Content Based Product Recommendation System

- **Objective**:

  The objective of this project is to develop a recommendation system that enhances user experience by suggesting relevant products based on historical interactions with products and available product data. The system aims to leverage machine learning techniques to analyse customer behaviour and provide personalised recommendations.

  The 2nd milestone focuses on implementing approaches like TF-IDF and cosine similarity approach and machine learning algorithms like XGBoost and Random Forest to provide personalized recommendations to users. The report focuses on the feature engineering, feature selection, training process and the output interpretation carried out for each approach.

- **Type of tool:**

  A recommendation engine will be developed based on the user behaviour, product interactions and pricing trends.

- **Tech Stack**

  Programming Language: Python

  Libraries & Frameworks: Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn

  Visualization Tools: Matplotlib, Seaborn

- **Data Collection:**

  The data used is open-source datasets available on Kaggle. The data describes the user's interactions on an ecommerce website/application, providing information regarding the time spent on the website, the products viewed/bought, price, name of the product, etc through the following attributes:

  event_time (timestamp of interaction)

  event_type (type of user interaction)

  product_id (unique identifier for products)

  category_id (unique category identifier)

  category_code (product classification)

  brand (product brand)

  price (product price)

  user_id (unique identifier for users)

user_session (unique identifier for user sessions)

Three csv files are used, 2 csv files contain information related to user's interactions with an electronic store/website and the other csv file contains information related to users' interactions with a multi category store.

- **Project Timeline**

Milestone 1: Data Collection, Preprocessing, and EDA (Feb 5, 2025 - Feb 21, 2025)

Week 1 (Feb 5 - Feb 11): Identify and acquire dataset, verify accessibility, and document dataset details.

Week 2 (Feb 12 - Feb 18): Handle missing values, address outliers, and apply feature scaling.

Week 3 (Feb 19 - Feb 21): Perform exploratory data analysis (EDA), generate visualizations, and identify patterns.

Milestone 2: Feature Engineering, Feature Selection, and Data Modeling (Feb 21, 2025 - March 21, 2025)

Week 4 (Feb 22 - Feb 28)**:** Engineer new features and encode categorical variables.

Week 5 (March 1 - March 7): Perform feature selection and dimensionality reduction.

Week 6 (March 8 - March 14): Split data into training and testing sets, train initial models.

Week 7 (March 15 - March 21): Tune hyperparameters and evaluate models using performance metrics.

Milestone 3: Evaluation, Interpretation, Tool Development, and Presentation (March 24, 2025 - April 23, 2025)

Week 8 (March 24 - March 30): Assess model performance and interpret results.

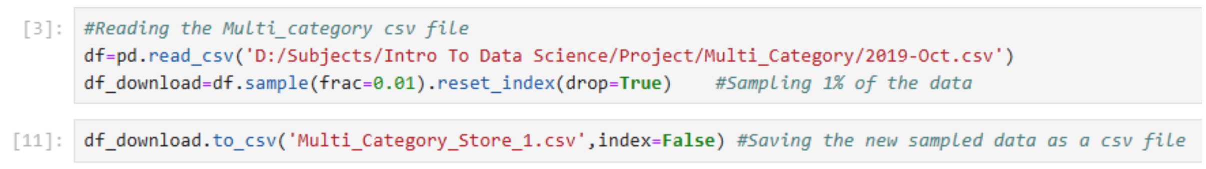
Week 9 (April 1 - April 7): Identify biases and refine models if needed.

Week 10 (April 8 - April 14): Develop an interactive tool (dashboard, conversational agent, or reporting system).

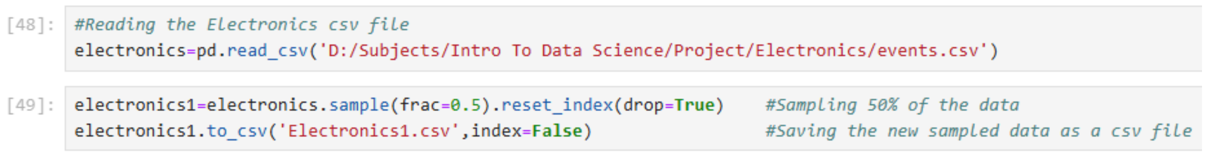
Week 11 (April 15 - April 23): Finalize report, prepare presentation, and deliver findings.

- **Data Preprocessing:**

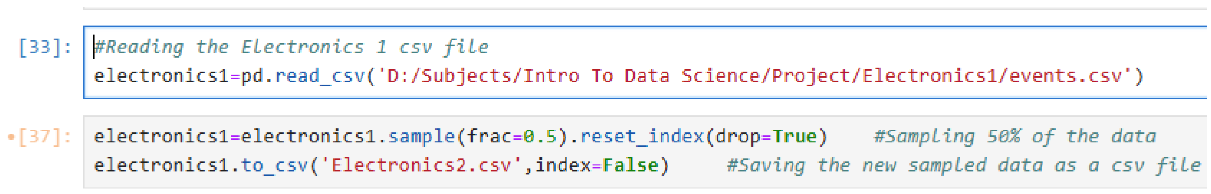
  Resizing the datasets to a smaller size for memory efficiency:



```
[3]:  #Reading the Multi_category csv file
      df=pd.read_csv('D:/Subjects/Intro To Data Science/Project/Multi_Category/2019-Oct.csv')
      df_download=df.sample(frac=0.01).reset_index(drop=True)    #Sampling 1% of the data

[11]: df_download.to_csv('Multi_Category_Store_1.csv',index=False) #Saving the new sampled data as a csv file
```



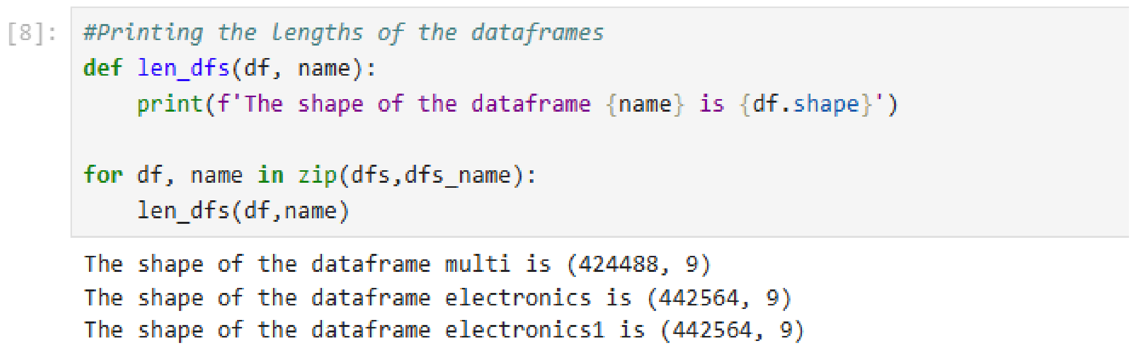
```
[48]: #Reading the Electronics csv file
      electronics=pd.read_csv('D:/Subjects/Intro To Data Science/Project/Electronics/events.csv')

[49]: electronics1=electronics.sample(frac=0.5).reset_index(drop=True)    #Sampling 50% of the data
      electronics1.to_csv('Electronics1.csv',index=False)                 #Saving the new sampled data as a csv file
```



```
[33]: #Reading the Electronics 1 csv file
      electronics1=pd.read_csv('D:/Subjects/Intro To Data Science/Project/Electronics1/events.csv')

•[37]: electronics1=electronics1.sample(frac=0.5).reset_index(drop=True)    #Sampling 50% of the data
       electronics1.to_csv('Electronics2.csv',index=False)      #Saving the new sampled data as a csv file
```

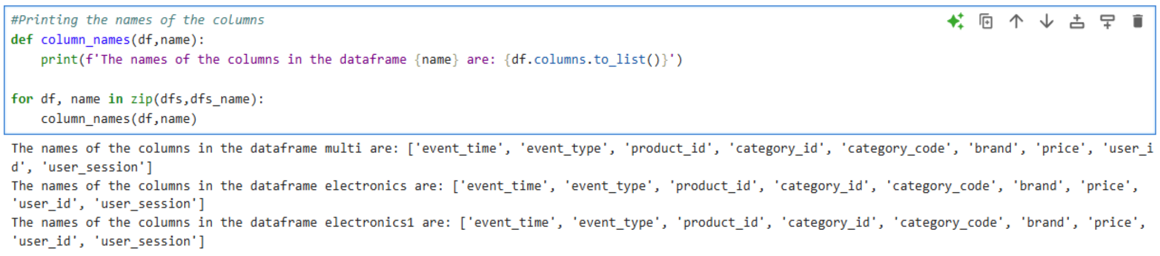
- **Statistical Analysis:**

  a. Printing the lengths of the dataframes:



```
[8]: #Printing the lengths of the dataframes
     def len_dfs(df, name):
         print(f'The shape of the dataframe {name} is {df.shape}')

     for df, name in zip(dfs,dfs_name):
         len_dfs(df,name)

     The shape of the dataframe multi is (424488, 9)
     The shape of the dataframe electronics is (442564, 9)
     The shape of the dataframe electronics1 is (442564, 9)
```

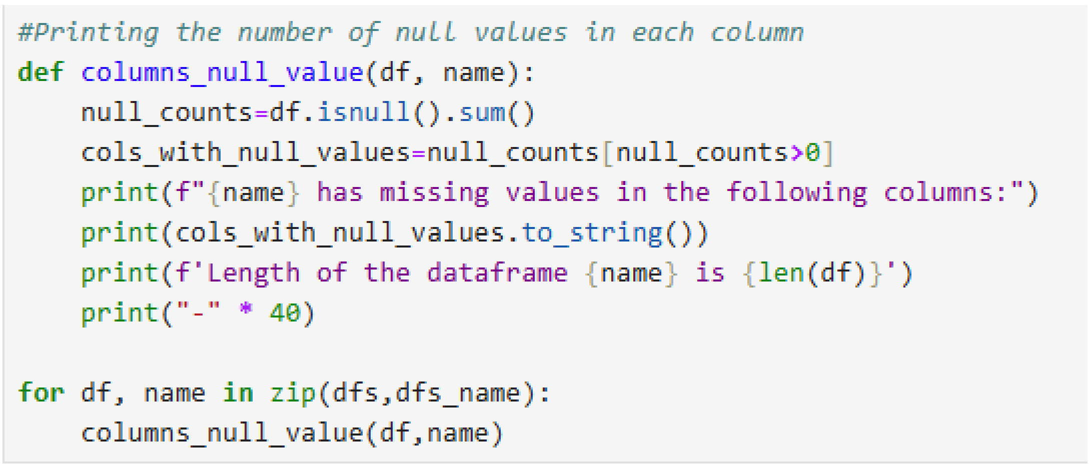
b. Printing the names of the columns in the dataframes:



```
#Printing the names of the columns
def column_names(df,name):
    print(f'The names of the columns in the dataframe {name} are: {df.columns.to_list()}')

for df, name in zip(dfs,dfs_name):
    column_names(df,name)
```

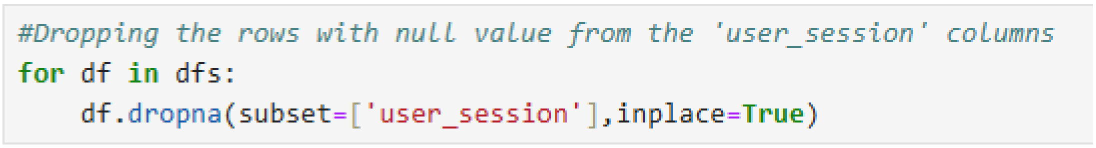
```
The names of the columns in the dataframe multi are: ['event_time', 'event_type', 'product_id', 'category_id', 'category_code', 'brand', 'price', 'user_i
d', 'user_session']
The names of the columns in the dataframe electronics are: ['event_time', 'event_type', 'product_id', 'category_id', 'category_code', 'brand', 'price',
'user_id', 'user_session']
The names of the columns in the dataframe electronics1 are: ['event_time', 'event_type', 'product_id', 'category_id', 'category_code', 'brand', 'price',
'user_id', 'user_session']
```

c. Printing the number of null values in each column:



```
#Printing the number of null values in each column
def columns_null_value(df, name):
    null_counts=df.isnull().sum()
    cols_with_null_values=null_counts[null_counts>0]
    print(f"{name} has missing values in the following columns:")
    print(cols_with_null_values.to_string())
    print(f'Length of the dataframe {name} is {len(df)}')
    print("-" * 40)

for df, name in zip(dfs,dfs_name):
    columns_null_value(df,name)
```

d. Dropping the rows with null values from the ‘user_session’ column:



```
#Dropping the rows with null value from the 'user_session' columns
for df in dfs:
    df.dropna(subset=['user_session'],inplace=True)
```

e. Checking if there are some rows with missing values in 'category_code' or 'brand' have the same 'category_id' as rows without missing values, instead of directly dropping the rows with null values:

```python
#checking if there are some rows with missing values in 'category_code' or 'brand' have the same 'category_id' as rows without missing values.
def identifying_common_ids(df, name):

    missing_rows = df[df['category_code'].isnull() | df['brand'].isnull()]

    non_missing_rows = df.dropna(subset=['category_code', 'brand'])

    matching_category_ids = missing_rows['category_id'].isin(non_missing_rows['category_id'])

    if matching_category_ids.any():
        print("Some rows with missing values in 'category_code' or 'brand' have the same 'category_id' as rows without missing values.")
    else:
        print("No rows with missing values in 'category_code' or 'brand' have a matching 'category_id' in rows without missing values.")

    for col in ['category_code', 'brand']:
        df[col] = df.groupby('category_id')[col].transform(lambda x: x.fillna(x.mode()[0] if not x.mode().empty else np.nan))

    print(len(missing_rows),len(non_missing_rows),len(matching_category_ids),matching_category_ids.value_counts())


    return df
# for df, name in zip(dfs,dfs_name):
#     identifying_common_ids(df,name)
dfs_filled = {name:identifying_common_ids(df, name) for df, name in zip(dfs, dfs_name)}
```

f. Checking if there are some rows with missing values in 'category_code' or 'brand' have the same 'product_id' as rows without missing values, instead of directly dropping the rows with null values:

```python
#checking if there are some rows with missing values in 'category_code' or 'brand' have the same 'product_id' as rows without missing values.
def identifying_common_ids_product(df, name):

    missing_rows = df[df['category_code'].isnull() | df['brand'].isnull()]

    non_missing_rows = df.dropna(subset=['category_code', 'brand'])

    matching_category_ids = missing_rows['product_id'].isin(non_missing_rows['product_id'])

    if matching_category_ids.any():
        print("Some rows with missing values in 'category_code' or 'brand' have the same 'category_id' as rows without missing values.")
    else:
        print("No rows with missing values in 'category_code' or 'brand' have a matching 'category_id' in rows without missing values.")

    for col in ['category_code', 'brand']:
        df[col] = df.groupby('category_id')[col].transform(lambda x: x.fillna(x.mode()[0] if not x.mode().empty else np.nan))

    print(len(matching_category_ids),matching_category_ids.value_counts())


for df, name in zip(dfs,dfs_name):
    identifying_common_ids_product(df,name)
```

g. Dropping the rows with null values that did not have a common ‘category_id’ or ‘product_id’ with the rows that do not have a null value:

```python
#dropping the null values
for name,df in dfs_filled.items():
    print(f'The number of null values in the columns of the dataframe {name} are:')
    print(df.isnull().sum())
    print('Dropping the null values\n')
    df.dropna(subset=['category_code','brand'],inplace=True)
```

h. Merging the dataframes together:

```python
#merging the three dataframes into a single one
merged_df=pd.concat([dfs_filled[dfs_name[0]],dfs_filled[dfs_name[1]],dfs_filled[dfs_name[2]]],ignore_index=True)
print(f'The length of the merged dataframe is {len(merged_df)}')
```

```
The length of the merged dataframe is 937263
```

i. Dropping the duplicate values from the merged dataframe:

```
#dropping the duplicate values
merged_df=merged_df.drop_duplicates(keep='first')
```

j. Performing statistical testing on the numerical columns to determine the distribution of the columns (normal or non-normal):

```
#checking the distribution type in the columns using statistical testing
for col in num_cols:
    stat, p= stats.shapiro(merged_df[col].dropna())
    print(f'{col}:p-value={p}')
    if p>0.05:
        print(f'{col} is normally distributed')
    else:
        print(f'{col} is not normally distributed')
```

