|  |
| --- |
| МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ |
| ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ  высшего образования |
| **«Национальный исследовательский ядерный университет «МИФИ»**  **(НИЯУ МИФИ)** |

**Отчет**

**по результатам выполнения задания**

**демонстрационного экзамена**

|  |  |
| --- | --- |
| Студент | Майоров Василий Юрьевич |
| Группа | С20-101 |
|  |  |
| Дата | 30.08.2023-30.11.2023 |

Москва 2023

**Оглавление**

[**I Исходные данные** 3](#_Toc143075002)

[**II Предобработка данных** 4](#_Toc143075003)

[**III Построение и исследование модели машинного обучения** 5](#_Toc143075004)

[**Заключение** 6](#_Toc143075005)

# **I Исходные данные**

Задача: Построение модели, кластеризующую полученный датасет.

Датасет представляет из себя таблицу, где 10000 людей распределны по 4 критериям: Возраст, пол, профессия и доход

Входные переменные: 'age', 'sex', 'profession'

Выходные переменные: 'income'

Количественные признаки : 'age', 'income'

Категориальные признаки: 'profession', 'sex'

Тип задачи: Кластеризация

Объем выборки: 10000

Статистика для age:

count 9521.000000

mean 34.965760

std 4.998853

min 18.000000

25% 32.000000

50% 35.000000

75% 38.000000

max 55.000000

Статистика для income:

count 10000.000000

mean 77494.399100

std 28965.823446

min 46042.000000

25% 51613.000000

50% 66724.000000

75% 111429.000000

max 200000.000000

Размах:

age 37.0

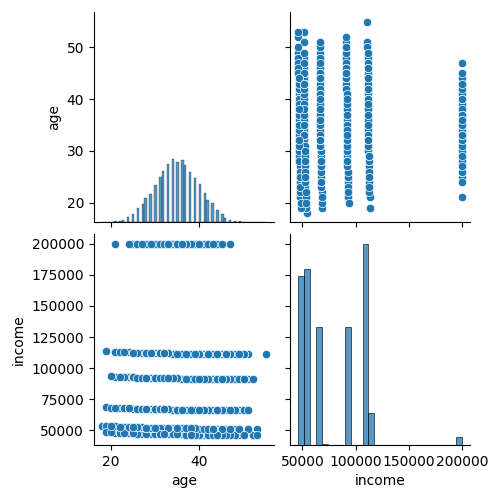
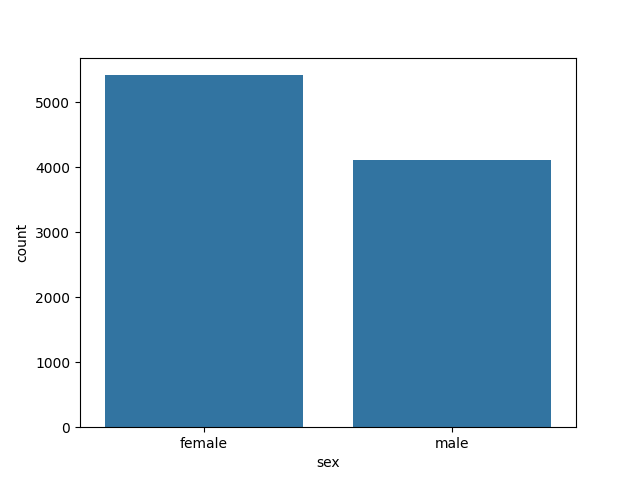
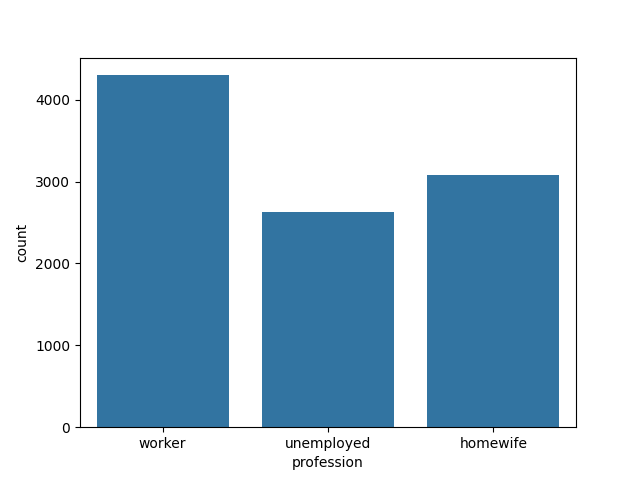
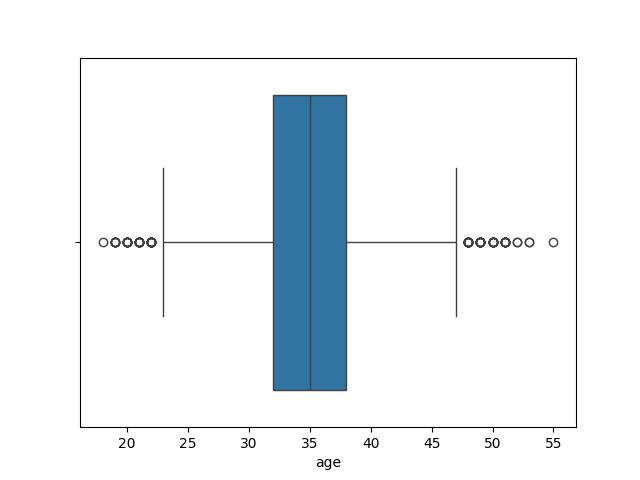
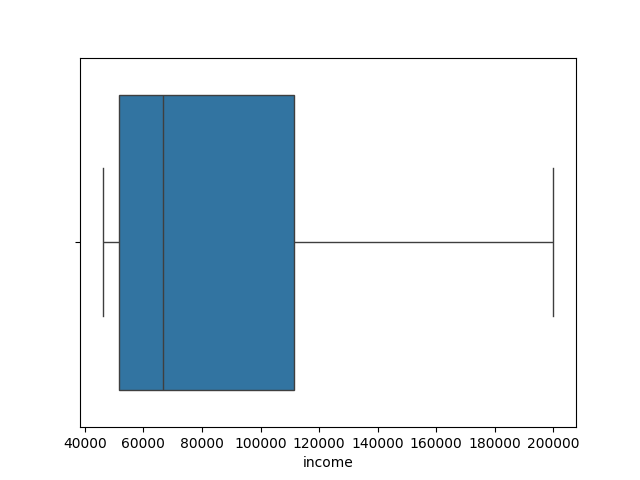
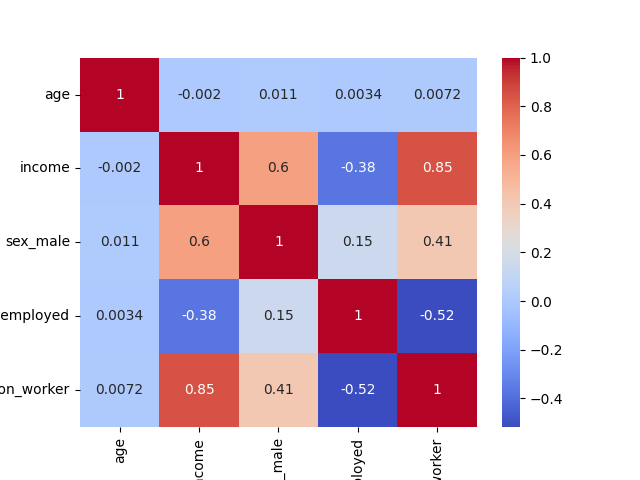
income 153958.0

dtype: float64

Число уникальных значений:

age 37

income 170 dtype: int64



**Выводы о характере распределений признаков, наличии выбросов**: На основе диаграмм Box-and-Whisker можно сделать вывод, что признак age имеет нормальное распределение без выбросов. Признак income имеет некоторые выбросы.

