Цель исследования: оценивание эффекта воздействия занятий спортом на заработную плату индивида.

1. Обоснование темы

1. Придумайте непрерывную зависимую (целевую) переменную (например, заработная плата или прибыль) и бинарную переменную воздействия (например, образование или факт занятий спортом).

Непрерывная зависимая переменная:	**Бинарная переменная воздействия:**
Заработная плата	Факт занятий спортом

2. Опишите, для чего может быть полезно изучение влияния переменной воздействия на зависимую переменную. В частности, укажите, как эта информация может быть использована бизнесом или государственными органами.

Изучать влияние факта занятий спортом на заработную плату индивида полезно как государству, так и бизнесу, так как данная информацию может помочь им функционировать более продуктивно и успешно.

Например, если Яндекс оценит эффект от занятий спортом на заработную плату своих сотрудников как положительный и значимый, возможно тогда, тренажёрных залов разного вида в их бизнес-центре станет еще больше. Спорт будет влиять не только на зарплату самого сотрудника, но и косвенно на составляющие успешного функционирования бизнеса в том числе (оптимизация управления персоналом, снижение расходов на здравоохранение, конкурентные преимущества).

Чем больше занимается спортом индивид, тем более высокую заработную плату он получает, потому что:

- **рост продуктивности** => готов трудиться больше, больше энергии => вносит больший вклад в деятельность компании=> больший выпуск => большая выручка => большая зарплата
- **снижение затрат на медицинское обслуживание работников,** так как спорт помогает сотрудникам поддерживать здоровье на высоком уровне => снижение издержек на здравоохранение =>

больший потенциальный выпуск компании => большая выручка => большая зарплата

- **повышение уровня стрессоустойчивости**=> быстро реагирует на внештатные ситуации, справляется с ними => задержек в функционировании бизнеса нет => больший выпуск по сравнению с конкурентами => большая выручка => большая зарплата

3. Обоснуйте наличие причинно-следственной связи между зависимой переменной и переменной воздействия. Приведите не менее 2-х источников из научной литературы, подкрепляющих ваши предположения.

   Причинно-следственная связь между фактом занятий спортом и более высокой заработной платой существует. Чем больше занимается спортом человек, тем более дисциплированным, энергичным, собранным он становится. Он успевает закрыть больше рабочих дел, работает эффективнее. Благодаря этому, в итоге, получает большую заработную плату.

   Источники из научной литературы, подкрепляющие предположения:

1) Dan-Olof Rooth,Work out or out of work — The labor market return to physical fitness and leisure sports activities,Labour Economics,Volume 18, Issue 3,2011,Pages 399-409,ISSN 0927-5371,https://doi.org/10.1016/j.labeco.2010.11.006.

Не используя другие контрольные переменные в модели линейной регрессии, помимо факта занятий спортом и возраста, в исследовании Рута было найдено, что при изменении факта занятий спортом на 100 п.п. (от 0 к 1), прибавка к прибыли будет около 7%. Если включить в модель ненаблюдаемые «семейные» переменные, влияющие на эту связь, с помощью модели фиксированных эффектов, эта премия снижается до 4%. Следует отметить, что премия за фитнес в этом случае точно соответствует отдачи примерно от 1,3 года трудового стажа.

Рис. 1. На рисунке видно, что заработок растет с ростом физической подготовки.

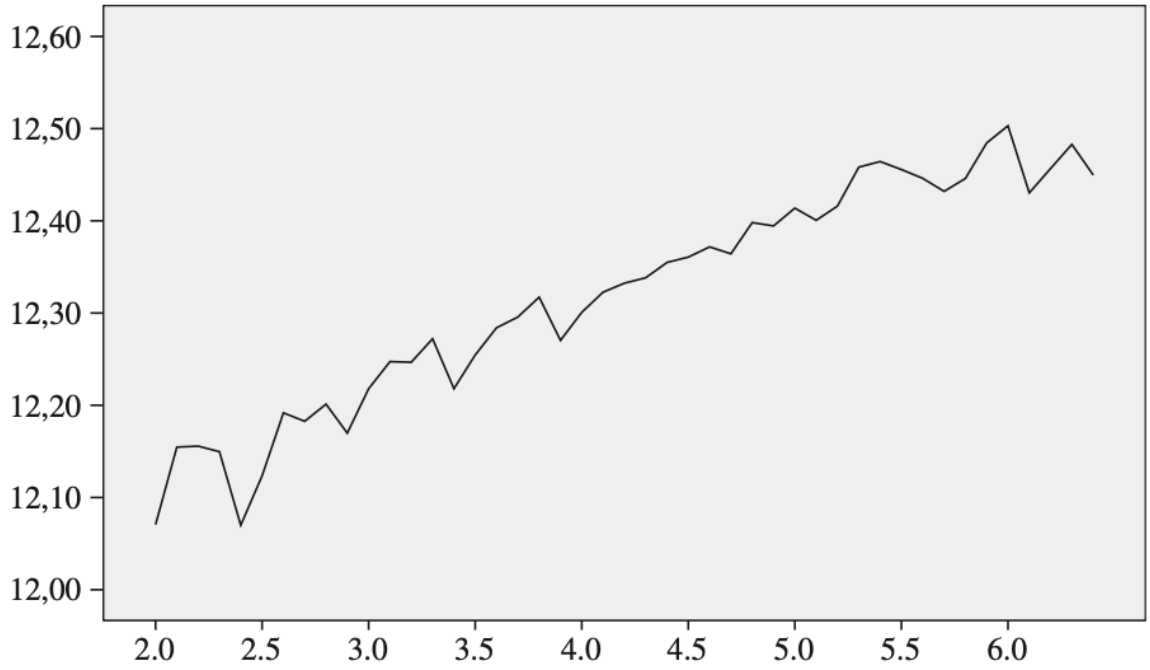


Fig. 1. (Log) Earnings and physical fitness. Total population. Note: The figure shows the mean log earnings for those with a value of 2.0, 2.1, 2.2 etc. up to 6.4 of the physical fitness variable. The physical fitness variable has a mean of 4.3, a standard deviation of 0.7, and is truncated at 2.0 and 6.4.

---

2) Kosteas, V.D. The Effect of Exercise on Earnings: Evidence from the NLSY. *J Labor Res* **33**, 225–250 (2012). https://doi.org/10.1007/s12122-011-9129-2

В статье Костеаса, исследующей приводят ли частые тренировки к более высокой заработной плате, были сделаны следующие важные выводы:

- По данным исследования, опубликованного в Journal of Labor Research, люди, регулярно занимающиеся физической активностью, зарабатывают на 5-10% больше своих менее активных коллег.
- В исследовании также отмечается, что важна не столько регулярность занятий, сколько их интенсивность.
- Те, кто занимается физическими упражнениями низкой интенсивности, зарабатывают в среднем $54,000 в год, тогда как

те, кто тренируется с умеренной или высокой интенсивностью, имеют доход $67,000 и $83,000 в год соответственно.

Рис. 2. На рисунке видно, что мужчины и женщины, которые часто занимаются спортом, получают большую заработную плату. При этом прирост этой зарплаты для мужчин будет меньшим (10.5%) по сравнению с женщинами (12.9%).

Table 3 Exercise frequency and earnings

	Men		Women	
	OLS	FE	OLS	FE
Rarely exercise	0.052[a] (0.027)	0.039 (0.026)	0.018 (0.026)	−0.013 (0.032)
Infrequent exercise	0.05[a] (0.028)	−0.016 (0.024)	0.014 (0.028)	0.0095 (0.037)
Moderate exercise	0.061* (0.025)	0.034 (0.025)	0.055* (0.027)	−0.033 (0.037)
Frequent exercise	0.105** (0.027)	0.034 (0.027)	0.129** (0.028)	0.079[a] (0.044)

4. Кратко опишите результаты предшествовавших исследований по схожей тематике и критически оцените методологию этих работ с точки зрения гибкости (жесткости предпосылок) использовавшихся методов эконометрического анализа.

