Predicting Skin Cancer Using Statistical Learning Models

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

Franklin Truong Jason Li Roy Cheng Keivan Bolouri
Felipe Duenas

2025-12-15

Table of contents

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.

Introduction

Skin cancer is one of the most common and preventable forms of cancer in the United States@hhs_skin_cancer, yet early detection remains crucial for improving survival outcomes. In the U.S. alone, nearly six million individuals are treated for skin cancer annually@cdc_skin_cancer, with melanoma accounting for approximately 97,000 new cases and over 8,000 deaths each year@cdc_skin_cancer. 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@skincancer_facts. 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

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 @kaiser2020. 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@Zou2025. 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 $\approx$ 0.74) compared to earlier tools, capturing $\sim$ 74% of of future melanomas by targeting the top 40% high-risk group@Whiteman2025). 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 @Fontanillas2021).

Data Analysis

Exploratory Data Analysis

We began by examining the overall structure and distributions in the training dataset. The response variable (Cancer) is binary (Benign or Malignant), and we observed a slight class

imbalance: benign lesions were more frequent than malignant lesions in the training data. This imbalance means that a naive classifier that predicts all cases as benign would achieve a non-trivial accuracy, so careful model evaluation is necessary. We also inspected the distribution of key predictors. For example, the average age of patients with malignant lesions appeared higher than that of patients with benign lesions, consistent with the fact that skin cancer risk increases with age. We explored categorical risk factors as well: certain exposure-related factors (such as indicators of high UV exposure or tanning habits) were somewhat more prevalent in malignant cases, though there was significant overlap between the benign and malignant groups for most individual predictors. Overall, no single predictor showed a dramatic separation between malignant and benign lesions in isolation, suggesting that multiple factors in combination would be needed for effective prediction. In addition, we checked for missing data and outliers during the exploratory phase. We found that the dataset's overall quality was high, with only a moderate amount of missing values (on the order of 7–8% missing per predictor). There was no evidence of extreme outliers that would require removal or transformation beyond standardization. The presence of some missing values and the lack of obvious one-variable predictors of cancer underlined the need for a robust modeling approach with proper data preprocessing, which we implemented as described below.

Data Cleaning and Missing Values

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.

Using these association measures, all predictors were ranked from strongest to weakest. Variables with very weak association to the response variable were discarded, while stronger predictors were retained. This process reduced the dataset from 51 variables to 35 total variables, including the response variable. The final set consisted of 20 numerical predictors and 14 categorical predictors, while maintaining all original observations. This variable selection process helped simplify the modeling stage, reduced multicollinearity, and improved model interpretability, while retaining the most informative predictors for classification.

