МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РФ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«НОВОСИБИРСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ

ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»  
(НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ, НГУ)

Факультет информационных технологий

Кафедра \_\_\_Компьютерных систем\_\_\_\_\_\_

Направление подготовки 09.03.01 Информатика и вычислительная техника

Направленность (профиль): Программная инженерия и компьютерные науки

**ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА БАКАЛАВРА**

**Шинкевича Вячеслава Сергеевича**

Тема работы:

**РаЗРАБОТКА КОМПЬЮТЕРНОЙ СИСТЕМЫ ПРОГНОЗИРОВАНИЯ НАЛОГОВЫХ ПОСТУПЛЕНИЙ и акцизов**

|  |  |
| --- | --- |
| **«К защите допущена»** | **Руководитель ВКР** |
| Заведующий кафедрой, | к.т.н., доцент каф. КС |
| к.т.н. |  |
| Пищик Б.Н./………… | Осипов А.Л./………... |
| (ФИО) / (подпись) | (ФИО) / (подпись) |
| «……»………………20…г. | «……»………………20…г. |
|  |  |

Новосибирск, 2024

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РФ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«НОВОСИБИРСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ

ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»  
(НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ, НГУ)

Факультет информационных технологий

Кафедра………Компьютерных систем…………………………………………

(название кафедры)

Направление подготовки 09.03.01 Информатика и вычислительная техника

Направленность (профиль): Программная инженерия и компьютерные науки

|  |
| --- |
| 1. УТВЕРЖДАЮ 2. Зав. кафедрой…Пищик Б.Н.… 3. (фамилия, И., О.) 4. …………………. 5. (подпись) 6. «19»\_\_октября\_\_2023г. |

**ЗАДАНИЕ**

**НА ВЫПУСКНУЮ КВАЛИФИКАЦИОННУЮ РАБОТУ БАКАЛАВРА**

|  |
| --- |
| Студент*у(ке)*…Шинкевичу Вячеславу Сергеевичу, группы 20205…………… |

(фамилия, имя, отчество, номер группы)

|  |
| --- |
| Тема Разработка компьютерной системы прогнозирования налоговых поступлений и акцизов |

(полное название темы выпускной квалификационной работы)

|  |  |
| --- | --- |
| утверждена распоряжением проректора по учебной работе от\_\_\_\_19 окт. 2023г\_\_№\_0367 | |
| Срок сдачи студентом готовой работы \_\_\_\_\_20 мая\_\_\_\_2024 г. | |
| Исходные данные (или цель работы): Разработать компьютерную систему с Web-интерфейсом для прогнозирования налоговых поступлений и акцизов | |
| Структурные части работы:   1. Анализ предметной области и сбор информации о налоговых поступлениях и акцизах 2. Проведение анализа данных с помощью регрессионных и нейросетевых алгоритмов прогнозирования 3. Разработка базы данных по налоговым поступлениям и акцизам 4. Разработка интерфейса в виде Web-приложения к базе данных для анализа информации, хранимой в базе данных 5. Сравнительный анализ различных алгоритмов прогнозирования и выбор наилучшего 6. Подготовка ВКР | |
|  | |
| Руководитель ВКР | Задание принял к исполнению |
| Должность и место работы в НГУ, | Шинкевич В.С./\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| к.т.н., доцент | (ФИО студента) / (подпись) |
| Осипов А,Л./\_\_\_\_\_\_\_\_\_\_\_\_\_ | «19»\_\_\_\_октября\_\_\_2023г. |
| (ФИ О) / (подпись) |  |
| «19» октября 2023г. |  |

Содержание

[Список терминов и сокращений 6](#_Toc167126012)

[Введение 8](#_Toc167126013)

[1. Анализ предметной области 10](#_Toc167126014)

[1.1. Анализ налогов и акцизов, поступающих в федеральный бюджет 11](#_Toc167126015)

[1.2. Анализ существующих решений 14](#_Toc167126016)

[1.3. Анализ существующих алгоритмов прогнозирования 16](#_Toc167126017)

[1.3.1. Экспоненциальное сглаживание 16](#_Toc167126018)

[1.3.2. Двойное экспоненциальное сглаживание 16](#_Toc167126019)

[1.3.3. Множественная линейная регрессия 17](#_Toc167126020)

[1.3.4. ARIMA 21](#_Toc167126021)

[1.3.5. SSA 22](#_Toc167126022)

[1.3.6. Выбор алгоритмов прогнозирования 24](#_Toc167126023)

[1.4. Анализ метрик для временных рядов 25](#_Toc167126024)

[2. Проектирование и разработка системы прогнозирования 27](#_Toc167126025)

[2.1. Проектирование базы данных 28](#_Toc167126026)

[2.2. Проектирование и реализация бэкенда 30](#_Toc167126027)

[2.2.1. Описание реализации дата-слоя 31](#_Toc167126028)

[2.2.2. Описание реализации слоя вычислений 32](#_Toc167126029)

[2.2.3. Описание реализации других слоев 35](#_Toc167126030)

[2.3. Проектирование и реализация фронтенда 39](#_Toc167126031)

[3. Анализ применения алгоритмов 48](#_Toc167126032)

[3.1. Прогноз для НДПИ 49](#_Toc167126033)

[3.2. Прогноз для налога на прибыль 50](#_Toc167126034)

[3.3. Прогноз для НДС 51](#_Toc167126035)

[3.4. Прогнозы акцизов 52](#_Toc167126036)

[3.5. Прогноз объема федерального бюджета 56](#_Toc167126037)

[3.5.1 Прогноз полностью через приложение 56](#_Toc167126038)

[3.5.2 Полуавтоматический прогноз 58](#_Toc167126039)

[4. Заключение 61](#_Toc167126040)

[Список литературы 63](#_Toc167126041)

[Приложение 68](#_Toc167126042)

# Список терминов и сокращений

1. **Бюджет** – это форма образования и расходования денежных средств, предназначенных для финансового обеспечения задач и функций государства и местного самоуправления.
2. **Госбюджет** является важным рычагом влияния на развитие производительных сил общества, реализации экономической политики государства, которая направлена на рост уровня жизни населения страны.
3. **Федеральный бюджет** выполняет следующие задачи:

* перераспределение ВВП;
* государственное регулирование и стимулирование экономики;
* финансовое обеспечение бюджетной сферы и осуществление социальной политики государства;
* контроль за образованием и использованием централизованных фондов денежных средств.

1. **Доходы федерального бюджета** – денежные средства, поступающие в бюджет в безвозмездном и безвозвратном порядке в соответствии с законодательством Российской Федерации в распоряжение органов государственной власти Российской Федерации, за исключением средств, являющихся источниками финансирования дефицита бюджета.
2. **Бюджетный кодекс –** кодифицированный нормативно-правовой акт, устанавливающий общие принципы бюджетного законодательства Российской Федерации, организации и функционирования бюджетной системы Российской Федерации, правовое положение субъектов бюджетных правоотношений, определяет основы бюджетного процесса и межбюджетных отношений в Российской Федерации, порядок исполнения судебных актов по обращению взыскания на средства бюджетов бюджетной системы Российской Федерации, основания и виды ответственности за нарушение бюджетного законодательства Российской Федерации.
3. **Налоговое законодательство** – совокупность нормативных правовых актов государства, регулирующих налоговое право.
4. **Бэкенд –** компонент системы, запускающийся без графического интерфейса, и отвечающий за взаимодействие с внешним миром и другими внутренними сервисами.
5. **Фронтенд –** графический компонент системы, которые взаимодействует с бэкендом.

# Введение

Важной государственной задачей является эффективное функционирование его институтов, что обеспечивается наполнением бюджета на разных уровнях. Налоги и акцизы являются основным инструментом для решения данной задачи. Прогнозируя налоги, на основе макроэкономических и налоговых индикаторов, можно предсказывать тренд развития экономики и, как следствие, страны.

При планировании любого бюджета, важна объективная точность, поэтому при прогнозировании важно предусмотреть возможность различных вариаций факторов, по которым рассчитываются целевые показатели, а также возможность прогнозировать различными методами.

Автоматизированная мульти-алгоритмическая компьютерная система позволит упростить процесс прогноза и улучшить его качество и точность благодаря одновременному использованию нескольких алгоритмов прогнозирования, в то время как в существующих системах используется единственный выбранный заранее алгоритм.

**Цель дипломной работы** – разработать автоматизированную систему прогнозирования налоговых поступлений и акцизов, совмещенную с базой данных для хранения истории изменения налоговых поступлений, с использованием существующих алгоритмов прогнозирования.

Для достижения данной цели были поставлены следующие **задачи:**

* Изучить существующие алгоритмы прогнозирования и выбрать из них несколько.
* Собрать данные по налогам и акцизам.
* Разработать базу данных для хранения данных по налогам и акцизам.
* Разработать систему с задокументированным API для прогнозирования с использованием выбранных алгоритмов, которая будет использовать данные из базы данных.
* Разработать Web-интерфейс для взаимодействия с системой.
* Проанализировать прогнозы, полученные выбранными алгоритмами прогнозирования, и с помощью этих прогнозных значений получить прогноз объема Федерального бюджета на следующий календарный период.

**Практическая ценность** заключается в реализации моделей прогнозирования и хранения данных для прогнозирования в единой системе, а также возможность добавления в будущем новых алгоритмов прогнозирования помимо тех, которые будут реализованы в этой работе.

**Научная новизна** работы состоит в новом алгоритме, который использует одновременно несколько алгоритмов прогнозирования для их сравнительного анализа и выбора наилучшего варианта с помощью критериев из математической статистики.

**Структура работы.** Работа состоит из введения, 3 глав и заключения.

В первой главе будут перечислены налоги и акцизы, прогнозом которых мы будем заниматься в данной работе, и также будут рассмотрены выбранные алгоритмы прогнозирования.

Вторая глава будет посвящена разработке автоматизированной системы прогнозирования, а именно: будет описана структура БД; будут описаны принципы работы системы прогнозирования и точки доступа для web-приложения; будет описана структура web-приложения, внешний вид экранов, функционал и т.д.

Третья глава будет посвящена вычислению прогнозных значений налогов и акцизов на следующий календарный период с помощью разработанной системы, а также получению прогноза объема федерального бюджета с использованием полученных значений налогов и акцизов.

В заключении сформулированы основные выводы работы.

# 1. Анализ предметной области

В современном российском государстве основными источниками поступлений федерального бюджета являются: налоговые поступления, неналоговые поступления, безвозмездные поступления.

Налоговые поступления (далее налоги) расходуются на социально важные государственные потребности, такие как: обеспечение национальной обороны, социальной политики, здравоохранения, развития культуры и спорта, образование и обслуживание государственного долга.

Налоги в рамках государства выполняют следующие основные функции:

* распределение ВВП между отраслями и гражданами;
* обеспечение доходной базы системы государства;
* влияние на спрос и предложение в экономике.

Именно поэтому планирование федерального бюджета, а, следовательно, и планирование поступлений налогов, важно для государства при планировании бюджета на следующий календарный период.

Самым основным регулятором внутреннего рынка государства, в условиях рыночной экономики, являются налоги. Налогом является обязательный, индивидуально безвозмездный платеж, взимаемый с организаций и физических лиц в форме отчуждения, принадлежащих им на праве собственности, хозяйственного ведения или оперативного управления денежных средств в целях финансового обеспечения деятельности государства и (или) муниципальных образований.

Большую часть поступлений в бюджет занимают налоговые поступления, поэтому в данной работе рассматривается задача их прогнозирования.

## 1.1. Анализ налогов и акцизов, поступающих в федеральный бюджет

Согласно Бюджетному кодексу и налоговому законодательству, федеральный бюджет наполняют следующие налоги: налог на добавленную стоимость (НДС), налог на доходы физических лиц (НДФЛ), налог на добычу полезных ископаемых (НДПИ), налог на прибыль (НПр), акцизы и другие налоги и сборы. На рис. 1 представлен график налоговых поступлений[[1]](#Анализ_налогов):



Рисунок 1 – График налоговых поступлений в федеральный бюджет за 2023 год

**Налог на добычу полезных ископаемых (НДПИ)** является основным по объему поступлений в Федеральный бюджет РФ. Налогоплательщиками являются предприятия, организации, индивидуальные предприниматели, которые пользуются природными ресурсами. Основной налоговой базой являются полезные ископаемые, производимые при добыче нефти, природного газа, угля и полезных руд.

**Налог на добавленную стоимость (НДС)** является косвенным и вторым налогом по объему поступлений, а также наиболее стабильным источником наполнения бюджета. Объектом налогообложения является реализация товаров и услуг. Поступления осуществляются за счет реализации продавцом товаров, услуг, имущественных прав покупателю.

**Налог на прибыль организаций**: основным объектом налогообложения является прибыль организаций и предприятий, получаемая в процессе их деятельности. Прибылью является разность суммы доходов и суммы расходов. Объем налога зависит от организации и ее финансовых результатов. Данный налог уплачивают все юридические лица (ОАО, ООО и тд), а также иностранные юридические лица и организации, которые получают доход от источника в РФ.

**Налог на доходы физических лиц (НДФЛ)**: объектом налогообложения являются доходы, получаемые резидентами или нерезидентами Российской Федерации. Доходы, которые облагаются данным налогом – это доходы, получаемые при заработке, а также доходы, полученные при продаже и сдаче имущества в аренду, в том числе, доходы от различных выигрышей.

**Акцизы** – это косвенный налог, объектом налогообложения которого является реализация подакцизных товаров. Согласно Бюджетному кодексу и налоговому законодательству, подакцизными товарами являются[[2]](#Подакцизные_товары):

1. этиловый спирт;
2. спиртосодержащие лекарственные и косметические средства;
3. алкогольная продукция;
4. табачная продукция;
5. автомобили;
6. автомобильный бензин;
7. дизельное топливо;
8. моторные масла;
9. средние дистилляты (углеводороды, продукты первичной и вторичной переработки нефти);
10. бензол;
11. авиационный керосин;
12. природный газ;
13. жидкости для электронных систем доставки никотина;
14. табак;
15. виноград, использованный для производства вина;
16. этан;
17. сталь жидкая.

Акцизы работают почти также как налог на добавленную стоимость – объем акциза закладывается продавцом в итоговую стоимость подакцизного товара. Поэтому факторами, влияющими на объем акцизных поступлений, являются суммарная стоимость реализованного подакцизного товара и акцизная ставка на этот товар.

Стоит отметить, что больше всего на суммарный объем акцизных поступлений влияют акцизы на:

1. табачную продукцию;
2. табак;
3. автомобильный бензин;
4. дизельное топливо;
5. алкогольная продукция;
6. авиационный керосин;
7. природный газ;
8. нефтяное сырье.

Сбор и контроль исполнения данных налоговых доходов является одной из стратегических функций Федеральной налоговой службы, так как данные поступления формируют государственный бюджет.

## 1.2. Анализ существующих решений

Существующие методы прогнозирования в экономике основываются на:

* отдельном сборе данных для прогнозирования и хранении их в отдельной системе;
* последующем построении прогноза с помощью различных алгоритмов и методик вручную или с помощью компьютерной системы.

Поэтому скорость получения прогноза такими способами довольно мала из-за большого количества ручной работы. К тому же, если возникнет потребность поменять алгоритм прогнозирования, потребуется повторно выполнять прогнозирование экономических факторов, а при большом объеме данных это может занять дни или даже недели. Помимо этого, повышается риск ошибки из-за человеческого фактора при переносе данных из одной системы в другую[[3-11](#Анализ_существующих_решений)].

В связи с вышеизложенным возникает необходимость в разработке новой, более совершенной системы. Это должна быть автоматизированная мульти-алгоритмическая система прогнозирования экономических факторов, совмещающая в себе систему для прогнозирования и систему для хранения данных, которая позволит значительно ускорить процесс прогнозирования за счет совмещения двух ранее раздельных систем в одной. Также эта система должна позволять:

* С минимальными усилиями добавлять новые алгоритмы прогнозирования в случае, если потребуется перейти со старых алгоритмов на новые.
* Анализировать применение алгоритмов прогнозирования к различным экономическим факторам для выбора оптимальных факторов и алгоритмов.

Использование такой системы позволит с высокой скоростью строить прогнозы и анализировать качество этих прогнозов с помощью методов математической статистики, математических ошибок и графического представления. Также использование этой системы позволит с легкостью добавлять и обновлять данные по экономическим факторам и использовать новые данные для прогнозов. Более подробно про реализацию этой системы будет рассказано в следующей главе.

## 1.3. Анализ существующих алгоритмов прогнозирования

Временной ряд — собранный в разные моменты времени статистический материал о значении каких-либо параметров (в простейшем случае одного) исследуемого процесса. Изменение объема налога или акциза по годам можно представить в виде временного ряда, и с помощью алгоритмов прогнозирования для временных рядов получать прогноз объема налога или акциза на следующий календарный период[[12](#Общий_анализ_временных_рядов)].

