Stats 101C - Final Project: Skin Cancer Status Prediction

Joey Lee (Lec 2)

Abstract

The goal of this project is to build a classification model that reliably predicts a binary outcome: whether a particular abnormal skin lesion suspected of being cancerous is malignant or benign. We found that the "best" model, balances predictive accuracy and parsimony, uses Generalized Additive Models (GAM) with 14 predictors: s(age), skin_tone, s(avg_daily_uv), sunscreen_freq, hat_use, clothing_protection, tanning_bed_use, outdoor_job, family_history, immunosuppressed, s(lesion_size_mm), number_of_lesions, skin_photosensitivity, and sunburns_last_year. The public Kaggle score is 0.60425.

I. Introduction

Skin cancer is the most common type of cancer afflicting people around the world. The main cause of the carcinogenesis is believed to be damage to DNA due to exposure to ultraviolet (UV) radiation. This is most commonly from sunlight, but can also come from exposure to tanning beds. Types of skin cancer include Basal Cell Carcinoma (BCC), Squamous Cell Carcinoma (SCC), and the least common but most lethal form: Melanoma.

Early detection and treatment is crucial to maximize patient survival. For example, 99% of melanoma patients whose disease was detected early in its course survived for at least 5 years. In recent years, machine learning has been leveraged to use patient characteristics and characteristics of abnormal skin lesions to detect skin cancer early at a stage when the disease is easier to treat.

We work with a synthetic skin cancer dataset constructed using epidemiological data. It consists of 70,000 rows, each of which represents a single patient suspected of having skin cancer. There are 50 columns total, 49 which are a mix of categorical and numerical predictors, and one binary categorical target variable: whether the lesion is malignant or benign. The predictor categories include: demographic, environmental and exposure, sun protection and skincare, biological and health, behavior and lifestyle, and miscellaneous categories. Before data preprocessing, the raw dataset contains 20 numerical predictors and 29 categorical predictors.

II. Data Preprocessing

A. Data Cleaning

In the raw dataset, the variable "outdoor_job" was categorized as numeric type and each value was either 0 or 1. To be consistent with the other binary predictors in the dataset, which are categorical,

“outdoor_job” was reassigned to be factor type, and each value became either “Yes” or “No” corresponding to 1 or 0, respectively.

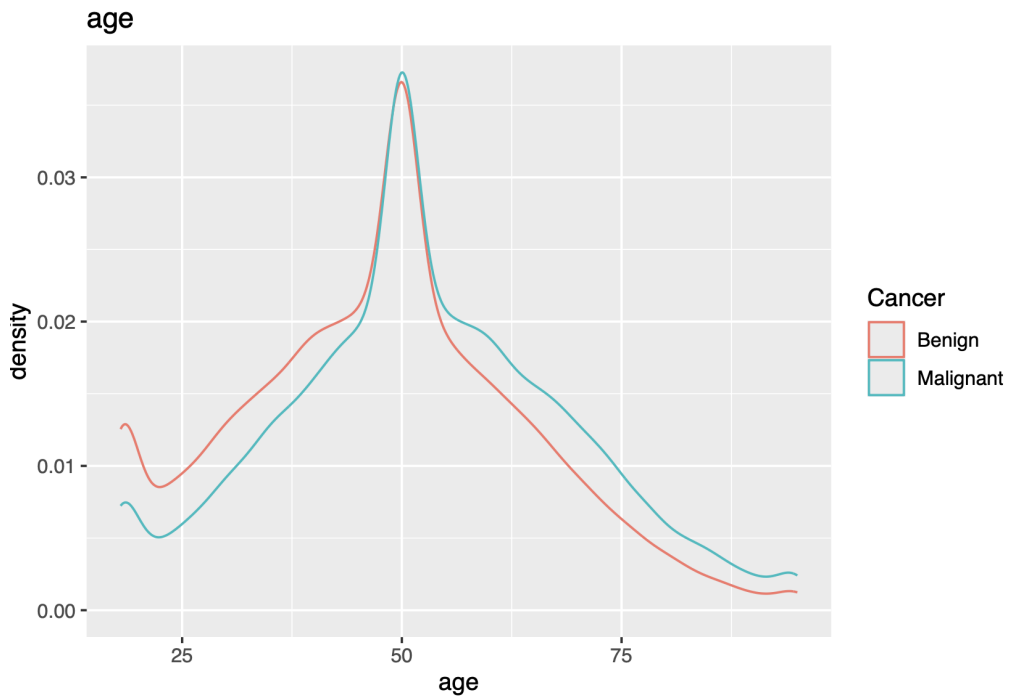
B. Imputation of Missing Values

The dataset was downloaded in pre-split training and test sets, consisting of 50,000 rows and 20,000 rows, respectively. The testing set contains the same variables as the training set, except for the target variable Cancer, which is omitted. We found that approximately 8% of the dataset contains missing values, which were evenly distributed throughout both categorical and numerical predictors in both the training and testing sets. Different imputation methods were experimented. These included simple imputation with median for numerical values and mode for categorical values, a random forest-based method using the package missRanger, and multiple imputation using the package MICE. For all models discussed in this report, multiple imputation using MICE was used. MICE uses predictive mean matching, which uses regression of known data values to predict the missing ones. The parameter “m” was set to produce 5 different imputed datasets. During model fitting, all candidate models were fitted onto all five datasets, and the results were pooled to obtain a single dataframe of testing set predictions. Compared to using a single imputed dataset, the pooling method preserves the uncertainty of the imputation process.

