Predicting Skin Cancer Status

Sarah Dias; Elizabeth Jiang; Melody Mao; Diandian Shi

(Lecture 1)

Introduction: The Skin Cancer Problem

Skin cancer occurs when UV-induced DNA damage causes skin cells to grow uncontrollably, forming tumors. It most often appears on sun-exposed areas and includes three main types: **Basal Cell Carcinoma**, **Squamous Cell Carcinoma**, and **Melanoma**, the most dangerous form.

The goal of this dataset is to **analyze predictors related to skin cancer risk** and to <u>classify</u> **each individual's cancer status as either** benign **or** malignant using 49 predictors across training and testing sets.

Table of contents

01

Data Cleaning & Preprocessing

The Skin Cancer Dataset

Training dataset:

50000 observations

49 predictors (17 numerical, 32 categorical)

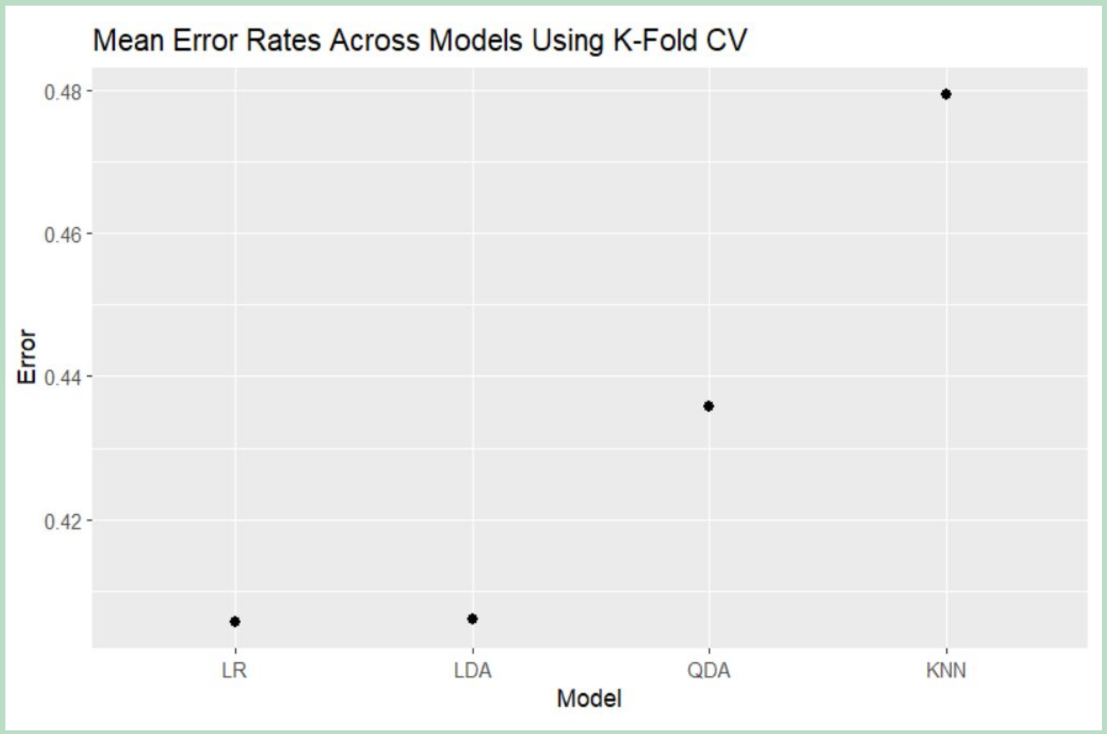
Note: manually adjusted "sunscreen_spf", "outdoor_job", and "zip_code_last_digit" to be factors as they are read in as numeric

Testing dataset:

20000 observations

CV Benchmarking for Imputation

- Ran a 5-fold CV for some basic models to get generalized idea about model performance
- LR and LDA perform best, suggesting a linear boundary in classification
- Used full LR model as benchmarking for subsequent imputation



