

Министерство образования Республики Беларусь
Учреждение образования «Белорусский государственный университет
информатики и радиоэлектроники»

Факультет компьютерных систем и сетей

Кафедра программного обеспечения информационных технологий

Дисциплина: Модели и методы обработки больших объёмов данных
(МиМОБОД)

Отчет по проекту
на тему:
«Анализ данных для цен на натуральный газ»

Выполнила:
студент гр. 156301
Бондарева Т. О.

Проверил:
доц. каф. информатики
Стержанов М. В.

Минск 2021

СОДЕРЖАНИЕ

1. Постановка задачи.....	3
2. Получение данных	4
3. Очистка датасета	6
4. Анализ данных.....	7
5. Прогнозирование цены.....	8
Выводы	10

1. ПОСТАНОВКА ЗАДАЧИ

В рамках проекта необходимо разработать скрапер получения данных из открытых источников и сохранить полученные данные в базу данных. Полученные данные необходимо отфильтровать и очистить при необходимости и сформировать датасеты и провести их анализ, путем построения графиков.

Для реализации поставленной задачи было принято решение провести анализ цен на природный газ. Данные для проведения анализа были взяты с сайта <https://datahub.io>. Данный сайт предоставляет данные по дням и месяцам. Для лучшей наглядности анализироваться будут данные по дням.

Также кроме анализа будет предсказана цена на основе собранных данных, используя простейшие модели машинного обучения.

2. ПОЛУЧЕНИЕ ДАННЫХ

Сайт <https://datahub.io/> предоставляет данные по различным тематикам. Данные по ценам на натуральный газ доступны по странице <https://datahub.io/core/natural-gas>. Внешний вид сайта представлен на рисунке 1.

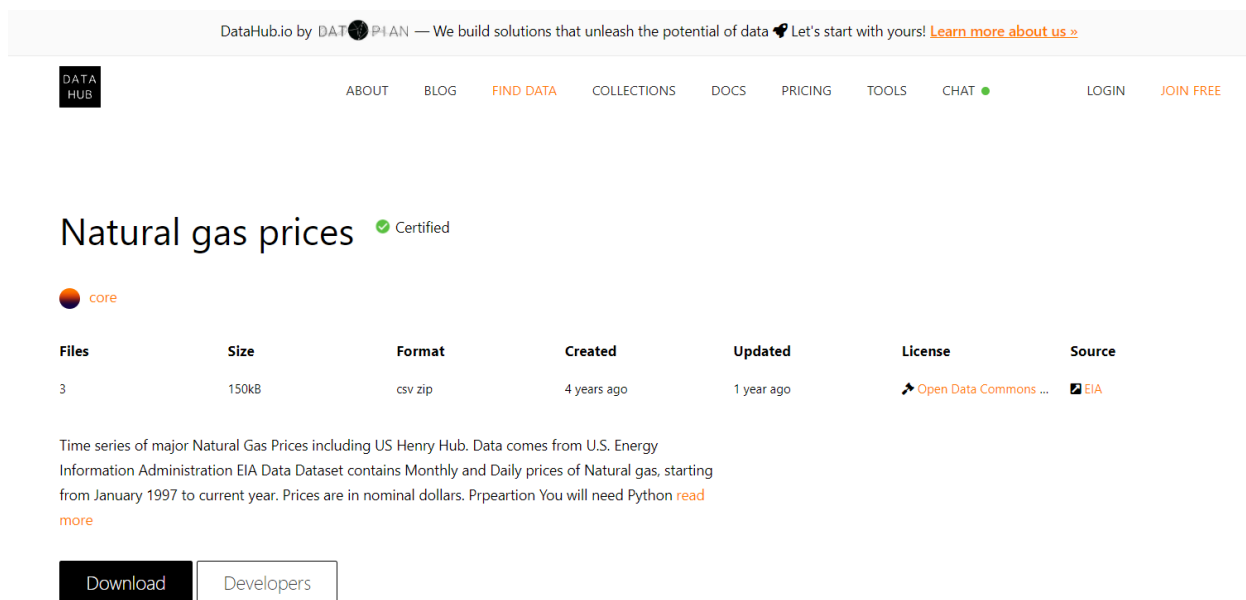


Рисунок 1 – Внешний вид страницы <https://datahub.io/core/natural-gas>

Данные по ценам на натуральный газ предоставляются в двух видах: данные по дням и данные по месяцам. Эти можно скачать уже готовые архивы данных в формате csv, json и zip(рисунок 2), а также можно найти ссылки, которые помогут получить дынные с помощью запроса.

