Predicting Skin Cancer Status

By: Hartejh Surikapuram, Anika Chowdhury, Lucy Borkowski, Justine Shu, Sriya Donepudi

Stats 101C Lecture 1

Table of contents

01 Introduction

02 Data/Data Cleaning

03 Methodology

05 Models

04 Results/Limitations

01

Introduction

Why Skin Cancer Prediction Matters

- Skin cancer is the most common cancer in the United States, affecting millions each year
- 1 in 5 Americans will develop skin cancer by age 70
- Over 2 people die every hour from skin cancer in the U.S.
- Melanoma survival is very high when caught early, but risk depends on many factors that are hard to assess manually
- Machine learning improves detection by combining these factors and spotting patterns that manual screening often misses.

https://www.skincancer.org/skin-cancer-information/skin-cancer-facts/

02

Data/Data Cleaning

Data

Dataset Summary

- Training data has 50,000 rows and 50 columns
- Testing data has 20,000 rows and 49 predictors (same features as training but no Cancer label)

Cancer Distribution

- Training set is roughly balanced:
  - Benign: 47.7 percent
  - Malignant: 52.3 percent

Baseline accuracy: 0.5226
Baseline error rate: 0.4774

Predictor Types

- 49 predictors total in the modeling set
- 18 numeric predictors
- 31 categorical predictors
- Some variables are technically numeric but represent categories
  - outdoor_job: coded 0/1 but represents Yes/No
  - zip_code_last_digit: values 0–9, but not a quantitative measure

03

Methodology

Methodology/Imputation

- Training set: 196,042 missing values (**7.84%**)
- Testing set: 78,256 missing values (**7.99%**)

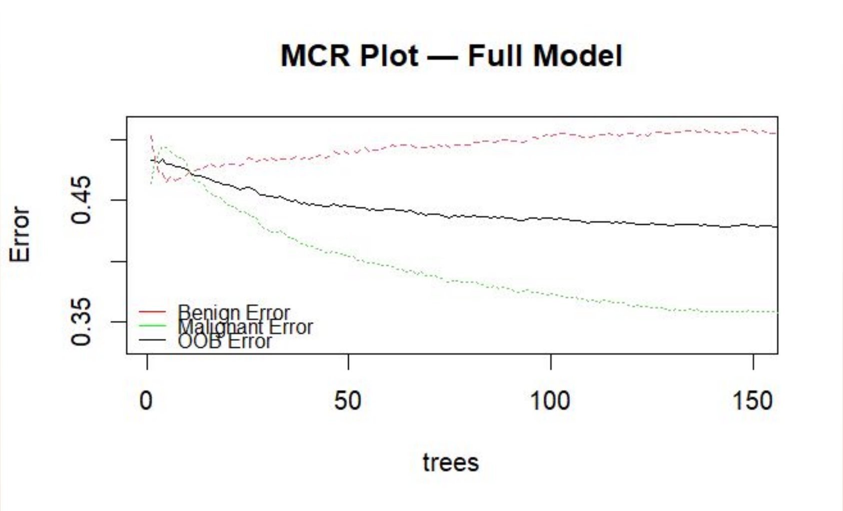
- Missingness is evenly spread and no predictor has more than 10 percent missing, so didn’t to drop any predictors due to excessive missingness

- We compared several approaches for handling missing values
    - ex. missForest, mice, and simple replacements like mean/median for numerical and mode for categorical data