Результаты предшествующих исследований: с увеличением частоты занятий спортом, заработная плата индивида увеличивается

Методология, что в первом и во втором исследовании - построение линейной регрессии (OLS, FE). Предполагается линейная связь между зависимой (зарплата) и независимыми переменными (возраст, факт занятий спортом). Хотя это необязательно так, для это мы попробуем использовать другие модели в будущем (RF, GB, KNN). Отсутствие мультиколлинеарности - так или иначе наши независимые переменные могут быть связаны между собой, поэтому говорить, что они не коррелируют - неправильно. Экзогенность данных - это невозможно, когда мы не имеем эксперимента, поэтому мы всегда будем иметь эндогенность в

наших данных, оценка будет смещена. Чтобы хотя бы чуть-чуть уменьшить это смещение, вместо обычной OLS, рассчитывается также модель FE.

1) Регрессионная модель: (Dan-Olof Rooth)

   Log Earningsij = a + b * Fitnessij + c * Xij + fj + eij.

Эта модель включает только 2 независимые переменные - факт занятия спортом и возраст (также переменную fixed effects)

Оценка коэффициента при спорте может быть смещена, так как могут существовать экзогенные неучтенные нами в модели переменные, которые коррелируют с Fitness.

2) Все это аналогично и для 2 исследования. (Kosteas) К тому же, можно добавить, что несмотря на то, что были учтены постоянные не меняющиеся с течением времени переменные (fixed effects), устойчивый и точный результат не был найден (non-significant results)

5. Придумайте хотя бы 3 контрольные переменные, по крайней мере одна из которых должна быть бинарной и хотя бы одна – непрерывной. Кратко обоснуйте выбор каждой из них.

   Придумайте бинарную инструментальную переменную и обоснуйте, почему она удовлетворяет необходимым условиям.

   Таблица № 1. Обоснование выбора контрольных переменных

Контрольные переменные	Обоснования выбора
Образование индивида (1 - 25 лет, берем как непрерывную)	Чем выше образование индивида => тем более компетентным он считается на рынке труда=> тем большую зарплату он получает (до определенного количества лет обучения)
Здоровье индивида (1 - здоров, 0 - болен)	#Около 50% прошедших диспансеризацию россиян имеют хронические заболевания (РБК)

(факт того, что человек здоров определяет его занятие спортом)	Если человек болен => фактически он трудиться меньше в связи с ограничениями здоровья=> меньшая заработная плата по сравнению с другими
Возраст (14 до 85 лет, берем как непрерывную)	Чем старше человек, тем большую заработную плату он склонен получать, но до определенного возраста
Наличие стресса (1 - есть, 0 - нет) (факт депрессии определяет занятие спортом)	Если у человека депрессия, он будет лениться, будет менее продуктивным => меньшая потенциальная заработная плата В 2023 году уровень рабочего стресса как очень высокий оценивали 13% опрошенных. Предположили, что в 2024 будет около 10%.

Таблица № 2. Удовлетворение необходимым условиям IV

Бинарная инструментальная переменная	**Удовлетворение необходимым условиям**
Отдаленность спортивного зала/ места для занятий спортом. (1 - близко, 0 - далеко)	• Сильно коррелирует с фактом занятий спортом (переменной воздействия): Если турник находиться рядом с домой, вероятность того, что человек пройдет мимо него очень мала • Косвенное влияние на заработную плату (целевая переменная): Чем ближе спортзал находится к месту проживания индивида, тем больше вероятность того, что он станет заниматься спортом чаще,

	тем выше шанс получения им более высокой заработной платы • Отсутствие корреляции IV с ошибкой в регрессии на факт занятий спортом Если мы можем сказать, что способности и другие характеристики, влияющие на зарплату индивида, могут коррелировать с наличием им занятий спортом, то мы точно может утверждать, что с нашей IV (отдаленность спортивного зала) этой корреляции не будет, она экзогенна.

2. Генерация и предварительная обработка данных

1. Опишите математически предполагаемый вами процесс генерации данных.

Генерация случайных величин из нормального и биномиального распределения с математическим ожиданием (loc) равным μ и стандартным отклонением (scale) равным σ для нормального распределения (X_i ~ N (μ,σ^2), i=1...n) и с порогом p для биномиального (X_i~ Bi(1,p), i=1...n). Эти величины (size=10000) генерируются при помощи функций scipy.stats.norm.rvs и np.random.binomialиз библиотек scipy and numpy.

$f(X_i|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ и

$P(X_i = k) = \binom{1}{k} p^k (1-p)^{1-k}$

2. Кратко обоснуйте предполагаемые направления связей зависимой переменной и переменной воздействия с контрольными переменными.

Таблица №3. Предполагаемые направления связей зависимой переменной с контрольными.

Зависимая переменная	Контрольная переменная	Предполагаемое направление связи
Заработная плата	Уровень образования	Положительная связь с затухающим эффектом (отдача в определенный период начинает уменьшаться) Чем выше у человека образование, тем больше отдача от него (но это как правильно до PHD, отдача от магистратуры выше отдачи от получения PHD)
Заработная плата	Здоровье	Положительная связь Чем более здоровый человек, тем больше часов он способен трудиться, тем выше его зарплата
Заработная плата	Возраст	Положительная связь с затухающим эффектом (отдача в определенный период начинает уменьшаться) Зарплата как правило растет до 40-45 лет, достигает пика, потом начинает постепенно уменьшаться
Заработная плата	Наличие стресса	Отрицательная связь Чем больше стресса, тем менее собран человек, тем меньшая трудоспособность и производительность, заработная плата падает

Таблица №4. Предполагаемые направления связей переменной воздействия с контрольными.

Переменная воздействия	Контрольная переменная	Предполагаемое направление связи
Факт занятий спортом	Уровень образования	Положительная Чем выше уровень образования, тем больше индивид думает о своем здоровье

		(self-awareness), тем больше занимается спортом
Факт занятий спортом	Здоровье	Положительная Чем более здоровый человек, тем больше он занимается спортом (когда болеет - не может заниматься -противопоказания) (факт здоровья определяет занятие спортом)
Факт занятий спортом	Возраст	Отрицательная Чем старше человек, тем меньше он занимается спортом
Факт занятий спортом	Наличие стресса	Отрицательная (факт депрессии определяет занятие спортом) Лень, нет настроения, чтобы заниматься

3. Симулируйте данные в соответствии с предполагаемым вами процессом и приведите корреляционную матрицу, а также таблицу со следующими описательными статистиками:

Рис. № 3. Корреляционная матрица

```
df.corr()
```

	sport_activities	education	health	age	stress	abilities	GYM_distance	wage
sport_activities	1.000000	0.080402	0.727078	0.260242	-0.262786	0.047658	0.880715	0.524711
education	0.080402	1.000000	-0.001107	-0.012890	-0.008016	-0.011750	0.058082	0.247725
health	0.727078	-0.001107	1.000000	0.004515	-0.009573	-0.010871	0.831320	0.453726
age	0.260242	-0.012890	0.004515	1.000000	-0.012468	-0.012372	0.127704	-0.075904
stress	-0.262786	-0.008016	-0.009573	-0.012468	1.000000	0.007752	-0.285617	-0.120076
abilities	0.047658	-0.011750	-0.010871	-0.012372	0.007752	1.000000	-0.019369	0.521231
GYM_distance	0.880715	0.058082	0.831320	0.127704	-0.285617	-0.019369	1.000000	0.503147
wage	0.524711	0.247725	0.453726	-0.075904	-0.120076	0.521231	0.503147	1.000000

Для непрерывных переменных: выборочное среднее, выборочное стандартное отклонение, медиана, минимум и максимум

Рис. № 4. Дескриптивная таблица для непрерывных переменных

```
df[['education','age','abilities','wage']].describe()
```

	education	age	abilities	wage
count	10000.000000	10000.000000	10000.000000	10000.000000
mean	11.401900	45.020600	25.051200	59.183011
std	5.399249	20.831557	9.817353	19.244841
min	4.000000	14.000000	1.000000	3.345368
25%	7.000000	28.000000	18.000000	45.376292
50%	11.000000	44.000000	25.000000	57.521862
75%	15.000000	61.000000	32.000000	71.628567
max	30.000000	80.000000	50.000000	151.704835

Для бинарных переменных: доля и количество единиц.

