2025 FALL STATS 101C LECTURE 1

Predicting Skin Cancer

Team Members:

Alexis Xie, Jiawen Yu, Joe Zhou, Kimya Jin, Yunjia Huang

Content Overview

01 Introduction

> **Skin Cancer:**
>
> The abnormal growth of skin cells that occurs when DNA damage triggers mutation. These mutations cause skin cells to multiply uncontrollably and form malignant tumors in the skin's layers.

Objective: To analyze datasets and construct a classification method to predict skin cancer status (Benign Vs. Malignant) based on various clinical, demographic, and lifestyle parameters.

5 million+ cases diagnosed in U.S. each year

Prevalence

1 in 5 Americans will develop it by 70

Lifetime Risk

$8 billion+ Annual treatment costs

Economic Cost

02 Data Cleaning

Dataset Overview

50,000
Observations

20 + 29
Variables

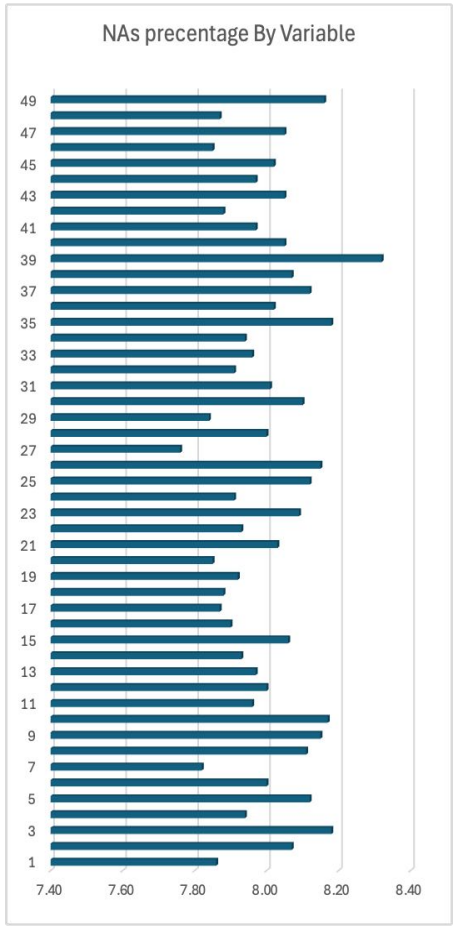
Deal with NA values

Individual parameters generally have an NA percentage between **7.76 and 8.83.**

> **Could not simply delete entries:**
> Entries with full information is 791/50000.
> *Too aggressive and less informative.*

Approach

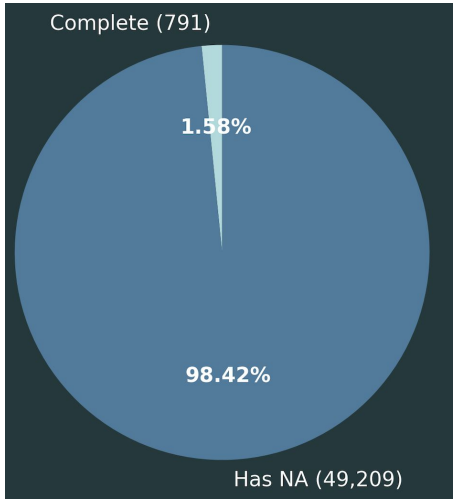
- **Categorical:** replace NAs with mode
- **Numerical:** replace NAs with mean



