knowledgehut upGrad

upGrad

# DATA SCIENCE HACKATHON

## HACKATHON TOPIC

## { E-COMMERCE PRODUCT CATEGORIZATION }

- RAJU KUMAR

## PROBLEM STATEMENT: ECOMMERCE PRODUCT CATEGORIZATION

**Objective**: Create a machine learning model for accurate classification of eCommerce products into categories based on product descriptions and numerical features.
Background:
Using the eCommerce domain product categorization a company can improve the search relevance, user experience, and inventory optimization. The manual process of categorization requires considerable time and effort and is also vulnerable to errors that bring about the data being enormously large. However, this machine learning has the potential to create an automatic system that will cut down labor and resources required while improving efficiency.

**Data Preparation:**

Test Data: Through a data set of extra product descriptions and numerical features process for predicting their categories be carried out.

**Feature Engineering:**

Text Features: Employ TF-IDF to vectorize the product description data so that they convert into vector representations.

**Model Development:**

Train-Test Split: Divide the training data into training and validation sets.
Model Training: Employ a LightGBM model to predict the product categories based on the integrated features of product descriptions and numerical data.
Model Evaluation: Assess the model's accuracy and confusion matrix on the validation set to determine its performance.

**Prediction and Results:**

Test Data Processing: Use the trained model to apply the test dataset to the prediction of product categories.
Output: Save your predictions in a CSV file and evaluate them to analyze the results.
Confusion Matrix: This method compares the model's predictions using the actual categories (if provided) for both training and test datasets.

**Data Sources:**

**Training Data:** train_product_data.csv - This data contains product descriptions, numerical features, and category labels.
**Test Data:** test_data.csv - The product descriptions are included along with numerical features that have not been labeled with category labels. Deliverables:
A trained model that can categorize products into different categories.
A CSV file that lists the predicted categories of the test data.
Presentations includes confusion matrices for the purpose of assessing the performance of the model.

**Challenges:**
Handling missing values in product descriptions and numerical features.
Addressing any class imbalance or feature mismatch issues.

**Future Enhancements:**
Incorporate additional features or external data sources to improve prediction accuracy.
Experiment with other machine learning algorithms or deep learning approaches.
Implement real-time classification and integration into an eCommerce platform.

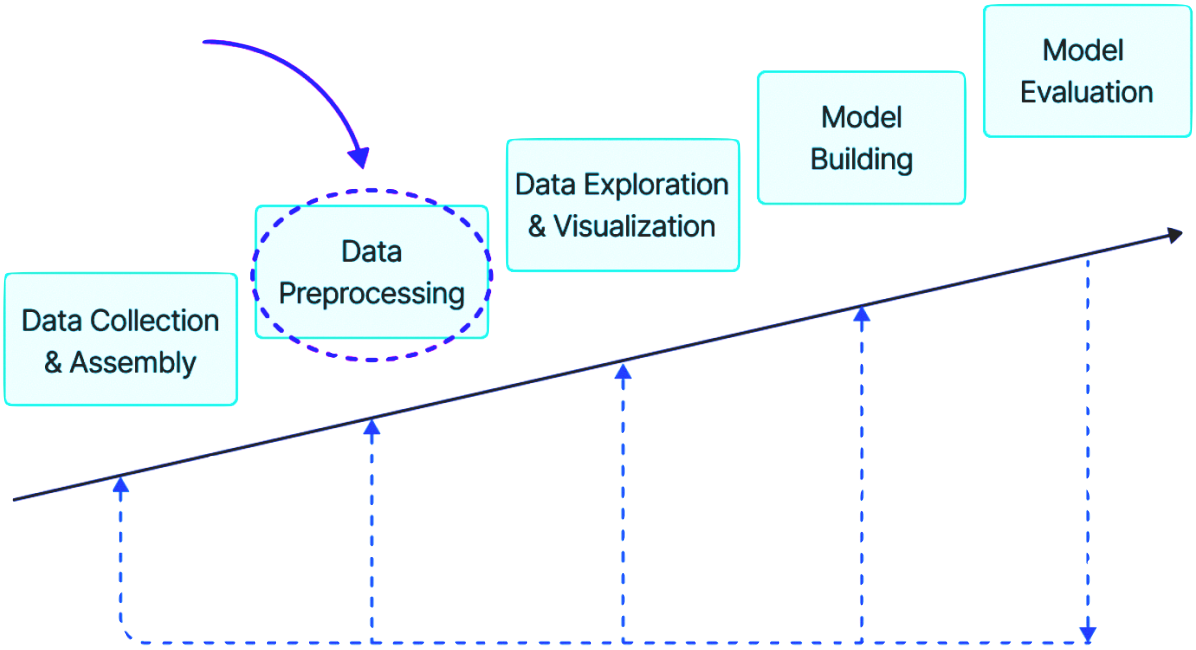
```
[ ] # displays the first few rows of the train_data.
    train_data.head()
```

