Stats 101C Final Project

Predictive Analysis of Skin Cancer Diagnostic

Group Members: Elizabeth Jiang, Melody Mao, Diandian Shi, Sarah Dias

Abstract

The goal of this Kaggle project was to predict an individual's skin cancer diagnosis using statistical learning models. This report details the building of our classification model, including data preprocessing and missing value imputation, feature selection, model selection, and fine-tuning. The final model is based on logistic regression and achieved a Kaggle score of 0.60725 with a final rank tied for 5th place.

I. Introduction

Skin cancer is the most commonly diagnosed cancer worldwide, accounting for approximately 30% of all cancer cases each year. Of the different subtypes, melanoma is the most aggressive and life-threatening, and has over 300,000 diagnoses each year. While early detection of skin cancer can lead to survival rates exceeding 90%, late-stage diagnosis significantly reduces treatment effectiveness and increases mortality risk. Current diagnostic techniques include visual skin examinations, dermoscopy, and histopathological biopsy. Although biopsies are highly accurate, they are invasive, time-consuming, and costly, while visual assessment and dermoscopic analysis are subject to variability in clinician expertise. Consequently, developing accurate and efficient predictive models for early skin-cancer detection is both significant and urgent.

In this Kaggle project, we used a skin-cancer diagnostic dataset consisting of demographic predictors, environmental and exposure predictors, sun protection and skin care predictors, biological and health predictors, behavioral predictors, and lifestyle and miscellaneous predictors. This gives a total of 49 predictors, with 17 numerical and 32 categorical, and the response variable being skin cancer type. The training dataset contains 50,000 observations labeled as benign or malignant, while the testing dataset includes 20,000

unlabeled cases for prediction. Our objective is to develop a classification model that accurately predicts the diagnosis of skin lesions by identifying the most informative variables, thereby improving early detection and supporting non-invasive diagnostic decision-making.

II. Data Preprocessing and Missing Value Imputation

Handling Numerical Categorical Variables

There are three number-based categorical variables in the dataset: “sunscreen_spf”, “outdoor_job”, and “zip_code_last_digit” that are read into R as numerical variables. To obtain imputations that accurately represented the set of possible values that these variables could have, these were manually adjusted to be factors before imputation. However, as "sunscreen_spf" and “zip_code_last_digit” have numerical values as variables, they were converted back to numerical variables when used in the downstream classification models. This reconversion provided for better predictive accuracy.

Initial Exploration

To obtain a general idea of the potential class division in predicting skin cancer and assess basic model performance, a 5-fold cross-validation was run using the full set of predictors on a few basic models: logistic regression (LR), linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), and K-nearest neighbors (KNN).

It can be seen that LR and LDA have the lowest error rates (Figure 1), suggesting a linear boundary in the separation of the two classes in our response variable. Because LDA is only more stable than LR when the sample size is small and the distribution is normal, the LR model with the full set of predictors was used as a benchmark for testing subsequent data imputation method approaches.