III. Exploratory Data Analysis and Feature Selection

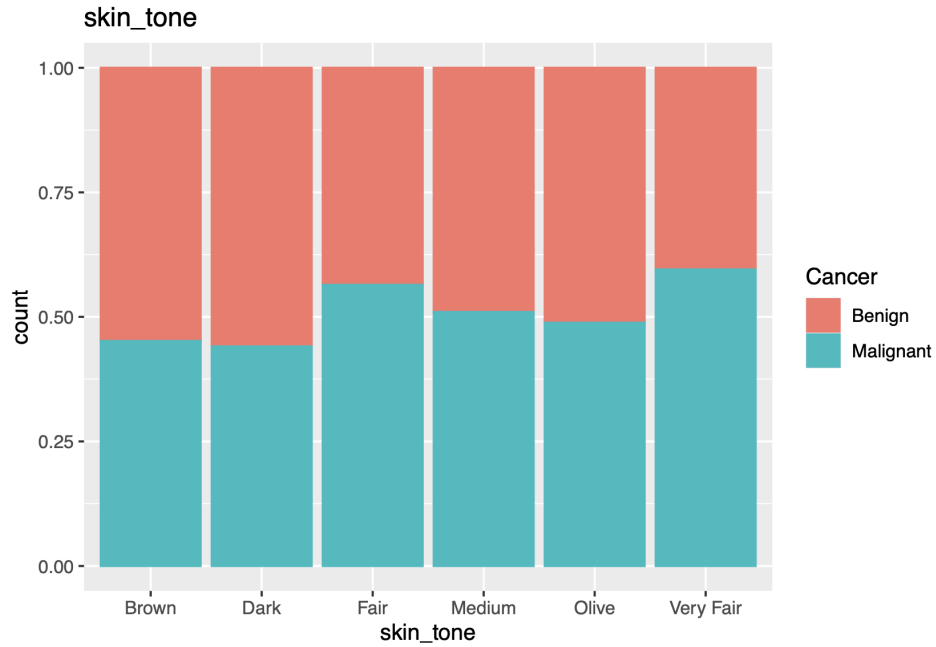
A. Numerical Predictors

Density plots were made of each candidate predictor to screen for predictors with distinct distributions Malignant and Benign classes. Few of the density plots showed distributions that were easy to distinguish, suggesting that there are few strong predictors among the candidates. Age was one of the best numerical predictors; the density plot below shows that “Benign” is much more likely for patients aged 20-45, and “Malignant” is more likely for patients older than 55.



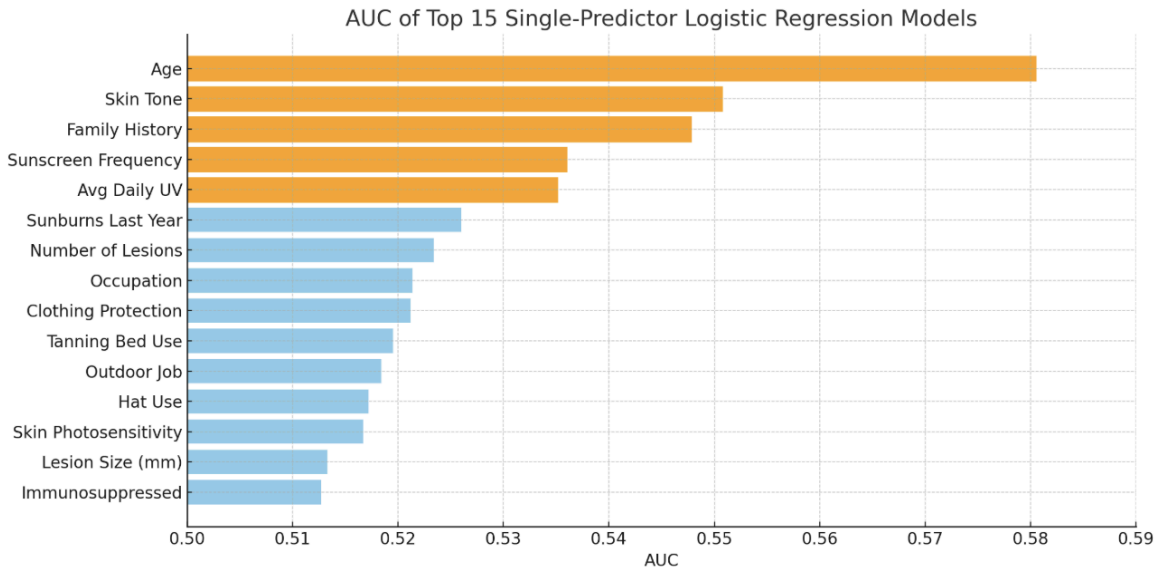
B. Categorical Predictors

For categorical predictors, stacked barplots were used to screen for "good" predictors. The distribution of skin tone, shown in the barplot below, was one of the more promising predictors, with fair and very fairskinned patients having higher proportions of malignancies than darker skin tones.



