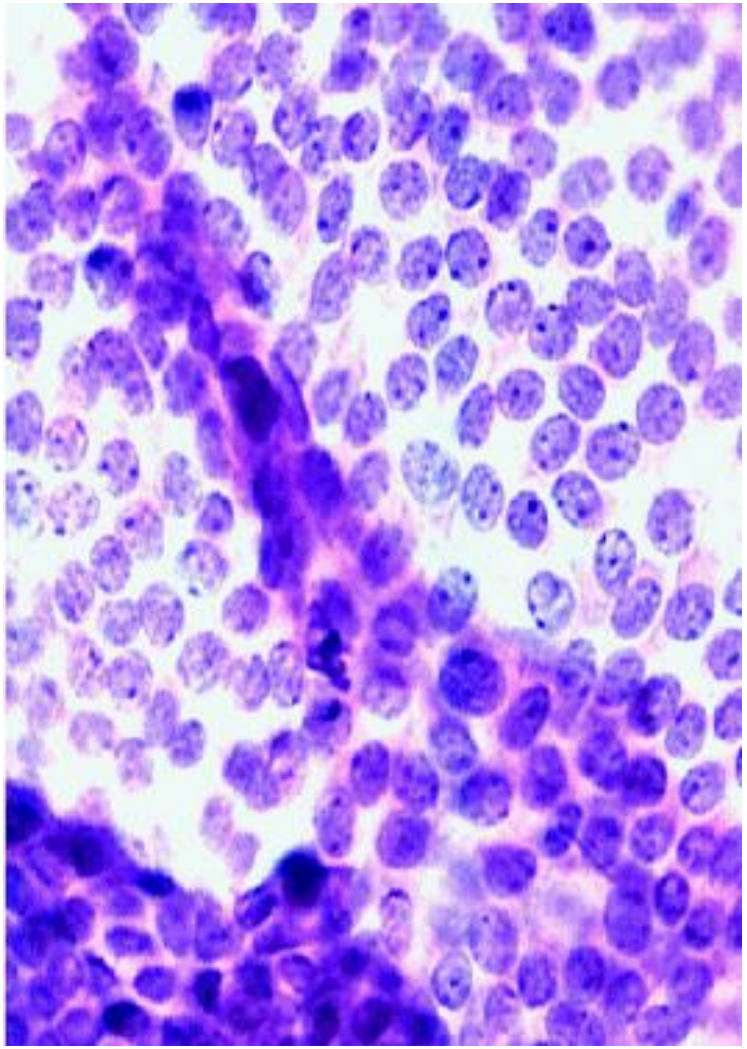
Predicting Breast Cancer Diagnosis

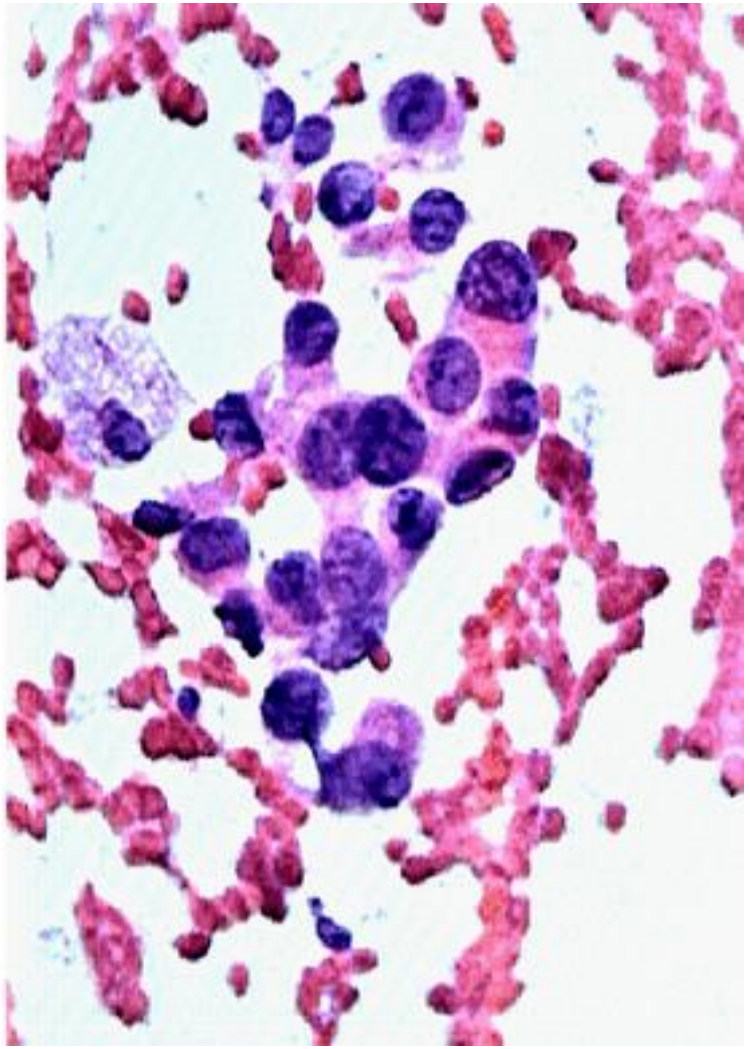
