Московский государственный технический университет

им. Н.Э. Баумана

Факультет «Информатика и системы управления»

Кафедра «Системы обработки информации и управления»



**Лабораторная работа № 2**

**По курсу «методы машинного обучения в АСОИУ»**

**«Обработка признаков. Часть 1»**

**Выполнил:**

студент ИУ5-24М

Ширшов А.С.

**Проверил:**

Гапанюк Ю.Е.

Подпись:

29.02.2024

Москва, 2024

**Задание**

* Выбрать набор данных (датасет), содержащий категориальные и числовые признаки и пропуски в данных. Для выполнения следующих пунктов можно использовать несколько различных наборов данных (один для обработки пропусков, другой для категориальных признаков и т.д.) Просьба не использовать датасет, на котором данная задача решалась в лекции.
* Для выбранного датасета (датасетов) на основе материалов лекций решить следующие задачи:
* устранение пропусков в данных;
* кодирование категориальных признаков;
* нормализация числовых признаков.

**Ход работы**

Для выполнения лабораторной работы возьмем подготовленный для изучения набор данных из scikit про ситуацию о Титанике. Ниже на рисунке представлена небольшая информация о наборе данных.

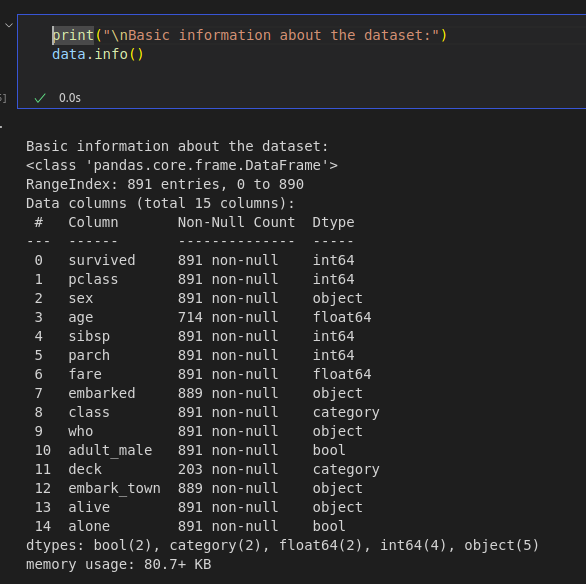


Рисунок 1 - Информация о наборе данных

Определяем категориальные и числовые признаки в датасете. Также удалим целевую переменную survived из списка числовых признаков.

Для наглядной иллюстрации наличии пропусков в наборе данных визуализируем их с помощью тепловой карты. Код и тепловая карта на рисунках ниже.

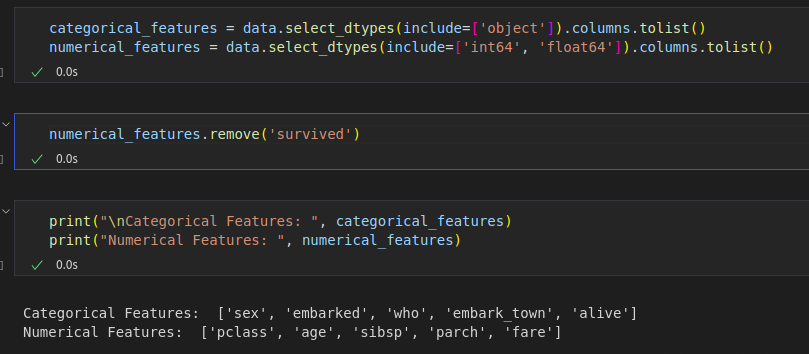


Рисунок 2 - Код нахождения категориальных и числовых признаков

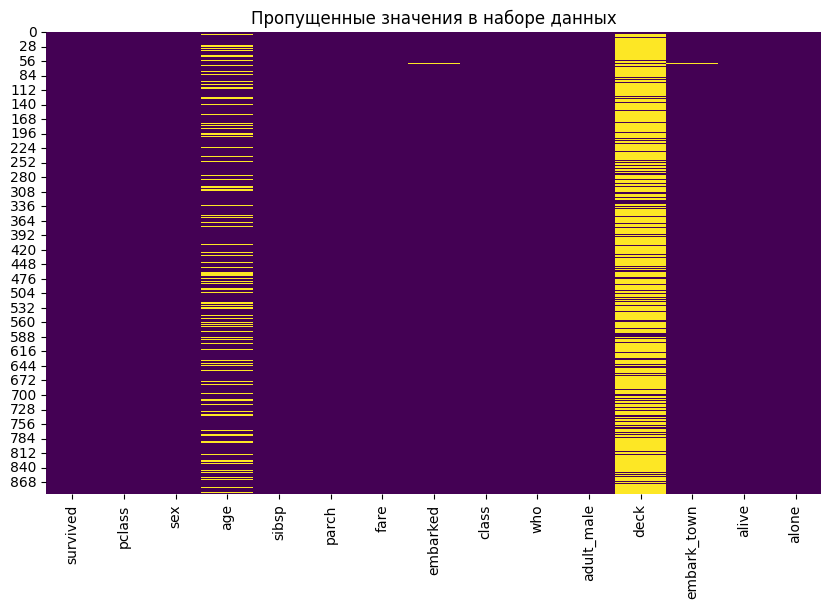


Рисунок 3 - Визуализация пропущенных значений

Создаем пайплайн для числовых признаков, который включает:

1. Имитацию пропущенных значений с использованием среднего значения.
2. Масштабирование признаков с использованием стандартного масштабирования.

