Predicting Breast Cancer Using Machine Learning

Classification Project

Flora Xinru Cheng

October 30, 2019

Motivation

Motivation

Motivation

Source: National Breast Cancer Foundation

Motivation

Motivation

Motivation

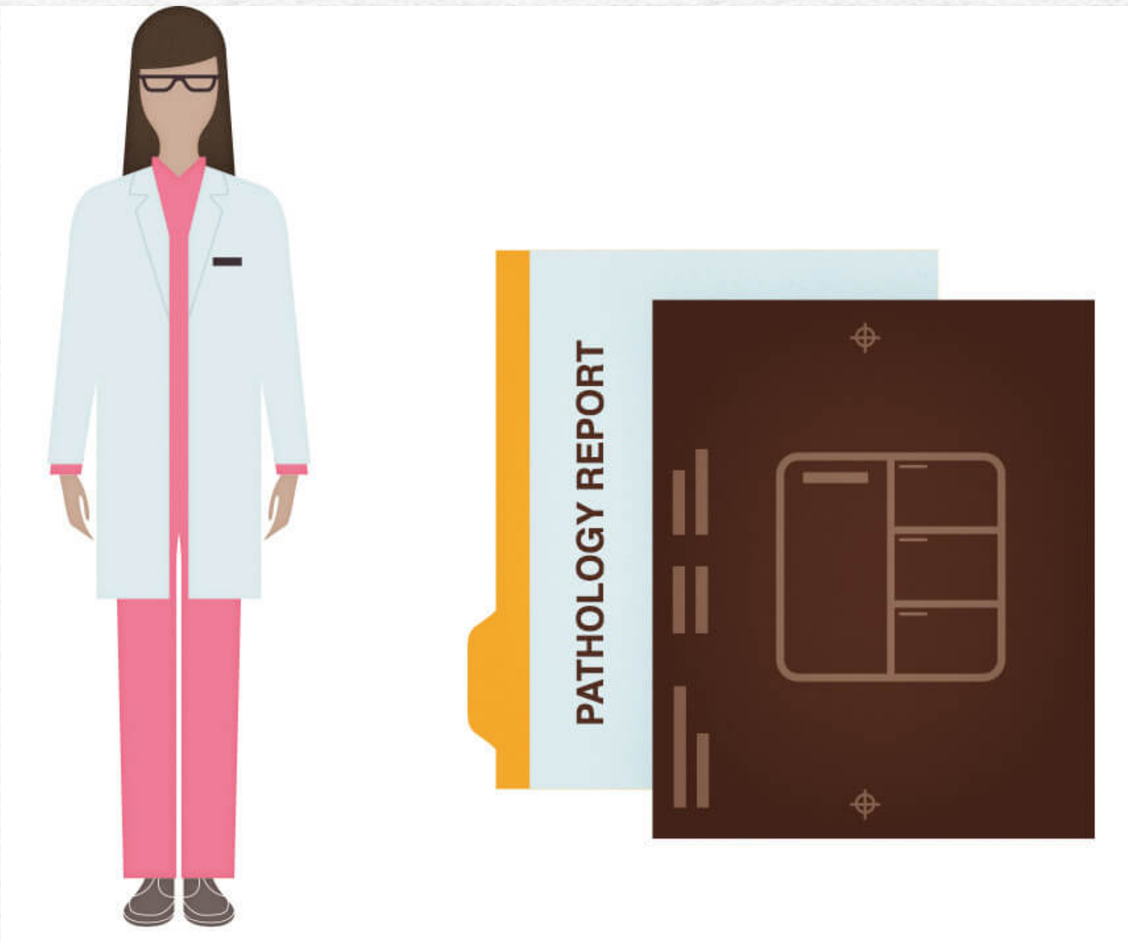
Early detection **survival rate** is **99%**