Рассмотрим наиболее используемые алгоритмы прогнозирования временных рядов и выберем из них наиболее оптимальные и подходящие.

### 1.3.1. Экспоненциальное сглаживание

Один из самых простых в реализации и использовании алгоритм прогнозирования, который хорошо подходит для временных рядов, значения которых с течением времени находятся вокруг какой-либо константы. Например, для рядов, значения которых в рассматриваемые периоды изменяются в диапазоне A+-e, где А – константы, а е – ширина коридора.

Формула алгоритма:



где:

* – период расчета,
* – сглаженное значение за период ,
* – реальное значение ряда за период ,
* – коэффициент сглаживания.

### 1.3.2. Двойное экспоненциальное сглаживание

Это улучшение алгоритма экспоненциального сглаживания, которое хорошо подходит для прогнозирования рядов с выраженным трендом изменения значений. Формула алгоритма:

где:

* t – период расчета,
* – сглаженное значение за период t,
* – реальное значение ряда за период t,
* – коэффициент сглаживания,
* – величина влияния тренда
* – коэффициент сглаживания тренда

### 1.3.3. Множественная линейная регрессия

Этот алгоритм позволяет найти зависимость целевого значения (в нашем случае объем налога или акциза) от нескольких факторов в виде коэффициентов. Формула выглядит следующим образом:

где – спрогнозированное значение, – значения факторов.

Для использования этого алгоритма необходимо предварительно выбрать несколько факторов, от которых может зависеть целевое значение и вычислить значения коэффициентов, а для получения прогноза целевого значения необходимо сделать прогноз каждого из факторов любым подходящим способом и затем подставить спрогнозированные значения факторов в формулу линейной регрессии вместе с полученными значениями коэффициентов[[13](#Множественная_регрессия)].

Коэффициенты регрессии можно искать разными способами. Один из них – с помощью нейронной сети. При подборе коэффициентов с помощью нейронной сети используется нейронная сеть определенного вида, представленная на рис. 2:

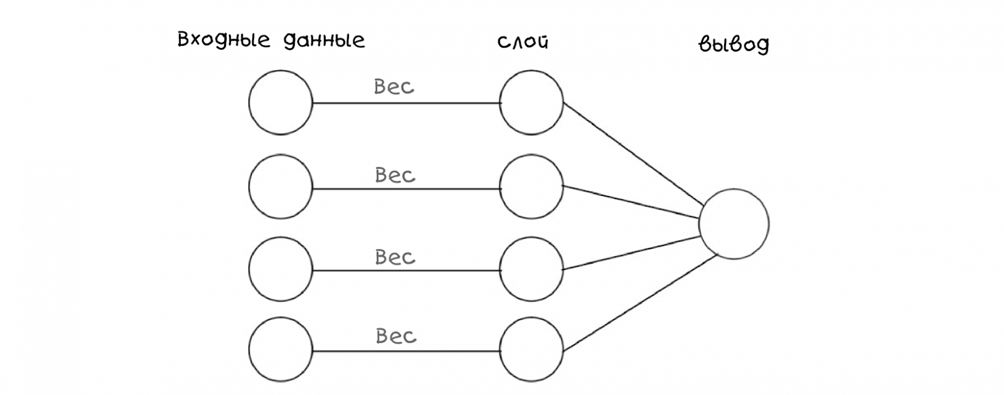


Рисунок 2 – Схема нейронной сети

где *входные данные* – значения факторов, *веса* – коэффициенты , *вывод* – наше спрогнозированное значение.

Алгоритм обучения нейронной сети:

1. Умножает входные данные на вес, связанный с определенным фактором (нейрон в контексте НС),
2. Суммирует перемноженные значения,
3. Применяет нормирующую функцию (функция активации):
4. Применяет нормирующую функцию к ожидаемому значению и получает ошибку:

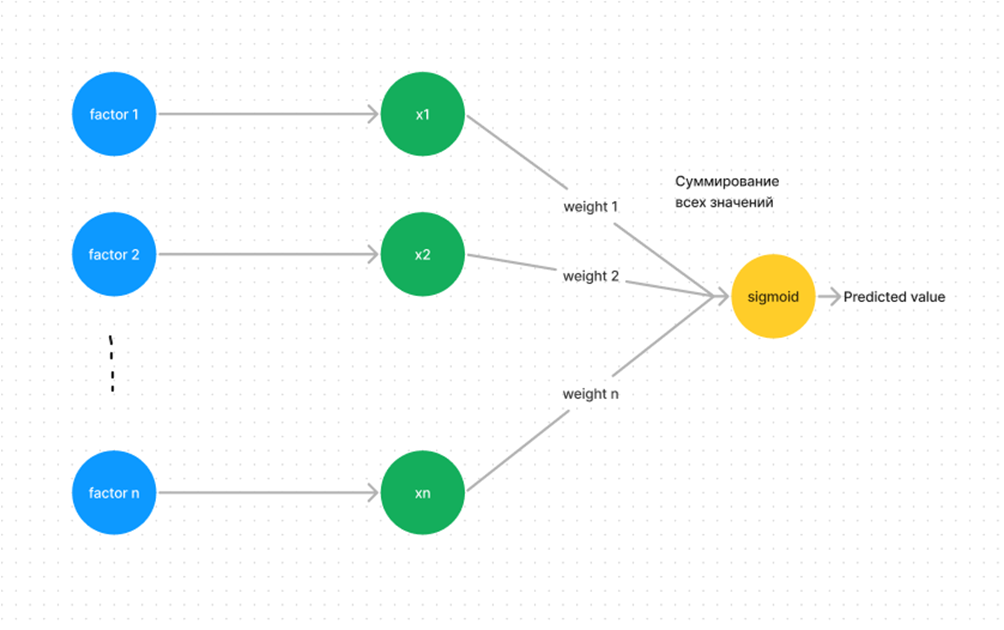
Степень у выражения будет меняться в зависимости от данных.

1. Изменяет веса в соответствии с формулой обратного распространения ошибки:

где:

* – ошибка из предыдущего пункта,
* – производная функции активации,
* – значение фактора, связанного с этим весом,
* – произвольный множитель, влияющий на скорость обучения НС. Выбирается экспериментально.

На рис. 3 показано графическое представление алгоритма:



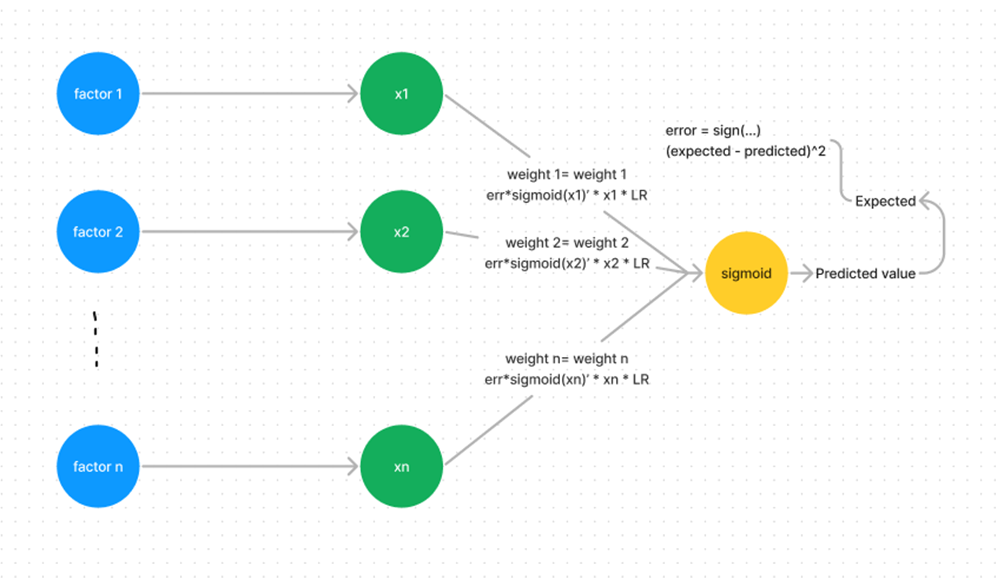


Рисунок 3 – Графическое изображение обучения нейросети и изменения весов нейронов

Данный алгоритм в среднем прогоняется 100000 раз для всего набора данных. Количество прогонов тоже выбирается экспериментально и важно выбрать его не слишком большим, иначе нейросеть найдет подходящие веса для конкретного набора данных и, когда полученные коэффициенты применятся к другим данным, может получиться слишком большое отклонение. Рис. 4 хорошо описывает эту ситуацию:

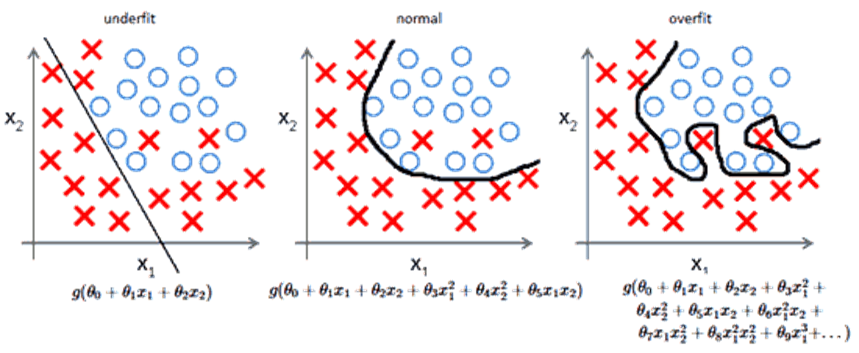


Рисунок 4 – Зависимость прогнозируемых значений от количества обучений на одном наборе данных

*Underfit* – НС недостаточно обучена, плохое прогнозирование.

*Normal* – НС обучена, приемлемое прогнозирование.

*Overfit* – НС переобучена, слишком точное прогнозирование, велика вероятность ошибок.

Чтобы понять стадию обучения НС будем использовать RMSE как индикатор. Сначала пять раз с нуля обучим НС на имеющихся данных и посмотрим какая будет получаться RMSE. Если отклонения в RMSE будут в районе 10-100, то можем считать, что НС обучена. Если же RMSE будут совпадать с точностью, то считаем что НС переобучена и уменьшаем количество итераций обучения.

После обучения НС мы сможем взять получившиеся веса как коэффициенты, a в выражении множественной линейной регрессии и использовать их для получения прогноза поступления на следующий календарный период[[14-15](#Нейронки)].

### 1.3.4. ARIMA

Этот алгоритм состоит из 3 компонент: интегрирующая модель (I), авторегрессия (AR) и скользящее среднее (MA).

Сначала к временному ряду применяется интегрирующая модель. Она позволяет из нестационарных рядов получать стационарные. Нестационарные ряды – такие ряды, которые содержат в себе тренд и/или сезонную составляющую. Стационарные ряды таких особенностей не имеют.

Модель I имеет параметр d, которые отвечает за порядок дифференцирования значений временного ряда. Дифференцирование происходит по следующему принципу:

Где:

* – разность порядка 1 для периода t
* – значение ряда за период t

Такое дифференцирование можно делать несколько раз пока не получим стационарный ряд. Разности больших порядков определяются рекуррентно:

Модели AR и MA используют такие разности порядка d и работают следующим образом. Модель AR использует предыдущие значения ряда для получения нового значения:

*,*

где p – число учитываемых предыдущих значений, называется порядком модели AR. Коэффициенты будут находиться по следующему принципу. Пусть в ряду 100 значений. Тогда коэффициенты можно найти следующим образом:

Модель MA использует значения предыдущих ошибок ряда для прогнозирования следующего значения:

где:

* – среднее значение ряда до периода t
* – разница между средним значением ряда и значением за период i (ошибка)
* q – число учитываемых предыдущих значений ошибок

Итоговая модель ARIMA примет вид:

После получения прогнозного значения необходимо сделать преобразование, обратное тому, которое делает интегрирующая модель, и таким образом мы найдем прогноз целевого значения[[16-18](#ARIMA)].

### 1.3.5. SSA

Базовый вариант метода состоит в:

* преобразовании одномерного ряда в многомерный с помощью однопараметрической сдвиговой процедуры (называют “Гусеница”);
* исследовании полученной многомерной траектории с помощью анализа главных компонент (сингулярного разложения);
* восстановлении (аппроксимации) ряда по выбранным главным компонентам.

Таким образом, результатом применения метода является разложение временного ряда на простые компоненты: медленные тренды, сезонные и другие периодические или колебательные составляющие, а также шумовые компоненты. Полученное разложение может служить основой прогнозирования как самого ряда, так и его отдельных составляющих.

Алгоритм допускает естественное обобщение на многомерные временные ряды, а также на случай анализа изображений. Также алгоритм содержит два входных параметра: длину “гусеницы” и число ее компонент, выбор которых существенно влияет на результат работы алгоритма.

Алгоритм работает следующим образом. За основу берется одномерный временной ряд (в нашем случае данные по объему налога или акциза) . Выбирается n – время жизни гусеницы, в диапазоне и вычисляется – длина гусеницы. Затем строится последовательность из n векторов следующего вида:

Обозначим . Будем называть Z нецентрированной матрицей наблюдений, порождённой гусеницей с временем жизни n.

Далее займемся анализом главных компонент ряда. Рассмотрим ковариационную матрицу полученной выборки: . Выполним ее svd-разложение: , где – диагональная матрица собственных чисел, и – ортогональная матрица собственных векторов.

Главные компоненты представляют собой функции, из которых состоит рассматриваемый ряд, и выглядят следующим образом: . После проведения анализа главных компонент выбираются несколько из них и используются для восстановления исходной матрицы наблюдений по этому поднабору главных компонент, т. е. для и вычисляется . Далее восстанавливаются исходные последовательности. В одномерном случае i-ая компонента восстановленного ряда есть среднее значение по i-ой диагонали восстановленной матрицы наблюдений Z'.

Как строится прогноз? Определим:

и

Также положим:

Тогда прогнозное значение ряда в точке N+1 вычисляется по формуле:

Преимущество метода состоит в том, что он может выявлять сразу несколько относительно простых составляющих, которые влияют на итоговое значение временного ряда, что позволяет получать более долгосрочные и точные прогнозы на основе имеющихся данных[[19-20](#SSA)].

### 1.3.6. Выбор алгоритмов прогнозирования

Для прогнозирования налоговых поступлений и акцизов будем использовать все алгоритмы, кроме экспоненциального сглаживания, так как экспоненциальное сглаживание хорошо работает только на стационарных рядах, а все ряды, которые мы будем исследовать в данной работе, стационарными не являются. Также стоит отметить, что для некоторых поступлений не будет использоваться множественная линейная регрессия ввиду недостаточного количества данных по факторам, влияющих на объем поступления.

# 1.4. Анализ метрик для временных рядов

Для оценки точности и качества прогнозов будут использоваться метрики соответствия временных рядов. К таким метрикам относятся:

* Корень из среднеквадратичной ошибки (RMSE). Классическая метрика сравнения двух временных рядов. Быстро возрастает при минимальных несоответствиях в прогнозе, что позволяет быстро заметить отклонения. Вычисляется по формуле:

где – реальное значение ряда, – спрогнозированное значение ряда

* Алгоритм динамической трансформации (ADT)[[21-22](#АДТ)]. Позволяет определять соответствие рядов с учетом их возможного сдвига по оси времени или оси значений. Значение соответствие вычисляется по алгоритму:
  1. Берутся два временных ряда Q длины n и C длины m
  2. Строится матрица расстояний d порядка , где
  3. Используя матрицу расстояний, строится матрица трансформации D, элементы которой вычисляются по формуле
  4. Строится некоторый оптимальный путь W, берущий начало в и заканчивающийся в
  5. Ищется такой путь, чтобы расстояние трансформации было минимальным – .
* Средняя абсолютная масштабированная ошибка (MASE)[[23](#MASE)]. Создана для того, чтобы убрать зависимость величины ошибки от значений сравниваемых рядов. Вычисляется по формуле:

где – реальное значение ряда, – спрогнозированное значение ряда.

Эти метрики будут использоваться для вычисления точности прогнозирования, а также для выбора наилучшего алгоритма прогнозирования для каждого из налогов.

# 2. Проектирование и разработка системы прогнозирования

В этой главе займемся разработкой системы. Система будет состоять из трёх компонент:

* **База данных (БД)** – будет хранить в себе все необходимые данные по налогам, акцизам и факторам, а также результаты прогнозов, вычисленные метрики и параметры.
* **Бэкенд** – будет взаимодействовать с БД и выполнять прогнозы с помощью реализованных алгоритмов.
* **Фронтенд** – графическая система для более удобного взаимодействия с бэкендом.

