SKIN CANCER PREDICTION

THE CHALLENGE

Skin cancer is one of the most common cancers, and early detection is essential for improving patient outcomes.

In this project, we aim to build a predictive model that determines whether a skin lesion is **Benign** or **Malignant**, using demographic, biological, and lifestyle features.

TABLE OF CONTENTS

01

DATA OVERVIEW

What does the data consist of?

THE DATA

#	age		gender		skin_tone		education		income		urban_rural
1 50.0k	NA	8%	Female	47%	Medium	28%	Some College	28%	NA	8%	Urban
	18	4%	Male	44%	Fair	25%	Bachelor's	26%	17644	0%	Suburban
	Other (44227)	88%	Other (4486)	9%	Other (23485)	47%	Other (23099)	46%	Other (45934)	92%	Other (11017)
1	46		Male		Olive		Bachelor's		62888		Urban
2	78		Female		Fair		High School		NA		NA
3	54		Male		NA		Some College		18635		Urban
4	73		NA		Medium		High School		61623		NA
5	39		Female		Medium		Bachelor's		NA		NA
6	49		Male		NA		NA		22178		Urban
7	53		Male		Brown		NA		NA		Urban
8	65		Male		NA		High School		48868		Urban
9	NA		Male		Medium		High School		39618		Urban
10	74		Female		Brown		Some College		43425		Urban
11	55		Male		Olive		Bachelor's		8933		Suburban
12	70		Male		Very Fair		Bachelor's		8970		Suburban
13	49		Female		Fair		NA		20209		Suburban
14	62		Male		Olive		NA		348729		Suburban
15	50		Female		Medium		High School		9230		Urban
16	33		Female		Fair		High School		22389		Urban
17	61		Female		Very Fair		High School		53267		Urban

50,000

TRAINING OBSERVATIONS

20

NUMERIC PREDICTORS

29

CATEGORICAL PREDICTORS

1

RESPONSE VARIABLE (CANCER)

02 METHODOLOGY

Data Cleaning, Tested Models, Process

OUR PROCESS

01
DATA CLEANING

Dealing with NA's and type conversion

02
FEATURE SELECTION

Diagnostic Plots and Multicollinearity

03
MODEL BUILDING

Modeling Training and Selection

04
MODEL ANALYSIS

Misclassification rates and CV accuracy

05
MODEL SELECTION

Optimal Accuracy and Model Interpretability

01
DATA CLEANING

NA VALUES

7.69%

of the data were NA values

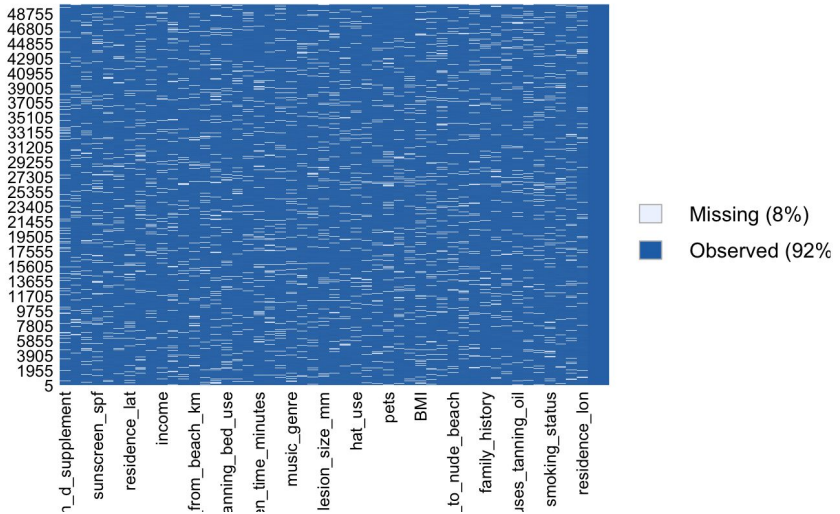
vitamin_d_supplement

had the most NA values with **4158** NAs

residence_lon

had the least number of NA values with **3878** NAs

MISSINGNESS MAP



IMPUTATION TECHNIQUES

01.

