STA 141C: Hair Loss Prediction

Mehul Tailang, Riyaadh Bukhsh, Talha Shafik, Nirmal Kaluvai

March 18, 2024

Abstract

In this study, we aim to decipher the relationship between various health factors and the probability of baldness in individuals. By conducting prediction analysis using several methods we begin understanding the origins of baldness more deeply, subsequently facilitating the creation of personalized interventions to improve patient outcomes through tailored strategies.

Background

Hair loss, medically known as alopecia, affects millions of individuals worldwide. It can occur due to various reasons, ranging from genetic predisposition to environmental factors. The common forms include androgenetic alopecia, alopecia areata, and telogen effluvium. Understanding the basis of hair loss is crucial for effective treatment and management. Genetics play a significant role in hair loss, following a hereditary pattern that is influenced by androgen levels. Research indicates that specific genes related to hair growth and hormone regulation are key contributors. External factors such as stress, nutritional deficiencies, and exposure to chemicals or pollutants can also lead to hair loss. Lifestyle choices, including smoking, poor diet, and inadequate hair care practices, have been linked to exacerbated hair thinning and loss. Various medical conditions, including thyroid disorders, autoimmune diseases, and hormonal imbalances, can cause hair loss. Additionally, treatments like chemotherapy can lead to significant hair loss, known as anagen effluvium. Noting all these various hair loss factors our project aims to investigate the multifaceted aspects of hair loss, focusing on identifying the predominant causes in different demographics.

Data Description

The primary objective is to model the likelihood of baldness based on a holistic set of predictors. These include, but are not limited to, genetic predispositions, hormonal fluctuations, underlying medical conditions, and lifestyle dynamics.

Dataset Overview

The dataset encompasses 999 individual records, each detailed with 12 significant variables that are postulated to influence hair loss. These variables are systematically categorized as follows:

Genetics: Hereditary factors contributing to hair loss.
Medical Conditions: Underlying health issues that may precipitate hair thinning or loss.
Nutritional Deficiencies: Impact of diet and nutrient intake on hair vitality.
Age: Correlation between age and hair density changes.
Environmental Factors: External conditions like pollution or climate affecting hair quality.
Weight Loss: The association between significant weight change and hair loss.
Hormonal Changes: Variations in hormonal levels affecting hair growth and retention.
Medications & Treatments: Effects of pharmaceuticals and therapeutic procedures on hair health.
Stress: Influence of psychological stress on hair condition.
Poor Hair Care Habits: Consequences of inadequate hair care practices.
Smoking: The potential detrimental effects of smoking on hair health.
Hair Loss: The outcome variable indicating the degree or probability of hair loss.

The dataset's comprehensive nature allows for an in-depth analysis of aspects of baldness, facilitating the development of a predictive model with nuanced insights into the various factors.

Exploratory Data Analysis

The exploratory data analysis uncovered a nearly even distribution between individuals exhibiting signs of balding and those who did not within the dataset. Further scrutiny was devoted to examining the correlations between age and various medical conditions, which emerged as significant factors in our analysis. Notably, conditions such as Seborrheic Dermatitis and Scalp Infections exhibited heightened prominence in older age groups, whereas Dermatosis and Eczema displayed greater significance at younger ages.

