

Домашняя работа, Машинное обучение

Лучинкина Ксения, Щудро Виктория

Март 2025

1 Обоснование темы

1.1 Придумайте непрерывную зависимую (целевую) переменную (например, за- работная плата или прибыль) и бинарную переменную воздействия (напри- мер, образование или факт занятий спортом).

- Целевая переменная - налоговые поступления правительства (Taxes)
- Переменная воздействия - уровень налоговой ставки (Rate) (1 - высокая, 0 - низкая).

1.2 Опишите, для чего может быть полезно изучение влияния переменной воздействия на зависимую переменную. В частности, укажите, как эта информация может быть использована бизнесом или государственными органами.

Для начала рассмотрим возможности применения влияния переменной воздействия на зависимую переменную для государственных органов:

Оптимизация налоговой политики - анализ налоговых поступлений в зависимости от уровня налоговой ставки может помочь определить, при каком уровне налоговой ставки поступления в бюджет максимизируются. Это важно для создания сбалансированной фискальной политики, предотвращающей как чрезмерное налогообложение, так и недобор налогов.

Прогнозирование последствий налоговых реформ - введение прогрессивной шкалы налогообложения (Progressive, контрольная переменная) или изменение ставок также могут повлиять на налоговые поступления. Оценка этих эффектов позволит минимизировать финансовые риски. Если высокий уровень неравенства (Inequality = 1, инструментальная переменная) усиливает негативные последствия высоких налогов, правительство может разрабатывать программы компенсации для уязвимых слоев населения.

Применение для бизнеса:

Прогнозирование налоговой нагрузки - компании могут учитывать изменения в налоговой политике для долгосрочного планирования финансовых потоков, ценообразования и инвестиционных решений. Например, при высокой налоговой ставке (Rate = 1) бизнес может пересматривать структуру расходов, увеличивать долю налоговых вычетов (используя на практике так называемые накопленные налоговые убытки, NOL) или менять стратегию найма. Выбор стратегии инвестирования - инвесторы учитывают налоговую среду при принятии решений о вложениях. Высокие налоги могут сделать страну менее привлекательной для бизнеса, тогда как умеренные налоговые ставки способствуют притоку капитала.

1.3 Обоснуйте наличие причинно-следственной связи между зависимой переменной переменной воздействия. Приведите не менее 2-х источников из научной литературы, подтверждающих ваши предположения.

Установление причинно-следственной связи между уровнем налоговой ставки (переменная воздействия) и налоговыми поступлениями правительства (зависимая переменная) является фундаментальным вопросом в экономической теории и практике. Согласно концепции кривой Лаффера,

предложенной американским экономистом Артуром Лаффером, существует нелинейная зависимость между налоговой ставкой и объемом налоговых поступлений (кривая иллюстрирует, что как чрезмерно низкие, так и чрезмерно высокие налоговые ставки могут приводить к снижению общих налоговых поступлений).

Эмпирические исследования подтверждают эту зависимость. Например, в работе *"Зависимость налоговых поступлений от налоговой нагрузки на основе закона Артура Лаффера"* анализируется практическое применение кривой Лаффера и подчёркивается, что оптимальный уровень налогообложения способствует максимизации налоговых поступлений, тогда как отклонения от этого уровня в любую сторону могут привести к их снижению.

Кроме того, в исследовании *"Эластичность кривой Лаффера и оптимизация налогов при оценке теневого сектора экономики"* рассматривается влияние гибкости фискальной политики на максимальные доходы государства.

Таким образом, научная литература подтверждает наличие причинно-следственной связи между уровнем налоговой ставки и объемом налоговых поступлений, подчеркивая в том числе и практическую важность выбора оптимального уровня налогообложения для обеспечения максимальных доходов бюджета.

1.4 Кратко опишите результаты предшествовавших исследований по схожей тематике и критически оцените методологию этих работ с точки зрения гибкости (жесткости предпосылок) использовавшихся методов эконометрического анализа. Объясните, в чем заключается преимущество и недостатки применяемых вами методов в сравнении с теми, что ранее использовались в литературе.

На основе найденной нами информации из доступных источников, были выбраны три ключевых исследования для детального анализа, а также еще несколько исследований, которые, как нам кажется, должны быть упомянуты в кратком формате. Основные статистические исследования на тему воздействия ставки налога на величину налоговых сборов:

- Исследование Канто, Джойнса и Лаффера (1981)
- Работа Гулсби, Холла и Каца (1999)
- Исследование Трабандта и Улига (2009)

Исследование Канто, Джойнса и Лаффера (1981)

Основные результаты:

Авторы разработали теоретическую модель, демонстрирующую существование оптимальной налоговой ставки, при которой государственные доходы максимизируются. Они использовали модель общего равновесия для оценки влияния налоговых изменений на доходы, применяя анализ временных рядов. Результаты показали, что снижение налоговых ставок может привести к увеличению налоговых поступлений, если исходные ставки превышают оптимальный уровень. Говоря о критической оценке модели, можно упомянуть следующие преимущества: во-первых,

использование модели общего равновесия позволяет учитывать взаимодействие различных экономических факторов и их влияние на налоговые поступления; во вторых, анализ временных рядов предоставляет эмпирическую основу для оценки динамических эффектов налоговых изменений. Однако стоит также отметить, что модель основана на ряде упрощающих предположений, таких как однородность налогоплательщиков и отсутствие внешних шоков, что может ограничивать применимость результатов в реальных условиях.

Работа Гулсби, Холла и Каца (1999)

Основные результаты:

Авторы исследовали поведенческие реакции налогоплательщиков на изменения налоговых ставок, используя метод разности разностей (метод разности разностей сравнивает изменения во времени между двумя группами: группа воздействия – подвергается изменению, контрольная группа – остается неизменной. DiD предполагает, что без вмешательства разница между этими группами оставалась бы неизменной, что называется допущением параллельных трендов). Авторы использовали данные за шесть десятилетий налоговых реформ в США, начиная с 1920-х годов, чтобы исследовать реакцию высокодоходных налогоплательщиков на изменения в предельных налоговых ставках. Основной акцент был сделан на анализе изменений в поведении налогоплательщиков в ответ на значительные налоговые реформы, такие как снижение максимальной предельной ставки с 91 до 70 процентов в 1964 году и дальнейшее снижение до 50 процентов в 1982 году. Для оценки эластичности налогооблагаемого дохода по отношению к предельной налоговой ставке авторы применили эконометрические модели, учитывающие возможные искажения, вызванные изменениями в налоговой политике. Они также рассмотрели временные лаги в реакции налогоплательщиков на изменения налоговых ставок, чтобы учесть возможные задержки в адаптации поведения. Исследование внесло немалый вклад в скептицизм относительно общей теории Лаффера, поскольку авторы пришли к выводу о том, что реакция налогоплательщиков на изменения налоговых ставок была незначительной, что ставит под сомнение практическую применимость кривой Лаффера в некоторых контекстах. Говоря о критической оценке методологии, можно отметить следующие преимущества: метод разности разностей позволяет контролировать влияние неизмеренных факторов, которые остаются постоянными во времени, улучшая точность оценок; использование микроуровневых данных обеспечивает детальное понимание индивидуальных реакций на налоговые изменения. Однако метод разности разностей предполагает, что в отсутствие изменения налоговых ставок разницы между группами остались бы постоянными, что может не соответствовать действительности.

Исследование Трабандта и Улига (2009)

Основные результаты:

Авторы оценили кривую Лаффера для США и стран ЕС-14, используя неоклассическую модель (стандартную неоклассическую модель экономического роста, включающую домохозяйства, фирмы и правительство). Они рассчитали, что оптимальные налоговые ставки (в модели введены различные виды налогов: на потребление, труд и капитал) различаются между регионами, и подчеркнули, что повышение налоговых ставок до уровня, максимизирующего доходы, может негативно сказаться на благосостоянии населения. Для приведения модели в соответствие с реальными экономическими условиями авторы откалибровали ее, используя эмпирические данные за 1995 и 2007 годы. Авторы произвели настройку параметров, таких как коэффициенты

производственной функции, показатели предпочтений домохозяйств и ставки налогообложения. Смоделировав экономику с различными уровнями налоговых ставок, авторы построили кривые Лаффера для каждого типа налога. Это позволило определить точки, при которых налоговые поступления достигают максимума. Оптимальные налоговые ставки по итогам исследования получились следующими: налоги на труд - оптимальная ставка составляет около 30 процентов для США и 40 процентов для стран ЕС-14. Налоги на капитал - оптимальная ставка составляет около 40 процентов для США и 35 процентов для стран ЕС-14. Касаемо критической оценки методологии, можно отметить следующее: Применение неоклассической модели роста позволяет учитывать долгосрочные эффекты налоговой политики на экономический рост и благосостояние, а также - международное сравнение предоставляет ценную информацию о различиях в налоговых системах и их влиянии на экономику. Однако у модели есть и недостатки - она может не учитывать институциональные и структурные различия между странами, что ограничивает точность сравнений.