Рис. № 5. Дескриптивная таблица для бинарных переменных

```
df_cat
```

	mean	share of 1
health	0.5088	5088
stress	0.1023	1023
sport_activities	0.4057	4057
GYM_distance	0.4172	4172

Указания:

- Необходимо сгенерировать не менее 1000 наблюдений (Число наблюдений n=10000)
- Доля единиц не должна быть меньше 0.1 ни для одной из бинарных переменных. (Health = 0.5, Stress = 0.1)

4. Разделите выборку на обучающую и тестовую. Тестовая выборка должна включать от 20% до 30% наблюдений.

Test_size = 20%

Рис. № 6. Разбиение выборки на обучающую и тестовую

```
[173] from sklearn.model_selection import train_test_split

    # Разделим целевую переменную и признаки
    target = df.loc[:, ['sport_activities']]                    # целевая переменная
    features = df.loc[:, df.columns.drop(['wage', 'GYM_distance', 'abilities', 'sport_activities'])]    # матрица признаков
    target = np.squeeze(target)                    # преобразуем из вектора столбца
                                                   # в одномерный массив
    # Разделим выборку на обучающую и тестовую
    features_train, features_test, target_train, target_test = train_test_split(
        features, target, test_size = 0.2, random_state = 777)

    # Сохраним число наблюдений обучающей и тестовой выборок
    n_train = len(target_train)
    n_test  = len(target_test)

    # Вернем исходную сортировку индексов
    features_train = features_train.reset_index(drop = True)
    target_train   = target_train.reset_index(drop = True)
    features_test  = features_test.reset_index(drop = True)
    target_test    = target_test.reset_index(drop = True)
```

3. Классификация

В каждом из заданий, если не сказано иного, необходимо использовать хотя бы 3 (на ваш выбор) из следующих методов: наивный Байесовский классификатор, метод ближайших соседей, случайный лес, градиентный бустинг и логистическая регрессия.

Будем использовать следующие методы: случайный лес, градиентный бустинг и KNN.

1. Отберите признаки, которые могут быть полезны при прогнозировании переменной воздействия и кратко обоснуйте выбор каждой из них. Не включайте в число этих признаков целевую переменную.

Целевая переменная: заработная плата индивида (ее временно опускаем)

Переменная воздействия: факт занятий спортом

Признаки, которые определяют переменную «занятия спортом»:

- **Образование** (сильно влияет на занятия спортом, когда человек образован, он знает о потребностях своего организма больше, заботиться о нем ответственнее - занимается спортом интенсивнее)
- **Здоровье** (возможность занятий спортом определяется уровнем здоровья, так, например, людям с тяжёлыми хроническими заболеваниями противопоказаны занятия спортом, сильная положительная связь)

- **Возраст** (чем старше человек, тем меньше он занимается спортом, так как предпочитает сидеть дома и играть в пасьянс на компьютере, сильная отрицательная связь)
- **Наличие стресса** (чем больше нервничает человек, тем больше он ленится, лежит в кровати, тем меньше он занимается спортом, отрицательная связь)

2. Выберите произвольные значения гиперпараметров, а затем оцените и сравните (между методами) точность прогнозов:

• на обучающей выборке.
• на тестовой выборке.
• с помощью кросс-валидации (используйте только обучающую выборку).

Таблица №5. Точность прогнозов для трех методов классификации

	Random forest	**Gradient boosting**	**KNN**
Произвольные значения выбранных гиперпараметров	Max_depth = 12 Max_features = ‘sqrt’ Max_samples = 500 Random_state =777 Criterion = ‘entropy’	N_estimators = 50 Max_depth = 3 Learning_rate = 0.5 Objective = 'binary:logistic' Random_state = 123	n_neighbors = 3 metric = "minkowski" p = 2
На обучающей выборке	ACC_train_rf = 0.9456	ACC_train_gb = 0.9459	0.94675
На тестовой выборке	ACC_test_rf = 0.939	ACC_test_gb = 0.938	0.9275
С помощью кросс-валидации	ACC_CV_total_rf = 0.9416	ACC_CV_total_gb= 0.9419	ACC_CV_total_knn= 0.9298

В целом все три метода дают близкие значения показателя Accuracy (точности прогнозов).

XGBoost показывает лучший результат по сравнению с Random Forest и KNN в общем.

KNN хуже всех предсказывает на тестовой выборке.

3. Для каждого метода с помощью кросс-валидации на обучающей выборке подберите оптимальные значения гиперпараметров (тюнинг).

   В качестве критерия качества используйте точность ACC. Результат представьте в форме таблицы, в которой для каждого метода должны быть указаны:

   - изначальные и подобранные значения гиперпараметров.
   - кросс-валидационная точность на обучающей выборке с исходными и подобранными значениями гиперпараметров.
   - точность на тестовой выборке с исходными и подобранными значениями гиперпараметров.

   Проинтерпретируйте полученные результаты и далее используйте методы с подобранными значениями гиперпараметров.

Таблица №6. Точность прогнозов для трех методов классификации и их параметры до/после тюнинга

Случайный лес

	Metric/Feature	Before Tuning	After Tuning
0	Accuracy on Test Data	0.939	0.9415
1	Accuracy on CV Data	0.941625	0.940125
2	bootstrap	True	True
3	ccp_alpha	0.0	0.0
4	class_weight	None	None
5	criterion	entropy	gini
6	max_depth	12	10
7	max_features	sqrt	sqrt
8	max_leaf_nodes	None	None
9	max_samples	500	None
10	min_impurity_decrease	0.0	0.0
11	min_samples_leaf	1	2
12	min_samples_split	2	5
13	min_weight_fraction_leaf	0.0	0.0
14	n_estimators	100	120
15	n_jobs	None	None
16	oob_score	False	False
17	random_state	777	777
18	verbose	0	0
19	warm_start	False	False

KNN

	Metric/Feature	Before Tuning	After Tuning
0	Accuracy on Test Data	0.9275	0.9345
1	Accuracy on CV Data	0.92975	0.938125
2	algorithm	auto	auto
3	leaf_size	30	30
4	metric	minkowski	minkowski
5	metric_params	None	None
6	n_jobs	None	None
7	n_neighbors	3	7
8	p	2	1
9	weights	uniform	uniform

Градиентный бустинг

	Metric/Feature	Before Tuning	After Tuning
0	Accuracy on Test Data	0.938	0.941
1	Accuracy on CV Data	0.941875	0.943875
2	objective	binary:logistic	binary:logistic
3	base_score	None	None
4	booster	None	None
5	callbacks	None	None
6	colsample_bylevel	None	None
7	colsample_bynode	None	None
8	colsample_bytree	None	None
9	device	None	None
10	early_stopping_rounds	None	None
11	enable_categorical	False	False
12	eval_metric	None	None
13	feature_types	None	None
14	gamma	None	None
15	grow_policy	None	None
16	importance_type	None	None
17	interaction_constraints	None	None
18	learning_rate	0.5	0.1
19	max_bin	None	None
20	max_cat_threshold	None	None
21	max_cat_to_onehot	None	None
22	max_delta_step	None	None
23	max_depth	3	4
24	max_leaves	None	None
25	min_child_weight	None	None
26	missing	NaN	NaN
27	monotone_constraints	None	None
28	multi_strategy	None	None
29	n_estimators	50	50
30	n_jobs	None	None
31	num_parallel_tree	None	None
32	random_state	123	123
33	reg_alpha	None	None
34	reg_lambda	None	None
35	sampling_method	None	None
36	scale_pos_weight	None	None
37	subsample	None	None
38	tree_method	None	None
39	validate_parameters	None	None
40	verbosity	None	None

Таблица №7. Интерпретация полученных результатов

Random forest	Gradient boosting	KNN
Модель с подобранными гиперпараметрами показывает точность немного ниже при проверке обоими способами. Такое возможно, если мы сначала случайно попали в достаточно хорошие параметры, но потом при переборе по сетке даже не включали в список возможных параметров потенциально лучшие комбинации. Однако если включить больше комбинаций, CV займет очень много времени.	Изменились некоторые опции параметров при переборе. При этом точность почти не изменилась при оценке обоими способами.	Изменились некоторые параметры при переборе. При этом точность повысилась при оценке обоими способами.