Motivation

Early detection **survival rate** is **99%**

Objective

Fine-needle aspiration biopsy

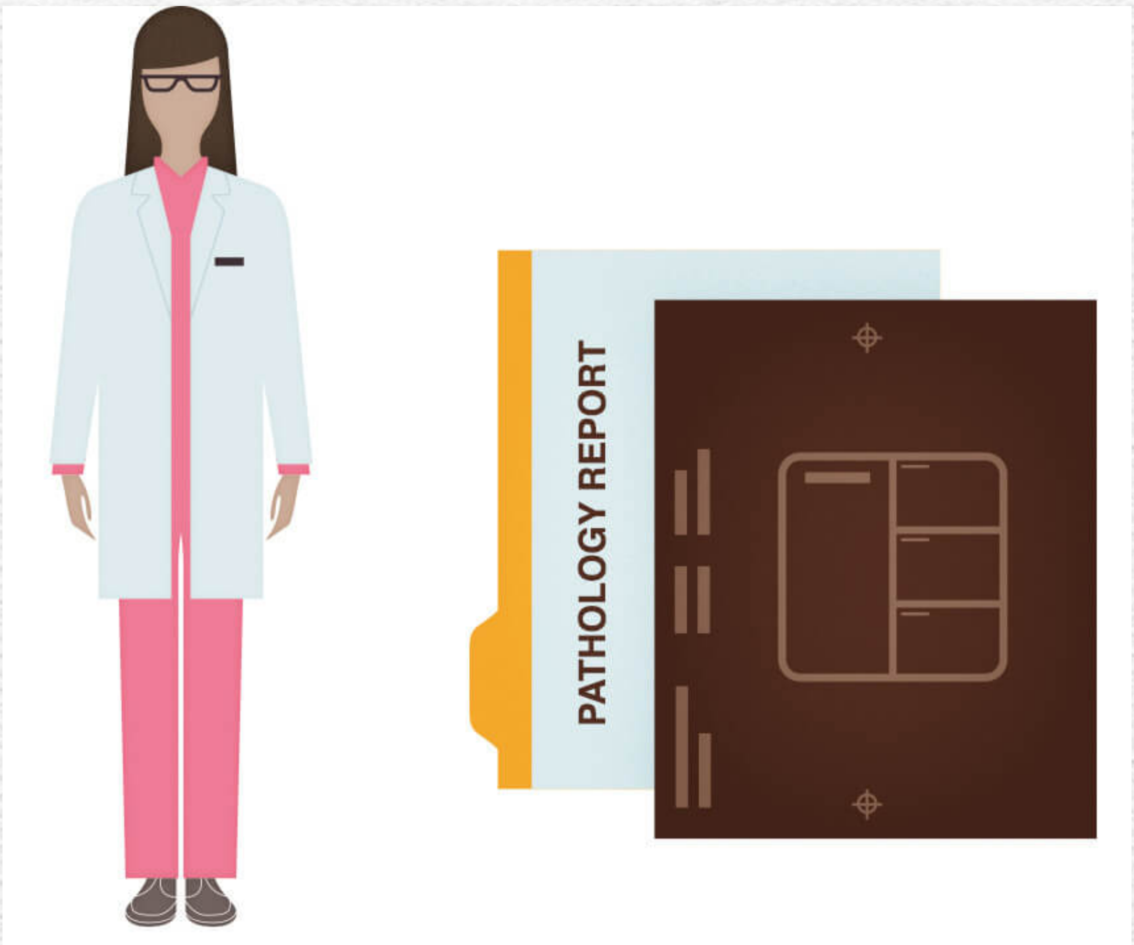


Objective

Fine-needle aspiration biopsy

1~2 weeks

Accuracy ~ 80 %

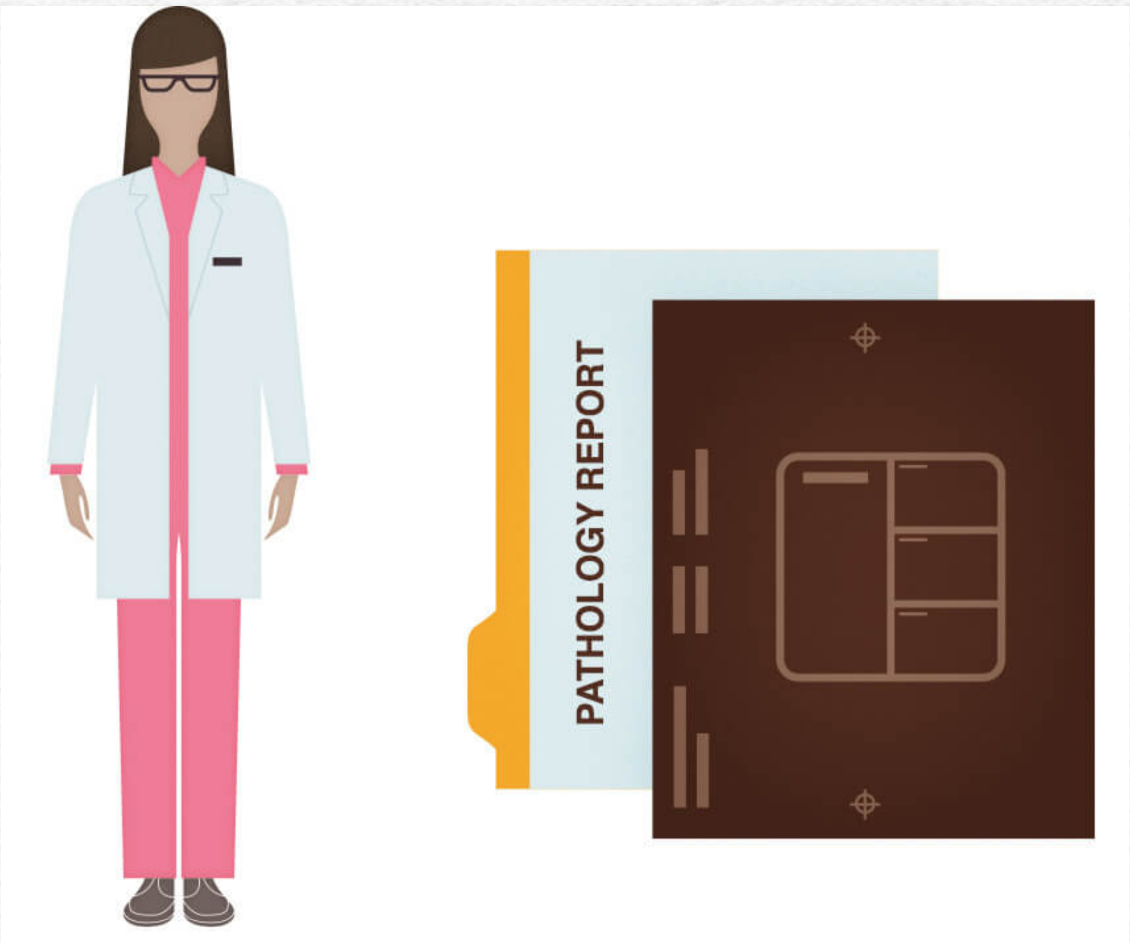


Objective

Fine-needle aspiration biopsy

1~2 weeks

Accuracy ~ 80 %



Improve diagnosis accuracy and speed with machine learning

Methodology

Methodology

Data

Methodology

Data

- Source: UCI, Breast Cancer Wisconsin Diagnostic Data Set

Methodology

Data

- Source: UCI, Breast Cancer Wisconsin Diagnostic Data Set
- 569 observations, 10 calculated features (area, symmetry, concavity…)

Methodology

Data

- Source: UCI, Breast Cancer Wisconsin Diagnostic Data Set
- 569 observations, 10 calculated features (area, symmetry, concavity…)
- Target feature: Diagnosis—Malignant or Benign

Methodology

Data

- Source: UCI, Breast Cancer Wisconsin Diagnostic Data Set
- 569 observations, 10 calculated features (area, symmetry, concavity…)
- Target feature: Diagnosis—Malignant or Benign

Models

Methodology

Data

- Source: UCI, Breast Cancer Wisconsin Diagnostic Data Set
- 569 observations, 10 calculated features (area, symmetry, concavity…)
- Target feature: Diagnosis—Malignant or Benign