Для остальных исследований, представлена следующая таблица с кратким анализом:

max width=				
Автор(ы)	Год	Предмет исследования	Результаты	Использованные переменные
Хсинг	1996	Оценка кривой Лаффера для США (1959–1991 гг.)	Оптимальная налоговая ставка составляет 32–35 процентов.	Налоговые ставки, налоговые поступления, макроэкономические показатели
Фейг и Макги	1983	Оценка кривой Лаффера для Швеции	Оптимальная налоговая ставка варьируется в пределах 54–62 процентов.	Предельные налоговые ставки, налоговые поступления
Равстейн и Вийлбриф	1988	Оценка кривой Лаффера для Нидерландов (1960–1985 гг.)	Оптимальная налоговая ставка в 1970 г. составила 66,9 процентов.	Налоговые ставки на прибыль, косвенные налоги, налоговые поступления
Хийжман и ван Офем	2005	Оценка кривой Лаффера для нескольких европейских стран	Оптимальная предельная ставка налога всегда ниже 36 процентов; повышение налогов увеличивает теневую экономику.	Налоговые ставки, уровень теневой экономики, макроэкономические показатели
Йоан	2012	Оценка совокупной кривой Лаффера для Румынии (1999–2009 гг.)	Увеличение налоговых ставок приводит к снижению налоговых поступлений.	Налоговые ставки, налоговые поступления, макроэкономические показатели

Таблица 1: Анализ исследований на рассматриваемую тему

1.5 Придумайте хотя бы 3 контрольные переменные, по крайней мере, одна из которых должна быть бинарной и хотя бы одна – непрерывной. Кратко обоснуйте выбор каждой из них.

- Deficit - дефицит государственного бюджета в млн долларов. Дефицит бюджета отражает уровень финансового состояния страны. Если государство испытывает большой дефицит, оно может стремиться повысить налоги для его сокращения. Однако высокая налоговая нагрузка может снижать экономическую активность, что, в свою очередь, может негативно сказаться на налоговых поступлениях.
- Progressive - наличие прогрессивной шкалы налогообложения (1 - есть, 0 - нет). Прогрессивная система налогообложения может изменять поведенческие реакции налогоплательщиков. Высокодоходные группы населения могут проявлять большую склонность к налоговой оптимизации или уклонению при увеличении ставок, что влияет на общие налоговые поступления. Наличие прогрессивного налогообложения также может влиять на структуру налоговой базы.
- Grey - уровень теневой экономики в структуре экономики в процентах. Чем выше уровень теневой экономики, тем сложнее собирать налоги, поскольку значительная часть экономической деятельности остается вне официальной налоговой системы.

1.6 Придумайте бинарную инструментальную переменную и обоснуйте, почему она удовлетворяет необходимым условиям.

- Инструментальная переменная - уровень неравенства в стране (Inequality) (1 - высокое, 0 - низкое).

Данная переменная удовлетворяет требованиям, так как обладает необходимой релевантностью (в странах с высокой политической стабильностью чаще сохраняются умеренные налоговые ставки, а в странах с нестабильностью могут происходить частые изменения налоговой политики, в том числе резкое повышение налогов) и экзогенностью (политическая стабильность сама по себе не определяет налоговые поступления, но влияет на налоговую политику, которая, в свою очередь, формирует объем поступлений).

1.7 В случае необходимости приведите дополнительные содержательные комментарии о целях, задачах, методологии и вкладе вашего исследования.

Ненаблюдаемые переменные, порождающие эндогенность:

- Culture - фискальная культура, склонность к уклонению от налогов.

2 Генерация и предварительная обработка данных

2.1 Опишите математически предполагаемый вами процесс генерации данных. Примечание: оценивается в том числе оригинальность предложенного вами процесса, поэтому, в частности, не рекомендуется использовать совсем простые линейные модели.

Процесс генерации данных подробно описан при генерации каждой переменной в заданиях ниже.

2.2 Обоснуйте предполагаемые направления связей зависимой переменной и переменной воздействия с контрольными переменными.

Визуализируем предполагаемые связи

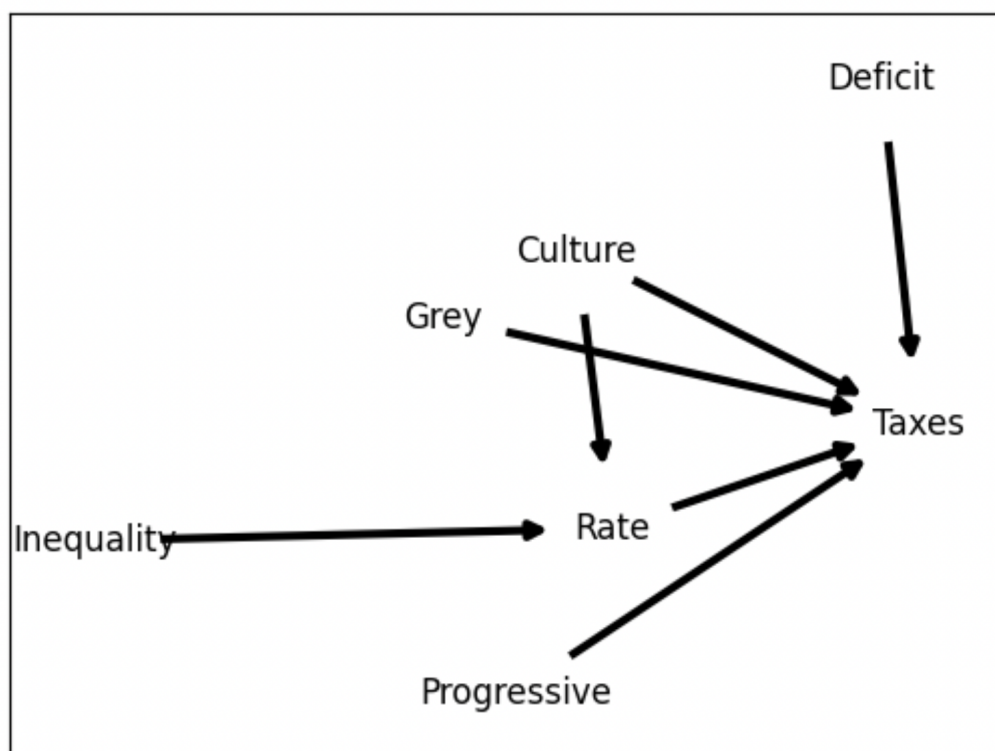


Рис. 1: Граф связей

- Налоговые поступления формируются под воздействием налоговой ставки, но также зависят от состояния экономики, теневого сектора, структуры налоговой системы и уровня доверия граждан к государству.
- Уровень налоговой ставки частично определяется уровнем неравенства в стране, что позволяет использовать его как инструментальную переменную.
- Государственные финансы, налоговая система и доля теневой экономики являются ключевыми контрольными переменными, которые могут влиять на взаимосвязь между налоговой ставкой и поступлениями.

- Фискальная культура остается ненаблюдаемой переменной, которая может вызывать эндогенность и затруднять точное измерение влияния налоговых ставок на налоговые поступления.

2.3 Симулируйте данные в соответствии с предполагаемым вами процессом и приведите корреляционную матрицу, а также таблицу со следующими описательными статистиками. Для непрерывных переменных: выборочное среднее, выборочное стандартное отклонение, медиана, минимум и максимум. Для бинарных переменных: доля и количество единиц.

Генерация контрольных переменных

Распределение дефицита государственного бюджета будет распределено согласно эмпирическим наблюдениям с сайта ОЭСР

