Predicting Skin Cancer Status

Joey Lee (Lec 2)

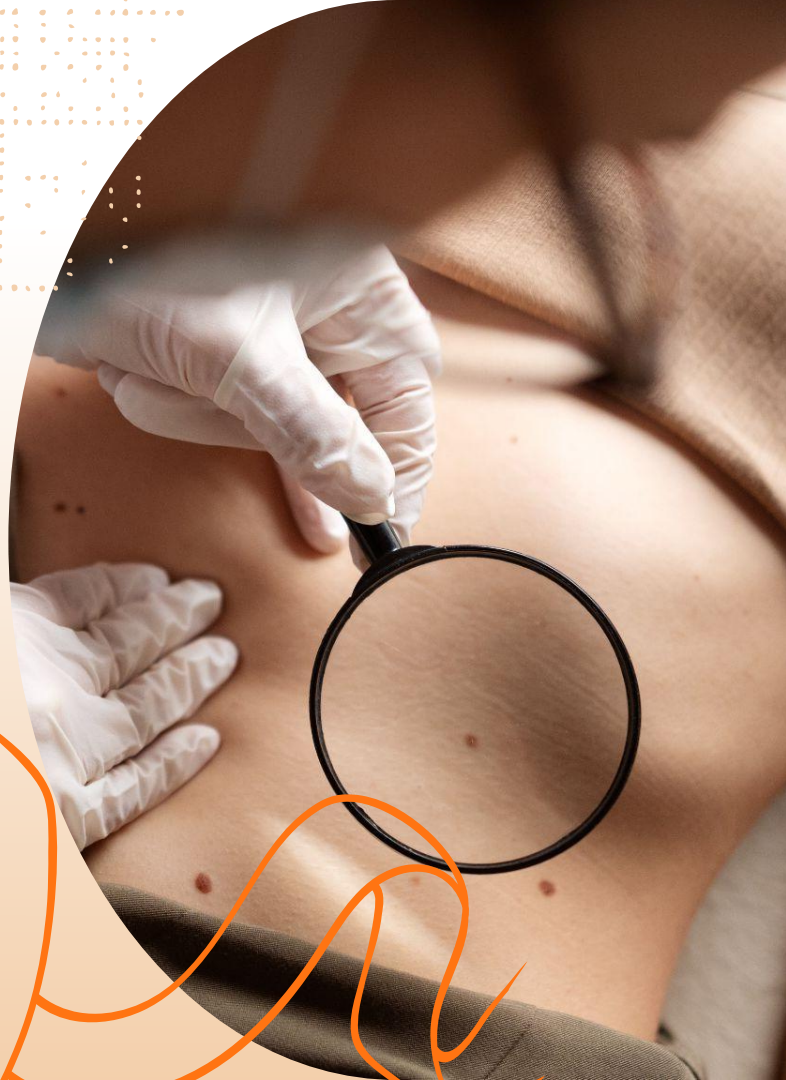


Table of contents

01 **Introduction**

Context, overview of dataset, data cleaning

02 **Preprocessing/EDA**

Data imputation, exploratory data analysis, hypothesis testing

03 **Modeling**

GAM, Logistic Regression, Random Forest, GBM

04 **Conclusion**

Winning model and project limitations

05 **Discussion**

Overview of testing performance

Introduction

Skin cancer is the most common type of cancer worldwide. It is estimated that around 20% of Americans will be affected by age 70.

Research shows that early detection can vastly increase survival rates. Early detection of melanoma increases 5-year survival rates to 99%. (Source: skincancer.org)

Introduction

This emphasizes the urgent need for reliable methods of detecting cancerous skin growths at a point when they can most easily be treated.

This project is an attempt to build a classification model that can reliably categorize abnormal skin growths into either “Malignant” or “Benign” based on characteristics of the patient and lesion(s).

Dataset Overview

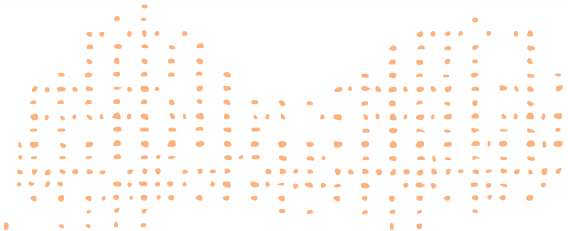
Observations

70,000 total
Each observation is one patient

(Training set: 50000, Testing set: 20000)

