Skincare Product Pricing & Recommendation System

Ritika Verma

Skincare Recommendation System

Mission

➢ **Predict** the price of products based on features
➢ **Features contribution** to the final price of a product
➢ **Recommendation** based on Customer's features

ML Vision

➢ **Scalability & Adaptability**: Models can easily be updated with new data to reflect changing market trends.
➢ **Data Insights** & **Decision Making**
➢ **Feature Importance**
➢ **Automation** & **Efficiency**

Table of contents

Dataset Introduction

●**Product** contains all beauty products (over 8,000) from the Sephora online store, including product and brand names, prices, ingredients, ratings, and all features.

●**User Reviews** (about 1 million on over 2,000 products) of all products from the Skincare category, including user appearances, and review ratings by other users

01

Price Analysis
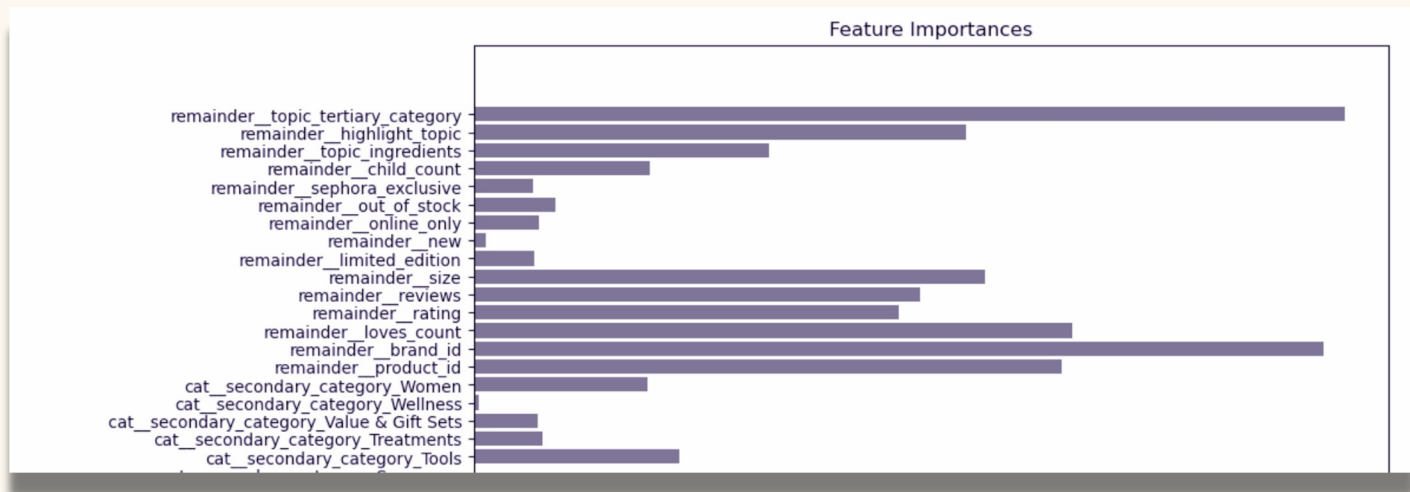
Data Preprocessing

Shape   [8494, 27]

product_id	product_name	brand_id	brand_name	loves_count	rating	reviews	size	variation_type	ingredients	price_usd	limited_edition	new
473671	Fragrance Discovery Set	6342	19-69	6320	3.6364	11.0	1.7	Unknown	['Capri Eau de Parfum:', 'Alcohol Denat. (SD A...	35.0	0	0
473668	La Habana Eau de Parfum	6342	19-69	3827	4.1538	13.0	3.4	Size + Concentration + Formulation	['Alcohol Denat. (SD Alcohol 39C), Parfum (Fra...	195.0	0	0
473662	Rainbow Bar Eau de Parfum	6342	19-69	3253	4.2500	16.0	3.4	Size + Concentration + Formulation	['Alcohol Denat. (SD Alcohol 39C), Parfum (Fra...	195.0	0	0

ColumnTransformer

Imputation, One-hot, Scaling

```
df.topic_ingredients.unique()
array([10, 54, 74,  0, 68,  9, 15, 28, 26, 29, 14, 41, 13, 66, 75, 37,  5,
       57,  2, 64, 48, 61,  7, 33, 67, 20, 72, 62, 55, 18, 17, 31, 27, 42,
       70, 25, 49, 63, 38, 77, 44, 22, 51, 35, 34, 16, 56,  4, 30,  1, 19,
        6, 40, 50, 32, 71, 45,  8, 78, 69, 58, 65, 21, 76, 39, 43, 47, 73,
        3, 60, 23])
```

