Министерство образования и науки Российской Федерации

Федеральное государственное автономное образовательное учреждение

высшего образования

«Южно-Уральский Государственный университет»

(национальный исследовательский университет)

Высшая школа электроники и компьютерных наук

Кафедра «Информационно-аналитическое обеспечение управления в социальных и экономических системах»

**Отчёт по преддипломной практике**

Выполнил:

Студент группы КЭ-221

Гарипов А.Р.

«26» апреля 2018 года

Проверил:

Клепиков М.О.

Работа защищена с оценкой

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

«\_\_\_»\_\_\_\_\_\_\_\_\_\_\_\_2018 г.

Челябинск 2018

**Оглавление**

[Введение 3](#_Toc513657436)

[1. Сбор данных 4](#_Toc513657437)

[2. Анализ и подготовка собранных данных 6](#_Toc513657438)

[2.1. Поиск выбросов в наименовании товара 8](#_Toc513657439)

[2.2. Поиск выбросов в тексте категории товара 10](#_Toc513657440)

[2.3. Поиск выбросов в тексте характеристик 12](#_Toc513657441)

[Заключение 16](#_Toc513657442)

[Библиографический список 17](#_Toc513657443)

# **Введение**

Компания Proanalytics занимается мониторингом цен на товары в интернет-магазинах, и предоставляет различные отчеты для магазинов конкурентов по собранным данным о товарах с различных сайтов. Для компании важно иметь собственный классификатор товаров, который на данный момент составляется вручную. Это представляет огромные затраты и большой объем рутинной работы. На данный момент существуют технологии, которые позволяют автоматизировать эту работу.

Цель преддипломной практики — осуществить сбор и подготовку данных для разработки классификатора товаров на основе машинного обучения.

# **Сбор данных**

Вся информация по товарам хранится в единой базе данных. В качестве СУБД используется MySQL.

Был проведен опрос стейкхолдеров, для выяснения того, как и какие данные собираются, в каких таблицах можно найти ценную информацию. Так же было необходимо выяснить насколько зашумленными могут быть данные. В итоге удалось выяснить следующее.

Всего накоплено 1770580 товаров привязанных к товару.

Данные, которые собираются с интернет-магазинов по товарам:

1. Наименование
2. Текст категории
3. Статус (В наличии / Нет в наличии)
4. Цена
5. Ссылка на изображение
6. Характеристики товаров
7. Описание
8. Артикул товара на сайте
9. Дата сбора
10. Ссылка на товар
11. Компания (интернет-магазин), с которой был собран товар
12. Бренд

У всех товаров собираются:

1. наименование
2. текст категории
3. статус
4. цена, ссылка на изображение
5. артикул товара на сайте
6. дата сбора
7. ссылка на товар
8. ссылка на изображение

И только для некоторых компаний собираются:

1. Характеристики товаров
2. Описание
3. Бренд

Среди товаров у который есть привязка к категории

товаров с характеристиками: 382087

С описанием: 209651

С брендами: 1284510

Данные о привязках верны, и без ошибок. Так утверждал стейкхолдер, но, как оказалось, этому утверждению нельзя доверять.

Для построения первой версии модели машинного обучения было решено использовать текстовую информацию, поскольку она наиболее полно характеризует категорию товара:

1. Наименование
2. Текст категории
3. Характеристики

Поскольку многие товары в обучающей выборке не имеют описания, эту информацию решено не включать в обучающую выборку.

Бренды также были исключены, поскольку они содержат много ошибок, и определяются автоматически через наименование товара.

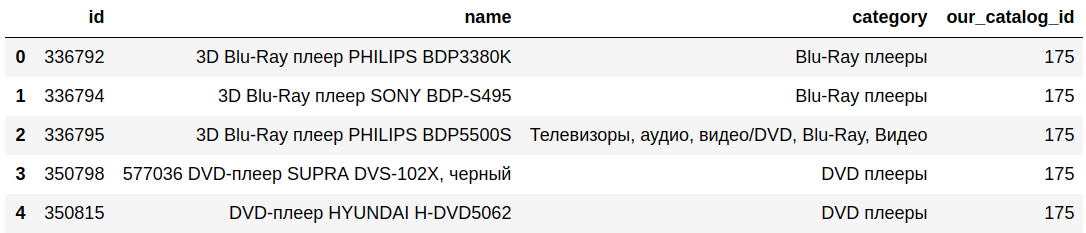
Картинки так же можно использовать, но хорошее качество в классификации они способны показать лишь при использовании глубокого обучения, на многослойных нейронных сетях, которые способны генерировать высокоуровневые признаки из изображений. Но для их обучения требуется большое количество изображений и вычислительных ресурсов. Поэтому для начала было решено их не использовать.

