Федеральное государственное автономное образовательное
учреждение высшего образования
**«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»**
Факультет экономических наук

КУРСОВАЯ РАБОТА

Прогнозирование экономических и финансовых показателей на основе текстов новостей

по направлению подготовки Экономика
образовательная программа «Экономика и статистика»

Работу выполнили:
Разуваев Никита Сергеевич
Группа: БСТ182,
Романенко Александр Сергеевич
Группа: БСТ196
Руководитель:
доцент, кандидат экономических
наук, Мамедли Мариам Октаевна

Москва 2021

Содержание

1 Введение

С развитием методов машинного обучения и обработки текстовой информации продвинутыми алгоритмами появляются возможности по извлечению новых данных из текстов. В частности, благодаря инструментам NLP (от англ. «Natural Language Processing») можно превращать новостные статьи, комментарии в социальных сетях, поисковые запросы и другую текстовую информацию в полезные признаки, которые могут быть использованы в качестве объясняющих переменных в моделях. По сути, эти признаки являются некоторыми замещающими переменными (proxy) для ненаблюдаемых процессов, таких как настроения, ожидания, доверие к экономике и т.п.

Использование признаков, основанных на текстовой информации, все больше привлекает как исследователей, так и бизнес и государственные институты. Ряд исследований центральных банков различных стран направлен на использование текстовых данных для улучшения прогнозирования макроэкономических и финансовых показателей. В нашей работе проводится обзор результатов некоторых статей по данной теме.

Основная цель исследования заключается в том, чтобы измерить, насколько новостные тексты могут быть полезны в текущей оценке и прогнозировании макроэкономических и финансовых индикаторов, выявить влияние экзогенных новостных переменных на работу алгоритмов машинного обучения.

Используя наработки других исследователей и модифицируя их методологию, будет проведен анализ значимости новостной информации для прогнозирования макроэкономических и финансовых показателей, определены подходы к построению автоматических новостных индексов, собраны тексты новостей информационных агентств «РБК» и «Лента.ру», сформирована выборка целевых переменных. Для прогнозирования временных рядов с новостями и без них были выбраны как классические линейные модели — лассо, гребневая и обычная регрессии, так и модели машинного обучения, такие как случайный лес, градиентный

бустинг и многослойный перцептрон.

1.1 Обзор литературы

1.1.1 Making text count: economic forecasting using newspaper text

В статье [Kalamara [и др.], 2020] авторы прогнозировали широкий спектр макроэкономических показателей с использованием текстов новостей. По их словам, улучшение прогнозирования от применения текстов новостей особенно видно в периоды рецессии и спада.

Авторы статьи используют данные с ресурса Dow Jones Factiva с применением фильтра «Экономические новости/Новости производства/Новости финансовых рынков/Новости товарных рынков/Новости корпораций». Они ограничили спектр данных из-за незначимости остальных тематик статей, а также для большей простоты вычислений. Начальным этапом работы с текстом стала его отчистка от пунктуации и стоп-слов, приведение к нижнему регистру и разделение текстов на N-граммы. Авторы использовали три категории методов для создания текстовых метрик: словарные с применением tf-idf, булевские (от англ. boolean) и методы из литературы по компьютерным наукам (в пример приводится VADER метрика). В альтернативу раннее созданным словарям они сформировали набор весов для каждого термина при помощи различных моделей машинного обучения (Ridge и Lasso регрессии, SVM, Random Forest). Затем авторы отделили текстовые временные ряды, взяв те, которые содержат в себе сентименты (или тональности/настроения), и те, в которых наблюдается неопределенность. Первые полезны тем, что они явно коррелируют с прогнозируемыми показателями, вторые также несут в себе экономический смысл. Исследователи определили ключевую значимость сентиментов во времени. Отклонение всех отобранных рядов от среднего сильно коррелировало с отклонением рядов, содержащих сентименты. «Неопределенные» ряды также были значимы, однако она проявлялась лишь в спорных экономических и политических ситуациях,

таких как вторжение в Ирак и Брексит.

В итоге, лучше всего себя показали метрики TF-IDF и «Stability». Вторая создана для выявления предпосылок к наступлению экономического спада, однако она основана на истории возникновения кризиса 2008-го года, что не всегда может подойти ввиду неоднородности факторов, из-за которых может возникнуть кризис.

В основу используемой модели легло 36-ти месячное скользящее окно. Любые преобразования данных, такие как, например, нормализация проводились с данными настоящего или прошлого, избегая использования данных из будущего (Look-ahead bias). Целью использования модели стал поиск тех ситуаций, когда модель с применением данных новостей справляется лучше иных. Использовалась предпосылка о том, что политик обладает новостной информацией в период t и хочет предсказать некоторый показатель в период $t + h$. Базовой моделью была выбрана авторегрессия первого порядка $AR(1)$, авторы прогнозировали ВВП, индекс безработицы, деловой индекс, потребление домашних хозяйств, инфляцию, производственные индексы. В сравнении с базовой, модель с наличием новостной информации существенно сократила метрику RMSE. При использовании авторегрессии с наличием экономических факторов в качестве признаков, новостная информация показывает себя значительно хуже и вносит меньший вклад в дисперсию целевой переменной. Модель с использованием метрики TF-IDF economy оказалась хуже, чем изначальная, тогда как использование метрики Stability привнесло улучшение прогнозной силы модели.

Основной частью анализа стало сравнение авторегрессии первого порядка с использованием метода наименьших квадратов и различных методов машинного обучения, включающих в себя новостные данные. В итоге, любая из моделей (Lasso, Ridge, нейронные сети, SVR) показала себя лучше базовой, новости оказались незначительным фактором только в Lasso регрессии, ввиду регуляризации модели вес новостного фактора обнулился. Тем не менее, при использовании нейронных сетей и Ridge регрессии новостной фактор остался существенным, доля вли-

яния новостей на дисперсию предсказания у данных моделей оказалась выше, чем у модели с алгоритмическими текстовыми метриками. Качество прогнозирования с использованием новостей также улучшилось, но не настолько, как было в случае сравнения обычной авторегрессией и моделью с использованием новостей. Данный факт можно объяснить определяющей важностью макроэкономических факторов при прогнозировании, однако уже на данном этапе стоит отметить, что данные новостей способны улучшить предсказание. Наибольший вклад новостей наблюдался при прогнозировании потребления и инвестиций в 9-ти месячном горизонте планирования, они оказались наиболее статистически значимыми в данным периоде. Авторы статьи сделали вывод, что гипотеза Шиллера о том, что новости скорее отражают мнения и прогнозы, а не реальную ситуацию, подтверждается.

Далее авторы статьи рассматривают ситуации, в которых использование новостей может иметь смысл, обеспечивает улучшение прогнозирования. Как оказалось, наиболее эффективный прогноз возникает при использовании новостного фактора в точках максимума и минимума прогнозируемого показателя, то есть в преддверии кризиса или экономического роста. После рецессии новостной фактор обеспечивает наибольший прирост в качестве прогнозирования, разность между RMSE (у базовой модели и моделей с новостными факторами) наиболее велика в период экономического кризиса 2008-го года. Также можно сделать вывод, что нейросетевые алгоритмы лучше всего используют новостные данные, разность между RMSE базовой модели и RMSE нейросетей достигает 30-ти пунктов, что является максимумом.

Авторы статьи делают вывод, что новостные данные могут надежно улучшить прогнозирование в горизонте от 3-х до 9-ти месяцев, а методы формирования признакового описания текстов, учитывающие как сентименты, так и неопределенность, могут учесть всю значимую информацию. Была выявлена зависимость между частотой упоминания однокоренных слова «economy» и возможностью резкого роста/спада прогнозируемого показателя. Словарь «Dictionary for Financial stability» показал

наилучшие результаты среди словарей, однако наиболее эффективным оказалось использование всех слов из текста во временном ряде и дальнейшее определение веса каждого из них моделью машинного обучения. Наилучшими моделями оказались нейронные сети и Ridge-регрессия, назначающие веса словам одновременно и не учитывающие нелинейные отношения между переменными.

1.1.2 Прогнозирование российских макроэкономических показателей на основе информации в новостях и поисковых запросах

[Ульянкин, 2020] сравнивает «ручные» и «автоматические», построенные при помощи методов машинного обучения, индексы экономической активности.

Ручные индексы, представляющие собой опросы домохозяйств об их текущем финансовом положении, ожиданиях цен, экономическом развитии, чаще всего отражают уровень доверия к экономике страны. Наиболее популярным из подобных показателей является Мичиганский индекс потребительских настроений, основанный на анализе ответов респондентов по поводу их ожиданий и текущей ситуации. Как оказалось, между макроэкономическими рядами и индексами доверия, схожими с Мичиганским, можно выработать взаимосвязь и доказать статистическую значимость последних при объяснении тех или иных составляющих реальных экономических показателей. Тем не менее, чаще всего создание и обработка выборки для составления ручных индексов является предельно ресурсоемким занятием, поэтому все большее число исследователей ставит перед собой задачу построения качественных автоматических индексов.

Одним из основных видов автоматических индексов является индекс новостей. Чаще всего применяют словарный метод, то есть определяют частоту упоминания определенных терминов. К примеру, многие выделяют такие слова, как «экономика» и «неопределенность» для выявления тренда.

В качестве данных для анализа автор статьи выбрал статистику Google Trends (масштабированные до предельного значения 100 поисковые запросы), также были взяты статьи из «РБК», «РИА Новости», ТАСС, «Интерфакса» и «Ленты» (новостные данные). Каждый текст был токенизирован и очищен от стоп-слов и знаков пунктуации. Также для предварительной обработки данных производилась лемматизация статей, она была выбрана ввиду некачественной работы алгоритмов стемминга с русским языком. Еще одним источником данных для построения автоматических индексов послужили комментарии из социальной сети «Вконтакте» со страниц новостных ресурсов.

Автор выбрал несколько ручных индексов для последующего сравнения: PMI, ИПН «Левада-центра», Индекс предпринимательской уверенности. Целевыми переменными стали Доходность по индексу РТС, Официальные курсы валют Банка России, ряды, связанные с вкладами и кредитами, а также набор макроэкономических рядов.

Индекс новостей строился в нескольких вариациях:

1) Частотные индексы новостей: каждый день считается число статей, содержащих в себе каждый из кризисных дескрипторов, затем взвешивают их с весами на основе корреляций дескрипторов.

2) Сентимент-индекс: подсчет сентимент-окраса статьи при помощи тонального словаря «Карта слов»

По итогу корреляционного анализа наблюдалась умеренная взаимосвязь автоматических индексов, корреляция находилась на уровне 0.3 - 0.4. Также автор заметил, что новостные индексы улавливают тренды раньше, нежели чем ручные, то есть при помощи новостей возможно предсказывать поведение ручных индексов, основанных на опросах. На основе теста на причинность по Грейнджеру новостные индексы оказались причинными для ручных индексов. Это можно связать с тем, что для участников опросов, то есть менеджеров и предпринимателей, первоисточником принятия решений и соответствующих ответов, данных на опросе, являются новости, то есть новостные индексы отражают информацию раньше ручных.

Для изучения прогнозной силы автоматических индексов использовалась модель ARIMA (авторегрессионное интегрированное скользящее среднее) с оптимальными параметрами, найденными при помощи минимизации информационного критерия Шварца. В качестве подсчета общей ошибки алгоритма использовался метод leave-one-out cross-validation on a rolling basis, то есть считалась ошибка (по функции потерь MAE) от 1-го до n-го наблюдения с шагом 1 (сначала для первого, потом для 1-го и 2-го, затем для 1-го, 2-го и 3-го и т.д.), затем вычислялась средняя ошибка прогноза. Также была построена аналогичная ARIMA модель, имеющая полученный раннее индекс в качестве экзогенно заданной переменной. Они сопоставлялись как в задаче текущей оценки прогноза (nowcasting) для анализа объясняющей способности модели, так и в задаче прогнозирования (forecasting) для выявления прогнозной силы модели.

