Data Science Proposal Report

Group 10

2391007 Minjeong Kim, 2391016 Suahn Lee, 2391017 Yoonji Lee

* Originality, clarity, and validity of the problem.

탈모를 겪는 인구는 매해 증가하고 있음. 그에 따라 탈모 용품 시장 또한 커지는 중. 왜냐면 탈모는 다른 질병과는 다르게 병원에서 치료를 받으려고 하지 않는 사람 다수. 스스로 식습관을 바꾸거나 헤어 관리 용품 등을 사용함으로써 탈모를 예방, 극복하려는 경향. [Androgenetic Alopecia Study Reveals Unmet Treatment Needs (dermatologytimes.com)](https://www.dermatologytimes.com/view/androgenetic-alopecia-study-reveals-unmet-treatment-needs)

우리는 이런 탈모인들을 겨냥한 서비스를 제공하려고 한다.

한 개인이 본인이 대머리가 될지 말지를 예측하기는 매우 어려움. 그걸 우리가 많은 양의 데이터를 토대로 가능하게 해줄 것.

병원에 가지 않고도 진단할 수 있음! 효과적으로.

탈모는 초기에 치료 부담이 더 적고 초기에 발견할수록 치료 효과가 높기 때문에 우리가 당신 대머리 될 것 같으니까 병원에 가서 치료 받아야한다 라고 해줄 것임. (병원에 방문 해서 약물치료 받으라고 권고하고 홈케어로는 이런 것들이 가능하다하면서 관련 제품, 보조제 추천)

* Quality of the dataset, model, and evaluation methods used.

Let me introduce Suahn. He is a 21-year-old university student. He has a bad diet, always sleeps late, never does work out, and doesn’t take any supplements. He is always stressed because of his assignments, smokes alot, and doesn’t have hair care routine. Suahn is at risk of becoming bald. Despite being aware of this possibility, Suahn hesitates to visit the hospital for treatment. Suahn feels embarrassed about seeking help for hair loss. Also, she believes that it might not actually lead to baldness.

If Cheolsoo continues this way, he will increasingly stress over his thinning hair, his confidence in his appearance will plummet, and he may even experience depression. By the time Cheolsoo recognizes the seriousness of the situation and seeks medical help, it would be too late, with hair follicles already having lost their function, leaving hair transplant surgery as the only option.

Fortunately, Suahn came across Group 10's Hair Loss Detection service. Hair Loss Detection evaluates whether a user is at risk of becoming bald and emphasizes the importance of visiting a hospital for those who need it. Additionally, it recommends home care methods that can be used alongside medical treatment. Thanks to Hair Loss Detection, Suahn began treatment in the early stages of hair loss, successfully avoiding the fate of becoming bald!

1. Business Value (1 slide)
   1. Background

According to the National Health Insurance Service of Korea, nearly 250,000 people sought medical treatment for "pathological alopecia" in 2021. "Pathological alopecia" refers to hair loss due to conditions such as dermatitis or scarring. Additionally, many people experience hair loss without a specific cause. A survey conducted by Embrain in early 2023 among 1,000 adults found that 17.2% of those in their 20s, 28.4% of those in their 30s, 29% of those in their 40s, and 33.3% of those in their 50s reported experiencing hair loss.

* KB Securities estimated the Korean domestic hair loss market to be worth 4 trillion won in 2023
* it is projected that the global market for hair loss-related products and treatments will reach approximately 400 trillion won by 2024.
* Medication treatment is only possible if the hair follicles are still alive
* The sooner the treatment is started, the more effective it will be
* When hair loss becomes severe, hair transplant surgery becomes the only viable option.
* Hair loss can also add to the emotional burden.

A service that aligns with these social currents is needed.

* 1. Problem Definition & Value Proposition

Problem Definition:

predict the likelihood of becoming bald and induce users to visit hospitals and buy products that are relevant to their hair condition

Value:

- aim to provide personalized product recommendations

- can keep track of the user’s hair condition

- helpful when getting diagnosed for alopecia (since users have records of the progress)

1. Available Data (1 slide)
   1. Data Acquisition

found it at kaggle(카글)

데이터셋 설명

<https://www.kaggle.com/datasets/amitvkulkarni/hair-health>

Dataset Overview

This dataset encompasses data on potential contributors to baldness in individuals. Each entry represents a distinct person, with columns detailing factors spanning genetics, hormonal fluctuations, medical ailments, treatments, nutrient deficiencies, stress levels, age, hair care practices, environmental exposures, smoking habits, weight fluctuations, and the presence or absence of baldness.

Column Descriptions:

1. Genetics: Indicates a familial history of baldness (yes or no).

2. Hormonal Changes: Reflects whether the individual has undergone hormonal shifts (yes or no).

3. Medical Conditions: Enumerates specific ailments linked to baldness like Alopecia Areata, Thyroid Problems, Scalp Infection, Psoriasis, Dermatitis, etc.

