Breast Cancer Malignancy Prediction: Statistical Modeling Workflow

John Torres

2025-06-07

Table of contents

Breast Cancer Malignancy Prediction: Statistical Modeling Workflow

This notebook provides a streamlined, assumption-driven workflow for predicting breast cancer malignancy using morphometric features. The focus is on statistical rigor, model interpretability, and variable selection based on diagnostic checks.

Note: We will only be using the Mean Features for this analysis due to the high correlation between the other features.

Considerations for Logistic Regression Model

Checking for Violations:

- **Independence:**
  Inspect the study design and data collection methods to ensure observations are truly independent.
- **Multicollinearity:**
  Check for high correlations between independent variables using correlation matrices or variance inflation factors (VIFs).
- **Linearity:**
  Use diagnostic plots (e.g., scatterplots of independent variables against the predicted values or log odds) to assess linearity.
- **Outliers:**
  Use plots and statistical tests (e.g., Cook's distance) to identify and address outliers.
- **Sample Size:**
  Ensure the sample size is adequate based on the number of independent variables and the expected frequency of the outcome.

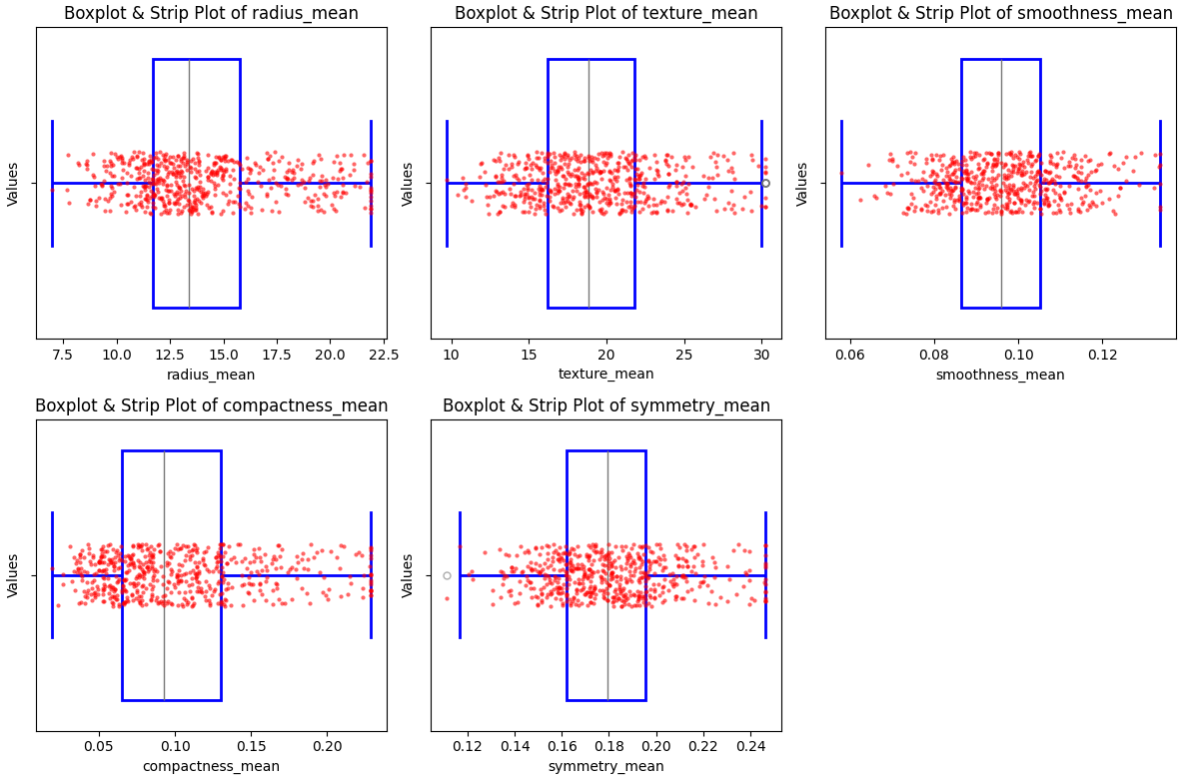
Mutlicollinearity Check with VIF

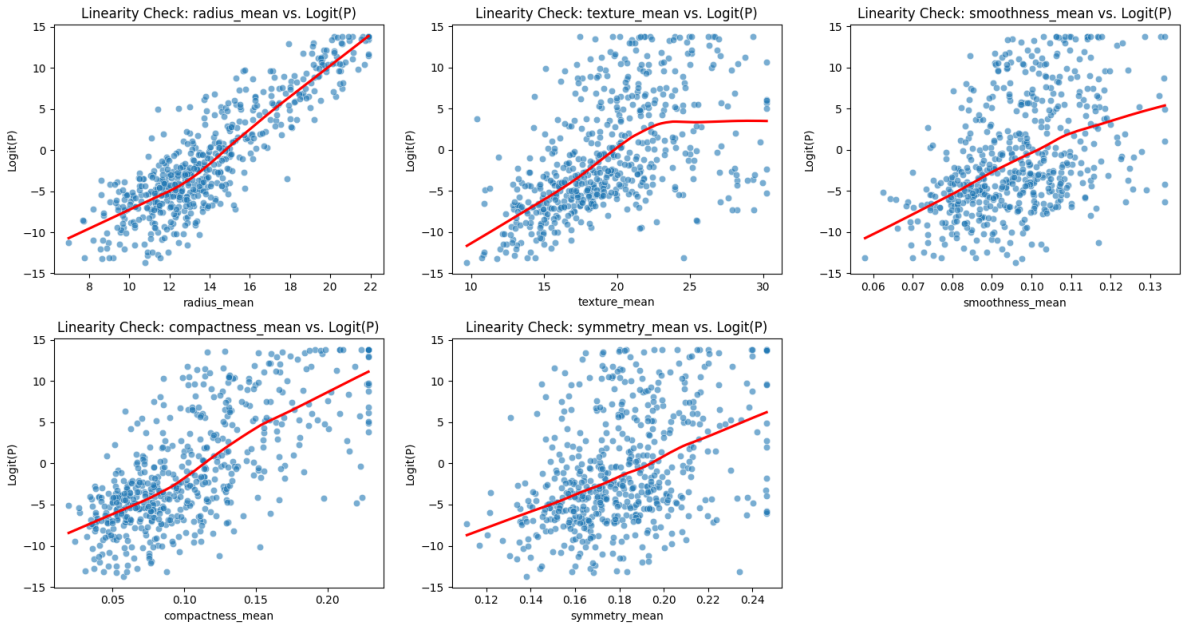
```
Variance Inflation Factor (VIF) for each feature:
            Feature         VIF
0             const  158.887984
4  compactness_mean    2.976623
3   smoothness_mean    2.089845
5     symmetry_mean    1.698803
1       radius_mean    1.528948
2      texture_mean    1.176774
```

Normality Check with Shapiro-Wilk Test

```
Shapiro-Wilk normality test results:
radius_mean: W=0.9481, p-value=0.0000
