РОССИЙСКИЙ УНИВЕРСИТЕТ ДРУЖБЫ НАРОДОВ

Факультет/институт: Искусственного интеллекта

Кафедра: Искусственного интеллекта

ОТЧЁТ О ЛАБОРАТОРНОЙ РАБОТЕ

по дисциплине «Прикладная статистика и анализ данных»

Лабораторная работа № 3

Тема: «Статистический вывод, бутстреп, важность признаков и SHAP»).

Студент(ка): Перевозчикова Валерия Дмитриевна, ЗФИмд-01-25, ст. б.: 1032259214

Преподаватель: Курашкин Сергей Олегович, Доцент

Дата выполнения:

«14» декабря 2025 г.

Оценка/подпись:

_____

Москва — 2025

СОДЕРЖАНИЕ

Оглавление

Введение 3

Теоретические основы 3

Описание данных и инструментов 5

Методика и план эксперимента 6

Результаты и их анализ 8

Выводы 11

Приложения 12

Введение

Цель работы: Построение полного цикла статистического моделирования с фокусом на оценке неопределенности и интерпретируемости моделей для анализа аэропортовых поездок такси Нью-Йорка.

Задачи работы:

1. Загрузить и предобработать данные NYC TLC за январь 2019 года

2. Построить и сравнить несколько моделей регрессии для прогнозирования стоимости за милю (`fare_per_mile`)

3. Оценить неопределенность моделей с помощью bootstrap методов

4. Проверить статистическую значимость с помощью permutation tests

5. Проанализировать важность признаков методами permutation importance и SHAP

6. Построить причинно-следственный граф (DAG) для анализа эффекта аэропортового статуса

Актуальность: В эпоху сложных моделей машинного обучения (градиентный бустинг, нейросети) критически важно не только достичь высокой точности, но и понимать, как модель принимает решения, оценивать надежность предсказаний и отделять корреляцию от причинности.

Используемый датасет: NYC TLC Trip Records (Yellow Taxi, январь 2019), подвыборка 100,000 поездок.

Модели: Линейная регрессия (базовая), Ridge регрессия, Random Forest, Gradient Boosting.

Ожидаемые результаты:

- Количественная оценка неопределенности метрик качества через доверительные интервалы

- Статистическое обоснование различий между моделями

- Интерпретация вклада признаков, особенно `is_airport_trip`

- Качественный анализ причинных связей

Теоретические основы

Статистический вывод и доверительные интервалы

В условиях, когда аналитическое распределение статистики неизвестно или предположения нарушены, используются методы ресемплинга. Пусть $\hat{\theta}$ - оценка параметра по выборке размера n. Доверительный интервал уровня $(1-\alpha)$ строится как $CI_{1-\alpha}(\theta) = [q_{\alpha/2}, q_{1-\alpha/2}]$, где q_p - p-квантиль бутстреп-распределения оценки.

Bootstrap

Непараметрический bootstrap генерирует B выборок с возвращением из исходных данных:

$$X_b^* = \{x_i^* : i = 1, \dots, n\}, \quad x_i^* \sim \text{Uniform}(x_1, \dots, x_n)$$

Для каждой выборки вычисляется статистика $\hat{\theta}_b^*$, а эмпирическое распределение $\{\hat{\theta}_b^*\}_{b=1}^B$ аппроксимирует выборочное распределение $\hat{\theta}$.

Percentile bootstrap: $CI = [\hat{\theta}_{(\alpha/2)}, \hat{\theta}_{(1-\alpha/2)}]$

Permutation Tests

Для проверки гипотезы H_0: "отсутствие эффекта" многократно переставляются метки или признаки, разрушая исходную зависимость. Тестовая статистика T вычисляется для каждой перестановки, формируя нулевое распределение. P-value:

$$p = \frac{1}{B} \sum_{b=1}^B I(|T_b^*| \ge |T_{\text{obs}}|)$$

Permutation Feature Importance

Важность признака j измеряется как ухудшение качества модели при разрушении связи признака с целевой переменной:

$$\text{Imp}(j) = Q(y, f(X)) - Q(y, f(X^{\pi(j)}))$$

где $X^{\pi(j)}$ - матрица признаков со случайно переставленными значениями x_j.

SHAP (SHapley Additive exPlanations)

На основе теории кооперативных игр, вклад признака j в предсказание для наблюдения x:

$$\phi_j(x) = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} [f_{S \cup \{j\}}(x) - f_S(x)]$$

где $f_S(x)$ - ожидаемое предсказание при известных только признаки из S.

Метрики качества

- $R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$ - доля объясненной дисперсии

- $RMSE = \sqrt{\frac{1}{n} \sum (y_i - \hat{y}_i)^2}$ - в единицах целевой переменной

- $MAE = \frac{1}{n} \sum |y_i - \hat{y}_i|$ - устойчивая к выбросам

Описание данных и инструментов

Источник данных

NYC Taxi and Limousine Commission (TLC) Trip Record Data, Yellow Taxi, январь 2019.