Повышенная сложность: подберите на обучающей выборке оптимальные значения гиперпараметров случайного леса ориентируясь на значение OOB (out-of-bag) ошибки. Сопоставьте гиперпараметры и точность на тестовой выборке для случайного леса в зависимости от того, используется кросс-валидация или OOB ошибка. Объясните преимущество OOB ошибки по сравнению с кросс- валидацей.

Таблица №8. Сравнение параметров и точности для CV и OOB

	CV	OOB
Best parameters Из тех параметров, что перебирались: изменилась max_depth, min_samples_split, min_samples_leaf.	'bootstrap' = True 'ccp_alpha' = 0.0 'class_weight' = None 'criterion' = 'gini' 'max_depth' = 10 'max_features' = 'sqrt' 'max_leaf_nodes' = None 'max_samples' = None 'min_impurity_decrease'= 0.0 'min_samples_leaf' = 2 'min_samples_split' = 5 'min_weight_fraction_leaf'= 0.0 'n_estimators' = 120 'n_jobs' = None 'oob_score' = False 'random_state' = 777 'verbose' = 0 'warm_start'= False	n_estimators = 120 max_depth = 15 min_samples_split = 2 min_samples_leaf =4 max_features = 'sqrt' oob_score = 0.94263
Accuracy on test	0.9415	0.94

Преимущество OOB ошибки - обычно быстрее, чем кросс-валидация, поскольку не нужно многократно обучать модель. Для каждой комбинации параметров модель обучалась однажды, но не 5 раз с теми же параметрами, как, например, в случае 5-fold CV.

Заметим, что на тесте точность меньше для OOB-подбора

4. Повторите предыдущий пункт, используя любой альтернативный критерий качества модели. Обоснуйте возможные преимущества и недостатки этого альтернативного критерия.

Будем использовать F1-score.

Преимущества F1-оценки:

- Учитывает баланс между точностью и полнотой: F1-оценка учитывает как точность (долю истинно положительных среди всех положительных предсказаний), так и полноту (долю истинно положительных среди всех реальных положительных классов), что делает ее хорошим компромиссом между этими двумя метриками.

- Чувствительна к дисбалансу классов: F1-оценка подходит для оценки моделей на несбалансированных наборах данных, где количество примеров в каждом классе значительно различается. Она помогает учитывать неравномерное распределение классов.

Недостатки F1-оценки:

- Не учитывает верно отрицательные: F1-оценка не учитывает верно отрицательные (True Negatives), поэтому может быть не подходящей для задач, где важна их доля.
- Зависимость от порога классификации: F1-оценка зависит от порога классификации, который используется для принятия решения о принадлежности к классу. Оптимальный порог может варьироваться в зависимости от конкретного приложения или требований бизнеса.

Таблица №9. Сравнение параметров и точности для CV и OOB

Random forest	Gradient boosting	KNN
Модель с подобранными гиперпараметрами показывает f1 немного ниже при проверке обоими способами. Такое возможно, если мы сначала случайно попали в достаточно хорошие параметры, но потом при переборе по сетке даже не включали в список возможных параметров потенциально лучшие комбинации. Однако если включить больше комбинаций, CV займет очень много времени.	F1 на тесте почти одинаковая, F1 по CV немного хуже после перебора гиперпараметров, но стоит учитывать, что мы используем очень узкую сетку из-за ограничения по времени обучения.	После тюнинга F1-мера повысилась для двух способов оценивания.

	Metric/Feature	Before Tuning	After Tuning
0	f1 on Test Data	0.924318	0.923171
1	f1 on CV Data	0.927024	0.918975
2	bootstrap	True	True
3	ccp_alpha	0.0	0.0
4	class_weight	None	None
5	criterion	entropy	gini
6	max_depth	12	None
7	max_features	sqrt	sqrt
8	max_leaf_nodes	None	None
9	max_samples	500	None
10	min_impurity_decrease	0.0	0.0
11	min_samples_leaf	1	1
12	min_samples_split	2	5
13	min_weight_fraction_leaf	0.0	0.0
14	n_estimators	100	130
15	n_jobs	None	None
16	oob_score	True	False
17	random_state	777	777
18	verbose	0	0
19	warm_start	False	False

	Metric/Feature	Before Tuning	After Tuning
0	f1 on Test Data	0.923362	0.926799
1	f1 on CV Data	0.927756	0.918975
2	objective	binary:logistic	binary:logistic
3	base_score	None	None
4	booster	None	None
5	callbacks	None	None
6	colsample_bylevel	None	None
7	colsample_bynode	None	None
8	colsample_bytree	None	None
9	device	None	None
10	early_stopping_rounds	None	None
11	enable_categorical	False	False
12	eval_metric	None	None
13	feature_types	None	None
14	gamma	None	None
15	grow_policy	None	None
16	importance_type	None	None
17	interaction_constraints	None	None
18	learning_rate	0.5	0.1
19	max_bin	None	None

	Metric/Feature	Before Tuning	After Tuning
0	f1 on Test Data	0.911854	0.920343
1	f1 on CV Data	0.912883	0.922994
2	algorithm	auto	auto
3	leaf_size	30	30
4	metric	minkowski	minkowski
5	metric_params	None	None
6	n_jobs	None	None
7	n_neighbors	3	7
8	p	2	2
9	weights	uniform	uniform

Повышенная сложность: дополнительно самостоятельно запрограммируйте не представленный в стандартных библиотеках критерий качества и используйте его для тюнинга гиперпараметров. Сравните результат стандартного и вашего критериев.

Создадим новый скорер как комбинацию с весом альфа двух предыдущих метрик.

Таблица №10. Сравнение методов с новым скорером

Random forest	Gradient boosting	KNN
После тюнинга качество модели чуть-чуть ухудшилось при оценке обоими способами. Это схоже с результатами стандарнтных критериев.	F1 на тесте почти одинаковая, F1 по CV тоже почти совпадает. Это почти не отличается от результатов стандартных метрик.	После тюнинга F1-мера повысилась для двух способов оценивания. Это совпадает со стандартными метриками.

	Metric/Feature	Before Tuning	After Tuning
0	Own metric on Test Data	0.931659	0.930336
1	Own metric on CV Data	0.934325	0.926987
2	bootstrap	True	True
3	ccp_alpha	0.0	0.0
4	class_weight	None	None
5	criterion	entropy	gini
6	max_depth	12	None
7	max_features	sqrt	sqrt
8	max_leaf_nodes	None	None
9	max_samples	500	None
10	min_impurity_decrease	0.0	0.0
11	min_samples_leaf	1	1
12	min_samples_split	2	5
13	min_weight_fraction_leaf	0.0	0.0
14	n_estimators	100	130
15	n_jobs	None	None
16	oob_score	True	False
17	random_state	777	777
18	verbose	0	0
19	warm_start	False	False

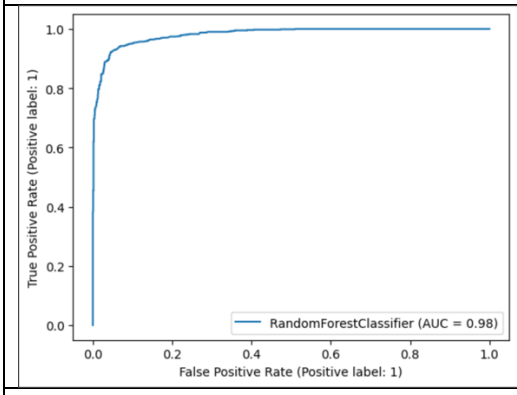
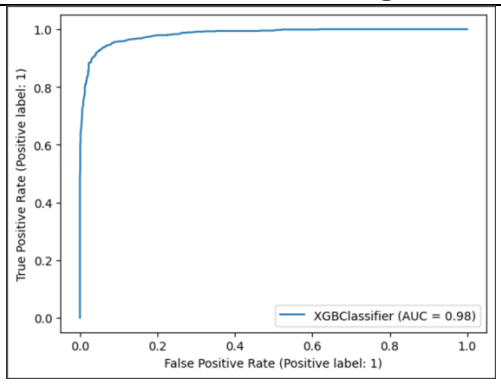
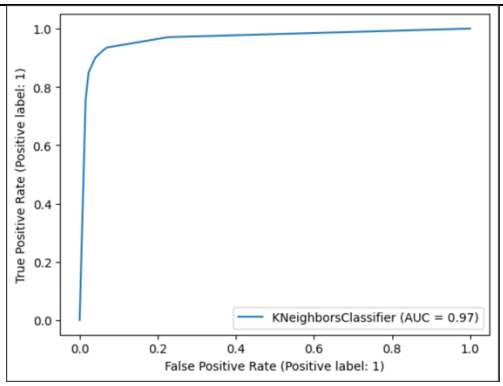
	Metric/Feature	Before Tuning	After Tuning
0	facc on Test Data	0.930681	0.9339
1	facc on CV Data	0.934815	0.937087
2	objective	binary:logistic	binary:logistic
3	base_score	None	None
4	booster	None	None
5	callbacks	None	None
6	colsample_bylevel	None	None
7	colsample_bynode	None	None
8	colsample_bytree	None	None
9	device	None	None
10	early_stopping_rounds	None	None
11	enable_categorical	False	False
12	eval_metric	None	None
13	feature_types	None	None
14	gamma	None	None
15	grow_policy	None	None
16	importance_type	None	None
17	interaction_constraints	None	None
18	learning_rate	0.5	0.1
19	max_bin	None	None