Вся система будет запускаться в окружении Docker в виде трёх контейнеров. Настройка контейнеров будет осуществляться в автоматическом режиме во время их сборки и запуска. За счет этого достигается простота развертывания системы на компьютере конечного пользователя. Далее будет описаны реализации каждой из компонент.

## 2.1. Проектирование базы данных

База данных будет работать на PostgreSQL и должна отвечать следующим требованиям:

1. Хранить данные по налогам, акцизам и факторам в удобном формате.
2. Хранить результаты прогнозов в удобном формате.
3. Хранить связанные с прогнозами данные: метрики, параметры.

Для достижения этих целей была спроектирована следующая база данных, представленная на рис. 5:

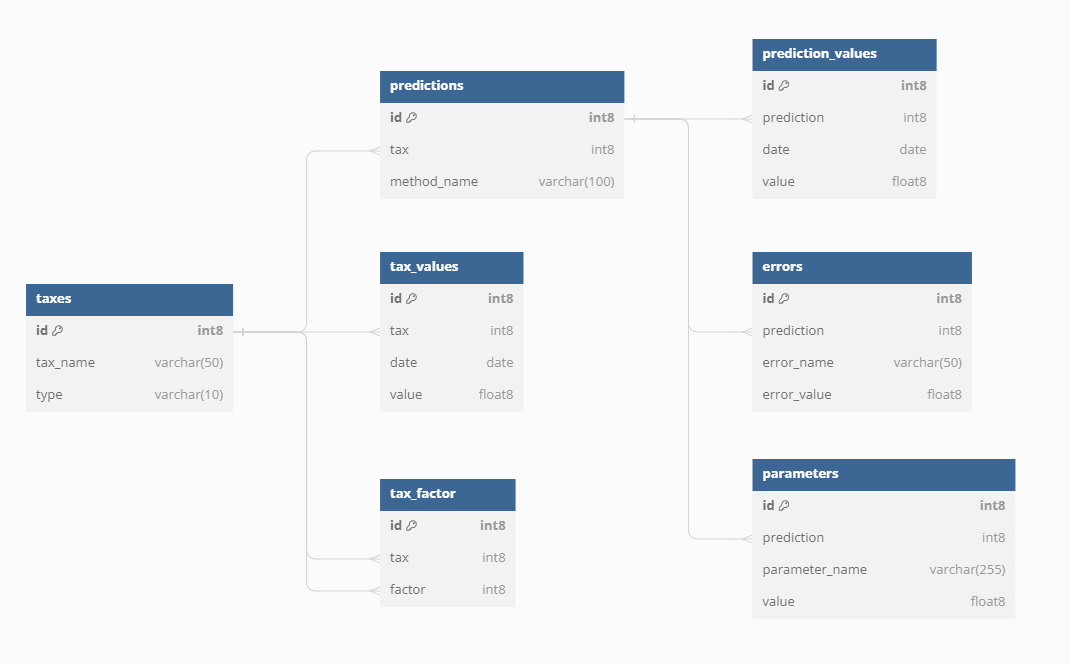


Рисунок 5 – Структура базы данных

В этой структуре главной таблицей является таблица taxes, описывающая имеющиеся в системе налоги и их тип – налог или фактор для налога. Тип влияет на то, как система будет взаимодействовать с этим налогом – либо позволять пользователю выполнять для этого налога прогнозы или использовать как фактор, либо использовать этот налог только как фактор, влияющий на какой-то другой налог. От этой таблицы зависят таблицы со значениями этих налогов, таблица связей налогов с факторами и таблица прогнозов.

В то же время, от таблицы прогнозов predictions зависят таблицы со значениями прогнозов, таблица с метриками прогнозов и таблица с параметрами прогнозов. Такая структура базы данных позволит экономно хранить прогнозы с большим количеством значений за счет того, что строки с названиями хранятся в отдельных таблицах и не дублируются для каждой записи.

Создание описанной структуры будет происходит с помощью бэкенда, поэтому более подробно об этом будет рассказано в подглаве про бэкенд.

## 2.2. Проектирование и реализация бэкенда

Бэкенд будет разрабатываться на языке Java с использованием фреймворка Spring Boot, так как этот фреймворк значительно упрощает создание web-приложений[[24](#SpringBOOT)]. Web-приложение позволяет осуществлять взаимодействие посредством HTTP запросов с помощью заранее описанного API. Эта возможность добавляет большую гибкость системе, так как конечный пользователь, при желании, сможет реализовать свой фронтенд или внедрить бэкенд составляющую в уже существующую систему.

Опишем требования к возможностям бэкенда:

1. Выполнять запрос на прогнозирование.
2. Выполнять запрос на получение всех прогнозов по налогу.
3. Добавлять, изменять, просматривать и удалять данные по налогам и факторам.
4. Получать параметры, с помощью которых был получен тот или иной прогноз.
5. Получать метрики, описывающие качество прогноза.

Для достижения этих целей реализуем следующую внутреннюю архитектуру приложения:

* **Бизнес-слой** – этот слой отвечает за прием запросов от пользователей. Состоит из нескольких контроллеров, в которых описаны точки доступа приложения, а также из классов, которые контроллеры принимают от пользователя или отправляют в ответ на запросы. Получив запрос, перенаправляет его в сервисы в дата-слое или слое вычислений, а результат запросов во внутренние сервисы возвращает пользователю, предварительно преобразовав в удобные для конечного пользователя сущности.
* **Дата-слой** – слой, отвечающий за взаимодействие с базой данных и за агрегацию данных в удобные для бэкенда сущности. Состоит из сервисов, к которым приходят запросы от других сервисов и которые возвращают удобные для бэкенда сущности, и из репозиториев, через которые осуществляются запросы к базе данных.
* **Слой вычислений** – слой, отвечающий за применение реализованных алгоритмов и вычисление метрик. Состоит из сервиса, к которому приходят запросы на прогнозирование, реализованных алгоритмов и реализованных метрик. Реализации алгоритмов и метрик должны реализовывать определенные интерфейсы, и за счет этого достигается простота расширения функционала: достаточно написать новый класс, который реализует нужный интерфейс. Под реализацией интерфейса подразумевается реализация методов, описанных интерфейсом.
* **Слой конфигурации** – слой, отвечающий за различные глобальные настройки бэкенда: адрес базы данных, имя пользователя и пароль для базы данных, формат описания API и так далее.
* **Слой конфигурации базы данных** – слой, отвечающий за поддержание необходимой структуры и минимального наполнения базы данных в момент запуска бэкенда. Необходим для того, чтобы дать бэкенду гарантию того, что структура базы данных будет именно такой, как мы ожидаем.

### 2.2.1. Описание реализации дата-слоя

Главными слоями в бэкенде являются дата-слой и слой вычислений, так как в них хранятся и обрабатываются самые важные части бэкенда.

Минимально рабочая структура дата-слоя, отвечающая требованиями, такова:

* Сущности
  + Сущности для налогов
  + Сущности для прогнозов
* Мапперы
* Репозитории
  + Репозитории для налогов
  + Репозитории для прогнозов
* Представления
  + Представления налогов
  + Представления прогнозов
* Сервисы
  + Общий data сервис
  + Сервис для налогов
  + Сервис для прогнозов

Сущности описывают каждую из реализованных таблиц для их использования в системе. Изменения этих сущностей отражаются соответствуют изменениям соответствующих записей в связанной таблице. Репозитории позволяют получать из таблиц данные и записывать значения нужных записей в сущности.

Сервисы для налогов и прогнозов используют эти репозитории, чтобы получать доступ к таблицам базы данных, а затем с помощью мапперов преобразовывают полученные данные в представления для дальнейшего использования другими слоями бэкенда. Преобразовывание в представления необходимо для того, чтобы уйти от записей таблиц и работать уже с классами Java.

Общий data сервис нужен для того, чтобы агрегировать данные из одного или двух сервисов в какой-то один набор данных для удобного использования этих данных в других слоях. Через него проходят все запросы к дата-слою от других слоев.

### 2.2.2. Описание реализации слоя вычислений

Слой вычислений должен принимать запросы на вычисление прогнозов, во время которых будут вычисляться прогнозы налогов, метрики прогнозов, а также параметры, если таковые реализованы в алгоритме.

Структура слоя вычислений:

* Алгоритмы
  + Интерфейс для алгоритма прогнозирования
  + Интерфейс для параметров алгоритма прогнозирования
  + Класс для хранения значений для прогнозирования
  + Папка с реализованными алгоритмами
* Метрики
  + Интерфейс для метрики прогнозирования
  + Папка с реализованными метриками
* Классы для агрегирования результата прогноза
* Сервис для управления реализованными алгоритмами и метриками (или сервис управления реализованными сущностями)
* Сервис прогнозирования

Взаимодействие с этим слоем осуществляется только через сервис прогнозирования. При получении запроса на прогноз, этот сервис использует data сервис из дата-слоя, чтобы получить данные по налогу и данные по связанным факторам, затем обращается к сервису управления реализованными сущностями, чтобы найти необходимый алгоритм и выполнить прогноз налога.

После выполнения прогноза, сервис прогнозирования вновь обращается к сервису управления реализованными сущностями, чтобы получить все реализованные метрики и вычислить метрики для нового прогноза. Затем из data сервиса он получает уже выполненные прогнозы для данного налога, чтобы определить, насколько новый прогноз лучше или хуже уже имеющихся. Затем все полученные данные упаковываются в классы для агрегирования результатов прогноза и возвращаются тому, кто запрашивал прогноз.

Когда пользователю возвращается прогноз, он не записывается сразу в базу данных. Это сделано для того, чтобы в случае, когда пользователю по каким-либо причинам не понравится получившийся прогноз, предыдущий прогноз, который устраивал пользователя, не затерялся.

Теперь расскажем про внедрение в систему новых алгоритмов и метрик. Чтобы реализовать новый алгоритм прогнозирования, необходимо создать класс, который реализует интерфейс для алгоритма прогнозирования, а также содержит внутри себя класс, который реализует интерфейс для параметров алгоритма прогнозирования. После создания класса его необходимо поместить в папку с реализованными алгоритмами, и система уже при следующем запуске будет знать про новый алгоритм и использовать его.

Интерфейс для алгоритма прогнозирования состоит из следующих функций:

* **makePrediction** – выполняет прогноз с использованием данных по налогу и связанными с этим налогом факторами;
* **getName** – позволяет получить названия алгоритма, которое будет использовать пользователь;
* **getParameters** – позволяет получить список параметров, необходимых алгоритму для получения прогноза. Возвращает класс, который реализует интерфейс для параметров алгоритма прогнозирования;
* **getDescription** – позволяет получить описание алгоритма, чтобы помочь пользователю понять что именно делает алгоритм или чтобы обьяснить смысл параметров;
* **getPredictionParameters** – возвращает список параметров, которые алгоритм получил при выполнении прогноза и которые несут смысловую нагрузку для пользователя.

Интерфейс для параметров алгоритма прогнозирования состоит из следующих функций:

* **getParametersNames** – возвращает список параметров, которые необходимо передать алгоритму для выполнения прогноза. Строка с названием параметра также содержит в себе описание возможных значений параметра;
* **parseParameters** – позволяет разобрать пришедшие строки на нужные значение параметров. Параметры передаются в виде строки, так как при выполнении запроса на прогноз можно указать сразу несколько значений параметра, чтобы алгоритм нашел наилучшее значение из перечисленных. Обычно несколько значений перечисляются через точку с запятой;
* **getParameterValues** – используется алгоритмом прогнозирования, чтобы получить переданные пользователем параметры.

Чтобы реализовать новую метрику, необходимо создать класс, который реализует интерфейс для метрики прогнозирования и поместить его в папку с реализованными метриками. Так же, как и с новым алгоритмом, при следующем запуске система начнет использовать новую метрику при выполнении прогнозов.

Интерфейс для метрики прогнозирования состоит из следующих функций:

* **calcError** – вычисляет значение метрики по реальным и спрогнозированным значениям налога;
* **getName** – позволяет получить название метрики, которое будет видеть пользователь в результате прогноза;
* **compareTo** – позволяет сравнить две метрики и узнать какая из них лучше.

### 2.2.3. Описание реализации других слоев

Другие слои бэкенда являются вспомогательными, поэтому их опишем кратко.

**Бизнес-слой** отвечает за прием запросов от клиента или фронтенда. Из него происходят обращения к дата-слою и слою прогнозирования. Состоит из двух контроллеров, первый принимает запросы на операции с данными по налогам, второй принимает запросы на операции с прогнозами. Перечислим все доступные точки доступа, а также опишем их:

1. /prediction/predict – выполняет прогноз для выбранного налога с использованием выбранного алгоритма и необходимых параметров;
2. /prediction/predicts/get – возвращает все прогнозы для указанного налога;
3. /prediction/algorithms – возвращает список реализованных алгоритмов с описанием и списком необходимых параметров;
4. /prediction/confirm – используется для подтверждения или отклонения прогноза;
5. /taxes/values – позволяет получить список значений для налога или фактора;
6. /taxes/tax/names – позволяет получить список налогов;
7. /taxes/factor/names – позволяет получить список факторов;
8. /taxes/tax/factors – позволяет получить список факторов, связанных с указанным налогом;
9. /taxes/save – позволяет сохранить изменение данных по налогу или фактору;
10. /taxes/tax/factor/add – позволяет добавить связь налога с фактором;
11. /taxes/tax/add – позволяет добавить новые данные по налогу или фактору;
12. /taxes/add/file – позволяет добавить новые данные по налогу или фактору с помощью переданного файла в формате .csv;
13. /taxes/delete – позволяет удалить данные по налогу или фактору;
14. /taxes/tax/factor/delete – позволяет удалить связь фактора с налогом.

Полное описание каждой из точек доступа (API) можно найти по адресу http://\*адрес компьютера, на котором запущен бэкенд\*:8080/swagger-ui/index.html#/. По этому адресу можно найти подробное описание каждой точки доступа, ее входные параметры, а также возвращаемые значения с примерами. Пример описания точки доступа для получения выполненных прогнозов для налога представлен на рис. 6:

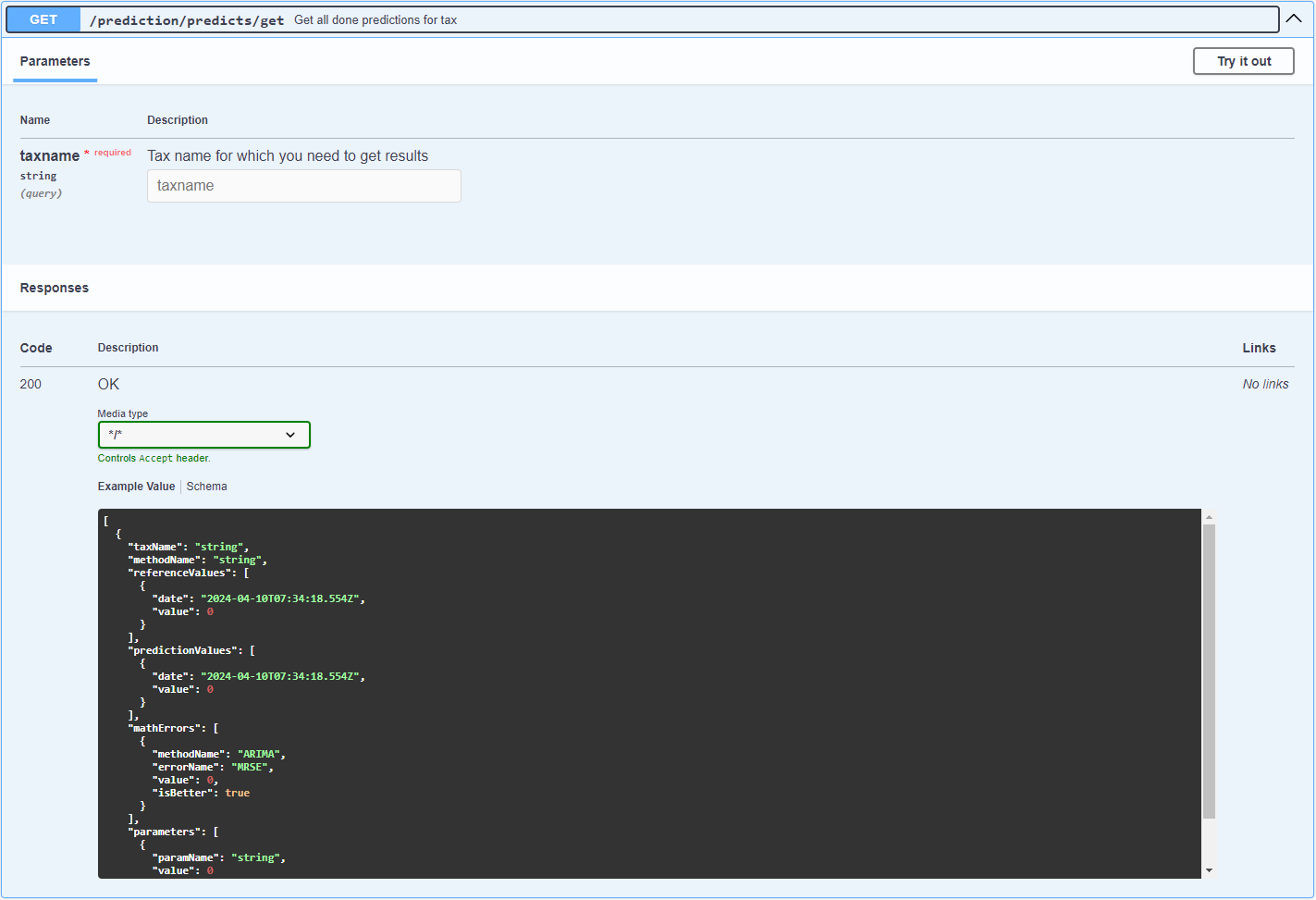


Рисунок 6 – Пример описания точки доступа

**Слой конфигурации** отвечает за подключение к базе данных, за настройку API. Состоит из нескольких классов-конфигураций и глобального класса настроек.