Predictor Variables

49
Continuous: 19,
Categorical, 30
(After data cleaning)



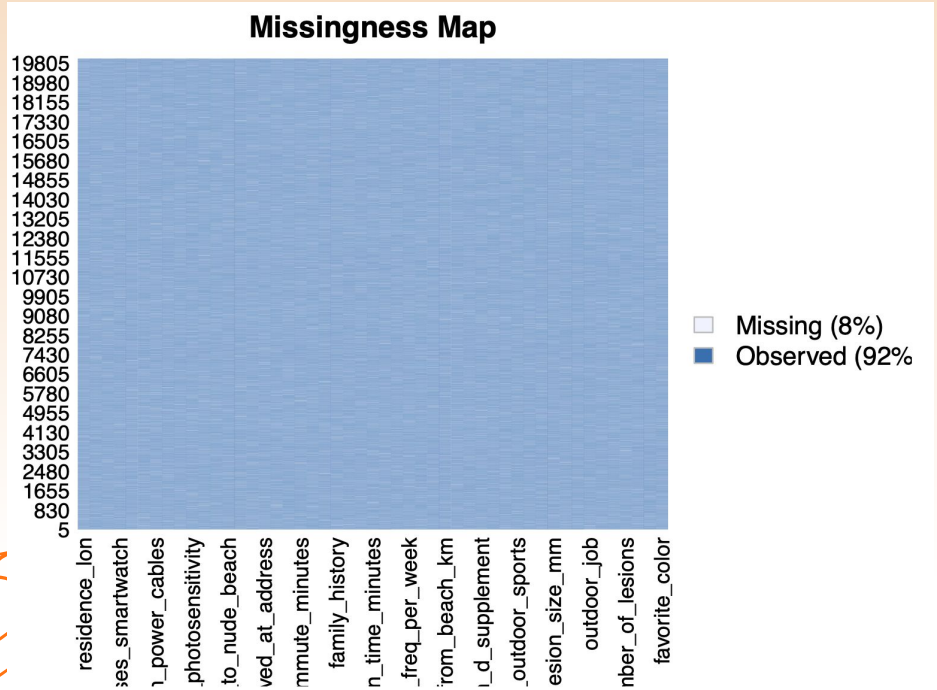
Response Variable

Binary categorical type, either “Malignant” or “Benign”

Data Preprocessing

Missing Data Handling

Missing Values



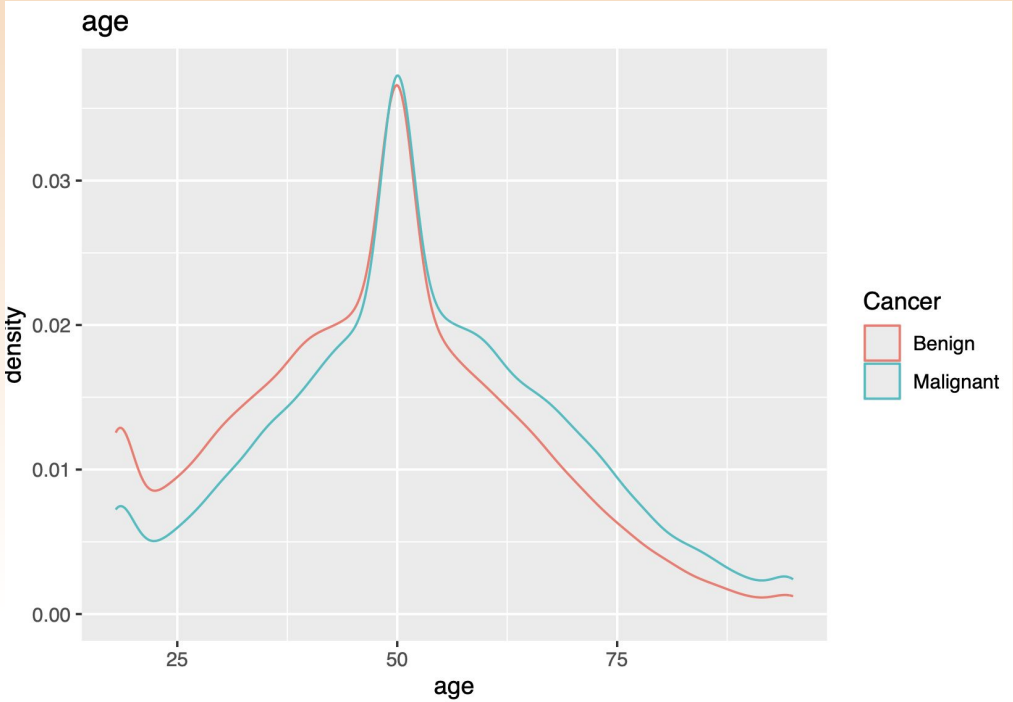
- Visualize missingness with library(Amelia)
- Around 8% of values in both train and test sets are missing