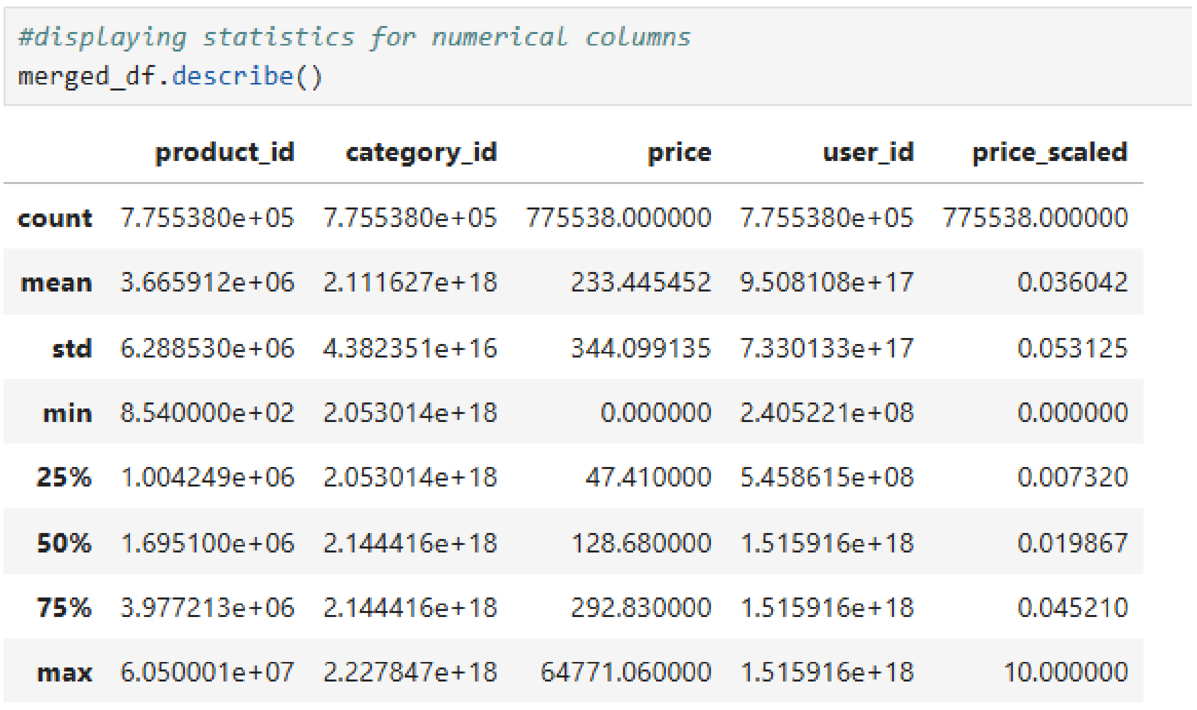
k. Detecting the outliers in the ‘price’ column:

```
#detecting outliers
def detect_outliers(df,col):
    q1=df[col].quantile(0.25)
    q3=df[col].quantile(0.75)
    iqr=q3-q1
    lower_bound=q1-1.5*iqr
    upper_bound=q3+1.5*iqr
    return df[(df[col]<lower_bound) | (df[col]>upper_bound)][col]

outliers=detect_outliers(merged_df,'price')

print(outliers)
```

l. Performing scaling operation on the ‘price’ column to reduce the effect of outliers on the data:

```
scaler=MinMaxScaler(feature_range=(0,10))
merged_df['price_scaled']=scaler.fit_transform(merged_df[['price']])
```

- **Data Visualization:**
  - **Exploratory Data Analysis**



```
#displaying statistics for numerical columns
merged_df.describe()
```

	product_id	category_id	price	user_id	price_scaled
count	7.755380e+05	7.755380e+05	775538.000000	7.755380e+05	775538.000000
mean	3.665912e+06	2.111627e+18	233.445452	9.508108e+17	0.036042
std	6.288530e+06	4.382351e+16	344.099135	7.330133e+17	0.053125
min	8.540000e+02	2.053014e+18	0.000000	2.405221e+08	0.000000
25%	1.004249e+06	2.053014e+18	47.410000	5.458615e+08	0.007320
50%	1.695100e+06	2.144416e+18	128.680000	1.515916e+18	0.019867
75%	3.977213e+06	2.144416e+18	292.830000	1.515916e+18	0.045210
max	6.050001e+07	2.227847e+18	64771.060000	1.515916e+18	10.000000

Fig. 1: Exploratory Data Analysis

  - **Histogram to show the distribution of the 'price' column:**

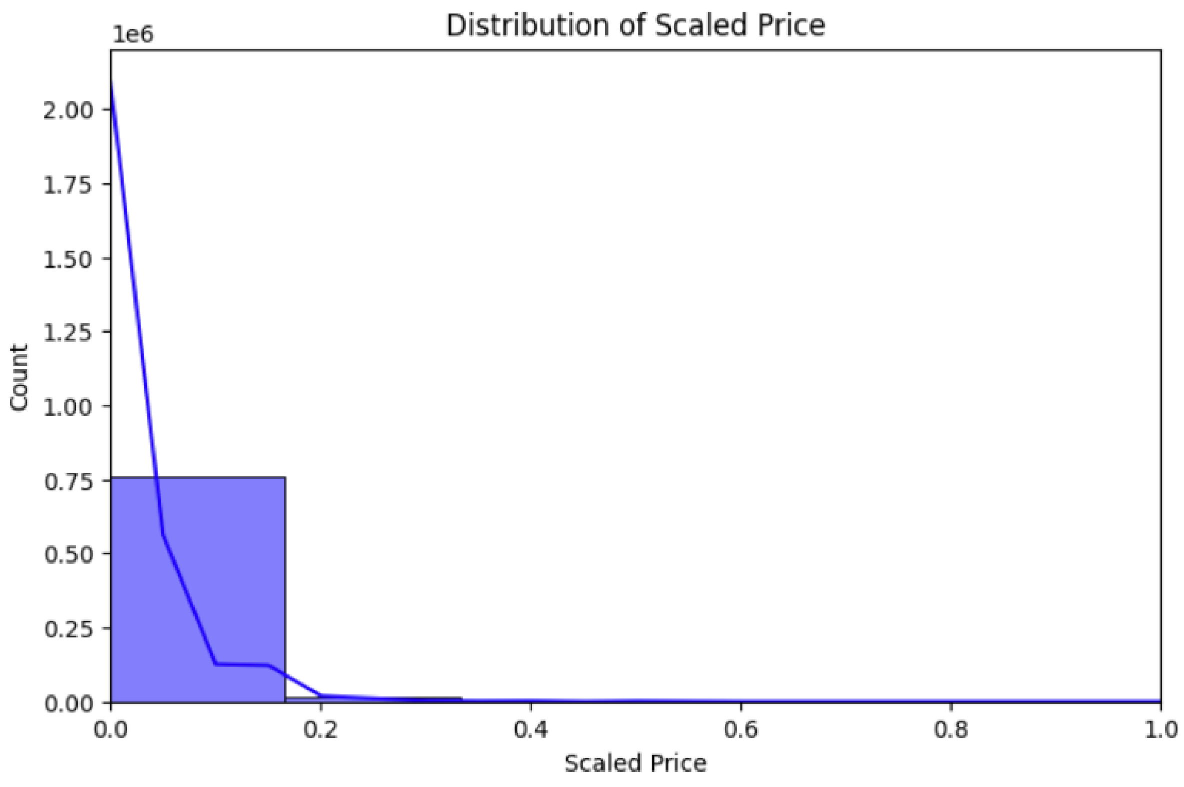


Fig. 2: Distribution of Price

The above graph displays the distribution of the ‘scaled_price’ column, from the graph it is evident that the data follows non-normal distribution which is why IQR method was used to determine outliers in the data.

- **Box plot to detect outliers in the data:**

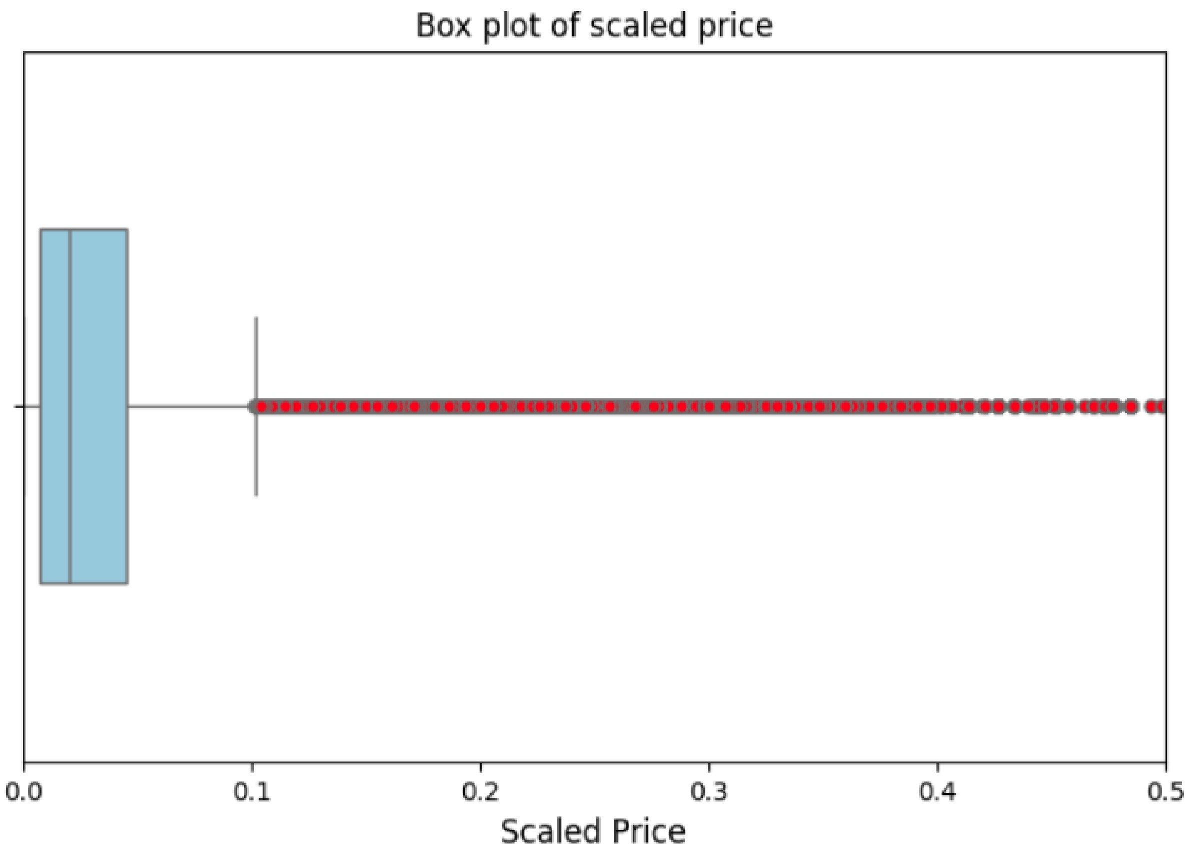


Fig. 3: Boxplot of Price column

The above box plot displays the range of values in the scaled price column, it’s median and the outliers are denoted by the red dots. From this graph it is visible that the ‘price’ column is highly skewed.

- **Distribution of values in the ‘event_type’ column:**

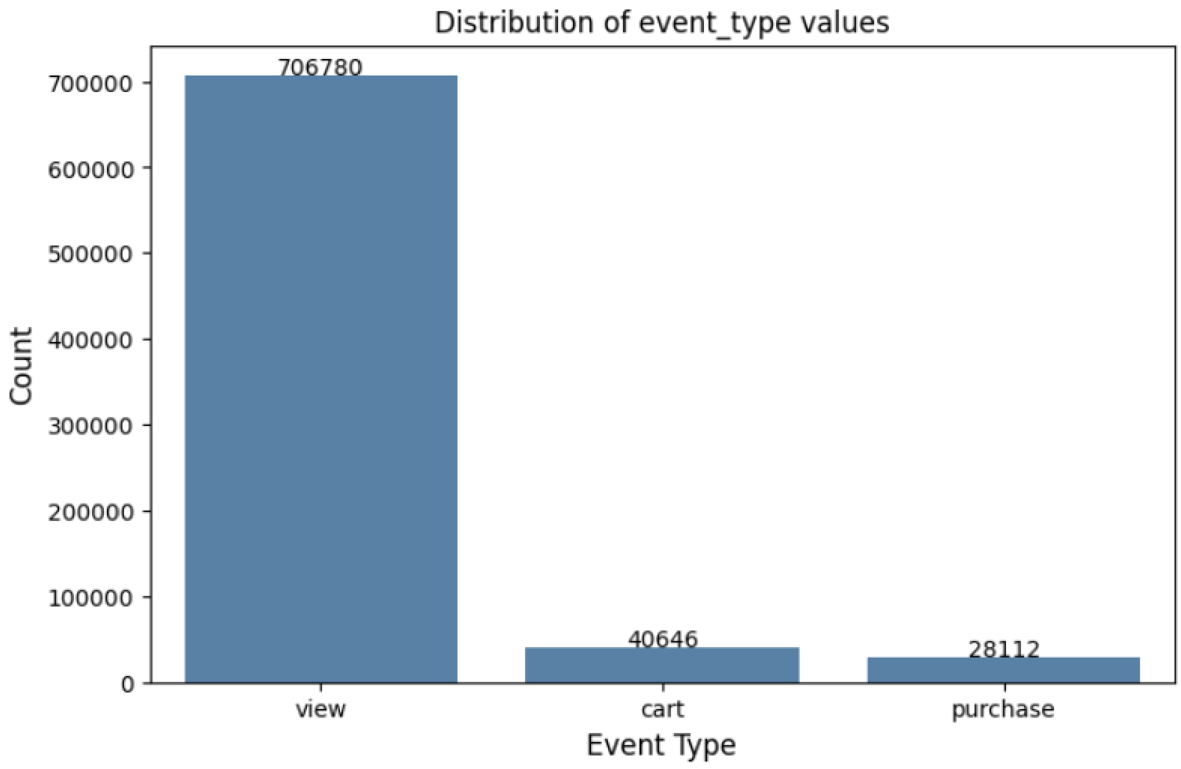


Fig. 4: Distribution of event_type column

The above bar chart describes the distribution of values in the ‘event_type’ column. There are three unique values ‘view’, ‘cart’ and ‘purchase’ which describe the interactions the user had with the product on the ecommerce application/website. It is evident from the graph that the data is highly biased towards the ‘view’ value.

- **Top 10 brands by number of products sold:**

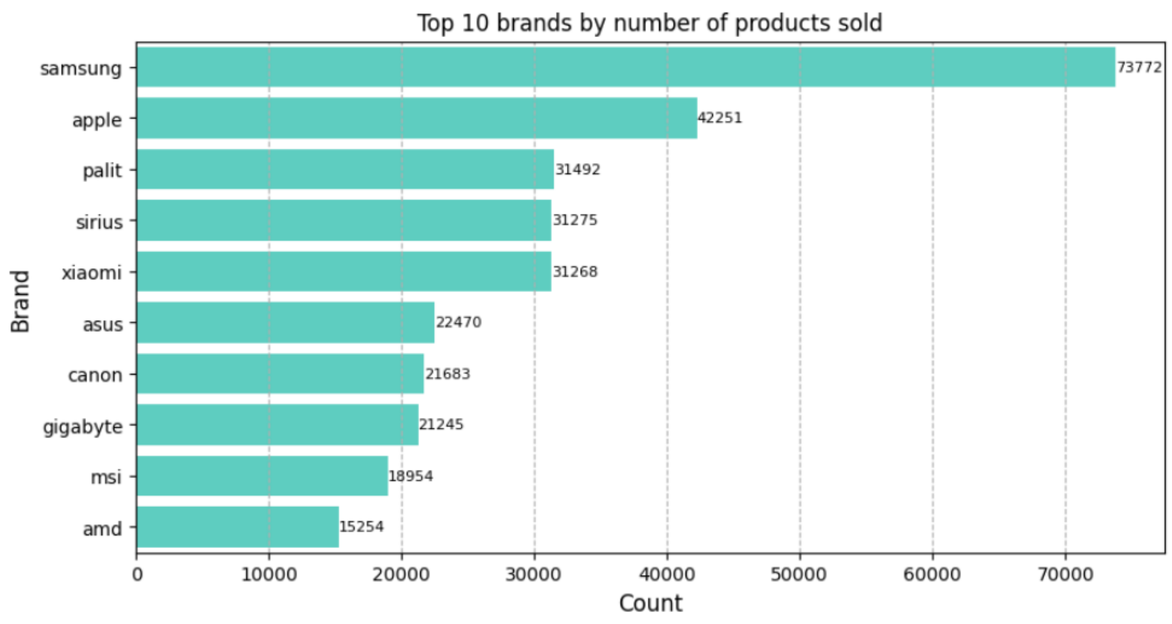


Fig. 5: Top 10 brands by number of products sold

The above horizontal bar chart denotes the top 10 brands whose products were bought/viewed by the users. From the data, we can conclude that the brand ‘samsung’ product were the most bought/viewed by the users, which will help the recommendation system while providing recommendations to the new users.