После отсеивания пропусков и лишних данных было получено 342012 товаров.

# **Анализ и подготовка собранных данных**

Прежде чем как либо обрабатывать данные, нужно их проанализировать. Нужно понять какие ошибки встречаются в данных. Возможно где-то может не хватать данных. Для начала найдем выбросы.

Для начала посмотрим как выглядят наименования и категории товаров. Для этого выведем 5 товаров в виде таблицы. Таблица представлена на рисунке 1.

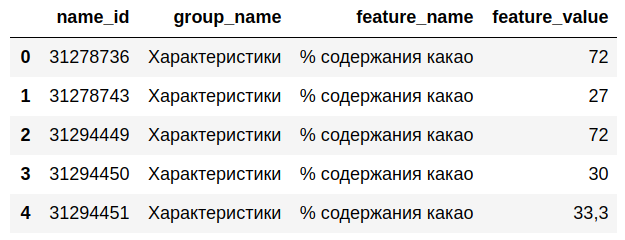
Рис. 1 - Наименования и категории товаров.

В таблице id — это идентификатор товара в базе, name — наименование, category — текст категории, our\_catalog\_id — идентификатор каталога.

Как мы видим наименования и текст категории содержат различный регистр символов, символы пунктуации, и прочие символы, которые необходимо заменить на пробел. Символы верхнего регистра необходимо перевести в нижний. Необходимо оставить только буквы, числа и пробелы. К тому же в тексте встречаются слова, состоящие только из чисел, такие слова тоже удалим.

Структура каталога имеет 4 уровня вложенности. Но товары относятся только к последнему 4-му уровню. Т.е. к первым трем уровням привязок товаров нет.

К тому же были собраны характеристики товаров. Эти характеристики иметь привязку один товар ко многим характеристикам. Т.е. несколько записей характеристик могут относиться к одному товару. Посмотрим записи собранных из базы характеристик. Характеристики представлены на рисунке 2.

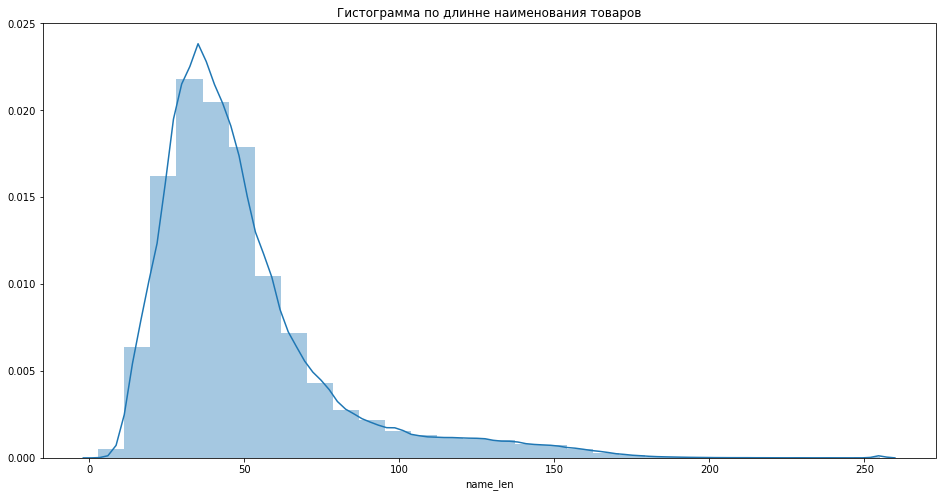
Рис. 2 – Характеристики.

Здесь group\_name — это наименование группы характеристик. Характеристики могут быть сгруппированы. feature\_name — название характеристики, feature\_value — значение характеристики.

