ML Special Assignment

- Munmun Sangtani, 20BEC073
- Siddharth Agarwal, 20BEC119

What we have developed?

E-Commerce Product Price Prediction

We aim to develop a pricing algorithm that automatically suggests optimal prices for products listed on online platforms.

Why we have developed this?

By providing users with suggested prices based on product attributes and user-inputted text descriptions, enhance the buying and selling experience can be enhanced and the volume of transactions can be increased.

How we have developed this?

we have presented a comparative study of two machine learning algorithms, LightGBM and Ridge Regression, for e-commerce price prediction tasks. Additionally, we have evaluated their performance using the RMSLE metric, which is particularly useful for predicting logarithmic values.

Flow of Program:

Importing dataset
Preparing the Corpus for Analysis
Explore Training Set
Create three new features from Categories (Main, Sub1, Sub2)
Create CSR_Matrix & Merge the Sparse Matrices
Data Cross validation - Holdout Method
Train with LGBM
Train with Ridge Regression
Predict on test set

Dataset

Train-id / test-id	id of the listing
Name	the title of listing.
condition-id	the condition of the items by seller
Category-name	category of listing
Brand-name	brand of listing
Price	the price for the item sold. Our Target variable we want to predict, which is in USD.
Shipping	1 if shipping fee paid by seller and 0 by buyer
Item-description	the full description of item.

Algorithms Used

Algorithms Used

- **LGBM** : It is a gradient boosting framework that uses a tree-based learning algorithm and is designed to be efficient, scalable, and accurate for large-scale datasets.
- rmsle: 0.657144374711778 7
- **Ridge Regression:** It is a regularization technique that adds a penalty term to the regression model's cost function to reduce the impact of multicollinearity on the model's coefficients.
- rmsle: 0.4704993399548869 6