Data Imputation

- Multivariate Imputation by Chained Equations, via library(mice), used to impute all missing values
- Creates 5 different imputed datasets
- Each candidate model was fitted on each of the 5 imputed datasets, and their predictions were pooled and averaged to produce the final test set predictions
- Preserves the uncertainty of the imputation process

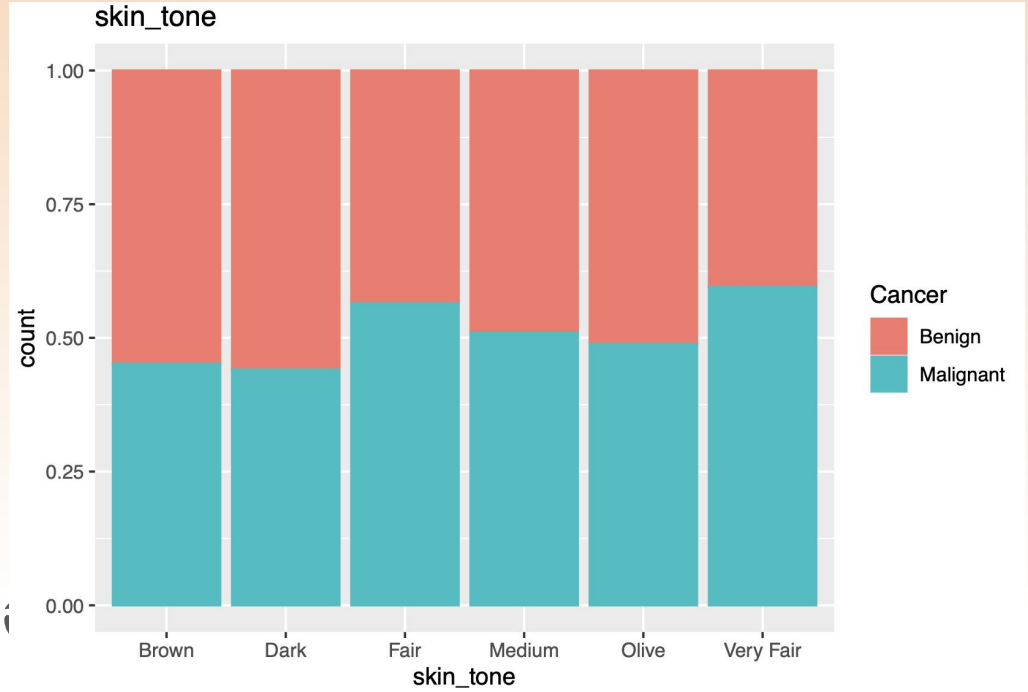
Exploratory Data Analysis

Some “good” predictors



- Age is one of the best numerical predictors
- Density plot shows that “Benign” is much more likely for patients aged 20-45, and “Malignant” is more likely for patients older than 55

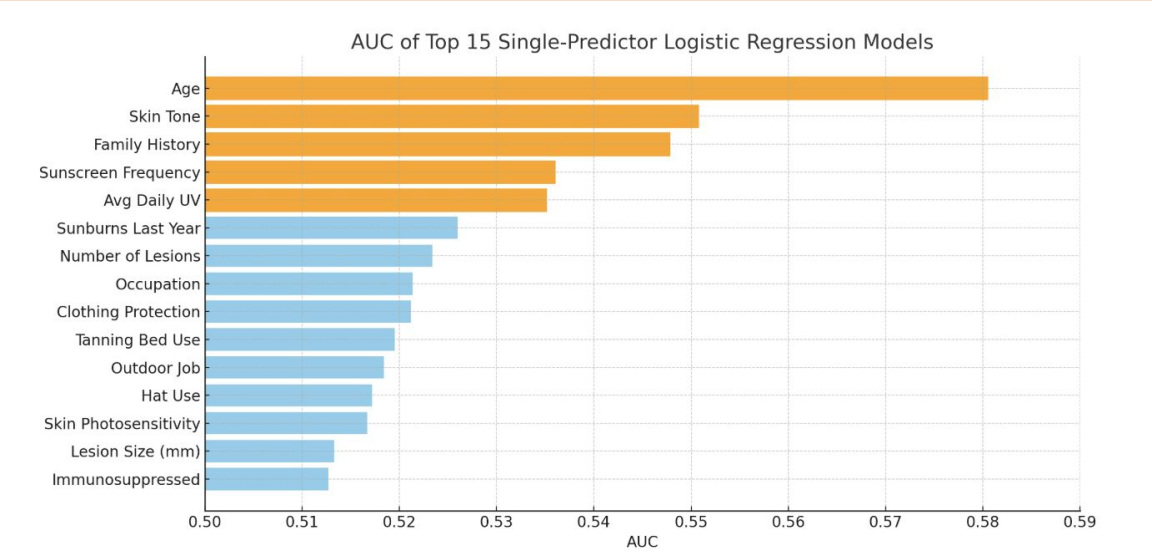
Some “good” predictors



- Skin tone is one of the best categorical predictors
- Fair and very fair-skinned patients have higher proportions of malignancies than patients with darker skin tones

