Машинное обучение в экономике. Проект. 2024–2025

Подойников Иван
Москва, 2025

Обоснование темы

Задание 1.1

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

Ответ

В данном исследовании в качестве целевой переменной выбран **годовой доход** (в тыс. руб.). Переменная воздействия представлена **участием в программах повышения квалификации** — бинарная переменная, где значение 1 означает, что индивид прошёл дополнительное обучение, а 0 — не прошёл.

Задание 1.2

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.

Ответ

Для бизнеса это позволит:

1. Оптимизировать затраты на обучение персонала.
2. Повысить эффективность управления человеческим капиталом.
3. Разработать целевые стратегии по повышению квалификации сотрудников, что может привести к росту производительности и, соответственно, доходов.

Для государственных органов результаты исследования могут помочь в:

1. Разработке программ поддержки профессионального развития.
2. Повышении конкурентоспособности экономики через улучшение квалификации рабочей силы.

Для образовательных учреждений и научных организаций результаты этой работы смогут:

1. Послужить фундаментом для академических исследований в области оценки эффективности образовательных интервенций.

2. Посодействовать развитию индивидуального профессионального роста, предоставляя специалистам информацию для планирования карьерного пути и повышения собственной конкурентоспособности.

Задание 1.3

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

Ответ

Повышение квалификации способствует развитию профессиональных навыков, что ведет к увеличению производительности и, как следствие, росту годового дохода. Данный причинно-следственный механизм поддерживается следующими источниками:

1. Влияет ли дополнительное профессиональное обучение на зарплаты российских работников?[1]

2. The Economic Effect of Gaining a New Qualification Later in Life[2]

Задание 1.4

Кратко опишите результаты предшествовавших исследований по схожей тематике и критически оцените методологию этих работ с точки зрения гибкости (жесткости предпосылок) использовавшихся методов эконометрического анализа. Объясните, в чем заключается преимущество и недостатки применяемых вами методов в сравнении с теми, что ранее использовались в литературе.

Ответ

1. Влияет ли дополнительное профессиональное обучение на зарплаты российских работников?

    а) Применение нескольких эконометрических методов (МНК, квантильная регрессия, метод «разности разностей») для проверки устойчивости результатов.

b) Контекстуальная релевантность исследования для российского рынка труда.

c) Ограниченная внешняя валидность, так как фокус на российском рынке затрудняет обобщение результатов для других стран.

d) Смешанные результаты: применение метода «разности разностей» не всегда выявляет статистически значимый эффект.

2. The Economic Effect of Gaining a New Qualification Later in Life

   a) Применение современных методов машинного обучения для решения проблемы.

   b) Обширные на национальном уровне(HILDA) данные позволяет обеспечить высокую степень репрезентативности

   c) Так же ограниченная внешняя валидность, поскольку фокус только на австралийском рынке.

   d) Сложность интерпретации некоторых методов машинного обучения.

   Особенности нашего метода:

   a) Отсутствие фокуса на определенном рынке.

   b) Системный подход к обработке, валидации данных, выбору метрик и моделей.

   c) Сложность интерпретации некоторых моделей машинного обучения.

Задание 1.5

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.

Ответ

В исследовании предлагается использовать следующие контрольные переменные:

1. **Опыт работы (X_1):** непрерывная переменная, измеряемая в годах. Опыт работы влияет на профессиональные навыки и, соответственно, на уровень дохода.

2. **Оценка когнитивных способностей (X_2):** непрерывная переменная, измеряемая по шкале от 50 до 100. Данная переменная отражает интеллектуальный потенциал, способствующий эффективному усвоению новых знаний.

3. **Наличие поддержки (X_3):** бинарная переменная, где 1 означает, что руководство поощряет или спонсирует участие в программах повышения квалификации, а 0 — нет. Это может влиять как на желание участвовать в программах повышения квалификации, так и на доход.

Задание 1.6

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

Ответ

В качестве инструментальной переменной предлагается использовать **Доступность корпоративных обучающих программ (Z)**. Эта переменная принимает значение:

- 1, если в регионе или компании доступны специализированные обучающие программы;
- 0, если такие программы отсутствуют.

Данная переменная удовлетворяет необходимым условиям:

- **Релевантность:** Наличие обучающих программ напрямую повышает вероятность участия в программах повышения квалификации.
- **Экзогенность:** Доступность программ не оказывает прямого влияния на годовой доход, если эффект проходит через участие в обучении.