	Metric/Feature	Before Tuning	After Tuning
0	facc on Test Data	0.919677	0.927672
1	facc on CV Data	0.921316	0.93056
2	algorithm	auto	auto
3	leaf_size	30	30
4	metric	minkowski	minkowski
5	metric_params	None	None
6	n_jobs	None	None
7	n_neighbors	3	7
8	p	2	2
9	weights	uniform	uniform

5. Постройте ROC-кривую для ваших моделей и сравните их по AUC на тестовой выборке.

Будем использовать модели с гиперпараметрами, подобранными по accuracy. По AUC однозначно выигрывают RF и XGBoost.

Таблица №11. Сравнение кривой ROC для 3х методов классификации

Random forest	Gradient boosting	KNN
		
AUC_test_rf = 0.9819	AUC_test_gb = 0.9828	AUC_test_knn = 0.9657

XGBoost немного лучше остальных методов.

Повышенная сложность: дополнительно выполните это задание для Байесовской сети.

Дополнительно: проделаем то же самое для Байесовской сети.

Структура байесовской сети определяется направленным ациклическим графом (DAG - directed acyclic graph). Поэтому, сперва необходимо сформировать DAG, указав предполагаемые направления причинно-следственных связей между переменными.

Рис. № 6. Направленный ациклический граф

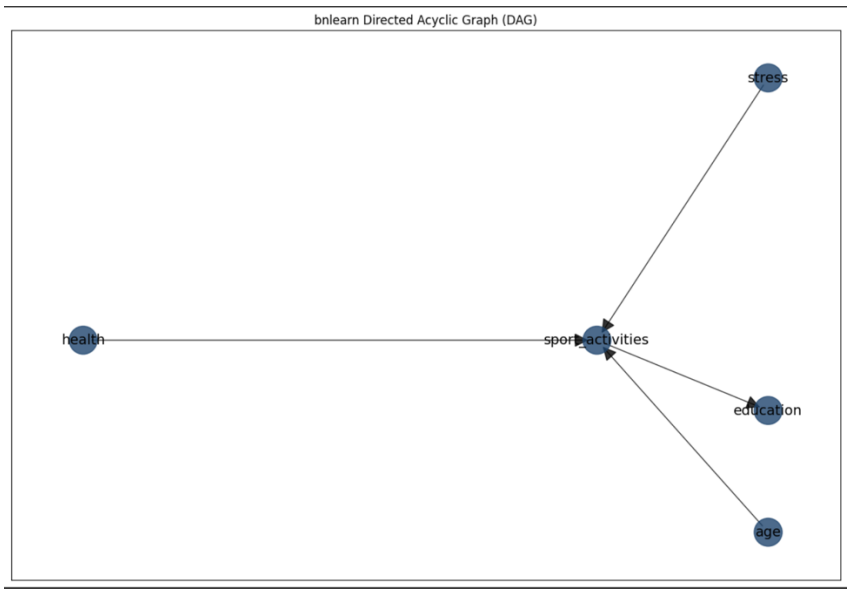
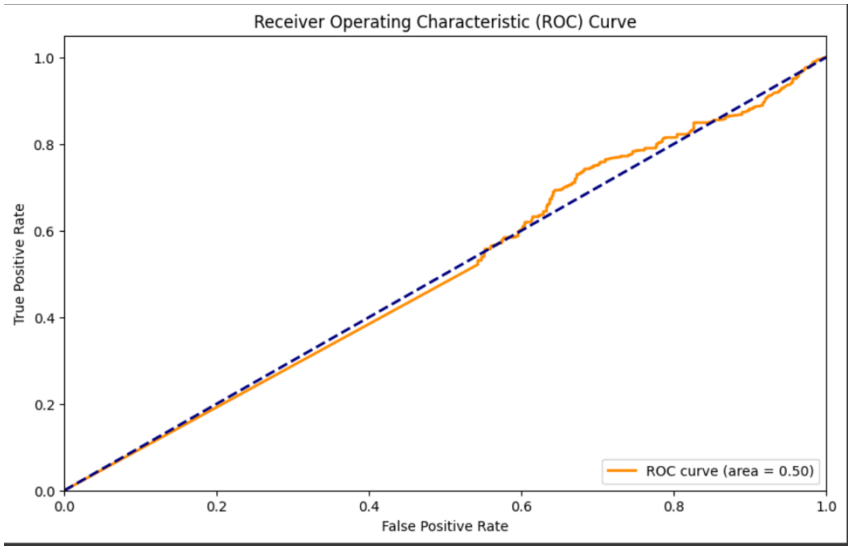


Рис. № 7. Кривая ROC



bn_AUC = 0.499

Скорее всего, автоматически подобраннный DAG неправильно описывает структуру зависимостей между переменными. Мы получили AUC сильно ниже, чем для других моделей, то есть BN не превосходит решения случайного выбора

6. Постройте матрицу путаницы и предположите цены различных видов прогнозов. Исходя из критерия максимизации прибыли на обучающей выборке подберите оптимальный порог прогнозирования для каждого из методов и сравните прибыли на тестовой выборке при соответствующих порогах.

   Результат представьте в форме таблицы, в которой должны быть указаны как AUC, так и прибыли (на тестовой выборке). Проинтерпретируйте полученный результат.

Таблица №12. Цены прогнозов, AUC и прибыли при оптимальных порогах прогнозирования

	Random forest	Gradient Boosting	KNN
Бизнес задача:	Оценка точности и ценности прогнозов о занятиях спортом поможет администрации эффективно распределить ресурсы и максимизировать положительный эффект программы на здоровье населения.		
TP	752	747	751
Бизнес значение:	Такие люди будут улучшать свое здоровье, снижать нагрузку на медицинскую систему. Цена: $500 - Это включает в себя экономию на медицинских расходах (предотвращение заболеваний), а также увеличение производительности труда за счет улучшенного здоровья.		
TN	1131	1135	1118
Бизнес значение:	Это помогает избежать необоснованных затрат на спортивные мероприятия для тех, кто не будет участвовать. Цена: $100 - Это экономия на затратах, связанных с предоставлением спортивных мероприятий, которые не будут востребованы.		
FP	54	50	67
Бизнес значение:	Ресурсы будут затрачены впустую на человека, который не будет заниматься спортом. Цена: -$200 - Это потерянные инвестиции в спортивные мероприятия и пропаганду, которые не принесут ожидаемых результатов.		
FN	63	68	64

Бизнес значение:	Потенциально активные граждане не получат нужного внимания и ресурсов, что приводит к упущенной возможности улучшения здоровья населения. Цена: -$300 - Это упущенная возможность улучшения здоровья граждан, что может привести к дополнительным медицинским расходам и потерям в производительности труда.		
AUC score	AUC_test_rf = 0.982	AUC_test_gb = 0.983	AUC_test_knn = 0.967
Profit	4587	4646	4498

Заметим, что AUC может служат прокси для прибыли, когда мы не знаем цены, но все равно хотим сравнить модели. Наилучший результат показывает градиентный бустинг.

