МОДЕЛЬ ПРОГНОЗИРОВАНИЯ ЭКОНОМИЧЕСКИХ ПОКАЗАТЕЛЕЙ ПРЕДПРИЯТИЯ С ПОМОЩЬЮ КОНФОРМНЫХ ПРЕДИКТОРОВ

Сафронов В.С., Дранко О.И.,
Институт проблем управления им. В.А. Трапезникова РАН, Москва, Россия
safronov@phystech.edu, oleg.dranko@gmail.com
Черноглазов И.М.,
Финансовый университет при Правительстве РФ, Москва, Россия
211305@edu.fa.ru
Черников А.А., Костиков К.О., Петрыкина И.Р., Никитин А.В., Аринин М.А., Брагин А.В.
МГТУ им. Баумана, Москва, Россия
list.kedra.79@gmail.com, kostickov.costya@yandex.ru, petrykina.i57@gmail.com, nikitinnvlalex@gmail.com, arinin.matvey@yandex.ru research@alexbragin.com

Аннотация. Представлена модель прогнозирования экономических показателей предприятий с использованием конформных предикторов для оценки уверенности прогноза. Конформные предикторы дают индивидуальные предсказательные интервалы, оценивая неопределенность данных и модели, что особенно важно в экономических прогнозах. Показано, что средняя уверенность моделей прогнозирования выручки не зависит от размера бизнеса, тогда как дисперсия уверенности выше для мало и среднего бизнеса. Это может использоваться для выявления некорректной отчетности. Также выявлено, что прогнозирование совокупной выручки группы предприятий в рамках одного ОКВЭД может повысить уверенность моделей.

Ключевые слова: прогнозирование, регрессионный анализ, финансовая отчетность, оценка уверенности, конформные предикторы, машинное обучение.

Введение

Управление крупномасштабными экономическими системами на любом уровне рассмотрения «предприятие – холдинг – отрасль» является важным аспектом для успешного и устойчивого управления страной. В значительной мере управление элементами экономической экосистемы опирается на построение моделей, способных связывать план по экономическим показателям, их фактические значения и эффект управленческих решений для последующего анализа стратегии управления. В контексте подобного моделирования и анализа возникает вопрос доверия к построенным моделям, т.к. отклонение фактических показателей от плана может происходить как результат управленческих решений, так и вследствие неточности моделей. Наличие метода численной оценки уверенности моделей позволило бы разделить отклонения показателей предприятия от прогнозных на смещения вследствие ошибок управления и на смещения вследствие несовершенства самих моделей прогноза.

В данной статье рассматривается подход к применению численной оценки уверенности прогностических моделей с применением интервалов прогноза (predictioninterval). В качестве задачи экономического моделирования рассматривается прогнозирование одного из ключевых параметров в крупномасштабной системе – выручки, как на уровне отдельных предприятий (выборка 16 тыс. предприятий РФ) и на уровне отрасли. Построение интервалов прогноза опирается на теорию конформных предикторов и позволяет получать одновременно как диапазон значений, в котором с определенной вероятностью находится истинное значение предсказываемой переменной (выручки), так и меру оценки уверенности модели, в данном случае регрессионной.

Важным замечанием является то, что конформные предикторы являются надстройкой над основной (регрессионной) моделью, и их задачей является построение диапазона(набора) значений целевой величины, а не задача непосредственной оценки значений. Другими словами, регрессионные модели дают на выходе точечный прогноз, а конформные предикторы строятся на основе общей с ними выпуклой оболочки и признакового пространства, и дают на выходе интервалы значений прогнозируемой величины, с известной вероятностью попадания в них прогнозных значений. Также необходимо отметить, что выбор в пользу конформных предикторов сделан поскольку они предоставляют теоретические гарантии точности прогнозов, дают индивидуальную оценку неопределенности для каждого прогноза, могут применяться с различными типами моделей и дают интерпретируемые результаты. В отличие от квантильной регрессии конформные предикторы предоставляют оценку доверительных интервалов для индивидуальных прогнозов, в то время как

квантильная регрессия фокусируется на оценке квантилей распределения целевой переменной. Метод конформного прогнозирования была разработан Владимиром Вовком, Александром Гаммерманом, Крейгом Сондерсом и Владимиром Вапником в 1996-1999 годах, сначала с использованием e-values [1], а затем с p-values [2,3,4]. В течение десятилетий Вовк и его коллеги разрабатывали теорию и приложения конформного прогнозирования. Ультимативный и постоянно обновляемый обзор научных публикаций, туториалов, курсов и практических инструментов по конформному прогнозированию может быть найден в репозитории AwsomeConformalPrediction [6].

Целью данной работы является анализ и разработка метода оценки уверенности моделей прогнозирования выручки на уровне отраслей и отдельных организаций.