C. Single-Predictor Logistic Regression

To screen for relationships between the patient features and the outcome variable, single-predictor logistic regression was fitted to each candidate predictor. This allows us to quantify the potential predictive power of each feature individually as a prelude to feature selection. Area under the curve (AUC), z-statistic and p-value from logistic regression was used as the metric of the strength of the predictor/response relationship. AUC ranges from 0.5 to 1, where 0.5 represents the result of random chance (equivalent merely guessing) and 1 represents maximum predictive signal. The results of the predictors with the top 15 AUC are shown below, with the top 5 predictors highlighted:



More statistical details, including z-statistics and p-values from single-predictor logistic regression, are provided in the appendix.

D. Hypothesis Testing

Statistical hypothesis testing, using two-sample independent t-tests for numerical predictors and chi-square test for categorical predictors, was also used as an aid to feature selection. The null hypothesis for both t-tests and chi-square tests implies that no significant difference exists between the means of a particular feature and the response variable, while the alternative hypothesis implies that a true difference exists between the means of the two distributions. Thus, a test statistic with a large absolute value as well as a small p-value (below 0.05) indicates that a particular candidate predictor would be useful to us. Below is a table showing the results of hypothesis testing on each candidate predictor:

Table 1: Hypothesis Testing Results for the 20 Most Significant Predictors (by p-value)

Variable	Test Statistic	p-value
age	-31.9190	2.200e-16
family_history	566.0900	2.200e-16
skin_tone	416.7900	2.200e-16
avg_daily_uv	-13.9320	2.200e-16
sunscreen_freq	220.3300	2.200e-16
immunosuppressed	198.3300	2.200e-16
number_of_lesions	-12.3080	2.200e-16
sunburns_last_year	-11.2970	2.200e-16
tanning_bed_use	95.7610	2.200e-16
clothing_protection	78.9570	2.200e-16
outdoor_job	114.9500	2.200e-16
occupation	111.4600	2.200e-16
skin_photosensitivity	61.5750	4.258e-14
hat_use	49.8010	8.807e-11
lesion_size_mm	-5.1691	2.361e-07
sunscreen_spf	4.0756	4.598e-05
years_lived_at_address	-2.3267	1.999e-02
residence_lon	1.9231	5.448e-02
income	-1.7597	7.847e-02
desk_height_cm	-1.6169	1.059e-01

From the results of single-predictor logistic regression and hypothesis testing, we can categorize the strength of the 49 candidate predictors into one of four categories: Top 5, Strong, Weak, and Noise:

Table 2: Predictor Strength Categories

Top 5	Strong	Weak	Noise
1. age	1. clothing_protection	1. income	Everything else
2. family_history	2. sunburns_last_year	2. urban_rural	
3. skin_tone	3. number_of_lesions	3. years_lived_at_address	
4. sunscreen_freq	4. tanning_bed_use	4. desk_height_cm	
5. avg_daily_uv	5. outdoor_job		
	6. occupation		
	7. skin_photosensitivity		
	8. immunosuppressed		
	9. hat_use		

IV. Modeling

For each statistical learning method (Logistic Regression, Generalized Additive Models, Random Forest, Gradient Boosting) to be discussed, four candidate models were fit, one on all 49 predictors and three on subsets of the top 14, top 10, and top 5 of the best predictors. The predictors chosen to be in each subset are shown below:

Table 3: Predictor Subsets for Model Fitting

14 Predictors	10 Predictors	5 Predictors
age	age	age
skin_tone	skin_tone	skin_tone
avg_daily_uv	family_history	family_history
sunscreen_freq	avg_daily_uv	avg_daily_uv
hat_use	sunscreen_freq	sunscreen_freq
clothing_protection	immunosuppressed	
tanning_bed_use	number_of_lesions	
outdoor_job	sunburns_last_year	
family_history	tanning_bed_use	
immunosuppressed	outdoor_job	
lesion_size_mm		
number_of_lesions		
skin_photosensitivity		
sunburns_last_year		

A. Logistic Regression

Logistic regression is a classification method that models the probability of a categorical outcome variable using a logistic function. The advantage of logistic regression is that it is simple to implement, fast, computationally cheap, and produces results that are relatively easy to interpret. However, unlike tree-based models, logistic regression can only work with linear decision boundaries in the feature space.

