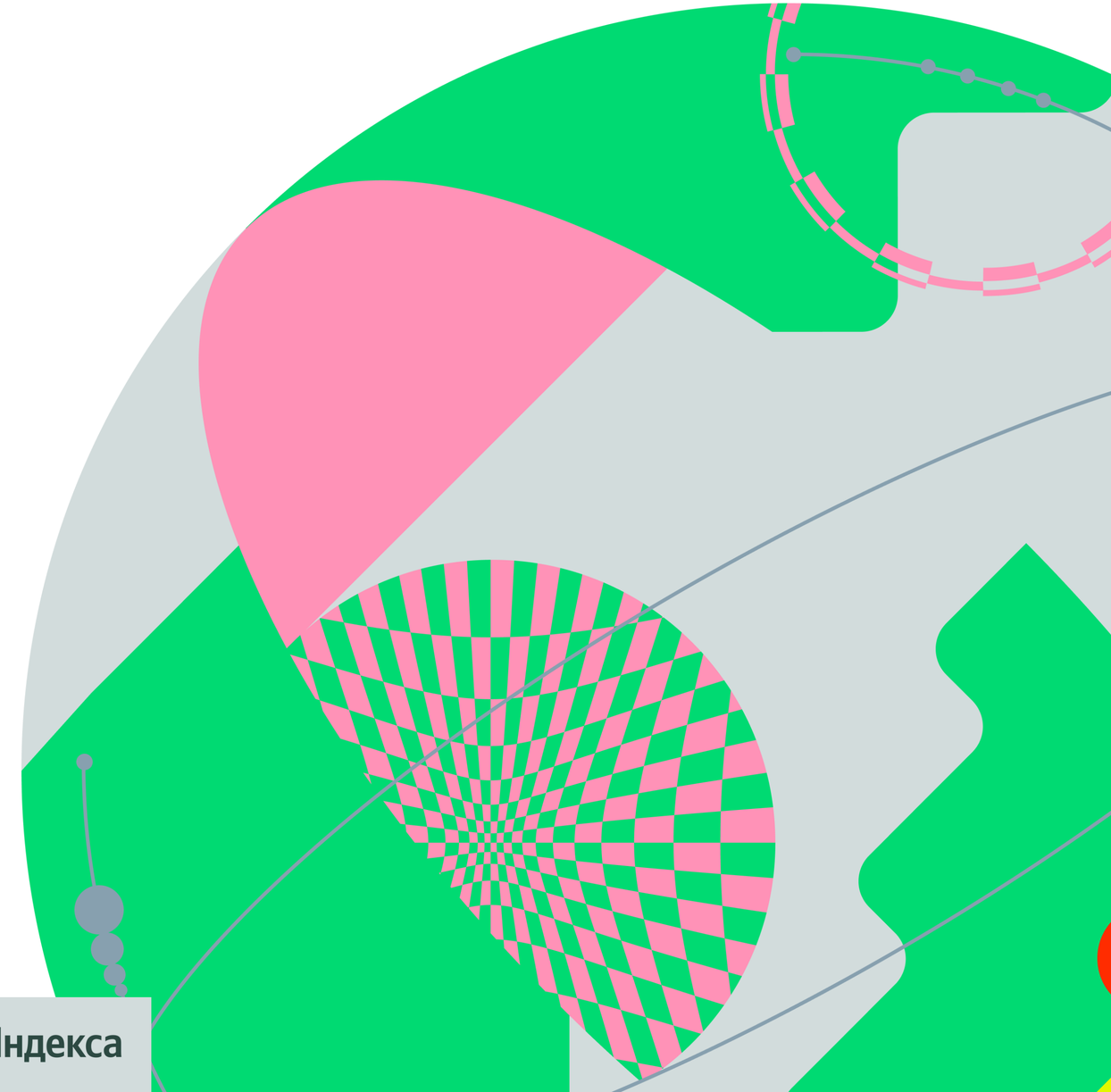
Регрессионное моделирование

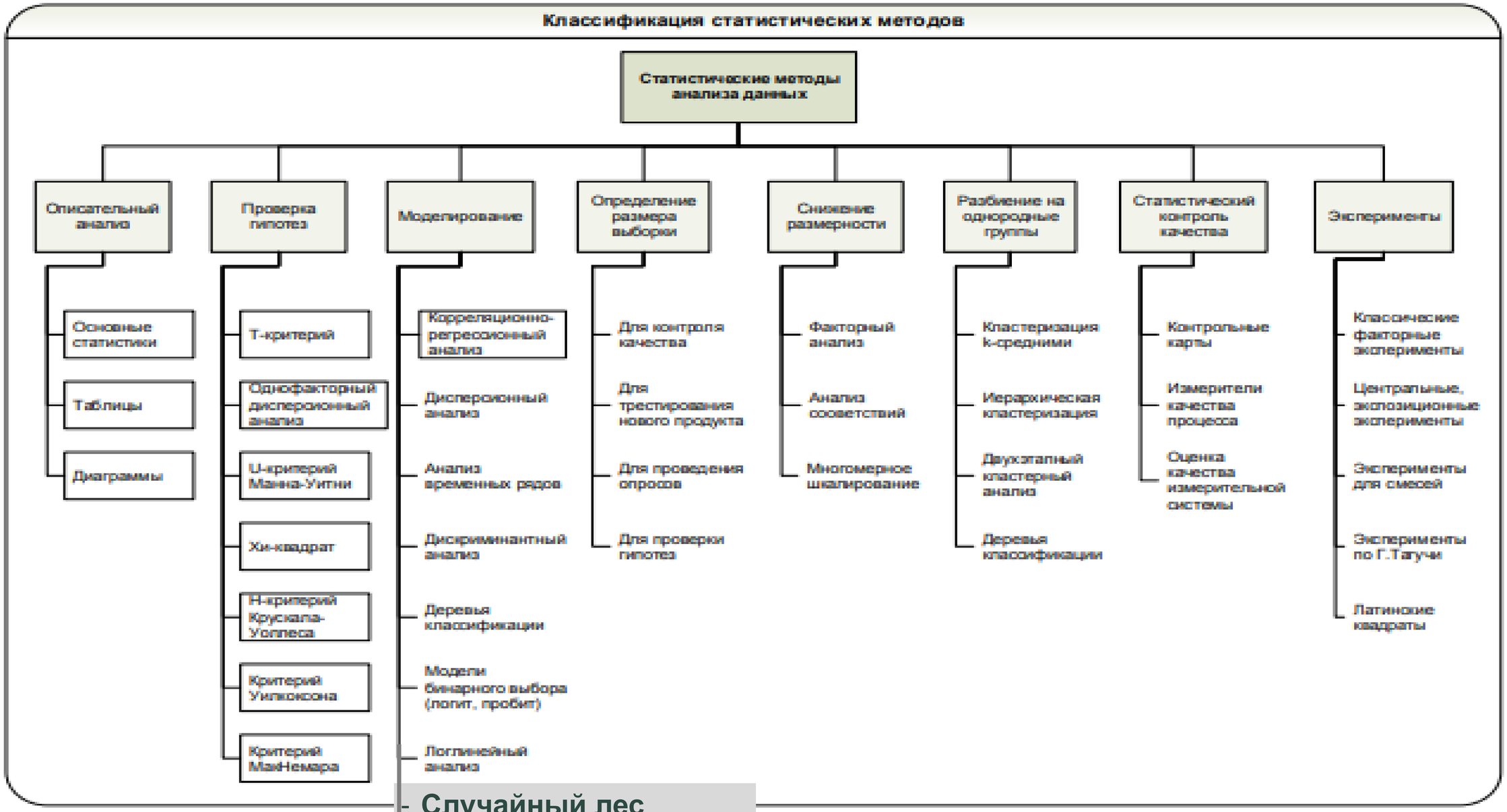


Академия Яндекса

Место регрессионного анализа в большой семье «Статистических Методов»

Друзья, мы рассмотрели с Вами уже несколько статистических тем – описательный анализ и проверки гипотез.

Статистическим методам подвластна любая проблематика!