Представлена модель прогнозирования экономических показателей предприятий с использованием конформных предикторов для оценки уверенности прогноза. Конформные предикторы обеспечивает возможность получения предсказательных интервалов, которые учитывают неопределенность данных и модели, что особенно важно в экономических прогнозах, где точность и надежность предсказаний критически важны для принятия управленческих решений. Актуальность исследования обусловлена необходимостью оценки уверенности регрессионных моделей, применяемых для прогнозирования экономических показателей предприятий. Новизна работы заключается в использовании исторических данных по экономическим показателям 16 тысяч предприятий за последние 10 лет. Результаты исследования показывают, что средняя уверенность модели прогнозирования выручки предприятий не зависит от оборотов компаний, но средний разброс (дисперсия) уверенности модели имеет статистически значимое превышение для компаний малого и среднего бизнеса. Последнее может использоваться для определения компаний с сомнительной отчетностью. Также выявлено, что прогнозирование совокупной выручки группы предприятий, относящихся к одному и тому же типу деятельности (ОКВЭД), может повысить уверенность моделей. На примере прогноза энергетической отрасли показано, что величина ошибки прогноза отрасли может снизиться на 10% относительно средней ошибки по отдельным предприятиям отрасли.

1. Методика исследования

Объектом исследования была оценка уверенности моделей прогнозирования экономических показателей предприятия. Прогнозирование выручки компаний осуществлялось с использованием современной регрессионной модели (алгоритм регрессии с градиентным бустингом) и конформных предикторов. Цели исследования:

1) Изучение применимости интервалов прогноза (predictioninterval) как меры уверенности модели экономических показателей предприятия, в частности проверка реализации теоретических гарантий конформных предикторов как подхода оценки уверенности (uncertaintyestimation) и поиск неочевидных свойств уверенность прогнозов.

2) Проверка наличия более и менее прогнозируемых компаний, изучение взаимосвязи надежности прогнозов с масштабами оборотов компаний

3) Анализ среднего уровня прогнозируемости компаний отрасли с прогнозируемость совокупной выручки по отрасли (по ОКВЭДу)

4) Свободный поиск неожиданных результатов

1.1. Последовательность обработки данных

Пайплайн обработки данных состоял из следующих шагов:

- Очистка выборки от компаний с пропусками и резкими скачками в выручке (более 300%)
- Разделение данных на обучающую и отложенную выборку (тест)
- Обучение регрессионной модели на основе градиентного бустинга для прогнозирования выручки
- Калибровка конформных предикторов как надстройки над регрессионной моделью для оценки уверенности прогнозов
- Получение конформных интервалов прогноза для тестовой выборки
- Изучение полученных оценок уверенности регрессионной модели

Программное обеспечение: для анализа данных был использован Python с библиотеками Pandas, NumPy, Scikit-learn, CatBoost и ConformalTights [10].

1.2. Регрессионная модель

Для построения регрессионной модели выручки компаний использовался регрессор на основе градиентного бустинга.

1. Инициализация модели простой функцией
2. Вычисляется градиент ошибки м/д текущей моделью и истинными значениями
3. На градиенте ошибки обучается новая модели, которая будет корректировать ошибку текущей модели.
4. Новая модель объединяется со старой, используя веса, которые определяются градиентом ошибки.
5. Шаги 2-4 повторяются до тех пор, пока не будет достигнуто желаемое качество

При этом, градиентный бустинг делит признаковое пространство на фрагменты, используя деревья решений (decisiontrees). Каждое дерево решений представляет собой простую функцию, которая делит признаковое пространство на два фрагмента. Для каждого фрагмента, дерево решений строит собственную аппроксимирующую функцию, которая представляет собой простую линейную или нелинейную функцию. Подробное описание работы алгоритма градиентного бустинга может быть найдено, например, в [6].

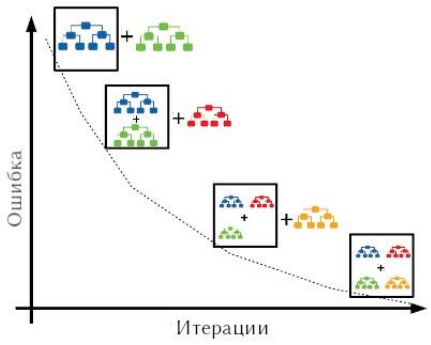

Рис. 1. Качественная картина поведения ошибки при использовании градиентного бустинга при добавлении новых слоев модели

1.3. Конформные предсказания

В этом разделе очень кратко изложена методология конформного прогнозирования для регрессии. Более подробно случай регрессии рассмотрен, например, в [7].

Беря за основу независимые и одинаково распределенные (i.i.d.) данные регрессии $Z_i = (X_i, Y_i)_{i=1}^{n}$, взятые из распределения P, где каждый Z_i состоит из результата Y_i и d-мерного входного вектора $X_i=(X_i(1),\ldots,X_i(d))$, мы стремимся предсказать результат Y_{n+1} для нового входного вектора X_{n+1}.
Конечная цель — построить интервал прогнозирования $C\subset\mathbb{R}^d\times\mathbb{R}$, то есть:

$$P\big(Y_{n+1} \in C(X_{n+1})\big) \geq 1-\alpha \ , \tag{1}$$

где α — заданный уровень ошибки, то есть доля истинных значений, не попавших в интервал предсказания. Вероятность $P(Y_{n+1}\in C(X_{n+1}))$ вычисляется на $n+1$ независимо распределенных выборках $Z_1,\ldots,Z_n,Z_{n+1}\sim P$. Для наблюдения $x\in\mathbb{R}^d$, $C(x)$ представляет собой множество возможных ответов $y\in\mathbb{R}$ таких, что $(x,y)\in C$. Диапазон предсказания должен иметь конечно-выборочную (неасимптотическую) достоверность без предположений о P.