Logistic regression was fit onto the 49, 14, 10, and 5 predictor subsets to produce four candidate models. Each model was used to predict on the testing data to produce four testing accuracies, which are shown in the table and plot below:

Table 4: Logistic Regression Model Performance

Number of Predictors	Testing Accuracy
49	0.60475
14	0.59685
10	0.59560
5	0.58975

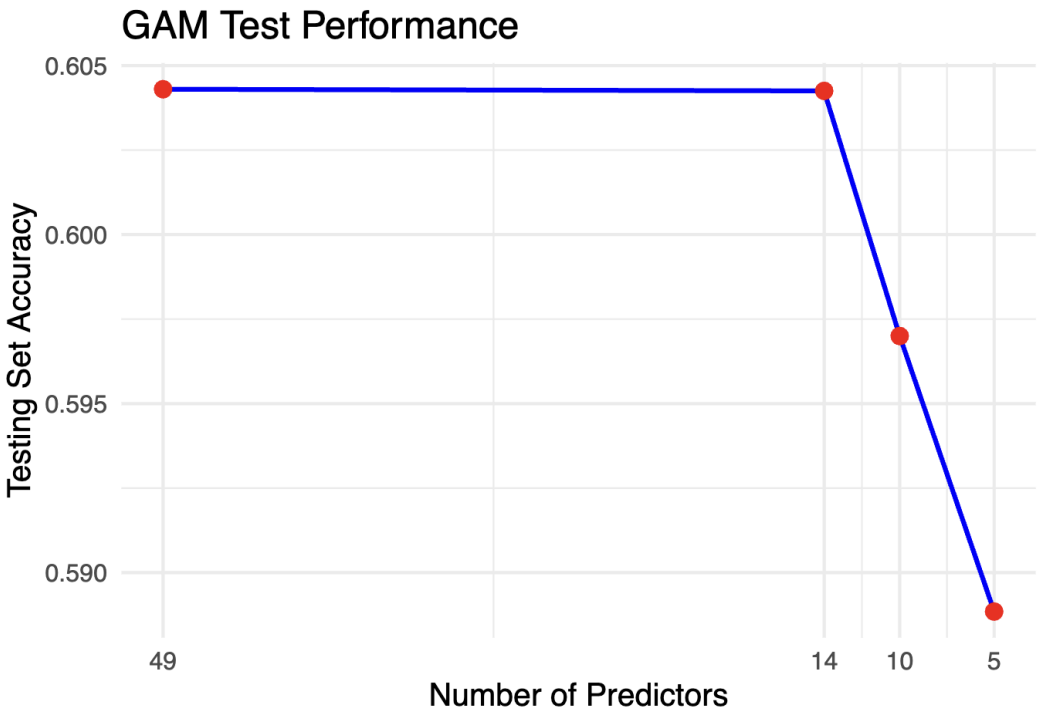
The full Logistic Regression model fit with all 49 predictors performed the best out of the 16 models discussed here, with a testing accuracy of 0.60475. However, since our goal is to maximize parsimony as well as testing accuracy, we did not select this as our best overall model. As we can see from the line plot above, the performance of Logistic Regression declined quickly as predictors were removed from the full model.

B. Generalized Additive Models (GAM)

Generalized Additive Models (GAM) is a moderately flexible statistical learning method, lying somewhere between logistic regression and tree-based modeling. Like logistic regression, it allows for interpretable results, while allowing modeling of nonlinear relationships using smoothed spline terms fitted on numeric predictors. We used the package mgcv to fit four GAM models. Spline functions were applied to age, avg_daily_uv, and lesion_size_mm. The testing accuracies of the four candidate models are shown below:

Table 5: GAM Model Performance

Number of Predictors	Testing Accuracy
49	0.60430
14	0.60425
10	0.59700
5	0.58885



GAM performed very well on our dataset, with test accuracy staying very stable when features were reduced from the full model to 14 predictors. However, accuracy declined precipitously and steadily when going from 14 to 10, and 10 to 5 predictors.