Повышенная сложность: предложите, содержательно обоснуйте и примените собственную, отличную от линейной функцию прибыли от прогнозов.

Предложим другую функцию прибыли. Государство сильно потеряет в доверии избирателей, если они увидят пустующие спорт. площадки, на которые вообще-то были потрачены их деньги. Это будет очевидная трата ресурсов впустую, поэтому введём штраф за большие отклонения FP от 0.

Порог сдвинулся в сторону 0. Это логично, ведь наша модель теперь очень боится выдавать FN, и будет их избегать. Значит, есть стимулы как можно больше вхождений классифицировать как 1, и порог сдвинется к 0.

Таблица №13. Новая функция прибыли от прогнозов - влияние на результаты 3х методов

Method	AUC Score	Profit linear	Profit nonlinear
RF	0.981853	4587	3133
XGB	0.982832	4646	4007
KNN	0.965674	4498	2608

В случае, когда мы сильно штрафуем за FN, пороги сдвигаются к 0. Лучший профит в случае нелинейной модели получаем в XGBoost оценивании.

7. Опишите предполагаемые связи между переменными в форме ориентированного ациклического графа (DAG). Обучите структуру Байесовской сети на обучающей выборке и сравните точность прогнозов вашего и обученного DAG на тестовой выборке.

Обученный DAG уже получили выше. Попробуем самостоятельно описать структуру сети.

Предполагаемая структура DAG:

Возраст → Образование:

Возраст влияет на образование, так как по мере взросления люди обычно проходят через различные уровни образовательных учреждений, от начальной школы до высших учебных заведений.

Возраст → Здоровье:

Возраст напрямую влияет на здоровье, так как с возрастом здоровье обычно меняется. Пожилые люди могут сталкиваться с большими проблемами со здоровьем по сравнению с молодыми.

Здоровье → Стресс:

Состояние здоровья значительно влияет на уровень стресса. Плохое здоровье может приводить к большему стрессу из-за физического дискомфорта и беспокойства о благополучии.

Здоровье → Занятия спортом:

Здоровье влияет на способность и мотивацию заниматься спортом. Хорошее здоровье обычно способствует более активному участию в спортивных мероприятиях.

Образование → Стресс:

Образование может влиять на уровень стресса. Более высокие образовательные требования могут приводить к увеличению стресса, в то время как лучшее образование может дать навыки для управления стрессом.

Стресс → Занятия спортом:

Уровень стресса может влиять на вероятность занятия спортом. Высокий стресс может снижать участие в спорте, в то время как некоторые люди могут использовать спорт для снижения стресса.

Рис. № 8. Направленный ациклический граф (обученная структура)

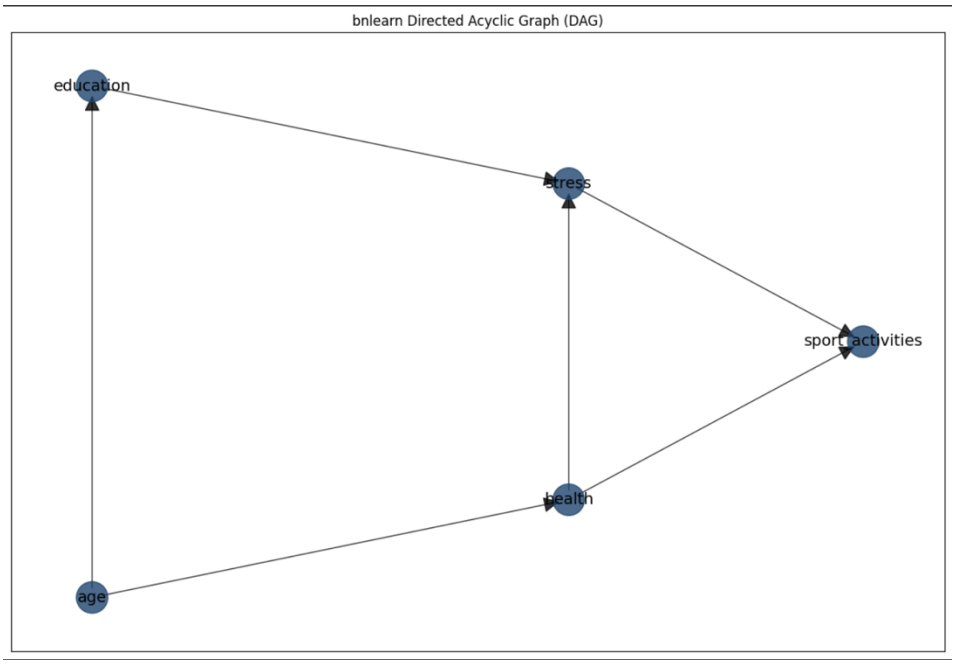


Таблица №14. Сравнение результатов, метод DAG

“Наш” DAG	Точность DAG с обученной структурой
0.899	0.936

Точность нашего DAG достаточно высокая, но она несколько уступает точности DAG с обученной структурой.

8. На основании проделанного анализа выберите лучший и худший из обученных классификаторов. Обоснуйте сделанный выбор.

Мы обучили случайный лес, градиентный бустинг и метод ближайших соседей + дополнительно построили классификатор на основании байесовских сетей. Первые три классификатора мы можем сравнить не только по точности и AUC, но и по прибылям и дополнительным метрикам качества. По точности, f1-мере и сбалансированной метрике 'точность/f1-мера' XGBoost показывал качество не хуже двух других моделей, тогда как KNN -- не лучше двух других моделей, как на оценке на тестовой выборке, так и на оценке с помощью CV. Подбор гиперпараметров обычно немного ухудшал качество случайного леса, почти не улучшал качество XGBoost, но немного улучшал качество KNN. Тем не менее, исходя

только из перечисленных мтерик качества, XGBoost оказывался лучшим классификатором, а KNN худшим.

При оптимальном пороге XGBoost также позволял получить наибольшую прибыль как в линейном, так и в нелинейном случае записи функции. KNN показывал наименьшую прибыль в обоих случаях. По AUC XGBoost тоже опережает другие модели, а KNN справляется немного хуже.

Вывод: лучшим классификатором стоит признать XGBoost с подобранными гиперпараметрами, худшим -- KNN. Байесовские сети показывали точность на тестовой выборке не лучше KNN.

9. Повышенная сложность: включите в анализ дополнительный метод классификации, не рассматривавшийся в курсе и не представленный в библиотеке scikit-learn. Опишите данный метод (принцип работы, преимущества и недостатки) и осуществите тюнинг гиперпараметров. Сопоставьте его точность на тестовой выборке с точностью лучшего из обученных вами ранее методов. (все методы, приходящие нам в голову, оказывались в scikit-learn, поэтому, к сожалению, этот пункт мы не сделали)

4. Регрессия

В каждом из заданий, если не сказано иного, необходимо использовать хотя бы 3 (на ваш выбор) из следующих методов: случайный лес, метод наименьших квадратов, метод ближайших соседей и градиентный бустинг.

Будем использовать следующие методы: случайный лес, градиентный бустинг и KNN.

1. Отберите признаки, которые могут быть полезны при прогнозировании целевой (зависимой) переменной. Не включайте в число этих признаков переменную воздействия.

   Целевая переменная: заработная плата индивида.

   Переменную воздействия опустим.

Полезные признаки при прогнозировании заработной платы: образование, здоровье, возраст, стресс. Способности также были бы полезны, но обычно мы их не наблюдаем.

2. Выберите произвольные значения гиперпараметров, а затем оцените и сравните (между методами) точность прогнозов с помощью RMSE и MAPE:

Таблица №15. Выбранные произвольные значения гиперпараметров:

Random Forest	Gradient Boosting	KNN
max_depth = 12 max_features = "sqrt" max_samples = 500 random_state = 777 criterion = 'friedman_mse'	n_estimators = 50 max_depth = 3 learning_rate = 0.5 reg_lambda = 0.01 random_state = 123	n_n_neighbors = 40 metric = "minkowski" p = 2

• на обучающей выборке
• на тестовой выборке
• с помощью кросс-валидации (используйте только обучающую выборку)

Таблица №16. MAPE & RMSE при выбранных произвольных значениях гиперпараметров:

Model	Train_RMSE	Train_MAPE	Test_RMSE	Test_MAPE	CV_RMSE	CV_MAPE
KNN	14.433137	0.219715	14.863763	0.224673	14.784337	0.225637
RF	14.099213	0.216197	14.857242	0.226944	14.774896	0.226231
XGBoost	14.262328	0.216963	14.842554	0.223293	14.816588	0.225427

Все модели дают очень похожие результаты на тесте и CV, тогда как на трейне KNN чуть хуже.