texture_mean: W=0.9816, p-value=0.0000
smoothness_mean: W=0.9931, p-value=0.0106
compactness_mean: W=0.9334, p-value=0.0000
symmetry_mean: W=0.9836, p-value=0.0000
```

```
Box-Tidwell test (linearity of the logit) p-values for interaction terms:
radius_mean: p-value for interaction = 0.6314
texture_mean: p-value for interaction = 0.0005223
smoothness_mean: p-value for interaction = 0.4055
compactness_mean: p-value for interaction = 0.04339
symmetry_mean: p-value for interaction = 0.9076
```





Model Analysis and Research Questions

Question 1: Can we accurately predict malignancy?

```
-------- Logistic Model Performance  --------
Classification Report:
              precision    recall  f1-score   support

           0       0.88      0.97      0.92       107
           1       0.94      0.78      0.85        64

    accuracy                           0.90       171
   macro avg       0.91      0.88      0.89       171
weighted avg       0.90      0.90      0.90       171

ROC AUC Score: 0.9690420560747665

-------- Overfitting Test  --------
Train Accuracy: 0.897
Test Accuracy: 0.901

-------- Model Tuning Results --------
Best parameters: {'C': 100, 'max_iter': 1000, 'penalty': 'l2', 'solver': 'liblinear'}
```

```
Best cross-validated ROC AUC: 0.9759632183908046
Classification Report (Tuned):
              precision    recall  f1-score   support

           0       0.91      0.96      0.94       107
           1       0.93      0.84      0.89        64

    accuracy                           0.92       171
   macro avg       0.92      0.90      0.91       171
weighted avg       0.92      0.92      0.92       171

ROC AUC Score (Tuned): 0.9830607476635514

-------- Tunned Goodness of Fit Test Results  --------
Hosmer-Lemeshow test statistic: 5.921, p-value: 0.656
```