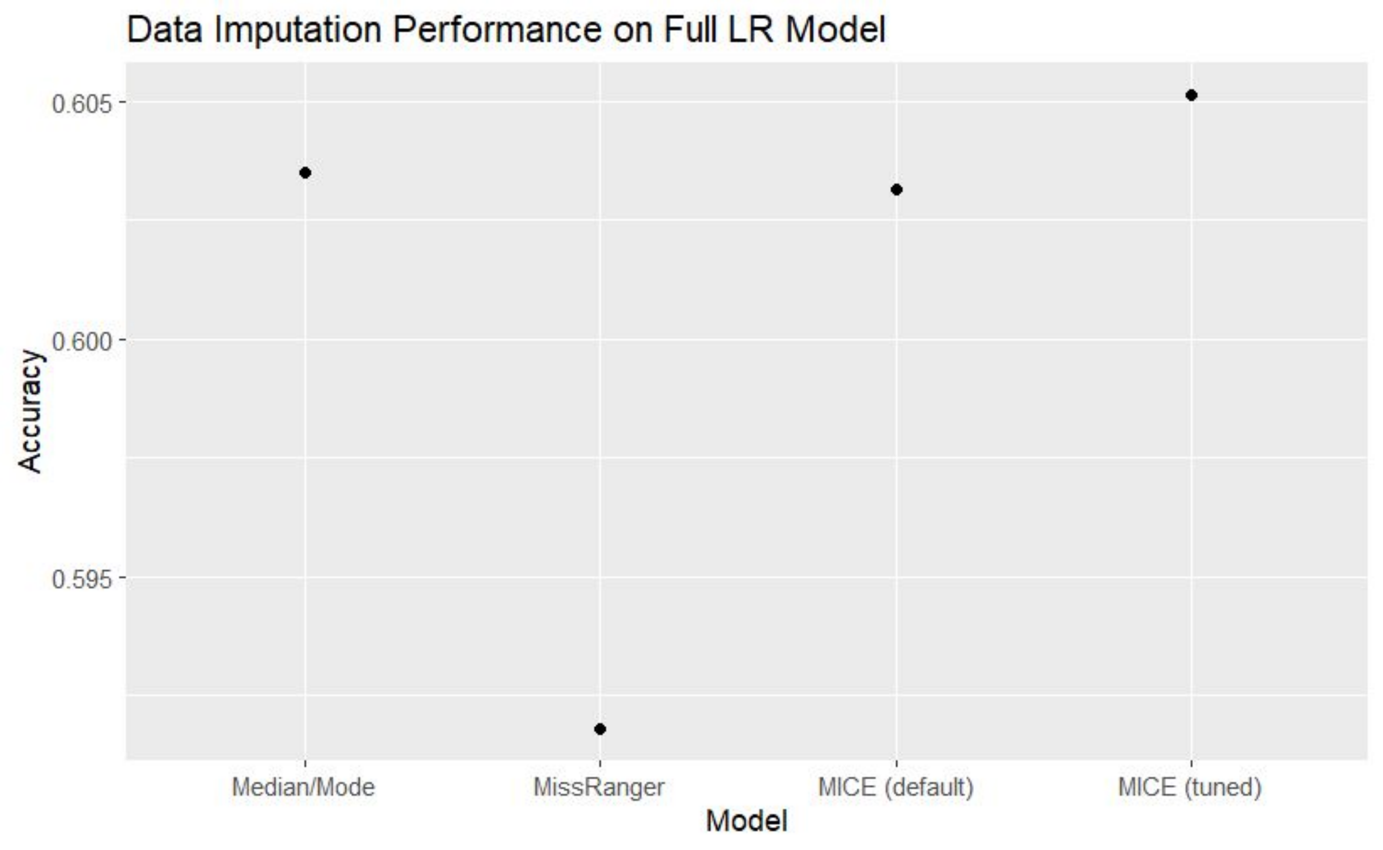
Missing Value Imputation

What we tried...

Median/Mode → MissRanger → MissForest → MICE

What worked:

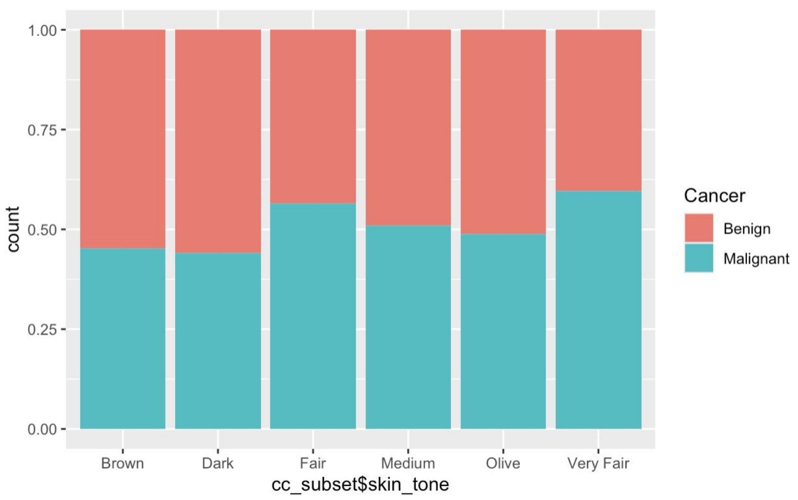
- MICE: iterative approach where each predictor is treated as response and modeling using the other variables as predictors
  - *Numerical* - pmm, *Binary* - logistic regression, *Unordered Categorical* - polytomous regression
  - Pros: robust and less biased, Cons: computationally expensive
- CV to tune parameters for most optimal imputation