Models

- 9 classification models in SciKit-Learn

Methodology

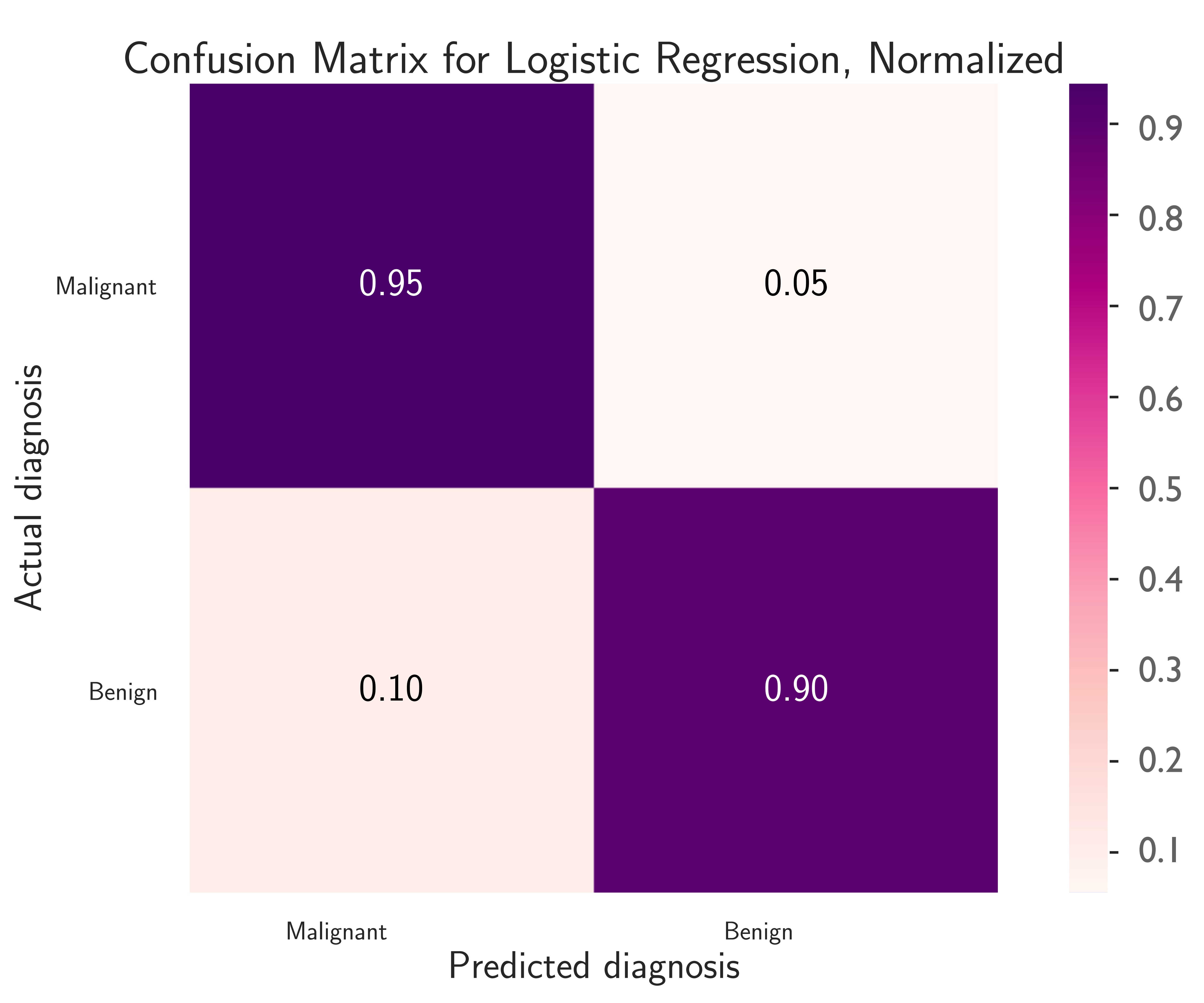
Data

- Source: UCI, Breast Cancer Wisconsin Diagnostic Data Set
- 569 observations, 10 calculated features (area, symmetry, concavity…)
- Target feature: Diagnosis—Malignant or Benign

Models

- 9 classification models in SciKit-Learn
  - Top 4: Logistic Regression, Gradient Boosting, Random Forest, SVM

Results



Confusion Matrix for Logistic Regression, Normalized
Actual diagnosis
Malignant
Benign
0.95
0.05
0.10
0.90
Malignant
Benign
Predicted diagnosis
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1

Conclusion

Conclusion

- Improved diagnosis accuracy and speed with machine learning

Conclusion

- Improved diagnosis accuracy and speed with machine learning
- Very good results (human accuracy ~ 80%)

Conclusion

- Improved diagnosis accuracy and speed with machine learning
- Very good results (human accuracy ~ 80%)
    - 95% prediction accuracy for malignant tumors

Conclusion

- Improved diagnosis accuracy and speed with machine learning
- Very good results (human accuracy ~ 80%)
  - 95% prediction accuracy for malignant tumors