Наибольшего увеличения разницы значений функций потерь удалось добиться при текущей оценке курса доллара, индекса промышленного производства, оборота розничной торговли, индекса РТС и депозитов физлиц. Высокая разница в MAE с применением новостного индекса оказалась при оценке курса доллара (1.06), индекса промышленного производства (0.63)

В задаче прогнозирования (forecasting) наиболее успешным оказался индекс новостей, ошибка при его использовании в среднем снижалась на 0.046, наиболее выгодным оказалось его использование при предсказании индекса промышленного производства и оборота розничной торговли: MAE уменьшилось на 1.18 и 0.34 пункта соответственно.

Автор также выявил, что от использования лагированного, а не real time показателя (именно такой и использовался в задаче «forecasting») сильнее всего растет в прогнозной силе именно новостной индекс, что можно связать с предсказательной сутью множества экономических статей.

2 Основная часть

2.1 Новостные данные

Новостные тексты изданий «РБК[1]» и «Лента.ру[2]» были собраны из открытых источников при помощи авторского парсера[3] (от англ. – parser). Были взяты статьи с 2000-го по 2020-й год, относящиеся к разделам «Экономика» и «Финансы». Ограничение на тематику статей введено по двум причинам. Во-первых, статьи других категорий могут нести семантический окрас, но не иметь экономической значимости, что может сильно зашумлять построенные индексы. Во-вторых, сбор новостей только по определенным тематикам существенно сокращает время вычислений.

Источник	РБК	Лента.ру
Количество статей	18339	97344
Период	с 2000-09 по 2020-10	с 2000-01 по 2020-11

Таблица 1: Новостные статьи

В таблице 1 указано число статей в источниках и периоды их публикаций. Заметим, что в указанный период времени число статей агентства «Лента.ру» более чем в 5 раз превышает количество статей издания «РБК». Это отражает специфику СМИ. Вероятно, «Лента.ру» публикует как оригинальные статьи, так и дублирует информацию из других источников, в то время как «РБК» концентрируется на оригинальных статьях и дублирует лишь наиболее важные новости из сторонних источников. Фактически это означает, что индексы, построенные на данных «Лента.ру», могут оказаться с одной стороны более полными, а с другой более зашумленными в сравнении с «РБК». Динамика количества публикаций также может отражать как внутреннюю конъюнктуру СМИ, так экономические настроения в стране. В периоды кризиса растет число новост-

[1] https://www.rbc.ru/

[2] https://lenta.ru/

[3] Ссылка на парсер и данные: https://www.kaggle.com/hardtype/rbc-economy-news

ных статей и, вероятно, спрос на них, что видно на рисунках **??** и **??** в приложении.

2.2 Целевые показатели

В качестве целевых переменных было отобрано 32 показателя. Среди них доходность по индексу РТС и IMOEX, средние значения курсов валют Банка России (за месяц), данные о средствах физических лиц на депозитах и жилищных ипотечных кредитах на первичном рынке по данным ЦБ, а также макроэкономические ряды, такие как месячная инфляция, безработица, импорт и т.п. Полный перечень используемых в работе показателей и источников данных отражен в таблице 2.

Обработка показателей представляет собой нетривиальную задачу. Основной принцип состоит в том, чтобы не допустить «подглядывания в будущее», о чем упоминается в статье [Kalamara [и др.], 2020], в то время как в статье [Ульянкин, 2020] данный момент упускается, показатели нормируются в шкалу от 0 до 100, что, вероятно, делается путем min-max нормирования:

$$x_t^* = \frac{x_t - x_{min}}{x_{max} - x_{min}} \cdot 100$$

С одной стороны, данная процедура неизбежно ведет к использованию данных из будущих периодов, если $min_t\{x_t\}$ и $max_t\{x_t\}$ будут определятся по всему множеству периодов $t = 0, 1, ..., T$. С другой стороны, если $min_t\{x_t\}$ и $max_t\{x_t\}$ будут определяться с помощью скользящего ($t \in \{t, t-1, ..., t-m\}$) или расширяющегося ($t \in \{t, t-1, ..., 0\}$) окна, может возникнуть ситуация, когда одно и то же значение преобразованного показателя x^* будет означать совершенно разные исходные значения.

Короткое название	Полное название	Кол-во (мес.)	Период	Источник
Макроэкономические показатели				
CPI_M_CHI	Индекс потребительских цен	337	01.1992 - 01.2021	sophist.hse.ru
IM_T_M	Импорт	311	01.1994 - 11.2020	
INVFC_M	Инвестиции в основной капитал	252	01.1994 - 12.2015	
IP2_EA_M	Индекс промышленного производства по ОКВЭД2	72	01.2014 - 12.2020	
RTRD_M_DIRI	Индекс реального оборота розничной торговли	312	01.1994 - 12.2020	
RTRD_M	Оборот розничной торговли	300	01.1995 - 12.2020	
UNEMPL_M_SH	Уровень безработицы	312	01.1994 - 12.2020	
WAG_M	Реальная заработная плата	323	01.1993 - 11.2020	
WAG_C_M	Средняя номинальная заработная плата	323	01.1993 - 11.2020	
Депозиты и ипотечный рынок				
DEP_FIZ_SUM	Депозиты физических лиц РФ в рублях	132	01.2007 - 12.2018	www.cbr.ru
IPOT_CNT	Количество жилищных кредитов, предоставленных резидентам РФ в рублях	121	01.2009 - 01.2019	www.cbr.ru
IPOT_VOLUME	Объем жилищных кредитов, предоставленных резидентам РФ в рублях	121	01.2009 - 01.2019	
IPOT_DEBT	Задолженность жилищных кредитов, предоставленных резидентам РФ в рублях	121	01.2009 - 01.2019	
Валютные курсы				
AVG_USD	Средний курс Доллара США~	253	01.2000-01.2021	www.cbr.ru
CLOSE_USD	Курс Доллара США на конец периода~	253	01.2000-01.2021	
AVG_EURO	Средний курс Евро~	253	01.2000-01.2021	
CLOSE_EURO	Курс Евро на конец периода~	253	01.2000-01.2021	
Индексы финансового рынка				
RTSI_M_OPEN	Средняя цена открытия индекса РТС~	253	01.2000-01.2021	www.moex.com
RTSI_M_CLOSE	Средняя цена закрытия индекса РТС~	253	01.2000-01.2021	
RTSI_M_MAX	Максимальная цена индекса РТС~	253	01.2000-01.2021	
RTSI_M_MIN	Минимальная цена индекса РТС~	253	01.2000-01.2021	
RTSI_M_VOLUME	Средний объем торгов индекса РТС	253	01.2000-01.2021	
MOEX_M_OPEN	Средняя цена открытия индекса МосБиржи~	253	01.2000-01.2021	www.moex.com
MOEX_M_CLOSE	Средняя цена закрытия индекса МосБиржи~	253	01.2000-01.2021	
MOEX_M_MAX	Максимальная цена индекса МосБиржи~	253	01.2000-01.2021	
MOEX_M_MIN	Минимальная цена индекса МосБиржи~	253	01.2000-01.2021	
MOEX_M_VOLUME	Средний объем торгов индекса МосБиржи	231	11.2002-01.2021	
Индексы настроений и деловой активности				
PMI_MANUFACTURING	Индекс деловой активности промышленного сектора	89	07.2012-11.2020	www.markiteconomics.com
PMI_SERVICES	Индекс деловой активности сферы услуг	89	07.2012-11.2020	
BCI_RUSSIA_OECD	Индекс предпринимательской уверенности по методике OECD	157	01.2008-01.2021	data.oecd.org
CCI_RUSSIA	Индекс потребительской уверенности по методике Левада-центра	141	03.2008-11.2020	www.levada.ru
BCI_RUSSIA_ROSSTAT	Индекс предпринимательской уверенности по методике Росстата	115	07.2011-01.2021	rosstat.gov.ru

Таблица 2: Целевые переменные

В нашей работе используется методика, указанная в статье Банка Ан-

глии. Данные нормируются скользящим окном $m = 24$ месяца:

$$x_t^* = \frac{x_t - \bar{x}_t}{\sigma_{x_t}}, \quad \bar{x}_t = \frac{1}{m}\sum_{i=t-m+1}^{t} x_i, \quad \sigma_{x_t}^2 = \frac{1}{m}\sum_{i=t-m+1}^{t} (x_i - \bar{x}_t)$$

Тем не менее, это не уравнивает ни среднее значение, ни дисперсию между разными показателями и абсолютно сопоставимыми они попрежнему не являются, однако лежат в примерно одинаковой, «стандартизованной» шкале.

Перед нормализацией показатели были приведены к цепному росту:

$$x_t^* = \frac{x_t}{x_{t-1}}, \quad t = 0, 1, ...$$

Отказ от приведения показателей год к году ($x_t^* = \frac{x_t}{x_{t-12}}, \quad t = 0, 1, ...$) в пользу месяца к месяцу обусловлен крайне низким качеством прогнозов моделей при такой обработке целевых переменных.

2.3 Обработка текста, создание новостных признаков

Для использования новостей как признаков для прогнозирования была проведена стандартная для NLP предварительная обработка текстов статей. Каждый из текстов был разбит на токены (в нашем случае на слова), удалены стоп-слова и пунктуация, применена лемматизация слов — каждое слово приведено к начальной форме. Альтернативным вариантом лемматизации является стемминг – выделение основы слова, но мы отказались от данного метода нормализации текста ввиду его малой пригодности для работы с русским языком. В качестве инструмента для лемматизации использовалась библиотека PyMystem3[4] [Segalovich, 2003].

Обработанные тексты новостей впоследствии использовались для создания признаковых описаний и индексов новостей:

[4]https://yandex.ru/dev/mystem/

2.3.1 TF-IDF

Основной принцип методики TF-IDF (Term Frequency - Inverted Document Frequency) заключается в создании признакового описания статьи на основе подсчета слов. Для каждого слова t из текста d рассчитывается относительная частота встречаемости в нем:

$$tf(t,d) = \frac{n_t}{\sum_k n_k},$$

Где n_t — количество повторений слова t в статье.

Далее каждое слово взвешивается с помощью «обратной частоты документа»:

$$idf(t) = ln\left(\frac{D+1}{df(t)+1}\right) + 1$$

Где D — количество документов (новостных статей в нашем случае), $df(t)$ — число документов, в которых встречается слово t. Итоговый вектор-признак для статьи l выглядит следующим образом:

$$f^{(l)}_{tf-idf} = (tf(t_1, d_{(l)}) \cdot idf(t_1), tf(t_2, d_{(l)}) \cdot idf(t_2), \ldots, tf(t_k, d_{(l)}) \cdot idf(t_k))$$

Чтобы получить вектор-признак для месяца M, в котором более 1-ой статьи, применяется усреднение:

$$f^{(M)}_{tf-idf} = \frac{1}{|l \in M|} \sum_{l \in M} f^{(l)}_{tf-idf}$$

При обучении модели с помощью TF-IDF применяется принцип скользящего окна — TF-IDF признаки формируются не на основе всей выборки, а только на последних $m = 12$ месяцах, за счет чего можно не беспокоится о «заглядывании в будущее». Кроме того, для отсечения лишнего шума в обучающую выборку входили лишь слова с $0.1 \leqslant df(t) \leqslant 0.9$.

На этом этапе можно заметить, что данный метод может быть не совсем применимым к нашей задаче – ведь частота используемости таких терминов, как «кризис», «рецессия» в новостных сводках может являть-

ся важным индикатором наступающего экономического спада, однако метрика IDF может существенно занизить важность данных слов ввиду их повсеместной встречаемости в различных статьях.

