

АНАЛИЗ ФАКТОРОВ, ВЛИЯЮЩИХ НА УПОТРЕБЛЕНИЕ ТАБАКА В США

Построение регрессионной модели

Цель и описание данных

Цель: выявить основные факторы, влияющие на уровень употребления табака и построить регрессионную модель, предсказывающую уровень употребления табака в США в зависимости от выявленных факторов.

Year	State	Smoke everyday	Smoke some days	Former smoker	Never smoked
2009	Vermont	12.5%	4.6%	31.6%	51.3%
2009	South Carolina	14.6%	5.7%	24.5%	55.2%
2006	Puerto Rico	7.7%	4.8%	16.8%	70.7%
2006	Nationwide (States and DC)	14.9%	5.1%	24.7%	54.3%
2008	Georgia	14.5%	5 %	21.4%	59.1%

Подготовка данных к анализу

- Пропуски отсутствуют
 - Исключены строки с общими показателями по группам субъектов
 - Проценты переведены в доли населения
 - Проверка корректности данных
 - Созданы dummy-переменные для штатов

Предварительный анализ

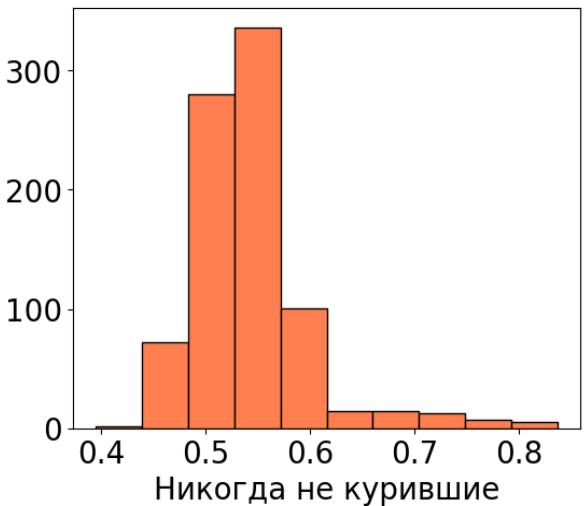
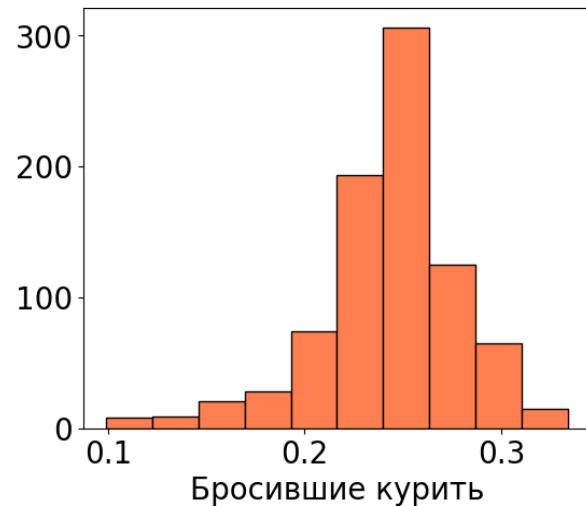
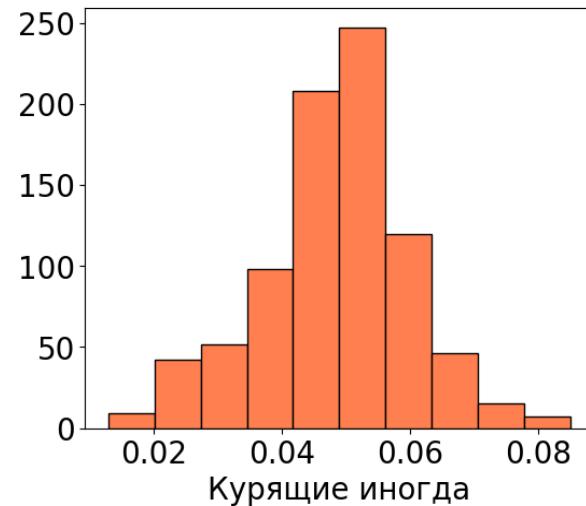
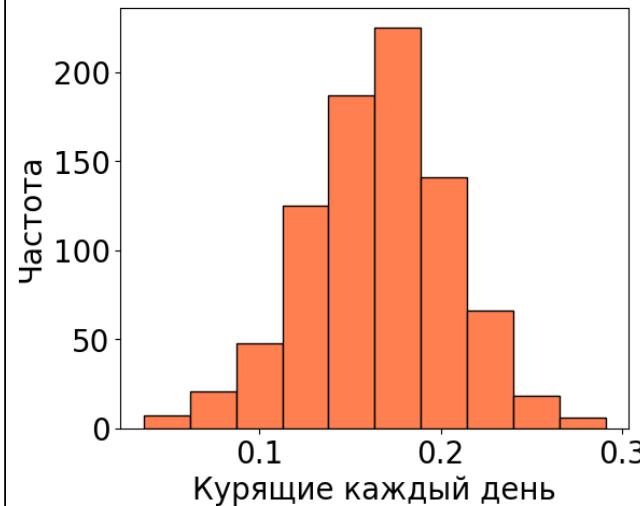
Целевой признак — “everyday” -
доля людей, курящих ежедневно