В конформном прогнозировании для построения интервала предсказания для Y_{n+1}используется следующий подход. Мы рассматриваем Y_{n+1} при новом результате X_{n+1}, причем (X_{n+1},Y_{n+1}) является независимой выборкой, взятой из P. Далее, учитывая предыдущее описание, строим интервал предсказания:

$$C_{pred}(X_{n+1}) = [\mu(X_{n+1}) - F_n^{-1}(1-\alpha), \mu(X_{n+1}) + F_n^{-1}(1-\alpha)], \tag{2}$$

где μ — оценка функции регрессии, а F_n — эмпирическое распределение разницы между наблюдаемыми значениями результата и прогнозируемым (т.е. распределение ошибки прогноза) $|Y_i-\mu(X_i)|$ для $i=1,\ldots,n$. Член$F_n^{-1}(1-\alpha)$ представляет собой $(1-\alpha)$-квантиль F_n. В случае большой выборки интервал приблизительно действителен, если значение функции оценки μнадежно. В частности, это означает, что оцененный $(1-\alpha)$-квантиль $F_n^{-1}(1-\alpha)$подобранного распределения

остатков должен быть близок к $(1-\alpha)$-квантилю невязок популяции $|Y_i-\mu(X_i)|$ для $i=1,\ldots,n$. Гарантия этого уровня точности для μ обычно требует надлежащих условий регулярности для базового распределения данных P и для самого μ. Эти условия включают наличие правильно указанной модели и выбор соответствующих параметров настройки модели. Ниже приведен псевдоалгоритм построения конформного интервала прогноза.

Алгоритм Конформное предсказание.

Вход:
Набор данных (X_i,Y_i) , $i=1,\ldots,n$
Обученная регрессионная модель μ
Уровень непокрытия/неконформности (допустимый уровень непопадания прогнозов в интервал предсказания) $\alpha\in(0,1)$
Выход: полоса предсказания, по $x\in\mathbb{R}^d$
Шаг 1: Случайным образом разделить $\{1,\ldots,n\}$
на подмножества одинакового размера I_1, I_2
Шаг 2: Обучить $\mu_{I1}=\mu(\{(X_i,Y_i):i\in I_1\})$
Шаг 3: Вычислить функцию оценки (например, невязку) $R_i=|Y_i-\hat{\mu}_{I1}(X_i)|:i\in I_2$
Шаг 4: Сортировка $\{R_i:i\in I_2\}$ в порядке возрастания $R_{(1)}\leq\cdots\leq R_{(n/2)}$
Шаг 5: Вычисление $d=R_{(k)}$, т.е. k-го наименьшего значения в $\{R_i:i\in I_2\}$, где $k=\lceil(1-\alpha)(n/2+1)\rceil$
Результат: $C_{pred}(x)=[\mu_{I1}(x)-d,\mu_{I1}(x)+d]$, для всех $x\in\mathbb{R}^d$

1.4. Исходные данные

В качестве исходных данных по выручке организаций используется информация из открытых источников по финансовой (бухгалтерской) отчётности организации. Комплект финансовой отчетности включает несколько документов, из которых для финансового моделирования используется форма №1 (баланс) и форма №2 (отчет о финансовых результатах).

В качестве источника использованы массивы открытых данных Росстата [8] и сервиса финансовой (бухгалтерской) отчетности Федеральной налоговой службы Российской Федерации (БФО ФНС) [9], по которым были известны объемы выручки компаний с 2012 по 2020 годы, ИНН, название, ОКВЭД. Начальная выборка содержала 16831 компанию, из анализа были удалены компании с пропусками в данных и скачками в выручке более чем в 3 раза от года к году, итоговый объем анализируемой выборки составил 10517 компаний.

