Минобрнауки России
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Волгоградский государственный технический университет»

Факультет Электроники и вычислительной техники

Кафедра Электронно-вычислительные машины и системы

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА
к курсовой работе (проекту)

по дисциплине Системы обработки больших данных

на тему: Исследование датасета Big Sales Data с использованием фреймворка Apache Spark

Студент Челядинов Дмитрий Владимирович
(фамилия, имя, отчество)

Группа САПР-1.1

Руководитель работы (проекта) _____ П.Д. Кравченя
(подпись и дата подписания) (инициалы и фамилия)

Члены комиссии:

_____ _____
(подпись и дата подписания) (инициалы и фамилия)

_____ _____
(подпись и дата подписания) (инициалы и фамилия)

_____ _____
(подпись и дата подписания) (инициалы и фамилия)

Нормоконтроллер _____ П.Д. Кравченя
(подпись и дата подписания) (инициалы и фамилия)

Волгоград 2025

Минобрнауки России
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Волгоградский государственный технический университет»

Факультет Электроники и вычислительной техники

Направление (специальность) Информатика и вычислительная техника

Кафедра Электронно-вычислительные машины и системы

Дисциплина Системы обработки больших данных

Утверждаю
Зав. кафедрой А.Е. Андреев

«____» _____ 20 ____ г.

ЗАДАНИЕ
на курсовую работу (проект)

Студент Челядинов Дмитрий Владимирович
(фамилия, имя, отчество)

Группа САПР-1.1

1. Тема: Исследование датасета Big Sales Data с использованием фреймворка Apache Spark

Утверждена приказом от «___» _____ 20 ___ г., № _____.

2. Срок представления работы (проекта) к защите «___» _____ 20 ___ г.

3. Содержание расчётно-пояснительной записки: РАЗВЕДОЧНЫЙ АНАЛИЗ ДАННЫХ С ИСПОЛЬЗОВАНИЕМ PYSPARK; МАШИННОЕ ОБУЧЕНИЕ НА БОЛЬШИХ ДАННЫХ.

4. Перечень графического материала: _____

5. Дата выдачи задания «___» _____ 20 ___ г.

Руководитель работы (проекта) _____ П.Д. Кравченя
(подпись и дата подписания) (инициалы и фамилия)

Задание принял к исполнению _____ Д.В. Челядинов
(подпись и дата подписания) (инициалы и фамилия)

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Актуальность данной курсовой работы обусловлена возрастающей потребностью в разработке эффективных методов машинного обучения для обработки больших данных [1–3] и извлечения полезных знаний из пользовательского контента [4]. Особую значимость приобретают комплексные подходы, сочетающие предварительную обработку данных и построение прогностических моделей на распределенных вычислительных платформах.

В связи с этим целью данной курсовой работы является исследование и реализация полного цикла машинного обучения на больших данных о книгах и пользовательских отзывах с использованием фреймворка Apache Spark [5; 6].

Для достижения поставленной цели выдвинуты следующие задачи:

1. Загрузка и первичное исследование структуры данных из распределенной файловой системы HDFS;
2. Выполнение базовых преобразований и очистки данных для подготовки к машинному обучению;
3. Применение алгоритмов машинного обучения на больших данных;
4. Анализ и оценка качества построенных прогностических моделей;
5. Визуализация результатов и подготовка выводов по исследованию.

В первом разделе рассмотрена более подробно постановка задачи и проведен обзор современных методов машинного обучения на больших данных. Во втором разделе описана методика предварительной обработки данных и выполнена верификация качества очистки. В третьем разделе представлены результаты применения алгоритмов машинного обучения, а также анализ эффективности построенных моделей. В заключении работы сформулированы общие выводы по результатам проведенного исследования.

1 РАЗВЕДОЧНЫЙ АНАЛИЗ ДАННЫХ С ИСПОЛЬЗОВАНИЕМ PYSPARK

1.1 Постановка задачи разведочного анализа

Разведочный анализ данных (Exploratory Data Analysis, EDA) является ключевым этапом при работе с большими данными [1; 2; 4] и определяет качество последующих шагов — очистки, трансформации и моделирования. Целью данной главы является проведение комплексного EDA большого набора данных о книгах и пользовательских отзывах с использованием возможностей фреймворка Apache Spark.

Для исследования применяются распределённые вычисления, что позволяет эффективно обрабатывать миллионы записей [5; 7] и анализировать данные в условиях ограничений по памяти и времени. Использование PySpark обеспечивает масштабируемость, а интеграция с HDFS — удобство работы с большими объёмами данных.

Основные задачи разведочного анализа заключаются в следующем:

- загрузка данных из распределённой файловой системы HDFS и формирование единого датафрейма;
- исследование структуры, схемы и качества данных;
- выявление пропусков, дубликатов и некорректных значений;
- преобразование типов данных и создание производных признаков;
- предварительная оценка распределений количественных признаков и анализа категориальных данных;
- подготовка очищенного и структурированного набора данных для последующего применения алгоритмов машинного обучения.

Результатом данного этапа является построение целостного представления о данных и формирование корректной основы для дальнейших шагов анализа и моделирования.

1.2 Описание исходного датасета

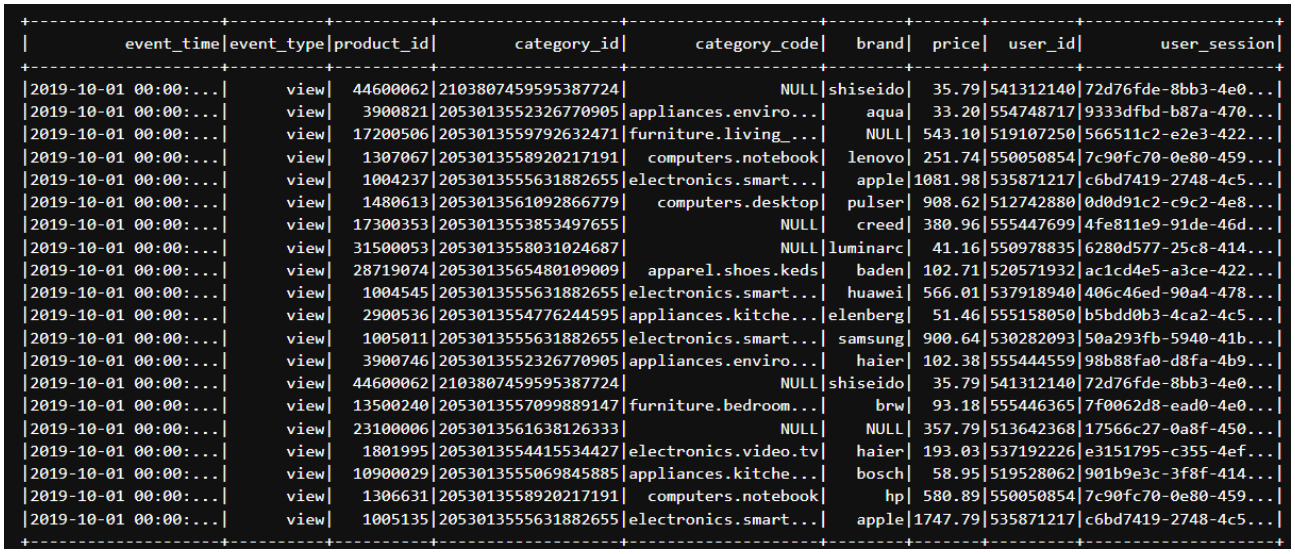
В работе используется датасет «eCommerce behavior data from multi category store», доступный на платформе Kaggle [8]. Набор данных состоит из одного датасета — 2019-Nov.csv. В дальнейшем его название измениться на dataset.csv

Совокупный объём данных составляет более 67 миллионов строк. Загруженные данные изначально имеют строковые типы, разнородные форматы идентификаторов и включают большое количество текстовых полей.