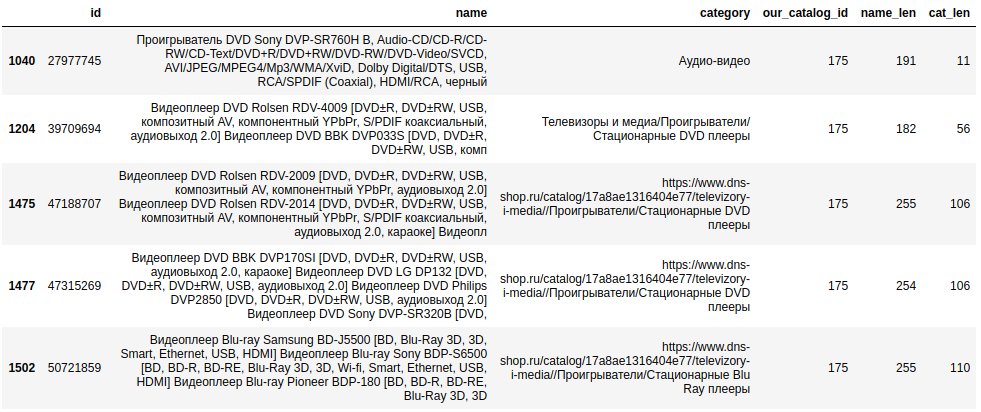
## **2.1. Поиск выбросов в наименовании товара**

В статистике выбросы – это значения, резко отличающиеся от других значений в собранном наборе данных. Выброс может указывать на аномалии в распределении данных или на ошибки при измерениях, поэтому зачастую выбросы исключаются из набора данных. Исключив выбросы из набора данных, вы можете прийти к неожиданным или более точным выводам. Поэтому необходимо уметь вычислять и оценивать выбросы, чтобы обеспечить надлежащее понимание статистических данных.

Для того чтобы найти выбросы, посмотрим на наименования с слишком малой длинной наименования и слишком большой. Для этого построим гистограмму, по которой можно увидеть распределение количества товаров из выборки по длине наименования. Она представлена на рисунке 3.

Рис. 3 –Гистограмма по длине наименования товара.

По гистограмме видно, что количество товаров, у которых длинна наименования меньше 180, резко уменьшается, практически до 0. Но в конце после 250 увеличивается. Это подозрительно. Посмотрим на товары у которых длинна наименования больше 180 (рис 4). Для примера возьмем 5 товаров.

Рис. 4 –Товары с длинными наименованиями.

Видно, что у этих товаров в наименовании присутствуют характеристики. К тому же у них присутствуют ошибки в тексте категории. Поскольку таких товаров немного, удалим их из выборки.

Далее посмотрим на слишком короткие наименования (рис. 5). Возьмем товары, с длиной наименования менее 4-х символов.

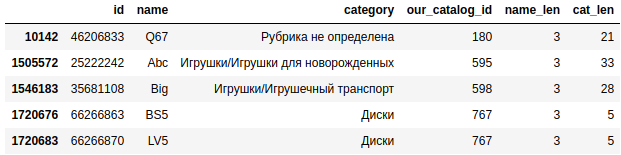


Рис. 5 –Товары с короткими наименованиями.

У этих товаров в наименовании только текст модели. Также удалим их, поскольку их не много, а текст модели никак не характеризует категорию.

Так же мы видим, что у товаров может быть категория «Рубрика не определена». Для товаров, где в тексте категории содержится «Рубрика не определена», будем оставлять пустой текст.

## **2.2. Поиск выбросов в тексте категории товара**

Для поиска выбросов в тексте категории построим такую же гистограмму, как в пункте выше, только для текста категории (рис.6).

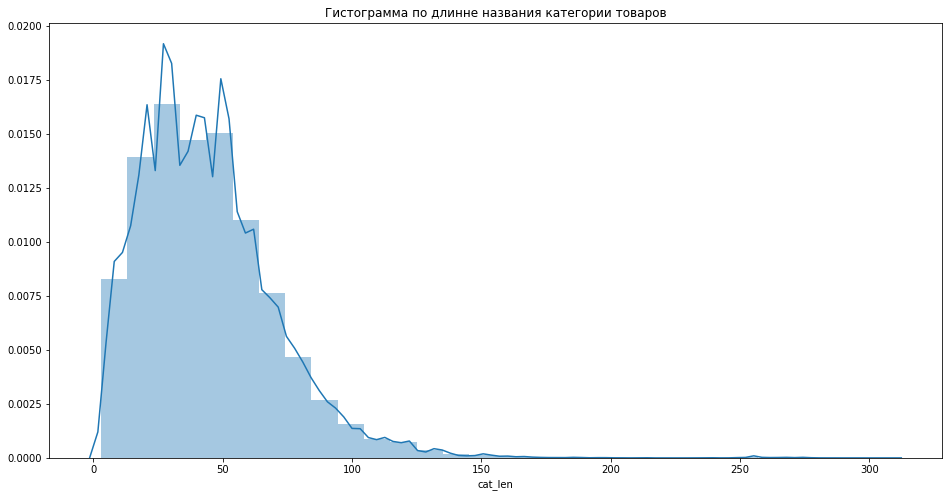


Рис. 6 –Гистограмма товаров по длине категорий.

