

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ  
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

**Факультет экономических наук**

**Образовательная программа «Прикладная экономика»**

Ан Ен Ми

Миткинов Валентин Евгеньевич

Шайдуллин Ансэль Ильгизович

**«Предсказание волатильности фондового рынка с использованием  
данных социальных сетей»**

Выпускная квалификационная работа - МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ  
по направлению подготовки 38.04.01 Экономика

**Рецензент**

**доцент, к.э.н.**

**Лапшин Виктор  
Александрович**

**Руководитель**

**доцент, к.э.н.**

**Мамедли**

**Мариам Октаевна**

Москва 2021

## АННОТАЦИЯ

Актуальность выбранной темы заключается в том, что в современном мире растет роль использования различных методов, при помощи которых можно анализировать риски и строить прогнозы. Материалы работы представляют практическую ценность для банковских организаций и инвесторов, которые могут применять указанную модель для прогнозирования рисков и угроз.

В работе была проанализирована реализованная волатильность акций ряда российских компаний, которые относятся к разным отраслям экономики: от сырьевых компаний – нефть, газ (Лукойл, Роснефть, Газпром, Новатэк) до компаний, занятых в сфере услуг – банковский сектор (Сбербанк), ИТ (Яндекс), телекоммуникации (МТС). Как выяснилось, разные компании (с разной направленностью) по-разному реагируют на настроения и/или новости в социальных сетях.

В работе так же были предложены практические рекомендации по выбору области применения полученных результатов и моделей.

*Ключевые слова:* тональность новостей, реализованная волатильность, фондовый рынок.

УДК: 004.852; 004.043; 336.76.066; 336.761.532; 336.763.215; 336.763.218

JEL: G17, G32

## СОДЕРЖАНИЕ

Введение.....	4
Обзор литературы.....	9
Глава 1 Теоретические основы построения модели на основе данных социальных сетей .....	13
1.1 Общие подходы к измерению внимания и настроения инвесторов .....	13
1.2 Социальные сети как источник измерения настроения инвесторов, причины выбора Twitter для проведения исследования .....	19
1.3 Модели тональности текстов .....	25
1.3.1 Общий подход к измерению тональностей текста и особенностей языка.....	27
1.3.2 Методы обработки естественного языка .....	31
Глава 2 Исследование волатильности на фондовом рынке и факторов, влияющих на ее динамику.....	60
2.1 Реализованная волатильность как альтернативный метод оценки динамики фондового рынка .....	60
2.2 Экономические и финансовые факторы, влияющие на фондовый рынок .....	64
2.3 Факторы внимания и настроения на основе данных Twitter и обучение модели тональности текстов .....	73
Глава 3 Краткосрочное прогнозирование реализованной волатильности .....	73
3.1 Теоретические основы используемых методов машинного обучения...	83
3.2 Обучение базовой и экономической модели и прогнозирование .....	96
3.3 Обучение модели настроения и прогнозирование .....	110
Заключение .....	118
Список литературы .....	123
Приложения .....	130

## **Введение**

*Актуальность выбранной темы* заключается в том, что в современном мире растет роль использования различных методов, при помощи которых можно анализировать риски и строить прогнозы. Данная тенденция наблюдается во многих сферах, в том числе и в финансовой. В целом, важность предложенной работы можно выразить в следующих тенденциях:

- Рост необходимости снижения риска финансовых потерь методом построения прогнозов на основе различных баз данных;
- Рост популярности социальных сетей, в т.ч. и как средства получения информации (баз данных);
- Наличие корреляции между настроениями в социальных сетях и волатильностью акций компаний.

Результаты работы можно применять в контексте: 1) теоретических знаний; 2) прикладных аспектов. Охват сфер применения результатов исследования также широк: от государственного и муниципального управления до анализа тональности новостей в средствах массовой информации.

Сегодня роль информации неуклонно растет. Объемы поступающей информации, по сравнению, с аналогичными показателями начала 21 века, значительно выросли. К сожалению, подобный рост не всегда коррелирует с качеством этой информации. Анализ тональности новостей позволяет выявить ключевые слова в тексте, и на основе данных делать выводы о качестве сообщения и/или новости.

Именно поэтому возникает потребность грамотной и эффективной обработки данных. На основе информации выше определим цели и задачи настоящего исследования.

*Цель работы* - на основе данных социальных сетей методами машинного обучения построить адекватную модель для оценки и прогнозирования волатильности фондового рынка.

*Основные задачи работы:*

- изучить базовые статьи по темам, посвященным прогнозированию волатильности (реализованной) акций на основе различных видов данных, выделить основные тезисы и методики работы;
- собрать данные по волатильности, экономическим и финансовым показателям; предварительно обработать данные;
- собрать данные из социальной сети Twitter, отфильтровать релевантные твиты;
- обучить классификатор ULMFiT тональности текста (библиотека fast.ai) на автоматически размеченных данных, классифицировать собранные твиты;
- сформировать переменные внимания (количественные характеристики) и настроения (тональность твитов) по данным социальных сетей
- построить пять линейных и нелинейных моделей – линейная регрессия, регрессия лассо, случайный лес, XGBoost модель (градиентный бустинг) и LightGBM (расширенный градиентный бустинг);
- обучить на in sample (обучающей) выборке вышеперечисленными методами три модели:
  - базовая модель (включены переменные реализованной волатильности и ее производных);
  - экономическую модель (также включены экономические и финансовые показатели);
  - модель настроения (также включены переменные внимания и настроения)
- спрогнозировать реализованную волатильность на день вперед на out-of-sample (тестовой) выборке; проанализировать степень влияния

экономических и финансовых переменных, переменных настроения и внимания;

- сравнить методы машинного обучения, выявить преимущества и недостатки;
- определить сферы возможного применения полученной модели;
- на основе полученных результатов сделать выводы и предложить способы для практического применения модели.

*Новизна исследования* заключается в следующем:

- построение новой модели, адаптированной для российского финансового рынка;
- поиск корреляции между зависимой и независимой переменной в условиях предложенных моделей. (То есть, как изменения в настроениях влияет на реализованную волатильность акций российских компаний);
- выбор наилучшей модели для прогнозирования реализованной волатильности фондового рынка на основе твитов на русском языке;
- используется большой набор переменных настроения и внимания, полученных из различных источников с большим объемом данных

В данной работе будут произведены попытки ответить на следующие вопросы:

- 1) влияют ли социальные сети, на будущую реализованную волатильность при контроле других экономических переменных?
- 2) какой тип оценки настроения или внимания наиболее значим?
- 3) отличается ли степень влияния данных Twitter на волатильность для компаний различных отраслей экономики?
- 4) в каких сферах можно применить полученные наработки, как их эффективно использовать?

*Для проведения исследования* использовались разные методы исследования: эмпирический, статистический, графический, теоретический, а также методы машинного обучения и нейронные сети. Также активно применялись инструменты программирования, был использован язык

программирования Python: библиотеки Numpy, Pandas, Matplotlib, SciKit-Learn, XGBoost, LightGBM, Seaborn, Statistics, StatsModels, Fast.ai, Csv, Snsccape, Shap, SciKit Optimize. Мы использовали сервис Google Collab при работе с трудозатратными вычислениями: он позволяет пользоваться удаленным доступом, подключенным к машине с GPU (графический процессор).

*Объект исследования* – акции российских компаний, их волатильность за 5-летний промежуток времени с 1 марта 2016 по 28 февраля 2021. В работе анализируется реализованная волатильность, которая позволяет снизить влияние резких флуктуаций, а, следовательно, и вероятность неправильной интерпретации результатов.

*Субъект исследования*. В данной работе были проанализированы различные вопросы, которые касаются тематики изменения динамики цен акций вследствие изменения тональности новостей в социальных сетях.

*Предметом работы является* анализ основных экономических, финансовых и других показателей, способных повлиять на реализованную волатильность фондового рынка. Как было отмечено выше, на данные показатели могут влиять и настроения в социальных сетях и в новостях.

При написании работы были проанализированы различные источники: учебные издания, электронные источники, статьи из газет и журналов. Также были изучены различные методы машинного обучения через платформы Coursera, Stepik, Open Data Science: Open Course “Machine Learning and Deep Learning”; курсы в открытом доступе МИТ Deep Learning School; GitHub. Для того чтобы увеличить достоверность работы были использованы статистические данные с сайтов Московской биржи (крупнейшего российского биржевого холдинга), InvestCom, а также данные с официальных государственных сайтов (сайт Центрального Банка и Росстата). Благодаря этим данным и иной полученной информации, удалось построить правильное и достоверное представление об исследуемом секторе экономики.

Структура данной исследовательской работы выглядит следующим образом:

- 1) В первой главе рассматриваются теоретические вопросы анализа тональности новостей, обсуждаются возможности проверки настроения инвесторов, анализируются социальные сети, на основе которых будет проводиться дальнейший анализ;
- 2) Во второй главе происходит выбор ключевых показателей для оценки влияния новостей и сообщений в социальных сетях на реализованную волатильность, оцениваются акции российских компаний;
- 3) В третьей главе выбирается базовая модель, оценивающая исследуемые параметры, строится краткосрочный прогноз на основе полученных данных, предлагаются возможные пути практического применения модели.

В заключении подводятся итоги проделанной работы, выявляются слабые места исследования, даются рекомендации по исправлению и/или улучшению полученных моделей.

## **Обзор литературы**

Проблема влияния тональности новостей на волатильность акций не раз поднималась в различных научных статьях. Согласно гипотезе эффективного рынка (Fama, 1970), прогнозирование доходности акций не должно быть возможным, поскольку рыночные цены будут отражать всю доступную информацию. С другой стороны, с начала 1990-х годов исследователи поведенческих финансов сообщают о растущем количестве эмпирических данных, показывающих, что фондовый рынок управляемся психологией инвесторов (Daniel, Hirshleifer и Teoh, 2002). Существуют различные объяснения этого открытия в области поведенческих финансов (Tseng, 2006), такие как, например, предвзятость неправильной атрибуции (неправильной интерпретации причинно-следственной связи, нарушение логической последовательности), согласно которой люди принимают рискованные решения в зависимости от состояния своего настроения (Johnson & Tversky, 1983). В данных работах даются первые попытки проанализировать влияние психологического фактора на экономические показатели.

Следующим этапом в оценке психологических факторов в экономике является оценка факторов чувств и эмоций. Переменные настроения и внимания имеют значительную предсказательную силу для будущей волатильности. Переменные, имеющие наибольшую прогностическую важность, - это показатели внимания инвесторов, такие как объем публикаций в социальных сетях или количество запросов в поисковых системах. Наблюдается наибольшее повышение точности прогнозов благодаря данным о настроениях и внимании для компаний с большой рыночной капитализацией и/или низкой долей институциональных инвесторов. Кроме того, результаты показывают, что данные микроблогов, поисковых систем и новостных статей имеют особую прогностическую силу, когда волатильность велика. То есть, наблюдается прямая корреляция

между величиной флуктуаций/колебаний волатильности и данными в социальных сетях и новостных порталах. Включение переменных настроения и внимания приводит к повышению точности прогнозов по сравнению с моделью, которая включает только экономические переменные.

Есть ряд исследований, в которых изучается, в какой степени переменные настроения и внимания, полученные из социальных сетей и других интернет-платформ, могут быть использованы для прогнозирования доходности финансового рынка. Некоторые исследования приходят к выводу, что настроения общественности, полученные из социальных сетей, таких как Twitter или LiveJournal, или объемы интернет-активности могут быть использованы для прогнозирования движений фондового рынка. Например, Da et al. (2011) показывают, что повышенное внимание инвесторов предсказывает рост цен в краткосрочной перспективе и разворот цен в долгосрочной перспективе.

В более позднем исследовании Sun et al. (2016) 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. В соответствующем исследовании Renault (2017) показано, что первое получасовое изменение настроений из сообщений, опубликованных на StockTwits, имеет предсказательную силу для последнего получасового возврата.

Другие исследования более скептически относятся к предсказательной способности социальных сетей (Antweiler & Frank, 2004; Oliveira et al., 2013; Tumarkin & Whitelaw, 2001). Анализ сообщений, размещенных на Yahoo! Finance and Raging Bull, Antweiler и Frank (2004) обнаружили, 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. Кроме того, Oliveira et al. (2013) не обнаружили доказательств предсказуемости доходности с помощью сообщений, размещенных на StockTwits. В целом, нет единого мнения относительно того, действительно ли можно предсказать доходность фондового рынка с помощью анализа настроений (Schoen et al., 2013). Тем не менее, есть основания предполагать, что настроения в социальных сетях способны влиять на стоимость той или иной акции. Отдельные события/явления так же способны оказывать влияние на фондовый рынок (см. новость о «курении» Илона Маска<sup>1</sup> в прямом эфире или его «сообщение о покупке криптовалюты»). Первая новость вызвала «бурю обсуждения/негодования» в Интернете, из-за чего стоимость акции компании Tesla упала. Вторая история связана с твитом Маска о покупке криптовалюты, что в свою очередь привело к росту заинтересованности криптовалютой.<sup>2</sup>

Точно так же несколько исследователей обнаружили, что настроения в социальных сетях, новости или объемы поисковых запросов могут использоваться для прогнозирования волатильности фондового рынка (Aouadi, Arouri & Teulon, 2013; Behrendt & Schmidt, 2018; Bordino et al ., 2012). Antweiler и Frank (2004) обнаружили, что настроение в биржевых сообщениях, размещенных на Yahoo! могут помочь предсказать волатильность фондового рынка. Wang et al. (2006) подчеркивают важность контроля над другими экономическими и финансовыми переменными при

---

<sup>1</sup> Один день Илона Маска после курения в прямом эфире. [Электронный ресурс]. Режим доступа: <https://tjournal.ru/internet/76217-odin-den-ilona-maska-posle-kureniya-marihuany-v-pryamom-efire> (дата обращения: 30.04.2021)

<sup>2</sup> Илон Маск купил Dogecoin, Tesla — биткоин. Главные новости недели. [Электронный ресурс]. Режим доступа: <https://www.rbc.ru/crypto/news/6026a44c9a794760559dddce> (дата обращения: 30.04.2021)

анализе влияния запаздывающих настроений на реализованную волатильность. Их исследование показывает, что индикаторы настроений, такие как коэффициент объема торговли ОЕХ, мало влияют на будущую волатильность после учета запаздывающей доходности.

Mao et al. (2011) 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. Используя набор данных по конкретным компаниям и макроэкономическим новостям, Ho et al. (2013) обнаружили, что новостные настроения оказывают значительное влияние на внутридневную волатильность ряда отдельных акций США (особенно это оказалось актуальным для акций компаний, задействованных в сфере IT). Аналогичным образом Aouadi et al. (2013), Dimpfl and Jank (2015) и Hamid and Heiden (2015) показывают, что объем поисковых запросов можно использовать для прогнозирования волатильности фондового рынка. На основе оценок настроений, рассчитанных для финансовых статей, полученных с платформы NASDAQ, Zhang et al. (2016) обнаружили, что рост настроений влияет на волатильность.

На основе оценок настроений, рассчитанных для финансовых статей, полученных с платформы NASDAQ, Zhang et al. (2016) обнаружили, что рост настроений влияет на волатильность, а также на объем торгов. See-To and Yang (2017) и Siganos et al. (2017) показывают, что за сильным расхождением в настроениях инвесторов следует увеличение дневной волатильности.

В исследовании Caporin and Poli (2017) показано, что переменные, связанные с новостями, можно использовать для получения улучшенных прогнозов волатильности. Используя оценки настроений, рассчитанных на

основе данных Twitter, Берендт и Шмидт (2018) показывают, что настроения влияют на внутридневную волатильность, но улучшения прогнозов не являются экономически значимыми.

Все предыдущие исследования рассматривают только ограниченный набор переменных-предикторов, часто без учета экономических переменных, и/или используют только ограниченный объем данных. Более того, часто анализируемый период времени довольно короткий. Например, исследования Caporin and Poli (2017) и Mao et al. (2011) не учитывают классические экономические и финансовые переменные. Таким образом, исходя из результатов данных работ, нельзя достоверно утверждать, что настроения в социальных сетях и/или в новостях обладают дополнительной прогностической силой (если только не учитываются другие экономические переменные, такие как кредитные спреды или волатильность опционов). Другой особенностью всех вышеперечисленных работ является тот факт, что ни в одной из них не упоминаются российские компании. К тому же в работах мало представлены компании, занятые в сырьевой сфере (например, нефтегазовая отрасль). В данном исследовании будут проанализированы исключительно российские компании (в том числе и занятые в сырьевых сферах).

Далее оценим некоторые подходы и методы, позволяющие дать предварительную оценку общему «тонусу» инвесторов, разберем способы количественной интерпретации настроений и внимания.

## **Глава 1 Теоретические основы построения модели на основе данных социальных сетей**

### **1.1 Общие подходы к измерению внимания и настроения инвесторов**

Существует несколько способов оценить степень влияния настроения инвесторов на уровень волатильности фондового рынка.

- Показатели финансового рынка (объем, VIX, спред TED и т. д.). Сpread TED – это разница между трехмесячным казначейским векселем и трехмесячным LIBOR (Лондонская межбанковская ставка предложения)<sup>3</sup> в долларах США. Простыми словами, LIBOR – это некая эталонная процентная ставка предложения на рынке межбанковских кредитов.
- Индекс настроений на основе опросов (например, индекс потребительского доверия на основе реакции на ситуацию). Индекс вычисляется как относительный показатель количества положительных/отрицательных ответов на общее число ответов.
- Поведение при поиске в Интернете. Для анализа заинтересованности можно использовать различные каналы для сбора статистической информации в Интернете, например, WordStat – позволяет вычислить количество запросов (по словам) за определенные промежутки времени.
- Факторы неэкономического характера (новости, погода, состояние здоровья).
- Текстовых данных о тональности из онлайн-ресурсов (например, Twitter).

Есть исследование, где для оценки тональности текстов использовался метод «мешка слов» и пяти различных словарей, а также учитывалась инерционность влияния новостного потока на поведение участников рынка. Эконометрическая методология исследования состояла из двух этапов: отбор потенциально значимых объясняющих переменных на

---

<sup>3</sup> ЛИБОР // Лас-Тунас — Ломонос. — М. : Большая российская энциклопедия, 2010. — С. 391. — (Большая российская энциклопедия : [в 35 т.] / гл. ред. Ю. С. Осипов ; 2004—2017, т. 17). — ISBN 978-5-85270-350-7.

базе метода эластичной сети и оценка параметров модели ARMAX-GARCH. Тональность новостей, наряду с фундаментальными экономическими факторами, также оказывает систематическое влияние на курс рубля, которое, тем не менее, зависит от их тематики: наибольшее значение имеют новости бизнес-тематики, в то время как политические сообщения в СМИ не оказывают статистически значимого влияния на рыночный курс рубля [Ulyankin, 2020].

В современной экономической литературе строится довольно много различных индексов экономической активности. Часть из них основана на опросах экономических агентов («ручные» индексы), часть – на неструктурированных данных из интернета («автоматические» индексы). При этом вопрос о том, какие из подходов оказываются самыми успешными, остается открытым.

«Автоматические» индексы строятся с помощью методов машинного обучения. В качестве исходных данных используются поисковые запросы, новостные статьи и комментарии пользователей под новостными постами из социальных сетей. Анализ получившихся индексов экономической активности показывает, что существует причинность по Грейнджею между поисковыми и новостными индексами, с одной стороны, и «ручными» индексами – с другой, при этом поисковые и новостные индексы являются причинами по Грейнджею для «ручных» индексов. Кроме того, оказывается, что поисковые и новостные индексы экономической активности лучше ручных индексов объясняют и прогнозируют набор выбранных для исследования макроэкономических переменных. Хорошая объясняющая способность данных индексов позволяет использовать их для научкистинга в условиях наличия лага в выходе макроэкономической статистики.

Для своевременной оценки состояния экономики и заблаговременного выявления кризисных ситуаций используются системы опережающих индикаторов. Они помогают осуществлять мониторинг и

прогнозирование деловой активности, а также ускоряют принятие государственными органами необходимых важных решений макроэкономической политики. Многие такие индикаторы строятся вручную в ходе опросов экономических агентов. Этот подход позволяет строить индекс экономической активности довольно редко (раз в месяц или даже в квартал). Начиная с Choi and Varian (2009) многие авторы пишут о том, что построение подобных индикаторов можно автоматизировать за счет обработки современными статистическими методами больших массивов данных, доступных в интернете. Обычно для автоматизации используют один из трех источников данных:

- поисковые запросы (индекс поиска);
- поток новостей (индекс новостей);
- поток комментариев из социальных сетей (индекс настроений).

Как правило, подобный индикатор строят, отталкиваясь от вручную отобранных дескрипторов, а затем проверяют, обладает ли этот индекс объясняющей способностью, то есть смотрят, насколько он помогает объяснить какой-либо определенный макроэкономический показатель. При этом часто в статьях используется какой-то один из подходов к построению подобных индексов, из-за чего остается непонятным, какой из возможных индексов позволяет получить наиболее точный прогноз рассматриваемого макроэкономического показателя.

Поисковые системы собирают статистику о том, что именно интересует пользователей. Частично эта статистика находится в открытом доступе. В 2010 г. поисковая компания Google заявила, что с помощью информации из поисковых запросов о трейлерах выходящих фильмов она может с точностью 94% предсказывать кассовость ленты в первые дни проката (Goel et al., 2010). Кроме того, поисковик научился использовать поисковые запросы, чтобы предсказывать распространение эпидемии гриппа (Ginsberg et al., 2009) и вспышек коронавируса (Li et al., 2020). Данные по поисковым запросам могут оказаться полезными для

конструирования индикаторов финансовой конъюнктуры. Так, британские исследователи McLaren and Shanbhogue (2011) показали, что между числом запросов по дескрипторам, связанным с безработицей, и фактической безработицей существует значимая связь. Авторы оценивали авторегрессию второго порядка в разностях, в которую в качестве объясняющей переменной добавляли индекс поиска. Запросы, связанные с занятостью, включались в модель без запаздывания, поскольку поиск работы осуществляется в большинстве случаев уже безработными. В результате оценивания в уравнении для безработицы был получен значимый коэффициент при переменной индекса поиска. Кроме того, модель с индексом показывала более низкие значения информационных критериев. Аналогичный индекс поиска McLaren and Shanbhogue (2011) выявляет и для цен на недвижимость (также оказался значим). В этом случае поисковые запросы включаются в модель с запаздыванием, так как после того, как человек нашел квартиру, ему нужно некоторое время на проведение переговоров и заключение сделки по ее покупке.

Похожие результаты были ранее получены D'Amuri (2009) на данных Италии и Choi and Varian (2009) – США. В работе Choi and Varian с использованием индексов поиска оцениваются сезонные AR-модели для таких показателей, как продажи автомобилей, заявки на пособие по безработице, дальность поездок при авиаперелетах, что в ряде случаев значительно повышает точность моделей. Информационные критерии указывают на предпочтительность расширенных спецификаций, включающих индексы поиска как дополнительные объясняющие переменные. Также расширенные спецификации обладают более сильными прогнозными способностями, измеренными метрикой RMSE (Root Mean Square Error – среднеквадратичная ошибка). Поисковые индексы в этой работе интерпретируются как желание найти работу, купить дом и т. п. В работе Столбова (2011) была предпринята попытка на основе поисковых запросов сконструировать индикатор, способный улучшить прогноз

изменения российского фондового индекса. Автор выяснил, какие термины пользователи наиболее интенсивно «гуглили» в России по теме финансовых рынков и страхования в самый разгар глобального финансового кризиса (в период с сентября 2008 г. по июнь 2009 г.). Отобранные дескрипторы были соотнесены с некоторыми экономическими и финансовыми терминами. Поисковые запросы по ним скачивались из Google Trends на месячной основе и включались с определенными весами в итоговый индекс. В качестве альтернативы рассматривался метод главных компонент (Principal Component Analysis, PCA), который позволял выделить в колебании индекса первую главную компоненту. В конечном счете, была установлена значимая связь между построенным индексом поиска и динамикой индекса РТС. При этом, как и в работах Choi and Varian, информационные критерии и RMSE указывали на предпочтительность расширенной спецификации модели.

Таблица 1

Пример конструкций и слов с негативной семантикой: отрицатели, усилители, уменьшители

№	Модели конструкций	Примеры
Отрицатели		
1	Не + глагол с положительной семантикой	Не писать, не заниматься, не читал
2	Не + глагол с негативной семантикой	Не воровать, не врать
3	Не + модальный глагол + глагол	Не мог выполнить (обещанное), не сумел объяснить
Усилители		
4	Слишком + слово с негативной семантикой	Слишком неуверенный
5	Очень + слово с негативной семантикой	Очень плохой
6	Крайне + слово с негативной семантикой	Крайне неприятный

Уменьшители		
7	Не + очень + слово	Не очень хороший
8	Почти + не + глагол	Почти не знает
9	Практически + не + глагол	Практически не соответствует

Источник: <https://na-journal.ru/3-2019-informacionnye-tehnologii/1916-opredelenie-tonalnosti-i-obektivnosti-novostnyh-tekstov-slovarnym-podhodom>

При анализе тональности новостных сообщений необходимо учитывать некоторые нюансы русского языка (см. табл.1). Алгоритм определения полярности новостного текста состоит из следующих этапов:

- 1) Поиск и выделение по всему тексту отрицателей, усилителей и уменьшителей тональности и кодируем парные слова, к примеру, такие как «отнюдь не» или слишком неграмотный» как 1 слово.
- 2) Проводим процедуру распознавания частей речи и лемматизации.
- 3) Проводим выделение в тексте позитивных и негативных слов.

Пример:

Погода была плохой (без частицы отрицателя тональность предложения (-1)).

Погода была не плохой (с частицей отрицателем, тональность предложения (+1)).

Погода была весьма неплохой (с увеличителем очень итоговая тональность (+1.5)).

- 4) Вычисляем полярность предложения и нормализованную сумму текста в промежутке [-1:1].

Схожую методику можно использовать для оценки настроения инвесторов.

## **1.2 Социальные сети как источник измерения настроения инвесторов, причины выбора Twitter для проведения исследования**

Социальная сеть - платформа, предназначенная для построения социальных взаимоотношений в Интернете.<sup>4</sup> В первую очередь, это сайты, разработанные с целью познакомить и собрать людей с общими интересами, дать им возможность общаться на различные темы, выкладывать и обсуждать фото и т.д. Первые социальные сети появились на западе. Самые популярные из них - это Facebook, Twitter, MySpace, Badoo, YouTube, Google+ и др.

Вскоре и в России стали появляться аналоги: ВКонтакте, Одноклассники.ru, Мой Мир@mail.ru, Гайдпарк, В кругу друзей, Привет.ru, Мой Круг и т. д. Социальная сеть ВКонтакте.ру — это сайт №1 в России, как по посещаемости, так и по популярности. В данном случае популярность определялась по количеству запросов в Яндексе (использовалось приложение WordStat). С нынешней ситуацией популярности социальных сетей можно ознакомиться в табл. 2. В настоящее время группы людей, объединенных в сеть, могут не только подключаться в реальном времени, но и иметь под рукой возможность создавать и получать доступ к большему объему информации. Благодаря быстрому мобильному оборудованию и простым в использовании технологиям социальных сетей (например, такие как Facebook и Twitter), большие группы населения могут организовывать общественные мероприятия и встречи, а также создавать или реагировать на изменения в мире.

Таблица 2

Количество упоминаний социальных сетей в поисковой системе «Яндекс» (были выбраны максимальные значения)

---

<sup>4</sup> Академический онлайн-словарь. Определение «социальной сети». [Электронный ресурс]. Режим доступа: <https://dic.academic.ru/dic.nsf/ruwiki/60759> (дата обращения: 04.05.2021)

Социальная сеть/мессенджер	Название	Количество упоминаний	Страна происхождения	Популярность сайта в мире (рейтинг) – по объему трафика. Разработано аналитическим ресурсом Ahrefs (по данным 2020 года). Источник: <a href="https://ahrefs.com/blog/most-visited-websites/">https://ahrefs.com/blog/most-visited-websites/</a>
Социальная сеть	VКонтакте (VK)	19 534 435 + 5 904 127	Россия	14
Социальная сеть	Одноклассники	10 701 921 (поиск осуществлялся по связке слов: «одноклассник и + моя страница»)	Россия	27
Мессенджер	Telegram	1 013 335	Россия	Не включен
Социальная сеть	Twitter	792 489	США	4
Вideoхостинг	TikTok	792 489	Китай	Не включен
Социальная сеть	Instagram	1 818 920	США	6
Вideoхостинг	YouTube	8 474 696	США	2
Мессенджер	WhatsApp	2 474 921	США	26
Мессенджер	Skype	237 293	Эстония	Не включен
Социальная сеть	Facebook	1 207 438	США	3
Мессенджер	zoom.us	1 068 713	США	21

Источник: построено автором на основе аналитического портала Wordstat Yandex.

Исследуемый период: 04.04.2021 - 04.05.2021. [Электронный ресурс]. Режим доступа: <https://wordstat.yandex.ru/>

Telegram, по большей части, подходит под определение «мессенджер». Напомним, что мессенджер – система мгновенного обмена сообщениями (IM, Instant messaging). Мессенджеры позволяют, не отвлекаясь на «инородные» приложения, акцентировать свое внимание только на сообщения/процесс коммуникации. Тем не менее, в последние годы (начиная примерно с 2017 года) в Telegram начали появляться сообщества, где представлены новости и различные события. Однако, в

данном мессенджере нет возможности быстрой коммуникации большого числа людей, объем представленной «важной» информации мал, поэтому данный канал не очень подходит для исследования.

Twitter - популярный сервис микроблогов в реальном времени, который позволяет пользователям обмениваться короткой информацией, известной как твиты, длина которых ограничена 140 символами<sup>5</sup>. Пользователи пишут твиты, чтобы выразить свое мнение по различным темам, касающимся их повседневной жизни. Twitter - идеальная платформа для получения общественного мнения по конкретным вопросам, так информация представлена в удобной, легкоредактируемой для анализа форме.<sup>6</sup> Коллекция твитов используется в качестве основного корпуса для анализа настроений, который относится к использованию анализа мнений или обработки естественного языка<sup>7</sup>.

Twitter, с 500 миллионами пользователей и миллионами сообщений в день, быстро стал ценным активом для организаций, позволяющим укреплять свою репутацию.<sup>8</sup> Данные Twitter становятся все более важным источником для описания финансовой динамики.<sup>9</sup> Twitter обеспечивает

---

<sup>5</sup> M.Rambocas, and J. Gama, “MarketingResearch:TheRoleof SentimentAnalysis”. The 5th SNA-KDD Workshop’11. Universityof Porto, 2013

<sup>6</sup> A. K. Jose, N. Bhatia, and S. Krishna, “Twitter Sentiment Analysis”. National Institute of Technology Calicut, 2010

<sup>7</sup> А. Пак и П. Пароубек, «Твиттер как корпус для анализа настроений и сбора мнений», специальный выпуск Международного журнала компьютерных приложений, Франция: Universitede Paris-Sud, 2010

<sup>8</sup> С. Лохманн, М. Берч, Х. Шмаудер и Д. Вайскопф, «Визуальный анализ содержимого микроблогов с использованием изменяющегося во времени выделения совпадений в облаках тегов», Ежегодная конференция VISVISUS. Германия: Штутгартский университет, 2012.

<sup>9</sup> А. Агарвал, Б. Се, И. Вовша, О. Рамбоу и Р. Пасонно, «Анализ настроений в данных Twitter», Ежегодные международные конференции. Нью-Йорк: Колумбийский университет, 2012 г.

поток детализированной информации. По сути, это канал в режиме реального времени, который включает не только основные новости, но и второстепенные события, которые при правильном моделировании могут предоставить предварительную информацию о рынке еще до появления основных новостных лент.

Недавние события также отразили важную роль социальных сетей. Одним из основных примеров является отчет Комиссии по ценным бумагам и биржам США. Он позволяет компаниям использовать Twitter для объявления ключевой информации в соответствии с Положением о добросовестном раскрытии информации.<sup>10</sup>

В 2013 году Hash Crash взломал аккаунт Twitter американского информационного агентства Associated Press, ложно раскрыл сообщение об атаке на Белый дом. В результате таких противоправных действий промышленный индекс Доу-Джонса упал на 145 пунктов за минуты. Данный факт может отражать и шаткость моделей, основанных на данных социальных сетей. Влияние внешних мошеннических действий/операций/манипуляций значительно сказывается на качестве прогнозов.

Следовательно, изучение хаотических и, казалось бы, непредсказуемых действий социальных сетей становится все более важной исследовательской задачей для организаций, ориентированных на будущее. От инвестиционных банков до военных аналитических центров организации все чаще изучают инновационные методологии, чтобы внимательно прислушиваться к ежедневному потоку пользовательского контента. Это делается, чтобы лучше понять поведение потребителей, предсказать и подготовиться к будущим тенденциям в населении и

---

<sup>10</sup> Х. Саиф, Я. Хе и Х. Алани, «SemanticSentimentAnalysis of Twitter», материалы семинара по извлечению информации и анализу сущностей в данных социальных сетей. Соединенное Королевство: Институт СМИ, 2011

принимать более обоснованные решения в области социально-экономической политики.

Используя большое количество пользовательских сообщений в Твиттере и соответствующие исследования из быстро развивающейся области поведенческой экономики и информационных технологий, исследователи пытаются внести свой вклад, улучшить качество моделей как с точки зрения качества, так и своевременности оценки и прогнозов ряда макроэкономических индикаторов и индексов финансового рынка. Все для того, чтобы создать лучшие возможности для принятия решений.

Платформа микроблогов, такая как Twitter, похожа на обычную платформу для ведения блогов, за исключением того факта, что сообщения короче. Twitter имеет ограниченное количество слов, которые предназначены для быстрой передачи информации или обмена мнениями<sup>11</sup>. Малый бизнес и даже некоторые крупные организации начинают использовать потенциал микроблогов как маркетинговый инструмент электронной коммерции. Платформа микроблогов была разработана несколько лет назад для продвижения внешнеторговых веб-сайтов с использованием зарубежной платформы микроблогов в качестве маркетинга. Платформа микроблогов позволила компаниям создавать имидж бренда, увеличивать продажи продуктов, разговаривать с потребителями для хорошего взаимодействия с ними и др.<sup>12</sup>

Зачастую информация, представленная в одной социальной сети, дублируется другими источниками. В этой связи достаточно рассмотреть одну крупную социальную сеть. Остальные источники будут коррелированы с ней. Среди представленных выше социальных сетей,

---

<sup>11</sup> Д. Байд, С. Голдер и Г. Лотан, «Твит, твит, ретвит: разговорные аспекты ретвита в твиттере», Системные науки (HICSS), 2010 г.

<sup>12</sup> Дж. Чжан, Ю. Ку, Дж. Коди и Ю. Ву, «Пример использования микроблогов на предприятии: использование, ценность и связанные вопросы», Материалы семинара по Web 2.0, 2010.

наибольшую репрезентативную и информационную силу имеет сеть Twitter. В главе 3 именно она будет использоваться для проведения расчетов.

### 1.3 Модели тональности текстов

Для начала определим теоретические основы идеи анализа тональности новостей. Анализ тональности — класс методов контент-анализа в компьютерной лингвистике, основная задача которого заключается в классификации текста по его настроению.<sup>13</sup> Целью анализа тональности является нахождение мнений в тексте и определение их свойств. В зависимости от поставленной задачи нас могут интересовать разные свойства, например:

1. автор — кому принадлежит это мнение;
2. тема — сфера, в которой задействована новость;
3. тональность — позиция автора относительно упомянутой темы

(обычно «положительная» или «отрицательная»). В современных работах можно встретить и «нейтральную» коннотацию.

Пример подобной оценки:

«Некоторые автопроизводители начинают более серьезно думать о цепочках поставок аккумуляторов. Tesla подписала соглашения о закупках с Glencore, которая добывает кобальт в Демократической Республике Конго. Илон Маск даже предложил добывать литий в Неваде с использованием новых технологий. Однако этот план встретил скептицизм в самой горнодобывающей промышленности; в более широком смысле ситуация выглядит достаточно несостоятельной, чтобы вмешаться политики в Америке и Европе. Начинается возвращение в пользу более интервенционистской политики в деятельности, которая могла бы

---

<sup>13</sup> Анализ тональности в русскоязычных текстах. [Электронный ресурс]. Режим доступа: <https://habr.com/ru/company/mailru/blog/516214/> (дата обращения: 30.03.2021)

рассматриваться как чисто коммерческая», - утверждает Родерик Эггерт из Американского института критических материалов»<sup>14</sup>.

- Автор: Родерик Эггерт;
- Тема: "инновационный менеджмент, логистика";
- Тональность: "негативная"

Методы анализа тональности новости могут носить как контролируемый (с учителем), так и неконтролируемый (без учителя) характер. Анализ настроений неконтролируемыми методами имеет следующие свойства:

- анализ тональности основан на основе словаря, который отображает заранее заданные списки положительных и отрицательных слов в твитах. Напомним, что в России существует так называемый «Словарь тональности русского языка». Далее можно будет ознакомиться с примерами с этого источника;
- окончательная оценкадается функцией положительного и отрицательного счета;
- с помощью векторизации текста представляются документы в векторном пространстве, создавая сопоставление терминов с идентификаторами терминов;
- можно применить несколько подходов, например, подход Bag of Words (BoW): здесь текст представлен как неупорядоченный набор слов, в котором учитывается каждое слово из твитов. В методе «Мешка слов» каждое слово имеет определенное положение в некотором числовом ряду. Если слово в тексте встречается n раз, то в данном ряду тому месту, где это слово зашифровано, будет присвоено

---

<sup>14</sup> Governments have identified commodities essential to economic and military security. [Электронный ресурс]. Режим доступа: <https://www.economist.com/finance-and-economics/2021/03/31/governments-have-identified-commodities-essential-to-economic-and-military-security> (дата обращения: 03.04.2021)

значение  $n$ . Если слово не встречается, то будет присвоен 0. Таким образом, можно получить строку, которая считает количество слов в предложении.

Существует несколько подходов по классификации тональности новостей. Компьютеры могут выполнять автоматический анализ цифровых текстов, используя элементы машинного обучения, такие как скрытый семантический анализ, метод опорных векторов, «мешок слов» и семантическая направленность в этой области. Анализ тональности может быть разделен на две отдельные категории:

- ручной (или анализ тональности экспертами);
- автоматизированный анализ тональности.

Различия между ними лежат в эффективности системы и точности анализа. В данной работе будут проанализированы 2 основных вида/способа анализа тональностей новостей.

### **1.3.1 Общий подход к измерению тональностей текста и особенностей языка**

#### **WordNet-Affect**

WordNet-Affect - это расширение доменов WordNet, включающее подмножество наборов синонимов, подходящих для представления аффективных понятий, связанных с аффективными словами. Подобно методу для меток домена, присвоим ряду наборов синхронизации WordNet одну или несколько аффективных меток (метки  $a$ ). В WordNet-Affect используются дополнительные эмоциональные метки для того, чтобы разделять синсеты в соответствии с их эмоциональной валентностью. Для этого определяются четыре эмоциональные метки: позитивная, негативная, неоднозначная и нейтральная.

Как уже отмечалось выше, каждое слово из «словаря тональности» может обладать разным эмоциональным окрасом. Но помимо эмоции,

сообщения могут обладать различными настроениями, чувствами, быть в разных в когнитивных, физических состояниях и т.д. Несмотря на то, что термины очень близки по смыслу, тем не менее, они имеют различительные особенности:

- 1) Чувства и эмоции. Согласно «Большому словарю психологических терминов» под эмоциями понимается субъективное состояние человека (и животных), вызванных реакцией на внешние и внутренние раздражители. Зачастую главным отличием эмоции от чувств признается их продолжительность: чувства более продолжительные, чем эмоции. Чувства определяются долгосрочной или среднесрочной оценкой, эмоции – краткосрочной.
- 2) Настроение - это эмоциональное состояние, которое влияет на поведение человека в определенный промежуток времени. Оно определяет нашу активность, умение взаимодействовать с миром, общий тонус.
- 3) Особенность. В данном случае подразумеваются некоторые выделяющиеся черты того или иного слова, способные повлиять на контекстуальное значение предложения и/или текста.
- 4) Когнитивное состояние.
- 5) Физическое состояние, выраженное самочувствием субъекта.
- 6) Гедонический сигнал.
- 7) Ситуации, вызывающие эмоции.
- 8) Эмоциональные отклики. Сторонняя реакция на поведение субъекта.
- 9) Поступки. Реальные действия субъекты, предпринятые под воздействием чувств и/или эмоций.

10) Отношение, позиция. Устойчивые взгляды на то или иное состояние окружающей среды, выраженное в придании объекту или субъекту некоторых оценочных качеств.<sup>15</sup>

Скомпилированная информация представлена в табл.3. Одна и та же единица языка может носить разную эмоциональную метку.

Таблица 3

Возможный эмоциональный окрас сообщений, полученных с социальных сетей

Эмоциональная метка	Пример
Эмоция (emotion)	сущ. раздражение, гл. бояться (fear)
Настроение (mood)	сущ. неприязнь (animosity), прил. любезный (amiable)
Особенность (trait)	сущ. агрессивность (aggressiveness), прил. конкурирующий (competitive)
Когнитивное состояние (cognitive state)	сущ. замешательство (confusion), прил. потрясенный (dazed)
Физическое состояние (physical state)	сущ. болезнь (illness), прил. выдохнувшийся (all in)
Гедонический сигнал (hedonic signal)	сущ. боль (hurt), сущ. мучение (suffering)
Ситуации, вызывающие эмоции (emotion-eliciting situation)	сущ. неловкость (awkwardness), сущ. безопасность (out of danger)
Эмоциональные отклики (emotional responses)	сущ. холодный пот (cold sweat), гл. дрожать (tremble)
Поступки (behaviour)	сущ. преступление (offense), прил. заторможенный (inhibited)
Отношение, позиция (attitude)	сущ. нетерпимость (intolerance), сущ. оборонительная позиция (defensive)
Чувство (sensation)	сущ. холод (coldness), гл. чувствовать (feel)

Источник: построено автором на основе материалов Д.Усталова «Анализ тональности текста на русском языке при помощи графовых моделей» (2012)

---

<sup>15</sup> Stefano Baccianella. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining // Proceedings of LREC : конференция. — 2010. — Р. 2200–2204.

## **SentiWordNet**

SentiWordNet — это лексический семантический тезаурус. Данная система является результатом процесса автоматического аннотирования каждого WordNet синсета в соответствии с его степенью позитивности, негативности и объективности. Каждому синонимическому ряду из WordNet присваивается три численных оценки, где каждая из этих оценок соответственно определяет объективную, позитивную или негативную составляющую синсета. Каждая из этих оценок принимает значения в интервале от 0 до 1, и в сумме они дают 1, то есть каждая из этих оценок может иметь ненулевое значение.

Процедура оценки немного громоздкая, поэтому порой бывает выгоднее использовать методики более низкого уровня, с более простыми алгоритмами.

## **SenticNet**

Главным назначением SenticNet является упрощение процедуры машинного распознавания концептуальной и эмоциональной информации, передаваемой с помощью естественного языка. Главным различием от других приложений будет то, что SentiWordNet и WordNet-Affect обеспечивают связывание слов и эмоциональных понятий на синтаксическом уровне, не позволяя выявлять смысловую составляющую. То есть, данный метод позволяет заглянуть «вглубь» языка, исследовать смысловые значения слов (в отличие от «мешка слов», где каждому элементу присваивается число, здесь анализируется само слово).

Таким образом, данная методика выходит на еще один уровень вверх. Теперь анализируются не только отдельные слова, их количественные и смысловые значения, а целые тексты, даже улавливается контекстуальное значение слова на основе смысловой нагрузки текста.

Для того чтобы представить данные SenticNet в машиночитаемом виде, то есть удобном для обработки компьютерными программами, данные

кодируются в RDF-триплеты (Resource Description Framework или «среда описания ресурса») с использованием синтаксиса XML.

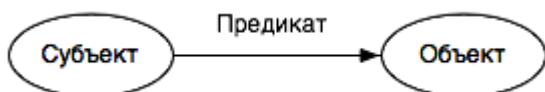


Рис.1. Триплет RDF. Источник: построено автором на основе материалов John Hebeler, Matthew Fisher, Ryan Blace, Andrew Perez-Lopez. Semantic Web Programming. — John Wiley & Sons, 2009. — 648 с. — ISBN 9780470418017.

Устройство триплета представлено на рис.1. Смысл заключается в том, что есть некоторая семантическая паутина, которая объединяет субъект с объектом посредством предиката (мост, «смысловой соединитель»). Единство этих трех компонентов и называется «триплетом». Например, выражение «трава зеленого цвета» можно перевести в формат RDF – триплета в следующем виде: «субъект» - трава; «объект» - зеленый; «предикат» - имеет цвет.<sup>16</sup>

### 1.3.2 Методы обработки естественного языка (Natural Language Processing, NLP)

Обработка естественного языка состоит из понимания лингвистики и методов машинного обучения. Для анализа текста необходимо понимание морфологии – форма слов, синтаксиса – связь между словами, семантики – значение слов, прагматики – условие употребления слов. Опираясь на знания вышеперечисленной структуры, можно переходить к пониманию смысла текста. В результате такого анализа текста (Natural Language Understanding, NLU) можно переходить к генерации текста (Natural

---

<sup>16</sup> Bo Pang, Lillian Lee. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts (англ.) // Proceedings of the Association for Computational Linguistics (ACL) : журнал. — 2004. — Р. 271–278.

Language Generation, NLG). Таким образом, процессы NLU и NLG в совокупности представляют NLP. Например, таким образом работают чат-боты: они принимают на вход вопрос, после этого искусственный интеллект распознает смысл вопроса, и на выходе чат-бот генерирует ответ на вопрос. Далее проанализируем машинный процесс обработки текстов.

### **Предобработка текста**

Для анализа естественного языка необходима предобработка текста. Данный этап в свою очередь можно разбить на ряд подопераций:

1. Токенизация – разбиение текста на токены. Токеном может быть слово, пунктуация, пробел и т. д. Таким образом, можно осуществлять токенизацию по словам, предложениям и др.. Здесь и далее под токенизацией будем иметь в виду токенизацию по словам.

После токенизации все встречающиеся слова в тексте образуют словарь. Возникает сложность в том, что такой словарь может достигать больших размеров. Для уменьшения размера словаря используются следующие шаги - нормализация слова и удаление слов.

2. Нормализация слова. У каждого слова существует множество форм – это разные падежи, единственное и множественное число, время глагола и т. д. Такие формы нужно приводить к одной форме для того, чтобы уменьшить размер словаря. Это можно осуществить двумя подходами: стемминг и лемматизация.

- Стемминг. Процесс удаления последних символов в словах для того, чтобы похожие по смыслу слова пришли в единую форму.

Например:

change	
changing	
changes	→
changed	chang
changer	

- Лемматизация. Приведение формы к начальной форме. Например, для существительного – это именительный падеж, единственное число.

change  
changing  
changes → change  
changed  
changer

### 3. Удаление слов:

- Удаление стоп-слов – предлогов, союзов, частиц и прочее. Такие слова являются неинформативными, поэтому нет смысла оставлять их для анализа.
- Удаление неинформативных слов или шаблонов. Например, при анализе писем такими словами будут являться слова «Кому», «От кого», «Тема» и т. д.<sup>17</sup>

## Примеры задач NLP

Методы обработки естественного языка применяются для решения задач различного характера.

1. Классификация текста: жанр, автор, страна и пр. Например, по комментарию к фильму можно классифицировать, каким является комментарий - положительным или отрицательным.
2. Исправление опечаток. Например, признаком того, является ли ошибка опечаткой, является то, на сколько близко находятся рядом корректная буква и буква, нажатая случайно.
3. Ранжирование поисковой выдачи.
4. Генерация текста, например, генерация стиха.

---

<sup>17</sup> Benjamin Snyder, Regina Barzilay. Multiple Aspect Ranking using the Good Grief Algorithm (англ.) // Proceedings of the Joint Human Language Technology/North American Chapter of the ACL Conference (HLT-NAACL) : конференция. — 2007. — P. 300–307.

5. Машинный перевод, например, Яндекс переводчик или Google Translate.
6. Диалоговые системы – Siri, Алиса и др..
7. Суммаризация текста – помогает сгенерировать краткое содержание текста.

## **Компьютерное представление слов**

Можно выделить три основных классических способа представления слов – One-hot encoding, TF-IDF и Embedding.

### **One-hot encoding**

Является самым простым способом кодирования слова. Каждое слово в словаре представляет из себя one-hot – вектор размерности длины словаря. Например, словарь имеет размерность 1000 слов. Тогда каждое  $i$ -тое слово будет представляться как вектор длины 1000,  $i$ -тая координата которого равняется единице, остальные координаты равняются нулю.

Данный способ имеет ряд существенных недостатков. Во-первых, векторы не отображают значения слов. Во-вторых, невозможно установить меру близости между словами. В-третьих, вектора имеют большую размерность, при этом содержат в себе минимальное количество информации.

### **TF-IDF (Term Frequency – Inverse Document Frequency)**

Данный подход по сравнению с предыдущим методом способен отображать важность слова для документа. Принцип работы метода следующий — если слово встречается в каком-либо документе часто, при этом встречаясь редко во всех остальных документах — это слово имеет большую значимость для того самого документа. TF — это частотность термина, которая измеряет, насколько часто термин встречается в

документе. IDF — это обратная частотность документов. Она измеряет непосредственно важность термина. Введем следующие обозначения:

$n_{dw}$  — число вхождений слова  $w$  в документ  $d$ ;

$N_w$  — число документов, содержащих слово  $w$ ;

$N$  — общее число документов;

$p(w, d) = N_w/N$  — вероятность выбрать документ, содержащий  $w$ ;

Вероятность встретить  $w$  в каждом случайно выбранном документе из  $n_{dw}$  штук:

$$p(w, d, n_{dw}) = (N_w/N)^{n_{dw}} \quad (1)$$

Таким образом, чем больше  $p(w, d, n_{dw})$  в формуле (1), тем менее значимо слово для данного документа. Введем функцию (2) от этой функции так, чтобы чем большее значение принимал функционал, тем более значимым было слово для документа:

$$-\log p(w, d, n_{dw}) = n_{dw} \cdot \log(N/N_w) = TF(w, d) \cdot IDF(w) \quad (2)$$

В целом множитель  $TF = n_{dw}$  сам по себе показывает то, на сколько слово значимо для документа (как было описано в методе Bag-of-Words). Однако, поправка  $IDF$  делает метрику более совершенной. Например, слово «the» встречается очень часто, тогда значение  $TF$  будет большим и говорит о значимости слова для документа, однако, это не так. Так как это слово встречается почти в любом тексте,  $N/N_w \approx 1$ , тогда  $IDF$  будет близким к нулю. И произведение множителей тоже будет близким к нулю, следовательно, слово «the» не значимо для документа.

Далее рассмотрим простой пример: пусть есть 2 документа, содержащих термины (табл. 4). Проанализируем важность термина «this» для обоих документов:

$$TF(\text{«this»}, d_1) = 0.2;$$

$$TF(\text{«this»}, d_2) \approx 0.14.$$

Таблица 4

Пример двух документов, входящих в них терминов и кол-ва вхождений терминов

<b>Документ 1, <math>d_1</math></b>		<b>Документ 2, <math>d_2</math></b>	
<b>Термин</b>	<b>Кол-во вхождений</b>	<b>Термин</b>	<b>Кол-во вхождений</b>
this	1	this	1
is	1	is	1
a	2	another	2
sample	1	example	3

Источник: построено автором на основе информации с сайта Habr. [Электронный ресурс]. Режим доступа: <https://habr.com/ru/post/542048/> (дата обращения: 03.05.2021)

Получаем, что для первого документа слово «this» более значимо, чем для второго. Однако, само слово по себе неинформативно для обоих документов. В этом случае срабатывает поправка  $IDF(\text{«this»}) = 0$  и говорит о том, что слово встречается во всех документах, а значит, не является специфическим ни для первого, ни для второго документа. Суммарно получаем, что  $TF \cdot IDF = 0$  для обоих документов.

В этом подходе документ будет представлять из себя вектор, координаты которого равны  $TF \cdot IDF$  для каждого слова в документе.

## Коллокация

TF-IDF можно использовать не только для слов, но и для N-грамм и коллокаций.

N-грамма – это последовательность из N слов, идущих подряд. Коллокация – это сочетание слов, не обязательно идущих подряд. Например, в предложении “This is a sentence” у нас есть следующие униграммы, биграммы и триграммы (см. табл. 5).

Таблица 5

## Пример униграмммы, биграммы, триграммы

униграммы:	this is a sentence
биграммы:	this is is a a sentence
триграмммы:	this is a is a sentence

Источник: построено автором на основе данных сайта на основе информации с сайта Habr. [Электронный ресурс]. Режим доступа: <https://habr.com/ru/post/332078/> (дата обращения: 03.05.2021)

### PMI (Pointwise Mutual Information)

Для словосочетаний слов (возьмем словосочетания из двух слов) вместе TF-IDF может быть использована метрика PMI, которая измеряет то, на сколько слова встретились вместе не случайно. В подсчете данной метрике используется скользящее окно фиксированной длины (в нашем случае длина равна двум, рис. 2), метрика вычисляется формулой (3).

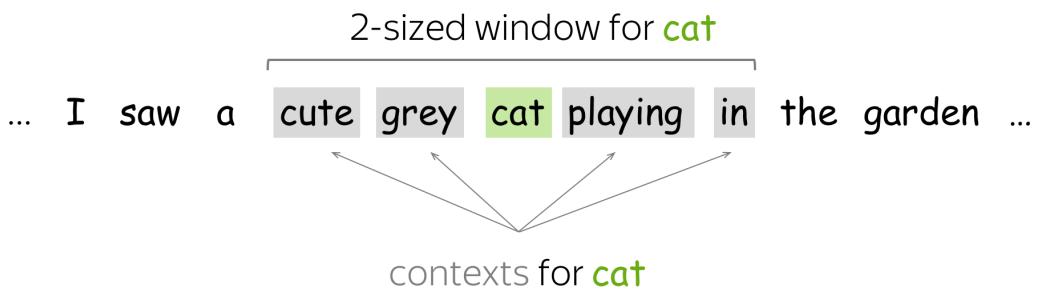


Рис. 2. Скользящее окно размера два. Слово «cat» является центральным, два слова слева и два слова справа являются контекстными. Источник: [https://lenavoita.github.io/nlp\\_course/word\\_embeddings.html](https://lenavoita.github.io/nlp_course/word_embeddings.html)

$p(u, v)$  – совместная вероятность встретить слова в тексте;

$p(u), p(v)$  – отдельная вероятность встретить слово  $u$  или  $v$  в тексте;

$n_{uv}$  – количество раз, когда слова  $u$  и  $v$  встретились вместе;

$$PMI = \log \frac{p(u, v)}{p(u)p(v)} = \log \frac{n_{uv} n}{n_u n_v} \quad (3)$$

От метрики  $PMI$  можно также взять линейный выпрямитель (Rectified Linear Unit, ReLu) для того, чтобы значения метрики были всегда неотрицательными, тогда формула (3) сводится к формуле (4).<sup>18</sup>

$$pPMI = \max(0, PMI) \quad (4)$$

## Embedding

Эмбеддинг – это закодированное представление слова, которое отражает контекст, в котором это слово используется. Рассмотрим один из способов эмбеддинга. Пусть:

$v(word_i)[j]$  – количество нахождений слов  $i$  и  $j$  рядом в датасете.

Тогда, например, возьмем пару слов «horse» и «car». Для них представление в векторной форме будет таким, что каждая координата вектора равна количеству нахождений слова рядом с другими словами из словаря:

$$\begin{aligned} v(word_1) &= [12, 1, 0, 10, 5, \dots] \\ &\text{horse} \quad \text{ride} \quad \text{whee} \quad \text{roof} \quad \text{hair} \quad \text{breed} \\ v(word_2) &= [20, 10, 0, 0, 1, \dots] \\ &\text{car} \quad \text{ride} \quad \text{whee} \quad \text{roof} \quad \text{hair} \quad \text{breed} \end{aligned}$$

Таким образом, такое представление передает значение и смысл слова. Однако, вектора будут иметь размерность размерности словаря минус один (исключаем  $v(word_i)[i]$ ). Для оптимизации размера векторов можно использовать методы понижения размерности, такие как метод главных компонент (PCA) или сингулярное разложение (SVD).

---

<sup>18</sup> Peter Turney. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews (англ.) // Proceedings of the Association for Computational Linguistics. — 2002. — P. 417–424. — arXiv:cs.LG/0212032.

В таком подходе существует ряд проблем. Во-первых, метод плохо описывает редкие слова, норма таких векторов значительно меньше других. Наоборот, слишком часто встречающиеся слова будут иметь большую норму весов. Во-вторых, метод требует больших вычислительных ресурсов на вычисление векторов и на уменьшение размерности этих векторов). В-третьих, при добавлении в датасет новых слов необходимо пересчитывать все вектора.

### Сингулярное разложение (SVD)

Далее рассмотрим, как работает сингулярное разложение (SVD) для данной задачи (рис. 3). Составим квадратную матрицу ( $N \times N$ ), каждая строка которой является представлением слова. Тогда столбцы такой матрицы представляют собой контекст слова, то есть информацию о том, с какими другими словами чаще употребляется наше слово. Тогда матрицу слева можем представить как произведение трех матриц размерами ( $N \times R$ ), ( $R \times R$ ), ( $R \times N$ ), где  $R$  мы задаем сами. Тогда строки матрицы  $V_d$  являются представлениями слов, когда слова являются центральными, а столбцы матрицы  $U_d^T$  являются представлениями слов, когда слова являются контекстными.

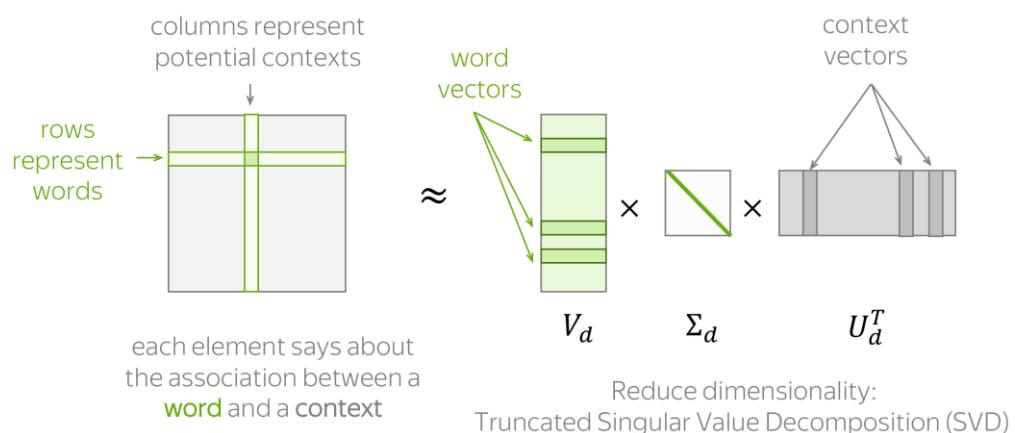


Рис. 3. Сингулярное разложение (SVD).