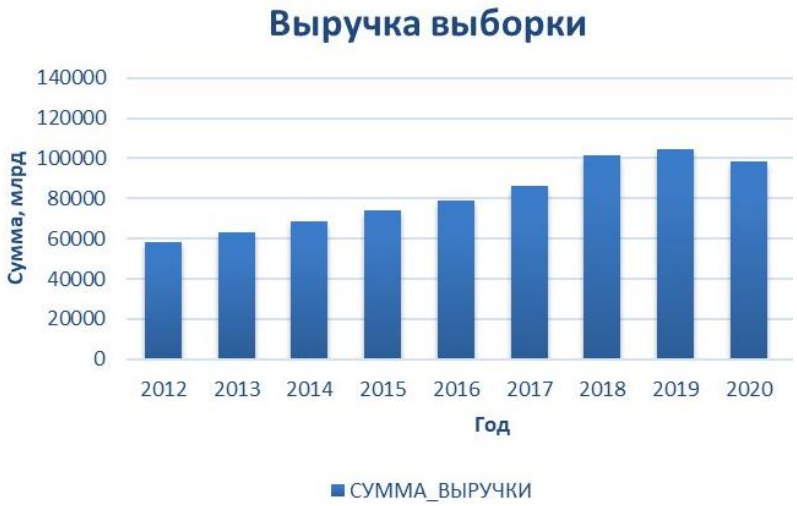

Рис. 2. Динамика совокупной выручки исходной выборки компаний

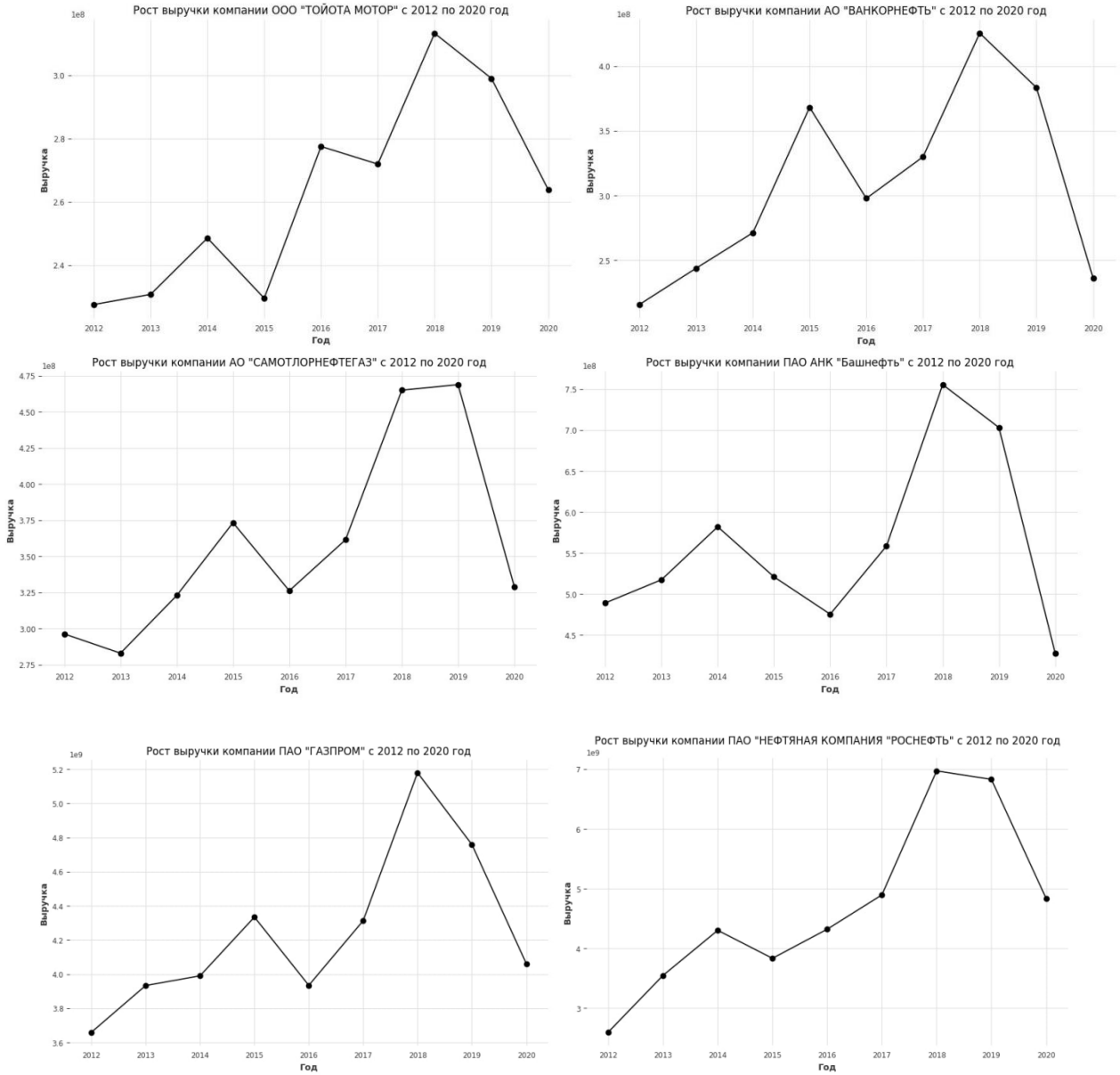

Рис. 3. Пример компаний с немонотонной динамикой выручки за период 2012-2020

При том, что совокупная сумма выручки компаний в выборке показывала в целом плавный рост (Рис.2), отдельные юрлица могли показывать более сложную динамику. На Рис.3 приведено несколько примеров компаний с немонотонным ростом, на Рис.4 – с монотонным. Факторы, влияющие на прогнозируемость компаний, могут быть индивидуальными для предприятий (условия поставок, износ оборудования), общими для сегмента экономики (логистика, спрос на продукцию, тарифы, санкции, сезонность), общими для всей экономики (инфляция, курсовая разница) [11]. Но само по себе наличие совокупности (кластеров) компаний с более или менее схожей динамикой выручки делает предпочтительным выбор регрессионной модели, которая могла бы учитывать такую кластерную структуру. Это стало причиной выбора в качестве регрессора модели на градиентном бустинге, т.