LdaModel(n_topics = 80)
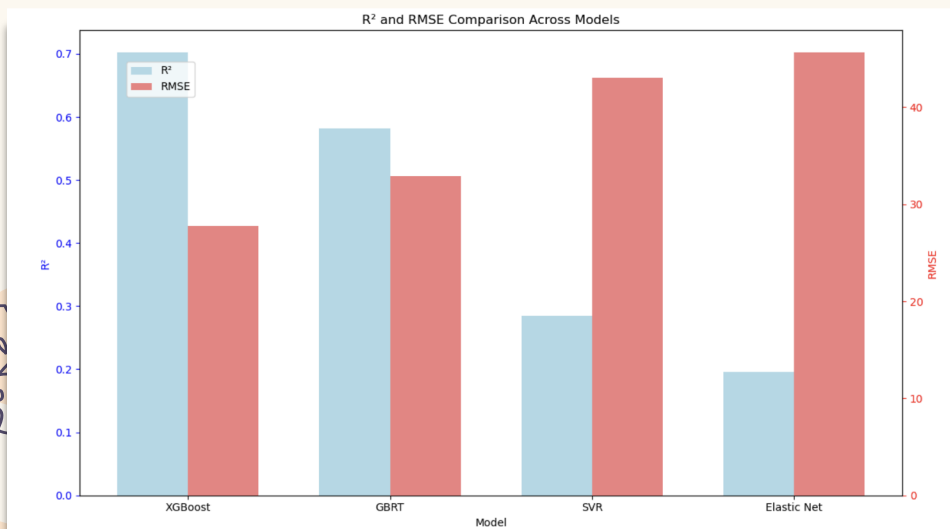
.feature_importances_(Regression)

Feature Importances

- remainder__topic_tertiary_category
- remainder__highlight_topic
- remainder__topic_ingredients
- remainder__child_count
- remainder__sephora_exclusive
- remainder__out_of_stock
- remainder__online_only
- remainder__new
- remainder__limited_edition
- remainder__size
- remainder__reviews
- remainder__rating
- remainder__loves_count
- remainder__brand_id
- remainder__product_id
- cat__secondary_category_Women
- cat__secondary_category_Wellness
- cat__secondary_category_Value & Gift Sets
- cat__secondary_category_Treatments
- cat__secondary_category_Tools

Regression Models
- ➢ **Linear Regression (ElasticNet)**
- ➢ **Random Forest**
- ➢ **Gradient Boosting**
- ➢ **XGBoost**
- ➢ **Neural Networks**

Evaluation Metrics (rmse & r2)

➤ For businesses, minimizing RMSE can directly translate to more accurate pricing strategies
➤ R² values suggest that the model captures most of the factors affecting prices

R² and RMSE Comparison Across Models

```
# Compile the model
optimizer = tf.keras.optimizers.Adam(learning_rate=1e-3)
model.compile(loss="mse", optimizer=optimizer, metrics=["RootMeanSquaredError"])
```

R² for XGBoost model: 0.7162772537035915
MSE for XGBoost model: 735.2925845545727
RMSE for XGBoost model: 27.11627895848862
MAD for XGBoost model: 14.958018596657038
MAPE for XGBoost model: 0.37343394056654633

Training and Validation MSE

XGBoost works best Price Prediction

—Someone Famous

03

Reviews Analysis

Data Preprocessing

```
df.shape
(1044434, 19)
```

Columns

| author_id	rating	is_recommended	review_text	review_title	skin_tone	eye_color	skin_type	hair_color	product_id	brand_name	price_usd
1741593524	5	1.0	I use this with the Nudestix "Citrus Clean Bal...	Taught me how to double cleanse!	NaN	brown	dry	black	P504322	NUDESTIX	19.0
31423088263	1	0.0	I bought this lip mask after reading the revie...	Disappointed	NaN	NaN	NaN	NaN	P420652	LANEIGE	24.0
5061282401	5	1.0	My review title says it all! I get so excited ...	New Favorite Routine	light	brown	dry	blonde	P420652	LANEIGE	24.0
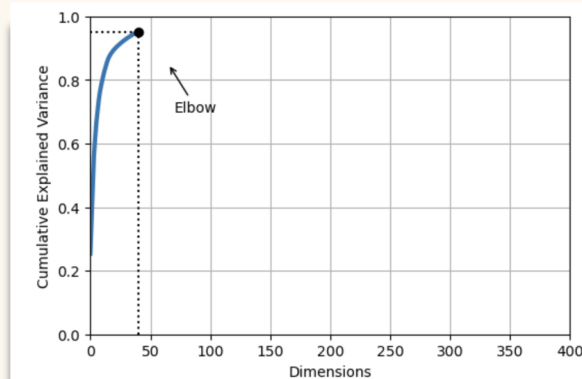
PCA(n_compement = 0.95)