2.3.2 Частотный индекс новостей

Для построения частотного индекса были взяты и дополнены токены-дескрипторы из статьи [Stolbov [и др.], 2011], которые также используются в работе [Ульянкин, 2020] [5]. Для каждой новостной статьи считалось суммарное количество токенов-дескрипторов. Затем происходило усреднение этого показателя для каждого месяца и отдельно для каждого СМИ.

Несмотря на различающуюся специфики новостных источников и разного числа статей, схожая динамика показателя прослеживается в обоих изданиях, исходя из данных рисунка 1. В приложении на рисунке **??** приводится сравнение динамики частотного индекса новостей с «ручными» индексами.

Рис. 1: Динамика частотного индекса новостей

2.3.3 Сентимент индекс новостей

Для построения сентимент индекса новостей была использована fasttext модель под названием «Dostoevsky»[6]. Модель была обучена на дан-

[5] «банк», «курс», «доллар», «евро», «цб», «ртс», «ммвб», «ипотека», «акция», «кредит», «пиф», «банкротство», «залог», «дефолт», «девальвация», «финансовый», «кризис», «спад», «рецессия», «крах», «дефицит», «безработица», «инфляция»

[6] https://github.com/bureaucratic-labs/dostoevsky

ных комментариев социальных сетей RuSentiment [Rogers [и др.], 2018]. Она классифицирует текст на 5 категорий тональностей: негативное настроение, позитивное настроение, нейтральное поведение, речевой акт (поздравления и благодарности), класс «пропустить». Для каждой категории модель определяет вероятность ее присутствия в статье. Для проведения классификации с помощью использовалась стандартная цепочка предварительной обработки текста.

Оценки вероятностей классов статей усреднялись на месячной основе, затем были построены индексы, считающие негативные, нейтральные и позитивные заголовки и тексты статей (Negative Title, Neutral Title, Positive Title, Negative Text, Neutral Text, Positive Text и т.д.). Композитный индекс для каждого издания был построен вычитанием средней вероятности негативной статьи из средней вероятности позитивной статьи:

$$f_{composite} = f_{positive} - f_{negative}$$

В статье [Ульянкин, 2020] утверждалось, что данная модель не подходит для оценки тональностей новостей, поскольку она была обучена на комментариях социальных сетей, где стиль и лексика речи отличаются от новостных. Перспективы использования данной библиотеки для нашей задачи может быть высок, однако тот факт, что она обучена на аннотированных данных из социальной сети «Вконтакте» может существенно снизить качество прогнозирования ввиду малой частоты использования экономических терминов в обучающей выборке.

На рисунках 2 и 3 отображены полученные индексы. Как и с предыдущим индексом можно видеть, что их динамика схожа между изданиями, однако волатильность показателя кажется невысокой. Вероятно, это обусловлено тем, что большинство статей модель классифицирует как нейтральные по тональности. Кроме того, за исключением выбросов композитный индекс включает только отрицательные значения, что может говорить о том, что большинство экономических новостей скорее имеют негативный характер, чем позитивный. Остальные графики

связанные с сентимент индексами вынесены в приложение.

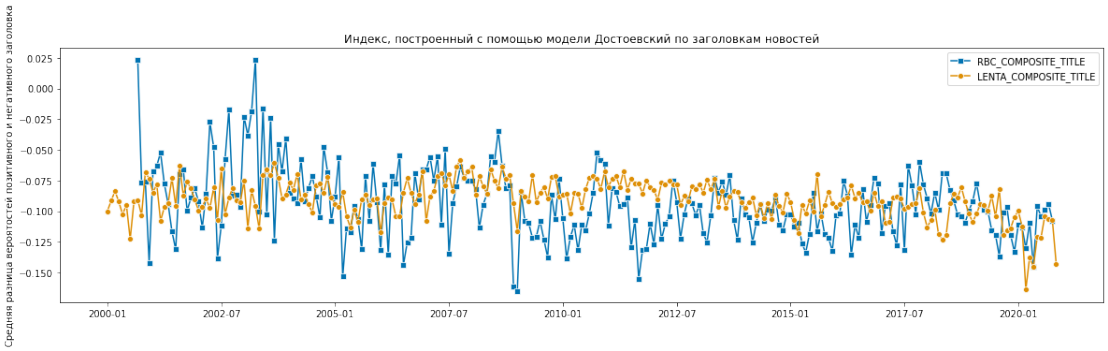

Рис. 2: Сентимент индекс заголовков новостей

Рис. 3: Сентимент индекс текстов новостей

2.3.4 Словарный индекс новостей

Словарный индекс новостей основан на применении данных Тонального словаря русского языка КартаСловСент[7]. В выборке словаря КартаСловСент представлены метки тональности для каждого входящего в него слова. Они разделяются на PSTV («позитивные»), NGTV («отрицательные») и NEUT («нейтральные»).

Методика применения построения словарного индекса заключается в подсчете суммы позитивных (Positive dict) и негативных (Negative dict) оценок слов по отдельности. Показатели усредняются на месячной основе, композитный индекс строится как разность позитивной и негативной

[7]https://github.com/dkulagin/kartaslov

среднемесячной оценки тональности стать:

$$f_{composit\_dict} = f_{positive\_dict} - f_{negative\_dict}$$

Обращаясь к рисунку 4 можно снова видеть схожую динамику показателей в разных издательствах. Сравнение с ручными индексами отражено в рисунке **??** приложения.

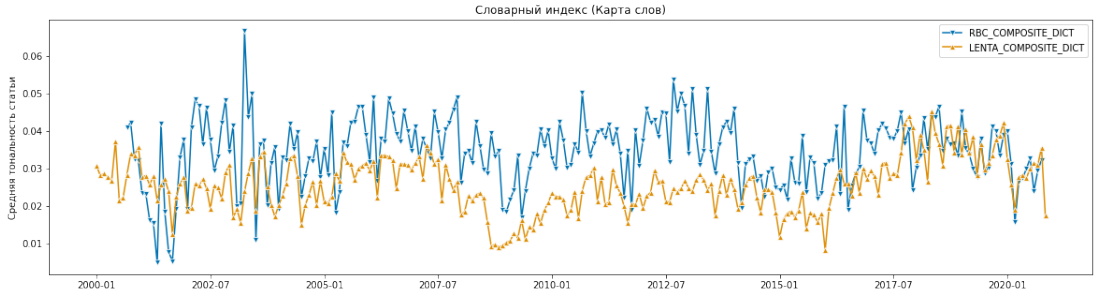

Рис. 4: Словарный индекс новостей

2.4 Корреляционный анализ

После построения новостных признаков была получена матрица парных корреляций между прогнозируемыми переменными и «автоматическими индексами новостей». В ходе ее анализа было выявлено, что наиболее серьезную и устойчивую корреляцию с частотными индексами имеют показатели, характеризующие ситуацию на валютном рынке страны. Среди них находятся средние цены доллара и евро в рублевом эквиваленте за период. Предварительно данный факт можно связать с тем, что валютные рынки наиболее остро реагируют экономические новости, цены быстро меняются в зависимости от тех или иных макроэкономических и финансовых индикаторов, представленных в новостях.

Стоит заметить, что построенный на основе Тонального словаря русского языка композитный индекс (на данных из источника «Лента») сильно коррелирует с показателями финансового рынка, то есть ценами открытия, закрытия и средними за период на акции из корзин IMOEX и

РТС. Аналогичная ситуация наблюдается и у схожего индекса, но полученного на данных статей из «РБК». Наиболее сильную корреляцию с «позитивными» индексами демонстрируют инвестиционные индексы и индекс предпринимательской активности.

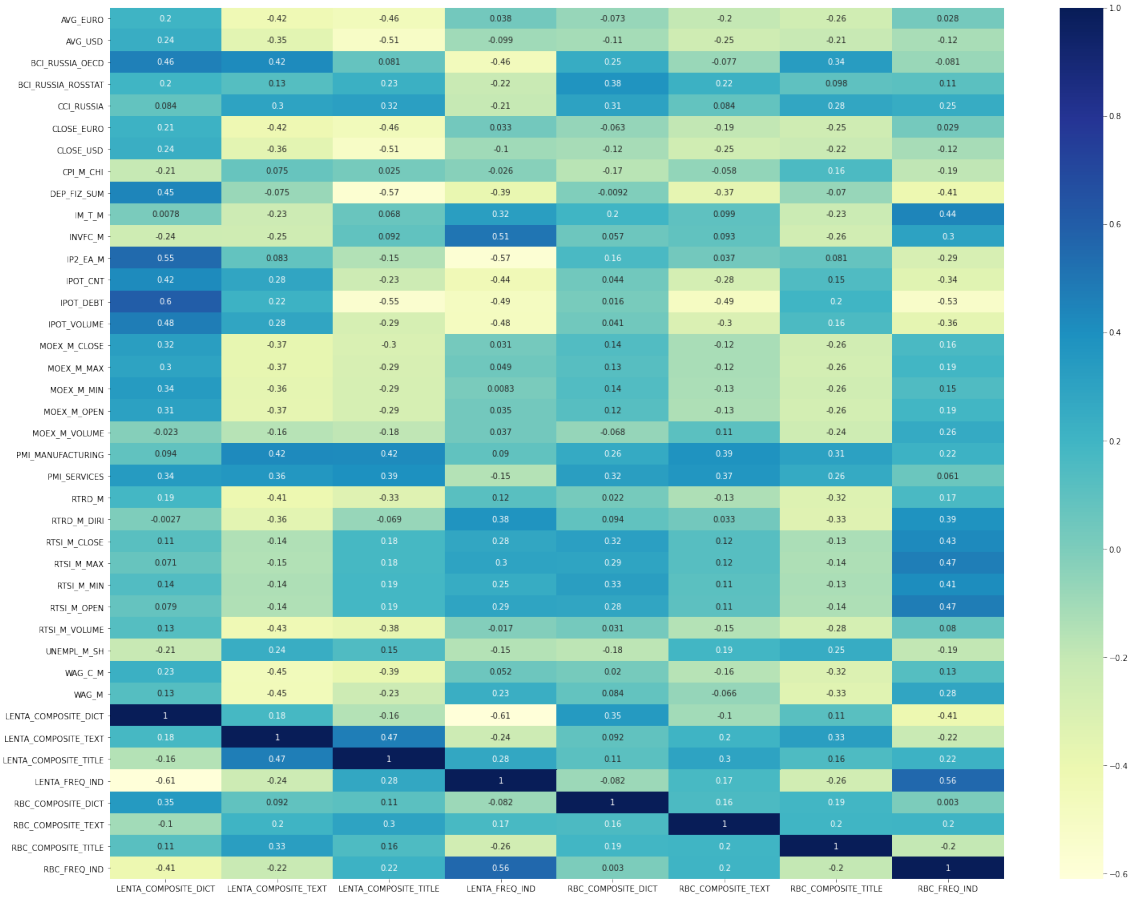

Рис. 5: Часть матрицы парных коэффициентов корреляции

Подводя итоги, наибольшую взаимосвязь с показателями финансового рынка демонстрируют композитные индексы, значения находятся на уровне 0.2 - 0.4. Возникает гипотеза о возможности частотных индексов качественно описать и спрогнозировать ситуацию на валютном рынке, а в роли предикторов для показателей предпринимательской и инвестиционной активности могут выступить «позитивные индексы».

Большинство коэффициентов корреляции являются значимыми на уровне 0.05, о чем говорит матрица их значимости, изображенная на рисунке 6.

