Predicting Skin Cancer Using Statistical Learning Models

A Comparative Analysis of Logistic, LASSO, and Elastic Net Regression on Structured Clinical Data

Franklin Truong[1], Jason Li[1], Roy Cheng[1], Felipe Duenas[1] and Keivan Bolouri[1]

1. University of California, Los Angeles, Department of Statistics and Data Science, 8125 Math Sciences Bldg, Los Angeles, CA 90095

Stats 101C — Lecture 2, Fall 2025 https://keivanbolouri.github.io/finalProject140XP/

Table of contents

List of Figures

List of Tables

1 Abstract

This project develops and evaluates statistical learning models to classify skin lesions as benign or malignant using structured (non-image) predictors. Using a skin cancer dataset with 50,000 training observations and 20,000 test observations, we performed exploratory analysis, handled missing values via median (numeric) and mode (categorical) imputation, standardized numerical features, and reduced dimensionality through association-based feature screening. We then trained and compared a baseline logistic regression model with regularized variants (LASSO and elastic net logistic regression), tuning hyperparameters via cross-validation to balance generalization and performance under slight class imbalance.

The final model was an elastic net logistic regression with mixing parameter $\alpha = 0.7$ and 34 predictors, selected based on leaderboard performance. On the held-out test set, the model achieved 60.508% accuracy, outperforming naive baselines but remaining modest for a diagnostic setting. Overall, results suggest that structured demographic, behavioral, environmental, and clinical variables provide limited discriminative signal without lesion imaging features, highlighting both the potential utility of tabular predictors and the limitations of relying on them alone for skin cancer detection.

2 Introduction

Skin cancer is one of the most common and preventable forms of cancer in the United States[1], yet early detection remains crucial for improving survival outcomes. In the U.S. alone, nearly six million individuals are treated for skin cancer annually[2], with melanoma accounting for approximately 97,000 new cases and over 8,000 deaths each year[2]. Early detection greatly improves patient prognosis – for example, the five-year survival rate for melanoma is over 99% when detected at an early stage, compared to much lower survival if the cancer has advanced[3]. While skin cancer can often be identified visually or through clinical examination, early diagnosis remains challenging due to the wide range of risk factors and lesion characteristics that must be considered.

In this project, we analyze a skin cancer dataset containing 50,000 training observations and 20,000 test observations, with a binary response variable (Cancer) indicating whether a lesion is benign or malignant. The dataset includes 50 predictor variables representing a broad range of factors, including demographic information, environmental measures, behavioral and lifestyle factors, and clinical attributes. These predictors provide a structured, non-image-based view of potential skin cancer risk factors.

The goal of this project is to determine whether structured (tabular) data can be used to distinguish between benign and malignant skin lesions. By applying statistical learning methods, we aim to identify which predictors are most informative and to evaluate how well different classification models perform on this task. We seek to select an approach that balances interpretability, generalization, and predictive accuracy, while also discussing the limitations of using only structured predictors for skin cancer classification

3 Literature Review

Recent studies have demonstrated the potential of machine learning models applied to structured clinical and demographic data (non-image features) for skin cancer risk prediction. For example, an XGBoost-based model using electronic health records and genetic/lifestyle factors from a 400,000-patient cohort achieved high accuracy in identifying skin cancer cases ($F_1 \approx 0.90$ in European-ancestry patients) and leveraged SHAP values to interpret nonlinear risk factor effects [4]. Another approach employed logistic regression to develop a nomogram with eight behavioral and dietary risk factors, showing good discrimination (AUC $\approx$ 0.8) and clinical utility for head and neck skin cancer prevention[5]. In melanoma-specific research, a new 16-factor risk model (MP16) was trained on a 41,000-person cohort and improved predictive accuracy (C-index ≈ 0.74) compared to earlier tools, capturing $\sim$ 74% of of future melanomas by targeting the top 40% high-risk group[6]). Finally, a large 2021 study combined survey-derived risk variables with polygenic risk scores to create composite risk metrics for melanoma and non-melanoma skin cancers, identifying top-percentile individuals with over tenfold higher risk to inform targeted screening [7]).