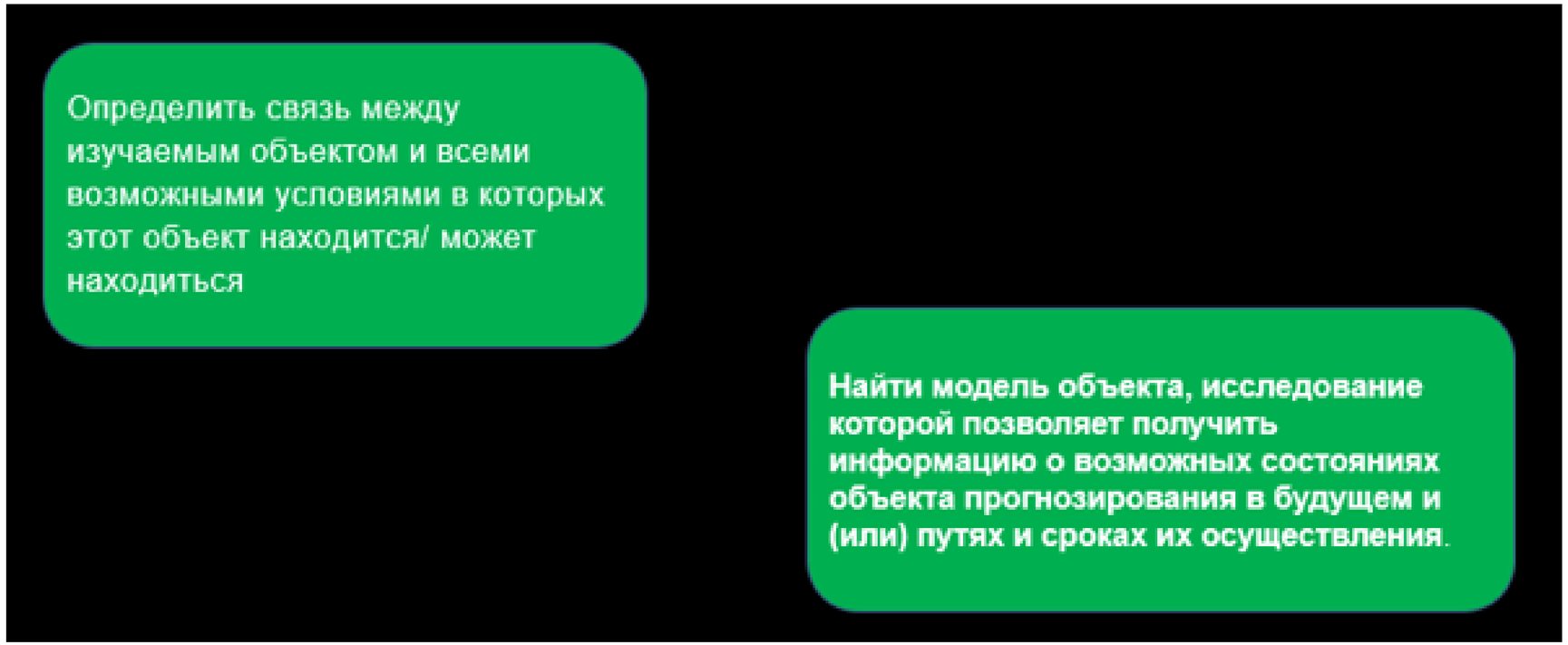


- **Случайный лес**
- **Бустинговые модели**
- **Нейронные сети**

Что значит построить модель

А речь сегодня пойдёт с классики жанра – регрессионное моделирование.

Ещё этот класс моделей называют «линейная регрессия». Его можно назвать «дедушкой». НО этот «старичок» часто выдаёт такие выкрутасы в танцах с данными, что современные ML-методы машинного обучения «нервно курят в сторонке». В частности, если наш заказчик кто-то из промышленности – большая вероятность, что мы будем применять линейную регрессию сначала. НО, обо всём по порядку.



Что значит построить модель

Что значит построить модель.

Объект моделирования

- Целевая переменная
- Зависимая переменная
- Отклик
- y

Условия состояния объекта и его внешнего окружения

- Факторы
- Фичи
- Независимые переменные
- X

Примеры задач

- Определить, какие факторы влияют на расход электроэнергии на предприятии и построить прогноз расходов электроэнергии на ближайший квартал.
- Планируется строительство нового торгового центра. Требуется спрогнозировать «проходимость» секций будущего торгового центра с целью обоснования ставки арендной платы и оптимальной площади помещений.
- На основе риэлтерской базы данных по реализованным объектам недвижимости построить прогноз стоимости квартиры с учетом площади, удобств, типа дома и других факторов.
- Выявить факторы, определяющие долю рынка торговой марки определенных товаров.
- При покупке автомобиля требуется выбрать такой автомобиль, который по истечении трех лет службы на вторичном рынке незначительно потеряет в цене.
- Построить прогноз продаж торговой сети на 2 месяца вперёд.
- Создать модель, которая будет распознавать заболевание по фото рентгена/МРТ/КТ

Типы моделирования

Что такое регрессия

Вот, например, мы с Вами решили начать свой бизнес по продаже мороженного. Начали достаточно успешно. Торгуем себе, собираем информацию о продажах. И по прошествии времени есть вот такой блок данных (Рис.1). Где продажи – это наше всё:)

	дни_продаж	продажи_шт	температура
0	1_день	20	17
1	2_день	25	20
2	3_день	30	25
3	4_день	10	13
4	5_день	28	25
5	6_день	18	17
6	7_день	32	29
7	8_день	10	13
8	9_день	23	23
9	10_день	15	16

```
import pandas as pd

# зададим табличку с продажами
df = pd.DataFrame(
          {'дни_продаж': ['1_день', '2_день', '3_день', '4_день', '5_день', '6_день', '7_день', '8_день', '9_день', '10_день'],
           'продажи_шт': [20, 25, 30, 10, 28, 18, 32, 10, 23, 15],
           'температура': [17, 20, 25, 13, 25, 17, 29, 13, 23, 16 ],
          }
                )
df
```

Что такое регрессия

Что с нашим бизнесом дальше?

Случилось вот что. Раньше нам просто давали товар на реализацию. А то, что не продавалось – дистрибьютор принимал назад без комментариев и убытки брал на себя. (срок годности этого вида мороженого – 2 дня. Т.е. если не продалось – выбрасывали, на следующий день – не продать уже). Это был пробный период продаж для нас, как начинающих предпринимателей. Но всё хорошее заканчивается рано или поздно(. Закончился и этот период. И теперь нам озвучили новые правила: мы сами должны покупать товар на реализацию. Т.е. если товар не продастся – это уже наши риски. Т.е. если закупим товар в бОльшом количестве , чем надо – попадём на убытки. Если в мЕньшем, чем надо – потеряем в продажах.

А это значит, что надо закупать товар максимально точно. Как это сделать? Нужно изучить историю продаж за прошлый период и построить прогноз на будущее. Что ж, начнём изучать семью Модельяни) Один из старейшин этой семьи – Мистер «Линейная регрессия». Его сейчас и рассмотрим.

Давайте построим график скатерплот по данным таблицы.

