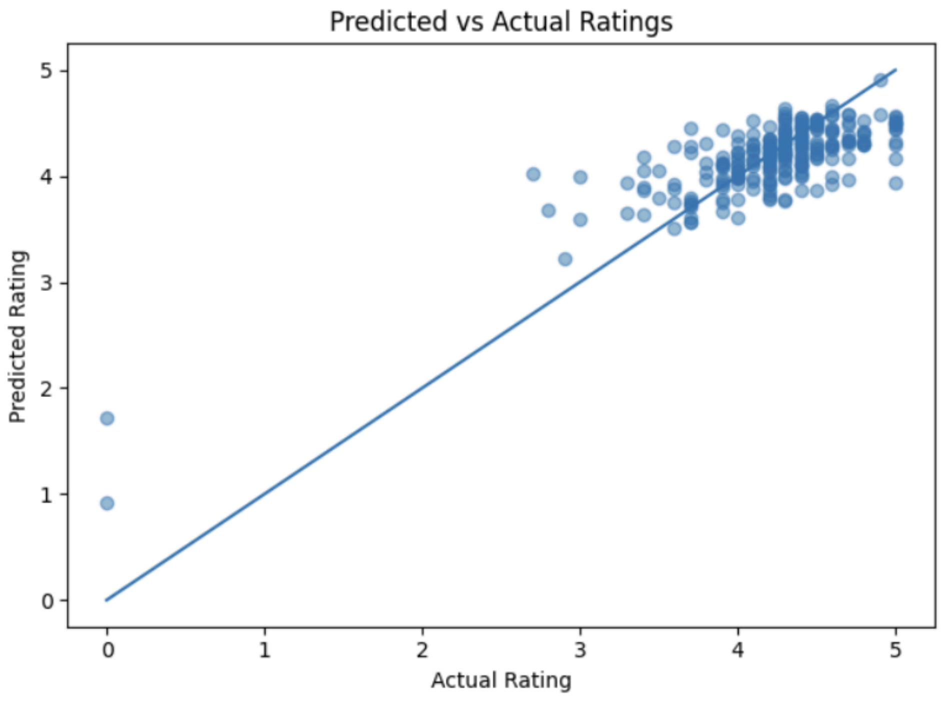
Model Development and Analysis Workflow

1. Data Preparation and Cleaning
The raw dataset of cosmetic formulas was first cleaned and reformatted so that each ingredient per product label was accurately represented. Special attention was given to standardizing ingredient names, handling missing values, and normalizing quantities, ensuring that the dataset was fully structured for model training. During this stage, several outliers were identified that could negatively impact model performance.

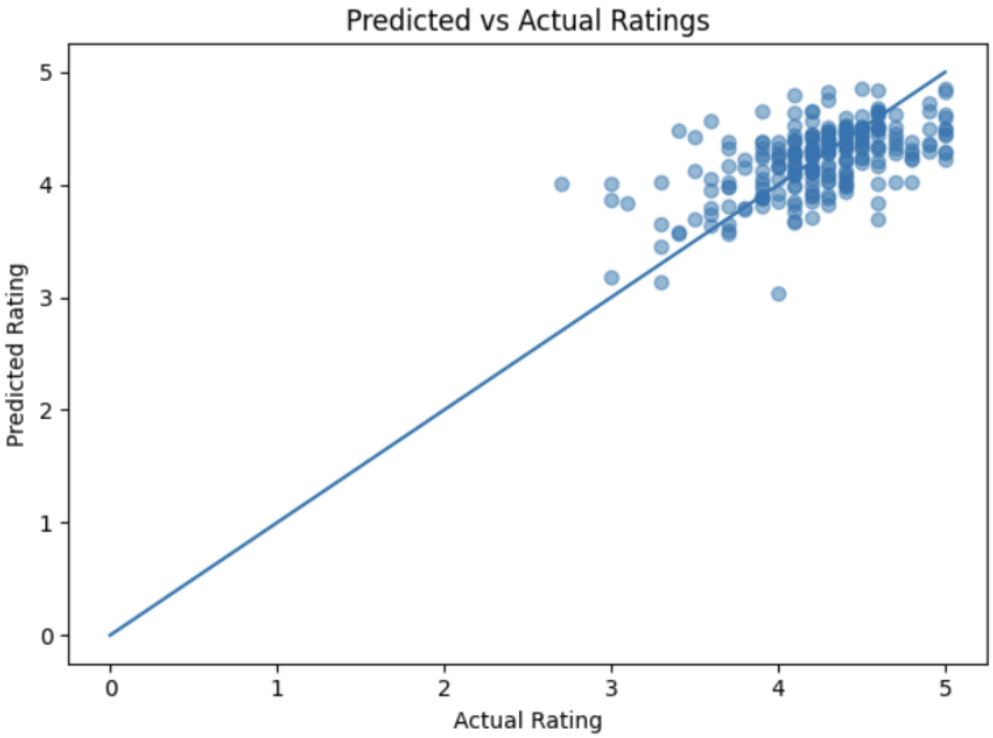
2. Feature Engineering
Key features were derived for each formula, including the average rank of ingredients (`Avg_Ingredient_Mean_Rank`) and the standard deviation of ingredient ranks (`Avg_Ingredient_Std_Rank`). These features captured the overall influence and variability of ingredients within a product.