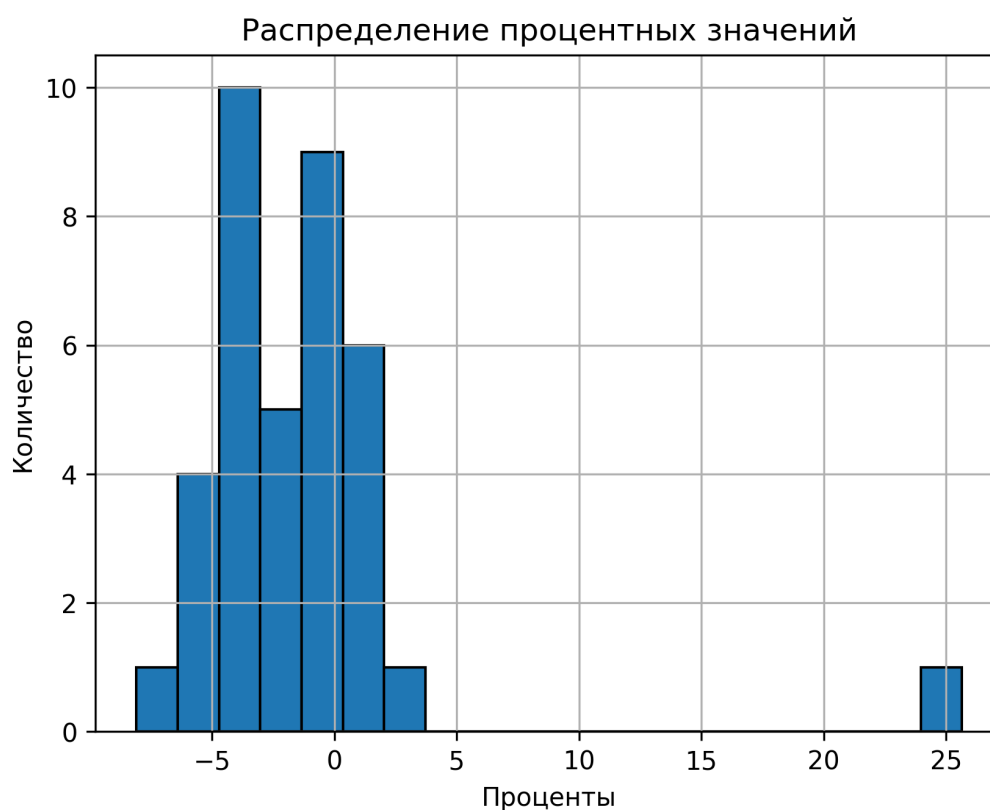


Рис. 2: Гистограмма дефицитов госбюджета

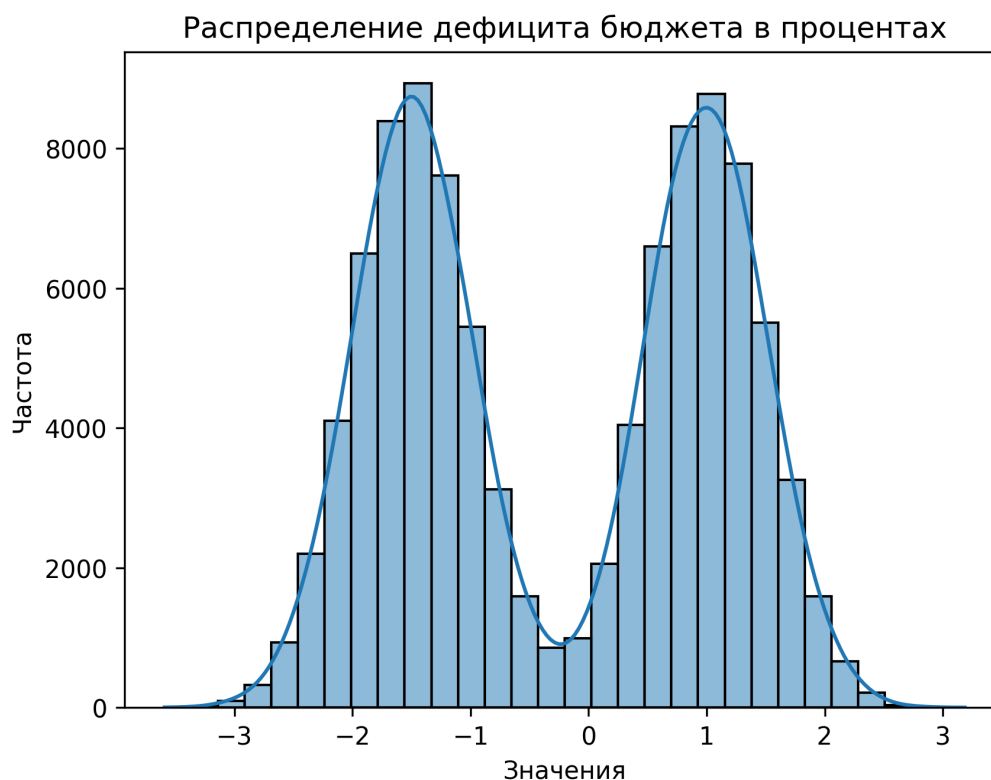


Рис. 3: Распределение дефицита бюджета в процентах

Генерация данных по наличию прогрессивной шкалы налогообложения будет происходить при помощи нормальной копулы для прогрессивной шкалы налогообложения и для уровня фискальной культуры

Генерация теневого уровня экономики. Будем считать, что в среднем в странах уровень теневой экономики составляет примерно 0,24 (24 процента от ВВП), со стандартным отклонением в 13 процентных пункта.

<https://cyberleninka.ru/article/n/statisticheskie-izmereniya-tenevoy-ekonomiki/viewer>

По данным из статьи видно, что данные хорошо ложатся на логнормальное распределение, поэтому по нему и будем моделировать данные в будущем.

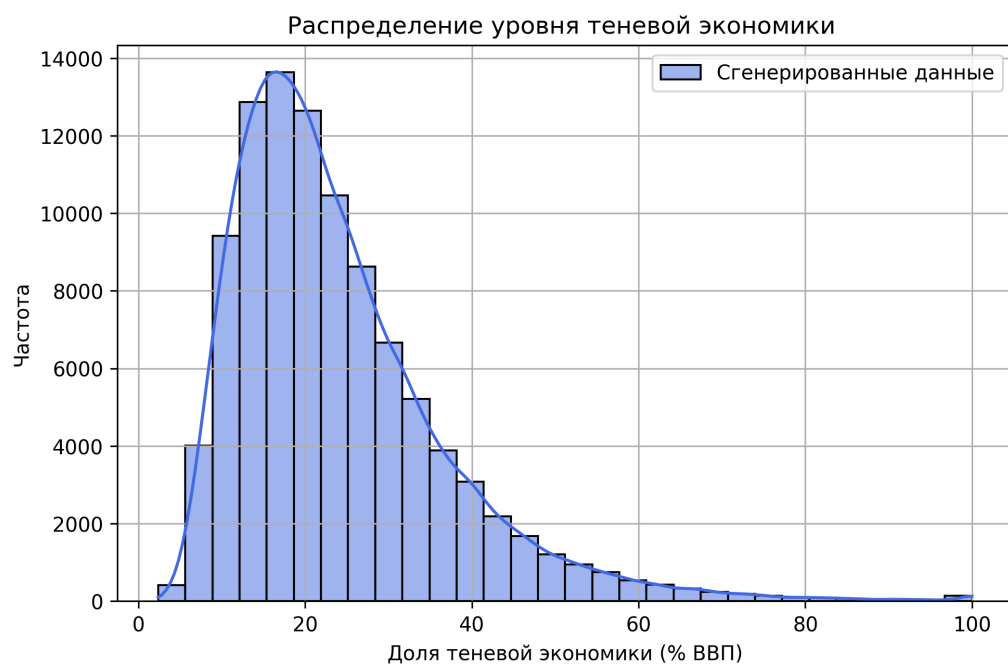


Рис. 4: Распределение уровня теневой экономики

Генерация ненаблюдаемой переменной

За ненаблюдаемую переменную мы взяли уровень фискальной культуры. Генерируем при помощи нормальной копулы, идею взяли тут <https://habr.com/ru/articles/145751/>