Видим, что есть аномальные, по длине категории, товары, при длине больше 150 символов. Посмотрим на эти товары.(рис.7)

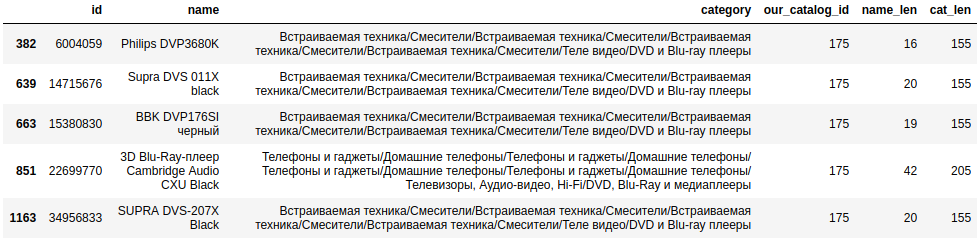


Рис. 7 –Товары с длинной категорией.

В этих товарах дубли категории. Удалим их, поскольку их также немного.

Посмотрим на слишком короткие категории с длиной текста менее 4-х символов (рис.8).

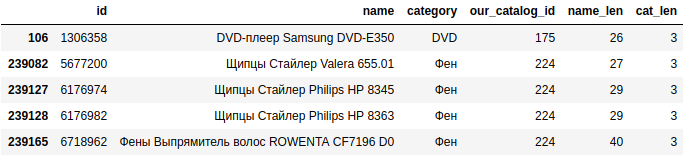


Рис. 8 –Товары с короткой категорией.

У этих товаров все нормально, поэтому их можно оставить.

## **Поиск выбросов в тексте характеристик**

Прежде чем искать выбросы, сразу удалим из характеристик различные спецсимволы, такие как «%», «-», «@», знаки пунктуации, и другие. Переведем все в нижний регистр. Также удалим слова, которые состоят только из чисел.

Построим гистограммы по названию групп характеристик. (рис.9)

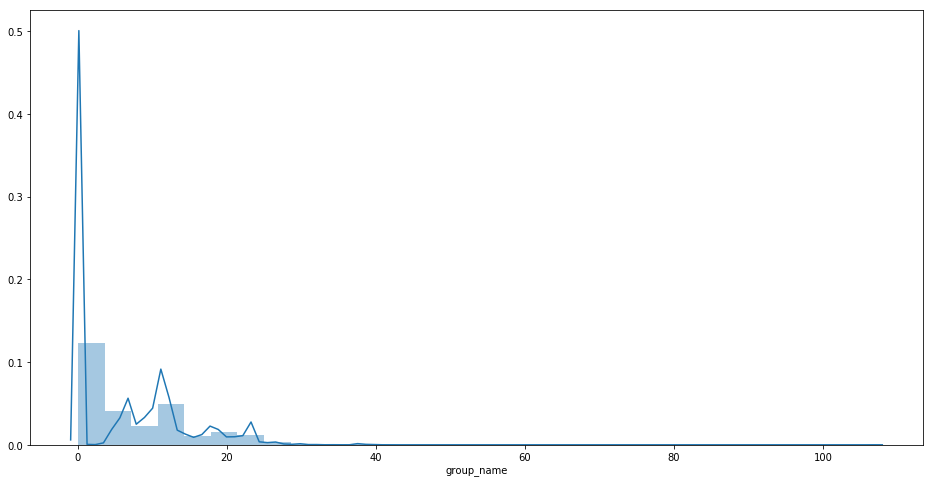


Рис. 9 - Гистограммы по названию групп характеристик.

Для наименования групп аномальными являются группы с длинной более 40 символов.

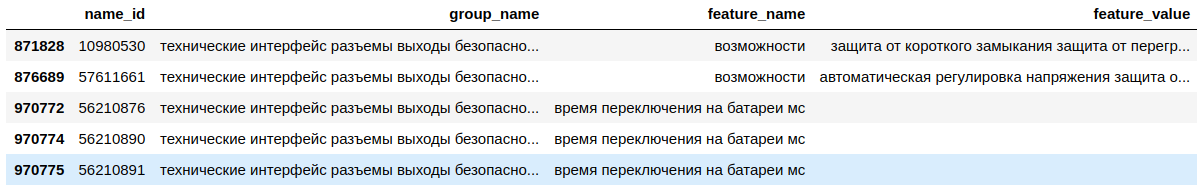


Рис. 10 – Длинные характеристики

Среди этих товаров мы видим, что названия групп собраны неправильно, эти характеристики с ошибками, удалим их.(рис.10)

Построим гистограмму по названию характеристик. (рис.11)



Рис. 11 – Гистограмма по названию характеристик

Видим, что при длине более 50-ти символов начинаются аномалии. Посмотрим на эти характеристики.

