яндекс». Построено автором.

Атрибут	Фактор
log_rv	Логарифм реализованной волатильности
IMOEX	Индекс IMOEX
MSCI	Индекс MSCI Russia
Bond_LR	Доходность облигации Россия 10-летние (годовая ставка), %
Bond_SR	Доходность облигации Россия годовые (годовая ставка), %
USD_RUB	Курс доллара к рублю, руб
yield	Доходность акции, %
volume_brand	Объем торгов, тыс. руб.
volume_MOEX	Объем торгов индекса МосБиржи, тыс. руб.
CPI	Индекс потребительских цен (ИПЦ)

Рисунок 4. Графики зависимостей между факторами внимания и настроения для компании «Яндекс» и распределения факторов. Построено автором.



Атрибут	Фактор
log_rv	Логарифм реализованной волатильности
att_invest	Прирост логарифма количества постов (финансы и инвестиции)
att_moex	Логарифм количества постов (пользовательские посты по МосБирже)
sent_invest	Среднее настроение, сглаженное (финансы и инвестиции)
std_invest	Стандартное отклонение настроения, сглаженное (финансы и инвестиции)
sent_general	Среднее настроение, сглаженное (новости и события)
std_general	Стандартное отклонение настроения, сглаженное (новости и события)
att_brand	Прирост логарифма количества постов (пользовательские посты по компании)
sent_brand	Среднее настроение, сглаженное (пользовательские посты по компании)
std_brand	Стандартное отклонение настроения, сглаженное (пользовательские посты по компании)



IV. ПРОГНОЗИРОВАНИЕ

Корреляции

Прикладная экономика

Таблица 7. Корреляция между объясняющими переменными и целевой зависимой переменной для каждой рассматриваемой компании. По строкам расположены факторы, по столбцам компании. Обозначения признаков аналогичны как в приложении 1. Таблица построена автором.

Фактор	Лукойл	МТС	Новатэк	Роснефть	Газпром	Яндекс	Сбербанк
Логарифм реализованной волатильности за предыдущий день	0,79	0,76	0,75	0,74	0,70	0,67	0,68
Логарифм реализованной волатильности за предыдущую неделю	0,77	0,78	0,75	0,74	0,69	0,62	0,68
Логарифм реализованной волатильности за предыдущую месяц	0,58	0,61	0,60	0,59	0,51	0,46	0,45
Индекс IMOEX	0,58	0,52	0,54	0,56	0,51	0,39	0,56
Индекс MSCI Russia	-0,15	-0,20	-0,16	-0,15	-0,20	-0,17	-0,19
Доходность облигации Россия годовые (годовая ставка), %	0,07	0,02	0,05	0,04	0,10	0,08	0,10
Доходность облигации России 10-летние (годовая ставка), %	0,12	0,13	0,13	0,11	0,18	0,19	0,20
Курс доллара к рублю, руб	0,47	0,45	0,46	0,44	0,44	0,34	0,47
Доходность акции, %	0,04	0,00	0,01	0,03	0,01	-0,08	-0,05
Объем торгов, тыс. руб.	0,64	0,51	0,62	0,67	0,55	0,36	0,57
Объем торгов индекса МосБиржи, тыс. руб.	0,55	0,50	0,50	0,49	0,44	0,40	0,48
Индекс потребительских цен (ИПЦ)	0,03	0,00	0,01	0,04	0,00	0,01	0,03
Среднее настроение, сглаженное (финансы и инвестиции)	0,07	0,09	0,08	0,10	0,05	0,03	0,04
Прирост логарифма количества постов (финансы и инвестиции)	0,10	0,12	0,11	0,14	0,09	0,14	0,07
Стандартное отклонение настроения, сглаженное (финансы и инвестиции)	0,06	0,06	0,03	0,03	0,05	0,10	0,08
Среднее настроение, сглаженное (новости и события)	-0,41	-0,27	-0,29	-0,37	-0,32	-0,15	-0,35
Стандартное отклонение настроения, сглаженное (новости и события)	0,38	0,21	0,25	0,32	0,29	0,13	0,31
Среднее настроение, сглаженное (пользовательские посты по компании)	-0,10	-0,01	-0,03	0,04	0,09	0,14	0,06
Стандартное отклонение настроения, сглаженное (пользовательские посты по компании)	0,05	-0,04	0,10	0,06	-0,01	-0,15	0,13
Прирост логарифма количества постов (пользовательские посты по компании)	0,21	-0,13	0,14	0,40	0,03	0,11	0,02
Логарифм количества постов (пользовательские посты по МосБирже)	0,36	0,32	0,32	0,38	0,35	0,27	0,34



IV. ПРОГНОЗИРОВАНИЕ

Базовая модель

Прикладная экономика

Таблица 8. Коэффициенты для базовой модели. Таблица построена автором.

Фактор	Яндекс	Сбербанк	МТС	Лукойл	Роснефть	Газпром	Новатэк
Свободный коэффициент	-1.469	-1.585	-1.058	-1.135	-1.121	-1.352	-0.993
Логарифм реализованной волатильности за предыдущий день	0.472***	0.376***	0.354***	0.496***	0.395***	0.446***	0.436***
Логарифм реализованной волатильности за предыдущую неделю	0.300***	0.510***	0.563***	0.408***	0.469***	0.380***	0.435***
Логарифм реализованной волатильности за предыдущий месяц	0.043	-0.070	-0.037	-0.041	0.002	0.016	0.004

Таблица 9. Значения среднеквадратичной ошибки MSE для базовой модели на тестовой выборке для каждой отдельной компании. Таблица построена автором.

