Анализ предсказаний курсов валют с помощью

модели машинного обучения на разный промежуток времени в зависимости от ключевых индикаторов на фондовом рынке

Программа:

Специализация: Искусственный интеллект. Цифровые Профессии

Студент: Эгипти Евгений Викторович

Оглавление

[Введение 1](#_Toc172787023)

[Глава 1 Теоретическая часть 3](#_Toc172787024)

[1.1 Выбор инструментов 3](#_Toc172787025)

[1.2 Поиск и загрузка датасетов 3](#_Toc172787026)

[1.3 Выбор модели машинного обучения 3](#_Toc172787027)

[1.4 Оценка качества предсказаний модели 4](#_Toc172787028)

[1.5 Концепция ноутбука 5](#_Toc172787029)

[Глава 2 Практическая часть 6](#_Toc172787030)

[2.1 Среда разработки 6](#_Toc172787031)

[2.2 Выбор датасетов 6](#_Toc172787032)

[2.3 Создание файла Jupiter Notebook 7](#_Toc172787033)

[2.4 Установка необходимых python-пакетов 7](#_Toc172787034)

[2.5 Импортируем нужные модули 8](#_Toc172787035)

[2.6 Загрузка датасетов 9](#_Toc172787036)

[2.7 Разведочный анализ данных (EDA) 12](#_Toc172787037)

[2.7.1 Посмотрим типы данных. 12](#_Toc172787038)

[2.7.2 Посмотрим пропуски данных. 13](#_Toc172787039)

[2.7.3 Посмотри статистику по дата фрейму. 13](#_Toc172787040)

[2.7.4 Выделим численные признаки от даты 15](#_Toc172787041)

[2.7.5 Создадим словарь признаков и пояснения 15](#_Toc172787042)

[2.7.6 Визуализируем распределение числовых признаков 15](#_Toc172787043)

[2.8 Посмотрим какие признаки имеют выбросы 17](#_Toc172787044)

[2.9 Создание новых признаков 19](#_Toc172787045)

[2.10 Проверка состояния данных 21](#_Toc172787046)

[2.11 Метрика, по которой будем оценивать качество работы модели 22](#_Toc172787047)

[2.12 Разбиение временного ряда на train и test 22](#_Toc172787048)

[2.13 Функция кросс-валидации 24](#_Toc172787049)

[2.14 Создание модели и дата фрейма для хранения результатов кросс-валидации 25](#_Toc172787050)

[2.15 Кросс-валидация данных 25](#_Toc172787051)

[2.15.1 С шагом в год 26](#_Toc172787052)

[2.15.2 С шагом в квартал 30](#_Toc172787053)

[2.15.3 С шагом в месяц 35](#_Toc172787054)

[2.15.4 С шагом в неделю 40](#_Toc172787055)

[2.16 Визуализация результатов кросс-валидации 44](#_Toc172787056)

[2.17 Важность признаков модели 47](#_Toc172787057)

[2.18 Средняя ошибка предсказаний модели в зависимости от длительности периода предсказаний 51](#_Toc172787058)

[2.19 Автоматизация процесса 54](#_Toc172787059)

[Заключение 55](#_Toc172787060)

[Список используемой литературы 56](#_Toc172787061)

Введение

Роль валютного курса в экономике переоценить сложно. Изменения валютного курса предопределяют статус той или иной страны в мировом экономическом пространстве, а вместе с ним ориентиры ее будущего развития, как во внешних, так и во внутренних связях.

Курс национальной валюты является важным ориентиром для принятия решений экономическими агентами, как на уровне населения, так и на уровне правительства. При высокой волатильности валюты возрастают издержки экспортно-импортных операций, растут цены на импортируемые товары. Это ведет к повышению цен на продукцию на внутреннем рынке. Чтобы защититься от роста цен и возможной девальвации, потребители увеличивают долю сбережений в иностранной валюте, что не способствует укреплению и стабилизации национальной. Резкие скачки национальной валюты дают повод населению снизить потребление, а предпринимателем сократить инвестиции в национальную экономику. Высокая волатильность препятствует бизнесу в построении долгосрочных стратегий развития, что приводит к снижению экспорта предприятиями и замедлением производства [1].

На данный момент машинное обучение является одной из наиболее развивающихся областей прикладной математики, позволяющих решать большой спектр задач предсказания и распознавания. Многие экономисты используют методы анализа данных для предсказания валютных курсов. Так, например, Martin Evans и Richard Lyons в своей статье «Micro-Based Exchange-Rate Forecasting» используют метод k ближайших соседей и метод опорных векторов для прогнозирования основных мировых валютных пар (EUR/USD, GBP/USD, USD/JPY) [4].

**Темой проекта** является анализ предсказаний курсов валют с помощью модели машинного обучения на разный промежуток времени в зависимости от ключевых индикаторов на фондовом рынке.

**Целью** будет являться хорошо интерпретируемые данные, показывающие с какой погрешностью модель машинного обучения может предсказывать курс валют по ключевым индикаторам на фондовом рынке на разный промежуток времени, а также показать важность ключевых индикаторов для модели машинного обучения.

**Будет решена проблема** вероятности недополучить прибыль или понести убытки от финансовых, торговых и кредитных операций из-за изменчивости соотношения валют**.** То есть снизятся валютные риски.

**Задачи:**

* Собрать и обработать данные необходимые для обучения и тестирования модели;
* Определить признаки (фичи) и ключевые параметры (таргеты);
* Выбрать модель машинного обучения;
* Определить уровень ошибки предсказаний модели на временном ряду, для разных отрезков времени;
* Вывести графики качества предсказаний и важности признаков;
* Автоматизировать процесс вычислений и отрисовки графиков при добавлении новых фичей и таргетов.

**Инструменты:**

JupyterLab [5; 18], Visual Code [11], Microsoft 365 Word [19], Zotero [16; 2; 15].

**Состав команды:** Эгипти Евгений Викторович

1. Теоретическая часть
   1. Выбор инструментов

В качестве инструментов для написания кода будем использовать бесплатные программы. JupiterLab [5] который входит в состав программы Anaconda [18]. JupiterLab это практически то же самое, что и Jupiter Notebook, только с расширенными возможностями. JupiterLab как и Jupiter Notebook позволяют выполнять каждую ячейку кода отдельно, при этом результат выполнения и все переменные сохраняются в ноутбуке. Это позволяет запускать ресурсоёмкие части кода один раз (например обучение модели), а дальше экспериментировать с результатами, без надобности запускать весь код с начала. Это очень удобно, по сколько, например процессы кросс-валидации могут занимать более часа.

Для оформления текстового описания работы будем использовать программу Microsoft 365 Word [19], которая позволяет применять широкие возможности для редактирования документов.

В бесплатной программе Visual Studio Code [11], можно просматривать ноутбук и копировать ячейки кода с сохранением полного форматирования, для последующего переноса этого текста в документ Word.

Бесплатная программа Zotero [16; 2; 15], поможет в оформлении библиографических ссылок.

* 1. Поиск и загрузка датасетов

Датасеты должны содержать информацию о ключевых индексах фондового рынка (фичах), например цены на нефть, драг. металлы, индексы бирж и т.д. и обменных кусах различных валют (таргеты). Эти данные должны быть привязаны ко времени торгов. Желательно, чтобы датасеты были бесплатными и читаемыми в Excel.

* 1. Выбор модели машинного обучения

Выберем модель машинного обучения на основе алгоритма случайного леса RandomForestRegressor из библиотеки Scikit-Learn [13]. Это регрессионная модель, которая может предсказывать числа и есть возможность визуализировать важность признаков для этой модели.

* 1. Оценка качества предсказаний модели

Качество модели будем оценивать по следующей метрике:

()

Где – среднее значение курса валют за определённый отрезок времени (неделя, месяц, квартал, год);

– предсказанное среднее значение курса валют за определённый отрезок времени (неделя, месяц, квартал, год);

– величина ошибки предсказанных средних значений в процентах.

Будем проверять эффективность работы модели на тестовых данных, то есть на тех данных которая модель не видела. Так как последовательность данных для нас важна, будем делать кросс-валидацию на временном ряду [7]. Ниже, рисунок 1, показывает принцип кросс-валидации на временном ряду.

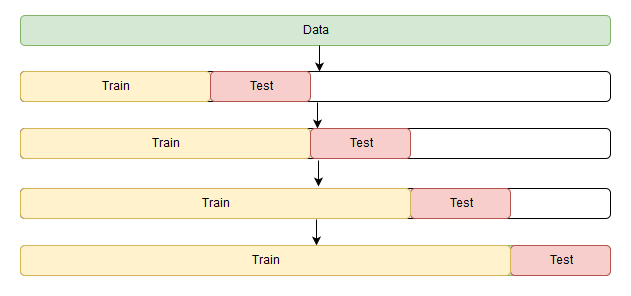


Рисунок 1. Принцип кросс-валидации на временном ряду

То есть берётся набор данных Data и разбивается на Train и Test, на Train модель тренируется, а на Test проверяет свою эффективность предсказания. Внутри Scikit-Learn есть готовая функция TimeSeriesSplit [14] которая разбивает выборку на train и test, но нам она не подойдёт, поскольку она будет разбивать выборку с фиксированным количеством элементов, а нам нужно ориентироваться на даты, а не на количество элементов. Разбивка будет осуществляться по годам, кварталам, месяцам, неделям. Для этого потребуется написать собственную функцию кросс-валидации.

* 1. Концепция ноутбука

Чтобы избавиться от ручного монотонного труда при загрузке, обработке датасетов, отрисовке графиков, нужно продумать программу так, чтобы максимально автоматизировать весь процесс.

Например, датасеты должны автоматически загружаться с какой ни будь директории, в соответствии с правилами загрузки.

Распределение на фичи и таргеты также должно происходить в автоматическом режиме. Например файлы, содержащие символ ‘\_’ автоматически будут инициализированы как таргет, все остальные как фичи.

На всех графиках должны в автоматическом режиме добавляться/удалятся фичи и таргеты. Если есть русская расшифровка таргета/фич, она также должна в автоматическом режиме вставляться в график. Возможна некая корректировка размера полотна графика в коде программы для улучшения вида графиков.

Процесс добавления/удаления фич и таргетов должен сводиться к добавлению или удалению датасета в директорию откуда производиться чтение, без вмешательства в код программы.

Кросс-валидация временного ряда должна выполняться с привязкой к датам с определённой длиной периода (год, квартал, месяц, неделя)

1. Практическая часть
   1. Среда разработки

В качестве среды разработки был использован JupyterLab [5], там больше возможностей по сравнению с Jupyter Notebook. Можно одновременно работать с несколькими файлами, использовать отладчик [10], если потребуется.

Данная среда разработки, а также другие инструменты входят в бесплатную программу Anaconda [18] которую можно скачать с сайта <https://www.anaconda.com/>.

Разработку проекта желательно вести в новом окружении, чтобы не возникло конфликта между версиями установленных и устанавливаемых пакетов. Для этого, открываем программу Anaconda и выбираем вкладку Environments, в открывшейся вкладке нажимаем кнопку Create, задаём название окружения и версию Python. Делаем активным наше окружение, в данном случае окружение diplom\_2024\_gb, см. рисунок 2.

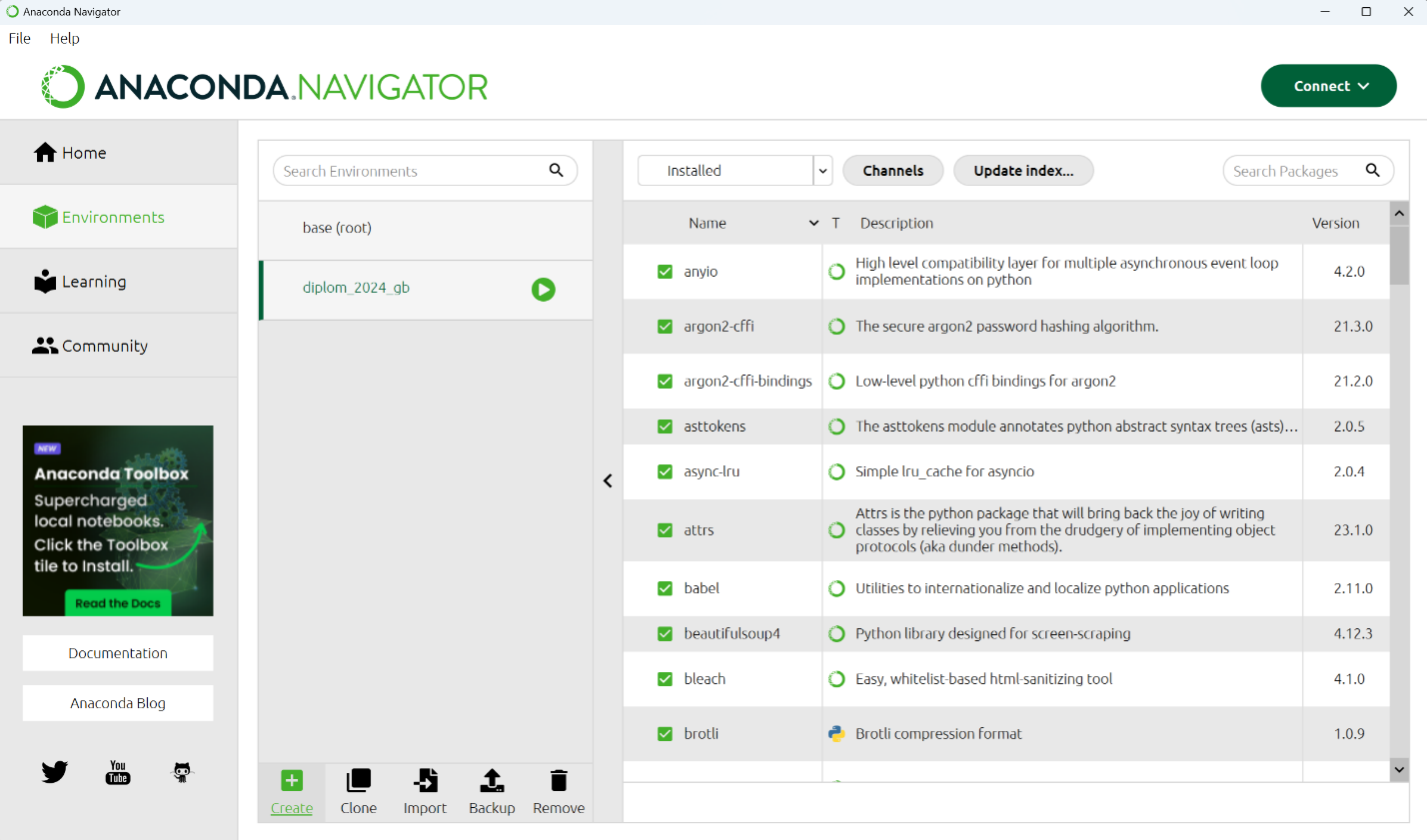


Рисунок 2. Выбор активного окружения diplom\_2024\_gb

После выбора окружения, нажимаем вкладку Home и выбираем среду разработки JupyterLab.

* 1. Выбор датасетов

Датасеты были взяты с сайта InvestFunds (<https://investfunds.ru/>), их можно бесплатно скачать в формате Excel, за выбранный промежуток времени.

В качестве ключевых индикаторов [8] были приняты следующие показатели:

* индекс Мосбиржи;
* индекс РТС;
* индекс S&P 500;
* платина (Банк России);
* нефть Brent;
* золото (Банк России);
* серебро (Банк России);
* палладий (Банк России).

В качестве таргетов (целевой переменной) были взяты следующие курсы валют [9]:

* доллар США / Российский рубль;
* Евро / Российский рубль;
* бивалютная корзина / Российский рубль;
* Британский фунт стерлингов / Российский рубль;
* Швейцарский франк / Российский рубль;
* Китайский юань / Российский рубль;
* Японская иена / Российский рубль;
  1. Создание файла Jupiter Notebook

После того как мы установили JupyterLab, выбрали рабочее окружение и открыли среду разработки (пункт 2.1), создадим файл Jupiter Notebook. Для этого нужно перейти меню File → New → Notebook.

Файл Jupiter Notebook состоит из ячеек, в которых может находиться код программы или пояснительный текст программы, написанный на языке Markdown. Код каждой ячейки может выполняться отдельно, а также есть возможность запустить выполнение всех ячеек. Результат выполнения ячейки может выводиться непосредственно после самой ячейки.

* 1. Установка необходимых python-пакетов

Для того чтобы использовать необходимые нам библиотеки (Pandas, NumPy, Matplotlib и т.д.) нужно установить соответствующие пакеты с помощью системы управления пакетами pip. Для этого в ячейки ноутбука запустим следующий код:

ячейка Jupiter Notebook

# Для каждого проекта я использую новое окружение,

# думаю это помогает избежать ошибок связанных с версиями пакетов

!pip install pandas

!pip install matplotlib

!pip install seaborn

!pip install numpy

!pip install scikit-learn

!pip install openpyxl

* 1. Импортируем нужные модули

После установки пакетов, нужно импортировать нужные нам для работы модули.