04

Models

Random Forest Full Model



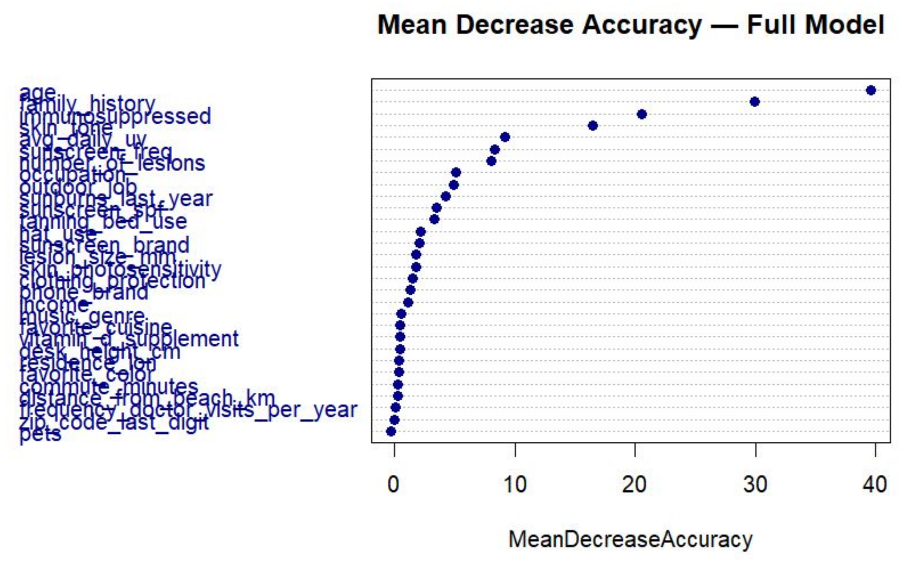
Full Model Training Confusion Matrix			
	Benign	Malignant	Class Error
Benign	11684	12184	0.5104743
Malignant	8656	17476	0.3312414

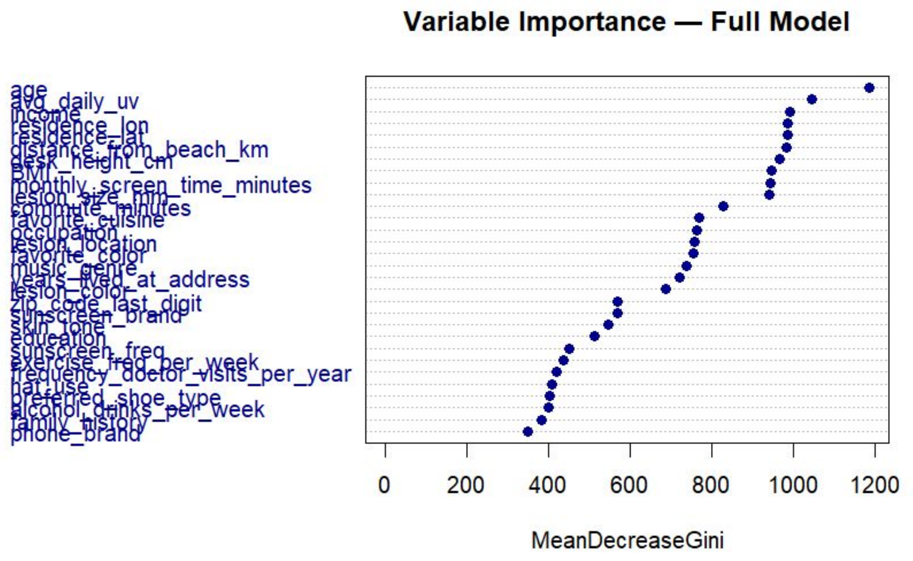
We used a random forest model using all of the predictors. For the training data we were able to see that the Malignant Error deceased as more trees were added while the Benign error increases, so the model was able to recognize the patterns for malignant better.

The full random forest model misclassified Benign 51% of the time and Malignant by 33%, this suggests that this model was biased to Malignant.