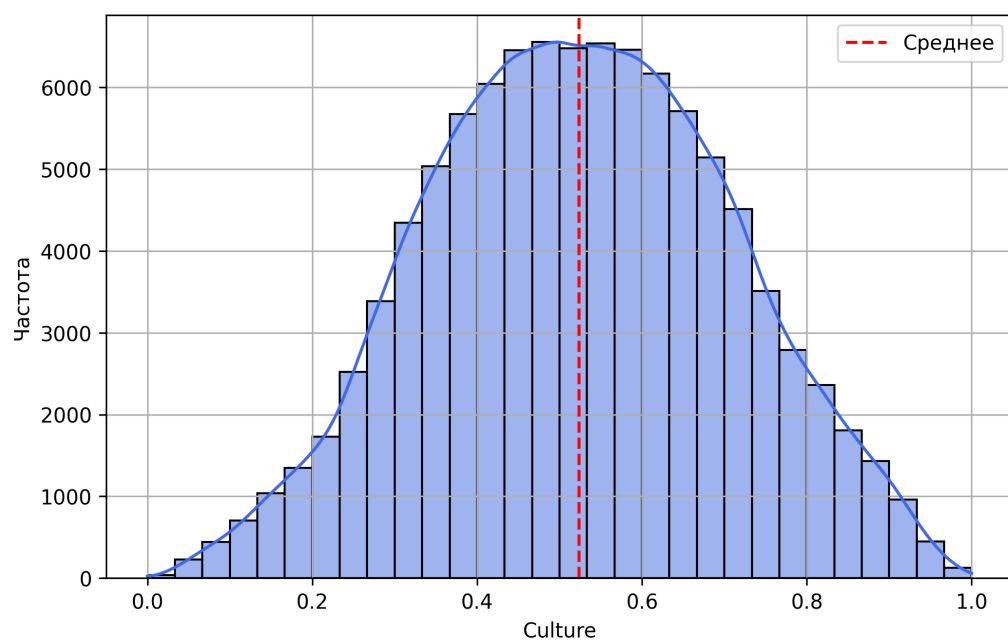


Рис. 5: Распределение уровня фискальной культуры

Генерация инструментальной переменной

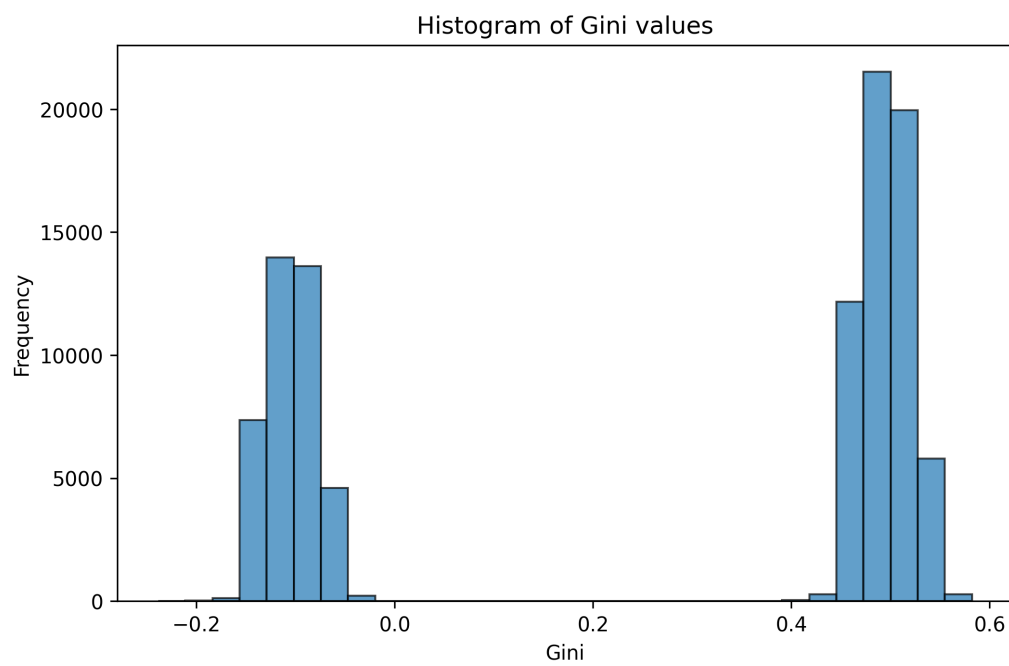


Рис. 6: Распределение уровня Индекса Джини

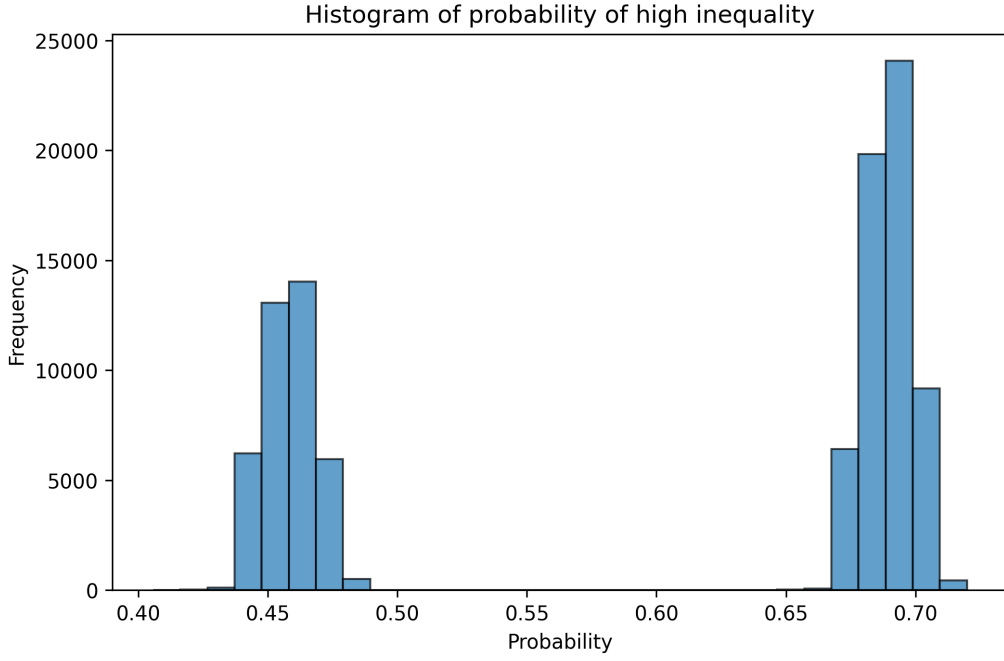


Рис. 7: Гистограмма вероятности высокого уровня неравенства

Это отражает две группы стран: развитые и развивающиеся.

Генерация переменной воздействия

Здесь необходимо рассмотреть два сценария: когда в стране низкое неравенство и когда в стране высокое неравенство.

Условные вероятности переменной воздействия (налоговой ставки) зависят от контрольных переменных, инструментальной переменной (уровня неравенства) и ненаблюдаемой переменной (фискальной культуры):

$$P(Rate_i = 1 | Deficit_i, Progressive_i, Grey_i, Culture_i, Inequality_i) =$$

$$F_{Logistic} \left(5 \times \ln(Culture_i + 1) + \sqrt{|Deficit_i|} + Progressive_i \times Grey_i - 20 + 4 \times Inequality_i \right) \quad (1)$$

где $F_{Logistic}$ — функция распределения стандартного логистического распределения.

Для краткости введем обозначение для условной вероятности высокой налоговой ставки при конкретном уровне неравенства:

$$P_k = P(Rate_i = 1 | Deficit_i, Progressive_i, Grey_i, Culture_i, Inequality_i = k), \quad k \in \{0, 1\} \quad (2)$$

Введем равномерную случайную величину $U_i \sim U(0, 1)$ и определим гипотетические переменные:

$$Rate_i^1 = I(P_1 \geq U_i) \quad (3)$$

$$Rate_i^0 = I(P_0 \geq U_i) \quad (4)$$

где

$$I() = \{1, 0, \quad (5)$$

Переменные $Rate_i^1$ и $Rate_i^0$ отражают потенциальный уровень налоговой ставки в зависимости от уровня неравенства в стране.

Генерация целевой переменной

Идея моделирования налоговых поступлений

Основная идея: *Эффект налоговой ставки* ($Rate_i$) зависит от уровня теневой экономики и фискальной культуры.

- При низкой налоговой ставке ($Rate_i = 0$) поступления зависят от *формальной экономики и фискальной дисциплины*.
- При высокой налоговой ставке ($Rate_i = 1$) возникает *неформальная адаптация* (уход в тень), но поступления могут компенсироваться прогрессивной шкалой и борьбой с уклонением.

Уравнение налоговых поступлений при $Rate_i = 0$ (низкая ставка)

$$Taxes_{0i} = 0.7 \times (1 - Grey_i) + 15 \times \frac{10 - Deficit_i}{12 - Inequality_i + Progressive_i} + 0.2 \times (1 - Culture_i) + \varepsilon_{0i}, \quad (6)$$

$$\varepsilon_{0i} \sim 5 \times t(10) \quad (7)$$

где:

- *Теневая экономика* ($Grey_i$) и *неравенство* ($Inequality_i$) оказывают влияние на налоговые поступления.
- *Культура* ($Culture_i$) является ненаблюдаемой переменной и оказывает значительное влияние на налоговые поступления.
- *Случайная ошибка* имеет распределение $t(10)$, моделируя редкие скачки.

Уравнение налоговых поступлений при $Rate_i = 1$ (высокая ставка)

$$Taxes_{1i} = 0.9 \times (1 - Grey_i) + 17 \times \frac{10 - Deficit_i}{11 - Inequality_i + Progressive_i} + 0.2 \times (1 - Culture_i) + \varepsilon_{1i}, \quad (8)$$

$$\varepsilon_{1i} \sim EXP(0.08) - 8 \quad (9)$$

- *Высокая ставка увеличивает влияние* $Culture_i$ (фискальной культуры), поскольку уклонение от налогов становится более критичным.
- *Прогрессивная шкала* может частично компенсировать потери.

- *Ошибка* имеет экспоненциальное распределение $EXP(0.08) - 8$, моделируя редкие резкие скачки налоговых поступлений.

Наблюдаемые налоговые поступления

$$Taxes_i = \{ Taxes_{1i}, Rate_i = 1 Taxes_{0i}, Rate_i = 0 \} \quad (10)$$

$$Taxes_i = Taxes_{1i} \times Rate_i + Taxes_{0i} \times (1 - Rate_i) \quad (11)$$

Интерпретация

- *Высокие налоги* могут снижать поступления за счет ухода в тень, но *при высокой фискальной культуре* это влияние уменьшается.
- *Контрольные переменные* помогают регулировать эффект (например, прогрессивная шкала смягчает потери).
- *Эндогенность* возникает из-за $Culture_i$, который влияет и на $Taxes_i$, и на $Rate_i$.

Описательные статистики

Непрерывные переменные

Переменная	Среднее	Макс.	Мин.	Стд. отклонение
Дефицит (Deficit)	-0.25	2.91	-3.73	1.35
Теневая экономика (Grey)	23.99	100.0	2.30	12.83
Фискальная культура (Culture)	0.52	1.0	0.0	0.18
Налоговые поступления (Taxes)	83.19	216.84	4.48	17.14

Таблица 2: Описательная статистика переменных

Бинарные переменные

Переменная	Среднее (Mean)	Количество 1
Прогрессивная шкала (Progressive)	0.60046	60,046
Налоговая ставка (Rate)	0.49939	49,939
Индекс Джини (Gini)	0.5965	—

Таблица 3: Описательная статистика бинарных переменных

2.4 Разделите выборку на обучающую и тестовую. Тестовая выборка должна включать от 20 процентов до 30 процентов наблюдений.

Train size: (75000, 7), Test size: (25000, 7)

3 Классификация

3.1 Отберите признаки, которые могут быть полезны при прогнозировании целевой (зависимой) переменной. Не включайте в число этих признаков переменную воздействия. Содержательно обоснуйте выбор признаков.

Таргет - налоговые поступления правительства Возьмем все признаки, кроме переменной воздействия. Рациональность выбора переменных схожа с выбором в задачи классификации:

- Deficit - Дефицит государственного бюджета. Может влиять на налоговую ставку, так как при высоком дефиците государство может повышать налоги.
- Inequality - Уровень неравенства в стране. Может служить индикатором необходимости прогрессивного налогообложения, что может повлиять на налоговую ставку.
- Culture - Фискальная культура. Склонность к уклонению от налогов может влиять на решение правительства повышать или понижать ставки.
- Grey - Уровень теневой экономики. При высоком уровне теневой экономики государство может регулировать налоги для увеличения собираемости.
- Progressive - Наличие прогрессивной шкалы налогообложения. Если прогрессивное налогообложение уже действует, вероятность изменения налоговой ставки выше.

