

Разработка информационной технологии обработки данных для анализа клиентских предпочтений в электронной коммерции с использованием алгоритмов Big Data

Духнов Константин Александрович

Приамурский государственный университет им. Шолом-Алейхема

Студент

Аннотация

Цель исследования заключалась в разработке информационной технологии для анализа клиентских предпочтений в электронной коммерции. Методика включала применение алгоритмов машинного обучения на основе данных из открытых источников, обработанных с использованием языка Python и соответствующих библиотек. Разработанная технология демонстрирует потенциал для улучшения клиентского опыта и увеличения прибыльности бизнеса.

Ключевые слова: анализ данных, машинное обучение, алгоритмы Big Data

Development of data processing information technology for analyzing customer preferences in e-commerce using Big Data algorithms

Dukhnov Konstantin Alexandrovich

Sholom-Aleichem Priamursky State University

Student

Abstract

The aim of the research was to develop an information technology for analyzing customer preferences in e-commerce. The methodology included the application of machine learning algorithms based on open-source data processed using the Python language and relevant libraries. The developed technology demonstrates the potential to improve customer experience and increase business profitability.

Keywords: data analysis, machine learning, Big Data algorithms

1 Введение

1.1 Актуальность исследования

Развитие электронной коммерции стимулирует активное накопление информации о предпочтениях потребителей. Благодаря развитию современных технологий хранения и обработки данных, стало возможным получение ценных инсайтов, позволяющих компаниям более точно адаптироваться к потребностям рынка. Одним из ключевых методов повышения прибыльности является анализ предпочтений потребителей с целью подбора наиболее релевантных товаров и услуг. Такой подход

способствует повышению уровня удовлетворенности клиентов и общего объёма продаж.

1.2 Обзор исследований

А. Замятин в книге «Интеллектуальный анализ данных» [1] приводит примеры использования анализа данных в различных сферах, включая индустрию развлечений. Один из интересных примеров касается компании, которая анализирует историю пользовательских запросов для предложения индивидуальных рекомендаций. Этот подход позволяет улучшить пользовательский опыт и повысить вовлеченность аудитории. С. Н. Лобанова в статье «Актуальные инструменты продвижения для стриминговых площадок» [2] анализирует современные методы продвижения онлайн-сервисов на российском рынке. В исследовании рассматриваются количественные и статистические данные об активности пользователей и эффективность различных рекламных стратегий. Выделяют баннерную рекламу как основной инструмент продвижения, указывая на ее высокую кликабельность и значимость для привлечения новых пользователей. Также изучаются поведенческие реакции пользователей на различные маркетинговые кампании, что помогает стриминговым сервисам формировать конкурентные преимущества. В исследовании М. И. Бовкун, Е.А Полоско, А.А. Ефремова [3] описывается использование структурированных и неструктурированных данных большого объёма для их дальнейшего использования в решении задач электронной коммерции. Ими замечено, что использование Big Data особенно эффективно в случае качественных наборов данных, которые упорядочены по определённым свойствам. Если такого ранжирования нет, то из этого следует пустая трата времени и ресурсов.

И. А. Валентов в своей статье [4] исследовании отмечает, что первостепенную важность в вопросе анализа данных занимает целеполагание, то есть специалисту Big Data следует точно осознавать, какие выводы он рассчитывает получить. В противном случае, усилия, потраченные на сбор, разметку и анализ данных не имеют смысла, однако, верные выводы способны окупиться в очень короткий срок.

М. Д. Гутниковым рассмотрен [5] потенциал инструментов Big Data для укрепления позиций Авито на рынке электронной коммерции. Были выявлены наиболее актуальные вопросы, связанные с внедрением BD на Авито, проведен опрос, отражающий отношение покупателей и продавцов Авито к возможностям BD.

1.3 Цель исследования

Целью исследования является разработка информационной технологии для анализа данных о клиентских предпочтениях в электронной коммерции.

2 Материалы и методы

Для исследования выбран датасет из платформы Kaggle, отвечающих критериям полноты и актуальности данных. Для реализации информационной

системы использован Python с модулями pandas, scikit-learn, numpy и matplotlib.

3 Результаты и обсуждения

Для анализа используется датасет «Fashion E-commerce User Data» [6] из платформы Kaggle, который содержит информацию о взаимодействии пользователей с платформой электронной коммерции. Датасет включает такие атрибуты, как идентификаторы пользователей, время сеансов, историю просмотров, данные о покупках и категориальную информацию о товарах. Анализ данных из этого набора позволяет выявить поведенческие модели клиентов, определить популярные категории товаров и разработать персонализированные рекомендации. Планируется проверка эффективности ряда алгоритмов, включая линейную регрессию, модели решающего дерева и случайного леса, для прогнозирования покупок и сегментации клиентов.