On Kaggle: Testing accuracy was 0.59835

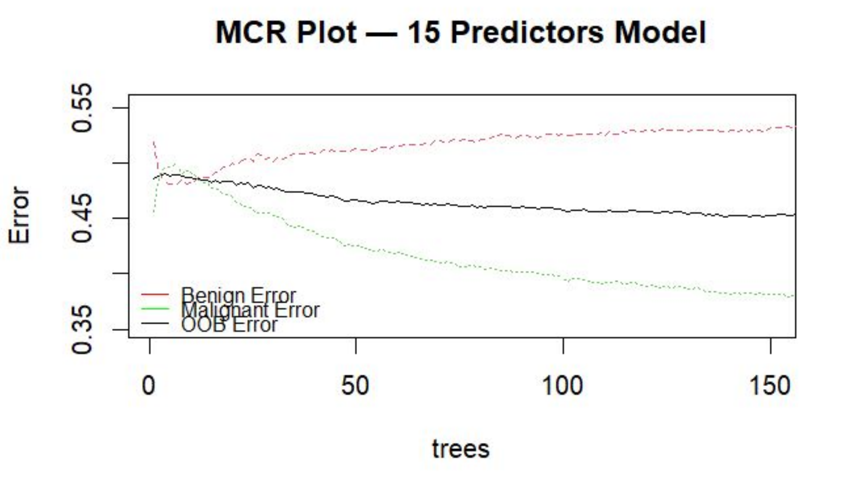
Random Forest Full Model


Mean Decrease Accuracy — Full Model


Variable Importance — Full Model

With so many predictors we wanted to keep influential predictors while removing those that were of low importance and only adding noise. So using this we wanted to perform feature reduction to potentially reduce overfitting.

Random Forest 15 Predictors Model



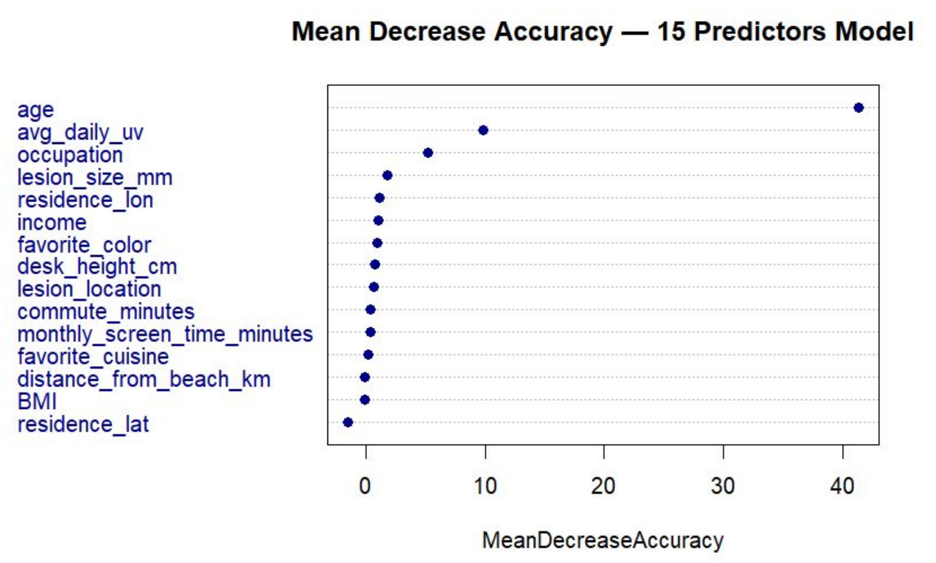
15 Predictor Model Training Confusion Matrix			
	Benign	Malignant	Class Error
Benign	10686	13182	0.5522876
Malignant	9161	16971	0.3505664