Источник: [https://lena-voita.github.io/nlp\\_course/word\\_embeddings.html](https://lena-voita.github.io/nlp_course/word_embeddings.html)

Для сингулярного разложения вместо метрики  $v(word_i)[j]$  может быть использована ранее рассмотренная метрика  $PMI$ , формула (3).

Также сингулярное разложение может быть использовано для подхода  $TF - IDF$ , тогда документ – будет являться центральным словом, а весь остальной текст – контекстом.

### Задача классификации текста

Признаками для задачи классификации будут являться представления слов: Bag-of-Words, TF-IDF, Embeddings. Для формирования признакового пространства часто используют инструменты дистрибутивной семантики, такие как Word2Vec, Glove, FastText и т. д.. Например, на входе мы имеем какое-либо эмбеддинги слов, далее эти эмбеддинги каким-либо образом аккумулируются (суммируются, усредняются и т. д.) на эмбеддинговом слое, результат передается в какую-либо нейронную сеть, и на выходе нейронной сети получаем класс.

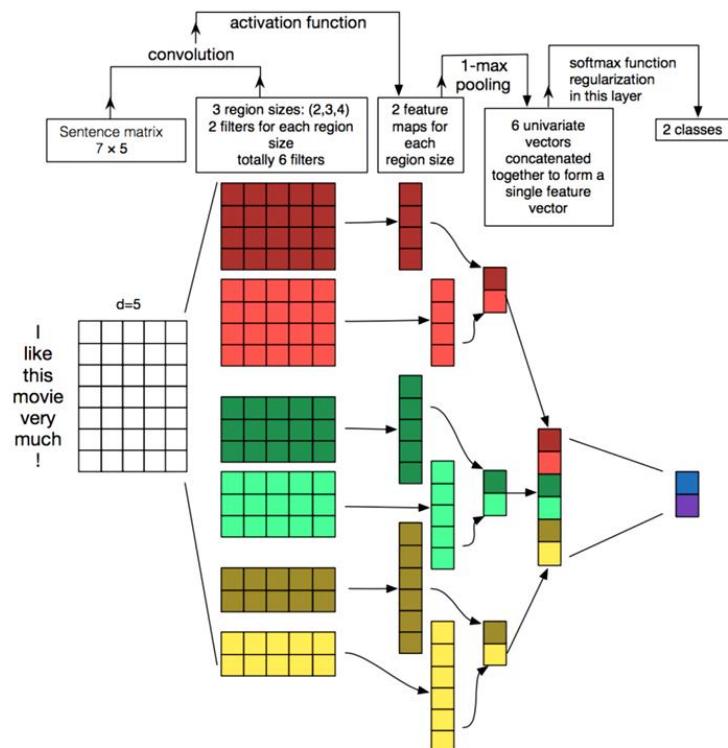


Рис. 4. Архитектура сверточной нейронной сети. Источник: Zhang et al. 2015.

В качестве нейронной сети может выступать, например, сверточная нейронная сеть (Convolutional Neural Network, CNN), рис. 4. Сверточная сеть работает только в том случае, когда на вход подаются данные одной размерности. Поэтому все предложения необходимо сводить к одной длине (длина самого длинного предложения), добавляя неинформативные токены в конец матрицы. Такой процесс называется паддингом (padding). На входе матрица предложения представляет из себя матрицу, где каждое слово пусть будет представлено вектором размерности пять. Всего мы имеем 7 слов в предложении. Далее слова сворачиваются по двум биграммам, двум триграммам и двум четыре-граммам. Для каждой из двух биграмм мы получаем 6 чисел (т.к. всего возможно 6 биграмм), для триграмм – 5 чисел, для четыре-грамм – 4 числа. На пулинговом слое происходит снижение размерности. Далее числа конкатенируются и на выходе мы получаем два класса (например, позитивный и негативный).

## Word2Vec

Word2Vec – совокупность моделей на основе искусственных нейронных сетей, предназначенных для получения векторных представлений слов на естественном языке. Основная идея модели – предсказывать контекст слова. В отличии от моделей, основанных на количественном подходе, данная модель обучается предсказывать контекст.

Контекст слова определяет в какой-то степени само слово, а значит, вектора центрального и контекстных слов должны быть похожи в каком-то смысле. Word2Vec — это итеративный метод. Его основная идея заключается в следующем:

1. взять огромный текстовый корпус;
2. перемещаться по тексту с помощью скользящего окна, сдвигаясь на одно слово за раз. На каждом шаге есть центральное слово и контекстные слова (другие слова в этом окне);

3. для центрального слова вычислить вероятности контекстных слов;
4. отрегулировать векторы, чтобы увеличить эти вероятности.

Задача состоит в том, чтобы предсказывать контекстные слова по центральному слову (рис. 5). Для этого мы должны максимизировать вероятности видеть данные контекстные слова на основе центрального слова.



Рис. 5. Скользящее окно размера два.  $P(w_{t+i}|w_t)$  – вероятность появления слова  $w_{t+i}$  в текущем окне при центральном слове  $w_t$ ,  $i \in \{\pm 1\} \cup \{\pm 2\}$ . Источник: [https://lenavoita.github.io/nlp\\_course/word\\_embeddings.html](https://lenavoita.github.io/nlp_course/word_embeddings.html)

Для каждой позиции  $t = 1, \dots, T$  в текстовом корпусе, Word2Vec предсказывает контекстные слова в окне размером  $m$  с учетом центрального слова  $w_t$ . Функция правдоподобия:

$$L(\theta) = \prod_{t=1}^T \prod_{-m \leq j \leq m, j \neq 0} P(w_{t+j}|w_t, \theta),$$

- где  $\theta$  – оптимизируемый параметр (в нашем случае – это все векторные представления слов, когда они центральные и когда они контекстные). Мы хотим максимизировать функцию правдоподобия или минимизировать функцию потерь. Функция потерь (5) – функция правдоподобия, взятая со знаком минус, прологарифмированная и усредненная по количеству окон:

$$J(\theta) = -\frac{1}{T} \sum_{t=1}^T \sum_{-m \leq j \leq m, j \neq 0} P(w_{t+j} | w_t, \theta) \quad (5)$$

Для каждого слова у нас будет два вектора:

- $v_w$  — когда это центральное слово;
- $u_w$  — когда это контекстное слово.

Тогда для центрального слова ( $c$  - центральное) и контекстного слова ( $o$  - внешнее слово) вероятность контекстного слова равна:

$$P(o|c) = \frac{\exp(u_o^T v_c)}{\sum_{w \in V} \exp(u_c^T v_w)} \quad (6)$$

- где  $u_o^T v_c$  — скалярное произведение, являющееся мерой похожести центрального и контекстного слова, чем больше мера, тем больше слова похожи друг на друга. Знаменатель дроби — нормализация по всему словарю. Формула (6) является SoftMax моделью.

Существует два варианта Word2Vec: Skip-Gram и CBoW (Continuous Bag of Words), рис. 6. Skip-Gram — это модель, которую мы рассматривали до сих пор: она предсказывает контекстные слова по центральному слову. CBoW (Непрерывный мешок слов) предсказывает центральное слово по суммы векторов контекста. Эта простая сумма векторов слов называется «мешком слов». Skip-Gram работает медленнее, чем CBoW, т. к. он предсказывает  $h$  распределений, когда CBoW предсказывает только одно распределение.

A)

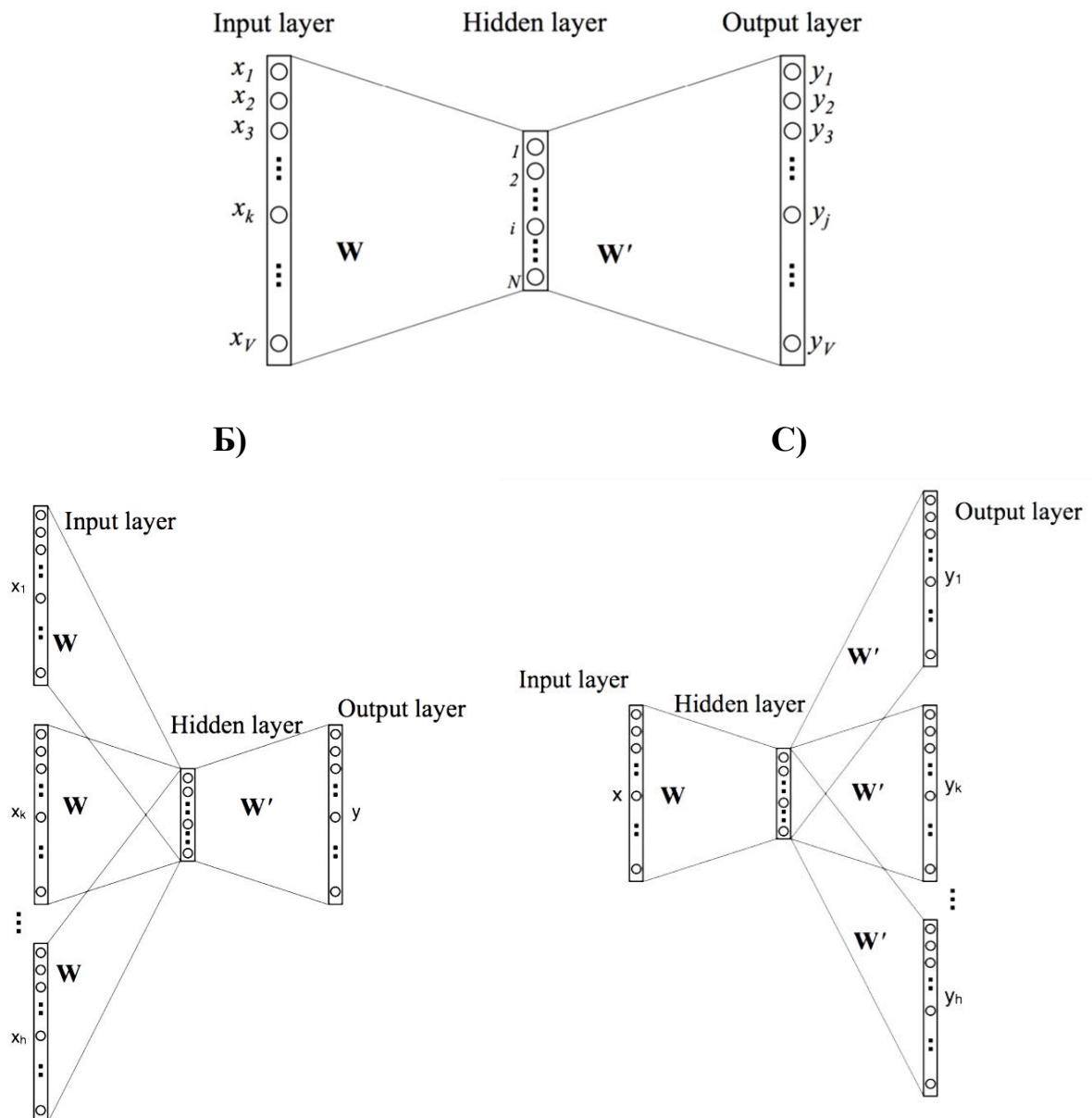


Рис. 6. Архитектура Word2Vec. А) CBoW: предсказывание центрального слова по одному контекстному слову  $i$  ( $h = 1$ ). На вход подаем вектор, где на  $i$ -том месте стоит 1, остальные нули. На выходе получаем вероятности для каждого слова из словаря быть центральным словом. Б) CBoW: предсказывание центрального слова по произвольному количеству контекстных слов. В) Skip-Gram: предсказывание контекстных слов по центральному слову. На выходе получаем столько распределений, сколько контекстных слов умещается в окне. Источник: <https://dyakonov.org/2018/08/28/интерпретации-чёрных-ящиков/>

## Оценка качества Embeddings

Оценить качество эмбеддингов можно с помощью оценки семантической близости между словами. Например, можно осуществить поиск близких слов (фрукты, страны, марки машин), для таких групп расстояния между векторными представлениями должны быть ближе внутри одной группы, чем между словами из разных групп.

Также можно осуществить поиск аналогий (пол: мужской и женский, король/королева, страна – столица (рис. 7), тогда можно выделить ось, вдоль которой слова меняются на их аналоги.

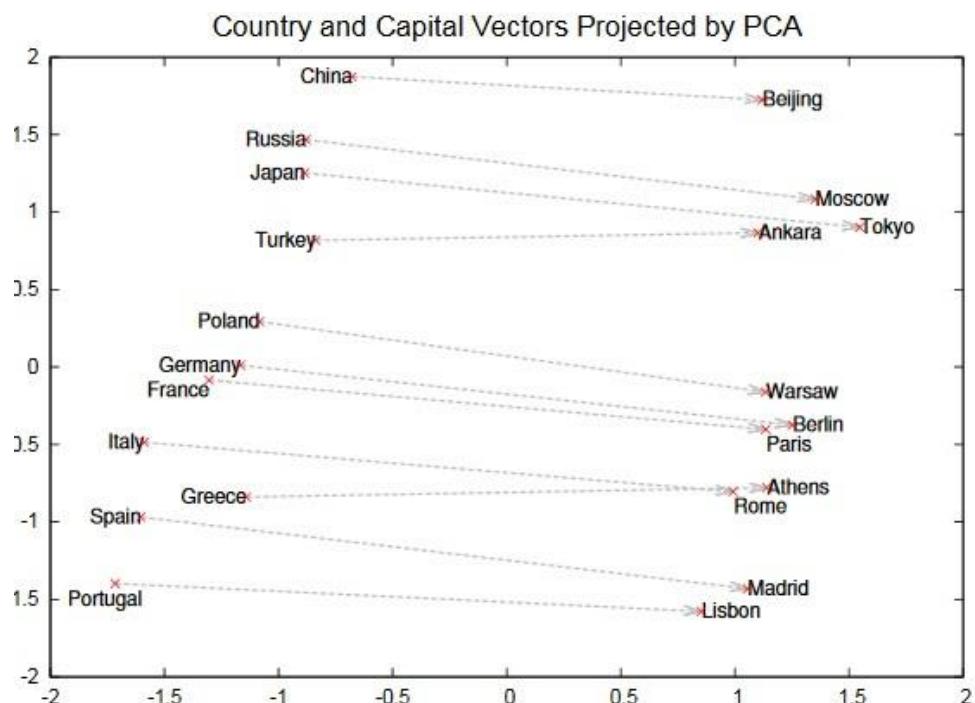


Рис. 7. Поиск аналогий, нахождение направления, вдоль которого стране сопоставляется ее столица. Визуализация сделана при помощи метода опорных векторов (PCA), пространство было сжато до размерности 2. Источник: <https://dyakonov.org/2018/08/28/интерпретации-чёрных-ящиков/>

Наконец, можно оценить качество эмбеддингов при решении NLP задачи (ранжирование, классификация).

## Рекуррентная нейронная сеть (Recurrent Neural Network, RNN)

В задаче классификации стоит задача по входной последовательности предсказать выходную последовательность, то есть оценить последовательность выходных векторов (формула 7):

$D$  – множество размеченных последовательностей  $(x, y)$ ;

$x = \{x_1, \dots, x_N\}$  – последовательность входных объектов;

$y = \{y_1, \dots, y_N\}$  – последовательность выходных объектов (классы);

$$\hat{Y} = \arg \max_Y p(Y|X) \quad (7)$$

Архитектура рекуррентной нейронной сети представлена на рис. 8 и содержит в себе последовательные ячейки. В каждой ячейке выполняются одни и те же операции,  $x_i$  – эмбеддинг слова  $i$ ,  $h_i$  – скрытое состояние (рекуррентное), сопоставляемое эмбеддингу. Вычисление скрытого состояния происходит на каждом шаге согласно равенству (формула (8)):

$$h_t = f(Vx_t + Wh_{t-1} + b) \quad (8)$$

- где  $f$  – некая нелинейная функция, которая берется от линейного преобразования  $x_t$  и  $h_{t-1}$ ,  $(Vx_t + Wh_{t-1})$  – взвешенная сумма входных весов,  $b$  – пороговое значение (сдвиг). После вычисления скрытого состояния вычисление класса происходит согласно равенству:

$$\hat{y} = g(Uh_t + \hat{b}) \quad (9)$$

- где  $g$  – некая нелинейная функция, которая берется от линейного преобразования  $h_t$ ,  $Uh_t$  – взвешенный вход,  $\hat{b}$  – пороговое значение. Далее рассмотрим возможные нелинейные преобразования.

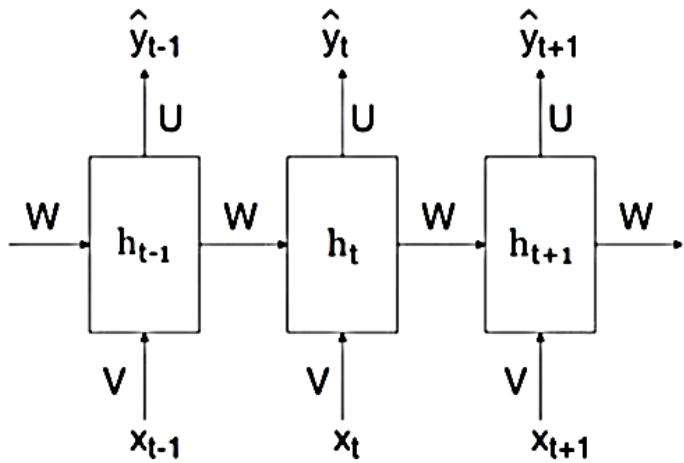


Рис. 8. Архитектура рекуррентной нейронной сети. Источник:  
<https://habr.com/ru/company/wunderfund/blog/331310/>

### Функции активации

Функция активации определяет выходное значение нейрона в зависимости от результата взвешенной суммы входов и порогового значения. Рассмотрим только три возможные функции активации, которые наиболее часто используются в RNN в скрытых слоях (формула (9)), рис. 9:

1. Sigmoid - сигмоидная или логистическая функция активации. Обобщенная функция логистической активации, используемая для мультиклассовой классификации, называется SoftMax. Функция активации сигмоида переводит входные данные в диапазоне  $[-\infty; +\infty]$  к диапазону в  $(0; 1)$ ;
2. tanh – гиперболический тангенс. Обычно именно она используется в RNN;
3. RELU – линейный выпрямитель. RELU обычно не используется в RNN, потому что они могут иметь очень большие выходы, поэтому можно ожидать, что они с большей вероятностью взорвутся, чем единицы, имеющие ограниченные значения.

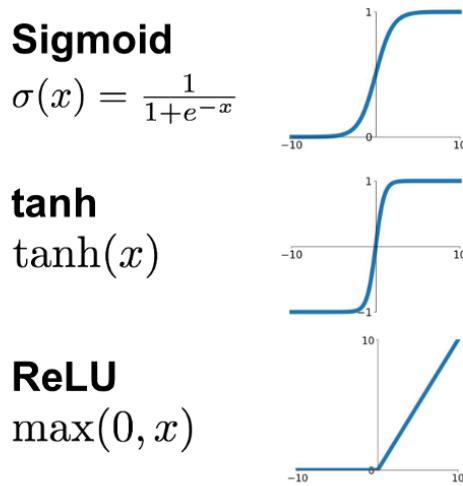


Рис. 9. Функции активации, используемые в RNN (источник: <https://www.machinelearningmastery.ru/complete-guide-of-activation-functions-34076e95d044/>).

В качестве функции активации в последнем слое (формула (10)) можно брать любую функцию активации, например, SoftMax. Softmax является популярным выбором для активации выходного слоя.

### Обучение нейронной сети

Нейронная сеть обучается путем минимизации функции потерь (10):

$$\sum_{t=1}^n L_t(y_t, \hat{y}_t) \rightarrow \min_{V, U, W, b, \hat{b}} \quad (10)$$

В качестве функции потерь может быть взята кросс-энтропия. Кросс-энтропия измеряет расхождение между двумя вероятностными распределениями:

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

В рекуррентных нейронных сетях может возникнуть две проблемы: взрыв градиента или его затухание. Например, возьмем первую производную по  $W$ :

$$\frac{dL_t}{dW} = \frac{\partial L_t}{\partial \hat{y}_t} \frac{\partial \hat{y}_t}{\partial h_t} \frac{dh_t}{dW};$$

$$\frac{dh_t}{dW} = \sum_{k=1}^t \left( \prod_{i=k+1}^t \frac{\partial h_i}{\partial h_{i-1}} \right) \frac{\partial h_k}{\partial W}.$$

Тогда, если под произведением несколько элементов будут устремлены к бесконечности или к нулю, произведение так же будет стремиться к бесконечности или к нулю:

Взрыв градиента $\prod_{i=k+1}^t \frac{\partial h_i}{\partial h_{i-1}} \rightarrow \infty$	Затухание градиента $\prod_{i=k+1}^t \frac{\partial h_i}{\partial h_{i-1}} \rightarrow 0$
-----------------------------------------------------------------------------------------------	----------------------------------------------------------------------------------------------

В первом случае при градиентном спуске мы можем вылететь за область оптимизации, во втором случае «застрять» на одной точке. Борьба с такими проблемами осуществляется или через Gradient Clipping для первой проблемы, или через модели LSTM (модель будет рассмотрена позже) и GRU (Gated Recurrent Unit) для второй.

### LSTM - Long Short-Term Memory Cell

В отличии от RNN с одним слоем, в LSTM содержится четыре слоя. Ключевой компонент LSTM – это состояние ячейки (cell state) – горизонтальная линия, проходящая по верхней части схемы, рис. 10. С помощью состояния ячейки можно доносить некоторую информацию, зависящую от оптимизируемых параметров, тем самым меняя эти параметры, мы можем менять и значение функции потерь.

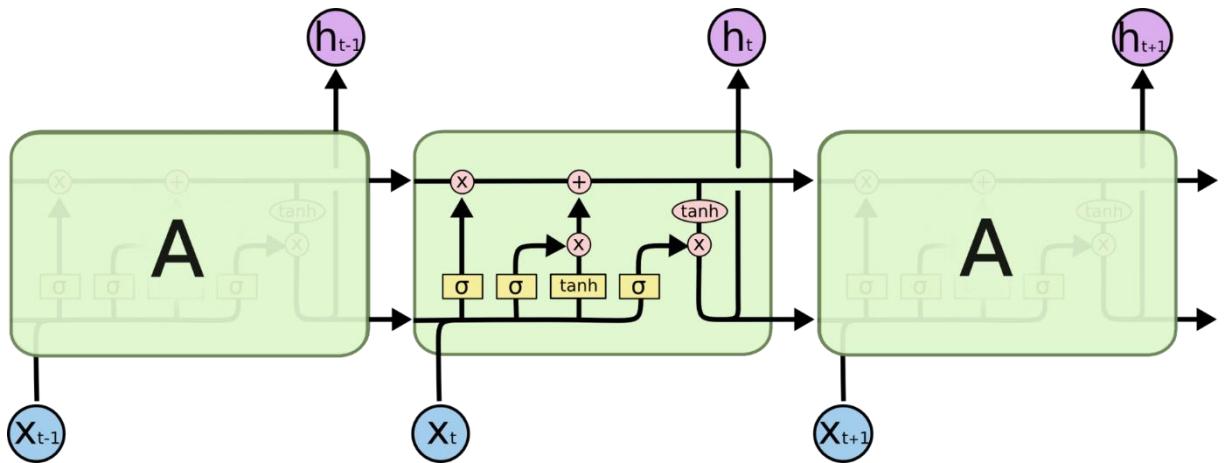


Рис. 10. Архитектура LSTM. Источник:  
<https://habr.com/ru/company/wunderfund/blog/331310/>

На первом шаге (рис. 11, А)) вход обрабатывается сигмоидой, после чего получившееся число от 0 до 1 говорит о том, на сколько важна информация со входа. Чем больше число, тем важнее информация. Это число умножается на  $C_{t-1}$  и передается в состояние ячейки. Этот слой называется «слоем фильтра забывания» (forget gate layer).

Следующий шаг (рис. 11, Б)) состоит из двух частей. Сначала «слой входного фильтра» (input gate layer) определяет, какие значения следует обновить. Потом вход обрабатывается с помощью гиперболического тангенса, т. е. генерируется некий сигнал и умножается на сигмоиду, полученную в первой части. Этот слой называется input modulation gate. После того, как мы определим, в какой степени нужно учитывать сгенерированный сигнал, результат прибавляется к состоянию ячейки.

На следующем шагу (рис. 11, В)) генерируется коэффициент через сигмоиду и он умножается на гиперболический тангенс от  $C_t$ . Этот слой называется output gate.

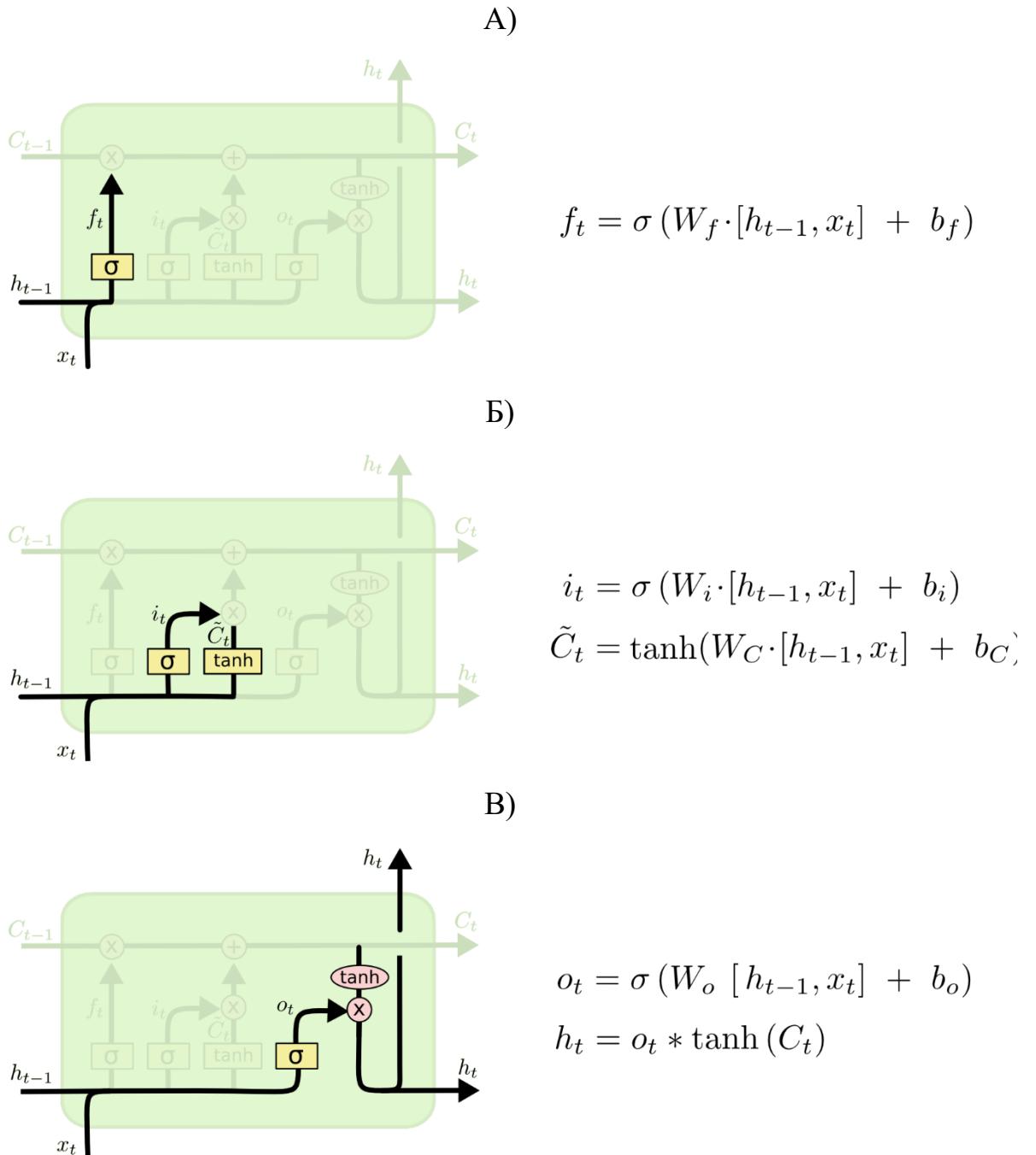


Рис. 11. Пошаговый принцип работы LSTM внутри одной ячейки. Источник:

<https://habr.com/ru/company/wunderfund/blog/331310>

### Average-Stochastic Gradient Descent (SGD) Weight-Dropped LSTM (AWD-LSTM)

AWD-LSTM – это тип рекуррентной нейронной сети, которая использует DropConnect для регуляризации, а также NT-ASGD (Non-monotonically Triggered Averaged SGD) для оптимизации, который возвращает среднее значение последних итераций весов. Дополнительные

используемые методы регуляризации включают в себя variable length backpropagation sequences, variational dropout, embedding dropout, weight tying, independent embedding/hidden size, activation regularization и temporal activation regularization.<sup>19</sup>

## Регуляризация DropConnect

DropConnect является обобщением DropOut. Поэтому сначала опишем, что такое регуляризация DropOut.

DropOut — это метод регуляризации для нейронных сетей, который отбрасывает узел (вместе с соединениями) во время обучения с заданной вероятностью (обычное значение  $p = 0.5$ ). Во время тестирования все узлы присутствуют, но с весами, умноженными на  $p$  (т. е. вместо  $\omega$  вес будет  $p\omega$ ), рис. 12. Идея состоит в том, чтобы предотвратить ко-адаптацию, когда нейронная сеть становится слишком зависимой от определенных соединений, поскольку это может приводить к переобучению. Интуитивно понятно, что DropOut можно рассматривать как создание неявного ансамбля нейронных сетей.

DropConnect обобщает Dropout путем случайного сброса весов, а не активаций (узлов) с вероятностью  $p$ , рис. 13. DropConnect похож на DropOut, поскольку он вводит динамическую разреженность в модель, но отличается тем, что разреженность зависит от весов, а не от выходных векторов слоя.

---

<sup>19</sup> Stephen Merity, Nitish Shirish Keskar, Richard Socher. Regularizing and Optimizing. Режим доступа: LSTM Language Models <https://arxiv.org/pdf/1708.02182.pdf>

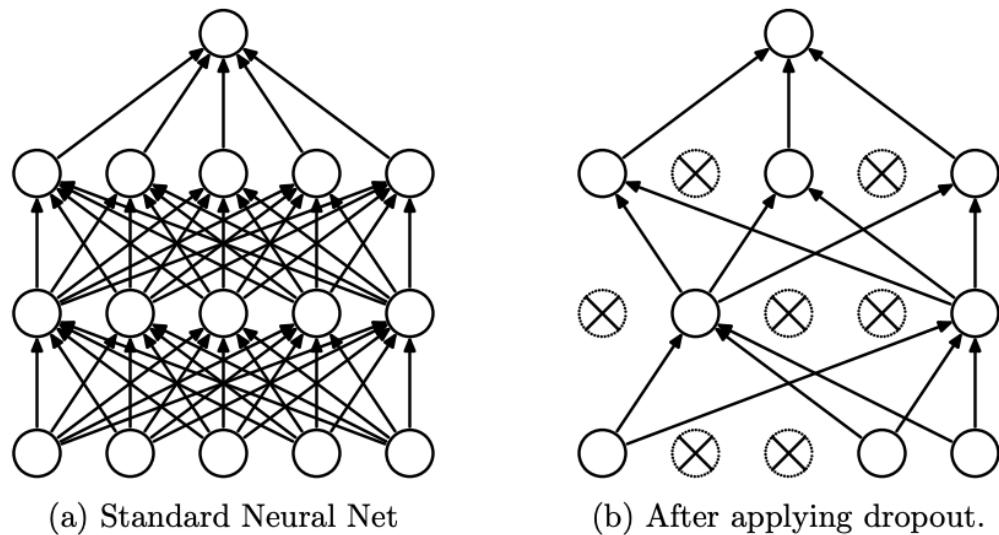


Рис. 12. Архитектура нейронной сети после применения регуляризации DropOut.

Источник: <https://paperswithcode.com/method/dropout#>

Для слоя при регуляризации DropConnect выход выражается как:

$$r = f((M \cdot W)v),$$

- где  $v$  – вход, подаваемый на слой,  $W$  – веса,  $M$  – бинарная матрица, в которой закодирована информация о соединениях,  $M_{ij} \sim Bernoulli(p)$ . Каждый элемент матрицы считается независимо для каждого примера при обучении, таким образом, для каждого примера создаются различные связи (соединения).

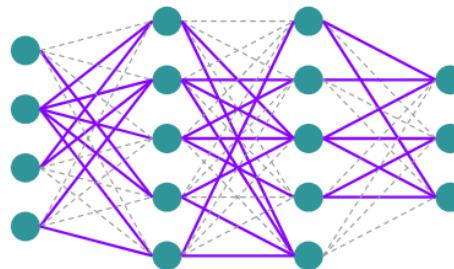


Рис. 13. Архитектура нейронной сети после применения регуляризации DropConnect.  
Источник: <https://towardsdatascience.com/12-main-dropout-methods-mathematical-and-visual-explanation-58cdc2112293>

## NT-ASGD - Non-monotonically Triggered Averaged SGD

NT-ASGD представляет собой метод усредненного стохастического градиентного спуска.<sup>20</sup> В обычном ASGD мы предпринимаем шаги, идентичные обычному SGD, но вместо того, чтобы возвращать последнюю итерацию в качестве решения, мы возвращаем усредненное значение:

$$\frac{1}{(K - T + 1)} \sum_{i=1}^T w_i,$$

- где  $K$  – общее количество итераций,  $T < K$  – определяемый пользователем параметр усреднения. Алгоритм этого метода оптимизации можно увидеть на рис. 14.

---

**Algorithm 1** Non-monotonically Triggered ASGD (NT-ASGD)

---

**Inputs:** Initial point  $w_0$ , learning rate  $\gamma$ , logging interval  $L$ , non-monotone interval  $n$ .

```

1: Initialize  $k \leftarrow 0, t \leftarrow 0, T \leftarrow 0, \text{logs} \leftarrow []$ 
2: while stopping criterion not met do
3:   Compute stochastic gradient  $\hat{\nabla}f(w_k)$  and take SGD
      step (1).
4:   if  $\text{mod}(k, L) = 0$  and  $T = 0$  then
5:     Compute validation perplexity  $v$ .
6:     if  $t > n$  and  $v > \min_{l \in \{t-n, \dots, t\}} \text{logs}[l]$  then
7:       Set  $T \leftarrow k$ 
8:     end if
9:     Append  $v$  to  $\text{logs}$ 
10:     $t \leftarrow t + 1$ 
11:  end if
12: end while
return  $\frac{\sum_{i=T}^k w_i}{(k-T+1)}$ 
```

---

Рис. 14. Алгоритм NT-ASGD. Источник: <https://paperswithcode.com/method/nt-asgd#>

---

<sup>20</sup> Stephen Merity, Nitish Shirish Keskar, Richard Socher. Regularizing and Optimizing. Режим доступа: LSTM Language Models <https://arxiv.org/pdf/1708.02182.pdf>

## **Transfer Learning (трансферное обучение) и языковая модель**

### **ULMFiT (Universal Language Model Fine-tuning)**

Трансферное обучение позволяет адаптировать заранее обученную модель/систему к конкретной задаче с использованием относительно небольшого объема данных. Примерами таких языковых моделей служат ULMFiT, ELMO, OpenAI Transformer, Google BERT.

Далее будет рассмотрена модель ULMFiT, так как именно ее мы используем в данной работе. Эта языковая модель использует архитектуру AWD-LSTM, рассмотренную ранее.

Пусть есть какая-то задача со статичным источником  $T_s$  в интересуемая задача  $T_t$ , и при помощи первой задачи мы хотим улучшить качество решения второй задачи. В качестве задачи  $T_s$  может быть взята, например, задача по моделированию языка, т. к. она охватывает множество аспектов, характеризующих язык и для такой задачи данные почти неограничены в количестве (может быть взят корпус больших размеров).<sup>21</sup> Тогда предобученная модель сможет легко адаптироваться под нашу задачу интереса. Метод так же удобен тем, что, во-первых, он работает для задач с документами различного размера, количества и типа целевой метки. Во-вторых, для всех задач архитектура едини и процесс обучения одинаков. В-третьих, метод не требует дополнительной разработки признаков (feature engineering) и предобработки данных.

Схема обучения такой модели представлена на рис. 15. Она состоит из трех этапов: предобучение языковой модели на общем корпусе, дообучение языковой модели на данных задачи, обучение классификатора текста.

---

<sup>21</sup> Howard J., Ruder S. Universal Language Model Fine-tuning for Text Classification. Режим доступа: <https://arxiv.org/pdf/1801.06146.pdf>

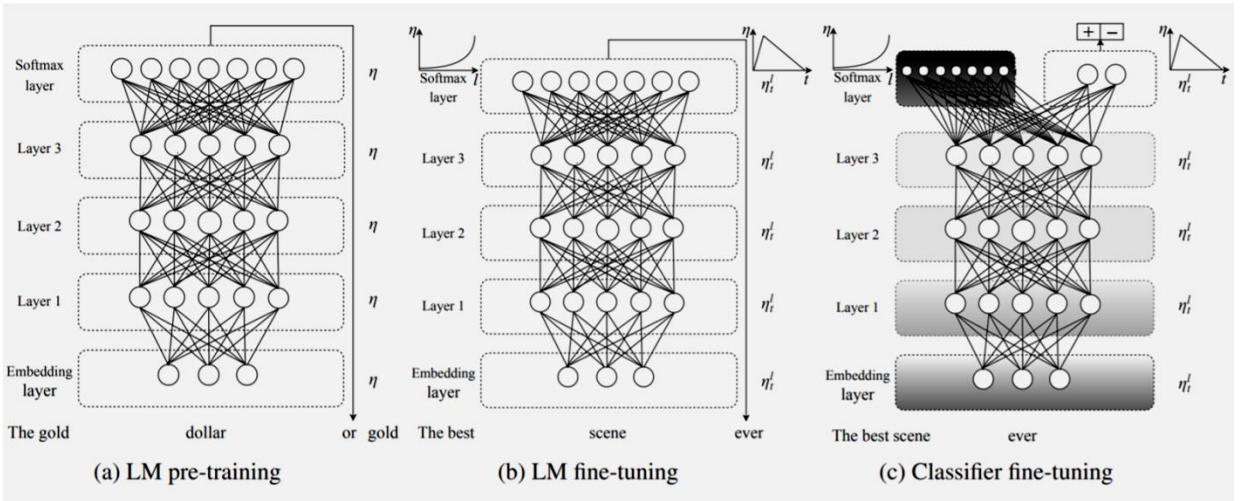


Рис. 15. Стадии обучения модели ULMFiT. Источник: Howard J., Ruder S. Universal Language Model Fine-tuning for Text Classification. Режим доступа: <https://arxiv.org/pdf/1801.06146.pdf>

### Предобучение языковой модели (LM pre-training)

Предобученная языковая модель, которая обучается на общем корпусе. Корпус должен охватывать общие свойства языка. Подробнее об используемом нами корпусе будет написано ниже. Этот этап самый дорогостоящий, но его нужно выполнить только единожды.

### 1. Дообучение языковой модели на данных задачи (LM fine-tuning)

Этап проводится независимо от того, насколько разнообразен общий корпус, т. к. данные в задаче интереса скорее всего имеют другое распределение. Модель донастраивается (fine tune) на данных задачи с помощью дискриминационной донастройки (Discriminative fine-tuning, ‘Discr’) и изменяющимся по наклонному треугольнику шагом (Slanted triangular learning rates, STLR).

**Дообучение (fine-tuning).** Суть дообучения заключается в размораживании последних слоев нейронной сети и их обучении. Таким образом, корректируются слои, которые имеют наиболее абстрактные представления. Производя дообучение только нескольких слоев, мы

уменьшаем риск переобучения. И самое главное, это позволяет сделать текущую модель ещё более подходящей к нашей задаче.<sup>22</sup>

### Дискриминационное дообучение (Discriminative fine-tuning).

Вместо использования одинаковой скорости обучения для всех слоев модели, ‘Discr’ позволяет нам настраивать каждый слой с различной скоростью. Для контекста обычный стохастический градиентный спуск модифицируется через параметр  $\theta$ . Обычно SGD на шаге  $t$  вычисляется как:

$$\theta_t = \theta_{t-1} - \eta \cdot \nabla_{\theta} J(\theta) \quad (11)$$

- где  $\eta$  – скорость обучения,  $\nabla_{\theta} J(\theta)$  – градиент функции потерь. В ‘Discr’  $\theta = \{\theta^1, \dots, \theta^L\}$  и  $\eta = \{\eta^1, \dots, \eta^L\}$ , где  $\theta^l$  – хранит в себе параметры модели на  $l$ -том слое,  $L$  – количество слоев,  $\eta^l$  – скорость обучения (шаг) на  $l$ -том слое. Тогда формула (11) сводится к формуле (12):

$$\theta_t^l = \theta_{t-1}^l - \eta^l \cdot \nabla_{\theta^l} J(\theta) \quad (12)$$

Эмпирически выявлено, что лучше всего сначала подбирать  $\eta^L$  на последнем слое, дообучая модель только в последнем слое, а далее брать  $\eta^{l-1} = \eta^l / 2.6$ .

**Изменяющийся по наклонному треугольнику шаг (STLR).** Чтобы модель могла быстро сойтись к подходящей области пространства для параметров, а затем только немного корректировать их значения, мы изменяем шаг по наклонному треугольнику (рис. 16).

---

<sup>22</sup> Как повысить точность ML-модели. [Электронный ресурс]. Режим доступа: <https://python-school.ru/fine-tuning/> (дата обращения: 04.05.2021)

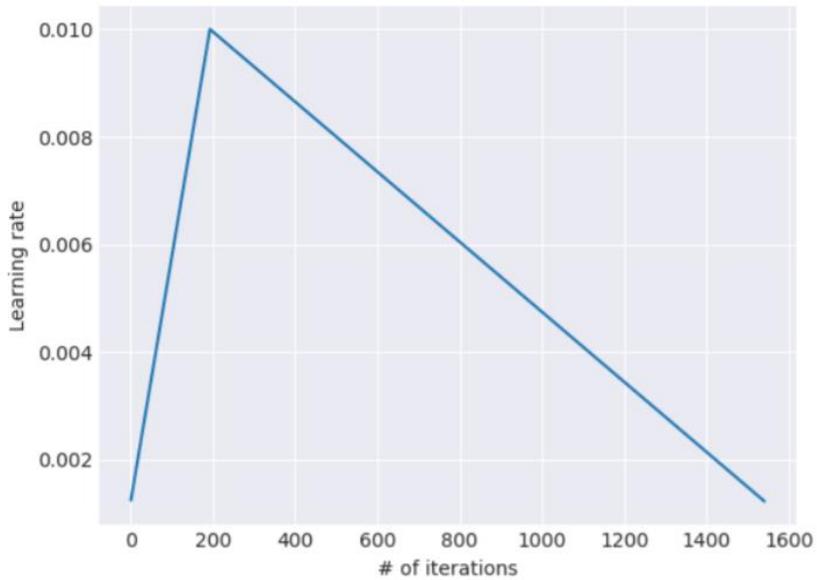


Рис. 16 Моделирование изменяющегося по наклонному треугольнику шага (STLR).

Источник: <https://python-school.ru/fine-tuning/>

$cut\_frac$  – доля итераций, на которой мы увеличиваем следующий шаг;  
 $cut = T \cdot cut_{frac}$ ;

$$p = \begin{cases} \frac{t}{cut}, & \text{if } t < cut \\ 1 - \frac{t - cut}{cut \cdot \left(\frac{1}{cut_{frac}} - 1\right)}, & \text{otherwise} \end{cases} \quad (13)$$

$$\eta_t = \eta_{max} \frac{1 + p(ratio - 1)}{ratio};$$

- где  $cut$  – это итерация, на которой функция шага начинает убывать,  $p$  – текущая доля итераций, в которые функция шага возрастила или убывала соответственно,  $ratio$  – отношение максимального шага к минимальному,  $\eta_t$  – шаг на итерации  $t$ . По умолчанию могут быть использованы следующие параметры:  $cut\_frac = 0.1$ ,  $\eta_{max} = 0.01$ ,  $ratio = 32$ .

## 2. Обучение классификатора текста (Classifier fine-tunning)

Для классификатора текста модель дополняется двумя линейными блоками. Для каждого из этих блоков используется Batch Normalization и DropOut, функцией активации для средних слоев служит линейный

выпрямитель (ReLU), для последнего слоя – SoftMax. Первый линейный слой принимает в качестве входных данных объединенные состояния последнего скрытого слоя. Классификатор настраивается с использованием постепенного размораживания (gradual unfreezing), ‘Discr’ и STLR.

**Постепенное размораживание (gradual unfreezing).** Вместо того, чтобы одновременно донастраивать все слои за раз, модель размораживается постепенно, начиная с последнего слоя, содержащего наименьшее количество общих знаний. После разморозки последнего слоя донастраиваются все незамороженные слои. Далее размораживается следующий слой и донастройка происходит еще раз. Так происходит слой за слоем. На рисунке 15 затемненные слои – слои, постепенно размораживающиеся, черный слой – замороженный слой.

Как можно заметить, существует множество способов анализа тональности текста, выше были перечислены основные из них. С некоторыми методами можно подробнее ознакомиться в главе 3, где представлены результаты самого исследования.

## **Глава 2 Исследование волатильности на фондовом рынке и факторов, влияющих на ее динамику**

В данной работе будут рассмотрены 7 различных российских компаний: Яндекс, Сбербанк, МТС, Лукойл, Роснефть, Газпром, Новатэк. Все компании котируются на Московской фондовой бирже. Выбор этих компаний мотивирован тем фактом, что переменные настроения и внимания могут по-разному влиять на будущую реализованную волатильность в зависимости от типа рассматриваемой акции. По этой причине мы выбираем компании из разных отраслей (например, Яндекс является компанией, предоставляющей услуги в сфере ИТ; банковские услуги предоставляет компания Сбербанк; в сфере нефтедобычи задействованы следующие компании - Лукойл, Роснефть; в сфере газодобычи – Газпром, Новатэк; телекоммуникационные услуги – МТС) – см. табл. 6. За исследуемый период возьмем период с 01.03.2016 по 28.02.2021.

Таблица 6

Выбор компаний для исследования

	<b>Отрасль</b>	<b>Компания</b>
1	ИТ	Яндекс
2	Банки	Сбербанк
3	Телеком	МТС
4	Нефтедобыча	Лукойл
5	Нефтедобыча	Роснефть
6	Газодобыча	Газпром
7	Газодобыча	Новатэк

Источник: построено автором

### **2.1 Реализованная волатильность как альтернативный метод оценки динамики фондового рынка**

Перед проведением основного анализа разберем, почему для исследования была выбрана именно гетерогенная авторегрессия реализованной волатильности (The Heterogeneous Autoregressive model of the Realized Volatility, HAR-RV, HAR). Концепция реализованной волатильности была введена в начале 2000-х годов (Andersen et al., 2001; Barndorff-Nielsen & Shephard, 2002). Идея состоит не в использовании моделей типа GARCH или общих стохастических моделей волатильности, а в вычислении оценок волатильности с использованием высокочастотных данных.

Дневная реализованная волатильность оценивается на основе внутридневной 5-минутной доходности с использованием оценки медианной реализованной волатильности (MedRV), введенной Андерсеном, Добревом и Шаумбургом (2012). Эта оценка используется потому, что она показала хорошую устойчивость к скачкам и «нулевой» доходности.

Используемая в исследовании модель HAR-RV основана на принципе поведения агентов на фондовом рынке. Они отличаются своим восприятием в зависимости от того, какие у них планы на инвестиции: кратко-, средне- или долгосрочные. Это влечет за собой наличие гетерогенности на рынке. Ключевая идея оценки состоит в том, что агенты различных временных горизонтов воспринимают, реагируют и влияют на разные компоненты волатильности (используем три компонента волатильности, различающихся временными горизонтами).

Кроме этого, здесь присущ каскад гетерогенности компонентов волатильности. Волатильность за долгосрочный временной период имеет сильное влияние на краткосрочную волатильность, обратное не верно. Это можно экономически обосновать: краткосрочные агенты учитывают уровень долгосрочной волатильности, т.к. он определяет ожидаемые тренды и риски.

Следует отметить также следующие характерные черты финансового рынка:<sup>23</sup>

1. «долгая память» - долгосрочная зависимость: реализованная волатильность показывает значимую автокорреляцию даже очень больших лагов. Чтобы учитывать эту особенность мы используем разные временные шкалы (для краткосрочного влияния используем дневной лаг, для среднесрочного – недельный, для долгосрочного – месячный);
2. «эффект рычага»: доходность отрицательно коррелирована с реализованной волатильностью. В частности, обычно отрицательные доходности в прошлом влекут за собой всплеск волатильности в настоящем;
3. «скачки»: цены подвержены резким скачкам. Такие скачки не часты, но непредсказуемы и вносят большой положительный вклад в будущую волатильность.

### **Медианная реализованная волатильность за день (MedRV)**

Мы рассматриваем одномерный процесс  $Y = \{Y_t\}_{0 \leq t \leq 1}$ ,  $Y_t$  – логарифм цены в дискретный момент времени  $t$  внутри одного дня<sup>24</sup>. В день имеется  $N + 1$  наблюдений цен в дискретном наборе точек  $0 \leq t_0 \leq \dots \leq t_N \leq 1$ , мы брали данные за каждые 5 минут внутри одного дня. Тогда для каждого 5-минутного интервала доходность и временной интервал определяются как  $\Delta Y_i = Y_{t_i} - Y_{t_{i-1}}$  и  $\Delta t_i = t_i - t_{i-1}$ ,  $i = 1, \dots, N$ .

---

<sup>23</sup> 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

<sup>24</sup> Andersen T. Jump-robust volatility estimation using nearest neighbor truncation. Journal of Econometrics (169) – 2012. 75 - 93

Интегрируемая дисперсия (integrated variance, IV) или непрерывная часть квадратичной вариации определяется как:

$$IV = \int_0^1 \sigma_u^2 du \quad (14)$$

- где  $\sigma$  – вариация процесса. Этую интегрируемую дисперсию можно оценить с помощью MedRV, или ее альтернативы – MinRV<sup>25</sup>:

$$\begin{aligned} \text{MinRV}_N &= \frac{\pi}{\pi - 2} \left( \frac{N}{N - 1} \right) \sum_{i=1}^{N-1} \min(|\Delta Y_i|, |\Delta Y_{i+1}|)^2 \\ \text{MedRV}_N &= \frac{\pi}{6 - 4\sqrt{3} + \pi} \left( \frac{N}{N - 2} \right) \sum_{i=2}^{N-1} \text{med}(|\Delta Y_{i-1}|, |\Delta Y_i|, |\Delta Y_{i+1}|)^2 \end{aligned} \quad (15)$$

Следует отметить, что функции  $\min$  и  $\text{med}$  устраняют большие скачки. Например, если большой скачок наблюдается в одной из двух переменных внутри оператора  $\min$ , он просто отбирает другую переменную. Поэтому в смещение оценки вносит вклад только количество прыжков, а не их размеры. Обе оценки являются состоятельными для интегрируемой дисперсии и при  $N \rightarrow \infty$  сводятся к (14) [Andersen T., Dobrev D., 2012].

Таким образом, мы выбрали оценку (15) для реализованной волатильности, в частности потому, что в работе [Francesco A., 2020] была взята эта оценка. Для построения моделей мы, как было написано ранее, взяли переменные реализованной волатильности, учитывающие кратко-, средне- и долгосрочное поведение динамики акций (см. табл.7).

Таблица 7

---

<sup>25</sup> Andersen T. Jump-robust volatility estimation using nearest neighbor truncation. Journal of Econometrics (169) – 2012. 75 - 93

Номер	Наименование фактора	Обозначение	Расчет
1.1	Реализованная волатильность в день $t$	$RV_d$	$\log RV_t^{(d)}$
1.2	Лагированная недельная волатильность за предыдущую	$RV_w$	$\frac{1}{5} \sum_{i=1}^5 \log RV_{t-i+1}^{(d)}$
1.3	Лагированная месячная волатильность за предыдущую	$RV_m$	$\frac{1}{22} \sum_{i=1}^{22} \log RV_{t-1+i}^{(d)}$

Источник: построено автором

## 2.2 Экономические и финансовые факторы, влияющие на фондовый рынок

В данной части работы будет производиться анализ экономических и финансовых показателей (30 переменных). Все показатели разделены на 5 больших групп: 1) переменные фондового рынка (7 переменных); 2) переменные рынка облигаций (5 переменных); 3) переменные обменного курса (5 переменных); 4) переменные ликвидности (2 переменных); 5) макроэкономические переменные (11 переменных). Помимо этого, все переменные разделены на две другие большие группы: 1) общие факторы – одинаковы для всех компаний; 2) специфичные факторы – отличны и характерны для каждой определенной компании. Мы столкнулись с проблемой в сборе факторов второй группы, т.к. такие данные как правила отсутствуют в открытом доступе, поэтому их число очень ограничено. Все данные ежедневные.

### Переменные фондового рынка

Переменные фондового рынка: этот набор включает широкий диапазон показателей. В работе среди основных показателей фондового рынка будут рассматриваться:

1) Индекс Московской Биржи. Данный показатель будет играть роль аналога показателя Dow Jones. Ценовой, взвешенный по рыночной капитализации (free-float) композитный индекс российского фондового рынка, включающий наиболее ликвидные акции крупнейших и динамично развивающихся российских эмитентов, виды экономической деятельности которых относятся к основным секторам экономики<sup>26</sup>;

2) Индекс волатильности российского рынка (RVI) - индикатор срочного рынка, который рассчитывается на основе волатильности фактических цен опционов на Индекс РТС. При расчёте индекса используются цены ближайшей и следующей за ней серий опционов со сроком до экспирации более 30 дней<sup>27</sup>;

3) Изменение в Индексе волатильности российского рынка в процентах.

4) Индекс MSCI. Фондовый индекс российского рынка, входящий в группу индексов развивающихся рынков MSCI Emerging Markets.<sup>28</sup>

5) Доходность акции, %. Показатель прибыли от ценных бумаг в процентном либо номинальном выражении. Представляет собой суммированную прибыль, поделенную на сумму вложений.

6) Объем торгов в руб. рассчитывается как суммарное число акций или контрактов (лотов), сменивших владельца за торговый период. Крупный объем торгов указывает на высокий интерес у участников рынка к данному финансовому инструменту.

---

<sup>26</sup> Индекс МосБиржи и Индекс РТС. [Электронный ресурс]. Режим доступа: <https://www.moex.com/ru/index/IMOEX> (дата обращения: 03.04.2021)

<sup>27</sup> Индекс волатильности российского рынка. [Электронный ресурс]. Режим доступа: <https://www.moex.com/ru/index/RVI> (дата обращения: 03.04.2021)

<sup>28</sup> Индекс MSCI. [Электронный ресурс]. Режим доступа: <https://ru.investing.com/indices/msci-russia> (дата обращения: 22.04.2021)

7) Dow Jones Industrial Average (DJ). Старейший из существующих американских рыночных индексов. Он был создан для отслеживания развития промышленной составляющей американских фондовых рынков.

8) Индекс РТС. Фондовый индекс, основной индикатор фондового рынка России.<sup>29</sup> Для расчета данного показателя рассчитывается группа акций.

С подробной информацией о данных переменных можно ознакомиться в табл.8.

Таблица 8  
Переменные фондового рынка

Номер	Наименование фактора	Обозначение	Общий / специфичный	Источник
2.1.1	Индекс IMOEX	211 IMOEX	Общий	<a href="https://ru.investing.com/indices/mcx-historical-data">https://ru.investing.com/indices/mcx-historical-data</a>
2.1.2	Индекс RVI	212_RVI_level	Общий	<a href="https://ru.investing.com/equities/retail-value-historical-data">https://ru.investing.com/equities/retail-value-historical-data</a>
2.1.3	Изменение в Индексе RVI, %	213_RVI_change	Общий	<a href="https://ru.investing.com/equities/retail-value-historical-data">https://ru.investing.com/equities/retail-value-historical-data</a>
2.1.4	Индекс MSCI Russia	214_MSCI	Общий	<a href="https://ru.investing.com/indices/msci-russia-historical-data">https://ru.investing.com/indices/msci-russia-historical-data</a>
2.1.5	Доходность акции, %	215_Return	Специфичный	<a href="https://ru.investing.com/">https://ru.investing.com/</a>
2.1.6	Объем торгов, руб.	216_Volume	Специфичный	<a href="https://ru.investing.com/">https://ru.investing.com/</a>
2.1.7	Dow Jones Industrial Average	216_DJ	Общий	<a href="https://ru.investing.com/indices/us-30-historical-data">https://ru.investing.com/indices/us-30-historical-data</a>
2.1.8	Индекс РТС	217_IRTS	Общий	<a href="https://ru.investing.com/indices/rtsi">https://ru.investing.com/indices/rtsi</a>

Источник: построено автором на основе данных финансового портала [finam.ru](http://finam.ru)

<sup>29</sup> Индекс РТС. [Электронный ресурс]. Режим доступа: <https://ru.investing.com/indices/rtsi> (дата обращения: 22.04.2021)

## **Переменные рынка облигаций**

Переменные рынка облигаций: этот набор состоит из процентных ставок, срочных спредов и премий за риск по облигациям. Данная категория представлена следующими факторами (подробнее в табл.9):

1) Доходность российских облигаций за 10 лет в процентах. Напомним, что основной показатель - доходность к продаже - рассчитывается по формуле:

$$ДП = ТД + ((НО - ЦП) / ЦП) \times (365 / В) \times 100 \%,$$

где: ДП - простая доходность к продаже (к погашению);

ТД - текущая доходность от купона;

НО - номинал облигации;

ЦП - цена покупки;

В - время от покупки до продажи (или погашения).

2) Годовая доходность российских облигаций (годовая ставка) в процентах.

3) Дневное изменение в доходности российских облигаций Россия за 10-лет (годовая ставка) в процентах.

4) Дневное изменение в доходности облигации Россия за год в процентах. В течение дня волатильность может иметь значительные флуктуации, которые при вычислении на более длительные периоды может не отображаться в данных.