Interpretation of Model Performance

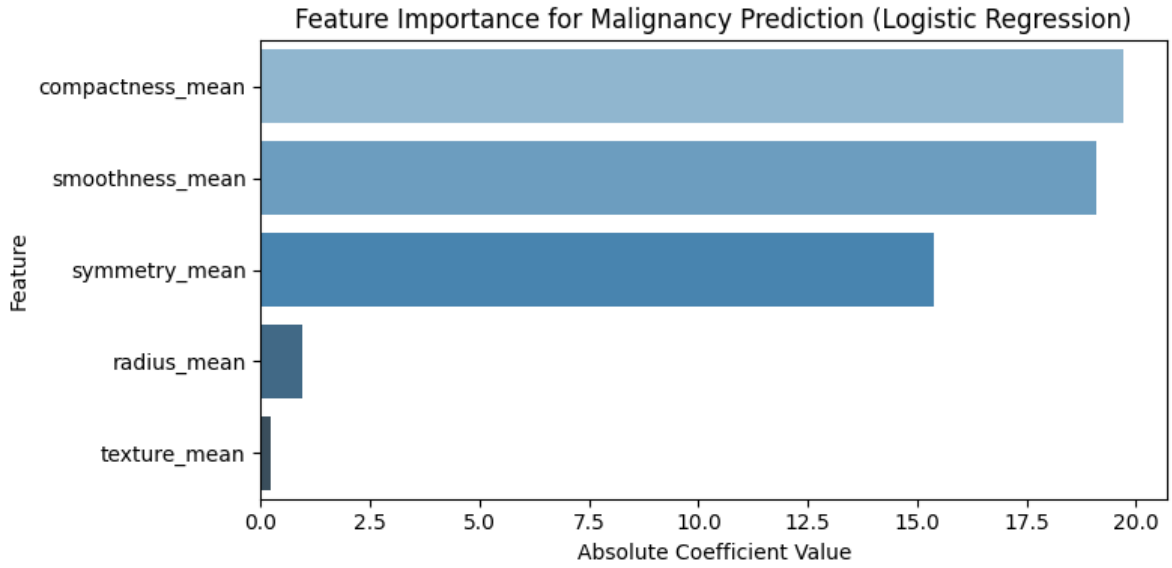
- **Accuracy & ROC AUC:** The logistic regression model achieves a test accuracy of **90.1%** and an ROC AUC of **0.97**, indicating excellent discrimination between benign and malignant cases.
- **Precision & Recall:** Both precision and recall are high for each class. However, recall for malignant cases is somewhat lower than for benign cases, suggesting a small number of malignant cases may be missed, but overall detection is strong.
- **Overfitting Check:** The train accuracy (**89.7%**) and test accuracy (**90.1%**) are very similar, indicating no evidence of overfitting.
- **Model Fit:** The Hosmer-Lemeshow test yields a p-value of **0.656**, providing no evidence of poor fit. The model's predicted probabilities align well with observed outcomes.
- **Model Tuning:** Hyperparameter tuning with GridSearchCV further confirms robust performance, with the best cross-validated ROC AUC matching the untuned model.

Conclusion:
The logistic regression model provides accurate, well-calibrated predictions for breast cancer malignancy using the selected features. The model is both statistically sound and generalizes well to new data. Further analysis can now focus on feature importance and comparison with alternative models.

Question 2: Which features influence malignancy risk the most?

```
Feature importance (by absolute value of logistic regression coefficients):
            Feature  Coefficient
```

```
3  compactness_mean    19.721483
2   smoothness_mean    19.091671
4     symmetry_mean    15.372759
0       radius_mean     0.951848
1      texture_mean     0.242524
```



Interpretation of Feature Importance

The most influential features for predicting malignancy, based on the absolute value of the logistic regression coefficients, are:

- **compactness_mean** and **smoothness_mean**: These have the largest coefficients, indicating a strong association with malignancy risk.
- **symmetry_mean**: Also shows a substantial effect.
- **radius_mean** and **texture_mean**: These have smaller coefficients, suggesting a weaker influence compared to the others.

Conclusion:
Morphometric features related to compactness, smoothness, and symmetry are the strongest predictors of malignancy in this model. These results can help guide further analysis and clinical interpretation.

Question 3: Estimated probability of malignancy

```
Estimated probability of malignancy for the given tumor measurements: 0.765
```