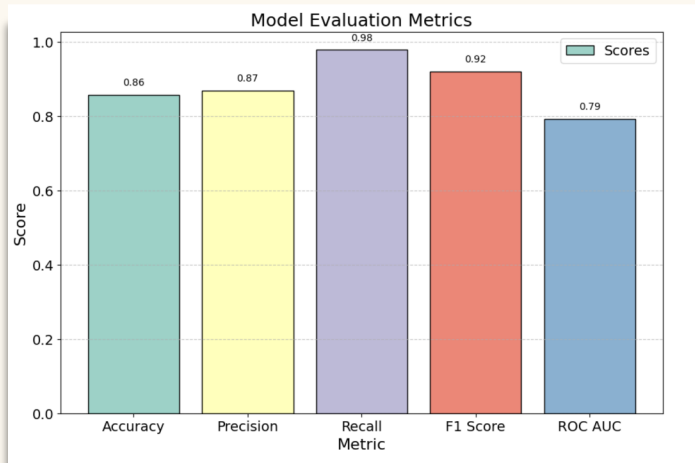# .feature_importances_

Classification Models

➢ **Logistic Regression**
➢ **Random Forest**
➢ **Gradient Boosting**
➢ **XGBoost**
➢ **Neural Networks**

Feature Importances from RandomForestClassifier
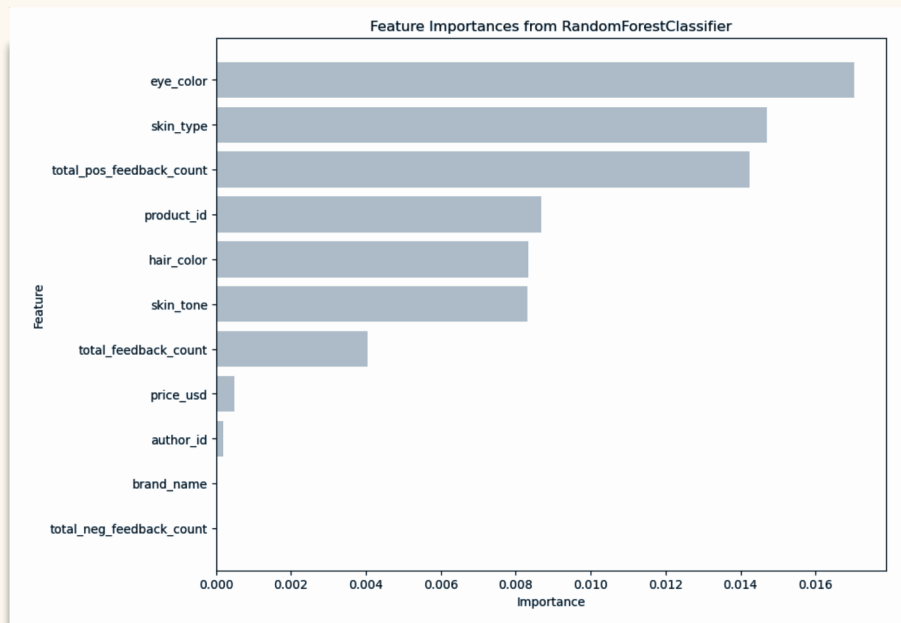
Evaluation Matrices

Best Model:
WeightedEnsemble_L3

roc_auc

➢ ROC AUC is less sensitive to class imbalance.
➢ true positives while minimizing the instances where uninteresting items are recommended false positives is key to user satisfaction and engagement

