****Московский государственный технический университет им. Н.Э. Баумана  
Кафедра «Системы обработки информации и управления»

Лабораторная работа №2  
по дисциплине  
«Методы машинного обучения»  
на тему

«Обработка признаков»

Выполнил:  
студент группы ИУ5-22М  
ЧжаоЛян

Москва — 2022 г.

**Цель лабораторной работы:**

Изучение продвинутых способов предварительной обработки данных для дальнейшего формирования моделей.

**Задание:**

1. 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.)
2. Для выбранного датасета (датасетов) на основе материалов лекций решить следующие задачи:
   * устранение пропусков в данных;
   * кодирование категориальных признаков;
   * нормализацию числовых признаков.

## Загрузка и первичный анализ данных

图片包含 表格

描述已自动生成

## Обработкa пропусков в данных

图形用户界面, 文本, 应用程序

描述已自动生成

图形用户界面, 文本, 应用程序

描述已自动生成

Датасет достаточно маленький, а доля пропущенных значений не очень большая. В таком случае можно поработать с внедрением пропущенных значений.

### Заполнение значений для одного признака

文本

描述已自动生成

文本

低可信度描述已自动生成

图形用户界面

描述已自动生成

图形用户界面, 文本, 应用程序, Word

描述已自动生成

Распределения одномодальные, поэтому можно использовать для импутации моду.

### Для категориальных признаков

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成

图表, 瀑布图

描述已自动生成

图表, 瀑布图

描述已自动生成

### KNN

图片包含 文本

描述已自动生成

图形用户界面

描述已自动生成

## кодирование категориальных признаков

文本

描述已自动生成

## label encoding

表格

描述已自动生成

## one-hot encoding

表格

描述已自动生成

图形用户界面, 应用程序, 表格

描述已自动生成

## Count (frequency) encoding

日历

中度可信度描述已自动生成

## Target (Mean) encoding

图形用户界面, 应用程序, 表格

中度可信度描述已自动生成

## Weight of evidence (WoE) encoding

表格

中度可信度描述已自动生成

文本

低可信度描述已自动生成 图形用户界面, 文本, 应用程序

描述已自动生成 图形用户界面, 条形图

中度可信度描述已自动生成

## нормализация числовых признаков

图表, 条形图, 瀑布图

描述已自动生成

## 图形用户界面, 文本, 应用程序 描述已自动生成 图表, 直方图 描述已自动生成

## Логарифмическое преобразование

图形用户界面, 图表, 直方图

描述已自动生成

## Обратное преобразование

图表

描述已自动生成

## Преобразование Бокса-Кокса

图表, 直方图

描述已自动生成

## Преобразование Йео-Джонсона

图表, 直方图

描述已自动生成

## Возведение в степень

图表

描述已自动生成

**Список литературы**

[1] Гапанюк Ю. Е. Лабораторная работа «Обработка пропусков в данных, кодирование категориальных признаков, масштабирование данных» [Электронный ресурс] // GitHub. –– 2019. –– Режим доступа: https://github.com/ugapanyuk/ml\_course/wiki/LAB\_MISSING (дата обращения: 05.04.2019).

[2] Team The IPython Development. IPython 7.3.0 Documentation [Electronic resource]//Read the Docs. — 2019. — Access mode: https://ipython.readthedocs.io/en/stable/ (online; accessed: 20.02.2019).

[3] Waskom M. seaborn 0.9.0 documentation [Electronic resource] // PyData. — 2018. — Access mode: https://seaborn.pydata.org/ (online; accessed: 20.02.2019).

[4] pandas 0.24.1 documentation [Electronic resource] // PyData. — 2019. — Access mode: http://pandas.pydata.org/pandas-docs/stable/ (online; accessed: 20.02.2019).

[5] Gupta L. Google Play Store Apps [Electronic resource] // Kaggle. — 2019. — Access mode: https://www.kaggle.com/lava18/google-play-store-apps (online; accessed:05.04.2019).