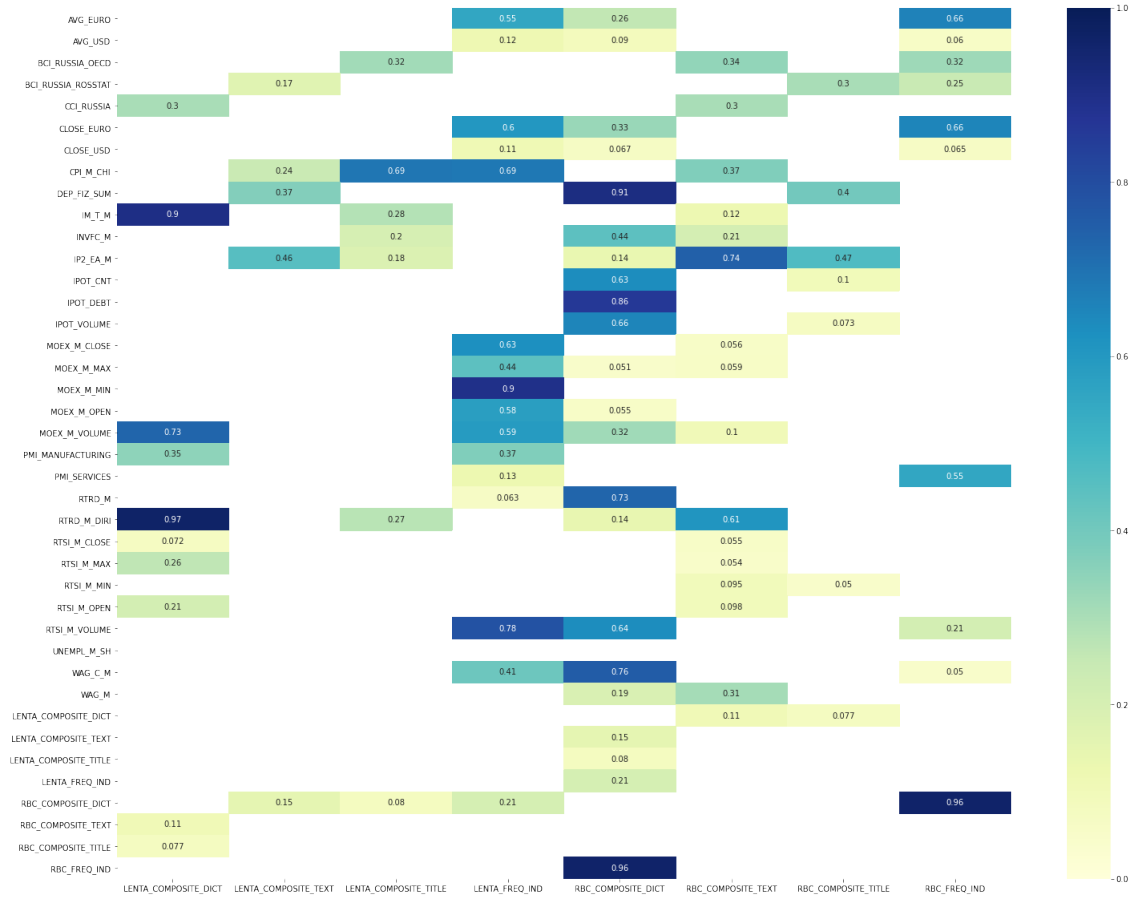

Рис. 6: Матрица $p - value$ парных коэффициентов корреляции для *необработанных* данных. Пропущенные ячейки означают значимость коэффициента на уровне 0.05

2.5 Причинность по Грейнджеру

Для того, чтобы проверить причинно-следственную зависимость между «автоматическими» индексами и целевыми показателями, воспользуемся тестом на причинность по Грейнджеру. Его суть состоит в том, чтобы узнать, предшествуют ли изменения X изменениям Y при условии того, что X влияет на Y и Y не вносит значимый вклад в прогноз X. При выполнении данного теста ставится нулевая гипотеза H_0 о том, что «X не влияет на Y», оценивается данное регрессионное уравнение:

$$y_t = \alpha_0 + \sum_{i=1}^{m}(\alpha_i y_{t-i}) + \sum_{i=1}^{m}(\beta_i x_{t-i}) + \varepsilon_t$$

При этом, изначальная нулевая гипотеза принимает следующий вид: $H_0 : \beta_1 = \ldots = \beta_m = 0$. Для ее проверки применяется F-тест. Далее рассмотрим ситуации с 1-м и 12-ю лагами:

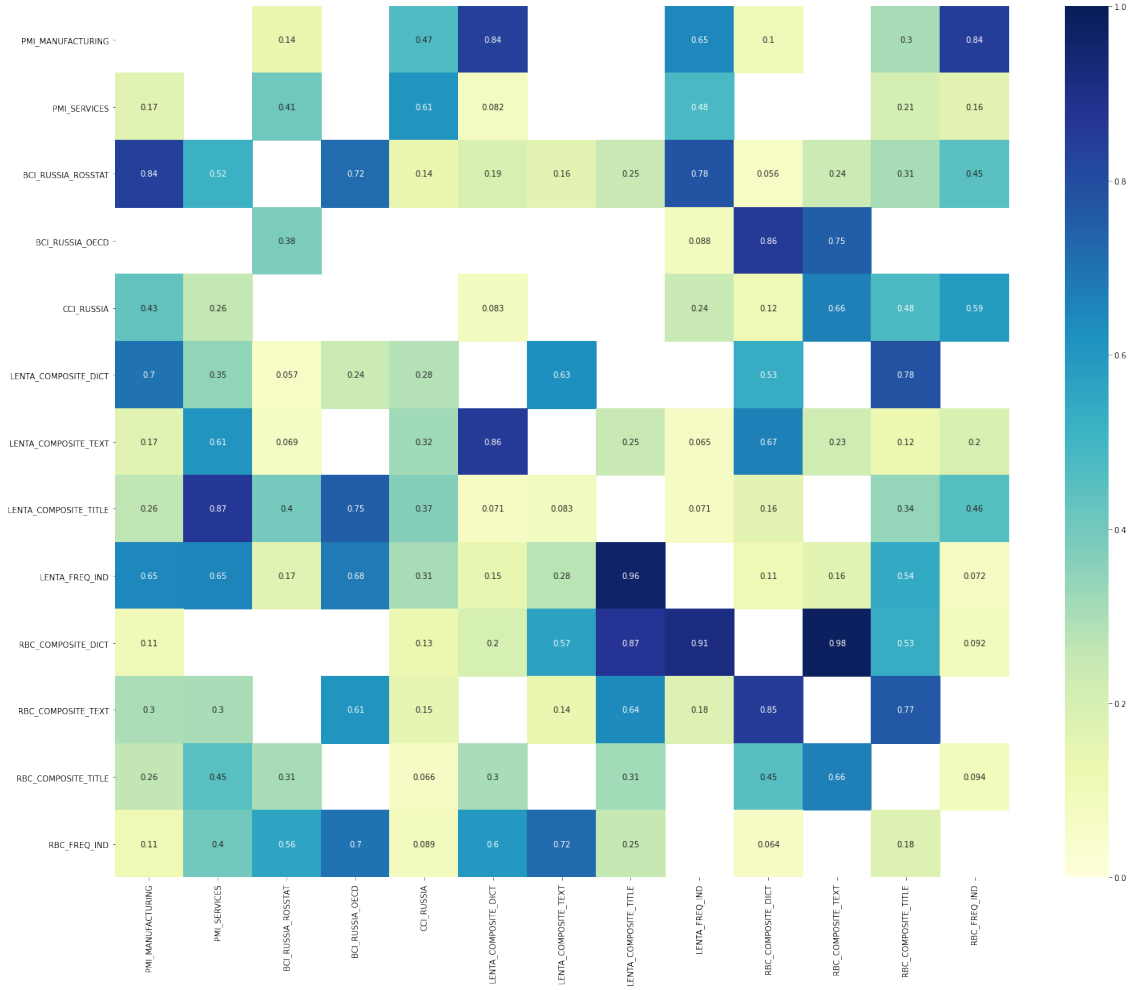

Рис. 7: Матрица $p - value$ тестирования причинности по Грейнджеру с **1**-м лагом *на необработанных* данных. Пропущенные ячейки означают, что временной ряд в столбце j является причиной по Грейнджеру временного ряда в строке i на уровне значимости 0.05

На рисунке 7 указаны значения p-value с 1-м лагом для каждой пары индексов. В данной ситуации частотные индексы не являются причинными для любых из «ручных», композитный индекс статей «РБК» (RBC Composite Dict), составленный по Тональному словарю, является причинным по Грейнджеру для обоих индикаторов предпринимательской уверенности (BCI Russia Rosstat и BCI Russia OECD), аналогичный ин-

декс, составленный при помощи RuSentiment модели (RBC Composite Title) оказался причинным только для ИПУ, полученного по методологии OECD (BCI Russia OECD).

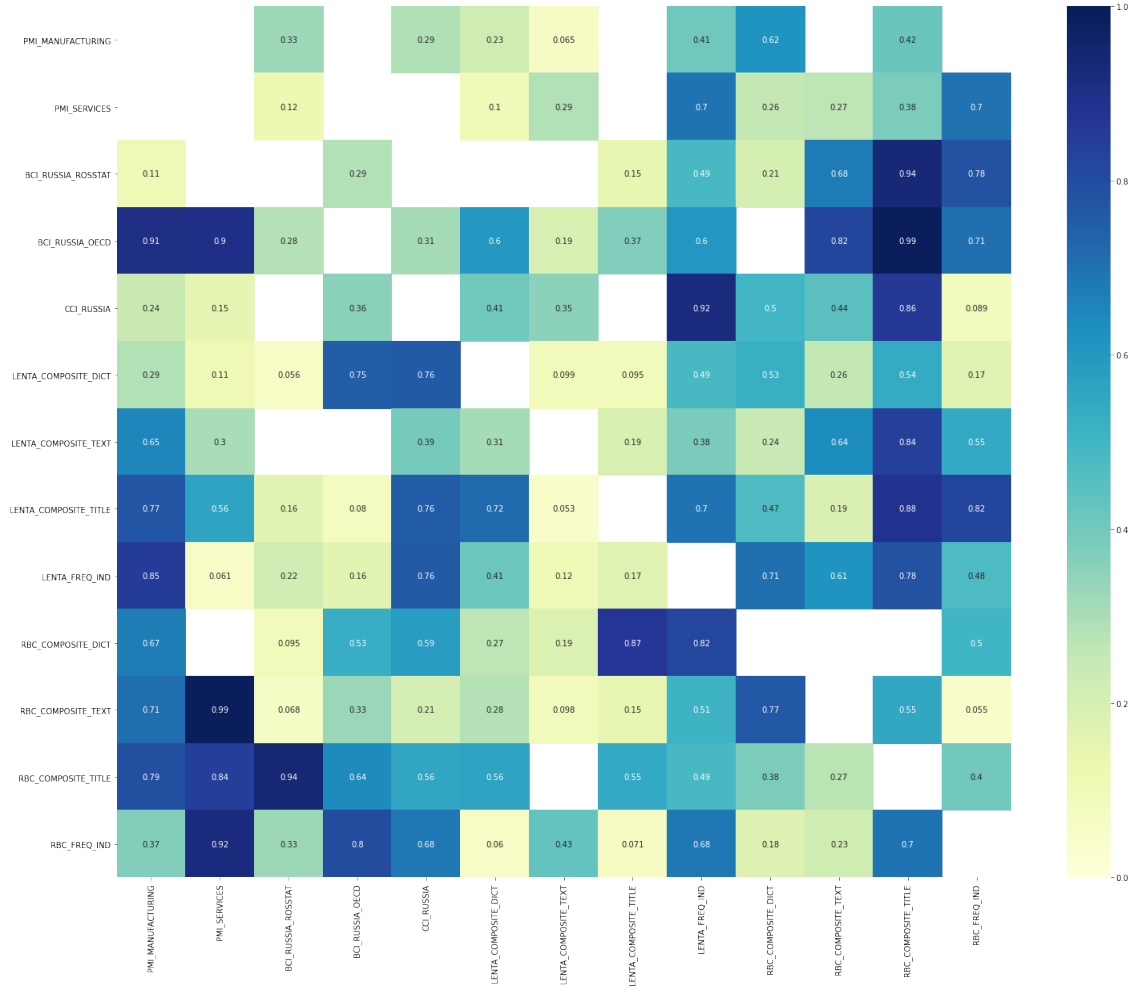

Рис. 8: Матрица $p - value$ тестирования причинности по Грейнджеру с **12**-ю лагами *на необработанных* данных. Пропущенные ячейки означают, что временной ряд в столбце j является причиной по Грейнджеру временного ряда в строке i на уровне значимости 0.05

Рисунок 8 отражает полученные при 12-ти лагах значения p-value. Они указывают на то, что композитный индекс статей «Лента» по RuSentiment модели (Lenta Composite Text) является причинным для ИПУ по методологии OECD, композитный индекс статей «РБК», полученный при помощи Тонального словаря, является причинным для показателя предпринимательской активности в сфере услуг (PMI Services).