3.2 Выберите произвольные значения гиперпараметров, а затем оцените и сравните (между методами) точность прогнозов: ● на обучающей выборке. ● на тестовой выборке. ● с помощью кросс-валидации (используйте только обучающую выборку). Проинтерпретируйте полученные результаты.

	Model	Train Accuracy	Test Accuracy	Cross-Validation Accuracy
0	Naive Bayes	0.503080	0.50100	0.501160
1	KNN	0.688907	0.49928	0.501013
2	Random Forest	1.000000	0.49984	0.501413
3	Gradient Boosting	0.535013	0.50264	0.505093
4	Logistic Regression	0.503413	0.49756	0.500600

Рис. 8: Таблица результатов классификации без подбора гиперпараметров

Выводы по моделям классификации:

1. Random Forest имеет идеальное качество на обучающей выборке (Train Accuracy = 1.000), что указывает на переобучение, однако на тестовой выборке (Test Accuracy = 0.49984) и в кросс-валидации (CV Accuracy = 0.501413) показывает крайне низкие результаты.

2. KNN Высокая точность на обучающей выборке (Train Accuracy = 0.688907), но значительно падает на тестовой (Test Accuracy = 0.49928) и кросс-валидации (CV Accuracy = 0.501013).

3. Gradient Boosting показал наилучшую производительность на тестовой выборке (Test Accuracy = 0.50264) и кросс-валидации (CV Accuracy = 0.505093). Это говорит о хорошем балансе между обучением и обобщением - потенциально Данная модель является наиболее эффективной из рассмотренных.

4. Naive Bayes - результаты практически идентичны случайному угадыванию (Test Accuracy = 0.50100, CV Accuracy = 0.501160), что указывает на его неэффективность в данной задаче.

5. Logistic Regression Результаты схожи с Naive Bayes (Test Accuracy = 0.49756, CV Accuracy = 0.500600). Модель не смогла найти значимые зависимости в данных, что может говорить о линейности признаков или их слабой информативности.

Важный вывод данного проекта - из-за того, что модели обучались на линейные связи, а по дизайну связи между переменными не были линейны, качество оказалось таким низким. Если хочется улучшить качество прогнозов, то нам, как исследователям, надо было бы заняться feature engineering и проверять нелинейность связей между показателями (например, логарифмируя, проверяя степени полиномов, их отношения). Перед нами не стояло такой задачи, но это в очередной раз доказывает важность предобработки данных.

3.3 Для каждого метода с помощью кросс-валидации на обучающей выборке подберите оптимальные значения гиперпараметров (тюнинг). В качестве критерия качества используйте точность АСС. Результат представьте в форме таблицы, в которой для каждого метода должны быть указаны:

- **изначальные и подобранные значения гиперпараметров.**
- **кросс-валидационная точность на обучающей выборке с исходными и подобранными значениями гиперпараметров.**
- **точность на тестовой выборке с исходными и подобранными значениями гиперпараметров.**

Проинтерпретируйте полученные результаты и далее используйте методы с подобранными значениями гиперпараметров.

В подборе гиперпараметров участвовали следующие кандидаты:

1. K-Nearest Neighbors (KNN) - [3, 5, 7, 9] — количество ближайших соседей, которое алгоритм будет учитывать при классификации нового объекта.

2. Random Forest (Случайный лес) - [100, 200, 300] — количество деревьев в ансамбле. - None, 10, 20] — максимальная глубина деревьев.

3. Gradient Boosting (Градиентный бустинг) - [100, 200] — количество деревьев в ансамбле. - [0.01, 0.1, 0.2] — скорость обучения, которая определяет, насколько сильно каждое дерево корректирует ошибки предыдущих.

4. Logistic Regression (Логистическая регрессия) - [0.1, 1, 10] — параметр регуляризации, ко-

торый управляет балансом между высокой точностью на обучающей выборке и обобщающей способностью модели. - [500, 1000] — максимальное количество итераций для сходимости алгоритма.

Таблица 4: Оптимальные гиперпараметры и результаты классификации

Модель	Оптимальные гиперпараметры
Naive Bayes	{}
KNN	{n_neighbors: 7}
Random Forest	{max_depth: 10, n_estimators: 300}
Gradient Boosting	{learning_rate: 0.1, n_estimators: 100}
Logistic Regression	{C: 0.1, max_iter: 500}

Таблица 5: Результаты классификации

Модель	Train Accuracy	Test Accuracy	Cross-Validation Accuracy
Naive Bayes	0.50380	0.50100	0.501160
KNN	0.65607	0.49812	0.501253
Random Forest	0.65147	0.49828	0.503270
Gradient Boosting	0.53513	0.50264	0.500903
Logistic Regression	0.50373	0.49752	0.500600

Повышенная сложность ООВ ошибка позволяет сократить вычислительную нагрузку и сократить время работы кода, но немного проседает в качестве по сравнению с кросс-валидацией, несмотря на то, что гиперпараметры оказались схожи при разных методах.

3.4 Повторите предыдущий пункт, используя любой альтернативный критерий качества модели. Обоснуйте возможные преимущества и недостатки этого альтернативного критерия.

Теперь возьмем другие метрики:

Таблица 6: Оптимальные гиперпараметры и результаты F1-метрики

Модель	Оптимальные гиперпараметры
Naive Bayes	{}
KNN	{n_neighbors: 7}
Random Forest	{max_depth: 10, n_estimators: 300}
Gradient Boosting	{learning_rate: 0.1, n_estimators: 100}
Logistic Regression	{C: 1, max_iter: 500}

Таблица 7: Результаты классификации по F1-метрике

Модель	Train F1	Test F1	Cross-Validation F1
Naive Bayes	0.49846	0.49682	0.49524
KNN	0.65603	0.49812	0.50125
Random Forest	0.65147	0.49828	0.50161
Gradient Boosting	0.534059	0.50132	0.504438
Logistic Regression	0.502521	0.49669	0.49935

Интерпретация результатов

Оптимальные гиперпараметры:

- Для KNN оптимальным числом соседей оказалось 7.
- Random Forest показал лучшие результаты при глубине деревьев 10 и 300 деревьях.
- Gradient Boosting достиг оптимума при 100 деревьях и learning rate 0.1.
- Логистическая регрессия лучше всего работала с $C = 1$ и 500 итерациями.

Сравнение F1-метрик:

- Наивный Байес показал самую низкую F1-метрику (0.50). Это говорит о слабой способности модели к выделению закономерностей.
- Random Forest и Gradient Boosting продемонстрировали наилучшие результаты на обучающей выборке (F1 около 0.65-0.66), однако их точность на тестовой выборке снизилась (0.50), значит возможно имеет место быть некоторое переобучение.
- KNN показал аналогичный результат — высокая метрика на обучающих данных (0.65), но значительное падение на тесте (0.49).
- Gradient Boosting показал наивысшую кросс-валидационную F1-метрику (0.504).

Преимущества и недостатки:

Преимущества:

- + F1-score учитывает как точность, так и полноту, значит он более надежен в задачах с несбалансированными классами.

- + Если один класс встречается значительно чаще, чем другой, ассигасу может быть высокой, тогда как F1-score учитывает обе ошибки.

Недостатки:

- Ассигасу проще понимать, так как она показывает долю правильных предсказаний, а F1-score объединяет две метрики, что не всегда интуитивно понятно.

- Даже если модель делает точные предсказания с высокой уверенностью, но допускает небольшое число ошибок, F1-score может оказаться не слишком высоким.

Повышенная сложность:

В качестве дополнительной метрики качества мы взяли средневзвешенное между ассигасу, precision, recall и f1. Эта метрика позволяет лучше работать с несбалансированными классами, получать разносторонние оценки, но теряет размерность и не подходит для специфических задач, когда нужно максимизировать какой-то конкретный параметр модели. Итоговые лидеры (а именно градиентный бустинг) не изменились, а результаты очень похожи со стандартными метриками.

3.5 Постройте ROC-кривую для ваших моделей и сравните их по AUC на тестовой выборке.