Feature Selection

Single-Predictor Logistic Regression



- Regress Cancer on each individual predictor
- Used to quantify predictive power of an individual feature
- Area under the curve (AUC), z-statistic and p-value from logistic regression was used as measure of predictor strength
- AUC ranges from 0.5 to 1, where 0.5 represents the result of random chance and 1 represents maximum predictive signal

Hypothesis Testing

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

- Two-sample independent t-tests for numerical predictors
- Chi-squared test for categorical predictors
- Test statistic with a large absolute value and a small p-value (below 0.05) indicates that a particular candidate predictor would be useful

Categorizing Predictors

Table 2: Predictor Strength Categories

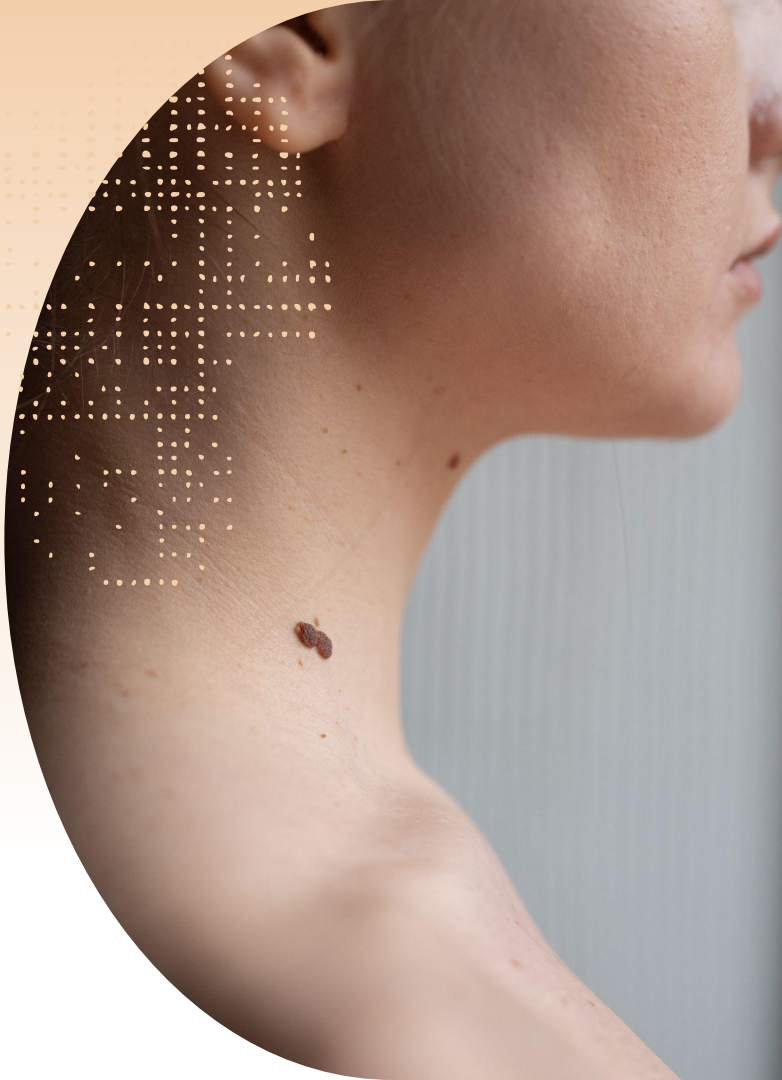
Top 5	Strong	Weak	Noise
1. age	1. clothing_protection	1. income	Everything else
2. family_history	2. sunburns_last_year	2. urban_rural	
3. skin_tone	3. number_of_lesions	3. years_lived_at_address	
4. sunscreen_freq	4. tanning_bed_use	4. desk_height_cm	
5. avg_daily_uv	5. outdoor_job		
	6. occupation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

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

- Top 5 (good predictive signal)
- Strong (some predictive signal)
- Weak (little predictive signal)
- Noise (no predictive signal)

Modeling

Generalized Additive Models, Logistic Regression, Random Forest, Gradient Boosting