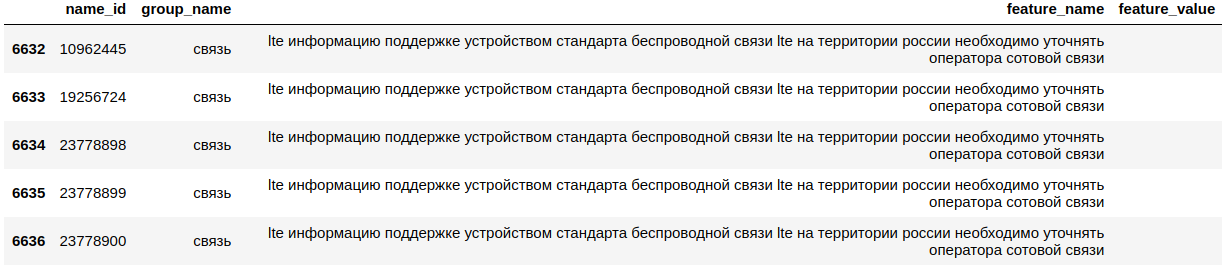


Рис. 12 – Длинные названия характеристик.

Видно, что у этих характеристик, название ошибочное (рис.12). Эти характеристики были неправильно собранны с интернет-магазинов. Удалим эти характеристики.

Теперь посмотрим гистограмму по длине значения характеристик (поле feature\_value).

Аномальные характеристики, начинаются при длине более 40 символов (рис. 13). Посмотрим на них (рис. 14).



Рис. 13 – Гистограмма по длине значения характеристик.

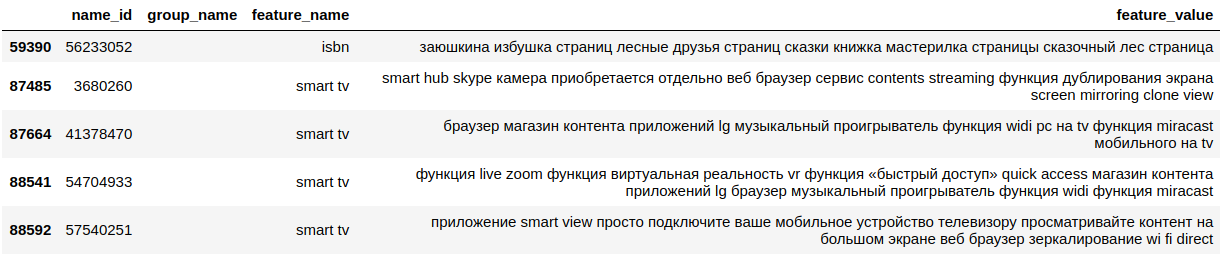


Рис. 14 – Длинные значения характеристик.

У этих товаров так же ошибки. Эти характеристики были собраны неправильно с интернет-магазинов.

Далее преобразуем все характеристики в мешок слов для каждого товара. Т.е. возьмем группу характеристик склеим её со всеми наименованиями и значениями характеристик.

Затем объединим мешки характеристик с наименованиями и посмотрим на полученную выборку (рис.15).



Рис. 15 – Полученная выборка.

Видно, что в тексте категорий остались лишние слова. На примере последний товар имеет слова «главная», «каталог», который нужно удалить из текста категории.

# **Заключение**

В результате работы практики были собраны необходимые данные, и осуществлен их анализ и найдены выбросы. Были найдены аномальные значения в тексте товаров и их характеристик.

Эти данные позволят создать классификатор товаров с применением машинного обучения.

# Библиографический список

1. Гудфеллоу Я., Бенджио И. Глубокое обучение / пер. с анг. А. А. Слинкина. – 2-е изд., испр. – М.: ДМК Пресс, 2018. – 652 с.: цв. ил.
2. Блог компании ГК ЛАНИТ - CRISP-DM: проверенная методология для DataScientist-ов.– 2017 - <https://habr.com/company/lanit/blog/328858/>
3. Введение в анализ текстовой информации с помощью Python и методов машинного обучения.– 2013 -<https://habr.com/post/205360/>
4. Seaborn documentation. - <http://seaborn.pydata.org>