| | uniq_id | crawl_timestamp | product_url | product_name | product_category_tree | pid | retail_price | discounted_price | image | is_FK_Advantage_product | description | product_rating | overall_rating | brand | product_specifications |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | c2d766ca982eca8304150849735ff9 | 2016-03-25 22:59:23 +0000 | http://www.flipkart.com/alisha-solid-women-s-c... | Alisha Solid Women's Cycling Shorts | Clothing | SRTEH2FF9KEDEFGF | 999.0 | 379.0 | ["http://img5a.flixcart.com/image/short/u/4/a/... | False | Key Features of Alisha Solid Women's Cycling S... | No rating available | No rating available | Alisha | {"product_specification"=>[{"key"=>"Number of ... |
| 1 | f449ec65dcbc041b6ae5e6a32717d01b | 2016-03-25 22:59:23 +0000 | http://www.flipkart.com/aw-bellies/p/itmaf4grg... | AW Bellies | Footwear | SHOEH4GRSUBJGZXE | 999.0 | 499.0 | ["http://img5a.flixcart.com/image/shoe/7/z/z/r... | False | Key Features of AW Bellies Sandals Wedges Heel... | No rating available | No rating available | AW | {"product_specification"=>[{"key"=>"Ideal For"... |
| 2 | 0973b37acd0c664e3de26e97e5571454 | 2016-03-25 22:59:23 +0000 | http://www.flipkart.com/alisha-solid-women-s-c... | Alisha Solid Women's Cycling Shorts | Clothing | SRTEH2F6HUZMQ66J | 699.0 | 267.0 | ["http://img5a.flixcart.com/image/short/6/2/h/... | False | Key Features of Alisha Solid Women's Cycling S... | No rating available | No rating available | Alisha | {"product_specification"=>[{"key"=>"Number of ... |
| 3 | ce5a6818f7707ed2b611ddcdbba61f5ad | 2016-03-25 22:59:23 +0000 | http://www.flipkart.com/alisha-solid-women-s-c... | Alisha Solid Women's Cycling Shorts | Clothing | SRTEH2FVVKRBAXHB | 1199.0 | 479.0 | ["http://img5a.flixcart.com/image/short/p/j/z/... | False | Key Features of Alisha Solid Women's Cycling S... | No rating available | No rating available | Alisha | {"product_specification"=>[{"key"=>"Number of ... |
| 4 | 29c6d290caa451f97b1c32df64477a2c | 2016-03-25 22:59:23 +0000 | http://www.flipkart.com/dill-bazaar-bellies-... | dill bazaar Bellies, Corporate Casuals, Casuals | Footwear | SHOEH3DZBFR8ESCK | 699.0 | 349.0 | ["http://img5a.flixcart.com/image/shoe/b/p/n/p... | False | Key Features of dill bazaar Bellies, Corpora... | No rating available | No rating available | dill bazaar | {"product_specification"=>[{"key"=>"Occasion",... |

```
[ ] # displays the first few rows of the test_data.
```

## WHAT IS DATA PREPROCESSING?

Data preprocessing refers to steps in transforming or encoding data into a format easy for machines to understand. This step should be done in order to make sure that the features of the data are as precise and correct as possible to be accurately and precisely interpreted by our models.




## WHY IS DATA PREPROCESSING IMPORTANT?

Most of the real-world data used for machine learning is messy, incomplete, and full of noise because of their variant origins. Running data mining algorithms on that kind of data will not provide quality results since they will not effectively identify the patterns. That is why preprocessing of data is important since it improves the quality in general of the data.

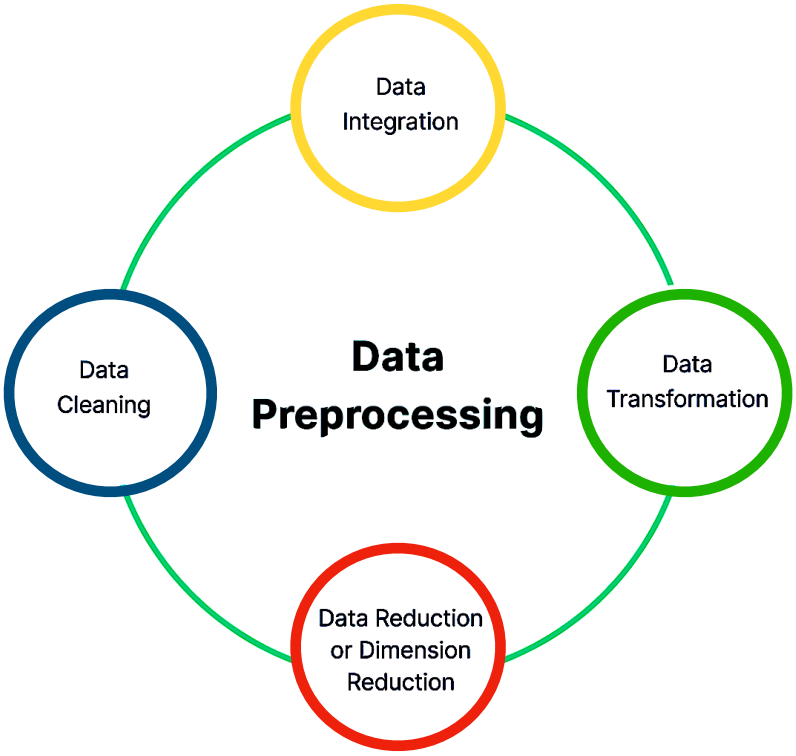
Handling Duplicates and Missing Values: These duplicate or missing values will distort overall statistics and insights from the data.
Manage Outliers and Inconsistencies: Outliers and inconsistent data points mislead the model's learning process, making wrong predictions. Quality Decisions Require Quality Data:

Quality is obtained from quality data. Preprocessing ensures we obtain this quality data, avoiding a situation where "Garbage In, Garbage Out."