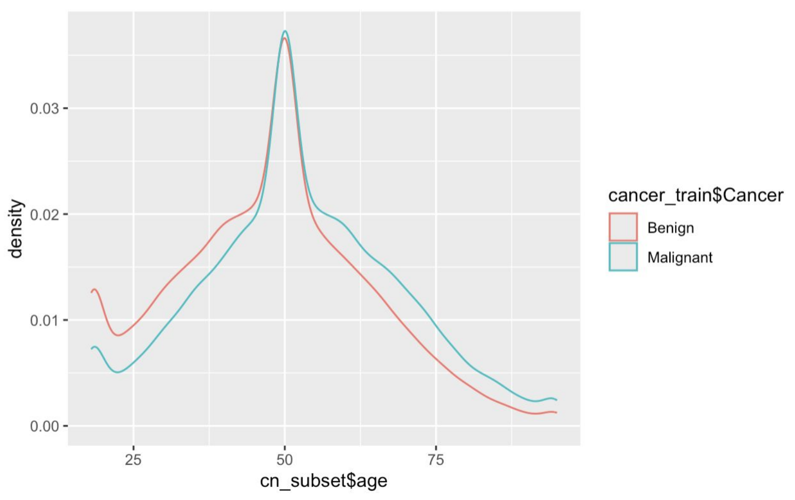

Data Imputation Performance on Full LR Model
0.605
0.600
0.595
Accuracy
Median/Mode
MissRanger
MICE (default)
MICE (tuned)
Model

02

Feature Selection

Method 1: Density Plots/Bar Charts





Pros

- Interpretable
- Allows variable comparison

Cons

- Uninformative
- Not reliable as a standalone

Results

- Kept 18 predictors, worsened accuracy by itself

Method 2: Manual Stepwise

Step 01: High Cardinality

Remove variables with **high number of unique levels** which may lead to overfitting (e.g, favorite color)

Step 03: Near 0 Variance

Predictors with **very little variability that only adds noise and has no predictive power**

Step 02: High VIF

(>5): represents **multicollinearity** (e.g, occupation).

Step 04: Summary

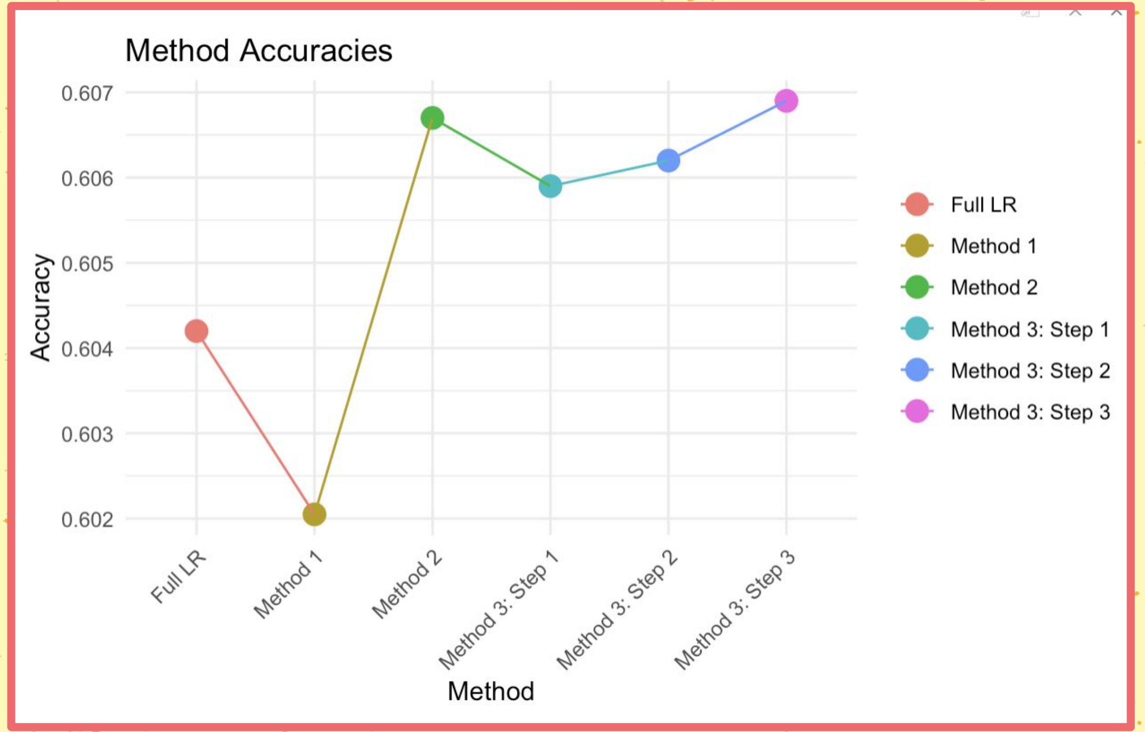
Overall our manual selection removed 11 predictors so far - we think we can improve this even more.