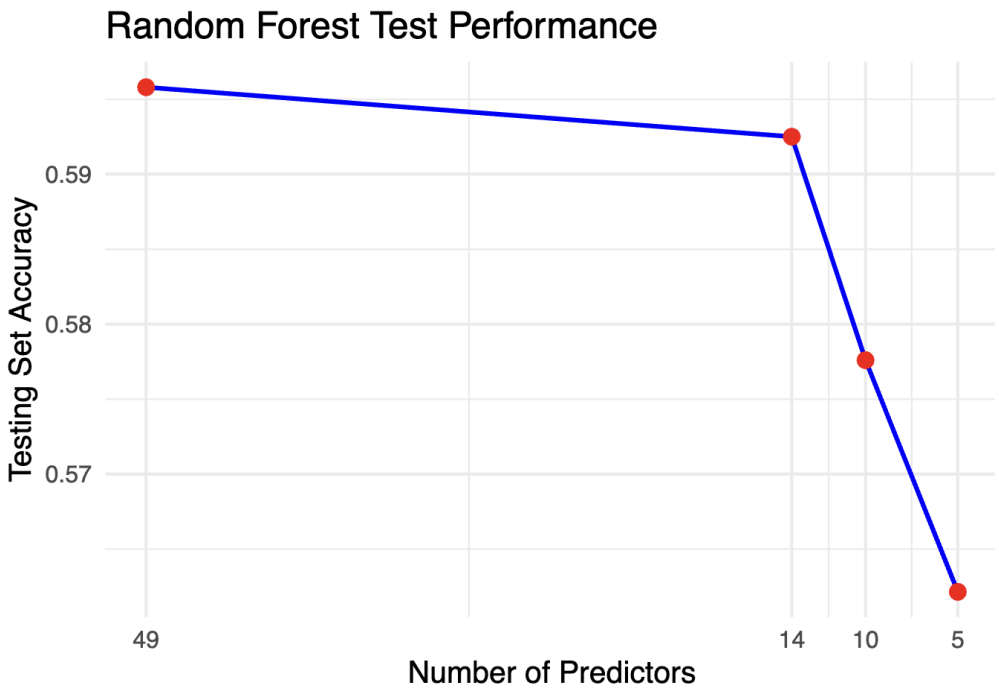
C. Random Forest

Random Forest is a tree-based learning method that excels at capturing complex nonlinear relationships. It uses individual decision trees trained on randomly selected subsets of the full predictor set. This results in trees that have high variance but little correlation. For classification, the final prediction is determined either using majority vote of each individual trees predictions, or using the average of each tree's predicted probabilities.

From a combination of experimentation and grid search, the hyperparameter set was chosen to be: num.trees = 1000, min.node.size = 5, mtry = floor(sqrt(ncol(train) - 1)).

Table 6: Random Forest Model Performance

Number of Predictors	Testing Accuracy
49	0.59580
14	0.59250
10	0.57760
5	0.56215



Surprisingly, Random Forest performed the worst out of all the learning methods discussed here, with testing accuracy rates well below those of other methods given a model fitted with the same predictors. Testing performance was stable when predictors were reduced to 14, though not as stable as GAM and GBM. Like all other methods, testing performance declined very quickly between 14 and 5 predictors. The Random Forest model fitted with 5 predictors was by far the worst performing model when testing accuracy is the metric.

D. Gradient Boosting Machines (GBM)

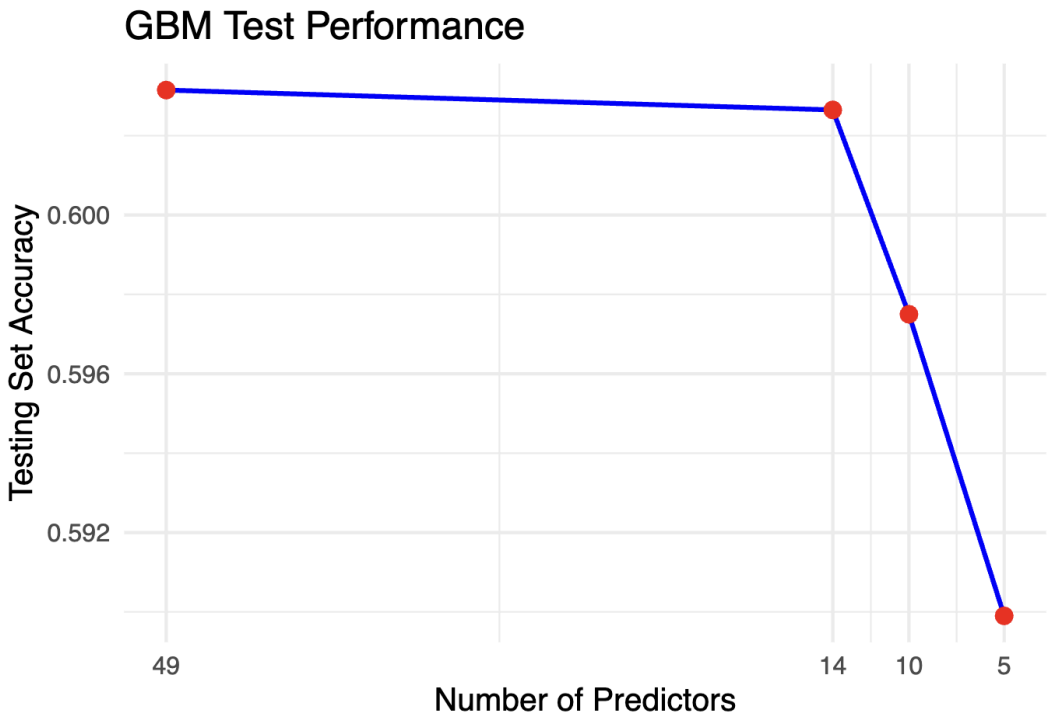
Gradient Boosting is another tree-based learning method, but unlike Random Forest, the trees are not trained independently. Instead, decision trees are trained sequentially on the residuals produced by the previous tree, with the goal of correcting that tree's errors. Like Random Forest, it excels at modeling nonlinearity and feature interactions, and is probably the most flexible methods of those discussed here.