ячейка Jupiter Notebook

# Импортируем необходимые библиотеки

import numpy as np

import pandas as pd

from matplotlib import pyplot as plt

import seaborn as sns

# Следующая магическая команда Jupyter Notebook нужна для того, чтобы графики

# отображались прямо в ноутбуке, а не в отдельном окне

%matplotlib inline

# Настройка более четкого отображения графиков

%config InlineBackend.figure\_format = 'svg'

# Модуль для зазбиения выборки на тренировочнию и тестовую

from sklearn.model\_selection import train\_test\_split

# Уберем warnings

import warnings

warnings.filterwarnings('ignore')

# Настройка формата вывода чисел float

pd.set\_option('display.float\_format', '{:.2f}'.format)

# Вычисление Z-score

from scipy import stats

from sklearn.preprocessing import LabelEncoder

from scipy.stats.mstats import winsorize

# Дата и время

from datetime import datetime

# Случайные числа

import random as rnd

# Алгоритм машинного обучения 'Метод случайного леса'

from sklearn.ensemble import RandomForestRegressor

# Кросс-валидация

from sklearn.model\_selection import cross\_validate

# Разбиение

from sklearn.model\_selection import KFold

# Метрика r2

from sklearn.metrics import r2\_score

# Пути файловой системы

from pathlib import Path

# Регулярные выражения

import re

# Округление числа в большую сторону

from math import ceil

* 1. Загрузка датасетов

Загрузка файлов в дата фрейм происходит в автоматическом режиме. Программа читает все файлы с расширением xlsx в директории заданной константой PATH\_DATASET [6]. Все файлы соединяются по полю Дата. Если число строк в файле меньше, чем MIN\_ROW, то такие данные не будут включены в итоговый дата фрейм. В дата фрейме название столбца будет соответствовать названию файла. Информация о ходе процесса выводится под ячейкой.

ячейка Jupiter Notebook

# Расположение данных

PATH\_DATASET = './Dataset'

# Минимальное разрешённое количество строк для загрузки

MIN\_ROW = 5000

# Итоговый датафрейм со всеми данными

data\_loaded = pd.DataFrame()

# Расположение файлов датасета

p = Path(Path.cwd() / PATH\_DATASET)

for obj in p.iterdir():

    if obj.is\_file():

        \*\_, name\_file\_all = str(obj).split('\\')

        name\_file, type\_file = name\_file\_all.split('.')

        if type\_file == 'xlsx':

            df\_temp = pd.read\_excel(f'{PATH\_DATASET}/{name\_file\_all}')

            # Переименуем стобцы

            df\_temp.rename(columns={'Значение': f'{name\_file}'}, inplace=True)

            # Нормализуем дату

            df\_temp['Дата'] = df\_temp['Дата'].dt.normalize()

            # Проверяем условие первой загрузки

            if data\_loaded.shape[1]:

                if df\_temp.shape[0] >= MIN\_ROW:

                    print(f'Обработан: {name\_file\_all} {df\_temp.shape}')

                    data\_loaded = pd.merge(data\_loaded, df\_temp, on='Дата', how='inner')

                    print(f'Размерность после merge {data\_loaded.shape}\n')

                else:

                    print(f'Отклонён, мало строк: {name\_file\_all}\n')

            else:

                data\_loaded = df\_temp

                print(f'Обработан: {name\_file\_all} {df\_temp.shape}\n')

результат выполнения

Обработан: chf\_rub-(банк-россии).xlsx (7303, 2)

Обработан: cny\_rub-(банк-россии).xlsx (5419, 2)

Размерность после merge (5419, 3)

Обработан: eur\_rub-(банк-россии).xlsx (7303, 2)

Размерность после merge (5419, 4)

Обработан: gbp\_rub-(банк-россии).xlsx (7303, 2)

Размерность после merge (5419, 5)

Обработан: jpy\_rub-(банк-россии).xlsx (7303, 2)

Размерность после merge (5419, 6)

Обработан: s-p-500.xlsx (6178, 2)

Размерность после merge (4303, 7)

Обработан: usd\_rub-(банк-россии).xlsx (7303, 2)

Размерность после merge (4303, 8)

Отклонён, мало строк: бивалютная-корзина\_rub.xlsx

Обработан: золото-(банк-россии).xlsx (6034, 2)

Размерность после merge (4285, 9)

Обработан: индекс-мосбиржи.xlsx (6130, 2)

Размерность после merge (4185, 10)

Обработан: нефть-brent.xlsx (6514, 2)

Размерность после merge (4185, 11)

Обработан: палладий-(банк-россии).xlsx (6838, 2)

Размерность после merge (4185, 12)

Обработан: платина-(банк-россии).xlsx (6912, 2)

Размерность после merge (4185, 13)

Обработан: ртс.xlsx (6128, 2)

Размерность после merge (4178, 14)

Обработан: серебро-(банк-россии).xlsx (6034, 2)

Размерность после merge (4178, 15)

Выведем первые пять строк дата фрейма.

ячейка Jupiter Notebook

# Посмотрим, что загрузилось

data\_loaded.head()

результат выполнения

Изображение выглядит как текст, снимок экрана, линия, Шрифт

Автоматически созданное описание

Рисунок 3. Вывод первых пяти строк дата фрейма

Проверим размер загруженных данных.

ячейка Jupiter Notebook

# Посмотрим размер

data\_loaded.shape

результат выполнения

(4178, 15)

В итоговый дата фрейм data\_loaded не был включен датасет бивалютная-корзина\_rub, по сколько он не удовлетворяет критерию минимального количества строк MIN\_ROW.

* 1. Разведочный анализ данных (EDA)

С помощью разведочного анализа (EDA) оценим качество загруженных данных.

* + 1. Посмотрим типы данных.

ячейка Jupiter Notebook

data\_loaded.dtypes

результат выполнения

Дата datetime64[ns]

chf\_rub-(банк-россии) float64

cny\_rub-(банк-россии) float64

eur\_rub-(банк-россии) float64

gbp\_rub-(банк-россии) float64

jpy\_rub-(банк-россии) float64

s-p-500 float64

usd\_rub-(банк-россии) float64

золото-(банк-россии) float64

индекс-мосбиржи float64

нефть-brent float64

палладий-(банк-россии) float64

платина-(банк-россии) float64

ртс float64

серебро-(банк-россии) float64

dtype: object

* + 1. Посмотрим пропуски данных.

ячейка Jupiter Notebook

# Проверка того, в каких столбцах отсутствуют значения

print(data\_loaded.isnull().sum(axis=0))

результат выполнения

Дата 0

chf\_rub-(банк-россии) 0

cny\_rub-(банк-россии) 0

eur\_rub-(банк-россии) 0

gbp\_rub-(банк-россии) 0

jpy\_rub-(банк-россии) 0

s-p-500 0

usd\_rub-(банк-россии) 0

золото-(банк-россии) 0

индекс-мосбиржи 0

нефть-brent 0

палладий-(банк-россии) 0

платина-(банк-россии) 0

ртс 0

серебро-(банк-россии) 0

dtype: int64

* + 1. Посмотри статистику по дата фрейму.

ячейка Jupiter Notebook

data\_loaded.describe()

результат выполнения

Изображение выглядит как текст, снимок экрана, Шрифт, линия

Автоматически созданное описание

Рисунок 4. Статистика по загруженному дата фрейму

* + 1. Выделим численные признаки от даты

ячейка Jupiter Notebook

# Создадим список численных признаков

num\_cols = [column for column in data\_loaded if not column=='Дата']

num\_cols

результат выполнения

['chf\_rub-(банк-россии)',

'cny\_rub-(банк-россии)',

'eur\_rub-(банк-россии)',

'gbp\_rub-(банк-россии)',

'jpy\_rub-(банк-россии)',

's-p-500',

'usd\_rub-(банк-россии)',

'золото-(банк-россии)',

'индекс-мосбиржи',

'нефть-brent',

'палладий-(банк-россии)',

'платина-(банк-россии)',

'ртс',

'серебро-(банк-россии)']

* + 1. Создадим словарь признаков и пояснения

Для того, чтобы на графиках отображалось более понятное и презентабельное название. Сделаем словарь, где название столбца будет соответствовать его более подробному описанию. Название столбца соответствует названию файла.

Это действие не обязательно, тогда вместо подробного описания на графике будет выводиться название столбца, которое будет соответствовать названию файла.

ячейка Jupiter Notebook

# Словарь признаков и их пояснения

explanations\_col = {

    'chf\_rub-(банк-россии)': 'Швейцарский франк / Российский рубль',

    'eur\_rub-(банк-россии)': 'Евро / Российский рубль',

    'gbp\_rub-(банк-россии)': 'Британский фунт стерлингов / Российский рубль',

    'jpy\_rub-(банк-россии)': 'Японская иена / Российский рубль',

    's-p-500': 'Индекс S&P 500',

    'usd\_rub-(банк-россии)': 'Доллар США / Российский рубль',

    'cny\_rub-(банк-россии)': 'Китайский юань / Российский рубль'

}

* + 1. Визуализируем распределение числовых признаков

ячейка Jupiter Notebook

# Создаём полотно

plt.figure(figsize=[11, 13])

# Общий заголовок для всех графиков

plt.suptitle('Распределение числовых признаков ',

                y=1.005,

                fontsize=19,

                fontweight='bold')

# Рассчитаем количество строк в графике

n\_row = ceil(len(num\_cols)/3)

for i, col in enumerate(num\_cols):

    plt.subplot(n\_row, 3, i+1)

    # Заголовок для графика

    # Вставляем русский перевод если он есть, или оригинальное название

    explanations = col + '\n' + explanations\_col[col] if explanations\_col.get(col) else '\n' + col

    plt.title(f'\n{explanations}', fontsize=10)

    # Задаём размер шривта и угол поворота текста для осей X и Y

    plt.xticks(fontsize=8, rotation=0)

    plt.yticks(fontsize=8, rotation=0)

    # Делаем размер шрифта по Y=5, не убирая название оси

    plt.ylabel('', fontsize=5)

    # Отрисовываем гистограмму

    plt.hist(data\_loaded[col])

# Автоматически уместить все элементы на полотне

plt.tight\_layout()

# Вывести графики на экран

plt.show()

результат выполнения

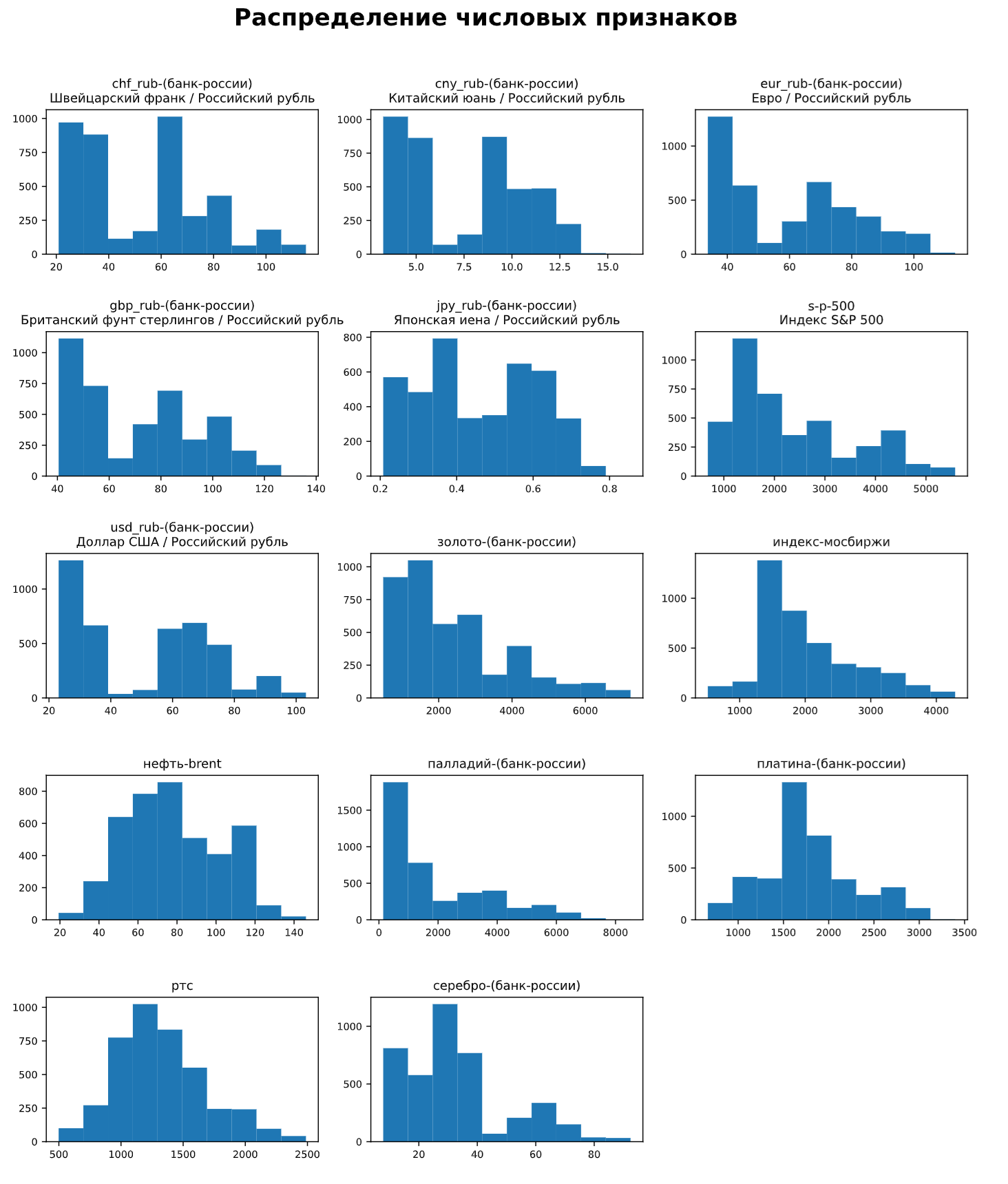


Рисунок 5. Распределение численных показателей дата фрейма

* 1. Посмотрим какие признаки имеют выбросы

ячейка Jupiter Notebook

def emission\_test(ds, col\_test, threshold\_val=3, explanations\_col=explanations\_col):

    '''

    Функция проводит тестирование столбцов на выбросы методом Z-score.

    :param ds: исследуемый датасет,

    :param col\_test: список колонок,

    :param threshold\_val: пороговое значение Z-score,

    :param explanations\_col: словарь признаков и их пояснения,

    :return: текстовый отчет и список выбросов.

    '''

    result = ''

    NUM = 2  # количество цифр после запятой

    outliers\_list = [] # лист выбросов

    for col in col\_test:

        # Вычисление Z-score

        z = np.abs(stats.zscore(ds[col]))

        # Установка порогового значения Z-score

        threshold = 3

        # Выявление выбросов на основе Z-score

        outliers = ds[col][z > threshold]

        if len(outliers) > 0:

            outliers\_list.append(outliers)

            # Вставляем русский перевод если он есть, или оригинальное название

            explanations = f' ({explanations\_col[col]})' if explanations\_col.get(col) else ''

            result += f'В столбце {col}{explanations},\n{len(outliers)} выбросов. Mean: {round(ds[col].mean(), NUM)}, ' \

            f'Min: {round(ds[col].min(), NUM)}, Max: {round(ds[col].max(), NUM)}, ' \

            f'Moda: {round(ds[col].mode()[0], NUM)}, Median: {round(ds[col].median(), NUM)} \n\n'

    return f'Выбросов нет. ' if result == '' else result, outliers\_list

ячейка Jupiter Notebook

def viev\_outliers(outliers\_list, col\_name):

    '''

    Функция выводит строки датафрейма, в которых есть выбросы.

    :param outliers\_list: список выбросов,

    :param col\_name: название признака (фичи),

    :return: индексы выбросов.

    '''

    for item in outliers:

        if item.name == col\_name:

            return item.index

ячейка Jupiter Notebook

# Выведем признаки которые имеют выбросы

text\_outliers, outliers = emission\_test(ds=data\_loaded, col\_test=num\_cols)

# Выведем текстовый отчёт по выбросам

print(text\_outliers)

результат выполнения

В столбце золото-(банк-россии),

11 выбросов. Mean: 2406.81, Min: 486.82, Max: 7231.12, Moda: 5993.16, Median: 2246.76

В столбце индекс-мосбиржи,

3 выбросов. Mean: 2017.93, Min: 513.62, Max: 4287.52, Moda: 647.8, Median: 1765.4

В столбце палладий-(банк-россии),

6 выбросов. Mean: 1889.29, Min: 144.32, Max: 8507.4, Moda: 3275.71, Median: 1235.6

В столбце платина-(банк-россии),

1 выбросов. Mean: 1775.77, Min: 664.78, Max: 3396.33, Moda: 2883.54, Median: 1714.11

В столбце ртс,

11 выбросов. Mean: 1325.14, Min: 498.2, Max: 2487.92, Moda: 982.94, Median: 1274.45

В столбце серебро-(банк-россии),

21 выбросов. Mean: 33.01, Min: 7.81, Max: 92.42, Moda: 32.11, Median: 31.27

Каких то аномалий в выбросах не наблюдается, данные корректны. При применении модели RandomForestRegressor сглаживание выбросов с помощью логарифмической функции или другими методами, результат предсказаний не улучшает. Оставим дата фрейм без изменений.

* 1. Создание новых признаков

Создадим дополнительные признаки год, квартал, месяц, неделя, день, которые в дальнейшем помогут нам при выполнении запросов.

ячейка Jupiter Notebook

def choice\_quarter(month):

    '''

    Функция определения квартала.

    :param month: номер месяца 1-12,

    :return: номер квартала 1-4, или -1, если ошибка.

    '''

    if month <= 3:

        return 1

    elif 4 <= month <= 6:

        return 2

    elif 7 <= month <= 9:

        return 3

    elif 10 <= month <= 12:

        return 4

    else:

        return -1

# Создадим дополнительные признаки (год, месяц, день, день недели (Пн-Вс), номер недели, квартал)

data\_loaded['Year'] = data\_loaded['Дата'].dt.year

data\_loaded['Month'] = data\_loaded['Дата'].dt.month

data\_loaded['Weekday'] = data\_loaded['Дата'].dt.weekday

data\_loaded['Day'] = data\_loaded['Дата'].dt.day

data\_loaded['Week'] = data\_loaded['Дата'].dt.isocalendar().week

data\_loaded['Quarter'] = data\_loaded['Month'].apply(lambda x: choice\_quarter(x))

data\_loaded.head()

результат выполнения

Изображение выглядит как текст, снимок экрана, линия, Шрифт

Автоматически созданное описание

Рисунок 6. Первые пять строк дата фрейма с созданными новыми признаками

* 1. Проверка состояния данных

Проверим за какие временные периоды собраны данные и если понадобиться удалим выбранные периоды.