Нельзя не упомянуть, что тест Грейнджера может лишь указать на вероятность причинно-следственной связи между переменными. При анализе ситуаций с различными вариантами лагов мы заметили, что в большинстве случаев возникает причинность для показателей предпринимательской уверенности, которые, в свою очередь, формируются при помощи опросов топ-менеджмента крупных фирм. Так как ответы на подобные опросы содержат суждения, основанные на личном опыте и информации из новостей, подобная взаимосвязь вполне обоснована. Таким образом, мы получили пары «автоматических» и «ручных» индексов, для которых существует возможность прогнозирования при помощи данных новостей, во всех указанных парах присутствуют композитные индикаторы, что может говорить о предсказательной силе последних.

2.6 Модели для прогнозирования и метрики качества

Основной целью нашего анализа является выделение тех ситуаций, когда применение признаков, построенных на данных из новостей, дает улучшение в объяснении (для задачи «nowcasting») или прогнозировании (для задачи «forecasting») целевых переменных. В качестве метрики для вычисления ошибки была выбрана MAE (Mean Absolute Error):

$$\frac{1}{n}\sum_{i=1}^{n}|a(x_i)-y_i|$$

Данный выбор основан на малой чувствительности к выбросам. Метрика качества была определена как отношение MAE в модели с использованием новостного признака к MAE в базовой модели, что представляет из себя следующее:

$$\frac{MAE_{news}}{MAE_{basic}}$$

В качестве алгоритмов машинного обучения, при помощи которых строились модели как с использованием новостных признаков, так и без,

взяты следующие:

1. Линейная регрессия (Linear Regression). В основе работы данного алгоритма лежит формула $y = w_0 + X^T w$, где X^T – матрица значений признаков, w – вектор-столбец коэффициентов регрессии, w_0 – свободный коэффициент. При помощи минимизации выбранной функции потерь (например, MSE в модели OLS) на обучающей выборке находятся коэффициенты w и свободный член, далее предсказания регрессии проверяются на отложенной тестовой выборке.

2. Ridge и Lasso регрессии, Elastic Net. Данные модели являются производными от линейной регрессии, в которых применяется регуляризация, то есть система «штрафов» для борьбы с переобучением. В Ridge регрессии применяется регуляризатор $\lambda||w||^2$, а Lasso регрессия борется с переобучением при помощи функционала $\lambda||w||$, где λ – настраиваемый гиперпараметр. Elastic Net работает схожим образом, сочетая в себе функционалы Ridge и Lasso регрессий, то есть регуляризации $\lambda_1||w||^2$ и $\lambda_2||w||$ применяются вместе. Таким образом, в данной модели появляются два гиперпараметра — сила регуляризации α и относительная важность L_1-регуляризации β. При обучении данной модели использовались параметры $\alpha = 0{,}001$ и $\beta = 0{,}5$. В моделях Ridge и Lasso параметр $\alpha_{Ridge} = 0{,}001$ и $\alpha_{Lasso} = 0{,}0001$ соответственно. Низкие значение α обусловлены тем, что при высоких значениях модели слишком сильно регуляризировались, в результате чего прогнозы были практически константными (среднее значение).

3. SVM (Support Vector Machine) модель. Данный алгоритм решает задачу минимизации следующей функции потерь:

$$\frac{1}{2}w^T w + C\sum_{i=1}^{n}(\xi_i) + C\sum_{i=1}^{n}(\hat{\xi}_i)$$

при определенных условиях. Основными преимуществами SVM являются возможность работать с нелинейными отношениями между переменными и малая чувствительность к выбросам. SVM обучался со сле-

дующими параметрами: [$epsilon = 0, C = 10000, max\_iter = 100000$, $tol = 0.1$], остальные — по умолчанию в соответствие с библиотекой scikit-learn[8].

4. Случайный лес (Random Forest). Являясь одним из самых популярных и часто используемых алгоритмов машинного обучения. Случайный лес с большой долей вероятности может гарантировать неплохие результаты применительно к нашей задаче. В его основе лежит композиция выбранных моделей, чаще всего применяются деревья решений. Для каждого дерева выбирается случайное подмножество данных при помощи бутстрапа (booststrap), каждый предикат строится на ограниченном количестве признаков, лучший выбирается по наименьшему значению критерия Джинни. Ответы алгоритма выбираются как средние по композиции. Главным преимуществом Случайного леса является низкий разброс (variance) - устойчивость модели к изменениям в обучающей выборке. Обучался Случайный лес с параметрами по умолчанию[9].

5. Градиентный бустинг (Gradient boosting). Данный алгоритм является одним из самых продвинутых и современных. Он работает по схожей со Случайным лесом логике, однако в нем каждая из N моделей композиции строится не независимо, а на ошибках предыдущих по следующей логике:

$$\frac{1}{l}\sum_{i=1}^{l}(b_N(x_i) - s_i) \to min$$

, где $b_N(x_i)$ – ответ следующей модели, а $s_i = y_i - a_{N-1}(x_i)$ - остатки. Градиентный бустинг способен демонстрировать низкое смещение (bias), то есть качественно приближать наилучшую модель из всех возможных. Тем не менее, при использовании данного алгоритма возникает риск переобучения, для исправления данного недочета часто вводится регуляризация модели. Обучение градиентного бустинга проводилось с параметрами по умолчанию[10].

[8] scikit-learn.org/.../sklearn.svm.LinearSVR

[9] scikit-learn.org/.../sklearn.ensemble.RandomForestRegressor

[10] scikit-learn.org/.../sklearn.ensemble.GradientBoostingRegressor

6. Многослойный перцептрон (MLP). В качестве нейросетевого алгоритма был выбран многослойный перцептрон с $L = 1$ скрытым слоем, $k^{(1)} = 64$ нейронами и функцией активации $ReLU$: $\sigma(x) = max(0, x)$. Основной принцип работы алгоритма заключается в построении линейных слоев, от которых берется нелинейная функция. Слой l и $l-1$ связаны следующим образом:

$$z^{(l-1)} = w_0^{(l-1)} + \sum_{i=1}^{k^{(l-1)}} \beta_i^{(l-1)} x_i^{(l-1)}$$
$$x^{(l)} = \sigma(z^{(l-1)})$$

Его обучение производилось методом оптимизации L-BFGS с шагом обучения $\eta = 0.01$. Остальные параметры оставлены по умолчанию[11].

Для улучшения обучения и стабильности работы перед использованием алгоритмов SVM и MLP входные данные дополнительно нормализовывались путем вычитания среднего и деления на стандартное отклонение.

2.7 Текущая оценка и прогнозирование

В данном разделе мы сравним качество работы выбранных алгоритмов машинного обучения с использованием новостных признаков и без них. При оценке моделей использовалось скользящее окно длинной $m = 24$ месяца, которое разбивается на обучающую часть, первые 12 месяцев, и тестовую — следующие 12 месяцев. Оценка метрики MAE осуществлялась путем конкатенации всех предсказаний модели на указанный горизонт прогнозирования (по всем скользящим окнам) и сравнения с исходными значениями. Как целевые, так и новостные показатели были нормализованы скользящим окном 24 месяца, как было описано в разделе целевых переменных.

Начнем с задачи nowcasting, решение которой формулируется следу-

[11] scikit-learn.org/.../sklearn.neural_network.MLPRegressor

ющим образом:

$$y_t = f_{ML}(y_{t-1}, x_t)$$

Результаты применения моделей в указанной задаче приведены на рисунке 9, рассмотрим их поочередно для каждой из групп целевых переменных:

1. Курсы валют. При прогнозировании средний цены и цены закрытия доллара США (AVG USD и CLOSE USD) наибольшего улучшения в качестве добилась MLP модель, оно достигает 30-ти пунктов. Практически все созданные нами «автоматические» индексы показали приемлемый результат в данном случае, однако наибольший прирост качества происходит при включении в модель композитных индексов, созданных на базе статей «Лента» (Lenta Composite Dict и Lenta Composite Title). Индекс Lenta Composite Text также показал хороший результат при прогнозировании средней цены доллара США, но не справился с задачей прогнозирования цены закрытия. Это может быть связано с тем, что основные части статей содержат в себе больше оценочных и прогнозных суждений, которые могут угадать тенденцию, но не последние изменения. При прогнозировании показателей, связанных с курсом евро, композитные индексы аналогично привнесли наибольшие улучшения в качество работы моделей, однако лучшими алгоритмами в данной ситуации стали Случайный лес и Градиентный бустинг.

2. Индикаторы финансового рынка. Применительно к задаче прогнозирования показателей доходности по индексу МосБиржи (MOEX M Open, MOEX M Close, MOEX M Max, MOEX M Min, MOEX M Volume), наибольшее улучшение происходило при работе со Случайным лесом и Бустингом, метрика качества показала лучшие результаты при применении частотных и композитных (с применением Тонального словаря) индексов. Похожие выводы были получены при прогнозировании показателей доходности по РТС (RTS M Open, RTS M Close, RTS M Max, RTS M Min, RTS M Volume), однако при их предсказании наилучшей моде-