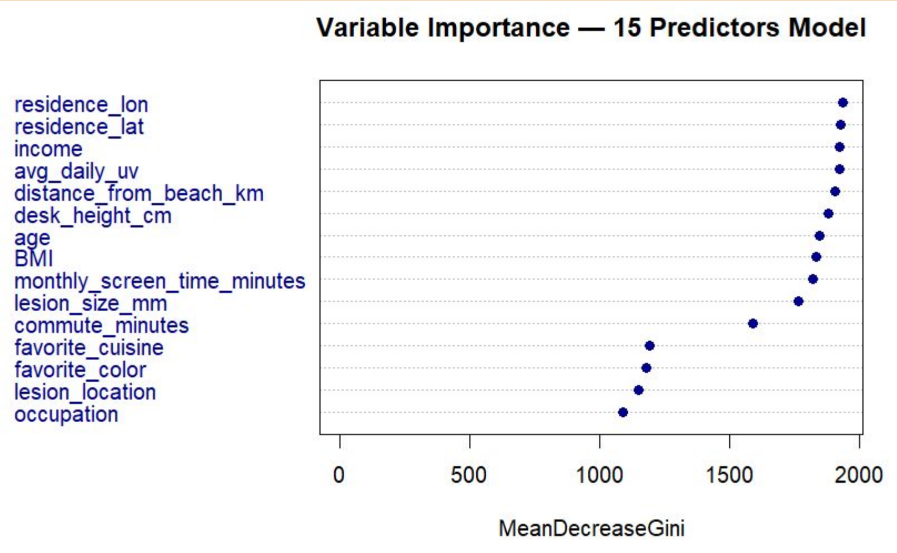
We redid the random forest model with the top 15 important predictors. We found that this model performed worse than the full model for both cancer types. Though as the trees increase Benign error continues to increase and Malignant decreases but less than the full model.

The 15 predictor model now has a Benign error of 55% and a Malignant error of 35% so we saw that the reduced model is weaker in classifying both classes.

On Kaggle: Testing accuracy was 0.56005

Random Forest 15 Predictors Model





The importance plots shows us that for prediction accuracy a patient’s age, UV exposure, and occupation seem to be most important and then several others have similar levels of impotence. We find that the predictors that organize the data the best are not the same that predict the best outcome.

GBM (Boosted Trees)

We fit a boosted tree model using gbm w/ a Bernoulli loss. GBM sequentially built shallow trees, each one correcting the errors of the previous ones.

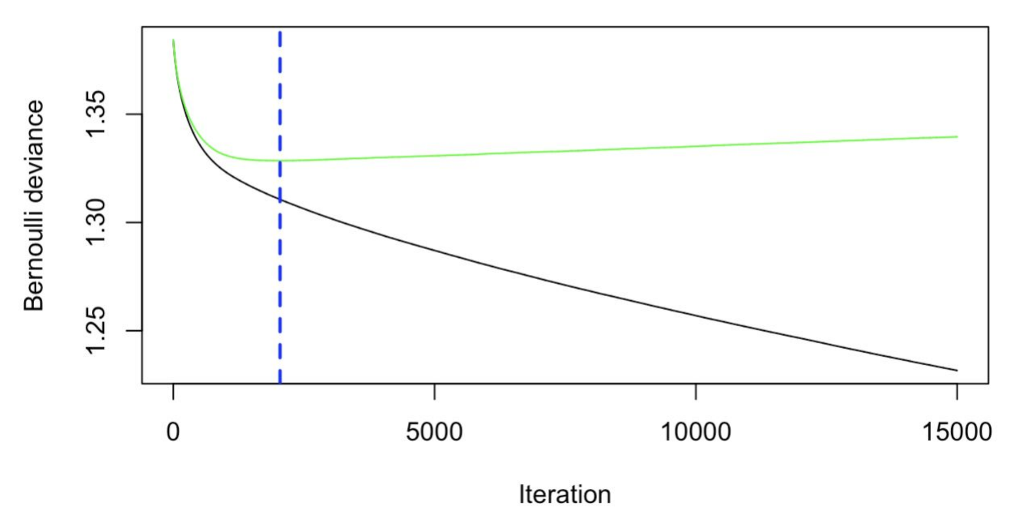
We tuned these hyperparameters:

- **# of trees** - bigger B for more trees -> higher accuracy
- **Interaction Depth** - small trees
- **Shrinkage** - learning rate, 0.1-0.2
- **Minimum observations per node** - 10
- **Subsample** - 1.0
- **10-fold CV** to select optimal # of trees (also tried 5 CV)