Predictor Subsets

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		

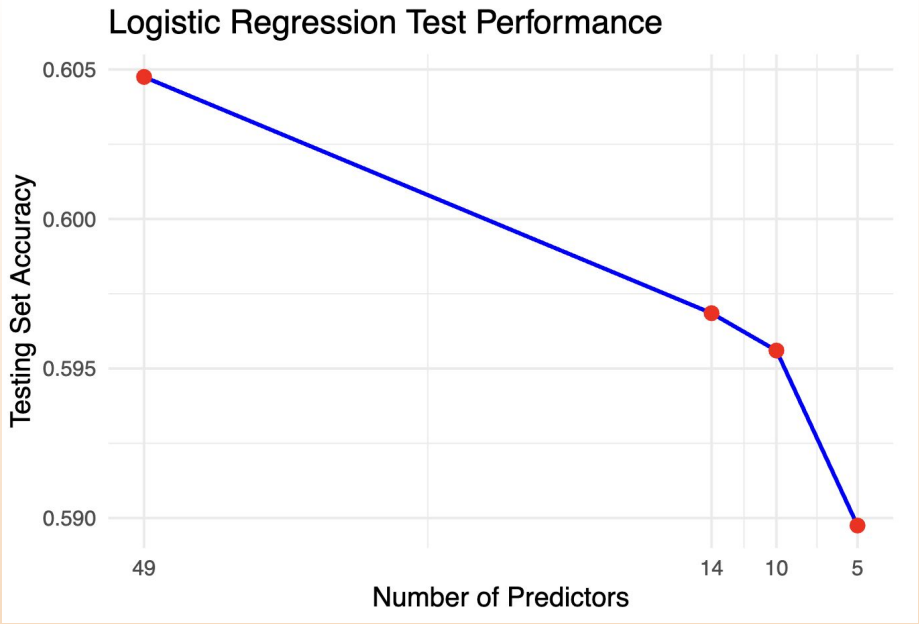
- Four candidate models were fit to each statistical learning method (Logistic Regression, Generalized Additive Models, Random Forest, Gradient Boosting)
- One on all 49 predictors (full model)
- Three on subsets of the top 14, top 10, and top 5 of the best predictors

Logistic Regression

Table 4: Logistic Regression Model Performance

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

- Full model had good performance, but test accuracy declined quickly as predictors were removed

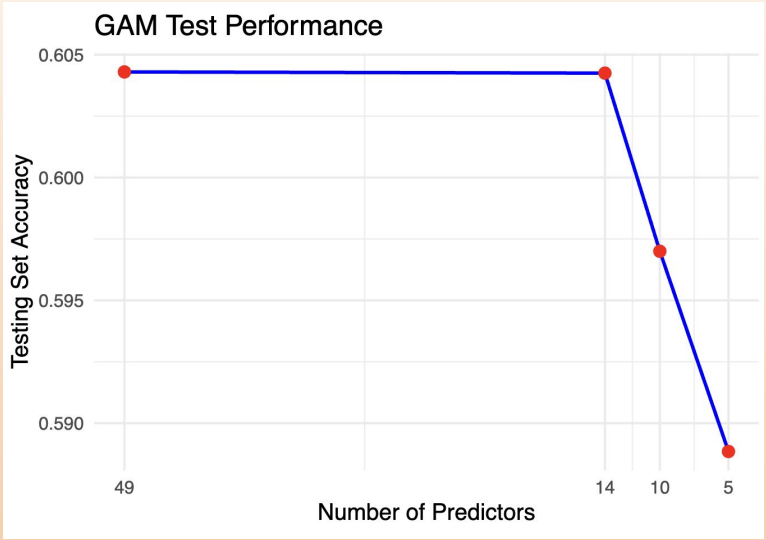


Generalized Additive Models (GAM)

- Allows moderate flexibility
- Performed well overall
- Testing accuracy robust when predictors were reduced from full model to 14
- However, accuracy declined precipitously when going from 14 to 10, and 10 to 5

Table 5: GAM Model Performance

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



Random Forest (ranger)