Датафрейм включает следующие ключевые признаки (таблица 1):

Таблица 1 – Описание признаков датасета

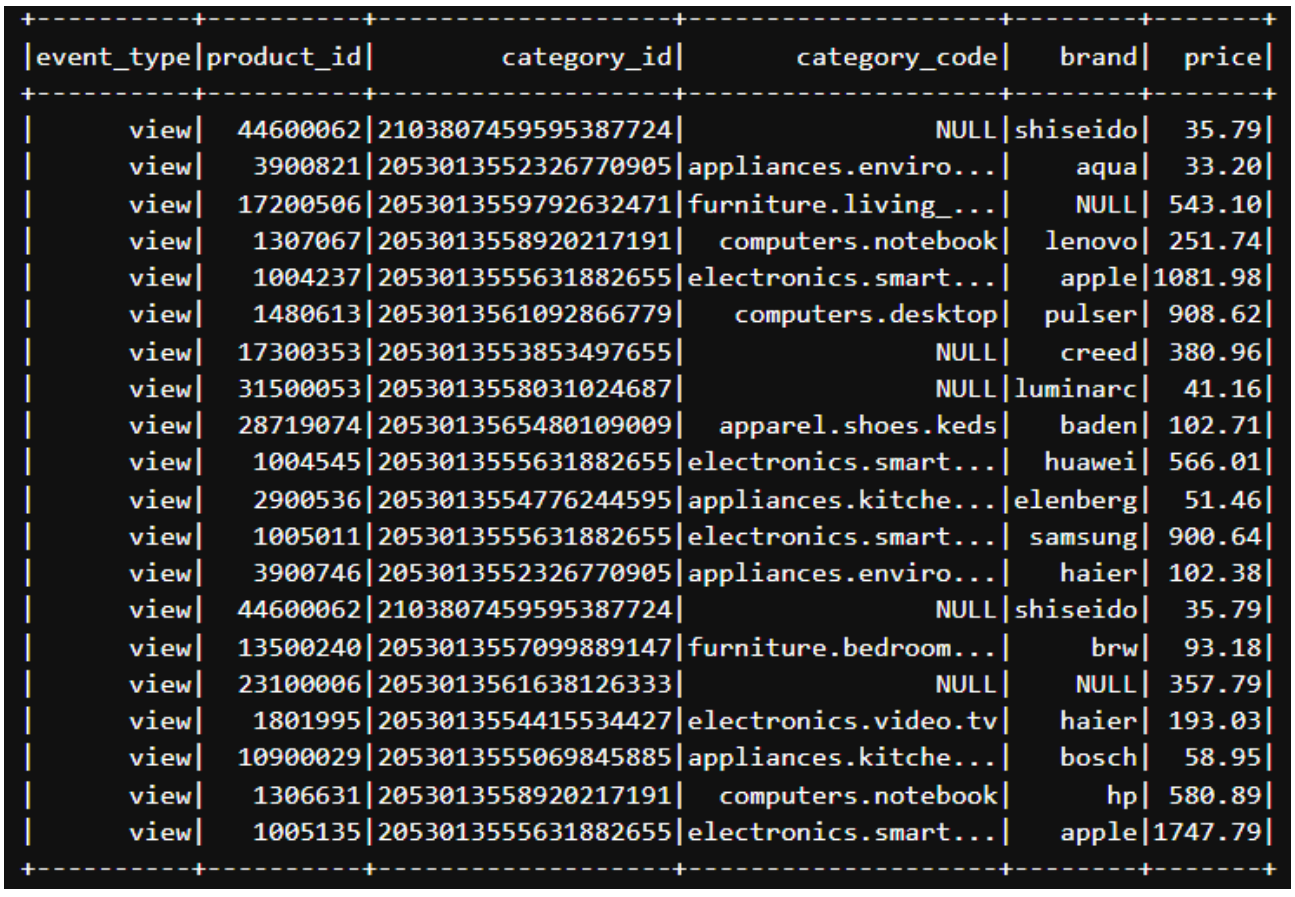
Признак	Описание
event_time	Время, когда произошло событие (UTC).
event_type	Вид события.
product_id	Идентификатор продукта.
category_id	Идентификатор категории продукта.
category_code	Таксономия категории товара (кодовое название), если это было возможно. Обычно указывается для значимых категорий и пропускается для различных видов аксессуаров.
brand	Строка с названием бренда.
price	Цена продукта.
user_id	Идентификатор пользователя.



event_time	event_type	product_id	category_id	category_code	brand	price	user_id	user_session
2019-10-01 00:00:...	view	44600062	2103807459595387724	NULL	shiseido	35.79	541312140	72d76fde-8bb3-4e0...
2019-10-01 00:00:...	view	3900821	2053013552326770905	appliances.enviro...	aqua	33.20	554748717	9333dfbd-b87a-470...
2019-10-01 00:00:...	view	17200506	2053013559792632471	furniture.living_...	NULL	543.10	519107250	566511c2-e2e3-422...
2019-10-01 00:00:...	view	1307067	2053013558920217191	computers.notebook	lenovo	251.74	550050854	7c90fc70-0e80-459...
2019-10-01 00:00:...	view	1004237	2053013555631882655	electronics.smart...	apple	1081.98	535871217	c6bd7419-2748-4c5...
2019-10-01 00:00:...	view	1480613	2053013561092866779	computers.desktop	pulser	908.62	512742880	0d0d91c2-c9c2-4e8...
2019-10-01 00:00:...	view	17300353	2053013553853497655	NULL	creed	380.96	555447699	4fe811e9-91de-46d...
2019-10-01 00:00:...	view	31500053	2053013558031024687	NULL	luminarc	41.16	550978835	6280d577-25c8-414...
2019-10-01 00:00:...	view	28719074	2053013565480109009	apparel.shoes.keds	baden	102.71	520571932	ac1cd4e5-a3ce-422...
2019-10-01 00:00:...	view	1004545	2053013555631882655	electronics.smart...	huawei	566.01	537918940	406c46ed-90a4-478...
2019-10-01 00:00:...	view	2900536	2053013554776244595	appliances.kitche...	elenberg	51.46	555158050	b5bdd0b3-4ca2-4c5...
2019-10-01 00:00:...	view	1005011	2053013555631882655	electronics.smart...	samsung	900.64	530282093	50a293fb-5940-41b...
2019-10-01 00:00:...	view	3900746	2053013552326770905	appliances.enviro...	haier	102.38	555444559	98b88fa0-d8fa-4b9...
2019-10-01 00:00:...	view	44600062	2103807459595387724	NULL	shiseido	35.79	541312140	72d76fde-8bb3-4e0...
2019-10-01 00:00:...	view	13500240	2053013557099889147	furniture.bedroom...	brw	93.18	555446365	7f0062d8-ead0-4e0...
2019-10-01 00:00:...	view	23100006	2053013561638126333	NULL	NULL	357.79	513642368	17566c27-0a8f-450...
2019-10-01 00:00:...	view	1801995	2053013554415534427	electronics.video.tv	haier	193.03	537192226	e3151795-c355-4ef...
2019-10-01 00:00:...	view	10900029	2053013555069845885	appliances.kitche...	bosch	58.95	519528062	901b9e3c-3f8f-414...
2019-10-01 00:00:...	view	1306631	2053013558920217191	computers.notebook	hp	580.89	550050854	7c90fc70-0e80-459...
2019-10-01 00:00:...	view	1005135	2053013555631882655	electronics.smart...	apple	1747.79	535871217	c6bd7419-2748-4c5...