## FEATURES IN MACHINE LEARNING

In machine learning, features are simply the individual independent variables that we feed into our models. They may be viewed as characteristics, or if you may will, attributes that describe the data and help the model in making predictions. For instance, these may comprise columns like Name, Age, Sex, Fare, etc., in a structured dataset such as a CSV file.

## STEPS TO DATA PREPROCESSING

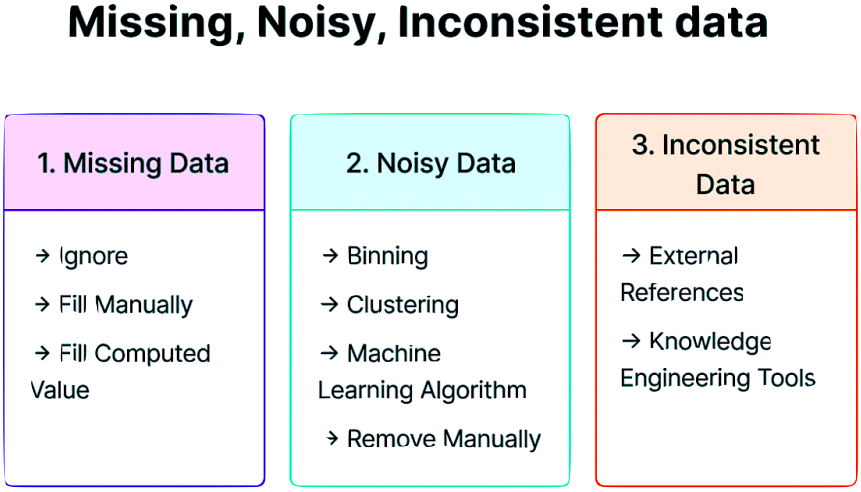



Let us discuss the four broad steps involved in data preprocessing:

### 1. DATA CLEANING

The process of correcting or removing incorrect, corrupted, duplicated, or incomplete data in a set of data.

- **Handling Missing Values:**



**Ignore Tuples:** This is particularly useful when there are too many missing values within a tuple and the dataset is large.

**Fill in Missing Values:** Some of the methods to fill up missing values include manual filling, prediction based on regression, and using mean/mode/median.
Dealing with

- **Noisy Data:**

  **Binning**: Data is divided into equal-sized bins, and through smoothening, some of the noise is removed by replacing values by mean, median, or boundary values.

  **Regression:** Regression techniques are used to fit all data points and smoothen out noise. One independent attribute is fitted using linear regression, while more than one are fitted using polynomial regression.

  **Clustering:** In this technique, data points are grouped into clusters. Those points that don't fit with any cluster can be treated as noise and removed.

- **Removing Outliers:**

  Those data points that are very far away from the other observations are called outliers and can be detected and removed using clustering.

```
[ ]  # Calculate IQR for retail_price
     Q1_retail = train_data['retail_price'].quantile(0.25)
     Q3_retail = train_data['retail_price'].quantile(0.75)
     IQR_retail = Q3_retail - Q1_retail
     lower_bound_retail = Q1_retail - 1.5 * IQR_retail
     upper_bound_retail = Q3_retail + 1.5 * IQR_retail

     # Identify outliers
     outliers_retail = (train_data['retail_price'] < lower_bound_retail) | (train_data['retail_price'] > upper_bound_retail)

     # Calculate IQR for discounted_price
     Q1_discounted = train_data['discounted_price'].quantile(0.25)
     Q3_discounted = train_data['discounted_price'].quantile(0.75)
     IQR_discounted = Q3_discounted - Q1_discounted
     lower_bound_discounted = Q1_discounted - 1.5 * IQR_discounted
     upper_bound_discounted = Q3_discounted + 1.5 * IQR_discounted

     # Identify outliers
     outliers_discounted = (train_data['discounted_price'] < lower_bound_discounted) | (train_data['discounted_price'] > upper_bound_discounted)

     # Combine outlier conditions
     outliers_combined = outliers_retail | outliers_discounted

     # Remove outliers
     cleaned_data = train_data[~outliers_combined]

     # Print the shape of the dataset before and after removing outliers
     print(f"Original dataset shape: {train_data.shape}")
     print(f"Cleaned dataset shape: {cleaned_data.shape}")

     # Display statistics for the cleaned dataset
     print("\nCleaned Retail Price Statistics:")
     print(cleaned_data['retail_price'].describe())

     print("\nCleaned Discounted Price Statistics:")
     print(cleaned_data['discounted_price'].describe())
```

## 2. DATA INTEGRATION

Data integration is the process of integration of data from several sources into one view, like a data warehouse. This is quite useful in solving real-world problems, like detecting medical conditions from CT scan images by integration of data from various medical sources.

- **Some of the issues are:**

  **Schema Integration and Object Matching:** Different formats and attributes may lead to difficulty in integration.

  **Removal of Redundant Attributes**: Same data may be present in more than one view; those duplicate data have to be identified and removed.

  **Data Value Conflict Resolution:** Treatment of inconsistencies in the values of data.

## 3. DATA TRANSFORMATION

Finally, after cleaning, we transform the data into appropriate form by applying different strategies.

- **Generalization:**

Low-level data is converted to high-level information through a concept hierarchy; for example, transforming city names into country names.

- **Normalization**:

  This is a process where numerical attributes are scaled to come within some specified range. It ensures homogeneity and brings better correlation among different data points. Different techniques of normalisation are min-max normalization, Z-score normalization, and decimal scaling.

  **Attribute Transformation:** Some new property values are derived from given attributes to make data mining easier. For example, date of birth can be transformed into a property like age group "senior citizen."