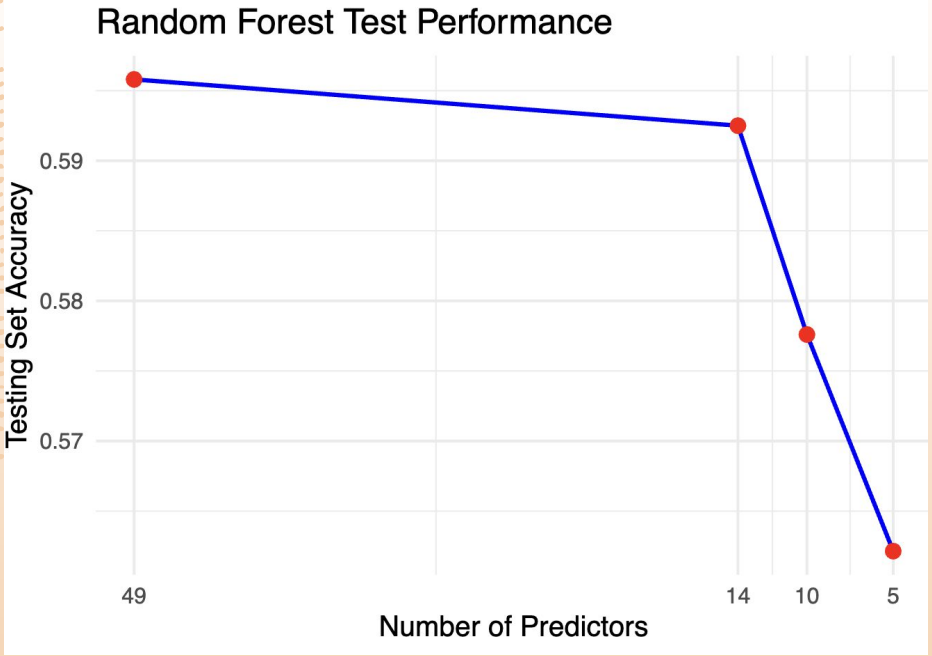


Table 6: Random Forest Model Performance

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

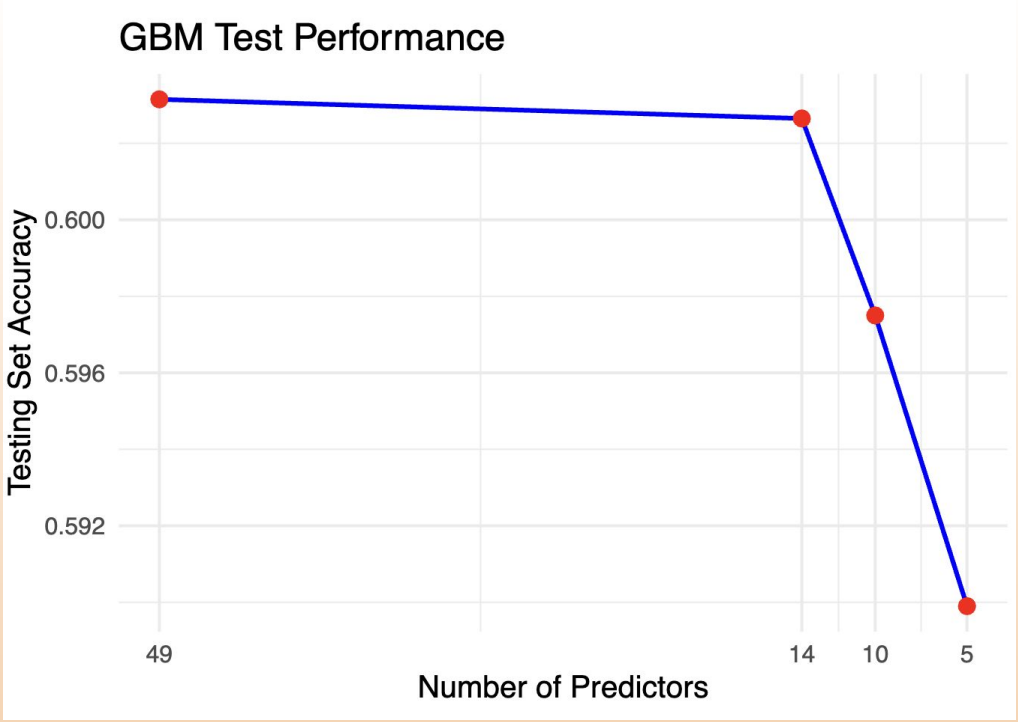
- Worst performer overall
- Testing accuracy rates well below those of other methods given a model fitted with the same predictors
- Random Forest model fitted with 5 predictors the worst performing model overall (when testing accuracy is the metric)

Gradient Boosting (GBM)