- **Top 10 products bought by users:**

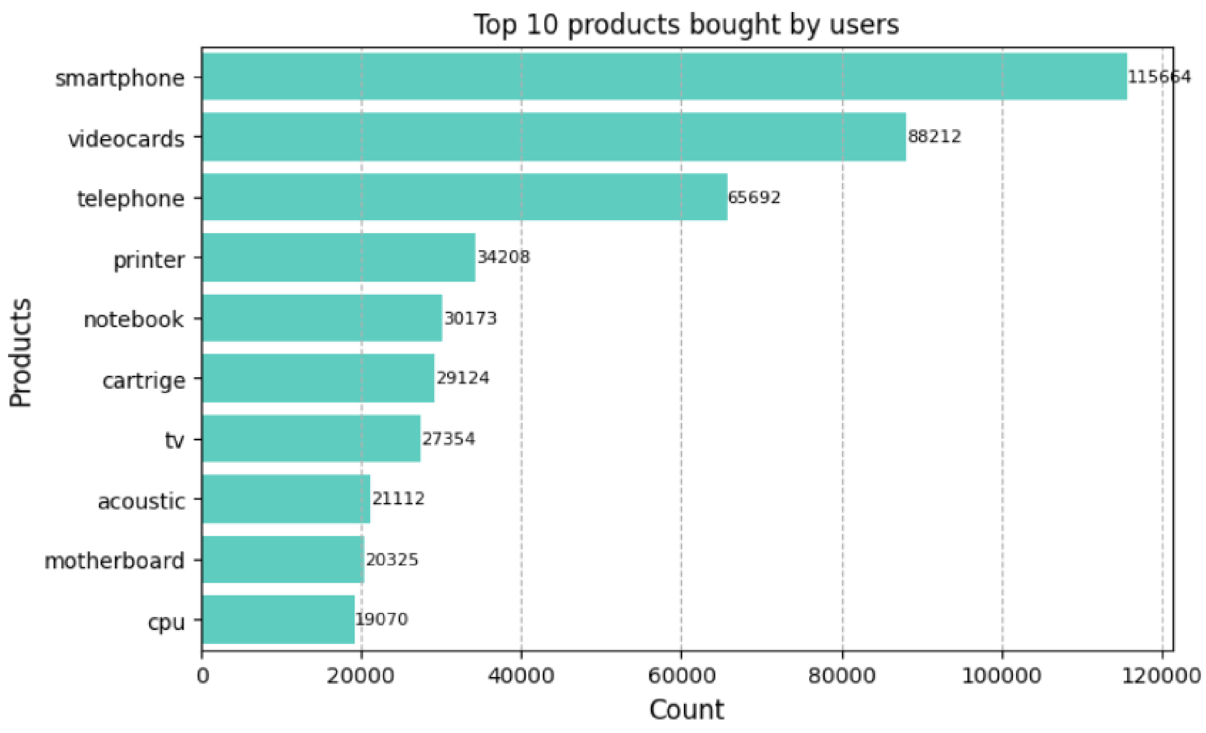


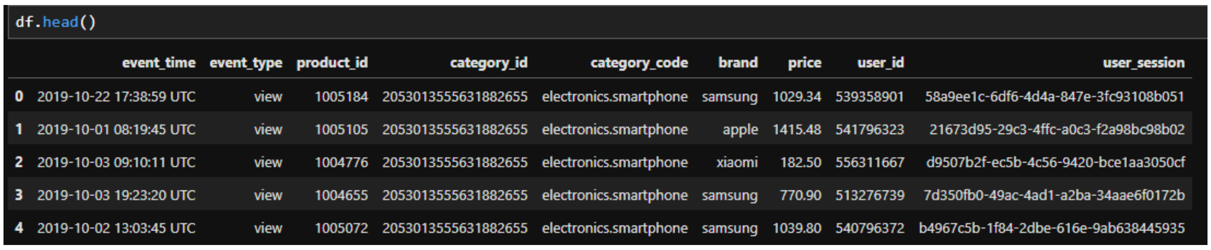
Fig. 6: Top 10 products bought by users

The above horizontal bar chart denotes the top 10 products that users by from the ecommerce website/application.

- **Feature Engineering:**
  Three different approaches have been used to create a recommendation system, TF-IDF vectorizer and Cosine similarity approach, XGBoost approach and the random forest approach. The process of feature engineering slightly differs for these approaches.

A. TF-IDF vectorizer and Cosine similarity:

  a) The 'merged_df' csv has the following columns:



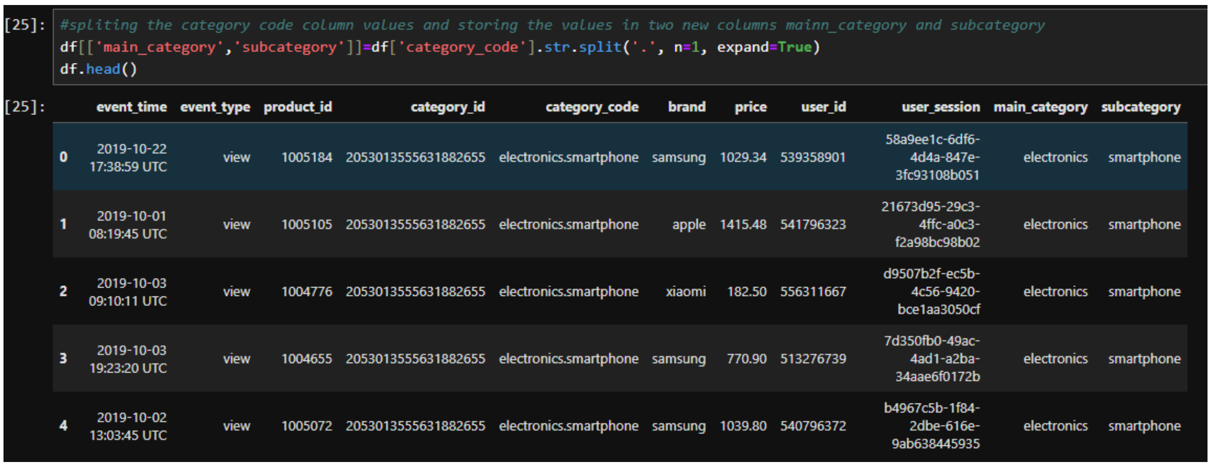
```
df.head()
```

	event_time	event_type	product_id	category_id	category_code	brand	price	user_id	user_session
0	2019-10-22 17:38:59 UTC	view	1005184	2053013555631882655	electronics.smartphone	samsung	1029.34	539358901	58a9ee1c-6df6-4d4a-847e-3fc93108b051
1	2019-10-01 08:19:45 UTC	view	1005105	2053013555631882655	electronics.smartphone	apple	1415.48	541796323	21673d95-29c3-4ffc-a0c3-f2a98bc98b02
2	2019-10-03 09:10:11 UTC	view	1004776	2053013555631882655	electronics.smartphone	xiaomi	182.50	556311667	d9507b2f-ec5b-4c56-9420-bce1aa3050cf
3	2019-10-03 19:23:20 UTC	view	1004655	2053013555631882655	electronics.smartphone	samsung	770.90	513276739	7d350fb0-49ac-4ad1-a2ba-34aae6f0172b
4	2019-10-02 13:03:45 UTC	view	1005072	2053013555631882655	electronics.smartphone	samsung	1039.80	540796372	b4967c5b-1f84-2dbe-616e-9ab638445935

Fig 7. Visualizing the dataframe

  b) Before building the recommendation system the following feature engineering steps were carried out:

    1. Creating 'main_category' and 'sub_category' columns from 'category_code' column:



```
[25]: #spliting the category code column values and storing the values in two new columns mainn_category and subcategory
      df[['main_category','subcategory']]=df['category_code'].str.split('.', n=1, expand=True)
      df.head()
```

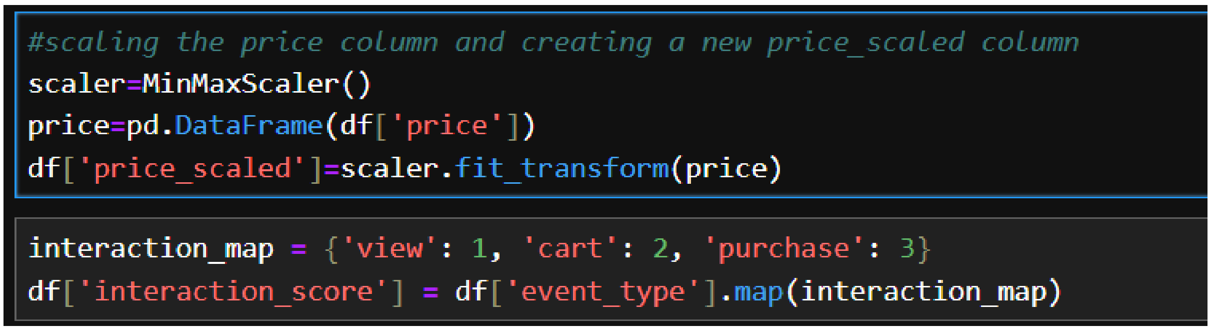
[25]:

	event_time	event_type	product_id	category_id	category_code	brand	price	user_id	user_session	main_category	subcategory
0	2019-10-22 17:38:59 UTC	view	1005184	2053013555631882655	electronics.smartphone	samsung	1029.34	539358901	58a9ee1c-6df6-4d4a-847e-3fc93108b051	electronics	smartphone
1	2019-10-01 08:19:45 UTC	view	1005105	2053013555631882655	electronics.smartphone	apple	1415.48	541796323	21673d95-29c3-4ffc-a0c3-f2a98bc98b02	electronics	smartphone
2	2019-10-03 09:10:11 UTC	view	1004776	2053013555631882655	electronics.smartphone	xiaomi	182.50	556311667	d9507b2f-ec5b-4c56-9420-bce1aa3050cf	electronics	smartphone
3	2019-10-03 19:23:20 UTC	view	1004655	2053013555631882655	electronics.smartphone	samsung	770.90	513276739	7d350fb0-49ac-4ad1-a2ba-34aae6f0172b	electronics	smartphone
4	2019-10-02 13:03:45 UTC	view	1005072	2053013555631882655	electronics.smartphone	samsung	1039.80	540796372	b4967c5b-1f84-2dbe-616e-9ab638445935	electronics	smartphone

Fig 8. Dataframe after splitting the column category_code

    This step will allow to recommend more details about the products like the main_category and subcategory (name of the product). Apart from this, using these two new columns the recommendation system will be more efficient and will be able to provide valid recommendations to the user.

    2. Adding the 'price_scaled' and 'interaction_score' columns:



```
#scaling the price column and creating a new price_scaled column
scaler=MinMaxScaler()
price=pd.DataFrame(df['price'])
df['price_scaled']=scaler.fit_transform(price)
```

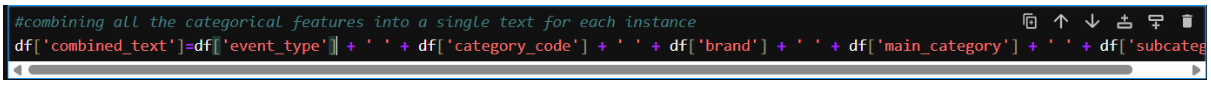
```
interaction_map = {'view': 1, 'cart': 2, 'purchase': 3}
df['interaction_score'] = df['event_type'].map(interaction_map)
```

Fig 9. Scaling and encoding the numerical and categorical columns

The ‘price’ column is scaled down to have a price range from 0 to 1, to decrease the training time and to deal with the outliers in the columns, scaling down these values will help mitigate the variance introduced by the outliers.

Adding interaction_score column quantifies user engagement, mapping event_type to values: 1 for "view," 2 for "cart," 3 for "purchase." This numeric representation reflects interest levels, enabling Random Forest and XGBRanker models to rank products by relevance. Interaction_score acts as target variable during training and ground truth for evaluation metrics like Precision@K and NDCG@K. Converting categorical data to continuous scale enhances personalization and accuracy, ensuring recommendations align with user behavior effectively.

3. Combining the categorical variables:



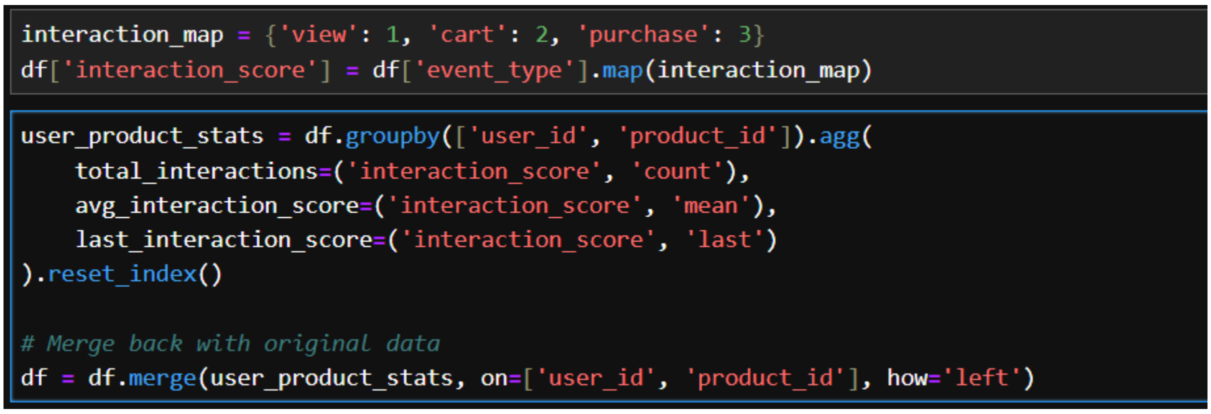
```
#combining all the categorical features into a single text for each instance
df['combined_text']=df['event_type'] + ' ' + df['category_code'] + ' ' + df['brand'] + ' ' + df['main_category'] + ' ' + df['subcateg
```

Fig 10. Combining the categorical columns

This is a crucial step as it transforms all the categorical variables into uniform strings, that enables TF-IDF vectorization for content-based recommendation. It enhances the similarity computation as the contextual relationships between event type and product attributes are captured. This feature engineering supports cosine similarity analysis, aligning product recommendations with user preferences effectively.