Future Work

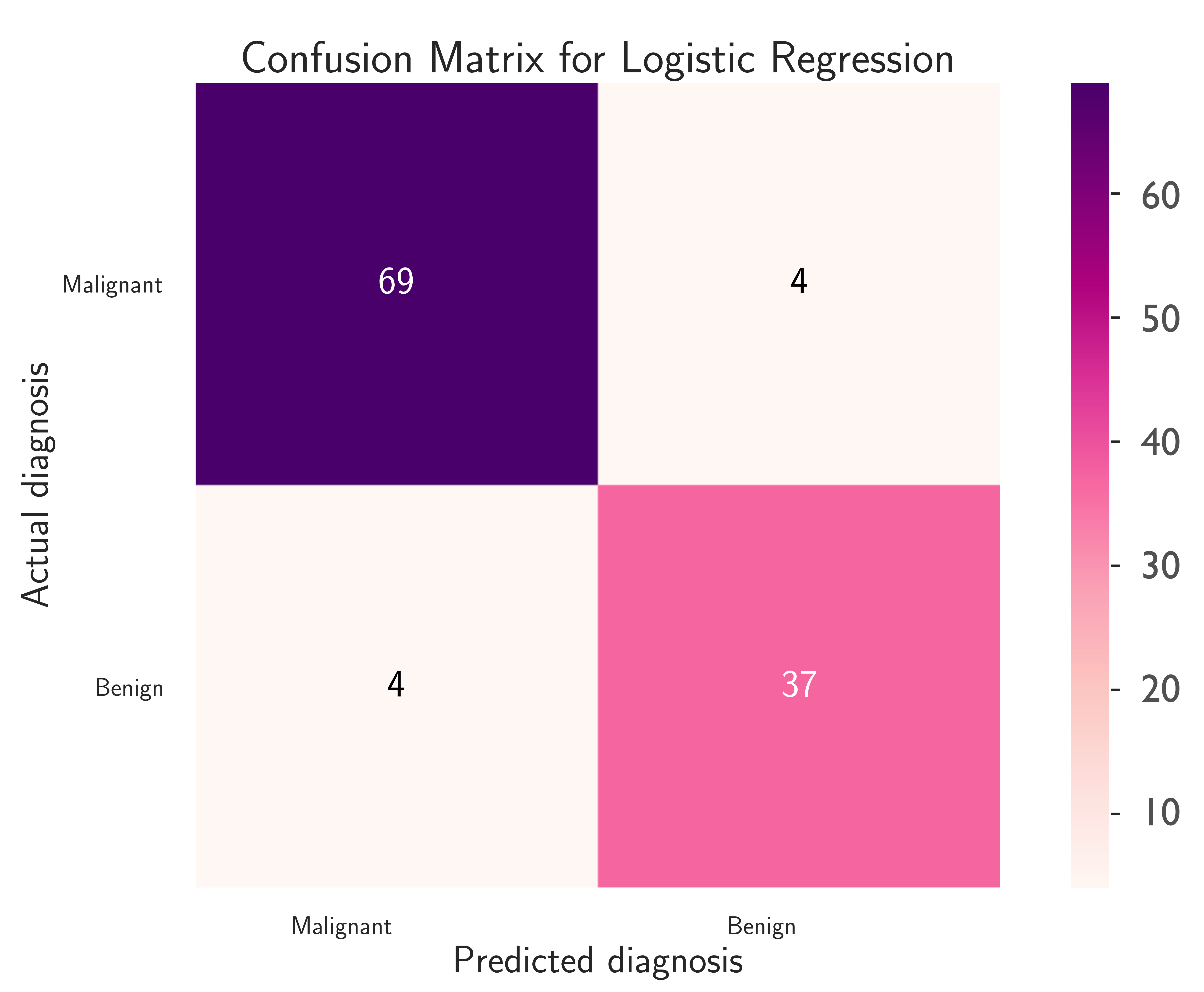
Conclusion

- Improved diagnosis accuracy and speed with machine learning
- Very good results (human accuracy ~ 80%)
  - 95% prediction accuracy for malignant tumors