Through a combination of experimentation and random search, the hyperparameter set was chosen to be: n.trees = 2000, interaction.depth = 3, shrinkage = 0.01, n.minobsinnode = 5, bag.fraction =

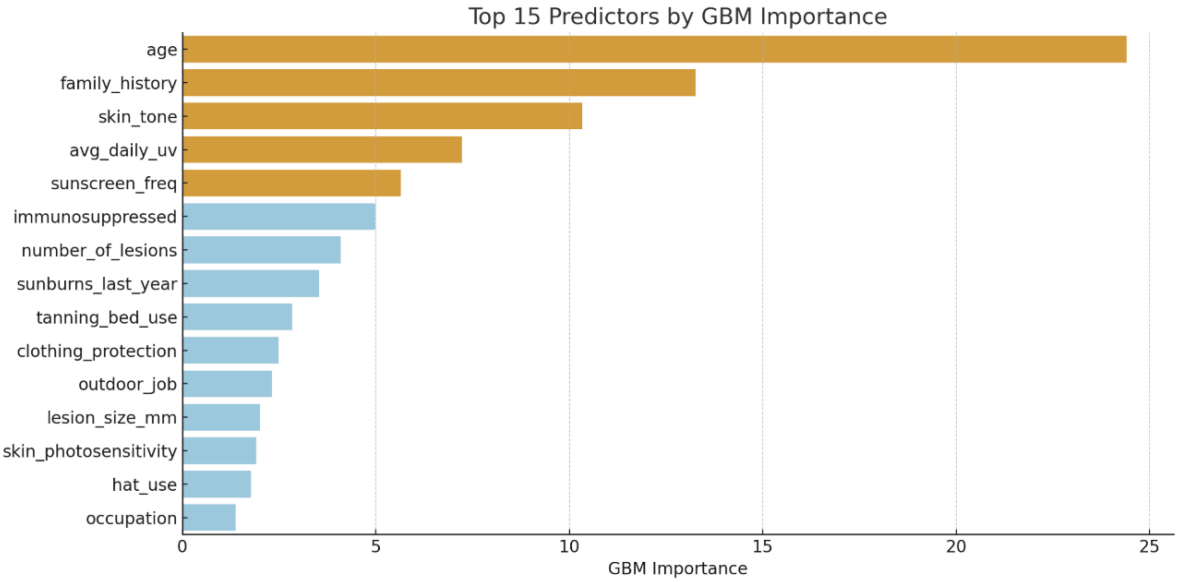
0.7. During model training, 5-fold cross validation was used to select the optimal number of trees, and the final model was fit with the best training iteration.

Table 7: GBM Model Performance

Number of Predictors	Testing Accuracy
49	0.60315
14	0.60265
10	0.59750
5	0.58990



GBM performed much better than the other tree-based method, Random Forest, with the trend of the Testing Accuracy vs Number of Predictors line following closely to that of GAM. Of the models trained on the 10-predictor subset, the GBM model performed the best. Of the models trained on the 14-predictor and 5-predictor subsets, the GBM models performed the second best. We also examined the feature importance of the predictors in the GBM model, plotted in the bar plot below, with the top 5 most important features highlighted in orange:

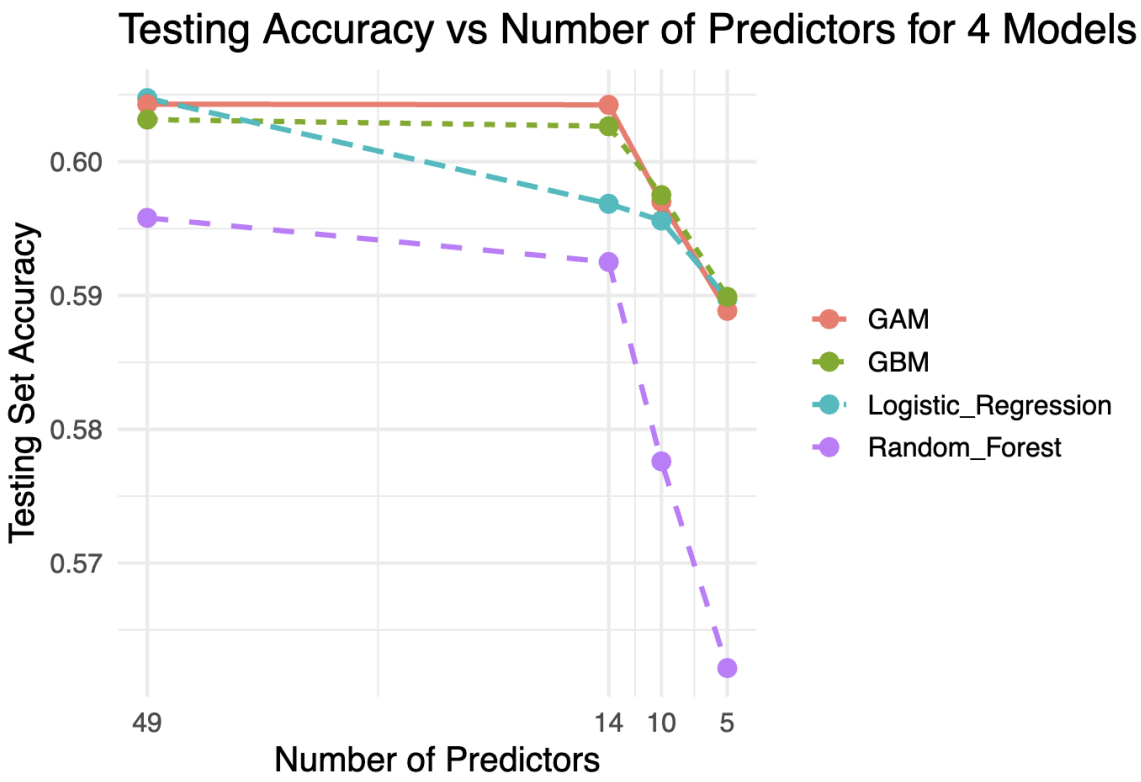


The top 5 features for GBM are the same as those selected using hypothesis testing and single-variable logistic regression as the strongest predictors during feature selection. This gives us confidence that age, family_history, skin_tone, avg_daily_uv, and sunscreen_freq offer the strongest predictive power for a binary classification model of Cancer trained on this dataset.

E. Other Models

Regularized logistic regression (LASSO, Ridge, Elastic Net), including bagged ridge trained on many bootstrapped samples, as well as XGBoost models and Support Vector Machines (SVM) were also tried. An SVM model fitted with the 14-predictor subset performed quite well with a testing accuracy of 0.60275, putting it second to GAM for models trained on that particular subset. Due to the computational expense of running SVM, we did not try training models on other subsets.

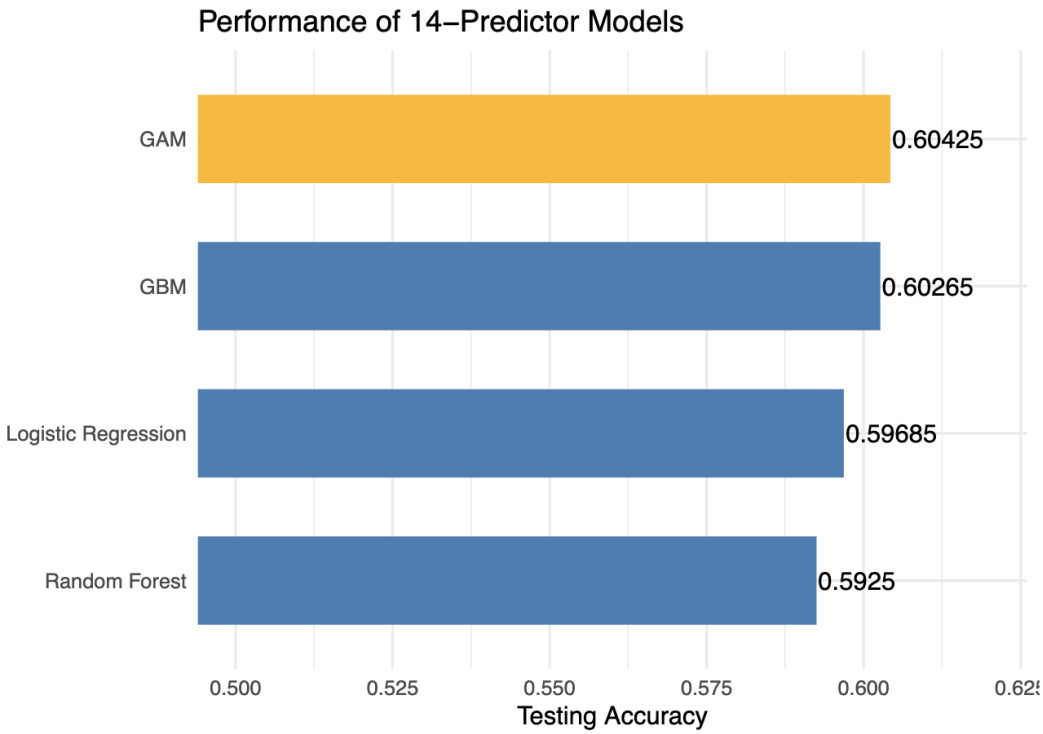
V. Discussion



From the line plot, we can see that the testing accuracy of every model decreases monotonically as predictors are reduced from 49 to 5. Most models, and especially GAM and GBM, stayed stable when going from 49 to 14 predictors. This suggests that most predictors eliminated in this transition were noise predictors that did not contribute actual signal to the model. In contrast, the slope of the decline from 14 to 10 predictors, and 10 to 5 predictors is steep, suggesting that predictors with actual signal were being eliminated. Random Forest declined the most between 49 and 5 predictors, and was the worse performing model overall.