Компания	MSE
Яндекс	0.326
Сбербанк	0.220
МТС	0.187
Лукойл	0.208
Роснефть	0.196
Газпром	0.254
Новатэк	0.247



IV. ПРОГНОЗИРОВАНИЕ

Прикладная экономика

Экономическая модель. Лассо регрессия

Таблица 10. Коэффициенты для регрессии лассо в экономической модели для каждой рассматриваемой компании. По строкам расположены факторы, по столбцам компании.
Таблица построена автором.

Фактор	Яндекс	Сбербанк	МТС	Лукойл	Роснефть	Газпром	Новатэк
Свободный коэффициент	-7,859	-8,607	-8,793	-8,355	-8,422	-8,575	-8,008
Логарифм реализованной волатильности за предыдущий день	0,338	0,166	0,241	0,285	0,243	0,229	0,274
Логарифм реализованной волатильности за предыдущую неделю	0,113	0,202	0,361	0,217	0,233	0,153	0,164
Логарифм реализованной волатильности за предыдущий месяц	0,047	0,007	0,003	0,063	0,047	0,074	0,030
Индекс IMOEX	0,057	0,180	0,106	0,173	0,187	0,141	0,091
Индекс MSCI Russia	0,000	-0,018	-0,046	-0,021	-0,025	-0,029	-0,003
Курс доллара к рублю, руб	0,012	0,049	0,037	0,047	0,037	0,044	0,056
Доходность облигации Россия годовые (годовая ставка), %	0,001	0,071	0,063	0,102	0,092	0,078	0,076
Доходность облигации России 10-летние (годовая ставка), %	0,095	0,082	0,101	0,045	0,047	0,051	0,023
Индекс потребительских цен (ИПЦ)	0,000	0,006	0,000	0,017	0,026	-0,003	0,000
Доходность акции, %	0,004	0,000	0,053	0,034	0,033	0,000	-0,016
Объем торгов, тыс. руб.	0,124	0,077	0,042	0,155	0,105	0,168	0,168
Объем торгов индекса МосБиржи, тыс. руб.	0,051	0,072	0,050	0,038	0,059	0,000	0,053



IV. ПРОГНОЗИРОВАНИЕ

Экономическая модель. Лассо регрессия и дополнительные модели

Прикладная экономика

Таблица 11. Значения среднеквадратичной ошибки MSE для экономической модели на тестовой выборке для каждой отдельной компании. По столбцам слева направо: регрессия лассо, случайный лес, экстремальный градиентный бустинг, расширенный градиентный бустинг. Таблица построена автором.

Компания	Lasso	RF	XGB	LGBM
Яндекс	0.160	0.123	0.144	0.132
Сбербанк	0.213	0.207	0.202	0.189
МТС	0.255	0.233	0.261	0.283
Лукойл	0.191	0.213	0.192	0.200
Роснефть	0.234	0.186	0.183	0.214
Газпром	0.233	0.231	0.219	0.223
Новатэк	0.263	0.223	0.238	0.250

Таблица 12. Уменьшение среднеквадратичной ошибки MSE на тестовой выборке в экономической модели относительно базовой модели, %. По столбцам слева направо: регрессия лассо, случайный лес, экстремальный градиентный бустинг, расширенный градиентный бустинг. Таблица построена автором.

Компания	Lasso	RF	XGB	LGBM
Яндекс	51%	62%	56%	60%
Сбербанк	3%	6%	8%	14%
МТС	-36%	-25%	-40%	-51%
Лукойл	8%	-2%	8%	4%
Роснефть	-19%	5%	7%	-9%
Газпром	8%	9%	14%	12%
Новатэк	-6%	10%	4%	-1%

- В плохом смысле выделяется только компания **МТС**, для нее экономические факторы не дали никакого прироста в точности.
- **Яндекс, Сбербанк и Газпром** хорошо отреагировали на добавление экономических факторов.
- **Лукойл, Роснефть и Новатэк** дали неоднозначные результаты, хотя для 3 из 4 моделей добавление экономических факторов оказалось положительное влияние на качество прогноза.
- В среднем, если не брать в учет МТС, экономическая модель дает **прирост в 13%** в качестве прогноза по сравнению с бенчмарком.



IV. ПРОГНОЗИРОВАНИЕ

Модель настроения. Лассо регрессия

Прикладная экономика

Таблица 13. Коэффициенты для регрессии лассо в модели настроения для каждой рассматриваемой компании. По строкам расположены факторы, по столбцам компании.
Таблица построена автором.