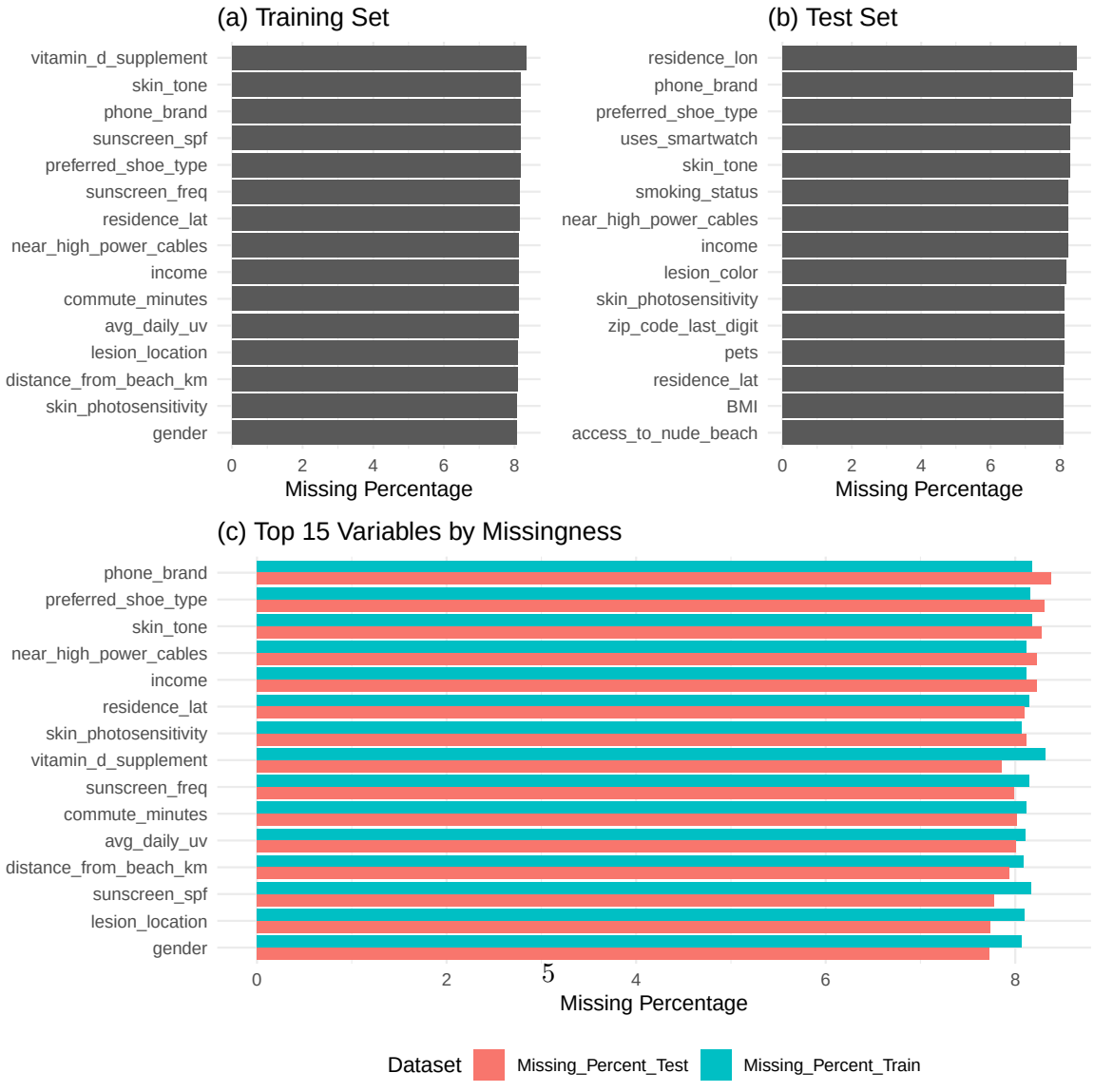
Table 1: Top ten variables with the highest proportion of missing values.

Variable	Missing (%)
vitamin_d_supplement	8.3
phone_brand	8.2
skin_tone	8.2
sunscreen_spf	8.2
preferred_shoe_type	8.2
sunscreen_freq	8.2
residence_lat	8.2
commute_minutes	8.1
near_high_power_cables	8.1
income	8.1

(a) Training set variables.

Variable	Missing (%)
residence_lon	8.5
phone_brand	8.4
preferred_shoe_type	8.3
uses_smartwatch	8.3
skin_tone	8.3
smoking_status	8.2
income	8.2
near_high_power_cables	8.2
lesion_color	8.2
skin_photosensitivity	8.1

(b) Test set variables.



Variable Selection

Given the large number of predictors, variable selection was necessary to reduce dimensionality and eliminate weak predictors that could introduce noise and increase the risk of overfitting. For categorical predictors, we used Cramér's V to measure the strength of association between each categorical variable and the cancer outcome. Cramér's V is based on the chi-squared statistic and produces values between 0 and 1, with larger values indicating stronger association. Here is a sample of the first 10 variables: For numerical predictors, we computed the point-biserial correlation, which measures the strength of association between a continuous predictor and a binary response variable. Numerical variables were ranked based on the absolute value of this correlation.

Table 4: Top 10 Categorical Predictors Ranked by Cramér's V (Training Data)

Variable	Cramér's V
family_history	0.104
skin_tone	0.087
sunscreen_freq	0.066
immunosuppressed	0.060
occupation	0.045
tanning_bed_use	0.041
clothing_protection	0.036
skin_photosensitivity	0.033
hat_use	0.028
lesion_location	0.011

Table 5: Top 10 Numerical Predictors Ranked by Point-Biserial Correlation (Training Data)

Variable	Point-Biserial Correlation
age	0.1355
avg_daily_uv	0.0599
number_of_lesions	0.0513
sunburns_last_year	0.0457
outdoor_job	0.0451
lesion_size_mm	0.0223
sunscreen_spf	-0.0174
years_lived_at_address	0.0103
income	0.0088
desk_height_cm	0.0083

```
[1] 35
```