B. XGBoost and Random Forest:

a. Aggregating user_product_stats:



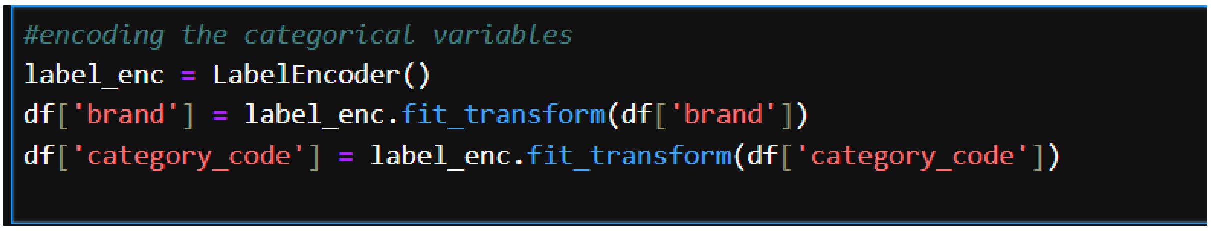
```
interaction_map = {'view': 1, 'cart': 2, 'purchase': 3}
df['interaction_score'] = df['event_type'].map(interaction_map)
```

```
user_product_stats = df.groupby(['user_id', 'product_id']).agg(
    total_interactions=('interaction_score', 'count'),
    avg_interaction_score=('interaction_score', 'mean'),
    last_interaction_score=('interaction_score', 'last')
).reset_index()

# Merge back with original data
df = df.merge(user_product_stats, on=['user_id', 'product_id'], how='left')
```

Fig 11. Feature engineering for XGBoost dataframe

This step involves grouping by user_id and product_id, calculating total_interactions as count, avg_interaction_score as mean, and last_interaction_score as latest value of interaction_score. This step captures user-product interaction patterns numerically and by combining them with behavioral features enhances the model input. Total_interactions reflects frequency, avg_interaction_score indicates average engagement, and last_interaction_score preserves recency.

b. Encoding Categorical variables:



```
#encoding the categorical variables
label_enc = LabelEncoder()
df['brand'] = label_enc.fit_transform(df['brand'])
df['category_code'] = label_enc.fit_transform(df['category_code'])
```

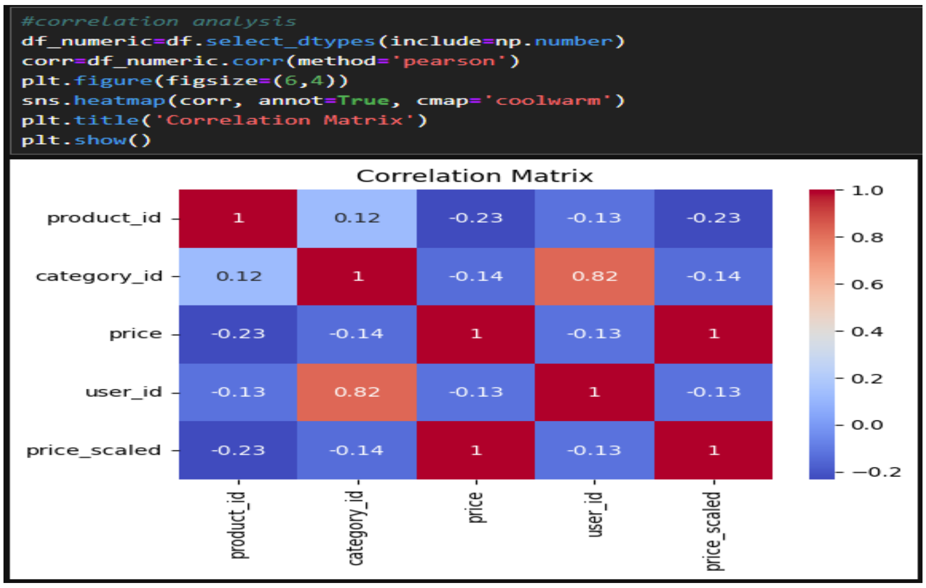
Fig 12. Encoding categorical variables

This is an important step as it involves converting the categorical variables(textual features) into numerical values, enabling machine learning models to process them effectively. The numeric format preserves categorical distinctions without assuming ordinality, facilitating feature use in Random Forest and XGBRanker.

- ## Feature Selection:

Correlation Analysis:

Majority of the columns in the dataframe are text/categorical columns, although checking the correlation among the numerical columns is an important step while building a content based recommendation system as it measures the strength and direction of these relationships.

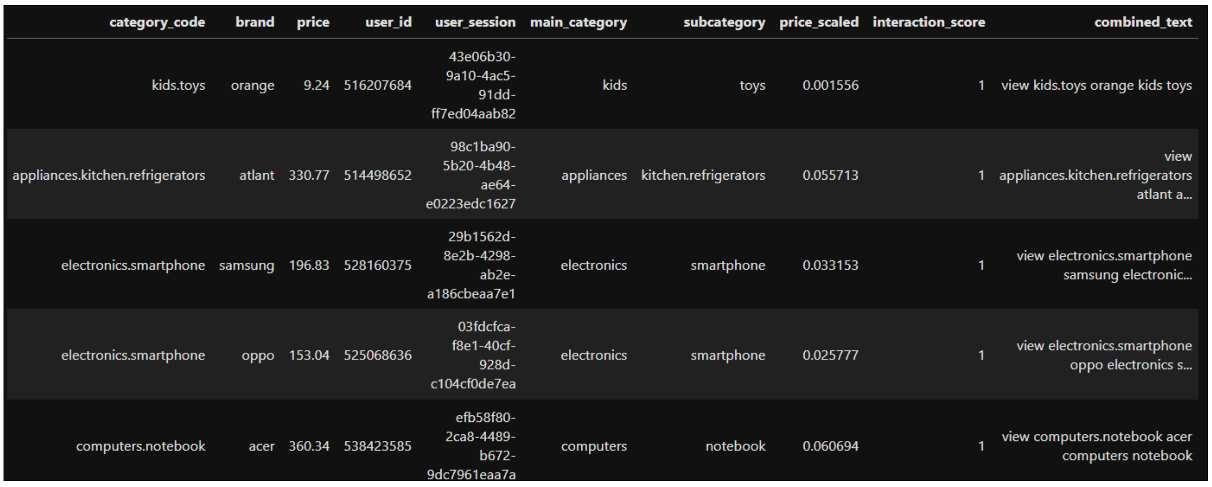


```
#correlation analysis
df_numeric=df.select_dtypes(include=np.number)
corr=df_numeric.corr(method='pearson')
plt.figure(figsize=(6,4))
sns.heatmap(corr, annot=True, cmap='coolwarm')
plt.title('Correlation Matrix')
plt.show()
```

Fig 13. Heatmap for feature selection

From the matrix, we observe that category_id and user_id have a strong positive correlation (0.82), indicating that user preferences may align closely with specific categories. Similarly, price and price_scaled are perfectly correlated, as expected since one is a scaled transformation of the other. Other variables, such as product_id and price, show weak correlations (-0.23), suggesting minimal linear dependence. This analysis is crucial for identifying redundant features and understanding relationships between variables, which can guide feature selection or engineering in building a content-based recommendation system.

Feature selection process differs a bit for TF-IDF approach and the XGBoost and Random Forest approach. For TF-IDF, apart from the above-mentioned features, the column 'combined_text' (mentioned in feature engineering) is considered as it is a combination of the most important features. It summarizes all the textual imformation.



category_code	brand	price	user_id	user_session	main_category	subcategory	price_scaled	interaction_score	combined_text
kids.toys	orange	9.24	516207684	43e06b30-9a10-4ac5-91dd-ff7ed04aab82	kids	toys	0.001556	1	view kids.toys orange kids toys
appliances.kitchen.refrigerators	atlant	330.77	514498652	98c1ba90-5b20-4b48-ae64-e0223edc1627	appliances	kitchen.refrigerators	0.055713	1	view appliances.kitchen.refrigerators atlant a...
electronics.smartphone	samsung	196.83	528160375	29b1562d-8e2b-4298-ab2e-a186cbeaa7e1	electronics	smartphone	0.033153	1	view electronics.smartphone samsung electronic...
electronics.smartphone	oppo	153.04	525068636	03fdcfca-f8e1-40cf-928d-c104cf0de7ea	electronics	smartphone	0.025777	1	view electronics.smartphone oppo electronics s...
computers.notebook	acer	360.34	538423585	efb58f80-2ca8-4489-b672-9dc7961eaa7a	computers	notebook	0.060694	1	view computers.notebook acer computers notebook

Fig 14. Displaying the 'combined_text' column

For the Random Forest and XGBoost approach, it is necessary to consider the features created in the feature engineering process like the total_interactions, avg_interaction_score and last_interaction_score.

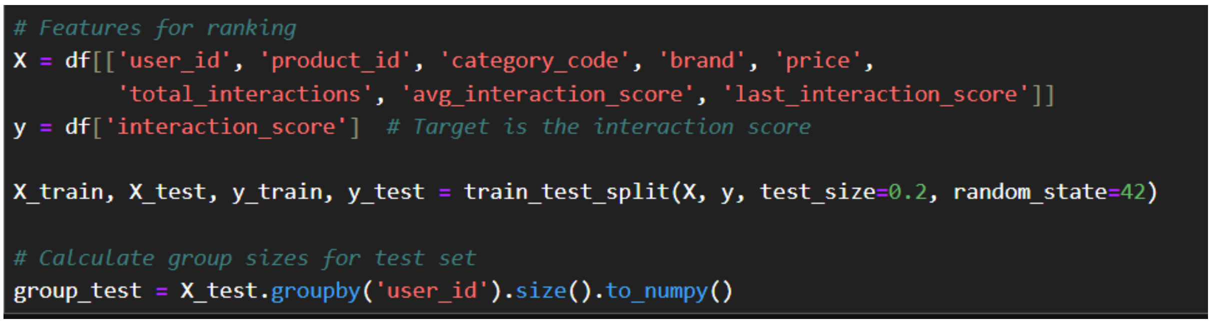
- ## Data Splitting

1. TF-IDF

For the TF-IDF approach there is no splitting of the dataset because of the way the products are recommended. In this approach, the cosine similarity of the products with each other is calculated and on the basis of highest cosine similarity value, the products are recommended. The greater the number of instances, more accurate the cosine similarity value and thus the recommendations are more relevant.

2. XGBoost

Before feeding the data to the XGBoost model, the dataset was split into training and testing sets. The train_test_split class was used provided by sklearn. 20% of the original dataset was reserved for testing.



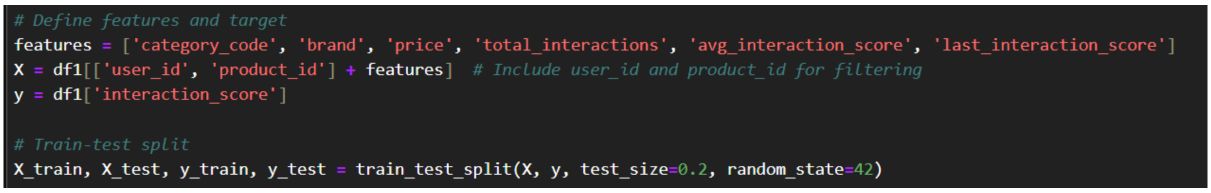
```
# Features for ranking
X = df[['user_id', 'product_id', 'category_code', 'brand', 'price',
        'total_interactions', 'avg_interaction_score', 'last_interaction_score']]
y = df['interaction_score']  # Target is the interaction score

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Calculate group sizes for test set
group_test = X_test.groupby('user_id').size().to_numpy()
```

Fig 15. Splitting the dataset into train and test sets

Computing group_test involves grouping X_test by user_id, calculating the size of each group, and converting results to a NumPy array. This process quantifies interactions per user in the test set, creating a group size vector. It is essential for XGBRanker's pairwise ranking objective, group_test ensures the model recognizes user-specific interaction counts, enabling accurate ranking of products within each user's context during evaluation, supporting effective recommendation generation. The same process is also carried out for the training set, discussed later in the training section.

3. Random Forest

Before feeding the data to the Random Forest model, the dataset was split into training and testing sets. The train_test_split class was used provided by sklearn. 20% of the original dataset was reserved for testing.



```
# Define features and target
features = ['category_code', 'brand', 'price', 'total_interactions', 'avg_interaction_score', 'last_interaction_score']
X = df1[['user_id', 'product_id'] + features]  # Include user_id and product_id for filtering
y = df1['interaction_score']

# Train-test split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

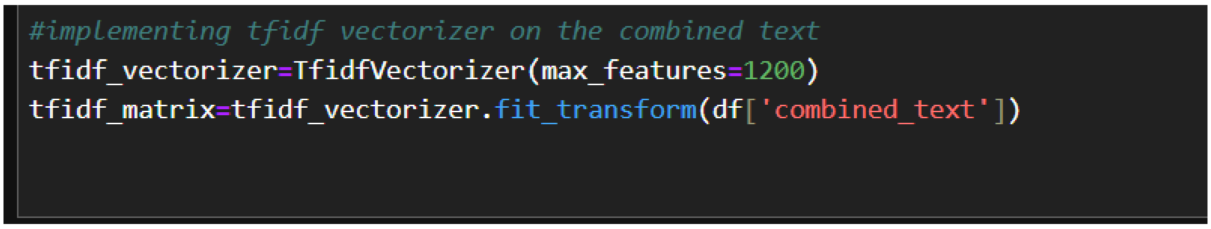
Fig 16. Splitting the dataset into train and test sets for random forest

• Model training

TF-IDF:

The process of training in this approach is a bit different from the traditional machine learning model training. Following steps were to taken to get the product recommendations for users

1. Implementing the TF-IDF vectorizer:

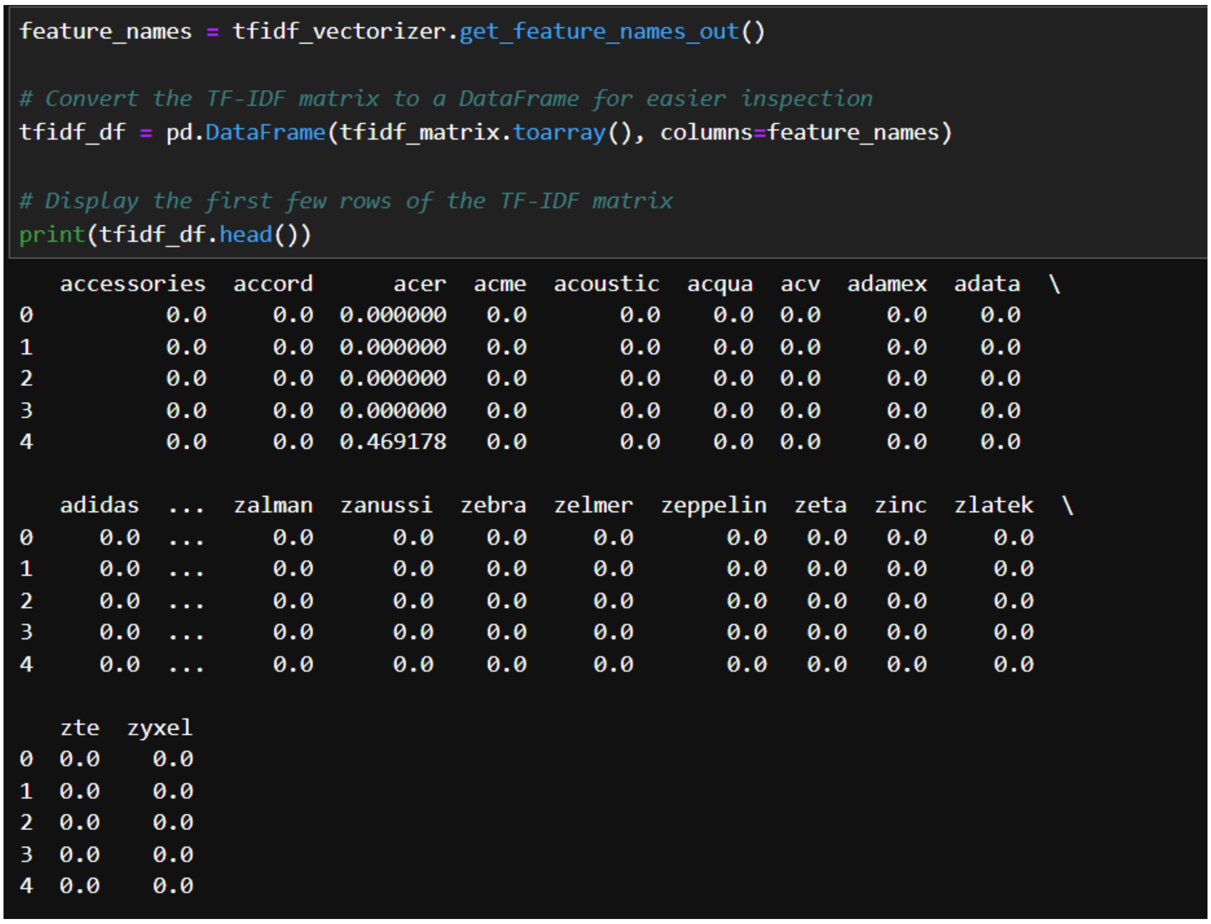