Рисунок 1 – Данные из датасета

В ходе анализа полей датасета, с помощью команды `df.select()` были использованы следующие поля: `event_type`, `product_id`, `category_id`, `category_code`, `brand`, `price`.



event_type	product_id	category_id	category_code	brand	price
view	44600062	2103807459595387724	NULL	shiseido	35.79
view	3900821	2053013552326770905	appliances.enviro...	aqua	33.20
view	17200506	2053013559792632471	furniture.living_...	NULL	543.10
view	1307067	2053013558920217191	computers.notebook	lenovo	251.74
view	1004237	2053013555631882655	electronics.smart...	apple	1081.98
view	1480613	2053013561092866779	computers.desktop	pulser	908.62
view	17300353	2053013553853497655	NULL	creed	380.96
view	31500053	2053013558031024687	NULL	luminarc	41.16
view	28719074	2053013565480109009	apparel.shoes.keds	baden	102.71
view	1004545	2053013555631882655	electronics.smart...	huawei	566.01
view	2900536	2053013554776244595	appliances.kitche...	elenberg	51.46
view	1005011	2053013555631882655	electronics.smart...	samsung	900.64
view	3900746	2053013552326770905	appliances.enviro...	haier	102.38
view	44600062	2103807459595387724	NULL	shiseido	35.79
view	13500240	2053013557099889147	furniture.bedroom...	brw	93.18
view	23100006	2053013561638126333	NULL	NULL	357.79
view	1801995	2053013554415534427	electronics.video.tv	haier	193.03
view	10900029	2053013555069845885	appliances.kitche...	bosch	58.95
view	1306631	2053013558920217191	computers.notebook	hp	580.89
view	1005135	2053013555631882655	electronics.smart...	apple	1747.79

Рисунок 2 – Данные из датасета после select

Структура данных была изучена с использованием команды `df.printSchema()`, что позволило определить типы полей и выявить их

потенциальную неоднородность. Так, поля `event_type`, `product_id`, `category_id`, `category_code`, `brand`, `price` и текстовые поля загружаются как строки, что указывает на возможное наличие разнородных форматов данных. Структура представлена на рисунке 3.

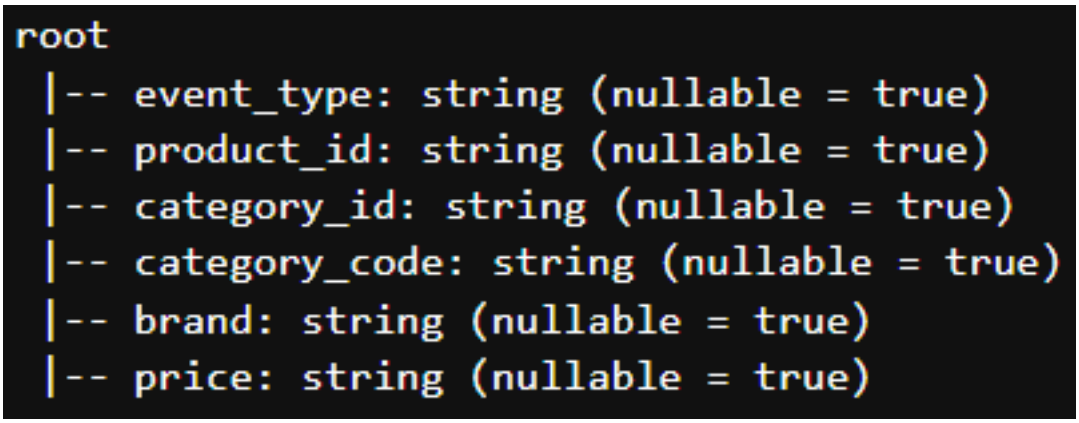


Рисунок 3 – Структура таблицы после загрузки данных

Более детальное изучение содержимого выполнялось уже на этапе разведочного анализа при помощи выборочного просмотра записей `df.show()`, анализа уникальных значений и регулярных выражений. Эти методы позволили установить, что:

- числовые идентификаторы различной длины (от 6 до 8 цифр);
- category_id имеет длинные числовые идентификаторы (19-20 цифр);
- только одно значение view во всех строках выборки;
- множество значений NULL;
- значительный разброс цен (от бюджетных товаров до премиальных);
- числовые значения с двумя десятичными знаками.

Параметры Spark-сессии были настроены с учётом объёма данных [1; 2]: увеличены объёмы памяти драйвера и исполнителей. Это обеспечивает стабильную работу при чтении и трансформации больших датафреймов. На рисунке 4 показана конфигурация SparkSession.

В конфиге, указанном ниже, последние строчки указывают на то, что используется hadoop:

```
def create_spark_configuration() -> SparkConf:
    """
```

```
    Создает и конфигурирует экземпляр SparkConf для приложения Spa

    Returns:
        SparkConf: Настроенный экземпляр SparkConf.
    """
    # Получаем имя пользователя
    user_name = "dchel"

    conf = SparkConf()
    conf.setAppName("Lab 1")
    conf.setMaster("local[*]")
    conf.set("spark.submit.deployMode", "client")
    conf.set("spark.executor.memory", "12g")
    conf.set("spark.executor.cores", "8")
    conf.set("spark.executor.instances", "2")
    conf.set("spark.driver.memory", "4g")
    conf.set("spark.driver.cores", "2")
    conf.set("spark.sql.catalog.spark_catalog.type", "hadoop")
    conf.set("spark.sql.catalog.spark_catalog.warehouse", f"hdfs:/
    conf.set("spark.sql.catalog.spark_catalog.io-impl", "org.apach

    return conf
```

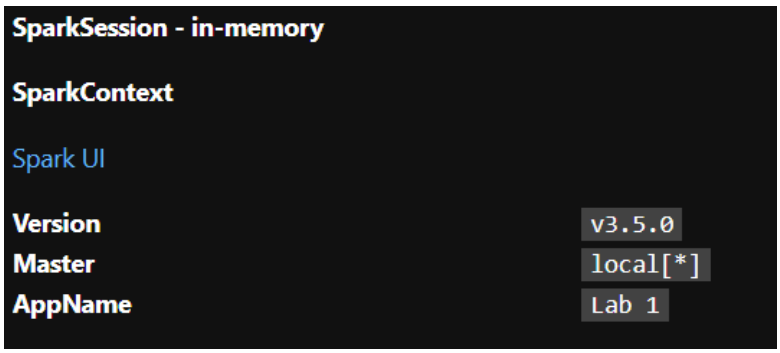


Рисунок 4 – Демонстрация работы сессии Spark

Также для работы с датасетом, он был предварительно загружен в hdfs. Обзор директории представлен на рисунке 5.

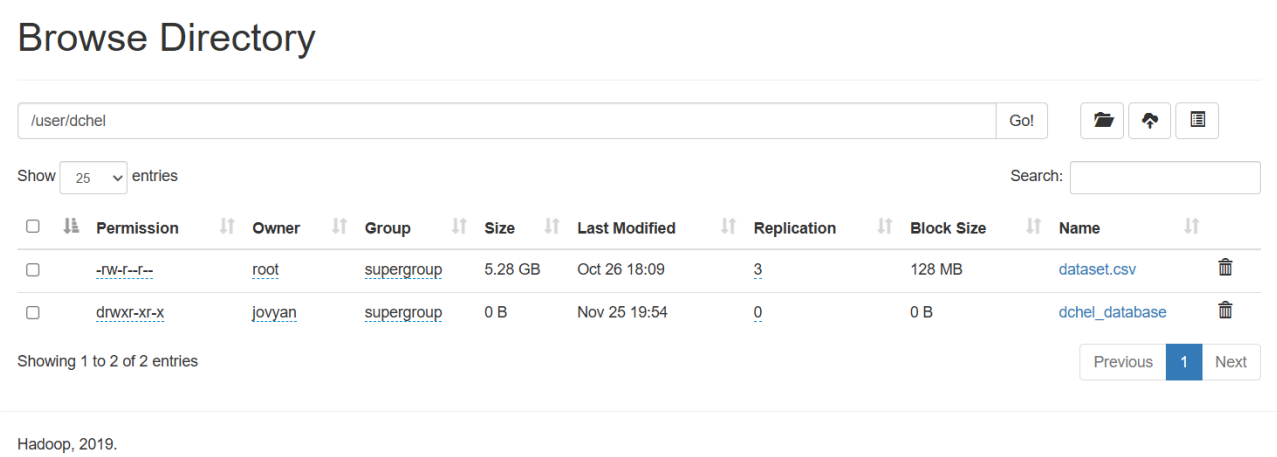


Рисунок 5 – Директория с данными для работы

1.3 Определение пропущенных значений и преобразование данных

Анализ полноты данных показал наличие пропусков в ряде столбцов, преимущественно в текстовых полях. Для оценки количества пропусков была использована служебная функция, выполняющая подсчёт `NULL`-значений в каждом столбце.