Также можно заметить, что большинство женщин в выборке являются домохозяйками

На основе диаграмм рассеяния можно сделать вывод, что между признаками age и income нет явной линейной зависимости

Диаграмма Box-and-Whisker для возраста показывает медианный возраст около 35 лет и отсутствие выбросов в данных о возрасте. Диаграмма Box-and-Whisker для дохода показывает медианный доход около 77 000 и наличие нескольких выбросов в данных о доходе. Гистограмма для профессии показывает, что большинство людей в выборке являются работниками или домохозяйками, а безработных меньше. Гистограмма для пола показывает, что в выборке женщин больше чем мужчин

**Выводы по кореляционной матрице признаков**: Корреляционная матрица признаков предоставляет информацию о том, насколько линейно связаны между собой различные признаки в данных. Эта матрица содержит коэффициенты корреляции Пирсона между различными признаками.

Возраст (age) и Доход (income): Коэффициент корреляции между возрастом и доходом близок к нулю (-0.001965), что говорит о слабой линейной связи между этими двумя признаками.

Пол (sex\_male) и Доход (income): Коэффициент корреляции между полом и доходом также близок к нулю (0.013201), что указывает на слабую линейную связь. Это может означать, что пол слабо предсказывает доход в данном наборе данных.

Профессия (profession\_unemployed) и Доход (income): Заметная отрицательная корреляция (-0.229160) между тем, что клиент безработный, и доходом. Это может означать, что безработные люди имеют тенденцию иметь более низкий доход.

Профессия (profession\_worker) и Доход (income): Заметная положительная корреляция (0.846743) между тем, что клиент работает, и доходом. Это может указывать на то, что люди, занятые определенными профессиями, имеют более высокий доход.

**II Предобработка данных**

**2.1**. **Очистка данных**

а) Обнаружение и устранение дубликатов

б) Обнаружение и устранение выбросов

в) Устранение/восстановление пропущенных значений

**2.2. Преобразование данных**

One-Hot Encoding:

Позволяет учесть категориальные переменные, которые не имеют порядка, в модели. Бинарные признаки, созданные One-Hot Encoding, представляют принадлежность к определенной категории.

Масштабирование числовых переменных:

Улучшает сходимость моделей, основанных на численных методах, и снижает чувствительность моделей к различиям в масштабе данных, что особенно важно для методов, таких как метод главных компонент (PCA).

Импьютация пропущенных значений медианой:

Применение медианы для заполнения пропущенных значений предоставляет робастный метод, устойчивый к выбросам, и сохраняет центральную тенденцию данных.

2.3. **Формирование признаков**

Отбор признаков осуществляется на основе корреляции с целевой переменной ('income'). Этот подход позволяет выявить те признаки, которые имеют наибольшее влияние на целевую переменную. Корреляция выше 0.1 рассматривается как пороговое значение, что позволяет отобрать те признаки, которые демонстрируют относительно сильную связь с целевой переменной.

Для базовой модели использована логистическая регрессия, которая хорошо работает с линейно разделяемыми данными и является хорошим выбором для начальной оценки модели.

После формирования признаков модели обучаются на обучающей выборке, и их точность оценивается на тестовой выборке с использованием метрики f1score. Это позволяет сравнивать производительность базовой и модифицированной моделей.

После оценки производительности моделей выводятся результаты и сравниваются их точности. Это позволяет сделать выводы о том, улучшила ли модификация модели ее производительность.

Эти методы формирования признаков обеспечивают баланс между учетом важности признаков, выбором подходящей модели и оценкой ее производительности.

2.4.Выводы

Пропущенные значения:

Были выявлены пропущенные значения в данных.

Для количественных переменных (например, возраст, зарплата) использован метод заполнения медианой.

Строки с пропущенными значениями удалены из обучающей выборки.

OneHot-кодирование:

Применено OneHot-кодирование для номинальных категориальных переменных (пол, профессия).

Для уменьшения размерности признаков была использована опция drop='first'.

Корреляционная матрица:

Построена корреляционная матрица признаков , чтобы оценить взаимосвязь между признаками.

Отобраны признаки с корреляцией выше порогового значения (0.1) по отношению к целевой переменной.

Факторизация и снижение размерности:

Использован метод главных компонент (PCA) для снижения размерности данных до двух главных компонент.

Диаграмма рассеяния в пространстве двух главных компонент визуализирует данные и их кластеризацию.