```
#implementing tfidf vectorizer on the combined text
tfidf_vectorizer=TfidfVectorizer(max_features=1200)
tfidf_matrix=tfidf_vectorizer.fit_transform(df['combined_text'])
```

Fig 17. Vectorizing the 'combined_test' column

The TfidfVectorizer is initialized with max_features=1200 (due to computational limitations) restricts vocabulary to 1200 key terms. The fit_transform methos is exeuted on combined_text transforms concatenated categorical data into a TF-IDF matrix, assigning weights based on term frequency and inverse document frequency. This process generates a sparse feature representation, capturing textual importance for training. Limiting features reduces dimensionality, enhancing computational efficiency while retaining critical information for similarity-based recommendation in the TF-IDF model training phase.

2. Extracting the feature names



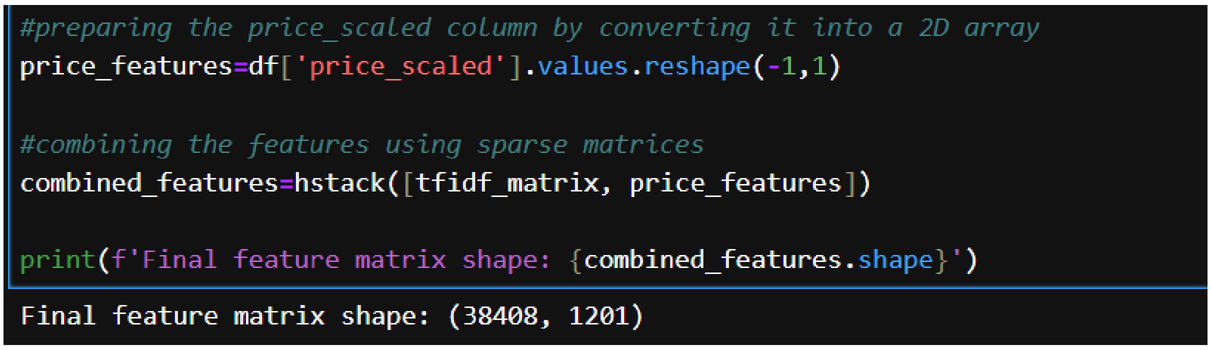
```
feature_names = tfidf_vectorizer.get_feature_names_out()

# Convert the TF-IDF matrix to a DataFrame for easier inspection
tfidf_df = pd.DataFrame(tfidf_matrix.toarray(), columns=feature_names)

# Display the first few rows of the TF-IDF matrix
print(tfidf_df.head())
```

```
   accessories  accord       acer  acme  acoustic  acqua  acv  adamex  adata  \
0          0.0     0.0  0.000000   0.0       0.0    0.0  0.0     0.0    0.0
1          0.0     0.0  0.000000   0.0       0.0    0.0  0.0     0.0    0.0
2          0.0     0.0  0.000000   0.0       0.0    0.0  0.0     0.0    0.0
3          0.0     0.0  0.000000   0.0       0.0    0.0  0.0     0.0    0.0
4          0.0     0.0  0.469178   0.0       0.0    0.0  0.0     0.0    0.0

   adidas  ...  zalman  zanussi  zebra  zelmer  zeppelin  zeta  zinc  zlatek  \
0     0.0  ...     0.0      0.0    0.0     0.0       0.0   0.0   0.0     0.0
1     0.0  ...     0.0      0.0    0.0     0.0       0.0   0.0   0.0     0.0
2     0.0  ...     0.0      0.0    0.0     0.0       0.0   0.0   0.0     0.0
3     0.0  ...     0.0      0.0    0.0     0.0       0.0   0.0   0.0     0.0
4     0.0  ...     0.0      0.0    0.0     0.0       0.0   0.0   0.0     0.0

   zte  zyxel
0  0.0    0.0
1  0.0    0.0
2  0.0    0.0
3  0.0    0.0
4  0.0    0.0
```

Fig 18. Displaying the feature vector

The feature_names from tfidf_vectorizer are extracted using get_feature_names_out(). The tfidf_matrix is then converted to a dense array and creating tfidf_df with feature_names as columns facilitates inspection. Displaying tfidf_df.head() shows the first few rows, revealing TF-IDF weights for each term per instance. This step aids in understanding feature representation, verifying vectorization output before training the TF-IDF recommendation model.

3. Combining tf-idf matrix with the numerical columns



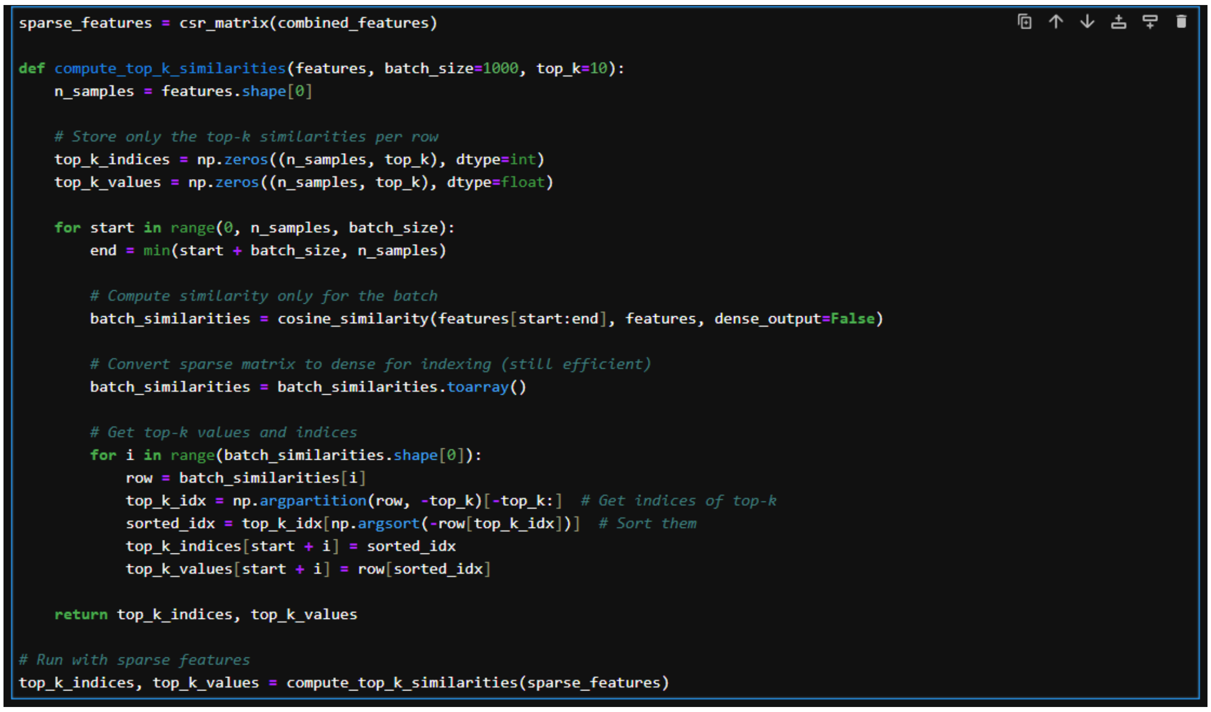
```
#preparing the price_scaled column by converting it into a 2D array
price_features=df['price_scaled'].values.reshape(-1,1)

#combining the features using sparse matrices
combined_features=hstack([tfidf_matrix, price_features])

print(f'Final feature matrix shape: {combined_features.shape}')
```

```
Final feature matrix shape: (38408, 1201)
```

Fig 19. Stacking feature vector with price_scaled (numerical feature)

Combining tfidf_matrix with price_scaled enhances feature set for TF-IDF training. Converting price_scaled to a 2D array via reshape(-1,1) ensures compatibility. Using hstack merges sparse tfidf_matrix with numerical price_features, creating combined_features. Printing combined_features.shape confirms the resulting matrix dimensions, integrating textual and price data. This step enriches the model with both categorical and numerical insights, improving recommendation quality in the TF-IDF training phase.

4. Computing the cosine similarity

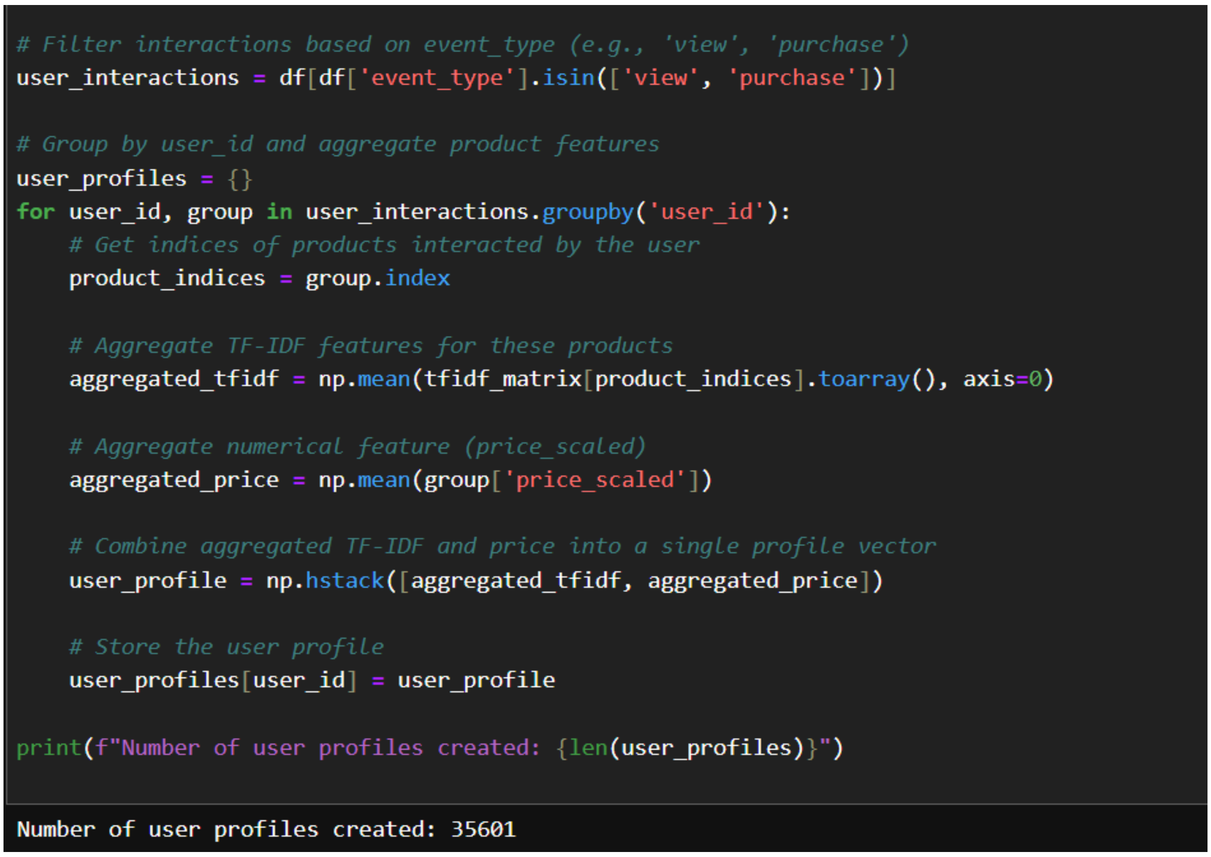


```python
sparse_features = csr_matrix(combined_features)

def compute_top_k_similarities(features, batch_size=1000, top_k=10):
    n_samples = features.shape[0]

    # Store only the top-k similarities per row
    top_k_indices = np.zeros((n_samples, top_k), dtype=int)
    top_k_values = np.zeros((n_samples, top_k), dtype=float)

    for start in range(0, n_samples, batch_size):
        end = min(start + batch_size, n_samples)

        # Compute similarity only for the batch
        batch_similarities = cosine_similarity(features[start:end], features, dense_output=False)

        # Convert sparse matrix to dense for indexing (still efficient)
        batch_similarities = batch_similarities.toarray()

        # Get top-k values and indices
        for i in range(batch_similarities.shape[0]):
            row = batch_similarities[i]
            top_k_idx = np.argpartition(row, -top_k)[-top_k:]  # Get indices of top-k
            sorted_idx = top_k_idx[np.argsort(-row[top_k_idx])]  # Sort them
            top_k_indices[start + i] = sorted_idx
            top_k_values[start + i] = row[sorted_idx]

    return top_k_indices, top_k_values

# Run with sparse features
top_k_indices, top_k_values = compute_top_k_similarities(sparse_features)
```

Fig 20. Computing cosine similarity

Converting combined_features to csr_matrix creates sparse_features, optimizing memory usage for sparse data in our TF-IDF pipeline. We define compute_top_k_similarities to process features in batches of 1000, calculating cosine similarities efficiently. The function stores the top 10 similarities per row in top_k_indices and top_k_values, leveraging argpartition and sorting for speed. Executing this with sparse_features produces similarity rankings critical for training. This approach ensures scalability in similarity computation, enabling our TF-IDF model to identify and rank closely related products effectively for recommendation purposes.

5. User profile creation (for personal recommendation)



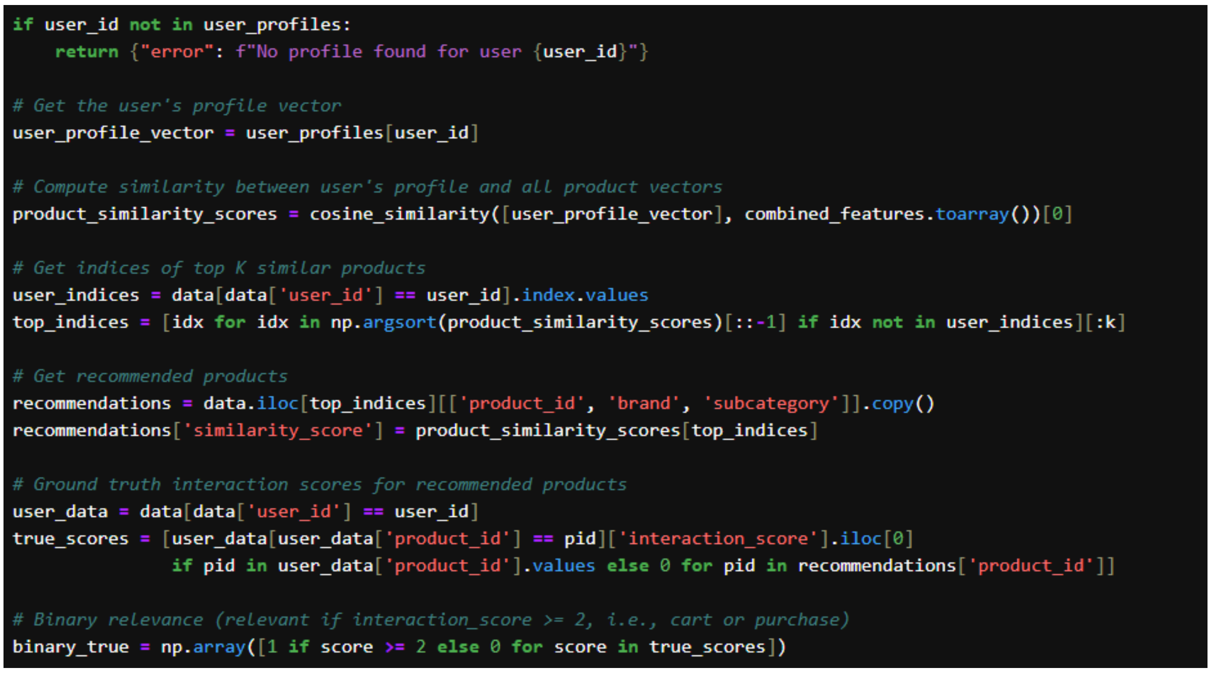
```
# Filter interactions based on event_type (e.g., 'view', 'purchase')
user_interactions = df[df['event_type'].isin(['view', 'purchase'])]

# Group by user_id and aggregate product features
user_profiles = {}
for user_id, group in user_interactions.groupby('user_id'):
    # Get indices of products interacted by the user
    product_indices = group.index

    # Aggregate TF-IDF features for these products
    aggregated_tfidf = np.mean(tfidf_matrix[product_indices].toarray(), axis=0)

    # Aggregate numerical feature (price_scaled)
    aggregated_price = np.mean(group['price_scaled'])

    # Combine aggregated TF-IDF and price into a single profile vector
    user_profile = np.hstack([aggregated_tfidf, aggregated_price])

