Predicting breast tumor malignancy based on nuclear feature extraction

Alice Yue
Springboard Data Science

Introduction

- Cancer detection and classification
- Importance of effective triage
- 60% of ICU triage decisions
- Errors in judgement
- Mortality rate
- Create more tools for triaging cancer patients

Task

- Classification model
- Fluid sample from the tumor
- Cell nuclei features
- Using data from the Diagnostic Wisconsin Breast Cancer Database (1993)
- Mean, standard error, extreme
- 63% malignant, 37% benign

Data

- Geometric/derived features
- Appeared to be strong correlations
- All positively correlated with malignancy
- Interpreting standard errors (SE features)
- Importance of SE features
- No benefit to model performance

Exploring feature importances

- Permutations done by subdividing training data
- Confirmed weakness of SE features
- Confirmed importance of extreme/worst measurements
- Starting point for feature selection

Preprocessing

- 80/20 split
- Built model on full feature space
- Built 3 classifiers:
  - Logistic regression
  - Stochastic Gradient Descent
  - Random Forest

Model selection

Classifier	Metrics

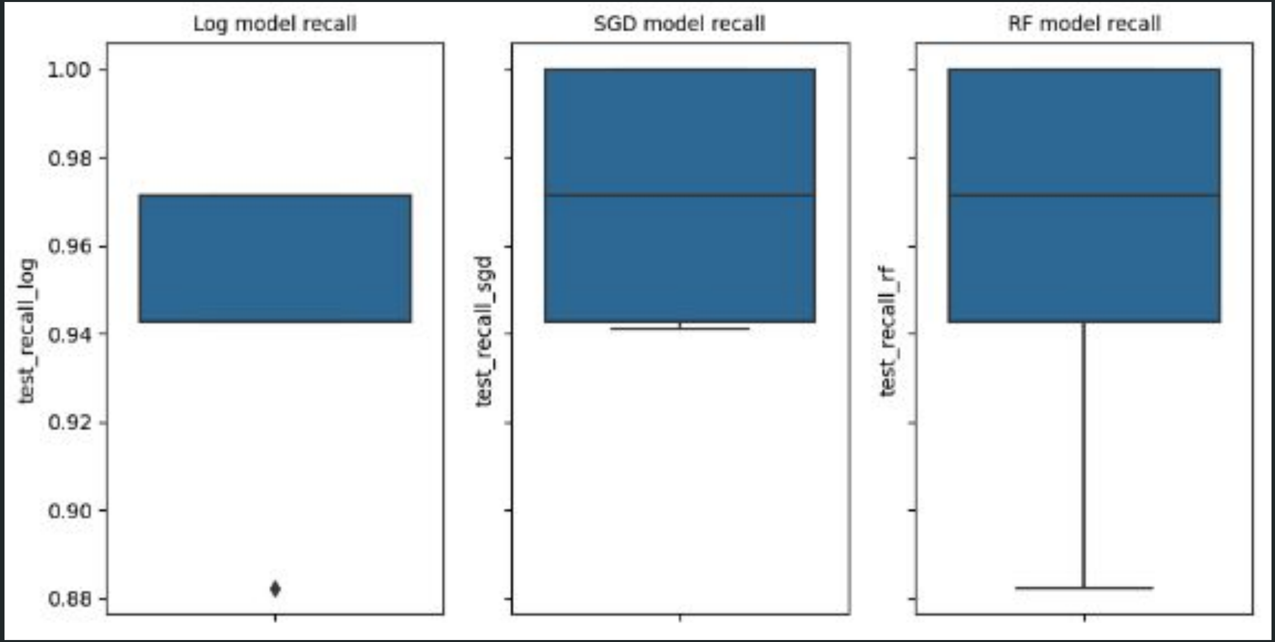
Logistic regression	ROC-AUC = 95.39
	recall = 94.74