Rachel Khoo

Thinkful Capstone 2

September 2020

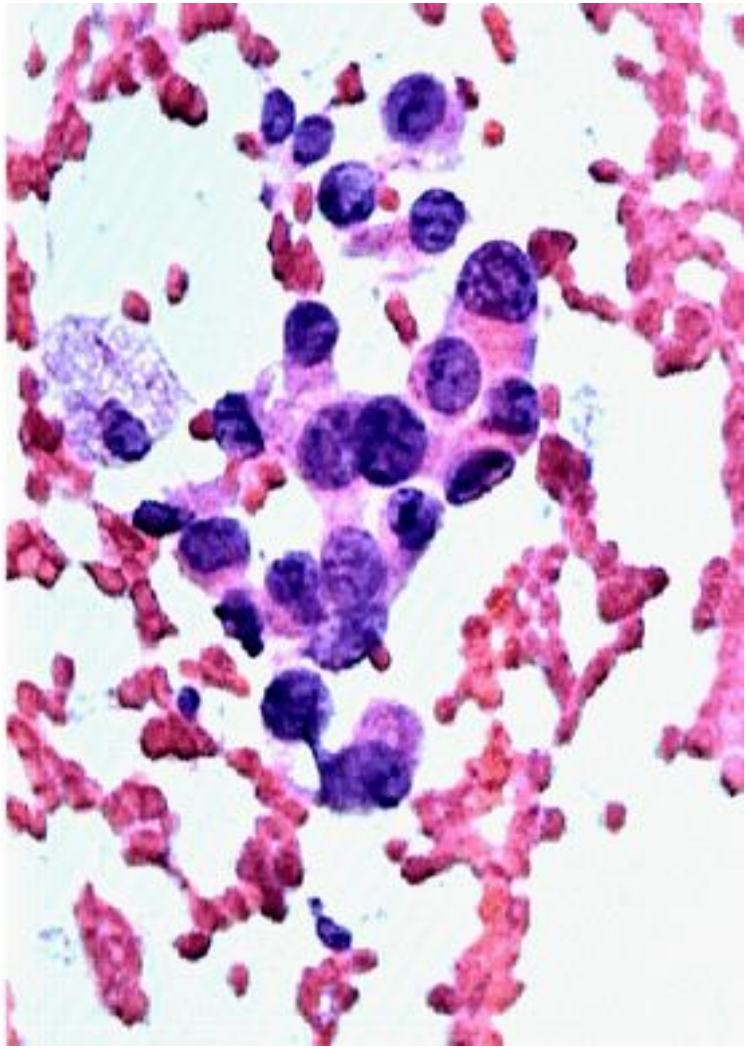
Who has cancer?