5) Разница между доходностью 10-летней облигации и годовой. Показатель используется для: 1) расчета динамики (роста/спада) стоимости облигации; 2) расчета изменения доходности акции за указанные периоды, а, следовательно, и выгодности вложения средств в эти акции.

Таблица 9  
Переменные рынка облигаций

<b>Номер</b>	<b>Наименование фактора</b>	<b>Обозначение</b>	<b>Общий / специфичный</b>	<b>Источник</b>
2.2.1	Доходность облигации России 10-летние (годовая ставка), %	221_OFZ_rate_LR	Общий	<a href="https://ru.investing.com/rates-bonds/russia-10-year-bond-yield-historical-data">https://ru.investing.com/rates-bonds/russia-10-year-bond-yield-historical-data</a>
2.2.2	Доходность облигации Россия годовые (годовая ставка), %	222_OFZ_rate_SR	Общий	<a href="https://ru.investing.com/rates-bonds/russia-1-year-bond-yield-historical-data">https://ru.investing.com/rates-bonds/russia-1-year-bond-yield-historical-data</a>
2.2.3	Дневное изменение в доходности облигации Россия 10-летние (годовая ставка), %	223_Return_OFZ_LR	Общий	<a href="https://ru.investing.com/rates-bonds/russia-10-year-bond-yield-historical-data">https://ru.investing.com/rates-bonds/russia-10-year-bond-yield-historical-data</a>
2.2.4	Дневное изменение в доходности облигации Россия годовые (годовая ставка), %	224_Return_OFZ_SR	Общий	<a href="https://ru.investing.com/rates-bonds/russia-1-year-bond-yield-historical-data">https://ru.investing.com/rates-bonds/russia-1-year-bond-yield-historical-data</a>
2.2.5	Разница между доходностью 10-летней облигации и годовой	225_Yield_diff_LR_SR	Общий	<a href="https://ru.investing.com/rates-bonds/russia-10-year-bond-yield-historical-data">https://ru.investing.com/rates-bonds/russia-10-year-bond-yield-historical-data</a> <a href="https://ru.investing.com/rates-bonds/russia-1-year-bond-yield-historical-data">https://ru.investing.com/rates-bonds/russia-1-year-bond-yield-historical-data</a>

Источник: построено автором на основе данных финансового портала finam.ru

### **Переменные обменного курса**

Переменные обменного курса: эта категория содержит набор переменных, которые отражают премию за риск и вариацию доходности на валютных рынках (Lustig, Roussanov, & Verdelhan, 2011, 2014). В частности, необходимо включить доходность обменных курсов между долларом США и четырьмя основными валютами (евро, швейцарский франк, британский фунт и японскую иену), коэффициент carry trade и средний форвардный дисконт, который измеряет разницу процентных ставок между долларом США и различными наборами иностранных валют. В данной работе будет проанализировано отношение рубля к основным валютам мира (доллар, евро, швейцарский франк, фунт стерлингов, японская иена). К сожалению, данных carry trade нет в бесплатном открытом доступе, поэтому в данном исследовании эти переменные не будут включены в модель. Подробнее в табл.10.

Таблица 10

## Переменные обменного курса

Номер	Наименование фактора	Обозначение	Общий / специфичный	Источник
2.3.1	Курс доллара к рублю, руб	231_USD/RUB	Общий	<a href="https://ru.investing.com/currencies/usd-rub-historical-data">https://ru.investing.com/currencies/usd-rub-historical-data</a>
2.3.2	Курс евро к рублю, руб	232_EUR/RUB	Общий	<a href="https://ru.investing.com/currencies/eur-rub-historical-data">https://ru.investing.com/currencies/eur-rub-historical-data</a>
2.3.3	Курс швейцарского франка к рублю, руб	233_CHF/RUB	Общий	<a href="https://ru.investing.com/currencies/chf-rub-historical-data">https://ru.investing.com/currencies/chf-rub-historical-data</a>
2.3.4	Курс фунта стерлингов к рублю, руб	234_GBP/RUB	Общий	<a href="https://ru.investing.com/currencies/gbp-rub-historical-data">https://ru.investing.com/currencies/gbp-rub-historical-data</a>
2.3.5	Курс японской иены к рублю, руб	235_JPY/RUB	Общий	<a href="https://ru.investing.com/currencies/jpy-rub-historical-data">https://ru.investing.com/currencies/jpy-rub-historical-data</a>

Источник: построено автором на основе данных финансового портала [finam.ru](http://finam.ru)

### Переменные ликвидности

Переменные ликвидности: этот набор включает переменные, которые фиксируют ликвидность на финансовых рынках. В частности, в качестве показателей ликвидности рассматриваются коэффициенты оборачиваемости (самой акции и двух основных фондовых индексов: MSCI World и Dow Jones Industrial Average), спред по умолчанию (измеряемый как разница между доходностями Baa- и Aaa- рейтинговых корпоративных облигаций), средний спред между для пяти основных валют и спред TED (разница между трехмесячной ставкой LIBOR и ставкой казначейских векселей). В качестве переменных ликвидности в работе были проанализированы следующие показатели:

- 1) Объем торгов индекса МосБиржи. Рассчитывается как суммарное число акций или контрактов (лотов), сменивших владельца за торговый период.
- 2) Объем торгов Dow Jones Industrial Average. Показатель, оценивающий объем торгов американского производства товаров. Используется как универсальный показатель развития сферы производства.

Показатель используется также как универсальное средство для анализа различных изменений (динамики) в экономике (см. табл.11).

Таблица 11

Переменные ликвидности

Номер	Наименование фактора	Обозначение	Общий / специфичный	Источник
2.4.1	Объем торгов индекса МосБиржи	241_IMOEX_vol	Общий	<a href="https://ru.investing.com/indices/mcx-historical-data">https://ru.investing.com/indices/mcx-historical-data</a>
2.4.2	Объем торгов Dow Jones Industrial Average	242_DJI_vol	Общий	<a href="https://ru.investing.com/indices/us-30-historical-data">https://ru.investing.com/indices/us-30-historical-data</a>

Источник: построено автором на основе данных финансового портала finam.ru

### Макроэкономические переменные

Макроэкономические переменные: эта категория включает широкий спектр макроэкономических временных рядов. Мы учитываем инфляцию, показатели производства, переменные рынка труда, цены на сырьевые товары, денежную массу, а также настроения потребителей и производителей, измеряемые в ходе опросов.

В данной работе для исследования были взяты следующие показатели (см. табл.12):

- 1) Индекс потребительских цен – индекс, который используется для измерения среднего уровня изменения цен на услуги и товары.
- 2) Ожидаемый индекс потребительских цен. Данный индекс анализирует прогнозы о росте/спаде потребительских цен в ближайшем будущем. Фактор важен, так как он чувствителен к колебаниям на фондовом рынке.
- 3) Объём промышленного производства в России. Измеряет выпуск производственных предприятий промышленности, добывающих отраслей и энергоснабжения.
- 4) Денежная масса, млрд рублей. Совокупность наличных денег, находящихся в обращении, и безналичных средств на счетах, которыми

располагают физические и юридические лица и государство. Объем денежной массы в экономике – один из ключевых элементов монетарной политики.

5) Первая разность денежной массы, млрд рублей. Данный показатель используется для расчета разницы между периодами  $t$  и  $t+1$ , таким образом легче проследить динамику изменения денежной массы.

6) Первая разность денежной массы сезонно скорректированной, млрд рублей. Показатель используется для исправления погрешности показателя выше.

7) Доходность индекса CRB (The Thomson Reuters/Jefferies CRB Index) - индикатор сырьевого рынка, который рассчитывается по котировкам 19 товаров. Индекс CRB был разработан для отслеживания общих ценовых тенденций на сырьевом рынке.

8) Уровень безработицы в процентах. Безработица определяется как отношение числа безработных к общему числу трудоспособного населения (сумма безработных и рабочих).

9) Индекс производственной активности PMI России. (National Association of Purchasing Managers - PMI index). Отчет представляет собой результаты опроса менеджеров по закупкам в сфере промышленности.<sup>30</sup>

10) Фьючерс на нефть Brent. Напомним, что фьючерс — производный финансовый инструмент на бирже купли-продажи базового актива, при заключении которого стороны договариваются только об уровне цены и сроке поставки.<sup>31</sup>

---

<sup>30</sup> Индекс производственной активности PMI России. [Электронный ресурс]. Режим доступа: <https://ru.investing.com/economic-calendar/russian-markit-manufacturing-pmi-1630> (дата обращения: 03.04.2021)

<sup>31</sup> Фьючерс. РБК. [Электронный ресурс]. Режим доступа: <https://quote.rbc.ru/dict/Futures> https://quote.rbc.ru/dict/Futures (дата обращения: 03.05.2021)

Таблица 12  
Макроэкономические переменные

Номер	Наименование фактора	Обозначение	Общий / специфичный	Источник
2.5.1	Индекс потребительских цен (ИПЦ)	251_CPI	Общий	<a href="https://ru.investing.com/economic-calendar/russian-cpi-1180">https://ru.investing.com/economic-calendar/russian-cpi-1180</a>
2.5.2	Ожидаемый индекс потребительских цен (ИПЦ)	252_CPI_Expected	Общий	<a href="https://ru.investing.com/economic-calendar/russian-cpi-1180">https://ru.investing.com/economic-calendar/russian-cpi-1180</a>
2.5.3	Объём промышленного производства в России	253_Indst_production	Общий	<a href="https://ru.investing.com/economic-calendar/russian-industrial-production-553">https://ru.investing.com/economic-calendar/russian-industrial-production-553</a>
2.5.4	Денежная масса, млрд рублей	254_M1	Общий	<a href="https://www.cbr.ru/statistics/macro_itm/dkfs/">https://www.cbr.ru/statistics/macro_itm/dkfs/</a>
2.5.5	Денежная масса сезонно скорректированная, млрд рублей	255_M1_SA	Специфичный	<a href="https://www.cbr.ru/statistics/macro_itm/dkfs/">https://www.cbr.ru/statistics/macro_itm/dkfs/</a>
2.5.6	Первая разность денежной массы, млрд рублей	256_FD_M2	Специфичный	<a href="https://www.cbr.ru/statistics/macro_itm/dkfs/">https://www.cbr.ru/statistics/macro_itm/dkfs/</a>
2.5.7	Первая разность денежной массы сезонно скорректированной, млрд рублей	257_FD_M2_SA	Общий	<a href="https://www.cbr.ru/statistics/macro_itm/dkfs/">https://www.cbr.ru/statistics/macro_itm/dkfs/</a>
2.5.8	Доходность индекса CRB	258_Return_CRB	Общий	<a href="https://ru.investing.com/indices/thomson-reuters--jefferies-crb-historical-data">https://ru.investing.com/indices/thomson-reuters--jefferies-crb-historical-data</a>
2.5.9	Уровень безработицы, %	259_Unempl	Общий	<a href="https://ru.investing.com/economic-calendar/russian-unemployment-rate-556">https://ru.investing.com/economic-calendar/russian-unemployment-rate-556</a>
2.5.10	Индекс производственной активности PMI России	2510_PMI	Общий	<a href="https://ru.investing.com/economic-calendar/russian-markit-manufacturing-pmi-1630">https://ru.investing.com/economic-calendar/russian-markit-manufacturing-pmi-1630</a>
2.5.11	Фьючерс на нефть Brent	_Brent_oil	Общий	<a href="https://ru.investing.com/commodities/brent-oil-historical-data">https://ru.investing.com/commodities/brent-oil-historical-data</a>

Источник: построено автором на основе данных финансового портала [finam.ru](http://finam.ru)

Все макроэкономические переменные, кроме доходности индекса CRB и фьючерса на нефть Brent доступны только с месячной периодичностью. Чтобы получить ежедневно изменяющиеся данные все эти факторы были линейно интерполированы.

## **2.3 Факторы внимания и настроения на основе данных Twitter и обучение модели тональности текстов**

### **2.3.1 Факторы внимания и настроения**

Для анализа реагирования новостей и сообщений в социальных сетях на волатильность фондового рынка была выбрана социальная сеть Twitter. Парсинг данных осуществлялся с помощью библиотеки Snsrape. Источники данных делятся на две группы:

1. паблики или новостные страницы – брали все посты;
2. пользовательские посты – отбирали посты по ключевым словам,

табл 13.

Таблица 13

Источники данных по постам Twitter

(1)	Новостные паблики	(1.1)	Общий	Финансы и инвестиции	InvestingRu Finam Finam signals RBC invest
		(1.2)	Общий	Новости и события	Meduza RBC Forbes Ria
(2)	Пользовательские посты	(2.1)	Общий	Посты, содержащие ключевые слова «московская биржа», «мосбиржа»	
		(2.2)	Специфичный	Посты, содержащие ключевые слова по компаниям	

Источник: построено автором на основе данных Twitter

Источник: <https://twitter.com/> (дата обращения: 01.05.2021)

В качестве новостных страниц Twitter были выбраны страницы, указанные в табл. 14. Среди каналов получения информации есть как многомиллионные (см. Forbes и Ria), так и каналы с немногочисленной аудиторией. Такой разброс объясняется прежде всего спецификой публикуемых новостей. Некоторые каналы имеют либеральное

направление, некоторые консервативное. Таким образом, удается охватить наиболее широкий круг настроений и интересов.

Таблица 14

Страницы в Twitter, выбранные для проведения исследования

Страница Twitter	Спецификация страницы	Ссылка на страницу	Аудитория
InvestingRu	Финансы и инвестиции	<a href="https://twitter.com/investingru">https://twitter.com/investingru</a>	7,7 тыс.
Finam		<a href="https://twitter.com/finam_blog">https://twitter.com/finam_blog</a>	12.5 тыс.
Finam signals		<a href="https://twitter.com/finamalert">https://twitter.com/finamalert</a>	6.8 тыс.
RBC invest		<a href="https://twitter.com/rbc_quote">https://twitter.com/rbc_quote</a>	31 тыс.
Meduza	Новости и события	<a href="https://twitter.com/meduzaproject">https://twitter.com/meduzaproject</a>	1.3 млн
RBC		<a href="https://twitter.com/ru_rbc">https://twitter.com/ru_rbc</a>	363 тыс.
Forbes		<a href="https://twitter.com/forbes">https://twitter.com/forbes</a>	16.6 млн
Ria		<a href="https://twitter.com/rianru">https://twitter.com/rianru</a>	2.7 млн

Источник: построено автором на основе данных социальной сети Twitter.

Источник: <https://twitter.com/> (дата обращения: 01.05.2021)

Поиск специфичных для компаний постов осуществлялся по следующим словам (в соответствии с компанией, для которой строится модель): «новатэк», «мтс», «газпром», «лукойл», «роснефть», «яндекс», «сбербанк», «сбер».

Настроение и внимание инвесторов – важный параметр, способный повлиять на волатильность той или иной акции. Под настроением инвесторов в работе будет пониматься отношение пользователей социальной сети к фондовому рынку в целом и к определенным компаниям, рассматриваемым в нашем исследовании. Под вниманием инвесторов будет пониматься общая заинтересованность инвестора к фондовому рынку и тому или иному проекту.<sup>32</sup>

---

<sup>32</sup> Индекс страха. Кто и как зарабатывает на нервозности инвесторов. РБК [Электронный ресурс]. Режим доступа:

Таким образом, переменные, построенные на данных социальных сетей, делятся на две группы: переменные внимания и переменные настроения. В то же время, как и экономические факторы, переменные делятся на общие и специфичные. Все данные ежедневные. Все факторы приведены в табл. 15.

Таблица 15

Переменные настроения и внимания. В источнике указан номер источника в соответствии с таблицей «Источники данных по постам Twitter»

Номер	Общий / специфичный	Внимание / настроение	Источник	Наименование фактора	Обозначение	
1	Общий	Внимание	(1.1)	Прирост логарифма количества постов	att_invest	
2			(2.1)	Логарифм количества постов	att_moex	
3		Настроение	(1.1)	Среднее настроение, сглаженное	sent_invest	
4				Стандартное отклонение настроения, сглаженное	std_invest	
5			(1.2)	Среднее настроение, сглаженное	sent_general	
6				Стандартное отклонение настроения, сглаженное	std_general	
7	Специфичный	Внимание	(2.2)	Прирост логарифма количества постов	att_brand	
8		Настроение		Среднее настроение, сглаженное	sent_brand	
9				Стандартное отклонение настроения, сглаженное	std_brand	

Источник: построено автором

Переменные 3-6, 8-9 из таблицы 15 сглаживались следующим образом:

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

Таким образом, учитывалось настроение не только за текущий день, но и за 1 и 2 дня до текущего дня. Для оценки переменных настроения мы

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

### **2.3.2 Обучение модели ULMFiT для определения тональности текста и измерение факторов настроения**

Чтобы измерить факторы настроения нами была обучена модель для прогнозирования тональностей ULMFiT. Подробно модель описана в пункте 1.3.2.

#### **Предобучение языковой модели (LM pre-training)**

В качестве предобученной модели мы взяли готовую модель Russian AWD-LSTM language model, которая обучалась на корпусе Taiga.<sup>33</sup> В корпусе Taiga содержится порядка 5 миллиардов слов и 77% художественных текстов, 19% стихов, 2% новостей, а также тексты взяты из социальных сетей, научной литературы, любительских поэм и проз. Для обучения модели была взята подвыборка текстов, источником которых являются: Arzamas, Fontaka, Interfax, KP, Lenta и NPlus1.<sup>34</sup>

#### **Дообучение языковой модели (LM fine-tuning)**

**Данные.** Дообучение предобученной модели проводилось на корпусе коротких текстов Юлии Рубцовой, находящемся в открытом доступе, на

---

<sup>33</sup> Russian AWD-LSTM language model. [Электронный ресурс]. Режим доступа: <https://github.com/noise-field/Russian-ULMFit> (дата обращения: 04.05.2021)

<sup>34</sup> Shavrina T., Shapovalova O. (2017) To the Methodology of Corpus Construction for Machine Learning: «Taiga» Syntax Tree Corpus and Parser. In proc. Of “CORPORA2017”, international conference , Saint-Petersbourg, 2017.

корпусе SentiRuVal 2016<sup>35</sup>, находящемся в открытом доступе, и на собранных нами данных (мы брали подвыборку всех твитов за весь исследуемый промежуток со страниц Meduza, RBC, Forbes, Ria).

Корпус Юлии Рубцовой состоит из коротких текстов, взятых из социальной сети Twitter. Он содержит базу автоматически размеченных твитов: 114 991 положительных и 111 923 отрицательных твитов за время с конца ноября 2013 года до конца февраля 2014 года.<sup>36</sup> Разметка данных производилась методом, предложенным Jonathon Read, 2005.<sup>37</sup> Тестовая выборка фильтровалась по следующим критериям: удалялись твиты, содержащие одновременно положительные и отрицательные тональности, удалялись дубликаты, удалялись короткие твиты длиной менее 40 символов.

Из корпуса SentiRuVal 2016 были взяты 2000 текстов (источник данных: Twitter), связанных с банками. В данном корпусе используется автоматическая разметка и содержит три класса: позитивные, нейтральные и негативные твиты, мы отобрали только позитивные и негативные текста. Оба корпуса объединили в один.

**Предобработка.** Перед обучением модели мы также провели процедуру предварительной обработки текста:

1. приведение текста к нижнему регистру;
2. замена буквы «ё» на «е»;

---

<sup>35</sup> Рубцова Ю. Автоматическое построение и анализ корпуса коротких текстов (постов микроблогов) для задачи разработки и тренировки тонового классификатора //Инженерия знаний и технологии семантического веба. – 2012. – Т. 1. – С. 109-116.

<sup>36</sup> Лукашевич, Н. В. Автоматический анализ тональности текстов по отношению к заданному объекту и его характеристикам. Электронные библиотеки. - 2015 - 18(3-4), 88-119.

<sup>37</sup> Read J. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. ACLstudent '05: Proceedings of the ACL Student Research WorkshopJune - 2005 - Pages 43–48

3. удаление ссылок на слово;
4. удаление упоминаний пользователей;
5. удаление знаков пунктуации.

Далее выборка балансируется (через метод stratify), чтобы количество позитивных и негативных твитов было одинаково (минимум из двух количеств). Далее выборка делилась на обучающую – 80% и тестовую – 20%, таблица 16.

Таблица 16

Размеры обучающей (слева) и тестовой (справа) выборки, «0» – метка, что твит негативный, «1» – позитивный.

Label		Label	
0	90926	0	22731
1	92490	1	23123

Источник: построено автором

Далее создавался токенизатор (Tokenizer). Токенизатор разбивает поток текста на токены, обычно путем поиска пробелов (табуляции, пробелов, новых строк). Помимо этого, токенизатор предобрабатывает токены согласно документации<sup>38</sup>, помечает слова, с которых начинается текст, слова после которых идет слово с заглавной буквы, слова, которых нет в словаре (в обучающей выборке), слова, которые повторяются в тексте несколько раз и прочее.

**Обучение.** Сначала мы подбирали параметры, функция оптимального значения скорости обучения приведена на рис. 17. Модель обучалась в два этапа:

---

<sup>38</sup> NLP data processing; tokenizes text and creates vocab indexes. [Электронный ресурс]. Режим доступа: <https://fastai1.fast.ai/text.transform.html> (дата обращения: 01.05.2021)

1. Все слои заморожены кроме последнего, проводится обучение в три эпохи. Эпоха – это прохождение всего датасета через нейронную сеть в прямом и обратном направлении один раз. Мы используем ограниченный датасет, чтобы оптимизировать обучение. Делается это с помощью градиентного спуска — итеративного процесса. Поэтому обновления весов после одного прохождения недостаточно. С увеличением числа эпох веса нейронной сети изменяются все большее количество раз. Кривая с каждый разом лучше подстраивается под данные, переходя последовательно из плохо обученного состояния в оптимальный. Если вовремя не остановиться, то может произойти переобучение.

2. Все слои разморожены, проводится обучение в пять эпох. При этом скорость обучения уменьшаем.

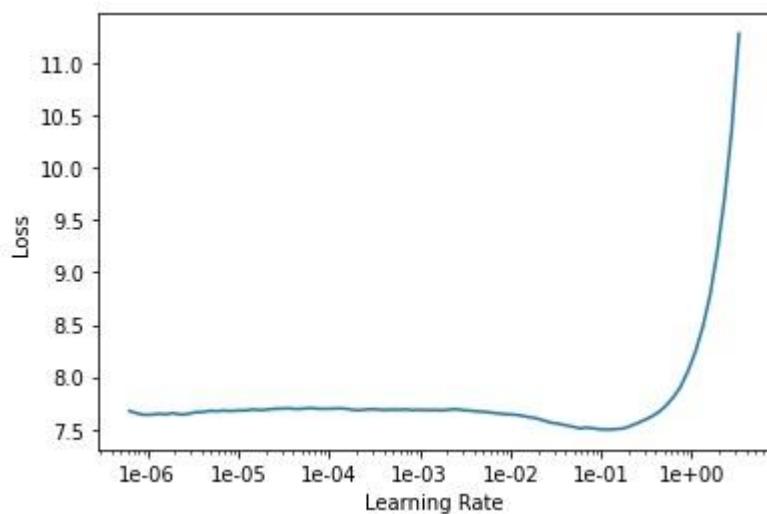


Рис. 17. Оптимальная скорость обучения в зависимости от итерации

На рисунке 17 продемонстрирована работа языковой модели. В темно-сером окне задано начало предложения и количество предсказываемых слов, в светло-сером окне видно, какой текст предсказывает наша модель.

```

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'

learn_lm.predict("акции компании", n_words=30)

'акции компании moderna снизились после одобрения китаем позор на критику google дурова распространение нового оружия способно укрепить военную базу оск в целом включая иран xxbos учёные предлагают способность технологий разделить людей на'

learn_lm.predict("финансирование компании", n_words=20)

'финансирование компании на годы начнет снижаться xxbos пою я думаю что люблю буду жалею что я выиграла xxbos линии московского метро ижевск'

```

Рис. 17. Результаты предсказания языковой модели при заданном начале предложения и количестве предсказываемых слов. Источник:

### Обучение классификатора текста (Classifier fine-tunning)

Обучение модели классификатора проводилось на размеченных данных из предыдущего пункта (на корпусе коротких текстов Юлии Рубцовой и на корпусе SentiRuVal 2016<sup>39</sup>). Выборка была разделена на тестовую – 20%, обучающую –  $0.8 \cdot 0.8 = 64\%$  и валидационную –  $0.8 \cdot 0.2 = 16\%$ . Подвыборки были разделены случайно, при этом соблюдалась балансировка классов (таким же образом, как в предыдущем пункте). Размеры выборок приведены в таблице 17.

Таблица 17

Размеры обучающей (слева), валидационной (в середине) и тестовой (справа) выборки, «0» – метка, что твит негативный, «1» – позитивный.

Label		Label		Label	
<b>0</b>	90926	<b>0</b>	22731	<b>0</b>	18186
<b>1</b>	92490	<b>1</b>	23123	<b>1</b>	18498

Источник: построено автором

<sup>39</sup> Рубцова Ю. Автоматическое построение и анализ корпуса коротких текстов (постов микроблогов) для задачи разработки и тренировки тонового классификатора //Инженерия знаний и технологии семантического веба. – 2012. – Т. 1. – С. 109-116.

Сначала мы подбирали гиперпараметры, функция оптимального значения скорости обучения приведена на рис. 18. Потом мы постепенно размораживали слой за слоем и обучали модель. Таким образом, мы получили итоговый обученный классификатор.

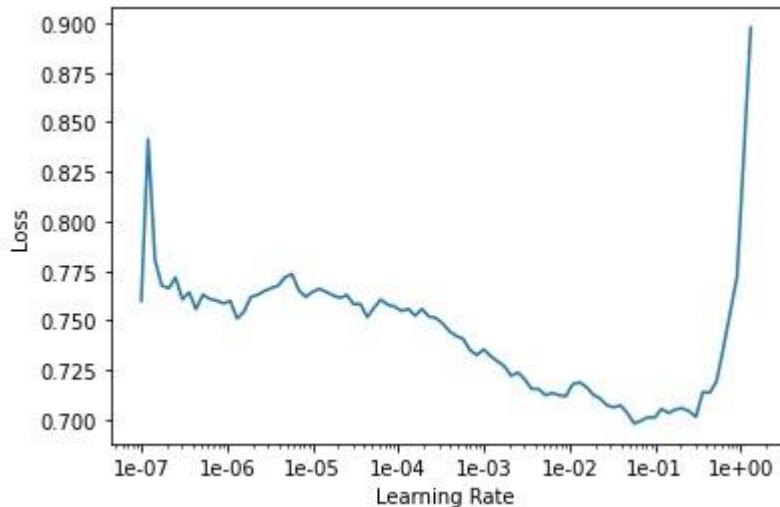


Рис. 18. Оптимальная скорость обучения в зависимости от итерации. Источник:  
построено автором

Чтобы оценить качество классификатора мы использовали отложенную тестовую выборку. Матрица ошибок (Confusion matrix) приведена на рис. 19, по строкам расположены реальные данные, по столбцам – прогнозные. Видно, что точность классификатора примерно равна 75%. Кроме этого, в таблице 18 приведены другие метрики качества: точность, полнота, F1 score. Точность – это доля объектов, принадлежащих классу относительно всех объектов, отнесенных классификатором к данному классу. Полнота – это доля объектов, отнесенная к данному классу классификатором относительно всех объектов класса. Нужно отметить, что в нашем классификаторе точность и полнота примерно одинакова, что является признаком того, что классификатор работает хорошо. С одной стороны, наш классификатор умеет правильно классифицировать примерно в 79% случаев для обоих классов. С другой стороны, наш классификатор

может выявить примерно 79% объектов каждого класса. F1 score учитывает сразу и точность и полноту и выражается следующим образом:

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

В нашем случае все три метрики приблизительно равны и F1 score также показывает неплохое качество классификатора.

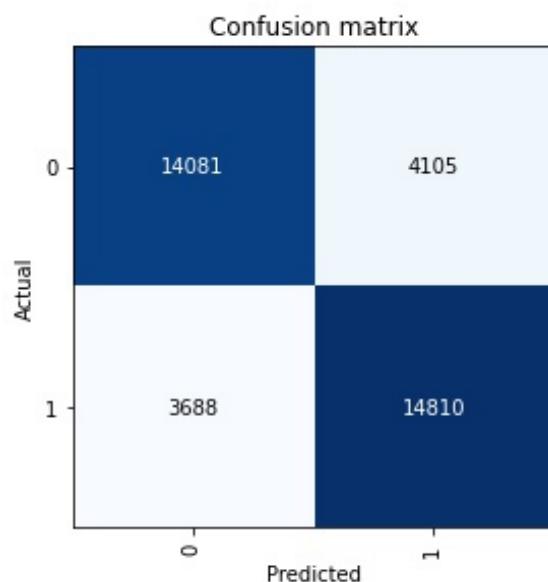


Рис. 19. Матрица ошибок классификатора для данных из тестовой выборки. Источник: построено автором

Таблица 18

Метрики качества: точность, полнота, F1 score для каждого класса по отдельности и их среднее значение

	<b>Точность</b>	<b>Полнота</b>	<b>F1 - score</b>
<b>Negative</b>	0.79245	0.77428	0.78326
<b>Positive</b>	0.78298	0.80063	0.79170
<b>Average</b>	0.78771	0.78745	0.78748

Источник: построено автором

## **Глава 3 Краткосрочное прогнозирование реализованной волатильности**

Для прогнозирования реализованной волатильности мы выбрали за основу гетерогенную авторегрессионную модель (The Heterogeneous Autoregressive model of the Realized Volatility, HAR-RV), далее просто HAR. В исследовании [Аганин А.Д., 2017] проводилось сравнение GARCH, ARFIMA и HAR-RV моделей по качеству одношагового прогноза реализованной волатильности на один день вперед на российском фондовом рынке. Было выявлено, что HAR-RV обладает наибольшей высокой прогнозной силой, для сравнения проводился тест Model Comparison Set (MCS).

HAR модель имеет ряд преимуществ, она способна воспроизводить свойства, присущие волатильности: «длинная память», устойчивость волатильности, «каскад волатильности» (долгосрочная волатильность имеет сильное влияние на краткосрочную волатильность, обратное не выполняется), «толстые хвосты». Также HAR экономически интерпретируема. По сути, наша модель – это авторегрессия первого порядка.

По мимо основной модели мы также построили еще четыре модели. Мы использовали следующие методы машинного обучения: случайный лес, экстремальный градиентный бустинг (XGBoost), расширенный градиентный бустинг (Light GBM).

### **3.1 Теоретические основы используемых методов машинного обучения**

#### **Линейная регрессия**

Линейная регрессия – линейная модель, наиболее часто используемая в эконометрическом анализе в силу ее интерпретируемости. Объясняемая

переменная в ней задается как линейная комбинация объясняющих ее факторов:

$$Y = XW + \varepsilon,$$

-  $Y$  – это вектор размерности  $n$ , где каждая координата – это значение  $i$ -го наблюдения зависимой переменной.  $X$  – это матрица размерности  $n \times m$ , где  $ij$ -тый элемент матрицы – это значение, которое принимает  $j$ -тый фактор в  $i$ -том наблюдении.  $W$  – это веса или коэффициенты, где  $i$ -тая координата показывает степень влияния  $i$ -того фактора.  $\varepsilon$  – вектор размерности  $n$ , где каждая координата – это ненаблюденная ошибка наблюдения. Всего факторов  $m = 1$ .

На модель накладываются условия Гаусса-Маркова. Обычно для поиска весов в качестве функции потерь используют среднеквадратичное отклонение истинных значений от прогнозных:

$$L(X, Y, W) = \frac{1}{2n} \sum_{i=1}^n (y_i - W^T x_i)^2 \rightarrow \min$$

Задача оптимизируется путем взятия первой производной:

$$\frac{\partial L}{\partial W} = \frac{1}{n} (-X^T Y + X^T X W) = 0$$

Тогда оцененные веса могут быть найдены явно и равны:

$$\hat{W} = (X^T X)^{-1} X^T Y$$

В нашей задаче мы используем регрессию лассо (LASSO, Least Absolute Shrinkage and Selection Operator), которая в свою очередь использует регуляризацию L1. Мы предполагаем наличие мультиколлинеарности признаков, вследствие чего может проявляться неустойчивость оценок коэффициентов многомерной линейной регрессии. По мимо этого, наличие большого количества факторов ухудшает предсказательную силу, так как в таком случае модель имеет все шансы на переобучение, рис. 20. Регуляризация L1, обладая особенностью занулять менее значимые признаки, поможет нам автоматически отобрать лучшие факторы, влияющие на целевую переменную. Таким образом, этот метод будет служить как метод понижения размерности.

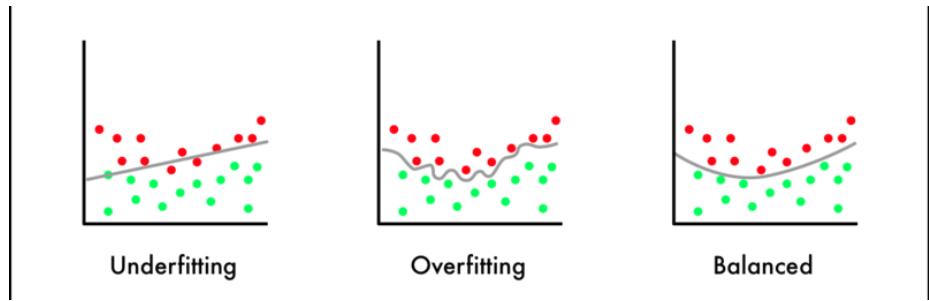


Рис. 20. Слева направо: недообученная модель, переобученная модель, оптимальная модель. Источник: [https://gb.ru/posts/deep\\_learning\\_guide](https://gb.ru/posts/deep_learning_guide)

Регуляризация штрафует за большие веса, увеличивая значение функции потерь. В этом случае функция потерь для слагаемого регуляризации прямо пропорциональна абсолютным значениям весов. Функция потерь выражается следующим образом:

$$L_{lasso}(X, Y, W) = \frac{1}{2n} \sum_{i=1}^n (y_i - W^T x_i)^2 + \lambda ||W||_1 \rightarrow \min,$$

- где  $||W||_1 = \sum_{j=1}^m |w_j|$  – норма весов,  $\lambda$  – коэффициент регуляризации (гиперпараметр), значение которого указывает на то, на сколько сильно мы хотим учитывать штрафующее правило.

Поиск оптимальных параметров осуществляется в неявном виде. Для хорошего приближения мы используем метод градиентного спуска (итеративный метод), где на каждой итерации веса пересчитываются согласно формуле:

$$W_t := W_t - \eta_t \frac{\partial L}{\partial W},$$

$$\eta_t = \frac{k}{t} \text{ (например, шаг можно использовать такой),}$$

- где  $\eta$  – скорость обучения (гиперпараметр), он отвечает за скорость спуска (т.е. за длину шага). То есть постепенно мы передвигаемся в сторону наискорейшего убывания функции (вдоль роста антиградиента функции  $-\frac{\partial L}{\partial W}$ ).

Возникает следующая проблема: все признаки обладают разным масштабом, то есть, например, одни факторы могут принимать значения 1

порядка, когда как другие параметры могут достигать 10 и более порядков. Тогда метод может разойтись, так как мы можем уйти за пределы области, содержащей минимум. Этого можно избежать либо путем выбора очень маленького шага, либо привести все признаки к одному масштабу. Логичнее и менее затратно выбрать второй путь решения. Процесс приведения признаков к одному масштабу называется стандартизацией и приводит все признаки к распределению с нулевым матожиданием и единичной дисперсией:

$$\bar{\mu}_{\cdot j} = \frac{1}{n} \sum_{i=1}^n x_{ij},$$

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

$$X^{new} = \frac{X - \bar{\mu}}{\bar{\sigma}}$$

## Случайный лес

Перед тем как перейти к модели случайного леса, стоит рассмотреть решающее дерево. Решающее дерево – это нелинейная модель. Структура простого решающего дерева для задачи классификации представлена на рис.21:

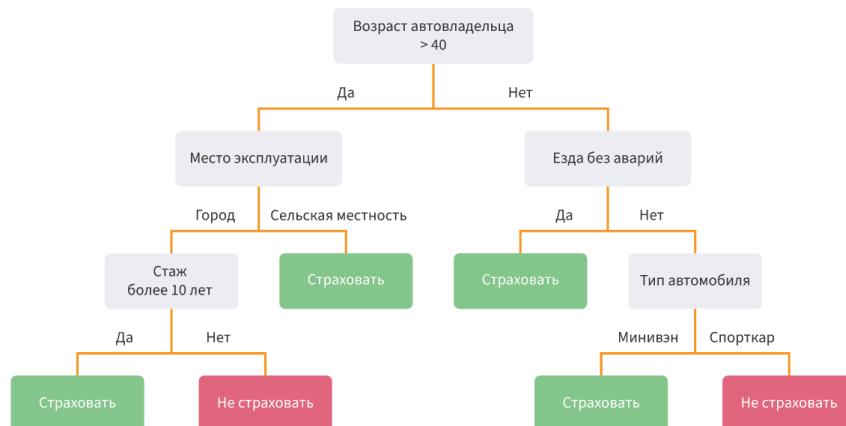


Рис. 21. Структура простого дерева решений, которое определяет, одобрить или нет страховку клиенту по его характеристикам. Источник: <https://loginom.ru/blog/decision-tree-p1>

Здесь узел – это внутренний узел проверки, корневой узел – это начальный узел дерева, лист – конечный узел дерева, решающее правило – это условие в узле. Таким образом, в отличие от узла, в листе содержится не правило, а подмножество объектов, удовлетворяющих всем правилам ветви, которая заканчивается данным листом. Условия в узлах обычно просты и обычно выглядят как сравнение значения признака  $x_j$  с заданным порогом  $t$ :

$$[x_j \leq t].$$

Для задачи регрессии в одномерном признаковом пространстве функция, описывающая данные выглядит как кусочно-постоянная функция (рис. 22.А)). По мере увеличения глубины дерева решающим деревьям свойственно переобучение (рис. 22.Б)).

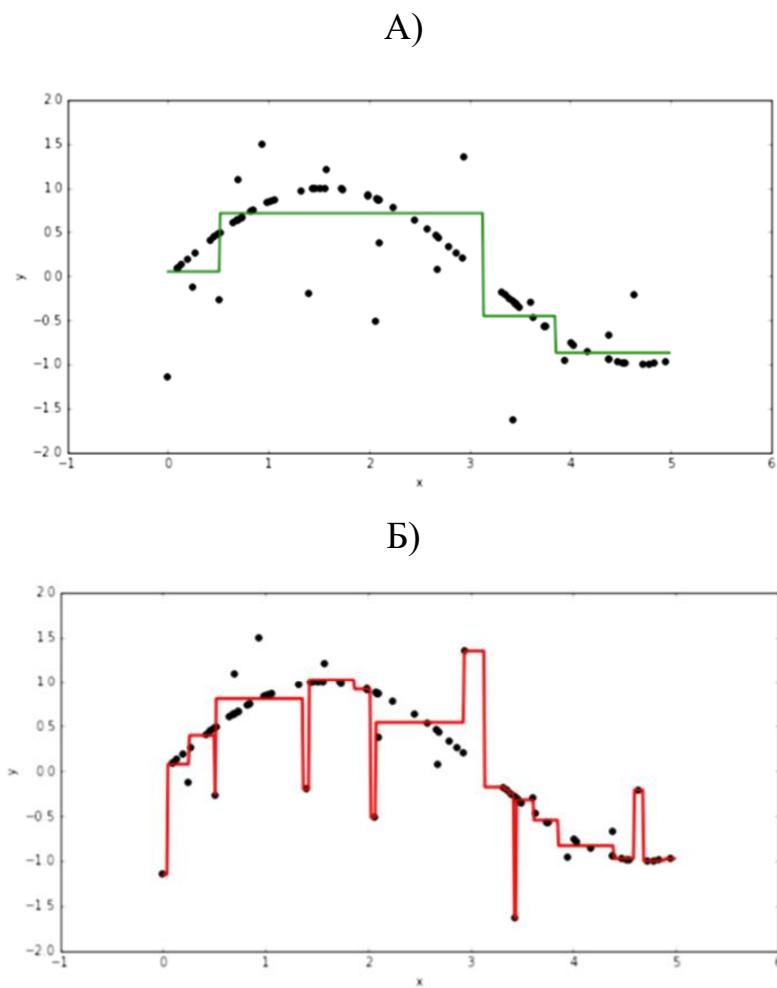


Рис. 22.Решающее дерево в задаче регрессии. Источник: Coursera

Решающее дерево – жадный алгоритм. Для начала выбирается корневой узел, в нем выборка делится на две части, каждая из которых затем далее разбивается еще на две в следующем узле. И так происходит до определенного момента. Параметры в правиле разбиения выбираются таким образом, чтобы свести критерий ошибки к минимуму:

$$Q(X_m, j, t) \rightarrow \min_{j,t}$$

- где  $X_m$  – это множество объектов выборки, попавшее в узел  $t$ . Параметры подбираются методом перебора, ведь множество значений  $j$ , и  $t$  ограничено, причем значений  $t$  столько же, сколько различных значений принимает признак  $x_j$ . Таким образом, множество  $X_m$  делится на два множества:

$$X_l = \{x \in X_m | [x_j \leq t]\} \text{ и } X_r = \{x \in X_m | [x_j > t]\}.$$

По мере дальнейшего разбиения вершин (узлов) наше дерево становится все большей глубины. Очередная вершина объявляется листком в том случае, если выполняется критерий остановки, который может быть, например, таким:

- в вершину попал только один или какое-то кол-во объектов обучающей выборки;
- все объекты выборки принадлежат одному классу (классификация);
- глубина дерева достигла заданного максимума.

Когда лист определен, выбирается значение прогноза, оптимальное для данной подвыборки. Для задачи регрессии и функционала – среднеквадратичной ошибки значением выбирается среднее по подвыборке:

$$a_m = \frac{1}{|X_m|} \sum_{i \in X_m} y_i.$$

В качестве критерия ошибки берем:

$$Q(X_m, j, t) = \frac{|X_l|}{|X_m|} H(X_l) + \frac{|X_r|}{|X_m|} H(X_r),$$

- где  $H(X)$  – это критерий информативности, он тем меньше, чем меньше разброс значений в выборке  $X$ . Для регрессии мерой разброса значений служит дисперсия, тогда:

$$H(X) = \frac{1}{|X|} \sum_{i \in X} (y_i - \bar{y}(X))^2,$$

$$\bar{y} = \frac{1}{|X|} \sum_{i \in X} y_i.$$

Решающие деревья могут очень хорошо описывать обучающие данные и сильно склонны к переобучению, а потому непригодны для прогноза. Но композиция из многих решающих деревьев, т.е. случайный лес – алгоритм, который не переобучается.

Идея рандомного леса состоит в следующем: мы обучаем множество алгоритмов  $b_i$ , далее объединяем их и усредняем, для регрессии ответ выглядит так:

$$a(X) = \frac{1}{N} \sum_{n=1}^N b_n(X)$$

Для того, чтобы  $b_i$  получались разными, необходимо, чтобы обучающие выборки для них отличались. Мы можем применить рандомизацию в выборе подвыборок для обучения алгоритма. В нашем исследовании был использован бутстррап. Суть этого метода состоит в том, что в выборке, длиной  $L$  выбираются  $L$  объектов с возвращением, то есть часть объектов может повторяться в подвыборке, а часть отсутствовать. В бутстрэпированной выборке содержится приблизительно 63% различных объектов первоначальной выборки. Далее поговорим о том, почему случайный лес улучшает качество прогноза.

На тестовых данных ошибка алгоритма содержит в себе три компоненты:

1. шум – характеристика самих данных, данная компонента присутствует всегда, даже если модель идеальна;

2. смещение – отклонение прогноза модели, усредненного по обучающим выборкам, от прогноза идеальной модели;
3. разброс – дисперсия всех ответов, полученных разными моделями (которые обучались на разных выборках).

Для решающих деревьев характерны низкое смещение, т.к. способны воспроизводить трудные закономерности в данных, и большой разброс, т.к. при малых изменениях обучающей выборки решающие деревья сильно изменяются. Для композиции деревьев так же свойственно низкое смещение, поскольку каждый алгоритм по отдельности обладает одинаковым низким смещением, а смещение композиции одинаково со смещением отдельного алгоритма. Разброс композиции отличен от разброса решающего дерева:

$$\left( \begin{array}{c} \text{разброс} \\ \text{композиции} \end{array} \right) = \frac{1}{N} \left( \begin{array}{c} \text{разброс одного} \\ \text{базового алгоритма} \end{array} \right) + \left( \begin{array}{c} \text{корреляция между} \\ \text{базовыми алгоритмами} \end{array} \right).$$

Таким образом, если алгоритмы будут некоррелированными, разброс композиции будет меньше в  $N$  раз. Чтобы уменьшить корреляцию базовых алгоритмов мы используем бэггинг. Бэггинг – рандомизация обучающей выборки, причем, чем подвыборка меньше, тем более независимы базовые алгоритмы.

Алгоритм построения случайного леса из  $N$  решающих деревьев:

1. С помощью бутстрата строим  $N$  случайных подвыборок;
2. Для каждой подвыборки  $\tilde{X}_n$  строим решающее дерево  $b_n$ , соблюдая следующие условия:
  - a. Дерево строится до конца, пока в каждом листе не окажется по одному объекту. Таким образом, получается переобученное дерево с нулевым смещением.
  - b. Выбор признака в узле рандомизирован: признак выбирается из случайного подмножества признаков, причем в каждом узле подмножество признаков разное и выбирается случайно каждый раз;

### 3. Объединяем деревья:

$$a(X) = \frac{1}{N} \sum_{n=1}^N b_n(X).$$

С ростом числа деревьев случайный лес не обучается, а ошибка на тестовой выборке с определенного числа выходит на асимптоту.

Для оценки качества может быть использована подвыборка, не вошедшая в бутстрэпированную подвыборку (67 % от первоначальной выборки). <sup>40</sup> Для каждого наблюдения  $x_i$  из первоначальной выборки вычисляется усредненный прогноз по тем деревьям, в которых это наблюдение отсутствует. Затем полученный прогноз сравнивается с истинным значением. Такой подход называется out-of-bag (OOB), и оценка выглядит так:

$$OOB = \sum_{i=1}^l L\left(y_i, \frac{1}{\sum_{n=1}^N [x_i \notin X_n]} \sum_{n=1}^N [x_i \notin X_n] b_n(x_i)\right).$$

### Экстремальный градиентный бустинг (XGBoost)

Бустинг – композитный алгоритм, в нем алгоритмы строятся последовательно, уменьшая ошибки уже построенной композиции алгоритмов. В качестве алгоритмов мы брали деревья малой глубины. В качестве ошибки мы брали среднеквадратичную ошибку:

$$MSE(a, X) = \frac{1}{l} \sum_{i=1}^l (a(x_i) - y_i)^2.$$

Первым шагом происходит обучение первого алгоритма, решается градиентным спуском:

$$b_1(x) = argmin_b \frac{1}{l} \sum_{i=1}^l (b(x_i) - y_i)^2.$$

---

<sup>40</sup> Композиции: бэггинг, случайный лес. [Электронный ресурс]. Режим доступа: <https://habr.com/ru/company/ods/blog/324402/#sverhsluchaynye-derevya> (дата обращения: 05.05.2021)

Вторым шагом строится алгоритм, композиция которого с первым алгоритмом имеет наименьшую ошибку:

$$b_2(x) = \operatorname{argmin}_b \frac{1}{l} \sum_{i=1}^l (b_1(x_i) + b(x_i) - y_i)^2 = \operatorname{argmin}_b \frac{1}{l} \sum_{i=1}^l (b(x_i) - (y_i - (b_1(x_i)))^2).$$

На шаге  $N$  построение алгоритма задается как:

$$b_N(x) = \operatorname{argmin}_b \frac{1}{l} \sum_{i=1}^l \left( b(x_i) - \left( y_i - \sum_{n=1}^{N-1} b_n(x_i) \right) \right)^2.$$

Процесс останавливается тогда, когда ошибка композиции достигает заданного значения.

### Градиентный бустинг

В градиентном бустинге первый базовый алгоритм  $b_0(x)$  необходимо инициализировать. Можно просто возвращать нуль или среднее значение зависимой переменной. Далее 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  $b_N(x)$ , который минимизирует функционал:

$$\sum_{i=1}^l L(y_i, a_{N-1}(x_i) + b_N(x)) \rightarrow \min_b.$$

Определим, какие значения должен принимать алгоритм  $b_N(x_i) = s_i$ , чтобы ошибка была минимальной.  $s$  называют вектором сдвига и его можно найти по направлению наискорейшего убывания функции потерь:

$$s = \begin{pmatrix} -L'_z(y_1, a_{N-1}(x_1)), \\ \dots \\ -L'_z(y_l, a_{N-1}(x_l)) \end{pmatrix}.$$

Компоненты этого вектора – значения, которые должен принимать алгоритм  $b_N(x)$ . Тогда задача сводится к обучению на размеченных данных, где вектор сдвигов – наша разметка:

$$b_N(x) = \operatorname{argmin}_b \frac{1}{l} \sum_{i=1}^l (b(x_i) - s_i)^2.$$

Этот метод предрасположен к переобучению и качество на тестовой выборке достигает максимума примерно к 10 итерации, далее начинает ухудшаться. Эта проблема разрешается заданием длины шага  $\eta \in (0,1]$  (гиперпараметр) вектора сдвига:

$$a_N(x) = a_{N-1}(x) + \eta b_N(x).$$

Данный гиперпараметр ищется следующим образом: фиксируется число итераций  $N$  и подбирается оптимальный шаг  $\eta$ . Гиперпараметр  $N$  ищется похожим способом, фиксируется шаг  $\eta$  и подбирается оптимальное количество итераций.

## XGBoost

Теперь перейдем к рассмотрению XGBoost, он отличается следующими особенностями:

1. базовый алгоритм приближает направление, посчитанное с учетом вторых производных функции потерь;
2. из функционала, находящего отклонение направление базового алгоритма, удалено деление на вторую производную (уменьшает стоимость вычислений);
3. в функционал добавляется регуляризация, большое количество листьев и норма коэффициентов штрафуются;
4. критерий информативности и критерий остановки зависят от оптимального вектора сдвига.

С одной стороны, как мы рассматривали ранее, пусть на шаге  $N - 1$  уже есть какая-то композиция алгоритмов, тогда на шаге  $N$  мы добавляем еще какой-то алгоритм  $b_N(x)$ :

$$a_N(x) = a_{N-1}(x) + b_N(x),$$

тогда производные функции потерь (сдвиги) равны:

$$s_i = -\frac{\partial L}{\partial z} \Big|_{z=a_{N-1}(x_i)}.$$

Оптимальный алгоритм находится из условия:

$$b_N(x) = \operatorname{argmin}_b \sum_{i=1}^l (b(x_i) - s_i)^2.$$

Раскроем квадрат и уберем последнее слагаемое, т.к. мы минимизируем по  $b$ :

$$\sum_{i=1}^l (b(x_i) - s_i)^2 = \sum_{i=1}^l (b^2(x_i) - 2s_i b(x_i) + s_i^2) = 2 \sum_{i=1}^l \left( \frac{1}{2} b^2(x_i) - s_i b(x_i) \right) \quad (16)$$

С другой стороны, находим оптимальный алгоритм, который должен минимизировать исходную функцию потерь на обучающей выборке:

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

Разложим в ряд Тейлора по второму аргументу в окрестности  $a_{N-1}(x_i)$  и уберем слагаемые, не зависящие от  $b$ :

$$\begin{aligned} & \sum_{i=1}^l L(y_i, a_{N-1}(x_i) + b_N(x_i)) = \\ & = \sum_{i=1}^l \left( L(y_i, a_{N-1}(x_i)) + \frac{\partial L}{\partial z} \Big|_{z=a_{N-1}(x_i)} \cdot b(x_i) + \frac{1}{2} \frac{\partial^2 L}{\partial z^2} \Big|_{z=a_{N-1}(x_i)} \cdot b^2(x_i) \right) = \\ & = \sum_{i=1}^l \left( -s_i \cdot b(x_i) + \frac{1}{2} h_i \cdot b^2(x_i) \right) \end{aligned} \quad (17)$$

Тогда (16) и (17) одинаковы, если положить  $h_i = 1$ . Это предположение о том, что вторые производные константные говорит о том, что функционал одинаково выпуклый, однако, это не совсем так, поэтому обобщим (1) до (2). Тогда задача сходится к следующей:

$$\sum_{i=1}^l \left( -s_i \cdot b(x_i) + \frac{1}{2} h_i \cdot b^2(x_i) \right) \rightarrow \min_b$$

Ранее мы говорили, что мы можем бороться с переобучением посредством уменьшения скорости обучения. Тут опишем альтернативный метод, который заключается в регуляризации. Деревья имеют вид:

$$b_N(x) = \sum_{i=1}^J b_j [x \in R_j],$$

- где  $J$  – это количество регионов, на которое дерево бьет пространство,  $b_j$  – это ответ в регионе,  $[x \in R_j]$  – это индикатор попадания объекта в регион. Тогда имеет смысл, во-первых, ограничивать  $\sum_{i=1}^J b_j^2$  и, во-вторых, количество регионов  $J$ , чтобы уменьшить переобучаемость. Тогда задача усложниться и будет выглядеть:

$$\sum_{i=1}^l \left( -s_i \cdot b(x_i) + \frac{1}{2} h_i \cdot b^2(x_i) \right) + \gamma J + \frac{\lambda}{2} \sum_{i=1}^J b_j^2 \rightarrow \min_b.$$

–  $\gamma$  – значение гиперпараметра показывает, на сколько нежелательно добавлять глубокое дерево,  $\lambda$  – гиперпараметр. Разобъем все суммы на листья:

$$\begin{aligned} & \sum_{j=1}^J \left( -b_j \sum_{x_i \in R_j} s_i + b_j^2 \frac{1}{2} \sum_{x_i \in R_j} h_i + \gamma + \frac{\lambda}{2} b_j^2 \right) = \\ & = \sum_{j=1}^J \left( -b_j \sum_{x_i \in R_j} s_i + b_j^2 \left( \frac{1}{2} \sum_{x_i \in R_j} h_i + \frac{\lambda}{2} \right) + \gamma \right) = \\ & = \sum_{j=1}^J \left( -b_j S_j + \frac{1}{2} b_j^2 (H_j + \lambda) + \gamma \right). \end{aligned} \quad (3)$$

Возьмем первую производную, приравняем к нулю, получим аналитическое решение для оптимальных значений в листьях  $b_j$ :

$$b_j = \frac{S_j}{H_j + \lambda} \quad (18)$$

Тогда ошибка функционала при подстановке (18) в (17) выглядит так:

$$H(b) = -\frac{1}{2} \frac{S_j^2}{H_j + \lambda} + \gamma J$$

$H(b)$  является критерием информативности, тогда если мы хотим разбить вершину  $R$  на  $R_1$  и  $R_2$ , то можно поставить задачу максимизации следующего функционала:

$$Q = H(R) - H(R_1) - H(R_2) \rightarrow \max.$$

### Расширенный градиентный бустинг (Light GBM)

Данный метод основан алгоритмах дерева решений, он разделяет лист дерева с наилучшим соответствием, тогда как другие алгоритмы повышения делят дерево по глубине или уровню, а не по листу (рис. 23). Таким образом, при выращивании на одном и том же листе в Light GBM, листовой алгоритм может уменьшить больше потерь, чем поуровневый алгоритм, и, следовательно, приводит к гораздо лучшей точности, что редко может быть достигнуто любым из существующих алгоритмов повышения.

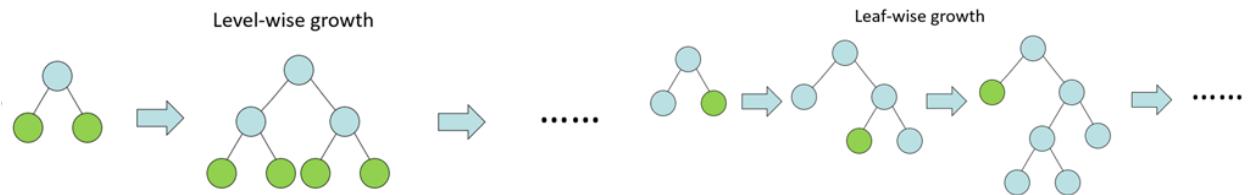


Рис. 23. Схематичная разница между Light GBM и другими алгоритмами, основанными на деревьях решений (источник: <https://www.machinelearningmastery.ru/lightgbm-vs-xgboost-which-algorithm-win-the-race-1ff7dd4917d/>)

Таким образом, дерево разбивается там, где критерий информативности меняется сильнее всего. Дерево может получиться несбалансированным и в этом методе логичнее ограничивать не глубину дерева, а количество листьев.

## 3.2 Обучение базовой и экономической модели и прогнозирование

## **Описание методики подбора гиперпараметров и оценки важности признаков**

В пунктах 3.2 и 3.3 нами рассмотрены три модели, которые различаются набором факторов. Мы имеем базовую модель (бенчмарк), экономическую модель и модель настроения. Первые две модели описаны в этом пункте, так как по сути являются основой для сравнения с третьей моделью. Базовая модель очень проста и включает в себя только первые лаги кратко-, средне- и долгосрочной реализованной волатильности, описание в пункте 2.1, экономическая модель является обобщением базовой и включает в себя также экономические и финансовые признаки, описанные в пункте 2.2. Итоговая модель настроения включает в себя и вышеупомянутые признаки, и признаки, построенные на данных социальной сети Twitter, описание признаков в пункте 2.3.

Для базовой модели мы строили линейную регрессию без использования регуляризации. Для экономической модели и модели настроения мы строили как и лассо регрессию с регуляризацией L1, так и случайный лес, экстремальный градиентный бустинг (XGBoost) и расширенный градиентный бустинг (Light GBM).

Далее мы привели сравнительный анализ моделей и анализ влияния факторов на реализованную волатильность:

1. Сравнение базовой, экономической моделей и модели настроения методом регрессии. Во-первых, мы хотим исследовать, даст ли добавление экономических показателей в базовую модель прирост в качестве прогноза. Во-вторых, мы хотим исследовать, как добавление признаков настроения и внимания улучшат качество прогноза по сравнению с базовой моделью, с экономической моделью.
2. Сравнение экономических моделей и моделей настроения всеми 4-мя методами между собой и друг с другом. Во-первых, мы хотим выяснить оптимальный метод машинного обучения для

экономической модели и модели настроения для каждой компании. Во-вторых, мы хотим проанализировать, в каких методах и для каких компаний модель настроения дает прирост в качестве прогноза в сравнении с экономической моделью. В-третьих, выбрать наилучший метод машинного обучения в пределах одной модели и сравнить лучшую экономическую модель и лучшую модель настроения для каждой отдельной компании.

