Identification of Hazardous Ingredients in Cosmetic Products for Reformulation

Olivia Folly-Gah
11/22/2024

Introduction

The cosmetics industry faces increasing scrutiny regarding the safety and sustainability of its products. Harmful ingredients, associated with cancer, reproductive harm, and environmental risks, remain a critical concern for consumers and regulators alike. My project aims to identify these hazardous chemicals in cosmetic products and analyze their suitability for coily, curly hair and melanin-rich skin. Using data from the California Department of Public Health (CDPH), I seek to uncover insights that can guide safer reformulation practices, particularly for consumers with these specific needs. This report details the progress made, challenges encountered, and future steps to advance this work.

Summary

In the last two weeks, I completed essential tasks, including data preprocessing, analyzing product discontinuation trends, and examining hazardous chemical concentrations in various product subcategories. By resolving inconsistencies and preparing the dataset, I enabled a deeper analysis of products containing hazardous chemicals. This analysis revealed spikes in product discontinuations, particularly around 2010 and 2016, which likely correlate with regulatory changes or shifts in consumer demand for safer cosmetics. Furthermore, I identified makeup products as having the highest concentrations of hazardous chemicals among the top 10 subcategories, emphasizing the need for targeted reformulation efforts in this area.

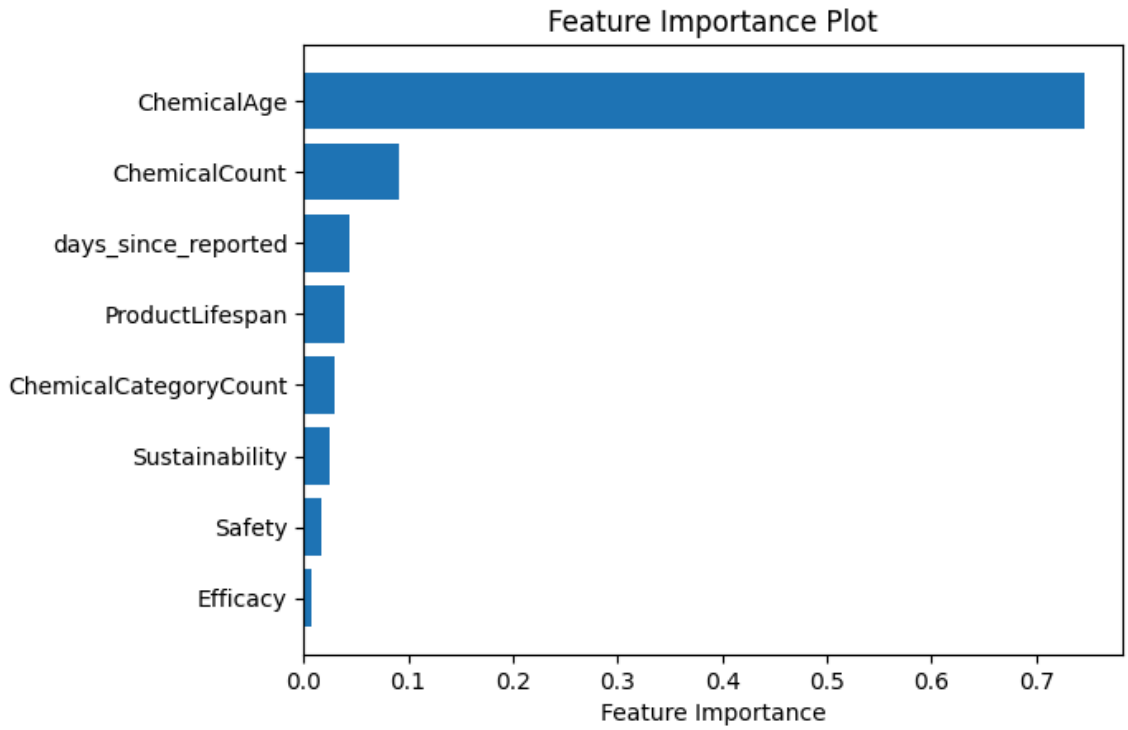

Figure 1. Feature Importances in Hazardous Ingredient Reformulation Model

The chart above illustrates the feature importances, highlighting the most influential variables for predicting the likelihood of hazardous ingredient reformulation in cosmetic products. Based on the model, the most significant feature is **ChemicalAge**, with an importance score of **0.7456**. This suggests that the age of chemicals within a product plays a critical role in the reformulation decision. Other features, such as **ChemicalCount** and **days_since_reported**, also contribute meaningfully, with scores of **0.0912** and **0.0443**, respectively. Conversely, **Efficacy** has the least influence on the model, with a score of **0.0073**, indicating minimal relevance to the reformulation process.

This ranking emphasizes the importance of chemical age and product composition in identifying products that may require reformulation due to hazardous ingredients. Such insights can guide industry efforts to prioritize reformulation in a more effective and targeted manner.

Progress and Milestones

I have made significant strides in data preprocessing, addressing inconsistencies and preparing the dataset for advanced analysis. Key milestones include identifying trends in product discontinuations and analyzing hazardous chemical concentrations across different subcategories. These efforts have laid the groundwork for understanding industry responses to regulatory changes and consumer demands. By highlighting makeup products as a high-risk category, I

have also identified a critical focus area for further research and reformulation efforts, particularly regarding their impact on coily, curly hair and melanin-rich skin.

Problem-Solving and Challenges

One major challenge involved managing an imbalanced dataset, as reformulated products were significantly outnumbered by non-reformulated ones. To address this, I employed stratified sampling during train-test splitting, ensuring balanced representation of both classes. Another challenge was aligning discontinuation trends with product lifecycle timelines, which required standardizing date formats and cleaning redundant entries. Lastly, distinguishing overlapping subcategories presented difficulties, but I implemented clear criteria to ensure consistent and meaningful analysis.

Technical Depth and Accuracy

The project employed statistical and machine learning methods to analyze trends and predict reformulation likelihood. For instance, time series analysis uncovered peaks in product discontinuation rates, while frequency-based methods ranked subcategories by hazardous chemical concentrations. I developed a machine learning model using logistic regression, achieving an accuracy score of **0.7** and an **ROC AUC score of 0.7**, to predict the likelihood of reformulation. These results reflect the model's precision but will be further validated with additional data to ensure reliability.

Future Plans and Goals

In the coming weeks, I plan to continue leveraging the dataset I am currently scrapping with my industry partner. I will return to my original objective of identifying which ingredients are hazardous or beneficial for coily, curly hair and melanin-rich skin. By refining the model's performance and applying additional resampling techniques to address dataset imbalance, I aim to enhance the accuracy of the reformulation predictions. These efforts will help identify safer and more suitable ingredients for these specific hair and skin types, which will drive more targeted reformulation practices within the cosmetics industry.

Conclusion

This phase of the project has provided key insights into the presence of hazardous ingredients and reformulation trends in the cosmetics industry. By combining data analysis, statistical methods, and machine learning, I have highlighted critical safety challenges and identified areas for improvement, particularly for coily, curly hair and melanin-rich skin. The next phase will build on these findings to deliver actionable recommendations for enhancing safety and sustainability in cosmetic formulations tailored to these consumer needs.