ячейка Jupiter Notebook

def not\_correct\_year(df=data\_loaded, min\_month=12, min\_day=200):

    '''

    Функция выдает список не корректных годов, где min\_month и min\_day меньше заданного.

    :param  df: исследуемый датафрейм,

    :param  min\_month: минимальное количество месяцев в году,

    :param  min\_day: минимальное количество дней в году,

    :return: возвращает список не корректных годов и текстовый отчёт.

    '''

    not\_corr\_year = []

    txt\_report = ''

    for year in df['Year'].unique():

        month\_quantity = len(df[df['Year']==year]['Month'].unique())

        day\_quantity = len(df[df['Year']==year]['Weekday'])

        if (month\_quantity < min\_month) or (day\_quantity < min\_day):

            not\_corr\_year.append(int(year))

        txt\_report += f'Год: {year}, кол. месяцев: {month\_quantity}, кол. дней: {day\_quantity}\n'

    return not\_corr\_year, txt\_report

# Протестируем датасет и запишем список не корректных годов и отчёт

not\_correct\_year\_list, report = not\_correct\_year()

print(report)

результат выполнения

Год: 2024, кол. месяцев: 7, кол. дней: 124

Год: 2023, кол. месяцев: 12, кол. дней: 237

Год: 2022, кол. месяцев: 12, кол. дней: 214

Год: 2021, кол. месяцев: 12, кол. дней: 233

Год: 2020, кол. месяцев: 12, кол. дней: 225

Год: 2019, кол. месяцев: 12, кол. дней: 232

Год: 2018, кол. месяцев: 12, кол. дней: 227

Год: 2017, кол. месяцев: 12, кол. дней: 232

Год: 2016, кол. месяцев: 12, кол. дней: 232

Год: 2015, кол. месяцев: 12, кол. дней: 231

Год: 2014, кол. месяцев: 12, кол. дней: 233

Год: 2013, кол. месяцев: 12, кол. дней: 232

Год: 2012, кол. месяцев: 12, кол. дней: 230

Год: 2011, кол. месяцев: 12, кол. дней: 232

Год: 2010, кол. месяцев: 12, кол. дней: 224

Год: 2009, кол. месяцев: 12, кол. дней: 230

Год: 2008, кол. месяцев: 12, кол. дней: 226

Год: 2007, кол. месяцев: 12, кол. дней: 227

Год: 2006, кол. месяцев: 8, кол. дней: 157

Видно, что за 2006 и 2024 собрана информация не полностью за год, но удалять эти года не будем просто будем иметь в виду этот момент при кросс-валидации.

* 1. Метрика, по которой будем оценивать качество работы модели

Метрика будет вычислять какова величина ошибки в процентах между тестовыми средними значениями за определённой период и предсказанными средними значениями, формула ниже [3].

()

Где – среднее значение курса валют за определённый отрезок времени (неделя, месяц, квартал, год);

– предсказанное среднее значение курса валют за определённый отрезок времени (неделя, месяц, квартал, год);

– величина ошибки предсказанных средних значений в процентах.

ячейка Jupiter Notebook

# Метрика

def mape(y\_test, y\_pred):

    '''

    Функция вычисляет ошибку в процентах от средних значений предсказанных и тестовых.

    :param  y\_test: тестовые значения таргета,

    :param  y\_pred: предсказанные значения таргета,

    :return: возвращает ошибку предсказания в процентах.

    '''

    return abs((y\_test.mean() - y\_pred.mean()) / y\_test.mean()) \* 100

* 1. Разбиение временного ряда на train и test

Функция-генератор split\_on\_time\_series разбивает временной ряд на train и test. Класс TimeSeriesSplit из библиотеки sklearn не подошел так ка он разбивает временной ряд по числу элементов [14]. Поэтому создадим собственную функцию, которая будет разбивать временной ряд привязываясь к датам [20; 12; 17]. Функция также исправляет мелкие орфографические ошибки в указании типа разбиений (Year, Quarter, Month, Week) и выводит ошибку если количество разбиений задано не корректно.

ячейка Jupiter Notebook

def split\_on\_time\_series(data, start\_year, number\_splits, type\_split, target\_name):

    '''

    Функция-генератор, разбивает временной ряд на train и test.

    :param  data: датафрем с данными,

    :param  start\_year: год с которого будем разбивать,

    :param  number\_splits: количество разбиений,

    :param  type\_split: тип разбиений (Year, Quarter, Month, Week),

    :param  target\_name: имя таргета (целевой переменной),

    :return: возвращает X\_train, y\_train, X\_test, y\_test, summary, time\_point\_test.

    '''

    # Удаляем все не латинские знаки, делаем первую букву заглавную, остальные обычные

    resample\_param = re.sub(r'[^a-zA-z]', '', type\_split).capitalize()

    resample\_code = ''

    # Создадим список валют

    currency\_list = [column for column in data if '\_' in column]

    # Тригер первого включения

    triger = False

    # Выставляем параметр для сводной таблицы

    if resample\_param == 'Year':

        resample\_code = 'YE'

    elif resample\_param == 'Quarter':

        resample\_code = 'BQE'

    elif resample\_param == 'Month':

        resample\_code = 'ME'

    elif resample\_param == 'Week':

        resample\_code = 'W'

    else:

        raise ValueError(f'Ошибка параметра "type\_split={type\_split}". Корректно: Year, Quarter, Month, Week')

    # Делаем датафрем независимым

    df = data.copy(deep=True)

    # Сделаем резервную копию даты

    df['Дата-copy'] = df['Дата']

    # Установим столбец даты в качестве индекса

    df = df.set\_index('Дата-copy')

    # Выберем данные после стартового года

    df = df.query(f'(Year >= {start\_year})')

    # Создадим X\_train и y\_train

    # Делаем датафрем независимым

    X\_train = data.copy(deep=True)

    X\_train = X\_train.query(f'(Year < {start\_year})')

    # Установим столбец даты в качестве индекса

    X\_train = X\_train.set\_index('Дата')

    # Создадим таргет

    y\_train = X\_train[target\_name]

    # Удалим все ненужные столбцы из X\_train

    X\_train.drop(columns=currency\_list, axis=1, inplace=True)

    # Проверка

    n\_row = df.resample(resample\_code).count().shape[0]

    if n\_row < number\_splits:

        raise ValueError(f'Столько данных с параметрами "start\_year={start\_year}", не найдём ((')

    # Если всё ок разбиваем на train и test

    for row in range(number\_splits):

        # Добавляем данные из предыдущего цикла, если он не первый

        if triger:

            X\_train = pd.concat([X\_train, X\_test]).sort\_index(ascending=False)

            y\_train = pd.concat([y\_train, y\_test]).sort\_index(ascending=False)

        else:

            triger = True

        # Читаем строчку сводной таблицы

        resample\_param\_list = ['Year']

        if not resample\_param in resample\_param\_list:

            resample\_param\_list.append(resample\_param)

        pivot\_table\_row = pd.DataFrame(df.resample(resample\_code).mean().iloc[row]).T[resample\_param\_list]

        # Соединяем таблицу и строку сводной таблицы

        result = df.merge(pivot\_table\_row, on=resample\_param\_list, how='right')

        result = result.set\_index('Дата')

        # Запишем таргет в отдельную переменную

        y\_test = result[target\_name]

        # Удалим все таргеты из датасета

        result.drop(columns=currency\_list, axis=1, inplace=True)

        X\_test = result

        # Инфа по данным

        summary = f'Train({X\_train.index.min().strftime('%d.%m.%Y')}-{X\_train.index.max().strftime('%d.%m.%Y')}, X.shape{X\_train.shape}, y.shape{y\_train.shape}), ' \

                  f'Test({X\_test.index.min().strftime('%d.%m.%Y')}-{X\_test.index.max().strftime('%d.%m.%Y')}, X.shape{X\_test.shape}, y.shape{y\_test.shape})'

        # Запишем переменные времени, для графика

        time\_point\_test = X\_test.index.min()

        yield X\_train, y\_train, X\_test, y\_test, summary, time\_point\_test

* 1. Функция кросс-валидации

Функция кросс-валидации обучает модель на train данных и выполняет предсказание на test данных, которая модель не видела. Разбивка на train и test происходит с учетом временного ряда и привязана к датам. Ошибка предсказания оценивается по формуле 2 (пункт 2.12). Данные, на которых будет тестироваться модель, могут быть разного объёма (год, квартал, месяц, неделя), это зависит от параметра type\_split. В качестве целевой переменной (таргета) указывается список валют – параметр currency. Чтобы была информация о ходе процесса, данные выводятся с помощью команды print. Также данные о кросс-валидации, сохраняются в дата фрейм Pandas, параметр cv\_info.

ячейка Jupiter Notebook

def cross\_validation\_time\_series(model, df, year, n\_split, type\_split, currency, cv\_info):

    '''

    Функция кросс-валидирует список валют на временном ряду.

    :param  model: ML модель которую будем использовать для кросс-валидации,

    :param  df: датафрейм с данными,

    :param  year: год с которого нужно начать кросс-валидацию,

    :param  n\_split: число кросс-валидаций,

    :param  type\_split: тип кросс валидаций (Year, Quarter, Month, Week),

    :param  currency: список полей валют,

    :param  cv\_info: дата фрейм pandas куда будет записываться статистика о кросс-валидациях,

    :return: возвращает дата фрейм статистики кросс-валидаций.

    '''

    for currency\_item in currency:

        for  X\_train, y\_train, X\_test, y\_test, summary, time\_point\_test in split\_on\_time\_series(df, year, n\_split, type\_split, currency\_item):

            # Обучим модель

            model.fit(X\_train, y\_train)

            # Выполним предикт

            y\_pred = model.predict(X\_test)

            # Посмотрим эффективность модели

            cv\_error = mape(y\_test, y\_pred)

            # Запишем статистику кросс валидации

            cv\_info.loc[len(cv\_info)]  = {'date':time\_point\_test, 'currency':currency\_item, 'type\_split':type\_split, 'error':cv\_error}

            # Сделал, чтобы выводилась статистика о ходе процесса

            print(f'{summary}, \nОшибка: {round(cv\_error, 2)} %,  Валюта: {re.sub(r'[^a-zA-z\_]', '', currency\_item)}\n')

    return cv\_info

* 1. Создание модели и дата фрейма для хранения результатов кросс-валидации

Создадим модель машинного обучения которая будет использоваться при кросс-валидации, а также дата фрейм где будут храниться результаты кросс-валидации.

ячейка Jupiter Notebook

# Создадим модель с количеством деревьев в лесу 2000, с максимальной глубиной залегания дерева 21

rf\_model = RandomForestRegressor(n\_estimators=2000, max\_depth=21, random\_state=42, criterion='squared\_error')

# Обновим список валют

currency\_list = [column for column in data\_loaded if '\_' in column]

# Датафрейм статистики кросс-валидации

cv\_info = pd.DataFrame({'date':[], 'currency':[], 'type\_split':[], 'error':[]})

* 1. Кросс-валидация данных

Будем производить кросс-валидацию данных на временном ряду с количеством проверок на тестовых данных равным 10. Тестовые данные будут взяты за год, квартал, месяц, неделю.

Оцениваться качество работы модели будет по метрике mape описанной в пункте 2.11, для каждой валюты отдельно. Взяв среднее значение этой метрики по 10 попыткам, получим более точные результаты. Более подробно о кросс-валидации на временном ряду описано в пункте 1.4.

* + 1. С шагом в год

Запустим кросс-валидацию данных data\_loaded с применением модели rf\_model, начиная с 2013 года, генерить тестовую и тренировочные выборки будем 10 раз, с шагом тестовой выборки 1 год, по каждой валюте отдельно из списка валют currency\_list, сохранять результаты кросс-валидации будем в дата фрейм cv\_info.

ячейка Jupiter Notebook

%%time

# Будем кросс-валидировать 10 раз по году

cv\_info = cross\_validation\_time\_series(rf\_model, data\_loaded, 2013, 10, 'Year', currency\_list, cv\_info)

В процессе кросс-валидации выводиться информация для Train и Test - диапазон дат выборки, размерность фичей и таргетов, ошибка предсказания рассчитывается по метрике mapе описанной в пункте 2.11, так же выводиться наименование курса валюты.

результат выполнения

Train(11.05.2006-28.12.2012, X.shape(1526, 14), y.shape(1526,)), Test(10.01.2013-30.12.2013, X.shape(232, 14), y.shape(232,)),

Ошибка: 5.29 %, Валюта: chf\_rub

Train(11.05.2006-30.12.2013, X.shape(1758, 14), y.shape(1758,)), Test(10.01.2014-30.12.2014, X.shape(233, 14), y.shape(233,)),

Ошибка: 18.19 %, Валюта: chf\_rub

Train(11.05.2006-30.12.2014, X.shape(1991, 14), y.shape(1991,)), Test(13.01.2015-30.12.2015, X.shape(231, 14), y.shape(231,)),

Ошибка: 11.26 %, Валюта: chf\_rub

Train(11.05.2006-30.12.2015, X.shape(2222, 14), y.shape(2222,)), Test(12.01.2016-30.12.2016, X.shape(232, 14), y.shape(232,)),

Ошибка: 4.85 %, Валюта: chf\_rub

Train(11.05.2006-30.12.2016, X.shape(2454, 14), y.shape(2454,)), Test(10.01.2017-29.12.2017, X.shape(232, 14), y.shape(232,)),

Ошибка: 7.86 %, Валюта: chf\_rub

Train(11.05.2006-29.12.2017, X.shape(2686, 14), y.shape(2686,)), Test(10.01.2018-28.12.2018, X.shape(227, 14), y.shape(227,)),

Ошибка: 3.66 %, Валюта: chf\_rub

Train(11.05.2006-28.12.2018, X.shape(2913, 14), y.shape(2913,)), Test(10.01.2019-30.12.2019, X.shape(232, 14), y.shape(232,)),

Ошибка: 6.55 %, Валюта: chf\_rub

Train(11.05.2006-30.12.2019, X.shape(3145, 14), y.shape(3145,)), Test(10.01.2020-30.12.2020, X.shape(225, 14), y.shape(225,)),

Ошибка: 13.72 %, Валюта: chf\_rub

Train(11.05.2006-30.12.2020, X.shape(3370, 14), y.shape(3370,)), Test(12.01.2021-30.12.2021, X.shape(233, 14), y.shape(233,)),

Ошибка: 0.6 %, Валюта: chf\_rub

Train(11.05.2006-30.12.2021, X.shape(3603, 14), y.shape(3603,)), Test(11.01.2022-30.12.2022, X.shape(214, 14), y.shape(214,)),

Ошибка: 6.0 %, Валюта: chf\_rub

Train(11.05.2006-28.12.2012, X.shape(1526, 14), y.shape(1526,)), Test(10.01.2013-30.12.2013, X.shape(232, 14), y.shape(232,)),

Ошибка: 10.15 %, Валюта: cny\_rub

Train(11.05.2006-30.12.2013, X.shape(1758, 14), y.shape(1758,)), Test(10.01.2014-30.12.2014, X.shape(233, 14), y.shape(233,)),

Ошибка: 16.41 %, Валюта: cny\_rub

Train(11.05.2006-30.12.2014, X.shape(1991, 14), y.shape(1991,)), Test(13.01.2015-30.12.2015, X.shape(231, 14), y.shape(231,)),

Ошибка: 10.55 %, Валюта: cny\_rub

Train(11.05.2006-30.12.2015, X.shape(2222, 14), y.shape(2222,)), Test(12.01.2016-30.12.2016, X.shape(232, 14), y.shape(232,)),

Ошибка: 2.82 %, Валюта: cny\_rub

Train(11.05.2006-30.12.2016, X.shape(2454, 14), y.shape(2454,)), Test(10.01.2017-29.12.2017, X.shape(232, 14), y.shape(232,)),

Ошибка: 8.63 %, Валюта: cny\_rub

Train(11.05.2006-29.12.2017, X.shape(2686, 14), y.shape(2686,)), Test(10.01.2018-28.12.2018, X.shape(227, 14), y.shape(227,)),

Ошибка: 6.05 %, Валюта: cny\_rub

Train(11.05.2006-28.12.2018, X.shape(2913, 14), y.shape(2913,)), Test(10.01.2019-30.12.2019, X.shape(232, 14), y.shape(232,)),

Ошибка: 6.36 %, Валюта: cny\_rub

Train(11.05.2006-30.12.2019, X.shape(3145, 14), y.shape(3145,)), Test(10.01.2020-30.12.2020, X.shape(225, 14), y.shape(225,)),

Ошибка: 11.95 %, Валюта: cny\_rub

Train(11.05.2006-30.12.2020, X.shape(3370, 14), y.shape(3370,)), Test(12.01.2021-30.12.2021, X.shape(233, 14), y.shape(233,)),

Ошибка: 3.73 %, Валюта: cny\_rub

Train(11.05.2006-30.12.2021, X.shape(3603, 14), y.shape(3603,)), Test(11.01.2022-30.12.2022, X.shape(214, 14), y.shape(214,)),

Ошибка: 5.99 %, Валюта: cny\_rub

Train(11.05.2006-28.12.2012, X.shape(1526, 14), y.shape(1526,)), Test(10.01.2013-30.12.2013, X.shape(232, 14), y.shape(232,)),