Table 6: Final Selected Predictors After Association-Based Variable Screening

No	Type	Variable	Selection Criterion
1	Categorical	family_history	Cramér's V > 0.05
2	Categorical	skin_tone	Cramér's V > 0.05
3	Categorical	sunscreen_freq	Cramér's V > 0.05
4	Categorical	immunosuppressed	Cramér's V > 0.05
5	Categorical	occupation	Cramér's V > 0.05
6	Categorical	tanning_bed_use	Cramér's V > 0.05
7	Categorical	clothing_protection	Cramér's V > 0.05
8	Categorical	skin_photosensitivity	Cramér's V > 0.05
9	Categorical	hat_use	Cramér's V > 0.05
10	Categorical	lesion_location	Cramér's V > 0.05
11	Categorical	favorite_cuisine	Cramér's V > 0.05
12	Categorical	music_genre	Cramér's V > 0.05
13	Categorical	lesion_color	Cramér's V > 0.05
14	Categorical	sunscreen_brand	Cramér's V > 0.05
1	Numerical	age	Abs. Point-Biserial Corr. > 0.02
2	Numerical	avg_daily_uv	Abs. Point-Biserial Corr. > 0.02
3	Numerical	number_of_lesions	Abs. Point-Biserial Corr. > 0.02
4	Numerical	sunburns_last_year	Abs. Point-Biserial Corr. > 0.02
5	Numerical	outdoor_job	Abs. Point-Biserial Corr. > 0.02
6	Numerical	lesion_size_mm	Abs. Point-Biserial Corr. > 0.02
7	Numerical	sunscreen_spf	Abs. Point-Biserial Corr. > 0.02
8	Numerical	years_lived_at_address	Abs. Point-Biserial Corr. > 0.02
9	Numerical	income	Abs. Point-Biserial Corr. > 0.02
10	Numerical	desk_height_cm	Abs. Point-Biserial Corr. > 0.02
11	Numerical	residence_lon	Abs. Point-Biserial Corr. > 0.02
12	Numerical	exercise_freq_per_week	Abs. Point-Biserial Corr. > 0.02
13	Numerical	zip_code_last_digit	Abs. Point-Biserial Corr. > 0.02
14	Numerical	distance_from_beach_km	Abs. Point-Biserial Corr. > 0.02
15	Numerical	alcohol_drinks_per_week	Abs. Point-Biserial Corr. > 0.02
16	Numerical	BMI	Abs. Point-Biserial Corr. > 0.02
17	Numerical	commute_minutes	Abs. Point-Biserial Corr. > 0.02
18	Numerical	frequency_doctor_visits_per_year	Abs. Point-Biserial Corr. > 0.02
19	Numerical	residence_lat	Abs. Point-Biserial Corr. > 0.02
20	Numerical	monthly_screen_time_minutes	Abs. Point-Biserial Corr. > 0.02

Model Selection

After preprocessing and feature selection, we proceeded to build predictive classification models on the reduced set of 34 predictors. We first fit a standard logistic regression model using all selected features as a baseline. The logistic regression model provides an interpretable baseline by estimating the odds of a lesion being malignant as a linear function of the predictors (with a logistic link). However, with 34 predictors, a basic logistic model can still risk overfitting and may include some predictors that contribute little to accuracy. All numeric features were standardized (centered to mean 0 and scaled to unit variance) prior to modeling so that they would be on comparable scales and to help the optimization of regularized models. Next, we applied regularization techniques to the logistic model to perform automatic variable selection and to improve generalization performance. In particular, we trained a LASSO logistic regression model, which is a logistic regression with an L1 penalty on the coefficients. The LASSO penalty tends to shrink the coefficients of less important features toward zero, effectively eliminating those features from the model if they are not strongly associated with the outcome. This can greatly simplify the model by selecting a sparse set of predictors. We expected the LASSO to be useful given the number of predictors and the likelihood that some of them, even after our filtering step, might still be only weakly related to the outcome. Finally, we trained an elastic net regularized logistic regression model, which generalizes the LASSO by using a combination of L1 and L2 penalties. The elastic net has two hyperparameters to tune: the mixing parameter (which controls the relative weight of L1 vs L2 regularization) and the regularization strength parameter . We utilized cross-validation to tune these hyperparameters and to select the best model. Specifically, we performed 3-fold cross-validation on the training set with a model training pipeline that included the standardization step and the logistic model with a given regularization setting. During cross-validation, we evaluated model performance using a classification metric that takes into account both sensitivity and specificity. (In practice, we optimized for the ROC AUC metric on the validation folds, which balances true positive rate and false positive rate, rather than simply using raw accuracy, to account for the slight class imbalance.) We searched over a grid of values for the LASSO and elastic net models, and for the elastic net we also tried different values of to find the optimal blend of L1 and L2 regularization. Based on the cross-validation results, the elastic net model achieved the best performance on the validation folds, outperforming both the baseline logistic regression and the purely L1-penalized model. The optimal hyperparameters found for the elastic net were a mixing parameter of approximately $\alpha \approx 0.7$ (indicating that a 70% L1 and 30% L2 penalty mix gave the best result) and a certain regularization strength that maximized the model's discriminative ability. The fact that $\alpha \approx 0.7$ was selected (rather than $\alpha = 1$ corresponding to pure LASSO) suggested that including some L2 penalty provided a benefit, likely by stabilizing the coefficient estimates for correlated predictors. After determining the best parameters, we refit the elastic net logistic regression on the entire training set using those parameters. This final model was then used to predict probabilities of malignancy for the lesions in the hold-out test set. We used a default 0.5 probability cutoff to classify a lesion as "Malignant" or "Benign" for the final submission. (We also briefly experimented with adjusting the clas-