Method 3: Stepwise Regression

Forwards AIC	Backwards BIC	Forwards BIC
20 predictors	**19** predictors	**14** predictors

Result: BIC was too strict, but Forwards AIC produced one of our winning models!

Results


Method Accuracies
0.607
0.606
0.605
0.604
0.603
0.602
Accuracy
Full LR
Method 1
Method 2
Method 3: Step 1
Method 3: Step 2
Method 3: Step 3
Method
Full LR
Method 1
Method 2
Method 3: Step 1
Method 3: Step 2
Method 3: Step 3

03

Model Selection and Fine Tuning

Exploration Roadmap

01
Classical Statistical Models

LR, LDA, QDA, KNN,
naive Bayes

02
"Black Box" Predictive Models

Random Forest,
XGBoost, Catboost

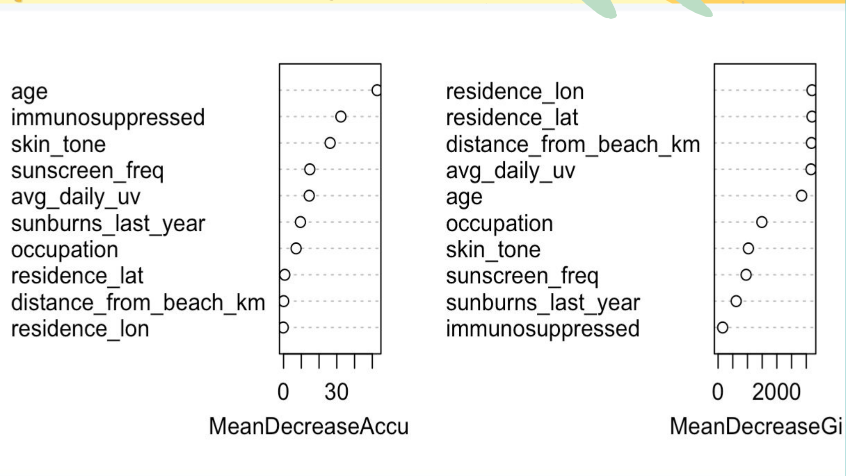
03
Best Model

Logistic Regression
(simpler is better!)

04
Regularization & Fine Tuning

Random Forest

- Fine tuned parameters using CV
  - *mtry*, *nodesize*, *ntree*
  - Accuracy improves with lower mtry and higher nodesize
- <u>Bias towards mode</u> = doesn't work well with unbalanced dataset
  - Slight accuracy improvement with prediction threshold adjustments
- Too computationally expensive to fine tune every parameter
- Conclusion : accuracy struggled to pass 0.6, does not fit well with our imputed dataset



Importance plot: Shows similar result as forward AIC but removing variables with low importance did not significantly improve the accuracy.

Logistic Regression

1

Threshold tuning

Try tuning decision threshold to match training distribution, found out that **0.5** is optimal.

2

Regularization

Elastic Net : median shrinkage between Lasso and Ridge
trainControl() : function used to perform CV to determine best alpha and lambda value

Performed better with our first MICE imputation, but worse with our second...