к. отличительной чертой этого семейства алгоритмов является адаптация аппроксимирующей функции под каждый кластер однотипных объектов внутри признакового пространства. Другими словами, выпуклая оболочка выборки бьется с помощью бустинга на выпуклыекомпакты, в которых строится своя регрессионная модель. Это позволило сконцентрировать фокус исследования на изучении свойств уверенности модели, т.к. функцию оценки целевой переменной (выручки) можно считать равномерно оптимизированной на всем протяжении векторного пространства.

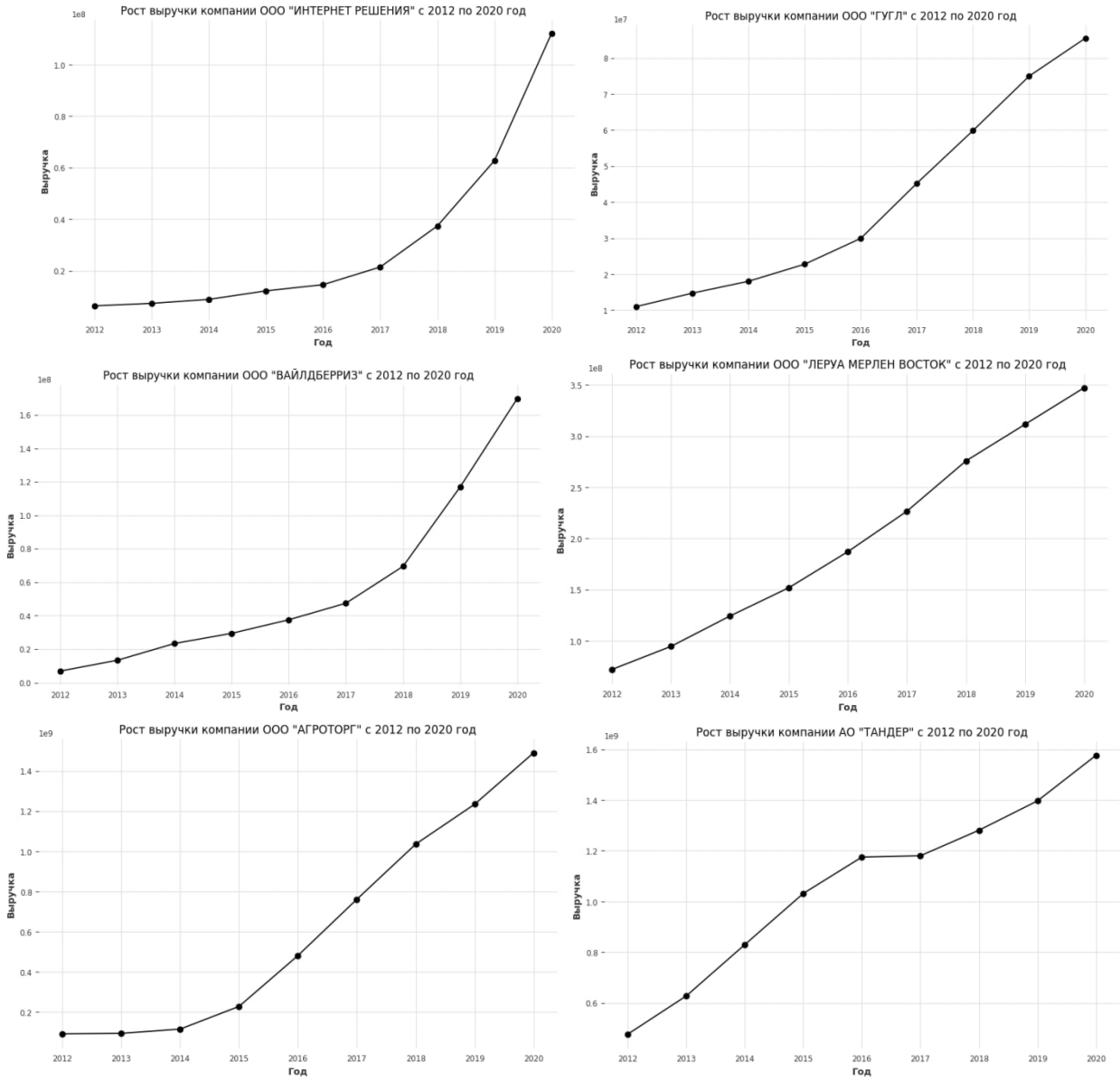

Рис. 4. Пример компаний с монотонной динамикой выручки за период 2012-2020

2. Эксперимент

Обучающая выборка содержала данные по выручке с 2012 по 2020 год. Входной вектор $\vec{x}$для выбранной компании содержал данные до 2019 года и был получен конкатенацией вектора отношений выручки от года к предыдущему и вектора отношений выручки по каждому году к начальному году выручки (для передачи в модель информации о динамике выручки):

$$\vec{\mathbf{x}} = \left[\frac{r_{2013}}{r_{2012}}, \frac{r_{2014}}{r_{2013}}, \ldots, \frac{r_{2019}}{r_{2018}}\right] \cup \left[\frac{r_{2014}}{r_{2012}}, \frac{r_{2015}}{r_{2012}}, \ldots, \frac{r_{2019}}{r_{2012}}\right]$$