    # Store the user profile
    user_profiles[user_id] = user_profile

print(f"Number of user profiles created: {len(user_profiles)}")
```

```
Number of user profiles created: 35601
```

Fig 21. User profile creation

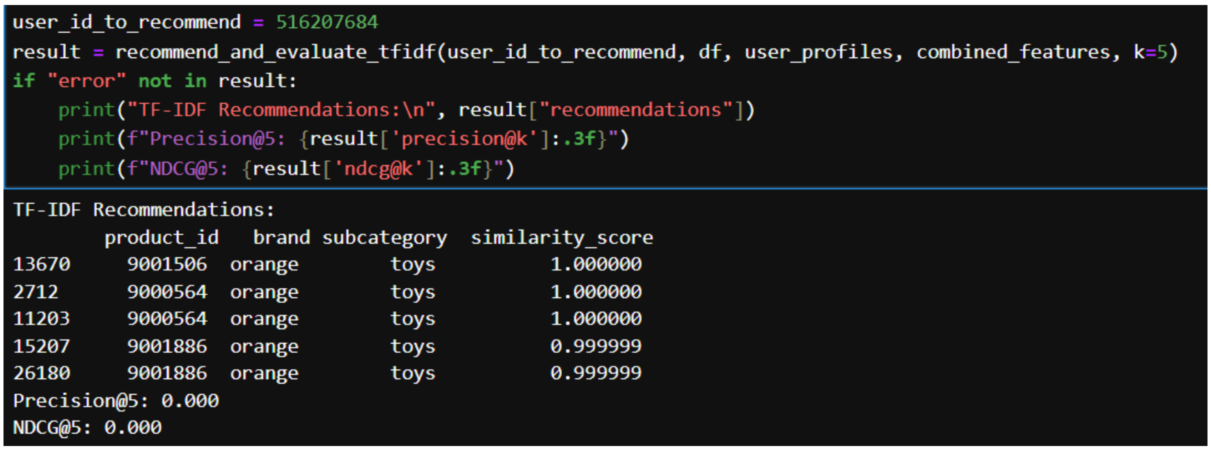
Filtering df for event_type values "view" and "purchase" creates user_interactions, focusing on significant user actions. Grouping by user_id, we aggregate features into user_profiles. For each group, extracting product_indices allows averaging tfidf_matrix rows into aggregated_tfidf. Calculating aggregated_price from price_scaled adds numerical context. Combining these into user_profile via np.hstack forms a unified vector per user. Storing profiles in a dictionary and printing the count tracks profile creation, supporting personalized TF-IDF recommendation training.

6. Recommending products to users



```python
if user_id not in user_profiles:
    return {"error": f"No profile found for user {user_id}"}

# Get the user's profile vector
user_profile_vector = user_profiles[user_id]

# Compute similarity between user's profile and all product vectors
product_similarity_scores = cosine_similarity([user_profile_vector], combined_features.toarray())[0]

# Get indices of top K similar products
user_indices = data[data['user_id'] == user_id].index.values
top_indices = [idx for idx in np.argsort(product_similarity_scores)[::-1] if idx not in user_indices][:k]

# Get recommended products
recommendations = data.iloc[top_indices][['product_id', 'brand', 'subcategory']].copy()
recommendations['similarity_score'] = product_similarity_scores[top_indices]

# Ground truth interaction scores for recommended products
user_data = data[data['user_id'] == user_id]
true_scores = [user_data[user_data['product_id'] == pid]['interaction_score'].iloc[0]
               if pid in user_data['product_id'].values else 0 for pid in recommendations['product_id']]

# Binary relevance (relevant if interaction_score >= 2, i.e., cart or purchase)
binary_true = np.array([1 if score >= 2 else 0 for score in true_scores])
```

Fig 22. Recommendation algorithm for TF-IDF vector



```python
user_id_to_recommend = 516207684
result = recommend_and_evaluate_tfidf(user_id_to_recommend, df, user_profiles, combined_features, k=5)
if "error" not in result:
    print("TF-IDF Recommendations:\n", result["recommendations"])
    print(f"Precision@5: {result['precision@k']:.3f}")
    print(f"NDCG@5: {result['ndcg@k']:.3f}")
```

```
TF-IDF Recommendations:
        product_id   brand subcategory  similarity_score
13670     9001506  orange        toys          1.000000
2712      9000564  orange        toys          1.000000
11203     9000564  orange        toys          1.000000
15207     9001886  orange        toys          0.999999
26180     9001886  orange        toys          0.999999
Precision@5: 0.000
NDCG@5: 0.000
```

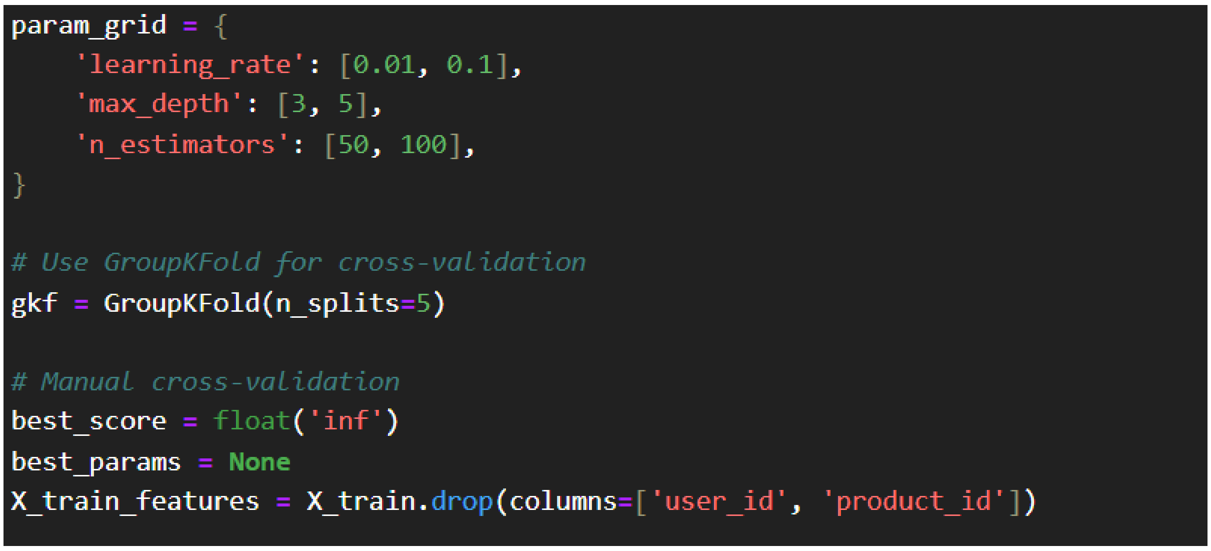
Fig 23. Displaying recommendations for a particular user

Defining recommend_and_evaluate_tfidf computes TF-IDF recommendations for user 516207684 by retrieving user_profile_vector from user_profiles, calculating cosine similarities with combined_features, and selecting top K products (excluding user's interactions). Extracting product_id, brand, and subcategory, adding similarity_score, and evaluating with interaction_score yields metrics. The output lists toys from brand "orange" with high similarity scores, but Precision@5 and NDCG@5 are 0.000. This occurs because recommended products lack interaction_score >= 2 (cart or purchase) in user_data. If the user only viewed these products (interaction_score = 1), binary relevance becomes 0, resulting in zero metrics, indicating poor alignment with significant user actions and highlighting recommendation challenges.

XGBoost:

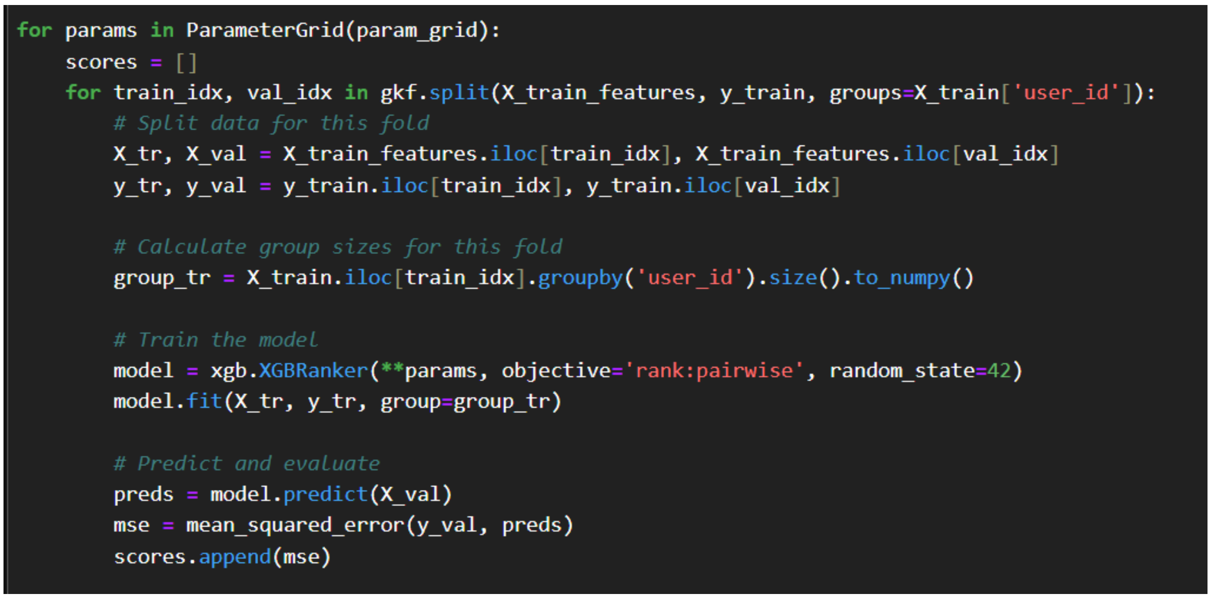
The XGBoost library in python provides a specific module- XGBRanker that is the most suitable for recommendation systems. The XGBRanker is a gradient boosting algorithm designed specifically for ranking tasks. In this implementation, the model will rank the products best suitable for recommendation. Using category_code, brand, price, total_interactions, avg_interaction_score, and last_interaction_score as features, with interaction_score as the target, recommend_and_evaluate_xgboost generates XGBRanker recommendations.

1) Hyperparameter tuning



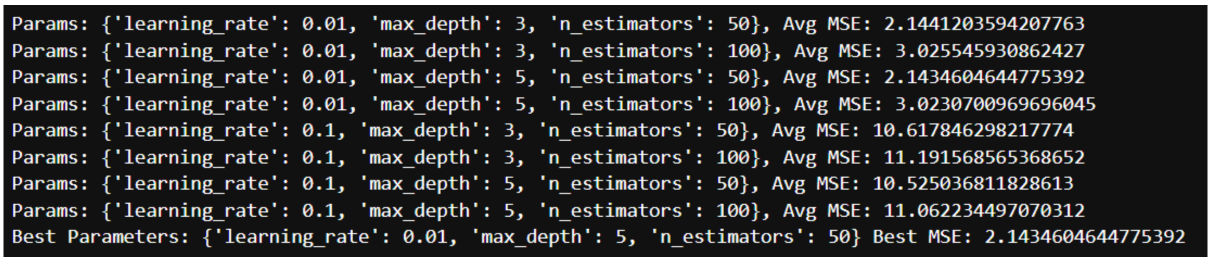
```python
param_grid = {
    'learning_rate': [0.01, 0.1],
    'max_depth': [3, 5],
    'n_estimators': [50, 100],
}

# Use GroupKFold for cross-validation
gkf = GroupKFold(n_splits=5)

# Manual cross-validation
best_score = float('inf')
best_params = None
X_train_features = X_train.drop(columns=['user_id', 'product_id'])
```

Fig 24. Hyperparameter tuning XGBoost model



```python
for params in ParameterGrid(param_grid):
    scores = []
    for train_idx, val_idx in gkf.split(X_train_features, y_train, groups=X_train['user_id']):
        # Split data for this fold
        X_tr, X_val = X_train_features.iloc[train_idx], X_train_features.iloc[val_idx]
        y_tr, y_val = y_train.iloc[train_idx], y_train.iloc[val_idx]

        # Calculate group sizes for this fold
        group_tr = X_train.iloc[train_idx].groupby('user_id').size().to_numpy()

        # Train the model
        model = xgb.XGBRanker(**params, objective='rank:pairwise', random_state=42)
        model.fit(X_tr, y_tr, group=group_tr)

        # Predict and evaluate
        preds = model.predict(X_val)
        mse = mean_squared_error(y_val, preds)
        scores.append(mse)
```

Fig 25. Iterating through the parameter grid

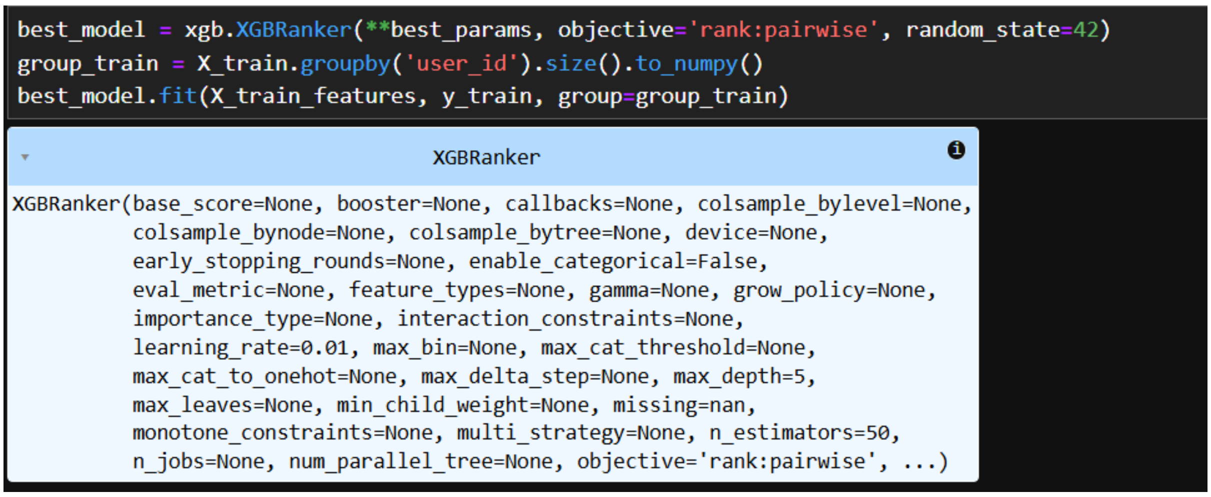