Logistic Regression (Continued)

3 Potential Interaction terms

Added interaction terms (such as age * avg_daily_uv) to mimic non-linearity but did not improve score.

4 Combination models

Tried taking the average of probabilities predicted by **glm and glmnet(ridge)** , but the predictions are not as good as using only glm (scored around 0.60480)

Kaggle Results for each Tuned Model

Best accuracy for each model; Sorted from lowest to highest performing

XGboost	RF	Catboost	LDA	Ensemble Glm & Glmnet	LR
0.59480 Median/Mode Max_depth = 5 Full model	0.59655 MICE Mtry = 2 Full model	0.59915 MICE 10-fold CV 15 var, based on feature imp	0.60400 Median/Mode 10-fold CV Full model	0.60470 MICE Scaled Averaged glm and glmnet prop	0.60690 MICE Unscaled Reduced model

Takeaway: Each model (when tuned) performed similarly, with no clear hero model except for LR

04. Final Results

Our Two Best Models: LR

Manual Selection

- **39 predictors**
  - More complex, greater variance
- **Accuracy** : 0.60690

AIC Selected

- **20 predictors**
  - Less complex, greater bias
- **Accuracy** : 0.60670

In terms of a better model, AIC wins in regards to simplicity. However, Kaggle only cares about performance...

Best Performing Model:

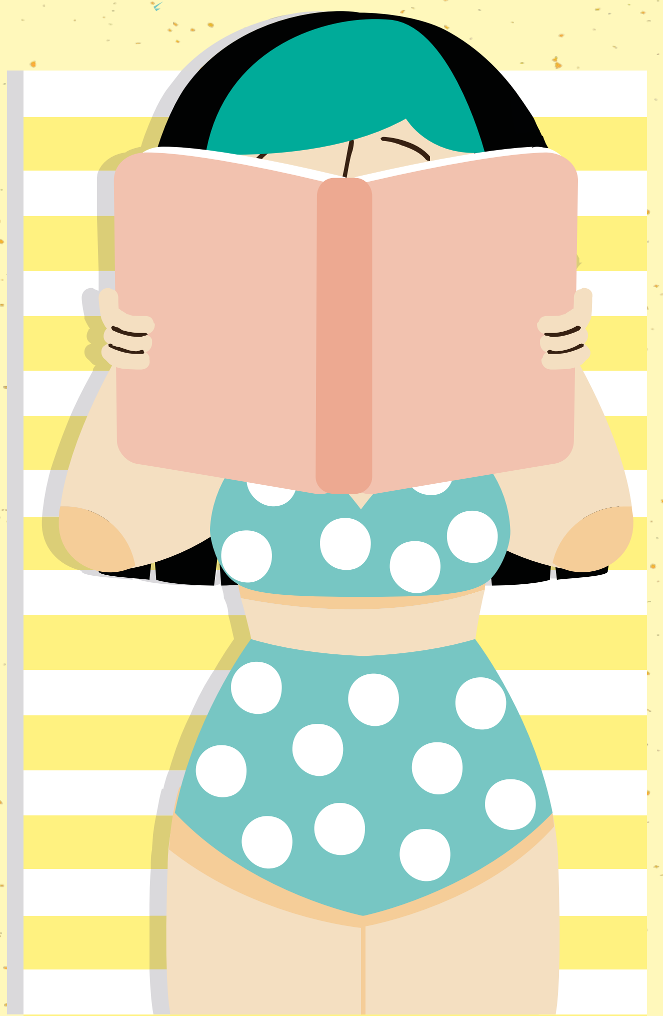
LR - 39P

Best Accuracy:

0.60690

Current Rank:

Top 5



Discussion

Takeaways :

- Simpler is better

Limitations :

- **Too many predictors =>** kept predictors that don't add a lot
- **Computationally expensive =>** MICE imputation method
- **Training accuracy not reflective of testing accuracy =>** increase/decrease in training doesn't translate to testing

Looking Forward :

- **Variable Selection :** Want to achieve similar results with reduced dimensions
- **More fine tuning and cross validation:** Are we actually using the best parameters for the models that we tried?
- **Boosting:** may be advantageous over random forest, which we could explore further!

Thanks!

Does anyone have any questions?

Acknowledgements:

- Professor Almohalwas' Lecture Slides
- Example Presentation by Jason Kwan and Kyle Leung