Упорядоченность данных по
критерию Вальда–Вольфовица :

$$p = 0.028 \rightarrow p = 0.899$$

	everyday	sometimes	former	never
mean	0.082801	0.024751	0.126400	0.298734
std	0.039424	0.011547	0.066231	0.163106
min	0.036000	0.011708	0.035652	0.056970
25%	0.056039	0.016386	0.087755	0.208191
50%	0.080637	0.024028	0.129143	0.306565
75%	0.100171	0.029876	0.157578	0.411886
max	0.145008	0.043076	0.223447	0.499911

Нормальность распределений и упорядоченность данных



$p = 0.136$

нормальное

$p = 0.028$

ненормальное

$p = 7.01e-45$

ненормальное

$p = 0$

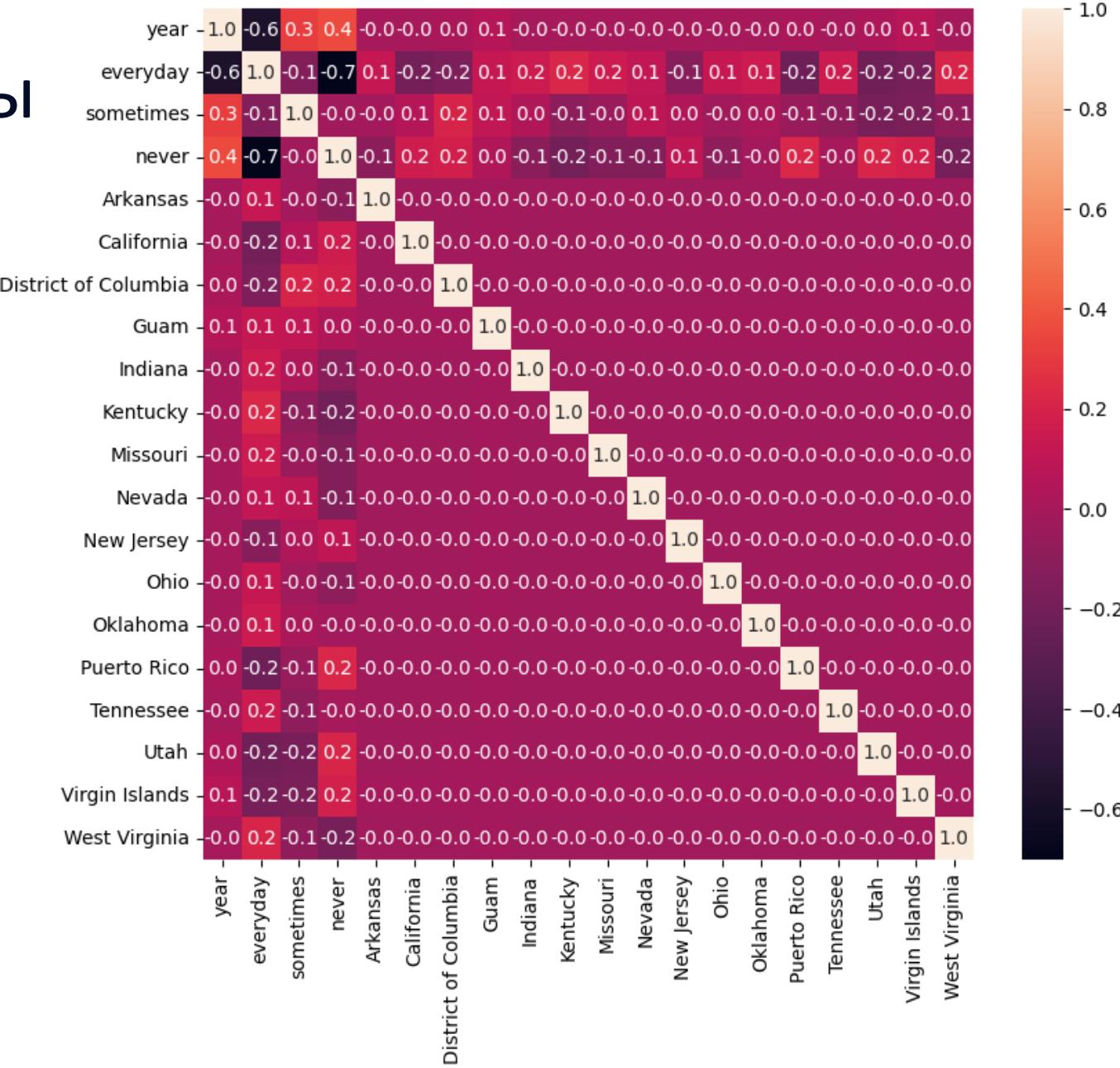
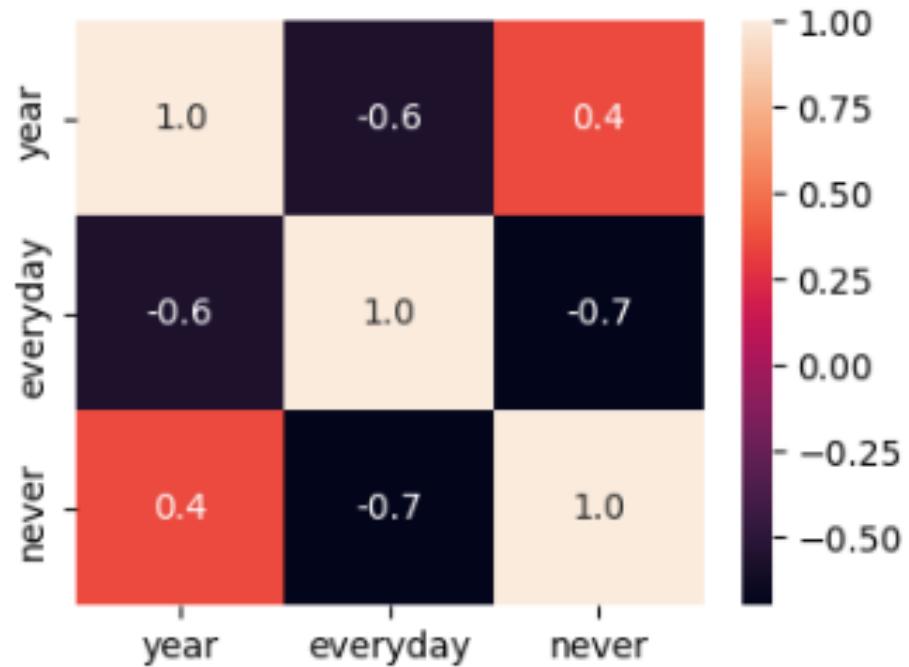
ненормальное