Задание 1.7

В случае необходимости приведите дополнительные содержательные комментарии о целях, задачах, методологии и вкладе вашего исследования.

Ответ не дается

Генерация и предварительная обработка данных

Задача 2.1

Опишите математически предполагаемый вами процесс генерации данных.

Ответ

В нашем исследовании мы моделируем данные следующим образом. Пусть имеются следующие переменные:

- **Контрольные переменные:**
  - X_1 (Опыт работы): генерируется из нормального распределения с средним 10 и стандартным отклонением 3, при этом отрицательные значения отсекаются, чтобы гарантировать положительность (например, $X_1 \sim \max\{N(10, 3^2), \epsilon\}$, где $\epsilon > 0$).
  - X_2 (Оценка когнитивных способностей): равномерно распределена на отрезке от 50 до 100, т.е. $X_2 \sim U(50, 100)$.
  - X_3 (Поддержка руководства): бинарная переменная, $X_3 \sim \text{Bernoulli}(0.5)$ (1 — есть, 0 — нет).
- **Инструментальная переменная:**
  - Z (Доступность корпоративных обучающих программ): бинарная переменная, $Z \sim \text{Bernoulli}(0.5)$.

Далее, вероятность участия в программах повышения квалификации (переменная воздействия D) определяется с помощью логистической модели, включающей контрольные переменные и инструмент Z:

$$p(D = 1) = \sigma\Big(-0.5 + 0.05X_1 + 0.03X_2 + 0.8X_3 + 0.7Z\Big),$$

где $\sigma(x) = \frac{1}{1+e^{-x}}$ — сигмоида. Затем переменная D генерируется как:

$$D \sim \text{Bernoulli}\Big(p(D = 1)\Big).$$

Наконец, зависимая переменная (годовой доход, Y) генерируется по следующей нелинейной модели:

$$Y = 20 + 15\,D + 3\sqrt{X_1} + 0.2X_2 + 0.5X_3 + 2\,(D \cdot \ln(X_2)) + 4\sin(X_1) + \varepsilon,$$

где $\varepsilon \sim N(0, 9)$ — случайная ошибка.

Задача 2.2

Обоснуйте предполагаемые направления связей зависимой переменной и переменной воздействия с контрольными переменными.

Ответ

В нашей модели предполагается, что контрольные переменные влияют как на вероятность участия в программах повышения квалификации (переменная D), так и на годовой доход (Y). Обоснование направлений связей представлено ниже:

- X_1 **(Опыт работы):** Более высокий уровень опыта, как правило, повышает шансы на участие в программах повышения квалификации, поскольку опыт может быть связан с желанием профессионального роста. При этом X_1 с помощью $\sqrt{X_1}$ и $\sin(X_1)$ учтены возможные нелийности эффекта.

- X_2 **(Оценка когнитивных способностей):** Более высокие когнитивные способности способствуют лучшему усвоению знаний и, следовательно, увеличивают вероятность участия в обучении. X_2 положительно влияет на Y напрямую, а также усиливает эффект участия в обучении (через взаимодействие $D \cdot \ln(X_2)$).

- X_3 **(Поддержка руководства):** Может играть важную роль в решении сотрудника принять участие в обучении, поскольку, когда начальство поощряет профессиональное развитие, сотрудник с большей вероятностью воспользуется предоставляемыми возможностями. Кроме того, это может иметь психологический эффект, который может привести к увеличению дохода.

- Z **(Инструментальная переменная):** Z влияет только на вероятность участия (D) и не включается напрямую в модель Y.

Задача 2.3

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

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

- Для бинарных переменных: доля и количество единиц.

Указания:

- Необходимо сгенерировать не менее 1000 наблюдений.
- Доля единиц не должна быть меньше 0.1 или больше 0.9 ни для одной из бинарных переменных.