  **Aggregation:** Aggregation of data, e.g., aggregation of sales data by month or year.

## 4. DATA REDUCTION

While managing huge data volumes, the task of how to reduce it in size without losing any key information becomes paramount for analysis and modeling.

- **Data Cube Aggregation:** The data is summarized into a more manageable form.
- **Dimensionality Reduction:** The number of redundant features can be reduced using techniques such as Principal Component Analysis.

- **Data Compression:** It reduces the size of the data through some encoding techniques. It may be lossless or lossy.

## WHY IS EXPLORATORY DATA ANALYSIS IMPORTANT?

The Exploratory Data Analysis (EDA) is one of the preliminary steps in the analytical process of a data set. At this stage, the data is visualised and analyzed in order to achieve a general idea of the main points of the data, see the patterns if any, and discover if there are relationships between the variables. The EDA techniques could include the use of Python libraries like Pandas, NumPy, Matplotlib Seaborn, and Plotly.

## 5. PREPROCESSING STEPS

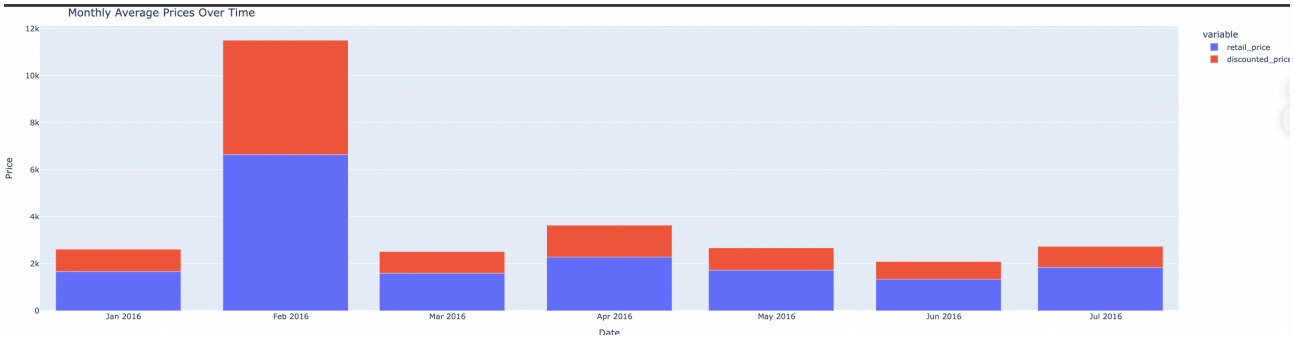
Handle Missing Values: Fill missing values in the description column with an empty string.

Feature Extraction:

**TF-IDF Vectorization:** Convert text descriptions into numerical features using TF-IDF vectorization.

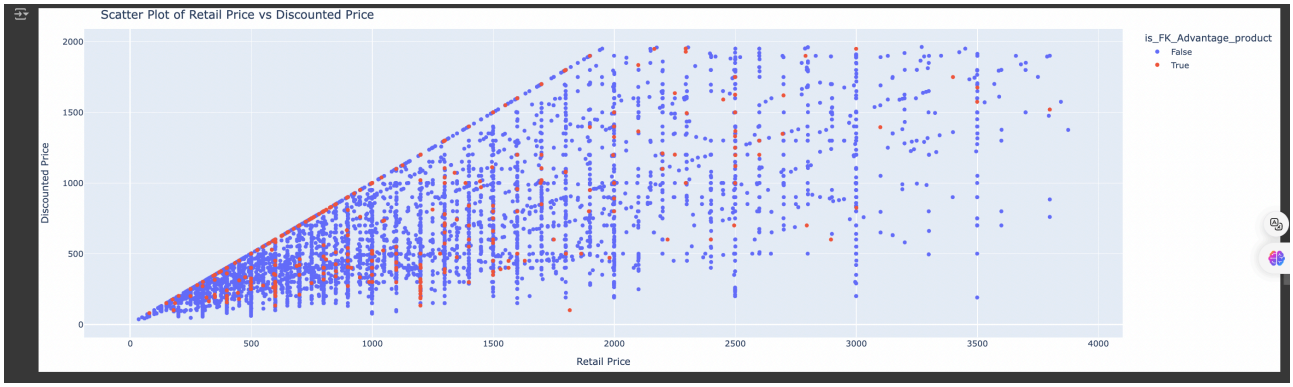
```
[ ] from sklearn.feature_extraction.text import TfidfVectorizer
    from sklearn.preprocessing import OneHotEncoder
    from scipy.sparse import hstack, csr_matrix

    # Convert 'description' to TF-IDF features
    tfidf = TfidfVectorizer(stop_words='english')
    X_description = tfidf.fit_transform(train_data['description'])

    # Prepare numerical features
    numerical_features = train_data[['retail_price', 'discounted_price']]

    # Convert categorical features to numerical
    categorical_features = train_data[['product_name', 'image', 'product_rating', 'overall_rating', 'brand', 'product_specifications']]
    encoder = OneHotEncoder(sparse=False, drop='first')
    categorical_features_encoded = encoder.fit_transform(categorical_features)

    # Combine all features
    X_combined = hstack([csr_matrix(numerical_features.values), categorical_features_encoded, X_description])
```