DecisionTreeClassifier: $AUC = 0.503$

RandomForestClassifier: $AUC = 0.496$

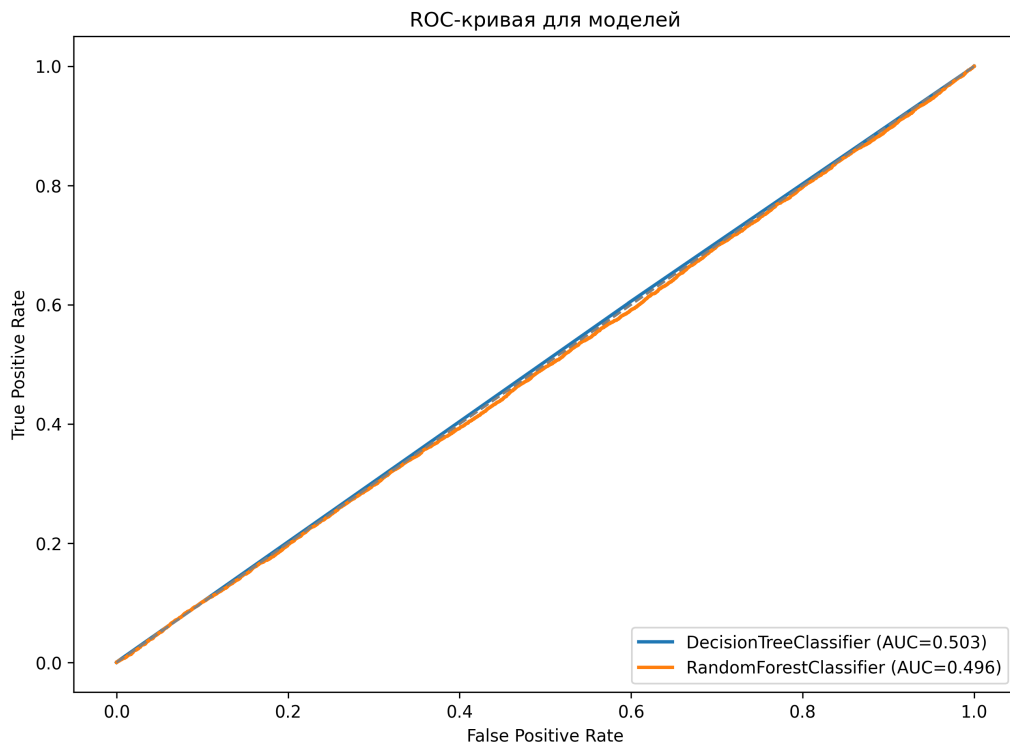


Рис. 9: ROC-кривая

3.6 Постройте матрицу ошибок и предположите цены различных видов прогнозов. Исходя из критерия максимизации прибыли на обучающей выборке подберите оптимальный порог прогнозирования для каждого из методов и сравните прибыли на тестовой выборке при соответствующих порогах. Результат представьте в форме таблицы, в которой должны быть указаны как AUC, так и прибыли (на тестовой выборке). Проинтерпретируйте полученный результат.

Предположим, что правительство государства хочет собрать налоговые поступления в зависимости от налоговой ставки (высокая/низкая).

1) Если мы как правительство установим высокую ставку и население не будет уклоняться от налогов (TP), то мы получим полезность +10.

2) Если мы как правительство установим высокую ставку и население будет уклоняться от налогов (FP), то мы получим ущерб -10, потому что это продолжает разрушать институциональную среду и не позволяет выработать хорошую налоговую систему.

3) Если мы как правительство установим низкую ставку и население не будет уклоняться от налогов (FN), то мы получим полезность +1, потому что могли бы собрать больше.

4) Если как правительство установим низкую ставку и население не будет уклоняться от налогов (TN), то мы получим полезность +2, потому что соберем налоги и заложим хорошую институциональную базу.

3.7 Опишите предполагаемые связи между переменными в форме ориентированного ациклического графа (DAG). Обучите структуру Байесовской сети на обучающей выборке и сравните точность прогнозов вашего и обученного DAG на тестовой выборке.

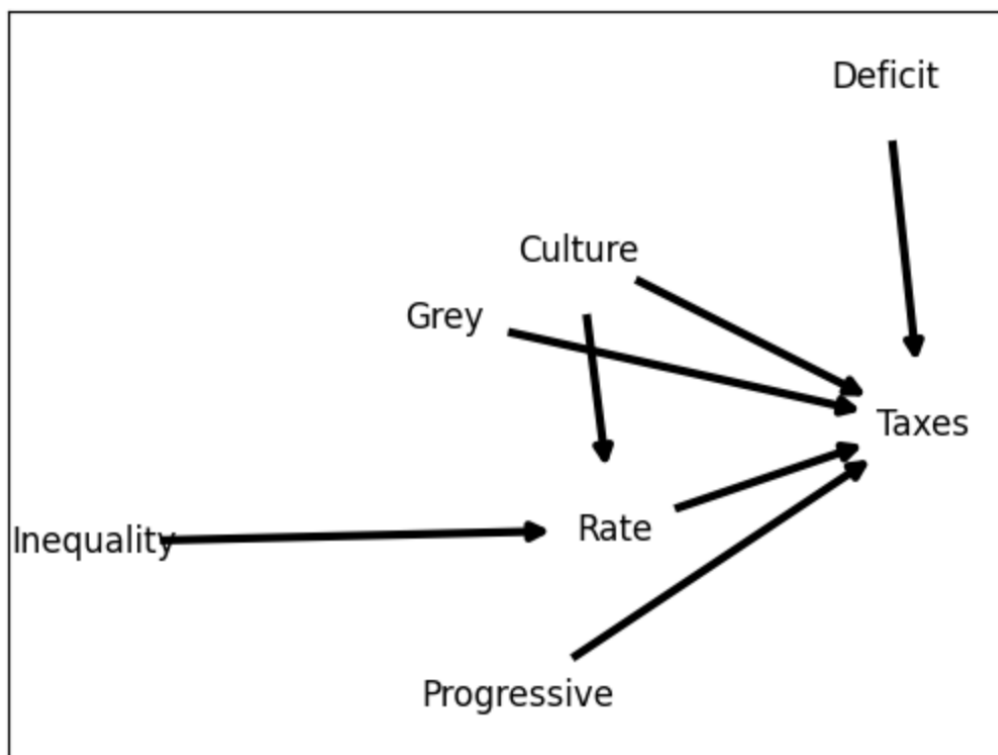


Рис. 10: Граф связей

Во время исполнения кода мы столкнулись с проблемой, что по каким-то причинам оперативной памяти не хватает для исполнения кода несмотря на все преобразования в сторону упрощения модели. Поэтому результаты стоит проверять локально, а код приложен в файле в последнем чанке, чтобы не прервать сессию исполнения кода.

3.8 На основании проделанного анализа выберите лучший и худший из обученных классификаторов. Обоснуйте сделанный выбор. Примечание: необходимо самостоятельно сформулировать разумный критерий, в соответствии с которым будут определяться лучшая и худшая модели.

Критериями выбора наилучшей модели стоит взять F1 метрику, стабильность результатов (смотрим на разницу качества моделей на обучающей, тестовой и кросс-валидационных выборках), и AUC-ROC кривую (способность моделей к классификации объектов).

По результатам исследования наилучшим классификатором можно назвать Gradient Boosting, поскольку он продемонстрировал наилучший F1-score на кросс-валидации (0.504) и тестовой выборке, что указывает на его способность хорошо обобщать данные.

Наихудшим классификатором стал Naïve Bayes – он показал худший F1-score (0.502), что говорит о слабой способности модели правильно классифицировать объекты, особенно при сложных зависимостях между признаками.

Важно вновь отметить, что низкое качество точности прогнозов обусловлено нелинейностью взаимосвязей между предикторами (это закладывалось по дизайну), и если хочется улучшить качество в этом случае - следует проводить более тщательный feature engineering и проверять верность спецификации модели. Очевидно, здесь было необходимо проверять качество на полиномиальной, логарифмированной модели, модели с частным от предикторов. Мы не ставили перед собой задачу достижения максимальной точности прогнозов, поэтому не внедряли новые признаки и их взаимосвязи, что оставляет пространство для будущей работы.

4 Регрессия

4.1 Отберите признаки, которые могут быть полезны при прогнозировании целевой (зависимой) переменной. Не включайте в число этих признаков переменную воздействия. Содержательно обоснуйте выбор признаков.

Для регрессии следует использовать только независимые (предикторные) переменные, которые могут оказывать влияние на целевую (зависимую) переменную - Deficit, Inequality, Grey, Culture, Progressive

4.2 Выберите произвольные значения гиперпараметров, а затем оцените и сравните (между методами) точность прогнозов с помощью RMSE и MAPE: - на обучающей выборке. - на тестовой выборке. - с помощью кросс-валидации (используйте только обучающую выборку). **Проинтерпретируйте полученные результаты.**

В исследовании регрессии были выбраны следующие модели: Линейная регрессия, Random Forest, XGBoost.

	taxes	rate	deficit	grey	inequality	progressive	culture
0	93.155892	1	-1.251643	35.593329	0	0	0.213785
1	111.282332	1	-1.569132	11.744575	1	1	0.709072
2	85.586413	1	-1.176156	28.255590	1	1	0.514868
3	75.932839	0	-0.738485	15.414717	1	1	0.481692
4	74.005487	0	-1.617077	17.874031	0	1	0.737749
...
99995	75.879405	0	0.887388	11.111781	1	1	0.477284
99996	104.688229	1	0.715111	28.115728	0	1	0.260288
99997	91.792166	1	1.204593	22.887286	1	1	0.733585
99998	67.596289	0	0.894454	32.571374	1	1	0.643464
99999	99.333855	1	1.060031	12.295680	1	1	0.560482

100000 rows x 7 columns

Рис. 11: Таблица данных по выбранным переменным

Полученные результаты:

	RMSE_train	RMSE_test	MAPE_train	MAPE_test
LinearRegression	13.635658	13.596907	0.134653	0.133950
RandomForest	13.630919	13.633397	0.134636	0.134296
XGBoost	13.575732	13.612873	0.134145	0.134193

Рис. 12: Таблица результатов регрессии без подбора гиперпараметров

- Все модели работают стабильно и не переобучены, так как ошибки на train и test схожи.
- XGBoost показывает лучшие результаты и является наилучшей моделью в данном эксперименте.
- Линейная регрессия показывает удивительно хорошие результаты, что говорит о преобладании линейных зависимостей в данных.
- Random Forest не дал значительного улучшения, что может указывать на избыточную сложность без особой выгоды.