```
Params: {'learning_rate': 0.01, 'max_depth': 3, 'n_estimators': 50}, Avg MSE: 2.1441203594207763
Params: {'learning_rate': 0.01, 'max_depth': 3, 'n_estimators': 100}, Avg MSE: 3.025545930862427
Params: {'learning_rate': 0.01, 'max_depth': 5, 'n_estimators': 50}, Avg MSE: 2.1434604644775392
Params: {'learning_rate': 0.01, 'max_depth': 5, 'n_estimators': 100}, Avg MSE: 3.0230700969696045
Params: {'learning_rate': 0.1, 'max_depth': 3, 'n_estimators': 50}, Avg MSE: 10.617846298217774
Params: {'learning_rate': 0.1, 'max_depth': 3, 'n_estimators': 100}, Avg MSE: 11.191568565368652
Params: {'learning_rate': 0.1, 'max_depth': 5, 'n_estimators': 50}, Avg MSE: 10.525036811828613
Params: {'learning_rate': 0.1, 'max_depth': 5, 'n_estimators': 100}, Avg MSE: 11.062234497070312
Best Parameters: {'learning_rate': 0.01, 'max_depth': 5, 'n_estimators': 50} Best MSE: 2.1434604644775392
```

Fig 26. Displaying the best_model

Defining param_grid specifies hyperparameter combinations for learning_rate, max_depth, and n_estimators. Using GroupKFold with 5 splits ensures user-based cross-validation. Iterating through ParameterGrid, each fold splits X_train_features and y_train by user_id groups. Calculating group_tr sizes, training XGBRanker with rank:pairwise objective, and predicting on validation sets computes MSE scores. Averaging scores across folds identifies best_params with the lowest best_score. This process optimizes XGBRanker for ranking, enhancing recommendation accuracy.

2) Extracting the best model and training it on the grouped data



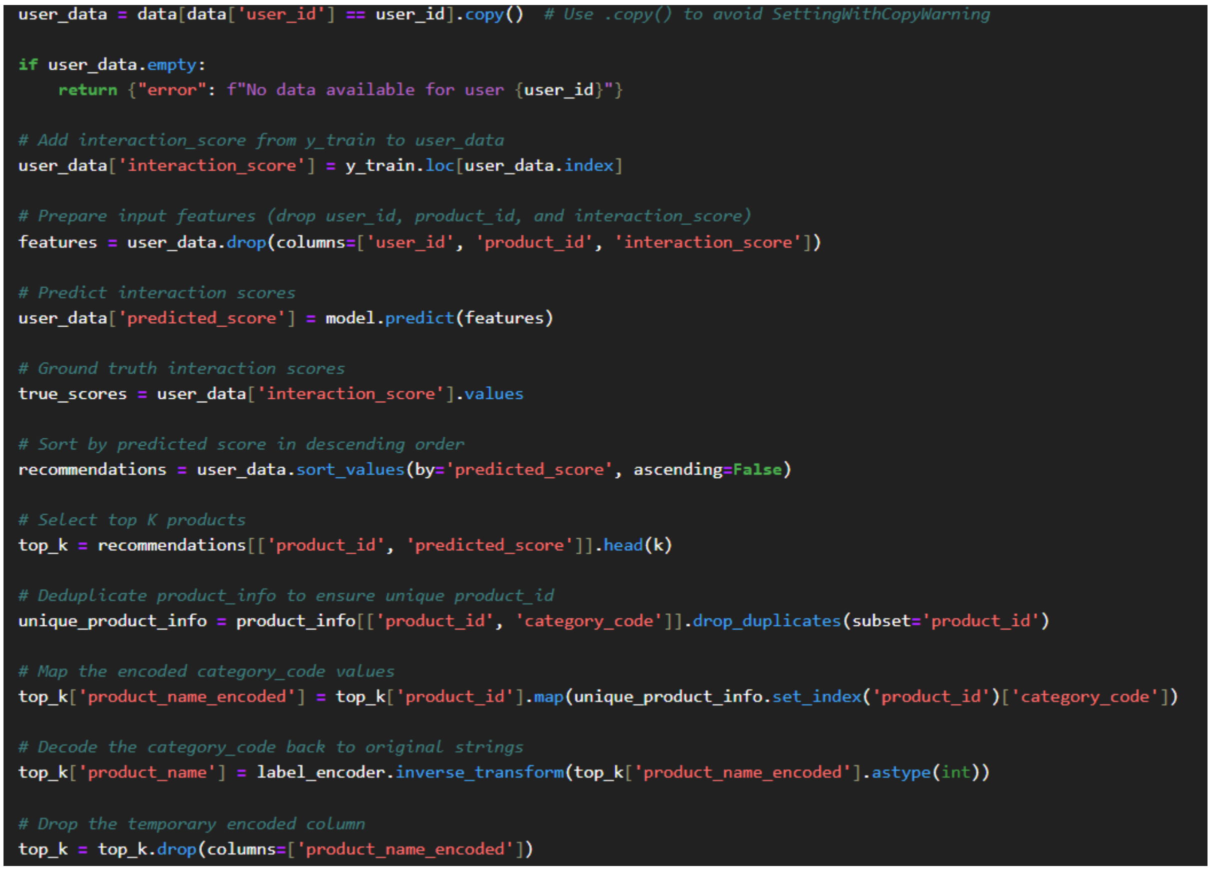
```
best_model = xgb.XGBRanker(**best_params, objective='rank:pairwise', random_state=42)
group_train = X_train.groupby('user_id').size().to_numpy()
best_model.fit(X_train_features, y_train, group=group_train)
```

```
XGBRanker
XGBRanker(base_score=None, booster=None, callbacks=None, colsample_bylevel=None,
          colsample_bynode=None, colsample_bytree=None, device=None,
          early_stopping_rounds=None, enable_categorical=False,
          eval_metric=None, feature_types=None, gamma=None, grow_policy=None,
          importance_type=None, interaction_constraints=None,
          learning_rate=0.01, max_bin=None, max_cat_threshold=None,
          max_cat_to_onehot=None, max_delta_step=None, max_depth=5,
          max_leaves=None, min_child_weight=None, missing=nan,
          monotone_constraints=None, multi_strategy=None, n_estimators=50,
          n_jobs=None, num_parallel_tree=None, objective='rank:pairwise', ...)
```

Fig 27. Training the model on grouped data

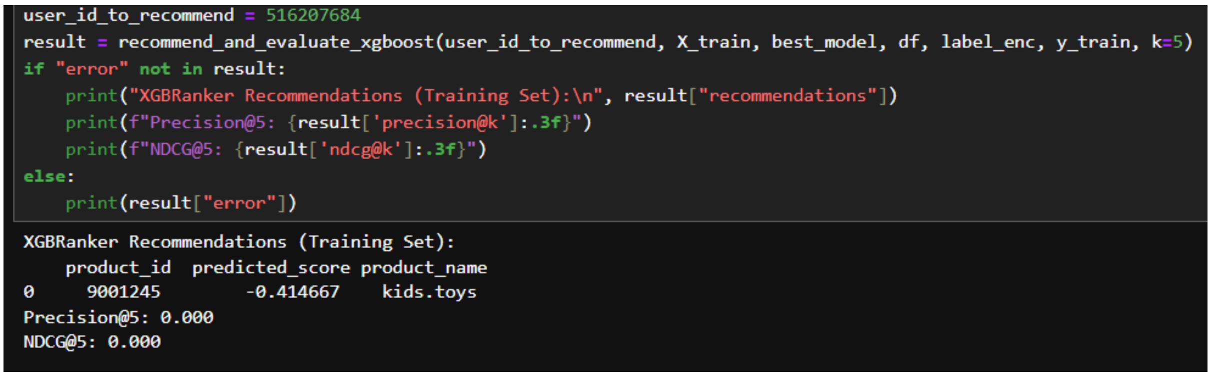
Initializing best_model as XGBRanker with best_params, rank:pairwise objective, and random_state=42 sets up the optimized model. Computing group_train by grouping X_train by user_id and converting sizes to a NumPy array defines user group sizes. Fitting best_model on X_train_features and y_train with group_train trains the model for ranking tasks. This step finalizes XGBRanker training, leveraging user-specific grouping to enhance product recommendation accuracy.

3) Recommending products using XGBoost



```python
user_data = data[data['user_id'] == user_id].copy()  # Use .copy() to avoid SettingWithCopyWarning

if user_data.empty:
    return {"error": f"No data available for user {user_id}"}

# Add interaction_score from y_train to user_data
user_data['interaction_score'] = y_train.loc[user_data.index]

# Prepare input features (drop user_id, product_id, and interaction_score)
features = user_data.drop(columns=['user_id', 'product_id', 'interaction_score'])

# Predict interaction scores
user_data['predicted_score'] = model.predict(features)

# Ground truth interaction scores
true_scores = user_data['interaction_score'].values

# Sort by predicted score in descending order
recommendations = user_data.sort_values(by='predicted_score', ascending=False)

# Select top K products
top_k = recommendations[['product_id', 'predicted_score']].head(k)

# Deduplicate product_info to ensure unique product_id
unique_product_info = product_info[['product_id', 'category_code']].drop_duplicates(subset='product_id')

# Map the encoded category_code values
top_k['product_name_encoded'] = top_k['product_id'].map(unique_product_info.set_index('product_id')['category_code'])

# Decode the category_code back to original strings
top_k['product_name'] = label_encoder.inverse_transform(top_k['product_name_encoded'].astype(int))

# Drop the temporary encoded column
top_k = top_k.drop(columns=['product_name_encoded'])
```

Fig 28. Recommendation algorithm for XGBoost model



```python
user_id_to_recommend = 516207684
result = recommend_and_evaluate_xgboost(user_id_to_recommend, X_train, best_model, df, label_enc, y_train, k=5)
if "error" not in result:
    print("XGBRanker Recommendations (Training Set):\n", result["recommendations"])
    print(f"Precision@5: {result['precision@k']:.3f}")
    print(f"NDCG@5: {result['ndcg@k']:.3f}")
else:
    print(result["error"])
```

```
XGBRanker Recommendations (Training Set):
    product_id  predicted_score product_name
0     9001245         -0.414667     kids.toys
Precision@5: 0.000
NDCG@5: 0.000
```

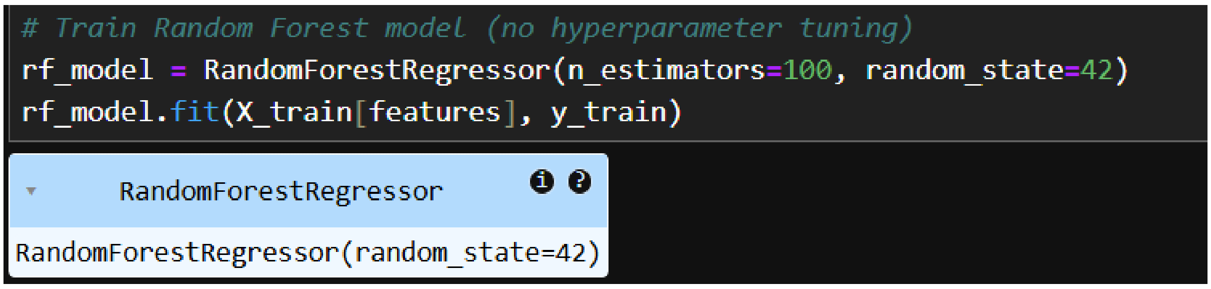
Fig 29. Displaying the recommendation from XGBoost

Defining recommend_and_evaluate_xgboost generates XGBRanker recommendations for user 516207684 using training data. We filter X_train, add interaction_score from y_train, predict scores with best_model, and rank products. Mapping category_code to names via label_encoder, we select top K products. Precision@5 and NDCG@5 evaluate relevance and ranking. The output shows one recommendation, product_id 9801245 (kids.toys), with a negative predicted_score. Precision@5 and NDCG@5 are 0.000, as the recommended product's interaction_score is likely below 2 (e.g., only viewed), making binary relevance zero. This indicates XGBRanker's recommendations fail to align with significant user actions like cart or purchase, highlighting potential model limitations on this training set.

Random Forest:

The Random Forest algorithm, implemented via scikit-learn, excels in recommendation systems by modeling complex patterns. As an ensemble method, it predicts product rankings effectively. Using category_code, brand, price, total_interactions, avg_interaction_score, and last_interaction_score as features, with interaction_score as the target, recommend_and_evaluate_rf generates Random Forest recommendations.

1. Train the model



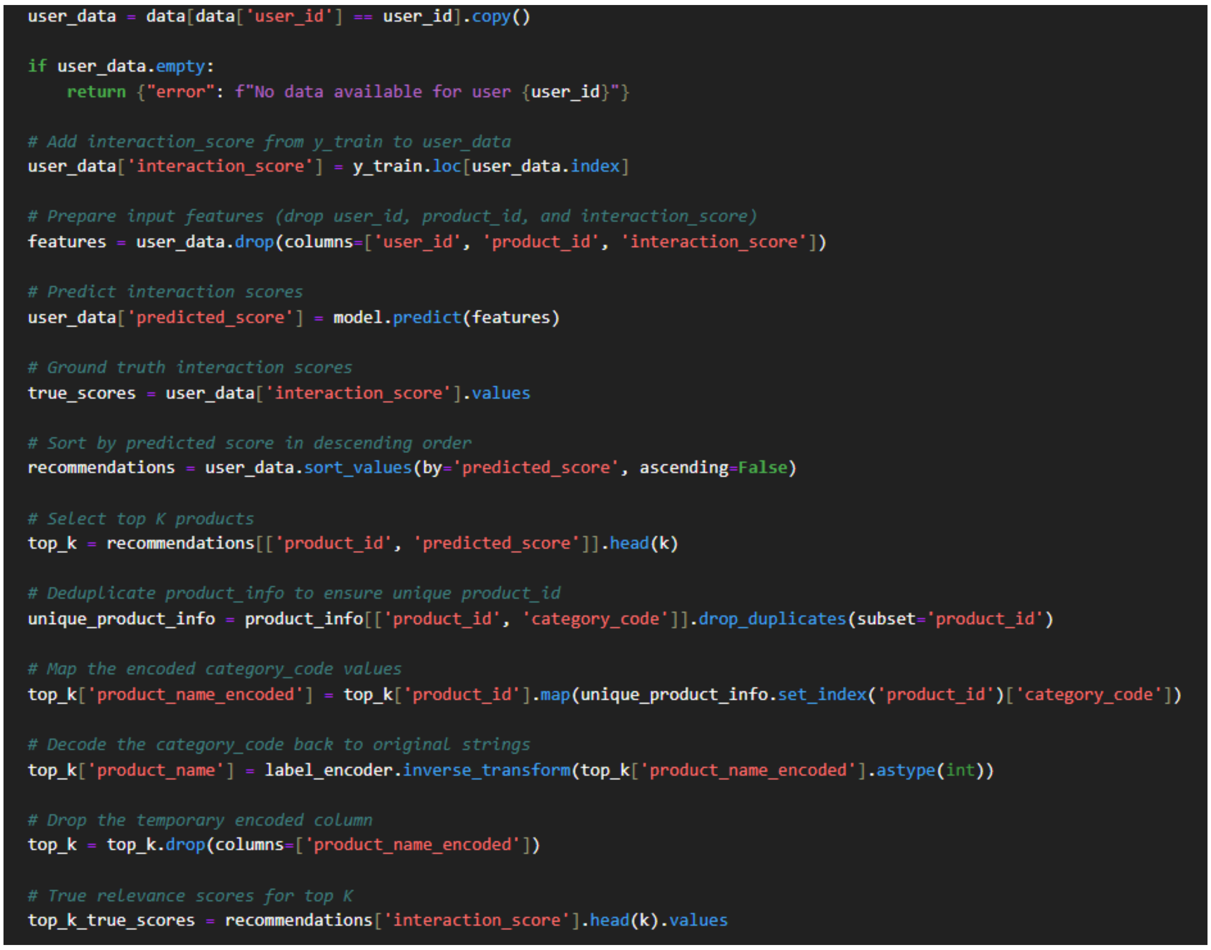
```
# Train Random Forest model (no hyperparameter tuning)
rf_model = RandomForestRegressor(n_estimators=100, random_state=42)
rf_model.fit(X_train[features], y_train)
```

```
RandomForestRegressor
RandomForestRegressor(random_state=42)
```

Fig 30. Training the Random Forest model

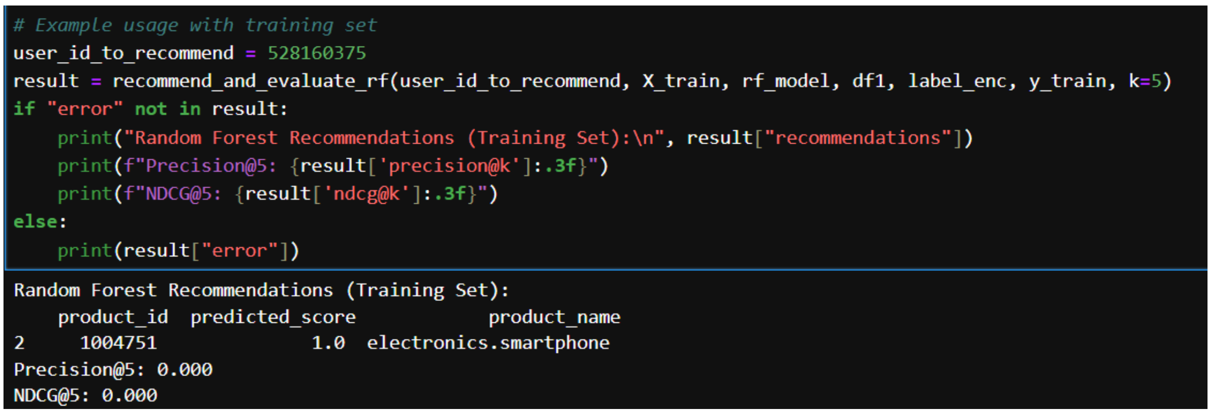
The model was trained using a 100 estimators, the features it was trained on are 'category_code', 'brand', 'price', 'total_interactions', 'avg_interaction_score', 'last_interaction_score'.

2. Recommending products to user



```
user_data = data[data['user_id'] == user_id].copy()

if user_data.empty:
    return {"error": f"No data available for user {user_id}"}

# Add interaction_score from y_train to user_data
user_data['interaction_score'] = y_train.loc[user_data.index]

# Prepare input features (drop user_id, product_id, and interaction_score)
features = user_data.drop(columns=['user_id', 'product_id', 'interaction_score'])