Прогнозируемая переменная $\mathbf{Y}$ – выручка компании в 2020 году, приведенная к выручке 2019 т.е. $\mathbf{Y} = \frac{\mathrm{r}_{2020}}{\mathrm{r}_{2019}}$, μ— оценка функции регрессии:

$$\mu(\vec{\mathrm{x}}) = \mathrm{CatBoostRegressor}(\vec{\mathrm{x}}) = \mathrm{Y}$$

Конечная цель:

$$P\big(Y \in C(X)\big) \geq 1 - \alpha,$$

то есть вероятность непопадания прогнозовY в интервалы$C(X)$ не должна превышать наперед заданный уровень толерантности к ошибкам α.

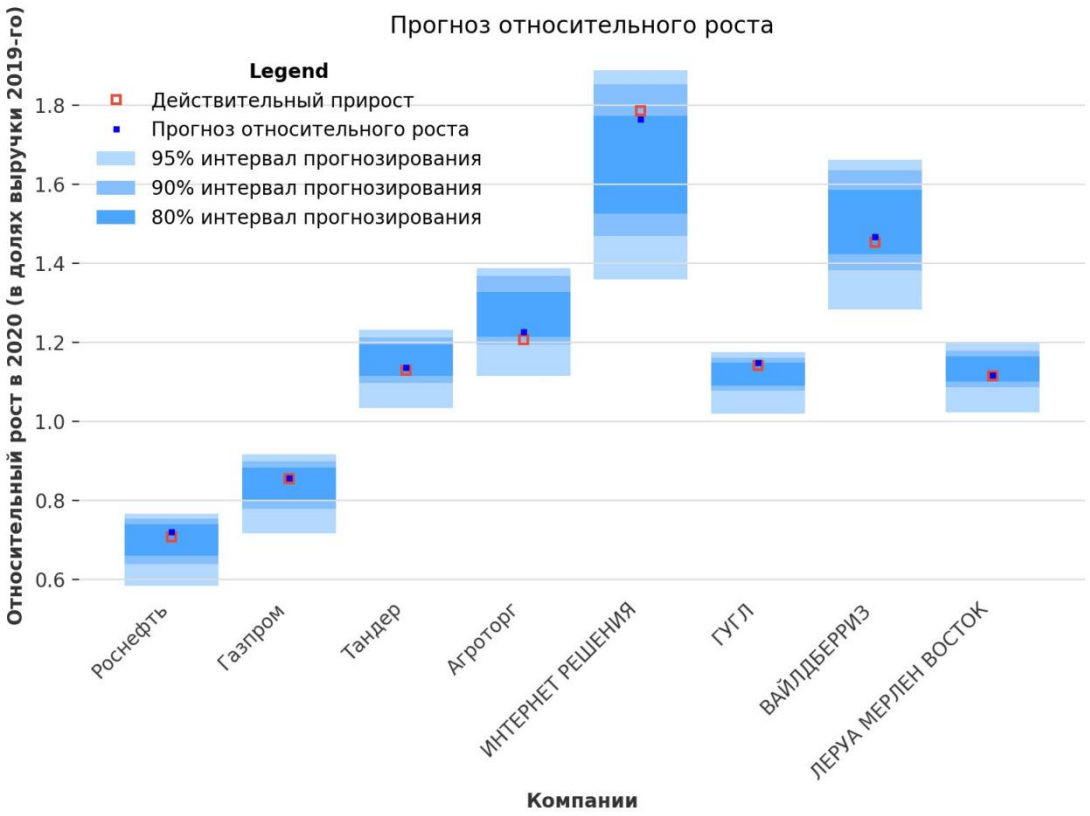

Рис. 5. Пример интервалов прогноза для некоторых компаний

3. Результаты

Первая проверявшаяся гипотеза, о соблюдении теоретической гарантии конформного прогнозирования выполнялась через прямой подсчет пропорции (не)попадания оценки функции регрессии (прогноза) в соответствующий интервал.

Таблица 1. Проверка теоретической гарантии конформного прогнозирования

Заданный порог ошибок (доля)	**Реальная доля ошибок**
0.05	0.05
0.1	0.11
0.2	0.23

Вторая проверявшаяся гипотеза касалась наличия более и менее надежно прогнозируемых компаний. Наличие выбросов и неоднородности в размерах интервалов прогноза должно приводить к высокой дисперсии интервалов прогноза.