URL: https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

Размерность и характеристики

- Исходный размер: 7,696,617 записей

- После фильтрации: 100,000 записей (подвыборка)

- Количество признаков: 7 после feature engineering

- Целевая переменная: `fare_per_mile` (стоимость за милю, $/mile)

Словарь данных

| Поле | Описание | Тип/шкала | Допустимые значения |

| trip_distance | Расстояние поездки в милях | Вещественная | (0.1, 30] |

| pickup_hour | Час начала поездки | Целая | 0-23 |

| pickup_dayofweek | День недели | Целая | 0-6 (0=Понедельник) |

| is_weekend | Признак выходного | Бинарная | 0/1 |

| is_airport_trip | Поездка из/в аэропорт | Бинарная | 0/1 |

| pickup_borough | Район пикапа | Категориальная | 6 районов |

| passenger_count | Количество пассажиров | Целая | 1-6 |

| fare_per_mile | Целевая: стоимость за милю | Вещественная | [2.74, 14.41] $ |

Предобработка данных

1. Фильтрация: Удалены поездки с расстоянием ≤ 0.1 мили, стоимостью ≤ $2.5, некорректным количеством пассажиров

2. Обработка выбросов: Удалены 1% экстремальных значений `fare_per_mile` с каждой стороны

3. Feature engineering:

- Создание временных признаков из timestamp

- Определение аэропортовых поездок по LocationID (JFK=132, LGA=138)

- Сопоставление районов через taxi_zones.csv

Разбиение данных

- Train/Test split: 80%/20%

- Стратификация по `is_airport_trip` для сохранения пропорции

- Размеры: Train - 78,402; Test - 19,601

Программные инструменты

- python
- numpy==1.24.3
- pandas==2.0.3
- scikit-learn==1.3.0
- matplotlib==3.7.2
- seaborn==0.12.2
- shap==0.42.1
- scipy==1.11.1
- networkx==3.1

Настройки воспроизводимости

- RANDOM_STATE = 42
- np.random.seed(RANDOM_STATE)

Методика и план эксперимента

Конвейер препроцессинга

```
numeric_transformer = Pipeline([
    ('scaler', StandardScaler())
])
categorical_transformer = Pipeline([
    ('onehot', OneHotEncoder(handle_unknown='ignore', sparse_output=False))
])
```

```
preprocessor = ColumnTransformer([
    ('num', numeric_transformer, numeric_features),
    ('cat', categorical_transformer, categorical_features)
])
```

Модели

- Базовая: LinearRegression с ограниченными признаками
- Ridge: Ridge регрессия с регуляризацией $\alpha=1.0$
- RandomForest: 100 деревьев, max_depth=10
- GradientBoosting: 100 деревьев, max_depth=5, learning_rate=0.1

Схема bootstrap

- Тип: Non-parametric bootstrap по объектам
- Количество репликаций: $B = 200$
- Оцениваемые статистики: R^2, RMSE, MAE, разность ошибок моделей
- Интервалы: Percentile 95% CI
- Процедура:
- Для $b = 1,...,B$:
  a. Сгенерировать выборку с возвращением размера n
  b. Обучить модель
  c. Вычислить метрики на out-of-bag данных
- Построить эмпирическое распределение метрик
- Вычислить квантили для CI

Перестановочные тесты

- Нулевая гипотеза H_0: Модель не лучше случайного угадывания
- Альтернатива H_1: Модель значимо лучше
- Тестовая статистика: R^2
- Количество перестановок: 100
- Критерий: p-value < 0.05

Permutation Feature Importance

- Модели: Все обученные модели
- Метрика: R^2
- Количество повторений: 10
- Оценка: Среднее уменьшение метрики ± стандартное отклонение

SHAP анализ

- Подвыборка: 1000 наблюдений из тестовой выборки
- Explainer: TreeExplainer для GradientBoosting

Графики:

- Summary plot (глобальная важность)
- Dependence plot для is_airport_trip
- Waterfall plot для экстремальных случаев

Блок-схема эксперимента

[Загрузка данных] → [Предобработка] → [Разбиение] → [Обучение моделей]

↓ ↓

[Статистика] [Кросс-валидация]

↓ ↓

[Визуализация] ← [Интерпретация] ← [Bootstrap] ← [Permutation tests]

↓ ↓

[DAG] [Выводы]

Результаты и их анализ

5.1 Качество моделей и доверительные интервалы

Таблица 1: Сравнение моделей по кросс-валидации

Модель	Средний R^2	Std R^2	Ранг
Gradient Boosting	0.7655	0.0044	1
Random Forest	0.7628	0.0043	2
Airport-only GB	0.4238	0.1044	3
Ridge	0.3381	0.0079	4
Linear (базовая)	0.3285	0.0082	5

Таблица 2: Bootstrap доверительные интервалы для лучшей модели

Метрика	Среднее	95% CI	Test значение
R^2 (OOB)	0.7650	[0.7594, 0.7706]	0.7568
RMSE ($)	1.00	[0.98, 1.01]	1.01

Интерпретация: Test R^2 (0.7568) не попадает в bootstrap CI [0.7594, 0.7706], что указывает на небольшое переобучение модели. Однако разница минимальна (0.0082), что приемлемо для практического применения.