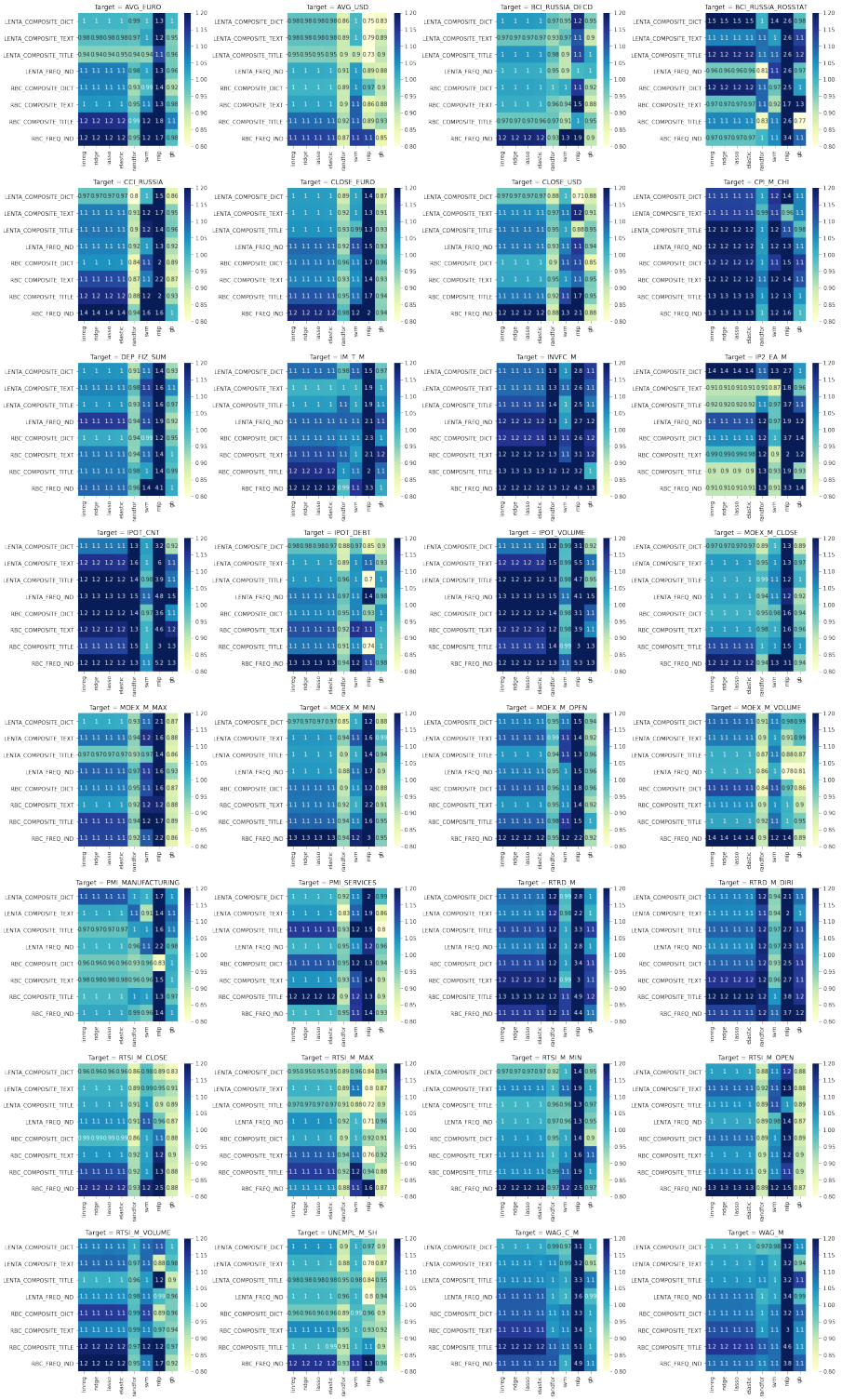

Рис. 9: Сравнение качества текущей оценки прогноза модели с новостями и без при помощи отношения $\frac{MAE_{news}}{MAE}$. Показатель меньше 1 означает, что при прочих равных добавление новостных данных в среднем улучшает качество прогнозирования.

лью оказалась MLP. В целом отношение MAE в упомянутых моделях упало на 0.05 - 0.2 пункта.

3. Индексы настроений и деловой активности. Наибольший прирост качества при прогнозировании индикаторов деловой активности (PMI Manufacturing и PMI Services) с использованием новостей продемонстрировали SVM и MLP в случае предсказания показателя, относящегося к производству, Случайный лес и Градиентный бустинг для сферы услуг. Улучшение достигало 0.2 пунктов, лучший результат также продемонстрировали композитные индексы. Как ни странно, для показателей предпринимательской уверенности, рассчитанных по разным методикам, результаты сильно разнятся. В случае расчета индекса по методике OECD (BCI Russia OECD), наилучшими оказались композитные индексы при применении SVM и Градиентного бустинга, тогда как при расчете по методики Росстата (BCI Russia Rosstat) лучший результат был продемонстрирован частотным индексом по статьям «Лента» и композитным по заголовкам «РБК».

4. Макроэкономические показатели. В задаче nowcasting для данного вида индикаторов лишь в некоторых случаях MAE в моделях с использованием новостей оказалось ниже MAE без них. Таким образом, качественное улучшение в предсказании произошло при прогнозировании индекса промышленного производства (IP2 EA M) с использованием всех видов регрессий, SVM и Случайного леса в сочетании с композитными индексами по RuSentiment модели, частотным индексом по статьям «РБК». Применение новостей при предсказании индекса реального оборота розничной торговли (RTRD M DIRI) оказалось действенным лишь для алгоритма SVM, а улучшение для задачи предсказания уровня безработицы произошло для всех моделей и «автоматических» индексов, среди них особым образом выделились Случайный лес и MLP, улучшение составило до 0.12-ти и 0.22-х пунктов соответственно. Таким образом, только немногие из рассмотренных макроэкономических рядов подходят для использования в качестве целевых переменных в задаче nowcasting с использованием новостных индикаторов.

5. Депозиты и ипотечный рынок. Среди показателей данной группы только для двух произошло падение MAE при включении новостного признака. Так, для суммы депозитов физических лиц (Dep Fiz Sum) наибольшее улучшение произошло при применении композитного индекса, основанного на Тональном словаре, с использованием алгоритма Случайный лес. Задолженность по ипотечным кредитам (Ipot Debt) лучшим образом предсказывается при помощи модели MLP и заголовков статей (Composite Title), значение метрики качества достигает 0.3-х.

Далее мы рассмотрим значения метрики качества для задачи forecasting с прогнозированием на 2, 5, 8 и 11 месяцев вперед. В данном случае алгоритмы машинного обучения решают следующую задачу:

$$y_{t+k} = f_{ML}(y_{t-1}, x_t)$$

, где $k = 2, 5, 8, 11$ соответственно.

На рисунке 10 указаны значения метрики качества для агрегированных MAE по указанным периодам. Обозначим общие тенденции в задаче прогнозирования:

1. Для показателей валютного рынка возможно применение композитных и частотных новостных индексов, они существенно улучшают прогнозирование в случае применения Случайного леса и Градиентного бустинга, также возможен прирост метрики качества при прогнозировании средней цены и цены закрытия доллара США с использованием модели MLP. В указанных случаях MAE с использованием новостей уменьшилось на 0.1 - 0.4 пункта по сравнению с «базовым» MAE, данный результат довольно значим.

2. Прогнозирование индикаторов фондового рынка также существенно улучшается при использовании Бустинга и Случайного леса с любыми индексами, в случае прогнозирования среднего объема торгов и максимальной цены по индексу РТС лучшие результаты демонстрирует модель MLP. В среднем качество предсказания выросло на 0.15 пунктов, в случае прогнозирования показателей РТС с использованием MLP оно

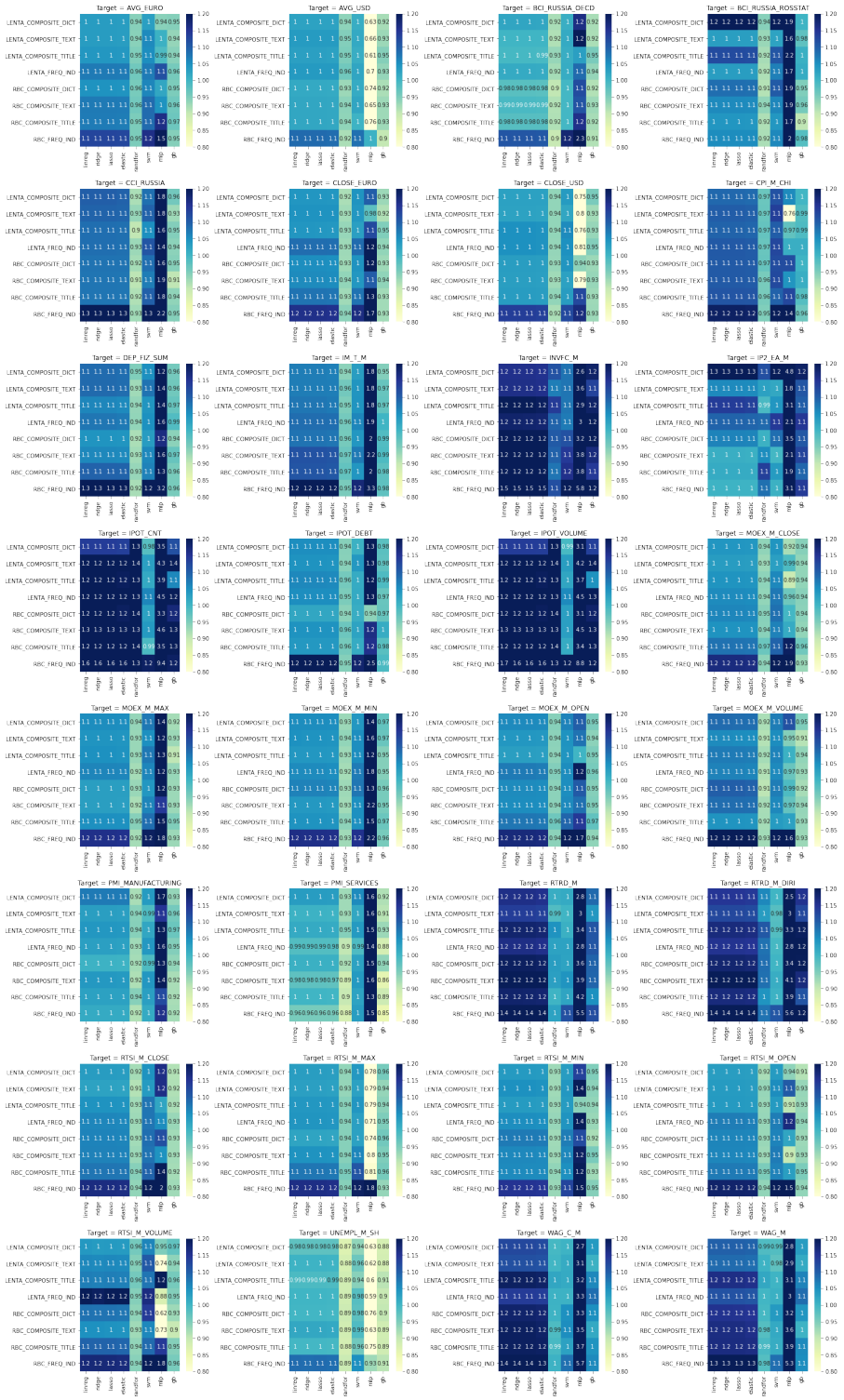

Рис. 10: Сравнение качества оценки прогноза модели на всех горизонтах прогнозирования ($k = 0, 1, ..., 11$) с новостями и без при помощи отношения $\frac{MAE_{news}}{MAE}$. Показатель меньше 1 означает, что при прочих равных добавление новостных данных в среднем улучшает качество прогнозирования.

достигало 0.3.

3. При рассмотрении индикаторов деловой активности и настроений бизнеса использование новостей также привносит уменьшение MAE примерно на 10 процентов, аналогично другим индексам наилучшими моделями оказываются Градиентный бустинг и Случайный лес.

4. В отличие от задачи nowcasting, макроэкономическими показателями, при прогнозировании которых с применением «автоматических» индексов качество предсказания улучшилось, оказались индекс потребительских цен (CPI M CHI), импорт (IM T M) и уровень безработицы (Unempl M SH). При использовании MLP, Бустинга и Случайного леса улучшение в качестве варьировалось от 0.5 до 0.4.

5. Качество прогнозирования показателей депозитного и ипотечного рынков при использовании новостей выросло незначительно, прирост в среднем составил 5 процентов.

В итоге проведенного анализа можно сделать вывод о том, что применение новостных индексов может быть целесообразно во многих задачах прогнозирования. Их наличие в определенных моделях машинного обучения повышает качество предсказания минимум на 5 процентов. Наиболее адаптированными для применения новостных признаков моделями оказались Градиентный Бустинг и Случайный лес. Во многих случаях именно они демонстрировали лучшие результаты. Также было выявлено, что для прогнозирования некоторых целевых переменных наилучшей моделью для включения в нее новостного признака может стать Многослойный перцептон (MLP), качество предсказания возрастает до 40-а процентов.