Ответ

Сгенерированы 1000 данных.
Корреляционная матрица(1), описательная статистика для непрерывных (1) и бинарных (2) переменных представлены ниже

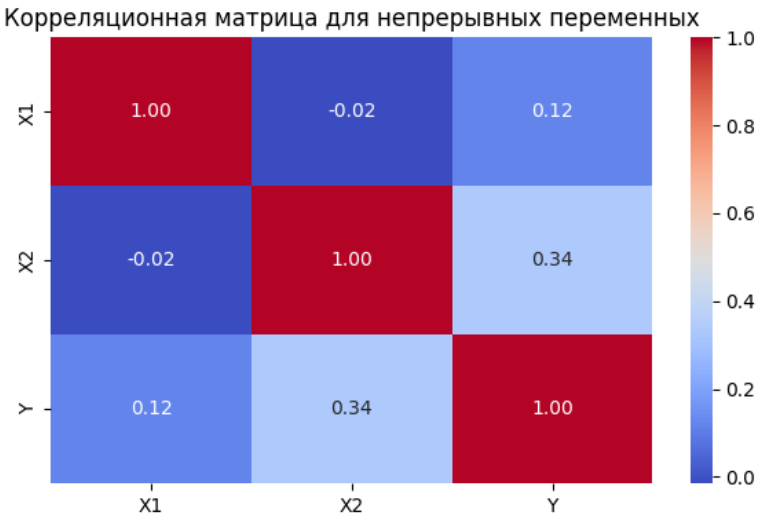

Рис. 1: Корреляционная матрица для переменных X1, X2 и Y.

Переменная	Выборочное среднее	Выборочное стандартное отклонение	Медиана	Минимум	Максимум
X1	10.01	2.97	10.01	0.92	18.73
X2	75.94	14.39	76.45	50.07	99.95
Y	56.72	13.67	51.88	28.43	87.31

Таблица 1: Описательные статистики для непрерывных переменных

Переменная	X3	Z	D
Доля единиц	0.51	0.50	0.41
Количество единиц	511.00	501.00	406.00

Таблица 2: Описательные статистики для бинарных переменных

Задача 2.4

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

Ответ

Размер обучающей выборки: (750, 6)
Размер тестовой выборки: (250, 6)

Задача 2.5

В случае необходимости проведите дополнительный анализ и приведине дополнительные комментарии о процессе генерации данных, описательных статистиках и т.д.

Ответ не дается

Классификация

Задача 3.1

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

Ответ

В типичном инструментальном анализе прогнозирование D — это первый этап. D прогнозируется на основании всех доступных предикторов, то есть X_1, X_2, X_3, Z. Это позволяет выделить ту часть вариации D, которая обусловлена экзогенными факторами.

Задача 3.2

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).

Проинтерпретируйте полученные результаты.

Ответ

Модели

1. Логистическая регрессия с параметрами: C=1, solver='lbfgs', max_iter=1000
2. KNN классификатор: n_neighbors=5
3. XGBClassifier: n_estimators=30, learning_rate=0.1, max_depth=3, eval_metric='logloss'

Результаты

Method	Train Accuracy	Test Accuracy	CV Accuracy
Logistic Regression	0.65	0.60	0.64
k-NN	0.70	0.53	0.55
XGBoost	0.69	0.59	0.61

Таблица 3: Классификация с базовыми параметрами

Выводы по таблице 3

- **Logistic Regression** демонстрирует достаточно ровные результаты: 65% на обучении, 60% на тесте и 64% в кросс-валидации. Это говорит о сравнительно хорошем обобщении и небольшой разнице между обучающей и тестовой точностью.
- **k-NN** показывает наивысшую точность (70%) на обучающей выборке, но при этом самый низкий результат (53%) на тесте и 55% в кросс-валидации. Это указывает на явное переобучение (overfitting), когда модель сильно адаптируется к обучающим данным, но хуже обобщает на новых данных.
- **XGBoost** даёт промежуточные результаты: 69% на обучении и 59% на тесте, при этом кросс-валидация (61%) также подтверждает, что модель недалека по качеству от логистической регрессии, но немного отстаёт по тестовой точности.

Задача 3.3

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.

Ответ

Результаты представлены в таблице 4

Method	Init Hyperparams	Tuned Hyperparams	CV Acc Init	CV Acc Tuned	Test Acc Init	Test Acc Tuned
Logistic Regression	C=1	C=1	0.64	0.64	0.60	0.60
k-NN	k=5	k=9	0.55	0.60	0.53	0.58
XGBoost	n_estimators=30, lr=0.1, max_depth=3	learning_rate=0.05, max_depth=3, n_estimators=30	0.61	0.63	0.59	0.56

Таблица 4: Оптимальные значения гиперпараметров (ACC)

Задача 3.4

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

Ответ

Предлагается F1-мера.
Преимущества:

1. Учитывает сразу и полноту и точность.
2. Позволяет учитывать как FP, так и FN ошибки.