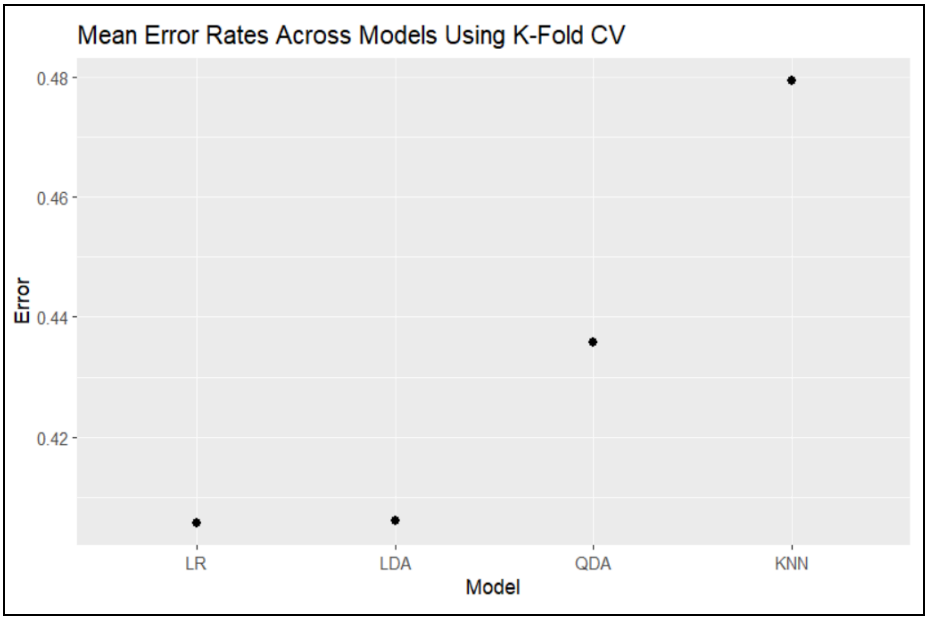

Figure 1: Mean error rates across different models using 5-fold CV

Missing Value Imputation

Both training and testing datasets had approximately 8% missingness distributed evenly across each predictor variable. As the missing values were artificially inserted and there was no evidence for the removal of any predictors, it was determined that all variables should be kept and imputed first.

As it was determined that the quality of data imputations played a significant role in the predictive accuracy of subsequent models, a considerable amount of time was spent testing and fine-tuning a number of data imputation strategies to determine the best one for the dataset. The simplest data imputation method was median and mode imputation, where numerical variables were imputed with their median value and categorical variables were imputed with the most frequent class. This method was surprisingly robust due to a lack of bias between training and testing data imputations, and was also much less computationally expensive compared to the other strategies we tried. The next approach tested was MissForest and MissRanger, a fast alternative to MissForest through chaining multiple random forests to impute the data. Although

more could have been done to fine-tune the parameters of the imputations, these strategies performed surprisingly poorly, being extremely computationally expensive and producing a worse testing accuracy than simple median/mode imputation. This is partially due to the bias of imputations being overtrained on the training dataset, therefore performing poorly on the testing dataset.

The imputation strategy that produced the best testing accuracy was Multiple Imputation by Chained Equations (MICE), an iterative approach where each predictor variable is treated as the response variable and modeled using the other variables as predictors. Predictive mean matching (PMM) is used for numerical variables, logistic regression is used for binary variables, and polytomous regression is used for unordered categorical variables. Through cross-validation, it was determined that the most optimal imputation parameters under a reasonable computation time were 7 imputations with 3 iterations, providing stability by having 7 separately imputed datasets, and preventing overtraining and overfitting by keeping the iteration number for each dataset at 3. The imputed MICE model obtained from training data was then used to impute the missing values in the testing data, producing 7 imputations of the testing dataset as well. Classifications were made by generating probability predictions for each testing dataset, computing the average probability, and making the final classification based on this average. Comparing the tuned model with an initial run with default parameters (5 imputations with 5 iterations), which generated a dataset with a testing accuracy similar to simple mean/mode imputation, the tuned parameters achieved a significantly higher accuracy at a similar computational expense (Figure 2). While this model provided for a robust and less biased method that ultimately produced the highest accuracy among the imputation methods that were tested, the large number of unordered categorical variables in the dataset meant that this strategy

was extremely time-consuming due to the computationally expensive nature of polytomous regression.