Фактор	Яндекс	Сбербанк	МТС	Лукойл	Роснефть	Газпром	Новатэк
Свободный коэффициент	-2,902	-3,368	-2,244	-2,817	-3,854	-3,023	-3,191
Логарифм реализованной волатильности за предыдущий день	0,428	0,320	0,349	0,423	0,329	0,356	0,390
Логарифм реализованной волатильности за предыдущую неделю	0,184	0,328	0,434	0,268	0,270	0,227	0,240
Логарифм реализованной волатильности за предыдущий месяц	0,071	0,000	0,000	0,024	0,000	0,113	0,022
Индекс IMOEX	0,000	0,000	0,000	0,000	0,000	0,000	0,000
Индекс MSCI Russia	0,000	0,000	0,000	0,000	0,000	0,000	0,000
Курс доллара к рублю, руб	0,000	0,000	0,000	0,000	0,000	0,000	0,000
Доходность облигации Россия годовые (годовая ставка), %	0,000	0,028	0,028	0,041	0,037	0,031	0,030
Доходность облигации России 10-летние (годовая ставка), %	0,056	0,047	0,064	0,027	0,021	0,028	0,011
Индекс потребительских цен (ИПЦ)	0,000	0,000	0,000	0,000	0,000	0,000	0,000
Доходность акции, %	0,004	-0,005	0,043	0,013	0,013	0,001	-0,012
Объем торгов, тыс. руб.	0,000	0,000	0,000	0,000	0,000	0,000	0,000
Объем торгов индекса МосБиржи, тыс. руб.	0,000	0,000	0,000	0,000	0,000	0,000	0,000
Среднее настроение, сглаженное (финансы и инвестиции)	0,000	0,000	0,000	0,000	0,000	0,000	0,000
Прирост логарифма количества постов (финансы и инвестиции)	0,049	0,000	0,035	0,000	0,026	0,008	0,000
Стандартное отклонение настроения, сглаженное (финансы и инвестиции)	0,000	0,000	0,000	0,000	0,000	0,000	0,000
Среднее настроение, сглаженное (новости и события)	0,000	0,000	0,000	0,000	0,000	0,000	0,000
Среднее настроение, сглаженное (пользовательские посты по компании)	0,000	0,000	0,000	0,000	0,000	0,000	0,000
Стандартное отклонение настроения, сглаженное (пользовательские посты по компании)	0,000	0,000	0,000	0,000	0,000	0,000	0,000
Прирост логарифма количества постов (пользовательские посты по компании)	0,000	0,000	0,000	0,091	0,052	0,000	0,026
Логарифм количества постов (пользовательские посты по МосБирже)	0,000	0,026	0,046	0,004	0,078	0,050	0,003



IV. ПРОГНОЗИРОВАНИЕ

Модель настроения. Лассо регрессия и дополнительные модели

Прикладная экономика

Таблица 14. Значения среднеквадратичной ошибки MSE для модели настроения на тестовой выборке для каждой отдельной компании. По столбцам слева направо: регрессия лассо, случайный лес, экстремальный градиентный бустинг, расширенный градиентный бустинг. Таблица построена автором.

Компания	Lasso	RF	XGB	LGBM
Яндекс	0.158	0.121	0.118	0.123
Сбербанк	0.206	0.200	0.183	0.208
МТС	0.222	0.239	0.267	0.284
Лукойл	0.212	0.184	0.186	0.195
Роснефть	0.218	0.182	0.192	0.185
Газпром	0.241	0.227	0.215	0.224
Новатэк	0.260	0.218	0.257	0.224

Таблица 15. Уменьшение среднеквадратичной ошибки MSE на тестовой выборке в модели настроения относительно базовой модели, %. По столбцам слева направо: регрессия лассо, случайный лес, экстремальный градиентный бустинг, расширенный градиентный бустинг.

Компания	Lasso	RF	XGB	LGBM
Яндекс	52%	63%	64%	62%
Сбербанк	6%	9%	17%	5%
МТС	-19%	-28%	-43%	-52%
Лукойл	-2%	12%	11%	6%
Роснефть	-11%	7%	2%	6%
Газпром	5%	11%	15%	12%
Новатэк	-5%	12%	-4%	9%

- В плохом смысле выделяется только компания **МТС**, для нее экономические факторы и факторы настроения не дали никакого прироста в точности.
- **Яндекс, Сбербанк и Газпром** хорошо отреагировали на добавление экономических факторов и факторов настроения.
- **Лукойл, Роснефть и Новатэк** дали неоднозначные результаты.
- В среднем, если не брать в учет МТС, модель настроения дает **прирост в 15%** в качестве прогноза по сравнению с бенчмарком.



IV. ПРОГНОЗИРОВАНИЕ

Сравнение экономической модели и модели настроения

Прикладная экономика

Таблица 16. Уменьшение среднеквадратичной ошибки MSE на тестовой выборке в модели настроения относительно экономической модели, %. По столбцам слева направо: регрессия лассо, случайный лес, экстремальный градиентный бустинг, расширенный градиентный бустинг. Таблица построена автором.