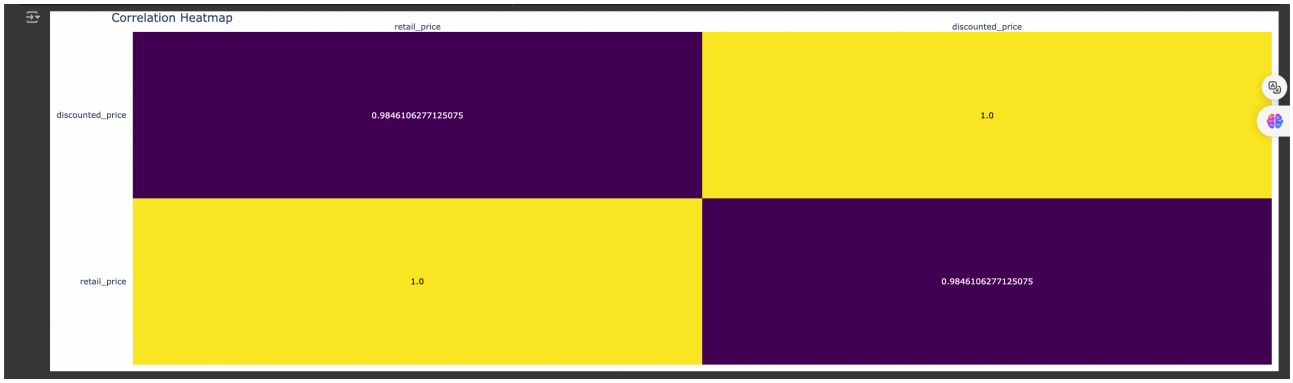
**Standard Scaling:** Scale numerical features (retail_price, discounted_price) to make sure they all have a mean of 0 and a standard deviation of 1.

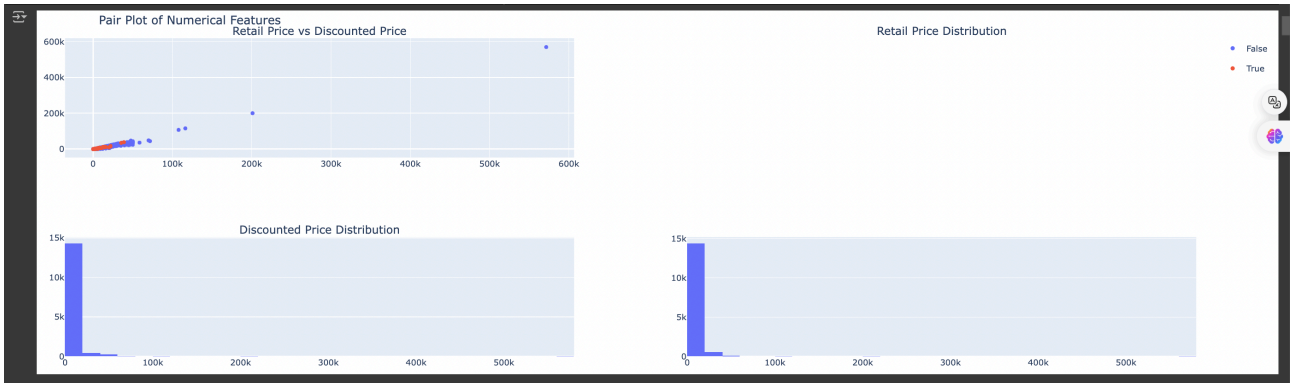


## 6. FEATURE ENGINEERING

Using hstack from Scipy, combine scaled numerical features and TF-IDF vectorized text features into one feature set.




## 7. MODEL TRAINING

Train-Test Split: Split the combined features and target variable into a training and a validation set.


Model Training: Train a LightGBM classifier on this training set.

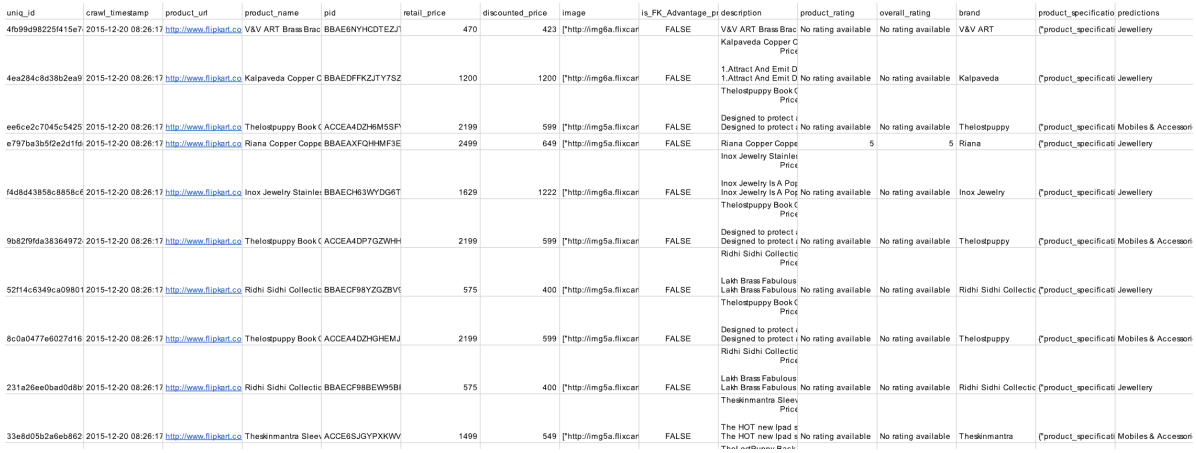
```
[32] from sklearn.ensemble import RandomForestClassifier, GradientBoostingClassifier, AdaBoostClassifier
     from sklearn.linear_model import LogisticRegression
     from sklearn.svm import SVC
     from sklearn.naive_bayes import MultinomialNB
     from sklearn.tree import DecisionTreeClassifier
     from sklearn.neighbors import KNeighborsClassifier
     from sklearn.ensemble import ExtraTreesClassifier
     from xgboost import XGBClassifier
     from lightgbm import LGBMClassifier
     from sklearn.metrics import classification_report, accuracy_score

     # Initialize models
     models = {
        "Random Forest": RandomForestClassifier(n_estimators=100, random_state=42),
        "Logistic Regression": LogisticRegression(max_iter=1000, random_state=42),
        "Support Vector Classifier": SVC(random_state=42),
        "Naive Bayes": MultinomialNB(),
        "Decision Tree": DecisionTreeClassifier(random_state=42),
        "K-Nearest Neighbors": KNeighborsClassifier(),
        "AdaBoost": AdaBoostClassifier(random_state=42),
        "Extra Trees": ExtraTreesClassifier(n_estimators=100, random_state=42),
        "LightGBM": LGBMClassifier(random_state=42)
     }

     # Train and evaluate models
     for name, model in models.items():
         print(f"Training {name}...")
         model.fit(X_train, y_train)
         y_pred = model.predict(X_val)
         accuracy = accuracy_score(y_val, y_pred)
         report = classification_report(y_val, y_pred)

         print(f"\n{name} - Accuracy: {accuracy}")
         print(f"\n{name} - Classification Report:\n{report}")
```