4. Medications & Treatments: Lists drugs or therapies potentially causing hair loss, such as Chemotherapy, Heart Medication, Antidepressants, Steroids, etc.

5. Nutritional Deficiencies: Details deficiencies like Iron, Vitamin D, Biotin, Omega-3 fatty acids, etc. are associated with hair loss.

6. Stress: Indicates stress level (low, moderate, or high).

7. Age: Denotes individual age.

8. Poor Hair Care Habits: Indicates negligent hair care practices (yes or no).

9. Environmental Factors: Notes exposure to environmental elements linked to hair loss (yes or no).

10. Smoking: Specifies a smoking habit (yes or no).

11. Weight Loss: Indicates significant weight reduction (yes or no).

12. Baldness (Target): binary variable indicating baldness presence (1) or absence (0).

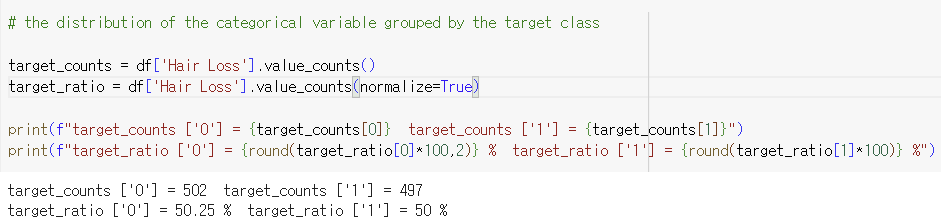
Purpose of the Dataset:

This dataset is designed for exploratory analysis, modeling, and predictive analytics endeavors, aiming to decipher the interplay between diverse factors and the probability of baldness in individuals.

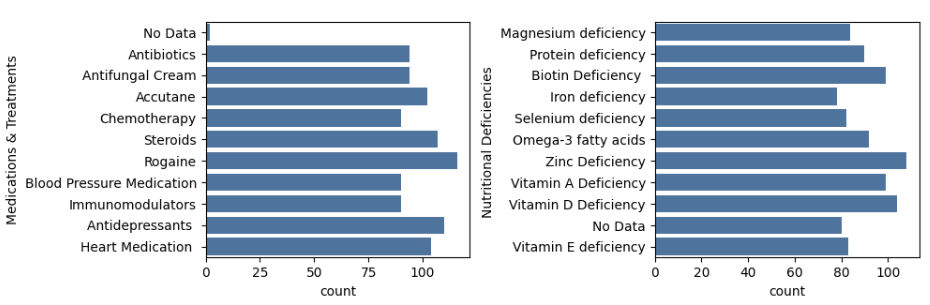
* 1. Data Review

data preprocessing/visualization

* visualization
* train\_test\_split 진행함



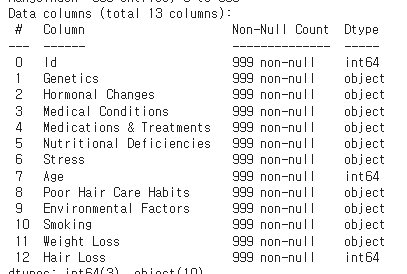
* target 반반으로 나눠져 있음



* 대부분의 column이 비슷한 count
* All other columns are also uniformly distributed
* “No Data” will be replaced by mode values

1. Formulation (2 slides)
   1. Input & Output Variables

column description



* 1. Model Selection

We blended following 5 methods :

* Decision Tree Classifier
* K-Nearest Neighbor
* AdaBoost
* Logistic Regression
* Latent Dirichlet Allocation

=> We opted for these methods because most of the variables are categorical

* 1. Model Validation

We chose to do K-Fold Cross Validation

K-Fold Cross Validation :

A validation method that systematically changes the subsets to measure the model's performance across all data

This method is time-consuming compared to other validation methods. However, we still have opted for it due to the relatively small size of our dataset.

Model Evaluation

* Why we selected F2 Score to test our model?
  + F2 score assigns more weight to recall.
  + In the given context, the mislabeling of a user as having a low likelihood of baldness (False Negative) carries a greater magnitude of loss compared to mislabeling them as having a high likelihood of baldness (False Positive).

1. Insight & Action (1 slide)
   1. Business Actions to Convert Insights into Value

* Display the user’s probability of becoming bald

Recommendation to visit a dermatology clinic

Personalized product recommendation

\*If users have a probability of 25% or higher, they are all considered to have the potential for becoming bald, and we recommend that they visit a dermatology clinic for treatment from a dermatologist. We will also provide personalized home-care product recommendations.

75~100% Assessed as severe hair loss

50~75% Assessed as hair loss at a certain stage of progression

25~50% Assessed as early-stage hair loss

0~25% - Assessed as a normal state

Because the likelihood of baldness is low, recommending only products without suggesting to visit dermatology clinics. (Users might incur unnecessary expenses if they visit dermatology clinics)