Недостатки:

1. Не учитывает стоимость ошибок разного типа.

Результаты представлены в таблице 5

Method	Init Hyperparams	Tuned Hyperparams	CV F1 Init	CV F1 Tuned	Test F1 Init	Test F1 Tuned
Logistic Regression	C=1	C=2	0.47	0.47	0.40	0.40
k-NN	k=5	k=9	0.44	0.45	0.48	0.43
XGBoost	n_estimators=30, lr=0.1, max_depth=3	learning_rate=0.2, max_depth=5, n_estimators=200	0.44	0.51	0.40	0.42

Таблица 5: Оптимальные значения гиперпараметров (F1)

Задача 3.5

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

Ответ

Результат представлен на графике 2

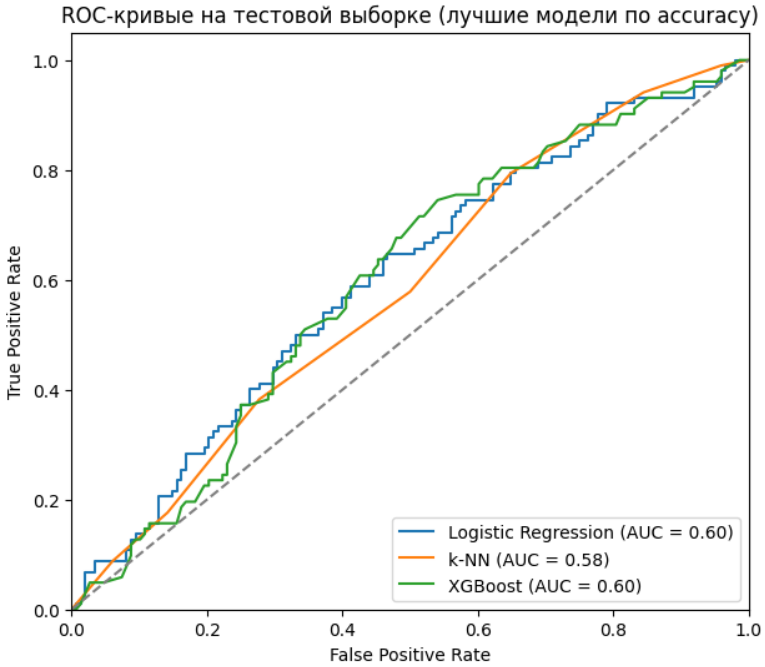

Рис. 2: ROC-AUC для моделей

Логистическая регрессия и **XGBoost** показывают одинаковый результат по метрике Accuracy: 0.6.
k-NN немного отстает по качеству, его метрика равна 0.58.

Задача 3.6

Постройте матрицу ошибок 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.

Ответ

Предполагаем такие цены прогнозов: $TP = 100,\ FP = -50,\ FN = -30,\ TN = 0$.
Равномерно распределим пороги в количестве 101 на отрезке $[0, 1]$.
Результат

Model	AUC Test	Best Threshold	Train Profit	Test Profit
Logistic Regression	0.60	0.31	10530	2480
k-NN	0.58	0.12	10800	3170
XGBoost	0.60	0.34	12390	2790

Таблица 6: Прибыли и AUC моделей

Вывод по таблице 6

1. AUC: Наилучшие модели — XGBoost и Logistic Regression
2. Best Threshold: Низкий порог у k-NN (0.12) говорит о том, что для максимизации прибыли модель предпочитает «агрессивнее» относить объекты к положительному классу.
3. Profit: У всех моделей наблюдается снижение Profit при переходе от тренировочной выборке к контрольной. Это типичный признак переобучения под малое количество данных. k-NN показывает небольшой отрыв по метрике.

Задача 3.7

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.

Ответ не дан

Задача 3.8

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

Ответ

В качестве основного критерия мы берём итоговую прибыль на тестовой выборке. Тогда:
Лучший классификатор — это k-NN, поскольку у него самая высокая тестовая прибыль (3170).
Худший классификатор — это Logistic Regression, так как его тестовая прибыль (2480) ниже, чем у двух остальных моделей.