**Слой конфигурации базы данных** использует библиотеку Liquibase, которая записывает в базу данных изменения, описанные в специальном xml-документе. Пример описания двух таблиц в этом файле с указанием зависимости одной таблицы от другой представлен на рис. 7:



Рисунок 7 – XML-документ для создания двух таблиц в базе данных при запуске бэкенда

## 2.3. Проектирование и реализация фронтенда

Фронтенд будет разрабатываться на языке TypeScript с использованием фреймворка React[[25](#React)]. Такое сочетание позволяет создавать сайты с динамическим наполнением страниц, что облегчает разработку.

Фронтенд (или по-другому – сайт) должен позволять конечному пользователю взаимодействовать с системой через удобный интерфейс, а не через API, что ускорит получение прогноза и операции с данными, а также уменьшит риск отправки некорректных данных, что тоже ускорит взаимодействие с приложением.

Фронтенд должен содержать следующие страницы:

* **Главная страница –** страница, которая будет встречать пользователя при первом заходе на сайт.
* **Страница выполнения прогноза** – на этой странице можно будет выбрать налог и алгоритм, с помощью которого пользователь будет выполнять прогноз. Когда будет выбран алгоритм прогнозирования, на странице появится описание алгоритма, а также список необходимых параметров с полями для их ввода.
* **Страница просмотра выполненных прогнозов** – на этой странице можно будет выбрать налог, а также алгоритм из списка тех, с помощью которых уже выполнялся прогноз для этого алгоритма. После этого на странице появится информация, связанная с этим прогнозом, а также графики и сводки, которые помогут лучше понять качество прогноза.
* **Страница для просмотра и изменения данных по налогам** – на этой странице можно будет: посмотреть и изменить данные по налогам; посмотреть и изменить данные по факторам; посмотреть с какими факторами связан тот или иной налог, добавить или удалить такие связи; добавить данные по налогу или фактору.

Для взаимодействия с бэкендом необходимы функции, которые будут использоваться на страницах для получения или отправки данных, а также сущности, которые будут использовать при получении и отправке данных.

Структура фронтенда:

* Сущности для взаимодействия с бэкендом
* Функции для взаимодействия с бэкендом
* Различные иконки и картинки, необходимые для страниц
* Код страниц

Теперь расскажем про внешний вид каждой из страниц. Внешний вид главной страницы представлен на рис. 8:

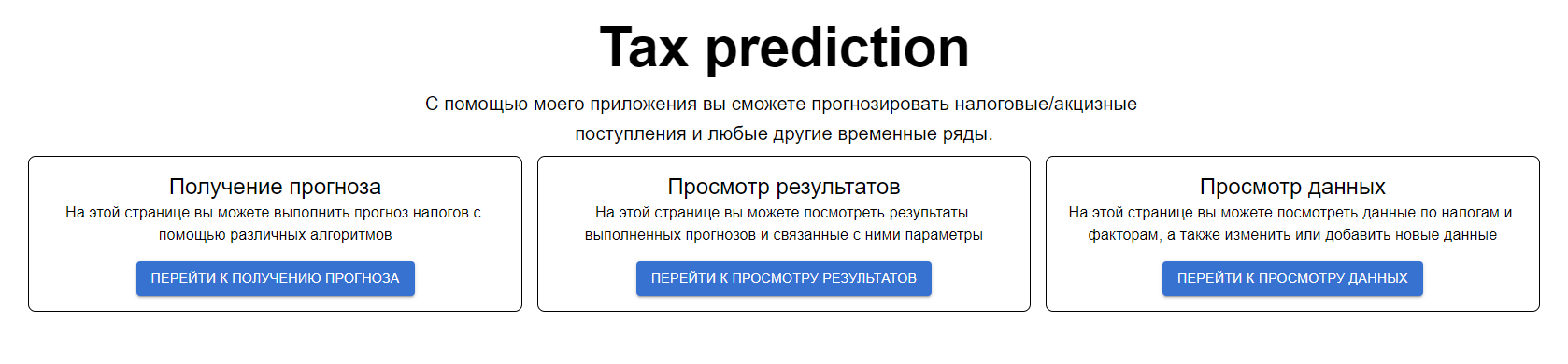


Рисунок 8 – Вид главной страницы

Здесь находится описание приложения, а также краткое описание возможностей с кнопками перехода на соответствующие страницы.

Внешний вид страницы выполнения прогноза представлен на рис. 9:

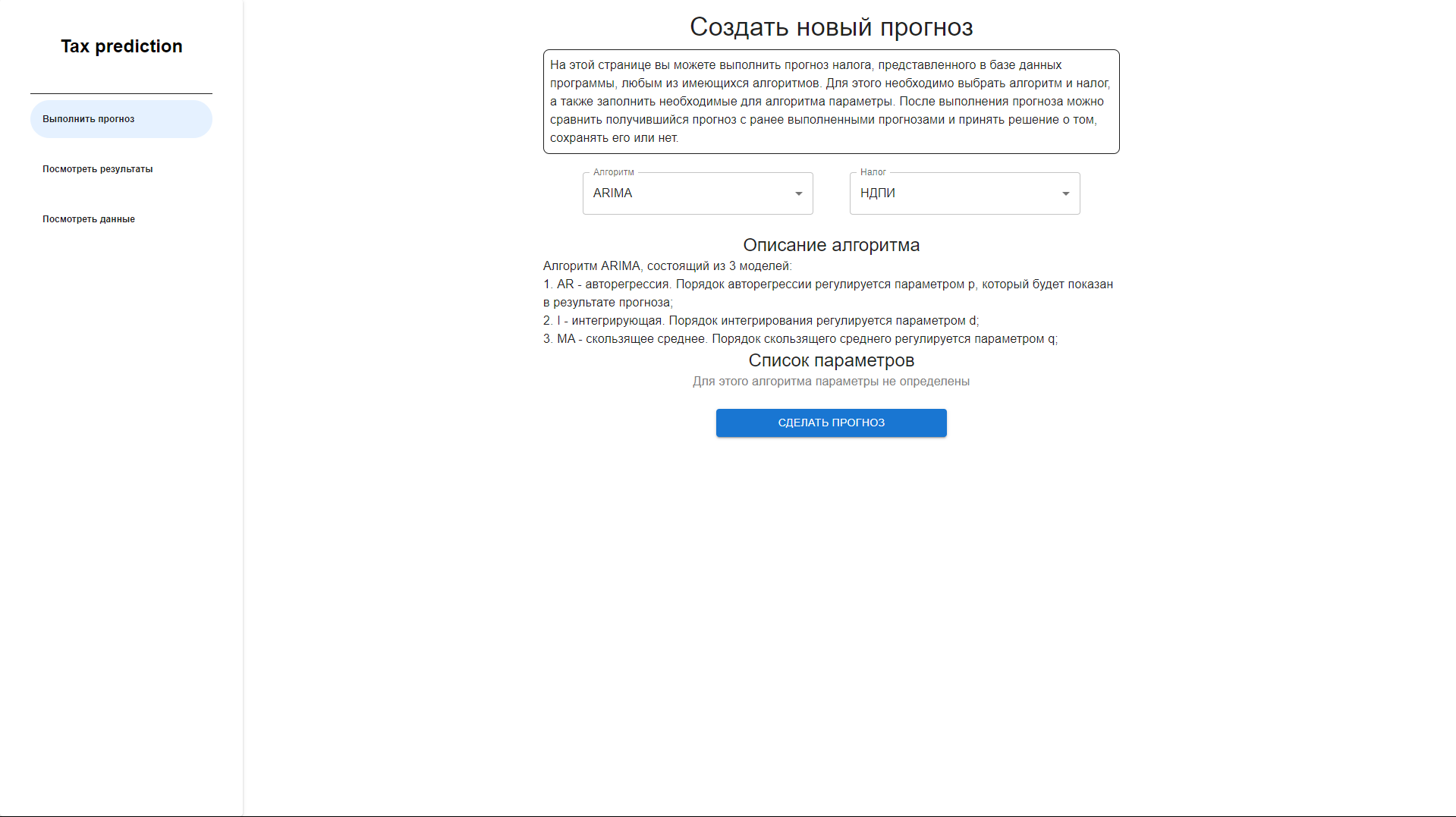


Рисунок 9 – Внешний вид страницы выполнения прогнозов

На этой и последующих страницах слева располагается меню для быстрого переключения между страницами. Помимо этого, на странице есть описание алгоритма получения прогноза. Для получения прогноза необходимо выбрать алгоритм прогнозирования и налог, для которого необходимо выполнить прогноз. После ввода необходимых параметров, если таковые требуются, необходимо нажать кнопку «Сделать прогноз». Спустя какое-то время на экране появится выполненный прогноз вместе с другими прогнозами, если таковые имеются:

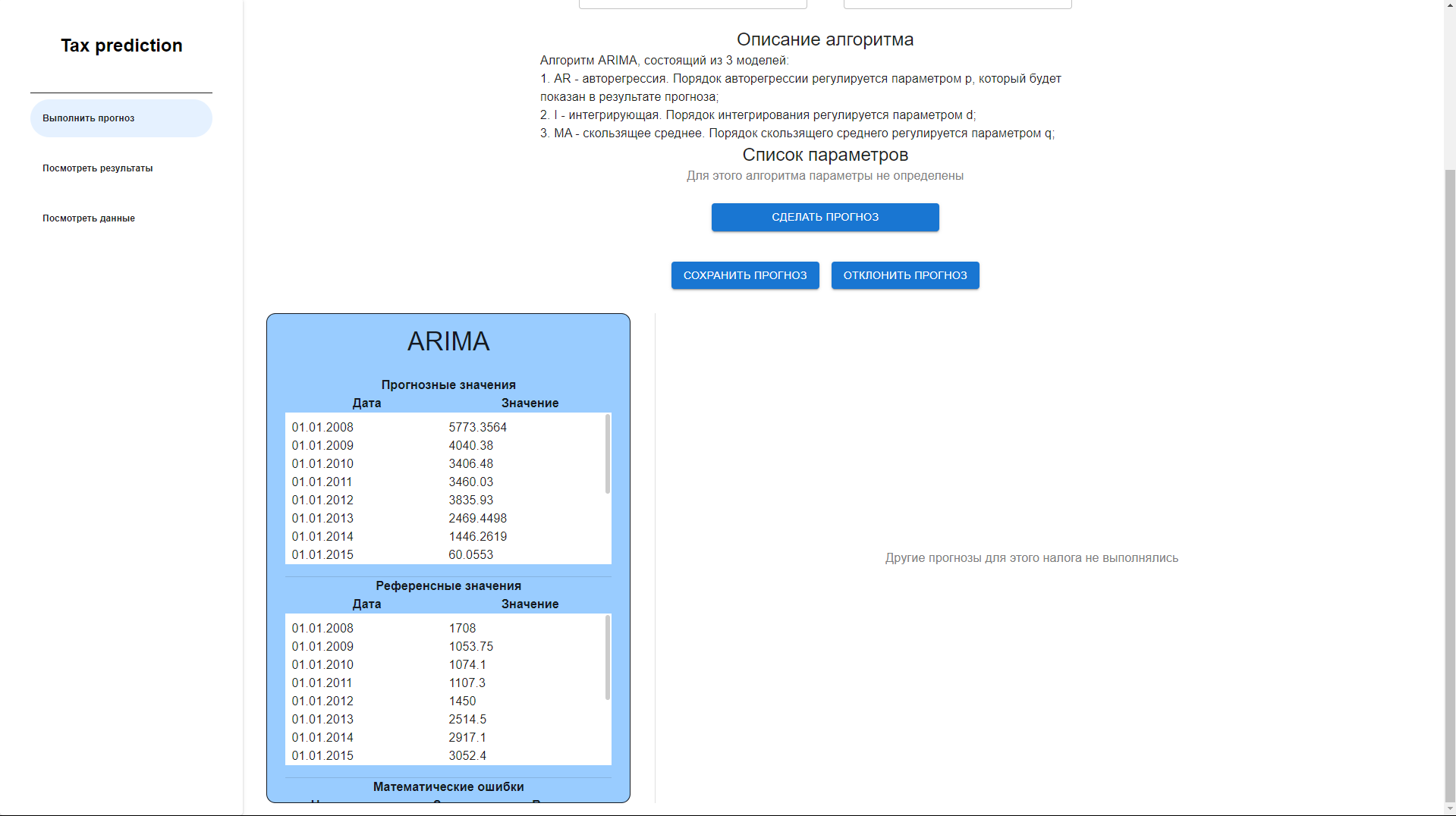


Рисунок 10 – Внешний вид страницы при получении прогноза

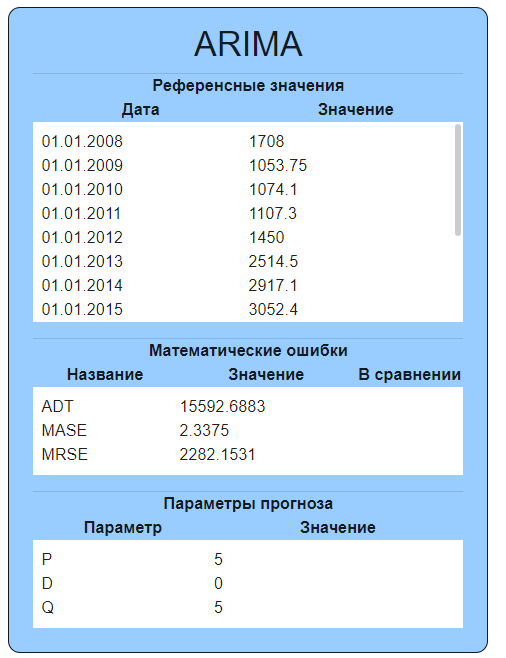
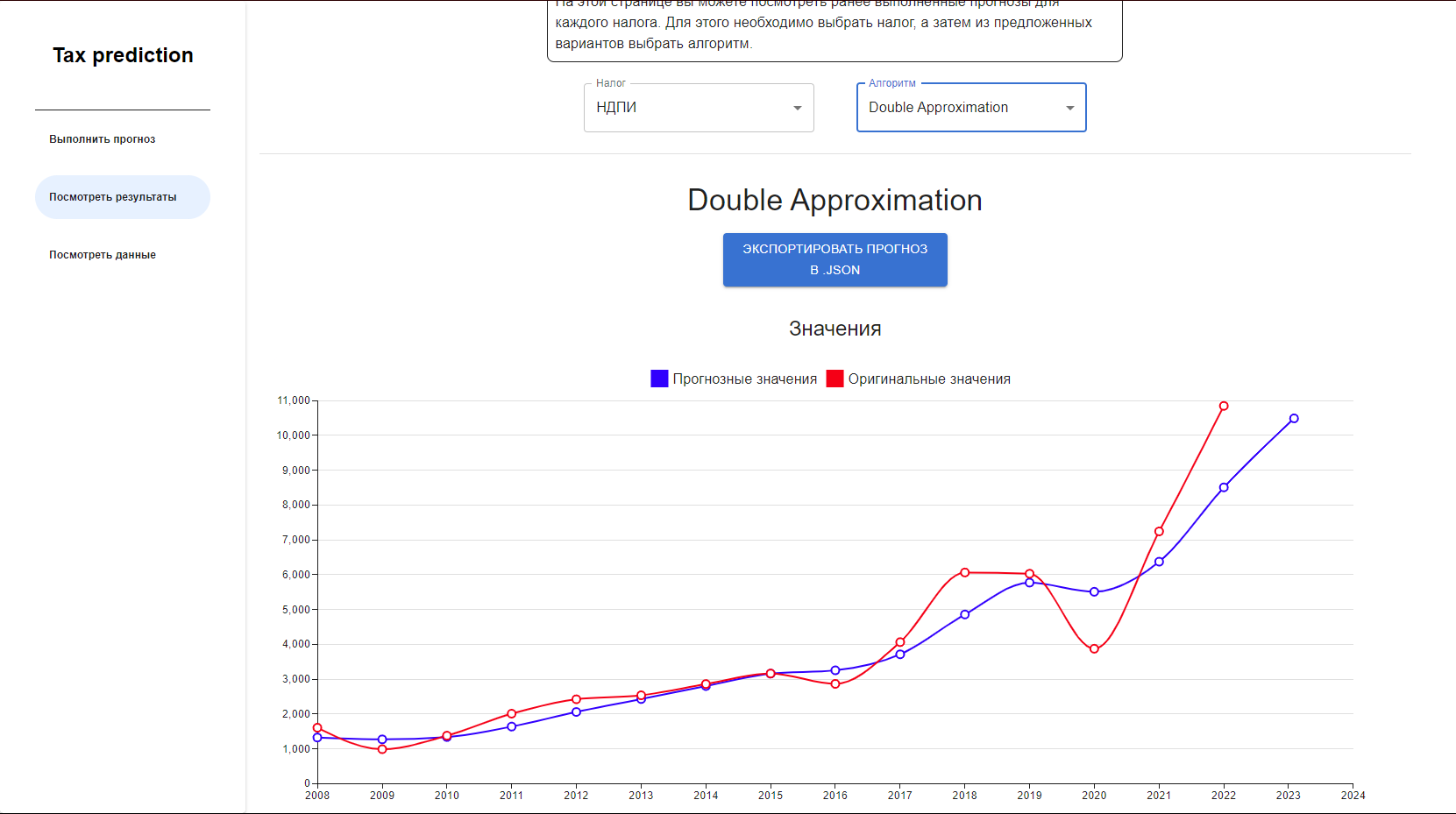


Рисунок 11 – Внешний вид одного из прогнозов

Возвращаемые результаты прогнозов показываются в краткой форме, достаточной для того, чтобы оценить его точность по сравнению с другими.