3. Model Training and Evaluation
The model was trained using statistical features of the ingredients to predict product ratings on a 0–5 scale. Initial training resulted in a prediction error of approximately ±0.4. After further cleaning, reformatting, and handling outliers, the model's performance improved to ±0.34. Calibration steps were applied to adjust for the model's tendency to slightly underestimate ratings, ensuring more accurate predictions across the dataset.

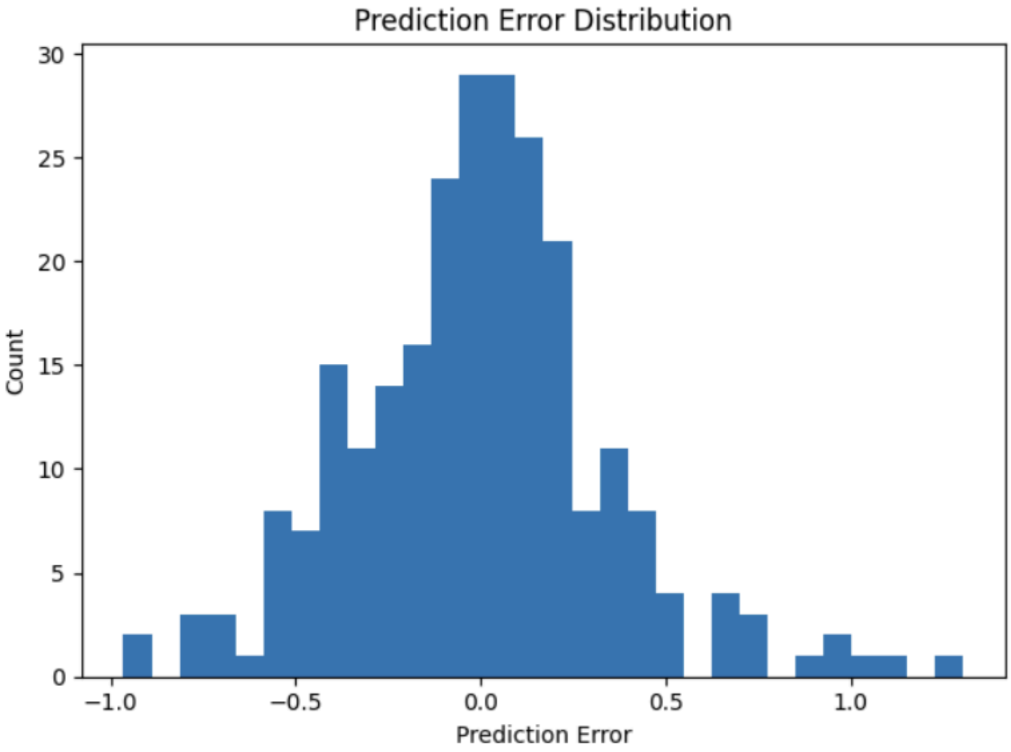
Before Removing the Outliers



After Removing the Outliers



Distribution of Outliers (No Concerning Trends Noticed)



4. Ingredient Impact Analysis

Using the trained model, we quantified the influence of individual ingredients:

Avg_Ingredient_Mean_Rank

- Top 5 ingredients that increase rating:
  - Packette Oryza Sativa Starch (+0.0277)
  - Cymbidium Grandiflorum Flower Extract (+0.0193)
  - Jasminum Sambac Extract (+0.0193)
  - Amaranthus Squalane Oil (+0.0153)
  - Capric Triglycerides (+0.0153)

- Top 5 ingredients that decrease rating:
  - Calcium Sulfate (−0.0515)
  - Sodium Acrylate Sodium Acryloyldimethyl Taurate (−0.0171)
  - Pleiogynium Timorense Fruit Extract (−0.0171)
  - Podocarpus Elatus Fruit Extract (−0.0171)
  - Grevillea Speciosa Flower Extract (−0.0171)

Avg_Ingredient_Std_Rank

- Top 5 ingredients that increase variability:
  - Calcium Sulfate (+0.0266)
  - Hordeum Vulgare Seed Extract (+0.0115)
  - Citrus Aurantium Dulcis Peel Powder (+0.0111)
  - Sodium Palmate Cocoate / Palm Kernelate (+0.0094)
  - Tetrasodium Pyrophosphate (+0.0093)

- Top 5 ingredients that decrease variability:
  - Cucurbita Pepo Seed Oil (−0.0407)
  - Passiflora Incarnata Oil (−0.0292)
  - Argania Spinosa Oil Organic (−0.0292)
  - Titanium Dioxide Sunscreen (−0.0264)
  - Coenzyme Q10 (−0.0258)

Interestingly, the model showed that skin type has minimal impact on predictions, which aligns with the fact that most commercial cosmetic products are designed to be compatible with all skin types.