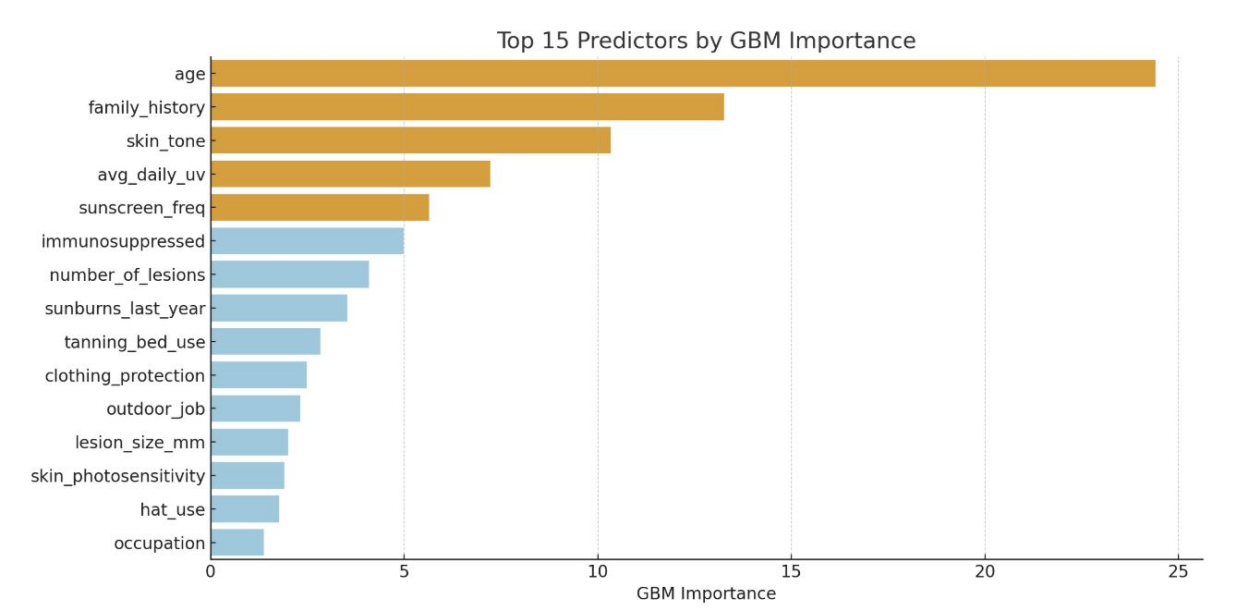
Table 7: GBM Model Performance

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

- Trains decision trees sequentially; most flexible method tried
- Performed better than RF, and similar to GAM in accuracy vs predictor trend



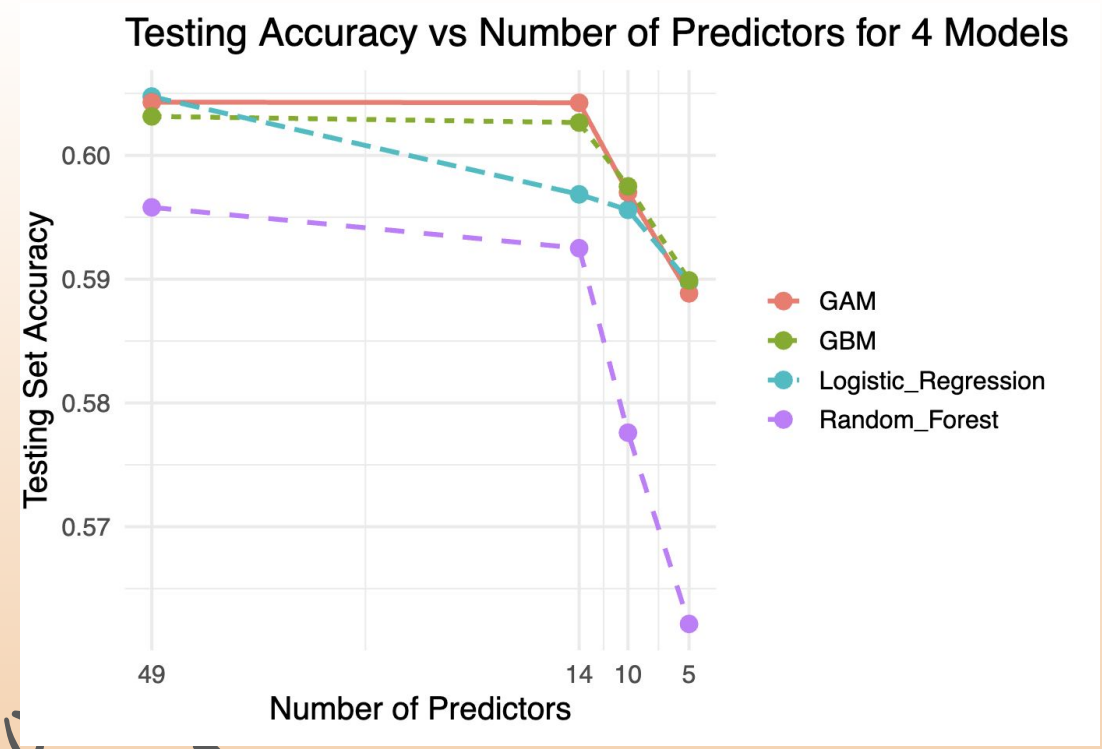
GBM: Feature Importance



Five most important features (in orange) are the same as those selected using hypothesis testing and single-variable logistic regression as the strongest predictors during feature selection