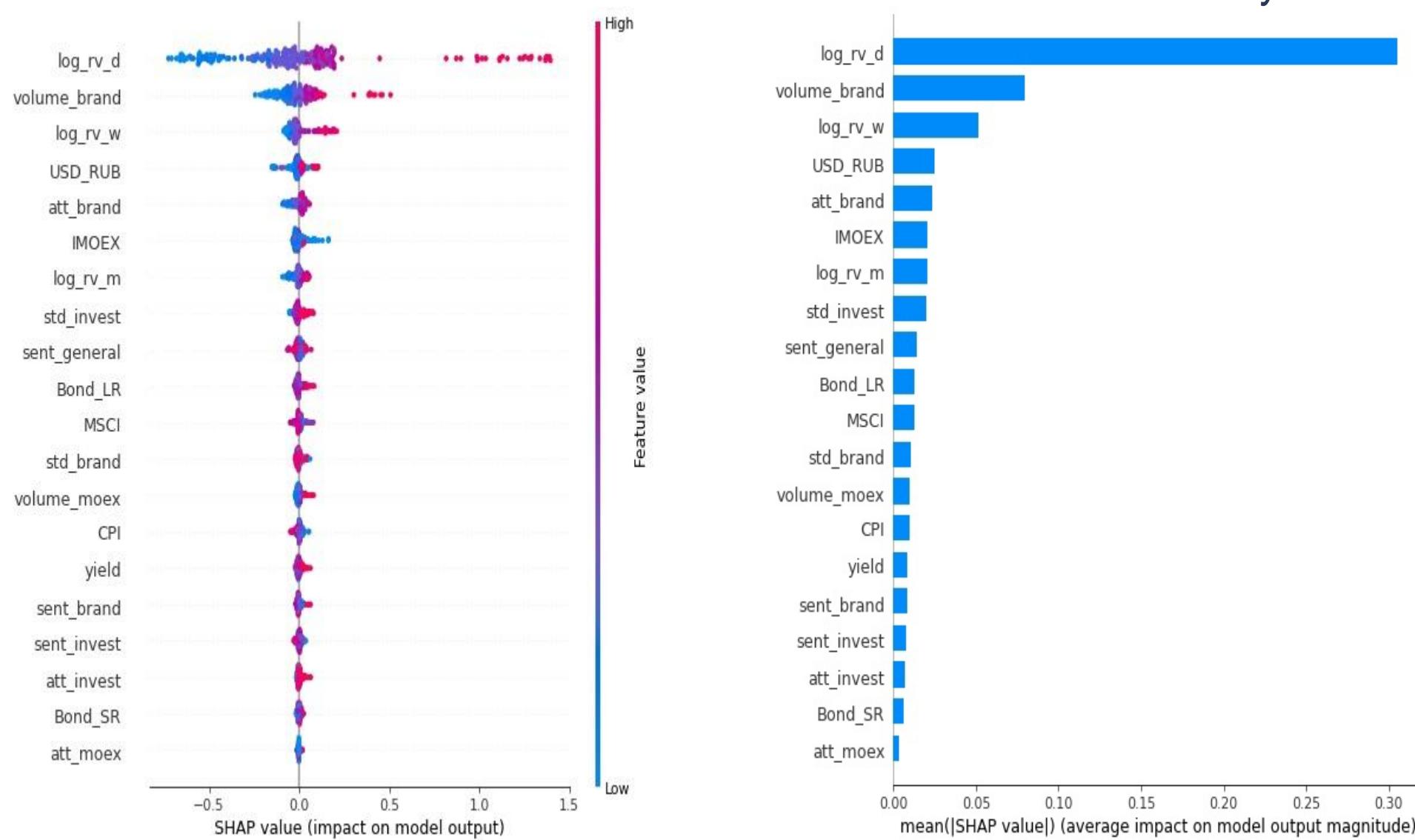
Компания	Lasso	RF	XGB	LGBM
Яндекс	1%	2%	18%	7%
Сбербанк	3%	3%	9%	-10%
МТС	13%	-3%	-2%	0%
Лукойл	-11%	14%	3%	3%
Роснефть	7%	2%	-5%	14%
Газпром	-3%	2%	2%	0%
Новатэк	1%	2%	-8%	10%

- Стоит отметить компанию **Яндекс**, каждый алгоритм с добавлением признаков по данным социальных сетей улучшил точность прогноза.
- В среднем для всех компаний добавление признаков настроения и внимания к экономической модели **улучшают** качество прогноза **на 3%**.

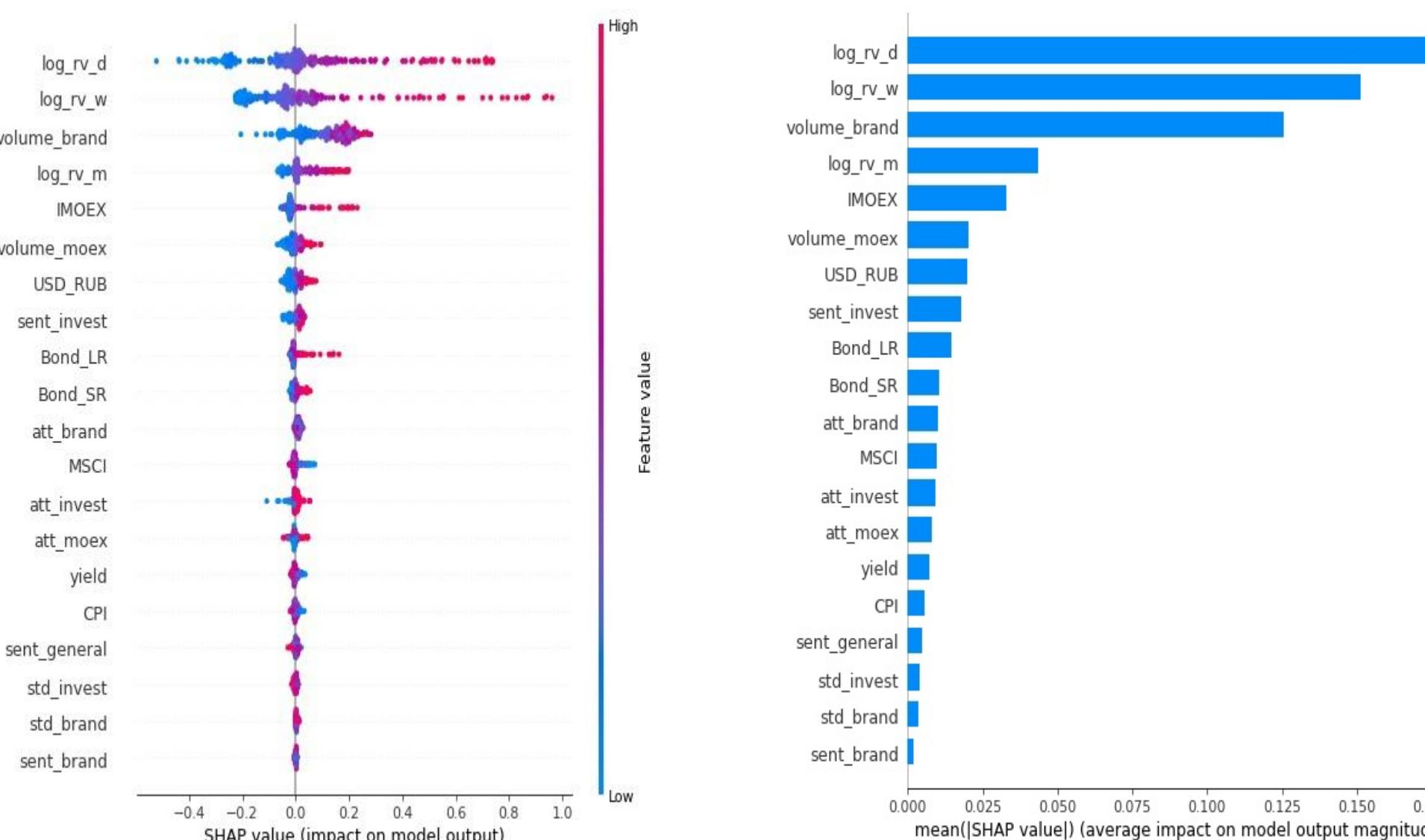
IV. ПРОГНОЗИРОВАНИЕ

Важность факторов в модели настроения для компании «Яндекс»