Ошибка: 5.05 %, Валюта: eur\_rub

Train(11.05.2006-30.12.2013, X.shape(1758, 14), y.shape(1758,)), Test(10.01.2014-30.12.2014, X.shape(233, 14), y.shape(233,)),

Ошибка: 16.35 %, Валюта: eur\_rub

Train(11.05.2006-30.12.2014, X.shape(1991, 14), y.shape(1991,)), Test(13.01.2015-30.12.2015, X.shape(231, 14), y.shape(231,)),

Ошибка: 0.19 %, Валюта: eur\_rub

Train(11.05.2006-30.12.2015, X.shape(2222, 14), y.shape(2222,)), Test(12.01.2016-30.12.2016, X.shape(232, 14), y.shape(232,)),

Ошибка: 1.22 %, Валюта: eur\_rub

Train(11.05.2006-30.12.2016, X.shape(2454, 14), y.shape(2454,)), Test(10.01.2017-29.12.2017, X.shape(232, 14), y.shape(232,)),

Ошибка: 2.16 %, Валюта: eur\_rub

Train(11.05.2006-29.12.2017, X.shape(2686, 14), y.shape(2686,)), Test(10.01.2018-28.12.2018, X.shape(227, 14), y.shape(227,)),

Ошибка: 7.02 %, Валюта: eur\_rub

Train(11.05.2006-28.12.2018, X.shape(2913, 14), y.shape(2913,)), Test(10.01.2019-30.12.2019, X.shape(232, 14), y.shape(232,)),

Ошибка: 7.34 %, Валюта: eur\_rub

Train(11.05.2006-30.12.2019, X.shape(3145, 14), y.shape(3145,)), Test(10.01.2020-30.12.2020, X.shape(225, 14), y.shape(225,)),

Ошибка: 10.42 %, Валюта: eur\_rub

Train(11.05.2006-30.12.2020, X.shape(3370, 14), y.shape(3370,)), Test(12.01.2021-30.12.2021, X.shape(233, 14), y.shape(233,)),

Ошибка: 1.03 %, Валюта: eur\_rub

Train(11.05.2006-30.12.2021, X.shape(3603, 14), y.shape(3603,)), Test(11.01.2022-30.12.2022, X.shape(214, 14), y.shape(214,)),

Ошибка: 15.41 %, Валюта: eur\_rub

Train(11.05.2006-28.12.2012, X.shape(1526, 14), y.shape(1526,)), Test(10.01.2013-30.12.2013, X.shape(232, 14), y.shape(232,)),

Ошибка: 2.1 %, Валюта: gbp\_rub

Train(11.05.2006-30.12.2013, X.shape(1758, 14), y.shape(1758,)), Test(10.01.2014-30.12.2014, X.shape(233, 14), y.shape(233,)),

Ошибка: 19.79 %, Валюта: gbp\_rub

Train(11.05.2006-30.12.2014, X.shape(1991, 14), y.shape(1991,)), Test(13.01.2015-30.12.2015, X.shape(231, 14), y.shape(231,)),

Ошибка: 9.43 %, Валюта: gbp\_rub

Train(11.05.2006-30.12.2015, X.shape(2222, 14), y.shape(2222,)), Test(12.01.2016-30.12.2016, X.shape(232, 14), y.shape(232,)),

Ошибка: 13.05 %, Валюта: gbp\_rub

Train(11.05.2006-30.12.2016, X.shape(2454, 14), y.shape(2454,)), Test(10.01.2017-29.12.2017, X.shape(232, 14), y.shape(232,)),

Ошибка: 4.42 %, Валюта: gbp\_rub

Train(11.05.2006-29.12.2017, X.shape(2686, 14), y.shape(2686,)), Test(10.01.2018-28.12.2018, X.shape(227, 14), y.shape(227,)),

Ошибка: 8.71 %, Валюта: gbp\_rub

Train(11.05.2006-28.12.2018, X.shape(2913, 14), y.shape(2913,)), Test(10.01.2019-30.12.2019, X.shape(232, 14), y.shape(232,)),

Ошибка: 7.7 %, Валюта: gbp\_rub

Train(11.05.2006-30.12.2019, X.shape(3145, 14), y.shape(3145,)), Test(10.01.2020-30.12.2020, X.shape(225, 14), y.shape(225,)),

Ошибка: 11.74 %, Валюта: gbp\_rub

Train(11.05.2006-30.12.2020, X.shape(3370, 14), y.shape(3370,)), Test(12.01.2021-30.12.2021, X.shape(233, 14), y.shape(233,)),

Ошибка: 5.23 %, Валюта: gbp\_rub

Train(11.05.2006-30.12.2021, X.shape(3603, 14), y.shape(3603,)), Test(11.01.2022-30.12.2022, X.shape(214, 14), y.shape(214,)),

Ошибка: 13.33 %, Валюта: gbp\_rub

Train(11.05.2006-28.12.2012, X.shape(1526, 14), y.shape(1526,)), Test(10.01.2013-30.12.2013, X.shape(232, 14), y.shape(232,)),

Ошибка: 12.72 %, Валюта: jpy\_rub

Train(11.05.2006-30.12.2013, X.shape(1758, 14), y.shape(1758,)), Test(10.01.2014-30.12.2014, X.shape(233, 14), y.shape(233,)),

Ошибка: 6.77 %, Валюта: jpy\_rub

Train(11.05.2006-30.12.2014, X.shape(1991, 14), y.shape(1991,)), Test(13.01.2015-30.12.2015, X.shape(231, 14), y.shape(231,)),

Ошибка: 9.23 %, Валюта: jpy\_rub

Train(11.05.2006-30.12.2015, X.shape(2222, 14), y.shape(2222,)), Test(12.01.2016-30.12.2016, X.shape(232, 14), y.shape(232,)),

Ошибка: 11.7 %, Валюта: jpy\_rub

Train(11.05.2006-30.12.2016, X.shape(2454, 14), y.shape(2454,)), Test(10.01.2017-29.12.2017, X.shape(232, 14), y.shape(232,)),

Ошибка: 2.93 %, Валюта: jpy\_rub

Train(11.05.2006-29.12.2017, X.shape(2686, 14), y.shape(2686,)), Test(10.01.2018-28.12.2018, X.shape(227, 14), y.shape(227,)),

Ошибка: 1.93 %, Валюта: jpy\_rub

Train(11.05.2006-28.12.2018, X.shape(2913, 14), y.shape(2913,)), Test(10.01.2019-30.12.2019, X.shape(232, 14), y.shape(232,)),

Ошибка: 2.27 %, Валюта: jpy\_rub

Train(11.05.2006-30.12.2019, X.shape(3145, 14), y.shape(3145,)), Test(10.01.2020-30.12.2020, X.shape(225, 14), y.shape(225,)),

Ошибка: 8.41 %, Валюта: jpy\_rub

Train(11.05.2006-30.12.2020, X.shape(3370, 14), y.shape(3370,)), Test(12.01.2021-30.12.2021, X.shape(233, 14), y.shape(233,)),

Ошибка: 4.12 %, Валюта: jpy\_rub

Train(11.05.2006-30.12.2021, X.shape(3603, 14), y.shape(3603,)), Test(11.01.2022-30.12.2022, X.shape(214, 14), y.shape(214,)),

Ошибка: 27.96 %, Валюта: jpy\_rub

Train(11.05.2006-28.12.2012, X.shape(1526, 14), y.shape(1526,)), Test(10.01.2013-30.12.2013, X.shape(232, 14), y.shape(232,)),

Ошибка: 4.1 %, Валюта: usd\_rub

Train(11.05.2006-30.12.2013, X.shape(1758, 14), y.shape(1758,)), Test(10.01.2014-30.12.2014, X.shape(233, 14), y.shape(233,)),

Ошибка: 16.43 %, Валюта: usd\_rub

Train(11.05.2006-30.12.2014, X.shape(1991, 14), y.shape(1991,)), Test(13.01.2015-30.12.2015, X.shape(231, 14), y.shape(231,)),

Ошибка: 12.09 %, Валюта: usd\_rub

Train(11.05.2006-30.12.2015, X.shape(2222, 14), y.shape(2222,)), Test(12.01.2016-30.12.2016, X.shape(232, 14), y.shape(232,)),

Ошибка: 1.67 %, Валюта: usd\_rub

Train(11.05.2006-30.12.2016, X.shape(2454, 14), y.shape(2454,)), Test(10.01.2017-29.12.2017, X.shape(232, 14), y.shape(232,)),

Ошибка: 7.68 %, Валюта: usd\_rub

Train(11.05.2006-29.12.2017, X.shape(2686, 14), y.shape(2686,)), Test(10.01.2018-28.12.2018, X.shape(227, 14), y.shape(227,)),

Ошибка: 5.14 %, Валюта: usd\_rub

Train(11.05.2006-28.12.2018, X.shape(2913, 14), y.shape(2913,)), Test(10.01.2019-30.12.2019, X.shape(232, 14), y.shape(232,)),

Ошибка: 2.86 %, Валюта: usd\_rub

Train(11.05.2006-30.12.2019, X.shape(3145, 14), y.shape(3145,)), Test(10.01.2020-30.12.2020, X.shape(225, 14), y.shape(225,)),

Ошибка: 6.42 %, Валюта: usd\_rub

Train(11.05.2006-30.12.2020, X.shape(3370, 14), y.shape(3370,)), Test(12.01.2021-30.12.2021, X.shape(233, 14), y.shape(233,)),

Ошибка: 0.98 %, Валюта: usd\_rub

Train(11.05.2006-30.12.2021, X.shape(3603, 14), y.shape(3603,)), Test(11.01.2022-30.12.2022, X.shape(214, 14), y.shape(214,)),

Ошибка: 6.21 %, Валюта: usd\_rub

CPU times: total: 22min 12s

Wall time: 32min 44s

* + 1. С шагом в квартал

Запустим кросс-валидацию данных data\_loaded с применением модели rf\_model, начиная с 2022 года, генерить тестовую и тренировочные выборки будем 10 раз, с шагом тестовой выборки 1 квартал, по каждой валюте отдельно из списка валют currency\_list, сохранять результаты кросс-валидации будем в дата фрейм cv\_info.

ячейка Jupiter Notebook

%%time

# Будем кросс-валидировать 10 раз по кварталу

cv\_info = cross\_validation\_time\_series(rf\_model, data\_loaded, 2022, 10, 'Quarter', currency\_list, cv\_info)

В процессе кросс-валидации выводиться информация для Train и Test - диапазон дат выборки, размерность фичей и таргетов, ошибка предсказания рассчитывается по метрике mapе описанной в пункте 2.11, так же выводиться наименование курса валюты.

результат выполнения

Train(11.05.2006-30.12.2021, X.shape(3603, 14), y.shape(3603,)), Test(11.01.2022-31.03.2022, X.shape(35, 14), y.shape(35,)),

Ошибка: 6.52 %, Валюта: chf\_rub

Train(11.05.2006-31.03.2022, X.shape(3638, 14), y.shape(3638,)), Test(01.04.2022-30.06.2022, X.shape(54, 14), y.shape(54,)),

Ошибка: 12.28 %, Валюта: chf\_rub

Train(11.05.2006-30.06.2022, X.shape(3692, 14), y.shape(3692,)), Test(01.07.2022-30.09.2022, X.shape(64, 14), y.shape(64,)),

Ошибка: 5.32 %, Валюта: chf\_rub

Train(11.05.2006-30.09.2022, X.shape(3756, 14), y.shape(3756,)), Test(03.10.2022-30.12.2022, X.shape(61, 14), y.shape(61,)),

Ошибка: 2.75 %, Валюта: chf\_rub

Train(11.05.2006-30.12.2022, X.shape(3817, 14), y.shape(3817,)), Test(10.01.2023-31.03.2023, X.shape(52, 14), y.shape(52,)),

Ошибка: 1.32 %, Валюта: chf\_rub

Train(11.05.2006-31.03.2023, X.shape(3869, 14), y.shape(3869,)), Test(03.04.2023-30.06.2023, X.shape(59, 14), y.shape(59,)),

Ошибка: 3.19 %, Валюта: chf\_rub

Train(11.05.2006-30.06.2023, X.shape(3928, 14), y.shape(3928,)), Test(03.07.2023-29.09.2023, X.shape(63, 14), y.shape(63,)),

Ошибка: 9.4 %, Валюта: chf\_rub

Train(11.05.2006-29.09.2023, X.shape(3991, 14), y.shape(3991,)), Test(02.10.2023-29.12.2023, X.shape(63, 14), y.shape(63,)),

Ошибка: 2.66 %, Валюта: chf\_rub

Train(11.05.2006-29.12.2023, X.shape(4054, 14), y.shape(4054,)), Test(03.01.2024-28.03.2024, X.shape(58, 14), y.shape(58,)),

Ошибка: 1.2 %, Валюта: chf\_rub

Train(11.05.2006-28.03.2024, X.shape(4112, 14), y.shape(4112,)), Test(01.04.2024-28.06.2024, X.shape(60, 14), y.shape(60,)),

Ошибка: 5.04 %, Валюта: chf\_rub

Train(11.05.2006-30.12.2021, X.shape(3603, 14), y.shape(3603,)), Test(11.01.2022-31.03.2022, X.shape(35, 14), y.shape(35,)),

Ошибка: 7.96 %, Валюта: cny\_rub

Train(11.05.2006-31.03.2022, X.shape(3638, 14), y.shape(3638,)), Test(01.04.2022-30.06.2022, X.shape(54, 14), y.shape(54,)),

Ошибка: 9.06 %, Валюта: cny\_rub

Train(11.05.2006-30.06.2022, X.shape(3692, 14), y.shape(3692,)), Test(01.07.2022-30.09.2022, X.shape(64, 14), y.shape(64,)),

Ошибка: 6.03 %, Валюта: cny\_rub

Train(11.05.2006-30.09.2022, X.shape(3756, 14), y.shape(3756,)), Test(03.10.2022-30.12.2022, X.shape(61, 14), y.shape(61,)),

Ошибка: 4.86 %, Валюта: cny\_rub

Train(11.05.2006-30.12.2022, X.shape(3817, 14), y.shape(3817,)), Test(10.01.2023-31.03.2023, X.shape(52, 14), y.shape(52,)),

Ошибка: 6.87 %, Валюта: cny\_rub

Train(11.05.2006-31.03.2023, X.shape(3869, 14), y.shape(3869,)), Test(03.04.2023-30.06.2023, X.shape(59, 14), y.shape(59,)),

Ошибка: 8.33 %, Валюта: cny\_rub

Train(11.05.2006-30.06.2023, X.shape(3928, 14), y.shape(3928,)), Test(03.07.2023-29.09.2023, X.shape(63, 14), y.shape(63,)),

Ошибка: 1.53 %, Валюта: cny\_rub

Train(11.05.2006-29.09.2023, X.shape(3991, 14), y.shape(3991,)), Test(02.10.2023-29.12.2023, X.shape(63, 14), y.shape(63,)),

Ошибка: 1.33 %, Валюта: cny\_rub

Train(11.05.2006-29.12.2023, X.shape(4054, 14), y.shape(4054,)), Test(03.01.2024-28.03.2024, X.shape(58, 14), y.shape(58,)),

Ошибка: 0.84 %, Валюта: cny\_rub

Train(11.05.2006-28.03.2024, X.shape(4112, 14), y.shape(4112,)), Test(01.04.2024-28.06.2024, X.shape(60, 14), y.shape(60,)),

Ошибка: 3.27 %, Валюта: cny\_rub

Train(11.05.2006-30.12.2021, X.shape(3603, 14), y.shape(3603,)), Test(11.01.2022-31.03.2022, X.shape(35, 14), y.shape(35,)),

Ошибка: 2.49 %, Валюта: eur\_rub

Train(11.05.2006-31.03.2022, X.shape(3638, 14), y.shape(3638,)), Test(01.04.2022-30.06.2022, X.shape(54, 14), y.shape(54,)),

Ошибка: 15.38 %, Валюта: eur\_rub

Train(11.05.2006-30.06.2022, X.shape(3692, 14), y.shape(3692,)), Test(01.07.2022-30.09.2022, X.shape(64, 14), y.shape(64,)),

Ошибка: 6.11 %, Валюта: eur\_rub

Train(11.05.2006-30.09.2022, X.shape(3756, 14), y.shape(3756,)), Test(03.10.2022-30.12.2022, X.shape(61, 14), y.shape(61,)),

Ошибка: 0.7 %, Валюта: eur\_rub

Train(11.05.2006-30.12.2022, X.shape(3817, 14), y.shape(3817,)), Test(10.01.2023-31.03.2023, X.shape(52, 14), y.shape(52,)),

Ошибка: 3.65 %, Валюта: eur\_rub

Train(11.05.2006-31.03.2023, X.shape(3869, 14), y.shape(3869,)), Test(03.04.2023-30.06.2023, X.shape(59, 14), y.shape(59,)),

Ошибка: 3.95 %, Валюта: eur\_rub

Train(11.05.2006-30.06.2023, X.shape(3928, 14), y.shape(3928,)), Test(03.07.2023-29.09.2023, X.shape(63, 14), y.shape(63,)),

Ошибка: 3.93 %, Валюта: eur\_rub

Train(11.05.2006-29.09.2023, X.shape(3991, 14), y.shape(3991,)), Test(02.10.2023-29.12.2023, X.shape(63, 14), y.shape(63,)),

Ошибка: 2.86 %, Валюта: eur\_rub

Train(11.05.2006-29.12.2023, X.shape(4054, 14), y.shape(4054,)), Test(03.01.2024-28.03.2024, X.shape(58, 14), y.shape(58,)),

Ошибка: 0.32 %, Валюта: eur\_rub

Train(11.05.2006-28.03.2024, X.shape(4112, 14), y.shape(4112,)), Test(01.04.2024-28.06.2024, X.shape(60, 14), y.shape(60,)),