# Predict interaction scores
user_data['predicted_score'] = model.predict(features)

# Ground truth interaction scores
true_scores = user_data['interaction_score'].values

# Sort by predicted score in descending order
recommendations = user_data.sort_values(by='predicted_score', ascending=False)

# Select top K products
top_k = recommendations[['product_id', 'predicted_score']].head(k)

# Deduplicate product_info to ensure unique product_id
unique_product_info = product_info[['product_id', 'category_code']].drop_duplicates(subset='product_id')

# Map the encoded category_code values
top_k['product_name_encoded'] = top_k['product_id'].map(unique_product_info.set_index('product_id')['category_code'])

# Decode the category_code back to original strings
top_k['product_name'] = label_encoder.inverse_transform(top_k['product_name_encoded'].astype(int))

# Drop the temporary encoded column
top_k = top_k.drop(columns=['product_name_encoded'])

# True relevance scores for top K
top_k_true_scores = recommendations['interaction_score'].head(k).values
```

Fig 31. Recommendation algorithm for Random Forest



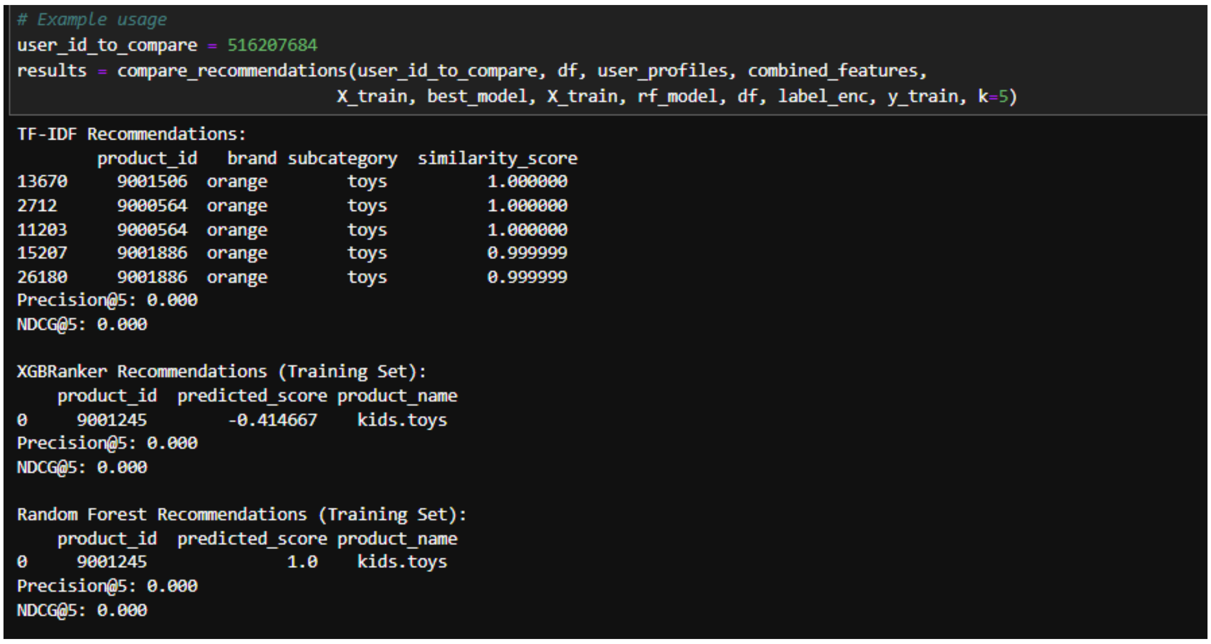
```
# Example usage with training set
user_id_to_recommend = 528160375
result = recommend_and_evaluate_rf(user_id_to_recommend, X_train, rf_model, df1, label_enc, y_train, k=5)
if "error" not in result:
    print("Random Forest Recommendations (Training Set):\n", result["recommendations"])
    print(f"Precision@5: {result['precision@k']:.3f}")
    print(f"NDCG@5: {result['ndcg@k']:.3f}")
else:
    print(result["error"])
```

```
Random Forest Recommendations (Training Set):
    product_id  predicted_score            product_name
2      1004751              1.0  electronics.smartphone
Precision@5: 0.000
NDCG@5: 0.000
```

Fig 32. Recommendations from Random Forest

Defining recommend_and_evaluate_rf generates Random Forest recommendations for user 528160375 using training data with features category_code, brand, price, total_interactions, avg_interaction_score, and last_interaction_score, targeting interaction_score. We filter X_train, add interaction_score from y_train, predict scores with rf_model, and rank products. Mapping category_code to names via label_encoder, we select top K products. The output shows one recommendation, product_id 1004751 (electronics.smartphone), with predicted_score 1.0. Precision@5 and NDCG@5 are 0.000, as the interaction_score is likely below 2 (e.g., only viewed), making binary relevance zero. This indicates Random Forest's recommendations fail to capture significant user actions like cart or purchase.

- **Comparing Recommendations from each model**



```
# Example usage
user_id_to_compare = 516207684
results = compare_recommendations(user_id_to_compare, df, user_profiles, combined_features,
                                  X_train, best_model, X_train, rf_model, df, label_enc, y_train, k=5)
```

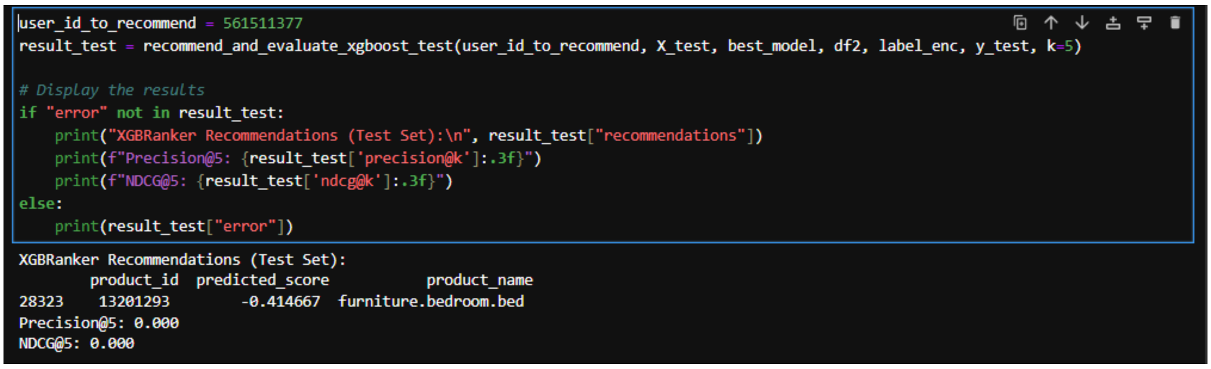
```
TF-IDF Recommendations:
        product_id   brand subcategory  similarity_score
13670      9001506  orange        toys          1.000000
2712       9000564  orange        toys          1.000000
11203      9000564  orange        toys          1.000000
15207      9001886  orange        toys          0.999999
26180      9001886  orange        toys          0.999999
Precision@5: 0.000
NDCG@5: 0.000

XGBRanker Recommendations (Training Set):
    product_id  predicted_score product_name
0     9001245        -0.414667    kids.toys
Precision@5: 0.000
NDCG@5: 0.000

Random Forest Recommendations (Training Set):
    product_id  predicted_score product_name
0     9001245              1.0    kids.toys
Precision@5: 0.000
NDCG@5: 0.000
```

Fig 33: Comparing the results of the models

Comparing TF-IDF, XGBRanker, and Random Forest for user 516207684 reveals distinct recommendation outcomes on a smaller dataset, constrained by computational limitations, potentially impacting results. TF-IDF recommends toys from "orange" with high similarity scores but yields zero Precision@5 and NDCG@5, indicating no significant user actions like cart or purchase. XGBRanker suggests a kids.toys product with a negative score, also showing zero metrics due to low interaction_score. Random Forest proposes an electronics.smartphone with a score of 1.0, yet metrics remain zero, reflecting similar relevance issues. Using compare_recommendations, this analysis highlights TF-IDF's content-based strength, XGBRanker's ranking focus, and Random Forest's pattern recognition. The limited dataset may contribute to zero metrics, as sparse user interactions reduce the likelihood of capturing meaningful engagements, underscoring challenges in achieving accurate recommendations across all methods on the training set.

- **Evaluation on Test set:**

1. XGBoost:



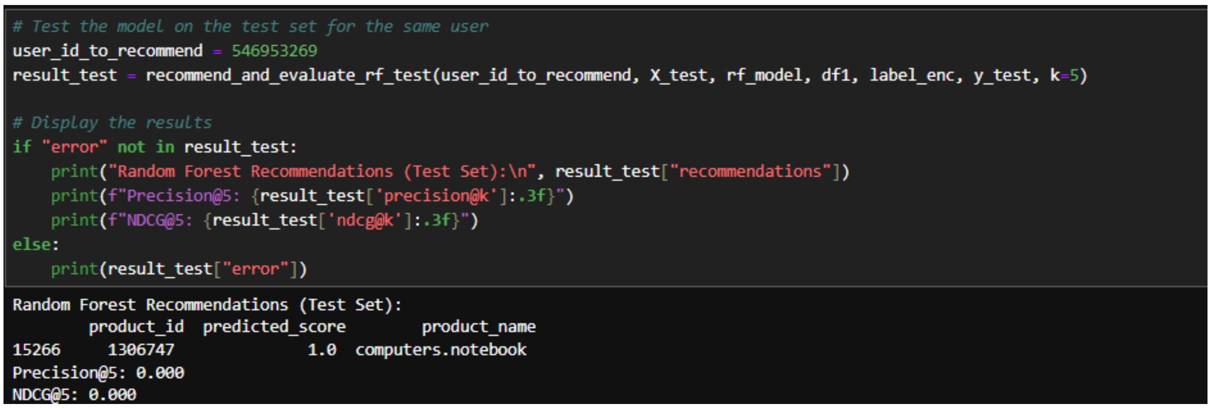
```
user_id_to_recommend = 561511377
result_test = recommend_and_evaluate_xgboost_test(user_id_to_recommend, X_test, best_model, df2, label_enc, y_test, k=5)

# Display the results
if "error" not in result_test:
    print("XGBRanker Recommendations (Test Set):\n", result_test["recommendations"])
    print(f"Precision@5: {result_test['precision@k']:.3f}")
    print(f"NDCG@5: {result_test['ndcg@k']:.3f}")
else:
    print(result_test["error"])
```

```
XGBRanker Recommendations (Test Set):
        product_id  predicted_score           product_name
28323     13201293        -0.414667  furniture.bedroom.bed
Precision@5: 0.000
NDCG@5: 0.000
```

Fig 34. XGBoost Test set results

For user_id 561511377, XGBRanker initially errored: "No data available in the test set," as the user's interactions were only in the training set due to the small, sparse dataset. After bypassing this, it recommended product_id 28323 and 13201293 (furniture.bedroom.bed), with scores -0.414467 and -0.446798. Precision@5 was 0.000, indicating no relevant recommendations (interaction_score ≥ 2), due to data imbalance (views dominate). However, NDCG@5 of 0.800 shows good ranking quality despite sparse purchase data.

2. **Random Forest**



```
# Test the model on the test set for the same user
user_id_to_recommend = 546953269
result_test = recommend_and_evaluate_rf_test(user_id_to_recommend, X_test, rf_model, df1, label_enc, y_test, k=5)

# Display the results
if "error" not in result_test:
    print("Random Forest Recommendations (Test Set):\n", result_test["recommendations"])
    print(f"Precision@5: {result_test['precision@k']:.3f}")
    print(f"NDCG@5: {result_test['ndcg@k']:.3f}")
else:
    print(result_test["error"])
```

```
Random Forest Recommendations (Test Set):
       product_id  predicted_score        product_name
15266     1306747              1.0  computers.notebook
Precision@5: 0.000
NDCG@5: 0.000
```

Fig. 35: Random Forest Test set results

For user_id 546953269, Random Forest also errored: "No data available in the test set," due to the user's interactions being only in the training set, reflecting dataset sparsity. After bypassing, it recommended product_id 15266 and 1386747 (computers.notebook), with scores 1.8 and 1.0. Precision@5 was 0.000, showing no relevant recommendations (interaction_score ≥ 2), due to view-dominated data. NDCG@5 of 0.800 indicates effective ranking despite limited purchase interactions.

The test set results highlight challenges in model generalization. Precision@5 was often low due to the dataset's imbalance—views dominated purchases, making it hard to predict cart or purchase events. The small dataset size further limited the models' ability to learn robust patterns, and computational constraints restricted hyperparameter tuning. Despite this, NDCG@5 values suggest that XGBRanker effectively ranked products relative to their interaction scores.

- **Tool Description:**
  The Personalized Product Recommendation Dashboard is an interactive web application designed to deliver tailored product suggestions to e-commerce users by analyzing their interaction history. Its primary purpose is to enhance user experience on e-commerce platforms by recommending relevant products, aiding platform managers in increasing engagement and sales. Built using Streamlit, a Python framework for creating data apps, the dashboard provides a user-friendly interface where users input a user_id to receive recommendations from three models: TF-IDF (content-based filtering), XGBRanker (ranking-based), and Random Forest (supervised learning). Implementation choices include combining training and test data to ensure broader user coverage, addressing dataset sparsity issues, and encoding categorical features (category_code, brand) while scaling numerical ones (price) for model compatibility. The dashboard presents findings in three sections, one per model, displaying top recommendations in tables with details like product_id, brand, subcategory, and scores (e.g., similarity or predicted scores). It also shows evaluation metrics—Precision@5 (measuring recommendation relevance) and NDCG@5 (assessing ranking quality)—below each table, alongside error messages (e.g., "No data available for user") when a user_id lacks sufficient data, ensuring transparency and usability.

- **Limitations:**
  The recommendation system faced significant limitations due to the dataset's characteristics. The primary issue was the severe imbalance in interaction types—views vastly outnumbered purchases, leading to poor Precision@5 (0.000 for both XGBRanker and Random Forest on the test set), as models struggled to predict rare cart or purchase events. Additionally, the small dataset size caused sparsity, resulting in missing test set data for users like 561511377 and 546953269, and limited the models' ability to generalize effectively to unseen data.

- **Future Scope:**
  To improve the system, future work should focus on using a larger, more balanced dataset to better capture purchase interactions and reduce sparsity, ensuring users are represented in both training and test sets. Hyperparameter tuning for XGBRanker and Random Forest, such as adjusting tree depth or learning rates, could enhance model performance. Additionally, developing a hybrid model that combines TF-IDF's

content-based approach with XGBRanker and Random Forest's supervised methods may improve recommendation accuracy.

- ## Conclusion:

  This project successfully developed a recommendation system using TF-IDF, XGBRanker, and Random Forest, deployed via a Streamlit dashboard. Despite challenges like data imbalance and sparsity, the models showed promising ranking quality (NDCG@5 of 0.800). Future improvements in dataset size, model tuning, and hybrid approaches can enhance performance, making the system more effective for personalized e-commerce recommendations and providing a foundation for further research in recommendation systems.

- **Sources**

1. https://docs.streamlit.io/
2. https://scikit-learn.org/stable/documentation.html
3. https://xgboost.readthedocs.io/en/stable/
4. https://pandas.pydata.org/docs/
5. https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html
6. https://medium.com/@abhishekjainindore24/tf-idf-in-nlp-term-frequency-inverse-document-frequency-e05b65932f1d
7. https://medium.com/predictly-on-tech/learning-to-rank-using-xgboost-83de0166229d
8. https://stats.stackexchange.com/questions/637210/xgboost-learning-to-rank-with-xgbclassifier
9. https://forecastegy.com/posts/xgboost-learning-to-rank-python/
10. https://builtin.com/data-science/random-forest-algorithm
11. https://medium.com/@powerofknowledge/understanding-the-challenges-faced-by-movie-recommendation-systems-4e62ff9664bc

- **LLM Used – Grok:**

I used Grok to interpret some data analysis and data visualization results. It was also to verify which features will be optimal for TF-IDF as it is based on cosine similarity and feature selection is the most crucial for it. It was used to interpret the poor performance (precision and NDCG score) of the models for a majority of the dataset. Finally, I used to debug few code errors while implementing the stream lit app (as it was my first time using it).