Mice

Fills in missing values by repeatedly predicting each variable from all the others, refining its guesses over several iterations.

02.

MissForest

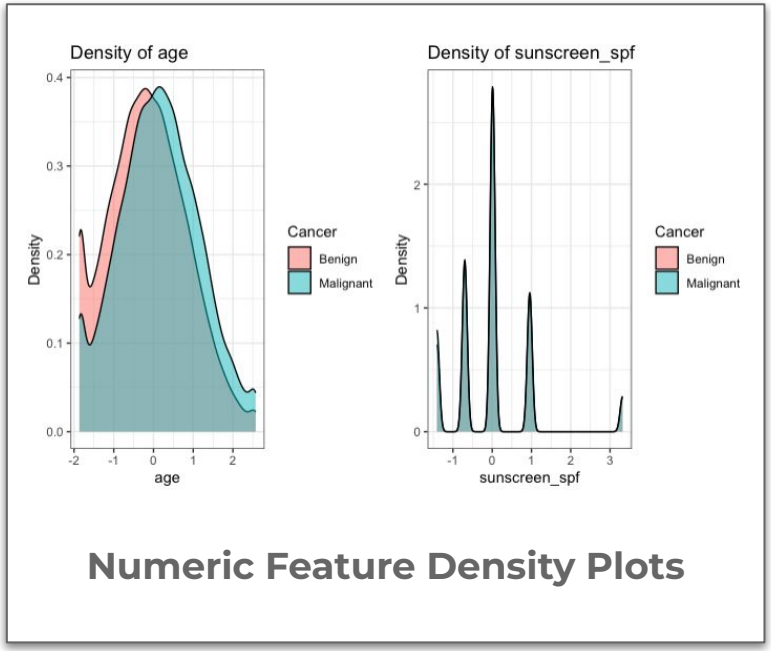
For each predictor, uses Random Forests with all other variables to fill in missing values.

03.

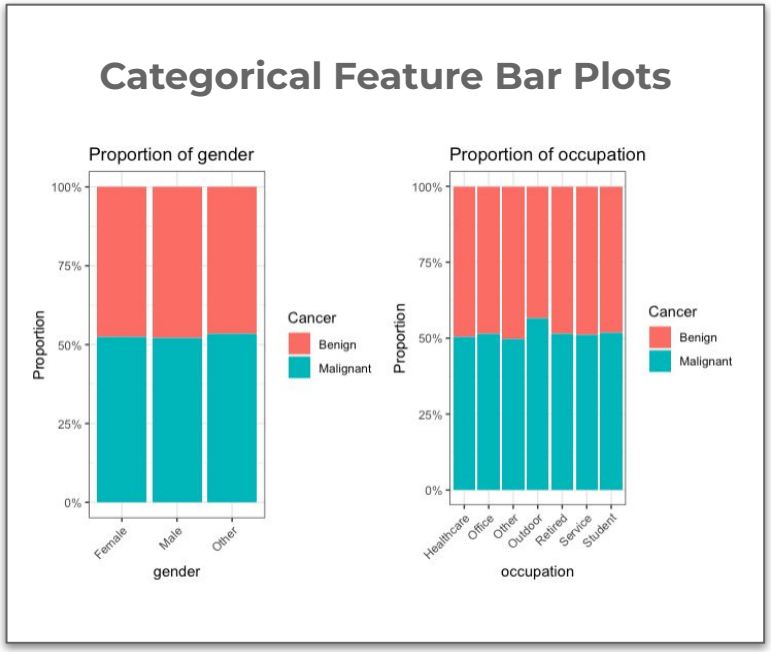
Mean/Mode

Directly calculating the mean (numeric) and mode (categorical) of each predictor and applying it to NA values.

FEATURE SELECTION



Numeric Feature Density Plots



02
FEATURE SELECTION

OTHER TECHNIQUES

ANOVA – Numeric

Ranked numeric predictors using ANOVA F-statistics, selecting features with the strongest separation.