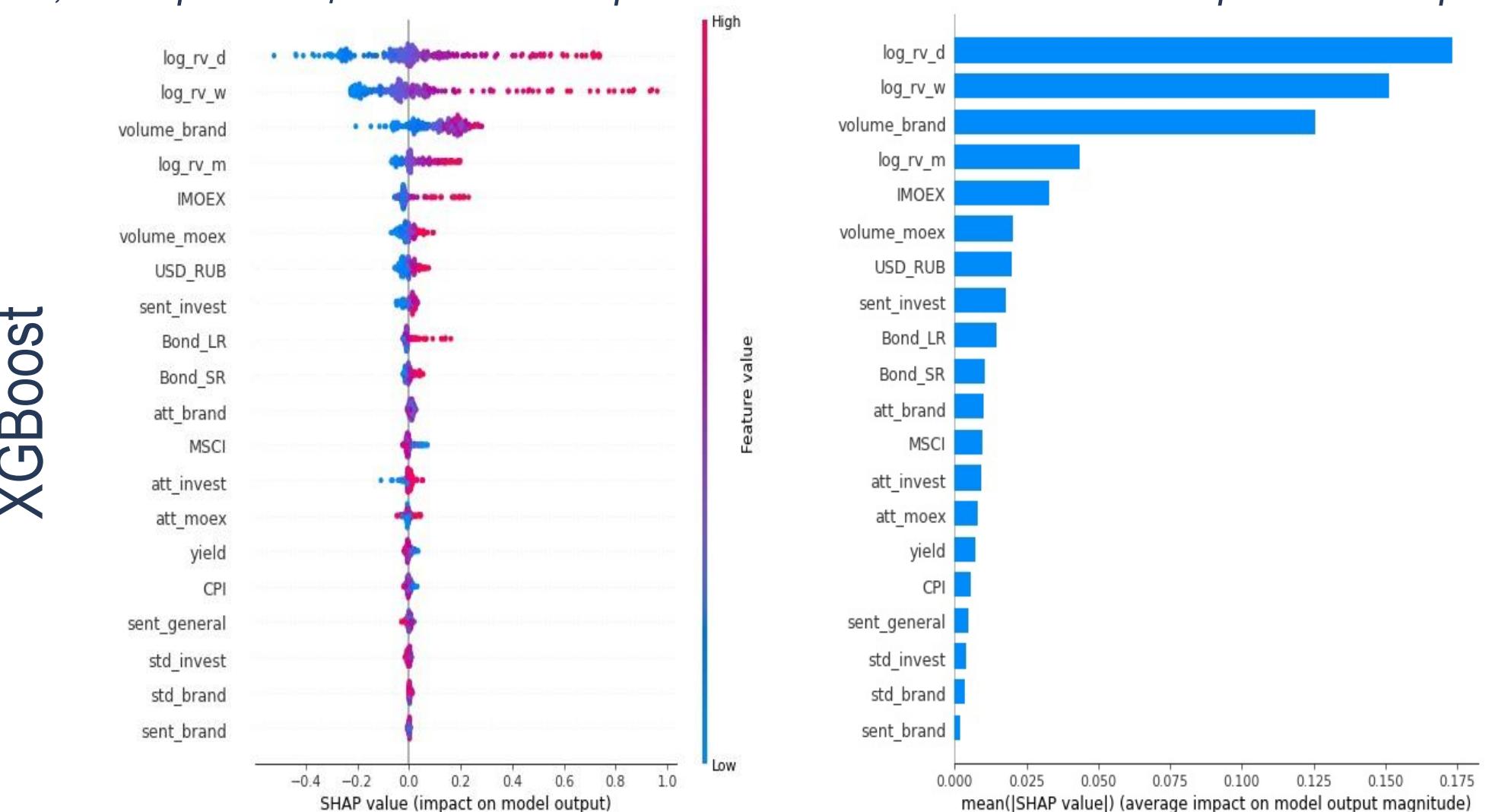
Случайный лес



Light GBM



XGBoost



Атрибут	Фактор
log_rv	Логарифм реализованной волатильности
att_invest	Прирост логарифма количества постов (финансы и инвестиции)
IMOEX	Индекс IMOEX
MSCI	Индекс MSCI Russia
Bond_LR	Доходность облигации России 10-летние (годовая ставка), %
Bond_SR	Доходность облигации Россия годовые (годовая ставка), %
USD_RUB	Курс доллара к рублю, руб
yield	Доходность акции, %
volume_brand	Объем торгов, тыс. руб.
volume_MOEX	Объем торгов индекса МосБиржи, тыс. руб.
CPI	Индекс потребительских цен (ИПЦ)
att_moex	Логарифм количества постов (пользовательские посты по МосБирже)
sent_invest	Среднее настроение, сглаженное (финансы и инвестиции)
std_invest	Стандартное отклонение настроения, сглаженное (финансы и инвестиции)
sent_general	Среднее настроение, сглаженное (новости и события)
std_general	Стандартное отклонение настроения, сглаженное (новости и события)
att_brand	Прирост логарифма количества постов (пользовательские посты по компании)
sent_brand	Среднее настроение, сглаженное (пользовательские посты по компании)
std_brand	Стандартное отклонение настроения, сглаженное (пользовательские посты по компании)



IV. ПРОГНОЗИРОВАНИЕ

Сравнение экономической модели и модели настроения

Прикладная экономика

Таблица 17. Лучшие алгоритмы прогнозирования для экономической модели и модели настроения (2, 4 столбцы), их среднеквадратичные ошибки на прогнозе (3, 5 столбцы), уменьшение среднеквадратичной ошибки MSE на тестовой выборке в модели настроения относительно экономической модели, % (последний столбец). Таблица построена автором.