Data Files

Download files in this dataset

File	Description	Size	Last changed	Download
daily		730kB		csv (730kB) , json (586kB)
monthly		33kB		csv (33kB) , json (27kB)
natural-gas.zip	Compressed versions of dataset. Inclu...	103kB		zip (103kB)

Рисунок 2 – Ссылки на скачивания данных

В ходе постановки задачи был выбран анализ данных по дням. Получить их можно по ссылке https://pkgstore.datahub.io/core/natural-gas/daily_json/data/2e630ca50c39a1a3cf6c3aff57a1b132/daily_json.json.

Формат получаемых данных представлен на рисунке 3.

Field information

Field Name	Order	Type (Format)
Date	1	date (%Y-%m-%d)
Price	2	number (default)

Рисунок 3 – Формат данных

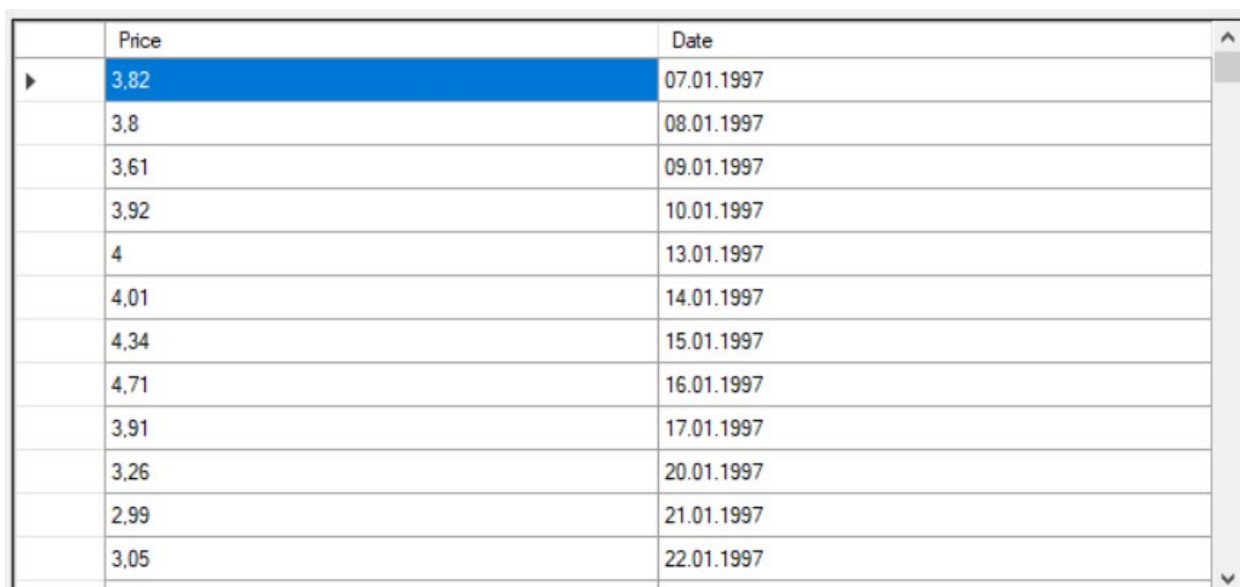
Алгоритм сбора информации:

1. Отправка GET запроса по ссылке https://pkgstore.datahub.io/core/natural-gas/daily_json/data/2e630ca50c39a1a3cf6c3aff57a1b132/daily_json.json.
2. Получения ответа на запрос.
3. Проверка статуса ответа.
4. Если статус ответа Success, то переходим к пункту 5, иначе генерируем исключения.
5. Чтение результат запроса.
6. Десериализация ответа в массив объектов, хранящих день и цену на натуральный газ за этот день.
7. Подключение к базе данных.
8. Цикл по элементам с проверкой на существование данного элемента в базе и занесением его, если такой не найден.
9. Вывод элементов базы данных в виде таблицы.