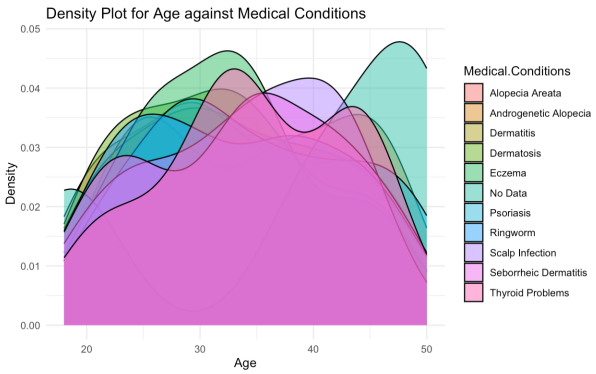

Figure 1: Stacked Density Plot

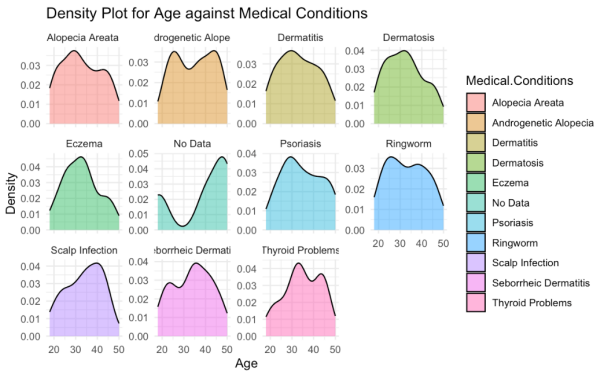

Figure 2: Individual Plots

We conducted an analysis to determine the significance of various features, gauged by their impact on reducing impurity within decision tree nodes. Features yielding substantial reductions in impurity, thereby partitioning the data into more coherent subsets, are deemed more influential. The graph depicted below illustrates this significance, where taller bars signify greater importance in predicting the target variable. These pivotal features contribute significantly to the overall predictive capacity of the random forest model. Conversely, features represented by shorter bars possess comparatively lesser importance in predicting the target variable. This graph serves as a valuable tool for discerning which variables wield the greatest influence in the decision-making process of the random forest model, providing insights into the critical determinants of the target outcome.

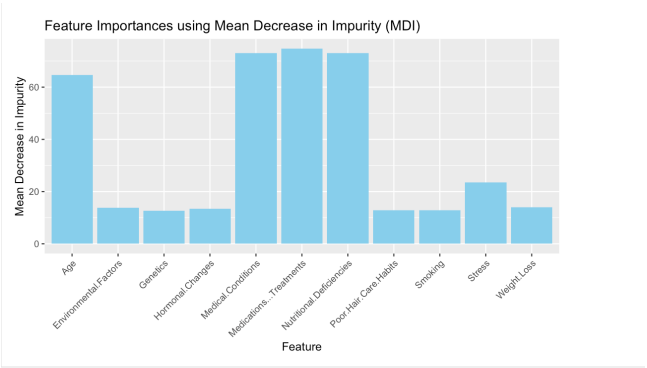

Figure 3: Feature Importance using Mean Decrease in Impurity

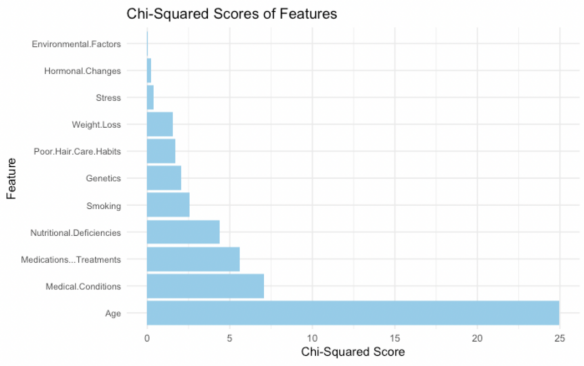

Figure 4: Feature Importance using Chi Squared Test

Proposed Methods and Assumptions

Our objective is to predict baldness probability using various statistical classifiers, focusing on their efficacy and the justification for their selection. We employ cross-validation, specifically k-fold cross-validation, to assess model performance and ensure robustness against overfitting.

Data Preparation

We quantified categorical variables to facilitate their integration into the models. The dataset, consisting of multiple predictors like genetic history, hormonal changes, and lifestyle choices, was divided into an 80% training set and a 20% testing set, with the latter also used for validation.

Logistic Regression

Why? Logistic regression is ideal for discriminative binary outcomes like baldness presence. It calculates the probability of occurrence by fitting data to a logit function.

Linear Discriminant Analysis (LDA)

Why? LDA is chosen for its efficiency in dimensionality reduction and its assumption of equal covariance matrices across groups. This method is suitable for datasets where the predictors are normally distributed, enhancing the interpretability of the model's classification.

Quadratic Discriminant Analysis (QDA)

Why? QDA allows for different covariance structures among classes, offering a more flexible approach than LDA when the equal covariance assumption does not hold. This method provides a better fit for datasets with complex relationships.

Random Forest

Why? An ensemble method that aggregates the outcomes of multiple decision trees, Random Forest is robust against overfitting and is effective for handling nonlinear data without making any specific distributional assumptions.

Model Selection and Validation

Cross-validation (k-fold): A resampling method used to evaluate the model on different subsets of the data, providing an estimate of the test error and reducing the variability in the assessment of model performance. It is especially crucial in our study to mitigate the risk of overfitting and to ensure that the model generalizes well to new data. We use k-fold cross-validation, as it strikes a balance between bias and variance, providing a reliable estimate of model performance.

Justification: These methods are selected to ensure a thorough examination of the predictive power of our variables and to construct a model that is not only accurate but also generalizable to unseen data. The choice of resampling method like k-fold cross-validation ensures that our model's performance is not merely a result of the specific way the data is split but holds across different segments of the data.