Discussion

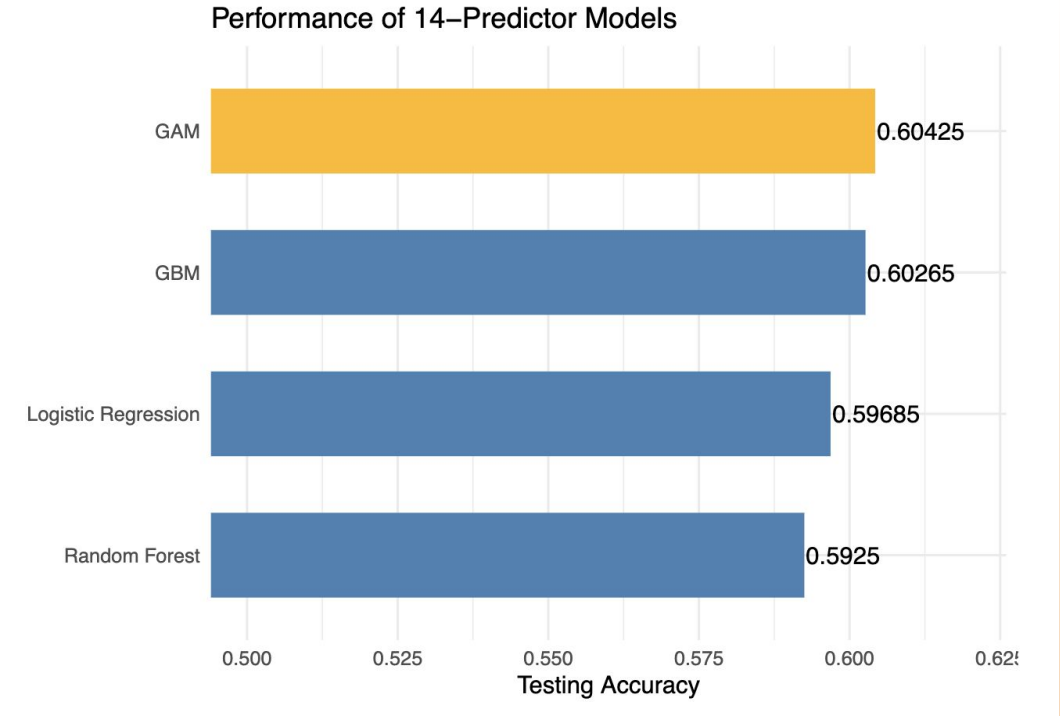
Overall Model Performance



- Testing accuracy of every model decreased monotonically as predictors are reduced from 49 to 5
- For most models, accuracy stayed stable when going from 49 to 14 predictors, suggesting that most predictors eliminated were noise predictors
- In contrast, the slope of the decline from 14 to 10 predictors, and 10 to 5 predictors is steep. This suggests that predictors with actual signal were being eliminated at these stages

Conclusion

Best Model: GAM with 14 Predictors



- **Goal**: Pick model that
  - maximizes test set accuracy
  - minimizes number of predictors
- Models trained on 14-predictor subset achieve the best balance of both goals
- Of these, GAM came out on top, while Random Forest did the worst

14

predictors

0.60425

Test Set Accuracy

Moderate

Flexibility

Best Model: GAM with 14 Predictors

Numerical

s(age)
s(avg_daily_uv)
s(lesion_size_mm)
number_of_lesions
sunburns_last_year

Categorical

skin_tone
sunscreen_freq
hat_use
clothing_protection
tanning_bed_use
outdoor_job
family_history
immunosuppressed
skin_photosensitivity

Limitations

Similar Model Performance

- Across learning methods, testing accuracies ranged from 0.58 to 0.60, i.e. a difference of only ~0.02
- This is well within fluctuations caused by a different random seed for seed-dependent methods, or different imputation methods
- Can't tell if differences in test accuracies reflect differences in actual model adequacy or just randomness

Generalizability

- Dataset is synthetic, can't generalize to real cases of skin cancer
- Actual machine learning for skin cancer probably relies on analyzing images of skin lesions more so than patient characteristics

Thanks!

Please keep this slide for attribution

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon** and infographics & images by **Freepik**