3. Для каждого метода с помощью кросс-валидации на обучающей выборке подберите оптимальные значения гиперпараметров (тюнинг). В качестве критерия качества используйте RMSE. Результат представьте в форме таблицы, в которой для каждого метода должны быть указаны:

- изначальные и подобранные значения гиперпараметров.
- кросс-валидационное значение RMSE на обучающей выборке с исходными и подобранными значениями гиперпараметров.
- значение RMSE на тестовой выборке с исходными и подобранными значениями гиперпараметров.

Таблица №17. Гиперпараметры модели до/после тюнинга и RMSE

Random Forest

	Metric/Feature	Before Tuning	After Tuning
0	RMSE on Test Data	14.857242	14.860656
1	RMSE on CV Data	14.774896	14.858173
2	MAPE on Test Data	0.226944	0.224351
3	MAPE on CV Data	0.226231	0.227108
4	bootstrap	True	True
5	ccp_alpha	0.0	0.0
6	criterion	friedman_mse	squared_error
7	max_depth	12	10
8	max_features	sqrt	sqrt
9	max_leaf_nodes	None	None
10	max_samples	500	None
11	min_impurity_decrease	0.0	0.0
12	min_samples_leaf	1	2
13	min_samples_split	2	7
14	min_weight_fraction_leaf	0.0	0.0
15	n_estimators	100	120

Модель с подобранными гиперпараметрами показывает лучшие результаты на тестовой выборке, но худшие при оценке по CV. Многие параметры модели изменились.

Gradient Boosting

	Metric/Feature	Before Tuning	After Tuning
0	RMSE on Test Data	14.842554	14.786599
1	RMSE on CV Data	14.816588	14.683641
2	MAPE on Test Data	0.223293	0.22383
3	MAPE on CV Data	0.225427	0.224228
4	objective	reg:squarederror	reg:squarederror
5	base_score	None	None
6	booster	None	None
7	callbacks	None	None
8	colsample_bylevel	None	None
9	colsample_bynode	None	None
10	colsample_bytree	None	None
11	device	None	None
12	early_stopping_rounds	None	None
13	enable_categorical	False	False
14	eval_metric	None	None
15	feature_types	None	None
16	gamma	None	None
17	grow_policy	None	None
18	importance_type	None	None
19	interaction_constraints	None	None

Изменились некоторые опции параметров при переборе. При этом качество модели стало чуть лучше при оценке обоими способами и по обеим метрикам.

KNN

	Metric/Feature	Before Tuning	After Tuning
0	RMSE on Test Data	14.863763	14.870343
1	RMSE on CV Data	14.784337	14.742299
2	MAPE on Test Data	0.224673	0.224739
3	MAPE on CV Data	0.225637	0.224982
4	algorithm	auto	auto
5	leaf_size	30	30
6	metric	minkowski	minkowski
7	metric_params	None	None
8	n_jobs	None	None
9	n_neighbors	40	50
10	p	2	1

Изменились некоторые пареметры при переборе. При этом качество модели почти не изменилось при оценке по обеим метрикам и обоими способами.

4. На основании проделанного анализа выберите лучший и худший из обученных классификаторов. Обоснуйте сделанный выбор.

Будем сравнивать модели после тюнинга по RMSE и MAPE. После тюнинга лучшее значение RMSE как на тестовой выборке, так и

на оценке по CV мы получили для XGBoost, то есть это лучшая модель. То же верно и для MAPE.

Ошибки на тестовой выборке выше у KNN, но на кросс-валидации у RF. В целом кросс-валидация учитывает больше информации о всей выборке, поэтому будем ориентироваться на нее при решении, какой метод хуже. В таком случае, RF показывает худший результат.

Вывод: лучший метод это XGBoost, худший - RF.

5. Повышенная сложность: включите в анализ дополнительный метод регрессии, не рассматривавшиеся в курсе и не представленный в библиотеке scikit- learn. Опишите данный метод (принцип работы, преимущества и недостатки) и осуществите тюнинг гиперпараметров. Сопоставьте его точность на тестовой выборке с точностью лучшего из обученных вами ранее методов. (все методы, приходящие нам в голову, оказывались в scikit-learn, поэтому, к сожалению, этот пункт мы не сделали)

5. Эффекты воздействия

При выполнении данного задания необходимо объединить обучающую и тестовую выборки в одну.

1. Математически запишите и содержательно проинтерпретируйте потенциальные исходы целевой переменной. Объясните, как они связаны с наблюдаемыми значениями целевой переменной.

В качестве инструментальной переменной имеется GYM_distance_{i}, переменной воздействия - $\text{Sport_activities}_{i}$, целевой переменной - $\text{Wage}_{i}Контрольные переменные включают в себя $ \text{Education}_{i},\text{Health}_{i},\text{Age}_{i},\text{Stress}_{i}$

Для того, чтобы впоследствии анализировать локальные средние эффекты воздействия LATE, необходимо различать

величину переменной воздействия $\text{Sport_activities}_{i}$ в зависимости от значения инструмента GYM_distance_{i}. Для этого рассмотрим ни от чего не зависящую равномерную случайную величину $U_{i}\sim U(0,1)$ и введем гипотетические переменные:

$$\text{Sport\_activities}_{1i} = I(P(\text{Sport\_activities}_i = 1|\text{Education}_i, \text{Health}_i, \text{Age}_i, \text{Stress}_i, \text{GYM\_distance}_i = 1) \geq U_i)$$

$$\text{Sport\_activities}_{0i} = I(P(\text{Sport\_activities}_i = 1|\text{Education}_i, \text{Health}_i, \text{Age}_i, \text{Stress}_i, \text{GYM\_distance}_i = 0) \geq U_i)$$

Наблюдаемый уровень занятия спортом можно выразить как:

$$"\{Sport\_activities\}\_\{i\} = \ \llbracket\{cases\}"\{Sport\_activities\}\_\{1i\}"\{, \text{если }\}"\{GYM\_distance\}\_\{i\} = 1\backslash\backslash\ "\{Sport\_activities\}\_\{0i\}"\{, \text{если }\}"\{GYM\_distance\}\_\{i\} = 0\rrbracket\ \{cases\}$$

Напомним, что к соблюдателям относятся те, у кого $\text{Sport_activities}_{1i} > \text{Sport_activities}_{0i}$, то есть занимаются спортом при $\text{GYM_distance}_{i}=1$ и не занимаются - при $\text{GYM_distance}_{i} = 0$.

Переменная воздействия (факт занятия спортом) положительно влияет на заработную плату с коэффициентом 2:

wage1 = 2 * alpha * sport_activities + g1 + error_wage1

wage0 = alpha * sport_activities + g0 + error_wage0

То есть мы подразумеваем, что люди, занимающиеся спортом в среднем, зарабатывают больше, что мы и будем анализировать в дальнейшем, смотря на оценки эффекта воздействия

2. Используя симулированные вами, но недоступные в реальных данных потенциальные исходы (гипотетические значения), получите оценки среднего эффекта воздействия, условных средних эффектов воздействия и локального среднего эффекта воздействия. Результаты представьте в форме таблицы.

Таблица №18. Эффекты воздействия

TE	ATE	LATE	ATET	CATE
[-43.05	-7.948	-8.92	-7.43	-18.18
-42.54				-43.80
-27.95				-28.28
27.16				28.34
-11.84				-14.17
-24.31				-6.71
-0.15				-7.27
10.66				19.35
-1.72				12.97
-30.11				-5.48

3. Оцените средний эффект воздействия как разницу в средних по выборкам тех, кто получил и не получил воздействие. Опишите недостатки соответствующего подхода с учетом специфики рассматриваемой вами экономической проблемы.