Внешний вид страницы просмотра выполненных прогнозов представлен на рис. 12:



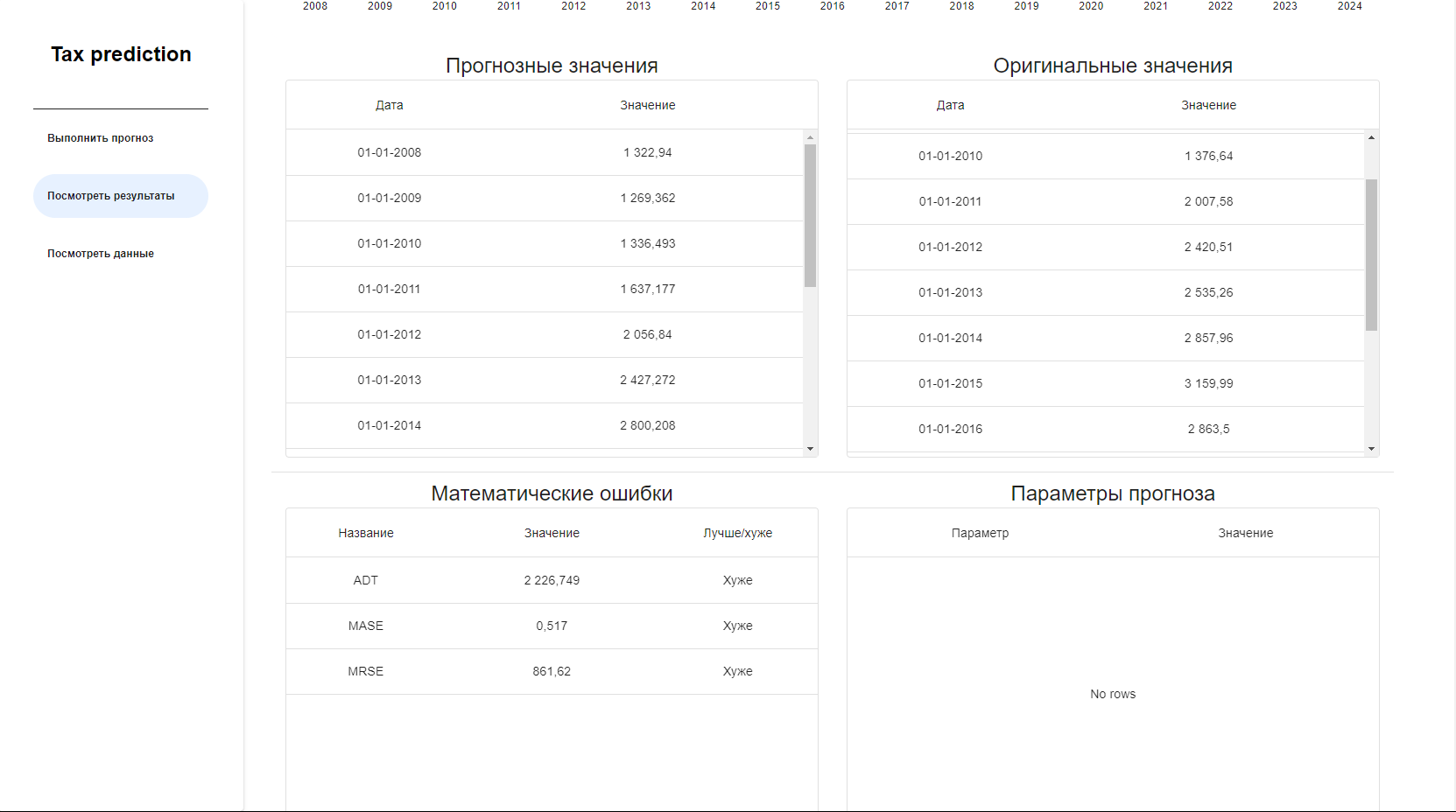


Рисунок 12 – Внешний вид страницы с результатами прогнозов

На этой странице, при выборе налога и алгоритма прогнозирования, показывается инфографика прогноза: график изменения значений, таблицы значений, таблицы параметров и метрик и так далее. На этой странице также есть возможность экспортировать выбранный прогноз в JSON формат, на случай если прогноз необходимо использовать где-то еще. После нажатия на кнопку «Экспортировать прогноз» скачается файл с названием «Названиеналога\_названиеалгоритма\_prediction.json». Выбор формата экспорта обусловлен популярностью формата и почти повсеместным его использованием в других системах.

Перейдем к внешнему виду страницы с данными по налогам и факторам. Данная страница содержит в себе 3 вкладки: для просмотра и изменения данных по налогам, для просмотра и изменения данных по факторам, для добавления новых данных. Внешний вид вкладки с данными по налогам представлен на рис. 13:



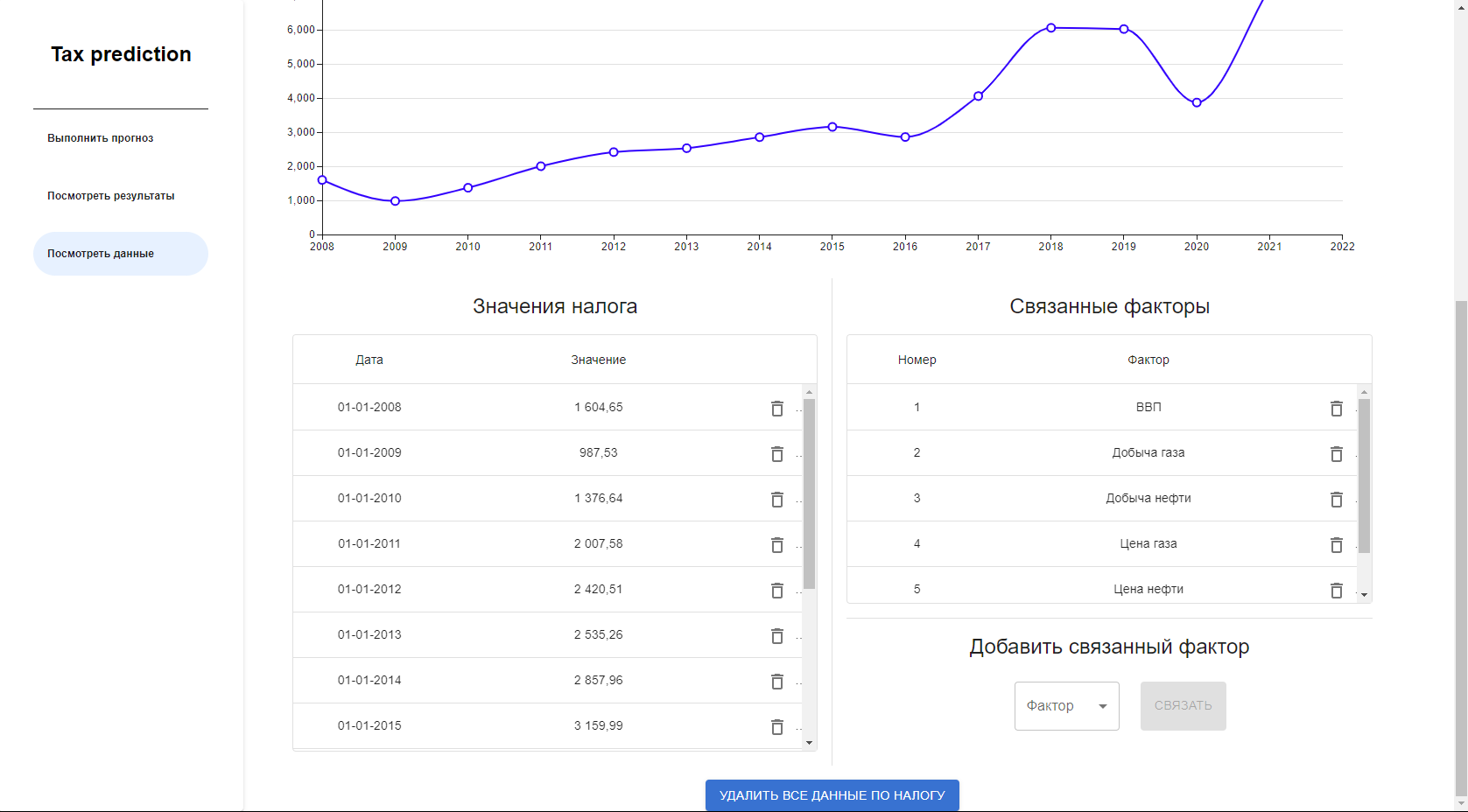
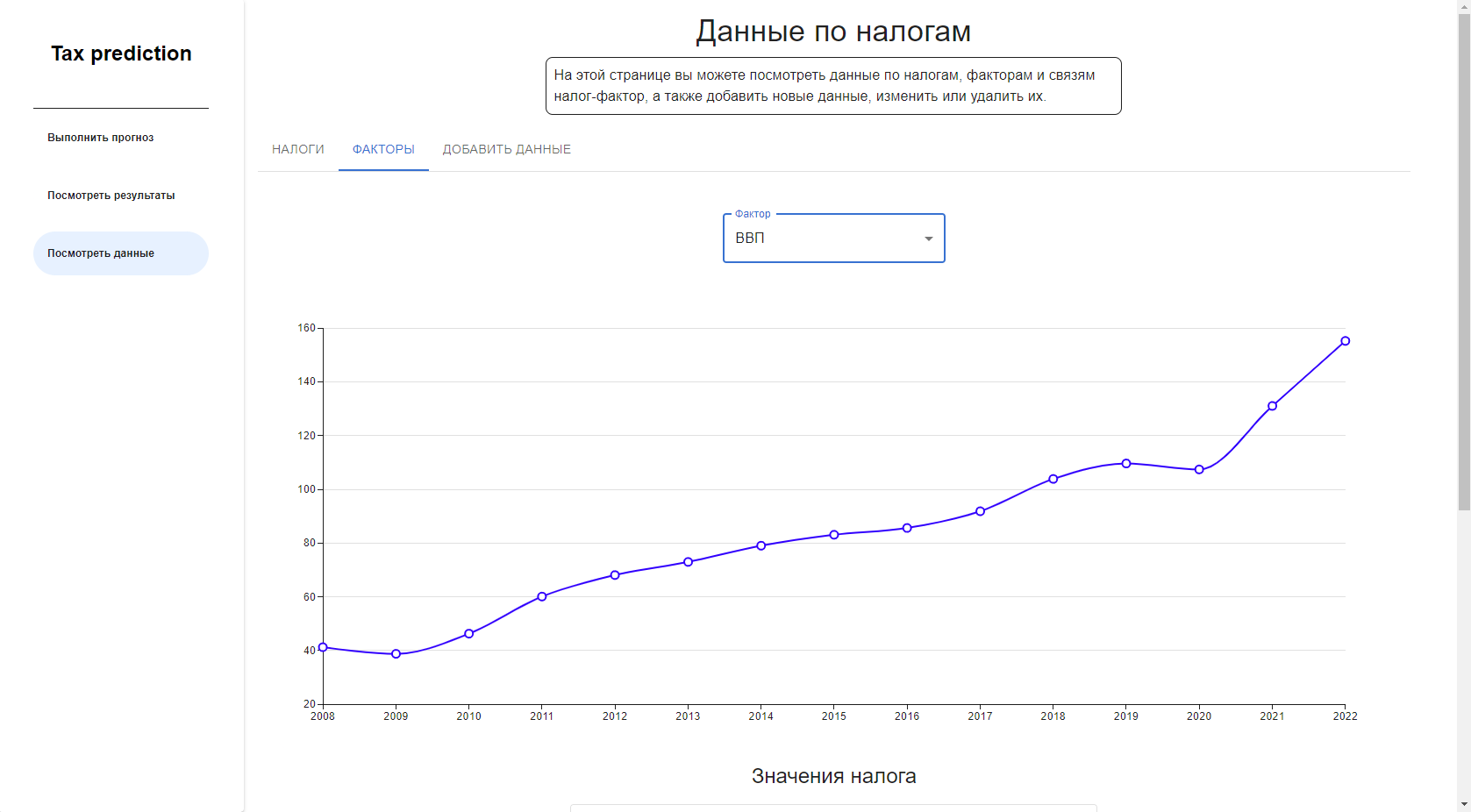


Рисунок 13 – Внешний вид страницы с данными по налогам

На этой вкладке необходимо выбрать налог, данные по которому необходимо посмотреть. После этого появится таблица с данными, в которой можно изменить или удалить данные, а также таблица связей этого налога с факторами.

Внешний вид вкладки с данными по факторам представлен на рис. 14:



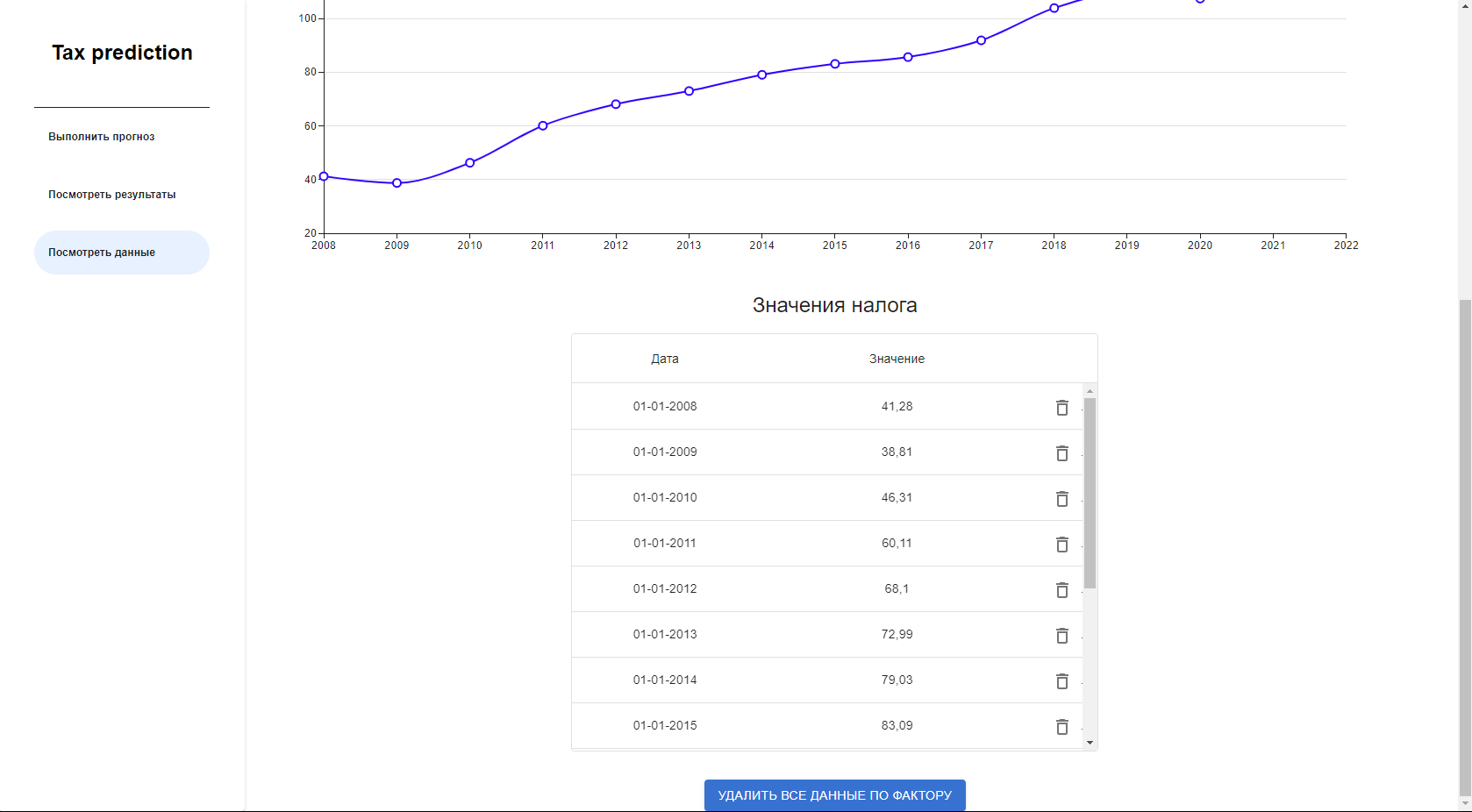


Рисунок 14 – Внешний вид вкладки с данными по факторам

На этой вкладке можно посмотреть, изменить или удалить данные по какому-либо фактору.