3. Помимо этого, мы хотим исследовать, какие рассмотренные нами факторы имеют наибольшее влияние на целевую переменную для каждого метода машинного обучения для экономической модели и модели настроения.

В качестве функции потерь мы взяли среднеквадратичную ошибку для всех моделей, так как целевая переменная имеет симметричное распределение.

### **Исследуемый период**

Изначально мы исследовали период в 5 лет с 01.03.2016 по 28.02.2021. Но позже мы сократили его до периода с 01.01.2020 по 28.02.2021. Это связано с рядом причин: во-первых, строя предварительные модели по 5-летнему периоду, было обнаружено, что в 2016–2019 гг. факторы настроения и внимания не оказывали никакого влияния на целевую переменную. Во-вторых, деля выборку на 5 равных подвыборок (для кросс-валидации) и обучаясь на каждой из них по отдельности, модели получались «разными». А значит, степень влияния факторов различна по всему временному горизонту, и обучаться на многолетнем периоде – плохая идея, так как влияние более давних периодов сильно ухудшало прогнозы в поздних периодах. Таким образом, модель переобучалась, получая избыточные данные. В-третьих, качество данных за 2016–2019 гг. сильно хуже, чем за 2020–2021 гг., особенно это заметно для данных социальных сетей. В ранних годах количество постов гораздо меньше, из-за чего

выборка данных за ранние годы вряд ли репрезентативна. К тому же, в экономических и финансовых показателях содержится много пропусков.

## Оптимизация гиперпараметров

Мы использовали метод поиска по сетке для подбора гиперпараметров в лasso регрессии, случайном лесе. А также метод байесовской оптимизации в экстремальном градиентном бустинге и в расширенном градиентном бустниге. Кросс-валидация аналогична и для поиска по сетке, и для байесовской оптимизации.

### 1. Поиск по сетке (Grid Search)

Гиперпараметр модели – это параметр, который устанавливается не во время обучения, а задается из вне. Для поиска оптимального значения таких параметров мы используем поиск по сетке. 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.

### 2. 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. Также метод требует меньшего количества выборок для изучения или получения наилучших значений.<sup>41</sup>

---

<sup>41</sup> Implementing Bayesian Optimization On XGBoost: A Beginner’s Guide. [Электронный ресурс]. Режим доступа: <https://analyticsindiamag.com/implementing-bayesian-optimization-on-xgboost-a-beginners-guide/> (дата обращения: 08.05.2021)

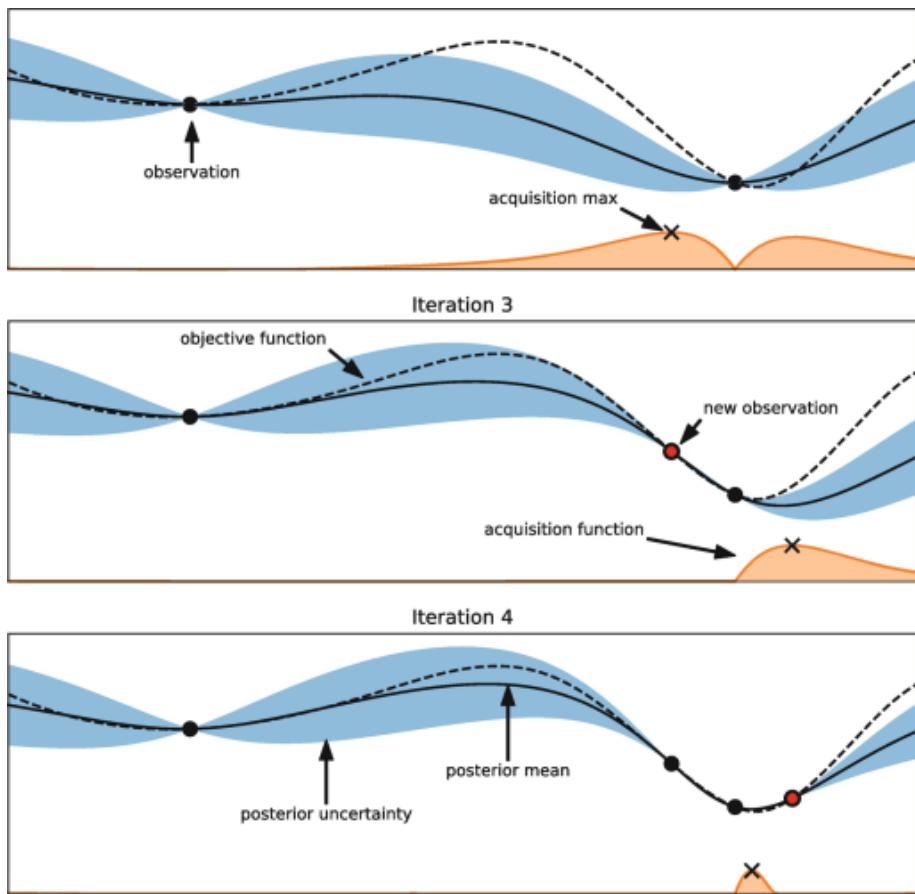


Рис. 24. Принцип работы байесовского оптимизатора. Источник:  
<https://habr.com/ru/company/skillfactory/blog/528240/>

На рисунке 24 осуществляется поиск истинной целевой функции, пунктирная линия. Пусть есть два гиперпараметра, один из них непрерывный. На второй итерации наблюдаются две чёрные точки, устанавливается суррогатная (регрессионная) модель, которая показана чёрной линией. Синяя область вокруг чёрной линии — это неопределенность. Кроме того, есть функция сбора. Эта функция — способ, которым исследуется пространство поиска для нахождения нового оптимального значения наблюдения. Другими словами, функция сбора помогает улучшить суррогатную модель и выбрать следующее значение. На изображении выше функция сбора данных показана в виде оранжевой кривой. Максимальное значение функции сбора означает, что неопределенность максимальна, а прогнозируемое значение невелико.

## Кросс-валидация (Cross Validation)

Для поиска гиперпараметров наша обучающая выборка должна разделяться еще на две: обучающую и валидационную. Мы делаем одинаковую процедуру для всех методов, где подбор гиперпараметра осуществляется через поиск по сетке и делаем разделение обучающей выборки методом кросс-валидации. Учитывая особенности нашего исследования (временные ряды), мы не можем пользоваться обычной перекрестной кросс-валидацией. В перекрестной кросс-валидации обучающая выборка делится на  $k$  количество фолдов (подвыборок, блоков), затем на  $k - 1$  блоках производится обучение модели, а  $-$ й блок используется для тестирования. Процедура повторяется  $k$  раз, при этом на каждом проходе для проверки выбирается новый блок, а обучение производится на оставшихся, рис. 25.

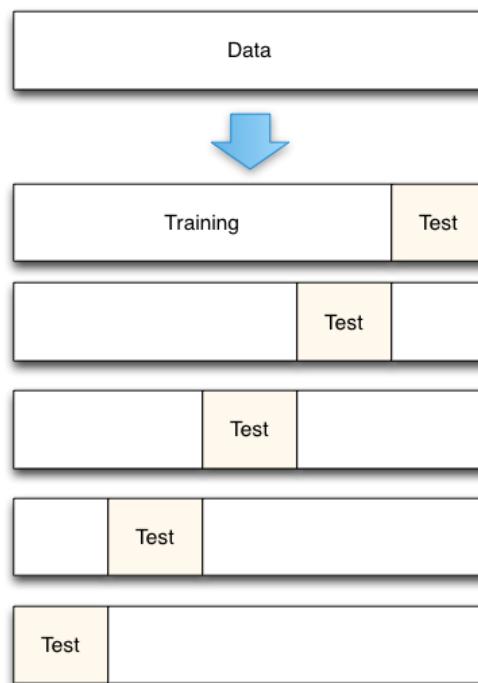


Рис. 25. Перекрестная кросс-валидация на 5 фолдов. Источник: <https://long-short.pro/post/kross-validatsiya-cross-validation-304/>

Для временных рядов такой способ деления не подходит, так как мы не можем использовать будущую информацию для прогноза в прошлом.

Поэтому мы должны брать в обучающую выборку только те данные, которые известны до той даты, на которую мы делаем прогноз.

Итак, мы исследуем период с 01.01.2020 по 28.02.2021. Выборку с 01.01.2021 по 28.02.2021 мы отложили для теста, при обучении модель никогда не видит ее. Далее выборку с 01.01.2020 по 31.12.2020 (250 рабочих дней) мы разделили на 12 фолдов для кросс-валидации, чтобы подобрать оптимальные гиперпараметры для модели. Кросс-валидация происходит следующим образом: обучаем модель на 9 первых фолдах, проводим валидацию на 10 фолде. Далее сдвигаемся вправо на один фолд: обучаем на 2-10 фолдах, валидируем на 11. Итого, мы повторяем процедуру для 3 непересекающихся (на валидации), но связанных выборок. Грубо говоря, мы двигаемся по скользящему окну, каждый раз имея одинаковую по размеру обучающую выборку в 9 фолдов, рис. 26.



Рис. 26. Деление на обучающую и тестовую (отложенную) выборку. Обучающая выборка делилась в свою очередь на 14 фолдов для кросс-валидации, чтобы подобрать оптимальные гиперпараметры. Источник: нарисовано автором

### Сетка для поиска гиперпараметров

Ниже в таблице приведены все гиперпараметры, которые мы подбирали для моделей и сетка для их поиска

Таблица 19

Сетка гиперпараметров

Гиперпараметр	Название	Интервал
<b>Регрессия лассо (Lasso)</b>		
Коэффициент регуляризации	alpha	[0.01, 10], шаг 0.5
<b>Случайный лес (RandomForestRegressor)</b>		

Минимальное количество объектов, необходимое для разделения внутреннего узла	<code>min_samples_split</code>	[2, 5], шаг 1
Минимальное количество объектов, необходимое в листе	<code>min_samples_leaf</code>	[1, 3], шаг 1
Количество признаков, которые следует учитывать при поиске лучшего разделения	<code>max_features</code>	$\left[d, \log_2 d, \frac{d}{3}\right]$ , $d$ – количество факторов
<b>XGBoost (XGBRegressor)</b>		
Количество деревьев с градиентным бустингом	<code>n_estimators</code>	[200, 500]
Максимальная глубина деревьев	<code>max_depth</code>	[3, 15]
Скорость обучения	<code>learning_rate</code>	[0.01, 1], ‘log-uniform’
Отношение размера подвыборки к размеру выборки	<code>subsample</code>	[0.2, 1], ‘uniform’
Минимальное снижение потерь, необходимое для дальнейшего разбиения на листовом узле дерева.	<code>gamma</code>	[1e-9, 0.5], ‘log-uniform’
Отношение столбцов подвыборки при построении каждого дерева.	<code>colsample_bytree</code>	[0.5, 1], ‘uniform’
<b>Light GBM (lgb)</b>		
Количество деревьев с градиентным бустингом	<code>n_estimators</code>	[200, 500]
Максимальная глубина деревьев	<code>max_depth</code>	[3, 15]
Скорость обучения	<code>learning_rate</code>	[0.01, 1], ‘log-uniform’
Отношение размера подвыборки к размеру выборки	<code>subsample</code>	[0.2, 1], ‘uniform’
Отношение столбцов подвыборки при построении каждого дерева.	<code>colsample_bytree</code>	[0.5, 1], ‘uniform’
максимальное количество ячеек, в которые будут помещены значения признаков	<code>max_bin</code>	[100, 1000]
Максимальное количество листьев на одном дереве	<code>num_leaves</code>	[2, 100]

Источник: Построено автором

## Методика оценки важности признаков и их влияния на целевую переменную

Для регрессии лассо мы имеем возможность получить значения весов для признаков. Поскольку все данные были стандартизированы (приведены к одному масштабу), имеет смысл сравнивать важность признаков по значениям их весов.

Для оценки значимости признаков в методах, основанных на решающих деревьях, мы использовали метрику SHAP (Shapley Additive

Explanations).<sup>42</sup> Этот метод считается несколько лучше традиционных методов scikit-learn (обычно тем важнее признак, чем большее количество раз он входит в углы деревьев), потому что многие из этих методов могут быть несостоительными, а это означает, что наиболее важным факторам не всегда может быть присвоен наивысший балл важности. SHAP в моделях на основе дерева может давать двум одинаково важным характеристикам разные оценки в зависимости от того, какой уровень разделения был выполнен с использованием этих факторов. Признаки, которые первыми разделяют модель, могут иметь большее значение. Метод оценивает, насколько важна модель, наблюдая, насколько хорошо модель работает с этим признаком и без него. Важно отметить, что SHAP вычисляет локальную важность признака для каждого наблюдения.

## **Базовая модель**

### **Данные**

Все признаки реализованной волатильности и целевая переменная были прологарифмированы, т.к. имели скошенное распределение, приложение 3.В. Это было сделано для того, чтобы мы могли брать в качестве функции потерь среднеквадратичное отклонение для всех методов машинного обучения. Мы брали везде MSE, в свою очередь, для того, чтобы мы могли сравнивать результаты между собой, применяя разные алгоритмы машинного обучения. Распределение логарифма дневной реализованной волатильности приведено для каждой компании в приложении 3.Д. Следует отметить, что гистограммы показывают симметричное, похожее на нормальное распределение признака, правые хвосты немножко «тяжеловаты» для всех компаний, наиболее часто встречающиеся значения находятся в

---

<sup>42</sup> A Novel Approach to Feature Importance — Shapley Additive Explanations. [Электронный ресурс]. Режим доступа: <https://towardsdatascience.com/a-novel-approach-to-feature-importance-shapley-additive-explanations-d18af30fc21b> (дата обращения: 08.05.2021)

пределах от -9 до -8. Все описательные статистики для всех рассматриваемых тут и далее факторов приведены в приложении 1.

Для базовой модели мы не делали предобработки данных, за исключением того, что все признаки (в т.ч. целевая переменная) были прологарифмированы. Признаки в моделях обозначены следующим образом:

Таблица 20  
Соответствие между старыми и новыми обозначениями для признаков реализованной волатильности

Название переменной	Обозначение	Обозначение, используемое в пункте 2.1
Логарифм реализованной волатильности	log_rv	
Логарифм реализованной волатильности за предыдущий день	log_rv_d	log RV_d
Логарифм реализованной волатильности за предыдущую неделю	log_rv_w	log RV_w
Логарифм реализованной волатильности за предыдущую месяц	log_rv_m	log RV_m

Источник: составлено автором

## Модель

Базовая модель строится на сумме трех AR(1) процессов: реализованная волатильность в день  $t$ , средняя недельная реализованная волатильность с дня  $t-4$  по день  $t$  и средняя месячная реализованная волатильность с дня  $t-21$  по день  $t$  (подробнее в п. 2.1) и выглядит следующим образом:

$$\log RV_{t+1}^{(d)} = c + \beta^{(d)} \log RV_t^{(d)} + \beta^{(w)} \log RV_t^{(w)} + \beta^{(m)} \log RV_t^{(m)} + \varepsilon_{t+1} =$$

$$= \log RV_{t+1}^{(d)} = c + (\log RV_t)' \beta_{RV} + \varepsilon_{t+1} \quad (19)$$

$$\log RV_t^{(w)} = \frac{1}{5} \sum_{i=1}^5 \log RV_{t-i+1}^{(d)}, \quad \log RV_t^{(m)} = \frac{1}{22} \sum_{i=1}^{22} \log RV_{t-i+1}^{(d)},$$

$-\log RV_t^{(d)}$  – оценка реализованной волатильности в день  $t$  методом MedRV,  $\varepsilon_{t+1}$  – ошибка предсказания,  $(\log RV_t)'$  – матрица признаков

волатильности размера  $n \times 3$ ,  $n$  – количество наблюдений,  $\beta_{RV}$  – трехмерный вектор-столбец весов.

Следует отметить, что все три признака для нас важны, поэтому для базовой модели построена обычная линейная регрессия (без регуляризации), чтобы избежать зануления любого из признаков.

## Результаты

Для всех 7 компаний краткосрочная и среднесрочная реализованные волатильности значимы на 1% уровне значимости. Средняя месячная реализованная волатильность – незначимый фактор ни для одной компании даже на 10% уровне значимости. Коэффициенты линейной регрессии для базовой модели приведены в таблице 22. Отметим, что средняя реализованная волатильность за месяц вносит значительно меньший вклад по сравнению с другими признаками. Наблюдается положительная корреляция между реализованной волатильностью за предыдущий день и предыдущую неделю с целевой переменной.

В таблице 21 приведены результаты прогноза модели на тестовой выборке. Данные результаты будем считать нашим бенчмарком.

Таблица 21

Значения среднеквадратичной ошибки MSE для базовой модели на тестовой выборке  
для каждой отдельной компании

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

Источник: построено автором

Таблица 22

Коэффициенты для базовой модели

Компания	intercept	log_rv_d	log_rv_w	log_rv_m

<b>Яндекс</b>	-1.469	0.472	0.300	0.043
<b>Сбербанк</b>	-1.585	0.376	0.510	-0.070
<b>МТС</b>	-1.058	0.354	0.563	-0.037
<b>Лукойл</b>	-1.135	0.496	0.408	-0.041
<b>Роснефть</b>	-1.121	0.395	0.469	0.002
<b>Газпром</b>	-1.352	0.446	0.380	0.016
<b>Новатэк</b>	-0.993	0.436	0.435	0.004

Источник: построено автором

## Экономическая модель

### Данные

Предварительно мы проанализировали все данные для экономических и финансовых показателей и выбрали 9 из них:

Таблица 23

Выбранные показатели для дальнейшего исследования. Соответствие между старыми и новыми обозначениями

Название переменной	Обозначение	Обозначение, используемое в пункте 2.2
Индекс IMOEX	IMOEX	211 IMOEX
Индекс MSCI Russia	MSCI	214 MSCI
Доходность облигаций России 10-летние (годовая ставка), %	Bond_LR	221_OFZ_rate_LR
Доходность облигации Россия годовые (годовая ставка), %	Bond_SR	222_OFZ_rate_SR
Курс доллара к рублю, руб	USD_RUB	231_USD/RUB
Доходность акции, %	yield	215_Return
Объем торгов, руб.	volume_brand	216_Volume
Объем торгов индекса МосБиржи	volume_MOEX	241 IMOEX_vol

Источник: построено автором

Во-первых, мы избавились от переменных, которые сильно коррелированы между собой, чтобы избежать мультиколлинеарности. Хотя, с одной стороны, мы используем регуляризацию, которая бы выбрала автоматически лучшую объясняющую переменную из коллинеарных, однако, в других моделях могут возникнуть проблемы. Во-вторых, мы не

брали в анализ переменные, имеющие много пропусков, так как, если бы мы включили все признаки сразу, осталось бы гораздо меньшее число объектов, не имеющих ни одного пропуска. В-третьих, большое количество факторов ухудшает прогноз, так как модель переобучается. Поэтому число исследуемых факторов было сокращено с 30 до 9.

Все экономические и финансовые показатели численные. Описательная статистика этих показателей, сила корреляции с целевой переменной и графики зависимостей каждого с каждым показателем приведены в приложении 1, 2, 4. В приложении 4 также приведены все распределения экономических и финансовых показателей в виде гистограмм, по которым видно, что в целом распределение признаков нормальное. На графиках зависимостей всех переменных большинство показателей попарно не коррелируют между собой, хотя для некоторых признаков виднеются линейные слабо выраженные зависимости. Таким образом, в итоговую экономическую модель вошли признаки, не имеющие пропусков и у нас нет необходимости в их заполнении.

В приложении 2 приведена таблица корреляций между целевой переменной и всеми факторами для экономической модели и модели настроения. Для удобства, зеленым цветом выделены те признаки, которые хоть в какой-то степени коррелируют с таргетом. Следует отметить, что 4 из 9 экономических показателя имеют корреляцию больше 0,3 для всех 7 компаний. Другие 5 из 9 факторов обладают меньшей корреляцией для всех компаний.

Предобработка данных заключалась в том, что мы сделали стандартизацию масштабов данных, описанную в пункте 3.1.

## **Модель**

Для экономической модели мы используем регрессию лассо. Так как, во-первых, мы хотим автоматически отбирать значимые признаки. Во-вторых, такая модель позволит налагать штрафы на очень маленькие веса.

Мы обобщаем базовую модель, добавляя экономические и финансовые показатели.

$$\log RV_{t+1}^{(d)} = c + (\log RV_t)' \beta_{RV} + E_t' \gamma_{eco} + \varepsilon_{t+1} \quad (20)$$

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

Добавляя регуляризацию, наша задача сводится к следующей оптимизационной задаче:

$$L_{lasso}(X_t, RV_{t+1}^{(d)}, W) = \left( \log RV_{t+1}^{(d)} - c - W' X_t \right)^2 + \lambda ||W||_1 \rightarrow \min_{c, W} \quad (21)$$

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

Во-вторых, мы попытались улучшить нашу экономическую модель, применив другие подходы машинного обучения: случайный лес, XGBoost и Light GBM.

## Результаты

В приложении 5.А приведены все коэффициенты для регрессии лассо. Для удобства цветом отмечены те атрибуты, которые регуляризация отбрала для модели. В среднем для всех компаний занулился только один коэффициент, остальные имеют вес. Некоторые признаки имеют близкий к нулю вес, и, как правило, это переменные, имеющие слабую корреляцию с таргетом. Регуляризация занулила веса следующих факторов для каждой компаний:

1. Яндекс: MSCI, CPI
2. Сбербанк: yield
3. МТС: CPI
4. Gazprom: yield

Для методов, основанных на решающих деревьях, в приложении 6.А приведены графики, в которых отображается информация о значимости признаков. Наиболее важные признаки для экономической модели – это признаки реализованной волатильности, индекс МосБиржи, объем торгов, объем торгов индекса МосБиржи, курс доллара к рублю, иногда в топе появляются доходности кратко- или долгосрочных облигаций.

В таблице 24 приведены результаты прогноза модели на тестовой выборке. Мы можем видеть, что ошибка в среднем стала меньше с добавлением экономических и финансовых показателей.

Таблица 24

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

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

Источник: построено автором

### 3.3 Обучение модели настроения и прогнозирование

#### Модель настроения

## **Данные**

Все показатели внимания и настроения численные. Описательная статистика этих показателей, сила корреляции с целевой переменной и графики зависимостей каждого с каждым показателем приведены в приложении 1, 2, 4. В приложении 4 так же приведены все распределения показателей в виде гистограмм, по которым видно, что в целом распределение признаков нормальное.

В приложении 2 приведена таблица корреляций между целевой переменной и всеми факторами для экономической модели и модели настроения. В среднем 2 из 9 признаков внимания и настроения имеют какую-то значимую корреляцию с целевой переменной. Стоит отметить, что переменная `att_moex` (логарифм количества постов, содержащих упоминание Мосбиржи) имеет корреляцию с таргетом больше 0,3 для 6 из 7 компаний.

Предобработка данных заключалась в том, что мы сделали стандартизацию масштабов данных, описанную в пункте 3.1.

## **Модель**

Во-первых, для того, чтобы сравнить, какой прирост в качестве прогноза дают факторы внимания и настроения, нами была построена лассо регрессия. Модель настроения описывается следующим образом:

$$\log RV_{t+1}^{(d)} = c + (\log RV_t)' \beta_{RV} + E_t' \gamma_{eco} + S_t' \theta_{sent} + \varepsilon_{t+1} \quad (22)$$

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

Добавляя регуляризацию, наша задача сводится к следующей оптимизационной задаче:

$$L_{lasso}\left(X_t, RV_{t+1}^{(d)}, W\right) = \left(\log RV_{t+1}^{(d)} - c - W'X_t\right)^2 + \lambda||W||_1 \rightarrow \min_{c,W} \quad (23)$$

- 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,  $\lambda$  – коэффициент регуляризации (гиперпараметр),  $||W||_1 = \sum_{j=1}^m |w_j|$  – норма весов.

Во-вторых, мы попытались улучшить нашу итоговую модель настроения, применив другие подходы машинного обучения: случайный лес, XGBoost и Light GBM.

## Результаты

В приложении 5.Б приведены все коэффициенты для регрессии лассо. Для удобства цветом отмечены те атрибуты, которые регуляризация отбрала для модели. Стоит отметить, что в сравнении с аналогичной экономической моделью, здесь большее количество признаков было занулено регуляризацией (в среднем по всем компаниям лассо отбрало 4 экономических показателя). Из факторов внимания и настроения регуляризация отбрала в среднем по компаниям 2 признака, очень часто это `att_moex` – логарифм количества постов, содержащих упоминание Мосбиржи и `att_invest` – прирост логарифма количества постов по инвестиционным страницам. Регуляризация отбрала следующие признаки настроения и внимания для каждой компании:

1. Яндекс: `att_invest`
2. Сбербанк: `att_invest`, `att_moex`
3. МТС: `att_invest`, `att_moex`
4. Лукойл: `att_brand`, `att_moex`
5. Роснефть: `att_invest`, `att_brand`, `att_moex`
6. Газпром: `att_invest`, `att_moex`
7. Новатэк: `att_brand`, `att_moex`

Отметим, что ни один из признаков, связанных с тональностью твитов не был выбран регрессией лассо. Все отобранные признаки связаны с изменением внимания (изменение в кол-ве постов).

Однако для моделей, основанных на деревьях, важность признаков по тональности существеннее чем для лассо регрессии. Очень часто std\_invest (стандартное отклонение в настроении по инвестиционным страницам) стоит в топе важности признаков для разных методов и компаний. Визуализация важности признаков на основе SHAP значений для RF, XGB, LGBM в приложении 6.В.

В таблице 25 приведены результаты прогноза модели на тестовой выборке. На тестовой выборке лучшие результаты дали модели Random Forest и XGBoost. Добавление факторов настроения и внимания не дали явный и сильно ощутимый прирост в улучшении прогноза, однако, большинство моделей уменьшило, хоть и в незначительной степени, ошибку на teste.

Таблица 25

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

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

Источник: построено автором

## Сравнение базовой, экономической моделей и модели настроения

В таблице 26 продемонстрировано сравнение базовой и экономических моделей, приведен процентный спад в ошибке на тестовой выборке. Те модели, которые дали лучшее качество прогноза относительно базовой модели выделены цветом.

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 Новатек 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.

Таблица 26  
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. Салатовым цветом отмечены методы, которые дали положительный прирост в качестве прогноза

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

Источник: построено автором

В таблице 27 продемонстрировано сравнение модели настроений и базовой модели. Результаты из таблицы 26 в целом аналогичны результатам из таблицы 27. В среднем, если не брать в учет МТС, модель настроения дает прирост в 15% в качестве прогноза по сравнению с бенчмарком.

Таблица 27

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. Салатовым цветом отмечены методы, которые дали положительный прирост в качестве прогноза

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

Источник: построено автором

Для более детального анализа следует сравнивать экономическую модель и модель настроений, таблица 28. Согласно таблице 28, отметим компанию Яндекс, каждый алгоритм с добавлением признаков по данным социальных сетей улучшил точность прогноза. Для Сбербанка, МТС, Лукойла, Роснефти, Газпрома и Новатэка добавление признаков внимания и настроения улучшил результат для 3 из 4 моделей. В среднем для всех компаний добавление признаков настроения и внимания к экономической модели улучшают качество прогноза на 3%.

Таблица 28

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. Салатовым цветом отмечены методы, которые дали положительный прирост в качестве прогноза

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

Источник: построено автором

Мы отобрали лучшую экономическую модель и лучшую модель настроения для дальнейшего сравнения. Лучшие алгоритмы прогнозирования для экономической модели и модели настроения, их ошибки на прогнозе и прирост в улучшении прогноза для модели настроения относительно экономической модели приведены в таблице 29.

Таблица 29

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).

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

Источник: построено автором

На основе таблицы 29 можно сделать следующий вывод: социальные сети в целом положительно влияют на точность прогноза. Уменьшение в MSE на teste незначительное (в среднем 3%), но все-таки присутствует. Лучшие результаты у компаний Яндекс и Сбербанк, что значит, признаки внимания и настроения оказывают более сильное влияние на них, чем на другие рассмотренные компании. Сравнение результатов по лучшим алгоритмам модели настроений с базовой в таблице 30.

Таблица 30

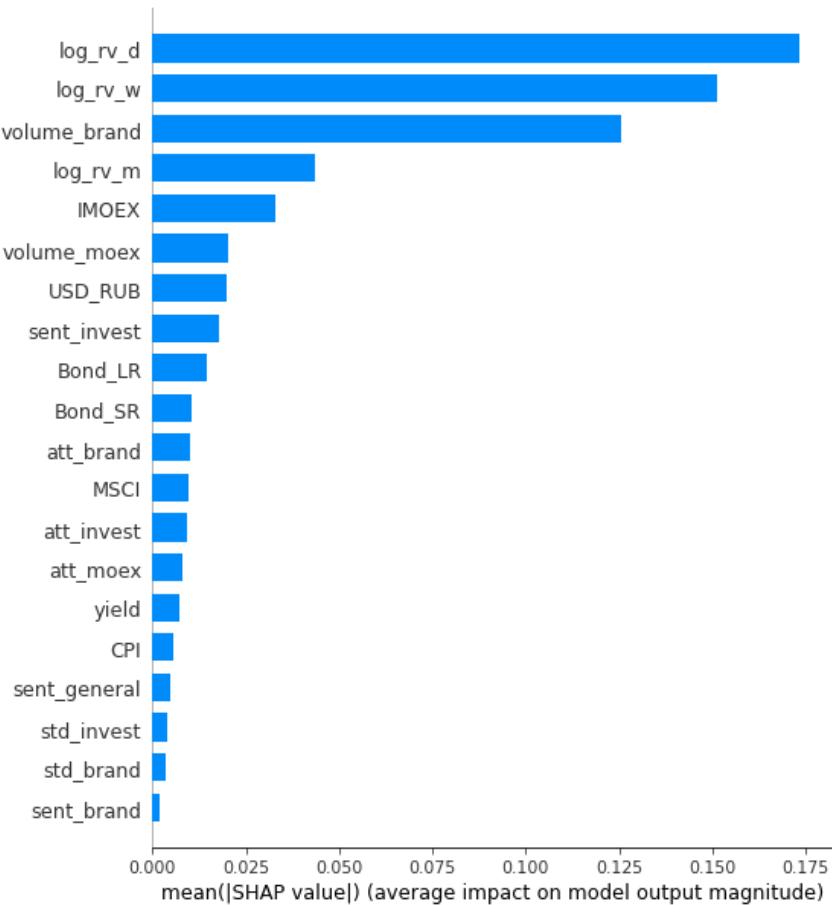
Уменьшение среднеквадратичной ошибки MSE на тестовой выборке в модели настроения относительно базовой модели, %

Компания	Уменьшение MSE, %
Яндекс	64%

<b>Сбербанк</b>	17%
<b>МТС</b>	-19%
<b>Лукойл</b>	12%
<b>Роснефть</b>	7%
<b>Газпром</b>	15%
<b>Новатэк</b>	12%

Источник: построено автором

Из таблицы 30 видно, что факторы внимания и настроения большее влияние оказывают на реализованную волатильность для компании Яндекс (IT), рассмотрим ее отдельно. Проанализируем важность признаков XGBoost для Яндекса, рисунок 27. Больший вклад вносят признаки `log_rv_d`, `log_rv_w` и `volume_brand` (объем торгов). Самые важные признаки по данным социальных сетей – `sent_invest` (тональность твитов на инвестиционных страницах) и `att_brand` (изменение в количестве твитов, содержащих упоминание компании), причем они занимают достаточно высокое место (5 и 8 соответственно).



**Рис. 27.** Важность признаков в модели настроения, построенной методом XGBoost для компании Яндекс. Чем больше бар соответствующего признака, тем большее значение он имеет. Источник: построено автором

В целом можно сделать вывод, что для компании Яндекс информация о настроении и внимании в социальных сетях имеет большую объясняющую силу. И факторы внимания и настроения улучшают качество прогноза в среднем по всем методам на 7% относительно экономической модели. Далее идут Сбербанк и МТС, для которых при выборе оптимального метода модель настроения улучшает качество прогноза на 3% в сравнении с экономической моделью. Компании, связанные с нефтедобычей и газодобычей хуже всего реагируют на факторы социальных сетей, и даже при лучшем методе дают всего прирост качества прогноза в 2% в среднем.

## **Заключение**

В ходе работы над исследованием удалось прийти к следующим выводам:

- Внимание и настроение инвесторов оказывают влияние на реализованную волатильность. Комментарии в социальных сетях также оказывают значимое влияние на акции компаний. При включении весов внимания и настроения в базовую модель с экономическими весами, качество модели растёт.

- Среди подходов измерения внимания и настроения инвесторов выделяют методы, основанные на:

- Различные показатели рынка - объем, VIX, спред TED и другие;
- Количественные и качественные опросы;
- Поисковые запросы в Интернете;
- Неэкономические факторы;
- Данные из социальных сетей и мессенджеров.

- 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.

- Существует множество способов проанализировать тональность новостей. Все методы можно классифицировать по разным признакам. В данном исследовании методы были классифицированы по 2 признакам: 1) по доле участия человека в процессе обработки: ручные и автоматические методы; 2) по источнику формирования метода: методы, основанные на «словарях тональности языка» (WordNet-Affect, SentiWordNet, SenticNet) и методы, основанные на принципе естественного языка (NLP). Обработка естественного языка состоит из понимания лингвистики и методов машинного обучения. Безусловно, существуют и другие способы классификации, но в данной работе были выбраны основные, именно они и применялись для изучения.

- Обработка естественного языка состоит из понимания лингвистики и методов машинного обучения. Для анализа естественного языка необходима предобработка текста. Данный этап разбивается на подоперации: токенизация, нормализация слова, уменьшение размера словаря (стемминг и лемматизация), компьютерное представление слов. Можно выделить три основных классических способа представления слов – One-hot encoding, TF-IDF и Embedding.

- В работе используются несколько вариантов применения нейронных сетей (например, рекуррентные нейронные сети (Recurrent Neural Network, RNN), трансферное обучение и др.).

- Для исследования волатильности оптимальнее всего брать реализованную волатильность. Выбор реализованной волатильность

определен тем, что в течение дня динамика флуктуаций цен акций может иметь несистемный характер. Для проведения анализа необходимо рассмотреть различные вариации реализованной волатильности (с разными лагами: день, неделя, месяц).

- В данной работе были рассмотрены 7 различных российских компаний: Яндекс, Сбербанк, МТС, Лукойл, Роснефть, Газпром, Новатэк. Все компании котируются на Московской фондовой бирже. Выбор этих компаний мотивирован тем фактом, что переменные настроения и внимания могут по-разному влиять на будущую реализованную волатильность в зависимости от типа рассматриваемой акции.

- В свою очередь, были проанализировано несколько групп показателей - экономических и финансовых. Все показатели были разделены на 5 больших групп: 1) переменные фондового рынка; 2) переменные рынка облигаций; 3) переменные обменного курса; 4) переменные ликвидности; 5) макроэкономические переменные. В свою очередь все переменные разделены на две другие большие группы: 1) общие факторы – одинаковы для всех компаний; 2) специфичные факторы – отличны и характерны для каждой определенной компании.

- В работе измерялись сигналы настроения и внимания, связанные с отдельными акциями и индексом фондового рынка, используются текстовые данные из двух социальных сетей (Twitter и StockTwits), статьи финансовых новостей, полученные с помощью RavenPack News Analytics, а также данные с поисковых систем (Google Trends). StockTwits - это платформа социальных сетей, где пользователи делятся информацией о рынке и отдельных акциях в форме коротких сообщений.

Полученные результаты могут быть применены в различных областях:

1) Государственное и муниципальное управление. Принципы оценки тональности новостей можно использовать для оценки отношения народа к

тому или иному государственному проекту. Это сократит затраты на проведение дорогостоящих опросов населения.

2) Управление бизнесом/менеджмент организации. На сегодняшний день в сферах управления остро стоит вопрос об эмоциональном и социальном интеллекте. Анализ тональности сообщений сотрудников (например, в онлайн-журнале отзывов о компании) позволит быстро отсортировать потоки информации и принимать эффективные решения. Отдельные результаты могут оказаться полезными и для конкретных секторов бизнеса, например, при анализе отзывов в сети о новом фильме.

3) Инвестиции.

4) Анализ данных. В работе предложены методологии анализа баз данных, которые могут заинтересовать специалистов в данной сфере.

5) Международные и отечественные СМИ. Средства массовой информации уже используют методики оценки тональности новостей.<sup>43</sup> Для анализа используется метод «мешка слов».

6) Маркетинг. Анализ настроений клиентов в сфере маркетинга позволяет с меньшими затратами оценить отношение клиента к продукту или к новым видам услуг и т.п.

Несмотря на качество полученных результатов, можно выделить несколько слабых мест, которые, в свою очередь, ставят задачу дополнительных исследований и совершенствований перед другими исследователями:

1) Можно проверить другие модели, которые могут показать более высокие значения качества.

---

<sup>43</sup> Афанасьев Д.О., Федорова Е.А., Рогов О.Ю. О влиянии тональности новостей в международных СМИ на рыночный курс российского рубля: текстовый анализ. <https://ej.hse.ru/data/2019/06/19/1485345307/%D0%A4%D0%B5%D0%B4%D0%BE%D1%80%D0%BE%D0%B2%D0%B0.pdf>

2) Найти размеченные данные по финансовым новостям, проанализировать их, сравнить с исходной моделью.

3) Можно проверить другие российские компании, сравнить их показатели с результатами работы.

Таким образом, данная работа может представлять интерес для разных групп исследователей.

## Список литературы

1. Аганин А.Д. Сравнение GARCH и HAR-RV моделей для прогноза реализованной волатильности на российском рынке. Прикладная эконометрика, 2017, т. 48, с. 63–84
2. Афанасьев Д.О., Федорова Е.А., Рогов О.Ю. О влиянии тональности новостей в международных СМИ на рыночный курс российского рубля: текстовый анализ. Экономический журнал ВШЭ. 2019. Т. 23. № 2. С. 264–289. HSE Economic Journal, 2019, vol. 23, no 2, pp. 264–289. <https://ej.hse.ru/data/2019/06/19/1485345307/Федорова.pdf>
3. Лукашевич, Н. В. Автоматический анализ тональности текстов по отношению к заданному объекту и его характеристикам. Электронные библиотеки. - 2015 - 18(3-4), 88-119.
4. Ульянкин Ф. Прогнозирование российских макроэкономических показателей на основе информации в новостях и поисковых запросах. Российская академия народного хозяйства и государственной службы (РАНХиГС)
5. Antweiler, W., & Frank, M. Z. (2004). Is all that talk just noise? The information content of internet stock message boards. *The Journal of Finance*, 59(3), 1259–1294.
6. Aouadi, A., Arouri, M., & Teulon, F. (2013). Investor attention and stock market activity: Evidence from France. *Economic Modelling*, 35, 674–681.
7. Barndorff-Nielsen, O. E., & Shephard, N. (2002). Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 64(2), 253–280.
8. Behrendt, S., & Schmidt, A. (2018). The Twitter myth revisited: Intraday investor sentiment, Twitter activity and individual-level stock return volatility. *Journal of Banking & Finance*, 96, 355–367.

9. Benjamin Snyder, Regina Barzilay. Multiple Aspect Ranking using the Good Grief Algorithm (англ.) // Proceedings of the Joint Human Language Technology/North American Chapter of the ACL Conference (HLT-NAACL) : конференция. — 2007. — P. 300–307.
10. Bo Pang, Lillian Lee. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts (англ.) // Proceedings of the Association for Computational Linguistics (ACL) : журнал. — 2004. — P. 271–278.
11. 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
12. Daniel, K., Hirshleifer, D., & Teoh, S. H. (2002). Investor psychology in capital markets: evidence and policy implications. *Journal of Monetary Economics*, 49(1), 139–209.
13. Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2), 383–417.
14. Mao, H., Counts, S., & Bollen, J. (2011). Predicting financial markets: Comparing survey, news, Twitter and search engine data.
15. Tseng, K. C. (2006). Behavioral finance, bounded rationality, neurofinance, and traditional finance. *Investment Management and Financial Innovations*, 3(4), 7–18.
16. John Hebel, Matthew Fisher, Ryan Blace, Andrew Perez-Lopez. Semantic Web Programming. — John Wiley & Sons, 2009. — 648 c. — ISBN 9780470418017.
17. Johnson, E. J., & Tversky, A. (1983). Affect, generalization, and the perception of risk. *Journal of Personality and Social Psychology*, 45(1), 20–31.
18. Oliveira, N., Cortez, P., & Areal, N. (2013). On the predictability of stock market behavior using StockTwits sentiment and posting volume. In L.

- Correia, L. P. Reis, & J. Cascalho (Eds.), Progress in artificial intelligence (pp. 355–365). Berlin, Heidelberg: Springer Berlin Heidelberg.
19. Peter Turney. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews (англ.) // Proceedings of the Association for Computational Linguistics. — 2002. — P. 417–424. — arXiv:cs.LG/0212032.
20. Read J. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. ACLstudent '05: Proceedings of the ACL Student Research WorkshopJune - 2005 - Pages 43–48
21. Renault, T. (2017). Intraday online investor sentiment and return patterns in the U.S. stock market. Journal of Banking & Finance, 84, 25–40.
22. Schoen, H., Gayo-Avello, D., Metaxas, P. T., Mustafaraj, E., Strohmaier, M., & Gloor, P. (2013). The power of prediction with social media. Internet Research, 23(5), 528–543.
23. Stefano Baccianella. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining // Proceedings of LREC : конференция. — 2010. — P. 2200–2204.
24. Strapparava S. and Valitutti A. Wordnet-affect: an affective extension of wordnet. In Proceedings of the 4th International Conference on Language Resources and Evaluation, Lisbon, 2004.
25. Sun, L., Najand, M., & Shen, J. (2016). Stock return predictability and investor sentiment: A high-frequency perspective. Journal of Banking & Finance, 73, 147–164.
26. Tumarkin, R., & Whitelaw, R. F. (2001). News or noise? Internet postings and stock prices. Financial Analysts Journal, 57(3), 41–51.
27. Wang, S., & Cui, H. (2013). Generalized F test for high dimensional linear regression coefficients. Journal of Multivariate Analysis, 117, 134–149.
28. Ulyankin, F. (2020). Forecasting Russian Macroeconomic Indicators Based on Information from News and Search Queries. Russian Journal of Money and Finance, 79(4), pp. 75–97. doi: 10.31477/rjmf.202004.75

## **Электронные ресурсы:**

29. Governments have identified commodities essential to economic and military security. [Электронный ресурс]. Режим доступа: <https://www.economist.com/finance-and-economics/2021/03/31/governments-have-identified-commodities-essential-to-economic-and-military-security> (дата обращения: 03.04.2021)
30. Академический онлайн-словарь. Определение «социальной сети». [Электронный ресурс]. Режим доступа: <https://dic.academic.ru/dic.nsf/ruwiki/60759> (дата обращения: 04.05.2021)
31. Анализ тональности в русскоязычных текстах. [Электронный ресурс]. Режим доступа: <https://habr.com/ru/company/mailru/blog/516214/> (дата обращения: 30.03.2021)
32. Аналитический портала Wordstat Yandex. [Электронный ресурс]. Режим доступа: <https://wordstat.yandex.ru/>
33. Данные об акциях компании «X5 retail group». [Электронный ресурс]. Режим доступа: <https://www.finam.ru/profile/raspiski/x5-retail-group/export/> (дата обращения: 27.04.2021)
34. Данные об акциях компании «Газпром». [Электронный ресурс]. Режим доступа: <https://www.finam.ru/profile/moex-akcii/gazprom/export/> (дата обращения: 26.04.2021)
35. Данные об акциях компании «Лукойл». [Электронный ресурс]. Режим доступа: <https://www.finam.ru/profile/moex-akcii/lukoil/export/> (дата обращения: 26.04.2021)
36. Данные об акциях компании «МТС». [Электронный ресурс]. Режим доступа: <https://www.finam.ru/profile/moex-akcii/mts/export/> (дата обращения: 26.04.2021)

37. Данные об акциях компании «Новатэк». [Электронный ресурс]. Режим доступа: <https://www.finam.ru/profile/moex-akcii/novatek/export/> (дата обращения: 26.04.2021)
- 38.Данные об акциях компании «Роснефть». [Электронный ресурс]. Режим доступа: <https://www.finam.ru/profile/moex-akcii/rosneft/export/> (дата обращения: 26.04.2021)
39. Данные об акциях компании «Сбербанк». [Электронный ресурс]. Режим доступа: <https://www.finam.ru/profile/moex-akcii/sberbank/export/> (дата обращения: 26.04.2021)
40. Данные об акциях компании «Яндекс». [Электронный ресурс]. Режим доступа: <https://www.finam.ru/profile/moex-akcii/pllc-yandex-nv/export/> (дата обращения: 26.04.2021)
- 41.Индекс потребительских цен (ИПЦ). [Электронный ресурс]. Режим доступа: <https://ru.investing.com/economic-calendar/russian-cpi-1180> (дата обращения: 26.04.2021)
- 42.Индекс производственной активности PMI России. [Электронный ресурс]. Режим доступа: <https://ru.investing.com/economic-calendar/russian-markit-manufacturing-pmi-1630> (дата обращения: 26.04.2021)
- 43.Денежная масса, млрд рублей. [Электронный ресурс]. Режим доступа: [https://www.cbr.ru/statistics/macro\\_itm/dkfs/](https://www.cbr.ru/statistics/macro_itm/dkfs/) (дата обращения: 26.04.2021)
- 44.Денежная масса сезонно скорректированная, млрд рублей. [Электронный ресурс]. Режим доступа: [https://www.cbr.ru/statistics/macro\\_itm/dkfs/](https://www.cbr.ru/statistics/macro_itm/dkfs/) (дата обращения: 26.04.2021)
- 45.Доходность индекса CRB. [Электронный ресурс]. Режим доступа: <https://ru.investing.com/indices/thomson-reuters---jefferies-crb-historical-data> (дата обращения: 26.04.2021)

46. ЛИБОР // Лас-Тунас — Ломонос. — М. : Большая российская энциклопедия, 2010. — С. 391. — (Большая российская энциклопедия : [в 35 т.] / гл. ред. Ю. С. Осипов ; 2004—2017, т. 17). — ISBN 978-5-85270-350-7.
47. Объём промышленного производства в России. [Электронный ресурс]. Режим доступа: <https://ru.investing.com/economic-calendar/russian-industrial-production-553> (дата обращения: 26.04.2021)
48. Ожидаемый индекс потребительских цен (ИПЦ). [Электронный ресурс]. Режим доступа: <https://ru.investing.com/economic-calendar/russian-cpi-1180> (дата обращения: 26.04.2021)
49. Первая разность денежной массы, млрд рублей. [Электронный ресурс]. Режим доступа: [https://www.cbr.ru/statistics/macro\\_itm/dkfs/](https://www.cbr.ru/statistics/macro_itm/dkfs/) (дата обращения: 26.04.2021)
50. Первая разность денежной массы сезонно скорректированной, млрд рублей. [Электронный ресурс]. Режим доступа: [https://www.cbr.ru/statistics/macro\\_itm/dkfs/](https://www.cbr.ru/statistics/macro_itm/dkfs/) (дата обращения: 26.04.2021)
51. Рубцова Ю. Автоматическое построение и анализ корпуса коротких текстов (постов микроблогов) для задачи разработки и тренировки тонового классификатора //Инженерия знаний и технологии семантического веба. – 2012. – Т. 1. – С. 109-116.
52. Саиф Х., Хе Я. и Алани Х., «SemanticSentimentAnalysis of Twitter», материалы семинара по извлечению информации и аналитике сущностей в данных социальных сетей. Соединенное Королевство: Институт СМИ, 2011.
53. Уровень безработицы, %. [Электронный ресурс]. Режим доступа: <https://ru.investing.com/economic-calendar/russian-unemployment-rate-556> (дата обращения: 26.04.2021)

54. Фьючерс на нефть Brent. [Электронный ресурс]. Режим доступа:  
<https://ru.investing.com/commodities/brent-oil-historical-data> (дата обращения: 26.04.2021)

## **Приложения**

### **ПРИЛОЖЕНИЕ 1**

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.

log\_rv\_d – логарифм дневной реализованной волатильности, log\_rv\_w – логарифм недельной реализованной волатильности, log\_rv\_m – логарифм месячной реализованной волатильности, IMOEX – индекс МосБиржи, MSCI - индекс MSCI Russia, Bond\_SR - доходность облигации Россия годовые (годовая ставка), Bond\_LR - доходность облигации Россия 10-летние (годовая ставка), USD\_RUB - курс доллара к рублю, CPI - индекс потребительских цен (ИПЦ), sent\_invest – среднее настроение по инвестиционным страницам, att\_invest – прирост логарифма количества постов по инвестиционным страницам, std\_invest – стандартное отклонение в настроении по инвестиционным страницам, sent\_general – среднее настроение по новостным страницам, std\_general – стандартное отклонение в настроении по новостным страницам, sent\_brand – среднее настроение по постам, содержащих имя компании, std\_brand – стандартное отклонение в настроении в постах, содержащих имя компании, att\_brand – прирост логарифма количества постов, содержащих имя компании, Yield – доходность компании, Volume\_brand – объем торгов компании, Volume\_MOEX – объем торгов индекса МосБиржи, att\_moex – логарифм количества постов, содержащих упоминание Мосбиржи.

### Компания «Яндекс»

	<b>log_r v_d</b>	<b>log_r v_w</b>	<b>log_r v_m</b>	<b>IMO EX</b>	<b>MSC I</b>	<b>Bond _SR</b>	<b>Bond _LR</b>	<b>USD_ RUB</b>	<b>CPI</b>	<b>sent_i nvest</b>	<b>att_in vest</b>	<b>std_i nvest</b>	<b>sent_ger eral</b>	<b>std_g enera l</b>	<b>sent_b rand</b>	<b>std_b rand</b>	<b>att_b rand</b>	<b>Yield</b>	<b>Volu me_b rand</b>	<b>Volu me_MOE X</b>	<b>att_m oex</b>
<b>count</b>	278	278	278	278	278	278	278	278	278	278	278	278	278	278	278	278	278	278	278	278	278
<b>mean</b>	-7,93	-7,93	-7,96	0,01	0	0,03	0,04	0,01	0	0,53	2,36	0,48	0,6	0,49	0,55	0,5	4,76	0,16	2612,7 1	13155, 94	0,4
<b>std</b>	0,78	0,63	0,51	0,01	0,03	2,97	1,52	0,01	0,01	0,09	1,15	0,04	0,03	0,01	0,04	0,01	0,29	2,8	1720,9 5	8658,8	0,55
<b>min</b>	-10,19	-9,1	-8,94	0	-0,15	-12,07	-5,88	0	-0,05	0,25	0	0,14	0,52	0,46	0,41	0,47	4,08	-13,34	370,63	961,46	0
<b>25%</b>	-8,38	-8,3	-8,26	0	-0,01	-1,3	-0,49	0	0	0,45	2,56	0,47	0,58	0,48	0,52	0,5	4,62	-1,34	1612,5	7765	0
<b>50%</b>	-8,04	-8,02	-8,02	0,01	0	-0,18	0	0,01	0	0,53	2,83	0,49	0,6	0,49	0,55	0,5	4,74	0,39	2225	11435	0
<b>75%</b>	-7,58	-7,72	-7,73	0,01	0,01	1,01	0,5	0,01	0	0,58	3,04	0,5	0,63	0,49	0,57	0,5	4,9	1,77	3200	15492, 5	0,69
<b>max</b>	-5,16	-5,7	-6,32	0,09	0,13	16,12	10,12	0,07	0,06	0,82	3,5	0,52	0,7	0,5	0,66	0,5	6,3	11,8	19730	69820	2,71

Источник: построено автором на основе данных компаний «Яндекс»

### Компания «Сбербанк»

	<b>log_r v_d</b>	<b>log_r v_w</b>	<b>log_r v_m</b>	<b>IMO EX</b>	<b>MSC I</b>	<b>Bond _SR</b>	<b>Bond _LR</b>	<b>USD_RU B</b>	<b>CPI</b>	<b>sent_i nvest</b>	<b>att_in vest</b>	<b>std_i nvest</b>	<b>sent_ger eral</b>	<b>std_g enera l</b>	<b>sent_b rand</b>	<b>std_b rand</b>	<b>att_b rand</b>	<b>Yield</b>	<b>Volu me_b rand</b>	<b>Volu me_MOE X</b>	<b>att_m oex</b>
<b>count</b>	289	289	289	289	289	289	289	289	289	289	289	289	289	289	289	289	289	289	289	289	289
<b>mean</b>	-8,59	-8,6	-8,64	0,01	0	0,05	0,04	0,01	0	0,53	2,32	0,48	0,6	0,49	0,45	0,49	3,29	0,05	73894 ,5	13121 ,22	0,4
<b>std</b>	0,73	0,62	0,54	0,01	0,03	2,94	1,51	0,01	0,01	0,09	1,18	0,04	0,03	0,01	0,1	0,02	0,44	2,23	39009 ,83	8632, 83	0,55

<b>min</b>	-10,26	-10,03	-9,84	0	-0,15	-12,07	-5,88	0	-0,05	0,25	0	0,14	0,52	0,46	0,23	0,41	1,95	-9,57	8760	961,46	0
<b>25%</b>	-9,04	-8,97	-8,94	0	-0,01	-1,28	-0,49	0	0	0,45	2,56	0,47	0,58	0,48	0,37	0,48	3,04	-1,01	47940	7740	0
<b>50%</b>	-8,68	-8,68	-8,67	0,01	0	0	0	0,01	0	0,52	2,83	0,49	0,6	0,49	0,46	0,5	3,3	0,04	66620	11420	0
<b>75%</b>	-8,26	-8,36	-8,45	0,01	0,01	1,01	0,5	0,01	0	0,58	3,04	0,5	0,62	0,49	0,51	0,51	3,53	1,29	87530	15560	0,69
<b>max</b>	-5,37	-6,16	-6,92	0,09	0,13	16,12	10,12	0,07	0,06	0,82	3,5	0,52	0,7	0,5	0,74	0,53	5,6	12,9	263040	69820	2,71

Источник: построено автором на основе данных компаний «Сбербанк»

### Компания «МТС»

	log_rv_d	log_rv_w	log_rv_m	IMO EX	MSC I	Bond_SR	Bond_LR	USD_RUB	CPI	sent_invest	att_i nvest	std_i nvest	sent_gener al	std_gen era l	sent_bran d	std_b rand	att_b rand	Yield	Volu me_b rand	Volu me_MOE X	att_moex	
<b>count</b>	289	289	289	289	289	289	289	289	289	289	289	289	289	289	289	289	289	289	289	289		
<b>mean</b>	-8,88	-8,88	-8,88	0,01	0	0,05	0,04	0,01	0	0,53	2,32	0,48	0,6	0,49	0,25	0,43	3,31	0,02	4218,85	1312,122	0,4	
<b>std</b>	0,88	0,77	0,67	0,01	0,03	2,94	1,51	0,01	0,01	0,09	1,18	0,04	0,03	0,01	0,07	0,05	0,37	1,43	2188,64	8632,83	0,55	
<b>min</b>	-10,54	-9,96	-9,62	0	-0,15	-	12,07	-5,88	0	-0,05	0,25	0	0,14	0,52	0,46	0,04	0,11	2,08	-7,53	530,92	961,46	0
<b>25%</b>	-9,41	-9,35	-9,37	0	-0,01	-1,28	-0,49	0	0	0,45	2,56	0,47	0,58	0,48	0,2	0,4	3,04	-0,65	2780	7740	0	
<b>50%</b>	-9,03	-9,09	-9,03	0,01	0	0	0	0,01	0	0,52	2,83	0,49	0,6	0,49	0,25	0,44	3,33	0,12	3560	11420	0	
<b>75%</b>	-8,55	-8,64	-8,7	0,01	0,01	1,01	0,5	0,01	0	0,58	3,04	0,5	0,62	0,49	0,3	0,46	3,53	0,72	5170	15560	0,69	
<b>max</b>	-5,52	-5,88	-6,61	0,09	0,13	16,12	10,12	0,07	0,06	0,82	3,5	0,52	0,7	0,5	0,45	0,5	4,54	6,83	14560	69820	2,71	

Источник: построено автором на основе данных компаний «МТС»

## Компания «Лукойл»

	<b>log_r v_d</b>	<b>log_r v_w</b>	<b>log_r v_m</b>	<b>IMO EX</b>	<b>MSC I</b>	<b>Bond _SR</b>	<b>Bond _LR</b>	<b>USD _RU B</b>	<b>CPI</b>	<b>sent_ invest</b>	<b>att_i nvest</b>	<b>std_i nvest</b>	<b>sent_ gener al</b>	<b>std_g enera l</b>	<b>sent_ bran d</b>	<b>std_b rand</b>	<b>att_b rand</b>	<b>Yield</b>	<b>Volu me_b rand</b>	<b>Volu me_MOE X</b>	<b>att_ moex</b>
<b>count</b>	289	289	289	289	289	289	289	289	289	289	289	289	289	289	289	289	289	289	289	289	
<b>mean</b>	-8,37	-8,39	-8,44	0,01	0	0,05	0,04	0,01	0	0,53	2,32	0,48	0,6	0,49	0,52	0,38	0,84	0	1555, 2	1312 1,22	0,4
<b>std</b>	0,88	0,81	0,77	0,01	0,03	2,94	1,51	0,01	0,01	0,09	1,18	0,04	0,03	0,01	0,25	0,15	0,65	2,71	889,0 6	8632, 83	0,55
<b>min</b>	- 10,89	- 11,27	-10,3	0	-0,15	- 12,07	-5,88	0	-0,05	0,25	0	0,14	0,52	0,46	0	0	0	- 18,65	123,1 2	961,4 6	0
<b>25%</b>	-8,92	-8,89	-8,91	0	-0,01	-1,28	-0,49	0	0	0,45	2,56	0,47	0,58	0,48	0,34	0,32	0	-1,3	964,4 8	7740	0
<b>50%</b>	-8,4	-8,43	-8,45	0,01	0	0	0	0,01	0	0,52	2,83	0,49	0,6	0,49	0,52	0,38	0,69	0,01	1360	1142 0	0
<b>75%</b>	-8,01	-8,11	-8,14	0,01	0,01	1,01	0,5	0,01	0	0,58	3,04	0,5	0,62	0,49	0,68	0,46	1,1	1,24	1870	1556 0	0,69
<b>max</b>	-5,18	-5,78	-6,45	0,09	0,13	16,12	10,12	0,07	0,06	0,82	3,5	0,52	0,7	0,5	1	0,71	3,89	15,42	7530	6982 0	2,71