Multicollinearity

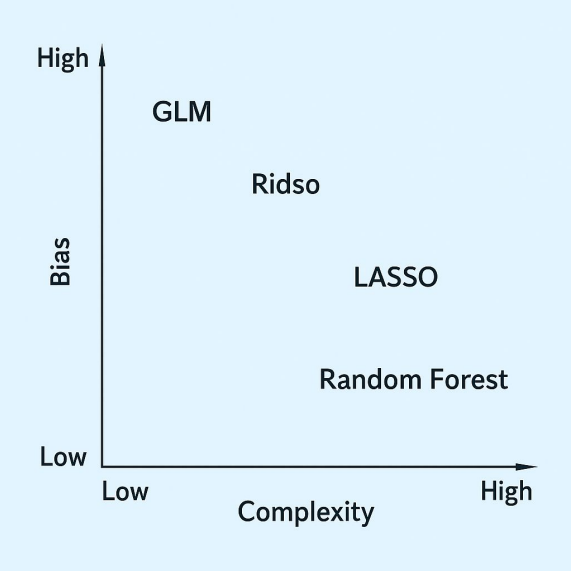
Outdoor_job vs Occupation

Graph from: U.S. Department of Labor

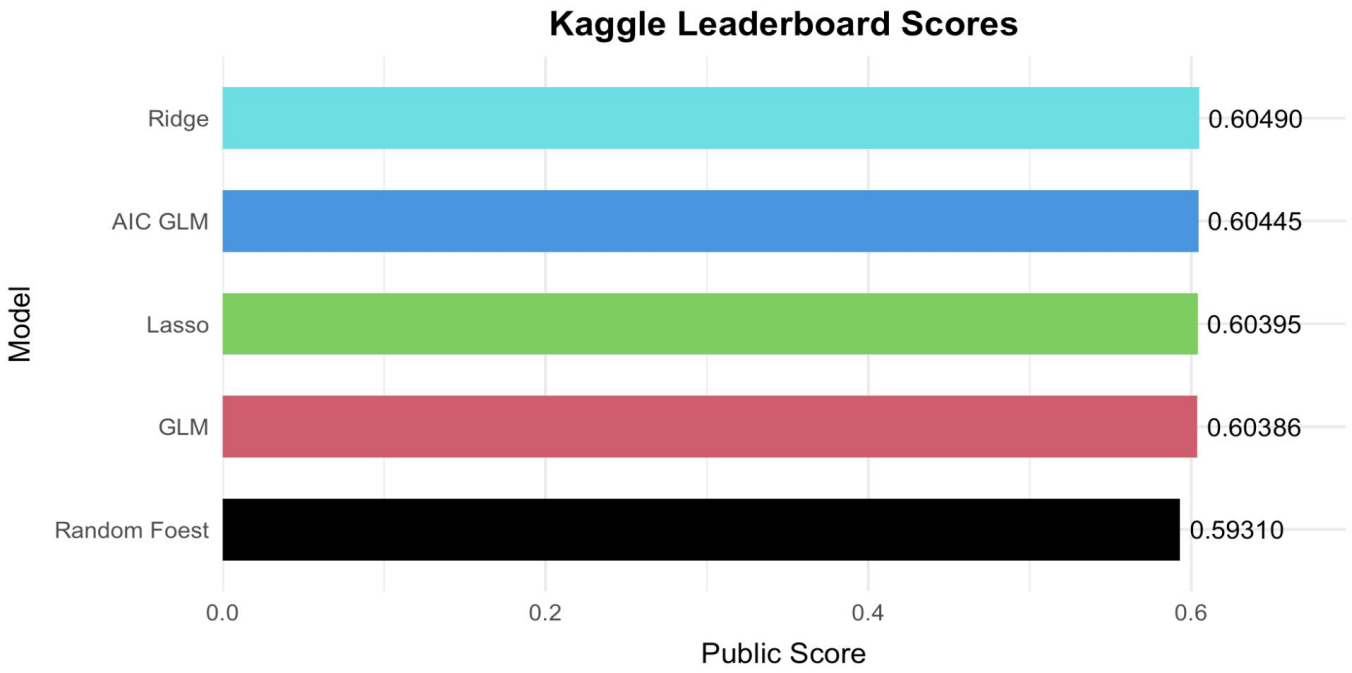


03 Data Modeling

Model	Characteristics / Pros
Logistic	Fast, stable, works good with linear relation
Ridge	L2 penalty (shrinks), works good with Multicollinearity
LASSO	L1 penalty, feature selection
Random Forest	Strong performance, flexible