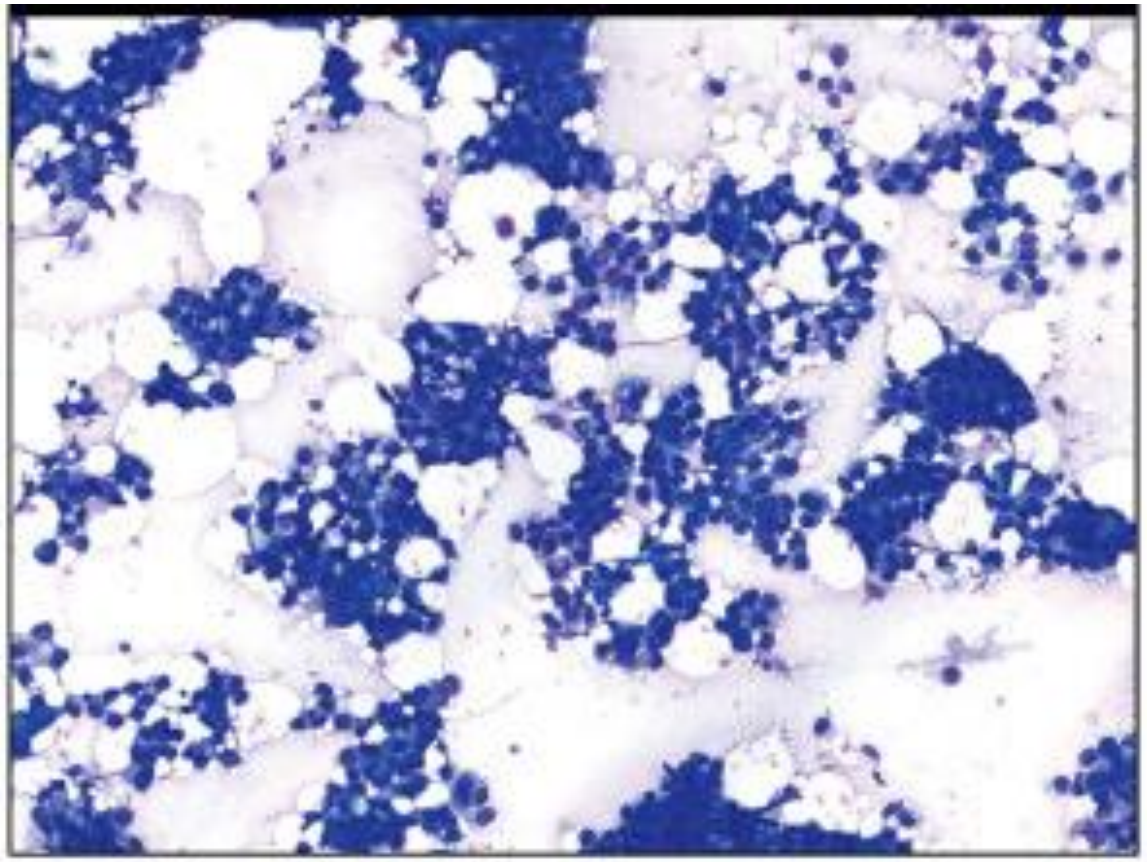
Eickhoff, Carsten. (2014). Crowd-powered experts: helping surgeons interpret breast cancer images. ACM International Conference Proceeding Series. 53-56. 10.1145/2594776.2594788.

Who has cancer?