Были применены следующие стратегии:

– удалены строки содержащие `NULL`-значения в столбцах `category_code` и `brand`;

– текстовый столбец `category_id` был удален с помощью команды `df.drop("category_id")`.

После очистки структура данных была расширена и дополнена новыми признаками. В частности, выполнено преобразование типов:

– численный перевод идентификаторов продуктов (`product_id`);

– перевод кодов категории в массив строк (`category_code`);

– численный перевод цены (`price`).

Дополнительно созданы следующие производные признаки:

– массив кодов категорий, полученный путём разбиения строки с использованием `split`;

– признак содержания вида продукта `contains_appliances`, `contains_computers`, `contains_electronics`, `contains_kitch` `contains_smartphone`;
– булевый признак дороговизны продукта `is_expensive`;
– булевый признак бюджетного продукта `is_budget`;
– булевый признак среднебюджетного продукта `is_mid_range`;
– кол-во категорий, которые охватывают продукт `category_count`;
– булевый признак просмотра продукта `is_view`;
– булевый признак добавление продукта в корзину `is_cart`;
– булевый признак покупки продукта `is_purchase`;
– класс продукта `price_range`;
– класс продукта в численном формате `price_range_numeric`.

Результаты продемонстрированы на рисунках 6 и 7.

```
root
 |-- event_type: string (nullable = true)
 |-- product_id: integer (nullable = true)
 |-- brand: string (nullable = true)
 |-- price: double (nullable = true)
 |-- contains_appliances: boolean (nullable = true)
 |-- contains_computers: boolean (nullable = true)
 |-- contains_electronics: boolean (nullable = true)
 |-- contains_kitchen: boolean (nullable = true)
 |-- contains_smartphone: boolean (nullable = true)
 |-- is_expensive: integer (nullable = false)
 |-- is_budget: integer (nullable = false)
 |-- is_mid_range: integer (nullable = false)
 |-- category_count: integer (nullable = true)
 |-- is_purchase: integer (nullable = false)
 |-- is_view: integer (nullable = false)
 |-- is_cart: integer (nullable = false)
 |-- price_range: string (nullable = false)
 |-- price_range_numeric: integer (nullable = false)
```

Рисунок 6 – Структура таблицы после обработки данных

	event_type	product_id	brand	price	contains_appliances	contains_computers	contains_electronics	contains_kitchen	contains_smartphone	is_expensive	is_budget	is_mid_range	category_count	is_purchase	is_view
0	cart	1002042	samsung	77.139999	False	False	True	False	True	0	0	1	2	0	0
1	cart	1002524	apple	513.450012	False	False	True	False	True	1	0	0	2	0	0
2	cart	1002524	apple	513.469971	False	False	True	False	True	1	0	0	2	0	0
3	cart	1002524	apple	531.409973	False	False	True	False	True	1	0	0	2	0	0
4	cart	1002524	apple	533.260010	False	False	True	False	True	1	0	0	2	0	0
5	cart	1002524	apple	540.270020	False	False	True	False	True	1	0	0	2	0	0
6	cart	1002536	apple	576.570007	False	False	True	False	True	1	0	0	2	0	0
7	cart	1002542	apple	488.790009	False	False	True	False	True	1	0	0	2	0	0
8	cart	1002544	apple	460.070007	False	False	True	False	True	1	0	0	2	0	0
9	cart	1002544	apple	460.309998	False	False	True	False	True	1	0	0	2	0	0

Рисунок 7 – Фрагмент данных после обработки

Все преобразования были объединены в функцию, позволяющую повторно применять трансформации к датафрейму. См. Приложение А.

1.4 Анализ распределений, выбросов и категориальных признаков

Для количественного признака `price`, была построена гистограмма с использованием Spark и последующей визуализацией в библиотеках Seaborn и Matplotlib.

Полученные результаты показывают:

– оценки пользователей смещены в сторону высоких значений (мода 4–5) (8);
– коэффициент полезности характеризуется бимодальным распределением, что отражает различия между отзывами без оценивания полезности и отзывами с активным пользовательским голосованием (**??**).

Рисунок 8 – Гистограмма распределения для price

На рисунках 9, 10 продемонстрировано распределение аномальных значений у price.

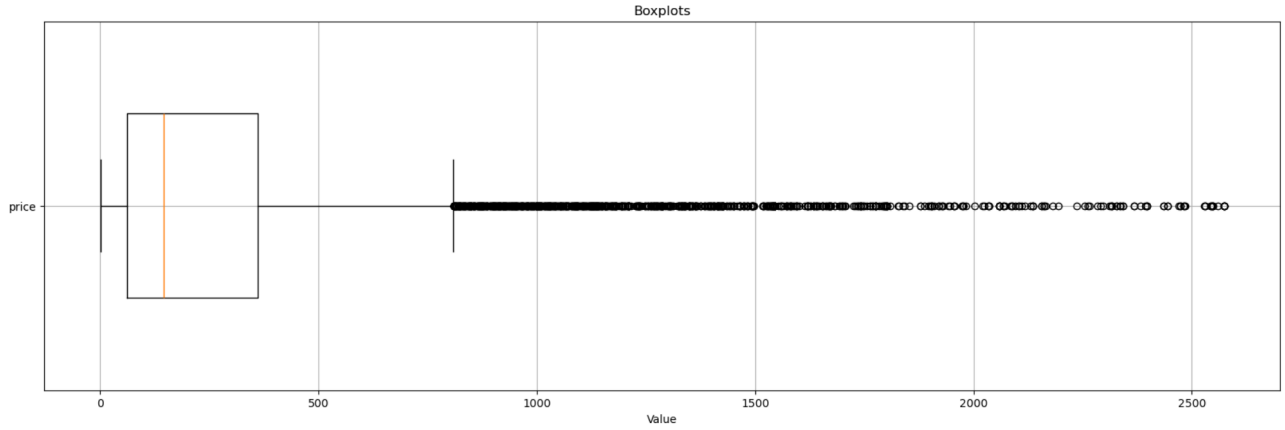


Рисунок 9 – Пример аномалий у price

Минимальное значение:	0.88
Среднее значение:	301.55
Среднеквадратичное отклонение:	390.91
Первый квартиль:	61.52
Медиана:	146.08
Третий квартиль:	360.11
Максимальное значение:	2574.07

Рисунок 10 – Расчетные значения у price

Для категориальных признаков `category_code`, `brand` были проанализированы частоты встречаемости. Самым популярным значением:

– `brand` стало samsung (рис. 11);

– `category_code` стало electronics (рис. 12).