Р-значения вычислены по критерию Жарка-Бера, уровень значимости 5%.

Поскольку распределения не являются нормальными, для корреляционного анализа будет применяться метод ранговой корреляции Спирмена.

Корреляционные матрицы

В матрицы включены факторы, демонстрирующие как минимум слабую тесноту связи с целевым признаком (по шкале Чеддока)



Модель, учитываящая штаты

Факторы: год, доля никогда не куривших, доля курящих иногда, субъекты: Пуэрто-Рико, Кентукки, Западная Вирджиния, Юта, Калифорния, Виргинские острова, Округ Колумбия, Миссури, Теннесси, Индиана, Оклахома, Арканзас, Гуам, Огайо, Нью-Джерси, Невада

При уровне значимости 5%:

$$R^2_{\text{train}} = 0.875, \quad R^2_{\text{test}} = 0.845$$

$$R^2_{\text{adj,train}} = 0.871, \quad R^2_{\text{adj,test}} = 0.825$$

$$\text{MAE} = 0.012, \quad \text{RMSE} = 0.016$$

Модель не переобучена

Модель с меньшим числом факторов

Факторы: год и доля никогда не куривших людей

При уровне значимости 5%:

$$R^2_{\text{train}} = 0.681, \quad R^2_{\text{test}} = 0.603$$

$$R^2_{\text{adj,train}} = 0.680, \quad R^2_{\text{adj,test}} = 0.598$$

$$\text{MAE} = 0.02, \quad \text{RMSE} = 0.023$$

Модель не переобучена

Интерпретация

- С каждым годом в США доля курящих ежедневно уменьшается на 3.7%
- При увеличении количества никогда не куривших людей на 1% доля ежедневно курящих уменьшается на 0.27%
- Кроме того доля ежедневно курящих меняется в зависимости от штата.

СПАСИБО ЗА ВНИМАНИЕ!

[Ссылка на проект в Google Colab](#)