## 8. MODEL EVALUATION

```
Decision Tree - Accuracy: 0.9606666666666667

Decision Tree - Classification Report:
                           precision    recall  f1-score   support

               Automotive       0.98      0.98      0.98       190
                Baby Care       0.69      0.63      0.66        38
     Bags, Wallets & Belts       0.96      0.84      0.90        32
                 Clothing       0.98      0.99      0.99      1070
                Computers       0.92      0.95      0.94        99
                 Footwear       0.96      0.94      0.95       214
Home Decor & Festive Needs       0.92      0.91      0.92       149
                Jewellery       0.98      0.99      0.99       645
          Kitchen & Dining       0.90      0.93      0.91       109
     Mobiles & Accessories       0.94      0.94      0.94       159
         Pens & Stationery       0.80      0.62      0.70        56
          Tools & Hardware       0.97      0.97      0.97        70
    Toys & School Supplies       0.85      0.85      0.85        54
                   Watches       0.97      1.00      0.98       115

                  accuracy                           0.96      3000
                 macro avg       0.92      0.90      0.91      3000
              weighted avg       0.96      0.96      0.96      3000
```

```
Random Forest - Accuracy: 0.9683333333333334

Random Forest - Classification Report:
                           precision    recall  f1-score   support

               Automotive       0.99      0.98      0.99       190
                Baby Care       1.00      0.53      0.69        38
     Bags, Wallets & Belts       0.96      0.78      0.86        32
                 Clothing       0.96      1.00      0.98      1070
                Computers       0.96      0.95      0.95        99
                 Footwear       0.99      0.96      0.97       214
Home Decor & Festive Needs       0.95      0.99      0.97       149
                Jewellery       0.99      1.00      1.00       645
          Kitchen & Dining       0.96      0.97      0.97       109
     Mobiles & Accessories       0.96      0.96      0.96       159
         Pens & Stationery       0.91      0.55      0.69        56
          Tools & Hardware       0.99      0.96      0.97        70
    Toys & School Supplies       0.77      0.85      0.81        54
                   Watches       0.99      0.96      0.97       115

                  accuracy                           0.97      3000
                 macro avg       0.96      0.89      0.91      3000
              weighted avg       0.97      0.97      0.97      3000
```

```
Extra Trees - Accuracy: 0.9716666666666667

Extra Trees - Classification Report:
                           precision    recall  f1-score   support

               Automotive       0.99      0.98      0.99       190
                Baby Care       0.95      0.55      0.70        38
     Bags, Wallets & Belts       1.00      0.75      0.86        32
                 Clothing       0.96      1.00      0.98      1070
                Computers       0.95      0.95      0.95        99
                 Footwear       0.99      0.96      0.97       214
Home Decor & Festive Needs       0.95      0.99      0.97       149
                Jewellery       1.00      1.00      1.00       645
          Kitchen & Dining       0.98      1.00      0.99       109
      Mobiles & Accessories       0.97      0.96      0.97       159
          Pens & Stationery       1.00      0.59      0.74        56
          Tools & Hardware       1.00      0.96      0.98        70
    Toys & School Supplies       0.78      0.93      0.85        54
                   Watches       1.00      0.99      1.00       115

                  accuracy                           0.97      3000
                 macro avg       0.97      0.90      0.92      3000
              weighted avg       0.97      0.97      0.97      3000
```