Для проверки работоспособности получения данных, код скрапера был покрыт тестами.

3. ОЧИСТКА ДАТАСЕТА

Датасет содержит 5953 элементов с 2 характеристиками: Date и Price(рисунок 4).



	Price	Date
▶	3,82	07.01.1997
	3,8	08.01.1997
	3,61	09.01.1997
	3,92	10.01.1997
	4	13.01.1997
	4,01	14.01.1997
	4,34	15.01.1997
	4,71	16.01.1997
	3,91	17.01.1997
	3,26	20.01.1997
	2,99	21.01.1997
	3,05	22.01.1997

Рисунок 4 – Данные из базы данных

Некоторые характеристики цены содержат значения null. Так как такие данные не подходят для анализа, то удалим их из датасета. Оставшиеся данные корректны и готовы к анализу.

4. АНАЛИЗ ДАННЫХ

Полученные данные представим в виде графика(рисунок 5).

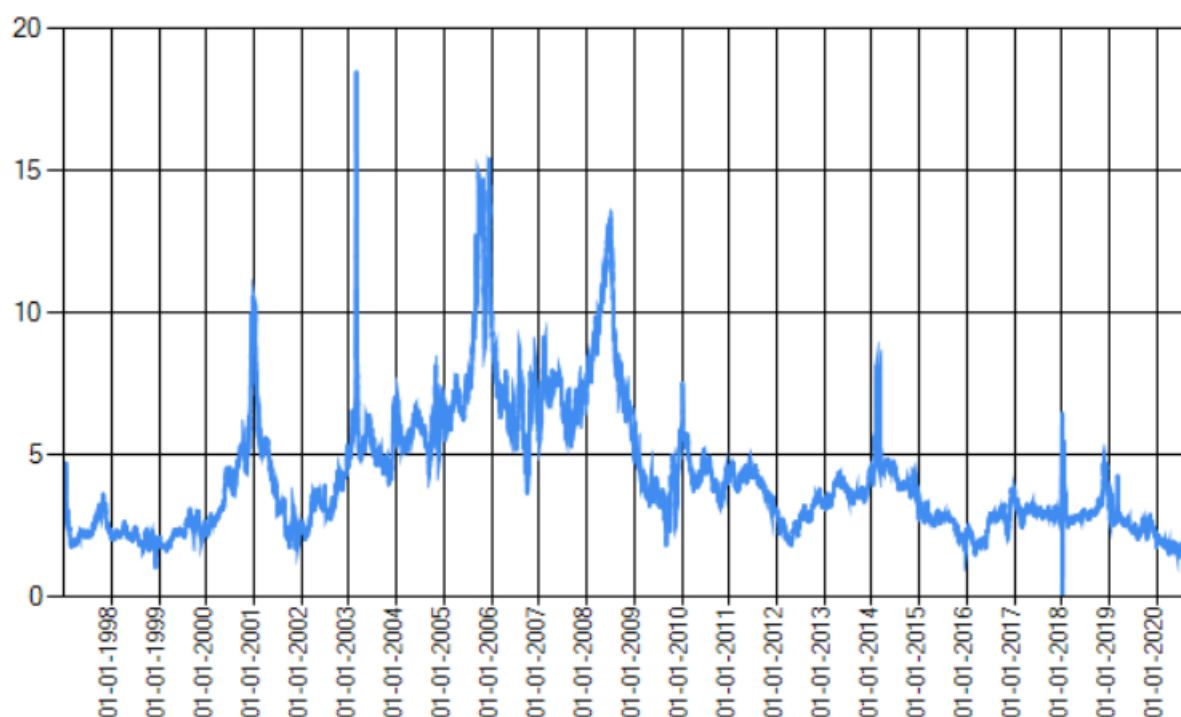


Рисунок 5 – Полученный датасет

В 2005 году наблюдается значительный скачок цен, поэтому рассмотрим его подробнее(рисунок 6).

