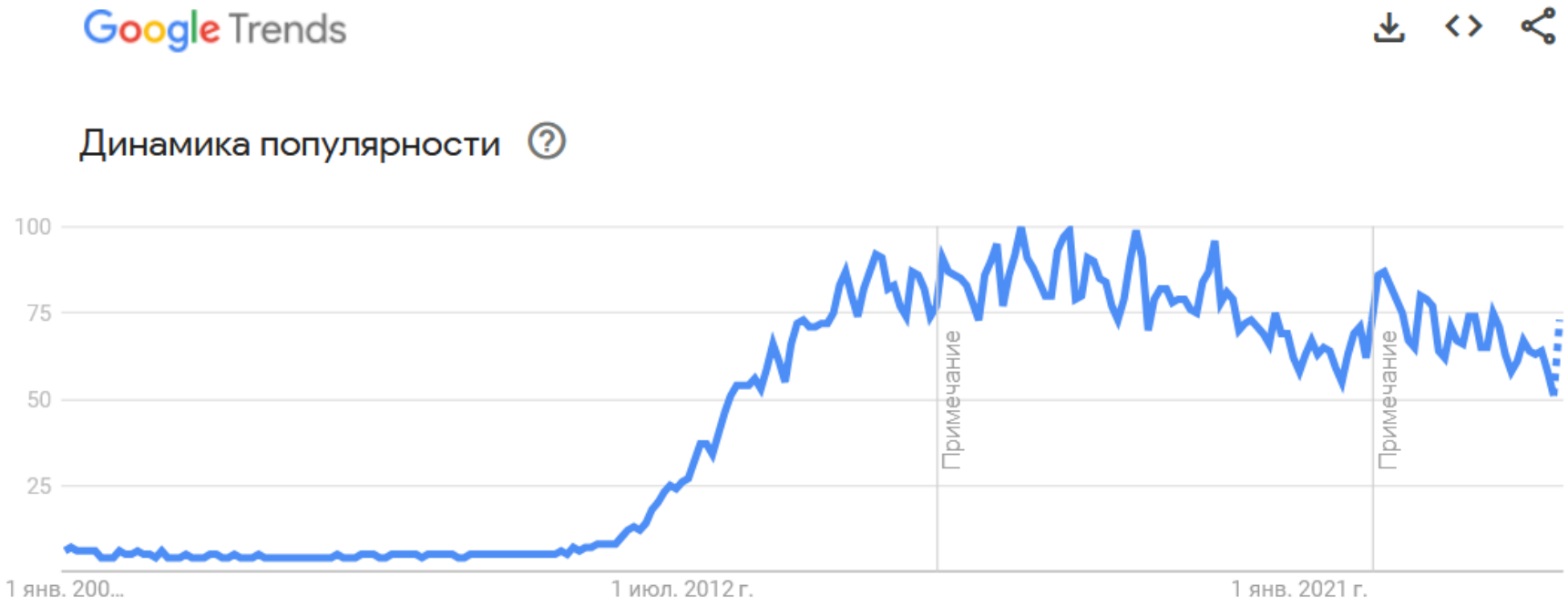
Машинное обучение

SKLearn

https://trends.google.com/



Термины

- Большие данные – Наука о данных
- Машинное обучение
- Искусственный интеллект
  - Artificial Intelligence
- Нейросети – Многослойные нейросети – Глубокое обучение

Machine Learning

- Supervised Learning
- Обучение с учителем
- Labelled data
- Размеченные данные
- Unsupervised Learning
- Обучение без учителя
- Reinforced learning
- Обучение с подкреплением

Профессии

- Data Engineer
- Data Analyst
- Data Scientist
- ML Engineer
- …

iPython

- Jupyter Notebook
  - Text Cell - Markdown
  - Code Cell

- Anaconda

- Google Colab
- Colab.research.google.com

Ячейки

- # Заголовок
- Пуск – Shift + Enter

- Github.com/valentin-arkov/
- Dataset-z
- Raw
- !wget <raw UML>
- !ls -la

Регрессия

pandas

- Import pandas as pd

- Dataframe

- GigaChat
- DeepSeek
- Perplexity
- Qwen

DataFrame

- df = pd.read….
- df
- df.
- df.dtypes int, float, object
- Уудалить два столбца drop
- Переименовать столбцы
- height weight hair
- df.hair df[“hair”]
- df.head() df.tail()

Описательная статистика

- df.describe()
- df.gender.value_counts()
- import matplotlib.pyplot as plt
- plt.hist
- plt.scatter
- plt.title
- plt.xlabel
- plt.show()

МНК

- Метод наименьших квадратов (МНК)
- Ordinary Least Squares (OLS)
- sklearn
- linear_models
- LinearRegression

SKLearn

- Sci Kit Learn
- Import
- Ordinary Least Squares
- Вес = f(рост)
- model = LinearRegression()
- model.fit(x, y)
- model.predict(x)
- model.coefs_ … .intercept_

Выбросы

- Аномалии

- Удалить строки
- По номеру строки
- По условию > 250

Регрессия на нейросети

- from sklearn.neural_network import MLPRegressor
- neuro = MLPRegressor(hidden_layer_sizes=(1,)
- neuro.fit(X, y)
- Y_predict = regr.predict(X)
- plt.scatter(x, y)
- plt.plot(x, y_predict)

Активация

- Функция активации / возбуждения нейрона

- activation='relu', 'tanh', 'identity'

Датасет

- 100 обычных людей
- X = 150 … 200
- Y = x – 100 ± 10

- 20 марафонцев
- X = 200 ± 10
- Y = 50 ± 10

- Линия регрессии «притягивается» к выбросам!

Перцептрон

- MLPRegression
- Multi Layer Perceptron

- Activation = “1”
- Solver – подобрать
- Hidden layers = (1,)
- Random_state

Colab

- Поделиться блокнотом
- Чтение
- Полный доступ

Decision Tree

- Дерево решений
  - Решающее дерево
- Яндекс картинки
  - Дерево решений WD40
- Import
- Model = decision…
- Fit
- Predict
- visualization

СРС - Зачет

- Boston Housing Dataset
  - Регрессия
- Titanic Survival Dataset
  - Классификация

- Форма на GitHub
  - Ссылка на чтение

Нелинейная модель

- MLPRegressor

- Hidden = (10,1,)
- Activation = tan

Ансамбль моделей

- Случайный лес
  - Несколько деревьев
- Random Forest Regressor
  - Количество деревьев
  - Глубина
  - «Консенсус-прогноз»

Качество модели

- Train-test-split
- Метрики качества

Выбросы

- Аномальные значения
  - Outliers
  - «Притягивают» к себе линию регрессии
- Основные, однородные объекты
  - x: 150 … 200 (n = 100)
  - y = x – 100 ± 10
- Выбросы
  - {190; 60} ± 10 (n = 20)

Робастная регрессия

- Robust Regression
- Устойчивая
- «грубая»
- Нечувствительная к выбросам