Ошибка: 2.27 %, Валюта: eur\_rub

Train(11.05.2006-30.12.2021, X.shape(3603, 14), y.shape(3603,)), Test(11.01.2022-31.03.2022, X.shape(35, 14), y.shape(35,)),

Ошибка: 5.03 %, Валюта: gbp\_rub

Train(11.05.2006-31.03.2022, X.shape(3638, 14), y.shape(3638,)), Test(01.04.2022-30.06.2022, X.shape(54, 14), y.shape(54,)),

Ошибка: 13.19 %, Валюта: gbp\_rub

Train(11.05.2006-30.06.2022, X.shape(3692, 14), y.shape(3692,)), Test(01.07.2022-30.09.2022, X.shape(64, 14), y.shape(64,)),

Ошибка: 9.22 %, Валюта: gbp\_rub

Train(11.05.2006-30.09.2022, X.shape(3756, 14), y.shape(3756,)), Test(03.10.2022-30.12.2022, X.shape(61, 14), y.shape(61,)),

Ошибка: 5.41 %, Валюта: gbp\_rub

Train(11.05.2006-30.12.2022, X.shape(3817, 14), y.shape(3817,)), Test(10.01.2023-31.03.2023, X.shape(52, 14), y.shape(52,)),

Ошибка: 3.79 %, Валюта: gbp\_rub

Train(11.05.2006-31.03.2023, X.shape(3869, 14), y.shape(3869,)), Test(03.04.2023-30.06.2023, X.shape(59, 14), y.shape(59,)),

Ошибка: 4.3 %, Валюта: gbp\_rub

Train(11.05.2006-30.06.2023, X.shape(3928, 14), y.shape(3928,)), Test(03.07.2023-29.09.2023, X.shape(63, 14), y.shape(63,)),

Ошибка: 4.0 %, Валюта: gbp\_rub

Train(11.05.2006-29.09.2023, X.shape(3991, 14), y.shape(3991,)), Test(02.10.2023-29.12.2023, X.shape(63, 14), y.shape(63,)),

Ошибка: 3.55 %, Валюта: gbp\_rub

Train(11.05.2006-29.12.2023, X.shape(4054, 14), y.shape(4054,)), Test(03.01.2024-28.03.2024, X.shape(58, 14), y.shape(58,)),

Ошибка: 0.95 %, Валюта: gbp\_rub

Train(11.05.2006-28.03.2024, X.shape(4112, 14), y.shape(4112,)), Test(01.04.2024-28.06.2024, X.shape(60, 14), y.shape(60,)),

Ошибка: 2.24 %, Валюта: gbp\_rub

Train(11.05.2006-30.12.2021, X.shape(3603, 14), y.shape(3603,)), Test(11.01.2022-31.03.2022, X.shape(35, 14), y.shape(35,)),

Ошибка: 2.69 %, Валюта: jpy\_rub

Train(11.05.2006-31.03.2022, X.shape(3638, 14), y.shape(3638,)), Test(01.04.2022-30.06.2022, X.shape(54, 14), y.shape(54,)),

Ошибка: 27.38 %, Валюта: jpy\_rub

Train(11.05.2006-30.06.2022, X.shape(3692, 14), y.shape(3692,)), Test(01.07.2022-30.09.2022, X.shape(64, 14), y.shape(64,)),

Ошибка: 8.99 %, Валюта: jpy\_rub

Train(11.05.2006-30.09.2022, X.shape(3756, 14), y.shape(3756,)), Test(03.10.2022-30.12.2022, X.shape(61, 14), y.shape(61,)),

Ошибка: 3.28 %, Валюта: jpy\_rub

Train(11.05.2006-30.12.2022, X.shape(3817, 14), y.shape(3817,)), Test(10.01.2023-31.03.2023, X.shape(52, 14), y.shape(52,)),

Ошибка: 3.38 %, Валюта: jpy\_rub

Train(11.05.2006-31.03.2023, X.shape(3869, 14), y.shape(3869,)), Test(03.04.2023-30.06.2023, X.shape(59, 14), y.shape(59,)),

Ошибка: 2.49 %, Валюта: jpy\_rub

Train(11.05.2006-30.06.2023, X.shape(3928, 14), y.shape(3928,)), Test(03.07.2023-29.09.2023, X.shape(63, 14), y.shape(63,)),

Ошибка: 10.2 %, Валюта: jpy\_rub

Train(11.05.2006-29.09.2023, X.shape(3991, 14), y.shape(3991,)), Test(02.10.2023-29.12.2023, X.shape(63, 14), y.shape(63,)),

Ошибка: 3.11 %, Валюта: jpy\_rub

Train(11.05.2006-29.12.2023, X.shape(4054, 14), y.shape(4054,)), Test(03.01.2024-28.03.2024, X.shape(58, 14), y.shape(58,)),

Ошибка: 2.04 %, Валюта: jpy\_rub

Train(11.05.2006-28.03.2024, X.shape(4112, 14), y.shape(4112,)), Test(01.04.2024-28.06.2024, X.shape(60, 14), y.shape(60,)),

Ошибка: 5.83 %, Валюта: jpy\_rub

Train(11.05.2006-30.12.2021, X.shape(3603, 14), y.shape(3603,)), Test(11.01.2022-31.03.2022, X.shape(35, 14), y.shape(35,)),

Ошибка: 6.08 %, Валюта: usd\_rub

Train(11.05.2006-31.03.2022, X.shape(3638, 14), y.shape(3638,)), Test(01.04.2022-30.06.2022, X.shape(54, 14), y.shape(54,)),

Ошибка: 15.88 %, Валюта: usd\_rub

Train(11.05.2006-30.06.2022, X.shape(3692, 14), y.shape(3692,)), Test(01.07.2022-30.09.2022, X.shape(64, 14), y.shape(64,)),

Ошибка: 4.29 %, Валюта: usd\_rub

Train(11.05.2006-30.09.2022, X.shape(3756, 14), y.shape(3756,)), Test(03.10.2022-30.12.2022, X.shape(61, 14), y.shape(61,)),

Ошибка: 2.06 %, Валюта: usd\_rub

Train(11.05.2006-30.12.2022, X.shape(3817, 14), y.shape(3817,)), Test(10.01.2023-31.03.2023, X.shape(52, 14), y.shape(52,)),

Ошибка: 2.81 %, Валюта: usd\_rub

Train(11.05.2006-31.03.2023, X.shape(3869, 14), y.shape(3869,)), Test(03.04.2023-30.06.2023, X.shape(59, 14), y.shape(59,)),

Ошибка: 0.66 %, Валюта: usd\_rub

Train(11.05.2006-30.06.2023, X.shape(3928, 14), y.shape(3928,)), Test(03.07.2023-29.09.2023, X.shape(63, 14), y.shape(63,)),

Ошибка: 6.14 %, Валюта: usd\_rub

Train(11.05.2006-29.09.2023, X.shape(3991, 14), y.shape(3991,)), Test(02.10.2023-29.12.2023, X.shape(63, 14), y.shape(63,)),

Ошибка: 2.32 %, Валюта: usd\_rub

Train(11.05.2006-29.12.2023, X.shape(4054, 14), y.shape(4054,)), Test(03.01.2024-28.03.2024, X.shape(58, 14), y.shape(58,)),

Ошибка: 1.52 %, Валюта: usd\_rub

Train(11.05.2006-28.03.2024, X.shape(4112, 14), y.shape(4112,)), Test(01.04.2024-28.06.2024, X.shape(60, 14), y.shape(60,)),

Ошибка: 1.79 %, Валюта: usd\_rub

CPU times: total: 32min 11s

Wall time: 49min 56s

* + 1. С шагом в месяц

Запустим кросс-валидацию данных data\_loaded с применением модели rf\_model, начиная с 2023 года, генерить тестовую и тренировочные выборки будем 10 раз, с шагом тестовой выборки 1 месяц, по каждой валюте отдельно из списка валют currency\_list, сохранять результаты кросс-валидации будем в дата фрейм cv\_info.

ячейка Jupiter Notebook

%%time

# Будем кросс-валидировать 10 раз по месяцу

cv\_info = cross\_validation\_time\_series(rf\_model, data\_loaded, 2023, 10, 'Month', currency\_list, cv\_info)

В процессе кросс-валидации выводиться информация для Train и Test - диапазон дат выборки, размерность фичей и таргетов, ошибка предсказания рассчитывается по метрике mapе описанной в пункте 2.11, так же выводиться наименование курса валюты.

результат выполнения

Train(11.05.2006-30.12.2022, X.shape(3817, 14), y.shape(3817,)), Test(10.01.2023-31.01.2023, X.shape(15, 14), y.shape(15,)),

Ошибка: 3.38 %, Валюта: chf\_rub

Train(11.05.2006-31.01.2023, X.shape(3832, 14), y.shape(3832,)), Test(01.02.2023-28.02.2023, X.shape(16, 14), y.shape(16,)),

Ошибка: 4.48 %, Валюта: chf\_rub

Train(11.05.2006-28.02.2023, X.shape(3848, 14), y.shape(3848,)), Test(01.03.2023-31.03.2023, X.shape(21, 14), y.shape(21,)),

Ошибка: 2.93 %, Валюта: chf\_rub

Train(11.05.2006-31.03.2023, X.shape(3869, 14), y.shape(3869,)), Test(03.04.2023-28.04.2023, X.shape(19, 14), y.shape(19,)),

Ошибка: 1.41 %, Валюта: chf\_rub

Train(11.05.2006-28.04.2023, X.shape(3888, 14), y.shape(3888,)), Test(02.05.2023-31.05.2023, X.shape(20, 14), y.shape(20,)),

Ошибка: 1.89 %, Валюта: chf\_rub

Train(11.05.2006-31.05.2023, X.shape(3908, 14), y.shape(3908,)), Test(01.06.2023-30.06.2023, X.shape(20, 14), y.shape(20,)),

Ошибка: 3.03 %, Валюта: chf\_rub

Train(11.05.2006-30.06.2023, X.shape(3928, 14), y.shape(3928,)), Test(03.07.2023-31.07.2023, X.shape(20, 14), y.shape(20,)),

Ошибка: 7.4 %, Валюта: chf\_rub

Train(11.05.2006-31.07.2023, X.shape(3948, 14), y.shape(3948,)), Test(01.08.2023-31.08.2023, X.shape(23, 14), y.shape(23,)),

Ошибка: 3.53 %, Валюта: chf\_rub

Train(11.05.2006-31.08.2023, X.shape(3971, 14), y.shape(3971,)), Test(01.09.2023-29.09.2023, X.shape(20, 14), y.shape(20,)),

Ошибка: 2.18 %, Валюта: chf\_rub

Train(11.05.2006-29.09.2023, X.shape(3991, 14), y.shape(3991,)), Test(02.10.2023-31.10.2023, X.shape(22, 14), y.shape(22,)),

Ошибка: 0.01 %, Валюта: chf\_rub

Train(11.05.2006-30.12.2022, X.shape(3817, 14), y.shape(3817,)), Test(10.01.2023-31.01.2023, X.shape(15, 14), y.shape(15,)),

Ошибка: 11.96 %, Валюта: cny\_rub

Train(11.05.2006-31.01.2023, X.shape(3832, 14), y.shape(3832,)), Test(01.02.2023-28.02.2023, X.shape(16, 14), y.shape(16,)),

Ошибка: 1.86 %, Валюта: cny\_rub

Train(11.05.2006-28.02.2023, X.shape(3848, 14), y.shape(3848,)), Test(01.03.2023-31.03.2023, X.shape(21, 14), y.shape(21,)),

Ошибка: 1.63 %, Валюта: cny\_rub

Train(11.05.2006-31.03.2023, X.shape(3869, 14), y.shape(3869,)), Test(03.04.2023-28.04.2023, X.shape(19, 14), y.shape(19,)),

Ошибка: 7.94 %, Валюта: cny\_rub

Train(11.05.2006-28.04.2023, X.shape(3888, 14), y.shape(3888,)), Test(02.05.2023-31.05.2023, X.shape(20, 14), y.shape(20,)),

Ошибка: 3.18 %, Валюта: cny\_rub

Train(11.05.2006-31.05.2023, X.shape(3908, 14), y.shape(3908,)), Test(01.06.2023-30.06.2023, X.shape(20, 14), y.shape(20,)),

Ошибка: 1.52 %, Валюта: cny\_rub

Train(11.05.2006-30.06.2023, X.shape(3928, 14), y.shape(3928,)), Test(03.07.2023-31.07.2023, X.shape(20, 14), y.shape(20,)),

Ошибка: 1.85 %, Валюта: cny\_rub

Train(11.05.2006-31.07.2023, X.shape(3948, 14), y.shape(3948,)), Test(01.08.2023-31.08.2023, X.shape(23, 14), y.shape(23,)),

Ошибка: 2.38 %, Валюта: cny\_rub

Train(11.05.2006-31.08.2023, X.shape(3971, 14), y.shape(3971,)), Test(01.09.2023-29.09.2023, X.shape(20, 14), y.shape(20,)),

Ошибка: 1.28 %, Валюта: cny\_rub

Train(11.05.2006-29.09.2023, X.shape(3991, 14), y.shape(3991,)), Test(02.10.2023-31.10.2023, X.shape(22, 14), y.shape(22,)),

Ошибка: 1.16 %, Валюта: cny\_rub

Train(11.05.2006-30.12.2022, X.shape(3817, 14), y.shape(3817,)), Test(10.01.2023-31.01.2023, X.shape(15, 14), y.shape(15,)),

Ошибка: 0.39 %, Валюта: eur\_rub

Train(11.05.2006-31.01.2023, X.shape(3832, 14), y.shape(3832,)), Test(01.02.2023-28.02.2023, X.shape(16, 14), y.shape(16,)),

Ошибка: 4.09 %, Валюта: eur\_rub

Train(11.05.2006-28.02.2023, X.shape(3848, 14), y.shape(3848,)), Test(01.03.2023-31.03.2023, X.shape(21, 14), y.shape(21,)),

Ошибка: 1.68 %, Валюта: eur\_rub

Train(11.05.2006-31.03.2023, X.shape(3869, 14), y.shape(3869,)), Test(03.04.2023-28.04.2023, X.shape(19, 14), y.shape(19,)),

Ошибка: 4.64 %, Валюта: eur\_rub

Train(11.05.2006-28.04.2023, X.shape(3888, 14), y.shape(3888,)), Test(02.05.2023-31.05.2023, X.shape(20, 14), y.shape(20,)),

Ошибка: 2.15 %, Валюта: eur\_rub

Train(11.05.2006-31.05.2023, X.shape(3908, 14), y.shape(3908,)), Test(01.06.2023-30.06.2023, X.shape(20, 14), y.shape(20,)),

Ошибка: 3.41 %, Валюта: eur\_rub

Train(11.05.2006-30.06.2023, X.shape(3928, 14), y.shape(3928,)), Test(03.07.2023-31.07.2023, X.shape(20, 14), y.shape(20,)),

Ошибка: 2.51 %, Валюта: eur\_rub

Train(11.05.2006-31.07.2023, X.shape(3948, 14), y.shape(3948,)), Test(01.08.2023-31.08.2023, X.shape(23, 14), y.shape(23,)),

Ошибка: 2.64 %, Валюта: eur\_rub

Train(11.05.2006-31.08.2023, X.shape(3971, 14), y.shape(3971,)), Test(01.09.2023-29.09.2023, X.shape(20, 14), y.shape(20,)),

Ошибка: 2.06 %, Валюта: eur\_rub

Train(11.05.2006-29.09.2023, X.shape(3991, 14), y.shape(3991,)), Test(02.10.2023-31.10.2023, X.shape(22, 14), y.shape(22,)),

Ошибка: 0.45 %, Валюта: eur\_rub

Train(11.05.2006-30.12.2022, X.shape(3817, 14), y.shape(3817,)), Test(10.01.2023-31.01.2023, X.shape(15, 14), y.shape(15,)),

Ошибка: 7.33 %, Валюта: gbp\_rub

Train(11.05.2006-31.01.2023, X.shape(3832, 14), y.shape(3832,)), Test(01.02.2023-28.02.2023, X.shape(16, 14), y.shape(16,)),

Ошибка: 0.8 %, Валюта: gbp\_rub

Train(11.05.2006-28.02.2023, X.shape(3848, 14), y.shape(3848,)), Test(01.03.2023-31.03.2023, X.shape(21, 14), y.shape(21,)),

Ошибка: 1.41 %, Валюта: gbp\_rub

Train(11.05.2006-31.03.2023, X.shape(3869, 14), y.shape(3869,)), Test(03.04.2023-28.04.2023, X.shape(19, 14), y.shape(19,)),

Ошибка: 3.51 %, Валюта: gbp\_rub

Train(11.05.2006-28.04.2023, X.shape(3888, 14), y.shape(3888,)), Test(02.05.2023-31.05.2023, X.shape(20, 14), y.shape(20,)),

Ошибка: 2.42 %, Валюта: gbp\_rub

Train(11.05.2006-31.05.2023, X.shape(3908, 14), y.shape(3908,)), Test(01.06.2023-30.06.2023, X.shape(20, 14), y.shape(20,)),

Ошибка: 4.85 %, Валюта: gbp\_rub

Train(11.05.2006-30.06.2023, X.shape(3928, 14), y.shape(3928,)), Test(03.07.2023-31.07.2023, X.shape(20, 14), y.shape(20,)),

Ошибка: 2.05 %, Валюта: gbp\_rub

Train(11.05.2006-31.07.2023, X.shape(3948, 14), y.shape(3948,)), Test(01.08.2023-31.08.2023, X.shape(23, 14), y.shape(23,)),