4 Data Processing

After completing the exploratory analysis, we examined the dataset for missing values. We found that the overall data quality was relatively high, with most predictors containing approximately 7–8% missing values. No predictor exceeded 10% missingness, so eliminating variables or observations would have resulted in unnecessary data loss. To address missing values in a consistent manner, we applied the following imputation strategy: Numerical variables were imputed using the median, which is robust to outliers. Categorical variables were imputed using the most frequent category (mode). This approach allowed us to preserve all 50,000 observations in the training dataset while ensuring that the data were suitable for modeling. After imputation, no missing values remained in the predictors used for analysis.

```
# A tibble: 50 x 3
   Variable              Missing_Count Missing_Percent
   <chr>                         <int>           <dbl>
 1 vitamin_d_supplement           4158            8.32
 2 phone_brand                    4092            8.18
 3 skin_tone                      4091            8.18
 4 sunscreen_spf                  4085            8.17
 5 preferred_shoe_type            4082            8.16
 6 sunscreen_freq                 4075            8.15
 7 residence_lat                  4074            8.15
 8 commute_minutes                4062            8.12
 9 near_high_power_cables         4061            8.12
10 income                         4058            8.12
# i 40 more rows
```

```
# A tibble: 49 x 3
   Variable              Missing_Count Missing_Percent
   <chr>                         <int>           <dbl>
 1 residence_lon                  1694            8.47
 2 phone_brand                    1677            8.38
 3 preferred_shoe_type            1662            8.31
 4 uses_smartwatch                1656            8.28
 5 skin_tone                      1655            8.28
 6 smoking_status                 1648            8.24
 7 income                          1647            8.23
 8 near_high_power_cables          1647            8.23
 9 lesion_color                    1634            8.17
10 skin_photosensitivity           1624            8.12
# i 39 more rows
```

```r
missing_report <- missing_train %>%
  rename(
    Missing_Count_Train   = Missing_Count,
    Missing_Percent_Train = Missing_Percent
  ) %>%
  left_join(
    missing_test %>%
      rename(
        Missing_Count_Test   = Missing_Count,
        Missing_Percent_Test = Missing_Percent
      ),
    by = "Variable"
  )
missing_report %>%
  summarise(
    Total_Variables = n(),
    Vars_Over_20pct_Train = sum(Missing_Percent_Train > 20, na.rm = TRUE),
    Vars_Over_30pct_Train = sum(Missing_Percent_Train > 30, na.rm = TRUE),
    Vars_Over_50pct_Train = sum(Missing_Percent_Train > 50, na.rm = TRUE),
    Vars_Over_20pct_Test  = sum(Missing_Percent_Test  > 20, na.rm = TRUE),
    Vars_Over_30pct_Test  = sum(Missing_Percent_Test  > 30, na.rm = TRUE),
    Vars_Over_50pct_Test  = sum(Missing_Percent_Test  > 50, na.rm = TRUE)
  )
```

```
# A tibble: 1 x 7
  Total_Variables Vars_Over_20pct_Train Vars_Over_30pct_Train
            <int>                 <int>                 <int>
1              50                     0                     0
# i 4 more variables: Vars_Over_50pct_Train <int>, Vars_Over_20pct_Test <int>,
#   Vars_Over_30pct_Test <int>, Vars_Over_50pct_Test <int>
```

4.0.1 Data Import and Cleaning

4.0.2 Data Import and Cleaning

4.0.3 Finalizing the Dataset for Analysis

5 Hypothesis Tests

5.1

5.1.1 Hypothesis 1: Civilian Harm Difference

6 Model Selection

6.0.1 Poisson dispersion test

7 Visualizations

7.1 Visualization (Test 1)

8 Conclusion

9 Author Contributions

10 Acknowledgments

11 References

[1] U.S. Department of Health and Human Services. Skin cancer: Quick facts from the surgeon general 2014. https://www.hhs.gov/surgeongeneral/reports-and-publications/skin-cancer/fact-sheet/index.html (accessed December 15, 2025).

[2] National Center for Chronic Disease Prevention and Health Promotion. Health and economic benefits of skin cancer interventions 2023. https://www.cdc.gov/nccdphp/priorities/skin-cancer.html (accessed December 15, 2025).

[3] The Skin Cancer Foundation. Skin cancer facts & statistics 2024. https://www.skincancer.org/skin-cancer-information/skin-cancer-facts/ (accessed December 15, 2025).

[4] Kaiser I, Pfahlberg AB, Uter W, Heppt MV, Veierød MB, Gefeller O. Risk Prediction Models for Melanoma: A Systematic Review on the Heterogeneity in Model Development and Validation. International Journal of Environmental Research and Public Health 2020;17:7919. https://doi.org/10.3390/ijerph17217919.

[5] Zou R, Lin Y, Da C, Liao G. Novel nomogram and decision curve analysis for predicting head and neck skin cancer risk. Scientific Reports 2025;15. https://doi.org/10.1038/s41598-025-25427-0.

[6] Whiteman DC, Olsen CM, Wang H, Law MH, Neale RE, Pandeya N. A Risk Prediction Tool for Invasive Melanoma. JAMA Dermatology 2025;161:1123. https://doi.org/10.1001/jamadermatol.2025.3028.

[7] Fontanillas P, Alipanahi B, Furlotte NA, Johnson M, Wilson CH, Agee M, et al. Disease risk scores for skin cancers. Nature Communications 2021;12. https://doi.org/10.1038/s41467-020-20246-5.