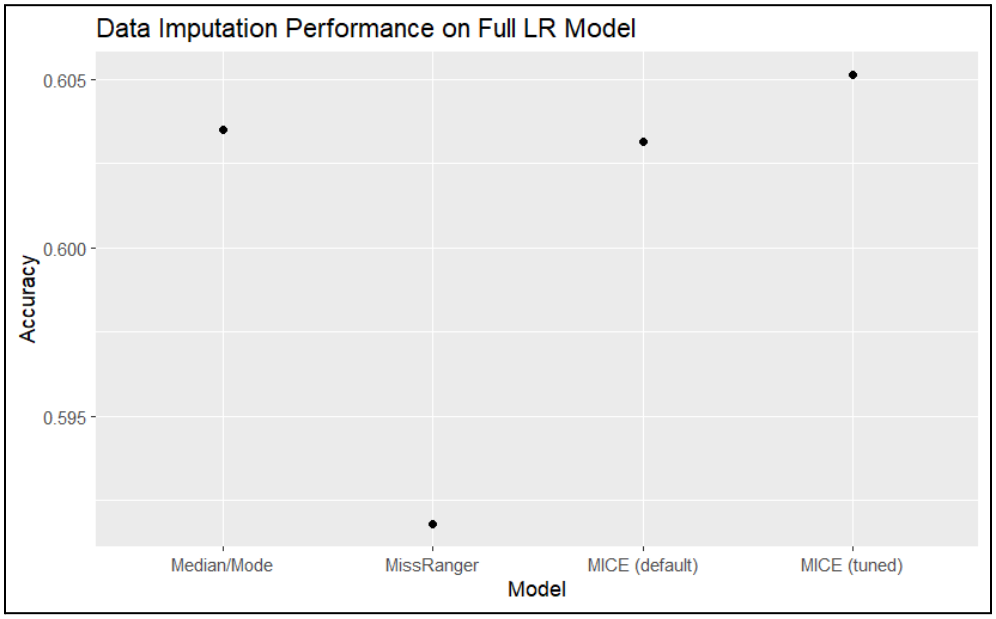

Figure 2: Comparison of testing accuracies between different imputation methods

III. Feature Selection

We evaluated a wide range of feature selection approaches, including but not limited to exploratory data analysis (EDA), hypothesis-based screening (chi-squared tests and two-sample *t*-tests), variable importance, and stepwise selection procedures (manual, automated stepwise, and regsubsets). In the interest of brevity, we focus here on our final "hero" feature selection method, followed by a concise discussion of how insights from alternative methods informed and refined our final predictor set. A logistic model was used as the baseline throughout the process.

EDA using Density Plots and Stacked Barcharts

We began our analysis with some exploratory data analysis, primarily through visualization tools such as density plots for numerical predictors and stacked bar charts for categorical predictors.

For density plots, strong predictors will typically exhibit clear (lateral) distributional separation between response classes. However, all density plots yielded inconclusive results, revealing minimal separation (Figure 3). Thus, it can be hypothesized that no individual numerical predictor had strong standalone predictive power. This finding motivated an investigation of whether or not interaction terms would improve our model; however, incorporating explicit interaction terms did not improve model performance in practice, suggesting that any joint effects were redundant.