Источник: построено автором на основе данных компании «Лукойл»

## Компания «Роснефть»

	<b>log_r v_d</b>	<b>log_r v_w</b>	<b>log_r v_m</b>	<b>IMO EX</b>	<b>MSC I</b>	<b>Bond _SR</b>	<b>Bond _LR</b>	<b>USD _RU B</b>	<b>CPI</b>	<b>sent_ invest</b>	<b>att_i nvest</b>	<b>std_i nvest</b>	<b>sent_ gener al</b>	<b>std_g enera l</b>	<b>sent_ bran d</b>	<b>std_b rand</b>	<b>att_b rand</b>	<b>Yield</b>	<b>Volu me_b rand</b>	<b>Volu me_MOE X</b>	<b>att_ moex</b>
--	----------------------	----------------------	----------------------	-------------------	------------------	---------------------	---------------------	--------------------------	------------	-------------------------	------------------------	------------------------	-------------------------------	------------------------------	-----------------------------	-----------------------	-----------------------	--------------	-------------------------------	------------------------------	----------------------

<b>count</b>	289	289	289	289	289	289	289	289	289	289	289	289	289	289	289	289	289	289	289	
<b>mean</b>	-8,41	-8,42	-8,46	0,01	0	0,05	0,04	0,01	0	0,53	2,32	0,48	0,6	0,49	0,49	0,43	1,24	0,08	9538, 44	1312 1,22
<b>std</b>	0,87	0,77	0,69	0,01	0,03	2,94	1,51	0,01	0,01	0,09	1,18	0,04	0,03	0,01	0,24	0,14	0,8	2,58	7512, 98	8632, 83
<b>min</b>	-10,4	-	10,52	-9,68	0	-0,15	-	12,07	-5,88	0	-0,05	0,25	0	0,14	0,52	0,46	0	0,04	0	- 909,6 16,91
<b>25%</b>	-8,94	-8,9	-8,93	0	-0,01	-1,28	-0,49	0	0	0,45	2,56	0,47	0,58	0,48	0,31	0,39	0,69	-1,04	4600	7740
<b>50%</b>	-8,5	-8,53	-8,52	0,01	0	0	0	0,01	0	0,52	2,83	0,49	0,6	0,49	0,49	0,45	1,1	-0,01	7440	1142 0
<b>75%</b>	-8,06	-8,07	-8,17	0,01	0,01	1,01	0,5	0,01	0	0,58	3,04	0,5	0,62	0,49	0,65	0,52	1,61	1,18	1145 0	1556 0
<b>max</b>	-5,44	-5,97	-6,51	0,09	0,13	16,12	10,12	0,07	0,06	0,82	3,5	0,52	0,7	0,5	0,97	0,68	4,48	12,99	4541 0	6982 0

Источник: построено автором на основе данных компаний «Роснефть»

### Компания «Газпром»

	<b>log_r v_d</b>	<b>log_r v_w</b>	<b>log_r v_m</b>	<b>IMO EX</b>	<b>MSC I</b>	<b>Bond _SR</b>	<b>Bond _LR</b>	<b>USD _RU B</b>	<b>CPI</b>	<b>sent_ inves t</b>	<b>att_i nvest</b>	<b>std_i nvest</b>	<b>sent_ gener al</b>	<b>std_g enera l</b>	<b>sent_ bran d</b>	<b>std_b rand</b>	<b>att_b rand</b>	<b>Yield</b>	<b>Volu me_b rand</b>	<b>Volu me_MOE X</b>	<b>att_ moex</b>	
<b>count</b>	289	289	289	289	289	289	289	289	289	289	289	289	289	289	289	289	289	289	289	289		
<b>mean</b>	-8,57	-8,58	-8,6	0,01	0	0,05	0,04	0,01	0	0,53	2,32	0,48	0,6	0,49	0,54	0,49	2,72	-0,04	5677 2,35	1312 1,22	0,4	
<b>std</b>	0,76	0,65	0,56	0,01	0,03	2,94	1,51	0,01	0,01	0,09	1,18	0,04	0,03	0,01	0,11	0,05	0,58	1,84	2930 5	8632, 83	0,55	
<b>min</b>	-	10,47	-	10,07	-9,76	0	-0,15	-	12,07	-5,88	0	-0,05	0,25	0	0,14	0,52	0,46	0,18	0,15	1,1	-9,17	6060
<b>25%</b>	-9,07	-8,97	-8,93	0	-0,01	-1,28	-0,49	0	0	0,45	2,56	0,47	0,58	0,48	0,48	0,49	2,3	-1,06	3772 0	7740	0	

<b>50%</b>	-8,65	-8,63	-8,64	0,01	0	0	0	0,01	0	0,52	2,83	0,49	0,6	0,49	0,55	0,5	2,71	-0,07	4904 0	1142 0	0
<b>75%</b>	-8,2	-8,32	-8,38	0,01	0,01	1,01	0,5	0,01	0	0,58	3,04	0,5	0,62	0,49	0,61	0,51	3,09	1,07	6700 0	1556 0	0,69
<b>max</b>	-5,44	-6,28	-6,93	0,09	0,13	16,12	10,12	0,07	0,06	0,82	3,5	0,52	0,7	0,5	0,89	0,66	4,57	5,7	2052 30	6982 0	2,71

Источник: построено автором на основе данных компании «Газпром»

### Компания «Новатэк»

	log_r_v_d	log_r_v_w	log_r_v_m	IMO_EX	MSC_I	Bond_SR	Bond_LR	USD_RUB	CPI	sent_invest	att_i_nvest	std_i_nvest	sent_general	std_general	sent_branch	std_branch	att_branch	Yield	Volume_branch	Volume_MOEX	att_moex
<b>count</b>	289	289	289	289	289	289	289	289	289	289	289	289	289	289	289	289	289	289	289	289	
<b>mean</b>	-8,01	-8,02	-8,06	0,01	0	0,05	0,04	0,01	0	0,53	2,32	0,48	0,6	0,49	0,6	0,27	0,28	0,04	1709, 53	1312 1,22	0,4
<b>std</b>	0,79	0,69	0,63	0,01	0,03	2,94	1,51	0,01	0,01	0,09	1,18	0,04	0,03	0,01	0,19	0,09	0,47	2,6	1086, 7	8632, 83	0,55
<b>min</b>	-10,04	-9,59	-9,25	0	-0,15	-12,07	-5,88	0	-0,05	0,25	0	0,14	0,52	0,46	0,05	0,05	0	-12,61	96,46	961,4 6	0
<b>25%</b>	-8,53	-8,47	-8,46	0	-0,01	-1,28	-0,49	0	0	0,45	2,56	0,47	0,58	0,48	0,55	0,26	0	-1,24	995,3 3	7740	0
<b>50%</b>	-8,1	-8,12	-8,11	0,01	0	0	0	0,01	0	0,52	2,83	0,49	0,6	0,49	0,65	0,26	0	-0,11	1400	1142 0	0
<b>75%</b>	-7,72	-7,78	-7,9	0,01	0,01	1,01	0,5	0,01	0	0,58	3,04	0,5	0,62	0,49	0,65	0,26	0,69	1,22	2040	1556 0	0,69
<b>max</b>	-4,96	-5,61	-6,14	0,09	0,13	16,12	10,12	0,07	0,06	0,82	3,5	0,52	0,7	0,5	0,96	0,66	3,58	16,29	7570	6982 0	2,71

Источник: построено автором на основе данных компании «Новатэк»

## ПРИЛОЖЕНИЕ 2

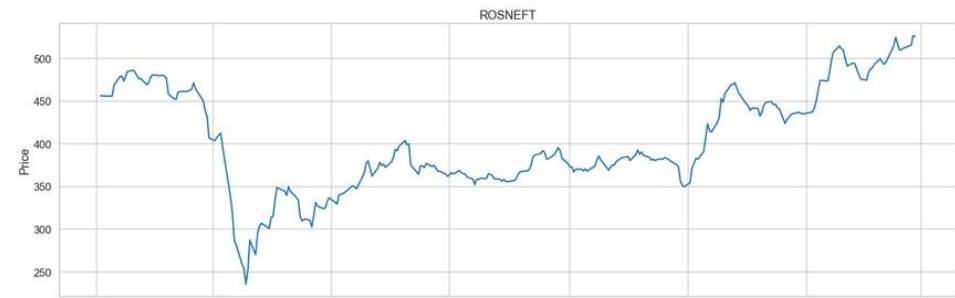
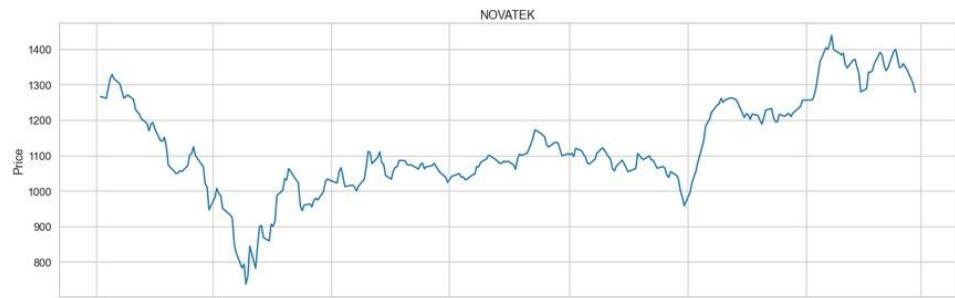
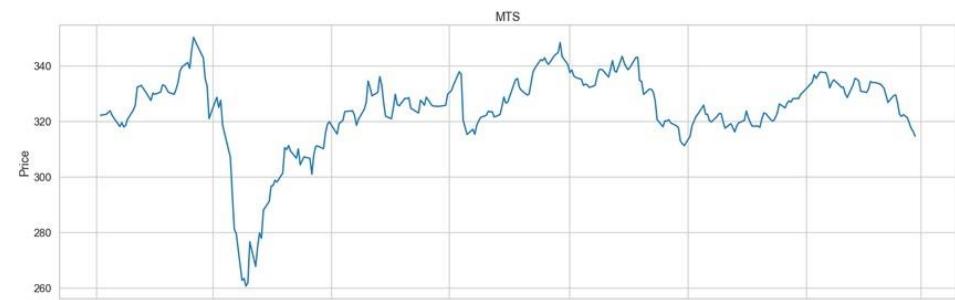
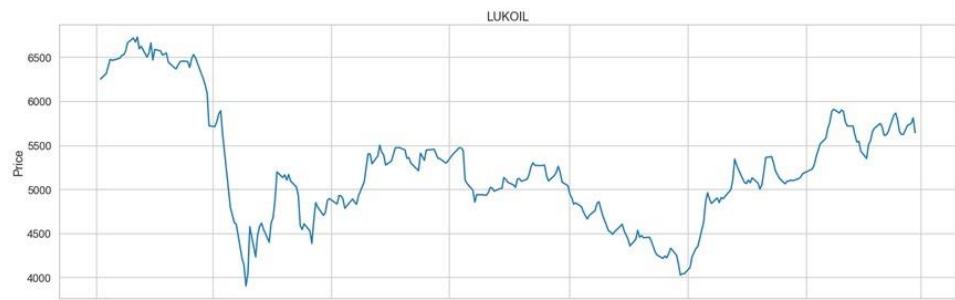
Корреляция между объясняющими переменными и целевой зависимой переменной для каждой рассматриваемой компании. По строкам расположены факторы, по столбцам компании. Обозначения признаков аналогичны как в приложении 1. Салатовым цветом выделены факторы, имеющие корреляцию с целевой переменной больше 0,3.

<b>Фактор</b>	<b>Лукойл</b>	<b>МТС</b>	<b>Новатэк</b>	<b>Роснефть</b>	<b>Газпром</b>	<b>Индекс</b>	<b>Сбербанк</b>
<b>log_rv_d</b>	0,79	0,76	0,75	0,74	0,70	0,67	0,68
<b>log_rv_w</b>	0,77	0,78	0,75	0,74	0,69	0,62	0,68
<b>log_rv_m</b>	0,58	0,61	0,60	0,59	0,51	0,46	0,45
<b>IMOEX</b>	0,58	0,52	0,54	0,56	0,51	0,39	0,56
<b>MSCI</b>	-0,15	-0,20	-0,16	-0,15	-0,20	-0,17	-0,19
<b>Bond_SR</b>	0,07	0,02	0,05	0,04	0,10	0,08	0,10
<b>Bond_LR</b>	0,12	0,13	0,13	0,11	0,18	0,19	0,20
<b>USD_RUB</b>	0,47	0,45	0,46	0,44	0,44	0,34	0,47
<b>CPI</b>	0,03	0,00	0,01	0,04	0,00	0,01	0,03
<b>sent_invest</b>	0,07	0,09	0,08	0,10	0,05	0,03	0,04
<b>att_invest</b>	0,10	0,12	0,11	0,14	0,09	0,14	0,07
<b>std_invest</b>	0,06	0,06	0,03	0,03	0,05	0,10	0,08
<b>sent_general</b>	-0,41	-0,27	-0,29	-0,37	-0,32	-0,15	-0,35
<b>std_general</b>	0,38	0,21	0,25	0,32	0,29	0,13	0,31
<b>sent_brand</b>	-0,10	-0,01	-0,03	0,04	0,09	0,14	0,06
<b>std_brand</b>	0,05	-0,04	0,10	0,06	-0,01	-0,15	0,13
<b>att_brand</b>	0,21	-0,13	0,14	0,40	0,03	0,11	0,02
<b>Yield</b>	0,04	0,00	0,01	0,03	0,01	-0,08	-0,05
<b>Volume_brand</b>	0,64	0,51	0,62	0,67	0,55	0,36	0,57
<b>Volume_MOEX</b>	0,55	0,50	0,50	0,49	0,44	0,40	0,48
<b>att_moex</b>	0,36	0,32	0,32	0,38	0,35	0,27	0,34

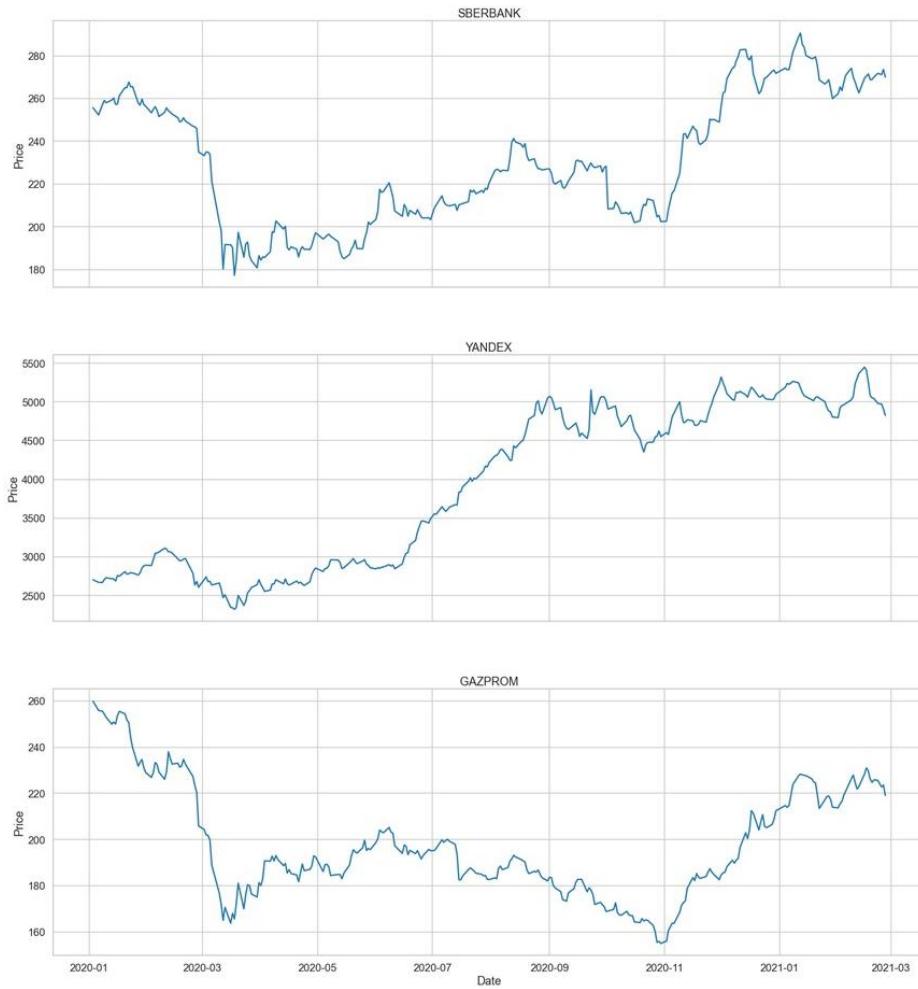
Источник: построено автором

### ПРИЛОЖЕНИЕ 3

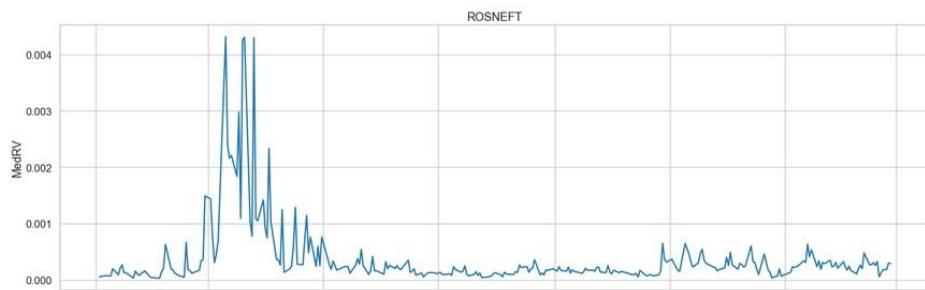
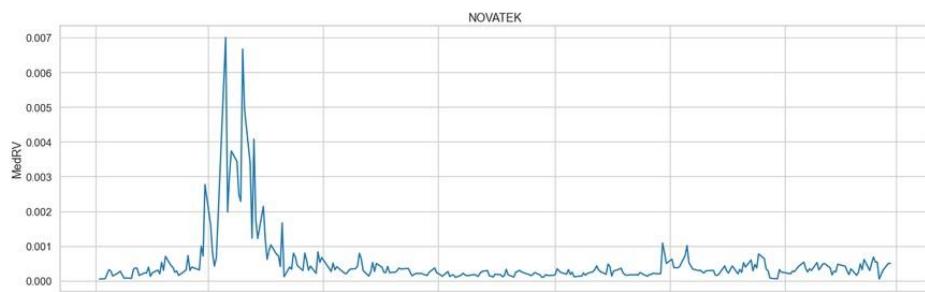
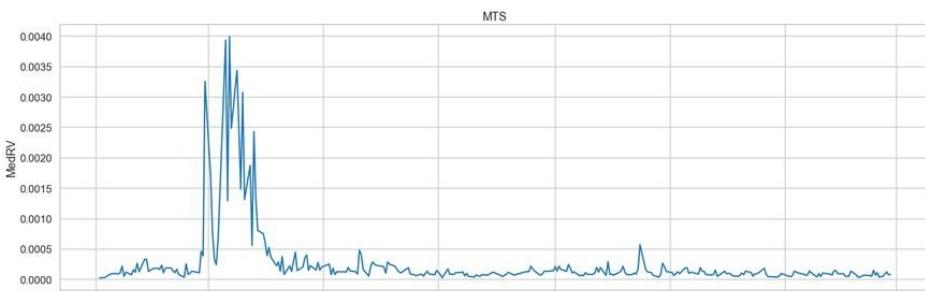
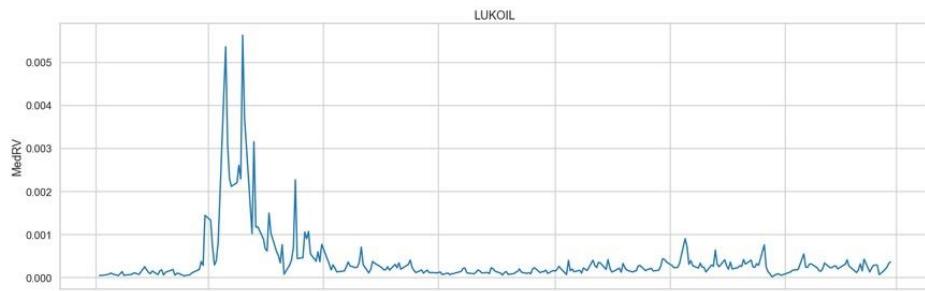
Временные ряды целевой переменной и ее распределение. А. Временной ряд цен акций всех компаний за рассматриваемый промежуток. Б. Временной ряд оценки реализованной волатильности (MedRV) для всех компаний. В. Распределение оценки реализованной волатильности для каждой компании. Г. Временной ряд логарифма оценки реализованной волатильности для всех компаний. Д. Распределение логарифма оценки реализованной волатильности для всех компаний.



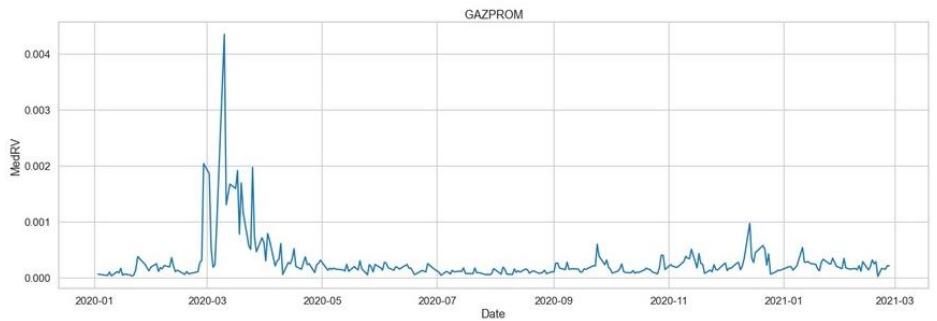
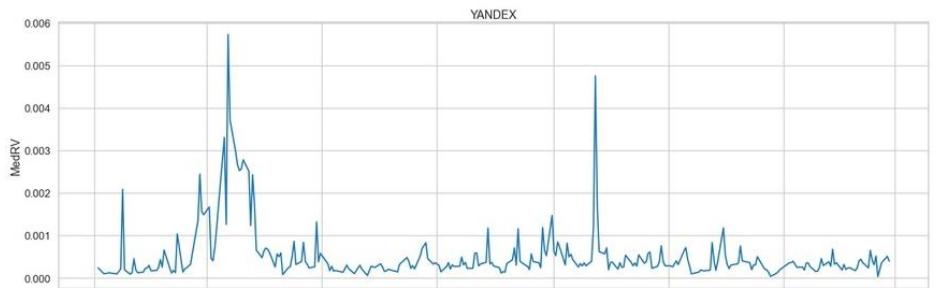
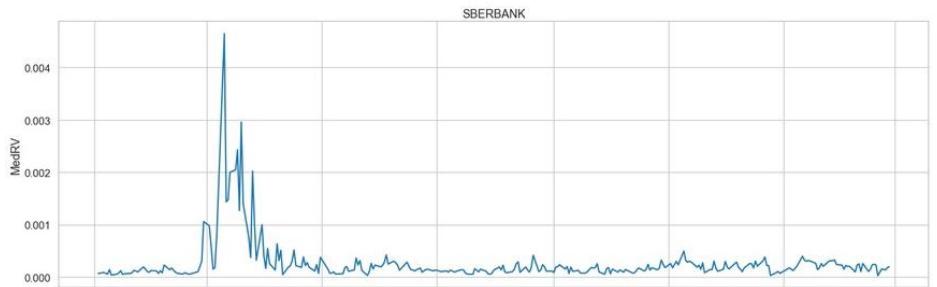
А. Временной ряд цен акций всех компаний за  
рассматриваемый промежуток



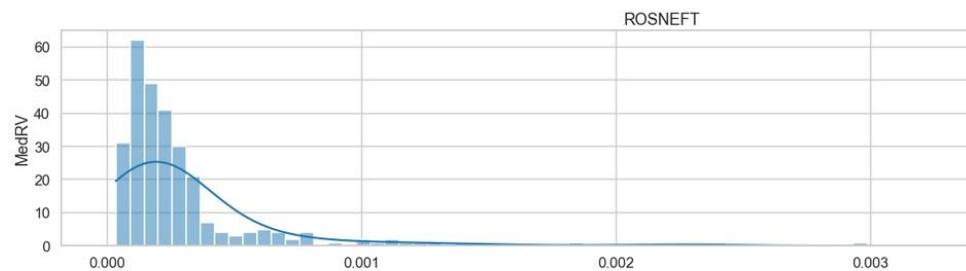
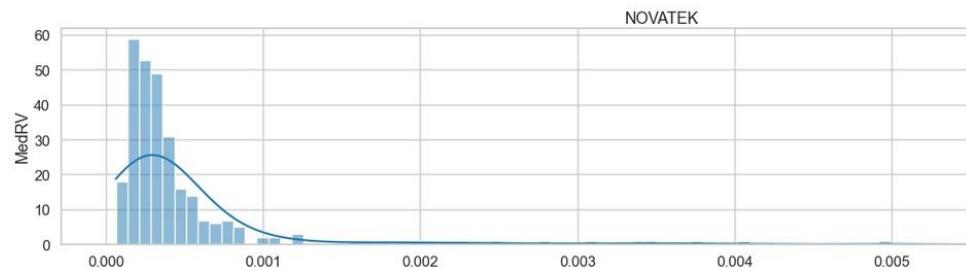
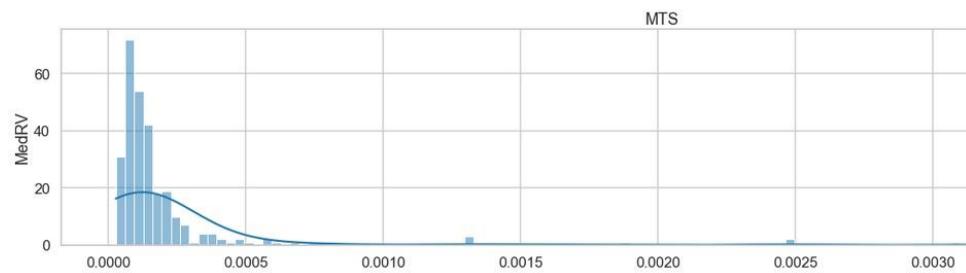
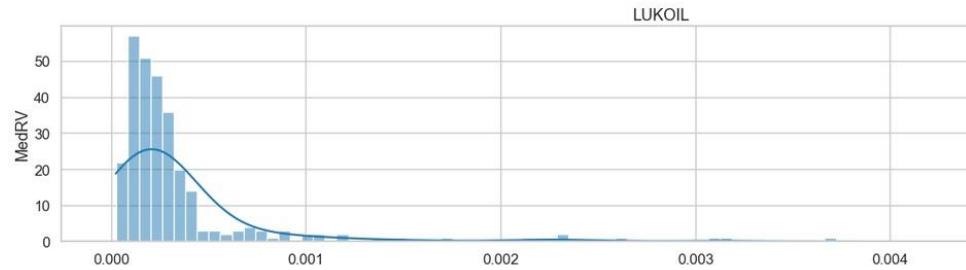
Источник: построено автором



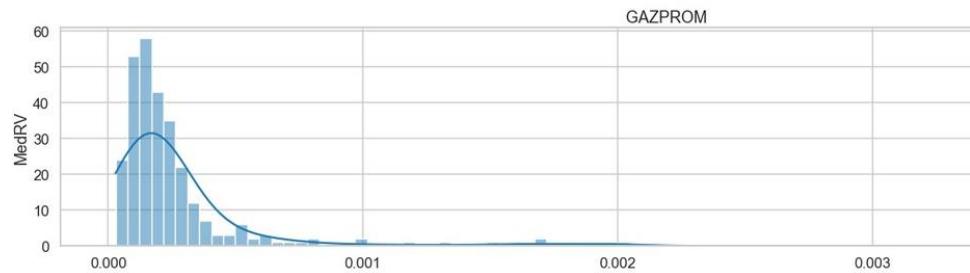
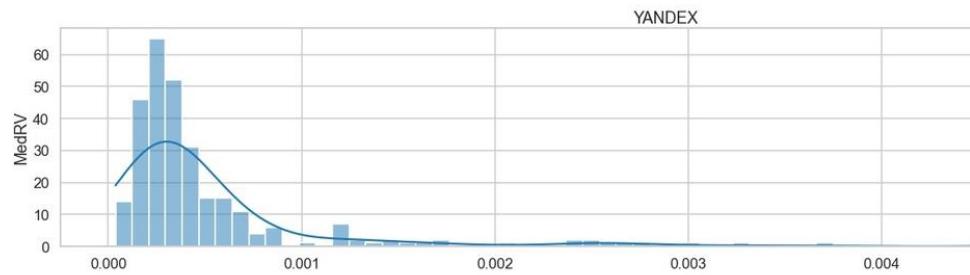
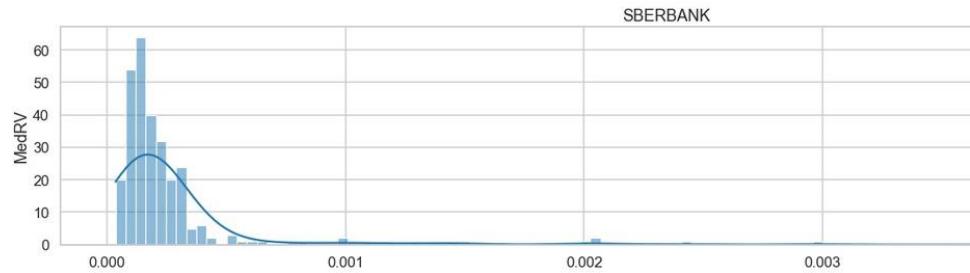
Б. Временной ряд оценки реализованной волатильности  
(MedRV) для всех компаний.



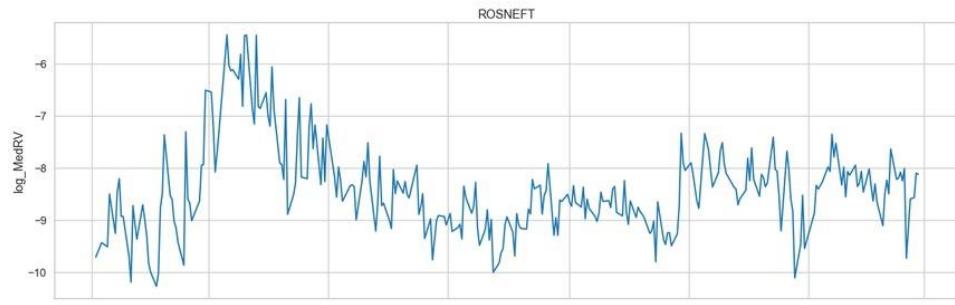
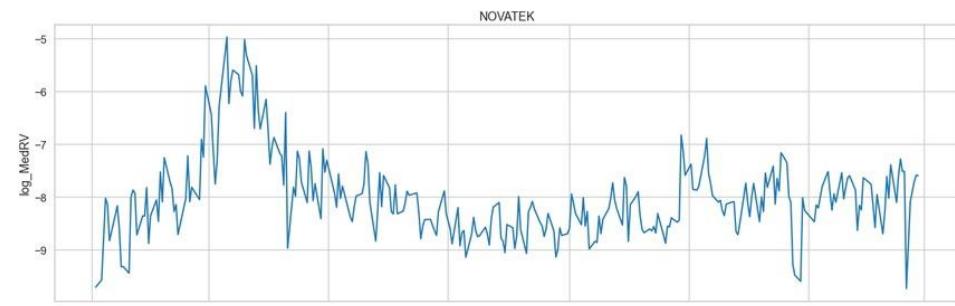
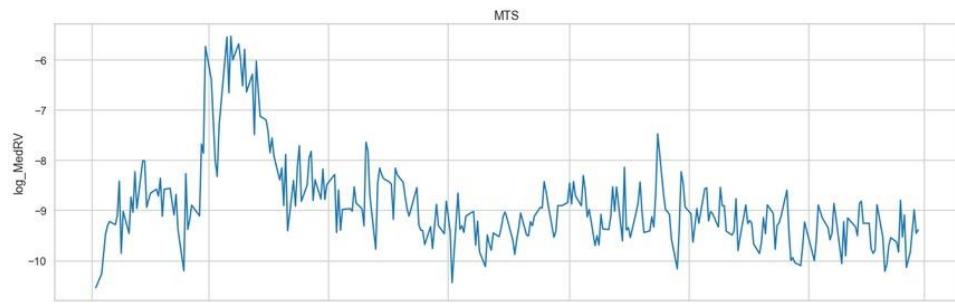
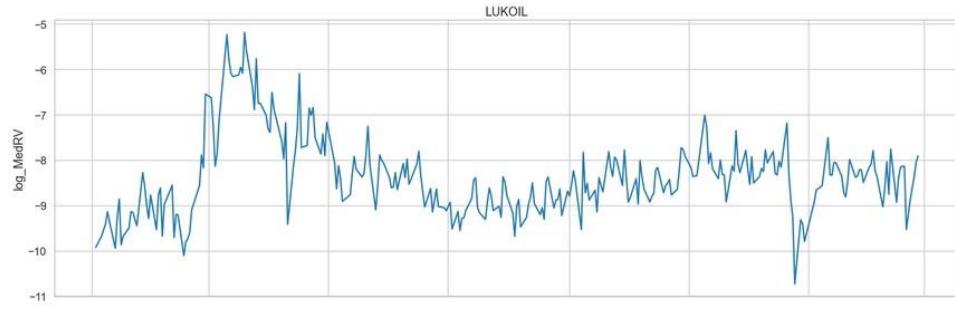
Источник: построено автором



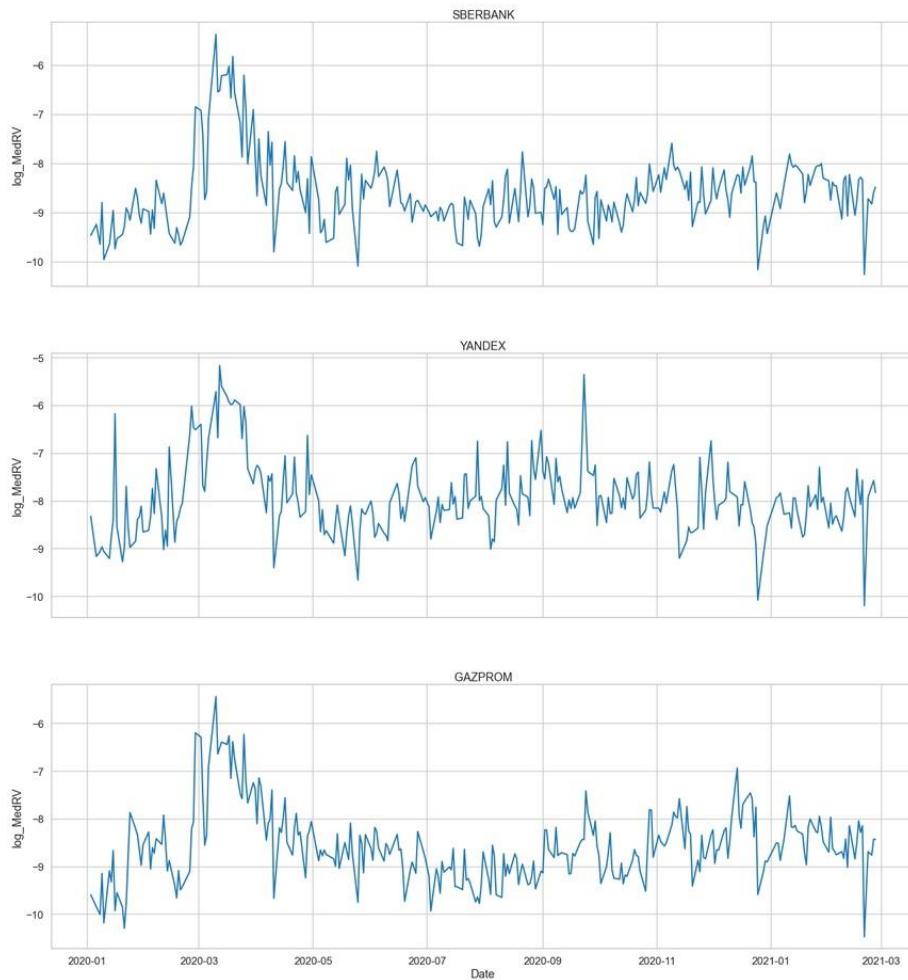
В. Распределение оценки реализованной волатильности для каждой компании.



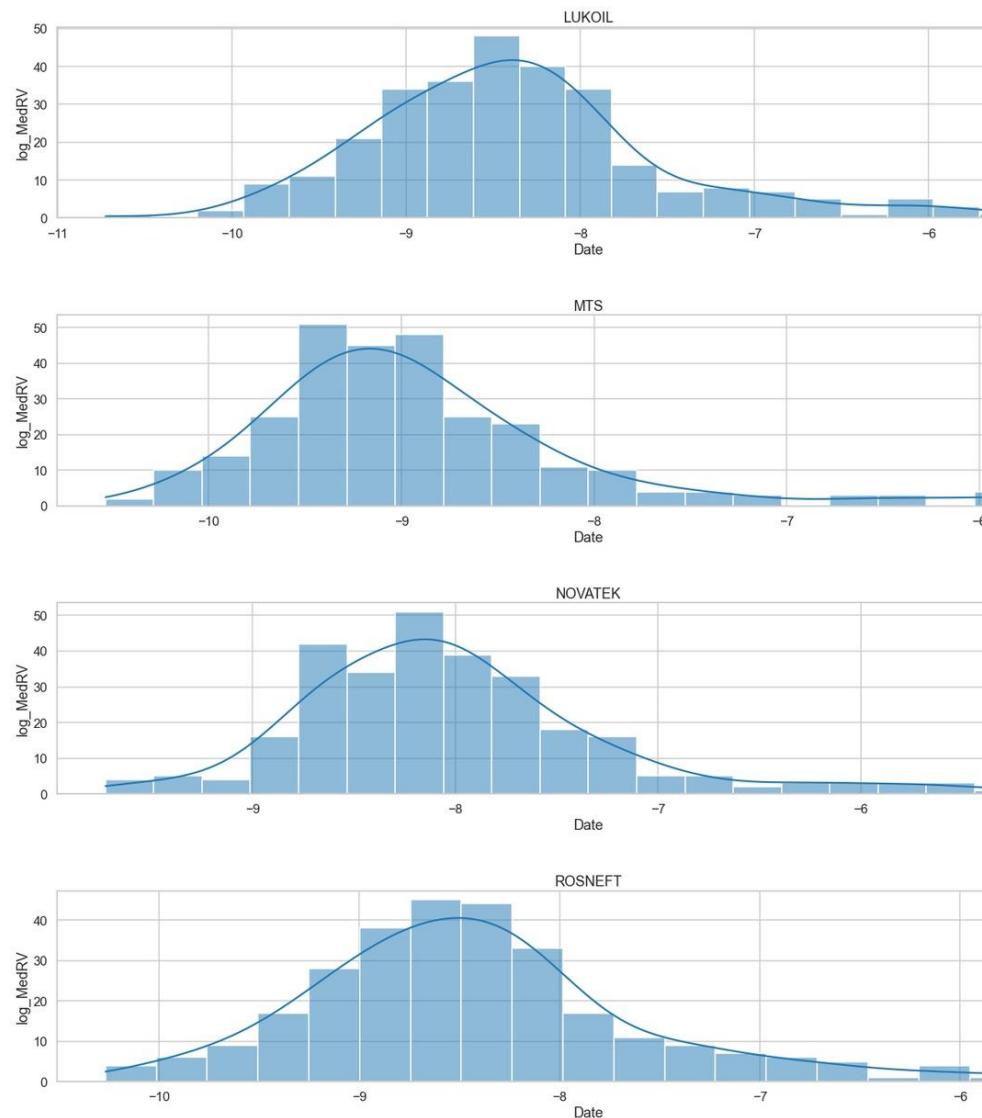
Источник: построено автором



Г. Временной ряд логарифма оценки реализованной волатильности для всех компаний.

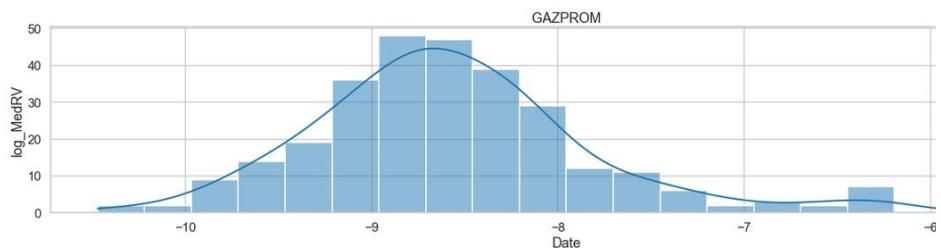
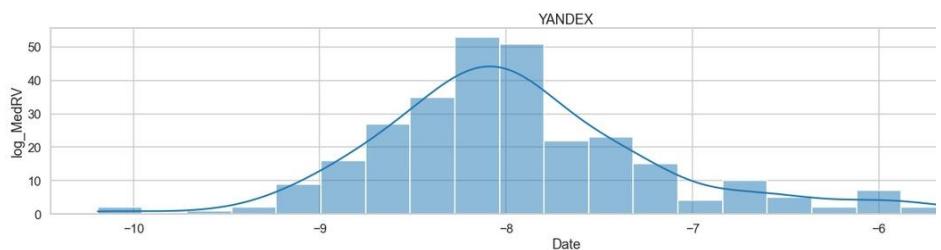


Источник: построено автором



Д. Распределение логарифма оценки реализованной

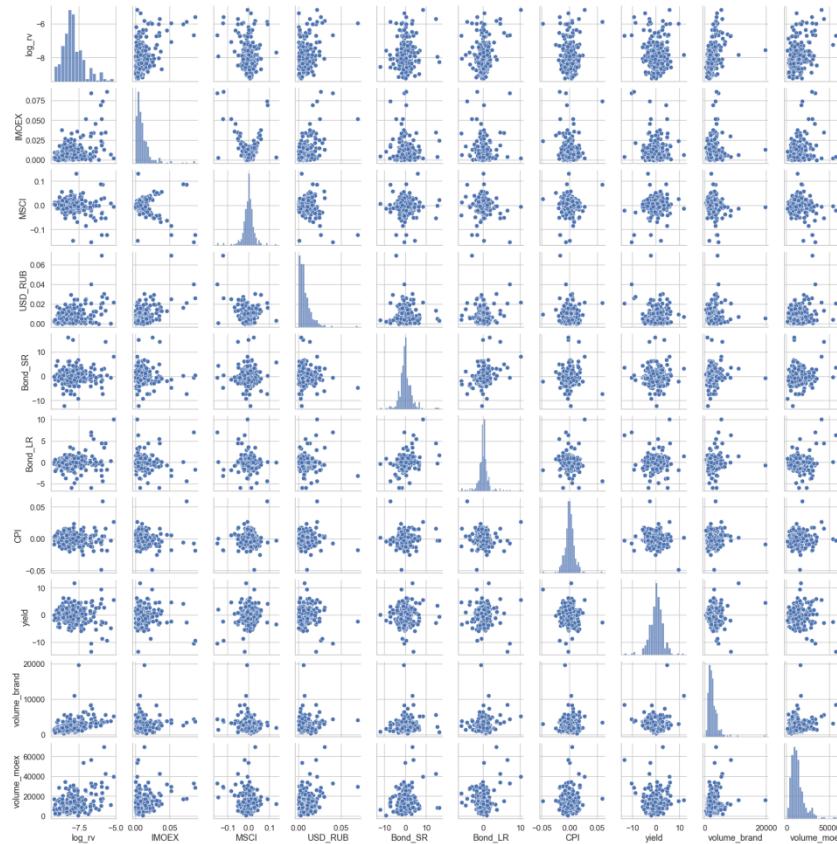
волатильности для всех компаний.



Источник: построено автором

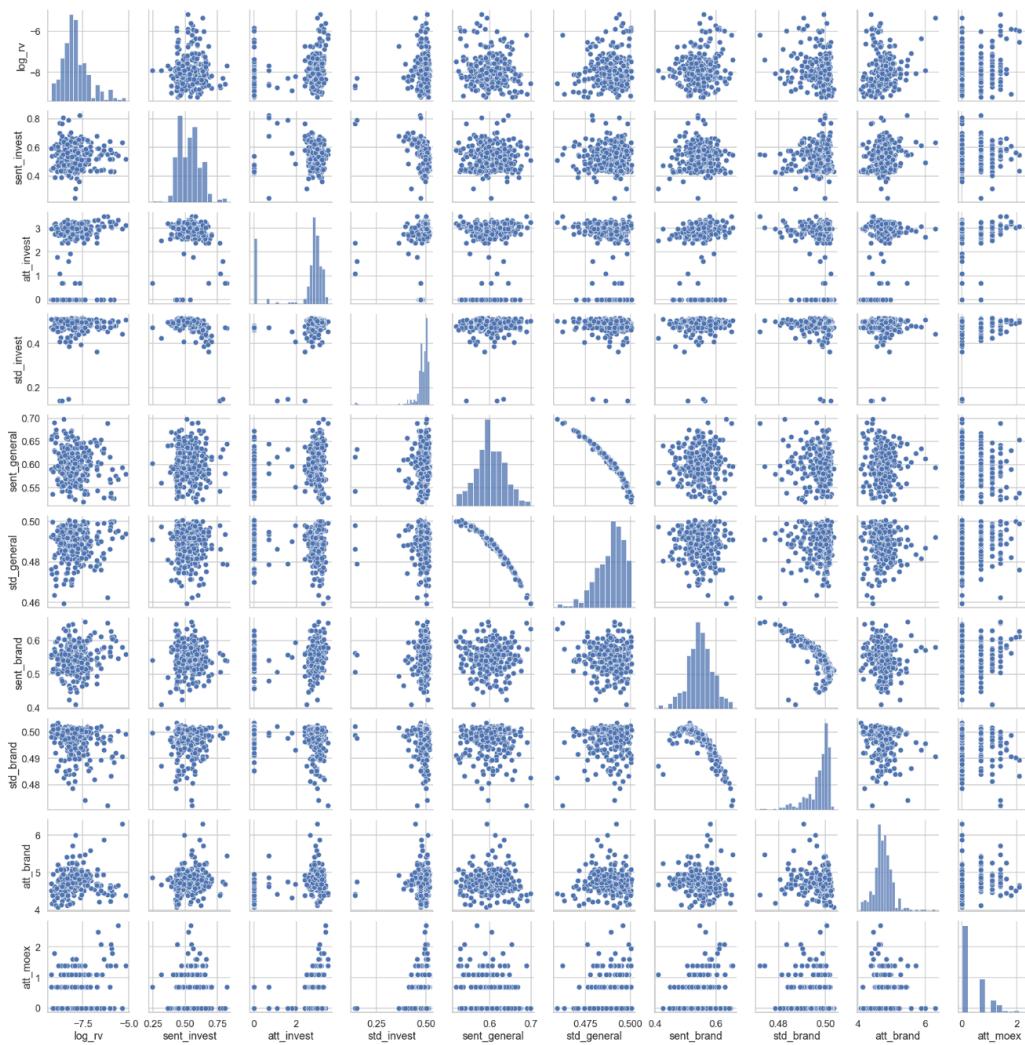
Графики зависимостей между экономическими и финансовыми переменными для каждой компании. Графики зависимостей между факторами настроения и внимания для каждой компании.