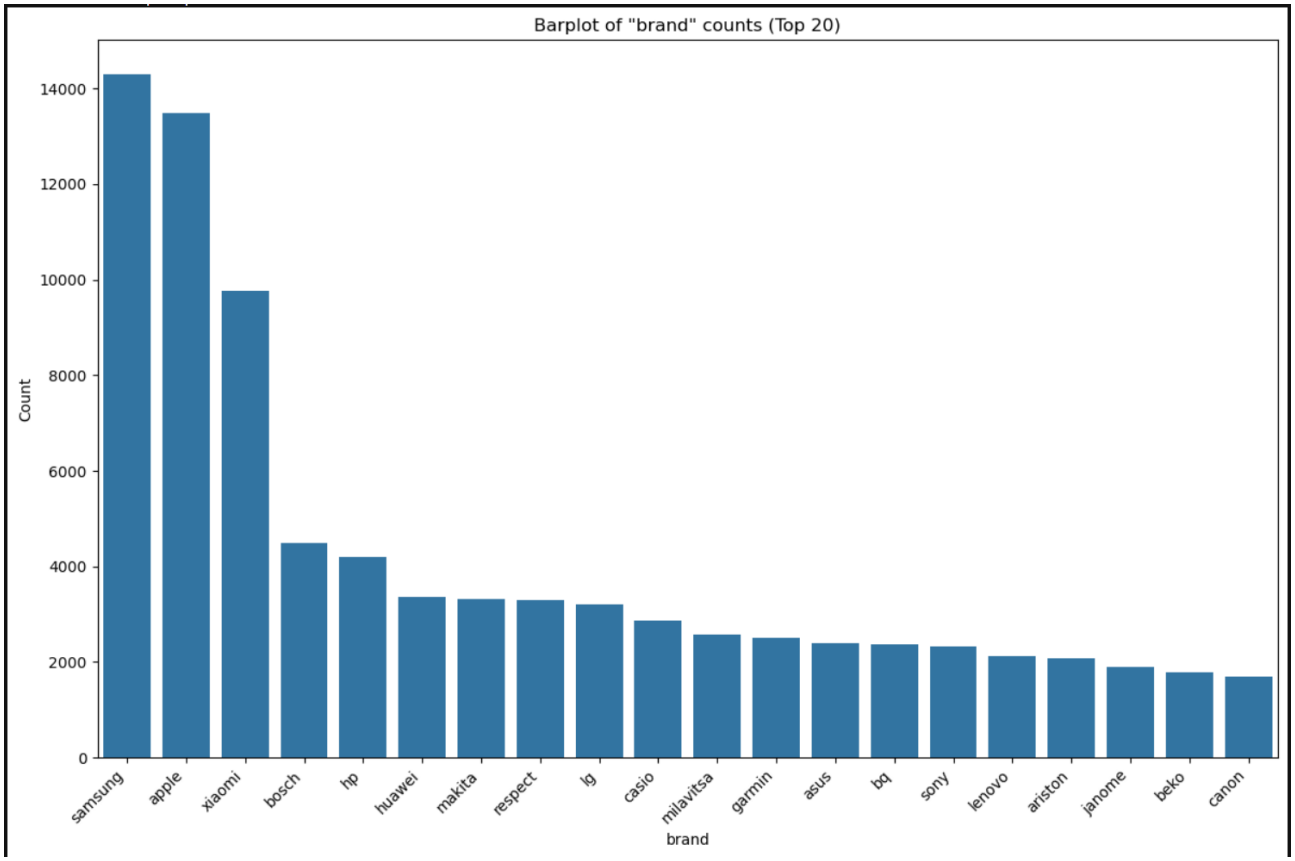


Рисунок 11 – Частоты для brand

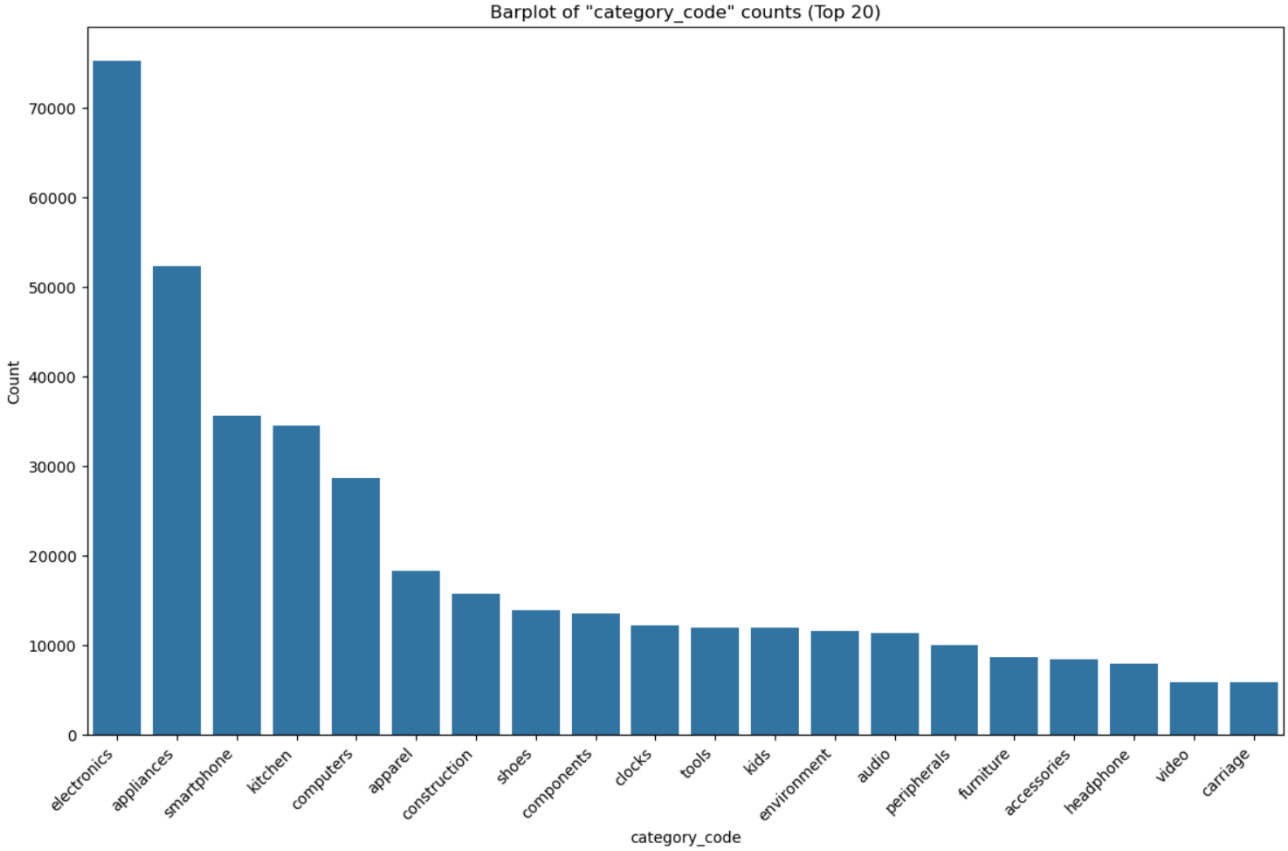


Рисунок 12 – Частоты для category_code

Проверка дубликатов показала, что некоторые товары имеют повторяющиеся записи — это связано с тем, что каждая запись соответствует отдельному пользовательскому действию. Такие дубликаты являются ожидаемыми и отражают структуру исходного набора данных. Поэтому для всего датасета была проведена дедубликация по `product_id`, `event_type` и `price` полям.