| uniq_id | crawl_timestamp | product_url | product_name | pid | retail_price | discounted_price | image | is_FK_Advantage_pi | description | product_rating | overall_rating | brand | product_specificatio | predictions |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 4fb99d98225f415e7 | 2015-12-20 08:26:17 | http://www.flipkart.co | V&V ART Brass Brac | BBAE6NYHCDTEZJ | 470 | 423 | ["http://img6a.flixcar | FALSE | V&V ART Brass Brac | No rating available | No rating available | V&V ART | ["product_specificati | Jewellery |
| 4ea284c8d38b2ea9 | 2015-12-20 08:26:17 | http://www.flipkart.co | Kalpaveda Copper C | BBAEDFFKZJTY7SZ | 1200 | 1200 | ["http://img6a.flixcar | FALSE | Kalpaveda Copper C<br>Price<br>1.Attract And Emit D | No rating available | No rating available | Kalpaveda | ["product_specificati | Jewellery |
| ee6ce2c7045c5425 | 2015-12-20 08:26:17 | http://www.flipkart.co | Thelostpuppy Book ( | ACCEA4DZH6M5SFY | 2199 | 599 | ["http://img5a.flixcar | FALSE | Thelostpuppy Book (<br>Price<br>Designed to protect | No rating available | No rating available | Thelostpuppy | ["product_specificati | Mobiles & Accessori |
| e797ba3b5f2e2d1fd | 2015-12-20 08:26:17 | http://www.flipkart.co | Riana Copper Coppe | BBAEAXFQHHMF3E | 2499 | 649 | ["http://img5a.flixcar | FALSE | Riana Copper Coppe | 5 | 5 | Riana | ["product_specificati | Jewellery |
| f4d8d43858c8858c6 | 2015-12-20 08:26:17 | http://www.flipkart.co | Inox Jewelry Stainle | BBAECH63WYDG6T | 1629 | 1222 | ["http://img6a.flixcar | FALSE | Inox Jewelry Stainle<br>Price<br>Inox Jewelry Is A Poc | No rating available | No rating available | Inox Jewelry | ["product_specificati | Jewellery |
| 9b82f9fda38364972 | 2015-12-20 08:26:17 | http://www.flipkart.co | Thelostpuppy Book ( | ACCEA4DP7GZWHH | 2199 | 599 | ["http://img5a.flixcar | FALSE | Thelostpuppy Book (<br>Price<br>Designed to protect | No rating available | No rating available | Thelostpuppy | ["product_specificati | Mobiles & Accessori |
| 52f14c6349ca09801 | 2015-12-20 08:26:17 | http://www.flipkart.co | Ridhi Sidhi Collectic | BBAECF98YZGZBV | 575 | 400 | ["http://img5a.flixcar | FALSE | Ridhi Sidhi Collectic<br>Price<br>Lakh Brass Fabulous | No rating available | No rating available | Ridhi Sidhi Collectic | ["product_specificati | Jewellery |
| 8c0a0477e6027d16 | 2015-12-20 08:26:17 | http://www.flipkart.co | Thelostpuppy Book ( | ACCEA4DZHGHEMJ | 2199 | 599 | ["http://img5a.flixcar | FALSE | Thelostpuppy Book (<br>Price<br>Designed to protect | No rating available | No rating available | Thelostpuppy | ["product_specificati | Mobiles & Accessori |
| 231a26ee0bad0d8b | 2015-12-20 08:26:17 | http://www.flipkart.co | Ridhi Sidhi Collectic | BBAECF98BEW95BI | 575 | 400 | ["http://img5a.flixcar | FALSE | Ridhi Sidhi Collectic<br>Price<br>Lakh Brass Fabulous | No rating available | No rating available | Ridhi Sidhi Collectic | ["product_specificati | Jewellery |
| 33e8d05b2a6eb862 | 2015-12-20 08:26:17 | http://www.flipkart.co | Theskinmantra Slee | ACCE6SJGYPXKWV | 1499 | 549 | ["http://img5a.flixcar | FALSE | Theskinmantra Slee<br>Price<br>The HOT new Ipad s | No rating available | No rating available | Theskinmantra | ["product_specificati | Mobiles & Accessori |

Use the validation set to evaluate the model based on an accuracy score and a confusion matrix.