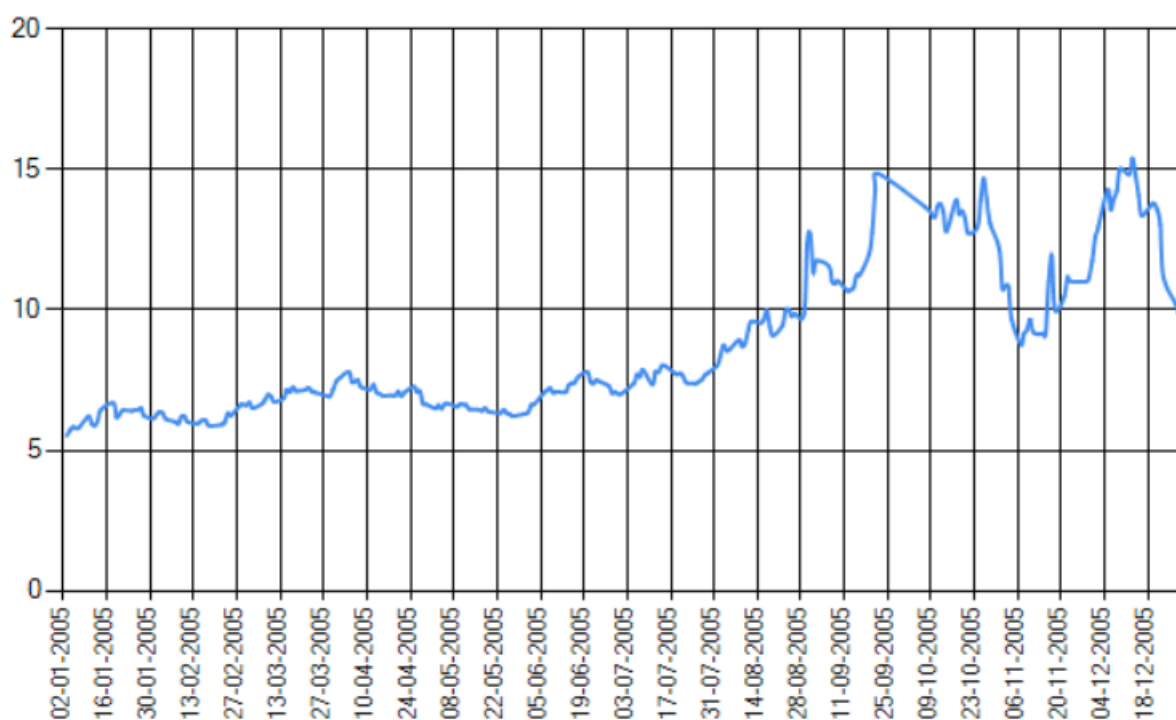


Рисунок 6 – Цены с 1 января по 31 декабря 2005 года

5. ПРОГНОЗИРОВАНИЕ ЦЕНЫ

В связи с тем что данная выборка – временной ряд, для предсказания применяется алгоритм «Анализ сингулярного спектра»(SSA). SSA разбивает временной ряд на набор основных компонентов. Эти компоненты можно интерпретировать как части одного сигнала: тенденции, шумы, сезонные колебания и многие другие факторы. Полученные компоненты реконструируются и применяются для прогнозирования значений на будущие периоды.

Датасет был разбит на две выборки: 30% и 70% - обучающая и тестовая соответственно.

Распределение реальных и предсказанных результатов для июля 2005 года можно увидеть на рисунке 7, где голубым цветом обозначена реальная цена на тот период, а красным – предсказанная.