Внешний вид вкладки для добавления данных представлен на рис. 15:

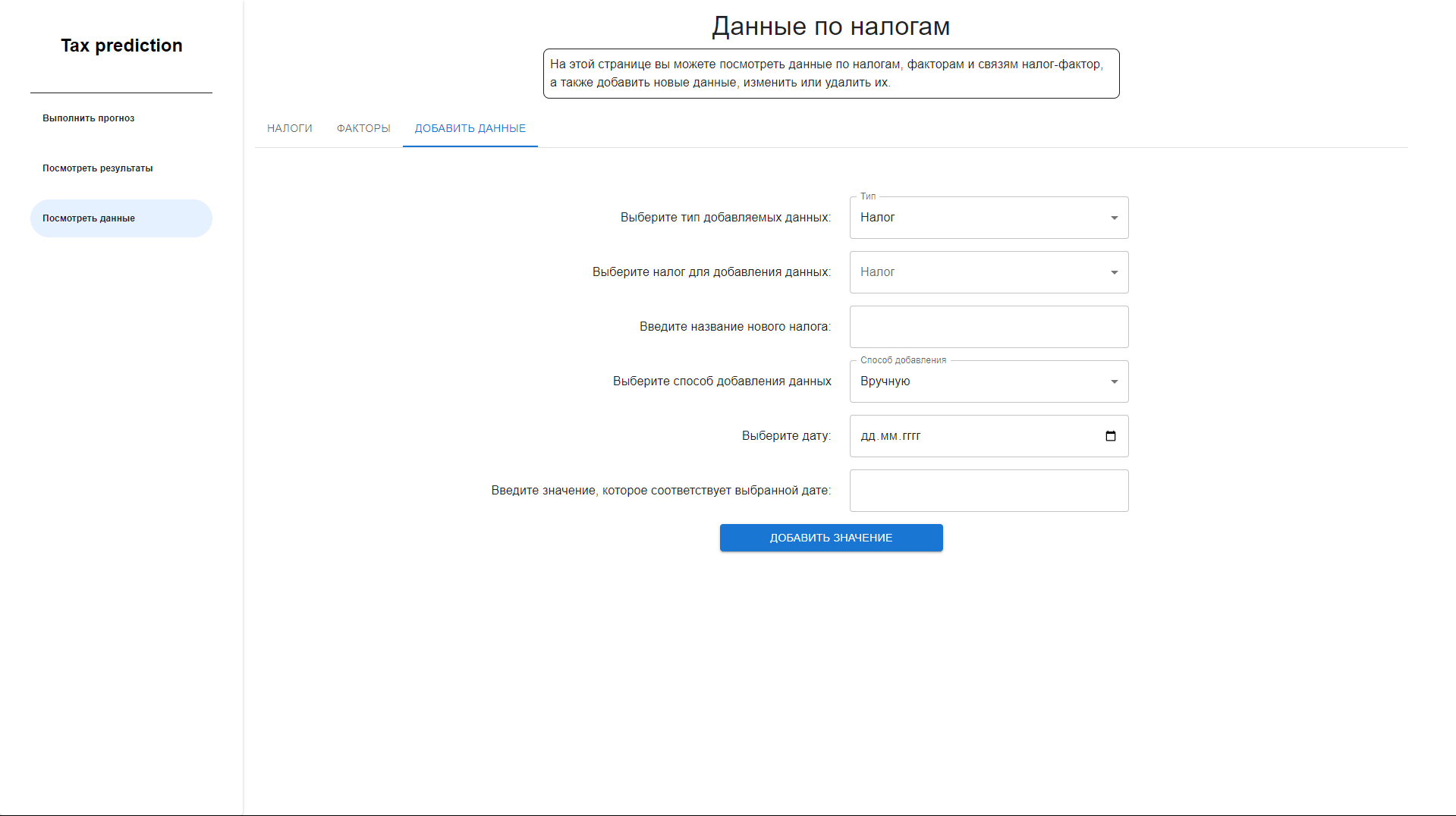


Рисунок 15 – Внешний вид вкладки для добавления данных

Для добавления данных необходимо сначала выбрать тип добавляемых данных. После этого в поле выбора налога/фактора появятся содержащиеся в базе данных налоги, к которым можно добавить новые данные. Также, если выбрать вариант «Новый налог», то появится поле для ввода названия нового налога.

После выбора налога, к которому будут добавляться данные, необходимо выбрать способ добавления данных: вручную или с помощью файла. При вводе данных вручную необходимо выбрать дату и значение, которое соответствует этой дате. При добавлении данных через файл, необходимо будет предоставить файл, который содержит в себе данные в виде множества csv строк формата ‘Дата;Значение’. Ввод данных вручную будет удобен, когда необходимо добавить одно новое значение к какому-либо налогу или фактору, а добавление через файл будет удобно, когда в систему необходимо добавить абсолютно новый налог или фактор, для которого накопилось достаточно много значений.

# 3. Анализ применения алгоритмов

Теперь применим разработанную систему для прогнозирования налогов и акцизов, а также для получения прогноза объёма федерального бюджета.

Для начала, необходимо добавить все необходимые данные в систему[[26-31](#Данные_для_прогнозов)]. После этого, на странице выполнения прогнозов на фронтенде, выполним прогноз каждым из реализованных алгоритмов для каждого из налогов и акцизов, а затем выберем лучший прогноз для каждого из них. Получив все необходимые прогнозы, мы сможем выполнить прогноз федерального бюджета.

Теперь рассмотрим лучший прогноз для каждого из налогов и акцизов, а именно:

1. НДПИ;
2. Налог на прибыль организаций;
3. НДС;
4. Акциз на табачную продукцию;
5. Акциз на табак;
6. Акциз на автомобильный бензин;
7. Акциз на дизельное топливо;
8. Акциз на алкогольная продукция;
9. Акциз на авиационный керосин;
10. Акциз на природный газ;
11. Акциз на нефтяное сырье.

НДФЛ не был включен в список, так как НДФЛ начал поступать в федеральный бюджет только с 2021 года и в очень малых количествах, поэтому его влияние несущественно по сравнению с другими налогами. Основная часть этого налога остается в регионах.

Для всех остальных налогов и акцизов прогноз будет выполняться на 2023 год.

## 3.1. Прогноз для НДПИ

Для НДПИ лучший прогноз получился с использованием алгоритма SSA. График прогноза изображен на рис. 16:



Рисунок 16 – Прогноз НДПИ с помощью SSA

Значение прогноза – 8557,928. Реальное значение – 9701,225. Отклонение – 11,7%.

Метрики прогнозов:

Таблица 1 – Метрики для прогнозов НДПИ

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **SSA** | Двойное сглаживание | Множественная регрессия | ARIMA |
| ADT | 698,57 | 2 226,749 | 2 774,967 | 11 224,078 |
| MASE | 0,061 | 0,517 | 0,521 | 1,833 |
| RMSE | 97,039 | 861,62 | 733,073 | 2475,618 |

Прогноз получился при следующих параметрах:

* Длина гусеницы – 9;
* Количество выбранных главных компонент – первые 4.

## 3.2. Прогноз для налога на прибыль

Для налога на прибыль лучший прогноз получился с использованием алгоритма SSA. График прогноза изображен на рис. 17:

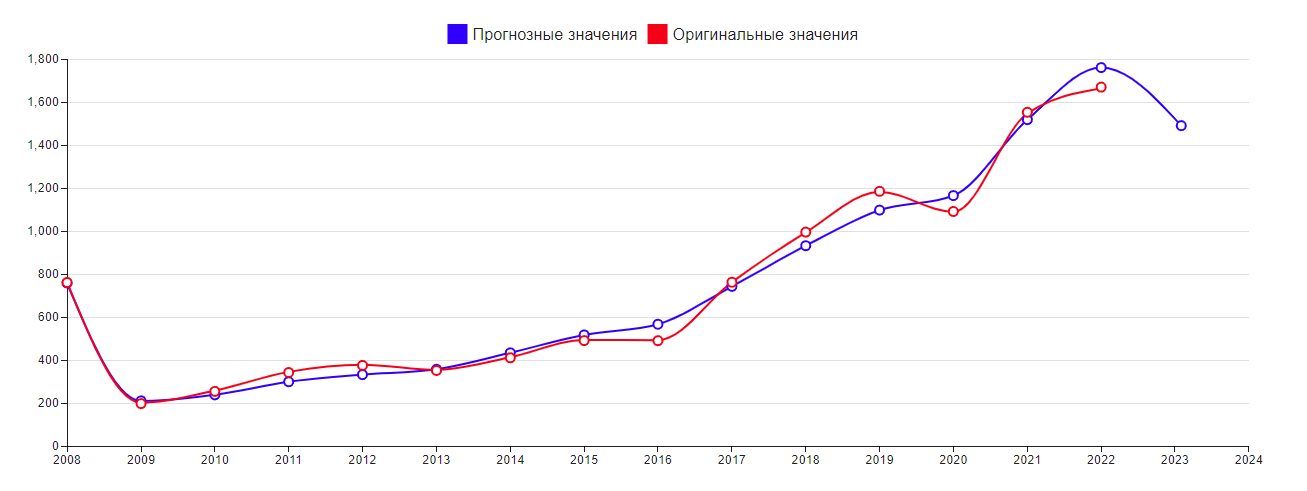


Рисунок 17 – Прогноз налога на прибыль с помощью SSA

Значение прогноза – 1490,27. Реальное значение – 1918,611. Отклонение – 22,3%.

Метрики прогнозов:

Таблица 2 – Метрики для прогнозов налога на прибыль

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **SSA** | Двойное сглаживание | Множественная регрессия | ARIMA |
| ADT | 217,436 | 659,573 | 771,357 | 2 821,177 |
| MASE | 0,253 | 0,697 | 0,649 | 2,913 |
| RMSE | 50,304 | 138,927 | 146,494 | 639,723 |

Прогноз получился при следующих параметрах:

* Длина гусеницы – 9;
* Количество выбранных главных компонент – 3. 1-я, 2-я и 4-я.

## 3.3. Прогноз для НДС

Для НДС лучший прогноз получился с использованием алгоритма SSA. График прогноза изображен на рис. 18:

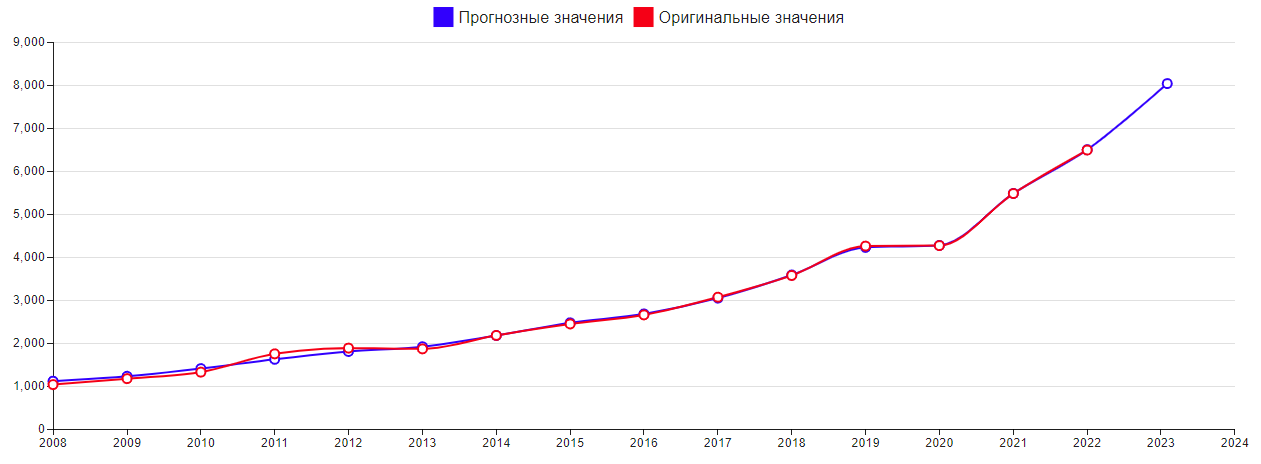


Рисунок 18 – Прогноз НДС с помощью SSA

Значение прогноза – 8034,562. Реальное значение – 7182,379. Отклонение – 11,8%.

Метрики прогнозов:

Таблица 3 – Метрики для прогнозов НДС

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **SSA** | Двойное сглаживание | Множественная регрессия | ARIMA |
| ADT | 441,296 | 797,31 | 1 125,639 | 4 330,35 |
| MASE | 0,102 | 0,432 | 0,638 | 3,674 |
| RMSE | 53,793 | 222,03 | 290,899 | 1 523,188 |

Прогноз получился при следующих параметрах:

* Длина гусеницы – 9;
* Количество выбранных главных компонент – 3. 1-я, 2-я и 4-я.

## 3.4. Прогнозы акцизов

Так как акцизов много, то результаты прогнозов будут представлены в кратком варианте. Для акцизов на табачные изделия, табак, керосин, дизель, бензин и алкоголь не выполнялся алгоритм множественной регрессии в связи с недостатком данных.