Компания	Базовая модель	Экономическая модель		Модель настроения		Уменьшение MSE, %
	MSE	Метод	MSE	Метод	MSE	
Яндекс	0.326	RF	0.123	XGB	0.118	4%
Сбербанк	0.220	LGBM	0.189	XGB	0.183	3%
Лукойл	0.208	XGB	0.191	RF	0.184	4%
Роснефть	0.196	XGB	0.183	RF	0.182	1%
Газпром	0.254	XGB	0.219	XGB	0.215	2%
Новатэк	0.247	RF	0.223	RF	0.218	2%

- Социальные сети в целом **положительно** влияют на точность прогноза.
- Уменьшение** в MSE на teste в среднем **3%**.
- Лучшие результаты у компаний **Яндекс и Сбербанк**, что значит, признаки внимания и настроения оказывают более сильное влияние на них, чем на другие рассмотренные компании.



V. ВЫВОДЫ

Прикладная экономика

- В последнее десятилетие растет значимость социальных сетей. Они используются уже не только как средство коммуникации между пользователями, но и как важный аналитический инструмент, позволяющий собирать количественную и качественную информацию. Twitter - идеальная платформа для получения общественного мнения по конкретным вопросам, так информация представлена в удобной, легкоредактируемой для анализа форме.
- Внимание и настроение инвесторов оказывают влияние на реализованную волатильность. Комментарии в социальных сетях также оказывают значимое влияние на акции компаний. При включении факторов внимания и настроения в экономическую модель, качество прогноза модели растёт.



V. ВЫВОДЫ

Критика

Прикладная экономика

Можно :

- проверить другие модели, которые могут показать более высокие значения качества. В данной работе использовался ряд инструментов и методов моделирования, которые могут дать высокие показатели качества. Однако, современные тенденции развития машинного обучения свидетельствуют о том, что вероятность возникновения новых, более совершенных методов анализа данных крайне высока. Поэтому результаты работы необходимо обновлять;
- найти размеченные данные по финансовым новостям, проанализировать их, сравнить с исходной моделью;
- проверить другие российские компании, сравнить их показатели с результатами работы;
- использовать данные компаний с сайтов Bloomberg и т.п..



НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ

Комментарии к рецензии

Разделы 1.2 и 1.3 полностью являются компиляцией местами не очень качественных, но дословных переводов кусков описаний соответствующих моделей, при этом источник заимствования текста указан только эпизодически — в то время, как оригинального авторского текста в этих разделах практически нет: это всё — компиляция дословных переводов.

В пункте 1.3.2 не использовался ни один иностранный источник, если он не указан в тексте. Весь текст был написан на основе прослушанных лекций. Местами (в преимущественном меньшинстве) использовались статьи с сайтов. Если это происходило, соответствующий источник также указан. Переводов в тексте нет.

В пункте 1.2 происходит анализ социальных сетей, обосновываются причины выбора Twitter в качестве основного источника получения данных (из данных социальных сетей). На эту подглаву приходится 9 источников. Возможно, мы неправильно оформили цитирования. В любом случае, все обозначенные куски текста выявлены Антиплагиатом. Мы в данном случае пользовались требованиями, которые загружены на странице программы.

Процент заимствований:



Вся глава посвящена теоретическим основам нашей темы. В главе 1 необходимо раскрыть основные понятия, которые задействованы в названии работы. Мы выявили следующие термины для разбора в 1 главе: Социальные сети, Оценка настроений инвесторов, Тональности текста. Для 2-ой главы использовались ключевые слова: Реализованная волатильность, Факторы внимания и настроения и др. Некоторые из этих ключевых слов были перенесены в Аннотацию (в абзац «Ключевые слова»). Если бы мы не рассмотрели эти темы, то получилось бы так, что наша работа недостаточно детализирована с точки зрения теории. Контролировать вероятность возникновения компиляций в 1-ой главе крайне сложно, оригинальные части там тоже есть и их достаточно много.

Комментарии к рецензии!

Текст на стр. 66 (формула без номера и пояснения к ней) иллюстрируют непонимание авторами основ финансовой экономики.

Так, упомянутая формула и пояснения к ней — плагиат с сайта finam.ru (без указания источника заимствования), однако плагиат совершенно не уместный: эта формула — для т. н. «простой доходности к погашению» и не применима для доходности к погашению, которая приведена в использованном авторами источнике данных (см. табл. 9).

Более того, судя по тексту пп. 1 и 3 на стр. 66, авторы не понимают разницы между доходностью облигации за 10 лет и доходностью к погашению 10-летней облигации.

Да, источник, действительно не указан (хотя указания на заимствования из сайта «finam.ru» есть как во всей главе, так и в списке источников). Примечание: этот кусок текста был определен Антиплагиатом, как неоригинальный, и уже учитывался при расчете оригинальности работы. Также признаем, что данная информация не совсем вписывается в общую канву рассуждения. Она дана, скорее, для напоминания (соответственное слово «Напомним» - см. стр. 66 - присутствует в тексте). Номера к данной формуле нет, так как она используется для справки и не влияет на дальнейшее рассуждение. Опять же, мы признаем, что, скорее всего, не стоило добавлять эту формулу в работу (к тому же она только подпортила статистику оригинальности).