Multicollinearity Filtering

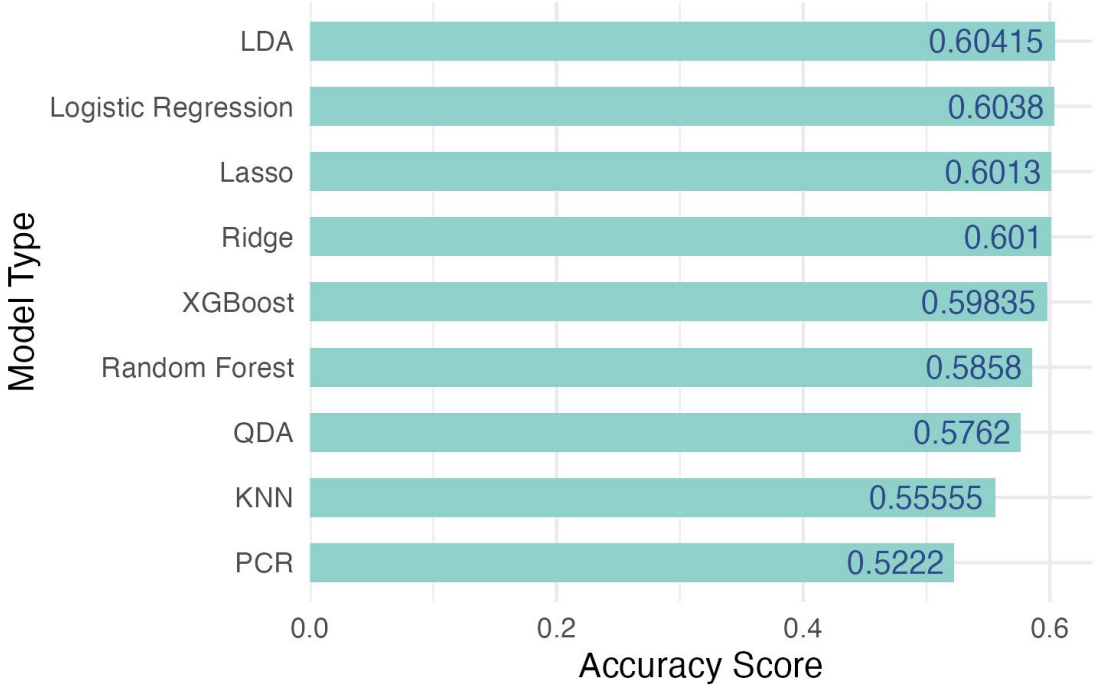
Removed numeric predictors under a certain correlation threshold to prevent redundant information.