Future Work

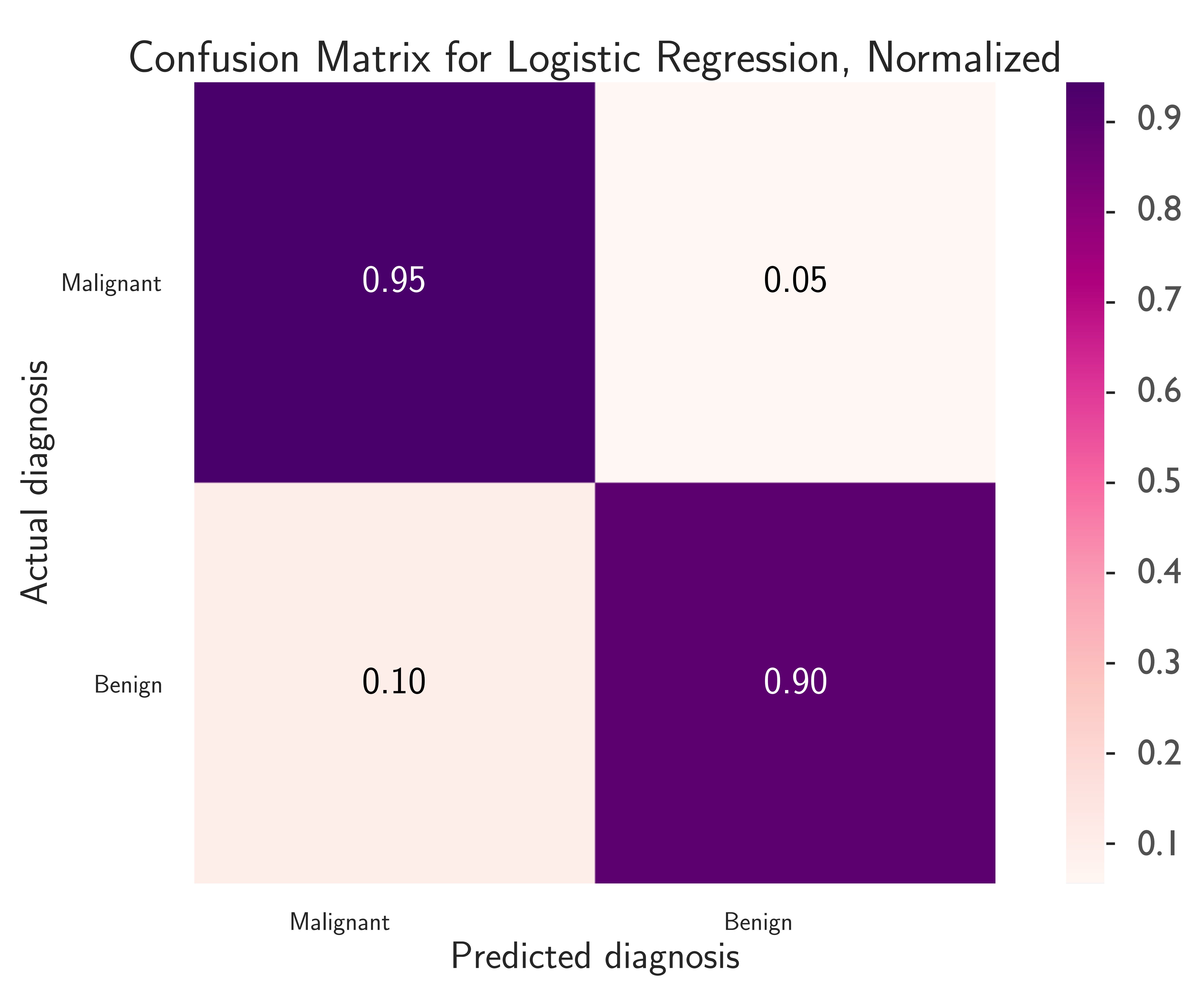
- Image classification — “biopsy to diagnosis”