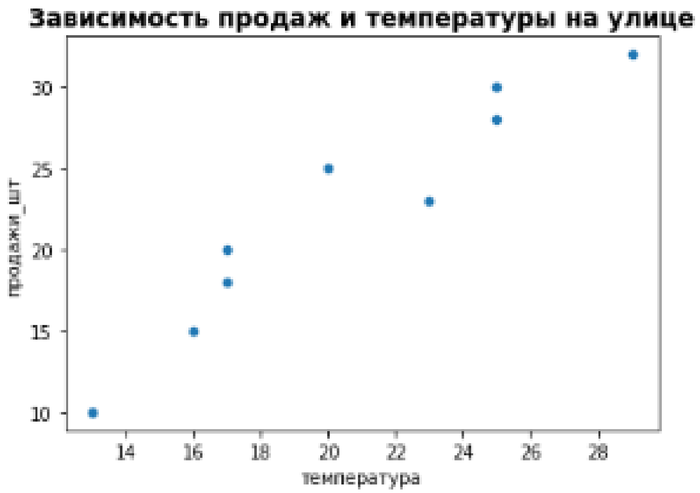
Что такое регрессия

```
: import seaborn as sns
  import matplotlib.pyplot as plt

  y=df['продажи_шт']
  x=df['температура']

  sns.scatterplot(data=df, y=y, x=x)
  plt.title('Зависимость продаж и температуры на улице', fontsize=14, weight='bold')
```

```
: Text(0.5, 1.0, 'Зависимость продаж и температуры на улице')
```

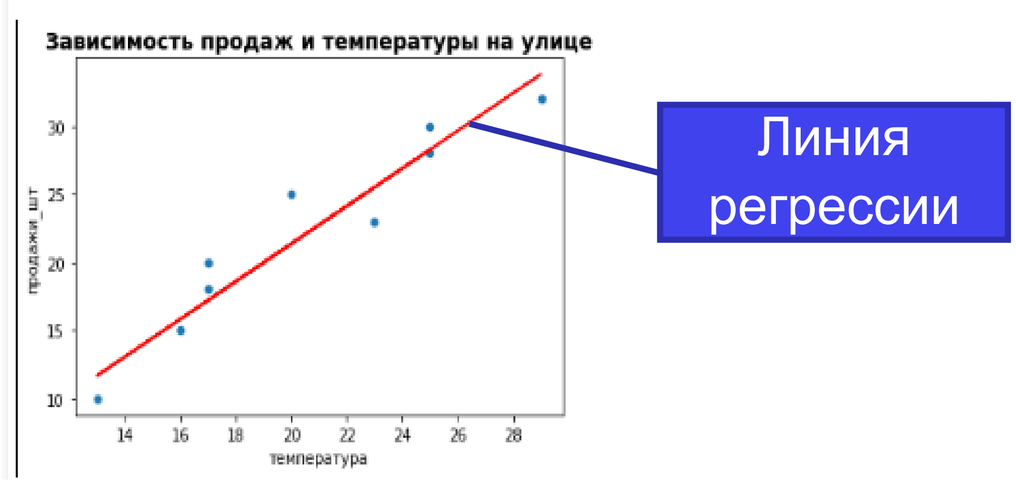


Всё понятно, чем выше температура, тем больше продажи. График показывает зависимость между ними. А можем ли мы так поставить вопрос: если завтра будет 25 градусов, то какое количество мороженого нам надо закупить, чтобы не было ни убытка от нехватки товара, ни убытка от того, что закупим много товара?

Если смотреть на график, то увидим, что при 25 градусах были продажи и 28 и 30 штук.

АААААА)
так сколько же завтра закупать?

Что такое регрессия



Так, разбираемся дальше. Ясно видно, что есть линейная зависимость между продажами и температурой. Эту зависимость в математике принято описывать называемой линией регрессии.

продажи_шт = -6.25 + 1.38*температура,

где (-6.25 и 1.38) – коэффициенты модели.

Причём вот что интересно в этой формуле. Мы можем сказать, что при увеличении температуры на 1 градус продажи мороженного увеличиваются на 1.38 шт. ☺ (Потом ещё рассмотрим этот вопрос про полтора землекопа)).

И что же мы имеем для нашего бизнеса: зная температуру на завтра мы сможем заказывать нужное количество мороженного и получать максимальную прибыль.

УРААА,
скоро купим «порше»
на шашлыки гонять

Академия Яндекса

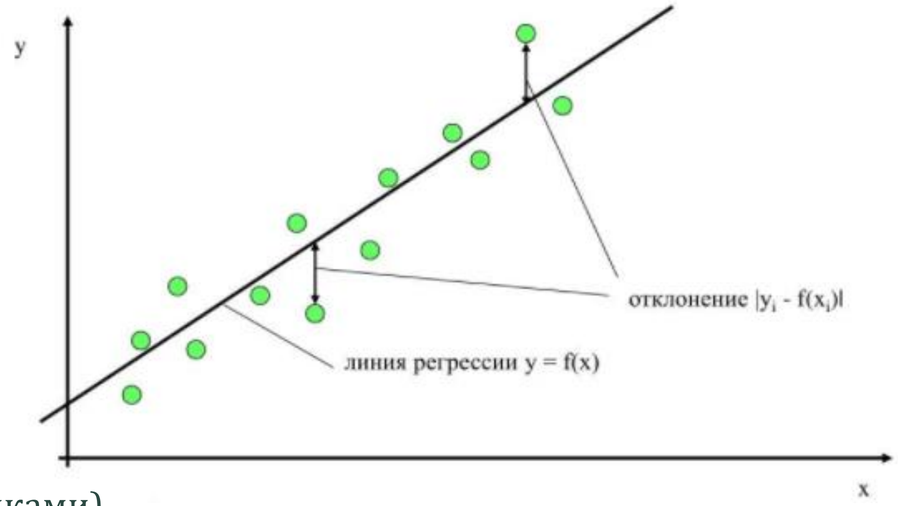
Что такое регрессия

Почему можно доверять этой линии, т.е. этому уравнению?