## 9. GENERATING PREDICTIONS

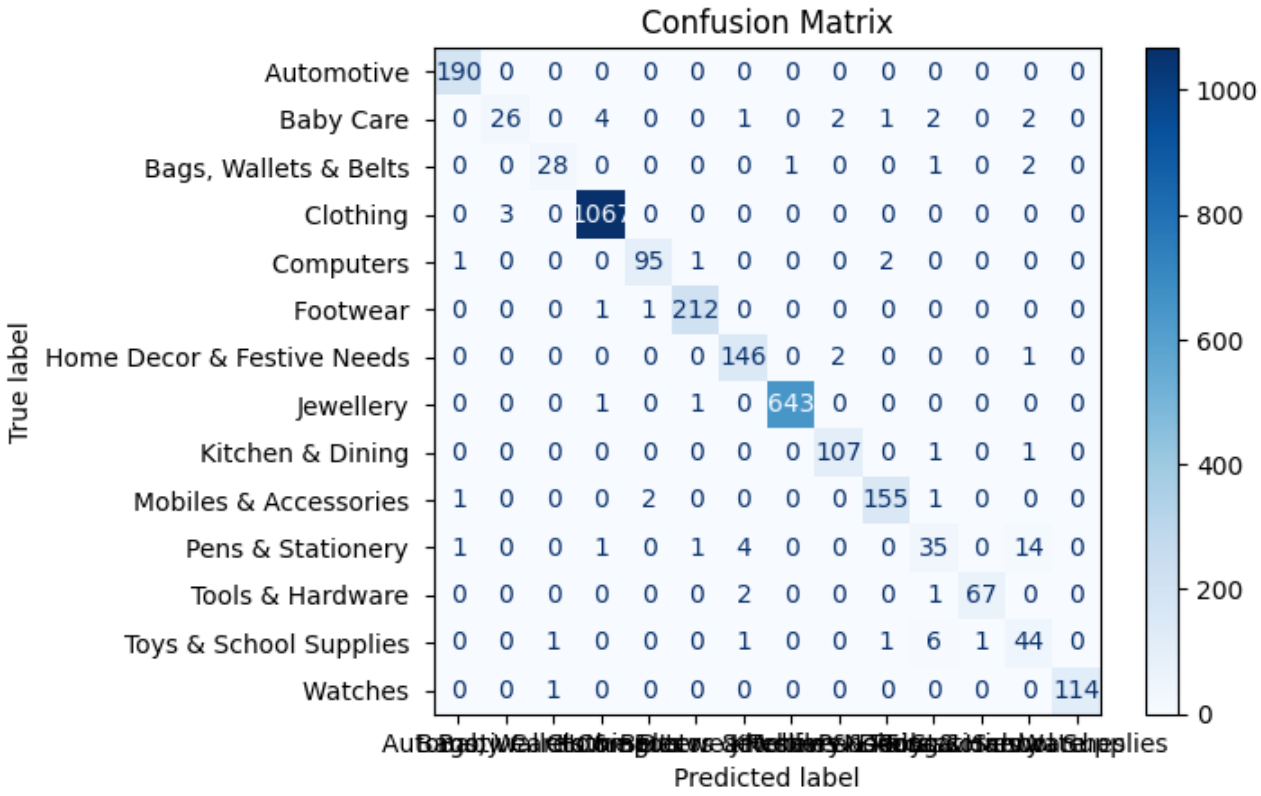
Test Data Preprocessing: Do all the preprocessing steps on test data.

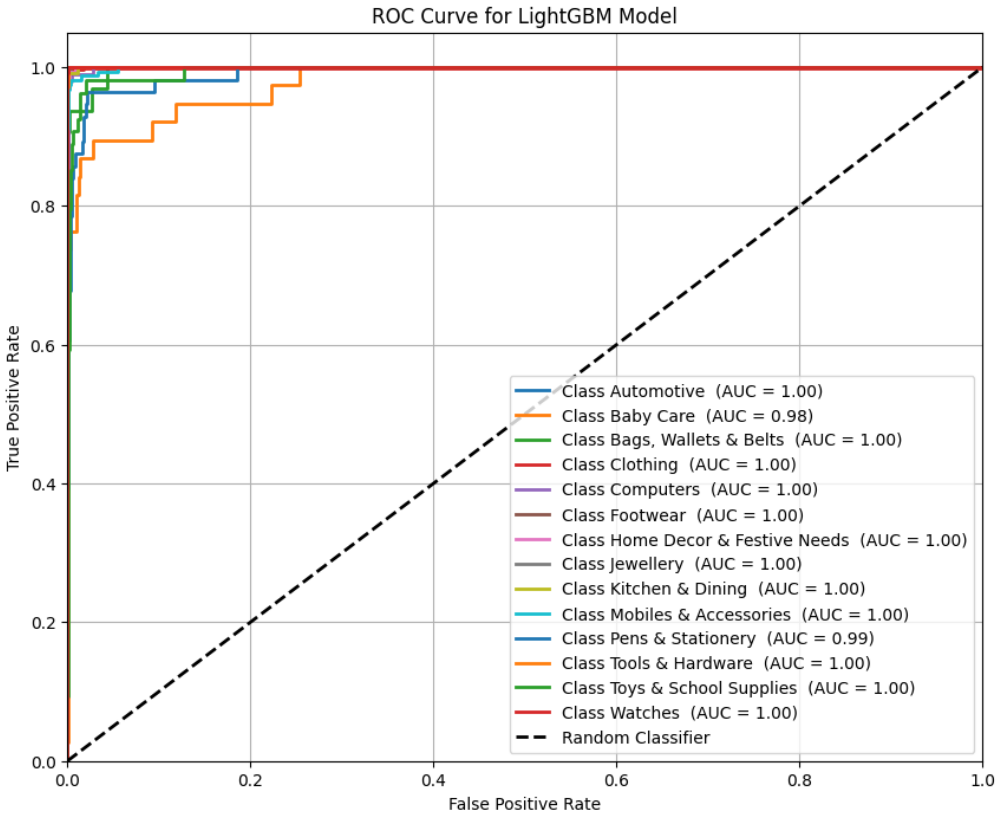
Generate Predictions: Generate the product category predictions of the test data using the above-trained model.

9. Save Predictions: These predictions would be stored in a CSV file.

## 10. CONFUSION MATRIX

The confusion matrix will plot the performance of the classification model against the actual and predicted categories.




ROC Curve for LightGBM Model
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Class Automotive (AUC = 1.00)
Class Baby Care (AUC = 0.98)
Class Bags, Wallets & Belts (AUC = 1.00)
Class Clothing (AUC = 1.00)
Class Computers (AUC = 1.00)
Class Footwear (AUC = 1.00)
Class Home Decor & Festive Needs (AUC = 1.00)
Class Jewellery (AUC = 1.00)
Class Kitchen & Dining (AUC = 1.00)
Class Mobiles & Accessories (AUC = 1.00)
Class Pens & Stationery (AUC = 0.99)
Class Tools & Hardware (AUC = 1.00)
Class Toys & School Supplies (AUC = 1.00)
Class Watches (AUC = 1.00)
Random Classifier