### Компания «Яндекс». Экономические и финансовые показатели



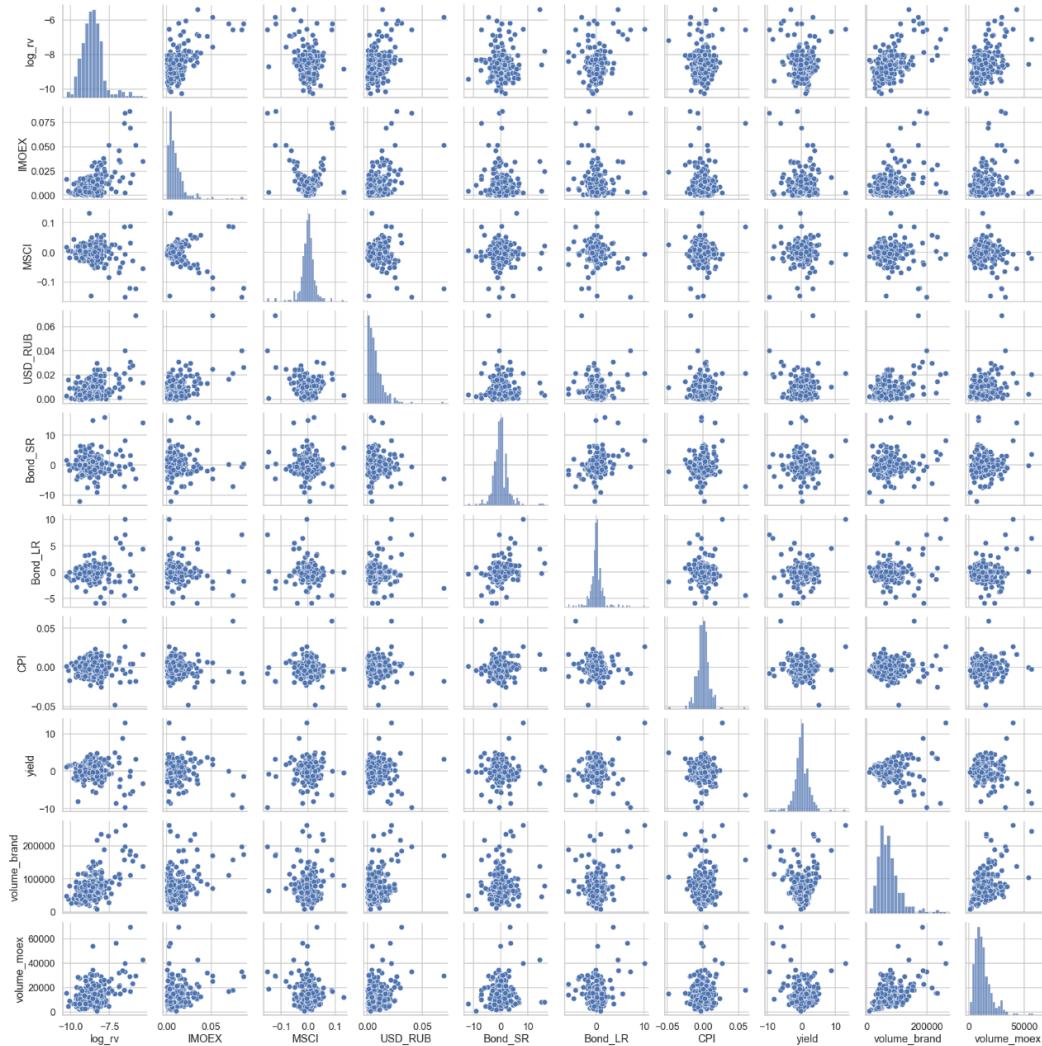
Источник: построено автором

## Компания «Яндекс». Показатели внимания и настроения



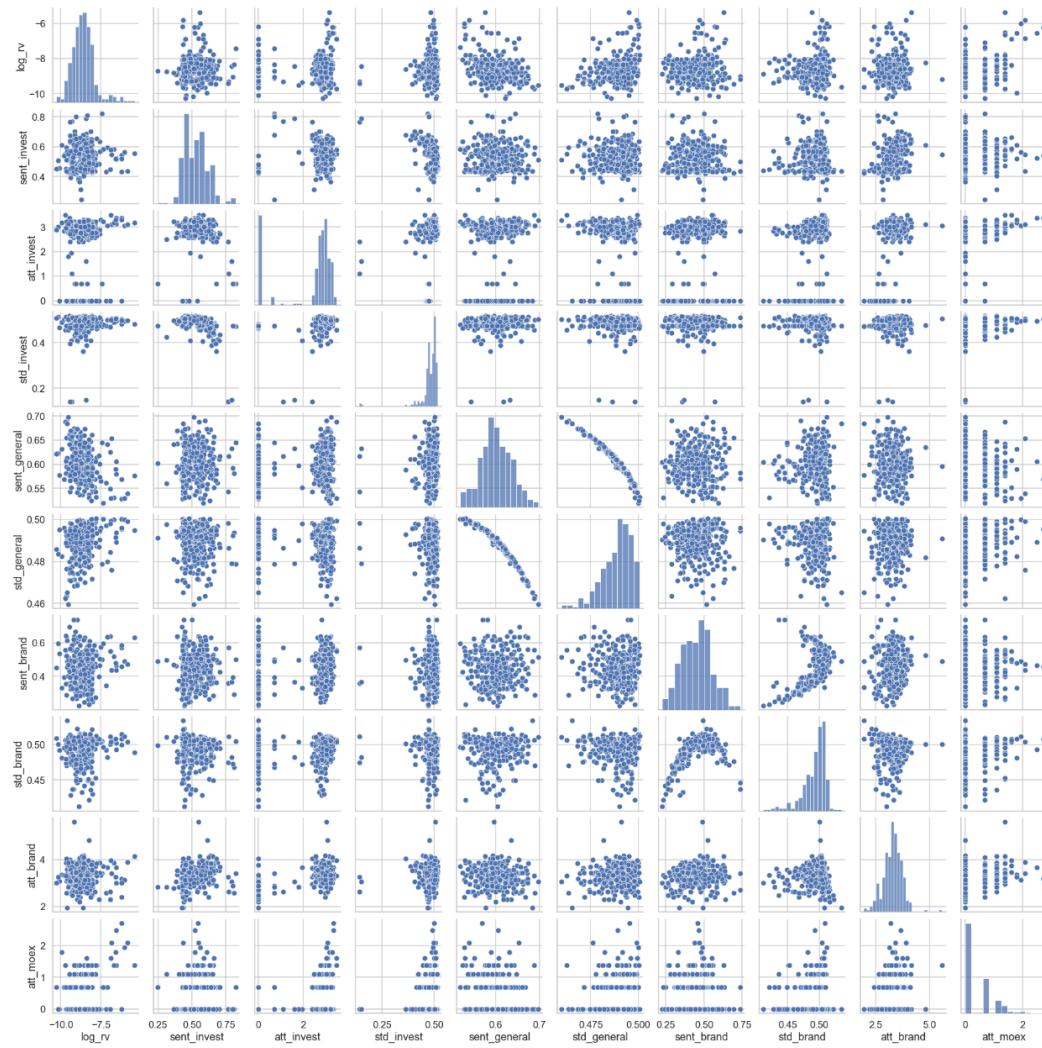
Источник: построено автором

## Компания «Сбербанк». Экономические и финансовые показатели



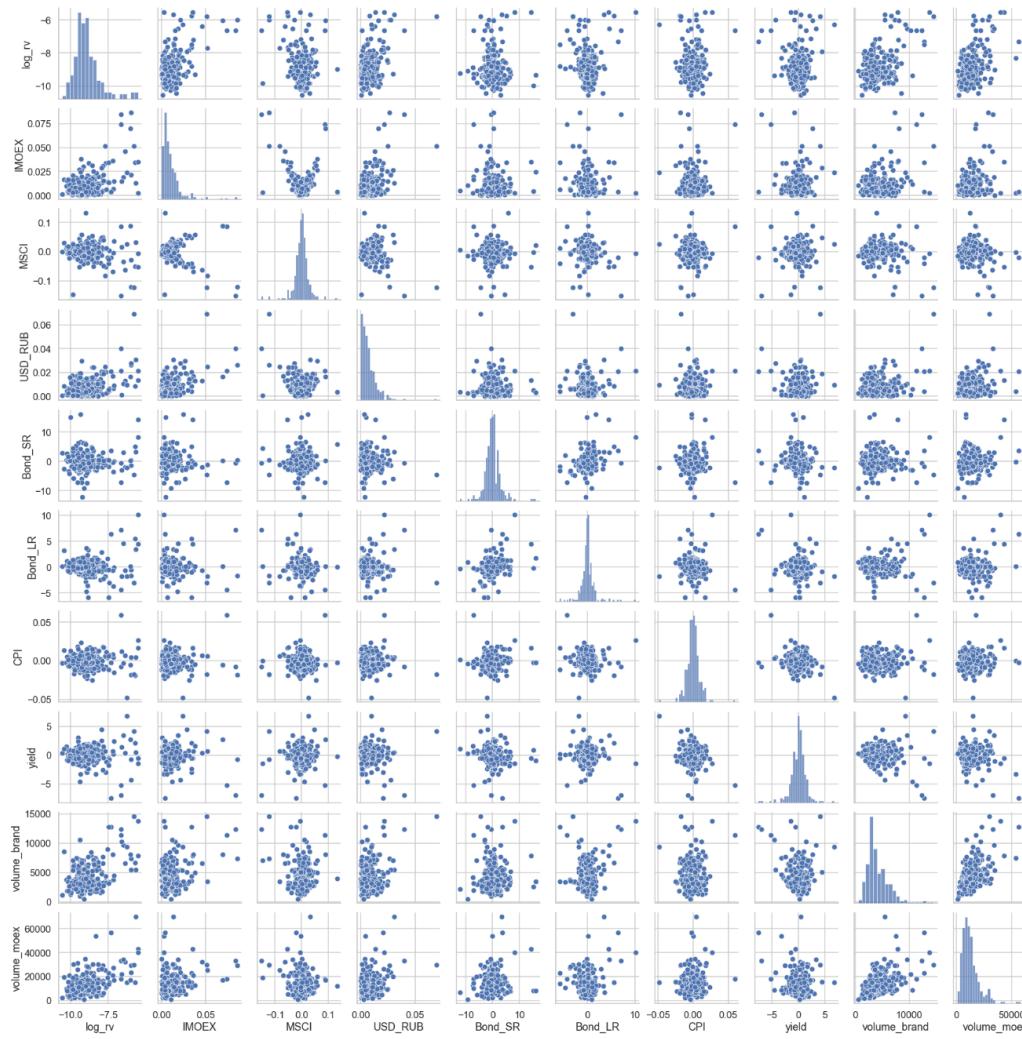
Источник: построено автором

## Компания «Сбербанк». Показатели внимания и настроения



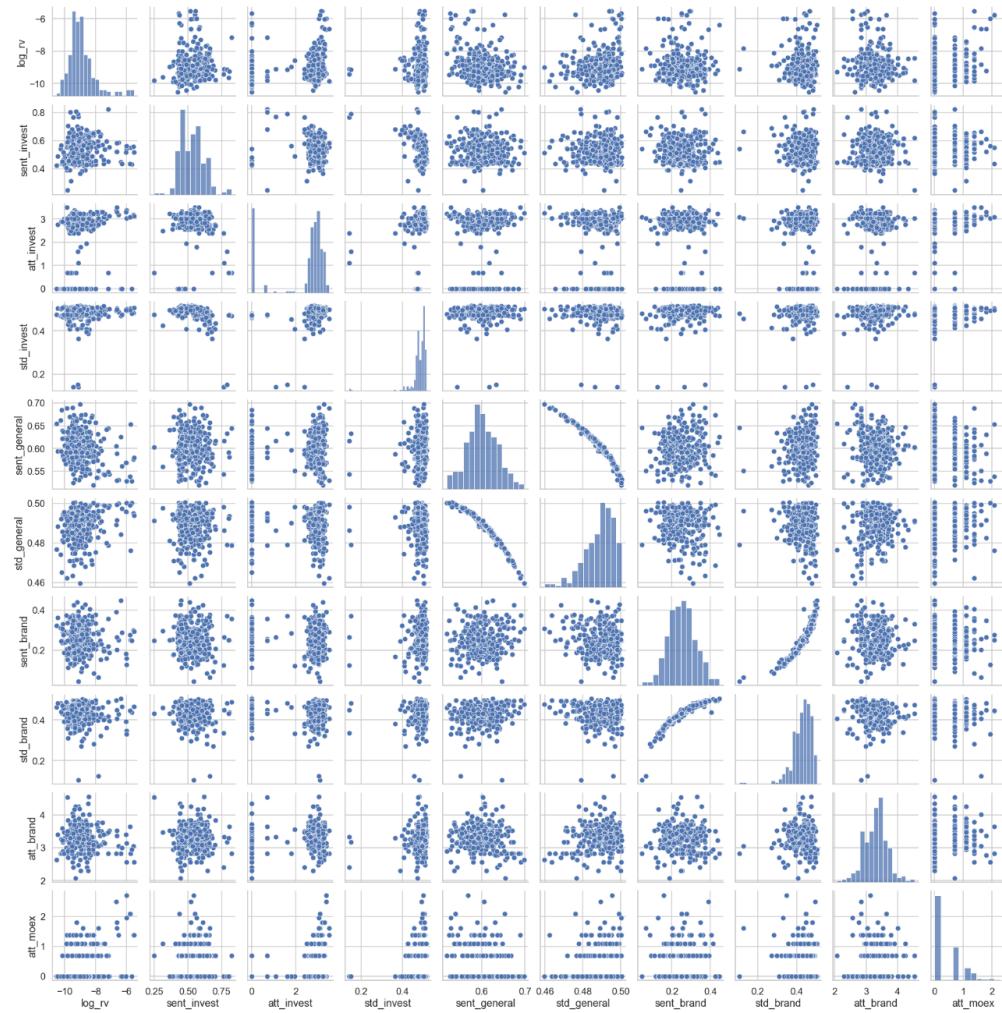
Источник: построено автором

## Компания «МТС». Экономические и финансовые показатели



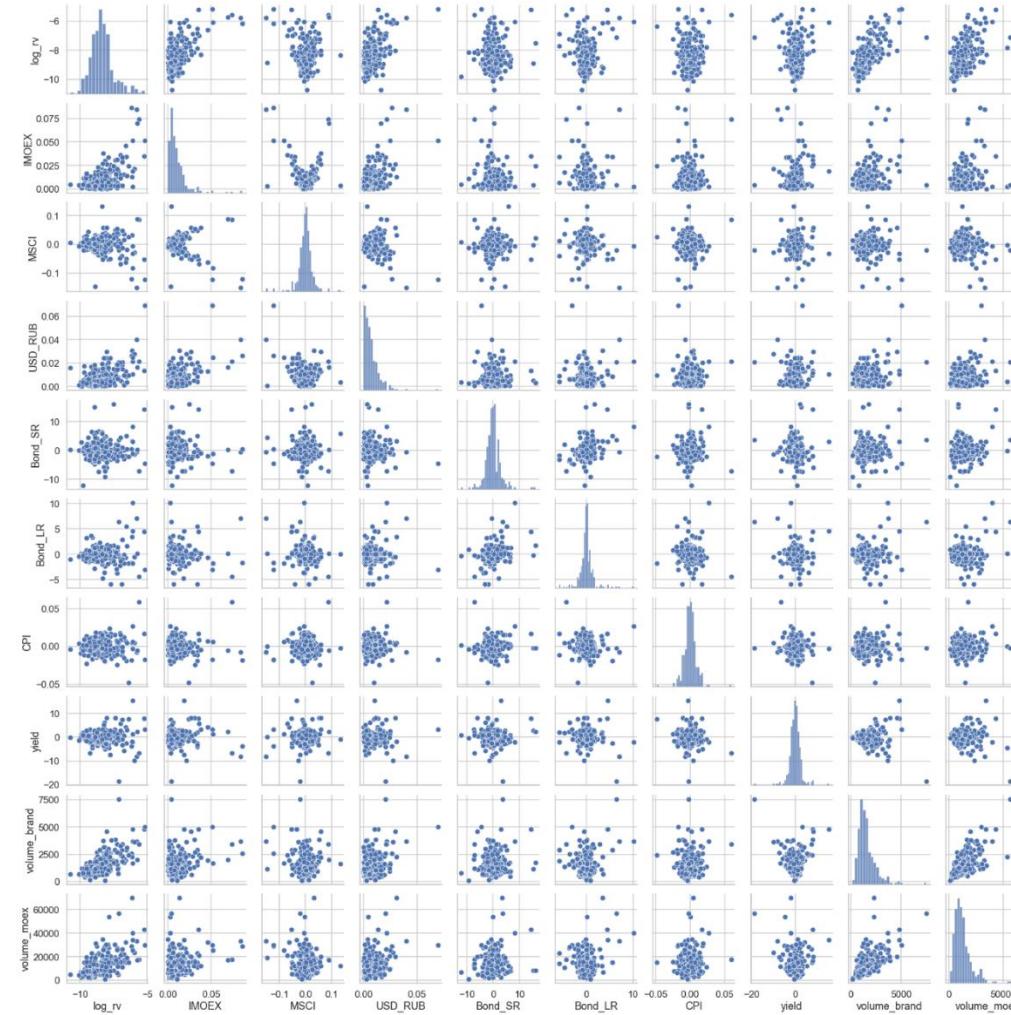
Источник: построено автором

## Компания «МТС». Показатели внимания и настроения



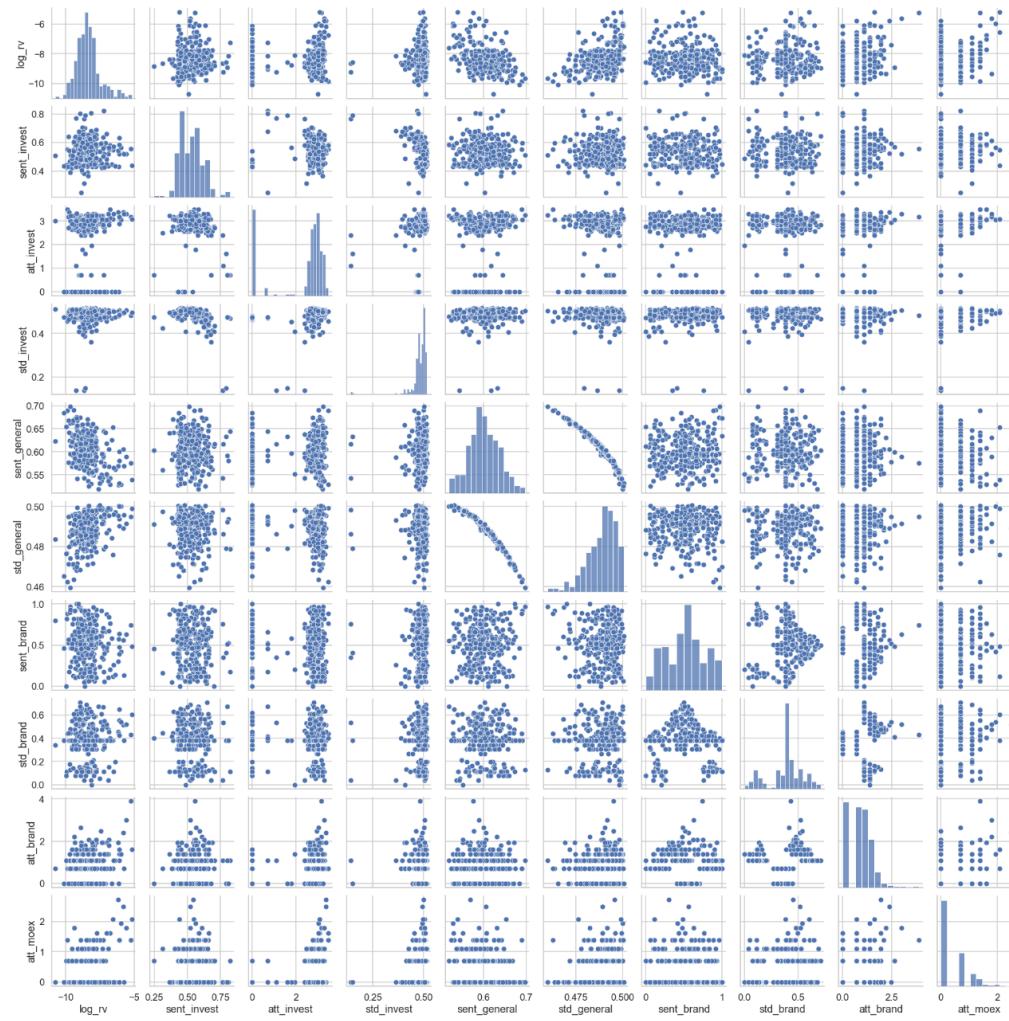
Источник: построено автором

## Компания «ЛУКОЙЛ». Экономические и финансовые показатели



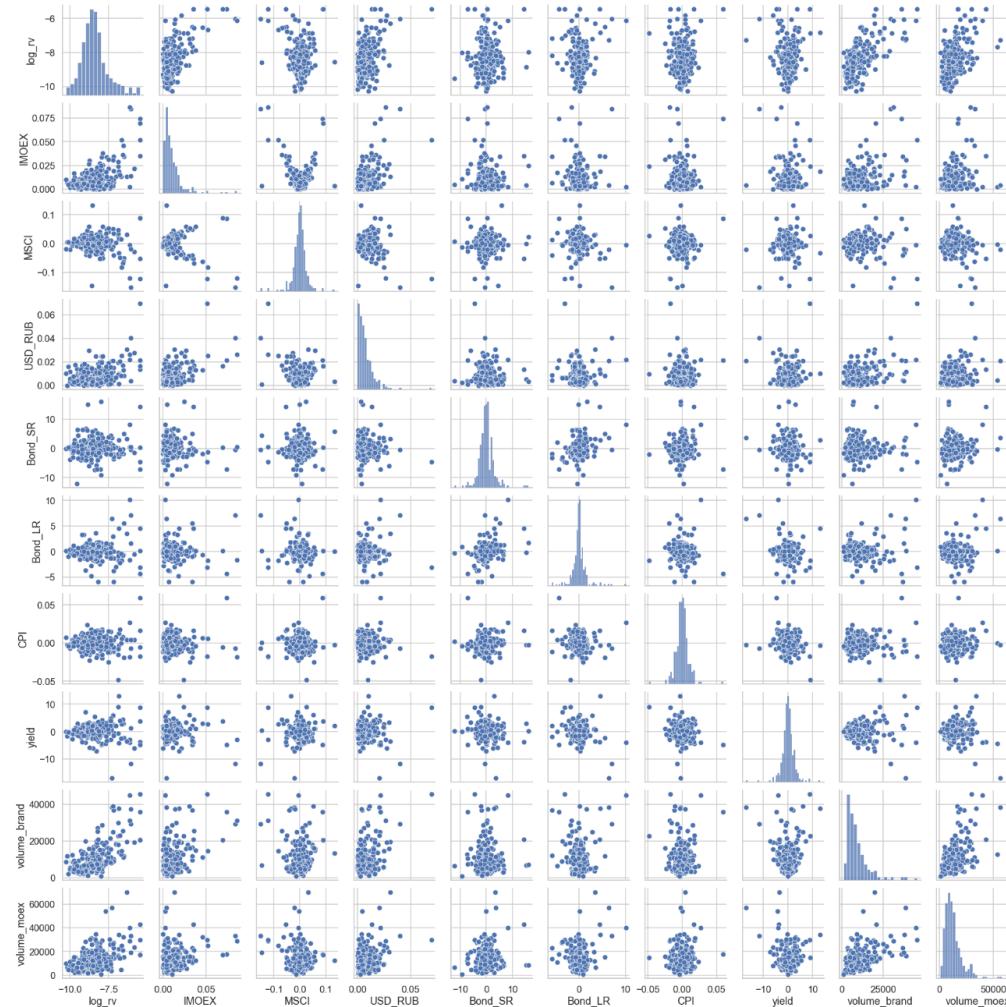
Источник: построено автором

## Компания «ЛУКОЙЛ». Показатели внимания и настроения



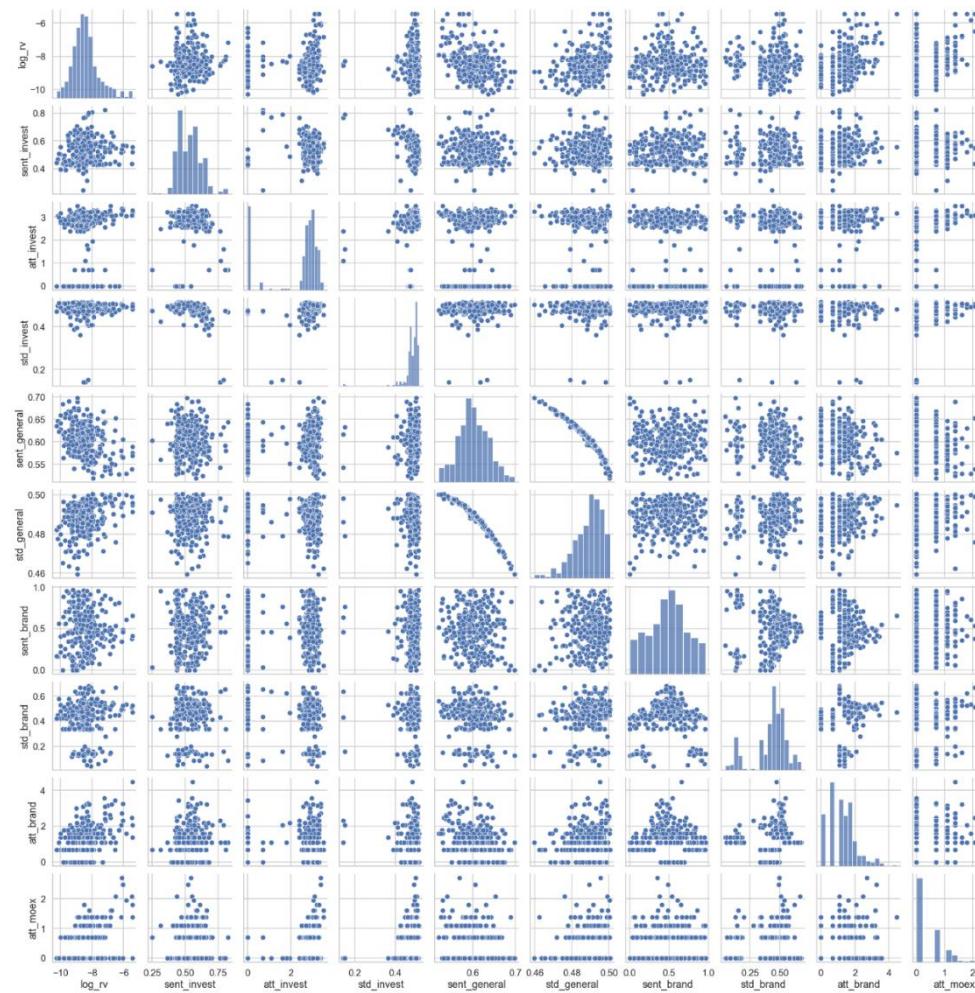
Источник: построено автором

## Компания «Роснефть». Экономические и финансовые показатели



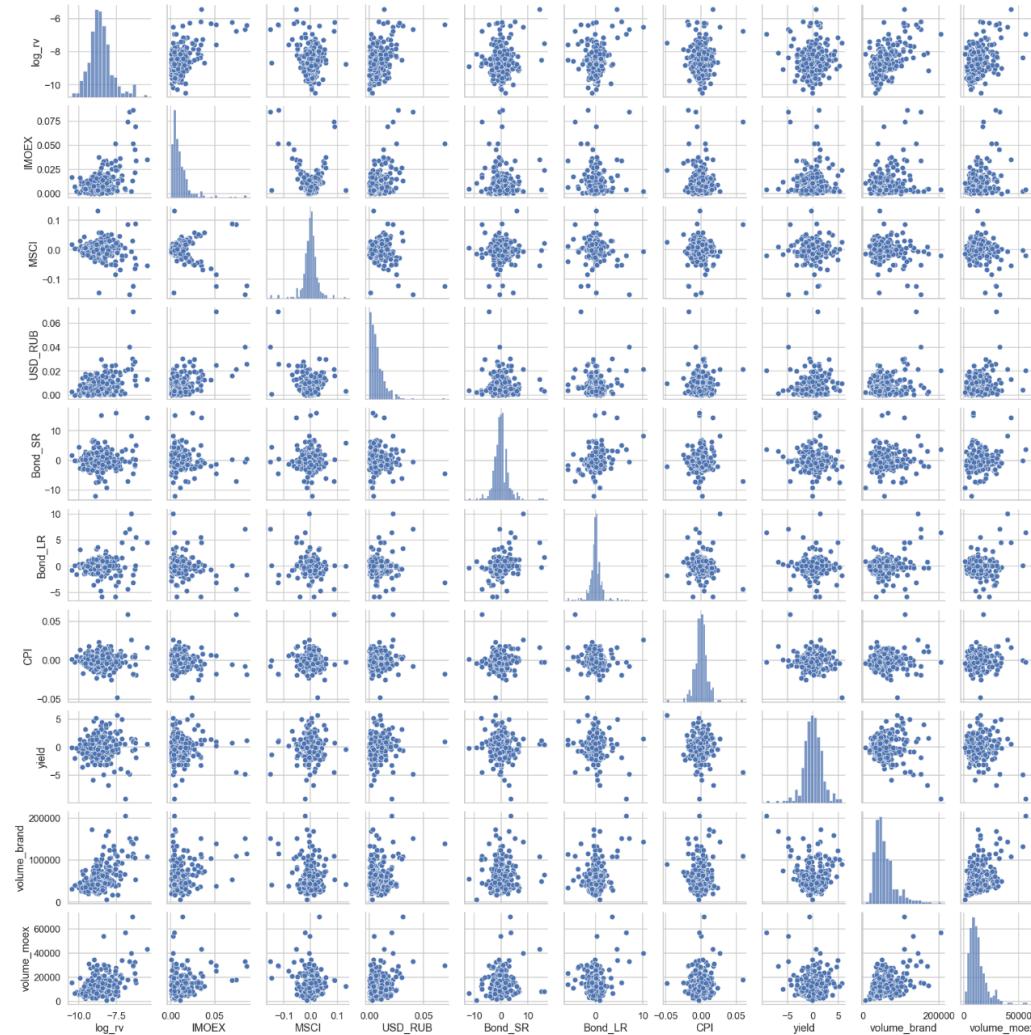
Источник: построено автором

## Компания «Роснефть». Показатели внимания и настроения



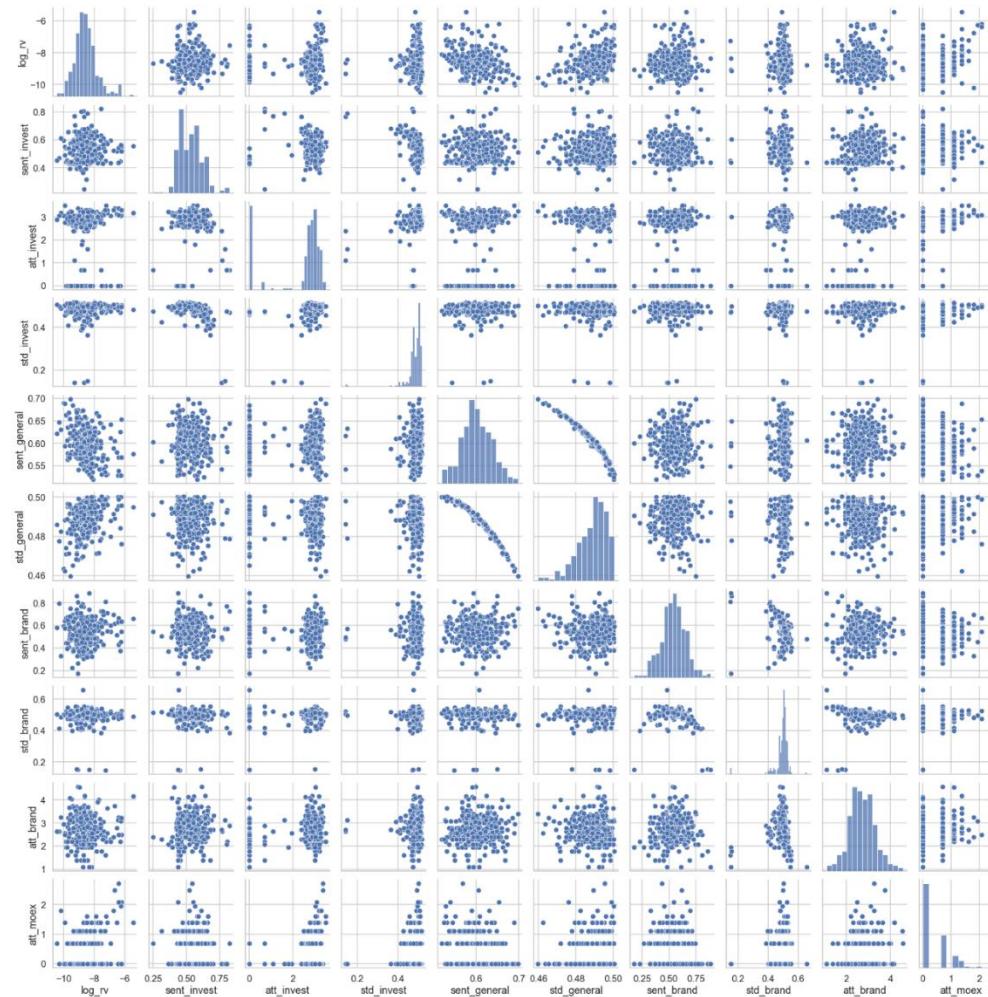
Источник: построено автором

## Компания «Газпром». Экономические и финансовые показатели



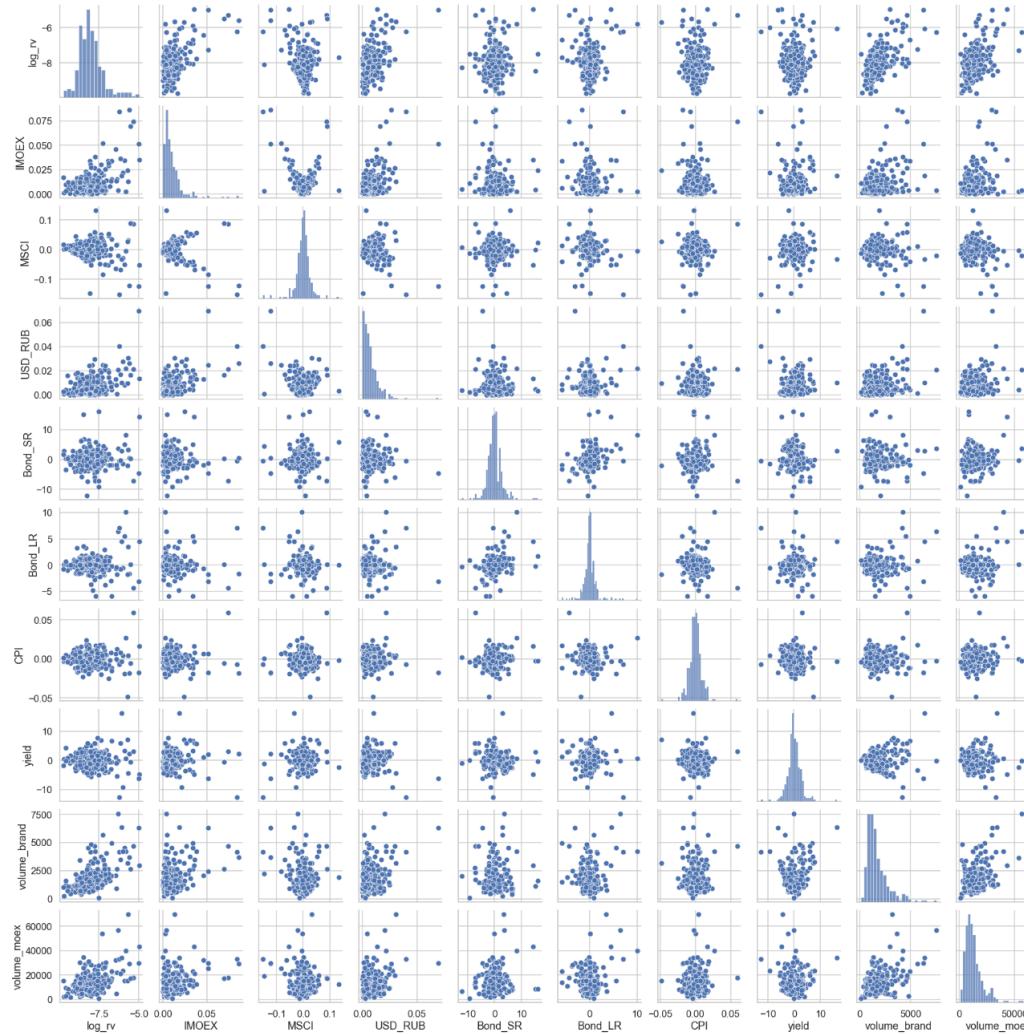
Источник: построено автором

## Компания «Газпром». Показатели внимания и настроения



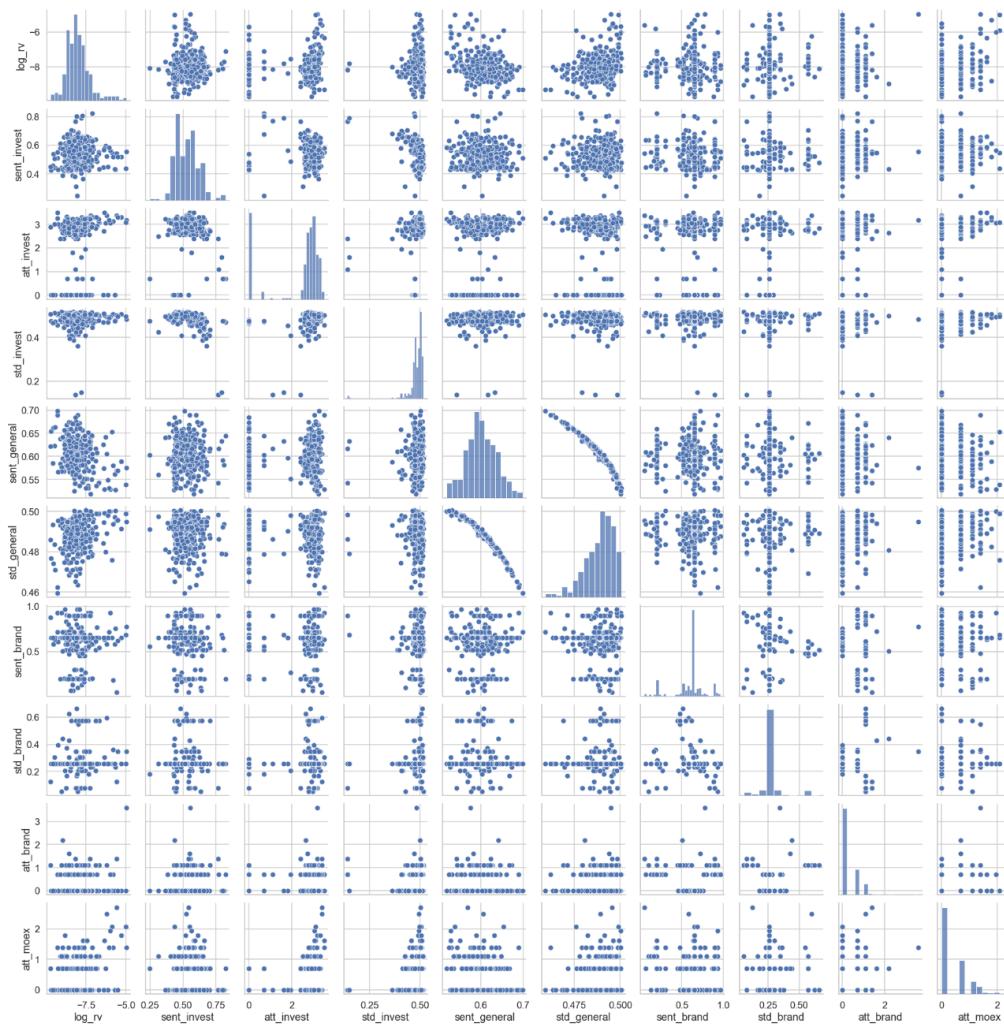
Источник: построено автором

## Компания «Новатэк». Экономические и финансовые показатели



Источник: построено автором

## Компания «Новатэк». Показатели внимания и настроения



Источник: построено автором

## ПРИЛОЖЕНИЕ 5

Коэффициенты для регрессии лассо в экономической модели (А) и модели настроения (Б) для каждой рассматриваемой компании. По строкам расположены факторы, по столбцам компаний. Обозначения признаков аналогичны как в приложении 1. Салатовым цветом выделены отобранные регуляризацией признаки

### А. Экономическая модель

<b>Компания</b>	<b>Яндекс</b>	<b>Сбербанк</b>	<b>МТС</b>	<b>Лукойл</b>	<b>Роснефть</b>	<b>Газпром</b>	<b>Новатэк</b>
<b>intercept</b>	-7,8585	-8,6068	-8,7931	-8,3548	-8,4216	-8,5753	-8,0076
<b>log_rv_d</b>	0,3377	0,1661	0,2408	0,2847	0,2431	0,2286	0,2743
<b>log_rv_w</b>	0,1128	0,2016	0,3611	0,2172	0,2332	0,1525	0,1643
<b>log_rv_m</b>	0,0470	0,0074	0,0034	0,0627	0,0473	0,0741	0,0303
<b>IMOEX</b>	0,0574	0,1796	0,1061	0,1729	0,1873	0,1412	0,0905
<b>MSCI</b>	0,0000	-0,0178	-0,0457	-0,0209	-0,0246	-0,0291	-0,0030
<b>USD_RUB</b>	0,0117	0,0488	0,0372	0,0466	0,0365	0,0444	0,0556
<b>Bond_SR</b>	0,0010	0,0709	0,0631	0,1019	0,0918	0,0782	0,0758
<b>Bond_LR</b>	0,0948	0,0824	0,1013	0,0452	0,0469	0,0509	0,0226
<b>CPI</b>	0,0000	0,0064	0,0000	0,0169	0,0263	-0,0033	-0,0001
<b>yield</b>	0,0040	0,0000	0,0533	0,0342	0,0331	0,0000	-0,0164
<b>volume_brand</b>	0,1243	0,0772	0,0418	0,1553	0,1053	0,1684	0,1676
<b>volume_moex</b>	0,0514	0,0721	0,0500	0,0382	0,0593	0,0001	0,0530

Источник: построено автором

## Б. Модель настроения

Компания	Яндекс	Сбербанк	МТС	Лукойл	Роснефть	Газпром	Новатэк
<b>intercept</b>	-2,9016	-3,3681	-2,2435	-2,8167	-3,8537	-3,0231	-3,1909
<b>log_rv_d</b>	0,4279	0,3204	0,3489	0,4232	0,3293	0,3555	0,3897
<b>log_rv_w</b>	0,1843	0,3280	0,4344	0,2678	0,2704	0,2274	0,2403
<b>log_rv_m</b>	0,0712	0,0000	0,0000	0,0236	0,0000	0,1133	0,0223
<b>IMOEX</b>	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
<b>MSCI</b>	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
<b>USD_RUB</b>	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
<b>Bond_SR</b>	0,0000	0,0284	0,0276	0,0405	0,0368	0,0312	0,0301
<b>Bond_LR</b>	0,0561	0,0473	0,0636	0,0265	0,0209	0,0282	0,0107
<b>CPI</b>	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
<b>yield</b>	0,0042	-0,0046	0,0432	0,0131	0,0129	0,0010	-0,0117
<b>volume_brand</b>	0,0001	0,0000	0,0000	0,0002	0,0000	0,0000	0,0002
<b>volume_moex</b>	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
<b>sent_invest</b>	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
<b>att_invest</b>	0,0492	0,0000	0,0351	0,0000	0,0262	0,0077	0,0000
<b>std_invest</b>	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
<b>sent_general</b>	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
<b>sent_brand</b>	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
<b>std_brand</b>	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
<b>att_brand</b>	0,0000	0,0000	0,0000	0,0912	0,0524	0,0000	0,0262
<b>att_moex</b>	0,0000	0,0257	0,0457	0,0036	0,0783	0,0499	0,0027

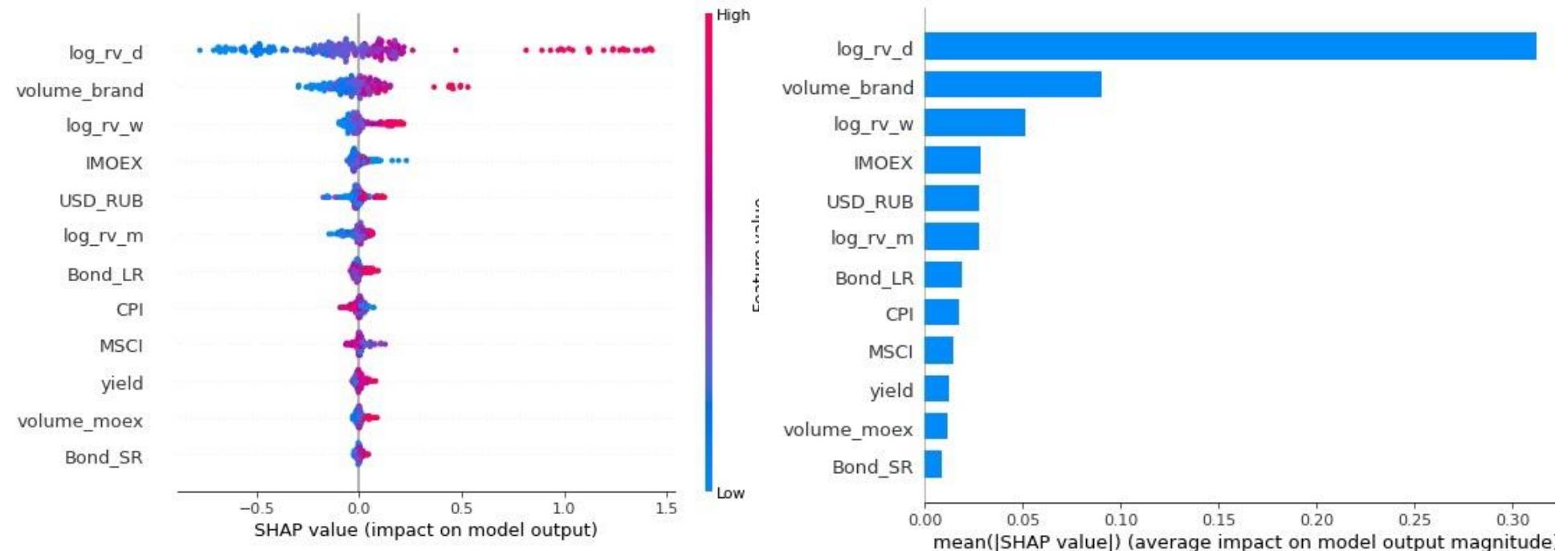
Источник: построено автором

## ПРИЛОЖЕНИЕ 6

Линейчатые диаграммы, отображающие важность признаков методом SHAP для экономической модели (А) и модели настроения (Б); для всех алгоритмов, основанных на решающих деревьях; для каждой компании. На графиках слева каждая точка обозначает одно наблюдение и значение вклада признака для данного объекта. На графиках справа отображены амплитуды, соответствующие важности признаков

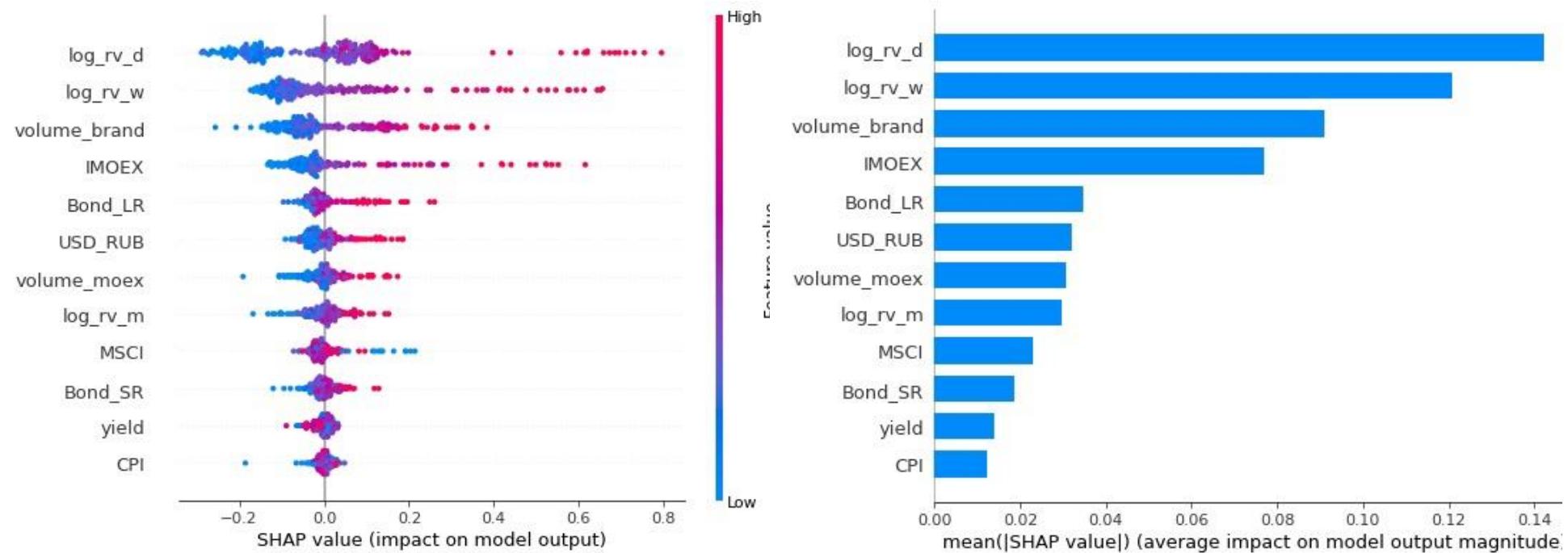
## A1. Экономическая модель. Случайный лес

Компания «Яндекс»