VI. Conclusion and Limitations

A. Best Model



Testing performance began to decline greatly for all models between 14 and 10 predictors, and, besides Logistic Regression, the performance of full models and 14-predictor models is mostly similar. The 14 predictor subset was chosen as the predictor-subset that optimizes both goals of testing accuracy and parsimony. Among this subset, GAM was the winner with a testing accuracy of 0.60425, followed relatively closely by GBM with a score of 0.60265.

B. Limitations

One limitation is that candidate models performed similarly on the testing data, with most models in the 0.58 to 0.60 range. The difference between model testing accuracies is small enough that fluctuations from using different random seeds, imputation methods, or different train/test splits could produce them. Thus, it's hard to conclusively say that the results we got reflect actual differences in model adequacy, or were just a result of random chance.

Another limitation for this project is generalizability. The skin cancer dataset uses synthetically generated tabular predictors rather than image-based predictors obtained from skin lesions of real patients. Though this tabular dataset works for obtaining general insight and inference about what patient characteristics are related to malignancy, it has little actual clinical application. Machine learning based on high-quality images of patient skin lesions would be preferable to produce models that would be suitable for predictive clinical use.

VII. Appendix

A. Results of Single-Predictor Logistic Regression on 20 Candidate Predictors

Predictor	Coefficient (by category)	z-stat (by category)	p-value (by category)	Training Accuracy	AUC
Age	0.0165790	31.31	< 2e-16	0.55966	0.5805
Family History	2nd-degree: -0.30925 None: -0.56297 Unknown: -0.43715	2nd-degree: -7.985 None: -23.256 Unknown: -8.942	2nd-degree: 1.41e-15 None: <2e-16 Unknown: <2e-16	0.53478	0.5478
Skin Tone	Dark: -0.03433 Fair: 0.46100 Medium: 0.22955 Olive: 0.14485 Very Fair: 0.57674	Dark: -0.616 Fair: 12.849 Medium: 6.465 Olive: 3.833 Very Fair: 14.041	Dark: 0.538034 Fair: < 2e-16 Medium: 1.01e-10 Olive: 0.000127 Very Fair: < 2e-16	0.5402	0.5507
Avg Daily UV	0.064749	13.867	< 2e-16	0.53084	0.5351
Sunscreen Freq	Never: 0.27048Often: -0.01620Rarely: 0.30241Sometimes: 0.19694	Never: 7.773Often: -0.569Rarely: 9.576Sometimes: 6.714	Never: 7.66e-15 ***Often: 0.5692Rarely: < 2e-16 ***Sometimes: 1.89e-11 ***	0.5342	0.536
Hat Use	Never: 0.173461Often: 0.053387Sometimes: 0.079027	Never: 6.267Often: 1.831Sometimes: 2.697	Never: 3.68e-10 ***Often: 0.06716 .Sometimes: 0.00699 **	0.52286	0.5172
Clothing Protection	Low: 0.21938Medium: 0.11196	Low: 8.744Medium: 4.644	Low: < 2e-16 ***Medium: 3.42e-06 ***	0.52632	0.5212
Tanning Bed Use	Occasionally: 0.16804Often: 0.31068Past: 0.10361	Occasionally: 5.954Often: 8.004Past: 3.726	Occasionally: 2.62e-09 ***Often: 1.20e-15 ***Past: 0.000195 ***	0.52264	0.5195
Outdoor Job	Yes: 0.251998	10.717	< 2e-16 ***	0.52286	0.5184
Immunosuppressed	Yes: 0.653922	13.885	< 2e-16 ***	0.52264	0.5127
Lesion Size (mm)	0.012739	5.166	2.4e-07	0.52264	0.5133
Number of Lesions	0.13347	12.186	< 2e-16	0.52264	0.5234
Skin Photosensitivity	Low: -0.22365Moderate: -0.13361	Low: -7.241Moderate: -4.047	Low: 4.45e-13 ***Moderate: 5.20e-05 ***	0.52264	0.5167
Sunburns Last Year	0.13136	11.234	< 2e-16	0.52264	0.526

Occupation	Office: 0.060511Other: 0.007005Outdoor: 0.285607Retired: 0.053856Service: 0.043963Student: 0.055941	Office: 1.914Other: 0.147Outdoor: 8.473Retired: 1.548Service: 1.310Student: 1.305	Office: 0.0556 .Other: 0.8830Outdoor: <2e-16 ***Retired: 0.1217Service: 0.1902Student: 0.1918	0.52264	0.5214
Income	3.302e-07	1.758	0.0788	0.52264	0.5045
Urban/Rural	Suburban: -0.02979Urban: -0.04811	Suburban: -1.059Urban: -1.851	Suburban: 0.2898Urban: 0.0641	0.52264	0.5042
Years Lived at Address	0.002601	2.324	0.0201	0.52264	0.5056
Participates in Outdoor Sports	Yes: -0.02633	-1.369	0.171	0.52264	0.5029
Desk Height (cm)	0.0014441	1.617	0.106	0.52264	0.5047