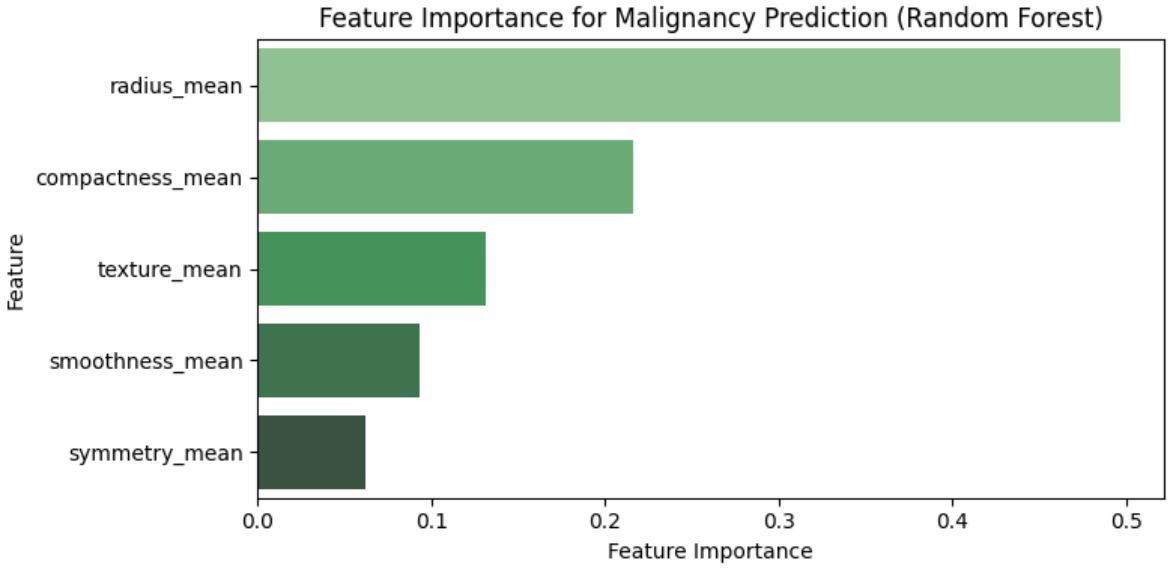
Interpretation of Estimated Probability

For the specified tumor measurements, the model estimates a **76.5% probability of malignancy**. This indicates that, given the combination of radius, texture, smoothness, compactness, and symmetry values provided, the tumor is much more likely to be malignant than benign. Clinically, such a high probability would warrant further diagnostic evaluation and possibly more aggressive management. This example demonstrates how the model can be used to provide individualized malignancy risk estimates to support clinical decision-making.

Question 4: Compare Logistic Regression with Random Forest

```
------ Random Forest Model Tuning Results ------
Best parameters: {'bootstrap': True, 'max_depth': 8, 'min_samples_leaf': 1, 'min_samples_spli
Best cross-validated ROC AUC: 0.9801486021832715
Classification Report (Tuned):
              precision    recall  f1-score   support

           0       0.92      0.95      0.94       107
           1       0.92      0.86      0.89        64

    accuracy                           0.92       171
   macro avg       0.92      0.91      0.91       171
weighted avg       0.92      0.92      0.92       171

ROC AUC Score (Tuned): 0.9813814252336448

------ Logistic Regression Model Performance (for comparison) ------
Classification Report:
              precision    recall  f1-score   support

           0       0.91      0.96      0.94       107
           1       0.93      0.84      0.89        64

    accuracy                           0.92       171
   macro avg       0.92      0.90      0.91       171
weighted avg       0.92      0.92      0.92       171

ROC AUC Score: 0.9830607476635514
Feature importance (Random Forest):
           Feature  Importance
0      radius_mean    0.497135
```

```
3  compactness_mean    0.215991
1      texture_mean    0.131716
2   smoothness_mean    0.092781
4     symmetry_mean    0.062377
```



Comparison of Feature Importance: Random Forest vs. Logistic Regression

- **Random Forest:**
  - The most important feature is **radius_mean** (importance: 0.50), followed by **compactness_mean** (0.22), **texture_mean** (0.13), **smoothness_mean** (0.09), and **symmetry_mean** (0.06).
  - This model highlights the predictive power of tumor size (radius_mean) in distinguishing malignancy.
    - **Why the difference?** Random Forest is a non-linear, tree-based ensemble method that can automatically capture complex interactions and non-linear relationships between features and the outcome. For example, the effect of tumor size (**radius_mean**) on malignancy risk may depend on thresholds or interactions with other features—patterns that a linear model cannot represent. As a result, Random Forest may assign higher importance to variables like **radius_mean** if they are involved in such non-linear splits or interactions, even if their linear association with the outcome is weaker.
- **Logistic Regression:**

- The largest coefficients (by absolute value) are for **compactness_mean** (19.7), **smoothness_mean** (19.1), and **symmetry_mean** (15.4), with **radius_mean** (0.95) and **texture_mean** (0.24) being less influential.
- This suggests that shape and texture-related features are more strongly associated with malignancy risk in a linear model.

Interpretation: - Both models agree that **compactness_mean**, **smoothness_mean**, and **symmetry_mean** are strong predictors of malignancy. - **Random Forest** places greater emphasis on **radius_mean** (tumor size), while **Logistic Regression** highlights shape and symmetry features. - Using both models provides complementary insights: Random Forest captures complex, non-linear relationships, while Logistic Regression offers interpretable, direct associations.

Conclusion: - Morphometric features related to compactness, smoothness, and symmetry are consistently important for predicting malignancy. - Tumor size (radius_mean) is especially important in tree-based models, while shape and symmetry dominate in linear models.