Всё, что выше описали и есть, можно сказать, регрессионный анализ. НО почему можно доверять этому уравнению? Почему именно такая линия, а не другая? Почему именно такие коэффициенты модели.

Ответы на эти вопросы и есть суть регрессионного анализа.

Скажем прямо – лучшая линия регрессии одна. Т.е. самая лучшая линия, которая описывала бы зависимость продаж и температуры – всего одна. Все остальные хуже. Достигается это в математике с использованием метода наименьших квадратов (МНК)



Основная идея его в следующем:

- Строим зависимость между откликом и фактором. (картинка с зелёными точками)
- Проводим любую линию (в табличке ниже «линия»)
- Ищем отклонения каждой точки графика от этой линии. (в табличке ниже «y – линия»))
- Возводим в квадрат это отклонение. И ищем сумму всех квадратов отклонений. (в табличке ниже «(y – линия)^2»
- Дальше, проводим другую линию на графике и повторяем п.п.3-4
- Дальше проводим третью линию и повторяем п.п.3.4. т.д….
- Та линия, у которой сумма квадратов отклонений минимальная будет и принято называть единственной линией регрессии. Т.е.

Найдена модель регрессии, которая позволит нам сроить прогнозы.

Что такое регрессия

Всё, что перечислено выше в пунктах, отражено в табличке, применительно к нашим данным.

	дни_продаж	продажи_шт	температура	линия	y - линия	(y - линия)^2
0	1_день	20	17	17.231959	2.768041	7.662052
1	2_день	25	20	21.376289	3.623711	13.131284
2	3_день	30	25	28.283505	1.716495	2.946355
3	4_день	10	13	11.706186	-1.706186	2.911069
4	5_день	28	25	28.283505	-0.283505	0.080375
5	6_день	18	17	17.231959	0.768041	0.589887
6	7_день	32	29	33.809278	-1.809278	3.273488
7	8_день	10	13	11.706186	-1.706186	2.911069
8	9_день	23	23	25.520619	-2.520619	6.353518
9	10_день	15	16	15.850515	-0.850515	0.723377

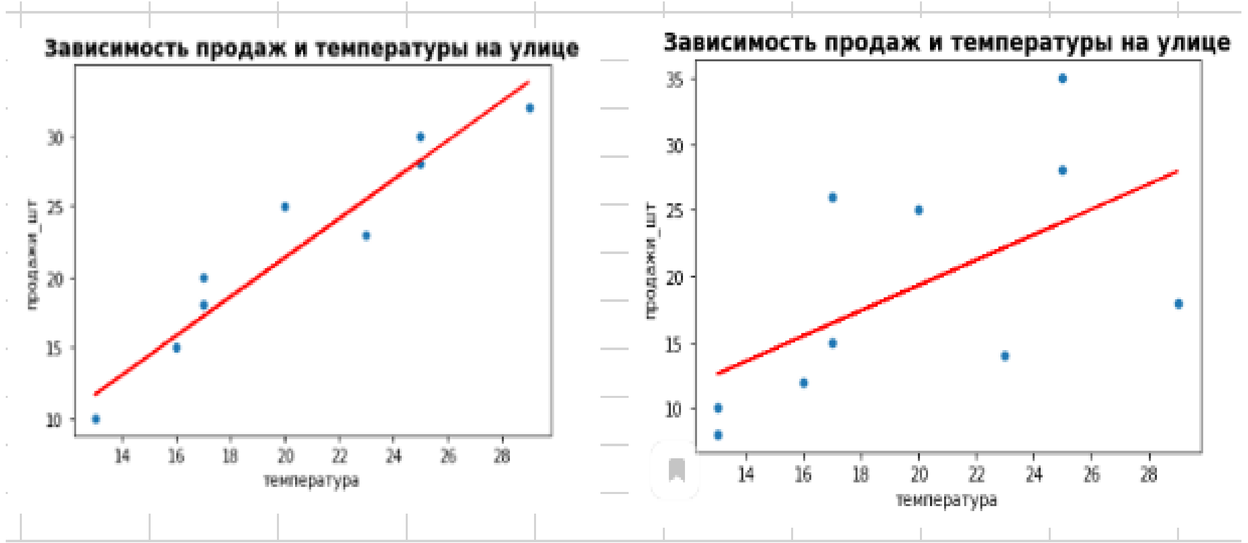
И вот та линия, у которой сумма по столбцу «(y – линия)^2» будет минимальной и есть наша линия регрессии. Единственная и не повторимая). Эта линия (известно из курса математики) описывается уравнением y = a0 + a*x, где a0 и a1 – коэффициенты линии.

И вот, благодаря рассмотренному методу наименьших квадратов, мы можем говорить о формуле, по которой можем прогнозировать нашу цель (продажи_шт), зная фактор продаж (температуру).

