Hackathon Topic:
Ecommerce Product Categorization

Problem Statement:

In the rapidly evolving world of eCommerce, accurate product categorization is crucial for ensuring seamless customer experiences, reducing search friction, and increasing product discoverability. However, the sheer volume of diverse products poses a significant challenge. Current classification systems struggle to handle ambiguities, unconventional naming conventions, and multi-language data. This hackathon aims to address these challenges by inviting participants to create innovative solutions that enhance product categorization efficiency, accuracy, and scalability.

Solution:

Develop a text classification model that categorizes products with maximum accuracy based on description of the product.

CK **by Chetan Kittali**

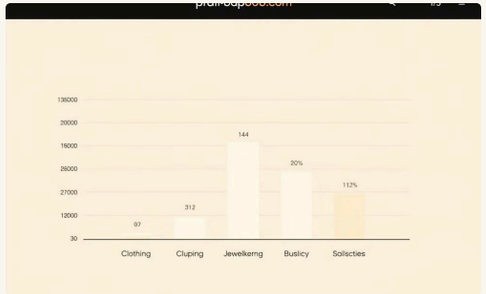
Problem Solving Approach

1. Load the dataset
2. Perform Exploratory Data Analysis to extract valuable business insights
3. Using NLTK for Data preprocessing and Feature Engineering.
4. Transform Textual Descriptions to numerical features using techniques like TF-IDF or word2Vec.
5. Perform train-test split using stratify sampling
6. Train and evaluate various suitable Machine Learning and Deep Learning models if needed.
7. Fine tune them if necessary, using Hyperparameter Tuning
8. Get our predictions.
9. Finally, we will compare the models to get the best performing model.

Key Insights

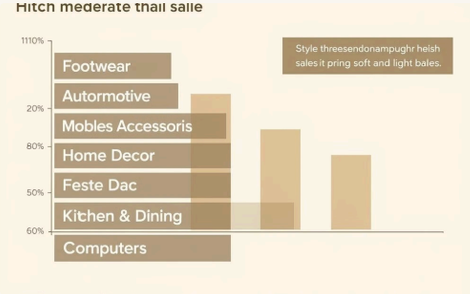
Price Correlation

Higher retail prices correspond to higher discount prices.



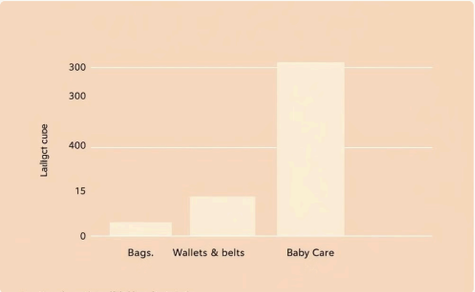
Top Sellers

Clothing and Jewellery are the highest selling product categories.



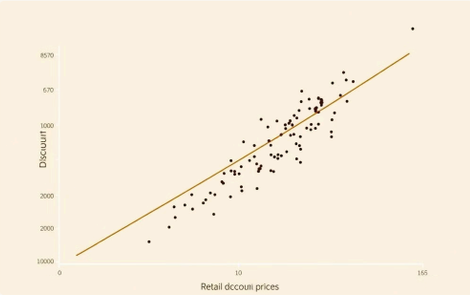
Moderate Sales

Moderate sales in various categories.



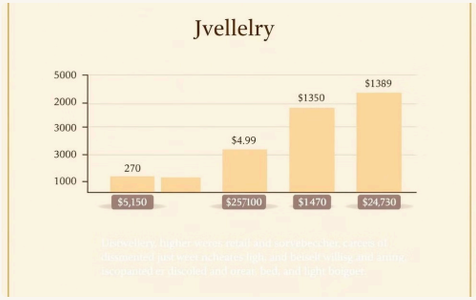
Low Sellers

Bags, Wallets & Belts, and Baby Care show the lowest sales.