Результаты прогнозов:

Таблица 4 – Результаты прогнозов акцизов

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Табак | Табачные изделия | Нефтяное сырье | Керосин | Дизель | Бензин | Газ | Алкоголь |
| Алгоритм | SSA | Двойное сглаживание | SSA | SSA | SSA | SSA | SSA | SSA |
| ADT | 14,2 | 198,421 | 767.5 | 13,092 | 42,067 | 55,232 | 86,85 | 79,462 |
| MASE | 0,442 | 0,635 | 0.186 | 0,394 | 0,251 | 0,284 | 0,464 | 0,735 |
| RMSE | 4,314 | 34,553 | 281.857 | 6,524 | 6,964 | 11,715 | 23,321 | 12,669 |
| Прогнозное значение | 31,167 | 690,6 | -64,143 | -169,424 | 192,839 | 190,804 | 306,631 | 12,201 |
| Реальное значение | 50,314 | 724,778 | -2912,459 | -120,66 | 129,635 | 133,988 | 73,288 | 40,44 |
| Отклонение | 38% | 4.7% | 97.7% | 40.4% | 48.7% | 42.4% | 318% | 69.8% |

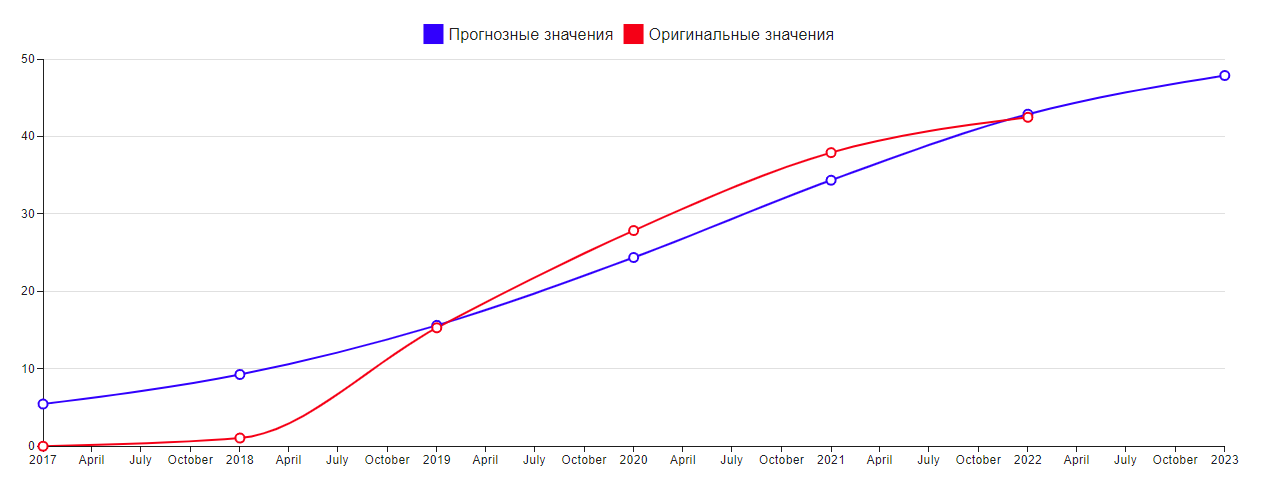
Параметры алгоритмов:

* **Табак**: Длина гусеницы – 4. Выбранные главные компоненты – 1-я;
* **Табачные изделия**: параметры для этого алгоритма не определены;
* **Нефтяное сырье**: Длина гусеницы – 3. Выбранные главные компоненты – 1-я;
* **Керосин**: Длина гусеницы – 3. Выбранные главные компоненты – 1-я;
* **Дизель**: Длина гусеницы – 9. Выбранные главные компоненты – первые 4;
* **Бензин**: Длина гусеницы – 9. Выбранные главные компоненты – первые 4;
* **Газ**: Длина гусеницы – 3. Выбранные главные компоненты – 1-я;
* **Алкоголь**: Длина гусеницы – 6. Выбранные главные компоненты – первые 2.

Большие отклонения прогнозов от реальных значений при прогнозе алгоритмом SSA по большей части обусловлены недостатком данных. Одни из самых низких отклонений получились для акцизов на бензин и дизель, данные по ним есть с той же даты, что и по налогам, с 2008 года, а, например, по налогу на нефтяное сырье данные есть только с 2018. Также стоит отметить, что значения акцизов как временной ряд ведут себя очень нестабильно из-за большого количества факторов, влияющих на объём акциза, а также из-за часто меняющейся ставки акциза. Из-за этого алгоритмам тяжело спрогнозировать новые значения.

Также стоит учитывать, что на момент написания работы федеральная налоговая служба еще уточняет данные по объемам акцизов за 2023 год, поэтому реальные значения могут изменится и прогнозы могут оказаться более точными.

Отмечу некоторые наблюдения по прогнозам. Для акциза на табак прогнозы алгоритмами двойного экспоненциального сглаживания и SSA получились очень близкими, и визуально и по метрикам:



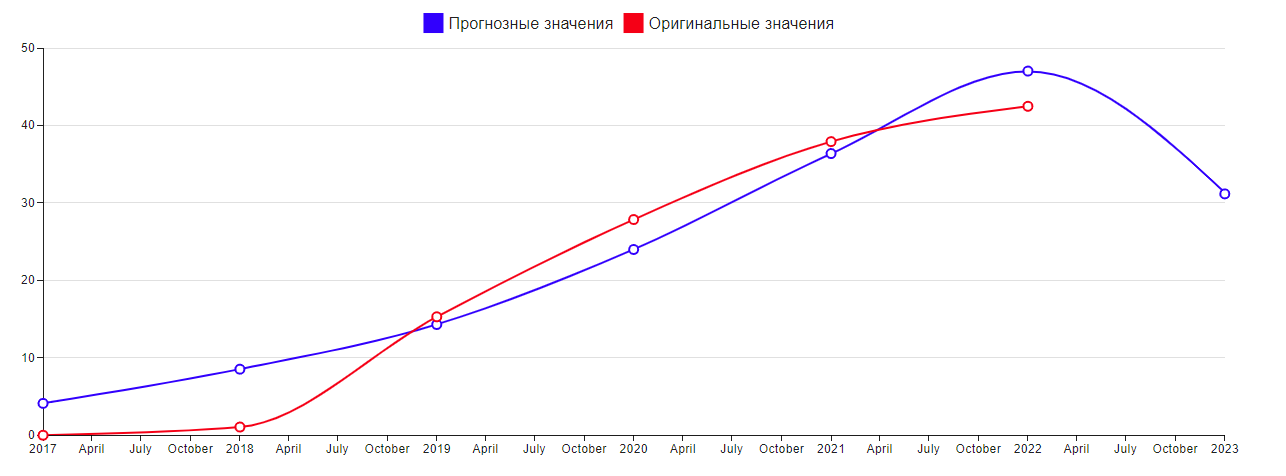


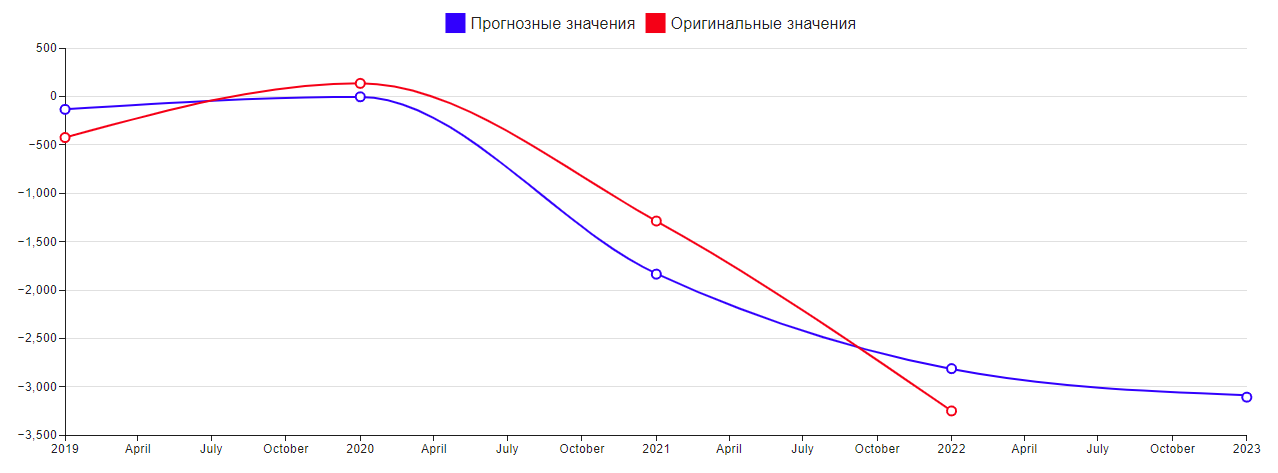
Рисунок 19 – Прогнозы для акциза на табак с помощью двойного сглаживания (выше) и SSA (ниже)

Таблица 5 – Метрики для прогнозов акциза на табак

|  |  |  |
| --- | --- | --- |
|  | Двойное сглаживание | SSA |
| ADT | 15,489 | 14,2 |
| MASE | 0,42 | 0,442 |
| MRSE | 4,511 | 4,314 |

Это говорит о том, что, при текущих данных, прогнозы этими алгоритмами для этого акциза можно считать равноценными, а при добавлении новых данных двойное сглаживание, возможно, даст лучший прогноз, чем SSA.

Похожая ситуация обстоит с акцизом на нефтяное сырье. Прогнозы с помощью множественной регрессии и SSA получились очень близкими по метрикам:



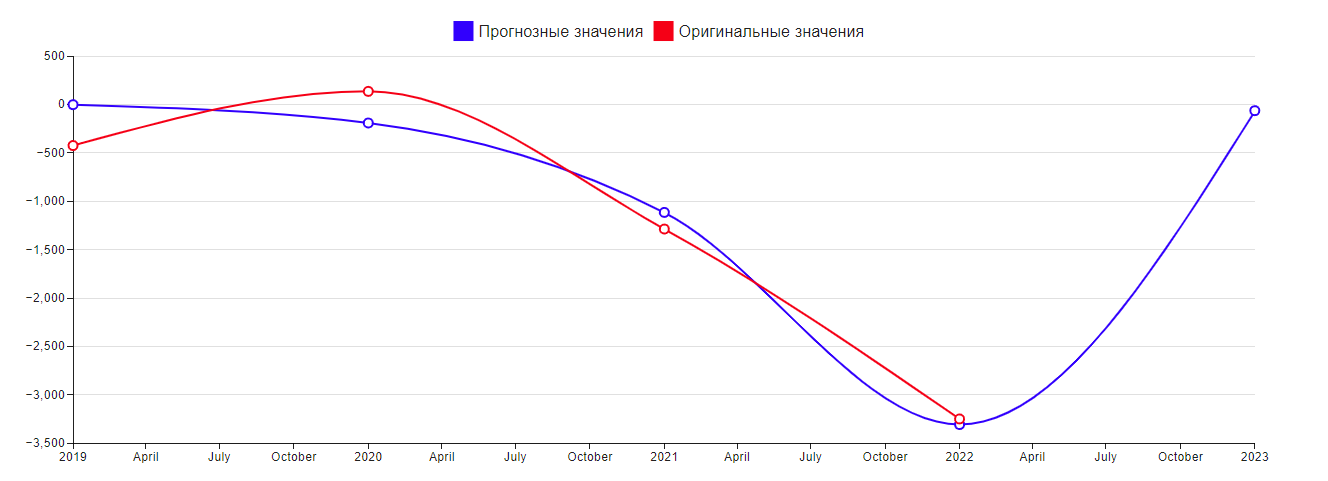


Рисунок 20 – Прогнозы акциза на нефтяное сырье с помощью множественной регрессии (выше) и SSA (ниже)

Таблица 6 – Метрики для прогнозов акциза на нефтяное сырье

|  |  |  |
| --- | --- | --- |
|  | Множественная регрессия | SSA |
| ADT | 778,354 | 767,5 |
| MASE | 0,269 | 0,186 |
| MRSE | 386,001 | 281,857 |

Из этого можно сделать вывод, что, при добавлении новых данных, прогноз с помощью множественной регрессии может дать более точный прогноз.

Для других акцизов подобной близости прогнозов замечено не было.

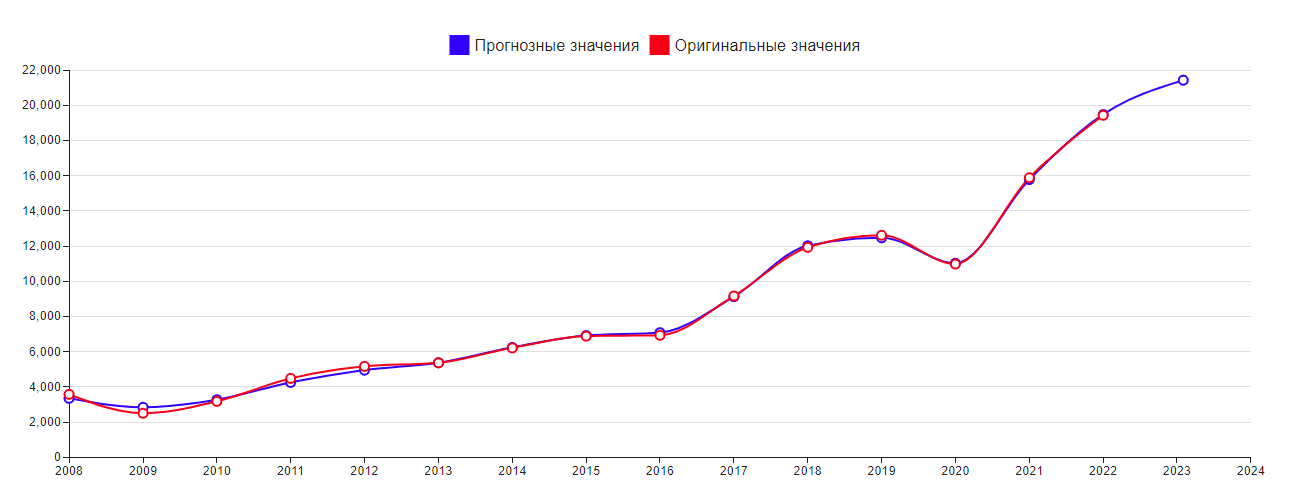
## 3.5. Прогноз объема федерального бюджета

Теперь используем все полученные прогнозы, чтобы получить прогноз федерального бюджета. Рассмотрим два варианта прогнозирования:

1. Выполним прогнозы всеми реализованными в приложении алгоритмами;
2. С помощью алгоритма множественной регрессии вычислим коэффициенты для факторов, возьмем лучшие прогнозы для этих факторов и посчитаем прогноз самостоятельно. Этот прогноз придется выполнять в полуавтоматическом режиме, т.к. реализованные алгоритмы прогнозирования не имеют доступа к уже выполненным прогнозам, а алгоритм множественной регрессии делает прогноз факторов с помощью двойного сглаживания.

### 3.5.1 Прогноз полностью через приложение

Начнем с прогнозов полностью через приложение. Хорошие прогнозы получились с помощью алгоритма SSA и множественной регрессии. Графики прогнозов изображены на рис. 21:



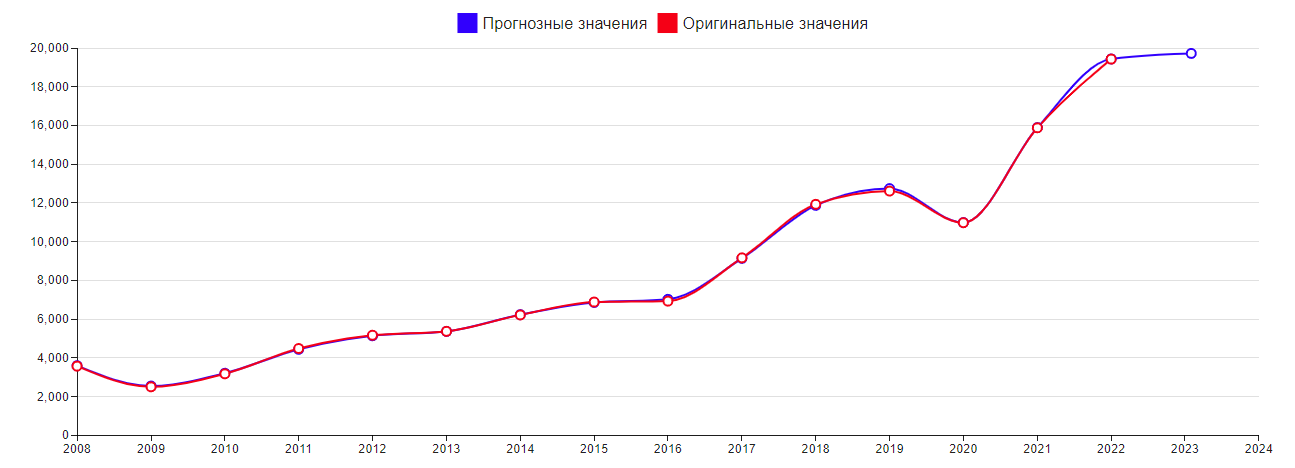


Рисунок 21 – Прогнозы федерального бюджета с помощью SSA (выше) и множественной регрессии (ниже)

Метрики получились следующие:

Таблица 7 – Метрики для прогнозов федерального бюджета

|  |  |  |
| --- | --- | --- |
|  | SSA | Множественная регрессия |
| ADT | 1 178,503 | 266,858 |
| MASE | 0,078 | 0,023 |
| RMSE | 150,944 | 47,922 |

А параметры такие:

* **SSA**: Длина гусеницы – 9. Выбранные главные компоненты – первые 4;
* **Множественная регрессия**: Коэффициент обучения – 0,005. Количество эпох – 500 000. Коэффициенты при факторах получились следующие:

Таблица 8 – Сводка коэффициентов

|  |  |
| --- | --- |
| Фактор | Коэффициент |
| НДПИ | 1,005 |
| НДС | 0,984 |
| Налог на прибыль | 1,042 |
| Акциз на табачную продукцию | 2,308 |
| Акциз на табак | 17,514 |
| Акциз на нефтяное сырье | 0,332 |
| Акциз на керосин | 3,072 |
| Акциз на дизель | -1,112 |
| Акциз на газ | -1,719 |
| Акциз на бензин | 0,775 |
| Акциз на алкоголь | 0,436 |

Полученные коэффициенты будут использованы при полуавтоматическом выполнении прогноза федерального бюджета в следующей главе.