В начале исследования создана модель машинного обучения для прогнозирования вероятности совершения покупок пользователями электронной коммерции. Основное внимание уделено анализу данных о поведении пользователей и их характеристиках, предоставленных в датасете.

Код состоит из нескольких последовательных этапов, начиная с предобработки данных и заканчивая обучением модели (Рис. 1).

```
7 # Загрузка данных
8 data = pd.read_csv('ds1.csv')
9
10 # Предобработка данных
11 # Выбор релевантных признаков для прогнозирования вероятности покупки
12 features = ['socialNbFollowers', 'socialNbFollows', 'socialProductsLiked',
13            'productsListed', 'productsSold', 'productsPassRate',
14            'productsWished', 'gender', 'hasAnyApp', 'daysSinceLastLogin',
15            'seniorityAsMonths']
16
17 # Целевая переменная: совершал ли пользователь покупки
18 # Создание бинарной целевой переменной
19
20 data['bought_something'] = data['productsBought'] > 0
21
22 # Кодирование категориальных переменных
23 categorical_features = ['gender', 'hasAnyApp']
24 label_encoders = {}
25
26 for feature in categorical_features:
27     le = LabelEncoder()
28     data[feature] = le.fit_transform(data[feature].astype(str))
29     label_encoders[feature] = le
30
31 # Заполнение пропущенных значений
32 data.fillna(0, inplace=True)
33
```

Рисунок 1 – Импорт и обработка данных

Для обработки категориальных данных применялись методы кодирования с использованием LabelEncoder. Числовые данные стандартизировались с помощью StandardScaler, что позволяет улучшить

производительность модели. Далее данные разделялись на обучающую и тестовую выборки.

На первом этапе из датасета были выбраны релевантные признаки, такие как количество подписчиков и подписок в социальных сетях, количество понравившихся продуктов, а также показатели активности пользователей, например, время с момента последнего входа и стаж использования платформы. Бинарная целевая переменная была определена на основе факта совершения хотя бы одной покупки (рис. 2).

```
38 # Разделение данных
39 X = scaled_features
40 y = data['bought_something']
41 X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
42
43 # Обучение модели
44 model = RandomForestClassifier(random_state=42, n_estimators=100)
45 model.fit(X_train, y_train)
46
47 # Прогнозирование
48 y_pred = model.predict(X_test)
49 y_proba = model.predict_proba(X_test)[:, 1]
50
51 # Оценка модели
52 print("Отчёт классификации:\n", classification_report(y_test, y_pred))
53 print("ROC AUC Score:", roc_auc_score(y_test, y_proba))
54
55 # Объяснение результатов
56 # Важность признаков
57 importances = model.feature_importances_
58 feature_importance_df = pd.DataFrame({'Feature': features, 'Importance': importances})
59 feature_importance_df = feature_importance_df.sort_values(by='Importance', ascending=False)
60 print("\nВажность признаков:\n", feature_importance_df)
61
62 # Сохранение модели и кодировщиков
63 import joblib
64 joblib.dump(model, 'purchase_prediction_model.pkl')
65 joblib.dump(scaler, 'scaler.pkl')
66 joblib.dump(label_encoders, 'label_encoders.pkl')
```

Рисунок 2 – Обработка данных и сохранение модели

На этапе обучения использовался классификатор Random Forest, который представляет собой ансамблевый метод, объединяющий результаты множества деревьев решений. После обучения модель была протестирована на тестовой выборке, а для оценки её качества использовались метрики, такие как precision, recall, f1-score и ROC AUC.

Результаты показывают высокую точность модели (accuracy – 95%), что свидетельствует о её способности различать пользователей, которые совершают покупки, и тех, кто этого не делает. Однако при более детальном анализе можно заметить значительную диспропорцию между классами. Модель демонстрирует высокую точность для пользователей, не совершавших покупок (precision – 0.96, recall – 0.99), но менее эффективна для постоянных покупателей (precision – 0.54, recall – 0.29). Это указывает на необходимость дальнейшей балансировки классов или применения других методов для повышения качества классификации меньшинств.

ROC AUC Score составляет 0.84, что свидетельствует о хорошей способности модели различать классы в общем случае. Анализ важности

признаков показал, что наибольшее влияние на предсказания оказывают такие признаки, как время с момента последнего входа (`daysSinceLastLogin`), стаж использования платформы (`seniorityAsMonths`) и количество добавленных в «избранное» продуктов (`productsWished`). Это говорит о важности активного участия пользователей в платформе и их заинтересованности в продуктах.