XGBoost performs best

XGBoost: {'xgb_classifier__gamma': 0,
'xgb_classifier__learning_rate': 0.05,
'xgb_classifier__max_depth': 6, 'xgb_classifier__n_estimators':
500, 'xgb_classifier__subsample': 0.7}
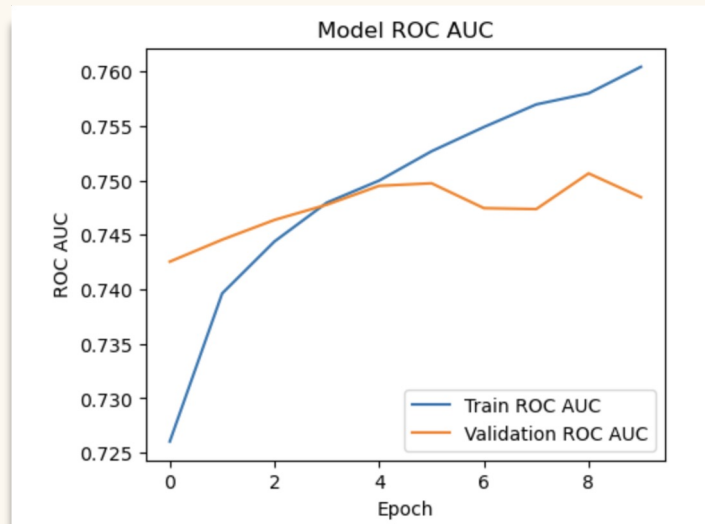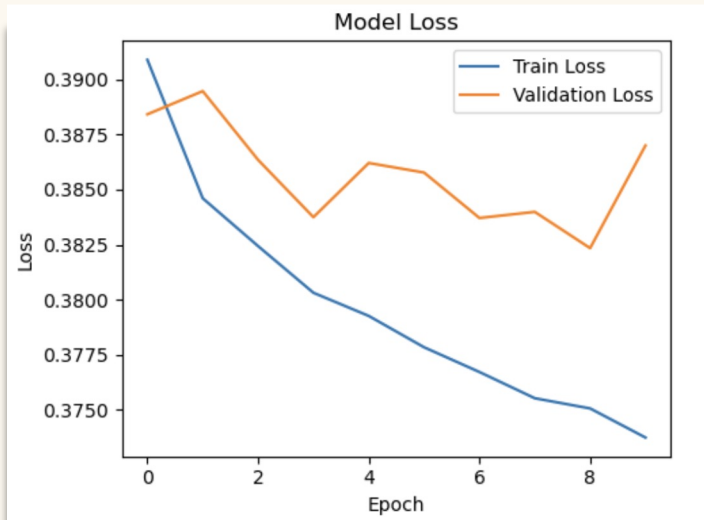Best score XGBoost: 0.7771145221516665
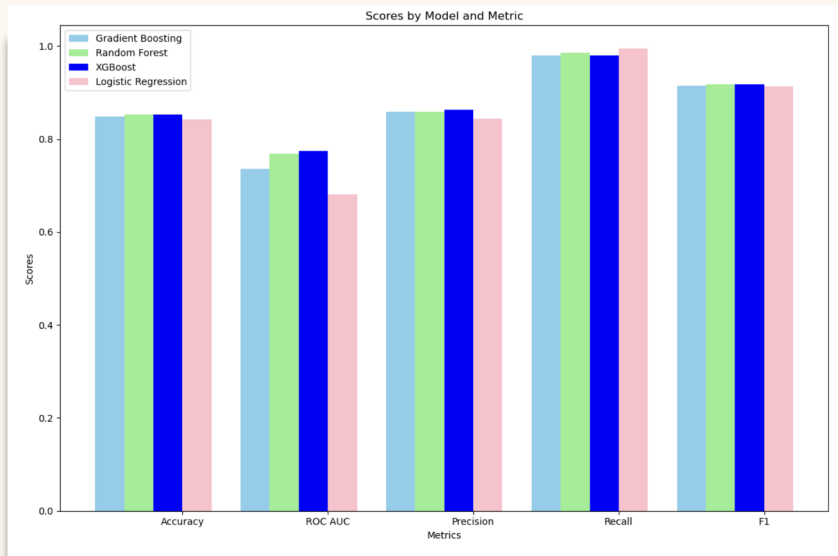
Neural Network Performance

```python
for _ in range(10):
    model.add(Dense(100))
    model.add(LeakyReLU(alpha=0.2))

# Add the output layer with a single unit for binary classification
model.add(Dense(1, activation='sigmoid'))
```

"XGBoostClassifier works best for Classification for Recommendation"

—Someone Famous

Key accomplishments

Product

- ❖ Few key factors for product prices are:
  - ➤ Brand, size
  - ➤ Love Count
  - ➤ Reviews provided
- ❖ Price category of the products can be predicted based on top key features

Recommendation

- ❖ Key factors for product recommended by the customers are:
  - ➤ Skin Type
  - ➤ Product
  - ➤ Positive Feedback Counts
- ❖ XGBoost Classifier is best model for our dataset.

Next Steps

Clustering

➤ **Objective**: Group users with similar preferences or behaviors.
➤ **Method**: Apply clustering algorithms (e.g., K-Means, DBSCAN) on user features such as demographic information, browsing history, purchase history, ratings given, etc.
➤ **Application**: After identifying clusters, recommend items that are popular or highly rated within a user's cluster. This approach assumes users in the same cluster will have similar tastes or needs.

Thank You

Stay Beautiful !!