Комментарии к рецензии!

Раздел 3.1 также является компиляцией источников из интернета, но на этот раз более качественной (что можно объяснить тем, что использованы в основном русскоязычные источники). Как и в других частях, ссылки на источники переработки практически не приведены (источники приведены только к рисункам, однако помимо рисунков из этих же источников заимствован и текст — притом зачастую дословно).

Ссылки не приведены, т.к. материал основан на прослушанных лекциях, а не на доступных письменных источниках. Картинки все взяты с интернета, потому что не вижу смысла рисовать схемы самостоятельно, когда всё это уже есть. Если текст непосредственно взят из сайта, он указан как источник.

Описание использованных переменных (на стр. 63--69) — прямой плагиат из работы Audrino F., Sigrist F., Ballinari D. The impact of sentiment and attention measures on stock market volatility //International Journal of Forecasting. – 2020. – Т. 36. – №. 2. – С. 334-357 (этого источника тоже нет в списке цитирования).

Мы не считаем, что вся информация на странице 63-69 является плагиатом. Это подтверждается тем, что некоторые переменные относятся к российской практике и не могли присутствовать в статье, которую указал рецензент.

Данный автор указан в списке цитирований, но с другой статьей: Corsi, F., Audrino, F. and Reno, R. (2012). HAR Modeling for Realized Volatility Forecasting. In: Handbook of Volatility Models and Their Applications. (pp. 363-382). New Jersey, USA: John Wiley & Sons, Inc. ISBN 9780470872512. В списке литературы статья тоже указана.

Комментарии к рецензии!

Таким образом, примерно 80 страниц из общего объема работы без приложений и списка литературы в 120 страниц не являются оригинальными. Более того, эти страницы — «вода»: их исключение из текста не приведет к усложнению понимания — а даже наоборот, к упрощению.

Да, действительно, не все источники оказались включенными в список литературы. В список литературы были включены основные работы. Тем не менее, список литературы не влияет на общую оригинальность работы. Все же, мы признаем свою ошибку, в дальнейшем будем указывать все источники, как в сносках, так и в списке литературы. В любом случае, список литературы содержит 54 источника, размещенных на 6 страницах. Думаю, добавить оставшиеся источники не составит большого труда. Мы в любой момент можем это сделать для улучшения качества работы.

Основная причина приведения подробного изложения используемых методов связана с тем, что мы используем методы и подходы, обычно не используемые в экономических исследованиях. И вряд ли каждый экономист поймет, какой смысл заложен в одних названиях моделей. Все описанные теоретические основы прямо связаны с исследованием. И все модели, которые описаны в тексте, используются в работе. Тем не менее, даже убрав это 80 страниц, наша работа продолжает вписываться в минимально необходимый объем.

В списке литературы указаны не все цитированные работы. Притом это не единичные работы. Так, наугад взятой стр. 11 цитируется 8 работ, 6 из которых не указаны в списке литературы — это работы Ho et al. (2013), Dimpfl and Jank (2015), Hamid and Heiden (2015), Zhang et al. (2016), See-To and Yang (2017), Siganos et al. (2017).

Комментарии к рецензии!

Оформление работы разнородно и местами небрежно: используются разные стили цитирования (подстраничные сноски и в конце документа; часть фамилий транслитерирована, часть оставлена на языке оригинала), подписей и т. п. На стр. 21 в сноске 5 очень странно расставлены пробелы: «U nivers ityof Porto» (скорее всего, это появилось в результате бездумного copy&paste).

Да, мы согласны, текст писал не один человек, и поэтому стиль текста может быть разным. Но, действительно, стоило уделить дополнительное время на полное редактирование всего текста.

Неправильное оформление сноски, возможно, объясняется ошибкой при переводе из формата Word в формат PDF. В оригинальном документе такой ошибки нет. В требованиях мы нашли указание, что ссылки на источники можно оформлять либо в виде сносок, либо в скобках. Показалось, что внутри работы допустим и тот и другой варианты, и это не нарушает допустимый формат.

Комментарии к рецензии!

Обзор литературы чрезмерно краток и неполон. В частности, в нём отсутствуют многие более новые работы, послужившие источником плагиата в других частях текста.

Обзор литературы в силу структуры нашей работы произведен на протяжении всей работы, а не только в пункте «Обзор литературы». И имеет место в пунктах 1.3.2, 2.1, 2.2, 2.3

В требованиях нет указаний на необходимый объем «Обзора литературы». В поддержку коллеги выше, могу отметить еще и то, что сложно предугадать заранее условия, которые не указаны в требованиях. На данный момент параграф с обзором литературы состоит из 5-и страниц (при минимально допустимом объеме ВКР в 75 страниц).

Мы согласны. Но, во-первых, в основной статье тоже использовались *ex ante* переменные. Во-вторых, они использовались исключительно в том смысле, чтобы посмотреть на то, влияют ли они на целевую переменную. Ведь, если видящие будущее переменные не имеют значимости даже с учетом того, что учитывают информацию о будущем, рассмотрение макроэкономических переменных вообще теряет смысл. В-третьих, для прогнозирования мы не использовали такого рода переменные.

Линейная интерполяция месячных данных к дневным приводит к использованию данных из будущего (для линейной интерполяции необходимы наблюдения за предыдущий и за следующий периоды). Таким образом, «все макроэкономические переменные, кроме доходности индекса CRB и фьючерса на нефть Brent» являются вперёдсмотрящими признаками.