Разработанная модель позволяет компаниям электронной коммерции предсказывать вероятность совершения покупок конкретными пользователями. Это может быть использовано для оптимизации маркетинговых стратегий, например, для таргетинга рекламных кампаний, предоставления индивидуальных скидок или рекомендаций. Анализ важности признаков помогает бизнесу лучше понимать поведение клиентов и выделять наиболее значимые аспекты их активности. В результате компания может не только увеличить продажи, но и повысить уровень удовлетворённости пользователей, предлагая им персонализированные решения.

В рамках работы написан ещё один скрипт (рис. 3). Он выполняет анализ клиентских предпочтений в электронной коммерции на основе предоставленного набора данных, включающего разнообразные метрики взаимодействия пользователей с платформой. Основная цель – выявить ключевые факторы, влияющие на предпочтения пользователей, и определить закономерности их поведения.

```
5 # Загрузка набора данных
6 data = pd.read_csv('ds1.csv')
7
8 # Вывод основной информации о наборе данных
9 print("Информация о наборе данных:")
10 data.info()
11 print("\nПервые строки набора данных:")
12 print(data.head())
13
14 # Выбор релевантных столбцов для анализа предпочтений
15 relevant_columns = [
16     'socialNbFollowers', 'socialNbFollows', 'socialProductsLiked',
17     'productsListed', 'productsSold', 'productsPassRate',
18     'productsWished', 'productsBought', 'gender',
19     'hasAnyApp', 'hasAndroidApp', 'hasIosApp',
20     'daysSinceLastLogin', 'seniorityAsMonths'
21 ]
22
23 # Создание поднабора данных для анализа
24 subset = data[relevant_columns]
25
26 # Проверка на наличие пропущенных значений
27 print("\nПропущенные значения:")
28 print(subset.isnull().sum())
29
30 # Обработка пропущенных значений (если есть)
31 subset = subset.dropna()
32
33 # Преобразование категориальных столбцов в числовые значения для корреляционного анализа
34 subset['gender'] = subset['gender'].astype('category').cat.codes
```

Рисунок 3 – Импорт и обработка данных

Скрипт начинается с загрузки и первичной обработки набора данных. На этапе подготовки производится выбор наиболее релевантных столбцов, отражающих социальное взаимодействие пользователей (подписчики,

подписки, понравившиеся товары), показатели активности (список товаров, продажи, успешность продаж), а также демографические данные (пол, наличие приложений, время с последнего входа).

Для выполнения корреляционного анализа и визуализации взаимосвязей между показателями категории переводятся в числовые значения. Например, пол пользователя кодируется числовыми идентификаторами. Далее вычисляется корреляционная матрица, которая позволяет определить, какие факторы имеют статистически значимые взаимосвязи. Для визуализации взаимосвязей используются графики, такие как scatterplot, иллюстрирующие зависимость между количеством понравившихся продуктов и числом подписчиков, а также между желанными и купленными товарами (рис. 4).