Регрессия

Задача 4.1

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

Ответ

Будем выбирать признаки исходя из их построения (Задание 1.5). Выбранные признаки: $X_1,\ X_2,\ X_3$. Z не включается в перечень, поскольку не имеет прямого влияния на Y по построению.

Задача 4.2

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

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

Проинтерпретируйте полученные результаты.

Ответ

Модели

1. k-NN: n_neighbors=5
2. XGBoost: n_estimators=30, learning_rate=0.1, max_depth=3
3. RandomForest: n_estimators=50, max_depth=5

Результат

Model	RMSE (train)	RMSE (test)	CV RMSE	MAPE (train)	MAPE (test)	CV MAPE
k-NN	10.23	13.90	12.40	0.15	0.22	0.19
XGBoost	10.86	12.59	11.89	0.18	0.21	0.20
RandomForest	10.32	12.97	11.89	0.17	0.21	0.19

Таблица 7: Регрессия с базовыми параметрами

Вывод по таблице 7

1. **k-NN**: заметное переобучение, худший RMSE на тесте (13.90) и самая высокая MAPE (0.22).

2. **XGBoost**: лучший RMSE на тесте (12.59), MAPE (0.21) такая же, как у RandomForest. Модель довольно стабильно ведёт себя на train, test и CV.

3. **RandomForest**: тестовый RMSE (12.97) немного хуже, чем у XGBoost, MAPE (0.21) совпадает с XGBoost. Модель достаточно стабильна.

Задача 4.3

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.

Проинтерпретируйте полученные результаты.

Ответ

Результат

Model	Init Hyperparams	Tuned Hyperparams	CV RMSE Init	CV RMSE Tuned	Test RMSE Init	Test RMSE Tuned
k-NN	n_neighbors=5	n_neighbors=3, weights='uniform'	12.40	13.58	13.90	14.10
XGBoost	n_estimators=30, learning_rate=0.1, max_depth=3	n_estimators=30, learning_rate=0.05, max_depth=3	11.89	11.95	12.59	12.59
RandomForest	n_estimators=50, max_depth=5	n_estimators=30, learning_rate=0.05, max_depth=3	11.89	11.86	12.97	12.80

Таблица 8: Регрессия с подобранными параметрами

Вывод по таблице 8

1. **k-NN**: Подбор параметров (добавление взвешивания) ухудшил качество.

2. **XGBoost**: Подбор параметров ничего не изменил.

3. **RandomForest**: Подбор параметров дал небольшое улучшение как на CV, так и на тестовой выборке.

Задача 4.4

На основании проделанного анализа выберите лучший и худший из обученных классификаторов (регрессоров?). Обоснуйте сделанный выбор.

Ответ

Лучший регрессор — XGBoost (наименьшая ошибка на тесте: 12.59).
Худший регрессор — k-NN (наибольшая ошибка на тесте: 14.10).

Задача 4.5

Повышенная сложность: включите в анализ дополнительный метод регрессии, не рассматривавшийся в курсе и не представленный в библиотеке scikitlearn. 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.

Ответ не дан

Задача 4.6

В случае необходимости проведите дополнительный анализ.

Ответ не дан

Эффекты воздействия

Задача 5.1

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

Ответ

Пусть для каждого наблюдения i $(i = 1, 2, ..., n)$ целевая переменная (исход) может принимать одно из двух потенциальных значений:

$$Y_i(1) = \text{результат (исход) для } i\text{-ого объекта, если } D_i = 1$$
$$Y_i(1) = \text{результат (исход) для } i\text{-ого объекта, если } D_i = 0$$

Это можно интерпретировать, как два исходных потенциальных исхода для каждого объекта. Однако фактически мы **не можем** наблюдать оба этих значения одновременно. Наблюдаемое (фактическое) значение Y_i определяется бинарной переменной воздействия D_i

$$Y_i = D_i \cdot Y_i(1) + (1 - D_i) \cdot Y_i(0)$$

Задача 5.2

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. Для ATE и LATE результаты представьте в форме таблицы, а для CATE постройте гистограмму или ядерную оценку функции плотности. Проинтерпретируйте полученные значения.