5.2 Распределения bootstrap оценок

Гистограмма показывает симметричное распределение R^2 с узким 95% доверительным интервалом. Стандартное отклонение 0.0030 указывает на высокую стабильность оценки качества модели.

5.3 Permutation tests

Результаты permutation test:

- Наблюдаемый R^2: 0.8120
- Максимальный R^2 на перемешанных данных: 0.2322
- p-value: 0.0000

Интерпретация: Модель статистически значима ($p < 0.05$). Все значения R^2 на перемешанных данных существенно ниже наблюдаемого, что подтверждает наличие реальных закономерностей в данных, а не случайных совпадений.

5.4 Важность признаков

Таблица 3: Permutation importance (Gradient Boosting)

Признак	Важность	Std	Ранг
trip_distance	1.3427	0.0134	1
pickup_hour	0.0769	0.0016	2

Признак	Важность	Std	Ранг
is_weekend	0.0168	0.0007	3
pickup_dayofweek	0.0105	0.0005	4
is_airport_trip	0.0039	0.0002	5
pickup_borough	0.0030	0.0003	6
passenger_count	-0.0003	0.0002	7

Интерпретация:

- trip_distance - самый важный признак (удаление снижает R^2 на 1.34)
- is_airport_trip занимает 5-е место с умеренной важностью 0.0039
- Признак passenger_count имеет отрицательную важность, возможно из-за переобучения

5.5 SHAP анализ

- Статистика SHAP для is_airport_trip:
- Средний |SHAP|: 0.0088
- Средний SHAP: 0.0005 (слабо положительный)
- Увеличивает стоимость: 88.1% случаев
- Уменьшает стоимость: 11.9% случаев

Взаимодействия is_airport_trip:

pickup_borough_Queens: отрицательная корреляция (-0.68) - в Queens эффект аэропорта ослабевает

trip_distance: отрицательная корреляция (-0.37) - на больших расстояниях эффект аэропорта меньше

5.6 Разрешение парадокса аэропортовых поездок

Исходный парадокс:

- Сырые данные: Аэропорт дешевле на 47.3% ($3.16 vs $6.00)
- Линейная модель: Коэффициент +1.17 (дороже)

Объяснение через SHAP:

- Гетерогенный эффект: в 88% случаев увеличивает, в 12% - уменьшает
- Конфаундинг расстоянием: аэропортовые поездки короче → ниже fare_per_mile
- При контроле расстояния проявляется реальный эффект

Выводы

Качество моделей: Gradient Boosting показывает наилучшее качество ($R^2 = 0.7568$) со стабильными оценками (95% CI: [0.7594, 0.7706]).

Статистическая значимость: Модель значимо лучше случайного угадывания ($p = 0.0000$) и значимо лучше линейных моделей (bootstrap сравнение).

Интерпретируемость:

- Самый важный признак - trip_distance (permutation importance = 1.3427)
- is_airport_trip имеет умеренную важность (5-е место, 0.0039)
- Эффект аэропорта гетерогенный: в основном увеличивает стоимость (88%), но в некоторых контекстах уменьшает
- Причинный анализ: Обнаружен сильный конфаундинг расстоянием. При контроле конфаундеров эффект аэропорта меняет знак с отрицательного на положительный.

Ограничения эксперимента

- Наблюдательные данные: Невозможно установить направление причинности
- Отсутствующие данные: Нет информации о платных дорогах, погодных условиях, точных тарифах
- Ограниченный период: Анализ только за январь 2019
- Географическая специфика: Результаты специфичны для Нью-Йорка

Направления дальнейшего развития

- Сбор дополнительных данных: Платные дороги, погода, события
- Причинные методы: Инструментальные переменные, Difference-in-Differences
- Пространственно-временной анализ: Учет географических и временных закономерностей
- Другие модели: Нейросетевые архитектуры, ансамбли
- Анализ справедливости: Проверка на дискриминацию в алгоритмах ценообразования

Практические рекомендации

- Для компаний такси: Учитывать аэропортовый статус при ценообразовании, но дифференцировать по районам и расстояниям

- Для регуляторов: Мониторить справедливость тарифов, учитывать конфаундинг при анализе
- Для исследователей: Использовать SHAP для интерпретации сложных моделей, применять bootstrap для оценки неопределенности

Заключение: Работа демонстрирует важность комплексного подхода к анализу данных, сочетающего предсказательное моделирование, оценку неопределенности и интерпретацию результатов. Полученные выводы имеют практическую значимость для транспортной отрасли и методологическую ценность для исследователей данных.

Приложения

Листинг кода:
https://colab.research.google.com/drive/12yY3i0tiY2Bxd_45Qp9ijeX8S3NS6S0e?usp=sharing