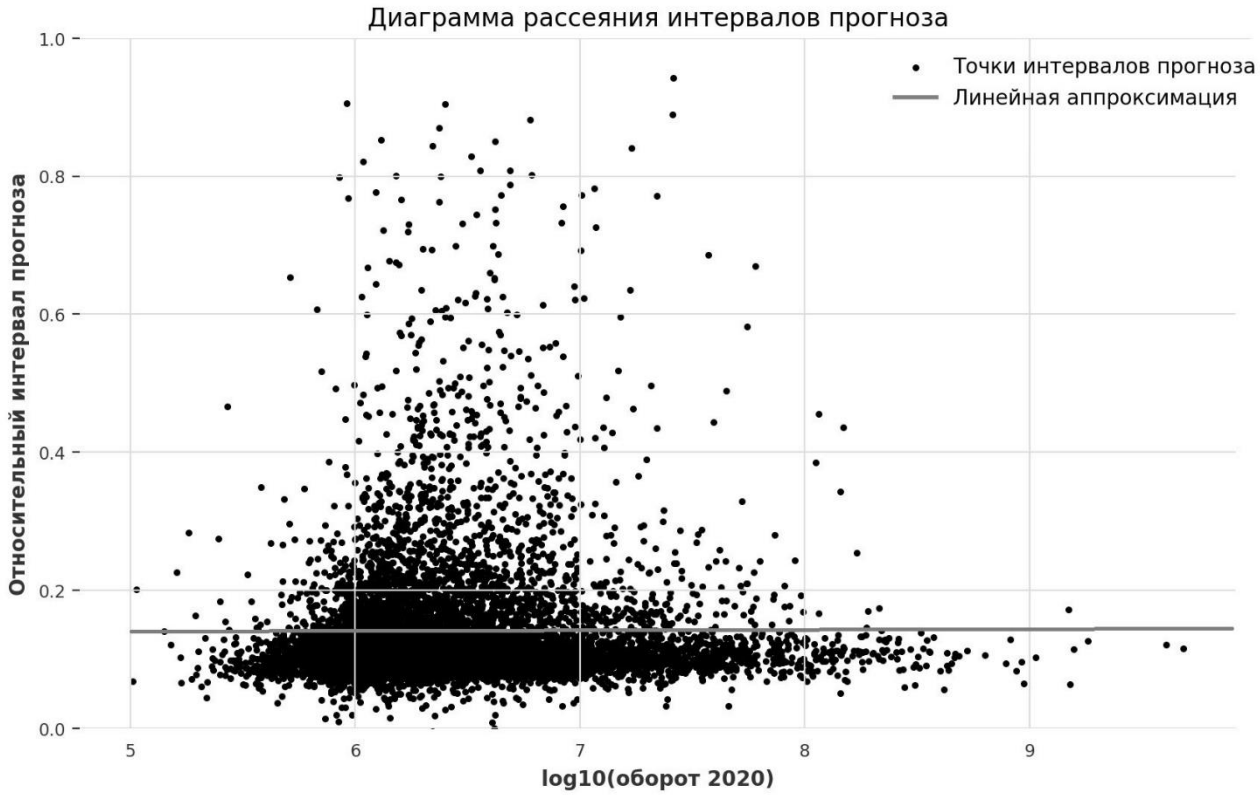

Рис. 6. Диаграмма рассеяния для интервалов прогноза по логарифму выручки компаний. На шкале абсцисс log_{10}от выручки в тысячах рублей

Для проверки третьей гипотезы Анализ диаграммы рассеяния (Рис.6) относительных интервалов прогноза показал, что зависимость длины интервала прогноза имеет очень слабый тренд, коэффициент полинома 1-й степени k=0.008. При этом дисперсия длины интервала прогноза показала нелинейное поведение, с наличием выбросов.