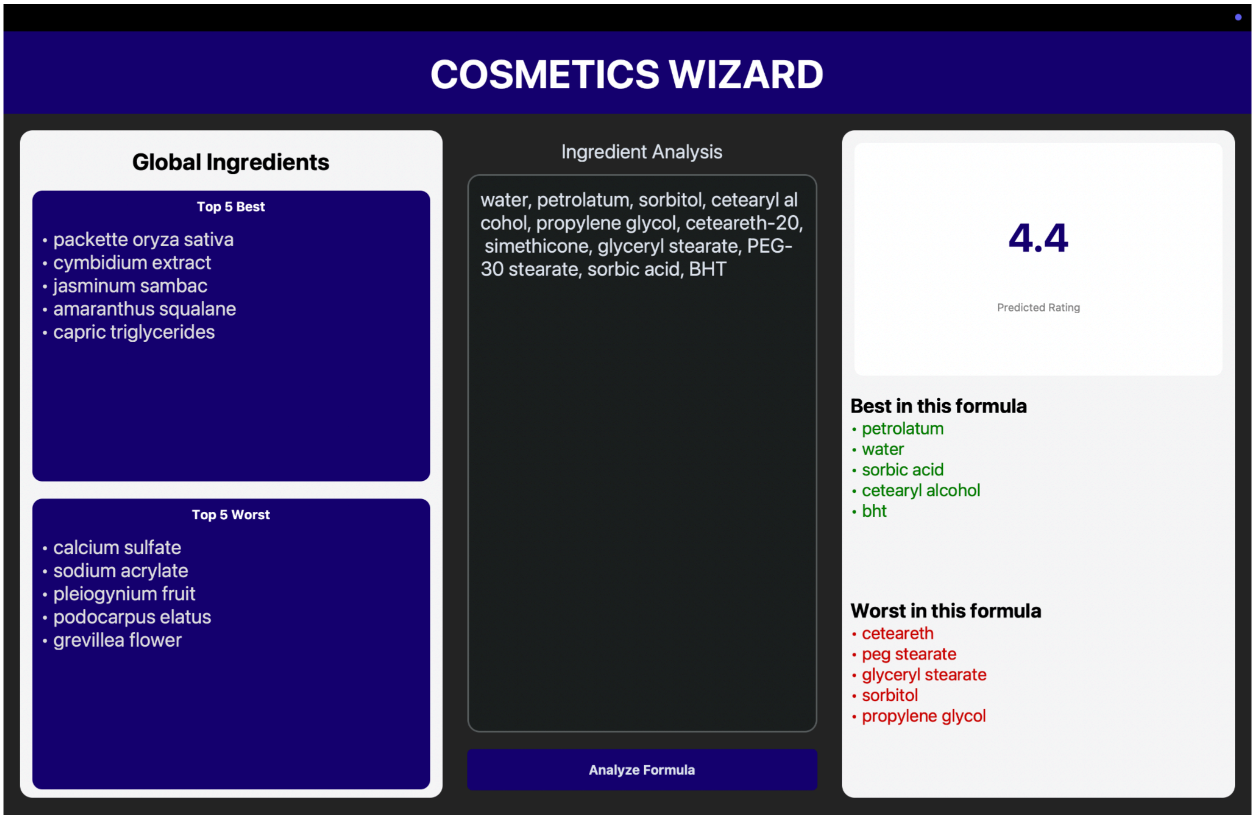
5. Before Outliers vs. After Outliers

Two scatterplots were generated to compare predictions: one including outliers and one after outlier removal. These visualizations demonstrate how outliers influenced the model and highlight the improvement in prediction consistency after data cleaning.

6. Average Prediction Error
A histogram of prediction errors was produced, showing an approximately normal distribution centered near zero. This indicates that the model's predictions are balanced and unbiased after cleaning and calibration.

7. Deployment Interface
A CustomTkinter-based interface was built to create a user-friendly application. Users can input a cosmetic formula and optionally specify skin type. The model outputs a safety rating (0–5) along with the top 5 ingredients positively and negatively impacting the score.

The Final Product with Demonstrated Use



Global Ingredients

Top 5 Best

- packette oryza sativa
- cymbidium extract
- jasminum sambac
- amaranthus squalane
- capric triglycerides

Top 5 Worst

- calcium sulfate
- sodium acrylate
- pleiogynium fruit
- podocarpus elatus
- grevillea flower

water, petrolatum, sorbitol, cetearyl alcohol, propylene glycol, ceteareth-20, simethicone, glyceryl stearate, PEG-30 stearate, sorbic acid, BHT

4.4

Predicted Rating

Best in this formula

- petrolatum
- water
- sorbic acid
- cetearyl alcohol
- bht

Worst in this formula

- ceteareth
- peg stearate
- glyceryl stearate
- sorbitol
- propylene glycol