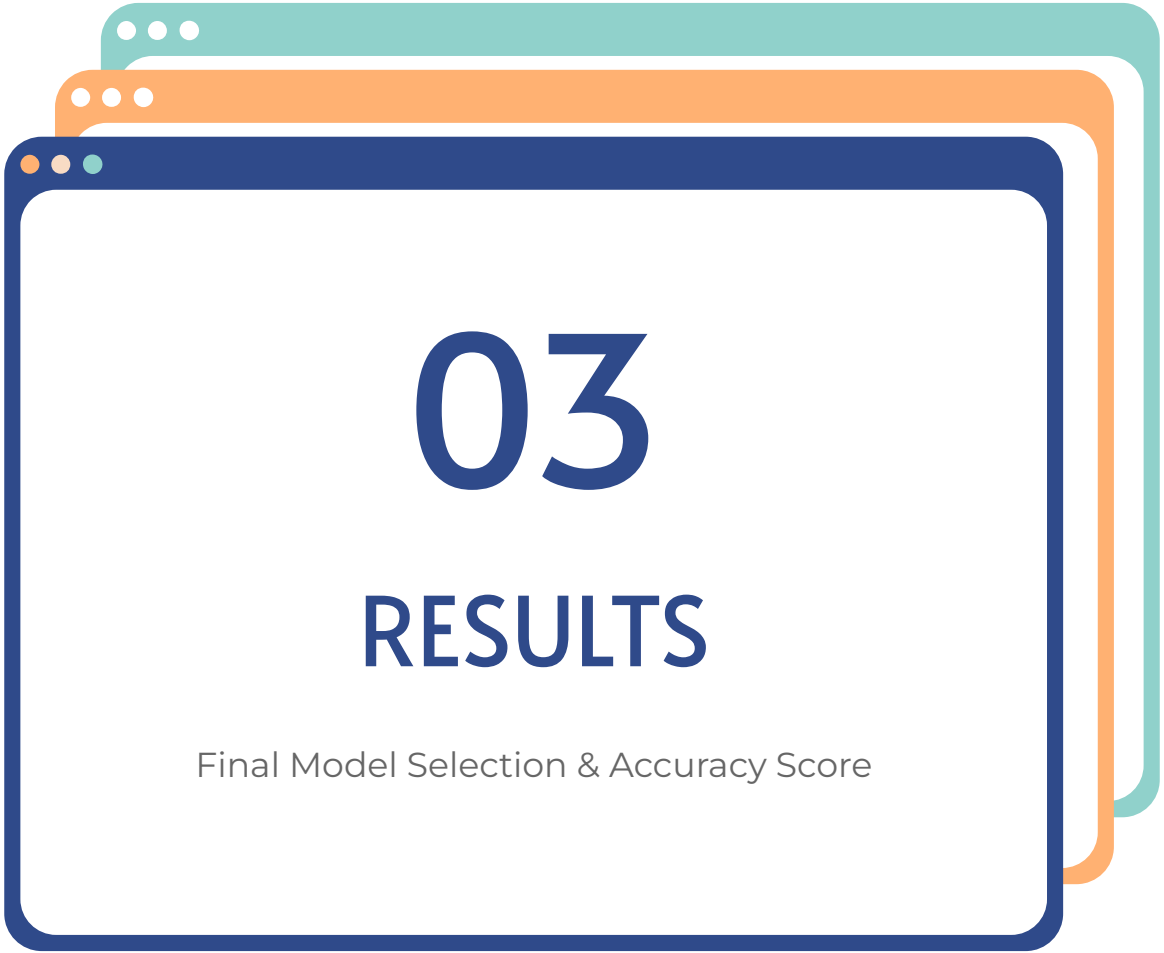
Chi-Square – Categorical

Ranked categorical features using chi-square significance testing, selecting features with the strong association with Cancer.

04
MODEL ANALYSIS

MODEL ACCURACY COMPARISON





03

RESULTS

Final Model Selection & Accuracy Score

FINAL LDA MODEL PIPELINE

STEP 1

UNUSEFUL PREDICTORS

Using **Lasso Shrinkage method**, we removed a subset of predictors that we deemed as “noise.”

STEP 2

We standardized all numerical variables using **z-score scaling** for models that are sensitive to scaling.

SCALE DATA

STEP 3

MULTICOLLINEARITY

We computed a **correlation matrix** and removed numeric predictors that are **too highly correlated** (> 0.8).

FINAL MODEL

MODEL TYPE

Linear Discriminant Analysis (LDA)

TRAIN

Accuracy:
0.6085
CV:
0.6059

TEST

Accuracy:
0.60415

04

DISCUSSION

Interpretability, Setbacks, Findings

OUR MOST IMPORTANT PREDICTORS

HEALTH HISTORY

Immunosuppressed, family history

SKIN/LIFESTYLE

Skin tone, sunscreen frequency, clothing protection, skin photosensitivity, tanning bed, outdoor job

DEMOGRAPHICS

Age, occupation

SETBACKS

IMPUTATION

Time-intensive with many parameters to adjust.

FEATURE SELECTION

Many combinations of “best” features.

FEATURE ENGINEERING

Did not seem to help our accuracy.

NEXT STEPS

IMPUTATION

Optimize parameter settings.

FEATURE SELECTION

Identify the most predictive set of variables.

FEATURE ENGINEERING

Experiment with more interaction effects.