Стоит отметить наличие определенной разницы в результатах при реализации задач nowcasting и forecasting. Получение текущей оценки (nowcasting) требует более тщательного подбора «автоматического» индекса, не для каждого из них вырастет качество предсказания, а при прогнозировании (forecasting) различия между значениями метрики качества с тем или иным новостным признаком малосущественны. Тем не менее, общим для двух указанных задач является тот факт, что не для

всех целевых переменных и моделей машинного обучения обосновано применение новостных признаков.

Наиболее подходящими для прогнозирования и получения текущей оценки целевыми переменными оказались показатели валютного и фондового рынка, индикаторы настроений бизнеса и деловой активности, некоторые из макроэкономических рядов, лучшие результаты продемонстрировало предсказание среднего курса доллара США, уровня безработицы и характеристик доходности по индексу МосБиржи. Именно для данных показателей использование новостных признаков было определено как соответствующее решение для прироста в качестве. Для задачи nowcasting наиболее оптимально использовать композитные индексы, тогда как частотные подходят в редких случаях, таких как прогнозирование курса доллара США.

2.7.1 Прогнозирование с TF-IDF

На рисунке 11 изображено сравнение моделей с TF-IDF признаками и без них в задаче текущей оценки прогноза[12]. Модели с новостными признаками могут улучшить качество текущей оценки почти для всех показателей, кроме среднего значения курса евро (AVG_EURO) и индекса промышленной активности (PMI_MANUFACTURING). Как и с другими новостными индексами лучшими для прогнозирования с новостями моделями являются градиентный бустинг и случайный лес. Остальные модели более подвержены факторам конкретного целевого или признакового показателей. Тем не менее, в некоторых ситуациях наибольший прирост качества происходит при использовании Многослойного перцептона.

Рисунок 12 показывает относительное улучшение/ухудшение прогноза для всех рассматриваемых горизонтов. Выводы относительно моделей-фаворитов при оценке будущего повторяются. Можно отметить, что при прогнозировании новости еще сильнее улучшают прогноз для индекса

[12]Модели ElasticNet и Gradient Boosting с TF-IDF признаками отсутствуют в связи с непреодолимыми техническими ограничениями

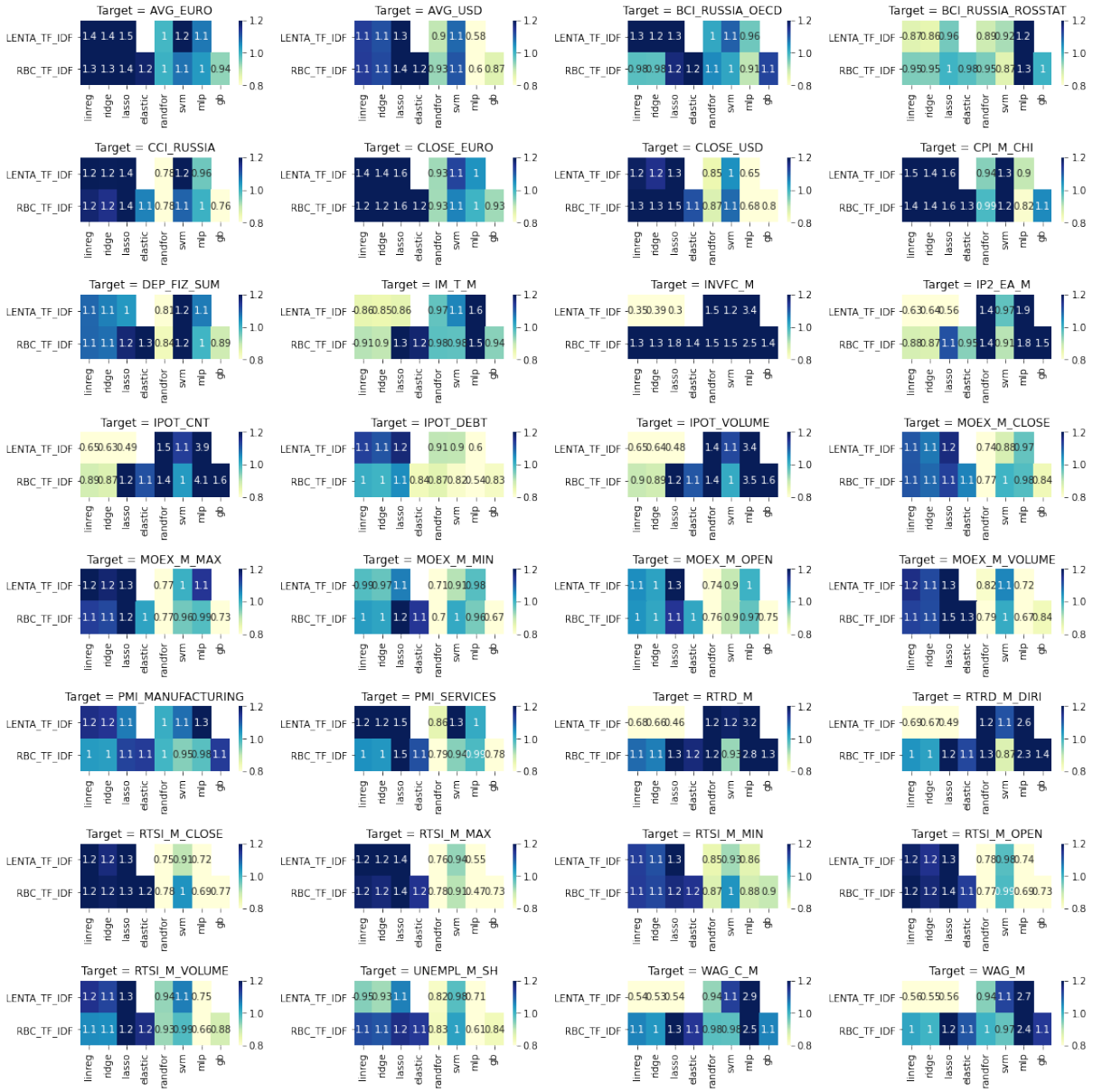

Рис. 11: Сравнение качества текущей оценки прогноза модели с TF-IDF признаками и без при помощи отношения $\frac{MAE_{news}}{MAE}$. Показатель меньше 1 означает, что при прочих равных добавление новостных данных в среднем улучшает качество прогнозирования.

предпринимательской уверенности по методике OECD (BCI_RUSSIA_OECD), для уровня безработицы (UNEMLP_M_SH), среднего курса доллара (AVG_USD) и отметки закрытия (CLOSE_USD) в сравнении с текущей оценкой, что может говорить о том, что новости могут также иметь долгосрочные инсайты.

2.7.2 Тест Диболда-Мариано

Тест Диболда-Мариано, предложенный в статье Diebold, Mariano, 1995, позволяет сравнить качество прогнозирования двух моделей. Считается разность остатков моделей:

$$d = \frac{1}{T} \sum_{i=t}^{T} |e_t^{(1)} - e_t^{(2)}|$$

Причем d может рассчитываться не только с помощью MAE, как в выражении выше, но и с помощью MSE, MAPE или Poly[13] критериев.

В нулевой гипотезе утверждается, что $H_0 \colon \mathbb{E}[d] = 0$. Для ее проверки составляется DW статистика с помощью автоковариационной функции и поправки Харвея, которая подчиняется распределению Стьюдента с $T - 1$ степенями свободы.

Результаты тестирования для текущей оценки прогноза (nowcasting) представлены на рисунке 13.

Хорошо видно, как тест себя показывает на показателе инвестиций в капитал INVFC_M, где вся матрица получилось замаскированной. Если обратиться к рисунку 9, то станет понятно, что нулевая гипотеза отвергается на уровне значимости 0.05, а значит, для данного показателя добавление новостных данных ухудшает MAE.

Как правило, если MAE модели с новостями улучшается примерно на 10%, то это изменение значимо на уровне 0.05. Наибольшее количество значимых улучшений при использовании новостных данных наблюдается для моделей Random Forest, Gradient Boosting. Данный вывод

[13] $d_k = \frac{1}{T} \sum_{t=1}^{T} (e_t^{(1)} - e_t^{(2)})^k$

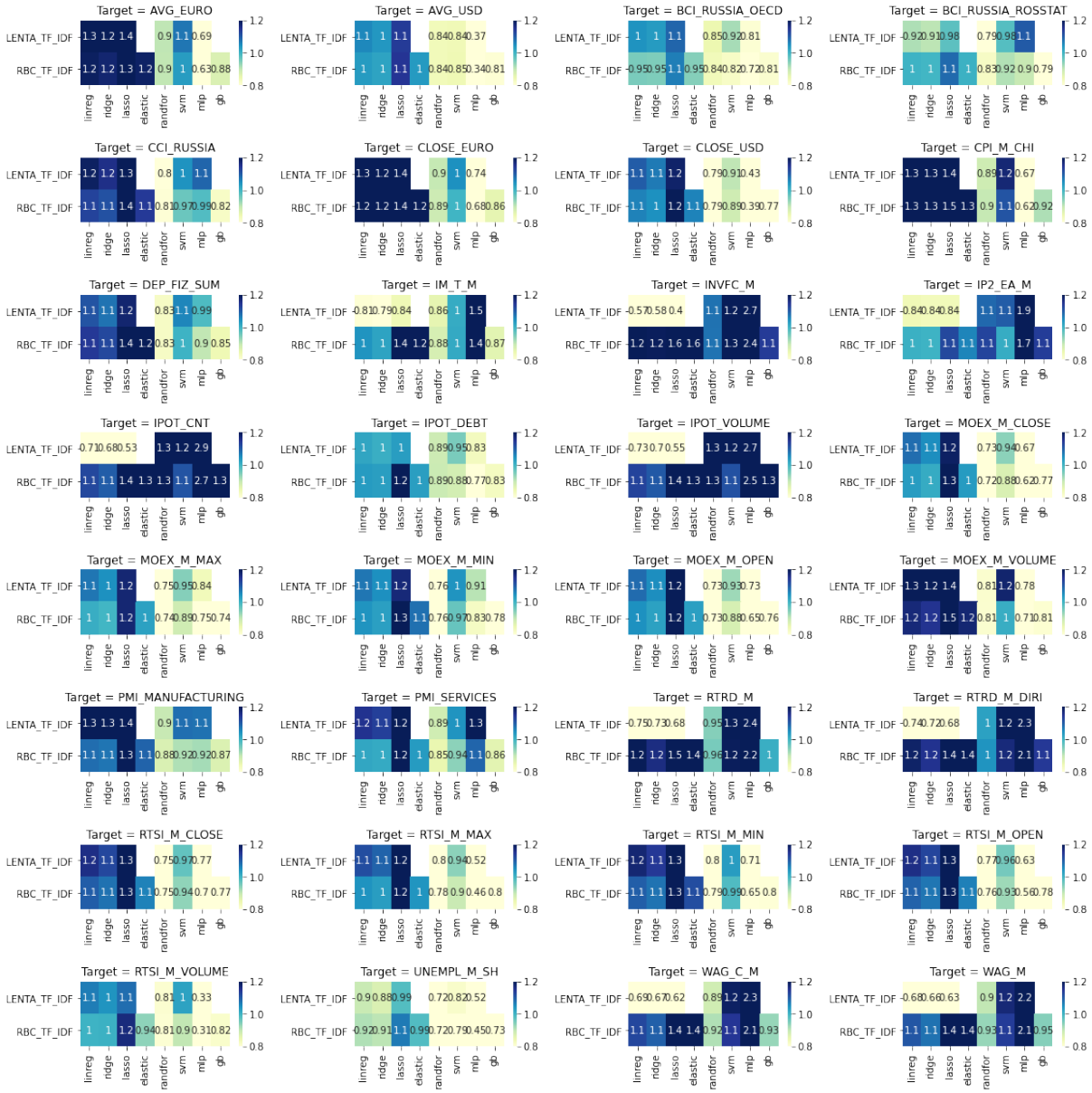

Рис. 12: Сравнение качества прогнозирования на всех горизонтах ($k = 0, 1, ..., 11$) модели с TF-IDF признаками и без при помощи отношения $\frac{MAE_{news}}{MAE}$. Показатель меньше 1 означает, что при прочих равных добавление новостных данных в среднем улучшает качество прогнозирования.