Ответ

Поскольку у нас имеется явная формула генерации Y, то мы можем сгенерировать $Y_i(1)$ и $Y_i(0)$.

$$Y(0) = 20 + 3\sqrt{X_1} + 0.2X_2 + 0.5X_3 + 4\sin(X_1) + \varepsilon$$
$$Y(1) = Y(0) + 15 + 2\ln(X_2)$$

Результат

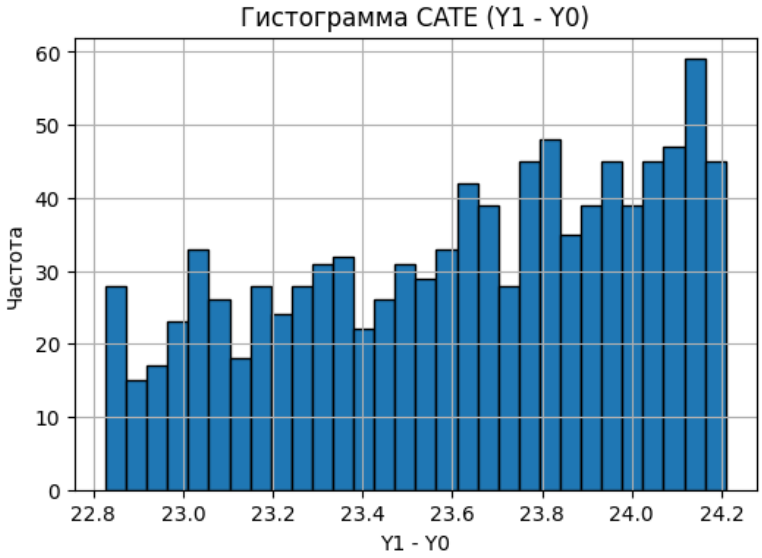

Рис. 3: Гистограмма CATE

Effect	Value
ATE	23.62
LATE	23.61

Таблица 9: Эффекты воздействия

Вывод по графику 3 и таблице 9

1. *Распределение CATE*

   a) Значения индивидуальных эффектов τ_i варьируются от 22.8 до 24.2.

   b) Диаграмма поднимается вверх, значит, большая часть эффектов лежит в верхней границе диапазона.

   c) Разброс довольно узкий (1.4), то есть все Y получают примерно одинаковый эффект от D.

2. *ATE*

   a) Средний эффект воздействия достаточно велик — 23.62 единицы.

   b) ATE близко к середине диапазона гистограммы CATE.

3. *LATE*

a) Локальный средний эффект воздействия среди собладателей практически совпадает со средним эффектом.

b) Нет статистически значимого различия в эффекте между подвыборкой комплаенеров и общей выборкой.

Задача 5.3

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.

Ответ

Оцененный средний эффект = 25.39.
Недостатки с учетом специфики рассматриваемой задачи.
Такой подсчет — это завышенная оценка реального эффекта воздействия, поскольку не учитываются никакие эффекты кроме самого факта участия D. Иными словами, иметь $D = 1$ могли иметь изначально те, кто "лучше"по другим критериям, однако модель этого не учитывает.

Задача 5.4

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.

Ответ

Результат представлен в таблице 10

Method	ATE
OLS	23.62
CondMeans	24.29
IPW	23.70
AIPW (Doubly Robust)	23.68
Double ML	23.69

Таблица 10: ATE

Ключевая предпосылка методов: допущение об условной независимости, согласно которому

$$\mathbb{E}(Y_{1i}|X_i, D_i = 1) = \mathbb{E}(Y_{1i}|X_i) \quad \mathbb{E}(Y_{0i}|X_i, D_i = 0) = \mathbb{E}(Y_{0i}|X_i)$$

Допущение об условной независимости соблюдается, если X_i отражают все факторы, которые могут статистически быть связаны с D_i и Y_{ji}.
Интерпретация оценок с помощью:

1. **Метода наименьших квадратов.**

   В предположении, что зависимость Y от D (при контроле за X) *линейна*, мы оцениваем модель:

   $$Y \;=\; \alpha + \tau\, D + f(X) + \varepsilon.$$

   Коэффициент при D (τ) трактуется как средний эффект воздействия (ATE).