Ошибка: 2.61 %, Валюта: gbp\_rub

Train(11.05.2006-31.08.2023, X.shape(3971, 14), y.shape(3971,)), Test(01.09.2023-29.09.2023, X.shape(20, 14), y.shape(20,)),

Ошибка: 2.16 %, Валюта: gbp\_rub

Train(11.05.2006-29.09.2023, X.shape(3991, 14), y.shape(3991,)), Test(02.10.2023-31.10.2023, X.shape(22, 14), y.shape(22,)),

Ошибка: 1.17 %, Валюта: gbp\_rub

Train(11.05.2006-30.12.2022, X.shape(3817, 14), y.shape(3817,)), Test(10.01.2023-31.01.2023, X.shape(15, 14), y.shape(15,)),

Ошибка: 5.82 %, Валюта: jpy\_rub

Train(11.05.2006-31.01.2023, X.shape(3832, 14), y.shape(3832,)), Test(01.02.2023-28.02.2023, X.shape(16, 14), y.shape(16,)),

Ошибка: 0.34 %, Валюта: jpy\_rub

Train(11.05.2006-28.02.2023, X.shape(3848, 14), y.shape(3848,)), Test(01.03.2023-31.03.2023, X.shape(21, 14), y.shape(21,)),

Ошибка: 2.39 %, Валюта: jpy\_rub

Train(11.05.2006-31.03.2023, X.shape(3869, 14), y.shape(3869,)), Test(03.04.2023-28.04.2023, X.shape(19, 14), y.shape(19,)),

Ошибка: 0.65 %, Валюта: jpy\_rub

Train(11.05.2006-28.04.2023, X.shape(3888, 14), y.shape(3888,)), Test(02.05.2023-31.05.2023, X.shape(20, 14), y.shape(20,)),

Ошибка: 5.31 %, Валюта: jpy\_rub

Train(11.05.2006-31.05.2023, X.shape(3908, 14), y.shape(3908,)), Test(01.06.2023-30.06.2023, X.shape(20, 14), y.shape(20,)),

Ошибка: 0.16 %, Валюта: jpy\_rub

Train(11.05.2006-30.06.2023, X.shape(3928, 14), y.shape(3928,)), Test(03.07.2023-31.07.2023, X.shape(20, 14), y.shape(20,)),

Ошибка: 4.88 %, Валюта: jpy\_rub

Train(11.05.2006-31.07.2023, X.shape(3948, 14), y.shape(3948,)), Test(01.08.2023-31.08.2023, X.shape(23, 14), y.shape(23,)),

Ошибка: 3.13 %, Валюта: jpy\_rub

Train(11.05.2006-31.08.2023, X.shape(3971, 14), y.shape(3971,)), Test(01.09.2023-29.09.2023, X.shape(20, 14), y.shape(20,)),

Ошибка: 0.74 %, Валюта: jpy\_rub

Train(11.05.2006-29.09.2023, X.shape(3991, 14), y.shape(3991,)), Test(02.10.2023-31.10.2023, X.shape(22, 14), y.shape(22,)),

Ошибка: 0.41 %, Валюта: jpy\_rub

Train(11.05.2006-30.12.2022, X.shape(3817, 14), y.shape(3817,)), Test(10.01.2023-31.01.2023, X.shape(15, 14), y.shape(15,)),

Ошибка: 7.9 %, Валюта: usd\_rub

Train(11.05.2006-31.01.2023, X.shape(3832, 14), y.shape(3832,)), Test(01.02.2023-28.02.2023, X.shape(16, 14), y.shape(16,)),

Ошибка: 3.89 %, Валюта: usd\_rub

Train(11.05.2006-28.02.2023, X.shape(3848, 14), y.shape(3848,)), Test(01.03.2023-31.03.2023, X.shape(21, 14), y.shape(21,)),

Ошибка: 2.22 %, Валюта: usd\_rub

Train(11.05.2006-31.03.2023, X.shape(3869, 14), y.shape(3869,)), Test(03.04.2023-28.04.2023, X.shape(19, 14), y.shape(19,)),

Ошибка: 2.29 %, Валюта: usd\_rub

Train(11.05.2006-28.04.2023, X.shape(3888, 14), y.shape(3888,)), Test(02.05.2023-31.05.2023, X.shape(20, 14), y.shape(20,)),

Ошибка: 1.73 %, Валюта: usd\_rub

Train(11.05.2006-31.05.2023, X.shape(3908, 14), y.shape(3908,)), Test(01.06.2023-30.06.2023, X.shape(20, 14), y.shape(20,)),

Ошибка: 3.34 %, Валюта: usd\_rub

Train(11.05.2006-30.06.2023, X.shape(3928, 14), y.shape(3928,)), Test(03.07.2023-31.07.2023, X.shape(20, 14), y.shape(20,)),

Ошибка: 3.42 %, Валюта: usd\_rub

Train(11.05.2006-31.07.2023, X.shape(3948, 14), y.shape(3948,)), Test(01.08.2023-31.08.2023, X.shape(23, 14), y.shape(23,)),

Ошибка: 4.0 %, Валюта: usd\_rub

Train(11.05.2006-31.08.2023, X.shape(3971, 14), y.shape(3971,)), Test(01.09.2023-29.09.2023, X.shape(20, 14), y.shape(20,)),

Ошибка: 0.24 %, Валюта: usd\_rub

Train(11.05.2006-29.09.2023, X.shape(3991, 14), y.shape(3991,)), Test(02.10.2023-31.10.2023, X.shape(22, 14), y.shape(22,)),

Ошибка: 1.08 %, Валюта: usd\_rub

CPU times: total: 35min 8s

Wall time: 50min 43s

* + 1. С шагом в неделю

Запустим кросс-валидацию данных data\_loaded с применением модели rf\_model, начиная с 2024 года, генерить тестовую и тренировочные выборки будем 10 раз, с шагом тестовой выборки 1 неделя, по каждой валюте отдельно из списка валют currency\_list, сохранять результаты кросс-валидации будем в дата фрейм cv\_info.

ячейка Jupiter Notebook

%%time

# Будем кросс-валидировать 10 раз по недели

cv\_info = cross\_validation\_time\_series(rf\_model, data\_loaded, 2024, 10, 'Week', currency\_list, cv\_info)

В процессе кросс-валидации выводиться информация для Train и Test - диапазон дат выборки, размерность фичей и таргетов, ошибка предсказания рассчитывается по метрике mapе описанной в пункте 2.11, так же выводиться наименование курса валюты.

результат выполнения

Train(11.05.2006-29.12.2023, X.shape(4054, 14), y.shape(4054,)), Test(03.01.2024-05.01.2024, X.shape(3, 14), y.shape(3,)),

Ошибка: 1.61 %, Валюта: chf\_rub

Train(11.05.2006-05.01.2024, X.shape(4057, 14), y.shape(4057,)), Test(08.01.2024-12.01.2024, X.shape(5, 14), y.shape(5,)),

Ошибка: 1.5 %, Валюта: chf\_rub

Train(11.05.2006-12.01.2024, X.shape(4062, 14), y.shape(4062,)), Test(16.01.2024-19.01.2024, X.shape(4, 14), y.shape(4,)),

Ошибка: 1.33 %, Валюта: chf\_rub

Train(11.05.2006-19.01.2024, X.shape(4066, 14), y.shape(4066,)), Test(22.01.2024-26.01.2024, X.shape(5, 14), y.shape(5,)),

Ошибка: 0.07 %, Валюта: chf\_rub

Train(11.05.2006-26.01.2024, X.shape(4071, 14), y.shape(4071,)), Test(29.01.2024-02.02.2024, X.shape(5, 14), y.shape(5,)),

Ошибка: 0.99 %, Валюта: chf\_rub

Train(11.05.2006-02.02.2024, X.shape(4076, 14), y.shape(4076,)), Test(05.02.2024-09.02.2024, X.shape(5, 14), y.shape(5,)),

Ошибка: 1.16 %, Валюта: chf\_rub

Train(11.05.2006-09.02.2024, X.shape(4081, 14), y.shape(4081,)), Test(12.02.2024-16.02.2024, X.shape(5, 14), y.shape(5,)),

Ошибка: 0.11 %, Валюта: chf\_rub

Train(11.05.2006-16.02.2024, X.shape(4086, 14), y.shape(4086,)), Test(20.02.2024-22.02.2024, X.shape(3, 14), y.shape(3,)),

Ошибка: 0.21 %, Валюта: chf\_rub

Train(11.05.2006-22.02.2024, X.shape(4089, 14), y.shape(4089,)), Test(26.02.2024-01.03.2024, X.shape(5, 14), y.shape(5,)),

Ошибка: 0.17 %, Валюта: chf\_rub

Train(11.05.2006-01.03.2024, X.shape(4094, 14), y.shape(4094,)), Test(04.03.2024-07.03.2024, X.shape(4, 14), y.shape(4,)),

Ошибка: 1.74 %, Валюта: chf\_rub

Train(11.05.2006-29.12.2023, X.shape(4054, 14), y.shape(4054,)), Test(03.01.2024-05.01.2024, X.shape(3, 14), y.shape(3,)),

Ошибка: 1.48 %, Валюта: cny\_rub

Train(11.05.2006-05.01.2024, X.shape(4057, 14), y.shape(4057,)), Test(08.01.2024-12.01.2024, X.shape(5, 14), y.shape(5,)),

Ошибка: 0.56 %, Валюта: cny\_rub

Train(11.05.2006-12.01.2024, X.shape(4062, 14), y.shape(4062,)), Test(16.01.2024-19.01.2024, X.shape(4, 14), y.shape(4,)),

Ошибка: 1.75 %, Валюта: cny\_rub

Train(11.05.2006-19.01.2024, X.shape(4066, 14), y.shape(4066,)), Test(22.01.2024-26.01.2024, X.shape(5, 14), y.shape(5,)),

Ошибка: 0.04 %, Валюта: cny\_rub

Train(11.05.2006-26.01.2024, X.shape(4071, 14), y.shape(4071,)), Test(29.01.2024-02.02.2024, X.shape(5, 14), y.shape(5,)),

Ошибка: 0.72 %, Валюта: cny\_rub

Train(11.05.2006-02.02.2024, X.shape(4076, 14), y.shape(4076,)), Test(05.02.2024-09.02.2024, X.shape(5, 14), y.shape(5,)),

Ошибка: 1.55 %, Валюта: cny\_rub

Train(11.05.2006-09.02.2024, X.shape(4081, 14), y.shape(4081,)), Test(12.02.2024-16.02.2024, X.shape(5, 14), y.shape(5,)),

Ошибка: 0.26 %, Валюта: cny\_rub

Train(11.05.2006-16.02.2024, X.shape(4086, 14), y.shape(4086,)), Test(20.02.2024-22.02.2024, X.shape(3, 14), y.shape(3,)),

Ошибка: 1.26 %, Валюта: cny\_rub

Train(11.05.2006-22.02.2024, X.shape(4089, 14), y.shape(4089,)), Test(26.02.2024-01.03.2024, X.shape(5, 14), y.shape(5,)),

Ошибка: 0.53 %, Валюта: cny\_rub

Train(11.05.2006-01.03.2024, X.shape(4094, 14), y.shape(4094,)), Test(04.03.2024-07.03.2024, X.shape(4, 14), y.shape(4,)),

Ошибка: 0.21 %, Валюта: cny\_rub

Train(11.05.2006-29.12.2023, X.shape(4054, 14), y.shape(4054,)), Test(03.01.2024-05.01.2024, X.shape(3, 14), y.shape(3,)),

Ошибка: 0.61 %, Валюта: eur\_rub

Train(11.05.2006-05.01.2024, X.shape(4057, 14), y.shape(4057,)), Test(08.01.2024-12.01.2024, X.shape(5, 14), y.shape(5,)),

Ошибка: 0.01 %, Валюта: eur\_rub

Train(11.05.2006-12.01.2024, X.shape(4062, 14), y.shape(4062,)), Test(16.01.2024-19.01.2024, X.shape(4, 14), y.shape(4,)),

Ошибка: 1.43 %, Валюта: eur\_rub

Train(11.05.2006-19.01.2024, X.shape(4066, 14), y.shape(4066,)), Test(22.01.2024-26.01.2024, X.shape(5, 14), y.shape(5,)),

Ошибка: 0.35 %, Валюта: eur\_rub

Train(11.05.2006-26.01.2024, X.shape(4071, 14), y.shape(4071,)), Test(29.01.2024-02.02.2024, X.shape(5, 14), y.shape(5,)),

Ошибка: 0.28 %, Валюта: eur\_rub

Train(11.05.2006-02.02.2024, X.shape(4076, 14), y.shape(4076,)), Test(05.02.2024-09.02.2024, X.shape(5, 14), y.shape(5,)),

Ошибка: 0.66 %, Валюта: eur\_rub

Train(11.05.2006-09.02.2024, X.shape(4081, 14), y.shape(4081,)), Test(12.02.2024-16.02.2024, X.shape(5, 14), y.shape(5,)),

Ошибка: 0.23 %, Валюта: eur\_rub

Train(11.05.2006-16.02.2024, X.shape(4086, 14), y.shape(4086,)), Test(20.02.2024-22.02.2024, X.shape(3, 14), y.shape(3,)),

Ошибка: 0.94 %, Валюта: eur\_rub

Train(11.05.2006-22.02.2024, X.shape(4089, 14), y.shape(4089,)), Test(26.02.2024-01.03.2024, X.shape(5, 14), y.shape(5,)),

Ошибка: 0.63 %, Валюта: eur\_rub

Train(11.05.2006-01.03.2024, X.shape(4094, 14), y.shape(4094,)), Test(04.03.2024-07.03.2024, X.shape(4, 14), y.shape(4,)),

Ошибка: 0.85 %, Валюта: eur\_rub

Train(11.05.2006-29.12.2023, X.shape(4054, 14), y.shape(4054,)), Test(03.01.2024-05.01.2024, X.shape(3, 14), y.shape(3,)),

Ошибка: 1.24 %, Валюта: gbp\_rub

Train(11.05.2006-05.01.2024, X.shape(4057, 14), y.shape(4057,)), Test(08.01.2024-12.01.2024, X.shape(5, 14), y.shape(5,)),

Ошибка: 0.56 %, Валюта: gbp\_rub

Train(11.05.2006-12.01.2024, X.shape(4062, 14), y.shape(4062,)), Test(16.01.2024-19.01.2024, X.shape(4, 14), y.shape(4,)),

Ошибка: 0.08 %, Валюта: gbp\_rub

Train(11.05.2006-19.01.2024, X.shape(4066, 14), y.shape(4066,)), Test(22.01.2024-26.01.2024, X.shape(5, 14), y.shape(5,)),

Ошибка: 0.09 %, Валюта: gbp\_rub

Train(11.05.2006-26.01.2024, X.shape(4071, 14), y.shape(4071,)), Test(29.01.2024-02.02.2024, X.shape(5, 14), y.shape(5,)),

Ошибка: 1.08 %, Валюта: gbp\_rub

Train(11.05.2006-02.02.2024, X.shape(4076, 14), y.shape(4076,)), Test(05.02.2024-09.02.2024, X.shape(5, 14), y.shape(5,)),

Ошибка: 0.96 %, Валюта: gbp\_rub

Train(11.05.2006-09.02.2024, X.shape(4081, 14), y.shape(4081,)), Test(12.02.2024-16.02.2024, X.shape(5, 14), y.shape(5,)),

Ошибка: 1.07 %, Валюта: gbp\_rub

Train(11.05.2006-16.02.2024, X.shape(4086, 14), y.shape(4086,)), Test(20.02.2024-22.02.2024, X.shape(3, 14), y.shape(3,)),

Ошибка: 1.05 %, Валюта: gbp\_rub

Train(11.05.2006-22.02.2024, X.shape(4089, 14), y.shape(4089,)), Test(26.02.2024-01.03.2024, X.shape(5, 14), y.shape(5,)),

Ошибка: 0.74 %, Валюта: gbp\_rub

Train(11.05.2006-01.03.2024, X.shape(4094, 14), y.shape(4094,)), Test(04.03.2024-07.03.2024, X.shape(4, 14), y.shape(4,)),

Ошибка: 0.52 %, Валюта: gbp\_rub

Train(11.05.2006-29.12.2023, X.shape(4054, 14), y.shape(4054,)), Test(03.01.2024-05.01.2024, X.shape(3, 14), y.shape(3,)),

Ошибка: 0.22 %, Валюта: jpy\_rub

Train(11.05.2006-05.01.2024, X.shape(4057, 14), y.shape(4057,)), Test(08.01.2024-12.01.2024, X.shape(5, 14), y.shape(5,)),

Ошибка: 0.17 %, Валюта: jpy\_rub

Train(11.05.2006-12.01.2024, X.shape(4062, 14), y.shape(4062,)), Test(16.01.2024-19.01.2024, X.shape(4, 14), y.shape(4,)),

Ошибка: 0.95 %, Валюта: jpy\_rub

Train(11.05.2006-19.01.2024, X.shape(4066, 14), y.shape(4066,)), Test(22.01.2024-26.01.2024, X.shape(5, 14), y.shape(5,)),

Ошибка: 0.53 %, Валюта: jpy\_rub

Train(11.05.2006-26.01.2024, X.shape(4071, 14), y.shape(4071,)), Test(29.01.2024-02.02.2024, X.shape(5, 14), y.shape(5,)),

Ошибка: 0.29 %, Валюта: jpy\_rub

Train(11.05.2006-02.02.2024, X.shape(4076, 14), y.shape(4076,)), Test(05.02.2024-09.02.2024, X.shape(5, 14), y.shape(5,)),

Ошибка: 0.83 %, Валюта: jpy\_rub

Train(11.05.2006-09.02.2024, X.shape(4081, 14), y.shape(4081,)), Test(12.02.2024-16.02.2024, X.shape(5, 14), y.shape(5,)),