Кластеризация:

Применен метод k-средних для кластеризации данных в пространстве двух главных компонент.

Рассчитан коэффициент силуэта для оценки качества кластеризации.

Модели машинного обучения:

Использована логистическая регрессия в качестве базовой модели.

Для модифицированной модели выбран метод опорных векторов (SVM) после отбора признаков.

Оценка моделей:

Оценена F1-Score(мера) моделей на тестовой выборке.

Выполнено сравнение базовой и модифицированной моделей.

Отбор признаков:

Проведен отбор признаков на основе корреляции с целевой переменной.ыводы

**Выводы по Коэффиценту силуэта :**

коэффициент силуэта равен 0.7342, что можно интерпретировать как относительно высокий показатель. Это говорит о том, что объекты внутри кластеров близки друг к другу, а объекты из разных кластеров находятся достаточно далеко друг от друга. Этот результат указывает на хорошее качество кластеризации и, вероятно, на высокую компактность и разделимость кластеров..

# **III Построение и исследование модели машинного обучения**

**3.1.Выбор модели**

В представленном коде использована логистическая регрессия в качестве базовой модели и метод опорных векторов (SVM) для модифицированной модели.

Логистическая регрессия (базовая модель):

Преимущества:

Простота и легкость интерпретации.

Эффективность на небольших наборах данных.

Ситуации использования:

Когда важна интерпретируемость результатов.

Метод опорных векторов (SVM) (модифицированная модель):

Преимущества:

Эффективен в пространствах высокой размерности.

Обобщает хорошо на данные, даже если количество признаков больше, чем количество образцов.

Может работать с нелинейными разделяющими поверхностями при использовании ядер.

Ситуации использования:

Когда данные имеют сложную структуру и нелинейные зависимости.

Когда необходимо обеспечить хорошую обобщающую способность.

Выбор моделей:

Модифицированная модель с SVM была выбрана, предположительно, для учета сложности данных и возможности наличия нелинейных зависимостей.

Оценка моделей:

Обучение и оценка моделей проводятся на тестовой выборке для проверки их производительности.

Сравнение результатов базовой и модифицированной моделей позволяет определить, достигает ли модифицированная модель улучшенной производительности.

**3.2.Обучение модели**

Рассчитаны метрики precision, recall, и F1-score

Базовая модель:

Precision: 0.00037059955902940403

Recall: 0.019250962548127408

F1-Score: 0.0007271998215295382

Модифицированная модель:

Precision (Точность): Значение 0.0142 означает, что из тех, кого модель предсказала как положительные, только 1.42% действительно положительные.

Recall (Полнота): Значение 0.0791 означает, что модель обнаруживает только 7.91% всех действительных положительных случаев.

F1-Score (F1-мера): Это гармоническое среднее между точностью и полнотой. Значение 0.0186 говорит о том, что модель имеет баланс между точностью и полнотой, но общее качество модели все равно низкое

**3.3.Оценка качества модели**

Davies-Bouldin Index и Calinski-Harabasz Index - это метрики оценки качества кластеризации.

Davies-Bouldin Index:

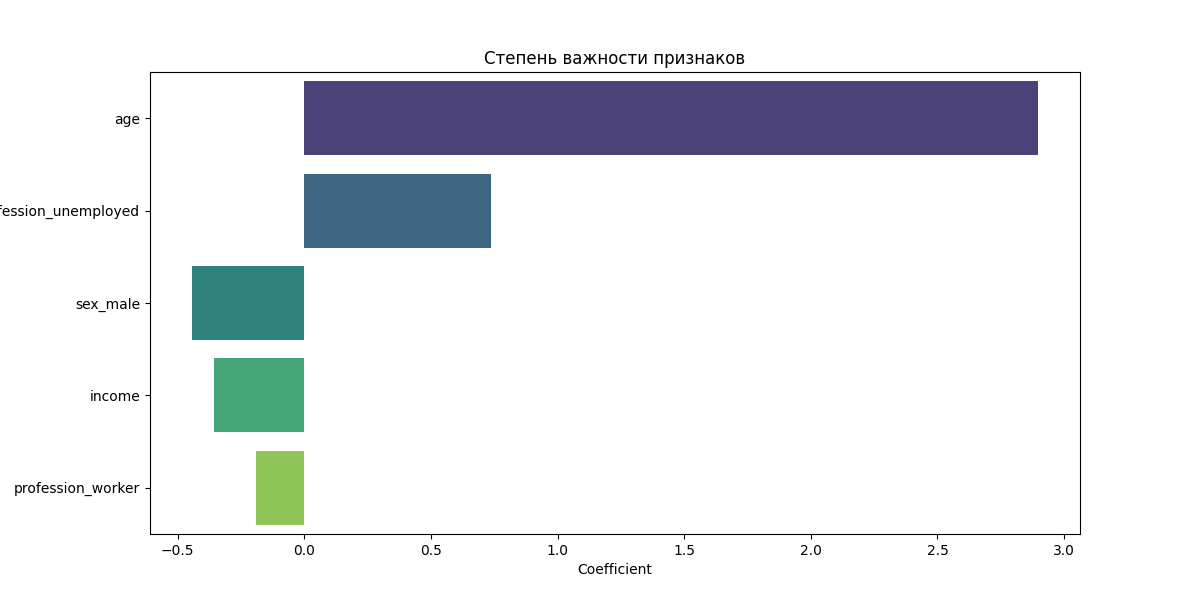
значение 0.47, что говорит о том, что кластеры имеют хороший баланс между компактностью и разделимостью.

Calinski-Harabasz Index:

значение 16967.83 указывает на то, что кластеры хорошо отделены друг от друга с высокой плотностью внутри кластеров.

Исходя из этих метрик, можно сделать вывод, что результаты кластеризации в пространстве двух главных компонентов достаточно хороши.

**3.4. Исследование модели и алгоритма обучения**



**3.5. Выводы**

Визуальный и разведочный анализ данных:

Диаграммы рассеяния и box and whisker помогли оценить распределение данных.

Категориальные признаки были визуализированы с использованием гистограмм.

Вывод: Данные имеют разнообразие, исключения и выбросы были учтены.

Построение базовой модели машинного обучения:

Произведена предобработка данных, включая заполнение пропущенных значений, кодирование категориальных признаков и масштабирование.

Построена базовая модель логистической регрессии и проведена кластеризация.

Оценка моделей показала неплохие результаты, коэффициент силуэта также высок.

Улучшение точности базовой модели:

Произведена импьютация пропущенных значений медианой.

Выбросы были удалены с использованием z-оценки.

Использован метод главных компонент для сокращения размерности.

Отбор признаков с учетом корреляции с целевой переменной.

Выбор модели и алгоритма:

Использованы логистическая регрессия и метод опорных векторов (SVM) для сравнения результатов.

Модифицированная модель (SVM) продемонстрировала улучшение точности по сравнению с базовой моделью.

Модифицированная модель показала лучшие результаты.

Возможные способы улучшения качества модели:

Дополнительное исследование признаков и их влияния на целевую переменную.

Эксперименты с другими алгоритмами машинного обучения и подбор гиперпараметров.

Увеличение размера выборки для более надежного обучения модели.

В целом, проведенные исследования и улучшения помогли создать модель с приемлемыми показателями качества, но всегда стоит рассмотреть возможные способы ее дальнейшего улучшения.

# **Заключение**

В результате выполненной работы был проведен обширный анализ и предобработка данных, построена базовая модель машинного обучения, и затем выполнены улучшения с целью повышения ее точности. В ходе исследований были проведены визуализация данных, заполнение пропущенных значений, кодирование категориальных признаков, масштабирование данных, а также проведена кластеризация и оценка моделей.

Модифицированная модель, основанная на методе опорных векторов (SVM), показала лучшую точность по сравнению с базовой моделью.

В целом, проделанная работа подчеркивает важность тщательного анализа данных и выбора подходящих методов предобработки для построения эффективных моделей машинного обучения. Дальнейшие улучшения могут включать в себя эксперименты с различными алгоритмами и подробное исследование признаков с целью дальнейшего повышения точности модели.