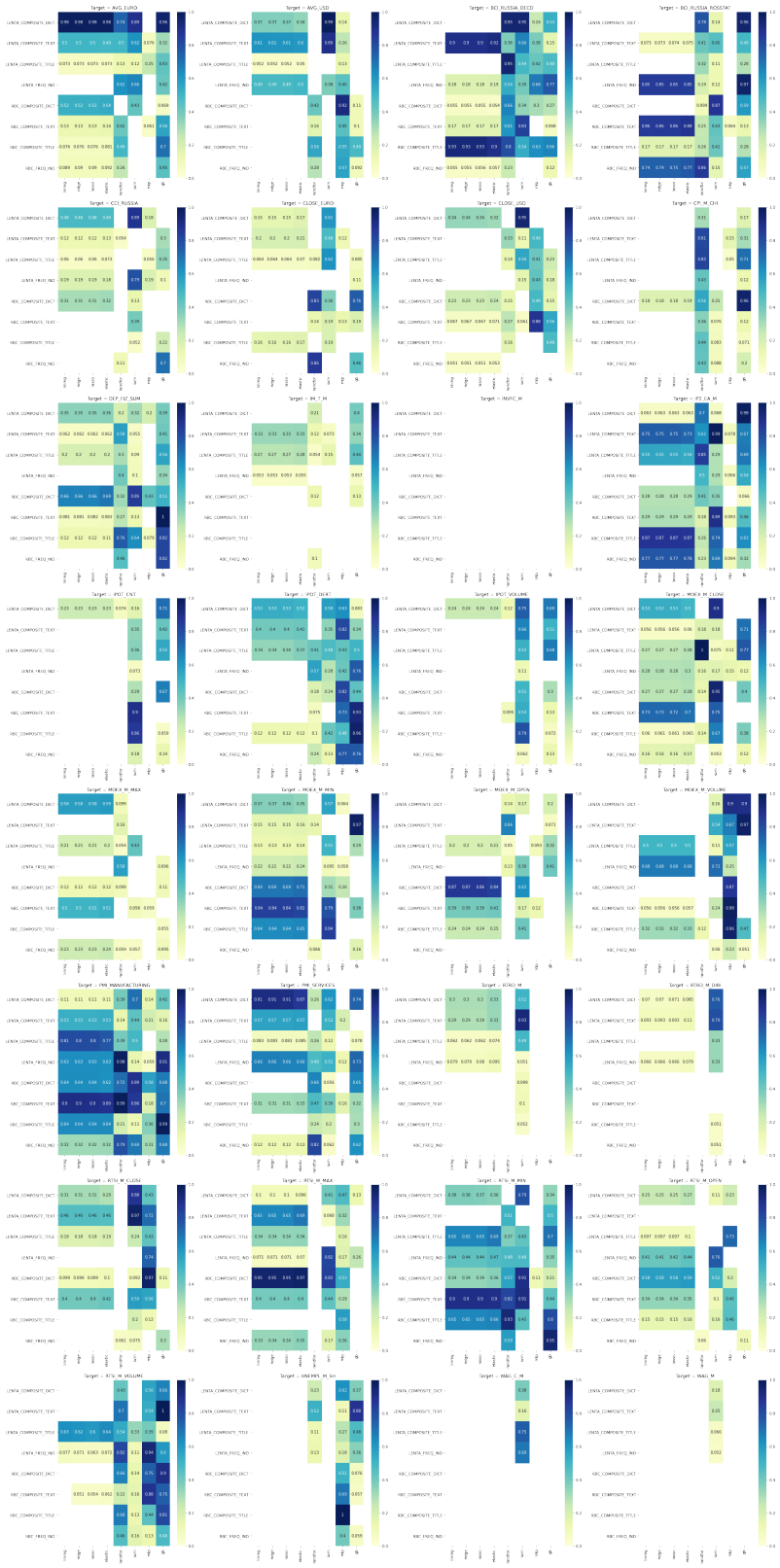

Рис. 13: Тест Диболда-Мариано для текущей оценки. Пропущенные ячейки означают значимые на уровне 0.05 отличия в текущей оценке (nowcasting) модели с новостями против модель без новостей

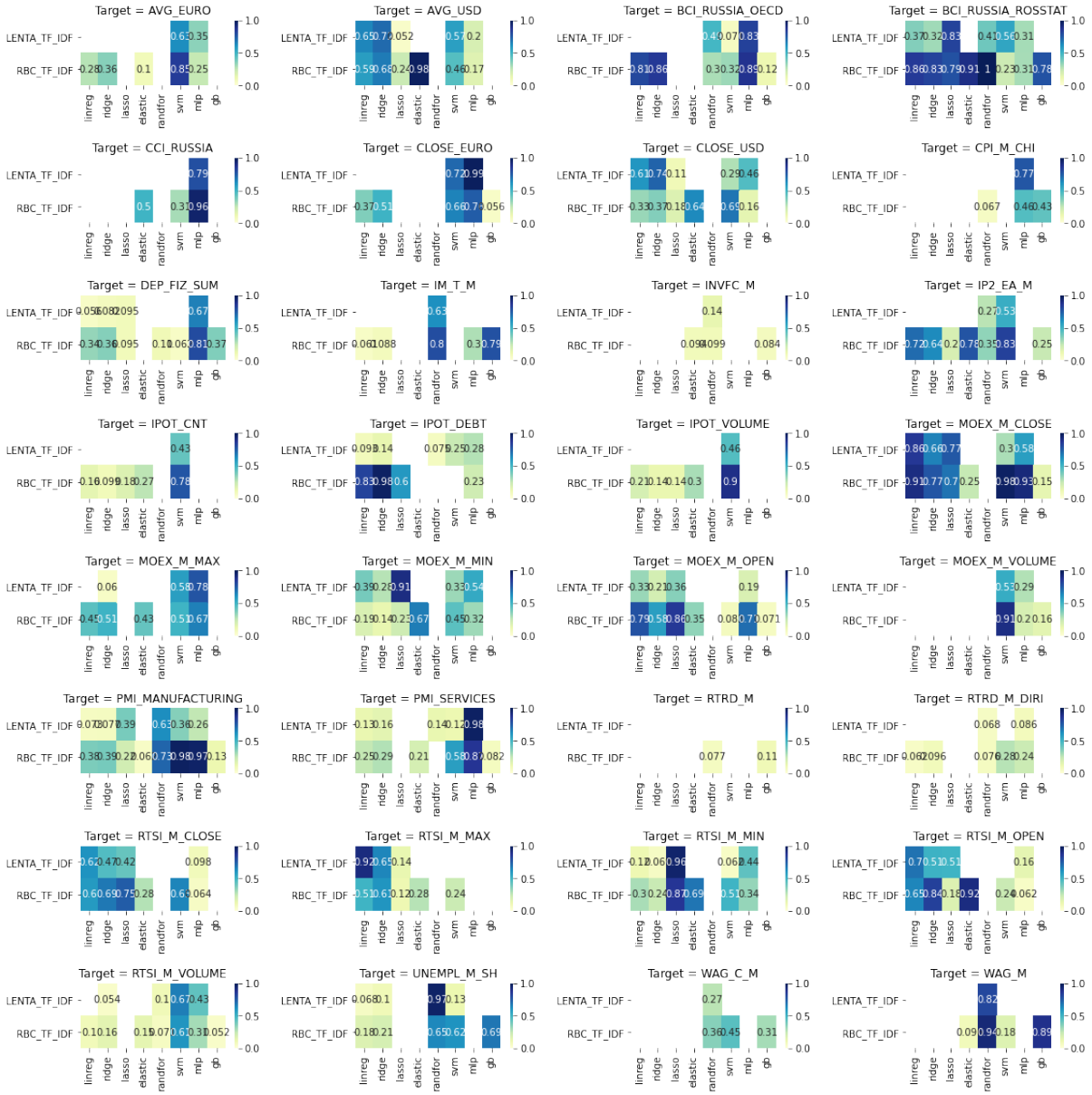

Рис. 14: Тест Диболда-Мариано для текущей оценки с помощью TF-IDF признаков. Пропущенные ячейки означают значимые на уровне 0.05 отличия в текущей оценке (nowcasting) модели с новостями против модель без новостей

также характерен и для прогнозирования в будущее.

3 Заключение

В данной работе мы построили несколько новостных признаков для их дальнейшего использования в моделях машинного обучения. Были использованы fast-text модель «Dostoevsky», Тональный словарь русского языка «КартаСловСент», применялся подсчет токенов для построения частотного индекса, также рассматривалось построение признаков на основе методики TF-IDF.

Некоторые из построенных композитных индексов оказались причинными по Грейнджеру для целевых переменных, значения которых исходили из различных опросов. Была сформулирована гипотеза о том, что это происходит ввиду существенного влияния экономических новостей на формирование мнения респондентов, которое так или иначе отражается в результатах опросов.

Оценка моделей с новостными признаками и без них позволила установить, что для некоторых моделей, в частности, Random Forest и Gradient Boosting добавление новостных признаков значимо улучшает прогноз и текущую оценку, уменьшая MAE приблизительно на 10% в терминах стандартизованных единиц. Данный факт также был подтвержден при проведении теста Диболда-Мариано.

Таким образом, новостная информация может содержать полезные признаки для прогнозирования и текущей оценки макроэкономических и финансовых показателей, часто изменения являются статистически значимыми.

Перспективным направлениям в улучшении как моделей, так и методов сбора новостной информации является глубинное обучение, однако в связи с техническими ограничениями применение нейронных сетей не нашло распространения. Именно поэтому модели машинного обучения все еще представляют интерес для изучения в прогнозировании.

Список литературы

Магнус Я. Р., *Катышев П. К.*, *Пересецкий А. А.* Эконометрика: начальный курс. — Дело, 2004.

Ульянкин Ф. Прогнозирование российских макроэкономических показателей на основе информации в новостях и поисковых запросах // Russian Journal of Money and Finance. — 2020. — Т. 79, № 4. — URL: https://rjmf.econs.online/2020/4/forecasting-macroeconomic-indicators-news-and-search-queries/.

Kalamara E., *Turrell A.*, *Redl C.*, *Kapetanios G.*, *Kapadia S.* Making text count: economic forecasting using newspaper text. — 2020. — URL: https://www.bankofengland.co.uk/working-paper/2020/making-text-count-economic-forecasting-using-newspaper-text.

Diebold F., *Mariano R.* Comparing predictive accuracy // Journal of business and Economics Statistics. — 1995. — Т. 13.

Rogers A., *Romanov A.*, *Rumshisky A.*, *Volkova S.*, *Gronas M.*, *Gribov A.* RuSentiment: An Enriched Sentiment Analysis Dataset for Social Media in Russian // Proceedings of the 27th International Conference on Computational Linguistics. — Santa Fe, New Mexico, USA : Association for Computational Linguistics, 08.2018. — С. 755—763. — URL: https://www.aclweb.org/anthology/C18-1064.

Segalovich I. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. // MLMTA. — 2003. — Т. 2003. — С. 273.

Stolbov M. [и др.]. Statistics of search queries in Google as an indicator of financial conditions // VOPROSY ECONOMIKI. — 2011. — Т. 11.