**Московский государственный технический университет им. Н.Э.Баумана**

|  |  |  |
| --- | --- | --- |
|  |  |  |

**Отчет по лабораторной работе №3 по курсу**

**Технологии машинного обучения**

«Обработка пропусков в данных, кодирование категориальных признаков, масштабирование данных»

2

(количество листов)

|  |  |
| --- | --- |
| ИСПОЛНИТЕЛЬ: |  |
| студент группы **ИУ5-62** | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
|  | (подпись) |
| **Писарчук Н.** | "2" марта 2019 г. |

Москва, МГТУ - 2019

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

### Задание:

1. Выбрать набор данных (датасет), содержащий категориальные 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 [лекции](https://github.com/ugapanyuk/ml_course/blob/master/common/notebooks/missing/handling_missing_norm.ipynb) решить следующие задачи:

* обработку пропусков в данных;
* кодирование категориальных признаков;
* масштабирование данных.

**Выполнение**:

Я выбрала датасет Wine Reviews, содержащий описание различных вин. Выполненное задание представлено в виде ноутбука Lab3–wine.ipynb.

Описание датасета:

country - The country that the wine is from

description

designation - The vineyard within the winery where the grapes that made the wine are from

points - The number of points WineEnthusiast rated the wine on a scale of 1-100 (though they say they only post reviews for wines that score >=80)

price - The cost for a bottle of the wine

province - The province or state that the wine is from

region\_1 - The wine growing area in a province or state (ie Napa)

region\_2 - Sometimes there are more specific regions specified within a wine growing area (ie Rutherford inside the Napa Valley), but this value can sometimes be blank

taster\_name

taster\_twitter\_handle

title - The title of the wine review, which often contains the vintage if you're interested in extracting that feature

variety - The type of grapes used to make the wine (ie Pinot Noir)

winery - The winery that made the wine

На основе проведенных вычислений для улучшения данных можно удалить следующие признаки:

1. region2 - это уточнение region1, в нем содержиться более 60% пропусков, поэтому его можно убрать
2. designayion имеет около трети пропусков
3. taster\_twitter\_handle - это уточнение taster\_name

Для выполнения операция масштабирования и нормализации в колонке price значения NaN были изменены на среднее значение с помощью Imputer из scikit learn.

**Выводы:**

Пропущенные значения в категориальных признаках надо заполнять наиболее часто встречающимся значением, а в числовые – средним значением.

Для преобразования категориальных признаков в числовые надо использовать кодирование категорий наборами бинарных значений - one-hot encoding или Pandas get\_dummies.