Retail vs Discount

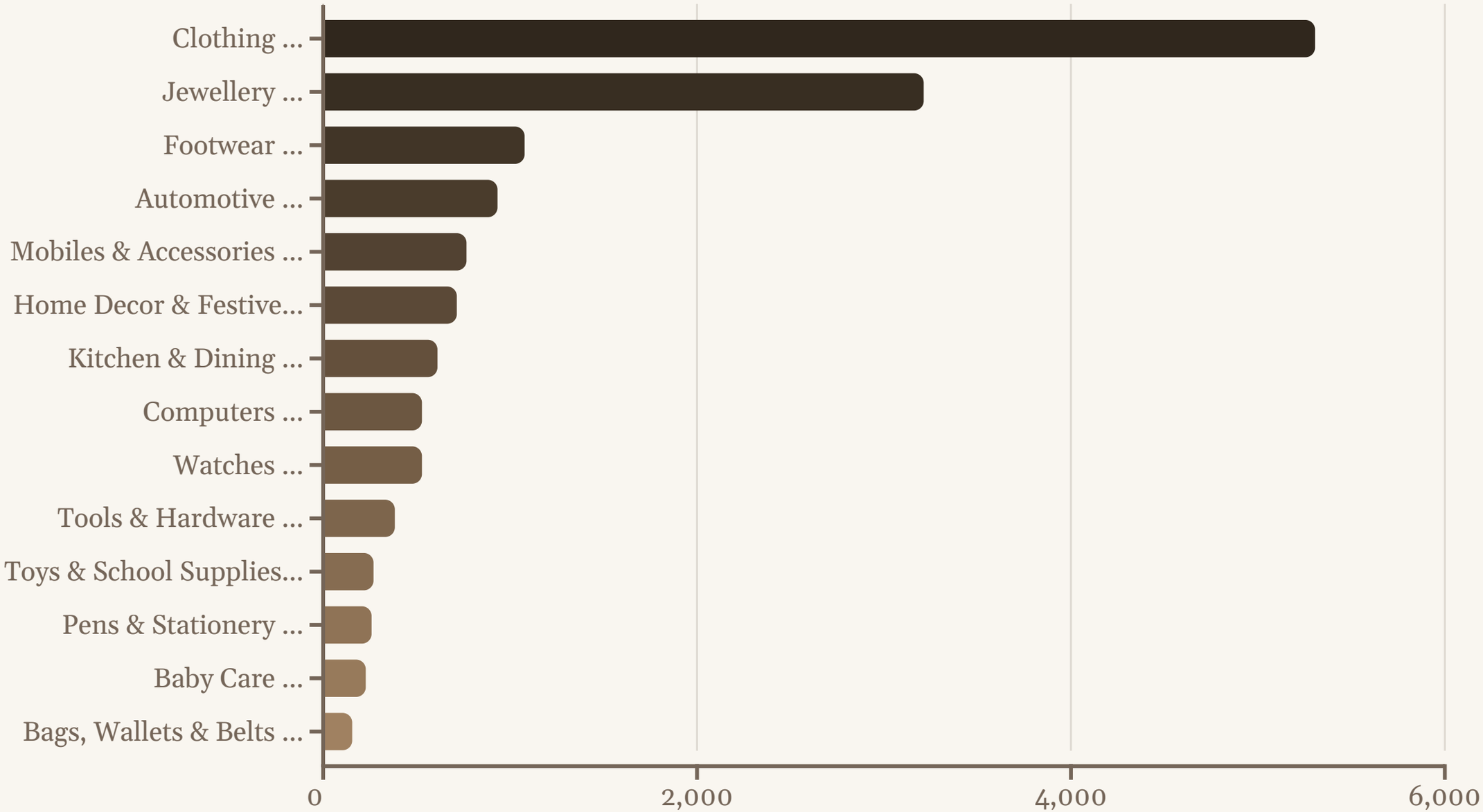
Higher retail price means higher discount price.



Jewellery Pricing

Jewellery has the highest average retail and discounted price.

Frequency counts of Product Category Tree



CHALLENGES

Problem: Class Imbalance

Solution: Employing techniques like stratified sampling and SMOTE oversampling to balance the dataset and improve model accuracy.

Problem: Natural Text Preprocessing and Converting to numeric data

Solution: Using NLTK and techniques like TF-IDF or word2Vec for preprocessing and feature extraction

Model Selection and Training

Models	Train Accuracy	Test Accuracy
SVM	99.78%	98.19
Logistic Regression	99.52%	98.04
Random Forest	99.90%	97.85
Decision Tree	99.90%	96.29
KNN Classifier	99.50%	95.11

Evaluation and Final Results

MODEL: Support Vector Machine (SVC) has achieved the highest accuracy among other.

98%

Accuracy

Our model achieved an impressive accuracy of 98% on a diverse dataset of product descriptions after fine tuning

Benefits of Accurate Categorization

1

Improved Customer Experience

Seamless browsing and navigation, making it easier for customers to find the products they're looking for.