GBM

- Green = CV error
- Black = training error
- Blue dotted line = optimal # of trees selected by CV ~ 2042
- Training error keeps decreasing -> gbm flexible
- CV error decr. Then rises -> overfitting
- Need many trees initially
- Min is around 2042 trees, model was refit using this to prevent overfitting

GBM Cross-validated Bernoulli Deviance vs. Number of Trees



GBM

Hyperparameters used for best model:

- **# of trees** - 8,000
- **Interaction Depth** - 5
- **Shrinkage** - 0.005
- **Minimum observations per node: 10**
- **Subsample** - 1.0
- **10-fold-CV** to select optimal # of trees

Different random splits and tuning changed results, with our team's best gbm performance: 60.415% accuracy score on Kaggle

Observed that model struggled with some predictor variables offering a weak signal and high variance in small trees even with bagging, long training time, marginal gains from additional tuning

Final Model

Final Model was a Logistic Regression Model

Explored several approaches including LASSO, Elastic Net, stepwise, subset selection, and multiple imputation methods (MissForest, MICE, median–mode).

Median–mode imputation showed the most stable performance across models.

Final model used a refined set of statistically meaningful predictors (based on p-value significance).

18 Predictors Including: Age, UV exposure, lesion characteristics, family history, photosensitivity, and protective behaviors.

Training Confusion Matrix		
	Benign	**Malignant**
Benign	11684	12184
Malignant	8656	17476

Training Accuracy: **60.76%**

Kaggle Accuracy: **60.545%**

05

Results/Limitations

Analysis

Model

Logistic Regression

classifies based on probability of a categorical response using predictors

Simplicity

Moderate

18 predictors on a size 50,000 dataset

Rank

#14

with accuracy of 0.60545

Analysis

Logistic regression model’s significant predictors for malignancy:

- **Demographics/Environment:** Age, Urban_Rural, Years_Lived_At_Address, Skin_Tone
- **Behavioral and Protective:** Hat_Use, Tanning_Bed_Use, Cloting_Protection, Sun_Burns_Last_Year
- **Medical Risk Factors:** Family_History, Immunosuppressed, Skin_Phtotosensitivity
- **Lesion-specific:** Lesion_Location, Number_Of_Lesions
- **Misc:** Music_Genre

Model performance: improved upon native benchmark by ~8%. (Baseline accuracy is 52.26%)

Discussion

Real World Application:

- Not reliable enough for clinical diagnosis
- Dermatology uses tools that are vastly more accurate (more than 90%)
- With a 60% accuracy, there would be many false negatives, which can lead to delayed treatment
- In conclusion, skin cancer prediction is highly complex and the data we currently have does not serve as an effective prediction method. Richer clinical features, image screening, and medically informed input is needed for accurate prediction.

Limitations

- Our best performing model on kaggle used data that replaced NA values with the median and mode, which may oversimplify the data and increase bias
- GBM and Random Forest models can be computationally expensive, especially with large data sets like ours
- All models risk overfitting if not tuned properly
- Random forest and GBM are harder to interpret in comparison to logistic regression
- Our best performing model was a logistic regression that struggles to capture complex, non-linear relationships and is impacted by multicollinearity
- All model results depend of train and test splits or random seeds, which affect reproducibility
- Some seemingly important variables were labeled as insignificant in our models, implying we may be overlooking certain effects / not capturing important information

Conclusion

- Our best model has an accuracy around 60.545%
- Our best model was a logistic regression using only significant variables, which is beneficial in its simplicity and interpretability
- Models like Random Forest and GBM performed similarly (slightly worse), but were more complex and harder to interpret and much more computationally expensive
- We have moderate accuracy, but as previously mentioned our data has potential bias from our data cleaning and our best model may not capture all complex relationships present in the data
- Overall, our best model provides a solid baseline for future research and model tuning.