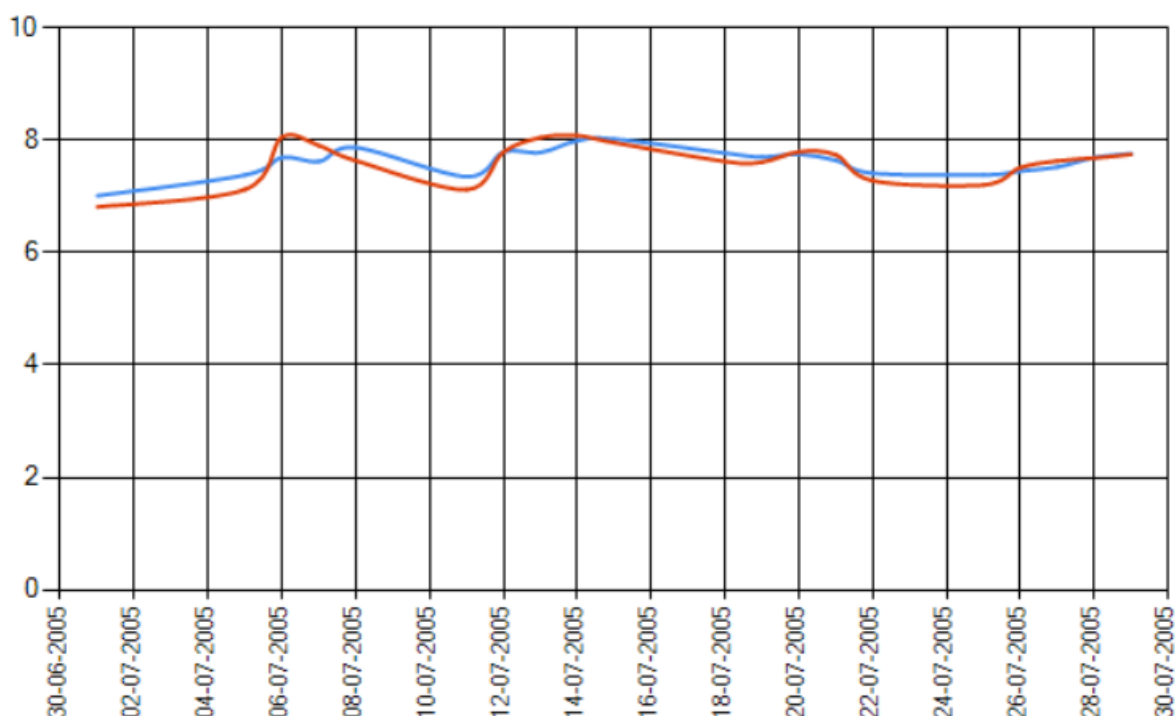


Рисунок 7 – Данные за июль 2005 года

Рассмотрим предельные значения за данный период. Желтым цветом обозначен верхний предел предсказанной цены, синим цветом – нижний предел. Данный график представлен на рисунке 8.