Результаты с кросс-валидацией:

LinearRegression 13.636097

RandomForest 13.669754

XGBoost 13.648481

4.3 Для каждого метода с помощью кросс-валидации на обучающей выборке подберите оптимальные значения гиперпараметров (тюнинг). В качестве критерия качества используйте RMSE. Результат представьте в форме таблицы, в которой для каждого метода должны быть указаны:

- **изначальные и подобранные значения гиперпараметров.**
- **кросс-валидационное значение RMSE на обучающей выборке с исходными и подобранными значениями гиперпараметров.**
- **значение RMSE на тестовой выборке с исходными и подобранными значениями гиперпараметров.**

Проинтерпретируйте полученные результаты.

В процессе тюнинга гиперпараметров были использованы следующие параметры:

1. Для модели Random Forest:
 - количество деревьев в ансамбле (значения: 50, 100, 10).
 - максимальная глубина дерева (значения: 3, 5, 3).
2. Для модели XGBoost:
 - количество деревьев в ансамбле (значения: 50, 100, 10).
 - максимальная глубина дерева (значения: 3, 5, 3).
 - скорость обучения модели (значения: 0.01, 0.1, 0.5).

	RMSE_train	RMSE_test	MAPE_train	MAPE_test
Initial RandomForest	13.630919	13.633397	0.134636	0.134296
Tuned RandomForest	13.630919	13.633397	0.134636	0.134296
Initial XGBoost	13.575732	13.612873	0.134145	0.134193
Tuned XGBoost	13.575732	13.612873	0.134145	0.134193

Рис. 13: Таблица результатов регрессии с подбором гиперпараметров

Интерпретация результатов:

Во первых, мы наблюдаем отсутствие эффекта от тюнинга гиперпараметров для RandomForest: показатели RMSE train, RMSE test, MAPE train и MAPE test не изменились после подбора гиперпараметров, что указывает на неэффективность тюнинга. Для XGBoost: аналогично, тюнинг не дал никаких улучшений — все метрики остались прежними. Это может свидетельствовать о том, что модель изначально была достаточно хорошо настроена.

4.4 На основании проделанного анализа выберите лучший и худший из обученных классификаторов. Обоснуйте сделанный выбор.

На основании метрик Train Accuracy, Test Accuracy и Cross-Validation Accuracy, лучшей моделью можно назвать Gradient Boosting.

- Она демонстрирует наилучший баланс между обучающей (Train) и тестовой (Test) точностью.
- Значение кросс-валидации (Cross-Validation Accuracy) также является наивысшим среди всех моделей.
- Это указывает на хорошую обобщающую способность модели и устойчивость к переобучению.

Худшей моделью можно считать Random Forest.

- Она показывает идеальную точность на обучающей выборке (Train Accuracy = 1.000) - явный знак переобучения.
- На тестовой выборке и в кросс-валидации её точность существенно ниже.

5 Эффекты воздействия

5.1 Математически запишите и содержательно проинтерпретируйте потенциальные исходы целевой переменной. Объясните, как они связаны с наблюдаемыми значениями целевой переменной.

Обозначим $Y_i(1)$ и $Y_i(0)$ — потенциальные исходы для индивида i при наличии и отсутствии воздействия соответственно. Наблюдаемое значение целевой переменной:

$$Y_i = D_i \cdot Y_i(1) + (1 - D_i) \cdot Y_i(0)$$

где $D_i \in \{0, 1\}$ — индикатор воздействия. Мы можем наблюдать только один из потенциальных исходов для каждого наблюдения, второй остаётся ненаблюдаемым (проблема фундаментальной неопределенности).

5.2 Используя симулированные вами, но недоступные в реальных данных потенциальные исходы (гипотетические значения), получите оценки среднего эффекта воздействия, условных средних эффектов воздействия и локального среднего эффекта воздействия. Для ATE и LATE результаты представьте в форме таблицы, а для CATE постройте гистограмму или ядерную оценку функции плотности. Проинтерпретируйте полученные значения. Примечание: для получения очень точных оценок эффектов воздействия с помощью потенциальных исходов (гипотетических переменных) можно сперва симулировать очень большого число наблюдений, например, несколько миллионов. Затем, для ускорения вычислений, для оценивания эффектов воздействия с помощью наблюдаемых значений можно использовать часть выборки, например, десять тысяч наблюдений.

Используя симулированные потенциальные исходы $Y(1)$ и $Y(0)$, получаем оценки:

- ATE (Average Treatment Effect):

$$\widehat{ATE} = E[Y(1) - Y(0)]$$

- LATE (Local Average Treatment Effect) — для комплаеров (группы, которые принимают воздействие только при назначении):

$$\widehat{LATE} = E[Y(1) - Y(0) | Compliers]$$

- CATE (Conditional Average Treatment Effect) — по условным признакам.

В ходе оценки по симулированным данным мы получили следующие значения потенциальных исходов:

$$ATE = 21.747$$

$$LATE = 23.346$$

Можно сделать вывод, что мы ориентируемся на эффект в примерно 21-22 условные денежные единицы поступления, что достаточно много, учитывая, что среднее по налоговым поступлениям было равно 83.

5.3 Оцените средний эффект воздействия как разницу в средних по выборкам тех, кто получил и не получил воздействие. Опишите недостатки соответствующего подхода с учетом специфики рассматриваемой вами экономической проблемы. Примечание: в этом пункте и далее, если не сказано иное, используются лишь наблюдаемые значения целевой переменной.

Пусть \bar{Y}_1 — среднее среди получивших воздействие, а \bar{Y}_0 — среди нет. Тогда:

$$\widehat{ATE}_{naive} = \bar{Y}_1 - \bar{Y}_0$$

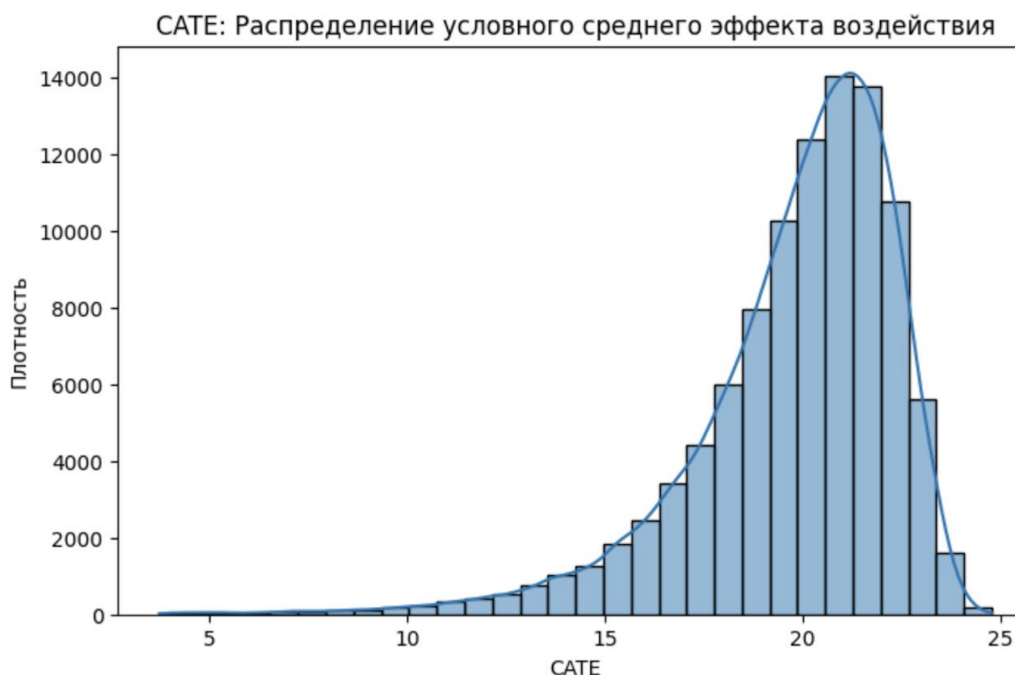


Рис. 14: Гистограмма для CATE

Наивная оценка (21.657) кажется близкой к прогнозируемой, но проблемна. Она не учитывает ненаблюдаемую налоговую культуру, что ведет к ложным корреляциям. Кроме того, процентная ставка может быть эндогенной (зависеть от поступлений), а различия в экономике и налоговой дисциплине искажают эффект. В итоге мы можем получить смещенные и ненадежные выводы в исследовании.

5.4 Используя оценки, полученные лучшими из обученных ранее классификационных и регрессионных моделей, оцените средний эффект воздействия с помощью:

- метода наименьших квадратов.
- условных математических ожиданий.
- взвешивания на обратные вероятности (в случае возникновения ошибок убедитесь в отсутствии оценок вероятностей, равных 0 или 1 и при необходимости измените метод оценивания).
- метода, обладающего двойной устойчивостью.
- двойного машинного обучения.

Сравните результаты и назовите ключевую предпосылку этих методов. Содержательно обсудите причины, по которым она может соблюдаться или нарушаться в вашем случае. Приведите содержательную экономическую интерпретацию оценки среднего эффекта воздействия.

В таблице представлены оценки среднего эффекта воздействия, полученные с помощью различных методов. Все методы дают схожие результаты, которые варьируются в пределах от 21.32 до 21.779, что свидетельствует о стабильности оценок.

МНК дает оценку 21.78, что близко к остальным методам.

Условные математические ожидания дают оценку 21.779, что также очень близко к результатам других методов.