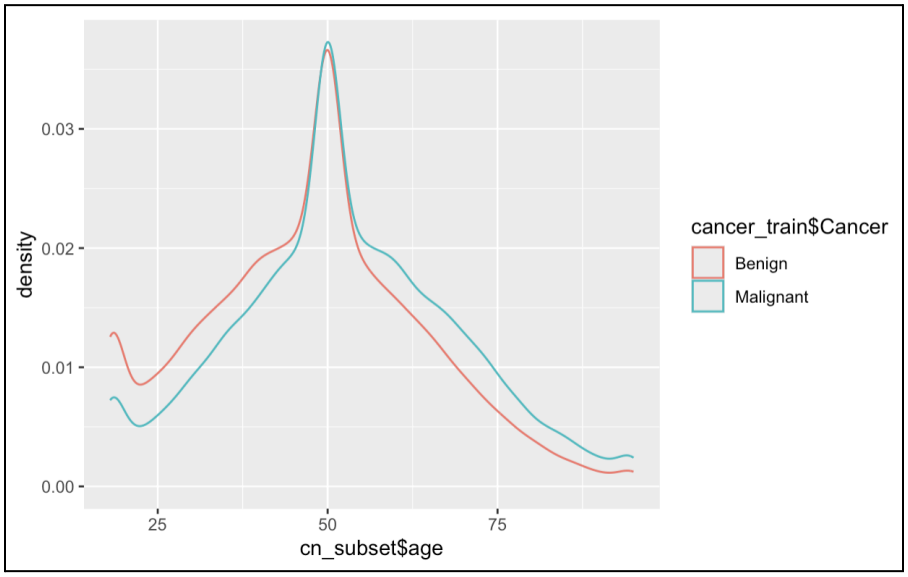

Figure 3: "Best" Density Plot of the age Predictor

For categorical predictors, stacked bar charts were used to visualize differences in response proportions across factor levels (Figure 4). While some predictors showed modest variation across categories, our findings were ultimately similar to our density plots: they were insufficient to identify strong marginal effects on predictive ability using our training data.

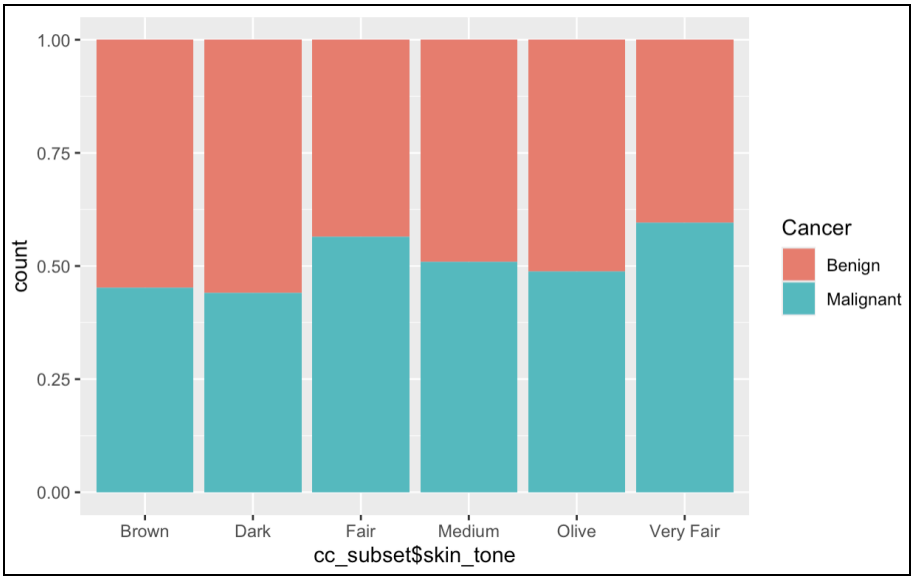

Figure 4: Best Stacked Bar Chart of the skin_tone Predictor

To confirm these observations, we fit a logistic regression model using predictors selected solely through EDA. As expected, this model significantly underperformed relative to our full model, reinforcing the limitations of relying exclusively on visual analysis. Nonetheless, EDA played a useful role in identifying overly complex categorical variables through factor levels.

Hero Method: Forward AIC

After testing a variety of different strategies, forward stepwise regression using AIC was both the most efficient and effective, yielding a performance comparable to our 39 predictor "best" model (0.6067 vs 0.6069) while cutting our predictors to 20. The final predictors include the following: *age, family_history, skin_tone, sunscreen_freq, avg_daily_uv, immunosuppressed, number_of_lesions, sunburns_last_year, outdoor_job, tanning_bed_use, clothing_protection, hat_use, skin_photosensitivity, lesion_size_mm, urban_rural, years_lived_at_address, desk_height_cm, uses_smartwatch, sunscreen_spf,* and *income*. These two models were combined in an ultimate aggregated ensemble discussed in later sections.

Further Adjustments and Discussion

With the results of our forward AIC selection, we sought to improve performance further while balancing parsimony, cross-referencing our results against previous tests.