Ошибка: 0.13 %, Валюта: jpy\_rub

Train(11.05.2006-16.02.2024, X.shape(4086, 14), y.shape(4086,)), Test(20.02.2024-22.02.2024, X.shape(3, 14), y.shape(3,)),

Ошибка: 0.95 %, Валюта: jpy\_rub

Train(11.05.2006-22.02.2024, X.shape(4089, 14), y.shape(4089,)), Test(26.02.2024-01.03.2024, X.shape(5, 14), y.shape(5,)),

Ошибка: 0.34 %, Валюта: jpy\_rub

Train(11.05.2006-01.03.2024, X.shape(4094, 14), y.shape(4094,)), Test(04.03.2024-07.03.2024, X.shape(4, 14), y.shape(4,)),

Ошибка: 0.89 %, Валюта: jpy\_rub

Train(11.05.2006-29.12.2023, X.shape(4054, 14), y.shape(4054,)), Test(03.01.2024-05.01.2024, X.shape(3, 14), y.shape(3,)),

Ошибка: 2.5 %, Валюта: usd\_rub

Train(11.05.2006-05.01.2024, X.shape(4057, 14), y.shape(4057,)), Test(08.01.2024-12.01.2024, X.shape(5, 14), y.shape(5,)),

Ошибка: 0.41 %, Валюта: usd\_rub

Train(11.05.2006-12.01.2024, X.shape(4062, 14), y.shape(4062,)), Test(16.01.2024-19.01.2024, X.shape(4, 14), y.shape(4,)),

Ошибка: 1.25 %, Валюта: usd\_rub

Train(11.05.2006-19.01.2024, X.shape(4066, 14), y.shape(4066,)), Test(22.01.2024-26.01.2024, X.shape(5, 14), y.shape(5,)),

Ошибка: 0.49 %, Валюта: usd\_rub

Train(11.05.2006-26.01.2024, X.shape(4071, 14), y.shape(4071,)), Test(29.01.2024-02.02.2024, X.shape(5, 14), y.shape(5,)),

Ошибка: 0.52 %, Валюта: usd\_rub

Train(11.05.2006-02.02.2024, X.shape(4076, 14), y.shape(4076,)), Test(05.02.2024-09.02.2024, X.shape(5, 14), y.shape(5,)),

Ошибка: 1.46 %, Валюта: usd\_rub

Train(11.05.2006-09.02.2024, X.shape(4081, 14), y.shape(4081,)), Test(12.02.2024-16.02.2024, X.shape(5, 14), y.shape(5,)),

Ошибка: 0.29 %, Валюта: usd\_rub

Train(11.05.2006-16.02.2024, X.shape(4086, 14), y.shape(4086,)), Test(20.02.2024-22.02.2024, X.shape(3, 14), y.shape(3,)),

Ошибка: 1.27 %, Валюта: usd\_rub

Train(11.05.2006-22.02.2024, X.shape(4089, 14), y.shape(4089,)), Test(26.02.2024-01.03.2024, X.shape(5, 14), y.shape(5,)),

Ошибка: 0.57 %, Валюта: usd\_rub

Train(11.05.2006-01.03.2024, X.shape(4094, 14), y.shape(4094,)), Test(04.03.2024-07.03.2024, X.shape(4, 14), y.shape(4,)),

Ошибка: 0.15 %, Валюта: usd\_rub

CPU times: total: 39min 36s

Wall time: 53min

* 1. Визуализация результатов кросс-валидации

Посмотрим на сколько ошибается модель машинного обучения в зависимости от длительности предсказания и валюты.

ячейка Jupiter Notebook

# Стиль легенды

legend\_font = {

    "size": 7,

    "family": "serif",

}

# Стиль заголовков

title\_font = {

    "fontsize": 12,

}

# Обновим список валют (на всякий случай)

currency\_list = [column for column in data\_loaded if '\_' in column]

# Задаём размер полотна

plt.figure(figsize=[12, 9])

# Общий заголовок для всех графиков

plt.suptitle('Ошибка предсказаний при кросс-валидации',

               y=0.97,

               fontsize=19,

               fontweight='bold')

# Строим 4-е графика 2х2

for i, type\_split\_item in enumerate(cv\_info['type\_split'].unique()):

    plt.subplot(2, 2, i+1)

    # Заголовок для графика

    split\_type\_dict = {

        'Year': 'годам',

        'Quarter':'кварталам',

        'Month':'месяцам',

        'Week':'неделям'}

    plt.title(f'кросс-валидация по {split\_type\_dict[type\_split\_item]}', fontdict=title\_font)

    # Задаём размер шривта и угол поворота текста для осей X и Y

    plt.xticks(fontsize=8, rotation=45)

    plt.yticks(fontsize=8, rotation=0)

    # Делаем размер шрифта по Y=8

    plt.ylabel('Ошибка в %', fontsize=8)

    # Задаём границы графика

    plt.ylim(-1.2, 30)

    # Задаём расстояние меду графиками

    plt.subplots\_adjust(wspace=0.3, hspace=0.5)

    # Отрисовываем графики

    for currency\_item in currency\_list:

        # Создаём датафрем по условию валюты и типа разбиения (год, месяц и тд.)

        df\_temp = cv\_info.query(f'(currency == "{currency\_item}") & (type\_split == "{type\_split\_item}")')

        # Запишем переменные в x и y

        x = df\_temp['date']

        y = df\_temp['error']

        # Вставляем русский перевод если он есть, или оригинал, если перевода нет

        legend\_text = explanations\_col[currency\_item] if explanations\_col.get(currency\_item) else currency\_item

        # Отрисовываем график и легенду

        plt.plot(x, y, label=legend\_text)

        # Выставляем стиль и положение легенды

        plt.legend(loc="upper right", prop=legend\_font)

# Вывести графики на экран

plt.show()

результат выполнения

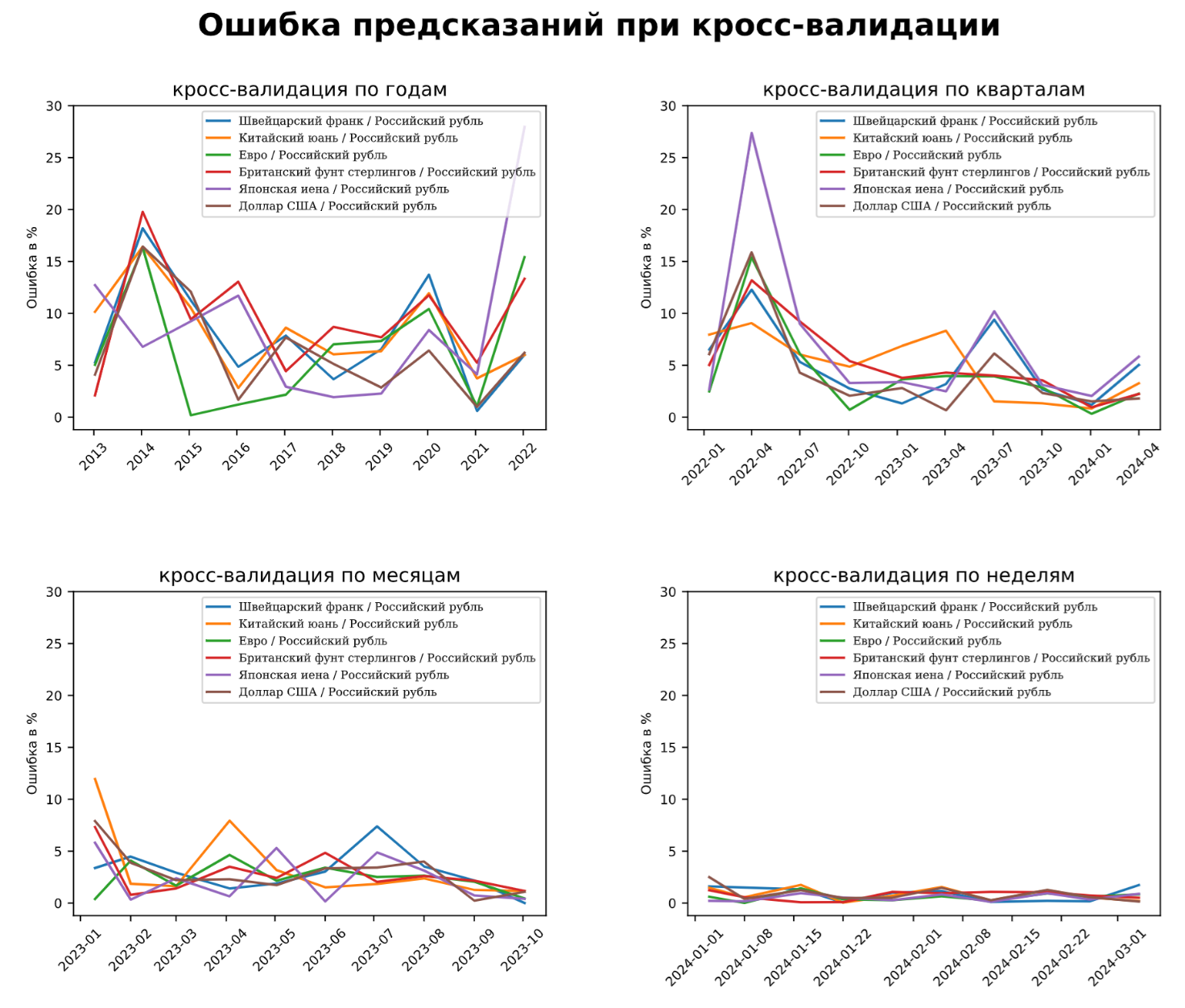


Рисунок 7. График результатов кросс-валидации

* 1. Важность признаков модели

Вычислим важность признаков для всех валют

ячейка Jupiter Notebook

%%time

# Словарь важности признаков

importance\_of\_signs = {}

# Пробежимся по всем валютам

for currency\_item in currency\_list:

    # Проитерируем функцию-генератор

    iterator = iter(split\_on\_time\_series(data\_loaded, 2024, 1, 'Year', currency\_item))

    X\_train, y\_train, X\_test, y\_test, \_, \_ = next(iterator)

    # Объединим все данные в общий датафрейм и отсортируем по индексу (индекс - это дата)

    X\_train = pd.concat([X\_train, X\_test]).sort\_index(ascending=False)

    y\_train = pd.concat([y\_train, y\_test]).sort\_index(ascending=False)

    # Обучим модель на всех данных за весь период

    rf\_model.fit(X\_train, y\_train)

    # Добавим в словарь важность признаков (ключ-валюта, значение-важность признаков)

    importance\_of\_signs[currency\_item] = rf\_model.feature\_importances\_

результат выполнения

CPU times: total: 4min 8s

Wall time: 5min 28s

ячейка Jupiter Notebook

# Посмотрим какие признаки в X\_train

X\_train

результат выполнения

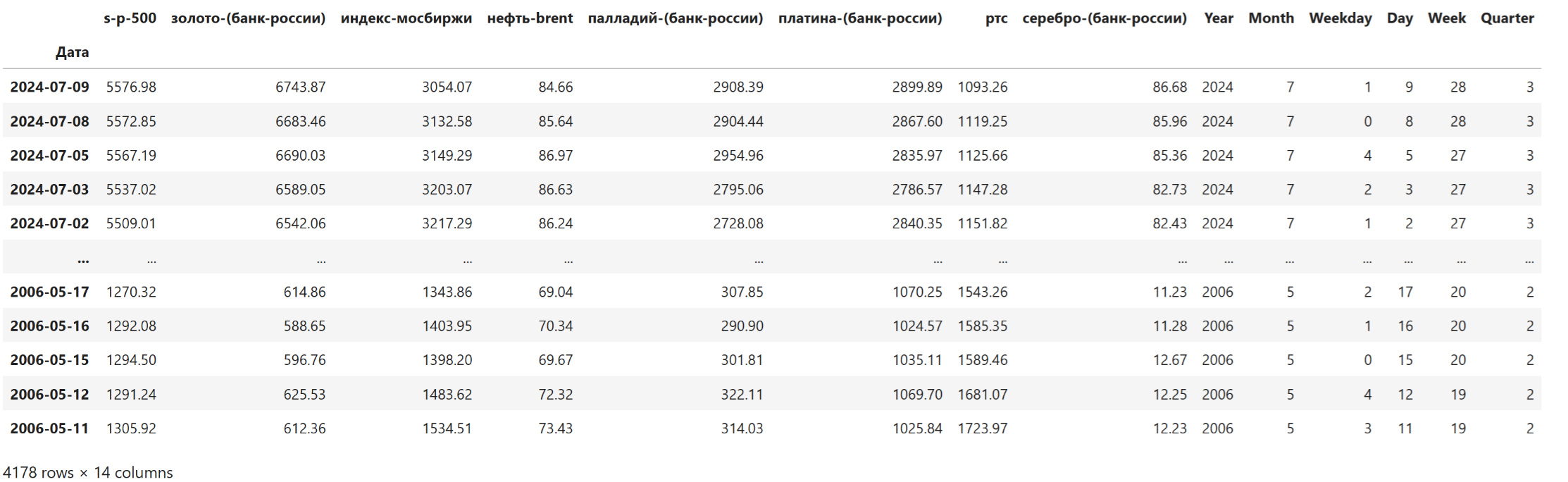


Рисунок 8. Дата фрейм с тренировочными данными

Отобразим важность признаков на графике

ячейка Jupiter Notebook

# Отрисуем графики важности для всех валют

# Стиль заголовков

title\_font = {

    "fontsize": 10,

}

# Задаём размер полотна

plt.figure(figsize=[13, 10])

# Рассчитаем количество строк в графике

n\_row = ceil(len(currency\_list)/2)

# Общий заголовок для всех графиков

plt.suptitle('Важность признаков для курса валюты',

               y=0.97,

               fontsize=19,

               fontweight='bold')

# Получим список всех фичей

feature\_names = X\_train.columns.values.tolist()

# Строим 4-е графика 2х2

for i, key in enumerate(importance\_of\_signs):

    plt.subplot(n\_row, 2, i+1)

    # Вставляем русский перевод если он есть, или оригинал, если перевода нет

    title\_text = explanations\_col[key] if explanations\_col.get(key) else key

    plt.title(title\_text, fontdict=title\_font)

    # Задаём размер шривта и угол поворота текста для осей X и Y

    plt.xticks(fontsize=8, rotation=0)

    plt.yticks(fontsize=8, rotation=0)

    # Делаем размер шрифта по Х=8 и задаём название оси

    plt.xlabel('Важность признаков', fontsize=8)

    # Задаём расстояние меду графиками

    plt.subplots\_adjust(wspace=0.6, hspace=0.45)

    # Отрисовываем графики

    plt.barh(feature\_names, importance\_of\_signs[key])

# Вывести графики на экран

plt.show()

результат выполнения

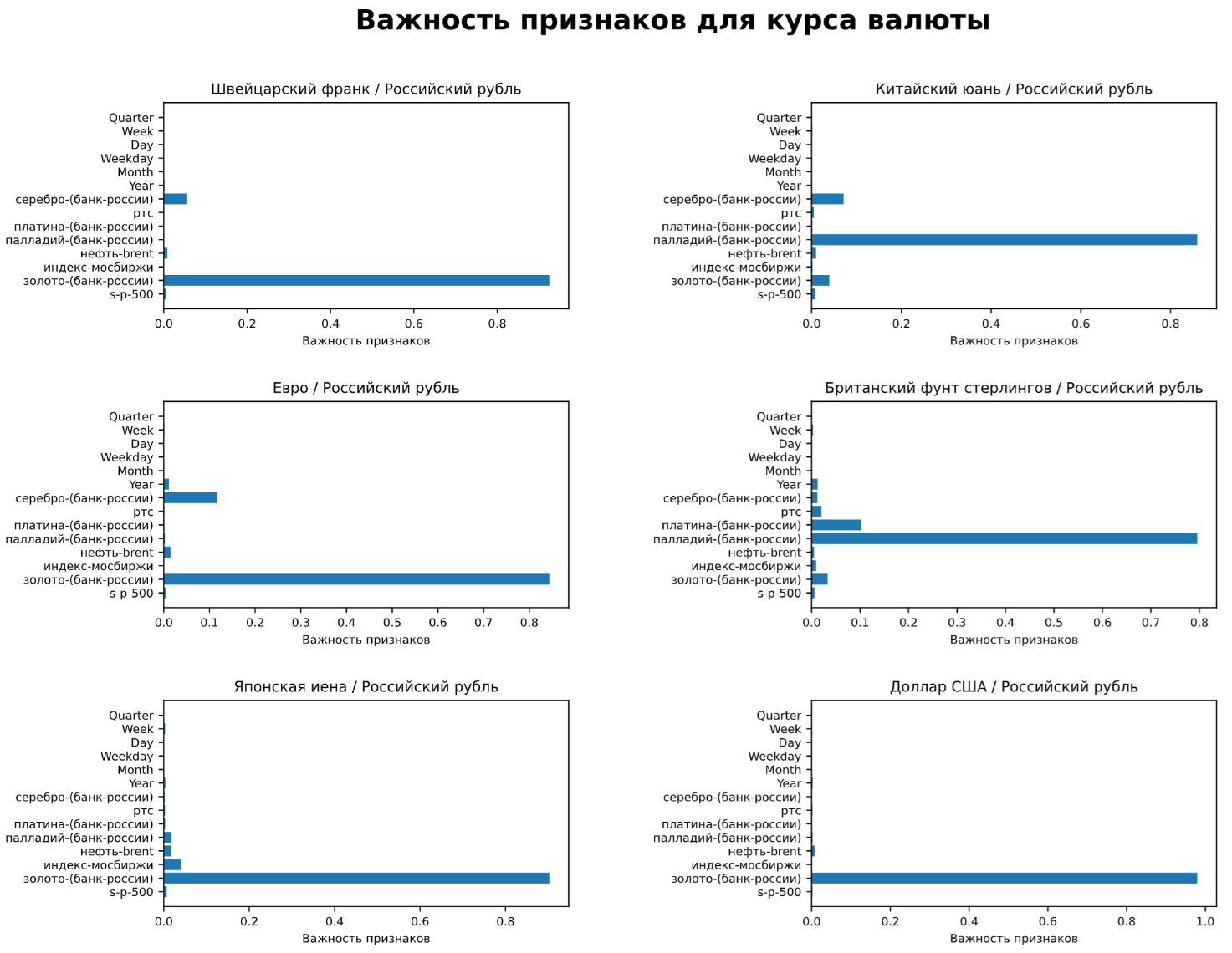


Рисунок 9. Важность признаков, влияющих на курс валюты

* 1. Средняя ошибка предсказаний модели в зависимости от длительности периода предсказаний

Посмотрим среднее значение из 10 предсказаний в зависимости от валюты и длительности предсказаний.