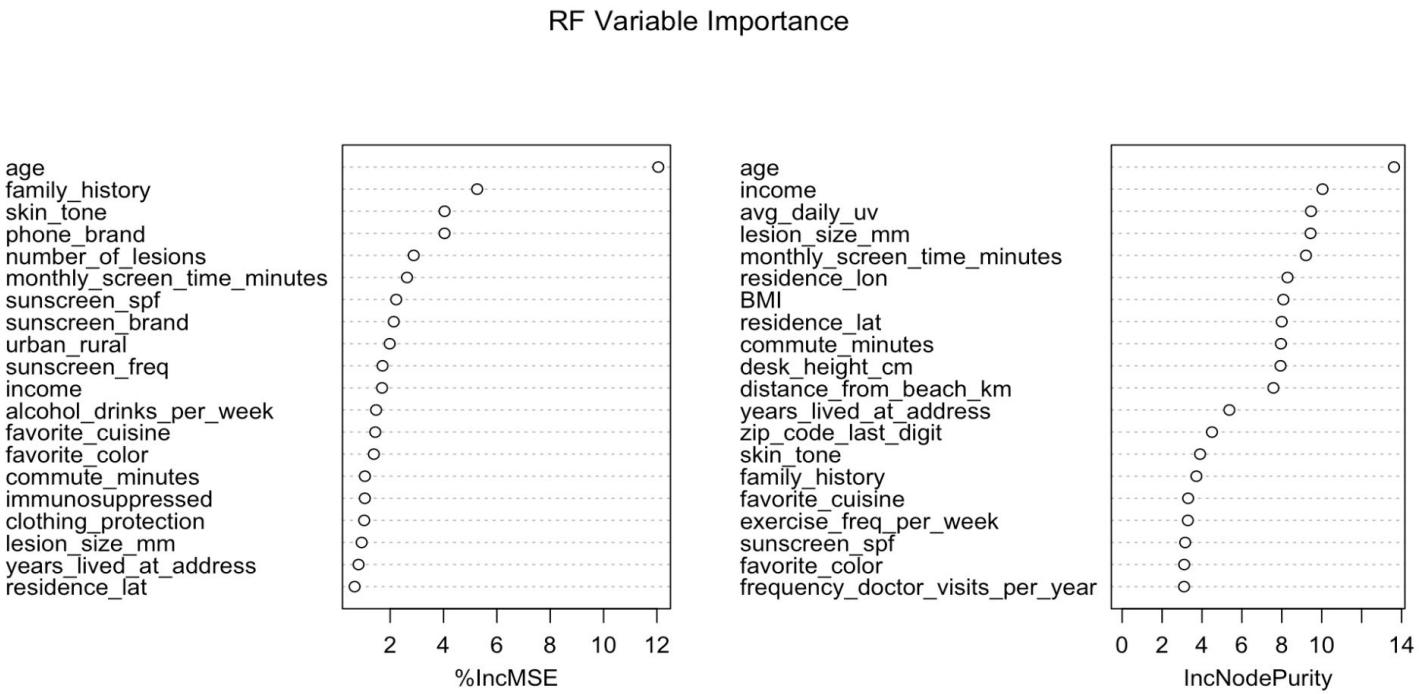
A: Full Model With Tuned Parameter(K-fold)



First attempt:

Ridge, Glm and Lasso out perform RF, suggest there is **linear relationship** in the our data. Also these scores suggest that **overfitting** are happened in our model, **Feature selection is needed.**

B: Checking Importance Plot, Aic Result




```
sunscreen_freqNever          2.938e-01  4.083e-02   7.197 6.14e-13 ***
sunscreen_freqOften          2.622e-02  2.948e-02   0.889 0.373772
sunscreen_freqRarely         3.402e-01  3.401e-02  10.004  < 2e-16 ***
sunscreen_freqSometimes      2.264e-01  3.158e-02   7.170 7.49e-13 ***
```

(Screen shot from AIC)

Summary:

The RF importance plot shows that most features have similar, low importance, meaning the model failed to capture strong nonlinear patterns.