For our numerical predictors, we compared the variables selected from forward stepwise regression with AIC with results from earlier two-sample *t*-tests and elastic net shrinkage (combining lasso and ridge). We found that nearly all numerically significant predictors identified by these approaches were included in the AIC-selected list, with the notable exception of *sunscreen_spf*. This was a crucial step in our process: adding back *sunscreen_spf* led to an improvement in test accuracy (from 0.60575 to 0.60670), all else constant. On the flip side, we found that removing seemingly "insignificant" predictors such as "desk_height_cm" resulted in a measurable decline in performance, highlighting the importance of consulting multiple methods.

For categorical predictors, we conducted chi-squared tests at the 5% significance level and compared results with variable importance rankings obtained from Random Forest. While the majority of forward AIC–selected categorical predictors aligned with these diagnostics, we removed *participates_in_outdoor_sports*. Although statistically significant in the chi-squared analysis, its inclusion did not improve predictive performance. We hypothesize that this variable may capture information overlapping with more informative numerical predictors such as *sunscreen_spf* or *avg_daily_uv*, highlighting a limitation of univariate categorical tests that do not account for relationships with numerical covariates. Additionally, a VIF test revealed that *outdoor_job* and *occupation* had high VIF scores (>5), indicating multicollinearity; therefore, the more complex *occupation* predictor was removed from our final set.

IV. Model Selection and Fine-Tuning

In exploring different imputation and variable selection strategies, it was important to determine the model that best captured the characteristics of the dataset. A broad range of

classification models was evaluated to identify the best-performing approach, including both classical statistical models such as logistic regression (LR), K-nearest neighbors (KNN), linear and quadratic discriminant analysis (LDA and QDA), and Naive Bayes. We also experimented with more complicated, flexible ensemble models such as Random Forest, XGBoost, and CatBoost. All models were evaluated using training accuracy, confusion matrices, and Kaggle leaderboard performance.

Our initial round of exploration indicated that the full logistic regression model produced the highest accuracy score, and LDA consistently outperformed QDA, strongly suggesting that the classification boundary was approximately linear. As a result, the more flexible models, such as KNN, Random Forest and XGBoost, generally had worse predictive power. These models are powerful when there are strong nonlinearities or interactions, but they tend to overfit and fail to generalize their training performance to testing data. Therefore, we picked logistic regression as our main focus due to its strong performance and simplicity. Believing that a well-tuned ensemble method could grant greater power, we also continued tuning a Random Forest model as a benchmark for comparison.

Logistic Regression

While there are no parameters to adjust in the building of a LR model, adjustments can be made to the classification threshold, which defines the probability cutoff used to assign class labels from predicted probabilities. After evaluating a range of thresholds and comparing error metrics, the optimal threshold yielding the best predictions was approximately 0.5 with some negligible fluctuations. Predictions made with the LR model don't tend to overfit and be biased towards one class over the other, so no adjustments for imbalance caused by bias are required.

Furthermore, due to the large number of predictor variables in the full dataset, we experimented with regularization strategies using glmnet() through implementing Lasso and Ridge as well as Elastic Net, which administers penalties at a balance between Lasso and Ridge. It was expected that the feature selection properties of Elastic Net and Lasso, and the general effect size shrinkage of these regularization techniques, would improve the predictive ability of the model. However, while the elastic net model showed slight improvements for one specific set of imputed data, applying regularization did not consistently improve standard LR. As regularization techniques were tested for both the full set of variables and a selected subset, it can be hypothesized that adding additional feature selection through regularization was removing potentially informative predictors.

Random Forest

Although simple logistic regression performed the best at an elementary evaluation, we wanted to fine-tune and test more powerful ensemble-based modeling techniques. As Random Forest is a flexible, nonparametric classification method capable of capturing nonlinear relationships and complex interactions among predictors, we chose to evaluate and fine-tune the model further to benchmark against the performance of the LR model.

An assessment of variable importance shows results similar to the top 6 forward stepwise AIC selected predictors, with “age” remaining the most important predictor in the model. However, the other chosen predictors were quite different.