Что такое регрессия

Точность линии регрессии

Ну хорошо. Линию регрессии нашли. И хотим строить прогноз продаж по ней. Но как узнать хороша ли найденная модель? Рассмотрим разные линии регрессии и два графика:



Что в них разного?

Верно, тот, что справа – имеет линию регрессию, которая хуже описывает данные. Т.е. точки вокруг линии расположены далеко. А на левом графике – близко к линии. И вот интуитивно можем сказать, что левый график лучше предсказывает продажи.

Графический анализ – хорош. Но хочется иметь ещё аналитический инструмент, чтобы утверждать о лучше/хуже. И такой инструмент есть. Он называется коэффициент детерминации.

Что такое регрессия

Мера проверки адекватности модели, построенной с помощью регрессии

Ранее, уже изучали коэффициент корреляции. И это поможет нам сейчас. Если кратко, то математически коэффициент детерминации равен коэффициенту корреляции в квадрате. Формула тут такая: R2 = R*R . Но есть ещё более понятная для понимания сути формула:

$$\sum (Y_i - \bar{Y})^2 = \sum (\hat{Y}_i - \bar{Y})^2 + \sum (Y_i - \hat{Y}_i)^2$$

Сумма квадратов относительно среднего **Y** **(общая)**

Сумма квадратов обусловленная регрессией **(объясненная)**

Сумма квадратов относительно регрессии **(необъясненная, остаток)**

$$SS_Y = SS_r + SS_o$$

КОЭФФИЦИЕНТ ДЕТЕРМИНАЦИИ

$$R^2 = SS_r / SS_Y \qquad 0 \le R^2 \le 1$$

Сутейно – это так можно понять. У нашей целевой – «продажи_шт» есть разброс - ну, у него продажи то 10 до 30. И мы ведь понимаем, что ответ на вопрос « почему то вниз, то вверх» бегут продажи зависит от различных факторов продаж.

И вот коэффициент детерминации показывает какая часть изменчивости (разброса) нашей целевой происходит из-за фактора, включённого в модель. Например, пусть на левой картинке выше R2=0.9, а на правой картинке R2=0.6. И говорят следующее: продажи на 90% и на 60% обусловлены температурой. А остальные 10% и 40% продаж - зависят от других факторов.

Каких? Тут надо уже другое исследование, но их тоже можно обнаружить и также вставить в модель.

Всё это время, мы рассматривали задачку, когда есть зависимая величина и одна независимая. Это простая регрессия. Но если используют несколько факторов для регрессии (когда R2 = 0.6 – конечно же надо дополнительно вводить ещё факторы) И тогда говорят о множественной регрессии.

Академия Яндекса

Что такое регрессия

Чаще всего все ваши модели потребуют использования нескольких факторов, а то и десятков факторов. И надо сказать, что идея нахождения самой лучшей линии регрессии для одномерного случая, справедливо и в множественной регрессии. Только вместо линии будет использоваться понятие поверхности (это мы рассматривать не будем). А остальное, включая историю с R2 -интерпретируется аналогично.

В общем виде регрессионное моделирование представляет собой уравнение вида:

$$Прогноз\ продаж = Константа + B_1 \cdot Фактор_1 + B_2 \cdot Фактор_2 + ... + B_k \cdot Фактор_k$$

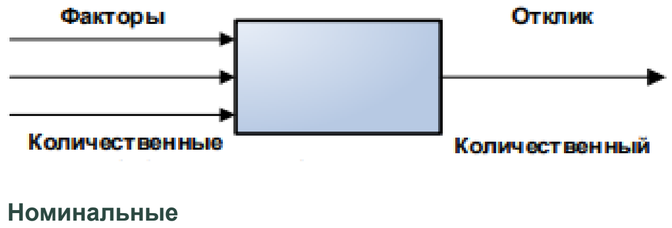
Или более кратко можем записать:

$$\hat{Y} = B_0 + B_1 X_1 + B_2 X_2 + ... + B_k X_k$$

Ограничения использования множественной линейной регрессии.

Всегда есть желание иметь «волшебную кнопку»: нажал её и построился самый точный прогноз будущего на все случаи жизни. Но жизнь сложнее и математика пока не в силах все её особенности учесть. Поэтому и использование регрессии возможно при определённых допущениях:

1) Требования к исходным данным:
Отклик – только количественная шкала
Факторы – количественная шкала или номинальная.



2) Для обеспечения приемлемой точности модели минимальный объем выборки не должен быть меньше величины «число факторов, умножить на 10». В нашем случае с мороженным: у нас 1 фактор, значит строк в базе данных должно быть не менее 10-20.

3) Факторы в модель для прогнозирования подставляются только в том диапазоне значений, в котором модель обучалась. Например, Фактор «температура воздуха» имеет диапазон обучения от -1 до +25. Это значит, что построить прогноз выручки со значением температуры +40 – мы не сможем. Т.е. модель-то всё посчитает (что с неё взять – просто формула). Но вот верить полученному прогнозу продаж надо осторожно. Т.к. наблюдений таких на обучении не было, увы.

4) Ну и такой момент: у нас название моделирования «линейная регрессия». По большому счёту, это значит, что зависимость между фактором и откликом должна быть линейная. Но на наших встречах мы увидим, как это можно эту проблемку решить и использовать себе на благо, если связь будет нелинейной.