AIC results show large differences across the levels in categorical variable — some are highly significant while others have no effect, indicating uneven influence within categorical predictors.

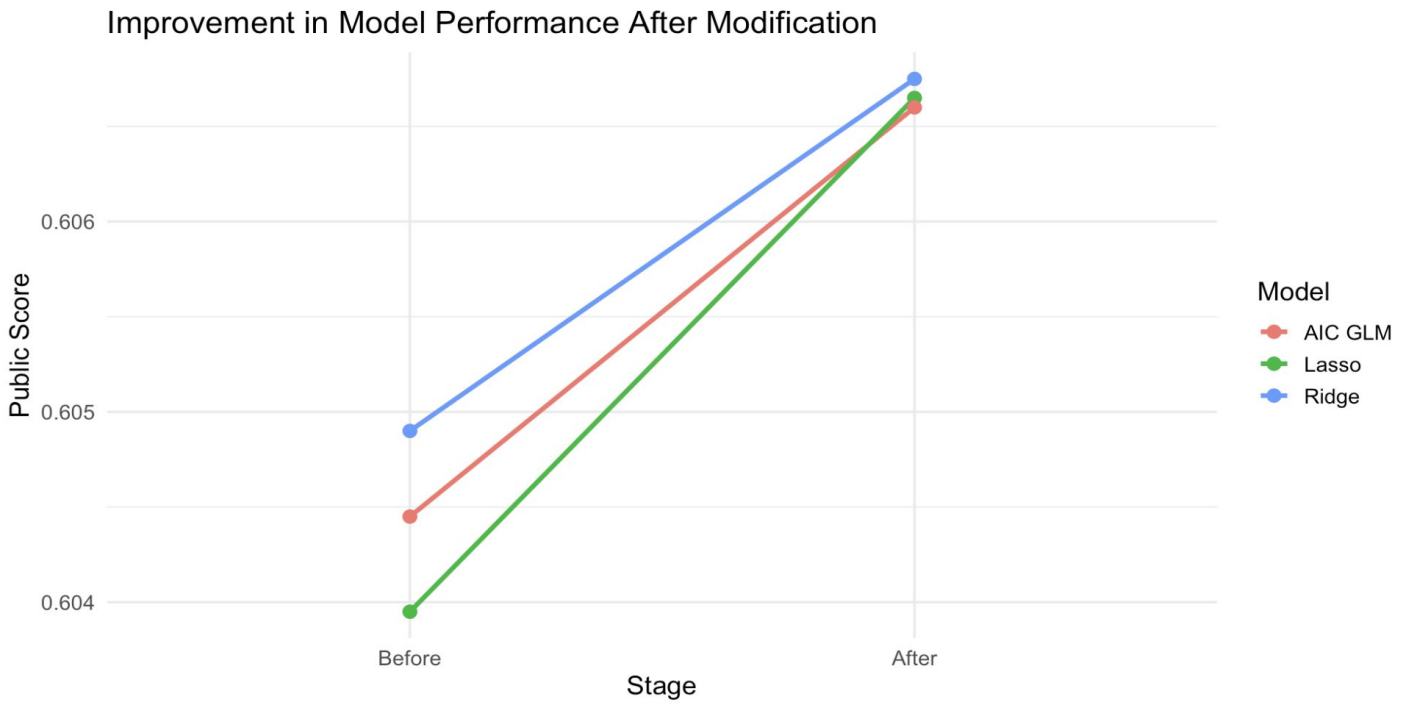
C: Feature Selection and Dummy Variable

sunscreen_freqRarely	sunscreen_freqAlways	sunscreen_freqNever	sunscreen_freqSometimes
1	0	0	0
0	1	0	0
1	0	0	0

Summary:

We selected 20 predictors based on variable importance and GLM results. Categorical predictors were converted into dummy variables according to their levels.