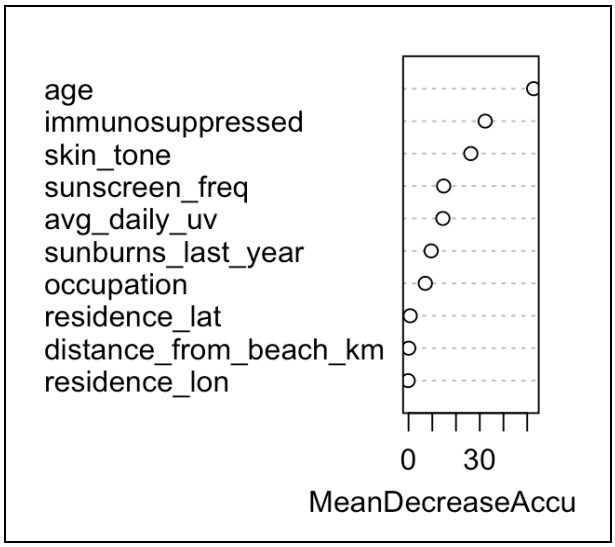

Figure 5: Top 10 predictor variables as ranked by mean decrease in accuracy

Random Forest offers several parameters that can be adjusted and fine-tuned to potentially drastically affect its predictive power. Some of these parameters include the number of variables considered at each split (mtry), the minimum node size (nodesize), and the number of trees (ntrees). Optimal values for these parameters were determined through cross-validation. Although mtry = 7, or the square root of the number of predictors, is typically the optimal parameter value, it was found that lower mtry values led to higher accuracies, with mtry = 4 being the most optimal. A larger nodesize and ntrees also had some marginal positive impact on the model, with the caveat of increasing computational expense. However, tuning overall did not help improve the performance of the Random Forest models, and their testing accuracy rates struggled to pass 0.6, surprisingly low in comparison to the much simpler LR model. The poor performance of Random Forest towards this dataset may be attributed to the class imbalance within our response variable, which tends to lead to biased predictions when using decision trees. Additionally, there were a large number of unordered categorical variables within the dataset that would have biased variable importance. It was not only more computationally expensive and complex, but there was also no reason to utilize this model further.

Overall Comparison of Various Models

The performance of some of the statistical models tested is summarized in Figure 6. It's clear that tree-based models do not not perform as well as the models assuming linear separation in the response variable. Considering all of the models we have tested, the standard logistic regression (LR) remains our best choice.