Допущение о независимости:

$$\begin{aligned} &\mathrm{E}(\mathrm{Wage}_{1i}|\mathrm{Sport\_activities}_i = 1) \\ &\quad = \mathrm{E}(\mathrm{Wage}_{1i}) \qquad \mathrm{E}(\mathrm{Wage}_{0i}|\mathrm{Sport\_activities}_i = 0) \\ &\quad = \mathrm{E}(\mathrm{Wage}_{0i}) \end{aligned}$$

Наивный подход предполагает оценивание ATE как средней разницы в зарплатах людей с высшим образованием и без высшего образования.

$$\widehat{\mathrm{ATE}_{\mathrm{naive}}} = \frac{1}{n_1} 1\sum i\colon \mathrm{Educ}_i = 1\mathrm{Wage}_{1i} \quad \frac{1}{n_0} \sum_{i:\mathrm{Educ}_i=0\mathrm{Wage}_{0i}}$$

Таблица №19. Эффект воздействия как разница средних

	Оценка
ATE	-7.948
ATE_naive	10.6329

Проблема: Факт увлечения спортом распределяется неслучайно между индивидами и зависит как от контрольных переменных, так и от целевой переменной. Социальные данные никогда не характеризуются какой-либо независимостью связей показателей: все скоррелировано и взаимосвязано

4. Используя оценки, полученные лучшими из обученных ранее классификационных и регрессионных моделей, оцените средний эффект воздействия с помощью:
    - метода наименьших квадратов.
    - условных математических ожиданий.
    - взвешивания на обратные вероятности (в случае возникновения ошибок убедитесь в отсутствии оценок вероятностей, равных 0 или 1 и при необходимости измените метод оценивания).
    - метода, обладающего двойной устойчивостью.
    - двойного машинного обучения.

Сравните результаты и назовите ключевую предпосылку этих методов. Содержательно обсудите причины, по которым она может соблюдаться или нарушаться в вашем случае. Приведите содержательную экономическую интерпретацию оценки среднего эффекта воздействия.

Повышенная сложность: включите дополнительный метод, не рассматривавшийся в курсе, и опишите его принцип работы, а также преимущества и недостатки по сравнению с другими методами.

Таблица № 20. Средние эффекты воздействия

	МНК	Условн.	Вероят.	2Уст.	2МО
Значение	-0.5337	1.5922	-3.7114	-1.1612	-2.2415

Главная предпосылка оценки эффектов воздействия – допущение о независимости, то есть условное матожидание зарплаты равно безусловному матожиданию зарплаты для каждого значения переменной воздействия. Нарушаться эта предпосылка может из-за отсутствия случайности при присваивании эффекта воздействия: в нашем случае это тот факт, что условная ожидаемая зарплата среди людей, занимающихся спортом не равна ожидаемой заплате занимающихся спортом людей из числа всей выборки.

Большинство методов продемонстрировало отрицательный эффект воздействия, что говорит об обратной связи заработной платы и факта занятия спортом

5. Оцените локальный условный эффект воздействия с помощью:

   • двойного машинного обучения без инструментальной переменной. • двойного машинного обучения с инструментальной переменной.

Сопоставьте результаты и объясните, в чем в вашем случае будет заключаться различие между средним эффектом воздействия и локальным средним эффектом воздействия. Приведите содержательную экономическую интерпретацию оценки локального среднего эффекта воздействия.

Повышенная сложность: воспользуйтесь также параметрической моделью, например, с помощью пакета switchSelection. Обсудите преимущества и недостатки такого подхода по сравненинию с двойным машинным обучением.

Таблица №21. Локальный эффект воздействия

	-Инструмент	+Инструмент
Значение	-2.322701	9.865216

С включением инструментальной переменной эффект воздействия становится положительным.

Допустим экзогенность (валидность) инструмента. Без введения дополнительных строгих допущений, например, о том что эффект воздействия Ti на Yi является одинаковым для всех индивидов, в общем случае оценить ATE не получится, но можно оценить локальный

средний эффект воздействия LATE, отражающий ATE среди соблюдателей.

6. Оцените условные средние эффекты воздействия с помощью:

• метода наименьших квадратов. • S-learner.
• T-learner.
• метода трансформации классов. • X-learner.

Сравните результаты и обсудите, насколько в вашем случае мотивированы применение метода X-learner. Опишите, как можно было бы использовать полученные вами оценки в бизнесе или при реализации государственных программ. Повышенная сложность: включите дополнительный метод, не рассматривавшийся в курсе и опишите его принцип работы, а также преимущества и недостатки по сравнению с другими методами.

Таблица №21. Условные эффекты воздействия

	МНК	S	T	X	Transform
0	0.0	2.452910	-2.196157	32.261574	-8.394755
1	0.0	-25.194430	-24.913964	15.208946	-26.019342
2	0.0	-10.810765	-15.586360	-39.034378	-34.573261
3	0.0	16.590565	28.427563	-28.266235	-29.828011
4	0.0	-8.095503	-3.713044	51.437386	1.457358
...	...	...	...	...	...
9995	0.0	-6.863930	6.746790	-19.181698	-19.869934
9996	0.0	1.774942	12.641007	-35.592121	-40.104492
9997	0.0	9.744814	12.767256	73.879318	12.659895
9998	0.0	-2.717137	-0.332256	66.461807	3.336450
9999	0.0	-7.014507	-9.153278	59.546806	-2.249998

X-learner не является необходимым, так как в данной работе группа воздействия содержит достаточное и сопоставимое число наблюдений. Полученные оценки можно использовать при внедрении программ продвижения спортивных мероприятий и здорового образа жизни.

7. Выберите лучшую модель оценивания условных средних эффектов воздействия, используя:

• истинные значения условных средних эффектов воздействия. • прогнозную точность моделей.
• псевдоисходы.

Проинтерпретируйте различия в результатах различных подходов.

Таблица №22. Сравнение на основе истинных значений

	MSE
LS	88.46
T-learner	94.778
S-learner	82.077
CT	633.56

Таблица №23. Сравнение на основе псевдоисходов

	MSE
LS	47276.81
T-learner	47488.849
S-learner	47337.127
CT	45985.764

На основе истинных значений наименьшей ошибкой обладает S-learner, в то время как на псевдоисходах трансформация классов имеет наименьшую метрику ошибки.

Таблица №24. Сравнение на основе ошибки моделей

	MSE
LS	207.646
T-learner	207.646
S-learner	186.858
CT	5299.1360

Исходя из прогнозных ошибок моделей наибольшей точностью обладает S-learner. Различия оценок заключаются в разном алгоритме оценок матожиданий, а также используются разные подвыборки (например, для оценки ошибки использовалась вся выборка)

8. Оцените средние эффекты воздействия и локальные средние эффекты воздействия используя худшие из обученных классификационных и регрессионных моделей. Сопоставьте результаты с теми, что были получены с помощью лучших моделей. Сдеайте вывод об устойчивости результатов к качеству используемых методов машинного обучения.

Таблица №25. Локальный эффект воздействия, сравнение

	-Инструмент	+Инструмент
Лучшая модель	-2.322701	9.865216
Худшая модель	-0.841517	9.559126

Таблица № 26. Средние эффекты воздействия, сравнение

	МНК	Условн.	Вероят.	2Уст.	2МО
Лучшая	-0.5337	1.5922	-3.7114	-1.1612	-2.2415
Худшая	0	-0.602559	NaN	NaN	-5.716958

В случае предсказания локальных эффектов худшие модели занижают эффекты воздействия, в то время как у средних эффектов ситуация иная. KNN выдает ошибки при прогнозе этих эффектов, а также они крайне непохожи на результаты, полученные лучшей моделью. Это говорит о высокой неустойчивости оценок к обучаемым моделям (оценкам условных матожиданий)

Ссылки:

1) Dan-Olof Rooth,Work out or out of work — The labor market return to physical fitness and leisure sports activities,Labour Economics,Volume 18, Issue 3,2011,Pages 399-409,ISSN 0927-5371,https://doi.org/10.1016/j.labeco.2010.11.006.

2) Kosteas, V.D. The Effect of Exercise on Earnings: Evidence from the NLSY. *J Labor Res* **33**, 225–250 (2012). https://doi.org/10.1007/s12122-011-9129-2