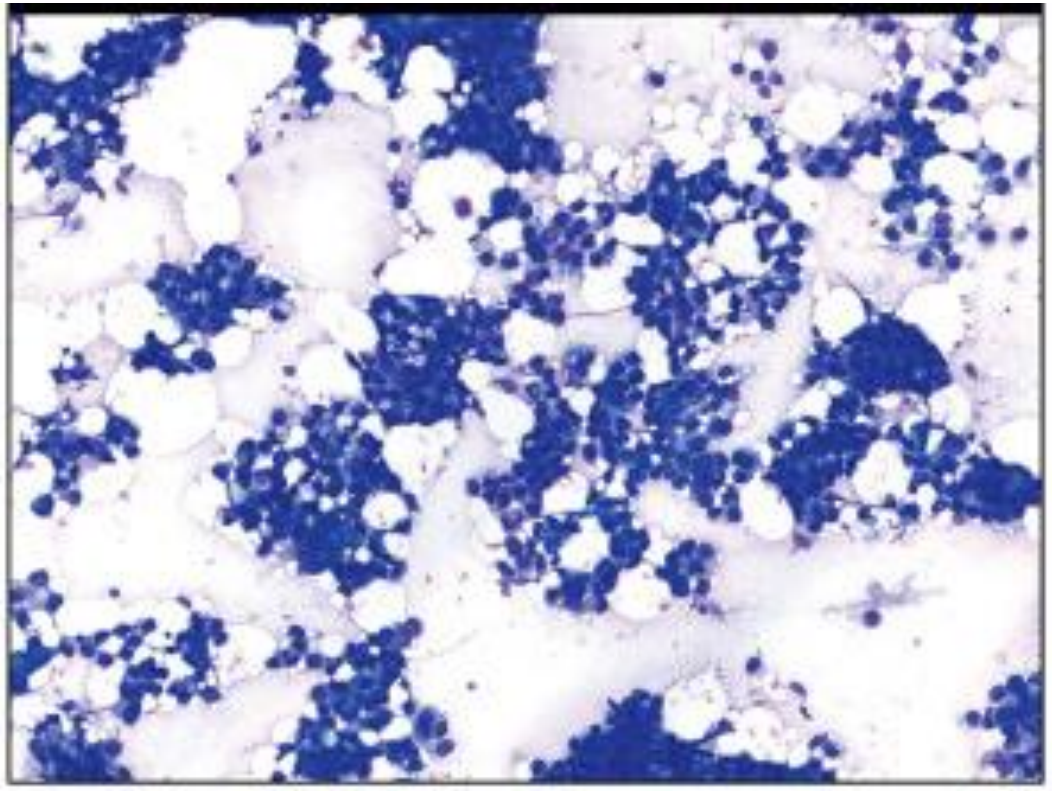
Eickhoff, Carsten. (2014). Crowd-powered experts: helping surgeons interpret breast cancer images. ACM International Conference Proceeding Series. 53-56. 10.1145/2594776.2594788.

Who has cancer?



https://www.captodayonline.com/cytopathology-standardized-reporting-breast-fnab-cytology/

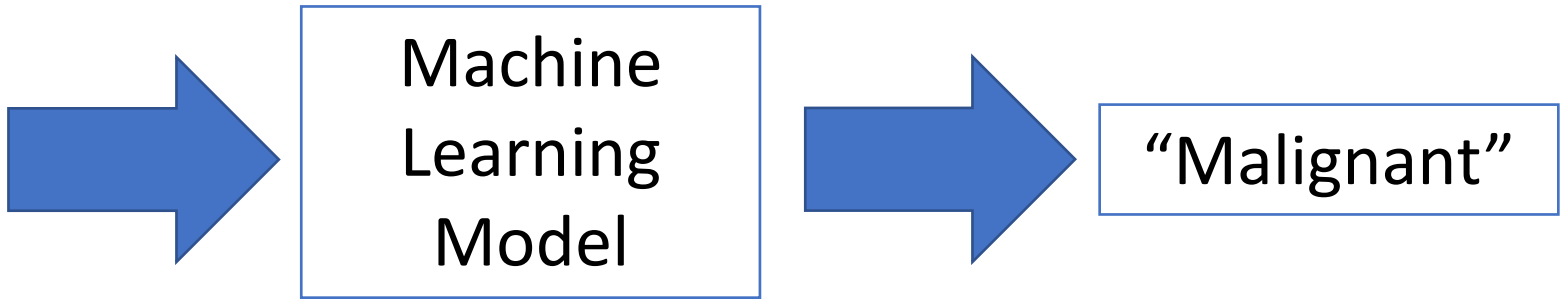
Machine Learning can make diagnosis easier