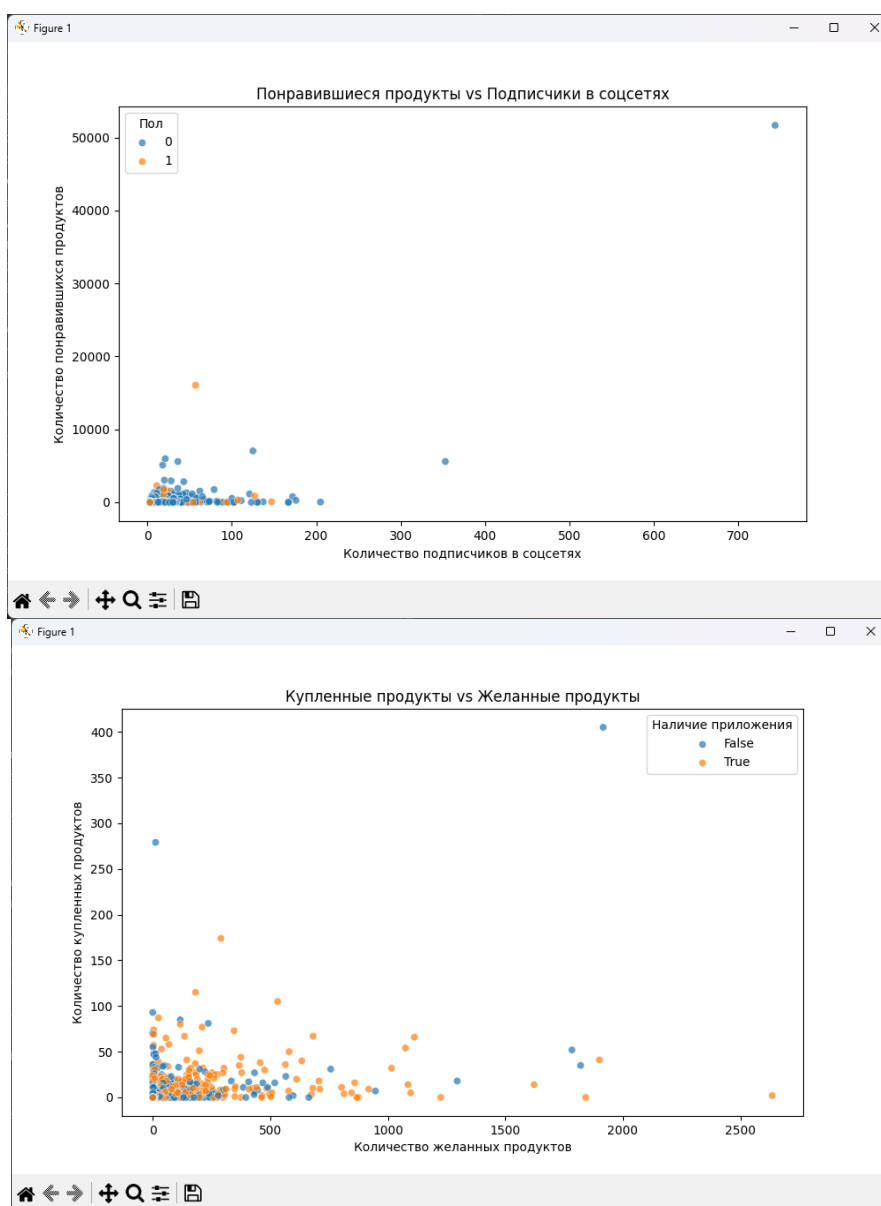


Рисунок 4 – Визуализация результатов

Завершающим этапом является проверка гипотез. В данном случае проведён Т-тест для сравнения двух групп пользователей: с наличием

приложения и без него. Результаты теста показывают статистически значимые различия в количестве купленных продуктов между этими группами.

Матрица корреляций выявляет наиболее сильные взаимосвязи между различными факторами. Например, значительная корреляция между количеством понравившихся продуктов и числом подписчиков в социальных сетях свидетельствует о том, что пользователи с большой аудиторией чаще выражают свои предпочтения.

Графики наглядно демонстрируют распределение данных и подтверждают гипотезу о взаимосвязи между активностью пользователя и его предпочтениями. Например, пользователи, которые отмечают больше желаемых продуктов, как правило, покупают больше товаров.

Результаты Т-теста подтверждают, что наличие мобильного приложения оказывает значительное влияние на покупательскую активность. Пользователи, использующие приложения, совершают больше покупок, что может свидетельствовать о повышении удобства использования платформы.

Данный анализ позволяет компаниям электронной коммерции глубже понять поведение пользователей и разработать стратегии, направленные на увеличение вовлечённости и повышения уровня продаж. Например, установив связь между наличием приложений и количеством покупок, можно сосредоточить усилия на улучшении функционала мобильных приложений. Анализ предпочтений помогает оптимизировать рекомендательные алгоритмы, что, в свою очередь, способствует удовлетворению потребностей клиентов и повышению их лояльности.

Весь код проекта, модели и инфографика опубликованы в GitHub репозитории [7].

4 Выводы

Разработанная технология позволила выделить инсайты, способствующие повышению конверсии или удовлетворенности клиентов.

Библиографический список

1. Замятин А. Интеллектуальный анализ данных. – Litres, 2022.
2. Лобанова С.Н. Актуальные инструменты продвижения для стриминговых площадок // Инновационная экономика: информация, аналитика, прогнозы. 2023. №2.
3. Бовкун М. И., Полоско Е. И., Ефремов А. А. Технологии BIG DATA в электронной коммерции // Наука и образование: актуальные вопросы, достижения и инновации. Пенза, 2022. С. 24.
4. Валентов И. А. Big Data в электронной коммерции // Современные технологии поддержки принятия решений в экономике: сборник трудов Всероссийской научно-практической конференции студентов, аспирантов и молодых ученых, 28-20 апреля 2014 г., г. Юрга. Томский политехнический университет, 2014. С. 215-216.

5. Гутников М. Д., Петров Д. С. Потенциал big data в E-commerce на примере компании Авито // Международный студенческий научный вестник. 2018. №. 6. С. 118-118.
6. Fashion E-commerce User Data URL:
<https://www.kaggle.com/datasets/utkarshx27/fashion-ecommerce-user-data>
(дата обращения: 26.12.2024)
7. customer_preference_analysis. URL:
https://github.com/kostyaaa20/customer_preference_analysis (дата обращения: 26.12.2024).