ячейка Jupiter Notebook

# Посмотрим, что записалось при кросс-валидации

cv\_info

результат выполнения

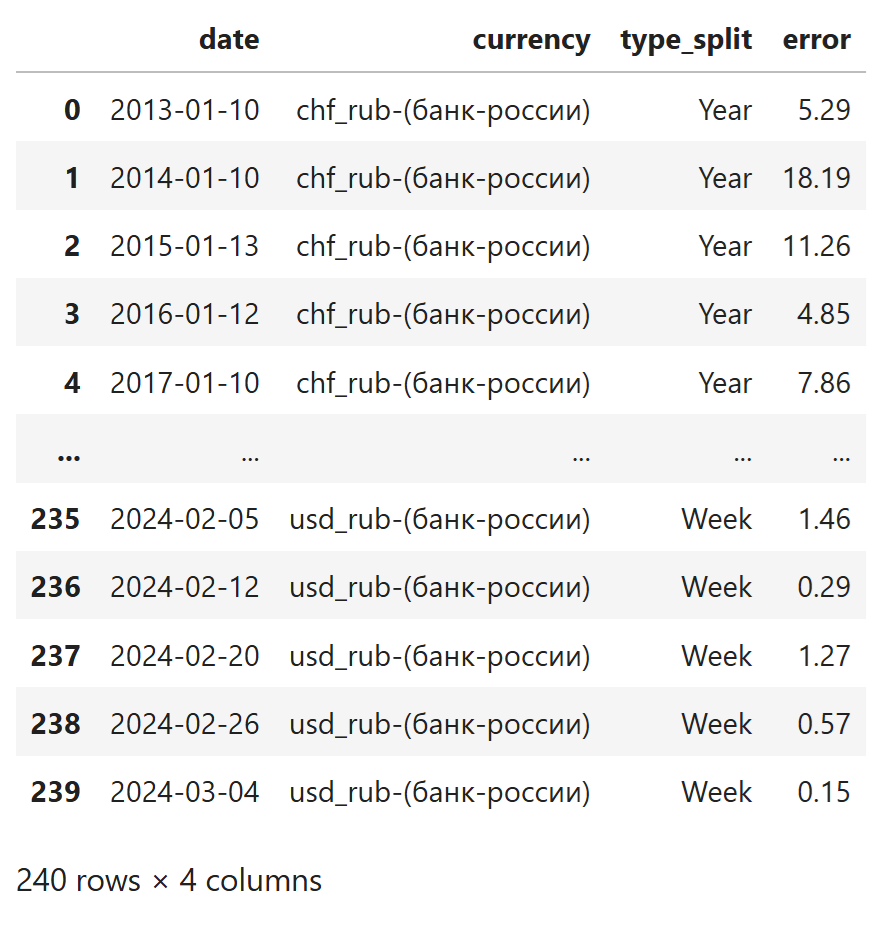


Рисунок 10. Дата фрейм с результатами кросс-валидации

ячейка Jupiter Notebook

# Сгруппируем средние значения в зависимости от типа разбиения и валюты

cv\_info\_mean = cv\_info.groupby(['type\_split', 'currency'], as\_index=False).agg({'error':'mean'}).rename(columns={'error':'error\_mean'}).sort\_values(by='error\_mean')

cv\_info\_mean

результат выполнения

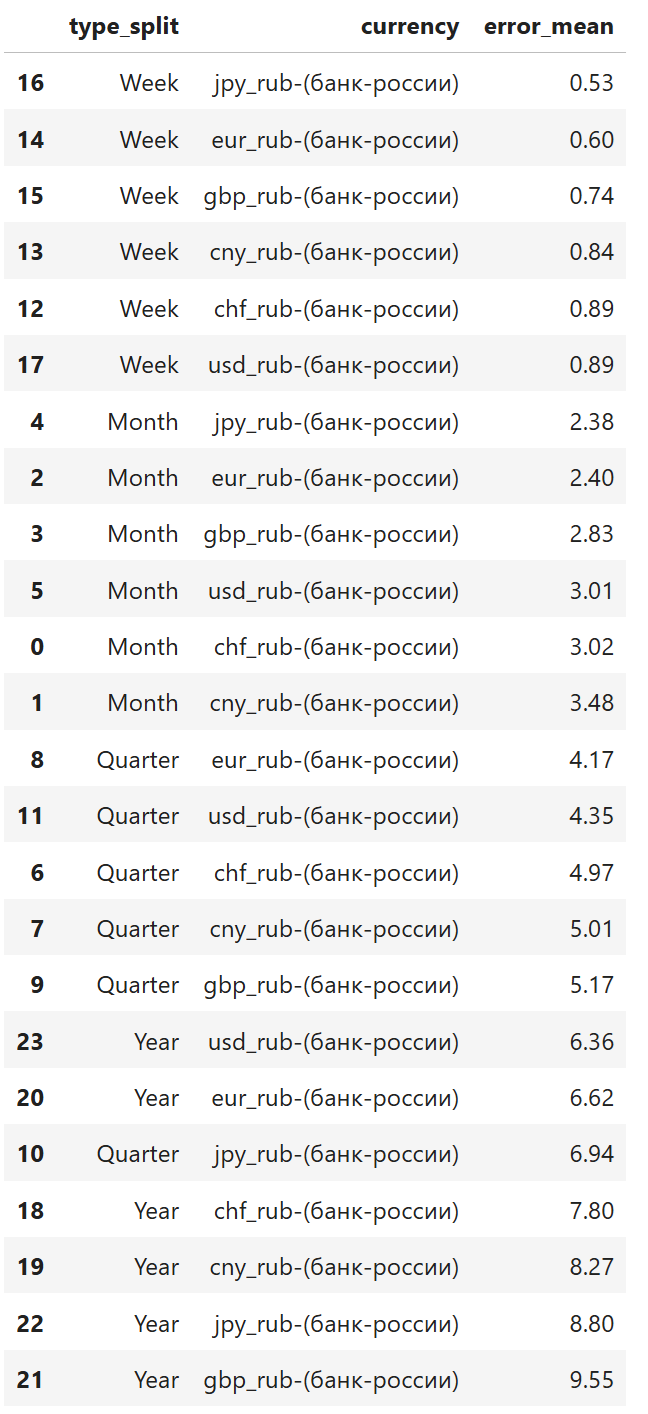


Рисунок 11. Дата фрейм со средними значениями ошибки предсказания, сгруппированный по длительности предсказания и курса валюты

ячейка Jupiter Notebook

# Отрисуем графики важности для всех валют

# Стиль заголовков

title\_font = {

    "fontsize": 10,

}

# Задаём размер полотна

plt.figure(figsize=[13, 10])

# Общий заголовок для всех графиков

plt.suptitle('Средняя ошибка предсказаний модели \n в зависимости от длительности предсказаний',

               y=1,

               fontsize=19,

               fontweight='bold')

# Рассчитаем количество строк в графике

n\_row = ceil(len(currency\_list)/2)

# Строим 4-е графика 2х2

for i, currency in enumerate(currency\_list):

    plt.subplot(n\_row, 2, i+1)

    # Вставляем русский перевод если он есть, или оригинал, если перевода нет

    title\_text = explanations\_col[currency] if explanations\_col.get(currency) else currency

    plt.title(title\_text, fontdict=title\_font)

    # Задаём размер шривта и угол поворота текста для осей X и Y

    plt.xticks(fontsize=8, rotation=0)

    plt.yticks(fontsize=8, rotation=0)

    # Делаем размер шрифта по Х=8 и задаём название оси

    plt.xlabel('Средняя ошибка предсказаний в %', fontsize=8)

    # Задаём расстояние меду графиками

    plt.subplots\_adjust(wspace=0.6, hspace=0.6)

    # Отрисовываем графики

    container = plt.barh(cv\_info\_mean.query(f'currency == "{currency}"')['type\_split'], round(cv\_info\_mean.query(f'currency == "{currency}"')['error\_mean'], 2))

    plt.bar\_label(container, padding=-10, color='black')

# Вывести графики на экран

plt.show()

результат выполнения

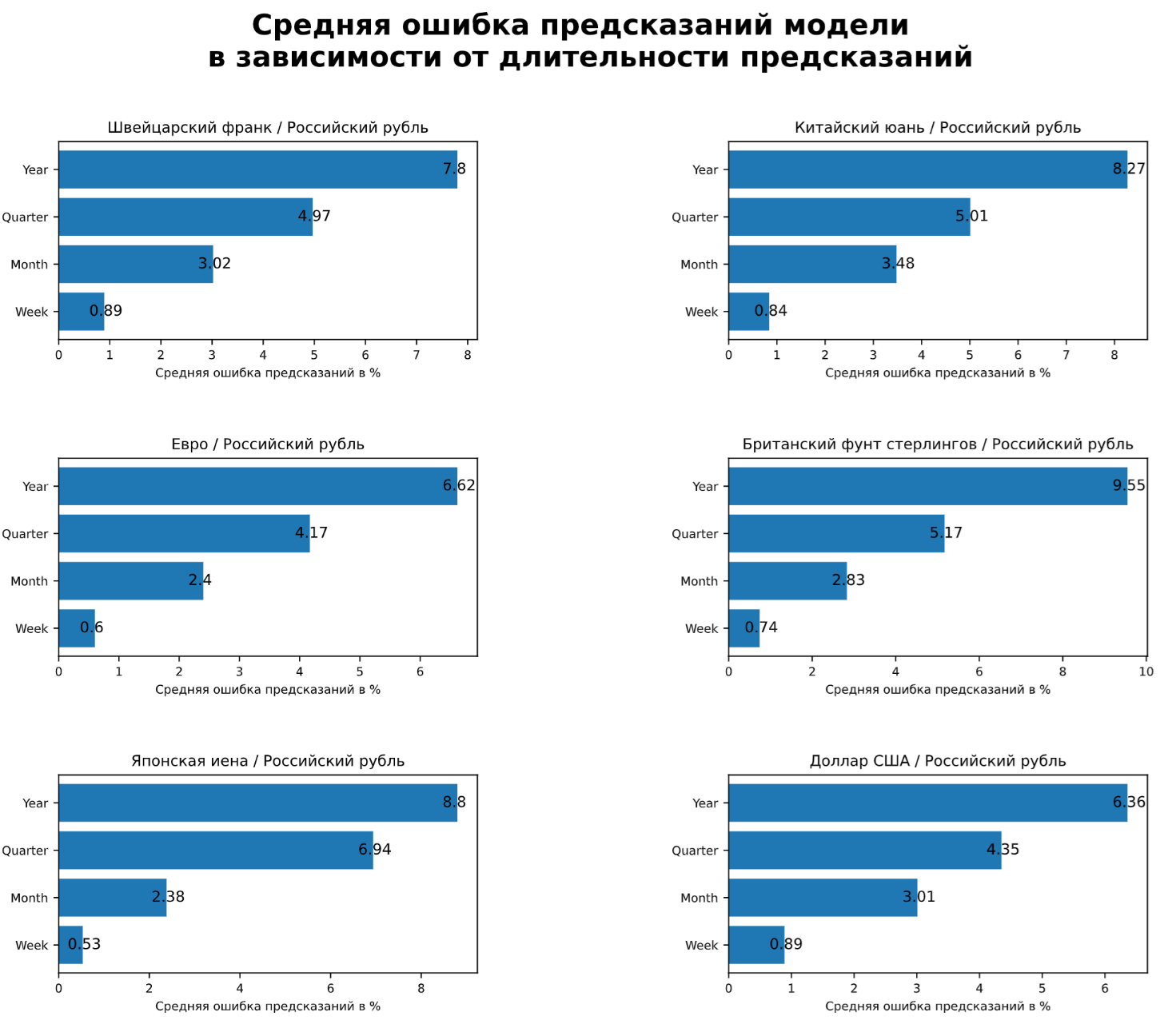


Рисунок 12. Среднее значение ошибки предсказания в зависимости от длительности предсказания

* 1. Автоматизация процесса

1. Заключение

Выполнив данное ДЗ, я научился писать скрипты для Linux, а также прокачал свои навыки в Word при оформлении данного ДЗ.

1. Список используемой литературы

1. Факторы, влияющие на формирование курса рубля Е. Я. Волков. – URL: https://view.officeapps.live.com/op/view.aspx?src=https%3A%2F%2Fwww.kubsu.ru%2Fsites%2Fdefault%2Ffiles%2Fusers%2F21365%2Fportfolio%2Fmvko\_i\_fo\_kursovaya.docx&wdOrigin=BROWSELINK (дата обращения: 11.07.2024) – Текст: электронный.

2. Zotero-style-for-Gost-R-7.0.100-2018 А. Мареев. – URL: https://github.com/ArtemMareev/Zotero-style-for-Gost-R-7.0.100-2018?tab=readme-ov-file.

3. Кросс валидация для временных рядов в python Прикладная статистика. – URL: https://www.youtube.com/watch?v=l2vhtyWp\_ck (дата обращения: 21.07.2024) – Текст: электронный.

4. Особенности методов машинного обучения для прогнозирования курса валют Е. В. Сангаджиева, Г. А. Оргдаева. – URL: https://files.scienceforum.ru/pdf/2020/5e04936d95941.pdf (дата обращения: 11.07.2024) – Текст: электронный.

5. Основы работы в Jupyter/Jupiter Notebook и JupyterLab — Python Tutorial И. Шамаев. – URL: https://python.ivan-shamaev.ru/jupiterlab-jupyter-notebook-install-python-tutorial/ (дата обращения: 19.07.2024) – Текст: электронный.

6. ДЗ «Погружение в Python». Семинар 7. Файлы и файловая система Е. В. Эгипти. – URL: https://github.com/uc20100/python\_2/blob/main/seminar\_7/file\_func/task\_home\_work.py (дата обращения: 21.07.2024) – Текст: электронный.

7. Cross-Validation на временном ряду ALEKSANDR. – URL: https://otus.ru/nest/post/483/ (дата обращения: 24.07.2024) – Текст: электронный.

8. Ключевые индикаторы InvestFunds. – URL: https://investfunds.ru/indicators/key-indexes/ (дата обращения: 20.07.2024) – Текст: электронный.

9. Курсы валют InvestFunds. – URL: https://investfunds.ru/indicators/currency/ (дата обращения: 20.07.2024) – Текст: электронный.

10. Debugger JupiterLab. – URL: https://jupyterlab.readthedocs.io/en/4.1.x/user/debugger.html (дата обращения: 19.07.2024) – Текст: электронный.

11. Site Visual Studio Code Microsoft. – URL: https://code.visualstudio.com/ (дата обращения: 24.07.2024) – Текст: электронный.

12. pandas.DataFrame.resample Pandas. – URL: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.resample.html (дата обращения: 21.07.2024) – Текст: электронный.

13. RandomForestRegressor Scikit-Learn. – URL: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html (дата обращения: 24.07.2024) – Текст: электронный.

14. TimeSeriesSplit sklearn. – URL: https://scikit-learn.org/stable/modules/generated/sklearn.model\_selection.TimeSeriesSplit.html (дата обращения: 21.07.2024) – Текст: электронный.

15. Оформление списка литературы по ГОСТ Р 7.0.100-2018 при помощи Зотеро Space G. – URL: https://www.youtube.com/watch?v=oT0ZpKnhUtI (дата обращения: 24.07.2024) – Текст: электронный.

16. Site Zotero Zotero. – URL: https://www.zotero.org/ (дата обращения: 24.07.2024) – Текст: электронный.

17. Как легко обрабатывать данные временных рядов. – URL: https://dfedorov.spb.ru/pandas/09.%20%D0%9A%D0%B0%D0%BA%20%D0%BB%D0%B5%D0%B3%D0%BA%D0%BE%20%D0%BE%D0%B1%D1%80%D0%B0%D0%B1%D0%B0%D1%82%D1%8B%D0%B2%D0%B0%D1%82%D1%8C%20%D0%B4%D0%B0%D0%BD%D0%BD%D1%8B%D0%B5%20%D0%B2%D1%80%D0%B5%D0%BC%D0%B5%D0%BD%D0%BD%D1%8B%D1%85%20%D1%80%D1%8F%D0%B4%D0%BE%D0%B2\_.html (дата обращения: 21.07.2024) – Текст: электронный.

18. Anaconda. – URL: https://www.anaconda.com/ (дата обращения: 19.07.2024) – Текст: электронный.

19. Word Microsoft 365. – URL: https://www.microsoft.com/ru-ru/microsoft-365/word?market=ru (дата обращения: 24.07.2024) – Текст: электронный.

20. Time series / date functionality Pandas. – URL: https://pandas.pydata.org/docs/user\_guide/timeseries.html (дата обращения: 21.07.2024) – Текст: электронный.