Appendix



Test set, cross-validation over 5 folds

Appendix

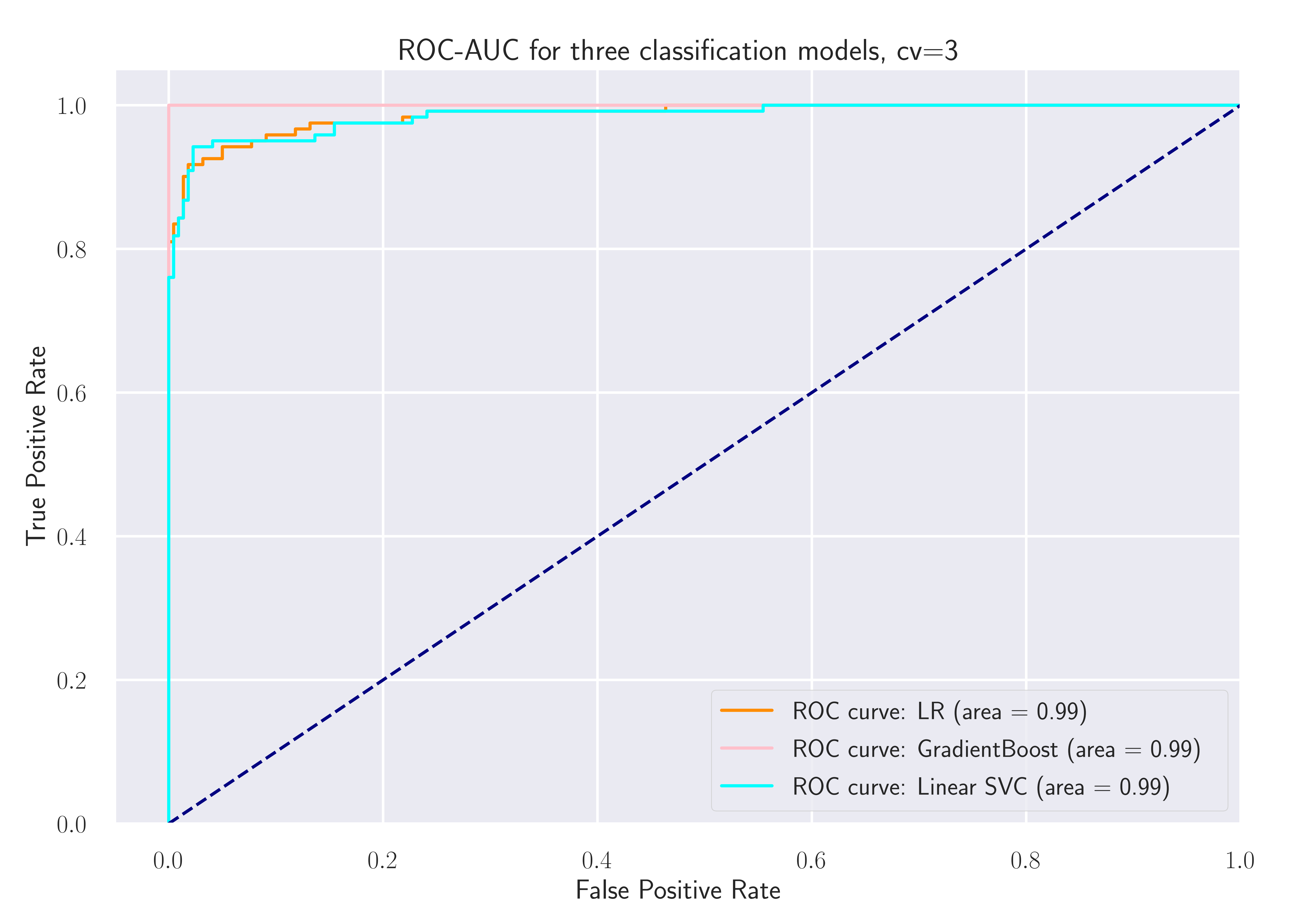


Normalization: divide prediction counts by the sum of each row

Appendix

Using 10 features

LR AUC score validation set 0.9884

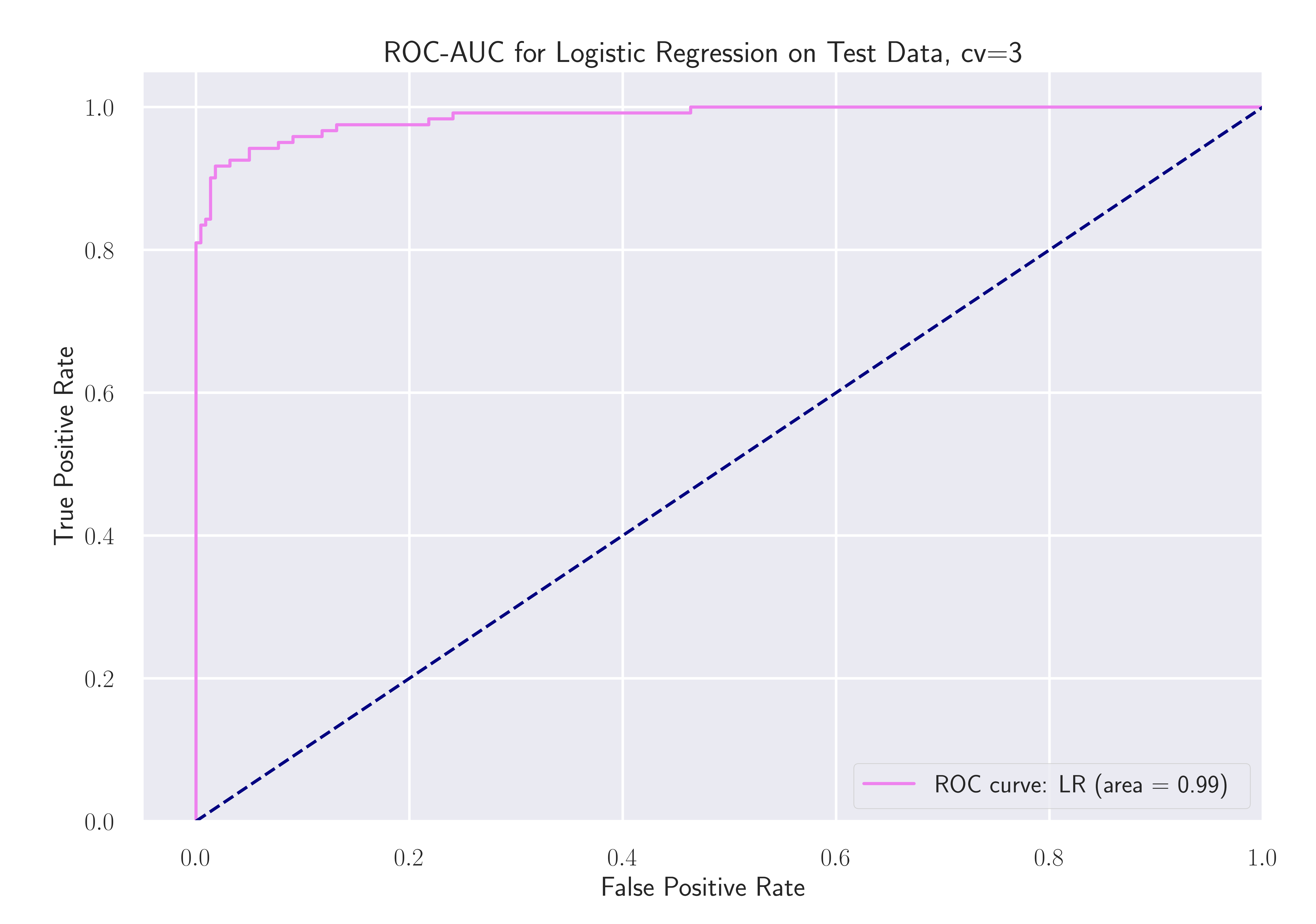


Comparison of the three models with highest roc-auc scores for validation set, with naïve model at 0.5 (dashed line)

Appendix

After feature engineering:
Validation set
0.9909

Test set
0.9870



Area under the curve for Logistic Regression on test data set after feature engineering, with naïve model at 0.5 (dashed line)

Appendix

Model	Metric	Top features	Weights (not normalized, 4 decimals)
LogisticRegression	coefficient	area	3.8253
		concave points	2.7111
		texture	1.6457
SVM - linear SVC	coefficient	area	2.4970
		concave points	1.2092
		texture	1.0927
RandomForest	feature importance	concave points	0.4975
		area	0.3071
		compactness	0.1300

Table showing the most impactful features for predicting diagnosis, for the models with the top roc-auc scores, after feature engineering. **Area** and **concave points** are the two most important features for our prediction.

Appendix

Appendix

Multicollinearity

Appendix

Multicollinearity

- Many features are strongly correlated (area, radius, perimeter; concavity and concave points)

Appendix

Multicollinearity

- Many features are strongly correlated (area, radius, perimeter; concavity and concave points)
  - leads to misleading feature importance values

Appendix

Multicollinearity

- Many features are strongly correlated (area, radius, perimeter; concavity and concave points)
  - leads to misleading feature importance values
- Decided to feature engineer and re-ran the models with fewer features

Appendix

Multicollinearity

- Many features are strongly correlated (area, radius, perimeter; concavity and concave points)
  - leads to misleading feature importance values
- Decided to feature engineer and re-ran the models with fewer features
- The metric for model selection did not vary much before and after feature engineering, suggesting those features did not have a big impact on our predictions

Appendix

AUC score for LR on validation set (10 features): 0.9884
LR on test set (after feature engineering, 5 features): 0.9870

Multicollinearity

- Many features are strongly correlated (area, radius, perimeter; concavity and concave points)
  - leads to misleading feature importance values
- Decided to feature engineer and re-ran the models with fewer features
- The metric for model selection did not vary much before and after feature engineering, suggesting those features did not have a big impact on our predictions