D: Model After data Processing



Summary:

Ridge still achieves the best performance. The L2 penalty appears to be a better fit for our selected variables

04 Result & Discussion

FINAL MODEL

Ridge

(lambda = 0.0524)

SCORE

0.60675

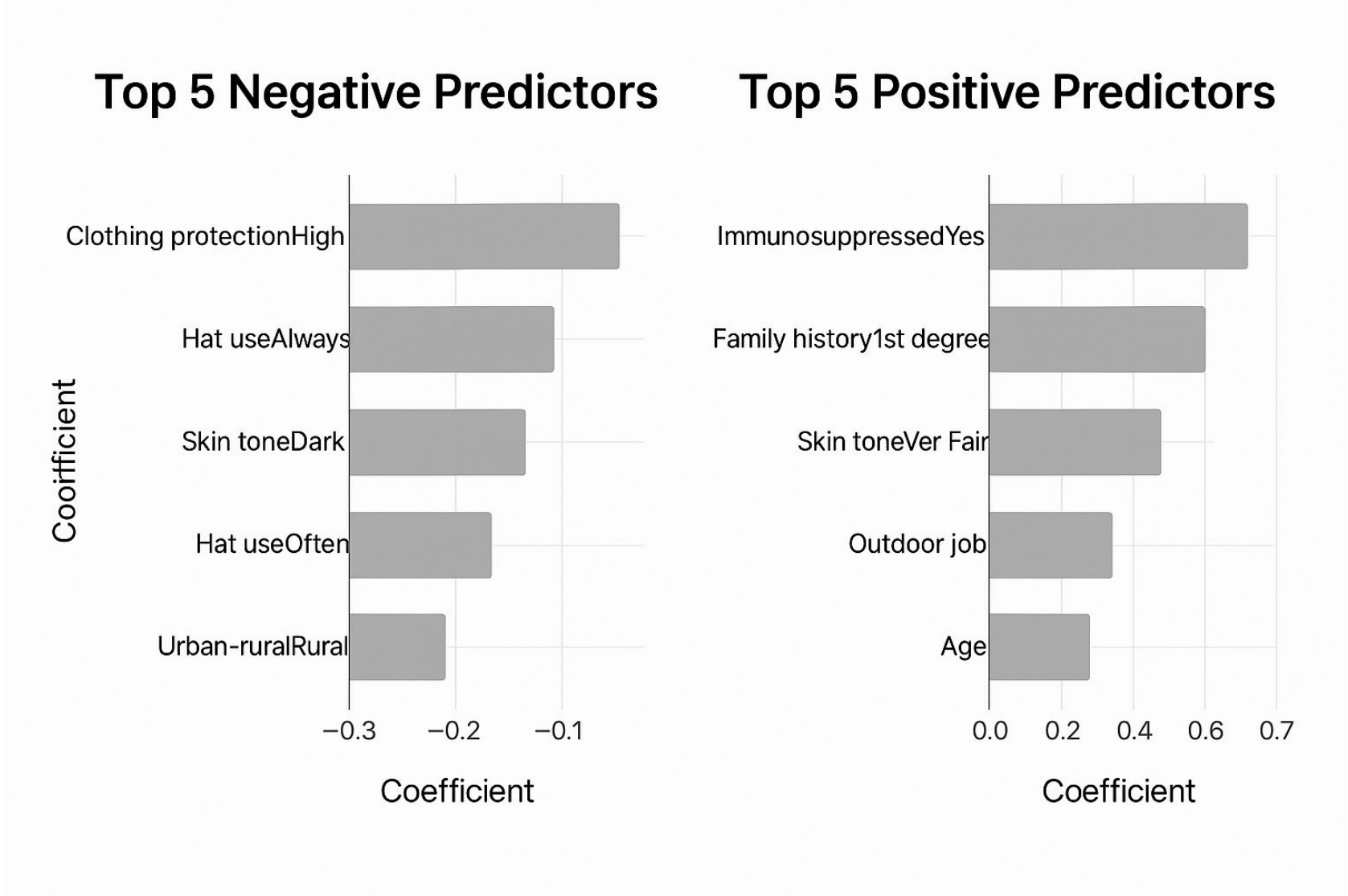
RANK

TOP 5!!

PREDICTORS

20

04 Result & Discussion



05 Limitation

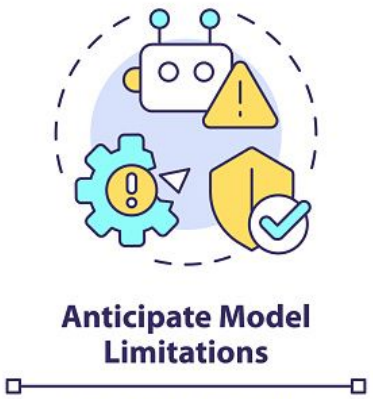
Data Cleaning Limitations

- **Using Mean Imputation (Numeric):**
  Underestimate variance, Reduce correlation, Loss of distribution shape.

- **Using Mode Imputation (Categorical):**
  Distortion of class proportions, Bias downstream models.

Model Limitations

- **Logistic:** Assumes linearity in the log-odds.
- **Ridge:** High dimensional issues.
- **LASSO:** Unstable with highly correlated features.
- **Random Forest:** Hard to construct, require large computational resources for hyperparameter CV.



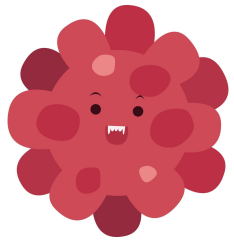
Conclusion

> Although advanced models like Random Forest are widely recognized for their predictive power, our analysis indicates that the **Ridge regression model** performed more effectively on this dataset.