### 3.5.2 Полуавтоматический прогноз

Для того чтобы выполнить прогноз, я возьму коэффициенты факторов из предыдущего подпункта, а также наилучший прогноз для каждого из факторов:

Таблица 9 – Сводка лучших прогнозов налогов и акцизов

|  |  |
| --- | --- |
| Фактор | Лучший прогноз фактора |
| НДПИ | 8 557,928 |
| НДС | 8 034,562 |
| Налог на прибыль | 1 490,27 |
| Акциз на табачную продукцию | 690,6 |
| Акциз на табак | 31,167 |
| Акциз на нефтяное сырье | -64,143 |
| Акциз на керосин | -169,424 |
| Акциз на дизель | 192,839 |
| Акциз на газ | 306,631 |
| Акциз на бензин | 190,804 |
| Акциз на алкоголь | 12,201 |

Теперь, зная прогнозы всех факторов и их коэффициенты, можно построить сравнительный график из трех прогнозов: SSA, Множественная регрессия и ручная множественная регрессия. Также на графике будет реальное значение объема федерального бюджета за 2023 год для сравнения этих прогнозов. График изображен на рис. 22:

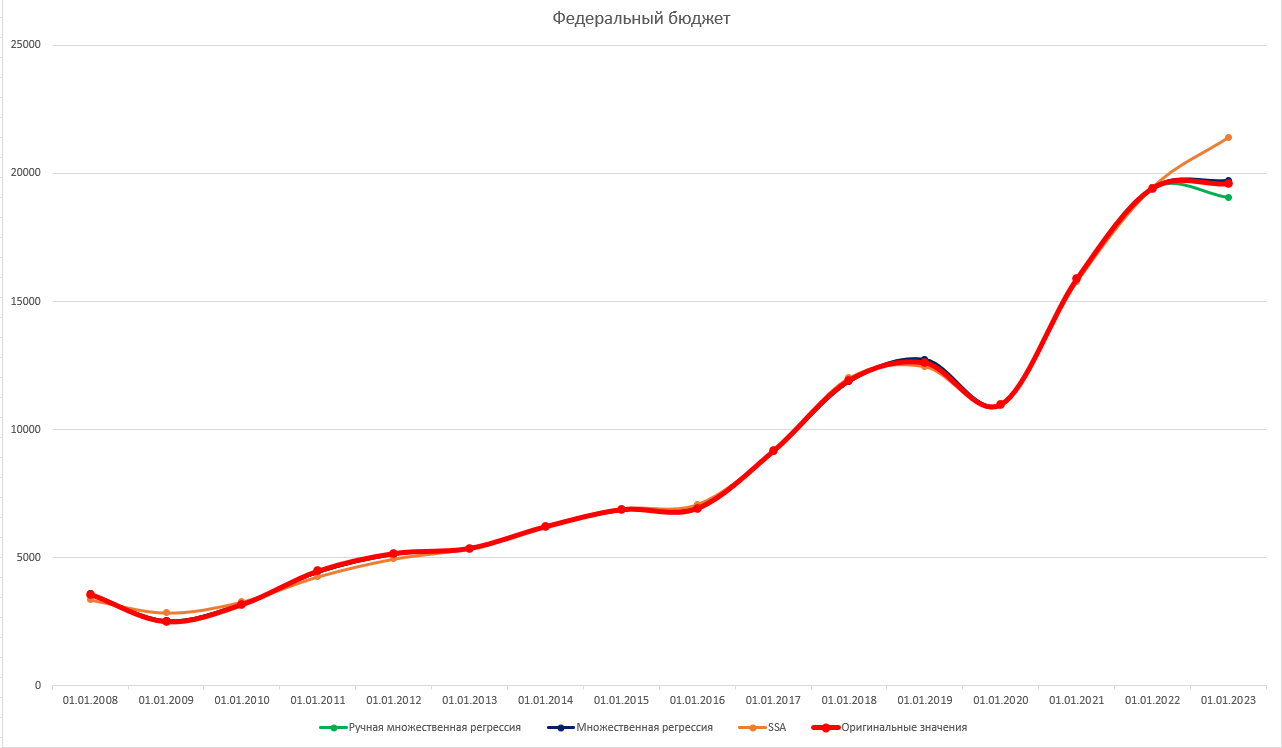


Рисунок 22 – Прогноз федерального бюджета с помощью трех алгоритмов и реальные значения федерального бюджета

Таблица 10 – Результаты прогнозов федерального бюджета

|  |  |  |  |
| --- | --- | --- | --- |
|  | SSA | Множественная регрессия | Ручная множественная регрессия |
| ADT | 1 178,503 | 266,858 | 266,858 |
| MASE | 0,078 | 0,023 | 0,023 |
| MRSE | 150,944 | 47,922 | 47,922 |
| Прогноз | 21 421,421 | 19 721,471 | 19 069,243 |
| Реальное значение | 19 600 | 19 600 | 19 600 |
| Отклонение | 9,3% | 0,6% | 2,7% |

На графике видно, что алгоритмы множественной регрессии практически совпадают с реальными значениями объема бюджета, а алгоритм SSA немного не точен, но также хорошо приближает реальные значения. Из таблицы 10 видно, что алгоритмы множественной регрессии и ручной множественной регрессии дали прогноз, наиболее близкий к реальным значениям, в то время как алгоритм SSA дал немного завышенный прогноз.

Из этого можно сделать вывод, что для получения быстрой оценки объема бюджета на следующий календарный период можно использовать прогноз с помощью алгоритма множественной регрессии прямо из приложения, а для более точной оценки выполнять полуавтоматический прогноз с использованием коэффициентов, полученных из алгоритма множественной регрессии, и выполнять прогноз с помощью SSA.

# 4. Заключение

При выполнении данной работы я провел анализ налоговой системы Российской Федерации и выяснил, какие налоги и акцизы оказывают наибольшее влияние на объем федерального бюджета. В процессе исследования был проведен обзор существующих теоретических и эмпирических исследований в области прогнозирования налоговых поступлений, были изучены основные методы и подходы, применяемые в данной области.

Я создал мультиалгоритмическую расширяемую автоматизированную систему прогнозирования налоговых поступлений, которая хранит в себе все необходимые данные, а также для которой можно реализовать новые алгоритмы и с легкостью внедрить их в систему. С использованием этой системы я сделал прогнозы трех видов налогов и восьми видов акцизов, которые затем были использованы при получении прогноза федерального бюджета.

Алгоритм SSA получился наиболее оптимальным для большинства налогов и акцизов. Из этого можно сделать вывод, что для быстрой оценки объема налога или акциза на следующий календарный период можно использовать алгоритм SSA, а прогнозы с помощью других алгоритмов можно выполнять для дополнительной проверки.

Выпускная квалификационная работа выполнена мной самостоятельно и с соблюдением правил профессиональной этики. Все использованные в работе материалы и заимствованные принципиальные положения (концепции) из опубликованной научной литературы и других источников имеют ссылки на них. Я несу ответственность за приведенные данные и сделанные выводы.

Я ознакомлен с программой государственной итоговой аттестации, согласно которой обнаружение плагиата, фальсификации данных и ложного цитирования является основанием для не допуска к защите выпускной квалификационной работы и выставления оценки «неудовлетворительно».

|  |  |
| --- | --- |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
|  |  |

« \_\_\_\_ »\_\_\_\_\_\_\_\_\_\_\_20 \_\_г.

# Список литературы

1. Аналитический портал ФНС России [Электронный ресурс] URL: <https://analytic.nalog.gov.ru/>
2. Список подакцизных товаров [Электронный ресурс] URL: <https://nalog.garant.ru/fns/nk/b89f3082384f3d024adf2f3a41be9756/>
3. Выдержка из научной статьи по экономике на тему “Прогнозирования в экономике” [Электронный ресурс] URL: <https://cyberleninka.ru/article/n/prognozirovanie-v-ekonomike/viewer>
4. Статья о реализации компьютерной системы прогнозирования на базе MS Excel [Электронный ресурс] URL: <https://fundamental-research.ru/ru/article/view?id=39174>
5. Антохина Ю.А., Колесников А.М., Медведева С.Н.,: Социально-экономическое прогнозирование, Пример используемых методов прогнозирования экономики на основе ВВП, стр. 57-75 [Учебное пособие] URL: <https://sev.msu.ru/wp-content/uploads/2020/02/UMP-Socialno-jekonomicheskoe-prognozirovanie.pdf>
6. Scott Cameron, Nick Spyropoulos, Aliya Saied-Tessier, Sandu Gusanu, Napoleon Barkis: A Review of Tax Revenue Forecasting Models for the Scottish Housing Market [Научная статья] URL: <https://www.gov.scot/binaries/content/documents/govscot/publications/research-and-analysis/2017/04/review-tax-revenue-forecasting-models-scottish-housing-market/documents/00516712-pdf/00516712-pdf/govscot%3Adocument/00516712.pdf>
7. Осипов А.Л.: Эконометрический анализ показателя уровня инновационного развития регионов РФ в условиях цифровой экономики // Наука Красноярья. 2023. Т. 12. [№ 1-3](https://elibrary.ru/contents.asp?id=50464780&selid=50464781). С. 7-11. [Научная статья] URL: <https://www.elibrary.ru/item.asp?id=50464781>
8. Glenn P. Jenkins, Chun-Yan Kuo, Gangadhar P. Shukla: Tax Analysis and Revenue Forecasting, Issues and Techniques [Научная статья] URL: <https://www.cri-world.com/publications/qed_dp_169.pdf>
9. Едронова В.Н., Акимов Н.Н.: Прогнозирование налоговых поступлений в субъекте Российской Федерации, стр. 51-54, ISSN 2311-8709 [Научный журнал] URL: <https://cyberleninka.ru/article/n/prognozirovanie-nalogovyh-postupleniy-v-subekte-rossiyskoy-federatsii/viewer>
10. Методика прогнозирования поступлений доходов, администрируемых ФНС России, в бюджет города Севастополя, стр. 55-87 [Документ] URL: <https://www.nalog.gov.ru/html/sites/www.rn92.nalog.ru/normativ/Metod_10042020.docx>
11. Стоянова Т.А., Туркова Е.В.: Современные наукоемкие технологии, Анализ и прогнозирование динамики налоговых поступлений региона на основе эконометрического моделирования временных рядов, стр. 93-97 [Научный журнал] URL: <https://www.isuct.ru/e-publ/snt/sites/ru.e-publ.snt/files/2013/03/snt_2013_n03_93.pdf>
12. Бокс., Дж., Дженкинс Г.: Анализ временных рядов, прогноз и управление, вып.1, стр. 39-101, 144-192 [Учебное пособие] URL: <https://www.studmed.ru/boks-dzh-dzhenkins-g-analiz-vremennyh-ryadov-prognoz-i-upravlenie_f23a4d04f73.html>
13. Alvin C. Rencher, William F. Christensen: Methods of multivariate analysis, ch. 10, Multivariate regression, p.339-383 [Учебное пособие]
14. Нейронные сети и способы их обучения [Электронный ресурс] URL: <https://neerc.ifmo.ru/wiki/index.php?title=%D0%9D%D0%B5%D0%B9%D1%80%D0%BE%D0%BD%D0%BD%D1%8B%D0%B5_%D1%81%D0%B5%D1%82%D0%B8,_%D0%BF%D0%B5%D1%80%D1%86%D0%B5%D0%BF%D1%82%D1%80%D0%BE%D0%BD>
15. Описание метода обратного распространения ошибки для обучения нейронных сетей [Электронный ресурс] URL: <https://neerc.ifmo.ru/wiki/index.php?title=%D0%9E%D0%B1%D1%80%D0%B0%D1%82%D0%BD%D0%BE%D0%B5_%D1%80%D0%B0%D1%81%D0%BF%D1%80%D0%BE%D1%81%D1%82%D1%80%D0%B0%D0%BD%D0%B5%D0%BD%D0%B8%D0%B5_%D0%BE%D1%88%D0%B8%D0%B1%D0%BA%D0%B8>
16. Применение алгоритма ARIMA для прогноза временных рядов [Электронный ресурс] URL: <https://www.mathnet.ru/links/8cd91f6365c7072583371e433f84b783/vyurv256.pdf>
17. Выдержка из научной статьи по математическому моделированию на тему «Методика построения модели ARIMA для прогнозирования динамики временных рядов» [Электронный ресурс] URL: <https://cyberleninka.ru/article/n/metodika-postroeniya-modeli-arima-dlya-prognozirovaniya-dinamiki-vremennyh-ryadov/viewer>
18. Charles Zaiontz: Finding AR(p) coefficients using Regression [Электронный ресурс] URL: <https://real-statistics.com/time-series-analysis/autoregressive-processes/finding-ar-coefficients-using-regression/>
19. Описание алгоритма SSA [Электронный ресурс] URL: <http://www.machinelearning.ru/wiki/index.php?title=%D0%9F%D1%80%D0%BE%D0%B3%D0%BD%D0%BE%D0%B7%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%B2%D1%80%D0%B5%D0%BC%D0%B5%D0%BD%D0%BD%D1%8B%D1%85_%D1%80%D1%8F%D0%B4%D0%BE%D0%B2_%D0%BC%D0%B5%D1%82%D0%BE%D0%B4%D0%BE%D0%BC_SSA_%28%D0%BF%D1%80%D0%B8%D0%BC%D0%B5%D1%80%29>
20. Пример построения алгоритма SSA в случае одномерного ряда [Электронный ресурс] URL: <http://www.machinelearning.ru/wiki/index.php?title=%D0%9C%D0%BD%D0%BE%D0%B3%D0%BE%D0%BC%D0%B5%D1%80%D0%BD%D0%B0%D1%8F_%D0%B3%D1%83%D1%81%D0%B5%D0%BD%D0%B8%D1%86%D0%B0%2C_%D0%B2%D1%8B%D0%B1%D0%BE%D1%80_%D0%B4%D0%BB%D0%B8%D0%BD%D1%8B_%D0%B8_%D1%87%D0%B8%D1%81%D0%BB%D0%B0_%D0%BA%D0%BE%D0%BC%D0%BF%D0%BE%D0%BD%D0%B5%D0%BD%D1%82_%D0%B3%D1%83%D1%81%D0%B5%D0%BD%D0%B8%D1%86%D1%8B_%28%D0%BF%D1%80%D0%B8%D0%BC%D0%B5%D1%80%29>
21. Алимова И.С., Соловьев В.Д., Батыршин И.З. Сравнительный анализ мер сходства, основанных на преобразовании скользящих аппроксимаций, в задачах классификации временных рядов. Труды ИСП РАН, том 28, вып. 6, 2016 г., стр. 207-222 [Электронный ресурс] URL: <https://www.mathnet.ru/links/f969895c9b4d61c4b8159120d73bfed3/tisp95.pdf>
22. Алимова И.С. Лекция о применении алгоритма динамической трансформации для анализа звуковых дорожек [Электронный ресурс] URL: <http://www.machinelearning.ru/wiki/images/c/c3/Digital_Signal_Processing%2C_lecture_6.pdf>
23. Philip Hans Franses: A note on the Mean Absolute Scaled Error [International Journal of Forecasting] URL: <https://pure.eur.nl/ws/portalfiles/portal/47633436/A-note-on-the-MASE-Revision-for-IJF.pdf>
24. Руководства по написанию web-приложения с помощью Spring Boot [Электронный ресурс] URL: <https://spring.io/guides>
25. Руководства по написанию сайтов с помощью TypeScript и React [Электронный ресурс] URL: <https://react.dev/learn>
26. Данные по формам налоговой отчетности [Электронный ресурс] URL: <https://www.nalog.gov.ru/rn77/related_activities/statistics_and_analytics/forms/>
27. Информация об изменении инфляции в РФ [Электронный ресурс] URL: <https://xn----ctbjnaatncev9av3a8f8b.xn--p1ai/%D1%82%D0%B0%D0%B1%D0%BB%D0%B8%D1%86%D1%8B-%D0%B8%D0%BD%D1%84%D0%BB%D1%8F%D1%86%D0%B8%D0%B8>
28. Информация о доходах организаций в РФ [Электронный ресурс] URL: <https://rosstat.gov.ru/storage/mediabank/oborot.htm>
29. Информация о количестве действующих организаций в РФ [Электронный ресурс] URL: <https://xn--h1ari.xn--p1ai/Main/StatisticalInformation>
30. Цена нефти Brent на бирже в долларах [Электронный ресурс] URL: <https://bcs-express.ru/kotirovki-i-grafiki/brent>
31. Объем ВВП РФ по годам [Электронный ресурс] URL: <https://rosstat.gov.ru/storage/mediabank/VVP_god_s_1995-2023.xlsx>

# Приложение

1. Репозиторий с кодом системы – <https://github.com/SlamperBOOM/Graduate-Work>