Алгоритм

Шаг 1
Создание теоретической модели

Шаг 2
Построение статистической модели

Шаг 3
Проверка качества модели

Шаг 4
Получение прогнозов

Шаг 5
Анализ ошибки прогноза

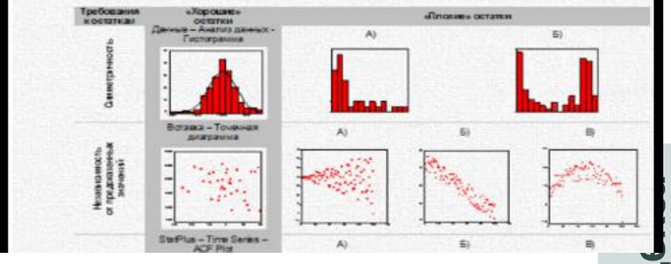
Факт	Прогноз	Остатки
112	96	16
111	110	1
98	105	-7
120	108	12
110	111	-1
119	111	8
89	98	-9
95	120	-25
96	110	-14
110	119	-9
105	89	16
108	105	3
111	108	3
120	96	24
110	110	0
119	105	14
89	108	-19

Остатки (ошибки, отклонения) = ФАКТ – МОДЕЛЬ

Основная идея анализа остатков:
остатки не должны содержать никаких закономерностей, то есть:

- Иметь симметричное (нормальное) распределение
- Иметь постоянный разброс
- Не иметь автокорреляции

Анализ остатков



Анализ ошибки прогноза. Регрессии

Факт	Прогноз	Остатки
112	96	16
111	110	1
98	105	-7
120	108	12
110	111	-1
119	111	8
89	98	-9
95	120	-25
96	110	-14
110	119	-9
105	89	16
108	105	3
111	108	3
120	96	24
110	110	0
119	105	14
89	108	-19
95	111	-16

- Зависящие от масштаба
  - MAE
  - MSE
  - RMSE
- Относительные
  - MAPE
  - SMAPE
- Не зависящие от масштаба
  - MASE

Что понадобится на практике

Построение линейной регрессии

1. Разбить на обучающую выборку и тестовую:

```
from sklearn.model_selection import train_test_split
```

```
# Разобьём на тренировочную выборку и тестовую
x_train, x_test, y_train, y_test = train_test_split(feature, target, test_size=0.2, random_state=RANDOM_STATE)
```

Что понадобится на практике

Построение линейной регрессии

2. Если надо – провести категоризацию

```
from sklearn.preprocessing import OneHotEncoder
```

`pd.get_dummies()` ! - **НЕ ИСПОЛЬЗУЕМ**

порода	тип_пастбища	порода_папы_быка
Вис Бик Айдиал	низменное	Соверин
РефлешнСоверинг	холмистое	Соверин
Вис Бик Айдиал	низменное	Айдиал

порода_РефлешнСоверинг	тип_пастбища_холмистое	порода_папы_быка_Соверин
0.0	0.0	1.0
1.0	1.0	1.0

```
ohe = OneHotEncoder(sparse=False, drop='first', handle_unknown="ignore")
ohe.fit(x_train[category_var])
```

Подробнее о методе: https://datagy.io/sklearn-one-hot-encode/

Что понадобится на практике

Построение линейной регрессии

3. Ищем модель на обучающей выборке

```
from sklearn.linear_model import LinearRegression
```

```
# Инициализация модели
lin_regr = LinearRegression()

# Обучение модели
lin_regr.fit(x_train_cat, y_train)
```

```
# Предсказание
prediction = lin_regr.predict(x_train_cat)
```

4. Считаем метрику на обучающей выборке

```
from sklearn.metrics import r2_score, mean_absolute_error, mean_squared_error, mean_absolute_percentage_error
```

```
def metrics_model(fact, prediction):
    '''
    Функция выводит метрики качества для регрессионных моделей (точность модели).
    '''

    print('R2   = ', round(r2_score(fact, prediction),2))
    print('MAPE = ', round(mean_absolute_percentage_error(fact, prediction)*100,2), 'процентов')
    print('MAE  = ', round(mean_absolute_error(fact, prediction),2))
   # print('MSE  = ', round(mean_squared_error(fact, prediction),2))
    print('RMSE = ', round(mean_squared_error(fact, prediction)**0.5,2))
```

Что понадобится на практике

Построение линейной регрессии

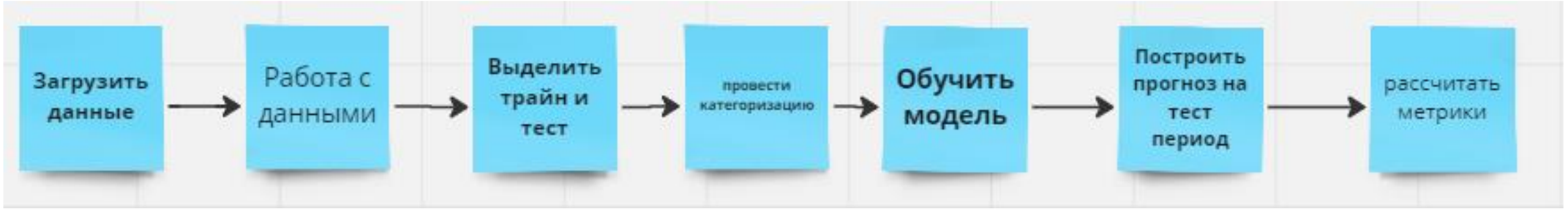
5. Ищем метрику на тестовой выборке

```
# Предсказание целевой
prediction_test = lin_regr.predict(x_test_cat)
```

```
print('R2   = ', round(r2_score(fact, prediction),2))
print('MAPE = ', round(mean_absolute_percentage_error(fact, prediction)*100,2), 'процентов')
```