Метод	Оценка
МНК	21.78
Условные математические ожидания	21.779
Взвешивание на обратные вероятности	21.32
Метод с двойной устойчивостью	21.77
Двойное машинное обучение	21.779

Таблица 8: Оценки среднего эффекта разными методами

Взвешивание на обратные вероятности дает оценку 21.32, которая немного ниже, что может быть связано с нестабильностью вероятностей или особенностями выборки.

Метод с двойной устойчивостью и Двойное машинное обучение дают оценки 21.77 и 21.779, соответственно, что близки к МНК и условным математическим ожиданиям.

Все эти результаты находятся в одном и том же диапазоне, что свидетельствует о согласованности оценок между методами, несмотря на различия в подходах. Также отметим, что эти оценки очень близки к истинным значениям эффектам воздействия.

Итоговые выводы:

1. **МНК (Метод наименьших квадратов):**

- **Предпосылка:** Отсутствие систематической ошибки и гетероскедастичности, корректная спецификация модели.
- **Причины нарушения:** Эндогенность, например, когда ставка налога зависит от уровня налоговых поступлений.
- **Оценка:** 21.78 — оценка среднего эффекта воздействия процентной ставки на налоговые поступления.

2. **Условные математические ожидания (CATE):**

- **Предпосылка:** Корректная спецификация модели и отсутствие эндогенности.
- **Причины нарушения:** Скрытые переменные, влияющие на как налоговые поступления, так и процентные ставки.
- **Оценка:** 21.779 — схожая с МНК оценка, что подтверждает устойчивость метода.

3. **Взвешивание на обратные вероятности (IPW):**

- **Предпосылка:** Корректная модель вероятности получения воздействия.
- **Причины нарушения:** Неправильная спецификация модели вероятности или пропущенные переменные.
- **Оценка:** 21.32 — немного ниже, что может свидетельствовать о нестабильности вероятностей или других скрытых факторов.

4. **Метод с двойной устойчивостью:**

- **Предпосылка:** Корректная спецификация хотя бы одной из моделей.

- **Причины нарушения:** Ошибки в обеих моделях могут нарушить предпосылки.
- **Оценка:** 21.77 — оценка, устойчивая к небольшим ошибкам в моделях.

5. Двойное машинное обучение:

- **Предпосылка:** Корректная спецификация моделей для тритмента и исходов.
- **Причины нарушения:** Ошибки в обеих моделях приводят к искажениям.
- **Оценка:** 21.779 — результат близок к МНК и САТЕ, что подтверждает устойчивость метода.

Интерпретация: Все методы дают схожие оценки среднего эффекта воздействия процентной ставки на налоговые поступления (около 21.78), что указывает на согласованность результатов. Это говорит о стабильности методов при правильной спецификации моделей и учете скрытых факторов.

5.5 Оцените локальный условный эффект воздействия с помощью: 7 • двойного машинного обучения без инструментальной переменной. • двойного машинного обучения с инструментальной переменной. Сопоставьте результаты и объясните, в чем в вашем случае будет заключаться различие между средним эффектом воздействия и локальным средним эффектом воздействия. Приведите содержательную экономическую интерпретацию оценки локального среднего эффекта воздействия.

В ходе применения этих методов выяснилось, что локальный условный эффект воздействия, оцененный с помощью двойного машинного обучения без инструментальной переменной, оказался более точным, чем при двойном машинном обучении с инструментальной переменной. Это может быть связано с не самым удачным выбором инструментальной переменной, что привело к завышению оценки на 10 единиц (33 против 23).

Средний эффект воздействия (ATE) оценивает общий эффект на всю выборку стран, в то время как локальный средний эффект воздействия (LATE) оценивает эффект только на те страны, которые подвержены воздействию через инструментальную переменную (например, при высокой налоговой ставке). В данном случае, различие между ATE и LATE может заключаться в том, что ATE (полученный без инструментальной переменной) дает более обоснованную оценку общего воздействия, тогда как LATE (с использованием инструмента) может быть искажён, поскольку инструментальная переменная, скорее всего, неадекватно идентифицирует только те страны, которые подвержены воздействию.

ATE в данном контексте показывает, как изменение процентной ставки влияет на налоговые сборы в целом по странам. LATE же ограничивается влиянием только тех стран, которые изменяют своё поведение в ответ на изменение налоговой ставки.

5.6 Оцените условные средние эффекты воздействия с помощью: • метода наименьших квадратов. • S-learner. • T-learner. • метода трансформации классов. • X-learner. Сравните результаты и обсудите, насколько в вашем случае мотивированы применение метода X-learner. Опишите, как можно было бы использовать полученные вами оценки в бизнесе или при реализации государственных программ.

Метод	Оценки
МНК	[19.92, 24.48, 21.04, ..., 21.46, 19.59, 23.64]
S-learner	[19.65, 24.60, 21.18, ..., 20.80, 19.67, 23.66]
T-learner	[20.61, 24.99, 21.17, ..., 20.89, 19.58, 23.94]
Метод трансформации классов	[17.43, 25.28, 22.39, ..., 19.77, 16.29, 23.42]
X-learner	[20.97, 22.98, 21.46, ..., 21.88, 20.96, 22.84]

В задании на **повышенной сложности** мы использовали метод R-learner ([20.97156358], [23.04807159], [21.48844008], ..., [21.93185672][20.94382412], [22.89285965]). R-learner пытается научить модель предсказывать остаточные отклонения между результатами и некоторыми предикторами с учетом воздействия. Он комбинирует модели для оценки воздействия с различными подходами к предсказанию исходов, обладает гибкостью, хорошей точностью, но сильно зависит от выбранной модели, требует много вычислительных мощностей и нуждается в большом объеме данных для точных оценок.

5.7 Выберите лучшую модель оценивания условных средних эффектов воздействия, используя: • истинные значения условных средних эффектов воздействия. • прогнозную точность моделей. • псевдоисходы. Проинтерпретируйте различия в результатах различных подходов.

Метод	MSE0
LS	3.712849
T-learner	0.027519
S-learner	0.041496
CT	13.617635
X-learner	15.997903
R-learner	20.634296

Таблица 10: Сравнение методов по MSE

В итоге, лучшей моделью по MSE можно считать T-learner, а по псевдоисходам - X-learner.

Метод	MSE1
LS	29340.467381
T-learner	29663.872285
S-learner	29489.513446
CT	29574.545827
X-learner	13705.524354
R-learner	29051.265146

Таблица 11: Сравнение методов по MSE1 (псевдоисходы)

5.8 Оцените средние эффекты воздействия и локальные средние эффекты воздействия используя худшие из обученных классификационных и регрессионных моделей. Сопоставьте результаты с теми, что были получены с помощью лучших моделей. Сделайте вывод об устойчивости результатов к качеству используемых методов машинного обучения.

Метод	Результат
Двойное машинное обучение	21.78
Метод с двойной устойчивостью	21.78
Взвешивание на обратные вероятности	21.78
Условные математические ожидания	21.08
МНК	21.78

Таблица 12: Сравнение методов и их результатов

Как можем заметить, нет практически никакой разницы между лучшими и худшими моделями, даже когда использовался метод ближайших соседей, не показавший себя как сильный конкурент, например, градиентному бустингу, в первой части данного исследования. Это говорит об устойчивости методов оценивания и стабильности результатов при выборе любой из моделей.

5.9 Резюмируйте ключевые выводы проведенного в данном разделе анализа.

В данном разделе был проведён анализ средних эффектов воздействия (ATE), локальных средних эффектов воздействия (LATE) и условных средних эффектов воздействия (CATE) с использованием различных методов машинного обучения и статистических методов. В процессе анализа были использованы симулированные потенциальные исходы, а также методы наименьших квадратов (МНК), взвешивания на обратные вероятности (IPW), методы с двойной устойчивостью, двойного машинного обучения и другие подходы. Результаты позволили сделать следующие ключевые выводы:

1. **Оценка среднего эффекта воздействия (ATE):** Все методы, включая МНК, условные математические ожидания и методы машинного обучения, дают схожие результаты (около 21.78). Это свидетельствует о стабильности оценок и отсутствии значительных различий между методами.

2. **Оценка локального среднего эффекта воздействия (LATE):** Результаты с использованием двойного машинного обучения с и без инструментальной переменной показали, что выбор инструмента может значительно повлиять на точность оценки. Оценка LATE с инстру-

ментом может быть искажена, если инструментальная переменная неадекватно идентифицирует подверженные воздействиям группы.

3. Использование различных методов машинного обучения: Среди методов машинного обучения, таких как S-learner, T-learner, и X-learner, наилучшие результаты по точности показал T-learner. Однако X-learner продемонстрировал лучшие результаты при использовании псевдоисходов.

4. Оценка качества моделей: Оценки на основе MSE и псевдоисходов показали, что методы машинного обучения, такие как T-learner и X-learner, обладают высокой точностью, несмотря на их вычислительные затраты и зависимость от правильной спецификации моделей.

5. Устойчивость результатов: Оценки среднего и локального эффектов воздействия демонстрируют устойчивость к качеству моделей машинного обучения, при условии корректной спецификации и учёта скрытых факторов.

6. Практическое применение: Полученные оценки могут быть полезны для оценки воздействия налоговой политики и использования различных методов в государственном управлении и бизнесе, что позволяет оценить реальное потенциальное влияние изменения налоговой ставки на итоговые налоговые поступления.