Results and Analysis

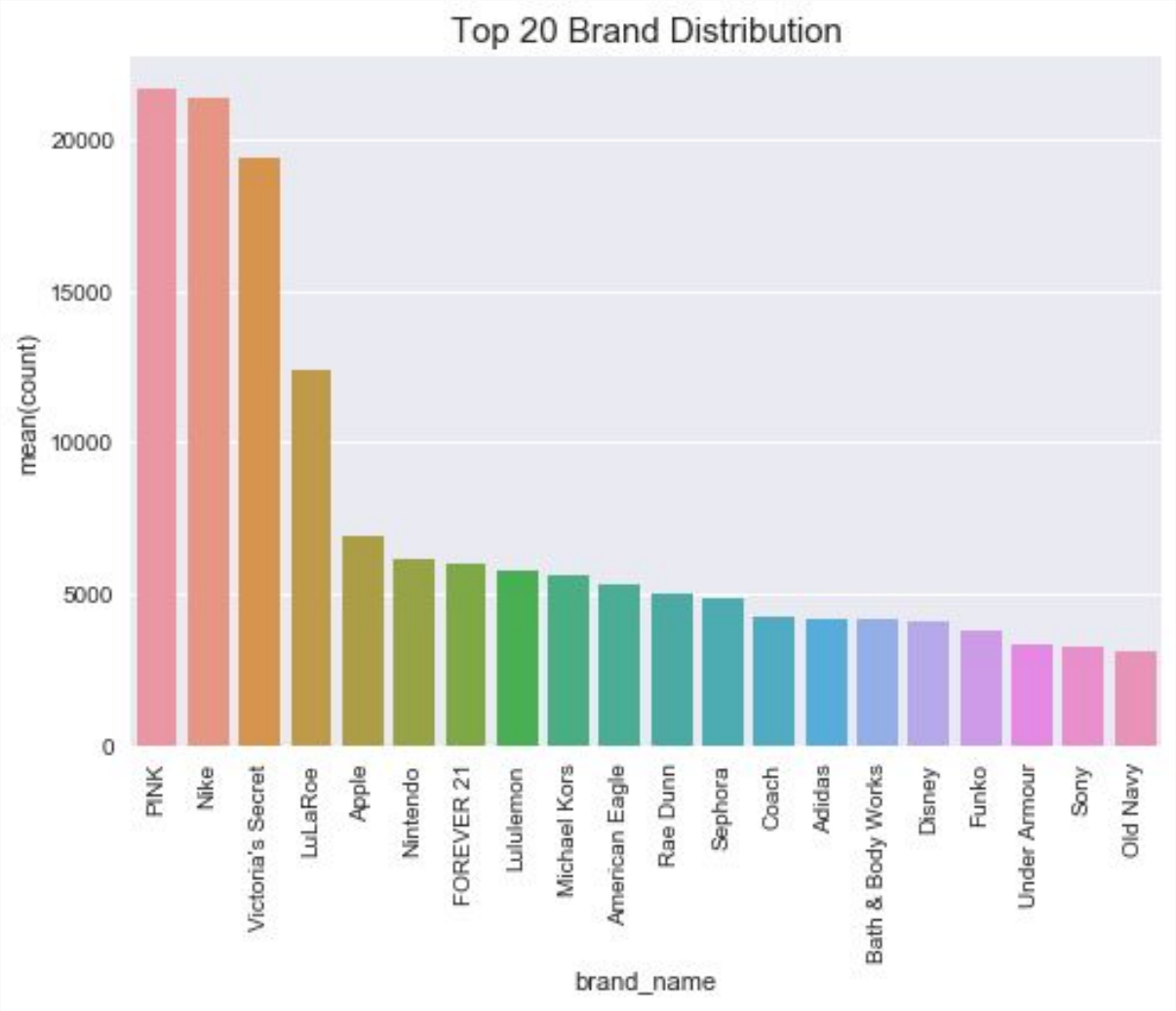
Price Distribution

- The mean price in the dataset is 26 Dollars
- The median price in the dataset is 17 Dollars
- The max price in the dataset is 2000 Dollars
- Due to the skewed dataset, the median price is a more reliable price to gauge off of.


Price Distrubtion With/Without Shipping
Price With Shipping
Price With No Shipping
Normalized Samples
0.04
0.03
0.02
0.01
0.00
0
50
100
150
200
250
Price