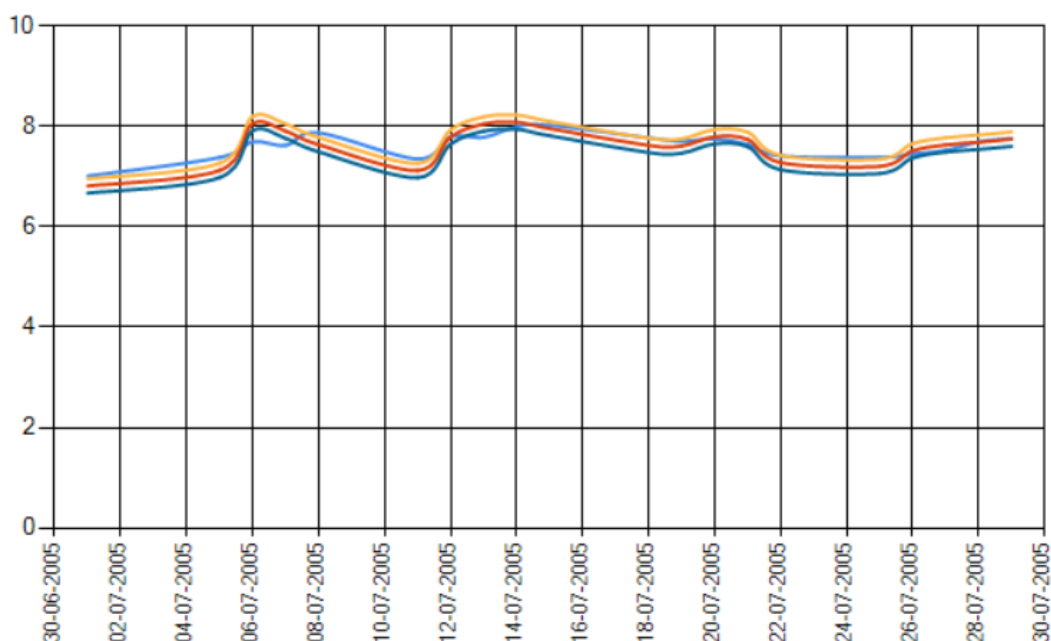


Рисунок 8 – Данные за июль 2005 года с предельными значениями предсказания

Для данного месяца прогнозирования показала очень приближенное значение к настоящим данным. Рассмотрим прогнозирование на более большем промежутке, например за 2015 год.

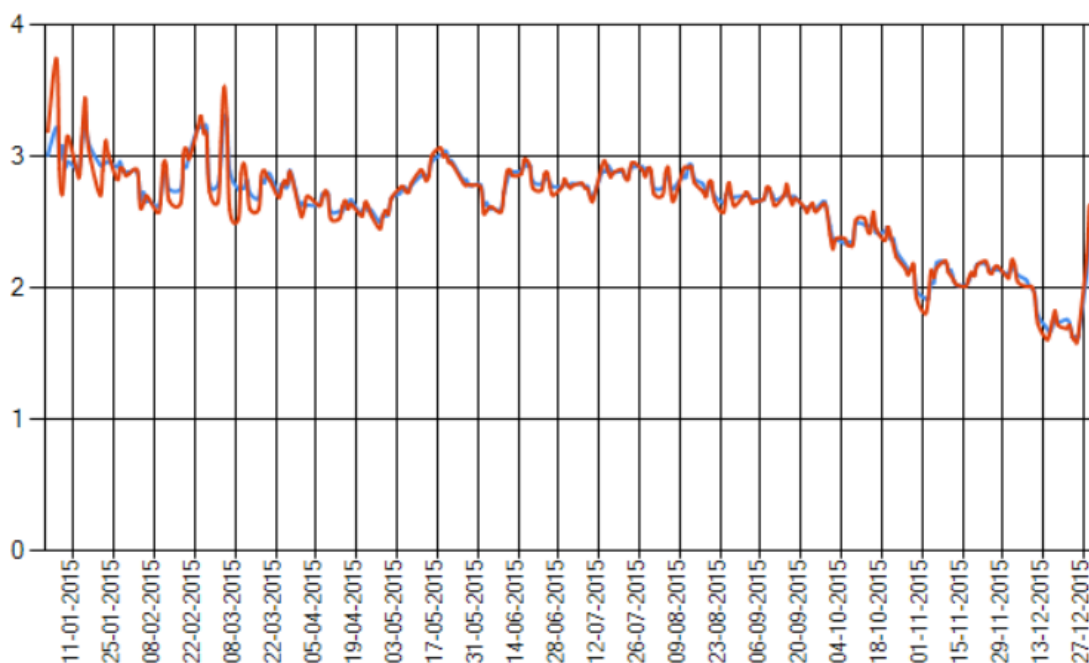


Рисунок 9 – Данные за июль 2015 года

Прогнозирование предсказывало более скачкообразное поведение в начале года, но погрешность также не велика, что говорит, что модель вполне может быть использована для предсказания цен.

ВЫВОДЫ

В проекте был реализован сбор данных о ценах на натуральный газ. Код скрапера был покрыт тестами. Данные были очищены и подготовлен датасет, выполнен анализ данных. Анализ данных был представлен в виде таблицы и наглядных графиков. Дополнительно были обучена модель по алгоритму «Анализ сингулярного спектра», который показал сравнительно неплохие результаты прогнозирования.