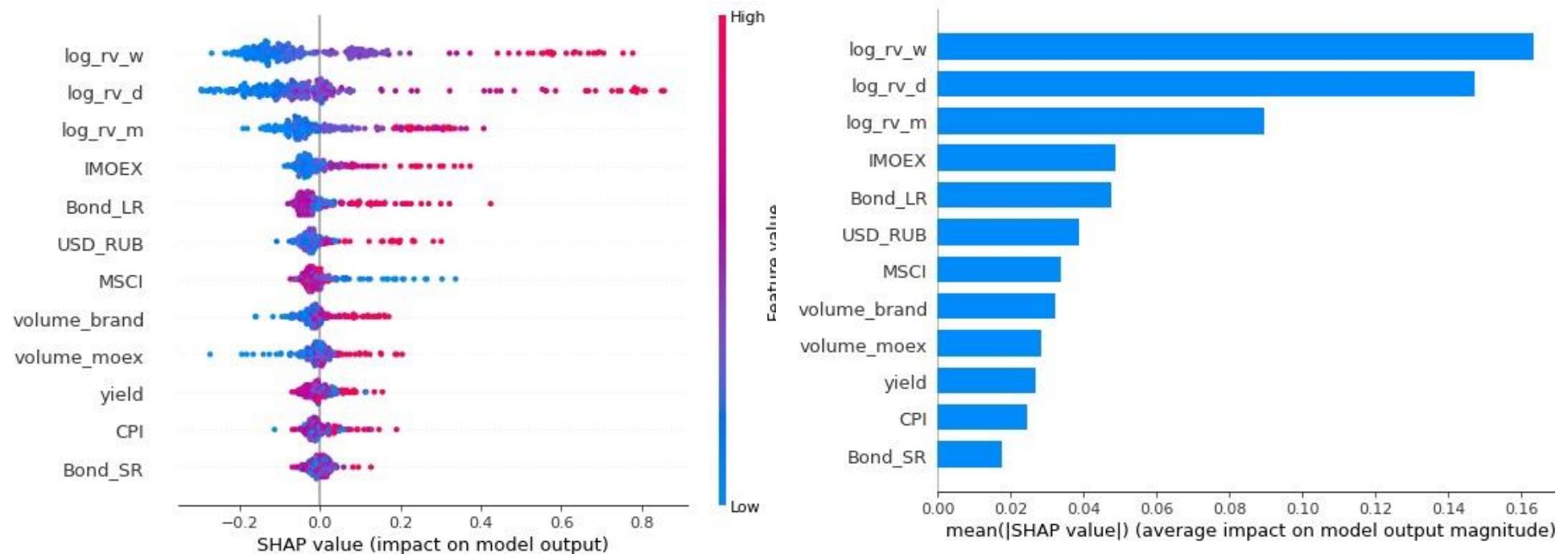
Источник: построено автором

## Компания «Сбербанк»



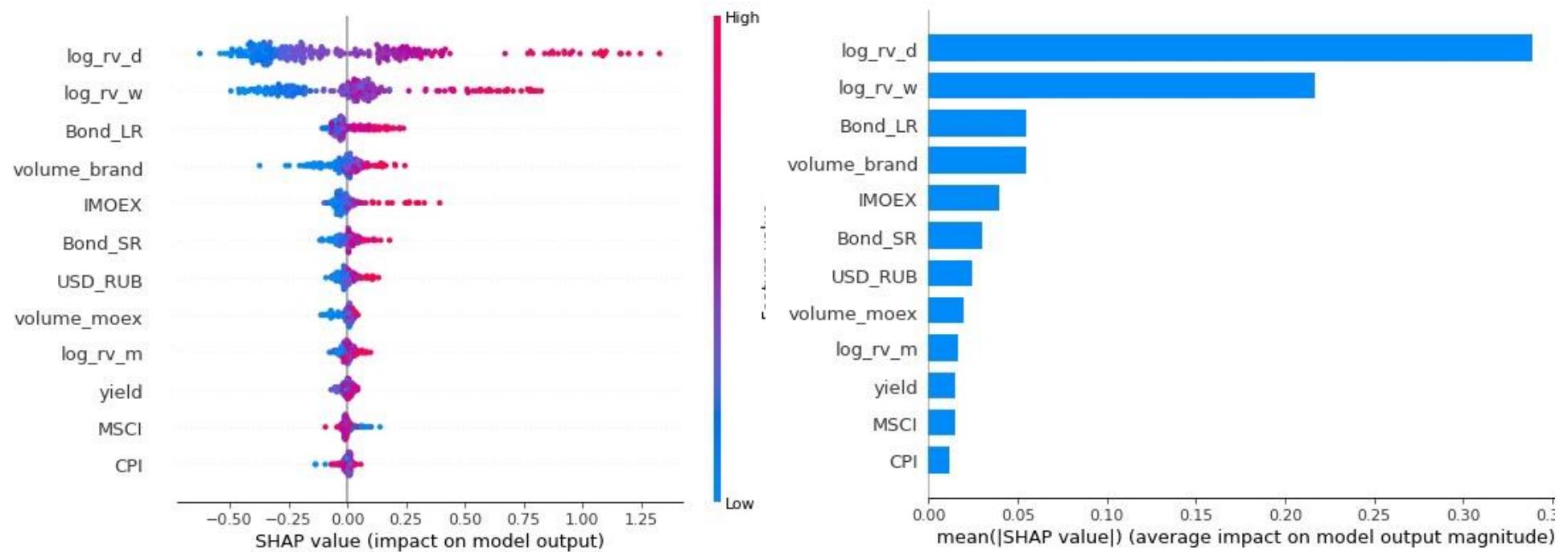
Источник: построено автором

## Компания «МТС»



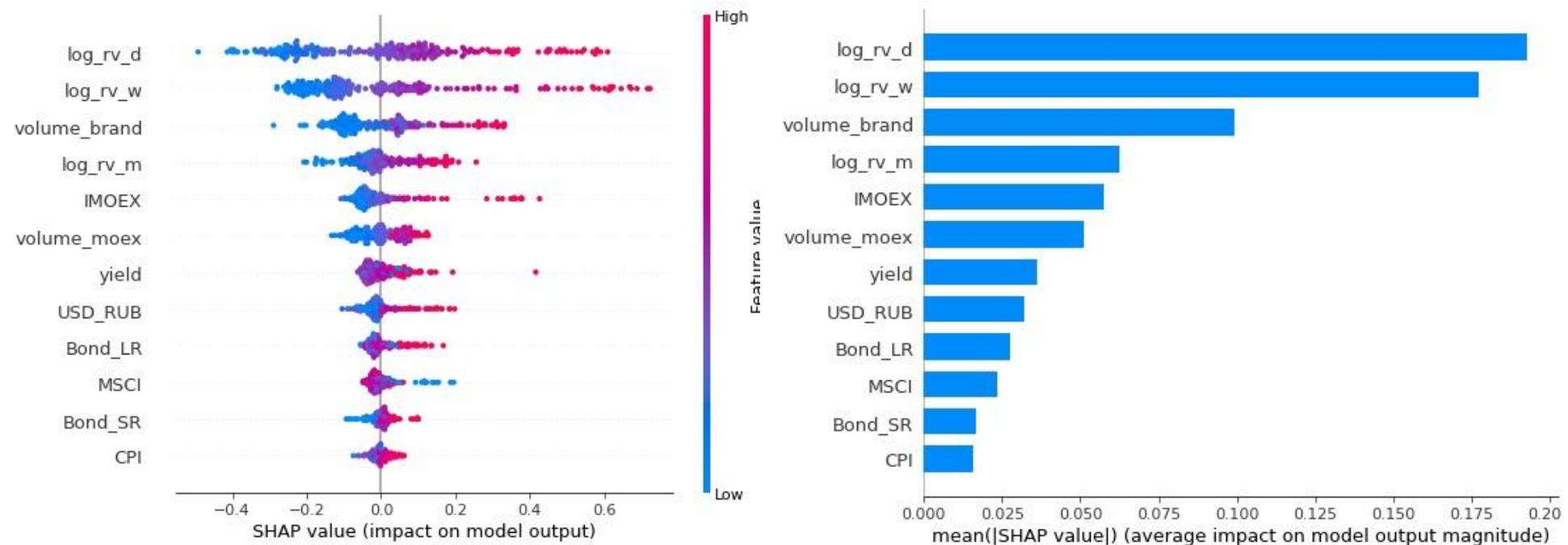
Источник: построено автором

## Компания «Лукойл»



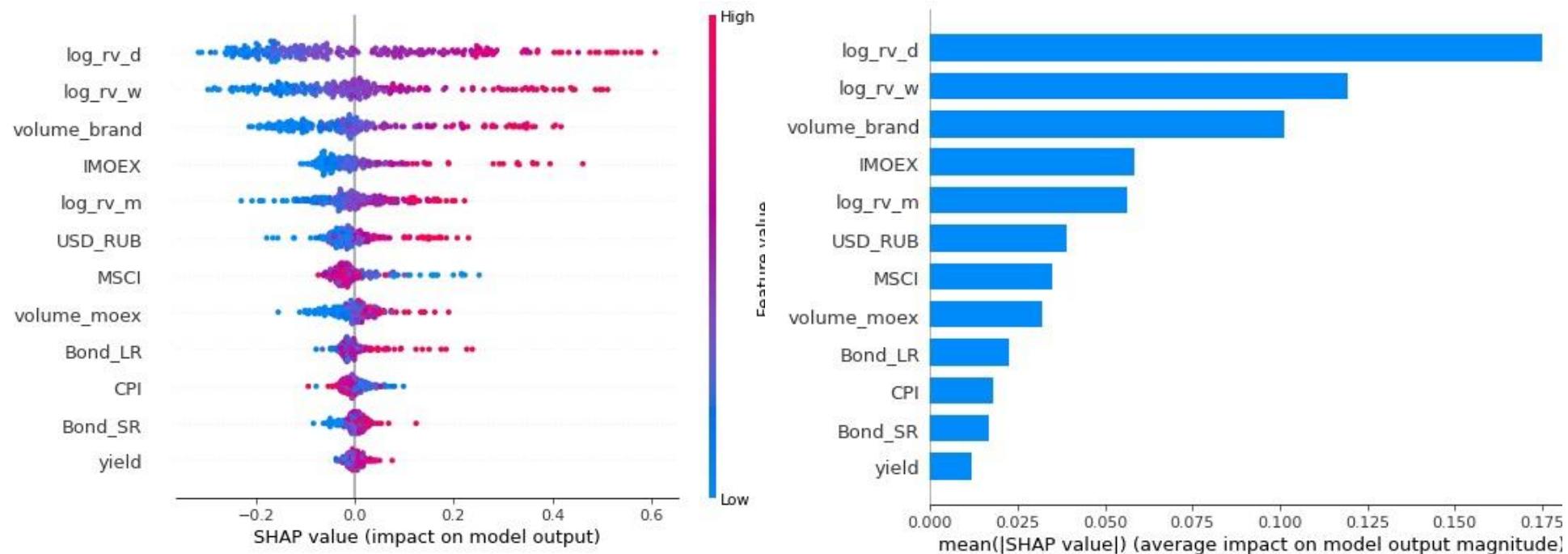
Источник: построено автором

## Компания «Роснефть»



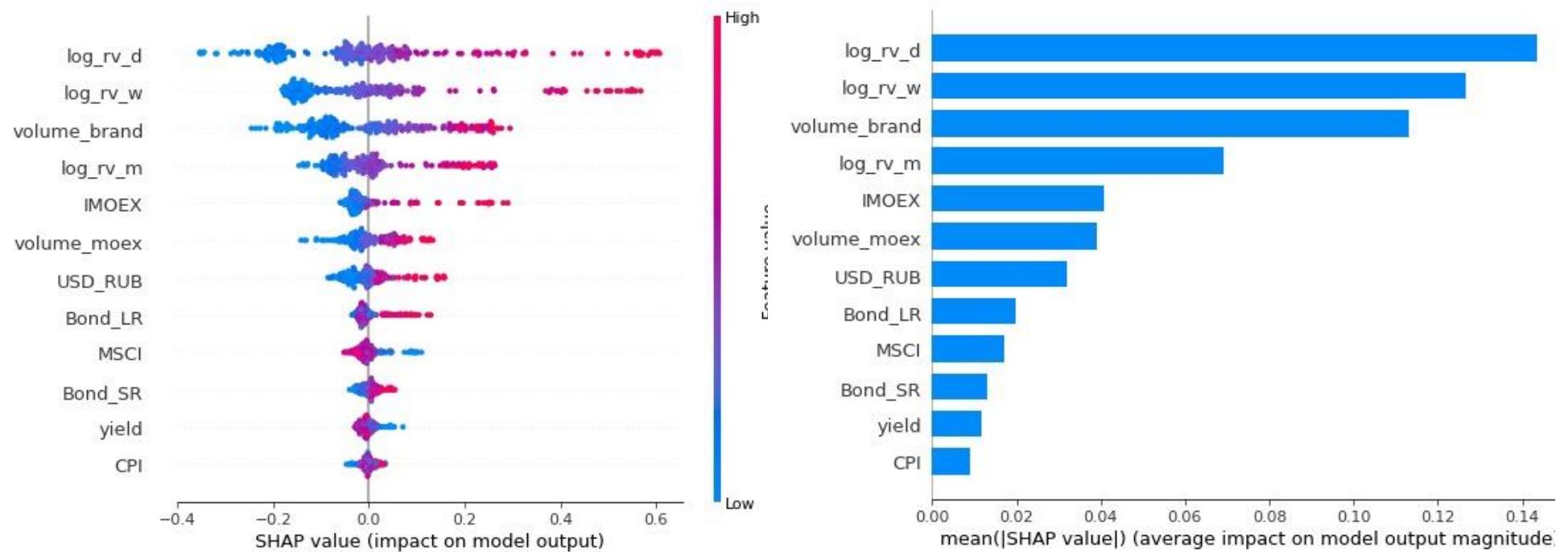
Источник: построено автором

## Компания «Газпром»



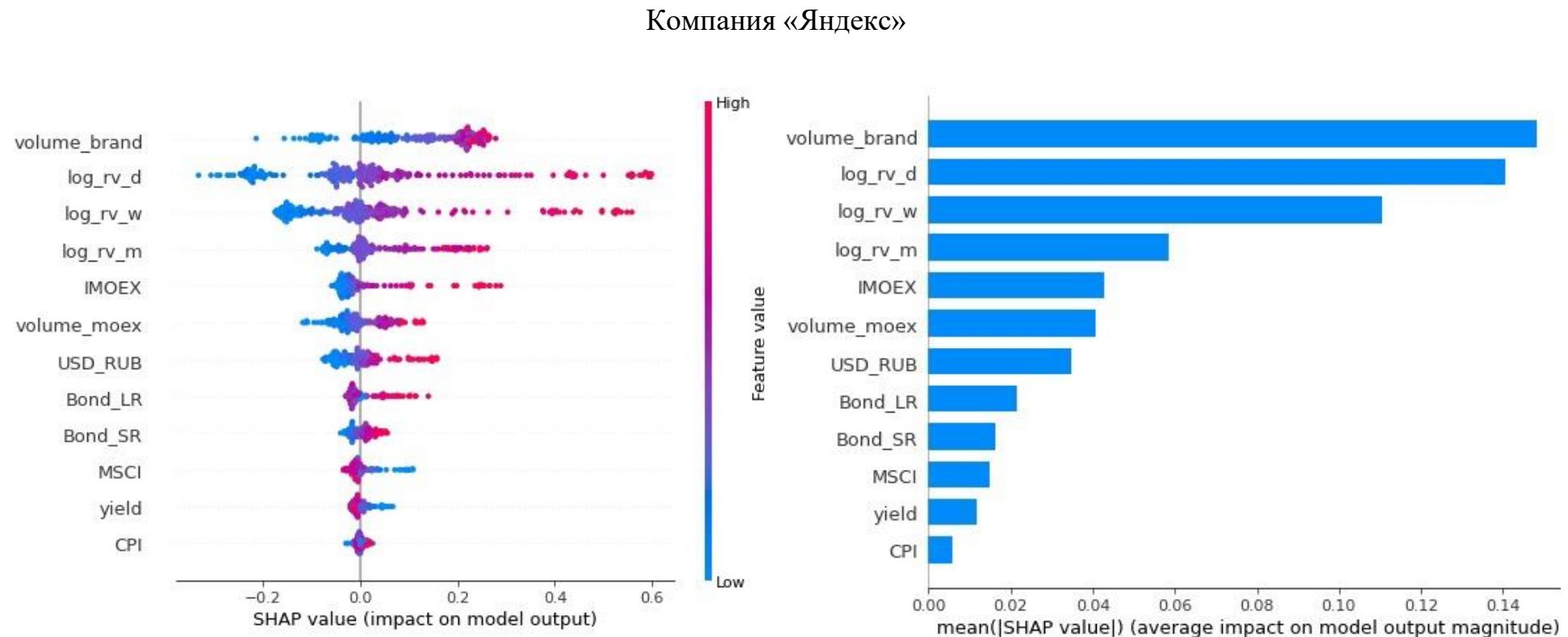
Источник: построено автором

## Компания «Новатэк»



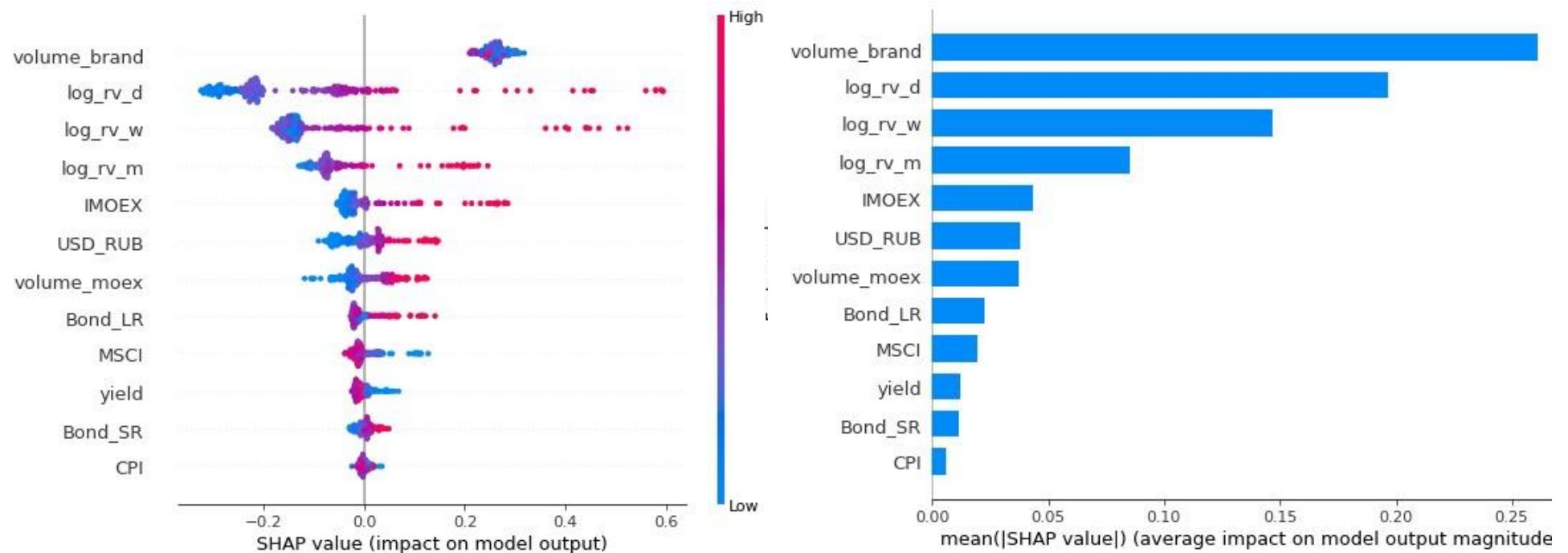
Источник: построено автором

## A2. Экономическая модель. XGBoost



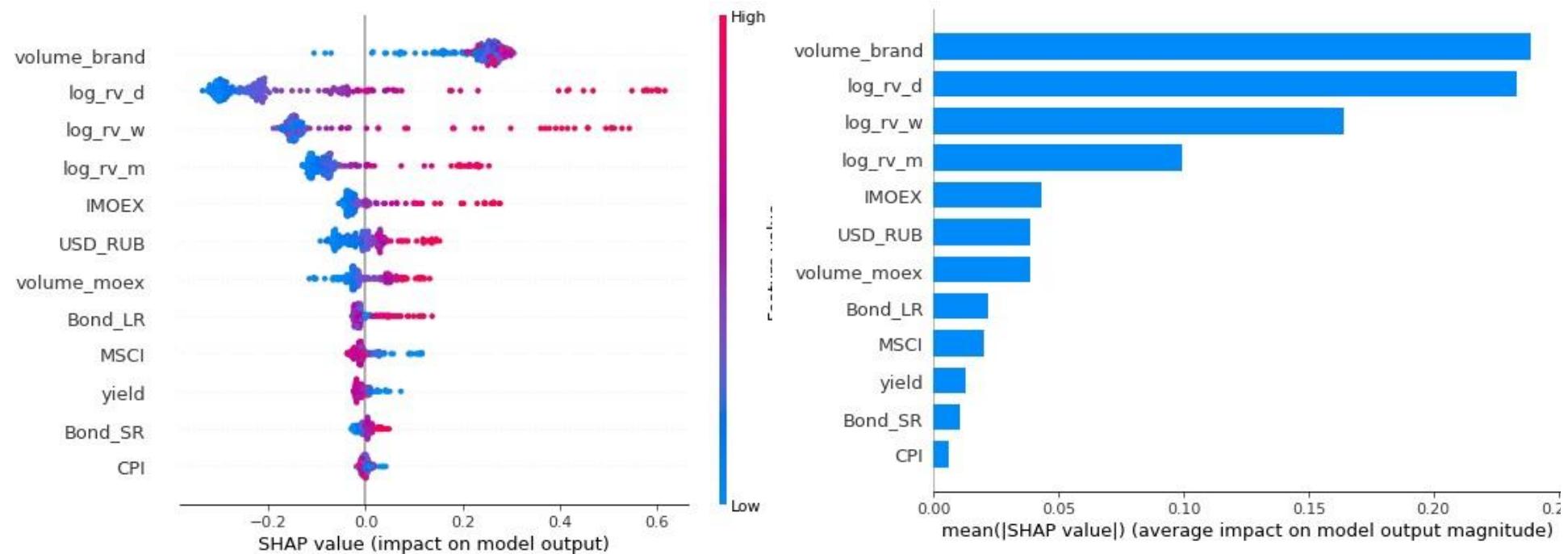
Источник: построено автором

## Компания «Сбербанк»



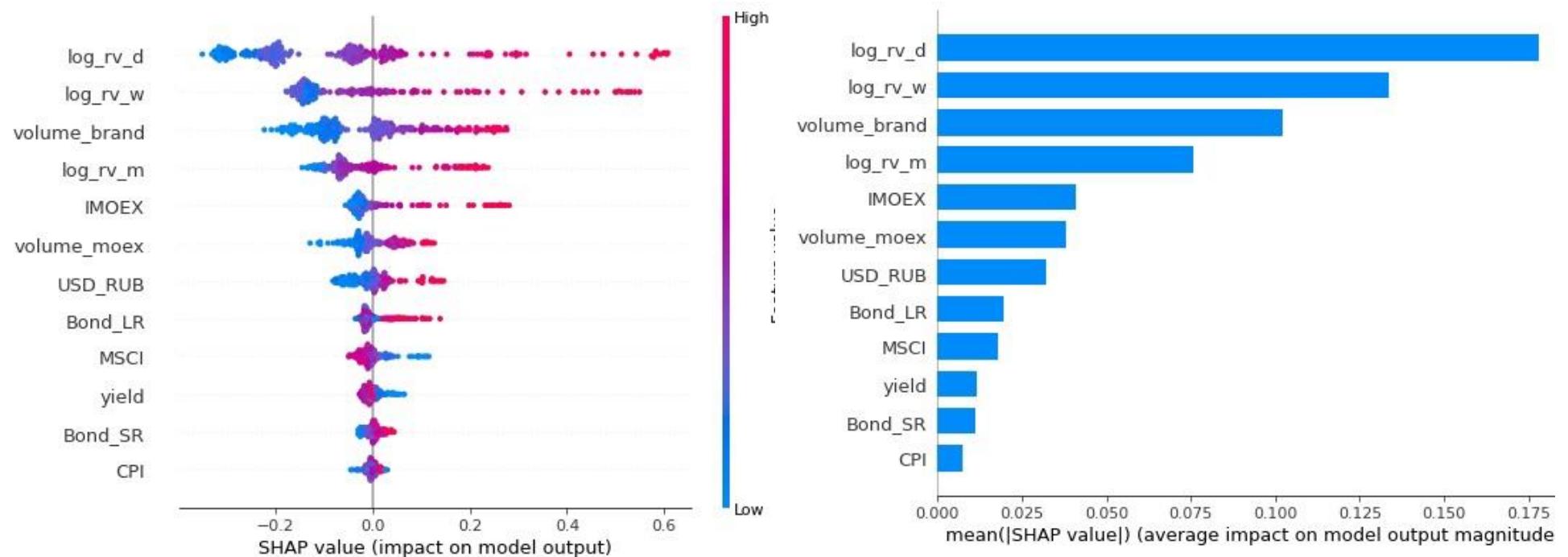
Источник: построено автором

## Компания «МТС»



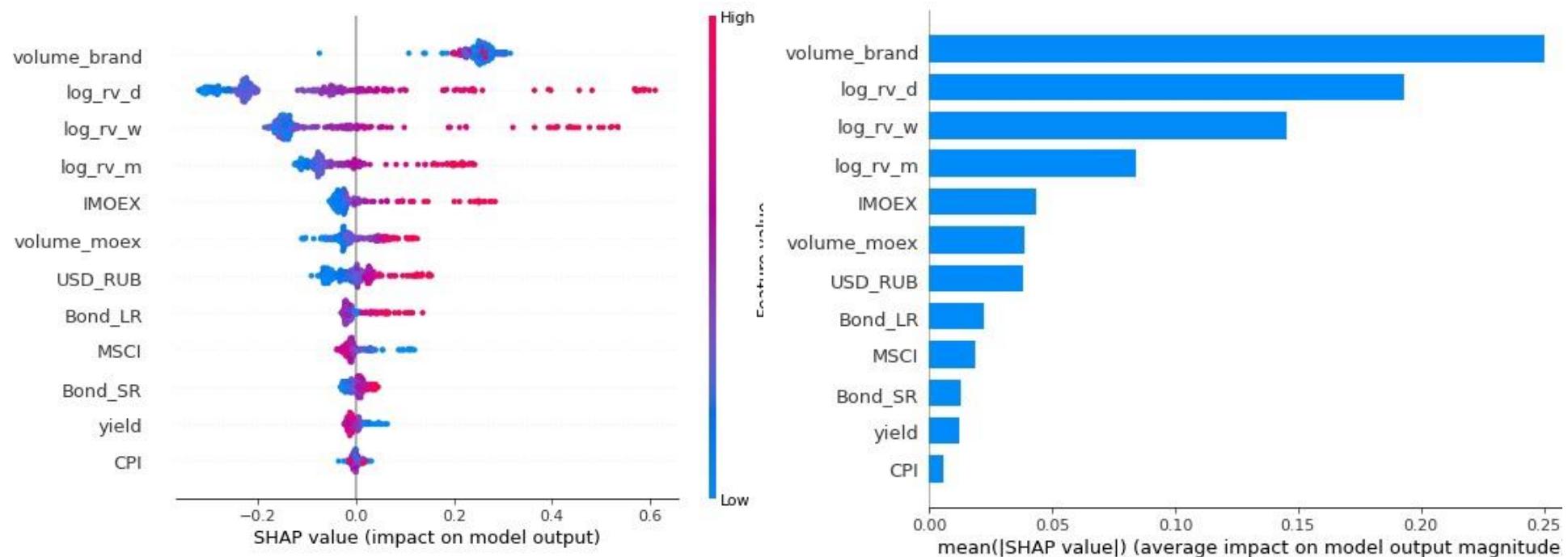
Источник: построено автором

## Компания «Лукойл»



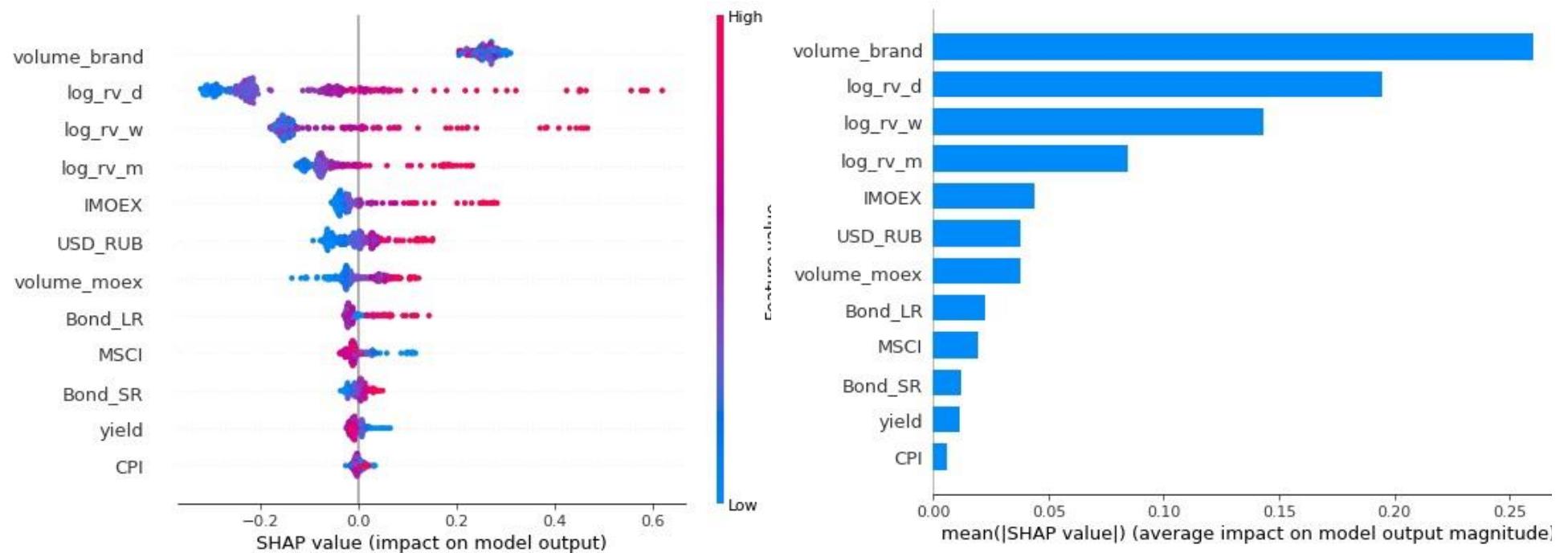
Источник: построено автором

## Компания «Роснефть»



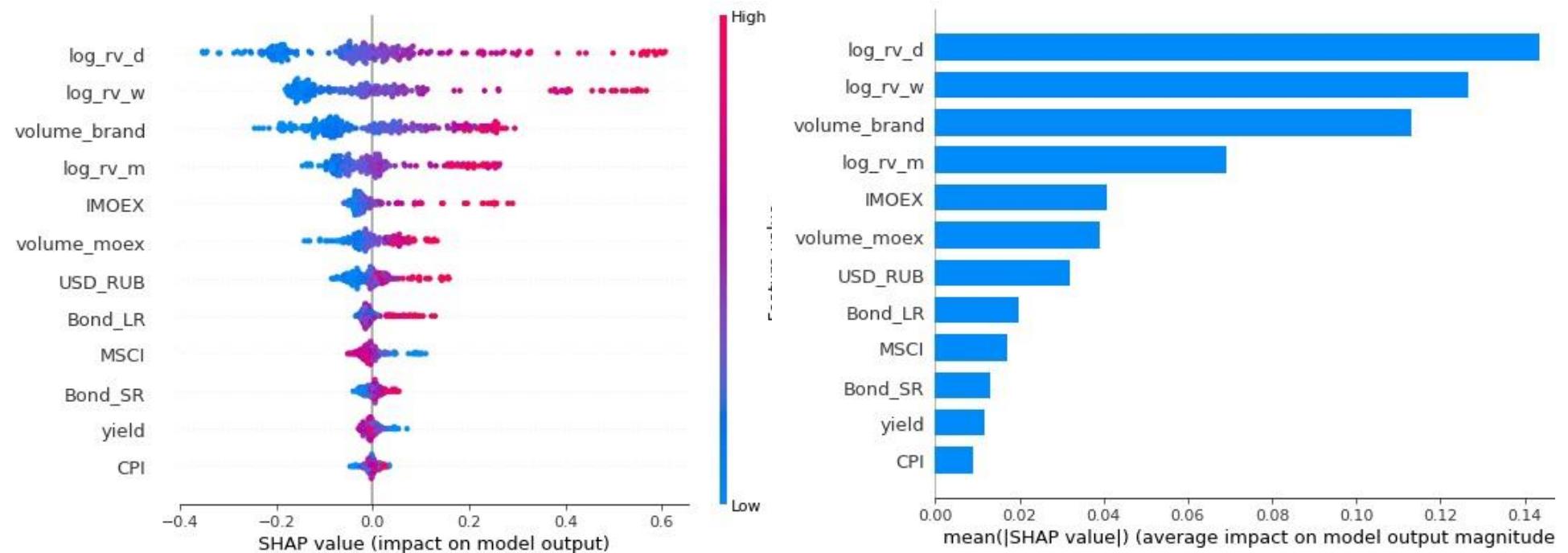
Источник: построено автором

## Компания «Газпром»



Источник: построено автором

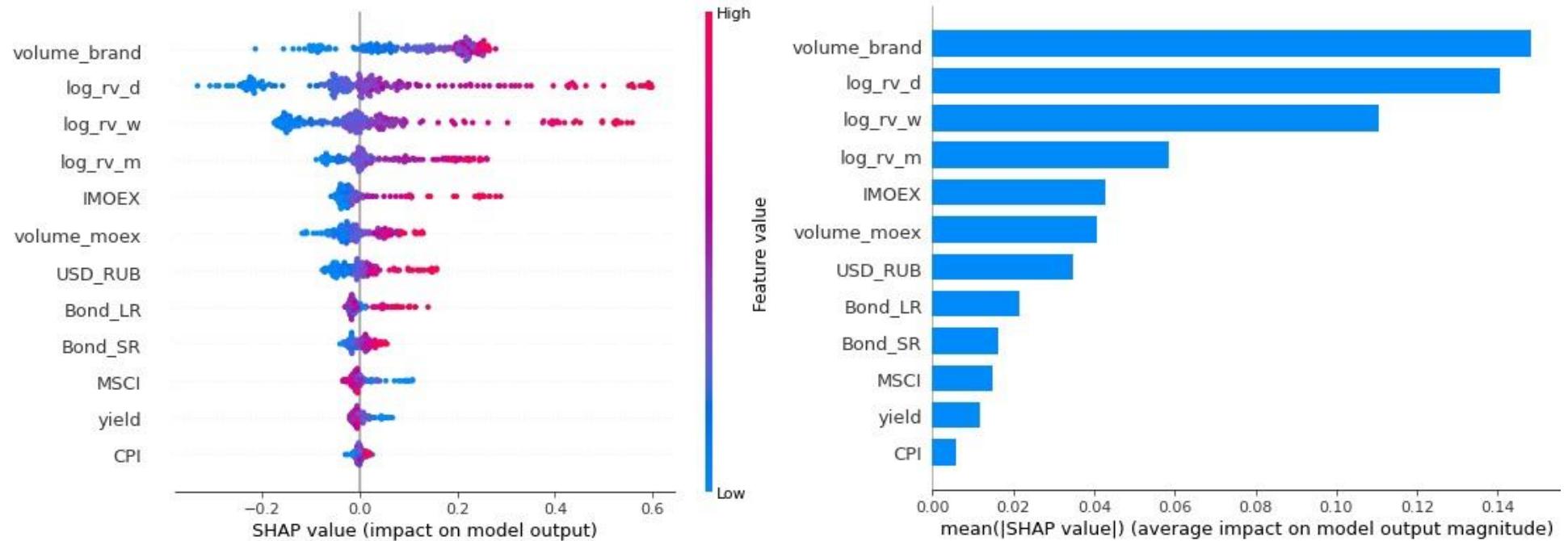
## Компания «Новатэк»



Источник: построено автором

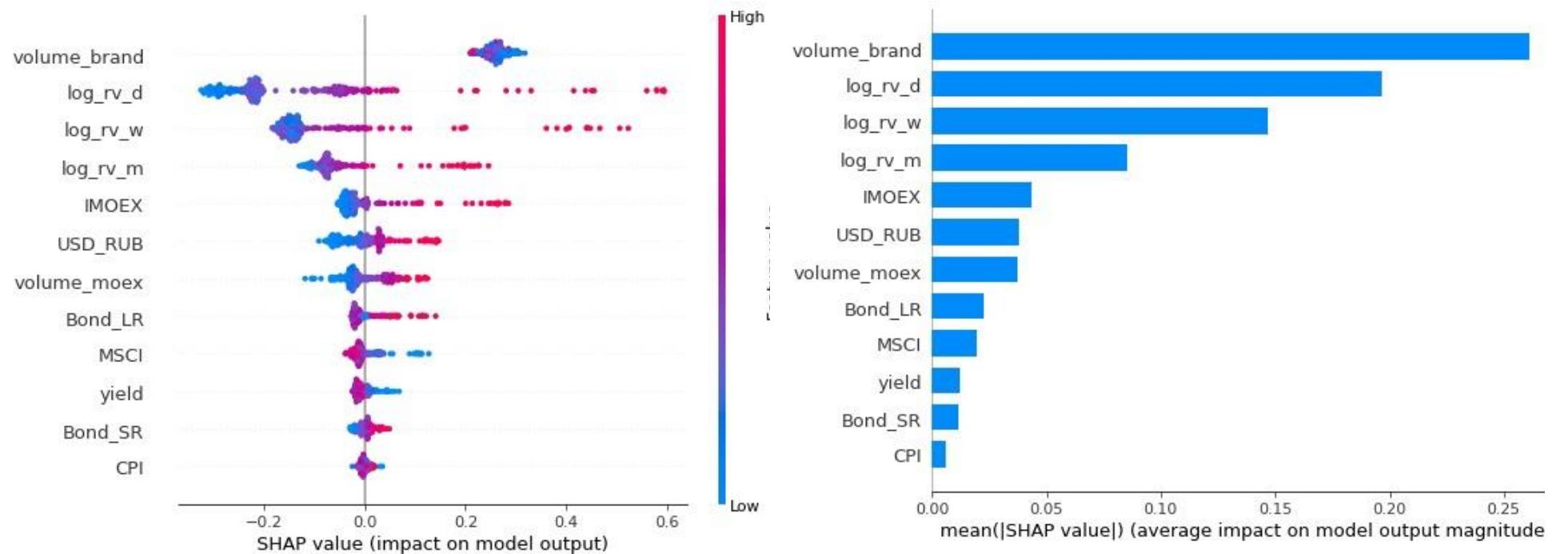
### A3. Экономическая модель. Light GBM

Компания «Яндекс»



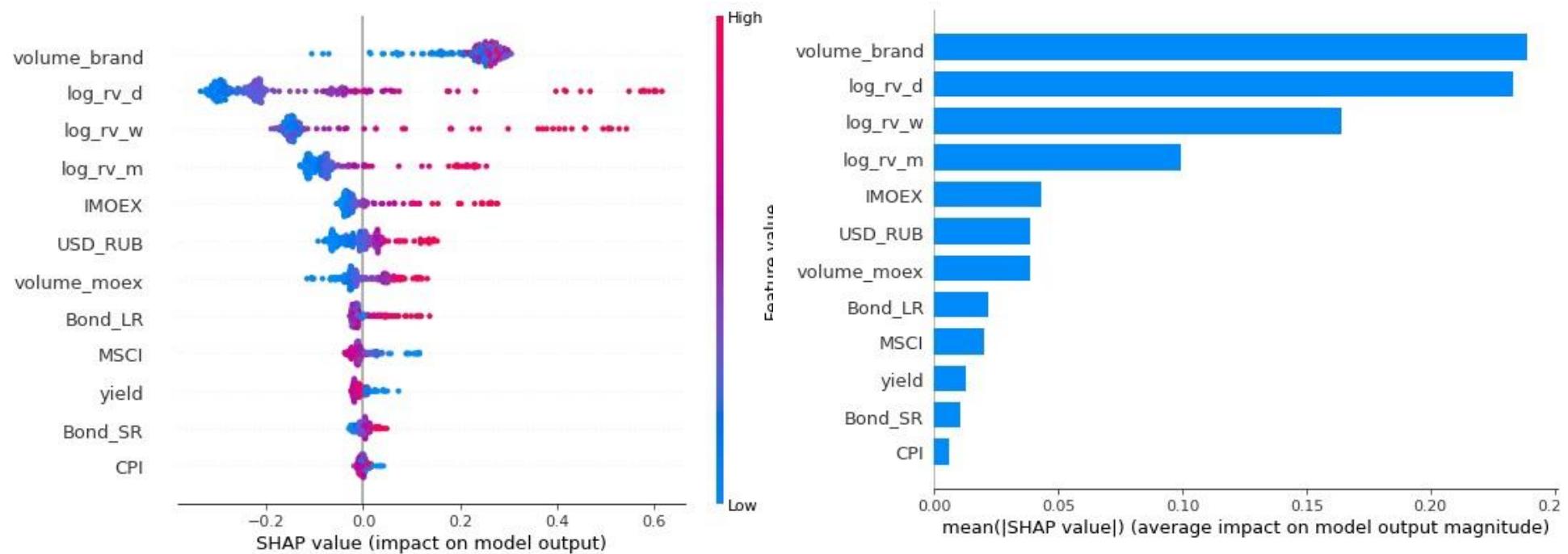
Источник: построено автором

## Компания «Сбербанк»



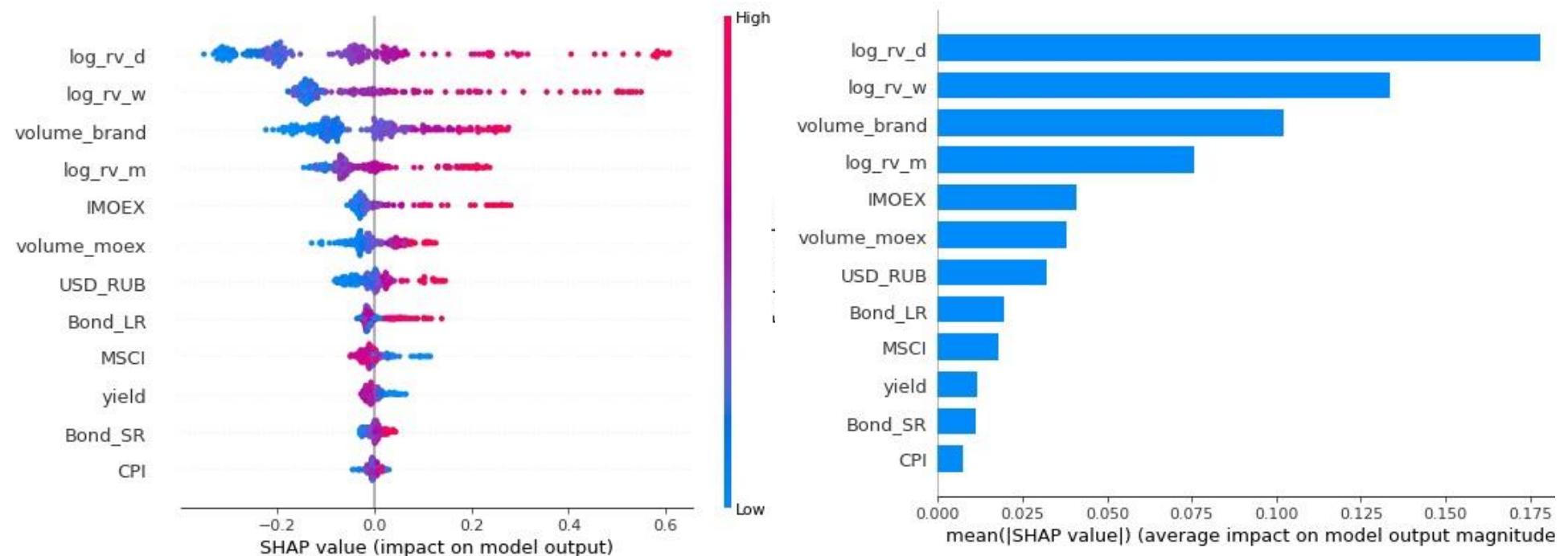
Источник: построено автором

## Компания «МТС»



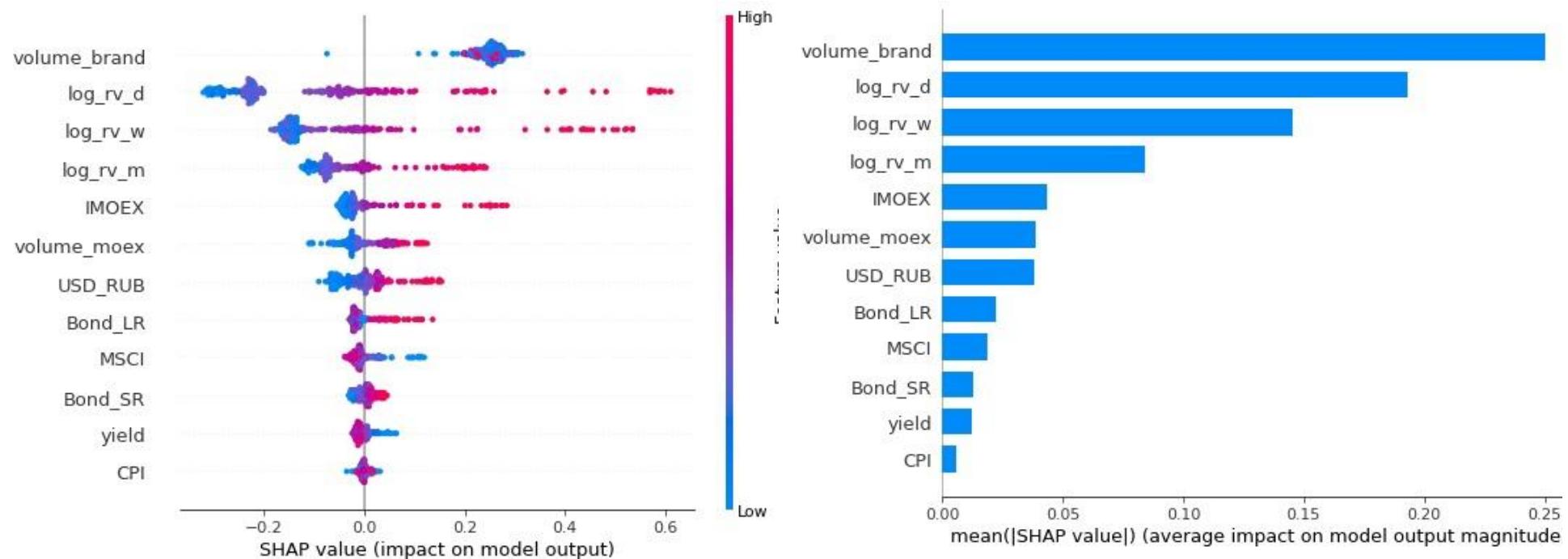
Источник: построено автором

## Компания «Лукойл»



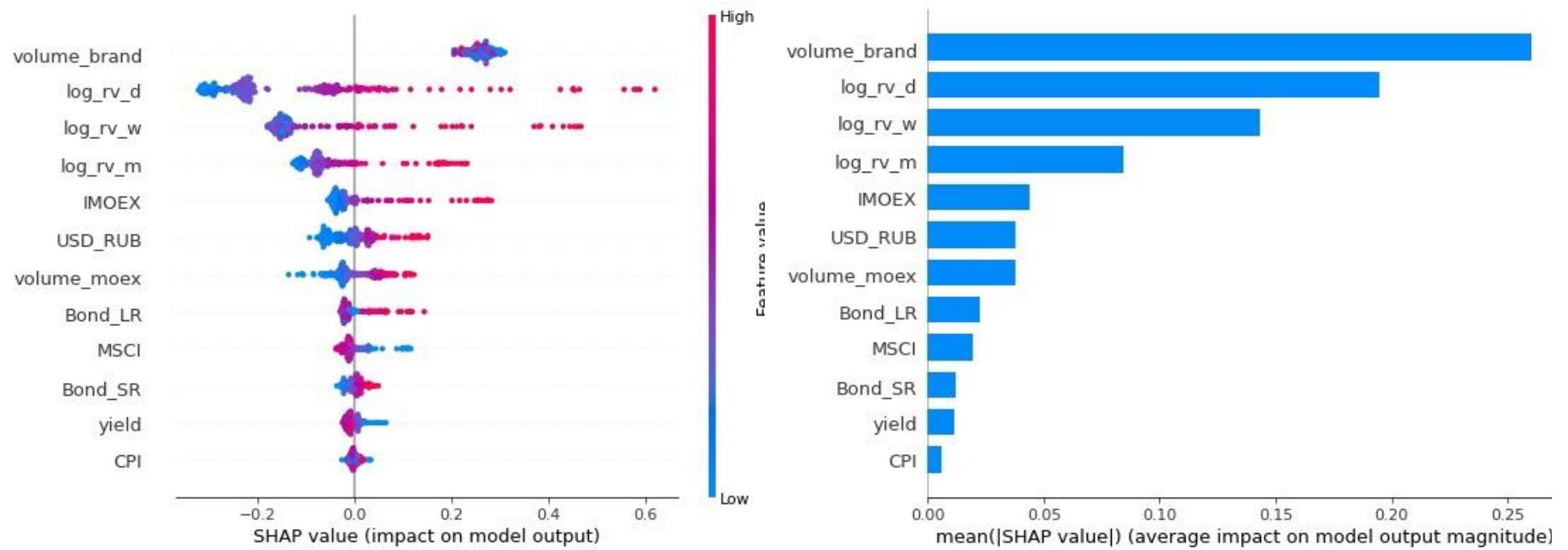
Источник: построено автором

## Компания «Роснефть»



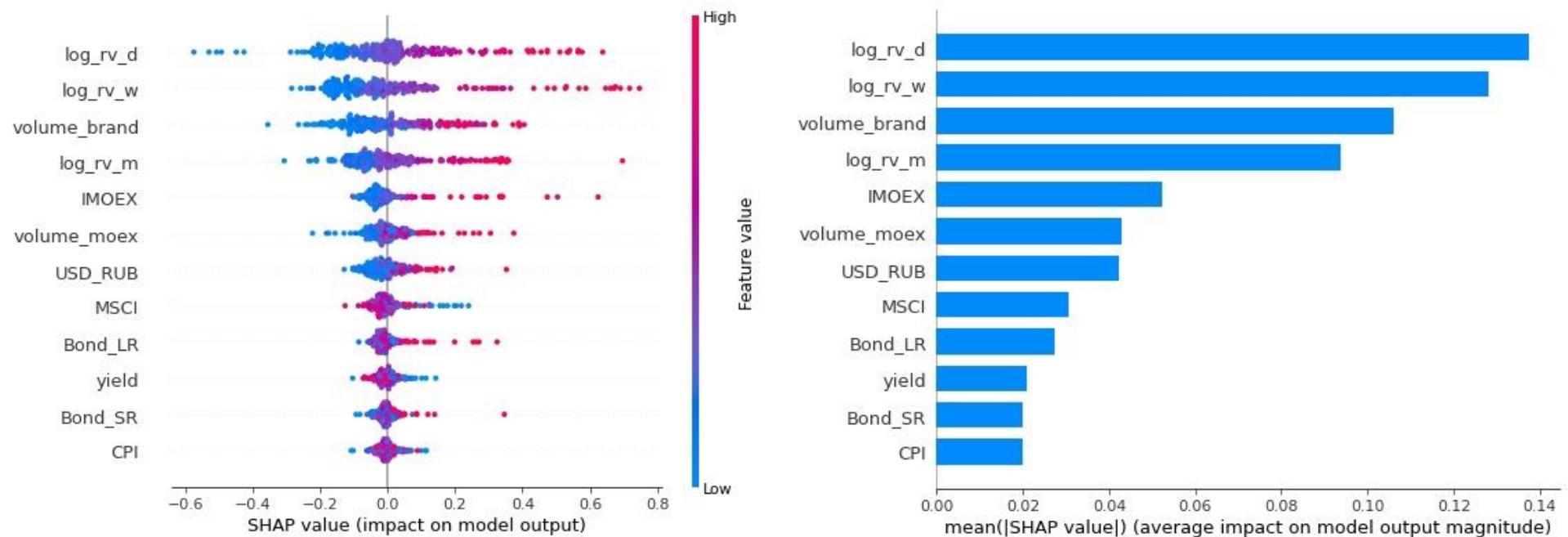
Источник: построено автором

## Компания «Газпром»



Источник: построено автором

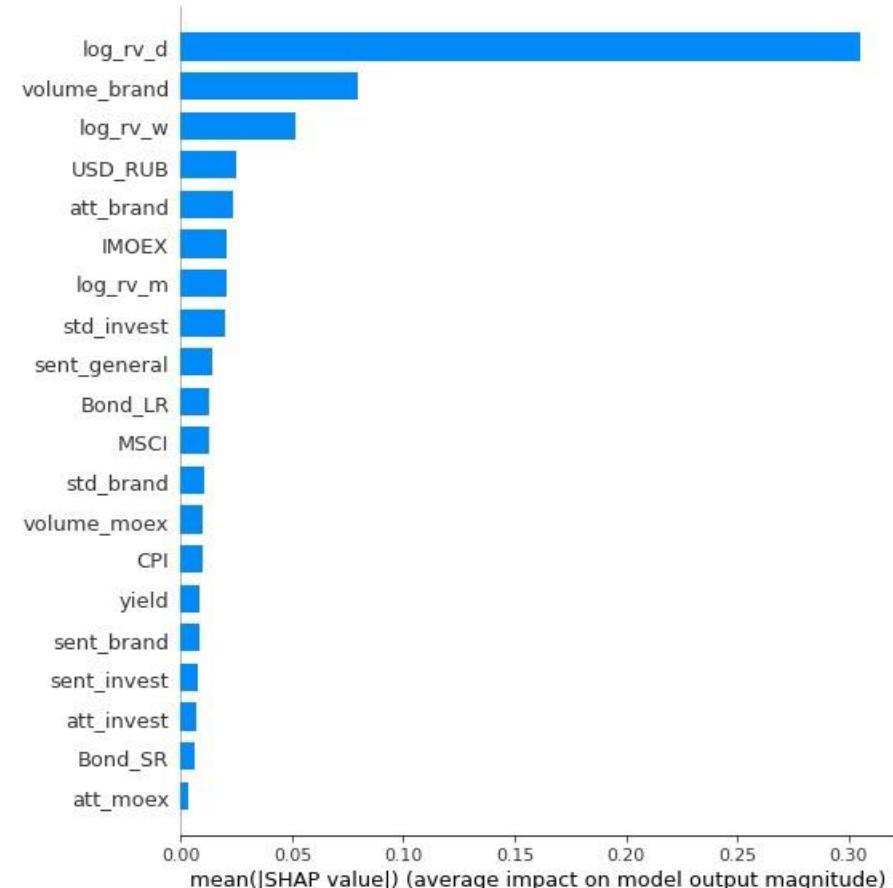
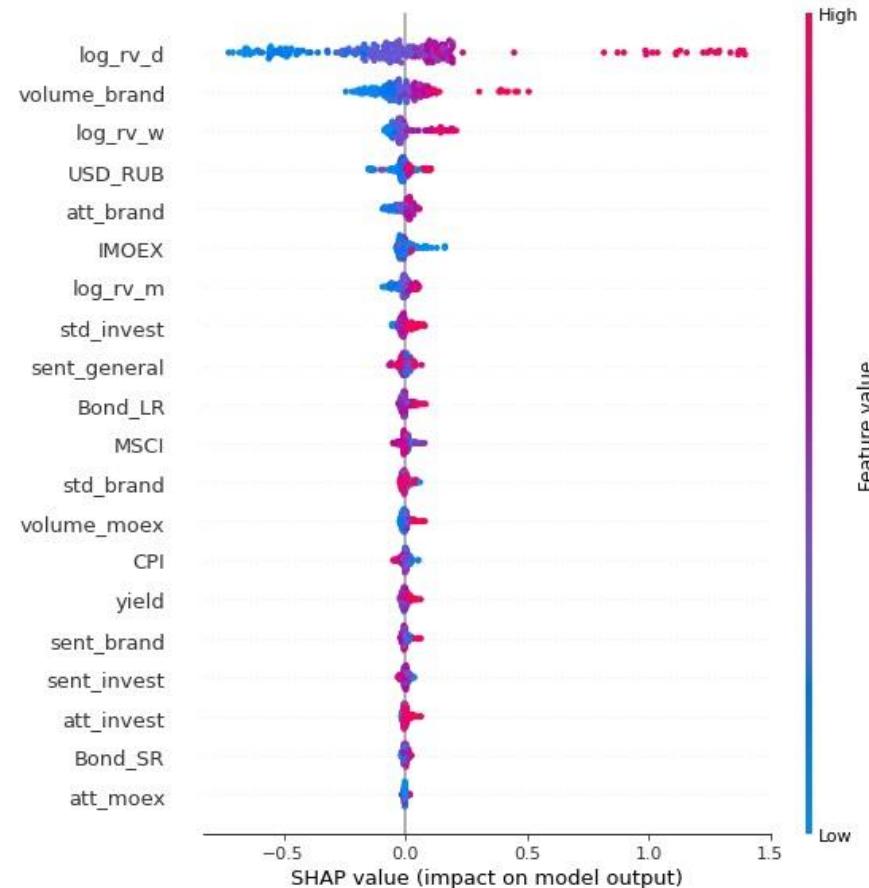
## Компания «Новатэк»



Источник: построено автором

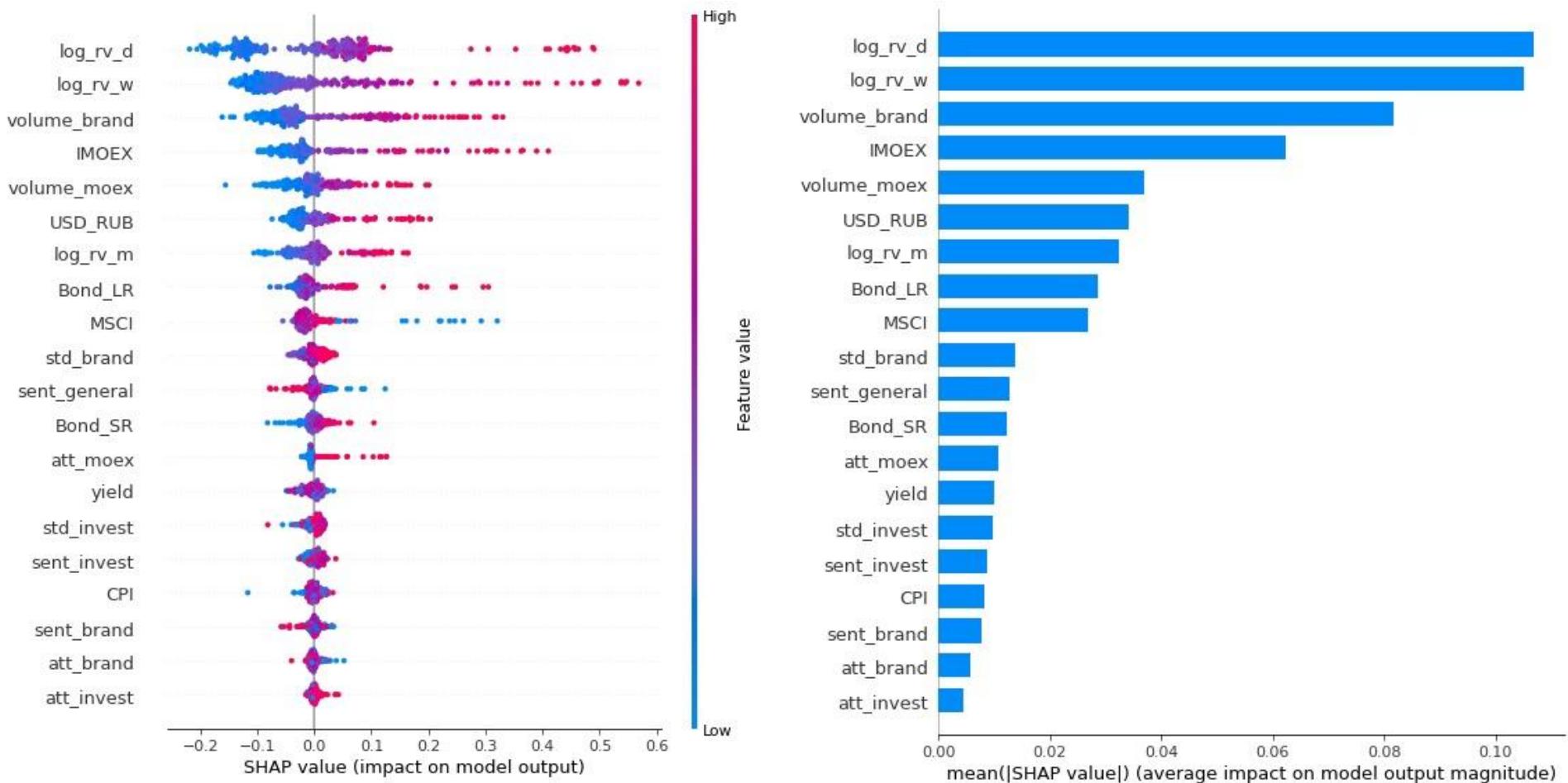
## Б1. Модель настроения. Случайный лес

Компания «Яндекс»



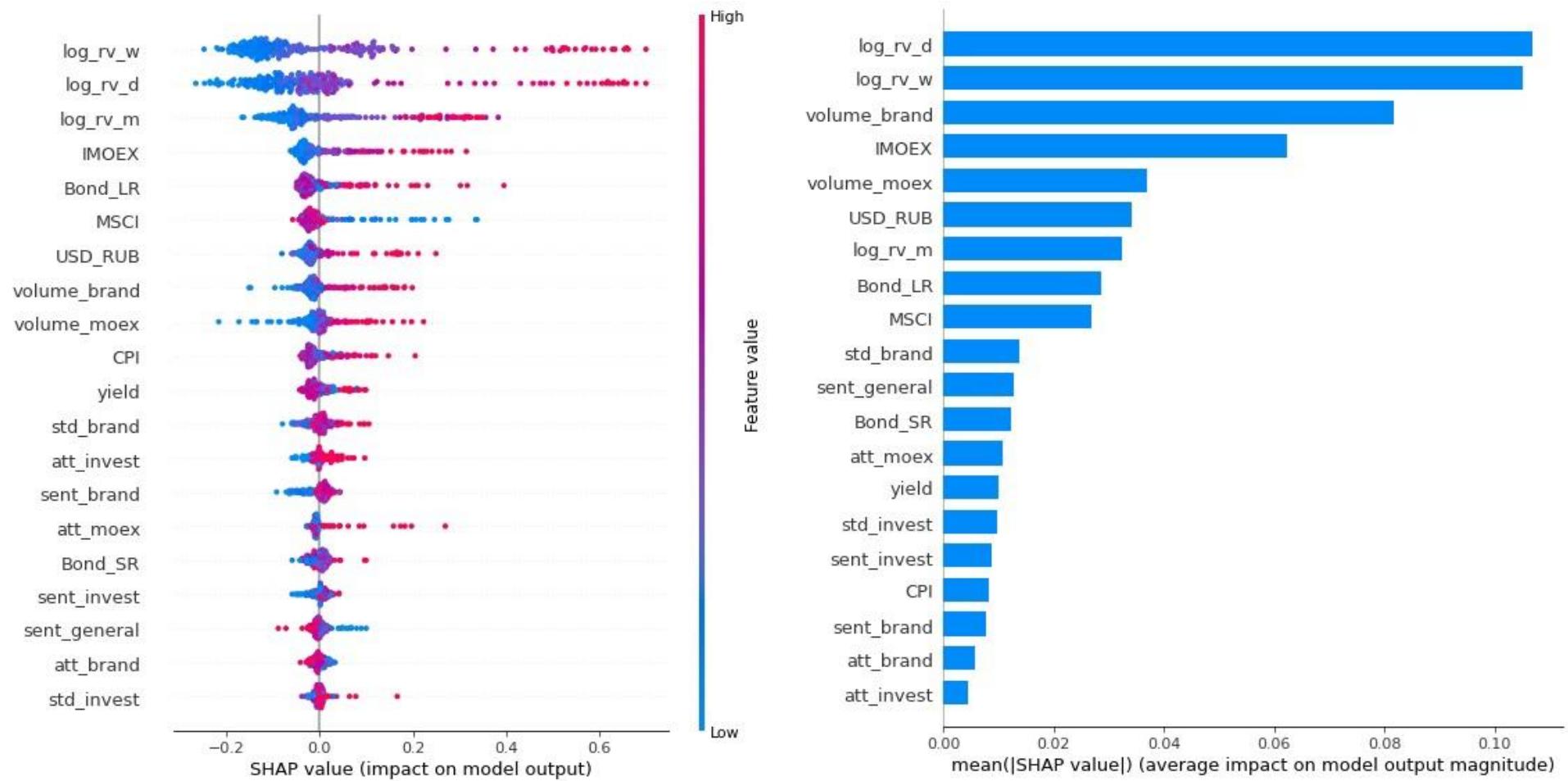
Источник: построено автором

## Компания «Сбербанк»



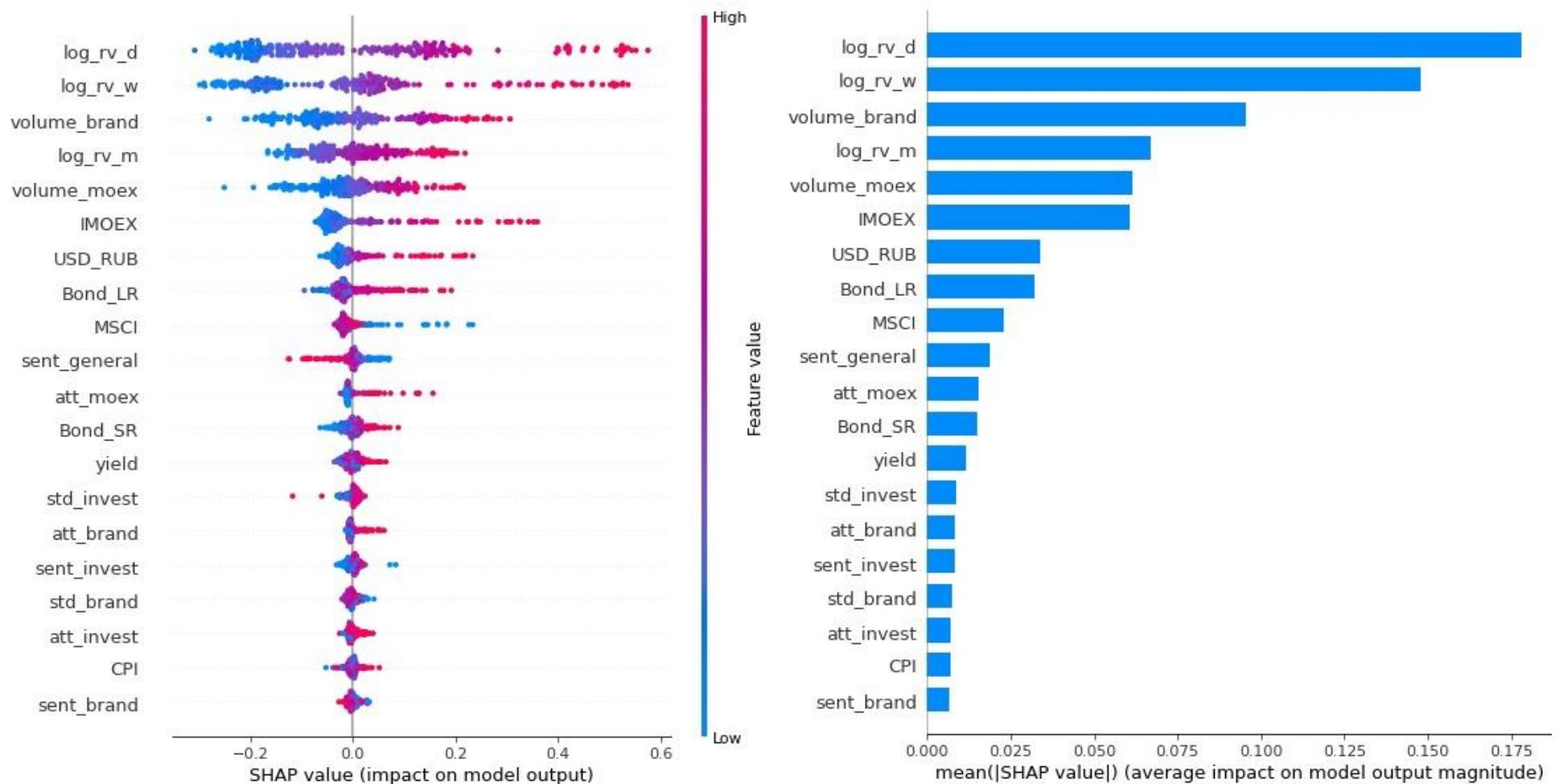
Источник: построено автором

## Компания «МТС»



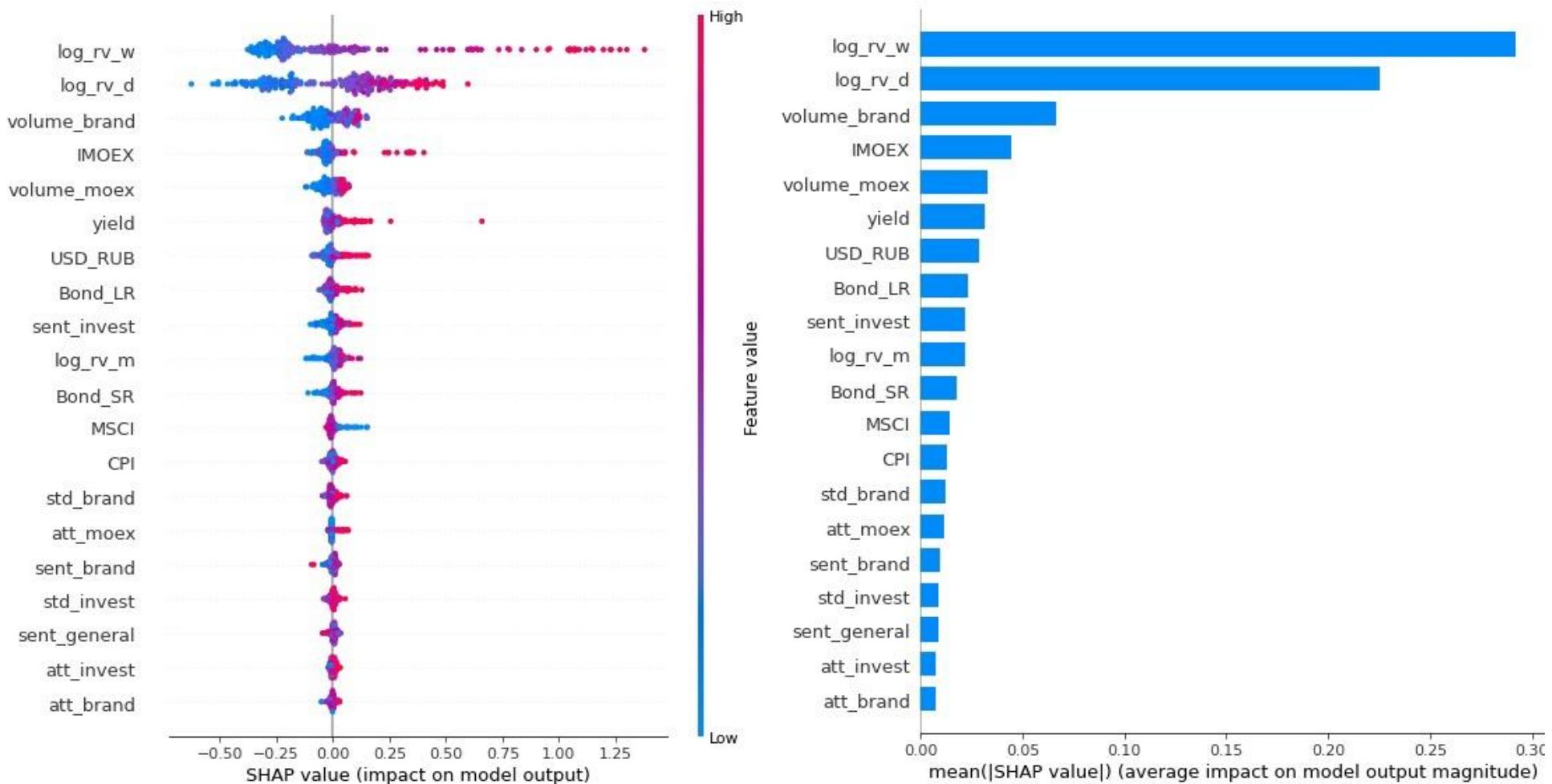
Источник: построено автором

## Компания «Лукойл»



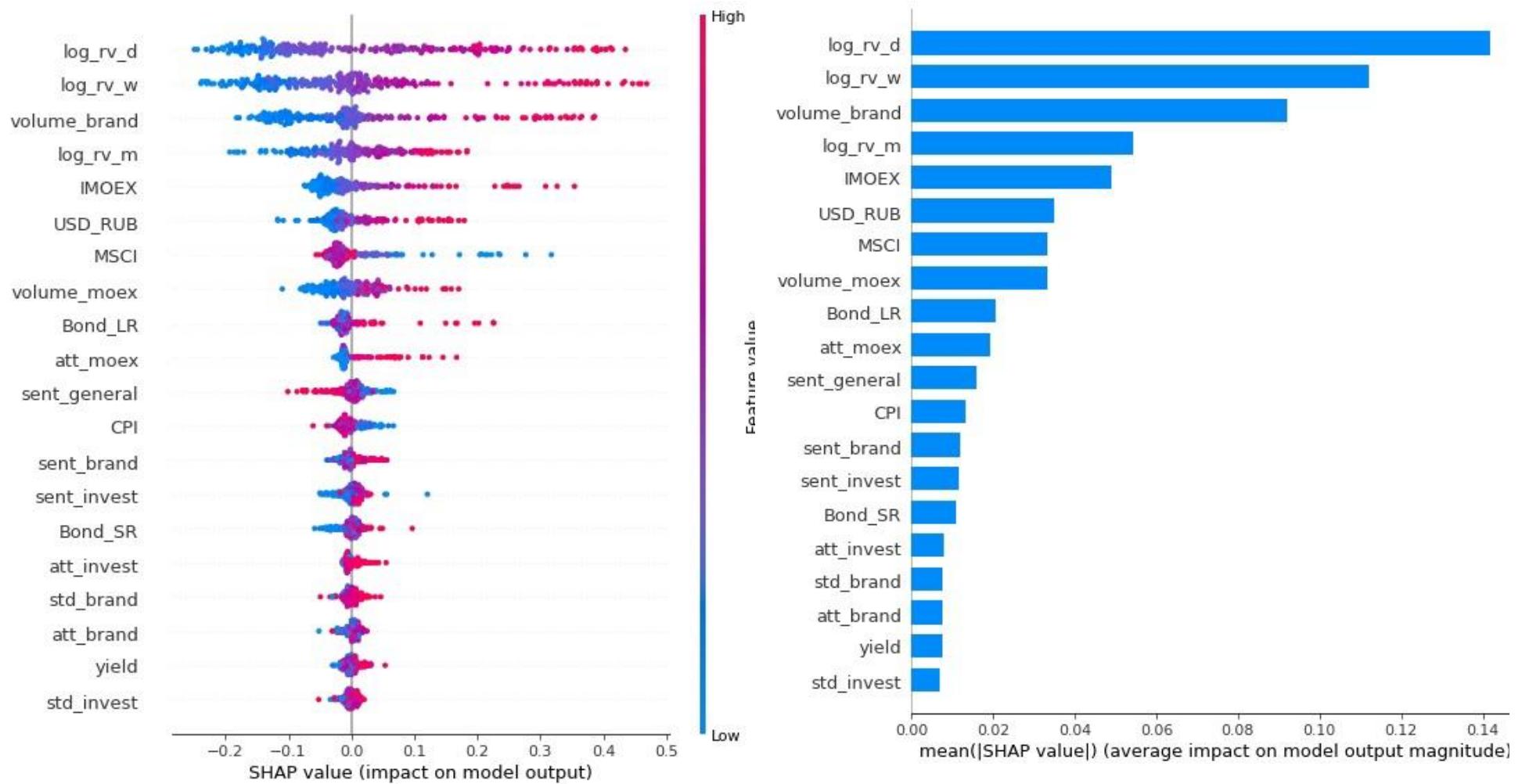
Источник: построено автором

## Компания «Роснефть»



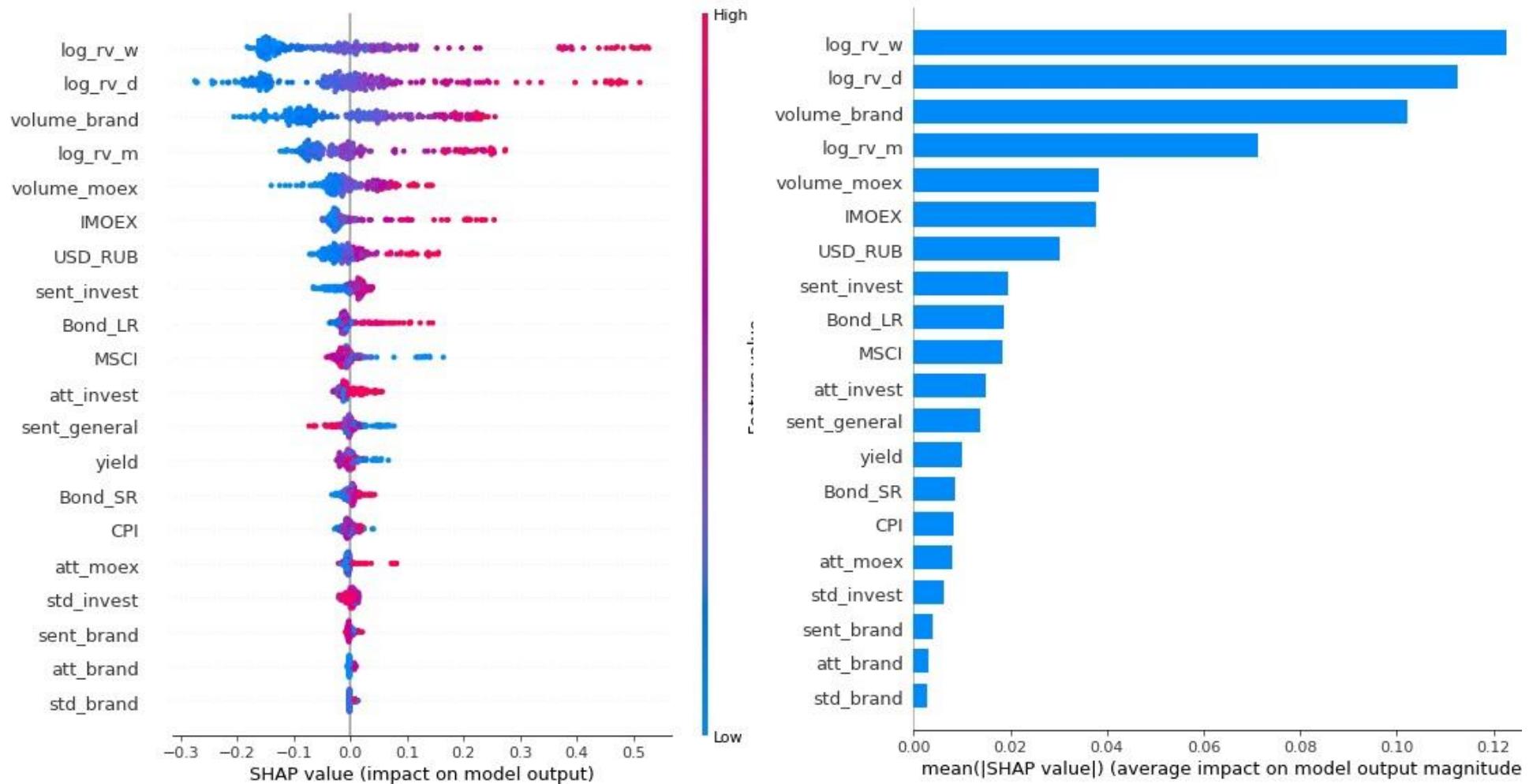
Источник: построено автором

## Компания «Газпром»



Источник: построено автором

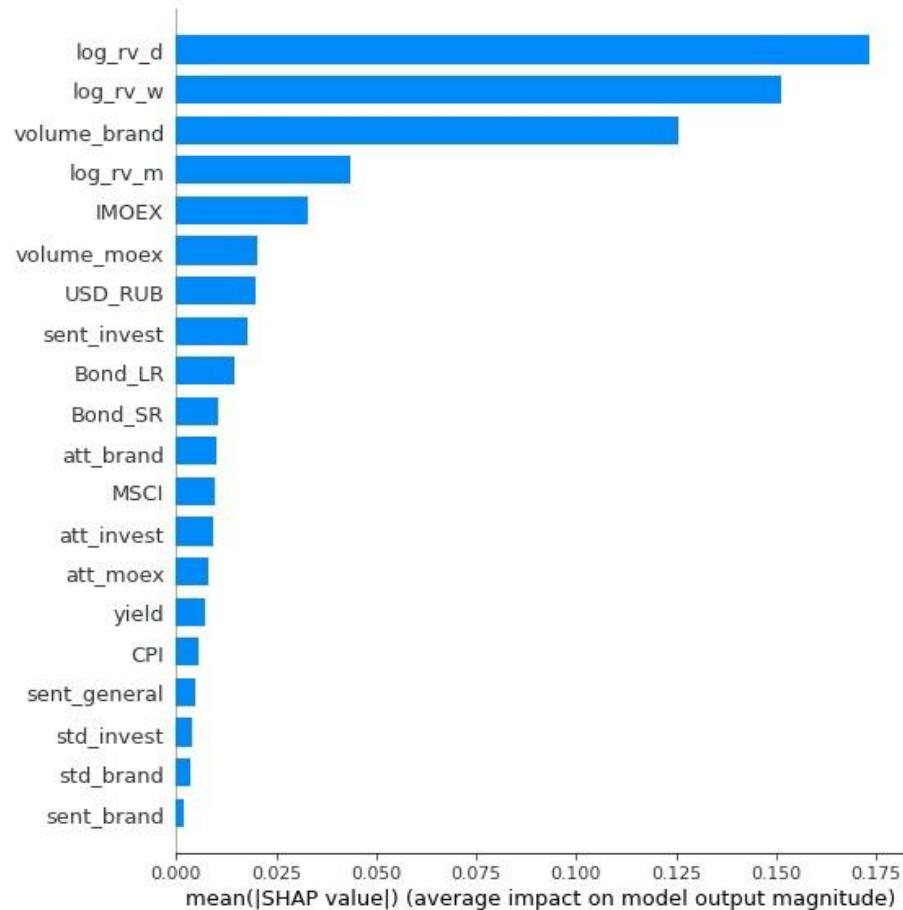
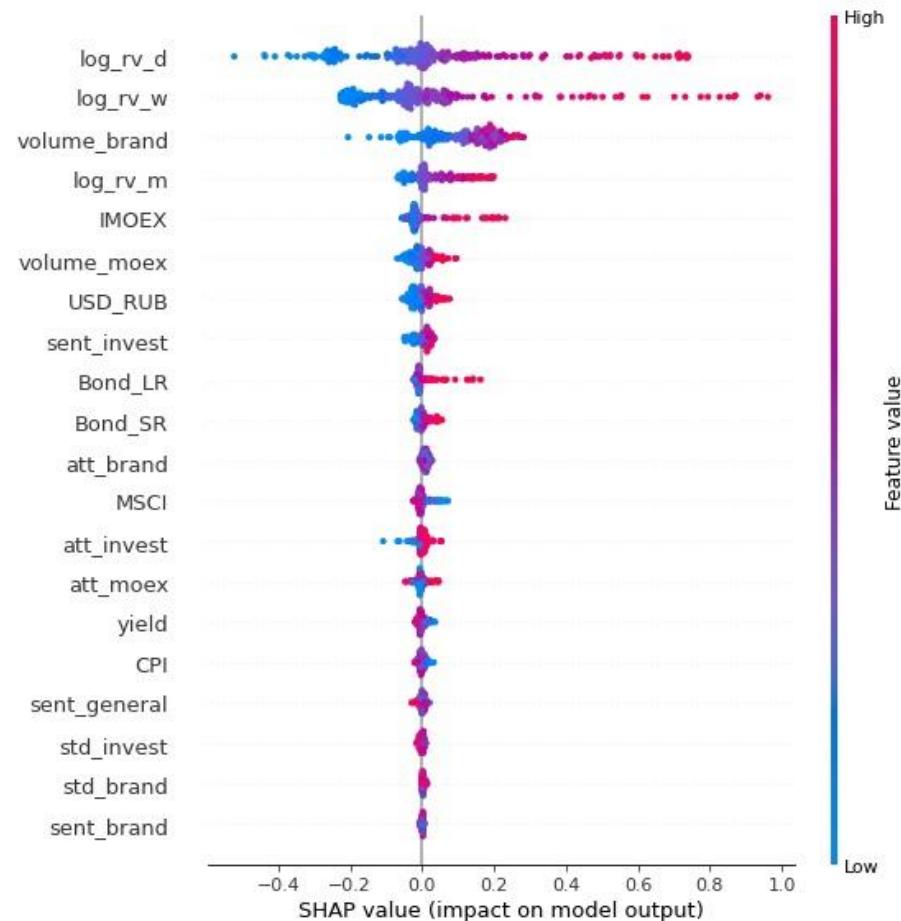
## Компания «Новатэк»