6. Принимаем решение о качестве прогноза

Алгоритм моделирования



Дополнительная литература

По линейной регрессии:

1. https://bibl.nngasu.ru/electronicresources/uch-metod/economic_statistics/859984.pdf
2. https://habr.com/ru/articles/514818/

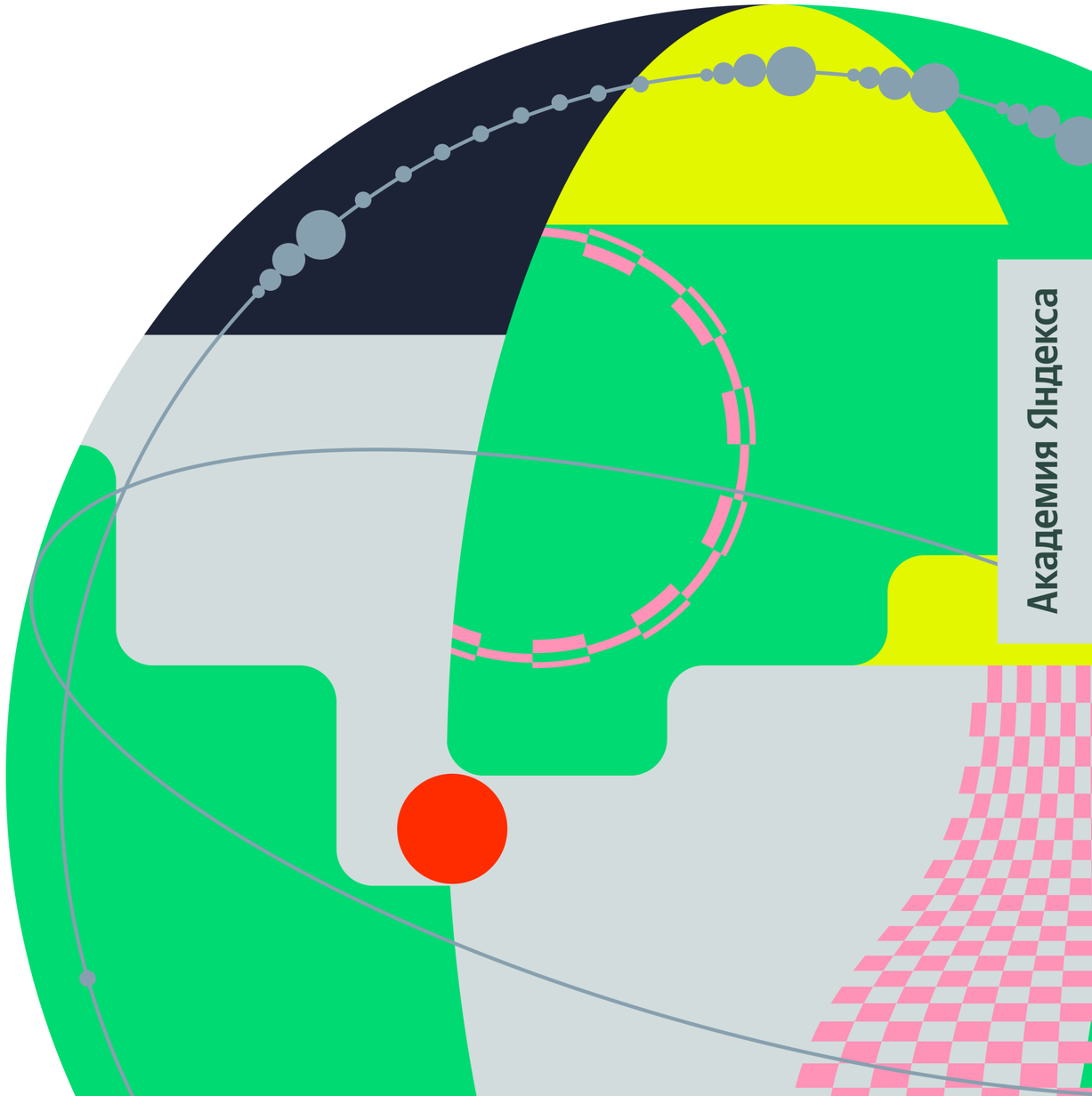
Реализация в питоне:

1. https://www.kaggle.com/code/muzafferdindar/linear-regression-in-python
2. https://www.kaggle.com/code/emineyetm/simple-linear-regression-using-python

О метриках: дополнительные слайды

По логистической регрессии: https://www.kaggle.com/code/prashant111/logistic-regression-classifier-tutorial

Спасибо за внимание



Академия Яндекса позволяет школьникам и студентам освоить востребованные IT-профессии по программам, разработанным экспертами компании