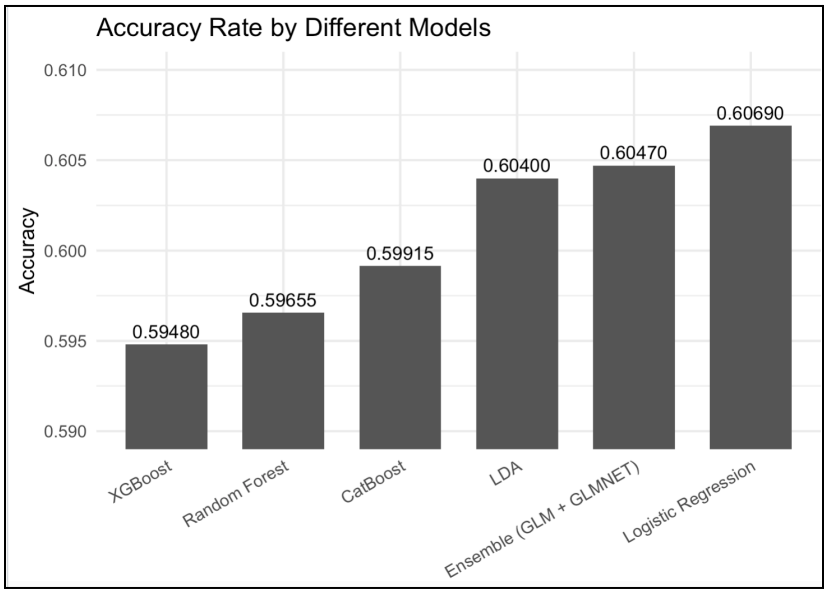

Figure 6: Comparison of accuracy rates across different models

Final Model Aggregation Prediction

Due to the nature of the competition being based entirely on accuracy, a final aggregate method using our best-performing models was used to give us a final accuracy boost. With the hypothesis that if the majority of models predict an individual as a certain class, there is a high probability that that individual is truly in that class, a separate classification submission was made by taking the mode predicted class for each individual across an aggregate of models. The two best-performing LR models were used, as well as a well-performing elastic net model to obtain an odd number of models for a majority vote. This theory proved to be true, as this final

aggregate model pushed our accuracy from 0.60670 and 0.60690 of our two best-performing models to 0.60725, a small but significant improvement.

V. Results and Discussion

Final Results

Our two best models were both logistic regression models with a classification threshold of 0.50; the only difference was the method of variable selection. One model used 20 AIC-selected variables, yielding a Kaggle accuracy score of 0.60670. The other model used 39 manually selected variables and scored 0.60690. Although this model had higher accuracy, as statisticians, we prefer the previous AIC model because of its simplicity and exclusion of noisy variables. This is also better in a real-world context, where getting data on extra variables can be timely and costly.

Limitations

Like every model, our model has limitations. One limitation is that our chosen data imputation method, MICE, is computationally expensive, which may not be a suitable choice when relying on slower technology or when predictions need to be made immediately in a healthcare setting. Another limitation is that our final list of predictors is relatively large (20 predictors is still a relatively complex model), and further feature selection may still be possible. Finally, decreases in training accuracy do not necessarily translate to decreases in testing, which means that this model may not perform as well on a different testing dataset.

Key Takeaways

Trying different imputation methods showed us that imputation has a strong effect on our overall baseline model accuracy, which makes it an important first step that should be experimented with before trying to build the best model. From model selection, we learned that

simpler is often better. Despite the appeal of flexible, complex models in their potential predictive power, logistic regression consistently had the best accuracy. Additionally, it is important to balance quantitative and qualitative reasoning and cross-reference results across multiple tests when performing feature selection. While predictors such as *zip_code_last_digit* were revealed by Lasso to be statistically significant, removing them had no significant impact on our testing results. On the other hand, seemingly irrelevant variables such as *desk_height_cm* were shown to be statistically significant in almost all of our tests, but removing them noticeably worsened our model performance. This suggests that each predictor explains a marginal proportion of variability in the data.

Improvements from Feedback

After watching the presentations of other groups, we determined that our data imputation was robust, but there were additional variable selection and tuning methods our group wanted to try with the hopes of improving our model. This included trying Box-Cox transformations, interaction terms, and engineered variables. Of these different techniques, engineered variables, a technique that the first-place team used, added onto the 20 AIC selected variables, had an accuracy of 0.6066, which was promising. Ultimately, however, any further attempts at improvement did not surpass our aggregate model.

Final Conclusion and Acknowledgements

Overall, this project provided a valuable opportunity for our team to collaborate and apply classification techniques to real-world data. Through this experience, we strengthened our analytical and teamwork skills, and we will carry these lessons forward in our future academic and professional work. We would like to express our sincere gratitude to Professor Almohalwas for his guidance and excellent instruction throughout the quarter.

References

Almohalwas, Akram. (2025). *Predicting Skin Cancer Status* Kaggle https://kaggle.com/competitions/predicting-skin-cancer-status

International Agency for Research on Cancer. (2023). *Skin cancer* World Health Organization https://www.iarc.who.int/cancer-type/skin-cancer/

van Buuren, S. (2018). *Flexible Imputation of Missing Data, Second Edition (2nd ed.)*. Chapman and Hall/CRC. https://doi.org/10.1201/9780429492259