SGD (scaled using StandardScaler)	ROC-AUC = 93.42
	recall = 89.47

Random Forest	ROC-AUC = **95.39**
	recall = **94.74**

Cross validation

- Recall scores on 5-fold cross validation



Thresholding

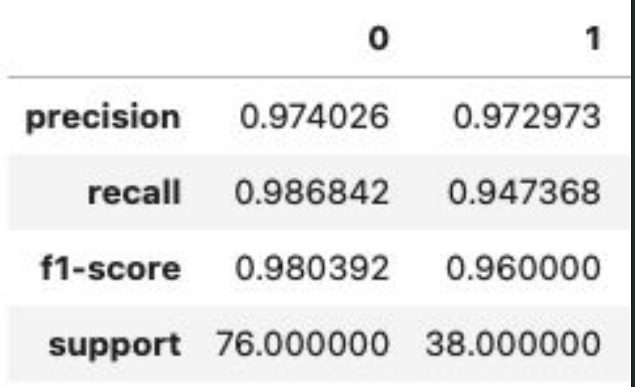
- Precision/recall tradeoff
  - Logistic regression
    - P = 88.1, R = 97.4
  - Random Forest
    - P = 97.3, R = 94.7
- 9.2% loss in precision vs. 2.7% loss in recall

Final model

- Random Forest
- Consider intended use
- Limited resources (time, beds, doctors)
- ICU triage decisions are made in the moment
- Preserve 9% precision

Final model

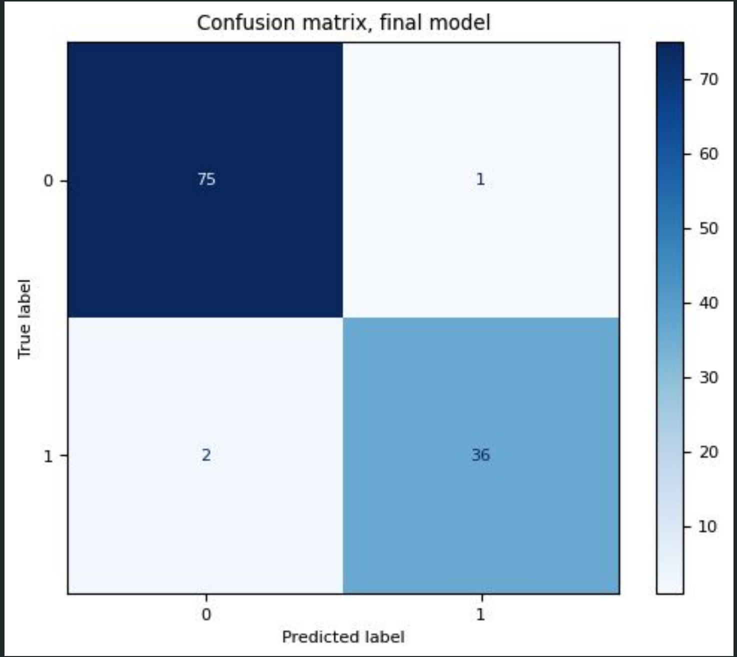
- Tested 5 subsets based on EDA
- Narrowed down to 13 features
- Extreme/worst measurements (8/13)
- Precision = 97.3
- Recall = 94.7



	0	1
precision	0.974026	0.972973
recall	0.986842	0.947368
f1-score	0.980392	0.960000
support	76.000000	38.000000

Final model

- 2 false negatives, 1 false positive (out of 114 cases)



Reflections

- Standard errors
  - Overall
  - Uncorrelated (discretization)
- Precision-recall tradeoff
- Support vector machine
- Stratification

Conclusion

- Tool for triage
  - Accurate diagnosis
  - Effective use of resources
- Larger systems
- Other types of cancer
- Predicting severity (regression)

References

Mangasarian, Street, & Wolberg. Nuclear Feature Extraction For Breast Tumor Diagnosis. International Symposium on Electronic Imaging Science and Technology, Volume 1905, pages 861-870. IS&T/SPIE, 1993, San Jose, California.

Van der Zee, Benoit, et al. Outcome of cancer patients considered for intensive care unit admission in two university hospitals in the Netherlands: the danger of delayed ICU admissions and off-hour triage decisions. National Center for Biotechnology Information, National Library of Medicine, 11 August 2021, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8357904/.