- Area: 1001.0
- Texture: 10.38
- Compactness: 0.27760
- Concavity: 0.3001

Machine Learning can make diagnosis easier

- Area: 1001.0
- Texture: 10.38
- Compactness: 0.27760
- Concavity: 0.3001

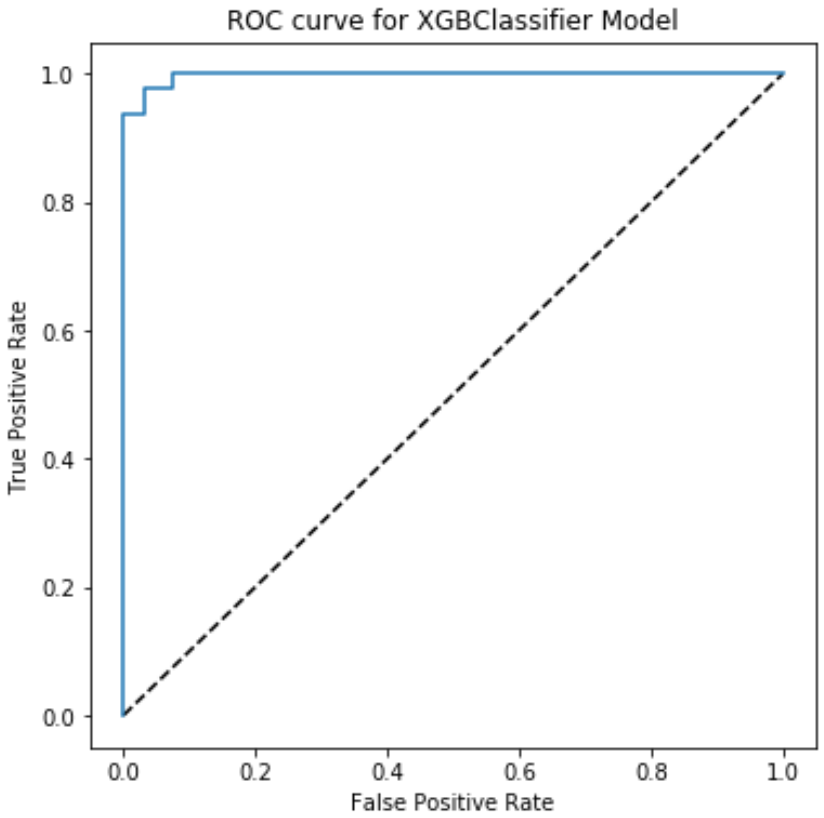


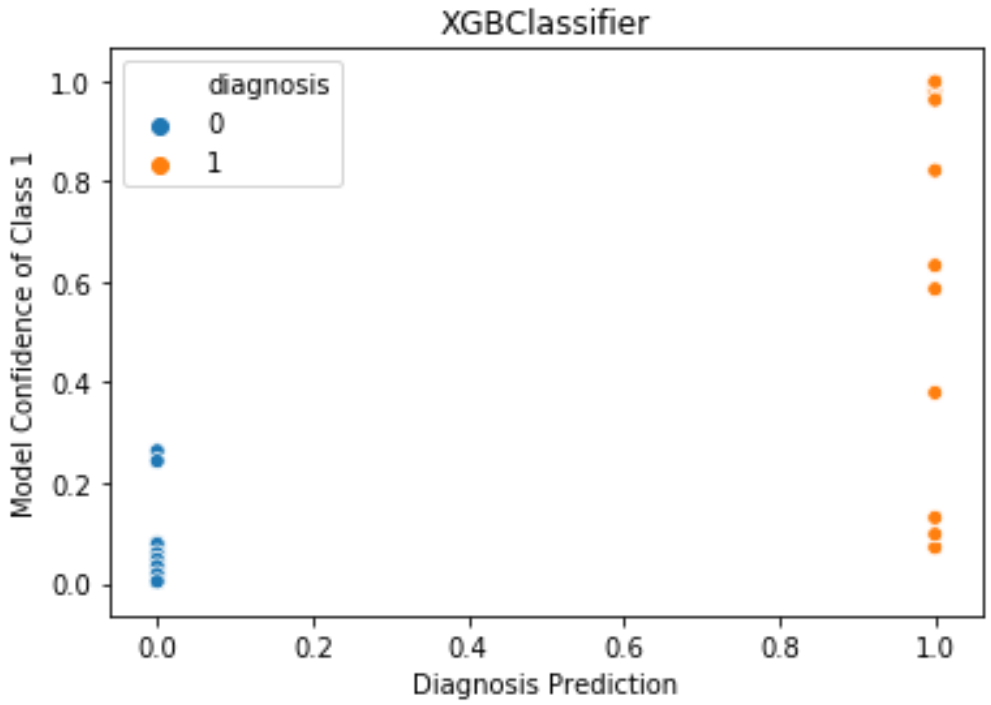