Комментарии к рецензии!

Результаты в таблице 26 нуждаются в уточнении. Помимо того, что необходимо проверить приведённые различия на статистическую значимость, меня очень смущает строка «МТС», в которой модель с большим количеством признаков показала ощутимое снижение качества — вплоть до 50%. Это может свидетельствовать об одном или нескольких обстоятельствах из перечисленных ниже:

- различные компании довольно сильно отличаются друг от друга (также косвенно подтверждается отличием Яндекса от остальных) — тогда полученные результаты нерепрезентативны;
- происходит сильное переобучение моделей — тогда остальные результаты тоже не заслуживают доверия

Указание на рассмотрение компаний с разной направленностью встречается в 5 частях работы: Аннотации (на рус. и англ. языках), Введении, Главе 2 (в нескольких местах), Главе 3

Ряд моделей просто не может переобучиться в силу своих свойств. В регрессии мы также включали регуляризацию, дабы избежать переобучения.

Изначально мы начинали работу с гипотезой о том, что компании отличаются друг от друга. И специально брали такие разрозненные компании.

В базовой модели использовалась линейная регрессия, в экономической — регрессия лассо. Если бы и там, и там была бы одна линейная регрессия, было бы и правда странно получить понижение качества модели при добавлении признаков на ОБУЧАЮЩЕЙ выборке. А у нас ошибка считается на тестовой. В основной статье так же присутствует понижение качества прогноза при добавлении атрибутов для ряда компаний. Так же стоит отметить, что модели независимые для каждой компании, и если на одну компанию факторы влияют сомнительно, не следует того что для других компаний результаты нерепрезентативны.

Комментарии к рецензии!

Методология эмпирического исследования, использованная авторами, по уровню ощутимо уступает методологии научных работ, процитированных или сплачиенных в ВКР. То есть, нельзя сказать, что авторы не были знакомы с современным положением дел в этой области (по меньшей мере, они открывали соответствующие статьи), однако использованная методология слишком проста. Так, например, уже упомянутая работа Audrino et al. (2020): она посвящена тому же самому вопросу, отсутствует в обзоре литературы, но является источником не оформленных заимствований, а главное — эмпирическое исследование в этой работе выполнено гораздо более полно и тщательно.

Как удалось выяснить, у данного автора (Audrino) довольно много работ (в том числе и по нашей теме). Опять же, рецензент не может достоверно утверждать, использовали ли мы те или иные тезисы именно из конкретной статьи. К тому же, некоторые тезисы (это выяснилось позднее) дублируются в разных работах. В современном мире крайне сложно изучить все статьи, в которых задействована та или иная тема (пусть даже и конкретного автора). Главная идея нашей работы заключалась в использовании данных российских компаний (об этом было сообщено сразу, во «Введении»). Добавлю, что кажется несколько странным аргумент об использовании методологии, которая встречается в некоторой статье. Плюс, нигде не нашел, что нельзя использовать некоторые аспекты методологии других работ. Конкретно в нашей работе есть ссылки на другие работы по схожей тематике.

Комментарии к рецензии!

Содержательная часть работы мала и содержит ряд методологических ошибок.

Мы не увидели ни одного замечания к используемой методологии, за исключением использования интерполяции. Стоит учитывать то, что никто из нас до этого года не был знаком с машинным обучением и питоном. И очень большие трудозатраты заняло обучение. Думаю, стоит учитывать то, что используемых методологий не найти в учебнике по эконометрике или временным рядам. И даже в курсах ФКН или на Coursera нет готовых кодов и аналогичных задач.

Содержательная часть работы заключается как минимум в анализе социальной сети Twitter, выборе показателей для анализа как экономических, финансовых показателей, так и показателей, относящихся к социальным сетям. К сожалению, нет ни единого комментария относительно Приложения с кодом (на 35 страниц). Напомним, что работы, где активно применялись языки программирования, должны оцениваться выше (при прочих равных). Также нет оценки Приложений, где освещены итоги работы. Хочется, отметить, что в нашем случае, Приложения играют существенную роль (все ссылки на Примечания были добавлены в ходе работы). Мы не стали добавлять их в основную часть в виду их обилия.

Итоговый комментарий

В этом учебном году мы перешли на новый формат сдачи ВКР (экспериментально). Предположительно, наша работа должна соответствовать: 1) выбранному НИС; 2) направлению обучения. Тема ВКР в данном случае удовлетворяет обоим требованиям.

Общий показатель плагиата в нашей работе составляет 7 процентов. Соглашусь, что при правильном оформлении цитирования, процент заимствований мог бы оказаться ниже (около 2-3%), так как цитирования в данной программе АП не учитываются за плагиат. Тем не менее, мы уже были «оштрафованы» за невключение цитат более низкой оригинальностью. Думаю, повторно «штрафовать» за то же самое не совсем корректно. К тому же, АП «штрафует» нас и за устойчивые словосочетания, длиной более 2 слов.

Подчеркнем, что если текст был заимствован из русскоязычной статьи, то Антиплагиат уже обнаружил это и снизил оригинальность. Если мы не ошибаемся, то очевидные компиляции АП «Высшей школы экономики» также обнаруживает. К сожалению, мы не можем изучить все существующие материалы по нашей теме ввиду их обилия и существования смежных тем.

В заключение работы присутствует критика. Мы понимаем, что работа неидеальна, что есть, куда расти, необходимы доработки. Но тема исследования довольно интересна и важна. Мелкие ошибки есть в любой работе. Мы гарантируем, что для дальнейшей работы исправим все существующие недочеты.

P.S. Также мы планируем опубликовать нашу работу в российском журнале с высоким импакт-фактором.