Источник: построено автором

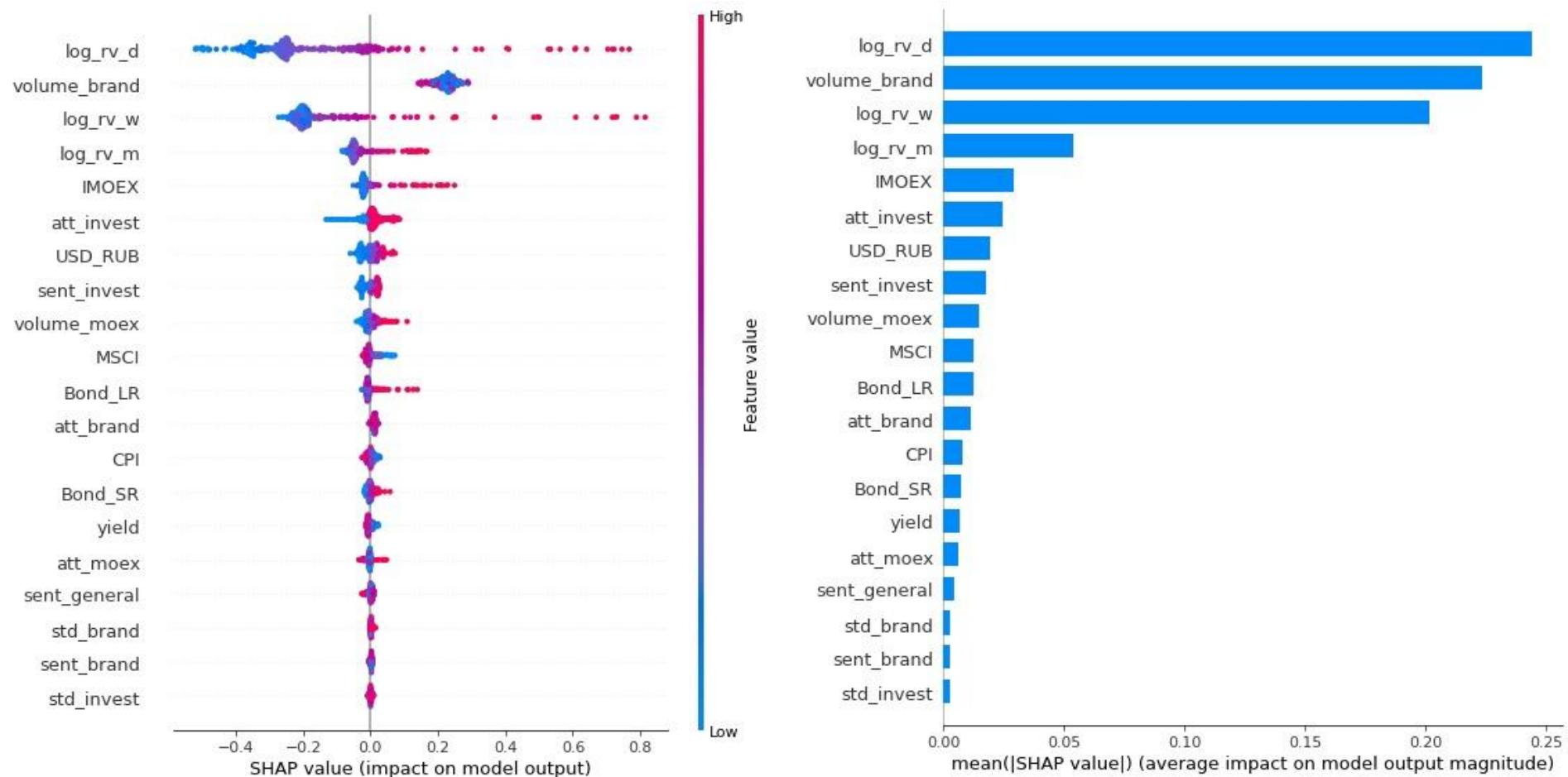
## Б2. Модель настроения. XGBoost

Компания «Яндекс»



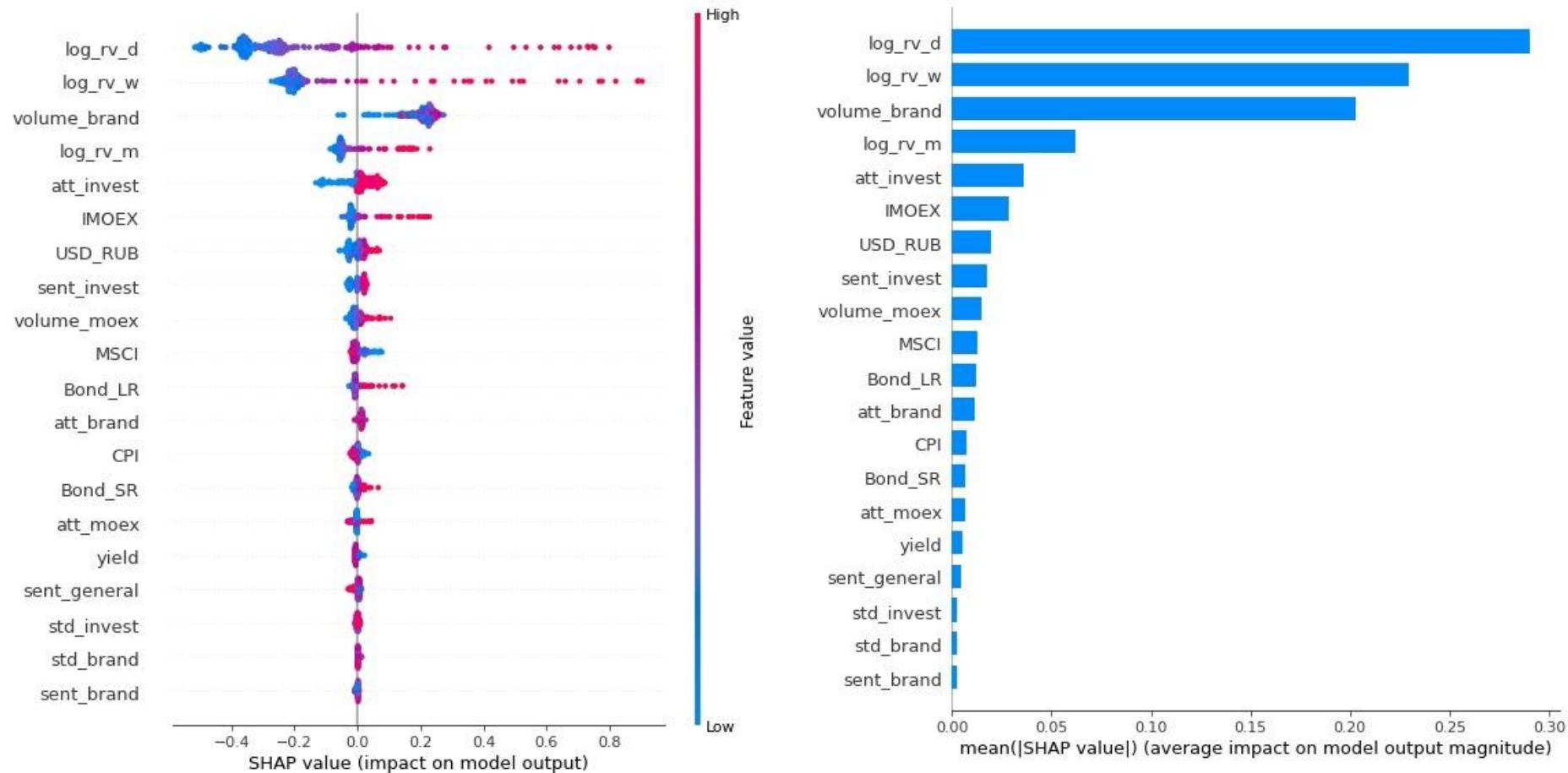
Источник: построено автором

## Компания «Сбербанк»



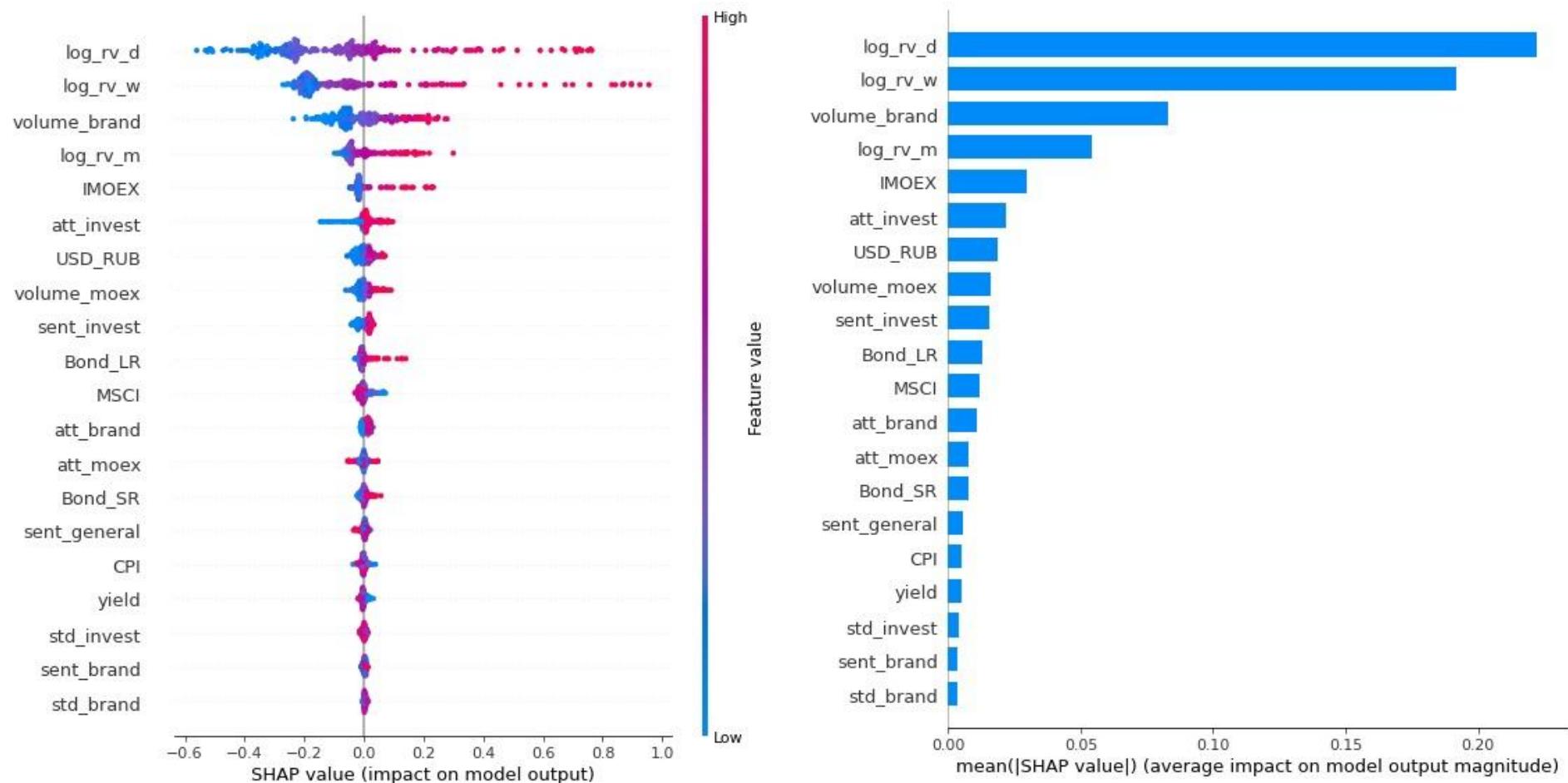
Источник: построено автором

## Компания «МТС»



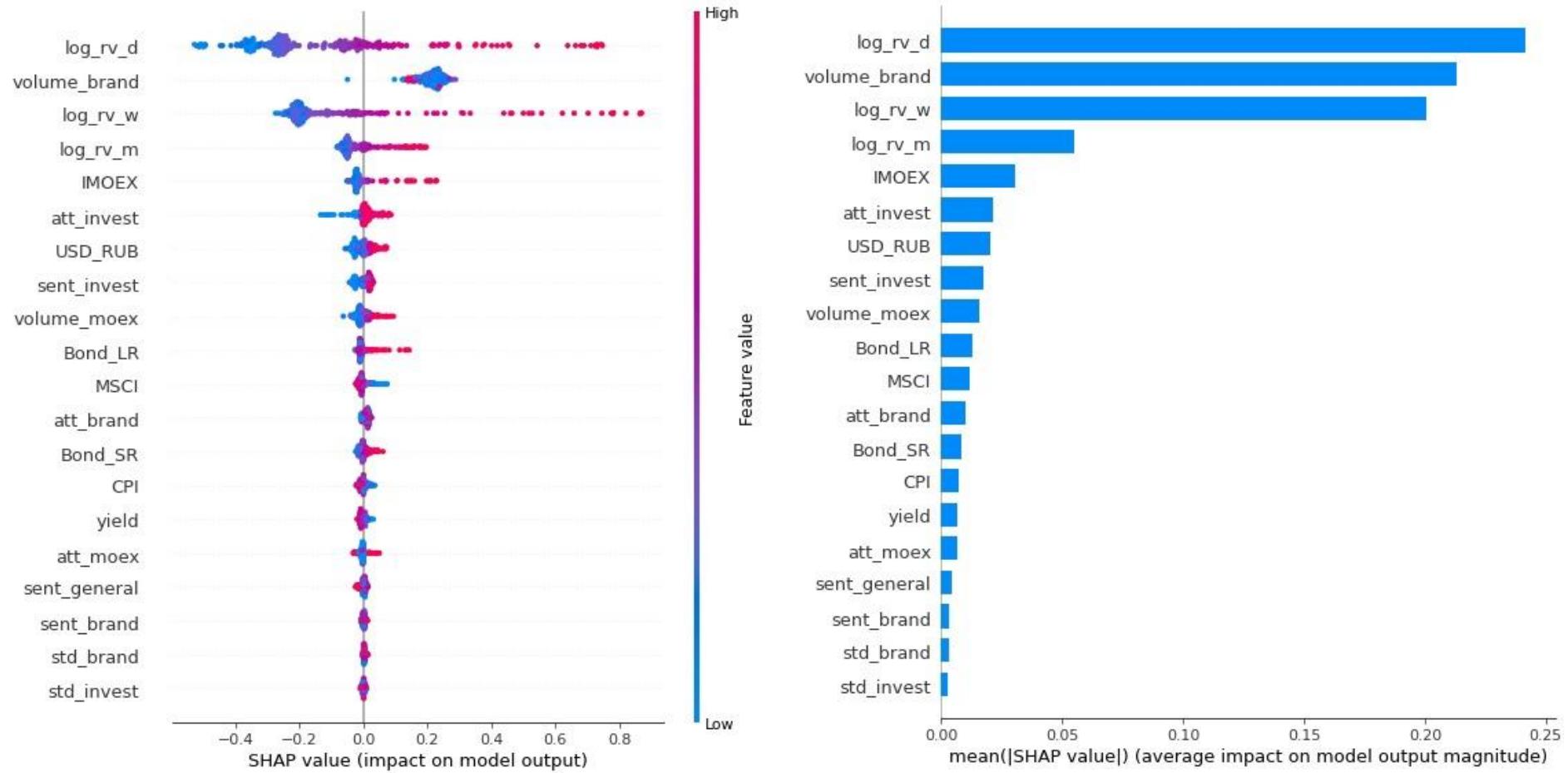
Источник: построено автором

## Компания «Лукойл»



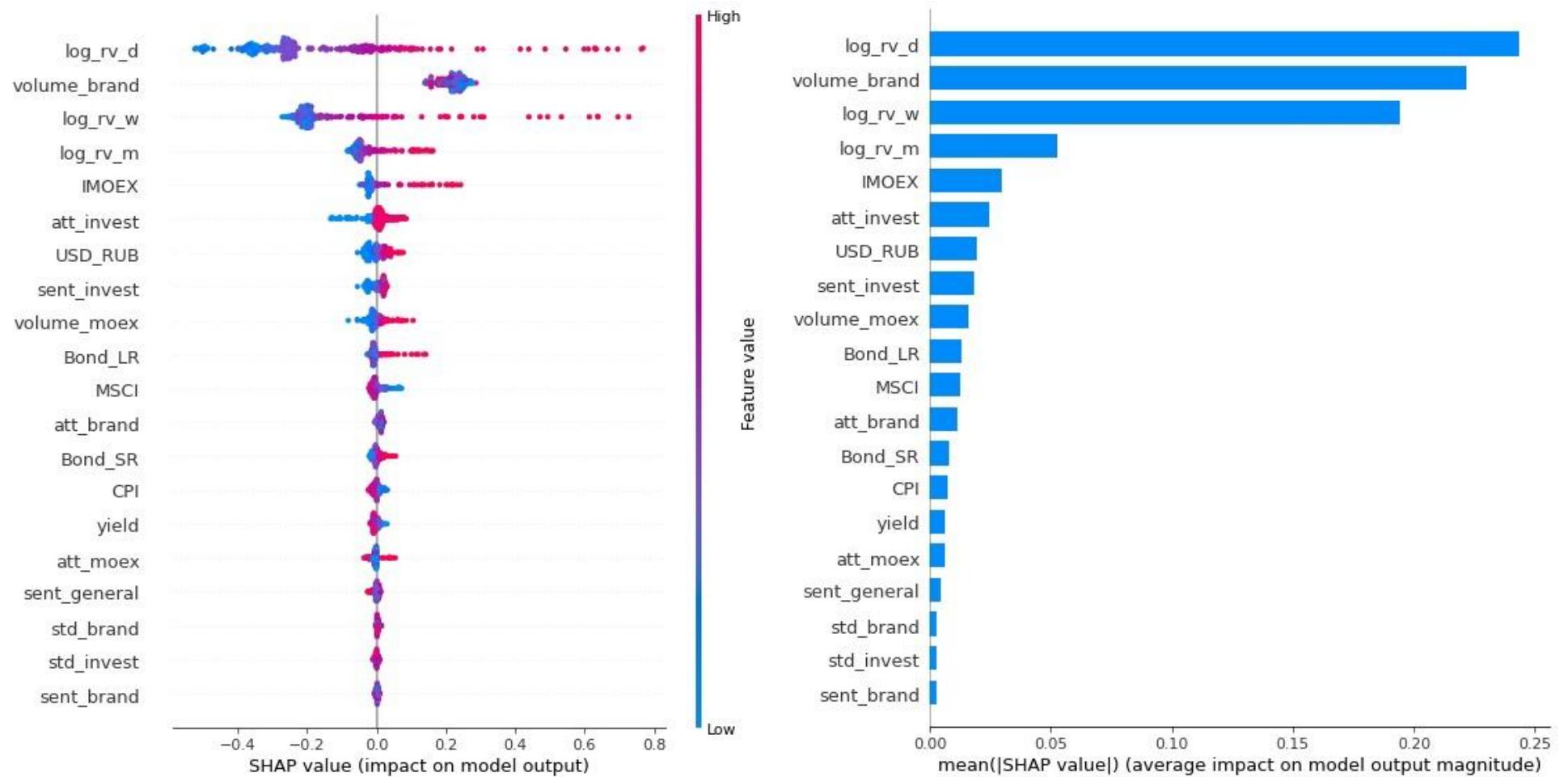
Источник: построено автором

## Компания «Роснефть»



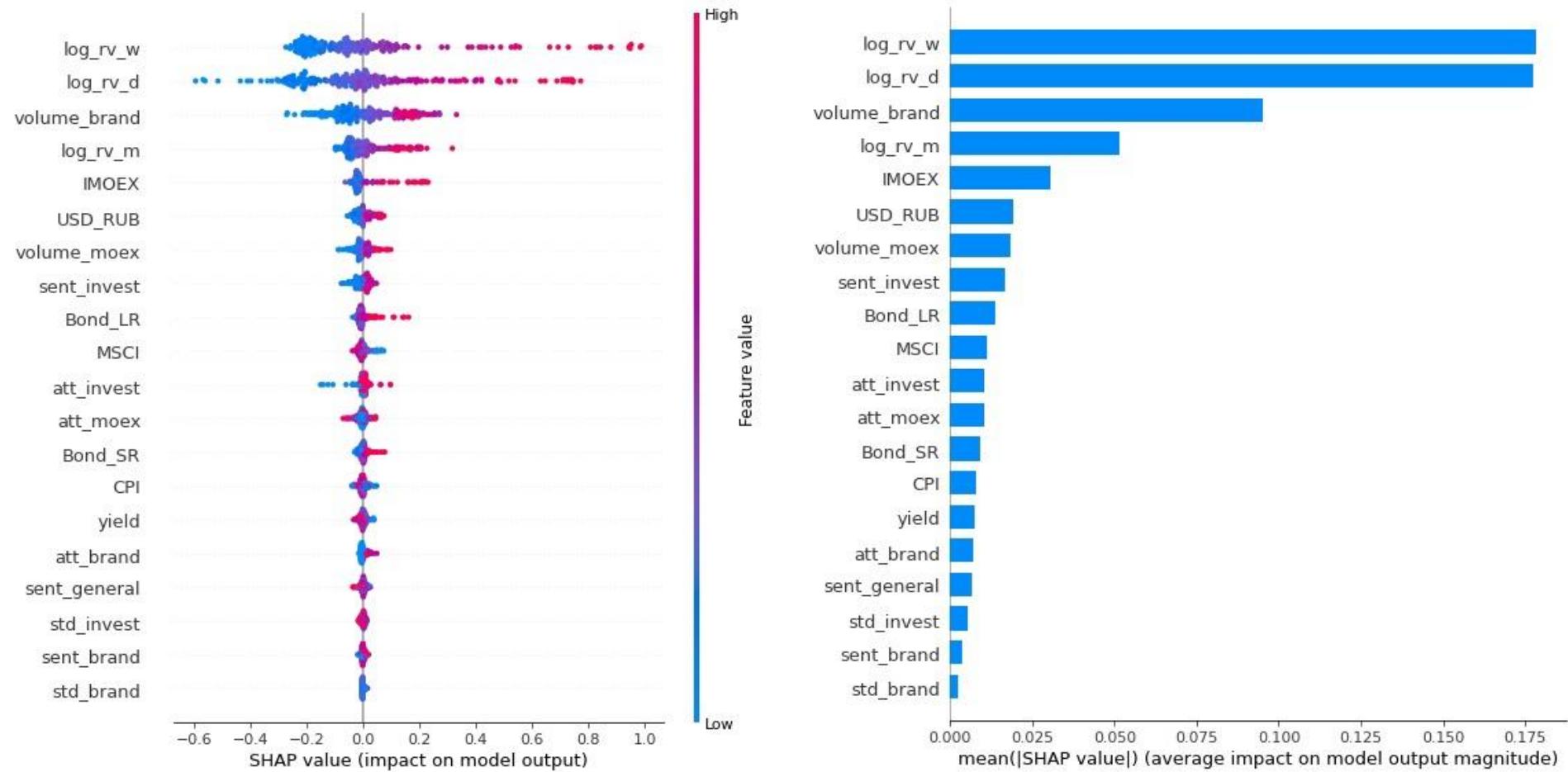
Источник: построено автором

## Компания «Газпром»



Источник: построено автором

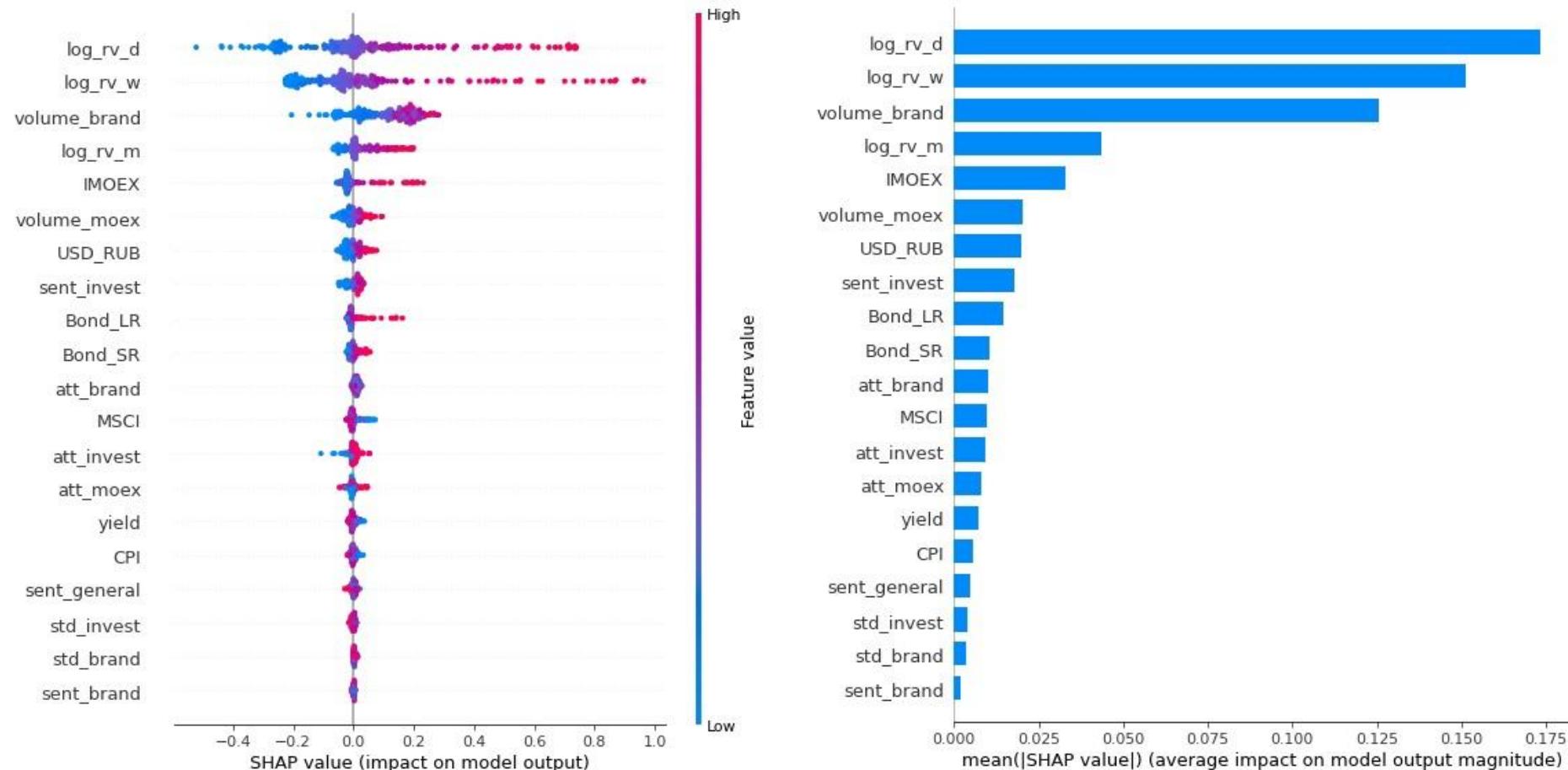
## Компания «Новатэк»



Источник: построено автором

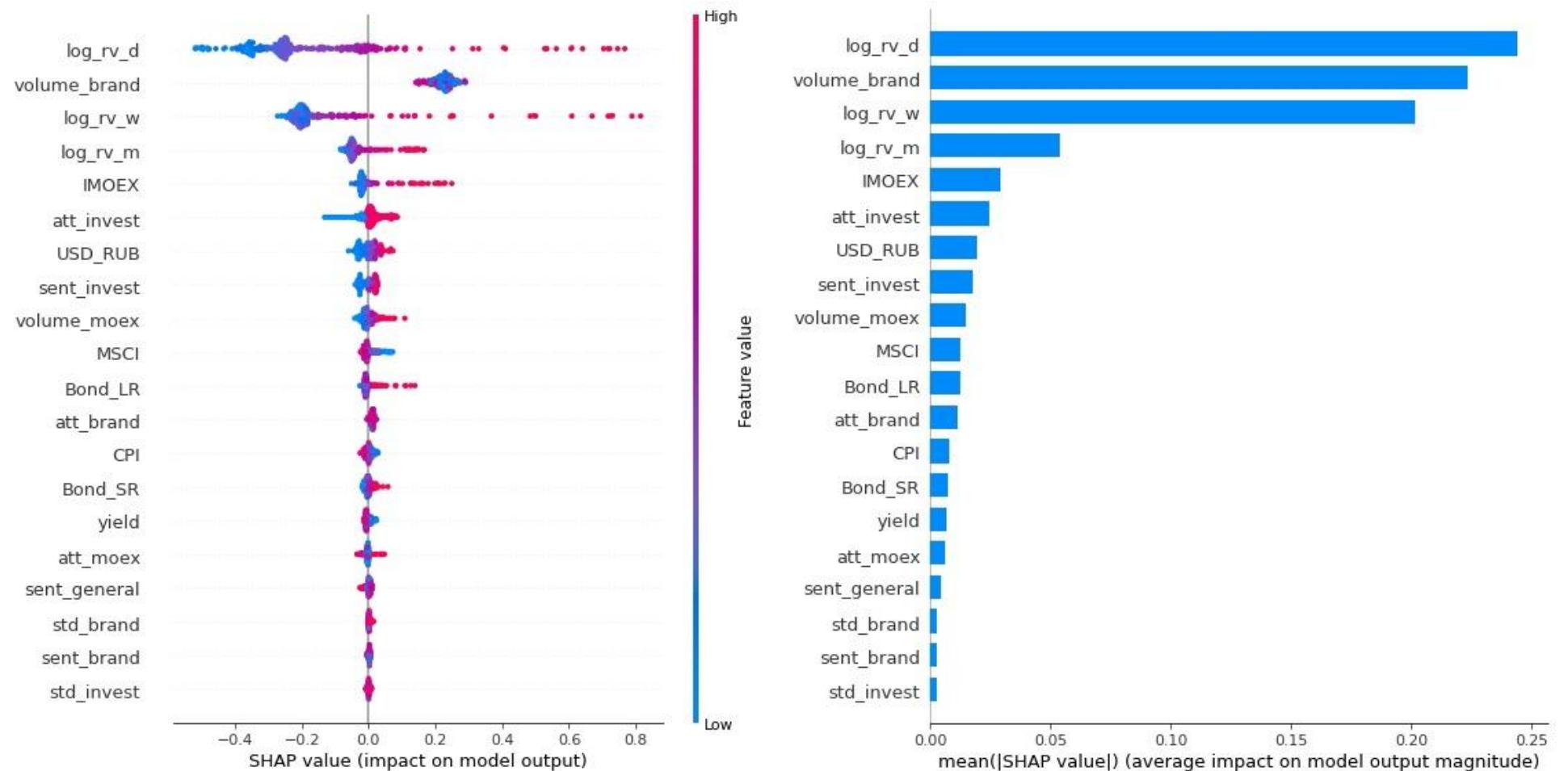
### Б3. Модель настроения. Light GBM

Компания «Яндекс»



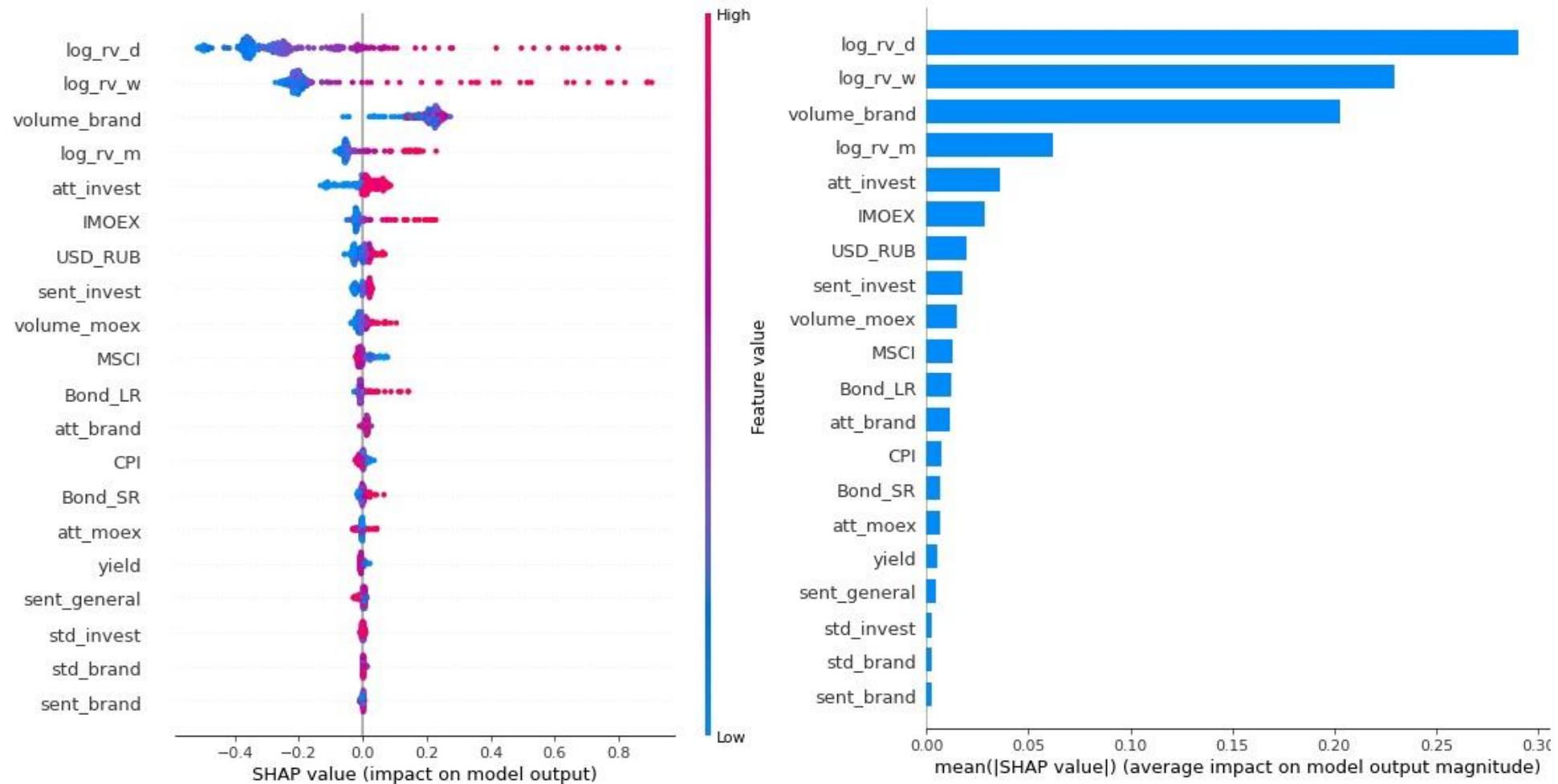
Источник: построено автором

## Компания «Сбербанк»



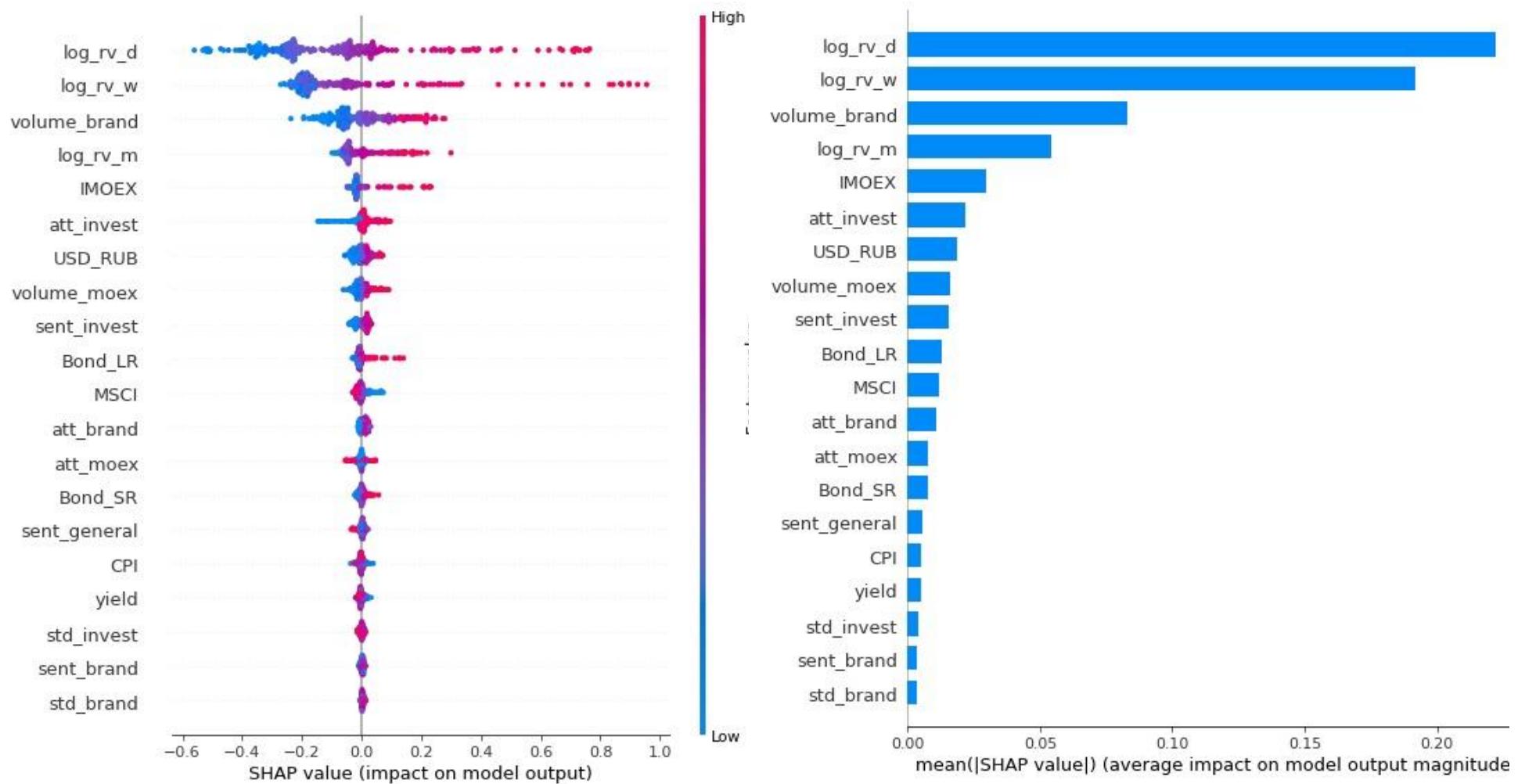
Источник: построено автором

## Компания «МТС»



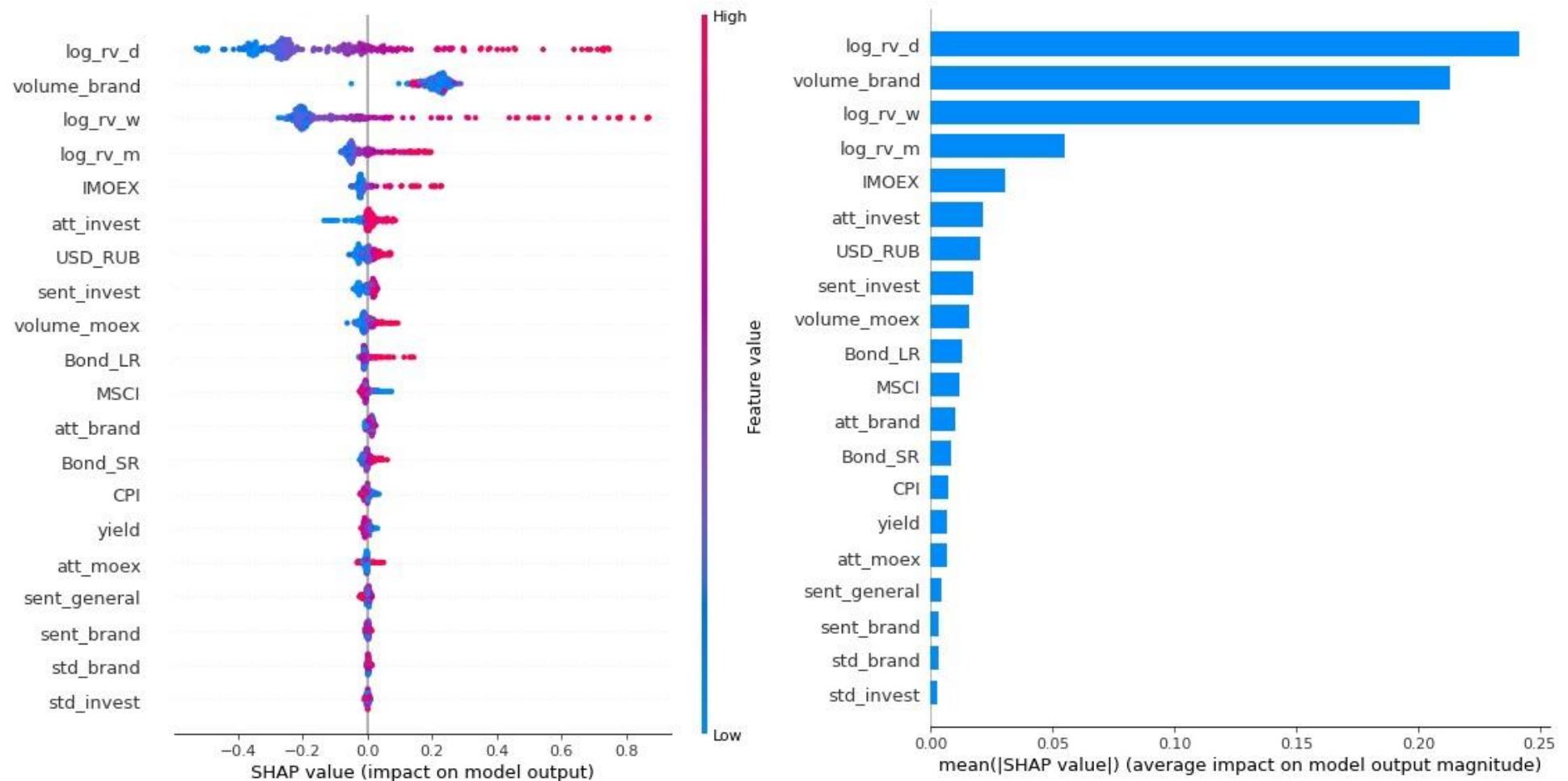
Источник: построено автором

## Компания «Лукойл»



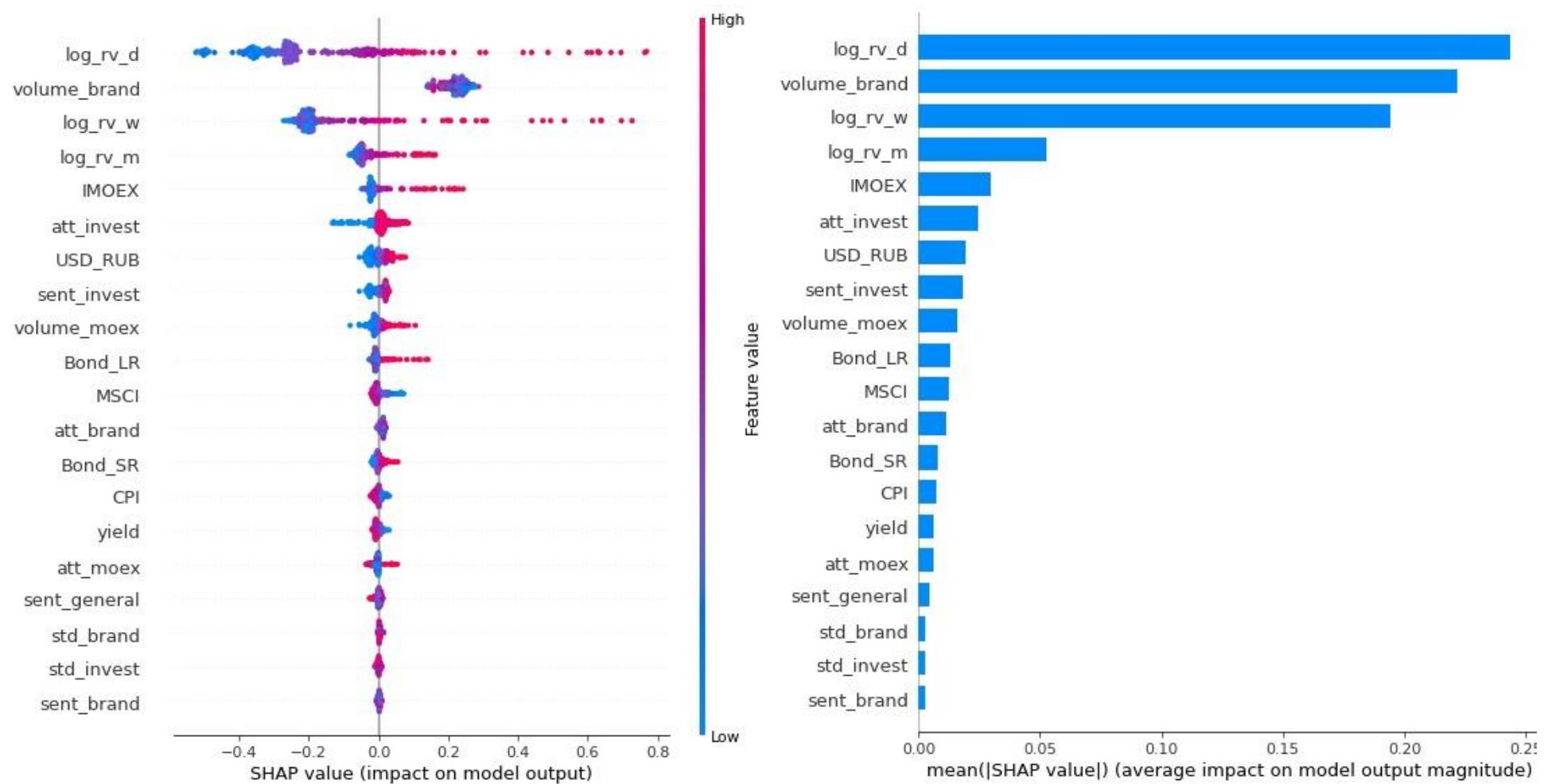
Источник: построено автором

## Компания «Роснефть»



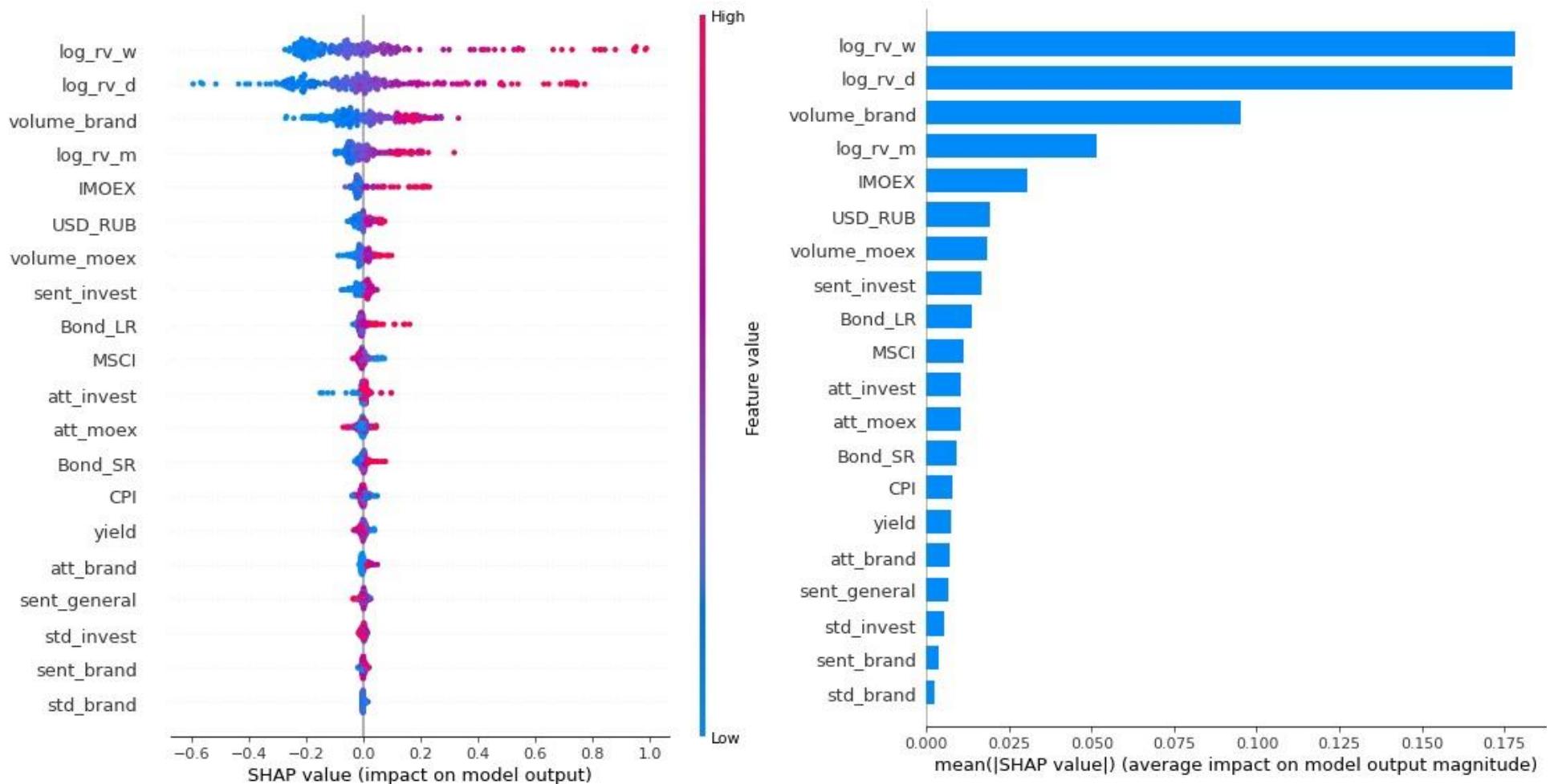
Источник: построено автором

## Компания «Газпром»



Источник: построено автором

## Компания «Новатэк»



Источник: построено автором

## ANNOTATION

The relevance of the chosen topic lies in the fact that in the modern world the role of using various methods is growing, with the help of which one can analyze risks and make forecasts. The materials of the work are of practical value for banking organizations and investors who can apply the specified model to predict risks and threats.

The work analyzed the volatility of shares of a number of Russian companies that belong to different sectors of the economy: from commodity companies - oil, gas (Lukoil, Rosneft, Gazprom, Novatek) to companies engaged in the service sector - the banking sector (Sberbank), IT (Yandex), telecommunications (MTS). As it turned out, different companies (with different focus) react differently to moods and / or news on social networks.

The work also offered practical recommendations on the choice of the field of application of the results and models.

Key words: news tone, realized volatility, stock market.

UDC: 004.852; 004.043; 336.76.066; 336.761.532; 336.763.215; 336.763.218

JEL: G17, G32