XGBClassifier is the best model

- Accuracy: 95%
- Recall: 96%

	Predicted Benign	**Predicted Malignant**
True Benign	68	4
True Malignant	2	40

XGBClassifier is confident and accurate



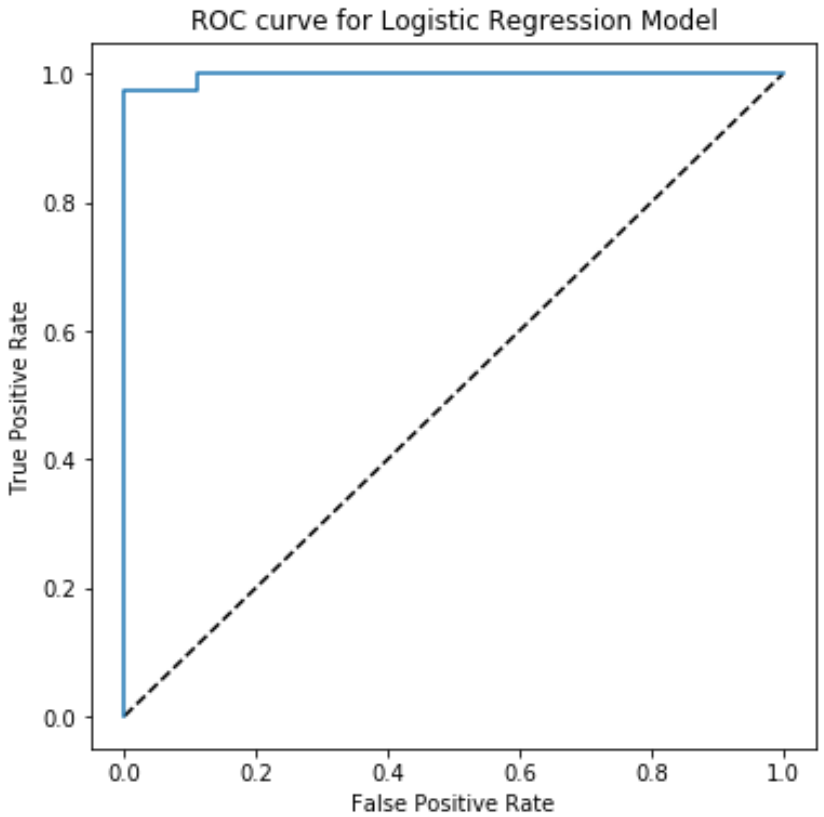


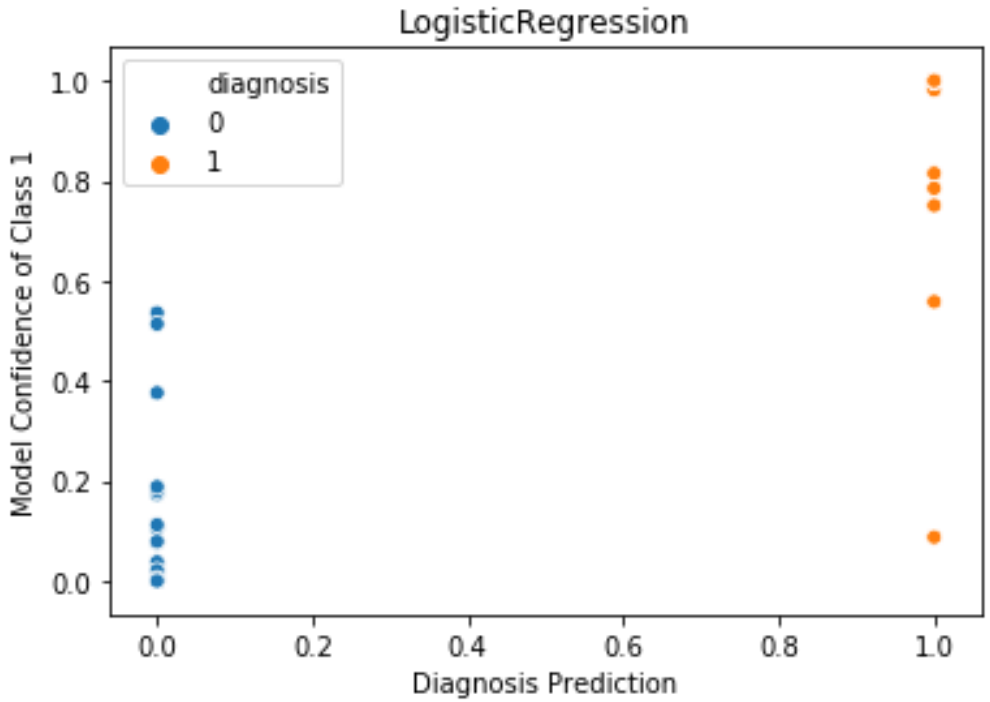