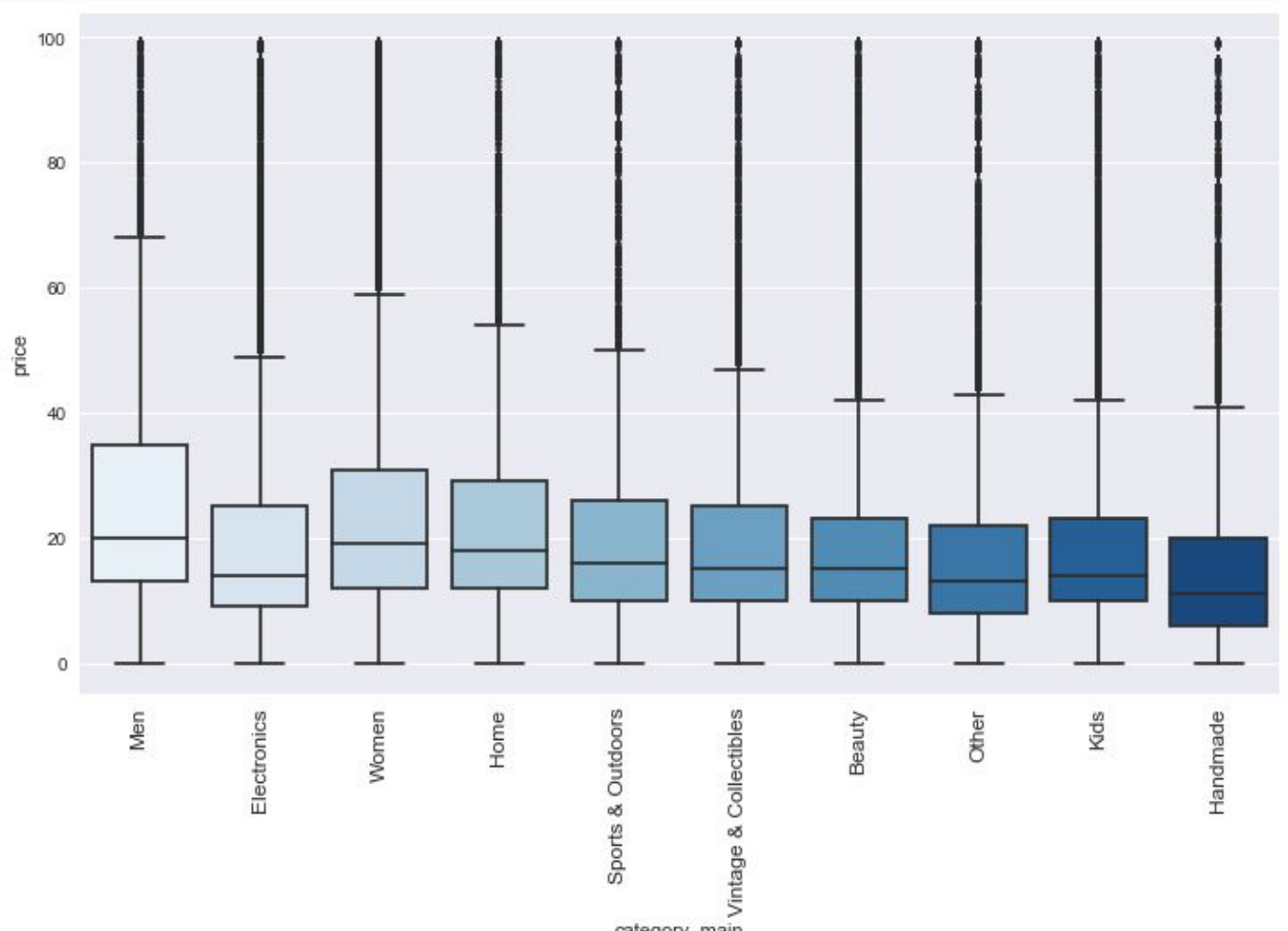
Top 20 brand distribution

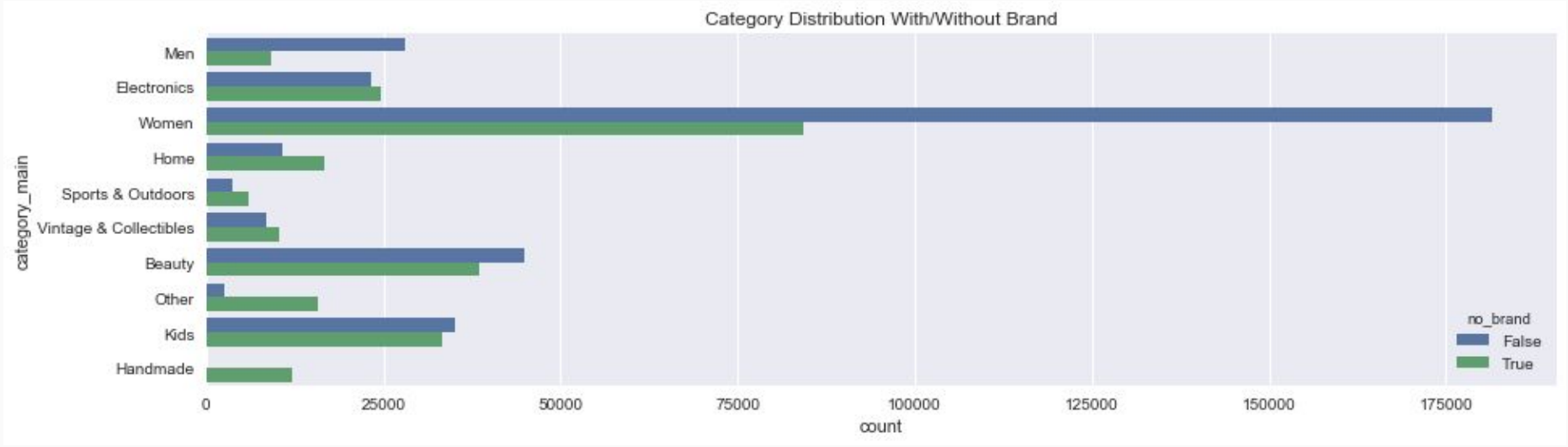
Majority of the top brands are clothing brands and electronics. PINK and Victoria Secret are among the top 3 brands and are typically towards female customers.


Top 20 Brand Distribution
mean(count)
20000
15000
10000
5000
0
PINK
Nike
Victoria's Secret
LuLaRoe
Apple
Nintendo
FOREVER 21
Lululemon
Michael Kors
American Eagle
Rae Dunn
Sephora
Coach
Adidas
Bath & Body Works
Disney
Funko
Under Armour
Sony
Old Navy
brand_name

Main Category

It is seen that majority of the distribution is taken by women and beauty. They take 56% of the distribution. The prices are evenly distributed across all categories. The Men category the only one that averages out the most.


price
100
80
60
40
20
0
Men
Electronics
Women
Home
Sports & Outdoors
Vintage & Collectibles
Beauty
Other
Kids
Handmade


Category Distribution With/Without Brand
category_main
Men
Electronics
Women
Home
Sports & Outdoors
Vintage & Collectibles
Beauty
Other
Kids
Handmade
0
25000
50000
75000
100000
125000
150000
175000
count
no_brand
False
True

Effect of word count

Effect of word count of item description on prices: It was observed that from about 0-300 words, there is a positive linear relationship. After that a gradual negative relationship was seen. It drops at about the 1000 word point. Overall, the word count did not have any significant role in deciding the price of an item


Price Distribution
price
140000
120000
100000
80000
60000
40000
20000
0
Samples
0
50
100
150
200
250
Price