WEB SCRAPING - EDA & MACHINE LEARNING

Navaneetha Krishnan

DA&DS -JULY 2025

Project Overview

- Scrape TV product data from Flipkart
- Clean & store it in a MySQL database
- Perform exploratory data analysis EDA
- Cluster prices and build classification models
- Perform Models and evaluate

Data Collection

- Target: Flipkart TV search results multiple pages
- Tools: Python, requests, BeautifulSoup
- Fields scraped: TV name, price, offers, raw info, ratings, deal type
- Pages scraped: 1 to 52 configurable

Data Cleaning

- Removed duplicates and null values
- Extracted structured fields from text TV_Name and Info_Raw: name,size_cm,size_inch,display,edition,model_id, sound_output, warranty_years
- Cleaned numeric fields : Removed symbol like - ₹,%.
- Converted fields to proper data types (int/float)
- Standardized unknown values and formatted categorical text

Storage & Retrieval

- Saved the cleaned TV_dataset as a CSV file
- Imported the CSV into MySQL database: webscrap, table: tv_dataset
- Retrieved the dataset using SQLAlchemy into Python
- Performed EDA, clustering, and machine learning models on the retrieved data
- Ensured smooth data flow from scraping , cleaning ,storage,analysis

Exploratory Data Analysis

- Distribution of Ratings, Prices, Offers
- Size variations cm & inch
- Display types like LED , QLED , OLED .
- Warranty, sound output, reviews & ratings
- Correlations between numerical features
- Price vs size & price vs display type

Key Insights

- Larger screen sizes inch & cm show significantly higher prices
- TVs with QLED, OLED, Neo QLED, Mini LED are priced much higher the others
- Items with more ratings usually have more reviews
- Most TVs offer 1-year warranty
- Very few offer more than 1 year
- Data distributions show market segmentation
- Price distribution shows three natural groups: Low, Mid, High
- Few extremely expensive TVs Premium OLED/QLED models

Clustering

- Used Elbow Method to identify the optimal number of clusters
- Formed three clusters: Low, Medium, High priced TVs
- Scaled price data to improve clustering accuracy
- Cluster patterns matched real pricing trends

Models & Performance

- Logistic Regression, SVM, KNN,XGBoost,Random Forest
- Low accuracy in Logistic Regression & SVC performed 20% & 66%
- KNN performed well Achieved 88% accuracy
- XGBoost struggled Predicted mostly one class Accuracy 66% but misleading
- Random Forest performed the best Handled non-linear relations, mixed features, and SMOTE-balanced data 99% accuracy Consistently classified all 3 price clusters correctly Further improved with hyperparameter tuning

Conclusions

- Random Forest is the best-performing model Achieved 99% accuracy for Price_Cluster prediction
- Useful business applications Price tier detection Low, Medium, High
- Can be deployed using Flask or Streamlit for live predictions
- Automate daily scraping & MySQL updates
- Add more features (brand, specs, resolution, etc.)
- Tune XGBoost and compare advanced models

THANK YOU