Logistic Regression: Accuracy isn't everything

- Accuracy: 96%
- Recall: 93%

	Predicted Benign	**Predicted Malignant**
True Benign	71	1
True Malignant	3	39

Logistic Regression is more confidently wrong





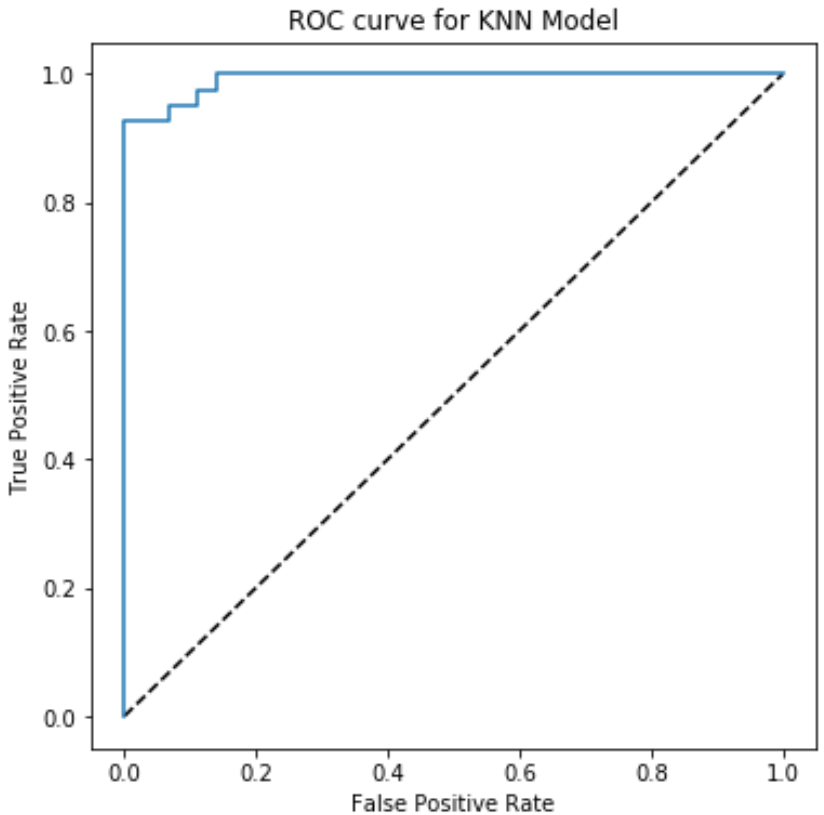
K Nearest Neighbors (KNN)