1.5 Выводы

В ходе проведённого разведочного анализа был сформирован целостный и очищенный набор данных, готовый для применения алгоритмов машинного обучения. Основными результатами являются:

- выполнена загрузка и объединение данных из HDFS средствами Apache Spark;
- исследована структура данных, выявлены и обработаны пропуски и аномалии;
- нормализованы текстовые поля и преобразованы числовые, бинарные признаки;
- сформированы новые признаки, повышающие информативность данных;
- проведён анализ распределений и выбросов в количественных характеристиках;
- выявлены особенности набора данных, связанные с дублированием записей по идентификатору продукта.

2 МАШИННОЕ ОБУЧЕНИЕ НА БОЛЬШИХ ДАННЫХ

В данной главе рассматриваются методы построения и оценки моделей машинного обучения в распределённой среде Apache Spark [1–3; 6]. Работа включает решение двух задач: прогнозирования числовой оценки цены товара (регрессия) и классификации бюджетного сегмента продукта. Все вычисления выполнялись с использованием фреймворка Apache Spark и библиотеки Spark ML [8; 9], обеспечивающих обработку данных объёмом около трёх миллионов записей.

2.1 Задача регрессии

2.1.1 Постановка задачи регрессии

Необходимо построить модель GBT регрессии для предсказания цены продукта (price) на основе доступных признаков. Цель: найти нелинейную зависимость между признаками и целевой переменной, минимизируя ошибку предсказания. Качество модели оценивается метриками RMSE и R^2 на тестовой выборке. Модель должна объяснить, какие факторы влияют на цену товара и насколько сильно.

В качестве признаков были выделены следующие группы:

```
binary_features = [
    "is_expensive",
    "is_budget",
    "is_mid_range",
    "is_purchase",
    "is_view",
    "is_cart",
    "contains_appliances",
    "contains_computers",
    "contains_electronics",
    "contains_kitchen",
```

```
    "contains_smartphone"
]

numeric_features = ["category_count"]

categorical_features = ["brand", "event_type",
                        "price_range", "price_range_numeric"]
```

Из анализа были исключены признаки, не влияющие на целевую переменную или потенциально приводящие к переобучению: идентификаторы (`category_id`, `event_time`, `user_id`, `user_session`).

event_time	event_type	product_id	category_id	category_code	brand	price	user_id	user_session
2019-10-01 00:00:...	view	44600062	2103807459595387724	NULL	shiseido	35.79	541312140	72d76fde-8bb3-4e0...
2019-10-01 00:00:...	view	3900821	2053013552326770905	appliances.enviro...	aqua	33.20	554748717	9333dfbd-b87a-470...
2019-10-01 00:00:...	view	17200506	2053013559792632471	furniture.living_...	NULL	543.10	519107250	566511c2-e2e3-422...
2019-10-01 00:00:...	view	1307067	2053013558920217191	computers.notebook	lenovo	251.74	550050854	7c90fc70-0e80-459...
2019-10-01 00:00:...	view	1004237	2053013555631882655	electronics.smart...	apple	1081.98	535871217	c6bd7419-2748-4c5...
2019-10-01 00:00:...	view	1480613	2053013561092866779	computers.desktop	pulser	908.62	512742880	0d0d91c2-c9c2-4e8...
2019-10-01 00:00:...	view	17300353	2053013553853497655	NULL	creed	380.96	555447699	4fe811e9-91de-46d...
2019-10-01 00:00:...	view	31500053	2053013558031024687	NULL	luminarc	41.16	550978835	6280d577-25c8-414...
2019-10-01 00:00:...	view	28719074	2053013565480109009	apparel.shoes.keds	baden	102.71	520571932	ac1cd4e5-a3ce-422...
2019-10-01 00:00:...	view	1004545	2053013555631882655	electronics.smart...	huawei	566.01	537918940	406c46ed-90a4-478...
2019-10-01 00:00:...	view	2900536	2053013554776244595	appliances.kitche...	elenberg	51.46	555158050	b5bdd0b3-4ca2-4c5...
2019-10-01 00:00:...	view	1005011	2053013555631882655	electronics.smart...	samsung	900.64	530282093	50a293fb-5940-41b...
2019-10-01 00:00:...	view	3900746	2053013552326770905	appliances.enviro...	haier	102.38	555444559	98b88fa0-d8fa-4b9...
2019-10-01 00:00:...	view	44600062	2103807459595387724	NULL	shiseido	35.79	541312140	72d76fde-8bb3-4e0...
2019-10-01 00:00:...	view	13500240	2053013557099889147	furniture.bedroom...	brw	93.18	555446365	7f0062d8-ead0-4e0...
2019-10-01 00:00:...	view	23100006	2053013561638126333	NULL	NULL	357.79	513642368	17566c27-0a8f-450...
2019-10-01 00:00:...	view	1801995	2053013554415534427	electronics.video.tv	haier	193.03	537192226	e3151795-c355-4ef...
2019-10-01 00:00:...	view	10900029	2053013555069845885	appliances.kitche...	bosch	58.95	519528062	901b9e3c-3f8f-414...
2019-10-01 00:00:...	view	1306631	2053013558920217191	computers.notebook	hp	580.89	550050854	7c90fc70-0e80-459...
2019-10-01 00:00:...	view	1005135	2053013555631882655	electronics.smart...	apple	1747.79	535871217	c6bd7419-2748-4c5...

Рисунок 13 – Фрагмент датафрейма с исходными данными

	event_type	product_id	brand	price	contains_appliances	contains_computers	contains_electronics	contains_kitchen	contains_smartphone	is_expensive	is_budget	is_mid_range	category_count	is_purchase	is_view
0	cart	1002042	samsung	77.139999	False	False	True	False	True	0	0	1	2	0	0
1	cart	1002524	apple	513.450012	False	False	True	False	True	1	0	0	2	0	0
2	cart	1002524	apple	513.469971	False	False	True	False	True	1	0	0	2	0	0
3	cart	1002524	apple	531.409973	False	False	True	False	True	1	0	0	2	0	0
4	cart	1002524	apple	533.260010	False	False	True	False	True	1	0	0	2	0	0
5	cart	1002524	apple	540.270020	False	False	True	False	True	1	0	0	2	0	0
6	cart	1002536	apple	576.570007	False	False	True	False	True	1	0	0	2	0	0
7	cart	1002542	apple	488.790009	False	False	True	False	True	1	0	0	2	0	0
8	cart	1002544	apple	460.070007	False	False	True	False	True	1	0	0	2	0	0
9	cart	1002544	apple	460.309998	False	False	True	False	True	1	0	0	2	0	0