2. **Условных математических ожиданий.**

   Формально,

   $$ATE \;=\; \frac{1}{n}\sum_{i=1}^{n}\Big(\mathbb{E}[Y_i \mid X_i, D_i = 1] \;-\; \mathbb{E}[Y_i \mid X_i, D_i = 0]\Big).$$

   Если мы можем *состоятельно* оценить $\mathbb{E}[Y \mid X, D]$ методами машинного обучения, то среднее разности прогнозов $\hat{m}_1(X_i) - \hat{m}_0(X_i)$ даёт оценку ATE.

3. **Взвешивание на обратные вероятности.**

   Пусть $p(X_i) = P(D = 1 \mid X_i)$. Тогда

$$ATE \;=\; \frac{1}{n}\sum_{i=1}^{n}\Big[\frac{D_i\,Y_i}{p(X_i)} \;-\; \frac{(1-D_i)\,Y_i}{1-p(X_i)}\Big].$$

   Здесь модель классификации $\hat{p}(X)$ даёт вероятности участия $D = 1$.

4. **Метод с двойной устойчивостью.**

   Данный подход сочетает идею IPW и модель $\mathbb{E}[Y \mid X, D]$. Формула:

$$\widehat{\tau}_i = \frac{D_i\big(Y_i - \hat{m}_1(X_i)\big)}{\hat{p}(X_i)} \;-\; \frac{(1-D_i)\big(Y_i - \hat{m}_0(X_i)\big)}{1-\hat{p}(X_i)} \;+\; \big[\hat{m}_1(X_i) - \hat{m}_0(X_i)\big].$$

   При усреднении по i получаем ATE.

5. **Двойное машинное обучение.**

   Для снижения переобучения мы используем *cross-fitting*: делим выборку на фолды (A,B), обучаем модели $(\hat{p}, \hat{m}_0, \hat{m}_1)$ на A и применяем к B, и наоборот. Это уменьшает смещение, возникающее при использовании мощных алгоритмов (RandomForest, XGBoost). Результат подставляем в формулу выше.

Задача 5.5

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.

Ответ

Результат представлен в таблице 11

Существование Z	LATE
Нет	24.68
Да	12.56

Таблица 11: LATE

- **Без инструмента** (локальный эффект): 24.68.

  Здесь мы оценивали средний эффект воздействия D на Y для некоторой подгруппы (или по условным средним), не используя инструмент. В результате получаем *локальную* оценку в размере 24.5, однако она может быть завышенной (или заниженной), поскольку не отражает эффект на соблюдателях.

- **С инструментом** (LATE): 12.56.

  Когда вводим бинарную инструментальную переменную Z, оцениваем *локальный средний эффект* (LATE) среди "соблюдателей" — тех, у кого D действительно меняется при изменении Z.

Интерпретация
Пусть D — участие в программе, Y — доход. Без инструмента мы видим эффект ≈ 24.68, но это лишь оценка на подгруппе. При использовании Z, оценка "чистого" эффекта падает до 12.56, поскольку учитываем только тех, чьё решение об участии действительно зависит от Z. Это и есть *локальный средний эффект* — более достоверная причинная прибавка дохода именно у "реагирующих" на инструмент.

Задача 5.6

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.

Ответ

Результат представлен в таблице 12

Method	AvgCATE
OLS	23.62
S-learner	18.89
T-learner	24.29
Class transformation	23.66
X-learner	24.15

Таблица 12: CATE

X-learner обычно используется, если группа воздействия мала, поскольку в таком случае оценка $\mathbb{E}(Y_i|X_i, D_i = 1)$ осложнена. В нашем случае применение метода X-learner не мотивировано, поскольку группа с $D = 1$ по мощности практически совпадает с группой с $D = 0$.

Применение в бизнесе

Позволяет таргетировать воздействие: применять маркетинговое действие к тем x, где $CATE$ достаточно высока.

Применение в государственных программах

Позволяет определить граждан, которые получат наибольшую пользу от субсидии, что позволит повысить эффективность распределения бюджета.