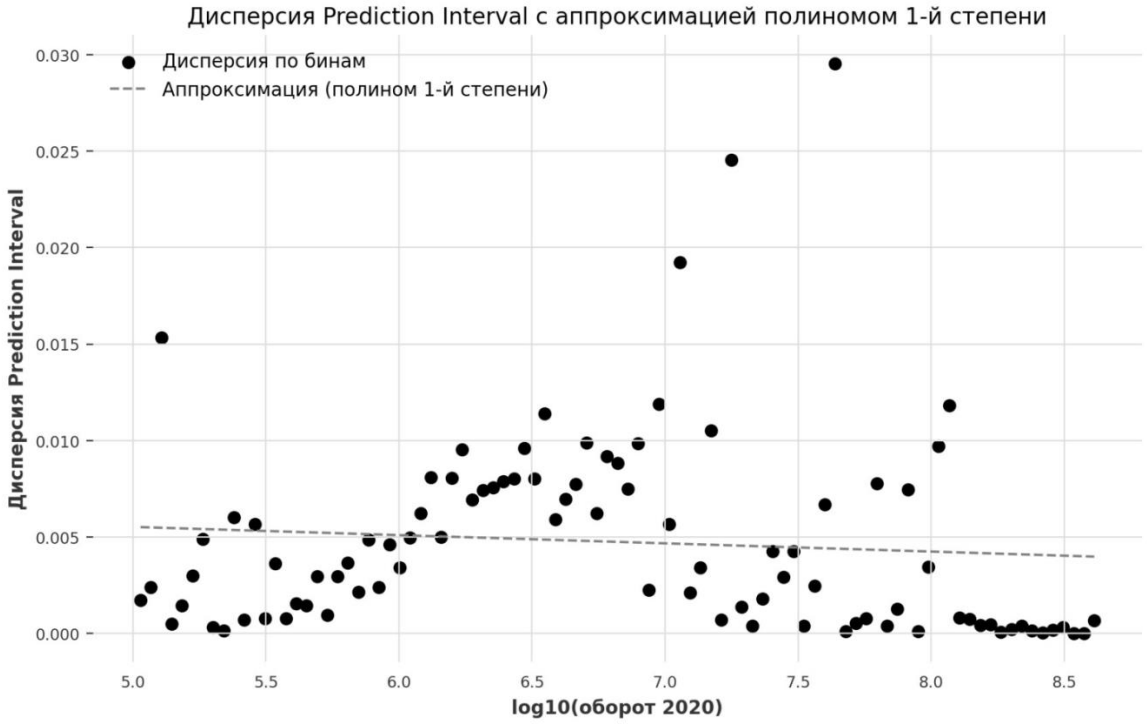

Рис. 7. Диаграмма рассеяния для дисперсии интервалов прогноза по логарифму выручки компаний (log_{10}от выручки в тысячах рублей). Для визуальной интерпретируемости шкала дохода разбита на 60 равных интервалов (бинов) и дисперсия усреднена в каждом из 60 бинов

При этом дисперсия длины интервала прогноза показала нелинейное поведение, с наличием выбросов (Рис.7). Заметен всплеск средней дисперсии для компаний с доходами от 1млрд до 10млрд в год (промежуток $6 < \log_{10} X < 7$). Такой разброс в уверенности прогноза указывает на то, что даже продвинутой модели регрессора сложно настраиваться на границу среднего и крупного бизнеса.

Также была построена отдельная модель на основе данных отдельно по энергетической отрасли (ОКВЭД 35). Размер среднего интервала прогноза, построенного по отельным предприятиям (с усреднением по предприятиям) составила 0.238, в то время как интервал прогноза для суммы всей выручки по отрасли оказался равен 0.185.

4. Заключение

Общие выводы по исследованию:

1) Конформные предикторы могут достигать заранее обозначенного исследователем уровня ошибок (проверка теоретической гарантии).
2) Оценка уверенности зависит от представления данных об экономических показателях.
3) Средняя оценка уверенности слабо зависит от оборота.
4) Для предприятий с оборотом 1-10 млрд/год отмечен повышенный уровень выбросов, можно интерпретировать как отсутствие единого распределения в данных (экономические причины, некорректная отчетность).
5) На примере энергетической отрасли – прогнозирование совокупной выручки может приводить к повышению уровня уверенности модели.

Дальнейшие исследования могут предполагать учет большего количества параметров моделируемых эффектов, сравнение уверенности от года к году, возможность калибровки конформного предиктора в расширенном пространстве относительно пространства регрессионной модели. Также требует проверки гипотеза о природе больших выбросов в уверенности модели как относящихся к некорректной отчетности компаний

Литература

1. *Gammerman* A., *Vovk V. and Vapnik V.* Learning by transduction, Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, 1998. vol. 14, pp. 148–155.
2. *Vovk V., Gammerman A. and Saunders C.* Machine-learning applications of algorithmic randomness, in International Conference on Machine Learning, 1999, pp. 444–453.
3. *Saunders C., Gammerman A. and Vovk V.* Transduction with confidence and credibility, 1999.
4. *Vovk V., Gammerman A. and Shafer G.* (2005), Algorithmic Learning in a Random World, Springer.
5. *Lei J.; G'Sell M.; Rinaldo A.; Tibshirani R.J.; Wasserman L.* Distribution-free predictive inference for regression. J. Am. Stat. Assoc., 2018. 113, 1094–1111.
6. A professionally curated list of awesome Conformal Prediction videos, tutorials, books, papers, PhD and MSc theses, articles and open-source libraries. URL: https://github.com/valeman/awesome-conformal-prediction.
7. *Ziegel E.R.* The elements of statistical learning. Technometrics. 2003, 45, 267–268.
8. Open data. Federal State Statistics Service. URL: https://rosstat.gov.ru/opendata/ (Accessed 20.03.2023).
9. State information resource of accounting (financial) statements. URL: https://bo.nalog.ru/ (Accessed 20.03.2023).
10. ConformalTightsPythonlibrary. URL: https://github.com/superlinear-ai/conformal-tights.
11. *Дранко О.И., Васильев М.В.* Двухуровневая модель прогнозирования доходов крупномасштабной энергетической системы // Датчики и системы. 2023. № 2 (267). С. 71-78.