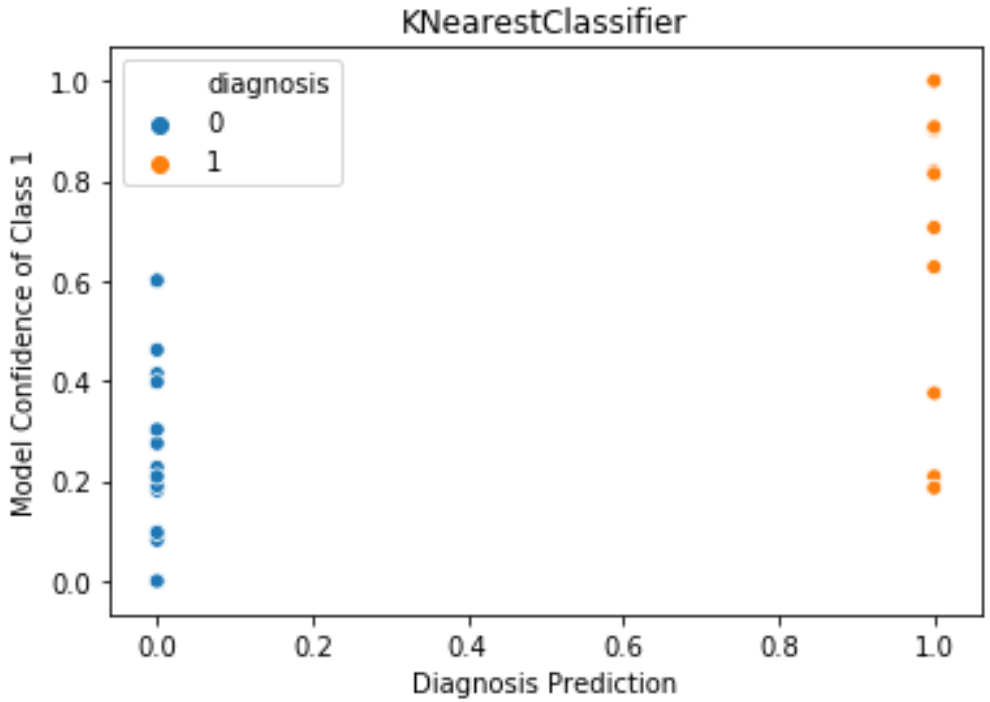
- Accuracy: 96%

- Recall: 91%

	Predicted Benign	**Predicted Malignant**
True Benign	71	1
True Malignant	4	38

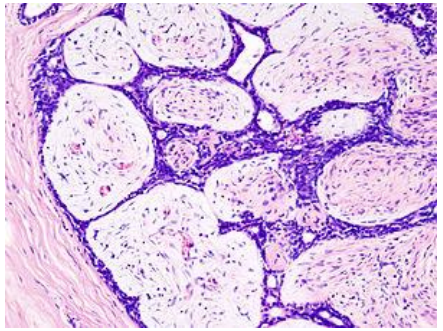
KNN


ROC curve for KNN Model
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0


KNearestClassifier
diagnosis
0
1
Model Confidence of Class 1
Diagnosis Prediction
0.0
0.2
0.4
0.6
0.8
1.0

Every model has limitations

- XGB
  - Can’t predict outside of sample
- Logistic Regression
  - Can be slow
- KNN
  - Slower
  - Not easily interpretable
  - Can’t predict outside of sample



XGBClassifier

?

How can we improve accuracy?

- More data
- Spend more time tuning hyperparameters
- PCA to reduce complexity redundancy