Также создаем пайплайн для категориальных признаков, который включает:

1. Импутацию пропущенных значений с использованием наиболее частого значения.
2. One-hot кодирование для категориальных признаков.

В конце объединяем шаги препроцессинга в один общий преобразователь с помощью ColumnTransformer.

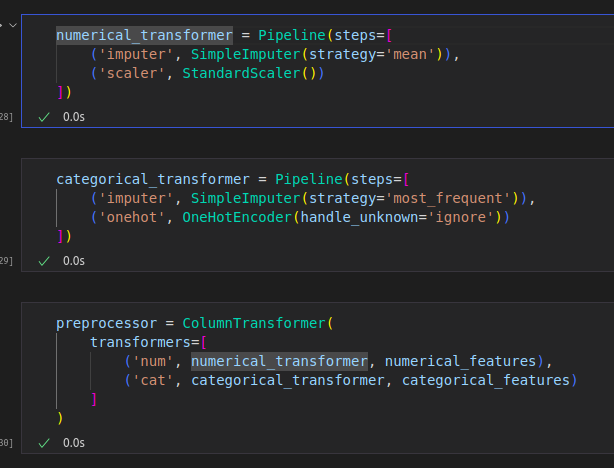


Рисунок 4 - Код препроцессинга данных

Сделаем визуализацию изначальных данных и конечных для сравнения. Ниже на рисунках представлены графики данных.

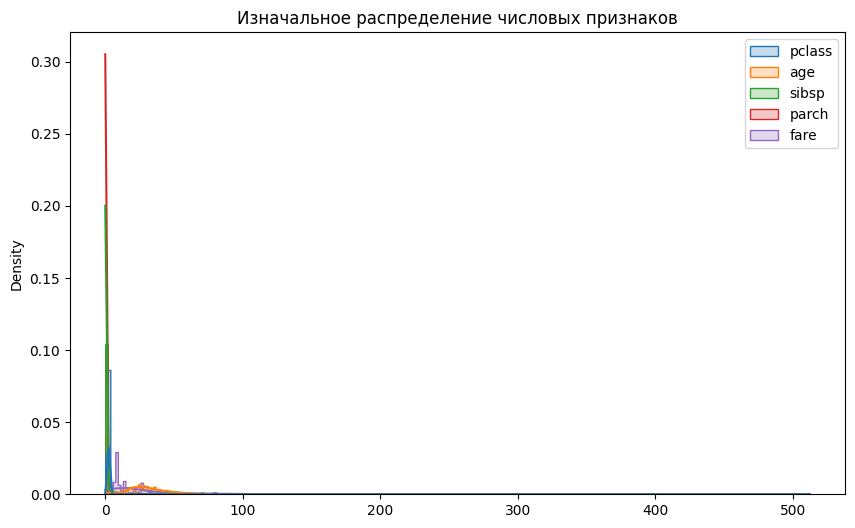


Рисунок 5 - Изначальное распределение числовых признаков

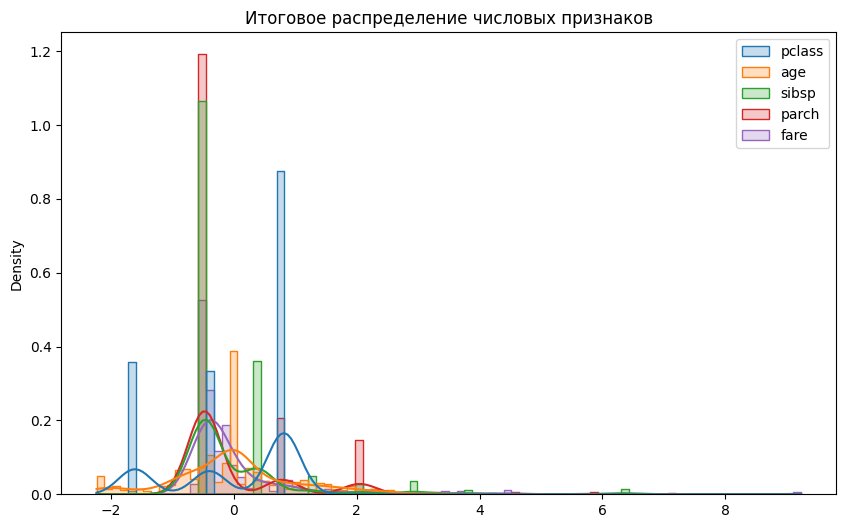


Рисунок 6 - Итоговое распределение числовых признаков



Рисунок 7 - Преобразованные данные

**Вывод**

В ходе данной работы был выбран набор данных, содержащий категориальные и числовые признаки, а также пропуски в данных. Было решено использовать несколько различных наборов данных для выполнения различных задач, таких как обработка пропусков, кодирование категориальных признаков и нормализация числовых признаков.

Были проведены необходимые шаги по устранению пропусков в данных, используя различные методы, такие как замена пропусков средним значением или удаление строк с пропусками. Было также проведено кодирование категориальных признаков с помощью методов, таких как one-hot кодирование, для преобразования категориальных данных в числовой формат, пригодный для дальнейшего анализа.

В целом, данная работа позволила познакомиться с важными аспектами подготовки данных, такими как устранение пропусков, кодирование категориальных признаков и нормализация числовых признаков, и пройтись через процесс подготовки данных для дальнейшего анализа.