Задача 5.7

Выберите лучшую модель оценивания условных средних эффектов воздействия, используя:

- истинные значения условных средних эффектов воздействия.
- прогнозную точность моделей.
- псевдоисходы. (не реализован)

Проинтерпретируйте различия в результатах различных подходов

Ответ

Результат

Method	MSE (true CATE)
OLS	0.37
S-learner	23.26
T-learner	1.65
Class transformation	102.23
X-learner	1.05

Таблица 13: Сравнение с true CATE

Method	MSE (true Y)
OLS	15.02
S-learner	22.74
T-learner	12.56
ClassTransformation	3580.44
X-learner	12.56

Таблица 14: Сравнение с true Y

CATE (таблица 13)

1. OLS дает наименьшую ошибку среди всех моделей.
2. T-learner и X-learner немного отстают по качеству.
3. Модели S-learner и Class Transformation можно считать невалидными, их прогнозы сильно смещены.

Y (таблица 14)

1. OLS снова дает отличную оценку.
2. X-learner и T-learner дают наилучший результат среди всех моделей.
3. Модели S-learner и Class Transformation вновь невалидны, их прогнозы сильно смещены.

Задача 5.8

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.

Ответ

Результаты

Method	ATE_best	ATE_worst
CondMeans	24.29	24.46
IPW	23.70	22.31
AIPW (Doubly Robust)	23.68	23.88
Double ML	23.69	23.77

Таблица 15: ATE (comparison)

Существование Z	LATE_best	LATE_worst
Нет	24.26	24.66
Да	12.56	23.53

Таблица 16: LATE (comparison)

Вывод

- **ATE (Таблица 15).**
  - Методы `CondMeans`, `AIPW` и `Double ML` показывают весьма близкие оценки между "best" и "worst" моделями, с разницей около 0.1–0.2. Это говорит о том, что их результаты *устойчивы* к вариациям моделей.
  - `IPW` (23.70 $\rightarrow$ 22.31) более чувствителен к ошибкам: при небольшом изменении оценки $p(X)$ итоговая оценка ATE может заметно измениться (примерно на 1.4).

  Таким образом, для ATE методы `CondMeans`, `AIPW` и `Double ML` выглядят *более стабильными*, тогда как `IPW` оказывается *чувствительным* к вариациям модели.

- **LATE (Таблица 16).**

– При *отсутствии* инструментальной переменной Z, оценки LATE меняются лишь с 24.26 до 24.66, то есть разница не превышает 0.4.

– При *наличии* Z колебания гораздо существеннее: 12.56 $\rightarrow$ 23.53, что означает разрыв порядка 11.

Следовательно, без Z оценка LATE оказывается *стабильной*, но при наличии инструмента колебания возрастают.

Задача 5.9

Резюмируйте ключевые выводы проведенного в данном разделе анализа.

Ответ

1. Эффект воздействия достаточно велик (23.62), при этом распределение CATE достаточно узкое (размах 1.4) и большая часть эффектов лежит в верхней границе диапазона (22.8 — 24.2) CATE.

2. Все модели оценки ATE показали примерно одинаковый результат (от 23.62 до 24.29), близкий к среднему эффекту воздействия.

3. Оценка локального условного эффекта воздействия, полученная с помощью метода двойного машинного обучения с инструментальной переменной, сильно смещена от LATE. Видимо из-за особенностей выбранных моделей ошибка на малом количестве данных оказывается слишком велика.

4. Среди моделей оценки условного среднего эффекта лучшими являются: OLS, T-learner и X-learner. Они показывают высокий результат как в CATE-тесте, так и в Y-тесте. Предполагаю, что оценка Class trasformation настолько смещена из-за особенностей предсказания вероятностей логистической регрессией: если классы были сильно разделимы, то посчитанные вероятности могли быть близки к 0 или 1, что могло сместить оценку.

5. Оценка средних эффектов воздействия с помощью худших моделей схожа с оценкой лучших моделей. Этого следовало ожидать, поскольку модели в предыдущих тестах давали достаточно близкие результаты. Наименее устойчивый алгоритм к выбору модели — IPW.

6. Оценка локального среднего эффекта воздействия с помощью худших моделей оказалась ближе к истинному LATE, чем оценка с помощью лучших моделей.

Список литературы

[1] Данил А Девятьяров and Людмила А Леонова. Влияет ли дополнительное профессиональное обучение на зарплаты российских работников? *Журнал экономической теории*, 20(3):621–640, 2023.

[2] Finn Lattimore, Daniel M Steinberg, and Anna Zhu. The economic effect of gaining a new qualification later in life. *arXiv preprint arXiv:2304.01490*, 2023.