However, the underlying mechanisms behind skin cancer development remain insufficiently explored within our feature scope. The current dataset may omit relevant biological, behavioral, or environmental predictors, limiting the model's ability to capture complex causal relationships.

Furthermore, when reducing predictor dimensionality, we relied on a fixed default of twenty variables rather than conducting a systematic evaluation across varying feature counts, which may have prevented the identification of the true optimal model configuration.

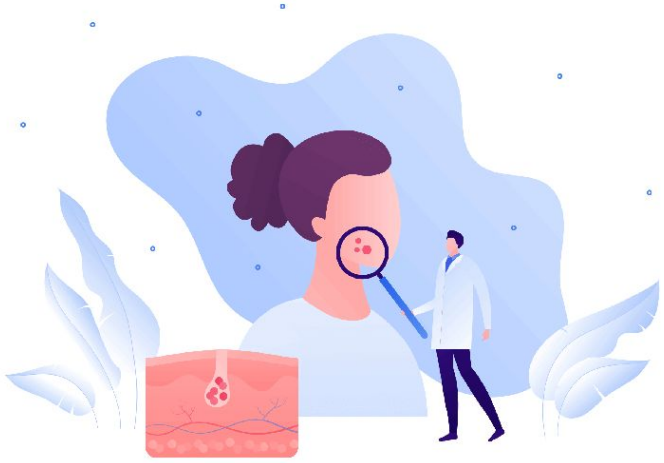
> **Future Research:** Should expand variable selection, incorporate additional medically relevant features, and rigorously test model performance across different predictor sets to ensure a more comprehensive and reliable understanding of skin cancer prediction.



06 References

Dataset:

“Predicting Skin Cancer Status.” Kaggle
www.kaggle.com/competitions/predicting-skin-cancer-status/data



Thank You

Thanking you for listening and participating.