The selected models and resampling technique aim to provide a comprehensive analysis of factors influencing hair loss, balancing between model complexity and predictive accuracy to derive meaningful insights that can guide future research and interventions in the domain of hair loss prevention and treatment.

Data Analysis and Main Results

Linear Discriminant Analysis (LDA)

We utilized LDA to create an additional model and similarly tested its performance on our test data.

	Predicted	
Actual	**Positive**	**Negative**
Positive	44	57
Negative	29	31

Table 1: LDA Confusion Matrix

Prediction Results The results of our tests showed that the model had an overall accuracy of 46.58%. Furthermore, we calculated a classification error of 53.41%, precision of 60.27%, and sensitivity of 43.56%.

Reduced Model: By analyzing the importance of each predictor variable, we developed a reduced model with only features that were identified to be significant. As shown in Figure 4, Age and Medical Conditions were tested to be the most important features based on the Chi-Squared Test. Using a reduced model with only these two features, we achieved an accuracy of 52.79% and a classification error of 47.21% which is an improvement from the previous model.

Quadratic Discriminant Analysis (QDA)

We employed QDA to predict the test data outcomes and compared these predictions to the actual values. In our classification scheme, responses of 0 and 1 represent the absence and presence of hair loss, respectively. QDA distinguishes itself from LDA by allowing each class to have its own covariance matrix. Below, we outline the theoretical framework of QDA and its application to our dataset:

1. **Assumptions:** The assumption under QDA is that data from each class follows a Gaussian distribution with class-specific mean vectors and covariance matrices, permitting variable variances and inter-variable correlations within each class.
2. **Estimating Parameters:** QDA estimates the mean vector and the covariance matrix for each class from the training data, which characterize the distribution's shape and orientation in the feature space.
3. **Classification:** For a new observation, QDA computes discriminant scores using the class-specific functions and assigns the class with the highest score.
4. **Model Training:** In the training phase, QDA estimates mean vectors and covariance matrices for the classes within the 'Hair.Loss' response variable.

Advantages and Limitations: QDA can capture more complex relationships due to class-specific covariance matrices, offering a potential advantage over LDA. However, it also requires sufficient data to reliably estimate these matrices and can overfit when the sample size is too small relative to the number of predictors. This limitation is reflected in our model's large MSE, suggesting that QDA may not have performed optimally given the sample size of our dataset.

Prediction Results Using the trained model, we predicted values on the testing data set and found the following results.

	Predicted	
Actual	**Positive**	**Negative**
Positive	42	55
Negative	43	59

Table 2: QDA Confusion Matrix

The model's overall accuracy was calculated to be 50.75%. We further assessed the model's sensitivity (43.29%), which indicates its ability to correctly identify positive instances, and its specificity (57.84%), which reflects the correct identification of negative instances. The precision was noted to be 49.41%, and the log loss was 0.4925.

Random Forest Classification

We opted not to conduct subset selection for our Random Forest model, as this is typically addressed during the training process. The algorithm constructs multiple decision trees on different subsets of the data, with each tree considering a random subset of predictors at each split. This randomness helps in decorrelating the trees and capturing a diverse set of features. Our decision tree provided an implicit ranking of variable importance, where features appearing higher in the tree, indicating earlier splits, are considered more important in predicting Hair Loss.

We identified different combinations of Age, Medical Treatments, and Nutritional Deficiencies as the most important features from the trees. This was cross-validated using a Gini Index and a mean impurity graph. The Gini index measures the impurity or disorder in a set of data points and reflects how much each feature contributes to the overall reduction in impurity when making splits in the decision trees. Similarly, the mean impurity determines feature importance based on how much they reduce impurity in the decision tree nodes.

According to our analysis, Nutritional Deficiencies, Medical Conditions, and Medical Treatments emerged as the top three most important variables. This aligns with our findings from the Gini index and mean impurity metrics, which also highlighted Age, Nutritional Deficiencies, Medical Conditions, and Medical Treatments as crucial predictors. After splitting up the model into test and training we had a 0.46 classification error

However, one drawback of our model is the potential for overfitting. With around 500 trees and a relatively

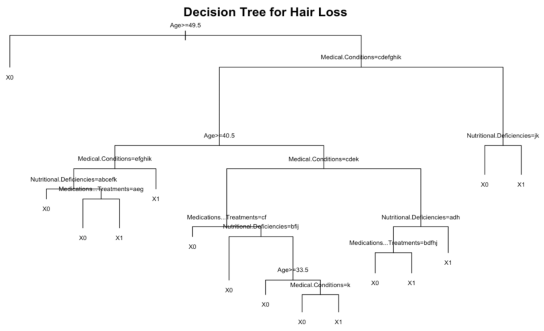

Figure 5: Random Forest Model

small dataset of only 1000 values, there is a risk that each tree may capture different aspects of noise in the data, resulting in an overly complex model. Additionally, the presence of irrelevant features may lead the model to fit to noise rather than underlying patterns, further exacerbating the risk of overfitting. To address this, we can explore regularization techniques by tuning parameters such as the minimum number of samples required to split a node and the minimum number of samples required at each leaf node in future iterations.