sification threshold to maximize accuracy on the training data, but ultimately the standard 0.5 threshold was chosen for evaluating test outcomes.) The final submitted model, therefore, was an elastic net logistic regression with $\alpha \approx 0.7$, applied to 34 selected features, predicting the probability of a lesion being malignant.

Results and Discussions

The final elastic net logistic regression model achieved a public accuracy of 0.606, which is substantially better than random guessing (approximately 0.50) and indicates meaningful predictive signal from the structured variables. However, an accuracy in the low 60% range remains modest for a medical diagnostic task, suggesting that non-image features alone provide limited discriminative power for skin cancer classification. This result highlights the inherent difficulty of distinguishing malignant from benign lesions using only demographic, environmental, and lifestyle variables, and suggests that incorporating lesion imaging or more informative clinical features would be necessary to achieve stronger predictive performance.

Conclusion and Limitations

Overall, despite our model ranking in the top 15 on the project leaderboard, this approach faced several important limitations. First, the model relied exclusively on structured, non-imaging variables and did not make use of any image-based features of the lesions. In real-world skin cancer detection, the appearance of the lesion (shape, color, border irregularity, etc.) is critical, and excluding that information severely limits diagnostic accuracy. Second, there was a class imbalance (fewer malignant cases than benign), which may have negatively impacted the model's ability to learn and detect the malignant class. Imbalanced data can lead models to favor the majority class (benign) and struggle to identify the minority class (malignant), as we observed with the lower sensitivity for malignancies. Finally, even after feature selection, we still included a couple dozen predictors that were derived from lifestyle and environmental data; many of these variables have only weak relationships with the outcome. The inclusion of numerous weak predictors can introduce noise and reduce the precision of the model's predictions, even with regularization to mitigate overfitting. This project demonstrates that supervised machine learning methods can be applied to distinguish between benign and malignant skin lesions using structured patient and lesion data. However, the moderate accuracy of the final model highlights the limitations of relying solely on non-image predictors for skin cancer classification. In future work, model performance could likely be improved by incorporating image data (such as dermoscopy or clinical photographs of the lesions) alongside the structured variables, since computer vision algorithms and dermatologists alike rely heavily on visual patterns to identify skin cancers. Additionally, exploring interaction terms or non-linear models (e.g. decision tree ensembles or neural networks) might capture more complex relationships between risk factors. Careful handling of class imbalance (through techniques such as

resampling, cost-sensitive learning, or adjusting decision thresholds) would also be important to improve malignant case detection. Nevertheless, this project provided valuable experience in the end-to-end process of predictive modeling with real-world data. We performed data cleaning and imputation to handle missing values, used statistical methods for feature selection, applied and tuned regularized logistic regression models, and evaluated the results on an independent test set. The exercise underscored the importance of domain knowledge (to understand what predictors might matter), the challenges of working with imperfect data, and the need to consider model limitations when interpreting results. Although our structured-data model alone is not ready for clinical use, it serves as a baseline and learning tool. It emphasizes that while non-image data can contribute some signal for skin cancer risk, truly effective skin cancer prediction will likely require integrating all available information — including the rich visual data that was beyond the scope of this project. The insights gained here set the stage for more comprehensive approaches in the future, combining both data-driven modeling and clinical expertise to improve early detection of skin cancer.

Author Contributions.

All authors contributed equally to the conception and design of the study, data preprocessing and analysis, model development, interpretation of results, and writing and revision of the manuscript. All authors read and approved the final version of the manuscript.

Acknowledgments

References

View Published Article