Рисунок 14 – Фрагмент датафрейма с обработанными данными

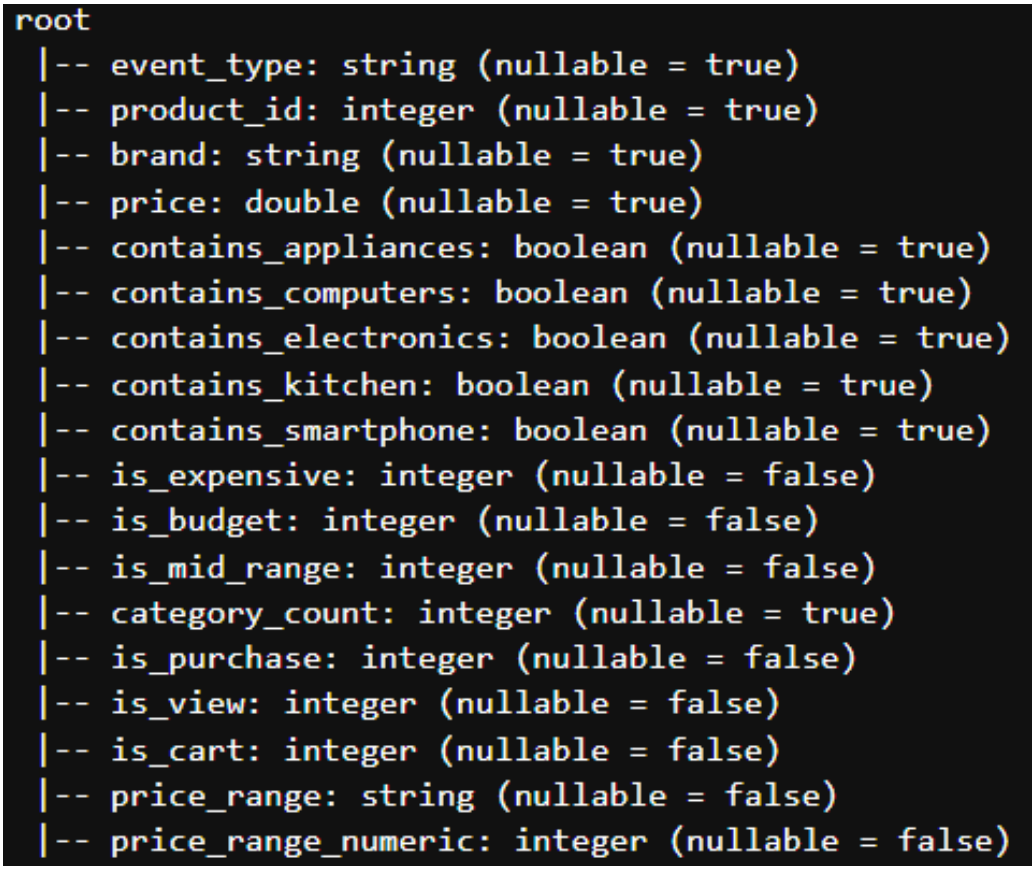


Рисунок 15 – Схема данных

Для оценки качества модели использовались метрики RMSE (Root Mean Square Error) и R^2 (коэффициент детерминации).

2.1.2 Решение задачи регрессии

Построение модели линейной регрессии начиналось с подготовки данных: датасет загружался из HDFS в формате Parquet, после чего из него выделялся небольшой сэмпл для последующей потоковой обработки. Основная выборка разделялась на тренировочную и тестовую части в соотношении 80/20.

```
df = spark.read.parquet("hdfs://namenode:9000/user/dchel/dchel_database/eCommerce_clear_data")
```

Рисунок 16 – Процесс загрузки данных из HDFS

Далее выполнялась предобработка признаков. Категориальные параметры преобразовывались с помощью StringIndexer. Все признаки объединялись в единый вектор посредством VectorAssembler.

На основе этих этапов формировался конвейер Spark ML, включающий индексацию, кодирование, масштабирование признаков и модель GBT регрессии. Для неё использовались параметры maxIter=100, maxDepth=5, regParam=0.01.

Оптимизация модели выполнялась с помощью 5-кратной кросс-валидации. Наилучшие результаты показала конфигурация с regParam=0.1, maxIter=100 и maxDepth=5, что соответствует комбинированной L1–L2 регуляризации.

2.1.3 Анализ полученных результатов регрессии

Модель была протестирована на отложенной тестовой выборке. Получены следующие значения метрик:

- RMSE = 65.9345;
- $R^2 = 0.7862$.

Высокое значение R^2 свидетельствует о сильной объясняющей способности модели.

2.2 Задача классификации с использованием LogisticRegression

2.2.1 Постановка задачи классификации

Вторая часть работы посвящена построению модели многоклассовой классификации, определяющей бюджетный сегмент (price_range_numeric). Целевой переменной является числовой индикатор бюджетной категории.

Требуется построить классификатор на основе Logistic Regression, предсказывающий бюджетный сегмент по доступным данным. Необходимо проанализировать работу модели на валидационной выборке, определить оптимальный порог принятия решения и представить модель, которая, гарантируя обнаружение не менее 60% всех полезных отзывов (Recall ≥ 0.60), обеспечивает при

этом наивысшую возможную долю верных предсказаний среди всех отмеченных как полезные (Precision).

2.2.2 Решение задачи классификации

В задаче классификации данные также загружались из HDFS. Потом разбивались на выборки (17).

```
# Обновим датафреймы train и test с новыми признаками
train_df, test_df = df.randomSplit([0.8, 0.2], seed=42)
print(train_df.count())
print(test_df.count())
```

```
151676
37625
```

Рисунок 17 – Объём выборок и кол-во экземпляров классов

На этапе предобработки категориальные признаки преобразовывались с помощью StringIndexer, после чего все признаки объединялись в общий вектор, необходимый для обучения модели. Нормализация числовых признаков проводилась с помощью StandardScaler.

Процесс обучения реализовывался через конвейер, включающий этапы подготовки данных и модель LogisticRegression. Для начального варианта использовались параметры maxIter=100, regParam=0.01, elasticNetParam=0.0 и family=”multinomial”.

Оптимизация гиперпараметров выполнялась с использованием 5-кратной кросс-валидации. Наилучшие результаты были достигнуты при maxIter=100, regParam=0.01, elasticNetParam=1.0.

2.2.3 Анализ полученных результатов классификации

Модель классификации продемонстрировала стабильные результаты: точность (Precision) составила 0.97, полнота (Recall) — 0.969, F1-мера — 0.968, а общая точность классификации достигла 0.969.

Матрица ошибок для выбранного порога показывает следующие значения: более 12 000 объектов были корректно классифицированы, около 2 000 — некорректно (см. рис. 18).

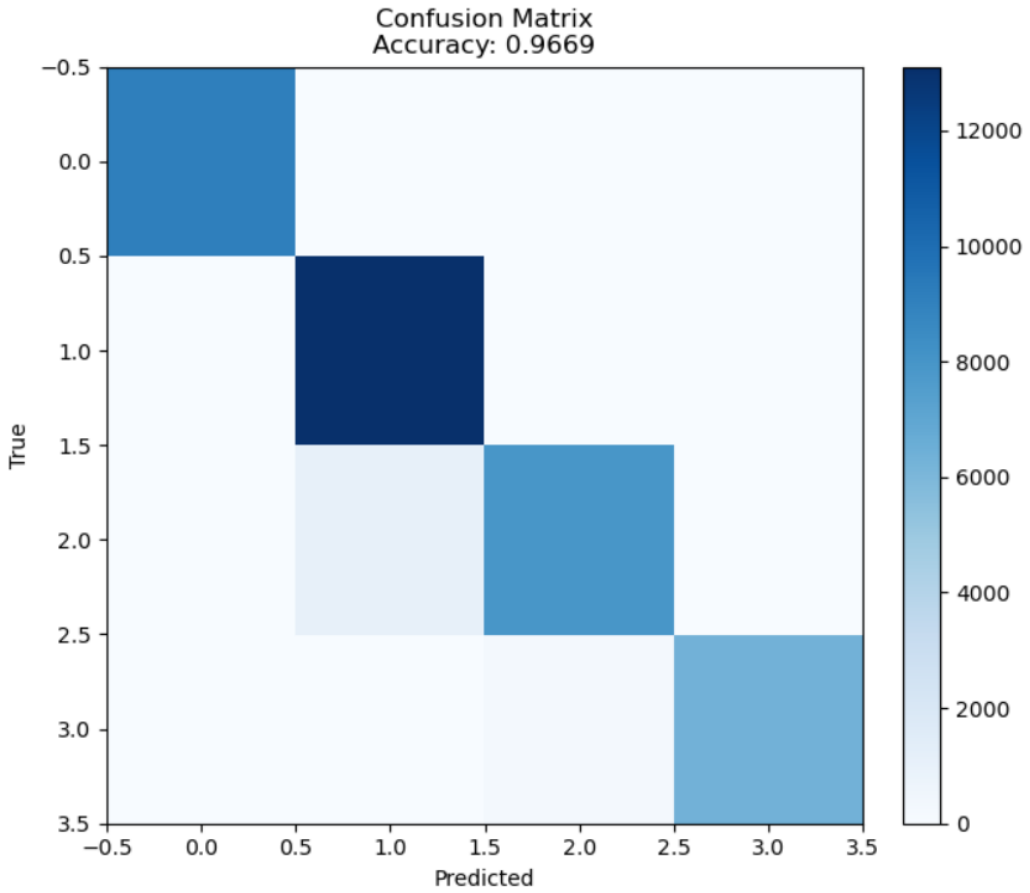


Рисунок 18 – Матрица ошибок

График распределения кол-ва верно предсказанных классификаций показывает следующее:

Рисунок 19 – Кол-во верно предсказанных классификаций

2.3 Выводы

В рамках работы были решены две задачи машинного обучения. GBT регрессия показала эффективность: значение R^2 (0.78) подтверждает, что признаков достаточно для точного прогнозирования цены продуктов. Задача классификации оказалась более успешной: модель достигла accurancy = 0.96, F1 = 0.96 и выполнила требуемый уровень полноты (Recall ≥ 60%).

В перспективе дальнейшее развитие связано с использованием текстовых признаков [10] (TF-IDF, Word2Vec, BERT), внедрением нейронных сетей и современных ансамблей, а также сокращением размерности категориальных признаков. Полученные результаты демонстрируют эффективность Spark ML при анализе продуктов электронной коммерции в условиях больших данных.

ЗАКЛЮЧЕНИЕ

В ходе выполнения курсовой работы был реализован полный цикл анализа больших данных и машинного обучения на примере датасета «eCommerce behavior data from multi category store» с использованием фреймворка Apache Spark. В рамках исследования проведён разведочный анализ данных, включающий загрузку, очистку и преобразование информации источника. Были обработаны пропущенные значения, созданы новые признаки, проанализированы распределения и выбросы, что обеспечило подготовку качественного набора данных для последующего моделирования.

Для решения поставленных задач были реализованы и протестированы две модели машинного обучения. GBT регрессия для прогнозирования пользовательских оценок показала устойчивую сходимость, высокое значение $R^2 = 0{,}78$. Модель логистической регрессии для классификации бюджетного сегмента продуктов продемонстрировала хорошее качество (accurancy = 0,96) и выполнила поставленное условие: полнота (Recall) не ниже 60% при точности (Precision) 68%.

Сравнительный анализ задач показал, что бинарные признаки, такие как is_expensive или is_budget, эффективнее используются для прогнозирования цены на продукт. Применение распределённых вычислений на платформе Apache Spark подтвердило свою эффективность при работе с большими объёмами данных, обеспечив масштабируемость и высокую производительность на всех этапах проекта.

Перспективы дальнейшего развития работы включают:

- расширение анализируемых пользовательских действий для построения полного воронки конверсии и расчета CTR (Click-Through Rate);
- обогащение данных профилями пользователей с применением RFM-анализа (Recency, Frequency, Monetary) и сегментацией по предпочтениям;

– внедрение временных и сезонных признаков для учета суточных, недельных и праздничных паттернов активности;
– разработку рекомендательных систем на основе контентной фильтрации (content-based) и коллаборативной фильтрации (collaborative filtering);
– анализ путей пользователей (Customer Journey) с применением марковских цепей для моделирования переходов между категориями;
– реализацию динамического ценообразования на основе анализа эластичности спроса и конкурентной среды;
– восстановление и структурирование иерархии категорий товаров для кросс-категорийного анализа;
– прогнозирование спроса с использованием моделей временных рядов (ARIMA/SARIMA, Prophet) и нейронных сетей (LSTM);
– проведение A/B-тестов для оптимизации алгоритмов рекомендаций и персонализации интерфейса;
– детекцию аномалий и мошеннической активности через анализ паттернов поведения пользователей.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Изучаем Spark: молниеносный анализ данных / Карау, Х. [и др.]. — ДМК Пресс, 2015. — 304 с. — ISBN 978-5-97060-274-6.
2. Изучаем Spark: Быстрый анализ данных / Дамджи, Дж. [и др.]. — ДМК Пресс, 2020. — 450 с. — ISBN 978-5-93700-102-8.
3. *Чемберс, Б.*, *Захария, М.* Spark: Полное руководство. — Питер, 2018. — 598 с. — ISBN 978-5-4461-0599-5.
4. *Koirala, Roshan*. Exploratory Data Analysis with pySpark. — 2020. — URL: https://github.com/roshankoirala/pySpark_tutorial/blob/master/Exploratory_data_analysis_with_pySpark.ipynb (visited on 09/19/2022).
5. *The Apache Software Foundation*. Официальный сайт Apache Spark. — 2022. — URL: https://spark.apache.org/ (visited on 09/19/2022).
6. Spark SQL: Relational Data Processing in Spark / Armbrust, Michael [et al.] // Proceedings of the ACM SIGMOD International Conference on Management of Data. — 2015. — С. 1383–1394. — DOI: 10.1145/2723372.2742797.
7. *Уайт, Т.* Hadoop: Подробное руководство. — 3-е изд. — Питер, 2013. — 672 с. — ISBN 978-5-496-00299-9.
8. *Kaggle*. eCommerce behavior data from multi category store. — 2019. — URL: https://www.kaggle.com/datasets/mkechinov/ecommerce-behavior-data-from-multi-category-store (visited on 12/16/2025).
9. *Tekdoğan, T.*, *Çakmak, A.* Benchmarking Apache Spark and Hadoop MapReduce on Big Data Classification // 2021 5th International Conference on Cloud and Big Data Computing. — ACM, 2022. — 15–20. — DOI: 10.1145/3481646.3481649.

10. Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics / Zaharia, Matei [et al.] // CIDR 2021. — 2021. — URL: https://www.cidrdb.org/cidr2021/papers/cidr2021_paper17.pdf (visited on 09/12/2024).

ПРИЛОЖЕНИЕ А

Обработка данных

```
def transform_dataframe(data: DataFrame) -> DataFrame:
    df = df.select(
        "event_type", "product_id", "category_id", "category_code"
    )
    data = data.withColumn("category_id", col("category_id").cast(
    data = data.withColumn("product_id", col("product_id").cast("I
    data = data.withColumn("price", col("price").cast("Float"))
    # Преобразуем строку в массив строк
    data = data.withColumn("category_code",
                           split(col("category_code"), r"\.")
                          )
    data = data.dropna(subset=["category_code", "brand"])
    df = df.withColumn("is_expensive", when(col("price") > 200, 1)
    df = df.withColumn("is_budget", when(col("price") < 50, 1).othe
    df = df.withColumn("is_mid_range", when((col("price") >= 50) &
    df = df.withColumn("category_count",
                   col("contains_appliances").cast("int") +
                   col("contains_computers").cast("int") +
                   col("contains_electronics").cast("int") +
                   col("contains_kitchen").cast("int") +
                   col("contains_smartphone").cast("int"))
    df = df.withColumn("is_purchase", when(col("event_type") == "p
    df = df.withColumn("is_view", when(col("event_type") == "view"
    df = df.withColumn("is_cart", when(col("event_type") == "cart"
    df = df.withColumn("price_range",
                   when(col("price") < 50, "budget")
                   .when(col("price") < 150, "affordable")
                   .when(col("price") < 300, "premium")
                   .otherwise("luxury"))

    # Создаем числовое представление для price_range
```

```
    df = df.withColumn("price_range_numeric",
                       when(col("price_range") == "budget", 1)
                       .when(col("price_range") == "affordable", 2)
                       .when(col("price_range") == "premium", 3)
                       .otherwise(4))
    return df
```