Prediction Results Using the trained model, we predicted values on the testing data set and found the following results.

	Predicted	
Actual	**Positive**	**Negative**
Positive	12	87
Negative	97	3

Table 3: Random Forest Confusion Matrix

Logistic Regression

Logistic regression was employed as a key analytical tool in our study to predict hair loss, leveraging its efficacy in binary outcome modeling. This subsection delineates the step-by-step methodology adopted to optimize the model selection and validate its predictive accuracy.

Initial Model and AIC-Based Selection Initiating with a baseline logistic regression model, 'Age' was considered as a preliminary significant predictor. To refine our model, we utilized the Akaike Information Criterion (AIC) through the `stepAIC` function, enabling forward stepwise selection. This process systematically identified the most contributive predictors to the model.

```
initial_model <- glm(Hair.Loss ~ Age, data = hair_train, family = binomial)
step_model <- stepAIC(initial_model, direction = "forward", ...)
```

Cross-Validation and Log Loss Estimation For assessing the model's predictive reliability, 10-fold cross-validation was implemented, focusing on log loss estimation. This metric, pivotal for evaluating predictive accuracy, assesses how well the model estimates the actual probabilities of the outcomes.

```
log_loss_estimate <- function(data, formula) { ... }
```

Forward Stepwise Selection with Log Loss Utilizing log loss as a performance metric, a forward stepwise selection was conducted. This iterative method enhanced the model by incorporating predictors that minimized the log loss, thereby optimizing predictive accuracy.

```
while (...) {
    ...
    mformula <- as.formula(...)
```

```
    losses[i] <- log_loss_estimate(hair_train, formula = mformula)
    ...
}
```

Final Model Evaluation The process culminated in a logistic regression model integrating 'Smoking', 'Age', and 'Genetics' as key predictors. This model's efficacy was subsequently appraised on a test dataset to affirm its predictive accuracy and general applicability.

```
logistic_model <- train(Hair.Loss ~ Smoking + Age + Genetics, ...)
```

Prediction Results Using the trained model, we predicted values on the testing data set and found the following results.

		Predicted	
		Positive	**Negative**
Actual	**Positive**	60	69
	Negative	25	45

Table 4: Logistic Regression Confusion Matrix

The model's overall accuracy was calculated to be 52.76%. We further assessed the model's sensitivity, which indicates its ability to correctly identify positive instances, at 70.59%, and its specificity, reflecting the correct identification of negative instances, also at 70.59%. The precision, the proportion of positive identifications that were correct, was noted to be 46.51%.

Summary Results and Discussion

Model Type	Overall Accuracy	Loss
LDA	52.79%	47.21%
QDA	50.75%	49.25%
Random Forest	54%	46%
Logistic Regression	52.7%	69%

Table 5: Model Summary

While Random Forest exhibits promising performance in predicting hair loss probability, we must be cautious about over fitting, particularly with smaller datasets such as ours. Given its ensemble nature to construct numerous decision trees, there exists a risk of overemphasizing noise over genuine patterns within the data. To address this concern, complementing Random Forest with alternative models like Logistic Regression or Linear Discriminant Analysis can provide additional insights and enhance predictive robustness. Despite the potential for over fitting, Random Forest's advantage is providing ease of interpretability to both patients and health care professionals. Therefore, while Random Forest remains a valuable tool, its integration with other models ensures a comprehensive approach to hair loss management, balancing predictive accuracy and interoperability.

Conclusion

In conclusion, while each model showcased varying degrees of success in predicting hair loss probability, Random Forest emerged as the most promising candidate. Its ability to effectively handle nonlinear relationships while maintaining high accuracy makes it the preferred choice for further exploration and refinement. Key variables such as age, medical conditions, medication treatments, and nutritional deficiencies were identified as significant predictors, warranting continued investigation in future studies aimed at personalized intervention strategies for mitigating hair loss.

References

[1] Kaggle, *Hair Health Dataset*, Available at: `https://www.kaggle.com/datasets/amitvkulkarni/hair-health?resource=download`.