2

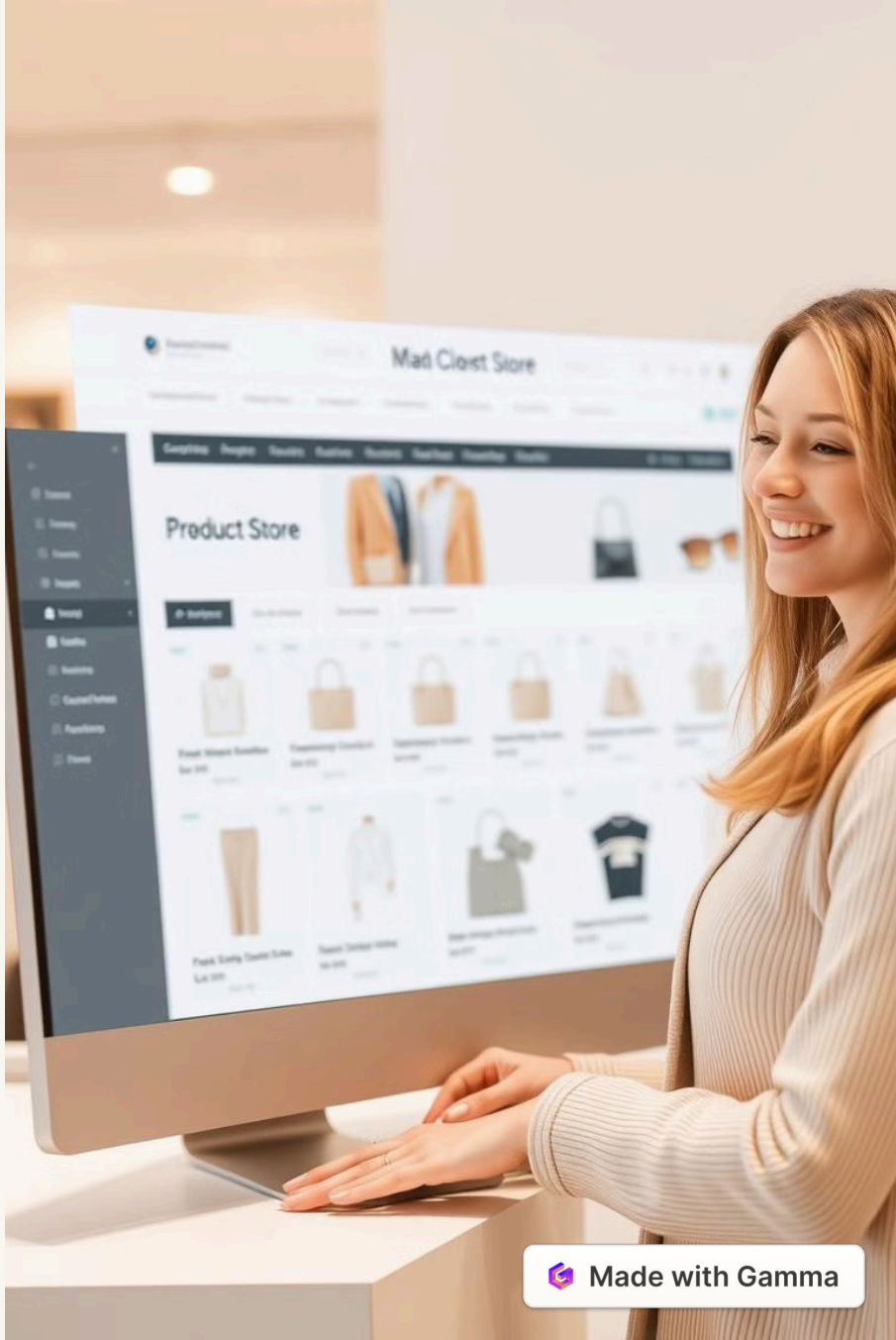
Increased Product Discoverability

Customers are exposed to a wider range of relevant products, leading to increased sales and customer satisfaction.

3

Reduced Search Friction

Customers can quickly find the products they want, reducing frustration and abandoned searches.



Conclusion and Next Steps

The development of accurate product categorization models is essential for the success of eCommerce platforms. Future research may explore the use of advanced deep learning models and the integration of visual features, such as product images.